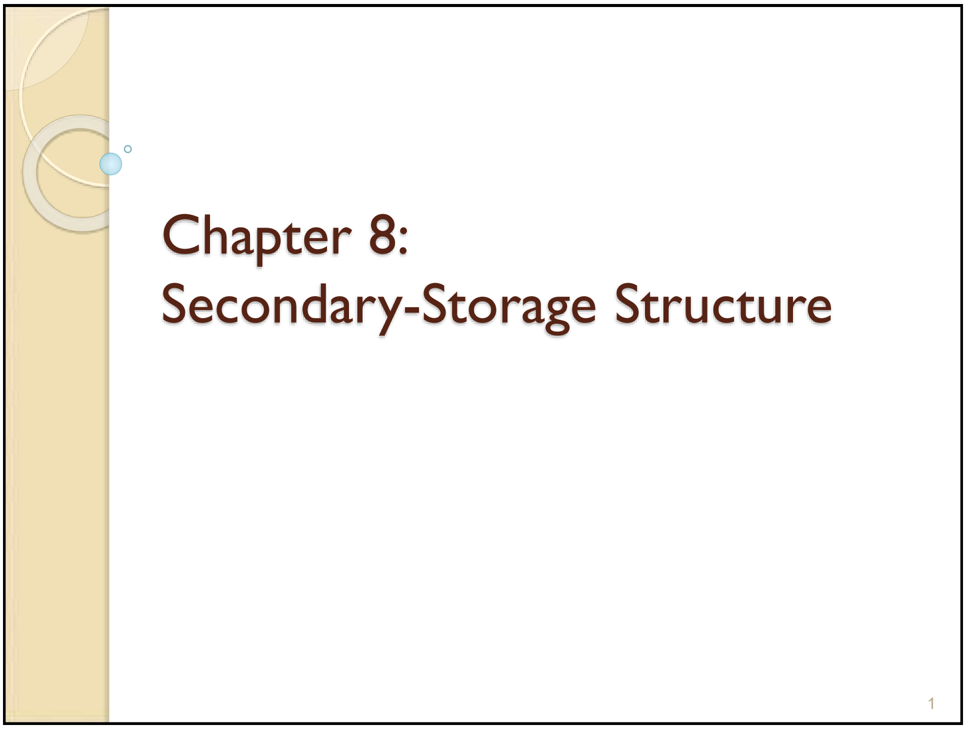
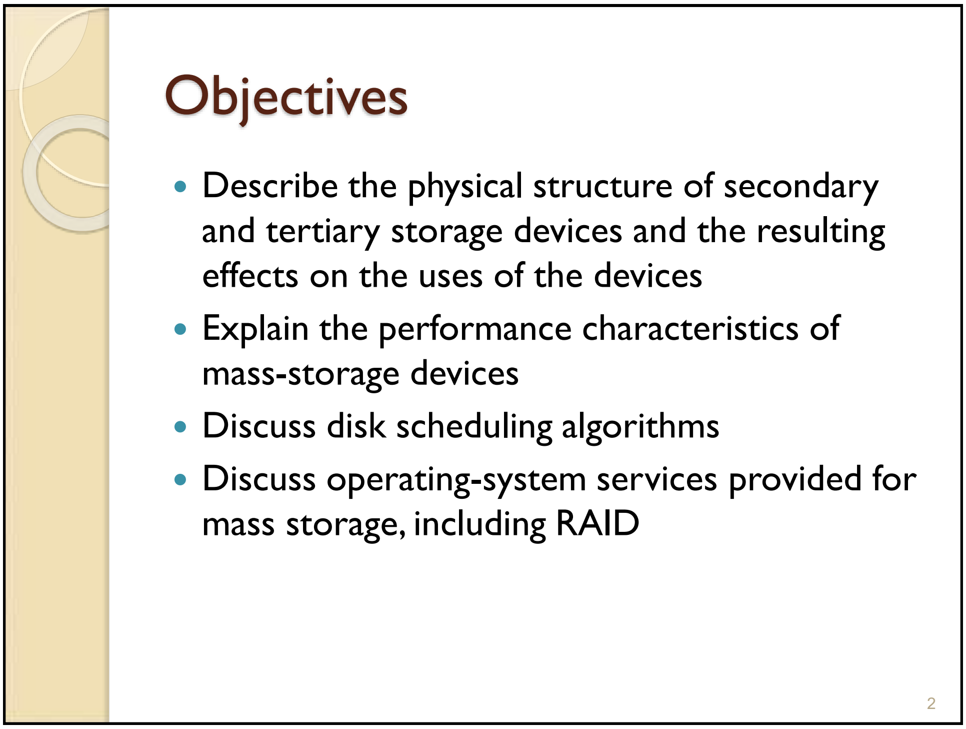


Chapter 8: Secondary-Storage Structure

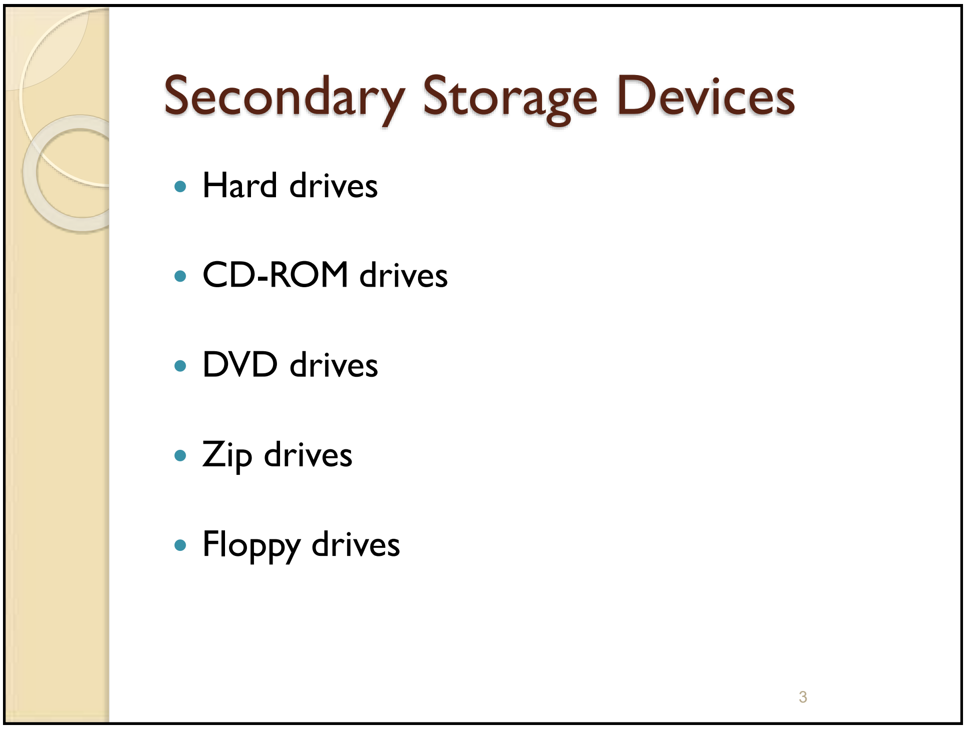
1



Objectives

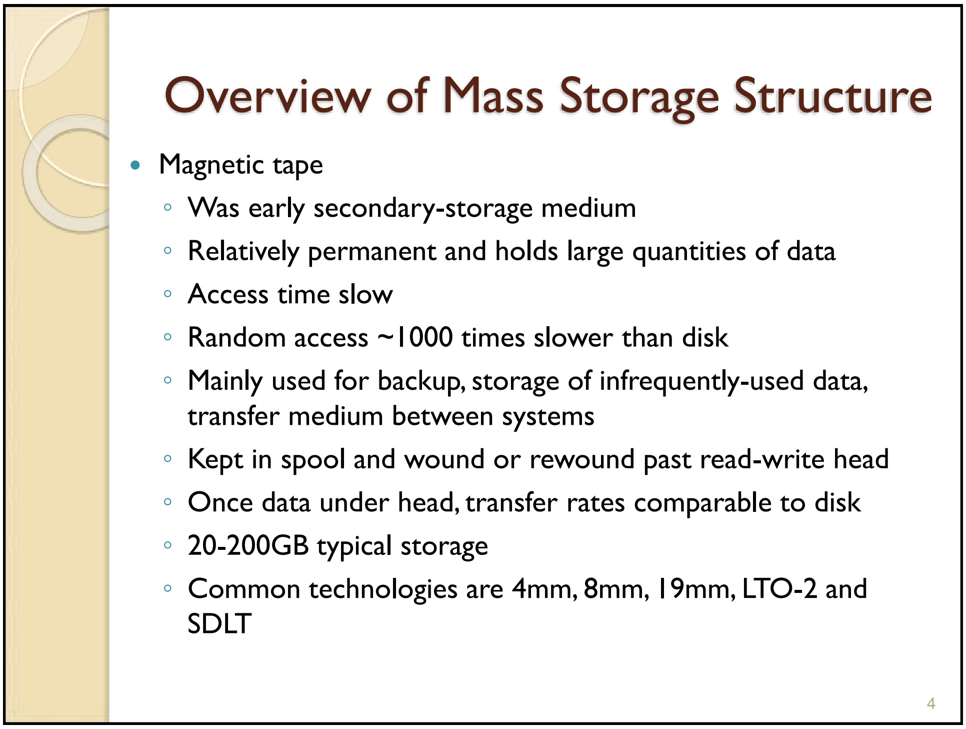
- Describe the physical structure of secondary and tertiary storage devices and the resulting effects on the uses of the devices
- Explain the performance characteristics of mass-storage devices
- Discuss disk scheduling algorithms
- Discuss operating-system services provided for mass storage, including RAID

2



Secondary Storage Devices

- Hard drives
- CD-ROM drives
- DVD drives
- Zip drives
- Floppy drives



Overview of Mass Storage Structure

- Magnetic tape
  - Was early secondary-storage medium
  - 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
  - 20-200GB typical storage
  - Common technologies are 4mm, 8mm, 19mm, LTO-2 and SDLT

