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

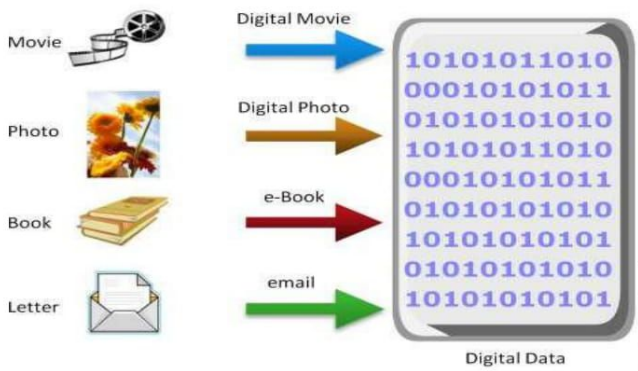


Figure 1-1: Digital data

- 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

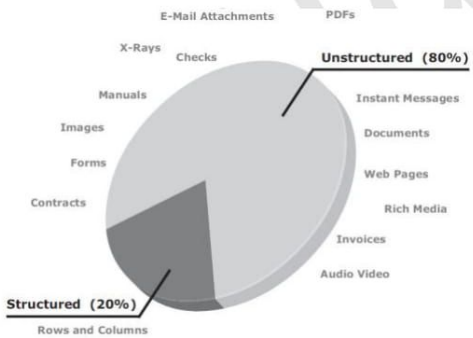


Figure 1-3: Types of data

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.

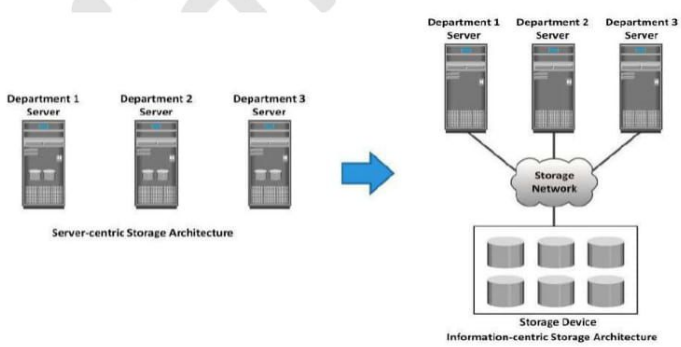


Figure 1-3: Evolution of storage architectures

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.

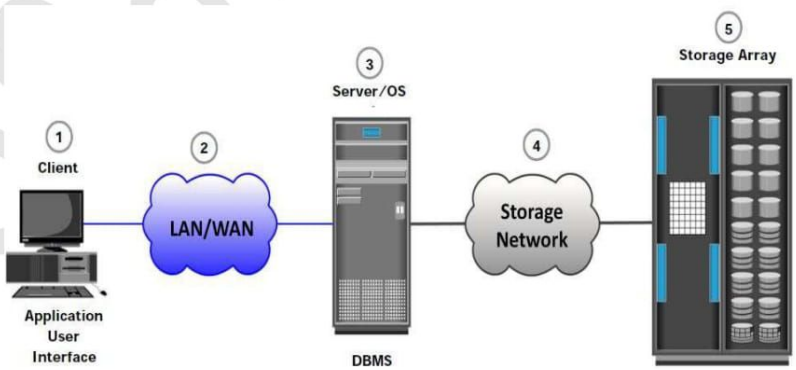


Figure 1-4: Example of an order processing-application

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).

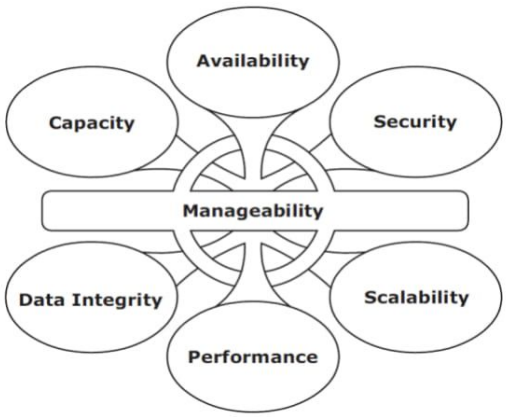


Figure 1-6: Key characteristics of data center elements

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

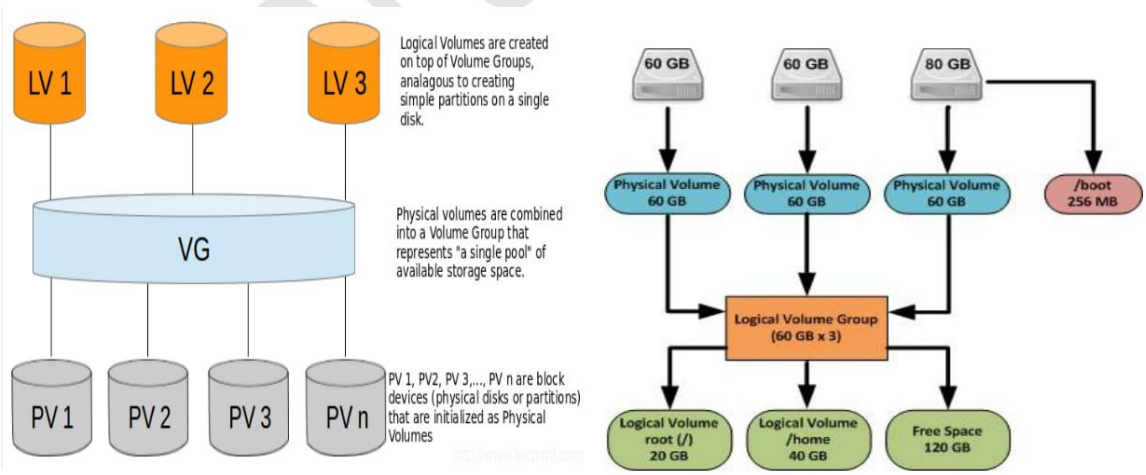


Figure: Components of LVM

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.

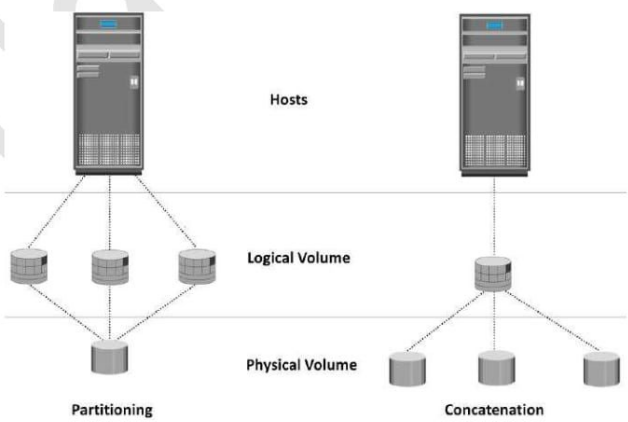


Figure 1-6: Disk partitioning and concatenation

- 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.

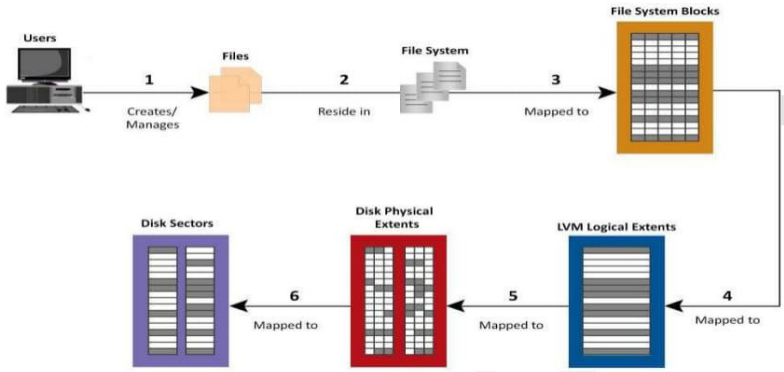


Figure 1-7: Process of mapping user files to disk storage

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.

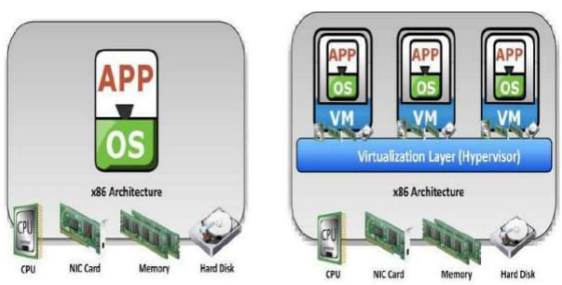


Figure 1-8: Server virtualization

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.

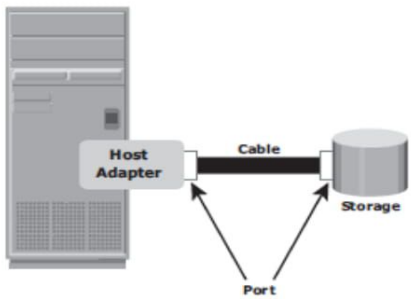


Figure 2-4: Physical components of connectivity

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:

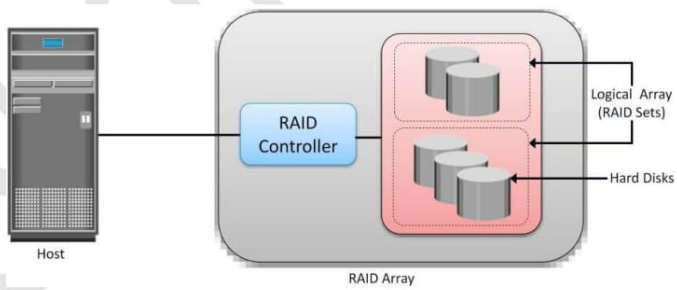
1. Management and control of disk aggregations
2. Translation of I/O requests between logical disks and physical disks
3. Data regeneration in the event of disk failures.

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** *********

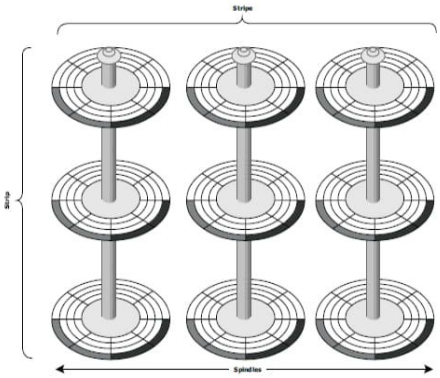
- RAID-levels are defined based on following 3 techniques:
  1) Striping (used to improve performance of storage)
  2) Mirroring (used to improve data-availability) and
  3) Parity (used to provide data-protection)
- The above techniques determine
  → performance of storage-device (i.e. better performance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

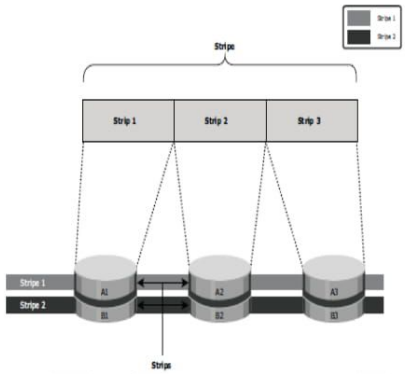
Some RAID-arrays use a combination of above 3 techniques. For example: Striping with mirroring, Striping with parity

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.

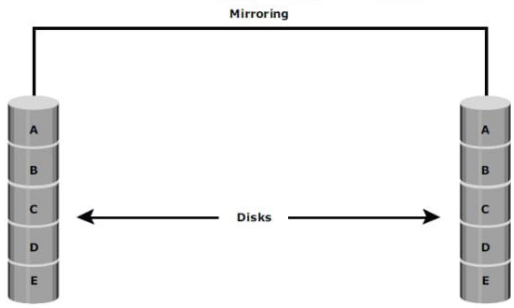


Figure 3-3: Mirrored disks in an array

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:

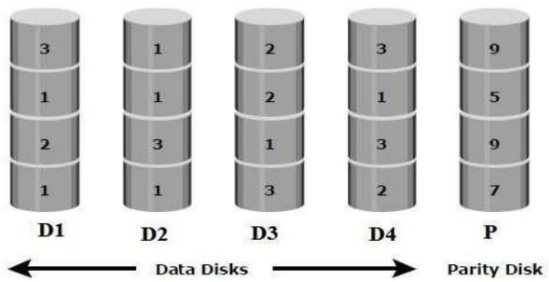
1) Decreases write-performance because
   → each write-request must perform 2 write-operations on the disks.
2) Duplication of data. Thus, amount of storage-capacity needed is twice amount of data being stored (E.g. To store 100GB data, 200GB disk is needed).
3) Considered expensive and preferred for 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.

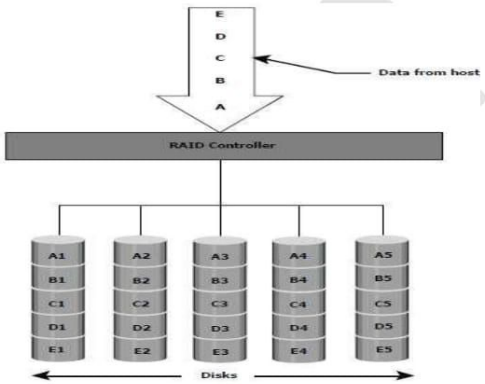
LEVELS	BRIEF DESCRIPTION
RAID 0	Striped array with no fault tolerance
RAID 1	Disk mirroring
RAID 3	Parallel access array with dedicated parity disk
RAID 4	Striped array with independent disks and a dedicated parity disk
RAID 5	Striped array with independent disks and distributed parity
RAID 6	Striped 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:

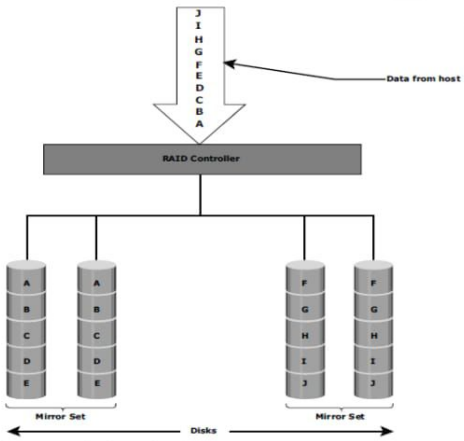
  1) Does not provide data-protection and data-availability in case of disk-failure.

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.

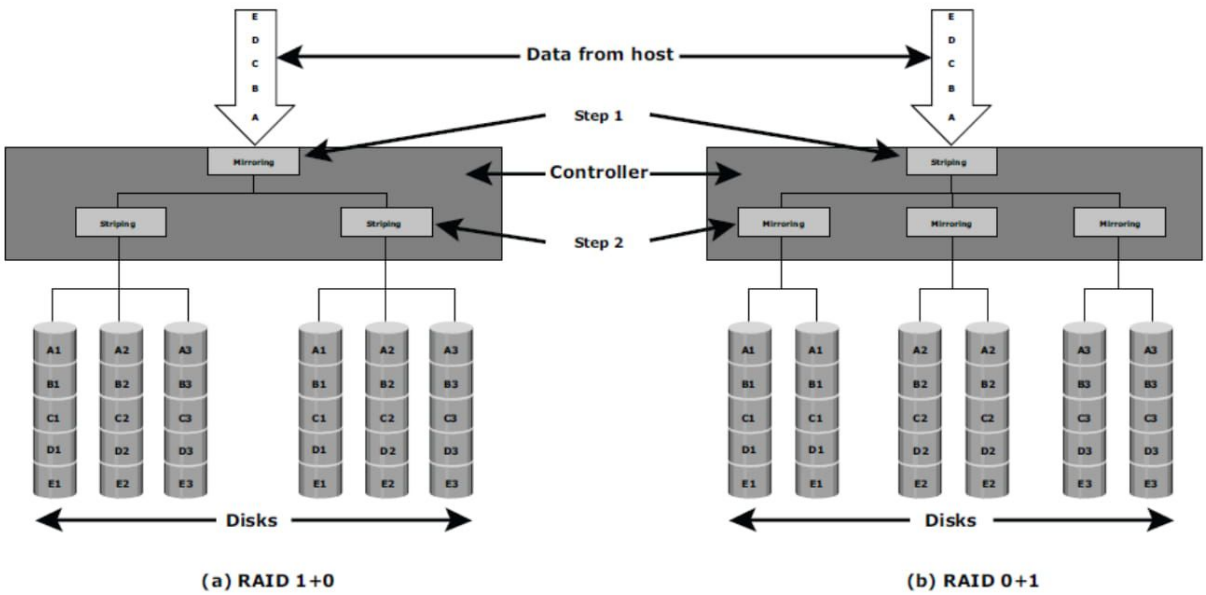


Figure 3-7: Nested RAID

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

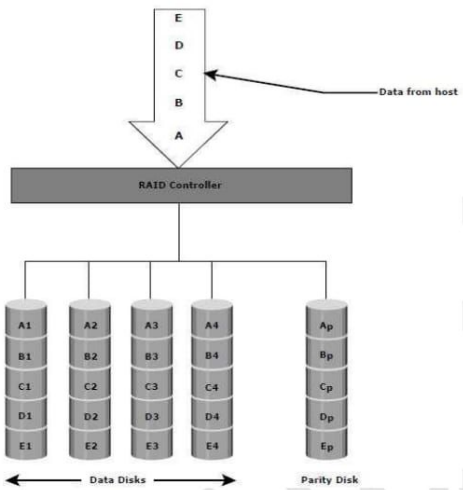
- ***Disadvantages of rebuild-operation:***
    1) Increased and unnecessary load on the surviving-disks.
    2) More vulnerable to a second disk-failure.

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.

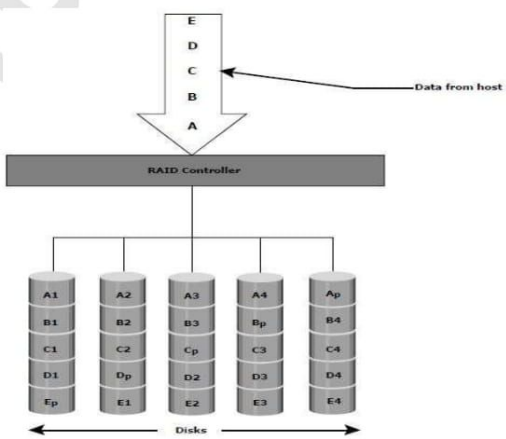
Solution: To overcome this problem, RAID-5 is proposed.

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.

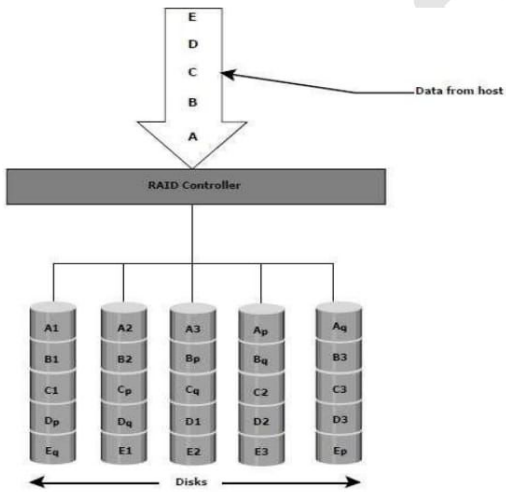
2) Preferred for RDBMS implementations in which database-admins can optimize data-access.



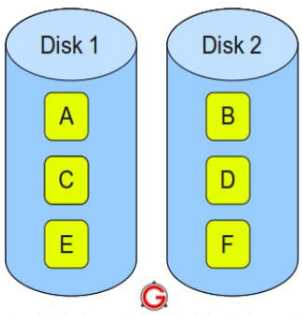
RAID-6

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

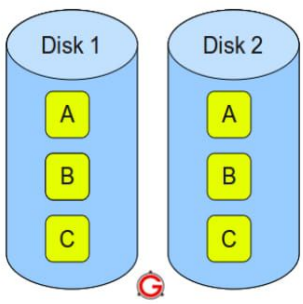
  → a second parity-element to enable survival in case of 2 disk-failures. (Figurebelow).
- Therefore, a RAID-6 implementation requires at least 4 disks.
- Similar to RAID-5, parity is distributed across all disks.
- Disadvantages: Compared to RAID-5,
  1. Write-penalty is more: RAID-5 writes perform better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



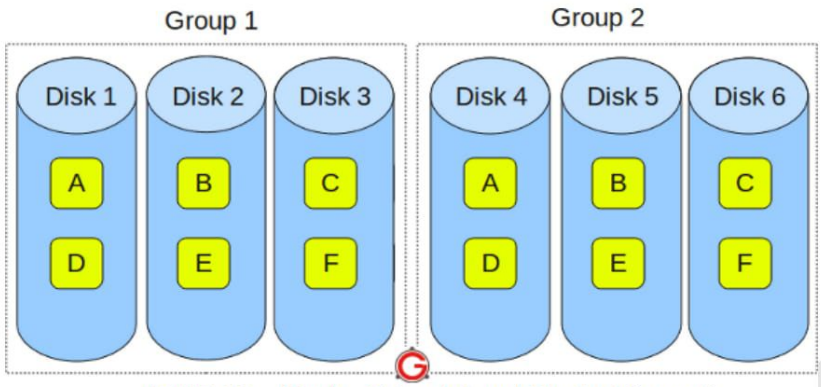
RAID Comparison: Summary



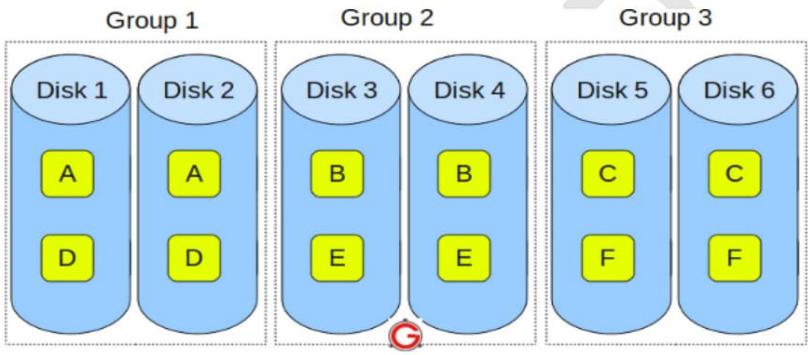
RAID 0 – Blocks Striped. No Mirror. No Parity.



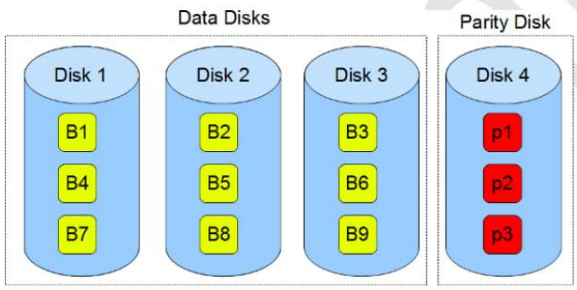
RAID 1 – Blocks Mirrored. No Stripe. No parity.



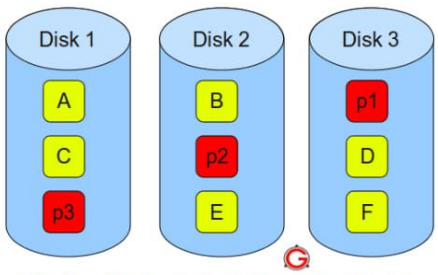
RAID 01 – Blocks Striped. (and Blocks Mirrored)



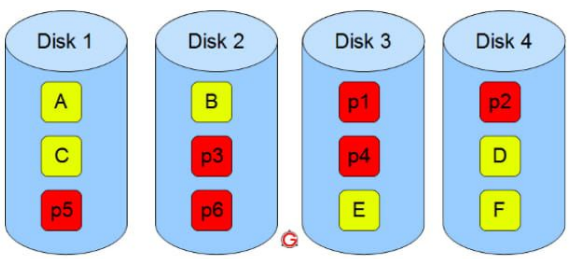
RAID 10 – Blocks Mirrored. (and Blocks Striped)



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.

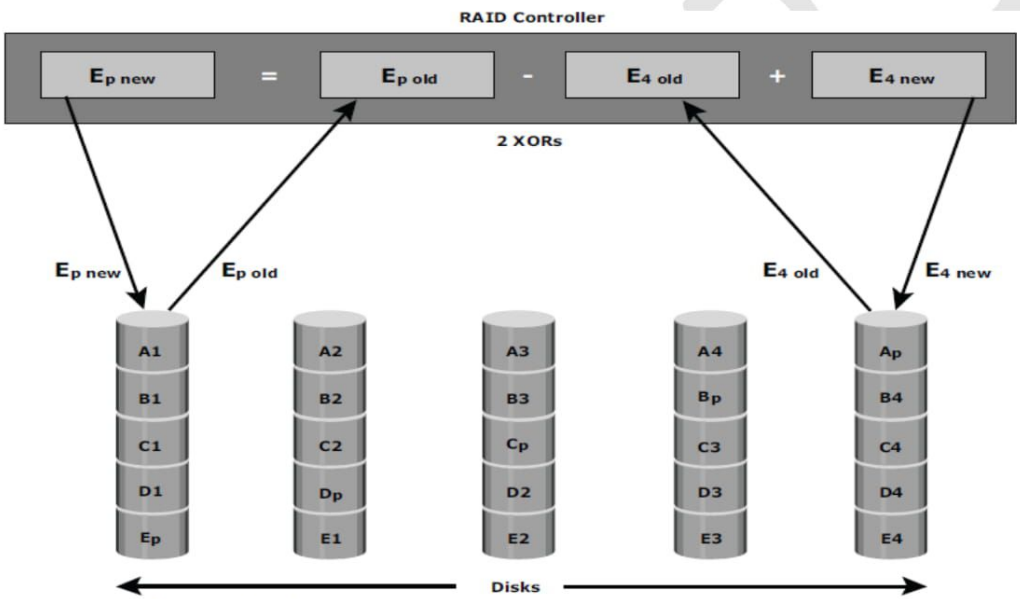


Figure 3-11: Write penalty in RAID 5

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.

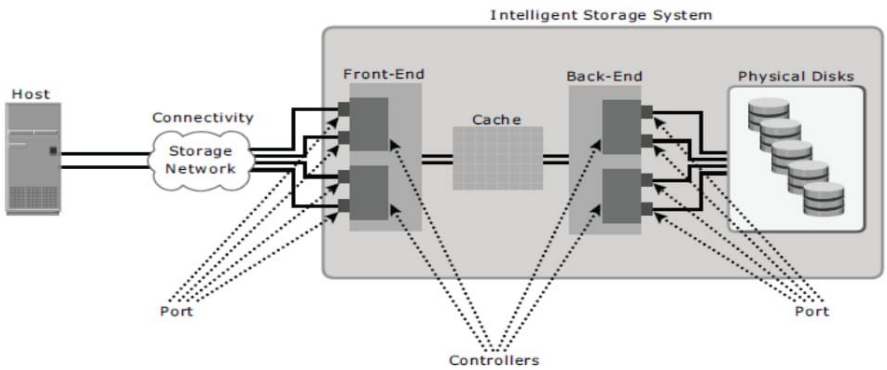


Figure 4-1: Components of an intelligent storage system

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.

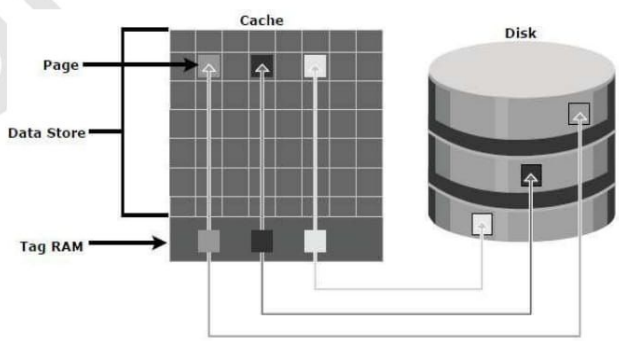


Figure: Structure of cache

- 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.

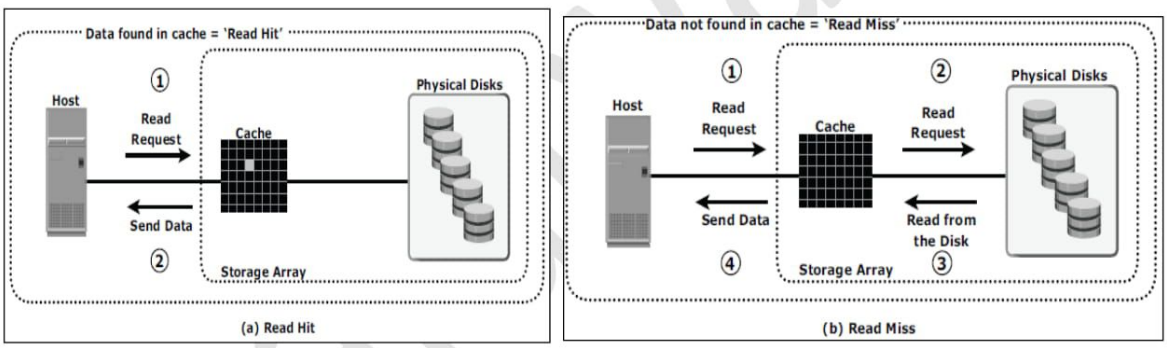
Now, the requested-data is read from the disk (Figure 1-23[b]). For this, the back-end controller → selects the appropriate disk and → retrieves the requested-data from the disk.

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).

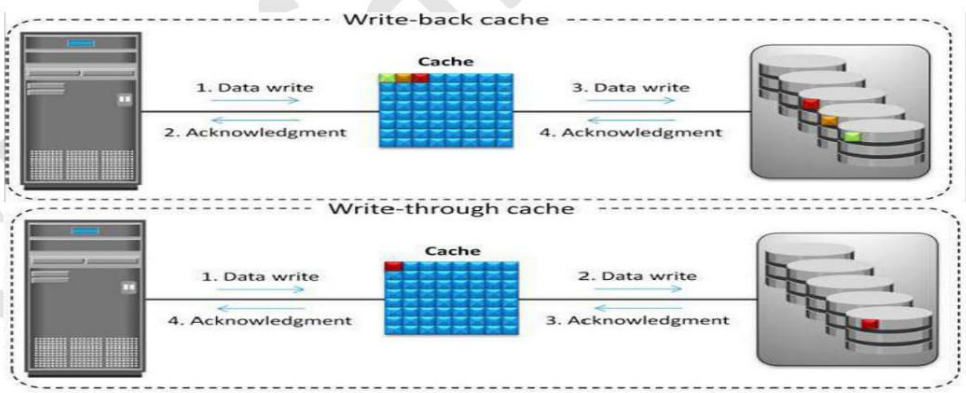
  hit-ratio =number of read-hits/ number of read-requests
- A higher hit-ratio means better read-performance.

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.

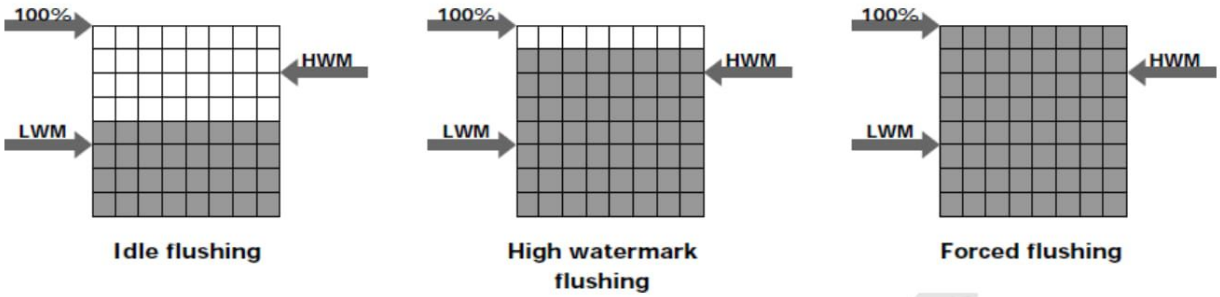


Figure 1-25: Types of flushing

- 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

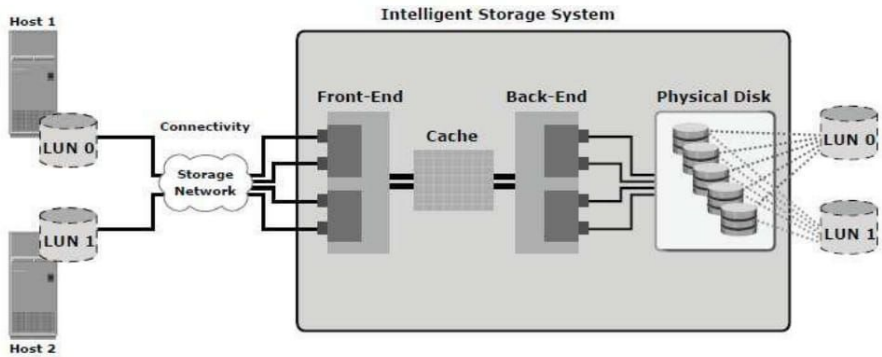


Figure 1-26: Logical-unit number

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

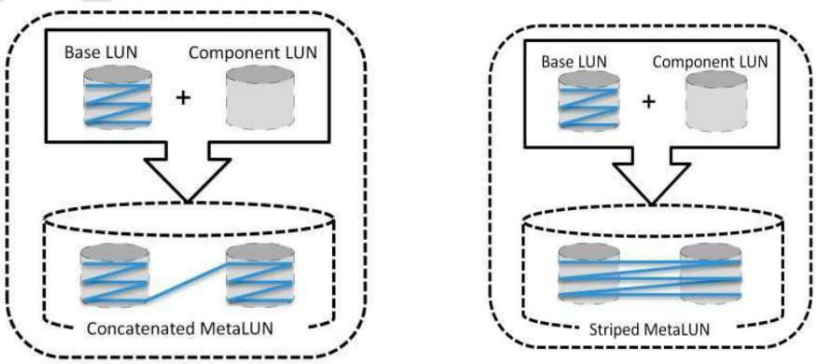


Figure 1-27: LUN Expansion

- 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

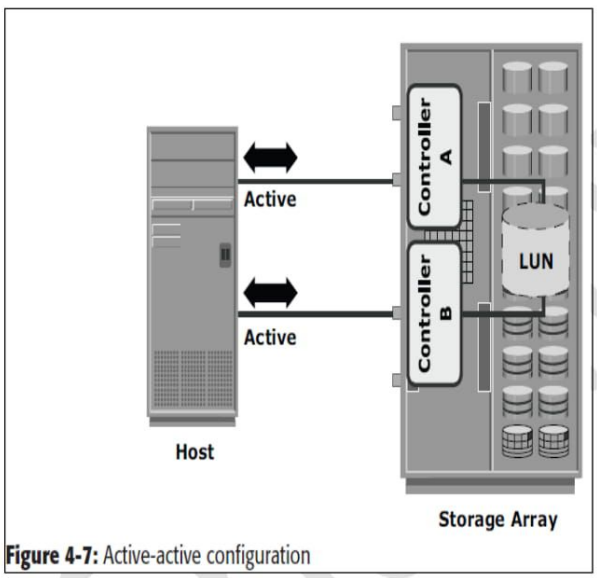


Figure 4-7: Active-active configuration

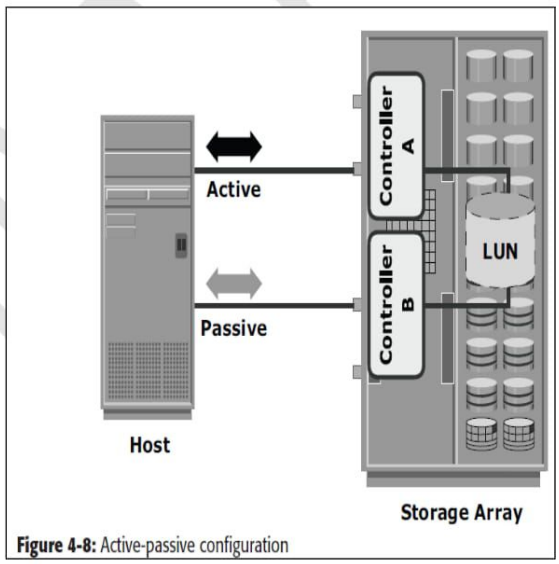


Figure 4-8: Active-passive configuration

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

MODULE 2

Syllabus: Storage Networking Technologies Fibre Channel Storage Area Networks: Components of FC SAN, FC connectivity, Fibre Channel Architecture, Zoning, FC SAN Topologies, Virtualization in SAN. IP SAN and FCoE: iSCSI, FCIP, FCoE. Network Attached Storage: Components of NAS, NAS I/O Operation, NAS File-Sharing Protocols, File-Level Virtualization, Object-Based Storage and Unified Storage: Object-Based Storage Devices, Content-Addressed Storage, Unified Storage.

Chapter 5: Fibre Channel Storage Area Networks

5.1 Fibre Channel: Overview

The FC architecture forms the fundamental construct of the FC SAN infrastructure. *Fibre Channel* is a high-speed network technology that runs on high-speed optical fiber cables and serial copper cables. The FC technology was developed to meet the demand for increased speeds of data transfer between servers and mass storage systems.

Explain the different components of FC SAN

5.3 Components of FC SAN

Node Ports

- In a Fibre Channel network, the end devices, such as hosts, storage arrays, and tape libraries, are all referred to as nodes.
- Each node is a source or destination of information.
- Each node requires one or more ports to provide a physical interface for communicating with other nodes.
- These ports are integral components of host adapters, such as HBA, and storage front-end controllers or adapters.
- In an FC environment a port operates in full-duplex data transmission mode with a transmit (Tx) link and a receive (Rx) link

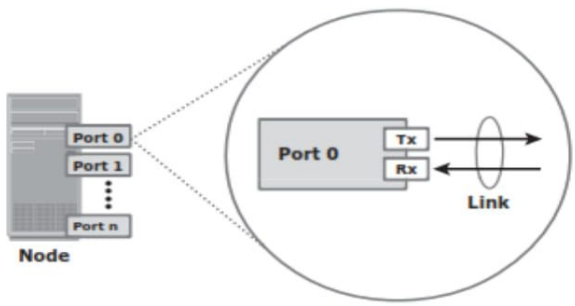


Figure 5-3: Nodes, ports, and links

Cables and Connectors

- SAN implementations use optical fiber cabling.
- Copper can be used for shorter distances for back-end connectivity because it provides an acceptable signal-to noise ratio for distances up to 30 meters.
- Optical fiber cables carry data in the form of light.
- There are two types of optical cables***: multimode and single-mode***.
- Multimode fiber (MMF) cable carries multiple beams of light projected at different angles simultaneously onto the core of the cable (see Figure 5-4 [a]).
- Based on the bandwidth, multimode fibers are classified as OM1 (62.5µm core), OM2 (50µm core), and laser-optimized OM3 (50µm core).
- In an MMF transmission, multiple light beams traveling inside the cable tend to disperse and collide.
- This collision weakens the signal strength after it travels a certain distance — a process known as modal dispersion.
- An MMF cable is typically used for short distances because of signal degradation (attenuation) due to modal dispersion.
- Single-mode fiber (SMF) carries a single ray of light projected at the center of the core
- These cables are available in core diameters of 7 to 11 microns; the most common size is 9 microns.
- In an SMF transmission, a single light beam travels in a straight line through the core of the fiber.
- The small core and the single light wave help to limit modal dispersion.
- Among all types of fiber cables, single mode provides minimum signal attenuation over maximum distance (up to 10 km).

Figure 5-4: Multimode fiber and single-mode fiber

- A ***connector*** is attached at the end of a cable to enable swift connection and disconnection of the cable to and from a port.
- A Standard connector (SC) (see Figure 5-5 [a]) and a Lucent connector (LC) (see Figure 5-5 [b]) are two commonly used connectors for fiber optic cables. (Both are push and pull connectors with the only difference in their size)
- Straight Tip (ST) is another fiber-optic connector, which is often used with fiber patch panels (see Figure 5.5 [c]).

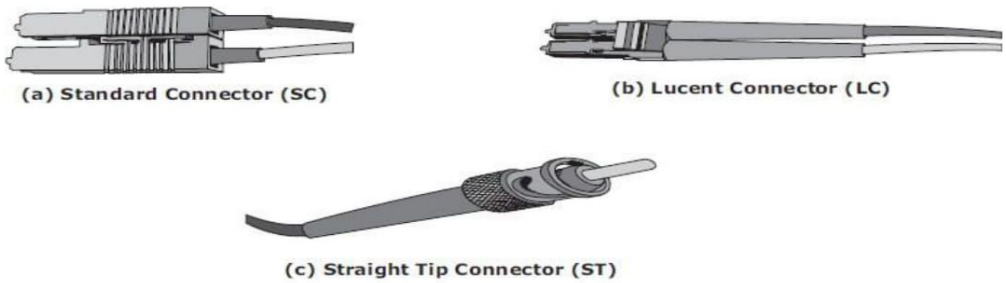


Figure 5-5: SC, LC, and ST connectors

Interconnect Devices

- FC hubs, switches, and directors are the interconnect devices commonly used in FC SAN.
- Hubs are used as communication devices in FC-AL implementations.
- Hubs physically connect nodes in a logical loop or a physical star topology.
- All the nodes must share the loop because data travels through all the connection points.
- Because of the availability of low-cost and high-performance switches, hubs are no longer used in FC SANs
- Switches are more intelligent than hubs and directly route data from one physical port to another.
- Therefore, nodes do not share the bandwidth. Instead, each node has a dedicated communication path.
- Directors are high-end switches with a higher port count and better fault tolerance capabilities.
- A port card or blade has multiple ports for connecting nodes and other FC switches

SAN Management Software

- SAN management software manages the interfaces between hosts, interconnect devices, and storage arrays.
- The software provides a view of the SAN environment and enables management of various resources from one central console.
- It provides key management functions, including mapping of storage devices, switches, and servers, monitoring and generating alerts for discovered devices, and zoning

*** Explain the different types of connectivity in FC Configuration***

5.4 FC Connectivity

Point-to-Point

- Point-to-point is the simplest FC configuration — two devices are connected directly to each other, as shown in Figure 5-6.
- This configuration provides a dedicated connection for data transmission between nodes.
- However, the point-to-point configuration offers limited connectivity, because only two devices can communicate with each other at a given time.
- Moreover, it cannot be scaled to accommodate a large number of nodes.
- Standard DAS uses point-to-point connectivity

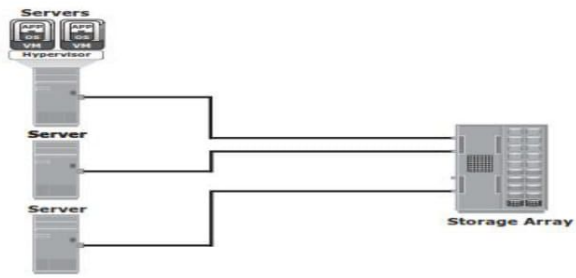


Figure 5-6: Point-to-point connectivity

Fibre Channel Arbitrated Loop

- In the FC-AL configuration, devices are attached to a shared loop.
- FC-AL has the characteristics of a token ring topology and a physical star topology.
- In FC-AL, each device contends with other devices to perform I/O operations.
- Devices on the loop must "arbitrate" to gain control of the loop.
- At any given time, only one device can perform I/O operations on the loop (see Figure 5-7).

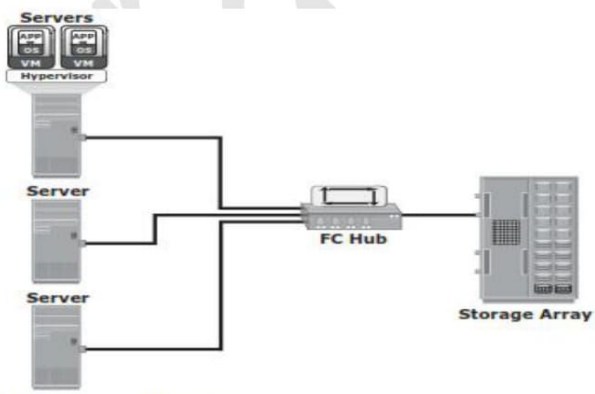


Figure 5-7: Fibre Channel Arbitrated Loop

Fibre Channel Switched Fabric

- FC-SW is also referred to as ***fabric connect***.
- A fabric is a logical space in which all nodes communicate with one another in a network.

- This virtual space can be created with a switch or a network of switches.
- Each switch in a fabric contains a unique domain identifier, which is part of the fabric's addressing scheme.
- In FC-SW, nodes do not share a loop; instead, data is transferred through a dedicated path between the nodes.
- Each port in a fabric has a unique 24-bit Fibre Channel address for communication. Figure 5-8 shows an example of the FC-SW fabric.
- In a switched fabric, the link between any two switches is called an ***Interswitch link (ISL).***
- ISLs enable switches to be connected together to form a single, larger fabric.
- ISLs are used to transfer host-to-storage data and fabric management traffic from one switch to another.
- By using ISLs, a switched fabric can be expanded to connect a large number of nodes.

