Operating System

Chapter 10: Mass Storage Systems

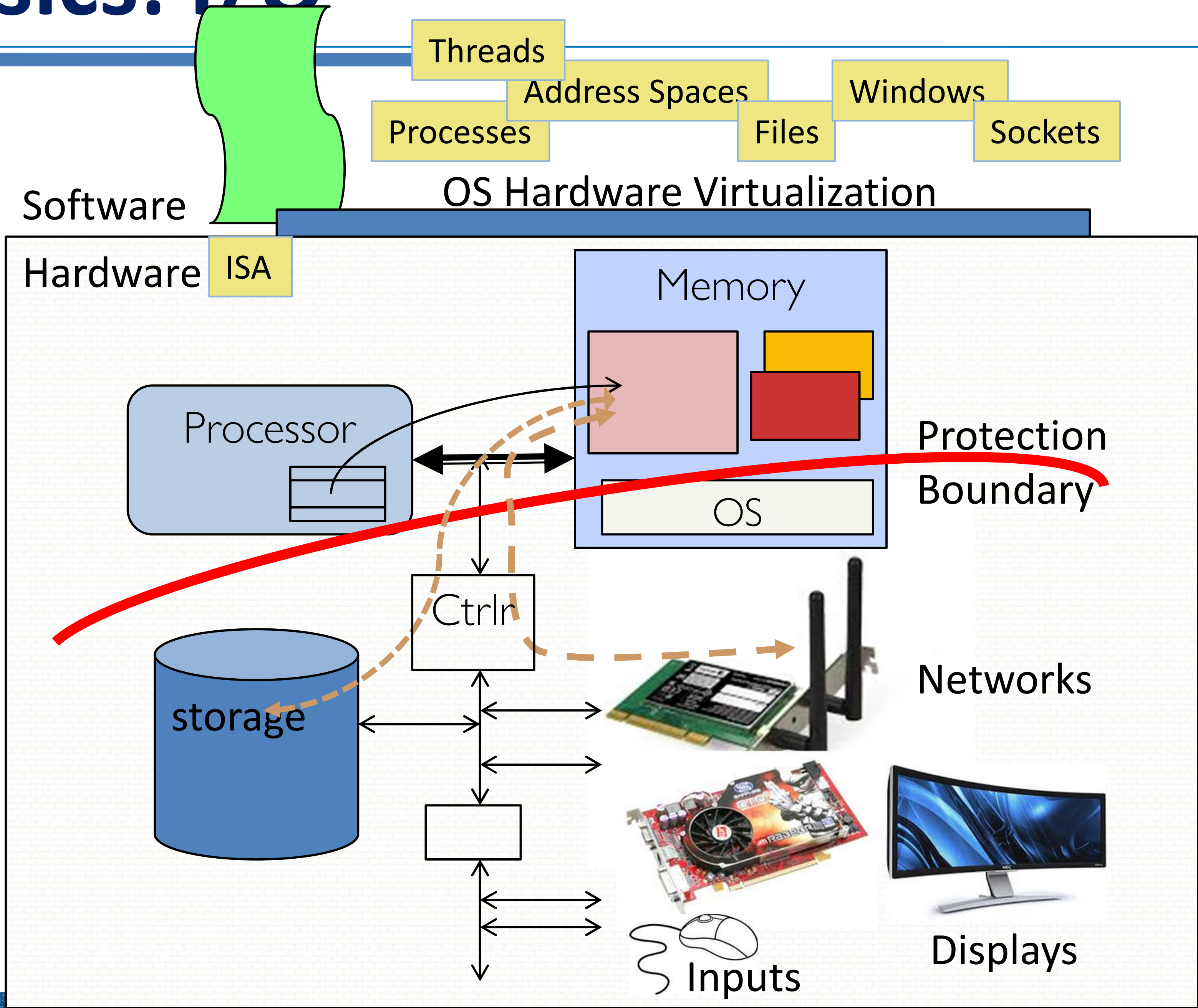
Objective

- To describe the physical structure of secondary storage devices and its effects on the uses of the devices.
- To explain the performance characteristics of mass-storage devices.
- To evaluate disk scheduling algorithms.
- To discuss operating-system services provided for mass storage, including RAID

