FUTURE VISION BIE

One Stop for All Study Materials

& Lab Programs

Future Vision

By K B Hemanth Raj

Scan the QR Code to Visit the Web Page

Or

Visit : https://hemanthrajhemu.github.io

Gain Access to All Study Materials according to VTU,
CSE – Computer Science Engineering,
ISE – Information Science Engineering,
ECE - Electronics and Communication Engineering
& MORE...

Join Telegram to get Instant Updates: https://bit.ly/VTU_TELEGRAM

Contact: MAIL: futurevisionbie@gmail.com

INSTAGRAM: www.instagram.com/hemanthraj_hemu/

INSTAGRAM: www.instagram.com/futurevisionbie/

WHATSAPP SHARE: https://bit.ly/FVBIESHARE

Introduction to Storage Area Network (SAN)

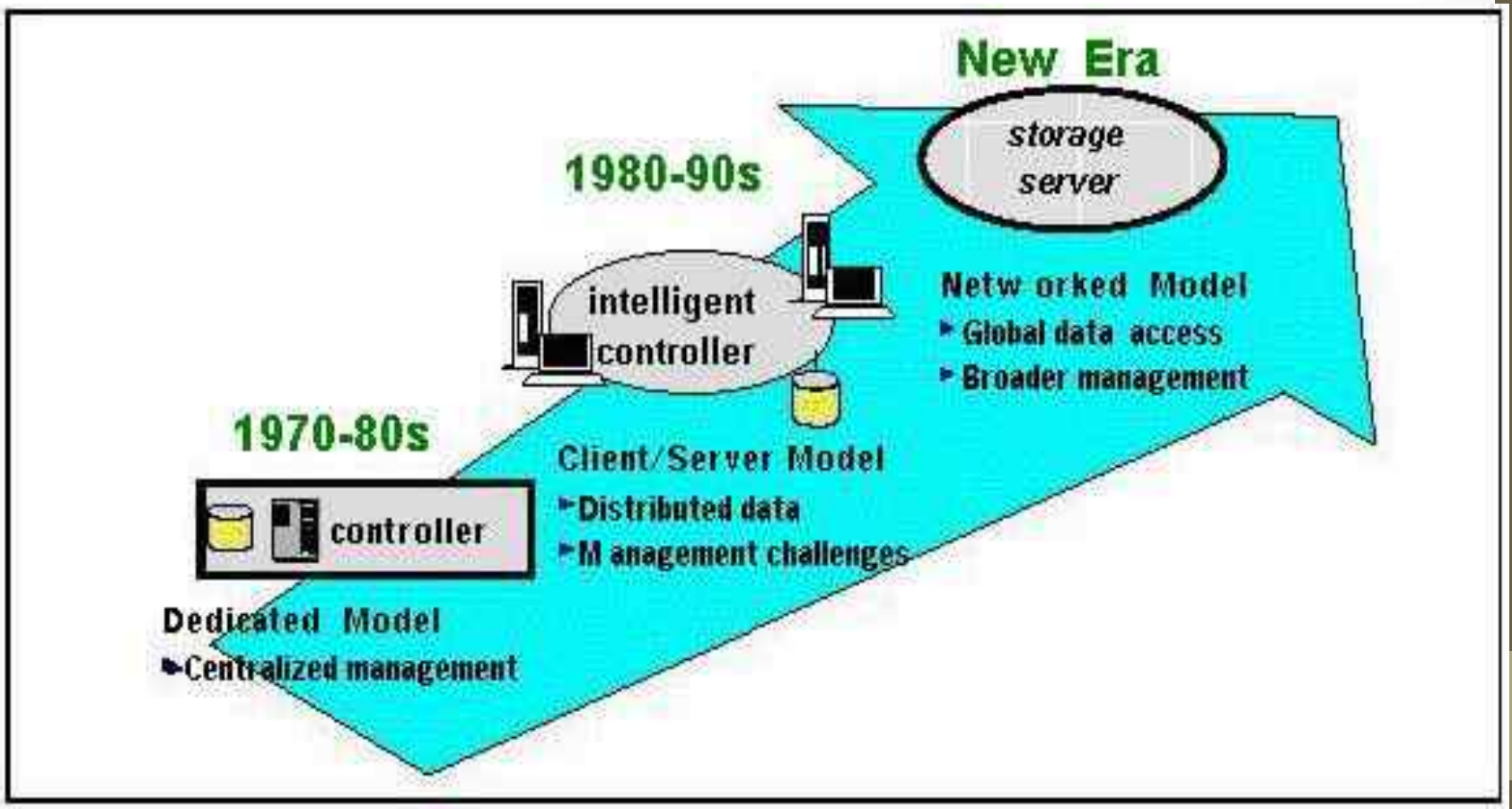
BMS
Institute of Technology and Management

https://hemanthrajhemu.github.io

What is SAN about

- Data is Asset
- How to Store Data?
- How to Access Data?
- How to Manage Data Storage?

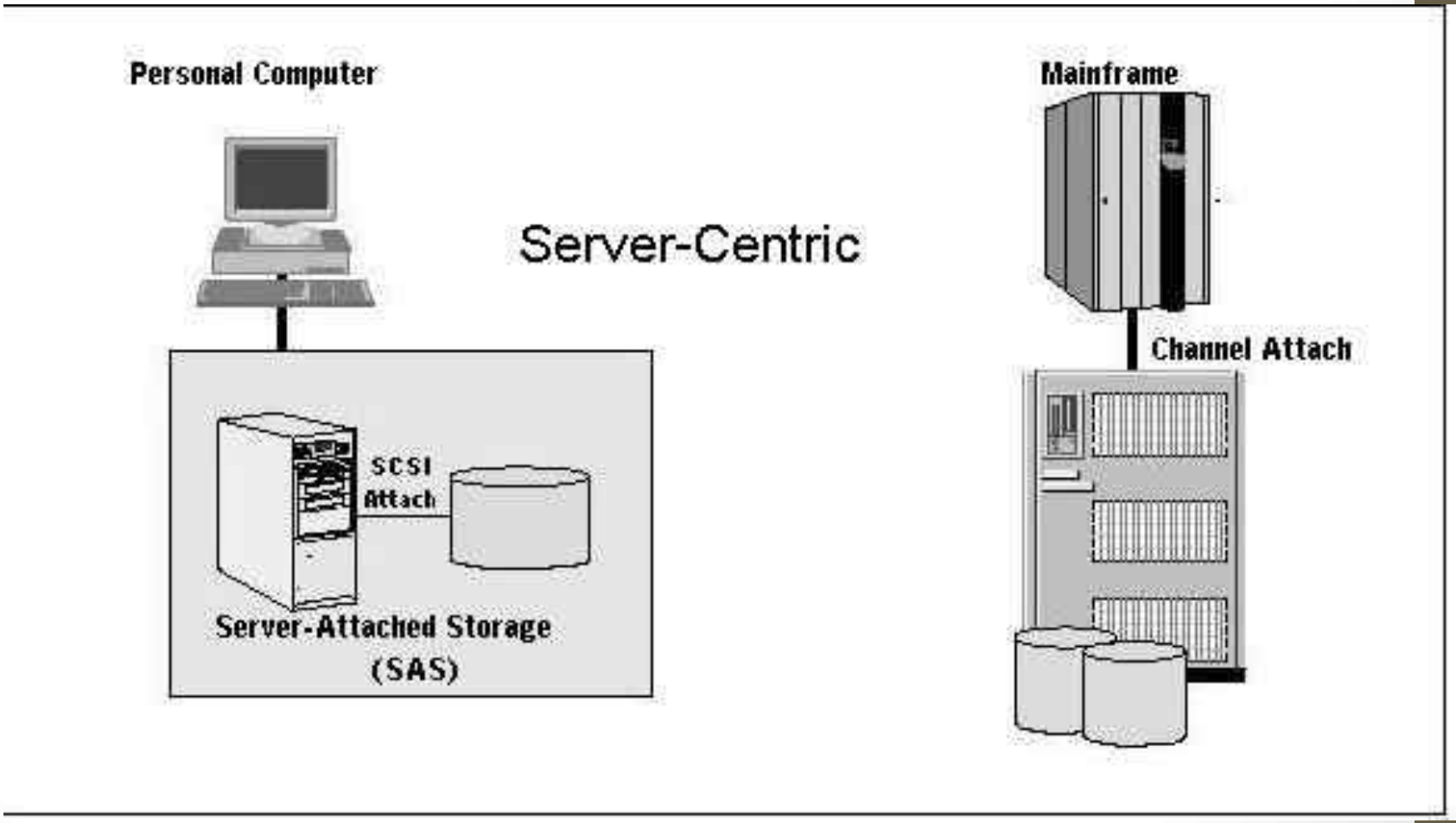
Evolution in Storage Architecture



Problem we are facing

- Scalability --Rapidly growing data volume
- Connectivity --Distributed data sharing
- 24/7 availability, no single point failure
- High performance
- Easy management

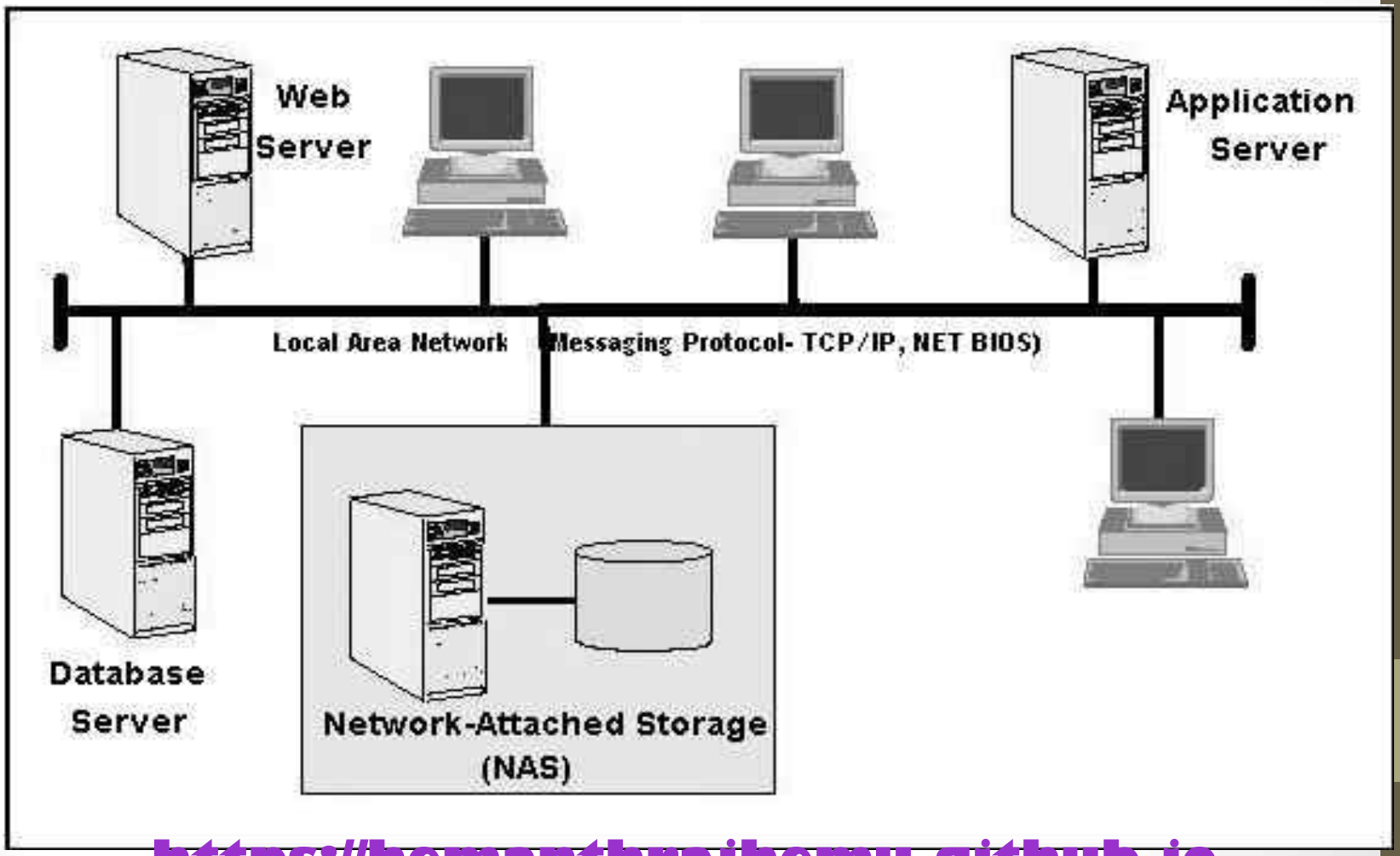
Server-Attached Storage (SAS)



