Module-1: Introduction Storage Systems

Syllabus: Storage System Introduction to Information Storage: Evolution of Storage Architecture, Data Center Infrastructure, Virtualization and Cloud Computing. Data Center Environment: Application, Host (Compute), Connectivity, Storage. Data Protection: RAID: RAID Implementation Methods, RAID Techniques, RAID Levels, RAID Impact on Disk Performance. Intelligent Storage Systems: Components of Intelligent Storage System, Storage Provisioning.

Text Book-1 Ch1: 1.2 to 1.4, Ch2: 2.1, 2.3 to 2.5, Ch3: 3.1, 3.3 to 3.5, Ch4: 4.1 and 4.2

Contents

Chapter 1
Introduction to Information Storage

Introduction

Information is increasingly important in our daily lives. We have become information dependents of the twenty-first century, living in an on-command, on-demand world that means need information when and where it is required.

Access the Internet every day to perform searches, participate in social networking, send and receive e-mails, share pictures and videos, and scores of other applications. Equipped with a growing number of content-generating devices, more information is being created by individuals than by businesses.

1.1 Information Storage

Organizations process data to derive the information required for their day-today operations. Storage is a repository that enables users to persistently store and retrieve this 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
  - → Student/Employee details
  - → Movie on video-tape
- The data can be generated **using a computer** and *stored in strings of binary numbers 0s and 1s*
- Data in 0s/1s form is called **digital-data.** (Figure 1-1).
- Digital-data is accessible by the user only after it is processed by a computer

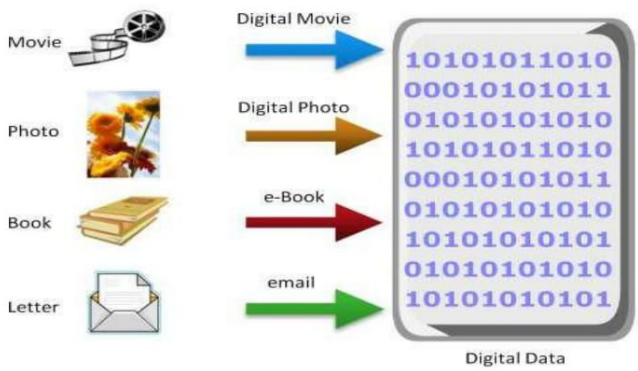


Figure 1-1: Digital data

- With the advancement of *computer* and *communication technologies*, the rate of data generation and sharing has increased exponentially.
- The following is a list of ***factors that have contributed to the growth of digital data:***

1) Increase in Data Processing Capabilities

- Modern computers provide a significant increase in data-processing and storage capabilities.
- This allows the conversion of various types of data (like book, photo or video) from conventional forms to digital-formats.

2) Lower Cost of Digital Storage

- With the advancement in technology, the cost of storage-devices has decreased; which provided the low-cost solutions and encouraged the development of less expensive data storage devices
- 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) Proliferation (Increase) of Smart Devices and Applications

- Proliferation of applications and smart devices: Smartphones, tablets, and newer digital devices, along with smart applications, have significantly contributed to the generation of digital content.

1.1.2 Types of Data

Data can be classified based on how it is stored and managed. There are 2 types

1. Structured
2. Unstructured (see Figure 1-3)

1. **Structured data** is organized in rows and columns (Table) format.
   - Applications can retrieve and process it efficiently.
   - Structured data is typically stored using a database management system (DBMS).
   - Example: Employee Database(excel/DB)
2. **Unstructured Data** can't be stored in rows and columns (Table)
   - Business applications find it difficult to query and retrieve data.
   - For example, customer contacts may be stored in various forms such as sticky notes, e-mail messages, business cards, or even digital format files such as .doc, .txt, and .pdf.
   - X-Rays, Images, Web Pages, Audio Video

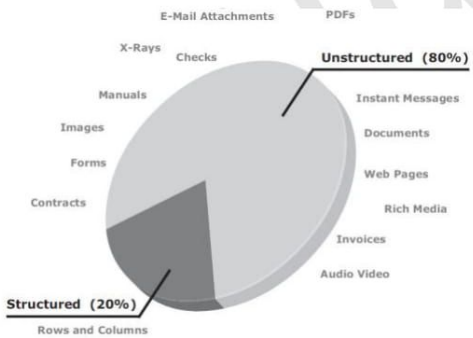


Figure 1-3: Types of data

1.1.3 Big Data

• It refers to data-sets whose sizes are beyond the capability of commonly used software tools to capture, store, manage and process 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.

The big data ecosystem consists the following:

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.

1.1.4 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.

Example 1, a **retailer** identifies **customers' preferred products** and **brand names** by analyzing their *purchase patterns* and *maintaining an inventory* of those products.

Effective data analysis not only extends its benefits to existing businesses, but also creates the potential for new business opportunities by using the information in creative ways.

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/digital camera, DVDs, CD-ROMs & hard-disks in computers.

1.2 Evolution of Storage Architecture

****** Explain the evolution of storage Architecture with a neat diagram ******

- In earlier days, organizations had 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 (Fig 1-3)

→ Evolution of open-systems

→ Affordability of open-systems and

→ Easy deployment of open-systems.

1. Server Centric Storage Architecture

Organizations have their own servers running the business applications. Storage devices are connected directly to the servers and are typically internal to the server. (Figure 1-3 (a))

- 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 → information-centric architecture.

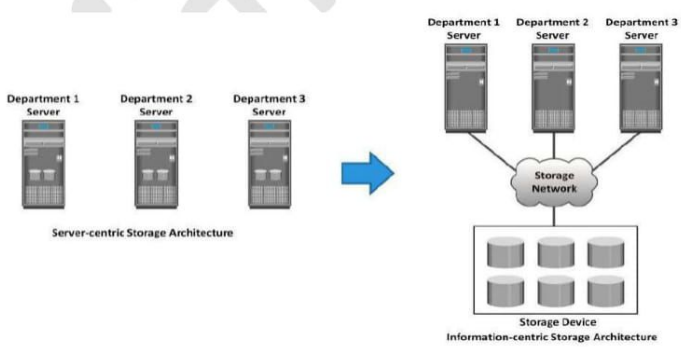


Figure 1-3: Evolution of storage architectures

2. Information Centric Architecture

- Storage is managed centrally and independent of servers. (Figure 1-3 (b))
- Storage is allocated to the servers "on-demand"
- Centrally managed stored devices are shared with 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:**
  - ✓ Information management is easier and cost-effective.
  - ✓ Storage technology even today continues to evolve.
  - ✓ Enables companies to consolidate & leverage their data to achieve highest return on information assets

1.3 Data Center Infrastructure

******* What is data center? List and explain the core elements of Data Center, explain online order transaction system with a diagram. *******

Organizations maintain data centers to provide centralized data processing capabilities across the enterprise. Data centers store and manage large amounts of mission-critical data.

The data center infrastructure includes **1) computers, 2) storage systems, 3) network devices, 4) dedicated power backups, 5) and environmental controls (such as air conditioning and fire suppression).**

1.3.1 Core Elements of Data Center

Five core-elements of a data-center:

1) Application: An application is a computer 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.

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 (hardware, firmware &software) 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 1-5 shows an order processing system that involves the five core elements of a data center and illustrates their functionality in a business process.

Step 1: A customer places an order through the AUI (Application User Interface Disk) of the order processing application software located on the client computer.

Step 2: The client connects to the server over the LAN and accesses the DBMS located on the server to update the relevant information such as the customer name, address, payment method, products ordered, and quantity ordered.

Step 3: The DBMS uses the server operating system to read and write this data to the database located on physical disks in the storage array.

Step 4: The Storage Network

→ provides the communication link between the server and the storage array and

→ transports the read or write commands between them.

Step 5: The storage array, after receiving the read or write commands from the server, store the data on physical disks.

