The Storage Hierarchy

Storage Technology and Trends

Module Outline

- **Storage Technologies and Trends**
  - **RAM**
  - **Nonvolatile Memories**
    - **Hard Disk Drives**
    - **Solid State Drives**
- **Module Summary**

IBM 350 Storage Unit

Storage Technologies and Trends

Random-Access Memory (RAM)

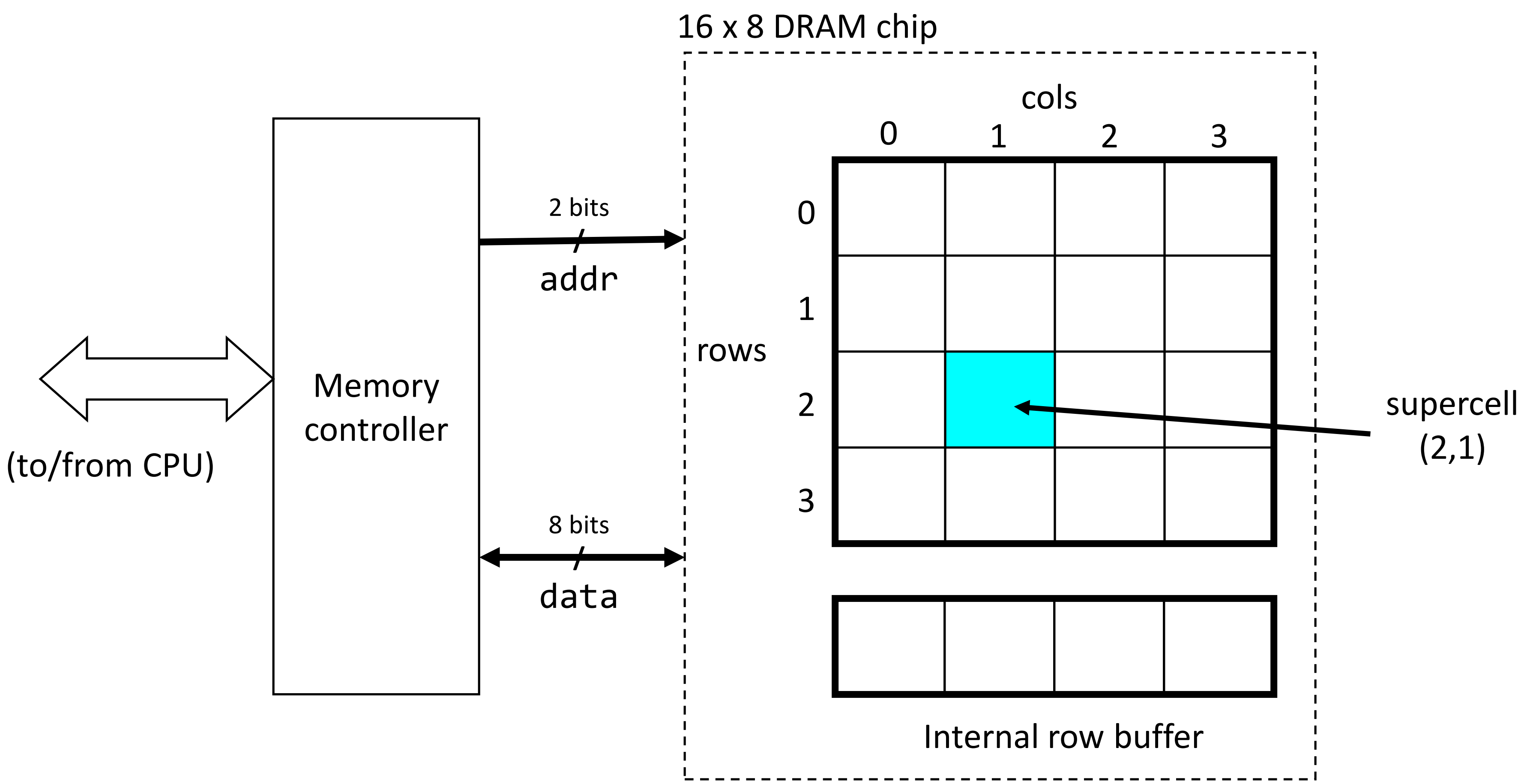
- Key features
  - RAM is traditionally packaged as a chip.
  - Basic storage unit is normally a cell (one bit per cell).
  - Multiple RAM chips form a memory.
- Static RAM (SRAM)
  - Each cell stores a bit with a four or six-transistor circuit.
  - Retains value indefinitely, as long as it is kept powered.
  - Relatively insensitive to electrical noise (EMI), radiation, etc.
  - Faster and more expensive than DRAM.
- Dynamic RAM (DRAM)
  - Each cell stores bit with a capacitor. One transistor is used for access
  - Value must be refreshed every 10-100 ms.
  - More sensitive to disturbances (EMI, radiation,…) than SRAM.
  - Slower and cheaper than SRAM.

SRAM vs DRAM Comparison

	Transistors per bit	Access time	Needs refresh?	Needs EDC?	Cost	Applications
SRAM	4 or 6	1X	No	Maybe	100x	Cache memories
DRAM	1	10X	Yes	Yes	1X	Main memories, frame buffers

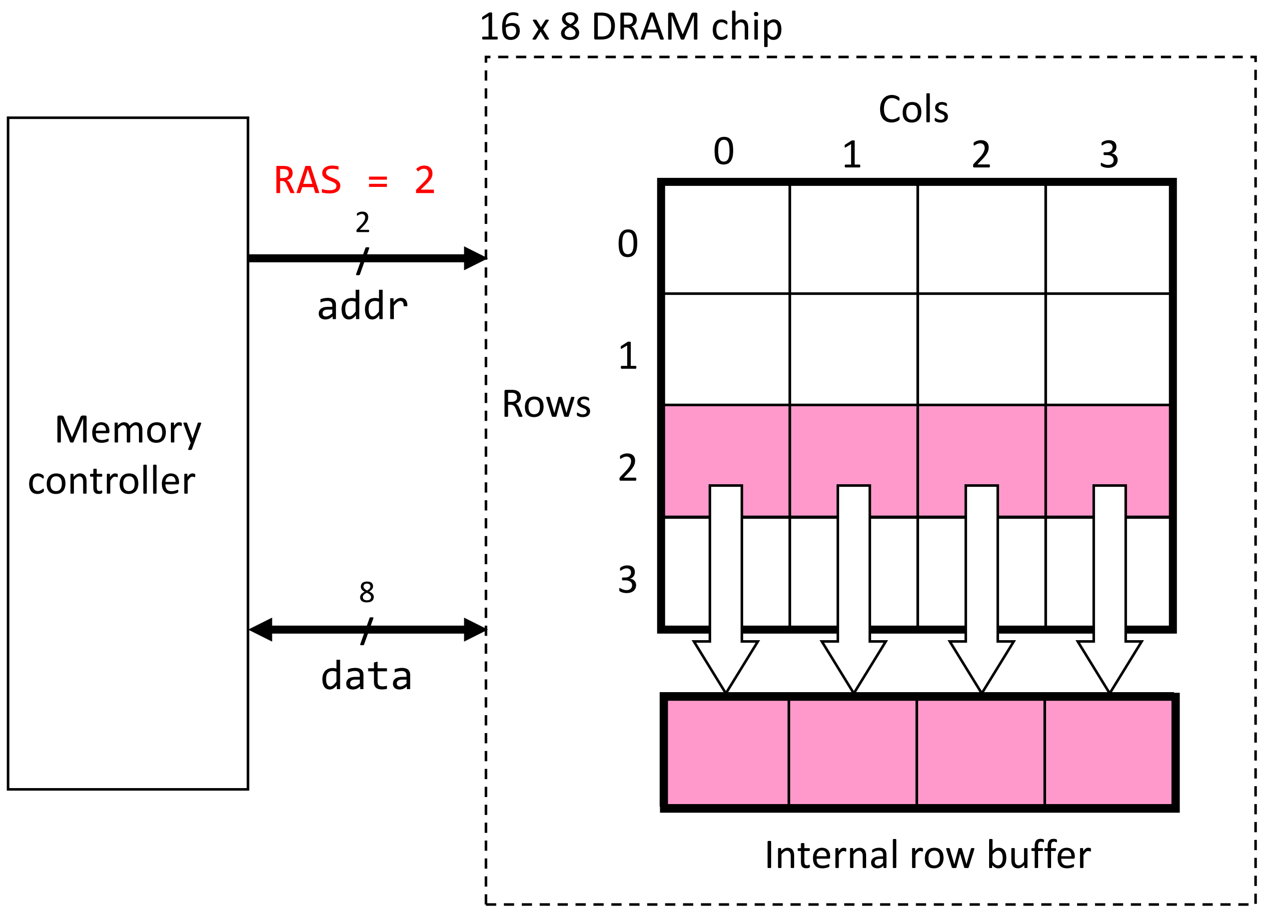
Conventional DRAM Organization

- *d* x *w* DRAM:
  - *dw* total bits organized as *d* supercells of size *w* bits



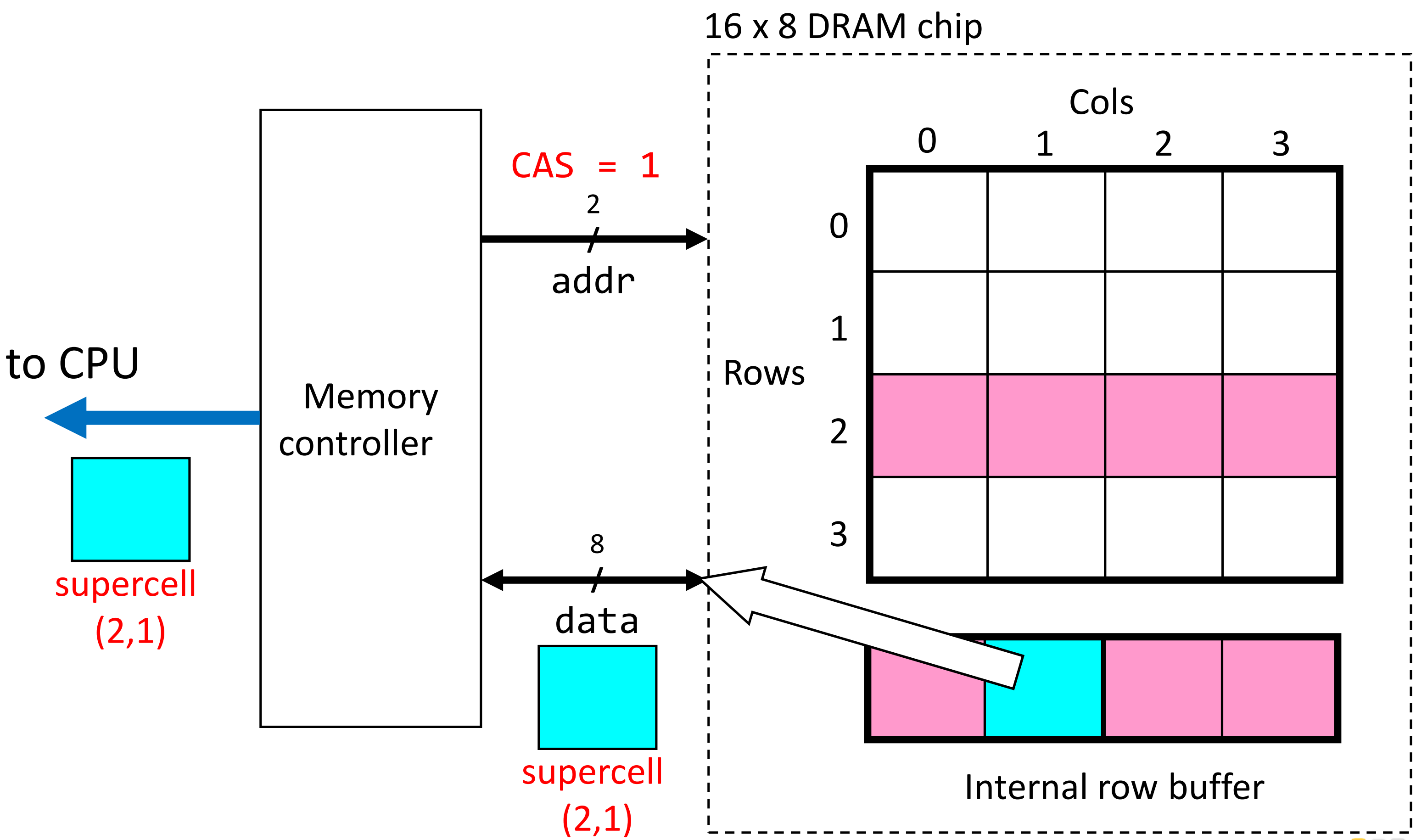
Reading DRAM Supercell (2,1)

- Step 1(a): Row access strobe (RAS) selects row 2.
- Step 1(b): Row 2 copied from DRAM array to row buffer.