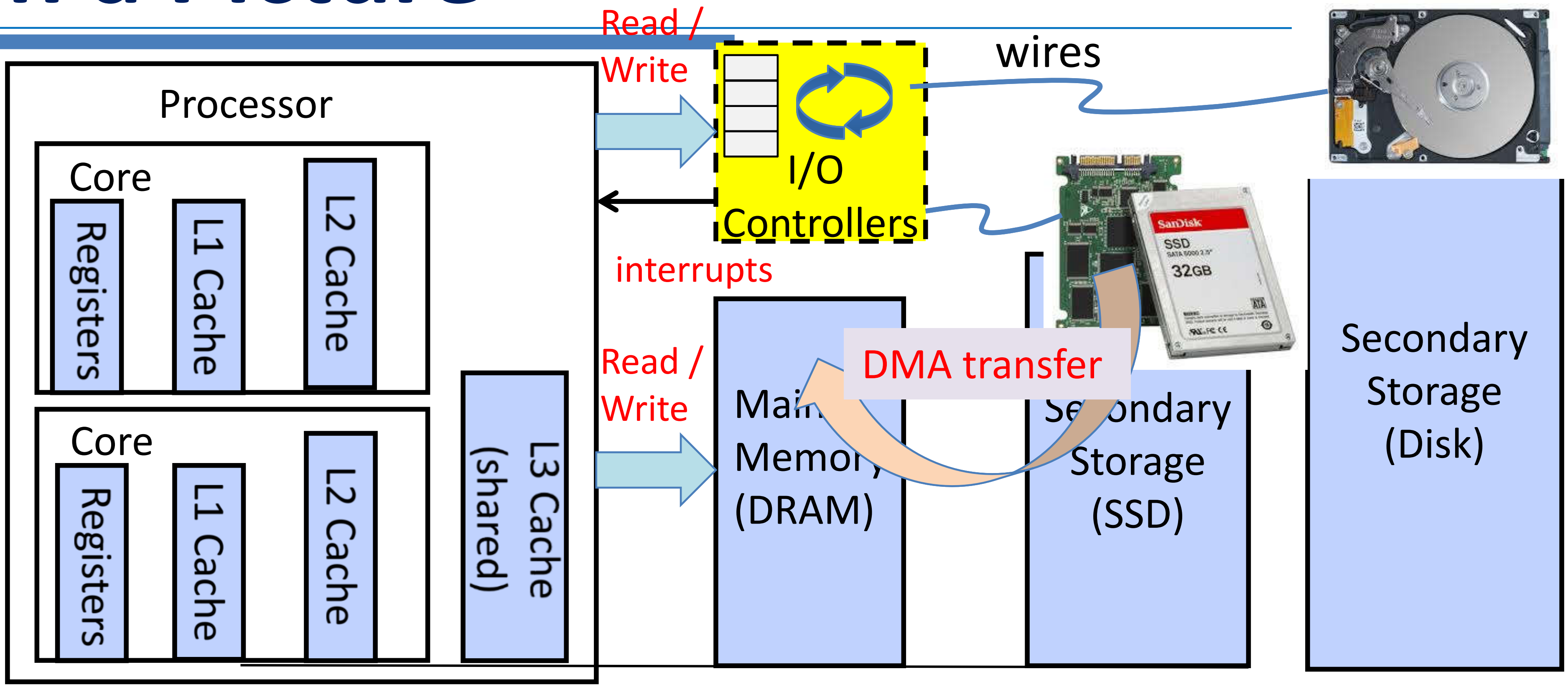
The Requirements of I/O

- So far in this course:
  - We have learned how to manage CPU and memory
- What about I/O?
  - Without I/O, computers are useless (disembodied brains?)
  - But… thousands of devices, each slightly different
    - How can we standardize the interfaces to these devices?
  - Devices unreliable: media failures and transmission errors
    - How can we make them reliable???
  - Devices unpredictable and/or slow
    - How can we manage them if we don’t know what they will do or how they will perform?

OS Basics: I/O


Threads
Address Spaces
Windows
Processes
Files
Sockets
Software
OS Hardware Virtualization
Hardware
ISA
Memory
Processor
OS
Protection
Boundary
Ctrlr
storage
Networks
Displays
Inputs

In a Picture



- I/O devices you recognize are supported by I/O Controllers
- Processors accesses them by reading and writing IO registers as if they were memory
  - Write commands and arguments, read status and results

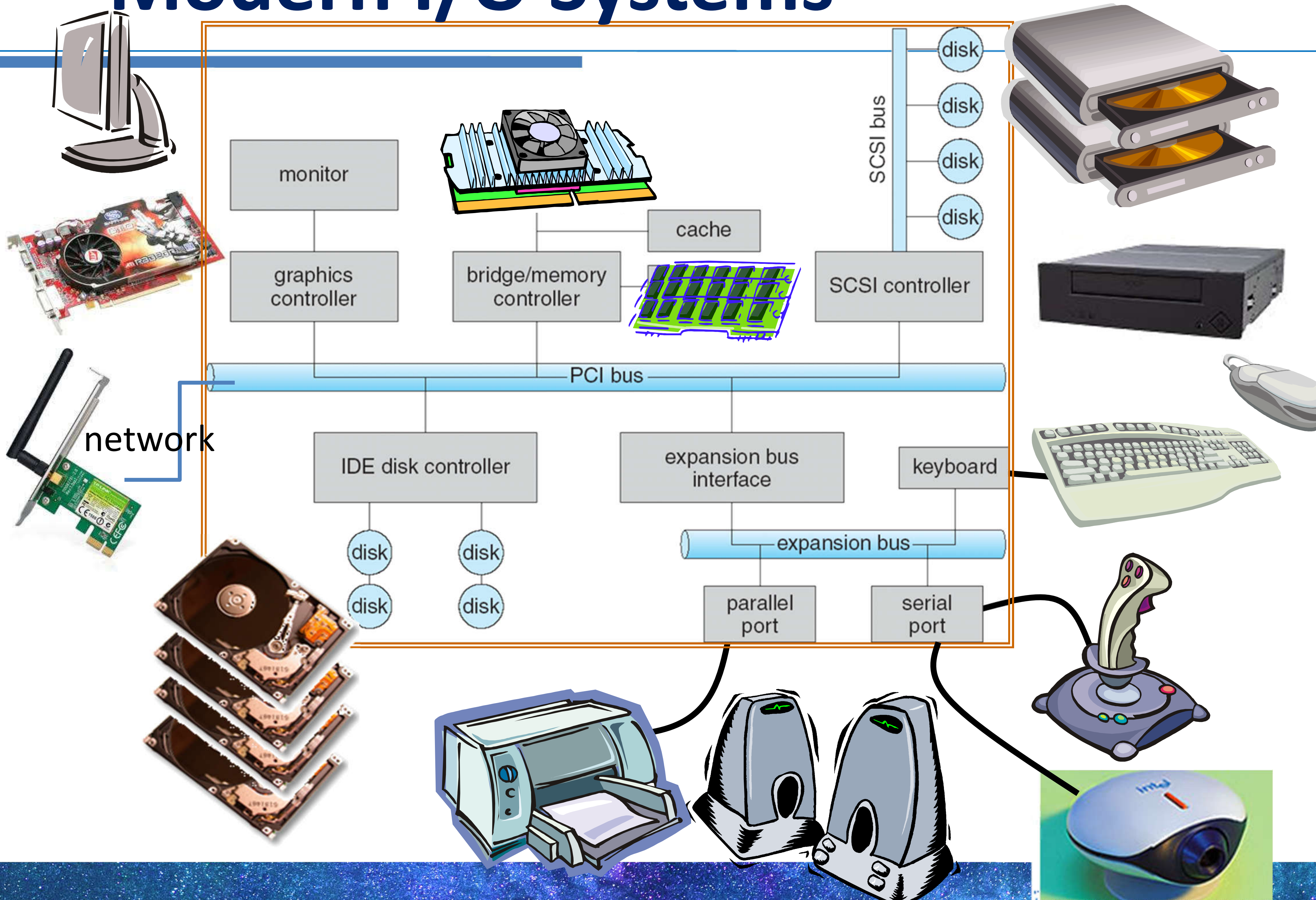
Operational Parameters for I/O

- Data granularity: Byte vs. Block
  - Some devices provide single byte at a time (e.g., keyboard)
  - Others provide whole blocks (e.g., disks, networks, etc.)