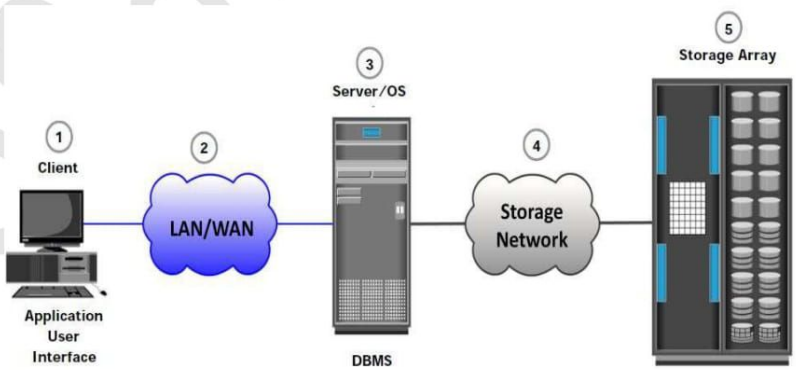


Figure 1-4: Example of an order processing-application

1.3.2 Key Requirements for Data Center Elements

******* Discuss the key characteristics of a data center, with a neat diagram *******

Are as follows (Figure 1-6).

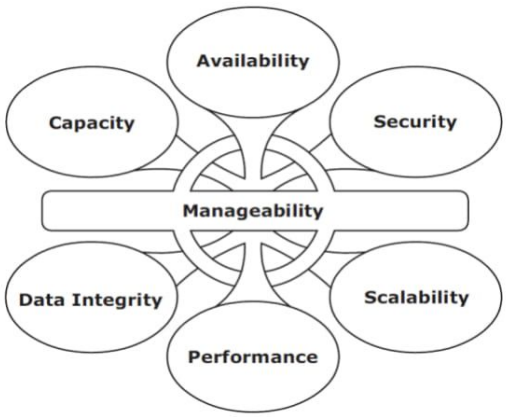


Figure 1-6: Key characteristics of data center elements

1) Availability

- In data-center, all core-elements must be designed to ensure availability.
- Data center should ensure the availability of information when required.
- If the users cannot access the data in time, then it will have negative impact on the company. Unavailability of information could cost millions of dollars per hour to businesses. (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/information,
  - Good polices & procedures must be used.
  - 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 include 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.
- Data center resources should scale based on requirements.

4) Performance

- All core-elements must be able to
  - Provide optimal-performance based on the required service level.
  - 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 stored and retrieved from disk exactly as it was received.
- For example: Parity-bit or ECC (error correction code).

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 or availability.
- Capacity must be managed by reallocation of existing-resources or by adding new resources

7) Manageability

- A data-center must perform all operations and activities in the most efficient manner
- Easy and integrated management of all its elements.
- Manageability is achieved through automation and 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**

- It is a continuous process of gathering information on various elements and services running in a datacenter.
- Following parameters are monitored:

i) Security

ii) Performance

iii) Accessibility

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 a process of providing hardware, software & other resources needed to run a datacenter.
- Resource management is done to meet capacity, availability, performance and security requirements.
- Main tasks are: i) *Capacity Planning* (future needs of both user & application will be addressed in most cost-effective way) ii) *Resource Planning* [process of evaluating & identifying required resources such as: Personnel (employees), Facility (site/plant), Technology (Artificial Intelligence, Deep Learning).]

1.4 Virtualization and Cloud Computing

Virtualization

- Virtualization is a technique of abstracting physical resource (Compute, Storage, Network) and appear as logical-resource.
- Virtualization existed in the IT-industry for several years in different forms.
- Virtualization enables
    - Pooling of resources, providing an aggregated view of the resource capabilities.

- *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).

- 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 (10GB) can be created from a storage-pool (100TB) and 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).

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/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.

Chapter 2

DATA CENTER ENVIRONMENT

The data flows from an application to storage through various components collectively referred as a *storage system environment.*

The three main components in this environment are the **1. Host 2. Connectivity 3. Storage.** These entities, along with their physical and logical components, facilitate data access.

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.

2.1 Components of a Storage System Environment

******** ***Explain Data Center Environment*** ********

Application

• An application is a computer 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 devices.

• 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: Example: e-mail

→ infrastructure management applications, Ex:enterprise resource planning (ERP), decision support system(DSS)

→ data protection applications, Ex: Resource Management, Backup

→ security applications. Ex: Authentication and antivirus applications.

Database Management System (DBMS)

A database is a structured way to store data in logically organized tables that are interrelated.

- The DBMS processes an application's request for data and instructs the OS to transfer the appropriate data from the storage.
    1) Helps to optimize the storage and retrieval of data.
    2) Controls the creation, maintenance and use of a database.

1. Host

- The computers on which these applications run are referred to as host or compute system.
- Users store and retrieve data through applications.
- Hosts can range from simple laptops, mobiles to complex clusters of servers.
- Physical Components
- A host has three key physical components:
    i) Central processing unit (CPU)
    ii) Storage, such as internal memory and disk devices
    iii) Input/output (I/O) devices

CPU: The CPU consists of four main components:

■ Arithmetic Logic Unit (ALU), Control Unit, Register, Level 1 (L1) cache

Storage There are two types of memory on a host:

■ Random Access Memory (RAM), Read-Only Memory (ROM)

I/O Devices: I/O devices enable sending and receiving data to and from a host. This communication may be one of the following types:

■ ***User to host communications:*** Handled by basic I/O devices, such as the keyboard, mouse, and monitor. These devices enable users to enter data and view the results of operations.

■ ***Host to host communications***: Enabled using devices such as a Network Interface Card (NIC) or modem.

■ ***Host to storage device communications:*** Handled by a Host Bus Adaptor (HBA). HBA is an application-specific integrated circuit (ASIC) board that performs I/O interface functions between the host and the storage, relieving the CPU from additional I/O processing workload.

**** *Explain logical components of host.* ****

Software.

The software includes

i) OS

ii) Device Drivers

iii) Logical volume manager (LVM)

iv) File System

v) Compute Virtualization

i) Operating System

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

Memory Virtualization

- Memory-virtualization is used to virtualize the physical-memory (RAM) of a host.
- It creates a Virtual Memory(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**
- The VMM (**virtual-memory manager)**

→ 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 called paging.
  - A paging

    → Moves inactive-pages onto the swap-file and

    → Brings inactive-pages back to the physical-memory when required.
  - Advantages:

    ¤ 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. Because → swap-file pages are allocated on the disk which is slower than physical-memory.

ii) Device Driver

- A device driver is special software that permits the operating system to interact with a specific device, such as a printer, a mouse, or a disk drive.
- 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

iii) Logical Volume Manager (LVM)

****** Explain LVM with a neat diagram. Explain Aggregation and Partition. Define physical**

volume, LVM, Logical Volume****

- The evolution of Logical Volume Managers (LVMs) enabled dynamic extension of file system capacity and efficient storage management.
- 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:
  - Partition a larger-capacity disk into virtual, smaller-capacity volumes (called **partitioning**) or aggregate several smaller disks to form a larger virtual volume. (**concatenation**.) These volumes are then presented to applications.
  - Disk partitioning was introduced to improve the flexibility and utilization of disk drives.
  - Provides optimized storage-access and simplifies storage-management.
  - Hides details about disk and location of data on the disk.
  - Enables admins to change the storage-allocation without interrupting normal-operations.
  - 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. These components are described in the following diagram

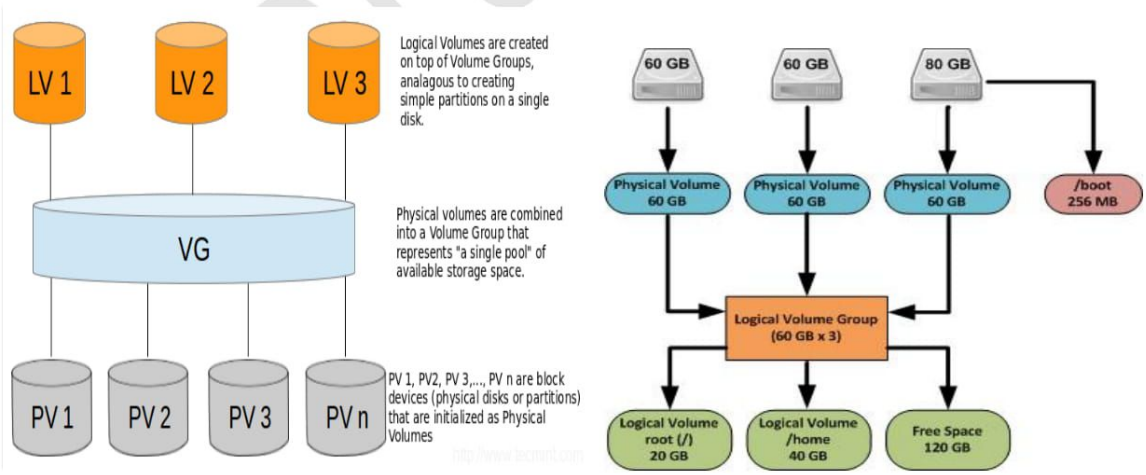


Figure: Components of LVM

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

2) **LVM :** Converts the physical storage provided by the physical volumes to a logical view of storage, which is then used by the OS and applications.

