BILKENT UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING
CS342 OPERATING SYSTEMS

LECTURE 12
MASS STORAGE
(CHAPTER 12)

Dr. İbrahim Körpeoğlu
http://www.cs.bilkent.edu.tr/~korpe

References

- *The slides here are adapted/modified from the textbook and its slides: Operating System Concepts, Silberschatz et al., 7th & 8th editions, Wiley.*

REFERENCES

- Operating System Concepts, 7th and 8th editions, Silberschatz et al. Wiley.
- Modern Operating Systems, Andrew S. Tanenbaum, 3rd edition, 2009.

Outline

- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Disk Attachment
- Stable-Storage Implementation
- Tertiary Storage Devices
- Operating System Issues
- Performance Issues

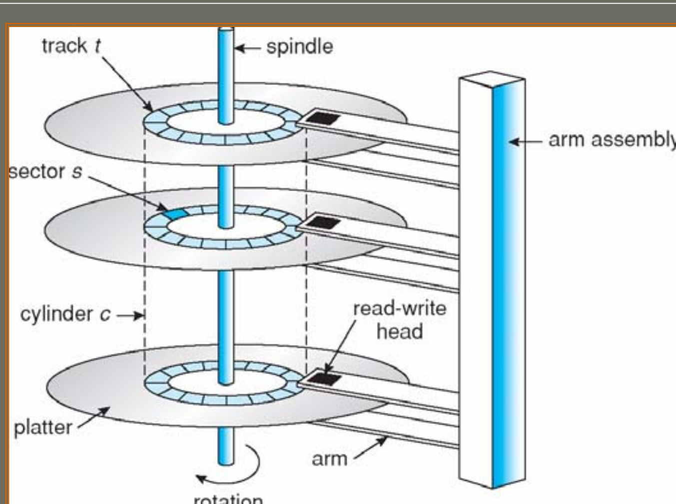
Objectives

- 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 operating-system services provided for mass storage, including RAID and HSM

Overview of Mass Storage Systems: Magnetic Disks

- 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

Moving-head Disk Mechanism



Overview of Mass Storage Systems: Magnetic Tapes

- Magnetic tape
  - Was early secondary-storage medium
  - Relatively permanent and holds large quantities of data
    - 20-200GB typical storage
  - Mainly used for backup, storage of infrequently-used data, transfer medium between systems

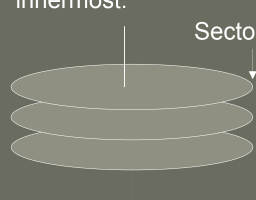
  - Access time slow
  - Random access ~1000 times slower than disk
  - Once data under head, transfer rates comparable to disk

  - Common technologies are 4mm, 8mm, 19mm, LTO-2 and SDLT

Disk Structure

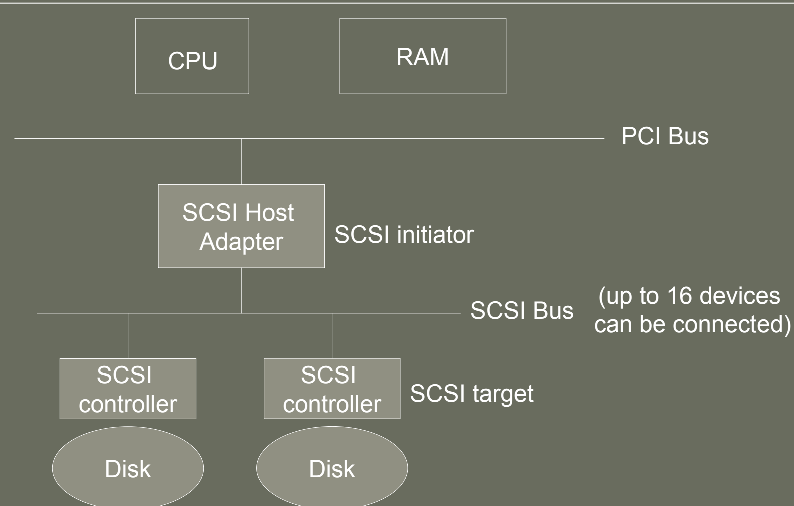
- Disk drives are addressed as large 1-dimensional arrays of *logical blocks*, where the logical block is the smallest unit of transfer.

- 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

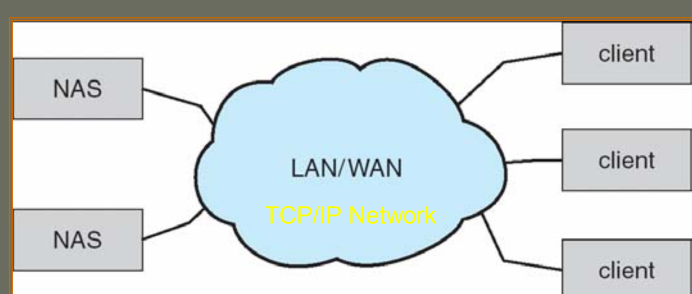
- Host-attached storage accessed through I/O ports talking to I/O busses
- Attachment technologies and protocols
  - IDE
  - SCSI
  - Fiber Channel
- SCSI itself is a bus, up to 16 devices on one cable, **SCSI initiator** requests operation and **SCSI targets** perform tasks
  - Each target can have up to 8 **logical units** (disks attached to device controller
- FC (fiber channel) is high-speed serial architecture
  - Can be switched fabric with 24-bit address space – the basis of **storage area networks** (**SANs**) 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)
- NFS and CIFS are common protocols used for network attached storage
  - We use those protocols to access remote storage that is connected to a network.
- Implemented via remote procedure calls (RPCs) between host and storage
- New iSCSI protocol uses IP network to carry the SCSI protocol
  - SCSI – SCSI bus/cable (local/host attached)
  - iSCSI – TCP/IP network (network attached)

