Operating Systems

Dr. J.M. Khalifeh

Mass-Storage Systems

Operating Systems

Dr. J.M. Khalifeh

Chapter 10: Mass-Storage Systems

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

Dr. J.M. Khalifeh Operating Systems

Objectives

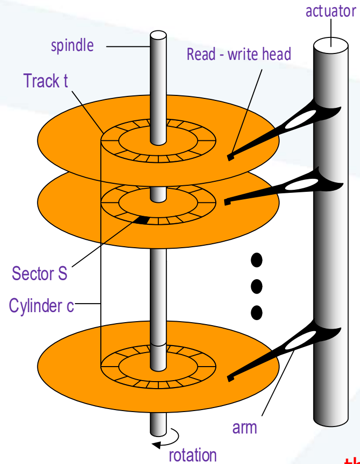
- ❖To describe the physical structure of secondary storage devices and its effects on the uses of the devices
- ❖To explain the performance characteristics of mass-storage devices
- ❖To evaluate disk scheduling algorithms
- ❖To discuss operating-system services provided for mass storage, including RAID

Dr. J.M. Khalifeh Operating Systems

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 -- 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, SAS, Firewire
  - ➢Host controller in computer uses bus to talk to disk controller built into drive or storage array

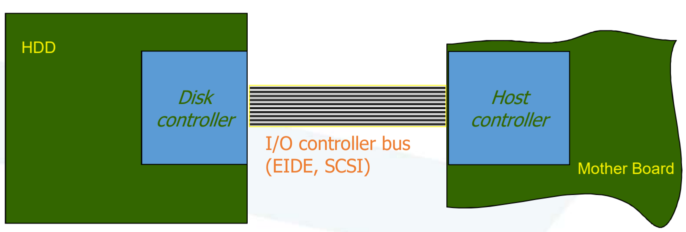
Moving-Head Disk Mechanism



- ❖Common platter diameters range from 1.8 to 5.25 inches
- ❖The surface of a platter is logically divided into circular *tracks* which are subdivided into *sectors*
- ❖The set of tracks that are at one arm position forms a *cylinder*
- ❖Most drives rotate 60 to 150 times per second

transfer rate
is the rate at which data flow between the drive and the computer

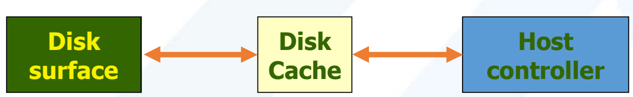
positioning time (random access time = the time to move the disk arm to the desired cylinder (*seek time)* + time for the desired sector to rotate to the disk head *(rotational latency)*



computer places a command into the host controller, typically using memory-mapped I/O ports

The host controller then sends the command via messages to the disk controller

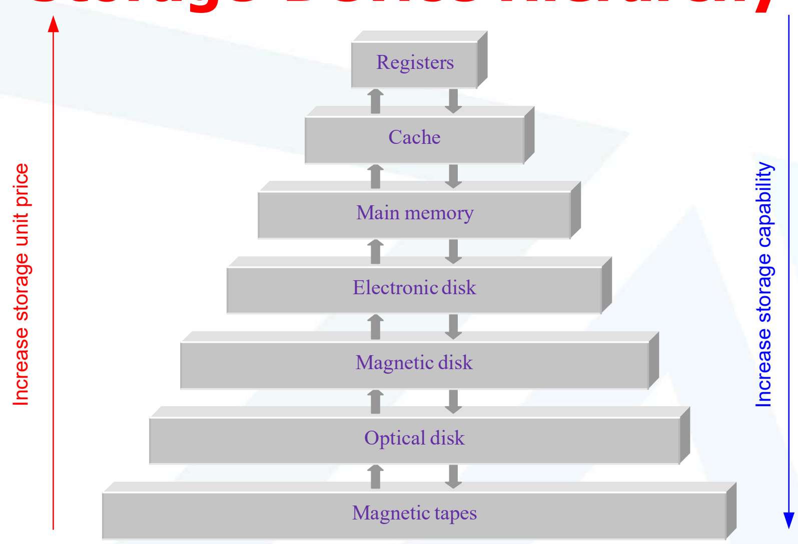
disk controller operates the disk-drive hardware to carry out the command



Storage Hierarchy

- Storage systems organized in hierarchy.
  - Speed
  - Cost
  - Volatility
- Speed of magnetic disk
  - *transfer rate* * size
  - *positioning time* (*random access time*)
    - ✓ *seek time* + *rotational latency*
- *Caching* – copying information into faster storage system; main memory can be viewed as a last *cache* for secondary storage.

