# Module I

# 1

# INTRODUCTION TO INFORMATION STORAGE

**Unit Structure**





## 1.0 OBJECTIVES

1. Differentiate between data and information and its processing
2. Interpret core elements of Data Centre
3. Evaluate different storage architectures
4. Understand classic, virtualized and cloud environments
5. Comprehend Virtualization and Cloud Computing

## 1.1 INTRODUCTION

Information is ever more important in our daily lives nowadays. We have become information dependent in this 21st century, living in an

on-command, on-demand world, which means, we need information as and when it is required. We use the Internet every day for surfing, participating in social networks, sending and receiving emails, sharing pictures and videos, and using other applications. Equipped with a growing number of content-generating devices, more information is created by individuals than by organizations (including business, governments, nonprofits and so on). Information created by individuals gains value when shared with others. When created, information resides locally on devices, such as cell phones, smart phones, tablets, cameras, and laptops. To be shared, this information needs to be uploaded to central data repositories (data centres) via networks. Although most of the information is created by individuals, it is stored and managed by a relatively small number of organizations. The importance, dependency, and volume of information for the business world also continue to grow at astounding rates. Businesses depend on fast and reliable access to information critical to their success. Examples of business processes or systems that rely on digital information include airline reservations, telecommunications billing, Internet commerce, electronic banking, credit card transaction processing, capital/stock trading, health care claims processing, life science research, and so on. The increasing dependence of businesses on information has amplified the challenges in storing, protecting, and managing data. Legal, regulatory, and contractual obligations regarding the availability and protection of data further add to these challenges. Organizations usually maintain one or more data centers to store and manage information. A data center is a facility that contains information storage and other physical information technology (IT) resources for computing, networking, and storing information. In traditional data centers, the storage resources are typically dedicated for each of the business units or applications. The proliferation of new applications and increasing data growth have resulted in islands of discrete information storage infrastructures in these data centers. This leads to complex information management and underutilization of storage resources. Virtualization optimizes resource utilization and eases resource management. Organizations incorporate virtualization in their data centers to transform them into virtualized data centers (VDCs). Cloud computing, which represents a fundamental shift in how IT is built, managed, and provided, further reduces information storage and management complexity and IT resource provisioning time. Cloud computing brings in a fully automated request-fulfilment process that enables users to rapidly obtain storage and other IT resources on demand. Through cloud computing, an organization can rapidly deploy applications where the underlying storage capability can scale-up and scale-down, based on the business requirements.

In this chapter, we will see the evolution of information storage architecture from a server-centric model to an information-centric model. We will also see an overview of virtualization and cloud computing.

## 1.2 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.

## 1.3 DATA

Data is a collection of raw facts from which conclusions might be drawn. Handwritten letters, a printed book, a family photograph, printed and duly signed copies of mortgage papers, a bank's ledgers, and an airline ticket are all examples that contain data. Before the advent of computers, the methods adopted for data creation and sharing were limited to fewer forms, such as paper and film. Today, the same data can be converted into more convenient forms, such as an e-mail message, an e-book, a digital image, or a digital movie. This data can be generated using a computer and stored as strings of binary numbers (0s and 1s), as shown in Figure 1.1. Data in this form is called digital data and is accessible by the user only after a computer processes it.

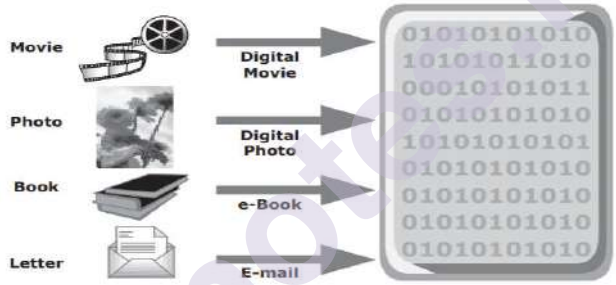



***Figure 1.1: Digital Data***

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

1. **Increase in data-processing capabilities**: Modern computers provide a significant increase in processing and storage capabilities. This enables the conversion of various types of content and media from conventional forms to digital formats.
2. **Lower cost of digital storage**: Technological advances and the decrease in the cost of storage devices have provided low-cost storage solutions. This cost benefit has increased the rate at which digital data is generated and stored.
3. **Affordable and faster communication technology**: The rate of sharing digital data is now much faster than traditional approaches. A handwritten letter might take a week to reach its destination, whereas it typically takes only a few seconds for an e-mail message to reach its recipient.

4. **Proliferation of applications and smart devices**: Smart phones, tablets, and newer digital devices, along with smart applications, have significantly contributed to the generation of digital content

   Inexpensive and easier ways to create, collect, and store all types of data, coupled with increasing individual and business needs, have led to accelerated data growth, popularly termed data explosion. Both individuals and businesses have contributed in varied proportions to this data explosion.

The importance and value of data vary with time. Most of the data created holds significance for a short term but becomes less valuable over time. This governs the type of data storage solutions used. Typically, recent data which has higher usage is stored on faster and more expensive storage. As it ages, it may be moved to slower, less expensive but reliable storage.

Following are some examples of research and business data:

1. **Customer data**: Data related to a company's customers, such as order details, shipping addresses, and purchase history.
2. **Product data**: Includes data related to various aspects of a product, such as inventory, description, pricing, availability, and sales.
3. **Medical data**: Data related to the healthcare industry, such as patient history, radiological images, details of medication and other treatment, and insurance information.
4. **Seismic data**: Seismology is a scientific study of earthquakes. It involves collecting data and processes to derive information that helps determine the location and magnitude of earthquakes.

## 1.4 TYPES OF DATA

Data can be classified as structured or unstructured (see Figure 1-2) based on how it is stored and managed. Structured data is organized in rows and columns in a rigidly defined format so that applications can retrieve and process it efficiently. Structured data is typically stored using a database management system (DBMS). Data is unstructured if its elements cannot be stored in rows and columns, which makes it difficult to query and retrieve by applications. For example, customer contacts that are stored in various forms such as sticky notes, e-mail messages, business cards, or even digital format files, such as .doc, .txt, and .pdf. Due to its unstructured nature, it is difficult to retrieve this data using a traditional customer relationship management application. A vast majority of new data being created today is unstructured. The industry is challenged with with new architectures, technologies, techniques, and skills to store, manage, analyze, and derive value from unstructured data from numerous sources.

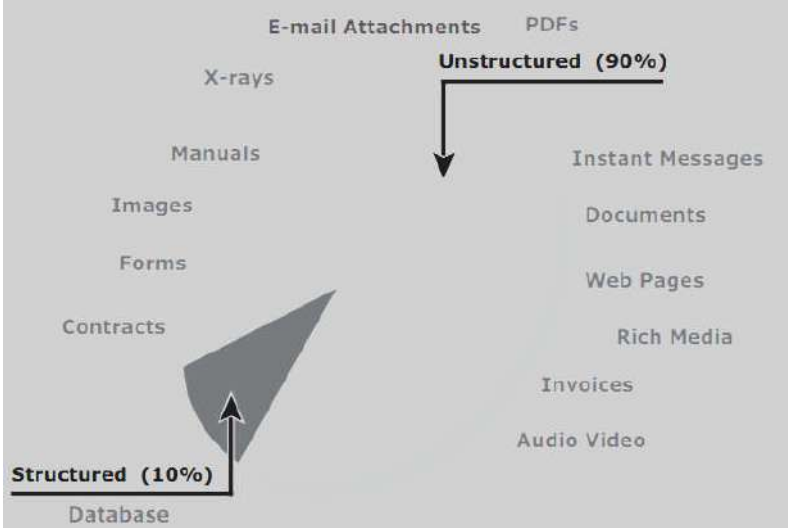



*Figure 1.2: Types of Data*

## 1.5 BIG DATA

Big data is a new and evolving concept, which 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. It includes both structured and unstructured data generated by a variety of sources, including business application transactions, web pages, videos, images, e-mails, social media, and so on. These data sets typically require real-time capture or updates for analysis, predictive modelling, and decision making. Significant opportunities exist to extract value from big data. The big data ecosystem (see Figure 1-3) consists of the following:

1. Devices that collect data from multiple locations and generate new data about this data (metadata).
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 and buyers who benefit from the information collected and aggregated by others in the data value chain.

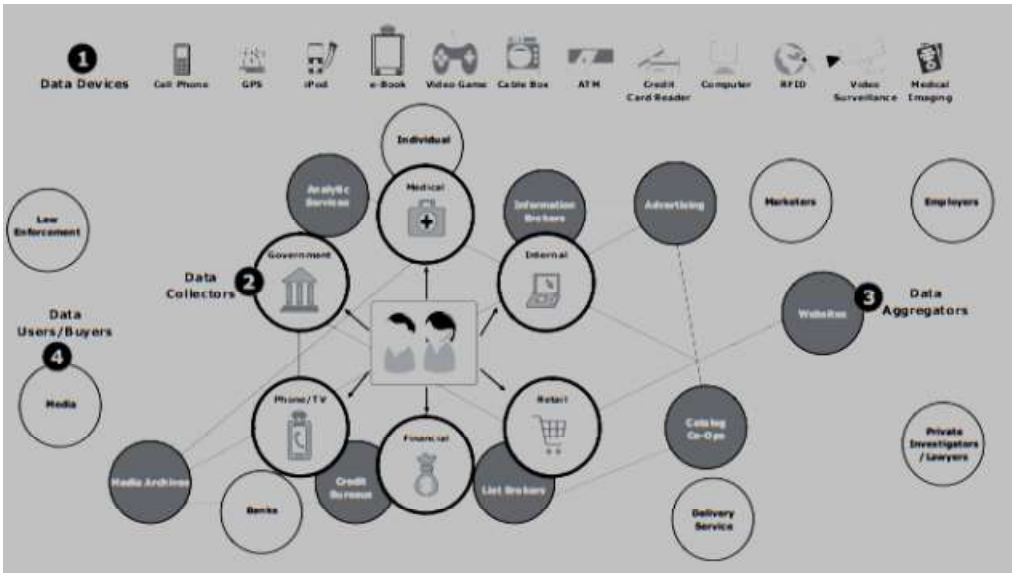



***Figure 1.3: Big Data Ecosystem***

Traditional IT infrastructure and data processing tools and methodologies are inadequate to handle the volume, variety, dynamism, and complexity of big data. Analysing big data in real time requires new techniques, architectures, and tools that provide high performance, massively parallel processing (MPP) data platforms, and advanced analytics on the data sets. Data science is an emerging discipline, which enables organizations to derive business value from big data. Data science represents the synthesis of several existing disciplines, such as statistics, math, data visualization, and computer science to enable data scientists to develop advanced algorithms for the purpose of analyzing vast amounts of information to drive new value and make more data-driven decisions.

## 1.6 INFORMATION

Data, whether structured or unstructured, does not fulfill any purpose for individuals or businesses unless it is presented in a meaningful form. Information is the intelligence and knowledge derived from data. Businesses analyze raw data to identify meaningful trends. On the basis of these trends, a company can plan or modify its strategy. For example, 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.

## 1.7 STORAGE

Data created by individuals or businesses must be stored so that it is easily accessible for further processing. In a computing environment, devices designed for storing data are termed storage devices or simply storage. The type of storage used varies based on the type of data and the

rate at which it is created and used. Devices, such as a media card in a cell phone or digital camera, DVDs, CD-ROMs, and disk drives in personal computers are examples of storage devices. Businesses have several options available for storing data, including internal hard disks, external disk arrays, and tapes.

## 1.8 EVOLUTION OF STORAGE ARCHITECTURE

Historically, organizations had centralized computers (mainframes) and information storage devices (tape reels and disk packs) in their data center. The evolution of open systems, their affordability, and ease of deployment made it possible for business units/departments to have their own servers and storage. In earlier implementations of open systems, the storage was typically internal to the server. These storage devices could not be shared with any other servers. This approach is referred to as server-centric storage architecture (see Figure 1-4 [a]). In this architecture, each server has a limited number of storage devices, and any administrative tasks, such as maintenance of the server or increasing storage capacity, might result in unavailability of information. The proliferation of departmental servers in an enterprise resulted in unprotected, unmanaged, fragmented islands of information and increased capital and operating expenses.

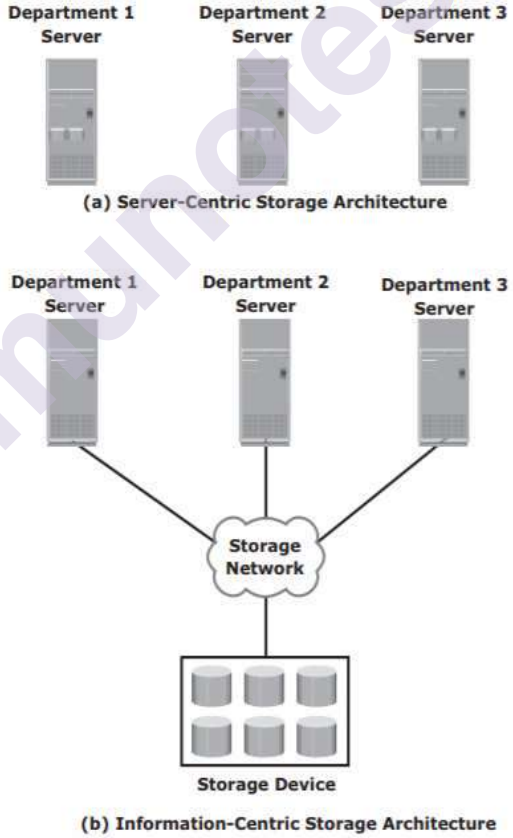



***Figure 1.4: Evolution of Storage Architecture***

To overcome these challenges, storage evolved from server-centric to information-centric architecture (see Figure 1-4 [b]). In this architecture, storage devices are managed centrally and independent of

servers. These centrally-managed storage devices are shared with multiple servers. When a new server is deployed in the environment, storage is assigned from the same shared storage devices to that server. The capacity of shared storage can be increased dynamically by adding more storage devices without impacting information availability. In this architecture, information management is easier and cost-effective. Storage technology and architecture continue to evolve, which enables organizations to consolidate, protect, optimize, and leverage their data to achieve the highest return on information assets.

## 1.9 DATA CENTER INFRASTRUCTURE

Organizations maintain data centers to provide centralized data-processing capabilities across the enterprise. Data centers house and manage large amounts of data. The data center infrastructure includes hardware components, such as computers, storage systems, network devices, and power backups; and software components, such as applications, operating systems, and management software. It also includes environmental controls, such as air conditioning, fi re suppression, and ventilation. Large organizations often maintain more than one data center to distribute data processing workloads and provide backup if a disaster occurs.

## 1.10 CORE ELEMENTS OF DATA CENTRE

Five core elements are essential for the functionality of a data center:

1. Application: A computer program that provides the logic for computing operations.
2. Database management system (DBMS): Provides a structured way to store data in logically organized tables that are interrelated
3. Host or compute: A computing platform (hardware, firmware, and software) that runs applications and databases
4. Network: A data path that facilitates communication among various networked devices
5. Storage: A device that stores data persistently for subsequent use

These core elements are typically viewed and managed as separate entities, but all the elements must work together to address data-processing requirements.

Figure 1-5 shows an example of an online order transaction system that involves the five core elements of a data center and illustrates their functionality in a business process.

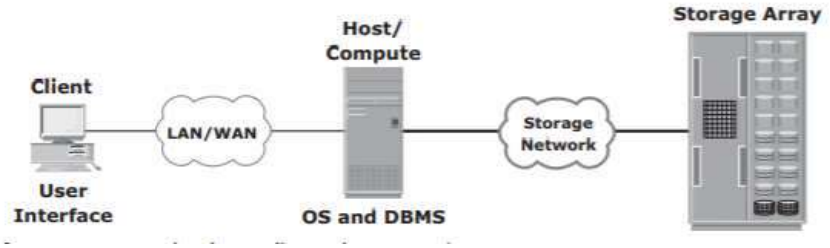



*Figure 1.5: Example of an online order transaction system*

A customer places an order through a client machine connected over a LAN/ WAN to a host running an order-processing application. The client accesses the DBMS on the host through the application to provide order-related information, such as the customer's name, address, payment method, products ordered, and quantity ordered. The DBMS uses the host operating system to write this data to the physical disks in the storage array. The storage networks provide the communication link between the host and the storage array and transports the request to read or write data between them. The storage array, after receiving the read or write request from the host, performs the necessary operations to store the data on physical disks.

## 1.11 KEY CHARACTERISTICS OF DATA CENTRE

Uninterrupted operation of data centers is critical to the survival and success of a business. Organizations must have a reliable infrastructure that ensures that data is accessible at all times. Although the characteristics shown in Figure 1-6 are applicable to all elements of the data center infrastructure, the focus here is on storage systems.

1. **Availability**: A data center should ensure the availability of information when required. Unavailability of information could cost millions of dollars per hour to businesses, such as financial services, telecommunications, and e-commerce.
2. **Security**: Data centers must establish policies, procedures, and core element integration to prevent unauthorized access to information
3. **Scalability**: Business growth often requires deploying more servers, new applications, and additional databases. Data center resources should scale based on requirements, without interrupting business operations.
4. **Performance**: All the elements of the data center should provide optimal performance based on the required service levels.
5. **Data integrity**: Data integrity refers to mechanisms, such as error correction codes or parity bits, which ensure that data is stored and retrieved exactly as it was received.
6. **Capacity**: Data center operations require adequate resources to store and process large amounts of data, efficiently. When capacity requirements increase, the data center must provide additional capacity without interrupting availability or with minimal disruption. Capacity

may be managed by reallocating the existing resources or by adding new resources.

7. **Manageability**: A data center should provide easy and integrated management of all its elements. Manageability can be achieved through automation and reduction of human (manual) intervention in common tasks.


*Figure 1.6: Key Characteristics of Data Centre*

## 1.12 MANAGING A DATA CENTRE

Managing a data center involves many tasks. The key management activities include the following:

1. **Monitoring**: It is a continuous process of gathering information on various elements and services running in a data center. The aspects of a data center that are monitored include security, performance, availability, and capacity.
2. **Reporting**: It is done periodically on resource performance, capacity, and utilization. Reporting tasks help to establish business justifications and chargeback of costs associated with data center operations.
3. **Provisioning**: It is a process of providing the hardware, software, and other resources required to run a data center. Provisioning activities primarily include resources management to meet capacity, availability, performance, and security requirements.

Virtualization and cloud computing have dramatically changed the way data center infrastructure resources are provisioned and managed. Organizations are rapidly deploying virtualization on various elements of data centers to optimize their utilization. Further, continuous cost pressure on IT and on-demand data processing requirements have resulted in the adoption of cloud computing.

## 1.13 VIRTUALIZATION AND CLOUD COMPUTING

Virtualization is a technique of abstracting physical resources, such as compute, storage, and network, and making them appear as logical resources. Virtualization has existed in the IT industry for several years and in different forms. Common examples of virtualization are virtual memory used on computer systems and partitioning of raw disks. Virtualization enables pooling of physical resources and providing an aggregated view of the physical resource capabilities. For example, storage virtualization enables multiple pooled storage devices to appear as a single large storage entity. Similarly, by using compute virtualization, the CPU capacity of the pooled physical servers can be viewed as the aggregation of the power of all CPUs (in megahertz). Virtualization also enables centralized management of pooled resources. Virtual resources can be created and provisioned from the pooled physical resources. For example, a virtual disk of a given capacity can be created from a storage pool or a virtual server with specific CPU power and memory can be configured from a compute pool. These virtual resources share pooled physical resources, which improves the utilization of physical IT resources. Based on business requirements, capacity can be added to or removed from the virtual resources without any disruption to applications or users. With improved utilization of IT assets, organizations save the costs associated with procurement and management of new physical resources. Moreover, fewer physical resources means less space and energy, which leads to better economics and green computing. In today's fast-paced and competitive environment, organizations must be agile and flexible to meet changing market requirements. This leads to rapid expansion and upgrade of resources while meeting shrinking or stagnant IT budgets. Cloud computing addresses these challenges efficiently. Cloud computing enables individuals or businesses to use IT resources as a service over the network. It provides highly scalable and flexible computing that enables provisioning of resources on demand. Users can scale up or scale down the demand of computing resources, including storage capacity, with minimal management effort or service provider interaction. Cloud computing empowers self-service requesting through a fully automated request-fulfilment process. Cloud computing enables consumption-based metering; therefore, consumers pay only for the resources they use, such as CPU hours used, amount of data transferred, and gigabytes of data stored. Cloud infrastructure is usually built upon virtualized data centers, which provide resource pooling and rapid provisioning of resources.

## 1.14 SUMMARY

- This chapter described the importance of data, information, and storage infrastructure.
- Meeting today's storage needs begins with understanding the type of data, its value, and key attributes of a data center.

- The evolution of storage architecture and the core elements of a data center covered in this chapter provided the foundation for information storage and management.
- The key elements of data centres are explained in the chapter.
- The emergence of virtualization has provided the opportunity to transform classic data centers into virtualized data centers.
- Cloud computing is further changing the way IT resources are provisioned and consumed.

## 1.15 REVIEW YOUR LEARNING

- Can you explain the need to information storage and retrieval?
- Explain how data are managed using Data Centre?
- Are you able to write the key elements of Data centres?
- Can you relate data and information in your data to day internet / application usage?

## 1.16 QUESTIONS

1. What is structured and unstructured data?
2. Explain the challenges of storing and managing unstructured data.
3. Discuss the benefits of information-centric storage architecture over server-centric storage architecture.
4. What are the attributes of big data? Research and prepare a presentation on big data analytics.
5. Research how businesses use their information assets to derive competitive advantage and new business opportunities.
6. Research and prepare a presentation on personal data management

## 1.17 FURTHER READING

- http://aad.tpu.ru/practice/EMC/Information%20Storage%20and%20Management-v.2.pdf
- https://nptel.ac.in/courses/106/108/106108058/
- https://nptel.ac.in/content/storage2/courses/106108058/lec%2007.pdf
- http://www.ictacademy.in/pages/Information-Storage-and-Management.aspx
- https://www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwiM8Kq6isHyAhUEkmYCHbJyDXAYABAAGgJzbQ&ae=2&ohost=www.google.com&cid=CAESQeD28QNmzUxhr6qtgEwm24g2Yc-TeMC_24a0sxeZf9MitA7QrS5Vz4VE3XfWSwFvX0iAKPoH4fT4QmSj7PhnMAQF&sig=AOD64_1Y3y_5vJpAZOJybqnNONsE6wNayQ&q&adurl&ved=2ahUKEwjvsaG6isHyAhXjxTgGHTvKBEEQ0Qx6BAgDEAE
- https://www.coursera.org/lecture/big-data-management/data-storage-RplBY

- https://www.coursera.org/courses?query=data%20storage
- https://www.coursera.org/lecture/technical-support-fundamentals/storage-RLNIZ
- https://www.coursera.org/learn/cloud-storage-big-data-analysis-sql

## 1.18 REFERENCES


1. Information Storage and Management: Storing, Managing and Protecting Digital Information in Classic, Virtualized and Cloud Environments, EMC, John & Wiley Sons, 2nd Edition, 2012\
2. Information Storage and Management, Pankaj Sharma
3. Information Technology Project Management, Jack T Marchewka
4. Information Storage and Management, I A Dhotre




❖ ❖ ❖ ❖

# 2

# DATA CENTRE ENVIRONMENT

**Unit Structure**

## 2.0 OBJECTIVES

1. Explain need and applications of virtualization
2. Differentiate Compute, Desktop and Memory Virtualization
3. Describe different Storage Media
4. Analyse Data Addressing

## 2.1 INTRODUCTION

Today, data centres are essential and integral parts of any business, whether small, medium, or large in size. The core elements of a data center are host, storage, connectivity (or network), applications, and DBMS that are managed centrally. These elements work together to process and store data. With the evolution of virtualization, data centers have also evolved from a classic data center to a virtualized data center (VDC). In a VDC, physical resources from a classic data center are pooled together and provided as virtual resources. This abstraction hides the complexity and limitation of physical resources from the user. By consolidating IT resources using virtualization, organizations can optimize their infrastructure utilization and reduce the total cost of owning an infrastructure. Moreover, in a VDC, virtual resources are created using software that enables faster deployment, compared to deploying physical resources in classic data centers. This chapter covers all the key components of a data center, including virtualization at compute, memory, desktop, and application. Storage and network virtualization is discussed later in the book. With the increase in the criticality of information assets to businesses, storage — one of the core elements of a data center — is recognized as a distinct resource. Storage needs special focus and attention for its implementation and management. This chapter also focuses on storage subsystems and provides details on components, geometry, and performance parameters of a disk drive. The connectivity between the host and storage facilitated by various technologies is also explained.

## 2.2 APPLICATION

An application is a computer program that provides the logic for computing operations. The application sends requests to the underlying operating system to perform read/write (R/W) operations on the storage devices. Applications can be layered on the database, which in turn uses the OS services to perform R/W operations on the storage devices. Applications deployed in a data centre environment are commonly categorized as business applications, infrastructure management applications, data protection applications, and security applications. Some examples of these applications are e-mail, enterprise resource planning (ERP), decision support system (DSS), resource management, backup, authentication and antivirus applications, and so on. The characteristics of I/Os (Input/Output) generated by the application influence the overall performance of storage system and storage solution designs.

***Application virtualization*** breaks the dependency between the application and the underlying platform (OS and hardware). Application virtualization encapsulates the application and the required OS resources within a virtualized container. This technology provides the ability to deploy applications without making any change to the underlying OS, fi le system, or registry of the computing platform on which they are deployed. Because virtualized applications run in an isolated environment, the underlying OS and other applications are protected from potential corruptions. There are many scenarios in which conflicts might arise if multiple applications or multiple versions of the same application are installed on the same computing platform. Application virtualization eliminates this conflict by isolating different versions of an application and the associated O/S resources.

## 2.3 DATABASE MANAGEMENT SYSTEMS

A database is a structured way to store data in logically organized tables that are interrelated. A database helps to optimize the storage and retrieval of data. A DBMS controls the creation, maintenance, and use of a database. The DBMS processes an application's request for data and instructs the operating system to transfer the appropriate data from the storage.

## 2.4 HOST (COMPUTE)

Users store and retrieve data through applications. The computers on which these applications run are referred to as hosts or compute systems. Hosts can be physical or virtual machines. A compute virtualization software enables creating virtual machines on top of a physical compute infrastructure. Compute virtualization and virtual machines are discussed later in this chapter. Examples of physical hosts

include desktop computers, servers or a cluster of servers, laptops, and mobile devices. A host consists of CPU, memory, I/O devices, and a collection of software to perform computing operations. This software includes the operating system, file system, logical volume manager, device drivers, and so on. This software can be installed as separate entities or as part of the operating system. The CPU consists of four components: Arithmetic Logic Unit (ALU), control unit, registers, and L1 cache. There are two types of memory on a host, Random Access Memory (RAM) and Read-Only Memory (ROM). I/O devices enable communication with a host. Examples of I/O devices are keyboard, mouse, monitor, etc. Software runs on a host and enables processing of input and output (I/O) data. The following section details various software components that are essential parts of a host system.

### 2.4.1 Operating Systems

In a traditional computing environment, an operating system controls all aspects of computing. It works between the application and the physical components of a compute system. One of the services it provides to the application is data access. The operating system also monitors and responds to user actions and the environment. It organizes and controls hardware components and manages the allocation of hardware resources. It provides basic security for the access and usage of all managed resources. An operating system also performs basic storage management tasks while managing other underlying components, such as the fi le system, volume manager, and device drivers. In a virtualized compute environment, the virtualization layer works between the operating system and the hardware resources. Here the OS might work differently based on the type of compute virtualization implemented. In a typical implementation, the OS works as a guest and performs only the activities related to application interaction. In this case, hardware management functions are handled by the virtualization layer.

***Memory Virtualization***

Memory has been, and continues to be, an expensive component of a host. It determines both the size and number of applications that can run on a host. Memory virtualization enables multiple applications and processes, whose aggregate memory requirement is greater than the available physical memory, to run on a host without impacting each other. Memory virtualization is an operating system feature that virtualizes the physical memory (RAM) of a host. It creates virtual memory with an address space larger than the physical memory space present in the compute system. The virtual memory encompasses the address space of the physical memory and part of the disk storage. The operating system utility that manages the virtual memory is known as the virtual memory manager (VMM). The VMM manages the virtual-to-physical memory mapping and fetches data from the disk storage when a process references a virtual address that points to data at the disk storage. The space used by the VMM on the disk is known as a swap space. A swap space (also known as page file or swap file) is a portion of the disk drive that appears to be physical memory to the operating system. In a virtual memory

implementation, the memory of a system is divided into contiguous blocks of fixed-size pages. A process known as paging moves inactive physical memory pages onto the swap file and brings them back to the physical memory when required. This enables efficient use of the available physical memory among different applications. The operating system typically moves the least used pages into the swap file so that enough RAM is available for processes that are more active. Access to swap fi le pages is slower than access to physical memory pages because swap fi le pages are allocated on the disk drive, which is slower than physical memory.

### 2.4.2 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. A device driver enables the operating system to recognize the device and to access and control devices. Device drivers are hardware-dependent and operating-system-specific.

### 2.4.3 Volume Manager

In the early days, disk drives appeared to the operating system as a number of continuous disk blocks. The entire disk drive would be allocated to the file system or other data entity used by the operating system or application. The disadvantage was lack of flexibility. When a disk drive ran out of space, there was no easy way to extend the file system's size. Also, as the storage capacity of the disk drive increased, allocating the entire disk drive for the file system often resulted in underutilization of storage capacity. The evolution of Logical Volume Managers (LVMs) enabled dynamic extension of file system capacity and efficient storage management. The LVM is software that runs on the compute system and manages logical and physical storage. LVM is an intermediate layer between the file system and the physical disk. It can partition a larger-capacity disk into virtual, smaller-capacity volumes (the process is called partitioning) or aggregate several smaller disks to form a larger virtual volume. (The process is called concatenation.) These volumes are then presented to applications. Disk partitioning was introduced to improve the flexibility and utilization of disk drives. In partitioning, a disk drive is divided into logical containers called logical volumes (LVs) (see Figure 2-1). For example, a large physical drive can be partitioned into multiple LVs to maintain data according to the fi le system and application requirements. The partitions are created from groups of contiguous cylinders when the hard disk is initially set up on the host. The host's fi le system accesses the logical volumes without any knowledge of partitioning and physical structure of the disk.

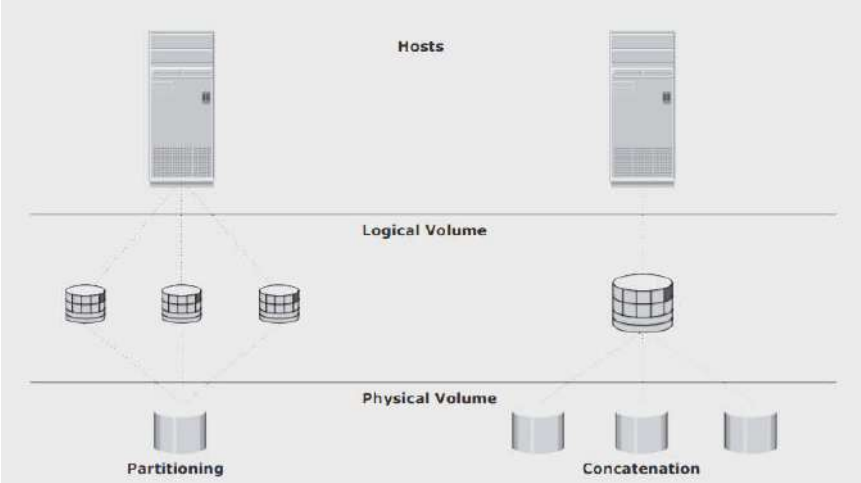



***Figure 2.1: Disk Partitioning and Concatenation***

Concatenation is the process of grouping several physical drives and presenting them to the host as one big logical volume (see Figure 2-1). The LVM provides optimized storage access and simplifies storage resource management. It hides details about the physical disk and the location of data on the disk. It enables administrators to change the storage allocation even when the application is running. The basic LVM components are physical volumes, volume groups, and logical volumes. In LVM terminology, each physical disk connected to the host system is a physical volume (PV). The LVM converts the physical storage provided by the physical volumes to a logical view of storage, which is then used by the operating system and applications. A volume group is created by grouping together one or more physical volumes. A unique physical volume identifier (PVID) is assigned to each physical volume when it is initialized for use by the LVM. Physical volumes can be added or removed from a volume group dynamically. They cannot be shared between different volume groups, which means that the entire physical volume becomes part of a volume group. Each physical volume is partitioned into equal-sized data blocks called physical extents when the volume group is created. Logical volumes are created within a given volume group. A logical volume can be thought of as a disk partition, whereas the volume group itself can be thought of as a disk. A volume group can have a number of logical volumes. The size of a logical volume is based on a multiple of the physical extents. The logical volume appears as a physical device to the operating system. A logical volume is made up of non-contiguous physical extents and may span multiple physical volumes. A file system is created on a logical volume. These logical volumes are then assigned to the application. A logical volume can also be mirrored to provide enhanced data availability.

### 2.4.4 File Systems

A file is a collection of related records or data stored as a unit with a name. A file system is a hierarchical structure of files. A file system enables easy access to data files residing within a disk drive, a disk partition, or a logical volume. A file system consists of logical structures

and software routines that control access to files. It provides users with the functionality to create, modify, delete, and access files. Access to files on the disks is controlled by the permissions assigned to the file by the owner, which are also maintained by the file system. A file system organizes data in a structured hierarchical manner via the use of directories, which are containers for storing pointers to multiple files. All file systems maintain a pointer map to the directories, subdirectories, and files that are part of the file system. Examples of common file systems are:

1. FAT 32 (File Allocation Table) for Microsoft Windows
2. NT File System (NTFS) for Microsoft Windows
3. UNIX File System (UFS) for UNIX
4. Extended File System (EXT2/3) for Linux

Apart from the files and directories, the file system also includes a number of other related records, which are collectively called the metadata. For example, the metadata in a UNIX environment consists of the super block, the in odes, and the list of data blocks free and in use. The metadata of a file system must be consistent for the file system to be considered healthy. A super block contains important information about the file system, such as the file system type, creation and modification dates, size, and layout. It also contains the count of available resources (such as the number of free blocks, in odes, and so on) and a flag indicating the mount status of the file system. An inode is associated with every file and directory and contains information such as the file length, ownership, access privileges, time of last access/modification, number of links, and the address of the data. A file system block is the smallest "unit" allocated for storing data. Each file system block is a contiguous area on the physical disk. The block size of a file system is fixed at the time of its creation. The file system size depends on the block size and the total number of file system blocks. A file can span multiple file system blocks because most files are larger than the predefined block size of the file system. File system blocks cease to be contiguous and become fragmented when new blocks are added or deleted. Over time, as files grow larger, the file system becomes increasingly fragmented. The following list shows the process of mapping user files to the disk storage subsystem with an LVM (see Figure 2-2):

1. Files are created and managed by users and applications.
2. These files reside in the file systems.
3. The file systems are mapped to file system blocks.
4. The file system blocks are mapped to logical extents of a logical volume.
5. These logical extents in turn are mapped to the disk physical extents either by the operating system or by the LVM.
6. These physical extents are mapped to the disk sectors in a storage subsystem. If there is no LVM, then there are no logical extents. Without LVM, file system blocks are directly mapped to disk sectors.

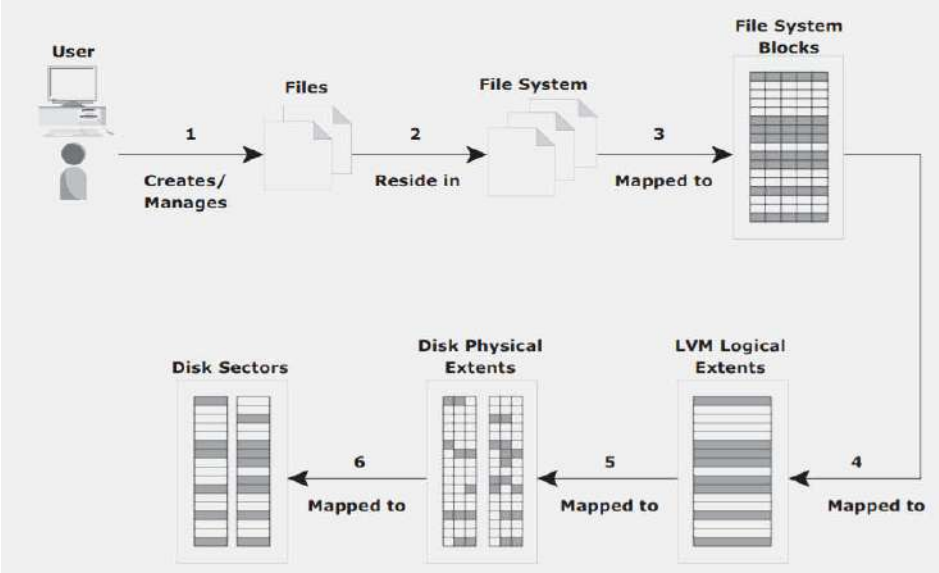



***Figure 2.2: Process of Mapping User Files to Disk Storage***

The file system tree starts with the root directory. The root directory has a number of subdirectories. A file system should be mounted before it can be used.

A file system can be either a journaling file system or a no journaling file system. No journaling file systems cause a potential loss of files because they use separate writes to update their data and metadata. If the system crashes during the write process, the metadata or data might be lost or corrupted. When the system reboots, the file system attempts to update the metadata structures by examining and repairing them. This operation takes a long time on large file systems. If there is insufficient information to re-create the wanted or original structure, the files might be misplaced or lost, resulting in corrupted file systems. A journaling file system uses a separate area called a log or journal. This journal might contain all the data to be written (physical journal) or just the metadata to be updated (logical journal). Before changes are made to the file system, they are written to this separate area. After the journal has been updated, the operation on the file system can be performed. If the system crashes during the operation, there is enough information in the log to "replay" the log record and complete the operation. Journaling results in a quick file system check because it looks only at the active, most recently accessed parts of a large file system. In addition, because information about the pending operation is saved, the risk of files being lost is reduced. A disadvantage of journaling file systems is that they are slower than other file systems. This slowdown is the result of the extra operations that have to be performed on the journal each time the file system is changed. However, the much-shortened time for file system checks and the file system integrity provided by journaling far outweighs its disadvantage. Nearly all file system implementations today use journaling. Dedicated file servers may be installed to manage and share a large number of files over a network. These file servers support multiple file systems and use

file-sharing protocols specific to the operating system — for example, NFS and CIFS.

## 2.5 COMPUTE VIRTUALIZATION

Compute virtualization is a technique for masking or abstracting the physical hardware from the operating system. 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). Each VM runs an operating system and application instance in an isolated manner. Compute virtualization is achieved by a virtualization layer that resides between the hardware and virtual machines. This layer is also called the hypervisor. The hypervisor provides hardware resources, such as CPU, memory, and network to all the virtual machines. Within a physical server, a large number of virtual machines can be created depending on the hardware capabilities of the physical 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. However, all VMs share the same underlying physical hardware in an isolated manner. From a hypervisor perspective, virtual machines are discrete sets of fi les that include VM configuration fi le, data files, and so on.

Typically, a physical server often faces resource-conflict issues when two or more applications running on the server have conflicting requirements. For example, applications might need different values in the same registry entry, different versions of the same DLL, and so on. These issues are further compounded with an application's high-availability requirements. As a result, the servers are limited to serve only one application at a time, as shown in Figure 2-3 (a). This causes organizations to purchase new physical machines for every application they deploy, resulting in expensive and inflexible infrastructure. On the other hand, many applications do not take full advantage of the hardware capabilities available to them. Consequently, resources such as processors, memory, and storage remain underutilized. Compute virtualization enables users to overcome these challenges (see Figure 2-3 [b]) by allowing multiple operating systems and applications to run on a single physical machine. This technique significantly improves server utilization and provides server consolidation.

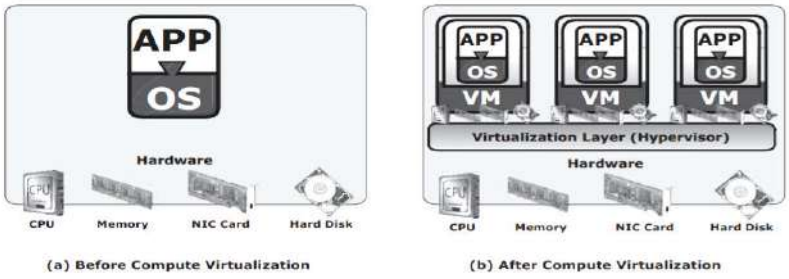



***Figure 2.3: Server Virtualization***

Server consolidation enables organizations to run their data center with fewer servers. This, in turn, cuts down the cost of new server acquisition, reduces operational cost, and saves datacenter floor and rack space. Creation of VMs takes less time compared to a physical server setup; organizations can provision servers faster and with ease. Individual VMs can be restarted, upgraded, or even crashed, without affecting the other VMs on the same physical machine. Moreover, VMs can be copied or moved from one physical machine to another without causing application downtime. Nondisruptive migration of VMs is required for load balancing among physical machines, hardware maintenance, and availability purposes.

***Desktop Virtualization***:

With the traditional desktop, the OS, applications, and user profiles are all tied to a specific piece of hardware. With legacy desktops, business productivity is impacted greatly when a client device is broken or lost. Desktop virtualization breaks the dependency between the hardware and its OS, applications, user profiles, and settings. This enables the IT staff to change, update, and deploy these elements independently. Desktops hosted at the data centerrun on virtual machines; users remotely access these desktops from a variety of client devices, such as laptops, desktops, and mobile devices (also called thin devices). Application execution and data storage are performed centrally at the data center instead of at the client devices. Because desktops run as virtual machines within an organization's data center, it mitigates the risk of data leakage and theft. It also helps to perform centralized backup and simplifies compliance procedures. Virtual desktops are easy to maintain because it is simple to apply patches, deploy new applications and OS, and provision or remove users centrally.

## 2.6 CONNECTIVITY

Connectivity refers to the interconnection between hosts or between a host and peripheral devices, such as printers or storage devices. The discussion here focuses only on the connectivity between the host and the storage device. Connectivity and communication between host and storage are enabled using physical components and interface protocols.

### 2.6.1 Physical Components of Connectivity

The physical components of connectivity are the hardware elements that connect the host to storage. Three physical components of connectivity between the host and storage are the host interface device, port, and cable (Figure 2-4). A host interface device or host adapter connects a host to other hosts and storage devices. Examples of host interface devices are host bus adapter (HBA) and network interface card (NIC). Host bus adaptor is an application-specific integrated circuit (ASIC) board that performs I/O interface functions between the host and storage, relieving the CPU from additional I/O processing workload. A host typically contains multiple HBAs. A port is a specialized outlet that

enables connectivity between the host and external devices. An HBA may contain one or more ports to connect the host to the storage device. Cables connect hosts to internal or external devices using copper or fiber optic media.

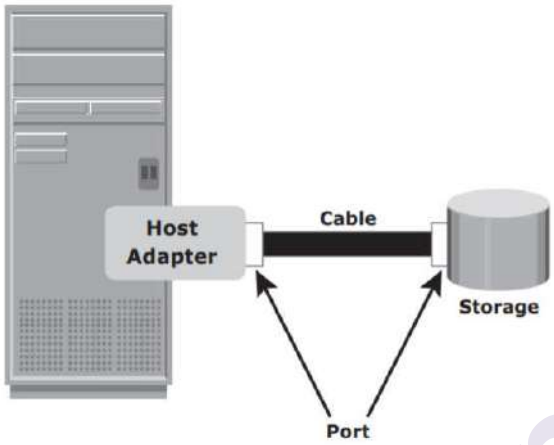



***Figure 2.4: Physical Components of Connectivity***

### 2.6.2 Interface Protocols

A protocol enables communication between the host and storage. Protocols are implemented using interface devices (or controllers) at both source and destination. The popular interface protocols used for host to storage communications are Integrated Device Electronics/Advanced Technology Attachment (IDE/ATA), Small Computer System Interface (SCSI), Fibre Channel (FC) and Internet Protocol (IP).

#### 2.6.2.1 IDE/ATA and Serial ATA

IDE/ATA is a popular interface protocol standard used for connecting storage devices, such as disk drives and CD-ROM drives. This protocol supports parallel transmission and therefore is also known as Parallel ATA (PATA) or simply ATA. IDE/ATA has a variety of standards and names. The Ultra DMA/133 version of ATA supports a throughput of 133 MB per second. In a master-slave configuration, an ATA interface supports two storage devices per connector. However, if the performance of the drive is important, sharing a port between two devices is not recommended. The serial version of this protocol supports single bit serial transmission and is known as Serial ATA (SATA). High performance and low-cost SATA has largely replaced PATA in newer systems. SATA revision 3.0 provides a data transfer rate up to 6 Gb/s.

#### 2.6.2.2 SCSI and Serial SCSI

SCSI has emerged as a preferred connectivity protocol in high-end computers. This protocol supports parallel transmission and offers improved performance, scalability, and compatibility compared to ATA. However, the high cost associated with SCSI limits its popularity among home or personal desktop users. Over the years, SCSI has been enhanced and now includes a wide variety of related technologies and standards. SCSI supports up to 16 devices on a single bus and provides data transfer

rates up to 640 MB/s (for the Ultra-640 version). Serial attached SCSI (SAS) is a point-to-point serial protocol that provides an alternative to parallel SCSI. A newer version of serial SCSI (SAS 2.0) supports a data transfer rate up to 6 Gb/s. This book's Appendix B provides more details on the SCSI architecture and interface.

#### 2.6.2.3 Fibre Channel

Fibre Channel is a widely used protocol for high-speed communication to the storage device. The Fibre Channel interface provides gigabit network speed. It provides a serial data transmission that operates over copper wire and optical fibre. The latest version of the FC interface (16FC) allows transmission of data up to 16 Gb/s.

#### 2.6.2.4 Internet Protocol (IP)

IP is a network protocol that has been traditionally used for host-to-host traffic. With the emergence of new technologies, an IP network has become a viable option for host-to-storage communication. IP offers several advantages in terms of cost and maturity and enables organizations to leverage their existing IP-based network. iSCSI and FCIP protocols are common examples that leverage IP for host-to-storage communication.

## 2.7 STORAGE

Storage is a core component in a data centre. A storage device uses magnetic, optic, or solid-state media. Disks, tapes, and diskettes use magnetic media, whereas CD/DVD uses optical media for storage. Removable Flash memory or Flash drives are examples of solid-state media.

In the past, tapes were the most popular storage option for backups because of their low cost. However, tapes have various limitations in terms of performance and management, as listed here:

1. Data is stored on the tape linearly along the length of the tape. Search and retrieval of data are done sequentially, and it invariably takes several seconds to access the data. As a result, random data access is slow and time-consuming. This limits tapes as a viable option for applications that require real-time, rapid access to data.
2. 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.
3. On a tape drive, the read/write head touches the tape surface, so the tape degrades or wears out after repeated use.
4. The storage and retrieval requirements of data from the tape and the overhead associated with managing the tape media are significant.

Due to these limitations and availability of low-cost disk drives, tapes are no longer a preferred choice as a backup destination for

enterprise-class data centers. Optical disc storage is popular in small, single-user computing environments. It is frequently used by individuals to store photos or as a backup medium on personal or laptop computers. It is also used as a distribution medium for small applications, such as games, or as a means to transfer small amounts of data from one computer system to another. Optical discs have limited capacity and speed, which limit the use of optical media as a business data storage solution. The capability to write once and read many (WORM) is one advantage of optical disc storage. A CD-ROM is an example of a WORM device. Optical discs, to some degree, guarantee that the content has not been altered. Therefore, it can be used as a low-cost alternative for long-term storage of relatively small amounts of fixed content that do not change after it is created. Collections of optical discs in an array, called a jukebox, are still used as a fixed-content storage solution. Other forms of optical discs include CD-RW, Blu-ray disc, and other variations of DVD. Disk drives are the most popular storage medium used in modern computers for storing and accessing data for performance-intensive, online applications. Disks support rapid access to random data locations. This means that data can be written or retrieved quickly for a large number of simultaneous users or applications. In addition, disks have a large capacity. Disk storage arrays are configured with multiple disks to provide increased capacity and enhanced performance.

Disk drives are accessed through predefined protocols, such as ATA, Serial ATA (SATA), SAS (Serial Attached SCSI), and FC. These protocols are implemented on the disk interface controllers. Earlier, disk interface controllers were implemented as separate cards, which were connected to the motherboard to provide communication with storage devices. Modern disk interface controllers are integrated with the disk drives; therefore, disk drives are known by the protocol interface they support, for example SATA disk, FC disk, and so on.

## 2.8 DISK DRIVE COMPONENTS

The key components of a hard disk drive are platter, spindle, read-write head, actuator arm assembly, and controller board (see Figure 2-5). I/O operations in a HDD are performed by rapidly moving the arm across the rotating fl at platters coated with magnetic particles. Data is transferred between the disk controller and magnetic platters through the read-write (R/W) head which is attached to the arm. Data can be recorded and erased on magnetic platters any number of times. Following sections detail the different components of the disk drive, the mechanism for organizing and storing data on disks, and the factors that affect disk performance.

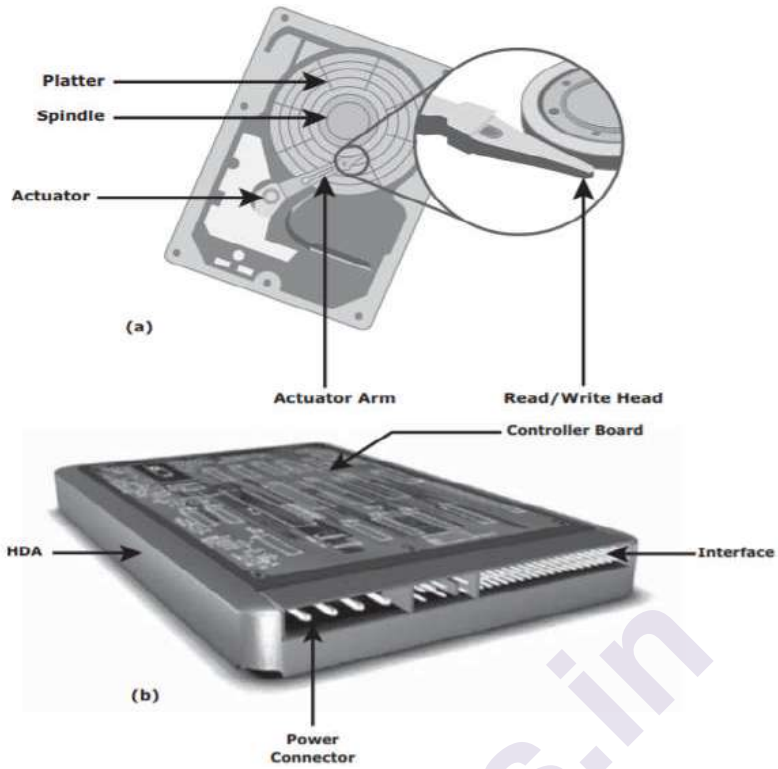



*Figure 2.5: Disk Drive Components*

### 2.8.1 Platter

A typical HDD consists of one or more flat circular disks called platters (Figure 2-6). The data is recorded on these platters in binary codes (0s and 1s). The set of rotating platters is sealed in a case, called the Head Disk Assembly (HDA). A platter is a rigid, round disk coated with magnetic material on both surfaces (top and bottom). The data is encoded by polarizing the magnetic area, or domains, of the disk surface. Data can be written to or read from both surfaces of the platter. The number of platters and the storage capacity of each platter determine the total capacity of the drive.

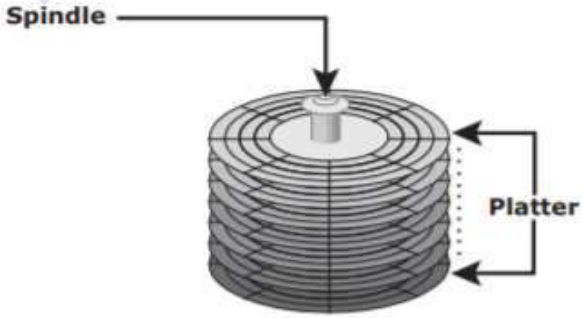



*Figure 2.6: Spindle and Platter*

### 2.8.2 Spindle

A spindle connects all the platters (refer to Figure 2-6) and is connected to a motor. The motor of the spindle rotates with a constant speed. The disk platter spins at a speed of several thousands of revolutions

per minute (rpm). Common spindle speeds are 5,400 rpm, 7,200 rpm, 10,000 rpm, and 15,000 rpm. The speed of the platter is increasing with improvements in technology, although the extent to which it can be improved is limited.

### 2.8.3 Read / Write Head

Read/Write (R/W) heads, as shown in Figure 2-7, read and write data from or to platters. Drives have two R/W heads per platter, one for each surface of the platter. The R/W head changes the magnetic polarization on the surface of the platter when writing data. While reading data, the head detects the magnetic polarization on the surface of the platter. During reads and writes, the R/W head senses the magnetic polarization and never touches the surface of the platter. When the spindle is rotating, there is a microscopic air gap maintained between the R/W heads and the platters, known as the head flying height. This air gap is removed when the spindle stops rotating and the R/W head rests on a special area on the platter near the spindle. This area is called the landing zone. The landing zone is coated with a lubricant to reduce friction between the head and the platter.

The logic on the disk drive ensures that heads are moved to the landing zone before they touch the surface. If the drive malfunctions and the R/W head accidentally touches the surface of the platter outside the landing zone, a head crash occurs. In a head crash, the magnetic coating on the platter is scratched and may cause damage to the R/W head. A head crash generally results in data loss.

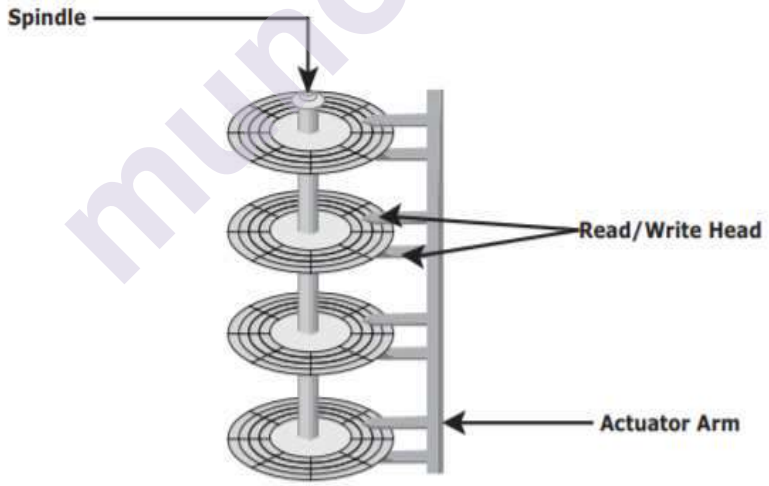



*Figure 2.7: Actuator Arm Assembly*

### 2.8.4 Actuator Arm Assembly

R/W heads are mounted on the actuator arm assembly, which positions the R/W head at the location on the platter where the data needs to be written or read (refer to Figure 2-7). The R/W heads for all platters on a drive are attached to one actuator arm assembly and move across the platters simultaneously.

### 2.8.5 Drive Controller Board

The controller (refer to Figure 2-5 [b]) is a printed circuit board, mounted at the bottom of a disk drive. It consists of a microprocessor, internal memory, circuitry, and firmware. The firmware controls the power to the spindle motor and the speed of the motor. It also manages the communication between the drive and the host. In addition, it controls the R/W operations by moving the actuator arm and switching between different R/W heads and performs the optimization of data access.

### 2.8.6 Physical Disk Structure

Data on the disk is recorded on tracks, which are concentric rings on the platter around the spindle, as shown in Figure 2-8. The tracks are numbered, starting from zero, from the outer edge of the platter. The number of tracks per inch (TPI) on the platter (or the track density) measures how tightly the tracks are packed on a platter. Each track is divided into smaller units called sectors. A sector is the smallest, individually addressable unit of storage. The track and sector structure is written on the platter by the drive manufacturer using a low-level formatting operation. The number of sectors per track varies according to the drive type. The first personal computer disks had 17 sectors per track. Recent disks have a much larger number of sectors on a single track. There can be thousands of tracks on a platter, depending on the physical dimensions and recording density of the platter.

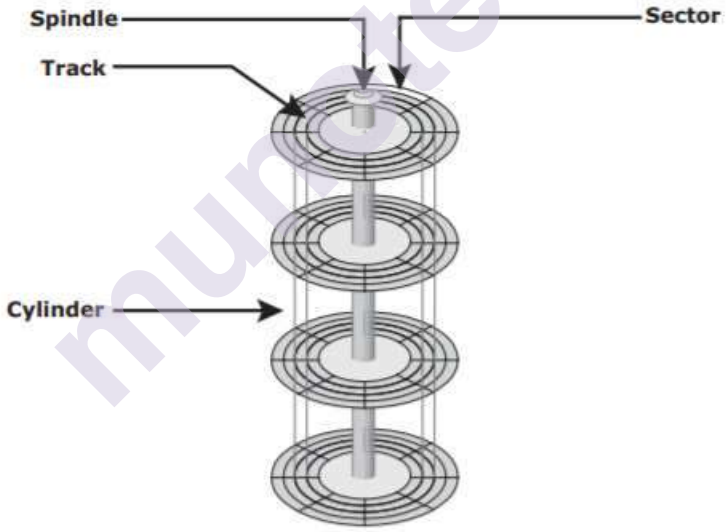



***Figure 2.8: Disk Structure: Sectors, Tracks and Cylinders***

Typically, a sector holds 512 bytes of user data, although some disks can be formatted with larger sector sizes. In addition to user data, a sector also stores other information, such as the sector number, head number or platter number, and track number. This information helps the controller to locate the data on the drive. A cylinder is a set of identical tracks on both surfaces of each drive platter. The location of R/W heads is referred to by the cylinder number, not by the track number.

### 2.8.7 Zone Bit Recording

Platters are made of concentric tracks; the outer tracks can hold more data than the inner tracks because the outer tracks are physically longer than the inner tracks. On older disk drives, the outer tracks had the same number of sectors as the inner tracks, so data density was low on the outer tracks. This was an inefficient use of the available space, as shown in Figure 2-9 (a). Zoned bit recording uses the disk efficiently. As shown in Figure 2-9 (b), this mechanism groups tracks into zones based on their distance from the center of the disk. The zones are numbered, with the outermost zone being zone 0. An appropriate number of sectors per track are assigned to each zone, so a zone near the center of the platter has fewer sectors per track than a zone on the outer edge. However, tracks within a particular zone have the same number of sectors.

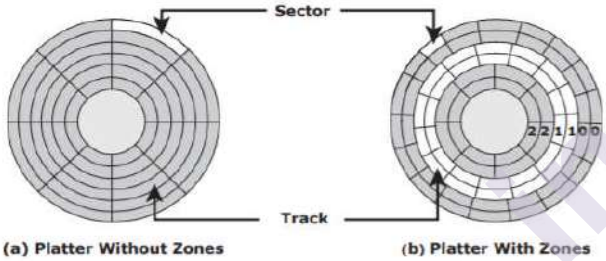



*Figure 2.9: Zoned Bit Recording*

### 2.8.8 Logical Block Addressing

Earlier drives used physical addresses consisting of the cylinder, head, and sector (CHS) number to refer to specific locations on the disk, as shown in Figure 2-10 (a), and the host operating system had to be aware of the geometry of each disk used. Logical block addressing (LBA), as shown in Figure 2-10 (b), simplifies addressing by using a linear address to access physical blocks of data. The disk controller translates LBA to a CHS address, and the host needs to know only the size of the disk drive in terms of the number of blocks. The logical blocks are mapped to physical sectors on a 1:1 basis.

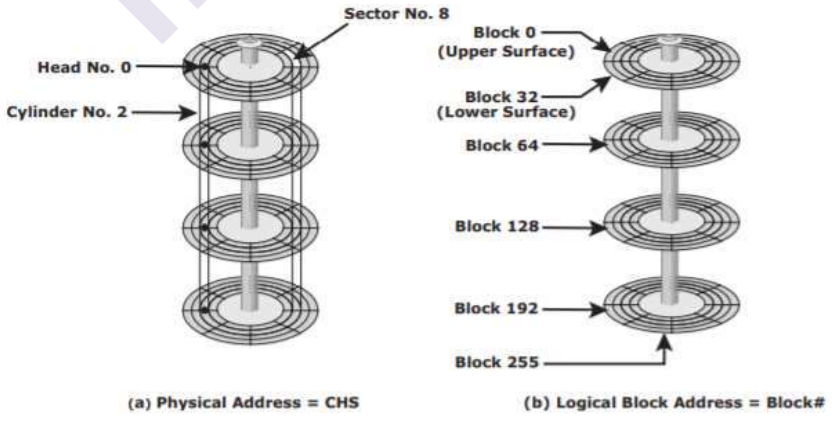



*Figure 2.10: Physical Address and Logical Block Address*

In Figure 2-10 (b), the drive shows eight sectors per track, eight heads, and four cylinders. This means a total of 8 × 8 × 4 = 256 blocks, so

the block number ranges from 0 to 255. Each block has its own unique address. Assuming that the sector holds 512 bytes, a 500 GB drive with a formatted capacity of 465.7 GB has in excess of 976,000,000 blocks.

## 2.9 DISK DRIVE PERFORMANCE

A disk drive is an electromechanical device that governs the overall performance of the storage system environment. The various factors that affect the performance of disk drives are discussed in this section.

### 2.9.1 Disk Service Time

Disk service time is the time taken by a disk to complete an I/O request. Components that contribute to the service time on a disk drive are seek time, rotational latency, and data transfer rate.

#### *Seek Time*

The seek time (also called access time) describes the time taken to position the R/W heads across the platter with a radial movement (moving along the radius of the platter). In other words, it is the time taken to position and settle the arm and the head over the correct track. Therefore, the lower the seek time, the faster the I/O operation. Disk vendors publish the following seek time specifications:

1. **Full Stroke**: The time taken by the R/W head to move across the entire width of the disk, from the innermost track to the outermost track.
2. **Average**: The average time taken by the R/W head to move from one random track to another, normally listed as the time for one-third of a full stroke.
3. **Track-to-Track**: The time taken by the R/W head to move between adjacent tracks.

Each of these specifications is measured in milliseconds. The seek time of a disk is typically specified by the drive manufacturer. The average seek time on a modern disk is typically in the range of 3 to 15 milliseconds. Seek time has more impact on the read operation of random tracks rather than adjacent tracks. To minimize the seek time, data can be written to only a subset of the available cylinders. This results in lower usable capacity than the actual capacity of the drive. For example, a 500 GB disk drive is set up to use only the first 40 percent of the cylinders and is effectively treated as a 200 GB drive. This is known as short-stroking the drive.

#### *Rotational Latency*

To access data, the actuator arm moves the R/W head over the platter to a particular track while the platter spins to position the requested sector under the R/W head. The time taken by the platter to rotate and position the data under the R/W head is called rotational latency. This

latency depends on the rotation speed of the spindle and is measured in milliseconds. The average rotational latency is one-half of the time taken for a full rotation. Similar to the seek time, rotational latency has more impact on the reading/writing of random sectors on the disk than on the same operations on adjacent sectors. Average rotational latency is approximately 5.5 ms for a 5,400-rpm drive, and around 2.0 ms for a 15,000-rpm (or 250-rps revolution per second) drive as shown here:
Average rotational latency for a 15,000 rpm (or 250 rps)
drive = 0.5/250 = 2 milliseconds

***Data Transfer Rate***

The data transfer rate (also called transfer rate) refers to the average amount of data per unit time that the drive can deliver to the HBA. It is important to first understand the process of read/write operations to calculate data transfer rates. In a read operation, the data first moves from disk platters to R/W heads; then it moves to the drive's internal buffer. Finally, data moves from the buffer through the interface to the host HBA. In a write operation, the data moves from the HBA to the internal buffer of the disk drive through the drive's interface. The data then moves from the buffer to the R/W heads. Finally, it moves from the R/W heads to the platters. The data transfer rates during the R/W operations are measured in terms of internal and external transfer rates, as shown in Figure 2-11.

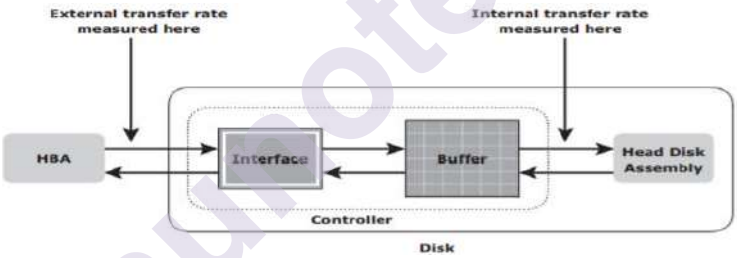



***Figure 2.11: Data Transfer Rate***

Internal transfer rate is the speed at which data moves from a platter's surface to the internal buffer (cache) of the disk. The internal transfer rate takes into account factors such as the seek time and rotational latency. External transfer rate is the rate at which data can move through the interface to the HBA. The external transfer rate is generally the advertised speed of the interface, such as 133 MB/s for ATA. The sustained external transfer rate is lower than the interface speed.

### 2.9.2 Disk I/O Controller Utilization

Utilization of a disk I/O controller has a significant impact on the I/O response time. To understand this impact, consider that a disk can be viewed as a black box consisting of two elements: n Queue: The location where an I/O request waits before it is processed by the I/O controller n Disk I/O Controller: Processes I/Os waiting in the queue one by one The I/O requests arrive at the controller at the rate generated by the application. This rate is also called the arrival rate. These requests are held in the I/O queue, and the I/O controller processes them one by one, as

shown in Figure 2-12. The I/O arrival rate, the queue length, and the time taken by the I/O controller to process each request determines the I/O response time. If the controller is busy or heavily utilized, the queue size will be large and the response time will be high.


***Figure 2.12: I/O Processing***

Based on the fundamental laws of disk drive performance, the relationship between controller utilization and average response time is given as Average response time (TR) = Service time (TS ) / (1 – Utilization) where TS is the time taken by the controller to serve an I/O. As the utilization reaches 100 percent — that is, as the I/O controller saturates — the response time is closer to infinity. In essence, the saturated component, or the bottleneck, forces the serialization of I/O requests, meaning that each I/O request must wait for the completion of the I/O requests that preceded it. Figure 2-13 shows a graph plotted between utilization and response time.

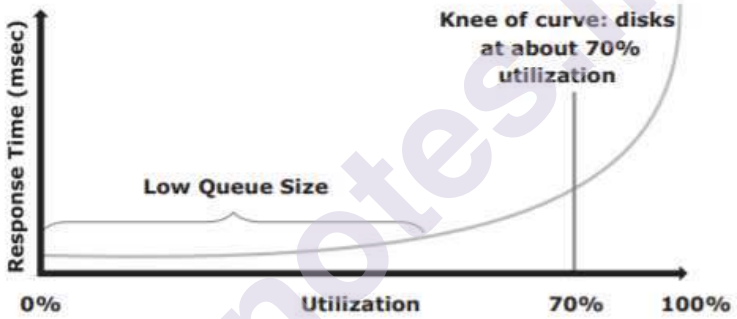



***Figure 2.13: Utilization Vs. Response Time***

The graph indicates that the response time changes are nonlinear as the utilization increases. When the average queue sizes are low, the response time remains low. The response time increases slowly with added load on the queue and increases exponentially when the utilization exceeds 70 percent. Therefore, for performance-sensitive applications, it is common to utilize disks below their 70 percent of I/O serving capability.

## 2.10 HOST ACCESS TO DATA

Data is accessed and stored by applications using the underlying infrastructure. The key components of this infrastructure are the operating system (or file system), connectivity, and storage. The storage device can be internal and (or) external to the host. In either case, the host controller card accesses the storage devices using predefined protocols, such as IDE/ATA, SCSI, or Fibre Channel (FC). IDE/ATA and SCSI are popularly used in small and personal computing environments for accessing internal storage. FC and iSCSI protocols are used for accessing data from an external storage device (or subsystems). External storage devices can be connected to the host directly or through the storage

network. When the storage is connected directly to the host, it is referred as direct-attached storage (DAS), which is detailed later in this chapter. Understanding access to data over a network is important because it lays the foundation for storage networking technologies. Data can be accessed over a network in one of the following ways: block level, file level, or object level. In general, the application requests data from the file system (or operating system) by specifying the filename and location. The file system maps the file attributes to the logical block address of the data and sends the request to the storage device. The storage device converts the logical block address (LBA) to a cylinder-head-sector (CHS) address and fetches the data. In a block-level access, the file system is created on a host, and data is accessed on a network at the block level, as shown in Figure 2-14 (a). In this case, raw disks or logical volumes are assigned to the host for creating the file system. In a fi le-level access, the file system is created on a separate file server or at the storage side, and the file-level request is sent over a network, as shown in Figure 2-14 (b). Because data is accessed at the file level, this method has higher overhead, as compared to the data accessed at the block level. Object-level access is an intelligent evolution, whereby data is accessed over a network in terms of self-contained objects with a unique object identifier.

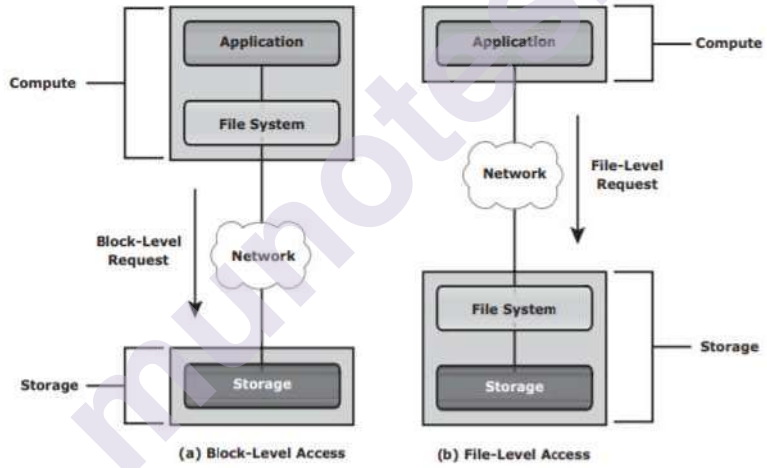



***Figure 2.14: Host Access to Storage***

## 2.11 DIRECT ATTACHED STORAGE

DAS is an architecture in which storage is connected directly to the hosts. The internal disk drive of a host and the directly-connected external storage array are some examples of DAS. Although the implementation of storage networking technologies is gaining popularity, DAS has remained suitable for localized data access in a small environment, such as personal computing and workgroups. DAS is classified as internal or external, based on the location of the storage device with respect to the host. In internal DAS architectures, the storage device is internally connected to the host by a serial or parallel bus (see Figure 2-15 [a]). The physical bus has distance limitations and can be sustained only over a shorter distance for highspeed connectivity. In addition, most internal buses can support

only a limited number of devices, and they occupy a large amount of space inside the host, making maintenance of other components difficult. On the other hand, in external DAS architectures, the host connects directly to the external storage device, and data is accessed at the block level (see Figure 2-15 [b]). In most cases, communication between the host and the storage device takes place over a SCSI or FC protocol. Compared to internal DAS, an external DAS overcomes the distance and device count limitations and provides centralized management of storage devices.

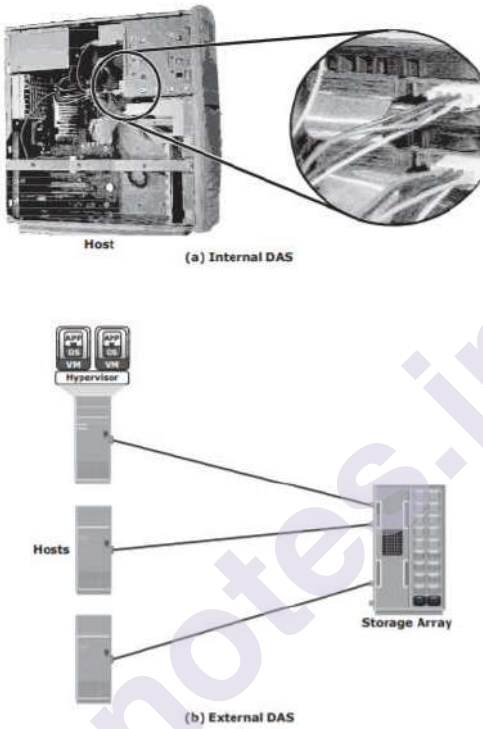



***Figure 2.15: Internal and External DAS Architecture***

### 2.11.1 DAS Benefits and Limitations

DAS requires a relatively lower initial investment than storage networking architectures. The DAS configuration is simple and can be deployed easily and rapidly. The setup is managed using host-based tools, such as the host OS, which makes storage management tasks easy for small environments. Because DAS has a simple architecture, it requires fewer management tasks and less hardware and software elements to set up and operate. However, DAS does not scale well. A storage array has a limited number of ports, which restricts the number of hosts that can directly connect to the storage. When capacities are reached, the service availability may be compromised. DAS does not make optimal use of resources due to its limited capability to share front-end ports. In DAS environments, unused resources cannot be easily reallocated, resulting in islands of over-utilized and under-utilized storage pools.

## 2.12 STORAGE DESIGN BASED ON APPLICATION REQUIREMENTS AND DISK PERFORMANCE

Determining storage requirements for an application begins with determining the required storage capacity. This is easily estimated by the size and number of file systems and database components used by applications. The I/O size, I/O characteristics, and the number of I/Os generated by the application at peak workload are other factors that affect disk performance, I/O response time, and design of storage systems. The I/O block size depends on the file system and the database on which the application is built. Block size in a database environment is controlled by the underlying database engine and the environment variables. The disk service time (TS ) for an I/O is a key measure of disk performance; TS , along with disk utilization rate (U), determines the I/O response time for an application. As discussed earlier in this chapter, the total disk service time (TS ) is the sum of the seek time (T), rotational latency (L), and internal transfer time (X):

TS = T + L + X

Consider an example with the following specifications provided for a disk:

- The average seek time is 5 ms in a random I/O environment; therefore, T = 5 ms.
- Disk rotation speed of 15,000 revolutions per minute or 250 revolutions per second — from which rotational latency (L) can be determined, which is one-half of the time taken for a full rotation or L = (0.5/250 rps expressed in ms).
- 40 MB/s internal data transfer rate, from which the internal transfer time (X) is derived based on the block size of the I/O — for example, an I/O with a block size of 32 KB; therefore X = 32 KB/40 MB.

Consequently, the time taken by the I/O controller to serve an I/O of block size 32 KB is (TS ) = 5 ms + (0.5/250) + 32 KB/40 MB = 7.8 ms.

Therefore, the maximum number of I/Os serviced per second or IOPS is (1/TS ) = 1/(7.8 × 10-3 ) = 128 IOPS.

Table 2-1 lists the maximum IOPS that can be serviced for different block sizes using the previous disk specifications.

*Table 2.1: IOPS Performed by Disk Drive*

| BLOCK SIZE | $T_s = T + L + X$ | $IOPS = 1/T_s$ |
|---|---|---|
| 4 KB | 5 ms + (0.5/250 rps) + 4 K/40 MB = 5 + 2 + 0.1 = 7.1 | 140 |
| 8 KB | 5 ms + (0.5/250 rps) + 8 K/40 MB = 5 + 2 + 0.2 = 7.2 | 139 |
| 16 KB | 5 ms + (0.5/250 rps) + 16 K/40 MB = 5 + 2 + 0.4 = 7.4 | 135 |
| 32 KB | 5 ms + (0.5/250 rps) + 32 K/40 MB = 5 + 2 + 0.8 = 7.8 | 128 |
| 64 KB | 5 ms + (0.5/250 rps) + 64 K/40 MB = 5 + 2 + 1.6 = 8.6 | 116 |

The IOPS ranging from 116 to 140 for different block sizes represents the IOPS that can be achieved at potentially high levels of utilization (close to 100 percent). As discussed in Section 2.7.2, the application response time, R, increases with an increase in disk controller utilization. For the same preceding example, the response time (R) for an I/O with a block size of 32 KB at 96 percent disk controller utilization is
$R = TS /(1 – U) = 7.8/(1 – 0.96) = 195$ ms

If the application demands a faster response time, then the utilization for the disks should be maintained below 70 percent. For the same 32-KB block size, at 70-percent disk utilization, the response time reduces drastically to 26 ms. However, at lower disk utilization, the number of IOPS a disk can perform is also reduced. In the case of a 32-KB block size, a disk can perform 128 IOPS at almost 100 percent utilization, whereas the number of IOPS it can perform at 70-percent utilization is 89 (128 x 0.7). This indicates that the number of I/Os a disk can perform is an important factor that needs to be considered while designing the storage requirement for an application.