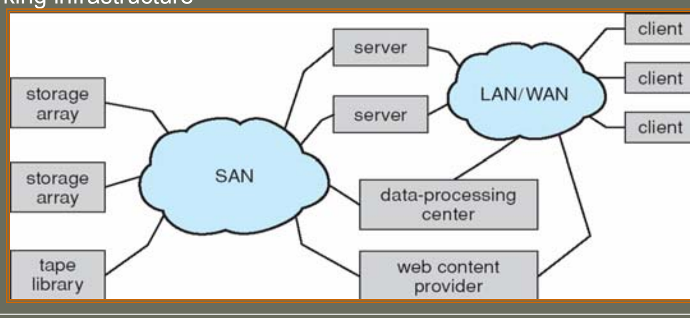
Network Attached Storage



NFS Protocol

Storage Area Network

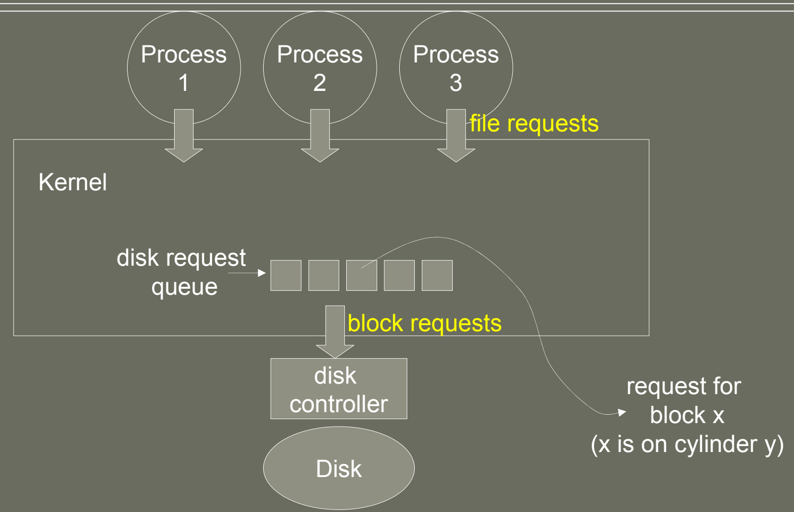
- Common in large storage environments (and becoming more common)
- Multiple hosts attached to multiple storage arrays – flexible
- Uses a different communication infrastructure (SAN) than the common networking infrastructure



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 to move the head to the cylinder containing the desired sector (block).
  - *Rotational latency* is the additional time waiting for the disk to rotate the desired sector to the disk head.
- Minimize seek time
- Seek time $\approx$ seek distance (between cylinders)
- 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 I/O queue



Disk Scheduling

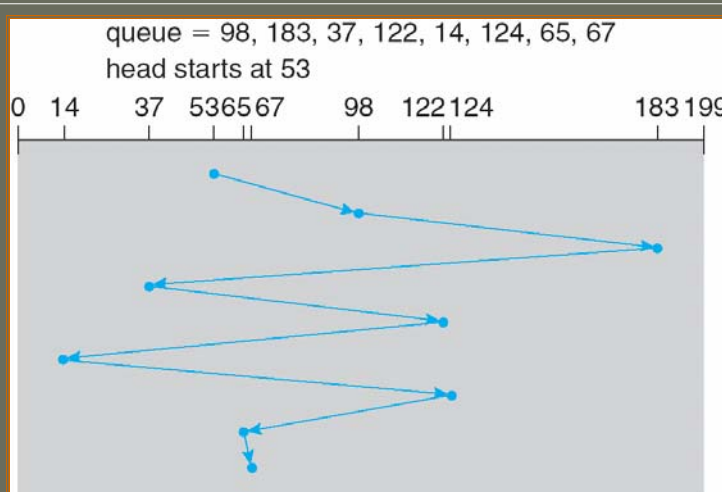
- Several algorithms exist to schedule the servicing of disk I/O requests.
- Assume disk has cylinders from 0 to 199.
- We illustrate them with a request queue. In the queue we have requests for blocks sitting in various cylinders. We just focus on the cylinder numbers.

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

Head pointer 53 (the head is currently on cylinder 53)

We have 8 requests queued. They are for blocks sitting on cylinders 98, 183, …

FCFS

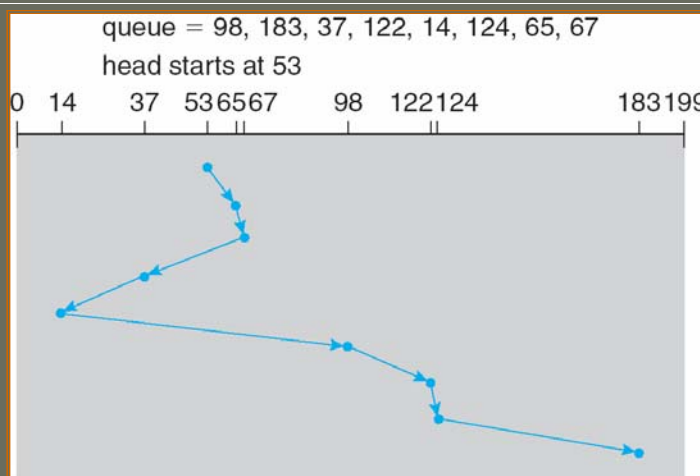


total head movement = 640 cylinders

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.