Storage-Device Hierarchy



Hard Disks

Operating Systems — Dr. J.M. Khalifeh

- Platters range from .85” to 14” (historically)
  - Commonly 3.5”, 2.5”, and 1.8”
- Range from 30GB to 3TB per drive
- Performance
  - Transfer Rate – theoretical – 6 Gb/sec
  - Effective Transfer Rate – real – 1Gb/sec
  - Seek time from 3ms to 12ms – 9ms common for desktop drives
  - Average seek time measured or calculated based on 1/3 of trac
  - Latency based on spindle speed
    - ✓ 1 / (RPM / 60) = 60 / RPM
  - Average latency = ½ latency

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

(From Wikipedia)

Hard Disk Performance

Operating Systems — Dr. J.M. Khalifeh

- Access Latency = Average access time = average seek time + average 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
- 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 + 0.1ms + transfer time =
  - Transfer time = 4KB / 1Gb/s * 8Gb / GB * 1GB / 10242KB = 32 / (10242) = 0.031 ms
  - Average I/O time for 4KB block = 9.27ms + .031ms = 9.301ms

Operating Systems

Dr. J.M. Khalifeh

Solid-State Disks

- Nonvolatile memory used like a hard drive
  - Many technology variations
- Can be more reliable than HDDs
- More expensive per MB
- Maybe have shorter life span
- Less capacity
- But much faster
- Busses can be too slow -> connect directly to PCI for example
- No moving parts, so no seek time or rotational latency

Operating Systems

Dr. J.M. Khalifeh

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

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

Operating Systems
Dr. J.M. Khalifeh

Disk Attachment

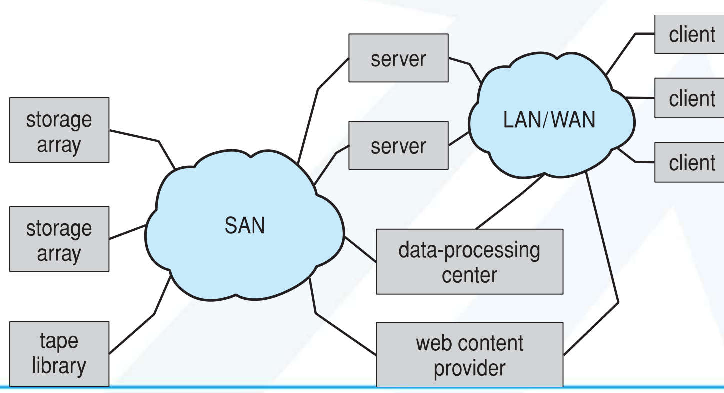
- Host-attached storage accessed through I/O ports talking to I/O busses
- SCSI itself is a bus, up to 16 devices on one cable, SCSI initiator requests operation and SCSI targets perform tasks
  - Each target can have up to 8 logical units (disks attached to device controller)
- FC is high-speed serial 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
- I/O directed to bus ID, device ID, logical unit (LUN)

Storage Array

- ❖Can just attach disks, or arrays of disks
- ❖Storage Array has controller(s), provides features to attached host(s)
  - ➢Ports to connect hosts to array
  - ➢Memory, controlling software (sometimes NVRAM, etc)
  - ➢A few to thousands of disks
  - ➢RAID, hot spares, hot swap (discussed later)
  - ➢Shared storage -> more efficiency
  - ➢Features found in some file systems
    - ✓ Snapshots, clones, thin provisioning, replication, deduplication, etc

Storage Area Network

- ❖Common in large storage environments
- ❖Multiple hosts attached to multiple storage arrays - flexible



Operating Systems
Dr. J.M. Khalifeh

Storage Area Network (Cont.)

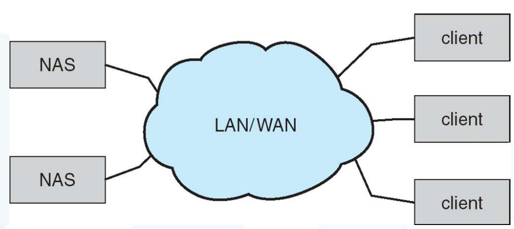
- SAN is one or more storage arrays
  - Connected to one or more Fibre Channel switches
- Hosts also attach to the switches
- Storage made available via LUN Masking from specific arrays to specific servers
- Easy to add or remove storage, add new host and allocate it storage
  - Over low-latency Fibre Channel fabric
- Why have separate storage networks and communications networks?
  - Consider iSCSI, FCoE

Operating Systems

Dr. J.M. Khalifeh

Network-Attached Storage

- Network-attached storage (NAS) is storage made available over a network rather than over a local connection (such as a bus)
  - 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
- iSCSI protocol uses IP network to carry the SCSI protocol
  - Remotely attaching to devices (blocks)



Operating Systems

Dr. J.M. Khalifeh

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 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.)