Therefore, the storage requirement for an application is determined in terms of both the capacity and IOPS. If an application needs 200 GB of disk space, then this capacity can be provided simply with a single disk. However, if the application IOPS requirement is high, then it results in performance degradation because just a single disk might not provide the required response time for I/O operations.

Based on this discussion, the total number of disks required (DR) for an application is computed as follows:
$D_R = \text{Max}\ (D_C, D_I)$

Where DC is the number of disks required to meet the capacity, and DI is the number of disks required to meet the application IOPS requirement. Let's understand this with the help of an example.

**Example:** Consider an example in which the capacity requirement for an application is 1.46 TB. The number of IOPS generated by the application at peak workload is estimated at 9,000 IOPS. The vendor specifies that a 146-GB, 15,000-rpm drive is capable of doing a maximum 180 IOPS. In this example, the number of disks required to meet the capacity requirements will be 1.46 TB/146 GB = 10 disks.

To meet the application IOPS requirements, the number of disks required is 9,000/180 = 50. However, if the application is response-time sensitive, the number of IOPS a disk drive can perform should be calculated based on 70- percent disk utilization. Considering this, the number of IOPS a disk can perform at 70 percent utilization is 180 * 0.7 = 126 IOPS. Therefore, the number of disks required to meet the application IOPS requirement will be 9,000/126 = 72.

As a result, the number of disks required to meet the application requirements will be Max (10, 72) = 72 disks.

The preceding example indicates that from a capacity-perspective, 10 disks are sufficient; however, the number of disks required to meet application performance is 72. To optimize disk requirements from a performance perspective, various solutions are deployed in a real-time environment. Examples of these solutions are disk native command queuing, use of fl ash drives, RAID, and the use of cache memory.

## 2.13 DISK NATIVE COMMAND QUEUING

Command queuing is a technique implemented on modern disk drives that determines the execution order of received I/Os and reduces unnecessary drive-head movements to improve disk performance. When an I/O is received for execution at the disk controller, the command queuing algorithms assign a tag that defines a sequence in which the commands should be executed. With command queuing, commands are executed based on the organization of data on the disk, regardless of the order in which the commands are received. The commonly used algorithm for command queuing is seek time optimization. Commands are executed based on optimizing read/write head movements, which might result in the reordering of commands. Without seek time optimization, the commands are executed in the order they are received. For example, as shown in Figure 2-16 (a), the commands are executed in the order A, B, C, and D. The radial movement required by the head to execute C immediately after A is less than what would be required to execute B. With seek time optimization, the command execution sequence would be A, C, B, and D, as shown in Figure 2-16 (b).

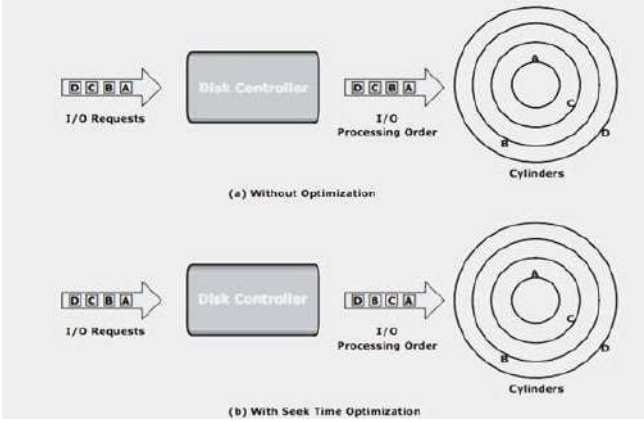



***Figure 2.16: Disk Command Queuing***

Access Time Optimization is another command queuing algorithm. With this algorithm, commands are executed based on the combination of seek time optimization and an analysis of rotational latency for optimal performance. Command queuing is also implemented on modern storage array controllers, which might further supplement the command queuing implemented on the disk drive.

## 2.14 INTRODUCTION TO FLASH DRIVE

With the growth of information, storage users continue to demand ever-increasing performance requirements for their business applications. Traditionally, high I/O requirements were met by simply using more disks. Availability of enterprise class fl ash drives (EFD) has changed the scenario. Flash drives, also referred as solid-state drives (SSDs), are new generation drives that deliver ultra-high performance required by performance-sensitive applications. Flash drives use semiconductor-based solid-state memory (fl ash memory) to store and retrieve data. Unlike conventional mechanical disk drives, flash drives contain no moving parts; therefore, they do not have seek and rotational latencies. Flash drives deliver a high number of IOPS with very low response times. Also, being a semiconductor-based device, flash drives consume less power, compared to mechanical drives. Flash drives are especially suited for applications with small block size and random-read workloads that require consistently low (less than 1 millisecond) response times. Applications that need to process massive amounts of information quickly, such as currency exchange, electronic trading systems, and real-time data feed processing benefit from flash drives. Compared to conventional mechanical disk drives, EFD provides up to 30 times the throughput and up to one-tenth the response time (<1ms compared with 6-10 ms). In addition, flash drives can store data using up to 38 percent less energy per TB than traditional disk drives, which translates into approximately 98 percent less power consumption per I/O.

Overall, flash drives provide better total cost of ownership (TCO) even though they cost more on $/GB basis. By implementing flash drives, businesses can meet application performance requirements with far fewer drives (approximately 20 to 30 times less number of drives compared to conventional mechanical drives). This reduction not only provides savings in terms of drive cost, but also translates to savings for power, cooling, and space consumption. Fewer numbers of drives in the environment also means less cost for managing the storage.

### 2.14.1 Components and Architecture of Flash Drives

Flash drives use similar physical form factor and connectors as mechanical disk drives to maintain compatibility. This enables easy replacement of a mechanical disk drive with a fl ash drive in a storage array enclosure. The key components of a fl ash drive are the controller, I/O interface, mass storage (collection of memory chips), and cache. The controller manages the functioning of the drive, and the I/O interface

provides power and data access. Mass storage is an array of nonvolatile NAND (negated AND) memory chips used for storing data. Cache serves as a temporary space or buffer for data transaction and operations. A fl ash drive uses multiple parallel I/O channels (from its drive controller to the fl ash memory chips) for data access. Generally, the larger the number of fl ash memory chips and channels, the higher the drive's internal bandwidth, and ultimately the higher the drive's performance. Flash drives typically have eight to 24 channels. Memory chips in fl ash drives are logically organized in blocks and pages. A page is the smallest object that can be read or written on a fl ash drive. Pages are grouped together into blocks. (These blocks should not be confused with the 512-byte blocks in mechanical disk drive sectors.) A block may have 32, 64, or 128 pages. Pages do not have a standard size; typical page sizes are 4 KB, 8 KB, and 16 KB. Because fl ash drives emulate mechanical drives that use logical block addresses (LBAs), a page spans across a consecutive series of data blocks. For example, a 4-KB page would span across eight 512-byte data blocks with consecutive addresses. In fl ash drives, a read operation can happen at the page level, whereas a write or an erase operation happens only at the block level.

### 2.14.2 Features of Enterprise Flash Drives

The key features of enterprise class fl ash drives are as follows: n NAND fl ash memory technology:

1. NAND memory technology is well suited for accessing random data. A NAND device uses bad block tracking and error-correcting code (ECC) to maintain data integrity and provide the fastest write speeds.
2. Single-Level Cell (SLC)-based fl ash: NAND technology is available in two different cell designs. A multi-level cell (MLC) stores more than one bit per cell by virtue of its capability to register multiple states, versus a single-level cell that can store only 1 bit. SLC is the preferred technology for enterprise data applications due to its performance and longevity. SLC read speeds are typically rated at twice those of MLC devices, and write speeds are up to four times higher. SLC devices typically have 10 times higher write erase cycles, compared to MLC designs. In addition, the SLC fl ash memory has higher reliability because it stores only 1 bit per cell. Hence, the likelihood for error is reduced.
3. Write levelling technique: An important element of maximizing a fl ash drive's useful life is ensuring that the individual memory cells experience uniform use. This means that data that is frequently updated is written to different locations to avoid rewriting the same cells. In EFDs, the device is designed to ensure that with any new write operation, the youngest block is used.

## 2.15 CONCEPT IN PRACTICE: VMWARE ESXI

VMware is the leading provider for a server virtualization solution. VMware ESXi provides a platform called hypervisor. The hypervisor abstracts CPU, memory, and storage resources to run multiple virtual machines concurrently on the same physical server. VMware ESXi is a hypervisor that installs on x86 hardware to enable server virtualization. It enables creating multiple virtual machines (VMs) that can run simultaneously on the same physical machine. A VM is a discrete set of files that can be moved, copied, and used as a template. All the files that make up a VM are typically stored in a single directory on a cluster file system called Virtual Machine File System (VMFS). The physical machine that houses ESXi is called the ESXi host. ESXi hosts provide physical resources used to run virtual machines. ESXi has two key components: VMkernel and Virtual Machine Monitor. VMkernel provides functionality similar to that found in other operating systems, such as process creation, file system management, and process scheduling. It is designed to specifically support running multiple VMs and provide core functionality such as resource scheduling, I/O stacks, and so on. The virtual machine monitor is responsible for executing commands on the CPUs and performing Binary Translation (BT). A virtual machine monitor performs hardware abstraction to appear as a physical machine with its own CPU, memory, and I/O devices. Each VM is assigned a virtual machine monitor that has a share of the CPU, memory, and I/O devices to successfully run the VM.

## 2.16 SUMMARY

- This chapter detailed the key elements of a data center environment — application, DBMS, host, connectivity, and storage.
- The data flows from an application to storage through these elements.
- Physical and logical components of these entities affect the overall performance of the application.
- Virtualization at different components of the data centre provides better utilization and management of these components.
- Storage is a core component in the data centre environment. The disk drive is the most popular storage device that uses magnetic media for accessing and storing data.
- Flash-based solid-state drives (SSDs) are a recent innovation, and in many ways, superior to mechanical disk drives.
- Modern disk storage systems use hundreds of disks to meet application performance requirements.
- Managing the capacity, performance, and reliability of these large numbers of disks poses significant challenges.

## 2.17 REVIEW YOUR LEARNING

- Can you explain hard disk structure?
- Can you explain what is Seek Time, Transfer Time, Access Time of data?
- Are you able to Virtualization?
- Explain Physical connectivity components used in data centre.
- Can you relate day to day data usage by application on real time basis?

## 2.18 QUESTIONS

1. Explain disk drive components with neat, labelled diagram.
2. Explain Direct Attached Storage. Explain its limitations and benefits.
3. Explain Disk Native Command Queuing.
4. Explain Components and Architectures of Flash Drives.
5. Explain working with VmwareEsxi.
6. What are the advantages of a virtualized data centre over a classic data centre?
7. An application specifies a requirement of 200 GB to host a database and other files. It also specifies that the storage environment should support 5,000 IOPS during its peak workloads. The disks available for configuration provide 66 GB of usable capacity, and the manufacturer specifies that they can support a maximum of 140 IOPS. The application is response time sensitive, and disk utilization beyond 60 percent does not meet the response time requirements. Compute and explain the theoretical basis for the minimum number of disks that should be configured to meet the requirements of the application.
8. Which components constitute the disk service time? Which component contributes the largest percentage of the disk service time in a random I/O operation?
9. The average I/O size of an application is 64 KB. The following specifications are available from the disk manufacturer: average seek time = 5 ms, 7,200 RPM, and transfer rate = 40 MB/s. Determine the maximum IOPS that could be performed with this disk for the application. Using this case as an example, explain the relationship between disk utilization and IOPS.
10. Refer to Question No. 9 based on the calculated disk service time, plot a graph showing the response time versus utilization, considering the utilization of the I/O controller at 20 percent, 40 percent, 60 percent, 80 percent, and 100 percent. Describe the conclusion that could be derived from the graph.
11. Research other elements of a data centre besides the core elements discussed in this chapter, including environmental control parameters such as HVAC (heat, ventilation, and air-condition), power supplies, and security

## 2.19 FURTHER READING

- http://aad.tpu.ru/practice/EMC/Information%20Storage%20and%20Management-v.2.pdf
- https://nptel.ac.in/courses/106/108/106108058/
- https://nptel.ac.in/content/storage2/courses/106108058/lec%2007.pdf
- http://www.ictacademy.in/pages/Information-Storage-and-Management.aspx
- https://www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwiM8Kq6isHyAhUEkmYCHbJyDXAYABAAGgJzbQ&ae=2&ohost=www.google.com&cid=CAESQeD28QNmzUxhr6qtgEwm24g2Yc-TeMC_24a0sxeZf9MitA7QrS5Vz4VE3XfWSwFvX0iAKPoH4fT4QmSj7PhnMAQF&sig=AOD64_1Y3y_5vJpAZOJybqnNONsE6wNayQ&q&adurl&ved=2ahUKEwjvsaG6isHyAhXjxTgGHTvKBEEQ0Qx6BAgDEAE
- https://www.coursera.org/lecture/big-data-management/data-storage-RplBY
- https://www.coursera.org/courses?query=data%20storage
- https://www.coursera.org/lecture/technical-support-fundamentals/storage-RLNIZ
- https://www.coursera.org/learn/cloud-storage-big-data-analysis-sql

## 2.20 REFERENCES


1. Information Storage and Management: Storing, Managing and Protecting Digital Information in Classic, Virtualized and Cloud Environments, EMC, John & Wiley Sons, 2nd Edition, 2012
2. Information Storage and Management, Pankaj Sharma
3. Information Technology Project Management, Jack T Marchewka
4. Information Storage and Management, I A Dhotre


# 3

# DATA PROTECTION

**Unit Structure**





## 3.0 OBJECTIVES

1. Explain basic data storage options and its components
2. Analyse data protection mechanisms using various RAID levels.

## 3.1 INTRODUCTION

In the late 1980s, rapid adoption of computers for business processes stimulated the growth of new applications and databases, significantly increasing the demand for storage capacity and performance. At that time, 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 because they could serve only a limited number of I/Os. Today's data centres house hundreds of disk drives in their storage infrastructure. Disk drives are inherently susceptible to failures due to mechanical wear and tear and other environmental factors, which could result in data loss. The greater the number of disk drives in a storage array, the greater the probability of a disk failure in the array. For example, consider a storage array of 100 disk drives, each with an average life expectancy of 750,000 hours. The average life expectancy of this collection in the array, therefore, is 750,000/100 or 7,500 hours. This means that a disk drive in this array is likely to fail at least once in 7,500 hours. RAID is an enabling technology that leverages multiple drives as part of a set that provides data protection against drive failures. In general, RAID implementations also improve the storage system performance by serving I/Os from multiple disks simultaneously. Modern arrays with flash drives also benefit in terms of protection and performance by using RAID. In 1987, Patterson, Gibson, and Katz at the University of California, Berkeley, published a paper titled "A Case for Redundant Arrays of Inexpensive Disks (RAID)." This paper described the use of small-capacity, inexpensive disk drives as an alternative to large-capacity drives common on mainframe computers. The term RAID has been redefined to refer to independent disks to reflect advances in the storage technology. RAID technology has now grown from an academic concept to an industry standard and is common implementation in today's storage arrays. This chapter details RAID technology, RAID levels, and different types of RAID implementations and their benefits.

## 3.2 RAID IMPLEMENTATION METHODS

The two methods of RAID implementation are hardware and software. Both have their advantages and disadvantages and are discussed in this section.

### 3.2.1 Software RAID

Software RAID uses host-based software to provide RAID functions. It is implemented at the operating-system level and does not use a dedicated hardware controller to manage the RAID array. Software RAID implementations offer cost and simplicity benefits when compared with hardware RAID. However, they have the following limitations:

1. **Performance**: Software RAID affects overall system performance. This is due to additional CPU cycles required to perform RAID calculations.
2. **Supported features**: Software RAID does not support all RAID levels.
3. **Operating system compatibility**: Software RAID is tied to the host operating system; hence, upgrades to software RAID or to the operating system should be validated for compatibility. This leads to inflexibility in the data-processing environment.

### 3.2.2 Hardware RAID

In hardware RAID implementations, a specialized hardware controller is implemented either on the host or on the array. Controller card RAID is a host-based hardware RAID implementation in which a specialized RAID controller is installed in the host, and disk drives are connected to it. Manufacturers also integrate RAID controllers on motherboards. A host-based RAID controller is not an efficient solution in a data center environment with a large number of hosts. The external RAID controller is an array-based hardware RAID. It acts as an interface between the host and disks. It presents storage volumes to the host, and the host manages these volumes as physical drives. The key functions of the RAID controllers are as follows:

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

## 3.3 RAID ARRAY COMPONENTS

A RAID array is an enclosure that contains several disk drives and supporting hardware to implement RAID. A subset of disks within a RAID array can be grouped to form logical associations called logical arrays, also known as a RAID, set or a RAID group (see Figure 3.1).

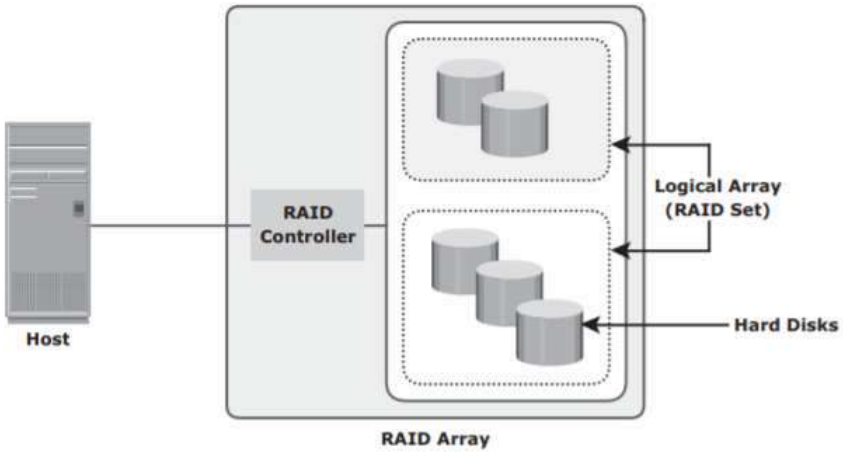



*Figure 3.1: Components od RAID array*

## 3.4 RAID TECHNIQUES

RAID techniques — striping, mirroring, and parity — form the basis for defining various RAID levels. These techniques determine the data availability and performance characteristics of a RAID set.

### 3.4.1 Striping

Striping is a technique to spread data across multiple drives (more than one) to use the drives in parallel. All the read-write heads work simultaneously, allowing more data to be processed in a shorter time and increasing performance, compared to reading and writing from a single disk. Within each disk in a RAID set, a predefined number of contiguously addressable disk blocks are defined as a strip. The set of aligned strips that spans across all the disks within the RAID set is called a stripe. Figure 3-2 shows physical and logical representations of a striped RAID set.

Strip size (also called stripe depth) describes the number of blocks in a strip and is the maximum amount of data that can be written to or read from a single disk in the set, if the accessed data starts at the beginning of the strip. All strips in a stripe have the same number of blocks. Having a smaller strip size means that data is broken into smaller pieces while spread across the disks.

Stripe size is a multiple of strip size by the number of data disks in the RAID set. For example, in a five-disk striped RAID set with a strip size of 64 KB, the stripe size is 320 KB(64KB * 5). Stripe width refers to the number of data strips in a stripe. Striped RAID does not provide any data protection unless parity or mirroring is used, as discussed in the following sections.

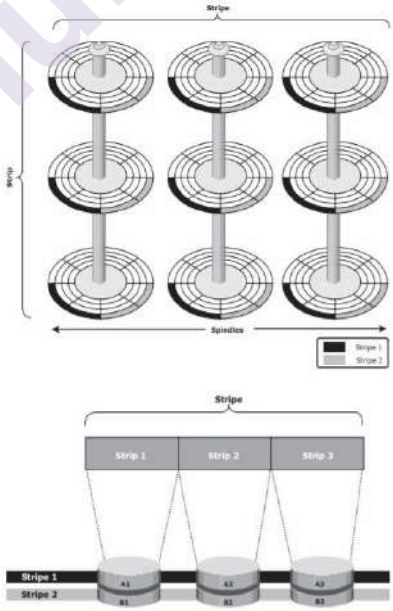



*Figure 3.2: Striped RAID Set*

### 3.4.2 Mirroring

Mirroring is a technique whereby the same data is stored on two different disk drives, yielding two copies of the data. If one disk drive failure occurs, the data is intact on the surviving disk drive (see Figure 3-3) and the controller continues to service the host's data requests from the surviving disk of a mirrored pair. When the failed disk is replaced with a new disk, the controller copies the data from the surviving disk of the mirrored pair. This activity is transparent to the host. In addition to providing complete data redundancy, mirroring enables fast recovery from disk failure. However, disk mirroring provides only data protection and is not a substitute for data backup. Mirroring constantly captures changes in the data, whereas a backup captures point-in-time images of the data. Mirroring involves duplication of data — the amount of storage capacity needed is twice the amount of data being stored. Therefore, mirroring is considered expensive and is preferred for mission-critical applications that cannot afford the risk of any data loss. Mirroring improves read performance because read requests can be serviced by both disks. However, write performance is slightly lower than that in a single disk because each write request manifests as two writes on the disk drives. Mirroring does not deliver the same levels of write performance as a striped RAID.

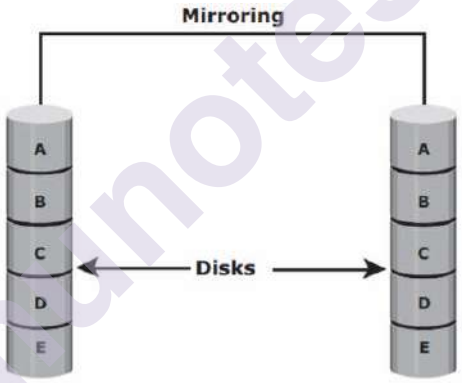



***Figure 3.3: Mirrored Disks in an array***

### 3.4.3 Parity

Parity is a method to protect striped data from disk drive failure without the cost of mirroring. An additional disk drive is added to hold parity, a mathematical construct that allows re-creation of the missing data. Parity is a redundancy technique that ensures protection of data without maintaining a full set of duplicate data. Calculation of parity is a function of the RAID controller.

Parity information can be stored on separate, dedicated disk drives or distributed across all the drives in a RAID set. Figure 3-4 shows a parity RAID set. The first four disks labelled "Data Disks," contain the data. The fifth disk, labelled "Parity Disk," stores the parity information, which, in this case, is the sum of the elements in each row. Now, if one of

the data disks fails, the missing value can be calculated by subtracting the sum of the rest of the elements from the parity value. Here, for simplicity, the computation of parity is represented as an arithmetic sum of the data. However, parity calculation is a bitwise XOR operation.

A bit-by-bit Exclusive -OR (XOR) operation takes two-bit patterns of equal length and performs the logical XOR operation on each pair of corresponding bits. The result in each position is 1 if the two bits are different, and 0 if they are the same. The truth table of the XOR operation is shown next. (A and B denote the inputs and C, the output after performing the XOR operation.) If any of the data from A, B, or C is lost, it can be reproduced by performing an XOR operation on the remaining available data. For example, if a disk containing all the data from A fails, the data can be regenerated by performing an XOR between B and C.

| A | B | C |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

Compared to mirroring, parity implementation considerably reduces the cost associated with data protection. Consider an example of a parity RAID configuration with five disks where four disks hold data, and the fifth holds the parity information. In this example, parity requires only 25 percent extra disk space compared to mirroring, which requires 100 percent extra disk space. However, there are some disadvantages of using parity. Parity information is generated from data on the data disk. Therefore, parity is recalculated every time there is a change in data. This recalculation is time-consuming and affects the performance of the RAID array. For parity RAID, the stripe size calculation does not include the parity strip. For example in a five (4 + 1) disk parity RAID set with a strip size of 64 KB, the stripe size will be 256 KB (64 KB * 4).

### 3.4.4 RAID Levels

Application performance, data availability requirements, and cost determine the RAID level selection. These RAID levels are defined on the basis of striping, mirroring, and parity techniques. Some RAID levels use a single technique, whereas others use a combination of techniques. Table 3-1 shows the commonly used RAID levels.

***Table 3.1: RAID Levels***

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

### 3.4.4.1 RAID 0

RAID 0 configuration uses data striping techniques, where data is striped across all the disks within a RAID set. Therefore, it utilizes the full storage capacity of a RAID set. To read data, all the strips are put back together by the controller. Figure 3-5 shows RAID 0 in an array in which data is striped across five disks. When the number of drives in the RAID set increases, performance improves because more data can be read or written simultaneously. RAID 0 is a good option for applications that need high I/O throughput. However, if these applications require high availability during drive failures, RAID 0 does not provide data protection and availability.

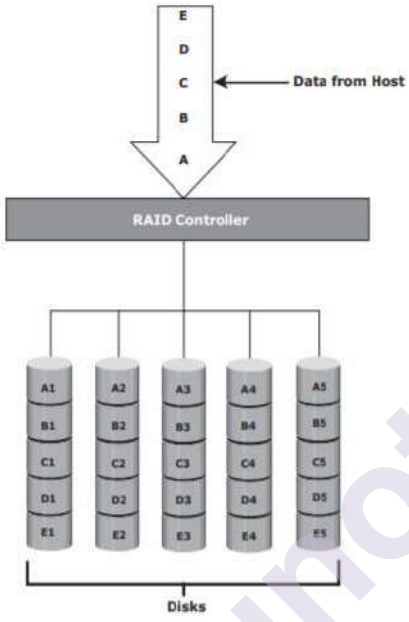



*Figure 3.5: RAID 0*

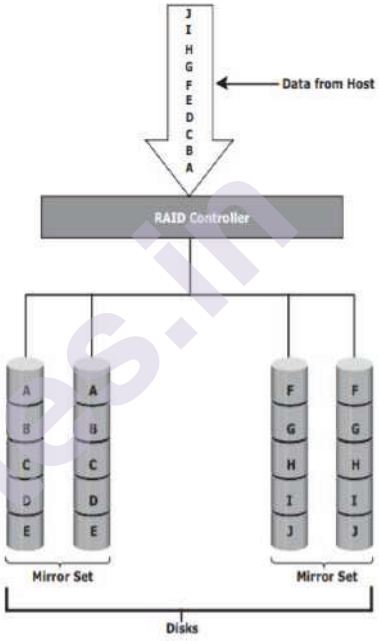



*Figure 3.6: RAID 1*

### 3.4.4.2 RAID 1

RAID 1 is based on the mirroring technique. In this RAID configuration, data is mirrored to provide fault tolerance (see Figure 3-6). A RAID 1 set consists of two disk drives and every write is written to both disks. The mirroring is transparent to the host. During disk failure, the impact on data recovery in RAID 1 is the least among all RAID implementations. This is because the RAID controller uses the mirror drive for data recovery. RAID 1 is suitable for applications that require high availability and cost is no constraint.

### 3.4.4.3 Nested RAID

Most data centers require data redundancy and performance from their RAID arrays. RAID 1+0 and RAID 0+1combine 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).

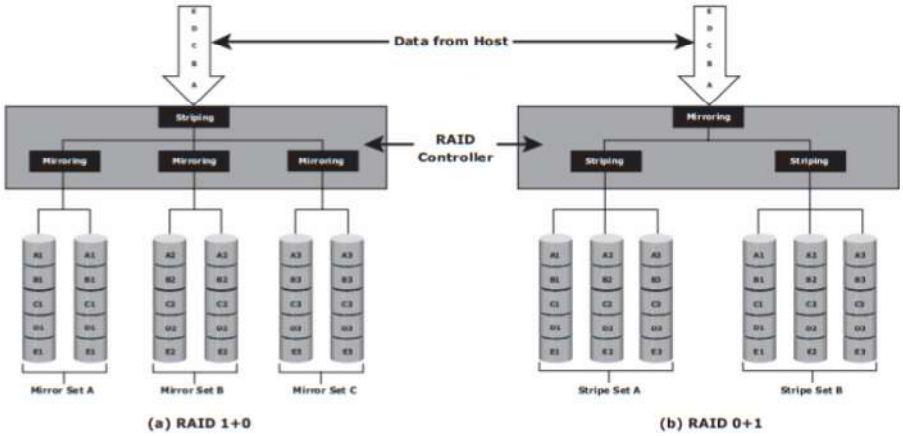



***Figure 3.7: Nested RAID***

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 with small, random, write-intensive I/Os. 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 with write intensive random-access workloads

A common misconception is that RAID 1+0 and RAID 0+1 are the same. Under normal conditions, RAID levels 1+0 and 0+1 offer identical benefits. However, rebuild operations in the case of disk failure differ between the two. 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 the data are striped across multiple disk drive pairs in a RAID set. When replacing a failed drive, only the mirror is rebuilt. In other words, 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. To understand the working of RAID 1+0, consider an example of six disks forming a RAID 1+0 (RAID 1 first and then RAID 0) set. These six disks are paired into three sets of two disks, where each set acts as a RAID 1 set (mirrored pair of disks). Data is then striped across all the three mirrored sets to form RAID 0. Following are the steps performed in RAID 1+0 (see Figure 3-7 [a]):

Drives 1+2 = RAID 1 (Mirror Set A)
Drives 3+4 = RAID 1 (Mirror Set B)
Drives 5+6 = RAID 1 (Mirror Set C)

Now, RAID 0 striping is performed across sets A through C. In this configuration, if drive 5 fails, then the mirror set C alone is affected. It still has drive 6 and continues to function and the entire RAID 1+0 array also keeps functioning. Now, suppose drive 3 fails while drive 5 was being

replaced. In this case the array still continues to function because drive 3 is in a different mirror set. So, in this configuration, up to three drives can fail without affecting the array, as long as they are all in different mirror sets. RAID 0+1 is also called a mirrored stripe. The basic element of RAID 0+1 is a stripe. This means that the process of striping data across disk drives is performed initially, and then the entire stripe is mirrored. In this configuration if one drive fails, then the entire stripe is faulted. Consider the same example of six disks to understand the working of RAID 0+1 (that is, RAID 0 first and then RAID 1). Here, six disks are paired into two sets of three disks each. Each of these sets, in turn, act as a RAID 0 set that contains three disks and then these two sets are mirrored to form RAID 1. Following are the steps performed in RAID 0+1 (see Figure 3-7 [b]):
Drives 1 + 2 + 3 = RAID 0 (Stripe Set A)
Drives 4 + 5 + 6 = RAID 0 (Stripe Set B)

Now, these two stripe sets are mirrored. If one of the drives, say drive 3, fails, the entire stripe set A fails. A rebuild operation copies the entire stripe, copying the data from each disk in the healthy stripe to an equivalent disk in the failed stripe. This causes increased and unnecessary I/O load on the surviving disks and makes the RAID set more vulnerable to a second disk failure.

### 3.4.4.4 RAID 3

RAID 3 stripes data for performance and uses parity for fault tolerance. Parity information is stored on a dedicated drive so that the data can be reconstructed if a drive fails in a RAID set. For example, in a set of five disks, four are used for data and one for parity. Therefore, the total disk space required is 1.25 times the size of the data disks. RAID 3 always reads and writes complete stripes of data across all disks because the drives operate in parallel. There are no partial writes that update one out of many strips in a stripe. Figure 3-8 illustrates the RAID 3 implementation. RAID 3 provides good performance for applications that involve large sequential data access, such as data backup or video streaming.

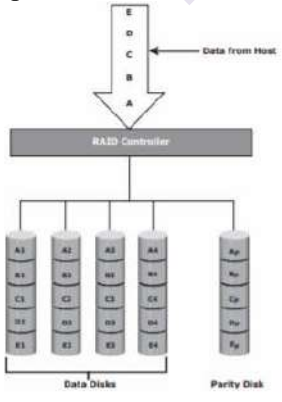



*Figure 3.8: RAID 3*

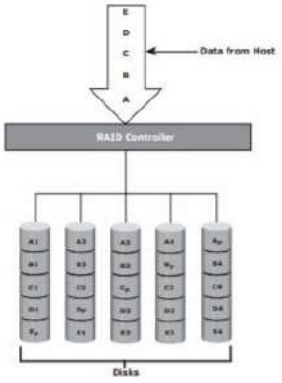



*Figure 3.9: RAID 5*

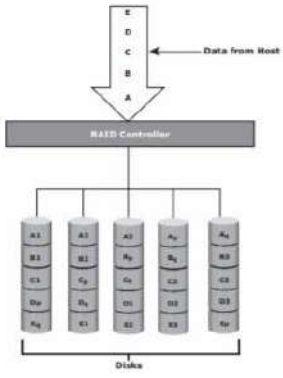



*Figure 3.10:RAID 6*

#### 3.4.4.5 RAID 4

Similar to RAID 3, RAID 4 stripes data for high performance and uses parity for improved fault tolerance. Data is striped across all disks except the parity disk in the array. Parity information is stored on a dedicated disk so that the data can be rebuilt if a drive fails. Unlike RAID 3, data disks in RAID 4 can be accessed independently so that specific data elements can be read or written on a single disk without reading or writing an entire stripe. RAID 4 provides good read throughput and reasonable write throughput.

#### 3.4.4.6 RAID 5

RAID 5 is a versatile RAID implementation. It is similar to RAID 4 because it uses striping. The drives (strips) are also independently accessible. The difference between RAID 4 and RAID 5 is the parity location. In RAID 4, parity is written to a dedicated drive, creating a write bottleneck for the parity disk. In RAID 5, parity is distributed across all disks to overcome the write bottleneck of a dedicated parity disk. Figure 3-9 illustrates the RAID 5 implementation.

RAID 5 is good for random, read-intensive I/O applications and preferred for messaging, data mining, medium-performance media serving, and relational database management system (RDBMS) implementations, in which database administrators (DBAs) optimize data access.

#### 3.4.4.7 RAID 6

RAID 6 works the same way as RAID 5, except that RAID 6 includes a second parity element to enable survival if two disk failures occur in a RAID set (see Figure 3-10). Therefore, a RAID 6 implementation requires at least four disks. RAID 6 distributes the parity across all the disks. The write penalty (explained later in this chapter) in RAID 6 is more than that in RAID 5; therefore, RAID 5 writes perform better than RAID 6. The rebuild operation in RAID 6 may take longer than that in RAID 5 due to the presence of two parity sets.

## 3.5 RAID IMPACT ON DISK PERFORMANCE

When choosing a RAID type, it is imperative to consider its impact on disk performance and application IOPS.

In both mirrored and parity RAID configurations, every write operation translates into more I/O overhead for the disks, which is referred to as a write penalty. In a RAID 1 implementation, every write operation must be performed on two disks configured as a mirrored pair, whereas in a RAID 5 implementation, a write operation may manifest as four I/O operations. When performing I/Os to a disk configured with RAID 5, the controller has to read, recalculate, and write a parity segment for every

data write operation. Figure 3-11 illustrates a single write operation on RAID 5 that contains a group of five disks.

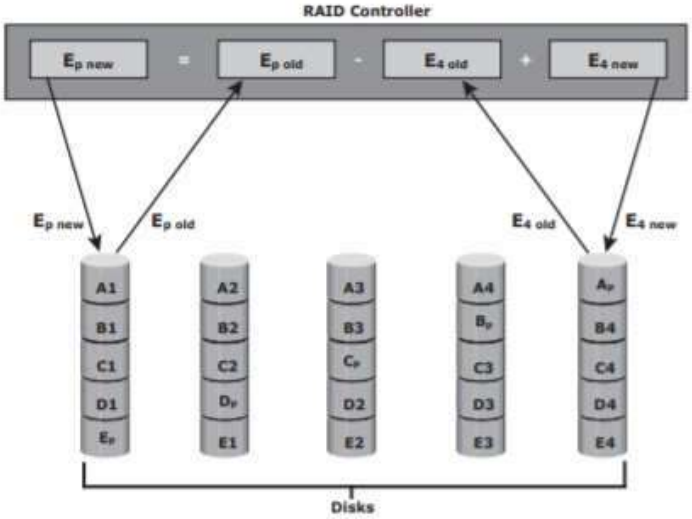



***Figure 3.11: Write Penalty in RAID 5***

The parity (P) at the controller is calculated as follows:
Ep = E1 + E2 + E3 + E4 (XOR operations)

Whenever the controller performs a write I/O, parity must be computed by reading the old parity (Ep old) and the old data (E4 old) from the disk, which means two read I/Os. Then, the new parity (Ep new) is computed as follows:
Ep new = Ep old – E4 old + E4 new (XOR operations)

After computing the new parity, the controller completes the write I/O by writing the new data and the new parity onto the disks, amounting to two write I/Os. Therefore, the controller performs two disk reads and two disk writes for every write operation, and the write penalty is 4.

In RAID 6, which maintains dual parity, a disk write requires three read operations: two parity and one data. After calculating both new parities, the controller performs three write operations: two parity and an I/O. Therefore, in a RAID 6 implementation, the controller performs six I/O operations for each write I/O, and the write penalty is 6.

### 3.5.1 Application IOPS and RAID Configuration

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 to compute 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 (reads + writes) = 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 manifests as two writes to the disks]

= 3,120 + 2* 2,080 = 3,120 + 4,160 = 7,280 IOPS

The computed disk load determines the number of disks required for the application. If in this example a disk drive with a specification of a maximum 180 IOPS needs to be used, the number of disks required to meet the workload for the RAID configuration would be as follows:

RAID 5: 11,440/180 = 64 disks

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

## 3.6 RAID COMPARISON

Following table shows the comparison between all RAID levels.

***Table 3.2: Comparison of Common RAID Types***

| RAID | MIN. DISKS | STORAGE EFFICIENCY % | COST | READ PERFORMANCE | WRITE PERFORMANCE | WRITE PENALTY | PROTECTION |
|---|---|---|---|---|---|---|---|
| 0 | 2 | 100 | Low | Good for both random and sequential reads | Good | No | No protection |
| 1 | 2 | 50 | High | Better than single disk | Slower than single disk because every write must be committed to all disks | Moderate | Mirror protection |
| 3 | 3 | [(n-1)/n] × 100 where n= number of disks | Moderate | Fair for random reads and good for sequential reads | Poor to fair for small random writes and fair for large, sequential writes | High | Parity protection for single disk failure |
| 4 | 3 | [(n-1)/n] × 100 where n= number of disks | Moderate | Good for random and sequential reads | Fair for random and sequential writes | High | Parity protection for single disk failure |
| 5 | 3 | [(n-1)/n] × 100 where n= number of disks | Moderate | Good for random and sequential reads | Fair for random and sequential writes | High | Parity protection for single disk failure |
| 6 | 4 | [(n-2)/n] × 100 where n= number of disks | Moderate but more than RAID 5. | Good for random and sequential reads | Poor to fair for random writes and fair for sequential writes | Very High | Parity protection for two disk failures |
| 1+0 and 0+1 | 4 | 50 | High | Good | Good | Moderate | Mirror protection |

## 3.7 HOT SPARES

A hot spare refers to a spare drive in a RAID array that temporarily replaces a failed disk drive by taking the identity of the failed disk drive. With the hot spare, one of the following methods of data recovery is

performed depending on the RAID implementation: n If parity RAID is used, the data is rebuilt onto the hot spare from the parity and the data on the surviving disk drives in the RAID set. n If mirroring is used, the data from the surviving mirror is used to copy the data onto the hot spare. When a new disk drive is added to the system, data from the hot spare is copied to it. The hot spare returns to its idle state, ready to replace the next failed drive. Alternatively, the hot spare replaces the failed disk drive permanently. This means that it is no longer a hot spare, and a new hot spare must be configured on the array. A hot spare should be large enough to accommodate data from a failed drive. Some systems implement multiple hot spares to improve data availability. A hot spare can be configured as automatic or user initiated, which specifies how it will be used in the event of disk failure. In an automatic configuration, when the recoverable error rates for a disk exceed a predetermined threshold, the disk subsystem tries to copy data from the failing disk to the hot spare automatically. If this task is completed before the damaged disk fails, the subsystem switches to the hot spare and marks the failing disk as unusable. Otherwise, it uses parity or the mirrored disk to recover the data. In the case of a user-initiated configuration, the administrator has control of the rebuild process. For example, the rebuild could occur overnight to prevent any degradation of system performance. However, the system is at risk of data loss if another disk failure occurs.

## 3.8 SUMMARY

- Individual disks are prone to failures and pose the threat of data unavailability.
- RAID addresses data availability requirements by using mirroring and parity techniques.
- RAID implementations with striping enhance I/O performance by spreading data across multiple disk drives, in addition to redundancy benefits.
- This chapter explained the fundamental constructs of striping, mirroring, and parity, which form the basis for various RAID levels.
- Selection of a RAID level depends on the performance, cost, and data protection requirements of an application.
- RAID is the cornerstone technology for several advancements in storage.
- The intelligent storage systems discussed in the next chapter implement RAID along with a specialized operating environment that offers high performance and availability.

## 3.9 REVIEW YOUR LEARNING

- Can you explain requirement of RAID protection?
- Can you explain RAID 0-6 Levels?
- Are you able to explain benefits of RAID levels?
- Are you able to explain impact of RAID on disk performance?

## 3.10 QUESTIONS

1. Why is RAID 1 not a substitute for a backup?
2. Research RAID 6 and its second parity computation.
3. Explain the process of data recovery in case of a drive failure in RAID 5.
4. What are the benefits of using RAID 3 in a backup application?
5. Discuss the impact of random and sequential I/Os in different RAID configurations.
6. An application has 1,000 heavy users at a peak of 2 IOPS each and 2,000 typical users at a peak of 1 IOPS each. It is estimated that the application also experiences an overhead of 20 percent for other workloads. The read/write ratio for the application is 2:1. Calculate RAID corrected IOPS for RAID 1/0, RAID 5, and RAID 6.
7. For Question 6, compute the number of drives required to support the application in different RAID environments if 10 K RPM drives with a rating of 130 IOPS per drive were used.
8. What is the stripe size of a five-disk RAID 5 set with a strip size of 32 KB? Compare it with the stripe size of a five-disk RAID 0 array with the same strip size.

## 3.11 FURTHER READING

- http://aad.tpu.ru/practice/EMC/Information%20Storage%20and%20Management-v.2.pdf
- https://nptel.ac.in/courses/106/108/106108058/
- https://nptel.ac.in/content/storage2/courses/106108058/lec%2007.pdf
- http://www.ictacademy.in/pages/Information-Storage-and-Management.aspx
- https://www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwiM8Kq6isHyAhUEkmYCHbJyDXAYABAAGgJzbQ&ae=2&ohost=www.google.com&cid=CAESQeD28QNmzUxhr6qtgEwm24g2Yc-TeMC_24a0sxeZf9MitA7QrS5Vz4VE3XfWSwFvX0iAKPoH4fT4QmSj7PhnMAQF&sig=AOD64_1Y3y_5vJpAZOJybqnNONsE6wNayQ&q&adurl&ved=2ahUKEwjvsaG6isHyAhXjxTgGHTvKBEEQ0Qx6BAgDEAE

- https://www.coursera.org/lecture/big-data-management/data-storage-RplBY
- https://www.coursera.org/courses?query=data%20storage
- https://www.coursera.org/lecture/technical-support-fundamentals/storage-RLNIZ
- https://www.coursera.org/learn/cloud-storage-big-data-analysis-sql-pdf/

## 3.12 REFERENCES


1. Information Storage and Management: Storing, Managing and Protecting Digital Information in Classic, Virtualized and Cloud Environments, EMC, John & Wiley Sons, 2nd Edition, 2012
2. Information Storage and Management, Pankaj Sharma
3. Information Technology Project Management, Jack T Marchewka
4. Information Storage and Management, I A Dhotre


❖❖❖❖

# Module II

# 4

# INTELLIGENT STORAGE SYSTEM

**Unit Structure**



## 4.0 INTELLIGENT STORAGE SYSTEM

Business-critical applications require high levels of performance, availability, security, and scalability. A hard disk drive is a core element of storage that governs the performance of any storage system. Some of the older disk array technologies could not overcome performance constraints due to the limitations of a hard disk and its mechanical components. 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 detailed in this chapter are the 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. These storage systems are configured with large amounts of memory called *cache* and use sophisticated algorithms to meet the I/O requirements of performance- sensitive applications.

**Components of An Intelligent Storage System**

An intelligent storage system consists of four key components: *front end, cache, back end,* and *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.

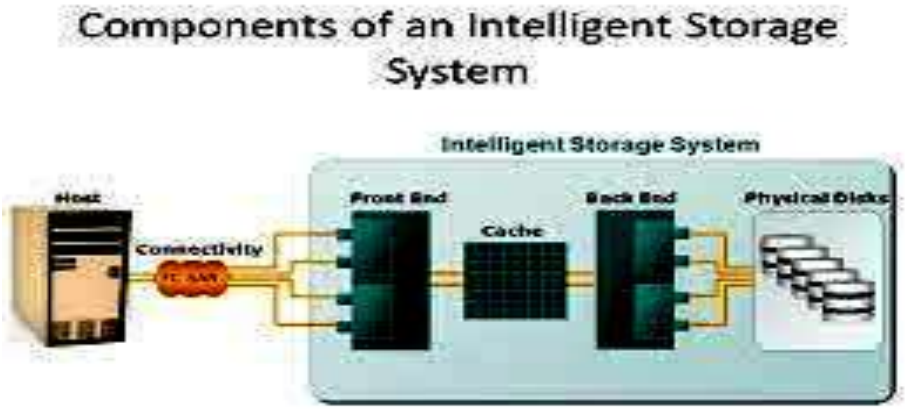



**Figure 4-1:** Components of an intelligent storage system

## 4.1 FRONT END

The front end provides the interface between the storage system and the host. It consists of two components: front-end ports and front-end controllers. The *front-end ports* enable hosts to connect to the intelligent storage system. Each front-end port has processing logic that executes the appropriate transport protocol, such as SCSI, Fibre Channel, or iSCSI, for storage connections. Redundant ports are provided on the front end for high availability.

*Front-end controllers* route data to and from cache via the internal data bus. When cache receives write data, the controller sends an acknowledgment message back to the host. Controllers optimize I/O processing by using command queuing algorithms.

**Front-End Command Queuing**

*Command queuing* is a technique implemented on front-end controllers. It determines the execution order of received commands and can reduce unnecessary drive head movements and improve disk performance. When a command is received for execution, the command queuing algorithms assigns a tag that defines a sequence in which commands should be executed. With command queuing, multiple commands can be executed concurrently based on the organization of data on the disk, regardless of the order in which the commands were received. The most commonly used command queuing algorithms are as follows:

- **First In First Out** (FIFO)**:** This is the default algorithm where commands are executed in the order in which they are received (Figure 4-2 [a]). There is no reordering of requests for optimization; therefore, it is inefficient in terms of performance.

■ **Seek Time Optimization:** Commands are executed based on optimizing read/write head movements, which may result in reordering of commands. Without seek time optimization, the commands are executed in the order they are received. For example, as shown in Figure 4-2(a), the commands are executed in the order A, B, C and D. The radial movement required by the head to execute C immediately after A is less than what would be required to execute B. With seek time optimization, the command execution sequence would be A, C, B and D, as shown in Figure 4-2(b).

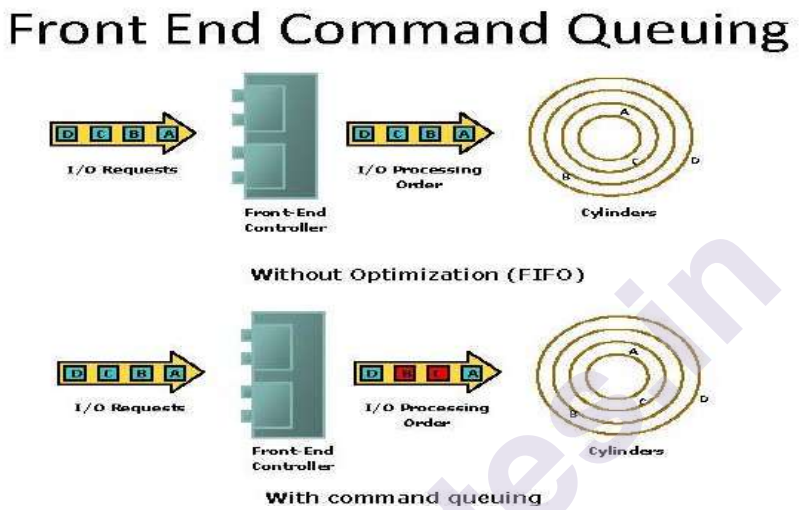



**Figure 4-2:** Front-end command queuing

■ **Access Time Optimization:** Commands are executed based on the combination of seek time optimization and an analysis of rotational latency for optimal performance.

Command queuing can also be implemented on disk controllers and this may further supplement the command queuing implemented on the front-end controllers. Some models of SCSI and Fibre Channel drives have command queuing implemented on their controllers.

## 4.2 CACHE

Cache is an important component that enhances the I/O performance in an intelligent storage system. Cache is semiconductor memory where data is placed temporarily to reduce the time required to service I/O requests from the host. Cache improves storage system performance by isolating hosts from the mechanical delays associated with physical disks, which are the slowest components of an intelligent storage system. Accessing data from a physical disk usually takes a few milliseconds because of seek times and rotational latency. If a disk has to be accessed by the host for every I/O operation, requests are queued, which results in a delayed response. 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**

Cache is organized into pages or slots, which is the smallest unit of cache allocation. The size of a cache page is configured according to the application I/Osize. Cache consists of the *data store* and *tag RAM*. The data store holds the data while tag RAM tracks the location of the data in the data store (see Figure 4-3)and in disk.

Entries in tag RAM indicate where data is found in cache and where the data belongs on the disk. Tag RAM includes a *dirty bit* flag, which indicates whether the data in cache has been committed to the disk or not. It also contains time-based information, such as the time of last access, which is used to identify cached information that has not been accessed for a long period and may be freed up.

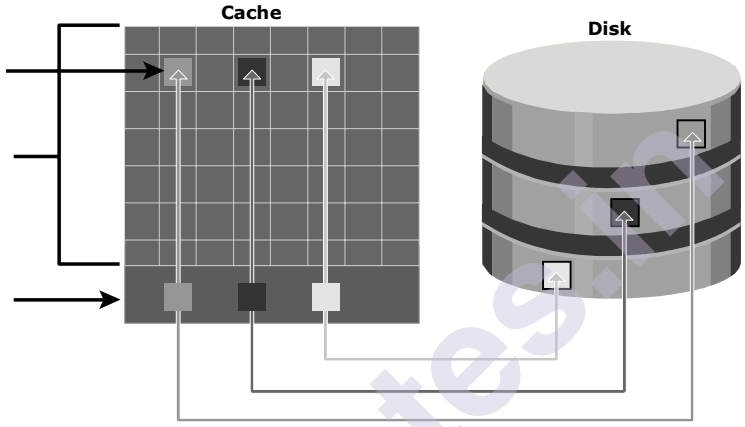



**Figure 4-3: Structure of cache**

**Read Operation with Cache**

When a host issues a read request, the front-end controller accesses the tag RAM to determine whether the required data is available in cache. If the requested data is found in the cache, it is called a *read cache hit* or *read hit* and data is sent directly to the host, without any disk operation (see Figure 4-4[a]). This provides a fast response time to the host (about a millisecond). If the requested data is not found in cache, it is called a *cache miss* and the data must be read from the disk (see Figure 4-4[b]). The back-end controller accesses the appropriate disk and retrieves the requested data. Data is then placed in cache and is finally sent to the host through the front-end controller. Cache misses increase I/O response time.

A *pre-fetch,* or *read-ahead,* algorithm is used when read requests are sequential. In a sequential read request, a contiguous set of associated blocks is retrieved. Several other blocks that have not yet been requested by the host can be read from the disk and placed into cache in advance. When the host subsequently requests these blocks, the read operations will be read hits. This process significantly improves the response time experienced by the host. The intelligent storage system offers fixed and variable pre-fetch sizes. In *fixed pre-fetch*, the intelligent storage system pre-fetches a fixed amount of data. It is most suitable when I/O sizes are uniform. In *variable pre-fetch*, the storage system pre-fetches an amount of data in multiples of the size of the host request. Maximum

pre-fetch limits the number of data blocks that can be pre-fetched to prevent the disks from being rendered busy with pre-fetch at the expense of other I/O.

Read performance is measured in terms of the *read hit ratio,* or the *hit rate*, usually expressed as a percentage. This ratio is the number of read hits with respect to the total number of read requests. A higher read hit ratio improves the read performance.

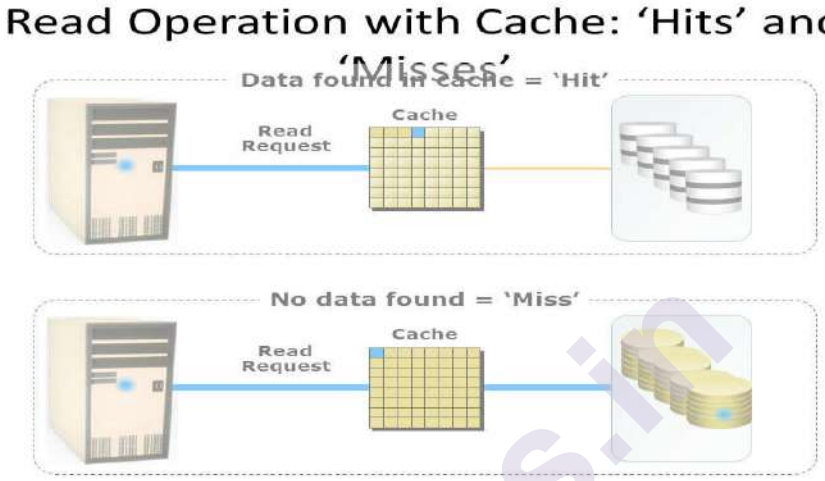



**Figure 4-4:** Read hit and read miss

***Write Operation with Cache***

Write operations with cache provide performance advantages over writing directly to disks. When an I/O is written to cache and acknowledged, it is completed in far less time (from the host's perspective) than it would take to write directly to disk. Sequential writes also offer opportunities for optimization because many smaller writes can be coalesced for larger transfers to disk drives with the use of cache.

A write operation with cache is implemented in the following ways:

**Write-back cache:** Data is placed in cache and an acknowledgment is sentto the host immediately. Later, data from several writes are committed (de-staged) to the disk. Write response times are much faster, as the write operations are isolated from the mechanical delays of the disk. However, uncommitted data is at risk of loss in the event of cache failures.

- **Write-through cache:** Data is placed in the cache and immediately writ- ten to the disk, and an acknowledgment is sent to the host. Because data is committed to disk as it arrives, the risks of data loss are low but writeresponse time is longer because of the disk operations.

Cache can be bypassed under certain conditions, such as very large size write I/O. In this implementation, if the size of an I/O request exceeds the pre- defined size, called *write aside size*, writes are sent to the disk directly to reducethe impact of large writes consuming a large cache area. This is particularly useful in an environment where cache resources are constrained and must bemade available for small random I/Os.

**Cache Implementation**

Cache can be implemented as either dedicated cache or global cache. With dedicated cache, separate sets of memory locations are reserved for reads and writes. In global cache, both reads and writes can use any of the available memory addresses. Cache management is more efficient in a global cache implementation, as only one global set of addresses has to be managed.

Global cache may allow users to specify the percentages of cache available for reads and writes in cache management. Typically, the read cache is small, but it should be increased if the application being used is read intensive. In other global cache implementations, the ratio of cache available for reads versus writes is dynamically adjusted based on the workloads.

**Cache Management**

Cache is a finite and expensive resource that needs proper management. Even though intelligent storage systems can be configured with large amounts of cache, when all cache pages are filled, some pages have to be freed up to accommodate new data and avoid performance degradation. Various cache management algorithms are implemented in intelligent storage systems to proactively maintain a set of free pages and a list of pages that can be potentially freed up whenever required:

- **Least Recently Used (LRU):** An algorithm that continuously monitors data access in cache and identifies the cache pages that have not been accessed for a long time. LRU either frees up these pages or marks them for reuse. This algorithm is based on the assumption that data which hasn't been accessed for a while will not be requested by the host. However, if a page contains write data that has not yet been committed to disk, data will first be written to disk before the page is reused.

- **Most Recently Used (MRU):** An algorithm that is the converse of LRU. In MRU, the pages that have been accessed most recently are freed up or marked for reuse. This algorithm is based on the assumption that recently accessed data may not be required for a while.

As cache fills, the storage system must take action to flush dirty pages (data written into the cahce but not yet written to the disk) in order to manage its availability. Flushing is the process of committing data from cache to the disk. On the basis of the I/O access rate and pattern, high and low levels called *watermarks* are set in cache to manage the flushing process. *High watermark (HWM)* is the cache utilization level at which the storage system starts high- speed flushing of cache data. *Low watermark (LWM)* is the point at which the storage system stops the high-speed or forced flushing and returns to idle flush behavior. The cache utilization level, as shown in Figure 4-5, drives the mode of flushing to be used:

- **Idle flushing:** Occurs continuously, at a modest rate, when the cache utilization level is between the high and low watermark.

- **High watermark flushing:** Activated when cache utilization hits the high watermark. The storage system dedicates some additional resources to flush- ing. This type of flushing has minimal impact on host I/O processing.
- **Forced flushing:** Occurs in the event of a large I/O burst when cache reaches 100 percent of its capacity, which significantly affects the I/O response time. In forced flushing, dirty pages are forcibly flushed to disk.

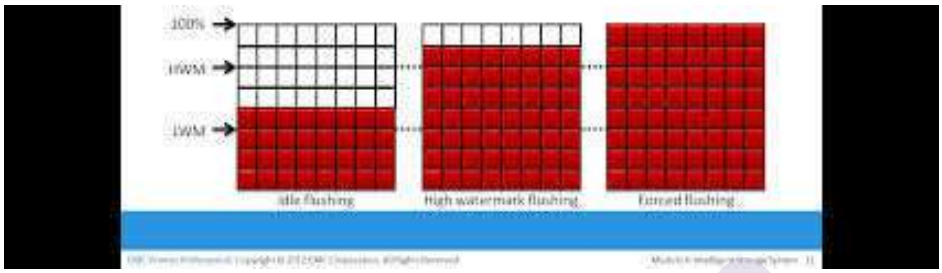



**Figure 4-5: Types of flushing**

### Cache Data Protection

Cache is volatile memory, so a power failure or any kind of cache failure will causethe loss of data not yet committed to the disk. This risk of losing uncommitted data held in cache can be mitigated using *cache mirroring* and *cache vaulting*:

- **Cache mirroring:** Each write to cache is held in two different memory locations on two independent memory cards. In the event of a cache failure, the write data will still be safe in the mirrored location and can be committed to the disk. Reads are staged from the disk to the cache; therefore, in the event of a cache failure, the data can still be accessed from the disk. As only writes are mirrored, this method results in betterutilization of the available cache.

In cache mirroring approaches, the problem of maintaining *cache coherency* is introduced. Cache coherency means that data in two different cache locations must be identical at all times. It is the responsibility of the arrayoperating environment to ensure coherency.

- **Cache vaulting:** Cache is exposed to the risk of uncommitted data loss dueto power failure. This problem can be addressed in various ways: power- ing the memory with a battery until AC power is restored or using battery power to write the cache content to the disk. In the event of extended power failure, using batteries is not a viable option because in intelligent storage systems, large amounts of data may need to be committed to numerous disks and batteries may not provide power for sufficient time to write each piece of data to its intended disk. Therefore, storage vendorsuse a set of physical disks to dump the contents of cache during power failure. This is called cache vaulting and the disks are called vault drives. When power is restored, data from these disks is written back to write cache and then 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. 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.

For high data protection and availability, storage systems are configured with dual controllers with multiple ports. Such configurations provide an alternate path to physical disks in the event of a controller or port failure. This reliability is further enhanced if the disks are also dual-ported. In that case, each disk port can connect to a separate controller. Multiple controllers also facilitate load balancing.

**Physical Disk**

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

## 4.3 LOGICAL UNIT NUMBER

Physical drives or groups of RAID protected drives can be logically split into volumes known as logical volumes, commonly referred to as *Logical Unit Numbers* (LUNs). The use of LUNs improves disk utilization. For example, without the use of LUNs, a host requiring only 200 GB could be allocated an entire 1TB physical disk. Using LUNs, only the required 200 GB would be allocated to the host, allowing the remaining 800 GB to be allocated to other hosts. In the case of RAID protected drives, these logical units are slices of RAID sets and are spread across all the physical disks belonging to that set. The logical units can also be seen as a logical partition of a RAID set that is presented to a host as a physical disk. For example, Figure 4-6 shows a RAID set consisting of five disks that have been sliced, or partitioned, into several LUNs. LUNs 0 and 1 are shown in the figure.

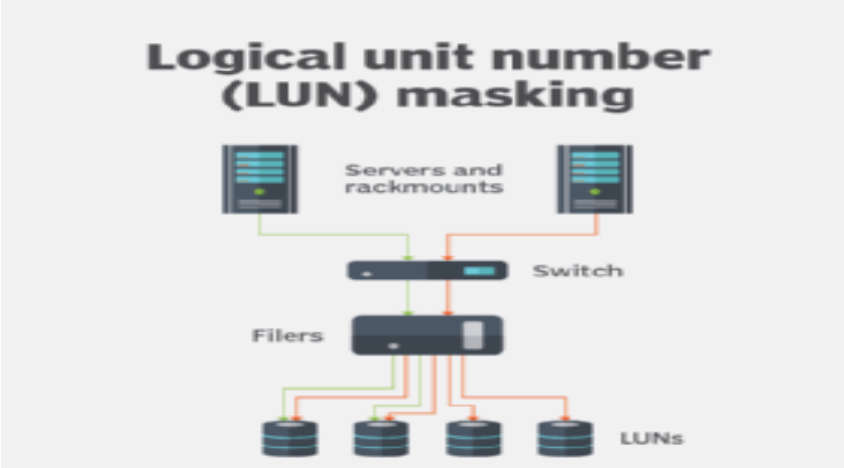



**Figure 4-6:** Logical unit number

Note how a portion of each LUN resides on each physical disk in the RAIDset. LUNs 0 and 1 are presented to hosts 1 and 2, respectively, as physical volumes for storing and retrieving data. Usable capacity of the physical volumesis determined by the RAID type of the RAID set.

The capacity of a LUN can be expanded by aggregating other LUNs with it. The result of this aggregation is a larger capacity LUN, known as a *meta- LUN*. The mapping of LUNs to their physical location on the drives is man- aged by the operating environment of an intelligent storage system.

## 4.4 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.

## 4.5 INTELLIGENT STORAGE ARRAY

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

- High-end storage systems
- Midrange storage systems

Traditionally, high-end storage systems have been implemented with *active-active arrays,* whereas midrange *storage systems* used typically in small- and medium- sized enterprises have been implemented with *active-passive arrays*. Active-passive arrays provide optimal storage solutions at lower costs. Enterprises make use of this cost advantage and implement active-passive arrays to meet specific applicationrequirements such as performance, availability, and scalability. The distinctions between these two implementations are becoming increasingly insignificant.

## 4.6 HIGH-END STORAGE SYSTEMS

High-end storage systems, referred to as *active-active arrays,* are generally 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 (see Figure 4-7).

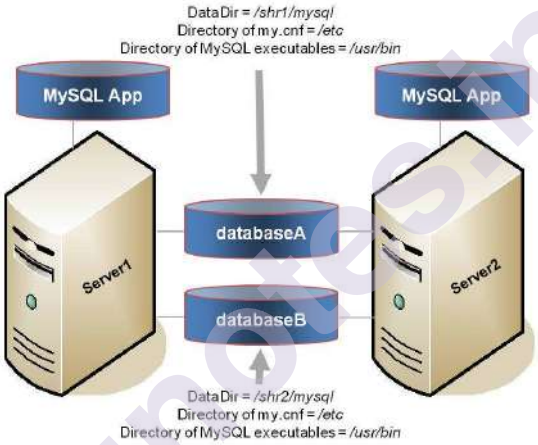



**Figure 4-7:** Active-active configuration

To address the enterprise storage needs, these arrays provide the followingcapabilities:

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

In addition to these features, high-end arrays possess some unique features and functionals that are required for mission-critical applications in large enterprises.

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

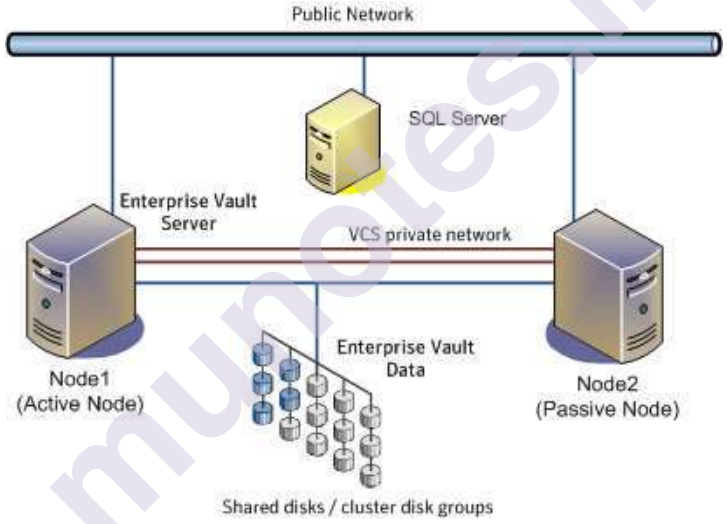



**Figure 4-8:** Active-passive configuration

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.

## 4.8 SUMMARY

This chapter detailed the features and components of the intelligent storage system — front end, cache, back end, and physical disks. The active-active and active-passive implementations of intelligent storage

systems were also described. An intelligent storage system provides the following benefits to an organization:

- Increased capacity
- Improved performance
- Easier storage management
- Improved data availability
- Improved scalability and flexibility
- Improved business continuity
- Improved security and access control

An intelligent storage system is now an integral part of every mission-critical data center. Although a high-end intelligent storage system addresses information storage requirements, it poses a challenge for administrators to share information easily and securely across the enterprise.

Storage networking is a flexible information-centric strategy that extends the reach of intelligent storage systems throughout an enterprise. It provides a common way to manage, share, and protect information.

## 4.9 QUESTIONS

1. Explain the components of an Intelligent Storage System.
2. What is front-end command queuing? Explain the common command queuing algorithms.
3. What is cache? Explain the structure of cache.
4. Explain the read and write cache operations.
5. How is cache implemented and managed?
6. Discuss the cache data protection.
7. What are solid state devices? Explain.
8. Explain the concept of Logical Unit number.
9. What are the two categories of intelligent storage systems?

## 4.10 REFERENCES

**1.** Data Center Virtualization Fundamentals, Gustavo Alessandro Andrade Santana, Cisco Press 1st Edition 2014.

# 5

# STORAGE AREA NETWORKS

**Unit Structure**



## 5.0 STORAGE CONSOLIDATION

Organizations are experiencing an explosive growth in information. This information needs to be stored, protected, optimized, and managed efficiently. Data center managers are burdened with the challenging task of providing low-cost, high-performance information management solutions. An effective information management solution must provide the following:

- **Just-in-time information to business users:** Information must be available to business users when they need it. The explosive growth in online storage, proliferation of new servers and applications, spread of mission-critical data throughout enterprises, and demand for 24 × 7 data availability are some of the challenges that need to be addressed.
- **Integration of information infrastructure with business processes:** The storage infrastructure should be integrated with various business pro- cesses without compromising its security and integrity.
- **Flexible and resilient storage architecture:** The storage infrastructure must provide flexibility and resilience that aligns with changing business requirements. Storage should scale without compromising performance requirements of the applications and, at the same time, the total cost of managing information must be low.

Direct-attached storage (DAS) is often referred to as a stovepiped storage environment. Hosts "own" the storage and it is difficult to manage

and share resources on these isolated storage devices. Efforts to organize this dispersed data led to the emergence of the storage area network (SAN). SAN is a high- speed, dedicated network of servers and shared storage devices. Traditionallyconnected over Fibre Channel (FC) networks, a SAN forms a single-storage pool and facilitates data centralization and consolidation. SAN meets the stor- age demands efficiently with better economies of scale. A SAN also provides effective maintenance and protection of data.

This chapter provides detailed insight into the FC technology on which a SAN is deployed and also reviews SAN design and management fundamentals.

## 5.1 FIBRE CHANNEL: OVERVIEW

The fibre channel methodology has means to implement three topologies: point-to-point links, arbitrated loops (shared bandwidth loop circuits), and bandwidth switched fabrics that provide SANs with the ability to do bandwidth multiplexing by supporting simultaneous data transmission between various pairs of devices. Any storage device on the loop can be accessed through a fibre channel switch (FCSW) or hub. The fibre channel switch can support entry-level (8–16 ports) to enterprise-level (64–128 ports) systems. Under the ANSI X3T11 standards regulation, up to 126 storage devices (nodes) can be linked in the fiber channel arbitrated loop (FC-AL) configuration, with the storage interface bandwidth about 100 Mbits/s for transferring large files. More than 70 companies, including industry-leading vendors of disk arrays and computer and networking systems, support the FC-AL voluntary standards. The FC-AL topology is used primarily to connect disk arrays and FC devices. Originally developed as the high-speed serial technology of choice for server– storage connectivity, the FC-AL methodology is extended to the FC-SL standard that supports isochronous and time-deterministic services, including methods of managing loop operational parameters and QoS definitions, as well as control. The FC-VI regulation establishes a fibre channel-virtual interface architecture (FC-VIA) mapping standard. See the chapter on fibre channels for more information about various implementations of the technology in various network configurations, including SANs.

Because of the high cost of the FC interconnect components and separation of storage and servers at the wide area network scale (resulting in slow capabilities of WAN– SANs with fibre channel), alternatives to FC technologies have been developed. The *ipStorage* technology (Barker & Massiglia, 2001, p. 187) employs TCP/IP as a storage interconnect. The Internet Engineering Task Force (IETF) has proposed the iSCSI (*Internet SCSI*) standards that address the issues of long distances (WAN-scale), reducing the interconnect cost, high security, and complex storage network topologies. The iSCSI is layered on top of the TCP/IP protocol hierarchy and can instantly access all modern transmission media and topologies.

TCP/IP and related protocols have been implemented in the server-based systems that allow the most general storage networks to be constructed with the iSCSI methodology. The main challenge is a reduction of the iSCSI processor overhead of operating iSCSI packets below the Fibre Channel overhead level.

**The SAN and Its Evolution**

A *storage area network (SAN)* carries data between servers (also known as *hosts*) and storage devices through fibre channel switches (see Figure 5-1). A SAN enables storage consolidation and allows storage to be shared across multiple servers. It enables organizations to connect geographically dispersed servers and storage.

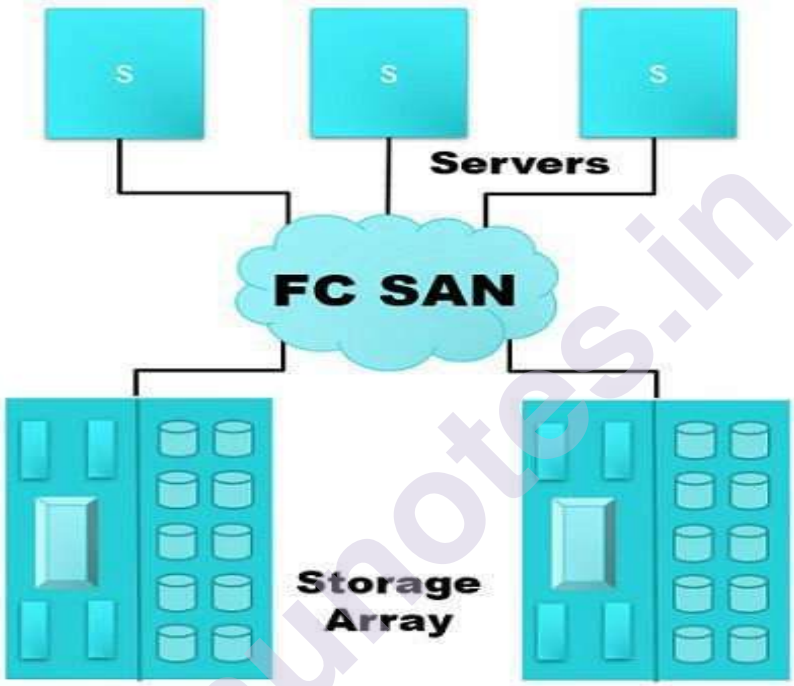



**Figure 5-1:** SAN implementation

A SAN can be considered as an extended and shared storage bus within a data center, consisting of various storage devices and specific interfaces (e.g., fibre channel, ESCON, HIPPI, SCSI, or SSA) rather than the Ethernet (Peterson, 1998). In order to be connected to the enterprise network, the SAN utilizes technologies similar to those of LANs and WANs: switches, routers, gateways, and hubs (see Figure 5-1). Wide area network carrier technologies, such as asynchronous transfer mode (ATM) or synchronous optical networks, can be used for remote archival data storing and backup. As an important element of modern distributed networking architectures of storage-centric enterprise information processing, SAN technology represents a significant step toward a fully networked secure data storage infrastructure that is radically different from traditional server-attached storage (Clark, 1999). The SAN represents a new segment of the information services industry called storage solution providers (SSP). However, isolated SANs cannot realize SSPs' services, such as real-time data replication, failover, storage hosting, and remote vaulting.

## 5.2 COMPONENTS OF SAN

A SAN consists of three basic components: servers, network infrastructure, and storage. These components can be further broken down into the following key elements: node ports, cabling, interconnecting devices (such as FC switches or hubs), storage arrays, and SAN management software.

### 5.2.1 Node Ports

In fibre channel, devices such as hosts, storage and tape libraries are all referred to as *nodes*. Each node is a source or destination of information for one or more nodes. Each node requires one or more ports to provide a physical interface for communicating with other nodes. These ports are integral components of an HBA and the storage front-end adapters. A port operates in full-duplex data transmission mode with a *transmit (Tx)* link and a *receive (Rx)* link (see Figure 5-3).

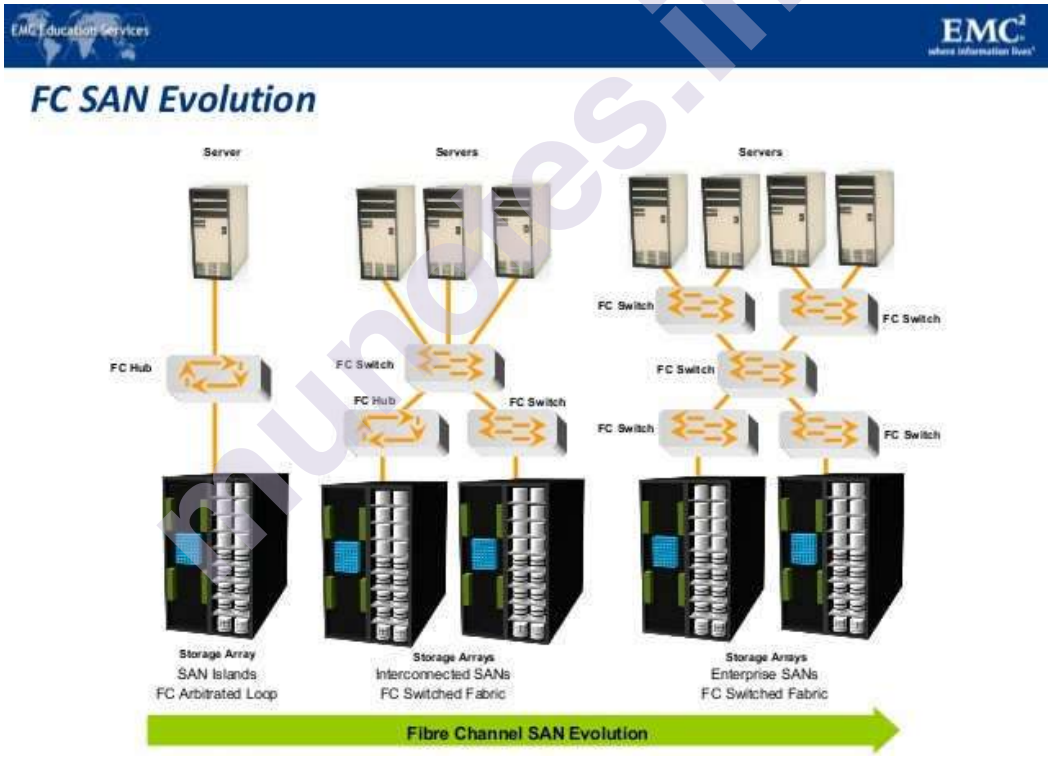



**Figure 5-2:** FC SAN evolution

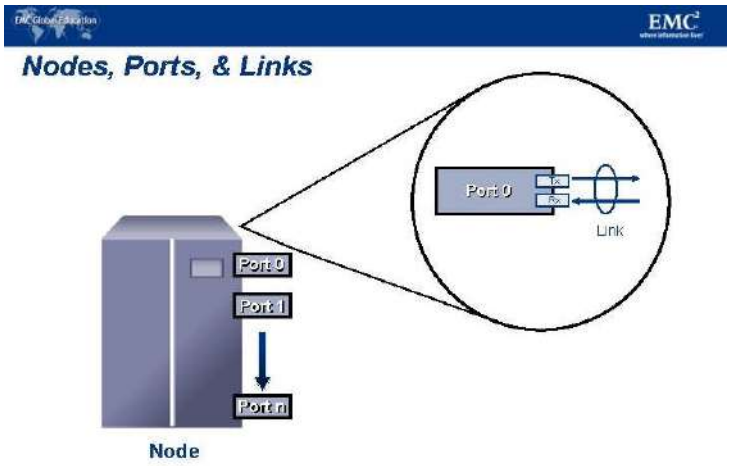



**Figure 5-3:** Nodes, ports, and links

### 5.2.2 Cabling

SAN implementations use optical fiber cabling. Copper can be used for shorter distances for back-end connectivity, as it provides a better 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, multi-mode and single-mode.