- Access pattern: Sequential vs. Random
  - Some devices must be accessed sequentially (e.g., tape)
  - Others can be accessed "randomly" (e.g., disk, cd, etc.)
    - Fixed overhead to start transfers
  - Some devices require continual monitoring
  - Others generate interrupts when they need service

- Transfer Mechanism: Programmed IO and DMA

Modern I/O Systems



Storage Devices

- Magnetic disks
  - Storage that rarely becomes corrupted
  - Large capacity at low cost
  - Block level random access (except for SMR – later!)
  - Slow performance for random access
  - Better performance for sequential access

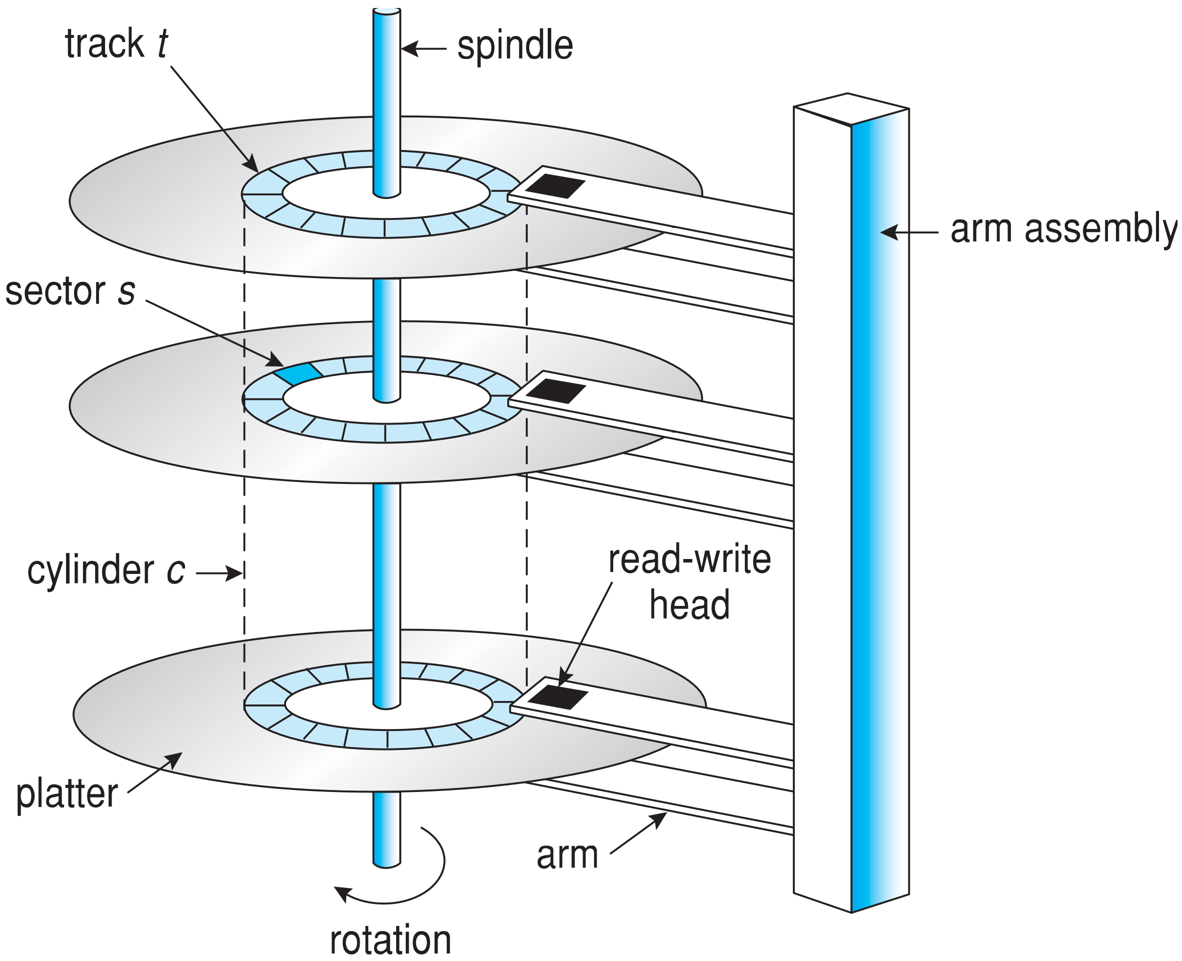
- Flash memory
  - Storage that rarely becomes corrupted
  - Capacity at intermediate cost (5-20x disk)
  - Block level random access
  - Good performance for reads; worse for random writes
  - Erasure requirement in large blocks
  - Wear patterns issue

Magnetic disk structure

Unit of Transfer: Sector
- Ring of sectors form a track
- Stack of tracks form a cylinder
- Heads position on cylinders

Disk Tracks ~ 1µm (micron) wide
- Wavelength of light is ~ 0.5µm
- Resolution of human eye: 50µm
- 100K tracks on a typical 2.5” disk

Separated by unused guard regions
- Reduces likelihood neighboring tracks are corrupted during writes (still a small non-zero chance)



Overview of mass storage structure

- **Magnetic disks**
  - Bulk of secondary storage of modern computers
  - **Transfer rate** is rate at which data flow between drive and computer
  - **Positioning time** (**random-access time**)
    - **Seek time:** time to move disk arm to desired cylinder
    - **Rotational latency:** time for desired sector to rotate under the disk head
  - **Head crash** results from disk head making contact with the disk surface -- That's bad

- Some drives attached to computer via **I/O bus**
  - Busses vary, including **EIDE**, **ATA**, **SATA**, **USB**, **Fiber Channel**, **SCSI, SAS, Firewire**
  - **Host controller** in computer uses bus to talk to **disk controller** built into drive or storage array

Magnetic disk performance

- Access Latency = Average access time = average seek time + average rotation latency
  - For fastest disk 3ms + 2ms = 5ms
  - For slow disk 9ms + 5.56ms = 14.56ms

- Average I/O time = average access time

  + (amount to transfer / transfer rate)

