COP 4610

Operating System Principles

Mass Storage

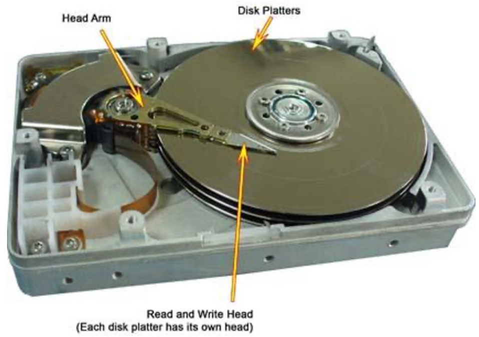
1

Mass-Storage Systems

- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Stable-Storage Implementation

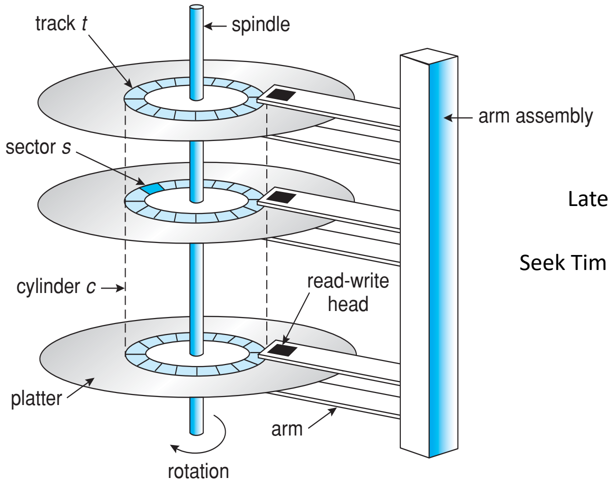
2

Hard Disks



Inside Hard Disk

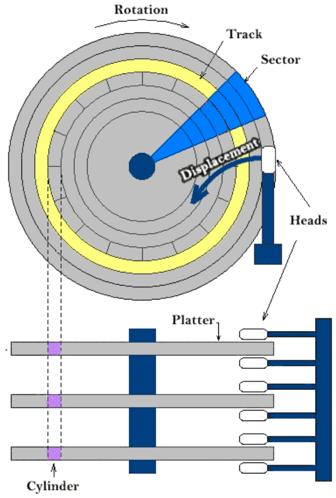
Moving-Head Disk Mechanism



Latency – Rotation time

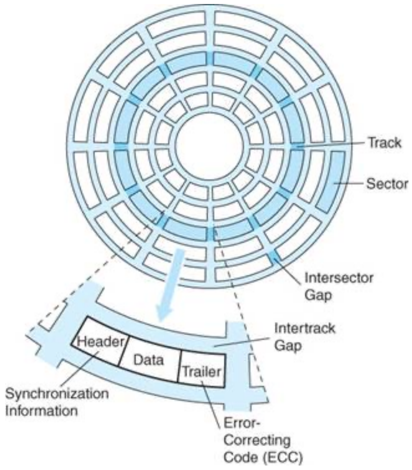
Seek Time – Time to move the head

Disk Structure



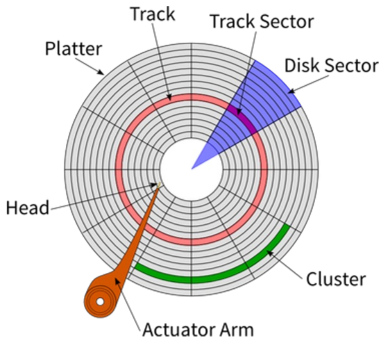
5

Disk Structure



6

Disk Structure



Term	Description
Platter	Magnet surface on which data is stored
Spindle	Platters are bound together around this component which has a motor
Track	Data is encoded in these concentric circles
Disk Head	This is used to read / write data to and from the platter
Disk Arm	This moves the head to the desired track

Overview of Mass Storage Structure

- **Magnetic disks** provide bulk of secondary storage of modern computers
  - Drives rotate at 60 to 250 times per second
  - **Transfer rate** is rate at which data flow between drive and computer
  - **Positioning time** (**random-access time**) is time to move disk arm to desired cylinder (**seek time**) and time for desired sector to rotate under the disk head (**rotational latency**)
  - **Head crash** results from disk head making contact with the disk surface
- Disks can be removable
- Drive attached to computer via **I/O bus**
  - Busses vary, including **EIDE**, **ATA**, **SATA**, **USB**, **Fibre Channel**, **SCSI, SAS, Firewire**
  - **Host controller** in computer uses bus to talk to **disk controller** built into drive or storage array

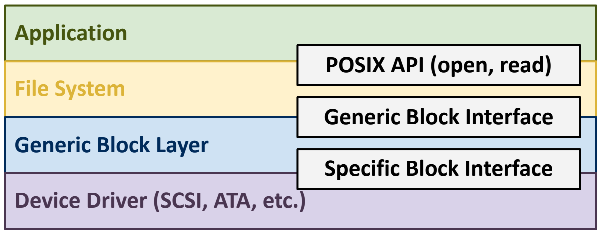
Data Transfer Modes

- There are three methods to transfer data from memory to and from disk storage:

Method	Description	Pros	Cons
Programmed I/O (PIO)	OS is directly involved data movement	Simple and works	Wastes CPU time polling
Interrupts	OS issues command and later handles interrupt on completion	Allows for overlap	Can be wasteful due to context switch, interrupt storm
Direct Memory Access (DMA)	OS programs DMA engine to handle data transfers	Allows for overlap	Need more hardware, synchronization

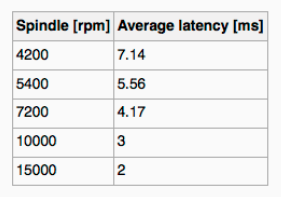
Data Transfer

- To interface with different devices, the **operating system** layers subsystems on top of different hardware specific **device drivers**:



Magnetic Disks

- Platters range from .85” to 14” (historically)
  - Commonly 3.5”, 2.5”, and 1.8”
- Range from 30GB to 20TB per drive
- Performance
  - **Seek time** from 3ms to 12ms – 9ms common for desktop drives
  - **Rotational latency** based on spindle speed
  - **Transfer Rate** – theoretical – 6 Gb/sec (effective Transfer Rate – real – 1Gb/sec)



