Manage Disk Storage on SANS

(HP World, Chicago, 2001)

Instructor: Jacob Farmer, Chief Technologist

Cambridge Computer Services, Inc.

About Cambridge Computer Services

- Over 10 years in the field of storage systems and storage management technologies.
  - Sales
  - Integration
  - Consulting
  - Training
- A large percentage of our business is subcontracted training and integration for industry giants (EMC, Compaq, Legato, etc.)
- Headquartered in Boston, MA
- Clients all over the world.

Other SAN-Related Activities

- Hired to write O’Reilly book on storage area networks and network attached storage.
  - Watch for it! Fall, 2001.
- Participating in Storage Network Industry Association SAN certification program.
  - Classes in Boston and on site all over the world
- Lectures at major conferences
  - HP World, Usenix LISA 2001, PC Expo SAN Summit, Disaster Recovery 2001, Contingency and Planning Management 2001.
- Private consulting and integration services.

Class Agenda

Chapter 1	SAN And SCSI Refresher
Chapter 2	Partitioning the SAN
Chapter 3	Disk Storage on a SAN
Chapter 4	Features and Benefits of Intelligent Disk Systems
Chapter 5	Comparing SAN to NAS
	Conclusions and Questions & Answers

Goals of This Class

- Solidify your understanding of basic SAN connectivity
- Get intimate with the problem of and solutions for SAN partitioning.
- Learn to compare SAN disk sharing technologies so that you can make educated purchasing and design decisions.

Chapter 1

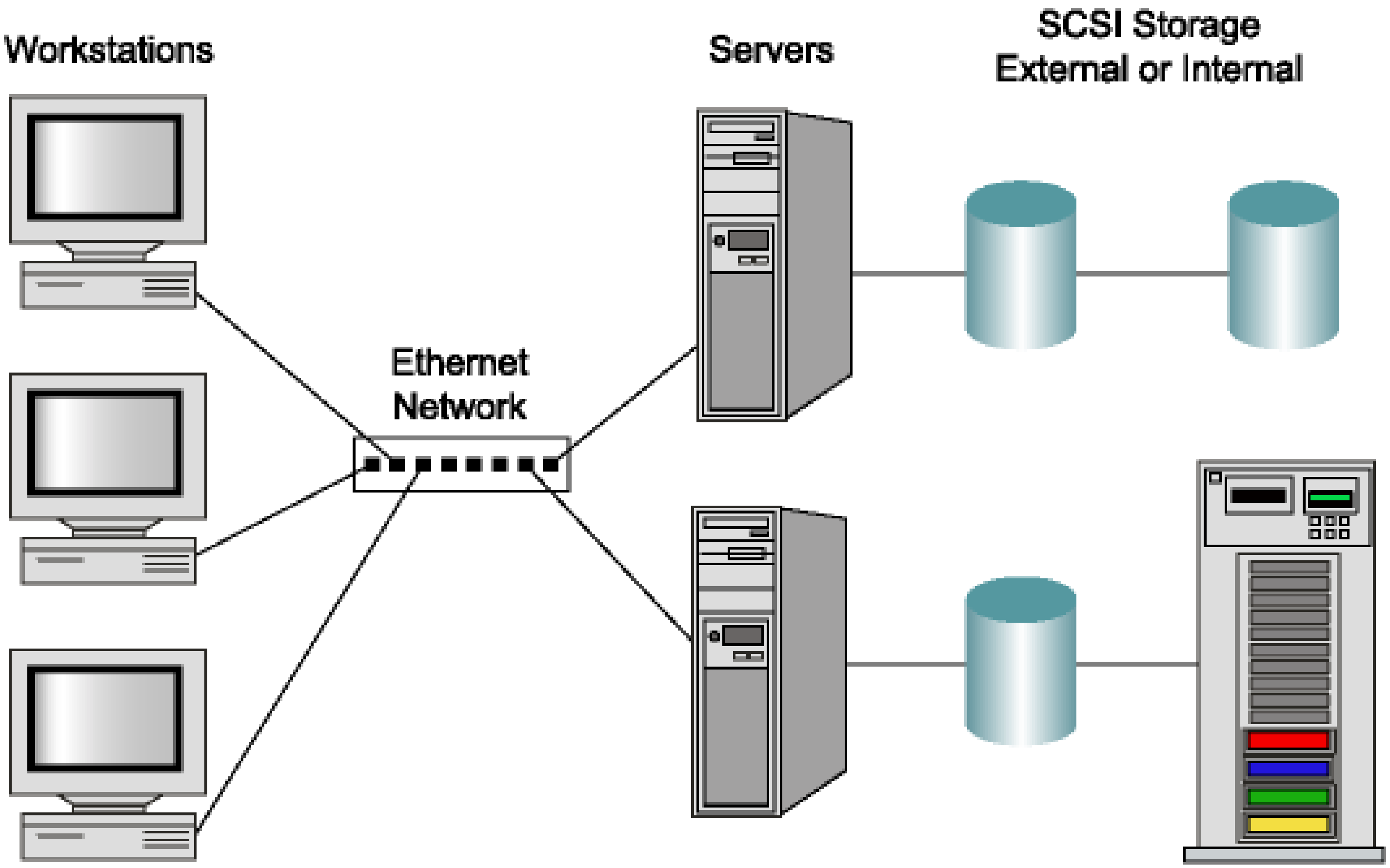
SAN and SCSI Refresher

SNIA’s SAN Definition

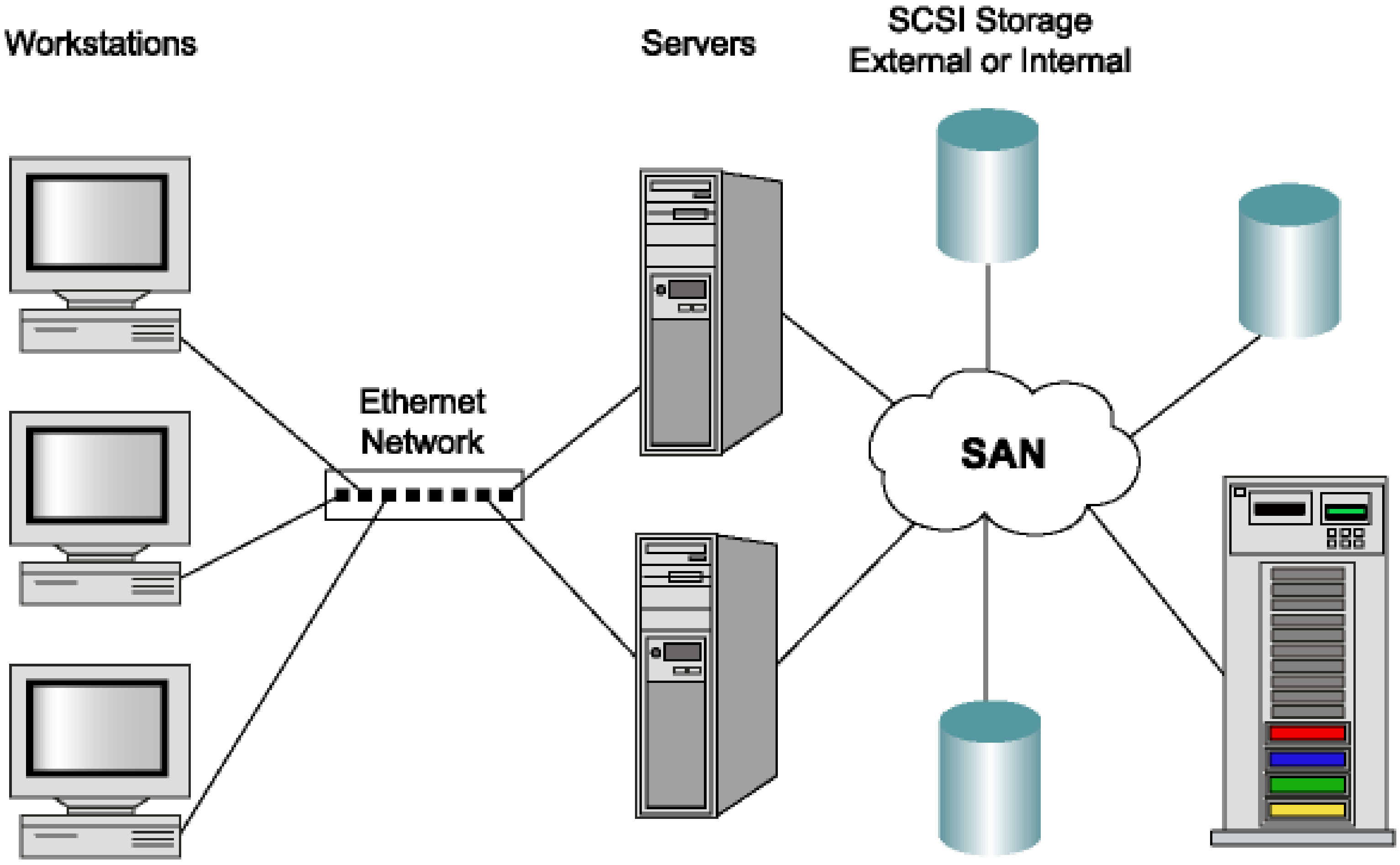
A network whose primary purpose is the transfer of data between computer systems and storage elements and among storage elements. Abbreviated SAN. A SAN consists of a communication infrastructure, which provides physical connections, and a management layer, which organizes the connections, storage elements, and computer systems so that data transfer is secure and robust.”

