Module-1:

STORAGE SYSTEMS

1.1 Information Storage

- Companies use data to derive information that is critical to their day-to-day operations.
- Storage is a repository that is used to store and retrieve the digital-data.

1.1.1 Data

- Data is a collection of raw facts from which conclusions may be drawn.
- Example:
  - → Handwritten-letters
  - → Printed book
  - → Photograph
  - → Movie on video-tape
- The data can be generated using a computer and stored in strings of 0s and 1s (Figure 1-1).
- Data in 0s/1s form is called digital-data.
- Digital-data is accessible by the user only after it is processed by a computer.

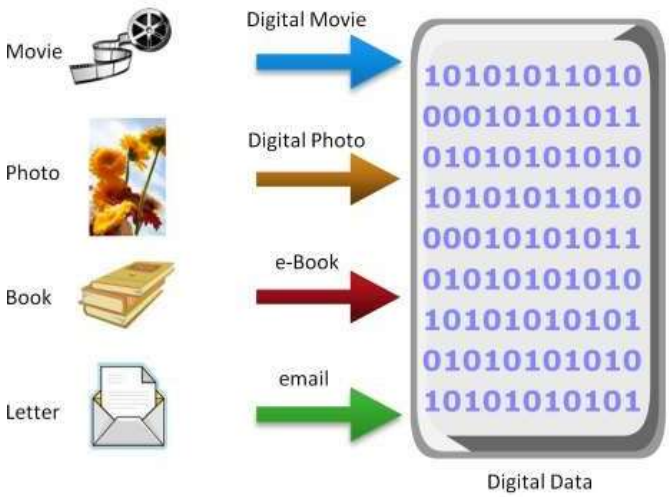

Figure 1-1: Digital data

- The factors contributing to the growth of digital-data are:

1) Increase in Data Processing Capabilities
- Modern computers provide a significant increase in data-processing capabilities.
- This allows conversion of various types of data (like book, photo or video) into digital-formats.

2) Lower Cost of Digital Storage
- With the advancement in technology, the cost of storage-devices have decreased.
- This cost-benefit has increased the rate at which data is being generated and stored.

3) Affordable and Faster Communication Technology
- Nowadays, rate of sharing digital-data is much faster than traditional approaches (e.g. postal)
- For example,
  - i) A handwritten-letter may take a week to reach its destination.
  - ii) On the other hand, an email message may take a few seconds to reach its destination.

4) Increase of Smart Devices and Applications
- Smartphones, tablets and smart applications have contributed to the generation of digital-content

1.1.2 Types of Data

- Data can be classified as structured or unstructured based on how it is stored & managed(Figure 1-2)
- Structured data is organized in table format.
  Therefore, applications can query and retrieve the data efficiently.
- Structured data is stored using a DBMS. (Table contains rows and columns).
- Unstructured data cannot be organized in table format.
  Therefore, applications find it difficult to query and retrieve the data.
- For example, customer contacts may be stored in various forms such as
  → Sticky notes
  → email messages
  → Business cards

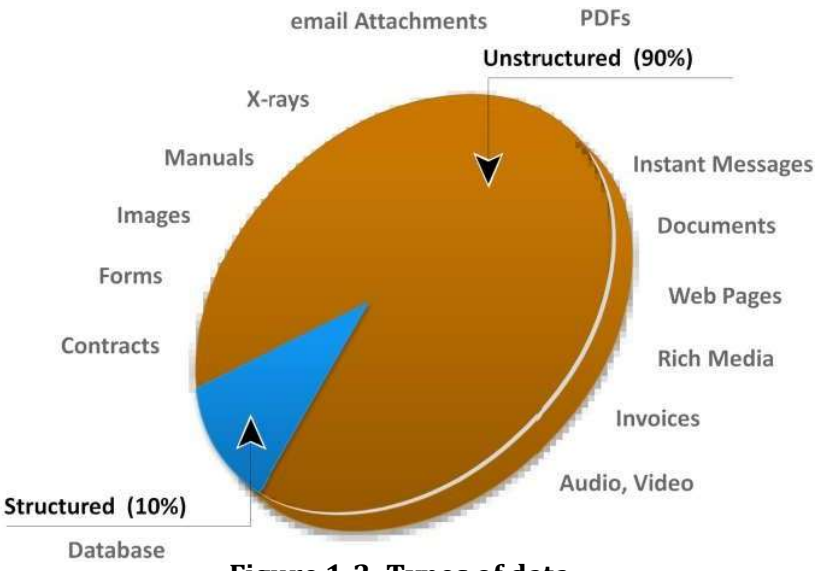

Figure 1-2: Types of data

1.1.3 Big Data

- It refers to data-sets whose sizes are beyond the capability of commonly used software tools.
- The software tools are used to store, manage, and process data within acceptable time limits.
- Big-data includes both structured- and unstructured-data.
- The data is generated by different sources such as
  → business application
  → web pages
  → videos
  → images
  → e-mails
  → social media
- These data-sets require real-time capture or updates for
  → analysis
  → predictive modeling and
  → decision making.
- Significant opportunities exist to extract value from big data.

1) Devices that collect data from multiple locations and also generate new data about this data.
2) Data collectors who gather data from devices and users.
3) Data-aggregators that compile the collected data to extract meaningful information.
4) Data users & buyers who benefit from info collected & aggregated by others in the data value chain.

1.1.3.1 Data Science

- Data Science is a discipline which enables companies to derive business-value from big-data.
- Data Science represents the synthesis of various existing disciplines such as
  - → statistics
  - → math
  - → data visualization and
  - → computer science.
- Several industries and markets currently looking to employ data science techniques include
  - → scientific research
  - → healthcare
  - → public administration
  - → fraud detection
  - → social media
  - → banks
- The storage architecture should
  - → be simple, efficient, and inexpensive to manage.
  - → provide access to multiple platforms and data sources simultaneously.

1.1.4 Information

- Information vs. Data:
  - i) Information is the intelligence and knowledge derived from data.
  - ii) Data does not fulfill any purpose for companies unless it is presented in a meaningful form.
- Companies need to analyze data for it to be of value.
- Effective data analysis extends its benefits to existing companies.
  Also, effective data analysis creates the potential for new business opportunities.
- Example: Job portal.
  - Job seekers post their resumes on the websites like Naukiri.com, LinkedIn.com, Shine.com
  - These websites collect & post resumes on centrally accessible locations for prospective employers
  - In addition, employers post available positions on these websites
  - Job-matching software matches keywords from resumes to keywords in job postings.
  - In this way, the search engine uses data & turns it into information for employers & job seekers.

1.1.5 Storage

- Data created by companies must be stored so that it is easily accessible for further processing.
- In a computing-environment, devices used for storing data are called as **storage-devices**.
- Example:
  - → Memory in a cell phone or digital camera
  - → DVDs, CD-ROMs and hard-disks in computers.

1.2 Evolution of Storage Architecture

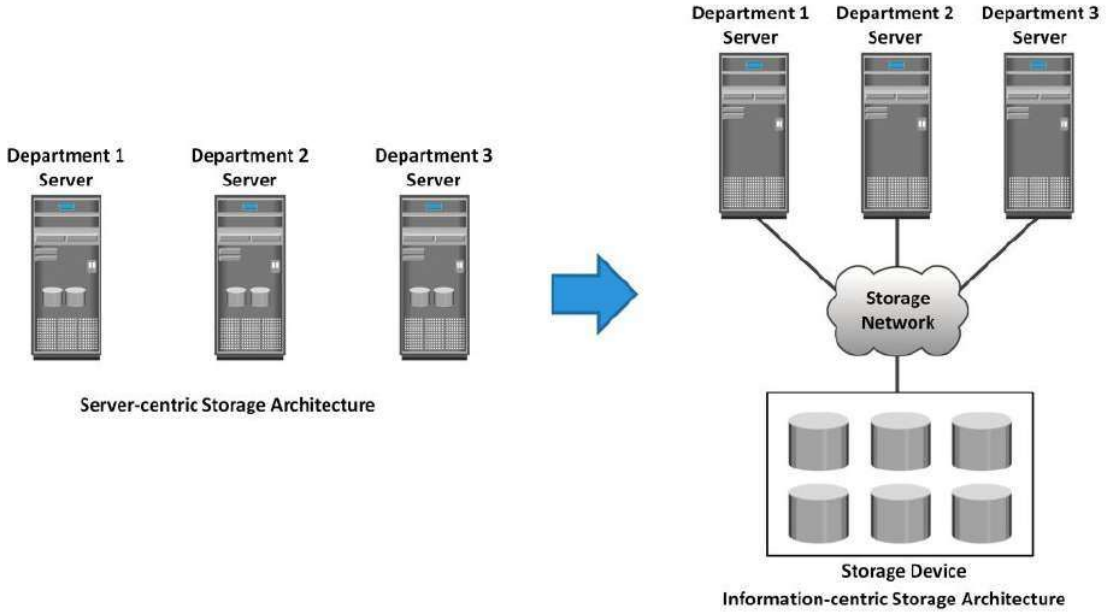

Figure 1-3: Evolution of storage architectures

1.2.1 Server Centric Storage Architecture

- In earlier days, companies had a data-center consisting of
  1) Centralized computers (mainframes) and
  2) Information storage-devices (such as tape reels and disk packs)
- Each department had their own servers and storage because of following reasons (Figure 1-3):
  → evolution of open-systems
  → affordability of open-systems and
  → easy deployment of open-systems.
- Disadvantages:
  1) The storage was internal to the server.
     Hence, the storage cannot be shared with any other servers.
  2) Each server had a limited storage-capacity.
  3) Any administrative tasks resulted in unavailability of information.
     The administrative tasks can be maintenance of the server or increasing storage-capacity
  4) The creation of departmental servers resulted in '
     → unprotected, unmanaged, fragmented islands of information and
     → increased capital and operating expenses.
- To overcome these challenges, storage evolved from
  → server-centric architecture to information-centric architecture.

1.2.2 Information Centric Architecture

- Storage is managed centrally and independent of servers.
- Storage is allocated to the servers "on-demand" from a shared-pool.
- A **shared-pool** refers to a group of disks.
- The shared-pool is used by multiple servers.
- When a new server is deployed, storage-capacity is assigned from the shared-pool.
- The capacity of shared-pool can be increased dynamically by
  → adding more disks without interrupting normal-operations.
- Advantages:

1) Information management is easier and cost-effective.
2) Storage technology even today continues to evolve.

1.3 Data Center Infrastructure
- Data-center provides centralized data processing capabilities to companies.

1.3.1 Core Elements of a Data Center
- Five core-elements of a data-center:

1) Application
- An application is a program that provides the logic for computing-operations.
- For example: Order-processing-application.
  Here, an Order-processing-application can be placed on a database.
  Then, the database can use OS-services to perform R/W-operations on storage.

2) Database
- DBMS is a structured way to store data in logically organized tables that are interrelated.
- Advantages:
  1) Helps to optimize the storage and retrieval of data.
  2) Controls the creation, maintenance and use of a database.

3) Server and OS
- A computing-platform that runs 1) applications and 2) databases.

4) Network
- A data-path that facilitates communication
  1) between clients and servers or
  2) between servers and storage.

5) Storage Array
- A device that stores data permanently for future-use.

Example: Figure 1-4 shows an Order-processing application
Step 1: A customer places an order through the AUI on the client-computer.
Step 2: The client accesses the DBMS located on the server to provide order-related information. (Order-related information includes customer-name, address, payment-method & product-ordered).
Step 3: The DBMS uses the server to write this data to the disks in the storage.
Step 4: The storage-network
→ provides the communication-link between server and storage and
→ transports the write-command from server to storage
Step 5: After receiving the write-command, the storage saves the data on disks.

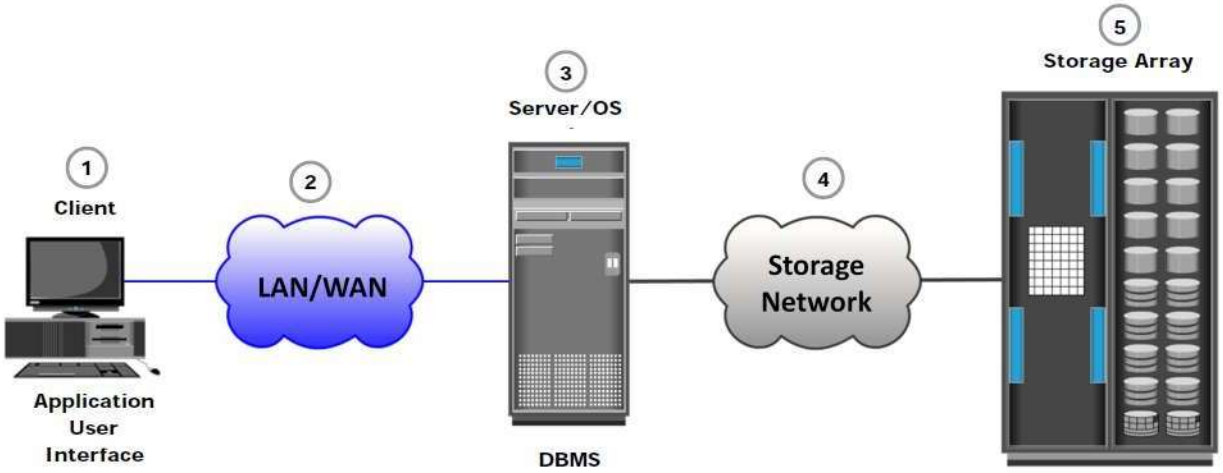

Figure 1-4: Example of an order processing-application

1.3.2 Key Characteristics for Data Center

Figure 1-5: Key characteristics of data center elements

1) Availability

- In data-center, all core-elements must be designed to ensure availability (Figure 1-5).
- If the users cannot access the data in time, then it will have negative impact on the company.
  (For example, if amazon server goes down for even 5 min, it incurs huge loss in millions).