Spindle [rpm]	Average latency [ms]
4200	7.14
5400	5.56
7200	4.17
10000	3
15000	2

(From Wikipedia)

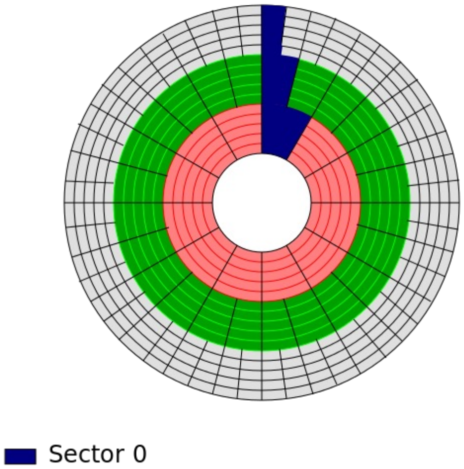
Magnetic Disk Performance

- **Access Latency = Average access time** = average seek time + average latency
  - For fastest disks: 3ms + 2ms = 5ms
  - For slow disks: 9ms + 5.56ms = 14.56ms
- **Average I/O time** = average access time + (amount to transfer / transfer rate) + controller overhead
- For example to transfer a 4KB block on a 7200 RPM disk with a 5ms average seek time, 1Gb/sec transfer rate with a .1ms controller overhead =
  - 5ms + 4.17ms + 4KB / 1Gb/sec + 0.1ms =
  - 9.27ms + 4 / 131072 sec =
  - 9.27ms + .12ms = 9.39ms

Disk Structure

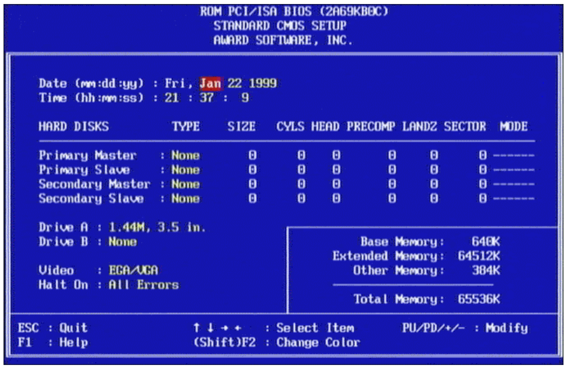
- Disk drives are addressed as large 1-dimensional arrays of **logical blocks**, where the logical block is the smallest unit of transfer
  - Low-level formatting creates **logical blocks** on physical media
- The 1-dimensional array of logical blocks is mapped into the sectors of the disk sequentially
  - Sector 0 is the first sector of the first track on the outermost cylinder
  - Mapping proceeds in order through that track, then the rest of the tracks in that cylinder, and then through the rest of the cylinders from outermost to innermost
  - Logical to physical address should be easy
    - Except for bad sectors
    - Non-constant # of sectors per track via constant angular velocity

Zone Bit Recording



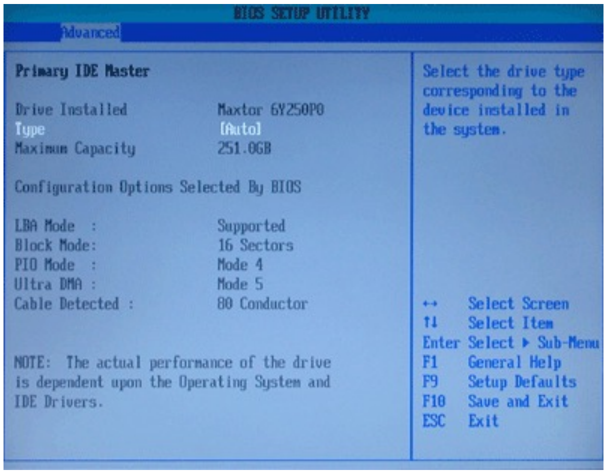
LBA vs. Non-LBA

- Logical Block Addressing
- Old school – must know disk geometry



```
                         ROM PCI/ISA BIOS (2A69KB0C)
                             STANDARD CMOS SETUP
                            AWARD SOFTWARE, INC.

  Date (mm:dd:yy) : Fri, Jan 22 1999
  Time (hh:mm:ss) : 21 : 37 :  9

  HARD DISKS          TYPE    SIZE   CYLS HEAD PRECOMP LANDZ SECTOR  MODE

  Primary Master    : None      0      0    0       0     0      0 ------
  Primary Slave     : None      0      0    0       0     0      0 ------
  Secondary Master  : None      0      0    0       0     0      0 ------
  Secondary Slave   : None      0      0    0       0     0      0 ------

  Drive A : 1.44M, 3.5 in.
  Drive B : None                                Base Memory:    640K
                                            Extended Memory:  64512K
  Video    : EGA/VGA                           Other Memory:    384K
  Halt On  : All Errors
                                               Total Memory:  65536K

ESC : Quit                ↑ ↓ → ←    : Select Item        PU/PD/+/- : Modify
F1  : Help                (Shift)F2  : Change Color
```

PNP – Plug and Play



```
                       BIOS SETUP UTILITY
        Advanced

 Primary IDE Master                           Select the drive type
                                              corresponding to the
 Drive Installed          Maxtor 6Y250P0      device installed in
 Type                     [Auto]              the system.
 Maximum Capacity         251.0GB

 Configuration Options Selected By BIOS

 LBA Mode   :             Supported
 Block Mode:              16 Sectors
 PIO Mode   :             Mode 4
 Ultra DMA  :             Mode 5
 Cable Detected :         80 Conductor        ↔     Select Screen
                                              ↑↓    Select Item
                                              Enter Select ▶ Sub-Menu
 NOTE:  The actual performance of the drive   F1    General Help
 is dependent upon the Operating System and   F9    Setup Defaults
 IDE Drivers.                                 F10   Save and Exit
                                              ESC   Exit
```