❖There are many sources of disk I/O request

- OS
- System processes
- Users processes

❖I/O request includes input or output mode, disk address, memory address, number of sectors to transfer

❖OS maintains queue of requests, per disk or device

❖Idle disk can immediately work on I/O request, busy disk means work must queue

- Optimization algorithms only make sense when a queue exists

Operating Systems Dr. J.M. Khalifeh

Operating Systems

Dr. J.M. Khalifeh

Disk Scheduling (Cont.)

- ❖Note that drive controllers have small buffers and can manage a queue of I/O requests (of varying “depth”)
- ❖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

Operating Systems

Dr. J.M. Khalifeh

FCFS

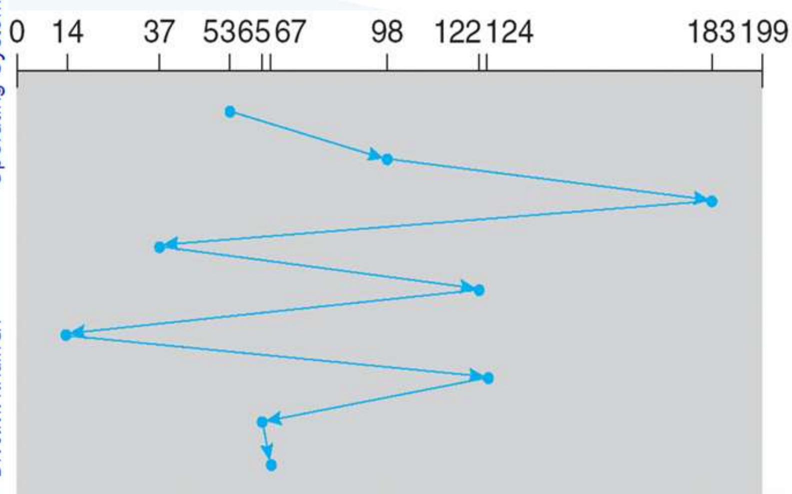


Illustration shows total head movement of 640 cylinders

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

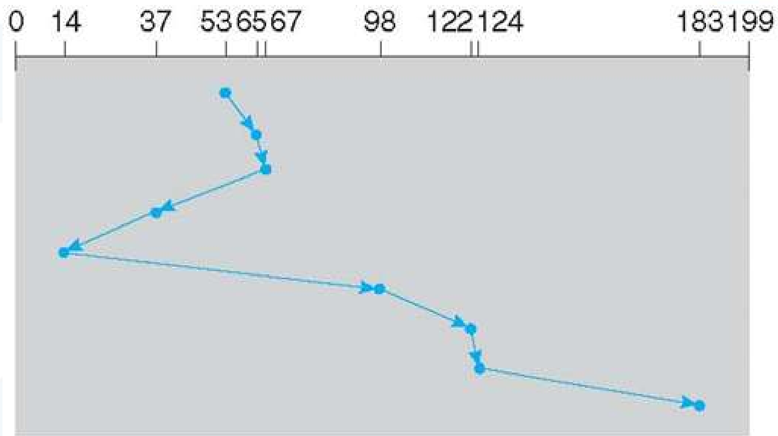
Head pointer 53

Operating Systems

Dr. J.M. Khalifeh

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



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

Head pointer 53

Operating Systems

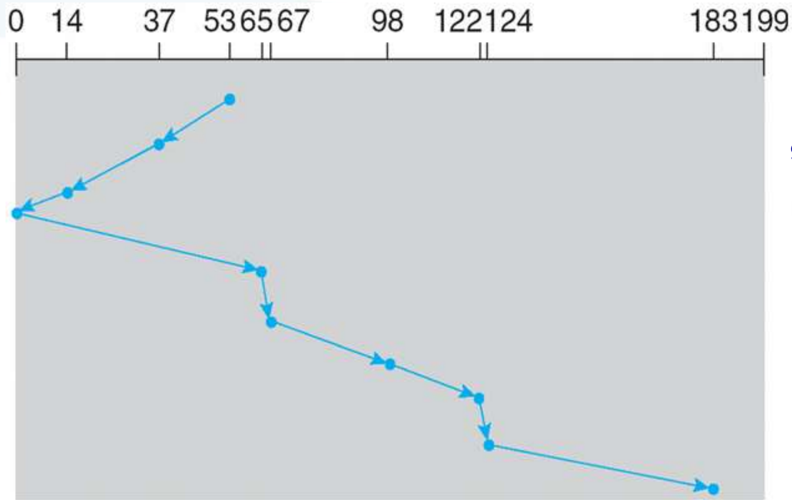
Dr. J.M. Khalifeh