Multi-mode 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, multi-mode fibers are classified as OM1 (62.5µm), OM2 (50µm) and laser optimized OM3 (50µm). 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 usually used for distances of up to 500 meters 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 (see Figure 5-4 (b)). These cables are available in diameters of 7–11 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 limits modal dispersion. Among all types of fibre cables, single-mode provides minimum signal attenuation over maximum distance (up to 10 km). A single-mode cable is used for long-distance cable runs, limited only by the power of the laser at the transmitter and sensitivity of the receiver.

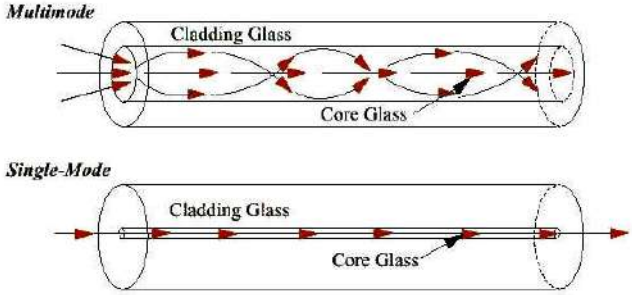



**Figure 5-4:** Multi-mode fiber and single-mode fiber

MMFs are generally used within data centers for shorter distance runs, while SMFs are used for longer distances. MMF transceivers are less expensive as compared to SMF transceivers.

A Standard connector (SC) (see Figure 5-5 (a)) and a Lucent connector (LC) (see Figure 6-5 (b)) are two commonly used connectors for fiber optic cables. An SC is used for data transmission speeds up to 1 Gb/s, whereas an LC is used for speeds up to 4 Gb/s. Figure 5-6 depicts a Lucent connector and a Standard connector.

A *Straight Tip (ST)* is a fiber optic connector with a plug and a socket that is locked with a half-twisted bayonet lock (see Figure 5-5 (c)). In the early days of FC deployment, fiber optic cabling predominantly used ST connectors. This connector is often used with Fibre Channel patch panels.


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

The Small Form-factor Pluggable (SFP) is an optical transceiver used in optical communication. The standard SFP+ transceivers support data rates up to 10 Gb/s.

### 5.2.3 Interconnect Devices

As the name suggests these devices are used for connection between hosts in the SAN environment and **Hubs, switches and directors** are the examples of interconnecting devices.

**Hubs** physically connect nodes in a logical loop or a physical star topology and are used as communication equipment in FC-AL applications.

**Switches** directly route data from 1 physical port to different one and are thus are more intelligent than hubs.

**Directors** works the same way as FC switches, but directors have higher port count and fault tolerance capacity. They are also larger than switches and are deployed for data center works.

### 5.2.3 Storage Arrays

**The main aim of any SAN network is to provide storage resources to its host**. SAN implementations supplement the standard features of storage arrays by providing high accessibility and redundancy, improved performance, business continuity, and multiple host connectivity.

The large storage capacities offered by modern storage arrays have been exploited in SAN environments for storage consolidation and centralization.

### 5.2.4 SAN Management Software

**SAN management application package handles the interfaces between hosts, interconnect devices, and storage arrays**.

SAN management software is very important as it allows the complete management of different resources from a single point and also gives the complete structure of the SAN environment.

It provides key management functions, as well as mapping of storage devices, switches, and servers, observance and generating alerts for discovered devices, and logical partitioning of the SAN, known as partitioning.



## 5.3 FC CONNECTIVITY

The FC architecture supports three basic interconnectivity options: point-to- point, arbitrated loop (FC-AL), and fabric connect.

### 5.3.1 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, as only two devices can communicate with each other at a given time. Moreover, it cannot be scaled to accommodate a large number of network devices. Standard DAS usess point-to-point connectivity.

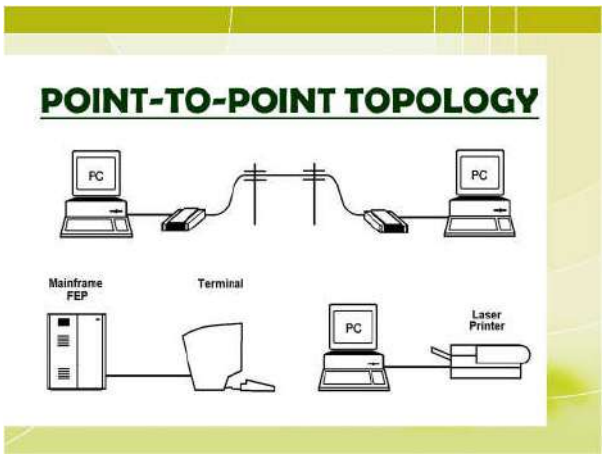



**Figure 5-6: Point-to-point topology**

### 5.3.2 Fibre Channel Arbitrated Loop

In the FC-AL configuration, devices are attached to a shared loop, as shown in Figure 6-7. 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/Ooperations. Devices on the loop must "arbitrate" to gain control of the loop. Atany given time, only one device can perform I/O operations on the loop.

As a loop configuration, FC-AL can be implemented without any interconnectingdevices by directly connecting one device to another in a ring through cables. However, FC-AL implementations may also use hubs whereby the arbitratedloop is physically connected in a star topology.

The FC-AL configuration has the following limitations in terms of scalability:

FC-AL shares the bandwidth in the loop. Only one device can perform I/O operations at a time. Because each device in a loop has to wait for itsturn to process an I/O request, the speed of data transmission is low in an FC-AL topology.

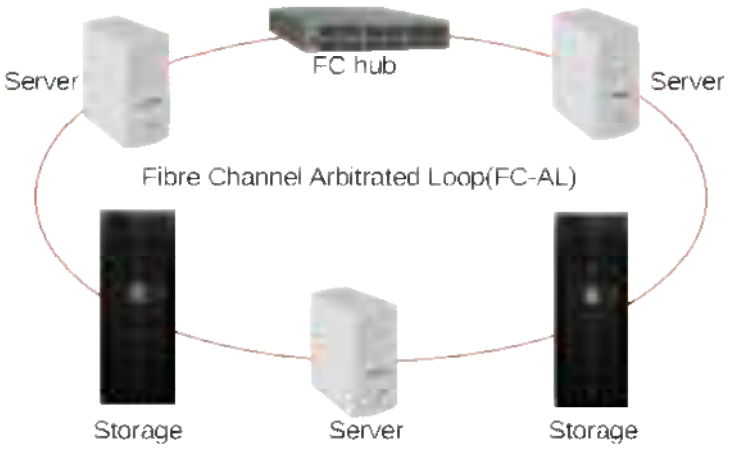



**Figure 5-7: Fibre Channel arbitrated loop**

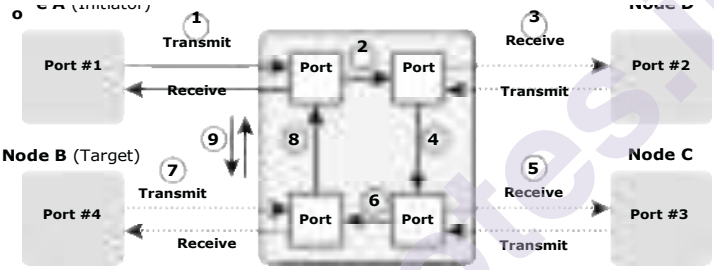
- FC-AL uses 8-bit addressing. It can support up to 127 devices on a loop.
- Adding or removing a device results in loop re-initialization, which can cause a momentary pause in loop traffic.

**FC-AL Transmission**

When a node in the FC-AL topology attempts to transmit data, the node sends an *arbitration (ARB)* frame to each node on the loop. If two nodes simultaneously attempt to gain control of the loop, the node with the highest priority isallowed to communicate with another node. This priority is determined on the basis of Arbitrated Loop Physical Address (AL-PA) and Loop ID, describedlater in this chapter.

When the initiator node receives the ARB request it sent, it gains control of theloop. The initiator then transmits data to the node with which it has establisheda virtual connection. Figure 5-8 illustrates the process of data transmission in an FC-AL configuration.

**N**




**FC Hub**

**Figure 5-8: Data transmission in FC-AL**

**Node A want to communicate with Node B**

1. High priority initiator, Node A inserts the ARB frame in the loop. ARB frame is passed to the next node (Node D) in the loop.
2. Node D receives high priority ARB, therefore remains idle. ARB is forwarded to next node (Node C) in the loop.
3. Node C receives high priority ARB, therefore remains idle. ARB is forwarded to next node (Node B) in the loop.
4. Node B receives high priority ARB, therefore remains idle and ARB is forwarded to next node (Node A) in the loop.
5. Node A receives ARB back; now it gains control of the loop and can start communicating with target Node B.

### 5.3.3 Fibre Channel Switched Fabric

**Fibre Channel** (**FC**) is a high-speed data transfer protocol providing in-order, lossless delivery of raw block data. Fibre Channel is primarily used to connect computer data storage to servers in storage area networks (SAN) in commercial data centers.

Fibre Channel networks form a switched fabric because the switches in a network operate in unison as one big switch. Fibre Channel typically runs on optical fiber cables within and between data centers, but can also run on copper cabling. Supported data rates include 1, 2, 4, 8, 16, 32, 64, and 128 gigabit per second resulting from improvements in successive technology generations.

There are various upper-level protocols for Fibre Channel, including two for block storage. Fibre Channel Protocol (FCP) is a protocol that transports SCSI commands over Fibre Channel networks.[3][4] FICON is a protocol that transports ESCON commands, used by IBM mainframe computers, over Fibre Channel. Fibre Channel can be used to transport data from storage systems that use solid-state flash memory storage medium by transporting NVMe protocol commands.

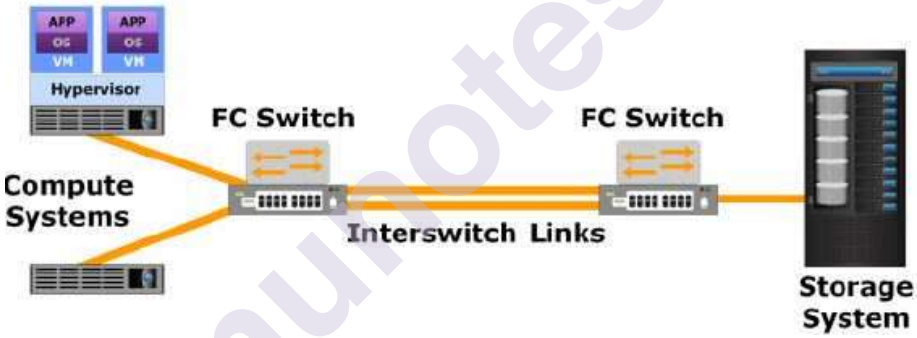



**Figure 5-9: Fibre Channel switched fabric**

When the number of tiers in a fabric increases, the distance that a fabric management message must travel to reach each switch in the fabric also increases. The 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 (detailed later in this chapter). Figure 6-10 illustrates two-tier and three-tier fabric architecture.

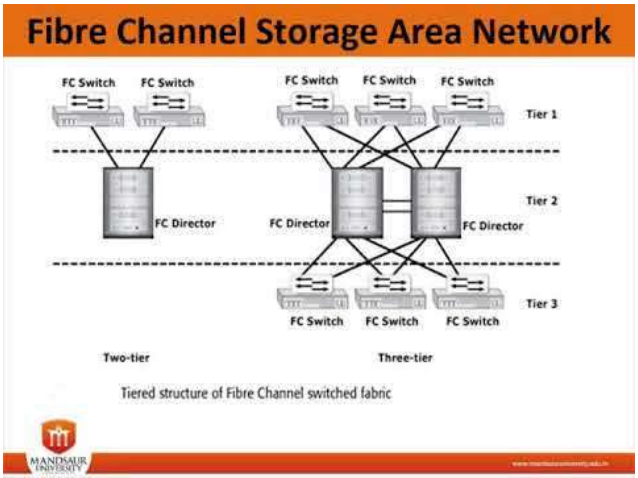



**Figure 5-10: Tiered structure of FC-SW topology**

**FC-SW Transmission**

FC-SW uses switches that are intelligent devices. They can switch data traffic from an initiator node to a target node directly through switch ports. Frames are routed between source and destination by the fabric.

As shown in Figure 5-11, if node B wants to communicate with node D, 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.

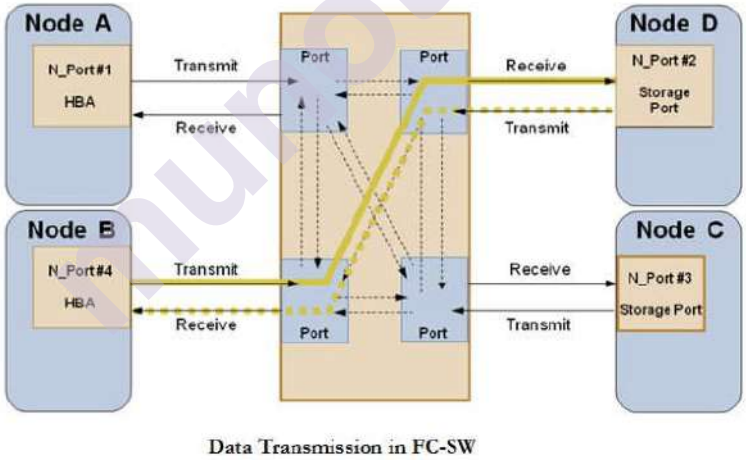



**Figure 5-11:** Data transmission in FC-SW topology

## 5.4 FIBRE CHANNEL PORTS

There are different types of Fibre Channel ports.

**Let's have a look**

**Quick Reference**

| Short Name | Descriptive Name | Device Type | Port Function |
|---|---|---|---|
| N-port | Node Port | Node | port used to connect a node to a Fibre Chanel switch |
| F-port | Fabric Port Switches | Switch | port used to connect the Fibre Channel fabric to a node |
| L-port | Loop Port Nodes | Node | port used to connect a node to a Fibre Channel loop |
| NL-port | Node Loop | Nodes | Node port which connects to both loops and switches |
| FL-port | Fabric + Loop Port | Switches | Switch port which connects to both loops and switches |
| E-port | Expansion Port | Switches | Used to cascade fibre channel switches together |
| G-port | General Port | Switches | General purpose port which can be configured to emulate other port type |
| U-port | Universal port | Switches | Intial port state on a switch before anything has connected and it changes personality to an operation state (E-port, F-port, fl-port) or a transitional state like a g-port |

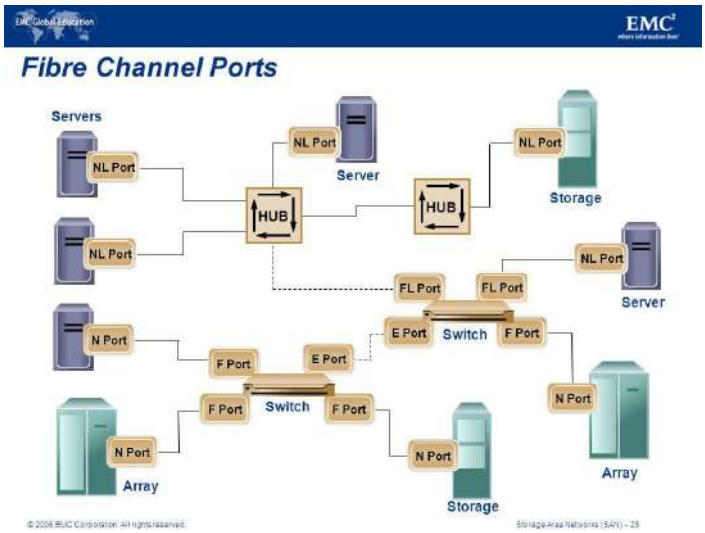



**Figure 5-12: Fibre channel ports**

## 5.5 FIBRE CHANNEL ARCHITECTURE

The FC architecture represents true channel/network integration with standard interconnecting devices. Connections in a SAN are accomplished using FC. Traditionally, transmissions from host to storage devices are carried out over channel connections such as a parallel bus. Channel technologies provide high levels of performance with low protocol overheads. Such performance is due to the static nature of channels and the high level of hardware and software integration provided by the channel technologies. However, these technologies suffer from inherent limitations in terms of the number of devices that can be connected and the distance between these devices.

*Fibre Channel Protocol (FCP)* is the implementation of serial SCSI-3 over an FC network. In the FCP architecture, all external and remote storage devices attached to the SAN appear as local devices to the host operating system. Thekey 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 over 15 million device addresses on a network.
- Exhibits the characteristics of channel transport and provides speeds up to 8.5 Gb/s (8 GFC).

The FC standard enables mapping several existing *Upper Layer Protocols (ULPs)* to FC frames for transmission, including SCSI, IP, High Performance Parallel Interface (HIPPI), Enterprise System Connection (ESCON), and AsynchronousTransfer Mode (ATM).

### 5.5.1 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 6-13 illustrates the fibre channel protocol stack.

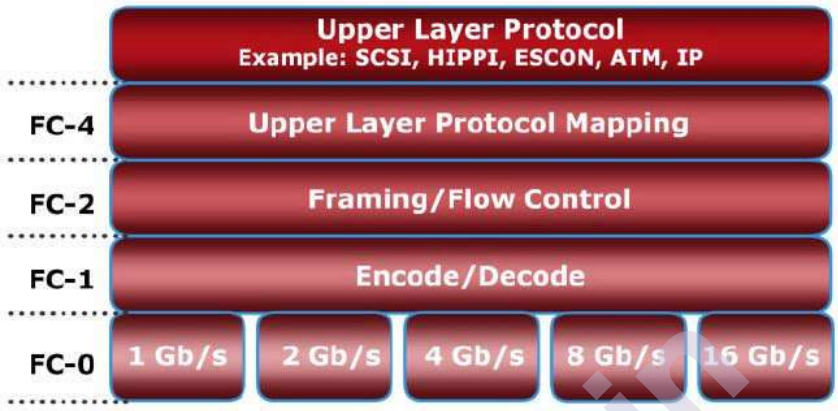



**Figure 5-13**: Fibre channel protocol stack

**FC-4 Upper Layer Protocol**

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 5-13). Some of the protocols include SCSI, HIPPI Framing Protocol, Enterprise Storage Connectivity (ESCON), ATM, and IP.

**FC-2 Transport Layer**

The FC-2 is the transport layer that contains the payload, addresses of the source and destination ports, and link control information. The FC-2 layer pro- vides 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 Transmission Protocol**

This layer defines the transmission protocol that includes serial encoding and decoding rules, special characters used, and error control. 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-0 Physical Interface**

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

### 5.5.2 Fibre Channel Addressing

An FC address is dynamically assigned when a port logs on to the fabric. The FC address has a distinct format that varies according to the type of node port in the fabric. These ports can be an N_port and an NL_port in a public loop, or an NL_port in a private loop.

The first field of the FC address of an N_port contains the domain ID of the switch (see Figure 6-14). This is an 8-bit field. Out of the possible 256 domain IDs, 239 are available for use; the remaining 17 addresses are reserved for specific services. For example, FFFFFC is reserved for the name server, and FFFFFE is reserved for the fabric login service. The maximum possible number of N_ports in a switched fabric is calculated as 239 domains □ 256 areas □ 256 ports = 15,663,104 Fibre Channel addresses.

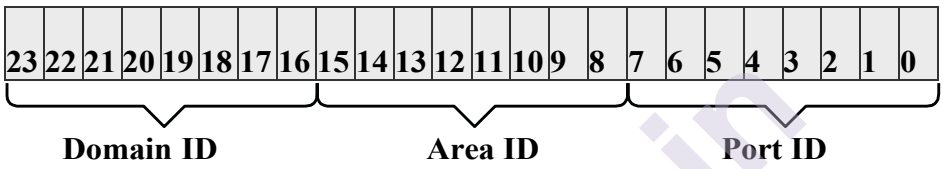


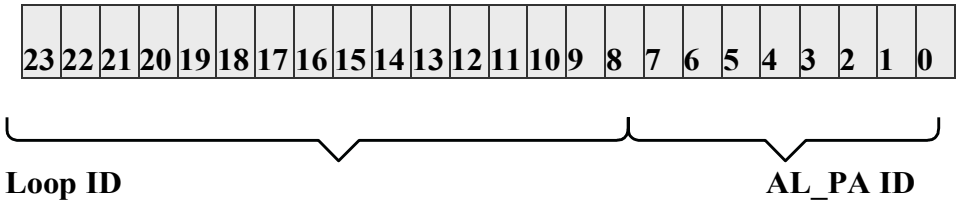

**Figure 5-14:** 24-bit FC address of N_port

The area ID is used to identify a group of F_ports. An example of a group of F_ports would be a card on the switch with more than one port on it. The last field in the FC address identifies the F_port within the group.

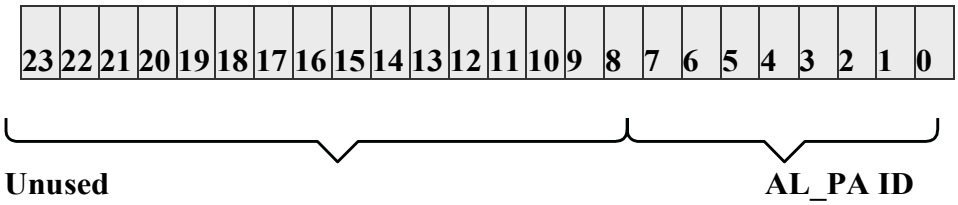
**FC Address of an NL_port**

The FC addressing scheme for an NL_port differs from other ports. The two upper bytes in the FC addresses of the NL_ports in a private loop are assigned zero values. However, when an arbitrated loop is connected to a fabric through an FL_port, it becomes a public loop. In this case, an NL_port supports a fabric login. The two upper bytes of this NL_port are then assigned a positive value, called a *loop identifier*, by the switch. The loop identifier is the same for all NL_ports on a given loop.

Figure 5-15 illustrates the FC address of an NL_port in both a public loop and a private loop. The last field in the FC addresses of the NL_ports, in both public and private loops, identifies the AL-PA. There are 127 allowable AL-PA addresses; one address is reserved for the FL_port on the switch.




**NL_Port, Public Loop**




**NL_Port, Private Loop**

**Figure 5-15:** 24-bit FC address of NL_port

**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). 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. WWNs are *burned* into the hardware or assigned through software. 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 cre- ated FC addresses for nodes. Figure 5-16 illustrates the WWN structure for an array and the 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 |
| | 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 |
| | Reserved 12 bits | | | Company ID 24 bits | | | | | | Company Specific 24 bits | | | | | |

**Figure 5-16:** World Wide Names

### 5.5.3 FC Frame

An FC frame (Figure 6-17) consists of five parts: *start of frame (SOF)*, *frame header*, *data field*, *cyclic redundancy check (CRC)*, and *end of frame (EOF)*.

The SOF and EOF act as delimiters. In addition to this role, the SOF is a flag that indicates whether the frame is the first frame in a sequence of frames.

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.

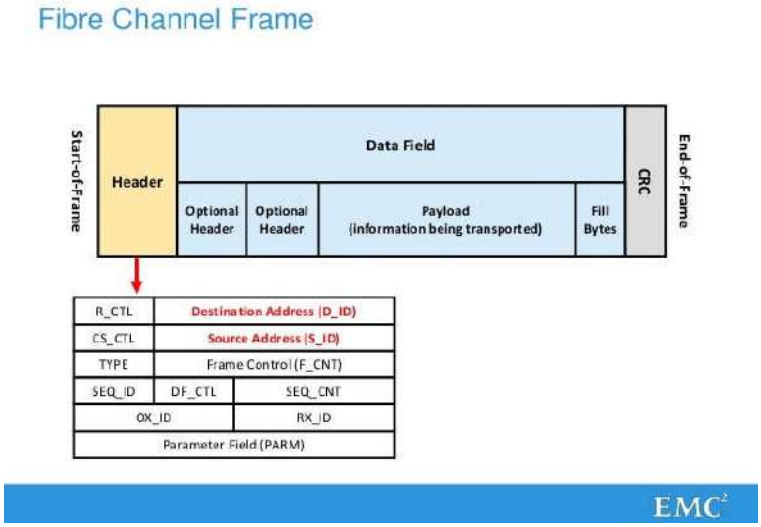



**Figure 5-17: FC frame**

The S_ID and D_ID are standard FC addresses for the source port and the destination port, respectively. The SEQ_ID and OX_ID identify the frame as a component of a specific sequence and exchange, respectively.

The frame header also defines the following fields:

- **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. In contrast, data frames carry the payload and are used for data transmission.
- **Class Specific Control (CS_CTL)**: This field specifies link speeds for class 1 and class 4 data transmission.
- **TYPE**: This field describes the upper layer protocol (ULP) to be carried on the frame if it is a data frame. However, if it is a link control frame, this field is used to signal an event such as "fabric busy." For example, ifthe TYPE is 08, and the frame is a data frame, it means that the SCSI willbe carried on an FC.
- **Data Field Control (DF_CTL)**: A 1-byte field that indicates the existence of any optional headers at the beginning of the data payload. It is a mechanism to extend header information into the payload.
- **Frame Control (F_CTL)**: A 3-byte field that contains control information related to frame content. For example, one of the bits in this field indicateswhether this is the first sequence of the exchange.

The data field in an FC frame contains the data payload, up to 2,112 bytes oforiginal data — in most cases, SCSI data. The biggest possible payload an FC frame can deliver is 2,112 bytes of data with 36 bytes of fixed overhead. A link control frame, by definition, has a payload of 0 bytes. Only

data frames carry a payload. The CRC checksum facilitates error detection for the content of the frame. This checksum verifies data integrity by checking whether the content of the frames was received correctly. The CRC checksum is calculated by the sender before encoding at the FC-1 layer. Similarly, it is calculated by the receiver after decoding at the FC-1 layer.

### 5.5.4 Structure and Organization of FC Data

In an FC network, data transport is analogous to a conversation between two people, whereby a frame represents a word, a sequence represents a sentence,and an exchange represents a conversation.

- **Exchange operation:** An exchange operation enables two N_ports to iden- tify and manage a set of information units. This unit maps to a sequence. Sequences can be both unidirectional and bidirectional depending upon the type of data sequence exchanged between the initiator and the target.
- **Sequence:** A sequence refers to a contiguous set of frames that are sent from one port to another. A sequence corresponds to an information unit, as defined by the ULP.
- **Frame:** A frame is the fundamental unit of data transfer at Layer2. Each frame can contain up to 2,112 bytes of payload.

### 5.5.5 Flow Control

Flow control defines the pace of the flow of data frames during data transmis- sion. FC technology uses two flow-control mechanisms: buffer-to-buffer credit(BB_Credit) and end-to-end credit (EE_Credit).

**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 transmitting port maintains a count of free receiver buffers and continues to send frames if the count is greater than 0. The BB_Credit mechanism provides frame acknowledgment through the *Receiver Ready (R_RDY)* primitive.

EE_Credit

The function of end-to-end credit, known as EE_Credit, is similar to that of BB_Credit. When an initiator and a target establish themselves as nodes communicat- ing with each other, they exchange the EE_Credit parameters (part of Port Login).

The EE_Credit mechanism affects the flow control for class 1 and class 2 trafficonly.

### 5.5.6 Classes of Service

The FC standards define different classes of service to meet the requirements of a wide range of applications. The table below shows three classes of servicesand their features (Table 6-1).

**Table 5-1:** FC Class of Services

| 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 acknowledgement | Acknowledged | Acknowledged | Not acknowledged |
| Multiplexing | No | Yes | Yes |
| Bandwidth utilization | Poor | Moderate | High |

Another class of services is *class F,* which is intended for use by the switches communicating through ISLs. Class F is similar to Class 2, and it provides notification of nondelivery of frames. Other defined Classes 4, 5, and 6 are used for specific applications. Currently, these services are not in common use.

## 5.6 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 (see Figure 5-18). When a device (host or storage array) logs onto a fabric, it is registered with the name server. When a port logs onto the fabric, it goes through a device discovery process with other devices registered in the name server. The zoning function controls this process by allowing only the members in the same zone to establish these link-level services.

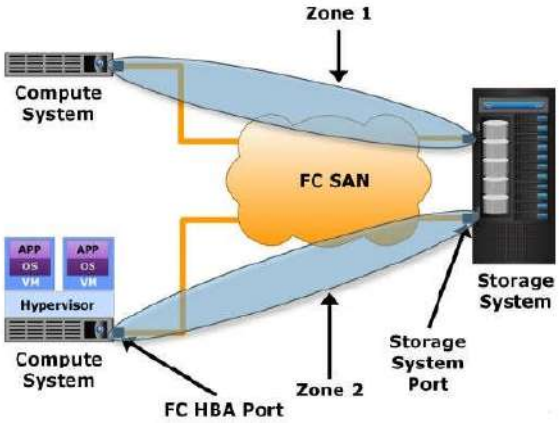



**Figure 5-18: Zoning**

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. Members, zones, and zone sets form the hierarchy defined in the zoning process (see Figure 5-19). *Members* are nodes within the SAN that can be included in a zone. *Zones* comprise a set of members that have access to one another. A port or a node can be a member of multiple zones. *Zone sets* comprise a group of zones that can be activated or deactivated as a single entity in a fabric. Only one zone set per fabric can be active at a time. Zone sets are also referred to as *zone configuration*

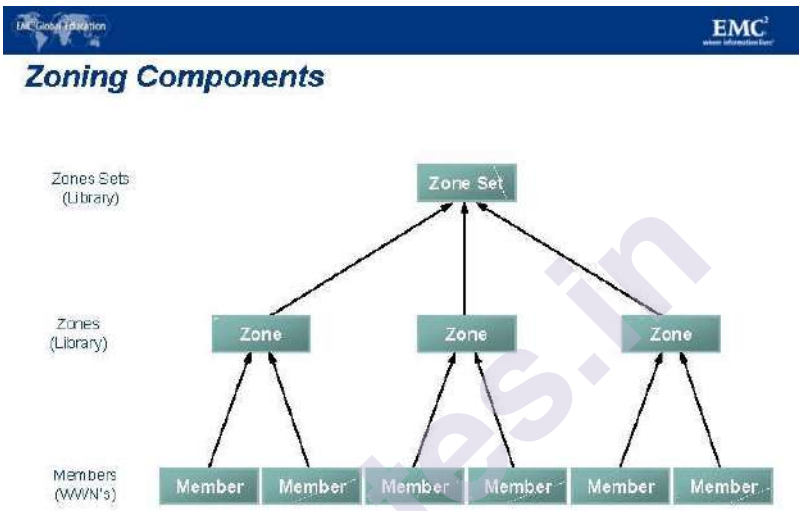



**Fig: 5-19:** Members, zones, and zone set

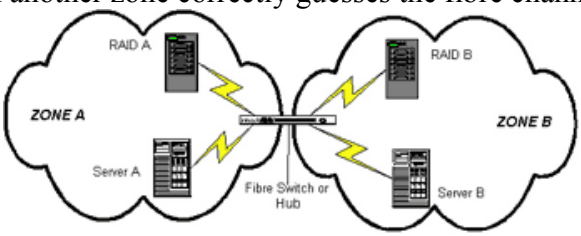
**Types of Zoning**
Zoning can be categorized into three types:

**Hard and Soft Zoning**

Hard zoning is zoning which is implemented in hardware. Soft zoning is zoning which is implemented in software.

Hard zoning physically blocks access to a zone from any device outside of the zone.

Soft zoning uses filtering implemented in fibre channel switches to prevent ports from being seen from outside of their assigned zones. The security vulnerability in soft zoning is that the ports are still accessible if the user in another zone correctly guesses the fibre channel address.



**WWN Zoning**

WWN zoning uses name servers in the switches to either allow or block access to particular World Wide Names (WWNs) in the fabric.

A major advantage of WWN zoning is the ability to recable the fabric without having to redo the zone information.

WWN zoning is susceptible to unauthorized access, as the zone can be bypassed if an attacker is able to spoof the World Wide Name of an authorized HBA.

**Port Zoning**

Port zoning utilizes physical ports to define security zones. A users access to data is determined by what physical port he or she is connected to.

With port zoning, zone information must be updated every time a user changes switch ports. In addition, port zoning does not allow zones to overlap.

Port zoning is normally implemented using hard zoning, but could also be implemented using soft zoning.

Figure 5-20 shows the three types of zoning on an FC network.

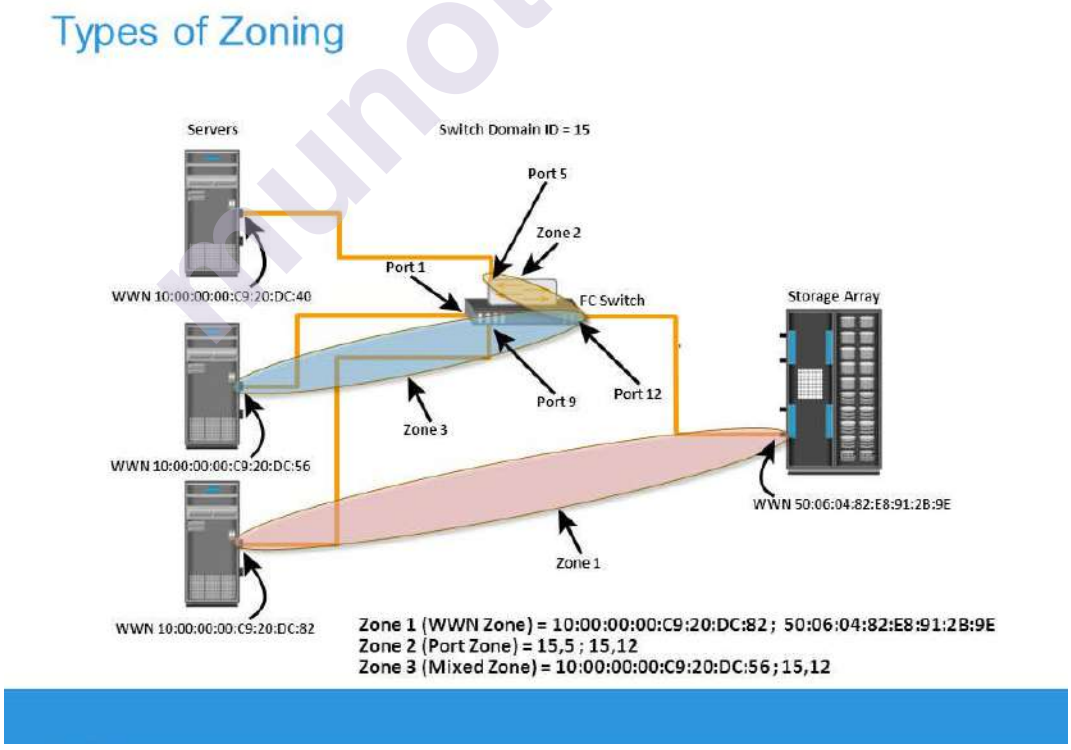



**Figure 5-20: Types of zoning**

Zoning is used in conjunction with LUN masking for controlling server access to storage. However, these are two different activities. Zoning takes place at the fabric level and LUN masking is done at the array level.

## 5.7 FIBRE CHANNEL LOGIN TYPES

**Fabric services define three login types:**

- Fabric login (FLOGI) is performed between an N_port and an F_port. To log on to the fabric, a device sends a FLOGI frame with the World Wide Node Name (WWNN) and World Wide Port Name (WWPN) parameters to the login service at the well-known FC address FFFFFE. In turn, the switch accepts the login and returns an Accept (ACC) frame with the assigned FC address for the device. Immediately after the FLOGI, the N_port registers itself with the local name server on the switch, indicating its WWNN, WWPN, and assigned FC address.

- Port login (PLOGI) is performed between an N_port and another N_port to establish a session. The initiator N_port sends a PLOGI request frame to the target N_port, which accepts it. The target N_port returns an ACC to the initiator N_port. Next, the N_ports exchange service parameters relevant to the session.

Process login (PRLI) is also performed between an N_port and another N_ port. This login relates to the FC-4 ULPs such as SCSI. N_ports exchange SCSI-3-related service parameters. N_ports share information about the FC-4 type in use, the SCSI initiator, or the target.

## 5.8 FC TOPOLOGIES

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

### 5.8.1 Core-Edge Fabric

In the *core-edge fabric* topology, there are two types of switch tiers in this fabric. The *edge tier* usually comprises switches and offers an inexpensive approach to adding more hosts in a fabric. The tier at the edge fans out from the tier at the core. The nodes on the edge can communicate with each other.

The *core tier* usually comprises enterprise directors that ensure high fabric availability. Additionally all traffic has to either traverse through or terminate at this tier. In a two-tier configuration, all storage devices are connected to the core tier, facilitating fan-out. The host-to-storage traffic has to traverse one and two ISLs in a two-tier and three-tier configuration, respectively. Hosts used for mission-critical applications can be connected directly to the core tier and consequently avoid traveling through the ISLs to process I/O requests from these hosts.

The core-edge fabric topology increases connectivity within the SAN while conserving overall port utilization. If expansion is required, an additional edge switch can be connected to the core. This topology can have different variations. In a *single-core topology*, all hosts are connected to the edge tier and all storage is connected to the core tier. Figure 5-21 depicts the core and edge switches in a single-core topology.

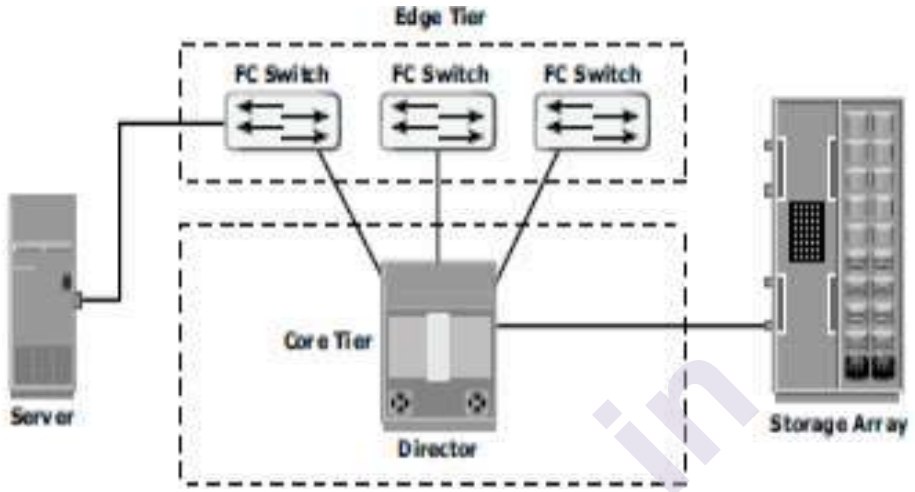



**Fig. 5-21 Single Core Topology**

A *dual-core topology* can be expanded to include more core switches. However, to maintain the topology, it is essential that new ISLs are created to connect each edge switch to the new core switch that is added. Figure 5-22 illustrates the core and edge switches in a dual-core topology.
Figure 5-22 Dual Core topology

**Edge Tier**

**FC Switch FC Switch FC Switch**

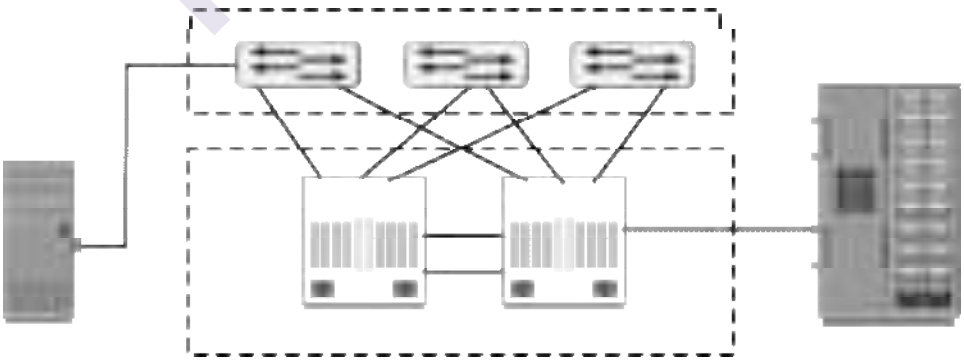


**Fig. 5-22**

**Benefits and Limitations of Core-Edge Fabric**

Thecoreedge fabric provides one-hop storage access to all storage in the system. Because traffic travels in a deterministic pattern (from the edge to the core), a core-edge provides easier calculation of ISL loading and traffic patterns. Because each tier's switch is used for either storage or hosts, one can easily identify which resources are approaching their capacity, making it easier to develop a set of rules for scaling and apportioning.

A well-defined, easily reproducible building-block approach makes rolling out new fabrics easier. Core-edge fabrics can be scaled to larger environments by linking core switches, adding more core switches, or adding more edge switches. This method can be used to extend the existing simple core-edge model or to expand the fabric into a compound or complex core-edge model.

However, the core-edge fabric may lead to some performance-related prob- lems because scaling a core-edge topology involves increasing the number of ISLs in the fabric. As more edge switches are added, the domain count in the fabric increases. A common best practice is to keep the number of host-to-storage hops unchanged, at one hop, in a core-edge. Hop count represents the total number of devices a given piece of data (packet) passes through. Generally a large hop count means greater the transmission delay between data traverse from its source to destination.

As the number of cores increases, it may be prohibitive to continue to main-tain ISLs from each core to each edge switch. When this happens, the fabric design can be changed to a compound or complex core-edge design.

### 5.8.2 Mesh Topology

In mesh topology each device is connected to every other device on the network through a dedicated point-to-point link. When we say dedicated, it means that the link only carries data for the two connected devices only. Let's say we have n devices in the network then each device must be connected with (n-1) devices of the network. Number of links in a mesh topology of n devices would be n(n-1)/2.

**Advantages of Mesh topology:**

1. No data traffic issues as there is a dedicated link between two devices which means the link is only available for those two devices.

2. Mesh topology is reliable and robust as failure of one link doesn't affect other links and the communication between other devices on the network.

3. Mesh topology is secure because there is a point-to-point link thus unauthorized access is not possible.

4. Fault detection is easy.

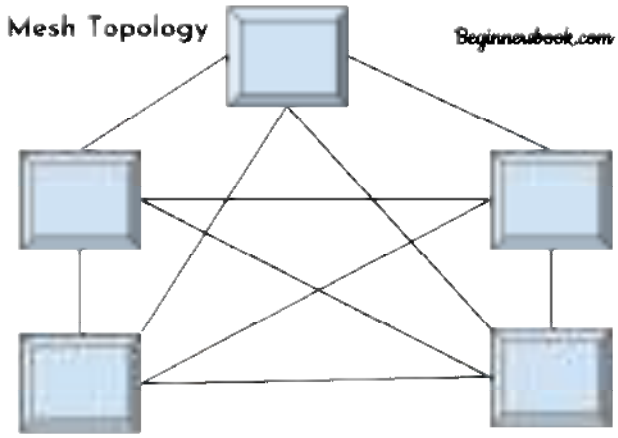



**Figure 5-23: Mesh Topology**

In a mesh topology, each switch is directly connected to other switches by using ISLs. This topology promotes enhanced connectivity within the SAN. When the number of ports on a network increases, the number of nodes that can participate and communicate also increases.

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. 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.

In a partial mesh topology, several hops or ISLs may be required for the traffic to reach its destination. Hosts and storage can be located anywhere in the fabric, and storage can be localized to a director or a switch in both mesh topologies. A full mesh topology with a symmetric design results in an even number of switches, whereas a partial mesh has an asymmetric design and may result in an odd number of switches. Figure 5-24 depicts both a full mesh and a partial mesh topology.

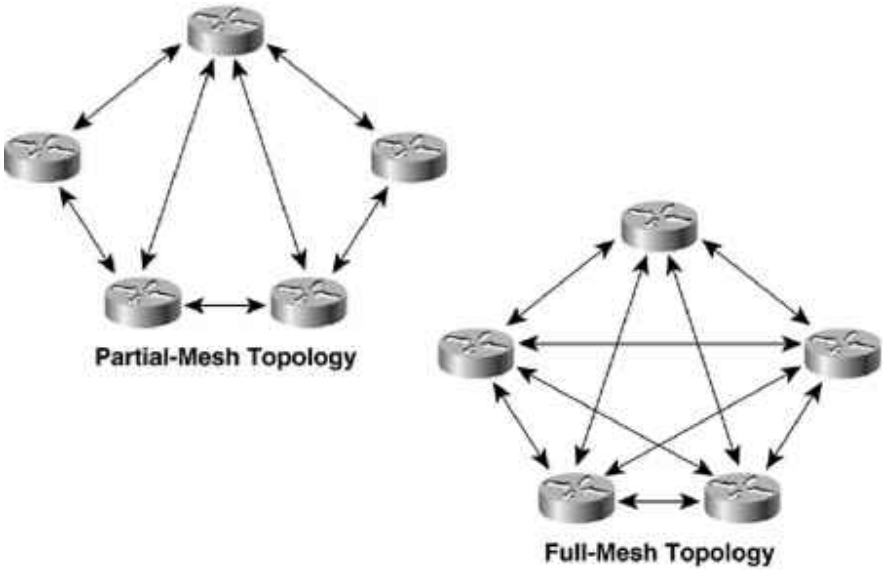



**Figure 5-24:- Partial and Full Mesh Topology**

## 5.9 SUMMARY

The SAN has enabled the consolidation of storage and benefited organizations by lowering the cost of storage service delivery. SAN reduces overall operational cost and downtime and enables faster application deployment. SANs and tools that have emerged for SANs enable data centers to allocate storage to an application and migrate workloads between different servers and storage devices dynamically. This significantly increases server utilization.

SANs simplify the business-continuity process because organizations are able to logically connect different data centers over long distances and provide cost-effective, disaster recovery services that can be effectively tested.

The adoption of SANs has increased with the decline of hardware prices and has enhanced the maturity of storage network standards. Small and medium size enterprises and departments that initially resisted shared storage pools have now begun to adopt SANs.

This chapter detailed the components of a SAN and the FC technology that forms its backbone. FC meets today's demands for reliable, high-performance, and low-cost applications.

The interoperability between FC switches from different vendors has enhanced significantly compared to early SAN deployments. The standards published by a dedicated study group within T11 on SAN routing, and the new product offerings from vendors, are now revolutionizing the way SANs are deployed and operated.

Although SANs have eliminated islands of storage, their initial implementation created islands of SANs in an enterprise. The emergence of the iSCSI and FCIP technologies, detailed in Chapter 6, has pushed the convergence of the SAN with IP technology, providing more benefits to using storage technologies.

## 5.10 QUESTIONS:

1. Give an overview of Fibre Channel.hub.
2. What is SAN? How is it implemented?
3. Explain the components of SAN.
4. What are the different types of connectors? Explain each.
5. Compare FC switch and FC Hub.
6. State and explain the different FC connectivity options.
7. Discuss the FC-SW data transmission.
8. What are the different fibre channel ports? Explain.
9. Explain the fibre channel protocol stack.
10. Discuss the fibre channel addressing.
11. With the help of a diagram, explain the FC frame.
12. Explain the flow control in FC technology.
13. What are the different classes of FC service? Explain each.
14. What is Zoning? What are its different types? Explain.
15. What are the different login types defined by fabric services? Explain each.
16. State and explain the different FC topologies to connect devices.

## 5.11 REFERENCES

**1.** Data Center Virtualization Fundamentals, Gustavo Alessandro Andrade Santana, Cisco Press 1st Edition 2014.

# 6

# IP SAN

**Unit Structure**



## 6.1 ISCSI PROTOCOL

Traditional SAN environments allow block I/O over Fibre Channel, whereas NAS environments allow file I/O over IP-based networks. Organizations need the performance and scalability of SAN plus the ease of use and lower TCO of NAS solutions. The emergence of IP technology that supports block I/O over IP has positioned IP for storage solutions.

IP offers easier management and better interoperability. When block I/O is run over IP, the existing network infrastructure can be leveraged, which is more economical than investing in new SAN hardware and software. Many long-distance, disaster recovery (DR) solutions are already leveraging IP-based networks. In addition, many robust and mature security options are now avail- able for IP networks. With the advent of block storage technology that leverages IP networks (the result is often referred to as IP SAN), organizations can extend the geographical reach of their storage infrastructure.

IP SAN technologies can be used in a variety of situations. Figure 6-1 illustrates the co-existence of FC and IP storage technologies in an organization where mission-critical applications are serviced through FC, and business- critical applications and remote office applications make use of IP SAN. Disaster recovery solutions can also be implemented using both of these technologies.

Two primary protocols that leverage IP as the transport mechanism are iSCSIand Fibre Channel over IP (FCIP).

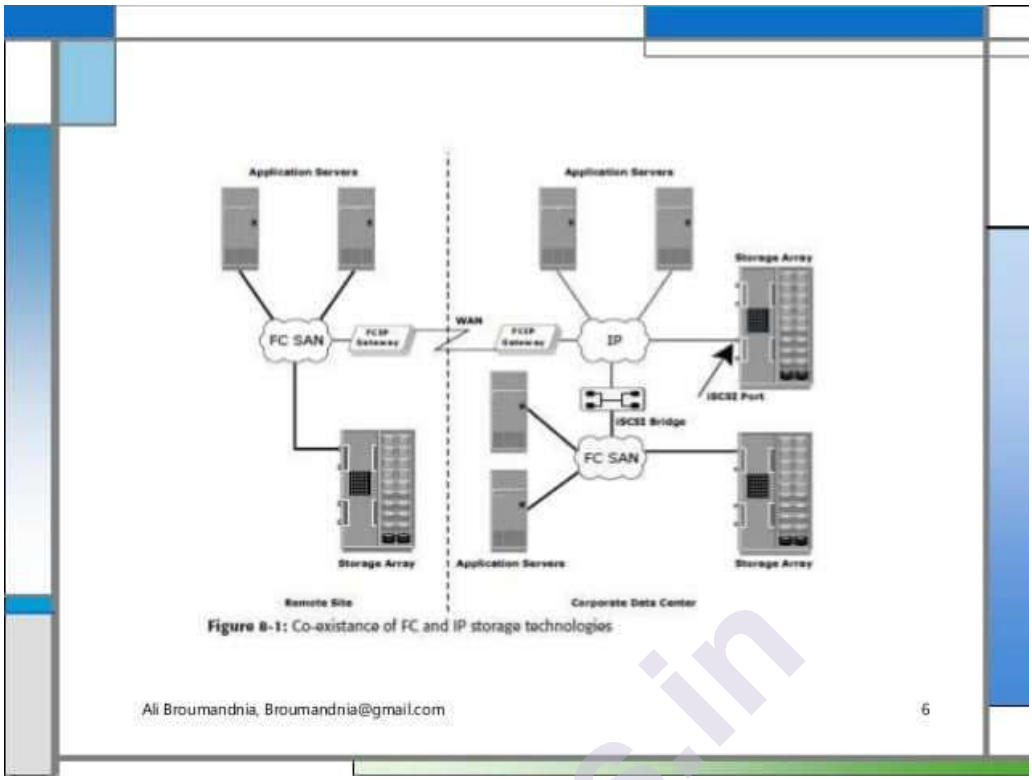
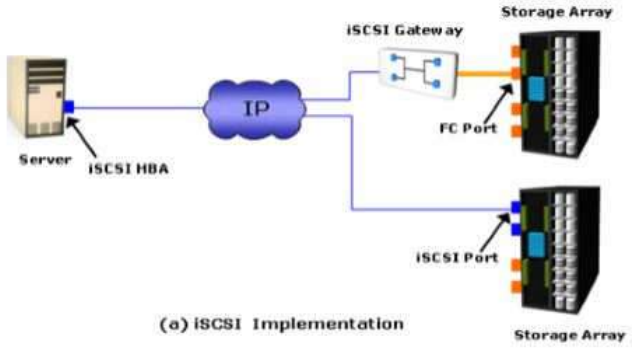


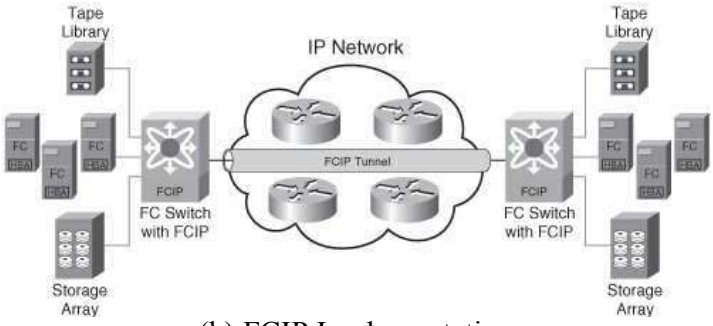

**Figure 6-1:** Co-existance of FC and IP storage technologies

iSCSI is the host-based encapsulation of SCSI I/O over IP using an EthernetNIC card or an iSCSI HBA in the host. As illustrated in Figure 6-2 (a), IP traffic is routed over a network either to a gateway device that extracts the SCSI I/O from the IP packets or to an iSCSI storage array. The gateway can then send the SCSI I/O to an FC-based external storage array, whereas an iSCSI storage arraycan handle the extraction and I/O natively.

FCIP uses a pair of bridges (FCIP gateways) communicating over TCP/IP as the transport protocol. FCIP is used to extend FC networks over distances and/or anexisting IP-based infrastructure, as illustrated in Figure 6-2 (b).

Today, iSCSI is widely adopted for connecting servers to storage because it is relatively inexpensive and easy to implement, especially in environments where an FC SAN does not exist. FCIP is extensively used in disaster-recoveryimplementations, where data is duplicated on disk or tape to an alternate site. This chapter describes iSCSI and FCIP protocols, components and topologies in detail.






(b) FCIP Implementation

**Figure 6-2:** iSCSI and FCIP implementation

**ISCSI**

iSCSI is an IP-based protocol that establishes and manages connections between storage, hosts, and bridging devices over IP. iSCSI carries block-level data over IP-based networks, including Ethernet networks and the Internet. iSCSI is built on the SCSI protocol by encapsulating SCSI commands and data inorder to allow these encapsulated commands and data blocks to be transportedusing TCP/IP packets.

### 6.1.1 Components of iSCSI

Host (initiators), targets, and an IP-based network are the principal iSCSI components. The simplest iSCSI implementation does not require any FC components. If an iSCSI-capable storage array is deployed, a host itself can act as an iSCSI initiator, and directly communicate with the storage over an IP network. However, in complex implementations that use an existing FC array for iSCSI connectivity, iSCSI gateways or routers are used to connect the existing FC SAN. These devices perform protocol translation from IP packets to FC packets and vice-versa, thereby bridging connectivity between the IP and FC environments.

### 6.1.2 iSCSI Host Connectivity

iSCSI host connectivity requires a hardware component, such as a NIC with asoftware component (iSCSI initiator) or an iSCSI HBA. In order to use the iSCSI protocol, a software initiator or a translator must be installed to route the SCSIcommands to the TCP/IP stack.

A standard NIC, a TCP/IP offload engine (TOE) NIC card, and an iSCSI HBAare the three physical iSCSI connectivity options.

A standard NIC 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. However, the NIC provides no external processing power, which places additional overhead on the host CPU because it is required to perform all the TCP/IP and iSCSI processing.

If a standard NIC is used in heavy I/O load situations, the host CPU may become a bottleneck. *TOE NIC* help alleviate this burden. A

TOE NIC offloads the TCP management functions from the host and leaves 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, requiring host CPU cycles.

An *iSCSI HBA* is capable of providing performance benefits, as it offloads the entire iSCSI and TCP/IP protocol stack from the host processor. Use of an iSCSI HBA is also the simplest way for implementing a boot from SAN environment via iSCSI. If there is no iSCSI HBA, modifications have to be made to the basic operating system to boot a host from the storage devices because the NIC needs to obtain an IP address before the operating system loads. The functionality of an iSCSI HBA is very similar to the functionality of an FC HBA, but it is the most expensive option.

A fault-tolerant host connectivity solution can be implemented using host- based multipathing software (e.g., EMC Power Path) regardless of the type of physical connectivity. Multiple NICs can also be combined via link aggregation technologies to provide failover or load balancing. Complex solutions may also include the use of vendor-specific storage-array software that enables the iSCSI host to connect to multiple ports on the array with multiple NICs or HBAs.

### 6.1.3 Topologies for iSCSI Connectivity

- The topologies used to implement iSCSI can be categorized into two classes:

1. Native
2. Bridged

- Native topologies do not have any FC components; they per-form all communication over IP. The initiators may be either directly attached to targets or connected using standard IP routers and switches.
- Bridged topologies enable the co-existence 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 SAN.

## 6.2 NATIVE ISCSI CONNECTIVITY:

1. If an iSCSI-enabled array is deployed, FC components are not needed for iSCSI connectivity in the native topology. In the example shown in Figure(a), the array has one or more Ethernet NICs that are connected to a standard Ethernet switch and configured with an IP address and listening port.
2. Once a client/ initiator is configured with the appropriate target information, it connects to the array and requests a list of available LUNs. A single array port can service multiple hosts or initiators as long as the array can handle the amount of storage traffic that the hosts generate. the array and requests a list of available LUNs. A single array port can service multiple hosts or initiators as long as the array can handle the amount of storage traffic that the hosts generate.

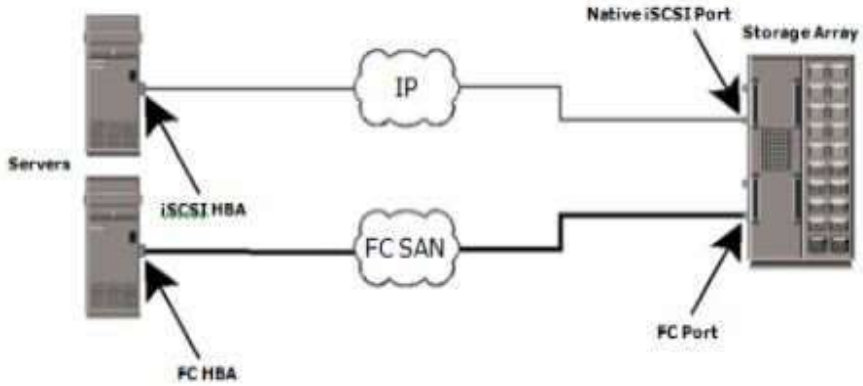



Figure (a): Native iSCSI Connectivity.

**1. Bridged iSCSI Connectivity:**

- A bridged iSCSI implementation includes FC components in its configuration. Following figure (b) illustrates an existing FC storage array used to service hosts connected through iSCSI.

1. The array does not have any native iSCSI capabilities—that is, it does not have any Ethernet ports. Therefore, an external device, called a bridge, router, gateway, or a multi-protocol router, must be used to bridge the communication from the IP network to the FC SAN.
2. In this configuration, the bridge device has Ethernet ports connected to the IP network, and FC ports connected to the storage. These ports are assigned IP addresses, similar to the ports on an iSCSI-enabled array.
3. The iSCSI initiator/host is configured with the bridge's IP address as its target destination. The bridge is also configured with an FC initiator or multiple initiators.

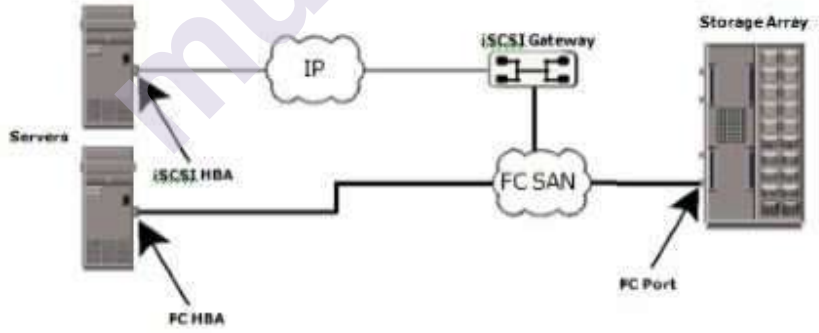



Figure (b): Bridged iSCSI Connectivity

**2. iSCSI Protocol Stack**

The architecture of iSCSI is based on the client/server model. Figure 6-4 displays a model of the iSCSI protocol layers and depicts the encapsulation order of SCSI commands for their delivery through a physical carrier.

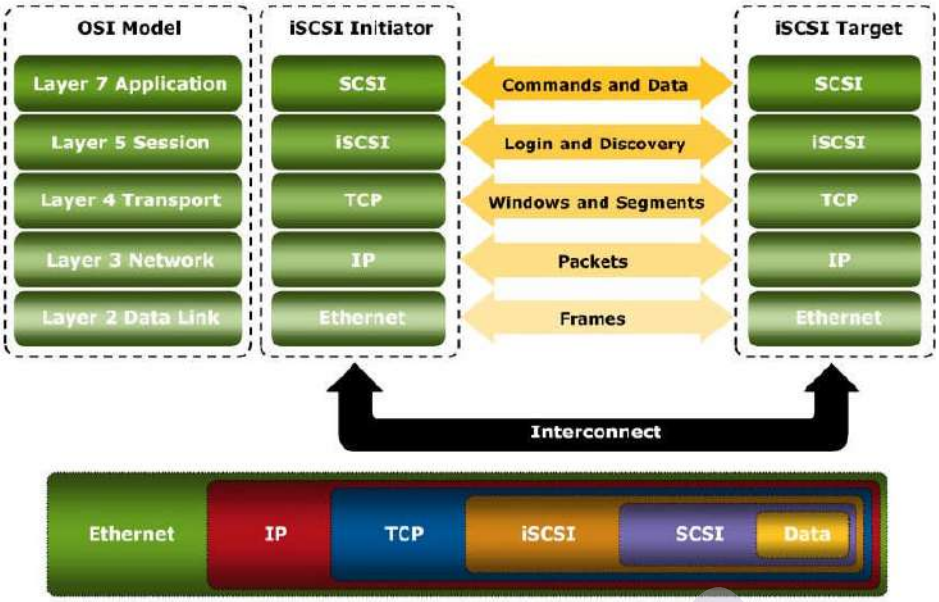



**Figure 6-4: iSCSI protocol stack**

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

iSCSI is the session-layer protocol that initiates a reliable session between adevice that recognizes SCSI commands and TCP/IP. The iSCSI session-layer interface is responsible for handling login, authentication, target discovery, and session management. TCP is used with iSCSI at the transport layer to providereliable service.

TCP is used to control message flow, windowing, error recovery, and retrans- mission. It relies upon the network layer of the OSI model to provide global addressing and connectivity. The layer-2 protocols at the data link layer of this model enable node-to-node communication for each hop through a separate physical network.

Communication between an iSCSI initiator and target is detailed next.

### 3. iSCSI Discovery

An initiator must discover the location of the target on a 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* and *internet Storage Name Service (iSNS)*.

In Send Targets discovery, the initiator is manually configured with the tar- get's network portal, which it uses to establish a discovery session with the iSCSI service on the target. The initiator issues the Send Targets command, and the target responds with the names and addresses of the targets availableto the host.

iSNS (see Figure 8-5) enables the 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.

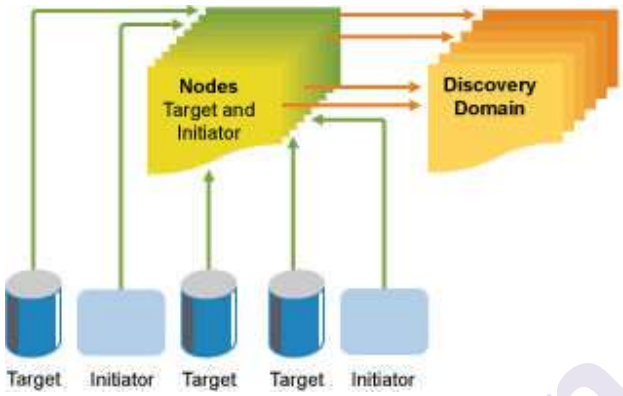



**Figure 6-5: Discovery using iSNS**

Discovery can also take place by using Service Location Protocol (SLP). However, this is less commonly used than SendTargets discovery and iSNS.

### 4. iSCSI Names

A unique worldwide iSCSI identifier, known as an *iSCSI name*, is used to name the initiators and targets within an iSCSI network to facilitate communication. The unique identifier can be a combination of department, application, manufacturer name, serial number, asset number, or any tag that can be used to recognize and manage a storage resource. There are two types of iSCSI names:

- **iSCSI Qualified Name (IQN):** An organization must own a registered domain name in order to generate iSCSI Qualified Names. This domain name does not have 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 transfer of domain names; the organization is required to have owned the domain name on that date. 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 of the storage device identifiers.

- **Extended Unique Identifier (EUI):** An EUI is a globally unique iden- tifier based on the IEEE EUI-64 naming standard. An EUI comprises the eui prefix followed by a 16-character hexadecimal name, such as eui.0300732A32598D26.

The 16-character part of the name includes 24 bits for the company name assigned by IEEE and 40 bits for a unique ID, such as a serial number. This allows for a more streamlined, although less user-friendly,

name string because the resulting iSCSI name is simply eui followed by the hexadecimal WWN.

In either format, the allowed special characters are dots, dashes, and blank spaces. The iSCSI Qualified Name enables storage administrators to assign meaningful names to storage devices, and therefore manage those devices more easily.

Network Address Authority (NAA) is an additional iSCSI node name type to enable worldwide naming format as defined by the InterNational Committee for Information Technology Standards (INCITS) T11 - Fibre Channel (FC) pro- tocols and used by Serial Attached SCSI (SAS). This format enables SCSI storage devices containing both iSCSI ports and SAS ports to use the same NAA-based SCSI device name. This format is defined by RFC3980, "T11 Network Address Authority (NAA) Naming Format for iSCSI Node Names."

**5. iSCSI Session**

An iSCSI session is established between an initiator and a target. A session ID (SSID), which includes an initiator ID (ISID) and a target ID (TSID), identifies a session. The session can be intended for one of the following:

- Discovery of available targets to the initiator and the location of a specific target on a network
- Normal operation of iSCSI (transferring data between initiators and targets)

TCP connections may be added and removed within a session. Each iSCSI connection within the session has a unique connection ID (CID).

**6. iSCSI PDU**

iSCSI initiators and targets communicate using iSCSI Protocol Data Units (PDUs). 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 6-6. The IP header provides packet-routing information that is used to move the packet across a network. The TCP header contains the information needed to guarantee the packet's delivery to the target. The iSCSI header describes how to extract SCSI commands and data for the target. iSCSI adds an optional CRC, known as the *digest*, beyond the TCP checksum and Ethernet CRC to ensure datagram integrity. The header and the data digests are optionally used in the PDU to validate integrity, data placement, and correct operation.

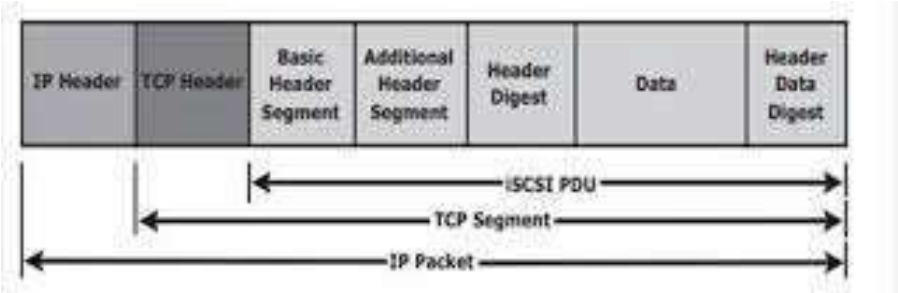



**Figure 6-6:** iSCSI PDU encapsulated in an IP packet

As shown in Figure 6-7, each iSCSI PDU does not correspond in a 1:1 relationshipwith an IP packet. Depending on its size, an iSCSI PDU can span an IP packet oreven coexist with another PDU in the same packet. Therefore, each IP packet andEthernet frame can be used more efficiently because fewer packets and frames arerequired to transmit the SCSI information.

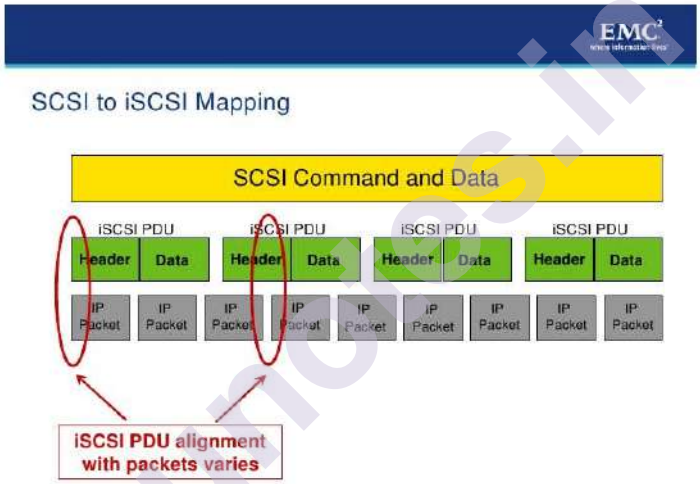



**Figure 6-7: Alignment of iSCSI PDUs with IP packets**

## 7. Ordering and Numbering

iSCSI communication between 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 to number all initiator-to-target command PDUs belonging to the session. This num- ber is used to ensure that every command is delivered in the same order in which it is transmitted, regardless of the TCP connection that carries the command inthe session.

Command sequencing begins with the first login command and the CmdSN is incremented by one for each subsequent command. The iSCSI target layer isresponsible for delivering the commands to the SCSI layer in the order of their CmdSN. This ensures the correct order of data and commands at a target even when there are multiple TCP connections between an initiator and the target using portal groups.

Similar to command numbering, a *status sequence number (StatSN)* is used to sequentially number status responses, as shown in Figure 6-8.

These unique numbers are established at the level of the TCP connection.

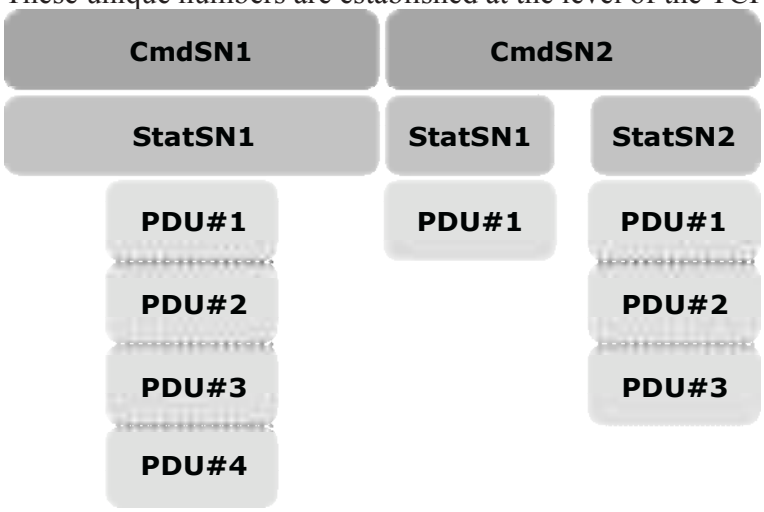



**Figure 6-8:** Command and status sequence number

A target sends the *request-to-transfer (R2T)* PDUs to the initiator when it is ready to accept data. *Data sequence number (DataSN)* is used to ensure in-order delivery of data within the same command. The DataSN and R2T sequence numbers are used to sequence data PDUs and R2Ts, respectively. Each of these sequence numbers is stored locally as an unsigned 32-bit integer counter defined by iSCSI. These numbers are communicated between the initiator and target in the appropriate iSCSI PDU fields during command, status, and data exchanges.

In the case of read operations, the DataSN begins at zero and is incremented by one for each subsequent data PDU in that command sequence. In the case of a write operation, the first unsolicited data PDU or the first data PDU in response to an R2T begins with a DataSN of zero and increments by one for each subsequent data PDU. R2TSN is set to zero at the initiation of the command and incremented by one for each subsequent R2T sent by the target for that command.

## 8. iSCSI Error Handling and Security

The iSCSI protocol addresses errors in IP data delivery. Command sequencing is used for flow control, the missing commands, and responses, and data blocks are detected using sequence numbers. Use of the optional digest improves com- munication integrity in addition to TCP checksum and Ethernet CRC.

The error detection and recovery in iSCSI can be classified into three levels: Level 0 = Session Recovery, Level 1 = Digest Failure Recovery and Level 2 = Connection Recovery. The error-recovery level is negotiated during login.

- **Level 0:** If an iSCSI session is damaged, all TCP connections need to be closed and all tasks and unfulfilled SCSI commands should be completed. Then, the session should be restarted via the repeated login.
- **Level 1:** Each node should be able to selectively recover a lost or damaged PDU within a session for recovery of data transfer. At this level,

identifica- tion of an error and data recovery at the SCSI task level is performed, and an attempt to repeat the transfer of a lost or damaged PDU is made.

- **Level 2:** New TCP connections are opened to replace a failed connection. The new connection picks up where the old one failed.

iSCSI may be exposed to the security vulnerabilities of an unprotected IP network. Some of the security methods that can be used are IPSec and authentication solutions such as Kerberos and CHAP (challenge-handshake authentication protocol).

## 6.3 FCIP

Organizations are now looking for new ways to transport data throughout the enterprise, locally over the SAN as well as over longer distances, to ensure that data reaches all the users who need it. One of the best ways to achieve this goal is to interconnect geographically dispersed SANs through reliable, high-speed links. This approach involves transporting FC block data over the existing IP infrastructure used throughout the enterprise.

The FCIP standard has rapidly gained acceptance as a manageable, cost- effective way to blend the best of two worlds: FC block-data storage and the proven, widely deployed IP infrastructure. FCIP is a tunneling protocol that enables distributed FC SAN islands to be transparently interconnected over existing IP-based local, metropolitan, and wide-area networks. As a result, organizations now have a better way to protect, store, and move their data while leveraging investments in existing technology.

FCIP uses TCP/IP as its underlying protocol. In FCIP, the FC frames are encapsulated onto the IP payload, as shown in Figure 6-9. FCIP does not manipulate FC frames (translating FC IDs for transmission).

When SAN islands are connected using FCIP, each interconnection is called an *FCIP link*. A successful FCIP link between two SAN islands results in a fully merged FC fabric.

