<u>UNIT-2</u>

STORAGE SYSTEMS ARCHITECTURE -

① <u>Intelligent disk subsystems overview</u> -

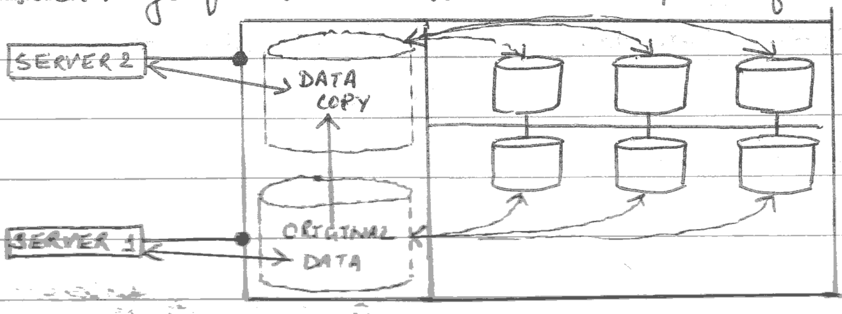
A large disk ~~sub set~~ subsystems that can store between a few hundred gigabytes and several ten petabytes of data, depending upon size and have functions such as high availability, high performance, instant copies and remote mirroring are available at a reasonable price. This disk subsystem is also known as Intelligent disk subsystems.

Intelligent disk subsystems represent the third level of complexity for controllers after JBODs and RAID arrays.

Some of the functions of Intelligent disk subsystems are -

(1) <u>Instant Copies</u> -

It can virtually copy data sets of several terabytes within a disk subsystem in a few seconds. <u>Virtual copying</u> means that disk ~~sys~~ subsystems fool the attached servers into believing that they are capable of copying such large quantities in such a short space of time.



<u>Space-efficient instant copy</u> only copies the blocks that were changed. These normally require considerably less storage space than the entire copy.

<u>Incremental instant copy</u> copies the data entirely for the first time then only those changes since the previous instant copy are copied.

<u>Reversal of instant copy</u> is used when the failure occurs by copy back the data on the productive hard disks per a second instant copy.

(2) <u>Remote mirroring</u> -

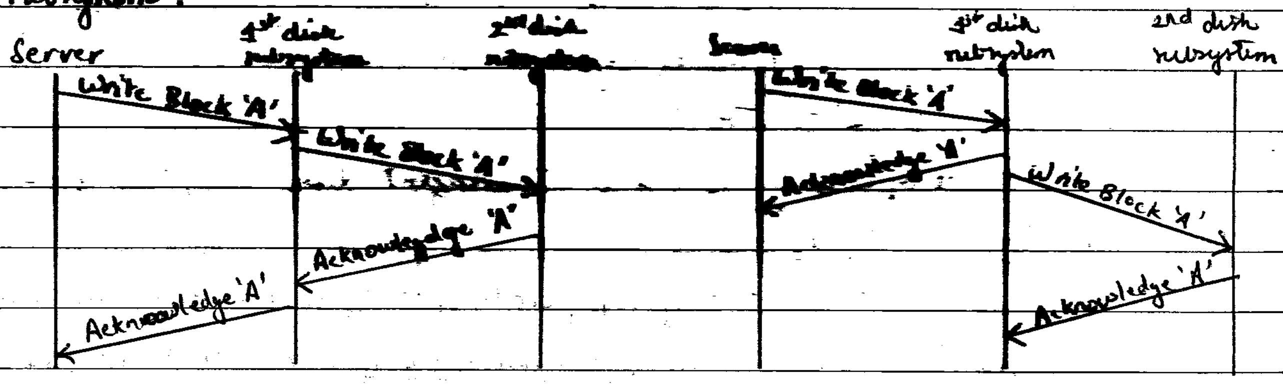
It offers protection against catastrophes by mirroring the data, or part of the data, independently onto a second sub system.

It is invisible to application servers and does not consume resources.

Two types of remote mirroring are <u>synchronous</u> and <u>asynchronous</u> remote mirroring.

In synchronous remote mirroring the first disk subsystem sends the data to the second disk subsystem first before it acknowledges a server's write command.

In asynchronous remote mirroring acknowledges a write command immediately, only then does it send the copy of the block to the second disk subsystem.



SYNCHRONOUS | ASYNCHRONOUS

Synchronous

Advantage -
Copy of data is always up-to-date

Disadvantage -
Increases the response time of 1st disk subsystem to the server.

Asynchronous

Advantage -
Rapid response time

Disadvantage -
Copy of data may not be up-to date

**** Rapid response time with mirroring over long distances can be achieved using the combination of synchronous and asynchronous remote mirroring.

(3) Consistency group -
Combine multiple instant copy pairs into one unit or combining multiple remote mirroring pairs ~~froms~~ forms a consistency group which the consistency of the data.

(4) LUN masking - (Logical Unit Number)
It limits the access to the hard disks that the disk subsystem exports to the connected server.
All hard disks (physical and virtual) that are visible outside the disk subsystem are known as LUN

Without LUN masking, a configuration error on one server can destroy the data of another server.