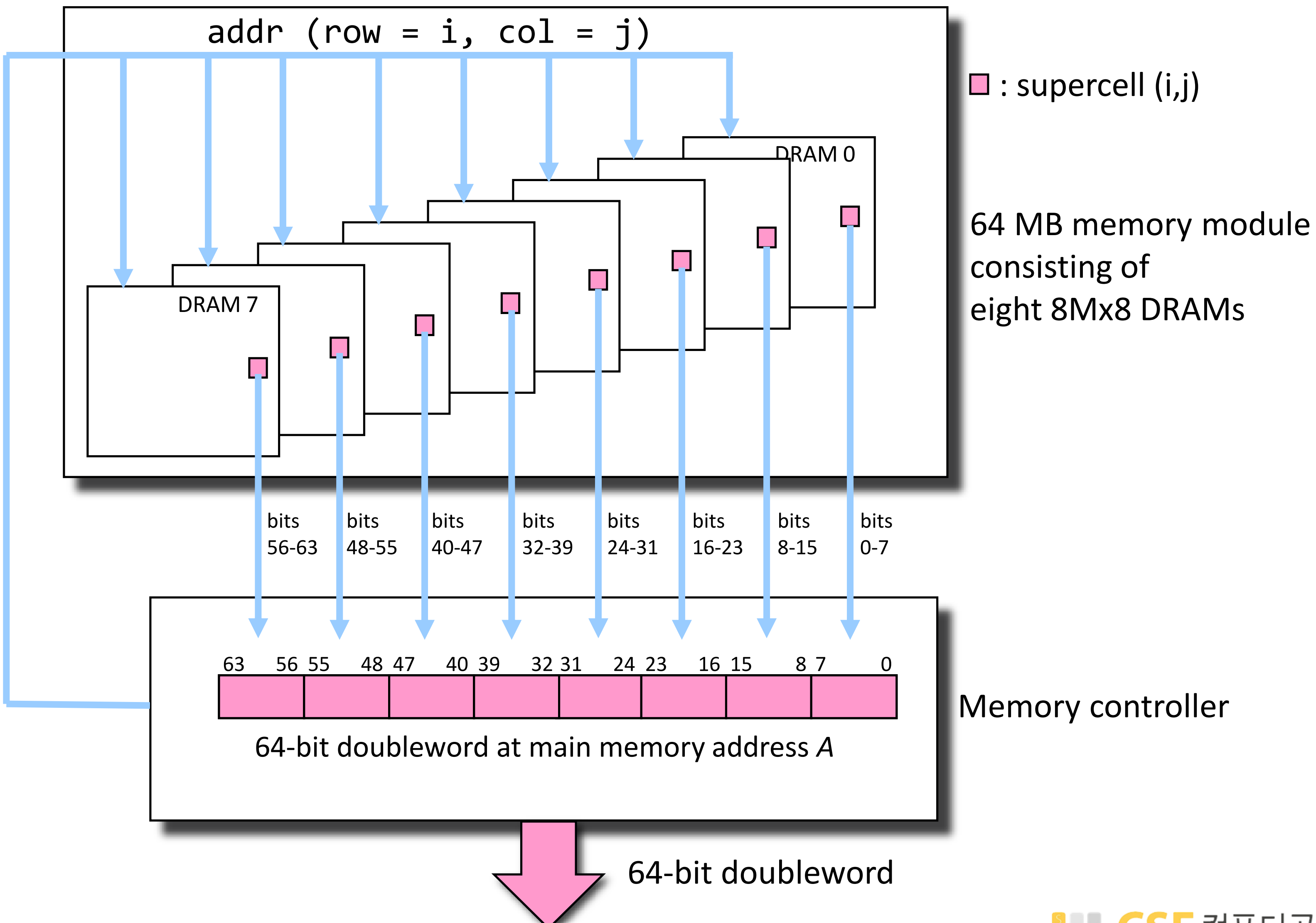


Reading DRAM Supercell (2,1)

- Step 2(a): Column access strobe (CAS) selects column 1.
- Step 2(b): Supercell (2,1) copied from buffer to data lines, and eventually back to the CPU.



Memory Modules



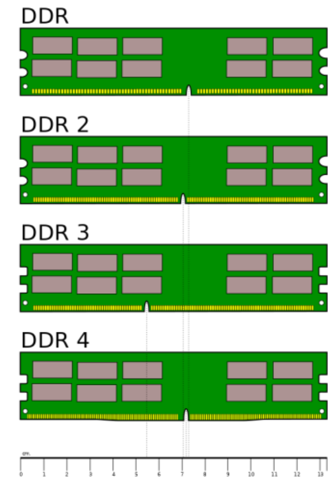
Enhanced DRAMs

- Basic DRAM cell has not changed since its invention in 1966.
  - Commercialized by Intel in 1970.

- DRAM cores with better interface logic and faster I/O :
  - Synchronous DRAM (SDRAM)
    - Uses a conventional clock signal instead of asynchronous control
    - Allows reuse of the row addresses (e.g., RAS, CAS, CAS, CAS)

  - Double data-rate synchronous DRAM (DDR SDRAM)
    - Double edge clocking sends two bits per bus cycle and pin
    - Different types distinguished by internal speed multiplier: DDR (2-fold), DDR2 (4-fold), DDR3/4 (8-fold), DDR5 (16-fold)
    - By 2010, standard for most server and desktop systems
    - Intel Core i7 supports only DDR3 SDRAM

source: [Wikipedia]



DDRX-Y?

- DDRX-Y
  - X: generation, Y: transfer rate (MT/s)
  - LP-DDR: Low Power DDR ([separate standard])

Gen	Year	Standards	Chip clock rate (MHz)	Data rate (MT/s)	Voltage (V)	Peak transfer rate (MB/s)	Module name
DDR1	2000	DDR−200 DDR−266 DDR−333 DDR−400	100 133 167 200	200 267 333 400	2.5 2.6	1600 2133 2667 3200	PC−1600 PC−2100 PC−2700 PC−3200
DDR2	2003	DDR2−400 DDR2−667 DDR2−800 DDR2−1066	100 167 200 267	400 667 800 1067	1.8	3200 5333 6400 8533	PC2−3200 PC2−5300 PC2−6400 PC2−8500
DDR3	2007	DDR3−800 DDR3−1600 DDR3−1886 DDR3−2133	100 200 233 266	800 1600 1867 2133	1.35	6400 12800 14933 17067	PC3−6400 PC3−12800 PC3−14900 PC3−17000
DDR4	2014	DDR4−1600 DDR4−2666 DDR4−3000 DDR4−3200	200 333 375 400	1600 2667 3000 3200	1.2	12800 20800 24000 25600	PC4−12800 PC4−20800 PC4−24000 PC4−25600
DDR5	2020	DDR5−3200 … DDR5−8000	200 … 500	3200 … 8000	1.1	25600 … 64000	PC5−25600 … PC5−64000

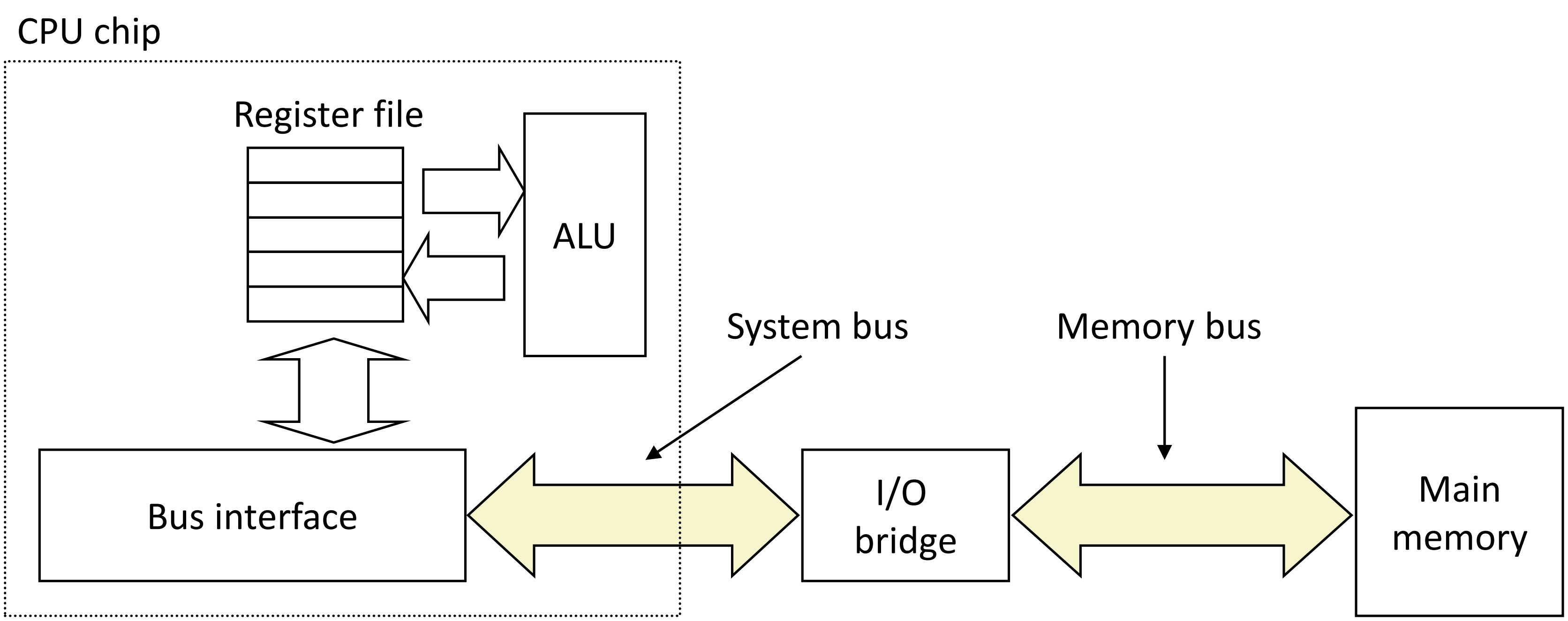
☞ [DDR4 memory on danawa.com]

Nonvolatile Memories

- DRAM and SRAM are volatile memories
  - Lose information if powered off.
- Nonvolatile memories retain value even if powered off
  - Read-only memory (ROM): programmed during production
  - Programmable ROM (PROM): can be programmed once
  - Erasable PROM (EPROM): can be bulk erased (UV, X-Ray)
  - Electrically erasable PROM (EEPROM): electronic erase capability
  - Flash memory: EEPROMs with partial (sector) erase capability
    - Wears out after about 100,000 erase operations.
- Uses for Nonvolatile Memories
  - Firmware programs stored in a ROM (BIOS, controllers for disks, network cards, graphics accelerators, security subsystems,…)
  - Solid state disks (replace rotating disks in thumb drives, smart phones, mp3 players, tablets, laptops,…)
  - Disk caches

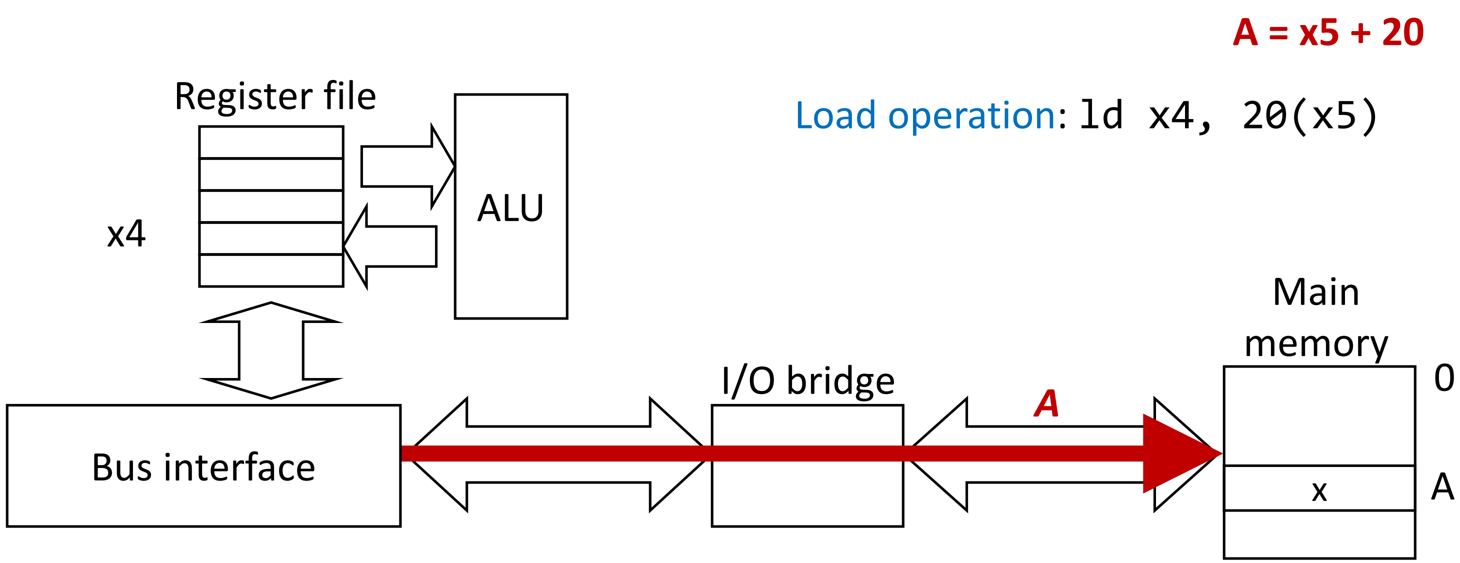
Traditional Bus Structure Connecting CPU and Memory

- A bus is a collection of parallel wires that carry address, data, and control signals.
- Buses are typically shared by multiple devices.