2) Security

- To prevent unauthorized-access to data,
  1) Good polices & procedures must be used.
  2) Proper integration of core-elements must be established.
- Security-mechanisms must enable servers to access only their allocated-resources on the storage.

3) Scalability

- It must be possible to allocate additional resources on-demand w/o interrupting normal-operations.
- The additional resources includes CPU-power and storage.
- Business growth often requires deploying
  → more servers
  → new applications and
  → additional databases.
- The storage-solution should be able to grow with the company.

4) Performance

- All core-elements must be able to
  → provide optimal-performance and
  → service all processing-requests at high speed.
- The data-center must be able to support performance-requirements.

5) Data Integrity

- Data integrity ensures that data is written to disk exactly as it was received.
- For example: Parity-bit or ECC (error correction code).
- Data-corruption may affect the operations of the company.

6) Storage Capacity

- The data-center must have sufficient resources to store and process large amount of data efficiently.
- When capacity-requirement increases, the data-center must be able
  → to provide additional capacity without interrupting normal-operations.
- Capacity must be managed by reallocation of existing-resources rather than by adding new resources

7) Manageability

- A data-center must perform all operations and activities in the most efficient manner.
- Manageability is achieved through automation i.e. reduction of human-intervention in common tasks.

1.3.3 Managing a Data Center

- Managing a data-center involves many tasks.
- Key management-tasks are: 1) Monitoring 2) Reporting and 3) Provisioning.

1) Monitoring is a process of continuous
→ collection of information and
→ review of the entire storage infrastructure (called as Information Storage System).

- Following parameters are monitored:
  i) Security
  ii) Performance
  iii) Accessibility and
  iv) Capacity.

2) Reporting is done periodically on performance, capacity and utilization of the resources.

- Reporting tasks help to
  → establish business-justifications and
  → establish chargeback of costs associated with operations of data-center.

3) Provisioning is process of providing h/w, s/w & other resources needed to run a data-center.

- Main tasks are: i) Capacity Planning and ii) Resource Planning.

i) Capacity Planning

➢ It ensures that future needs of both user & application will be addressed in most cost-effective way

ii) Resource Planning

➢ It is the process of evaluating & identifying required resources such as
→ Personnel (employees)
→ Facility (site or plant) and
→ Technology (Artificial Intelligence, Deep Learning).

1.4 Virtualization and Cloud Computing

1.4.1 Virtualization

- Virtualization is a technique of abstracting & making physical-resource appear as logical-resource.
- The resource includes compute, storage and network.
- Virtualization existed in the IT-industry for several years in different forms.

Form-1

- Virtualization enables
  → pooling of resources and
  → providing an aggregated view of the resource capabilities.

1) Storage virtualization enables
   → pooling of multiple small storage-devices (say ten thousand 10GB) and
   → providing a single large storage-entity (10000*10=100000GB = 100TB).
2) Compute-virtualization enables
   → pooling of multiple low-power servers (say one thousand 2.5GHz) and
   → providing a single high-power entity (1000*2.5=2500GHz = 2.5THz).

Form-2

- Virtualization also enables centralized management of pooled-resources.
- Virtual-resources can be created from the pooled-resources.
- For example,
  → virtual-disk of a given capacity(say 10GB) can be created from a storage-pool (100TB)
  → virtual-server with specific power (2.5GHz) can be created from a compute-pool (2.5THz)
- Advantages:
  1) Improves utilization of resources (like storage, CPU cycle).
  2) Scalable
  ➢ Storage-capacity can be added from pooled-resources w/o interrupting normal-operations.
  3) Companies save the costs associated with acquisition of new resources.
  4) Fewer resources means less-space and -energy (i.e. electricity).
  5)

1.4.2 Cloud Computing

- Cloud-computing enables companies to use IT-resources as a service over the network. For example:
  - CPU hours used
  - Amount of data-transferred Gigabytes of data-stored
- Advantages:
  1) Provides highly scalable and flexible computing-environment.
  2) Provides resources on-demand to the hosts.
  3) Users can scale up or scale down the demand of resources with minimal management-effort.
  4) Enables self-service requesting through a fully automated request-fulfillment process.
  5) Enables consumption-based metering. .'. consumers pay only for resources they use.
     For example: Jio provides 11Rs plan for 400MB
  6) Usually built upon virtualized data-centers, which provide resource-pooling.

1.5 Data Center Environment

- The data flows from an application to storage through various components collectively referred as a data-center environment.
- The five main components in this environment are
  1) Application
  2) DBMS
  3) Host
  4) Connectivity and
  5) Storage.
- These entities, along with their physical and logical-components, facilitate data-access.

1.5.1 Application

- An application is a program that provides the logic for computing-operations.
- It provides an interface between user and host. (R/W --> read/write)
- The application sends requests to OS to perform R/W-operations on the storage.
- Applications can be placed on the database.
  Then, the database can use OS-services to perform R/W-operations on the storage.
- Applications can be classified as follows:
  → business applications
  → infrastructure management applications
  → data protection applications
  → security applications.
- Some examples of the applications are:
  → e-mail
  → enterprise resource planning (ERP)
  → backup
  → antivirus
- Characteristics of I/Os generated by application influence the overall performance of storage-device.
- Common I/O characteristics are:
  → Read vs. Write intensive
  → Sequential vs. Random
  → I/O size

1.5.2 DBMS

- DBMS is a structured way to store data in logically organized tables that are inter-related.
- The DBMS
  - → processes an application's request for data and
  - → instructs the OS to transfer the appropriate data from the storage.
- Advantages:
  1) Helps to optimize the storage and retrieval of data.
  2) Controls the creation, maintenance and use of a database.

1.5.3 Host

- Host is a client- or server-computer that runs applications.
- Users store and retrieve data through applications. (hosts --> compute-systems)
- Hosts can be physical- or virtual-machines.
- Example of host includes
  - → desktop computers
  - → servers
  - → laptops
  - → smartphones.

A host consists of
1) CPU
2) Memory
3) I/O devices
4) Software.

The software includes

i) OS
ii) Device-drivers
iii) Logical volume manager (LVM)
iv) File-system

- The software can be installed individually or may be part of the OS.

1.5.3.1 Operating System (OS)

- An OS is a program that acts as an intermediary between
  - → application and
  - → hardware-components.
- The OS controls all aspects of the computing-environment.
- Data-access is one of the main service provided by OS to the application.
- Tasks of OS:
  1) Monitor and respond to user actions and the environment.
  2) Organize and control hardware-components.
  3) Manage the allocation of hardware-resource (simply the resource).
  4) Provide security for the access and usage of all managed resources.
  5) Perform storage-management tasks.
  6) Manage components such as file-system, LVM & device drivers.

1.5.3.1.1 Memory Virtualization

- Memory-virtualization is used to virtualize the physical-memory (RAM) of a host.
- It creates a VM with an address-space larger than the physical-memory space present in computer.
- The virtual-memory consists of
  - → address-space of the physical-memory and
  - → part of address-space of the disk-storage.
- The entity that manages the virtual-memory is known as the **virtual-memory manager (VMM)**.

- The VMM
  → manages the virtual-to-physical-memory mapping and
  → fetches data from the disk-storage
- The space used by the VMM on the disk is known as a swap-space.
- A **swap-space** is a portion of the disk that appears like physical-memory to the OS.
- The memory is divided into contiguous blocks of fixed-size pages. (VM --> virtual-memory)

Paging

- A paging
  → moves inactive-pages onto the swap-file and
  → brings inactive-pages back to the physical-memory when required.
- Advantages:
  1) Enables efficient use of the available physical-memory among different applications.
  ¤ Normally, the OS moves the least used pages into the swap-file.
  ¤ Thus, sufficient RAM is provided for processes that are more active.
- Disadvantage:
  1) Access to swap-file pages is slower than physical-memory pages. This is because
  → swap-file pages are allocated on the disk which is slower than physical-memory.

1.5.3.2 Device Driver

- It is a special software that permits the OS & hardware-component to interact with each other.
- The hardware-component includes printer, a mouse and a hard-drive.
- A device-driver enables the OS to
  → recognize the device and
  → use a standard interface to access and control devices.
- Device-drivers are hardware-dependent and OS-specific.

1.5.3.3 Logical Volume Manager (LVM)

- LVM is a software that
  → runs on the host and
  → manages the logical- and physical-storage.
- It is an intermediate-layer between file-system and disk.
- Advantages:
  1) Provides optimized storage-access.
  2) Simplifies storage-management. (PVID --> Physical-Volume IDentifier)
  3) Hides details about disk and location of data on the disk.
  4) Enables admins to change the storage-allocation without interrupting normal-operations.
  5) Enables dynamic-extension of storage-capacity of the file-system.
- The main components of LVM are: 1) Physical-volumes 2) Volume-groups and 3) Logical-volumes.

1) Physical-Volume (PV): refers to a disk connected to the host.

2) Volume-Group (VG): refers to a group of one or more PVs.

- A unique PVID is assigned to each PV when it is initialized for use.
- PVs can be added or removed from a volume-group dynamically.
- PVs cannot be shared between different volume-groups.
- The volume-group is handled as a single unit by the LVM.
- Each PV is divided into equal-sized data-blocks called **physical-extents**.

3) Logical-Volume (LV): refers to a partition within a volume-group.

- Logical-volumes vs. Volume-group
  i) LV can be thought of as a disk-partition.
  ii) Volume-group can be thought of as a disk.
- The size of a LV is based on a multiple of the physical-extents.
- The LV appears as a physical-device to the OS.
- A LV is made up of non-contiguous physical-extents and may span over multiple PVs.

- A file-system is created on a LV.
- These LVs are then assigned to the application.
- A LV can also be mirrored to improve data-availability.

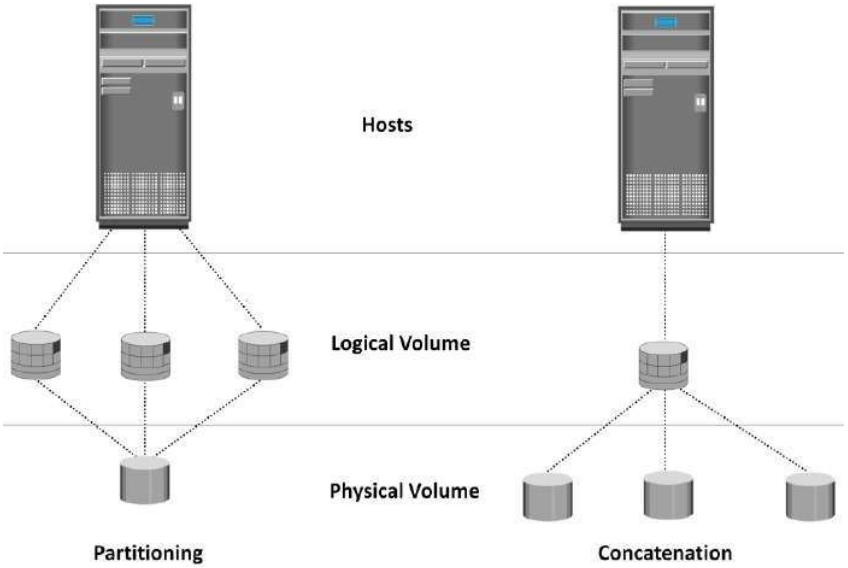

Figure 1-6: Disk partitioning and concatenation

- It can perform partitioning and concatenation (Figure 1-6).

1) Partitioning

- A larger-capacity disk is partitioned into smaller-capacity virtual-disks.
- Disk-partitioning is used to improve the utilization of disks.

2) Concatenation

- Several smaller-capacity disks are aggregated to form a larger-capacity virtual-disk.
- The larger-capacity virtual-disk is presented to the host as one big logical-volume.