Source: Storage Networking Industry Association
http://www.snia.org

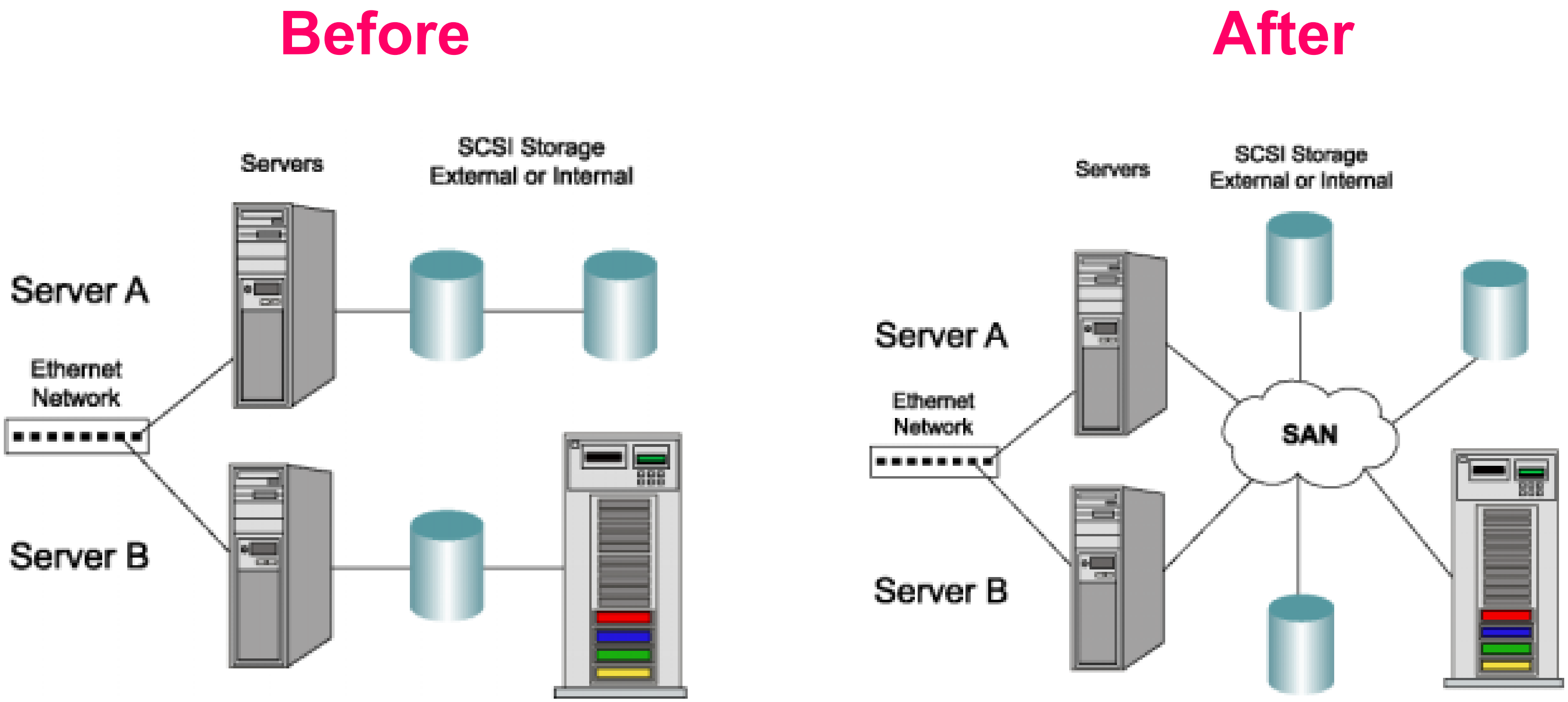
Storage Infrastructure Today



The Same Components as a SAN

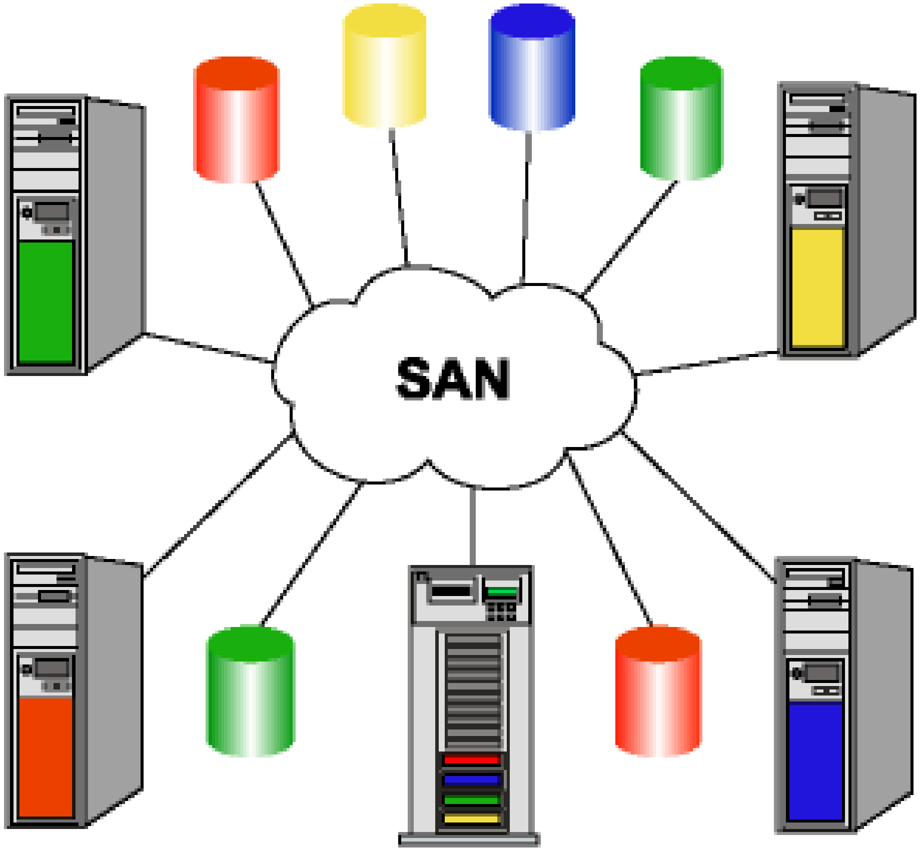


Traditional LAN vs. SAN



A SAN simply connects multiple hosts to a common set of storage devices.

SAN: A Practical Definition



SCSI in a Star Configuration

Important:

Fibre Channel is a form of SCSI

SCSI = Command Set + Data Transport

- SCSI can be separated into
  - Data transport
  - SCSI command protocol

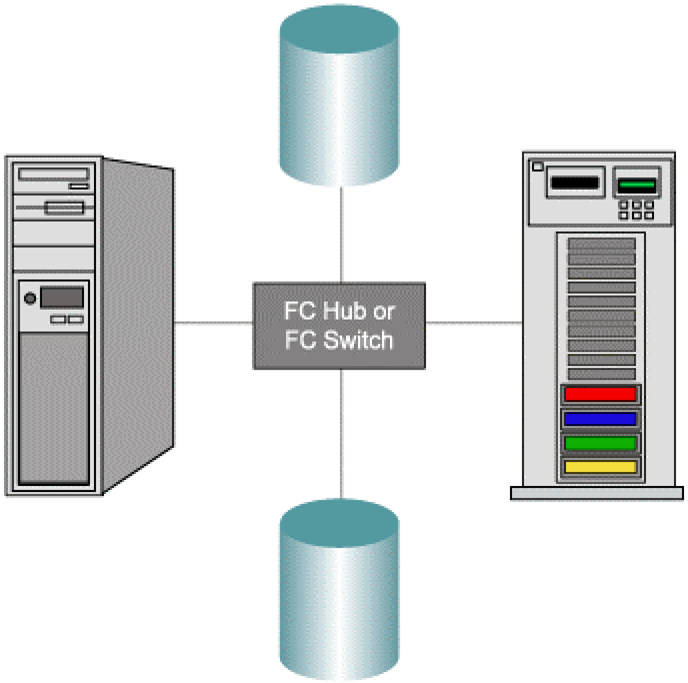
- The SCSI protocol can be run on alternative data transports.
  - Similar to the way that TCP/IP and other network protocols can run over Ethernet, Token Ring, FDDI, etc.

Fibre Channel = Serial SCSI

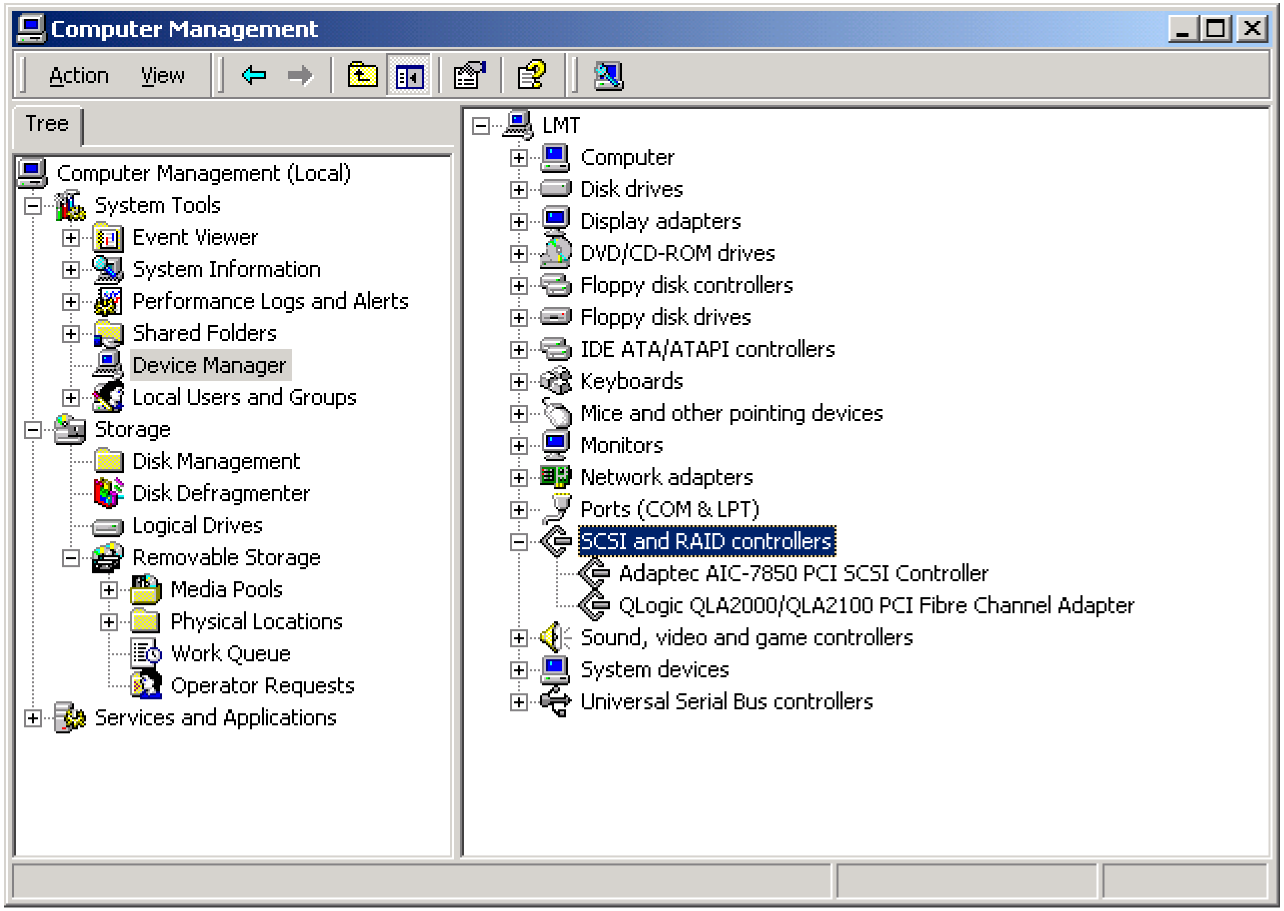
- Industry anticipated problems of parallel SCSI bus.
- ANSI released SCSI-3 standard in 1994
  - Standard for serial SCSI in a star topology
  - Also known as “SCSI-3 Serial”
- Fibre channel uses Fibre Channel Protocol (FCP), a version of the SCSI-3 serial standard

Fibre Channel = SCSI in a Star

- Ethernet evolved from a bus architecture (10Base-2) to a star architecture (10Base-T & 100Base-T)
- With fibre channel, SCSI has followed suit.