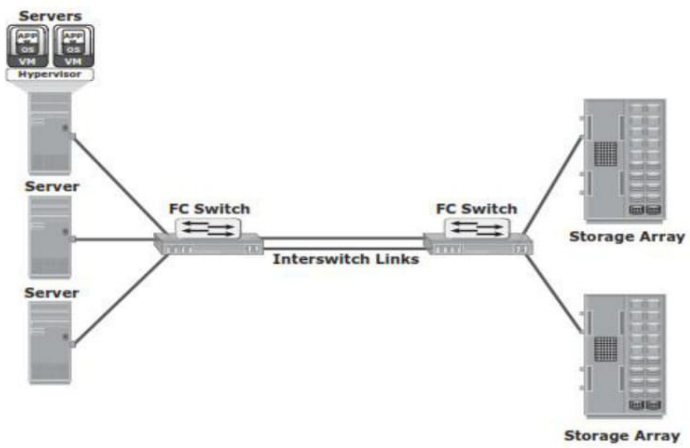


Figure 5-8: Fibre Channel switched fabric

- A fabric can be described by the number of tiers it contains
- When the number of tiers in a fabric increases, the distance that the fabric management traffic must travel to reach each switch also increases.
- This increase in the distance also increases the time taken to propagate and complete a fabric reconfiguration event, such as the addition of a new switch or a zone set propagation event.
- Figure 5-9 illustrates two- tier and three-tier fabric architecture.

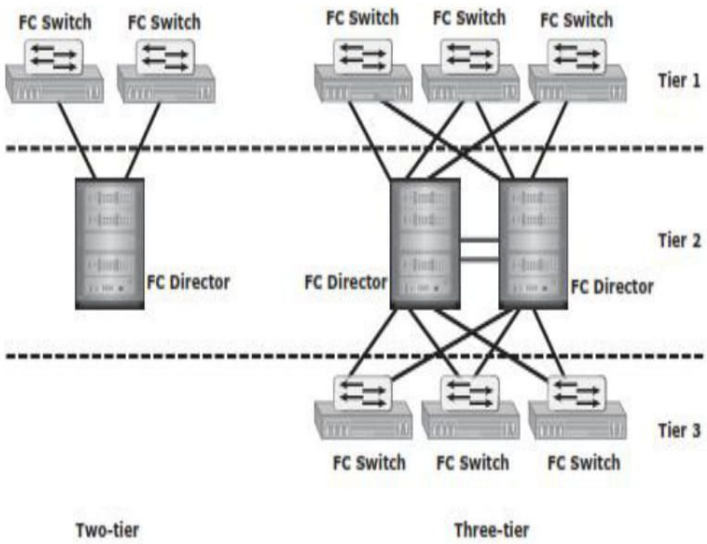


Figure 5-9: Tiered structure of Fibre Channel switched fabric

FC-SW Transmission

- FC-SW uses switches that can switch data traffic between nodes directly through switch ports.
- Frames are routed between source and destination by the fabric.
- As shown in Figure 5-10, if node B wants to communicate with node D, the nodes should individually login first and then transmit data via the FC-SW.
- This link is considered a dedicated connection between the initiator and the target.

** *Different types of Fabric ports***

5.6 Switched Fabric ports

Types

- **N_Port:** An end point in the fabric. This port is also known as the node port. Typically, it is a host port (HBA) or a storage array port connected to a switch in a switched fabric.
- **E_Port:** A port that forms the connection between two FC switches. This port is also known as the expansion port.
  The E_Port on an *FC* switch *con*nects to the E_Port of another FC switch in the fabric through ISLs.
- **F_Port:** A port on a switch that connects an N_Port. It is also known as a fabric port.
- **G_Port:** A generic port on a switch that can operate as an E_Port or an F_Port and determines its functionality automatically during initialization.

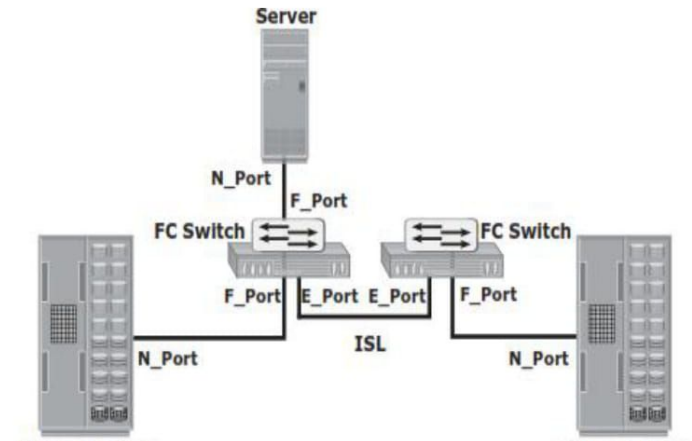


Figure 5-11: Switched fabric ports

5.6 Fibre Channel Architecture

- Traditionally, host computer operating systems have communicated with peripheral devices over channel connections, such as ESCON and SCSI.
- Channel technologies provide high levels of performance with low protocol overheads.
- Such performance is achievable due to the static nature of channels and the high level of hardware and software integration provided by the channel technologies.
- The FC architecture represents true channel/network integration and captures some of the benefits of both channel and network technology.
- FC SAN uses the Fibre Channel Protocol (FCP) that provides both channel speed for data transfer with low protocol overhead and scalability of network technology.

The key advantages of FCP are as follows:

- Sustained transmission bandwidth over long distances.
- Support for a larger number of addressable devices over a network. Theoretically, FC can support more than 15 million device addresses on a network.
- Support speeds up to 16 Gbps (16 GFC).

Explain the different layers in FCP stack with the neat diagram

Fibre Channel Protocol Stack

- It is easier to understand a communication protocol by viewing it as a structure of independent layers.
- FCP defines the communication protocol in five layers: FC-0 through FC-4 (except FC-3 layer, which is not implemented).
- In a layered communication model, the peer layers on each node talk to each other through defined protocols. Figure 5-12 illustrates the Fibre Channel protocol stack.

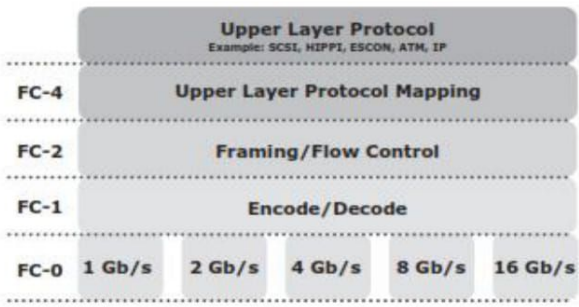


Figure 5-12: Fibre Channel protocol stack

FC-4 Layer

- FC-4 is the uppermost layer in the FCP stack.
- This layer defines the application interfaces and the way Upper Layer Protocols (ULPs) are mapped to the lower FC layers.
- The FC standard defines several protocols that can operate on the FC-4 layer (see Figure 5-12).
- Some of the protocols include SCSI, High Performance Parallel Interface (HIPPI) Framing Protocol, Enterprise Storage Connectivity (ESCON), Asynchronous Transfer Mode (ATM), and IP.

FC-2 Layer

- The FC-2 layer provides Fibre Channel addressing, structure, and organization of data (frames, sequences, and exchanges).
- It also defines fabric services, classes of service, flow control, and routing.

FC-1 Layer

- The FC-1 layer defines how data is encoded prior to transmission and decoded upon receipt.
- At the transmitter node, an 8-bit character is encoded into a 10-bit transmissions character.
- This character is then transmitted to the receiver node.
- At the receiver node, the 10-bit character is passed to the FC-1 layer, which decodes

the 10-bit character into the original 8-bit character.

- FC links with speeds of 10 Gbps and above use 64-bit to 66bit encoding algorithms.
- The FC-1 layer also defines the transmission words, such as FC frame delimiters, which identify the start and end of a frame and primitive signals that indicate events at a transmitting port.
- In addition to these, the FC-1 layer performs link initialization and error recovery.

FC-0 Layer

- FC-0 is the lowest layer in the FCP stack. T
- his layer defines the physical interface, media, and transmission of bits.
- The FC-0 specification includes cables, connectors, and optical and electrical parameters for a variety of data rates.
- The FC transmission can use both electrical and optical media.

Fibre Channel Addressing

- An FC address is dynamically assigned when a node port logs on to the fabric. The FC address has a distinct format, as shown in Figure 5-13.
- The addressing mechanism provided here corresponds to the fabric with the switch as an interconnecting device.

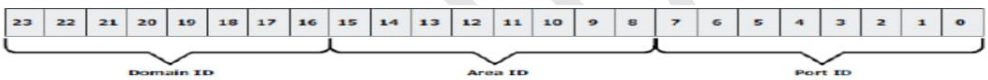


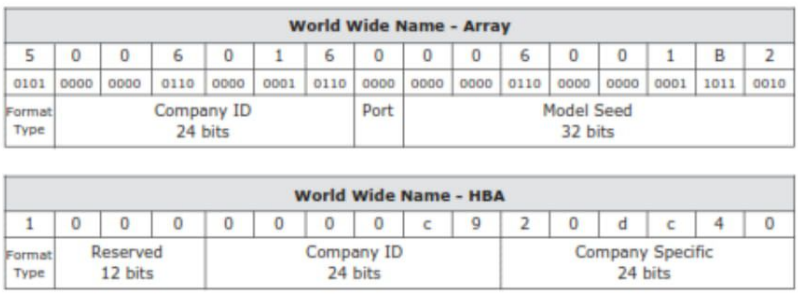
Figure 6-14: 24-bit FC address of N_port

- The first field of the FC address contains the domain ID of the switch.
- A domain ID is a unique number provided to each switch in the fabric.
- Although this is an 8-bit field, there are only 239 available addresses for domain ID because some addresses are deemed special and reserved for fabric management services. For example, FFFFFC is reserved for the name server, and FFFFFE is reserved for the fabric login service.
- The area ID is used to identify a group of switch ports used for connecting nodes.
- An example of a group of ports with a common area ID is a port card on the switch.
- The last field, the port ID, identifies the port within the group.
- Therefore, the maximum possible number of node ports in a switched fabric is calculated as: 239 domains *256 areas *256 ports = 15,663,104

World Wide Names

- Each device in the FC environment is assigned a 64-bit unique identifier called the World-Wide Name (WWN).

- The Fibre Channel environment uses two types of WWNs: ***World Wide Node Name (WWNN) and World-Wide Port Name (WWPN).***
- WWNs are burned into the hardware or assigned through software.
- Unlike an FC address, which is assigned dynamically, a WWN is a static name for each device on an FC network.
- WWNs are similar to the Media Access Control (MAC) addresses used in IP networking
- Several configuration definitions in a SAN use WWN for identifying storage devices and HBAs.
- The name server in an FC environment keeps the association of WWNs to the dynamically created FC addresses for nodes.
- Figure 5-14 illustrates the WWN structure examples for an array and an HBA.



World Wide Name - Array															
5	0	0	6	0	1	6	0	0	0	6	0	0	1	B	2
0101	0000	0000	0110	0000	0001	0110	0000	0000	0000	0110	0000	0000	0001	1011	0010
Format Type	Company ID 24 bits						Port	Model Seed 32 bits							

World Wide Name - HBA															
1	0	0	0	0	0	0	0	c	9	2	0	d	c	4	0
Format Type	Reserved 12 bits			Company ID 24 bits						Company Specific 24 bits					

Figure 5-14: World Wide Names

FC Frame

- An FC frame (Figure 5-15) consists of five parts***: start of frame (SOF), frame header, data field, cyclic redundancy check (CRC), and end of frame (EOF).***

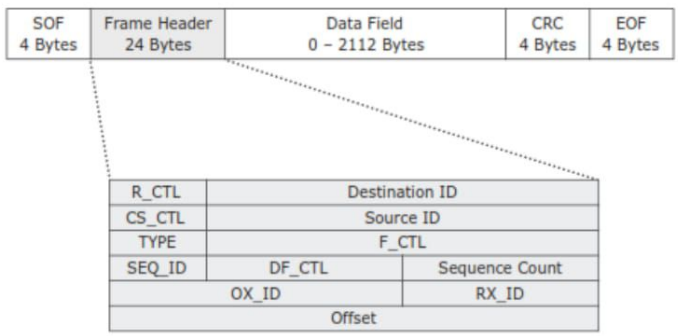


Figure 5-15: FC frame

- The frame header is **24 bytes long and contains addressing information for the frame.**
- It includes the following information: Source ID (S_ID), Destination ID (D_ID), Sequence ID (SEQ_ID), Sequence Count (SEQ_CNT), Originating Exchange ID (OX_ID), and Responder

Exchange ID (RX_ID), in addition to some control fields

- The frame header also defines the following fields:

1. Routing Control (R_CTL): This field denotes whether the frame is a link control frame or a data frame.

Link control frames are nondata frames that do not carry any payload. These frames are used for setup and messaging.

Data frames carry the payload and are used for data transmission.

2. Class Specific Control (CS_CTL): This field specifies link speeds for class 1 and class 4 data transmission.

3. TYPE: This field describes the upper layer protocol (ULP) to be carried on the frame if it is a data frame.

4. Data Field Control (DF_CTL): A 1-byte field that indicates the existence of any optional headers at the beginning of the data payload.

5. Frame Control (F_CTL): A 3-byte field that contains control information related to frame content. [First sequence]

Structure and Organization of FC Data

- A frame represents a word, a sequence represents a sentence, and an exchange represents a conversation.
    1. **Exchange operation:** An exchange operation enables two N_ports to identify and manage a set of information units. This unit maps to a sequence. Sequences can be both unidirectional and bidirectional.
    2. **Sequence:** A sequence refers to a contiguous set of frames that are sent from one port to another.
    3. **Frame:** A frame is the fundamental unit of data transfer at Layer 2. Each frame can contain up to 2,112 bytes of payload.

Flow Control

- Flow control defines the pace of the flow of data frames during data transmission.

1. BB_Credit: FC uses the *BB_Credit* mechanism for hardware-based flow control.

- BB_Credit controls the maximum number of frames that can be present over the link at any given point in time.
- In a switched fabric, BB_Credit management may take place between any two FC ports.
- The BB_Credit mechanism provides frame acknowledgment through the *Receiver Ready (R_RDY)* primitive.

2. EE_Credit: When an initiator and a target establish themselves as nodes communicating with each other, they exchange the EE_Credit parameters.

- Provides flow control class 1 and class 2 traffic

Classes of Service

- Classes of services to meet the requirements of wide range of applications

	CLASS 1	CLASS 2	CLASS 3
Communication type	Dedicated connection	Nondedicated connection	Nondedicated connection
Flow control	End-to-end credit	End-to-end credit B-to-B credit	B-to-B credit
Frame delivery	In order delivery	Order not guaranteed	Order not guaranteed
Frame acknowl-edgement	Acknowledged	Acknowledged	Not acknowledged
Multiplexing	No	Yes	Yes
Bandwidth utilization	Poor	Moderate	High

** *Explain zoning. Which are the different types of zoning***

5.9 Zoning

- Zoning is an FC switch function that enables nodes within the fabric to be logically segmented into groups that can communicate with each other.
- When a device (host or storage array) logs onto a fabric, it is registered with the name server.
- The zoning function controls the process by allowing only the members in the same zone to establish these link-level services.
- Multiple zone sets may be defined in a fabric, but only one zone set can be active at a time.
- A zone set is a set of zones and a zone is a set of members. A member may be in multiple zones. A member may be in multiple zones.

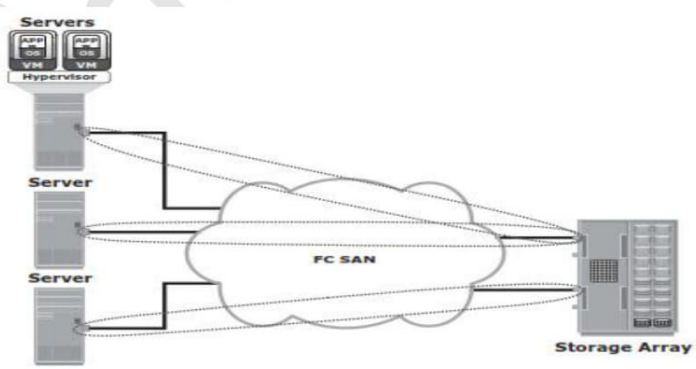
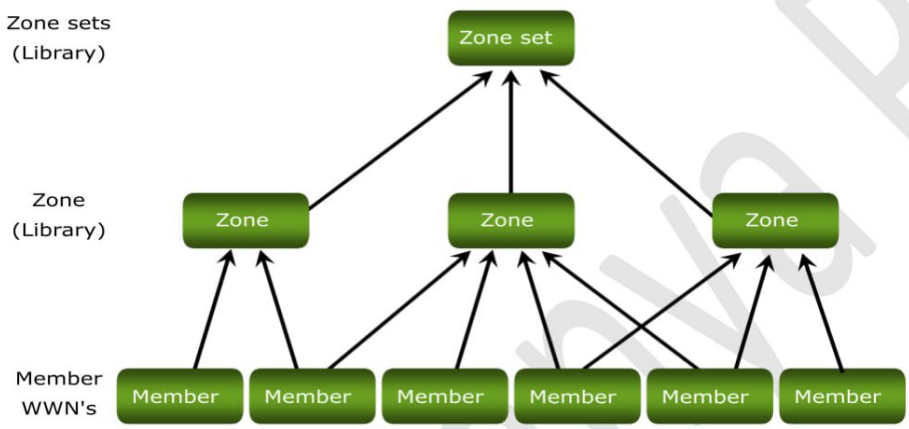


Figure 5-17: Zoning

- Zone members, zones, and zone sets form the hierarchy defined in the zoning process (see Figure 5-18).
- A zone set is composed of a group of zones that can be activated or deactivated as a

single entity in a fabric.

- Multiple zone sets may be defined in a fabric, but only one zone set can be active at a time.
- Members are nodes within the SAN that can be included in a zone.
- ***Switch ports, HBA ports, and storage device ports*** can be members of a zone.
- A port or node can be a member of multiple zones.
- Nodes distributed across multiple switches in a switched fabric may also be grouped into the same zone.
- Zone sets are also referred to as zone configurations.
- Zoning provides control by allowing only the members in the same zone to establish communication with each other.



Types of Zoning

Zoning can be categorized into three types:

1. Port zoning: It uses the FC addresses of the physical ports to define zones. The FC address is dynamically assigned when the port logs on to the fabric. Therefore, any change in the fabric configuration affects zoning.

Port zoning is also called *hard zoning*. Although this method is secure, it requires updating of zoning configuration information in the event of fabric reconfiguration.

2. WWN zoning: It uses World Wide Names to define zones. WWN zoning is also referred to as *soft zoning*.

A major advantage of WWN zoning is its flexibility. It allows the SAN to be recabled without reconfiguring the zone information.

3. Mixed zoning: It combines the qualities of both WWN zoning and port zoning. Using mixed zoning enables a specific port to be tied to the WWN of a node.

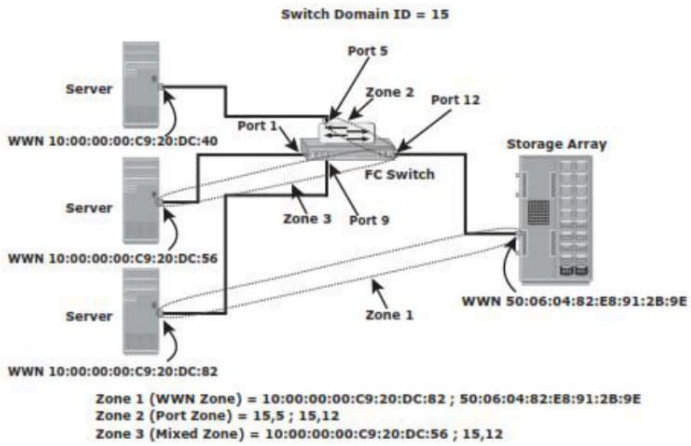


Figure 5-19: Types of zoning

Explain the FC topologies

5.10 FC SAN Topologies

Fabric design follows standard topologies to connect devices. Core-edge fabric is one of the popular topologies for fabric designs. Variations of core-edge fabric and mesh topologies are most commonly deployed in FC SAN implementations

Mesh Topology

- A mesh topology may be one of the two types: full mesh or partial mesh.
- In a full mesh, every switch is connected to every other switch in the topology.
- A full mesh topology may be appropriate when the number of switches involved is small
- A typical deployment would involve up to four switches or directors, with each of them servicing highly localized host-to-storage traffic.
- In a full mesh topology, a maximum of one ISL or hop is required for host-to-storage traffic.
- However, with the increase in the number of switches, the number of switch ports used for ISL also increases.
- This reduces the available switch ports for node connectivity.
- In a partial mesh topology, several hops or ISLs may be required for the traffic to reach its destination.
- Partial mesh offers more scalability than full mesh topology.

- However, without proper placement of host and storage devices, traffic c management in a partial mesh fabric might be complicated and ISLs could become overloaded due to excessive traffic aggregation.

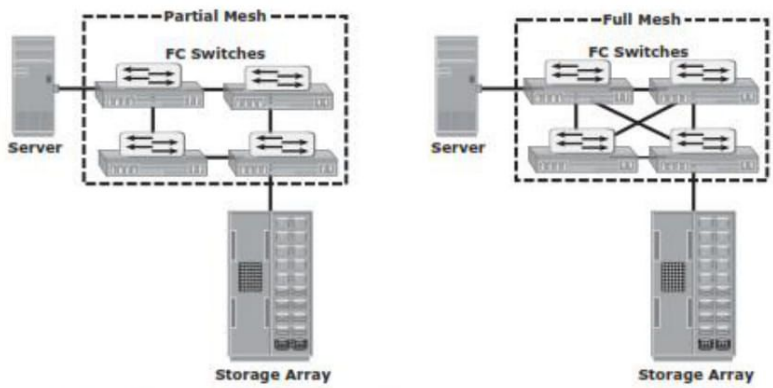


Figure 5-20: Partial mesh and full mesh topologies

Core-Edge Fabric

- The core-edge fabric topology has two types of switch tiers.
- The edge tier is usually composed of switches and offers an inexpensive approach to adding more hosts in a fabric.
- Each switch at the edge tier is attached to a switch at the core tier through ISLs.
- The core tier is usually composed of enterprise directors that ensure high fabric availability.
- In addition, typically all traffic must either traverse this tier or terminate at this tier.
- In this configuration, all storage devices are connected to the core tier, enabling host-to-storage traffic to traverse only one ISL.
- Hosts that require high performance may be connected directly to the core tier and consequently avoid ISL delays.
- In core-edge topology, the edge-tier switches are not connected to each other.
- The core edge fabric topology increases connectivity within the SAN while conserving the overall port utilization.
- If fabric expansion is required, additional edge switches are connected to the core.
- The core of the fabric is also extended by adding more switches or directors at the core tier.
- Based on the number of core-tier switches, this topology has different variations, such as, single-core topology (see Figure 5-21) and dual-core topology (see Figure 5-22).

- To transform a single-core topology to dual-core, new ISLs are created to connect each edge switch to the new core switch in the fabric.

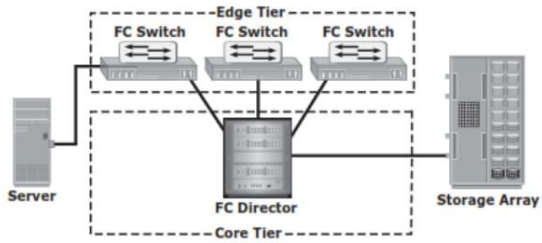


Figure 5-21: Single-core topology

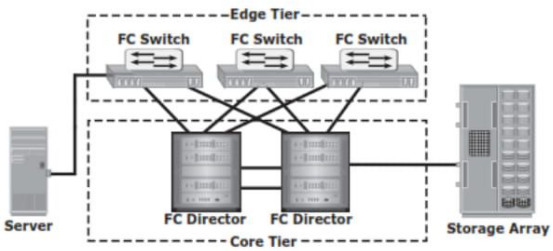


Figure 5-22: Dual-core topology

Benefits and Limitations of Core-Edge Fabric

- Benefits:
- The core-edge fabric provides maximum one-hop storage access to all storage devices in the system.
- Because traffic travels in a deterministic pattern (from the edge to the core and vice versa), a core-edge provides easier calculation of the ISL load and traffic patterns.
- Core-edge fabrics are scaled to larger environments by adding more core switches and linking them, or adding more edge switches.
- This method enables extending the existing simple core-edge model or expanding the fabric into a compound or complex core-edge model.
- Hop count represents the total number of ISLs traversed by a packet between its source and destination.
- A common best practice is to keep the number of host-to-storage hops unchanged, at one hop, in a core-edge.
- Limitations:
- Generally, a large hop count means a high data transmission delay between the source and destination.
- As the number of cores increases, it is prohibitive to continue to maintain ISLs from each core to each edge switch.

- When this happens, the fabric design is changed to a compound or complex core-edge design

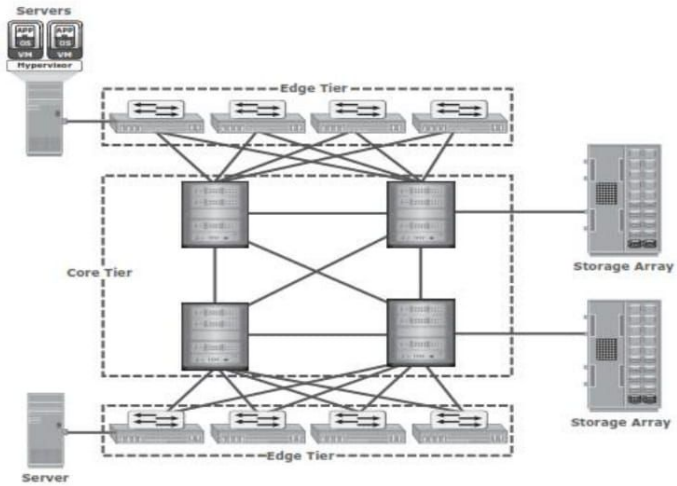


Figure 5-23: Compound core-edge topology

** *Discuss virtualization in SAN***

Virtualization in SAN

- Block-level storage virtualization aggregates block storage devices (LUNs) and enables provisioning of virtual storage volumes, independent of the underlying physical storage.
- A virtualization layer, which exists at the SAN, abstracts the identity of physical storage devices and creates a storage pool from heterogeneous storage devices.
- Virtual volumes are created from the storage pool and assigned to the hosts.
- Instead of being directed to the LUNs on the individual storage arrays, the hosts are directed to the virtual volumes provided by the virtualization layer.
- For hosts and storage arrays, the virtualization layer appears as the target and initiator devices, respectively.
- The virtualization layer maps the virtual volumes to the LUNs on the individual arrays.
- The hosts remain unaware of the mapping operation and access the virtual volumes as if they were accessing the physical storage attached to them.
- Typically, the virtualization layer is managed via a dedicated virtualization appliance to which the hosts and the storage arrays are connected.
- Figure 5 24 illustrates a virtualized environment. It shows two physical servers, each of which has one virtual volume assigned.
- Previously, block-level storage virtualization provided nondisruptive data migration only within a data center.
- The new generation of block-level storage virtualization enables non disruptive data

migration both within and between data centers.

- It provides the capability to connect the virtualization layers at multiple data centers.
- The connected virtualization layers are managed centrally and work as a single virtualization layer stretched across data centers (see Figure 5-25).
- This enables the federation of block storage resources both within and across data centers. The virtual volumes are created from the federated storage resources.

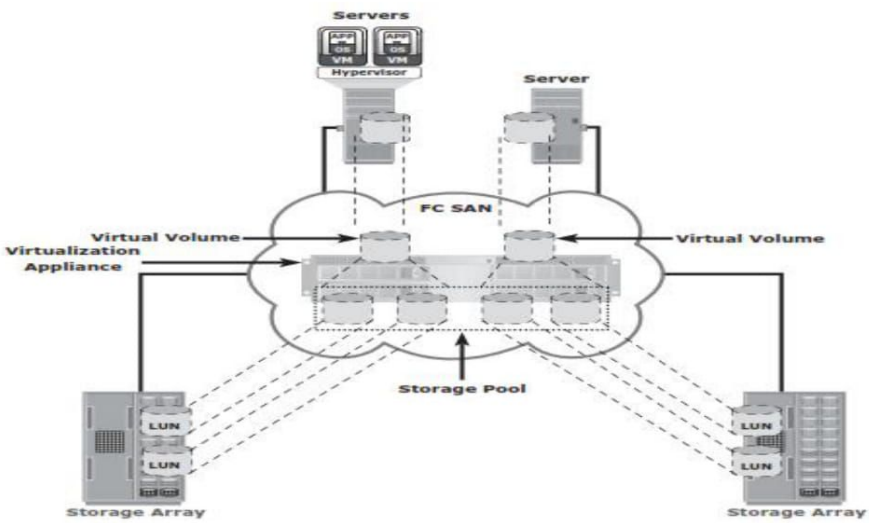


Figure 5-24: Block-level storage virtualization

Virtual SAN (VSAN)

- Virtual SAN (also called virtual fabric) is a logical fabric on an FC SAN, which enables communication among a group of nodes regardless of their physical location in the fabric.
- In a VSAN, a group of hosts or storage ports communicate with each other using a virtual topology defined on the physical SAN.
- Multiple VSANs may be created on a single physical SAN. Each VSAN acts as an independent fabric with its own set of fabric services, such as name server, and zoning.
- Fabric-related configurations in one VSAN do not affect the traffic in another.

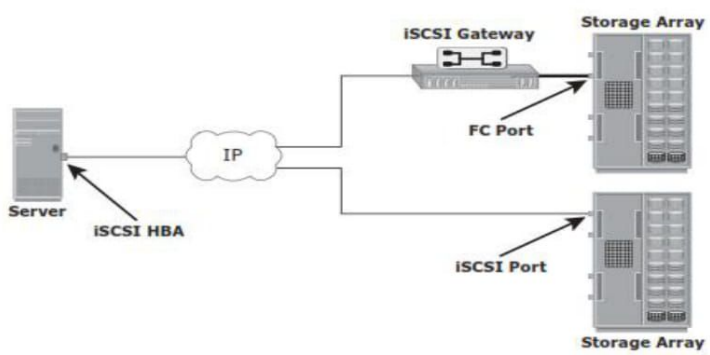


Figure 6-1: iSCSI implementation

Chapter 6: IP SAN and FCoE

6.1 iSCSI

iSCSIis an IP based protocol that establishes and manages connections between host and storage over IP, as shown in Figure 6-1. iSCSI encapsulates SCSI commands and data into an IP packet and transports them using TCP/IP. iSCSI is widely adopted for connecting servers to storage because it is relatively inexpensive and easy to implement, especially in environments in which an FC SAN does not exist.

Components of iSCSI

- An initiator (host), target (storage or iSCSI gateway), and an IP-based network are the key iSCSI components
- If an iSCSI-capable storage array is deployed, then a host with the iSCSI initiator can directly communicate with the storage array over an IP network
- However, in an implementation that uses an existing FC array for iSCSI communication, an iSCSI gateway is used
- These devices perform the translation of IP packets to FC frames and vice versa, thereby bridging the connectivity between the IP and FC environments.

iSCSI Host Connectivity

- A standard NIC with software iSCSI initiator, a TCP offload engine (TOE) NIC with software iSCSI initiator, and an iSCSI HBA are the three iSCSI host connectivity options.
- The function of the iSCSI initiator is to route the SCSI commands over an IP network.
- A standard NIC with a software iSCSI initiator is the simplest and least expensive connectivity option.
- It is easy to implement because most servers come with at least one, and in many cases two, embedded NICs.
- It requires only a software initiator for iSCSI functionality.
- Because NICs provide standard IP function, encapsulation of SCSI into IP packets and decapsulation are carried out by the host CPU.
- This places additional overhead on the host CPU.
- If a standard NIC is used in heavy I/O load situations, the host CPU might become a bottleneck. TOE NIC helps alleviate this burden.
- A TOE NIC offloads TCP management functions from the host and leaves only the iSCSI functionality to the host processor.
- The host passes the iSCSI information to the TOE card, and the TOE card sends the information to the destination using TCP/IP.
- Although this solution improves performance, the iSCSI functionality is still handled by a software

initiator that requires host CPU cycles.

iSCSI Topologies

- Two topologies of iSCSI implementations are native and bridged.
- Native topology does not have FC components.
- The initiators may be either directly attached to targets or connected through the IP network.
- Bridged topology enables the coexistence of FC with IP by providing iSCSI-to-FC bridging functionality.
- For example, the initiators can exist in an IP environment while the storage remains in an FC environment

Native iSCSI Connectivity

FC components are not required for iSCSI connectivity if an iSCSI-enabled array is deployed. In Figure 6-2 (a), the array has one or more iSCSI ports configured with an IP address and is connected to a standard Ethernet switch.

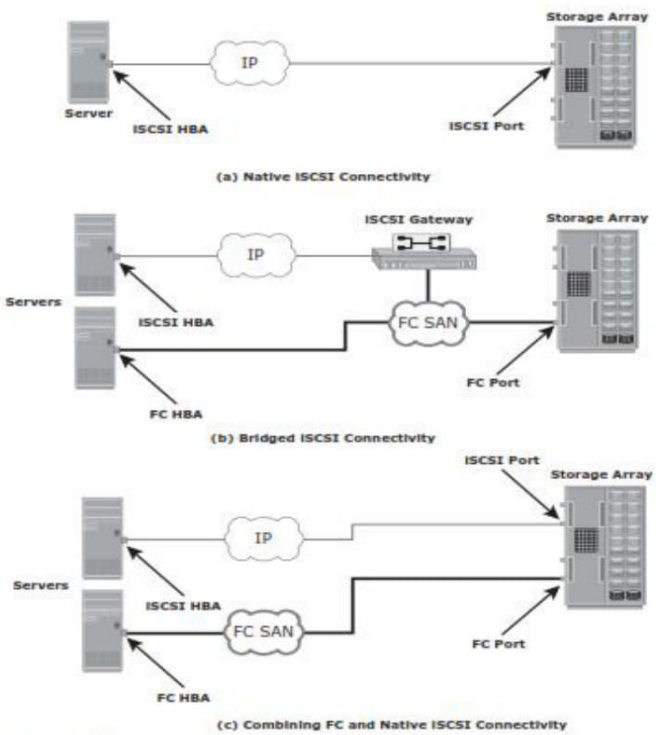


Figure 6-2: iSCSI Topologies

Bridged iSCSI Connectivity

A bridged iSCSI implementation includes FC components in its configuration. Figure 6-2 (b) illustrates iSCSI host connectivity to an FC storage array.

- In this case the array doesnot have any iSCSI ports.

- Therefore an external device called a gateway or a multiprotocol router must be used to facilitate the communication between the iSCSI host and FC storage.
- The gateway converts IP packets to FC frames and vice versa.
- The bridge devices contain both FC and Ethernet ports to facilitate the communication between the FC and IP environments.

Combining FC and Native iSCSI Connectivity

The most common topology is a combination of FC and native iSCSI. Typically, a storage array comes with both FC and iSCSI ports that enable iSCSI and FC connectivity in the same environment, as shown in Figure 6-2 (c).

iSCSI Protocol Stack

Figure 6-3 displays a model of the iSCSI protocol layers and depicts the encapsulation order of the SCSI commands for their delivery through a physical carrier.

- SCSI is the command protocol that works at the application layer of the Open System Interconnection (OSI) model.
- The initiators and targets use SCSI commands and responses to talk to each other.
- The SCSI command descriptor blocks, data, and status messages are encapsulated into TCP/IP and transmitted across the network between the initiators and targets.

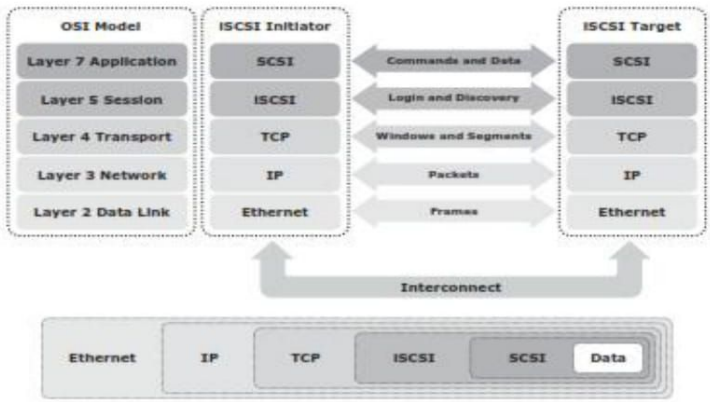


Figure 6-3: iSCSI protocol stack

iSCSI PDU

- A protocol data unit (PDU) is the basic "information unit" in the iSCSI environment.
- The iSCSI initiators and targets communicate with each other using iSCSI PDUs.
- This communication includes establishing iSCSI connections and iSCSI sessions, performing iSCSI discovery, sending SCSI commands and data, and receiving SCSI status.
- All iSCSI PDUs contain one or more header segments followed by zero or more data segments.
- The PDU is then encapsulated into an IP packet to facilitate the transport.

- A PDU includes the components shown in Figure 6-4.
- The IP header provides packet-routing information to move the packet across a network.
- The TCP header contains the information required to guarantee the packet delivery to the target.
- The iSCSI header (basic header segment) describes how to extract SCSI commands and data for the target. iSCSI adds an optional CRC, known as the digest, to ensure datagram integrity.
- This is in addition to TCP checksum and Ethernet CRC.
- The header and the data digests are optionally used in the PDU to validate integrity and data placement.
- As shown in Figure 6-5, each iSCSI PDU does not correspond in a 1:1 relationship with an IP packet.
- Depending on its size, an iSCSI PDU can span an IP packet or even coexist with another PDU in the same packet.

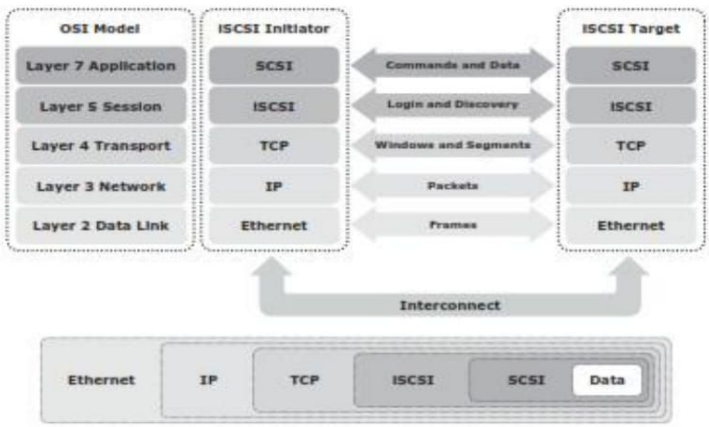


Figure 6-3: iSCSI protocol stack

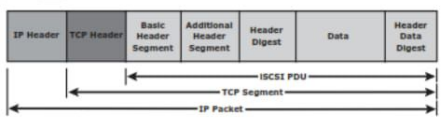


Figure 6-4: iSCSI PDU encapsulated in an IP packet

A message transmitted on a network is divided into a number of packets. If necessary, each packet can be sent by a different route across the network. Packets can arrive in a different order than the order in which they were sent. IP only delivers them; it is up to TCP to organize them in the right sequence. The target extracts the SCSI commands and data on the basis of the information in the iSCSI header.

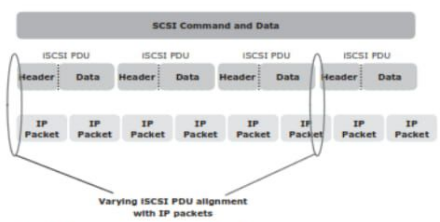


Figure 6-5: Alignment of iSCSI PDUs with IP packets

iSCSI Discovery

An initiator must discover the location of its targets on the network and the names of the targets available to it before it can establish a session. This discovery can take place in two ways: Send Targets discovery or internet Storage Name Service (iSNS).

- iSNS (see Figure 6-6) enables automatic discovery of iSCSI devices on an IP network.
- The initiators and targets can be configured to automatically register themselves with the iSNS server.
- Whenever an initiator wants to know the targets that it can access, it can query the iSNS server for a list of available targets.

iSCSI Names

A unique worldwide iSCSI identifier, known as an iSCSI name, is used to identify the initiators and targets within an iSCSI network to facilitate communication.

Following are two types of iSCSI names commonly used:

- iSCSI Qualified Name (IQN): An organization must own a registered domain name to generate iSCSI Qualified Names. This domain name does not need to be active or resolve to an address. It just needs to be reserved to prevent other organizations from using the same domain name to generate iSCSI names. A date is included in the name to avoid potential conflicts caused by the transfer of domain names. An example of an IQN is iqn.2008- 02.com.example:optional_string. The optional_string provides a serial number, an asset number, or any other device identifiers. An iSCSI Qualified Name enables storage administrators to assign meaningful names to iSCSI devices, and therefore, manage those devices more easily.
- Extended Unique Identifier (EUI): An EUI is a globally unique identifier based on the IEEE EUI-64 naming standard. An EUI is composed of the eui prefix followed by a 16-character hexadecimal name, such as eui.0300732A32598D26.

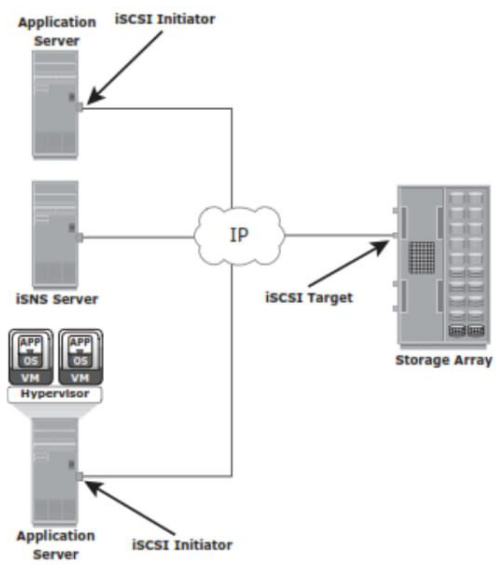


Figure 6-6: Discovery using iSNS

iSCSI Session

An iSCSI session is established between an initiator and a target, as shown in Figure 6-7. A session is identified by a session ID (SSID), which includes part of an initiator ID and a target ID. The session can be intended for one of the following:

- The discovery of the available targets by the initiators and the location of a specific target on a network
- The normal operation of iSCSI (transferring data between initiators and targets)

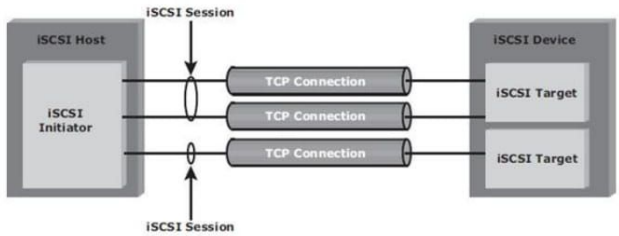


Figure 6-7: iSCSI session

iSCSI Command Sequencing

- The iSCSI communication between the initiators and targets is based on the request- response command sequences.
- A command sequence may generate multiple PDUs.
- A command sequence number (CmdSN) within an iSCSI session is used for numbering all initiator-to-target command PDUs belonging to the session.
- This number ensures that every command is delivered in the same order in which it is transmitted, regardless of the TCP connection that carries the command in the session.

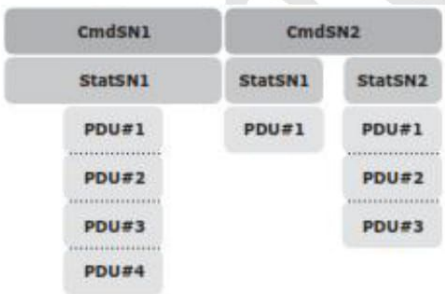


Figure 6-8: Command and status sequence number

6.2 FCIP

FC SAN provides a high-performance infrastructure for localized data movement. Organizations are now looking for ways to transport data over a long distance between their disparate SANs at multiple geographic locations. One of the best ways to achieve this goal is to interconnect geographically dispersed SANs through reliable, high- speed links.

FCIP Protocol Stack

The FCIP protocol stack is shown in Figure: 6-9. Applications generate SCSI commands and data, which are processed by various layers of the protocol stack.