SSTF

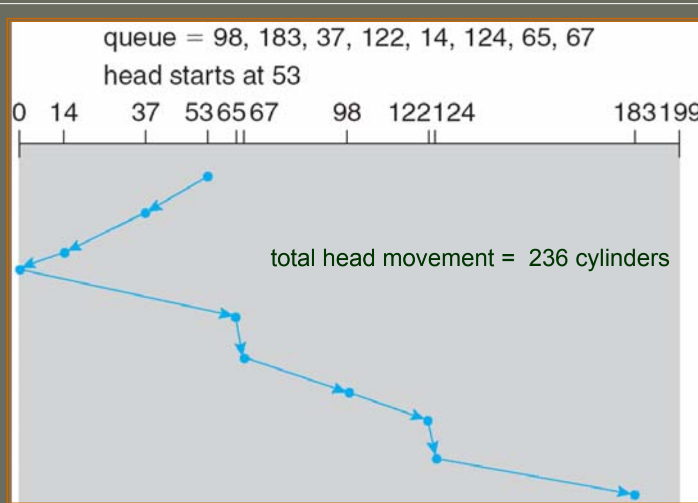


total head movement = 236 cylinders

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.
- Sometimes called the *elevator algorithm*.

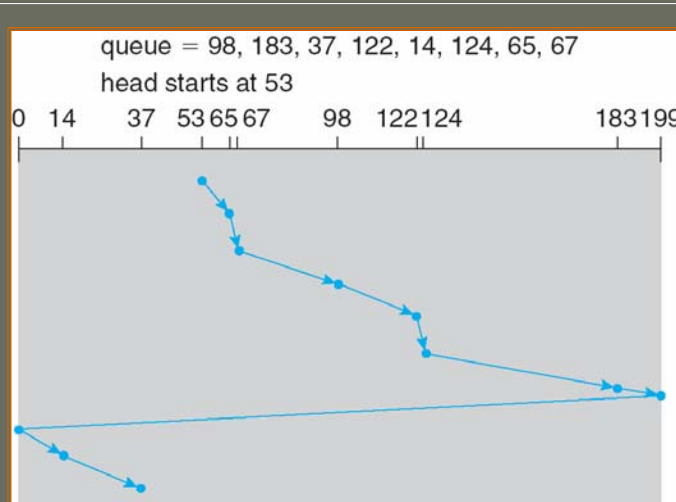
SCAN



C-SCAN

- 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.

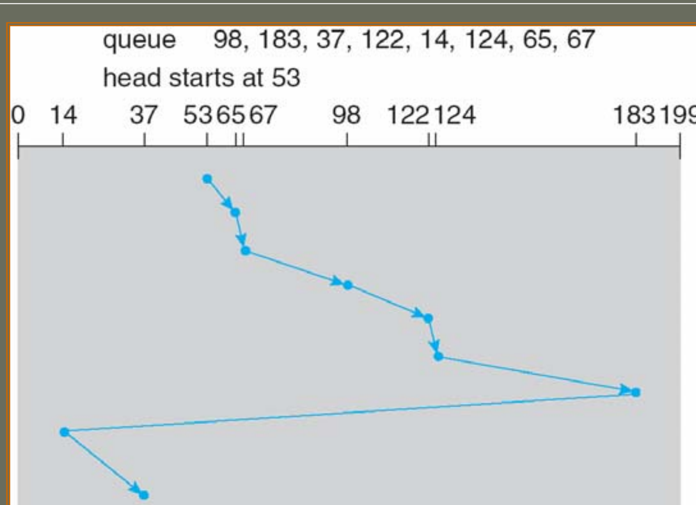
C-SCAN



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); and goes to the first request in the other end of the disk.

C-LOOK



Selecting a Disk-Scheduling Algorithm

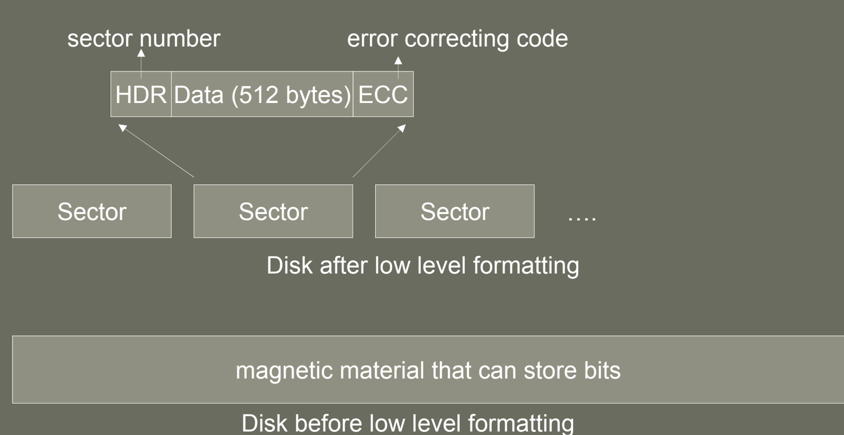
- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk.
- Performance depends on the number and types of requests.
- Requests for disk service can be influenced by the file-allocation method.
- The disk-scheduling algorithm should be 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.

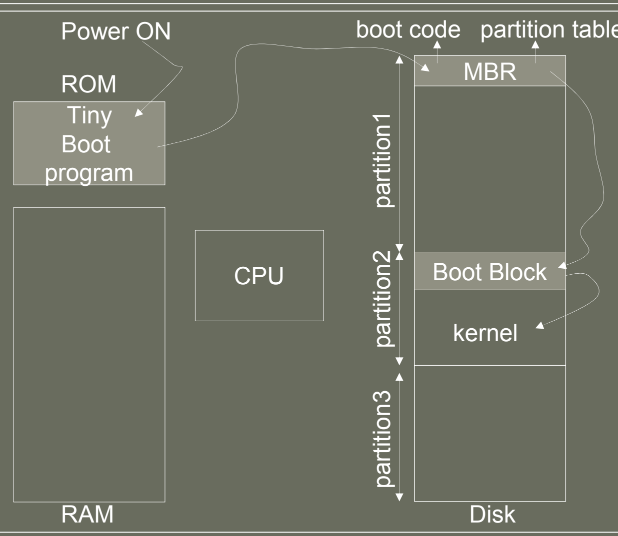
- To use a disk to hold files, the operating system still needs to record its own data structures on the disk.
  - *Partition* the disk into one or more groups of cylinders.
  - *Logical formatting* or "making a file system".