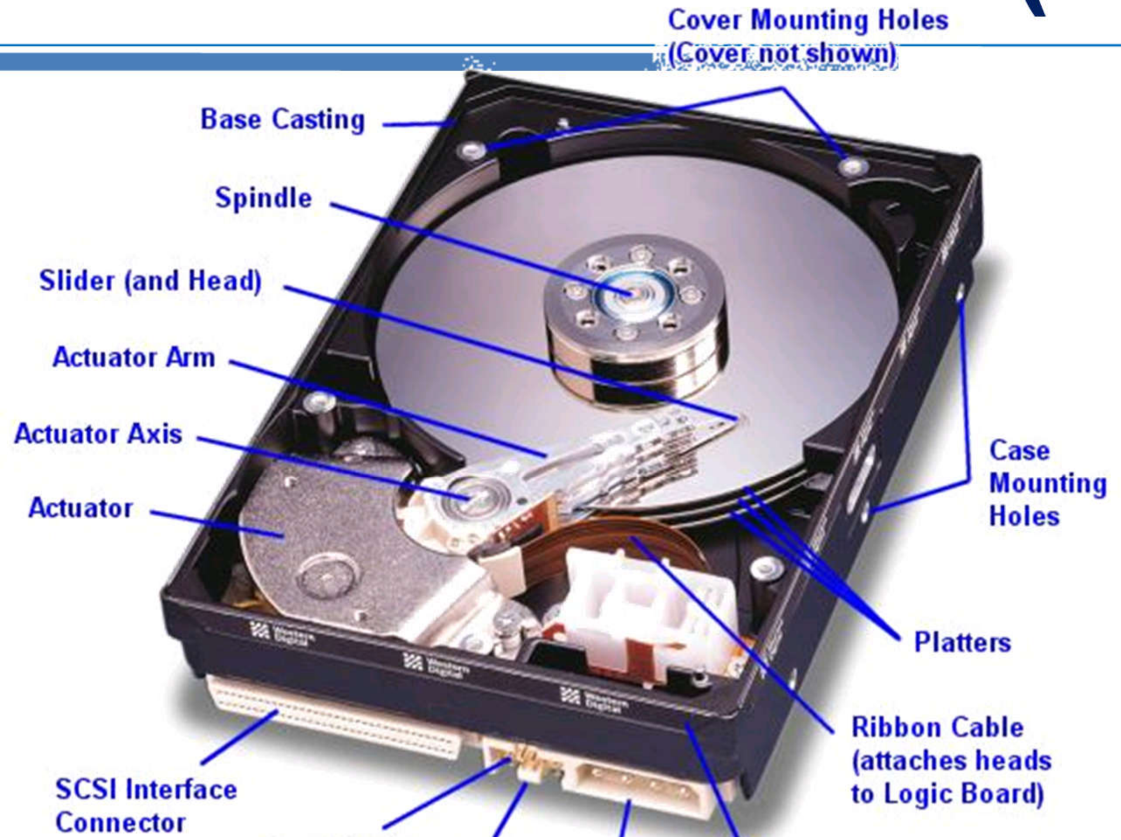
  + controller overhead

1st commercial disk drive

1956
IBM RAMDAC computer
included the IBM Model 350
disk storage system

5M (7 bit) characters
50 x 24” platters
Access time = < 1 second

Hard Disk Drives (HDDs)



Western Digital Drive
http://www.storagereview.com/guide/

Read/Write Head Side View

IBM/Hitachi Microdrive

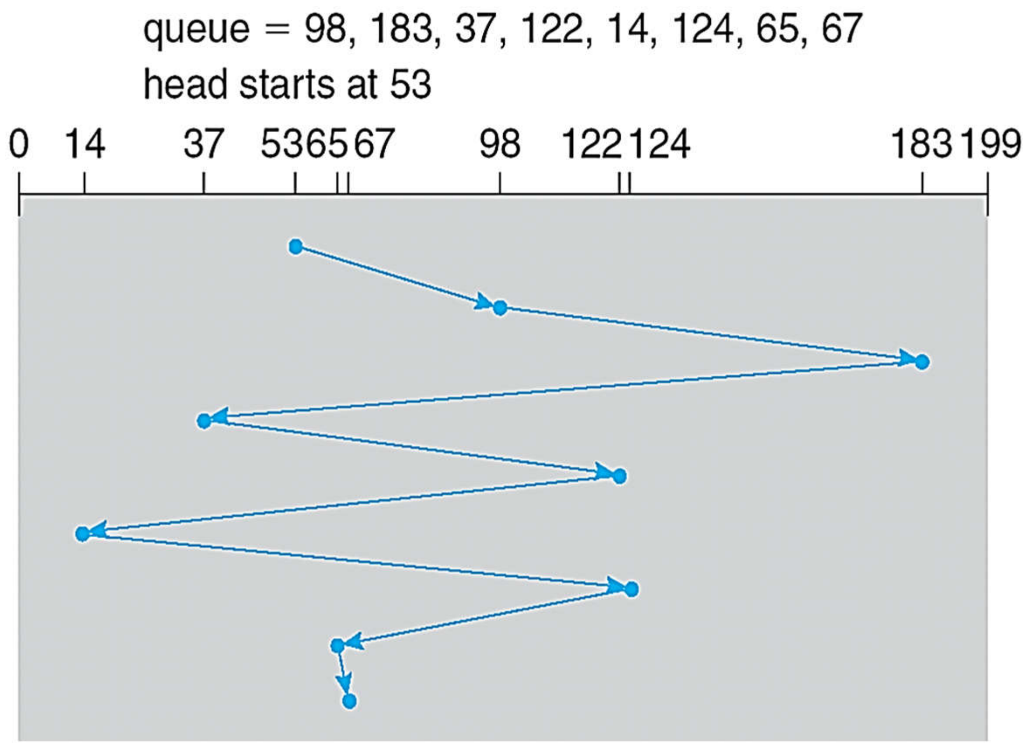
IBM Personal Computer/AT (1986)
- 30 MB hard disk - $500
- 30-40ms seek time
- 0.7-1 MB/s (est.)

Disk scheduling

- The operating system is responsible for using hardware efficiently — for the disk drives, this means having a fast access time and disk bandwidth
- Minimize seek time
- Seek time ≈ seek distance
- Disk **bandwidth**
  - The total number of bytes transferred, divided by the total time between the first request for service and the completion of the last transfer

FCFS (First come first serve)

Response to request queue based on FCFS
Head movement = 640 cylinders

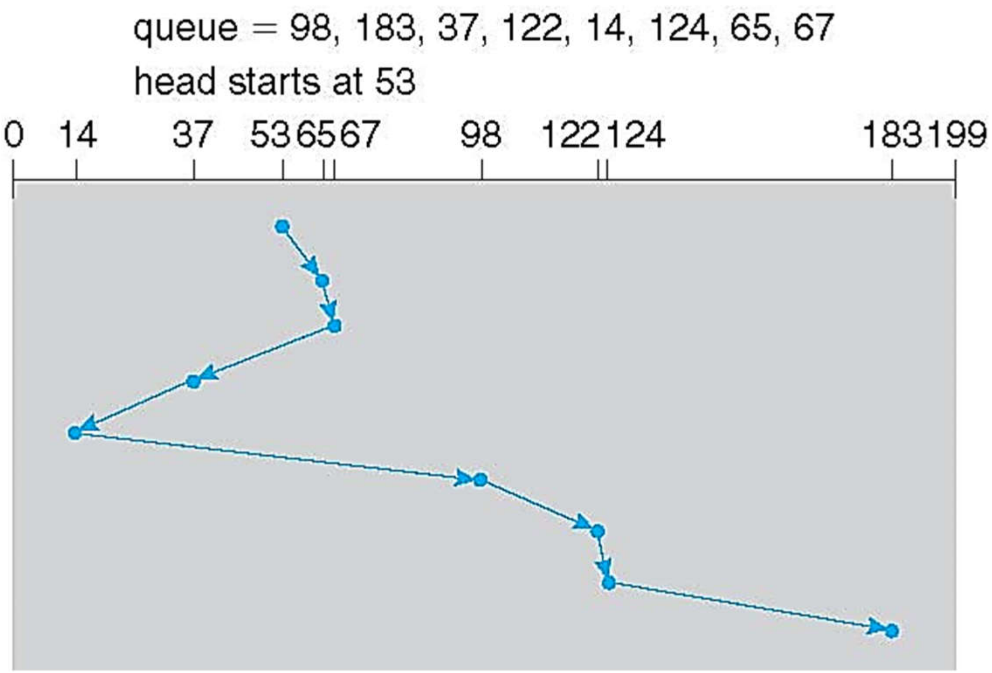


SSTF (Shortest Seek Time First)