Overview of Mass Storage Structure

- **Magnetic disks -** provide bulk of secondary storage of modern computers
  - Drives rotate at 60 to 200 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 -- That's bad
- Disks can be removable
- Drive attached to computer via **I/O bus**
  - Busses vary, including **EIDE, ATA, SATA, USB, Fibre Channel, SCSI**
  - **Host controller** in computer uses bus to talk to **disk controller** built into drive or storage array

Hard Drive

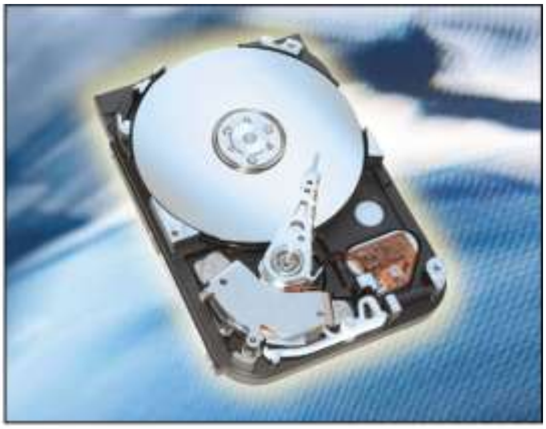


Figure 1-16 Hard drive with sealed cover removed

EIDE Technology

- Used by most hard drives, CD-ROM drives, and DVD drives
- Can accommodate up to four EIDE devices on one system

IDE Connectors on a Motherboard

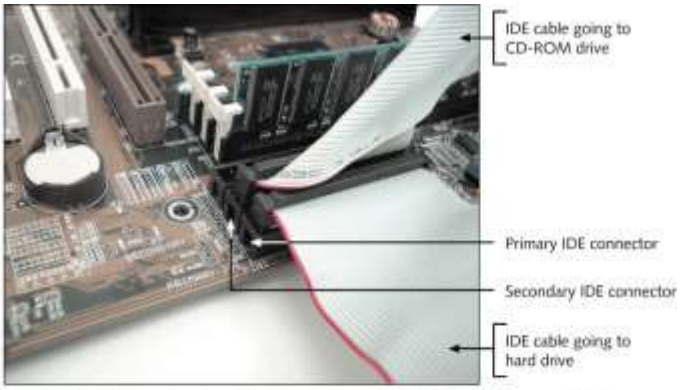

IDE cable going to CD-ROM drive
Primary IDE connector
Secondary IDE connector
IDE cable going to hard drive

Figure 1-17 A motherboard usually has two IDE connectors, each of which can accommodate two devices; a hard drive usually connects to the motherboard using the primary IDE connector

IDE Connectors on a Motherboard (continued)

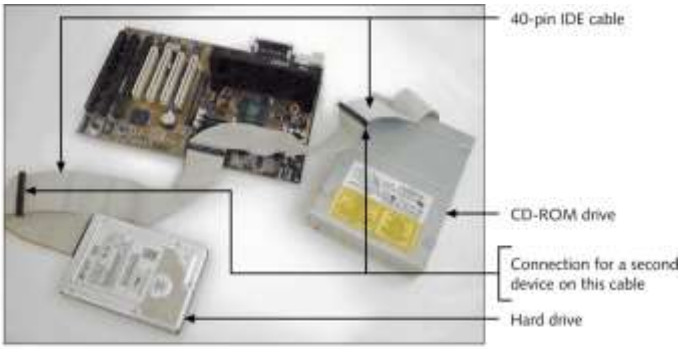


Figure 1-18 Two IDE devices connected to a motherboard using both IDE connections and two cables

9

9

IDE Connectors on a Motherboard

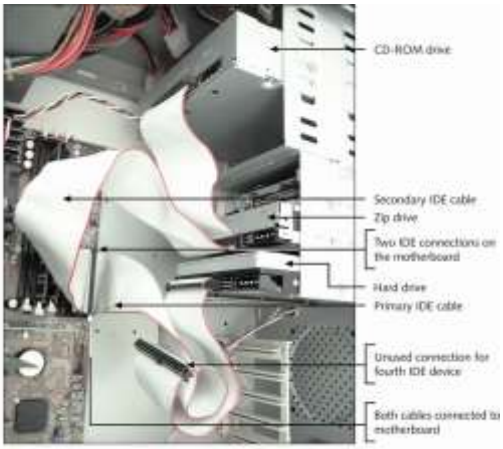
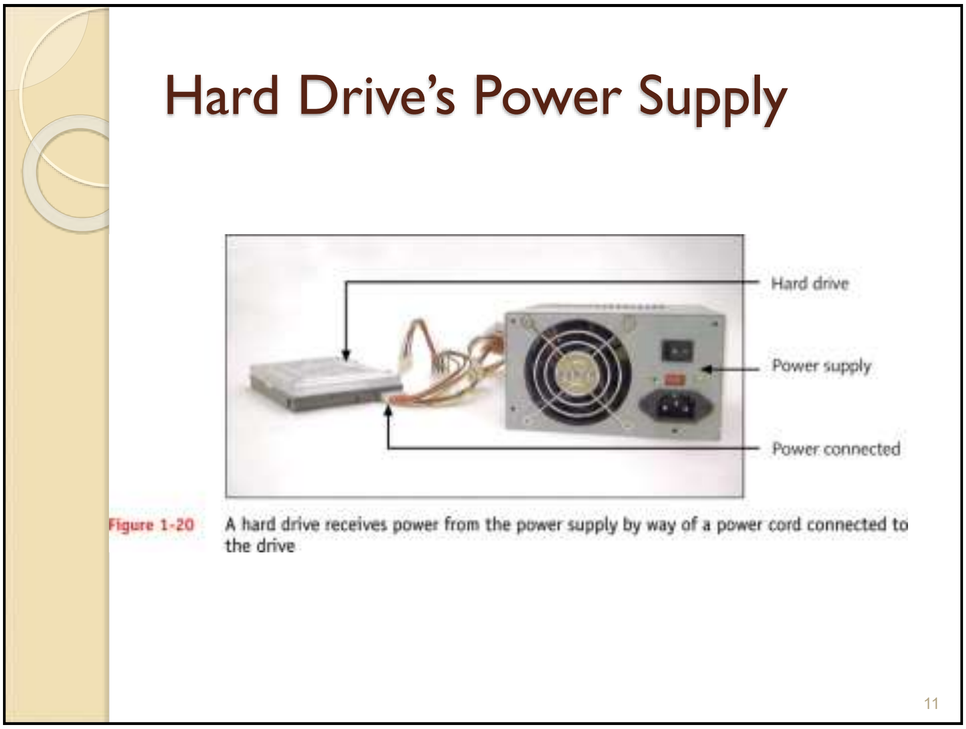

CD-ROM drive
Secondary IDE cable
Zip drive
Two IDE connections on the motherboard
Hard drive
Primary IDE cable
Unused connection for fourth IDE device
Both cables connected to motherboard

Figure 1-19 This system has a CD-ROM and a Zip drive sharing the secondary IDE cable and a hard drive using the primary IDE cable

10

10

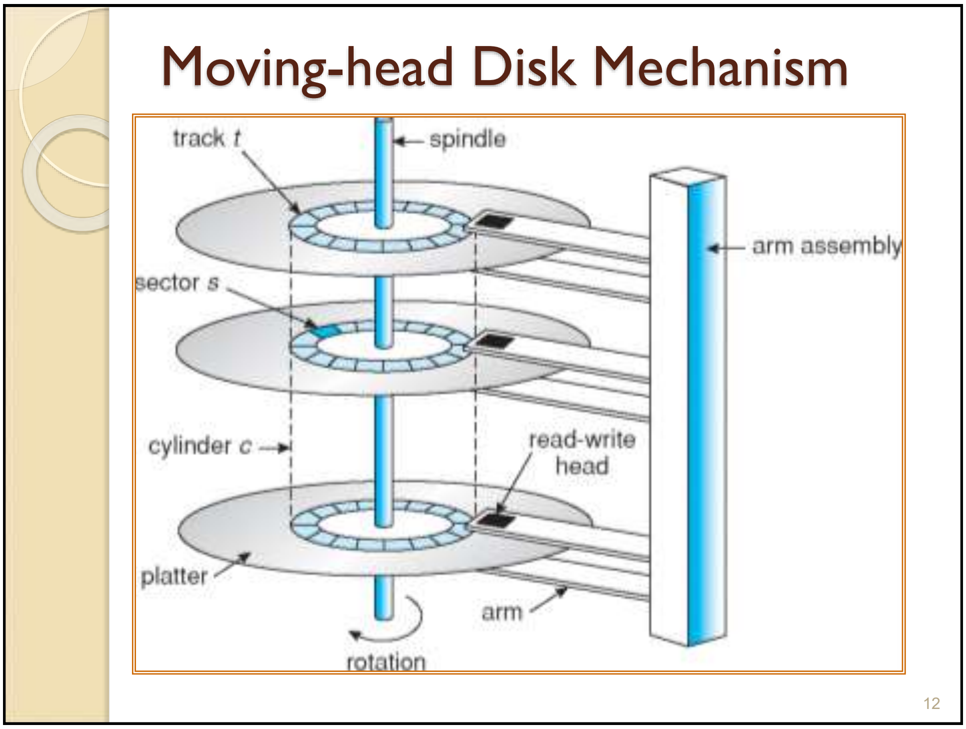


Hard Drive's Power Supply

Figure 1-20 A hard drive receives power from the power supply by way of a power cord connected to the drive