Low Level Formatting



Boot Process

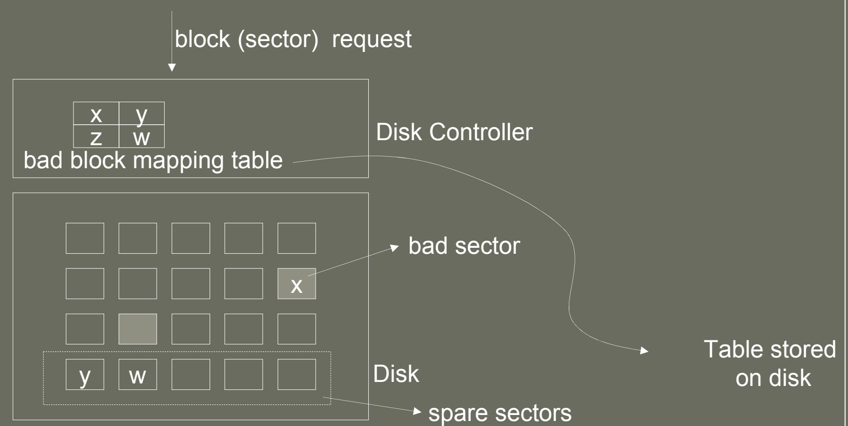
1. Boot code in ROM is run; it brings MBR into memory and starts MBR boot code
2. MBR boot code runs; looks to partition table; learns about the boot partition; brings and starts the boot code in the boot partition
3. Boot code in boot partition loads the kernel sitting in that partition



Bad Blocks

- Disk disk sectors (blocks) may become defective. Can no longer store data.
- System should not put data there.
- A bad block X can be remapped to a good block Y
- Whenever OS tries to access X, disk controller accesses Y.
- Some sectors (blocks) of disk can be reserved for this mapping.
  - This is called sector sparing.

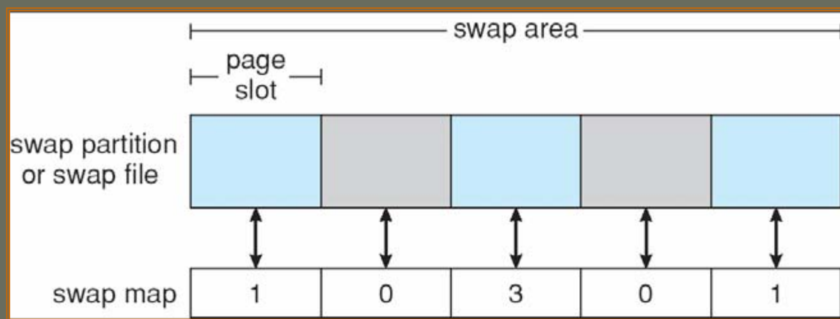
Bad Blocks



Swap Space Management

- Swap-space — Virtual memory uses disk space as an extension of main memory.
- Swap-space can be carved out of the normal file system, or, more commonly, it can be in a separate disk partition.
- Swap-space management
  - 4.3BSD allocates swap space when process starts; holds *text segment* (the program) and *data segment.*
  - Kernel uses *swap maps* to track swap-space use.
  - Solaris 2 allocates swap space only when a page is forced out of physical memory, not when the virtual memory page is first created.

Data Structures for Swapping on Linux Systems



RAID Structure

- **RAID** – multiple disk drives provides **reliability** via **redundancy**.
- RAID is arranged into six different levels.
- RAID: Redundant Array of Inexpensive Disks

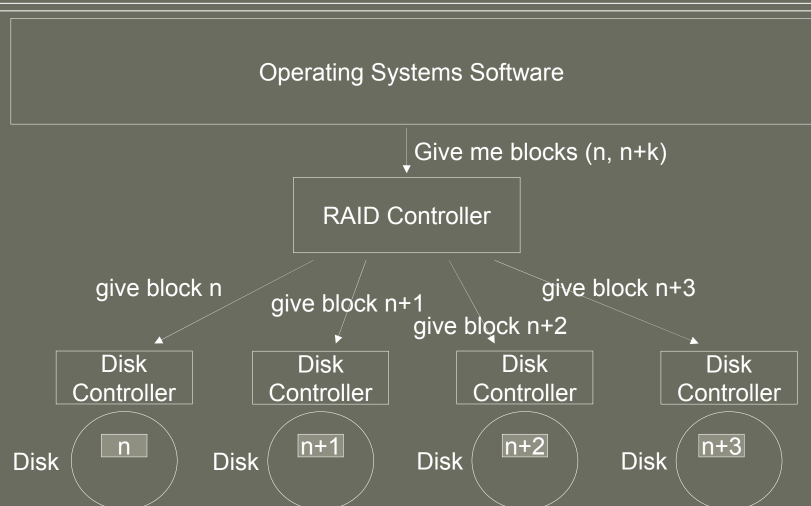
If you have many disks, the probability of one of them failing becomes higher.