Two types of LUN masking are -

(1) Port-based LUN masking is the 'poor man's LUN masking in which all servers connected to the disk subsystem via the same port see the same disks.

(2) Server-based LUN masking offers more flexibility in which every server sees only the hard disks assigned to it.

(2) Contrast of Integrated Vs. Modular Array -

Integrated (Monolithic) array } -

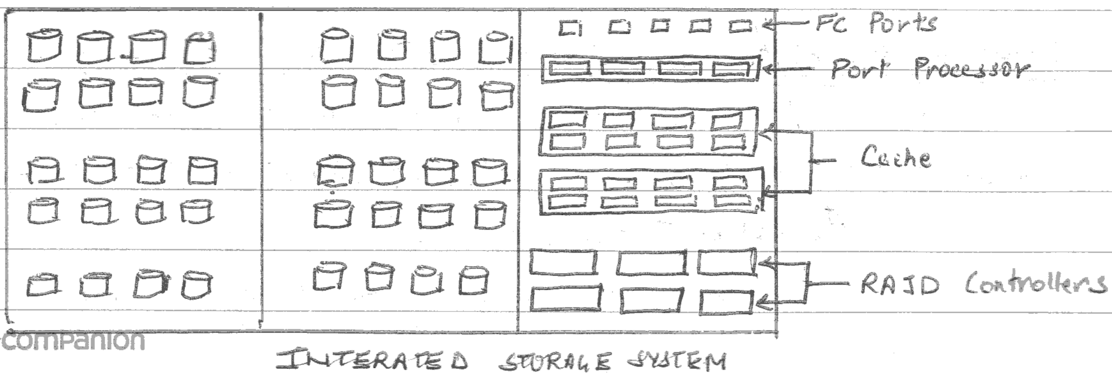
It is also known as enterprise arrays or cache centric arrays.

Integrated storage systems are generally aimed at the enterprise level, centralizing data in a powerful system with hundred of ~~two~~ drives.

This system is contained within a single frame or interconnected frames (for expansion) and scale to support increases in connectivity, performance and capacity as required.

Characteristics -

(1) Large storage capacity
(2) Large amount of cache to temporarily store I/Os before writing to disk
(3) Redundant components for improved data protection and availability
(4) Many built in features to make them more robust and fault tolerant
(5) Usually connect to mainframes or very powerful open systems hosts.
(6) Multiple front end ports to provide connectivity to multiple servers.
(7) Multiple back end Fibre Channel or SCSI RAID controllers to manage disk processing
(8) Expensive (applicable to only most mission critical applications)



INTERATED STORAGE SYSTEM

Modular array –

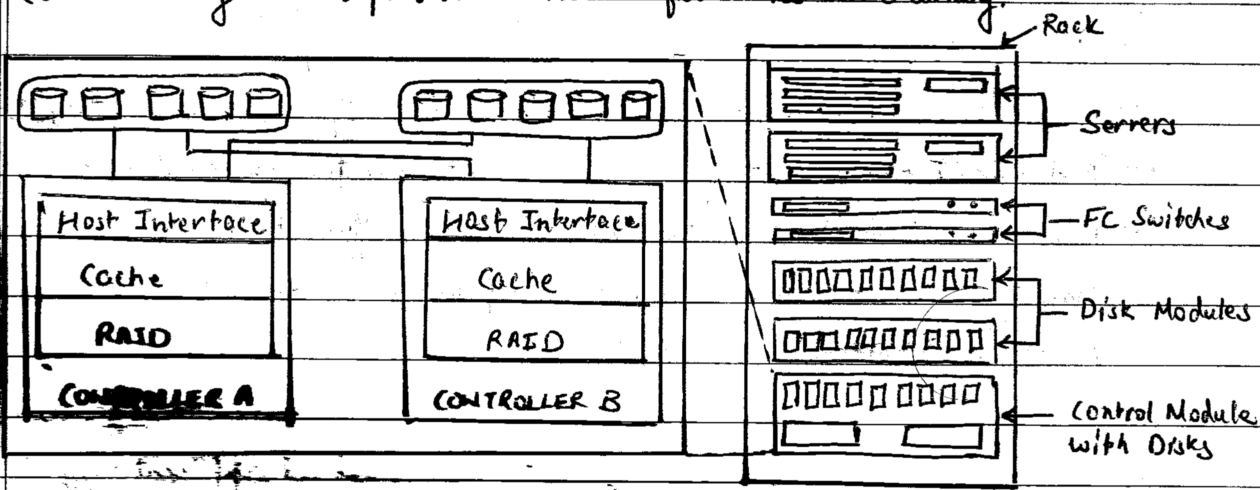
Modular storage systems provide storage to a smaller number of Windows or Unix servers than larger integrated storage systems.

Modular storage systems are typically designed with two controllers each of which contains host interfaces, cache, RAID processors, and disk drive interfaces.

Sometimes called as midrange or departmental storage systems.