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 236 cylinders
- ❖But note that if requests are uniformly dense, largest density at other end of disk and those wait the longest

Dr. J.M. Khalifeh

SCAN (Cont.)



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

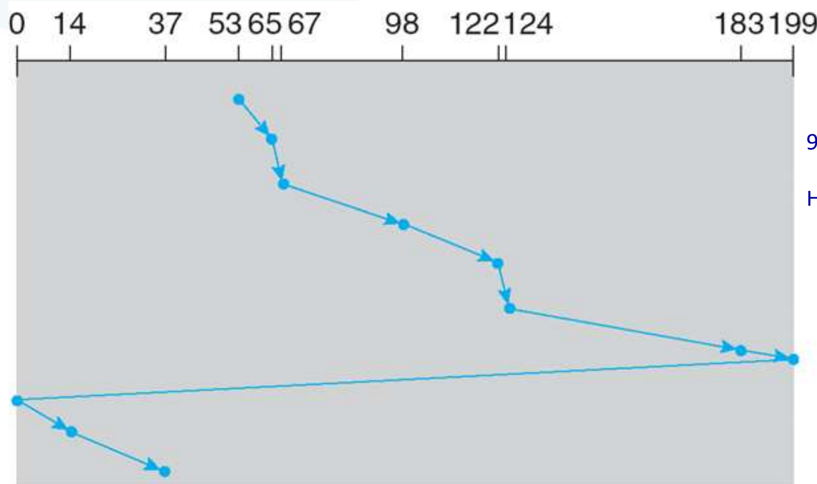
Head pointer 53

Dr. J.M. Khalifeh

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
- Total number of cylinders?

C-SCAN (Cont.)



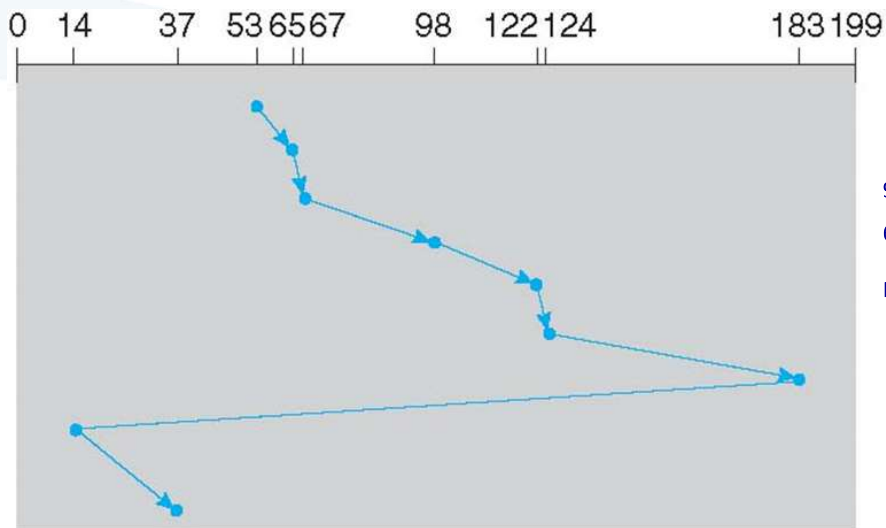
98, 183, 37, 122, 14, 124, 65, 67

Head pointer 53

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.
- ❖Total number of cylinders?

C-LOOK (Cont.)



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

Head pointer 53

Operating Systems
Dr. J.M. Khalifeh

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
  - Less starvation
- Performance depends on the number and types of requests
- Requests for disk service can be influenced by the file-allocation method
  - And metadata layout
- 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
- What about rotational latency?
  - Difficult for OS to calculate
- How does disk-based queueing effect OS queue ordering efforts?