We need to organize data for Reliability and Performance

RAID

- 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.
  - Improves performance
- 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



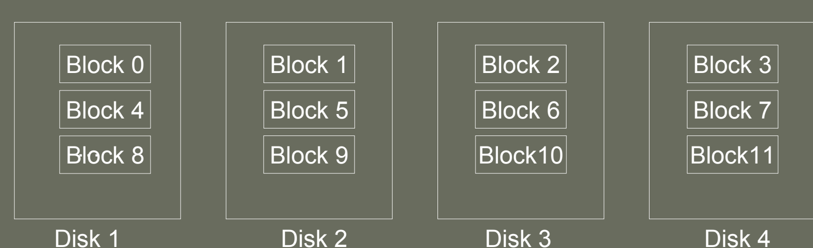
Different RAID Organizations (called Levels)

- RAID Level 0: block level striping (non redundant)
- RAID Level 1: mirroring
- RAID Level 2: bit level striping + error correcting codes
- RAID Level 3: bit level striping + parity
- RAID Level 4: block level striping + parity
- RAID Level 5: block level striping + distributed parity
- ….

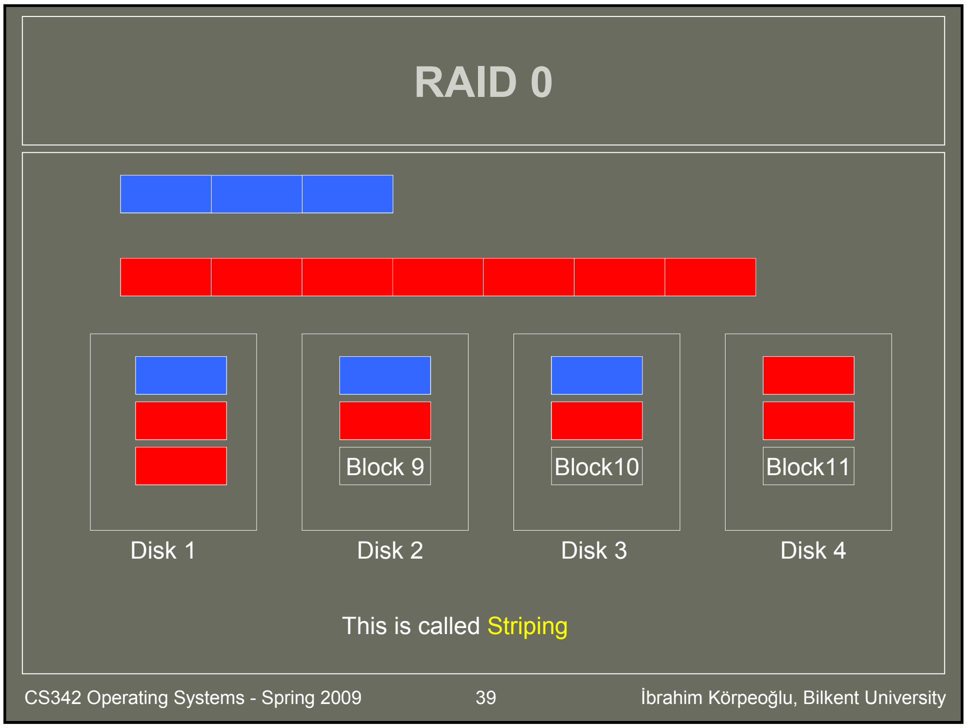
RAID 0

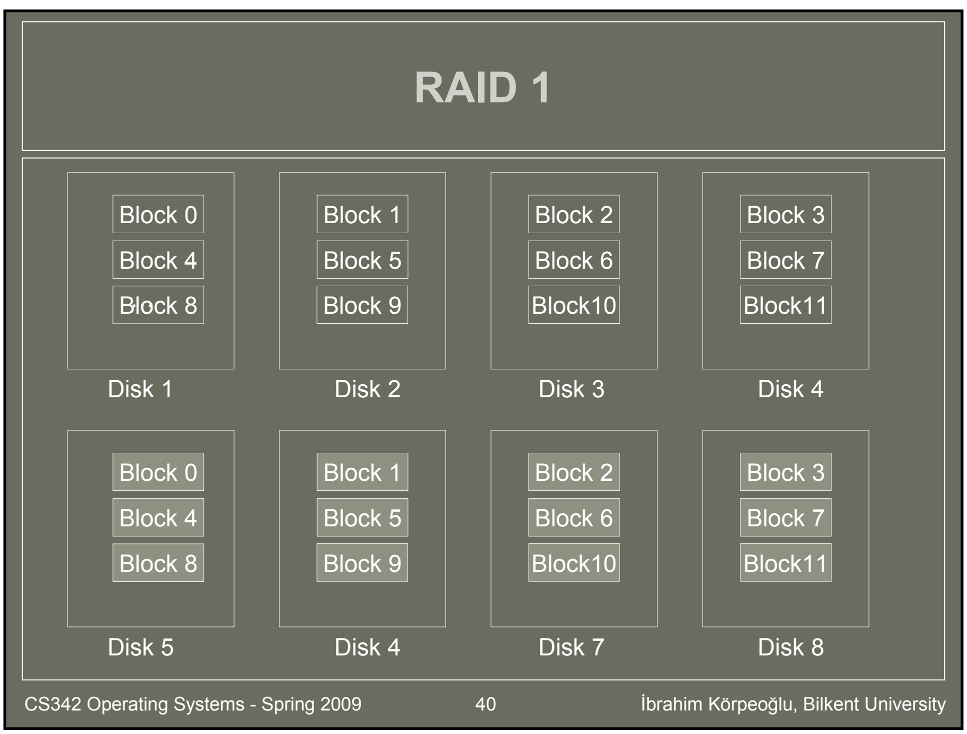
data: in blocks; adjacent blocks go into different disks

One block can be one or more sectors: k sectors (k can be 0,1,…)



No redundancy;
parallel read for large data transfers (larger than block size)