Characteristics –

(1) Small companies / department level.
(2) Smaller disk capacity and less global cache
(3) Takes up less floor space and cost less
(4) Can start with a smaller number of disks and scale as needed.
(5) Fewer front end ports for connection to servers.
(6) Performance can degrade as capacity increases
(7) Cannot connect to mainframes
(8) Limited redundancy and connectivity
(9) Usually have separate controllers from the disk array.



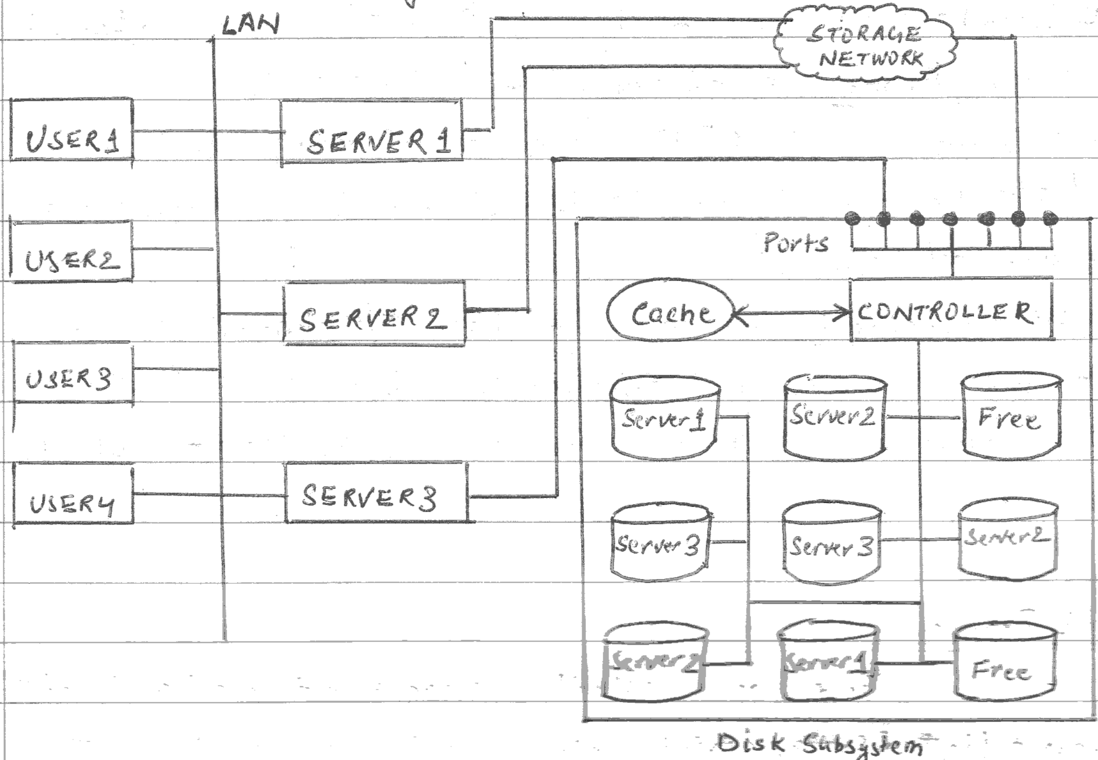
(3) Component architecture of Intelligent Disk Subsystems –

Servers are connected to the disk subsystems via the ports using standard I/O techniques such as SCSI (Small Computer System Interface), Fibre Channel SAN or Internet SCSI (iSCSI).

Internally, the disk subsystem consists of hard disks, a controller, a ca

ATA → Advance Technology Attachment

and ~~internall~~ internally I/O channels.



COMPONENT ARCHITECTURE OF INTELLIGENT DISK SUBSYSTEMS

Hard disks - For most applications, medium-size harddisks are sufficient. Only for appliations with higher performance, smaller hard disks be considered.

Controller - In most disk subsystems, there is a controller between the connection ports and hard disks. The controller can significantly increase the data availability and data access performance.

Cache - It is used in an attempt to accelerate read and write accesses to the server.

Internally I/O Channels - Standard I/O techniques such as SCSI, Fibre Channel, increasingly Serial ATA (SATA), Serial Attached SCSI (SAS) and Serial storage Architecture (SSA) are being used for internal I/O channels between connection ports and controller as well as between controller and internal hard disks.

Four types of cabling are done-

(1) Active - Individual physical hard disk are connected via only ~~only~~ one I/O channel.

mycompanion

(2) Active/Passive - Individual hard disks are connected via ~~a~~ two I/O channels. ~~A~~ Second I/O channel is used only when first I/O channel is failed.