SAS -- How to Share Data

- Each has own copy
  - scalability: Poor
  - availability: OK
  - performance: OK
  - management: how to keep data sync?
  - Connectivity: NA
- One copy, share
  - OK
  - single point failure
  - not that good
  - how to make back up without affecting service?
  - System dependent

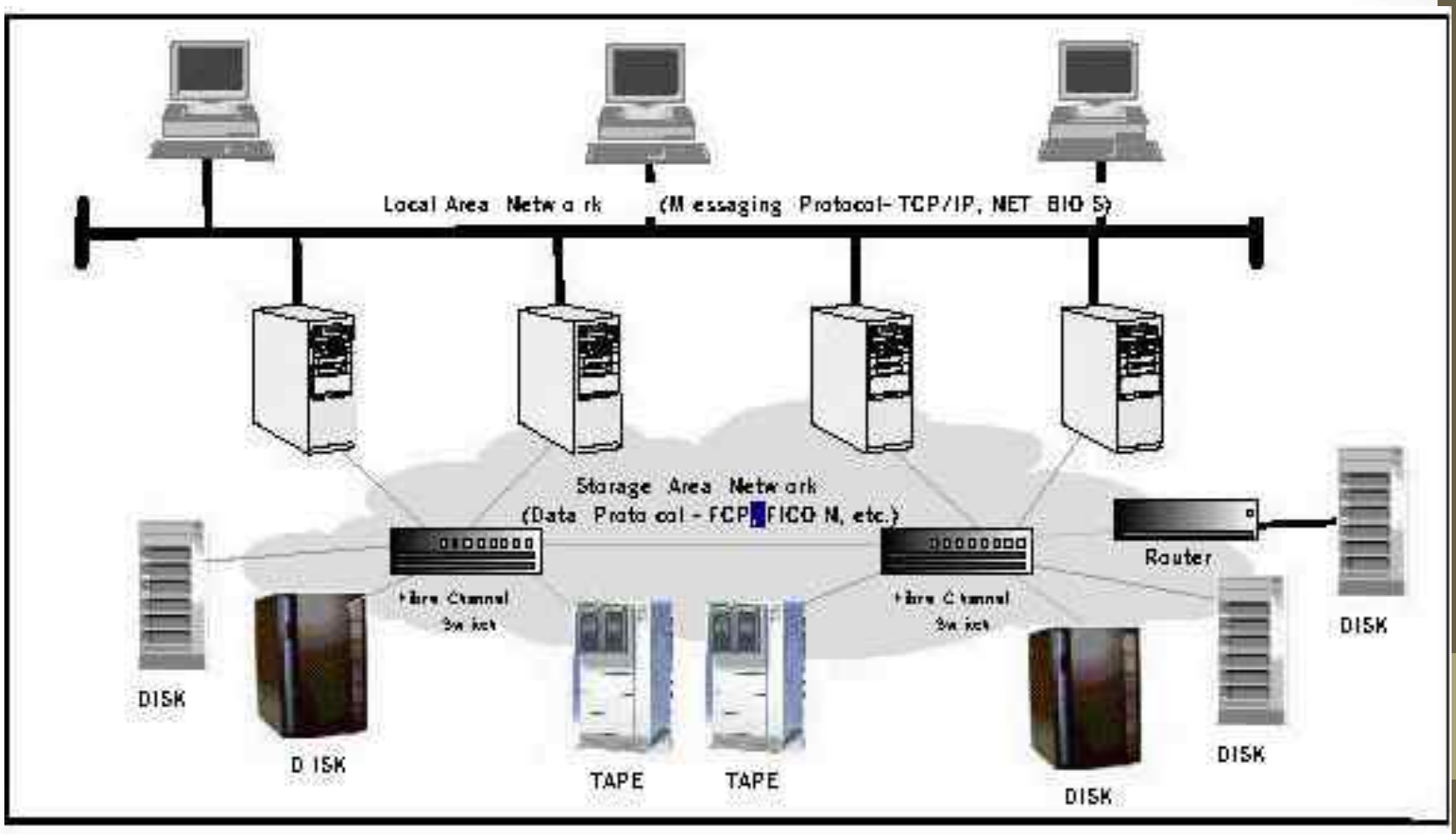
Network-Attached Storage(NAS)



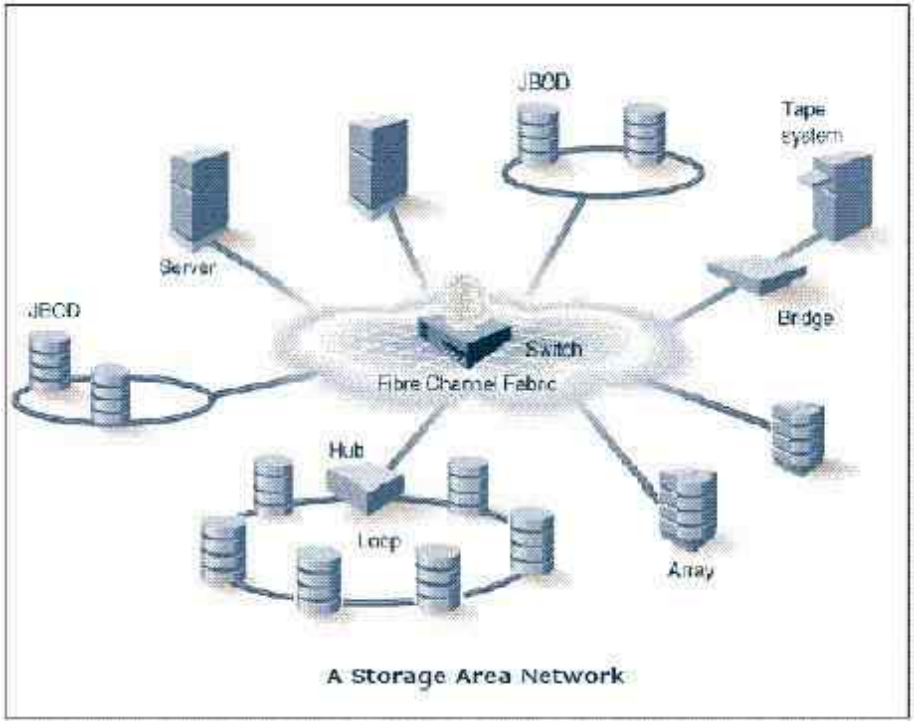
NAS

- Scalability: good
- Availability: as long as the LAN and NAS device work, generally good
- Performance: limited by speed of LAN, traffic conflicts, inefficient protocol
- Management: OK
- Connection: homogeneous vs. heterogeneous

Storage Area Network (SAN)



Storage Area Network (SAN)



- SAN is created by using the Fibre Channel to link peripheral devices such as disk storage and tape libraries

SAN vs. NAS

- Dedicated Fibre Channel Network for Storage
- More efficient protocol
- ==> higher availability
- ==> reduce traffic conflict
- ==> longer distance (up to 10 km)

Fibre Channel

- provides high-performance, any-to-any interconnection.
  - server to server
  - server to storage
  - storage to storage
- combines the characteristics of networks (large address space, scalability) and I/O channels (high speed, low latency, hardware error detection) together.

Benefits of SAN

- Scalability ==> Fibre Channel networks allow the number of attached nodes to increase without loss of performance because as switches are added, switching capacity grows. The limitations on the number of attached devices typical of channel interconnection disappears.

Benefits of SAN

- High Performance ==> Fibre Channel fabrics provide a switched 100Mbytes/second full duplex interconnect.
- Storage Management ==> SAN-attached storage allows the entire investment in storage to be managed in a uniform way.

Benefits of SAN

- Decoupling Servers and Storage
  - the servers can be upgraded while leaving storage in place.
  - Storage can be added at will and dynamically allocated to servers without downtime.

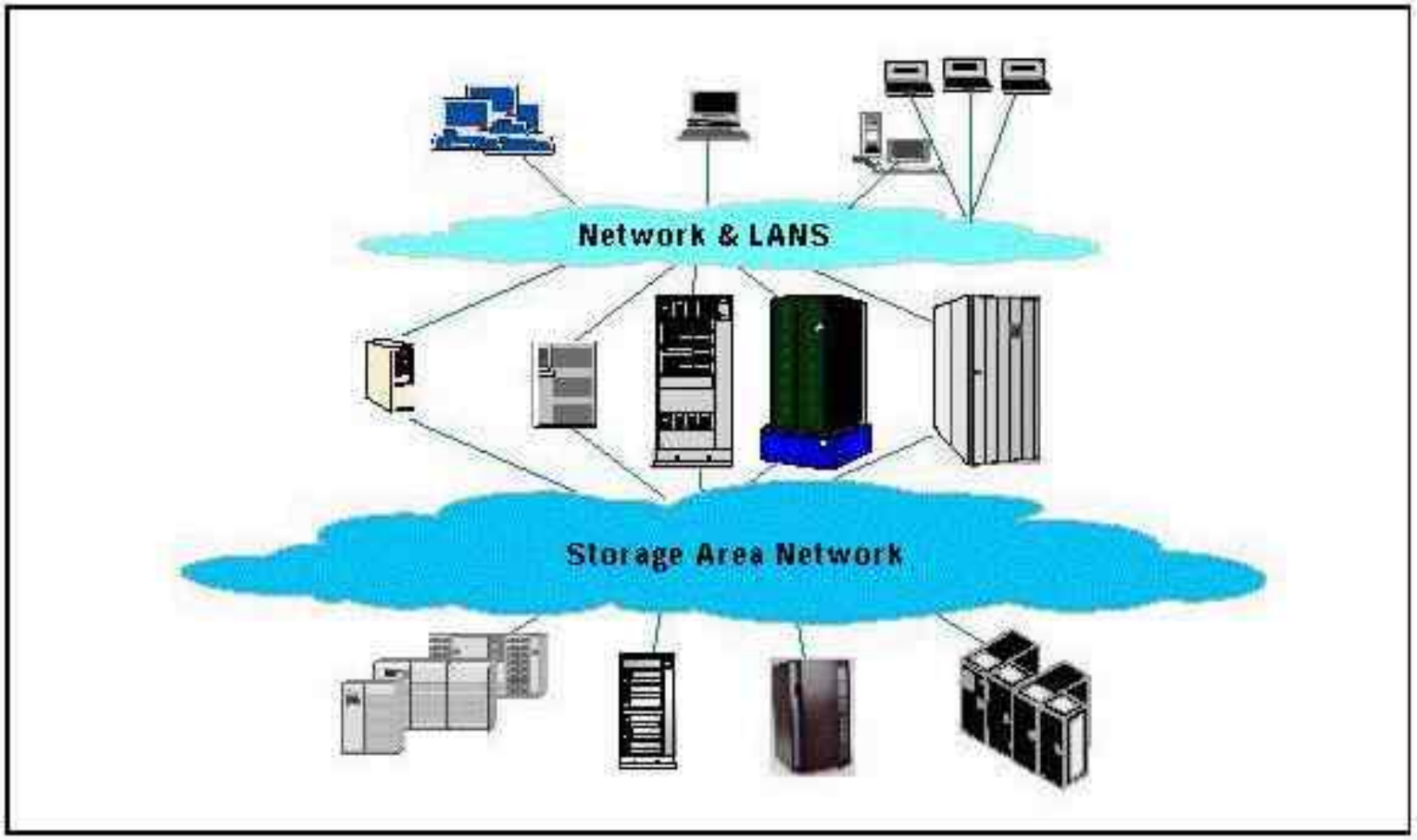
Easy Migration to SAN

- Host Bus Adapters (HBAs) -- connect servers to the SAN
- Fibre Channel storage -- connects directly to the SAN
- SCSI-FC bridge -- allows SCSI (disk and tape) components to be attached to the SAN
- SAN Network Components -- Fibre Channel switches

Summary

- SAN is a high-speed network that allows the establishment of direct connection between storage devices and processors (servers) centralized to the extent supported by the distance of Fibre Channel.