1.5.3.4 File System

- A file is a collection of related-records stored as a unit with a name. (say employee.lst)
- A file-system is a structured way of storing and organizing data in the form of files.
- File-systems enable easy access to data-files residing within
  - → disk-drive
  - → disk-partition or
  - → logical-volume.
- A file-system needs host-based software-routines (API) that control access to files.
- It provides users with the functionality to create, modify, delete and access files.
- A file-system organizes data in a structured hierarchical manner via the use of directories (i.e. folder)
- A **directory** refers to a container used for storing pointers to multiple files.
- All file-systems maintain a pointer-map to the directories and files.
- Some common file-systems are:
  - → FAT 32 (File Allocation Table) for Microsoft Windows
  - → NT File-system (NTFS) for Microsoft Windows
  - → UNIX File-system (UFS) for UNIX
  - → Extended File-system (EXT2/3) for Linux

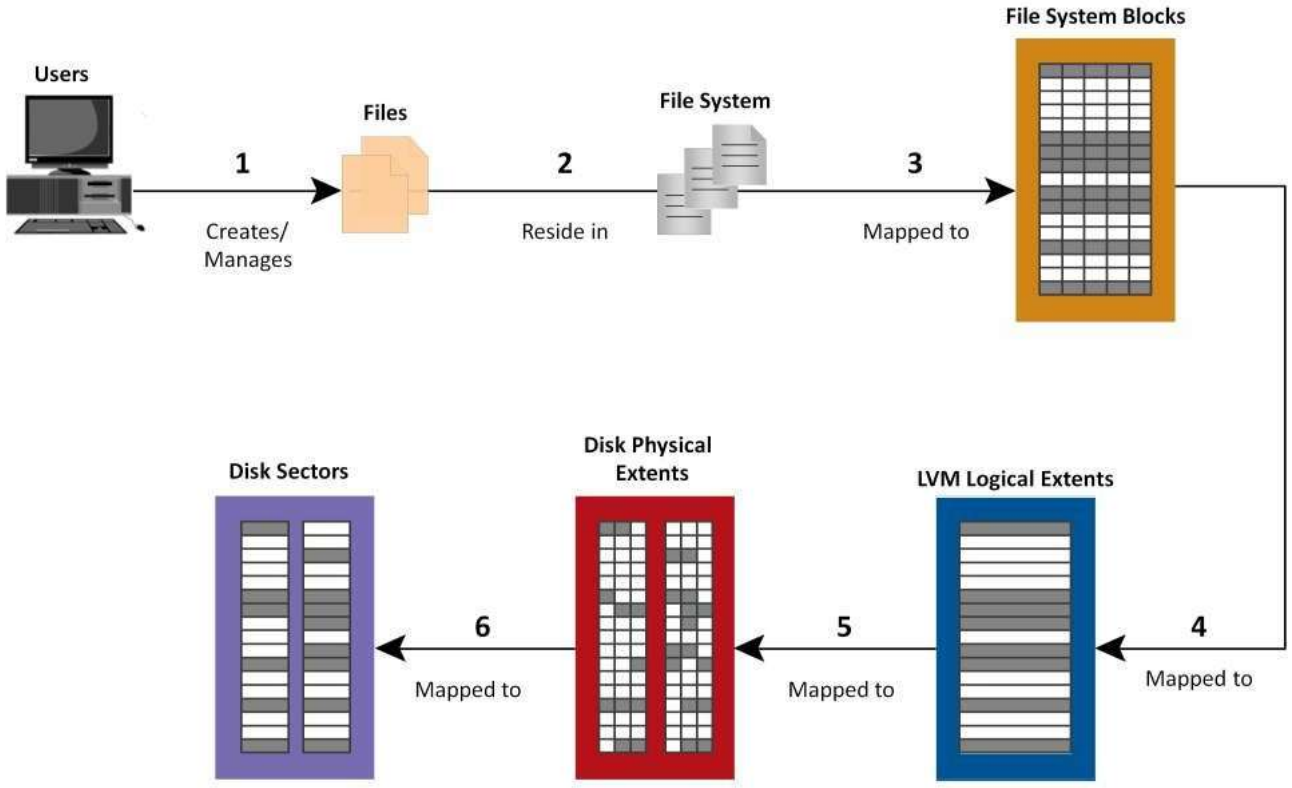

Figure 1-7: Process of mapping user files to disk storage

- Figure 1-7 shows process of mapping user-files to the disk-storage with an LVM:
  1) Files are created and managed by users and applications.
  2) These files reside in the file-system.
  3) The file-system are mapped to file-system blocks.
  4) The file-system blocks are mapped to logical-extents.
  5) The logical-extents are mapped to disk physical-extents by OS or LVM.
  6) Finally, these physical-extents are mapped to the disk-storage.

1.5.3.5 Compute Virtualization

- Compute-virtualization is a technique of masking(or abstracting) the physical-hardware from the OS.
- It can be used to create portable virtual-computers called as **virtual-machines** (VMs).
- A VM appears like a host to the OS with its own CPU, memory and disk (Figure 1-8).
  However, all VMs share the same underlying hardware in an isolated-manner
- Compute-virtualization is done by virtualization-layer called as **hypervisor**.
- The hypervisor
  → resides between the hardware and VMs.
  → provides resources such as CPU, memory and disk to all VMs.
- Within a server, a large no. of VMs can be created based on the hardware-capabilities of the server.
- Advantages:
  1) Allows multiple-OS and applications to run concurrently on a single-computer.
  2) Improves server-utilization.
  3) Provides server-consolidation.

Because of server-consolidation, companies can run their data-center with fewer servers Advantages of server-consolidation:

i) Cuts down the cost for buying new servers.
ii) Reduces operational-cost.
iii) Saves floor- and rack-space used for data-center.

4) VM can be created in less time when compared to setting up the actual server.
5) VM can be restarted or upgraded without interrupting normal-operations.
6) VM can be moved from one computer to another w/o interrupting normal-operations.

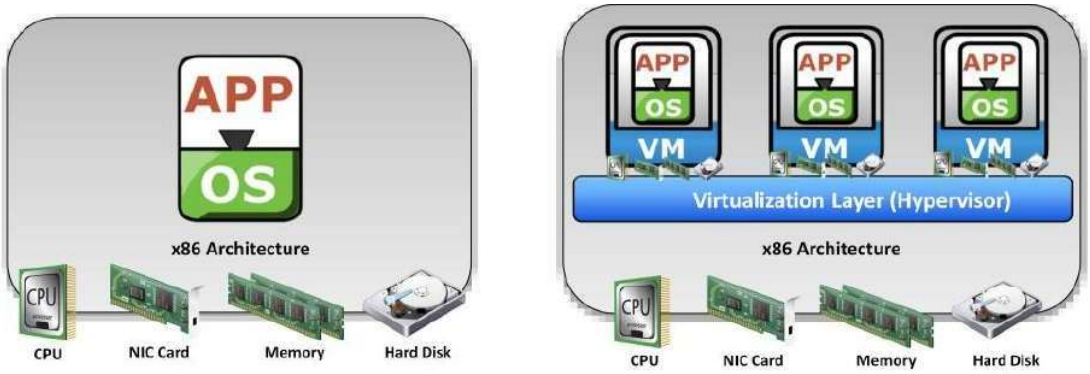

Figure 1-8: Server virtualization

1.5.4 Connectivity

- **Connectivity** refers to interconnection between host and peripheral-devices such as storage-devices
- Components of connectivity is classified as: 1) Physical-Components and 2) Interface Protocols

1.5.4.1 Physical Components

- Physical-components refers to hardware-components used for connection between host & storage.
- Three components of connectivity are (Figure 1-9):
  1) Host interface device
  2) Port and
  3) Cable

1) **Host Interface Device** is used to connect a host to other hosts and storage-devices.
   - Example:
     HBA (host bus adapter)
     NIC (network interface card).
   - HBA is an ASIC board that performs I/O-operations between host and storage.
   - Advantage:
     1) HBA relieves the CPU from additional I/O-processing workload.
   - A host typically contains multiple HBAs. (ASIC --> application-specific integrated circuit).
2) **Port** refers to a physical connecting-point to which a device can be attached.
   - An HBA may contain one or more ports to connect the host to the storage-device.
3) **Cable** is used to connect hosts to internal/external devices using copper-wire or optical-fiber.

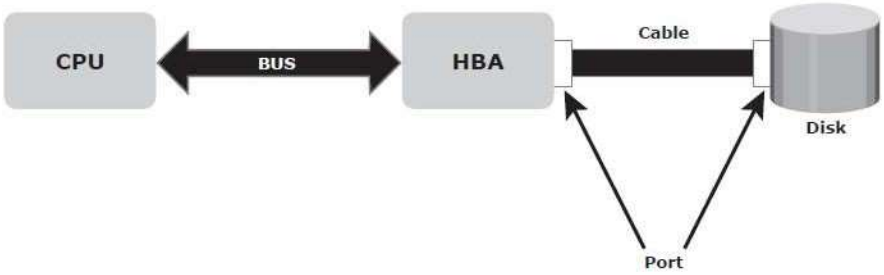

Figure 1-9: Physical components of connectivity

1.5.4.2 Interface Protocol

- Interface-Protocol enables communication between host and storage.
- Protocols are implemented using interface-devices (or controllers) at both source and destination.
- The popular protocols are:
    1) IDE/ATA (Integrated Device Electronics/Advanced Technology Attachment)
    2) SCSI (Small Computer System Interface)
    3) FC (Fibre Channel) and
    4) IP (Internet Protocol).

1.5.4.2.1 IDE/ATA

- It is a standard interface for connecting storage-devices inside PCs (Personal Computers).
- The storage-devices can be disk-drives or CD-ROM drives.
- It supports parallel-transmission. Therefore, it is also known as Parallel ATA (PATA).
- It includes a wide variety of standards.
    1) Ultra DMA/133 ATA supports a throughput of 133 Mbps.
    2) In a master-slave configuration, ATA supports 2 storage-devices per connector.
    3) Serial-ATA (SATA) supports single bit serial-transmission. .
    4) SATA version 3.0 supports a data-transfer rate up to 6 Gbps.

1.5.4.2.2 SCSI

- It has emerged as a preferred protocol in high-end computers.
- Compared to ATA, SCSI
    → supports parallel-transmission and
    → provides improved performance, scalability, and compatibility.
- Disadvantage:
    1) Due to high cost, SCSI is not used commonly in PCs.
- It includes a wide variety of standards.
    1) SCSI supports up to 16 devices on a single bus.
    2) SCSI provides data-transfer rates up to 640 Mbps (for the Ultra-640 version).
    3) SAS (Serial Attached SCSI) is a point-to-point serial protocol.
    4) SAS version 2.0 supports a data-transfer rate up to 6 Gbps.

1.5.4.2.3 Fibre Channel

- It is a widely used protocol for high-speed communication to the storage-device.
- Advantages:
    1) Supports gigabit network speed.
    2) Supports multiple protocols and topologies.
- It includes a wide variety of standards.
    1) It supports a serial data-transmission that operates over copper-wire and optical-fiber.
    2) FC version 16FC supports a data-transfer rate up to 16 Gbps.

1.5.4.2.4 IP

- It is a protocol used for communicating data across a packet-switched network.
- It has been traditionally used for host-to-host traffic.
- '.' of new technologies, IP network has become a feasible solution for host-to-storage communication
- Advantages:
    1) Reduced cost & maturity
    2) Enables companies to use their existing IP-based network.
- Common example of protocols that use IP for host-to-storage communication: 1) iSCSI and 2) FCIP

1.5.5 Storage

- A storage-device uses magnetic-, optical-, or solid-state-media.
    1) Disk, tape and diskette uses magnetic-media for storage.
    2) CD/DVD uses optical-media for storage.
    3) Flash drives uses solid-state-media for storage.

1) **Tapes** are a popular storage-device used for backup because of low cost.

- Disadvantage:
    i) Data is stored on the tape linearly along the length of the tape.
    - Search and retrieval of data is done sequentially.
    - As a result, random data-access is slow and time consuming.
    - Hence, tapes is not suitable for applications that require real-time access to data.

    ii) In shared environment, data on tape cannot be accessed by multiple applications simultaneously.
    - Hence, tapes can be used by one application at a time.

    iii) On a tape-drive, R/W-head touches the tape-surface.
    - Hence, the tape degrades or wears out after repeated use.

    iv) More overhead is associated with managing the tape-media because of
        → storage and retrieval requirements of data from the tape.

2) **Optical-disk** is popular in small, single-user computing-environments.

- It is used
    → to store data like photo, video
    → as a backup-medium on PCs.
- Example:
    CD-RW
    Blu-ray disc
    and DVD.
- It is used as a distribution medium for single applications such as games.
- It is used as a means of transferring small amounts of data from one computer to another.
- Advantages:
    1) Provides the capability to write once and read many
        (WORM). For example: CD-ROM
    2) Optical-disks, to some degree, guarantee that the content has not been altered.
- Disadvantage:
    1) Optical-disk has limited capacity and speed. Hence, it is not used as a business storage-solution
- Collections of optical-discs in an array is called as a **jukebox**.
    The jukebox is used as a fixed-content storage-solution.

3) **Disk-drives** are used for storing and accessing data for performance-intensive, online applications.

- Advantages:
    1) Disks support rapid-access to random data-locations.
        Thus, data can be accessed quickly for a large no. of simultaneous applications.
    2) Disks have a large capacity.
    3) Disk-storage is configured with multiple-disks to provide
        → increased capacity and
        → enhanced performance.

4) **Flash drives** uses semiconductor media. (Flash drives --> Pen drive)

- Advantages:
  1) Provides high performance and
  2) Provides low power-consumption.

1.6 RAID Implementation Methods

- RAID stands for Redundant Array of Independent Disk.
- RAID is the way of combining several independent small disks into a single large-size storage.
- It appears to the OS as a single large-size disk.
- It is used to increase performance and availability of data-storage.
- There are two types of RAID implementation 1) hardware and 2) software.
- RAID-controller is a specialized hardware which
  → performs all RAID-calculations and
  → presents disk-volumes to host.
- Key functions of RAID-controllers:
  1) Management and control of disk-aggregations.
  2) Translation of I/O-requests between logical-disks and physical-disks.
  3) Data-regeneration in case of disk-failures.

1.6.1 Software-RAID

- It uses host-based software to provide RAID functions.
- It is implemented at the OS-level.
- It does not use a dedicated hardware-controller to manage the storage-device.
- Advantage:
  1) Provides cost- and simplicity-benefits when compared to hardware-RAID.
- Disadvantages:
  1) Decreased Performance
  - RAID affects overall system-performance.
  - This is due to the additional CPU-cycles required to perform RAID-calculations.

  2) Supported Features
  - RAID does not support all RAID-levels.

  3) OS compatibility
  - RAID is tied to the host-OS.
  - Hence, upgrades to RAID (or OS) should be validated for compatibility.

1.6.2 Hardware-RAID

- It is implemented either on the host or on the storage-device.
- It uses a dedicated hardware-controller to manage the storage-device.

1) Internal-Controller

- A dedicated controller is installed on a host.
- Disks are connected to the controller.
- The controller interacts with the disks using PCI-bus.
- Manufacturers integrate the controllers on motherboards.
- Advantage:
  1) Reduces the overall cost of the system.
- Disadvantage:
  1) Does not provide the flexibility required for high-end storage-devices.