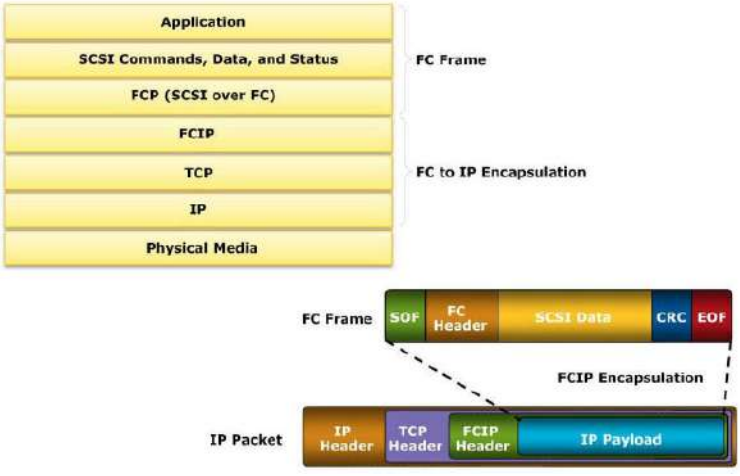
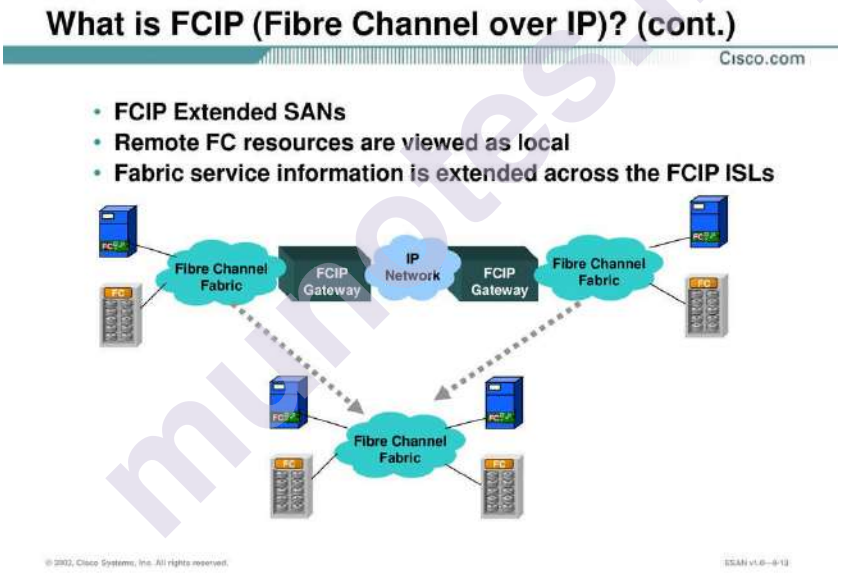



**Figure 6-9: FCIP encapsulation**

## 1. FCIP Topology

An FCIP environment functions as if it is a single cohesive SAN environment. Before geographically dispersed SANs are merged, a fully functional layer 2 network exists on the SANs. This layer 2 network is a standard SAN fabric. These physically independent fabrics are merged into a single fabric with an IP link between them.

An FCIP gateway router is connected to each fabric via a standard FC connection (see Figure 6-10). The fabric treats these routers like layer 2 fabric switches. The other port on the router is connected to an IP network and an IP address is assigned to that port. This is similar to the method of assigning an IP address to an iSCSI port on a gateway. Once IP connectivity is established, the two independent fabrics are merged into a single fabric. When merging the two fabrics, all the switches and routers must have unique domain IDs, and the fabrics must contain unique zone set names. Failure to ensure these requirements will result in a segmented fabric. The FC addresses on each side of the link are exposed to the other side, and zoning or masking can be done to any entity in the new environment.




## 2. FCIP Performance and Security

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

From the perspective of performance, multiple paths to multiple FCIP gate- ways from different switches in the layer 2 fabric eliminates single points of failure and provides increased bandwidth. In a scenario of extended distance, the IP network may be a bottleneck if sufficient bandwidth is not available. In addition, because FCIP creates a unified fabric, disruption in the underlying IP network can cause instabilities in the SAN environment. These include a segmented fabric, excessive RSCNs, and host timeouts.

The vendors of FC switches have recognized some of the drawbacks related to FCIP and have implemented features to provide additional stability, such as the capability to segregate FCIP traffic into a separate virtual fabric.

Security is also a consideration in an FCIP solution because the data is transmitted over public IP channels. Various security options are available to protect the data based on the router's support. IPSec is one such security measure that can be implemented in the FCIP environment.

## 6.4 SUMMARY

iSCSI has enabled IT organizations to gain the benefits of storage networking architecture at reasonable costs. Storage networks can now be geographically distributed with the help of hybrid IP SAN technology, which enhances storage utilization across enterprises. FCIP has emerged as a solution for implementing viable business continuity across enterprises.

Because IP SANs are based on standard Ethernet protocols, the concepts, security mechanisms, and management tools are familiar to administrators. This has enabled the rapid adoption of IP SAN in organizations. The block-level I/O requirements of certain applications that cannot be made with NAS can be targeted for implementation with iSCSI.

This chapter detailed the two IP SAN technologies, iSCSI and FCIP. The next chapter focuses on CAS, another important storage networking technology that addresses the online storage and retrieval of content and long-term archives.

## 6.5 QUESTIONS:

1. How do FC and IP storage technologies coexist? Explain.
2. What is iSCSI? What are its components?
3. Explain the iSCSI host connectivity.
4. State and explain the topologies for iSCSI connectivity.
5. Explain the iSCSI protocol stack.
6. How does discovery take place in iSCSI?
7. State and explain the two types of iSCSI names.
8. With the help of a diagram, explain the iSCSI PDU encapsulated in an IP packet.
9. Explain the ordering and numbering in iSCSI.
10. How are errors and security handled in iSCSI? Explain.
11. What is FCIP? Explain.
12. Explain the FCIP topology.
13. Discuss the FCIP performance and security.

## 6.6 REFERENCES

**1.** Data Center Virtualization Fundamentals, Gustavo Alessandro Andrade Santana, Cisco Press 1st Edition 2014.

# 7

# NETWORK-ATTACHED STORAGE

**Unit Structure**

## 7.0 OBJECTIVE

In this chapter we will study Network-based file sharing which support flexibility to share files over long distances among a large number of users.A NAS device provides file-serving functions such as storing, retrieving, and accessing files for applications and clients. We will see benefits of NAS in detail. There are different ways to access and process a file in networking. We will study two main components of NAS (NAS head and storage).

NAS I/O operation and NAS implementation is covered in this chapter. Most NAS devices support multiple file-service protocols to handle file I/O requests to a remote file system. As we know that NFS and CIFS are the common protocols for file sharing. We will also cover facts that affect NAS Performance. Tremendous growth of unstructured data has posed new challenges to IT administrators and storage managers.These challenges demand a smarter approach to manage unstructured data based on its content rather than metadata about its name, location, and so on.*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.This chapter details object-based storage, its components, and operation. It also details *content addressed storage* (CAS), a special type of OSD. Further, this chapter covers the components and data access method in unified storage.

## 7.1 INTRODUCTION

File sharing means providing common file access to more than one user.One method for file sharing iscopying files to portable media such as CD, DVD, or USB drives and providingthem to all users who want to access it. But this method is not suitable when common file to be shared among large number of users at different locations.This problem can be solved by Network-based file sharing which provides the flexibility to share files over long distances among a large number of users. Client-Server technology is used for file sharing over a network. File servers are used to store files to be shared among users.

These servers are either connected to direct-attached storage (DAS) orstorage area network (SAN)-attached storage. But a SAN is a poor choice if an organization lacks the financial resources to purchase, deploy and maintain it.And if there is a lot of traffic in the storage area network, then operations will be extremely slow. NAS devices are rapidly becoming popular with enterprise and small businesses in many industries as an effective, scalable, low-cost storage solution.

An NAS device is a storage device connected to a network that allows storage and retrieval of data from a central location for authorized network users and varied clients. NAS devices are flexible and scale out,

meaning that as you need additional storage, you can add to what you have. NAS is like having a private cloud in the office. It's faster, less expensive and provides all the benefits of a public cloud on site, giving you complete control.

## 7.2 GENERAL-PURPOSE SERVERS VERSUS NAS DEVICES

A NAS device provides file-serving functions such as storing,retrieving, and accessing files for applications and clients. A general-purpose server can be used to host any application because it runs a general-purpose operating system. Unlike a general-purpose server, a NAS device is dedicated to file-serving. It has specialized operating system dedicated to file serving by using industry-standard protocols, as shown in Figure 7-1.

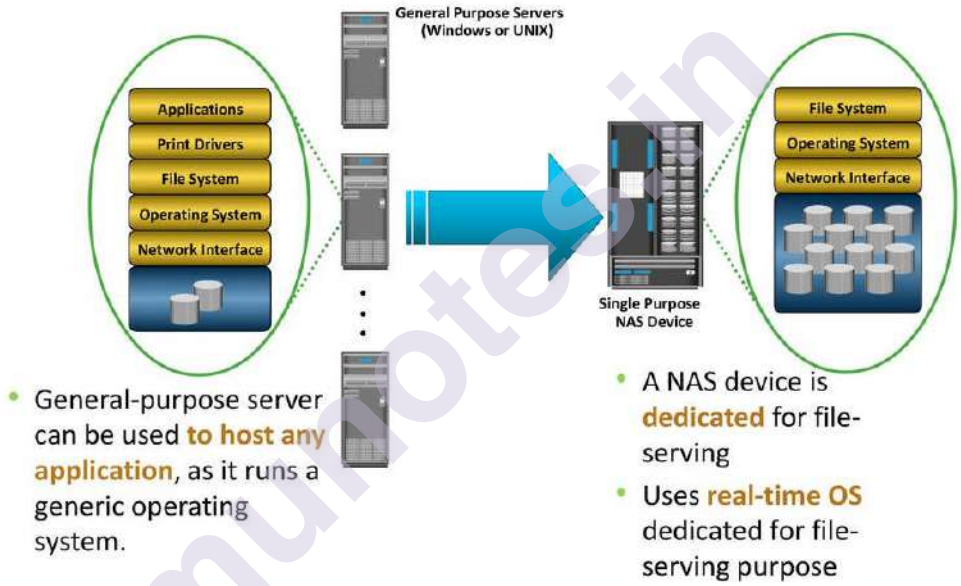



***Figure 7-1General-Purpose Servers versus NAS Devices***

## 7.3 ADVANTAGES/BENEFITS OF NAS

NAS has the following benefits:

- **Comprehensive access to information:** NAS enables efficient file sharing and supports many-to-one and one-to-many configurations. The many-to-one configuration enables a NAS device to serve many clients simultaneously. The one-to-many configuration enables one client to connect with many NAS devices simultaneously.
- **Improved efficiency:** NAS delivers better performance compared to a general-purpose file server because NAS uses an operating system specialized for file serving.

- **Improved flexibility:** NAS is compatible with clients on both UNIX and Windows platforms.
- **Centralized storage:** NASminimizes data duplication on client workstations by centralizing data storage. It also provides better data protection.
- **Simplified management:** To manage file systems efficiently NAS Provides a centralized console.
- **Scalability:** Because of the high-performance and low-latency,the devices of NAS are scalable and can be easily accessed remotely.
- **High availability:** NAS offers efficient replication and recovery options, enabling high data availability. A NAS device supports clustering technology for failover.
- **Security:**NAS ensures security, user authentication, and file locking with industry-standard security schemas.
- **Low cost:** NAS uses commonly available and inexpensive Ethernet components.
- **Ease of deployment:** Configuration at the client is minimal, because the clients have required NAS connection software built in.

## 7.4 FILE SYSTEMS AND NETWORK FILE SHARING

A file system is a process that manages how and where data on a storage disk is stored, accessed and managed.Many file systems maintain a file access table to simplify the process of searching and accessing files.

### 7.4.1 Accessing a File System

Before using a file system, it must be mounted. During the boot process, the operating system mounts a local file system. The mount process creates a link between the file system on the NAS and the operating system on the client. The operating system organizes files and directories in a tree-like structure and grants the privilege to the user to access this structure. The tree is rooted at a mount point.The mount point is named using operating system conventions. Users and applications can access the entire tree from the root to the leaf nodes as file system permissions allow. Files are located at leaf nodes, and directories and subdirectories are located at intermediate roots. The file system is uncounted when access to the file system terminated. Figure 7-2 shows general directory structure.

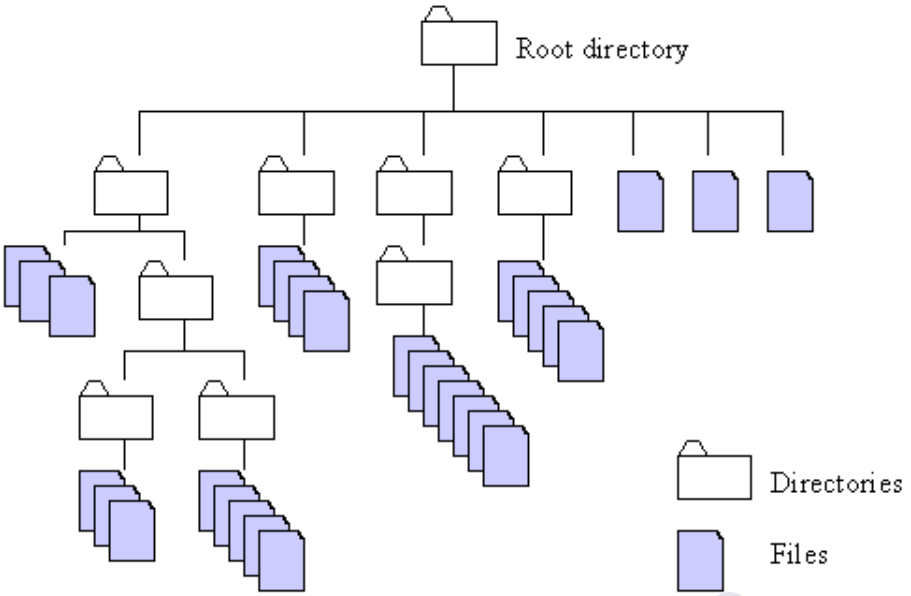



***Figure 7-2: General Directory Structure***

### 7.4.2 Network File Sharing

Network file sharing refers to storing and retrieving files over a network. In afile-sharing environment, the creator or owner ofa file determines the type of access like read, write, execute, append, etc. to be given to other users. When multiple users try to access a shared file simultaneously a locking scheme is provided to maintain data integrity and, at the same time, make this sharing possible.

Some examples of file-sharing methods are file transfer protocol (FTP),Distributed File System (DFS) and the peer-to-peer (P2P) model.

*FTP* is a client-server protocol for transmitting file over a network. AnFTP server and an FTP client communicate with each other using TCP/IP protocol connections. It's also one of the oldest protocols in use today and is a convenient way to move files around.FTP is not a secure method ofdata transfer because it uses unencrypted data transfer over a network. Secure Shell (SSH) adds security to the original FTP specification, which is referred to as Secure FTP (SFTP).

A Distributed File System (DFS) is a file system that is distributed on multiple file servers or multiple locations. It allows programs to access or store isolated files as they do with the local ones, allowing programmers to access files from any network or computer. Standard client-server file sharing protocols enable the owner of a file to set the required type of access, such as read-only or read-write, for a particular user or group of users. Using this protocol, the clients mount remote file systems that are available on dedicated file servers.

A *peer-to-peer* (P2P) file sharing model uses a peer-to-peer network.P2P enables client machines to directly share files with each other over a network. Clients use a file sharing software that searches for other

peer clients. This differs from the client-server model that uses file servers to store files for sharing.

## 7.5 COMPONENTS OF NAS

A NAS device has 2 main components:

- **NAS head**- The NAS head contains below components:
  - CPU and memory
  - One or more network interface cards (NICs), which provide connectivity to the client network.
  - An optimized operating system for managing the NAS functionality.
  - NFS, CIFS, and other protocols for file sharing.
  - Industry-standard storage protocols and ports to connect and manage physical disk resources.
- **Storage**-Storage contains files to be shared. The storage could be external to the NAS device and shared with other hosts.

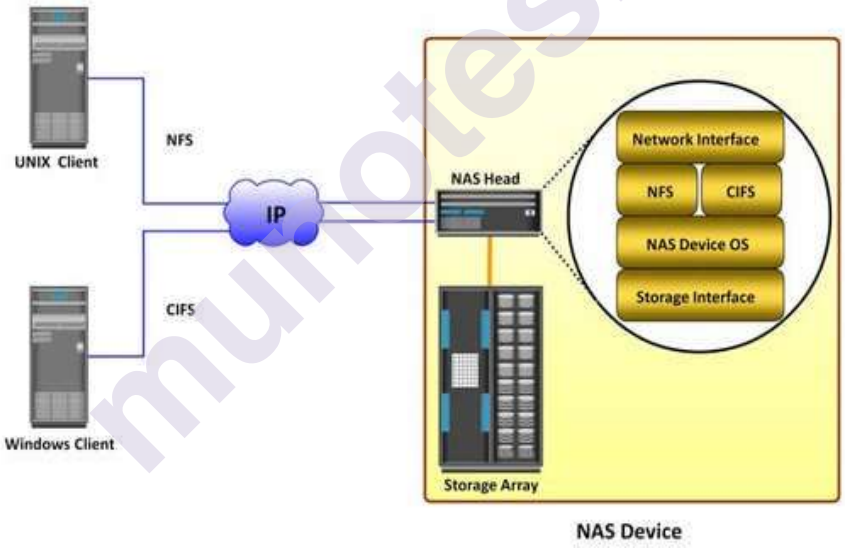



***Figure 7-3 NAS Components***

## 7.6 NAS I/O OPERATION

When a client request for file, 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 fileI/O into block-level I/O to retrieve data. The process of handling I/Os in a NASenvironment is as follows:

1. The 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.

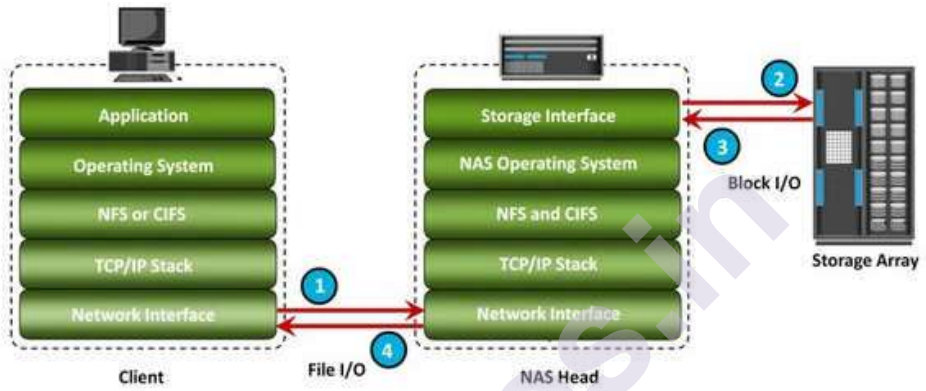



***Figure 7-4 NAS I/O Operation***

## 7.7 NAS IMPLEMENTATIONS

There are three ways of NAS implementations:

1. Unified NAS
2. Gateway NAS
3. scale-out NAS

### 7.7.1 Unified NAS

The *unified* NAS is a combination of NAS and SAN approaches. The *unified* NAS combines NAS-based and SAN-based data access with common storage platform and provides a common interface for managing both the environments. Unified NAS performs file serving and storing of file data, along with providing access to block-level data. It supports both CIFS and NFS protocols for file access and iSCSI and FC protocols for block level access. Due to consolidation of NAS-based and SAN-based access on a single storage platform, unified NASreduces an organization's infrastructure and management costs.

A unified NAS contains one or more NAS heads and storage in a single system. NAS heads are connected to the storage controllers (SCs), which provide access to the storage. These storage controllers also provide connectivity to iSCSI and FC hosts. The storage may consist of different drive types, such as SAS, ATA, FC, and flash drives, to meet different workload requirements.

**Unified NAS Connectivity**

Each NAS head in a unified NAS has front-end Ethernet ports, which connect to the IP network. The front-end ports provide connectivity to the clients and service the file I/O requests. Each NAS head has back-end ports, to provide connectivity to the storage controllers. iSCSI and FC ports on a storage controller enable hosts to access the storage directly or through a storage network at the block level. Figure 7-5 illustrates an example of unified NAS connectivity.

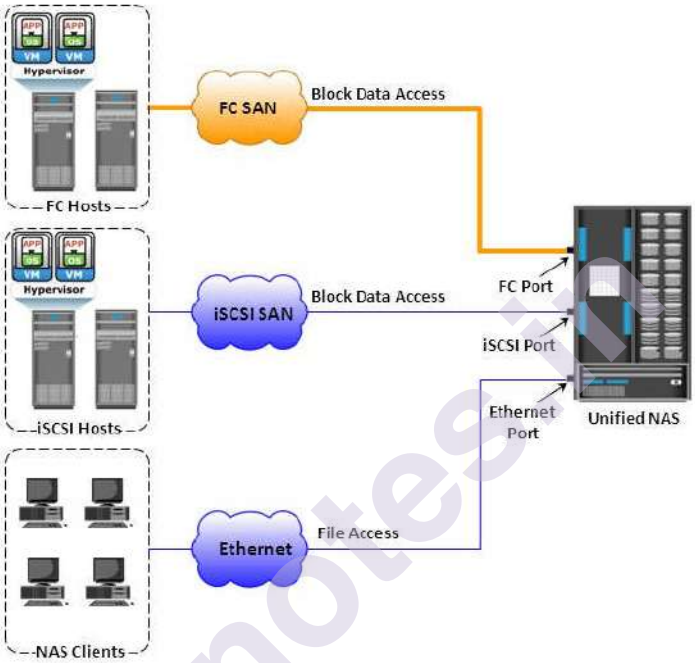



***Figure 7-5 Unified NAS Connectivity***

## 7.7.2 Gateway NAS

In a *gateway* implementation, the NAS device has external storage, and there is separate managing interface for the NAS device and storage. A gateway NAS device consists of one or more NAS heads and uses external and independently managed storage. Similar to unified NAS, the storage is shared with other applications that use block-level I/O. Management functions in this type of solution are more complex than those in a unified NASenvironment because there are separate administrative tasks for the NAShead and the storage. A gateway solution can use the FC infrastructure, suchas switches and directors for accessing SAN-attached storage arrays or direct attached storage arrays.

The gateway NAS is more scalable compared to unified NAS because NASheads and storage arrays can be independently scaled up when required. For example, NAS heads can be added to scale up the NAS device performance. When the storage limit is reached, it can scale up, adding capacity on the SAN, independent of NAS heads. Similar to a

unified NAS, a gateway NAS also enables high utilization of storage capacity by sharing it with the SAN environment.

**Gateway NAS Connectivity**

In a gateway solution, the front-end connectivity is similar to that in a unified storage solution. Communication between the NAS gateway and the storage system is achieved through a traditional FC SAN. To deploy gateway NAS solution, some factors must be considered, like multiple paths for data, redundant fabrics, and load distribution. Figure 7-6 illustrates an example of gateway NAS connectivity.

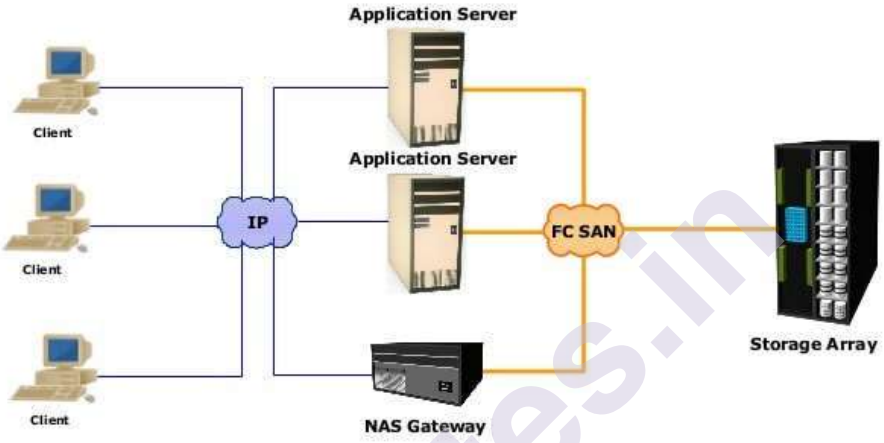



***Figure 7-6 Gateway NAS Connectivity***

Implementation of both unified and gateway solutions requires analysis of the SAN environment. This analysis is required to determine the feasibility of combining the NAS workload with the SAN workload. Analyze the SAN to determine whether the workload is primarily read or write, and if it is random or sequential. Also determine the predominant I/O size in use. Typically, NAS workloads are random with small I/O sizes. Introducing sequential workload with random workloads can be disruptive to the sequential workload. Therefore, it is recommended to separate the NAS and SANdisks. Also, determine whether the NAS workload performs adequately with the configured cache in the storage system.

## 7.7.3 Scale-Out NAS

The *scale-out* NAS implementation combines multiple nodes to form a cluster NAS system. A node may consist of either the NAS head or storage or both. Scale-out NAS enables grouping multiple nodes together to constructa clustered NAS system. A scale-out NAS provides the capability to scale its resources by simply adding nodes to a clustered NAS architecture. The cluster works as a single NAS device and is managed centrally. Nodescan be added to the cluster, when more performance or more capacity is needed, without causing any downtime. Scale-out NAS provides the flexibility to use many nodes of moderate performance and availability characteristics to produce a total system that

has better aggregate performance and availability. It also provides ease of use, low cost, and theoretically unlimited scalability.

Scale-out NAS creates a single file system that runs on all nodes in the cluster. All information is shared among nodes, so the entire file system is accessible by clients connecting to any node in the cluster. As data is sent from clients to the cluster, the data is divided and allocated to different nodes in parallel. When a client sends a request to read a file, the scale-out NAS retrieves the appropriate blocks from multiple nodes, recombines the blocks into a file, and presents the file to the client. As nodes are added, the file system grows dynamically and data is evenly distributed to every node. Each node added to the cluster increases the aggregate storage, memory, CPU, and network capacity. Hence, cluster performance also increases.

Scale-out NAS supports heavy concurrent ingest workloads—blending capacity, performance, and simplicity to provision storage. By scaling within a single cluster, memory and network resources are optimized across operations. In a scale-out architecture, new hardware can be easily added and configured to support changing business requirements.

**Scale-Out NAS Connectivity**

Scale-out NAS clusters use separate internal and external networks for back-endand front-end connectivity, respectively. An internal network provides connections for intracluster communication, and an external network connection enables clients to access and share file data. Each node in the cluster connects to the internal network.

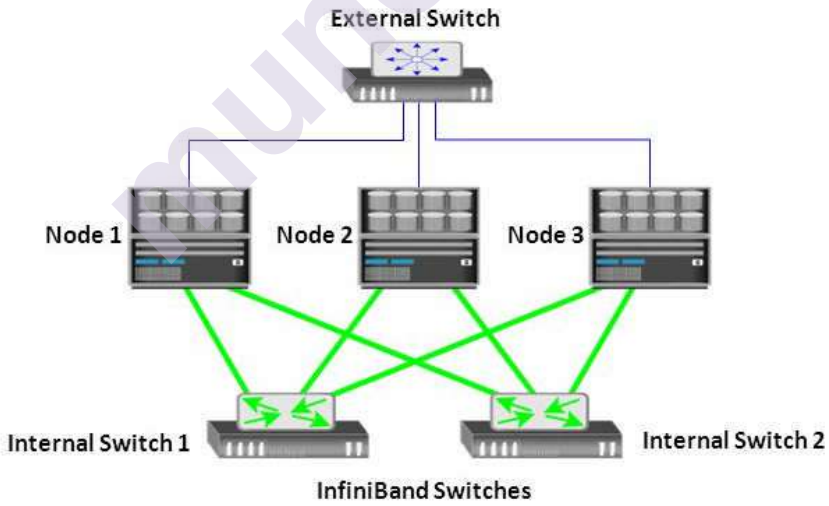



***Figure 7-7 Scale-Out NAS Connectivity***

The internal network offers high throughput and low latency and uses high-speed networking technology, such as InfiniBand or Gigabit Ethernet. To enable clients to access a node, the node must be connected to the external Ethernet network. Redundant internal or external networks may be used for high availability. Figure 7-7 illustrates an example of scale-out NAS connectivity.

## 7.8 NAS FILE-SHARING PROTOCOLS

Network-attached storage (NAS) is file-level computer data storage server connected to a computer network providing data access to a heterogeneous group of clients. NAS not only operates as a file server, but is specialized for this task either by its hardware, software, or configuration of those elements.

Two common NAS file sharing protocols are:

1. NFS – Network File System Protocol Traditional UNIX environment file sharing protocol
2. CIFS – Common Internet File System Protocol Traditional Microsoft environment file sharing protocol, based upon the Server Message Block Protocol.

### 7.8.1 NFS

NFS is a client/server application that enables a computer user view and optionally store and update files on a remote computer as though they were on the user's own computer. It uses Remote Procedure Calls (RPC) to communicate between computers. Following operations can be done:

- Searching files and directories
- Opening, reading, writing to, and closing a file
- Changing file attributes
- Modifying file links and directories

The user's system requires an NFS client to connect to the NFS server. Since the NFS server and client use TCP/IP to transfer files, TCP/IP must be installed on both systems. Currently, three versions of NFS are in use:

- **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.
- **NFS version 3 (NFSv3):** The most commonly used version, which usesUDP or TCP, and is based on the stateless protocol design. It includes extra features like a 64-bit file size, asynchronous writes, and additional file attributes.
- **NFS version 4 (NFSv4):** Uses TCP and is based on a tasteful protocol design.It offers enhanced security.

### 7.8.2 CIFS

CIFS is client/server application protocol, which enables client's programs make requests for files and services on remote computers on the Internet. CIFS is a public (or open) variation on Microsoft's Server Message Block (SMB) protocol. Like SMB, CIFS runs at a higher level

than, and uses the Internet's TCP/IP protocol. CIFS is viewed as a complement to the existing Internet application protocols such as the File Transfer Protocol (FTP) and the Hyper Text Transfer Protocol (HTTP). The CIFS protocol allows the client to:

i) Get access to files that are local to the server and read and write to them
ii) Share files with other clients using special locks
iii) Restore connections automatically in case of network failure
iv) Use Unicode file names

In general, CIFS gives the client user better control of files than FTP. It provides a potentially more direct interface to server programs than currently available through a Web browser and the HTTP protocol.

CIFS runs over TCP/IP and uses DNS (Domain Naming Service) for name resolution. These file system protocols allow users to share file data across different operating environments as well as provide a means for users to transparently migrate from one operating system to another.

## 7.9 FACTORS AFFECTING NAS PERFORMANCE

NAS uses IP network, bandwidth, and IP-related latency problems to affect NAS performance. Network congestion is one of the most critical latency sources in the NAS environment. Additional factors affecting NAS performance at various levels are:

- **The number of hops -** The number of hops used in a NAS system can affect the speed and performance. As you increase the number of hops, latency also increases as there is a requirement for IP processing in every hop. This will eventually cause a delay at the router and affect the overall performance.
- **File/directory lookup and metadata requests -** The NAS device files are accessed by NAS clients, and the process that leads to the correct file or directory in the system can affect NAS performance and cause delays. This delay can happen because of various reasons, including deep directory structures or bad file system layout. If the disk system is over-utilized, you will witness a considerable degradation in the performance. If you look to get past these issues, you must consider shifting to a flattened directory structure.
- **Over utilized NAS devices -** When multiple clients are working on a NAS platform and try to access multiple files simultaneously, the utilization levels of the NAS device shoot up and slow down the entire framework. The utilization statistics will aid you in understanding the levels in which you are running. This issue is also a result of a flawed file system.
- **Authentication with a directory service such as LDAP -** Bandwidth is a very pertinent factor to be considered when it comes to the authentication service. The authentication requests cannot be

accommodated if there is a lack of adequate bandwidth and other vital resources. The overall process can lead to an increase in latency, especially when the authentication takes place.

- **Active Directory, or NIS -** It is essential to have at least a single machine to take up the role of a NIS server in any network. The system could face bandwidth issues, and there will be an increase in latency in the case of Active directories or NIS.
- **Retransmission -** Sometimes, in NAS systems, sets of data that do not reach the respective destination get retransmitted, and this is one of the main reasons for the increase in latency. Retransmission can be attributed to several reasons, such as buffer overflows, link errors, and flow control mechanisms.
- **Overutilized clients -** Overutilization of clients in CIFS or NFS systems is again a parameter that adds to latency. In these cases, the client might require more time for processing the transmissions from the server.
- **Over utilized routers and switches -** Considering overutilized clients, the devices involved also get overutilized. Such a network takes more time to respond than a device that is functioning optimally.

## 7.10 FILE-LEVEL VIRTUALIZATION

` **Virtualization on block level** means that storage capacity is made available to the operating system or the applications in the form of virtual disks. **Virtualization on file level** means that the virtualization entity provides virtual storage to the operating systems or applications in the form of files and directories. Implementation of file-level virtualization is common in NAS or file-server environments. It provides non-disruptive file mobility to optimize storage utilization.

In **block-level storage**, a storage device such as a hard disk drive (HDD) is identified as something called a storage *volume*. A storage volume can be treated as an individual drive, a “block”. This gives a server's operating system the ability to have access to the raw storage sections. The storage blocks can be modified by an administrator, adding more capacity, when necessary, which makes block storage fast, flexible, and reliable. 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 simplifies file mobility. It provides user or application independence from the location where the files are stored. **File-level storage** is a type of storage that has a file system installed directly onto it where the storage volumes appear as a hierarchy of files to the server, rather than blocks. This is different from block type storage, which doesn't have a default file system and needs to have an administrator create one in order for non-administrator users to navigate and find data.

Figure 7-8 illustrates a file-serving environment before and after the implementation of file-level virtualization.

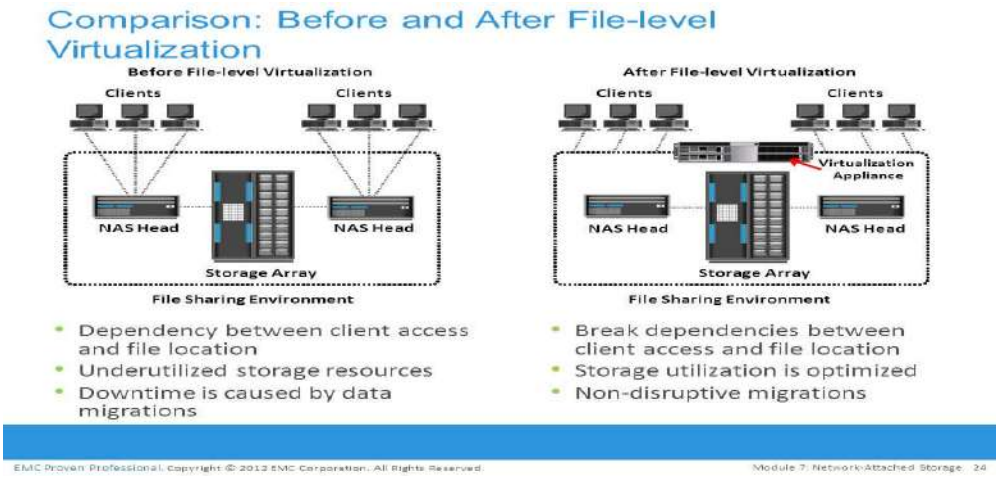



***Figure 7-8 File-serving environment before and after file-level virtualization***

One benefit of using file storage is that it is easier to use. Most people are familiar with file system navigation as opposed to storage volumes found in block-level storage, where more knowledge about **partitioning** is required to create volumes.

## 7.11OBJECT-BASED AND UNIFIED STORAGE

We know that more than90 percent of data generated is unstructured. This growth of unstructured data has created challenges to IT administrators and storage managers. With this growth, traditional NAShas become inefficient. Data growth adds high overhead to the network-attached storage (NAS). NASalso manages large amounts of metadata generated by hosts, storage systems, and individual applications. This adds to the complexity and latency in searching and retrieving files. These challenges demand a smarter approach to manage unstructured data based on its content rather than metadata about its name, location, and so on. *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. Object storage systems allow retention of massive amounts of unstructured data. Unified storage has emerged as a great solution that consolidates block, file, and object-based access within one unified platform. It supports multiple protocols for data access and can be managed using a single management interface.

## 7.12 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 datain 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 7-9).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). Each object stored in the system is identified by a unique ID called the *object ID*.

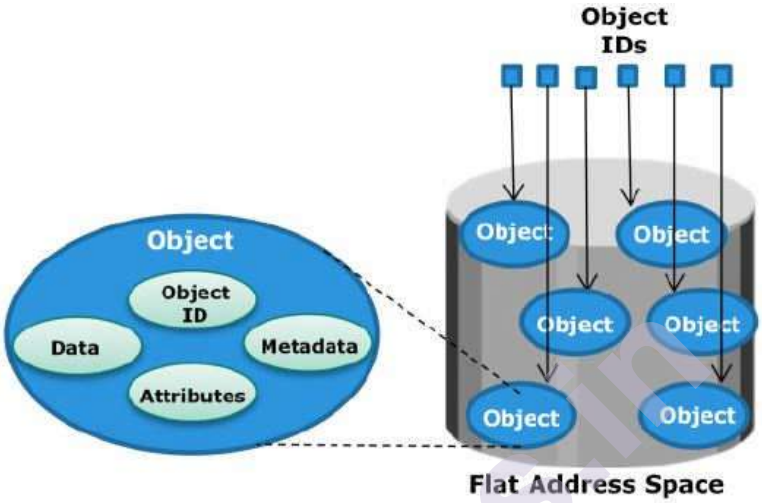



***Figure 7-9 Object-Based Storage Device***

### 7.12.1 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.

The file system has two components:

- 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. After the object is stored, the OSD sends an acknowledgment to the application server. The OSD storage component manages all the required

low-level storage and space management functions. It also manages security and access control functions for the objects.

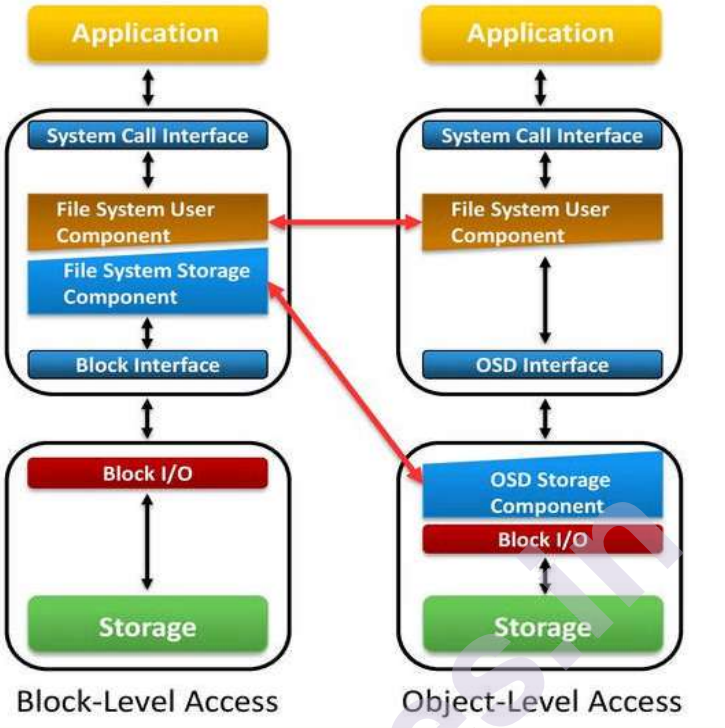



***Figure 7-10Traditional Vs. Object-Based Storage***

## 7.12.2 Components of OSD

The OSD system is typically composed of three key components: nodes, private network, and storage. Figure 7-11 illustrates the components of OSD.

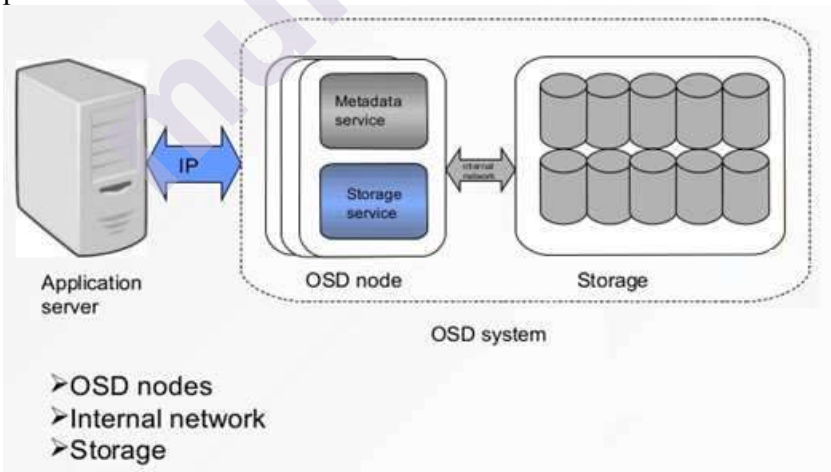



***Figure 7-11 Key components of OSD***

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. It has two key services:

- The metadata service is responsible for generating the object ID from the contents (and can also include other attributes of data) of a file, & 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 which provides node-to-node connectivity and node-to-storage connectivity. The application server accesses the node to store and retrieve data over an external network.

OSD typically uses low-cost and high-density disk drives to store the objects. As more capacity is required, more disk drives can be added to the system.

### 7.12.3 Object Storage and Retrieval in OSD

The process of storing objects in OSD is illustrated in Figure 7-12.

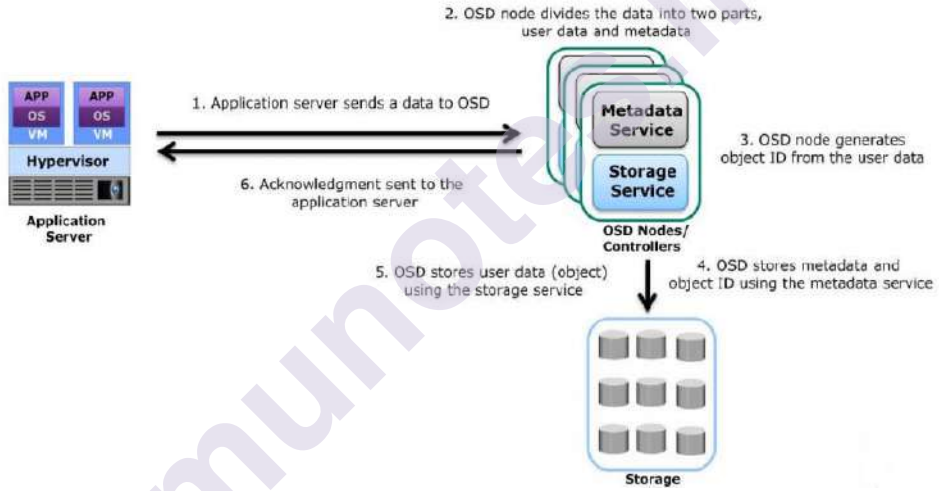



***Figure 7-12 Storing objects on OSD***

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.

After an object is stored successfully, it is available for retrieval. A user accesses the data stored on OSD by the same filename. The application server retrieves the stored content using the object ID. This process is transparent to the user.

The process of retrieving objects in OSD is illustrated in Figures 7-13.

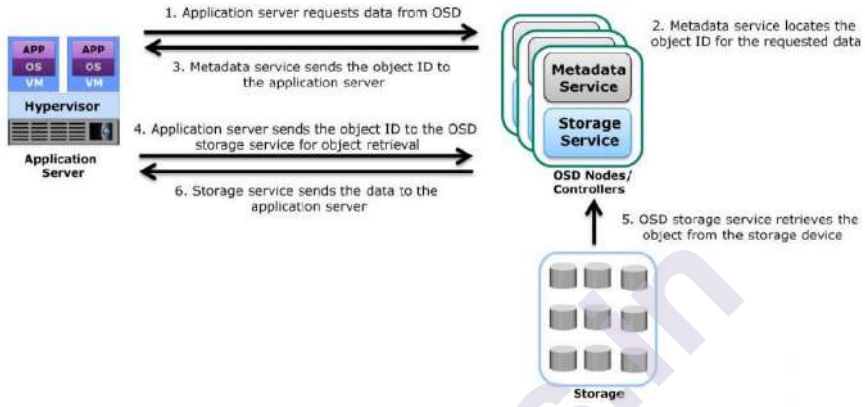



***Figure 7-13 Object retrieval from an OSD***

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.

## 7.12.4 Benefits of Object-Based Storage

Object-based storage devices for unstructured data provide numerous benefits over traditional storage solutions. The key benefits of object-based storage are as follows:

- **Security and reliability**. OSD make use of 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 attribute allows objects to be shared across heterogeneous platforms locally or remotely.

- **Scalability:** Due to the use of FL at address space, object-based storage can handle large amounts of data without impacting performance. Keep adding data, forever. There's no limit.
- **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.
- **Reduction in cost.** Due to the scale-out nature of object storage, it's less costly to store all your data.
- **Faster data retrieval.** Due to the categorization structure of object storage, and the lack of folder hierarchy, you can retrieve your data much faster.

### 7.12.5 Common Use Cases for Object-Based Storage

There are multiple use cases for object storage. For example, it can assist you in the following ways:

- **Deliver rich media.** Define workflows by leveraging industry-leading solutions for managing unstructured data. Reduce your costs for globally distributed rich media.
- **Manage distributed content.** Optimize the value of your data throughout its lifecycle and deliver competitive storage services.
- **Embrace the Internet of Things (IoT).** Manage machine-to-machine data efficiently, support artificial intelligence and analytics, and compress the cost and time of the design process.

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 and a web interface to access storage resources &provides inherent security, scalability, and automated data management. OSD supports web service access via *representational state transfer* (REST) and *simple object access protocol* (SOAP).

## 7.13 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 an optimized and centrally managed storage solution. Data access in CAS differs from other OSD devices. In CAS, the application server accesses the CAS device only via the CAS API running on the application server. However, the way CAS stores data is similar to the other OSD systems.

CAS provides following key features 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.
- **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
- **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 & provides both local and remote protection to it. In the local protection option, data objects are either mirrored or parity protected. In mirror protection, two copies of the data object are stored on two different nodes in the same cluster.

- **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 detects and repairs corrupted objects and alerts the administrator about the potential problem.
- **Audit trails:** CAS keeps track of management activities and any access or disposition of data. Audit trails are mandated by compliance requirements.

## 7.14 CAS USE CASES

Organizations have deployed CAS solutions to solve several business challenges. Two solutions are given below.

### 7.14.1 Healthcare Solution: Storing Patient Studies

Large healthcare centers examine hundreds of patients every day and generate large volumes of medical records. Each record might be composed of one or more images that range in size from approximately 15 MB for a standard digital X-ray to more than 1 GB for oncology studies. The patient records a restored online for a specific period of time for immediate use by the attending physicians. Even if a patient's record is no longer needed, compliance requirements might stipulate that the records be kept in the original format for several years.

Medical image solution providers offer hospitals the capability to view medical records, such as X-ray images, with acceptable response times and resolution to enable rapid assessments of patients. Figure 7-14 illustrates the use of CAS in this scenario. Patients' records are retained on the primary storage for 60 days after which they are moved to the CAS system. CAS facilitates long-term storage and at the same time, provides immediate access to data, when needed.

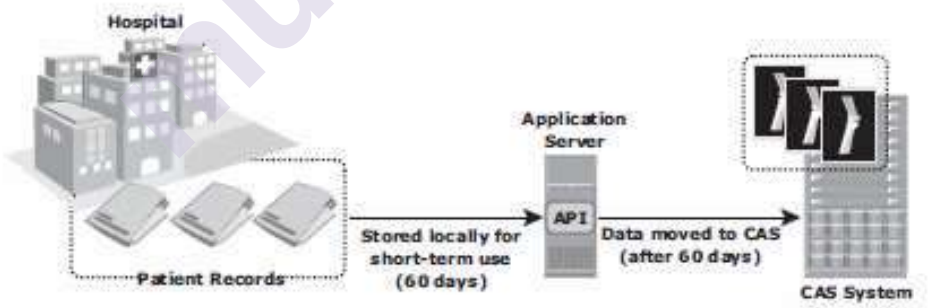



***Figure 7-14 Storing patient studies on a CAS system***

#### 7.14.2 Finance Solution: Storing Financial Records

In a typical banking scenario, images of checks, each approximately 25 KB in size,are created and sent to archive services over an IP network. A check imaging service provider might process approximately 90 million check images per month. Typically, check images are actively processed in transactional systems for about 5 days.

For the next 60 days, check images may be requested by banks or individual consumers for verification purposes; beyond 60 days, access requirements drop drastically. Figure 7-15 illustrates the use of CAS in this scenario. The check images are moved from the primary storage to the CAS system after 60 days, and can be held there for long term based on retention policy. Check imagingis one example of a financial service application that is best serviced with CAS. Customer transactions initiated by e-mail, contracts, and security transaction records might need to be kept online for 30 years; CAS is the preferred storage solution in such cases.

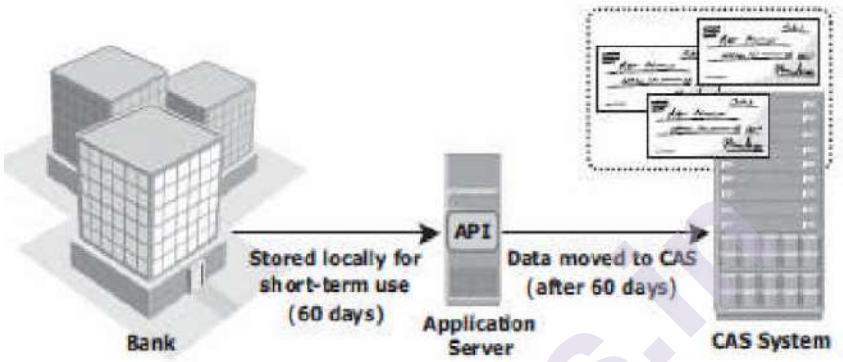



***Figure 7-15 Storing financial records on a CAS system***

## 7.15 UNIFIED STORAGE

A unified storage architecture, which combines block-level and file-level access in a single storage system. It supports multiple protocols, such as CIFS, NFS, iSCSI, FC, etc.

### 7.15.1 Components of Unified Storage

A unified storage system consists of the following key components: storage controller, NAS head, OSD node, and storage. Figure 7-16 illustrates the block diagram of a unified storage platform.

The unified controller provides the functionalities of block storage, file storage, and object storage. It contains iSCSI, FC, FCoE, and IP front-end ports for direct block access to application servers and file access to NAS clients.

For **block-level access**, the controller configures LUNs and presents them to application servers and the LUNs presented to the application server appear as local physical disks. A file system is configured on these LUNs at the server and is made available to applications for storing data. For **NAS clients**, the controller configures LUNs and creates a file system on these LUNs and creates a NFS, CIFS, or mixed share, and exports the share to the clients.

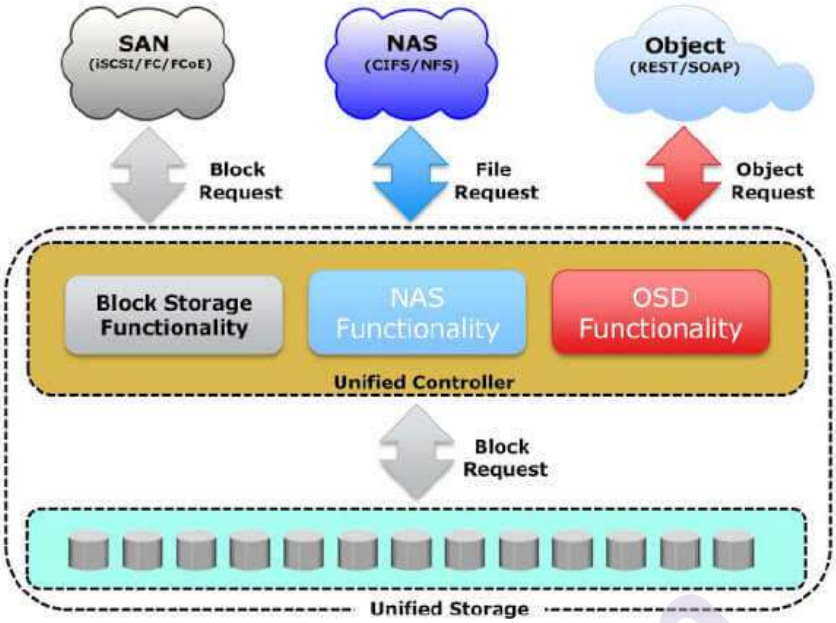



***Figure 7-16 Unified Storage Platform***

The *OSD node* accesses the storage through the storage controller using a FCor FCoE connection. The LUNs assigned to the OSD node appear as physical disks configured by the OSD nodes, enabling them to store the data from the web application servers.

## *Data Access from Unified Storage*

In a unified storage system, block, file, and object requests to the storage travel through different I/O paths. Figure 7-16 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 request is converted into a block request, and forwarded to 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 request received by OSD is converted into a block request, and is send to the disk through the storage controller. The controller in turn processes the block request and responds back to the OSD node, which in turn provides the requested object to the web application server.

## 7.16 SUMMARY

File sharing means providing common file access to more than one user.Client-Server technology is used for file sharing over a network.NAS devices are rapidly becoming popular with enterprise and small businesses in many industries as an effective, scalable, low-cost storage solution.A NAS device provides file-serving functions such as storing, retrieving, and accessing files for applications and clients.A file system is a process that manages how and where data on a storage disk is stored, accessed and managed. Many file systems maintain a file access table to simplify the process of searching and accessing files.

When a client request for file, NAS provides file-level data access to its clients. File I/O is a high-level request that specifies the file to be accessed.The *unified* NAS is a combination of NAS and SAN approaches. In a *gateway* implementation, the NAS device has external storage, and there is separate managing interface for the NAS device and storage.The *scale-out* NAS implementation combines multiple nodes to form a cluster NAS system.*NFS* is a client-server protocol for file sharing that is commonly used on UNIX systems.CIFS is a client-server application protocol that enables client programs to make requests for files and services on remote computers over TCP/IP.

NAS uses IP network; therefore, bandwidth and latency issues associated with IP affect NAS performance.Object-based storage provides a scalable, self-managed, protected, and shared storage option.CAS is an object-based storage device designed for secure online storage and retrieval of fixed content.Unified storage consolidates block, file, and object access into one storage solution. It supports multiple protocols, such as CIFS, NFS, iSCSI, FC, FCoE, REST, and SOAP.

## 7.17 REVIEW QUESTIONS

1. What are advantages of NAS (Network-Attached Storage)?
2. What are components of NAS (Network-Attached Storage)? Explain with diagram.
3. Explain NAS input-output operation.
4. What are different types of NAS implementations? Explain any in detail?
5. Explain Unified NAS implementation.
6. Explain gateway NAS implementation.
7. Explain scale-out NAS implementation.
8. Explain NFS protocol for file sharing.
9. Explain CIFS protocol for file sharing.
10. What are different factors that affect NAS performance at different levels?
11. Explain File-level virtualization.
12. Compare File level virtualization with block level virtualization.

13. What is Object-Based Storage Device? Explain Architecture Object-Based Storage.
14. What is Object-Based Storage Device? Explain Components of Object-Based Storage.
15. Show the process of storing and retrieving objects in OSD with diagram.
16. What are key benefits of object-based storage?
17. What is CAS (Content-Addressed Storage)? Explain The key features of CAS.
18. What is unified storage? What are different components of unified storage? How Data is accessed from Unified Storage?

## 7.18 REFERENCES


- Information Storage and Management: Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments by Somasundaram Gnanasundaram and Alok Shrivastava, 2nd Edition Publisher: John Wiley & Sons.
- https://www.ippartners.com.au/it-news/defining-storage-area-networks-sans-network-attached-storage-nas-and-unified-storage/
- https://www.mycloudwiki.com
- https://www.netapp.com
- https://whatis.techtarget.com
- https://www.ques10.com/p/2557/explain-nas-file-sharing-protocols/
- https://www.promax.com/blog/factors-that-impact-remote-nas-performance




❖❖❖❖

# 8

# INTRODUCTION TO BUSINESS CONTINUITY

**Unit Structure**





## 8.0 OBJECTIVES

This chapter describes the factors that affect information availability and the consequences of information unavailability. It also explains the key parameters that govern any Business Continuity (BC) strategy and the roadmap to develop an effective BC plan. We will study BC planning life cycle also.

## 8.1INTRODUCTION

In modern times, continuous access to information is a must for the smooth functioning of business operations. Unavailability of information cost is greater than ever, and outages in key industries cost millions of dollars per hour. Threats to information availability, such as natural disasters, unplanned occurrences, and planned occurrences, could result in the inaccessibility of information & becomes critical for businesses to define an appropriate strategy that can help them overcome these crises.

Business continuity is an important process to define and implement these strategies.

**Business continuity** is the advance planning and preparation undertaken to ensure that an organization will have the capability to operate its critical business functions during emergency events. Events can include natural disasters, a business crisis, pandemic, workplace violence, or any event that results in a disruption of your business operation. It is important to remember that you should plan and prepare not only for events that will stop functions completely but for those that also have the potential to adversely impact services or functions.

Common technology services designed for business continuity consist of cloud data backups, cloud-based disaster recovery as a service (DRaaS) for infrastructure outages, and managed security plans that protect against increasingly sophisticated cyber attacks.

## 8.2 INFORMATION AVAILABILITY

Information availability (IA) refers to the ability of the infrastructure to function according to business expectations during its specified time of operation. Information availability ensures that people (employees, customers, suppliers, and partners) can access information whenever they need it. Information availability can be defined with the help of reliability, accessibility and timeliness

- **Reliability:** This reflects a component's ability to function without failure, under stated conditions, for a specified amount of time.
- **Accessibility:** This is the state within which the required information is accessible at the right place, to the right user.
- **Timeliness:** Defines the exact moment or the time window (a particular time of the day, week, month, and/or year as specified) 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.

### 8.2.1 Causes of Information Unavailability

Various planned and unplanned incidents result in data unavailability. 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.

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.

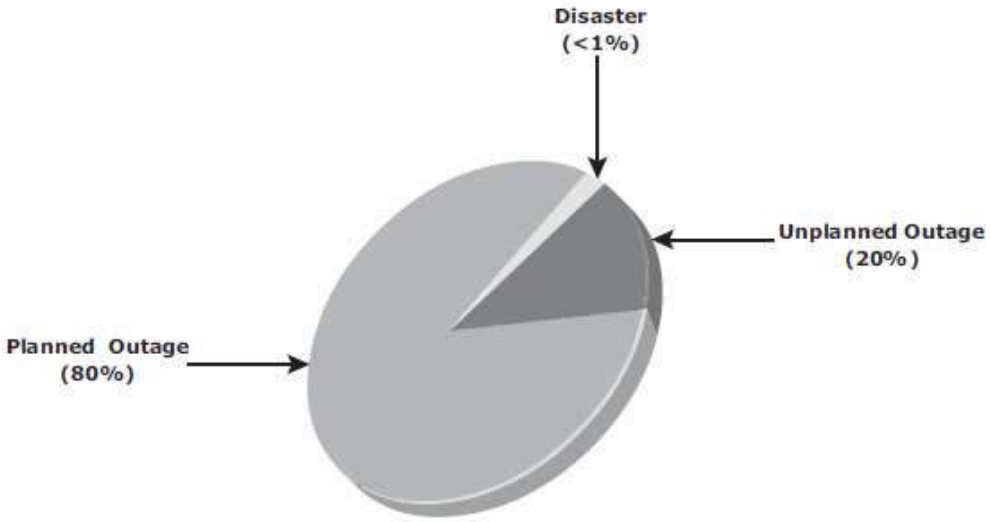



***Figure 8.1 disruptors of data availability***

### 8.2.2 Consequences of Downtime

**Downtime** is the amount of time during the agreed service time that the service is not available. Information unavailability or downtime outcome in loss of productivity, 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, and investment loss. Poor financial performance affects revenue recognition, cash flow, discounts, payment guarantees, 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 = average productivity loss per hour + average 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)

The average downtime cost per hour may also include estimates of projected revenue loss due to other consequences, such as damaged reputations, and the additional cost of repairing the system.

### 8.2.3 Measuring Information Availability

IA relies on the availability of both physical and virtual components of a data center & failure of these might disrupt IA. A failure is the termination of a component's capability to perform a required function. The component's capability can be restored by performing an external corrective action, such as a manual reboot, repair, or replacement of the failed component(s). Proactive risk analysis, performed as part of the BC planning process, considers the component failure rate and average repair time, which are measured by mean time between failure (MTBF) and mean time to repair (MTTR):

- **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.
- **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 the required spares and personnel are available. MTTR includes the total time required to do the following activities: Detect the fault, mobilize the maintenance team, diagnose the fault, obtain the spare parts, repair, test, and restore the data.

  Figure 8-2 illustrates the various information availability metrics that represent system uptime and downtime.

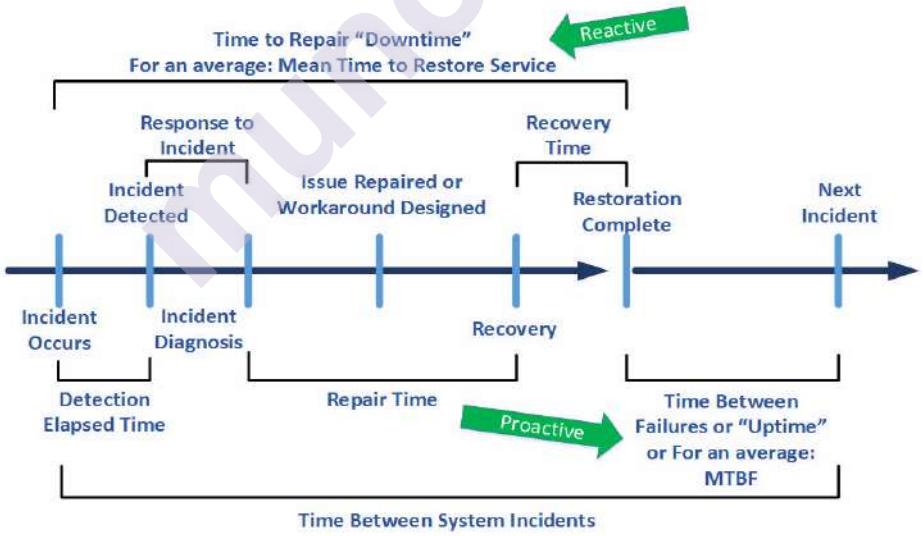



***Figure 8-2 Information Availability metrics***

IA is the time period during which 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 = system uptime/(system uptime + system downtime)*

Where *system uptime* is the period of time during which the system is in an accessible state; when it is not accessible, it is termed as *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.

## 8.3 BC TERMINOLOGY

- **Disaster recovery:** Refers to the process of restoring operations, including information technology resources, within a predetermined timeframe. The disaster recovery plan is a critical component of the business continuity plan.
- **Disaster restart:** This is the process of restarting business operations with mirrored consistent copies of data and applications.
- **Recovery-Point Objective (RPO):** This is the point in time to which systems and data must be recovered after an outage. A large RPO signifies high tolerance to information loss in a business. Based on the RPO, organizations plan for the frequency with which a backup or replica must be made. For example, if the RPO is 6 hours, backups or replicas must be made at least once in 6 hours. Figure 8-3 (a) shows various RPOs and their corresponding ideal recovery strategies.
  - **RPO of 24 hours:** Backups are created at an offsite tape library every midnight. The corresponding recovery strategy is to restore data from the set of last backup tapes.
  - **RPO of 1 hour:** Shipping database logs to the remote site every hour.
  - **RPO in the order of minutes:** Mirroring data asynchronously to a remote site
  - **Near zero RPO:** Mirroring data synchronously to a remote site

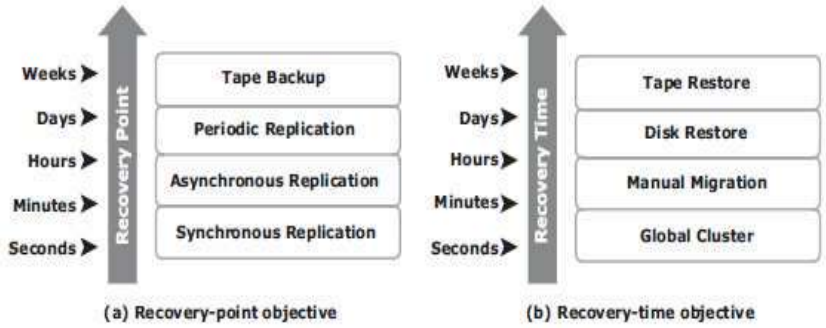



***Figure 8-3 Strategies to meet RPO and RTO targets***

- **Recovery-Time Objective (RTO):** The time within which systems and applications must be recovered after an outage. Businesses can

optimize disaster recovery plans after defining the RTO for a given system. For example, if the RTO is 2 hours, it requires disk-based backup because it enables a faster restore than a tape backup. However, for an RTO of 1 week, tape backup will likely meet the requirements. Some examples of RTOs and the recovery strategies to ensure data availability are listed here (refer to Figure 8-3 [b]):

- **RTO of 72 hours:** Restore from tapes available at a cold site.
- **RTO of 12 hours:** Restore from tapes available at a hot site.
- **RTO of few hours:** Use of data vault at a hot site
- **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 is kept as standby to take over completely, as required.

## 8.4 BC PLANNING LIFE CYCLE

BC planning must follow a disciplined approach like any other planning process. From the conceptualization to the realization of the BC plan, a life cycle of activities can be defined for the BC process. The BC planning life cycle includes five stages (see Figure 8-4):

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

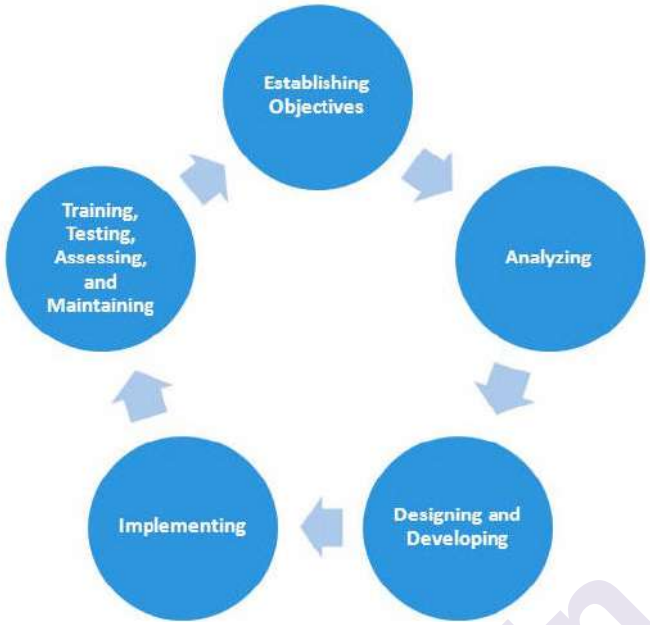



***Figure 8-4 BC Planning Life Cycle***

**1. Establish objectives:**

- Determine BC requirements.
- Estimate the scope and budget to achieve requirements.
- Select a BC team that includes subject matter experts from all areas of the business, whether internal or external.
- Create BC policies.

**2. Analysis:**

- Collect information on data profit, business processes, infrastructure support, dependencies, and frequency of using business infrastructure.
- Conduct a Business Impact Analysis (BIA).
- Identify critical business processes and assign recovery priorities.
- Perform risk analysis for critical functions and create mitigation strategies.
- Perform cost benefit analysis for available solutions based on the mitigation strategy.
- Evaluate options.

**3. Design and develop:**

- 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 solutions.

- Develop emergency response procedures.
- Detail recovery and restart procedures.

**4. Implement:**

- 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. Train, test, assess, and maintain:**

- 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.

## 8.5 FAILURE ANALYSIS

Failure analysis involves analyzing both the physical and virtual infrastructure components to identify systems that are susceptible to a single point of failure and implementing fault-tolerance mechanisms.

### 8.5.1 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 8-5 depicts a system setup in which an application, running on a VM, provides an interface to the client and performs I/O operations. The client is connected to the server through an IP network, and the server is connected to the storage array through an FCconnection.

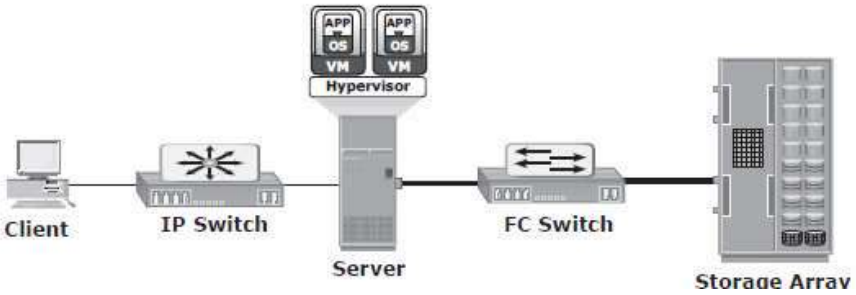



***Figure 8-5 single point of failure***

In a setup in which each component must function as required to ensure data availability, the failure of a single physical or virtual component causes the unavailability of an application & results in disruption of business operations. For example, failure of a hyper visor can affect all the running VMsand the virtual network, which are hosted on it. In the setup shown in Figure9-5, several single points of failure can be identified. A VM, a hypervisor, an HBA/NIC on the server, the physical server, the IP network, the FC switch, the storage array ports, or even the storage array could be a potential single point of failure.

### 8.5.2 Resolving Single Points of Failure

To reduce single points of failure, systems are designed with redundancy,such that the system fails only if all the components in the redundancy groupfail. This ensures that the failure of a single component does not affect data availability. Data centers follow stringent guidelines to implement fault tolerance for uninterrupted information availability. Careful analysis is performed to eliminate every single point of failure. The example shown in Figure 8-6represents all enhancements in the infrastructure to mitigate single points of failure:

- Configuration of redundant HBAs at a server to mitigate single HBA failure.
- Configuration of NIC teaming at a server allows protection against single physical NIC failure. It allows grouping of two or more physical NICs and treating them as a single logical device. With NIC teaming, if one of the underlying physical NICs fails or its cable is unplugged,the traffic is redirected to another physical NIC in the team.
- Configuration of redundant switches to account for a switch failure.
- Configuration of multiple storage array ports to mitigate a port failure
- RAID and hot spare configuration to ensure continuous operation in the event of disk failure
- Implementation of a redundant storage array at a remote site to mitigate local site failure

- Implementing server (or compute) clustering, a fault-tolerance mechanism whereby two or more servers in a cluster access the same set of data volumes. Clustered servers exchange a *heartbeat* to inform each other about their health. If one of the servers or hyper visors fails, the other server or hyper visor can take up the workload.
- Implementing a VM Fault Tolerance mechanism ensures BC in the event of a server failure. This technique creates duplicate copies of each VM on another server so that when a VM failure is detected, the duplicate VM can be used for failover.

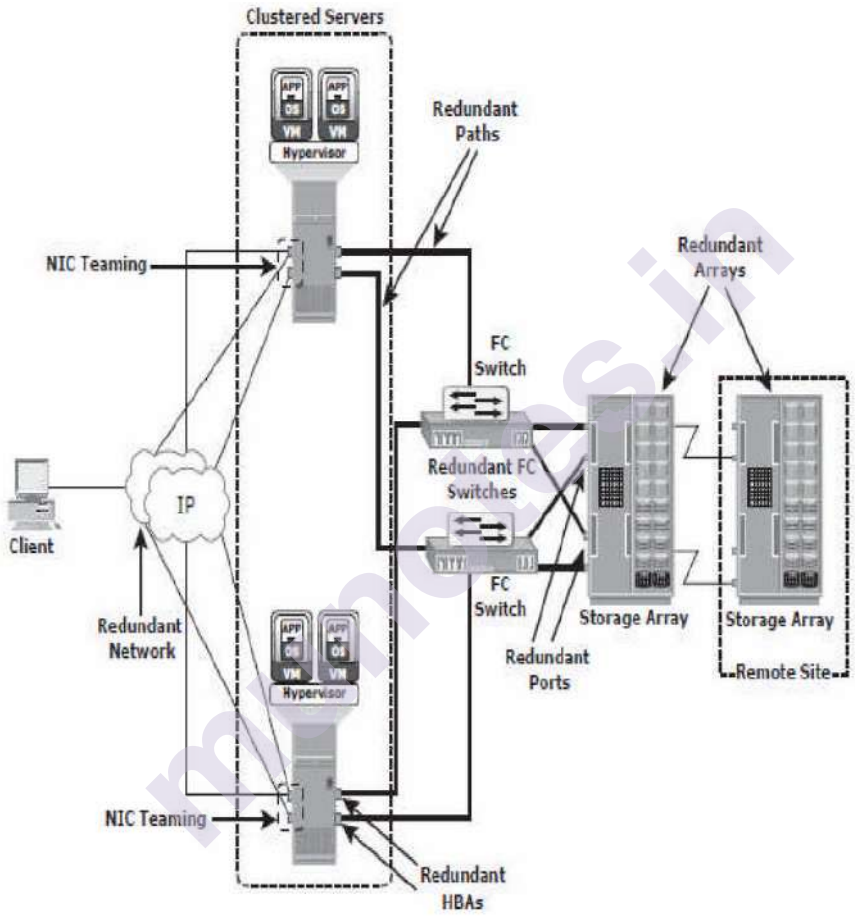



***Figure 8-6 Resolving single points of failure***

### 8.5.3 Multipathing Software

Configuration of multiple paths increases the data availability through path failover. If servers are configured with one I/O path to the data, there will beno access to the data if that path fails. Redundant paths to the data eliminate the possibility of the path becoming a single point of failure. Multiple paths to data also improve I/O performance through load balancing among the pathsand maximize server, storage, and data path utilization.

In practice, merely configuring multiple paths does not serve the purpose. Even with multiple paths, if one path fails, I/O does not reroute unless the system recognizes that it has an alternative path. Multipathing software provides the functionality to recognize and utilize alternative I/O paths to data, also manages the load balancing by distributing I/Os to all available, active paths.

Multipathing software intelligently manages the paths to a device by sendingI/O down the optimal path based on the load balancing and failover policy settingfor the device. It also takes into account path usage and availability before deciding the path through which to send the I/O. If a path to the device fails,it automatically reroutes the I/O to an alternative path.In a virtual environment, multipathing is enabled either by using thehypervisor's built-in capability or by running a third-party software module,added to the hypervisor.

## 8.6 BUSINESS IMPACT ANALYSIS

A **business impact analysis (BIA)** is a process that helps an organization determine and evaluate the potential effects of a problem on its operations.A *business impact analysis* (BIA) identifies which business units, operations, and processes are essential to the survival of the business. It evaluates the financial, operational, and service impacts of a disruption to essential business processes. A BIA includes the following set of tasks:

- Determine the business areas.
- For each business area, identify the key business processes critical to its operation.
- Determine the attributes of the business process in terms of applications, databases, and hardware and software requirements.
- Estimate the costs of failure for each business process.
- Calculate the maximum tolerable outage and define RTO and RPO for each business process.
- Establish the minimum resources required for the operation of business processes.
- Determine recovery strategies and the cost for implementing them.
- Optimize the backup and business recovery strategy based on business priorities.
- Analyze the current state of BC readiness and optimize future BC planning.

## 8.7 BC TECHNOLOGY SOLUTIONS

After analyzing the business impact of an outage, designing the appropriate solutions to recover from a failure is the next important activity. One or more copies of the data are maintained using any of the

following strategies so that data can be recovered or business operations can be restarted using an alternative copy:

- **Backup:** Data backup is a predominant method of ensuring data availability. The frequency of backup is determined based on RPO, RTO, and the frequency of data changes.
- **Local replication:** Data can be replicated to a separate location within the same storage array. The replica is used independently for other business operations. Replicas can also be used for restoring operations if data corruption occurs.
- **Remote replication:** Data in a storage array can be replicated to another storage array located at a remote site. If the storage array is lost due to a disaster, business operations can be started from the remote storage array.

## 8.8 SUMMARY

*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. *Information availability* (IA) refers to the ability of an IT infrastructure to function according to business expectations during its specified time of operation.Information unavailability or downtime results in loss of productivity, lossof revenue, poor financial performance, and damage to reputation.The business impact of downtime is the sum of all losses sustained as a result of a given disruption.A *business impact analysis* (BIA) identifies which business units, operations, and processes are essential to the survival of the business.

## 8.9 REVIEW QUESTIONS

1. What is Information Availability? What are causes of Information Unavailability?
2. What are causes of Information Unavailability? Explain effect of Information Unavailability on business.
3. Explain Life Cycle of BC Planning.
4. What is Single Point of Failure? How resolve Single Point of Failure? Explain with example.
5. What is *business impact analysis* (BIA)? What are different set of BIA tasks.

## 8.10 REFERENCES


- Information Storage and Management: Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments by Somasundaram Gnanasundaram and Alok Shrivastava, 2nd Edition Publisher: John Wiley & Sons.
- http://www.sis.pitt.edu/lersais/research/sahi/resources/labs/drp/Lab_IR_DR_BC_Planning_BC.pdf
- https://en.wikipedia.org/wiki/Business_continuity_planning
- https://www.ques10.com/p/20620/what-is-information-availability-and-information-u/
- https://www.inap.com/blog/business-continuity


❖❖❖❖

# 9

# BACKUP AND ARCHIVE

**Unit Structure**

## 9.0 OBJECTIVES

This chapter includes details about the purposes of the backup, backup and recovery considerations, backup methods, architecture, topologies, and backup targets. Backup optimization using data deduplication and backup in a virtualized environment are also covered in the chapter. Further, this chapter covers types of data archives and archiving solution architecture.