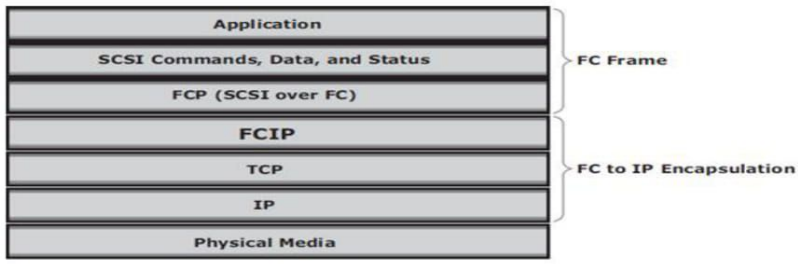


Figure 6-9: FCIP protocol stack

The FCIP layer encapsulates the Fibre Channel frames onto the IP payload and passes them to the TCP layer (see Figure 6-10). TCP and IP are used for transporting the encapsulated information across Ethernet, wireless, or other media that support the TCP/IP traffic.

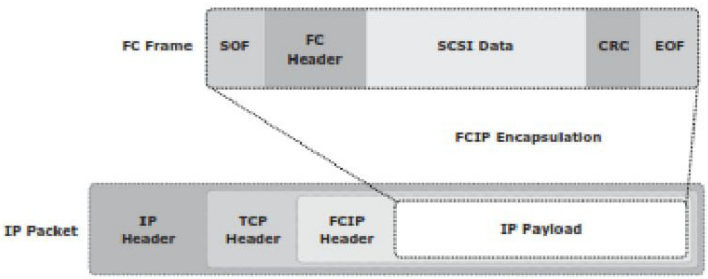


Figure 6-10: FCIP encapsulation

FCIP Topology

In an FCIP environment, an FCIP gateway is connected to each fabric via a standard FC connection (see Figure 6-11). The FCIP gateway at one end of the IP network encapsulates the FC frames into IP packets. The gateway at the other end removes the IP wrapper and sends the FC data to the layer 2 fabric.

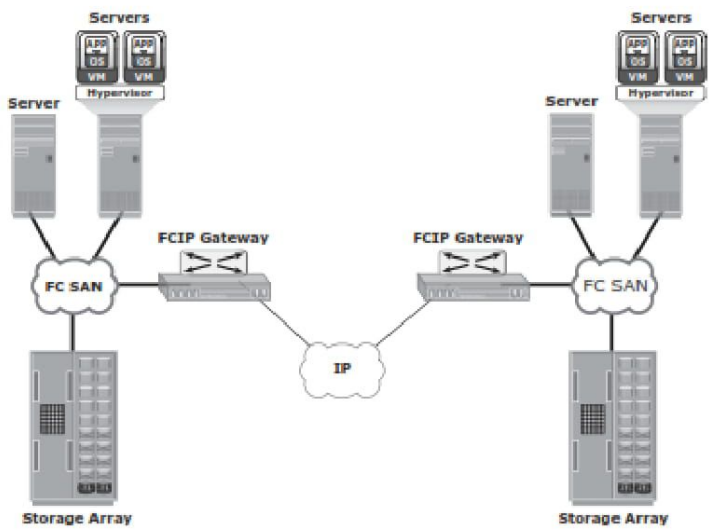


Figure 6-11: FCIP topology

FCIP Performance and Security

Performance, reliability, and security should always be taken into consideration when implementing storage solutions. The implementation of FCIP is also subject to the same considerations.

- Configuring multiple paths between FCIP gateways eliminates single point of failure and provides increased bandwidth
- In the case of extended distance the IP network might be a bottleneck is sufficient bandwidth is not available
- Because FCIP creates a unified fabric disruption in the underlying network can cause instabilities in the SAN environment.

6.3 FCoE

Data centers typically have multiple networks to handle various types of I/O traffic — for example, an Ethernet network for TCP/IP communication and an FC network for FC communication. TCP/IP is typically used for client-server.

Fibre Channel over Ethernet (FCoE) protocol provides consolidation of LAN and SAN traffi c over a single physical interface infrastructure. FCoE helps organizations address the challenges of having multiple discrete network infrastructures.

I/O Consolidation Using FCoE

The key benefit of FCoE is I/O consolidation. Figure 6-12 represents the infrastructure before FCoE deployment. Here, the storage resources are accessed using HBAs, and the IP network resources are accessed using NICs by the servers. Typically, in a data center, a server is configured with 2 to 4 NIC cards and redundant HBA cards. If the data center has hundreds of servers, it would require a large number of adapters, cables, and switches. This leads to a complex environment, which is difficult to manage and scale

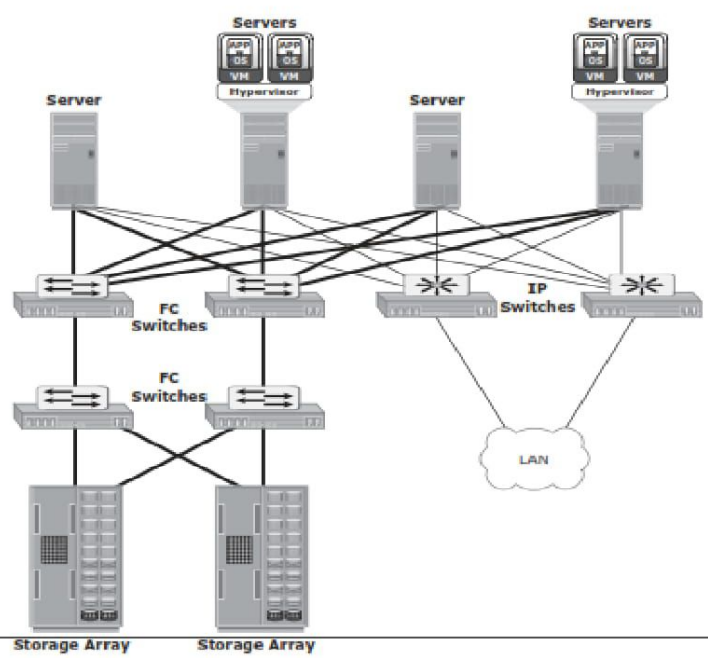


Figure 6-12: Infrastructure before using FCoE

Figure 6-13 shows the I/O consolidation with FCoE using FCoE **switches and Converged Network Adapters (CNAs).** A CNA replaces both HBAs and NICs in the server and consolidates both the IP and FC traffic. This reduces the requirement of multiple network adapters at the server to connect to different networks. Overall, this reduces the requirement of adapters, cables, and switches. This also considerably reduces the cost and management overhead.

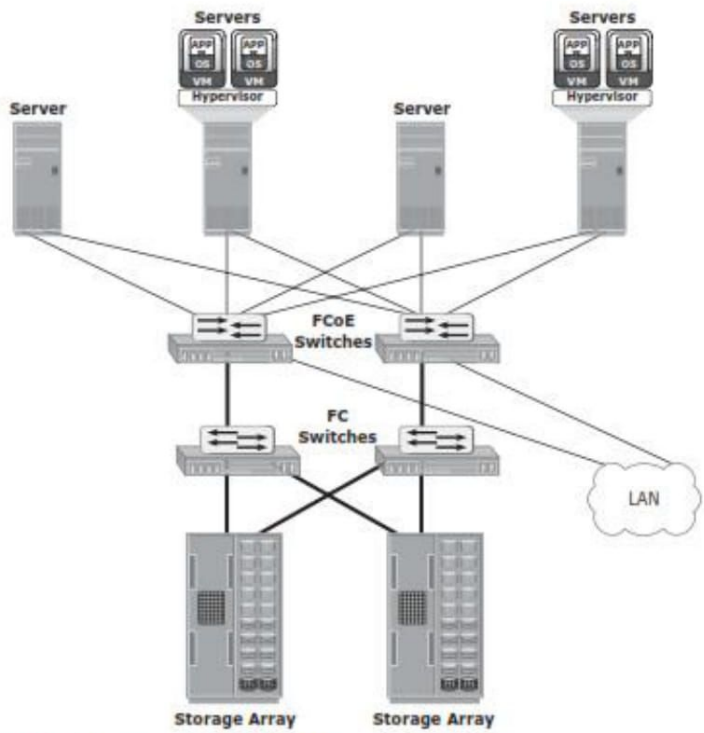


Figure 6-13: Infrastructure after using FCoE

Components of an FCoE Network

- Converged Network Adapter (CNA)
- Cables
- FCoE switches

Converged Network Adapter

- A CNA provides the functionality of both a standard NIC and an FC HBA in a single adapter and consolidates both types of traffic.
- CNA eliminates the need to deploy separate adapters and cables for FC and Ethernet communications, thereby reducing the required number of server slots and switch ports.
- As shown in Figure 6-14, a CNA contains separate modules for 10 Gigabit Ethernet, Fibre Channel, and FCoE Application Specific Integrated Circuits (ASICs).
- The FCoE ASIC encapsulates FC frames into Ethernet frames.
- One end of this ASIC is connected to 10GbE and FC ASICs for server connectivity, while the other end provides a 10GbE interface to connect to an FCoE switch.

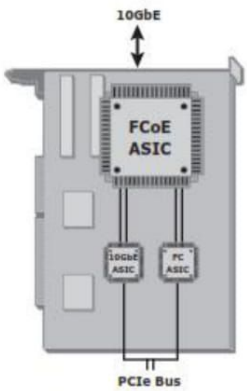


Figure 6-14: Converged Network Adapter

Cables

Currently two options are available for FCoE cabling: Copper based Twinax and standard fiber optical cables.

- A Twinax cable is composed of two pairs of copper cables covered with a shielded casing.
- The Twinax cable can transmit data at the speed of 10 Gbps over shorter distances up to 10 meters.
- Twinax cables require less power and are less expensive than fiber optic cables.

FCoE Switches

An FCoE switch has both Ethernet switch and Fibre Channel switch functionalities.

- The FCoE switch has a Fibre Channel Forwarder (FCF), Ethernet Bridge, and set of Ethernet ports and optional FC ports, as shown in Figure 6-15.
- The function of the FCF is to encapsulate the FC frames, received from the FC port,into the FCoE frames and also to de-encapsulate the FCoE frames, received from the Ethernet Bridge, to the FC frames.

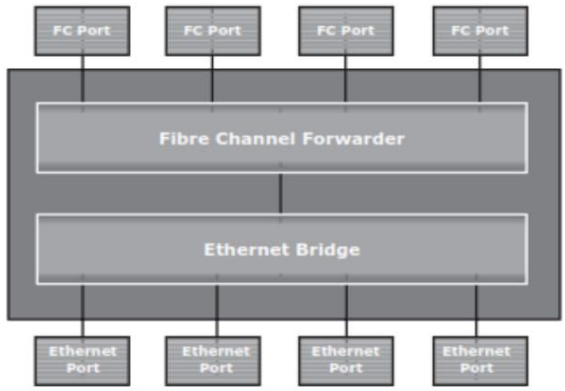


Figure 6-15: FCoE switch generic architecture

- Upon receiving the incoming traffic, the FCoE switch inspects the Ethertype (used to indicate which protocol is encapsulated in the payload of an Ethernet frame) of the incoming frames and uses that to determine the destination.
- If the Ethertype of the frame is FCoE, the switch recognizes that the frame contains an FC payload and forwards it to the FCF.
- From there, the FC is extracted from the FCoE frame and transmitted to FC SAN over the FC ports.
- If the Ethertype is not FCoE, the switch handles the traffi c as usual Ethernet traffi c and forwards it over the Ethernet ports.

FCoE Frame Structure

- An FCoE frame is an Ethernet frame that contains an FCoE Protocol Data Unit. Figure 6- 16 shows the FCoE frame structure.
- The first 48-bits in the frame are used to specify the destination MAC address, and the next 48-bits specify the source MAC address.
- The 32- bit IEEE 802.1Q tag supports the creation of multiple virtual networks (VLANs) across a single physical infrastructure.
- FCoE has its own Ethertype, as designated by the next 16 bits, followed by the 4-bit version field.
- The next 100-bits are reserved and are followed by the 8-bit Start of Frame and then the actual FC frame.
- The 8-bit End of Frame delimiter is followed by 24 reserved bits. The frame ends with the final 32-bits dedicated to the Frame Check Sequence (FCS) function that provides error detection for the Ethernet frame.

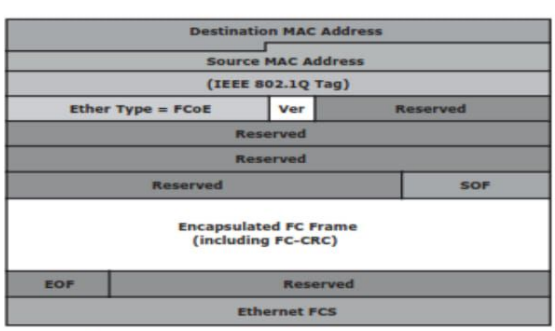


Figure 6-16: FCoE frame structure

FCoE Frame Mapping

The encapsulation of the Fibre Channel frame occurs through the mapping of the FC frames onto Ethernet, as shown in Figure 6-17. Fibre Channel and traditional networks have stacks of layers where each layer in the stack represents a set of functionalities. The FC stack consists of fi ve layers: FC-0 through FC-4. Ethernet is typically considered as a set of protocols that operates at the physical and data link layers in the seven layer OSI stack. The FCoE protocol specification replaces the FC-0 and FC-1 layers of the FC stack with Ethernet.

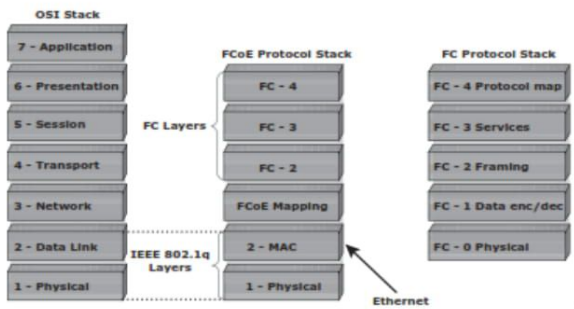


Figure 6-17: FCoE frame mapping

FCoE Enabling Technologies

- Conventional Ethernet is lossy in nature, which means that frames might be dropped or lost during transmission.
- Converged Enhanced Ethernet (CEE), or lossless Ethernet, provides a new specification to the existing Ethernet standard that eliminates the lossy nature of Ethernet.
- This makes 10 Gb Ethernet a viable storage networking option, similar to FC.
- Lossless Ethernet requires certain functionalities, they are:
    - Priority-based flow control
    - Enhanced transmission selection
    - Congestion Notification
    - Data center bridging exchange protocol

Priority-Based Flow Control (PFC)

- PFC provides a link level flow control mechanism.
- PFC creates eight separate virtual links on a single physical link and allows any of these links to be paused and restarted independently.
- PFC enables the pause mechanism based on user priorities or classes of service.
- Enabling the pause based on priority allows creating lossless links for traffic, such as FCoE traffic.
- This PAUSE mechanism is typically implemented for FCoE while regular TCP/IP traffic continues to drop frames.
- Figure 6-18 illustrates how a physical Ethernet link is divided into eight virtual links and

allows a PAUSE for a single virtual link without affecting the traffic for the others.

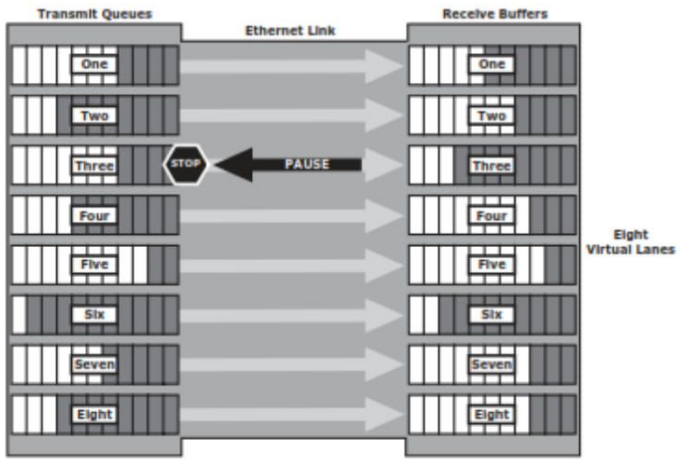


Figure 6-18: Priority-based flow control

Enhanced Transmission Selection (ETS)

- Enhanced transmission selection provides a common management framework for the assignment of bandwidth to different traffic classes, such as LAN, SAN, and Inter Process Communication (IPC).
- When a particular class of traffic does not use its allocated bandwidth, ETS enables other traffic classes to use the available bandwidth.

Congestion Notification (CN)

- Congestion notification provides end-to-end congestion management for protocols, such as FCoE, that do not have built-in congestion control mechanisms.
- Link level congestion notification provides a mechanism for detecting congestion and notifying the source to move the traffic flow away from the congested links.
- Link level congestion notification enables a switch to send a signal to other ports that need to stop or slow down their transmissions.
- The process of congestion notification and its management is shown in Figure 6-19, which represents the communication between the nodes A (sender) and B (receiver).

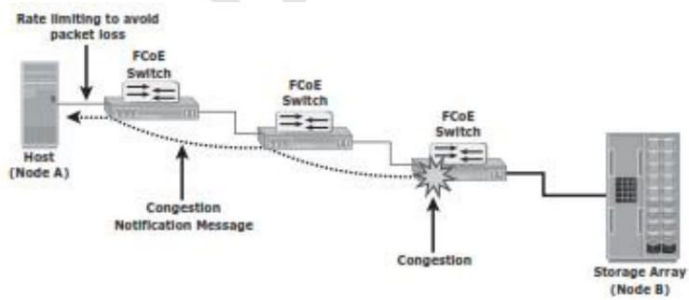


Figure 6-19: Congestion Notification

Chapter 7: Network-Attached Storage

Network-attached storage (NAS) is a dedicated, high-performance file sharing and storage device. NAS enables its clients to share files over an IP network.

- NAS provides the advantages of server consolidation by eliminating the need for multiple file servers.
- It also consolidates the storage used by the clients onto a single system, making it easier to manage the storage.
- A NAS device uses its own operating system and integrated hardware and software components to meet specific file-service needs.
- Its operating system is optimized for file I/O and, therefore, performs file I/O better than a general-purpose server.
- As a result, a NAS device can serve more clients than general-purpose servers and provide the benefit of server consolidation.

7.4 Components of NAS

A NAS device has two key components: NAS head and storage (see Figure 7-3). In some NAS implementations, the storage could be external to the NAS device and shared with other hosts. The NAS head includes the following components:

- CPU and memory
- One or more network interface cards (NICs), which provide connectivity to the client network. Examples of network protocols supported by NIC include Gigabit Ethernet, Fast Ethernet, ATM, and Fiber Distributed Data Interface (FDDI).
- An optimized operating system for managing the NAS functionality. It translates file-level requests into block-storage requests and further converts the data supplied at the block level to file data.
- NFS, CIFS, and other protocols for file sharing
- Industry-standard storage protocols and ports to connect and manage physical disk resources

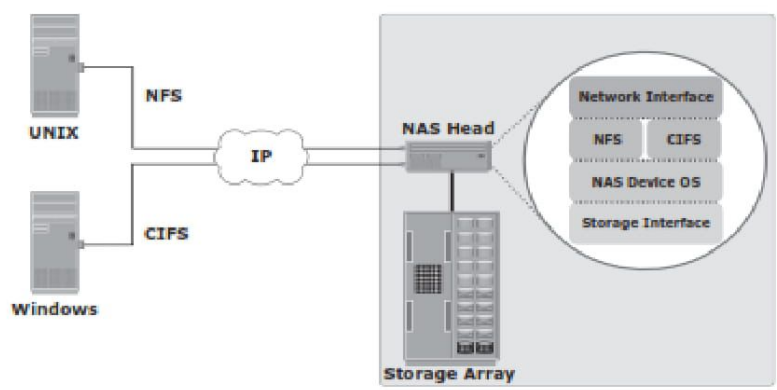


Figure 7-3: Components of NAS

7.5 NAS I/O Operation

NAS provides file-level data access to its clients. File I/O is a high-level request that specifies the file to be accessed. For example, a client may request a file by specifying its name, location, or other attributes. The NAS operating system keeps track of the location of files on the disk volume and converts client file I/O into block-level I/O to retrieve data. The process of handling I/Os in a NAS environment is as follows:

1. The requestor (client) packages an I/O request into TCP/IP and forwards it through the network stack. The NAS device receives this request from the network.
2. The NAS device converts the I/O request into an appropriate physical storage request, which is a block-level I/O, and then performs the operation on the physical storage.
3. When the NAS device receives data from the storage, it processes and repackages the data into an appropriate file protocol response.
4. The NAS device packages this response into TCP/IP again and forwards it to the client through the network. Figure 7-4 illustrates this process.

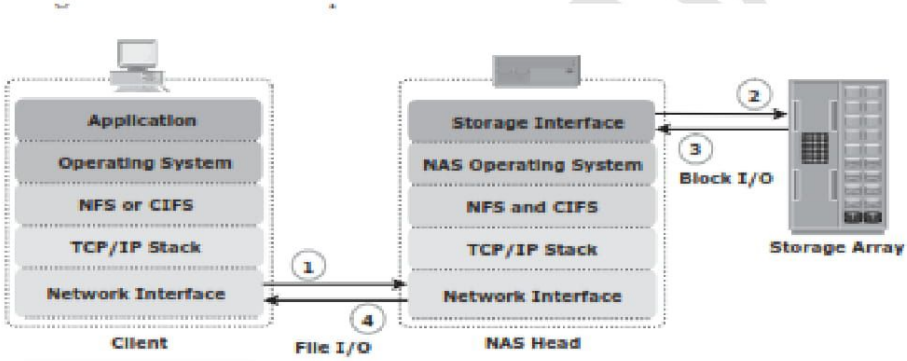


Figure 7-4: NAS I/O operation

7.7 NAS File-Sharing Protocols

Most NAS devices support multiple file-service protocols to handle file I/O requests to a remote file system. ***NFS and CIFS are the common protocols for file sharing***. NAS devices enable users to share file data across different operating environments and provide a means for users to migrate transparently from one operating system to another.

NFS

NFS is a client-server protocol for file sharing that is commonly used on UNIX systems.

- NFS was originally based on the connectionless User Datagram Protocol (UDP). It uses a machine-independent model to represent user data.

- It also uses Remote Procedure Call (RPC) as a method of inter-process communication between two computers.
- The NFS protocol provides a set of RPCs to access a remote fi le system for the following operations:
    - Searching files and directories
    - Opening, reading, writing to, and closing a file
    - Changing file attributes
    - Modifying file links and directories

- NFS creates a connection between the client and the remote system to transfer data.
- NFS (NFSv3 and earlier) is a *stateless protocol*, which means that it does not maintain any kind of table to store information about open fi les and associated pointers.
- Therefore, each call provides a full set of arguments to access files on the server.

- Currently, three versions of NFS are in use:
    1. ***NFS version 2 (NFSv2):*** Uses UDP to provide a stateless network connection between a client and a server. Features, such as locking, are handled outside the protocol.
    2. ***NFS version 3 (NFSv3):*** The most commonly used version, which uses UDP or TCP, and is based on the stateless protocol design. It includes some new features, such as a 64-bit file size, asynchronous writes, and additional fi le attributes to reduce refetching.
    3. ***NFS version 4 (NFSv4):*** Uses TCP and is based on a stateful protocol design. It offers enhanced security. The latest NFS version 4.1 is the enhancement of NFSv4 and includes some new features, such as session model, parallel NFS (pNFS), and data retention.

CIFS

CIFS is a client-server application protocol that enables client programs to make requests for files and services on remote computers over TCP/IP. It is a public, or open, variation of Server Message Block (SMB) protocol.

- It uses file and record locking to prevent users from overwriting the work of another user on a file or a record.
- It supports fault tolerance and can automatically restore connections and reopen files that were open prior to an interruption. The fault tolerance features of CIFS depend on whether an application is written to take advantage of these features.

7.9 File-Level Virtualization

- File-level virtualization eliminates the dependencies between the data accessed at the file level and the location where the files are physically stored.
- Implementation of file-level virtualization is common in NAS or file-server environments.

- It provides non-disruptive file mobility to optimize storage utilization.
- Before virtualization, each host knows exactly where its file resources are located.
- This environment leads to underutilized storage resources and capacity problems because files are bound to a specific NAS device or file server.
- It may be required to move the files from one server to another because of performance reasons or when the file server fills up.
- Moving files across the environment is not easy and may make files inaccessible during file movement.
- Moreover, hosts and applications need to be reconfigured to access the file at the new location.
- This makes it difficult for storage administrators to improve storage efficiency while maintaining the required service level.
- File-level virtualization simplifies file mobility.
- It provides user or application independence from the location where the files are stored. File-level virtualization creates a logical pool of storage, enabling users to use a logical path, rather than a physical path, to access files.
- File-level virtualization facilitates the movement of files across the online file servers or NAS devices.
- This means that while the files are being moved, clients can access their files non-disruptively.
- Clients can also read their files from the old location and write them back to the new location without realizing that the physical location has changed.
- A global namespace is used to map the logical path of a file to the physical path names. Figure 7-9 illustrates a file-serving environment before and after the implementation of file-level virtualization.

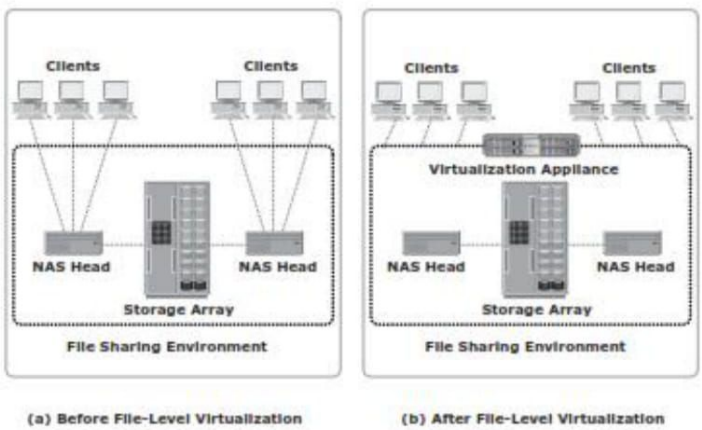


Figure 7-9: File-serving environment before and after file-level virtualization

Chapter 8: Object-Based and Unified Storage

Object-based storage is a way to store file data in the form of objects based on its content and other attributes rather than the name and location.

8.1 Object-Based Storage Devices

- An OSD is a device that organizes and stores unstructured data, such as movies, office documents, and graphics, as objects.
- Object-based storage provides a scalable, self- managed, protected, and shared storage option.
- OSD stores data in the form of objects.
- OSD uses flat address space to store data. Therefore there is no hierarchy of directories and files; as a result, a large number of objects can be stored in an OSD system (see Figure 8-1).

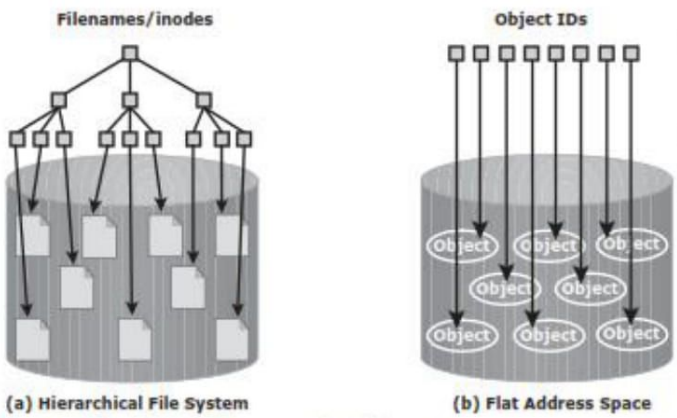


Figure 8-1: Hierarchical file system versus flat address space

- An object might contain user data, related metadata (size, date, ownership, and so on), and other attributes of data (retention, access pattern, and so on); see Figure 8-2.
- Each object stored in the system is identified by a unique ID called the object ID.
- The object ID is generated using specialized algorithms such as hash function on the data and guarantees that every object is uniquely identified.

Figure 8-2: Object structure

Object-Based Storage Architecture

- An I/O in the traditional block access method passes through various layers in the I/O path.
- The I/O generated by an application passes through the file system, the channel, or network and reaches the disk drive.
- When the file system receives the I/O from an application, the file system maps the incoming I/O to the disk blocks.
- The block interface is used for sending the I/O over the channel or network to the storage device.
- The I/O is then written to the block allocated on the disk drive. Figure 8-3 (a) illustrates the block- level access.

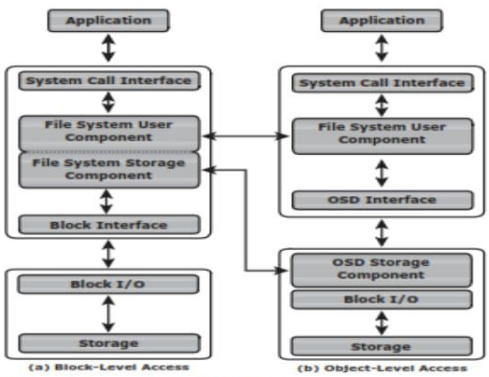


Figure 8-3: Block-level access versus object-level access

- The file system has two components: user component and storage component.
- The user component of the file system performs functions such as hierarchy management, naming, and user access control.
- The storage component maps the files to the physical location on the disk drive.
- When an application accesses data stored in OSD, the request is sent to the file system user component.
- The file system user component communicates to the OSD interface, which in turn sends the request to the storage device.
- The storage device has the OSD storage component responsible for managing the access to the object on a storage device. Figure 8-3 (b) illustrates the object-level access.
- After the object is stored, the OSD sends an acknowledgment to the application server.

Components of OSD

The OSD system is typically composed of three key components:

- Nodes
- private network and
- storage

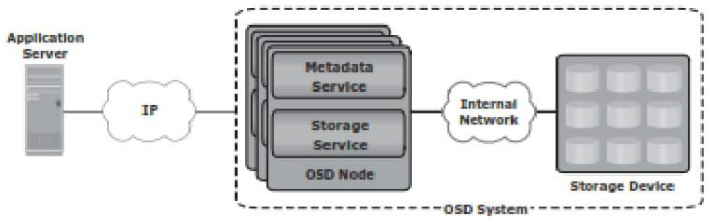


Figure 8-4: OSD components

- The OSD system is composed of one or more *nodes*.
- A node is a server that runs the OSD operating environment and provides services to store, retrieve, and manage data in the system.
- The OSD node has two key services: metadata service and storage service.
- The metadata service is responsible for generating the object ID from the contents (and can also include other attributes of data) of a file.
- It also maintains the mapping of the object IDs and the file system namespace.
- The storage service manages a set of disks on which the user data is stored.
- The OSD nodes connect to the storage via an internal network.
- The internal network provides node-to-node connectivity and node-to-storage connectivity.

Object Storage and Retrieval in OSD

The process of storing objects in OSD is illustrated in Figure 8-5. The data storage process in an OSD system is as follows:

1. The application server presents the file to be stored to the OSD node.
2. The OSD node divides the file into two parts: user data and metadata.
3. The OSD node generates the object ID using a specialized algorithm. The algorithm is executed against the contents of the user data to derive an ID unique to this data.
4. For future access, the OSD node stores the metadata and object ID using the metadata service.
5. The OSD node stores the user data (objects) in the storage device using the storage service.
6. An acknowledgment is sent to the application server stating that the object is stored

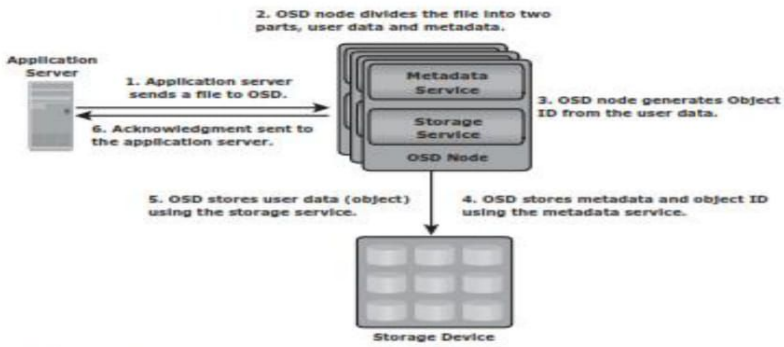


Figure 8-5: Storing objects on OSD

The process of retrieving objects in OSD is illustrated in Figures 8-6. The process of data retrieval from OSD is as follows:

1. The application server sends a read request to the OSD system.
2. The metadata service retrieves the object ID for the requested file.
3. The metadata service sends the object ID to the application server.
4. The application server sends the object ID to the OSD storage service for object retrieval.
5. The OSD storage service retrieves the object from the storage device.
6. The OSD storage service sends the file to the application server.

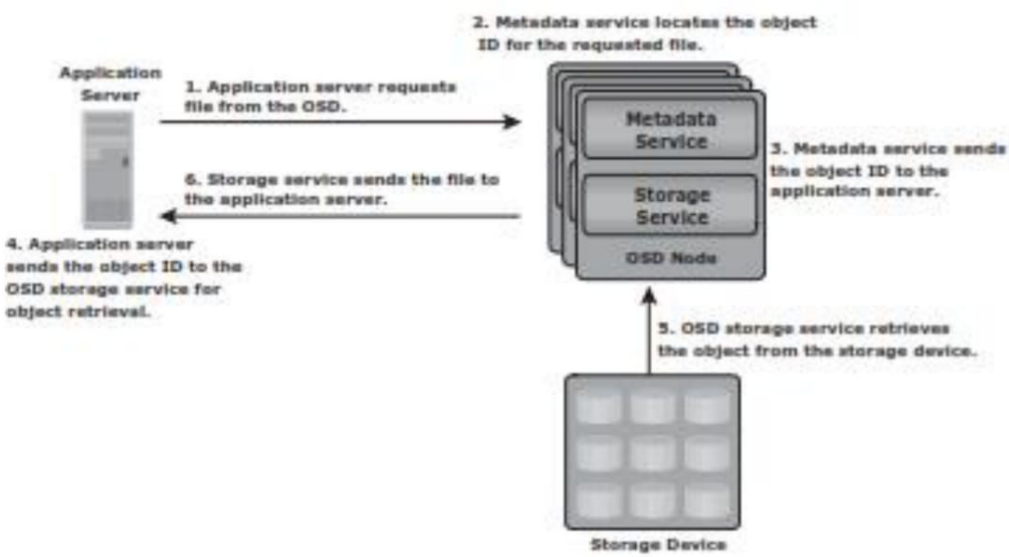


Figure 8-6: Object retrieval from an OSD system

Benefits of Object-Based Storage

The key benefits of object-based storage are as follows:

- Security and reliability: Data integrity and content authenticity are the key features of object-based storage devices. OSD uses specialized algorithms to create objects that provide strong data encryption capability. In OSD, request authentication is performed at the storage device rather than with an external authentication mechanism.

- Platform independence: Objects are abstract containers of data, including metadata and attributes. This feature allows objects to be shared across heterogeneous platforms locally or remotely. This platform-independence capability makes object-based storage the best candidate for cloud computing environments.
- Scalability: Due to the use of flat address space, object-based storage can handle large amounts of data without impacting performance. Both storage and OSD nodes can be scaled independently in terms of performance and capacity.
- Manageability: Object-based storage has an inherent intelligence to manage and protect objects. It uses self-healing capability to protect and replicate objects. Policy-based management capability helps OSD to handle routine jobs automatically.

Common Use Cases for Object-Based Storage

- A **data archival solution** is a promising use case for OSD. Data integrity and protection is the primary requirement for any data archiving solution. Traditional archival solutions — CD and DVD-ROM — do not provide scalability and performance. OSD stores data in the form of objects, associates them with a unique object ID, and ensures high data integrity. Along with integrity, it provides scalability and data protection. These capabilities make OSD a viable option for long term data archiving for fixed content.
- **Content addressed storage (CAS)** is a special type of object-based storage device purposely built for storing fixed content.
- Another use case for OSD is **cloud-based storage. OSD** uses a web interface to access storage resources. OSD provides inherent security, scalability, and automated data management. It also enables data sharing across heterogeneous platforms or tenants while ensuring integrity of data. These capabilities make OSD a strong option for cloud-based storage. Cloud service providers can leverage OSD to offer storage-as-a-service.

OSD supports web service access via representational state transfer (REST) and simple object access protocol (SOAP).

8.2 Content-Addressed Storage

CAS is an object-based storage device designed for secure online storage and retrieval of fixed content. CAS stores user data and its attributes as an object. The stored object is assigned a globally unique address, known as a content address (CA).

CAS provides all the features required for storing fixed content. The key features of CAS are as follows:

- Content authenticity: It assures the genuineness of stored content. This is achieved by generating a unique content address for each object and validating the content address for stored objects at regular intervals. Content authenticity is assured because the address assigned to each object is as unique as a fingerprint. Every time

an object is read, CAS uses a hashing algorithm to recalculate the object's content address as a validation step and compares the result to its original content address. If the object fails validation, CAS rebuilds the object using a mirror or parity protection scheme.

- Content integrity: It provides assurance that the stored content has not been altered. CAS uses a hashing algorithm for content authenticity and integrity. If the fixed content is altered, CAS generates a new address for the altered content, rather than overwrite the original fixed content.
- Location independence: CAS uses a unique content address, rather than directory path names or URLs, to retrieve data. This makes the physical location of the stored data irrelevant to the application that requests the data.

  Single-instance storage (SIS): CAS uses a unique content address to guarantee the storage of only a single instance of an object. When a new object is written, the CAS system is polled to see whether an object is already available with the same content address. If the object is available in the system, it is not stored; instead, only a pointer to that object is created.
- Retention enforcement: Protecting and retaining objects is a core requirement of an archive storage system. After an object is stored in the CAS system and the retention policy is defined, CAS does not make the object available for deletion until the policy expires.
- Data protection: CAS ensures that the content stored on the CAS system is available even if a disk or a node fails. CAS provides both local and remote protection to the data objects stored on it.
- Fast record retrieval: CAS stores all objects on disks, which provides faster access to the objects compared to tapes and optical discs.
- Load balancing: CAS distributes objects across multiple nodes to provide maximum throughput and availability.
- Scalability: CAS allows the addition of more nodes to the cluster without any interruption to data access and with minimum administrative overhead.
- Event notification: CAS continuously monitors the state of the system and raises an alert for any event that requires the administrator's attention. The event notification is communicated to the administrator through SNMP, SMTP, or e-mail.
- Self-diagnosis and repair: CAS automatically detect and repairs corrupted objects and alerts the administrator about the potential problem. CAS systems can be configured to alert remote support teams who can diagnose and repair the system remotely.
- Audit trails: CAS keeps track of management activities and any access or disposition of data. Audit trails are mandated by compliance requirements.

8.4 Unified Storage

Unified storage consolidates block, file, and object access into one storage solution. It supports multiple protocols, such as CIFS, NFS, iSCSI, FC, FCoE, REST (representational state transfer), and SOAP (simple object access protocol).

Components of Unified Storage

A unified storage system consists of the following key components: storage controller, NAS head, OSD node, and storage. Figure 8-9 illustrates the block diagram of a unified storage platform.

- **The storage controller** provides block-level access to application servers through iSCSI, FC, or FCoE protocols.
- It contains iSCSI, FC, and FCoE front-end ports for direct block access.
- The storage controller is also responsible for managing the back-end storage pool in the storage system.
- The controller configures LUNs and presents them to application servers, NAS heads, and OSD nodes.
- A **NAS head** is a dedicated file server that provides file access to NAS clients.
- A *NAS head* is a dedicated fi le server that provides file access to NAS clients.
- The NAS head is connected to the storage via the storage controller typically using a FC or FCoE connection.
- The system typically has two or more NAS heads for redundancy
- The **OSD node** accesses the storage through the storage controller using a FCor FCoE connection.

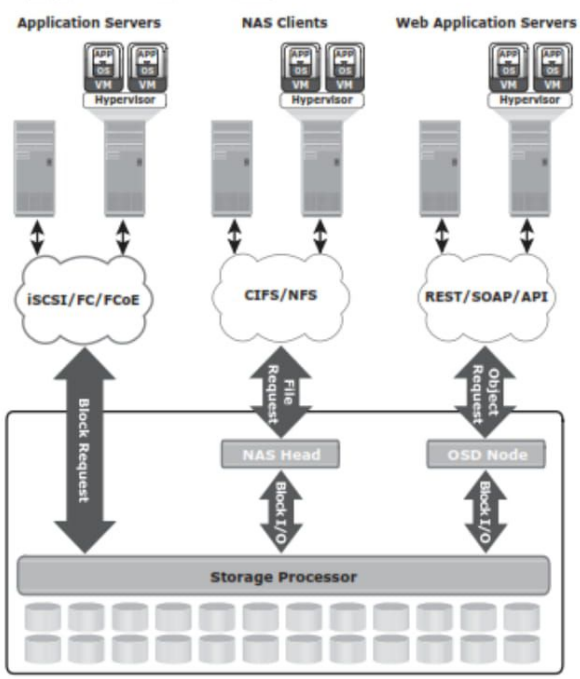


Figure 8-9: Unified storage platform

Data Access from Unified Storage

In a unified storage system, block, fi le, and object requests to the storage travel through different I/O paths. Figure 8-9 illustrates the different I/O paths for block, file, and object access.

- ***Block I/O request:*** The application servers are connected to an FC, iSCSI, or FCoE port on the storage controller. The server sends a block request over an FC, iSCSI, or FCoE connection. The storage processor (SP) processes the I/O and responds to the application server.
- ***File I/O request:*** The NAS clients (where the NAS share is mounted or mapped) send a file request to the NAS head using the NFS or CIFS protocol. The NAS head receives the request, converts it into a block request, and forwards it to the storage controller. Upon receiving the block data from the storage controller, the NAS head again converts the block request back to the file request and sends it to the clients.
- ***Object I/O request:*** The web application servers send an object request, typically using REST or SOAP protocols, to the OSD node. The OSD node receives the request, converts it into a block request, and sends it to the disk through the storage controller.

****

Module-3: Backup, Archive, and Replication

Syllabus: Backup, Archive and Replication Introduction to Business Continuity: Information Availability, BC Terminology, BC Planning Lifecycle, Failure Analysis, BC Technology Solutions. Backup and Archive: Backup Methods, Backup Topologies, Backup Targets, Data Deduplication for Backup, Backup in Virtualized Environments, Data Archive. Local Replication: Replication Terminology, Uses of Local Replicas, Local Replication Technologies, Local Replication in a Virtualized Environment. Remote Replication: Remote Replication Technologies, Three-Site Replication, Remote Replication and Migration in a Virtualized Environment.

Chapter 9: Introduction to Business Continuity

Introduction

- Continuous access to information is a must for the smooth functioning of business operations today.
- There are many threats to information availability, such as natural disasters (e.g., flood, fire, earthquake), unplanned occurrences (e.g., cybercrime, human error, network and computer failure), and planned occurrences (e.g., upgrades, backup, restore) that result in the inaccessibility of information.
- ***Business continuity (BC)*** is an integrated and enterprise-wide process that includes all activities (internal and external to IT) that a business must perform to mitigate the impact of planned and unplanned downtime.
- BC entails preparing for, responding to, and recovering from a system outage that adversely affects business operations. It involves proactive measures, such as business impact analysis and risk assessments, data protection, and security, and reactive countermeasures, such as disaster recovery and restart, to be invoked in the event of a failure.
- The **goal of a business continuity solution is to ensure the "information availability"** required to conduct vital business operations.

9.1 Information Availability

Information availability (IA) refers to the ability of the infrastructure to function according to

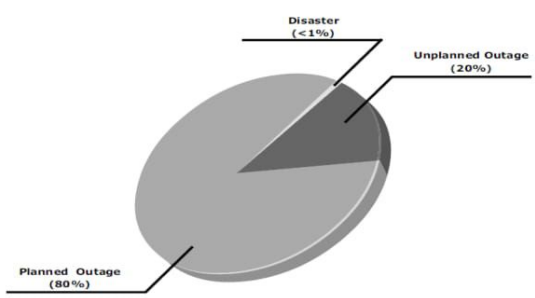
business expectations during its specified time of operation and ensure that people (Employees, customers, partners) can access information whenever they need it. I.A. can be defined with the help of following terms:

- **Accessibility:** This ensures that the required information is accessible at the right place, to the right user.
- **Reliability:** Information should be reliable and correct in all aspects. It is "the same" as what was stored, and there is no alteration or corruption to the information.
- **Timeliness:** Defines the exact moment or the time window (a particular time of the day) during which information must be accessible. For example, if online access to an application is required between 8:00 am and 10:00 pm each day, any disruptions to data availability outside of this time slot are not considered to affect timeliness.

9.1.1 Causes of Information Unavailability

Various planned and unplanned incidents result in data unavailability.