Selects the request with the minimum seek time from the current head position

SSTF scheduling is a form of SJF scheduling; may cause starvation of some requests

Head movement = 208 cylinders

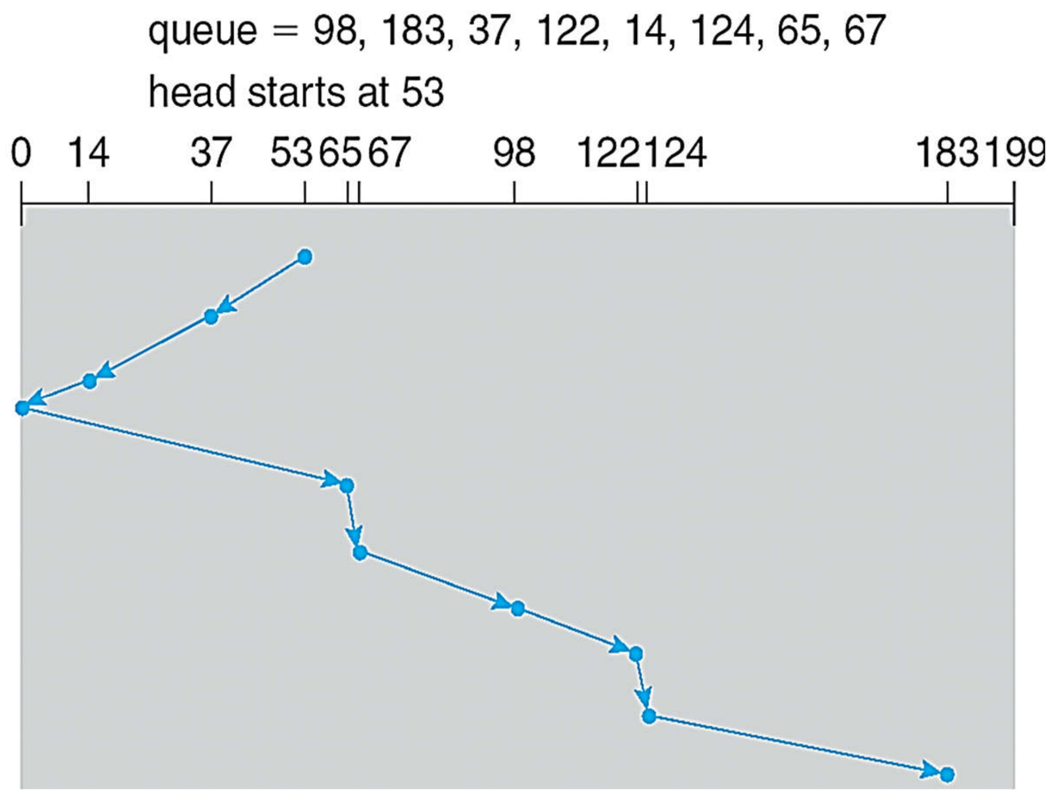


SCAN

- The disk arm starts at one end of the disk, and moves toward the other end
  - Servicing requests until it gets to the other end of the disk, where
  - The head movement is reversed and servicing continues.

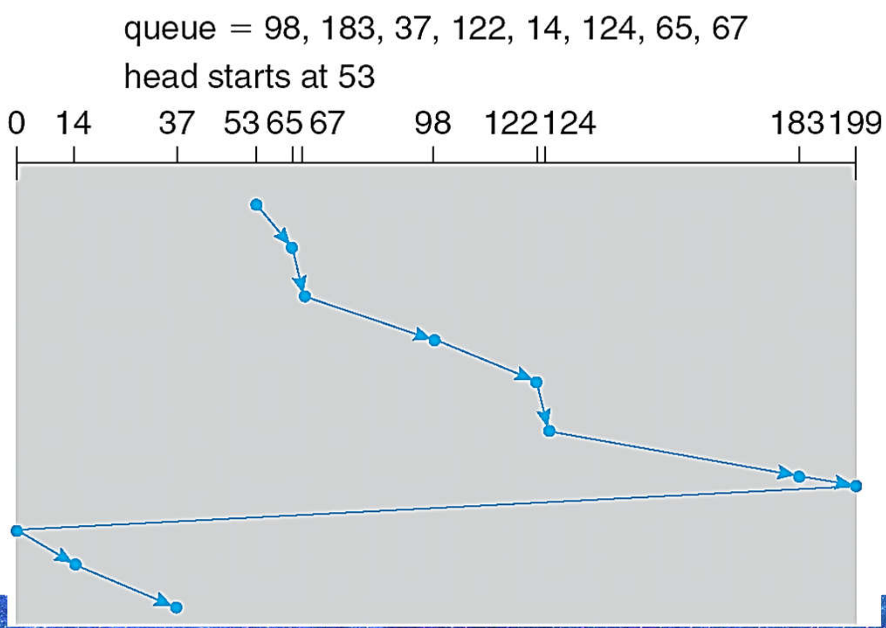
- SCAN algorithm sometimes called the elevator algorithm

- **If requests are uniformly dense, largest density at other end of disk and those wait the longest**



C-SCAN (Circular SCAN)

- Provides a more uniform wait time than SCAN
- The head moves from one end of the disk to the other, servicing requests as it goes
  – When it reaches the other end, however, it immediately returns to the beginning of the disk, without servicing any requests on the return trip
- Treats the cylinders as a circular list that wraps around from the last cylinder to the first one