2) External-controller

- The external-controller is an array-based hardware-RAID.
- It acts as an interface between host and disks.
- It presents storage-volumes to host, which manage the drives using the supported protocol.

1.7 RAID Array Components

- A RAID-array is a large container that holds (Figure 1-10):
    1) RAID-controller (or simply the controller)
    2) Number of disks
    3) Supporting hardware and software.

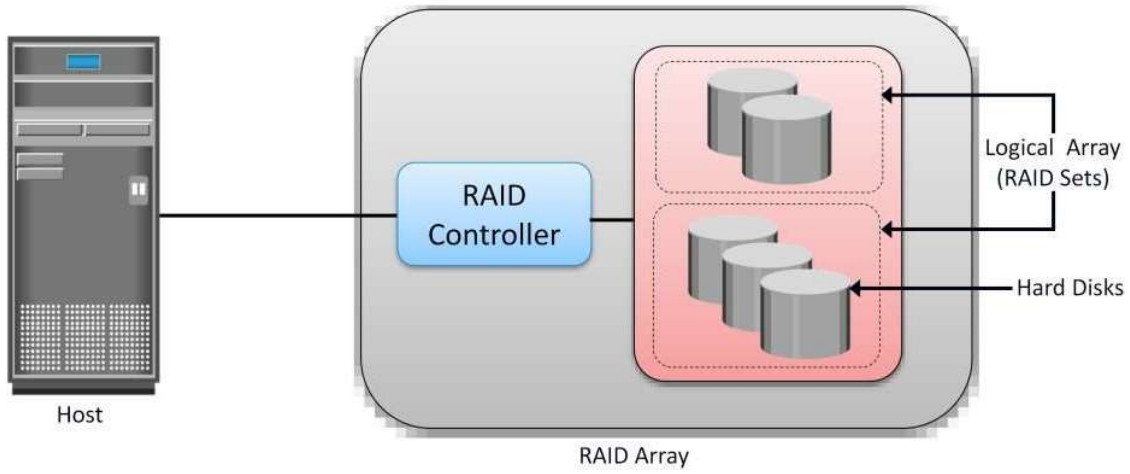

Figure 1-10: Components of RAID array

- The **logical-array** is a subset of disks grouped to form logical-associations.
- Logical-arrays are also known as a **RAID-set**. (or simply the set).
- Logical-array consists of logical-volumes (LV).
- The OS recognizes the LVs as if they are physical-disks managed by the controller.

1.8 RAID Techniques

- RAID-levels are defined based on following 3 techniques:
    1) Striping (used to improve performance of storage)
    2) Mirroring (used to improve data-availability) and
    3) Parity (used to provide data-protection)
- The above techniques determine
    → performance of storage-device (i.e. better performance --> least response-time)
    → data-availability
    → data-protection
- Some RAID-arrays use a combination of above 3 techniques.
    For example: Striping with mirroring
    Striping with parity

1.8.1 Striping

- Striping is used to improve performance of a storage-device.
- It is a technique of splitting and distribution of data across multiple disks.
- Main purpose: To use the disks in parallel.
- It can be bitwise, byte-wise or block wise.
- A **RAID-set** is a group of disks.

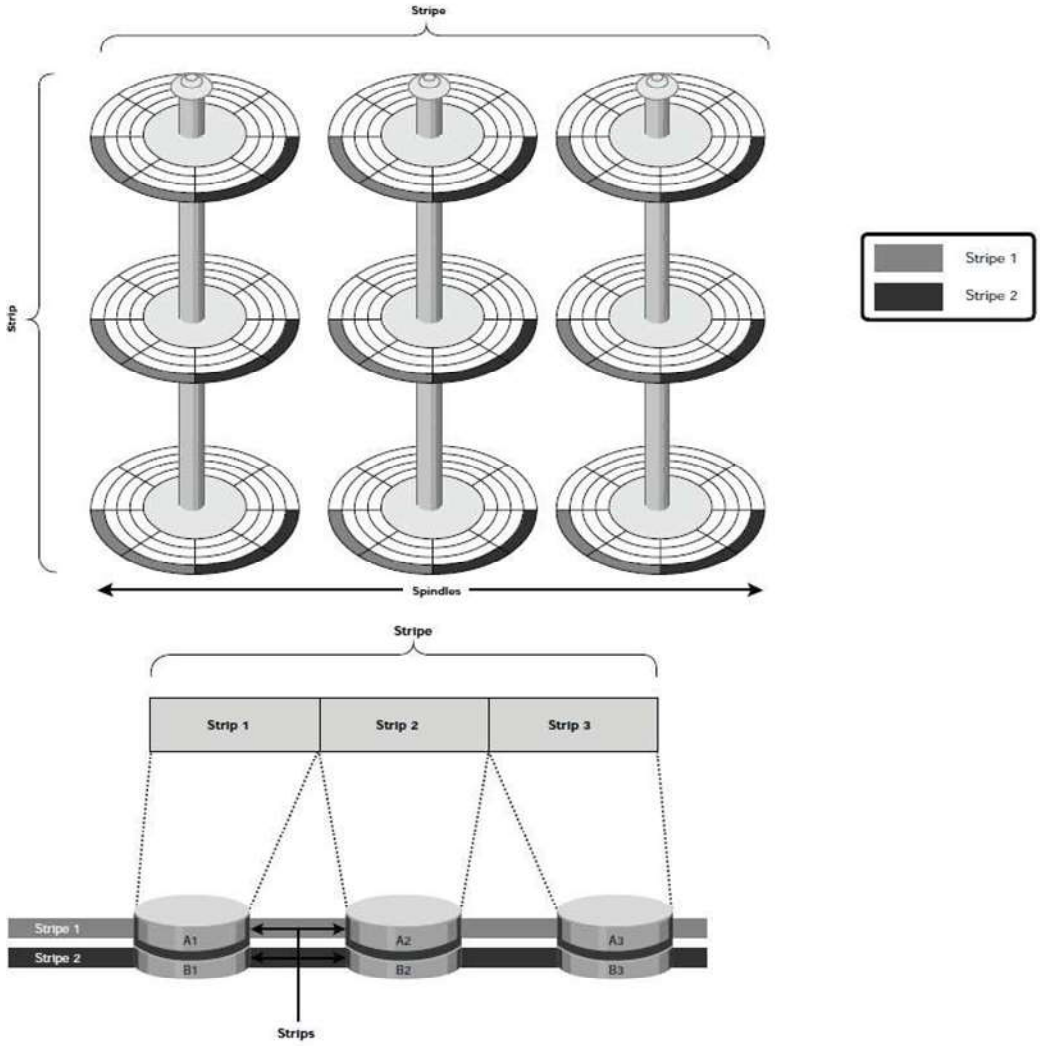

Figure 1-11: Striped RAID set

- In each disk, a predefined number of strips are defined.
- **Strip** refer to a group of continuously-addressable-blocks in a disk.
- **Stripe** refer to a set of aligned-strips that spans all the disks. (Figure 1-11)
- **Strip-size** refers to maximum amount-of-data that can be accessed from a single disk.
  In other words, strip-size defines the number of blocks in a strip.
- In a stripe, all strips have the same number of blocks.
- **Stripe-width** refers to the number of strips in a stripe.
- Striped-RAID does not protect data. To protect data, parity or mirroring must be used.
- Advantage:
  1) As number of disks increases, the performance also increases. This is because
     → more data can be accessed simultaneously.

(Example for stripping: If one man is asked to write A-Z the amount of time taken by him will be more as compared to 2 men writing A-Z because from the 2 men, one man will write A-M and another will write N-Z at the same time so this will speed up the process)

1.8.2 Mirroring

- Mirroring is used to improve data-availability (or data-redundancy).
- All the data is written to 2 disks simultaneously. Hence, we have 2 copies of the data.
- Advantages:
  1) Reliable
     - Provides protection against single disk-failure.
     - In case of failure of one disk, the data can be accessed on the surviving-disk (Figure 1-12).
     - Thus, the controller can still continue to service the host's requests from surviving-disk.
     - When failed-disk is replaced with a new-disk, controller copies data from surviving-disk to new-disk
     - The disk-replacement activity is transparent to the host.
  2) Increases read-performance because each read-request can be serviced by both disks.
- Disadvantages:
  1) Decreases write-performance because
     → each write-request must perform 2 write-operations on the disks.
  2) Duplication of data. Thus, amount of storage-capacity needed is twice amount of data being stored (E.g. To store 100GB data, 200GB disk is needed).
  3) Considered expensive and preferred for mission-critical applications (like military application).
- Mirroring is not a substitute for data-backup. Mirroring vs. Backup
  1) Mirroring constantly captures changes in the data.
  2) On the other hand, backup captures point-in-time images of data.

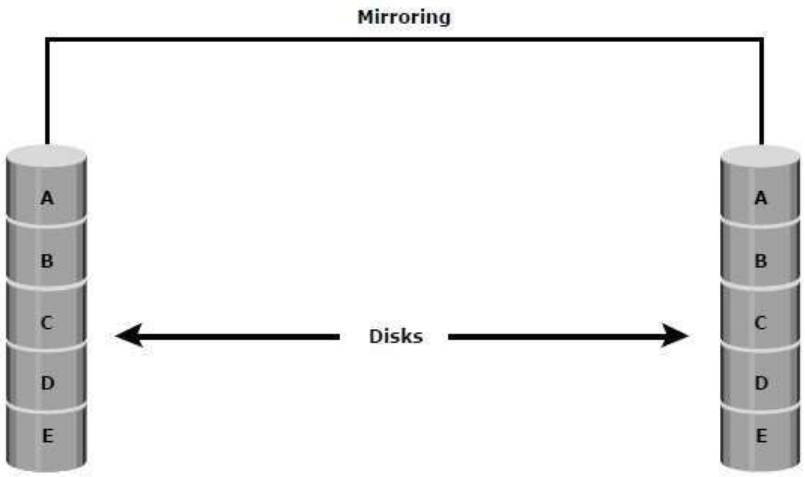

Figure 1-12: Mirrored disks in an array

1.8.3 Parity

- Parity is used to provide data-protection in case of a disk-failure.
- An additional disk is added to the stripe-width to hold parity.
- In case of disk-failure, parity can be used for reconstruction of the missing-data.
- Parity is a technique that ensures protection of data without maintaining a duplicate-data
- Parity-information can be stored on
  → separate, dedicated-disk or
  → distributed across all the disks.
- For example (Figure 1-13):
  Consider a RAID-implementation with 5 disks (5*100 GB = 500 GB).
  1) The first four disks contain the data (4*100 = 400GB).
  2) The fifth disk stores the parity-information (1*100 = 100GB).

Parity vs. Mirroring

i) Parity requires 25% extra disk-space. (i.e. 500GB disk for 400GB data).
ii) Mirroring requires 100% extra disk-space.(i.e. 800GB disk for 400GB data).

- The controller is responsible for calculation of parity.
- Parity-value can be calculated by

  P = D1 + D2 + D3 + D4

  where D1 to D4 is striped-data across the set of five disks.
- Now, if one of the disks fails (say D1), the missing-value can be calculated by

  D1 = P - (D2 + D3 + D4)

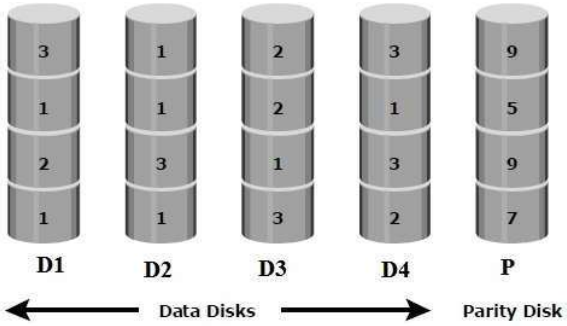

Figure 1-13: Parity RAID

- Advantages:
  1) Compared to mirroring, parity reduces the cost associated with data-protection.
  2) Compared to mirroring, parity consumes less disk-space. In previous example,
     i) Parity requires 25% extra disk-space. (i.e. 500GB disk for 400GB data).
     ii) Mirroring requires 100% extra disk-space. (i.e. 800GB disk for 400GB data).
- Disadvantage:
  1) Decreases performance of storage-device.
     For example:
     - ¤ Parity-information is generated from data on the disk.
     - ¤ Therefore, parity must be re-calculated whenever there is change in data.
     - ¤ This re-calculation is time-consuming and hence decreases the performance.

1.9 RAID Levels

LEVELS	BRIEF DESCRIPTION
RAID 0	Striped array with no fault tolerance
RAID 1	Disk mirroring
RAID 3	Parallel access array with dedicated parity disk
RAID 4	Striped array with independent disks and a dedicated parity disk
RAID 5	Striped array with independent disks and distributed parity
RAID 6	Striped array with independent disks and dual distributed parity
Nested	Combinations of RAID levels. Example: RAID 1 + RAID 0

Table 1-1: Raid Levels

1.9.1 RAID-0

- RAID-0 is based on striping-technique (Figure 1-14).
- Striping is used to improve performance of a storage-device.
- It is a technique of splitting and distribution of data across multiple disks.
- Main purpose:
  To use the disks in parallel.
- Therefore, it utilizes the full storage-capacity of the storage-device.
- Read operation: To read data, all the strips are combined together by the controller.
- Advantages:
  1) Used in applications that need high I/O-throughput. (Throughput --> Efficiency).
  2) As number of disks increases, the performance also increases. This is because
     → more data can be accessed simultaneously.
- Disadvantage:
  1) Does not provide data-protection and data-availability in case of disk-failure.