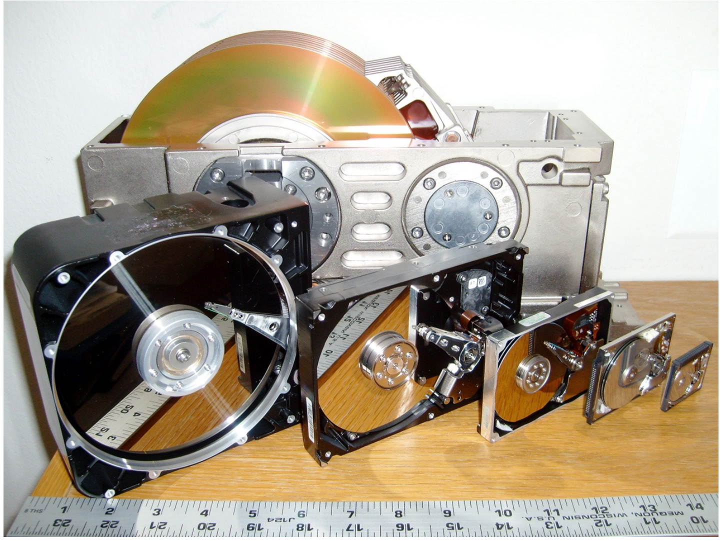
The First Commercial Disk Drive

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

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

17

Modern Disk Drives



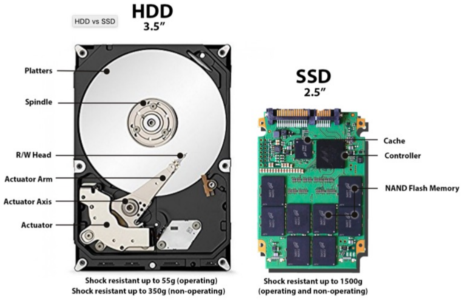
18

SSDs (Solid State Disks): Overview

Rather than use **slow spinning platters**, we can use **fast non-volatile NAND** memory to store data.

- **Faster**
- **Energy efficient**

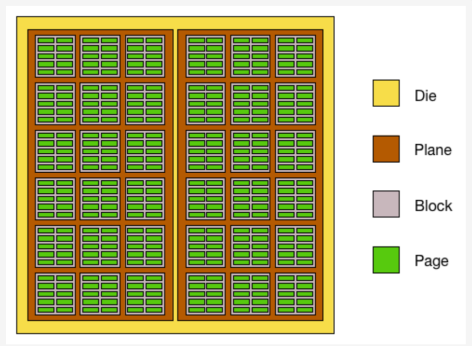
- **More expensive**
- **Unknown reliability**



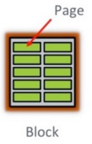
Shock resistant up to 55g (operating)
Shock resistant up to 350g (non-operating)

Shock resistant up to 1500g
(operating and non-operating)

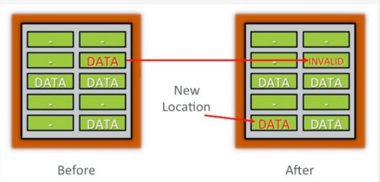
SSD Structure



NAND Flash Die Layout (image courtesy of AnandTech)



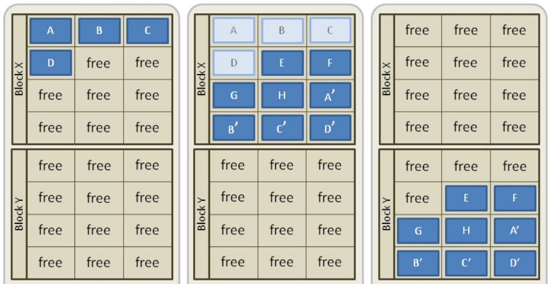
Operation	Area
Read	Page
Program (Write)	Page
Erase	**Block**



Updating a page in NAND flash. Note that the new page location does not need to be within the same block, or even the same flash die. It is shown in the same block here purely for ease of drawing.

SSDs: Garbage Collection

- Rather than overwrite data, we always write new pages of data and periodically collect any **garbage**:



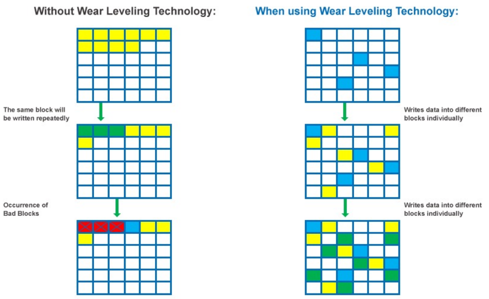
1. Four pages (A-D) are written to a block (X). Individual pages can be written at any time if they are currently free (erased).

2. Four new pages (E-H) and four replacement pages (A'-D') are written to the block (X). The original A-D pages are now invalid (stale) data, but cannot be overwritten until the whole block is erased.

3. In order to write to the pages with stale data (A-D) all good pages (E-H & A'-D') are read and written to a new block (Y) then the old block (X) is erased. This last step is *garbage collection*.

SSDs: Wear Leveling

Since each individual memory cell has a **limited number of write cycles**, we must avoid repeated writing to the same area.



Without Wear Leveling Technology:

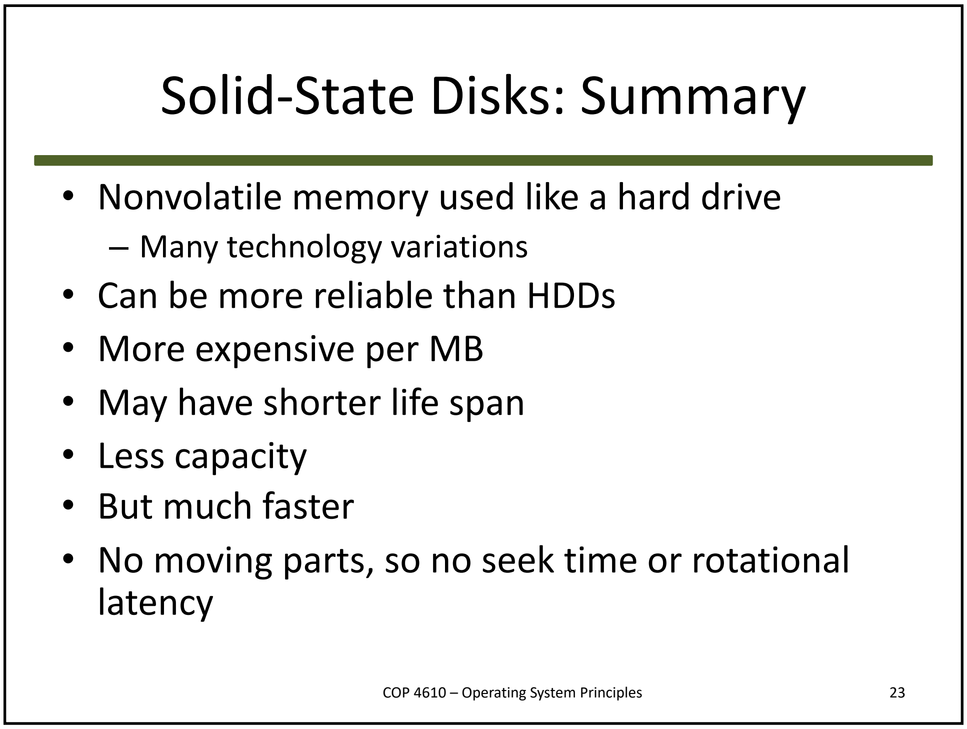
When using Wear Leveling Technology:

The same block will be written repeatedly

Occurrence of Bad Blocks

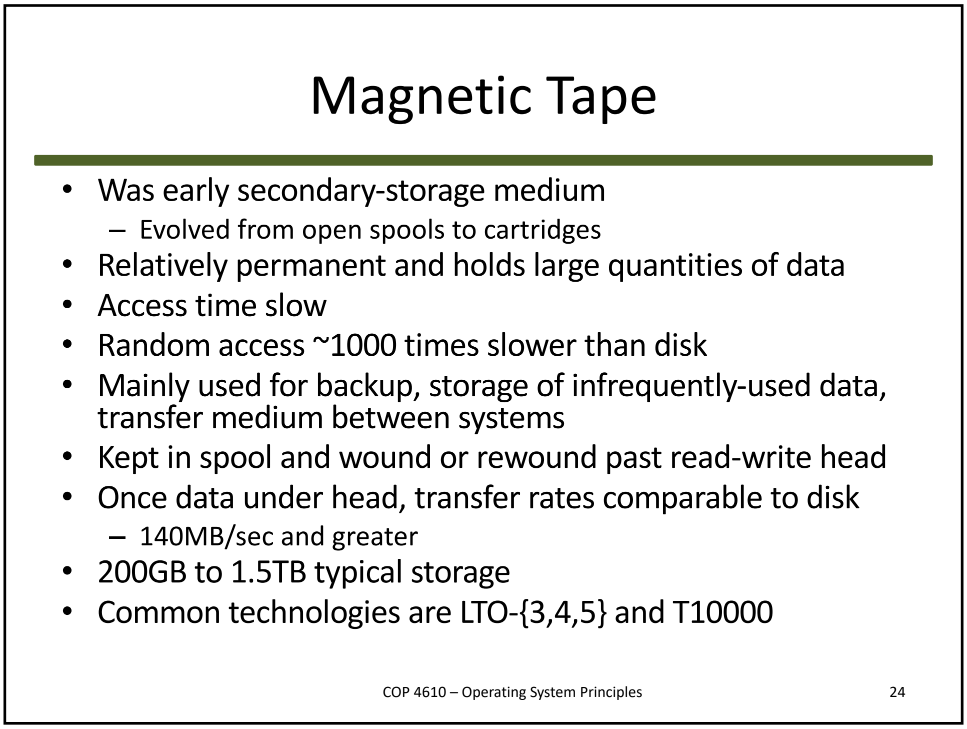
Writes data into different blocks individually

Writes data into different blocks individually



Solid-State Disks: Summary

- Nonvolatile memory used like a hard drive
  - Many technology variations
- Can be more reliable than HDDs
- More expensive per MB
- May have shorter life span
- Less capacity
- But much faster
- No moving parts, so no seek time or rotational latency



Magnetic Tape

- Was early secondary-storage medium
  - Evolved from open spools to cartridges
- Relatively permanent and holds large quantities of data
- Access time slow
- Random access ~1000 times slower than disk
- Mainly used for backup, storage of infrequently-used data, transfer medium between systems
- Kept in spool and wound or rewound past read-write head
- Once data under head, transfer rates comparable to disk
  - 140MB/sec and greater
- 200GB to 1.5TB typical storage
- Common technologies are LTO-{3,4,5} and T10000

Magnetic Tapes

MAN056 [RF] © www.visualphotos.com

Disk Scheduling

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

Disk Scheduling (Cont.)

- Several algorithms exist to schedule the servicing of disk I/O requests
- The analysis is true for one or many platters
- We illustrate scheduling algorithms with a **request queue** (0-199)

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

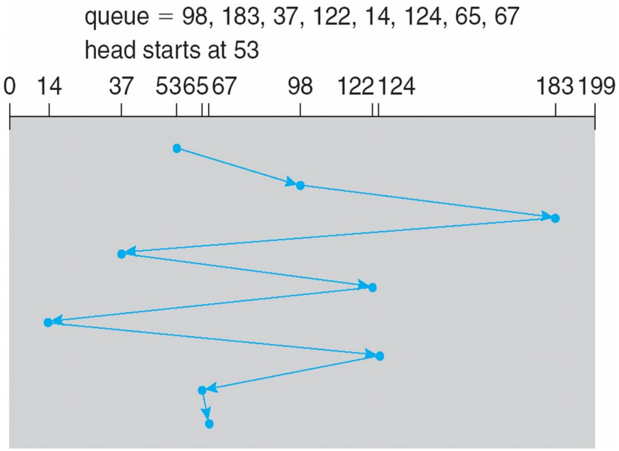
Head pointer 53

Disk Scheduling (Cont.)

- Idle disk can immediately work on I/O request
- Busy disk means work must queue
  - Optimization algorithms only make sense when a queue exists
- Note that drive controllers have small buffers and can manage a queue of I/O requests

FCFS

Illustration shows total head movement of 640 cylinders



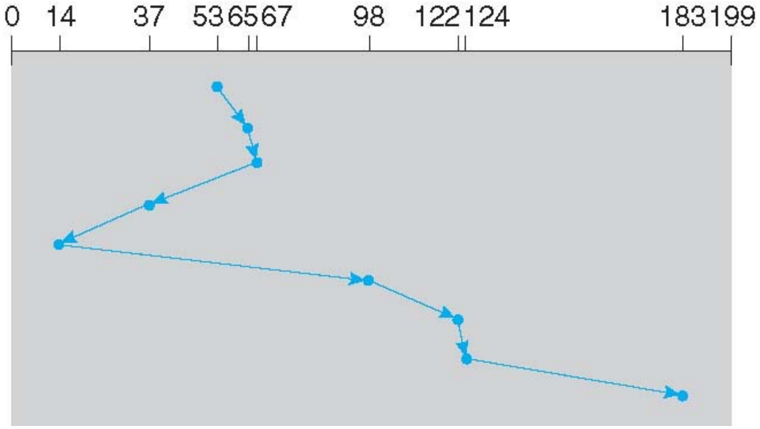
SSTF

- **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
- Illustration shows total head movement of 236 cylinders

SSTF (Cont.)