11

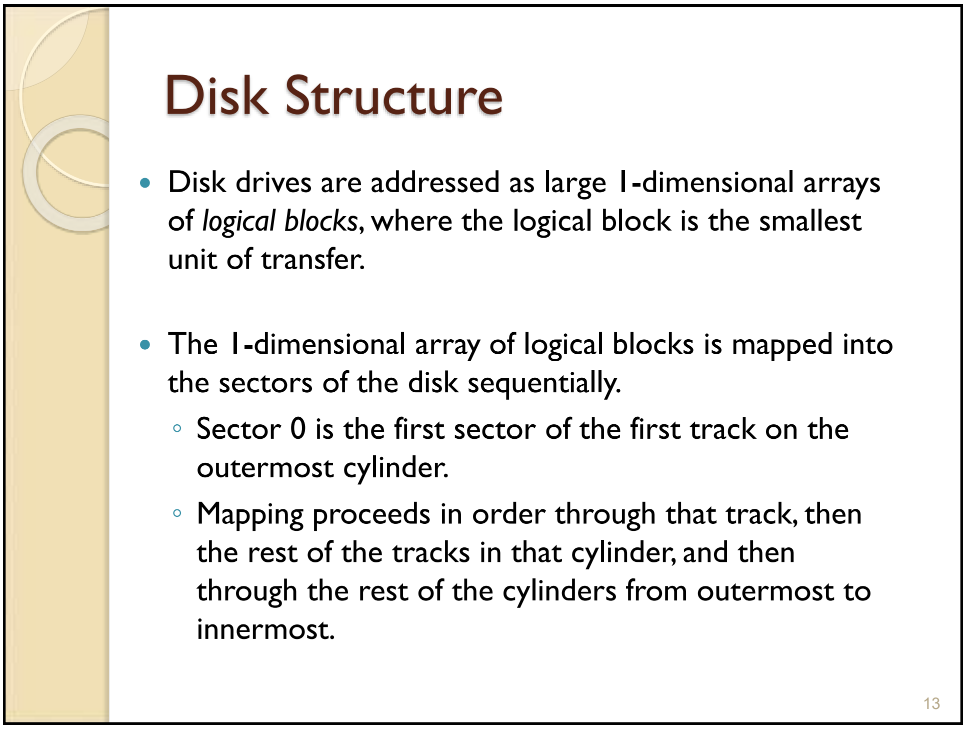
11



Moving-head Disk Mechanism

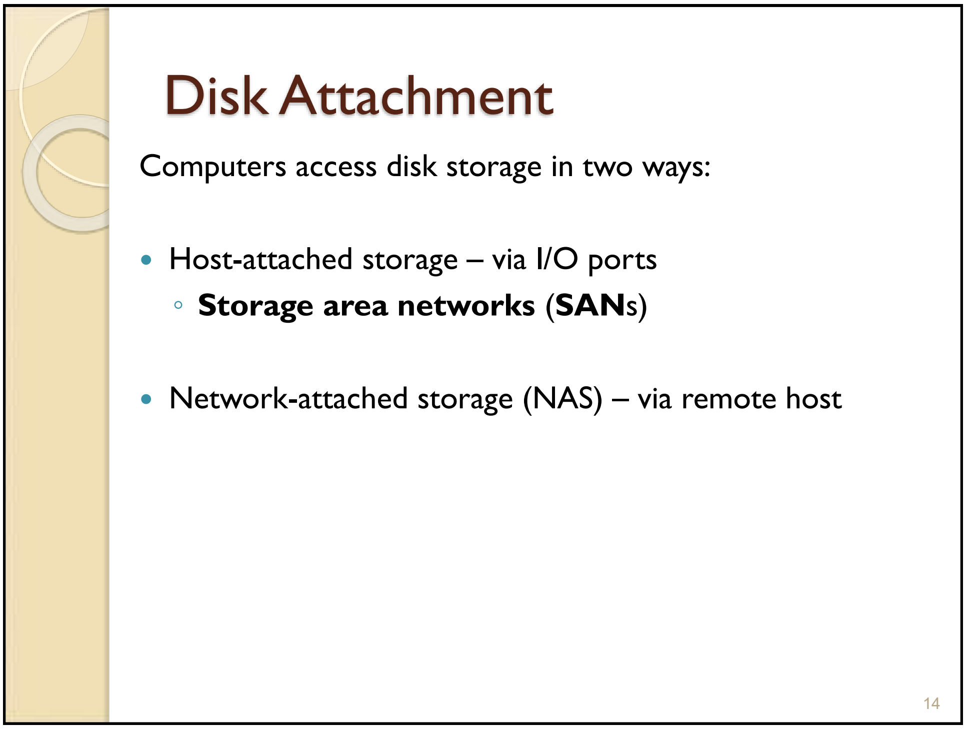
12

12



Disk Structure

- Disk drives are addressed as large 1-dimensional arrays of *logical blocks*, where the logical block is the smallest unit of transfer.
- 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.

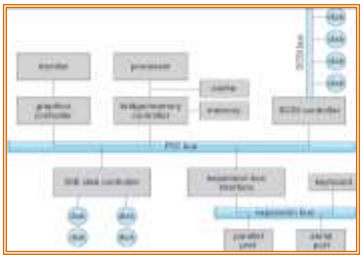


Disk Attachment

Computers access disk storage in two ways:

- Host-attached storage – via I/O ports
  - **Storage area networks** (**SAN**s)
- Network-attached storage (NAS) – via remote host

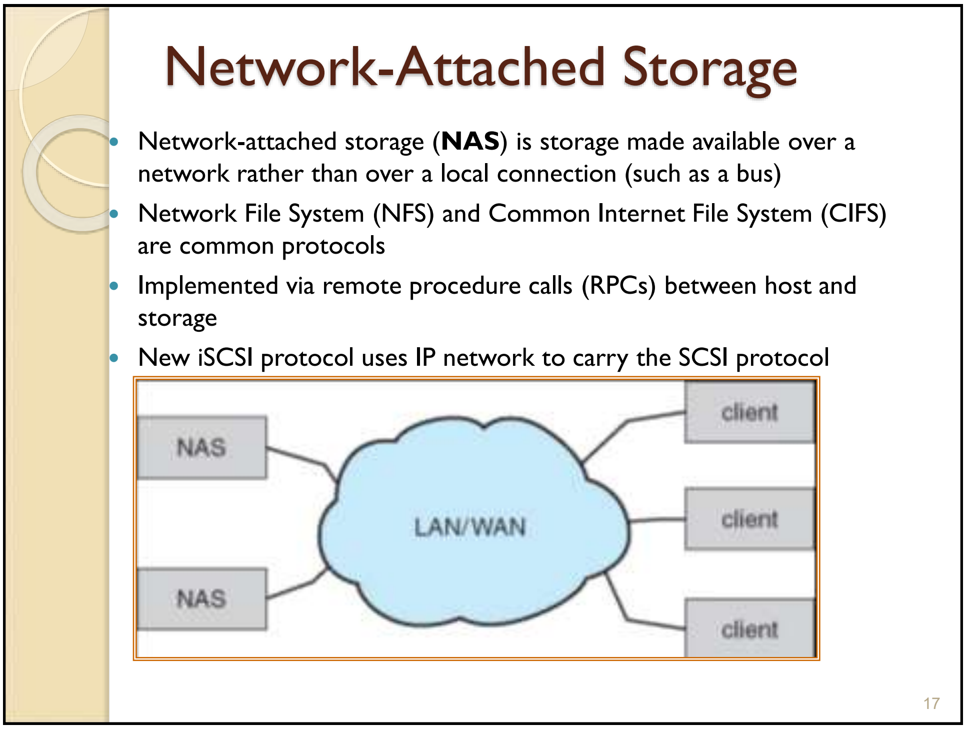
Host-Attached Storage



- Host-attached storage accessed through local I/O ports talking to I/O buses
  - Common architecture – IDE, ATA, SATA (newer, with simplified cabling)
- SCSI itself is a bus, up to 16 devices on one cable, **SCSI initiator** (controller card) requests operation and **SCSI targets** perform tasks
  - SCSI targets can be up to max 15 storage devices
  - Each target can have up to 8 **logical units** (disks attached to device controller

Host-Attached Storage

- FC (Fibre Channel) is high-speed serial architecture
  - can operate over optical fiber or over 4-conductor copper cable
  - Two variants:
    - Can be switched fabric with 24-bit address space – the basis of **storage area networks** (**SAN**s) in which many hosts attach to many storage units
    - Can be **arbitrated loop** (**FC-AL**) of 126 devices

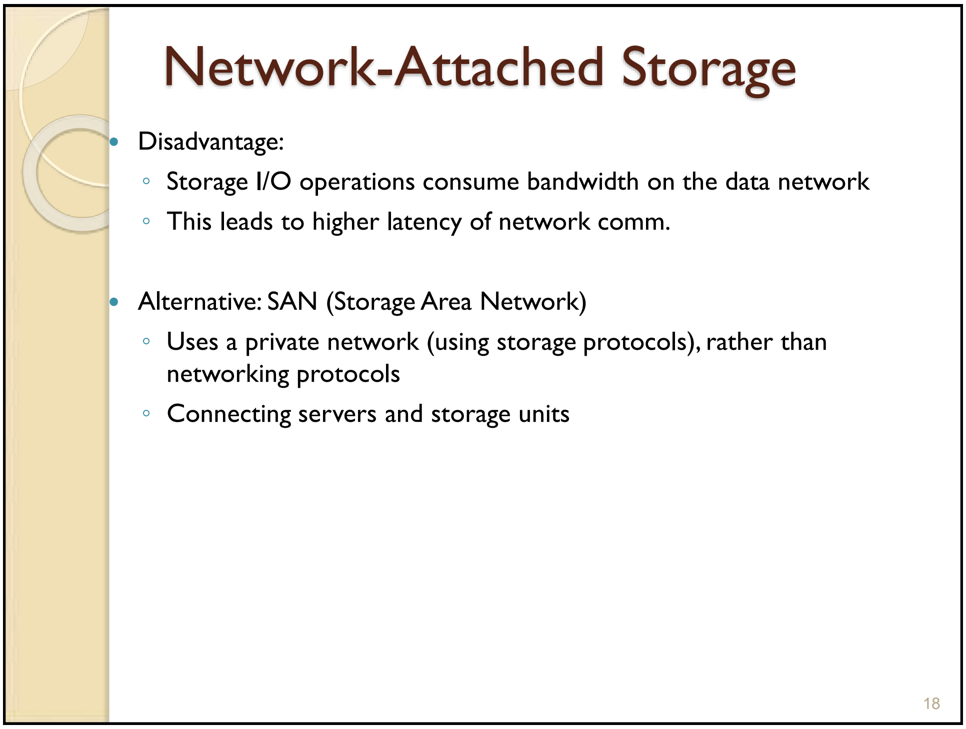


Network-Attached Storage

- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
- Network File System (NFS) and Common Internet File System (CIFS) are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage
- New iSCSI protocol uses IP network to carry the SCSI protocol