1. ***Planned outages*** include installation/ integration/ maintenance of new hardware, software upgrades or patches, taking backups, application and data restores, facility operations (renovation and construction), and refresh/migration of the testing to the production environment.
2. ***Unplanned outages*** include failure caused by database corruption, component failure, and human errors.
   - Another type of incident that may cause data unavailability is natural or man-made disasters such as flood, fire, earthquake, and contamination.
   - As illustrated in Figure 9-1, the majority of outages are planned. Planned outages are expected and scheduled, but still cause data to be unavailable.
   - Statistically, less than 1 percent is likely to be the result of an unforeseen disaster.



9.1.2 Consequences of Downtime

- Information unavailability or downtime results in loss of productivity, loss of revenue, poor financial performance, and damage to reputation.
- Loss of productivity includes reduced output per unit of labor, equipment, and capital.
- Loss of revenue includes direct loss, compensatory payments, future revenue loss, billing loss, & investment loss. Poor financial performance affects revenue recognition, cash flow, discounts, credit rating, and stock price.
- Damages to reputations may result in a loss of confidence or credibility with customers, suppliers, financial markets, banks, and business partners.

The business impact of downtime is the sum of all losses sustained as a result of a given disruption. An important metric, ***average cost of downtime per hour***, provides a key estimate in determining the appropriate BC solutions.

It is calculated as follows:

Average cost of downtime per hour = Avg productivity loss / hour + Avg revenue loss per hour

Where:

Productivity loss per hour = (total salaries and benefits of all employees per week)/ (average number of working hours per week)

Average revenue loss per hour = (total revenue of an organization per week)/ (average number of hours per week that an organization is open for business)

11.1.2 Measuring Information Availability

- Information availability relies on the availability of the hardware and software components of a data center. Failure of these components might disrupt information availability.
- A failure is the termination of a component's ability to perform a required function. The component's ability can be restored by performing an external corrective action, such as a manual reboot, a repair, or replacement of the failed component(s).
- Repair involves restoring a component to a condition that enables it to perform a required function within a specified time by using procedures and resources.
- Proactive risk analysis performed as part of the BC planning process considers the component failure rate and average repair time, which are measured by MTBF and MTTR:

1. **Mean Time Between Failure (MTBF):** It is the average time available for a system or component to perform its normal operations between failures. It is the measure of system or component reliability and is usually expressed in hours.
2. **Mean Time to Repair (MTTR):** It is the average time required to repair a failed component. While calculating MTTR, it is assumed that the fault responsible for the failure is correctly identified and that the required spares and personnel are available.

   Fault is a physical defect at the component level, which may result in data unavailability. MTTR includes the time required to do the following: detect the fault, mobilize the maintenance team, diagnose the fault, obtain the spare parts, repair, test, and resume normal operations. Figure 9-2 illustrates the various information availability metrics that represent system uptime and downtime.

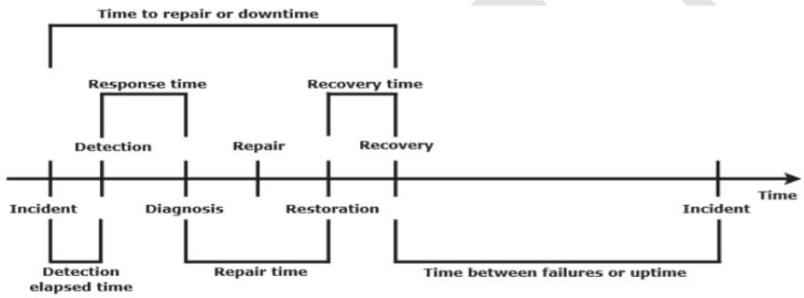


Figure 9-2: Information availability metrics

IA is the ***fraction of a time period*** that a system is in a condition to perform its intended function upon demand. It can be expressed in terms of system uptime and downtime and measured as the amount or percentage of system uptime:

$$IA = \text{system uptime} / (\text{system uptime} + \text{system downtime})$$

In terms of MTBF and MTTR, IA could also be expressed as

$$IA = MTBF / (MTBF + MTTR)$$

Uptime per year is based on the exact timeliness requirements of the service, this calculation leads to the number of "9s" representation for availability metrics.

Table 11-1 lists the approximate amount of downtime allowed for a service to achieve certain levels of 9s availability.

Table 11-1: Availability Percentage and Allowable Downtime

UPTIME (%)	DOWNTIME (%)	DOWNTIME PER YEAR	DOWNTIME PER WEEK
98	2	7.3 days	3 hr 22 minutes
99	1	3.65 days	1 hr 41 minutes
99.8	0.2	17 hr 31 minutes	20 minutes 10 sec
99.9	0.1	8 hr 45 minutes	10 minutes 5 sec
99.99	0.01	52.5 minutes	1 minute
99.999	0.001	5.25 minutes	6 sec
99.9999	0.0001	31.5 sec	0.6 sec

9.2 BC Terminology

This section introduces and defines common terms related to BC operations

1. **Disaster recovery:** This is the coordinated process of restoring systems, data, and the infrastructure required to support ongoing business operations after the disaster occurs.
    - It is the process of restoring a previous copy of the data and applying logs or other necessary processes to that copy to bring it to a known point of consistency.
    - Once all recoveries are completed, the data is validated to ensure that it is correct.
2. **Disaster restart:** This is the process of restarting business operations with mirrored consistent copies of data and applications.
3. **Recovery-Point Objective (RPO):** This is the point in time to which systems and data must be recovered after an outage. It defines the amount of data loss that a business can endure.
    - A large RPO signifies high tolerance to information loss in a business.
    - Based on the RPO, organizations plan for the minimum frequency with which a backup or replica must be made. Example: if the RPO is six hours, backups or replicas must be made at least once in 6 hours.
    - Figure 11-2 shows various RPOs and their corresponding ideal recovery strategies.
    - An organization can plan for an appropriate BC technology solution on the basis of the RPO it sets. For example:

a. **RPO of 24 hours:** This ensures that backups are created on an offsite tape drive every midnight. The corresponding recovery strategy is to restore data from the set of last backup tapes.

b. **RPO of 1 hour:** This ship database logs to the remote site every hour. The corresponding recovery strategy is to recover the database at the point of the last log shipment.

c. **RPO in the order of minutes:** Mirroring data asynchronously to a remote site.

d. **Near zero RPO:** This mirrors data synchronously to a remote site.

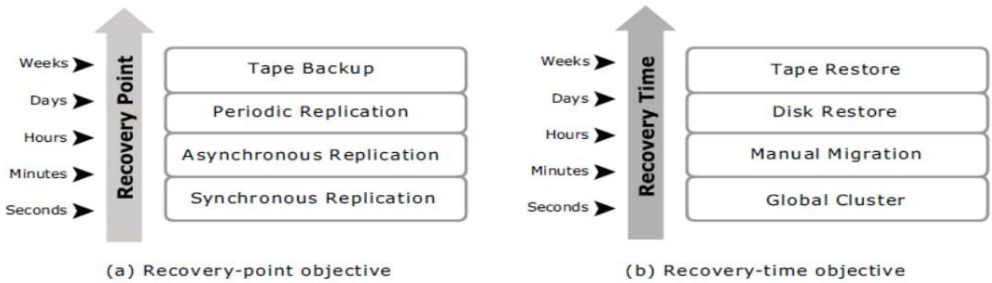


Figure 11-2: Strategies to meet RPO and RTO targets

4. **Recovery-Time Objective (RTO):** The time within which systems, applications, or functions must be recovered after an outage.

   It defines the amount of downtime that a business can endure and survive.

   Businesses can optimize disaster recovery plans after defining the RTO for a given data center or network. For example, if the RTO is two hours, then use a disk backup because it enables a faster restore than a tape backup.

   However, for an RTO of one week, tape backup will likely meet requirements. Some examples of RTOs and the recovery strategies to ensure data availability are listed below (Figure 11-2):

a. **RTO of 72 hours:** Restore from backup tapes at a cold site.

b. **RTO of 12 hours:** Restore from tapes at a hot site.

c. **RTO of 4 hours:** Use a data vault to a hot site.

d. **RTO of 1 hour:** Cluster production servers with controller-based disk mirroring.

e. **RTO of a few seconds:** Cluster production servers with bidirectional mirroring, enabling the applications to run at both sites simultaneously.

Data vault: A repository at a remote site where data can be periodically or continuously copied (either to tape drives or disks) so that there is always a copy at another site

Hot site: A site where an enterprise's operations can be moved in the event of disaster. It is a site with the required hardware, operating system, application, and network support to perform business operations, where the equipment is available and running at all times.

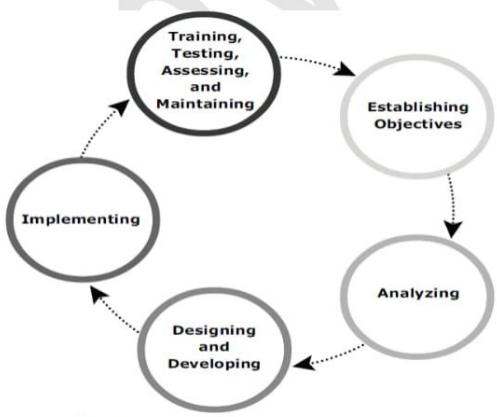
Cold site: A site where an enterprise's operations can be moved in the event of disaster, with minimum IT infrastructure and environmental facilities in place, but not activated

Server Clustering: A group of servers and other necessary resources coupled to operate as a single system. Clusters can ensure high availability and load balancing. Typically, in failover clusters, one server runs an application and updates the data, and another server is kept as standby to take over completely, as required. Server clustering provides load balancing by distributing the application load evenly among multiple servers within the cluster.

11.3 BC Planning Lifecycle

BC planning must follow a disciplined approach like any other planning process. From the conceptualization to the realization of the BC plan, a lifecycle of activities can be defined for the BC process. The BC planning lifecycle includes five stages (see Figure 11-3):

1. Establishing objectives
2. Analyzing
3. Designing and developing
4. Implementing
5. Training, testing, assessing, and maintaining



e 11-3: BC planning lifecycle

Several activities are performed at each stage of the BC planning lifecycle, including the following key activities:

1. ***Establishing objectives***

- Determine BC requirements.
- Estimate the scope and budget to achieve requirements.
- Select a BC team by considering subject matter experts from all areas of the business,

whether internal or external.

- Create BC policies.

2. ***Analyzing***

- Collect information on data profiles, business processes, infrastructure support, dependencies, and frequency of using business infrastructure.
- Identify critical business needs and assign recovery priorities.
- Create a risk analysis for critical functions and mitigation strategies.
- Conduct a Business Impact Analysis (BIA).
- Create a cost and benefit analysis based on the consequences of data unavailability.
- Evaluate options.

3. ***Designing and developing***

- Define the team structure and assign individual roles and responsibilities. For example, different teams are formed for activities such as emergency response, damage assessment, and infrastructure and application recovery.
- Design data protection strategies and develop infrastructure.
- Develop contingency scenarios.
- Develop emergency response procedures.
- Detail recovery and restart procedures.

4. ***Implementing***

- Implement risk management and mitigation procedures that include backup, replication, and management of resources.
- Prepare the disaster recovery sites that can be utilized if a disaster affects the primary data center.
- Implement redundancy for every resource in a data center to avoid single points of failure.

5. ***Training, testing, assessing, and maintaining***

- Train the employees who are responsible for backup and replication of business-critical data on a regular basis or whenever there is a modification in the BC plan.
- Train employees on emergency response procedures when disasters are declared.
- Train the recovery team on recovery procedures based on contingency scenarios.

- Perform damage assessment processes and review recovery plans.
- Test the BC plan regularly to evaluate its performance and identify its limitations.
- Assess the performance reports and identify limitations.
- Update the BC plans and recovery/restart procedures to reflect regular changes within the data center.

9.3 Failure Analysis

Failure analysis involves analyzing the data center to identify systems that are susceptible to a single point of failure and implementing fault-tolerance mechanisms such as redundancy.

9.4.1 Single Point of Failure

- A single point of failure refers to the failure of a component that can terminate the availability of the entire system or IT service.

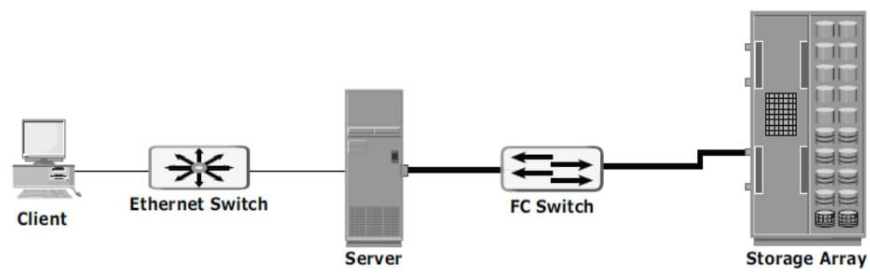


Figure 11-4: Single point of failure

- Figure 11-4 illustrates the possibility of a single point of failure in a system with various components: server, network, switch, and storage array.
- The figure depicts a system setup in which an application running on the server provides an interface to the client and performs I/O operations.
- The client is connected to the server through an IP network, the server is connected to the storage array through a FC connection, an HBA installed at the server sends or receives data to and from a storage array, and an FC switch connects the HBA to the storage port.
- In a setup where each component must function as required to ensure data availability.
- The failure of single component causes the failure of entire data center/application resulting

in disruption of business operations.

- In this example, several single points of failure can be identified. i.e. HBA on the server, the server itself, the IP network, FC switch, the storage array ports, or even storage array can become potential single point of failure.
- To avoid this situation, it is essential to implement fault tolerance mechanism.

9.4.2 Resolving Single Point of Failure: Fault Tolerance

- To overcome single point of failure, systems are designed with redundancy such that system will fail only if all the components in the redundancy group fail.
- This ensures that the failure of a single component does not affect data availability. Figure 11-5 illustrates the fault-tolerant implementation of the system just described (and shown in Figure 11-4).

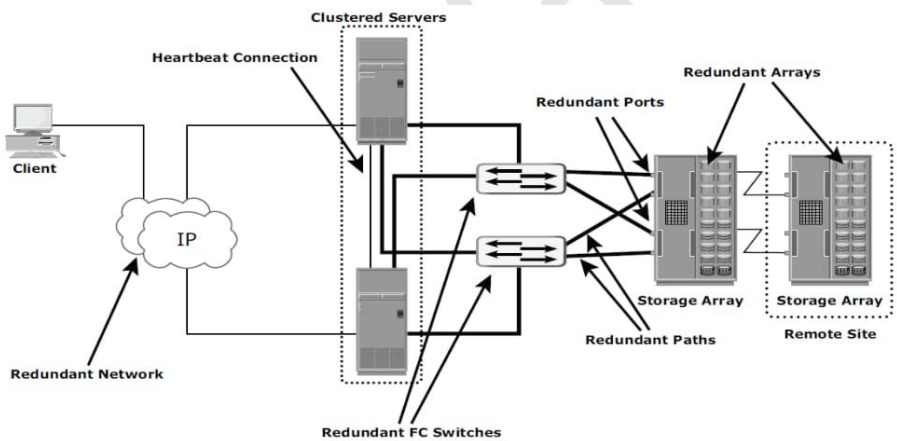


Figure 11-5: Implementation of fault tolerance

- Careful analysis is performed to eliminate every single point of failure, in the example figure below, all enhancements are done in the infrastructures to overcome single points of failure such as

  1. Configuration of multiple HBAs to mitigate single HBA failure.

  2. Configuration of multiple fabrics to account for a switch failure.

  3. Configuration of multiple storage array ports to enhance the storage array's availability.

4. RAID configuration to ensure continuous operation in the event of disk failure.

5. Implementing a storage array at a remote site to mitigate local site failure.

6. Implementing server (host) clustering, a fault-tolerance mechanism whereby two or more servers in a cluster access the same set of volumes. Clustered servers exchange *heartbeats* to inform each other about their health. If one of the servers fails, the other server takes up the complete workload.

9.4.3 Multipathing Software

- Configuring multiple paths increases the data availability through path failover.
- . If servers are configured with one I/O path to the data there will be no access to the data if that path fails. Redundant paths eliminate the path to become single points of failure.
- Multiple paths to data also improve I/O performance through load sharing and maximize server, storage, and data path utilization.

9.5 BC Technology Solutions

After analyzing the business impact of an outage, designing appropriate solutions to recover from a failure is the next important activity. Using one of the strategies data can be recovered and business operations can be restarted using an alternate copy:

o Backup/Restore

Backup to tape has been a predominant method to ensure business continuity

Frequency of backup is depending on RPO/RTO requirements

o Local Replication

Data from the production devices is copied to replica devices within the same array

The replicas can then be used for restore operations in the event of data corruption or other events

o Remote Replication

Data from the production devices is copied to replica devices on a remote array

In the event of a failure, applications can continue to run from the target device

Chapter 10

Backup and Recovery

- *A backup* is a copy of production data, created and retained for the sole purpose of recovering deleted or corrupted data.
- Data archiving is the process of moving data that is no longer actively used, from primary storage to a low-cost secondary storage.
- Backup Purpose
  1. Disaster Recovery
  2. Operational Recovery
  3. Archival

10.1 Backup Methods

There are two methods; cold backup and hot backup. They are based on the state of the application when the backup is performed.

1. **Hot backup**

- In a hot backup, the application is up-and-running with users accessing their data during the backup process. This method is also referred as an **online backup.**
- Hot backup of online data is challenging because data is actively used and changed. If a file is open, it is normally not backed up during the backup process. In such situations, an open file agent is required to back up the open file.
- These agents interact directly with the operating system or application and enable the creation of consistent copies of open files.
- In database environments, the use of open file agents is not enough, because the agent should also support a consistent backup of all the database components.

- Consistent backups of databases can also be done by using a cold backup. This requires the database to remain inactive during the backup.
- Hot backup is used in situations where it is not possible to shut down the database. This is facilitated by ***database backup agents*** that can perform a backup while the database is active.

Disadvantages of a hot backup

- The agents usually affect the overall application performance.

2. Cold backup

- In a cold backup, the application is shut down during the backup process. Hence this method is referred as offline backup.
- Consistent data backup can be done using cold backup due to it requires the database to remain inactive during the backup.
- The **disadvantage** of a cold backup is that the database is inaccessible to users during the backup process.

❖ A ***point-in-time (PIT)*** copy method is deployed in environments where the impact of downtime from a cold backup or the performance resulting from a hot backup is unacceptable. A pointer-based PIT copy consumes only a fraction of the storage space and can be created very quickly. The PIT copy is created from the production volume and used as the source for the backup. This reduces the impact on the production volume.

❖ In a disaster recovery environment, ***bare-metal recovery (BMR)*** refers to a backup in which all metadata, system information, and application configurations are appropriately backed up for a full system recovery.

Backup Architecture and Process

- **Backup client**

  Sends backup data to backup server or storage node

- **Backup server**

  Manages backup operations and maintains backup catalog

- **Storage node**

  Responsible for writing data to backup device

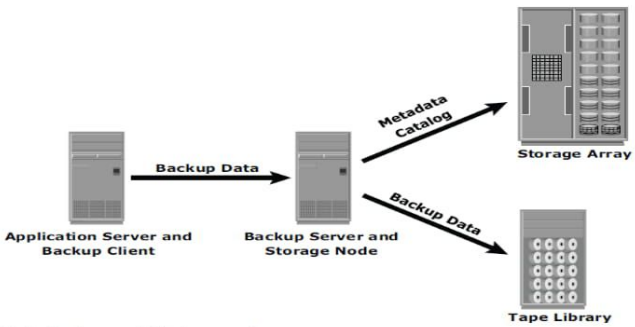


Figure 12-4: Backup architecture and process

Backup and Restore Operations

When a backup process is initiated, significant network communication takes place between the different components of a backup infrastructure. The backup server initiates the backup process for different clients based on the backup schedule configured for them. For example, the backup process for a group of clients may be scheduled to start at 3:00 am every day.

The backup server coordinates the backup process with all the components in a backup configuration (see Figure 12-5). The backup server maintains the information about backup clients to be contacted and storage nodes to be used in a backup operation. The backup server retrieves the backup-related information from the backup catalog and, based on this information, instructs the storage node to load the appropriate backup media into the backup devices. Simultaneously, it instructs the backup clients to start scanning the data, package it, and send it over the network to the assigned storage node. The storage node, in turn, sends metadata to the backup server to keep it updated about the media being used in the backup process. The backup server continuously updates the backup catalog with this information.

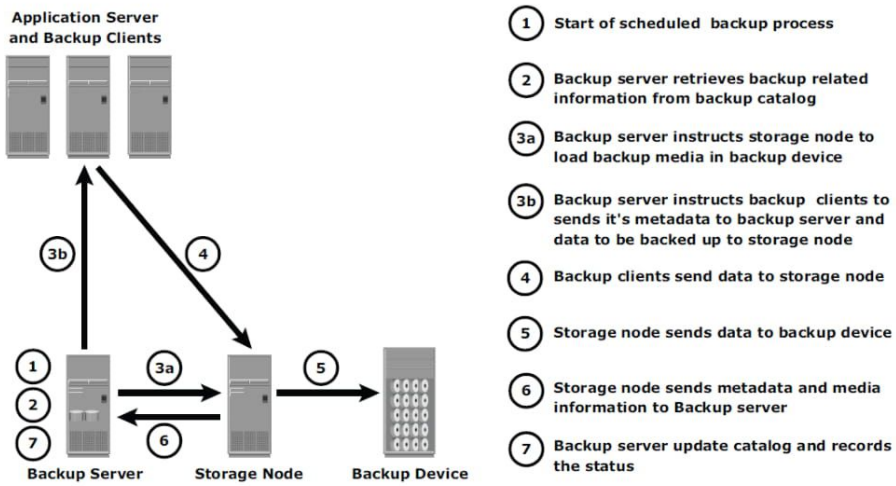


Figure 12-5: Backup operation

After the data is backed up, it can be restored when required. A restore process must be manually initiated. Some backup software has a separate application for restore operations. These restore applications are accessible only to the administrators.

Figure 12-6 depicts a restore process.

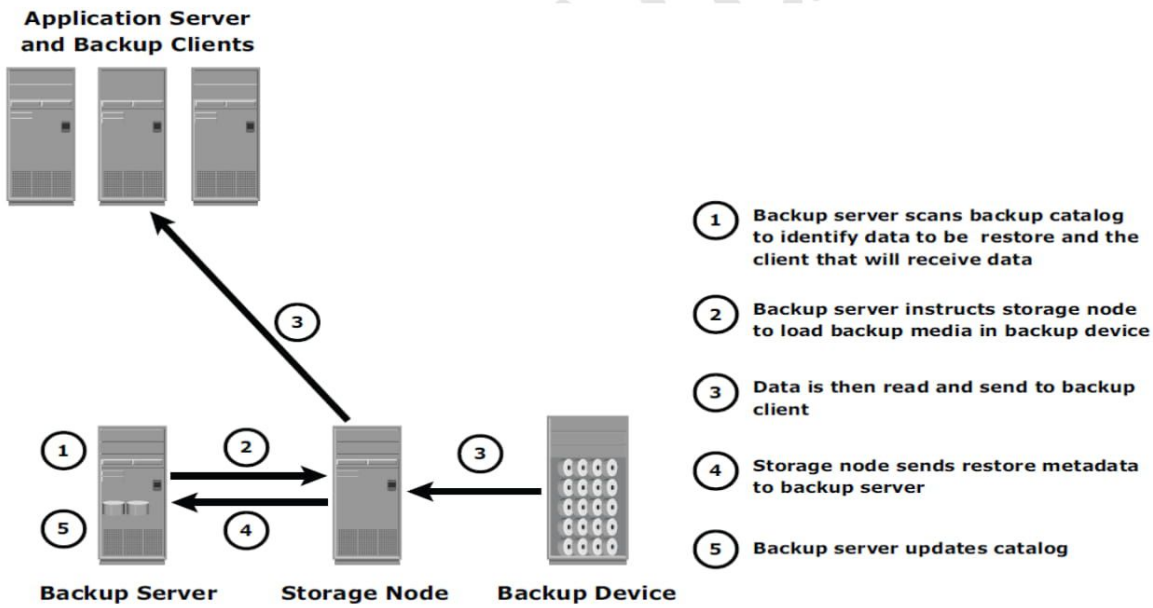


Figure 12-6: Restore operation

Upon receiving a restore request, an administrator opens the restore application to view the list of clients that have been backed up. While selecting the client for which a restore request has been made, the administrator also needs to identify the client that will receive the restored data. Data can be restored on the same client for whom the restore request has been made or on any other client.

The administrator then selects the data to be restored and the specified point in time to which the data has to be restored based on the RPO. Note that because all of this information comes from the backup catalog, the restore application must also communicate to the backup server.
The administrator first selects the data to be restored and initiates the restore process. The backup server, using the appropriate storage node, then identifies the backup media that needs to be mounted on the backup devices. Data is then read and sent to the client that has been identified to receive the restored data.

10.2 Backup Topologies

Three basic topologies are used in a backup environment:

1. Direct attached backup,
2. LAN based backup,
3. SAN based backup.
4. A mixed topology is also used by combining LAN based and SAN based topologies.

1. Direct attached backup

- In a direct-attached backup, the storage node is configured on a backup client, and the backup device is attached directly to the client.
- Only the metadata is sent to the backup server through the LAN.
- This configuration frees the LAN from backup traffic.

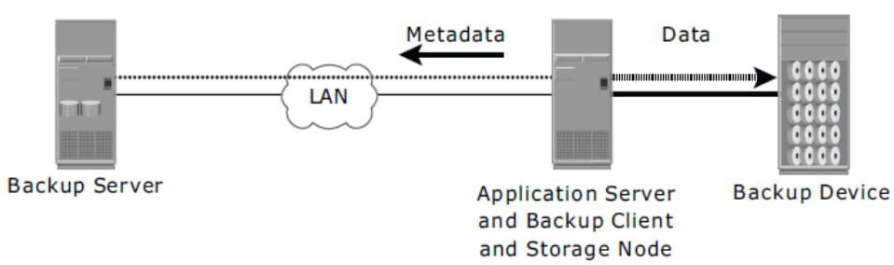


Figure 12-7: Direct-attached backup topology

- Figure 12-7 depicts use of a backup device that is not shared, the backup device is directly attached and dedicated to the backup client.
- As the environment grows, there will be a need for central management of all backup

devices and sharing of backup devices to optimize costs.

- **Disadvantage:** Not possible to share the backup devices among multiple servers.
- Network-based topologies (LAN-based and SAN-based) provide the solution to optimize the utilization of backup devices.

2. LAN Based Backup

- In a LAN-based backup, the clients, backup server, storage node, and backup device are connected to the LAN.
- The data to be backed up is transferred from the backup client (source) to the backup device (destination) over the LAN, which might affect network performance.

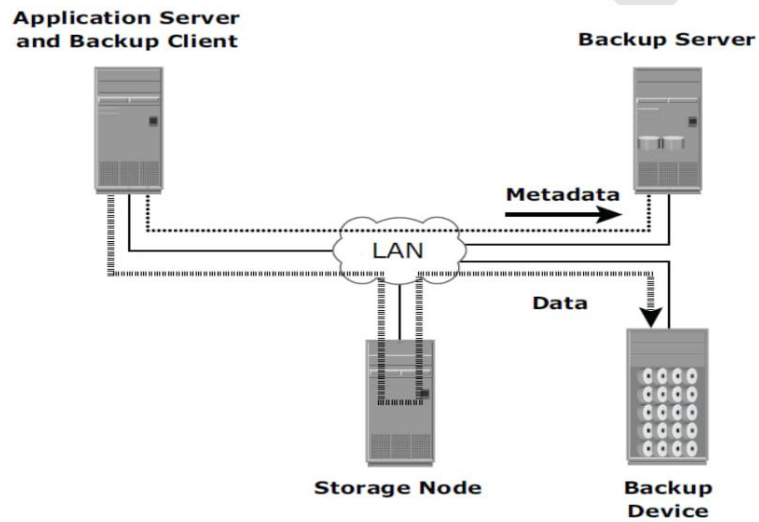


Figure 12-8: LAN-based backup topology

- This impact can be minimized by adopting a number of measures, such as configuring separate networks for backup and installing dedicated storage nodes for some application servers.

3. SAN Based Backup

- A SAN-based backup is also known as a **LAN-free backup.**
- This topology is most appropriate solution when a backup device needs to be shared among the clients. In this case the backup device and clients are attached to the SAN.

- A client sends the data to be backed up to the backup device over the SAN, only the backup metadata is transported over the LAN.
- Figure 10-9 illustrates a SAN-based backup.

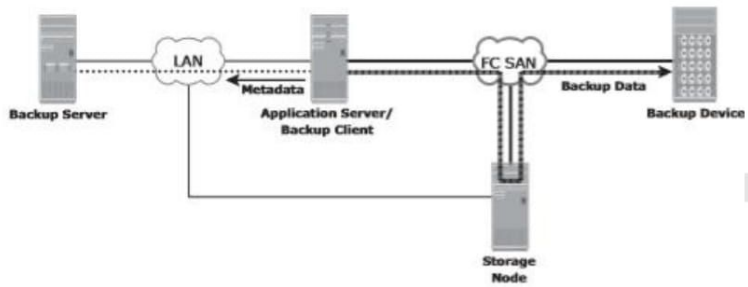
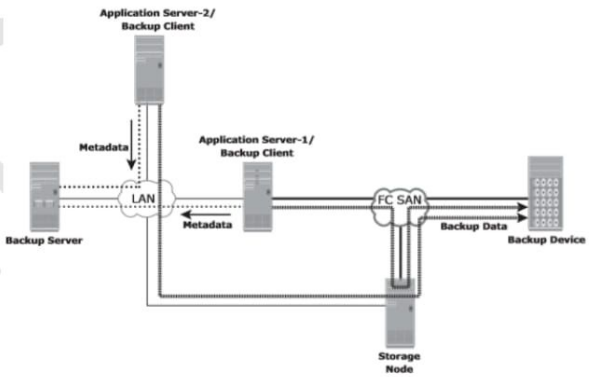


Figure 10-9: SAN-based backup topology

- In this example, a client sends the data to be backed up to the backup device over the SAN. Therefore, the backup data traffic is restricted to the SAN, and only the backup metadata is transported over the LAN. The volume of metadata is insignificant when compared to the production data; the LAN performance is not degraded in this configuration.

4. Mixed topology

- The mixed topology uses both the LAN-based and SAN-based topologies, as shown in Figure



- This topology might be implemented for several reasons, including cost, server location, reduction in administrative overhead, and performance considerations.

10.3 Backup Targets

A wide range of technology solutions are currently available for backup targets. Tape and disk libraries are the two most commonly used backup targets.

10.3.1 Backup to Tape

- ✓ Tapes, a low-cost technology, are used extensively for backup.
- ✓ Tape drives are used to read/write data from/to a tape cartridge. Tape drives are referred to as sequential, or linear, access devices because the data is written or read sequentially.
- ✓ Tape mounting is the process of inserting a tape cartridge into a tape drive. The tape drive has motorized controls to move the magnetic tape around, enabling the head to read or write data.

10.3.1 Physical Tape Library

- o The physical tape library provides housing and power for a large number of tape drives and tape cartridges, along with a robotic arm or picker mechanism.
- o The backup software has intelligence to manage the robotic arm and entire backup process.

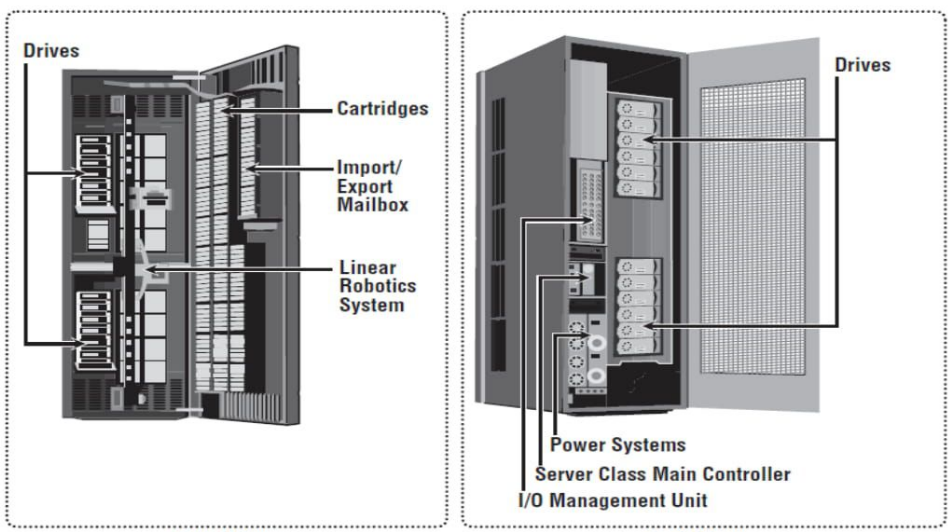


Figure 12-15: Physical tape library

- o Tape drives read and write data from and to a tape. Tape cartridges are placed in the slots when not in use by a tape drive. Robotic arms are used to move tapes between

cartridge slots and tape drives. Mail or import/export slots are used to add or remove tapes from the library without opening the access doors.

- When a backup process starts, the robotic arm is instructed to load a tape to a tape drive. This process adds the delay.
- The time taken to position the heads and validate header information is called load or ready time.
- The tape receives the backup data and stores in its internal buffer as blocks. The speed of the tape drives can also be adjusted to match the data transfer rate.
- To improve performance tape drive using multiple streaming; writes data from multiple streams on a single tape to keep the drive busy.
- Tape drive *streaming* or *multiple streaming* writes data from multiple streams on a single tape to keep the drive busy. Shown in Figure 12-16, multiple streaming improves media performance, but it has an associated disadvantage. The backup data is interleaved because data from multiple streams is written on it. Consequently, the data recovery time is increased.

Figure 12-16: Multiple streams on tape media

Limitations of Tape

- Tapes are primarily used for long-term offsite storage because of their low cost.
- Tapes must be stored in locations with a controlled environment to ensure preservation of the media and prevent data corruption.
- Data access in a tape is sequential, which can slow backup and recovery operations. Physical transportation of the tapes to offsite locations also adds management overhead.

10.3.3 Backup to Disk

Advantage over Backup to tape

- ❖ Disks have now replaced tapes as the primary device for storing backup data because of their performance advantages. Backup-to-disk systems offer ease of implementation, reduced cost, and improved quality of service. Apart from performance benefits in terms of data transfer rates, disks also offer faster recovery when compared to tapes.

- o Backing up to disk storage systems offers clear advantages due to their inherent random access and RAID-protection capabilities.

- o Backup to disk copies the data temporarily before transferring or staging it to tapes, this enhances the performance.

- o Some backup products allow for backup images to remain on the disk for a period of time even after they have been staged. This enables a much faster restore.

- o Recovering from a full backup copy stored on disk and kept onsite provides the fastest recovery solution.

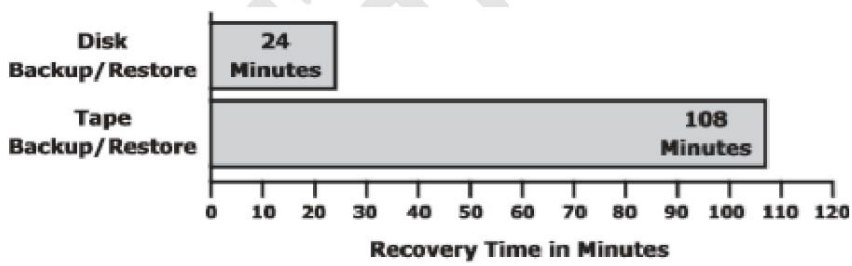


Figure 10-17: Tape versus disk restore

- o The above figure shows the comparison between disks that supports 800 users with a 75 MB mailbox and a 60GB database. It is observed that disk took 24 mins compared to the restore from tape, which took 108 mins for the same environment.

3. Backup to Virtual Tape

Virtual tapes are disk drives emulated and presented as tapes to the backup software. The key benefit of using a virtual tape is that it does not require any additional modules, configuration,

or changes in the legacy backup software. This preserves the investment made in the backup software.

Virtual Tape Library

- Components of Virtual Tape Library (VLT) are same as physical tape drive library, except that the majority of the components are presented as virtual resources.
- Backup software; there is no difference between physical tape library and a virtual tape library.
- Virtual tape libraries use disks as backup media.
- Emulation software has a database with a list of virtual tapes, and each virtual tape is assigned space on a LUN.
- Similar to physical tape library, a robot mount is virtually performed when a backup process starts in a virtual tape library; it does not involve any mechanical delays as in physical tape library. Even the load and ready time is much less than the physical tape library.
- After the virtual tape is mounted and the virtual tape drive is positioned, the virtual tape is ready to be used, and backup data can be written to it. In most cases the data is written immediately.
- Compared to physical tapes, virtual tapes offer better single stream performance, better reliability, and random disk access characteristics.
- Backup and restore operations are online and provide faster backup and recovery.
- The steps to restore data are similar to physical tape library but the restore operation is nearly instantaneous. Even though virtual tapes are based on disks, which provide random access, they still emulate the tape behavior.

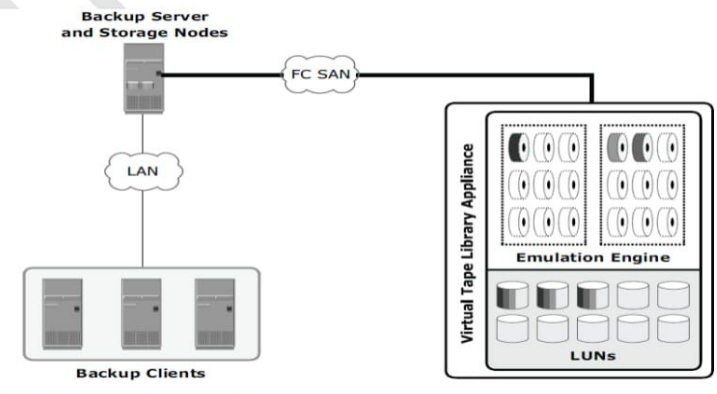


Figure 12-18: Virtual tape library

Advantages

- Using virtual tape offers several advantages over both physical tapes and disks. Compared to physical tape, virtual tape offers better single stream performance, better reliability, and random disk access characteristics.
- Backup and restore operations are sequential by nature, but they benefit from the disk's random-access characteristics because they are always online and ready to be used, improving backup and recovery times.
- Virtual tape does not require the usual maintenance tasks associated with a physical tape drive, such as periodic cleaning and drive calibration.
  - Compared to back up-to-disk devices, virtual tapes offer easy installation and administration and inherent offsite capabilities.

Data Deduplication for Backup

Data deduplication is the process of identifying and eliminating redundant data. Data deduplication helps to reduce the storage requirement for backup, shorten the backup window, and remove the network burden. It also helps to store more backups on the disk and retain the data on the disk for a longer time.

Data Deduplication Methods

There are two methods of deduplication: file level and subfile level.

File-level deduplication (also called ***single-instance storage***)

- It detects and removes redundant copies of identical files.
- It enables storing only one copy of the file; the subsequent copies are replaced with a pointer that points to the original file.
- File-level deduplication is simple and fast but does not address the problem of duplicate content inside the files.

Subfile deduplication

- It breaks the file into smaller chunks and then uses a specialized algorithm to detect

redundant data within and across the file. There are two forms of subfile deduplication:

- The **fixed-length block deduplication** divides the files into fixed length blocks and uses a hash algorithm to find the duplicate data.
- In **variable-length segment deduplication**, if there is a change in the segment, the boundary for only that segment is adjusted, leaving the remaining segments unchanged. This method vastly improves the ability to find duplicate data segments compared to fixed-block.

Data Deduplication Implementation

1. Source based data deduplication

- Source-based data deduplication eliminates redundant data at the source before it transmits to the backup device.
- Source-based data deduplication can dramatically reduce the amount of backup data sent over the network during backup processes.
- It provides the benefits of a shorter backup window and requires less network bandwidth.
- There is also a substantial reduction in the capacity required to store the backup images.
- Source-based deduplication increases the overhead on the backup client, which impacts the performance of the backup and application running on the client.

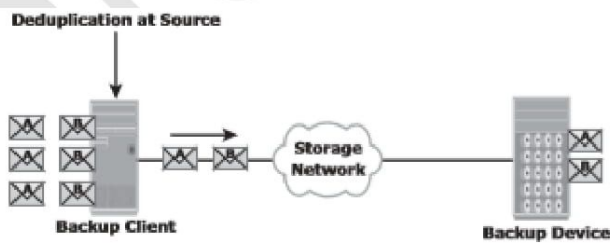


Figure 10-19: Source-based data deduplication

2. Target based data deduplication

- Target-based data deduplication occurs at the backup device, which offloads the backup client from the deduplication process.
- In this case, the backup client sends the data to the backup device and the data is

deduplicated at the backup device, either immediately (inline) or at a scheduled time (post-process).

- Backup data needs to be transferred over the network, which increases network bandwidth requirements.

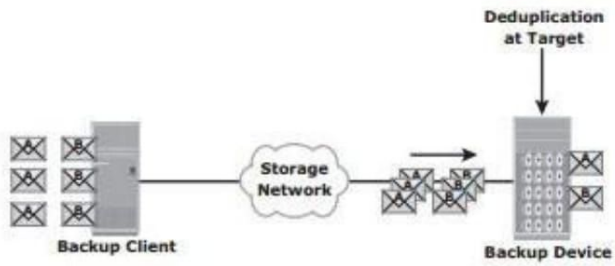


Figure 10-20: Target-based data deduplication

3. Inline deduplication

- It performs deduplication on the backup data before it is stored on the backup device. Hence, this method reduces the storage capacity needed for the backup.
- Inline deduplication introduces overhead in the form of the time required to identify and remove duplication in the data. So, this method is best suited for an environment with a large backup window.

4. Post-process deduplication

- It enables the backup data to be stored or written on the backup device first and then deduplicated later.
- This method is suitable for situations with tighter backup windows.
- Post-process deduplication requires more storage capacity to store the backup images before they are deduplicated.

Backup in Virtualized Environment

In a virtualized environment, it is imperative to back up the virtual machine data (OS, application data, and configuration) to prevent its loss or corruption due to human or technical errors. There are two approaches for performing a backup in a virtualized environment: the traditional backup

approach and the image-based backup approach.

1. Traditional Backup approach

- A backup agent is installed either on the virtual machine (VM) or on the hypervisor.
- If the backup agent is installed on a VM, the VM appears as a physical server to the agent.

Figure 10-21: Traditional VM backup

- The backup agent installed on the VM backs up the VM data to the backup device. The agent does not capture VM files, such as the virtual BIOS file, VM swap file, logs, and configuration files. Therefore, for a VM restore, a user needs to manually re-create the VM and then restore data onto it.
- VM files are backed up by performing a file system backup from a hypervisor. This approach is relatively simple because it requires having the agent just on the hypervisor instead of all the VMs.
- The backup should be performed when the server resources are idle or during a low activity period on the network.

2. Image-based backup approach

- It operates at the hypervisor level and takes the snapshot of the VM. It creates a copy of the guest OS and all the associated with, including the VM state and application configurations.
- The backup is saved as a single file called an 'image' and this image is mounted on the separate physical machine-proxy server, which acts as a backup client.
- Image based backup enables quick restoration of a VM.

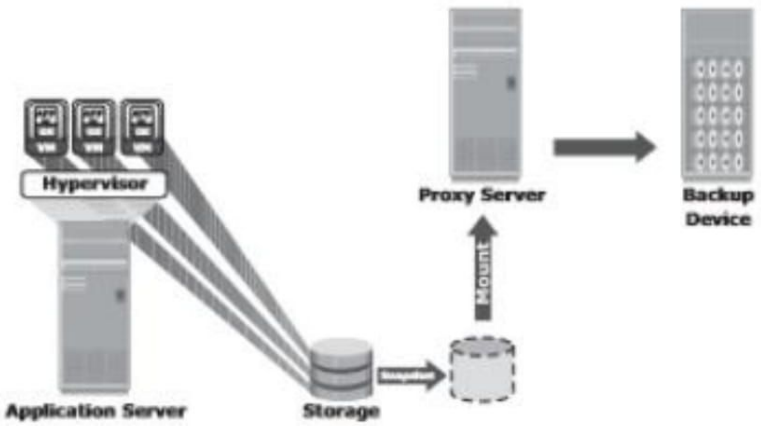


Figure 10-22: Image-based backup

Data Archive

In the life cycle of information, data is actively created, accessed, and changed. As data ages, it is less likely to be changed and eventually becomes "fixed" but continues to be accessed by applications and users. This data is called fixed content. X-rays, e-mails, and multimedia files are examples of fixed content. A repository where fixed content is stored is known as an archive.