3) **Volume-Group (VG):** Refers to a group of one or more PVs.

- Combining of multiple individual hard drives and/or disk partitions into a single **volume** group (VG)
- A unique PVID (physical volume identifier) 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**.

4) **Logical-Volume (LV):** Refers to a partition within a volume-group.

- Large physical drive can be portioned into multiple LVs to maintain data according to the file system and application requirements.
- Logical-volumes V/S Volume-group
  - i) LV can be thought of as a disk-partition.
  - ii) Volume-group can be thought of as a disk.

- 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.

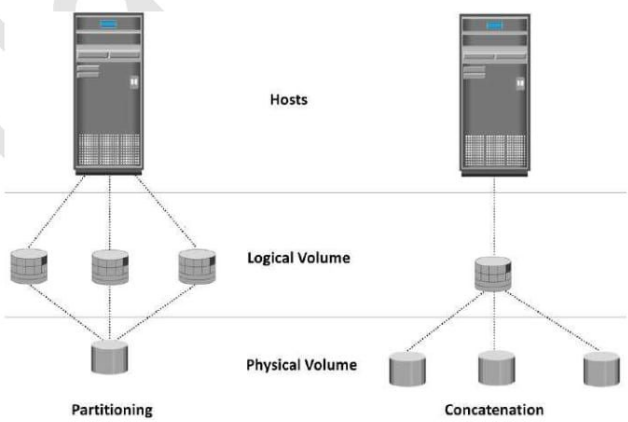


Figure 1-6: Disk partitioning and concatenation

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

1. **Partitioning**
   - A larger-capacity disk drive is partitioned (divided) into smaller-capacity virtual-disks called **logical volumes (LVs).**
   - Disk-partitioning is used to improve the utilization of disks.
2. **Concatenation**
   - Process of grouping several smaller-capacity physical disks are aggregated (grouping) to form a larger-capacity virtual-disk (Logical volume).
   - The larger-capacity virtual-disk is presented to the host as one big logical-volume.

iv) 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
- 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 shows process of mapping user-files to the disk-storage with an LVM:
  - Files are created and managed by users and applications.

- These files reside in the file-system.
- The file-system are mapped to file-system blocks.
- The file-system blocks are mapped to logical-extents.
- The logical-extents are mapped to disk physical-extents by OS or LVM.
- Finally, these physical-extents are mapped to the disk-storage.

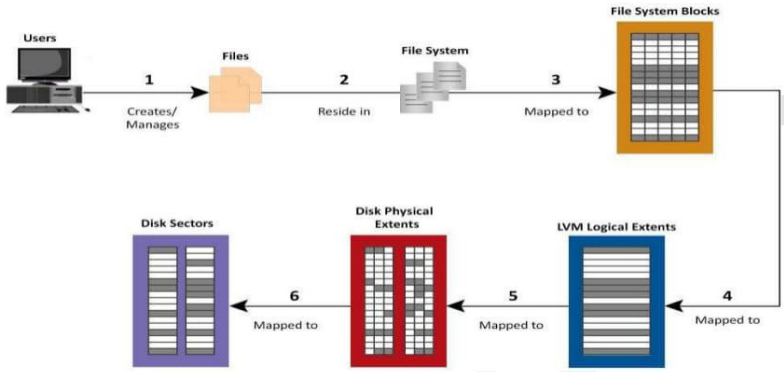


Figure 1-7: Process of mapping user files to disk storage

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).
- It enables multiple operating systems to run concurrently on single or clustered physical machines. This technique enables creating portable virtual compute systems called virtual machines (VMs).
- Compute-virtualization is done by virtualization-layer called as **hypervisor**.
- 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
- 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.
- A virtual machine is a logical entity but appears like a physical host to the operating system,

with its own CPU, memory, network controller, and disks.

- In **server virtualization** servers are limited to serve only one application at a time (Fig- 2-3 (a).
- **Disadvantages:**
  - ***Expensive and inflexible infrastructure:*** Organizations purchase new physical machines for every application they deploy.
  - ***Underutilization of resources:*** Many applications do not take full advantage of the hardware capabilities available to them. resources such as processors, memory, and storage remain underutilized.
- **Compute virtualization** enables multiple operating systems and applications to run on a single physical machine. This technique significantly improves server utilization and provides server consolidation.
- **Advantages:**
  - ✓ Allows multiple-OS and applications to run concurrently on a single-computer.
  - ✓ Improves server-utilization. And 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.
  - ✓ VM can be created in less time when compared to setting up the actual server.
  - ✓ VM can be restarted or upgraded without interrupting normal-operations.
  - ✓ VM can be moved from one computer to another w/o interrupting normal-operations.

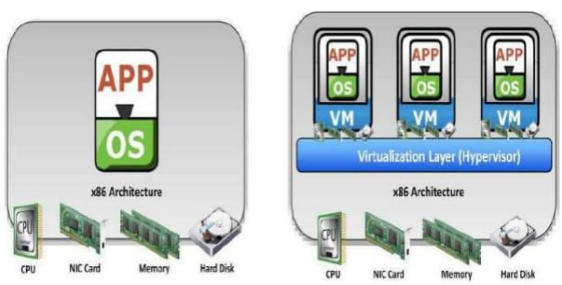


Figure 1-8: Server virtualization

2.2. Connectivity

****** With a neat diagram explain the different components of connectivity ******

Connectivity refers to the interconnection between **hosts or between a host and any other peripheral devices**, such as ***printers or storage devices***.

The components of connectivity in a storage system environment can be classified as **physical and logical.**

A. ***Physical components*** are the hardware elements that connect the host to storage

B. ***Logical components*** of connectivity are the protocols used for communication between the host and storage.

Physical Components of Connectivity

The three physical components of connectivity between the host and storage are host interface device/host adapter, Port, and Cable.

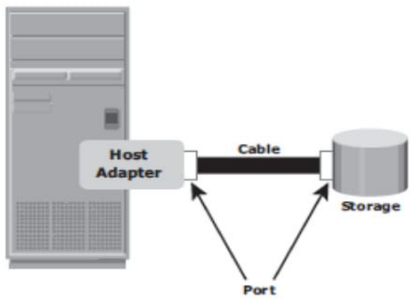


Figure 2-4: Physical components of connectivity

1. A **host interface device** or host adapter connects a host to other hosts and storage devices. **Examples**: host bus adapter (HBA), network interface card (NIC).

- HBA is an (application-specific integrated circuit) ASIC board that performs I/O-operations between host and storage.
    - Advantage:
        - HBA relieves the CPU from additional I/O-processing workload.

2. The ***port*** is a specialized outlet that enables connectivity between the host and external devices.

- 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. ***Cables*** connect hosts to internal or external devices using copper or fiber optic media.

Physical components communicate across a bus by sending bits (control, data, and address) of data between devices.

These bits are transmitted through the bus in either of the following ways:

■ **Serially:** Bits are transmitted sequentially along a single path. This transmission can be unidirectional or bidirectional.

■ **In parallel:** Bits are transmitted along multiple paths simultaneously. Parallel can also be bidirectional.

Logical Components of Connectivity: **Interface Protocol**

****** List and explain the Interface Protocols ******

- Interface-Protocol enables communication between host and storage.
- Protocols are implemented using interface-devices (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).

PCI

- PCI is a specification that standardizes how PCI expansion cards, such as network cards or modems, exchange information with the CPU.
- PCI provides the interconnection between the CPU and attached devices.
- The plug-and-play functionality of PCI enables the host to easily recognize and configure new cards and devices.

1. IDE/ATA

- IDE/ATA is the most popular interface protocol used on modern disks. This protocol offers excellent performance at relatively low cost.
- 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) Serial-ATA (SATA) supports single bit serial-transmission. .
    3) SATA version 3.0 supports a data-transfer rate up to 6 Gbps.

2. SCSI

SCSI 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) SAS (Serial Attached SCSI) is a point-to-point serial protocol.
    3) SAS version 2.0 supports a data-transfer rate up to 6 Gbps.

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.