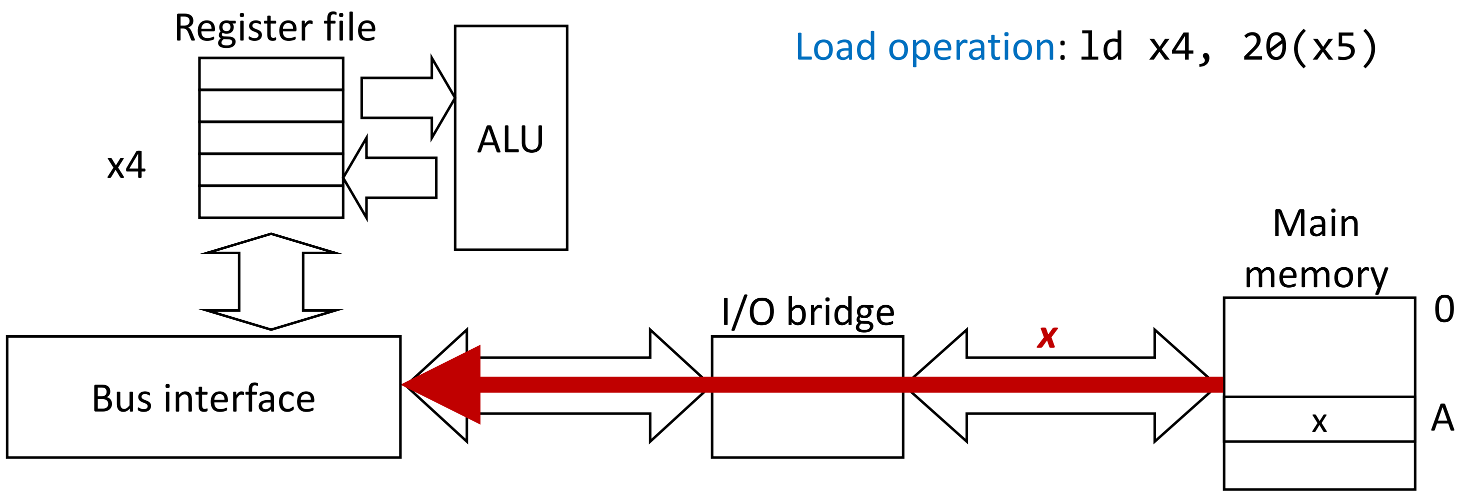
Memory Read Transaction (1)

- CPU computes and places memory address A on the memory bus.



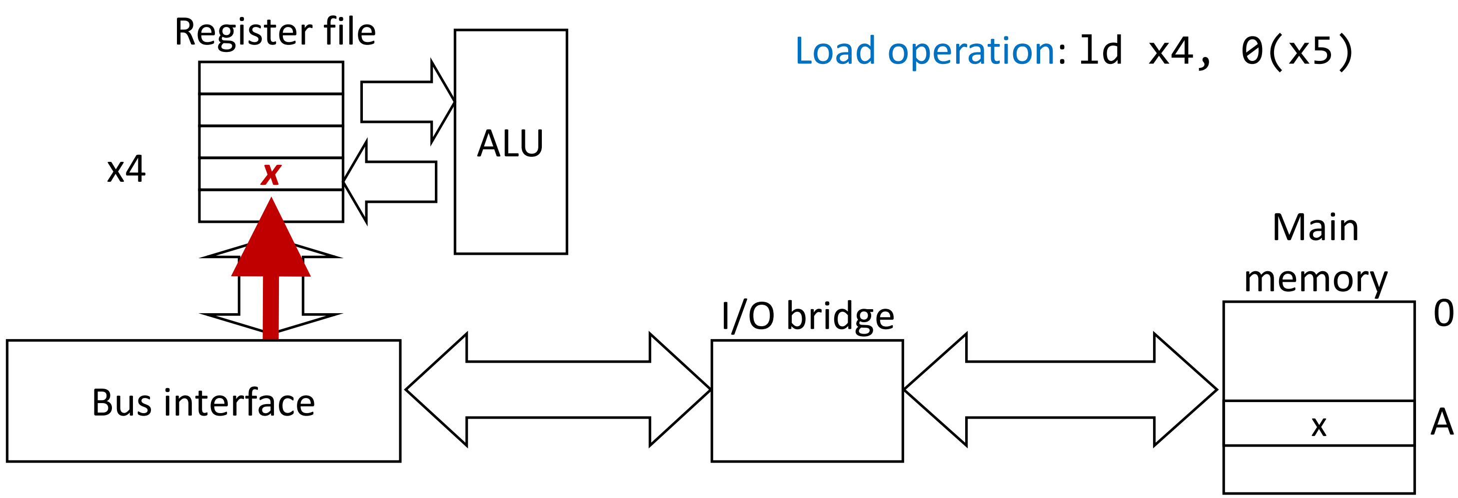
Memory Read Transaction (2)

- Main memory reads A from the memory bus, retrieves word x, and places it on the bus.



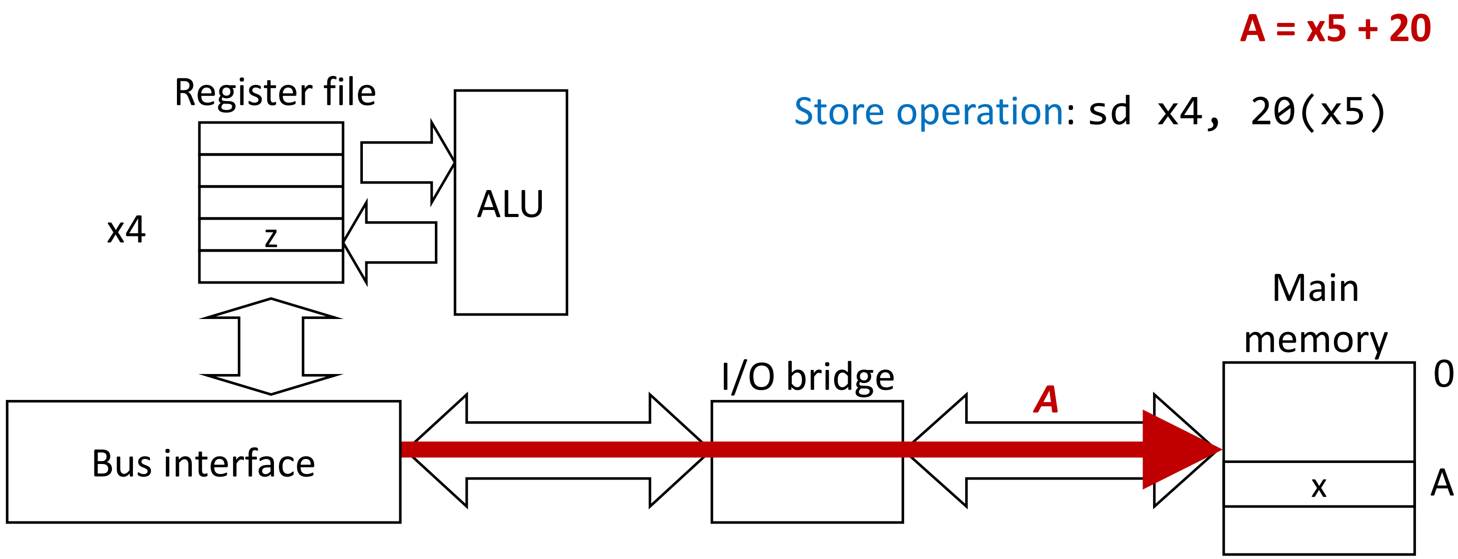
Memory Read Transaction (3)

- CPU read word x from the bus and copies it into register x4.



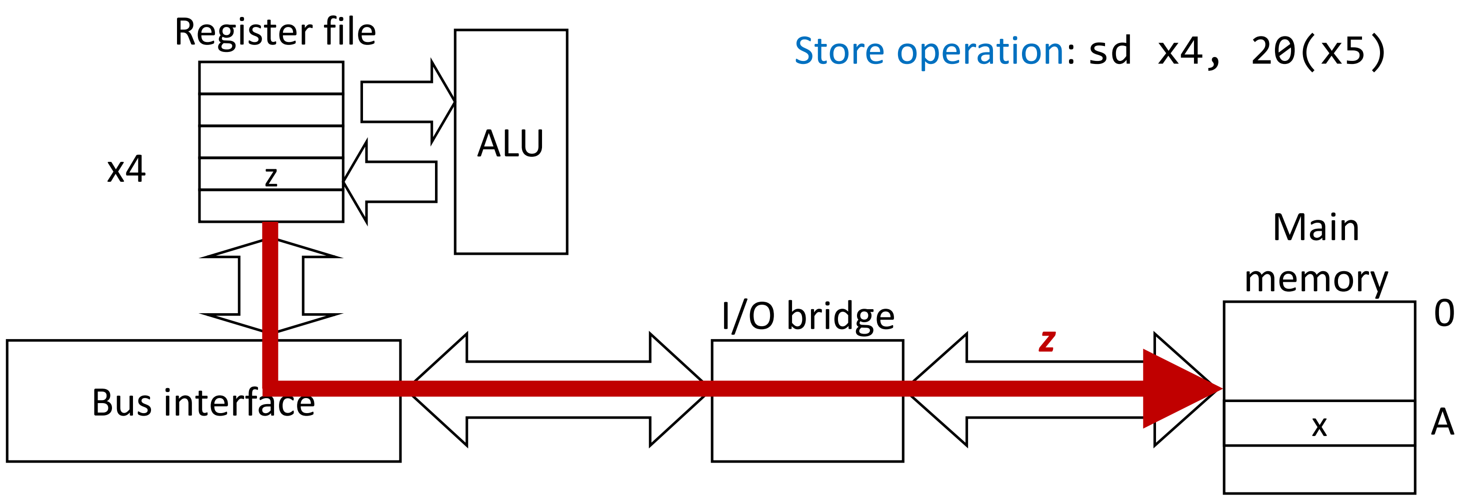
Memory Write Transaction (1)

- CPU places address A on bus. Main memory reads it and waits for the corresponding data word to arrive.



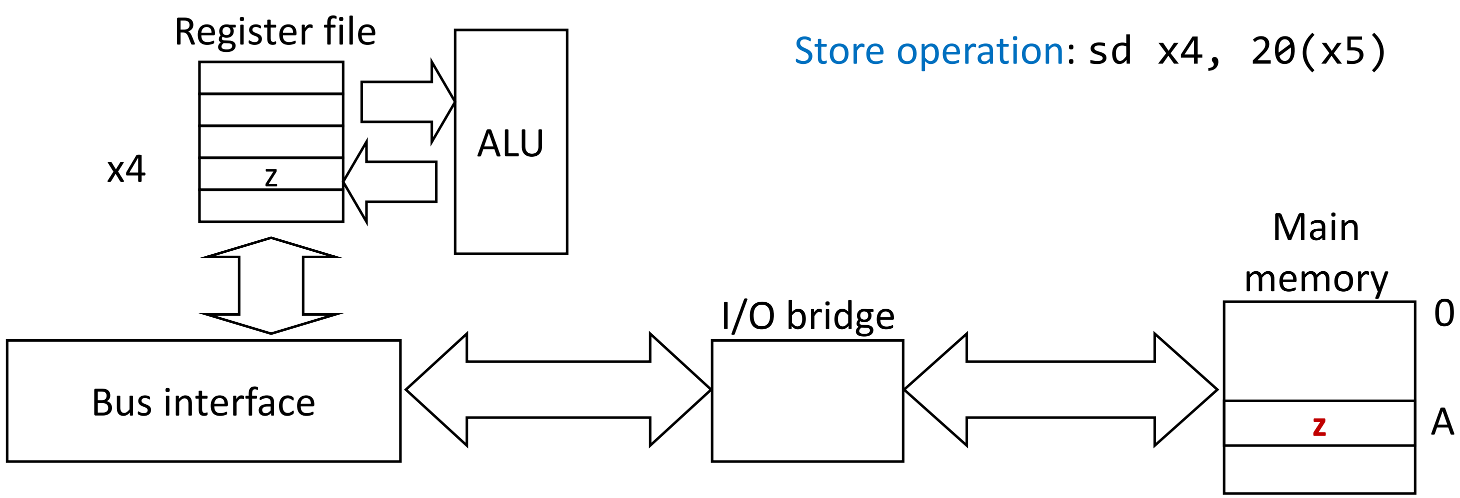
Memory Write Transaction (2)

- CPU places data word *z* on the bus.



Memory Write Transaction (3)

- CPU Main memory reads data word *z* from the bus and stores it at address A.



Hard Disk Drives

Image courtesy of Western Digital

What's Inside A Disk Drive?