An archive can be implemented as an online, nearline, or offline solution:

- **Online archive:** A storage device directly connected to a host that makes the data immediately accessible.
- **Nearline archive:** A storage device connected to a host, but the device where the data is stored must be mounted or loaded to access the data.
- **Offline archive:** A storage device that is not ready to use. Manual intervention is required to connect, mount, or load the storage device before data can be accessed.

Chapter 11

Local Replication

Replication Terminology

The common terms used to represent various entities and operations in a replication environment are listed below:

- **Source:** A host accessing the production data from one or more LUNs on the storage array is called a production host, and these LUNs are known as source LUNs (devices/volumes), production LUNs, or simply the source.
- **Target:** A LUN (or LUNs) on which the production data is replicated, is called the target LUN or simply the target or replica.
- **Point-in-Time (PIT) and continuous replica:** Replicas can be either a PIT or a continuous copy.
    - The PIT replica is an identical image of the source at some specific timestamp.
    - The continuous replica is in-sync with the production data at all times.
- **Recoverability and restart ability:**
    - Recoverability enables restoration of data from the replicas to the source if data loss or corruption occurs.
    - Restart ability enables restarting business operations using the replicas.

Uses of Local Replicas

- **Alternative source for backup:. The** local replica contains an exact point-in-time (PIT) copy of the source data, and therefore can be used as a source to perform backup operations. This alleviates the backup I/O workload on the production volumes. Another benefit of using local replicas for backup is that it reduces the backup window to zero.
- **Fast Recovery:**
    - If data loss or data corruption occurs on the source, a local replica might be used to recover the lost or corrupted data.
    - If a complete failure of the source occurs, some replication solutions enable a replica to be used to restore data onto a different set of source devices, or production can be restarted on the replica.

- In either case, this method provides faster recovery and minimal RTO compared to traditional recovery from tape backups.

- **Decision-support activities, such as reporting or data warehousing:** Running the reports using the data on the replicas greatly reduces the I/O burden placed on the production device. The data-warehouse application may be populated by the data on the replica and thus avoid the impact on the production environment.
- **Testing Platform:** Local replicas are also used for testing new applications or upgrades.
- **Data migration:** Data migrations are performed for various reasons, such as migrating from a smaller capacity LUN to one of a larger capacity for newer versions of the application.

Local Replication Technologies

Host-based, storage array-based and network-based replications are the major technologies used for local replication.

Host-Based Local Replication

LVM-based replication and file system (FS) snapshot are two common methods of host-based local replication.

LVM-Based Replication

- In *LVM-based replication*, the logical volume manager is responsible for creating and controlling the host-level logical volumes.
- An LVM has three components: physical volumes (physical disk), volume groups, and logical volumes.
  - A *volume group* is created by grouping one or more physical volumes.
  - *Logical volumes* are created within a given volume group.
- Each *logical block* in a logical volume is mapped to two physical blocks on two different

physical volumes, as shown in figure

- An application write to a logical volume is written to the two physical volumes by the LVM device driver. This is also known as *LVM mirroring*.

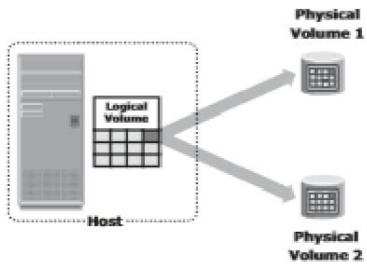


Figure 11-5: LVM-based mirroring

Advantages

- ✓ The LVM-based replication technology is not dependent on a vendor-specific storage system. No additional license is required to deploy LVM mirroring

Limitation

- × Every write generated by an application translates into two writes on the disk, and thus, an additional burden is placed on the host CPU. This can degrade application performance.
- × Tracking changes to the mirrors and performing incremental resynchronization operations is also a challenge because all LVMs do not support incremental resynchronization.

File system snapshot

A file system (FS) snapshot is a pointer-based replica, this snapshot can be implemented by either FS or by LVM. It uses the Copy on First Write (CoFW) principle to create snapshots.

- When a snapshot is created, a bitmap and block map is created in the metadata of the Snap FS. The bitmap is used to keep track of blocks that are changed on the production FS after the snap creation. The bitmap is used to keep track of blocks that are changed on the production FS after the snap creation.

- In a CoFW mechanism, if a write I/O is issued to the production FS for the first time after the creation of a snapshot, the I/O is held and the original data of production FS corresponding to that location is moved to the Snap FS. Then, the write is allowed to the production FS. The bitmap and blockmap are updated accordingly.
- To read from the Snap FS, the bitmap is consulted. If the bit is 0, then the read is directed to the production FS. If the bit is 1, then the block address is obtained from the block map, and the data is read from that address on the Snap FS.
- Figure below illustrates the write operations to the production file system.
- A write data "C" occurs on block 3 at the production FS, which currently holds data "c"' The snapshot application holds the I/O to the production FS and first copies the old data "c" to an available data block on the Snap FS.
- The bitmap and block map values for block 3 in the production FS are changed in the snap metadata.
- The bitmap of block 3 is changed to 1, indicating that this block has changed on the production FS. The blockmap of block 3 is changed and indicates the block number where the data is written in Snap FS, (in this case block 2).
- After this is done, the I/Os to the production FS are allowed to complete.
- Any subsequent writes to block 3 on the production FS occur as normal, and it does not initiate the CoFW operation.
- Similarly, if an I/O is issued to block 4 on the production FS to change the value of data "d" to "D," the snapshot application holds the I/O to the production FS and copies the old data to an available data block on the Snap FS.
- Then it changes the bitmap of block 4 to 1, indicating that the data block has changed on the production FS.

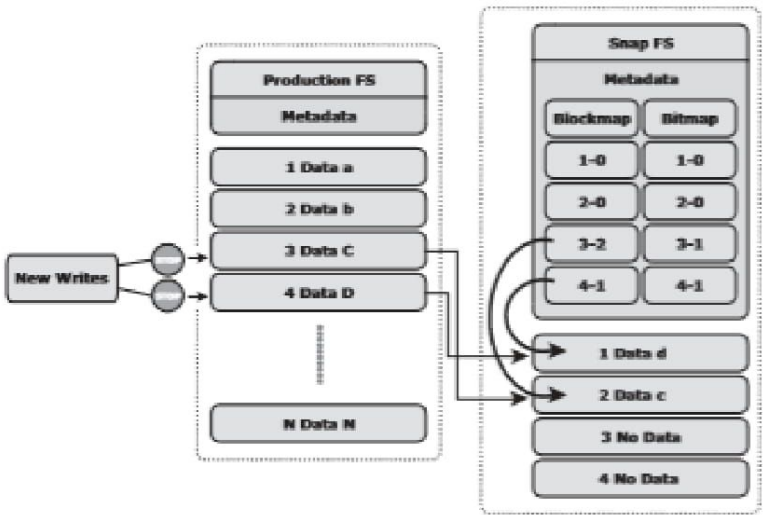


Figure 11-6: Write to production FS

- The blockmap for block 4 indicates the block number where the data can be found on the Snap FS, in this case, data block 1 of the Snap FS
- After this is done, the I/O to the production FS is allowed to complete.

Storage Array-Based Local Replication

• In this, the array-operating environment performs the local replication process. The host resources, such as the CPU and memory are not used in the replication process.

• In this replication process, the required number of replica devices should be selected on the same array and then data should be replicated between the source-replica pairs.

• Below figure shows a storage array-based local replication, where the source and target are in the same array and accessed by different hosts.

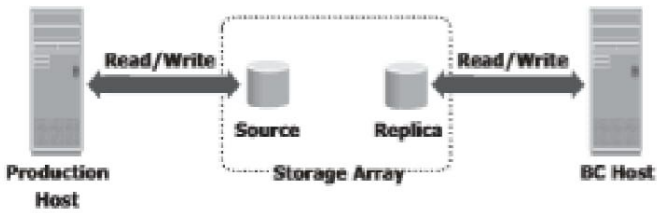
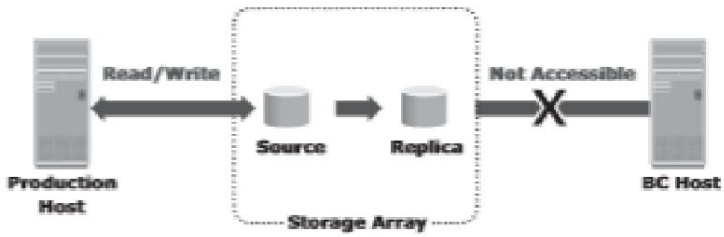


Figure 11-7: Storage array-based local replication

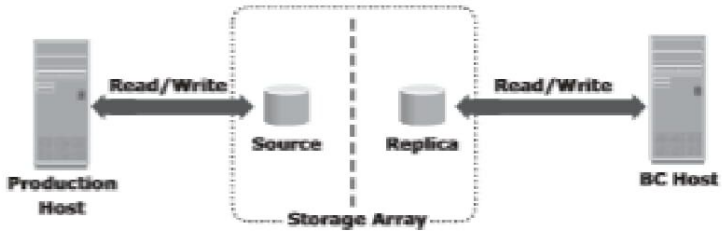
- Storage array-based local replication is implemented in three ways:
  - Full-Volume mirroring
  - Pointer-based full-volume replication
  - Pointer-based virtual replication

Full-Volume Mirroring

- In full-volume mirroring, the target is attached to the source and established as a mirror of the source
- Figure (a) the data is copied to the source to the target. New updates to the source are also updated to the target i.e., both source and target contains identical data, the target can be considered as a mirror of the source.
- Figure (b) shows full-volume mirroring when the target is detached from the source. Both the source and the target can be accessed for read and write operations by the production and business continuity hosts respectively.



(a) Full Volume Mirroring with Source Attached to Replica



(b) Full Volume Mirroring with Source Detached from Replica

Figure 11-8: Full-volume mirroring

Pointer-Based Full-Volume Replication

• Similar to full-volume, this technology can provide full copies of the source data on the targets.

• Unlike full-volume mirroring, the target is immediately accessible by the BC host after the replication session is activated.

• Pointer-based, full-volume replication can be activated in either

1. Copy on First Access (CoFA) mode
2. Full Copy mode.

• In either case, at the time of activation, a protection bitmap is created for all data on the source devices. The protection bitmap keeps track of the changes at the source device.

• The pointers on the target are initialized to map the corresponding data blocks on the source.

• After replication, the data is copied from source to target only when the following condition occurs;

- A write I/O is issued to a specific address on the source for the first time.
- A read or write I/O is issued to a specific address on the target for the first time.

• When a write is issued to the source for the first time after replication session activation, the original data at that address is copied to the target. After this operation, the new data is updated on the source. This ensures that the original data at the point-in-time of activation is preserved on the target (see Figure 11-9).

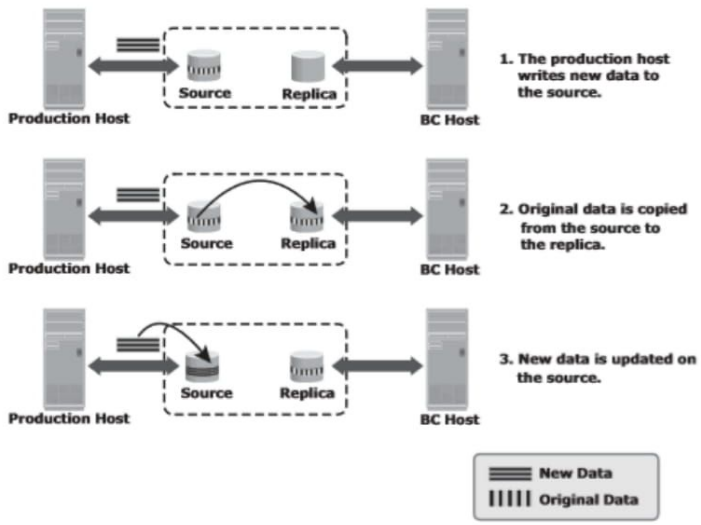


Figure 11-9: Copy on first access (CoFA) – write to source

• When a read is issued to the target for the first time after replication session activation, the original data is copied from the source to the target and is made available to the BC host (see Figure 11-10).

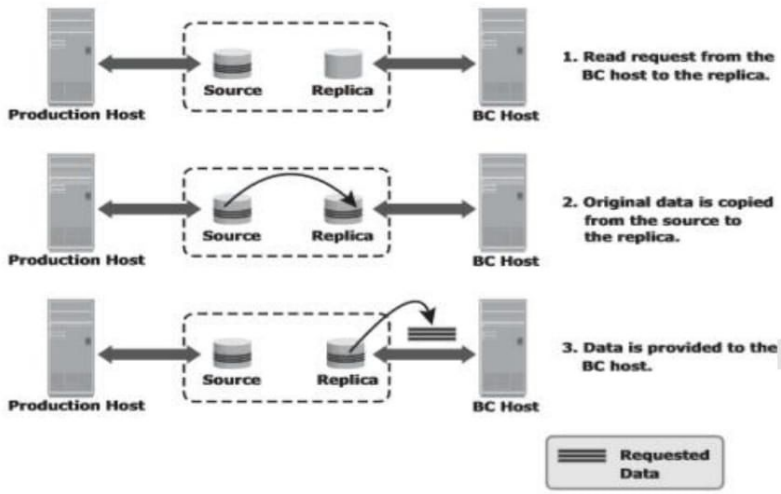


Figure 11-10: Copy on first access (CoFA) – read from target

When a write is issued to the target for the first time after the replication session activation, the original data is copied from the source to the target. After this, the new data is updated on the target (see Figure 11-11).

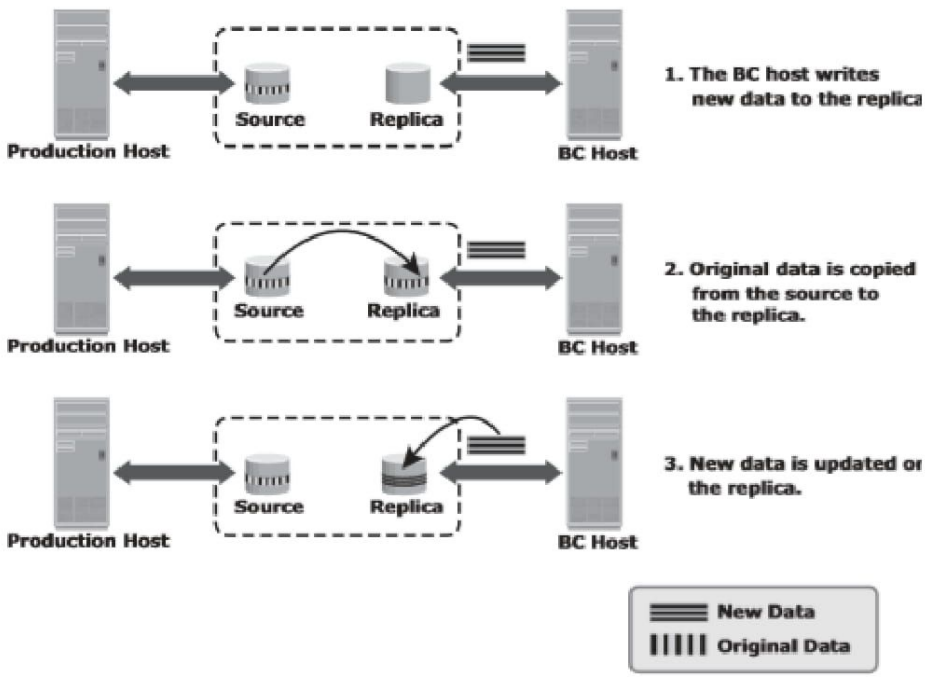


Figure 11-11: Copy on first access (CoFA) – write to target

In all cases, the protection bit for that block is reset to indicate that the original data has been copied over to the target. The pointer to the source data can now be discarded.

Subsequent writes to the same data block on the source, and reads or writes to the same data blocks on the target, do not trigger a copy operation (and hence are termed Copy on First Access).

2. Full Copy Mode

On session start, the entire contents of the Source device are copied to the Target device in the background. If the replication session is terminated, the target will contain all the original data from the source at the PIT of activation. Target can be used for restore and recovery In CoFA mode, the target will only have data was accessed until termination, and therefore it cannot be used for restore and recovery Most vendor implementations provide the ability to track changes:

- Made to the Source or Target
- Enables incremental re-synchronization

Pointer-Based Virtual Replication

- In *pointer-based virtual replication*, at the time of session activation, the target contains pointers to the location of data on the source.
- The target does not contain data, at any time. Hence, the target is known as a *virtual replica*. Similar to pointer-based full-volume replication, a protection bitmap is created for all data on the source device, and the target is immediately accessible. Granularity can range from 512-byte blocks to 64 KB blocks or greater.
- When a write is issued to the source for the first time after session activation, original data at that address is copied to a predefined area in the array. This area is generally termed the *save location*. The pointer in the target is updated to point to this data address in the save location. After this, the new write is updated on the source. This process is illustrated in Figure 13-10.

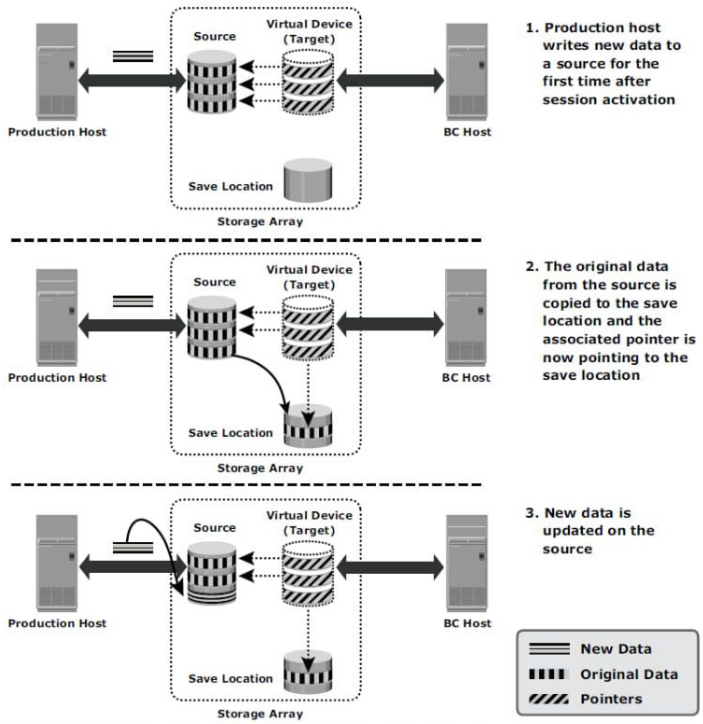


Figure 13-10: Pointer-based virtual replication – write to source

- When a write is issued to the target for the first time after session activation, original data is copied from the source to the save location and similarly the pointer is updated to data in save location. Another copy of the original data is created in the save location before the new write is updated on the save location. This process is illustrated in Figure 13-11.
- When reads are issued to the target, unchanged data blocks since session activation are read from the source. Original data blocks that have changed are read from the save location.
- Pointer-based virtual replication uses CoFW technology. Subsequent writes to the same data block on the source or the target do not trigger a copy operation.
- Data on the target is a combined view of unchanged data on the source and data on the save location. Unavailability of the source device invalidates the data on the target. As the target only contains pointers to data, the physical capacity required for the target is a fraction of the source device. The capacity required for the save location depends on the amount of expected data change.

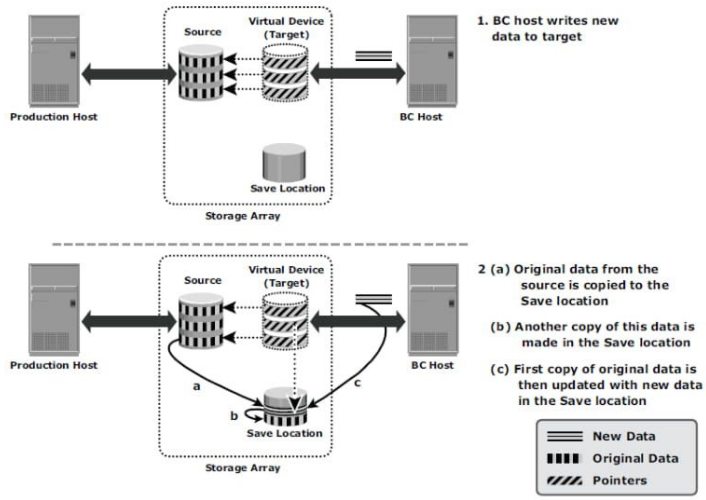


Figure 13-11: Pointer-based virtual replication – write to target

Network-Based Local Replication

- In network-based replication, the replication occurs at the network layer between the hosts and storage arrays.
- Network-based replication combines the benefits of array-based and host-based replications.
- By offloading replication from servers and arrays, network-based replication can work across a large number of server platforms and storage arrays, making it ideal for highly heterogeneous environments. Continuous data protection (CDP) is a technology used for network-based local and remote replications.

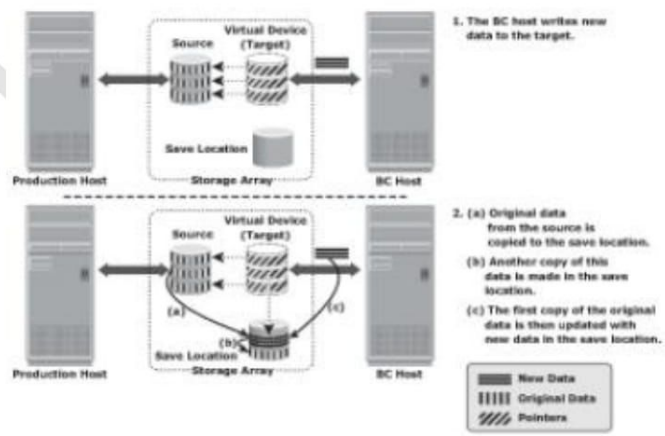


Figure 11-13: Pointer-based virtual replication – write to target

Local Replication in a Virtualized Environment

- Local replication of VMs is performed by the **hypervisor** at the compute level. However, it can also be performed at the **storage level** using **array-based local replication**, similar to the physical environment.
- In the **array-based method**, the LUN on which the VMs reside is replicated to another LUN in the same array.
- For **hypervisor-based** local replication, two options are available: **VM Snapshot and VM Clone.**
- **VM Snapshot** captures the state and data of a running virtual machine at a specific point in time. The VM state includes VM files, such as BIOS, network configuration, and its power state (powered-on, powered-off, or suspended).
- The VM data includes all the fi les that make up the VM, including virtual disks and memory. A VM Snapshot uses a separate delta file to record all the changes to the virtual disk since the snapshot session is activated.
- Snapshots are useful when a VM needs to be reverted to the previous state in the event of logical corruptions.
- Reverting a VM to a previous state causes all settings configured in the guest OS to be reverted to that PIT when that snapshot was created.
- **Challenges:** It does not support data replication if a virtual machine accesses the data by using raw disks.
- Using the hypervisor to perform snapshots increases the load on the compute and impacts the compute performance.
- **VM Clone** is another method that creates an identical copy of a virtual machine. When the cloning operation is complete, the clone becomes a separate VM from its parent VM.
- The clone has its own MAC address, and changes made to a clone do not affect the parent VM. Similarly, changes made to the parent VM do not appear in the clone.
- **Advantages:** VM Clone is a useful method when there is a need to deploy many identical VMs.
- Installing guest OS and applications on multiple VMs is a time-consuming task; VM Clone

helps to simplify this process.

Chapter 12

Remote Replication

Remote replication is the process to create replicas of information assets at remote sites (locations). Remote replication helps organizations mitigate the risks associated with regionally driven outages resulting from natural or human-made disasters

Remote Replication Technologies

Remote replication of data can be handled by the hosts or storage arrays. Other options include specialized network-based appliances to replicate data over the LAN or SAN.

Host-Based Remote Replication

Host-based remote replication uses the host resources to perform and manage the replication operation.

There are two basic approaches to host-based remote replication:

1. Logical volume manager (LVM) based replication
2. Database replication via log shipping.

LVM-Based Remote Replication

- LVM-based remote replication is performed and managed at the volume group level. Writes to the source volumes are transmitted to the remote host by the LVM.
- The LVM on the remote host receives the writes and commits them to the remote volume group. Prior to the start of replication, identical volume groups, logical volumes, and file systems are created at the source and target sites.
- Initial synchronization of data between the source and replica is performed.
- One method to perform initial synchronization is to back up the source data and restore the data to the remote replica.
- Alternatively, it can be performed by replicating over the IP network. Until the completion

of the initial synchronization, production work on the source volumes is typically halted.

- After the initial synchronization, production work can be started on the source volumes and replication of data can be performed over an existing standard IP network (Fig 12-5).

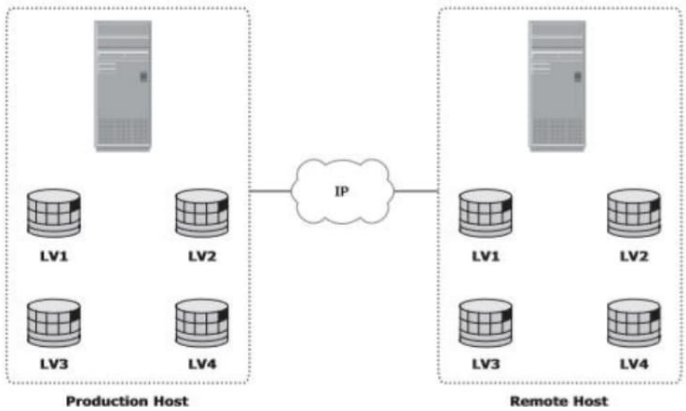


Figure 12-5: LVM-based remote replication

- LVM-based remote replication supports both synchronous and asynchronous modes of replication. If a failure occurs at the source site, applications can be restarted on the remote host, using the data on the remote replicas.
- LVM-based remote replication is independent of the storage arrays and therefore supports replication between heterogeneous storage arrays.
- Systems are shipped with LVMs, so additional licenses and specialized hardware are not typically required.

Host-Based Log Shipping

- Database replication via log shipping is a host-based replication technology supported by most databases.
- Transactions to the source database are captured in logs, which are periodically transmitted by the source host to the remote host (see Figure 12-6).
- The remote host receives the logs and applies them to the remote database.
- Prior to starting production work and replication of log fi les, all relevant components of the source database are replicated to the remote site. This is done while the source database is shut down. After this step, production work is started on the source database.

- The remote database is started in a standby mode. Typically, in standby mode, the database is not available for transactions.
- All DBMSs switch log fi les at preconfigured time intervals or when a log file is full.
- The current log fi le is closed at the time of log switching, and a new log fi le is opened. When a log switch occurs, the closed log fi le is transmitted by the source host to the remote host.

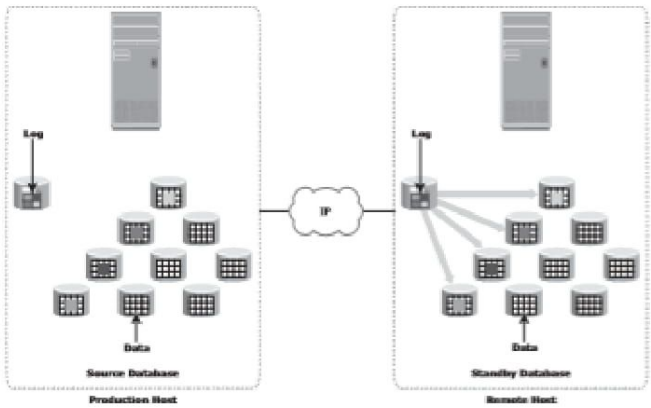


Figure 12-6: Host-based log shipping

- The remote host receives the log and updates the standby database. This process ensures that the standby database is consistent up to the last committed log.
- RPO at the remote site is finite and depends on the size of the log and the frequency of log switching.

Storage Array-Based Remote Replication

- In storage array-based remote replication, the array-operating environment and resources perform and manage data replication. This relieves the burden on the host CPUs, which can be better used for applications running on the host.
- A source and its replica device reside on different storage arrays.
- Data can be transmitted from the source storage array to the target storage array over a shared or a dedicated network.
- Replication between arrays may be performed in **synchronous, asynchronous, or disk-buffered modes.**

1. Synchronous Replication Mode

- In array-based synchronous remote replication, writes must be committed to the source and the target prior to acknowledging "write complete" to the production host.
- Additional writes on that source cannot occur until each preceding write has been completed and acknowledged. Figure 12-7 shows the array-based synchronous remote replication process

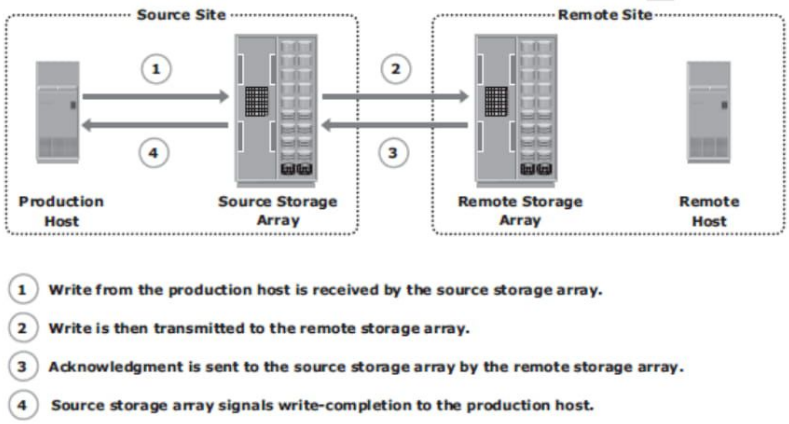
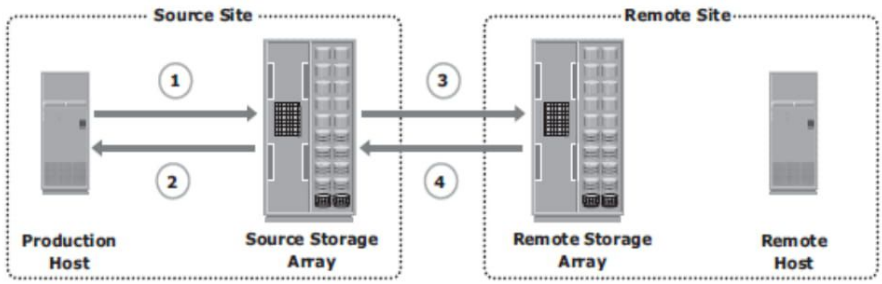


Figure 12-7: Array-based synchronous remote replication

2. Asynchronous Replication Mode

- In array-based asynchronous remote replication mode, as shown in Figure 12-8, a write is committed to the source and immediately acknowledged to the host.
- Data is buffered at the source and transmitted to the remote site later.
- The source and the target devices do not contain identical data at all times.
- The data on the target device is behind that of the source, so the RPO in this case is not zero.



1. The production host writes to the source storage array.
2. The source array immediately acknowledges the production host.
3. These writes are then transmitted to the target array.
4. After the writes are received by the target array, it sends an acknowledgment to the source array.

Figure 12-8: Array-based asynchronous remote replication

3. Disk-Buffered Replication Mode

- Disk-buffered replication is a combination of local and remote replication technologies.
- A consistent PIT local replica of the source device is first created.
- This is then replicated to a remote replica on the target array. Figure 12-9 shows the sequence of operations in a disk-buffered remote replication.
- At the beginning of the cycle, the network links between the two arrays are suspended, and there is no transmission of data.
- While production application runs on the source device, a consistent PIT local replica of the source device is created. The network links are enabled, and data on the local replica in the source array transmits to its remote replica in the target array.
- After synchronization of this pair, the network link is suspended, and the next local replica of the source is created.
- Optionally, a local PIT replica of the remote device on the target array can be created.

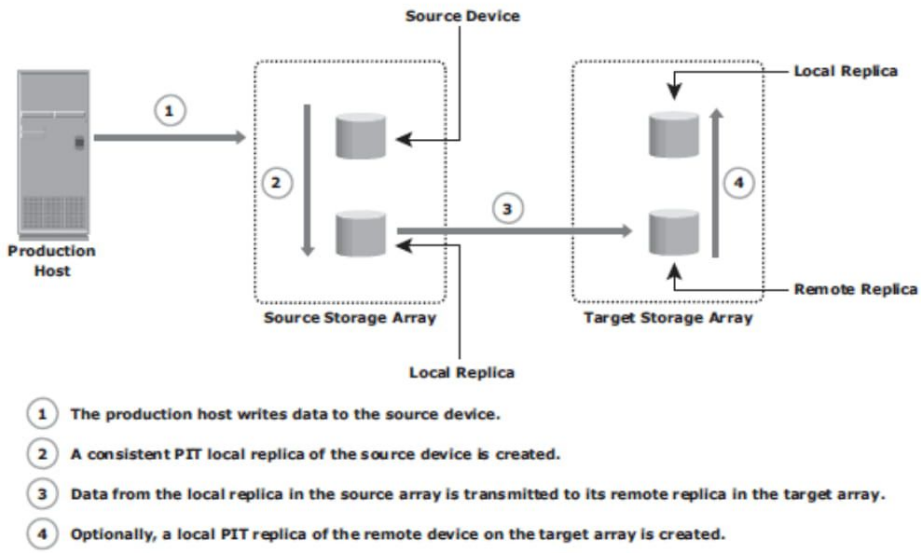
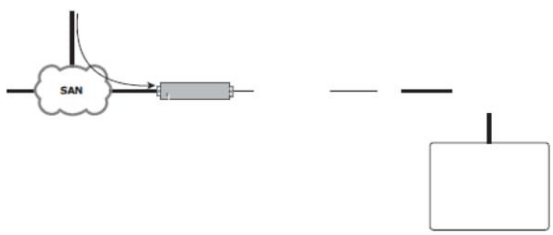


Figure 12-9: Disk-buffered remote replication

Network-Based Remote Replication

In network-based remote replication, the replication occurs at the network layer between the host and storage array.

In normal operation, CDP remote replication provides any-point-in-time recovery capability, which enables the target LUNs to be rolled back to any previous point in time. Similar to CDP local replication, CDP remote replication typically uses a journal volume, CDP appliance, or CDP software installed on a separate host (host-based CDP), and a write splitter to perform replication between sites. The CDP appliance is maintained at both source and remote sites. Figure 12-10 describes CDP remote replication.



- In this method, the replica is synchronized with the source, and then the replication process

starts. After the replication starts, all the writes from the host to the source are split into two copies.

- One of the copies is sent to the local CDP appliance at the source site, and the other copy is sent to the production volume.
- After receiving the write, the appliance at the source site sends it to the appliance at the remote site. Then, the write is applied to the journal volume at the remote site.
- For an asynchronous operation, writes at the source CDP appliance are accumulated, and redundant blocks are eliminated.
- Then, the writes are sequenced and stored with their corresponding timestamp. The data is then compressed, and a checksum is generated. It is then scheduled for delivery across the IP or FC network to the remote CDP appliance.
- After the data is received, the remote appliance verifies the checksum to ensure the integrity of the data. The data is then uncompressed and written to the remote journal volume.
- As a next step, data from the journal volume is sent to the replica at predefined intervals.

Three-Site Replication

- Three-site replication mitigates the risks identified in two-site replication.
- In a three-site replication, data from the source site is replicated to two remote sites.
- Replication can be synchronous to one of the two sites, providing a near zero-RPO solution, and it can be asynchronous or disk buffered to the other remote site, providing a finite RPO.

Three-site remote replication can be implemented as

1. Cascade/multihop
2. Triangle/multitarget solution.

Three-Site Replication — Cascade/Multihop

- In the cascade/multihop three-site replication, data fl ows from the source to the intermediate storage array, known as a bunker, in the first hop, and then from a bunker to

a storage array at a remote site in the second hop.

- Replication between the source and the remote sites can be performed in two ways: **synchronous + asynchronous or synchronous + disk buffered**.
- Replication between the source and bunker occurs synchronously, but replication between the bunker and the remote site can be achieved either as disk-buffered mode or asynchronous mode.

Synchronous + Asynchronous

- This method employs a combination of synchronous and asynchronous remote replication technologies.
- Synchronous replication occurs between the source and the bunker.
- Asynchronous replication occurs between the bunker and the remote site. The remote replica in the bunker acts as the source for asynchronous replication to create a remote replica at the remote site. Figure 12-11 (a) illustrates the synchronous + asynchronous method.

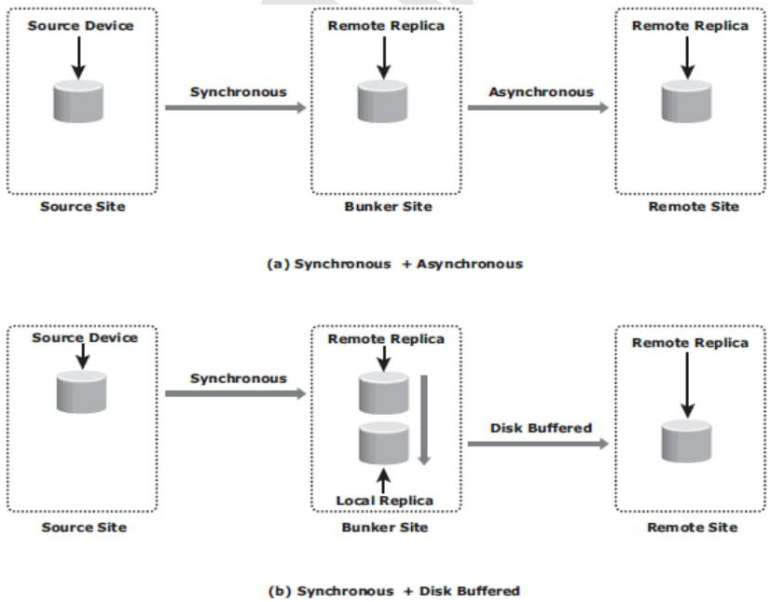


Figure 12-11: Three-site remote replication cascade/multihop

Synchronous + Disk Buffered

- This method employs a combination of local and remote replication technologies.

- Synchronous replication occurs between the source and the bunker: a consistent PIT local replica is created at the bunker.
- Data is transmitted from the local replica at the bunker to the remote replica at the remote site. Optionally, a local replica can be created at the remote site after data is received from the bunker. Figure 12-11 (b) illustrates the synchronous + disk buffered method.
- In this method, a minimum of four storage devices are required (including the source) to replicate one storage device.
- The other three devices are the synchronous remote replica at the bunker, a consistent PIT local replica at the bunker, and the replica at the remote site.
- RPO at the remote site is usually in the order of hours for this implementation.

Three-Site Replication — Triangle/Multitarget

- In three-site triangle/multitarget replication, data at the source storage array is concurrently replicated to two different arrays at two different sites, as shown in Figure 12-12. The source-to-bunker site (target 1) replication is synchronous with a near-zero RPO.
- The source-to-remote site (target 2) replication is asynchronous with an RPO in the order of minutes.
- The distance between the source and the remote sites could be thousands of miles.
- This implementation does not depend on the bunker site for updating data on the remote site because data is asynchronously copied to the remote site directly from the source.
- The triangle/multitarget configuration provides consistent RPO unlike cascade/ multihop solutions in which the failure of the bunker site results in the remote site falling behind and the RPO increasing.

Benefits

- ✓ The ability to failover to either of the two remote sites in the case of source-site failure, with disaster recovery (asynchronous) protection between the bunker and remote sites.
- ✓ Resynchronization between the two surviving target sites is incremental. Disaster recovery protection is always available if any one-site failure occurs.

Remote Replication and Migration in a Virtualized Environment

- Virtual machine migration is another technique used to ensure business continuity in case of hypervisor failure or scheduled maintenance.
- VM migration is the process to move VMs from one hypervisor to another without powering off the virtual machines.
- VM migration also helps in load balancing when multiple virtual machines running on the same hypervisor contend for resources.

Two commonly used techniques for VM migration are

1. **Hypervisor-to-hypervisor**
2. **Array-to-array migration.**

1. **In hypervisor-to-hypervisor VM migration**

   - The entire active state of a VM is moved from one hypervisor to another. Figure 12-14 shows hypervisor-to hypervisor VM migration.
   - This method involves copying the contents of virtual machine memory from the source hypervisor to the target and then transferring the control of the VM's disk fi les to the target hypervisor.

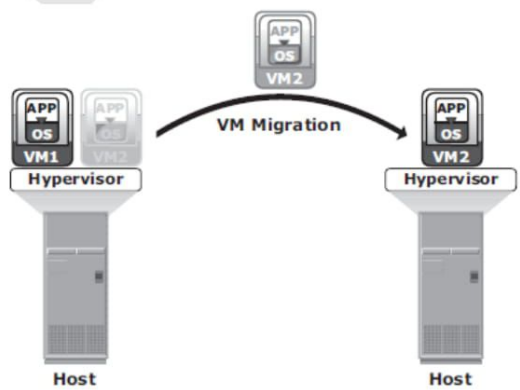


Figure 12-14: Hypervisor-to-hypervisor VM migration

   - Because the virtual disks of the VMs are not migrated, this technique requires both source and target hypervisor access to the same storage.

In array-to-array VM migration

- Virtual disks are moved from the source array to the remote array. This approach enables the administrator to move VMs across dissimilar storage arrays.
- Figure 12-15 shows array-to-array VM migration. Array-to-array migration starts by copying the metadata about the VM from the source array to the target.

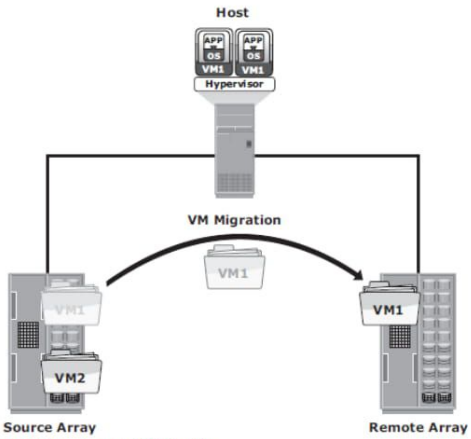


Figure 12-15: Array-to-array VM migration

- The metadata essentially consists of configuration, swap, and log files. After the metadata is copied, the VM disk fi le is replicated to the new location.

Sample Questions

1. Describe the failure analysis in BC. Briefly explain BC technology solution
2. With a neat diagram explain the steps involved in backup and restore operation.
3. What is information availability? Explain how information availability is defined and measured.
4. What is BC. Explain BC planning life cycle with a neat diagram
5. Explain the reasons for which backup is performed
6. What is BC? Explain the BC terminology in detail.
7. What is data deduplication? Explain the implementation of data deduplication.
8. Explain failure analysis.
9. Explain the factors used for measuring information availability.
10. Explain backup method and architecture.
11. Explain backup and restore operations.

12. Explain backup topologies. With a neat diagram
13. Explain backup technologies.
14. Explain backup in virtualized Environment with a neat diagram
15. Explain the different backup targets with comparison.
16. Explain local replication technologies.
17. Compare local replication technologies.
18. Explain modes of remote replication.
19. Briefly explain remote replication technologies.
20. Explain network infrastructure over remote replication.
21. What is local replication. Explain host based local replication technologies
22. Explain array based local replication technologies with a neat diagram
23. Differentiate between pointer based full volume replication and pointer based virtual replication methodology
24. Explain in detail local replication in Virtualized Environment.
25. Define remote replication. Explain remote replication technologies.
26. Explain with a neat diagram storage array based remote replication
27. Explain Synchronous + Asynchronous and Synchronous + Disk Buffered methods of three-site replication with neat diagram.
28. Explain Remote Replication and Migration in a Virtualized Environment.

Module 4

Syllabus: Cloud Computing and Virtualization Cloud Enabling Technologies, Characteristics of Cloud Computing, Benefits of Cloud Computing, Cloud Service Models, Cloud Deployment Models, Cloud Computing Infrastructure, Cloud Challenges and Cloud Adoption Considerations. **Virtualization Appliances**: Black Box Virtualization, In-Band Virtualization Appliances, Outof-Band Virtualization Appliances, High Availability for Virtualization Appliances, Appliances for Mass Consumption. **Storage Automation and Virtualization**: Policy-Based Storage Management, Application-Aware Storage Virtualization, Virtualization-Aware Applications.

Text Book-1 Ch13: 13.1 to 13.8. Text Book-2 Ch9: 9.1 to 9.5 Ch13: 13.1 to 13.3

Cloud Computing

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction

13.1 Cloud Enabling Technologies

Grid computing, utility computing, virtualization, and service-oriented architecture are enabling technologies of cloud computing.