NAS
NAS
LAN/WAN
client
client
client

17



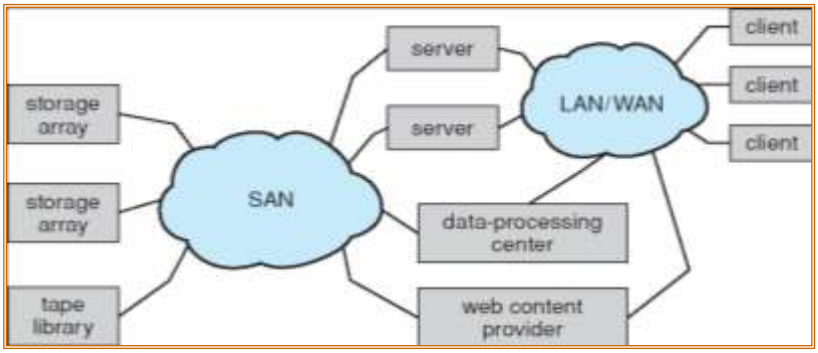
Network-Attached Storage

- Disadvantage:
  - Storage I/O operations consume bandwidth on the data network
  - This leads to higher latency of network comm.
- Alternative: SAN (Storage Area Network)
  - Uses a private network (using storage protocols), rather than networking protocols
  - Connecting servers and storage units

18

Storage Area Network

- Common in large storage environments (and becoming more common)
- Multiple hosts attached to multiple storage arrays – flexible
- Storage can be dynamically allocated to hosts. If storage for 1st host is running low, SAN can allocate more storage to that host.



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.
- Access time has two major components
  - ***Seek time*** is the time for the disk arm to move the heads to the cylinder containing the desired sector.
  - ***Rotational latency*** is the additional time waiting for the disk to rotate the desired sector to the disk head.
- 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.

Timing of a Disk I/O Transfer

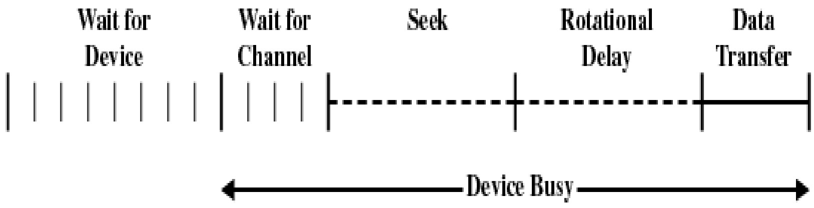


Figure 11.6 Timing of a Disk I/O Transfer

Disk Scheduling

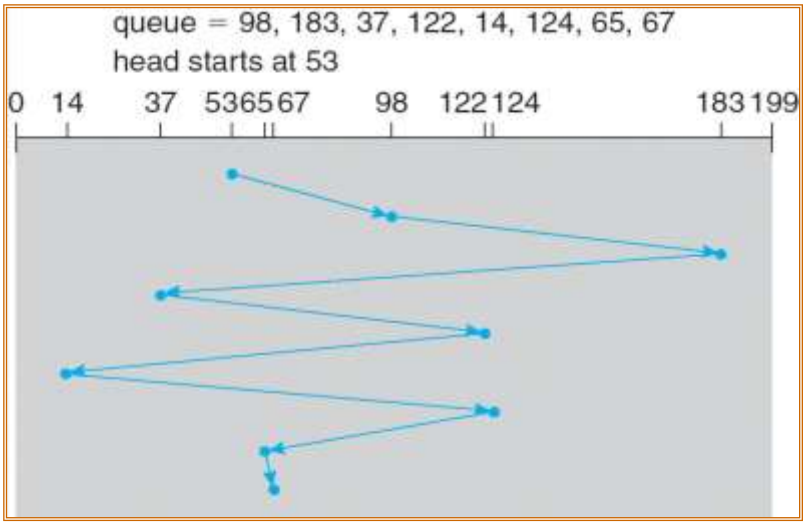
- Several algorithms exist to schedule the servicing of disk I/O requests.
- We illustrate them with a request queue (0-199).

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

Head pointer 53

FCFS / FIFO

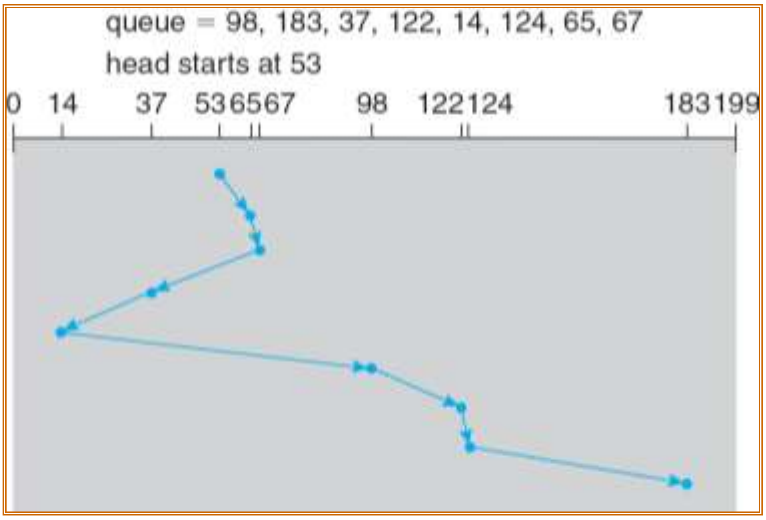
Illustration shows total head movement of 640 cylinders.



Shortest Seek Time First (SSTF)

- 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.)



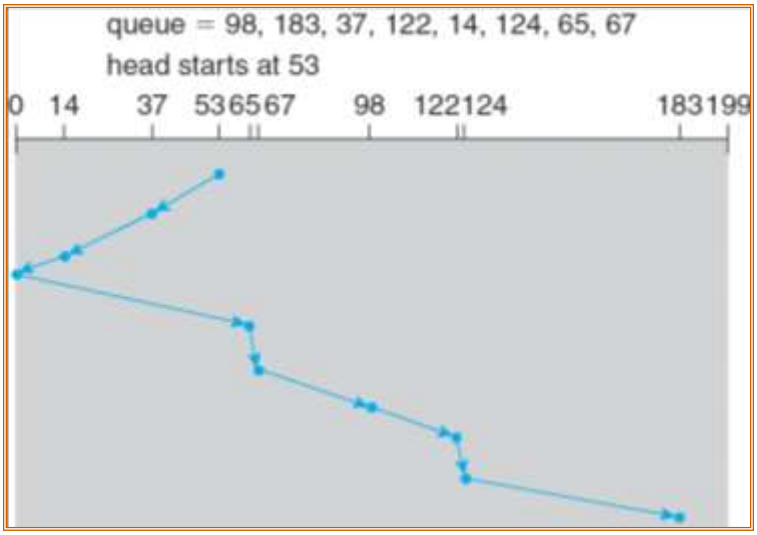
queue = 98, 183, 37, 122, 14, 124, 65, 67

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.
- Sometimes called the *elevator algorithm*.
- Illustration shows total head movement of 208 cylinders.

SCAN (Cont.)

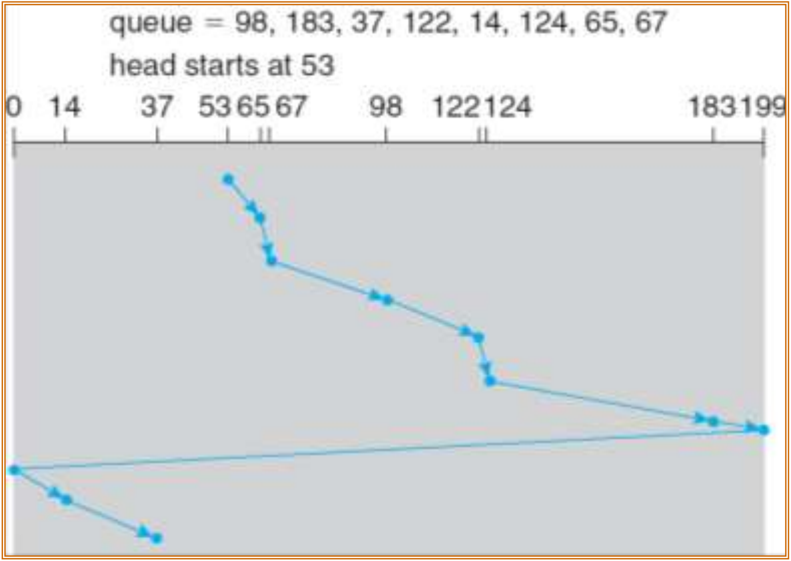

queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53

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.

C-SCAN (Cont.)