4. Internet Protocol (IP)

- It is a protocol used for communicating data across a packet-switched network. It has been traditionally used for host-to-host traffic.
- 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 use IP for host-to-storage communication: 1) iSCSI and 2) FCIP

2.3. Storage

****** Explain the different storage devices with an example ******

A storage device uses magnetic, optic solid state media.

1. **Disks, tapes, and diskettes** use **magnetic media**.
2. **CD-ROM** is an example of a storage device that uses **optical media**
3. **Removable flash memory card** is an example of **solid-state media**.

1. ***Tapes*** are a popular storage media used for backup because of their relatively low cost.

Tape has the following limitations:

■ 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 are not suitable for applications that require real-time access to data.

■ In a shared computing environment, data stored on tape cannot be accessed by multiple applications simultaneously, restricting its use to one application at a time.

■On a tape drive, the read/write head touches the tape surface, so the tape degrades or wears out after repeated use.

■ The storage and retrieval requirements of data from tape and the overhead associated with managing tape media are significant.

2. ***Optical disk storage*** 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.

Chapter 3

DATA PROTECTION: RAID

Introduction

In the late 1980s, data was stored on a single large, expensive disk drive called ***Single Large Expensive Drive (SLED)***. Use of single disks could not meet the required performance levels, due to their limitations.

RAID is an enabling technology that leverages multiple disks as part of a set, which provides data protection against HDD failures. In general, RAID implementations also improve the I/O performance of storage systems by storing data across multiple HDDs.

- 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.

Implementation of RAID

There are two types of RAID implementation, hardware and software. Both have their merits and demerits.

********** **Explain the 2 types how the RAID can be implemented** **********

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.

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:** Reduces the overall cost of the system.
- **Disadvantage:** 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.

Key functions of RAID controllers are:

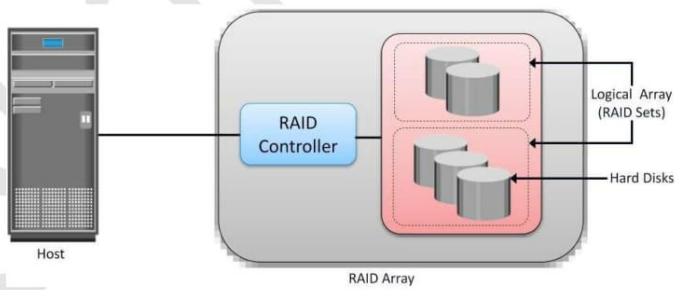
1. Management and control of disk aggregations
2. Translation of I/O requests between logical disks and physical disks
3. Data regeneration in the event of disk failures.

RAID Array Components

******* Explain with a neat diagram RAID array Components *********

A RAID array is an enclosure that contains a number of HDDs and the supporting hardware and software to implement RAID. RAID array components are shown in the below figure

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



HDDs inside a RAID array are contained in smaller sub-enclosures. These sub-enclosures, or **physical arrays**, hold a fixed number of HDDs, and also include other supporting hardware, such as power supplies.

- 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

RAID Techniques

******** **Illustrate 3 different RAID techniques with a suitable diagram** *********

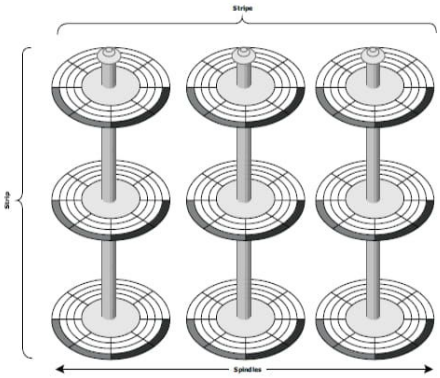
- RAID-levels are defined based on following 3 techniques:
  1) Striping (used to improve performance of storage)
  2) Mirroring (used to improve data-availability) and
  3) Parity (used to provide data-protection)
- The above techniques determine
  → performance of storage-device (i.e. better performance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

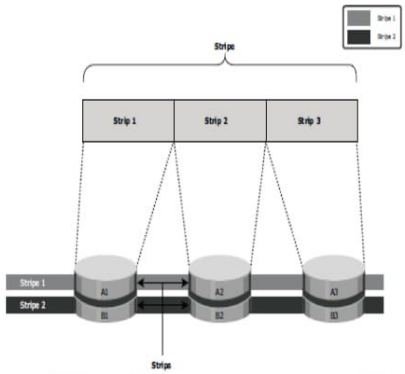
Some RAID-arrays use a combination of above 3 techniques. For example: Striping with mirroring, Striping with parity

1. Stripping

- 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. 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** (***stripe depth***) **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. Striping significantly improves I/O performance.
- **Advantage:** As number of disks increases, the performance also increases. 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)





Mirroring

- ***Mirroring*** is a technique whereby data is stored on two different HDDs, yielding two copies of data. In the event of one HDD failure, the data is intact on the surviving HDD
- 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.

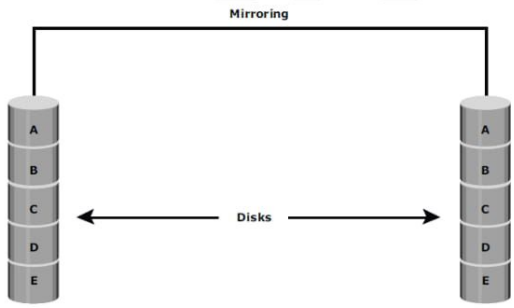


Figure 3-3: Mirrored disks in an array

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:

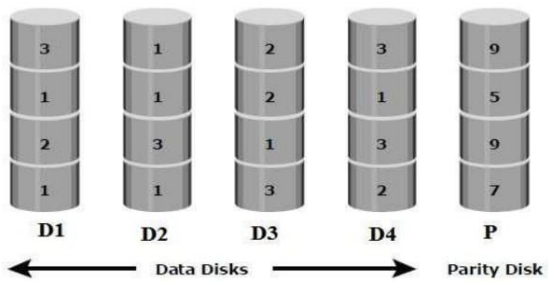
1) Decreases write-performance because
   → each write-request must perform 2 write-operations on the disks.
2) Duplication of data. Thus, amount of storage-capacity needed is twice amount of data being stored (E.g. To store 100GB data, 200GB disk is needed).
3) Considered expensive and preferred for mission-critical applications (like military application).
4) 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.
   3) Mirroring constantly captures changes in the data, whereas a backup captures point-in-time images of data.

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 / distributed across all the disks.

The computation of parity is represented as a simple arithmetic operation on the data. Parity calculation is a *bitwise XOR* operation. Calculation of parity is a function of the RAID controller. 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

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

- 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,
     - ✓ Parity requires 25% extra disk-space. (i.e. 500GB disk for 400GB data).
     - ✓ Mirroring requires 100% extra disk-space. (i.e. 800GB disk for 400GB data).
- Disadvantage: Decrease performance of storage-device: 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.

RAID Levels

RAID levels are defined on the basis of **1) striping**, **2) mirroring**, and **3) parity** techniques. These techniques determine the data availability and performance characteristics of an array. Some RAID arrays use one technique, whereas others use a combination of techniques.

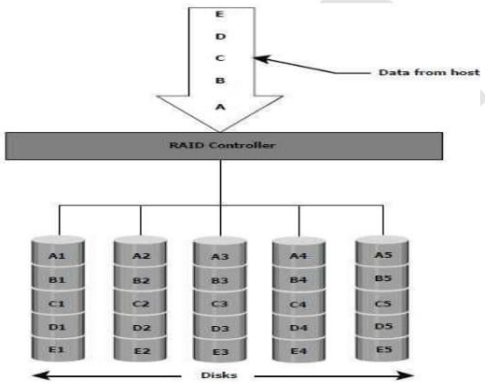
LEVELS	BRIEF DESCRIPTION
RAID 0	Striped array with no fault tolerance
RAID 1	Disk mirroring
RAID 3	Parallel access array with dedicated parity disk
RAID 4	Striped array with independent disks and a dedicated parity disk
RAID 5	Striped array with independent disks and distributed parity
RAID 6	Striped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

********** Explain different RAID levels **********

RAID 0

- RAID-0 is based on striping-technique. Striping is used to improve performance of a storage-device.
- In a RAID 0 configuration, data is striped across the HDDs in a RAID set.
- 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:

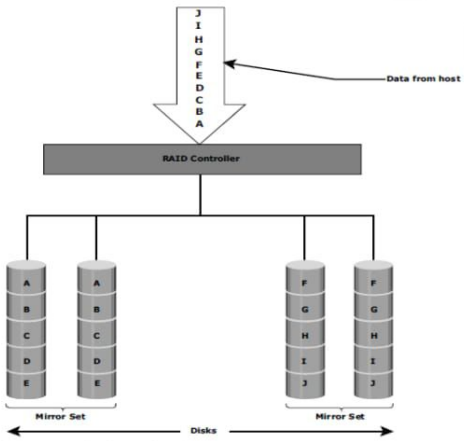
  1) Does not provide data-protection and data-availability in case of disk-failure.