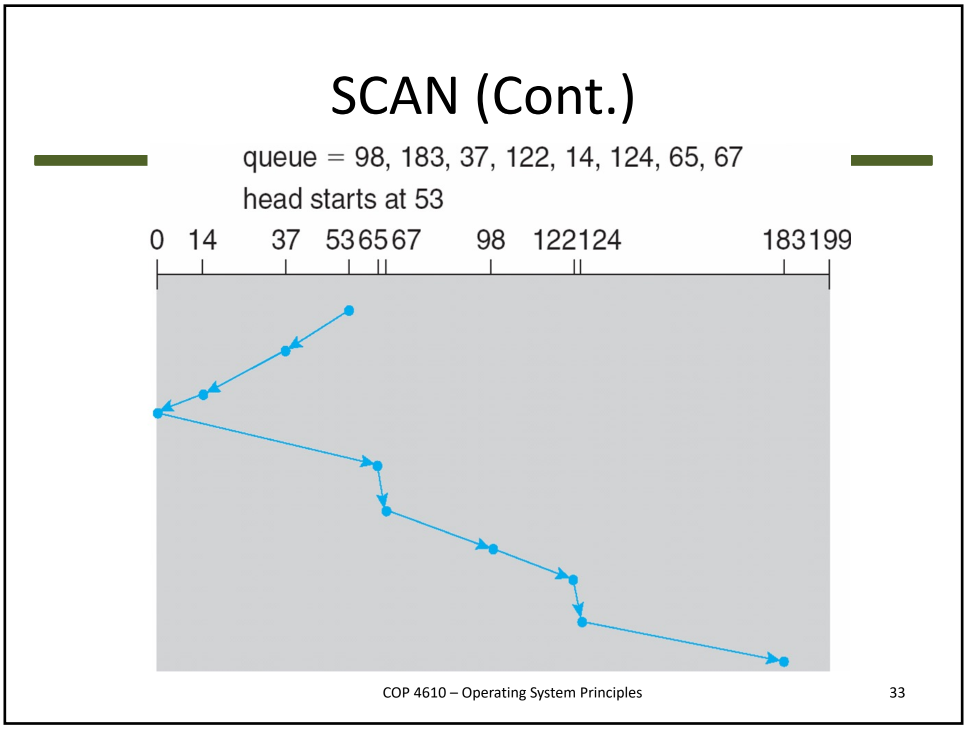
queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53



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**
- Illustration shows total head movement of 208 cylinders

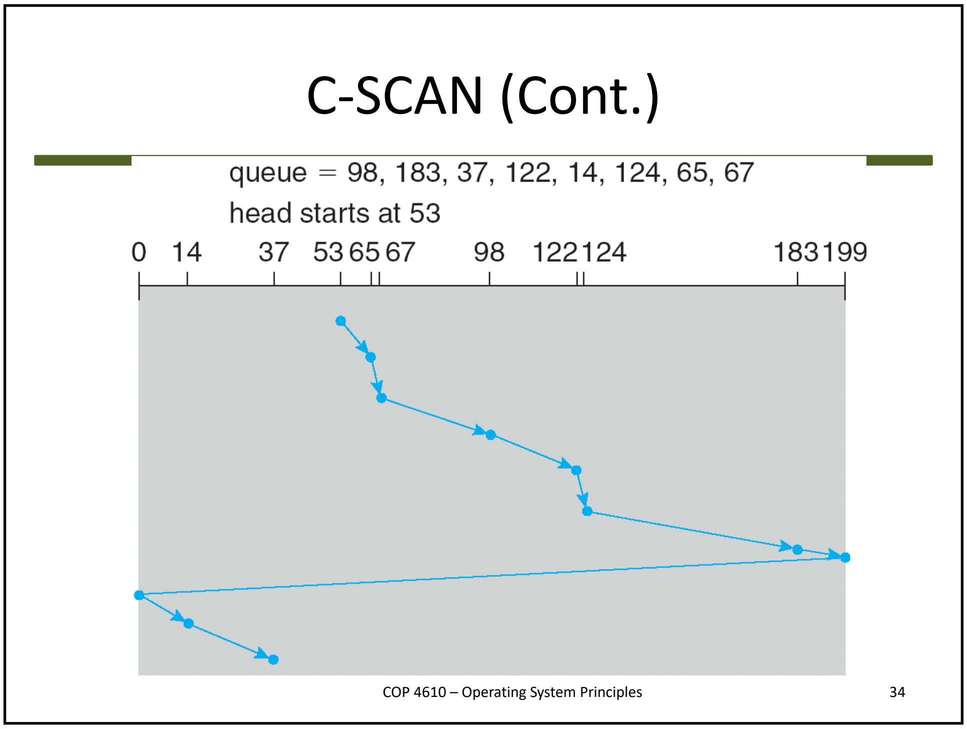


SCAN (Cont.)