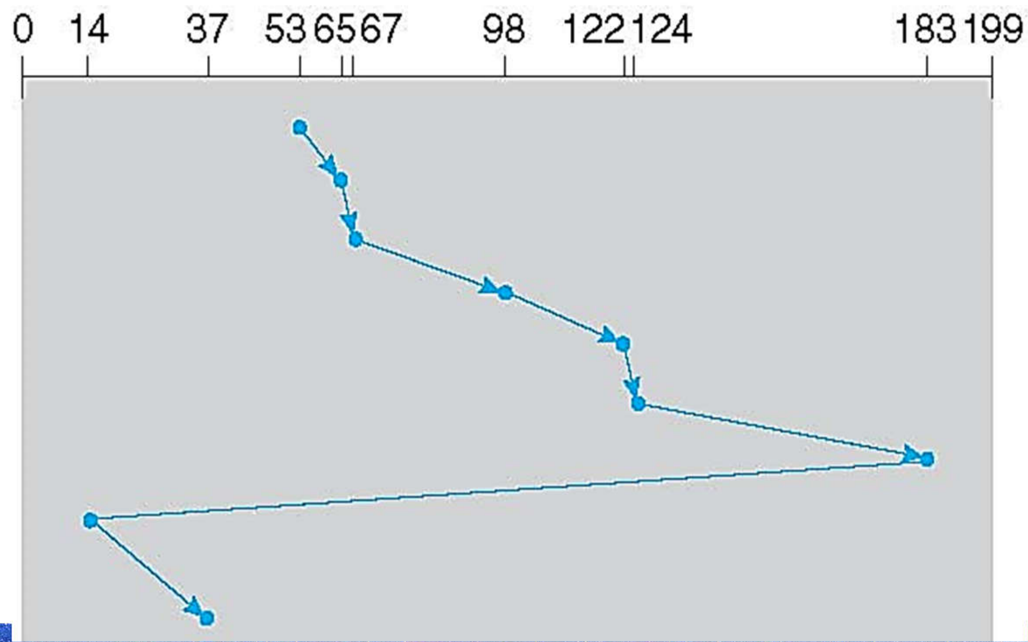
LOOK and C-LOOK (Circular LOOK)

- LOOK a version of SCAN
- C-LOOK a version of C-SCAN

- Arm only goes as far as the last request in each direction, then reverses direction immediately, without first going all the way to the end of the disk

queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53



Which one is better?

- SSTF is common and has a natural appeal: good performance
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk: less starvation
- Performance depends on the number and types of requests

Disk Attachment

Disk attachment

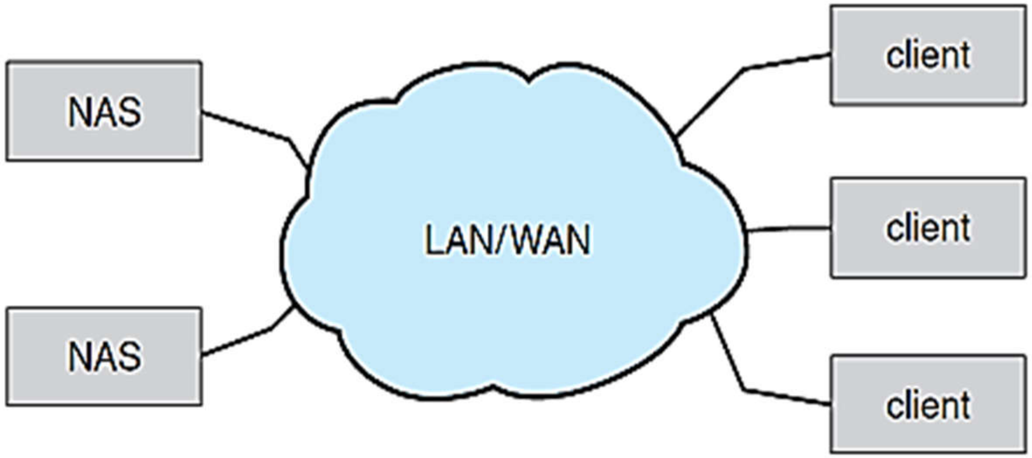
- 1) Host-attached storage
- 2) Network-attached storage (NAS)
- 3) Storage-area network (SAN)

1) Host-attached storage

- Accessed through PC I/O ports talking to I/O busses
- The typical desktop PC uses an I/O bus architecture, called IDE or ATA
  - Maximum of 2 drivers per I/O bus
  - A newer, similar protocol that has simplified cabling is SATA
- SCSI is a bus, up to 16 devices on one cable
- Fiber Channel (FC) is high-speed serial architecture

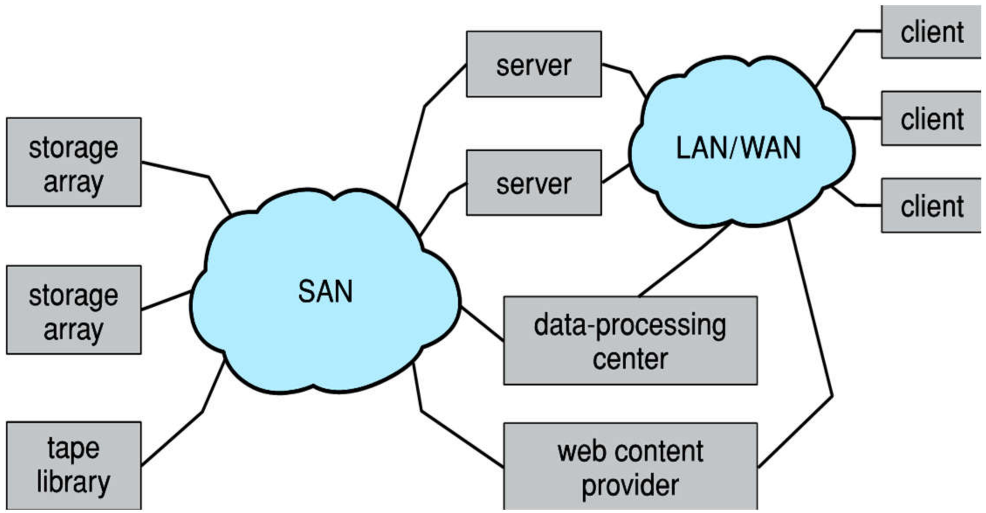
2) Network-attached storage (NAS)

- Storage over a network rather than a local connection
- Remotely attaching to file systems
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage over typically TCP or UDP on IP network



3) Storage-area network (SAN)

- A method for large storage environments
- Multiple hosts attached to multiple storage arrays
- Storage arrays and Hosts are connected to one or more Fiber Channel Switches



Disk Management

Disk management

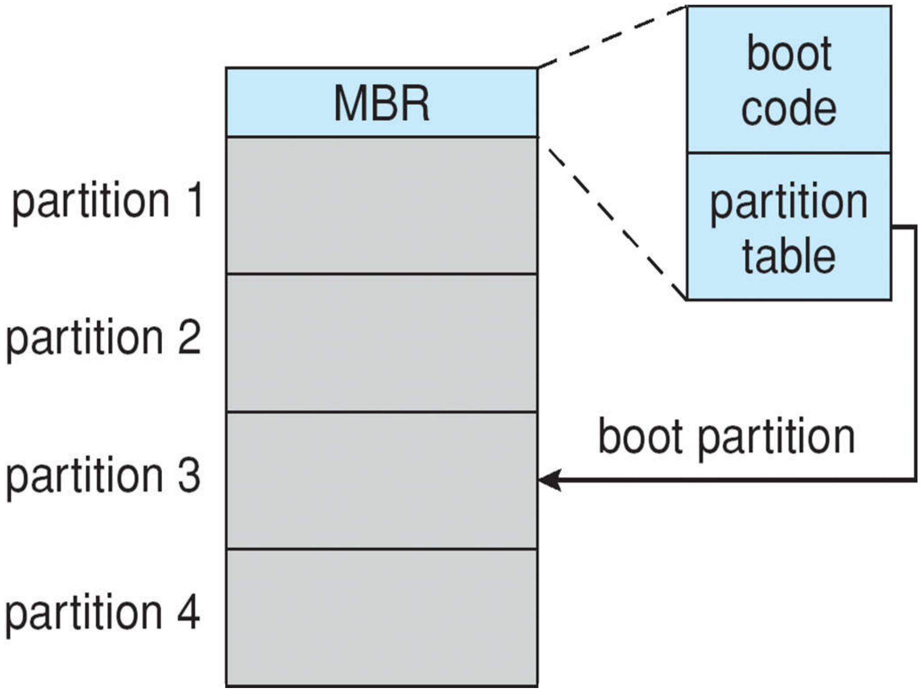
- Disk formatting
- Boot block
- Bad blocks