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 (military application)

Nested RAID

Most data centers require data redundancy and performance from their RAID arrays.

RAID 0+1 and **RAID 1+0** combine the **performance benefits of RAID 0** with the **redundancy benefits of RAID 1**.

They use striping and mirroring techniques and combine their benefits. These types of RAID require an even number of disks, the minimum being four (see Figure 3-7). It requires an even-number of disks. Minimum no. of disks = 4.

RAID 1+0 is also known as RAID 10 (Ten) or RAID 1/0. Similarly, RAID 0+1 is also known as RAID 01 or RAID 0/1.

RAID 1+0 performs well for workloads that use small, random, write-intensive I/O.

Some applications that benefit from RAID 1+0 include the following:

1. High transaction rate Online Transaction Processing (OLTP)
2. Large messaging installations
3. Database applications that require high I/O rate, random access, and high availability.

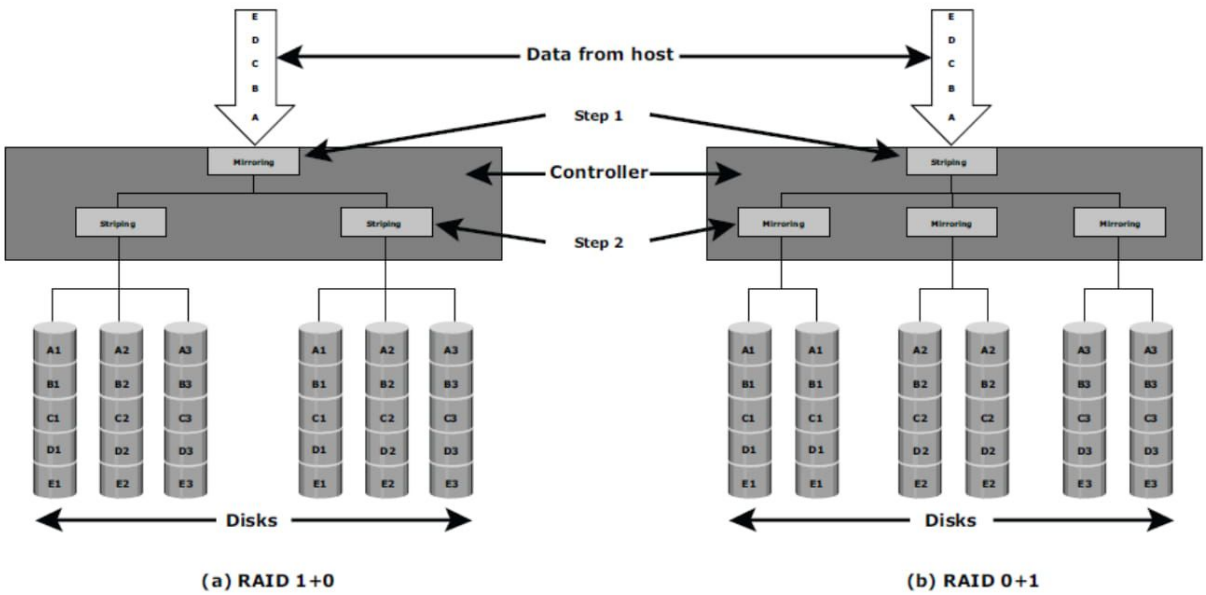


Figure 3-7: Nested RAID

Common misunderstanding is that RAID-10 and RAID-01 are the same. But they are totally different

RAID 1+0 is also called *striped mirror*. The basic element of RAID 1+0 is a mirrored pair, which means that data is first mirrored and then both copies of data are striped across multiple HDDs in a RAID set. When replacing a failed drive, only the mirror is rebuilt, i.e. the disk array controller

uses the surviving drive in the mirrored pair for data recovery and continuous operation. Data from the surviving disk is copied to the replacement disk.

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.

RAID 0+1 is also called *mirrored stripe*. The basic element of RAID 0+1 is a stripe. This means that the process of striping data across HDDs is performed initially and then the entire stripe is mirrored. If one drive fails, then the entire stripe is faulted. A rebuild operation copies the entire stripe, copying data from each disk in the healthy stripe to an equivalent disk in the failed stripe.

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

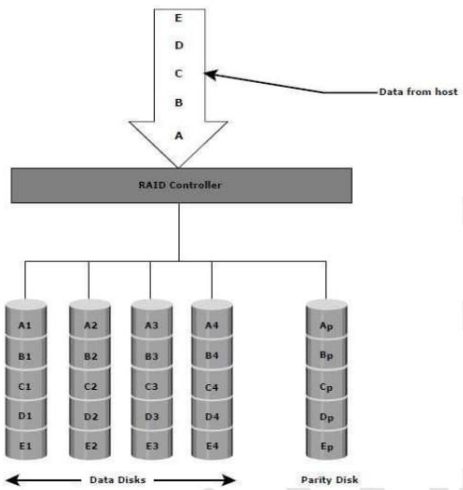
- ***Disadvantages of rebuild-operation:***
    1) Increased and unnecessary load on the surviving-disks.
    2) More vulnerable to a second disk-failure.

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:
  - 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.

There are no partial writes that update one out of many strips in a stripe

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.

RAID-5

Problem: In RAID-3 and RAID-4, parity is written to a dedicated-disk. If parity-disk fails, then it loses entire backup.

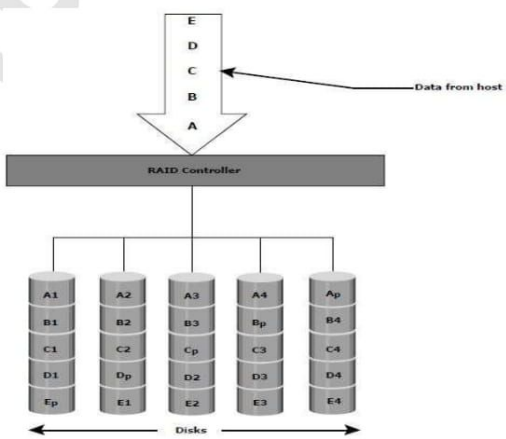
Solution: To overcome this problem, RAID-5 is proposed.

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.

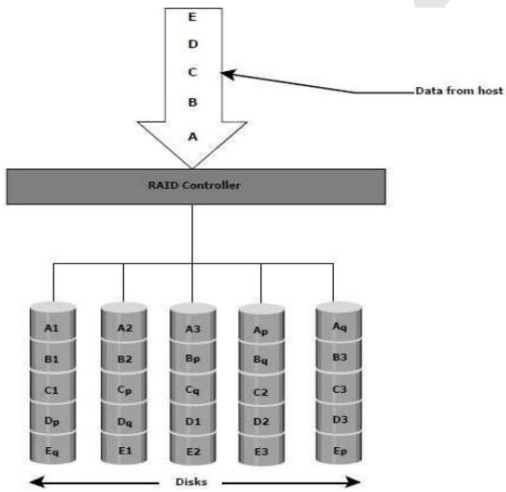
2) Preferred for RDBMS implementations in which database-admins can optimize data-access.