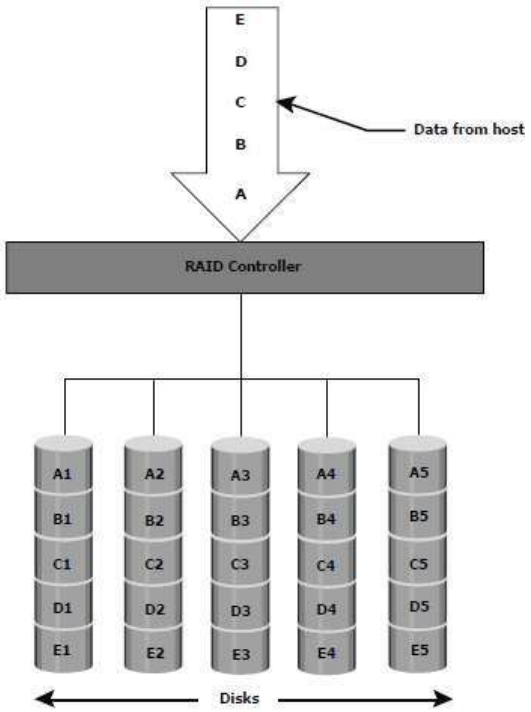

Figure 1-14: RAID 0

1.9.2 RAID-1

- RAID-1 is based on mirroring-technique.
- Mirroring is used to improve data-availability (or data-redundancy).
- Write operation: The data is stored on 2 different disks. Hence, we have 2 copies of data.
- Advantages:
  1) Reliable
     - Provides protection against single disk-failure.
     - In case of failure of one disk, the data can be accessed on the surviving-disk (Figure 1-15).
     - Thus, the controller can still continue to service the host's requests from surviving-disk.
     - When failed-disk is replaced with a new-disk, controller copies data from surviving-disk to new-disk
     - The disk-replacement activity is transparent to the host.
  2) Increases read-performance because each read-request can be serviced by both disks.
- Disadvantages:
  1) Decreases write-performance because
     → each write-request must perform 2 write-operations on the disks.
  2) Duplication of data.
     Thus, amount of storage-capacity needed is twice amount of data being stored
     (E.g. To store 100 GB data, 200 GB disk is required).
  3) Considered expensive and preferred for mission-critical applications (like military application)

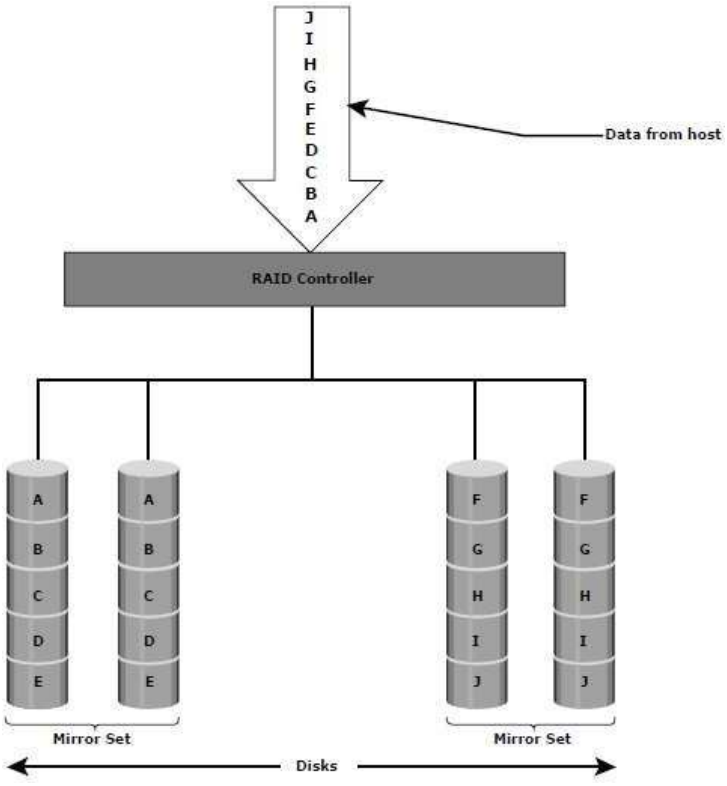

Figure 1-15: RAID-1

1.9.3 Nested-RAID

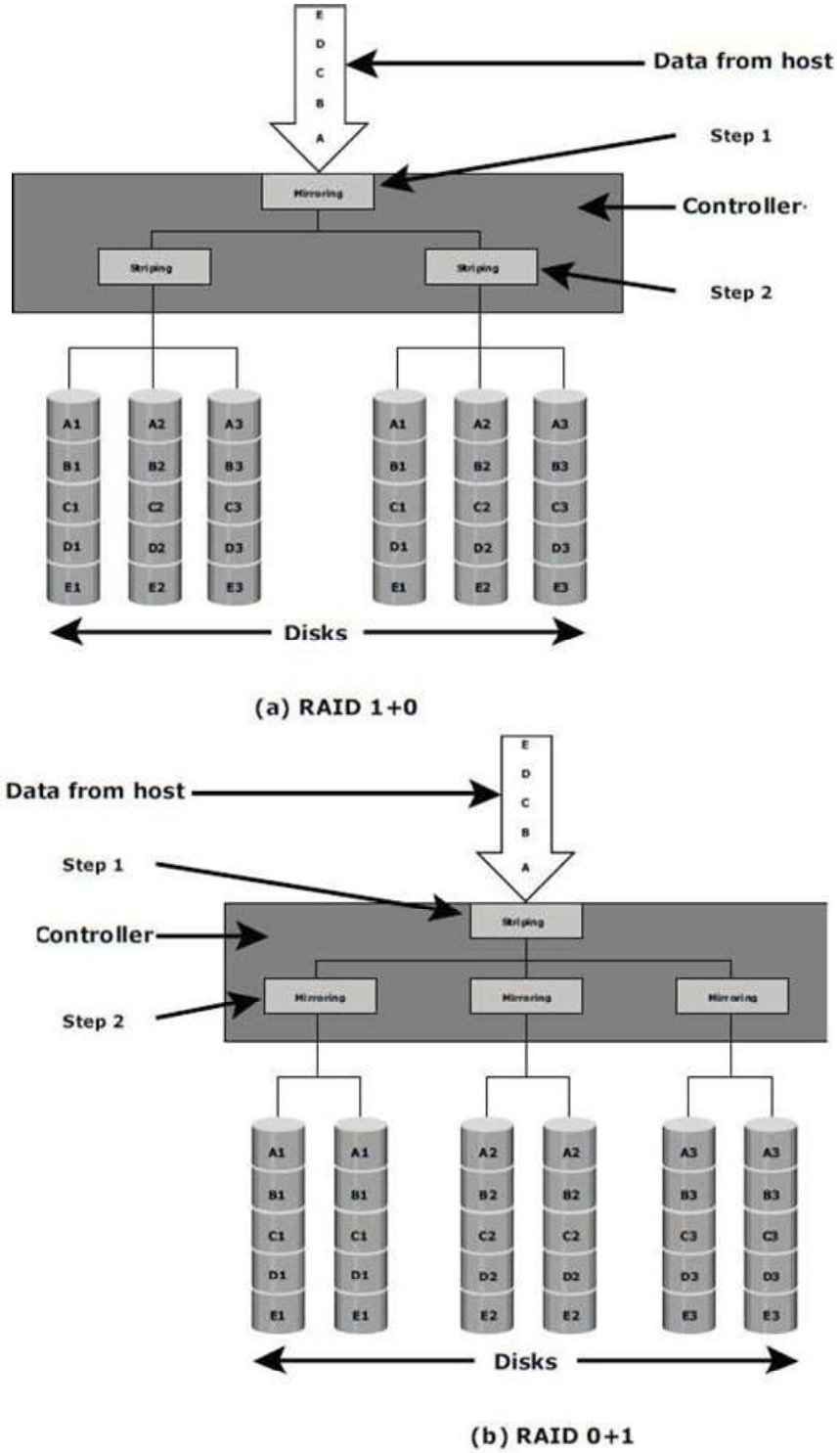

Figure 1-16: Nested-RAID

- Most data-centers require data-availability & performance from their storage-devices (Figure 1-16).
- RAID-01 and RAID-10 combines
  → performance-benefit of RAID-0 and
  → availability-benefit of RAID-1. (RAID-10 is also known as RAID-1+0)

- It uses mirroring- and striping-techniques.
- It requires an even-number of disks. Minimum no. of disks = 4.
- Some applications of RAID-10:
    1) High transaction rate OLTP (Online Transaction Processing)
    2) Large messaging installations
    3) Database applications that require
        → high I/O-throughput
        → random-access and
        → high-availability.
- Common misunderstanding is that RAID-10 and RAID-01 are the same. But, they are totally different

    1) RAID-10
    - RAID-10 is also called **striped-mirror**.
    - The basic element of RAID-10 is a mirrored-pair.
        1) Firstly, the data is mirrored and
        2) Then, both copies of data are striped across multiple-disks.

    2) RAID-01
    - RAID-01 is also called **mirrored-stripe**
    - The basic element of RAID-01 is a stripe.
        1) Firstly, data are striped across multiple-disks and
        2) Then, the entire stripe is mirrored.
- Advantage of rebuild-operation::

  1) Provides protection against single disk-failure.
  - In case of failure of one disk, the data can be accessed on the surviving-disk (Figure 1-15).
  - Thus, the controller can still continue to service the host's requests from surviving-disk.
  - When failed-disk is replaced with a new-disk, controller copies data from surviving-disk to new-disk
- Disadvantages of rebuild-operation:
    1) Increased and unnecessary load on the surviving-disks.
    2) More vulnerable to a second disk-failure.

1.9.4 RAID-3

- RAID-3 uses both striping & parity techniques.
    1) Striping is used to improve performance of a storage-device.
    2) Parity is used to provide data-protection in case of disk-failure.
- Parity-information is stored on separate, dedicated-disk.
- Data is striped across all disks except the parity-disk in the array.
- In case of disk-failure, parity can be used for reconstruction of the missing-data.
- For example (Figure 1-17):
    - Consider a RAID-implementation with 5 disks (5*100GB = 500GB).
        1) The first 4 disks contain the data (4*100 = 400GB).
        2) The fifth disk stores the parity-information (1*100 = 100GB).
    - Therefore, parity requires 25% extra disk-space (i.e. 500GB disk for 400GB data).
- Advantages:
    1) Striping is done at the bit-level.
        Thus, RAID-3 provides good bandwidth for the transfer of large volumes of data.
    2) Suitable for video streaming applications that involve large sequential data-access.
- Disadvantages:
    1) Always reads & writes complete stripes of data across all disks '.' disks operate in parallel.
    2) There are no partial writes that update one out of many strips in a stripe.

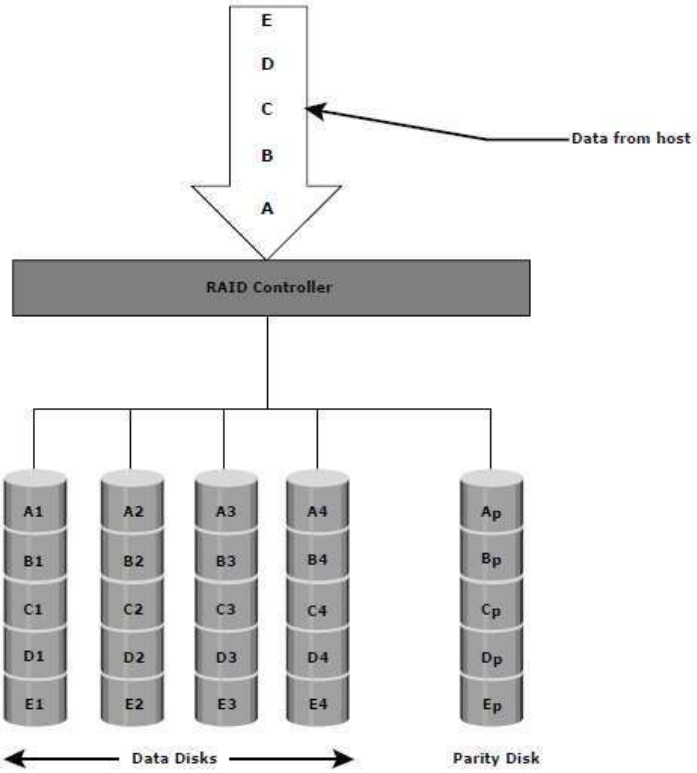

Figure 1-17: RAID-3

1.9.5 RAID-4

- Similar to RAID-3, RAID-4 uses both striping & parity techniques.
  1) Striping is used to improve performance of a storage-device.
  2) Parity is used to provide data-protection in case of disk-failure.
- Parity-information is stored on a separate dedicated-disk.
- Data is striped across all disks except the parity-disk.
- In case of disk-failure, parity can be used for reconstruction of the missing-data.
- Advantages:
  1) Striping is done at the block-level.
     Hence, data-element can be accessed independently.
     i.e. A specific data-element can be read on single disk without reading an entire stripe
  2) Provides
     → good read-throughput and
     → reasonable write-throughput.

1.9.6 RAID-5

- Problem:
  In RAID-3 and RAID-4, parity is written to a dedicated-disk.
  If parity-disk fails, we will lose our entire backup.

  Solution: To overcome this problem, RAID-5 is proposed.
  In RAID-5, we distribute the parity-information evenly among all the disks.
- RAID-5 similar to RAID-4 because
  it uses striping and the drives (strips) are independently accessible.
- Advantages:

1) Preferred for messaging & media-serving applications.
2) Preferred for RDBMS implementations in which database-admins can optimize data-access.