RAID 0
Block 9
Block10
Block11
Disk 1
Disk 2
Disk 3
Disk 4
This is called Striping
CS342 Operating Systems - Spring 2009
39
İbrahim Körpeoğlu, Bilkent University


RAID 1
Block 0
Block 4
Block 8
Block 1
Block 5
Block 9
Block 2
Block 6
Block10
Block 3
Block 7
Block11
Disk 1
Disk 2
Disk 3
Disk 4
Block 0
Block 4
Block 8
Block 1
Block 5
Block 9
Block 2
Block 6
Block10
Block 3
Block 7
Block11
Disk 5
Disk 4
Disk 7
Disk 8
CS342 Operating Systems - Spring 2009
40
İbrahim Körpeoğlu, Bilkent University

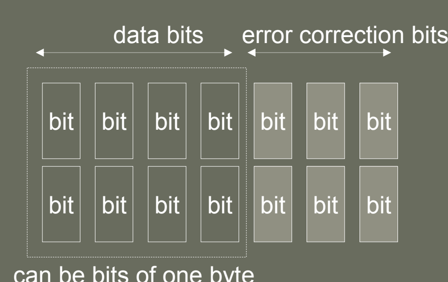
RAID 1

- We are just mirroring the disks (copying one disk to another one).
- Without striping: no performance gain, except for reads (doubled read rate)
- Reliability provided. If one disk fails, data can be recover from other disk.
- If there are originally N disks; we need N more disks to mirror
  – Quite costly in terms of disks required. This cost is for reliability

overhead/data = 1/1

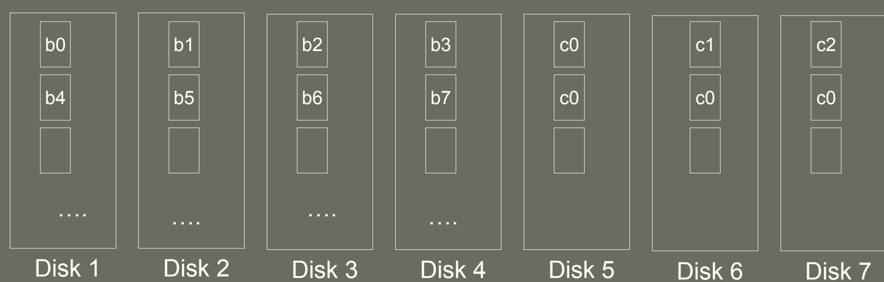
RAID 2

- Bit level striping.
- Error correcting codes used.
  – For example every 4 bit is protected with 3 redundant bits. If one of these 4 bits is in error, we can understand which one it is and correct it using 3 other code bits.
  – Hamming codes can be used.



overhead/data = 3/4

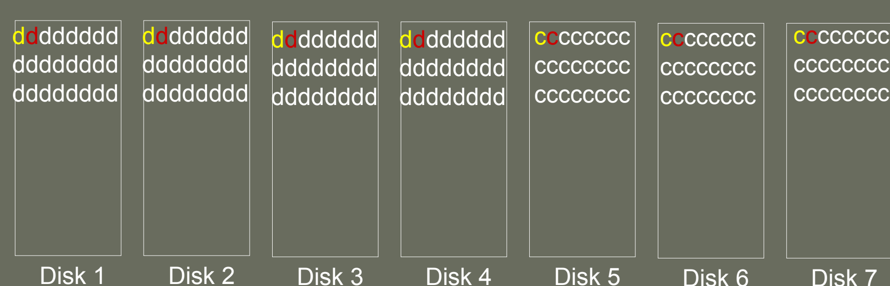
RAID 2 organization



bx: data bits

cx: control bits

RAID 2

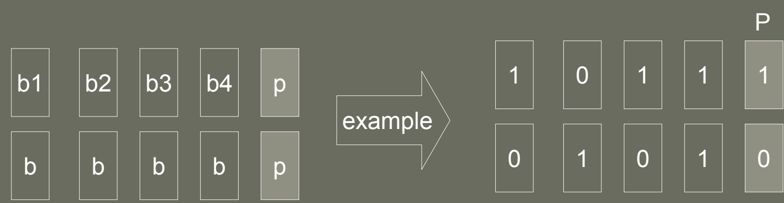


d: data bit

c: control bit

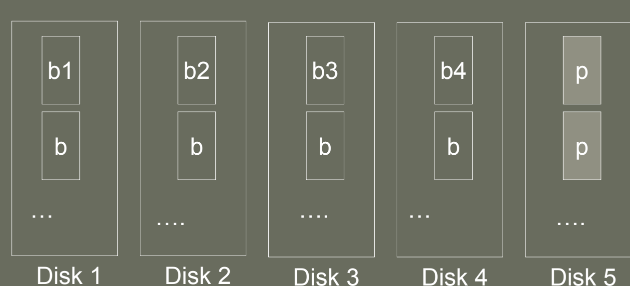
RAID 3

- Improved on RAID 2 in terms of space efficiency
- Only one control bit is used for k data bits
  - That control bit is a parity bit
  - compute the parity of k bits and store it in the parity bit.
  - k can be 4, 8, …
- This is enough to detect and correct one bit errors