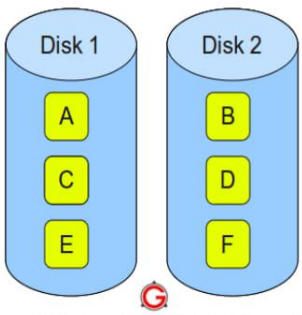
RAID-6

- RAID-6 is similar to RAID-5 except that it has

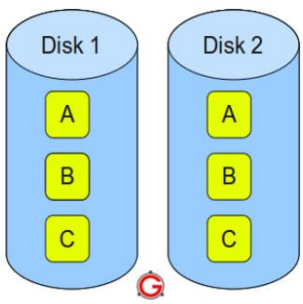
  → a second parity-element to enable survival in case of 2 disk-failures. (Figurebelow).
- Therefore, a RAID-6 implementation requires at least 4 disks.
- Similar to RAID-5, parity is distributed across all disks.
- Disadvantages: Compared to RAID-5,
  1. Write-penalty is more: RAID-5 writes perform better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



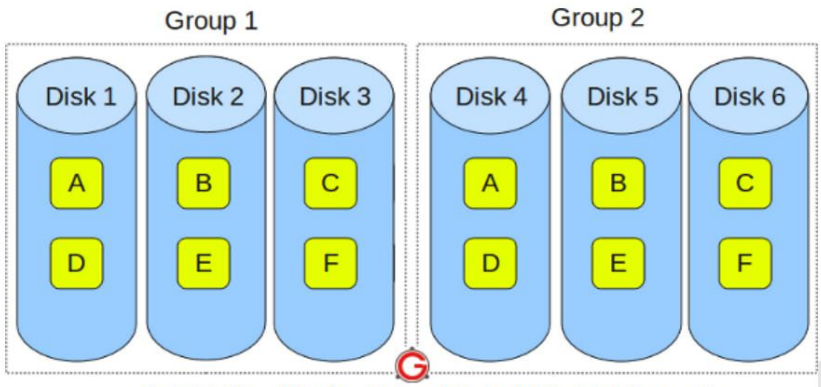
RAID Comparison: Summary



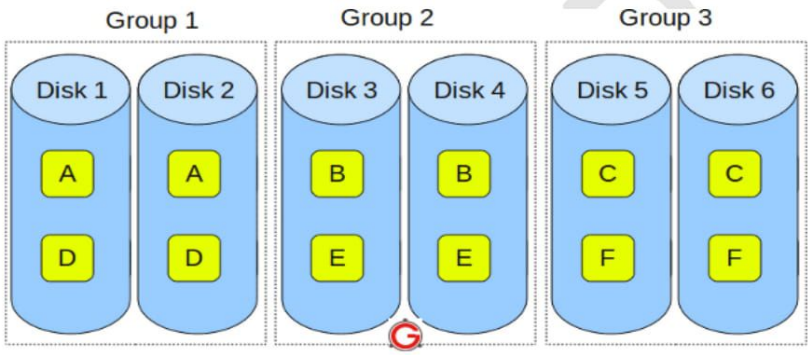
RAID 0 – Blocks Striped. No Mirror. No Parity.



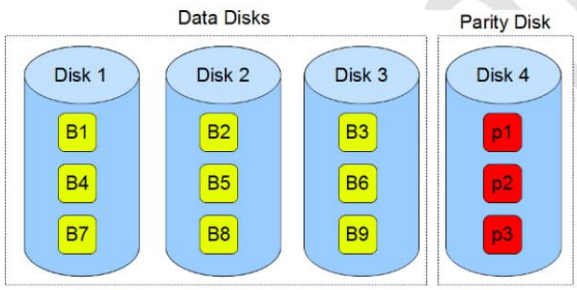
RAID 1 – Blocks Mirrored. No Stripe. No parity.



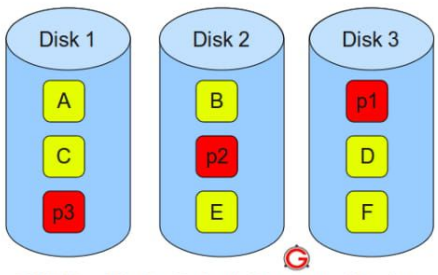
RAID 01 – Blocks Striped. (and Blocks Mirrored)



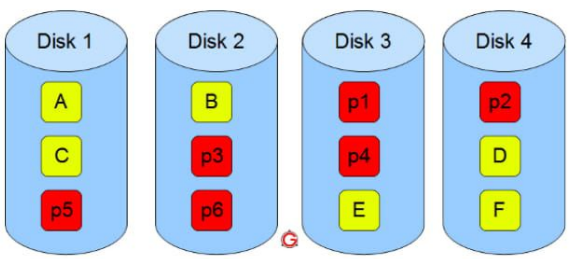
RAID 10 – Blocks Mirrored. (and Blocks Striped)



RAID 3 – Bytes Striped. (and Dedicated Parity Disk)



RAID 5 – Blocks Striped. Distributed Parity.



RAID 6 – Blocks Striped. Two Distributed Parity.

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

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: $Ep = E1 + E2 + E3 + E4$ 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 (Ep old) & old-data (E4 old) from the disk. This results in 2 read-operations

The new parity (Ep new) can be calculated by: Ep new = Ep 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.

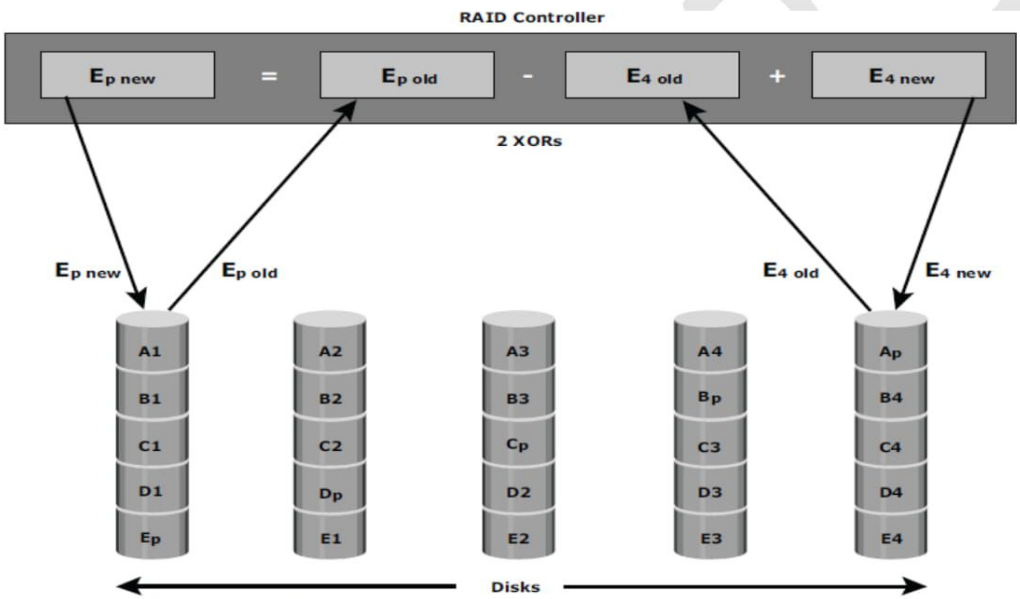


Figure 3-11: Write penalty in RAID 5

Application IOPS and RAID Configurations

When deciding the number of disks required for an application, it is important to consider the impact of RAID based on IOPS generated by the application. The total disk load should be computed by considering the type of RAID configuration and the 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 percent of them being reads.

The disk load in RAID 5 is calculated as follows:

RAID 5 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

The disk load in RAID 1 is calculated as follows:

RAID 1 disk load = 0.6 × 5,200 + **2** × (0.4 × 5,200) [because every write manifest as two writes to the disks]

= 3,120 + 2 × 2,080

= 3,120 + 4,160

= 7,280 IOPS

Computed disk load determines the number of disks required for the application.

If in this example an HDD with a specification of a maximum 180 IOPS for the application needs to be used, the number of disks required to meet the workload for the RAID configuration as follows:

RAID 5: 11,440 / 180 = 64 disks

RAID 1: 7,280 / 180 = 42 disks (approximated to the nearest even number)

Chapter 4 - Intelligent Storage System

Introduction

RAID technology made an important contribution to enhancing storage performance and reliability, but hard disk drives even with a RAID implementation could not meet performance requirements of today's applications.

With advancements in technology, a new breed of storage solutions known as an *intelligent storage system* has evolved. The intelligent storage systems are feature-rich RAID arrays that provide highly optimized I/O processing capabilities. These arrays have an operating environment that controls the management, allocation, and utilization of storage resources.

Components of an Intelligent Storage System

****** **Explain the components of intelligent storage system with a neat diagram** ******

An intelligent storage system consists of four key components:

1) *Front end,*
2) *Cache,*
3) *Back end,*
4) *Physical disks.*

Figure 4-1 illustrates these components and their interconnections.

An I/O request received from the **host** at the **front-end port** is processed through **cache** and the **back end**, to enable storage and retrieval of data from the **physical disk.**

A read request can be serviced directly from cache if the requested data is found in cache.