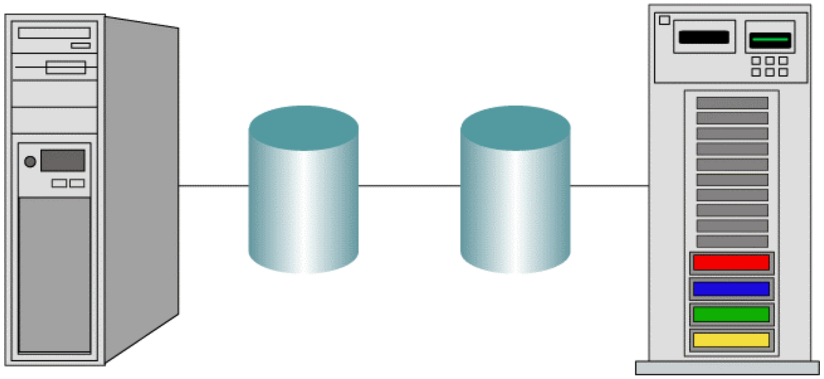


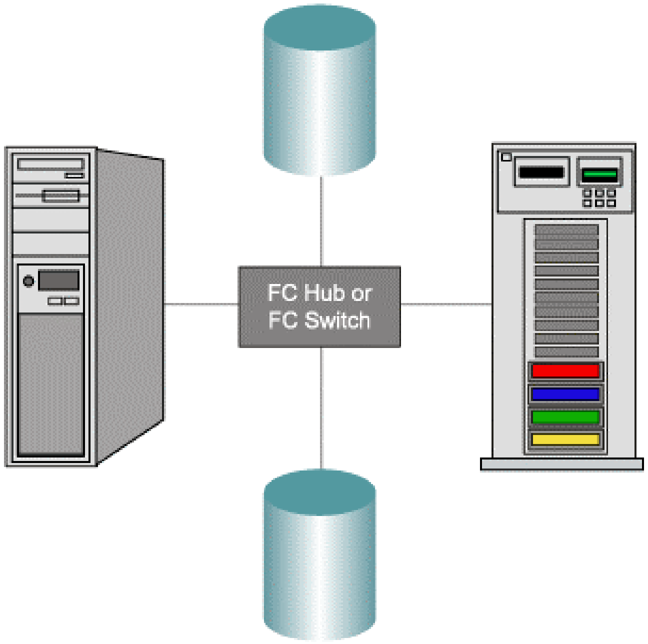
Fibre Channel HBA Installation



Fibre Channel: Host’s Perspective

- Host computers can not tell the difference between Fibre channel and SCSI.
- Fibre channel storage is installed and configured identically to SCSI storage





Review of SCSI Addressing

- The operating system uses three things to identify SCSI devices:
  - SCSI Channel
  - SCSI ID
  - SCSI LUN
- Most of the time the ID is the differentiator
  - Many SCSI systems only have one channel
  - Most of the time the LUN is set to zero
- But: all three factors make up the complete address

LUNs are a Subset of the SCSI ID

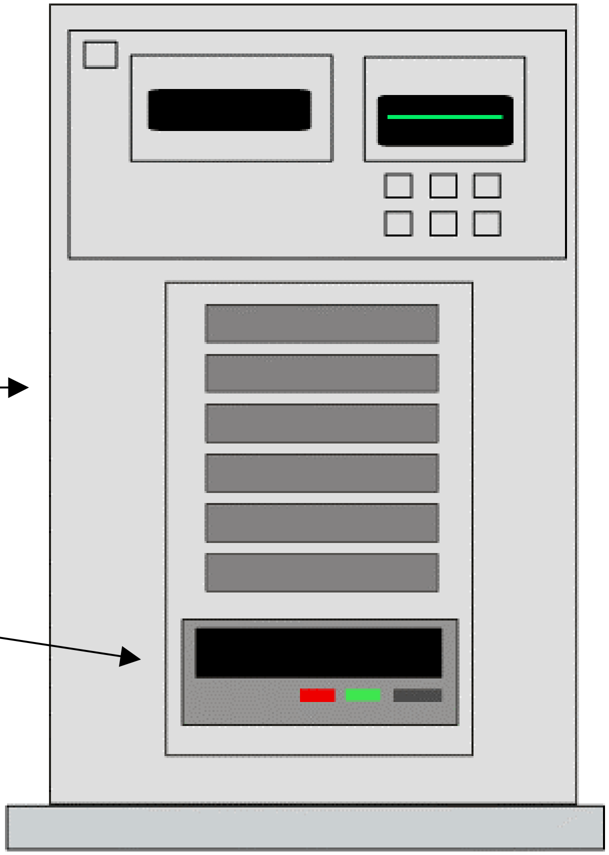
- SCSI uses IDs to distinguish devices on the same bus.
- LUNs (Logical Unit Numbers) are a subset of SCSI IDs
  - SCSI ID = Street Address
  - LUN = Apartment Number

SCSI Type	# IDs	ID Range	LUNs per ID
Narrow SCSI (8 Bit)	8 ID's	0-7	8 (64 total)
Wide SCSI (16 Bit)	16 ID's	0-15	Unlimited*

**The actual number of LUNs depends on operating system and driver support.*

Example: SCSI LUNs in a Tape Loader

Tape Library
SCSI ID = 3

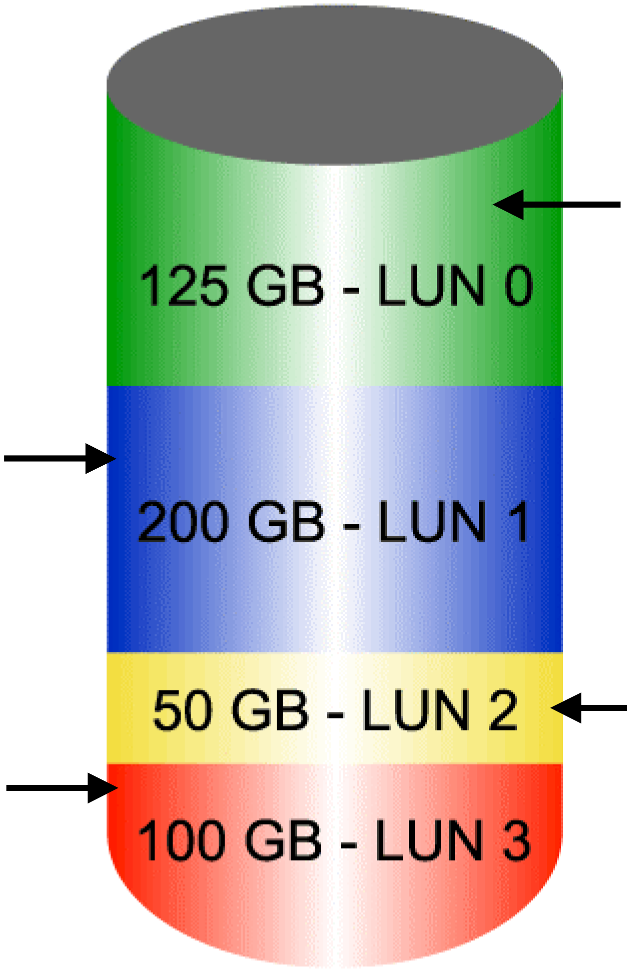


Robotics = ID3, LUN 0

Tape Drive = ID 3, LUN 1

Example: SCSI LUNs in a RAID Array

RAID System
SCSI ID = 1



125 GB Partition: ID 1, LUN 0

200 GB Partition: ID 1, LUN 1

50 GB Partition: ID 1, LUN 2

100 GB Partition: ID 1, LUN 3

Fibre Channel Addressing

World Wide Name (WWN)

- 64-bit unique name
- Similar to Ethernet MAC address
- Tied to hardware (assigned to ports and nodes)
- Usually assigned by the IEEE (each manufacturer is assigned a range)
- Globally unique
- Most reliable way to address a specific device

Chapter 3

Partitioning the SAN

The Disk I/O Path

Switch Zoning

LUN Masking

Partitioning The SAN

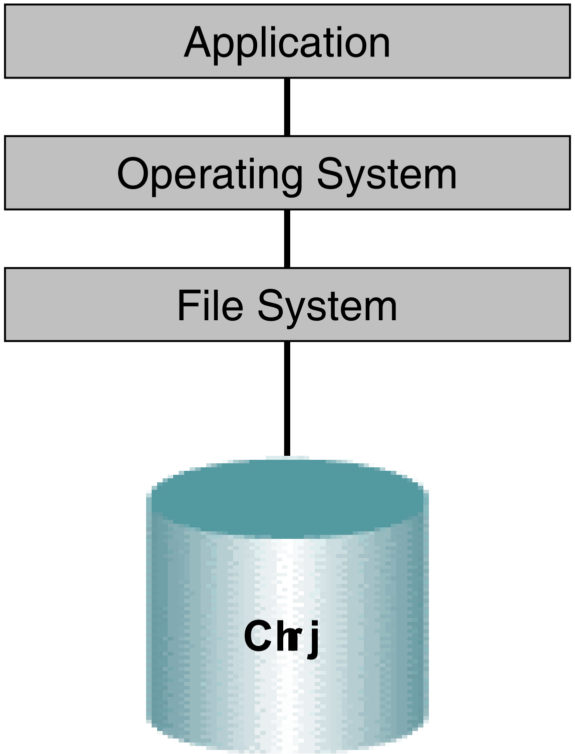
- Partitioning the SAN involves designating which devices on the SAN have access to other devices.

- Big SANs are broken down into mini, virtual SANs.
  - What would happen if one did not partition the SAN?

Where Can We Insert a Virtualizer?