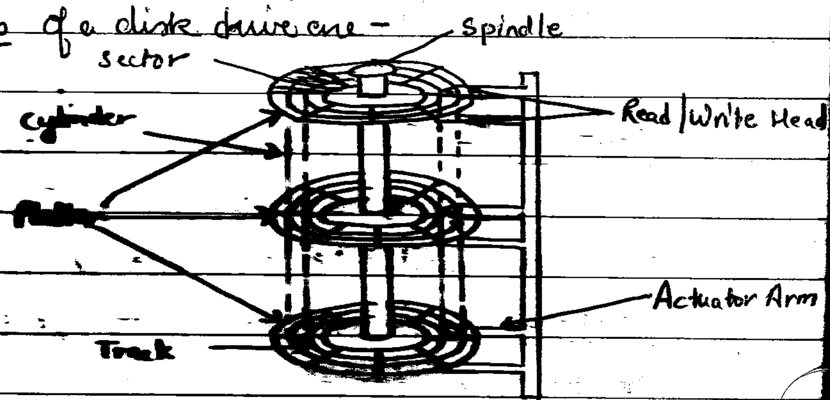
(3) Active/~~Passive~~ Active (no load sharing) - Both I/O channels are used. First group is addressed via 1st I/O channel and second group is addressed via 2nd I/O channel. If one I/O channel fails, both groups are addressed via the other I/O channel.

(4) Active/Active (load sharing) — Both I/O channels are used. The controller divides the load dynamically between the two I/O channels. If one I/O channel fails, then the communication goes through the other channel only.

(4) Disk Physical Structure -

A disk drive uses a rapidly moving arm to read and write data across a flat platter coated with magnetic particles. Data is transferred from the magnetic platter through the R/W head ~~and controller~~ to the computer. Several platters are assembled together with the R/W head and controller. Key components of a disk drive are -

(1) Platter
(2) Spindle
(3) Read/Write Head
(4) Actuator arm assembly
(5) ~~Arm~~ Controller



→ Properties & Specifications -

(1) Platter - The data is recorded on these platters in binary codes (0s and 1s). The set of rotating platters is sealed in a case, called a Head Disk Assembly (HDA).

A platter is a rigid, round disk coated with magnetic material on both surfaces (top and bottom). The data is encoded by polarizing the magnetic area, or domains, on the disk surface. Data can be written to or read from both the surfaces of the platter.

Number of platter and its capacity gives total capacity of the drive.

(2) Spindle – It connects all the platters and is connected to a motor. The motor of the spindle rotates with a constant speed.

The disk platter spins at a speed of several thousands of revolutions per minute (rpm).

(3) Read/Write Heads – It read and write data from or to a platter. Drives have two read/write heads per platter, one for each surface of the platter.

R/W head changes the magnetic polarization on the surface of the platter when writing data. ~~It det~~ R/W head detects the magnetic polarization on the surface of the platter when reading data.

Head flying height – Air gap between R/W head and the platter.

Landing zone – Resting zone of R/W head when spindle stops.

Head crash – When R/W head accidentally touches the surface results in data loss.

(4) Actuator Arm Assembly – The R/W head are mounted on it which positions the R/W head at the location on the platter where the data needs to be written or read.

The R/W heads for all platters on a drive are attached to one actuator arm assembly and move across the platters simultaneously.

(5) Controller – It is a printed circuit board, mounted at the bottom of a disk drive. It consists of a microprocessor, internal memory, circuitry and firmware.

The firmware controls power and speed of the spindle motor, manages communication between the drive and the host, controls the R/W operations by moving the actuator arm and switching between different R/W heads. The firmware performs the optimization of data access.

Data on the disk are recorded on tracks, which are concentric rings on the platter. ~~on~~ The track is numbered, starting from zero, from the outer edge of the platter. The number of track per inch (TPI) on ~~the~~ platter (or the track density) measures how tightly the tracks are packed on a platter.

Each track is divided into smaller units called <u>sectors</u>. A sector is the smallest, individually addressable unit of storage.

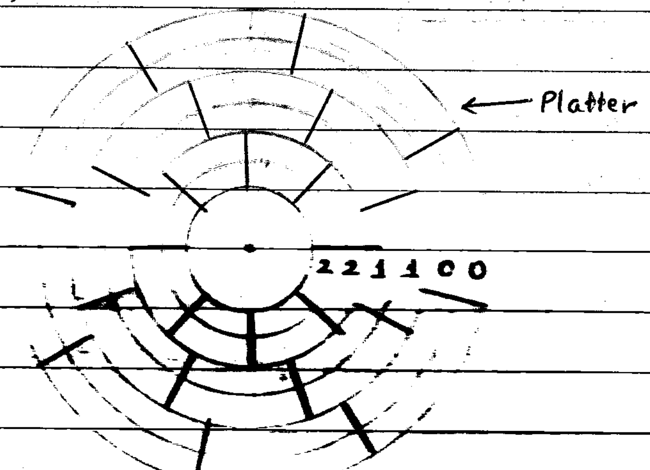
A <u>cylinder</u> is the set of identical tracks on both surfaces of each drive platter. The location of drive heads is referred to by cylinder number, not by track number.

<u>Zone Bit Recording</u> -

It utilizes the disk efficiency. This mechanism groups tracks into zone based on their distance from the centre of the disk.

The zone are numbered, with the outermost zone being zone 0.

An appropriate number of sectors per track are assigned to each zone, so a zone near the centre of the platter has fewer sectors per track than a zone on the outer edge. However, tracks within a particular zone have the same number of sectors.