## 9.1 INTRODUCTION

Companies and people are very dependent on data. Whereas a person cannot survive without air, water, and food, businesses cannot survive without data. Hence data is very important in business. A backup is a copy of important data that is stored on an alternative location, so it can be recovered if deleted or it becomes corrupted that can be used to protect organizations against data loss. Organizations are facing problem in the task of backing up because of heavy increasing amount of data. This task becomes more challenging with the growth of information, stagnant IT budgets, and less time for taking backups. Moreover, organizations need a quick restore of backed up data to meet business service-level agreements (SLAs).

To implement a successful backup and recovery solution we have to evaluate the various backup methods with their recovery considerations.

Organizations generate and maintain large volumes of fixed data content. This fixed content is rarely accessed after a period of time and needs to be retained for several years to meet regulatory compliance. Accumulation of this data on the primary storage increases the overall storage cost to the organization. Further, this increases the amount of data to be backed up, which in turn increases the time required to perform the backup.

Data archiving is the process of moving data that is no longer actively used to a separate storage device for long-term retention. Basically, it is stored in low-cost secondary storage. Data archiving reduces the amount of data to be backed up and hence time is also reduced.

## 9.2 BACKUP PURPOSE

There are 3 purposes of data backups:

1. Disaster recovery
2. Operational recovery
3. Archival

### 9.2.1 Disaster Recovery

One purpose of backups is to address disaster recovery needs. Disaster recovery relies upon the replication of data and computer processing in an off-premises location not affected by the disaster. When servers go down because of a natural disaster, equipment failure or cyber-attack, a business needs to recover lost data from a second location where the data is backed up. RTO (Recovery Time Objective) is defined as the time it takes for an organization's IT infrastructure to come back online and be fully functional post a disaster. RPO (Recovery Point Objective) reflects the number of transactions lost from the time of the event up to the full recovery of the IT infrastructure.

When tape-based backup is used as a disaster recovery option, the backup tape media is shipped and stored at an offsite location. Later, these tapes can be recalled for restoration at the disaster recovery site.Ideally, an organization can transfer its computer processing to that remote location as well in order to continue operations. This allows organizations to bring production systems online in a relatively short period of time if a disaster occurs.

### 9.2.2 Operational Recovery

Unlike Disaster recovery, operational recovery deals with more "routine" kinds of failures. The events that require recovery are, in this case, smaller in impact. While Disaster Recovery deals with high-impact events affecting the IT infrastructure, operational recovery deals with errors that appear in a business' daily life: an accidental file deletion, a code error, a file that is wrongly saved etc.For example, it is common for a user to accidentally delete an important e-mail or for a file to become corrupted, which can be restored using backup data.

### 9.2.3 Archival

Backups are also performed to address archival requirements. An archive is frequently used to ease the burden on faster and more frequently accessed data storage systems. Older data that is unlikely to be needed often is put on systems that don't need to have the speed and accessibility of systems that contain data still in use. Archival storage systems are usually less expensive, as well, so a strong motivation is to save money on data storage.

## 9.3 BACKUP CONSIDERATIONS

All enterprises need reliable, efficient data protection and recovery – it is fundamental to business survival. In recent years, mid-sized businesses and distributed enterprises have been largely underserved by data protection hardware due to its high cost at a time when enterprises struggle to keep up with the rising tide of data.

Backup speed is important but recovery speed is where the data protection plan proves its worth. In the event of a system crash or disaster, how fast can critical data be recovered and accessed? Businesses should establish a recovery time objective (RTO) and a recovery point objective (RPO). By establishing an RTO and RPO, a business can maximize business continuity by creating a tiered data protection system that ensures the least possible loss of the most important data

Another consideration is the retention period, which defines the duration for which a business needs to retain the backup copies. Some data is retained for years and some only for a few days. For example, data backed up for archival is retained for a longer period than data backed up for operational recovery.

Organizations must also consider the granularity of backups. The development of a backup strategy must include a decision about the most appropriate time for performing a backup to minimize any disruption to production operations. The location, size, number of files, and data compression should also be considered because they might affect the backup process. Location is an important consideration for the datato be backed up. Consider a data warehouse environment that uses the backup data from many sources. The backup process must address these sources for transactional and content integrity.

The file size and number of files also influence the backup process. Backing up large-size files (for example, ten 1 MB files) takes less time, compared to backing up an equal amount of data composed of small-size files (for example, ten thousand 1 KB files).

Data compression and data deduplication are widely used in the backup environment because these technologies save space on the media. Many backup devices have built-in support for hardware-based data compression. Some data, such as application binaries, do not compress well, whereas text data does compress well.

## 9.4 BACKUP GRANULARITY

Backup granularity describes the level of detail characterizing backup data.Backup granularity depends on business needs and the required RTO/RPO.Based on the granularity, backups can be categorized as full, incremental and cumulative (differential). Most organizations use a combination of these three backup types to meet their backup and recovery requirements. Figure 9-1 shows the different backup granularity levels.

\

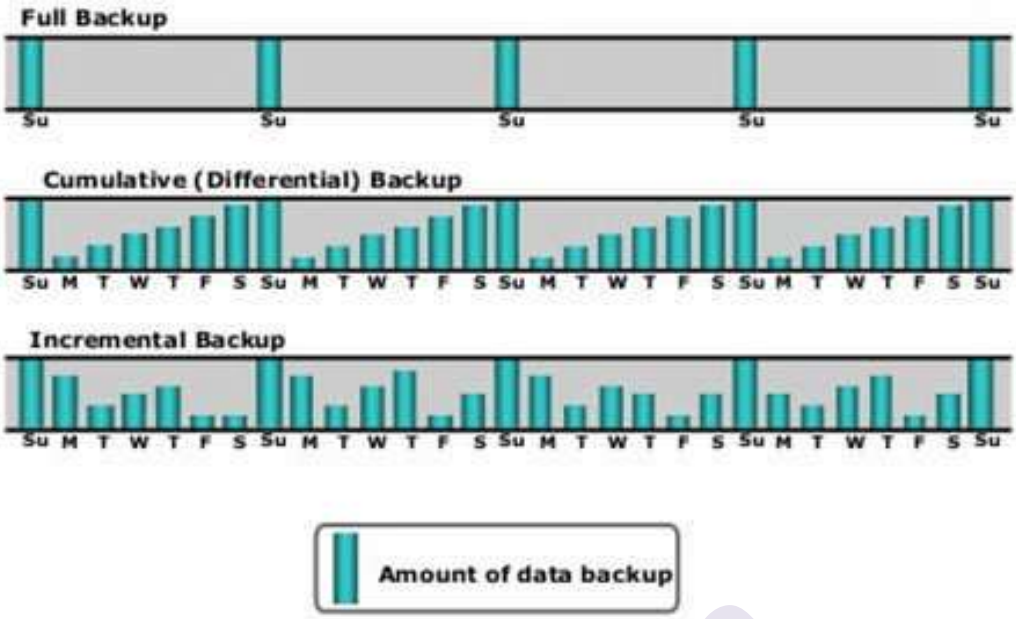



***Figure 9-1 Backup Granularity***

- Full backup: It is a backup of the complete data on the production volumes at a certain point in time. A full backup copy is created by copying the data on the production volumes to a secondary storage device. It provides a faster recovery but requires more storage space and time to back up.
- Incremental backup: It copies the data that has changed since the last full or incremental backup, whichever has occurred more recently. This is much faster because the volume of data backed up is restricted to changed data, but it takes longer to restore.
- Cumulative or differential backup: It copies the data that has changed since the last full backup. This method takes longer than incremental backup but is faster to restore.

Restore operations vary with the granularity of the backup. A full backup provides a single repository from which the data can be easily restored. The process of restoration from an incremental backup requires the last full backup and all the incremental backups available until the point of restoration. A restore from a cumulative backup requires the last full backup and the most recent cumulative backup.

Figure 9-2 shows an example of restoring data from incremental backup.

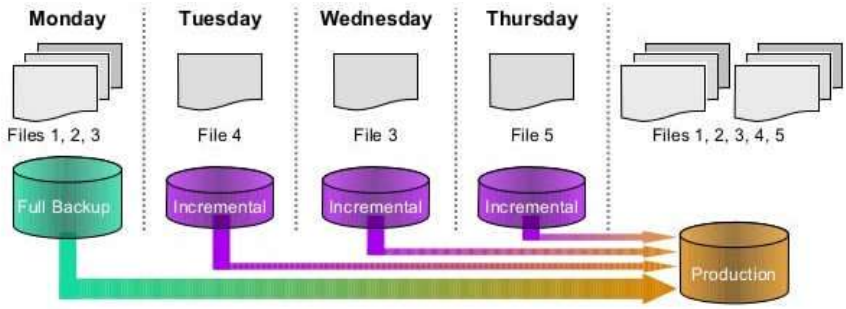



***Figure 9-2 Restoring data from incremental backup***

In this example, a full backup is performed on Monday evening. Each day after that, an incremental backup is performed. On Tuesday, a new file (File4 in the figure) is added, and no other files have changed. Consequently, only File4 is copied during the incremental backup performed on Tuesday evening. On Wednesday, no new files are added, but File3 has been modified. Therefore, only the modified File3 is copied during the incremental backup on Wednesday evening. Similarly, the incremental backup on Thursday copies only File5. OnFriday morning, there is data corruption, which requires data restoration from the backup. The first step toward data restoration is restoring all data from the full backup of Monday evening. The next step is applying the incremental backups of Tuesday, Wednesday, and Thursday. In this manner, data can be successfully recovered to its previous state, as it existed on Thursday evening.

Figure 9-3 shows an example of restoring data from cumulative backup.

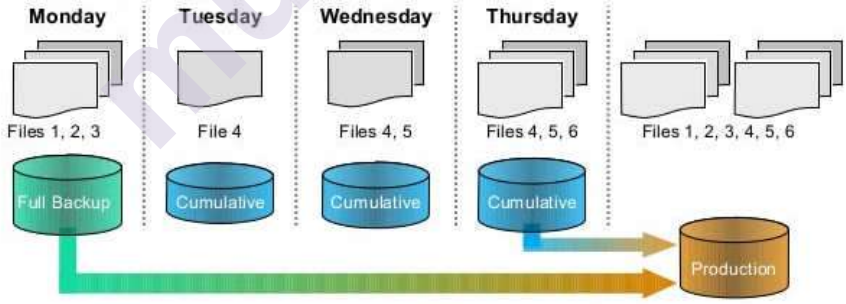



***Figure 9-3Restoring data fromcumulative backup***

In this example, a full backup of the business data is taken on Monday evening.Each day after that, a cumulative backup is taken. On Tuesday, File 4 is added and no other data is modified since the previous full backup of Monday evening. Consequently, the cumulative backup on Tuesday evening copies onlyFile 4. On Wednesday, File 5 is added. The cumulative backup taking place onWednesday evening copies both File 4 and File 5 because these files have been added or modified since the last full backup. Similarly, on Thursday, File 6 isadded. Therefore, the

cumulative backup on Thursday evening copies all three files: File 4, File 5, and File 6. On Friday morning, data corruption occurs that requires data restoration using backup copies. The first step in restoring data is to restore all the data from the full backup of Monday evening. The next step isto apply only the latest cumulative backup, which is taken on Thursday evening.In this way, the production data can be recovered faster because its needs only two copies of data — the last full backup and the latest cumulative backup.

## 9.5 RECOVERY CONSIDERATIONS

The retention period is a key consideration for recovery, derived from an RPO. For example, users of an application might request to restore the application data from its backup copy, which was created a month ago. This determines the retention period for the backup. Therefore, the minimum retention period of this application data is one month. However, the organization might choose to retain the backup for a longer period of time because of internal policies or external factors, such as regulatory directives.

If the recovery point is older than the retention period, it might not be possible to recover all the data required for the requested recovery point. Long retention periods can be defined for all backups, making it possible to meet any RPOwithin the defined retention periods. However, this requires a large storage space, which translates into higher cost. Therefore, while defining the retention period, analyze all the restore requests in the past and the allocated budget.

RTO relates to the time taken by the recovery process. To meet the defined RTO, the business may choose the appropriate backup granularity to minimize recovery time. In a backup environment, RTO influences the type of backup media that should be used. For example, a restore from tapes takes longer to complete than a restore from disks.

## 9.6 BACKUP METHODS

Hot backup and cold backup are the two methods deployed for a backup. A hot backup is performed whilst users are still logged into a system, whereas a cold backup is done with all users offline. The reason for performing hot backups is that it minimizes downtime on a day-to-day basis, which is especially useful for systems that require 24/7 operation. The issue with hot backups is that if data is changed whilst the backup is being performed there may be some inconsistencies, such as the previous state of the file being included in the backup rather than the latest one. Hot backups also take up computer resources, so machine and server performance can be affected during backups.

Cold backups, sometimes known as offline backups, are the safest way to backup data as no files can be changed during the backup. Cold backups can be performed on a copy of data too, such as that stored in an offsite repository. The benefit of cold backups is that the backup can't be affected by live viruses or hacking attempts. They also won't be affected by power surges, making them the most reliable way to backup your data. Obviously, the downside is that during this time no users can access the system. It can also take longer to recover from a disaster with cold backups as moving the data from the cold backup site to being fully operational can cause delays.

Consistent backups of databases can also be done by using a cold backup. This requires the database to remain inactive during the backup. Of course, the disadvantage of a cold backup is that the database is inaccessible to users during the backup process.

Hot backups should be used when downtime has to be as low as possible (When you have a low RTO) and cold backups should be used when no users have to access the system. You don't have to use just one backup method of course. You could run hot backups throughout the week and then perform a cold backup on Friday evenings or over the weekend when users won't be using the system. Depending on the data set size, a cold backup may only take an hour or less, which may not cause any disruption to some businesses.

To ensure consistency, it is not enough to back up only the production data for recovery. Certain attributes and properties attached to a file, such as permissions, owner, and other metadata, also need to be backed up. These attributes are as important as the data itself and must be backed up for consistency.

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. BMR builds the base system, which includes partitioning, the file system layout, the operating system, the applications, and all the relevant configurations. The base system is recovered first by BMR before starting the recovery of data files. Some BMR technologies — for example server configuration backup (SCB) — can recover a server even onto dissimilar hardware.

## 9.7 BACKUP ARCHITECTURE

The common and widely used Backup Architecture is based on the Server-Client model. Figure 9-4 illustrates the backup architecture. Any backup architecture is composed of the following four components.

- Backup Servers
- Backup Clients
- Media Servers
- Backup Destinations/Targets

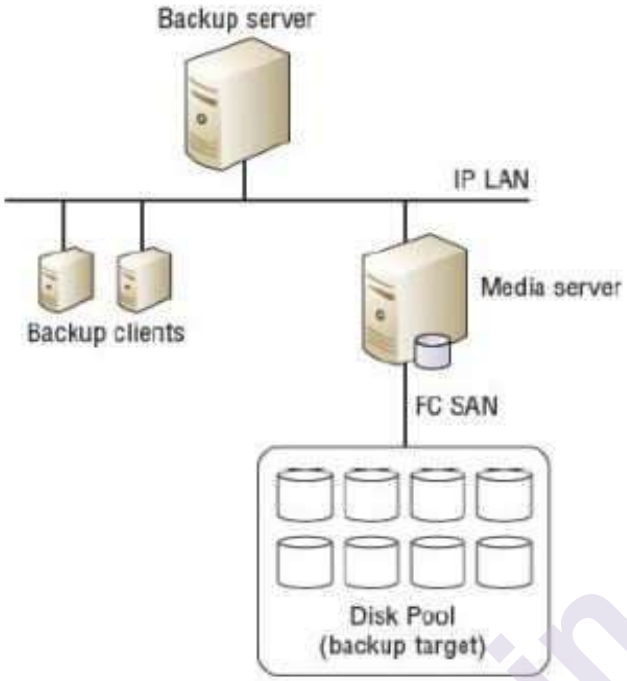



***Figure 9-4 Backup Architecture***

The **backup server** manages the backup operations and maintains the backup database, which contains information about the backup configuration and backup metadata. The backup configuration contains information about when to run backups, which client data to be backed up, and so on. The backup metadata contains information about the backed up data.

The role of a **backup client** is to gather the data that is to be backed up and send it to the backup server. The backup client can be installed on application servers, mobile clients, and desktops. It also sends the tracking information to the backup server.

**Media Servers** connect to the backup destinations and make it available to backup clients so that they can send data to the backup target. In IBM TSM terminology, media servers are referred as Primary Library Manager and other TSM servers as Library Clients. The media servers controls one or more backup devices. Backup devices may be attached directly or through a network to the Media Servers. The Media Servers sends the tracking information about the data written to the backup device to the backup server. Typically, this information is used for recoveries. For example, a media server might be connected to a pool of storage over an FC network and make that storage available to backup clients over an SAN.

A wide range of **backup destinations/targets** are currently available such as tape, disk, and virtual tape library. Traditional backup solutions primarily used tape as a backup destination and modern backup approaches tend to use disk based pools which are shared over SAN or LAN. Disk arrays can also be used as virtual tape libraries to combine the

benefits of Disk and Tape. Now, organizations can also back up their data to the cloud storage. Many service providers offer backup as a service that enables an organization to reduce its backup management overhead.

## 9.8 BACKUP AND RESTORE OPERATIONS

A significant network communication is built between different components of a backup infrastructure when backup operation is initiated. The backup operation is typically initiated by a server, but it can also be initiated by a client. The backup server initiates the backup process for different clients based on the backup schedule configured for them. For example, the backup for a group of clients may be scheduled to start at 11:00 p.m. every day.

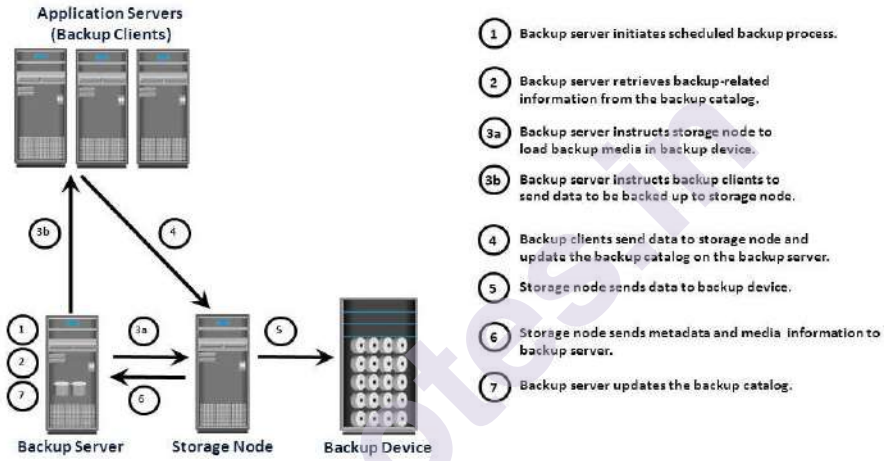



***Figure 9-5 Backup Operation***

The backup server coordinates the backup process with all the components in a backup environment (see Figure 9-5). Here the regularity of maintaining the backup is run by backup server. 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 gather the data to be backed up and send it over the network to the assigned storage node. After the backup data is sent to the storage node, the client sends some backup metadata (the number of files, name of the files, storage node details, and so on) to the backup server. The storage node receives the client data, organizes it, and sends it to the backup device. The storage node then sends additional backup metadata (location of the data on the backup device, time of backup, and so on) to the backup server. The backup server updates the backup catalog with this information.

After the data is backed up, it can be restored when required. A restore process must be manually initiated from the client. Some backup software has a separate application for restore operations. These restore

applications are usually accessible only to the administrators or backup operators. Figure 10-6 shows are store 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. Because all this information comes from the backup catalog, the restore application needs to communicate with the backup server.

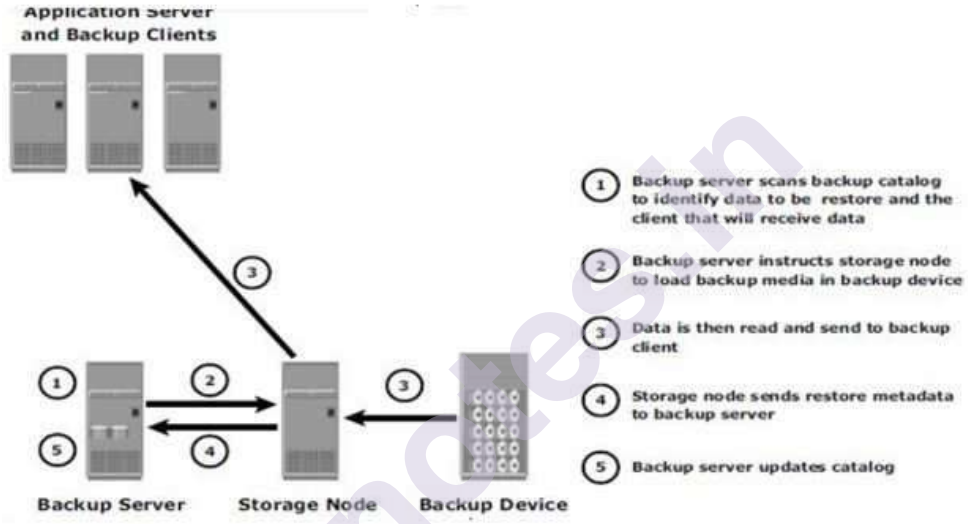



***Figure 9-6 Restore Operation***

The backup server instructs the appropriate storage node to mount the specific backup media onto the backup device. Data is then read and sent to the client that has been identified to receive the restored data. Some restorations are successfully accomplished by recovering only the requested production data. For example, the recovery process of a spreadsheet is completed when the specific file is restored. In database restorations, additional data, such as log files, must be restored along with the production data. This ensures consistency for the restored data.

## 9.9 BACKUP TOPOLOGIES

There are 4 topologies are used in a backup environment:

1. direct-attached backup
2. LAN-based backup
3. SAN-based backup
4. Mixed backup

**1- Direct-attached backup:**

In the direct connection backup mode, the backup data is directly backed up from the host to tape, without going through the LAN. The backup task is initiated by the backup client and directly backs up the data to the tape device connected to the client. In this model, we cannot perform centralized management and it is difficult to expand the existing environment. The main advantage of this backup topology is fast speed, and tape devices can maximize their own I/O speed. Because the tape device is closely connected to the data source and provided exclusively for the host, the speed of backing up and restoring data can be optimized. The disadvantage of direct-attached backups is that backups consume host I/O bandwidth, memory, and CPU resources, so they affect the performance of the host and its applications. In addition, direct-attached backups have distance limitations, especially when using short-range connections such as SCSI. The example in Figure 9-7 shows that the backup device is directly attached and dedicated to the backup client.

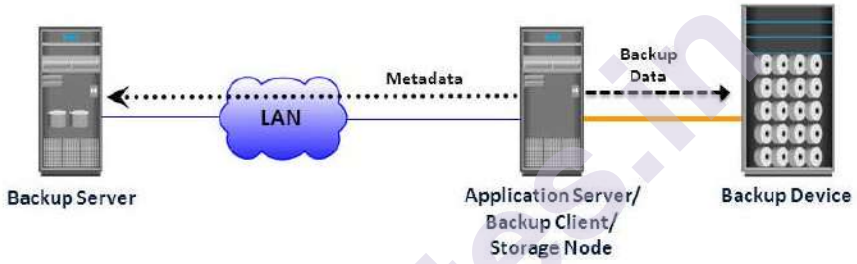



***Figure 9-7 Direct-attached Backup***

**2- LAN-based backup:**

In the LAN-based backup mode, the backup data is backed up from the host to the tape through the LAN. The backup server acts as a control center to control all backup tasks(see Figure 9-8). In this mode, we can perform centralized management but the high load rate of the LAN may be a problem because all data will pass through the LAN. The main advantage of this backup topology is the ability to centrally manage backup and tape resources, thereby improving operational efficiency. The disadvantage is that the backup process may affect the production system, client network, and applications because it consumes CPU, I/O bandwidth, LAN bandwidth, and memory.

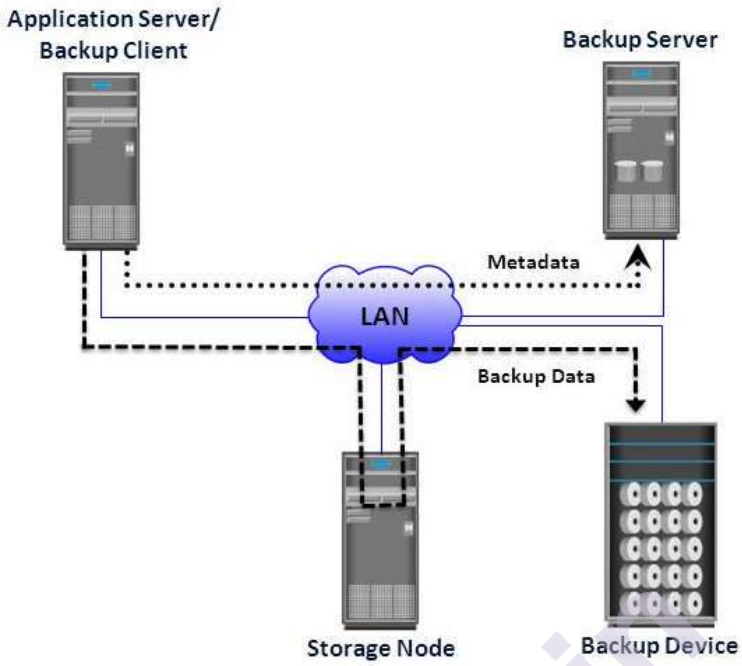



***Figure 9-8LAN-Based Backup***

**4- SAN-based backup:** In the SAN-based backup mode (LAN-Free), backup data is transferred through the SAN, and the LAN is only used to transfer metadata. The backup metadata contains information about the file being backed up, such as the file name, backup time, file size and permissions, file owner, and tracking information used to quickly locate and restore data. SAN-based backup optimizes the entire backup process, including providing optical fiber performance, high reliability, long distance, no LAN to transmit backup data, no need for a dedicated backup server, and high-performance backup and recovery. This model can provide better backup performance and more simplified management, but requires additional investment in facility construction. Figure 9-9 illustrates a SAN-based backup.

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.

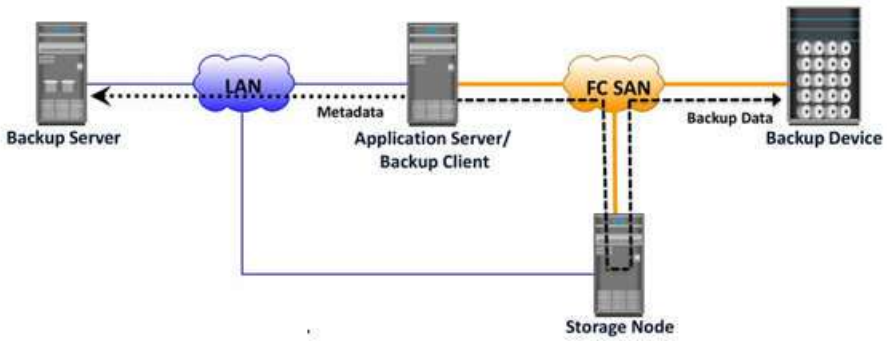



***Figure 9-9SAN-Based Backup***

**4- Mixed topology:** The *mixed topology* uses both the LAN-based and SAN-based topologies, as shown in Figure 9-10. This topology might be implemented for several reasons, including cost, server location, reduction in administrative overhead, and performance considerations.

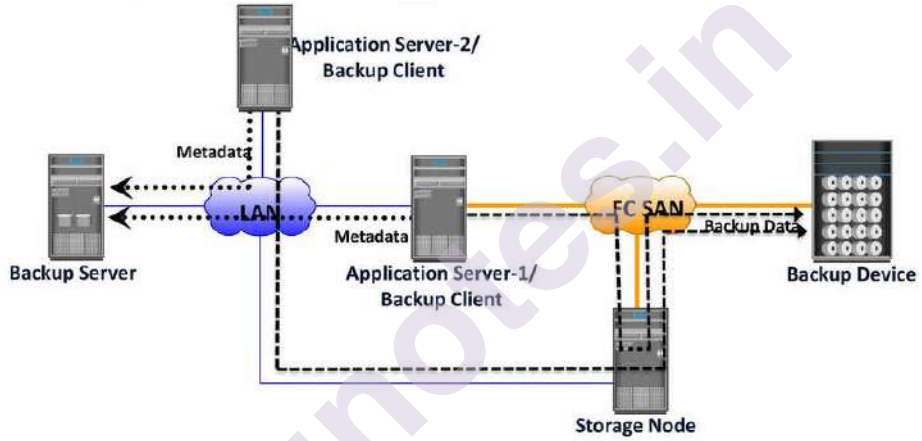



***Figure 9-10Mixed Backup***

## 9.10 BACKUP IN NAS ENVIRONMENTS

The use of a NAS head imposes a new set of considerations on the backup and recovery strategy in NAS environments. It heads use a proprietary operating system and file system structure that supports multiple file-sharing protocols. In the NAS environment, backups can be implemented in different ways: server based, server less, or using Network Data Management Protocol (NDMP).Common implementations are NDMP 2-way and NDMP 3-way.

### 9.10.1 Server-Based and Serverless Backup

In an *application server-based backup*, the NAS head retrieves data from a storage array over the network and transfers it to the backup client running on the application server. The backup client sends this data to the storage node, which in turn writes the data to the backup device. This results in overloading the network with the backup data and using application server resources to move the backupdata. Figure 9-11 illustrates server-based backup in the NAS environment.

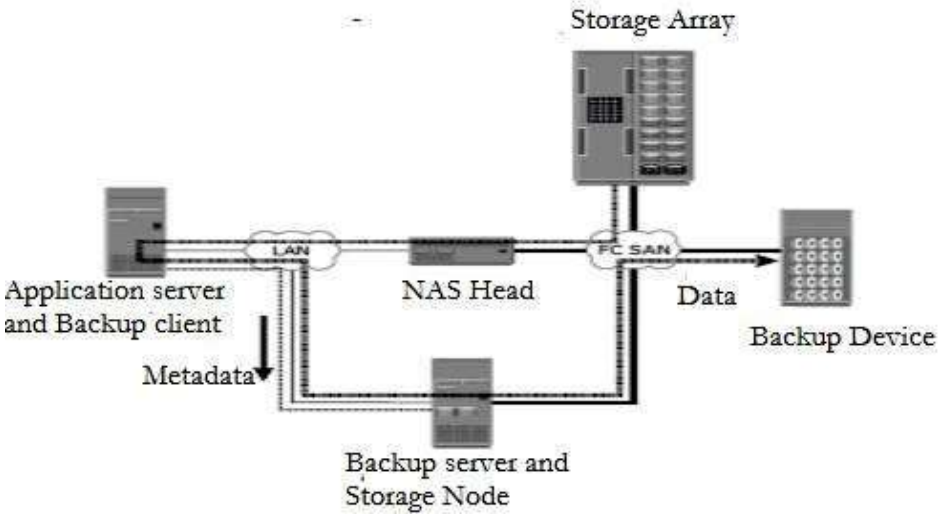



***Figure9-11 Server Based Backup in NAS environment***

In a *serverless backup*, the network share is mounted directly on the storage node. This avoids overloading the network during the backup process and eliminates the need to use resources on the application server. Figure 9-12 illustrates server less backup in the NAS environment. In this scenario, the storage node, which is also a backup client, reads the data from the NAS head and writes it to the backup device without involving the application server. Compared to the previous solution, this eliminates one network hop.

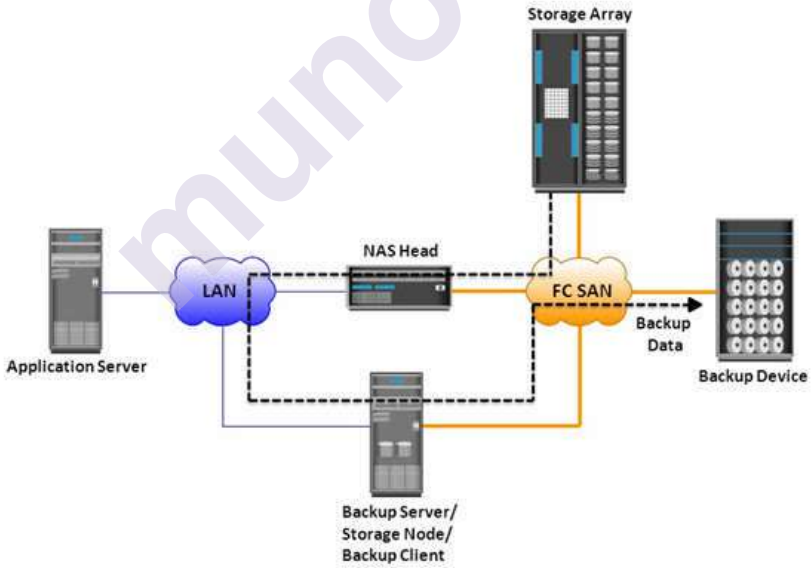



***Figure9-12 Serverless Backup in NAS environment***

### 9.10.2 NDMP-Based Backup

NDMP is a protocol designed for efficient NAS backups. It is similar to serverless backups where the data can be sent directly from the NAS device to the backup device without having to pass through a backup

media server. As the amount of unstructured data continues to grow exponentially, organizations face the daunting task of ensuring that critical data on NAS systems are protected. Most NAS heads run on proprietary operating systems designed for serving files. To maintain its operational efficiency generally it does not support the hosting of third-party applications such as backup clients.

NDMP is an industry-standard TCP/IP-based protocol specifically designed for a backup in a NAS environment. It communicates with several elements in the backup environment (NAS head, backup devices, backup server, and so on) for data transfer and enables vendors to use a common protocol for the backup architecture. Data can be backed up using NDMP regardless of the operating system or platform. NDMP backs up and restores data without losing the data integrity and file system structure with respect to different rights and permission in different file systems. NDMP optimizes backup and restore by leveraging the high-speed connection between the backup devices and the NAS head. In NDMP, backup data is sent directly from the NAS head to the backup device, whereas metadata is sent to the backup server.

The key components of an NDMP infrastructure are NDMP client and NDMP server. NDMP client is the NDMP enabled backup software installed as add-on software on backup server. The NDMP server has two components: data server and media server. The backup operation occurs as follows:

- Backup server uses NDMP client and instructs the NAS head to start the backup.
- The NAS head uses its data server to read the data from the storage.
- The NAS head then uses its media server to send the data read by the data server to the backup device

The actual backup data is either directly transferred to backup device (NDMP 2-way) or through private backup network (NDMP 3-way), by the NAS head.

- NDMP 2-way (Direct NDMP method) – In this method, the backup server uses NDMP over the LAN to instruct the NAS head to start the backup. The data to be backed up from the storage is sent directly to the backup device. In this model, network traffic is minimized on the production network by isolating backup data movement from the NAS head to a locally attached backup device. During the backup, metadata is transferred via NDMP over the LAN to the backup server. During a restore operation, the backup server uses NDMP over the LAN to instruct the NAS to start restoring files. Data is restored from the locally attached backup device.

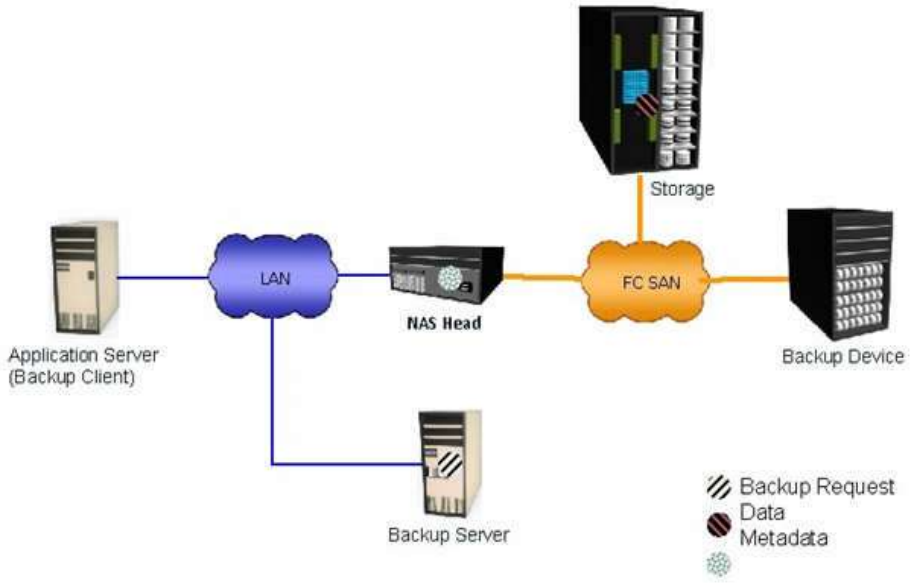



***Figure9-13NDMP2-way Backup***

- NDMP 3-way (Remote NDMP method) – In this method, the backup server uses NDMP over the LAN to instruct the NAS head to start backing up data to the backup device attached to NAS head. These NAS devices can be connected over a private backup network to reduce the impact on the production LAN network. During the backup, the metadata is sent via NDMP by the NAS head to the backup server over the production LAN network. NDMP 3-way is useful when there are limited backup devices in the environment. It enables the NAS head to control the backup device and share it with other NAS heads by receiving backup data through NDMP.

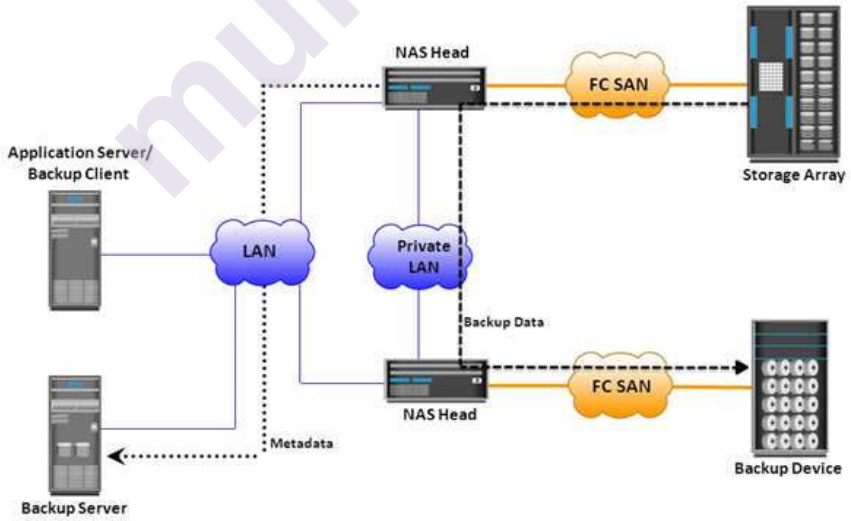



***Figure9-14NDMP3-way Backup***

## 9.11 BACKUP TARGETS

There are different devices are available for backup targets. Tape and disk libraries are the two most commonly used backup targets. In the past, tape technology was the predominant target for backup due to its low cost. But performance and management limitations associated with tapes and the availability of low-cost disk drives have made the disk a viable backup target.

### 9.11.1 Backup to Tape

Tapes, a low-cost solution, are used extensively for backup. Tape drives are used to read/write data from/to a tape cartridge (or cassette).Tape drives are referred to as sequential, or linear, access devices because the data is written or read sequentially. A tape cartridge is composed of magnetic tapes in a plastic enclosure. *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.

Several types of tape cartridges are available. They vary in size, capacity, shape, density, tape length, tape thickness, tape tracks, and supported speed.

***Physical Tape Library***

A tape library is a high-capacity storage system used for storing, retrieving, reading from and writing to tape cartridges. A tape library contains racks of cartridges and multiple tape drives with a robotic system used for automatically changing tape cartridges. A filing system that uses a barcode reader or an RF scanner allows the tape library to find the correct tape to load either for writing or for reading. Figure 9-15 shows a physical tape library.

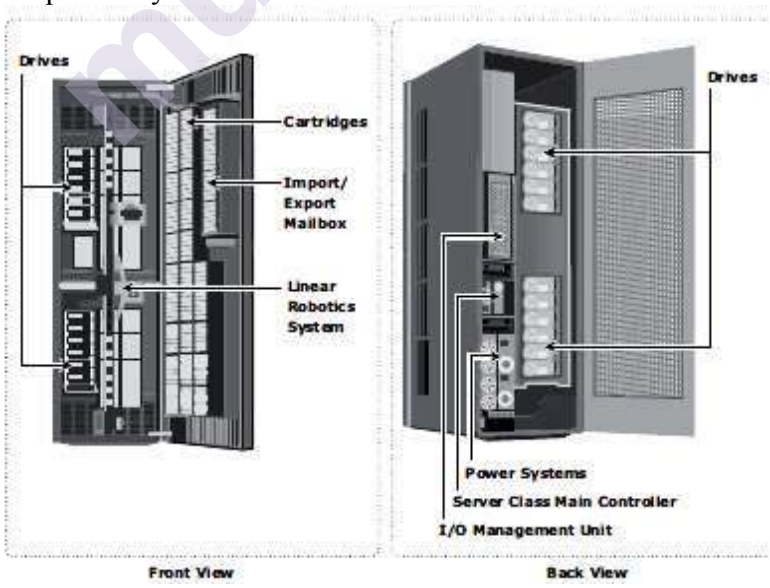



***Figure 9-15 Physical Tape Library***

*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 (refer to Figure 9-15 Front View).

When a backup process starts, the robotic arm is instructed to load a tape to a tape drive. This process adds delay to a degree depending on the type of hardware used, but it generally takes 5 to 10 seconds to mount a tape. After the tape is mounted, additional time is spent to position the heads and validate header information. This total time is called *load to ready time*, and it can vary from several seconds to minutes. The tape drive receives backup data and stores the data in its internal buffer. This backup data is then written to the tape in blocks. During this process, it is best to ensure that the tape drive is kept busy continuously to prevent gaps between the blocks. This is accomplished by buffering the data on tape drives. The speed of the tape drives can also be adjusted to match data transfer rates.

Tape drive *streaming* or *multiple streaming* writes data from multiple stream son a single tape to keep the drive busy. As shown in Figure 9-16, 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 because all the extra data from the other streams must be read and discarded while recovering a single stream.

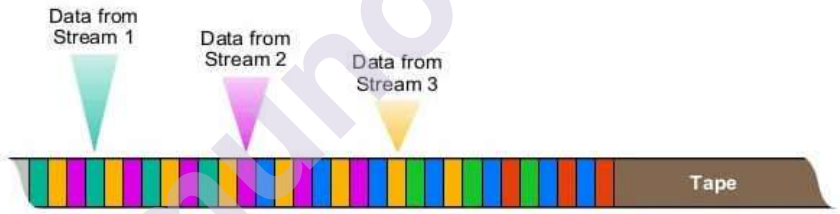



***Figure 9-16 Multiple streams on tape media***

Many times, even the buffering and speed adjustment features of a tape drive fail to prevent the gaps, causing the "*shoe shining effect*" or "*back hitching*." *Shoe shining* is the repeated back and forth motion a tape drive makes when there is an interruption in the backup data stream. For example, if a storage node ends data slower than the tape drive writes it to the tape, the drive periodically stops and waits for the data to catch up. After the drive determines that there is enough data to start writing again, it rewinds to the exact place where the last write took place and continues. This repeated back-and-forth motion not only causes a degradation of service, but also excessive wear and tear to tapes.

When the tape operation finishes, the tape rewinds to the starting position and it is uncounted. The robotic arm is then instructed to move the unmountedtape back to the slot. When a file or a group of files require restores, the tape must move to that file location sequentially before it can

start reading. This process can take a significant amount of time, especially if the required files are recorded at the end of the tape. Modern tape devices have an indexing mechanism that enables a tape to be fast forwarded to a location near the required data.

Tapes are extensively used for the on-premises and long-term off-site retention of data. Installing a new tape system takes a tremendous investment. We know that tapes are not that costly, but it may be very expensive to transport the tapes from one data center to the other safely.

### 9.11.2 Backup to Disk

backup-to-disk has several advantages over traditional tape backup for both technical and business reasons. Backup-to-disk systems offer ease of implementation, reduced cost, speed and improved quality of service. With continued improvements in storage devices to provide faster access and higher storage capacity, a prime consideration for backup and restore operations, backup-to-disk will become more prominent in organizations.

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. Figure 9-17 shows a recovery scenario comparing tape versus diskin a Microsoft Exchange environment that supports 800 users with a 75 MBmailbox size and a 60 GB database. As shown in the figure, a restore from the disk took 24 minutes compared to the restore from a tape, which took 108 minutes for the same environment.

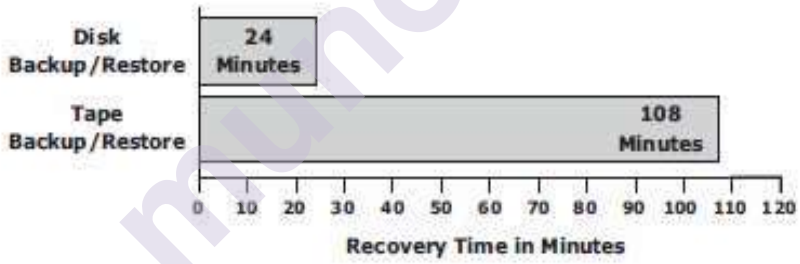



***Figure 9-17 Tape verses disk restore***

Recovering from a full backup copy stored on disk and kept onsite provides the fastest recovery solution. Using a disk enables the creation of full backups more frequently, which in turn improves RPO and RTO.

Disk-based backups generally provide better data security than tape. Physically accessing hard disks contained in a drive array is harder than gaining access to tapes in cold storage. Physical disks in drive arrays are usually monitored closely. Furthermore, data contained in disk-based backups is usually spread across multiple drives in what is called Redundant Array of Independent Disks (RAID systems). The complete set of data required to reconstruct a virtual machine or many virtual machines may be spread across multiple hard disks in the RAID group.

### 9.11.3 Backup to Virtual Tape

A virtual tape library (VTL) is a technology for data backup and recovery that uses tape libraries or tape drives along with their existing software for backup. The virtual tape library system emulates the former magnetic tape devices and data formats, but performs much faster data backups and recovery. It is able to avoid the data streaming problems that often occur with tape drives as a result of their slow data transfer speeds.

***Virtual Tape Library***

A *virtual tape library* (VTL) has the same components as that of a physical tape library, except that the majority of the components are presented as virtual resources. For the backup software, there is no difference between a physical tape library and a virtual tape library. Figure 9-18 shows 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 ona LUN. A virtual tape can span multiple LUNs if required. File system awareness is not required while backing up because the virtual tape solution typically uses raw devices.

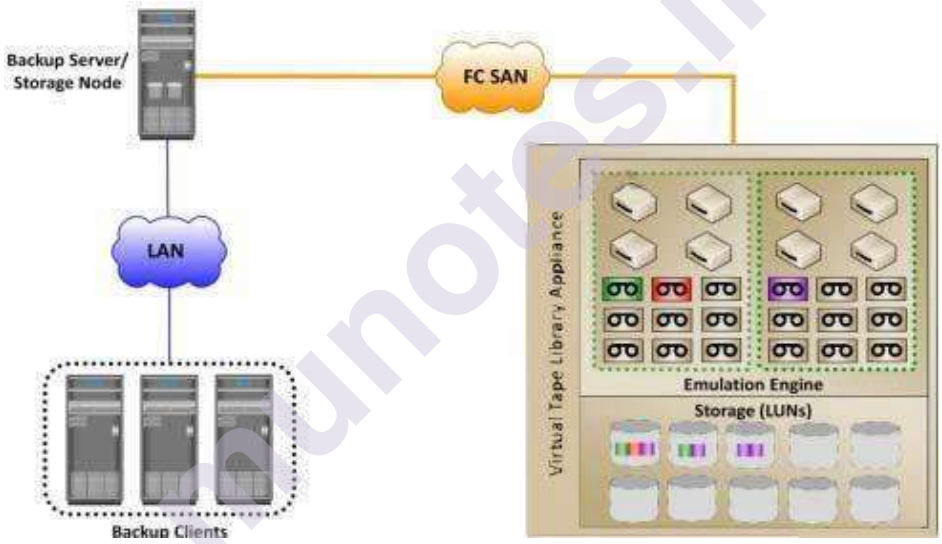



***Figure 9-18 Virtual Tape Library***

Similar to a physical tape library, a robot mount is virtually performed when a backup process starts in a virtual tape library. However, unlike a physical tape library, where this process involves some mechanical delays, in a virtual tape library it is almost instantaneous. Even the *load to ready* time is much less than in a 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. Inmost cases, data is written to the virtual tape immediately. Unlike a physical tape library, the virtual tape library is not constrained by the sequential access and shoe shining effect. When the operation is complete, the backup software issues a rewind command. This rewind is

also instantaneous. The virtual tape is then uncounted, and the virtual robotic arm is instructed to move it back to a virtual slot.

The steps to restore data are similar to those in a 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.

A virtual tape library appliance offers a number of features that are not available with physical tape libraries. Some virtual tape libraries offer *multiple emulation engines* configured in an active cluster configuration. An engine is a dedicated server with a customized operating system that makes physical disks in the VTL appear as tapes to the backup application. With this feature, one engine can pick up the virtual resources from another engine in the event of any failure and enable the clients to continue using their assigned virtual resources transparently.

Data replication over IP is available with most of the virtual tape library appliances. This feature enables virtual tapes to be replicated over an inexpensiveIP network to a remote site. As a result, organizations can comply with offsite requirements for backup data. Connecting the engines of a virtual tape library appliance to a physical tape library enables the virtual tapes to be copied onto the physical tapes, which can then be sent to a vault or shipped to an offsite location.

## 9.12 DATA DEDUPLICATION FOR BACKUP

Traditional backup solutions do not provide any inherent capability to prevent duplicate data from being backed up. Earlier back up leads to a lot of duplicate data. Backing up duplicate data results in unnecessary consumption of resources, such as storage space and network bandwidth.

Deduplication is also one of the Storage Capacity Optimization techniques which will identify the duplicate data and making sure that duplicate data is not stored again. Storage systems that implement deduplication technique achieve this by inspecting data and checking whether copies of the data already exist in the system. If a copy of this data already exists, instead of storing additional copies, pointers are used to point to the copy of the data.

For example, a typical email system might contain 50 instances of the same 2 MB file attachment. If the email platform is backed up or archived, all 50 instances are saved, requiring 100 MB storage space. With data deduplication, only one instance of the attachment is actually stored; each subsequent instance is just referenced back to the one saved copy reducing storage and bandwidth demand to only 2 MB.

Technologies, such as data deduplication, improves storage efficiency and reduces the amount of data that needs to be transmitted over the network. This not only enhances backup speed but also frees up space for additional files, which in turn leads to significant cost savings over time. By eliminating duplicate copies, dedupe optimizes storage capacity,

increases on-appliance retention and reduces the landing zone space required for backups.

### 9.12.1 Data Deduplication Methods

There are two methods of deduplication:

1. File level deduplication
2. Subfile leveldeduplication

**1- File-level deduplication:** File-level deduplication, also called Single-Instance storage, compares files to be archived with the ones already stored. It detects and removes redundant copies of identical files. While storing a file, its attributes are checked against an index, if it is unique, it is stored, and if not, only a pointer (stub) is created to an existing similar file. This is simple and fast but does not address the problem of the duplicate content within the file.

**2- Subfile deduplication:** Sub-file deduplication breaks the file into smaller chunks (contiguous blocks of data) and then uses a specialized algorithm to detect redundant data within and across a file. As a result, it eliminates duplicate data across files. There are 2 methods of sub-file deduplication:

- **Fixed-Length block:** In this process, a file is divided into fixed-length blocks, and a hash algorithm is used to find redundant data.
- **Variable-Length Segment:** It's an alternative that divides a file into chunks of different sizes leading to dedupe efforts to achieve better results.

### 9.12.2 Data Deduplication Implementation

Deduplication for backup can be implemented in 2 ways:

1. Source-Based Data Deduplication
2. Target-Based Data Deduplication

***1-Source-Based Data Deduplication***

In this scenario deduplication performed on client side i.e. before it transmitted By processing the data before transmitting we can reduce the transmitted amount of data and therefore it reduces the network bandwidth and this less bandwidth is required for the backup software. Deduplication on source side uses the engine at client side which checks for the duplication against the deduplication index which is located on the backup server. This is done with the help of the backup agent who is aware of the deduplication which is located at the client side and who is responsible for backs up only unique data or blocks. And those unique blocks of data will be transmitted to the disk. The result of this kind of technology i.e. source based deduplication improves bandwidth as well as the storage utilization. Figure 9-19 shows source-based data deduplication.

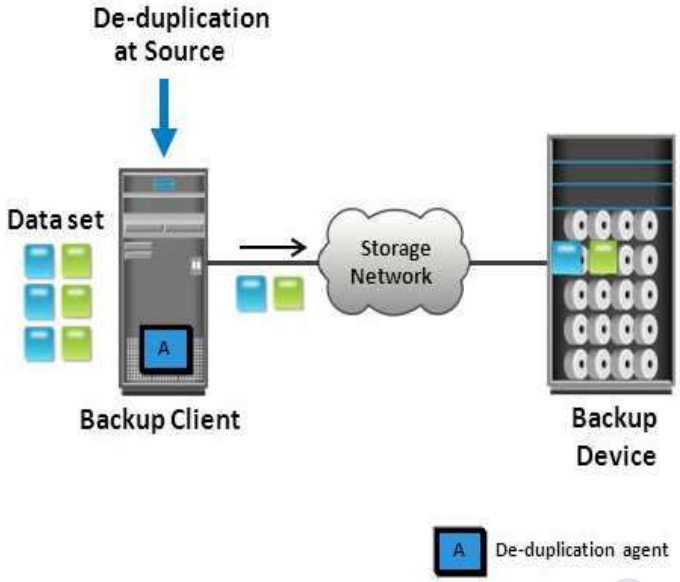



***Figure 9-19 Source-based data deduplication***

Source-based deduplication increases the overhead on the backup client, which impacts the performance of the backup and application running on the client.Source-based deduplication might also require a change of backup software ifit is not supported by backup software.

***2-Target-Based Data Deduplication***

Target-based deduplication, sometimes referred to as hardware-based deduplication or Destination side deduplication which is widely used in current backup environments. In target-based deduplication, the process of deduplication occurs on the target machine, such as a deduplicating backup appliance. These appliances tend to be purpose-built appliances with their own CPU, RAM, and persistent storage (disk). This approach relieves the host (source) of the burden of deduplicating, but it does nothing to reduce network bandwidth consumption between source and target. It is also common for deduplicating backup appliances to deduplicate or compress data being replicated between pairs of deduplicating backup appliances. Figure 9-20 shows target-based data deduplication.

Some manufacturers now tout the ability to deliver source and target-based deduplication under a single management framework. In this scenario, all backup workloads are controlled and optimized from a single console and a common disk storage appliance maintains all the backup data. While this seems like the logical progression for architecting deduplication into data center environments, it has not been deployed on the same scale to date as homogenous source and target-based solutions.

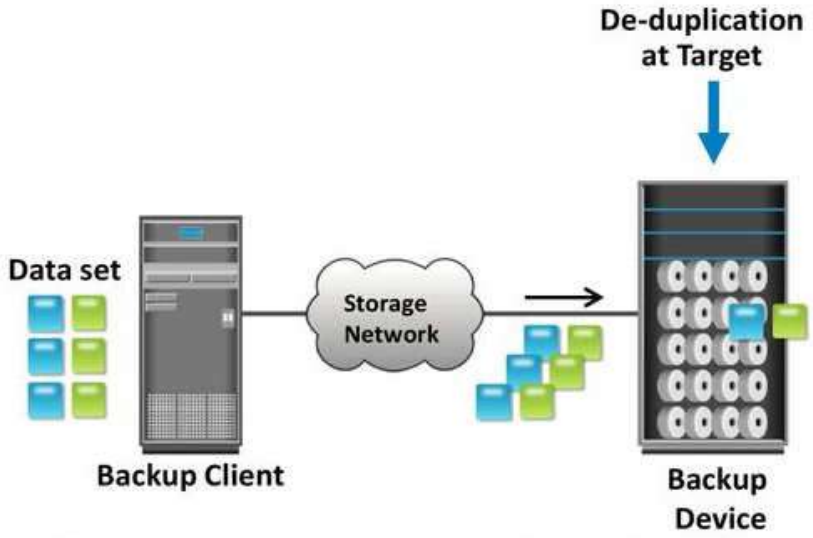



***Figure 9-20Target-based data deduplication***

## 9.13 BACKUP IN VIRTUALIZED ENVIRONMENTS

In a virtualized environment, it is vital 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.

In the *traditional backup approach*, a backup agent is installed either on the virtual machine (VM) or on the hypervisor. Figure 9-21 shows the traditional VM backup approach. If the backup agent is installed on a VM, the VM appears as a physical server to the agent. The backup agent installed on the VM backsup 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.

If the backup agent is installed on the hypervisor, the VMs appear as aset of files to the agent. So, VM files can be 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 traditional backup method can cause high CPU utilization on the server being backed up.

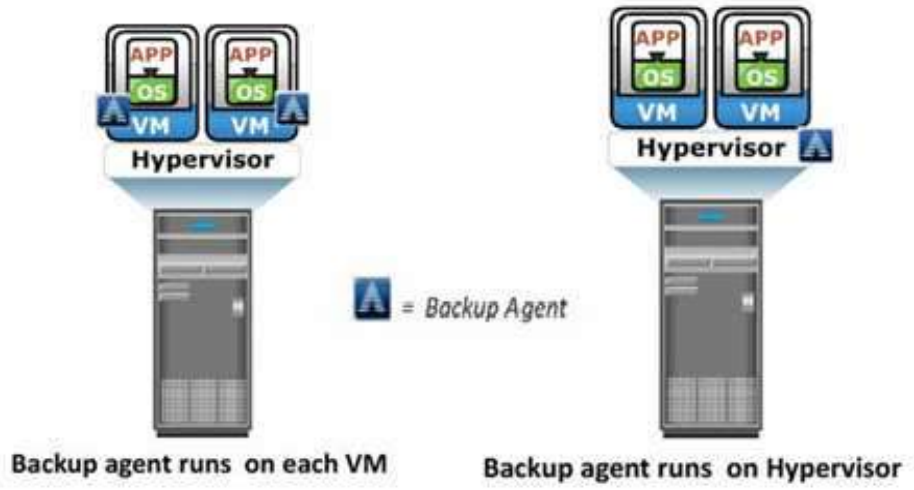



***Figure 9-21 Traditional VM Backup***

In the traditional approach, the backup should be performed when the server resources are idle or during a low activity period on the network. Also consider allocating enough resources to manage the backup on each server when a largenumber of VMs are in the environment.

*Image-based backup* operates at the hypervisor level and essentially takes a snapshot of the VM. It creates a copy of the guest OS and all the data associated with it(snapshot of VM disk files), including the VM state and application configurations. The backup is saved as a single file called an "image," and this image is mountedon the separate physical machine–proxy server, which acts as a backup client.The backup software then backs up these image files normally. (see Figure 9-22).This effectively offloads the backup processing from the hypervisor and transfers the load on the proxy server, thereby reducing the impact to VMs running on the hypervisor. Image-based backup enables quick restoration of a VM.

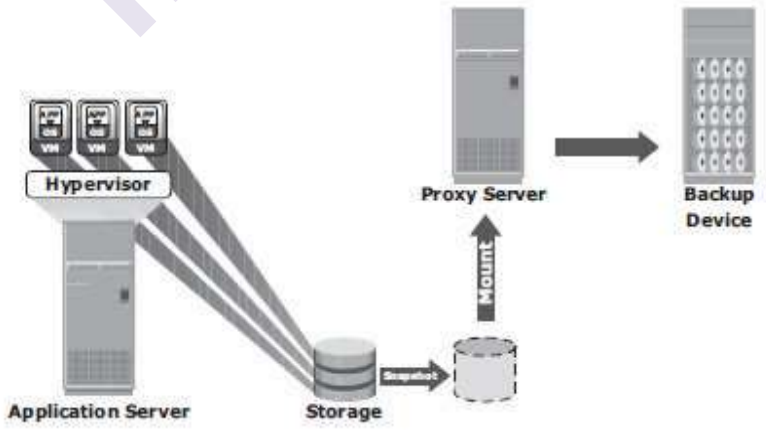



***Figure 9-22Image-based Backup***

The use of deduplication techniques significantly reduces the amount of datato be backed up in a virtualized environment. The effectiveness of deduplicationis identified when VMs with similar configurations are deployed in a datacenter. The deduplication types and methods used in a virtualized environment are the same as in the physical environment.

## 9.14 DATA ARCHIVE

Data is accessed and modified at varying frequencies between the time it is created and discarded. Some data frequently changes, for example, data accessed by an Online Transaction Processing (OLTP) application. Another category of data is fixed content, which defines data that cannot be changed. X-rays and pictures are examples of fixed content data. It is mandatory for all organizations to retain some data for an extended period of time due to government regulations and legal/contractual obligations. Some examples of fixed content asset include electronic documents, e-mail messages, Web pages, and digital media. A repository where fixed content is stored is known as an archive.

It can be implemented as online, near line, or offline based on the means of access:

- **Online archive:** The storage device is directly connected to the host to make the data immediately available. This is best suited for active archives.
- **Nearline archive**: The storage device is connected to the host and information is local, but the device must be mounted or loaded to access the information.
- **Offline archive:** The storage device is not directly connected, mounted, or loaded. Manual intervention is required to provide this service before information can be accessed.

An archive is often stored on a write once read many (WORM) devices, such as a CD-ROM. These devices protect the original file from being overwritten. Some tape devices also provide this functionality by implementing file locking capabilities in the hardware or software. Archives implemented using tape devices and optical disks involve many hidden costs. The traditional archival process using optical disks and tapes is not optimized to recognize the content, so the same content could be archived several times. Additional costs are involved in offsite storage of media and media management. Tapes and optical media are also susceptible to wear and tear.

*Content addressed storage* (CAS) is disk-based storage that has emerged as an alternative to tape and optical solutions. CAS meets the demand to improve data accessibility and to protect, dispose of, and ensure service-level agreements(SLAs) for archive data.

## 9.15 ARCHIVING SOLUTION ARCHITECTURE

Archiving solution architecture has three key components: archiving agent, archiving server, and archiving storage device (see Figure 9-23).*An archiving agent* is software installed on the application server. The agent is responsible for scanning the data that can be archived based on the policy defined on the archiving server. Next data is identified for archiving, it is sent to the archiving server by the agent. Then the original data on the application server is replaced with a stub file, which contains the address of the archived data. The size of this file is small and significantly saves space on primary storage, & is used to retrieve the file from the archive storage device.

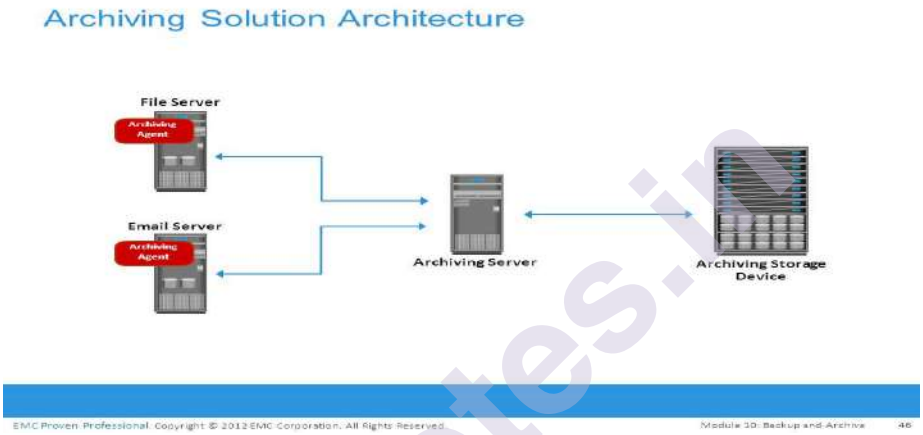



***Figure 9-23 Archiving solution architecture***

*An archiving server* is software installed on a host that enables administrators to configure the policies for archiving data. Policies can be defined based on file size, file type, or creation/ modification/access time. The archiving server receives the data to be archived from the agent and sends it to the archive storage device.

*An archiving storage device* stores fixed content. Different types of storage media options such as optical, tapes, and low-cost disk drives are available for archiving.

## 9.16 SUMMARY

A*backup* is an additional copy of production data, created and retained for thesole purpose of recovering lost or corrupteddata. *Data archiving* is the process of moving data that is no longer actively used,from primary storage to a low-cost secondary storage. Backup granularity depends on business needs and the required RTO/RPO.Based on the granularity, backups can be categorized as full, incremental andcumulative. Hot backup and cold backup are the two methods deployed for a backup.A backup system commonly uses the client-server architecture with a backup server and multiple backup clients.Three basic

topologies are used in a backup environment: direct-attached backup,LAN-based backup, and SAN-based backup.*NDMP* is an industry-standard TCP/IP-based protocol specifically designedfor a backup in a NAS environment.

## 9.17 REVIEW QUESTIONS

1. What is backup? What are purposes of backups?
2. What are backup considerations? Explain in detail.
3. Explain different backup granularity levels in detail.
4. Explain methods deployed for a backup in detail.
5. Explain Backup Architecture with diagram.
6. Explain Backup operation with diagram.
7. Explain restore operation with diagram.
8. Explain different backup topologies in detail.
9. Explain Server-Based and Serverless Backup in detail.
10. Explain NDMP-Based Backup with diagram.
11. Explain backup to tape in detail.
12. Explain backup to disk in detail.
13. Explain backup to virtual tape in detail.
14. What is Data deduplication? Explain different data deduplication methods.
15. How backup is done in Virtualized Environments?
16. What is data archive? How it can be implemented?
17. Explain Archiving Solution Architecture with diagram.

## 9.18 REFERENCES


- Information Storage and Management: Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environmentsby Somasundaram Gnanasundaram and Alok Shrivastava, 2nd Edition Publisher: John Wiley & Sons.
- https://cloudian.com/guides/data-backup/data-archive/
- https://cloudian.com/guides/data-backup/data-backup-in-depth/
- https://www.mycloudwiki.com/san/backup-methods/
- https://spanning.com/blog/types-of-backup-understanding-full-differential-incremental-backup/
- https://www.oo-software.com/en/different-methods-for-data-backups
- https://helpcenter.veeam.com/docs/backup/vsphere/backup_architecture.html?ver=110
- https://www.jigsawacademy.com/blogs/cloud-computing/deduplication/


❖❖❖❖

# Unit IV

# 10

# LOCAL REPLICATION

**Unit Structure**

## 10.0 OBJECTIVES:

This chapter would make you understand the following concepts:

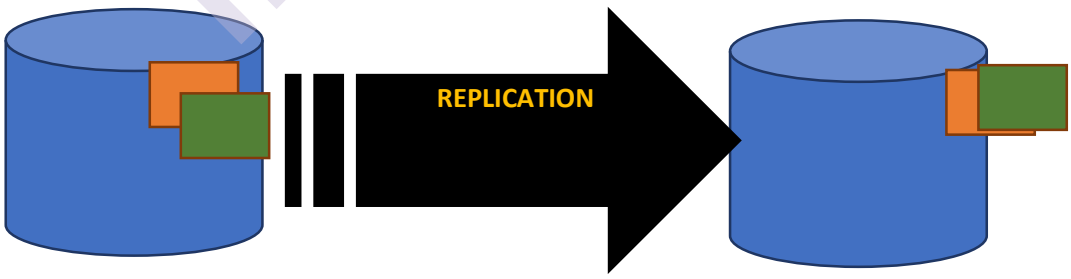
- Local replication and the uses of local replicas
- Data Consistency considerations when replicating file systems and databases
- Different replication technologies: Host-based and Array-based
- Creating multiple replicas

## 10.1 INTRODUCTION:

In this digital environment data storage and management are becoming more complex. It is difficult for organization to adopt real time data integration strategies and will help for better management of large volumes of complex data. The primary aim of the replication is to improve data availability and accessibility with the help of cloud storage.

Data Replication is used for creating one or more exact copies of storage database as well as for the purpose of fault tolerance. Replication provides recoverability and restart ability.

Recoverability is a service that enables the restoration of remote machine data with the help of cloud-based system. To avoid database loss or database corruption we used the recoverability service in cloud storage. It gives a business recover from any disaster, for the purpose of recovery time objective (RTO) and recovery point objective (RPO) as part of their disaster recovery plan. Replication is the process of reproducing data and Replica is the exact copy.Replication can be classified into two major categories namely Local Replication and Remote Replication. Local Replication is the replicating data within the same array or the same data center.




**Source** **Replica (Target)**

**Figure: 10.1 Local Replication**

## 10.2 USES OF LOCAL REPLICAS:

- Alternate source for backup

  For the backup purpose local replica maintains exact point-in-time (PIT) copy of the source database. The various operations and services are available for backup.

- Fast Recovery

  If data corrupt and loss on the source side then local replica can be used to recover the corrupted or lossy data.

- Decision support

  Reporting is the main aim in Decision support which will reduce the input/output pressure on the production device.

- Testing Platform

  If the test gives the successful result, then upgradation can be implemented on the production side.

- Data migration

  Data migration is for smaller capacity of data to larger capacity of data.

## 10.3 DATA CONSISTENCY

Consistency ensures the usability of replica. It can be achieved with the various ways for file system and database before creating the replica. Consistent replica ensures that the data buffered in the host is captured on the disk when the replica is created.

The data staged in the cache and not yet committed to the disk should be flushed before taking the replica. Storage array operating environment takes care of flushing its cache before the replication operation is started.

### 10.3.1 Replicated file system consistency

File system is two types namely offline and online. Offline file system is Un-mount file system while Online file system is flushing host buffers.

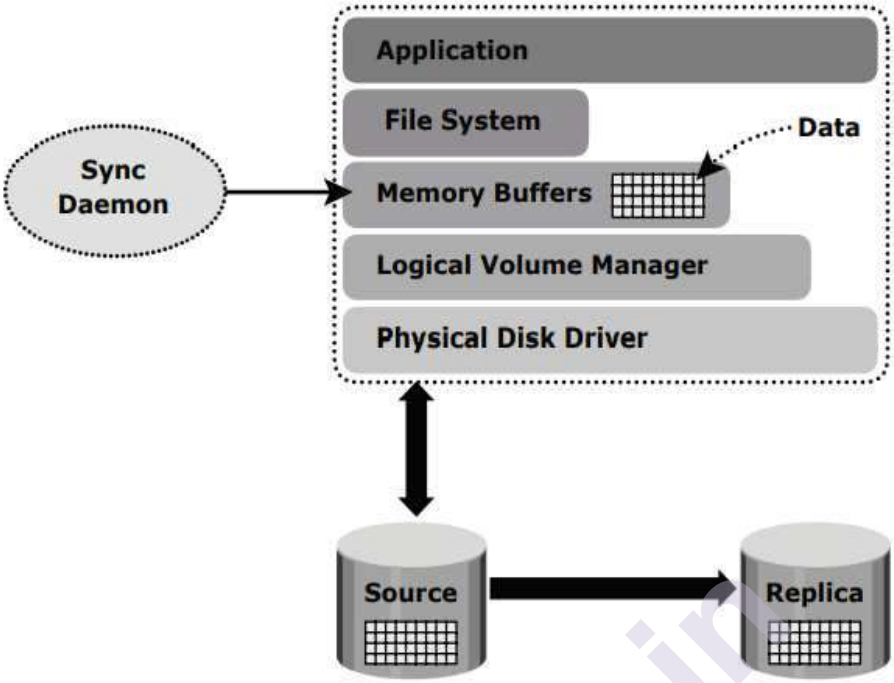



**Figure 10.3.1: File System replication**

Buffer data in host memory of the file system is useful to improve application response time.

Data written periodically with the help of buffered information in to the disk. In Unix operating systems, the sync daemon is the process where it flushes the buffer data into the disk at the given intervals and some-times in between the given intervals replica may be created, Therefore the use of host memory buffer is it flushed to ensure data consistency on the replica, before its creation.

In the figure 10.3.1 shows the flushing of buffer to its source, which is then replicated. If the host memory buffers, are not flushed, data on the replica will not contain the information that was buffered in the host. Data is automatically mounted to the replica and buffers would be flushed. File system replication process is completed, the replica file system can be useful for different operations.

### 10.3.2 Replicated database consistency

Database consistency is of two types offline and online. Offline database consistency will give the result shutdown database and Online database system will give two types of result a) Using dependent write input and output principles b) Holding input and outputs to source before creating replica.

Database can be stored in various files, file systems as well as various devices. The aim of replicated consistently is to ensure that the replica is restorable and restart-able.

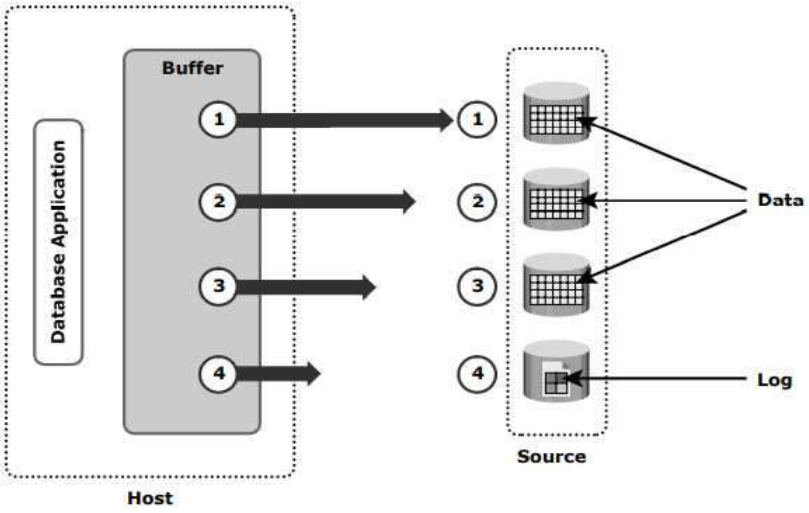



**Figure 10.3.2 Dependent write consistency on sources**

If the database is offline then there is no operations of input and output, no updates will occur during offline so replica will be consistent.

If the database is online then there is availability of input and output operations. Whenever transaction occurs the database will also be updated continuously. In online mode database backup is also consistent when the changes made to the database. It requires additional steps for taking backup and restore. We can do these backup process automatically for reducing human error and alleviating administrative work. Most of the database support online or hot backups. When the database is in the hot backup mode, there will be increased logging activity of that time.

Steps/sequence of operations in a hot backup mode.

1) To issue a database checkpoint to flush buffers to disk and place the database in hot backup mode.
2) Copy of Point-in-time (PIT) is taken out for the hot backup mode.
3) Logs are collected and then applied to the replica to restore the database consistently.

Figure 10.3.2 shows the process of flushing the buffer from host to source: Input/Outputs. The processes 1 to 4 must be complete, if the transaction to be completed successfully. Input/Output 4 is depending on Input/Output 3 will give the result only after completing the process 3. Input/Output 3 is depending on Input/Output 2, which will be depend on Input/Output 1. Each Input/Output completes only after completion of the previous Input/Outputs.

Dependent write Input/Output Principle

- Dependent Write: A write input/output that will not be issued by an application until a prior related write input/output has completed.
  - Logical dependency, not a time dependency
- Inherent in all database management systems (DBMS)
  - E. g. page (data) write is dependent write Input/output based on a successful log write
- Necessary for protection against local outages
  - Power failures create a dependent write consistent image
  - A restart transforms the dependent write consistent to transitionally consistent
    - i.e. Committed transactions will be recovered, in-flight transaction will be discarded.

During replica creation all the writes to the source devices are get captured on the replica devices for ensuring data consistency.
Figure 10.3.3 shows the process of replication from source to replica, Input/Output processes 1 to 4 must be carried out for the data to be consistent on the replica.

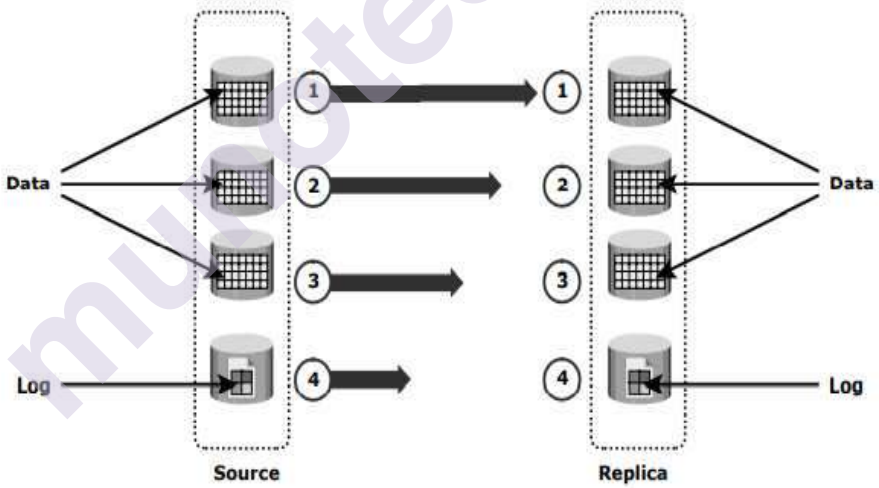



**Figure 10.3.3 Dependent write consistency on replica**

Point-in-Time (PIT) copy for multiple devices created very quickly. Input/Output transaction 3 and 4 were copied to the replica devices, bout input/output transactions 1 and 2 were not copied. In this case, the data on the replica is not consistent with the data on source. If the data associated with the transaction will be not available on replica, then replica must be inconsistent.

