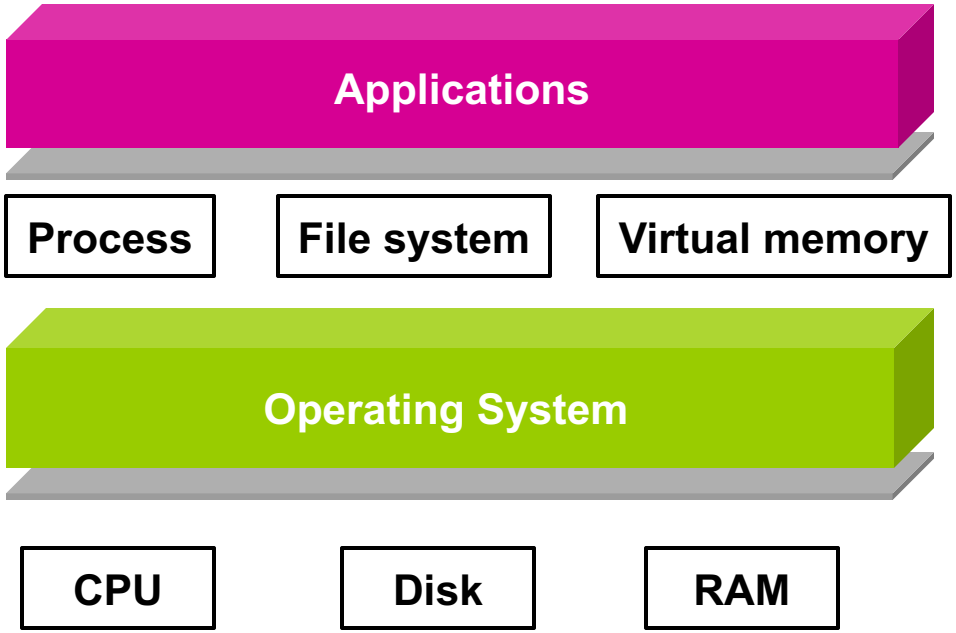
CS 153
Design of Operating Systems

Fall 19

Lecture 14: Disk Drives

Instructor: Chengyu Song

OS Abstractions



File Systems Agenda

- First, we'll discuss properties of physical disks
  - Structure, Performance
  - Scheduling
- Then we'll discuss how to build file systems (next time)
  - Abstraction:
    - » Files, Directories, Sharing, Protection
  - Implementation
    - » File System Layouts
    - » File Buffer Cache
    - » Read Ahead

Disks and the OS

- Disks are messy physical devices:
  - Errors, bad blocks, missed seeks, etc.
- OS's job is to hide this mess from higher level software
  - Low-level device control (initiate a disk read, etc.)
- OS provides higher-level abstractions (files, databases, etc.)
  - OS maps them to the device and implements policies to promote
    - » Performance, reliability, protection, …

What's Inside A Disk Drive?

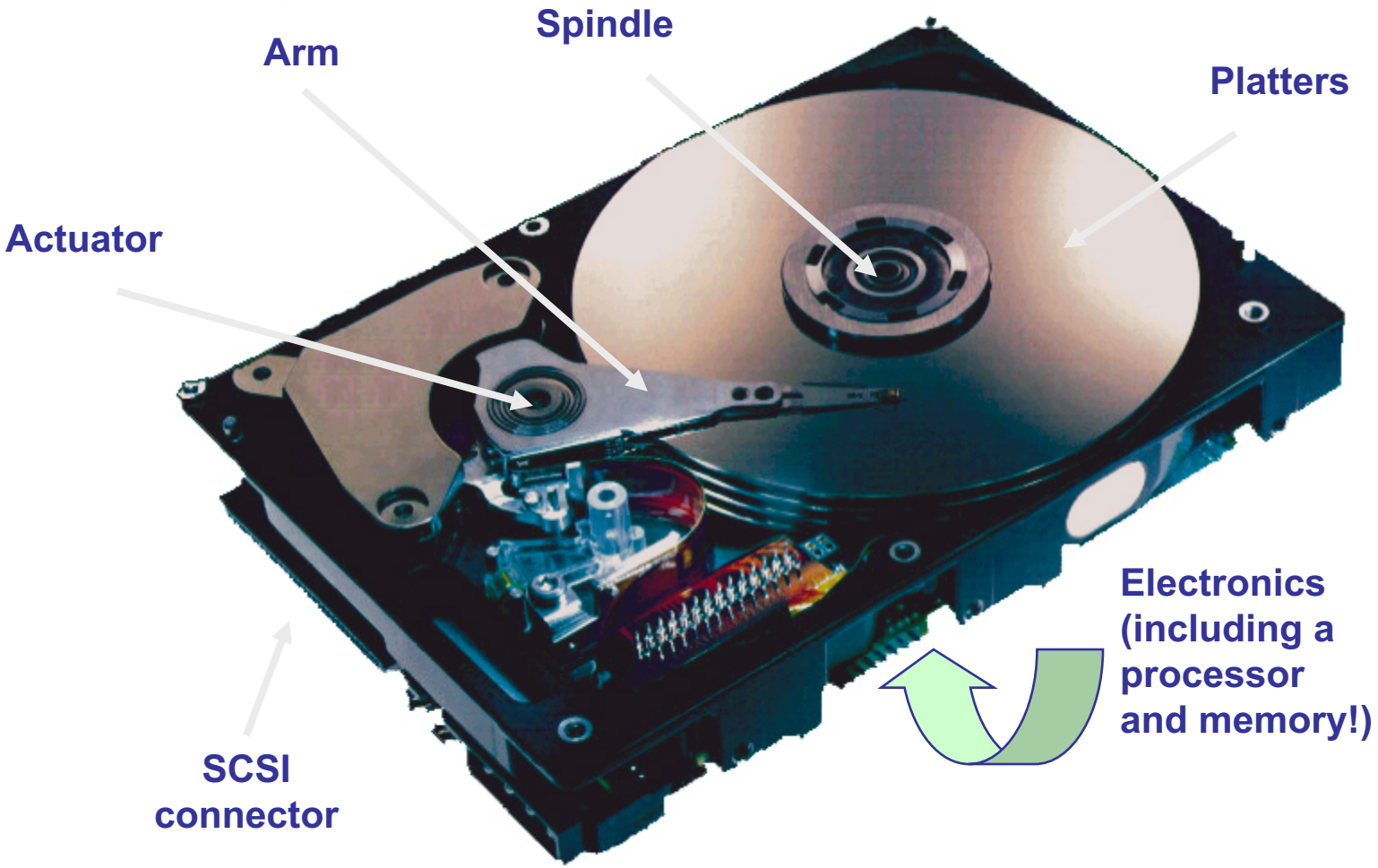
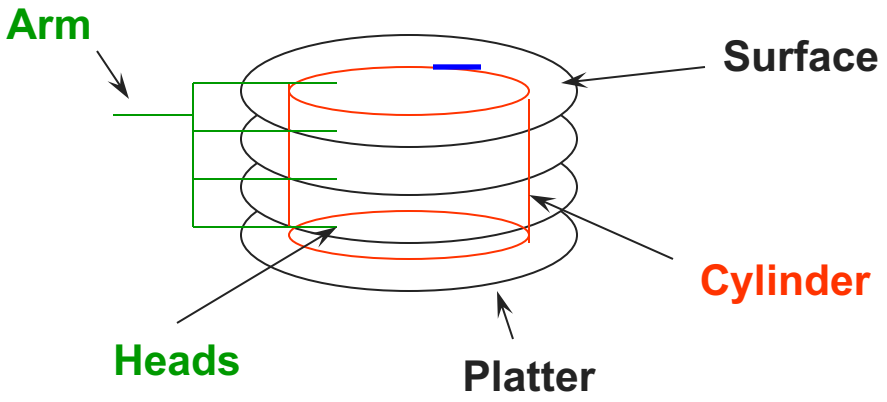
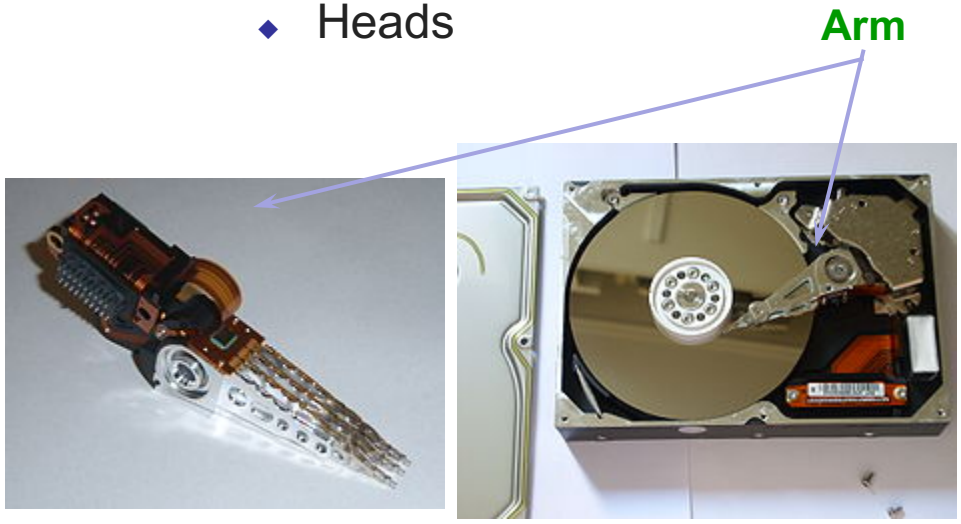