<u>Logical Block Addressing (LBA)</u> -

It simplifies addressing by a linear address to access physical blocks of data. The disk controller translates LBA to a CHS (Cylinder, Head and Sector) address, and the host only needs to know the size of the disk drive in terms of the number of blocks.

→ <u>Disk Drive Performance</u> -

<u>Disk service time</u> is the time taken by a disk to complete and I/O request. Components that contribute to service time on the disk drive are -

Date ___/___/___

Page ________

Seek time (access time) -

It describes the time taken to position the R/W heads across the platters with a radial movement.

Disk vendors publish the following seek time specifications -

(1) Full Stroke - The time taken by R/W head to move across the entire width of the disk, from the innermost track to the outermost track.

(2) Average - Average time taken by R/W head to move from one random track to another, normally listed as the time for one-~~tow~~third of a full stroke.

(3) Track-to-Track - Time taken by R/W head to move between adjacent tracks.

Each of these specification are measured in milliseconds.

Rotational latency -

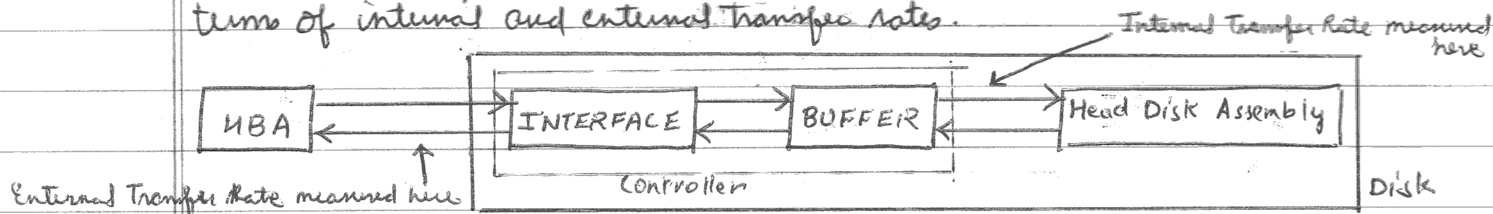
The time taken by the platter to rotate and position the data under the R/W head is called rotational latency.

Average rotational latency is one-half of the time taken for a full rotation.

Data Transfer Rate -

It refers to the amount of data per unit time that the drive can deliver to the ~~HBA~~ ~~HDA (Head disk assembly)~~ HBA (Host Bus Adapter)

The data transfer rates during the R/W operations are measured in terms of internal and external transfer rates.



(5) RAID levels and parity algorithms -

RAID levels are defined on the basis of striping, mirroring and parity techniques. These techniques determine the data availability and performance characteristics of an array.

→ ~~mirroring~~ Striping - A RAID set is a group of disks. Within each disk, a predefined number of contiguously addressable disk blocks are defined as strips. The set of aligned strips that span across all the disk within RAID set is called a stripe.

Stripe size (Stripe depth) → Number of ~~blocks~~ blocks in a strip.

Stripe width → Number of data strips in a stripe.

→ Mirroring – It is a technique whereby data is stored on two different HDDs, yielding two copies of data.

→ Parity – It is the method of protecting striped data from HDD failure without the cost of mirroring. An additional HDD is added to the stripe width to hold parity, a mathematical construct that allows re-creation of the missing data.

Parity Algorithm – Parity calculation is done using a bitwise XOR operation.

Calculation of parity is a function of the RAID controller.

Parity is recalculated everytime there is a change in data.

→ ~~RAID (0)~~ RAID 0 (Striped array with no fault tolerance) –

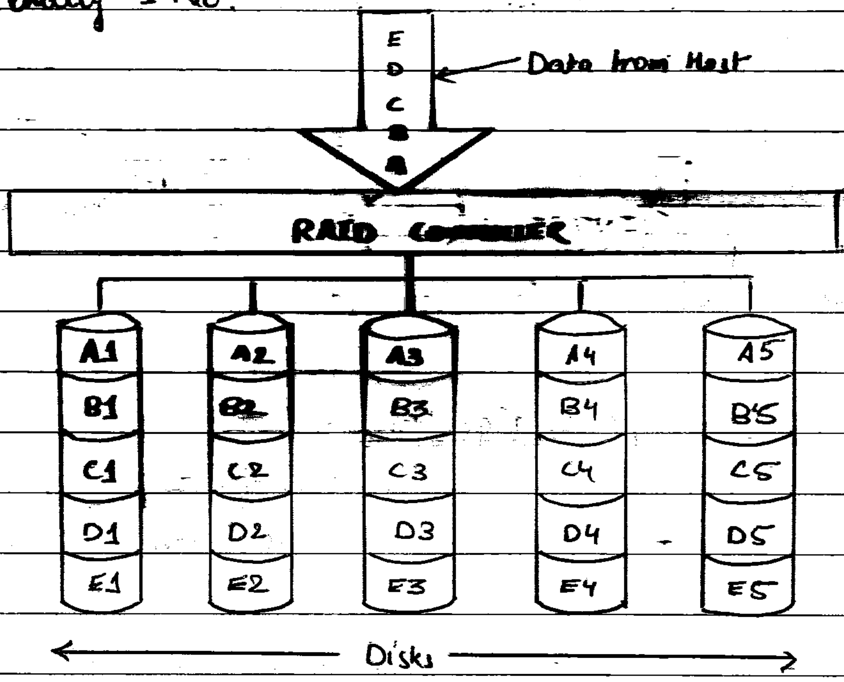
Minimum disk = 2 cost = low

Storage Efficiency = 100%

Read Performance → very good for both random and sequential read

Write Performance → very good

Write Penalty = No.