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53

COP 4610 – Operating System Principles 33

33

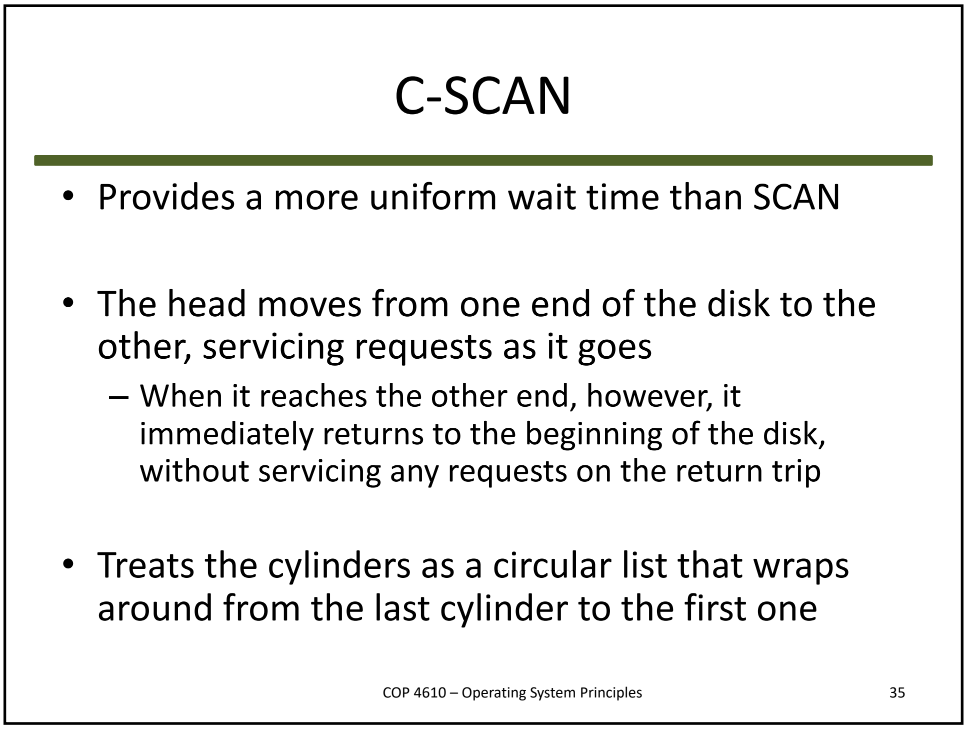


C-SCAN (Cont.)

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53

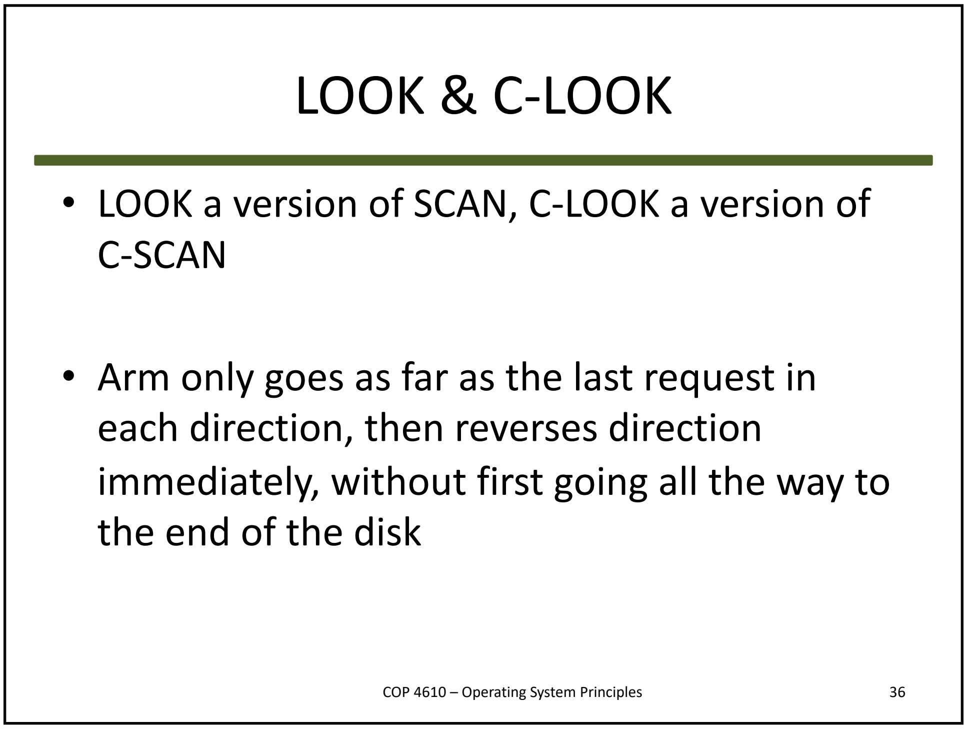
COP 4610 – Operating System Principles 34

34



C-SCAN

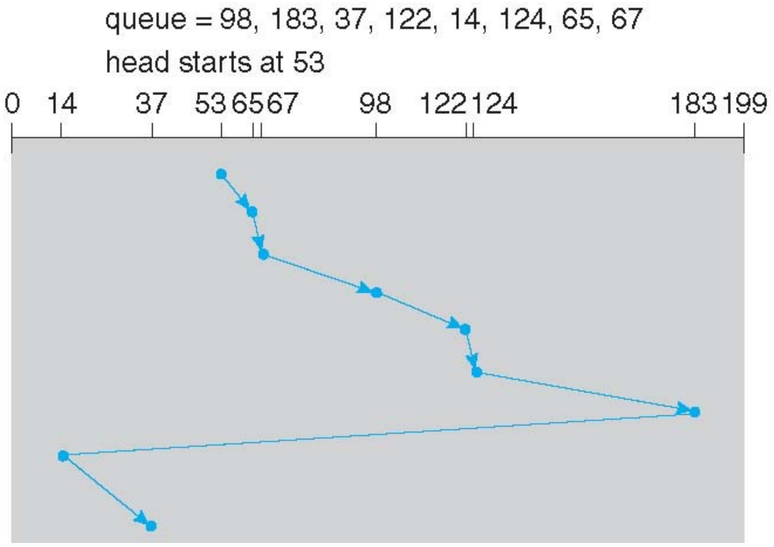
- Provides a more uniform wait time than SCAN
- The head moves from one end of the disk to the other, servicing requests as it goes
  - When it reaches the other end, however, it immediately returns to the beginning of the disk, without servicing any requests on the return trip
- Treats the cylinders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



LOOK & C-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

C-LOOK (Cont.)