- *Grid computing* is a form of distributed computing that enables the resources of numerous heterogeneous computers in a network to work together on a single task at the same time. Grid computing enables parallel computing and is best for large workloads.
- *Utility computing* is a service-provisioning model in which a service pro- vider makes computing resources available to customers, as required, and charges them based on usage. This is analogous to other utility services, such as electricity, where charges are based on the consumption.

- *Virtualization* is a technique that abstracts the physical characteristics of IT resources from resource users. It enables the resources to be viewed and managed as a pool and lets users create virtual resources from the pool. Virtualization provides better flexibility for provisioning of IT resources compared to provisioning in a non-virtualized environment. It helps optimize resource utilization and delivering resources more efficiently.
- *Service Oriented Architecture* (SOA) provides a set of services that can com- municate with each other. These services work together to perform some activity or simply pass data among services.

13.2 Characteristics of cloud computing

A computing infrastructure used for cloud services must have certain capabili-ties or characteristics. According to NIST, the cloud infrastructure should have five essential characteristics:

On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed, automatically without requiring human interaction with each service provider. A cloud service provider publishes a service catalogue, which contains information about all cloud services available to consumers. The service catalogue includes information about service attributes, prices, and request processes. Consumers view the service catalogue via a web-based user

interface and use it to request for a service. Consumers can either lever- age the "ready-to-use" services or change a few service parameters to customize the services.

Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (for example, mobile phones, tablets, laptops, and workstations).

Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multitenant model, with different physical and virtual resources dynamically assigned

and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (for example, country, state, or data center). Examples of resources include storage, processing, memory, and network bandwidth.

Rapid elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward com- mensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstrac- tion appropriate to the type of service (for example, storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

13. 3 Benefits of Cloud Computing

Cloud computing offers the following key benefits:

- **Reduced IT cost:** Cloud services can be purchased based on pay-per-use or subscription pricing. This reduces or eliminates the consumer's IT capital expenditure (CAPEX).
- **Business agility:** Cloud computing provides the capability to allocate and scale computing capacity quickly. Cloud computing can reduce the time required to provision and deploy new applications and services from months to minutes. This enables businesses to respond more quickly to market changes and reduce time-to-market.
- **Flexible scaling:** Cloud computing enables consumers to scale up, scale down, scale out, or scale in the demand for computing resources easily. Consumers can unilaterally and automatically scale computing resources without any interaction with cloud service providers. The flexible service provisioning capability of cloud computing often provides a sense of unlimited scalability to the cloud service consumers.
- **High availability:** Cloud computing has the capability to ensure resource availability at varying levels depending on the consumer's policy and priority. Redundant infrastructure components (servers, network paths, and storage equipment, along with clustered software) enable fault tolerance for cloud deployments. These techniques can encompass multiple data centers located in different geographic regions, which prevents data unavailability due to regional failures.

13.4 Cloud Service Models

According to NIST, cloud service offerings are classifi ed primarily into three models: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS).

13.4.1 Infrastructure-as-a-Service

IaaS is the base layer of the cloud services stack (see Figure 13-1 [a]). It serves as the foundation for both the SaaS and PaaS layers.

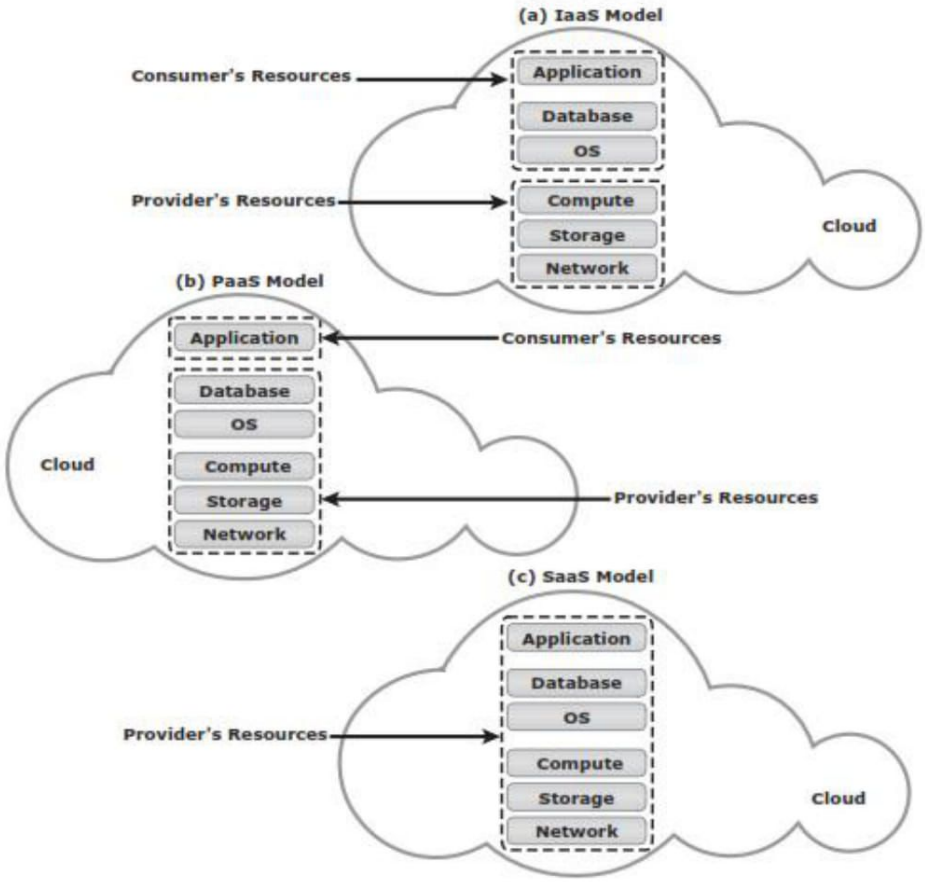


Figure 13-1: IaaS, PaaS, and SaaS models

Amazon Elastic Compute Cloud (Amazon EC2) is an example of IaaS that provides scalable compute capacity, on-demand, in the cloud. It enables consumers to leverage Amazon's massive computing infrastructure with no up-front capital investment.

13. 4. 2 Platform-as-a-Ser vice

The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, braries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. (See Figure 13-1[b]).

PaaS is also used as an application development environment, offered as aservice by the cloud service provider. The consumer may use these platforms to code their applications and then deploy the applications on the cloud.

13. 4. 3 Software-as-a-Ser vice

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (for example, web-based e-mail), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings(See Figure 13-1[c]).

In a SaaS model, applications, such as customer relationship management (CRM), e-mail, and instant messaging (IM), are offered as a service by the cloud service providers. The cloud service providers exclusively manage the required computing infrastructure and software to support these services.

13.5 Cloud Deployment Models

According to NIST, cloud computing is classified into four deployment models — public, private, community, and hybrid — which provide the basis for how cloud infrastructures are constructed and consumed.

13.5.1 Public Cloud

In a ***public cloud*** model, the cloud infrastructure is provisioned for open usemby the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

Consumers use the cloud services offered by the providers via the Internet and pay metered usage charges or subscription fees. An advantage of the public cloud is its low capital cost with enormous scalability. However, for consumers, these benefits come with certain risks: no control over the resources in the cloud, the security of confidentiality data, network performance, and interoperability issues. Popular public cloud service providers are Amazon, Google and

Salesforce.com. Figure 13-2 shows a public cloud that provides cloud services organizations and individuals.

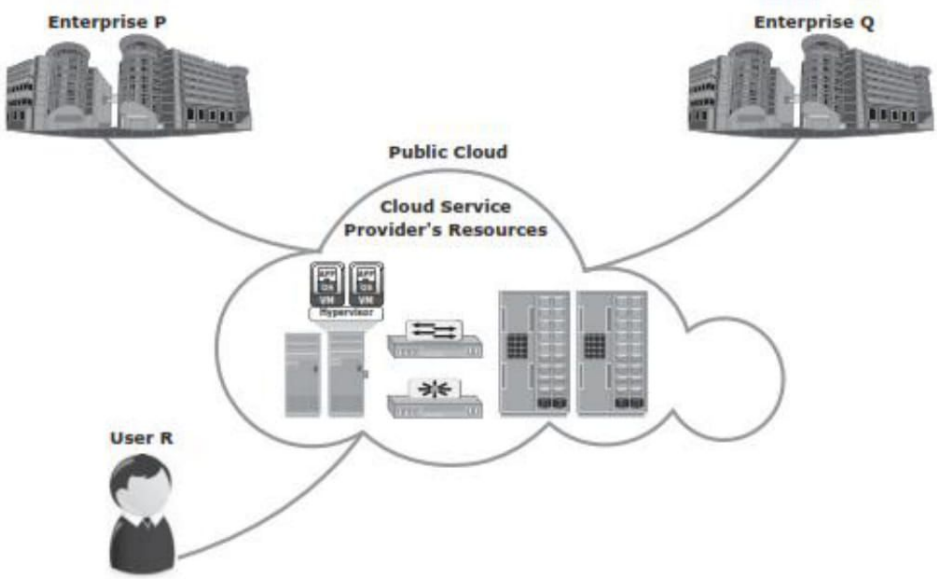


Figure 13-2: Public cloud

13.5.2 Private Cloud

In a *private cloud* model, the cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (for example, business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises. Following are two variations to the private cloud model:

On-premise private cloud: The on-premise private cloud, also known as internal cloud, is hosted by an organization within its own data centers (see Figure 13-3 [a]). This model enables organizations to standardize their cloud service management processes and security, although this model has limitations in terms of size and resource scalability. Organizations would also need to incur the capital and operational costs for the physical resources. This is best suited for organizations that require complete control over their applications, infrastructure configurations, and security mechanisms.

Externally hosted private cloud: This type of private cloud is hosted external to an organization (see Figure 13-3 [b]) and is managed by a third- party organization. The third-party

organization facilitates an exclusive cloud environment for a specifi c organization with full guarantee of privacy and confidentiality.

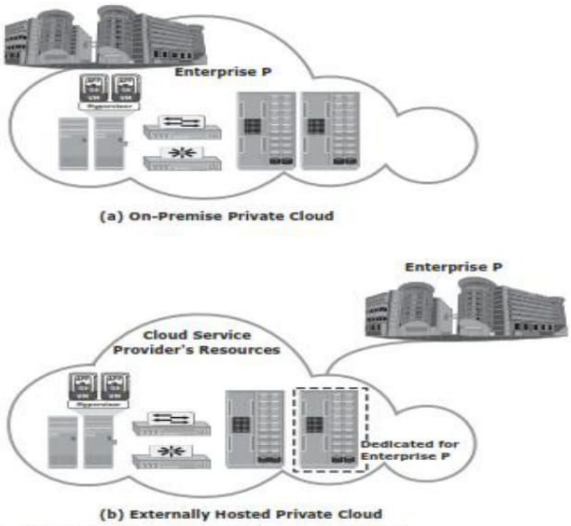


Figure 13-3: On-premise and externally hosted private clouds

13.5.3 Community Cloud

In a *community cloud* model, the cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (for example, mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises. (See Figure 13-4).

In a community cloud, the costs spread over to fewer consumers than a public cloud. Hence, this option is more expensive but might offer a higher level of privacy, security, and compliance. The community cloud also offers organizations access to a vast pool of resources compared to the private cloud.

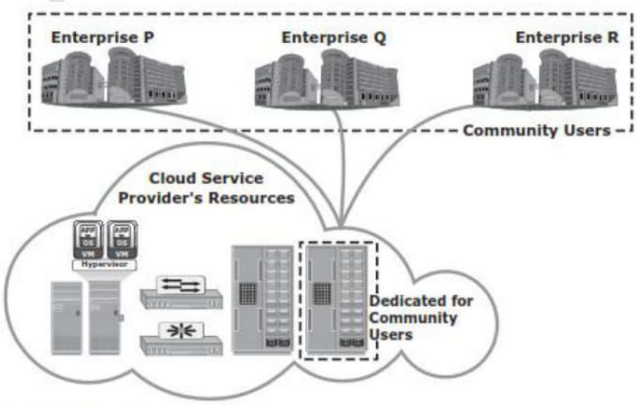


Figure 13-4: Community cloud

13.5.4 Hybrid Cloud

In a *hybrid cloud* model, the cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (for example, cloud bursting for load balancing between clouds).

The hybrid model allows an organization to deploy less critical applications and data to the public cloud, leveraging the scalability and cost-effectiveness of the public cloud. The organization's mission-critical applications and data remain on the private cloud that provides greater security. Figure 13-5 shows an example of a hybrid cloud.

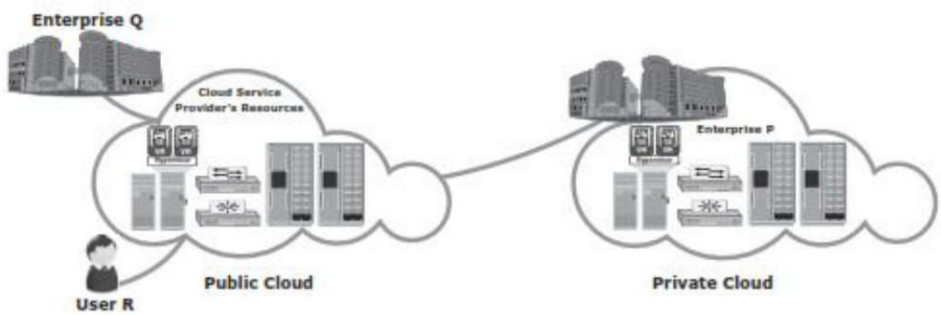


Figure 13-5: Hybrid cloud

13.6 Cloud Computing Infrastructure

A cloud computing infrastructure is the collection of hardware and software that enables the five essential characteristics of cloud computing. Cloud computing infrastructure usually consists of the following layers:

- Physical infrastructure
- Virtual infrastructure
- Applications and platform software
- Cloud management and service creation tools

13.6.1 Physical Infrastructure

The physical infrastructure consists of physical computing resources, which include physical servers, storage systems, and networks. Physical servers are connected to each other, to the storage systems, and to the clients via networks, such as IP, FC SAN, IP SAN, or FCoE networks.

Cloud service providers may use physical computing resources from one or more data centers to provide services. If the computing resources are distributed across multiple data centers, connectivity must be established among them. The connectivity enables the data centers in different locations to work as a single large data center. This enables migration of business applications and data across data centers and provisioning cloud services using the resources from multiple data centers.

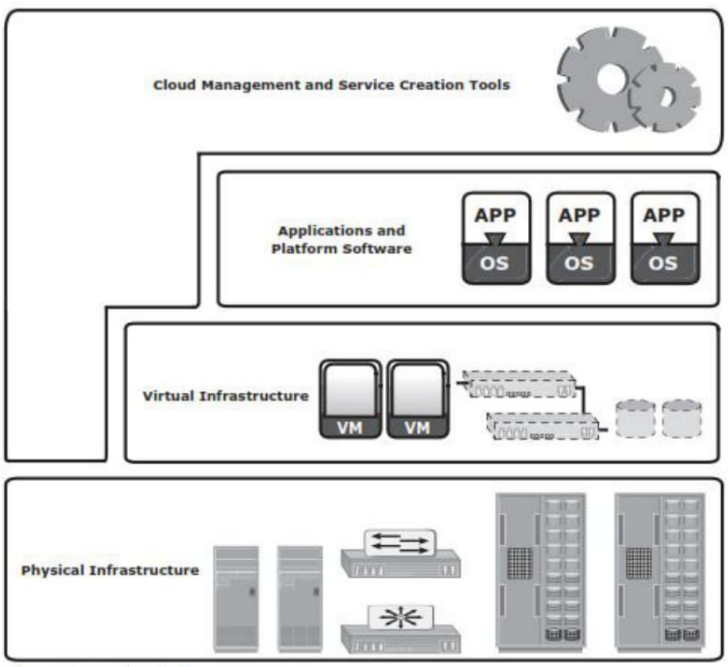


Figure 13-6: Cloud infrastructure layers

13.6.2 Virtual Infrastructure

- Cloud service providers employ virtualization technologies to build a virtual infrastructure layer on the top of the physical infrastructure.
- Virtualization enables fulfilling some of the cloud characteristics, such as resource pooling
  and rapid elasticity. It also helps reduce the cost of providing the cloud services.

- Virtualization abstracts physical computing resources and provides a consolidated view of the resource capacity.
- The
- Consolidated resources are managed as a single entity called a *resource pool*. For example, a resource pool might group CPUs of physical servers within a cluster.
- The capacity of the resource pool is the sum of the power of all CPUs (for example, 10,000 megahertz) available in the cluster.
- In addition to the CPU pool, the virtual infrastructure includes other types of resource pools, such as memory pool, network pool, and storage pool.
- Apart from resource pools, the virtual infrastructure also includes *identity pools*, such as VLAN ID pools and VSAN ID pools.
- Virtual infrastructure also includes virtual computing resources, such as virtual machines, virtual storage volumes, and virtual networks.
- These resources obtain capacities, such as CPU power, memory, network bandwidth, and storage space from the resource pools.
-

13.6.3 Applications and Platform Software

- This layer includes a suite of business applications and platform software, such as the OS and database.
- Platform software provides the environment on which business applications run.
- Applications and platform software are hosted on virtual machines to create SaaS and PaaS.

13.6.4 Cloud Management and Service Creation Tools

The cloud management and service creation tools layer includes three types of software:

- Physical and virtual infrastructure management software
- Unified management software
- User-access management software

The physical and virtual infrastructure management software is offered by the vendors of various infrastructure resources and third-party organizations. For example, a storage array has its own management software. Similarly, network and physical servers are managed independently using network and compute management software respectively.

- This software provides interfaces to construct a virtual infrastructure from the underlying physical infrastructure.

- ***Unified management software*** interacts with all standalone physical and virtual infrastructure management software.
- It collects information on the existing physical and virtual infrastructure configurations, connectivity, and utilization.

- Unified management software compiles this information and provides a consolidated view of infrastructure resources scattered across one or more data centers.
- It allows an administrator to monitor performance, capacity, and availability of physical and virtual resources centrally.
- Unified management software also provides a single management interface to configure physical and virtual infrastructure and integrate the compute (both CPU and memory), network, and storage pools.
- The key function of the unified management software is to automate the creation of cloud services.
- It enables administrators to define service attributes such as CPU power, memory, network bandwidth, storage capacity, name and description of applications and platform software, resource location, and backup policy.
- When the unified management software receives consumer requests for cloud services, it creates the service based on predefined service attributes.
- The ***user-access management software*** provides a web-based user interface to consumers.
- Consumers can use the interface to browse the service catalogue and request cloud services.
- The user-access management software authenticates users before forwarding their request to the unified management software.
- It also monitors allocation or usage of resources associated to the cloud service instances.
- Based on the allocation or usage of resources, it generates a chargeback report.
- The chargeback report is visible to consumers and provides transparency between consumers and providers.

13.7 Cloud Challenges

13.7.1 Challenges for Consumers

- **Security and regulation**
- Consumers are indecisive to transfer control of sensitive data
- Regulation may prevent organizations to use cloud services
- **Network latency**
- Real time applications may suffer due to network latency and limited bandwidth
- **Supportability**
- Service provider might not support proprietary environments
- Incompatible hypervisors could impact VM migration
- **Vendor lock-in**
- Restricts consumers from changing their cloud service providers
- Lack of standardization across cloud-based platforms

Business-critical data requires protection and continuous monitoring of its access. If the data moves to a cloud model other than an on-premise private cloud, consumers could lose absolute control of their sensitive data. Although most of the cloud service providers offer enhanced data security, consumer might not be willing to transfer control of their business-critical data to the cloud.

Cloud service providers might use multiple data centers located in different countries to provide cloud services. They might replicate or move data across these data centers to ensure high availability and load distribution. Consumers may or may not know in which country their data is stored.

Cloud services can be accessed from anywhere via a network. However, network latency increases when the cloud infrastructure is not close to the access point. A high network latency can either increase the application response time or cause the application to timeout. This can be addressed by implementing stringent Service Level Agreements (SLAs) with the cloud service providers.

Another challenge is that cloud platform services may not support consumers' desired applications. For example, a service provider might not be able to support highly specialized or proprietary environments, such as compatible Oss and preferred programming languages, required to develop and run the consumer's application. Also, a mismatch between hypervisors could impact migration of virtual machines into or between clouds.

Another challenge is vendor lock-in: the difficulty for consumers to change their cloud service provider. A lack of interoperability between the APIs of different cloud service providers could also create complexity and high migration costs when moving from one service provider to another.

13.7.2 Challenges for Providers

- **Service warranty and service cost**
- Resources must be kept ready to meet unpredictable demand
- Hefty penalty, if SLAs are not fulfilled
- **Complexity in deploying vendor software in the cloud**
- Many vendors do not provide cloud-ready software licenses
- Higher cost of cloud-ready software licenses
-
- **No standard cloud access interface**
- Cloud consumers want open APIs
- Need agreement among cloud providers for standardization

Cloud service providers usually publish a service-level agreement (SLA) so that their consumers know about the availability of service, quality of service, downtime compensation, and legal and regulatory clauses. Alternatively, customer-specific SLAs may be signed between a cloud service provider and a consumer. SLAs typically mention a penalty amount if cloud service providers fail to provide the service levels. Therefore, cloud service providers must ensure that they have adequate resources to provide the required levels of services. Because the cloud resources are distributed and service demands fluctuate, it is a challenge for cloud service providers to provision physical resources for peak demand of all consumers and estimate the actual cost of providing the services.

Many software vendors do not have a cloud-ready software licensing model. Some of the software vendors offer standardized cloud licenses at a higher price compared to traditional licensing models. The cloud software licensing complexity has been causing challenges in deploying vendor software in the cloud. This is also a challenge to the consumer.

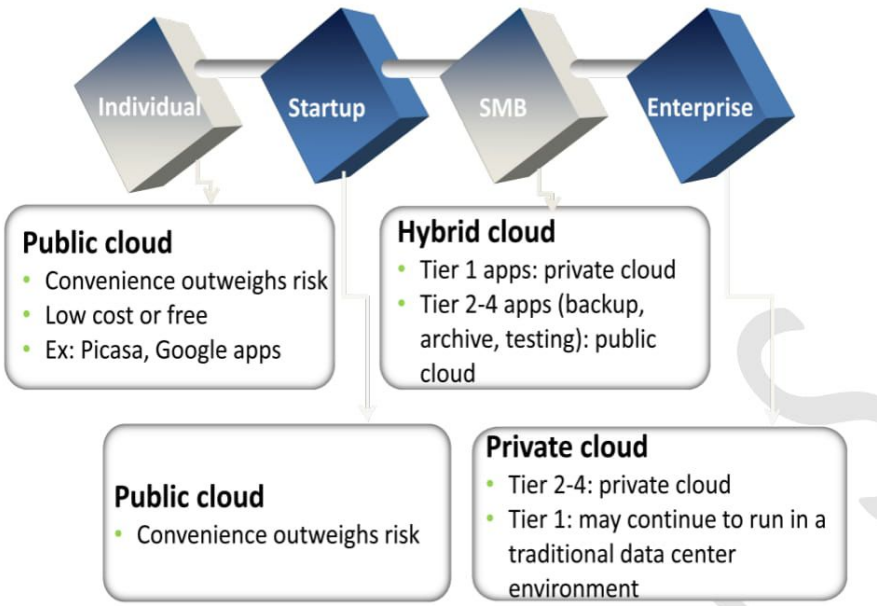
Cloud service providers usually offer proprietary APIs to access their cloud. However, consumers might want open APIs or standard APIs to become the tenant of multiple clouds. This is a challenge for cloud service providers because this requires agreement among cloud service providers.

13.8 Cloud Adoption Considerations

Cloud Adoption Considerations

- CIOs/IT Managers seeking to move to the cloud face several questions:
  - Which deployment model fits organization's requirements?
    - Private, public, hybrid
  - Which are the applications suitable for cloud?
  - How do I choose the cloud service provider?
  - Is the cloud infrastructure capable of providing the required Quality of Service (QoS)?
    - Performance, availability, and security
  - What is the financial benefit in adopting cloud?

Selection of a deployment model:



Application suitability:

- Some key questions to ask before migrating a consumer application to the public cloud:
  - Is the application compatible to cloud platform software? Is it a legacy application?
  - Is the application proprietary and mission-critical? Does the application provide competitive advantage?
  - Is the application workload network traffic intensive? Will application performance be impacted by network latency and limited network bandwidth?
  - Does the application communicate with other data center resources or applications?

Financial advantage:

- Require analysis of financial benefits in adopting cloud
- Consider CAPEX and OPEX to deploy and maintain own infrastructure versus cloud-adoption cost

Selection of a cloud service provider:

- Some key questions to ask before selecting a provider:

- How long and how well has the provider been delivering the services?
- How well does the provider meet the organization's current and future requirements?
- How easy is it to add or remove services?
- How easy is it to move to another provider, when required?
- What happens when the provider upgrades their software? Is it forced on everyone? Can you upgrade on your own schedule?
- Does the provider offer the required security services?
- Does the provider meet your legal and privacy requirements?
- Does the provider have good customer service support?

Service-level agreement (SLA):

- Consumers should check whether the QoS attributes meet their requirements
- SLA is a contract between the cloud service provider and consumers that defines QoS attributes
  - Attributes examples: throughput, uptime, and so on

VIRTUALIZATION APPLIANCES

- A virtual appliance is a software application residing and operating in a preconfigured virtual environment or platform.
- Storage virtualization appliances offer a means to pool storage assets and automate data replication, snapshot and other storage operation.
- Device with plug and play capability.
- Virtual appliances are accessed remotely by users and do not require locally-installed hardware.
- Variety of protocols like fibre channel, ISCSI, IP are supported by the host to access the appliances

9.1 Black box virtualization

- Virtualization appliances may exists as hardware processing platform that run an OS and Software and provide a variety of interconnects for attaching to SCSI, Fibre Channel or iSCSI environments.
- The appliance is attached to the network as a peer end device or inserted in-line between storage and servers.
- Appliance architecture vary from vendor to vendor.
- It can accommodate a variety of heterogeneous OS, host platform, storage target from different vendors.
- For enterprise data center the connectivity is done to the **fibre channel SAN's**
- For medium and small business the connectivity is done to **SCSI storage devices or a mix of parallel SCSI and Fibre channel array.**

9.2 In-Band Virtualization appliances

- Storage virtualization works on **control information** about where virtualized data storage actually exists and the transport of the actual data itself to and from the virtualized storage.
- **In-band virtualization appliances combine control information and data transport over the same path.**
- In a networked environment the control information in the form of meta data and the data is transmitted along the same path.
- In-line or in-band virtualization may be implemented in an number of ways depending on the type of storage targets and block access protocols used.

- For Fibre channel environments, in-band virtualization may split the storage network into 2 separate fabrics :
  - Host connectivity to the appliance
  - Appliance connectivity to storage targets.
- Server platforms have no direct access to physical storage arrays but communicate through the virtualization appliances.
- The virtualization appliance appear as a storage target to the fabric switch
- Fabric switch connects the server and presents target ID's and virtualized LUN's to the server fabric.
- For physical storage devices on the storage fabric the **appliances appears as an initiator** proxying multiple servers.

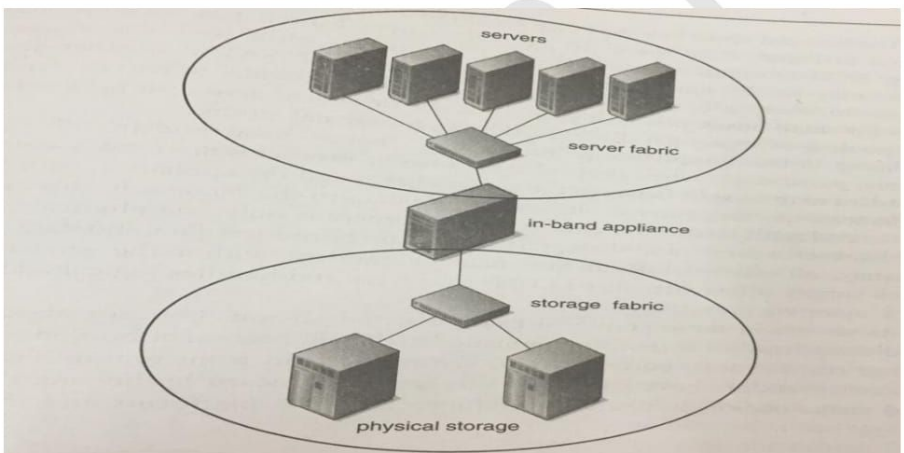


Figure 9.1 An in-band storage virtualization appliance may sit directly between separate Fibre Channel fabrics to physically segregate servers from the storage targets.

- **Ethernet attached host in-band virtualization.**
- **Block storage access over ethernet** is accomplished using iSCSI or vendor specific IP block protocols

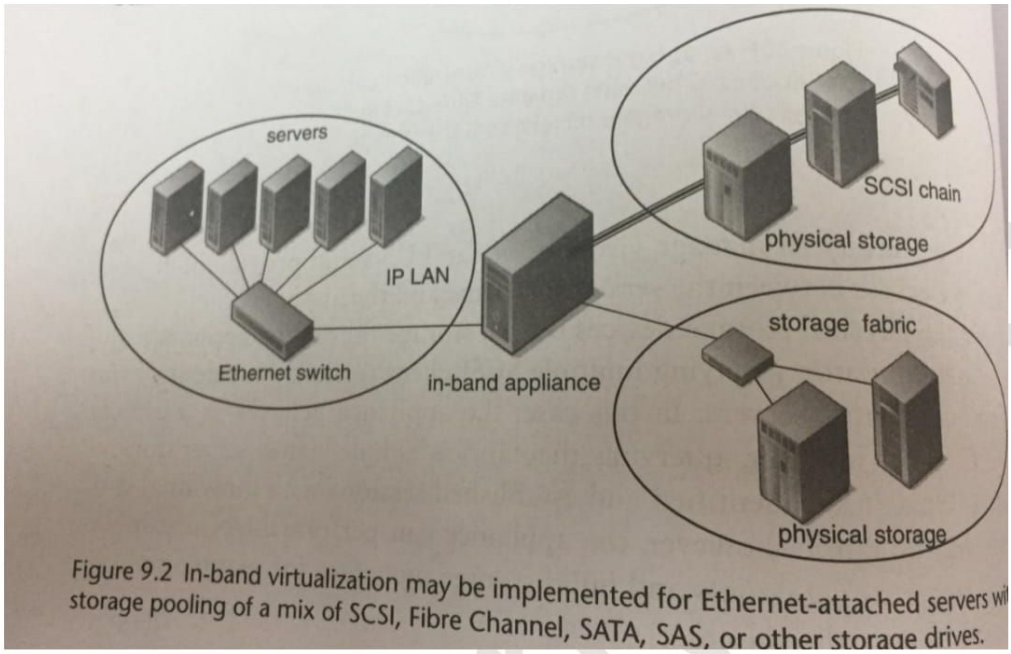


Figure 9.2 In-band virtualization may be implemented for Ethernet-attached servers wi storage pooling of a mix of SCSI, Fibre Channel, SATA, SAS, or other storage drives.

- **Backend storage access** may be through fibre channel, parallel SCSI, SATA or other disk protocols.
- Some in-band appliance solutions may require **host resident software drivers** particularly when proprietary protocols are used.
- Control metadata may be split between the server client and the appliance.
- Alternately all control information may be resident within the appliance.
- Advantage
  - Ability to enforce physical separation between the servers and the storage. The appliance acts as an intermediate between initiator and target.
  - It prevents servers from independently discovering and attaching to SAN based storage assets.
  -
- Dis Advantage

- In in-band virtualization the appliance itself will become a bottleneck for storage transaction *particularly as the traffic load from multiple servers increases.*
- As a solution, in-band virtualization software is hosted on multi process PC platforms and large amount of Cache memory is maintained.

9.3 Out-of-band of virtualization appliances

- Out-of band virtualization appliances use **separate path for control information and data** and thus place the appliance outside the primary path between the servers and the storage.
- An out-of band appliance may attach to an existing fibre channel SAN as a peer on the storage network.
- The control path is between the appliance and each SAN attached server switched directly through the fabric.
- Storage data doesn't pass through the appliance but traverses through the fabric directly between the storage and the servers.

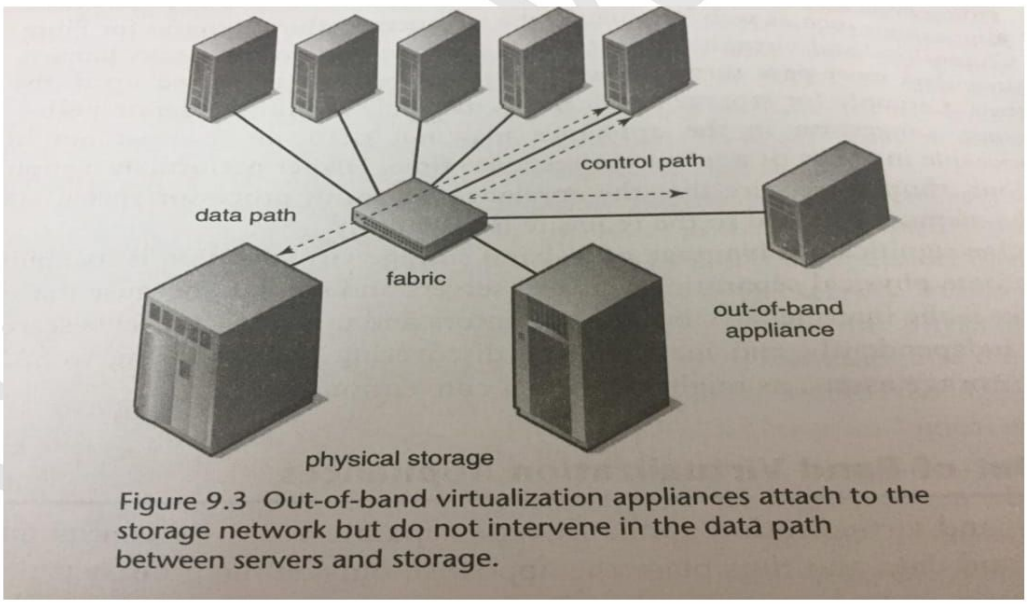


Figure 9.3 Out-of-band virtualization appliances attach to the storage network but do not intervene in the data path between servers and storage.

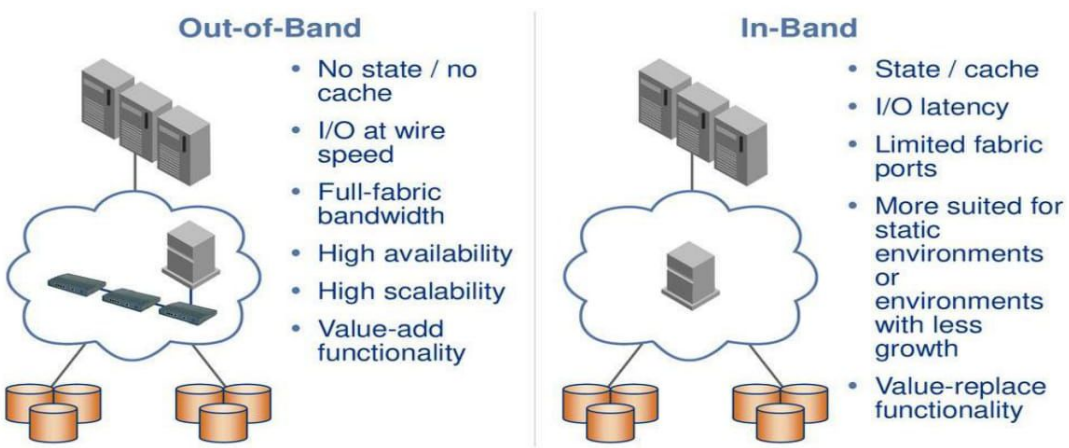
- Out-of band virtualization performance is exclusively dependent on the performance of the SAN switch and the end devices.
- Out-of band virtualization appliances may also require host resident drivers : to maintain the virtualization mapping generated by the appliance and to pass exceptions to the appliance for processing.

- Example: A 256 KB write operation to a striped virtual volume may need to be broken 4 separate 64KB write operations directed to 4 different storage targets.
- **Host-resident software** is required for the server to access the virtualized storage across the network.

Advantage:

- This configuration eliminates the potential bottle neck issue, since only small amount of metadata are exchanged between the appliances and servers.
- Out-of band virtualization appliance require NO significant changes to the SAN infrastructure.
- An out-of band appliance attaches to the SAN discovers the storage assets, configures storage pools and corresponding LUN's and distribute block address mapping metadata to the client servers.

Comparison of Virtualization Architectures



9.4 High availability for virtualization appliances

- Single instance of the virtualization appliance can't meet high availability requirements.
- Therefore implemented distributed appliances solution that enables the task of one appliance to be assumed by another.
- An **active/passive configuration** provides failover in the event that a primary appliance or path is down.

- An **active/active configuration** is preferable, since it utilizes both the appliances and can provide the load balancing in addition to failure.
- A fully redundant in-band solution duplicates all connections, switches and appliances.
- Servers are provided with redundant HBA and thus safeguarded in the event of adapter card failure.
- The server fabric is built with 2 switches each of which has its own links to each server and to the virtualization appliances.
- The storage fabric provide dual pathing between the storage array and redundant switches.
- Both the storage fabric and the server fabric are dual linked with each of the appliances.
- The failure of either appliance or any link to the server or storage will result in failover to the appropriate alternate path or appliance.

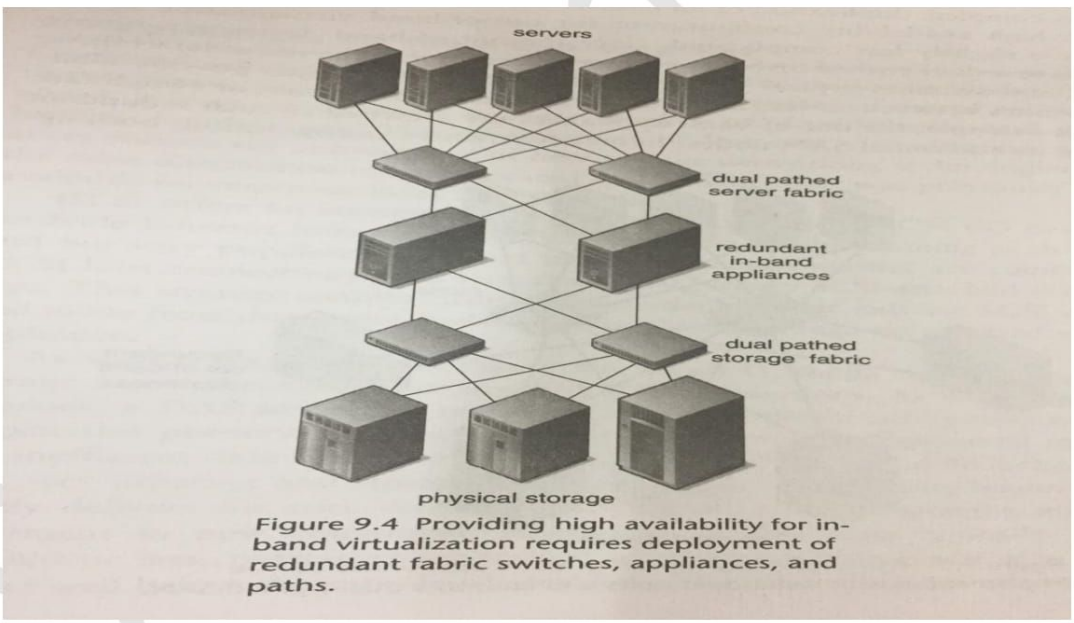


Figure 9.4 Providing high availability for in-band virtualization requires deployment of redundant fabric switches, appliances, and paths.

- It's less complicated compared to in band. The out-of band appliances are attached to each redundant switches.
- Both the in-band and out –of –band high availability configuration could be supplemented with the additional high availability SAN options such as server clustering, virtualization services such as data replications and snapshots.

- Each increment in high availability builds more complexity into the total installation and requires deliberate design and configuration

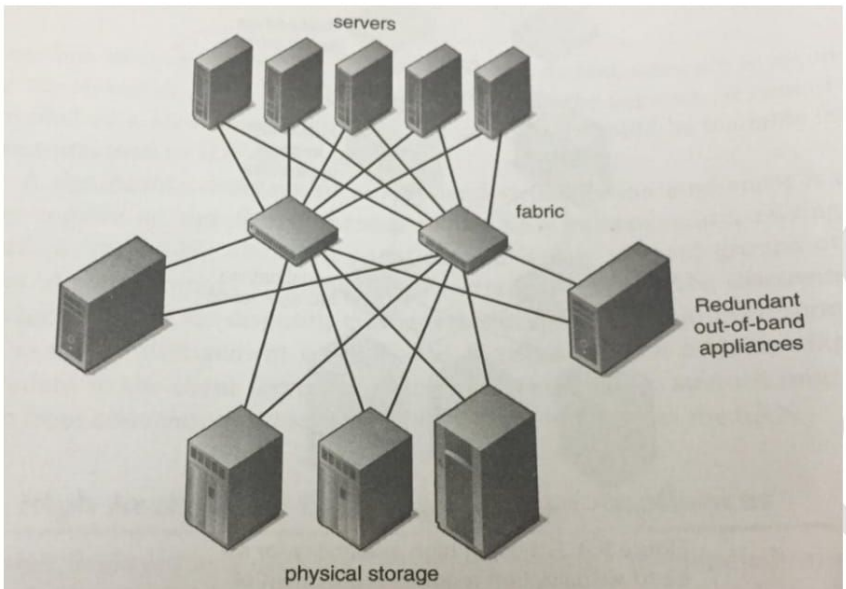


Figure 9.5 High availability for out-of-band appliances can be implemented with redundant units and links to a dual-pathed SAN.

9.5 Appliances for mass consumption

- The combination of iSCSI and virtualization technology is enabling low-cost but sophisticated shared storage solutions.

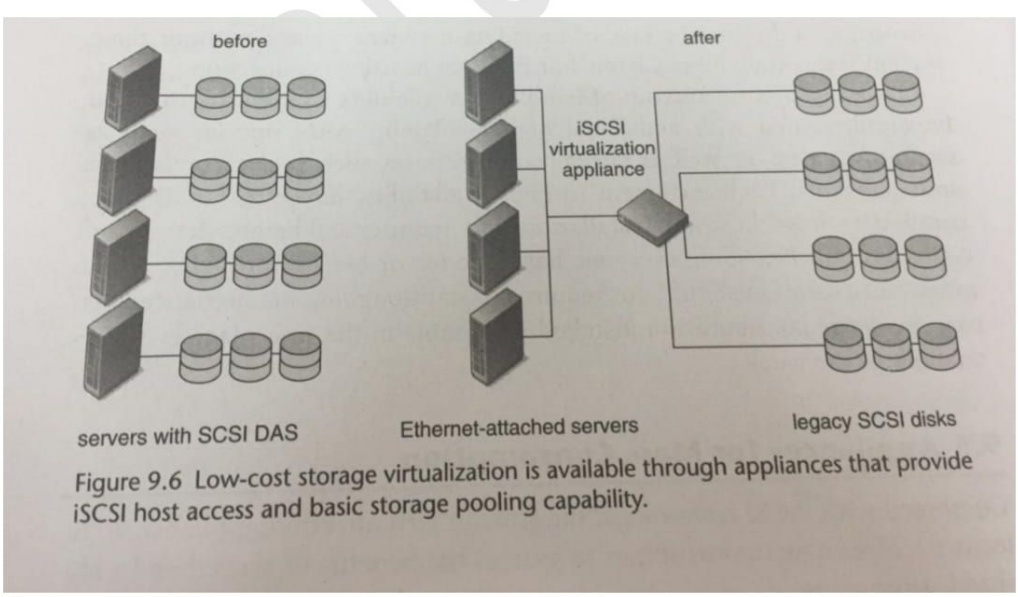


Figure 9.6 Low-cost storage virtualization is available through appliances that provide iSCSI host access and basic storage pooling capability.

- Economical iSCSI virtualization appliances may repurpose legacy direct-attached storage to provide virtualized shared storage.
- Low-cost iSCSI virtualization appliances provide a migration path from small to large shared storage networking.

Model Questions:

1. **What is cloud computing? Explain the benefits and characteristics of cloud computing?**
2. **Explain the various cloud service models available**
3. **Explain the cloud computing infrastructure in detail**
4. **Explain In-band and Out-of Band virtualization Appliances**

MODULE 4
Chapter 13:
STORAGE AUTOMATION AND VIRTUALIZATION

Policy-based management

Policy-based management is a broad category that includes a diversity of IT resources such as applications, computer platforms, networks, and storage.