Another method to ensure the consistency is to make sure that write Input/Output to all sources devices is held during the creation of replica. This creates a consistent image on the replica. If the input/output

operation is held for too long time then databases and applications can time out automatically.

## 10.4 LOCAL REPLICATION TECHNOLOGIES

There are two major technologies for Local Replication namely,

**1) Host-Based**

Example of Host-Based local replication technologies are file system replication and LVM (Logical Volume managers)-Based replication.

**2) Storage-based**

Full-volume mirroring, Pointer-based full-volume replication and pointer-based virtual replication can be implemented with the help of Storage array-based replication.

### 10.4.1 Local Replication with Host-Based

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

#### 10.4.1.1 Replication with LVM-Based

- In Logical Volume managers (LVM)-Based replication, the logical volume manager is responsible for creating and controlling the host-level logical volumes. Logical Volume managers 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. A volume group can have multiple logical volumes. Logical Volume Managers-Based replication, each logical block in a logical volume is mapped to two physical blocks on two different physical volumes. An application write to a logical volume is written to the two physical volumes by the Logical Volume Managers device drivers. This is known as LVM mirroring. Mirrors can be split, and the data contained therein can be independently accessed.

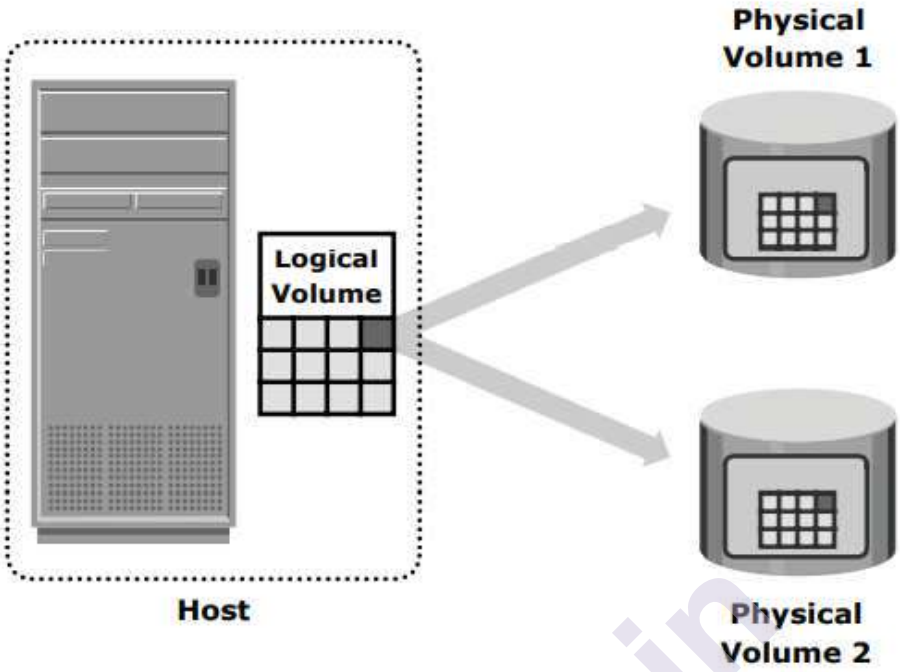



**Figure 10.4.1.1 LVM-Based Replication**

**10.4.1.2 Advantages of LVM-Based Replication**

LVM-Based Replication technology is not dependent on vendor specific storage system. It is a part of the operating system and does not require any additional license to deploy the LVM applications.

**10.4.1.2 Limitations of LVM-Based Replication**

An application generated write translates into two writes on the device. i.e. disk, because of these writes the additional burden on the host CPU will come and effect of this burden will decrease the application performance. There are two volume groups, we can use only one host group at any given time. To trace the changes to the mirrors is challenging in LVM. Performing incremental synchronization operation is challenging in LVM. Replica and the source both groups are stored on the same volume group, so replica itself may not available if there is an error in the both volume group.

If server fails, then both replica as well as source both volume groups are not available until the server will come back in online mode.

**10.4.1.4 File system snapshot**

File system snapshot is pointer-based replica. It requires a fraction of space used by the original file system. It can be used FS itself or for LVM. It uses the principle Copy on First Write (CoFW).While creating a snapshot, a block-map and a bitmap are created in the metadata on the snapshot file system. The use of bitmap is to keep track of blocks that are changed on the production file system after creation of the snapshot. The use of block-map is for addressing purpose where data is to be read when

the data is accessed from the snapshot file system. If the bit is 'zero' then the read operation is directed to the production file system. If the bit is 'one' then the block address is got from block-map and data is read from that address.

### 10.4.2 Local Replication with Storage Array-Based

Array operating environment performs the local replication process. Host resources (CPU, Memory) are not useful in the replication process. For Business operation an alternate host is used for replica which is useful for replica for accessing the data. Storage-array-based replication process, number of replica devices required, it must be selected on the same array and then data is replicated between source-replica pairs. Database is divided over multiple physical volumes and to ensure all replicated devices must be consistent PIT copy of the database.

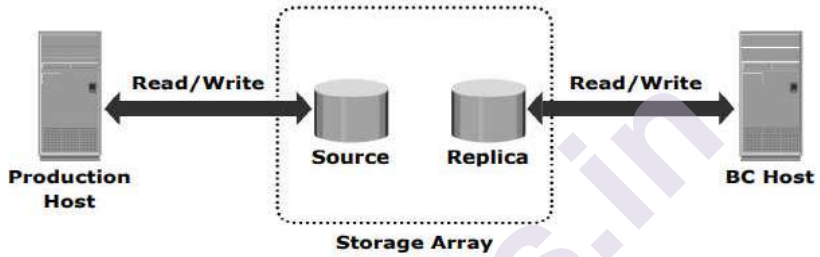



**Figure 10.4.2 Storage-array-based replication**

Figure 10.4.2 shows source and target storage array based local replication are in the same array and accessed by the different hosts. Storage-array-based local replication is get categorized in following ways.

- Full Volume Mirroring
- Pointer-Based Full Volume replication
- Pointer-Based virtual replication
- Replica devices is also known as target devices; which is accessed by Business operation Host.

#### 10.4.2.1 Full Volume Mirroring

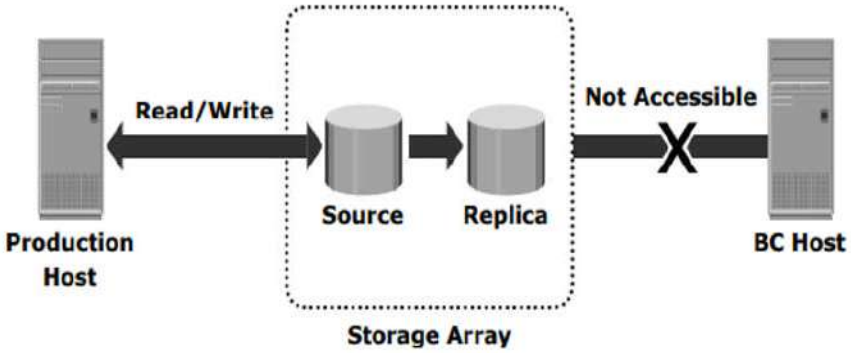



**Figure 10.4.2.1 (a): Full volume mirroring with source attached to replica**

The production host is attached to the source and then establishment of replica of the source is automatically generated. Data which is exist on source is get copied into target or replica. Source and the replica is on the same storage array. Whenever data is get updated on source it is automatically updated on target as well. Both source and target data is identical data. Target is also known as mirror of the source so it is known as replica. During synchronization, target is attached to the source, that time target is not available to any other host. Only production host can have accessed to the source.

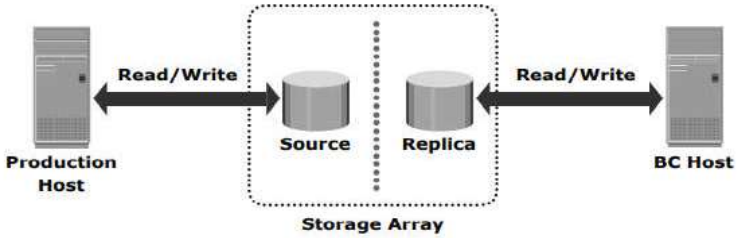



**Figure 10.4.2.1 (b) Full volume mirroring with source detached from replica**

After completion of the synchronization step, then the target detached from the source and is available for the Business continuity operations or any other host as well. The figure shows the full volume mirroring when the target is detached from the source. Source and target get accessed for the operations namely read and write by the production host. After splitting source and target it will be the Point-in-time(PIT) copy of the source. The source is detached from the target that Point-in-time (PIT) of a replica is determined by that time.

Example If the point-in-time for the target is 5.00 pm it means the detachment from source to target time is also 5.00 pm

Changes of source and replica can be tracked, after detachment of each point-in-time. In full volume mirroring, during synchronization process, the target is not accessible, till detachment from the source host. If database is large then it will take longer time.

### 10.4.2.2 Pointer-Based full Volume Replication

It is an alternative method to full volume mirroring where data is management is done by using pointers. Data is generated in bitmap for keeping track. It provides full copies of source data on the target. No need to wait for data synchronization to and detached of. Activation time is defines with the help of PIT copy of source.

Pointer-Based replication can be activated by different ways.

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

In both mode, the activation time created a protection bitmap for all data on the source devices.

Source is already having a data which is being mirror into a replica where target which is not being used.

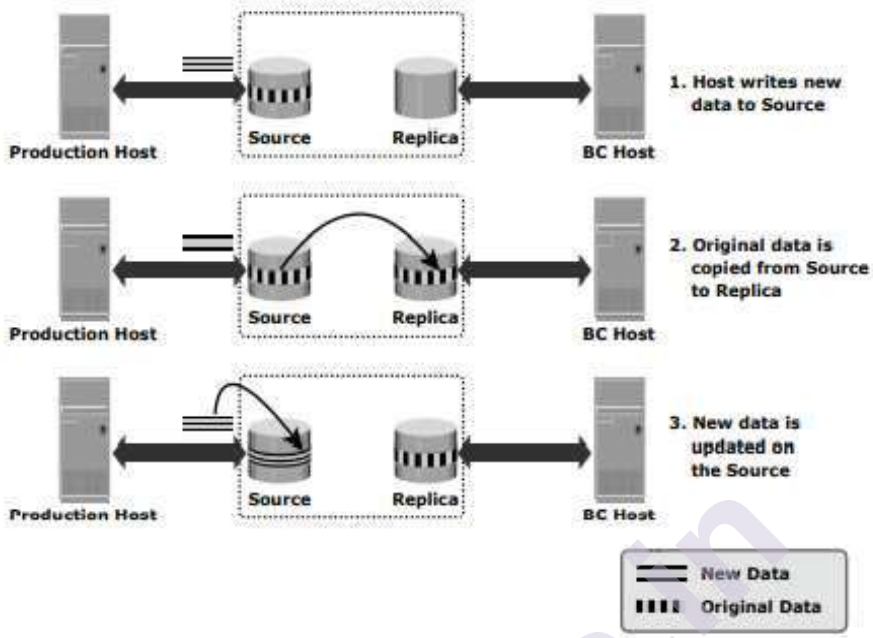



**Figure 10.4.2.2 Copy on first access (CoFA) – Write to source**

So original data is copied from source to replica. In case of new data is coming from source that time target is also update the data. So that full mirroring is done. While updating or coping data the source is waiting until all data is get copied from source to target.

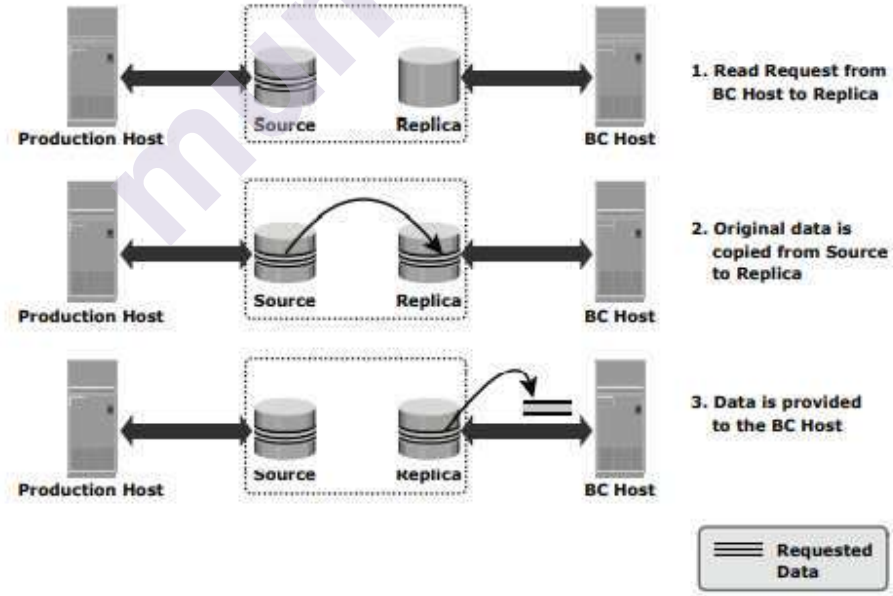



**Figure 10.4.2.2 Copy on first access (CoFA) – read from target**

Read operation request from the Business host to the target at first time access after the transaction activation the replication process started.

i.e. data is copied from the source to the target so that it will be available to host.

Write operation instructions given to the target that time the replication process started and the data is get copied from source to the target. After completion of this process, new data which is updated on the target get copied from source to replica.

In both read and write operation, protection bit for that block is reset. It shows the data has been copied from source to the target. Pointer to the data on the source side is get discarded.

Copy operation will not work during read and write operation on the same data block. so copy operation do not triggered so introduce the term Copy on First Access.

In case of replication Process termination, then the target device has all data that we can access till termination, that time not able to access entire data of the source at the point-in-time. Data on the target cannot be used as well as cannot be restore. It is not a full replication at source side.

Full copy mode, the name states all data is get copied from source to the target on backend side. If no need of accessing of data required for entire data is get copied block at the target side.

Entire data from the source is copied to the target in full copy mode. During process of replication termination, target side will have entire data from source side at the point-in-time of activation. This means the target is a responsible for recovery, restore and other business continuity operation.

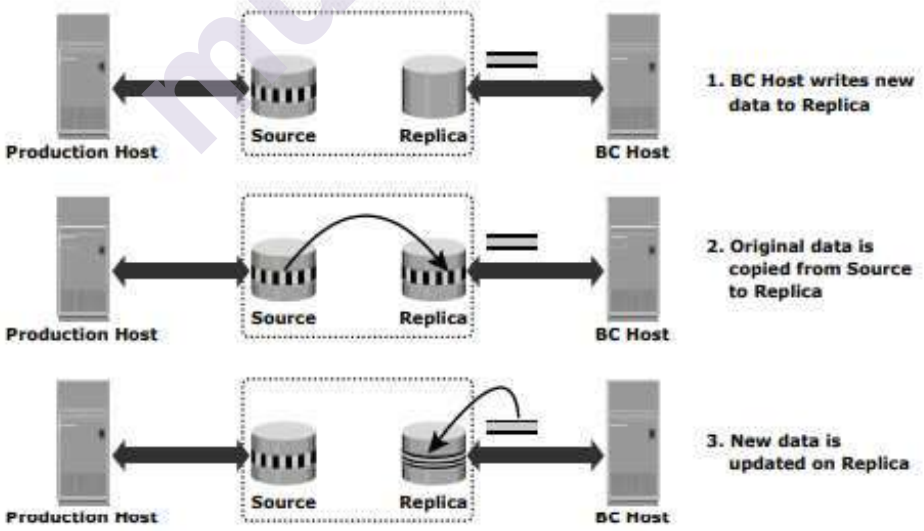



**Figure 10.4.2.2 Copy on First Access (CoFA) – write to target**

The main difference between pointer based full copy mode and full volume mirroring is that the target is accessible immediately after the activation of transaction in full copy mode. Opposite of that, one has to wait for the process of synchronization and detachment to access the target in full-volume mirroring.

**10.4.2.3 Full-Volume Replication**

Pointer-based full volume replication and full volume mirroring both technologies require the target devices as large as source devices.

In both technologies in full copy mode can provide incremental resynchronization or restore capacity.

**10.4.2.4 Pointer-Based Virtual Replication**

In case of pointer based virtual replication, the target contains pointer which is pointing to the location of data where data at the source side.

Target does not contain data. Target is also known as virtual replica.

Pointer based virtual replication, the data which is available at source device, protection bitmap is created. Granularity can range from 512 byte blocks to 64 KB blocks or more.

During write operation, that time data is get copied to the predefined area in the array. This area is also known as store location.

The pointer at the target side is updated that time address of data is get saved on location. After this process the next action of write is updated on the source side.

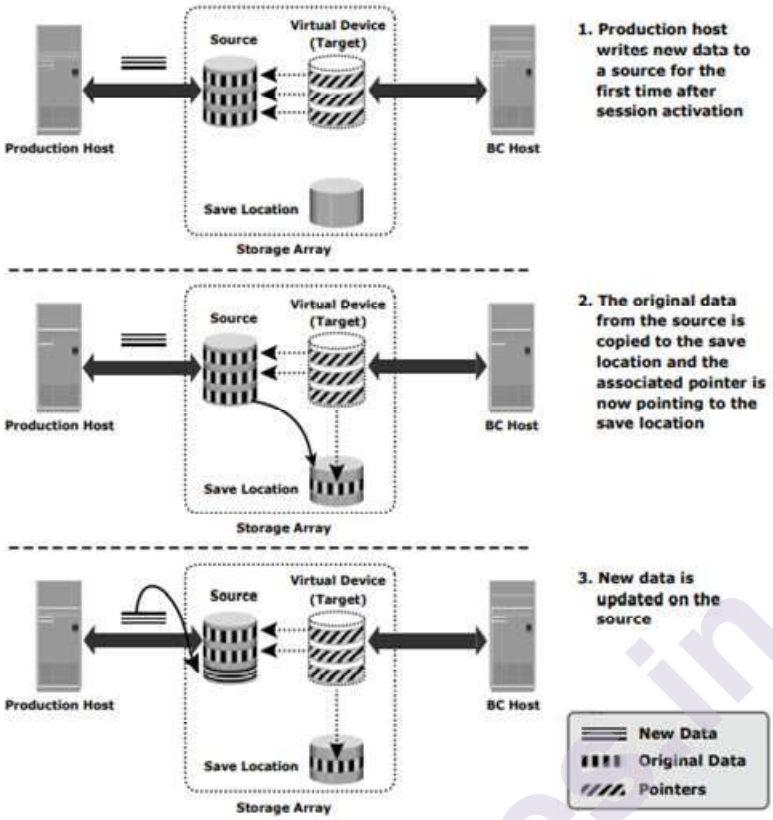



**Figure 10.4.2.4a) Pointer-Based virtual replication – write to source**

For write operation instructions issued to the target first time after the activation of session. Original data is get copied from source to the location data. Pointer is get updated at the location where data is saved. One more copy of the original data is get created at the location where data is get saved before the new write instruction is get updated on the location where data is get saved.

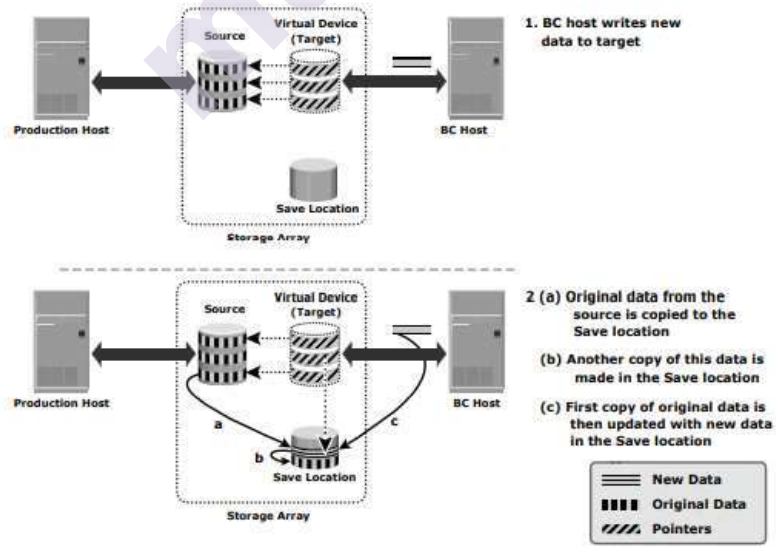



**Figure 10.4.2.4 b) Pointer-based virtual replication – write to target**

Read instruction issued to the target, data will not change the data blocks because of source read operation when the session activation. CoFW technology is useful for pointer based virtual replication. Write operation on the same data block is not trigger a copy operation on the same data block at source or target side. Combined data view at target side for unchanged data of the source side and the data which is saved at location side. Invalidates the data of the target side as there is no availability of source devices.

#### 10.4.2.5 Network-Based Virtual Replication

In network-based replication, the process of replication occurs at the network layer. Replication between two devices i.e. from server and the storage system. From server and the storage systems, network-based replication can work for a large volume of servers and the storage systems. It will work for different heterogeneous environments. Mostly used network-based replication technique is continuous data protection (CDP).

### 10.4.3 Continuous Data Protection (CDP)

Continuous data protection is a solution for network-based replication. It provides the capacity to restore the data at virtual machine platforms. It will not work like traditional data loss and recovery system, in traditional data protection, limited number of recovery of data occurs and in case of loss, the system can be rollback only to the last available recovery point.

Continuous Data Protection is opposite of transitional data loss and recovery. It tracks all the changes to the production volumes and maintains consistent point-in-time. CDP is useful for local as well as remote replication of data. Data can be replicated more than two sites using synchronous and asynchronous replication. CDP supports the duplication, compression (WAN optimization techniques) to reduce bandwidth requirements, and also optimal utilization of bandwidth.

#### 10.4.3.1 Local Replication Operations

**CDP Components**

1) Journal volume – It uses to store the data which has changed on the production volume at the time of replication process activated.

   Journal contains metadata and data with the operation rollback and recovery points.

   The space required for Journal is get configured by using how far back the recovery point can go.

2) CDP appliance – It is an intelligent hardware platform which work on CDP software and manages local and remote data replication.

3) Write splitter – It is useful to write the server or host and it splits each write into two copies. Write splitting work at host or storage system level.

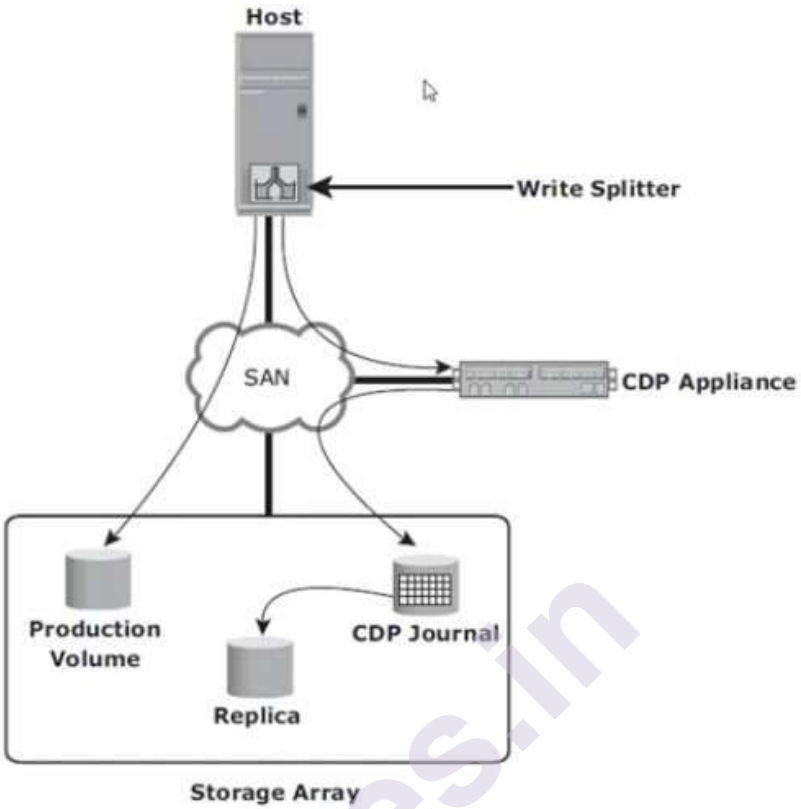
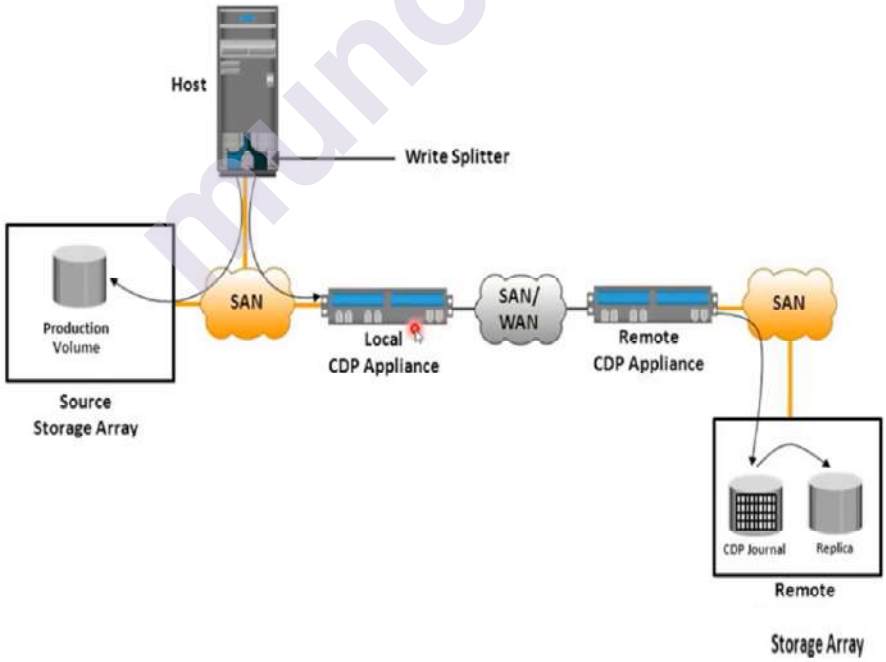



**Figure 10.4.3.1 Continuous Data Protection Local and remote replication operation**




CDP local and remote replication operations working model will use the different CDP components like write splitter for the use of deployment at the server level.

In CDP replication replica is synchronized with the source side, and then the process of replication is initialized. After the process of replication initialization, all the writes are get divided into two copies from the source i.e. production volume.

One copy is sent to the source site at local CDP appliance and the another one is sent to the production volume, after sending the copy the next step is at source site, local appliance writes the data to the journal and then data in turn is written to the local replica.

If the stored file is accidently corrupted or deleted then local journal enables to recover the application data at the point in time.

The local and remote replication operations are similar in network-based CDP replication.

**Comparison of Local Replication Technologies**

| Parameter | Full-volume mirroring | Pointer-based full-volume replication | Pointer-based virtual replication |
|---|---|---|---|
| Performance impact on source side | No impact | CoFA mode- Some impact<br>Full copy – no impact | Very high impact |
| Size of target | At least the same as the source side | At least the same as the source side | Small fraction of the source side. |
| Accessibility of source for restoration | Not required | CoFA mode- required<br>Full copy- not required | Required |
| Accessibility to target | Only after synchronization and detachment from the source | Immediately accessible | Immediately accessible |



**10.4.3.2 Tracking changes to source and Replica restore**

The main aim of local replication is point-in-time copy for data recovery and restore operations, during recovery and restore operations the target must be updated.

At the time of replica creation, the bitmap is created at block of data. One bit per block of data. Bit of the source and target is set to ‘zero’.

Updating at the source or target side then flagged by setting the appropriate bit to ‘one’ in the bitmap block data.

Different operations like resynchronization and restore requires then the source bitmap and target bitmap is operated using the Logical OR operation.

All bitmap block data is also get modified either at source or target side. It enables the copy of replica of all blocks between the source and target side. Data movement is depend on the resynchronization or the restore operations.

If resynchronization, then changes to the target are overwritten with the corresponding blocks data from the source. In the given diagram block no 3,4 and 8 on the target from the left side.

If restore, then updated data to the source are overwritten with the corresponding blocks data from the target. In the given diagram block 1,4 and 6 on the source.

In both the operations (resynchronization or restore) changes of data to the source side or the target side cannot be simultaneously happened.

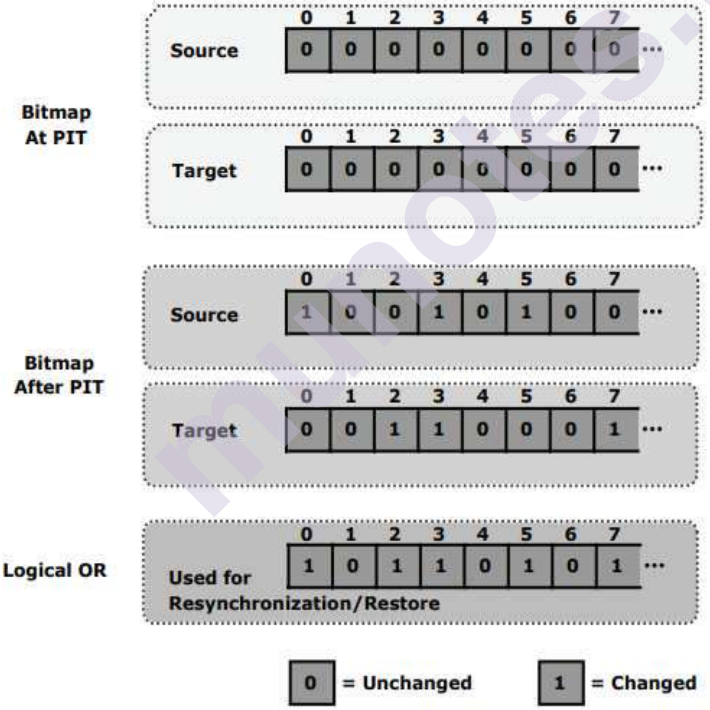



**Figure 10.4.3.2 Tracking changes**

### 10.4.3.3 Creating multiple Replicas

In Storage-array-based replication Process source devices are used to maintain replication relationships with multiple targets. Changes at source side made first and then each target side can be tracked. Incremental resynchronization of the target enabled. Point-in-time copy

can be used for various Business Continuity operations and at the time of restore operation.

Figure shows an example where copy is created every six hours from the same source.

If the data on the source side is get loss or corrupted then the data can be restore from the latest Point-in-Time copy.

Storage-Array-based replication process also enables the creation of multiple concurrent Point-in-time replicas. All replicas will contain the same data. One or more of the replicas can be set aside for the restore and recovery operations. Other replicas are used for the decision support activities.

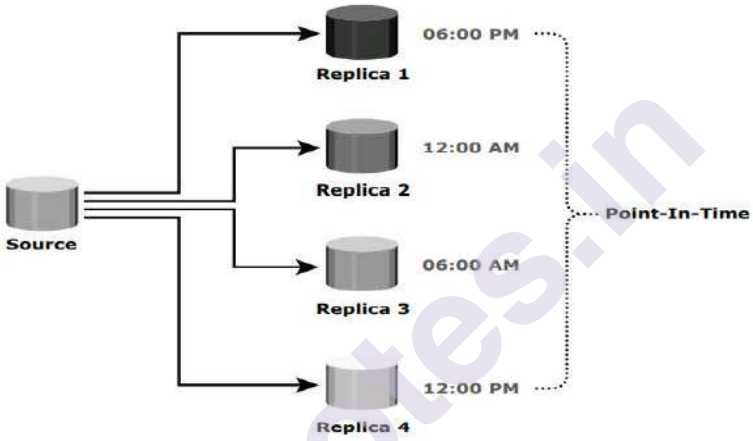



**Figure 10.4.3.3 Multiple replicas created at different point-in-time**

## 10.5 SUMMARY

This chapter gives detailed study about local replication which contains local replication terminology, uses of local replication, Replica consistency, Consistency of a replicated file systems and database. Host based local replication, LVM based replication its advantages as well as limitations. Storage array-based local replication with full-volume mirroring, pointer-based, full-volume replication, Pointer-based virtual replication, network-based local replication. Continuous data protection, CDP, local replication operation, tracking changes to source and replica, creating multiple replicas.

## 10.6 EXERCISES

1) Discuss Local replication technologies in detail.
2) What is the importance of recoverability and consistency in local replication?

3) What are the considerations for performing backup from a local replica?

4) Discuss about multiple replica

5) What are the uses of a local replica in various business operations.

## 10.7 REFERENCES:


Information storage and management: storing, managing and protecting digital information in Classic, Virtualized and Cloud Environments, EMC author, by Joh Wiley and Sons 2nd edition 2012.

https://books.google.co.in/books?id=PU7gkW9ArxIC&printsec=frontcover&dq=information+storage+and+management&hl=en&newbks=1&newbks_redir=1&sa=X&ved=2ahUKEwjjx_nakNPxAhWy4zgGHWUpCjcQ6AEwAHoECAsQAg

https://books.google.co.in/books?id=sCCfRAj3aCgC&printsec=frontcover&dq=information+storage+and+management&hl=en&newbks=1&newbks_redir=1&sa=X&ved=2ahUKEwjjx_nakNPxAhWy4zgGHWUpCjcQ6AEwAXoECAIQAg

**https://www.youtube.com/watch?v=H0T6a2ok6zw**


❖ ❖ ❖ ❖

# 11

# REMOTE REPLICATION

**Unit Structure**





## 11.0 OBJECTIVES:

This chapter deals with remote replication processes of creating replicas with reference to the remote locations. To study the remote replication may be either synchronous or asynchronous, Replication occurs occur in the three different places namely host or server, storage array, or in the network.

## 11.1 INTRODUCTION:

The process in which creating replicas of information assets at remote location is called as Remote replication. Organizations mitigate the risk related to regionally driven outages resulting from human-made or natural disasters using remote replicas. It can also be used in business operations like that of local replicas. The source is the infrastructure where the information assets are stored at primary site whereas target is referred to the infrastructure where the replica is stored at the remote site. Source hosts or target hosts are the hosts that access the source or target respectively. In this chapter we will study about various remote replication technologies, with the important steps to plan and design proper remote replication solutions. Also, this chapter describes network requirements and management considerations in the remote replication process.

Concepts-

1) Synchronous and Asynchronous Replication
2) LVM Based Replication
3) Host based Log Shipping
4) Disk-Buffered Replication
5) Three-Site Replication

## 11.1.1 MODES OF REMOTE REPLICATION

1) Synchronous
2) Asynchronous

**Synchronous Replication –**

The process of copying data over a local area network (LAN) or storage area network (SAN) or wide area network (WAN) so that there are many copies of the data. It writes data to the primary and secondary sites at the same time so that data remains current between the sites. Writes must be executed by the source and the target, before declaring "write complete" to the host. Until each preceding write has been completed and acknowledged, additional writes on the source cannot occur which also ensures data replicates all time and is identical on the source. After this the writes are sent to the remote locations exactly in the same order in which they were received by the source. Thus, write order is always maintained. If there is a failure of the source site, it provides zero or nonzero RPO and lowest RTO.

But with any kind of synchronous remote application, the application response increases. The distance between the sites, available bandwidth, and infrastructure of network connectivity decides the degree of impact on the response time. The distance over which the synchronous replication can be deployed depends on application's ability to tolerate extension in response time. Mostly, it is deployed for range of less than 200 KM (125miles) between two sites.

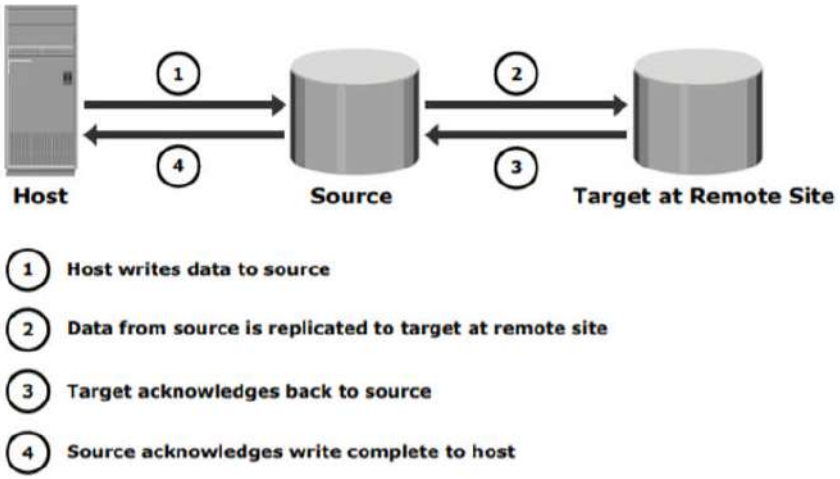



**Figure 11.1 Synchronous Replication**

## 11.2 REMOTE REPLICATION TECHNOLOGY

Remote replication of data can be organised by the storage arrays or the hosts. There are a few other options which include special appliances to replicate data over the SAN or the LAN, and replication on storage arrays over SAN.

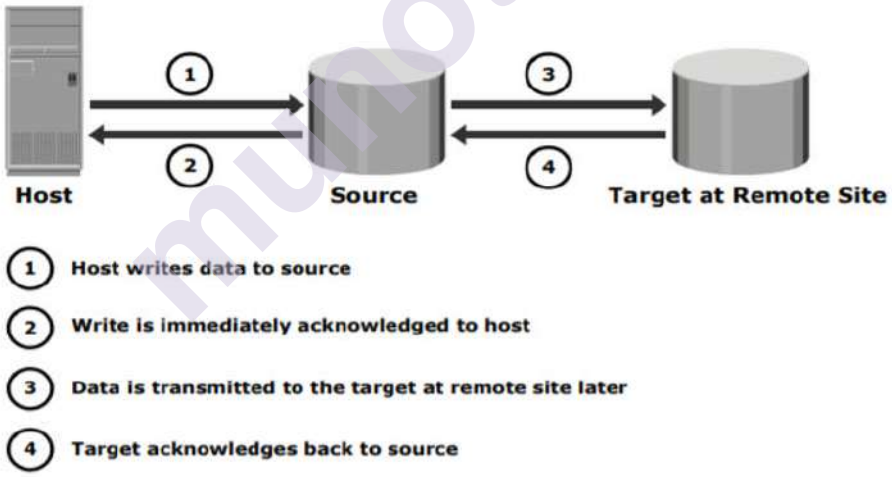



**Figure 11.2 Asynchronous Replication**

### 11.2.1 Remote Replication with Host-Based

In this type of remote replication, it uses one or many components of the host to manage and perform the operations. There are two basic ways to advance towards host-based remote replication : Database replication ( via log shipping) and LVM based replication.

### 11.2.2 Remote Replication with LVM Based

This type of replication is managed and performed at volume group levels. Firstly it writes to the source volumes and then transmits to the remote host by LVM. After this the LVM on the remote host recieves , writes and then commits them to the remote volume groups.

Before the starting of the application , identical volume groups, logical volumes and files systems are created at source and target sites. Initially the coordination of the data between source and replica can be performed in many different ways. One of them is to have backup the source data to tape and then restore the data to the remote replica. Another option is , it can be performed by replicating over IP network. Upto completion of the initial synchronization, production work on the source volumes is naturally paused. After synchronization, production work can be resumed on the source volumes and replication of the data can be achieved on the existing standard IP network.

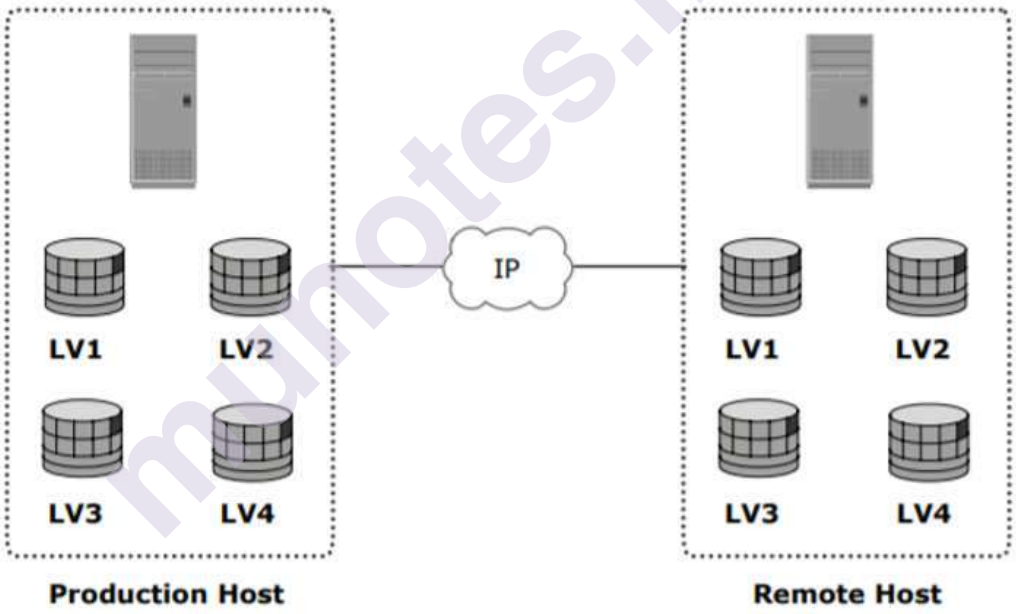



**Figure 11.2.2 LVM-based Remote Replication**

Both synchronous and asynchronous modes of data transfer are supported by LVM-based remote replication. In asynchronous mode, writes are line up in a log file at the source and then sent to the remote host in respective order in which they were received. The size of log file regulates the RPO at the remote site. In case of network failure, writes continue to gather in the log file. In case if the log file gets filled before the failure is determined, then a full resynchronization is required upon network availability. In case of a failure at source site, using the data on the remote replicas, applications can be started again on the remote host.

LVM-based remote replication removes the need for SAN infrastructure. It is independent of the storage arrays and types of disks at remote sites and source. Most of the operating systems are shipped with LVMs, so that supplementary licenses and specific hardware are not required.

The replication process makes an addition overhead on the host CPUs. The CPU resources which are there on the source host are shared between replication tasks and applications, may cause performance deprivation of the application.

As remote host is involved in the replication process, it has to be uninterruptedly active and available. Particularly in the case of applications using federated databases, LVM-based remote replication is not accurate to the mark.

## 11.3 HOST-BASED LOG SHIPPING

Database replication via log shipping which is supported by most of the databases is a host-based replication technology. Inside logs transactions to the source database are captured, and then periodically sent by the source host to the remote host. The logs are then received by the remote host and applied to the remote database.

Before starting the production work and replicating of log files, all appropriate component of the source database are replicated to remote site. The process is done when the source database is shut down.

In the next step, the production work is started on source database. The remote database is started on a standby mode. The database is not available for transactions in standby mode. Few executions allow reading and writing from standby database.

All DBMS's switch log files are configured before time intervals, or when a log file is saturated. The present log file is shut down at the time of log switching and a new log file is opened. When there is a log switch, the log which was shut down gets sent from source host to remote host. The remote host receive the log and updates standby database.

The procedure ensures that the standby database is reliable till the end of the log. RPO at the remote site is limited and rest on on the size of the log and occurrence of log switching. Provided that the network bandwidth, latency and rate of update available, and frequency of log switching should be considered when defining the best size of the log file.

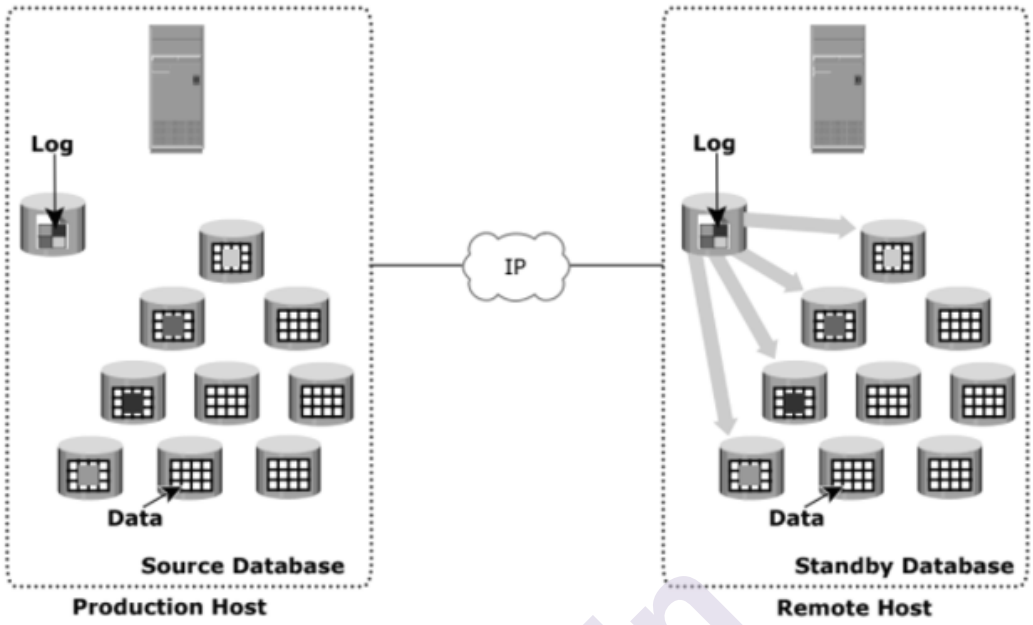



**Figure 11.3 Host-Based Log Shipping**

As the source host doesn't provide each and every update and buffer them, this alleviates the burden on the source host CPU. The existing standard IP network, same as that of LVM-Based remote replication can be used for replicating log files. Host-based log shipping does not rule accurate, mainly in the case of applications using federated databases.

## 11.4 STORAGE ARRAY BASED REMOTE REPLICATION

In storage array-based remote replication, the array operating environment and resources perform and oversee information replication. This soothes the weight on the host computer chips, which can be better used for running an application. A source and its imitation gadget dwell on various capacity exhibits. In different executions, the capacity regulator is utilized for both the host and replication responsibility. Information can be communicated from the source stockpiling cluster to the objective stockpiling exhibit over a common or a devoted organization.

Replication between clusters might be acted in coordinated, asynchronous, or circle cushioned modes. Three-site far off replication can be carried out utilizing a blend of coordinated mode and offbeat mode, just as a mix of simultaneous mode and circle cradled mode.

### 11.4.1 Remote Replication with Synchronous Replication Mode

In array based synchronous remote replication, composes should be focused on the source and the objective before recognizing "compose total" to the host. Extra composes on that source can't happen until each

previous compose has been finished and recognized. The cluster based coordinated replication measure is displayed in Figure 11.4.1.

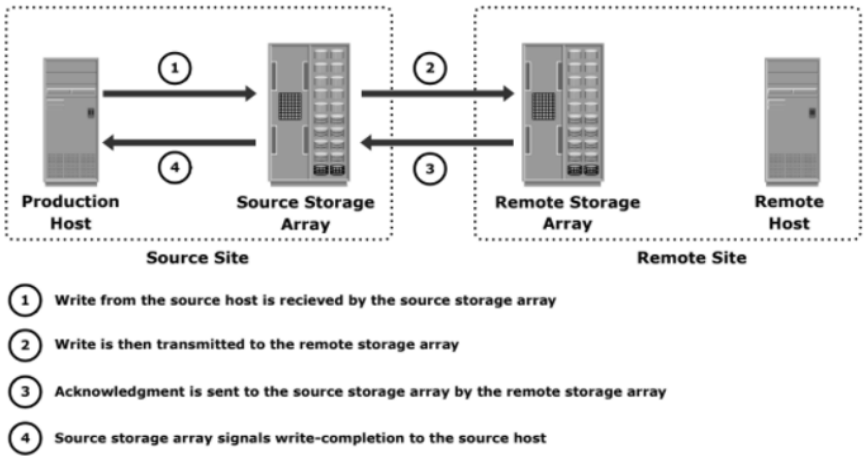



**Figure 11.4.1 a) Array-based synchronous remote replication**

On account of coordinated replication, to upgrade the replication cycle and to limit the effect on application reaction time, the compose is put on store of the two clusters. The savvy stockpiling exhibits can de-stage these keeps in touch with the fitting plates later.

On the off chance that the replication joins fizzle, replication is suspended; in any case, creation work can proceed with continuous on the source stockpiling cluster. The cluster working climate can monitor the composes that are not sent to the far-off capacity exhibit. At the point when the organization joins are re-established, the gathered information can be communicated to the far-off capacity exhibit. During the hour of organization interface blackout, if there is a disappointment at the source site, some info will be disappeared and the RPO at target will be non-zero. For Synchronous remote replication, network bandwidth equivalent to or larger than maximum written workload between the sites should be provided every time. Figure 14-6 demonstrates the write workload (expressed in MB/s) overtime. The “Max” line indicated in Figure 14-6 demonstrates the required bandwidth that must be provided for synchronous replication. Bandwidths less than max write workload result in an intolerable upsurge in application response time.

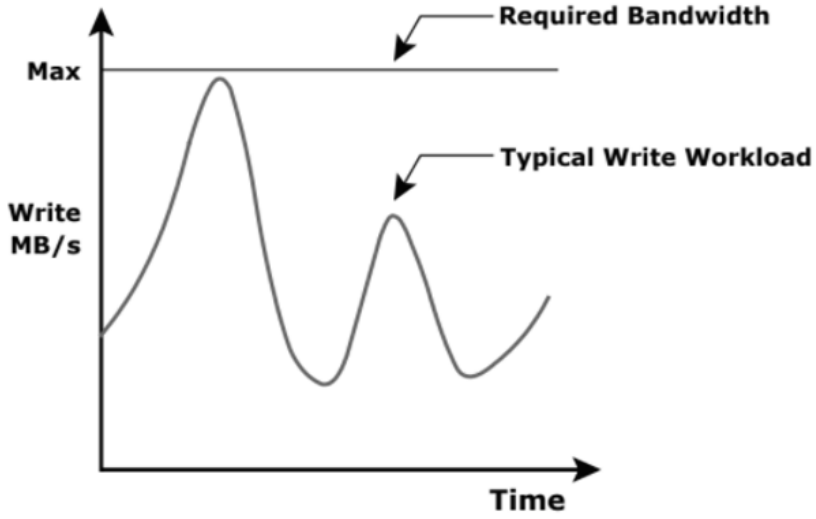



**Figure 11.4.1 b) Network bandwidth requirement for synchronous replication**

### 11.4.2 Remote Replication with Asynchronous Replication Mode

In exhibit based offbeat far off replication mode, displayed in Figure 11.4.2, a compose is focused on the source and quickly recognized to the host. Information is supported at the source and communicated to the distant site later. The source and the objective gadgets don't contain indistinguishable information consistently. The information on the objective gadget is behind that of the source, so the RPO for this situation isn't zero.

Like coordinated replication, asynchronous replication composes are set in store on the two exhibits and are later de-arranged to the suitable circles.

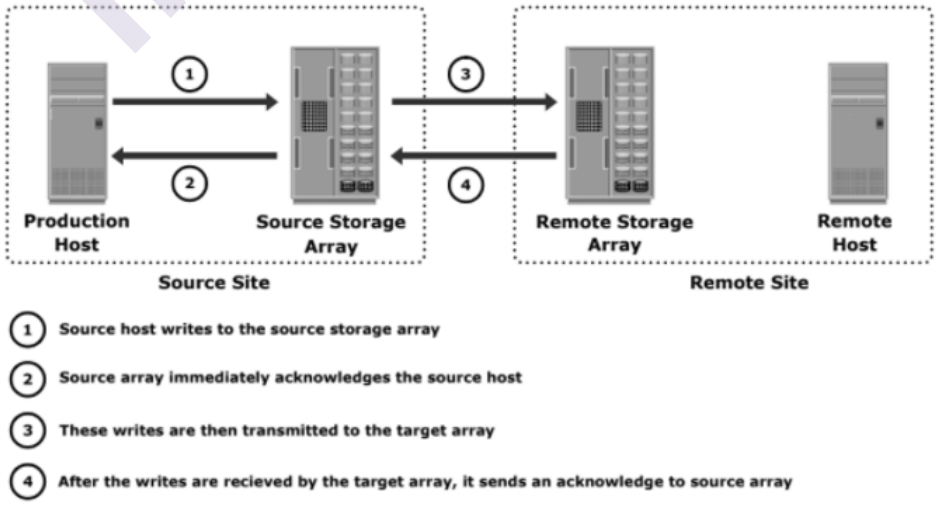



**Figure 11.4.2 a) Array-based asynchronous remote replication**

A few executions of asynchronous far-off replication keep up with compose requesting. A period stamp and succession number are appended to each compose when it is gotten by the source. Composes are then communicated to the far-off exhibit, where they are focused on the far-off copy in the specific request in which they were buffered at the source. This certainly ensures consistency of information on the far-off imitations. Different executions guarantee consistency by utilizing the ward compose standard inborn to most DBMSS. The composes are supported for a predefined timeframe. Toward the finish of this length, the support is shut, and another cradle is opened for ensuing composes. All writes in the shut cushion are sent together and focused on the far-off copy.

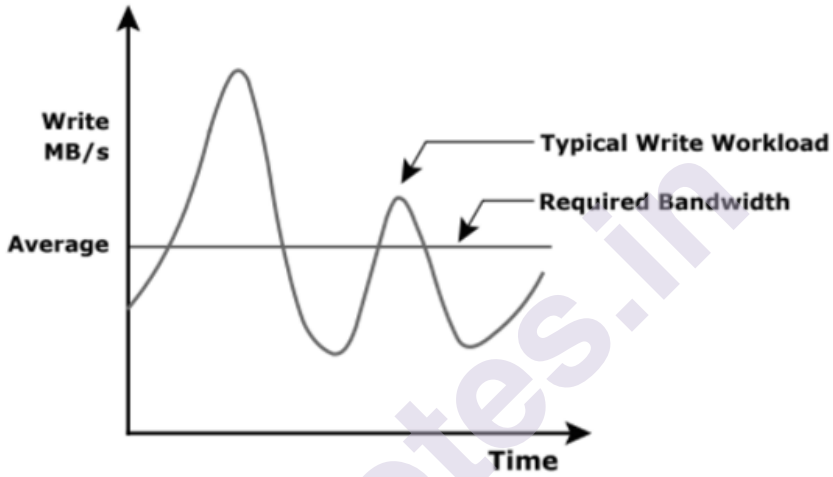



**Figure 11.4.2 b) Network bandwidth requirement for asynchronous replication**

Asynchronous remote replication gives network data transfer capacity cost savings, as just transfer speed equivalent to or more noteworthy than the normal compose responsibility is required, as addressed by the "Normal" line in Figure 14-8. During times when the compose responsibility surpasses the normal transmission capacity, adequate cushion space must be arranged on the source stockpiling exhibit to hold these composes.

### 11.4.3 Disk-Buffered Replication Mode

Disk-buffered replication is a mix of nearby and far off replication technology. A predictable PIT neighbourhood copy of the source gadget is first made. This is then repeated to a distant reproduction on the objective exhibit.

The grouping of activities in a circle cradled distant replication is displayed in Figure 11.4.3. Toward the start of the cycle, the organization joins between the two clusters are suspended and there is no transmission of information. While creation application is running on the source gadget, a reliable PIT neighbourhood copy of the source gadget is made. The

organization joins are empowered, and information on the nearby reproduction in the source exhibit is communicated to its distant imitation in the objective cluster. After synchronization of this pair, the organization connect is suspended and the following nearby copy of the source is made. Alternatively, a neighbourhood PIT reproduction of the distant gadget on the objective exhibit can be made. The recurrence of this pattern of tasks relies upon accessible connection transmission capacity and the information change rate on the source gadget.

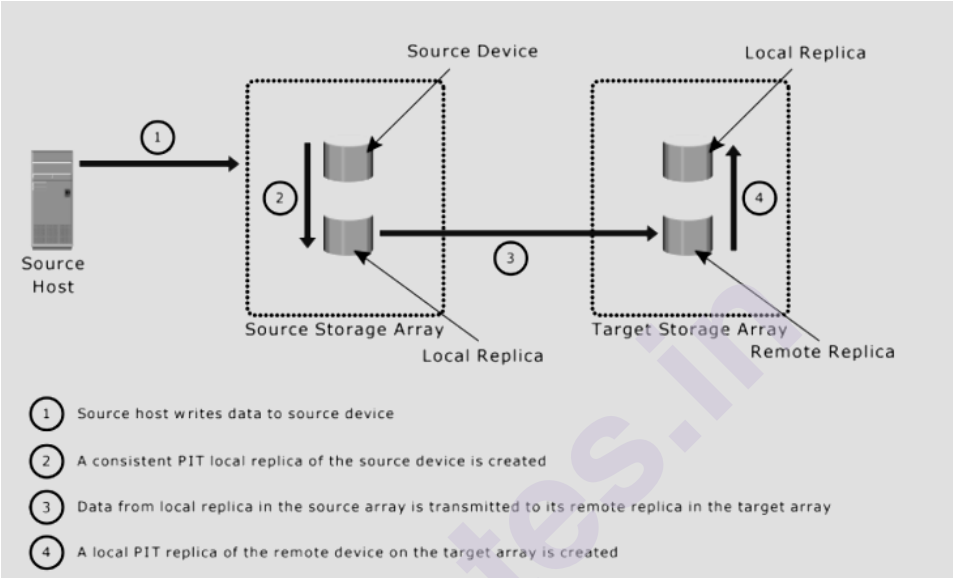



**Figure 11.4.3 Disk-buffered remote replication**

Exhibit based replication innovations can follow changes made to the source and target gadgets. Thus, all resynchronization activities should be possible gradually.

For instance, a nearby imitation of the source gadget is made at 10:00 AM and this information is sent to the far-off copy, which requires one hour to finish. Changes made to the source gadget after 10:00 AM are followed. Another imitation of the source gadget is made at 11:00 AM by applying track changes between the source and neighbourhood reproduction (10:00 AM duplicate). During the following pattern of transmission (11:00 AM information), the source information has moved to 12:00 PM The neighbourhood copy in the far-off cluster has the 10:00 AM information until the 11:00 AM information is effectively sent to the distant imitation. In the event that there is a disappointment at the source site preceding the fulfilment of transmission, then, at that point the most pessimistic scenario RPO at the far-off site would be two hours (as the distant site has 10:00 AM information).

### 11.4.4 Network Based Remote Replication Mode

For the Network Based Remote replication mode, the replication occurs at the network layer between the server and the storage systems. By divesting replication from server and storage systems, network-based

replication can work across a huge number of server platforms and storage systems, making it perfect for extremely diverse surroundings. One of the most widely used Network based replication technique is the Continuous Data Protection (CDP).

### 11.4.5 CDP Remote Replication

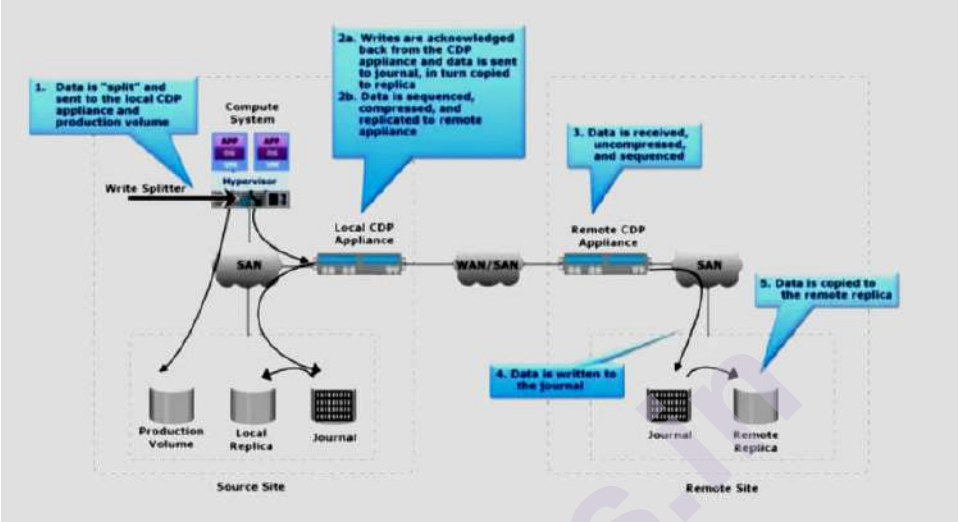



**Figure 11.4.5 CDP Remote Replication**

Continuous data protection (CDP) is a network-based replication key which delivers the ability to restore data and VMs to any preceding PIT. Traditional data protection technologies offer a restricted number of retrieval points. Suppose there is a data loss, system can be moved back to the preceding accessible retrieval point. However, CDP paths all the variations to the production volumes and regulates constant point-in-time images.

This is how the CDP is made to restore data to any previous PIT. CDP is supported by both the local and the remote replication of data and VM to meet functioning and adversity recovery respectively. In CDP application, data can be replicated to additional sites using synchronous and asynchronous replication. CDP chains various WAN optimization techniques (deduplication, compression) to decrease bandwidth necessities, and also optimally uses the accessible bandwidth.

## 11.5 THREE-SITE REPLICATION

In synchronous and asynchronous replication, below usual circumstances the load is successively at the source site. Processes at the source site will notbe interrupted by any disappointment to the target site or to the network used for replication. The replication process resumes as soon as the link or target site issues are resolved. The source site endures to function deprived of any remote protection. If disappointment happens at the source site through this time, RPO will be extended.

In synchronous replication, source and target sites are typically within 200KM (125 miles) of to each other. Henceforth, in the occurrence of a regional disaster, both the source and the target sites could turn out to be inaccessible. This leads to extended RPO and RTO because the past recognized decent copy of data would have to come from alternative source, such as offsite tape library. Regional disaster will not disturb the target site in asynchronous replication, as the sites are naturally kilometres apart.

If the source site be unsuccessful, production can be transferred to the target site, but there will be no remote protection till the let-down is determined.

Three-site replication is used to ease the dangers recognized in two-site replication. In three-site replication, data from the source site is replicated to two remote data centres. Replication can be synchronic to one of the two data centres, providing a zero RPO solution. It can be asynchronous or disk buffered to the other remote data centre, providing a limited RPO. Three-site remote replication can be applied as a cascade/multi-hop or a triangle/multi-target solution.

### 11.5.1 Three-Site Replication-Cascade/Multi-Hop

In the cascade/multi-hop mode of replication, data streams from source to transitional storage array, known as a bunker, in the initial hop and then from a bunker to a storage array at a remote site in the next hop. Replication amongst the source and the bunker occur synchronously, but replication between bunker and remote site can be attained in two ways: disk-buffered mode or asynchronous mode.

### 11.5.2 Three-Site Replication- Synchronous+ Asynchronous

This method services a combination of synchronous and asynchronous remote replication technologies. Synchronous replication happens between the source and the bunker. Asynchronous replication occurs between the bunker and the remote site. The remote replica inside the bunker performs as the source for the asynchronous replication to generate remote replica at the remote site. Figure 11.5.2 a) demonstrates the synchronous + asynchronous method.

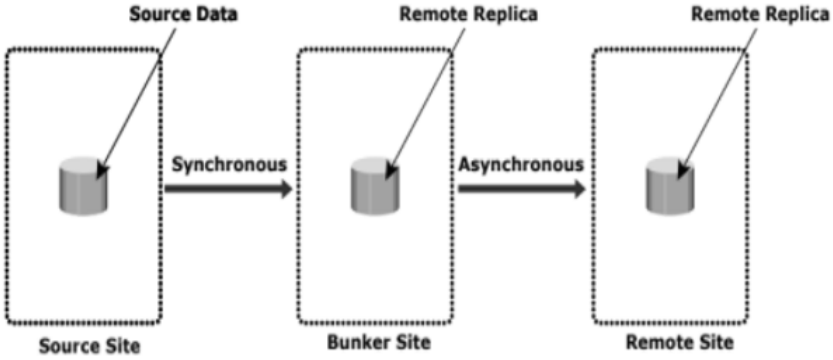



**Figure 11.5.2 a) Three-site replication with synchronous + asynchronous**

RPO on the remote site is generally on the instruction of minutes in this application. In this method, at least of three storage devices are mandatory (including the source) to replicate one storage device. The devices comprising a synchronous remote replica at the bunker and the asynchronous replica at the remote are the additional two devices.

Suppose if there is tragedy at the source, processes are unsuccessful over to the bunker site with zero or near zero data damage. Unlike the synchronous two-site situation, there is still remote protection available at the third site. The RPO among the bunker and third site can be on the order of minutes.

If there is a tragedy at the bunker site or if there is a network link disappointment between the source and bunker sites, the source site will remain to operate as normal but deprived of any remote replication. This situation is very similar to two-site replication when a disappointment/tragedy occurs at the target site. The apprises to the remote site cannot occur due to the disappointment in the bunker site. Hence, the data at the remote site keeps dropping behind, but the advantage here is that if the source miss the mark during this time, operations can be continued at the remote site. RPO at the remote site rest on on the time difference between the bunker site disappointment and source site disappointment.

A regional disaster in three-site cascade/multi-hop replication is identical to a source site disappointment in two-site asynchronous replication. Operations will failover to the remote site with an RPO on the order of minutes. There is no remote protection till the regional tragedy is determined. Local replication technologies could be used at the remote site during this time.

If a disaster occurs at the remote site, or if the network links between the bunker and the remote site be unsuccessful, the source site remains to work as normal with disaster recovery protection provided at the bunker site.

### 11.5.2 Three-Site Replication- Synchronous + Disc Buffered

This method services a mixture of local and remote replication technologies.

Synchronous replication occurs between the source and the bunker: A constant PIT local replica is formed at the bunker. Data is conducted from the local replica at the bunker to the remote replica at the remote site. A local replica can be formed at the remote site after data is established from the bunker. Figure 11.5.2 b) illustrates the synchronous + disk buffered method.

In this technique, atleast four storage devices are mandatory (including the source) to replicate one storage device. The remaining three devices are the synchronous remote replica at the bunker, a steady PIT local replica at the bunker, and the replica at the remote site. RPO at the remote site is regularly in the order of hours in this application. For example, if a local replica is created at 10:00 am at the bunker and it takes an hour to conduct this data to the remote site, variations made to the remote replica at the bunker since 10:00 am are traced. Therefore only one hour's worth of data has to be resynchronized between the bunker and the remote site during the subsequent cycle. RPO in this case will also be two hours, alike to disk-buffered replication.

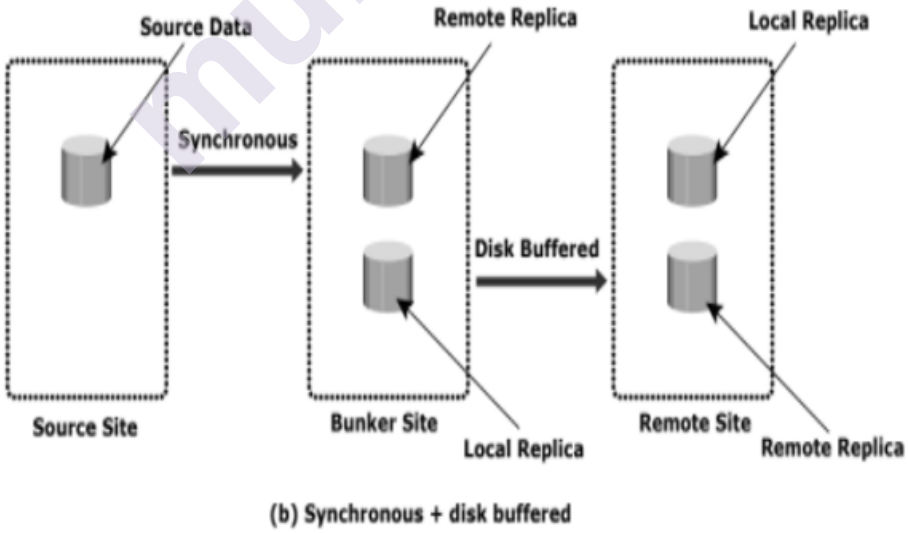



**Figure 11.5.2 b) Three-site replication with synchronous + disk buffered**

The procedure of making the consistent PIT replica at the bunker and incrementally informing the remote replica and the local replica at the remote site occurs uninterruptedly in a cycle. This procedure can be automatic and controlled from the source.

### 11.5.3 Three-Site Replication- Triangle/Multitarget

In this three-site triangle/multi-target replication, data at the source storage array is simultaneously replicated to two dissimilar arrays. 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 of minutes. The distance between the source and the remote site could be miles away. This type of configuration does not depend on the bunker site for informing data on the remote site, because data is asynchronously derivative to the remote site directly from the source.

The important benefit of three-site triangle/multi-target replication is the capability to failover to whichever of the two remote sites in the case of source site disappointment, with tragedy recovery (asynchronous) protection between them. Resynchronization between the two persisting target sites is incremental. Tragedy retrieval protection is always available in the occurrence of any one site disappointment.

During normal operations all three sites are accessible and the load is at the source site. At any given prompt, the data at the bunker and the source is alike.

The data at the remote site is overdue data at the source and the bunker. The replication network links between the bunker and remote sites will be in place but not in practice. Thus, during normal operations there is no data drive between the bunker and remote arrays. The alteration in the data between the bunker and remote sites are tracked, so that in the event of a source site tragedy, operations can be continued at the bunker or the remote sites with incremental resynchronization between the sites.

### 11.5.4 Data Migration Solution

Specialized replication techniques are Data migration and mobility solution, it is useful to enable or creating remote PIT copies. These copies help for data migration, mobility, content distribution and disaster recovery. Data migration solution gives the result of moves data between various heterogeneous storage arrays. Data may move from one array to the other array over SAN or WAN. This technology is useful for application and server operating system which is independently because the replication operations are performed by any one storage array.

Data mobility refers to moving data between various heterogeneous storage arrays for performance, cost or other reason. In push operation data move from the array (control) to the remote array, that time control devices are act like a source, while remote device is target. In pull operation, data is moved from the remote array to the control array to

the remote device, that time remote device act like the source, and the control device is the target.

Data migration solutions perform push and pull operations for data movement.

Remote replication and migration in a virtualized environment are the data migration solution. Virtual migration is a process where movement of virtual machines from one hypervisor to another virtual machines without power cutting at the virtual machines.

### 11.5.5 Remoter replication and migration in a virtualized environment

In a virtualized environment, all Virtual Machine data and VM configuration file residing on the storage array of the primary site are replicated to the storage array at the remote site. This process is visualized to the Virtual Machines. The Logical unit numbers are get replicated between the two sites using the storage array replication methodology. It may be either synchronous for limited distance, near to zero RPO or asynchronous for more distance, non-zero RPO.

Virtual Machine migration is a technique used for ensuring Business Continuity for the hypervisor failure or for maintenance which is scheduled. It is also useful for load balancing at the time of multiple virtual machines are running on the same hypervisor.

There are two commonly used techniques for virtual migration namely hypervisor-to-hypervisor and array-to-array migration

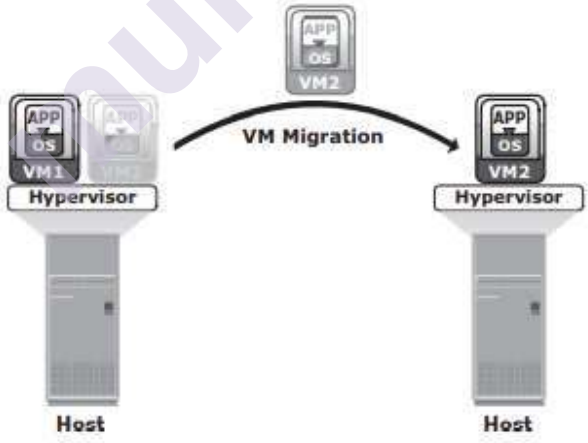



**Figure 11.5.5 a) Hypervisor-to-hypervisor VM migration**

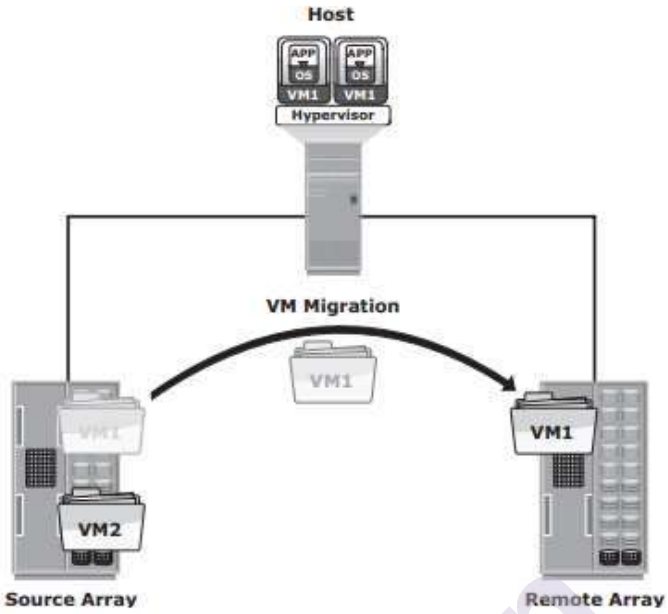



**Figure 11.5.5 b) Array-to array VM migration**

## 11.6 SUMMARY

This chapter gives the detail study of remote replication. Remote replication enables various business operations to the rapidly restarted at a remote site. Replica of the source data as well as target site which is useful for backup and testing purpose. It is also helping for data repurposing. Examples are report generation, decision support, and data warehousing. By using remote replication ensuring the improvements in production performance at the source.

Remote replication is also useful for data center migrations, without much disturbance to the production operations because the applications accessing where the source data is not at all affected.

Chapter also describes the different types of remote replication solutions: the distance between the primary site and the remote site which replication technology solutions for deploy with synchronous and asynchronous replications.

## 11.7 EXERCISES

1) What is remote replication? Explain in detail modes of remote replication.
2) Discuss about remote replication technologies
3) Give the difference between synchronous replication mode and asynchronous replication mode.

4) Discuss the effects of a bunker failure in a three-site replication for the following implementation:
   a) Multidrop - synchronous + disk buffered
   b) Multidrop – synchronous + asynchronous
   c) Multi-target
5) Discuss the remote replication and migration in a virtualized environment solution.

## 11.8 REFERENCES:

Information storage and management: storing, managing and protecting digital information in Classic, Virtualized and Cloud Environments, EMC author, by Joh Wiley and Sons 2nd edition 2012.
https://books.google.co.in/books?id=PU7gkW9ArxIC&printsec=frontcover&dq=information+storage+and+management&hl=en&newbks=1&newbks_redir=1&sa=X&ved=2ahUKEwjjx_nakNPxAhWy4zgGHWUpCjcQ6AEwAHoECAsQAg
https://books.google.co.in/books?id=sCCfRAj3aCgC&printsec=frontcover&dq=information+storage+and+management&hl=en&newbks=1&newbks_redir=1&sa=X&ved=2ahUKEwjjx_nakNPxAhWy4zgGHWUpCjcQ6AEwAXoECAIQAg

# 12

# CLOUD COMPUTING

**Unit Structure**





## 12.0 OBJECTIVES

Cloud computing is a model for on demand network access with the help of computing resources like networks, storage, servers,

applications and services, which can be use with various services with minimal economy scale.

## 12.1 INTRODUCTION

Cloud computing is the services provided on computing devices which requires network of remote servers hosted on internet with the various services like servers, storage, software, networking, databases, analytics and intelligence on the internet ("the cloud") for faster innovation, allocation of flexible resources as well as economies of the scale.

In short Cloud computing is useful for manipulating, configuring and accessing the various applications online. It also gives the services like online data storage, infrastructure and various applications.

Cloud refers to Internet or Network. Cloud is something which is present at remote location.

Computing refers to service provider over network. May be Public or Private networks.