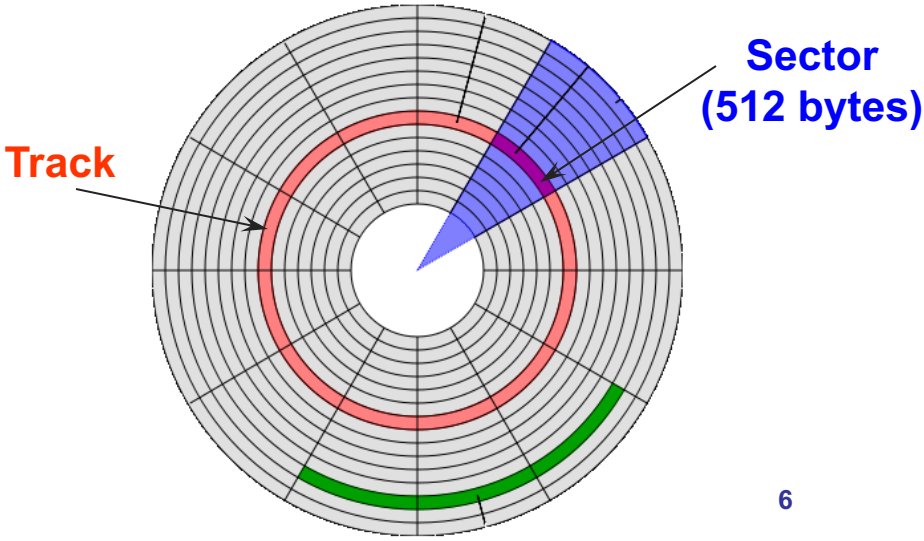
Image courtesy of Seagate Technology

Physical Disk Structure

- Disk components
  - Platters
  - Surfaces
  - Tracks
  - Sectors
  - Cylinders
  - Arm
  - Heads

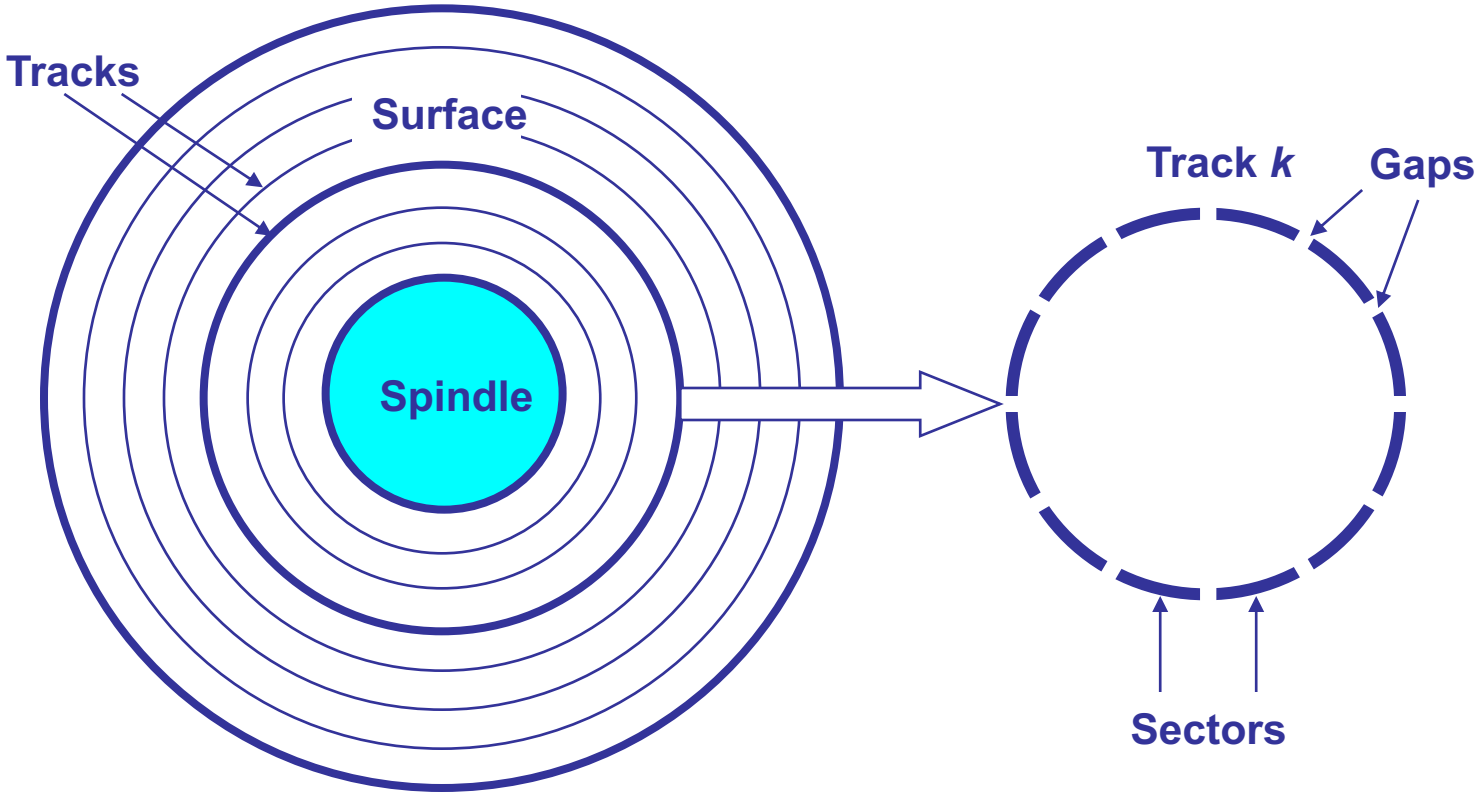






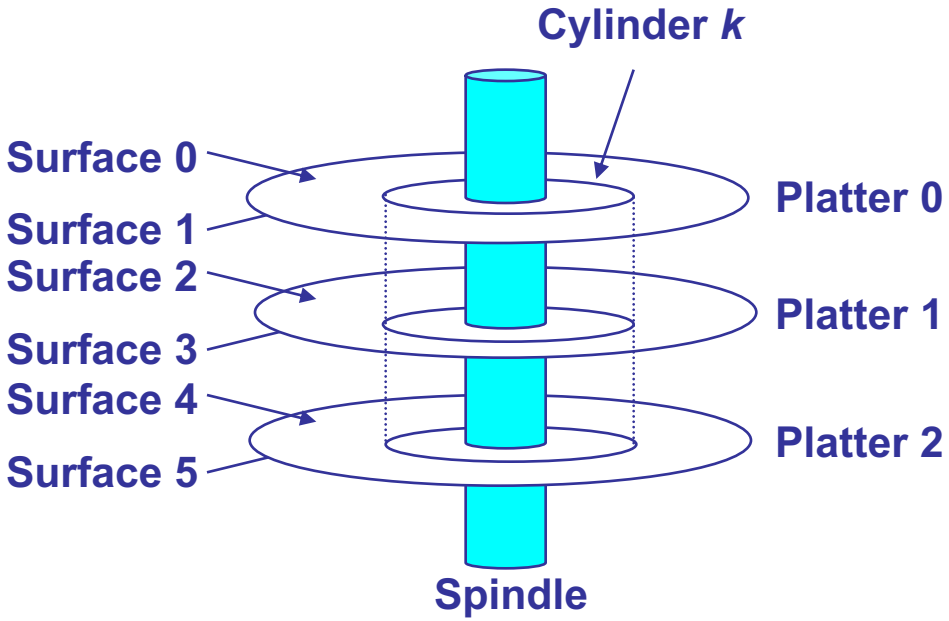
Disk Geometry

- Disks consist of platters, each with two surfaces.
- Each surface consists of concentric rings called tracks.
- Each track consists of sectors separated by gaps.