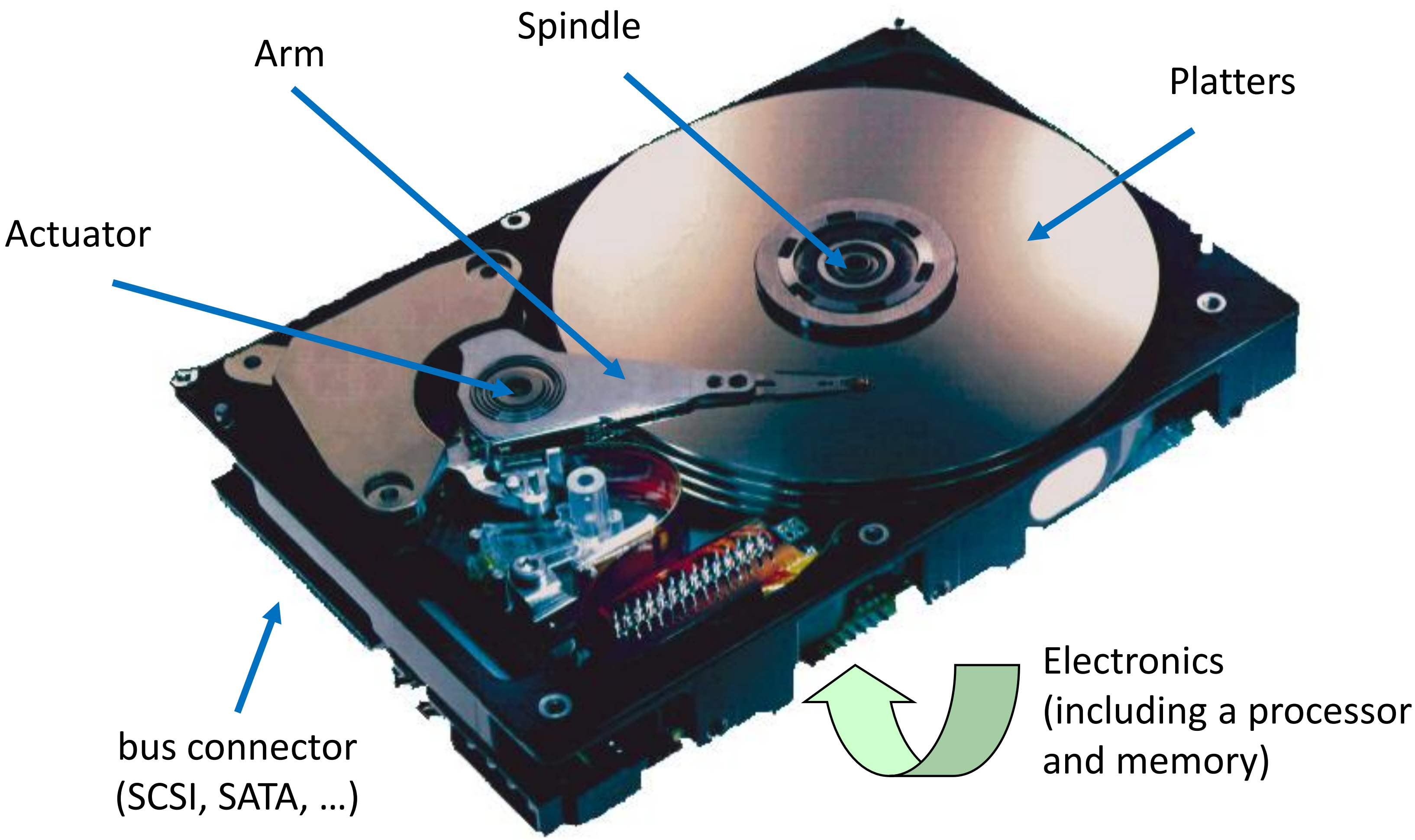
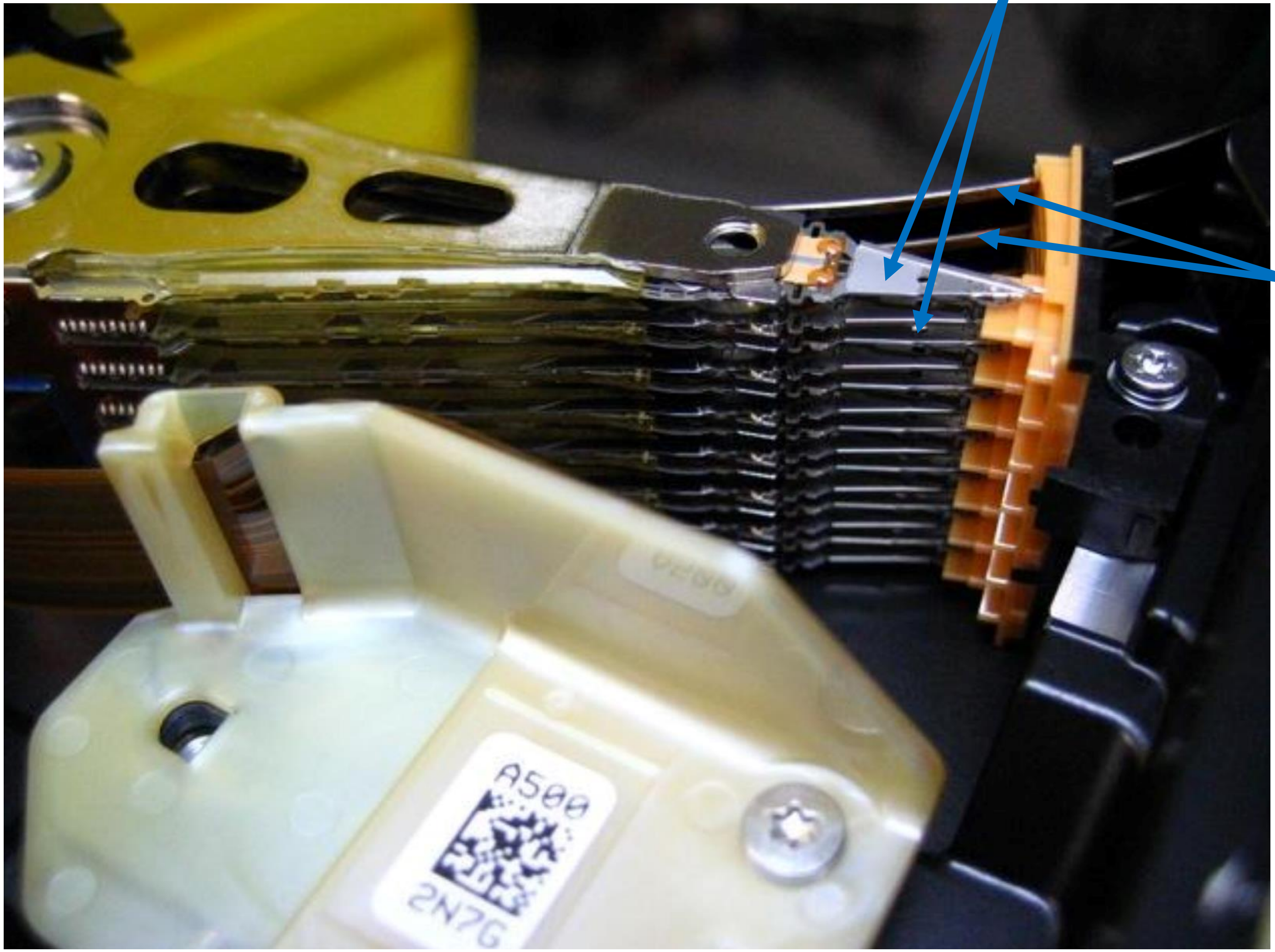


Image courtesy of Seagate Technology

What's Inside A Disk Drive?

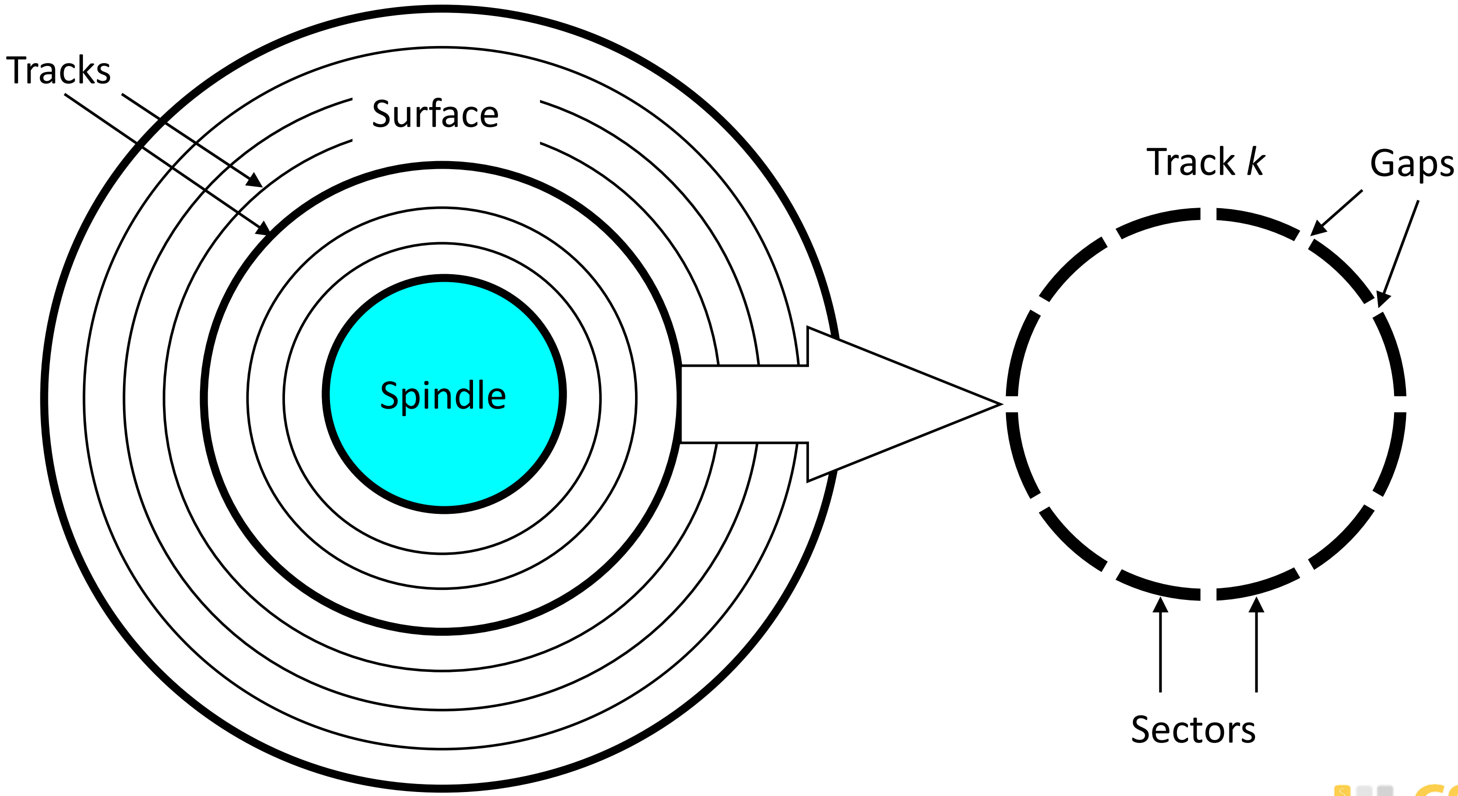
Read/Write Heads



Platters

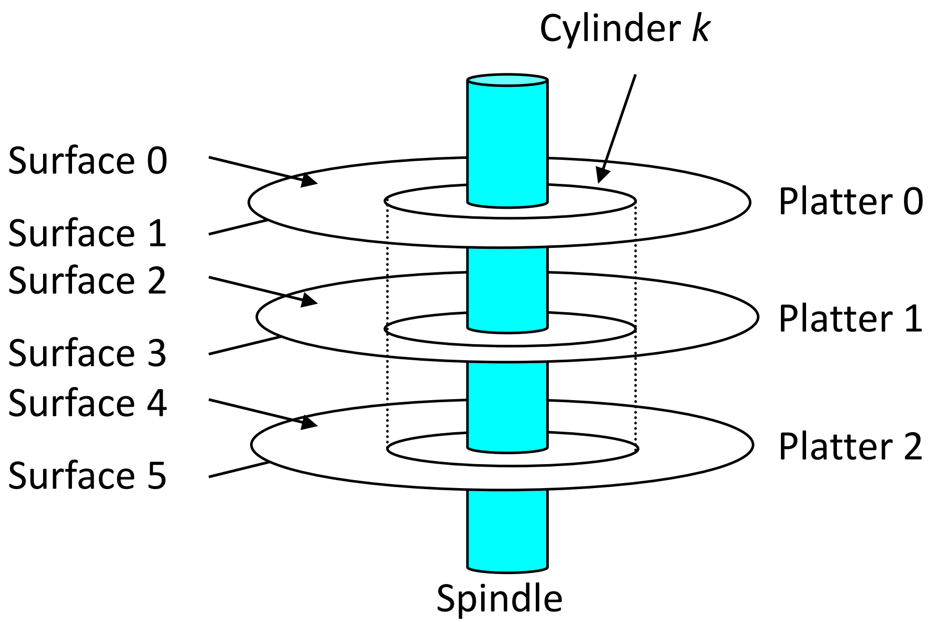
Disk Geometry

- Disks consist of platters, each with two surfaces.
- Each surface consists of concentric rings called tracks.
- Each track consists of sectors separated by gaps.
- Each sector contains an equal number of data bits (typically 512 bytes)



Disk Geometry (Muliple-Platter View)

- Aligned tracks form a cylinder.



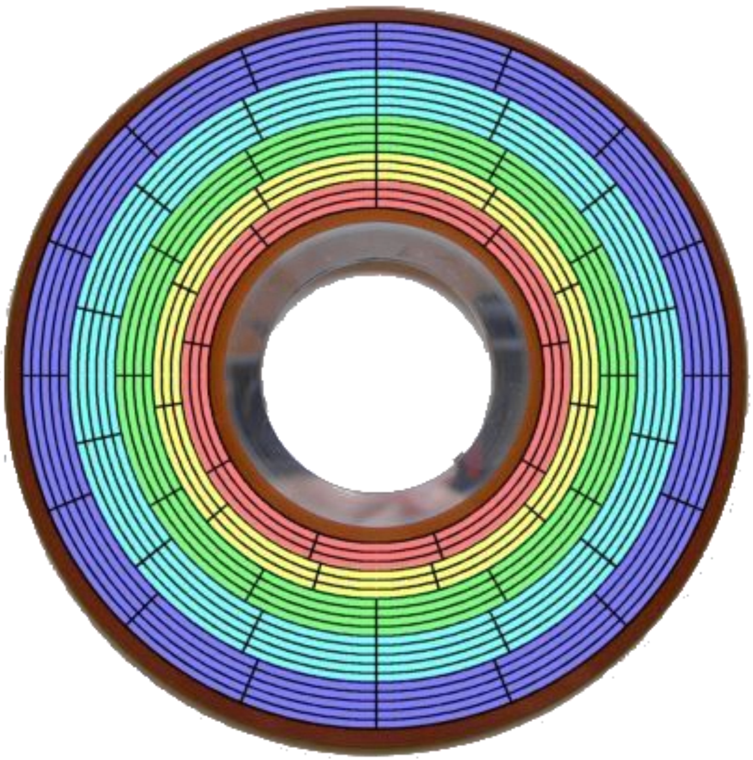
Disk Capacity

- Capacity: maximum number of bits that can be stored.
  - Vendors express capacity in units of gigabytes (GB), where 1 GB = 10^9 Bytes

- Capacity is determined by these technology factors:
  - Recording density (bits/in): number of bits that can be squeezed into a 1 inch segment of a track.
  - Track density (tracks/in): number of tracks that can be squeezed into a 1 inch radial segment.
  - Areal density (bits/in2): product of recording and track density.