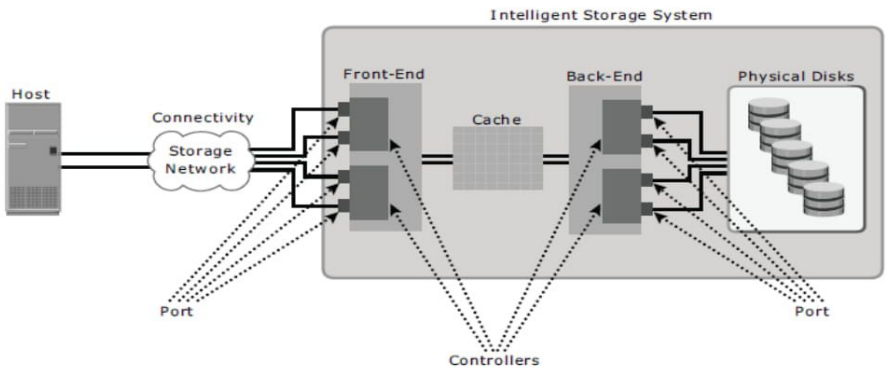


Figure 4-1: Components of an intelligent storage system

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 Controller

- Front-end port
  - → Route data to and from cache via the internal data bus.
  - → 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.

Cache

- Cache is a semiconductor-memory where 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.
- Performance is improved by separating hosts from mechanical-delays associated with disks. Rotating-disks are slowest components of a storage. This helps to improve seek-time & rotational-latency.
- Accessing data from cache takes less than a millisecond. Write data is placed in cache and then written to disk. After the data is securely placed in cache, the host is acknowledged immediately.

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.

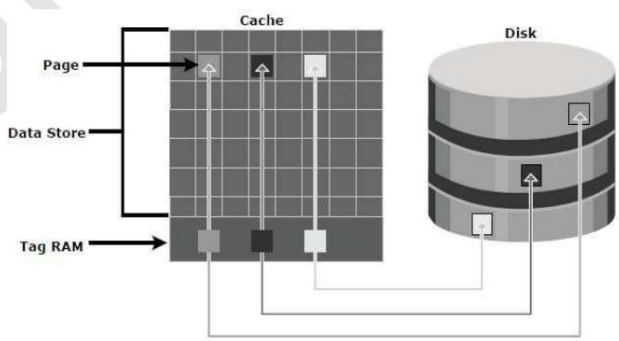


Figure: Structure of cache

- Cache consists of 2 main components

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.

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:
  - Read-Hit
  - Read-Miss &
  - 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.

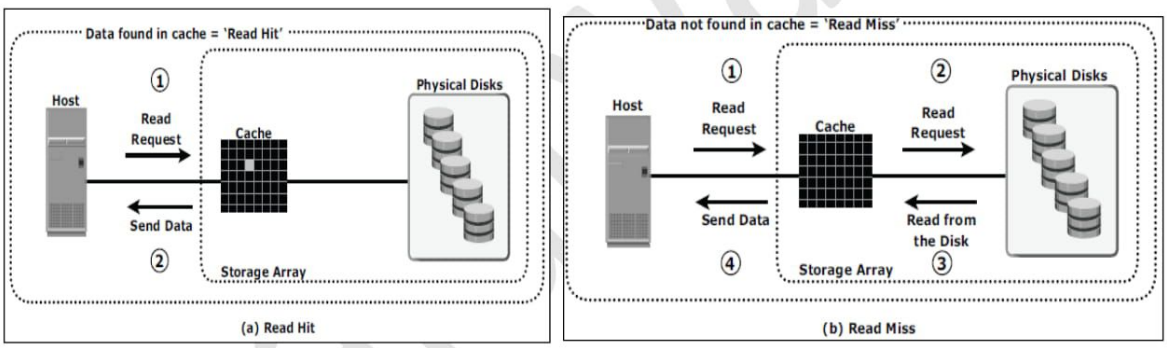
Now, the requested-data is read from the disk (Figure 1-23[b]). For this, the back-end controller → selects the appropriate disk and → retrieves the requested-data from the disk.

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.



3) Pre-Fetch (or Read-Ahead)

- A pre-fetch algorithm can be used when read-requests are sequential.
- Here is how it works:

In advance, a continuous-set of data-blocks will be **read** from the disk and placed into cache.

When host subsequently requests the blocks, data is immediately sent from cache to host.

- Advantage: 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).

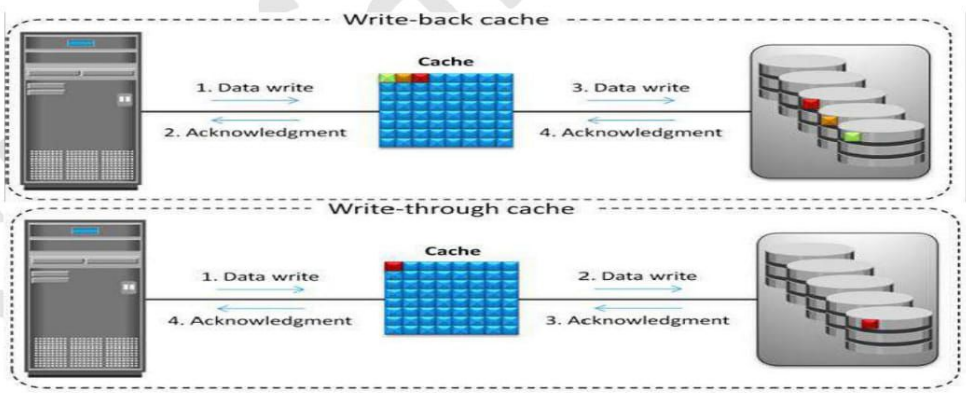
  hit-ratio =number of read-hits/ number of read-requests
- A higher hit-ratio means better read-performance.

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



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

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

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 because data is copied from cache to disk as soon as it arrives.
- Disadvantage: 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
- Suitable for applications where cache-capacity is limited and cache is used for small random-requests

Cache Implementation

1) Dedicated-cache or 2) Global-cache.

- In **dedicated-cache**, separate set of memory-locations are reserved for read and write-operations
- In **global-cache**, same set of memory-locations can be used for both read- &write-operations.

1) Global-cache is more efficient when compared to dedicated-cache. 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).

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.

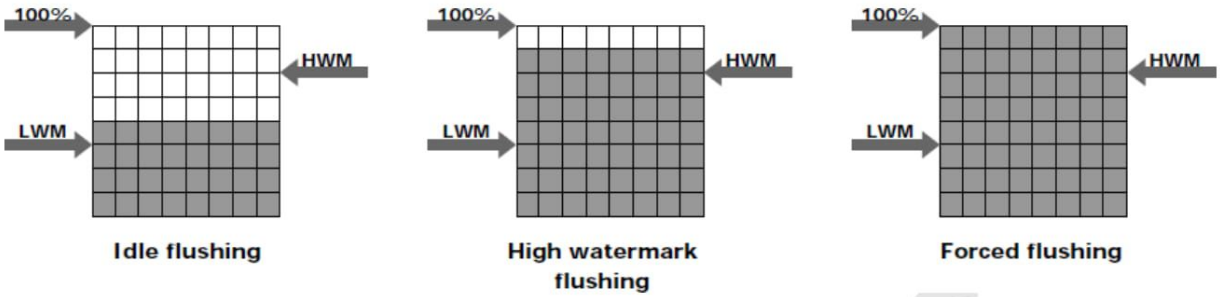


Figure 1-25: Types of flushing

- 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: The storage-device dedicates some additional resources to flushing.
- Advantage: 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: Affects the response-time.
- Advantage: The dirty-pages are forcibly flushed to disk.

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.

Other Solution to solve these problems: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: 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.

Back End

The *back end* provides an interface between cache and the physical disks. It consists of two components: back-end ports **and** 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 back end controls data transfers between cache and the physical disks. From cache, data is sent to the back end and then routed to the destination disk.

Physical disks are connected to ports on the back end. The back end controller communicates with the disks when performing reads and writes and also provides additional, but limited, temporary data storage.

The algorithms implemented on back-end controllers provide error detection and correction, along with RAID functionality.

Physical Disk

A physical disk stores data persistently. Disks are connected to the back-end with either SCSI or a Fibre Channel interface . An intelligent storage system enables the use of a mixture of SCSI or Fibre Channel drives and IDE/ATA drives.