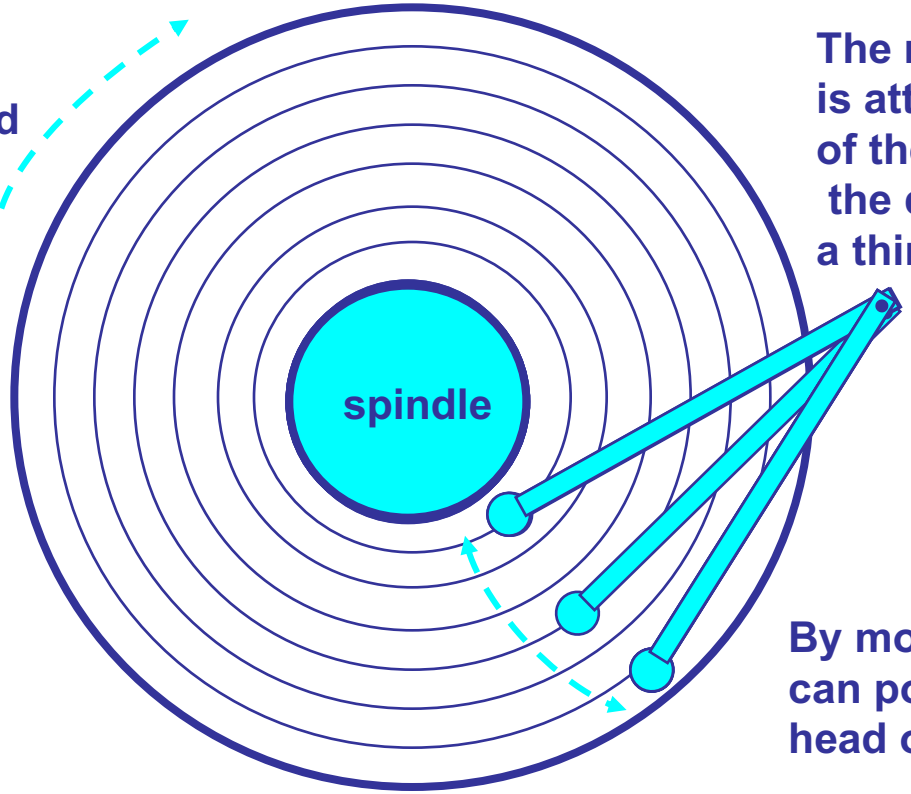
Disk Geometry (Muliple-Platter View)

- Aligned tracks form a cylinder.



Disk Operation (Single-Platter View)

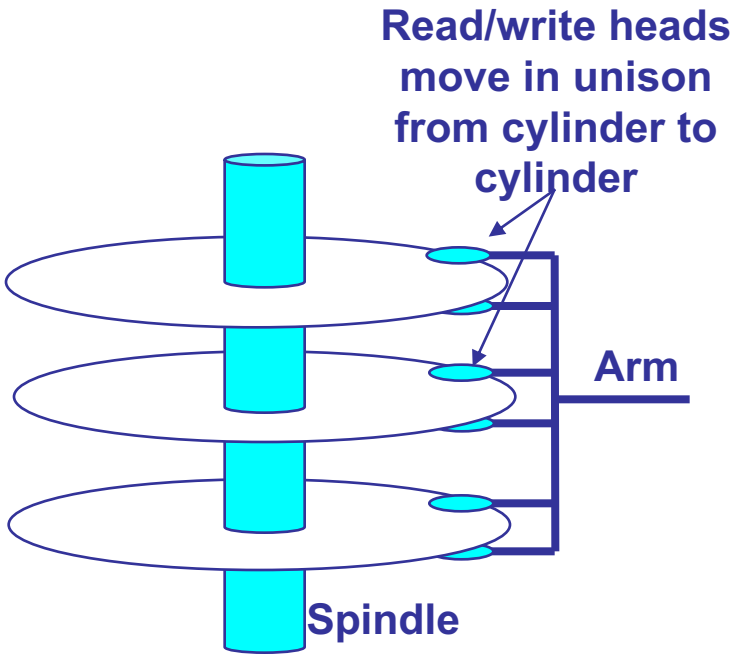
The disk surface spins at a fixed rotational rate



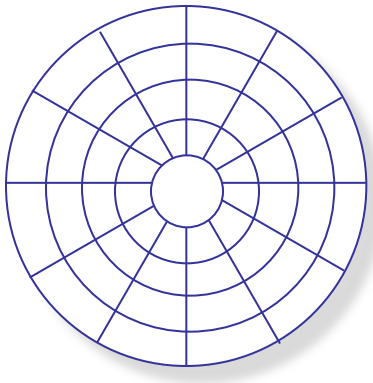
The read/write *head* is attached to the end of the *arm* and flies over the disk surface on a thin cushion of air.

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

Disk Operation (Multi-Platter View)



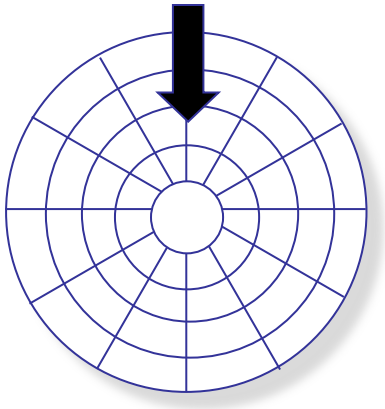
Disk Structure - top view of single platter