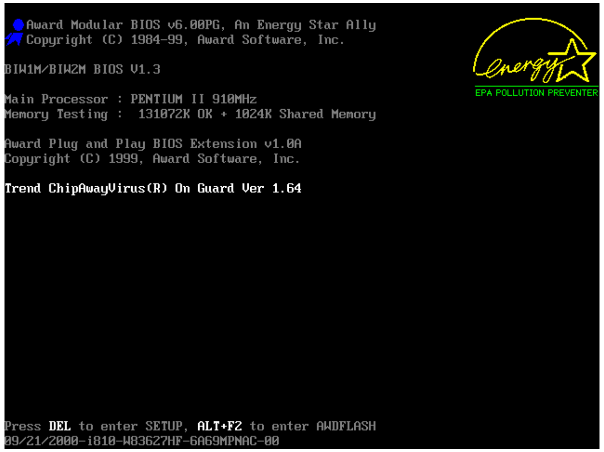
Disk formatting

- **Low-level formatting**, or **physical formatting**
  - Dividing a disk into sectors
  - Each sector can hold header information, plus data, plus error correction code (**ECC**)
  - Usually 512 bytes of data but can be selectable
- OS data structures to save files
  - **Partition** the disk into one or more groups of cylinders, each treated as a logical disk
  - **Logical formatting** or "making a file system"
  - To increase efficiency most file systems group blocks into **clusters**
    - Disk I/O done in blocks
    - File I/O done in clusters

Boot block

- Boot block initializes system
  - The bootstrap is stored in ROM
  - **Bootstrap loader** program stored in boot blocks of boot partition





Bad blocks

- The controller maintains a list of bad blocks on the disk
- The list is initialized during the low-level formatting at the factory and is updated over the file of the disk
- Sector sparing or forwarding: replacing each bad sector logically with one of the spare sectors.

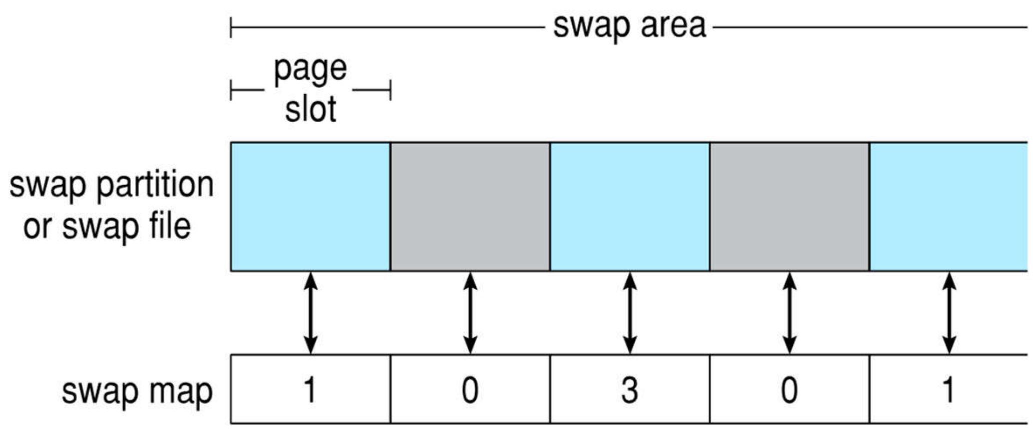
Swap space management

Swap-space management

- Swap-space — Virtual memory uses disk space as an extension of main memory

- Less common now due to memory capacity increases

- Swap-space
  - Normal file system, OR separate disk partition (raw)



RAID Structures

RAID

- RAID – Redundant Array of Inexpensive Disks

➢ Multiple disk drives provides reliability via **redundancy**

- Increases the **mean time to failure**

RAID Levels

RAID level 0 refers to disk arrays with striping at the level of blocks but without any redundancy.

RAID level 1 refers to disk mirroring.

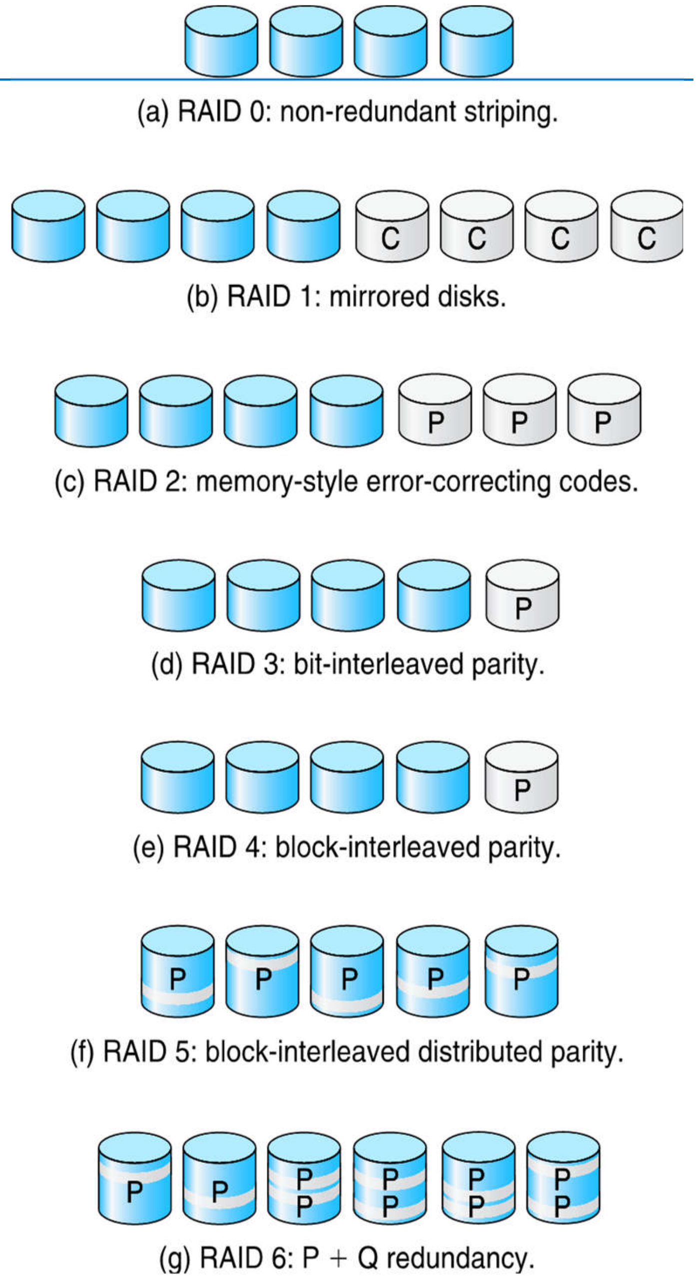
RAID level 2 is also known as memory-style error-correcting-code (ECC) organization.