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

head starts at 53

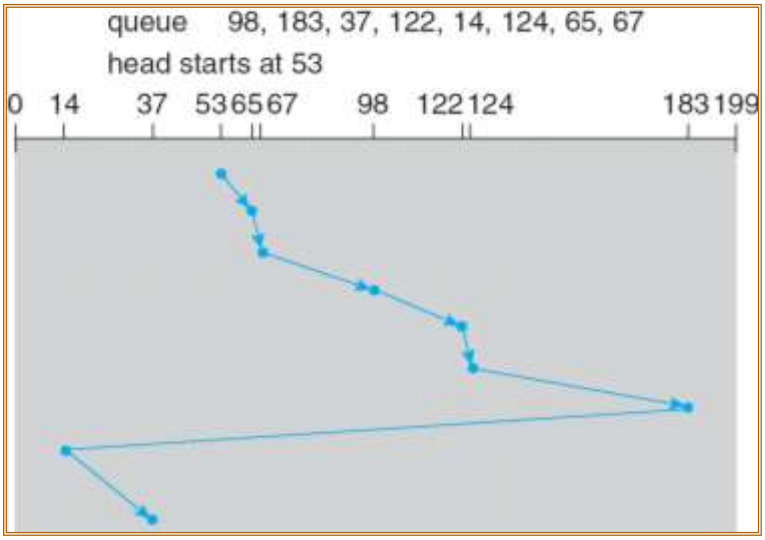
C-LOOK

- 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.

LOOK

- Version of SCAN
- Arm only goes as far as the last request in each direction, then the head movement is reversed and servicing continues, without first going all the way to the end of the disk.

C-LOOK (Cont.)

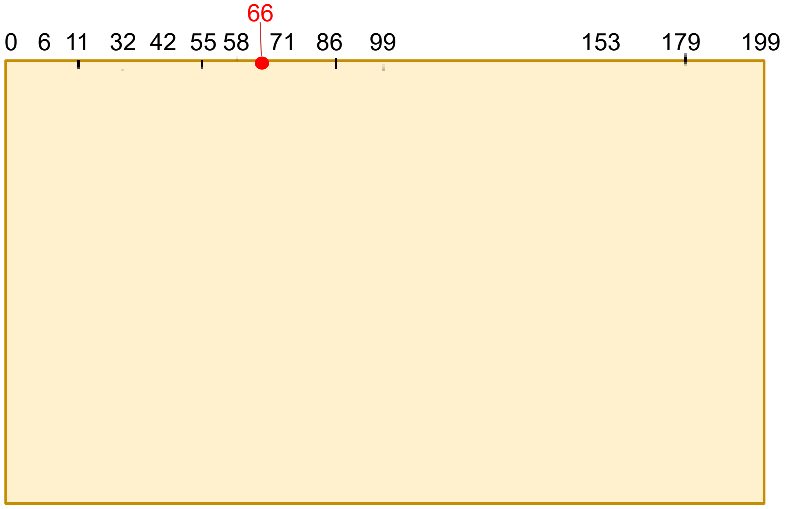


Example

- Disk with 200 cylinders (0-199)
- Request queue :

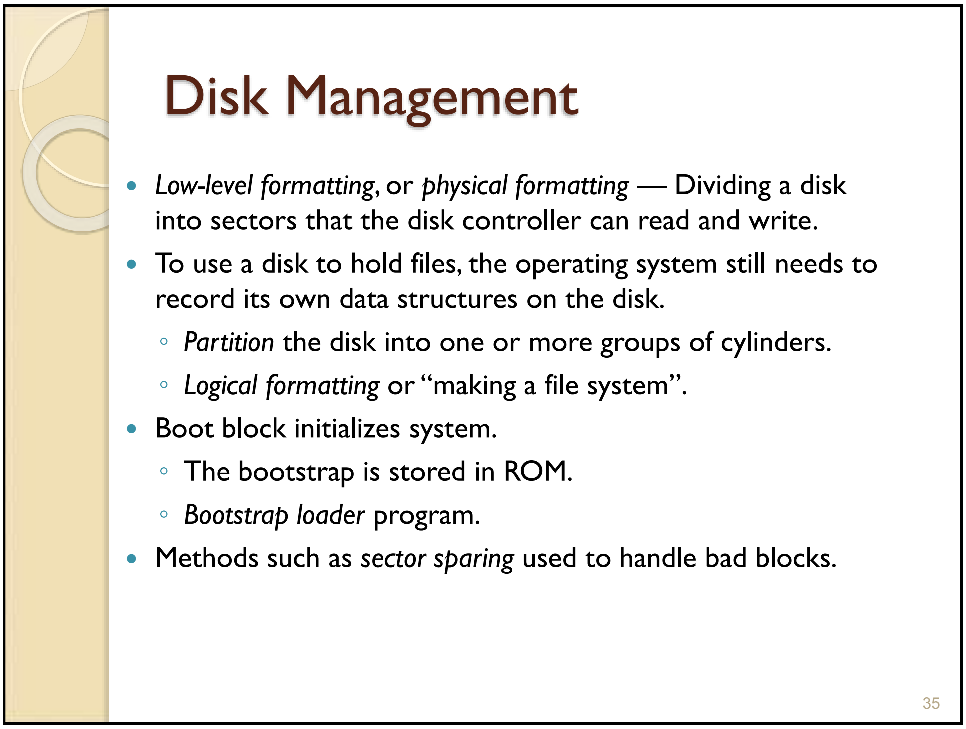
  55, 32, 6, 99, 58, 71, 86, 153, 11, 179, 42
- Disk head is currently at 66.
- Previous request serviced was 48.
- Show the track serviced using each of the following disk scheduling algorithms:
  - FCFS
  - SSTF
  - SCAN
  - ▯ C-SCAN
  - ▯ LOOK
  - ▯ C-LOOK
- What is the total head movement for each scheduling policy?

Request queue : 55, 32, 6, 99, 58, 71, 86, 153, 11, 179, 42
Disk head is currently at 66.
Previous request serviced was 48.



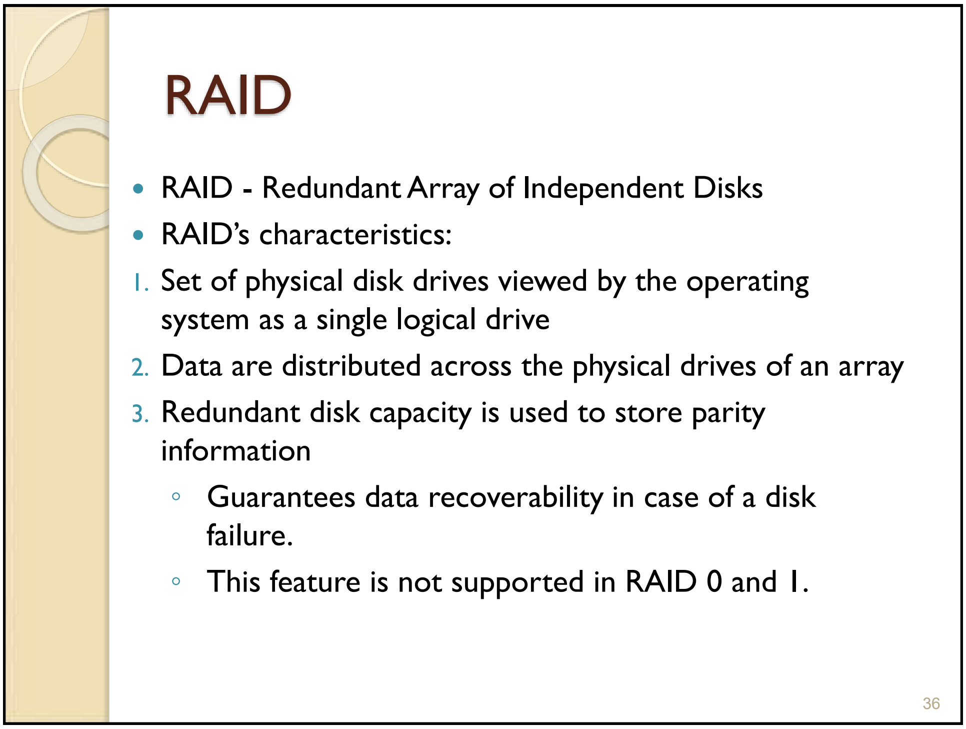
Selecting a Disk-Scheduling Algorithm

- SSTF is common and has a natural appeal.
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk.
- Performance depends on the number and types of requests.
- Requests for disk service can be influenced by the file-allocation method.
- The disk-scheduling algorithm should be written as a separate module of the operating system, allowing it to be replaced with a different algorithm if necessary.
- Either SSTF or LOOK is a reasonable choice for the default algorithm.



Disk Management

- *Low-level formatting*, or *physical formatting* — Dividing a disk into sectors that the disk controller can read and write.
- To use a disk to hold files, the operating system still needs to record its own data structures on the disk.
  - *Partition* the disk into one or more groups of cylinders.
  - *Logical formatting* or "making a file system".
- Boot block initializes system.
  - The bootstrap is stored in ROM.
  - *Bootstrap loader* program.
- Methods such as *sector sparing* used to handle bad blocks.



RAID

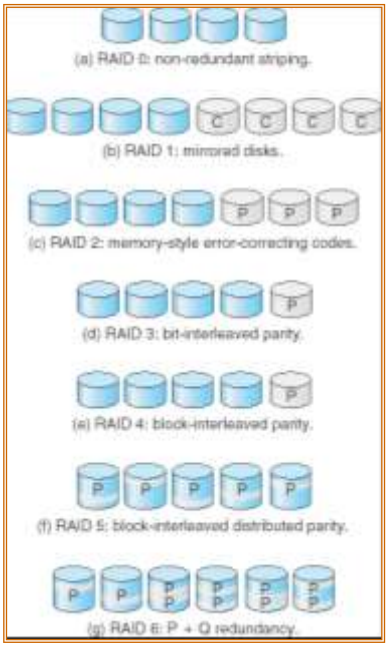
- RAID - Redundant Array of Independent Disks
- RAID's characteristics:

1. Set of physical disk drives viewed by the operating system as a single logical drive
2. Data are distributed across the physical drives of an array
3. Redundant disk capacity is used to store parity information
   - Guarantees data recoverability in case of a disk failure.
   - This feature is not supported in RAID 0 and 1.