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

head starts at 53

Selecting a Disk-Scheduling Algorithm

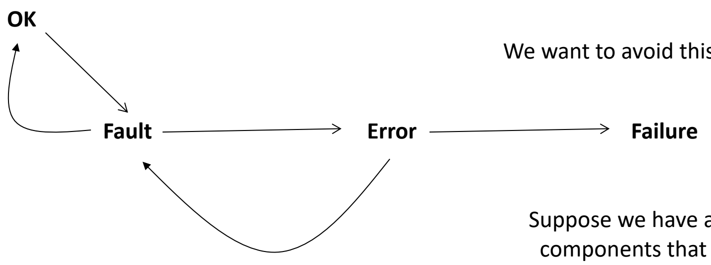
- **SSTF** is common and has a natural appeal for low load disks (quickly go to next request)
- LOOK, C-LOOK, SCAN and C-SCAN perform better for systems that place a heavy load on the disk (no starvation, more predictable delays)
- Performance depends on the number and types of requests
- Requests for disk service can be influenced by the file-allocation method

RAID Structure

- RAID – redundant array of inexpensive/independent disks
  - Using multiple disk drives provides reliability via **redundancy**

- Increases the **mean time to failure**
- **Mean time to repair –** exposure time when another failure could cause data loss
- **Mean time to data loss** based on above factors

Fault Tolerance



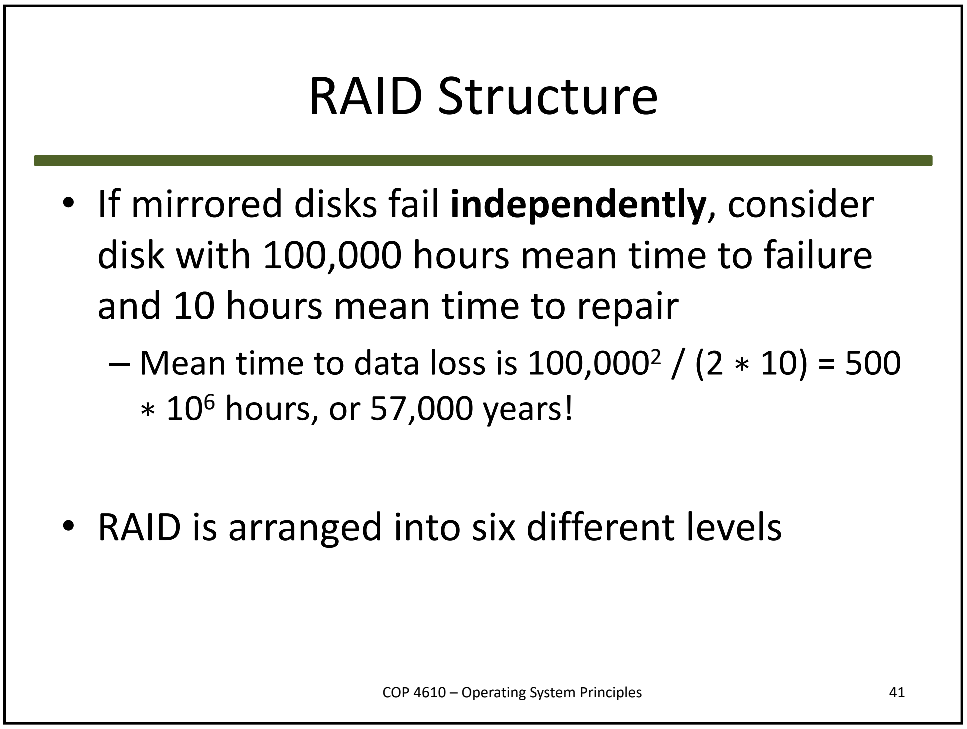
We want to avoid this

Suppose we have a system with 2 components that all must work

$P_{success}^2$

Complexity is the bane of fault tolerance

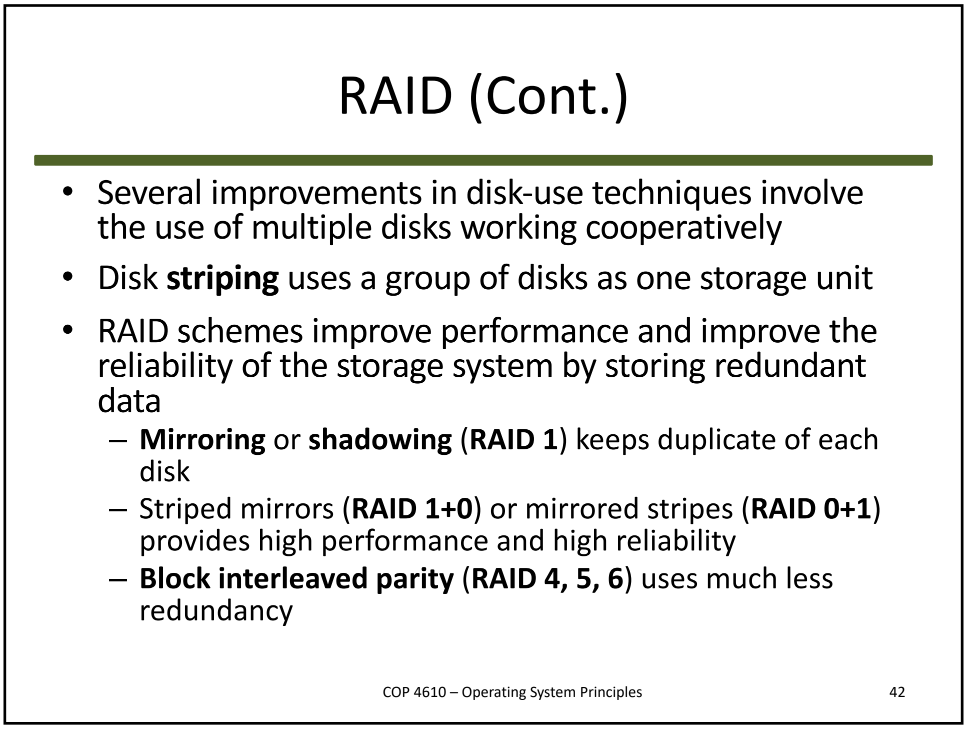
90% components → 81%



RAID Structure

- If mirrored disks fail **independently**, consider disk with 100,000 hours mean time to failure and 10 hours mean time to repair
  - Mean time to data loss is $100{,}000^2 / (2 * 10) = 500 * 10^6$ hours, or 57,000 years!

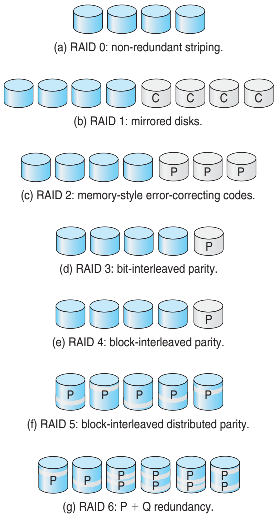
- RAID is arranged into six different levels



RAID (Cont.)