overhead/data = 1/4

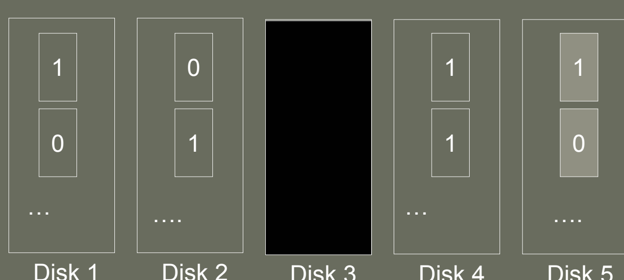
RAID 3



RAID 3

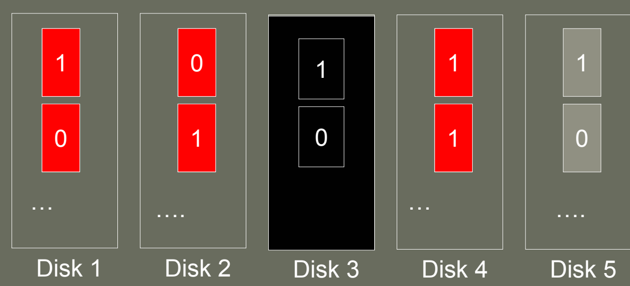
RAID 3

Let one disk fail! How can we recover its data



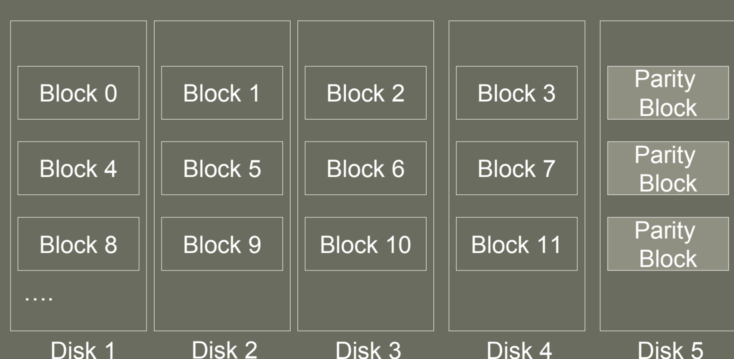
Look to disks 1, 2, 4, and 5. compute the parity and according to that generate the content of disk 3.

RAID 3



RAID 4

- Uses block level striping (like RAID 0)
- But uses an additional disk to store parity block

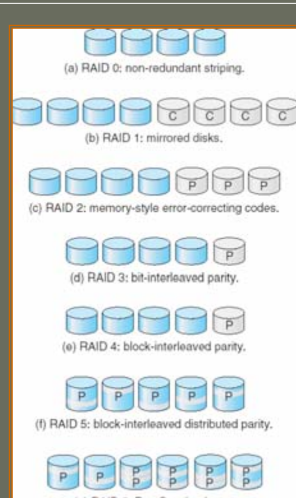


RAID 5

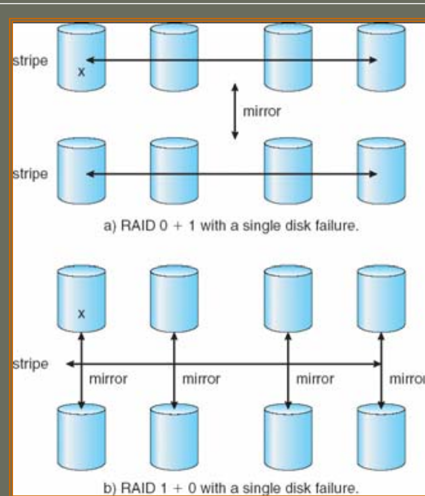
- Parity blocks are distributed on other disks.
- Load on parity disk is distributed in this way

Disk 1	Disk 2	Disk 3	Disk 4	Disk 5
Block 0	Block 1	Block 2	Block 3	Parity Block
Block 4	Block 5	Block 6	Parity Block	Block 7
Block 8	Block 9	Parity Block	Block 10	Block 11
….				

RAID Levels



RAID Levels



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;

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.

Removable Disks

- A magneto-optic disk records data on a rigid platter coated with magnetic material.
- 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.

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.

Operating System Issues

- Major OS jobs are to manage physical devices and to present a virtual machine abstraction to applications

- For hard disks, the OS provides two abstraction:
  – Raw device – an array of data blocks.
  – File system – the OS queues and schedules the interleaved requests from several applications.

Application Interface

- Most OSs handle removable disks almost exactly like fixed disks — a new cartridge is formatted and an empty file system is generated on the disk.
- Tapes are presented as a raw storage medium, i.e., and application does not not open a file on the tape, it opens the whole tape drive as a raw device.
  - Usually the tape drive is reserved for the exclusive use of that application .
  - Since the OS does not provide file system services, the application must decide how to use the array of blocks.
  - Since every application makes up its own rules for how to organize a tape, a tape full of data can generally only be used by the program that created it.

Tape drives

- The basic operations for a tape drive differ from those of a disk drive.
- **locate** positions the tape to a specific logical block (corresponds to **seek**).
- The **read position** operation returns the logical block number where the tape head is.
- Tape drives are "append-only" devices; updating a block in the middle of the tape also effectively erases everything beyond that block.
- An EOT mark is placed after a block that is written.

File Naming

- The issue of naming files on removable media is especially difficult when we want to write data on a removable cartridge on one computer, and then use the cartridge in another computer.
- Contemporary OSs generally leave the name space problem unsolved for removable media, and depend on applications and users to figure out how to access and interpret the data.
- Some kinds of removable media (e.g., CDs) are so well standardized that all computers use them the same way.

Hierarchical Storage Management (HSM)

- A hierarchical storage system extends the storage hierarchy beyond primary memory and secondary storage to incorporate tertiary storage — usually implemented as a jukebox of tapes or removable disks.
- Usually incorporate tertiary storage by extending the file system.
  - Small and frequently used files remain on disk.
  - Large, old, inactive files are archived to the jukebox.
- HSM is usually found in supercomputing centers and other large installations that have enormous volumes of data.