Disk Management

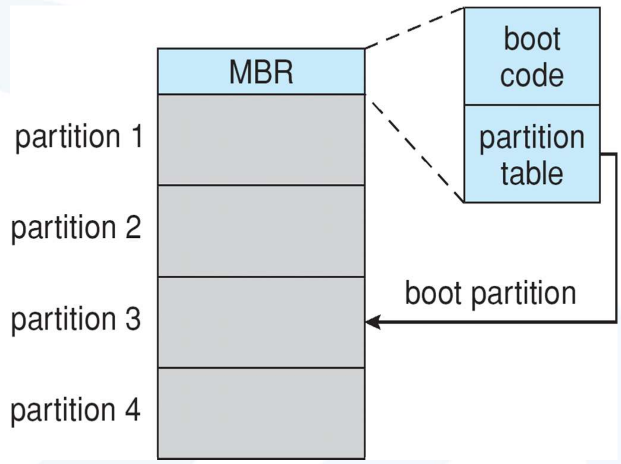
- ❖Low-level formatting, or physical formatting — Dividing a disk into sectors that the disk controller can read and write
  - ➢ Each sector can hold header information, plus data, plus error correction code (ECC)
  - ➢ Usually 512 bytes of data but can be selectable
- ❖To use a disk to hold files, the operating system still needs to record its own data structures on the disk
  - ➢ Partition the disk into one or more groups of cylinders, 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

Disk Management (Cont.)

- ❖Raw disk access for apps that want to do their own block management, keep OS out of the way (databases for example)
- ❖Boot block initializes system
  - ➢The bootstrap is stored in ROM
  - ➢Bootstrap loader program stored in boot blocks of boot partition
- ❖Methods such as sector sparing used to handle bad blocks

Operating Systems
Dr. J.M. Khalifeh

Booting from a Disk in Windows

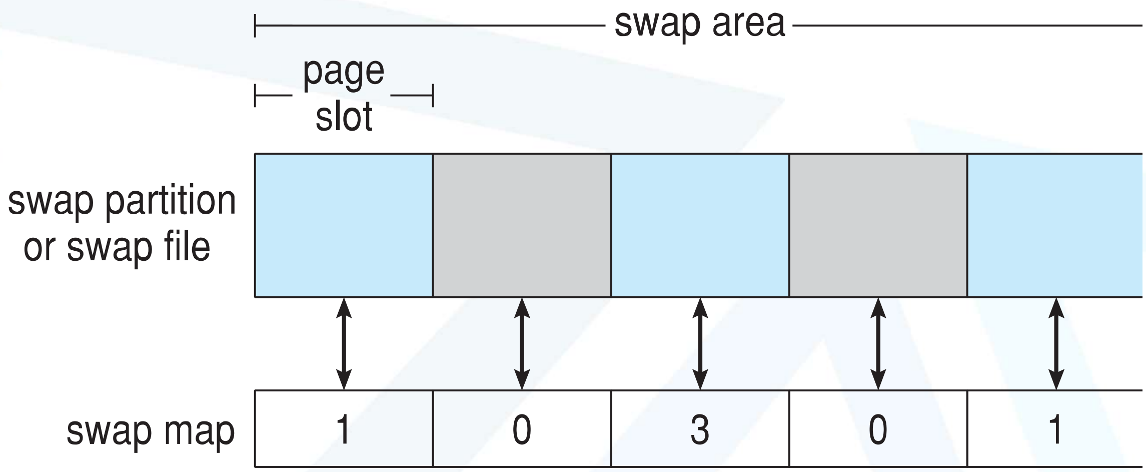


Operating Systems
Dr. J.M. Khalifeh

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 can be carved out of the normal file system, or, more commonly, it can be in a separate disk partition (raw)
- 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 dirty page is forced out of physical memory, not when the virtual memory page is first created
    - File data written to swap space until write to file system requested
    - Other dirty pages go to swap space due to no other home
    - Text segment pages thrown out and reread from the file system as needed
- What if a system runs out of swap space?
- Some systems allow multiple swap spaces