- You can’t teach an old operating system a new trick
- Must support legacy storage devices
- New technology must be “inserted”
  - Fortunately, disk I/O can be sliced into layers of abstraction

Basic Disk I/O Path



Application

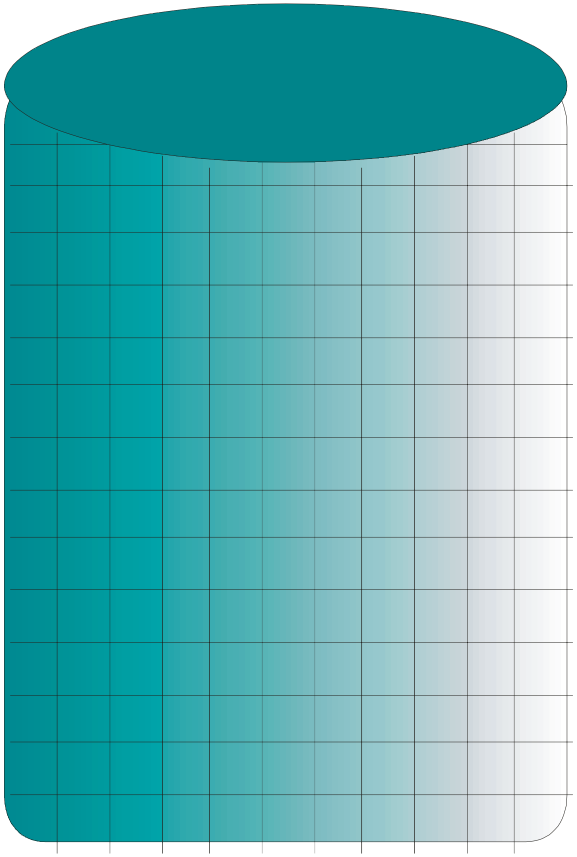
Operating System

File System

Chrj

Logical Blocks

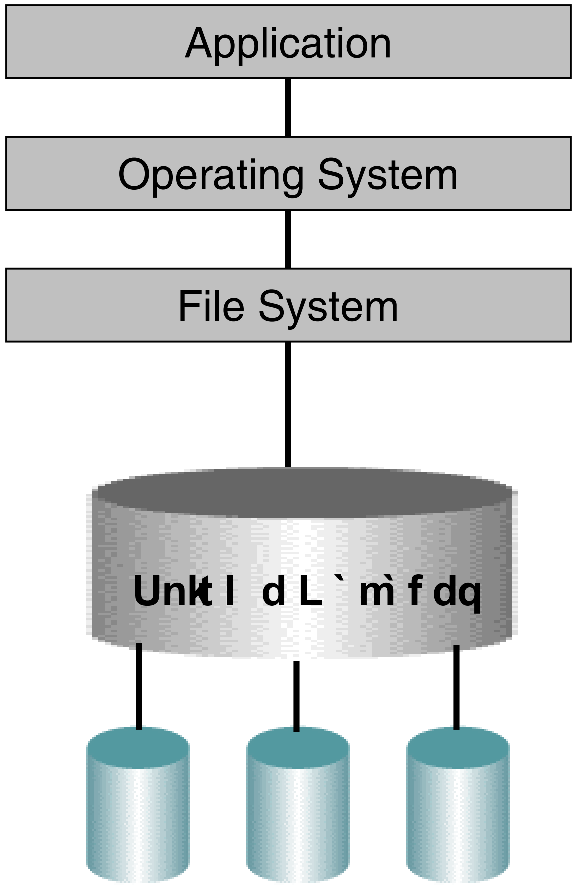
The logical block is the smallest unit of measurement for data storage. Storage devices (disks, tape, etc.) are represented as a bunch of logical blocks.



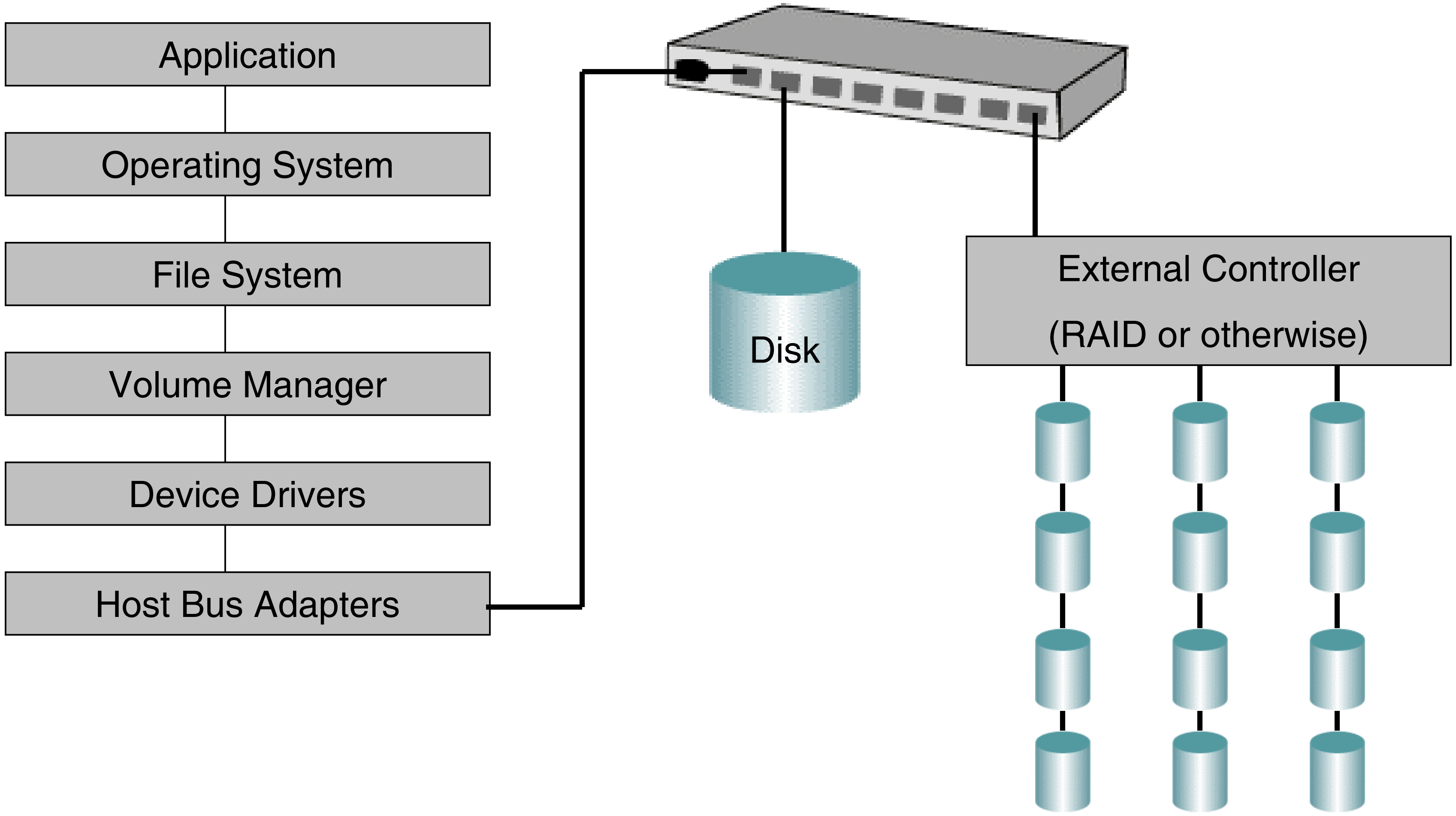
Files and File Systems

- Logical Blocks are organized into files by the file system.
- Examples of file systems:
  - FAT and FAT-32 (DOS, Windows 95, Windows 98)
  - NTFS (Windows NT, Windows 2000)
  - UFS (Unix File System)
  - ISO 9660 – CD-ROM file system
- Third Party File Systems – Veritas

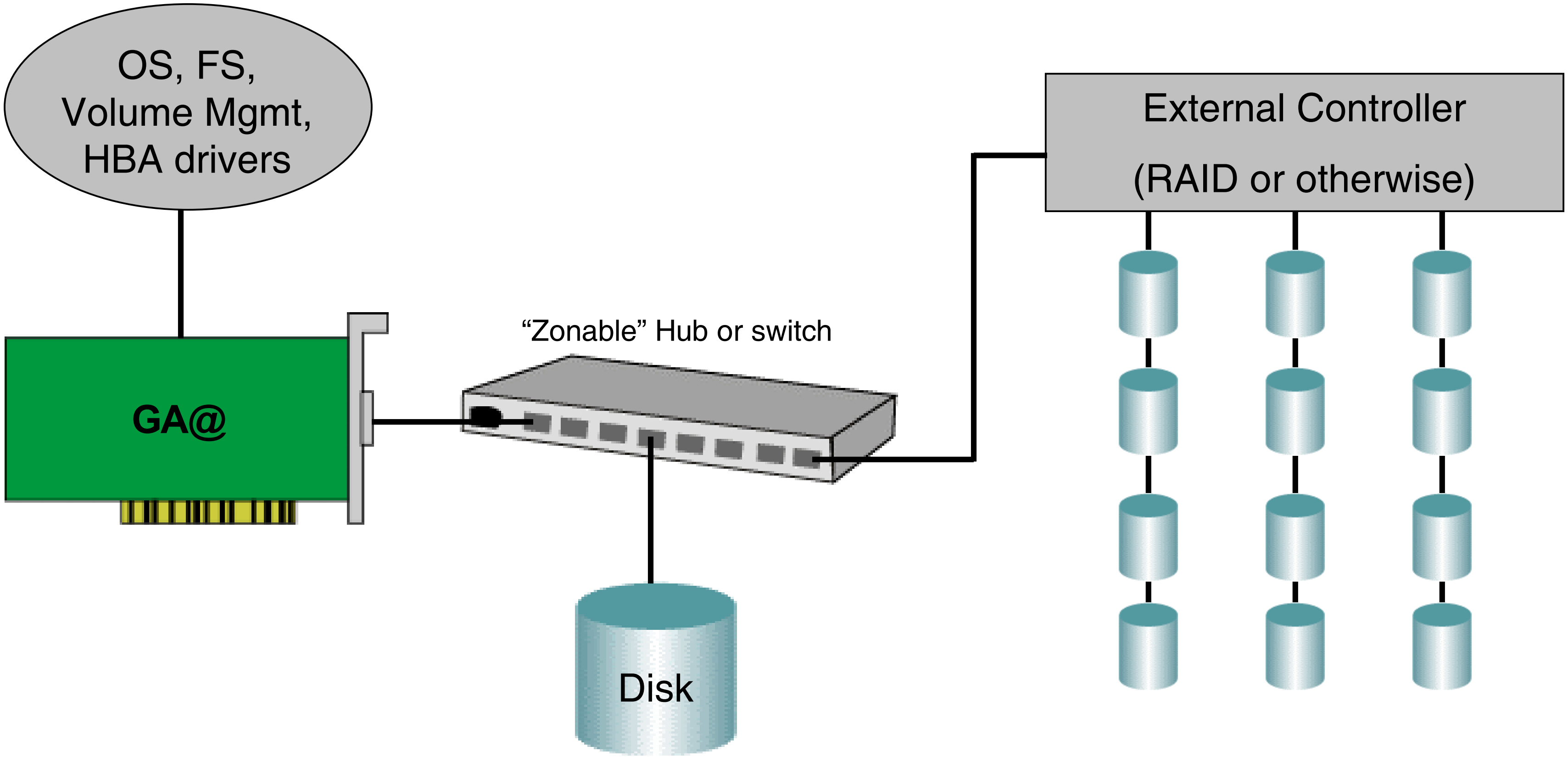
Abstraction With Volume Manager



More Detailed View Including Hardware



Simplified View of Data I/O Path



Summary: Virtualizer Insertion Points

- Software running on the computer
- Firmware on the host bus adapter
- At the switch or hub
- Inside an intelligent storage device
  - e.g. RAID sub-system
- By inserting an intelligent device such as a SCSI-FC router anywhere in the physical data path

Simplified Summary: Virtualizer Insertion Points

- At the host computer
- At the disk system
- Somewhere in the middle

Question: Which makes the most sense?