Storage Area Network



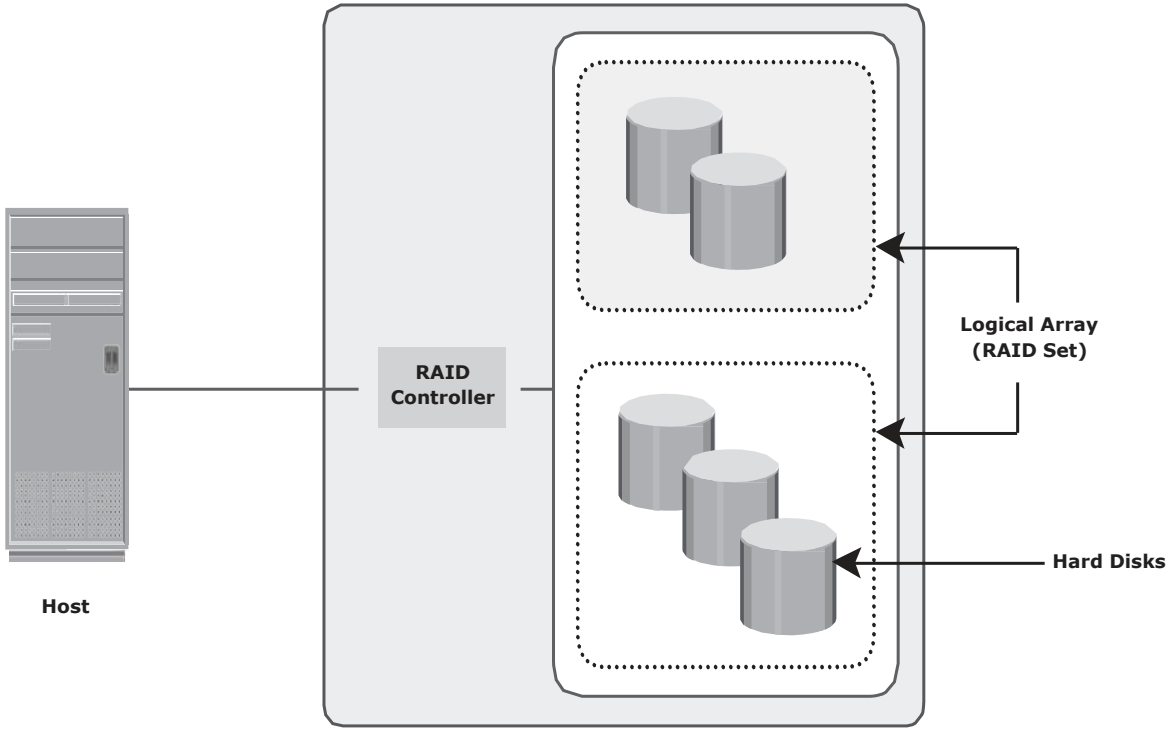
https://hemanthrajhemu.github.io

RAID Array Components

RAID Array Components

A *RAID array* is an enclosure that contains a number of 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*

Components of a RAID array



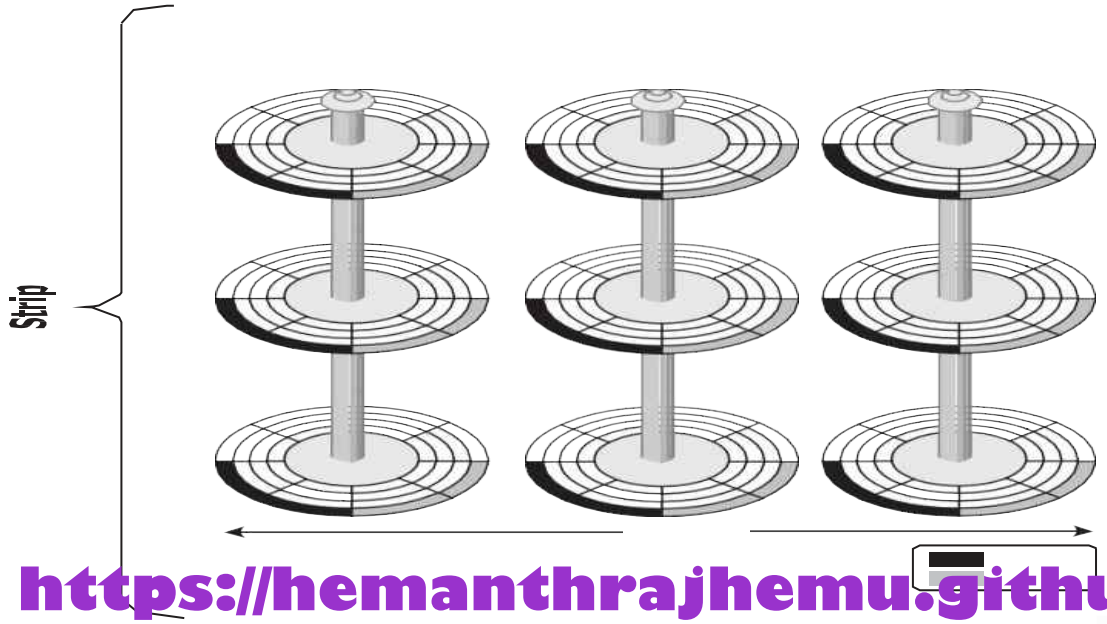
RAID techniques — striping, mirroring, and parity — form the basis for defining various RAID levels. These techniques determine the data availability and performance characteristics of a RAID set.

Striping

Striping

Striping is a technique to spread data across multiple drives (more than one) to use the drivesin parallel. Allthe read-writeheadsworksimultaneously, allowing more data to be processed in a shorter time and increasing performance, compared to reading and writing from a single disk.

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, assuming that the accessed data starts at the beginning of the strip.



Intelligent Storage Systems

Chapter Objectives

After completing this module, you will be able to:

- Describe components of intelligent storage system
- List benefits of intelligent storage system
- Explain intelligent cache algorithms and protection
- Describe implementation of intelligent storage system
  - High-end storage array
  - Midrange storage array

What is an Intelligent Storage System

Intelligent Storage Systems are RAID arrays that are:

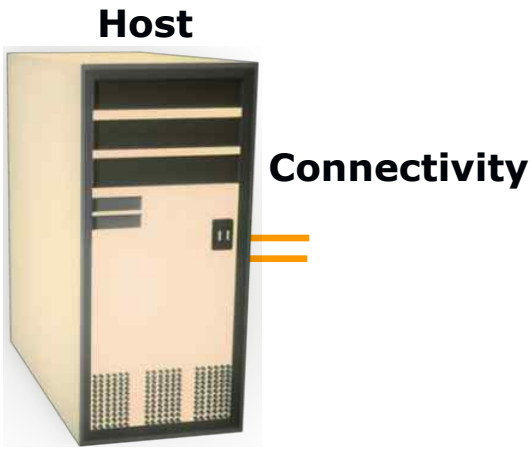
- Highly optimized for I/O processing
- Have large amounts of cache for improving I/O performance
- Have operating environments that provide:
  - Intelligence for managing cache
  - Array resource allocation
  - Connectivity for heterogeneous hosts
  - Advanced array based local and remote replication options

Benefits of an Intelligent Storage System

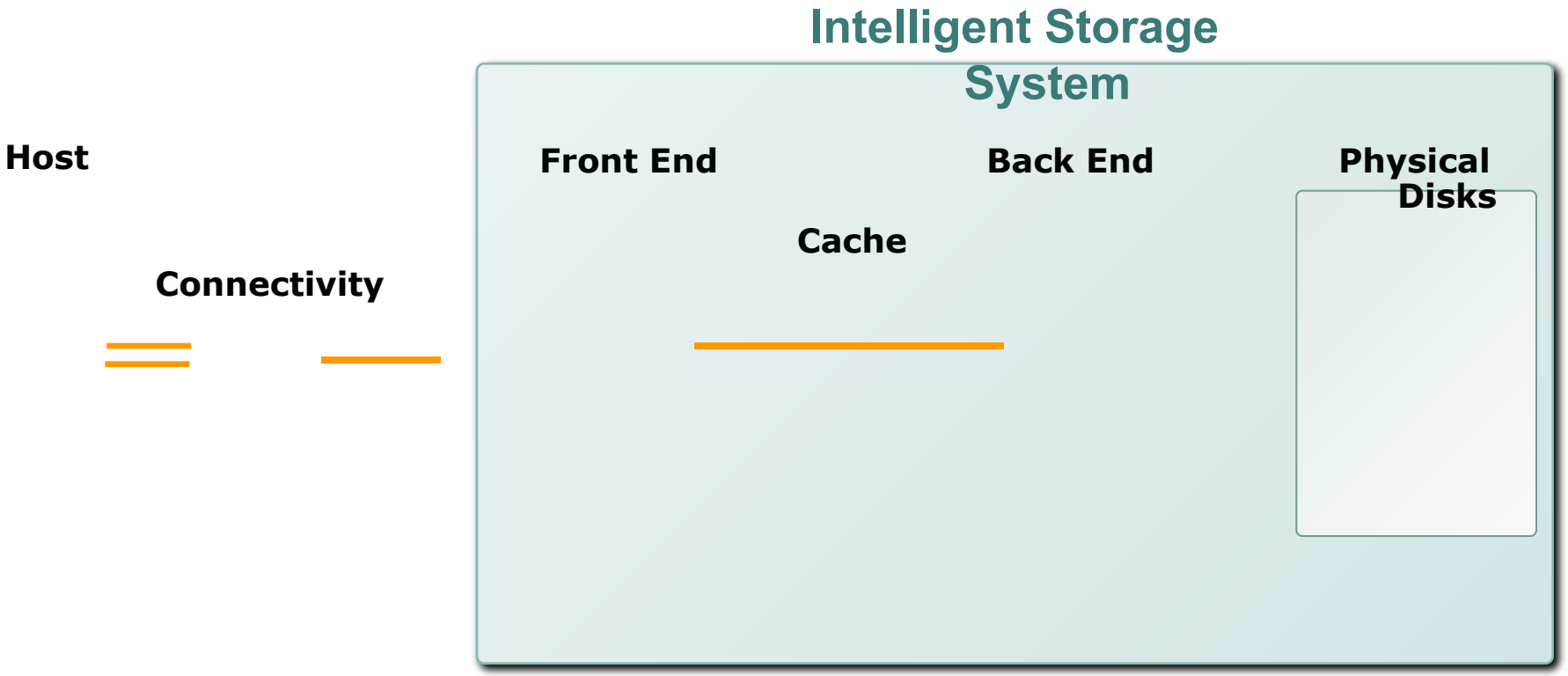
Intelligent storage system provides several benefits over a collection of disks in an array (JBOD) or even a RAID arrays:

- Increased capacity
- Improved performance
- Easier data management
- Improved data availability and protection
- Enhanced Business Continuity support
- Improved security and access control