Data Structures for Swapping on Linux Systems



Operating Systems
Dr. J.M. Khalifeh

RAID Structure

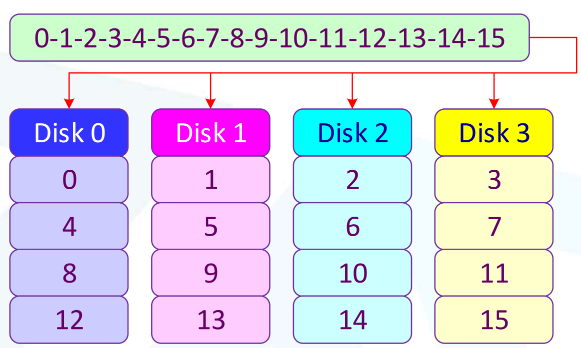
- RAID – redundant array of inexpensive disks
  - multiple disk drives provides reliability via redundancy
- Increases the mean time to failure
- Mean time to repair – exposure time when another failure could cause data loss
- Mean time to data loss based on above factors
- If mirrored disks fail independently, consider disk with 1300,000 mean time to failure and 10 hour mean time to repair
  - Mean time to data loss is 100, 0002 / (2 ∗ 10) = 500 ∗ 106 hours, or 57,000 years!
- Frequently combined with NVRAM to improve write performance
- Several improvements in disk-use techniques involve the use of multiple disks working cooperatively

RAID (Cont.)

- Disk striping uses a group of disks as one storage unit
- RAID is arranged into six different levels
- 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 within a storage array can still fail if the array fails, so automatic replication of the data between arrays is common
- Frequently, a small number of hot-spare disks are left unallocated, automatically replacing a failed disk and having data rebuilt onto them

RAID-0 (Stripping)

- **Reliability:** 0
  There is no duplication of data. Hence, a block once lost cannot be recovered.
- **Capacity:** N*B
  The entire space is being used to store data. Since there is no duplication, N disks each having B blocks are fully utilized.



RAID-0 (Stripping)

❖**Advantages**

- It is easy to implement.
- It utilizes the storage capacity in a better way.

❖**Disadvantages**

- A single drive loss can result in the complete failure of the system.
- Not a good choice for a critical system.

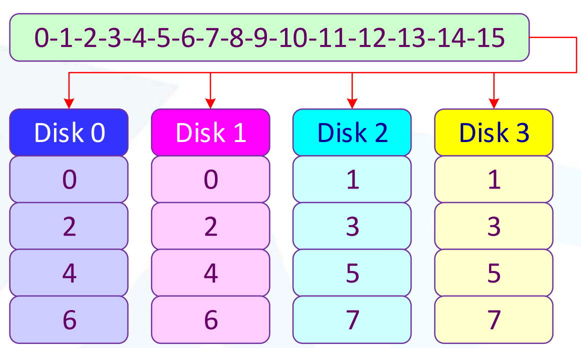
RAID-1 (Mirroring)

❖**Reliability:** 1 to N/2

1 disk failure can be handled for certain because blocks of that disk would have duplicates on some other disk. If we are lucky enough and disks 0 and 2 fail, then again this can be handled as the blocks of these disks have duplicates on disks 1 and 3. So, in the best case, N/2 disk failures can be handled.

❖**Capacity:** N*B/2

Only half the space is being used to store data. The other half is just a mirror of the already stored data.



RAID-1 (Mirroring)

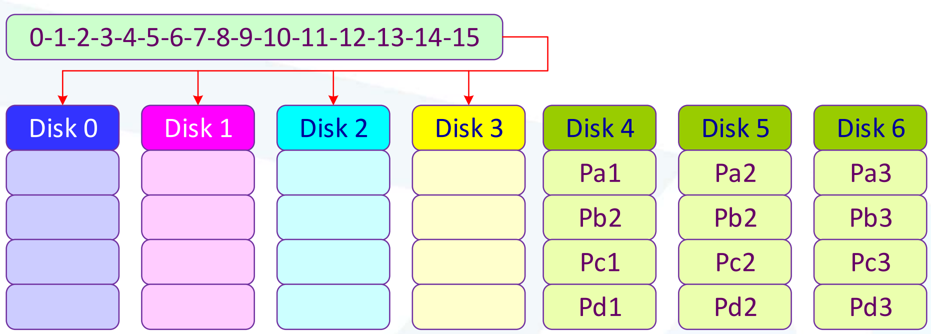
❖**Advantages**

- It covers complete redundancy.
- It can increase data security and speed.

❖**Disadvantages**

- It is highly expensive.
- Storage capacity is less.

RAID-2 (Bit-Level Stripping with Dedicated Parity)



❖Is rarely used in practice, stripes data at the bit (rather than block) level, and uses a Hamming code for error correction. The disks are synchronized by the controller to spin at the same angular orientation.