Two Main Technologies for Partitioning

- **Switch Zoning**
  - Layer 2 filtering
  - Protocol independent
- **LUN Masking**
  - SCSI Protocol filtering at the specific device level
  - Same as LUN Assignment

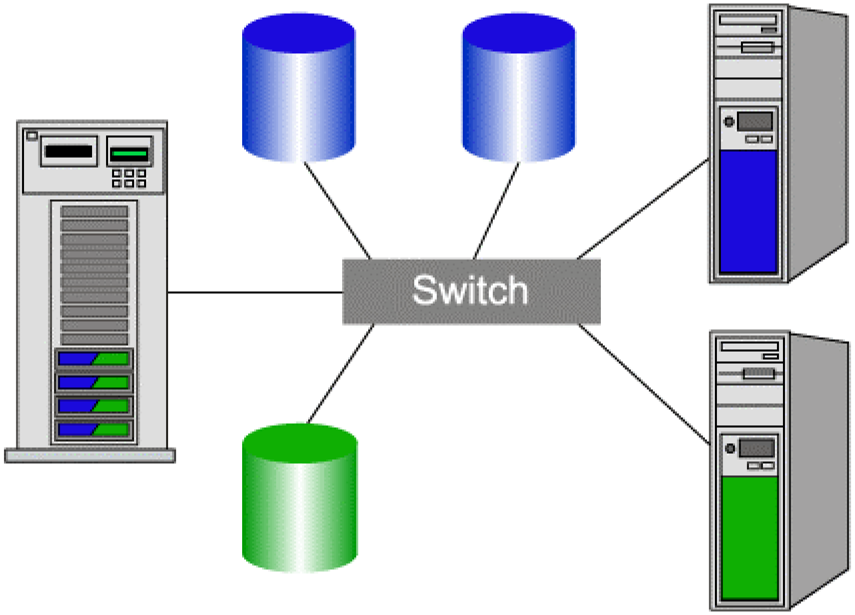
Switch Zoning

- Makes mini virtual SANs out of all of the devices on the SAN
- Can be configured by port number or WWN (world wide name)
- Any given device can be a member of multiple zones

- Note: Zoning is often sold as an option.

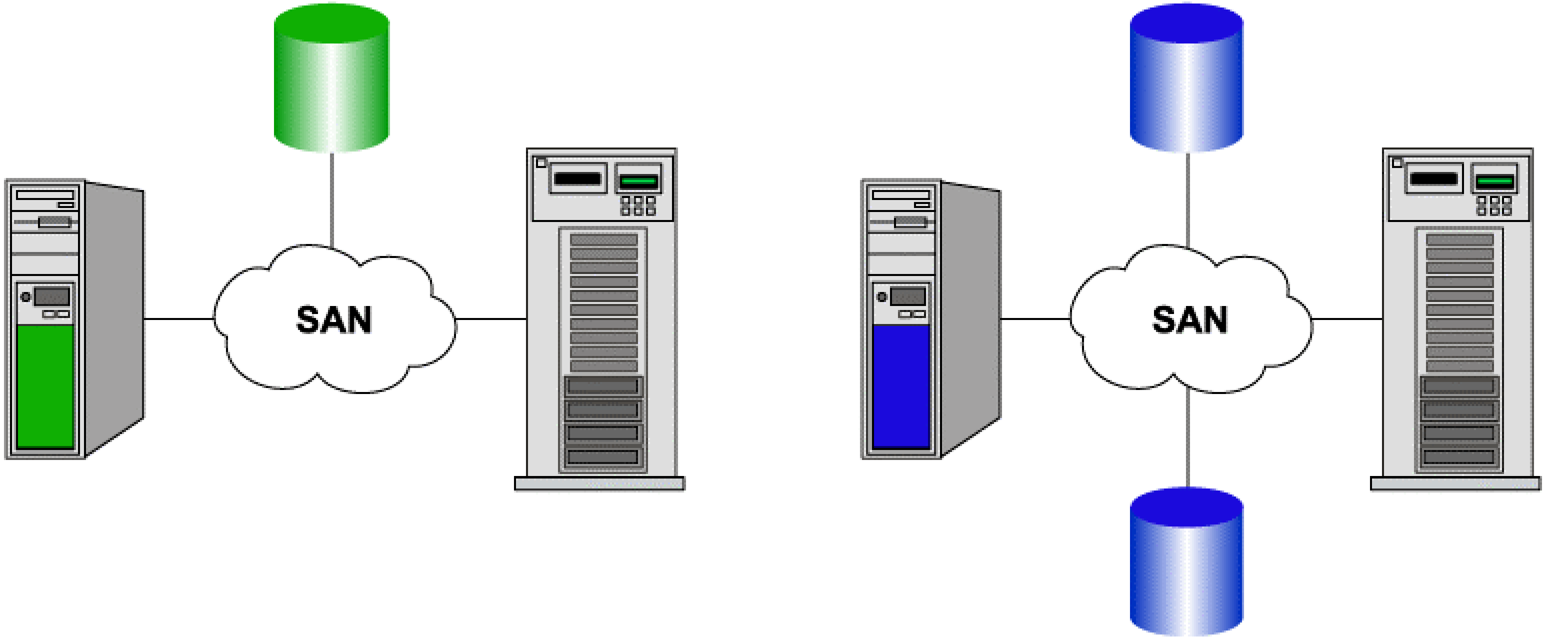
Switch Zoning

A single device can be in multiple zones:



- **Blue devices: Zone 1**
  - Blue hard drives cannot be used by green server
  - Green hard drive cannot be used by blue server.
- **Green devices: Zone 2**
  - Green server can only use the green disk
- **Tape library is in both Zone 1 and Zone 2**
  - Both servers can share the tape library

Switch Zoning



The switch creates two smaller "virtual" SANs

Zoning Limitations

- Switch zoning is OSI Layer 2
- Not SCSI protocol aware
- Only recognizes device IDs (not LUN-aware)
- Cannot sub-divide a single device
  - Cannot be used for sharing a central disk array or tape library
- Zoning alone is not usually enough
- Often sold separately

LUN Masking

- LUN: Logical Unit Number
- Subset of the SCSI or FC_AL or Fabric ID
  - If ID = street address, LUN = apartment number
- Sub-Partitions in a RAID system are usually presented as LUNs
- Individual tape drives in a library are presented as LUNs

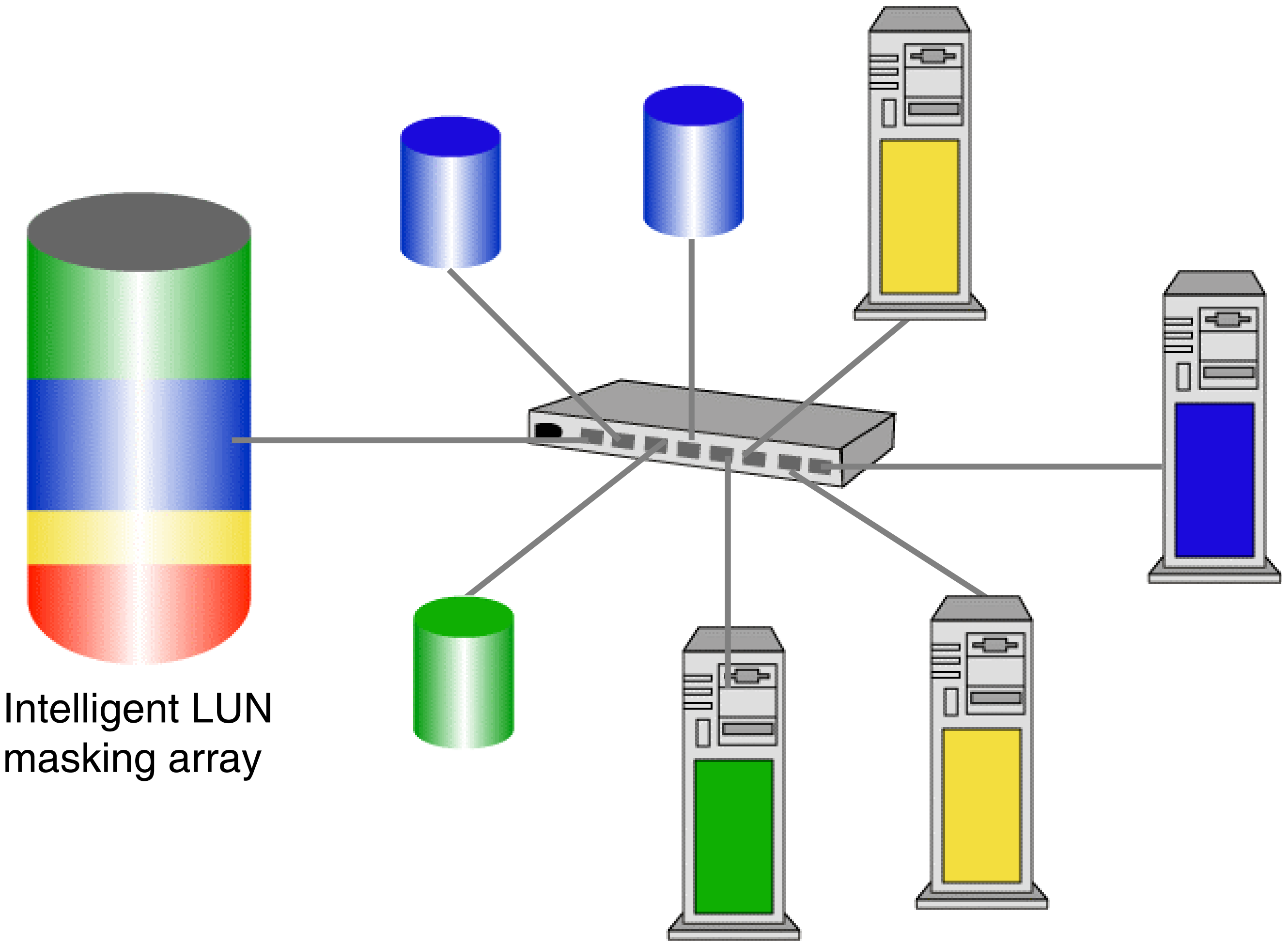
Possible LUN Masking Locations

- Host bus adapter
- Disk system
- Somewhere in the middle
  - FC-SCSI Router
- Not likely to be on the switch because switch is layer 2 device that is not SCSI protocol aware.