Surface organized into tracks

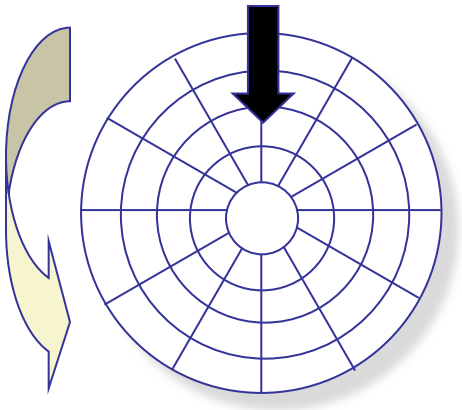
Tracks divided into sectors

Disk Access



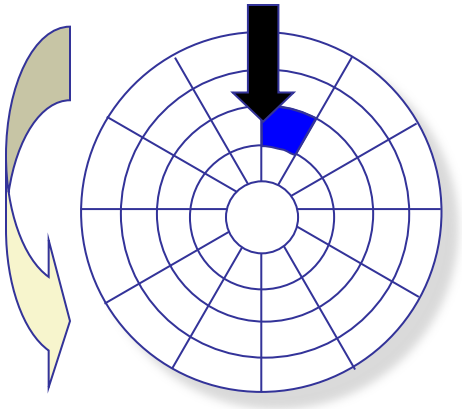
Head in position above a track

Disk Access



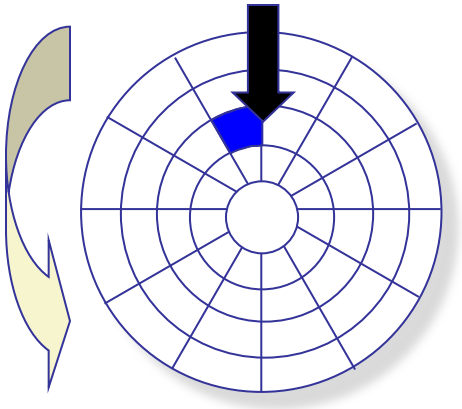
Rotation is counter-clockwise

Disk Access – Read



About to read blue sector

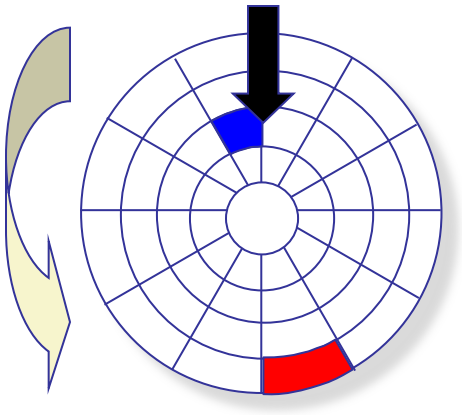
Disk Access – Read



After **BLUE** read

After reading blue sector

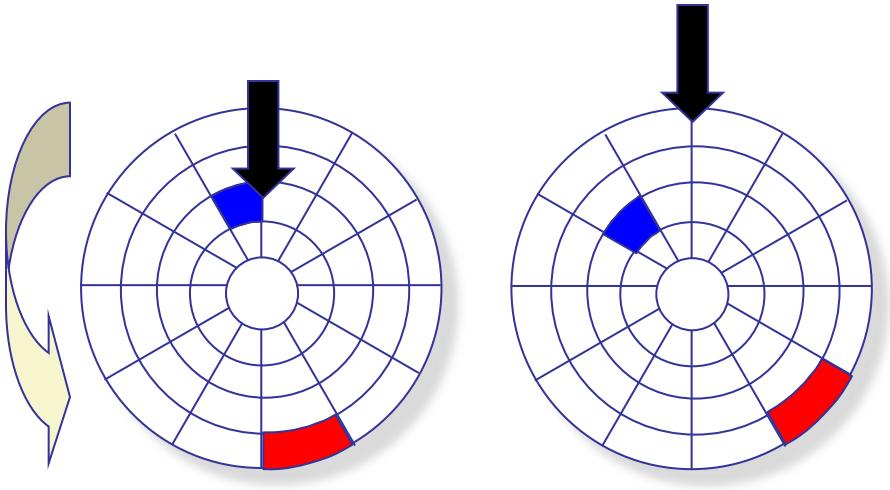
Disk Access – Read



After **BLUE** read

Red request scheduled next

Disk Access – Seek

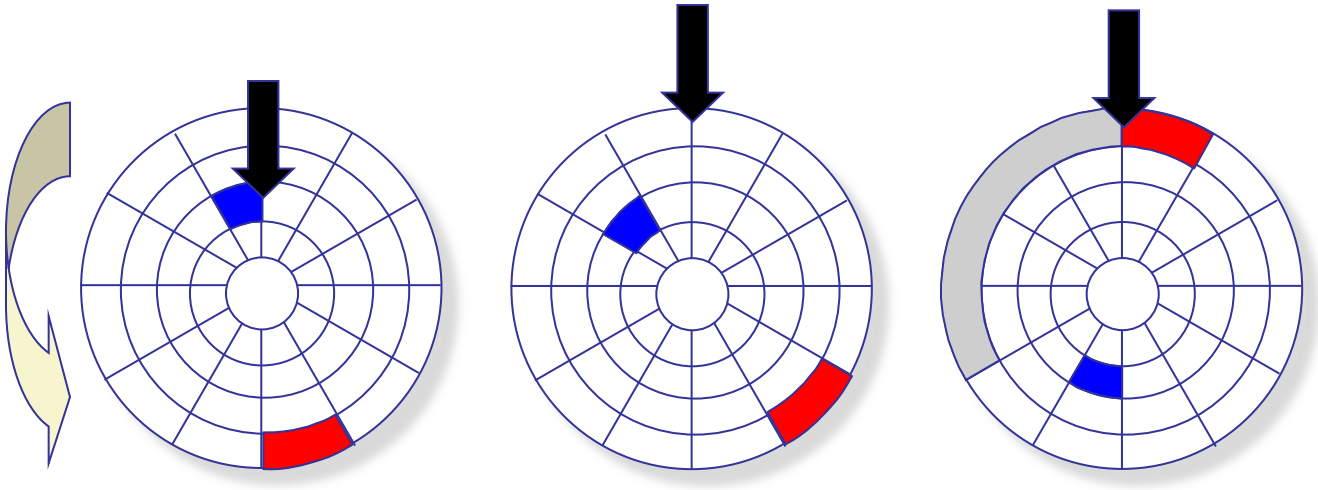


After BLUE read

Seek for RED

Seek to red's track

Disk Access – Rotational Latency



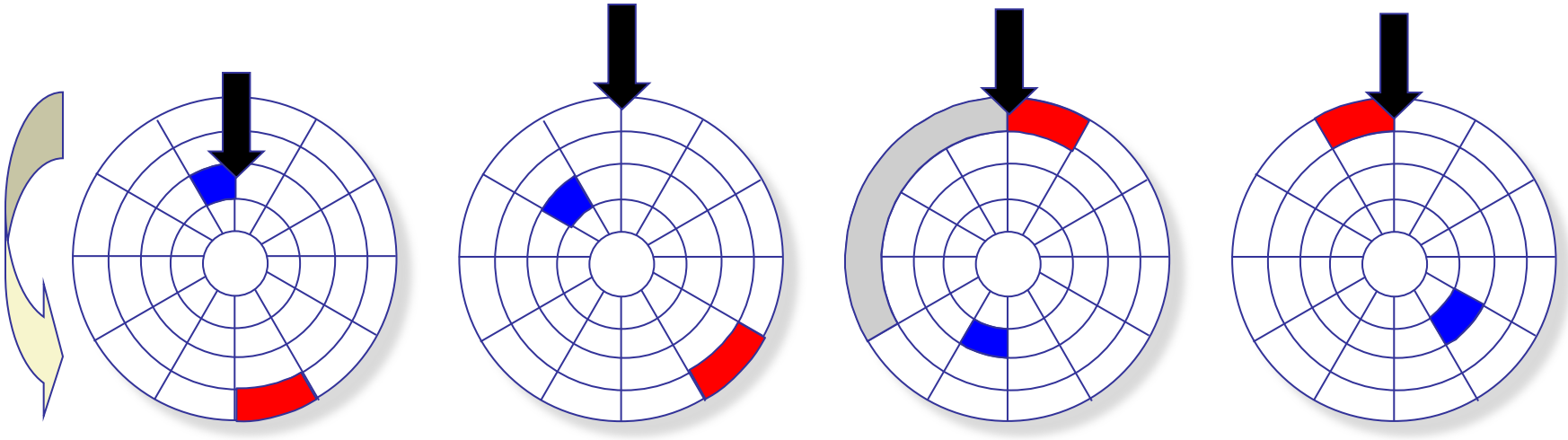
After **BLUE** read

Seek for **RED**

Rotational latency