RAID Structure

- **RAID** – multiple disk drives provides **reliability** via **redundancy**.
- RAID is arranged into six different levels – does not imply hierarchical relationship.
- 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* keeps duplicate of each disk.
    - *Block interleaved parity* 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.

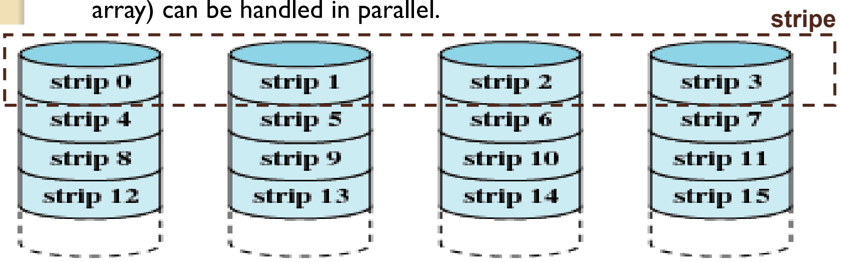
(e) RAID 4: block-interleaved parity.

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

(g) RAID 6: P + Q redundancy.

RAID 0

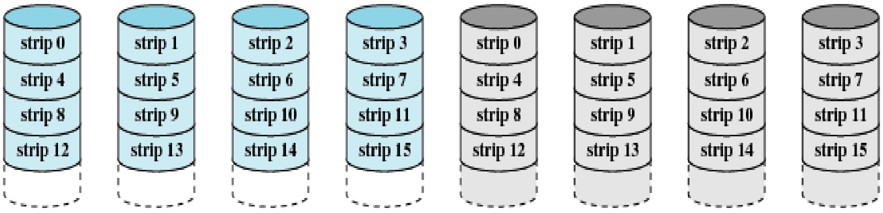
- Not a true member of the RAID family – not support redundancy.
- User and system data are distributed across all disks in the array.
- Advantage over using single large disk – if two I/O requests are pending for two different blocks of data, they may be on different disks → the two requests can be issued in parallel, reducing I/O queuing time.
- Advantage of strip layout – if a single I/O request consists of multiple logically contiguous strips, then up to *n* strips (in an *n*-disk array) can be handled in parallel.



(a) RAID 0 (non-redundant)

RAID 1

- Redundancy is achieved by simply duplicating the data.
- Each logical strip is mapped to two separate physical disks – each disk in the array has mirror disk that contains the same data.
- Advantages:
  - A read request can be serviced by either of the two disks, whichever one involves minimum access time.
  - A write request requires both corresponding strips to be updated, but can be done in parallel.
  - Recovery from failure is simple – from the mirror disk.

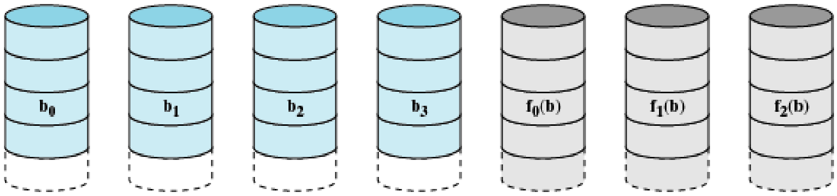


(b) RAID 1 (mirrored)

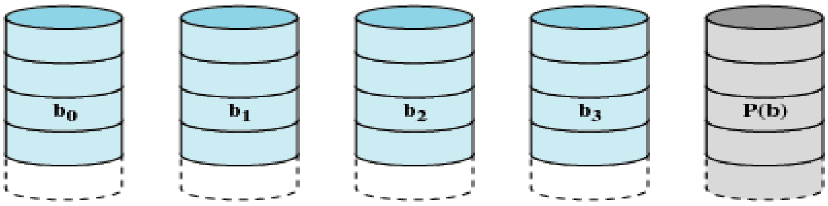
RAID 1 (cont.)

- Main disadvantage: Cost – it requires twice the disk space.
- Thus, this configuration is limited to drives that store system software and data, and other highly critical files.
  - Provides real-time backup, so that in disk failure all the critical data is still immediately available.
- Can achieve high I/O request rates if the bulk of the requests are reads – can double RAID 0 performance.
- But if many write requests, then no significant performance gain over RAID 0.

RAID 2 & 3 – Parallel access

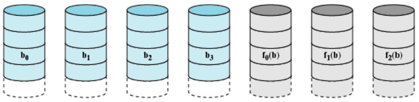


(c) RAID 2 (redundancy through Hamming code)



(d) RAID 3 (bit-interleaved parity)

RAID 2

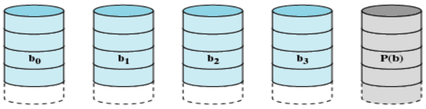


(c) RAID 2 (redundancy through Hamming code)

- RAID 2 and 3 – parallel access techniques
  - all member disks participate in the execution of every access I/O request.
  - The spindles of the individual drives are synchronized so that each disk head is in the same position on each disk at a given time.
- Data striping in RAID 2 & 3, the strips are very small (a single byte or word).
- Error-correcting code is calculated across corresponding bits on each disk, and the bits of the code are stored in the corresponding bit positions on multiple parity disks – a Hamming code is used – able to correct single-bit errors and detect double-bit errors.
- Although RAID 2 requires fewer disks than RAID 1, it is rather costly.
- Only effective where many disk errors occur – therefore never been implemented, as individual disks and disk drives are highly reliable.

RAID 3

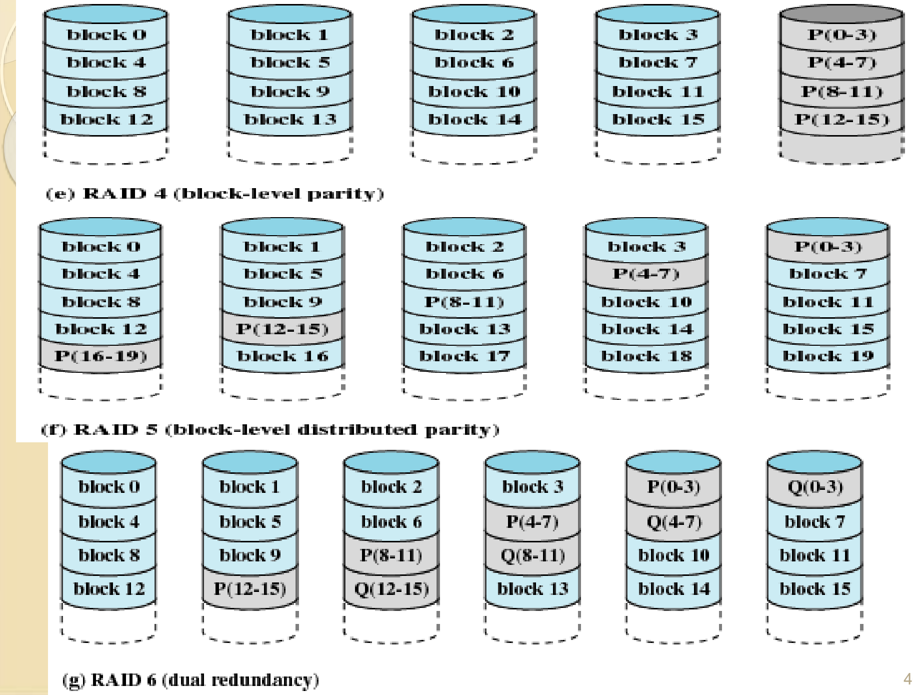
- Organized in a similar fashion to RAID 2, except that it requires only a single redundant disk no matter how large the disk array.
- Instead of an error-correcting code, a simple parity bit is computed for the set of individual bits in the same position on all of the data disks.
- In drive failure, the parity drive is accessed and data is reconstructed from the remaining devices. Once the failed drive is replaced, the missing data can be restored on the new drive and operation resumed.
- Performance: data are striped in very small strips → very high data transfer rates, esp. in large transfers.



(d) RAID 3 (bit-interleaved parity)

RAID 4, 5, 6

- Use independent access technique
  - Each member disk operates independently.
  - Separate I/O requests can be satisfied in parallel.
- Thus,
  - more suitable for applications that require high I/O request rates.
  - Less suitable for applications that require high data transfer rates.
- Data striping: relatively large strips.



block 0	block 1	block 2	block 3	P(0-3)
block 4	block 5	block 6	block 7	P(4-7)
block 8	block 9	block 10	block 11	P(8-11)
block 12	block 13	block 14	block 15	P(12-15)

(e) RAID 4 (block-level parity)

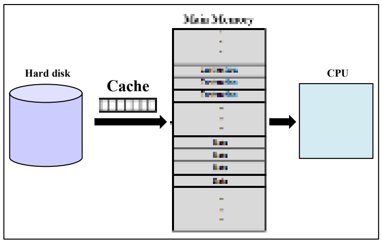
block 0	block 1	block 2	block 3	P(0-3)
block 4	block 5	block 6	P(4-7)	block 7
block 8	block 9	P(8-11)	block 10	block 11
block 12	P(12-15)	block 13	block 14	block 15
P(16-19)	block 16	block 17	block 18	block 19

(f) RAID 5 (block-level distributed parity)

block 0	block 1	block 2	block 3	P(0-3)	Q(0-3)
block 4	block 5	block 6	P(4-7)	Q(4-7)	block 7
block 8	block 9	P(8-11)	Q(8-11)	block 10	block 11
block 12	P(12-15)	Q(12-15)	block 13	block 14	block 15

(g) RAID 6 (dual redundancy)