Date ___/___/___

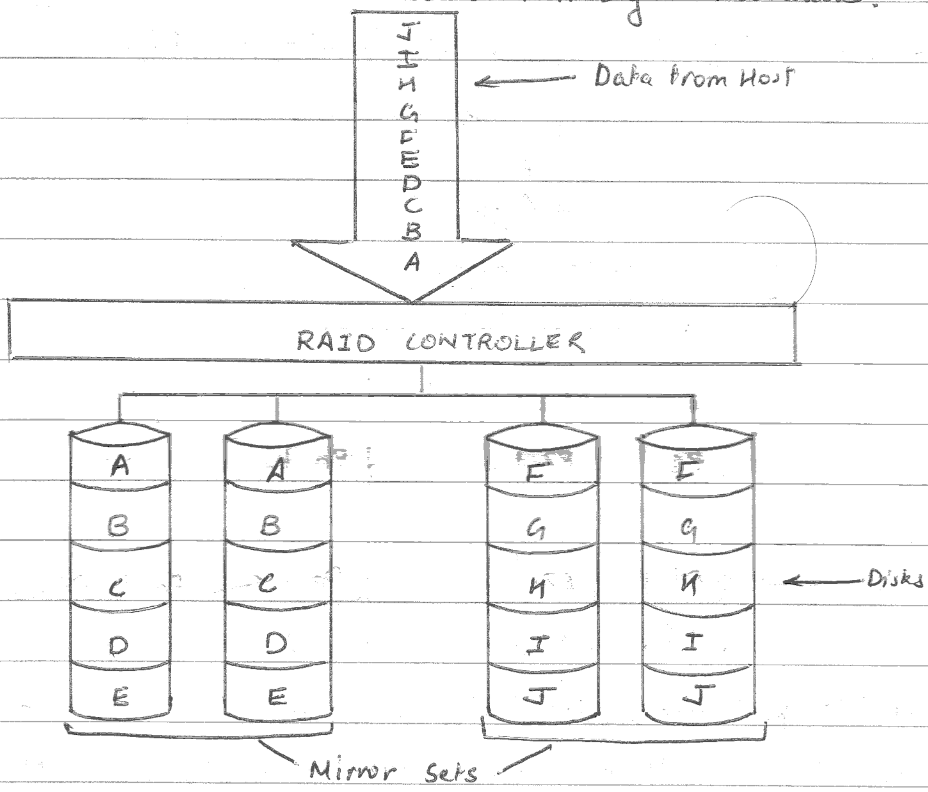
Page ________

→ RAID 1 (Disk mirroring) -

Minimum disk = 2, Storage efficiency = 50%, Cost = High

Read Performance → Good. Better than a single disk

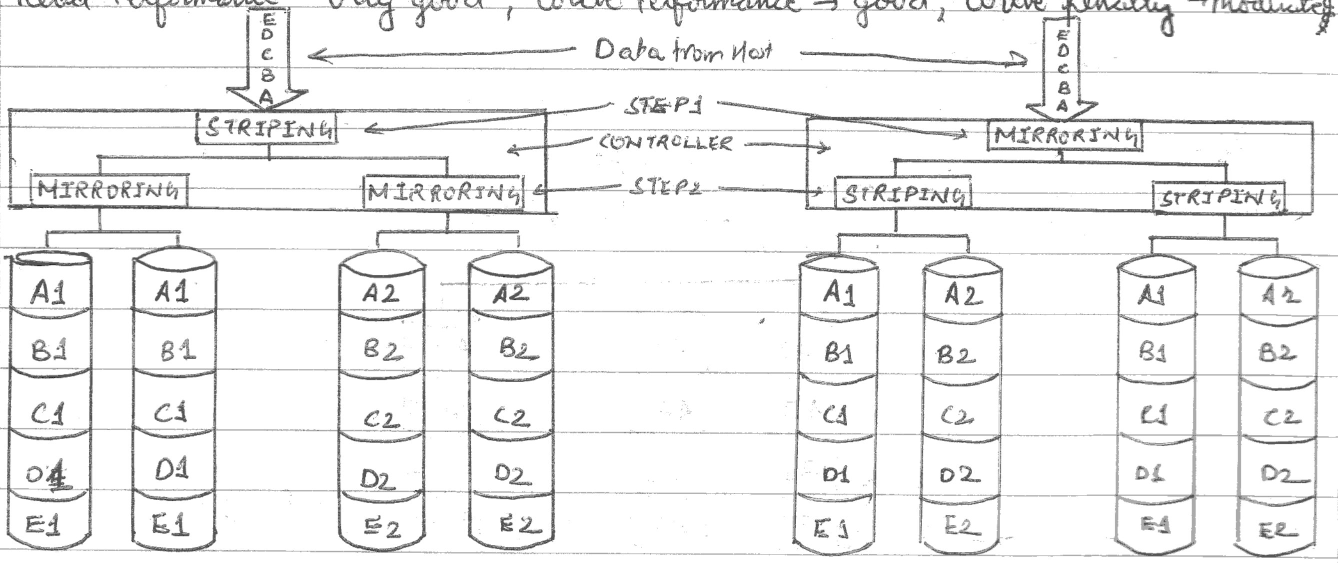
Write Performance → Good. Slower than a single disk, as every write must be committed to all disks. Write Penalty → Moderate.