- Several improvements in disk-use techniques involve the use of multiple disks working cooperatively
- Disk **striping** uses a group of disks as one storage unit
- RAID schemes improve performance and improve the reliability of the storage system by storing redundant data
  - **Mirroring** or **shadowing** (**RAID 1**) keeps duplicate of each disk
  - Striped mirrors (**RAID 1+0**) or mirrored stripes (**RAID 0+1**) provides high performance and high reliability
  - **Block interleaved parity** (**RAID 4, 5, 6**) uses much less redundancy

RAID Levels



(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.

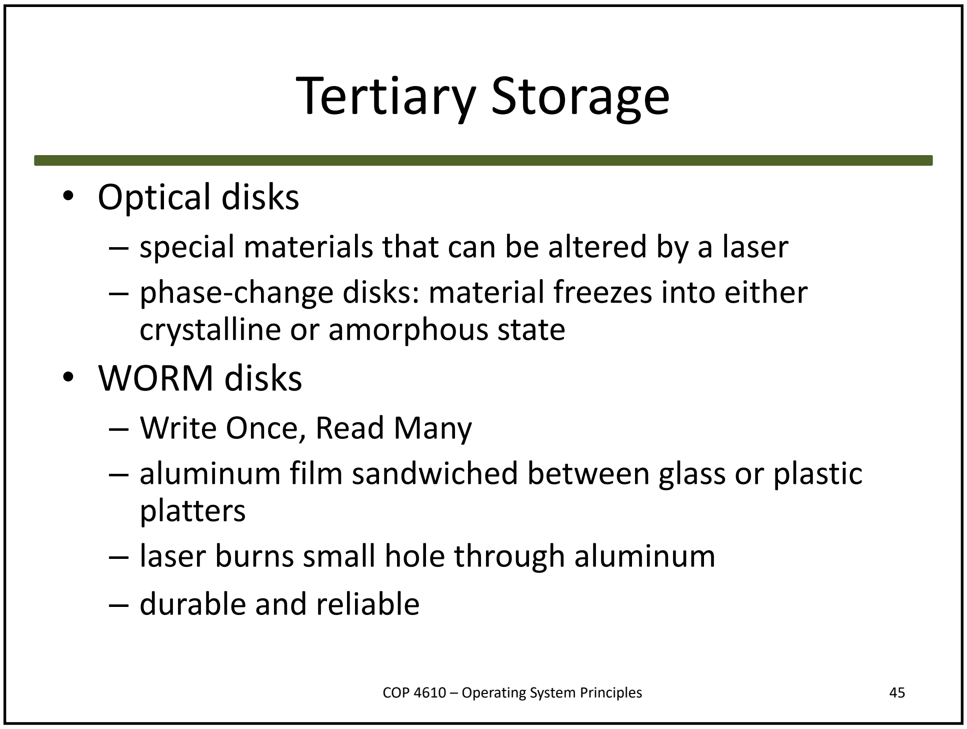
(e) RAID 4: block-interleaved parity.

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

(g) RAID 6: P + Q redundancy.

Tertiary Storage

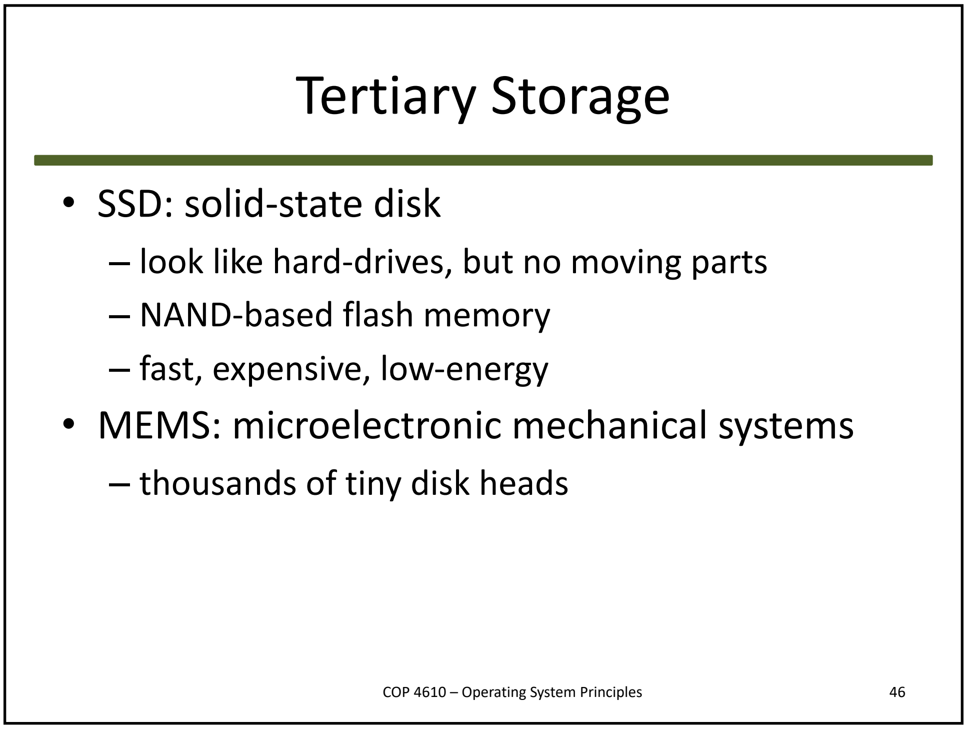
- Low cost
- “Removable media”
- Floppy disks
  - – thin flexible disk coated with magnetic material, enclosed in protective plastic case
- Magneto-optic disks
  - – rigid platter coated with magnetic material
  - – laser heat used to amplify a large/weak magnetic field to record a bit; also used to read data (Kerr effect)
  - – very resistant to head crashes



Tertiary Storage

- Optical disks
  - special materials that can be altered by a laser
  - phase-change disks: material freezes into either crystalline or amorphous state
- WORM disks
  - Write Once, Read Many
  - aluminum film sandwiched between glass or plastic platters
  - laser burns small hole through aluminum
  - durable and reliable

45



Tertiary Storage

- SSD: solid-state disk
  - look like hard-drives, but no moving parts
  - NAND-based flash memory
  - fast, expensive, low-energy
- MEMS: microelectronic mechanical systems
  - thousands of tiny disk heads

46