Disk Capacity

- Modern disks partition tracks into disjoint subsets called recording zones
  - Each track in a zone has the same number of sectors, determined by the circumference of innermost track.
  - Each zone has a different number of sectors/track



Computing Disk Capacity

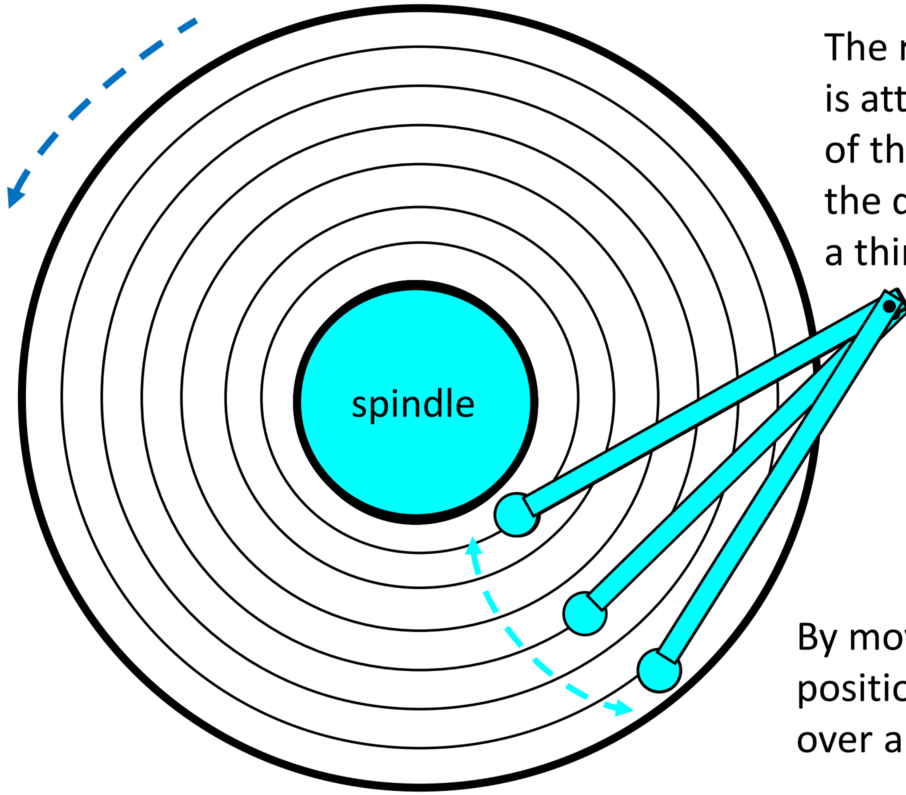
- Capacity = (# bytes/sector) x (avg. # sectors/track) x
  (# tracks/surface) x (# surfaces/platter) x
  (# platters/disk)

- Example:
  - 512 bytes/sector
  - 300 sectors/track (on average)
  - 20,000 tracks/surface
  - 2 surfaces/platter
  - 5 platters/disk

- Capacity = 512 x 300 x 20000 x 2 x 5
  = 30,720,000,000
  = 30.72 GB

Disk Operation (Single-Platter View)

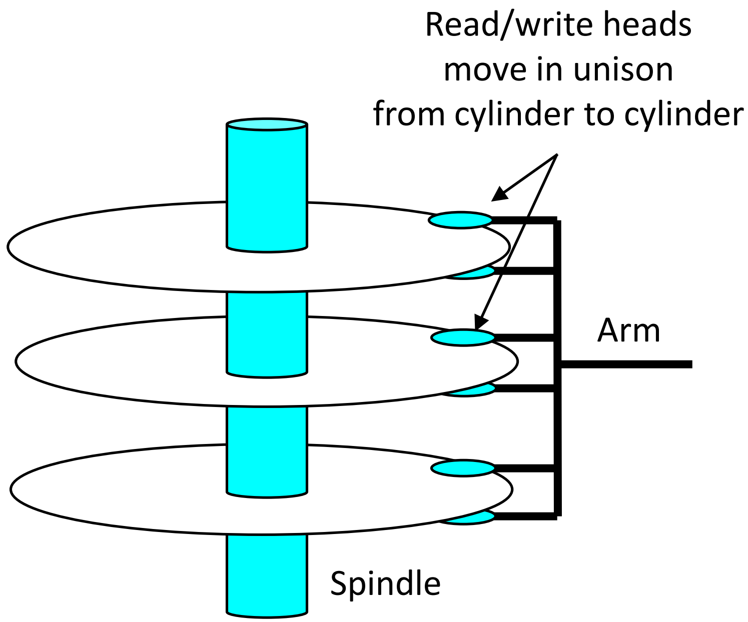
The disk surface spins at a fixed rotational rate



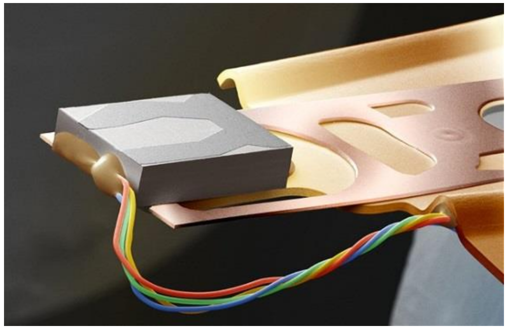
The read/write *head* is attached to the end of the *arm* and flies over the disk surface on a thin cushion of air.

By moving radially, the arm can position the read/write head over any track (*seek*).

Disk Operation (Multi-Platter View)



Read/Write Head



human hair
75 micron

smoke particle
2.5 micron

dust particle
4 micron

disk surface

surface - disk head
0.1 micron
80km/h

$PM_{2.5}$

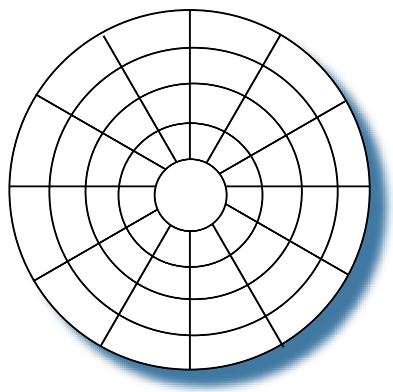
초미세먼지
<= 2.5 micron

PM_{10}

미세먼지
<= 10 micron

CSE 컴퓨터공학부
Department of Computer Science & Engineering

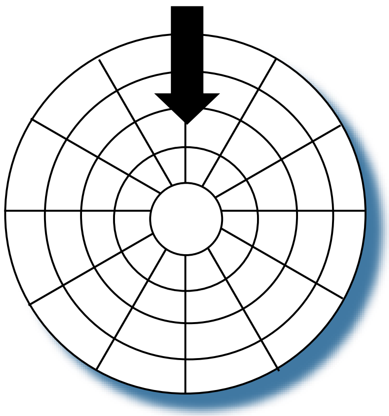
Disk Structure - top view of single platter