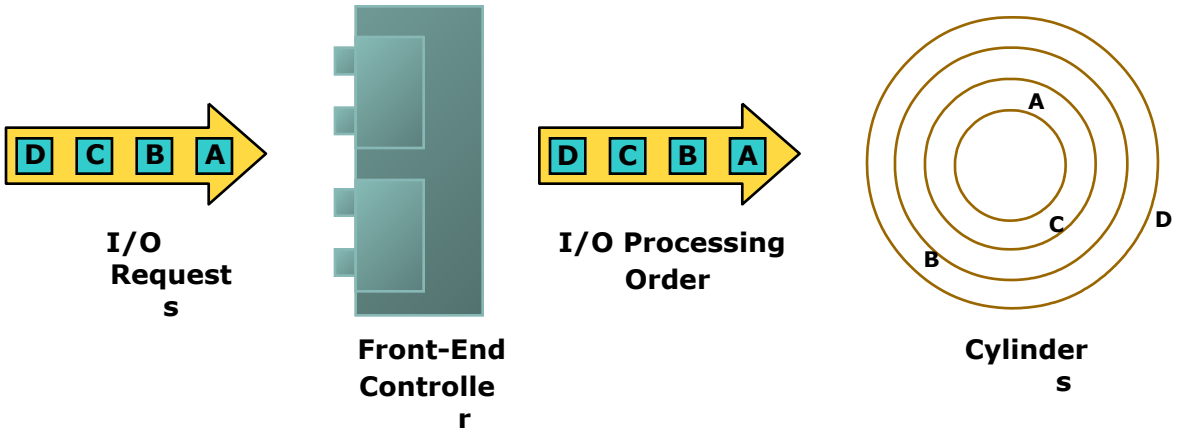
Components of an Intelligent Storage System



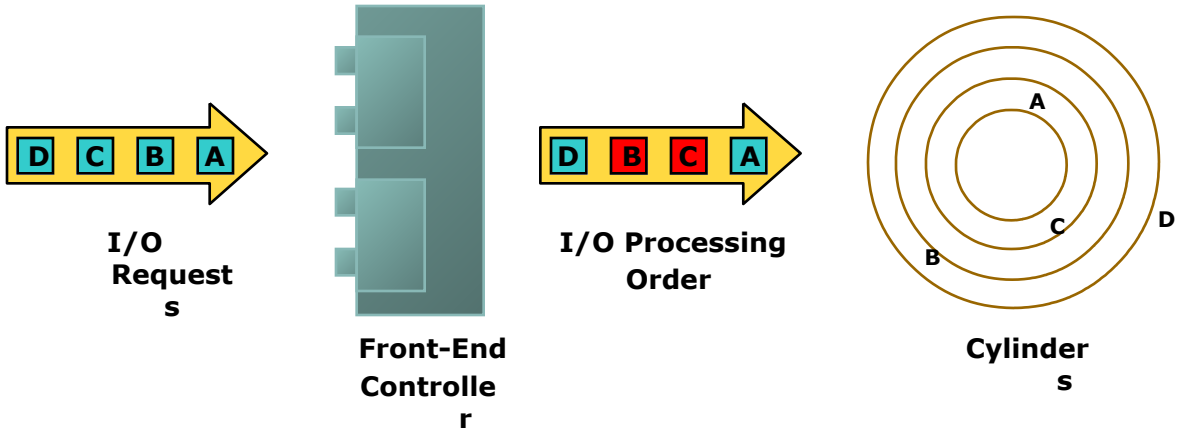
Intelligent Storage System: Front End



Front End Command Queuing

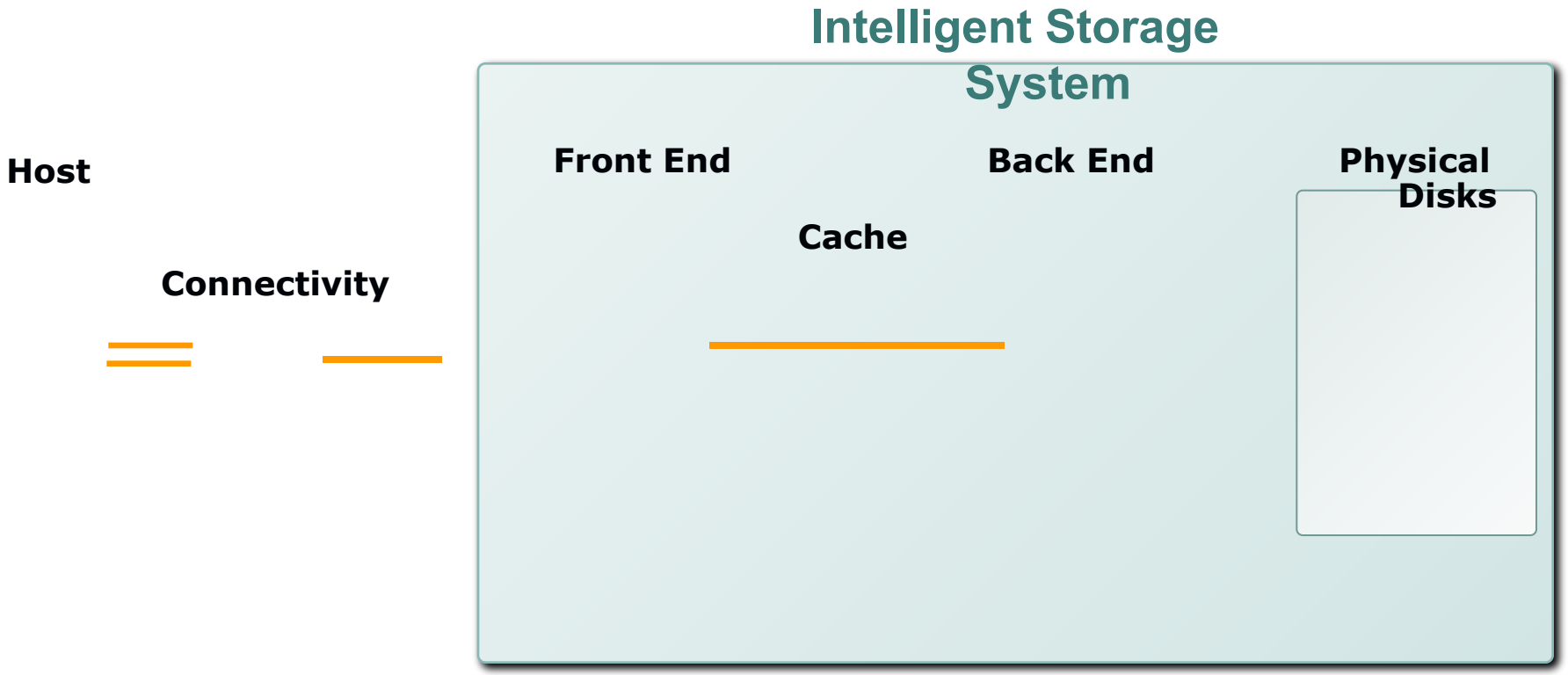


Without Optimization (FIFO)



With command queuing

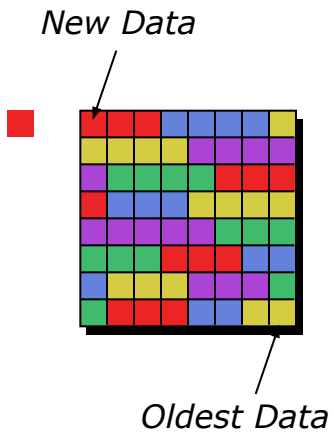
Intelligent Storage System: Cache



Write Operation with Cache

Read Operation with Cache: ‘Hits’ and ‘Misses’

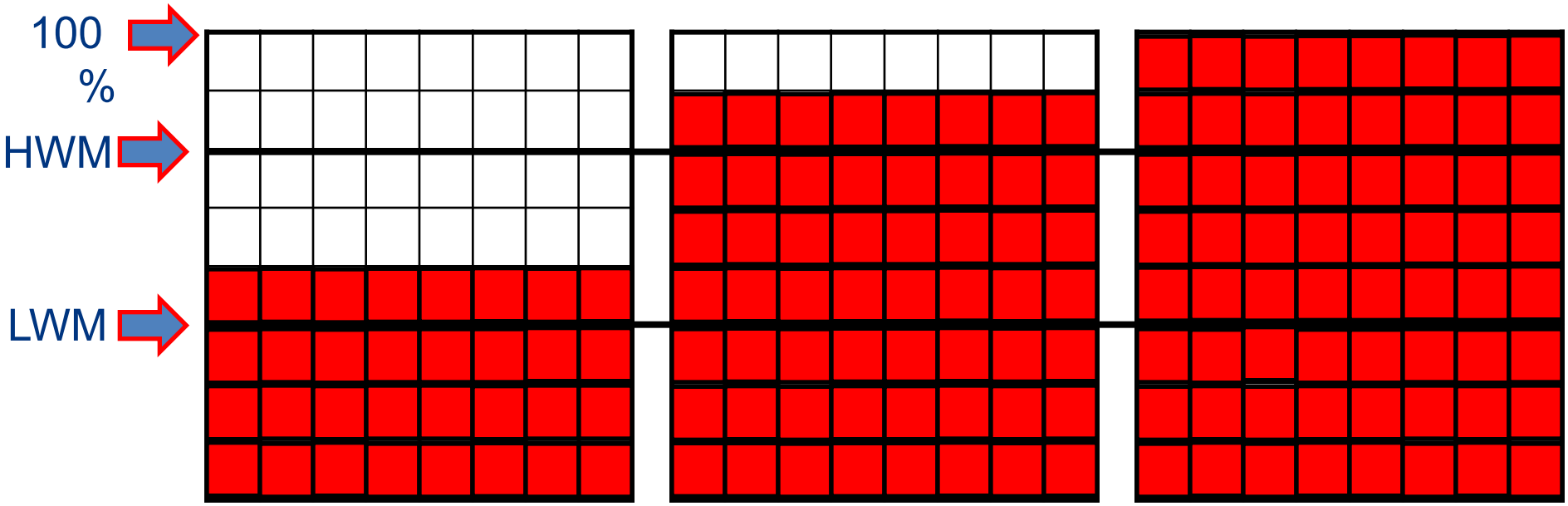
Cache Management: Algorithms



- Least Recently Used (LRU)
  - Discards least recently used data
- Most Recently Used (MRU)
  - Discards most recently used data

Cache Management: Watermarking

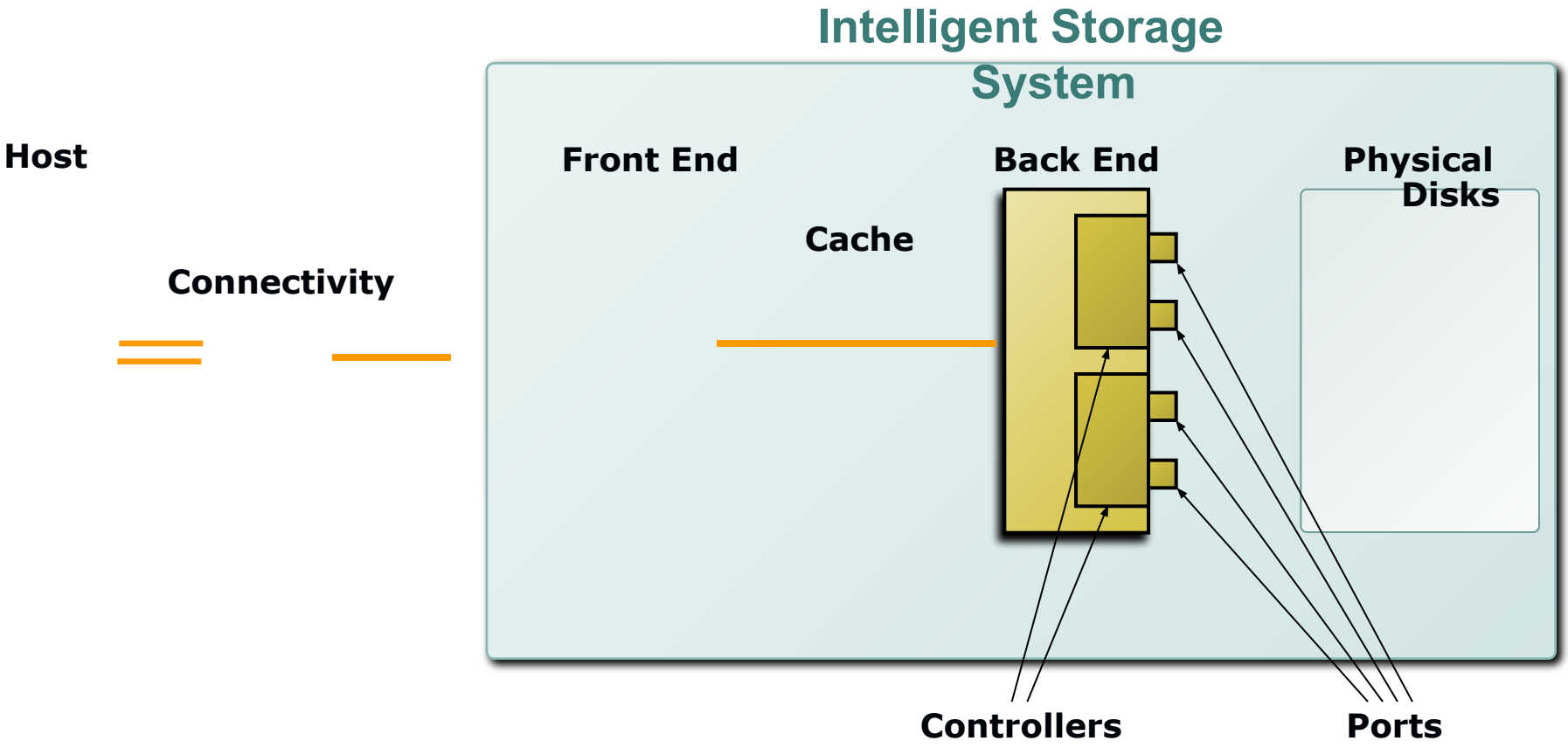
- Manage peak I/O requests “bursts” through flushing/de-staging
  - Idle flushing, High Watermark flushing and Forced flushing
- For maximum performance:
  - Provide headroom in write cache for I/O bursts