Wait for red sector to rotate around

Disk Access – Read



After BLUE read

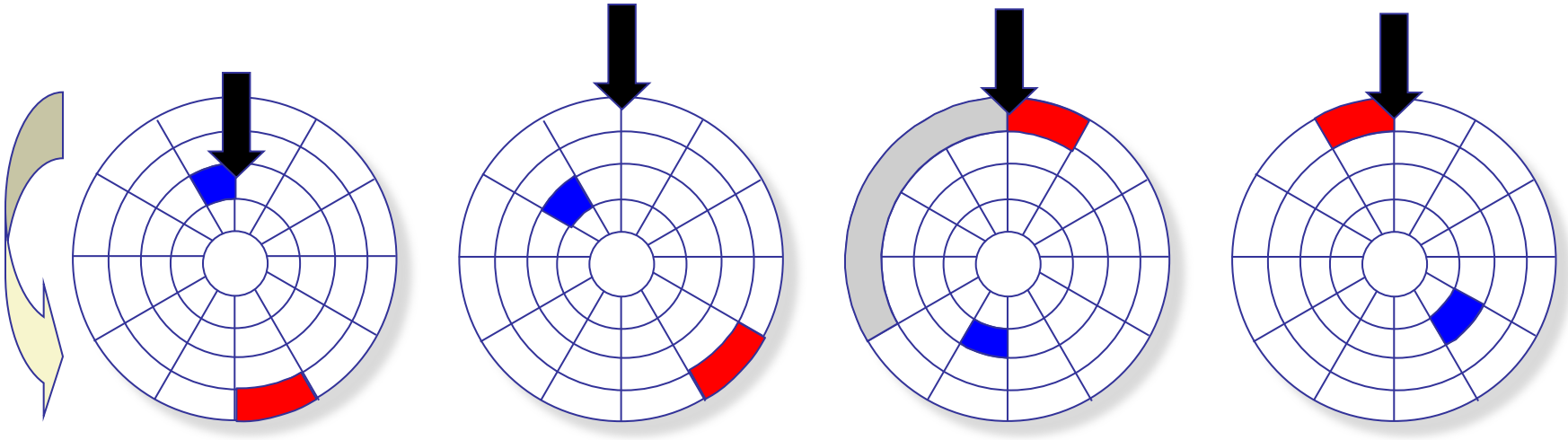
Seek for RED

Rotational latency

After RED read

Complete read of red

Disk Access – Service Time Components



Disk Access Time

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

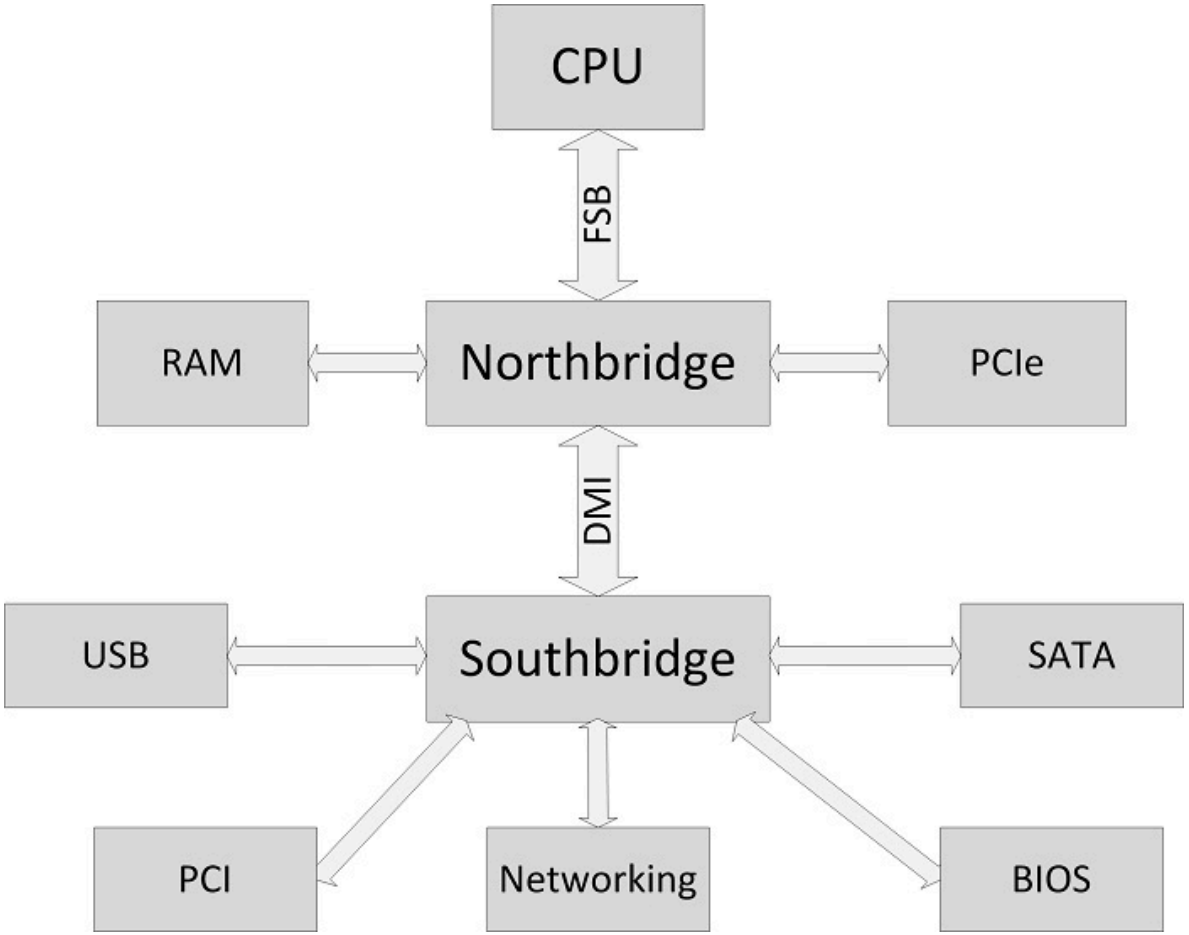
Disk Access Time Example

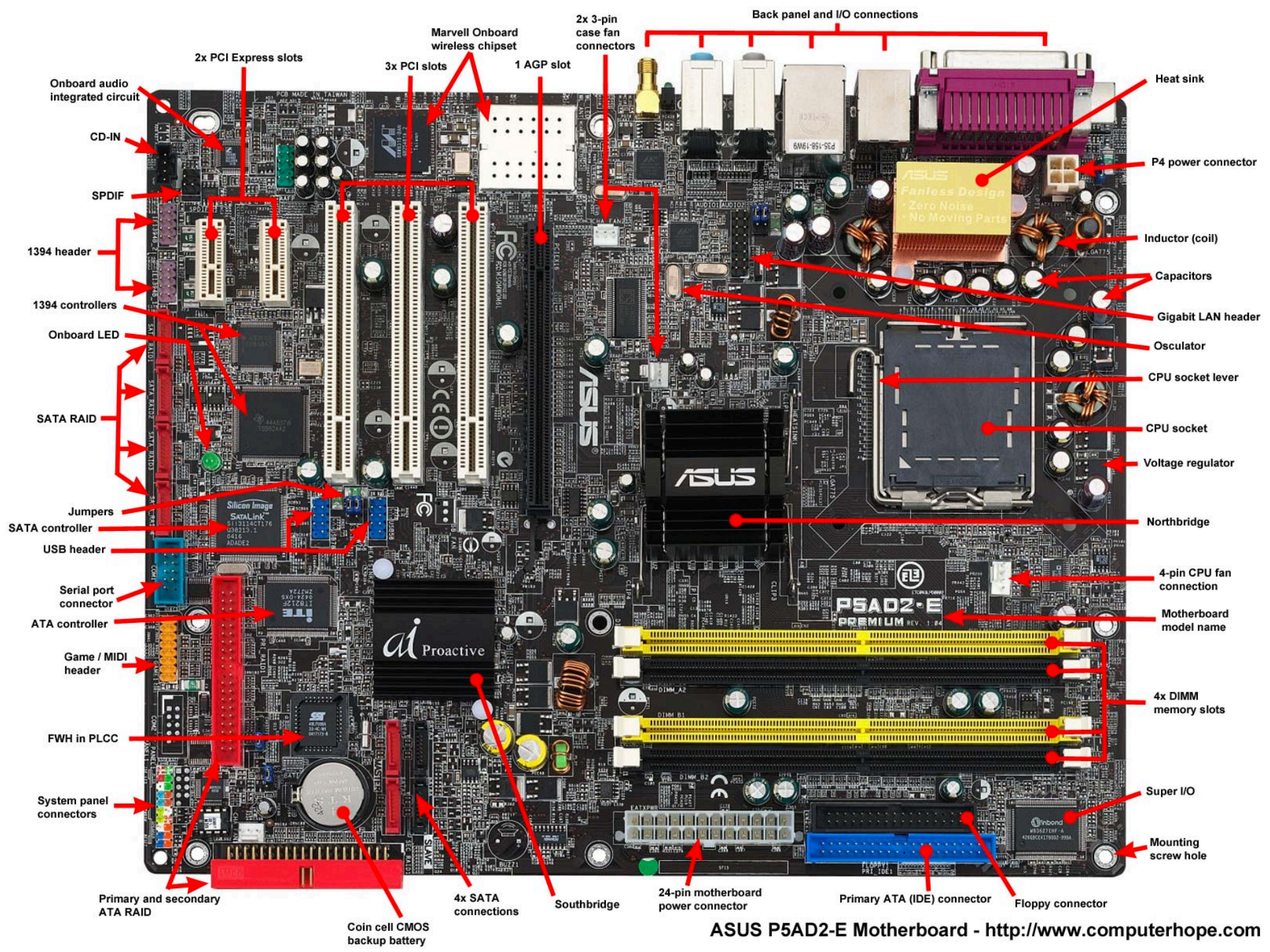
- Given:
  - Rotational rate = 7,200 RPM
  - Average seek time = 9 ms.
  - Avg # sectors/track = 400.
- Derived:
  - Tavg rotation = 1/2 x (60 secs/7200 RPM) x 1000 ms/sec = 4 ms.
  - Tavg transfer = 60/7200 RPM x 1/400 secs/track x 1000 ms/sec = 0.02 ms
  - Taccess = 9 ms + 4 ms + 0.02 ms
- Important 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 then DRAM.