Speed

- Two aspects of speed in tertiary storage are bandwidth and latency.
- Bandwidth is measured in bytes per second.
  - Sustained bandwidth – average data rate during a large transfer; # of bytes/transfer time.
    Data rate when the data stream is actually flowing.
  - Effective bandwidth – average over the entire I/O time, including **seek** or **locate**, and cartridge switching.
    Drive's overall data rate.
    - Effective bandwidth <= sustained bandwidth

Speed

- Access latency – amount of time needed to locate data.
  - Access time for a disk – move the arm to the selected cylinder and wait for the rotational latency; < 35 milliseconds.
  - Access on tape requires winding the tape reels until the selected block reaches the tape head; tens or hundreds of seconds.
    - Generally say that random access within a tape cartridge is about a thousand times slower than random access on disk.
- The low cost of tertiary storage is a result of having many cheap cartridges share a few expensive drives.
- A removable library is best devoted to the storage of infrequently used data, because the library can only satisfy a relatively small number of I/O requests per hour.

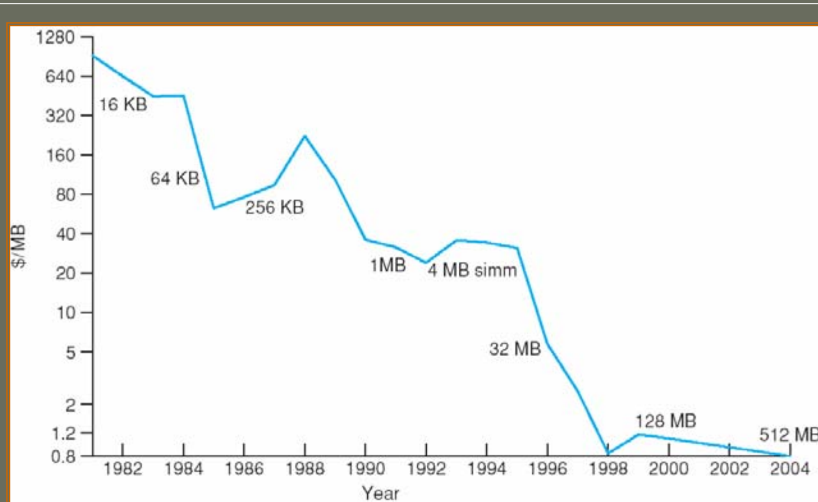
Reliability

- A fixed disk drive is likely to be more reliable than a removable disk or tape drive.
- An optical cartridge is likely to be more reliable than a magnetic disk or tape.
- A head crash in a fixed hard disk generally destroys the data, whereas the failure of a tape drive or optical disk drive often leaves the data cartridge unharmed.

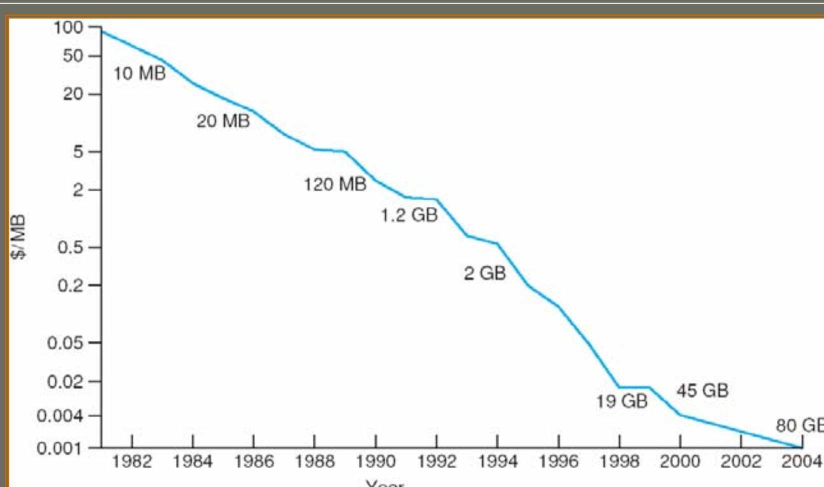
Cost

- Main memory is much more expensive than disk storage
- The cost per megabyte of hard disk storage is competitive with magnetic tape if only one tape is used per drive.
- The cheapest tape drives and the cheapest disk drives have had about the same storage capacity over the years.
- Tertiary storage gives a cost savings only when the number of cartridges is considerably larger than the number of drives.

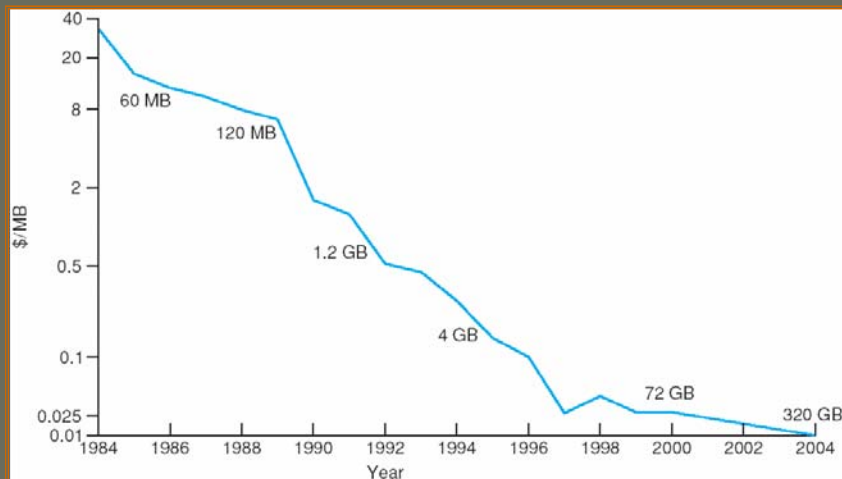
Price per Megabyte of DRAM, From 1981 to 2004



Price per Megabyte of Magnetic Hard Disk, From 1981 to 2004



Price per Megabyte of a Tape Drive, From 1984-2000



End of lecture