Surface organized into tracks

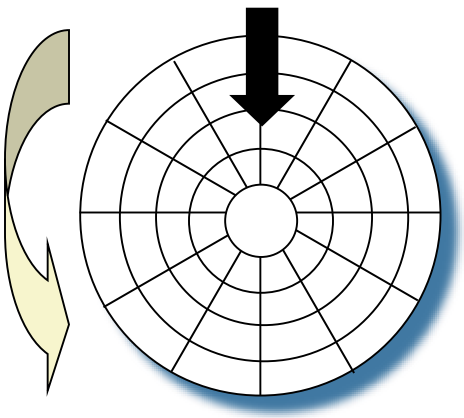
Tracks divided into sectors

Disk Access



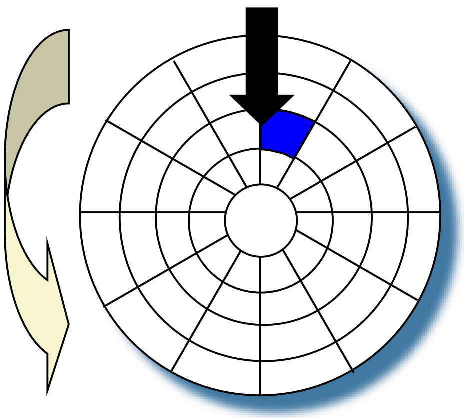
Head in position above a track

Disk Access



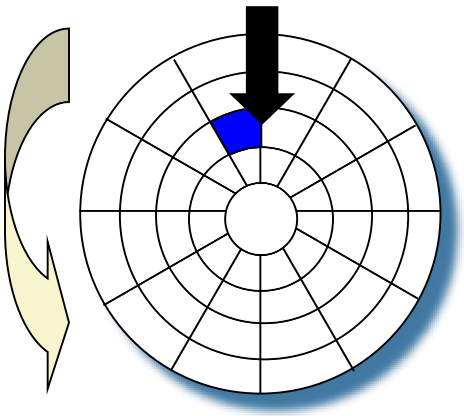
Rotation is counter-clockwise

Disk Access – Read



About to read blue sector

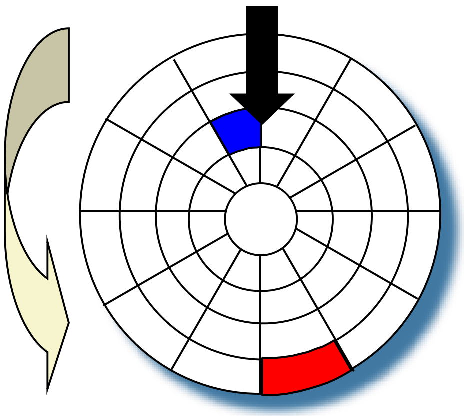
Disk Access – Read



After BLUE read

After reading blue sector

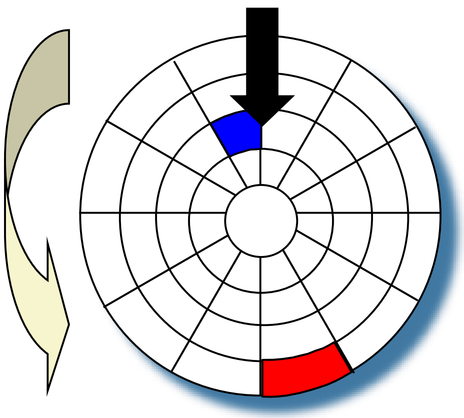
Disk Access – Read



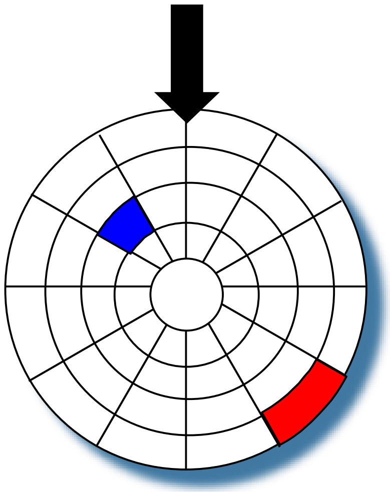
After BLUE read

Red request scheduled next

Disk Access – Seek



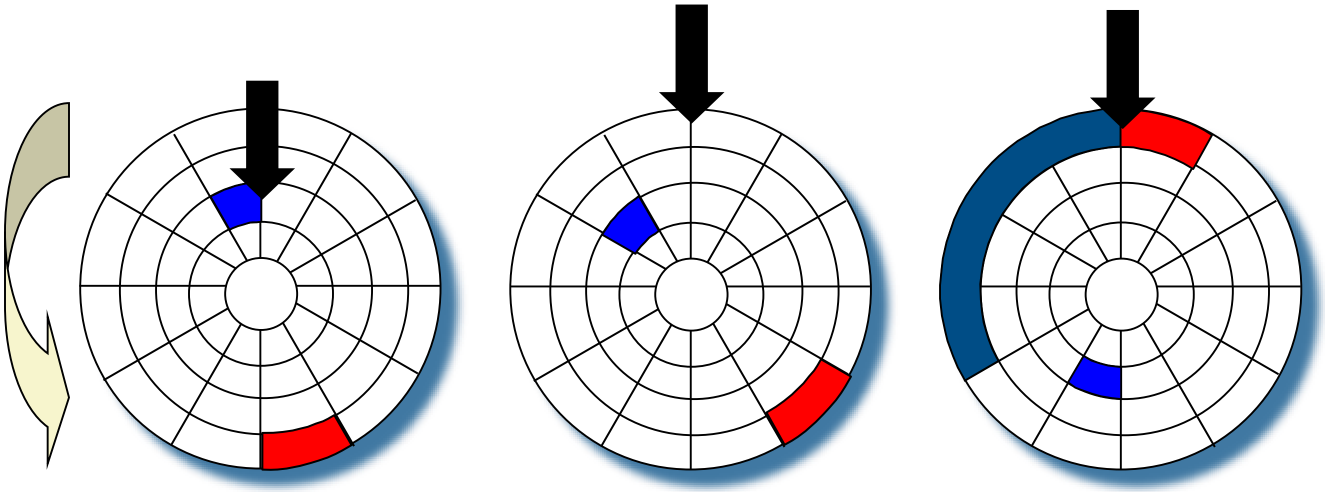
After BLUE read



Seek for RED

Seek to red's track

Disk Access – Rotational Latency



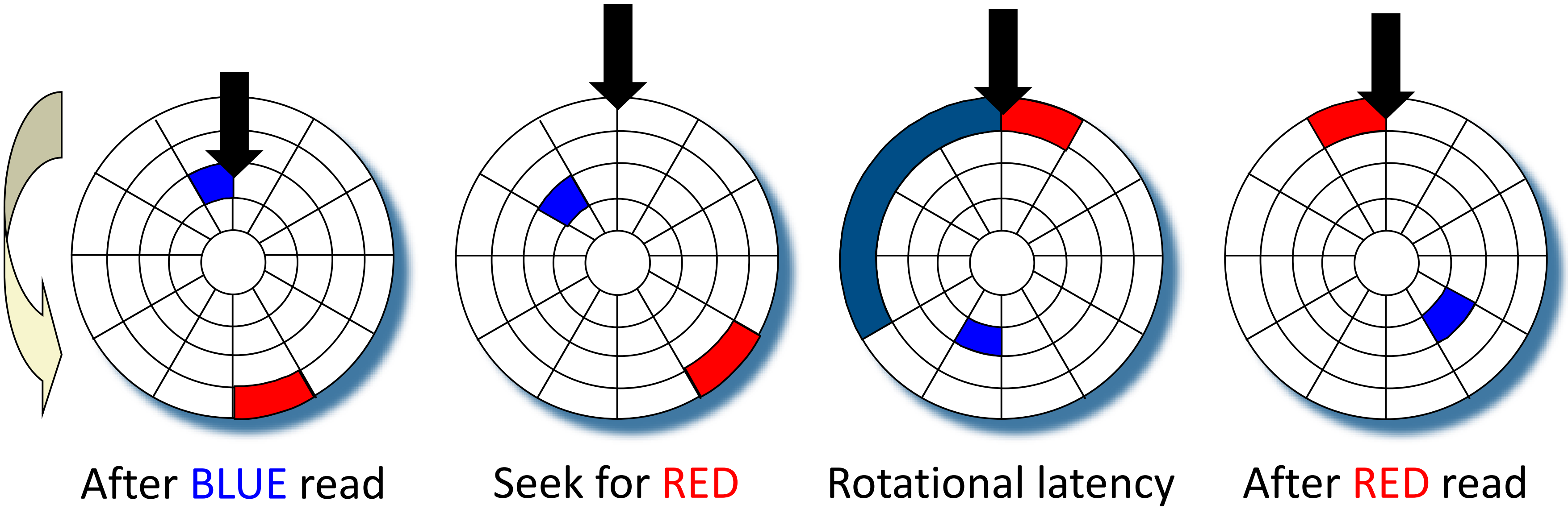
After BLUE read

Seek for RED

Rotational latency

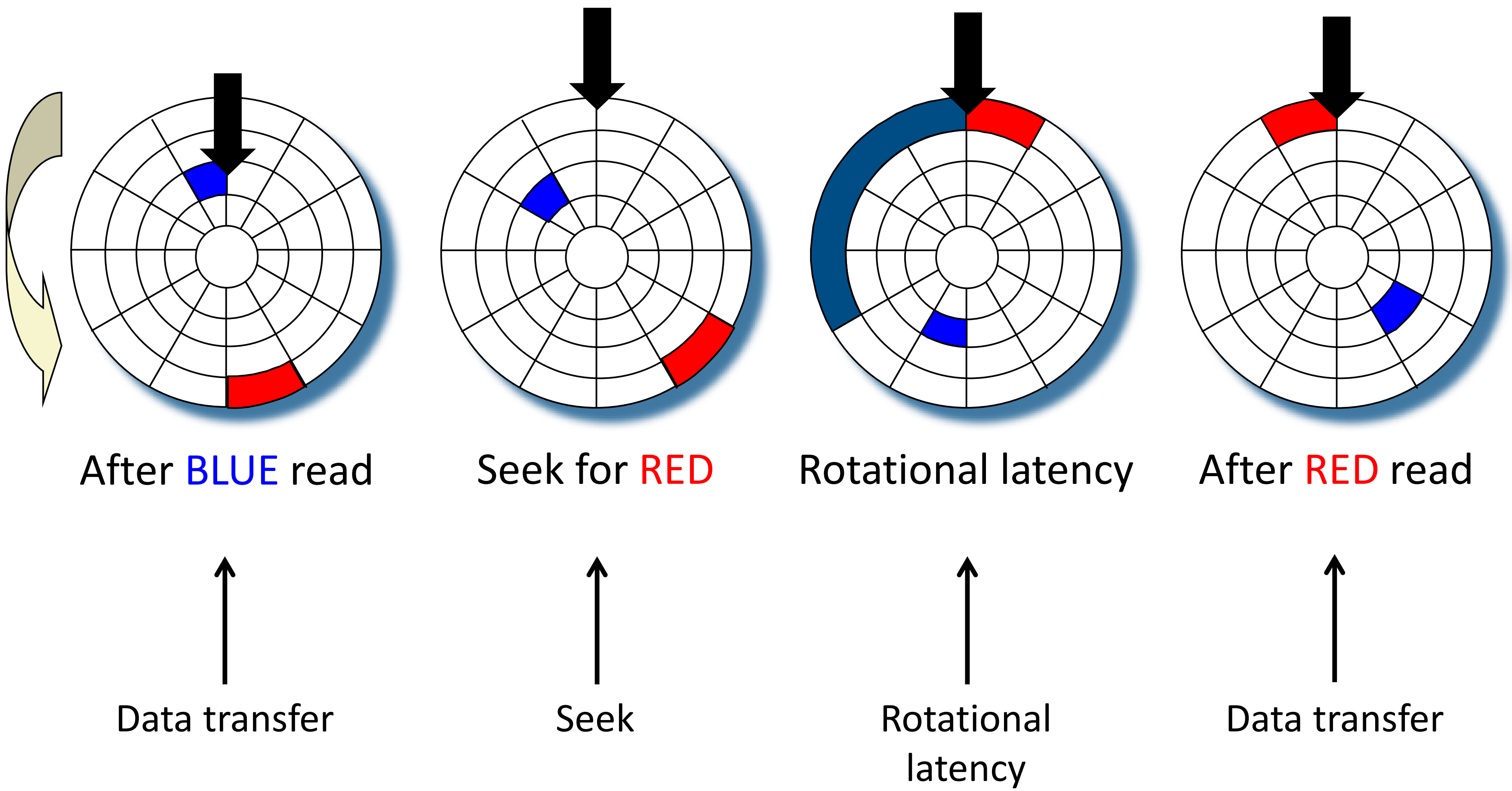
Wait for red sector to rotate around

Disk Access – Read



Complete **read** of red

Disk Access – Service Time Components



CSE 컴퓨터공학부 Department of Computer Science & Engineering

Disk Access Time

- Average time to access some target sector approximated by :
  - $T_{access} = T_{avg\ seek} + T_{avg\ rotation} + T_{avg\ transfer}$
- Seek time ($T_{avg\ seek}$)
  - Time to position heads over cylinder containing target sector.
  - Typical $T_{avg\ seek}$ is 3—9 ms
- Rotational latency ($T_{avg\ rotation}$)
  - Time waiting for first bit of target sector to pass under r/w head.
  - $T_{avg\ rotation}$ = 1/2 x 1/RPMs x 60 sec/1 min
  - Typical $T_{avg\ rotation}$ = 7200 RPMs
- Transfer time ($T_{avg\ transfer}$)
  - Time to read the bits in the target sector.
  - $T_{avg\ transfer}$ = 1/RPM x 1/(avg # sectors/track) x 60 secs/1 min.

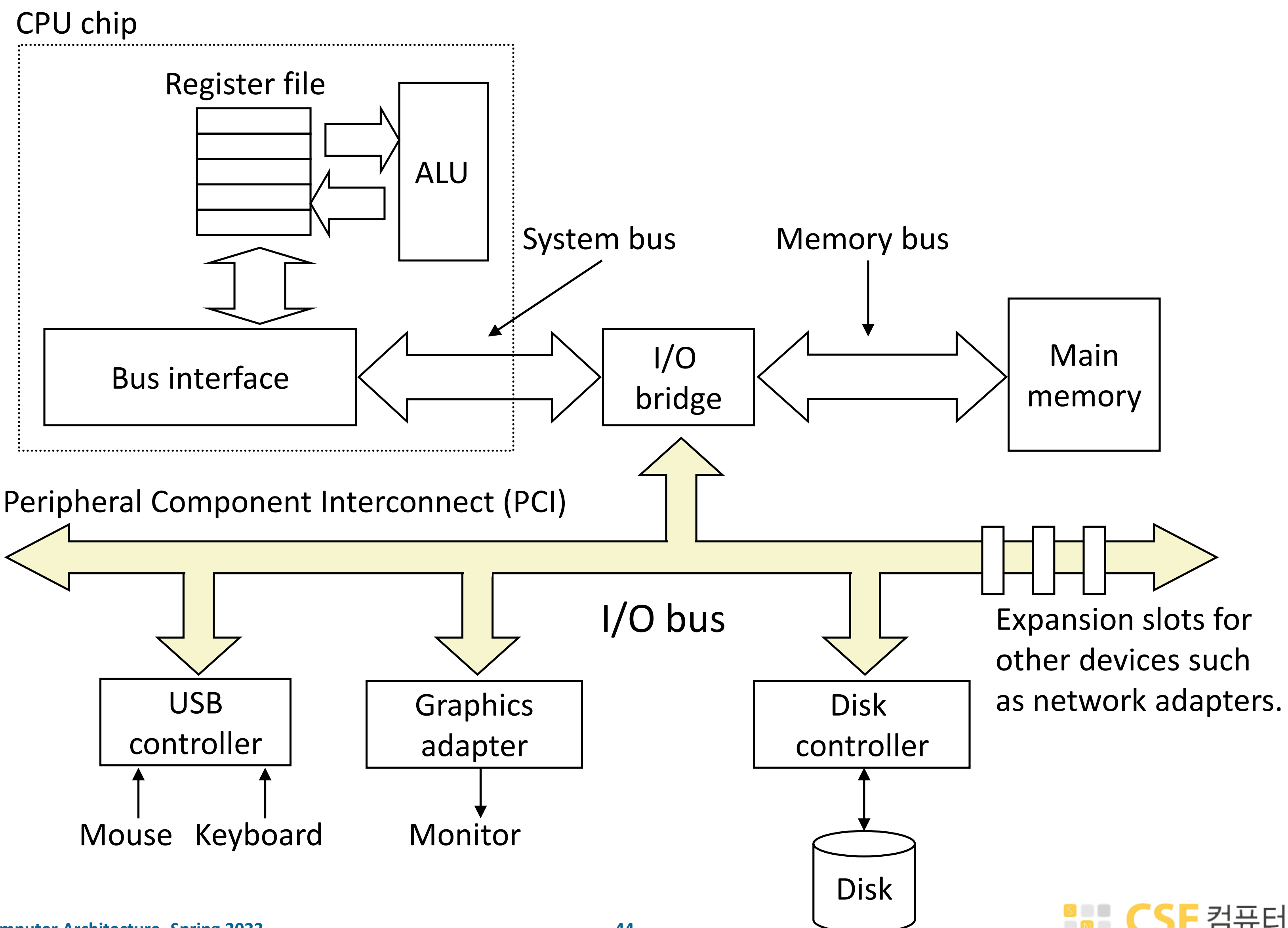
Disk Access Time Example

- Given:
  - Rotational rate = 7,200 RPM
  - Average seek time = 9 ms
  - Avg # sectors/track = 400
- Derived:
  - $T_{avg\ rotation}$ = 1/2 x (60 secs/7200 RPM) x 1000 ms/sec = 4 ms.
  - $T_{avg\ transfer}$ = 60/7200 RPM x 1/400 secs/track x 1000 ms/sec = 0.02 ms
  - T_{access} = 9 ms + 4 ms + 0.02 ms
- Important points:
  - Access time dominated by seek time and rotational latency.
  - First bit in a sector is the most expensive, the rest are free.
  - SRAM access time is about 4 ns/doubleword, DRAM about 60 ns
    - Disk is about 40,000 times slower than SRAM,
    - 2,500 times slower than DRAM.