RAID-2 (Bit-Level Stripping with Dedicated Parity)

❖ **Advantages**

- ➢ Extremely high data transfer rate
- ➢ Disk failure has an insignificant impact on throughput
- ➢ Low ratio of ECC (Parity) disks to data disks, that means high efficiency

❖ **Disadvantages**

- ➢ Cannot handle simultaneous requests
- ➢ Internal error correction complexity of the Hamming Code, offers little advantage over parity.
- ➢ RAID 2 is rarely implemented for commercial purpose

Operating Systems
Dr. J.M. Khalifeh

RAID-3 (Byte-Level Stripping with Dedicated Parity)

❖ It consists of byte-level striping with dedicated parity striping.

❖ At this level, we store parity information in a disc section and write to a dedicated parity drive.

❖ Whenever failure of the drive occurs, it helps in accessing the parity drive, through which we can reconstruct the data.

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

Disk 0	Disk 1	Disk 2	Disk 3
0	1	2	P(0,1,2)
3	4	5	P(3,4,5)
6	7	8	P(6,7,8)
9	10	11	P(9,10,11)

RAID-3 (Byte-Level Stripping with Dedicated Parity)

❖**Advantages**

- ➢Data can be transferred in bulk.
- ➢Data can be accessed in parallel.

❖**Disadvantages**

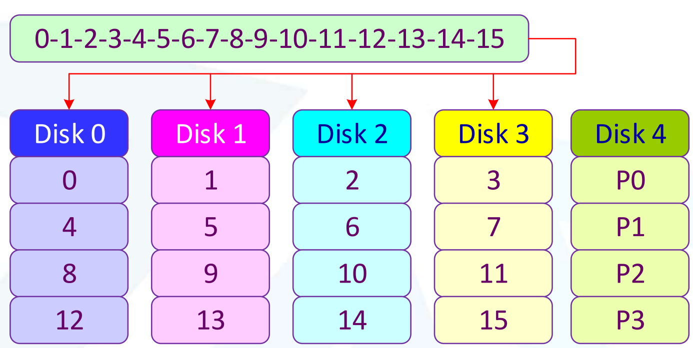
- ➢It requires an additional drive for parity.
- ➢In the case of small-size files, it performs slowly.

RAID LEVEL 4: Block Level Striping with Dedicated Parity Disk

❖consists of block-level striping with a dedicated parity disk.

❖Parity for same rank blocks is generated on Writes, recorded on the parity disk and checked on Reads.

❖RAID 4 requires at least 3 drives to implement.



RAID LEVEL 4: Block Level Striping with Dedicated Parity Disk

- **Reliability:** 1

  RAID-4 allows recovery of at most 1 disk failure (because of the way parity works). If more than one disk fails, there is no way to recover the data.

- **Capacity:** (N-1)*B

  One disk in the system is reserved for storing the parity. Hence, (N-1) disks are made available for data storage, each disk having B blocks.

Operating Systems Dr. J.M. Khalifeh

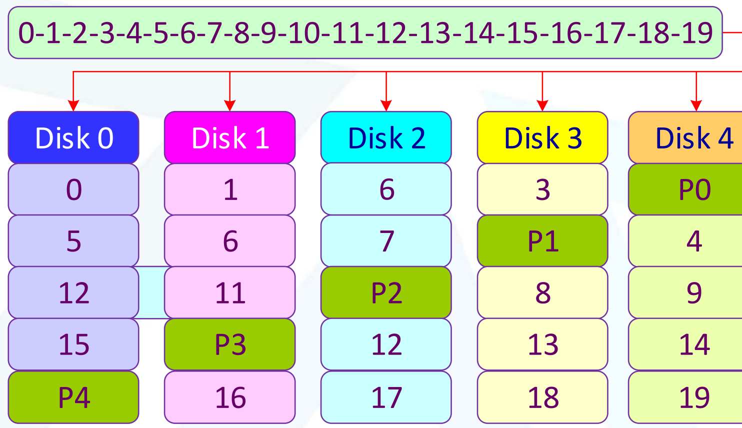
RAID LEVEL 4: Block Level Striping with Dedicated Parity Disk

- **Advantages**
  - Very high Read data transfer rate
  - High random read rate
  - The ratio of Parity disks to data disks is low, which means higher efficiency
- **Disadvantages**
  - Controller design is fairly complex
  - Worst write transaction rate
  - Data rebuild is difficult and inefficient during disk failure

Operating Systems
Dr. J.M. Khalifeh

RAID-5 (Block-Level Stripping with Distributed Parity)

- ❖consists of block-level striping with the parity information being distributed among the drives.
- ❖It requires that all drives but one be present to operate.
- ❖RAID 5 requires at least 3 drives to implement.