→ Nested RAID (Combination of Raid levels - Raid 1 + Raid 0 or Raid 0 + Raid 1)

Minimum disk = 4, Storage efficiency = 50%, Cost = High

Read Performance → Very good, Write Performance → good, Write penalty → moderate



RAID 0+1 RAID 1+0

Date __/__/__

Page ______

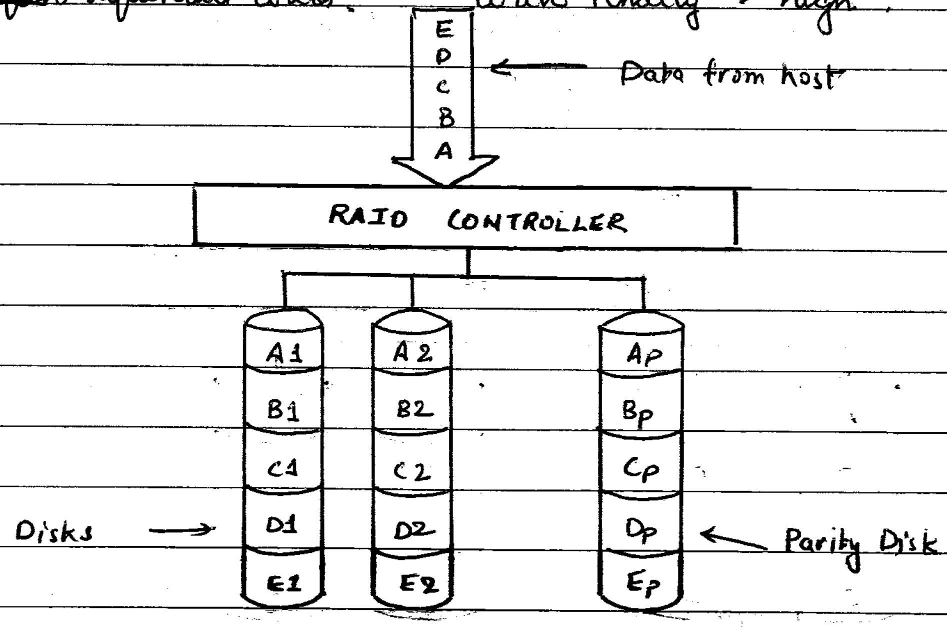
RAID 3 (Parallel access array with dedicated parity disk) –

Minimum disks = 3, Storage Efficiency = $((n-1) * 100)/n$, $n \to$ No. of disks

Cost = Moderate

Read Performance → Good for random reads, and very good for sequential reads

Write Performance → Poor to fair for small random writes. Good for large, ~~sequen~~ sequential writes. Write Penalty → High.



RAID 4 (Striped array with independent disks and a dedicated parity disk) –

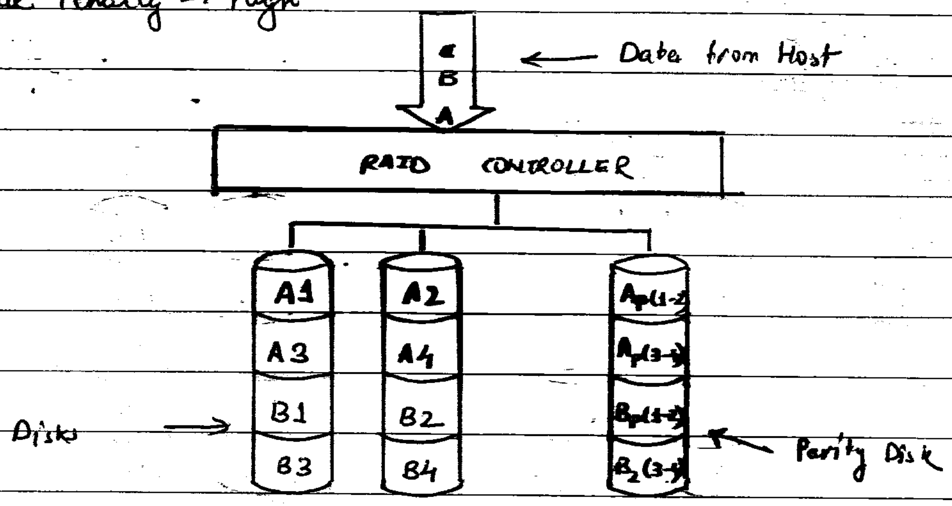
Minimum disks = 3, storage efficiency = $((n-1) * 100)/n$, $n \to$ No. of disks

Cost = Moderate.