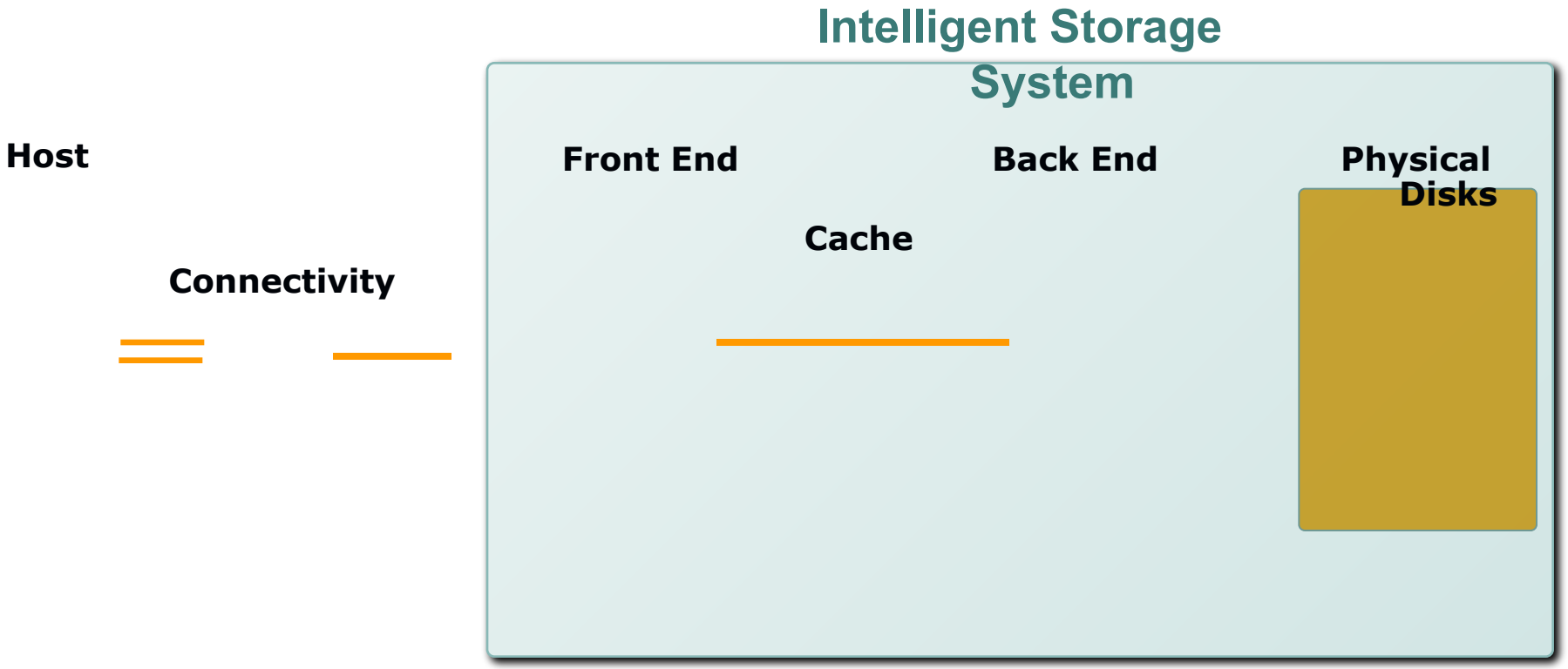
Cache Data Protection

- Protecting cache data against failure:
  - Cache mirroring
    - Each write to the cache is held in two different memory locations on two independent memory cards
  - Cache vaulting
    - Cache is exposed to the risk of uncommitted data loss due to power failure

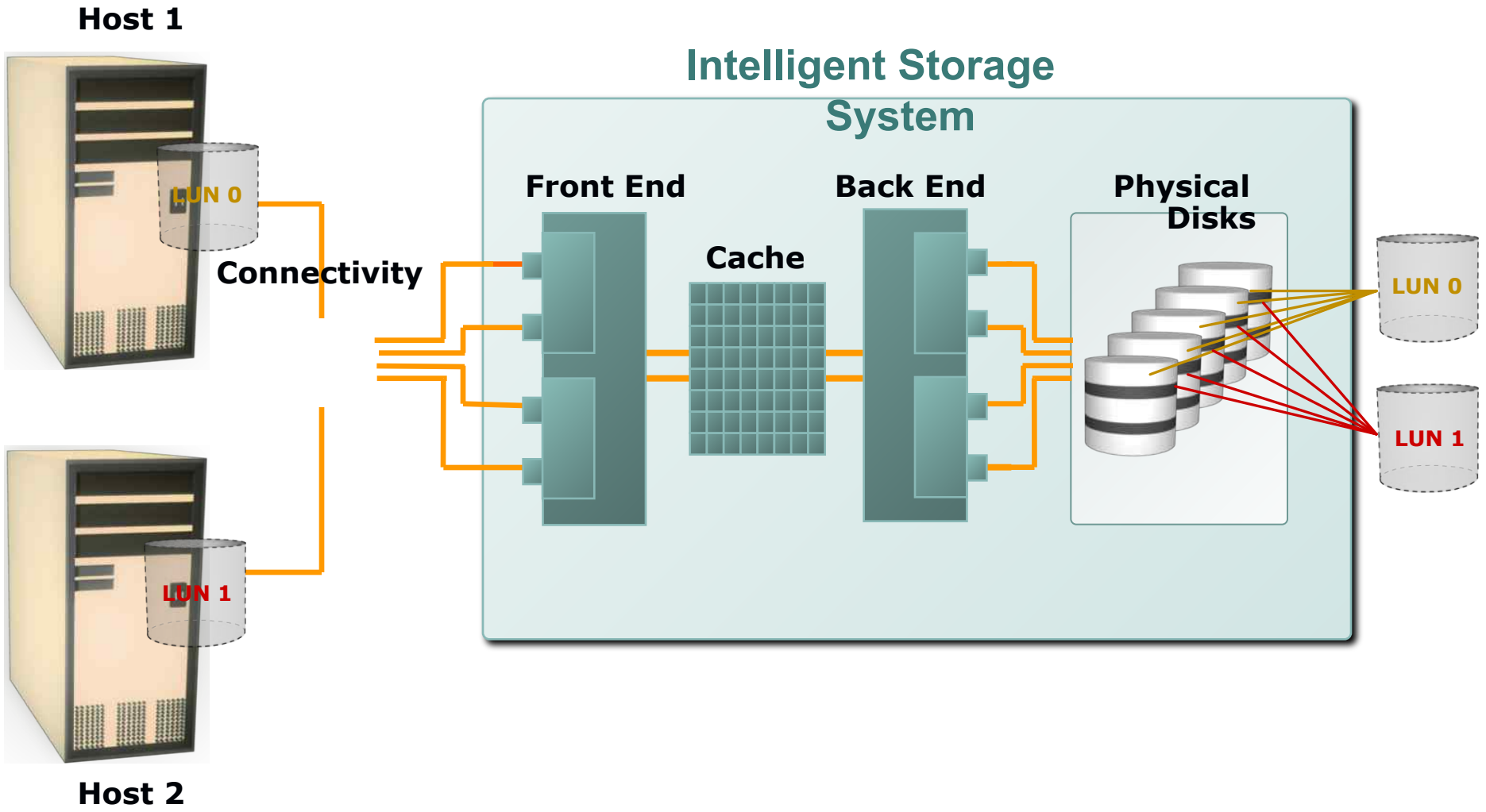
Intelligent Storage System: Back End



Intelligent Storage System: Physical Disks



What the Host Sees – RAID Sets and LUNs(Logical unit numbers)

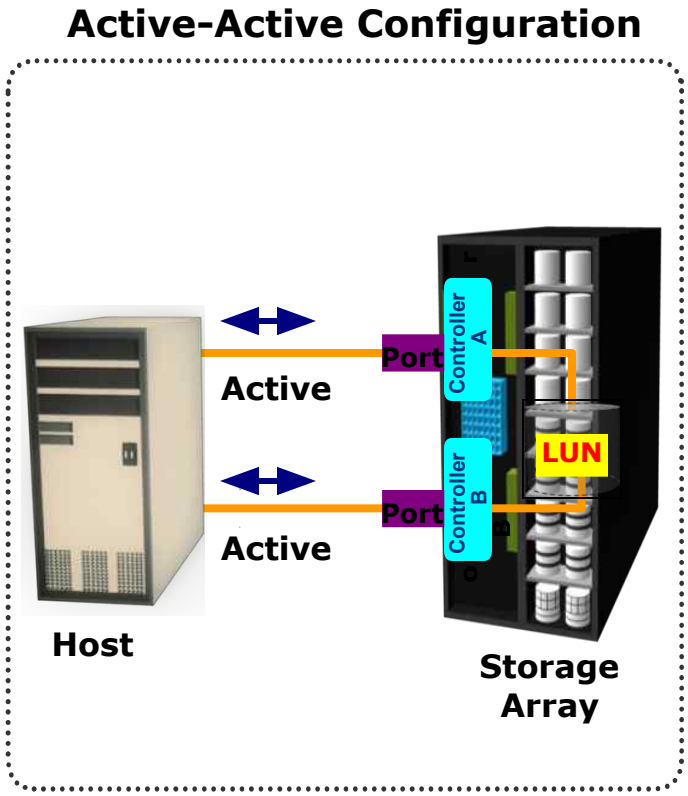


LUN Masking

- LUN masking is access control mechanism
- Process of masking LUNs from unauthorized access
- Implemented on storage arrays
- S*torage group* logical entity that contains one or more LUNs and one host

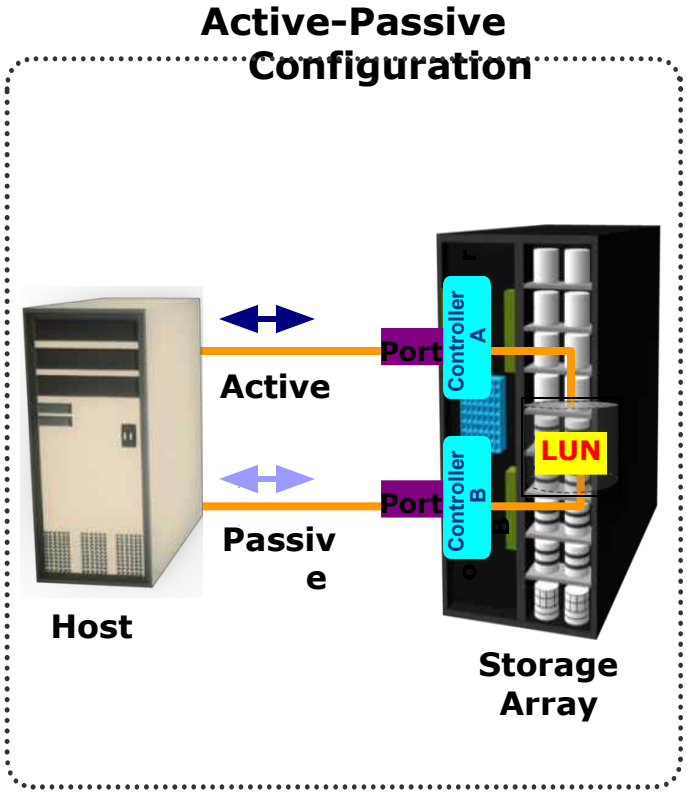
ISS Implementation: High-end Storage Systems