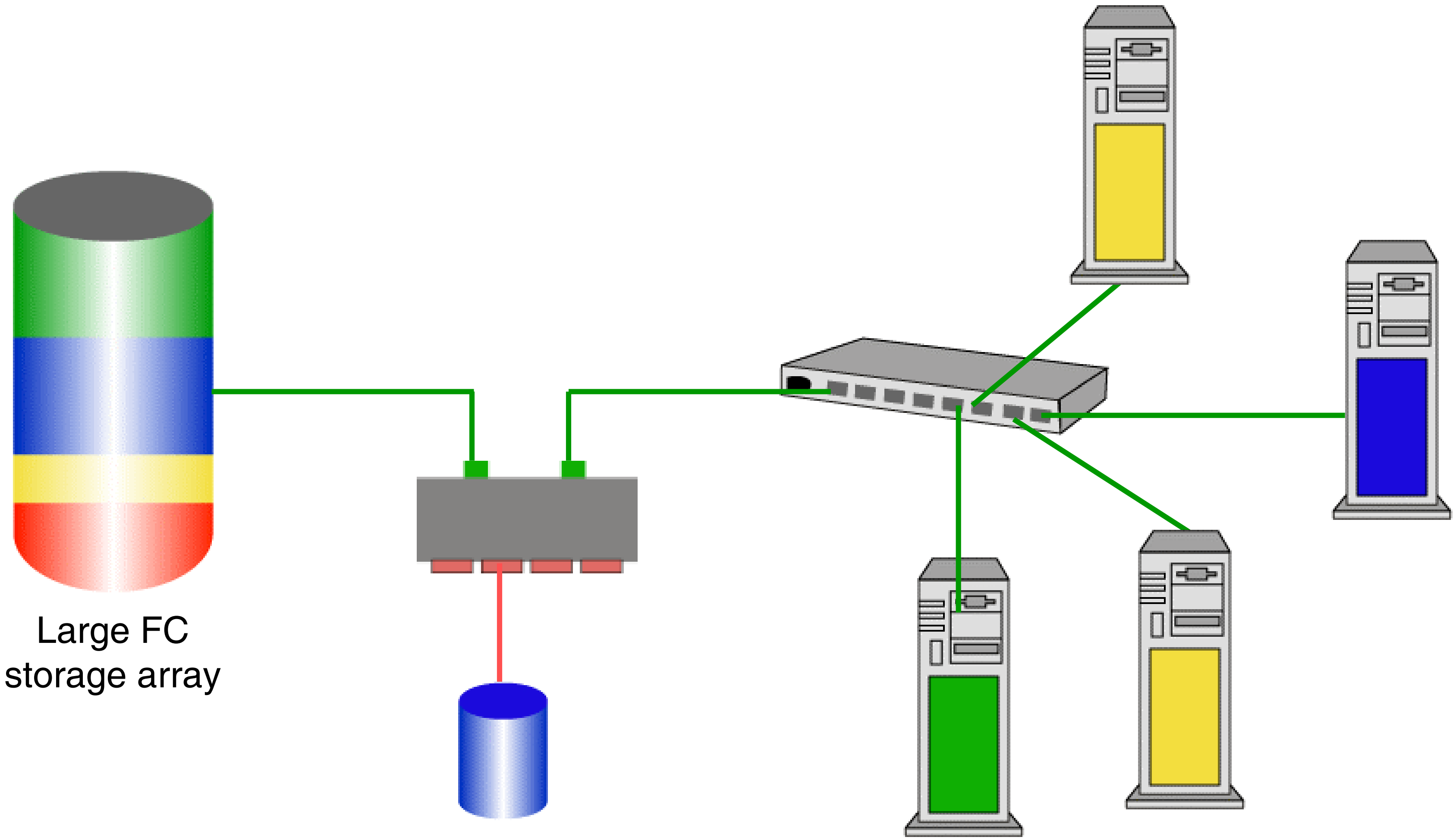
Devices Capable of LUN Masking

- Host Bus Adapter (driver or firmware)
- Fibre Channel - SCSI Routers
  - Usually used for connecting tape drives to SAN
  - Can be used for connecting hosts and disks
- Intelligent Disk Controllers
  - Like EMC Symmetrix and Modern Arrays
- Disk Virtualizers (a.k.a. “SAN Appliances”)

LUN Masking on Disk Controller


Intelligent LUN masking array

LUN Masking with FC-SCSI Router



Summary of Partitioning

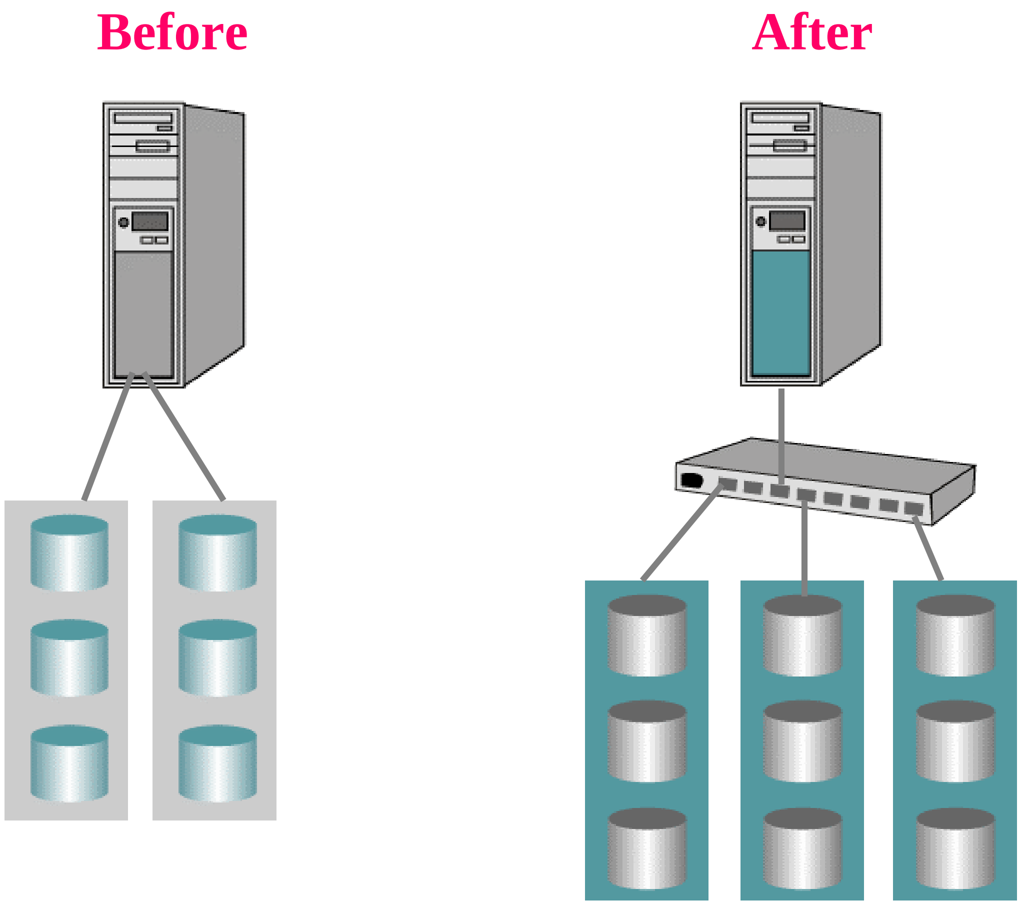
- **Switch Zoning**
  - Macro view division of SAN into logical mini SANs.
  - Happens at the switch, if the switch can do it.
- **LUN Masking**
  - Detailed sub-division of resources
  - Happens at the target, initiator, or somewhere in between. Many devices can do it.

Chapter 3

Disk Storage on the SAN

1) Single Zones and Simple Zoning
2) LUN Masking
3) In-Band (Symmetrical) Disk Virtualization
4) Out-of-Band (Asymmetrical) Disk Virtualization
5) SAN File Systems

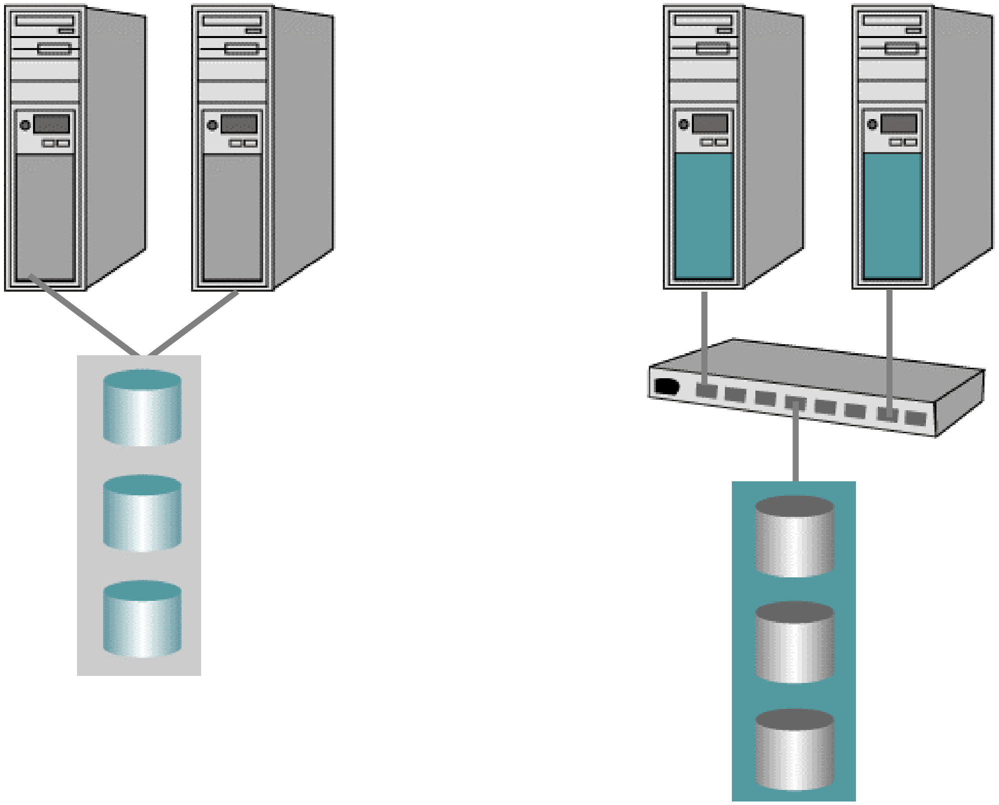
No Partitioning: Fibre Channel in lieu of Parallel SCSI