Disk Cache



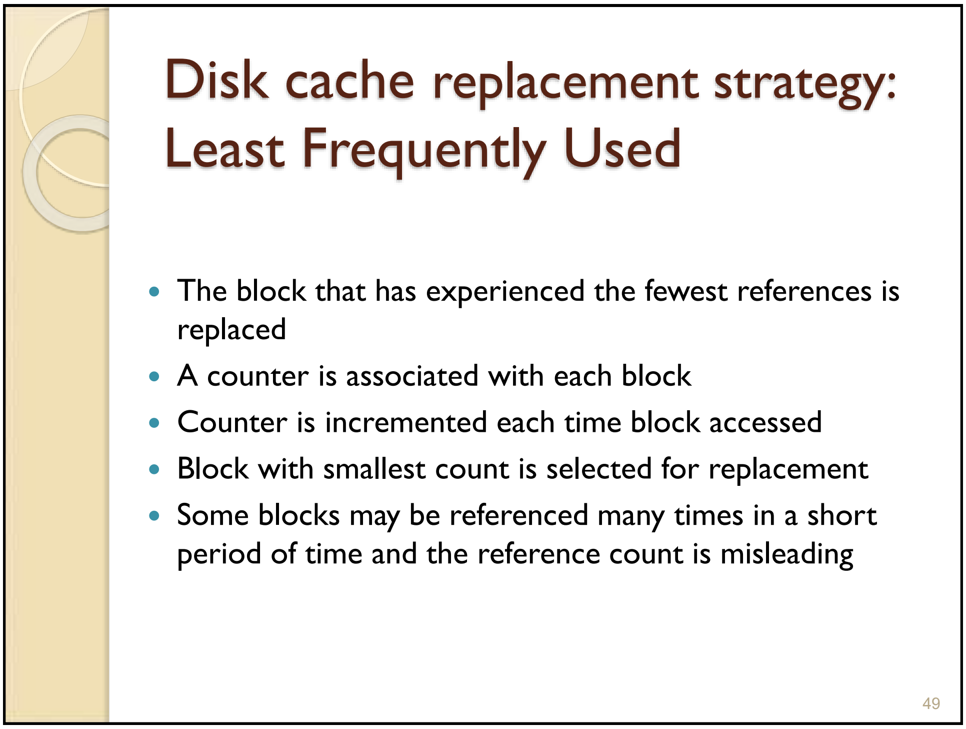
- ‘Cache memory’ – smaller and faster than MM
  - Interposed between MM and CPU

Disk cache

- Buffer in main memory for disk sectors
- Contains a copy of some of the sectors on the disk
- If disk cache is full => replacement strategy
  - Least recently used (LRU)
  - Least frequently used (LFU)

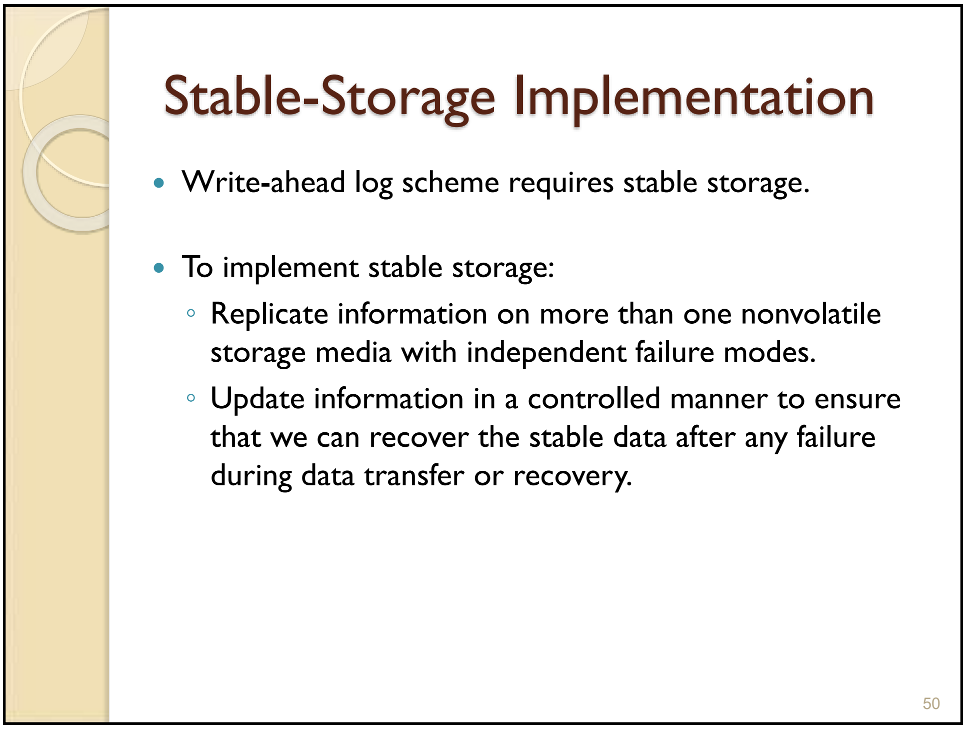
Disk cache replacement strategy: Least Recently Used

- The block that has been in the cache the longest with no reference to it is replaced
- The cache consists of a stack of blocks
- Most recently referenced block is on the top of the stack
- When a block is referenced or brought into the cache, it is placed on the top of the stack
- The block on the bottom of the stack is removed when a new block is brought in
- Blocks don’t actually move around in main memory
- A stack of pointers is used



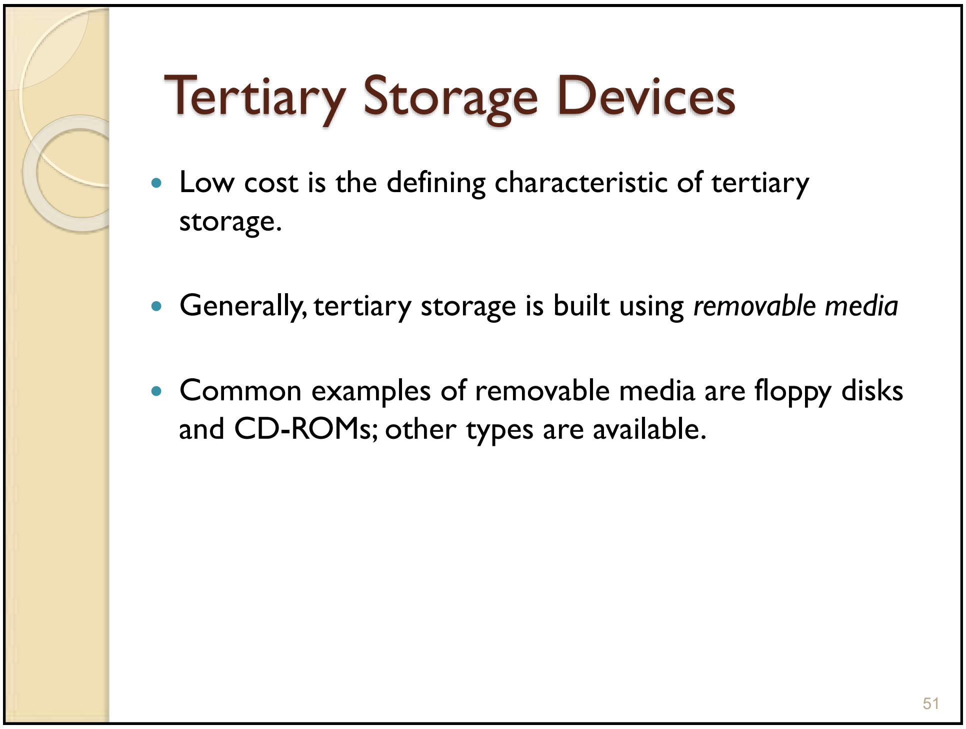
Disk cache replacement strategy: Least Frequently Used

- The block that has experienced the fewest references is replaced
- A counter is associated with each block
- Counter is incremented each time block accessed
- Block with smallest count is selected for replacement
- Some blocks may be referenced many times in a short period of time and the reference count is misleading



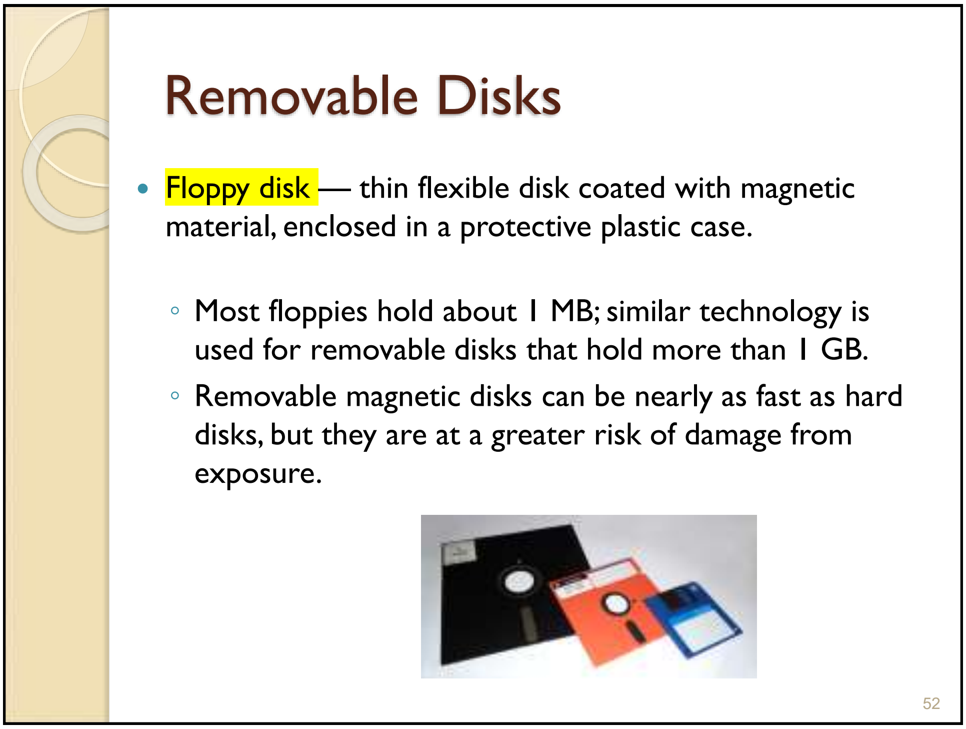
Stable-Storage Implementation

- Write-ahead log scheme requires stable storage.
- To implement stable storage:
  - Replicate information on more than one nonvolatile storage media with independent failure modes.
  - Update information in a controlled manner to ensure that we can recover the stable data after any failure during data transfer or recovery.