Operating Systems
Dr. J.M. Khalifeh

RAID-5 (Block-Level Stripping with Distributed Parity)

❖**Reliability:** 1

RAID-5 allows recovery of at most 1 disk failure (because of the way parity works). If more than one disk fails, there is no way to recover the data. This is identical to RAID-4.

❖**Capacity:** (N-1)*B

Overall, space equivalent to one disk is utilized in storing the parity. Hence, (N-1) disks are made available for data storage, each disk having B blocks.

RAID-5 (Block-Level Stripping with Distributed Parity)

❖ **Advantages**

- Highest Read data transfer rate
- Somewhat slower Write data transfer rate
- Upon failure of a single drive, subsequent reads can be calculated from the distributed parity; such that no data is lost.
- Excellent security and decent performance

❖ **Disadvantages**

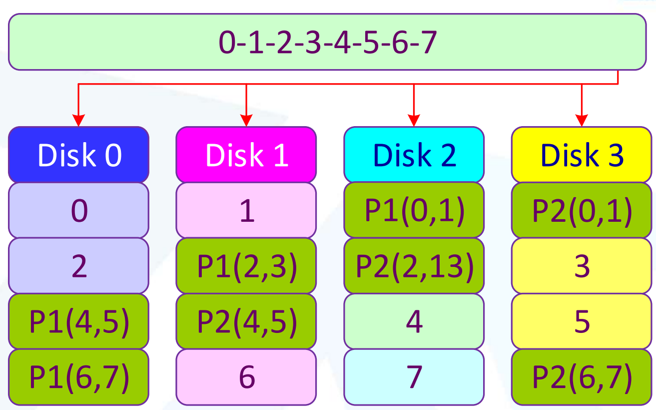
- Controller design is most complex
- Drive failures has an effect on throughput, which is still acceptable
- Rebuilding in the event of disk failure is difficult

Operating Systems
Dr. J.M. Khalifeh

RAID-6 (Block-Level Stripping with two Parity Bits)

❖ RAID 6 extends RAID 5 by using block-level striping with two parity blocks, instead of one, distributed across all member disks.

❖ RAID 6 requires at least 4 drives to implement.



RAID-6 (Block-Level Stripping with two Parity Bits)

❖**Advantages**

- ➢Very fast Read data transfer rate
- ➢Extremely high data fault tolerance
- ➢Can sustain two drive failures because of the two parity checks, data is still accessible.

❖**Disadvantages**

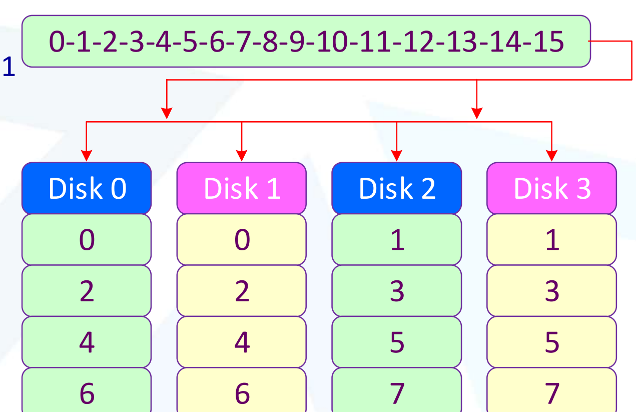
- ➢Transaction rate equal to a single disk drive at best (if spindles are synchronized)
- ➢Parity calculation overhead slows down the Write data transaction rate
- ➢Requires n+2 drives to implement because of dual parity scheme

RAID level 10 – combining RAID 1 & RAID 0

❖It is possible to combine the advantages (and disadvantages) of RAID 0 and RAID 1 in one single system.

❖This is a nested or hybrid RAID configuration.

❖It provides security by mirroring all data on secondary drives while using striping across each set of drives to speed up data transfers.



RAID level 01 – combining RAID 0 & RAID 1

- It is possible to combine the advantages (and disadvantages) of RAID 0 and RAID 1 in one single system.
- This is a nested or hybrid RAID configuration.
- It provides security by mirroring all data on secondary drives while using striping across each set of drives to speed up data transfers.