Before

After

No Partitioning:
Simple SAN for Clustering



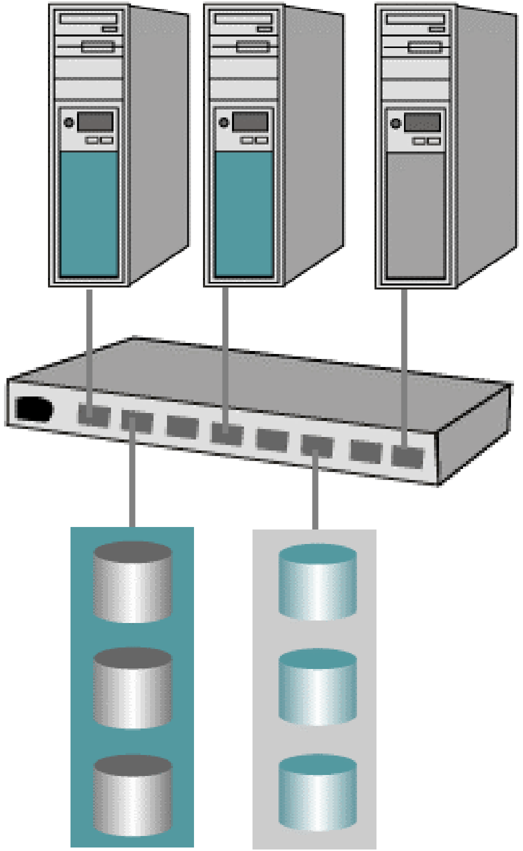
1) Switch Zoning

- Partition by zoning at the switch level
- Servers and storage devices are privately zoned to one another
- Advantages
  - Secure
  - Reliable
  - Relatively inexpensive
- Disadvantages
  - No real device sharing or resource optimization

Partitioning With Switch Zoning

Two teal servers (probably a cluster) share the teal disk array

One gray server has exclusive use of the gray disk array.



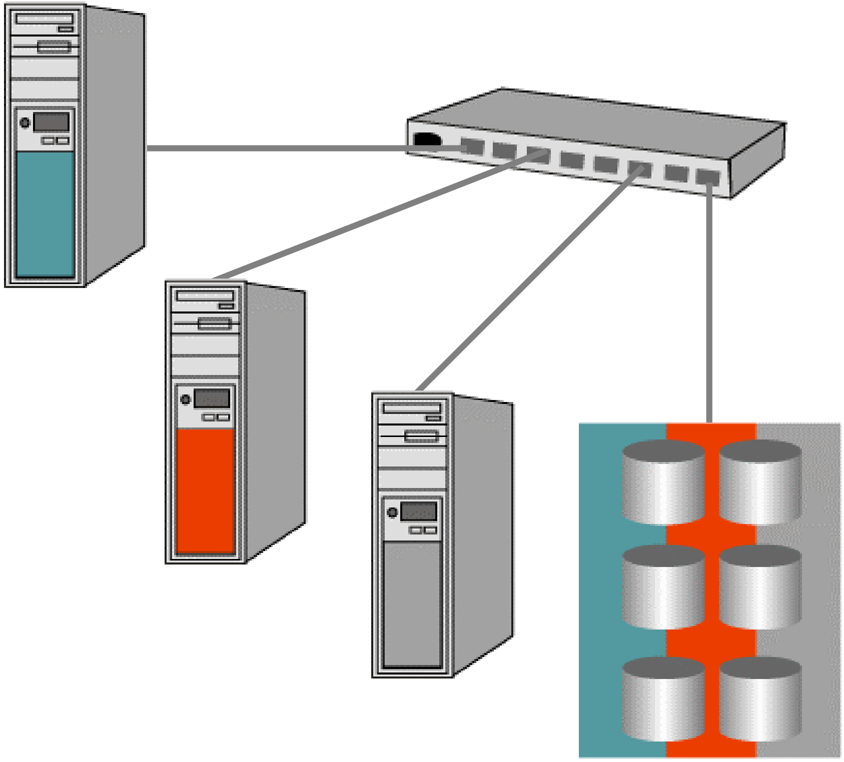
2) LUN Masking to Central Disk Array

LUN masking can occur at several locations

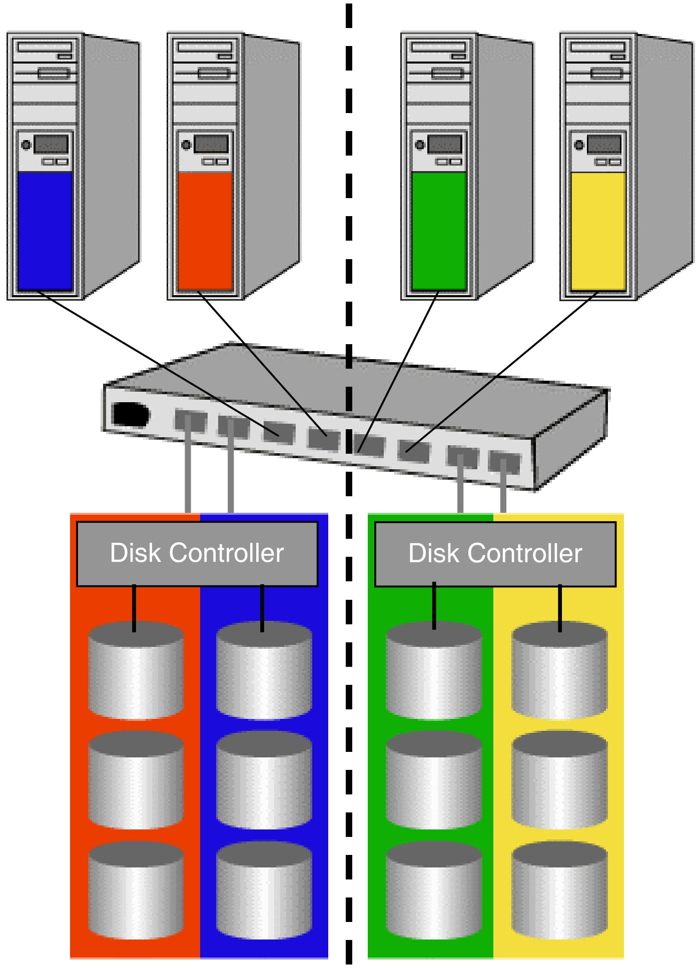
Host: HBA firmware or software drivers

Disk: With intelligent array controller

In Between: Via LUN masking SCSI router (not shown)



Hybrid of Zoning and LUN Masking

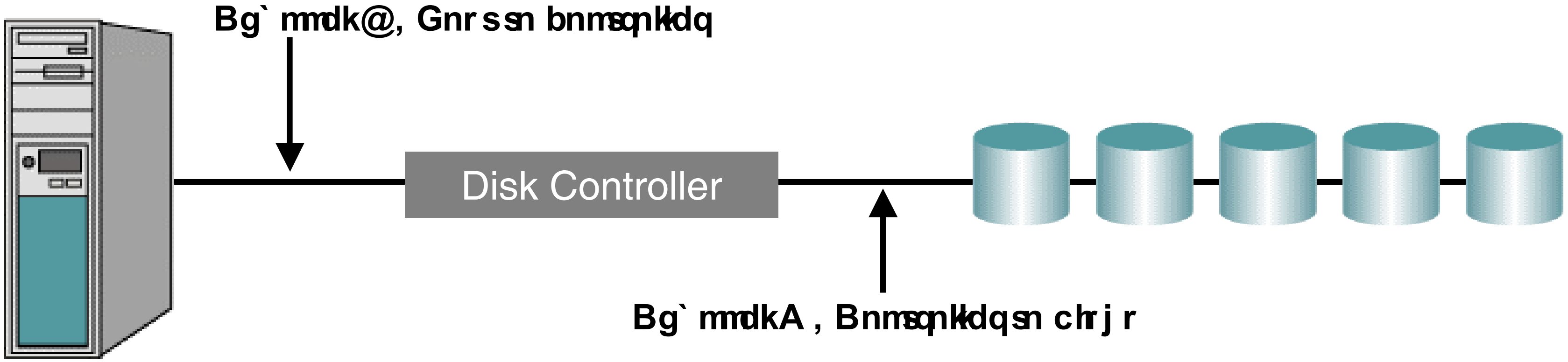


Advanced Technologies for Disk Sharing

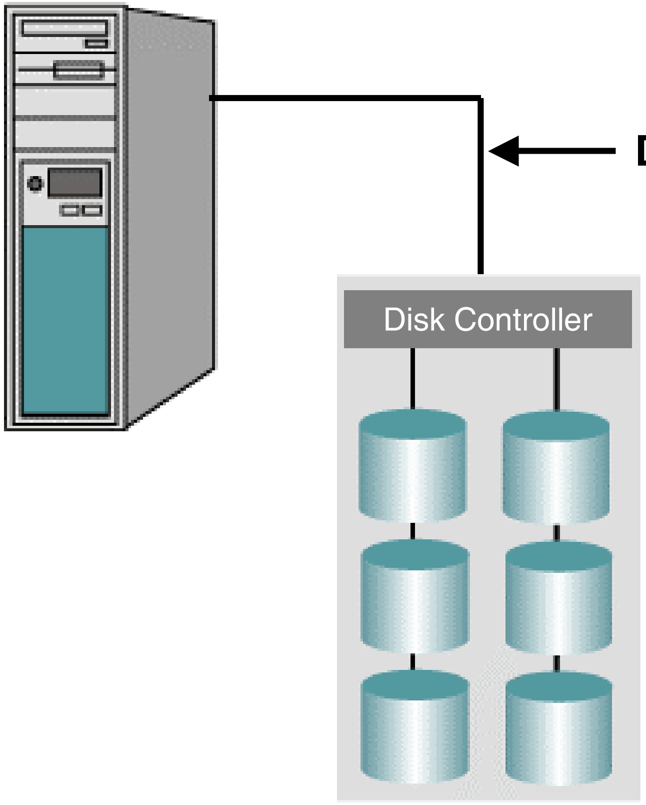
- 3) In-band (symmetrical) disk virtualization
- 4) Out-of-band (asymmetrical) disk virtualization
- 5) SAN File Systems

In-Band RAID Controllers & Disk Virtualizers

- Some computational device sits between the disk and the host in the path of the data.
- Five 75 GB drives behind the RAID controller look like one big 300 GB drive.