- Also referred as Active-active arrays
  - I/O's are serviced through all the available path
- Following are high-end array capabilities:
  - Large storage capacity
  - Huge cache to service host I/Os
  - Fault tolerance architecture
  - Multiple front-end ports and support to interface protocols
  - High scalability
  - Ability to handle large amounts of concurrent I/Os
- Designed for large enterprises



Midrange Storage Systems

- Also referred as Active-passive arrays
  - Host can perform I/Os to LUNs only through active paths
  - Other paths remain passive till active path fails
- Midrange array have two controllers, each with cache, RAID controllers and disks drive interfaces
- Designed for small and medium enterprises
- Less scalable as compared to high-end array



Chapter Summary

Key points covered in this chapter:

- Intelligent Storage Systems features
- Components of Intelligent Storage Systems
- Cache management algorithms
- Intelligent Storage System implementation
  - High-end storage array
  - Mid range storage array

BMS
Institute of Technology and Management

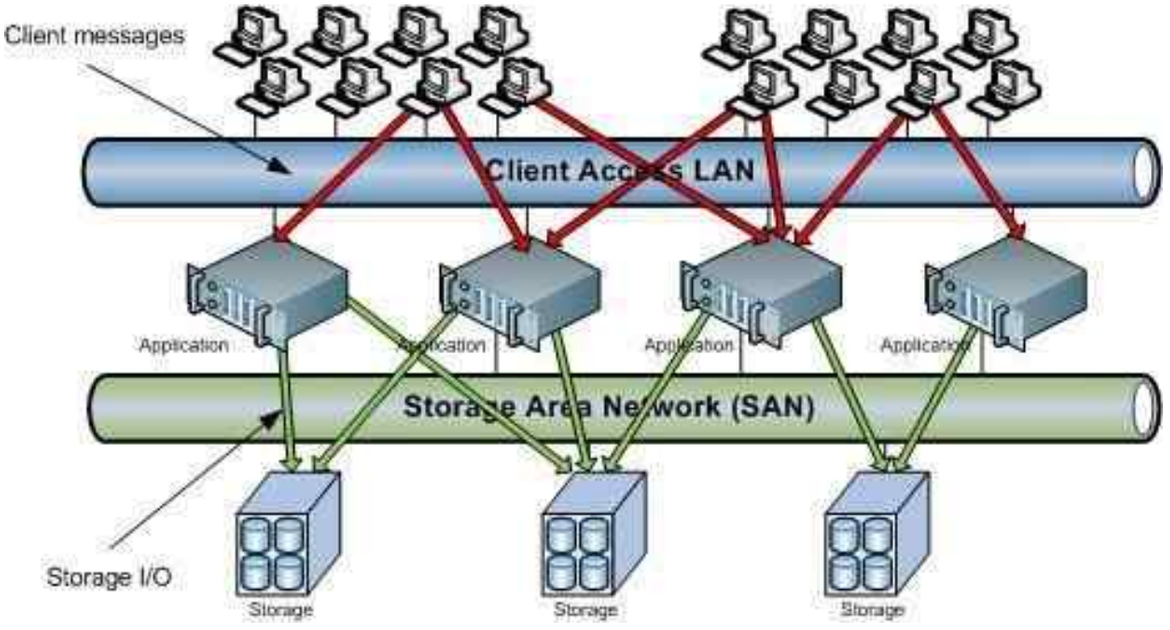
Avalahalli, Doddaballapur Main Road, Bengaluru – 560064

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Storage Area Networks

(17CS754)

SANs are primarily used to access storage devices, such as disk arrays and tape libraries from servers so that the devices appear to the operating system as direct- attached storage.



STORAGE AREA NETWORKS
[As per Choice Based Credit System (CBCS) scheme]
(Effective from the academic year 2017 - 2018) SEMESTER – VII

Subject Code	17CS754	IA Marks	40
Number of Lecture Hours/Week	3	Exam Marks	60
Total Number of Lecture Hours	40	Exam Hours	03

CREDITS – 03

Module – 1	**Teaching Hours**
Storage System Introduction to evolution of storage architecture, key data centre Elements, virtualization, and cloud computing. Key data centre elements – Host (or compute), connectivity, storage, and application in both classic and virtual Environments. RAID implementations, techniques, and levels along with the Impact of RAID on application performance. Components of intelligent storage systems and virtual storage provisioning and intelligent storage system Implementations.	**8 Hours**
Module – 2	
Storage Networking Technologies and Virtualization Fibre Channel SAN components, connectivity options, and topologies including access protection mechanism „zoning", FC protocol stack, addressing and operations, SAN-based virtualization and VSAN technology, iSCSI and FCIP(Fibre Channel over IP) protocols for storage access over IP network, Converged protocol FCoE and its components, Network Attached Storage (NAS) - components, protocol and operations, File level storage virtualization, Object based storage and unified storage platform.	**8 Hours**
Module – 3	
Backup, Archive, and Replication This unit focuses on information availability and business continuity solutions in both virtualized and non-virtualized environments. Business continuity terminologies, planning and solutions, Clustering and multipathing architecture to avoid single points of failure, Backup and recovery - methods, targets and topologies, Data deduplication and backup in virtualized environment, Fixed content and data archive, Local replication in classic and virtual environments, Remote replication in classic and virtual environments, Three-site remote replication and continuous data protection	**8 Hours**
Module – 4	
Cloud Computing Characteristics and benefits This unit focuses on the business drivers, definition, essential characteristics, and phases of journey to the Cloud. ‚Business drivers for Cloud computing, Definition of Cloud computing, Characteristics of Cloud computing, Steps involved in transitioning from Classic data center to Cloud computing environment Services and deployment models, Cloud infrastructure components, Cloud migration considerations	**8 Hours**
Module – 5	
Securing and Managing Storage Infrastructure This chapter focuses on framework and domains of storage security along with covering security. implementation at storage networking. Security threats, and countermeasures in various domains Security solutions for (Fiber channel)FC-SAN, IP-SAN, and NAS	**8 Hours**

managing various information infrastructure components in classic and virtual environments, Information lifecycle management (ILM) and storage tiering, Cloud service management activities	

Course outcomes: The students should be able to:

- Identify key challenges in managing information and analyze different storage networking technologies and virtualization
- Explain components and the implementation of NAS
- Describe CAS architecture and types of archives and forms of virtualization
- Illustrate the storage infrastructure and management activities

Question paper pattern:

The question paper will have ten questions.
There will be 2 questions from each module.
Each question will have questions covering all the topics under a module.
The students will have to answer 5 full questions, selecting one full question from each module.

Text Books:

1. Information Storage and Management, Author :EMC Education Services, Publisher: Wiley ISBN: 9781118094839
2. Storage Virtualization, Author: Clark Tom, Publisher: Addison Wesley Publishing Company ISBN: 9780321262516

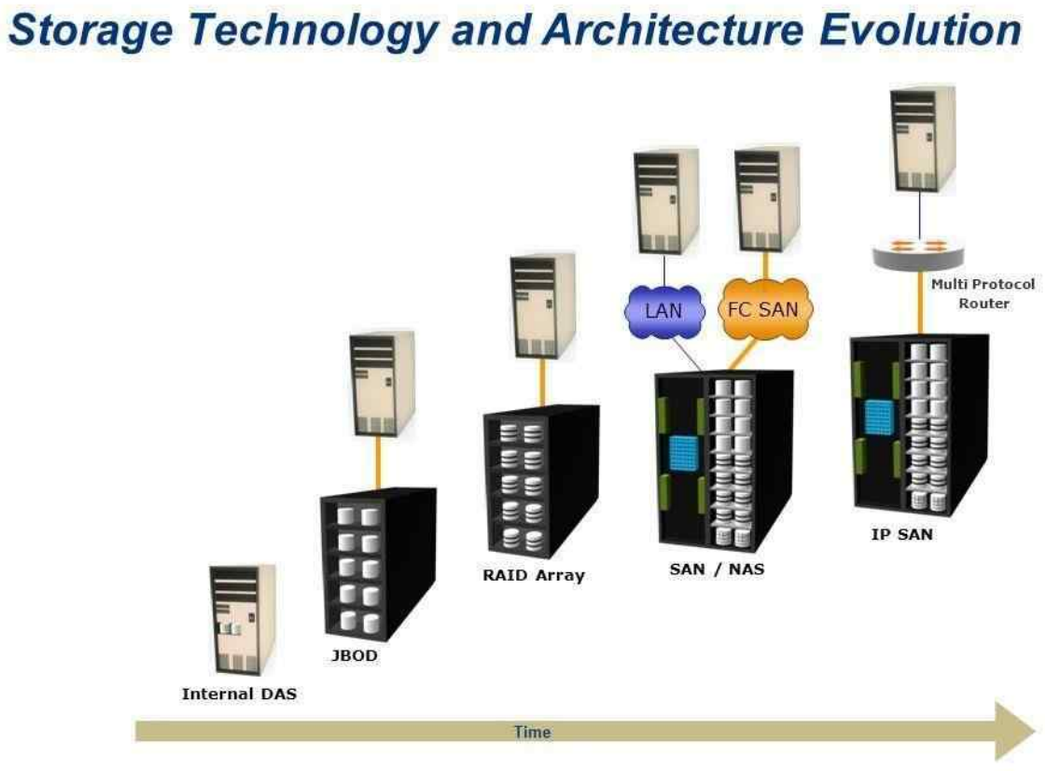
Table of Content

Module-1

Introduction to evolution of storage architecture:

Introduction

Information is increasingly important in our daily lives. We have become information dependents of the twenty-first century, living in an on-command, on-demand world that means we need information when and where it is required. We access the Internet every day to perform searches, participate in social networking, send and receive e-mails, share pictures and videos, and scores of other applications. Equipped with a growing number of content-generating devices, more information is being created by individuals than by businesses.



Key data centre Elements:

Uninterrupted operation of data centers is critical to the survival and success of a business. It is necessary to have a reliable infrastructure that ensures data is accessible at all times. While the requirements, , are applicable to all elements of the data centre infrastructure, our focus here is on storage systems.

1**Availability:** All data center elements should be designed to ensure accessibility. The inability of users to access data can have a significant negative impact on a business.

2**Security:** Polices, procedures, and proper integration of the data center core elements that will prevent unauthorized access to information must be established. In addition to the security measures for client access, specific mechanisms must enable servers to access only their allocated resources on storage arrays.

3**Scalability:** Data center operations should be able to allocate additional processing capabilities or storage on demand, without interrupting business operations. Business growth often requires deploying more servers, new applications, and additional databases. The storage solution should be able to grow with the business.