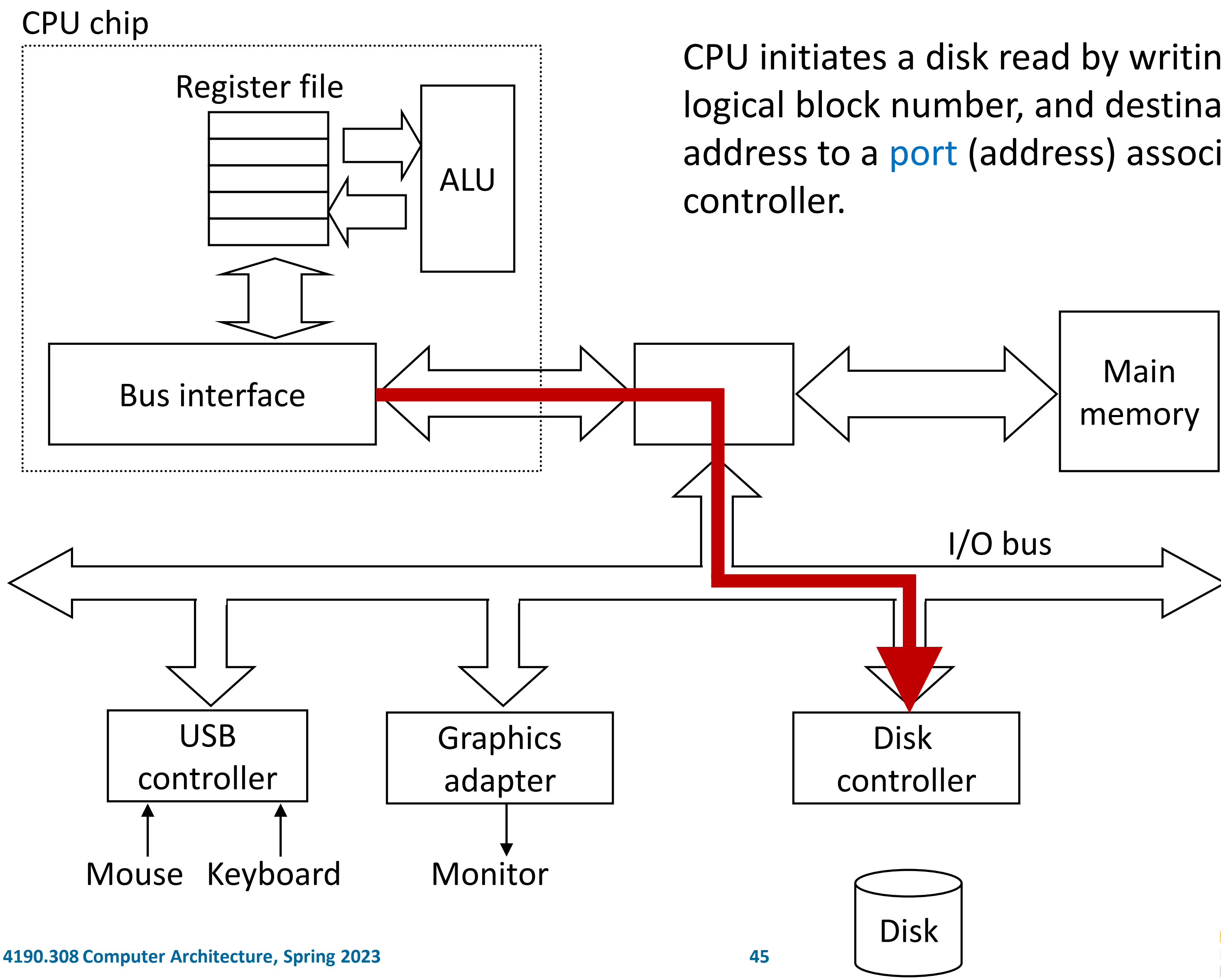
Logical Disk Blocks

- Modern disks present a simpler abstract view of the complex sector geometry:
  - The set of available sectors is modeled as a sequence of b-sized logical blocks (0, 1, 2, ...)
- Mapping between logical blocks and actual (physical) sectors
  - Maintained by hardware/firmware device called disk controller.
  - Converts requests for logical blocks into (surface,track,sector) triples.
- Allows controller to set aside spare cylinders for each zone.
  - Accounts for the difference in “formatted capacity” and “maximum capacity”.

I/O Bus


CPU chip
Register file
ALU
System bus
Memory bus
Bus interface
I/O
bridge
Main
memory
Peripheral Component Interconnect (PCI)
I/O bus
Expansion slots for
other devices such
as network adapters.
USB
controller
Graphics
adapter
Disk
controller
Mouse
Keyboard
Monitor
Disk

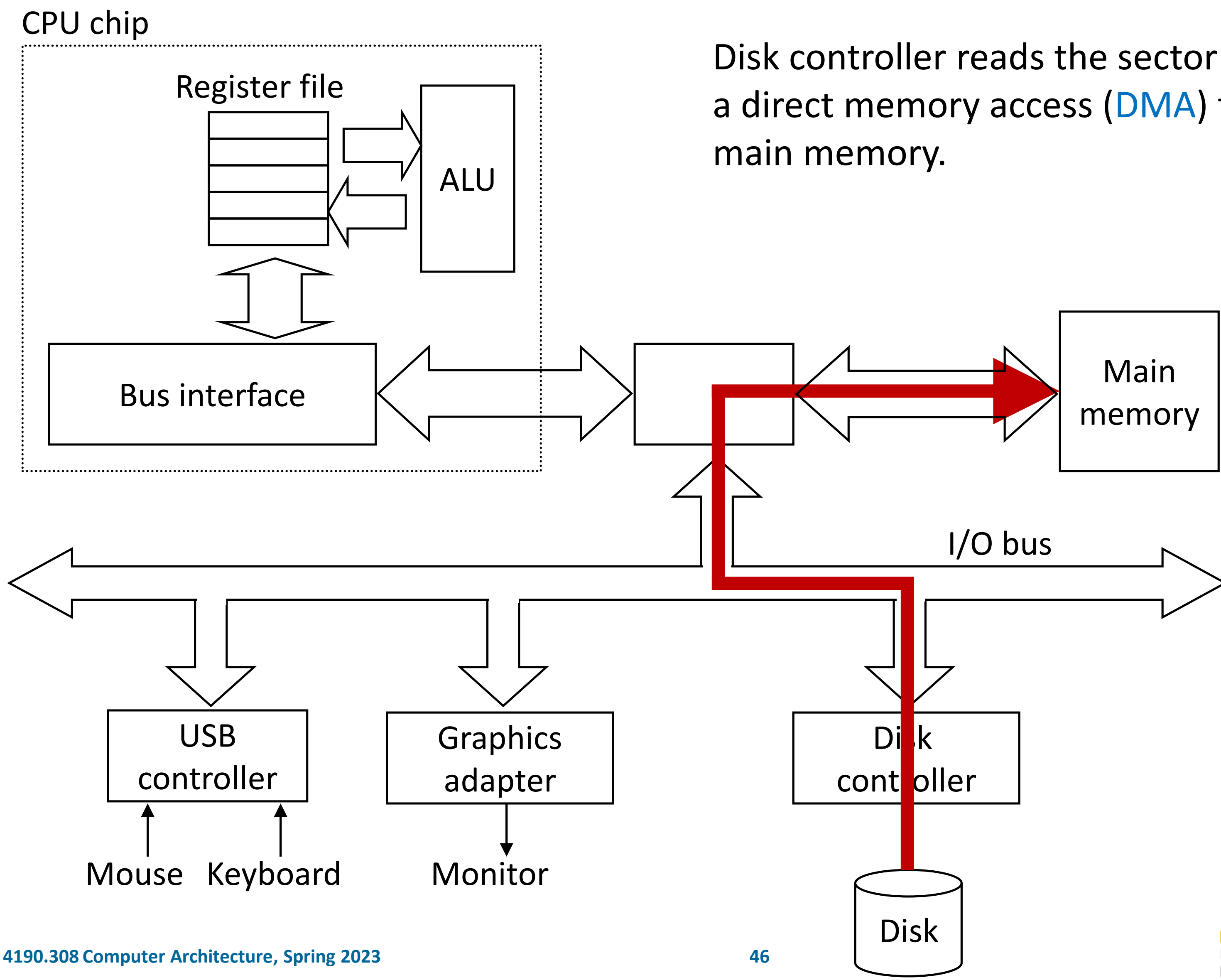
Reading a Disk Sector (1)



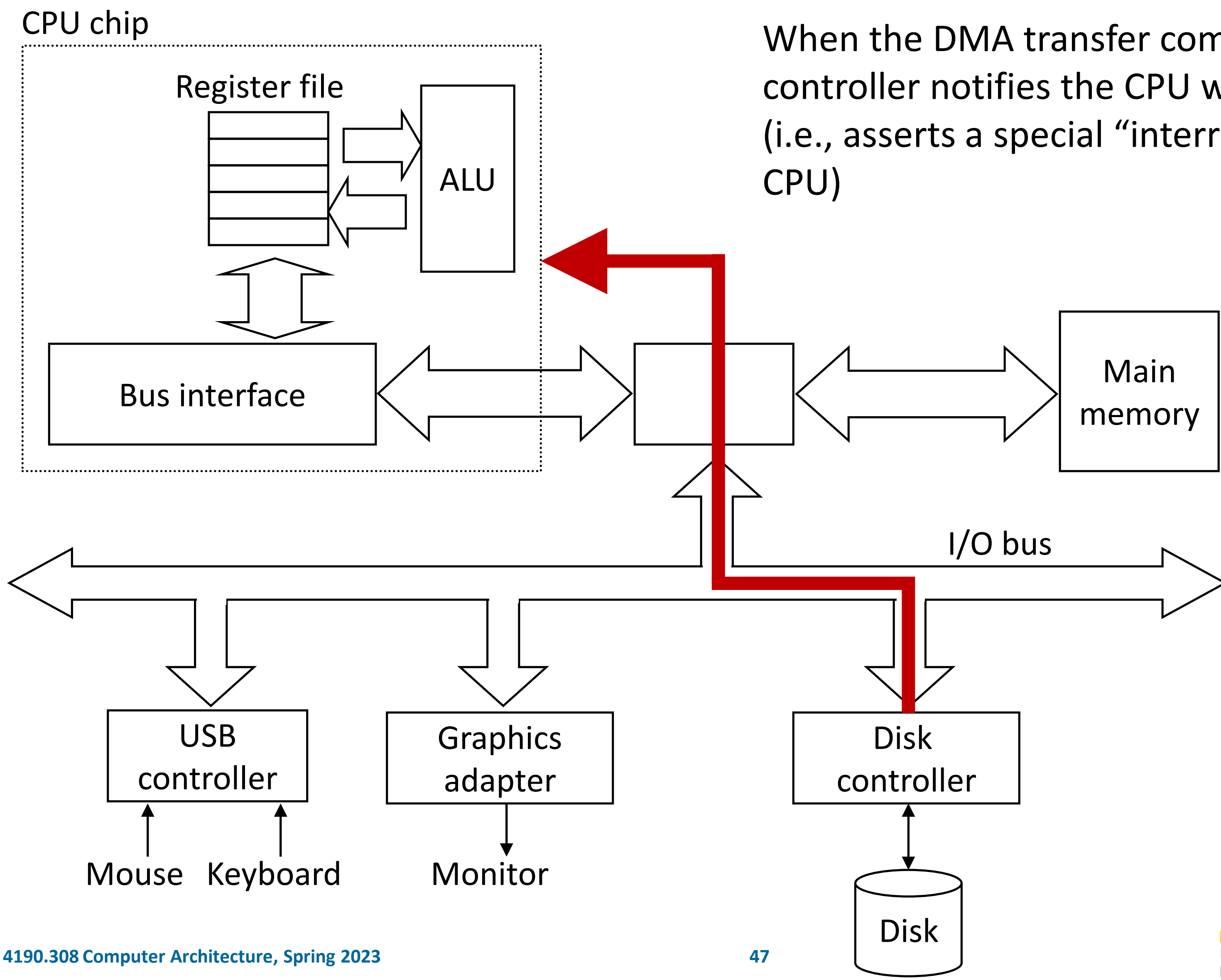
CPU initiates a disk read by writing a command, logical block number, and destination memory address to a port (address) associated with disk controller.

Reading a Disk Sector (2)

Disk controller reads the sector and performs a direct memory access (DMA) transfer into main memory.