### 12.1.1 CLOUD ENABLING TECHNOLOGIES

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

- Grid computing – It is emerging enabling technology. Useful for distributed systems and the network or Internet. At the same time, it enables to work on heterogeneous computers in a network for working together on a single task. It is also known as parallel computing. Grid computing is best for large workloads.
- Utility computing – It is a service provider model. As per the requirement of the customer Service provider prepares computing resources available to the customer, charges depending on the demand services and usage.
- Virtualization – It is a technique which allows to share single physical instance of an application or resource among multiple organizations or customers. It works on multiple operating system and applications on the same server at the same time. Virtualization is the process of creating a virtual or logical view of a server operating system, a storage device or networking services. The technology uses in virtualization is known as a virtual machine monitor (VM).
- Service Oriented Architecture (SOA) – It provides a various service that can communicate with each other on the network. Various services work together to run various activities.

## 12.2 CHARACTERISTICS OF CLOUD COMPUTING

1. On Demand self-service – Various services like email, server service or application network can be provided without requiring any interaction with each service provider.

   Cloud service providers gives the services on demand self-services like Microsoft, IBM and saleforce.com, Amazon Web Service.
2. Broad Network Access – Cloud capabilities and capacity are available over the network and we can access through standard mechanism that promote use by different heterogeneous clients like mobile phones, laptops.

3. Resource pooling – Service providers resources get pooled to serve multiple consumer requirements with different physical and virtual resources, dynamically assigned and reassigned to consumer demand. Resources consist of storage allocation, processing, network bandwidth and memory.

4. Rapid elasticity – To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

   Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand.

5. Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g. 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.

## 12.3 BENEFITS OF CLOUD COMPUTING

Cloud computing offers the following benefits:

- Reduced Infrastructure cost – Cloud services can be purchased which is based on pay-as-per-usage or subscription pricing. This reduces or eliminates the consumer's IT capital expenditure.

- Business Agility - Cloud computing provides the capability to allocate and scale computing capacity. Cloud computing reduces the time and cost to deploy the new applications and services from months to minutes. It enables the businesses to respond more quickly to market changes and reduce time-to-market.

- Flexible scaling – One of the major benefits of cloud computing for any business which has opted cloud computing can increase or decrease the bandwidth as per requirement.

- High availability – Cloud computing has the capacity to ensure resource availability at varying levels depending on the consumer's demand.

## 12.4 CLOUD SERVICE MODELS

Cloud service models consist of three types Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS).

**12.4.1 Infrastructure-as-a-Service:** Service includes provision of hardware and software for processing, the data storage, networks and any required infrastructure for deployment of operating systems and applications which would normally be needed in a data-center managed by the user.The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems and deployed applications and possibly limited control of select networking components e.g host firewalls.

IaaS is the base layer of the cloud services stack. It serves as the foundation for both SaaS and PaaS layers.

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.

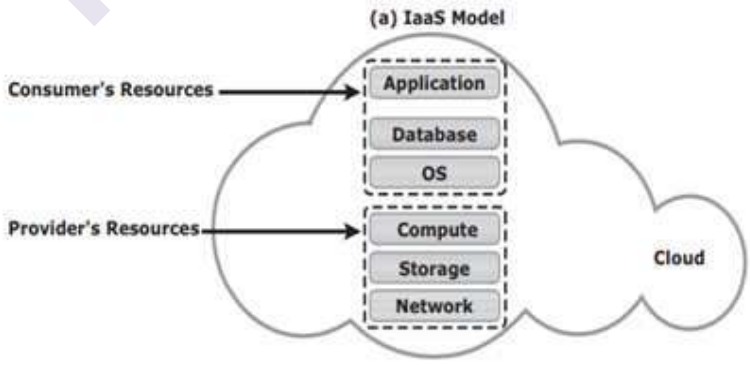



**Figure 12.4.1 Infrastructure-as-a-Service**

**12.4.2 Platform-as-a-Service –** PaaS is a cloud offering that provides infrastructure for development and deployment of applications. It provides the middleware, development tools, and artificial intelligence to create powerful applications. PaaS services gives the bundled together with the network infrastructure and storage services. With PaaS we can enables faster time to market, multiplatform development and easy collaboration.

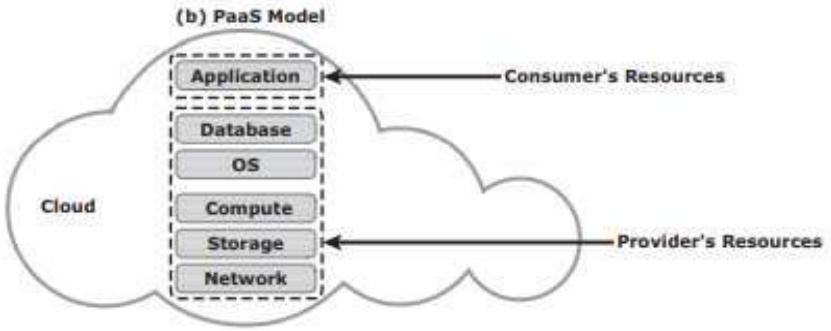



**Figure 12.4.2 Platform as a Service**

**12.4.3 Software-as-a-service -** SaaS is a model in which software provides the services on demand of the consumer. It is known as software + services as well as it is also known as on-demand software and web-hosted or web-based software. SaaS application are usually used by the user with the help of web browser. SaaS has a model for various business applications like messaging software, payroll processing of employees, CAD software, accounting software, customer relationship management (CRM), DBMS software, Enterprise resource planning (ERP), Geographic information systems (GIS), management information systems (MIS) and many more various applications.

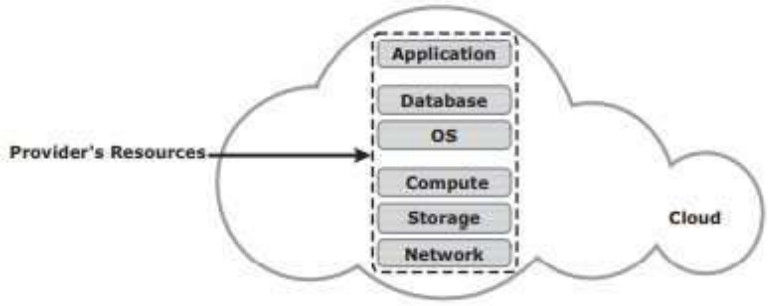



**Figure 12.4.3 a) Software as a Service(SaaS Model)**

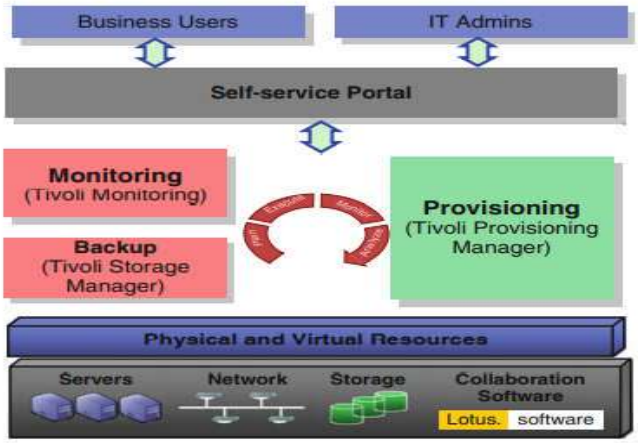



**Figure b): SaaS cloud**

**Table: IaaS, PaaS and Saas**

| Service type | IaaS | PaaS | SaaS |
|---|---|---|---|
| Service category | VM Rental, Online Storage | Online Operating Environment, Online Database, Online Message Queue | Application and Software Rental |
| Service Customization | Server Template | Logic Resource Template | Application Template |
| Service Provisioning | Automation | Automation | Automation |
| Service accessing and Using | Remote Console, Web 2.0 | Online Development and Debugging, Integration of Offline Development Tools and Cloud | Web 2.0 |
| Service monitoring | Physical Resource Monitoring | Logic Resource Monitoring | Application Monitoring |
| Service level management | Dynamic Orchestration of Physical Resources | Dynamic Orchestration of Logic Resources | Dynamic Orchestration of Application |
| Service resource optimization | Network Virtualization, Server Virtualization, Storage Virtualization | Large-scale Distributed File System, Database, Middleware etc | Multi-tenancy |
| Service measurement | Physical Resource Metering | Logic Resource Usage Metering | Business Resource Usage Metering |
| Service integration and combination | Load Balance | SOA | SOA, Mashup |
| Service security | Storage Encryption and Isolation, VM Isolation, VLAN, SSL/SSH | Data Isolation, Operating Environment Isolation, SSL | Data Isolation, Operating Environment Isolation, SSL, Web Authentication and Authorization |

## 12.5 CLOUD DEPLOYMENT MODELS

Cloud computing is classified into four different deployment models namely public, private, community and hybrid. Cloud Deployment models provide the services on the basis of how cloud infrastructure is organised and usage of it.

**12.5.1 Public cloud –** In a public cloud model, cloud infrastructure is provision for the general public usage. It may be manged, owned and operated by a business academic, or government organization, or some combination of them. It exists on the premises of the cloud service provider.

Consumers can use the various cloud services offered by the providers via the Internet and pay as per usage charges or subscription fees. The main advantage of public cloud is its less capital cost with high scalability. For consumers, these benefits come with few risks where no control over the various infrastructure or resources in the cloud, the network performance, security of confidential data and interoperability problem may occur. Examples of public cloud service providers are Google, Amazon and Saleforce.com all these business community uses public cloud that provides various cloud services to organizations and individuals.

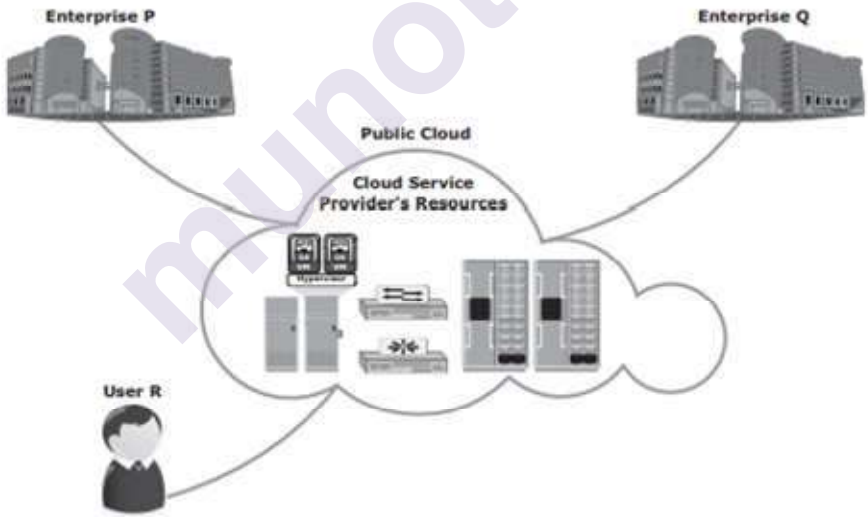



**Figure 12.5.1 Public Cloud**

**12.5.2 Private Cloud –** In private cloud model, cloud infrastructure is provision for the exclusive use by the individual organization for multiple consumers.

Example It may use for owned, managed and operated by the various business community or organization, where third party

involvement or few combinations of them. There are two different variations to the private cloud model namely On-premise private cloud and Externally hosted private clouds.

a) **On-premise private cloud –** It is also known as internal cloud, which is hosted by an organization within its own data centers. It is useful for management or standardization of various cloud service processes as well security. It is having limitations about size and resource scalability. It is very useful for the organization which require complete control over their applications, configurations of infrastructure and security purpose.

b) **Externally hosted private cloud**– This type of private cloud is useful for hosting the externals to business organizations. It may be managed by third party as well. Third party organization gives the various facilities an exclusive cloud environment for a specific organization with full guarantee of confidentiality and privacy.

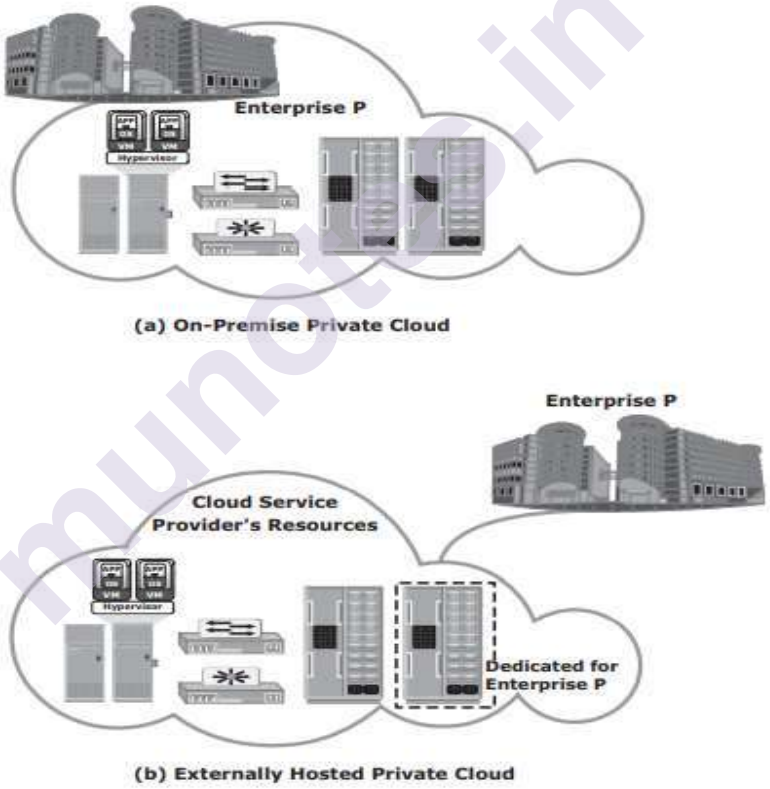



**Figure 12.5.2 On-Premise and externally hosted private clouds.**

**12.5.3 Community Cloud –** In community cloud the various services provision for infrastructure for exclusive use by a specific community of consumers from various business organizations that have shared with various services like security requirements, compliance considerations, policy etc. It can be owned, managed and operated by one or more of the organizations in the community, a third party or combination of both as well. It is cost effective because cloud is shared in various organizations or a community. It is compatible with every user so it is scalable and flexible.

Security is more in community cloud as compare to public cloud but less secure than the private cloud. In Community cloud slow adaptation of data. It may be not good choice for some organization. Responsibilities of sharing among organizations is very difficult.

**Example:** Government organization in India may share the computing infrastructure in the cloud to manage data.

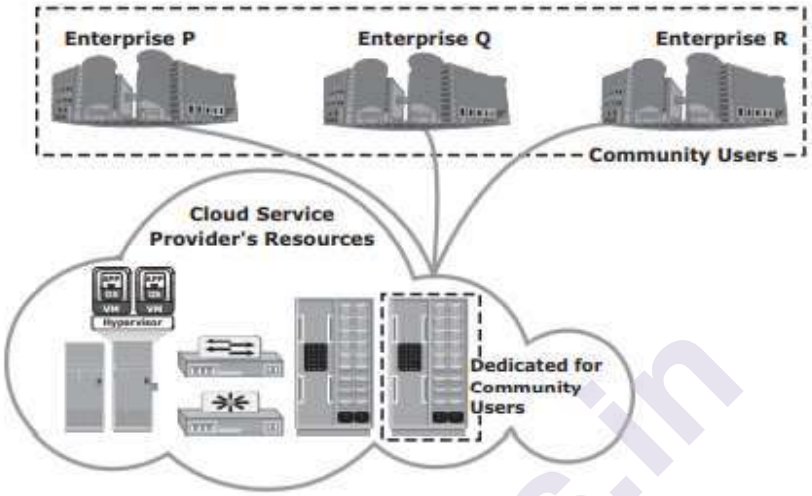



**Figure 12.5.3 Community cloud**

**12.5.4 Hybrid Cloud –** In hybrid cloud model, It is a combination of two or more cloud infrastructure may be private, public or community. Hybrid cloud = public cloud + private cloud. These combination of public and private create a unified, automated, and well managed computing environment. The activities which are critical are performed by private cloud whereas non-critical activities are performed by public cloud. Hybrid cloud is useful in finance, universities and healthcare centers. It is secure because of private cloud and flexible because of public cloud. It is cheaper than private cloud. It helps in cost cutting with the parameter as infrastructure and application support. It accepts all demands of company related with need of space, memory and system. Networking issues , Reliability and Infrastructure compatibility are the issues in Hybrid cloud.

Examples of hybrid cloud are Microsoft, Google, Amazon, Cisco and NetApp.

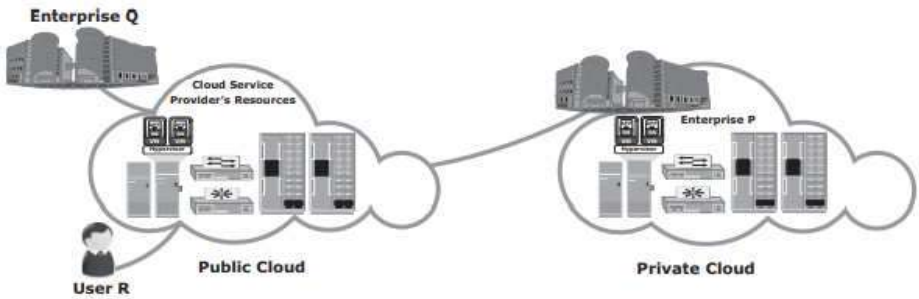



**Figure 12.5.4 Hybrid cloud**

## 12.6 CLOUD COMPUTING INFRASTRUCTURE

It is a collection of software as well as hardware that gives the five essential characteristics of cloud computing. Cloud computing infrastructure consist of various layers:

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

**12.6.1 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 with the help of networks, such as FC SAN, IP, IP SAN, or FCoE networks. Cloud service providers may use physical computing resources from one or more data centers to provide services. Computing resources are get distributed across various data centers, connectivity must be established between them.

**12.6.2 Virtual Infrastructure –** Cloud service provider 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 helps to reduce the cost of providing the cloud services. Some cloud service providers may not have completely virtualized their physical infrastructure, but they are adopting virtualization for better efficiency and optimization.

**12.6.3 Applications and Platform Software –** This layer includes a suite of business applications and platform software, such as the operating system and database. Platform software provides the environment on which business applications can run i.e. VM. Applications and platform software are hosted on virtual machines to create SaaS and PaaS. For SaaS, both the application and platform software are provided by cloud service providers. In case of PaaS, only the platform software is provided by cloud service providers; consumers export their applications to the cloud. In short Software platform service get provided by cloud.

**12.6.4 Cloud Management and Service Creation Tools –** The cloud management and service creation tools, this layer consists of three types of software;

1) Physical and virtual infrastructure management software
2) Unified management software
3) User-access management software

All these three types of software interact with each other for the provision of cloud services

1) **Physical and virtual infrastructure management software –** The physical and virtual infrastructure management software is offered by the vendors of various infrastructure resources and third-party organizations. 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 physical infrastructure. It depends on vendor's perspective or third party perspective.

2) **Unified management software –** 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 provides a single management interface to configure physical and virtual infrastructure and calculate both CPU, memory, network, and storage pools. By using configuration commands to respective physical and virtual infrastructure management software, which executes the instructions. The main function of unified management software is to automate the creation of cloud services. It enables administration to define service attributes like power of CPU, memory, storage capacity, bandwidth of network, name and description of applications and resource location, platform software and backup policy. If any request comes from consumer's side then the unified management software it will create a service based on the pre-defined service parameters.

3) **User-access management software –** It is web-based user interface to consumers. Consumers can use the browsing the service catalogue and request cloud services. The user-access management software authenticates users before forwarding their request to the unified management software. It is useful to monitor allocation or utilization of resources associated to the cloud service instances. On the basis of allocations of resources, it generates a charge back report. It is useful for consumer and the service provider.

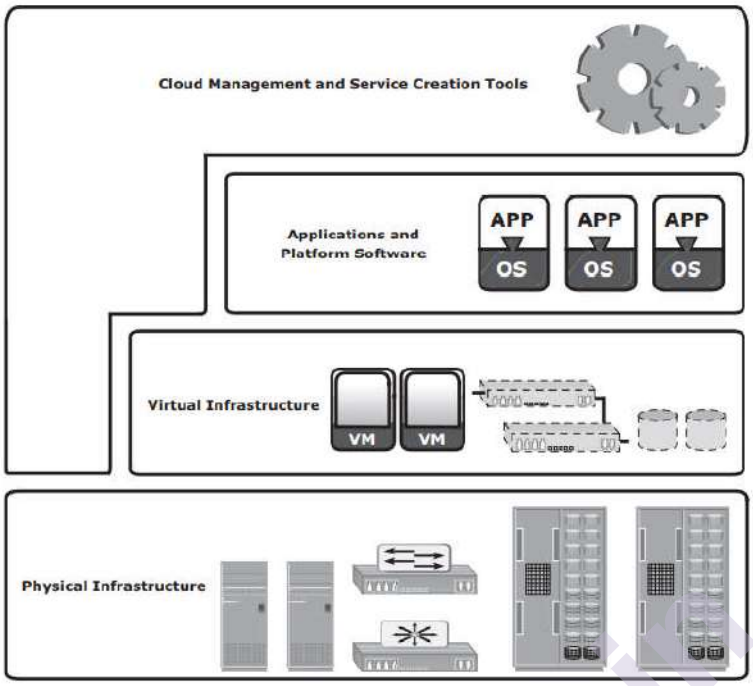



**Figure 12.6 Cloud infrastructure layers**

## 12.7 CLOUD CHALLENGES

Cloud computing is associate with cloud service, consumer and service providers and all of them have been facing few challenges.

### 12.7.1 Challenges for Consumers

Cloud services providers gives the usage of multiple data security, while consumers may not be able to transfer control of business data which is critical to the cloud. Cloud services providers are using multiple data centers which are located at different countries to give the cloud services, These data get replicated or move between different data centers for high availability and for load distribution services, where as consumers may not be able to give these services. Some of the cloud service providers give the options to the consumers to select the location for storing their data. It may create problem in data protection as well as data privacy concerns and regulatory compliance requirements like EU data protection directive and U.S. safe harbor program, it may create adaptation of cloud computing challenges for the consumers. Cloud services may not support for consumer's desired expectations to run the applications on cloud, might be because of highly specialized or not compatible operating system, programming languages which is required to develop or run the applications. Vendor lock-in option may occur difficulties for consumers when they want to change their cloud service provider. High migration cost while moving the data from one service provider to the another, cost effect may also the challenge for consumers.

### 12.7.2 Challenges for Providers

Cloud service providers might not provide every time the various service levels. Most of the software vendors does not have cloud-ready software licensing model. Few software vendors give the standardized cloud license with high cost as compared to the traditional licensing cost. Cloud software licensing complexity may cause the challenges while deploying the software at vendor's side in the cloud. Cloud resources are get distributed and service demands also changes as per vendor's requirement so cloud service providers have the provision of physical resources for peak demand of all vendors or consumers and to calculate the actual of the services which is going to provide by service provider. Agreement between the cloud service providers and the tenant of multiple clouds may create a challenge.

## 12.8 CLOUD ADOPTION CONSIDERATIONS

Cloud adoption has some key points as follows:

a) Selection of a deployment model: Convenience versus risk is a key factor for selection of on a cloud adoption. Selection of right cloud for deployment model.

   Public cloud is useful for individuals or for start-up businesses, in case of public cloud cost reduction offered by it but the security as well as availability risk in the cloud. Small or medium business organizations will not be willing for deployment of the online transaction processing in the public cloud as customer data and service levels may impact their business, in that case customer can select hybrid cloud for business operations. For backup, archive and testing attributes can be deploy with the help of public cloud.

b) Application suitability: All applications are not good for a public cloud so this may be incompatibility between the cloud platform software and the consumer applications, or for business organizations as well. They are basically designed, developed and maintained in house. Because of high risk organizations are not ready to move all applications to the public cloud. These applications are good for the on-premise private cloud.

c) Financial advantage: While adopting the cloud financial benefits provides a cost savings. The analysis of cost -saving shows the comparision between total cost of ownership (TCO) and the return on investment (ROI) in the cloud cost benefits. Calculating the expenditures for infrastructure resources, business organization must include the capital expenditure (CAPEX) which contains cost of storage, servers, operating systems, application, real estate, network equipment and operation expenditure (OPEX) which contains personnel, backup, power and cooling, maintenance and so on. The

cloud adaptation cost includes the cost of migrating the cloud, compliance and security, subscription fees etc.

d) Selection of a cloud service provider: Selection of a cloud service provider is very important for a public cloud. It depends on the services which are provided to the consumers. Security, privacy requirements, rules and regulations should check while selecting the service provider with good customer services support.

e) Service-level agreement (SLA): Quality of service (QoS) is an important factor in cloud service such as throughput and uptime is also the services of cloud. QoS is a part of an service level agreement between the consumer and the provider. Before adopting the cloud service, every consumer have to check whether the QoS meets with their requirements or not.

## 12.9 SUMMARY

Cloud computing chapter gives the detail study about the cloud characteristics, benefits, services, deployment models and infrastructure as well. It also describes cloud challenges, cloud challenges for consumers, cloud challenges for providers and adaptation considerations.

## 12.10 EXERCISE

1) What are the various characteristics of cloud computing?

2) Discuss different cloud challenges.

3) Explain benefits of cloud computing.

4) Discuss cloud adaptation in detail

5) Discuss different cloud services in detail.

## 12.11 REFERENCES

Information storage and management: storing, managing and protecting digital information in Classic, Virtualized and Cloud Environments, EMC author, by Joh Wiley and Sons 2nd edition 2012.

https://books.google.co.in/books?id=PU7gkW9ArxIC&printsec=frontcover&dq=information+storage+and+management&hl=en&newbks=1&newbks_redir=1&sa=X&ved=2ahUKEwjjx_nakNPxAhWy4zgGHWUpCjcQ6AEwAHoECAsQAg

https://books.google.co.in/books?id=sCCfRAj3aCgC&printsec=frontcover&dq=information+storage+and+management&hl=en&newbks=1&newbks_redir=1&sa=X&ved=2ahUKEwjjx_nakNPxAhWy4zgGHWUpCjcQ6AEwAXoECAIQAg

https://www.slideshare.net/golujain/characteristics-of-cloud-computing-as-per-nist

https://www.zoho.com/creator/paas/

https://en.wikipedia.org/wiki/Software_as_a_service

https://www.javatpoint.com/community-cloud

https://www.javatpoint.com/hybrid-cloud

❖ ❖ ❖ ❖

# Unit V

# 13

# SECURING THE STORAGE INFRASTRUCTURE

**Unit Structure**





## 13.0 OBJECTIVES

After going through this unit, you will be able to learn

Basic storage security implementations, such as the security architecture and protection mechanisms in FC-SAN, NAS, and IP-SAN, are covered in this chapter.

In addition, in virtualized and cloud systems, this chapter discusses new security considerations.

Further, this chapter describes the additional security considerations in virtualized and cloud environments.

## 13.1 INTRODUCTION

Important data, such as intellectual property, personal identities, and financial transactions, is routinely processed and stored in storage arrays accessible via the network. As a result, storage is now more vulnerable to a variety of security threats which have the potential to damage business-critical data and disrupt critical services. In both traditional and virtualized data centres, securing storage infrastructure has become an essential part of the storage management process. It's a time-consuming but crucial process for maintaining and securing sensitive data.

Because organisations have less control over shared IT infrastructure and the enforcement of security controls, storage security in a public cloud environment is more complicated. Furthermore, multitenancy in a cloud environment allows multiple users to share resources, such as storage. Data may be tampered across tenants as a result of such sharing.

## 13.2 INFORMATION SECURITY FRAMEWORK

The fundamental information security architecture is designed to accomplish four security objectives: confidentiality, integrity, and availability (CIA), as well as responsibility. All security standards, procedures, and controls required to minimise threats in the storage infrastructure environment are included in this framework.

### 13.2.1 Confidentiality:

Ensures that information is kept private and that only authorised users have access to it. Users that need access to information must be authenticated. Data in transit (data sent over a network) and data at rest (data stored on a primary storage device, backup media, or in archives) can both be encrypted to ensure privacy. Confidentiality necessitates the implementation of traffic flow protection mechanisms as part of the security protocol, in addition to preventing unauthorised users from accessing information. These safeguards often include the concealment of source and destination addresses, the frequency with which data is delivered, and the volume of data sent.

### 13.2.2 Integrity:

Ensures that the data hasn't been tampered. Integrity protection necessitates the detection and prevention of unwanted data tampering or

destruction. For both data and systems, ensuring integrity necessitates methods such as error detection and correction.

### 14.2.3 Availability:

This assures that authorized users have consistent and timely access to these systems' systems, data, and applications. Protection against unwanted data deletion and denial of service is required for availability.

The availability of sufficient resources to deliver a service is also implied by availability.

### 13.2.4 Accountability:

All events and actions that occur in the data centre infrastructure must be accounted for. For security purposes, the accountability service keeps a trail of occurrences that can be audited or traced afterwards.

## 14.3 RISK TRIAD

Threats, assets, and vulnerabilities are all part of the risk triangle. When a threat agent (an attacker) exploits an existing vulnerability to compromise an asset's security services, for example, if a sensitive document is sent over an unsecured channel without any protection, an attacker may get unauthorised access to the document and violate its confidentiality and integrity. This could result in a loss of revenue for the company. In this case, the risk of business loss derives from an attacker using the vulnerability of unencrypted communication to gain access to the document and tamper with it.

Organizations rely on vulnerabilities to control risks because they can't completely eradicate threat agents that arrive in numerous forms and sources to their assets. Organizations can use countermeasures to lessen the likelihood of attacks and the severity of their consequences.

The first step in determining the scope of potential threats and dangers in an IT infrastructure is to conduct a risk assessment. The procedure evaluates risk and aids in the identification of appropriate controls to reduce or eliminate it. Risk assessment helps prioritise security measures investment and provisioning based on the value of assets. Threats to an IT system must be examined along with potential vulnerabilities and existing security controls to determine the likelihood of an unfavourable event occurring.

The impact of an adverse incident on important business activities is used to determine its severity. IT assets and resources can be ascribed a relative value of criticality and sensitivity based on this research. For example, a high-criticality value could be assigned to an IT system component if an assault on it could result in the entire shutdown of mission-critical services.

The three essential parts of the risk triangle are examined in the following sections:

Assets, threats and vulnerabilities are considered from the perspective of risk identification and control analysis.

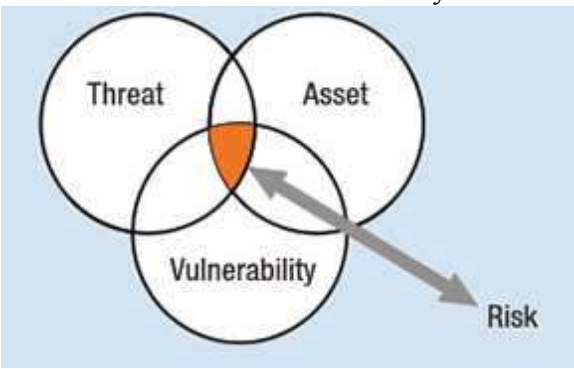



**Fig . Risk Tried**

**13.3.1 Assets:**

Information is one of an organization's most valuable assets. Hardware, software, and other infrastructure components necessary to access the information are examples of other assets. Organizations must design a set of parameters to ensure that resources are available to authorised users and trustworthy networks in order to protect these assets. Storage resources, network infrastructure, and organisational policies are all affected by these elements.

There are two goals to security measures. The first goal is to make sure that authorised users may readily access the network. It should also be dependable and stable in a variety of environments and usage volumes. The second goal is to make it difficult for potential attackers to gain access to the system and compromise it.

Unauthorized access, viruses, worms, trojans, and other harmful software programmes should all be protected by the security procedures. To reduce the amount of potential security threats, security solutions should include choices to encrypt vital data and terminate unnecessary services. The security strategy must ensure that operating system and other software updates are installed on a regular basis. Simultaneously, it must provide sufficient redundancy in the form of replication and mirroring of production data to prevent catastrophic data loss in the case of a data compromise. All users are informed about the policies controlling network use in order for the security system to function properly.

Two main factors can be used to assess the success of a storage security methodology. One, the cost of putting the system in place should be a small percentage of the value of the data being secured. Two, a potential attacker should pay a high price in terms of money, effort, and time.

### 13.3.2 Threats:

Threats are attacks that could be launched against an IT infrastructure. There are two types of attacks: active and passive. Attempts to acquire unauthorised access to a system are known as passive attacks. They put information confidentiality in risk. Data alteration, denial of service (DoS), and repudiation assaults are examples of active attacks. They expose the integrity, availability, and accountability of data.

- An unauthorised user tries to change information for malevolent purposes in a data modification attack. A data alteration attack might target data in transit or data at rest. Data integrity is compromised by these attacks. Attacks that disable legitimate users' access to resources and services are known as denial of service (DoS) attacks. In most cases, these assaults do not entail gaining access to or altering information. Instead, they put data availability at risk. A DoS attack occurs when a network or website is deliberately flooded in order to impede lawful access to authorised users.

- Repudiation is an attack on the information's accountability. It tries to offer misleading information by impersonating someone or denying the occurrence of an event or a transaction. A repudiation assault, for example, would entail executing an activity and then destroying any evidence that could be used to show the identity of the user (attacker) who carried it out. Circumventing the reporting of security events or tampering with the security log to mask the attacker's identity are examples of repudiation attacks.

### 13.3.3 Vulnerability:

Access points to information are frequently open to prospective assaults. Each path may contain a number of access points that grant varying levels of access to the storage resources. It's critical to put in place suitable security controls at all stages along an access path. Defense in depth refers to putting security measures in place at each access point along each access path.

If one component of security is compromised, security advocates adopting numerous security measures to lessen the risk of security risks. It's also known as a "layered security technique." Security allows for more time to notice and respond to an attack because there are several security measures in place at various levels. A security breach's breadth or impact can be reduced due to this.

When determining the amount to which an environment is vulnerable to security threats, three elements must be considered: attack surface, attack vector, and work factor. The many entry points that an attacker can use to start an assault are referred to as the attack surface. Each component of a storage network has the potential to be a source of

vulnerability. An attacker can leverage all of the component's external interfaces, such as the hardware and management interfaces, to carry out numerous assaults. The attacker's attack surface is made up of these interfaces. If enabled, even unused network services can become part of the attack surface.

An attack vector is a step or a set of steps that must be followed to complete an attack. For example, an attacker might use a flaw in the management interface to launch a snoop attack, in which the attacker changes the storage device's configuration to allow traffic to be accessed from another host. The data in transit can be snooped via this diverted traffic.

The amount of time and effort necessary to exploit an attack vector is referred to as the work factor. When attempting to retrieve sensitive information, for example, attackers evaluate the time and effort required to carry out a database attack. This could entail figuring out who has access to what, figuring out the database schema, and developing SQL queries. Instead, they may select a less effort-intensive technique to exploit the storage array by connecting to it directly and reading from the raw disc blocks, based on the work factor.

Organizations can apply specific control measures after assessing the environment's susceptibility. Any control mechanisms should take into account all three parts of infrastructure: people, process, and technology, as well as their interactions. The first stage in securing persons is to determine and confirm their identity. Selective controls for their access to data and resources can be implemented based on their identify. Processes and procedures are the primary determinants of any security measure's effectiveness. The procedures should be based on a thorough awareness of environmental concerns, as well as the relative sensitivity of various types of data and the requirements of various stakeholders for data access. The adoption of technology is neither cost-efficient nor aligned with the priorities of enterprises without an effective methodology. Finally, for the technology or controls to be effective, they must assure compliance with the processes, policies, and people. The goal of these security systems is to reduce vulnerability by lowering attack surfaces and increasing work factors. Technical or nontechnical controls can be used. Nontechnical controls are normally implemented through administrative and physical controls, whereas technical controls are usually done using computer systems. Security and personnel policies, as well as standard procedures, are examples of administrative controls that govern the safe execution of diverse operations. Setting up physical barriers, such as security guards, fences, or locks, are examples of physical controls.

**Controls are classified as:**

**Preventative**

The preventive control aims to prevent an attack; the detective control determines whether an assault is underway; and the remedial controls are executed after an attack is discovered. Preventive measures stop vulnerabilities from being exploited, preventing or reducing the impact of an attack.

**Detective, or corrective based**

Detective controls identify attacks and trigger preventative or corrective controls, whereas corrective controls decrease the impact of an attack. An Intrusion Detection/Intrusion Prevention System (IDS/IPS) example, is a detective control that assesses whether an attack is in progress and then seeks to stop it by terminating a network connection or activating a firewall rule to restrict traffic.

## 13.4 STORAGE SECURITY DOMAINS

Storage devices that are connected to a network increase the risk level and are more vulnerable to network-based security risks. However, as storage settings become more networked, storage devices are becoming increasingly vulnerable to security attacks from a variety of sources. To protect a storage networking environment, specific controls must be established. This necessitates a more in-depth examination of storage networking security as well as a thorough understanding of the access paths to storage resources. If a specific path is unlawful and needs to be blocked by technical controls, make sure these measures aren't compromised. If each component in the storage network is regarded a potential access point, the attack surface of all of these access points must be examined in order to determine the vulnerabilities associated with them.

Access paths to data storage can be grouped into three security domains to identify dangers that apply to a storage network: application access, administration access, and backup, replication, and archive. The three security domains of a storage system environment are depicted in Figure

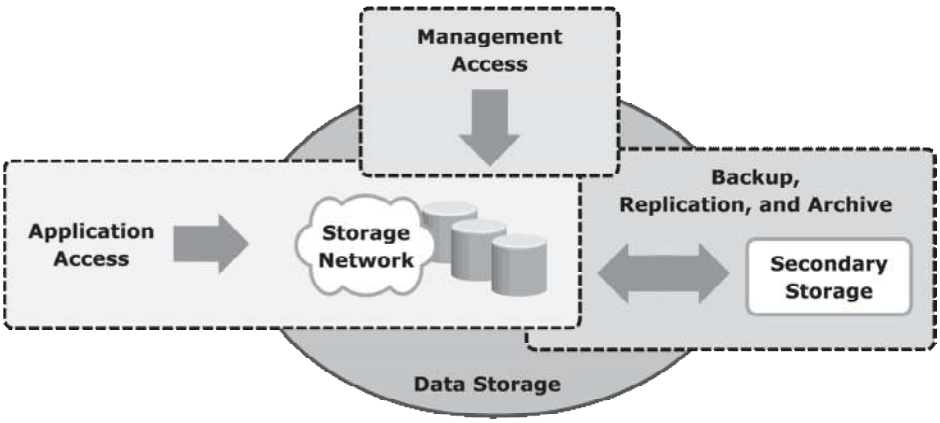



**Fig. Storage security domains**

Application access to stored data via the storage network is the first security domain. The second security domain covers management access to storage and connection devices, as well as the data they contain.

Storage administrators who configure and administer the environment use this domain the most. Backup, replication, and archive access make up the third domain. The backup media, like the access points in this domain, has to be protected.

Identify the existing dangers inside each of the security domains and classify the threats depending on the type of security services—availability, confidentiality, integrity, and accountability to safeguard the storage networking environment. The next stage is to choose and apply appropriate controls as dangers are discovered.

### 13.4.1 Securing the Application Access Domain

The application that access data through a file system or a database interface may be included in the application access domain.

Identifying dangers in the environment and implementing suitable controls is a crucial step in securing the application access domain. Physical security is also a significant concern when it comes to preventing media theft. In a storage networking context, Fig. depicts application access. All V1 volumes are accessible to Host A, and all V2 volumes are accessible to Host B. These volumes are classified as confidential, restricted, or public, depending on the level of access. In this scenario, some conceivable dangers include host A spoofing the identity or elevating to host B's credentials to get access to host B's resources. Another hazard is an unauthorised host gaining network access; the attacker on this host may attempt to mimic the identity of another host and tamper with data, probe the network, or launch a DoS assault. In addition, any sort of media theft could jeopardise security. These threats can pose

several serious challenges to the network security; therefore, they need to be addressed.

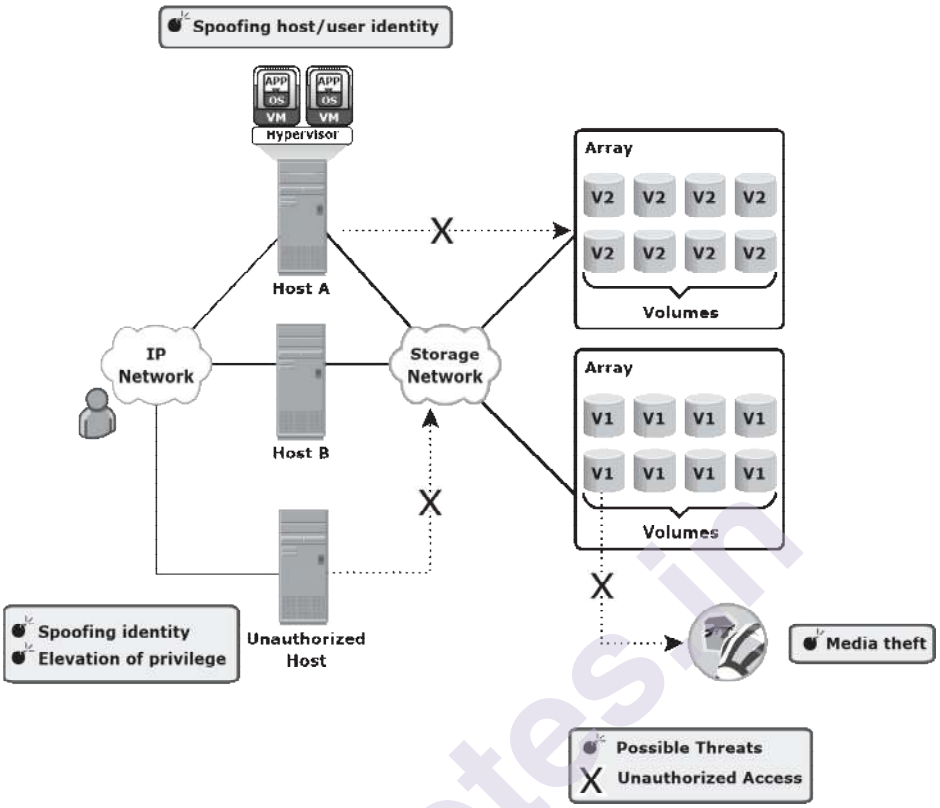



**Fig. : Security threats in an application access domain**

**Controlling User Access to Data**

User's access to data is controlled by access control services. The hazards of spoofing host identity and increasing host privileges are mitigated by these services. Both of these issues jeopardise data security and integrity.

User and host authentication (technical control) and authorization are the access control procedures used in the application access domain (administrative control). These mechanisms may exist outside the storage network's boundaries, necessitating the interconnection of various systems with other enterprise identity management and authentication systems, such as systems that provide strong authentication and authorization to protect user identities from spoofing. Access control lists can be created on NAS devices to limit user access to specified files. Information Rights Management (IRM), which specifies which users have what rights to a document, is used by the Enterprise Content Management applications to enforce data access. Authenticating a node when it tries to connect to a network is the first step towards restricting access at the host level.

Authentication procedures used by different storage networking technologies, such as iSCSI, FC, and IP-based storage, include Challenge-

Handshake Authentication Protocol (CHAP), Fibre Channel Security Protocol (FC-SP), and IPSec, respectively.

After a host has been authenticated, the next step is to establish security controls for the storage resources that the host is authorised to access, such as ports, volumes, or storage pools. Zoning is a switch control strategy that divides the network into certain data traffic pathways; LUN masking defines which hosts have access to which storage devices. Some devices allow you to map a host's WWN to a specific FC port and then to a specific LUN. The most secure method is to connect the WWN to a physical port.

Finally, administrative controls such as defined security rules and standards must be applied. Administrative controls must be audited on a regular basis to ensure that they are working properly. Significant events are logged on all participating devices to enable this. Unauthorized access to event logs should be avoided because they may fail to achieve their objectives if the logged content is subjected to unauthorised modifications by an attacker.

**Protecting the Storage Infrastructure**

Protecting the storage infrastructure from unauthorised access entails safeguarding all of the infrastructure's components. Unauthorized modification of data in transit that compromises data integrity, denial of service that compromises availability, and network surveillance that compromises confidentiality are all concerns that security rules for securing the storage infrastructure address.

There are two types of security controls for securing the network: network infrastructure integrity and storage network encryption. A fabric switch function that assures fabric integrity is one of the controls for assuring infrastructure integrity. This is accomplished by preventing unauthorised hosts from being added to the SAN fabric. The usage of IPSec for safeguarding IP-based storage networks and FC-SP for protecting FC networks are two storage network encryption technologies.

Root or administrator privileges for a specific device are not granted to every user in a secure storage environment. Instead, role-based access control (RBAC) is used to assign users the privileges they need to carry out their jobs. A role can be used to indicate a work function, such as an administrator. Privileges are linked to roles, and people gain access to these privileges as a result of their roles.

When defining data centre protocols, it's also a good idea to think about administrative controls like "separation of roles." A clear division of responsibilities ensures that no single person can both specify and carry out an action. The person who permits the creation of administrative accounts, for example, should not be the same person who uses them. In

the following part, we'll go through how to secure management access in further depth.

Storage system management networks should be logically separated from other enterprise networks. This segmentation is necessary to make management easier and to improve security by limiting access to components that are part of the same segment. IP network segmentation is enforced, for example, at Layer 3 with the use of routers and firewalls, and at Layer 2 with the use of VLANs and port-level security on Ethernet switches.

Finally, physical access to the device console and FC switch cabling must be managed to ensure the storage infrastructure's security. If an unauthorised user physically gains access to a device, all other specified security mechanisms fail, and the equipment becomes unreliable.

**Data Encryption**

Protecting data kept within storage arrays is the most crucial part of data security. At this level, threats include data manipulation, which undermines data integrity, and media theft, which jeopardises data availability and confidentiality. Encrypt the data on the storage media or the data before it is transmitted to the disc to guard against these risks. It's also crucial to choose a strategy for ensuring that data that's been wiped at the end of its life cycle is totally erased from discs and can't be rebuilt for malevolent purposes.

Data should be encrypted as close as feasible to its source. If encryption on the host device is not practicable, an encryption appliance can be used to encrypt data at the storage network's point of entry. Encryption devices that encrypt data between the host and the storage media can be implemented on the fabric. These techniques can secure data in transit as well as data at rest on the target device.

Adding antivirus scans and file extension controls to NAS devices can improve data integrity even more. The use of MD5 or SHA-256 cryptographic algorithms ensures data integrity in the case of CAS by detecting any changes in content bit patterns. Furthermore, before the disc is removed, the data erasure service verifies that the data has been completely overwritten by bit sequence. The data classification policy of an organisation determines whether the disc should be wiped before being discarded and the level of erasure required based on regulatory requirements.

### 13.4.2 Securing the Management Access Domain

Every device in the storage network has management access, whether it's for monitoring, provisioning, or controlling storage resources. The majority of management software includes a CLI, a system management console, or a web-based interface. Because the damage that can be produced by employing storage management applications is

significantly broad, it is critical to implement adequate controls for safeguarding these applications. Figure shows a storage networking architecture in which production hosts are connected to a SAN fabric and access production storage array A, which is replicated with remote storage array B. On Host A, this configuration also contains a storage management platform. An unauthorised host spoofing the user or host identity to operate the storage arrays or network is a possible hazard in this setting. An illegal host, for example, could get management access to remote array B.

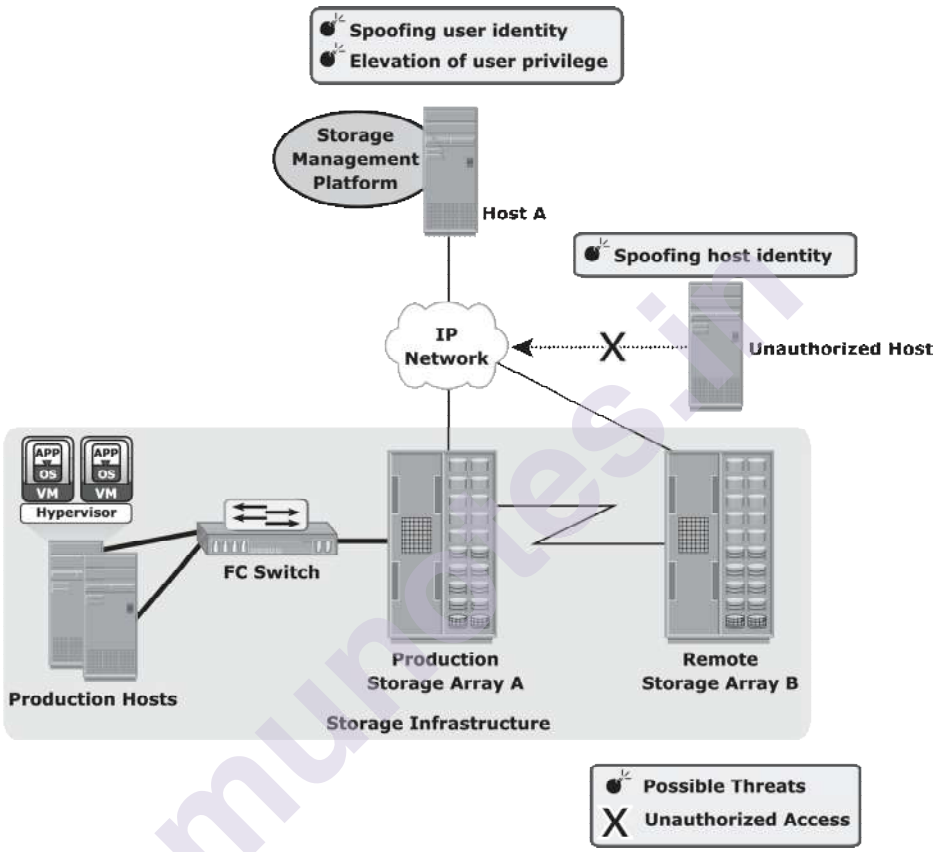



**Fig. Security threats in a management access domain**

Allowing administration access over an external network raises the risk of an unauthorised host or switch connecting to the network. In these situations, putting in place adequate security measures precludes some types of remote communication from taking place. Using secure communication methods, such as Secure Shell (SSH) or Secure Sockets Layer (SSL)/Transport Layer Security (TLS), protects against these threats effectively. Unauthorized access and changes to the infrastructure can be detected using event log monitoring. Event logs should be kept outside of shared storage systems so that they can be examined if the storage is hacked.

The available security controls on the storage management platform must be confirmed, and these controls must be adequate to secure

the total storage environment. An attacker cannot alter the entire storage array and cause unbearable data loss by reformatting storage media or making data resources unavailable unless the administrator's identity and role are protected against spoofing attempts.

**Controlling Administrative Access**

Controlling administrative access to storage tries to prevent an attacker from impersonating an administrator or increasing privileges to get administrative access. Both of these dangers jeopardise the security of data and equipment. Administrative access restriction and other auditing approaches are used to impose responsibility of users and processes in order to protect against these dangers. For each storage component, access control should be implemented. It may be necessary to integrate storage devices with third-party authentication directories, such as Lightweight Directory Access Protocol (LDAP) or Active Directory, in various storage setups.

According to security best practises, no single person should have complete authority over the system. If an administrator user is required, the number of activities that require administrative permissions should be kept to a minimum. Instead, RBAC should be used to allocate various administrative duties. Auditing logged events is a crucial control mechanism for tracking an administrator's operations. However, access to administrative log files and their content must be protected. Deploying a reliable Network Time Protocol on each system that can be synchronized to a common time is another important requirement to ensure that activities across systems can be consistently tracked. In addition, having a Security Information Management (SIM) solution supports effective analysis of the event log files.

**Protecting the Management Infrastructure**

Encrypting management traffic, implementing management access rules, and following IP network security best practises are all ways to protect the management network architecture. The usage of IP routers and Ethernet switches to restrict traffic to certain devices is one of these best practises. The threat of an unauthorised device connecting to the network and gaining access to the management interfaces is reduced by restricting network activity and access to a small number of hosts. Access controls need to be enforced at the storage-array level to specify which host has management access to which array. Some storage devices and switches can restrict management access to particular hosts and limit the commands that can be issued from each host.

Encrypting management traffic, implementing management access rules, and following IP network security best practises are all mechanisms to protect the management network architecture. IP routers and Ethernet switches are used to restrict traffic to certain devices as part of these best practises. The threat of an unauthorised device connecting to the network and gaining access to management interfaces is reduced by restricting

network activity and access to a small number of hosts. To summarize, security enforcement must focus on the management communication between devices, confidentiality and integrity of management data, and availability of management networks and devices.

### 13.4.3 Securing Backup, Replication, and Archive

The third domain to protect against an attack is backup, replication, and archive. A backup, as described in Chapter 10, is copying data from a storage array to backup media like tapes or discs. Backup security is complicated, and it relies on backup software that connects to storage arrays. It also depends on how the storage infrastructures at the primary and secondary locations are configured, especially with remote backup solutions that use a remote tape device or array-based remote replication. Organizations must ensure that the disaster recovery (DR) site maintains the same level of security for the backed up data. Protecting the backup, replication, and archive infrastructure requires addressing several threats, including spoofing the legitimate identity of a DR site, tampering with data, network snooping, DoS attacks, and media theft. Such threats represent potential violations of integrity, confidentiality, and availability. Figure depicts a generic remote backup configuration in which data on a storage array is duplicated to secondary storage at the DR site through a DR network. Threats at the transmission layer must be addressed in a remote backup system where the storage components are separated by a network. Otherwise, an attacker can impersonate the backup server's identity and request that the host transfer its data. An illegal host posing as the backup server may result in a remote backup to an unauthorised and unknown location. Furthermore, attackers can utilise the DR network connection to manipulate with data, snoop on the network, and launch a denial-of-service attack against the storage devices. The physical threat of a backup tape being lost, stolen, or misplaced, especially if the tapes contain highly confidential information, is another type of threat. Backup-to-tape applications are vulnerable to severe security implications if they do not encrypt data while backing it up.

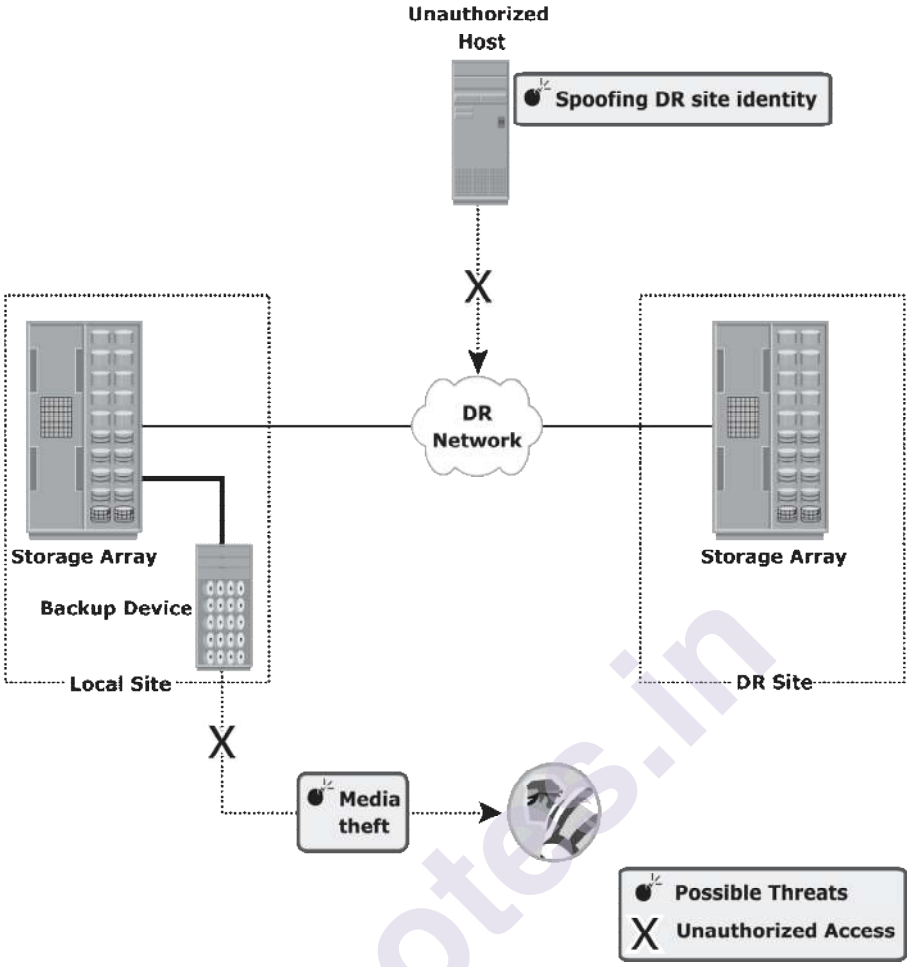



Figure: Security threats in a backup, replication, and archive environment

## 13.5 SECURITY IMPLEMENTATIONS IN STORAGE NETWORKING

The following discussion details some of the basic security implementations in FC SAN, NAS, and IP-SAN environments.

### 13.5.1 FC SAN

In comparison to IP-based networks, traditional FC SANs have a built-in security advantage. An FC SAN can be thought of as a private, isolated network with fewer nodes than an IP network. Consequently, FC SANs impose fewer security threats. With converged networks and storage consolidation, however, this picture has changed, driving rapid expansion and mandating designs for big, sophisticated SANs that cover several sites across the organisation. For FC SANs, there is currently no one complete security solution available. Many security methods in FC SANs have evolved from their IP networking counterparts, resulting in mature security solutions. Fibre Channel Security Protocol (FC-SP) standards (T11 standards), published in 2006, align security mechanisms and algorithms between IP and FC interconnects. These standards

describe protocols to implement security measures in a FC fabric, among fabric elements and N-Ports within the fabric. They also include guidelines for authenticating FC entities, setting up session keys, negotiating the parameters required to ensure frame-by-frame integrity and confidentiality, and establishing and distributing policies across an FC fabric.

**FC SAN Security Architecture**

Because of their size and complexity, storage networking setups are a potential target for illegal access, theft, and misuse. As a result, security tactics are built on the defence in depth principle, which calls for numerous levels of security to be integrated. This ensures that the assets under protection are not jeopardised if one of the security controls fails. Figure depicts the several levels (zones) of a storage networking environment that must be guarded, as well as the security solutions that can be used. FC SANs not only suffer from certain risks and vulnerabilities that are unique, but also share common security problems associated with physical security and remote administrative access. In addition to implementing SAN-specific security measures, organizations must simultaneously leverage other security implementations in the enterprise. Lists a variety of protection techniques that must be adopted in different security zones. List certain security procedures that aren't exclusive to SAN but are ubiquitous data centre techniques. Two-factor authentication, for example, is commonly used; in a basic implementation, it entails the use of a username/password as well as an additional security component, such as a smart card, for authentication.

**Basic SAN Security Mechanisms**

LUN masking and zoning, switch-wide and fabric-wide access control, RBAC, and logical partitioning of a fabric (Virtual SAN) are the most commonly used SAN security methods.

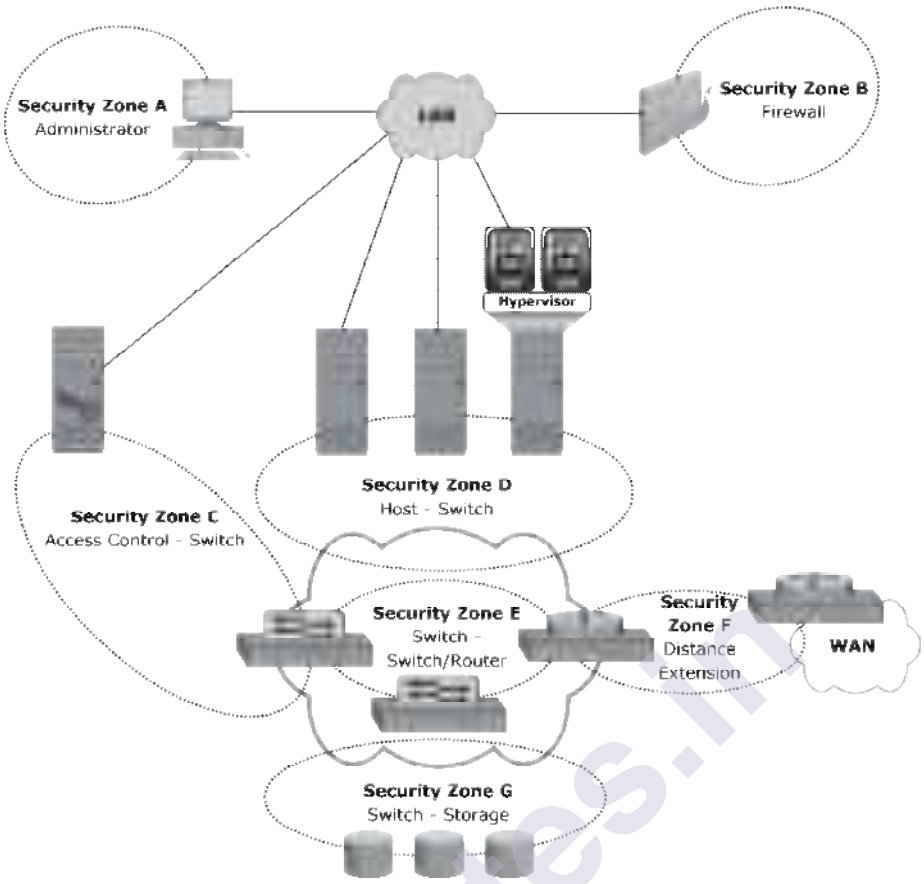



**Fig. FC SAN security architecture**

**Table 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. |

| SECURITY ZONES | PROTECTION STRATEGIES |
|---|---|
| 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. |

**LUN Masking and Zoning**

The primary SAN security measures used to defend against unwanted access to storage are LUN masking and zoning. LUN masking and zoning are discussed in length in Chapters 4 and 5. The WWPNs of the source HBAs are used to mask the LUNs supplied to a front end storage port in standard LUN masking implementations on storage arrays. A more powerful variation of LUN masking may be available on occasion, with masking based on source FC addresses. It has a technique for locking down a node port's FC address to its WWN. In security-conscious environments, WWPN zoning is the favoured option.

**Securing Switch Ports**

Additional security methods, such as port binding, port lockdown, port lockout, and persistent port disable, can be enabled on switch ports in addition to zoning and LUN masking. Only the appropriate switch port can connect to a node for fabric access, and port binding limits the number of devices that can attach to a particular switch port. WWPN spoofing is reduced, but not eliminated, by port binding. 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. Some variants ensure that the port role is restricted to only FL-Port, F-Port, E-Port, or a combination of these. Persistent port disable prevents a switch port from being enabled even after a switch reboot.

**Switch-Wide and Fabric-Wide Access Control**

The requirement to adequately manage SAN security grows as enterprises expand their SANs locally or over longer distances. Access

control lists (ACLs) on the FC switch and fabric binding on the fabric can be used to ensure network security.

The device connection control and switch connection control policies are both included in ACLs. The device connection control policy defines which HBAs and storage ports are allowed to connect to the fabric, preventing illegal devices from doing so. The switch connection control policy, likewise, specifies which switches are permitted to join the fabric, prohibiting illegal switches from doing so.

Role-based access control enhances SAN security by blocking unauthorised management actions on the fabric. It allows the security administrator to grant roles to users when they log into the fabric, allowing them to designate specified privileges or access permissions. The zone admin job, for example, has the ability to edit the fabric's zones, whereas a basic user can only see fabric-related data like port types and logged-in nodes.

**Logical Partitioning of a Fabric: Virtual SAN**

VSANs allow you to create numerous logical SANs from a single physical SAN. They enable the construction of bigger consolidated fabrics while maintaining the needed security and isolation. In a VSAN, logical partitioning is depicted in Figure

By populating each VSAN with switch ports, the SAN administrator can construct unique VSANs. The switch ports are split among two VSANs in this example: 10 and 20 for the Engineering and HR divisions, respectively. Despite the fact that they share physical switching equipment with other divisions, they can be controlled as separate 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.

Because management and control traffic on the SAN — which may include RSCNs, zone set activation events, and more — does not cross VSAN boundaries, VSANs reduce the impact of disruptive events. As a result, VSANs are a cost-effective way to create segregated physical fabrics. By isolating fabric events and giving permission control within a single fabric, they contribute to information availability and security.

**13.5.2 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.

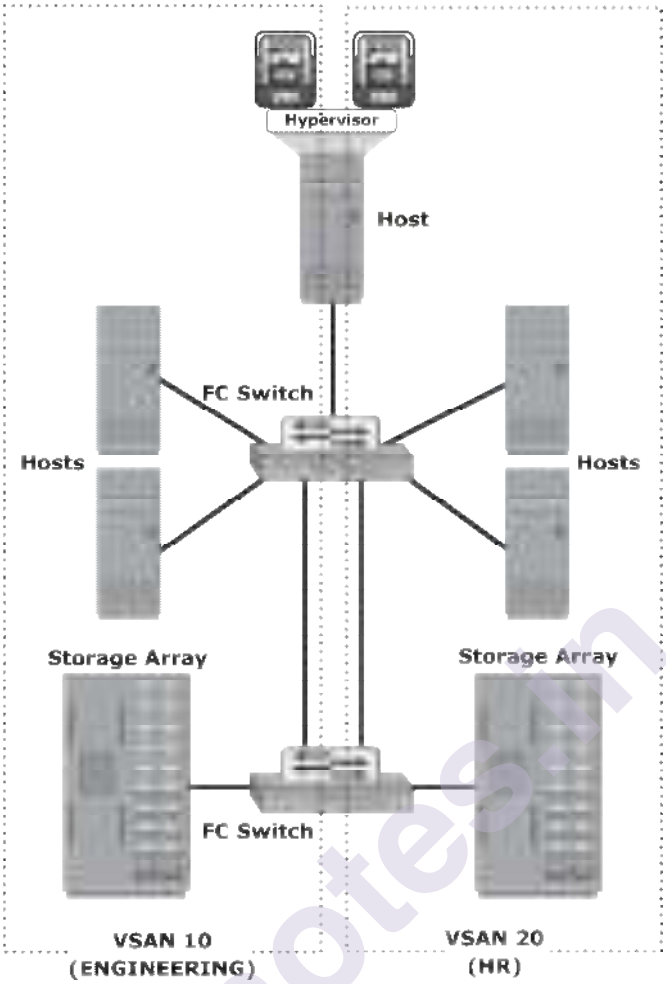



**Fig. : Securing SAN with VSAN**

Permissions and access control lists (ACLs) are the first line of defence for NAS resources, limiting access and sharing. These permissions are applied in addition to the behaviours and attributes that come standard with files and folders. Other authentication and authorisation systems, such as Kerberos and directory services, are also used to confirm network users' identities and define their privileges. Firewalls safeguard storage infrastructure from unauthorised access and malicious assaults in the same way.

**NAS File Sharing: Windows ACLs**

ACLs are divided into two categories in Windows: discretionary access control lists (DACLs) and system access control lists (SACLs) (SACLs). The access control is determined by the DACL, sometimes known as the ACL. If auditing is enabled, the SACL specifies which accesses must be audited.

Windows also supports the concept of object ownership in addition to these ACLs. The owner of an item has hard-coded permissions to that object, which are not need to be granted explicitly in the SACL. Each

object's owner, SACL, and DACL are all statically stored as attributes. Windows also offers the functionality to inherit permissions, which allows the child objects existing within a parent object to automatically inherit the ACLs of the parent object.

ACLs are also applied to directory objects known as security identifiers (SIDs). These are automatically generated by a Windows server or domain when a user or group is created, and they are abstracted from the user. In this way, though a user may identify his login ID as "User1," it is simply a textual representation of the true SID, which is used by the underlying operating system. Internal processes in Windows refer to an account's SID rather than the account's username or group name while granting access to an object. ACLs are set by using the standard Windows Explorer GUI but can also be configured with CLI commands or other third-party tools.

**NAS File Sharing: UNIX Permissions**

A user is an abstraction in the UNIX operating system that defines a logical entity for assigning ownership and operation privileges to the system. A user might be a person or a computer programme. Regardless of whether it is a person, a system action, or a device, a UNIX system is only aware of the user's privileges to execute specific operations on the system and identifies each user by a user ID (UID) and a username. In UNIX, users can be organized into one or more groups. The concept of group serves the purpose to assign sets of privileges for a given resource and sharing them among many users that need them. For example, a group of people working on one project may need the same permissions for a set of files.

UNIX permissions define the operations that any ownership relation can do with respect to a file. To put it another way, these permissions define what the owner, the owner group, and everyone else can do with the file. Three bits are used to describe access permissions for any given ownership relationship. Read (r) access is indicated by the first bit, write (w) access by the second bit, and execute (x) access by the third bit. Because UNIX defines three ownership relations (Owner, Group, and All), each ownership relationship requires a triplet (defining the access permission), resulting in nine bits. Each bit has two options: set or clear. When displayed, a set bit is marked by its corresponding operation letter (r, w, or x), a clear bit is denoted by a dash (-), and all are put in a row, such as rwxr-xr-x. In this example, the owner can do anything with the fi le, but group owners and the rest of the world can read or execute only. When displayed, a character denoting the mode of the fi le may precede this nine-bit pattern. For example, if the fi le is a directory, it is denoted as "d"; and if it is a link, it is denoted as "l."

**NAS File Sharing: Authentication and Authorization**

Standard file-sharing protocols, such as NFS and CIFS, are used by NAS devices in a file-sharing environment. As a result, authentication

and authorization on NAS devices are implemented and supported in the same way they are in a UNIX or Windows file sharing environment.

Authentication needs a login credential lookup on a Network Information System (NIS) server in a UNIX environment to validate the identity of a network user. A Windows domain controller, which houses the Active Directory, authenticates a Windows client in the same way. 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  depicts the authentication process in a NAS environment.

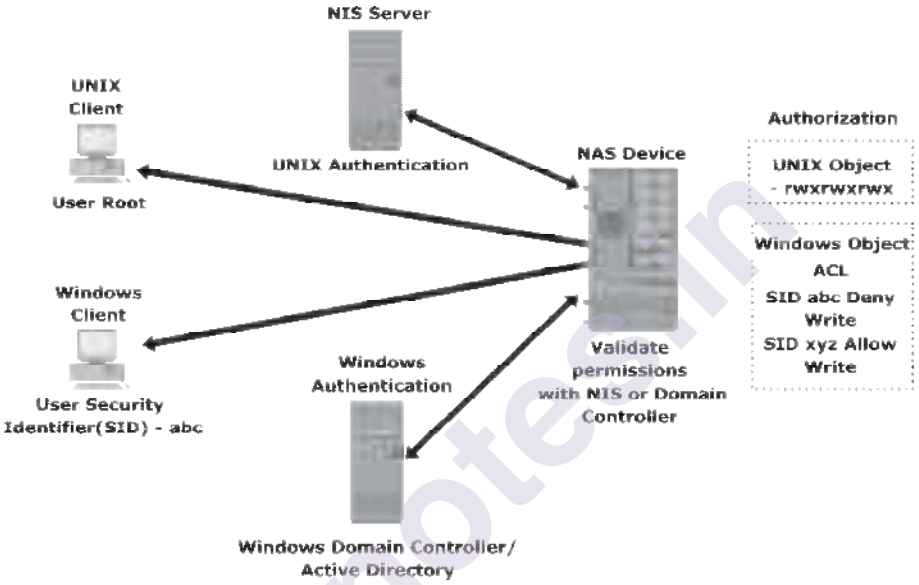



**Fig: Securing user access in a NAS environment**

User privileges in a network are defined by authorization. The authentication methods used by UNIX and Windows users are very different. UNIX files employ mode bits to define access rights for owners, groups, and other users, but Windows files use an ACL to allow or deny certain permissions to a specific user for a specific file.

NAS devices allow both of these approaches for UNIX and Windows users, however when UNIX and Windows users access and share the same data, complications develop. The integrity of both permission procedures must be preserved if the NAS device supports multiple protocols. A way of mapping UNIX rights to Windows and vice versa is provided by NAS device suppliers, allowing for a multiprotocol environment to be supported. When developing a NAS solution, keep in mind the complexity of multiprotocol support. Validate the domain controller and NIS server connectivity and bandwidth at the same time.
Kerberos

Kerberos is a network authentication protocol that uses secret-key cryptography to offer strong authentication for client/server applications. It employs cryptography to allow a client and server to establish their identity over an unsecured network connection. After proving their identities, the client and server might choose to encrypt all of their communications to preserve privacy and data integrity.

In Kerberos, authentications occur between clients and servers. The client gets a ticket for a service and the server decrypts this ticket by using its secret key. Any entity, user, or host that gets a service ticket for a Kerberos service is called a Kerberos client. 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. In Kerberos, users and servers for which a secret key is stored in the KDC database are known as principals.

Kerberos is usually used in a NAS context to authenticate against a Microsoft Active Directory domain, but it may also be used to perform security functions in UNIX settings. The steps in the Kerberos authentication procedure are represented in Fig. :

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. (This step is not explicitly shown in Figure)
2. The KDC responds with an encrypted Ticket Granting Ticket (TGT) and an encrypted session key. TGT has a limited validity period. TGT can be decrypted only by the KDC, and the client can decrypt only the session key.
3. When the client requests a service from a server, it sends a request, consisting of the previously generated TGT, encrypted with the session key 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 hosting the service.
6. The client then sends the service ticket to the server that houses the required 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 tracks the client activity, such as file locking, as long as the session is active.

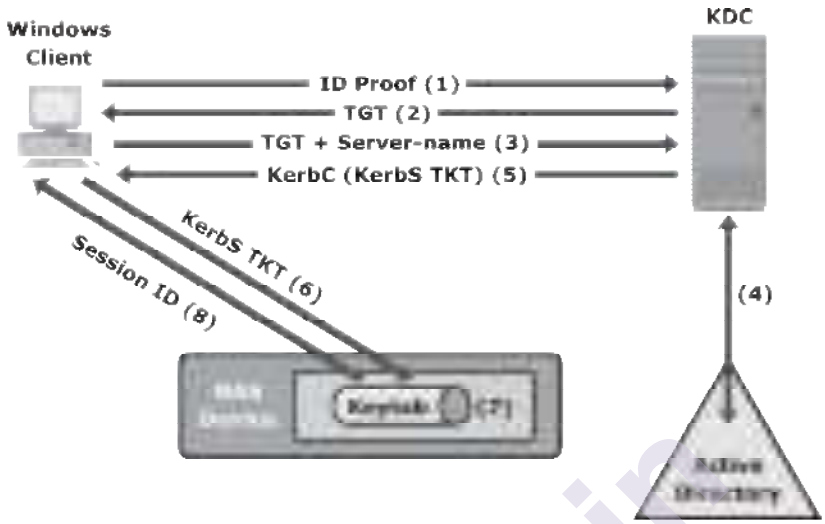



**Figure Kerberos authorization**

**Network-Layer Firewalls**

Because NAS systems use the IP protocol stack, they are susceptible to a variety of assaults launched via the public IP network. In NAS setups, network layer firewalls are used to defend the NAS devices from various security concerns. These network-layer firewalls can inspect network packets and compare them to a set of security rules that have been established. Packets that do not comply with a security rule are dropped and do not proceed to their destination. A source address (network or host), a destination address (network or host), a port, or a combination of those parameters can be used to create rules (source IP, destination IP, and port number).

Figure depicts a typical firewall implementation. A demilitarized zone (DMZ) is commonly used in networking environments. A DMZ provides a means to secure 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.

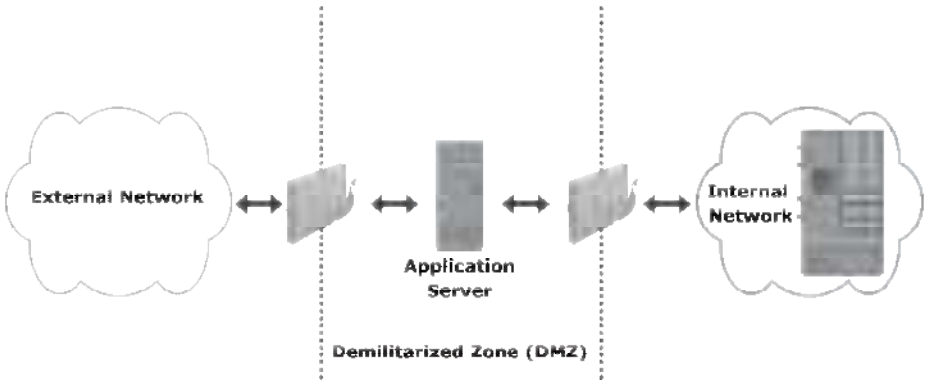



**Figure : Securing a NAS environment with a network-layer firewall**

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 accessing data stored on a NAS device, which may be located on the internal private network. A secure design would serve only data to internal and external applications through the DMZ.

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 accessing data stored on a NAS device, which may be located on the internal private network. A secure design would serve only data to internal and external applications through the DMZ.