RAID level 3 or bit-interleaved parity organization, improves on level 2 by taking into account a single parity bit can be used for error correction as well as for detection

RAID level 4 or block-interleaved parity organization, uses block-level striping and in addition keeps a parity block.

RAID level 5 or block-interleaved distributed parity, spreads data and parity among all *N*+1 disks.

RAID level 6 also called the **P + Q redundancy scheme,** stores extra redundant information to guard against multiple disk failures



(a) RAID 0: non-redundant striping.

(b) RAID 1: mirrored disks.

(c) RAID 2: memory-style error-correcting codes.

(d) RAID 3: bit-interleaved parity.

(e) RAID 4: block-interleaved parity.

(f) RAID 5: block-interleaved distributed parity.

(g) RAID 6: P + Q redundancy.

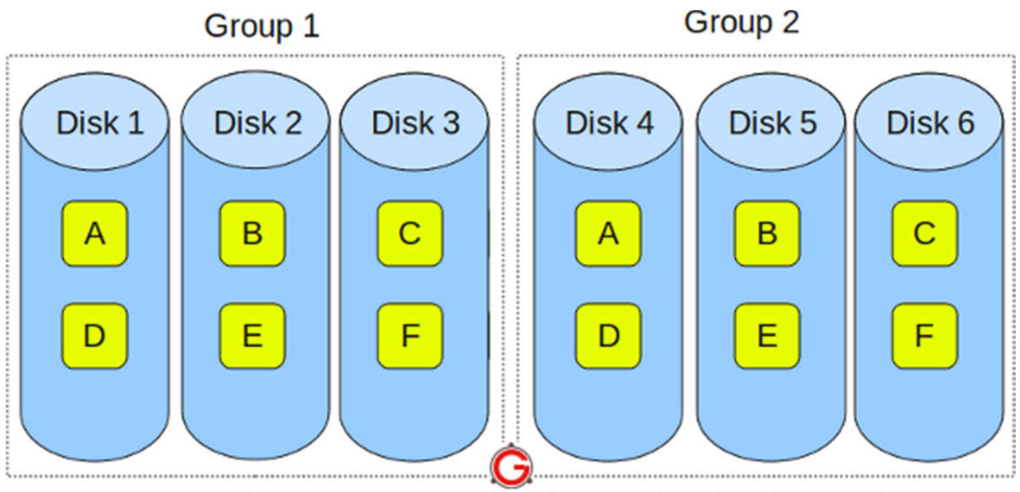
RAID (0 + 1) and (1 + 0)

- RAID level 0 + 1 refers to a combination of RAID levels 0 and 1 .
- RAID 0 provides the performance, while RAID 1 provides the reliability

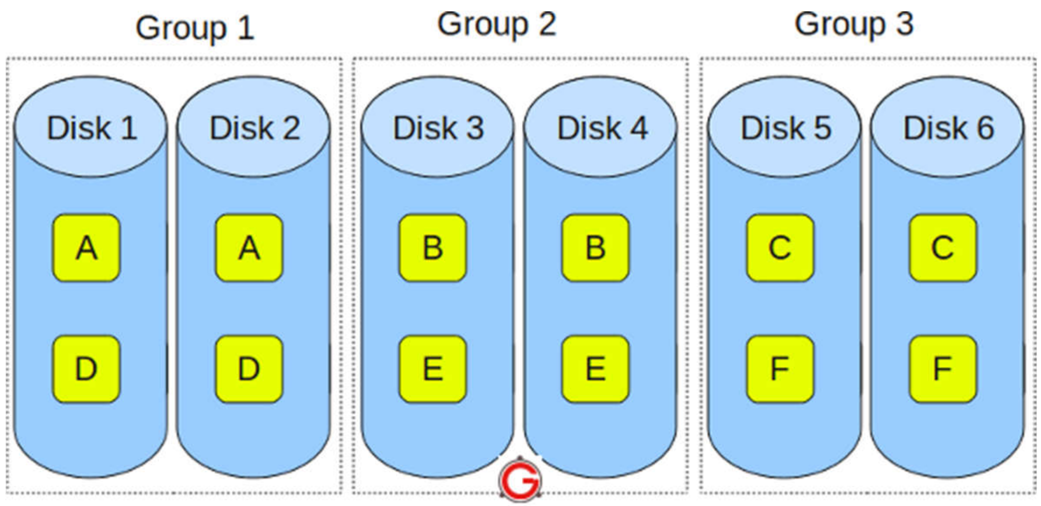
RAID 01: a set of disks are striped, and then the stripe is mirrored to another, equivalent stripe

RAID 10: disks are mirrored in pairs and then the resulting mirrored pairs are striped

•Performance on both RAID 10 and RAID 01 will be the same.
•The storage capacity on these will be the same.
•The main difference is the fault tolerance level.



RAID 01 – Blocks Striped. (and Blocks Mirrored)



RAID 10 – Blocks Mirrored. (and Blocks Striped)

Stable-Storage Implementation

Stable-storage implementation

- Stable storage: data is never lost (due to failures, etc)
- Write-ahead log (WAL) scheme requires stable storage
- In a system using WAL, all modifications are written to a log before they are applied.

What are failure's effects?

1.Successful completion

The data were written correctly on disk

2.Partial failure

A failure occurred in the midst of transfer, so only some of the sectors were written with the new data, and the sector being written during the failure may have been corrupted

3.Total failure

The failure occurred before the disk write started, so the previous data values on the disk remain intact

How to implement stable storage?

- If failure occurs during block write, recovery procedure restores block to consistent state
  - System maintains 2 physical blocks per logical block
    1. Write to 1st physical
    2. When successful, write to 2nd physical
    3. Declare complete only after second write completes successfully

➢ Systems frequently use NVRAM (Non-Volatile RAM) as one physical to accelerate

Questions?