Tertiary Storage Devices

- Low cost is the defining characteristic of tertiary storage.
- Generally, tertiary storage is built using *removable media*
- Common examples of removable media are floppy disks and CD-ROMs; other types are available.



Removable Disks

- Floppy disk — thin flexible disk coated with magnetic material, enclosed in a protective plastic case.
  - Most floppies hold about 1 MB; similar technology is used for removable disks that hold more than 1 GB.
  - Removable magnetic disks can be nearly as fast as hard disks, but they are at a greater risk of damage from exposure.

Floppy Drive Connection

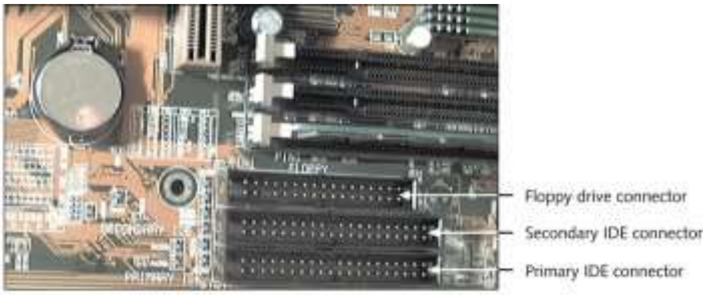


Figure 1-21 A motherboard usually provides a connection for a floppy drive cable

Floppy Drive Connection (continued)

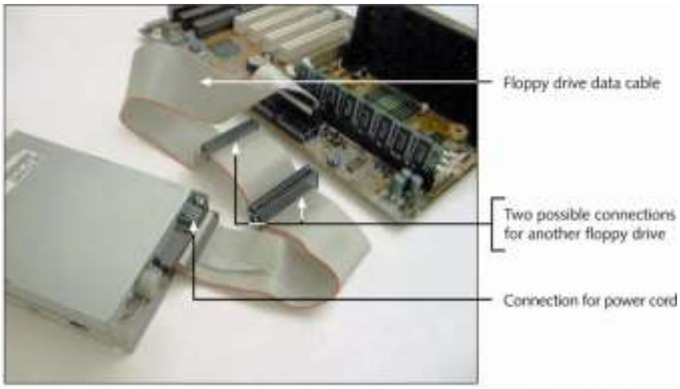
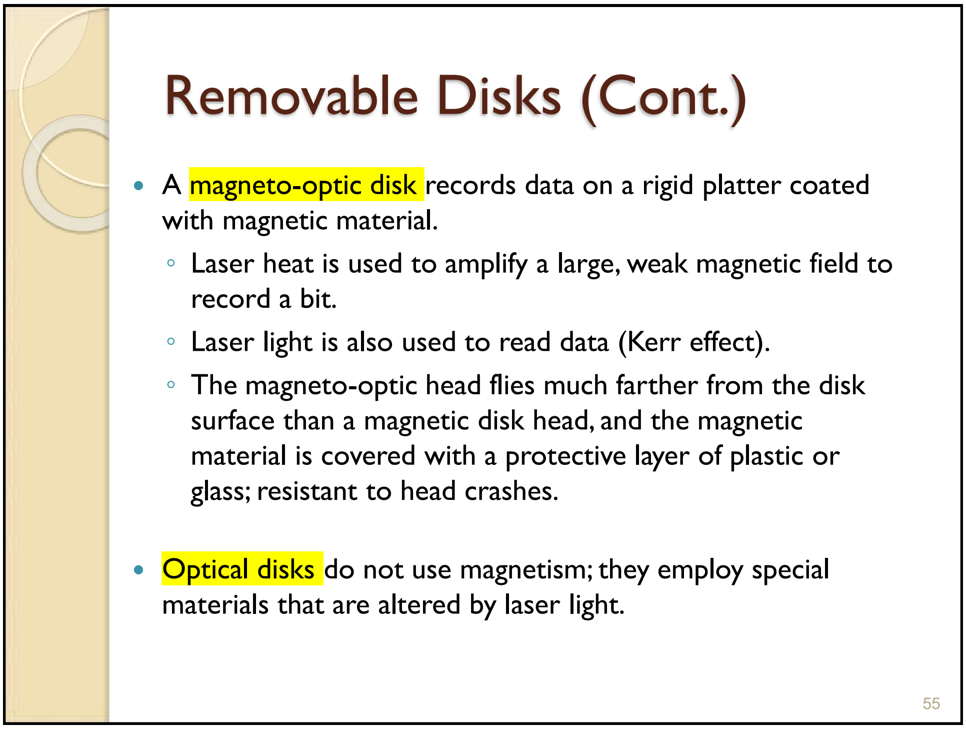

Floppy drive data cable
Two possible connections for another floppy drive
Connection for power cord

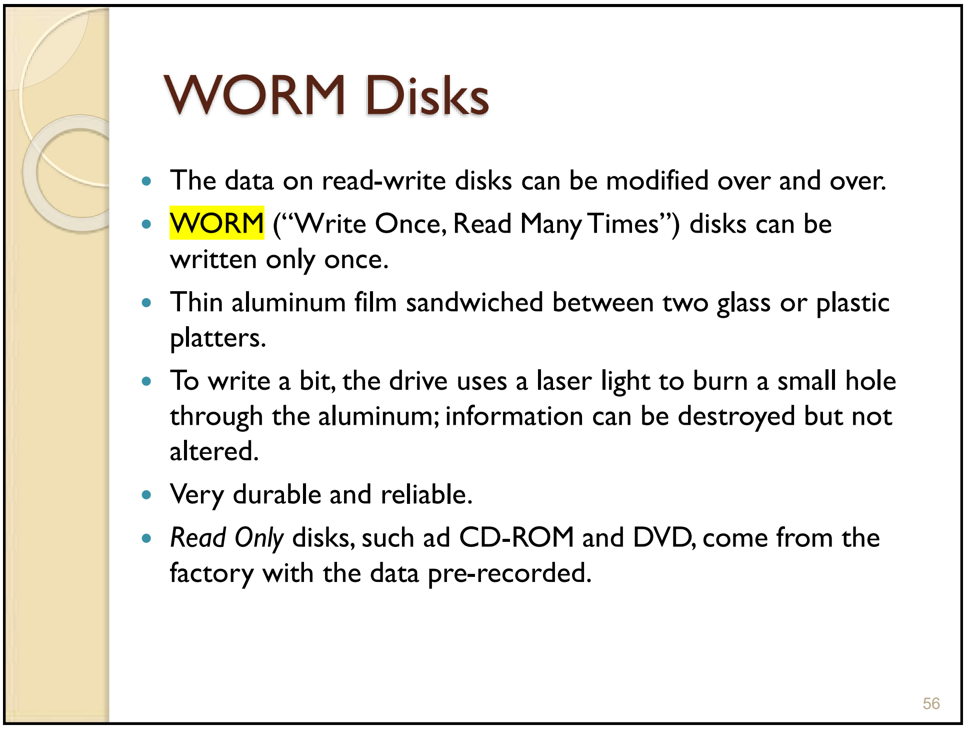
Figure 1-22 One floppy drive connection on a motherboard can support one or two floppy drives



Removable Disks (Cont.)

- A magneto-optic disk records data on a rigid platter coated with magnetic material.
  - Laser heat is used to amplify a large, weak magnetic field to record a bit.
  - Laser light is also used to read data (Kerr effect).
  - The magneto-optic head flies much farther from the disk surface than a magnetic disk head, and the magnetic material is covered with a protective layer of plastic or glass; resistant to head crashes.

- Optical disks do not use magnetism; they employ special materials that are altered by laser light.



WORM Disks

- The data on read-write disks can be modified over and over.
- WORM ("Write Once, Read Many Times") disks can be written only once.
- Thin aluminum film sandwiched between two glass or plastic platters.
- To write a bit, the drive uses a laser light to burn a small hole through the aluminum; information can be destroyed but not altered.
- Very durable and reliable.
- *Read Only* disks, such ad CD-ROM and DVD, come from the factory with the data pre-recorded.

CD-ROM Drive Connection

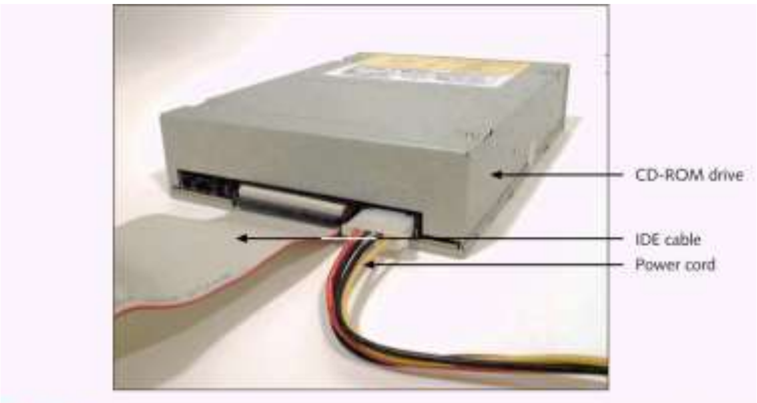


Figure 1-23 Most CD-ROM drives are EIDE devices and connect to the motherboard by way of an IDE data cable

Tapes

- Compared to a disk, a tape is less expensive and holds more data, but random access is much slower.
- Tape is an economical medium for purposes that do not require fast random access, e.g., backup copies of disk data, holding huge volumes of data.
- Large tape installations typically use robotic tape changers that move tapes between tape drives and storage slots in a tape library.
  - stacker – library that holds a few tapes
  - silo – library that holds thousands of tapes
- A disk-resident file can be *archived* to tape for low cost storage; the computer can *stage* it back into disk storage for active use.

Tapes

59

59

Epilogue:
A+ Installing HD
A+ Optimizing and Protecting HD
A+ Floppy Drive

60

60