- It is used for aligning underlying technology to business requirements
- **Policy-based management incorporates <u>three basic elements</u>:**

1. Measurement of actual behavior

2. Evaluation of that behavior against predefined *rules* or goals 3. Enforcement through behavior modification.

- **Goal of policy management initiatives:**
- **Automate IT operations on the basis of specific criteria that align with higher level business requirements**
- **Regulatory compliance example: archived customer information be secured and confidential and retrievable within 24 hours. Identify which transactions are candidates for special treatment, and then to enforce data handling that meets the desired requirements.**

Policy definition may be provided -by an upper layer management platform

- But **policy enforcement** requires a tight integration of management and the complex environment of compute resources such as network, and storage that supports data transactions.
- Ideally, this integration is provided by a common management interface that combines both management frameworks and a wide spectrum of infrastructure equipment.

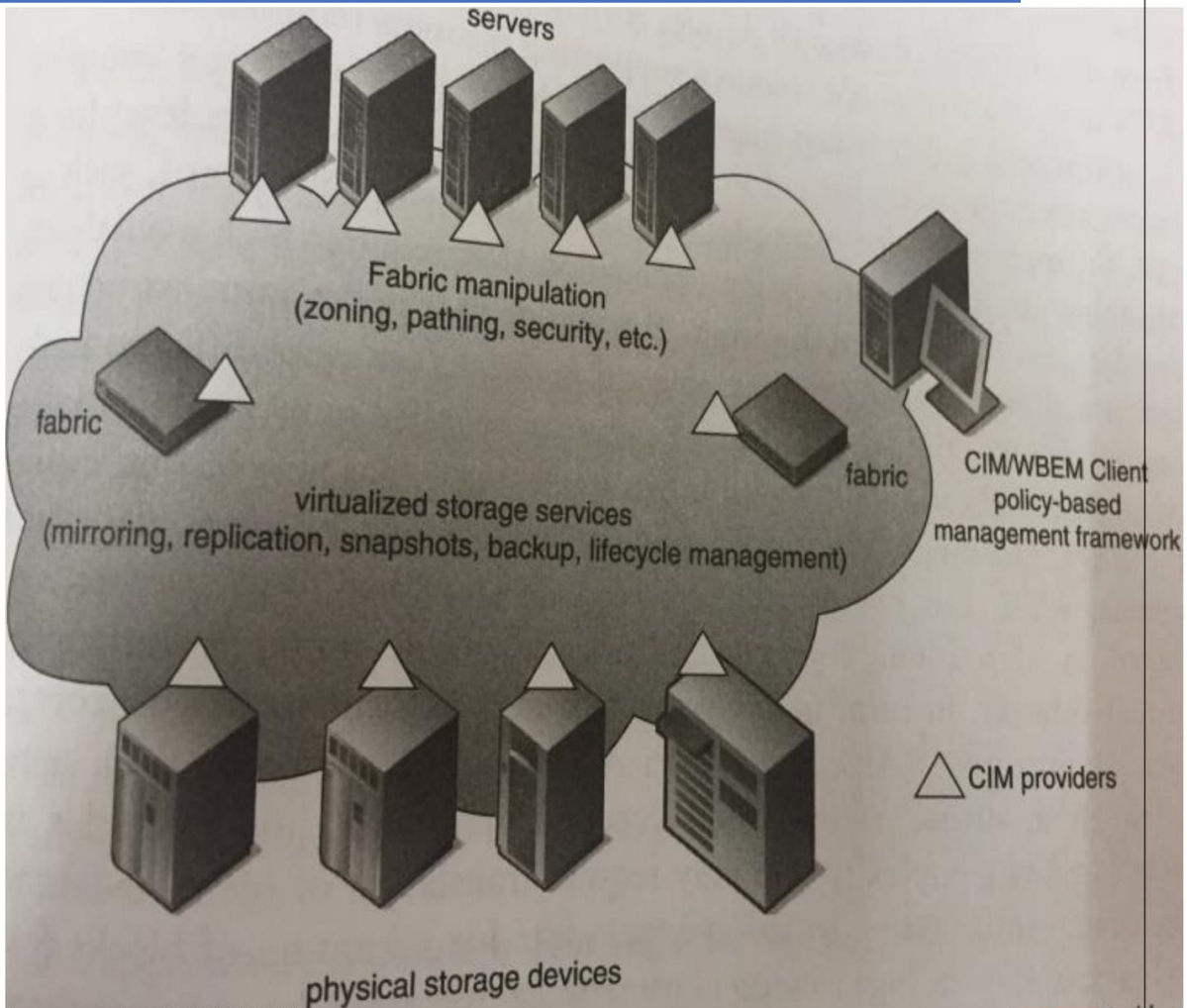
- The effort to define a common management interface for storage is being led by the Storage Networking Industry Association (SNIA).
- The SNIA Storage Management Interface Specification (SMI-S) is based on the common information model (CIM), which was originally developed by another industry group, the Desktop Management Task Force (DMTF).
- The SNIA storage management interface specification (SMI-S) establishes common management structure for heterogeneous SAN's.
- CIM : ***Defines management objects for a wide diversity of network and compute resource. It is managed through web based enterprise management (WBEM) protocol.***

CIM schema includes policy classes for automating IT processes, defining policies and policy execution.

- **Applying CIM/WBEM to storage HBAs, SAN switches, and storage devices requires:**
- **Creation of profiles with classes whose attributes reflect the unique capabilities of each type of product.**
- **A profile for a Fibre Channel switch, for example, may specify parameters for port statistics, device configuration, topology, regardless of manufacturer.**

- **SMI-S Common features**
  - **RAID definition and LUN management,**
  - **Storage virtualization through active management of storage pools**
  - **Mirroring and snapshots between storage systems.**
- The CIM Schema, for example, provides active management method calls such
- CreateOrModify Storage Pool( ) and CreateOrModifyElement From Storage Pool() to generate and resize virtual pools and virtual volumes from them.

- As shown in the diagram CIM/WBEM implementation requires that Physical devices such as HBA's, switches, storage arrays, and tape systems offer standards which are compliant with CIM providers
- Such that CIM can map vendor as specific features generic capabilities
- Example: Techniques for configuring RAID levels, may vary from vendor to vendor, but the CIM provider translates generic- RAID configuration instructions from a CIMWBEM management framework into appropriate commands for a particular vendor.



- Policy-based management must be flexible enough to accommodate changing business requirements as well as changes to the underlying infra-structure.
- Data archiving, for example, may be a phase of a particular policy for data handling.
- If the **archive infrastructure** is changed from disk-to-tape to disk -to-disk-to-tape, that change should be transparent to the upper layer application policy but recognized by lower layer policy objects that interface more directly with physical assets.

- **As shown in the diagram Policy-based management transforms the physical SAN into a collection of services supporting business application requirements.**

- **The physical connectivity of the SAN, pathing, zones, LUNs, and virtual volumes are <u>transparent to policy definition.</u>**

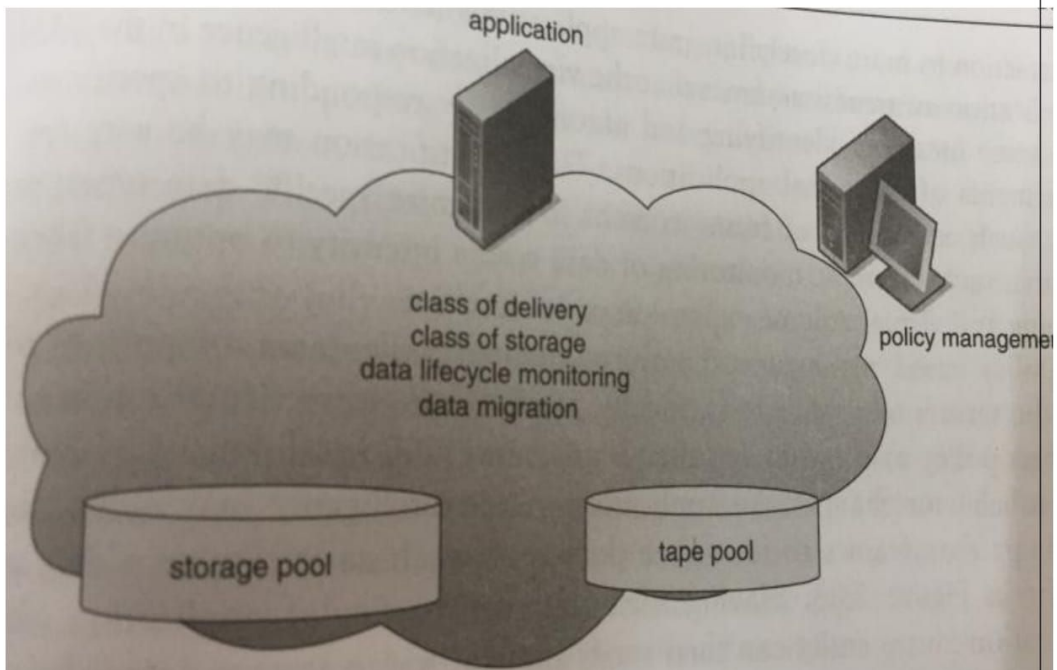


Figure 13.2 Policy-based management is predicated on a hierarchy of policy objec that govern different aspects of a virtualized infrastructure. Services under policy enforcement must be coordinated to accomplish the overall policy goal.

Application awareness

- **Application awareness** assumes that the virtualization intelligence in the SAN has some means of identifying and automatically responding to specific re-quirements of individual applications.
- That identification may be very spe-cific, such as
  - analysis of frame content to recognize specific data types,
- Or ***generic,*** such as simple monitoring of data traffic intensity to optimize fabric pathing and virtual volume expansion or contraction.
- Policy-based management provides a foundation for tighter integration of applications and storage virtualization.
- An application-aware virtualization entity must identify specific application data types and launch the appropriate policies for data handling.
- Application-aware virtualization must respond to changing application needs, such as capacity requirements and lifecycle management.

- Application aware virtualization automates policy association and then implements additional policies as application behavior changes.
- An application-aware intelligence may index into transport data frames to identify a data type, such as streaming video, as shown in Figure 13.3
  - Once the application is identified application aware entity can then verify that the video stream data is being serviced as per the defined policies.

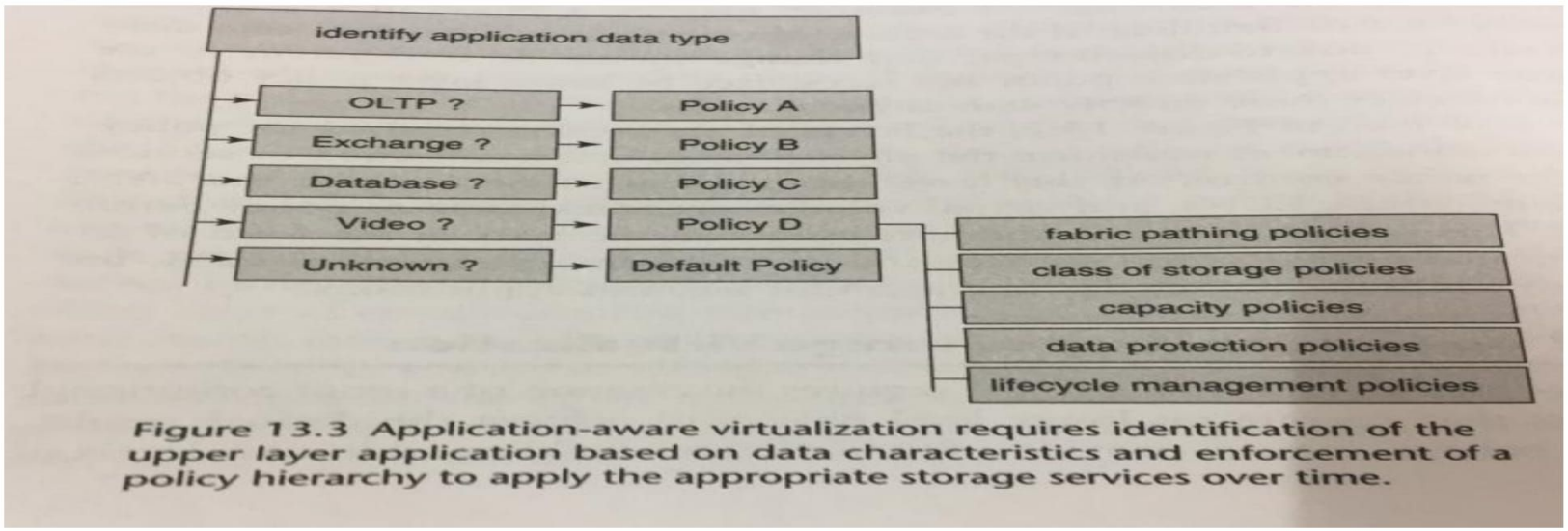


Figure 13.3 Application-aware virtualization requires identification of the upper layer application based on data characteristics and enforcement of a policy hierarchy to apply the appropriate storage services over time.

Virtualization awareness

- **Virtualization awareness within applications simplifies the task of linking applications to storage policy enforcement.**

- Virtualization APIs with an operating system enable upper layer applications to communicate requirements to storage virtualization entities in the SAN.
- Virtualization-awareness facilitates the integration of applications and infrastructure by enabling the application to define its own storage requirements.
- Virtualization-aware applications expand the scope of storage intelligence beyond the SAN .

- As shown in the figure the **storage services requirements such as**
  - **Availability,**
  - **Class of storage,**
  - **Archiving**
- **May be communicated to the SAN via configuration parameters**
- **Configuration parameters are loaded when the application establishes its connection to the SAN.**
- **Those parameters may be processed by the operating system, which in turn leverages the appropriate APIs to communicate with SAN-based virtualization entity.**

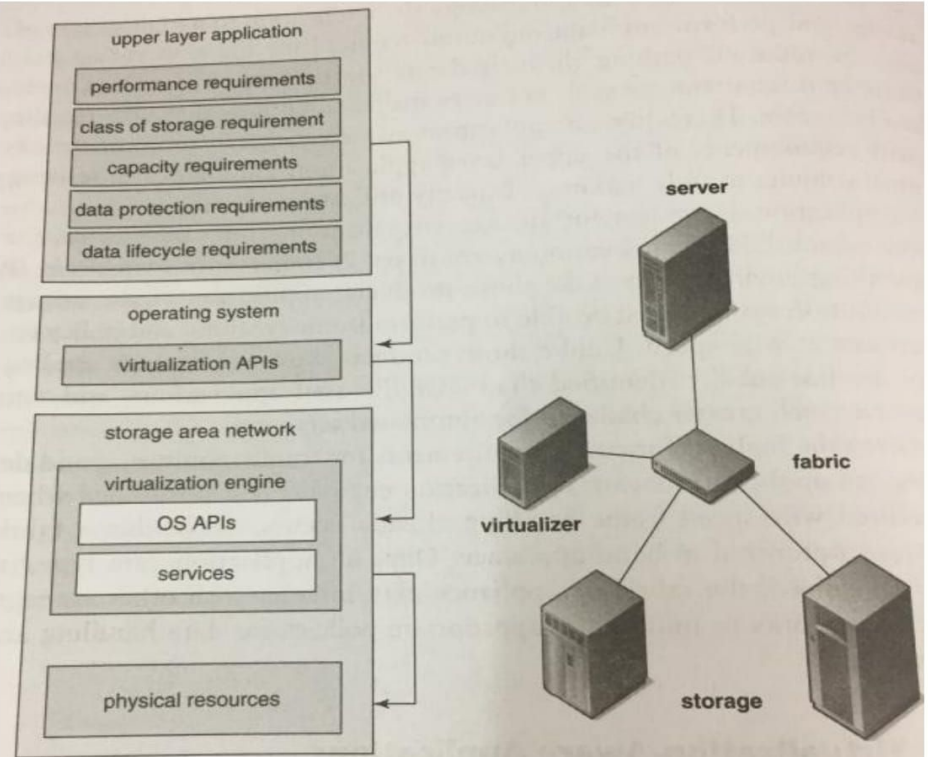


Figure 13.4 A virtualization-aware application may communicate its requirements to the operating system virtualization APIs, which in turn notifies the SAN-based virtualization engine. This configuration puts the application in control of its supporting storage infrastructure.

MODULE 5

Syllabus: Securing and Managing Storage Infrastructure Securing and Storage Infrastructure: Information Security Framework, Risk Triad, Storage Security Domains, Security Implementations in Storage Networking, Securing Storage Infrastructure in Virtualized and Cloud Environments. Managing the Storage Infrastructure Monitoring the Storage Infrastructure, Storage Infrastructure Management activities, Storage Infrastructure Management Challenges, Information Lifecycle management, Storage Tiering

Chapter 14: Securing and managing storage infrastructure

14.1 Information Security Framework

The basic information security framework is built to achieve four security goals: confidentiality, integrity, and availability (CIA), along with accountability.

Confidentiality:

- Provides the required secrecy of information and ensures that only authorized users have access to data. This requires authentication of users who need to access information.
- Data in transit (data transmitted over cables) and data at rest (data residing on a primary storage, backup media, or in the archives) can be encrypted to maintain its confidentiality.

Integrity:

- Ensures that the information is unaltered.
- Ensuring integrity requires detection of and protection against unauthorized alteration or deletion of information.
- Ensuring integrity stipulates measures such as error detection and correction for both data and systems.

Availability:

- This ensures that authorized users have reliable and timely access to systems, data, and applications residing on these systems.
- Availability requires protection against unauthorized deletion of data and denial of service .
- Availability also implies that sufficient resources are available to provide a service.

Accountability service:

- Refers to accounting for all the events and operations that take place in the data center infrastructure.
- The accountability service maintains a log of events that can be audited or traced later for the purpose of security.

Risk triad

- Raid defines risk in terms of threats, assets, and vulnerabilities.
- Risk arises when a threat agent (an attacker) uses an existing vulnerability to compromise the security services of an asset, for example, if a sensitive document is transmitted without any protection over an insecure channel, an attacker might get unauthorized access to the document and may violate its confidentiality and integrity.
- This may, in turn, result in business loss for the organization.
- In this scenario potential business loss is the risk, which arises because an attacker uses the vulnerability of the unprotected communication to access the document and tamper with it.
- To manage risks, organizations primarily focus on vulnerabilities because they cannot eliminate threat agents that appear in various forms and sources to its assets.
- Organizations can enforce countermeasures to reduce the possibility of occurrence of attacks and the severity of their impact.

14. 2.1 Assets

- Information is one of the most important *assets* for any organization.
- Other assets include hardware, software, and other infrastructure components required to access the information.
- To protect these assets, organizations must develop a set of parameters to ensure the availability of the resources to authorized users and trusted networks.
- These parameters apply to storage resources, network infrastructure, and organizational policies.
- **Security methods have two objectives:**
- The first objective is to ensure that the network is easily accessible to authorized users.
- It should also be reliable and stable under disparate environmental conditions and volumes of usage.
- The second objective is to make it difficult for potential attackers to access and compromise the system.

14.2.2 Threats

- Threats are the potential attacks that can be carried out on an IT infrastructure.
- These attacks can be classified as active or passive.
- ***Passive attacks*** are attempts to gain unauthorized access into the system.

- They pose threats to confidentiality of information.
- ***Active attacks*** include data modification, denial of service (DoS), and repudiation attacks.
- They pose threats to data integrity, availability, and accountability.
- In a **data modification attack**, the unauthorized user attempts to modify information for malicious purposes.
- A modification attack can target the data at rest or the data in transit. These attacks pose a threat to data integrity.

Denial of service **(DoS)** attacks prevent legitimate users from accessing resources and services.

- These attacks generally do not involve access to or modification of information.
- Instead, they pose a threat to data availability.
- The intentional flooding of a network or website to prevent legitimate access to authorized users is one example of a DoS attack.

Repudiation is an attack against the accountability of information.

- It attempts to provide false information by either impersonating someone or denying that an event or a transaction has taken place.
- For example, a repudiation attack may involve performing an action and eliminating any evidence that could prove the identity of the user (attacker) who performed that action.
- Repudiation attacks include circumventing the logging of security events or tampering with the security log to conceal the identity of the attacker.

Passive Attacks

Eavesdropping: When someone overhears a conversation, the unauthorized access to this information is called eavesdropping.

Snooping: This refers to accessing another user's data in an unauthorized way. In general, snooping and eavesdropping are synonymous.

14. 2. 3Vulnerability

- The paths that provide access to information are often vulnerable to potential attacks.
- Each of the paths may contain various access points, which provide different levels of access to the storage resources.
- It is important to implement adequate security controls at all the access points on an access path.
- Implementing security controls at each access point of every access path is known as *defense in depth*.

- Defense in depth recommends using multiple security measures to reduce the risk of security threats if one component of the protection is compromised.
- It is also known as a "layered approach to security."
- Because there are multiple measures for security at different levels, defense in depth gives additional time to detect and respond to an attack.
- *Attack surface*, *attack vector*, and *work factor* are the three factors to consider when assessing the extent to which an environment is vulnerable to security threats.
- ***Attack surface*** refers to the various entry points that an attacker can use to launch an attack.
- An ***attack vector*** is a step or a series of steps necessary to complete an attack vector example, an attacker might exploit a bug in the management interface to execute a snoop attack whereby the attacker can modify the configuration of the storage device to allow the traffic to be accessed from one more host.
- ***Work factor*** refers to the amount of time and effort required to exploit an attack vector.
- For example, if attackers attempt to retrieve sensitive information, they consider the time and effort that would be required for executing an attack on a database.
- Based on the roles they play, controls are categorized as preventive, detective, and corrective.
- The preventive control attempts to prevent an attack; the detective control detects whether an attack is in progress; and after an attack is discovered, the corrective controls are implemented.

14.3 Storage Security Domains

To identify the threats that apply to a storage network, access paths to data storage can be categorized into three security domains: *application access*, *management access*, and *backup, replication, and archive*. Figure 14-1 depicts the three security domains of a storage system environment.

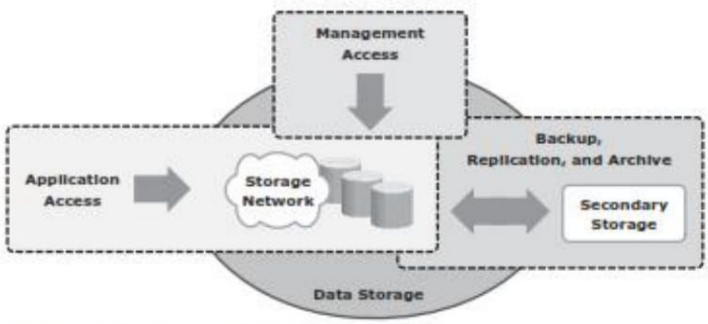


Figure 14-1: Storage security domains

- The first security domain involves application access to the stored data through the storage network.

- The second security domain includes management access to storage and interconnect devices and to the data residing on those devices.
- This domain is primarily accessed by storage administrators who configure and manage the environment.
- The third domain consists of backup, replication, and archive access. Along with the access points in this domain, the backup media also needs to be secured.

14.3.1 Securing the Application Access Domain

The *application access domain* may include only those applications that access the data through the file system or a database interface.

- An important step to secure the application access domain is to identify the threats in the environment and appropriate controls that should be applied.
- Implementing physical security is also an important consideration to prevent media theft.
- Figure 14-2 shows application access in a storage networking environment.

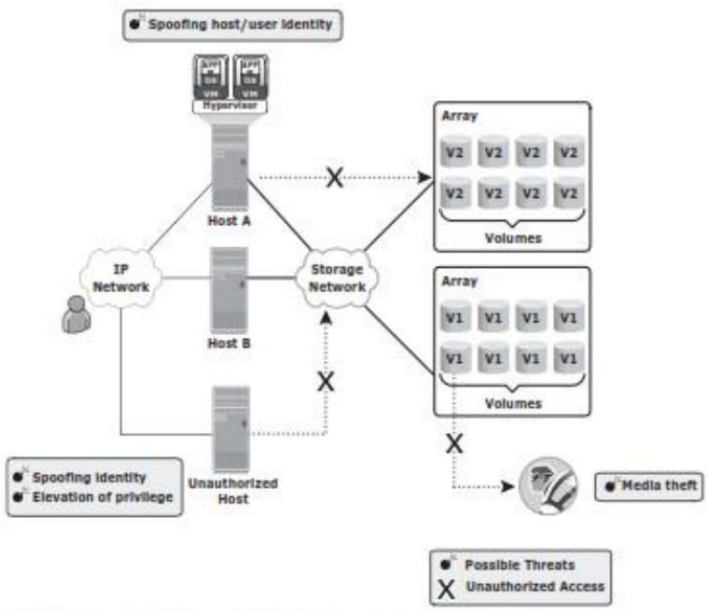


Figure 14-2: Security threats in an application access domain

- Host A can access all V1 volumes; host B can access all V2 volumes.
- These volumes are classified according to the access level, such as confidential, restricted, and public.
- Some of the possible threats in this scenario could be host A spoofing the identity or elevating to the privileges of host B to gain access to host B's resources.

- Another threat could be that an unauthorized host gains access the network; the attacker on this host may try to spoof the identity of another host and tamper with the data, snoop the network, or execute a DoS attack.
- Also any form of media theft could also compromise security. These threats can pose several serious challenges to the network security; therefore, they need to be addressed.

Controlling User Access to Data

- Access control mechanisms used in the application access domain are user and host authentication (technical control) and authorization (administrative control).
- These mechanisms may lie outside the boundaries of the storage network and require various systems to interconnect with other enterprise identity management and authentication systems, for example, systems that provide strong authentication and authorization to secure user identities against spoofing.
- NAS devices support the creation of *access control lists* that regulate user access to specific files.
- The Enterprise Content Management application enforces access to data by using Information Rights Management (IRM) that specifies which users have what rights to a document.
- Different storage networking technologies, such as iSCSI, FC, and IP-based storage, use various authentication mechanisms, such as Challenge-Handshake Authentication Protocol (CHAP), Fibre Channel Security Protocol (FC-SP), and IPSec, respectively, to authenticate host access.
- After a host has been authenticated, the next step is to specify security controls for the storage resources, such as ports, volumes, or storage pools, that the host is authorized to access.
- *Zoning* is a control mechanism on the switches that segments the network into specific paths to be used for data traffic;
- *LUN masking* determines which hosts can access which storage devices.

Protecting the Storage Infrastructure

- The security controls for protecting the network fall into two general categories: *network infrastructure integrity* and *storage network encryption*.

- Controls for ensuring the infrastructure integrity include a fabric switch function that ensures fabric integrity.
- This is achieved by preventing a host from being added to the SAN fabric without proper authorization.
- Storage network encryption methods include the use of IPSec for protecting IP-based storage networks, and FC-SP for protecting FC networks.
- In secure storage environments, root or administrator privileges for a specific device are not granted to every user.
- Instead, *role-based access control* (RBAC) is deployed to assign necessary privileges to users, enabling them to perform their roles
- It is also advisable to consider administrative controls, such as "separation of duties," when defining data center procedures. Clear separation of duties ensures that no single individual can both specify an action and carry it out.
- For example, the person who authorizes the creation of administrative accounts should not be the person who uses those accounts.

Data Encryption

- Data should be encrypted as close to its origin as possible.
- If it is not possible to perform encryption on the host device, an encryption appliance can be used for encrypting data at the point of entry into the storage network.
- Encryption devices can be implemented on the fabric that encrypts data between the host and the storage media.
- These mechanisms can protect both the data at rest on the destination device and data in transit.

14.3.2 Securing the Management Access Domain

- Implementing appropriate controls for securing storage management applications is important because the damage that can be caused by using these applications can be far more extensive.
- Figure 14-3 depicts a storage networking environment in which production hosts are connected to a SAN fabric and are accessing production storage array A, which is connected to remote storage array B for replication purposes.

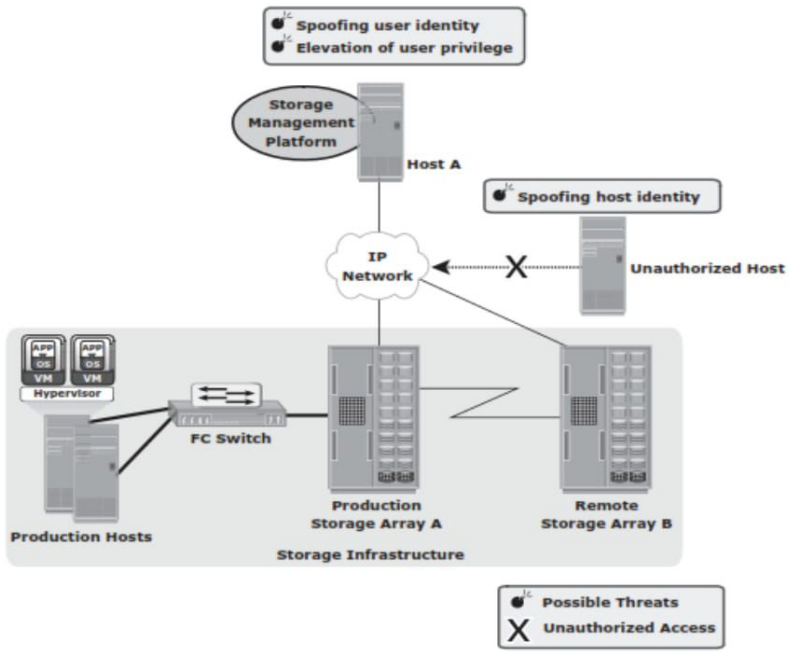


Figure 14-3: Security threats in a management access domain

- Further, this configuration has a storage management platform on Host A.
- A possible threat in this environment is an unauthorized host spoofing the user or host identity to manage the storage arrays or network.
- For example, an unauthorized host may gain management access to remote array B.
- Providing management access through an external network increases the potential for an unauthorized host or switch to connect to that network.
- Using secure communication channels, such as Secure Shell (SSH) or Secure Sockets Layer (SSL)/Transport Layer Security (TLS), provides effective protection against these threats. Event log monitoring helps to identify unauthorized access and unauthorized changes to the infrastructure.

Controlling Administrative Access

- Controlling administrative access to storage aims to safeguard against the threats of an attacker spoofing an administrator's identity or elevating privileges to gain administrative access.
- Both of these threats affect the integrity of data and devices.
- To protect against these threats, administrative access regulation and various auditing techniques are used to enforce accountability of users and processes.
- Access control should be enforced for each storage component.
- In some storage environments, it may be necessary to integrate storage devices with third-party authentication directories, such as Lightweight Directory Access Protocol (LDAP) or Active Directory.

Protecting the Management Infrastructure

- Mechanisms to protect the management network infrastructure include encrypting management traffic, enforcing management access controls, and applying IP network security best practices.
- These best practices include the use of IP routers and Ethernet switches to restrict the traffic to certain devices.
- Access controls need to be enforced at the storage-array level to specify which host has management access to which array.
- A separate private management network is highly recommended for management traffic.
- If possible, management traffic should not be mixed with either production data traffic or other LAN traffic used in the enterprise.
- Unused network services must be disabled on every device within the storage network.
- This decreases the attack surface for that device by minimizing the number of interfaces through which the device can be accessed.

14.3.3 Securing Backup, Replication, and Archive

BURA is the third domain that needs to be secured against attack. BURA is complex and is based on the BURA software accessing the storage arrays.

Organizations must ensure that the DR site maintains the same level of security for the backed up data. Protecting the BURA infrastructure requires addressing several threats, including spoofing the legitimate identity of a DR site, tampering with data, network snooping, DoS attacks, and media theft.

Figure 15-4 illustrates a generic remote backup design whereby data on a storage array is replicated over a disaster recovery (DR) network to a secondary storage at the DR site.

In a remote backup solution where the storage components are separated by a network, the threats at the transmission layer need to be countered.

Otherwise, an attacker can spoof the identity of the backup server and request the host to send its data. The unauthorized host claims to be the backup server at the DR site, which may lead to a remote backup being performed to an unauthorized and unknown site. In addition, attackers can use the connection to the DR network to tamper with data, snoop the network for authentication data, and create a DoS attack against the storage devices.

Security Implementations in Storage Networking

14.4.1 FC SAN

- Traditional FC SANs enjoy an inherent security advantage over IP-based networks.
- An FC SAN is configured as an isolated private environment with fewer nodes than an IP network. Consequently, FC SANs impose fewer security threats.
- Many FC SAN security mechanisms have evolved from their counterpart in IP networking, thereby bringing in matured security solutions.

FC SAN Security Architecture

- Storage networking environments are a potential target for unauthorized access, theft, and misuse because of the vastness and complexity of these environments.
- Therefore, security strategies are based on the *defense in depth* concept, which recommends multiple integrated layers of security.
- This ensures that the failure of one security control will not compromise the assets under protection.

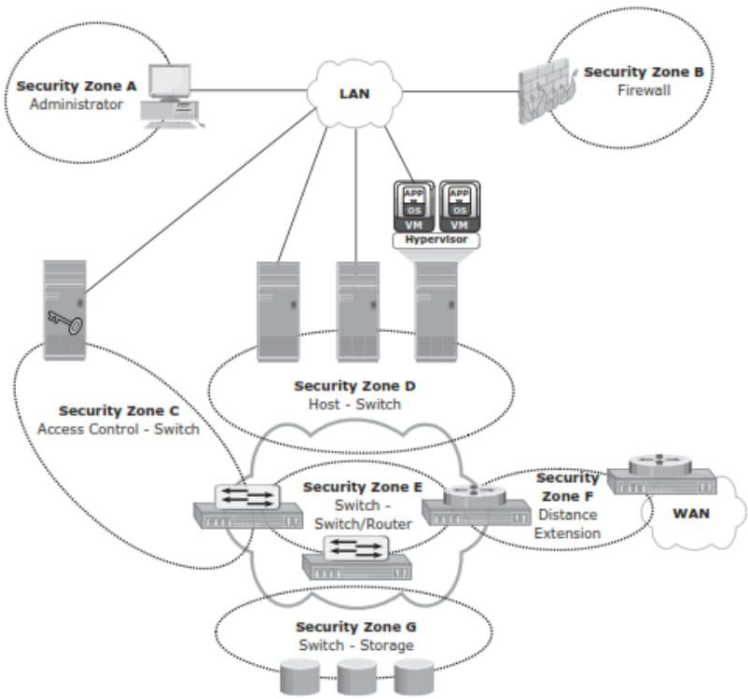


Figure 14-5: FC SAN security architecture

- Figure 14-5 illustrates various levels (zones) of a storage networking environment that must be secured and the security measures that can be deployed.

Table 14-1: Security Zones and Protection Strategies

SECURITY ZONES	PROTECTION STRATEGIES
Zone A (Authentication at the Management Console)	(a) Restrict management LAN access to authorized users (lock down MAC addresses); (b) implement VPN tunneling for secure remote access to the management LAN; and (c) use two-factor authentication for network access.
Zone B (Firewall)	Block inappropriate traffic by (a) filtering out addresses that should not be allowed on your LAN; and (b) screening for allowable protocols, block ports that are not in use.
Zone C (Access Control-Switch)	Authenticate users/administrators of FC switches using Remote Authentication Dial In User Service (RADIUS), DH-CHAP (Diffie-Hellman Challenge Handshake Authentication Protocol), and so on.
Zone D (Host to switch)	Restrict Fabric access to legitimate hosts by (a) implementing ACLs: Known HBAs can connect on specific switch ports only; and (b) implementing a secure zoning method, such as port zoning (also known as hard zoning).
Zone E (Switch to Switch/Switch to Router)	Protect traffic on fabric by (a) using E_Port authentication; (b) encrypting the traffic in transit; and (c) implementing FC switch controls and port controls.
Zone F (Distance Extension)	Implement encryption for in-flight data (a) FC-SP for long-distance FC extension; and (b) IPSec for SAN extension via FCIP.
Zone G (Switch to Storage)	Protect the storage arrays on your SAN via (a) WWPN-based LUN masking; and (b) S_ID locking: masking based on source FC address.

Basic SAN Security Mechanisms

Most commonly used SAN security methods.

1. LUN masking and zoning
2. switch-wide and fabric-wide access control
3. RBAC
4. logical partitioning of a fabric (Virtual SAN)

LUN Masking and Zoning

- LUN masking and zoning are the basic SAN security mechanisms used to protect against unauthorized access to storage.
- The standard implementations of LUN masking on storage arrays mask the LUNs presented to a frontend storage port based on the WWPNs of the source HBAs.
- A stronger variant of LUN masking may sometimes be offered whereby masking can be done on the basis of source FC addresses.
- It offers a mechanism to lock down the FC address of a given node port to its WWN.
- *WWPN zoning* is the preferred choice in security-conscious environments.
- *Hard zoning* or *port zoning* is the mechanism of choice in security-conscious environments. Unlike soft zoning or WWPN zoning, it actually filters frames to ensure that only authorized zone members can communicate.

- However, it lacks one significant advantage of WWPN zoning: The zoning configuration must change if the source or the target is relocated across ports in the fabric. There is a trade-off between ease of management and the security provided by WWPN zoning and port zoning.

Securing Switch Ports

- Apart from zoning and LUN masking, additional security mechanisms, such as port binding, port lockdown, port lockout, and persistent port disable, can be implemented on switch ports.
- *Port binding* limits the number of devices that can attach to a particular switch port and allows only the corresponding switch port to connect to a node for fabric access.
- Port binding mitigates but does not eliminate WWPN spoofing.
- *Port lockdown* and *port lockout* restrict a switch port's type of initialization.
- Typical variants of port lockout ensure that the switch port cannot function as an E_Port and cannot be used to create an ISL, such as a rogue switch.
- *Persistent port disable* prevents a switch port from being enabled even after a switch reboot.

Switch-Wide and Fabric-Wide Access Control

- Network security can be configured on the FC switch by using *access control lists* (ACLs) and on the fabric by using fabric binding.
- **ACLs** incorporate the device connection control and switch connection control policies.
- The device connection control policy specifies which HBAs and storage ports can be a part of the fabric, preventing unauthorized devices from accessing it.
- Similarly, the switch connection control policy specifies which switches are allowed to be part of the fabric, preventing unauthorized switches from joining it.
- ***Fabric binding*** prevents an unauthorized switch from joining any existing switch in the fabric.
- It ensures that authorized membership data exists on every switch and any attempt to connect any switch in the fabric by using an ISL causes the fabric to segment.

Logical Partitioning of a Fabric: Virtual SAN

- VSANs enable the creation of multiple logical SANs over a common physical SAN.
- They provide the capability to build larger consolidated fabrics and still maintain the required security and isolation between them.
- Figure 14-6 depicts logical partitioning in a VSAN.

- The SAN administrator can create distinct VSANs by populating each of them with switch port
- In the example, the switch ports are distributed over two VSANs: 10 and 20 — for the Engineering and HR divisions, respectively.
- Although they share physical switching gear with other divisions, they can be managed individually as standalone fabrics.
- Zoning should be done for each VSAN to secure the entire physical SAN. Each managed VSAN can have only one active zone set at a time.

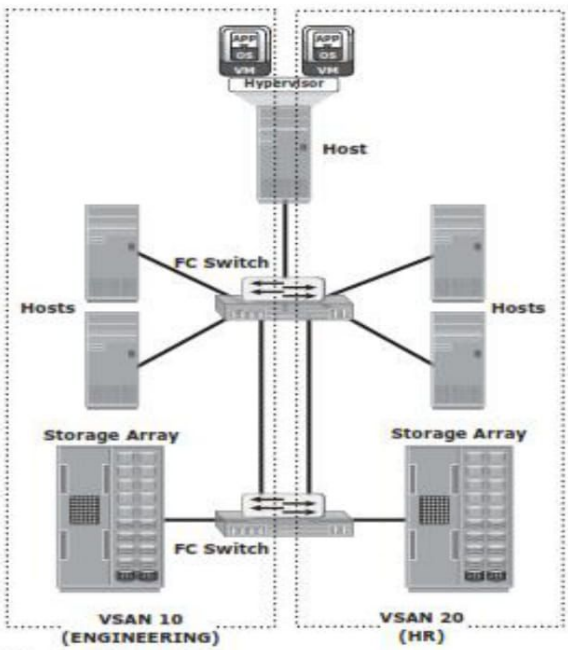


Figure 14-6: Securing SAN with VSAN

NAS

NAS is open to multiple exploits, including viruses, worms, unauthorized access, snooping, and data tampering. Various security mechanisms are implemented in NAS to secure data and the storage networking infrastructure.

NAS File Sharing: Windows ACLs

- Windows supports two types of ACLs: *discretionary access control lists (DACLs)* and *system access control lists (SACLs)*. The DACL, commonly referred to as the ACL, is used to determine access control.
- The SACL determines what accesses need to be audited if auditing is enabled. In

addition to these ACLs, Windows also supports the concept of object ownership.

- The owner of an object has hard-coded rights to that object, and these rights do not have to be explicitly granted in the SACL.

NAS File Sharing: UNIX Permissions

- For the UNIX operating system, a *user* is an abstraction that denotes a logical entity for assignment of ownership and operation privileges for the system.
- A user can be either a person or a system operation. UNIX permissions specify the operations that can be performed by any ownership relation with respect to a file.
- In simpler terms, these permissions specify what the owner can do, what the owner group can do, and what everyone else can do with the file.

Authentication and Authorization

- In a file-sharing environment, NAS devices use standard file-sharing protocols, NFS and CIFS. Authentication requires verifying the identity of a network user and therefore involves a login credential lookup on a Network Information System (NIS) server in a UNIX environment. Similarly, a Windows client is authenticated by a Windows domain controller that houses the Active Directory.
- The Active Directory uses LDAP to access information about network objects in the directory, and Kerberos for network security.
- NAS devices use the same authentication techniques to validate network user credentials. Figure 15-7 depicts the authentication process in a NAS environment.
- Authorization defines user privileges in a network. The authorization techniques for UNIX users and Windows users are quite different.
- UNIX files use mode bits to define access rights granted to owners, groups, and other users, whereas Windows uses an ACL to allow or deny specific rights to a particular user for a particular file.

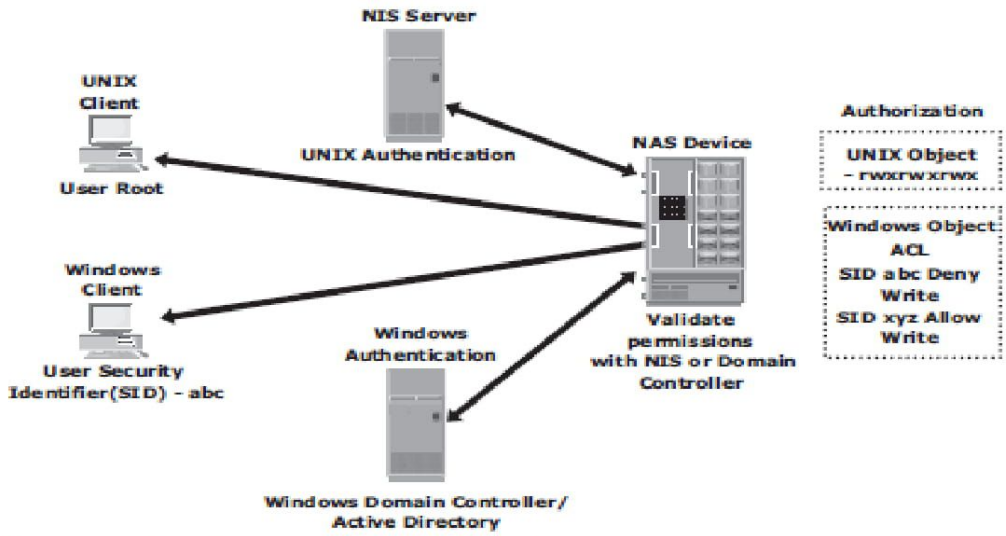


Figure 14-7: Securing user access in a NAS environment

Kerberos

- Kerberos is a network authentication protocol. It is designed to provide strong authentication for client/server applications by using secret-key cryptography.
- The term *Kerberos server* generally refers to the Key Distribution Center (KDC).
- The KDC implements the Authentication Service (AS) and the Ticket Granting Service (TGS). The KDC has a copy of every password associated with every principal, so it is absolutely vital that the KDC remain secure.
- The Kerberos authorization process shown in Figure 15-8 includes the following steps:

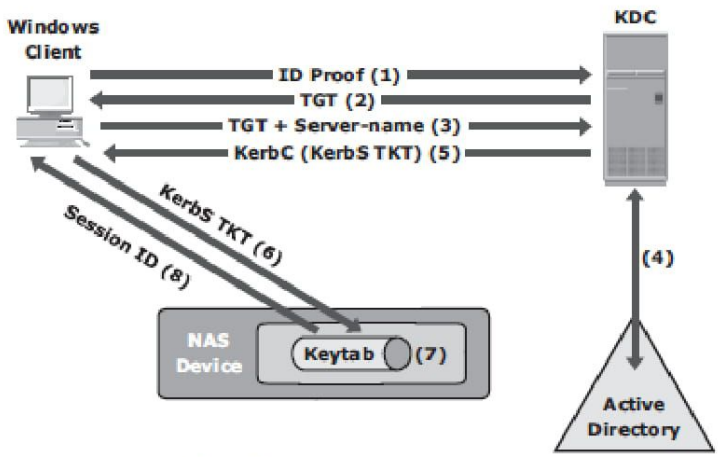


Figure 14-8: Kerberos authorization

1 The user logs on to the workstation in the Active Directory domain (or forest) using an ID

and a password. The client computer sends a request to the AS running on the KDC for a Kerberos ticket. The KDC verifies the user's login information from Active Directory. (Note that this step is not explicitly shown in Figure 15-8.)