Logical Disk Blocks

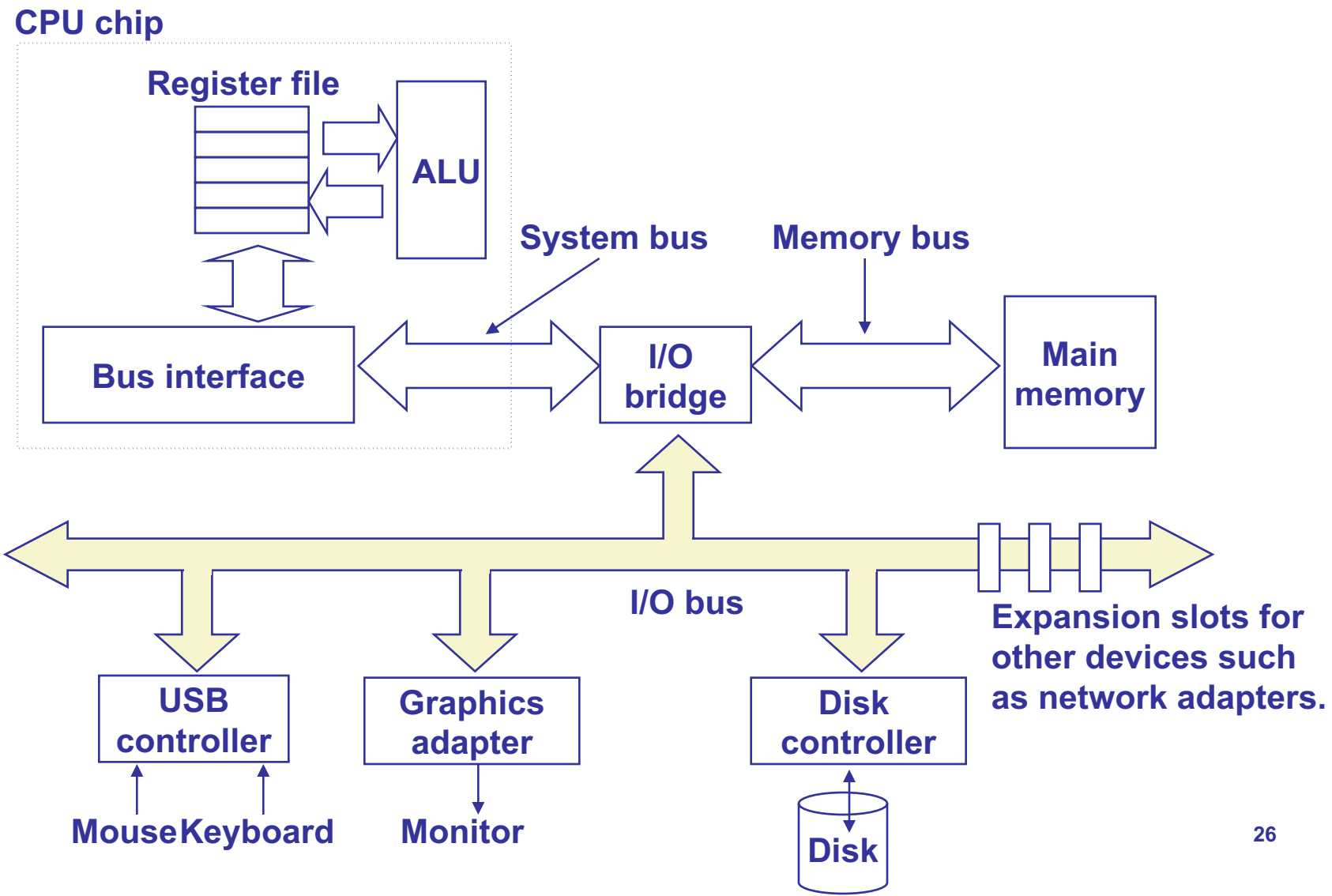
- Modern disks present a simpler abstract:
  - The set of available sectors is modeled as a sequence of b-sized logical blocks (0, 1, 2, ...)
- Mapping between logical and actual (physical) sectors
  - Maintained by a device called disk controller.
  - Converts requests for logical blocks into (surface,track,sector)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