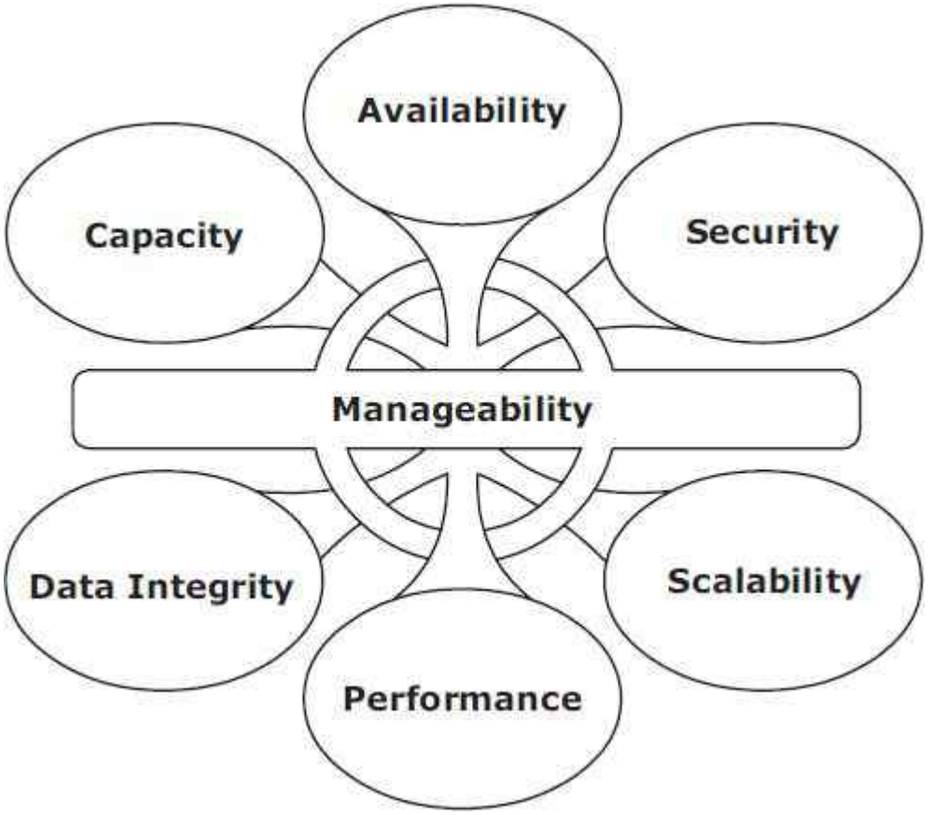
4**Performance:** All the core elements of the data center should be able to provide optimal performance and service all processing requests at high speed. The infrastructure should be able to support performance requirements.

5**Data integrity:** Data integrity refers to mechanisms such as error correction codes or parity bits which ensure that data is written to disk exactly as it was received. Any variation in data during its retrieval implies corruption, which may affect the operations of the organization.

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 be able to provide additional capacity without interrupting availability, or, at the very least, with minimal disruption.

Capacity may be managed by reallocation of existing resources, rather than by adding new resources.

7**Manageability:** A data center should perform all operations and activities in the most efficient manner. Manageability can be achieved through automation and the reduction of human (manual) intervention in common tasks.



Virtualization and cloud computing:

What Is Cloud Computing?

The National Institute of Standards defines cloud computing as "enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services)."

To be a cloud, NIST has determined it must have the following five essential characteristics:

- On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage.
- Broad network access: Capabilities are available over the network through multiple clients and devices.
- Resource pooling: The provider‟s computing resources are pooled to serve numerous consumers using a multi-tenant model.
- Rapid elasticity: Users can add or reduce capacity through software
- Measured service: Automatic control and optimization of resources detailing who is using what and how much.

Without those five essential characteristics, it is technically not a cloud.
The cloud model is comprised of three service models:

- Software as a Service (SaaS): The consumer can use the provider‟s applications running on a cloud infrastructure.
- Platform as a Service (PaaS): The consumer can deploy on the cloud infrastructure, applications created using programming languages, libraries, services or tools supported by the provider.
- Infrastructure as a Service (IaaS): The consumer can provision processing, storage, networks, and other computer resources to deploy and run arbitrary software.

There are four deployment models of the cloud:

- Private Cloud: The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers.
- Community Cloud: The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations).
- Public Cloud: The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization.
- Hybrid Cloud: The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public).

What Is Virtualization?

Contrary to what some believe, virtualization is not cloud computing. It is, however, a fundamental technology that makes cloud computing work. While cloud computing and virtualization rely on similar models and principles, they are intrinsically different.
Simply put, virtualization can make one resource act like many, while cloud computing lets different users access a single pool of resources.
With virtualization, a single physical server can become multiple virtual machines, which are essentially isolated pieces of hardware with plenty of processing, memory, storage, and network capacity.
Each virtual machine can run independently while sharing the resources of a single host machine because they‟ve been loaded into hypervisors. Hypervisors, also known as the abstraction layer, are used to separate physical resources from their virtual environments. Once resources are pooled together, they can be divided across many virtual environments as needed.

Cloud Computing vs. Virtualization

Deciding which to implement for your business depends on the type of business and the requirements you have.

For smaller companies, cloud computing is easier and more cost-effective to implement. Resources are accessed via the Internet rather than added to the network.

Many small businesses are turning to the cloud for applications such as customer relationship management (CRM), hosted voice over IP (VoIP) or off-site storage. The cost of using the cloud is much lower than implementing virtualization. Cloud computing also offers easier installation of applications and hardware, access to software they couldn‟t otherwise afford, and the ability to try software before they buy it. It requires a small investment to implement a cloud-based application.

For some businesses, virtualization is the smarter choice and can save money in several different ways:

- Adding many guests to one house maximizes resources, which means the business needs fewer servers. This cuts down on operational costs.
- Fewer servers mean fewer people to look after and manage servers. This helps to consolidate management, thereby reducing costs.
- Virtualization also adds another layer of protection for business continuity, since virtual machines will limit the damage to itself.

Software RAID

Software RAID uses host-based software to provide RAID functions. It is imple- mented 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:

- **Performance:** Software RAID affects overall system performance. This is duetoadditional CPU cyclesrequiredto perform RAID calculations.
- **Supported features:** Software RAID does not support all RAID levels.
- **Operating system compatibility:** Software RAID is tied to the host operat- ing 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.

Hardware RAID

In *hardware RAID* implementations, a specialized hardware controller is implemented either on the host or on the array.

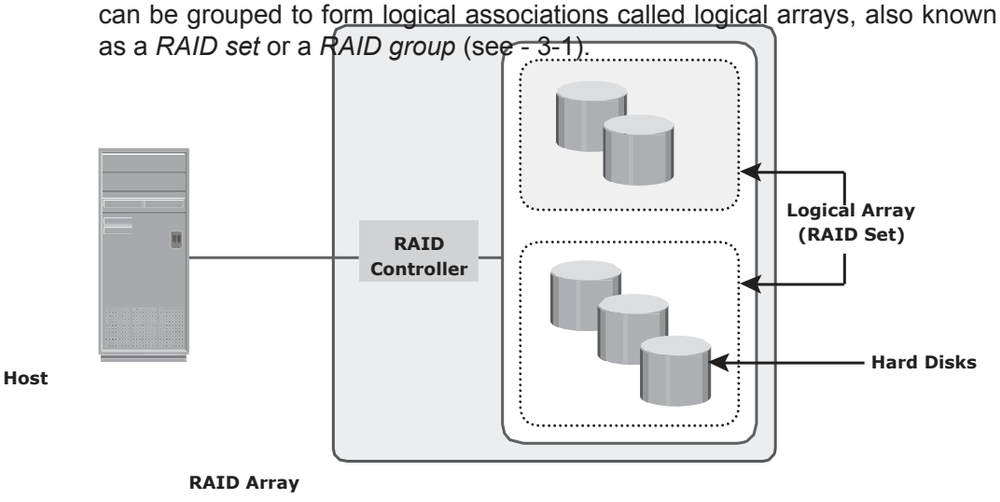
Controller card RAID is a host-based hardware RAID implementation in which a specialized RAID controller is installed in the host, and disk drives are con- nected to it. Manufacturers also integrate RAID controllers on motherboards. A host-based RAID controller is not an efficient solution in a data center environ- ment 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:

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

RAID Array Components

A *RAID array* is an enclosure that contains a number of 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 - 3-1).



- 3-1: Components of a RAID array

RAID Techniques

RAID techniques — striping, mirroring, and parity — form the basis for defin-

ing various RAID levels. These techniques determine the data availability and performance characteristics of a RAID set.

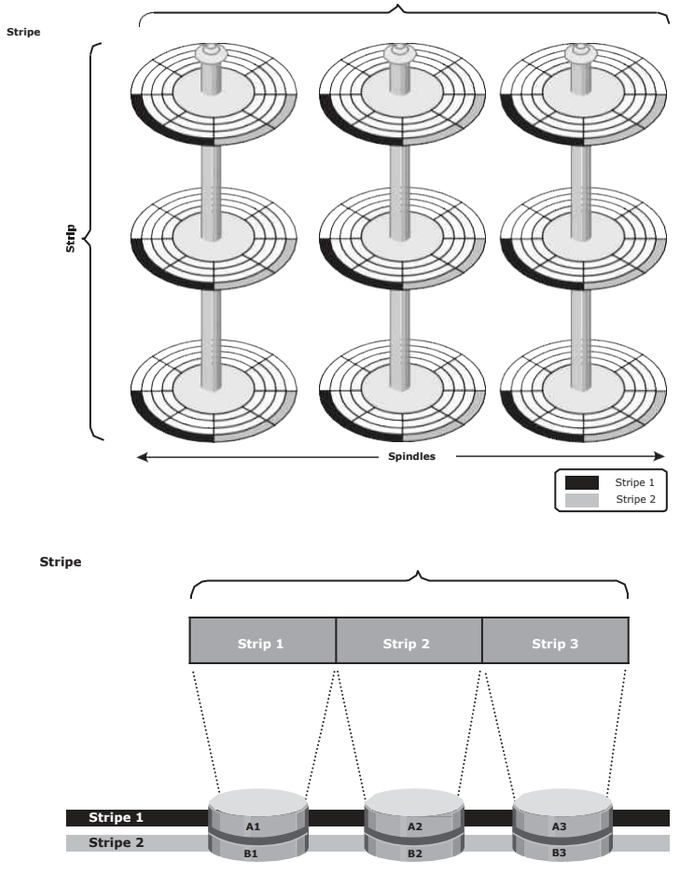
Striping

Striping is a technique to spread data across multiple drives (more than one) to use the drivesin parallel. Allthe read-writeheadsworksimultaneously, allowing

more data to be processed in a shorter time and increasing performance, com- pared to reading and writing from a single disk.

Within each disk in a RAID set, a predefined number of contiguously address- able 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*. - 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, assuming that 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.



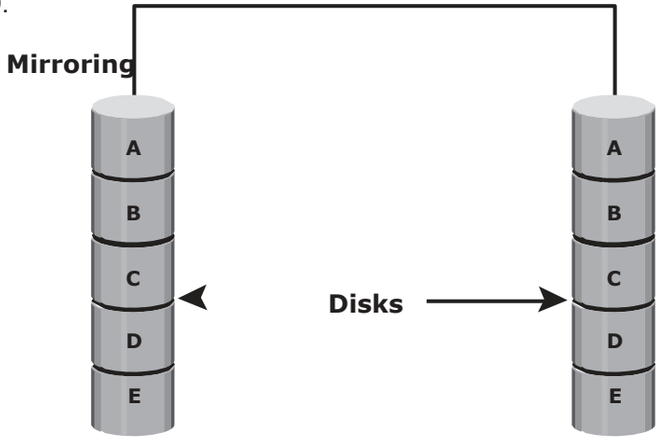
- 3-2: Striped RAID set

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 - 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 fromthe survivingdisk of the mirroredpair. This activityis transparent to the host. In addition to providing complete data redundancy, mirroring enables fast recovery from disk failure. However, disk mirroring provides only data pro- tection and is not a substitute for data backup. Mirroring constantly captures changes inthe data, whereas abackupcaptures point-in-time images of the data.

Mirroring involvesduplicationof data— the amount of storagecapacityneeded 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.

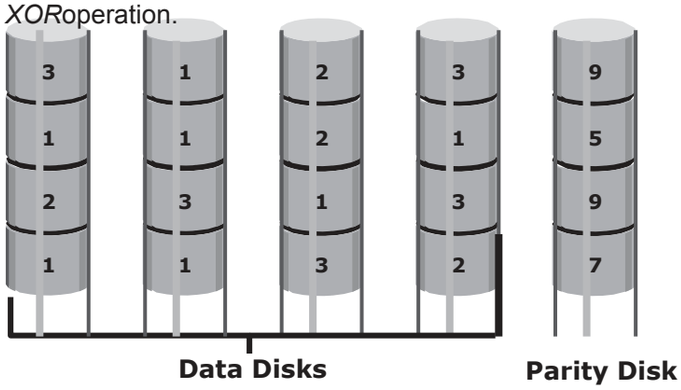


- 3-3: Mirrored disks in an array

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 distrib- uted across all the drives in a RAID set. - 3-4 shows a parity RAID set. The first four disks, labeled —Data Disks,ll contain the data. The fifth disk, labeled
—Parity Disk,ll 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 arith- metic sum of the data. However, parity calculation is a *bitwise XOR*operation.