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.

Traditional Storage Provisioning

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 (Fig) : 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#)**.
- LUNs hide the organization and composition of the set from the hosts. The use of LUNs improves disk-utilization. For example, Without using LUNs, a host requiring only 200 GB will be allocated an entire 1 TB disk.

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

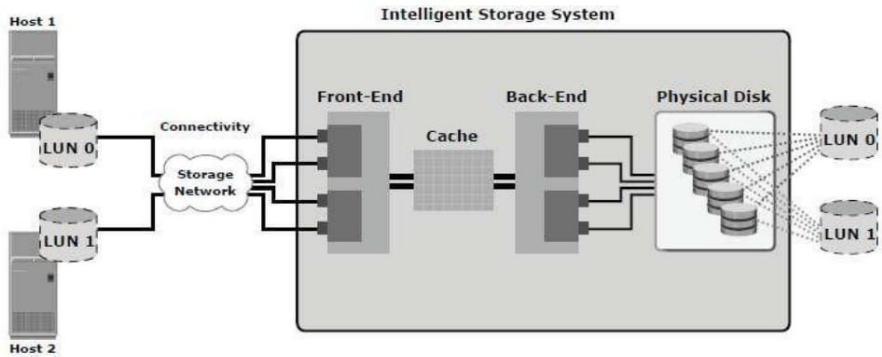


Figure 1-26: Logical-unit number

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: The expansion is quick.
- Disadvantages: Does not provide any performance-benefit.

1) 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

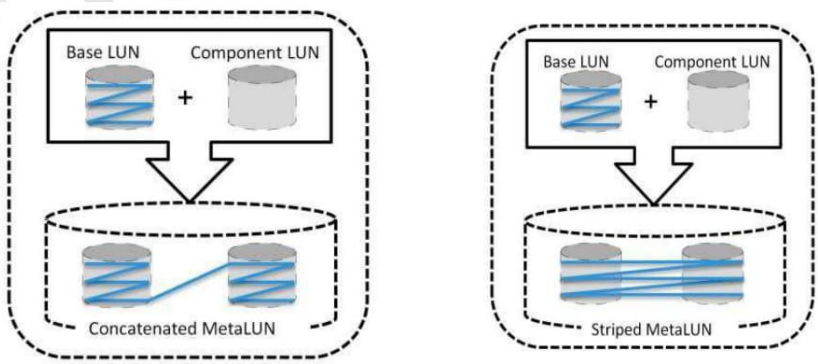


Figure 1-27: LUN Expansion

- 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

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, like 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:

LUN Masking: *LUN masking* is a process that provides data access control by defining which LUNs a host can access. LUN masking function is typically implemented at the front end

controller. This ensures that volume access by servers is controlled appropriately, preventing unauthorized or accidental use in a distributed environment.

For example, consider a storage array with two LUNs that store data of the sales and finance departments. Without LUN masking, both departments can easily see and modify each other's data, posing a high risk to data integrity and security. With LUN masking, LUNs are accessible only to the designated hosts.

Intelligent Storage Array

Intelligent storage systems generally fall into one of the following two categories:

1. High-end storage systems

2. Midrange storage systems

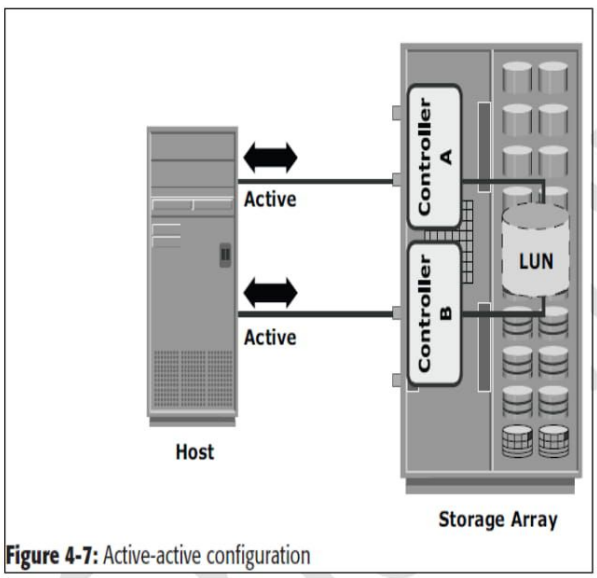


Figure 4-7: Active-active configuration

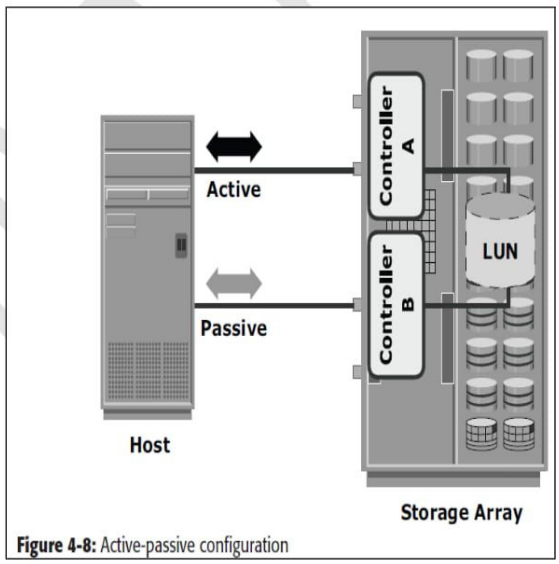


Figure 4-8: Active-passive configuration

High-end Storage Systems

High-end storage systems, referred to as ***active-active arrays***, are aimed at large enterprises for centralizing corporate data. These arrays are designed with a large number of controllers and cache memory. An active-active array implies that the host can perform I/Os to its LUNs across any of the available paths (Figure).

To address the enterprise storage needs, these arrays provide the following capabilities:

1. Large storage capacity

2. Large amounts of cache to service host I/Os optimally
3. Fault tolerance architecture to improve data availability
4. Connectivity to mainframe computers and open systems hosts
5. Availability of multiple front-end ports and interface protocols to serve a large number of hosts
6. Availability of multiple back-end Fibre Channel or SCSI RAID controllers to manage disk processing
7. Scalability to support increased connectivity, performance, and storage capacity requirements
8. Ability to handle large amounts of concurrent I/Os from a number of servers and applications
9. Support for array-based local and remote replication

Midrange Storage System

Midrange storage systems are also referred to as ***active-passive arrays*** and they are best suited for small- and medium-sized enterprises. In an active-passive array, a host can perform I/Os to a LUN only through the paths to the owning controller of that LUN. These paths are called ***active paths.*** The other paths are passive with respect to this LUN.

As shown in Figure 4-8, the host can perform reads or writes to the LUN only through the path to controller A, as controller A is the owner of that LUN. The path to controller B remains passive and no I/O activity is performed through this path.

Midrange arrays are designed to meet the requirements of small and medium enterprises; therefore, they host less storage capacity and global cache than active-active arrays. There are also fewer front-end ports for connection to servers. However, they ensure high redundancy and high performance for applications with predictable workloads. They also support array-based local and remote replication.

Question Bank

SI.NO.	Questions
1	Define server centric IT architecture and storage centric IT architecture with advantages and limitations. Explain evaluation storage architecture.
2	Discuss the key characteristics of a data center, with a neat diagram.
3	Explain different RAID levels with their advantages and disadvantages
4	Explain briefly how parity blocks are calculated in RAID4 and RAID5. How RAID5 overcomes limitations of RAID4?
5	With a neat diagram, explain the architecture of intelligent disk storage system[dss].
6	Define the two main goals of RAID. What is a RAID level and explain the use of hot spare disks for all RAID levels?
7	Explain RAID 0 and 1 level or Block-by-Block striping and mirroring?
8	Explain the connectivity protocols, logical components of data center
9	Compare the principle of operation in RAID 0+1 and RAID 10 level?
10	Explain the types of Intelligent disk storage system
11	Describe the two types of caches are designed to accelerate write and read accesses to physical hard disks?
12	Describe RAID levels with reference to nested RAID, RAID 3, RAID 5 with neat diagram.
13	With a neat diagram, explain the components of Intelligent Storage System with LUN. Define LUN masking
14	With a neat diagram, explain the structure of read and write operations in cache.
15	Explain the various techniques on the basis of which RAID levels are defined.
16	List and explain the components of storage environment
17	With a neat diagram explain the LVM