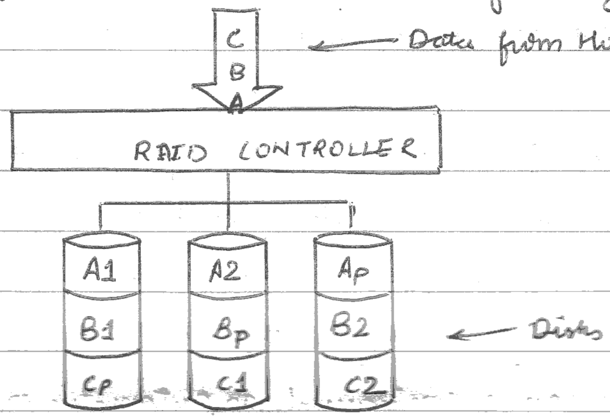
Read Performance → Very good for random reads. Good to very good for sequential ~~writes~~ reads

Write Performance → Poor to fair for random ~~reads~~ writes. Fair to good for sequential writes.

Write Penalty → High

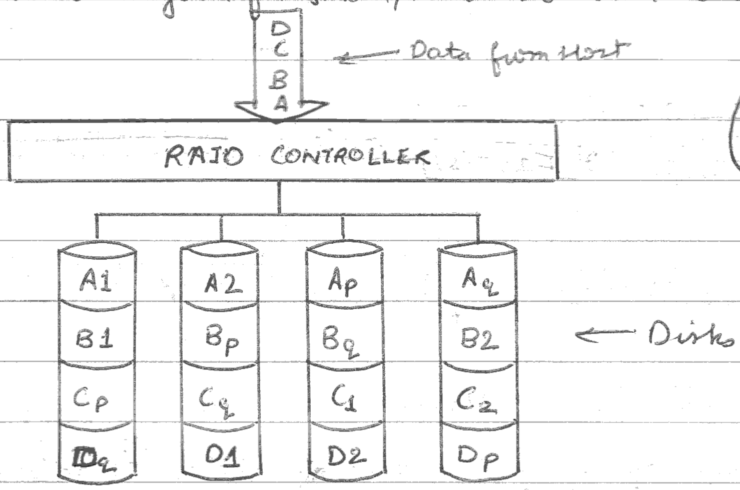


RAID 5 (Striped array with independent disks and distributed parity) –
Minimum Disks = 3, Storage Efficiency = $((n-1)*100)/n$, $n \rightarrow$ No. of disks
Cost = Moderate.
Read Performance → Very good for random reads. Good for sequential reads
Write Performance → Fair for random writes. Slower due to parity overhead. Fair to good for sequential writes. Write Penalty → High.



(It is useful when one disk is failed)

RAID 6 (Striped array with independent disks and dual distributed parity) –
Minimum disk = 4, Storage Efficiency = $((n-2)*100)/n$, $n \rightarrow$ No. of disks
Cost = Moderate but more than RAID 5.
Read Performance → Very good for random reads. Good for sequential reads.
Write Performance → Good for small, random writes. Write Penalty → Very high



(It is useful when two disks are failed)

(6) ~~Hot Sp~~ **Hot Spares** –
It refers to a spare HDD in a RAID array that temporarily replaces a failed HDD of a RAID set. A hot spare takes the identity of the failed HDD in the array.

~~LVM → Logical Volume Management~~

Date ___/___/___

Page ______

One of the following methods of data recovery is performed depending ~~on~~ the RAID implementation –

(1) If parity RAID is used, then the data is rebuilt onto the hot spare from the parity and the data on the surviving HDDs in the RAID set.

(2) If mirroring is used, then the data from the surviving mirror is used to copy the data.

When the failed HDD is replaced with a new HDD, one of the following takes place –

(1) The hot spare replaces the new HDD permanently. This means that it is no longer a hot spare, and a new hot spare must be configured on the array.

(2) When a new HDD is added to the system, data from the hot spare is copied to it. The hot spare returns to its idle state, ready to replace the next failed drive.

A hot spare can be configured as <u>automatic</u> or <u>user-initiated</u>.

(7) <u>Front end to host storage provisioning, mapping and operation</u> –

Front end provides the interface between the storage system and the host.

Storage provisioning is the process of assigning storage, usually the form of server disk drive space, ~~or~~

<u>Mapping of host file to storage</u> –

HOST → (configures/manages) → FILE(S) → (reside in) → FILE SYSTEM FILES → (mapped by a file system to) → FILE SYSTEM BLOCKS → (residing on) → LVM logical Extents → (Mapped by LVM to) → DISK PHYSICAL EXTENTS → (consisting of) → DISK SECTORS → Mapped by disk storage subsystem

<u>Storage operations</u> are done by either by operating system or by the LVM.

<u>LVM</u> (Logical Volume Management) –

It is a system for managing logical volumes, or file systems, that is much more advanced and flexible than the traditional method of partitioning a disk into one or more segments and formatting that partition with a filesystem