- 3-4: ParityRAID

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).

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.

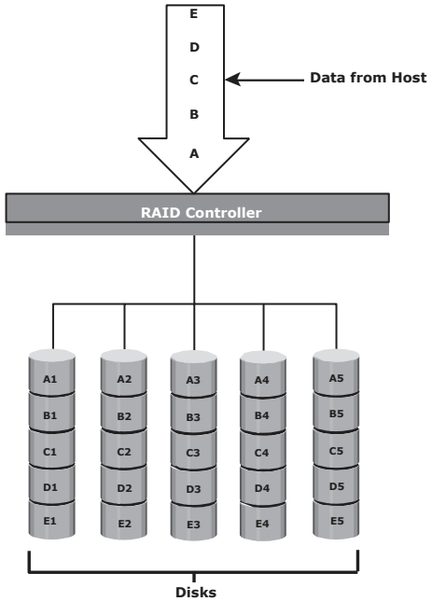
Table 3-1: Raid Levels

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

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. - 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 moredata can beread or written simultaneously. RAID 0 isa 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.

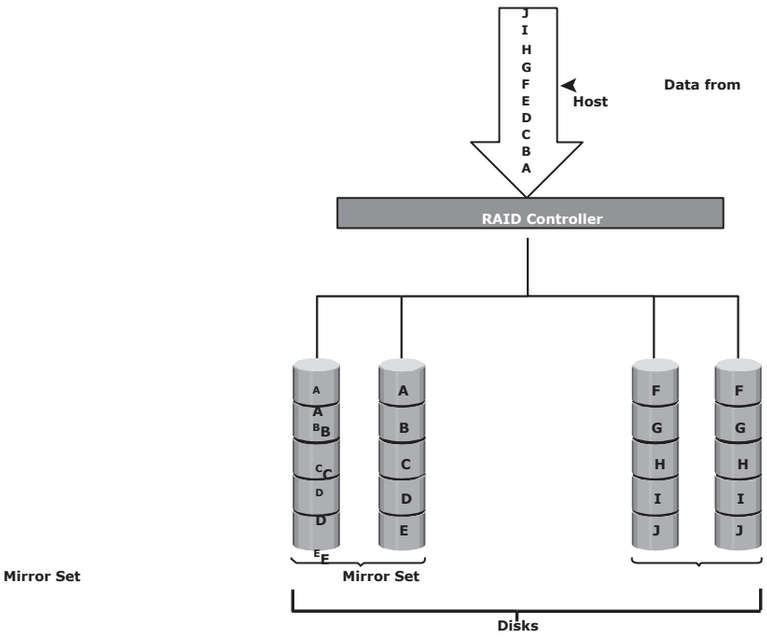


- 3-5: RAID 0

RAID 1

RAID 1 is based on the mirroring technique. In this RAID configuration, data is mirrored to provide fault tolerance (see - 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

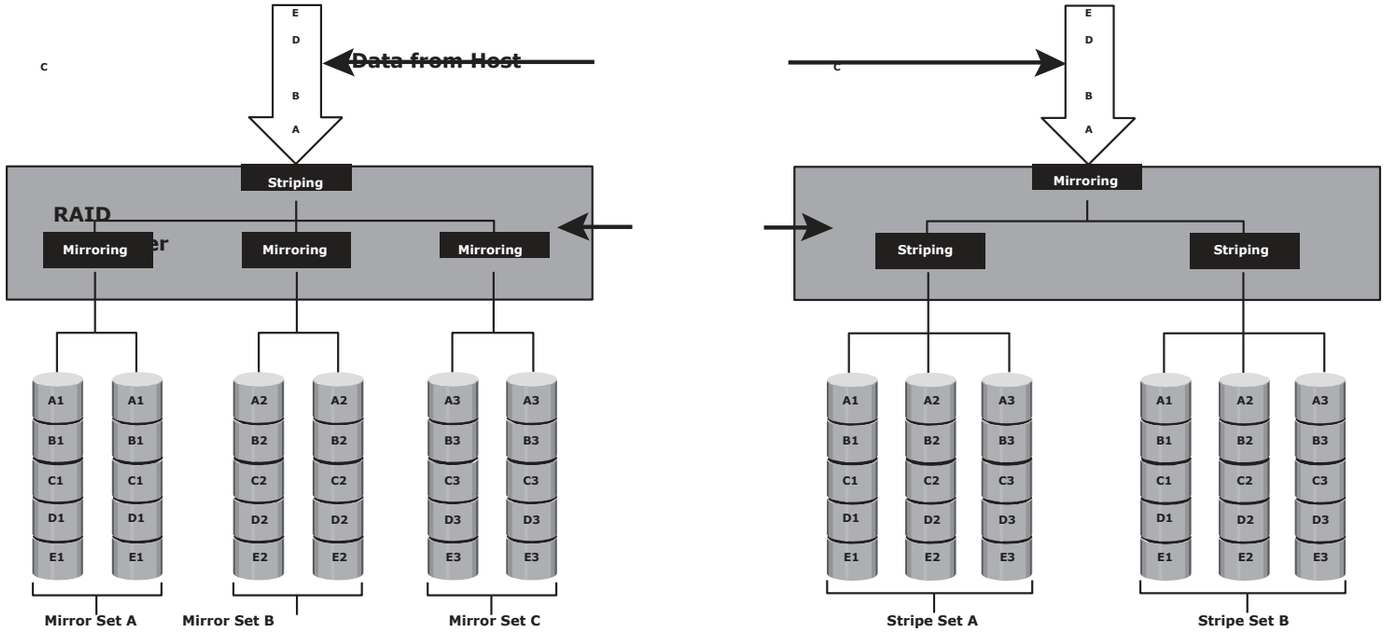
uses the mirror drive for data recovery. RAID 1 is suitable for applications that require high availability and cost is no constraint.



- 3-6: RAID 1

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 - 3-7).



(a) RAID 1+0 **(b) RAID 0+1**

- 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:

- High transaction rate Online Transaction Processing (OLTP)
- Large messaging installations
- 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 - 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 per- formed 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 - 3-7 [b]):

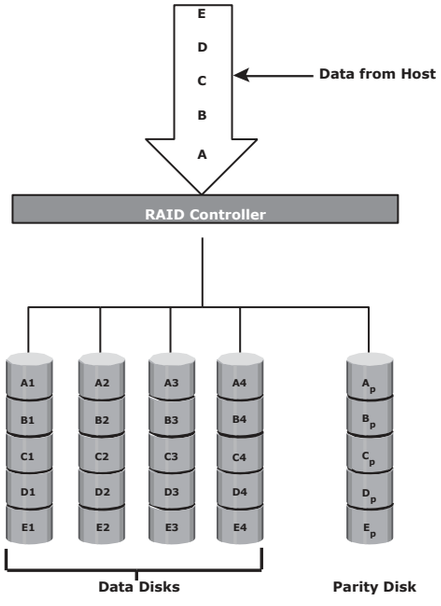
Drives 1 + 2 + 3 = RAID 0 (Stripe Set A)
Drives 4 + 5 + 6 = RAID 0 (Stripe Set B)

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.

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. - 3-8 illustrates the RAID 3 implementation.

RAID 3 provides good performance forapplicationsthat involvelarge sequen- tial data access, such as data backup or video streaming.



- 3-8: RAID 3

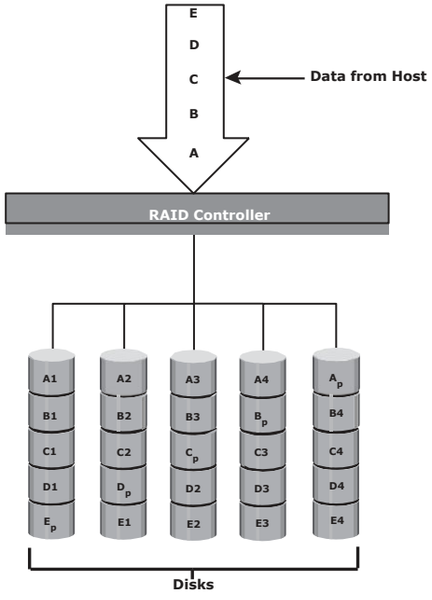
RAID 4

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 reason- able write throughput.

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. - 3-9 illustrates the RAID 5 implementation.

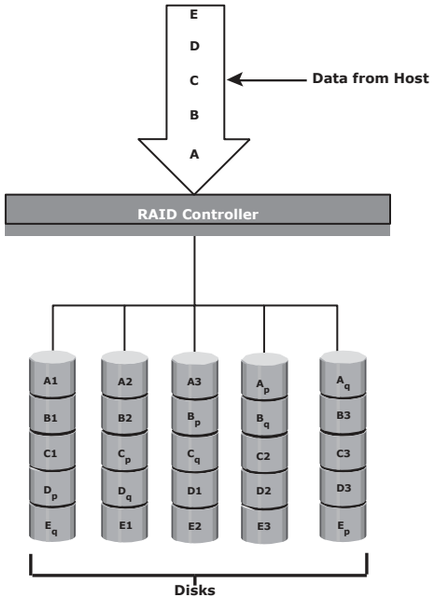


- 3-9: RAID 5

RAID 5 is good for random, read-intensive I/Oapplications 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.

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 - 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.



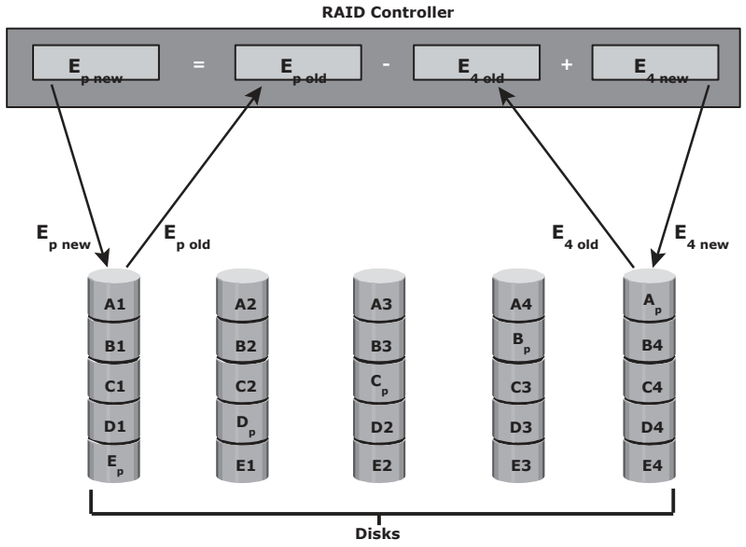
- 3-10: RAID 6

3.2 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 ina RAID 5 implementa- tion, 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.

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



- 3-11: Write penalty in RAID 5

The parity (P) at the controller is calculated as follows:

$$E_p = E_1 + E_2 + E_3 + E_4 \text{ (XOR operations)}$$

Whenever the controller performs a write I/O, parity must be computed by reading the old parity (E_p old) and the old data ($E_{4\ old}$) from the disk, which means two read I/Os. Then, the new parity (E_p new) is computed as follows:

$$E_{p\ new} = E_{p\ old} – E_{4\ old} + E_{4\ new} \text{ (XOR operations)}$$

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.

Application IOPS and RAID Configurations

When deciding the number of disks required for an application, it is important to consider the impact of RAID based on IOPS generated by the application. The total disk load should be computed by considering the type of RAID configura- tion 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)

RAID Comparison

Table 3-2 compares the common types of 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

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:

- If parity RAIDisused, thedataisrebuiltontothehotsparefromthe parity and the data on the surviving disk drives in the RAID set.
- 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 perma- nently. 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 auto- matically. 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 perfor- mance. However, the system is at risk of data loss if another disk failure occurs.