2. The KDC responds with a TGT (TKT is a key used for identification and has limited validity period). It contains two parts, one decryptable by the client and the other by the KDC.

3. When the client requests a service from a server, it sends a request, consist of the previously generated TGT and the resource information, to the KDC.

4. The KDC checks the permissions in Active Directory and ensures that the user is authorized to use that service.

5. The KDC returns a service ticket to the client. This service ticket contains fields addressed to the client and to the server that is hosting the service.

6. The client then sends the service ticket to the server that houses the desired resources.

7. The server, in this case the NAS device, decrypts the server portion of the ticket and stores the information in a key tab file. As long as the client's Kerberos ticket is valid, this authorization process does not need to be repeated. The server automatically allows the client to access the appropriate resources.

8. A client/server session is now established. The server returns a session ID to the client, which is used to track client activity, such as file locking, as long as the session is active.

Network-Layer Fire walls

- These network- layer firewalls are capable of examining network packets and comparing them to a set of configured security rules. Packets that are not authorized by a security rule are dropped and not allowed to continue to the requested destination.
- Rules can be established based on a source address (network or host), a destination address (network or host), a port, or a combination of those factors (source IP, destination IP, and port number). The effectiveness of a firewall depends on how robust and extensive the security rules are.
- Figure 15-9 depicts a typical firewall implementation. Demilitarized zone (DMZ) is commonly used in networking environments.
- A DMZ provides a means of securing internal assets while allowing Internet-based access to various resources. In a DMZ environment, servers that need to be accessed through the

Internet are placed between two sets of firewalls. Application-specific ports, such as HTTP or FTP, are allowed through the firewall to the DMZ servers.

- However, no Internet-based traffic is allowed to penetrate the second set of firewalls and gain access to the internal network. The servers in the DMZ may or may not be allowed to communicate with internal resources.
- In such a setup, the server in the DMZ is an Internet-facing Web application that is accessing data stored on a NAS device, which may be located on the internal private network.

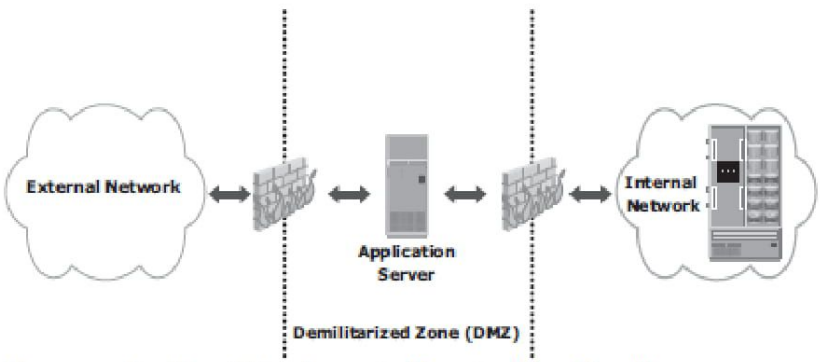


Figure 14-9: Securing a NAS environment with a network-layer firewall

IP SAN

- The Challenge-Handshake Authentication Protocol (CHAP) is a basic authentication mechanism that has been widely adopted by network devices and hosts.
- CHAP provides a method for initiators and targets to authenticate each other by utilizing a secret code or password. CHAP secrets are usually random secrets of 12 to 128 characters.
- The secret is never exchanged directly over the wire; rather, a one-way hash function converts it into a hash value, which is then exchanged.
- A hash function, using the MD5 algorithm, transforms data in such a way that the result is unique and cannot be changed back to its original form. Figure 15-10 depicts the CHAP authentication process.
- If the initiator requires reverse CHAP authentication, the initiator authenticates the target by using the same procedure.
- The CHAP secret must be configured on the initiator and the target. A CHAP entry, comprising the name of a node and the secret associated with the node, is maintained by the target and the initiator.

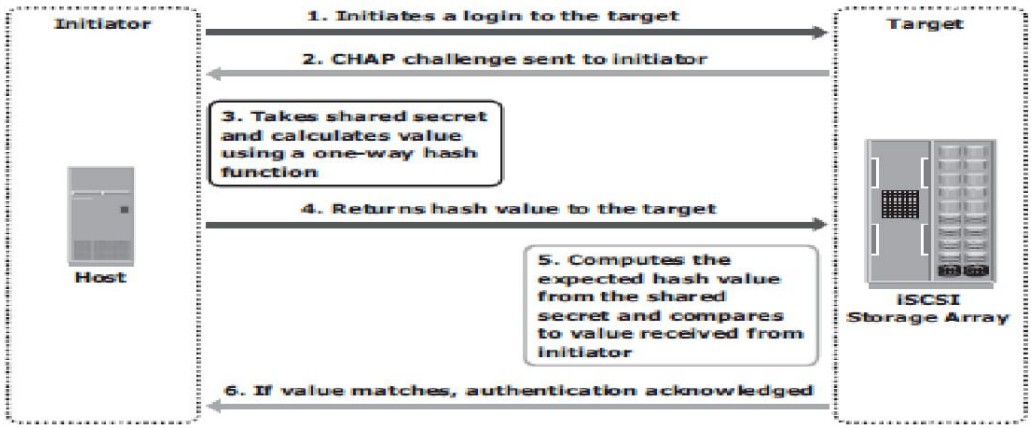


Figure 14-10: Securing IPSAN with CHAP authentication

- The same steps are executed in a two-way CHAP authentication scenario. After these steps are completed, the initiator authenticates the target.
- If both authentication steps succeed, then data access is allowed. CHAP is often used because it is a fairly simple protocol to implement and can be implemented across a number of disparate systems. iSNS discovery domains function in the same way as FC zones.
- Discovery domains provide functional groupings of devices in an IP-SAN. In order for devices to communicate with one another, they must be configured in the same discovery domain.
- State change notifications (SCNs) tell the iSNS server when devices are added or removed from a discovery domain. Figure 15-11 depicts the discovery domains in iSNS.

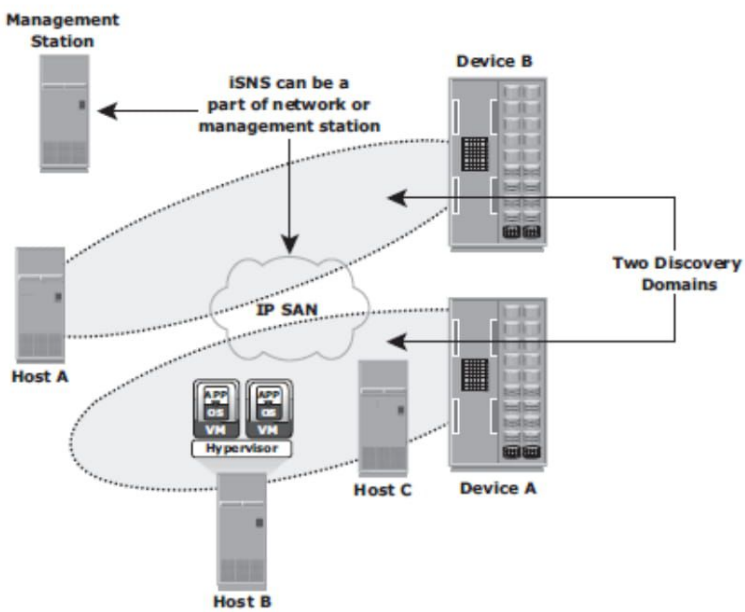


Figure 14-11: Securing IPSAN with iSNS discovery domains

Securing Storage Infrastructure in Virtualized and Cloud Environments

- These environments have additional threats due to multitenancy and lack of control over the cloud resources
- Virtualization-specific security concerns are common for all cloud models
- In public clouds, there are additional security concerns, which demand specific countermeasures
- Clients have less control to enforce security measures in public clouds
- Difficult for cloud service provider(CSP) to meet the security needs of all the clients

Security Concerns

- Multitenancy: Enables multiple independent tenants to be serviced using the same set of storage resources.
- Business critical data of one tenant is accessed by other competing tenants who run applications using the same resources.
- Velocity-of-attack: existing security threat in the cloud spreads more rapidly and has a larger impact than that in the traditional data center environments
- Information assurance for users ensures confidentiality, integrity, and availability of data in the cloud
- Data privacy is also a major concern in a virtualized and cloud environment.

Security Measures

- **Security at the computer Level:**
  - Enforce the security of physical server, VMs, and hypervisor
  - Physical server: user authentication and authorization mechanism
  - Hypervisor: security-critical hypervisor updates should be installed regularly
- **Security at the network level:**
  - To minimize vulnerabilities at the network layer: firewall, intrusion detection, DMZ, encryption.
  - Virtual firewall
  - Provides packet filtering and monitoring of the VM-to-VM traffic
  - DMZ and data encryption
- **Security at the storage Level:**
  - Adequate security measures at computer & network levels helps to ensure storage security
  - Access control: to regulate which users and processes access data on storage system
  - Data encryption: encrypt backup, store encryption keys separately from data
  - Use separate LUNs for VM configuration files and VM data
  - Segregate VM traffic from management traffic

Chapter 2

Managing the Storage Infrastructure

Monitoring the Storage Infrastructure

- Monitoring provides the performance and accessibility status of various components. It also enables administrators to perform essential management activities.
- Monitoring also helps to analyze the utilization and consumption of various storage infrastructure resources. This analysis facilitates capacity planning, forecasting, and optimal use of these resources.
- The major storage infrastructure components that should be monitored
  - Servers, databases and applications
  - Network (SAN and IP)
  - Storage arrays
- Each of these components should be monitored for accessibility, capacity, performance and security.

Monitoring Parameters

- **Accessibility** refers to the availability of a component to perform a desired operation. A component is said to be accessible when it is functioning without any fault at any given point in time.
- **Monitoring hardware components** (e.g., a SAN interconnect device, a port, an HBA, or a disk drive) or **software components** (e.g.,a database instance) for accessibility involves checking their availability status by listening to pre-determined alerts from devices. For example, a port may go down resulting in a chain of availability alerts.
- A storage infrastructure uses redundant components to avoid a single point of failure. Failure of a component may cause an outage that affects application availability, or it may cause serious performance degradation even though accessibility is not compromised. For example, an HBA failure can restrict the server to a few paths for access to data devices in a multipath environment, potentially resulting in degraded performance.
- Continuously monitoring for expected accessibility of each component and reporting any deviations helps the administrator to identify failing components and plan corrective

action to maintain SLA requirements.

- **Capacity** refers to the amount of storage infrastructure resources available. Examples of capacity monitoring include
  - Examining the free space available on a file system or a RAID group
  - The mailbox quota allocated to users
  - The numbers of ports available on a switch.
- Inadequate capacity may lead to degraded performance or affect accessibility or even application/service availability.
- Capacity monitoring ensures uninterrupted data availability and scalability by averting outages before they occur. For example, if a report indicates that 90 percent of the ports are utilized in a particular SAN fabric, a new switch should be added if more arrays and servers need to be installed on the same fabric.
- Capacity monitoring is preventive and predictive, usually leveraged with advanced analytical tools for trend analysis. These trends help to understand emerging challenges, and can provide an estimation of time needed to meet them.
- **Performance monitoring** evaluates how efficiently different storage infrastructure components are performing and helps to identify bottlenecks.
- Performance monitoring usually measures and analyzes behavior in terms of response time or the ability to perform at a certain predefined level.
- It also deals with utilization of resources, which affects the way resources behave and respond.
- Performance measurement is a complex task that involves assessing various components on several interrelated parameters. The number of I/Os to disks, application response time, network utilization, and server CPU utilization are examples of performance monitoring.
- **Security monitoring** helps to tracks unauthorized configuration changes of storage infrastructure elements. For example, security monitoring tracks and reports the initial zoning configuration performed and all subsequent changes.
- Physical security of a storage infrastructure is also continuously monitored using badge readers, biometric scans, or video cameras.

Components Monitored

Hosts, networks, and storage are the components within the storage environment that should be monitored for accessibility, capacity, performance, and security.

Hosts

- **Accessibility**
  - Hardware components: HBA, NIC, graphic card, internal disk
  - For example, an application crash due to host hardware failure can cause instant unavailability of the data to the user. Servers are used in a cluster to ensure high availability
  - Status of various processes/applications
- **Capacity**
  - File system utilization
  - For example capacity monitoring helps in estimating the file system's growth rate and predicting when it will reach 100 percent. Accordingly, the administrator can extend (manually or automatically) the file system's space proactively to prevent a failure resulting from a file system being full.
  - Database: Table space/log space utilization
  - User quota
- **Performance**
  - CPU and memory utilization
  - For example, if a server running an application is experiencing 80 percent of CPU utilization server may be running out of processing power, which can lead to degraded performance and slower response time. Administrators can take upgrading or adding more processors, shifting the workload to different servers, and restricting the number of simultaneous client access.
  - Memory utilization is measured by the amount of free memory available. Insufficient memory leads to excessive swapping and paging on the disk, which in turn affects response time to the applications.
  - Transaction response times
- **Security**
  - Login and authorization
  - For example, an administrator can block access to an unauthorized user if multiple login failures are logged.
  - Physical security (Data center access)

Storage Network

Uninterrupted access to data over the storage network depends on accessibility of the physical and logical components in storage network. Physical components of a storage network include elements such as switches, ports, cables, GBICs, and power supplies. The logical components include constructs, such as zones and fabrics. Any failure in the physical or logical components may cause data unavailability.

- **Accessibility**
  - Fabric errors, zoning errors, GBIC failure

- Device status/attribute Change
- Processor cards, fans, power supplies

- **Capacity**
  - ISL and port utilization
  - Availability of ports on a switch, number of available ports in the entire fabric, utilization of ISLs, individual ports, and each interconnect device in the fabric
- **Performance**
  - Connectivity ports
    - Link failures, Loss of signal, Link utilization
  - Connectivity devices
    - Port statistics
- Measuring receive or transmit link utilization metrics, which indicate how busy the switch port is, based on expected maximum throughput. Heavily used ports can cause queuing delays on the server.
- For IP networks, monitoring performance includes monitoring network latency, packet loss, bandwidth utilization for I/O, network errors, and collisions.
- **Security**
- Storage network security monitoring provides information for any unauthorized change to the configuration of the fabric—for example, changes to the zone policies that can affect data security. Login failures and unauthorized access to switches for performing administrative changes should be logged and monitored continuously
  - Zoning and LUN Masking
  - Administrative Tasks and physical security
  - Authorized Access, Strict Passwords

Storage

- **Accessibility**
  - All Hardware components
  - For example, the failure of a replication task affects disaster recovery capabilities. Some storage arrays also provide the capability to send a message to the vendor's support center in the event of hardware or process failures, referred to as a *call home*.
  - Array Operating Environment
    - RAID processes
    - Environmental Sensors
    - Replication processes
- **Capacity**
- Capacity monitoring of a storage array enables the administrator to respond to storage needs as they occur. Information about fan-in or fan-out ratios and the availability of front-end ports is useful when a new server is given access to the storage array.
  - Configured/un-configured capacity
  - Allocated/unallocated storage
  - Fan-in/fan-out ratios
- **Performance**
  - Performance metrics, such as utilization rates of the various storage array components, I/O response time, and cache utilization. A high utilization rate of

storage array components may lead to performance degradation.
    - FE and BE utilization/throughput
    - I/O profile, response time, cache metrics
- Security
    - Monitoring security helps to track unauthorized configuration of the storage array or corruption of data and ensures that only authorized users are allowed to access it.
    - Physical and administrative security

Monitoring Examples

A storage infrastructure requires implementation of an end-to-end solution to actively monitor all the parameters of its components.

Accessibility Monitoring

Failure of any component may affect the accessibility of one or more components due to their interconnections and dependencies, or it may lead to overall performance degradation.

Consider an implementation in a storage infrastructure with three servers: H1, H2, and H3. All the servers are configured with two HBAs, each connected to the storage array through two switches, SW1 and SW2, as shown in below figure. The three servers share two storage ports on the storage array. Path failover software has been installed on all three servers.

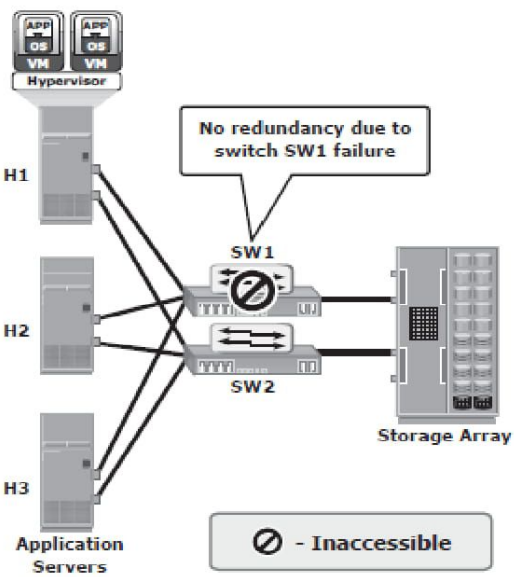


Figure 15-1: Switch failure in a storage infrastructure

If one of the *storage array ports* fails, all the storage volumes that were accessed through the switch connected to that port may become unavailable, depending on the type of storage array. If the storage volume becomes unavailable, path failover software initiates a path failover. However, due to redundant ports, the servers continue to access data through another switch, SW2. The servers H1, H2, and H3 may experience degraded performance due to an increased load on the path through SW2.

In the same example, if a single HBA fails on server H1, the server experiences path failure as shown in figure. However, due to redundant HBAs, H1 can still access the storage device but it may experience degraded application response time (depends on I/O load).

Capacity Monitoring

In the figure shown below the servers is allocated storage on the storage array. When a new server is deployed in this configuration, the applications on the new servers have to be given access to the storage devices from the array through switches SW1 and SW2.

Monitoring the available capacity on the array helps to proactively decide whether the array can provide the required storage to the new server. Other considerations include the availability of ports on SW1 and SW2 to connect to the new server as well as the availability of storage ports to connect to the switches.

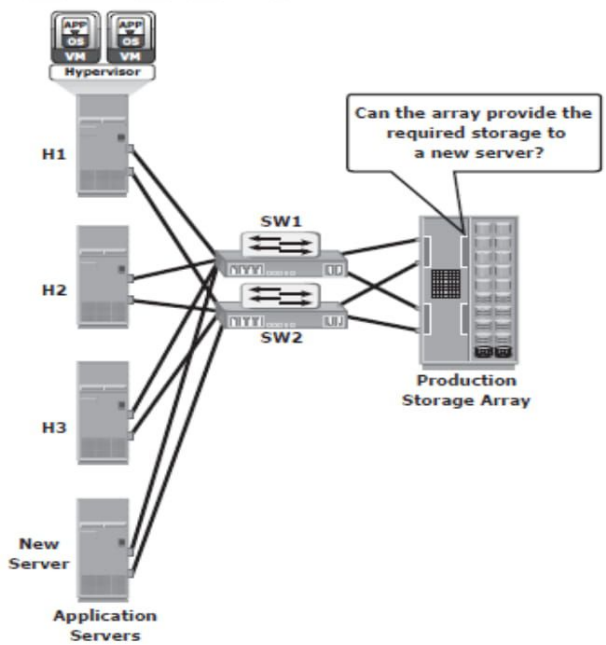


Figure 15-2: Monitoring storage array capacity

The following example illustrates the importance of monitoring file system capacity on servers. If file system capacity monitoring is not implemented, as shown in figure, and the file system is full, the application most likely will not function properly.

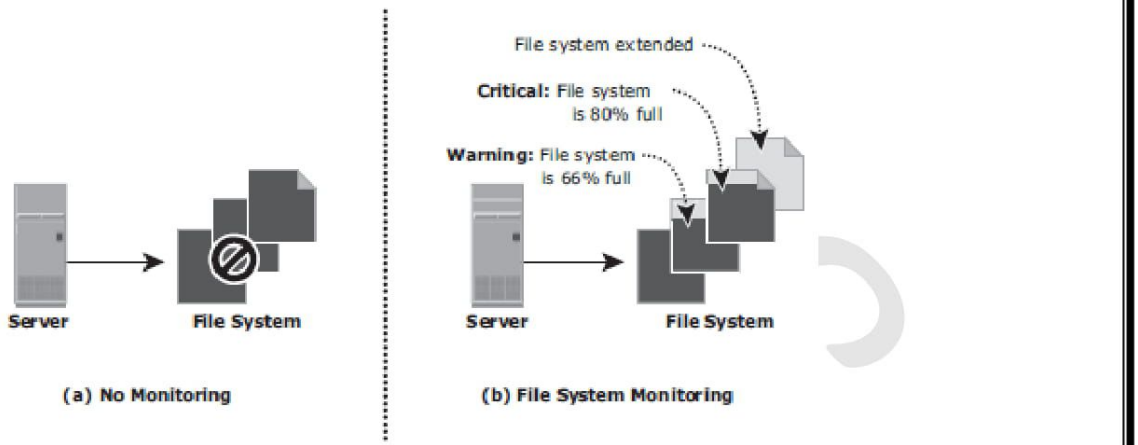


Figure 15-3: Monitoring server file system space

Monitoring can be configured to issue a message when thresholds are reached on file system capacity. For example, when the file system reaches 66 percent of its capacity a warning message is issued, and a critical message when the file system reaches 80 percent of its capacity.

Performance Monitoring

In the example shown below, servers H1, H2, and H3 (with two HBAs each) are connected to the storage array through switches SW1 and SW2. The three servers share the same storage ports on the storage array.

A new server H4 running an application with high work load, has to be deployed to share the same storage ports as H1, H2, and H3.

Monitoring array port utilization ensures that the new server does not adversely affect the performance of the other servers. In this example, utilization for the shared ports is shown by the solid and dotted lines in the line graph for the storage ports.

Notice that the port represented by a solid line is close to 100 percent utilization. If the actual utilization of both ports prior to deploying the new server is closer to the dotted line, there is room to add the new server. Otherwise, deploying the new server will affect the performance of all servers.

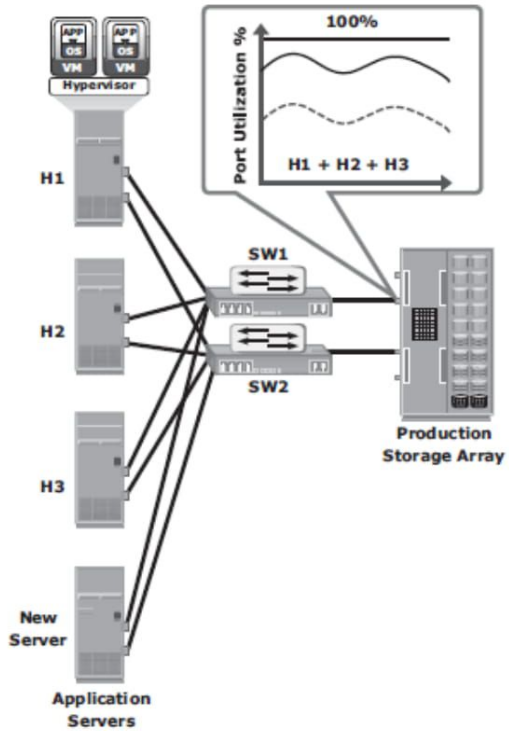


Figure 15-4: Monitoring array port utilization

Most servers offer tools that enable interactive monitoring of server CPU usage. For example, Windows Task Manager displays CPU and memory usage, as shown in slide. These interactive tools are useful only when a few servers need to be managed.

A storage infrastructure requires performance monitoring tools that are capable of monitoring many servers simultaneously.

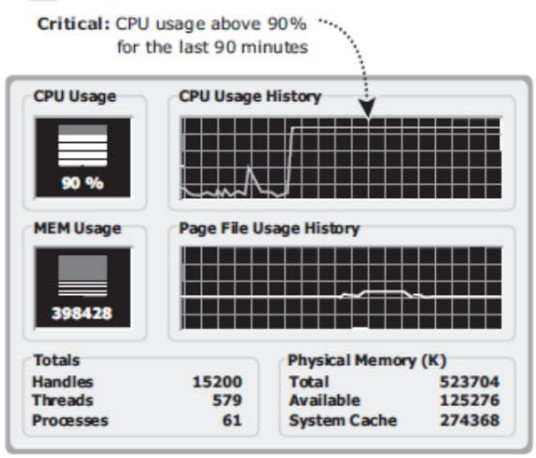


Figure 15-5: Monitoring the CPU and memory usage of a server

Security Monitoring

In this example shown below, the storage array is shared between two workgroups, WG1 and WG2. The data of WG1 should not be accessible by WG2. Likewise, WG2 should not be accessible by WG1. A user from WG1 may try to make a local replica of the data that belongs to WG2.

However, if this action is not monitored or recorded, it is difficult to track such a violation of security protocols.

Conversely, if this action is monitored, a warning message can be sent to prompt a corrective action or at least enable discovery as part of regular auditing operations.

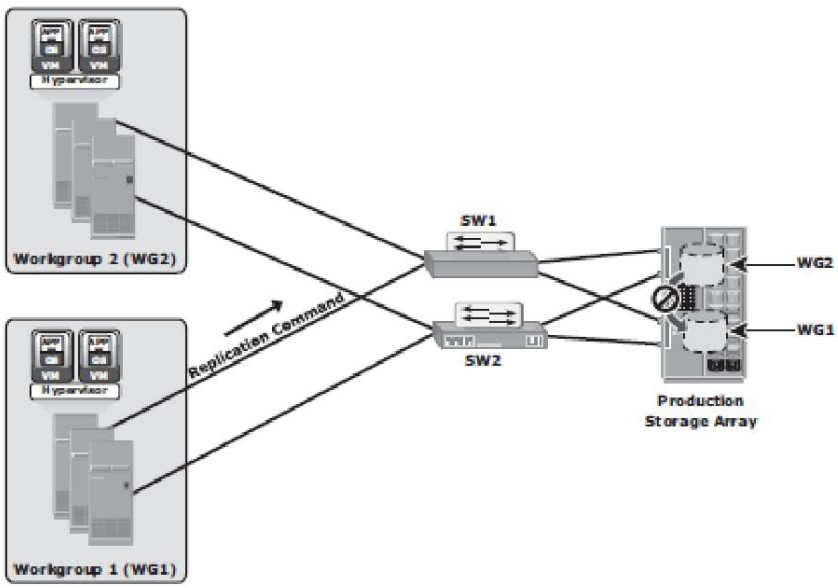


Figure 15-6: Monitoring security in a storage array

Alerts

- Alerting is an integral part of monitoring
- Monitoring tools enables administrators to assign different severity levels for different events
- Level of alerts based on severity
    - **Information alert**: Provide useful information and may not require administrator intervention
        - Creation of zone or LUN
    - **Warning alerts**: Require administrative attention
        - File systems becoming full/Soft media errors
    - **Fatal alert**: Require immediate administrative attention
        - Power failures/Disk failures/Memory failures/Switch failures

Storage Infrastructure Management Activities

The key storage infrastructure management activities performed in a data center can be broadly categorized into availability management, capacity management, performance management, security management, and reporting.

Availability Management

- Establishing guidelines for all configurations based on service levels
- To ensure high availability by:
    - Eliminating single points of failure deploy/configure
        - Deploying Two or more HBAs
        - Multipathing software with path failover capability, and server clustering
        - RAID protection
        - Redundant Fabrics
    - Configuring data backup and replication
    - Deploying virtualized environment

Capacity Management

- Ensures adequate availability of resources based on their service level requirements
- Capacity management provides capacity analysis, comparing allocated storage to forecasted storage on a regular basis.
- Manages resource allocation
- Key activities
    - Trend and Capacity analysis: actual utilization of allocated storage and rate of consumption
    - Storage provisioning
    - Examples
        - Host: Host configuration and file system/DB management
        - SAN: Unused Ports and Zoning
        - Storage: Device configuration and LUN Masking

Performance Management

- Configure/design for optimal operational efficiency
- Helps to identify the performance of storage infrastructure components. This analysis provides the information — whether a component is meeting expected performance

levels. Several performance management activities are initiated for the deployment of an application or server in the existing storage infrastructure.

- Performance analysis
    - Identify bottlenecks
    - Fine tuning for performance enhancement
- Key activities
    - Host: Volume management, database/application layout
    - SAN: Designing sufficient ISLs with adequate bandwidth
    - Storage Array: Choice of RAID type and layout of devices (LUNs) and choice of front-end ports

Security Management

- Prevent unauthorized activities or access
- For example, while deploying an application or a server, the security management tasks include managing user accounts and access policies, that authorizes users to perform role-based activities.
- Key activities
    - Server:
        - Creation of user logins, user privileges
    - SAN:
        - Configuration of zoning to restrict unauthorized HBA's
    - Storage Array:
        - LUN masking prevents data corruption on the storage array by restricting host access to a defined set of logical devices

Reporting

- Reporting on a storage infrastructure involves keeping track and gathering information from various components/processes
- This information is compiled to generate reports for trend analysis, capacity planning, chargeback, performance, and to illustrate basic configuration of storage infrastructure components
- Also used to provide information for Capacity, Availability, Security and Performance Management

Storage Management Examples

Example 1: Storage Allocation to a New Server/Host

Consider a deployment of the new RDBMS server to the existing non-virtualized SAN environment. Following are the management activities carried out to allocate storage to new host.

Storage array management activities: The administrator needs to configure new volumes on the array then assign those volumes to the array front-end ports. Administrator also need to configure LUN masking on the storage array by assigning new servers and volumes to the storage group.

Server management activities: The installation and configuration of the HBA hardware (at least two to ensure redundancy) and driver has to be performed on the server before it can be physically connected to the SAN. Server reconfiguration may be required, depending on the operating system installed on the server, so it can recognize the new devices. Optional multipathing software can be installed on the server, which might require additional configuration. The volume management tasks on the host involve the creation of volume groups, logical volumes, and file systems. On the application side, whether it is a database or any other type of application, administrator tasks include installation of the database or the application on the logical volumes or file systems that were created. Figure 15-7 illustrates the activities performed on a server, a SAN, and a storage array for the allocation of storage to a new server.

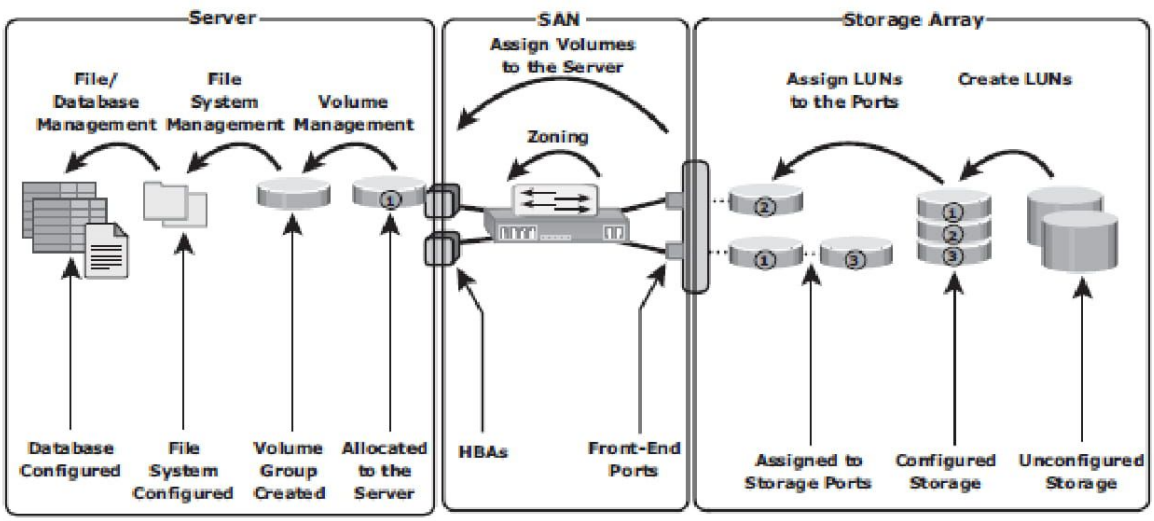


Figure 15-7: Storage allocation tasks

SAN management activities: The administrator configures the fabric's zoning policies for the new server's HBA, allowing the host to access the storage array port via the specific HBA port. This operation should probably be done at two or more fabrics to ensure redundant paths between the hosts and the storage array. The switches should have free ports available for the new server, and the array port utilization is validated against the required I/O performance of the server if the port is shared between many servers.

Example 2: File System Space Management

To prevent a fi le system from running out of space, administrators need to perform tasks to offload data from the existing fi le system. This includes deleting unwanted files or archiving data that is not accessed for a long time. Alternatively, an administrator can extend the file system to increase its size and avoid an application outage. The dynamic extension of file systems or a logical volume depends on the operating system or the logical volume manager (LVM) in use. Figure 15-8 shows the steps and considerations for the extension of file systems in the flow chart.

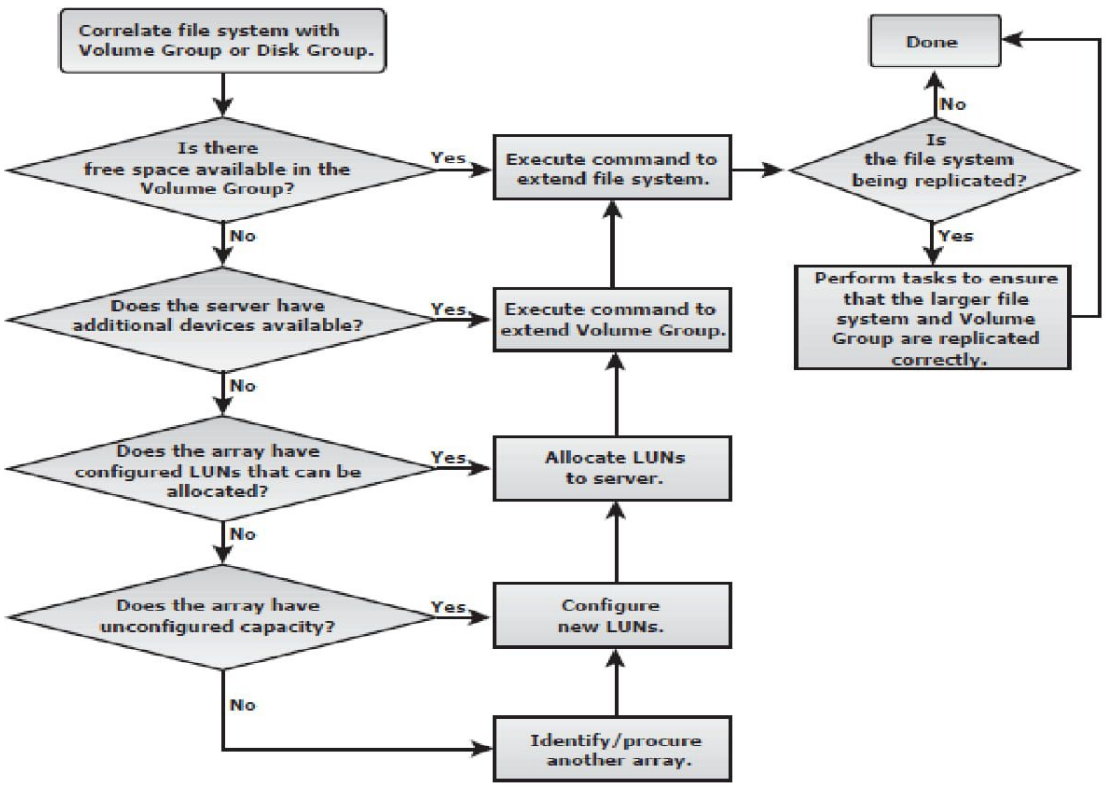


Figure 15-8: Extending a file system

Storage Infrastructure Management Challenges

Monitoring and managing today's complex storage infrastructure environment has become very challenging due to the number and variety of storage arrays, networks, servers, databases, and applications.

- Variety of storage devices varying in capacity, performance, and protection methodologies
- Storage infrastructures deploy both SAN and IP networks
- Servers with different operating systems such as UNIX, LINUX, Windows, or mainframe

These products and services from multiple vendors may have interoperability issues which add complexity in managing storage infrastructure. All of these components are provided with vendor-specific tools to manage and monitor them.

Information Lifecycle Management

The key challenges that exist in today's data centers:

- **Exploding digital universe**: The rate of information growth is increasing exponentially. Creating copies of data to ensure high availability and repurposing has contributed to the multifold increase of information growth.
- **Increasing dependency on information**: The strategic use of information plays an important role in determining the success of a business and provides competitive advantages in the marketplace.
- **Changing value of information:** Information that is valuable today might become less important tomorrow. The value of information often changes over time.

When information is first created, it often has the highest value and is accessed frequently. As the information ages, it is accessed less frequently and is of less value to the organization.

- For example, in a sales order application, the value of the information (customer data) changes from the time the order is placed until the time that the warranty becomes void (see Figure 15-11).
- High value → new sales order and process to deliver product
- Medium or low value → after order fulfillment → data can transfer to less expensive secondary storage
- No value → warranty becomes void → can dispose the data

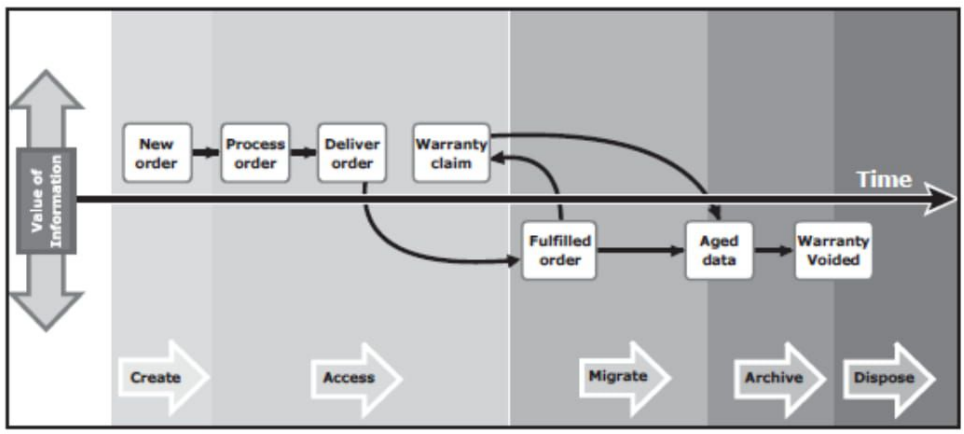


Figure 15-11: Changing value of sales order information

Information Lifecycle Management (ILM) is a proactive strategy that enables an IT organization to effectively manage information throughout its life cycle based on predefined business policies. From data creation to data deletion, ILM aligns the business requirements and processes with service levels in an automated fashion. This allows an IT organization to optimize the storage infrastructure for maximum return on investment.

Implementing an ILM strategy has the following key benefits that directly address the challenges of information management:

- **Lower Total Cost of Ownership (TCO):** By aligning the infrastructure and management costs with information value. As a result, resources are not wasted, and complexity is not introduced by managing low-value data at the expense of high-value data.
- **Simplified management**: By integrating process steps and interfaces with individual tools and by increasing automation
- **Maintaining compliance**: By knowing what data needs to be protected for what length of time
- **Optimized utilization:** By deploying storage tiering

Storage Tiering

It is a technique of establishing a hierarchy of storage types and identifying the candidate data to relocate to the appropriate storage type to meet service level requirements at a minimal cost.

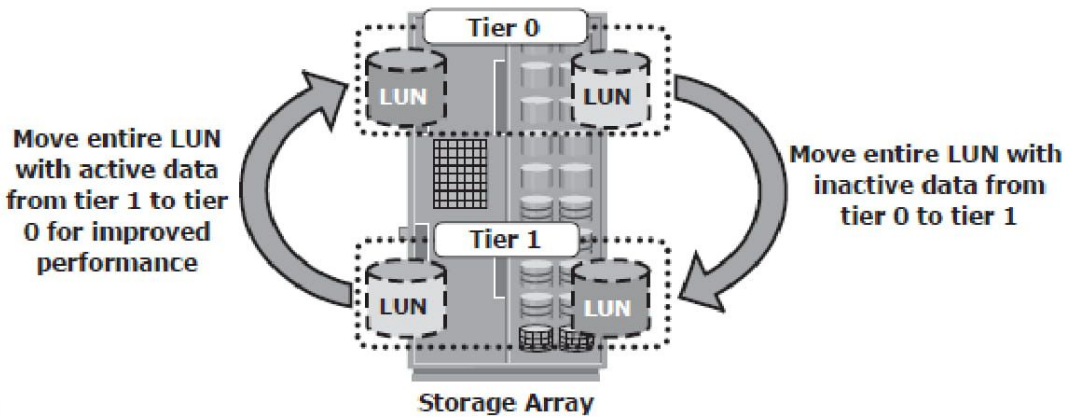
- Each tier has different levels of protection, performance, and cost
- Efficient storage tiering requires defining tiering policies

- Storage tiering implementations are:
  - Manual storage tiering
  - Automated storage tiering
- Data movement occurs between tiers
  - Within a storage array (Intra-array)
  - Between storage arrays (Inter-array)

Intra-array Storage Tiering

The process of storage tiering within a storage array is called *intra-array storage tiering*. It enables the efficient use of SSD, FC, and SATA drives within an array and provides performance and cost optimization. The goal is to keep the SSDs busy by storing the most frequently accessed data on them, while moving out the less frequently accessed data to the SATA drives.

- LUN tiering
  - Moves entire LUN from one tier to another
  - Does not give effective cost and performance benefits



(a) LUN Tiering

- Sub-LUN tiering
  - A LUN is broken down into smaller segments and tiered at that level
  - Provides effective cost and performance benefits

Inter-array Storage Tiering

The process of storage tiering between storage arrays is called inter-array storage tiering. Inter-array storage tiering automates the identification of active or inactive data to relocate them to different performance or capacity tiers between the arrays. Figure 15-14 illustrates an example of

a two-tiered storage environment. This environment optimizes the primary storage for performance and the secondary storage for capacity and cost.

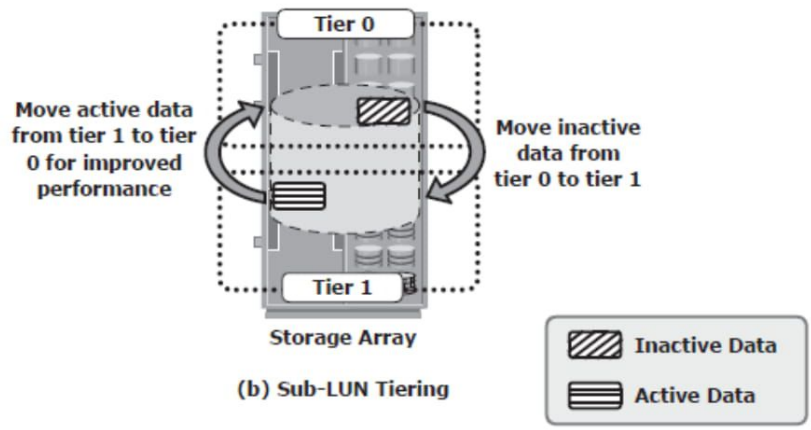


Figure 15-12: Implementation of intra-array storage tiering

Cache Tiering

- Enables creation of a large capacity secondary cache using SSDs
- Enables tiering between DRAM cache and SSDs (secondary cache)
- Most reads are served directly from high performance tiered cache

Benefits

- Enhances performance during peak workload
- Non-disruptive and transparent to applications

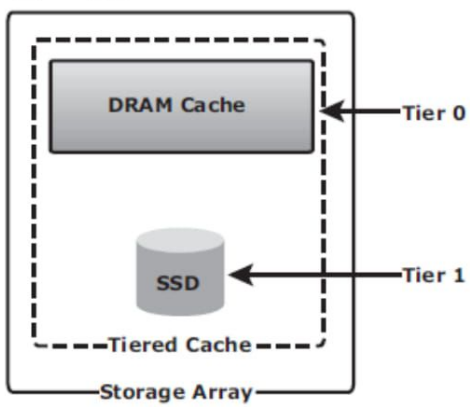


Figure 15-13: Cache tiering

********** END**********