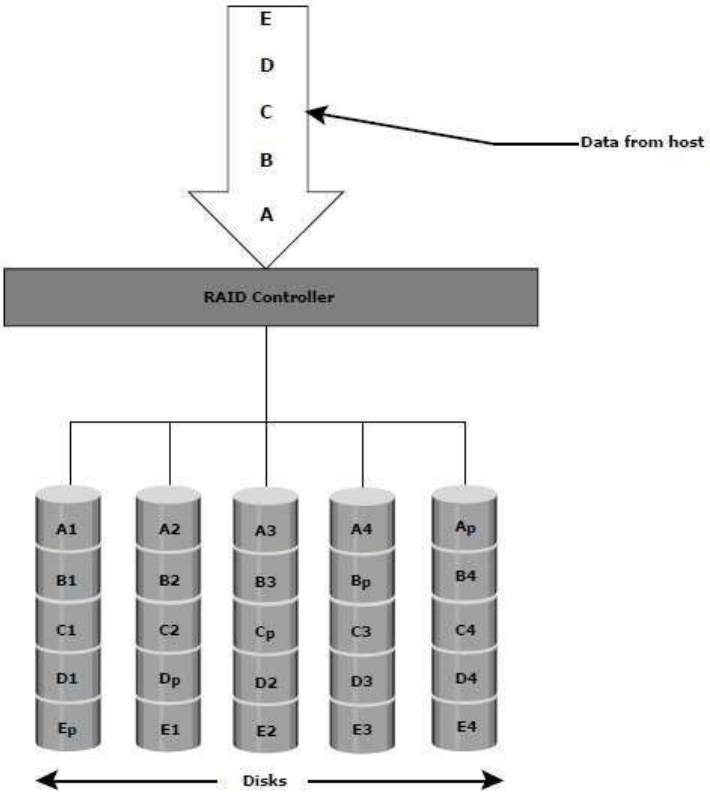

Figure 1-18: RAID-5

1.9.7 RAID-6

- RAID-6 is similar to RAID-5 except that it has
  → a second parity-element to enable survival in case of 2 disk-failures. (Figure 1-19).
- Therefore, a RAID-6 implementation requires at least 4 disks.
- Similar to RAID-5, parity is distributed across all disks.
- Disadvantages:
  Compared to RAID-5,
  1) Write-penalty is more. .'. RAID-5 writes perform better than RAID-6
  2) The rebuild-operation may take longer time. This is due to the presence of 2 parity-sets.

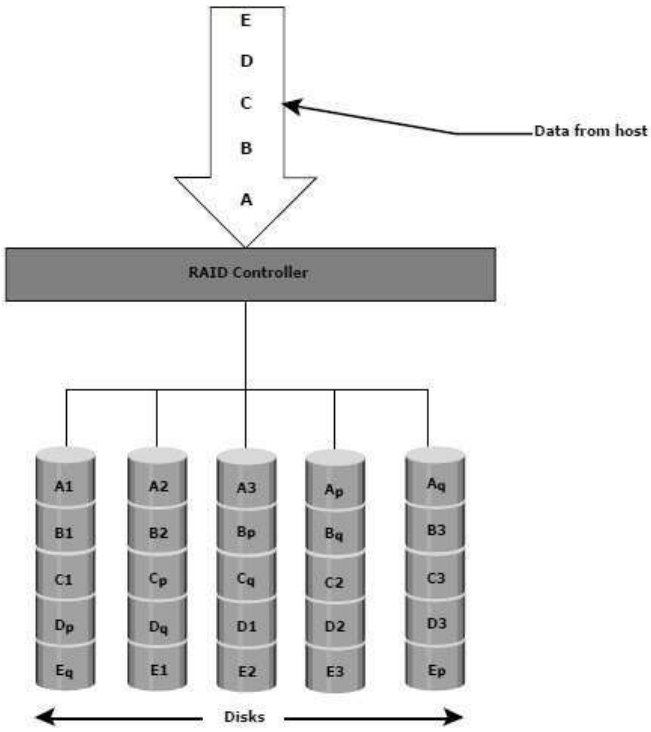

Figure 1-19: RAID-6

1.10 RAID Impact on Disk Performance

- When choosing a RAID-type, it is important to consider the impact to disk-performance.
- In both mirrored and parity-RAIDs,
  each write-operation translates into more I/O-overhead for the disks. This is called **write-penalty**
- Figure 1-20 illustrates a single write-operation on RAID-5 that contains a group of five disks.
  1) Four disks are used for data and
  2) One disk is used for parity.
- The parity (Ep) can be calculated
  by: $E_p = E1 + E2 + E3 + E4$
  Where, E1 to E4 is striped-data across the set of five disks.
- Whenever controller performs a write-operation, parity must be computed by
  → reading old-parity ($E_{p\ old}$) & old-data ($E4_{\ old}$) from the disk. This results in 2 read-operations
- The new parity ($E_{p\ new}$) can be calculated
  by: $E_{p\ new} = E_{p\ old} - E4_{\ old} + E4_{\ new}$
- After computing the new parity, controller completes write-operation by
  → writing the new-data and new-parity onto the disks. This results in 2 write-operations.
- Therefore, controller performs 2 disk reads and 2 disk writes for each write-operation.
- Thus, in RAID-5, the write-penalty = 4.

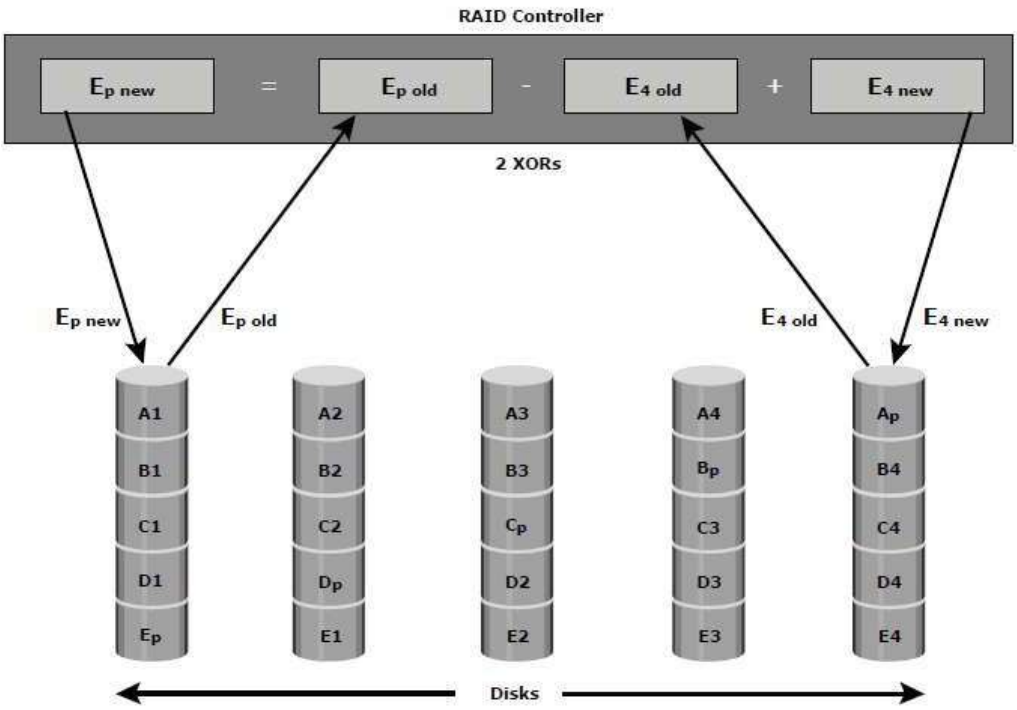

Figure 1-20: Write penalty in RAID-5

1.10.1 Application IOPS and RAID Implementations

- **Input Output per Second (IOPS)** refers to number of reads and writes performed per second.
- When deciding no. of disks for application, it is important to consider impact of RAID based on IOPS.
- The total disk-load depend on
  1) Type of RAID-implementation (RAID-0, RAID-1 or RAID 3) and
  2) Ratio of read compared to write from the host.
- The following example illustrates the method of computing the disk-load in different types of RAID.
- Consider an application that generates 5,200 IOPS, with 60% of them being reads.

Case 1: RAID-5

- The disk-load is calculated as follows:

Disk-load = 0.6 × 5,200 + 4 × (0.4 × 5,200) [because the write-penalty for RAID-5 is 4]
= 3,120 + 4 × 2,080
= 3,120 + 8,320
= 11,440 IOPS

Case 2: RAID-1

- The disk-load is calculated as follows:

Disk-load = 0.6 × 5,200 + 2 × (0.4 × 5,200) ['.' every write results as 2 writes to disks]
= 3,120 + 2 × 2,080
= 3,120 + 4,160
= 7,280 IOPS

- The disk-load determines the number of disks required for the application.
- If disk has a maximum 180 IOPS for the application, then number of disks required is as follows:

RAID-5: 11,440 / 180 = 64 disks
RAID-1: 7,280 / 180 = 42 disks (approximated to the nearest even number)

1.11 RAID Comparison

RAID	MIN. DISKS	STORAGE EFFICIENCY %	COST	READ PERFORMANCE	WRITE PERFORMANCE	WRITE PENALTY
0	2	100	Low	Very good for both random and sequential read	Very good	No
1	2	50	High	Good. Better than a single disk.	Good. Slower than a single disk, as every write must be committed to all disks.	Moderate
3	3	(n-1)*100/n where n= number of disks	Moderate	Good for random reads and very good for sequential reads.	Poor to fair for small random writes. Good for large, sequential writes.	High
4	3	(n-1)*100/n where n= number of disks	Moderate	Very good for random reads. Good to very good for sequential writes.	Poor to fair for random writes. Fair to good for sequential writes.	High
5	3	(n-1)*100/n where n= number of disks	Moderate	Very good for random reads. Good for sequential reads	Fair for random writes. Slower due to parity overhead. Fair to good for sequential writes.	High
6	4	(n-2)*100/n where n = number of disks	Moderate but more than RAID 5	Very good for random reads. Good for sequential reads.	Good for small, random writes (has write penalty).	Very High
1+0 and 0+1	4	50	High	Very good	Good	Moderate

Table 1.2: Comparison of different RAID Types

INTELLIGENT STORAGE SYSTEM

1.12 Components of an Intelligent Storage System (ISS)

- ISS is a feature-rich RAID-array that provides highly optimized I/O-processing capabilities.
- To improve the performance, storage-device provides
  - → large amount of cache
  - → multiple paths (storage-system --> storage-device)
- It handles the management, allocation, and utilization of storage-capacity.
- A storage-device consists of 4 components (Figure 1-21):
  1) Front-end
  2) Cache

3) Back-end and
4) Physical-disk (or simply the disk).

- A RW-request is used for reading and writing of data from the disk.
  1) Firstly, a read-request is placed at the host.
  2) Then, the read-request is passed to front-end, then to cache and then to back-end.
  3) Finally, the read-request is passed to disk.
- A read-request can be serviced directly from cache if the requested-data is available in cache.

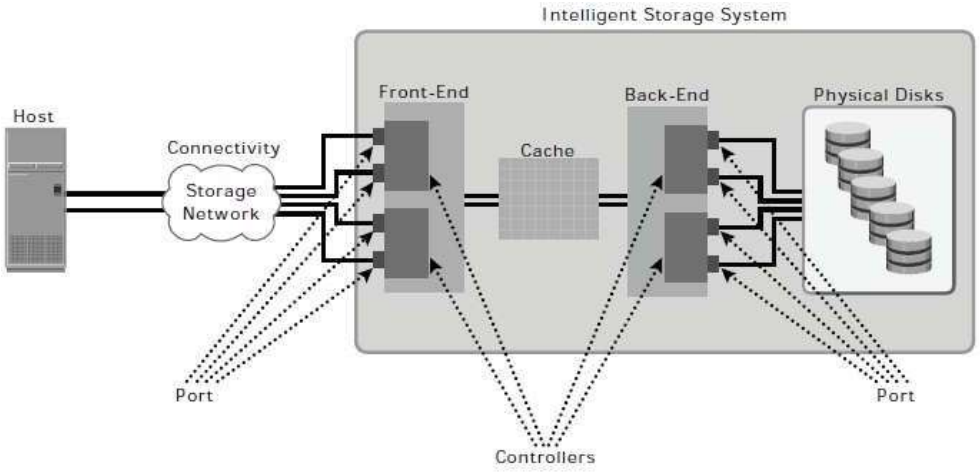

Figure 1-21: Components of an intelligent storage system

1.12.1 Front End

- Front-end provides the interface between host and storage.
- It consists of 2 components: 1) front-end port and 2) front-end controller.

1) Front-End Port

- Front-end port is used to connect the host to the storage.
- Each port has processing-logic that executes appropriate transport-protocol for storage-connections
- Transport-protocol includes SCSI, FC, iSCSI and FCoE.
- Extra-ports are provided to improve availability.

2) Front-End Port

- Front-end port
  → receives and processes I/O-requests from the host and
  → communicates with cache.
- When cache receives write-data, controller sends an acknowledgment back to the host.
- The controller optimizes I/O-processing by using command queuing algorithms.