### 13.5.3 IP SAN

The basic security measures utilised in IP SAN settings are described in this section. The Challenge-Handshake Authentication Protocol (CHAP) is a fundamental authentication technique that network devices and hosts have extensively embraced. CHAP uses a secret code or password to allow initiators and targets to verify each other's identity. CHAP secrets are usually 12 to 128 character long and random. The secret is never transferred directly over the communication channel; instead, it is converted into a hash value and then exchanged using a one-way hash function. 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 depicts the CHAP authentication process.

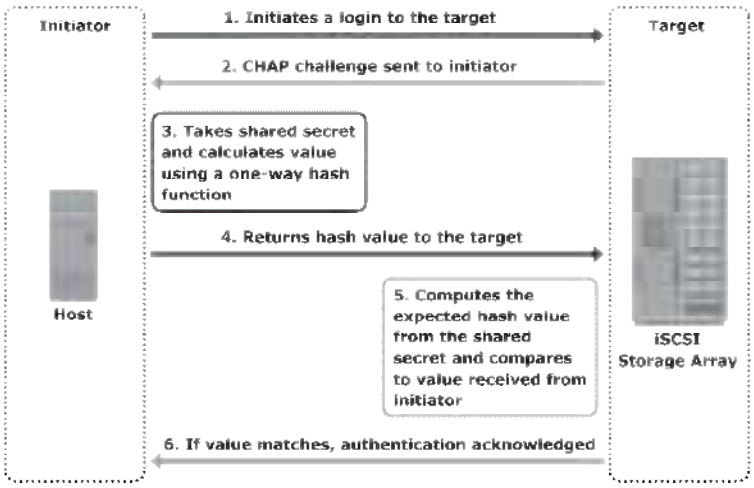



**Fig. A Securing IPSAN with CHAP authentication**

If the initiator requires reverse CHAP authentication, the initiator uses the same technique to authenticate the target. The initiator and the target must both know the CHAP secret. The target and the initiator each keep a CHAP entry that contains the name of a node and the secret connected with it. 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. For devices to communicate with one another, they must be configured in the same discovery domain. State change notifications (SCNs) inform the iSNS server when devices are added to or removed from a discovery domain. Figure B depicts the discovery domains in iSNS.

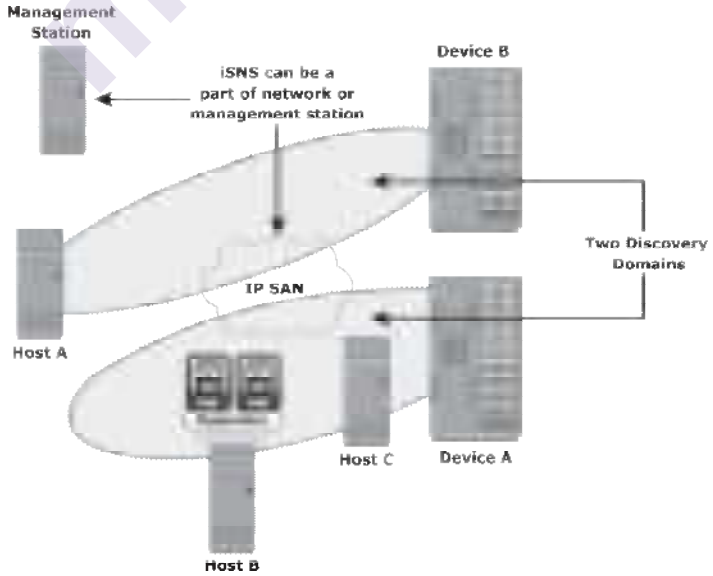



**Fig. B: Securing IPSAN with iSNS discovery domains**

## 13.6 SUMMARY

The continuing expansion of the storage network has exposed data center resources and storage infrastructures to new vulnerabilities. IP-based storage networking has exposed storage resources to traditional network vulnerabilities. Data aggregation has also increased the potential impact of a security breach. In addition to these security challenges, compliance regulations continue to expand and have become more complex. Data center managers are faced with addressing the threat of security breaches from both within and outside the organization.

Organizations are adopting virtualization and cloud as their new IT model. However, the key concern preventing faster adoption is security. The cloud has more vulnerabilities compared to a traditional or virtualized data center. This is because cloud resources are shared among multiple consumers. Also the consumers have limited control over the cloud resources. Cloud service providers and consumers are facing threat of security breaches in the cloud environment.

This chapter detailed a framework for storage security and provided mitigation methods that can be deployed against identified threats in a storage networking environment. It also detailed the security architecture and protection mechanisms in SAN, NAS, and IP-SAN environments. Further, this chapter touched on the security concerns and measures in a virtualized and cloud environment.

## 13.7 REVIEW YOUR LEARNING

- Can explain information security framework?
- Explain RISK TRIAD
- Can you write STORAGE SECURITY DOMAINS
- Explain Securing the Application Access Domain
- How to Protecting the Storage Infrastructure
- Explain details some of the basic security implementations in FC SAN, NAS, and IP-SAN environments

## 13.8 QUESTIONS

1. What is Information Security Framework
2. Explain part of the risk triangle.
3. Write a short note on
    a. Assets
    b. Threats
    c. Vulnerability
4. Explain how storage devices that are connected to a network increase the risk level.
5. Describe protecting the storage infrastructure.

6. How securing the management access domain
7. Explain FC SAN Security Architecture
8. Explain how VSANS allow you to create numerous logical sans from a single physical SAN
9. How NAS is open to multiple exploits, including viruses, worms, unauthorized access, snooping, and data tampering?
10. What are the various security mechanisms are implemented in NAS for storage networking infrastructure?
11. What Various security mechanisms are implemented in NAS to secure data?
12. Describe NAS File Sharing: Authentication and Authorization
13. Network-Layer Firewalls
14. How to securing IPSAN with CHAP authentication?

## 13.8 FURTHER READING

- https://nptel.ac.in/content/storage2/courses/106108058/lec%2007.pdf
- http://www.ictacademy.in/pages/Information-Storage-and-Management.aspx
- https://www.coursera.org/lecture/technical-support-fundamentals/storage-RLNIZ

## 13.9 REFERENCES


1. Information Storage and Management: Storing, Managing and Protecting Digital Information in Classic, Virtualized and Cloud Environments, EMC, John & Wiley Sons, 2nd Edition, 2012
2. Information Storage and Management, Pankaj Sharma




❖ ❖ ❖ ❖

# 14

# SECURING STORAGE INFRASTRUCTURE IN VIRTUALIZED AND CLOUD ENVIRONMENTS

**Unit Structure**





## 14.0 OBJECTIVES

After going through this chapter, you will be able to learn

- Security threads, Data storage in virtualized and cloud contexts, and RSA and VMware Security Products.
- Also learn monitoring the storage infrastructure

## 14.1 INTRODUCTION

So far, this chapter has solely discussed security threats and countermeasures in a traditional data centre. These hazards and countermeasures apply to data storage in virtualized and cloud contexts as well. However, due to multitenancy and a lack of control over cloud resources, virtualized and cloud computing environments pose new vulnerabilities to an organization's data. A public cloud poses higher security risks than a private cloud, necessitating additional safeguards.

Because cloud users (consumers) in a public cloud typically have limited control over resources, enforcing security methods by consumers is comparably difficult. From a security perspective, both consumers and cloud service providers (CSP) have several security concerns and face multiple threats. Security concerns and security measures are detailed next.

## 14.2 SECURITY CONCERNS

Virtualization and cloud computing are being increasingly adopted by businesses, but there are significant security risks. Multitenancy, attack velocity, information assurance, and data privacy are the four main security problems.By virtue of virtualization, multitenancy allows numerous independent tenants to share the same set of storage resources. Despite the advantages of multitenancy, it remains a major security problem for both users and service providers. The attack surface is increased when numerous VMs are co-located on a single server and share the same resources. It's possible that a tenant's business-critical data is accessed by other tenants running apps on the same resources.

Velocity-of-attack refers to a situation in which any 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. Also, the cloud user needs assurance that all the users operating on the cloud are genuine and access the data only with legitimate rights and scope.

Data privacy is also a major concern in a virtualized and cloud environment. A CSP needs to ensure that Personally Identifiable Information (PII) about its clients is legally protected from any unauthorized disclosure.

### 14.2 .2 Security Measures

At the computing, network, and storage layers, security mechanisms can be introduced. In virtualized and cloud systems, these security measures performed at three layers minimize the risks.

At the Compute Level, Security Securing a computational infrastructure entails ensuring the actual server, hypervisor, virtual machine, and guest operating system are all secure (OS running within a virtual machine). Implementing user authentication and authorization systems is part of physical server security. These techniques identify users and grant them server access privileges. . These mechanisms identify users and provide access privileges on the server. To minimize the attack surface on the server, unused hardware components, such as NICs, USB ports, or drives, should be removed or disabled.

For all the VMs running on it, a hypervisor is a single point of security failure. Rootkits and viruses put on a hypervisor make antivirus software deployed on the guest OS difficult to detect. Security-critical hypervisor updates should be installed on a regular basis to protect against attacks. In addition, the hypervisor management system must be safeguarded. Malicious assaults and management system infiltration can affect all current VMs and allow attackers to generate new ones. Only authorised administrators should have access to the management system. A second firewall must also be deployed between the management system and the rest of the network. VM isolation and hardening are some of the common security mechanisms to effectively safeguard a VM from an attack. VM isolation helps to prevent a compromised guest OS from impacting other guest OSs. VM isolation is implemented at the hypervisor level. Apart from isolation, VMs should be hardened against security threats. Hardening is a process to change the default configuration to achieve greater security.

Apart from the measures to secure a hypervisor and VMs, virtualized and cloud environments also require further measures on the guest OS and application levels.

**Security at the Network Level**

Firewalls, intrusion detection, demilitarised zones (DMZs), and data-in-flight encryption are some of the most important network security mechanisms.

A firewall prevents unauthorised access to networks while allowing only lawful communications. A firewall can also protect hypervisors and virtual machines (VMs) in a virtualized and cloud environment. If remote administration is allowed on a hypervisor, for example, a firewall should restrict access to all remote administration interfaces. VM-to-VM traffic is likewise protected by a firewall. A Virtual Firewall can be used to provide this firewall function (VF). A VF is a firewall service running entirely on the hypervisor. A VF provides packet filtering and monitoring of the VM-to-VM traffic. A VF gives visibility and control over the VM traffic and enforces policies at the VM level.

Intrusion Detection (ID) is the process to detect events that can compromise the confidentiality, integrity, or availability of a resource. An ID System (IDS) automatically analyses events to check whether an event or a sequence of events match a known pattern for anomalous activity, or whether it is (statistically) different from most of the other events in the system. It generates an alert if an irregularity is detected. DMZ and data encryption are also deployed as security measures in the virtualized and cloud environments. However, these deployments work in the same way as in the traditional data center.

**Security at the Storage Level**

Compromises at the compute, network, and physical security layers pose major vulnerabilities to storage systems in virtualized and cloud settings. This is due to the fact that storage systems are only accessible via computing and network infrastructure. To ensure storage security, suitable security mechanisms should be in place at the compute and network levels.

Common security mechanisms that protect storage include the following: nAccess control methods to regulate which users and processes access the data on the storage systems.Zoning and LUN-masking,Encryption of data-at-rest (on the storage system) and data-in-transit. Data encryption should also include encrypting backups and storing encryption keys separately from the data. nData shredding that removes the traces of the deleted data

Apart from these methods, employing VSANs to isolate different types of traffic improves the security of storage systems even further. In the case of hypervisor-based storage, additional security measures are required to safeguard the storage. Separate LUNs for VM components and VM data may be required for hypervisors using clustered file systems that support multiple VMs.

## 14.3 CONCEPTS IN PRACTICE: RSA AND VMWARE SECURITY PRODUCTS

RSA, EMC's security division, is the leading provider of security, risk, and compliance solutions, assisting businesses in overcoming their most difficult and sensitive security challenges.

For virtualized and cloud settings, VMware provides secure and reliable virtualization solutions. RSA SecureID, RSA Identity and Access Management, RSA Data Protection Manager, and VMware vShield are all covered in this area.

### 14.3.1 RSA SecureID

Two-factor authentication with RSA SecurID adds an extra layer of protection, ensuring that only authorised users have access to systems and data. RSA SecurID is based on two factors: what the user knows (password or PIN) and what the user owns (an authenticator device). It is far more trustworthy than reusable passwords when it comes to user authentication. Every 60 seconds, it generates a new one-time password code, making it difficult for anyone other than the legitimate user to enter the proper token code at any one time. Users combine their secret Personal Identification Number (PIN) with the token code that appears on their SecurID authenticator display at the time to gain access to their resources. The result is a unique, one-time password to assure a user's identity.

### 14.3.2 RSA Identity and Access Management

Through access management, the RSA Identity and Access Management product manages identity, security, and access restrictions for physical, virtual, and cloud-based environments. It allows trusted identities to engage with systems and access in a secure and free manner. RSA Access Manager and RSA Federated Identity Manager are two products in the RSA Identity and Access Management family. RSA Access Manager allows businesses to manage authentication and authorization policies for a large number of users, online web portals, and application resources from a single location. Access Manager provides seamless user access with single sign-on (SSO) and preserves identity context for greater security. RSA Federated Identity Manager enables end users to collaborate with business partners, outsourced service providers, and supply-chain partners or across multiple offices or agencies all with a single identity and logon.

### 14.3.3 RSA Data Protection Manager

RSA Data Protection Manager enables deployment of encryption, tokenization, and enterprise key management simply and affordably. The RSA Data Protection Manager family is composed of two products: Application Encryption and Tokenization and Enterprise Key Management. nApplication Encryption and Tokenization with RSA Data Protection Manager helps to achieve compliance with regulations related to PII by quickly embedding the encryption and tokenization of sensitive data and helping to prevent data loss. It begins at the point of creation and ensures that data is encrypted throughout transmission and storage. n Enterprise key management is a simple solution for encrypting keys at the database, file server, and storage levels. Its goal is to make encryption deployment in the organisation as simple as possible. It also assists in ensuring that information is adequately secured and accessible at all times during its life cycle.

### 14.3.4 VMware vShield

The VMware vShield family includes three products:

- **vShield App**

VMware vShield App is a hypervisor-based application-aware firewall solution. It protects applications in a virtualized environment from network-based threats by providing visibility into network communications and enforcing granular policies with security groups. VMware vShield App observes network activity between virtual machines to define and refine firewall policies and secure business processes through detailed reporting of application traffic. For a virtualized environment.

- **vShield Edge**

VMware vShield Edge delivers comprehensive perimeter network protection. It is installed as a virtual appliance and acts as a network security gateway for all virtualized hosts. It offers a variety of services,

including firewall, VPN, and DHCP (Dynamic Host Configuration Protocol).

- **vShield Endpoint.**

VMware vShield Endpoint is a hardened special security VM with antivirus software from a third party. VMware vShield Endpoint streamlines and accelerates antivirus and antimalware deployment because antivirus engine and signature fi les are updated only within the special security VM. VMware vShield Endpoint improves VM performance by offloading file scanning and other tasks from VMs to the security VM. It prevents antivirus storms and bottlenecks associated with multiple simultaneous antivirus and antimalware scans and updates. It also satisfies audit requirements with detailed logging of antivirus and antimalware activities.

## *14.4 MONITORING* THE STORAGE INFRASTRUCTURE

Monitoring is one of the most important aspects that forms the basis for managing storage infrastructure resources. Monitoring provides the performance and accessibility status of various components. It also enables administrators to perform essential management activities. This monitoring also aids in the analysis of storage infrastructure resource use and consumption. This research aids capacity planning, forecasting, and the most efficient use of these resources. The ambient characteristics of the Storage infrastructure, such as heating and power supply, are also monitoring.

### 14.4.1 Monitoring Parameters

Accessibility, capacity, performance, and security should all be monitored in storage infrastructure components. The availability of a component to perform its desired action throughout a given time period is referred to as accessibility. Checking the availability status of hardware components (for example, a port, an HBA, or a disc drive) or software components (for example, a database) requires evaluating the warnings issued by the system. A port failure, for example, may result in a cascade of availability warnings.

To avoid a single point of failure, a storage infrastructure employs redundant components. A component failure might result in an outage that affects application availability, or it could result in performance deterioration even if accessibility is not affected. Continuously monitoring each component's anticipated accessibility and reporting any deviations aids the administrator in identifying malfunctioning components and planning remedial action to meet SLA criteria.

The quantity of storage infrastructure resources available is referred to capacity. Examining the free space available on a file system or a RAID group, mailbox quotas assigned to users, or the number of ports accessible on a switch are all examples of capacity monitoring.

Insufficient capacity causes performance degradation or possibly application/service outage. By preventing failures before they happen, capacity monitoring assures continuous data availability and scalability. For example, if 90% of the ports in a given city are in use,If more arrays and servers need to be placed on the same fabric, this might suggest that a new switch is needed. Analytical tools are commonly used in capacity monitoring to do trend analysis. These patterns assist in determining future resource requirements as well as estimating deployment times.

Performance monitoring assesses the efficiency of various storage infrastructure components and aids in the identification of bottlenecks. Performance monitoring examines and assesses behaviour in terms of response time or capacity to execute at a certain level. It also deals with resource usage, which has an impact on how resources behave and respond. Performance measurement is a difficult process that entails evaluating multiple components based on a number of interconnected factors. The total number of Disk I/O, application response time, and network use are all factors to consider and server-CPU usage are two examples of performance metrics that should be considered.

Unauthorized access, whether inadvertent or intentional, may be tracked and prevented by monitoring a storage infrastructure for security. Unauthorized configuration modifications to storage infrastructure resources can be tracked with security monitoring. Security monitoring, for example, keeps track of and reports on the initial zoning configuration as well as any later changes.

### 14.4.2 Components Monitored

The components of the storage environment that should be monitored for accessibility, capacity, performance, and security include hosts, networks, and storage. Physical or virtualized components can be used.

**Hosts**

The accessibility of a host is determined by the state of availability of its hardware components and software processes. A host's NIC failure, for example, might render the host inaccessible to its users. Server clustering is a system that ensures high availability in the event of a server failure.

It's critical to keep track of a host's file system capacity usage to ensure that apps have enough storage space. The loss of file system space causes application availability to be disrupted. Monitoring aids in estimating the rate of expansion of the file system and predicting when it will reach 100%.As a result, the administrator can proactively expand the file system's space (manually or automatically) to avoid application downtime. Virtual provisioning technology is used.

Allows for efficient storage capacity management, although it is extremely reliant on capacity monitoring.

The basic goal of host performance monitoring is to keep track of how much of various server resources, such as CPU and memory, are being used. For example, if a server running an application consistently sees 80 percent CPU use, the server may be running out of processing power, resulting in decreased performance and slower response times. Administrators can address the issue by upgrading or adding additional CPUs, as well as redistributing the burden across various servers. To satisfy performance needs in a virtualized environment, more CPU and memory may be dynamically assigned to VMs from the pool, if available.

On servers, security monitoring entails recording login failures as well as the execution of illegal programmers or software activities. The danger identified informs proactive actions against unauthorized access to the systems. An administrator, for example, can disable a user's access if numerous login failures are recorded.

**Storage Network**

To guarantee that communication between the server and the storage array is not disrupted, storage networks must be monitored. Access to data across the storage network is contingent on the physical and logical components of the storage network being accessible. Switches, ports, and cables are the physical components of a storage network. Constructs such as zones are among the logical components. Data is unavailable when one or more physical or logical components fail. Errors in zoning, such as supplying the wrong WWN for a port, result in the port being unable to be accessed, possibly preventing a host from accessing its storage.

The number of available ports in the fabric, the usage of the inter switch connections, or individual ports, and each interconnect device in the fabric are all monitored during capacity monitoring in a storage network. Capacity monitoring offers all of the necessary data for future fabric resource planning and optimization.

Monitoring the storage network's performance allows you to analyze individual component performance and detect network bottlenecks. Monitoring port performance, for example, entails calculating the receiver or transmit link usage metrics, which show how busy the switch port is. I/O queuing on the server might be caused by heavily utilizing ports, resulting in poor performance.

Network latency, packet loss, bandwidth usage for I/O, network faults, packet retransmission rates, and collisions are all things to keep an eye on when it comes to IP networks.

Storage network security monitoring detects any unauthorized modifications to the fabric's configuration, such as changes to zone policies that might compromise data security. Login failures and illegal access to switches for administrative modifications should be continually documented and monitored.

**Storage**

For its physical components and different operations, the storage array's accessibility should be checked. Individual component failure does not normally impair the accessibility of storage arrays because they are often built with redundant components. Failure of any process in the storage array, on the other hand, might cause business activities to be disrupted or jeopardized. The failure of a replication task, for example, has an impact on disaster recovery capabilities. If hardware or process issues occur, certain storage arrays include the ability to transmit messages to the vendor's support center, known as a call home.

A storage array's capacity monitoring allows the administrator to anticipate storage demands based on capacity usage and consumption patterns. The administrator can use information regarding unconfirmed and unallocated storage space to determine if a new server can be given storage capacity from the storage array.

Various performance measures, such as utilization rates of various storage array components, I/O response time, and cache utilization, can be used to monitor a storage array. A component of a storage array that is overworked, for example, may cause performance deterioration.

A storage array is often a shared resource that is vulnerable to security attacks. Monitoring security helps track unlawful storage array configuration and guarantees that only authorized users have access to it.

### 14.4.3 Monitoring Examples

A storage infrastructure necessitates the installation of an end-to-end system that actively monitors all of its components' parameters. Early detection and preemptive alerting guarantee that essential assets continue to provide uninterrupted service. Furthermore, the monitoring tool should assess the consequences of a failure and determine the root cause of symptoms.

**Accessibility Monitoring**

Because of their linkages and interdependence, the failure of one component might influence the accessibility of another. Consider the following scenario: H1, H2, and H3 are three servers in a storage system. Each and every server as illustrated in Figure A, are configured with two HBAs, each linked to the production storage array by two switches, SW1 and SW2. On the storage array, all of the servers share two storage ports, and multipath software is installed on all of them

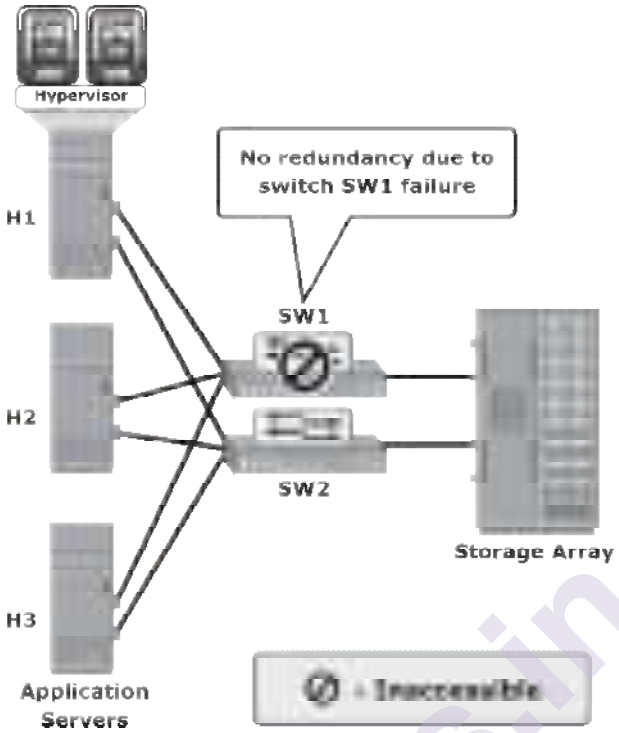



Figure A: Switch failure in a storage infrastructure

The multipath software begins a path failover if one of the switches (SW1) fails, and all of the servers continue to access data through the other switch, SW2. However, because there is no redundant switch, a second switch failure may render the array inaccessible. Monitoring for accessibility allows an administrator to identify a switch failure and take necessary action before another happens.

In most situations, the administrator is notified of a failing component's symptoms and can take action before it fails.

**Capacity Monitoring**

Servers H1, H2, and H3 are linked to the production array by two switches, SW1 and SW2, in the situation depicted in Figure B Each of the servers is unique

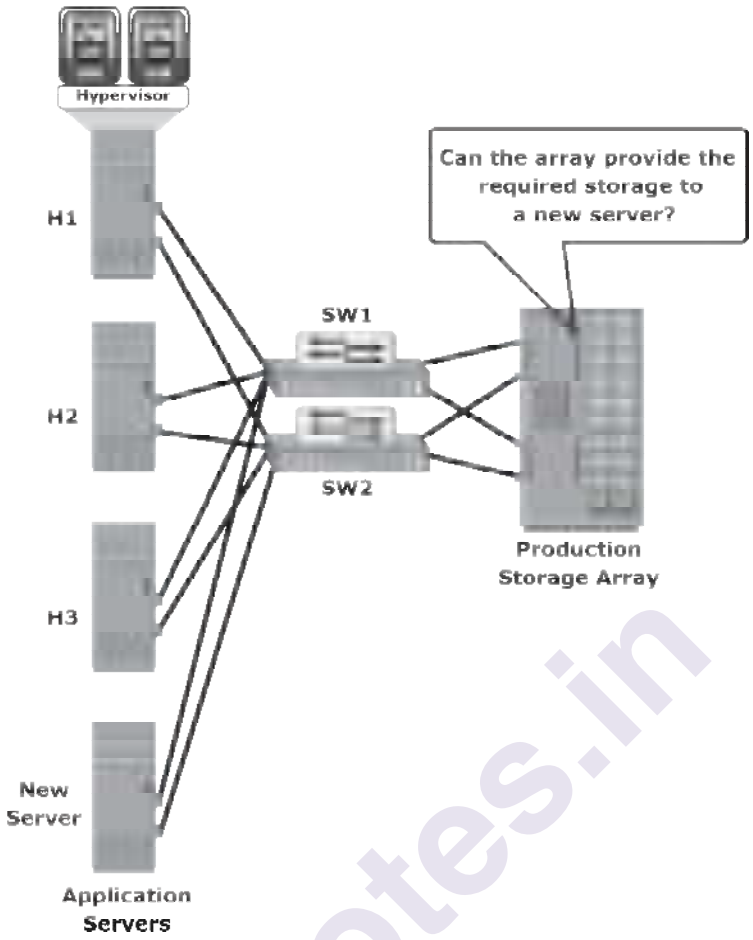
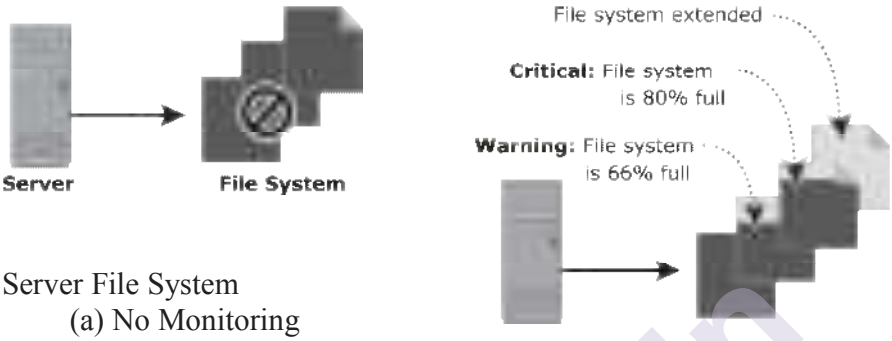



**Figure B: Monitoring storage array capacity**

Is the amount of storage on the storage array that has been assigned. In this configuration, when a new server is deployed, the applications on the new server must be provided storage capacity from the production storage array. Monitoring the array's available capacity (configurable and unallocated) allows you to decide ahead of time if the array will be able to offer enough storage for the new server. Also, monitoring the number of ports available on SW1 and SW2 also aids in determining whether the new server can be connected to the switches.

The following example demonstrates the significance of fi le system capacity monitoring on fi le servers. Figure C depicts the environment of a file system when it is full, resulting in application outage when capacity is not available.

Monitoring is in place. When capacity thresholds on the file system are surpassed, monitoring may be trusted to send a notification. A warning message is sent when the file system reaches 66 percent of its capacity, and a critical message is sent when the file system reaches 80 percent of its capacity (see Figure C. This allows the administrator to take steps to expand the file system before it reaches its maximum capacity. Monitoring the file system in advance can help prevent application outages caused by a shortage of file system capacity.




Server File System

(a) No Monitoring

(b) File System Monitoring

Figure C: Monitoring server file system space

**Performance Monitoring**

Figure shows how important it is to keep track on the performance of storage arrays. Switches SW1 and SW2 link servers H1, H2, and H3 (each with two HBAs) to the storage array in this example. To access LUNs, the three servers use the same storage ports on the storage array. . A new server with a high-workload application must be installed to use the same storage port as H1, H2, and H3.

Monitoring array port utilization guarantees that the new server does not have a negative impact on the other servers' performance. The solid and dotted lines in the graph depict use of the shared storage port in this case. If the port usage prior to deploying the new server is close to 100%, then installing the new server is not advised since it may have an influence on the existing server's performance.

The other servers' performance However, if the port's usage prior to the new server's deployment is closer to the dotted line, there is room to add a new server.

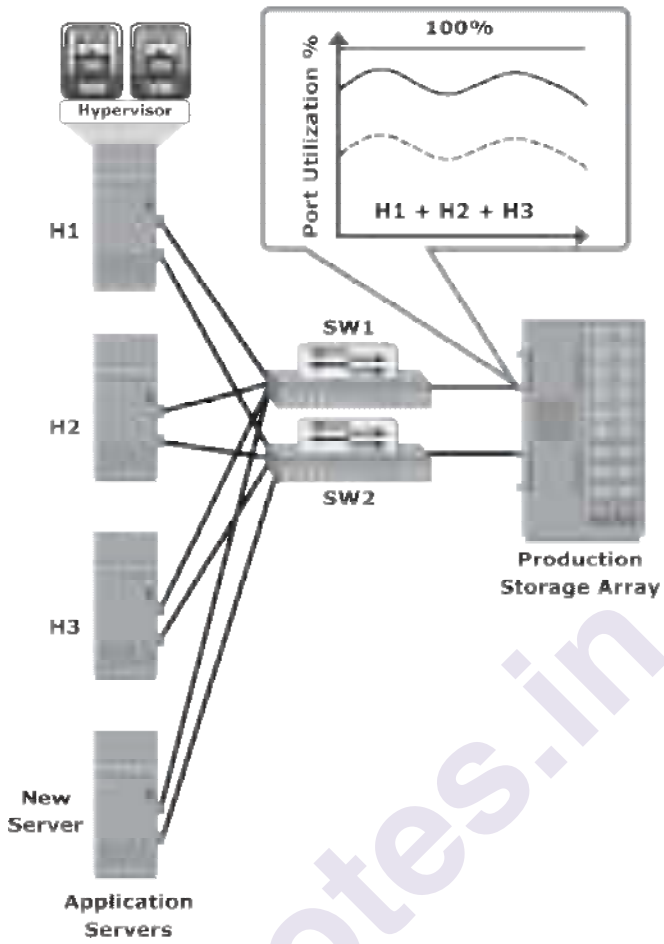



**Figure: Monitoring array port utilization**

Most servers come with tools that allow you to keep track of your server's CPU utilization. For example, as illustrated in Figure 14-5, Windows Task Manager displays CPU and memory utilization. These technologies, on the other hand, are ineffective in monitoring hundreds of servers in a data center. Intelligent performance monitoring solutions capable of simultaneously monitoring numerous servers are required in a data center setting.

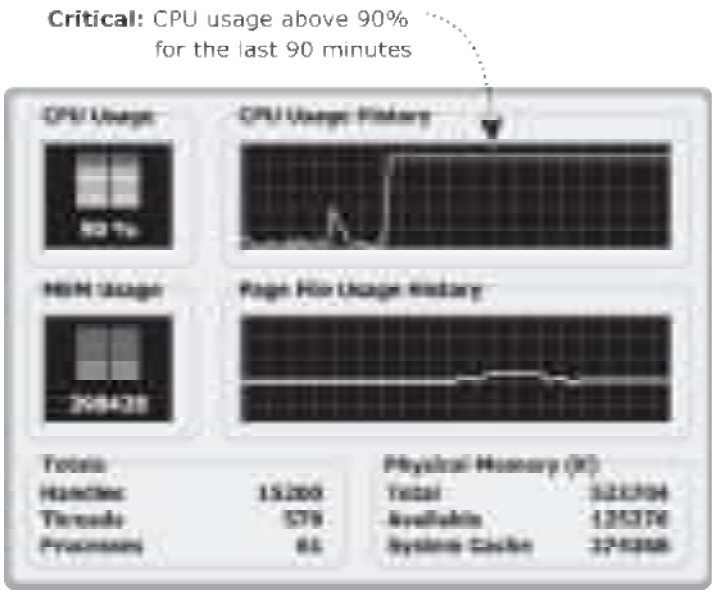


Figure: Monitoring the CPU and memory usage of a server

**Security Monitoring**

Figure shows how important it is to keep an eye on the security of a storage array.

The storage array is shared in this case by two workgroups, WG1 and WG2. WG1's data should not be available to WG2, and vice versa. A user from WG1 may attempt to create a local duplicate of data from WG2. It is difficult to detect such a breach of information security if the action is not observed or recorded. If this activity is monitored, a warning message can be delivered to suggest a remedial action or at the very least enable detection as part of routine auditing.

The tracking of login attempts at the host is an example of host security monitoring. If the login ID and password entered are correct, the login is permitted; otherwise, the login attempt fails. If this activity is monitored, a warning message can be provided to suggest a remedial action or, at the very least, allow detection as part of routine auditing activities. Monitoring login attempts at the host is an example of host security monitoring. If the login ID and password are correct, or if the login attempt fails, the login is permitted.

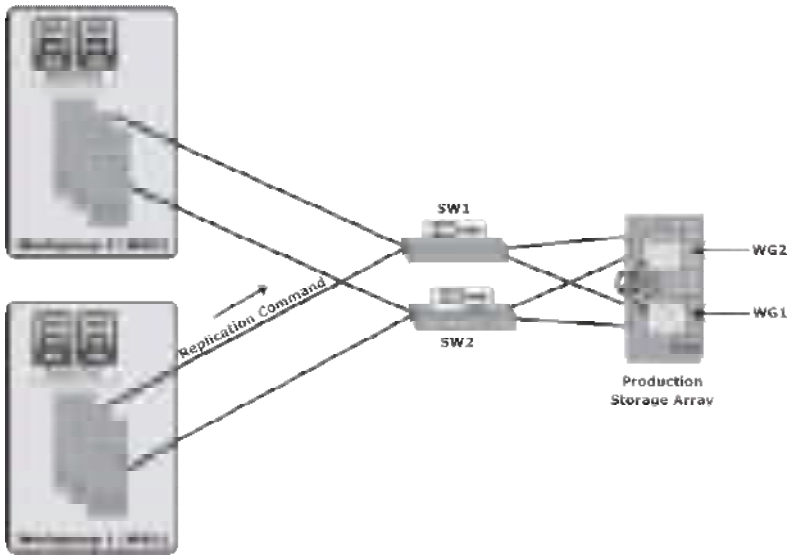



**Figure Monitoring security in a storage array**

### 14.4.4 Alerts

Event alerting is an important element of monitoring. Alerting keeps administrators informed about the state of different components and processes — for example, failures of power, discs, memory, or switches, which might have a negative influence on service availability and necessitate prompt administrative action. Other events, such as a fi le system approaching its capacity limit or a soft media fault on discs, are considered warning indicators and may necessitate administrator intervention.

Administrators can use monitoring tools to designate severity levels based on the severity of the detected issue. An alert is sent to the administrator, a script is launched, or an incident ticket is filed if a situation with a specific severity level occurs. Informational notifications to life-threatening warnings are all possible. Information alerts give useful information without requiring the administrator's participation. An example of an information alert is the establishment of a zone or LUN. Warning alerts necessitate administrative action in order to contain the notified circumstance. Accessibility is unaffected. If an alert shows that the number of soft media defects on a disc is reaching a predefined threshold value, the administrator can determine whether the disc should be replaced. Because the situation may compromise overall performance, security, or availability, fatal alerts demand rapid response. If a disc fails, for example, the administrator must ensure that it is replaced as soon as possible.

Administrators can respond quickly and proactively to problems thanks to continuous monitoring and automatic alerts. The information provided by alerting assists administrators in prioritizing their reaction to occurrences.

## 14.5 QUESTIONS

1. What are the layers minimize the risks in virtualized and cloud systems.
2. State and explain most important network security mechanisms.
3. Write a short note on **RSA SecureID**
4. What are the default products included in VMware vShield family.
5. Explain why monitoring is one of the most important aspects that forms the basis for managing storage infrastructure resources?
6. Why components of the storage environment that should be monitored?

## 14.6 FURTHER READING

- https://sites.google.com/site/testwikiforfirstciscolab/shd/14-securing-the-storage-infrastructure
- http://www.ictacademy.in/pages/Information-Storage-and-Management.aspx

## 14.7 REFERENCES


1. Information Storage and Management: Storing, Managing and Protecting Digital Information in Classic, Virtualized and Cloud Environments, EMC, John & Wiley Sons, 2nd Edition, 2012
2. Information Storage and Management, Pankaj Sharma


# 15

# STORAGE INFRASTRUCTURE MANAGEMENT ACTIVITIES

**Unit Structure**





## 15.0 OBJECTIVES:

After completing this chapter you will be able to :

- Lean major storage infrastructure components that should be monitored.
- Lean Storage Infrastructure Management Challenges and its solutions
- Also learn Storage Tiering

## 15. 1 INTRODUCTION

Unprecedented data growth, application proliferation, business process complexity, and need Information growth, application proliferation, business process complexity, and needs have never been greater. Information availability 24 hours a day, seven days a week has increased the demand on storage infrastructure.

Managing storage infrastructure efficiently is essential for companies to solve these issues and maintain business continuity. To reach the needed service level, comprehensive storage infrastructure management necessitates the use of intelligent tools and procedures. Performance twseaking, data protection, access control, central auditing, and satisfying compliance requirements are all possible with these technologies. They also guarantee that current resources are consolidated and efficiently utilizing reducing the need for continuous infrastructure investment. The management process establishes processes for addressing diverse activities, such as incidents, problems, and change requests, in an efficient manner. Because of the interdependence of the components, it is critical to manage not only the individual components, but the infrastructure as a whole.

Information Life-cycle Management (ILM), for example, is a storage infrastructure management strategy that optimizing storage investment while achieving service requirements. ILM assists in the management of information depending on its business value.

Managing storage infrastructure necessitates a variety of tasks, including access, capacity, performance, and security management. All of these actions are intertwined and should be examined in order to achieve the best results rate of return on investment. The paradigm of storage infrastructure management has shifted drastically as a result of virtualization technologies.

The return on investment monitoring and control of storage infrastructure is covered in this chapter. It also explains the industry requirements for creating storage resource management software. This chapter also covers ILM, its advantages, and storage tiering.

## 15.2 STORAGE INFRASTRUCTURE MANAGEMENT ACTIVITIES

The growing complexity of managing storage infrastructures is due to the rapid expansion of information, proliferation of applications, heterogeneous infrastructure, and high service-level requirements. Storage virtualization and additional technologies, including as data duplication and compression, virtual provisioning, federated storage access, and storage tiring, have, nevertheless, emerged.

Availability management, capacity management, performance management, security management, and reporting are some of the major storage infrastructure management tasks done in a data center.

### 15.2.1 Availability Management

Establishing a good guideline based on specified service levels to assure availability is a crucial responsibility in availability management. All availability-related issues for components or services are managed as part of availability management to guarantee that service standards are fulfilled. Provisioning redundancy at all levels, including components, data, and even locations, is a crucial task in availability management. When a server is installed to support a key business function, for example, high availability is required. Two or more HBAs, multipath software, and server clustering are typically used to accomplish this. At least two separate fabrics and switches with built-in redundancy must be used to link the server to the storage array. Furthermore, storage arrays should include built-in redundancy for various components and enable both local and distant replication.

### 15.2.2 Capacity Management

The objective of capacity management is to guarantee that resources are available in sufficient quantities to meet service level requirements. Capacity management also include capacity optimization based on cost and future requirements. On a regular basis, capacity management does a capacity analysis that compares allocated storage to anticipated storage.

It also offers trend analysis based on consumption rates, which must be balanced against storage acquisition and deployment schedules. Capacity management is an example of storage provisioning. It entails tasks like establishing RAID sets and LUNs and assigning them to the host. Another form of capacity management is enforcing capacity quotas for users. By allocating a fixed number of user quotas, users are prevented from exceeding the given capacity.

Data duplication and compression technologies have decreased the quantity of data that has to be backed up, and therefore the amount of storage space that needs to be managed.

### 15.2.3 Performance Management

All components' operating efficiency is ensured by performance management. Performance analysis is a crucial activity that aids in the identification of storage infrastructure component performance. This study determines whether or not a component achieves its performance targets.

When deploying a new application or server in an existing storage system, certain performance management tasks must be completed. Every component must be verified to ensure that it meets the service level requirements for performance. For example, operations on the server, such as volume configuration and database architecture, can be optimized to achieve the desired performance levels. It is necessary to fine-tune the

application layout, the configuration of numerous HBAs, and clever multipath software. Designing and deploying sufficient ISLs in a multi switch fabric with sufficient bandwidth to achieve the desired performance levels are among the performance management responsibilities on a SAN. When considering end-to-end performance, storage array configuration duties include selecting the proper RAID type, LUN layout, front-end ports, back-end ports, and cache configuration.

### 15.2.4 Security Management

The security management activity's main goal is to maintain information confidentiality, integrity, and availability in both virtualized and non-virtualized settings. Unauthorized access to storage infrastructure components is prevented by security management. Security administration duties, for example, include maintaining user accounts and access policies that enable users to execute role-based activities when deploying an application or a server. Configuring zoning to prevent an illegal HBA from accessing particular storage array ports is one of the security management responsibilities in a SAN system. In a SAN system, security management duties include configuring zoning to prevent an illegal HBA from accessing certain storage array ports. Similarly, a storage array's security management duty includes LUN masking, which limits a host's access to only the LUNs that are intended.

### 15.2.5 Reporting

Keeping track of and obtaining data from multiple components and processes is part of reporting on a storage system. Trend analysis, capacity planning, chargeback, and performance reports are generated using this data. Capacity planning reports provide current and historical statistics on storage, file systems, database tables-pace, ports, and other resources. Device allocation, local or distant replicas, and fabric configuration are all included in the configuration and asset management reports. This report also includes a detailed inventory of all the equipment, including purchase dates, lease status, and maintenance records. Detailed information regarding the performance of various storage infrastructure components may be found in performance reports.

### 15.2.6 Storage Infrastructure Management in a Virtualized Environment

The complexity of storage infrastructure management has been substantially reduced thanks to virtualization technologies. In reality, the flexibility and simplicity of management of virtualization at all layers of the IT infrastructure are major reasons for its widespread adoption.

Storage virtualization has allowed for dynamic data transfer and storage volume expansion. Storage volumes may be dynamically extended to suit capacity and performance needs without causing any disruptions. Data may be moved both inside and between data centers because virtualization removes the link between the storage volumes displayed to the host and its actual storage. While reconfiguring the physical environment, this has made the administrator's responsibilities easier.

Another innovation that has altered the infrastructure management cost and complexity picture is virtual storage provisioning. Storage capacity is allocated in advance in traditional provisioning in anticipation of future expansion. Because growth is uneven, some users or apps may reach capacity limits, while others may have surplus capacity that goes unused. Virtual provisioning can help to solve this problem and make capacity management easier. Storage is allocated from the shared pool to hosts on-demand in virtual provisioning. This enhances storage capacity usage and, as a result, simplifies capacity management.

Network management efficiency has also benefited from virtualization. VSANs and VLANs make the administrator's job simpler by conceptually isolating separate networks rather than physically separating them using management tools.

On a same physical network, many virtual networks may be established, and node reconfiguration can be done fast without any physical modifications. It also addresses some of the security concerns that could arise in a traditional setting.

Compute virtualization on the host side has made host deployment, configuration, and migration easier than in a physical environment. Virtualization of compute, application, and memory resources has enhanced provisioning while also contributing to high resource availability.

### 15.2.7 Storage Management Examples

Examples of various storage management actions are provided in the next section.

Example 1: Storage Allocation to a New Server/Host

Consider adding a new RDBMS server to your existing non-virtualized storage environment. Before the server is physically connected to the SAN, the administrator must first install and configure the HBAs and device drivers as part of storage management operations. Multi-path software can be installed on the server as an option, although this may need additional configuration. The SAN should also be linked to storage array ports.

The administrator must then execute zoning on the SAN switches in order to grant the new server access to the storage array ports through its HBAs. The new server's HBAs should be connected to various switches and zoned with distinct array ports to guarantee redundant routes between the server and the storage array.

In addition, the administrator must configure LUNs on the array and allocate these LUNs to the front-end ports of the storage array. In addition, the storage array is configured using LUN masking, which blocks access to LUNs by a single server.

Depending on the operating system used, the server then finds the LUNs assigned to it via a bus rescan procedure or, in certain cases, a server reboot. A volume manager can be used to manage the host's logical volumes and file systems. The number of logical volumes or file systems that must be built is determined by how the storage will be used by a database or application. The administrator's job also include setting up a database or application on the newly formed logical volumes or file systems. The final step is to enable the database or program to use the additional file system space. The actions conducted on a server, a SANs, and a storage array for the allocation of storage to a new server are depicted in Figure

Provisioning storage to a VM that runs an RDBMS in a virtualized environment necessitates distinct administrative activities. A physical link must be created between the physical server that hosts the VMs and the storage array through the SANs, just like in a non-virtualized environment. A VSAN may be trusted to transmit data between the physical server and the storage array at the SANs level. The VSAN separates this storage traffic from the rest of the SAN's traffic. Additionally, the administrator can set up zoning within the VSANs.

Administrators must build thin LUNs from the shared storage pool and assign these thin LUNs to the storage array front-end ports on the storage side. LUNs masking is required on the storage array, just as it is in a physical environment.

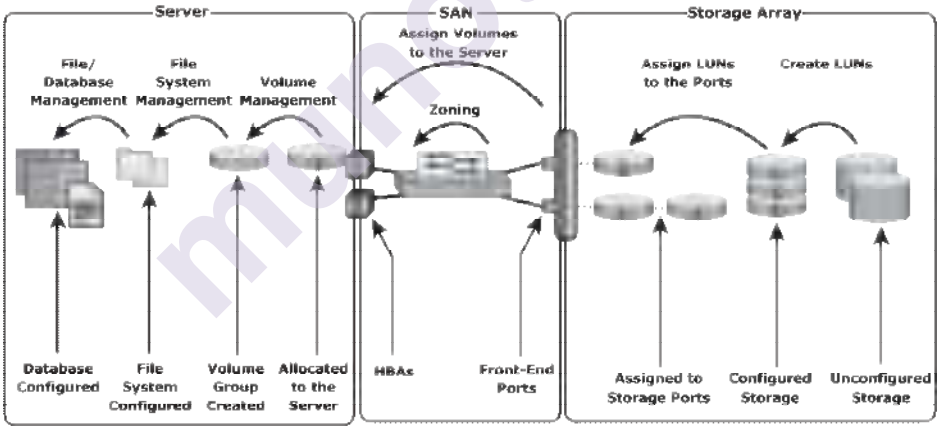



Figure Storage allocation tasks

The hypervisor detects the allocated LUNs on the physical server. To store and manage VM files, the hypervisor constructs a logical volume and file system. The administrator next builds a virtual machine (VM) and installs the operating system and database management system on it. The hypervisor produces a virtual disc file and other VM files in the hypervisor file system when constructing the VM. The RDBMS data is stored on the virtual disc file, which appears to the VM as a SCSI drive. Alternatively, the hypervisor can activate virtual provisioning and allocate a thin virtual

disc to the VM. Multi-paths generally available natively on hypervisors. A third-party multi-path program can be loaded on the hypervisor if desired.
Example 2: File System Space Management

Administrators must do activities to offload data from an existing fi le system to avoid a fi le system from running out of capacity. This might involve removing files that are no longer needed or archiving data that hasn't been viewed in a long time.

Alternatively, an administrator can expand the size of the file system by extending it and prevent a service interruption. A dynamic extension of fi le systems, often known as.

The operating system or logical volume manager (LVM) in use determines the logical volume. In the flow chart, Figure  depicts the procedures and considerations for extending fi le systems.

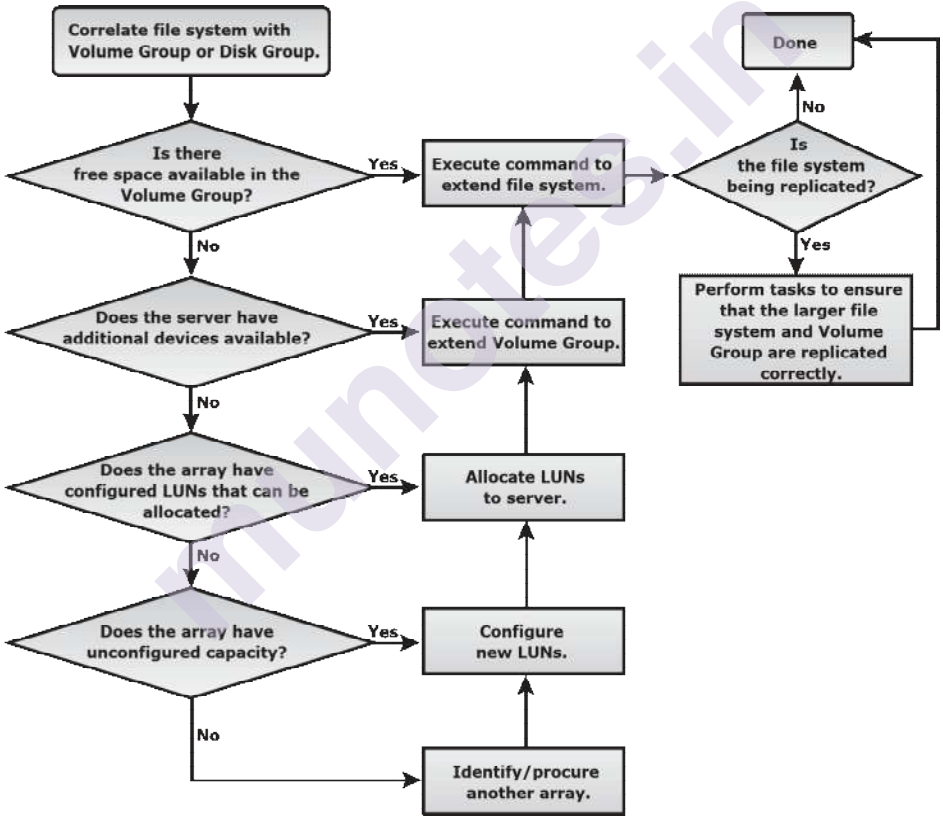



Figure : Extending a file system

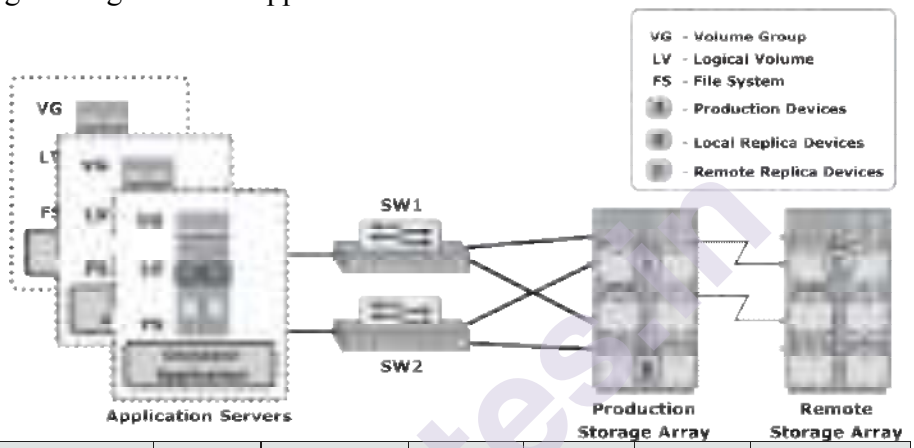
Example 3: Chargeback Report

The storage infrastructure management activities required to produce a chargeback report are explored in this example.

A configuration in a storage infrastructure is shown in Figure . Three Two switches, SW1 and SW2, link servers with two HBAs to a storage array.SW1 and SW2 On each of the servers, individual

departmental applications operate. Local and distant duplicates are created using array replication technology. The production device is labelled A, the local duplicate device is labelled B, and the final device is labelled C.

A chargeback analysis for each department is used to generate a report detailing the precise amount of storage resources consumed by each application. If the billing unit is based on the quantity of raw storage (usable capacity plus protection supplied) estimated for a department's application, the precise amount of raw space estimated for each application must be disclosed. A example report is shown in Figure  The information about two people is shown in the report. Payroll 1 and Engineering 1 are two apps.




| Application | Storage (GB) | Production Storage Raw (GB) | Local Replica Storage Raw (GB) | Remote Replica Storage Raw (GB) | Total Storage Raw (GB) | Chargeback Cost $ 5/Raw (GB) |
|---|---|---|---|---|---|---|
| Payroll_1 | 100 | 200 | 100 | 125 | 425 | $ 2125 |
| Engineering_1 | 200 | 250 | 200 | 250 | 700 | $ 3500 |

Figure : Chargeback report

The first step in calculating chargeback charges is to match the application to the precise quantity of raw storage required for that application.

The Payroll 1 application storage space is tracked from file systems to logical volumes, volume groups, and LUNs on the array, as shown in Figure . The storage space utilized for local replication and distant replication is also identified when the applications are duplicated. The program uses Source Vol1 and Vol2 in the example given (in the production array).

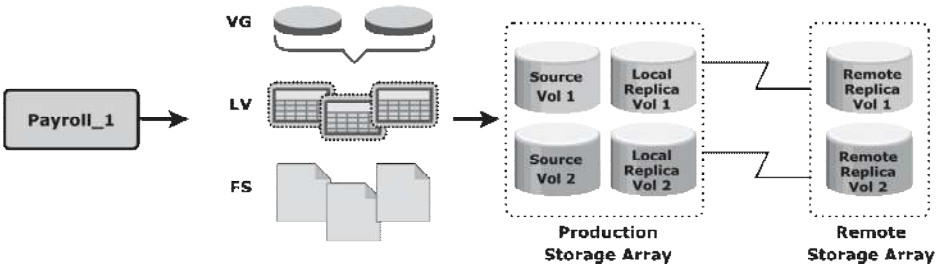



Figure : Correlation of capacity configured for an application

After the array devices have been identified, the amount of storage assigned to the program may be readily calculated. Consider the following scenario: Source Vol1 and Vol2 are both 50 GB in size, and the application is allotted 100 GB (50 + 50)

For local replication, 100 GB is allotted, while for distant replication, 100 GB is allotted. Based on the RAID protection that is utilize for various array devices, the raw storage configured for the application is calculated from the assigned storage.

The raw space needed by the production volumes of the Payroll 1 application is 200 GB if they are RAID 1-protected. If the local copies are on unprotected volumes and the distant replicas are protected using a RAID 5 configuration, the local replica will utilized 100 GB of raw space and the remote replica will use 125 GB. As a result, the Payroll 1 program uses a total raw capacity of 425 GB. The overall storage cost for the Payroll 1 application will be $2,125 (assuming a $5 per GB storage cost). To create the chargeback report, this process must be done for each application in the business.

Chargeback reports can be expanded to include the cost of additional resources in the setup, such as the number of switch ports, HBAs, and array ports. Data center managers utilized chargeback reports to ensure that storage customers are informed of the prices of the services they have requested.

## 15.3 STORAGE INFRASTRUCTURE MANAGEMENT CHALLENGES

It's difficult to keep track of and manage today's complicated storage infrastructure. This is due to the environment's variety in terms of storage arrays, networks, servers, databases, and applications. Heterogeneous storage arrays, for example, differ in terms of capacity, performance, security, and architecture.

Each component in a data center is generally equipped with vendor-specific management tools. Because the tools may not be compatible, understanding the overall condition of the environment is difficult in an environment with many tools in an ideal world,

management tools would bring together data from all components in one location. Such tools give a complete picture of the environment, as well as a speedier root cause analysis and resolution of alarms.

## 15.4 DEVELOPING AN IDEAL SOLUTION

An ideal solution would provide actionable information into the overall infrastructure's condition as well as root cause investigation for each failure. In a multi-vendor storage environment, this solution should also enable central monitoring and administration, as well as an end-to-end view of the storage infrastructure.

The capacity to connect one component's activity with the behaviour of another is a benefit of end-to-end monitoring. In many situations, examining each component separately may not be sufficient to determine the root of the problem. The central monitoring and management system should collect data from all components and handle it via a single user interface. It must also offer a way for administrators to be notified of various occurrences via e-mail and Simple Network Management Protocol (SNMP) traps. It should also be able to create monitoring reports and perform automatic task automation routines.

By utilize common APIs, data model language, and taxonomy, the optimal solution must be built on industry standards. This allows policy-based management to be implemented across heterogeneous devices, services, applications, and deployed topologies.

The SNMPs protocol was formerly the industry standard for managing multi-vendor SANs systems. SNMPs, on the other hand, was insufficient for delivering the level of detail necessary to control the SANs environment. The unavailability of automatic discovery functions and weak modeling constructs are some inadequacy of SNMPs in a SANs environment. Despite these drawbacks, SNMPs continues to play a significant role in SAN administration, even as newer open storage SAN management standards develop to better monitor and control storage settings.

### 15.4.1 Storage Management Initiative

The Storage Networking Industry Association (SNIA) has been working on a project to provide a standard storage management interface. Storage Management Initiative-Specification is a specification produced by SNIA (SMI-S). The Web-Based Enterprise Management (WBEM) technology and the Distributed Management Task Force's (DMTF) Common Information Model are used to create this standard. The goal of the effort was to provide extensive interoperability and administration across heterogeneous storage and SAN components. Visit www.snia.org for additional details.

Users and sellers alike will benefit from SMI-S. It creates a standardized, abstracted model to which the physical and logical components of a storage system may be mapped. This paradigm is used by management program for standardized, end-to-end control of storage resources, such as storage resource management, device management, and data management.

Device software developers may use SMI-S to provide a unified object model that includes data about controlling a wide range of storage and SAN components. SMI-S-compliant devices make policy-based storage management framework implementation and acceptance easier, quicker, and more widespread. Furthermore, SMI-S eliminates the requirement for manufacturers to build their own management interfaces, allowing them to focus on value-added functionality.

### 15.4.2 Enterprise Management Platform

An enterprise management platform (EMPs) is a collection of tools that work together to manage and monitor a company's storage infrastructure. These apps include unified frameworks that allow for end-to-end control of both real and virtual resources.

These apps can keep an eye on storage infrastructure components and send out alerts when something goes wrong. These warnings can be shown on a console with the defective component highlighted in a different color, or they can be set to send an e-mail. In addition to monitoring, an EMP includes administration capability, which can be built right into the EMPs or launched through the component manufacturer's own management application.

An EMP also makes it simple to schedule activities that must be done on a regular basis, such as resource provisioning, configuration maintenance, and problem investigation. To make storage infrastructure management easier, these systems include comprehensive analytical, remedial, and reporting capabilities. EMCs Control Center and EMCs pro-sphere are instances of EMPs, as stated in section 15.7 "Concepts in Practice."

## 15.5 INFORMATION LIFECYCLE MANAGEMENT

If information is not handled properly, it may be costly in both traditional data centers and virtualised settings. To handle information effectively, you'll need more than just the tools. You'll also need a good management plan. This strategy should address the following key challenges that exist in today’s data centers.

- Expand digital universe: Information is growing at an exponential rate. The multi-fold rise in information growth has been attributed to the creation of copies of data to ensure high availability and reuse.

- Increasing dependency on information: The strategic use of data is critical to a company's success and gives competitive advantages in the marketplace.
- Changing value information: Information that is useful today may be less valuable tomorrow. Information's worth fluctuates a lot throughout time.

Understanding the value of information throughout its life cycle is crucial to developing a plan to tackle these issues. When information is originally produced, it has the greatest value and is often accessed. Information becomes less valuable to the business as it matures and is accessed less regularly.

Understanding the value of information aids in the deployment of suitable infrastructure in response to changing information value.

For example, the value of information (customer data) in a sales order application varies from the moment the purchase is placed until the warranty is invalid . When a firm receives a new sales order and processes it to deliver the goods, the information has the greatest value. Customer data does not need to be available for real-time access once the order has been fulfilled. Until a warranty claim or another event necessitates its use, the firm can shift this data to less costly secondary storage with reduced performance. The firm can discard the information once the warranty has expired.

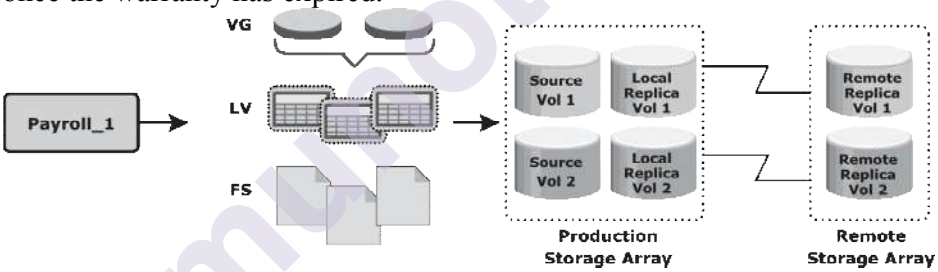



Figure: Changing value of sales order information

Information Life-cycle Management (ILM) is a proactive technique that allows an IT company to efficiently manage information throughout its life cycle while adhering to predefined business standards. ILM automates the alignment of business objectives and procedures with service levels from data generation through data disposal. This enables a company's storage system to be optimized for optimum return on investment. The following main benefits of implementing an ILM approach that directly address the issues of information management:

- Lower Total Cost of Ownership (TCO): Infrastructure and administrative expenses are aligned with the value of information. As a consequence, resources aren't squandered, and complexity isn't added by handling low-value data over high-value data.

- Simplified management: Process stages and interfaces with separate tools are integrated, and automation is increased.
- Maintaining compliance: By understanding what data must be protected for how long.
- Optimized utilization: By deploying storage tiering.

## 15.6 STORAGE TIERING

Storage tiering is a method of organizing several storage types into a hierarchy (tiers). This allows for the cost-effective storage of the correct data in the proper tier based on service level needs. Each tier offers varying degrees of security, performance, and pricing. High-performance solid-state drives (SSDs) or FC drives, for example, can be designated as tier 1 storage for often accessible data, whereas low-cost SATA drives can be designated as tier 2 storage for less frequently accessed data. Application performance is improved by storing frequently used data on SSD or FC. Moving less-frequently accessed data to SATA can free up storage capacity and lower storage costs in high-performance SSDs. This data flow is governed by defined tiering regulations. Tiering policies can be based on a variety of factors, including file type, size, frequency of access, and so on. For example, if a policy specifies, "Move the files that haven't been accessed in the previous 30 days to the lower tier," all files that meet this criteria are moved to the lower tier.

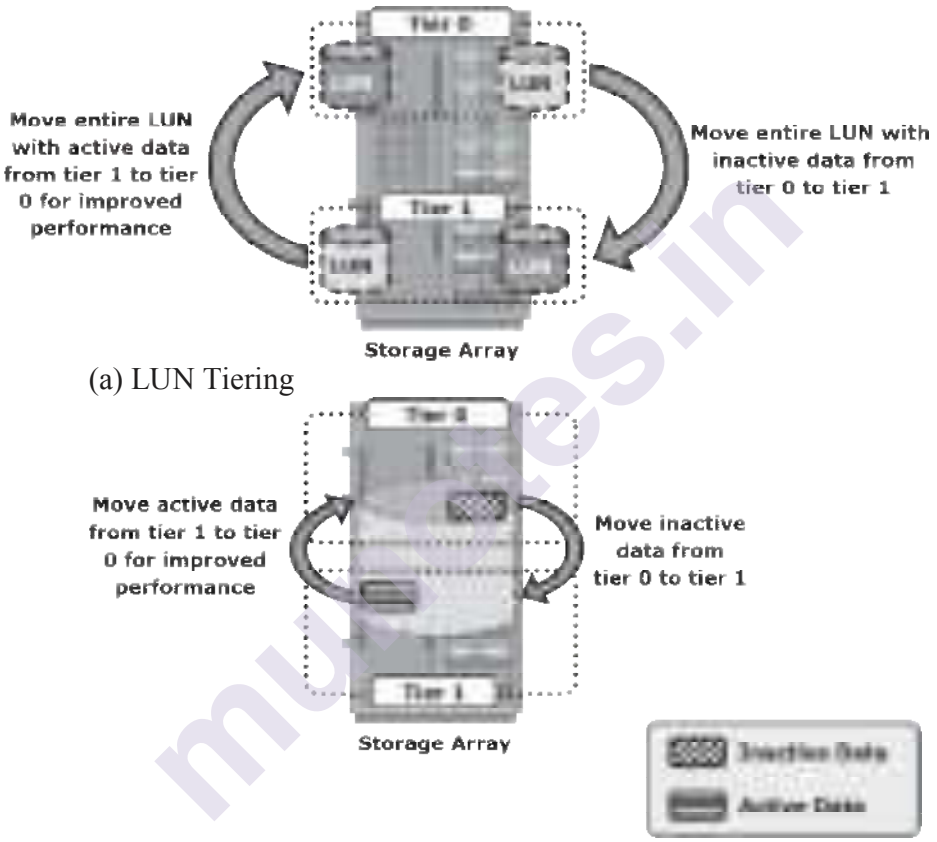
Storage tiering can be carried out manually or automatically. The conventional technique of manual storage tiering is for the storage administrator to monitor the storage workloads on a regular basis and shift the data across the tiers. Manual storage tiering is difficult and time-consuming. Automated storage tiering streamlines the storage tiering process by moving data across levels in a non-disruptive manner. The application workload is proactively monitored in automated storage tiering, and active data is automatically transferred to a higher performance tier, while idle data is automatically moved to a greater capacity, lower performance tier. Within (intra-array) or between (inter-array) storage arrays, data can be moved across levels.

### 15.6.1 Intra-Array Storage Tiering

Intra-array storage tiering is the technique of tiering storage within a storage array. It optimize speed and cost by allowing the efficient usage of SSD, FC, and SATA devices within an array. The objective is to keep SSDs busy by keeping the most often accessed data on them and transferring less frequently accessed data to SATA drives. The objective is to keep SSDs busy by keeping the most often accessed data on them and transferring less frequently accessed data to SATA drives. Data transfer between tiers can be done at the LUNs or sub-LUNs level. Implementing a layered cache can boost speed even further. The next sections cover LUNs tiering, sub-LUNs tiering, and cache tiering.

Storage tiering has traditionally been done at the LUNs level, when a whole LUNs is moved from one storage tier to another In that LUNs, this movement comprises both active and inactive data. This approach is ineffective in terms of cost and performance.

Storage tiering is now possible at the sub-LUNs level .A LUNs is split down into smaller parts and rated at that level in sub-LUNs level tiering. The value proposition of automated storage tiring is substantially enhanced when data is moved at a finer granularity, such as 8 MB. At the sub-LUNs level, tiring effectively transfers active data to faster drives while less active data is moved to slower drives.




(a) LUN Tiering

**(b) Sub-LUN Tiering**

**Figure : Implementation of intra-array storage tiering**

As illustrated in Figure  tiering can also be done at the cache level. A big cache in a storage array boosts speed by storing a large quantity of frequently requested data in the cache, allowing most reads to be delivered directly from it. Configuring a big cache in the storage array, on the other hand, is more expensive.

Utilizing the SSDs on the storage array is another option for increasing the cache capacity. SSDs are utilized as a large capacity secondary cache in cache tiering to allow tiering between DRAM (primary cache) and SSDs (secondary cache). Server flash-caching is a second tier of caching in which a flash-cache card is put in the server to improve the application's performance even more.

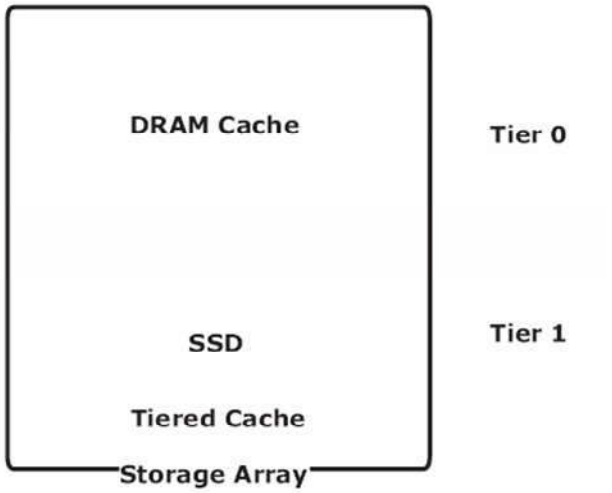



**Figure: Cache tiering**

### 15.6.2 Inter-Array Storage Tiering

Inter-array storage tiering is the process of tiering storage between storage arrays. Inter-array storage tiering automates the identification of active and inactive data in order to move it between the arrays to different performance or capacity tiers. . A two-tiered storage environment is illustrated in Figure . The primary storage is optimized for performance, while the secondary storage is optimized for capacity and cost. The policy engine, which may be software or hardware and is where rules are defined, allows data to be moved from main to secondary storage when it is inactive or seldom accessed. The policy engine produces a tiny space-saving stub fi le in the main storage for each archived fi le that refers to the data on the secondary storage. When a user attempts to access a file from its original location on main storage, the user is transparently given the file from secondary storage.

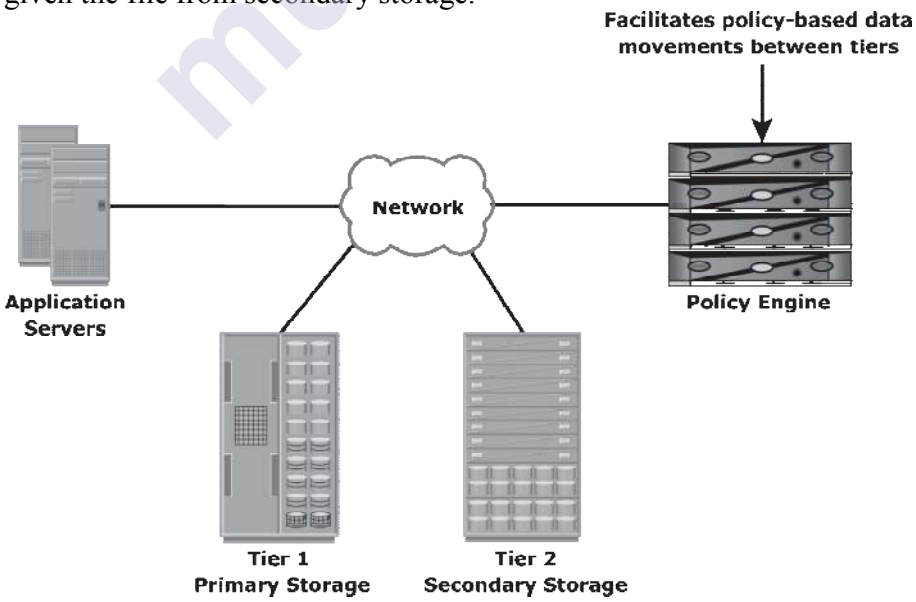



**Figure Implementation of inter-array storage tiering**

## 15.7 CONCEPTS IN PRACTICE: EMC INFRASTRUCTURE MANAGEMENT TOOLS

Due to the huge quantity of heterogeneous resources in today's world, businesses are having difficulty managing their IT infrastructure. Physical resources, virtualized resources, and cloud resources are all possibilities. EMC provides a variety of tools to meet a variety of business needs. EMC Control Center and Pro-Sphere are software suites that can control storage infrastructure from end to end, while EMC Unisphere is software that manages EMC storage arrays like VNX and VNX. The V-block infrastructure is managed using EMC Unified Infrastructure Manager (UIM) (cloud resources). Visit www.emc.com for additional details.

### 15.7.1 EMC Control Center and Pro-sphere

EMC Control-Center is a collection of storage resource management (SRM) products that work together to manage a multi-vendor storage environment. It aids in the management of a big, complicated storage environment that spans all layers and includes hosts, storage networks, storage, and virtualization.

Storage planning, provisioning, monitoring, and reporting are just a few of the features Control-Center offers. It supports the implementation of an ILM strategy by offering full storage infrastructure management. It also gives you a complete picture of your networked storage infrastructure, including SAN, NAS, and host storage resources, as well as a virtual environment. It has a central administration console, new component discovery, quota management, event management, root cause analysis, and charge back capabilities. Access control, data confidentiality, data integrity, logging, and auditing are all built-in security capabilities in Control-Center. It has a straightforward, user-friendly UI that provides insight into the ensuing complicated connections. To find the components in the environment, Control-Center employs an agent.

EMC Pro-Sphere is also storage resource management software designed to match the needs of today's cloud computing environment. In a virtual and cloud environment, EMC Pro-Sphere boosts productivity and service standards. The following major features are included in Pro-Sphere:

- End-to-end visibility: It gives insight into the intricate interactions between items in big, virtual systems using a simple, easy-to-use interface.
- Multi-site management: Pro-Sphere's federated design collects data from several sites and simplifies data administration between data centers from a single console. Pro-Sphere is controlled via a web browser, allowing for convenient remote management via the Internet.

- Improved productivity in growing virtual environments: Pro-Sphere provides Smart Organizes, a unique technology that groups items with similar characteristics into a user-defined group for administrative purposes. This allows IT to manage assets or create data collection policies using a policy-based approach.
- Fast, easy, and efficient deployment: Agent-less discovery eliminates the burden of deploying and managing host agents. Pro-Sphere is packaged as a virtual appliance that can be installed in a short time.
- Delivery of IT as a service: Service levels may now be monitored from the host to the storage layers with Pro-Sphere. This enables companies to maintain constant service levels at an appropriate price-performance ratio while offering IT-as-a-service to fulfil business objectives.

### 15.7.2 EMC Uni-sphere

EMC Uni-sphere is a unified storage management platform that lets you manage EMC VNX and EMC VNX storage arrays with simple user interfaces. Uni-sphere is web-based and allows storage arrays to be managed remotely. The following are some of Uni-sphere's major features:

- Offers unified storage management for files, blocks, and objects Allows all devices in a management domain to be accessed with a single sign-on.
- Supports automatic storage tiering and ensures that data is kept in the appropriate tier to maximize performance and minimize costs.
- Allows you to handle both real and virtual components.

### 15.7.3 EMC Unified Infrastructure Manager (UIM)

For V-blocks, EMC Unified Infrastructure Manager is a unified management solution. (Chapter 13 discusses V-block.) It allows for the configuration of V-block infrastructure resources as well as the activation of cloud services. It provides a single user interface for managing numerous V-blocks, removing the need to configure compute, network, and storage individually using various virtual infrastructure management systems. UIM provides a dashboard that displays how the V-block infrastructure is managed and resources are allocated. This allows an administrator to keep track of the configuration and usage of the V-block infrastructure resources, as well as plan for future capacity needs. A topology or map view of the V-block infrastructure is also provided by UIM, allowing an administrator to easily find and understand the linkages of the V-block infrastructure components and services. It has an alerts interface that allows administrators to view warnings for V-block infrastructure resources and associated services that have been affected by issues. Configuration. It validates compliance with configuration best practices. It also prevents conflicting resource identity assignments, for

example, accidentally assigning a MAC address to more than one virtual NIC

## 15.8 SUMMARY

The proliferation of data, its criticality, and organizations' rising reliance on digital data are resulting in bigger, more sophisticated storage systems. These infrastructures are becoming more difficult to manage. If a catastrophic failure happens, poorly managed storage systems can put the entire organization at risk.

The operations of monitoring and managing the storage infrastructure were covered in this chapter. This chapter also covered Information Life-cycle Management and its advantages, as well as storage tiering. Its advantages and storage tiering.

## 15.9 EXERCISES

1. Explain the storage infrastructure management activities in details.
2. What is the role of Storage Infrastructure Management in a Virtualized Environment
3. Write a short note on Storage Management with example.
4. What are the challenges of Storage Infrastructure Management.
5. Explain Information Lifecycle Management
6. Describe Storage Tiering in details.
7. Explain Intra-Array Storage Tiering
8. Explain Inter-Array Storage Tiering

## 15.10 REFERENCES

Information storage and management: storing, managing and protecting digital information in Classic, Virtualized and Cloud Environments, EMC author, by Joh Wiley and Sons 2nd edition 2012.