I/O and disk in the system

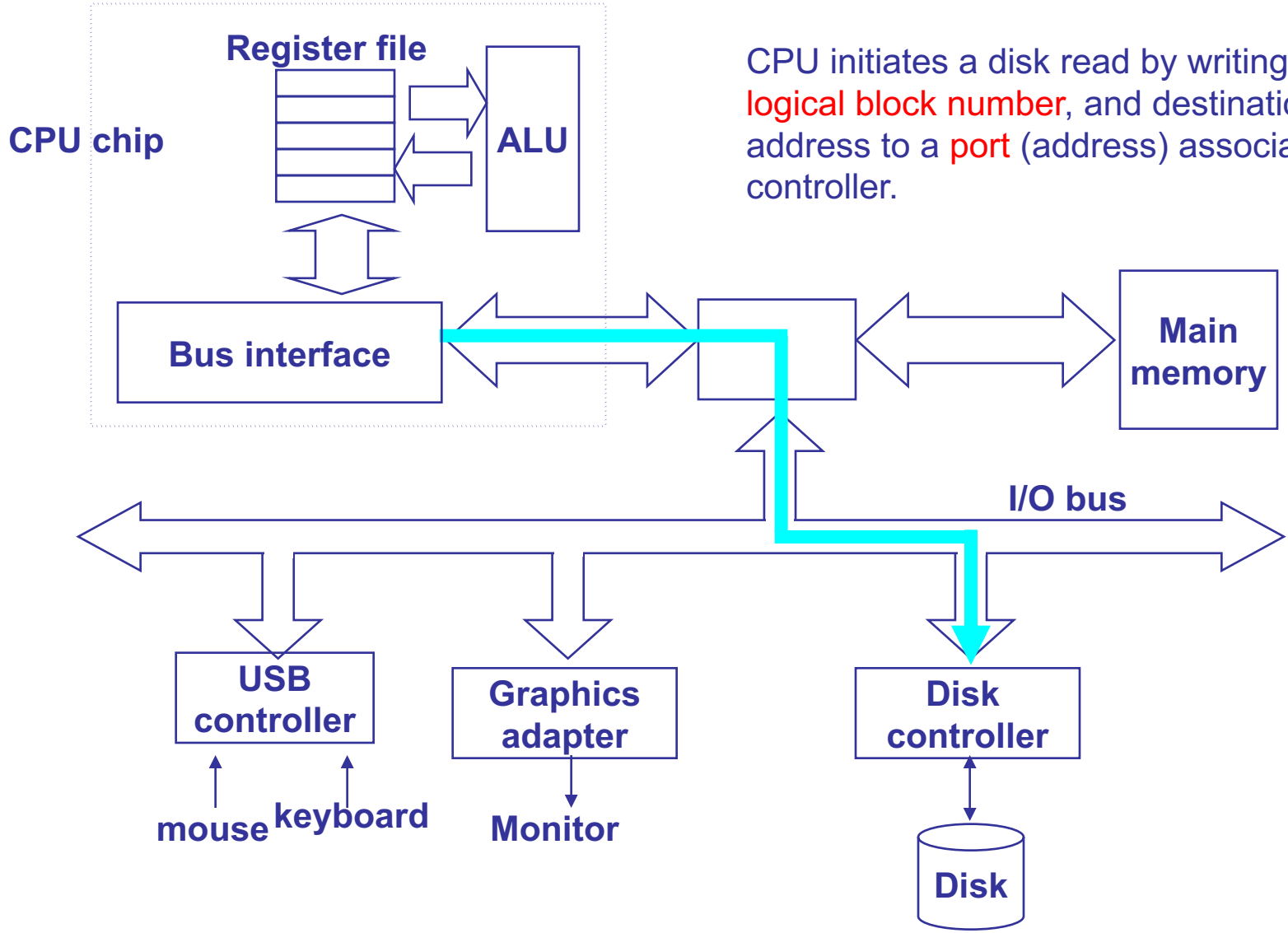



Back panel and I/O connections
2x 3-pin case fan connectors
Marvell Onboard wireless chipset
2x PCI Express slots
3x PCI slots
1 AGP slot
Onboard audio integrated circuit
CD-IN
SPDIF
1394 header
1394 controllers
Onboard LED
SATA RAID
Jumpers
SATA controller
USB header
Serial port connector
ATA controller
Game / MIDI header
FWH in PLCC
System panel connectors
Primary and secondary ATA RAID
Coin cell CMOS backup battery
4x SATA connections
Southbridge
24-pin motherboard power connector
Primary ATA (IDE) connector
Floppy connector
Heat sink
P4 power connector
Inductor (coil)
Capacitors
Gigabit LAN header
Osculator
CPU socket lever
CPU socket
Voltage regulator
Northbridge
4-pin CPU fan connection
Motherboard model name
4x DIMM memory slots
Super I/O
Mounting screw hole
ASUS P5AD2-E Motherboard - http://www.computerhope.com

I/O Bus

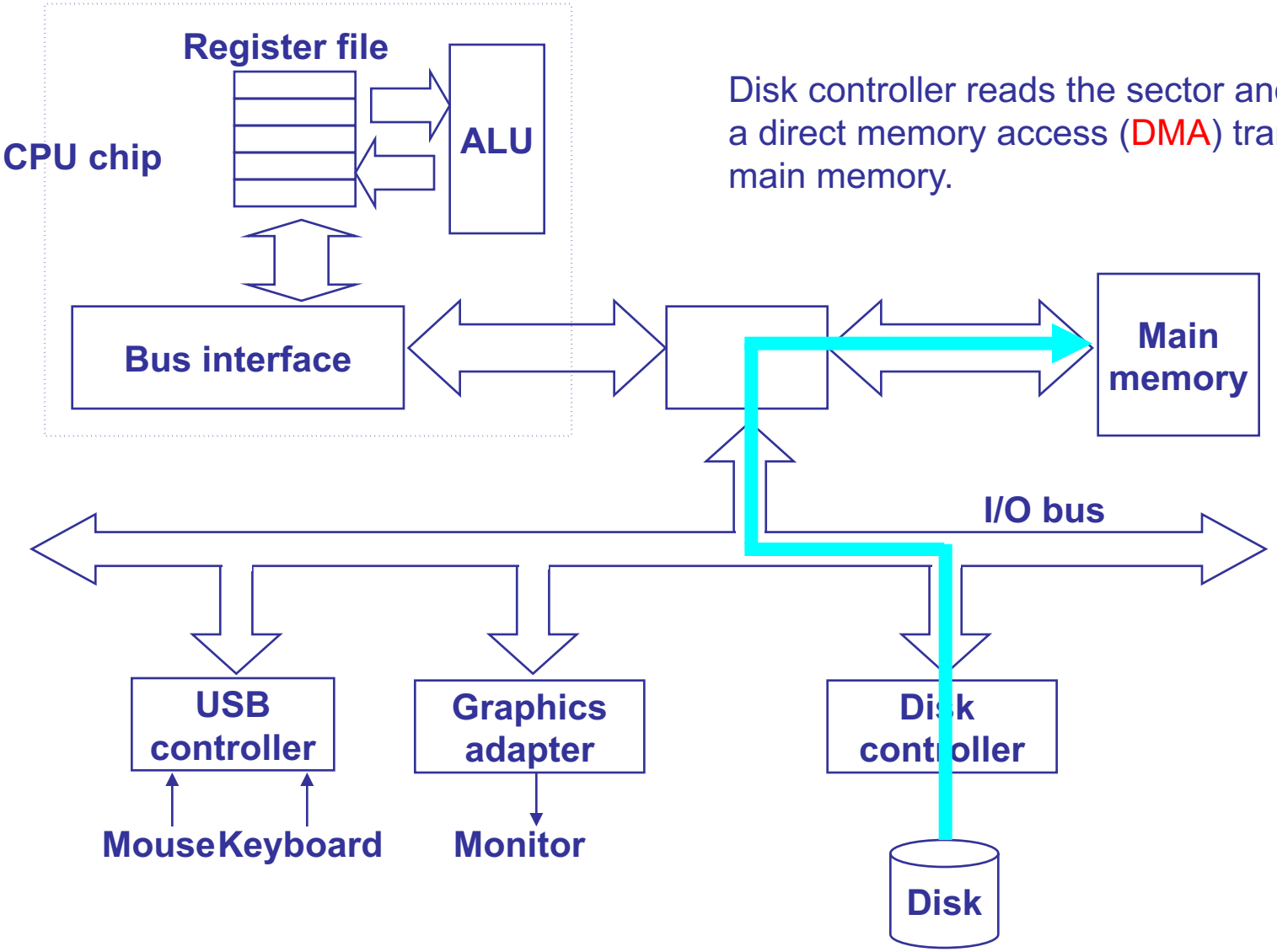

CPU chip
Register file
ALU
System bus
Memory bus
Bus interface
I/O bridge
Main memory
I/O bus
Expansion slots for other devices such as network adapters.
USB controller
Graphics adapter
Disk controller
Mouse
Keyboard
Monitor
Disk

Reading a Disk Sector (1)



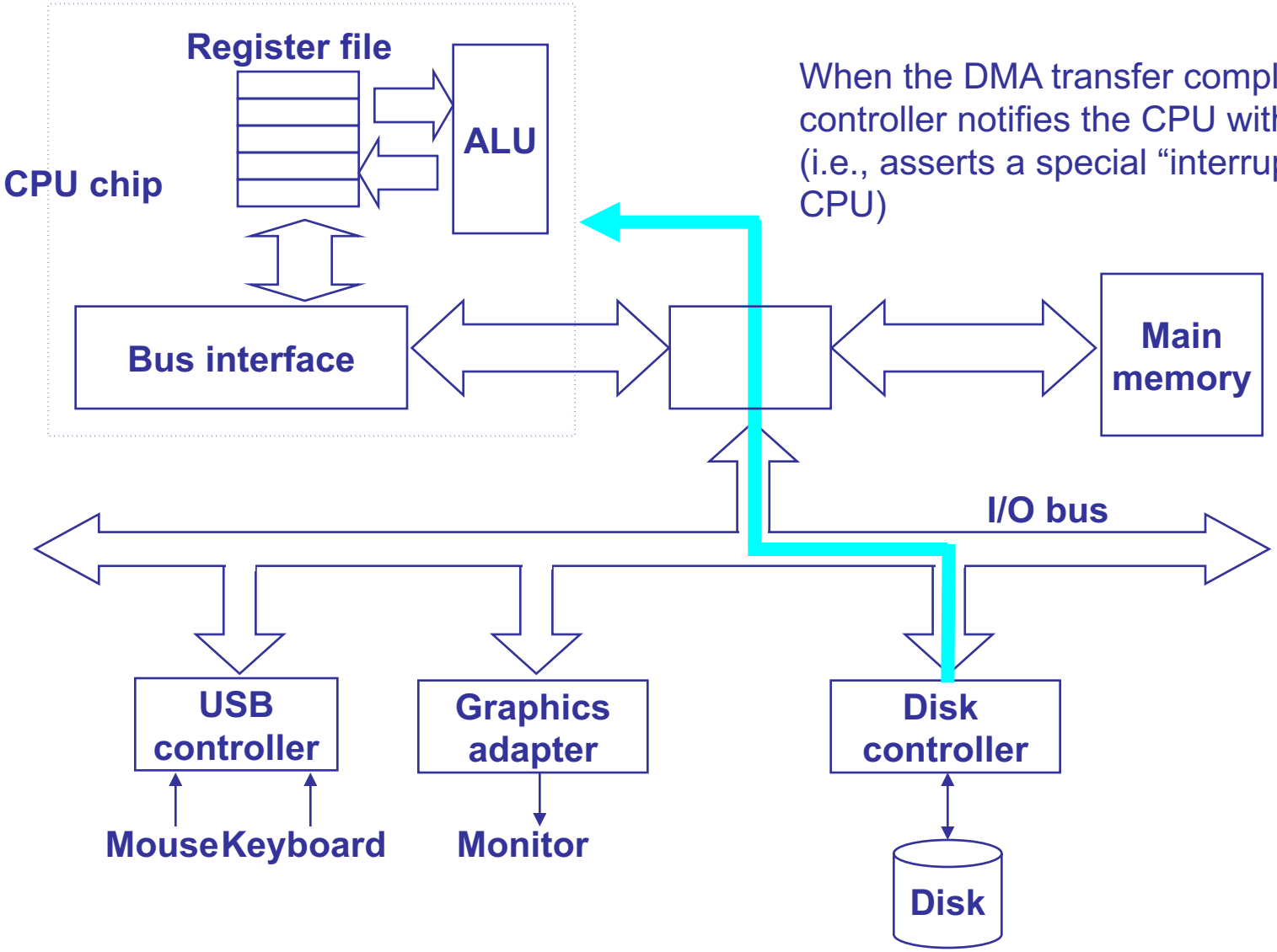
CPU initiates a disk read by writing a command, logical block number, and destination memory address to a port (address) associated with disk controller.