1.12.2 Cache

- Cache is a semiconductor-memory.
- Advantages:
  1) Data is placed temporarily in cache to reduce time required to service I/O-requests from host For example:
  Reading data from cache takes less time when compared to reading data directly from disk (Analogy: Travelling from Chikmagalur to Hassan takes less time when compared to travelling from Dharwad to Hassan. Thus, we have
  Host = Hassan Cache = Chikmagalur Disk = Dharwad).
  2) Performance is improved by separating hosts from mechanical-delays associated with disks. Rotating-disks are slowest components of a storage. This is '.' of seek-time & rotational-latency

1.12.2.1 Structure of Cache

- A cache is partitioned into number of pages.
- A page is a smallest-unit of cache-memory which can be allocated (say 1 KB).
- The size of a page is determined based on the application's I/O-size.

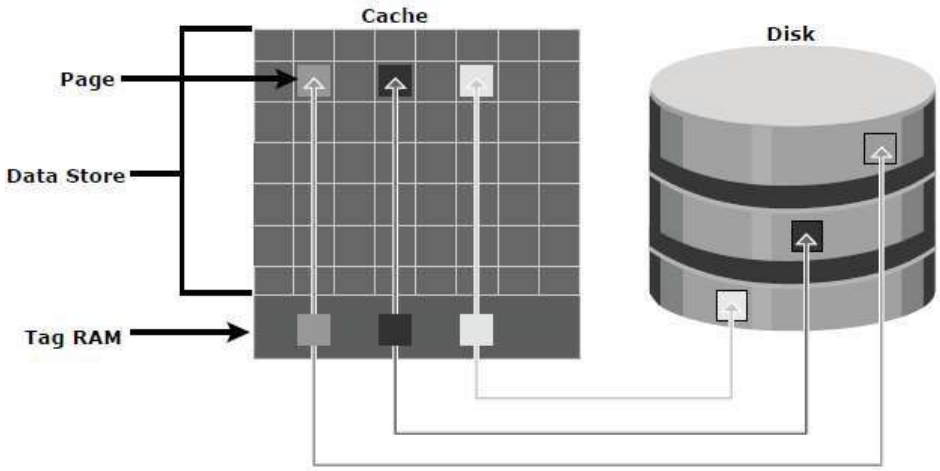

Figure 1-22: Structure of cache

- Cache consists of 2 main components (Figure 1-22):

1) Data Store

➢ Data-store is used to hold the data-transferred between host and disk.

2) Tag RAM

➢ Tag-RAM is used to track the location of the data in data-store and disk.

➢ It indicates

→ where data is found in cache and

→ where the data belongs on the disk.

➢ It also consists of i) dirty-bit flag ii) Last-access time

i) **Dirty-bit flag** indicates whether the data in cache has been committed to the disk or not.

i.e. 1 --> committed (means data copied successfully from cache to disk) 0

--> not committed

ii) **Last-access time** is used to identify cached-info that has not been accessed for a long-time

Thus, data can be removed from cache and the memory can be de-allocated.

1.12.2.2 Read Operation with Cache

- When host issues a read-request, the controller checks whether requested-data is available in cache
- A read-operation can be implemented in 3 ways:

1) Read-Hit
2) Read-Miss &
3) Read-Ahead.

1) Read-Hit

➢ Here is how it works:

1) A read-request is sent from the host to cache.

If requested-data is available in cache, it is called a **read-hit.**

2) Then, immediately the data is sent from cache to host. (Figure 1-23[a]).

➢ Advantage:

1) Provides better response-time. This is because

→ the read-operations are separated from the mechanical-delays of the disk.

2) Read-Miss

- Here is how it works:
  1) A read-request is sent from the host to cache.
     If the requested-data is not available in cache, it is called a **read-miss.**
  2) Then, the read-request is forwarded from the cache to disk.
     Now, the requested-data is read from the disk (Figure 1-23[b]).
     For this, the back-end controller
     → selects the appropriate disk and
     → retrieves the requested-data from the disk.
  3) Then, the data is sent from disk to cache.
  4) Finally, the data is forwarded from cache to host.
- Disadvantage:
  1) Provides longer response-time. This is because of the disk-operations.

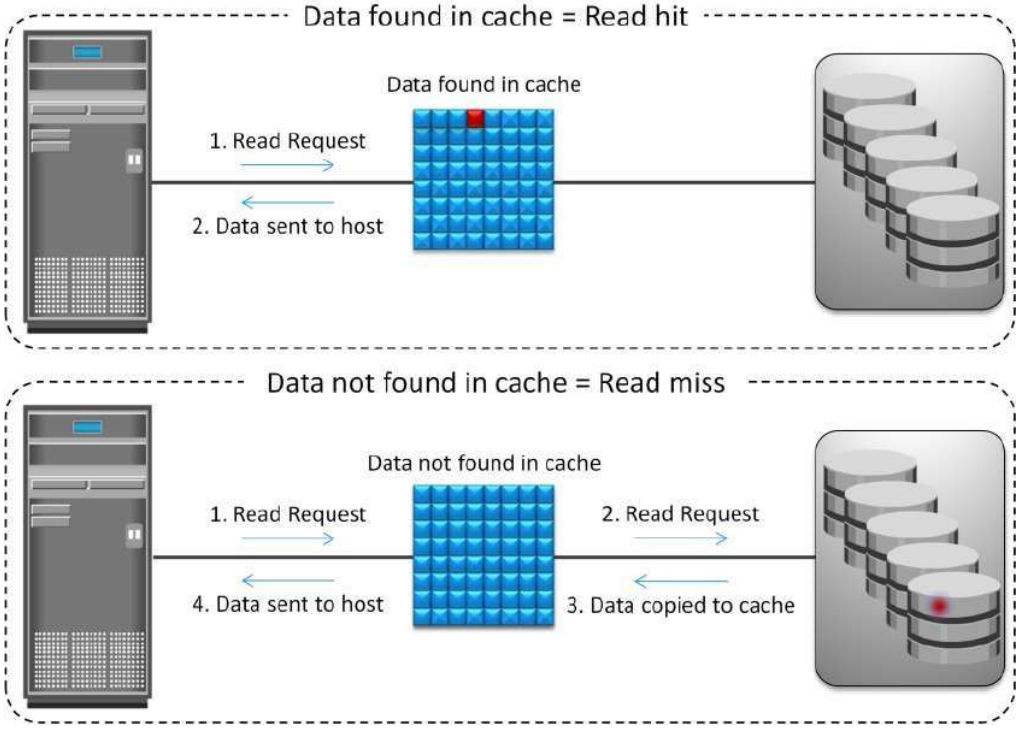

Figure 1-23: Read hit and read miss

3) Pre-Fetch (or Read-Ahead)

- A pre-fetch algorithm can be used when read-requests are sequential.
- Here is how it works:
  1) In advance, a continuous-set of data-blocks will be
     → read from the disk and
     → placed into cache.
  2) When host subsequently requests the blocks, data is immediately sent from cache to host.
- Advantage:
  1) Provides better response-time.
- The size of prefetch-data can be i) fixed or ii) variable.

i) Fixed Pre-Fetch

¤ The storage-device pre-fetches a fixed amount of data. (say 1*10 KB =10 KB).

¤ It is most suitable when I/O-sizes are uniform.

ii) Variable Pre-Fetch

¤ The storage-device pre-fetches an amount of data in multiples of size of host-request. (say 4*10 KB =40 KB)

Read-Hit-Ratio

- Read-performance is measured in terms of the read-hit-ratio (or simply hit-ratio).

$$\text{hit-ratio} = \frac{\text{number of read-hits}}{\text{number of read-requests}}$$

- A higher hit-ratio means better read-performance.

1.12.2.3 Write Operation with Cache

- Write-operation (Figure 1-24):

  Writing data to cache provides better performance when compared to writing data directly to disk.
- In other words, writing data to cache takes less time when compared to writing data directly to disk.
- Advantage:

  Sequential write-operations allow optimization. This is because

  → many smaller write-operations can be combined to provide larger data-transfer to disk via cache

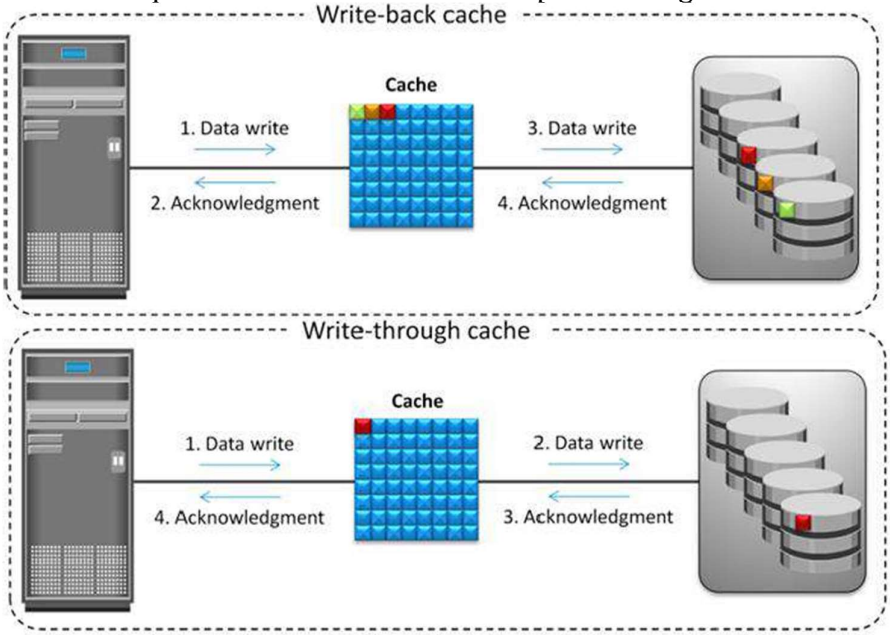

Figure 1-24: Write-back Cache and Write-through Cache

- A write-operation can be implemented in 2 ways: 1) Write-back Cache & 2) Write-through Cache

1) Write Back Cache

- Here is how it works:
  1) Firstly, a data is placed in the cache.
  2) Then, immediately an acknowledgment is sent from cache to host.
  3) Later after some time, the data is forwarded from cache to disk.
  4) Finally, an acknowledgment is sent from disk to cache.
- Advantage:
  1) Provides better response-time. This is because
     → the write-operations are separated from the mechanical-delays of the disk.
- Disadvantage:
  1) In case of cache-failure, there may be risk-of-loss of uncommitted-data.

2) Write Through Cache

- Here is how it works:
  1) Firstly, a data is placed in the cache.
  2) Then, immediately the data is forwarded from cache to disk.
  3) Then, an acknowledgment is sent from disk to cache.
  4) Finally, the acknowledgment is forwarded from cache to host.
- Advantage:
  1) Risk-of-loss is low. This is '.' data is copied from cache to disk as soon as it arrives.
- Disadvantage:
  1) Provides longer response-time. This is because of the disk-operations.

Write Aside Size

- Write-aside-size refers to maximum-size of I/O-request that can be handled by the cache.
- If size of I/O-request exceeds write-aside-size, then data is written directly to disk bypassing cache
- Advantage:
  Suitable for applications where cache-capacity is limited and cache is used for small random-requests

1.12.2.4 Cache Implementation

- Cache can be implemented as either 1) dedicated-cache or 2) global-cache.
  1) In **dedicated-cache**, separate set of memory-locations are reserved for read and write-operations
  2) In **global-cache**, same set of memory-locations can be used for both read- and write-operations.
- Advantages of global-cache:
  1) Global-cache is more efficient when compared to dedicated-cache. This is because
     → only one global-set of memory-locations has to be managed.
  2) The user can specify the percentage of cache-capacity used for read- and write-operation. (For example: 70% for read and 30% for write).

1.12.2.5 Cache Management

- Cache is a finite and expensive resource that needs proper management.
- When all cache-pages are filled, some pages have to be freed-up to accommodate new data.
- Two cache-management algorithms are:

  1) Least Recently Used (LRU)
  - Working principle: Replace the page that has not been used for the longest period of time.
  - Based on the assumption:
    data which hasn't been accessed for a while will not be requested by the host.

  2) Most Recently Used (MRU)
  - Working principle: Replace the page that has been accessed most recently.
  - Based on the assumption:
    recently accessed data may not be required for a while.
- As cache fills, storage-device must take action to flush dirty-pages to manage availability.
- A **dirty-page** refers to data written into the cache but not yet written to the disk.
- **Flushing** is the process of committing data from cache to disk.
- Based on access-rate and -pattern of I/O, watermarks are set in cache to manage flushing process.
- Watermarks can be set to either high or low level of cache-utilization.

  1) High watermark (HWM)
  - The point at which the storage-device starts high-speed flushing of cache-data.

  2) Low watermark (LWM)
  - The point at which storage-device stops high-speed flushing & returns to idle flush behavior.

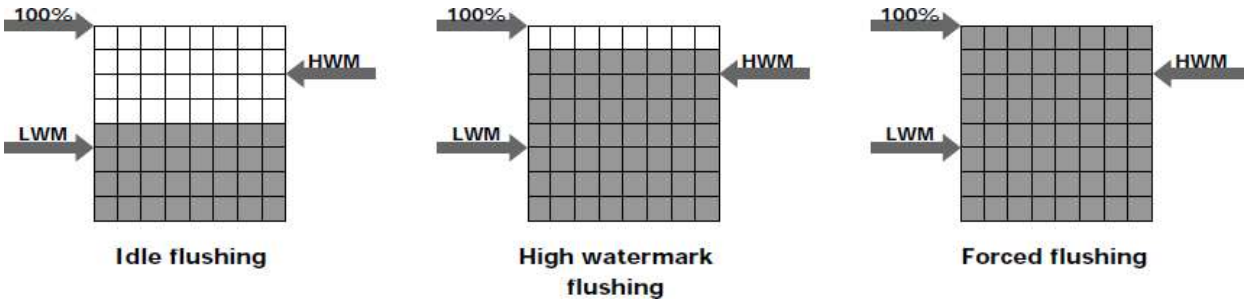

Figure 1-25: Types of flushing

- The cache-utilization level drives the mode of flushing to be used (Figure 1-25):

1) Idle Flushing occurs at a modest-rate when the level is between the high and low watermarks

2) High Watermark Flushing occurs when the cache utilization level hits the high watermark.