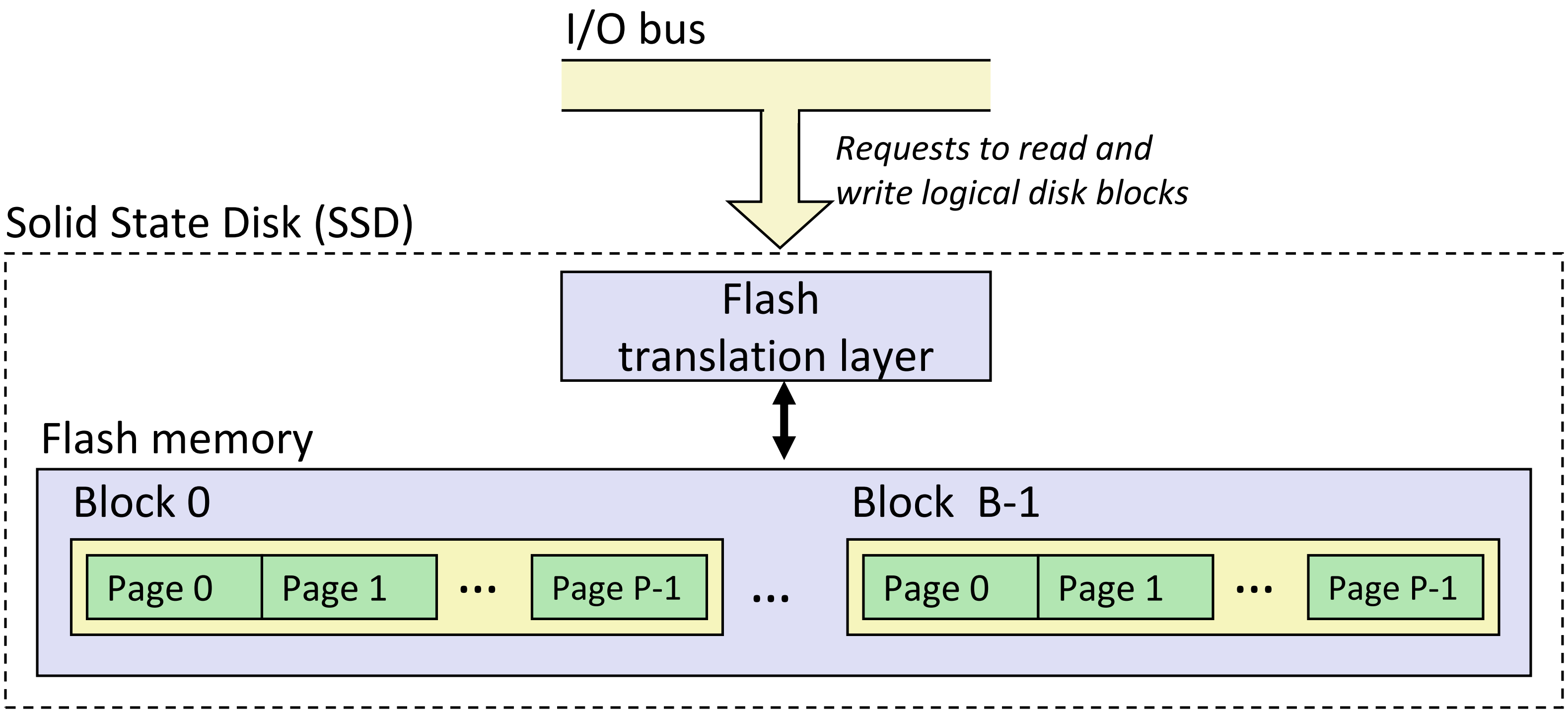


Reading a Disk Sector (3)



When the DMA transfer completes, the disk controller notifies the CPU with an interrupt (i.e., asserts a special "interrupt" pin on the CPU)

Solid State Disks (SSDs)



- Pages: 512KB to 4KB, Blocks: 32 to 128 pages
- Data read/written in units of pages.
- Page can be written only after its block has been erased
- A block wears out after ~100,000 writes.

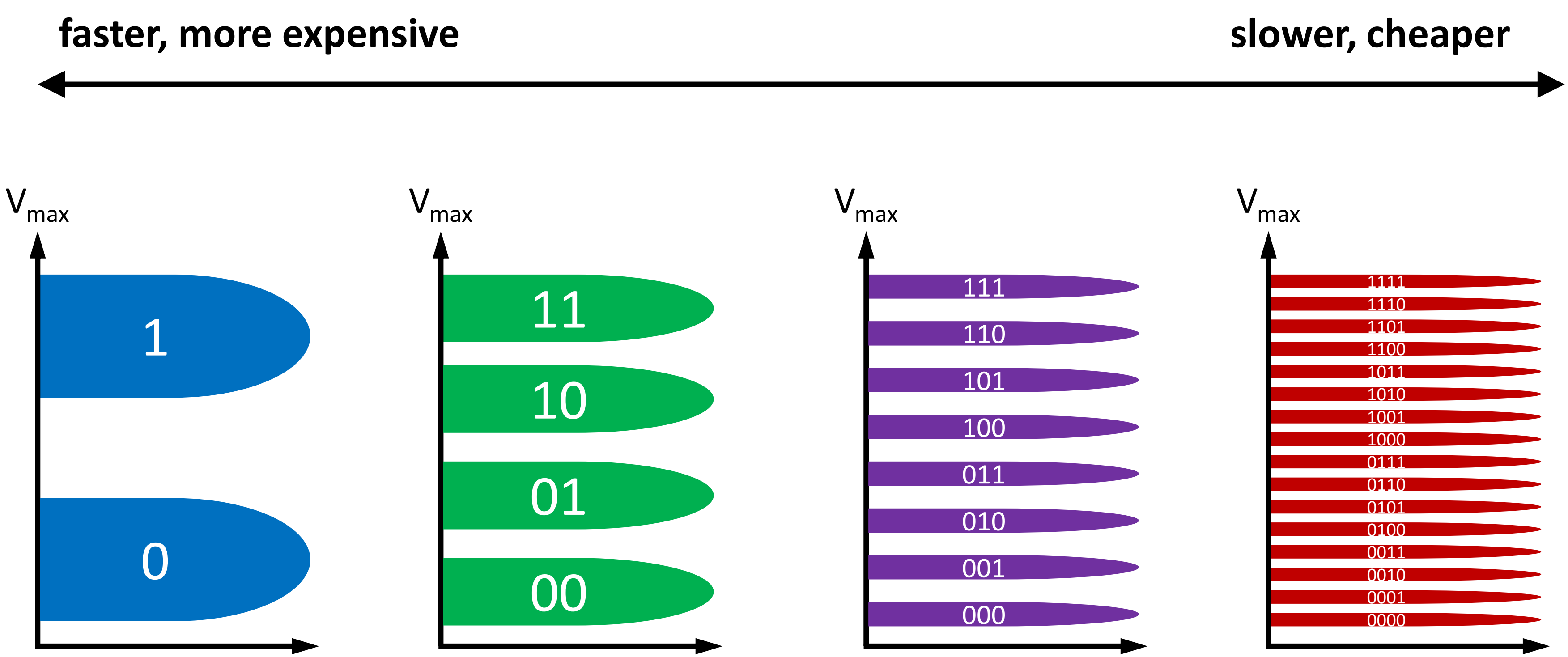
SSD Performance Characteristics

- Why are random writes so slow?

Sequential read throughput	250 MB/s	Sequential write throughput	170 MB/s
Random read throughput	140 MB/s	Random write throughput	14 MB/s
Rand read access	30 us	Random write access	300 us

- Erasing a block is slow (around 1 ms)
- Write to a page triggers a copy of all useful pages in the block
  - Find an used block (new block) and erase it
  - Write the page into the new block
  - Copy other pages from old block to the new block

SSD Storage Technology: Multi-Level Cells



SLC
one bit per cell

MLC
two bits per cell

TLC
three bits per cell

QLC
four bits per cell

PLC (five bits per cell) currently under development

SSD Tradeoffs vs Rotating Disks

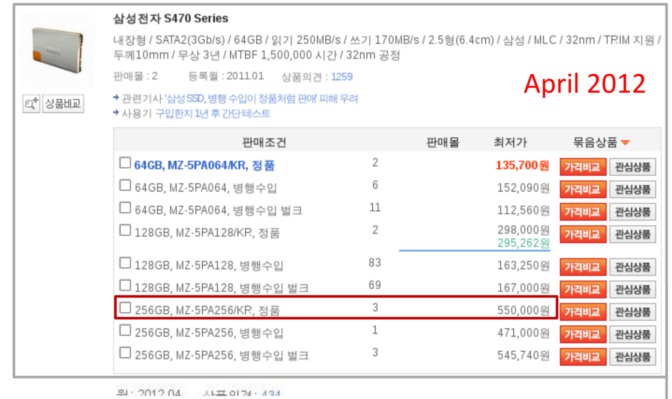
- Advantages
  - No moving parts → faster, less power, more rugged
- Disadvantages
  - Have the potential to wear out
    - Mitigated by "wear leveling logic" in flash translation layer
    - e.g. Intel X25 guarantees 1 petabyte (10^{15} bytes) of random writes before they wear out
  - More expensive
    - Used to be 100x more expensive
    - Today (2023/05/10) on danawa.co.kr:
      - Hancast 256GB SDD: 16'800won 66 KRW/GB
      - Seagate 8TB HDD: 135'980won 17 KRW/GB

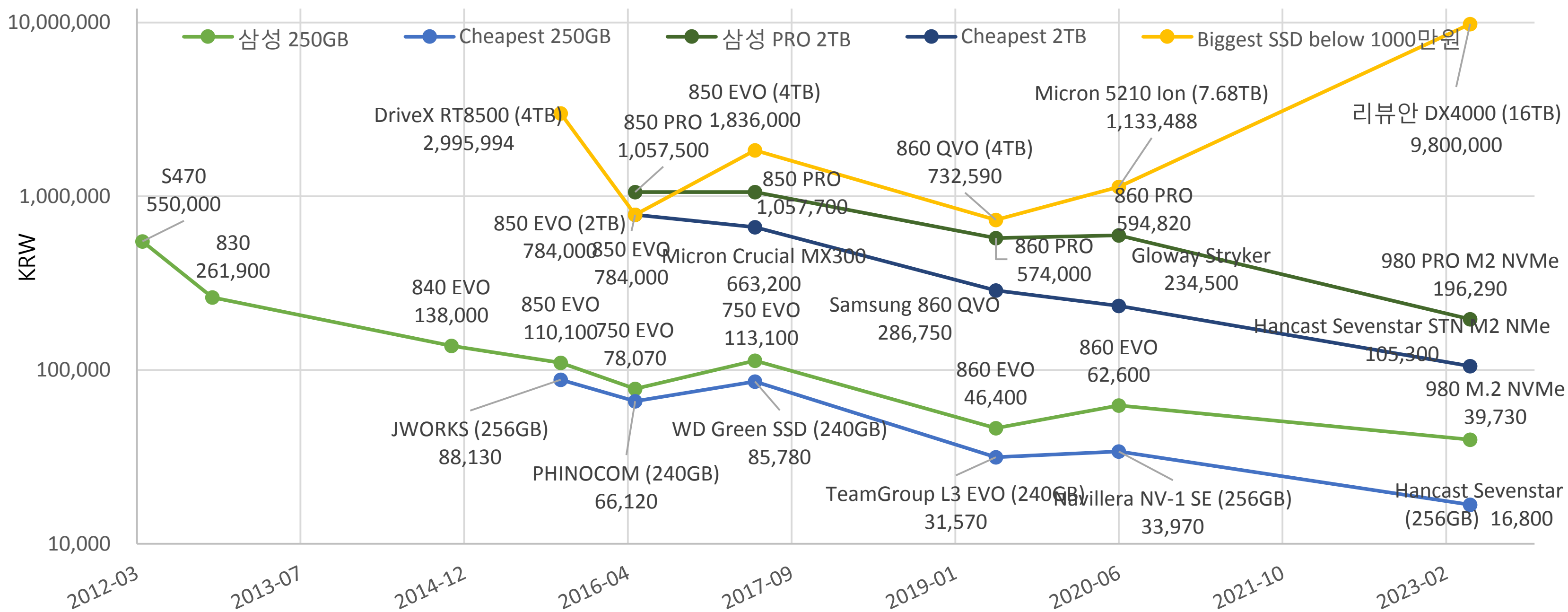
Only factor 4 now!

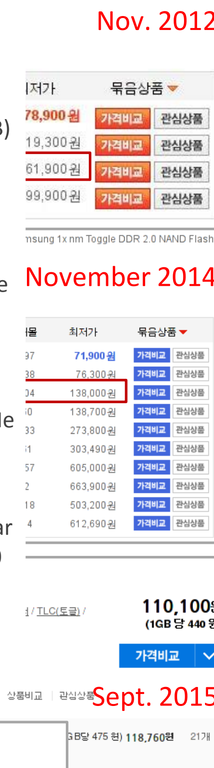
- Applications
  - MP3 players, smart phones, laptops
  - Today also the standard desktops and servers

SSD: Price Development

- Price of SSDs while teaching computer architecture (Fall 2012 ~ now)







CSE 컴퓨터공학부 Department of Computer Science & Engineering

Storage Trends

SRAM

Metric	1980	1985	1990	1995	2000	2005	2010	2020	2023	2010:1980
$/MB	19,200	2,900	320	256	100	75	60			320x
access (ns)	300	150	35	15	3	2	1.5			200x

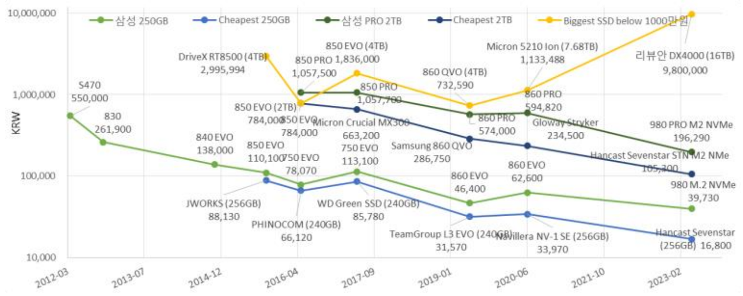
DRAM

Metric	1980	1985	1990	1995	2000	2005	2010	2020	2023	2020:1980
$/MB	8,000	880	100	30	1	0.1	0.06	0.004	0.0026	2,000,000x
access (ns)	375	200	100	70	60	50	40	20	20	18x
typical size (MB)	0.064	0.256	4	16	64	2,000	8,000	16,000	32,000	500,000x

Disk

Metric	1980	1985	1990	1995	2000	2005	2010	2020	2023	2020:1980
$/MB	500	100	8	0.3	0.01	0.005	0.0003	0.00002	0.000013	40,000,000x
access (ms)	87	75	28	10	8	4	3	3	3	29x
typical size (MB)	1	10	160	1,000	20,000	160,000	1,500,000	4,000,000	8,000,000	8,000,000x

→ it is easier to increase density than to decrease access time



Module Summary

Module Summary

- **Volatile vs nonvolatile storage**
  - SRAM and DRAM
  - magnetic disk drives, solid state drives
- **It is easier to increase capacity than to reduce latency**
- **Extreme developments over the past 40 years**