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)

Disks Heterogeneity

- Seagate Barracuda Pro 3.5" (workstation)
  - capacity: 2 TB - 10 TB (w/ 256 MB cache)
  - rotational speed: 7200 RPM
  - sequential read performance: 250 MB/s
  - seek time (average): 9 ms
- Seagate Barracuda 2.5" (laptop)
  - capacity: 500 GB – 5 TB (w/ 128 MB cache)
  - rotational speed: 5400 RPM
  - sequential read performance: 140 MB/s
  - seek time (average): 14 ms
- Seagate Firecuda 3.5" (gaming)
  - capacity: 1 TB - 2 TB (w/ 8 GB MLC NAND & 64 MB cache)
  - rotational speed: 7200 RPM
  - sequential performance is similar but IOPS is much better

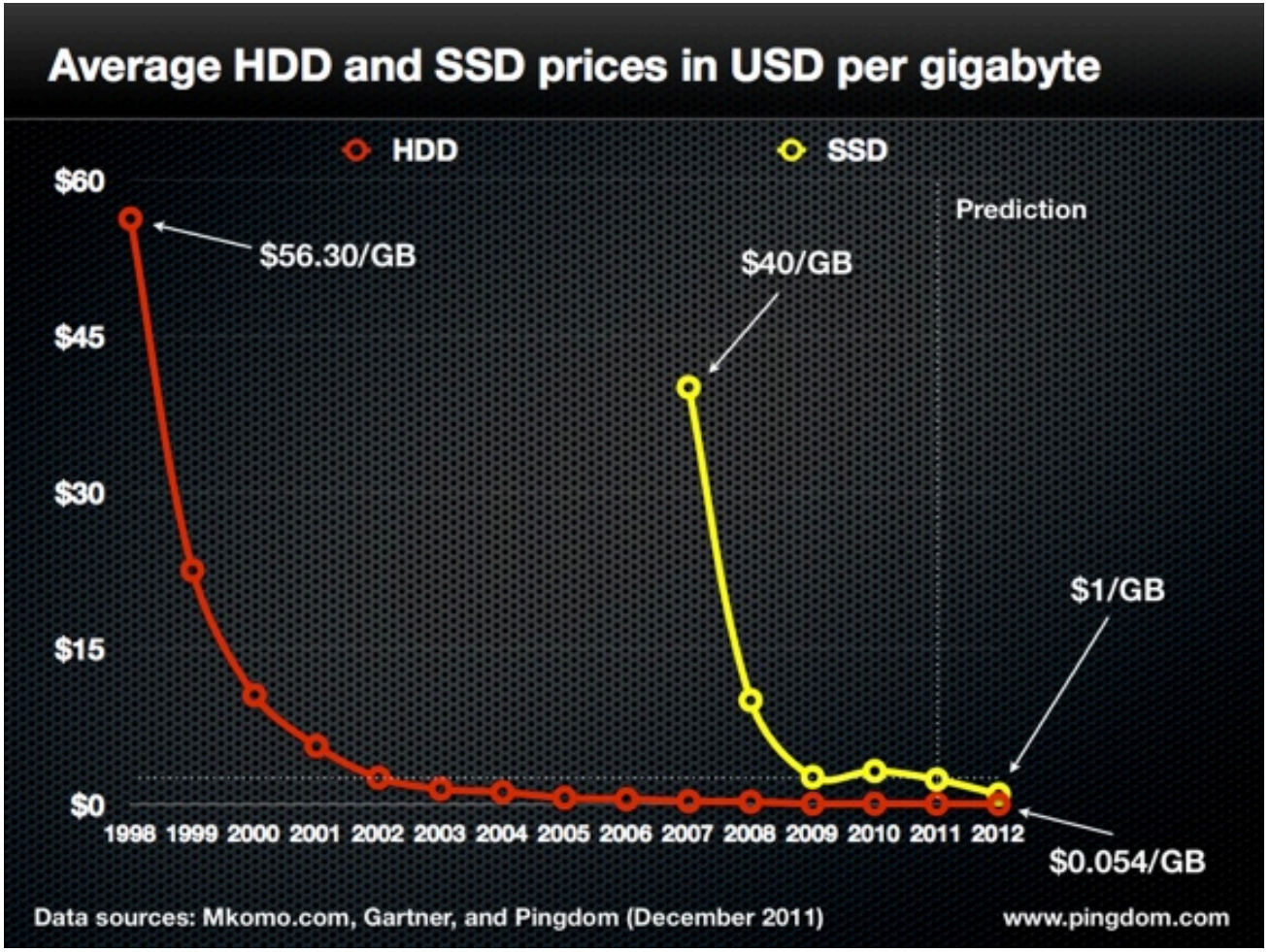
SSD/NVM

- Nonvolatile memories (NVMs) retain value
  - Read-only memory (ROM): programmed during production
  - Programmable ROM (PROM): can be programmed once
  - Eraseable PROM (EPROM): can be bulk erased (UV, X-Ray)
  - Electrically eraseable PROM (EEPROM): electronic erase
  - Flash memory (SSD, stick): EEPROMs with partial (sector) erase capability
    - » Wears out after about 100,000 erasings (SLC), 6,000 - 40,000 erasings (3D MLC), 1,000 - 3,000 (3D TLC), 100 - 1,000 (3D QLC).
  - Phase Change Memories (PCMs): Intel Optane
    - » More endurable, but also wear out

SSD Performance

- Samsung 860 QVO (3D QLC, SATA): $108
  - Seq Read/Write: 550/520 MB/s
  - 4KB Rand Read/Write: 96K/89K IOPS
- HP EX920 (3D TLC, NVMe): $135
  - Seq Read/Write: 3200/1800 MB/s
  - 4KB Rand Read/Write: 350K/250K IOPS
- Samsung 970 Pro (3D MLC, NVMe): $350
  - Seq Read/Write: 3500/2700 MB/s
  - 4KB Rand Read/Write: 500K/500K IOPS
- Intel Optane 905P (PCM, NVMe): $1,200
  - Seq Read/Write: 2600/2200 MB/s
  - 4KB Rand Read/Write: 575K/550K IOPS

Contrarian View



Disk Scheduling

- Because seeks are so expensive (milliseconds!), OS schedules requests that are queued waiting for the disk
  - FCFS (do nothing)
    - » Reasonable when load is low
    - » Does nothing to minimize overhead of seeks
  - SSTF (shortest seek time first)
    - » Minimize arm movement (seek time), maximize request rate
    - » Favors middle blocks, potential starvation of blocks at ends
  - SCAN (elevator)
    - » Service requests in one direction until done, then reverse
    - » Long waiting times for blocks at ends
  - C-SCAN
    - » Like SCAN, but only go in one direction (typewriter)

Disk Scheduling (2)

- In general, unless there are request queues, disk scheduling does not have much impact
  - Important for servers, less so for PCs
- Modern disks often do the disk scheduling themselves
  - Disks know their layout better than OS, can optimize better
  - Ignores, undoes any scheduling done by OS