- Disadvantage:
  1) The storage-device dedicates some additional resources to flushing.
- Advantage:
  1) This type of flushing has minimal impact on host.

3) Forced Flushing occurs in the event of a large I/O-burst when cache reaches 100% of its capacity.

- Disadvantage:
  1) Affects the response-time.
- Advantage:
  1) The dirty-pages are forcibly flushed to disk.

1.12.2.6 Cache Data Protection

- Cache is volatile-memory, so cache-failure will cause the loss-of-data not yet committed to the disk.
- This problem can be solved in various ways:
  1) Powering the memory with a battery until AC power is restored or
  2) Using battery-power to write the cached-information to the disk.
- This problem can also be solved using following 2 techniques: 1) Cache Mirroring
  2) Cache Vaulting

1) Cache Mirroring

i) Write Operation

- Each write to cache is held in 2 different memory-locations on 2 independent memory-cards.
- In case of cache-failure, the data will be still safe in the surviving-disk.
- Hence, the data can be committed to the disk.

ii) Read Operation

- A data is read to the cache from the disk.
- In case of cache-failure, the data will be still safe in the disk.
- Hence, the data can be read from the disk.
- Advantage:
  1) As only write-operations are mirrored, this method results in better utilization of available cache
- Disadvantage:
  The problem of cache-coherency is introduced.
  Cache-coherency means data in 2 different cache-locations must be identical at all times.

2) Cache Vaulting

- It is process of dumping contents of cache into a dedicated disk during a power-failure.
- A disk used to dump the contents of cache are called **vault-disk**.

Write Operation

- When power is restored,
  → data from vault-disk is written back to the cache and

→ then data is written to the intended-disks.

1.12.3 Back End

- The back-end
  → provides an interface between cache and disk.
  → controls data-transfer between cache and disk.
- Write operation:
  From cache, data is sent to the back-end and then forwarded to the destination-disk.
- It consists of 2 components: 1) back-end ports and 2) back-end controllers.

1) Back End Ports

- Back End Ports is used to connect the disk to the cache.

2) Back End Controllers

- Back End Controllers is used to route data to and from cache via internal data-bus.

- The controller
  → communicates with the disks when performing read- and write-operations and
  → provides small temporary data-storage.
- The controllers provides
  → error-detection and -correction (e.g. parity)
  → RAID-functionality.

Dual Controller

- To improve availability, storage-device can be configured with dual-controllers with multiple-ports In case of a port-failure, controller provides an alternative path to disks.
- Advantage:
  1) Dual-controllers also facilitate load-balancing.

Dual Port Disk

- The availability can be further improved if the disks are also dual-ported.
  In this case, each disk-port can be connected to a separate controller.

1.12.4 Physical Disk

- A disk is used to store data persistently for future-use.
- Disks are connected to the back-end using SCSI or FC.
- Modern storage-devices provide support for different type of disks with different speeds.
- Different type of disks are: FC, SATA, SAS and flash drives (pen drive).
- It also supports the use of a combination of flash, FC, or SATA.

1.13 Storage Provisioning

- It is process of assigning storage-capacity to hosts based on performance-requirements of the hosts.
- It can be implemented in two ways: 1) traditional and 2) virtual.

1.13.1 Traditional Storage Provisioning

1.13.1.1 Logical Unit (LUN)

- The available capacity of RAID-set is partitioned into volumes known as **logical-units (LUNs)**.
- The logical-units are assigned to the host based on their storage-requirements.
- For example (Figure 1-26)
  LUNs 0 and 1 are used by hosts 1 and 2 for accessing the data.
- LUNs are spread across all the disks that belong to that set.
- Each logical-unit is assigned a unique ID called a **logical-unit number (LUN#)**.
- Advantages:
  1) LUNs hide the organization and composition of the set from the hosts.
  2) The use of LUNs improves disk-utilization. For example,
     i) Without using LUNs, a host requiring only 200 GB will be allocated an entire 1 TB disk.

ii) With using LUNs, only the required 200 GB will be allocated to the host. This allows the remaining 800 GB to be allocated to other hosts.

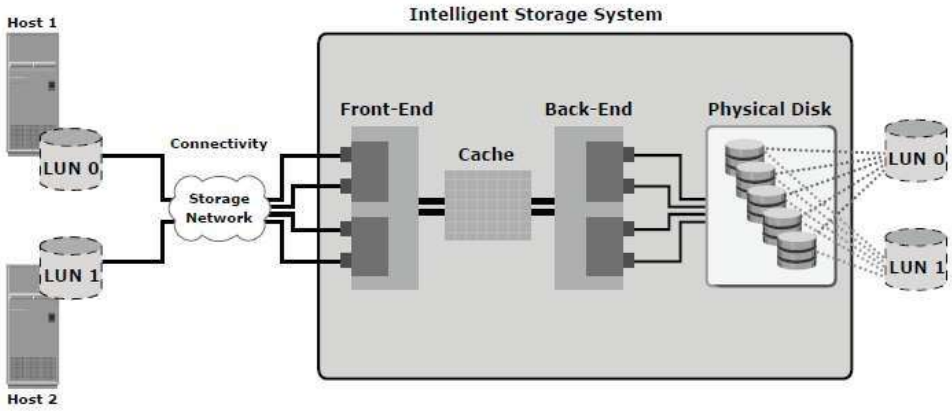

Figure 1-26: Logical-unit number

1.13.1.1.1 LUN Expansion: MetaLUN

- MetaLUN is a method to expand logical-units that require additional capacity or performance.
- It can be created by combining two or more logical-units (LUNs).
- It consists of
  - i) base-LUN and
  - ii) one or more component-LUNs.
- It can be either concatenated or striped (Figure 1-27).

1) Concatenated MetaLUN

- The expansion adds additional capacity to the base-LUN.
- The component-LUNs need not have the same capacity as the base-LUN.
- All LUNs must be either protected (parity or mirrored) or unprotected (RAID 0). For example, a RAID-0 LUN can be concatenated with a RAID-5 LUN.
- Advantage:
  1) The expansion is quick.
- Disadvantage:
  1) Does not provide any performance-benefit.

2) Striped MetaLUN

- The expansion restripes the data across the base-LUN and component-LUNs.
- All LUNs must have same capacity and same RAID-level.
- Advantage:
  1) Expansion provides improved performance due to the increased no. of disks being striped

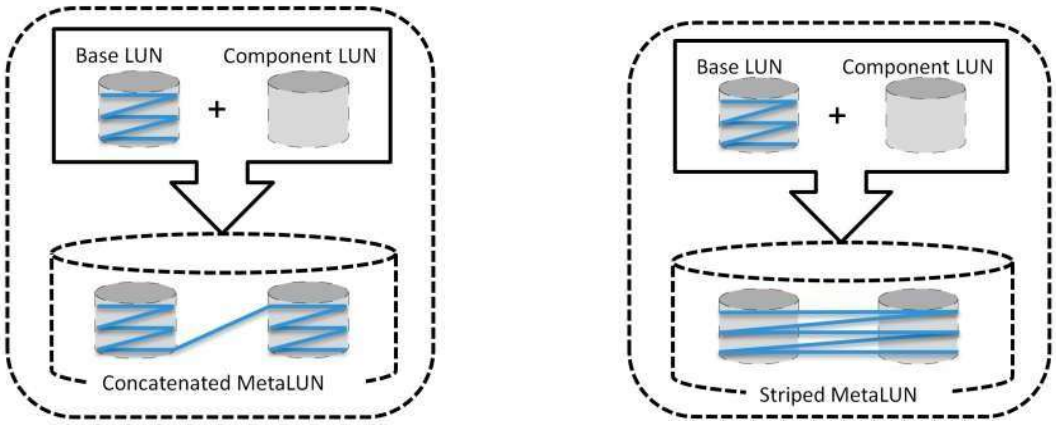

Figure 1-27: LUN Expansion

- Advantages of traditional storage-provisioning:
  1) Suitable for applications that require predictable performance.
  2) Provides full control for precise data-placement.
  3) Allows admins to create logical-units on different RAID-groups if there is any workload-contention

1.13.2 Virtual Storage Provisioning

- Virtual-provisioning uses virtualization technology for providing storage for applications.
- Logical-units created using virtual-provisioning is called thin-LUN to distinguish from traditional LUN.
- A host need not be completely allocated a storage when thin-LUN is created.
- Storage is allocated to the host "on-demand" from a shared-pool.
- A shared-pool refers to a group of disks.
- Shared-pool can be
  → homogeneous (containing a single drive type) or
  → heterogeneous (containing mixed drive types, such as flash, FC, SAS, and SATA drives).
- Advantages:
  1) Suitable for applications where space-consumption is difficult to forecast.
  2) Improves utilization of storage-space.
  3) Simplifies storage-management.
  4) Enables oversubscription.
     Here, more capacity is presented to the hosts than actually available on the storage-array
  5) Scalable:
     Both shared-pool and thin-LUN can be expanded, as storage-requirements of the hosts grow
  6) Sharing:
     Multiple shared-pools can be created within a storage-array. A shared-pool may be shared by multiple thin LUNs.
  7) Companies save the costs associated with acquisition of new storage.

1.14 Types of Intelligent Storage System

- Storage-devices can be classified into 2 types:
  1) High-end storage-system
  2) Midrange storage-system

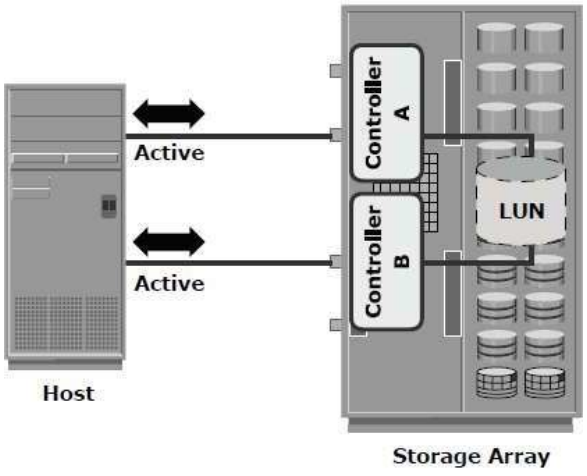

Figure 1-28: Active-active configuration

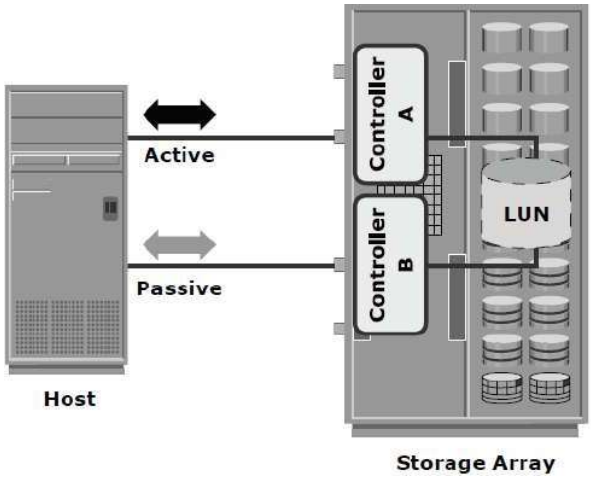

Figure 1-29: Active-passive configuration

1.14.1 High End Storage System

- High-end storage-device is also known as **active-active array**.
- It is suitable for large companies for centralizing corporate-data. (e.g. Big Bazaar)
- An active-active array implies that
  → the host can transfer the data to its logical-units using any of the available paths (Figure 1-28)
- It provides the following capabilities:
  1) Large number of controllers and cache.
  2) Multiple front-end ports to serve a large number of hosts.
  3) Multiple back-end controllers to perform disk-operations optimally. The controller includes FC and SCSI RAID.
  4) Large storage-capacity.
  5) Large cache-capacity to service host's requests optimally.
  6) Mirroring technique to improve data-availability.
  7) Interoperability: Connectivity to mainframe-computers and open-systems hosts.
  8) Scalability to support following requirements:
     → increased connectivity
     → increased performance and
     → increased storage-capacity
  9) Ability to handle large amount of concurrent-requests from a no. of servers and applications.
  10) Support for array-based local- and remote-replication.
  11) Suitable for mission-critical application (like military).

1.14.2 Midrange Storage System

- Midrange storage-device is also referred to as **active-passive array**.
- It is suitable for small- and medium-sized companies for centralizing corporate-data.
- It is designed with 2 controllers.
- Each controller contains
  → host-interfaces

→ cache
→ RAID-controllers, and
→ interface to disks.

- It provides the following capabilities:
  1) Small storage-capacity
  2) Small cache-capacity to service host-requests
  3) Provides fewer front-end ports to serve a small number of hosts.
  4) Ensures high-availability and high-performance for applications with predictable workloads.
  5) Supports array-based local- and remote-replication.
  6) Provides optimal storage-solutions at a lower cost.
- In an active-passive array,
  1) A host can transfer data to its logical-units only through the path owned by the controller. These paths are called **active-paths**.
  2) The other paths are passive with respect to this logical-units. These paths are called **passive-paths**.
- As shown in Figure 1-29,
  - The host can transfer the data to its LUNs only through the path owned by the controller-A. This is because controller-A is the owner of this LUN.
  - The path to controller-B remains passive and no data-transfer is performed through this path.