Usually Packaged Like This



Dwsdqm`k RBRH'EB(Bnmmdbshnmr

Hmsdqm`k Oqhu`sd RBRH'EB(Bnmmdbshnmr

Common Examples:

EMC Symmetrix and Clariion

IBM Shark

Hitachi, Xiotech, Dell, Compaq (some), etc.

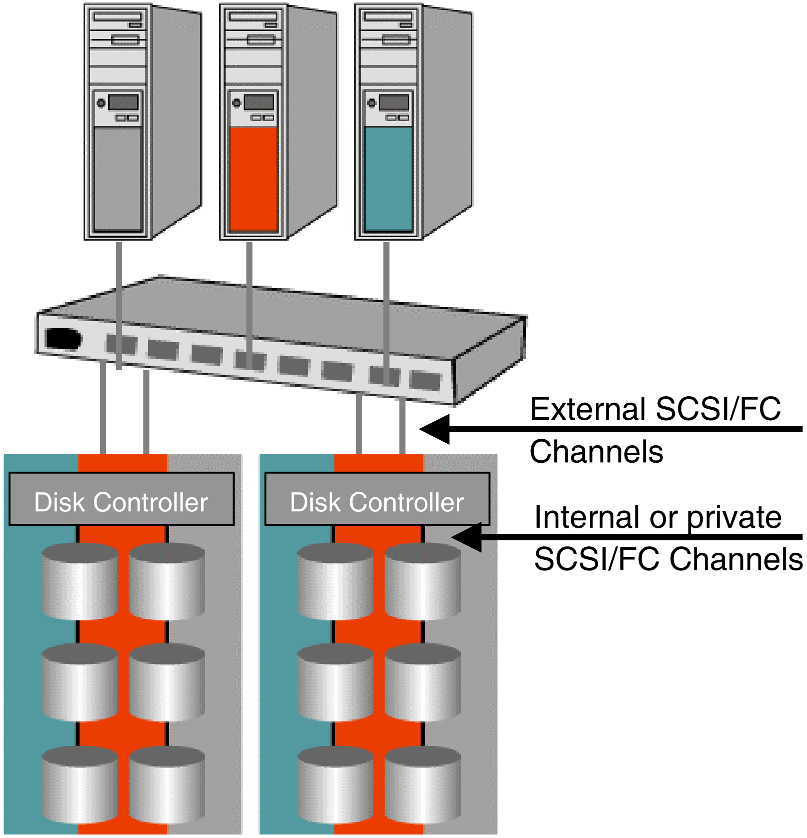
Benefits of In-Band Controllers

- Disk controller acts like a firewall for disks
  - LUNs must be specifically assigned to hosts via WWN
- Administration is centralized at the disk sub-system.
- Opportunity to insert sophisticated caching
- Opportunity to add another layer of abstraction in the path from application to data

In-Band Control: Considerations

- Not all RAID controllers are capable of LUN masking
- RAID is a tiny subset of disk virtualization
  - This issue will be covered in more depth later in the course
- Many RAID systems are marketed as disk virtualizers. Don’t be fooled.

Possible Problems: Traditional In-Band Disk Virtualization


External SCSI/FC Channels
Internal or private SCSI/FC Channels
Disk Controller
Disk Controller

1) Disk subsystem has maximum disk capacity. One pays a premium for proprietary disk cabinetry

2) Disk controller becomes potential I/O bottleneck

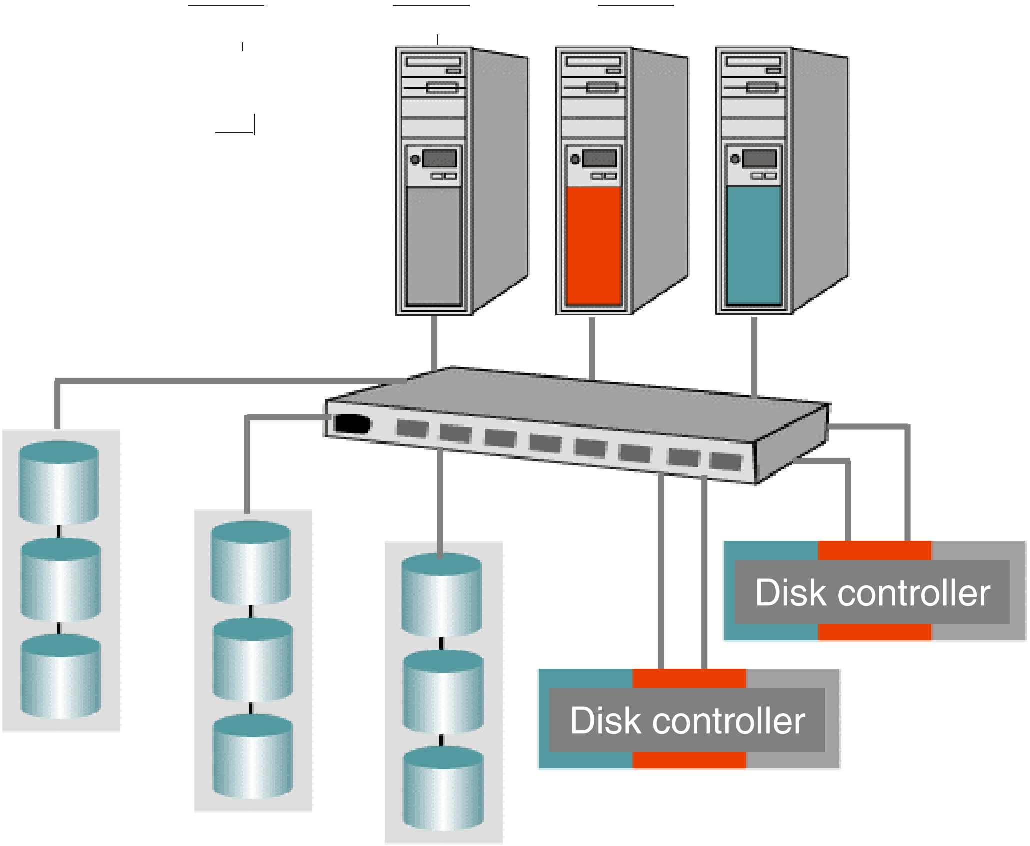
3) More capacity and more I/O only by adding additional units

4) Central point of failure

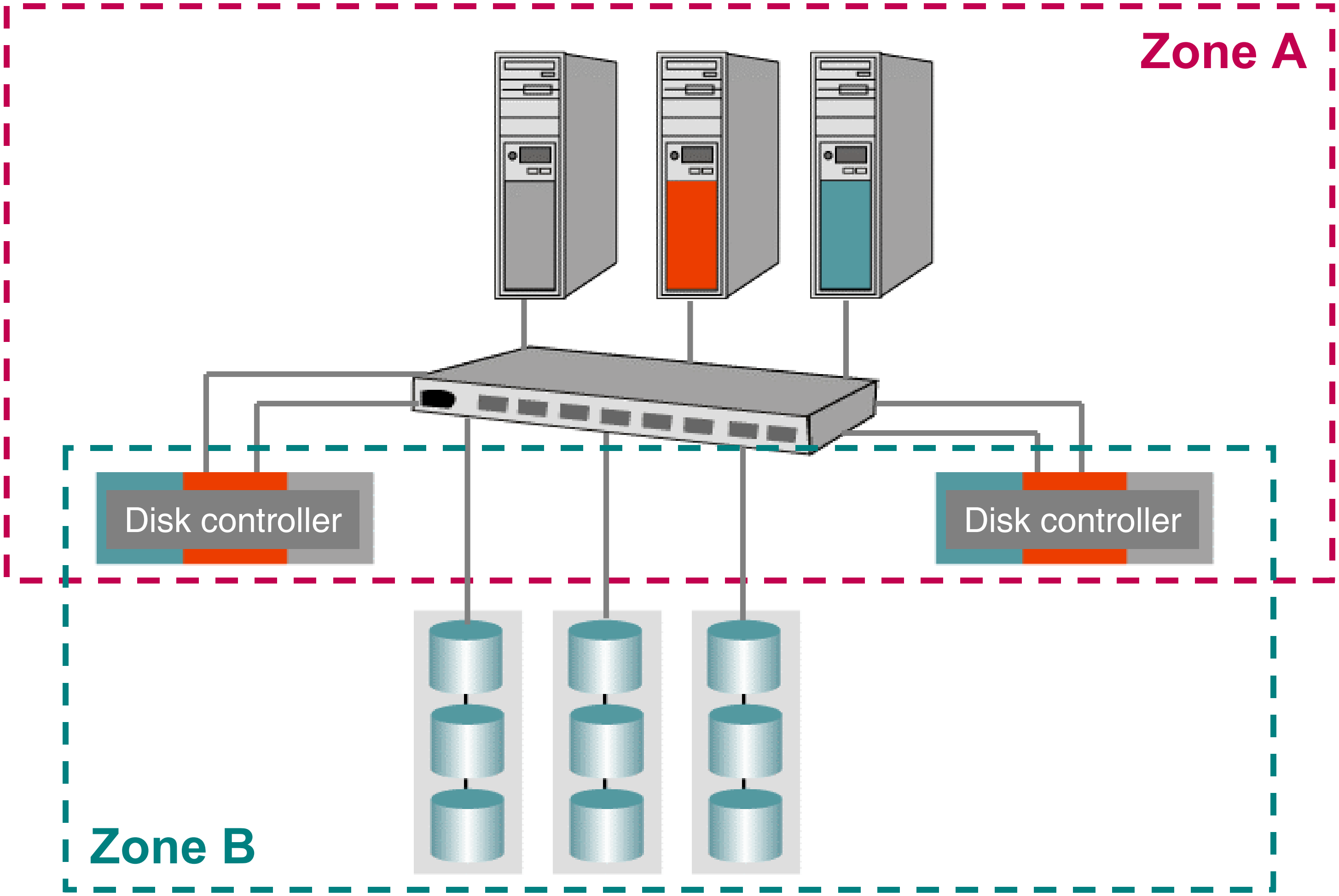
Solution: Separate Disk and Controller

Replace internal private disk channel with “virtual SAN” by zoning the switch

Disk I/O must still pass through the controllers, but becomes open and scalable.



3) Scalable In-Band Disk Virtualization



Open System Disk Virtualization: Benefits

- Use any disk systems for the disks
  - Legacy disk devices
  - JBODs and RAID systems
  - Even EMC Symmetrix, IBM Shark, etc.
- Use ordinary computers for the disk controllers
- Scale disk independent of I/O
- Centrally administer all storage resources

Scaling in Performance

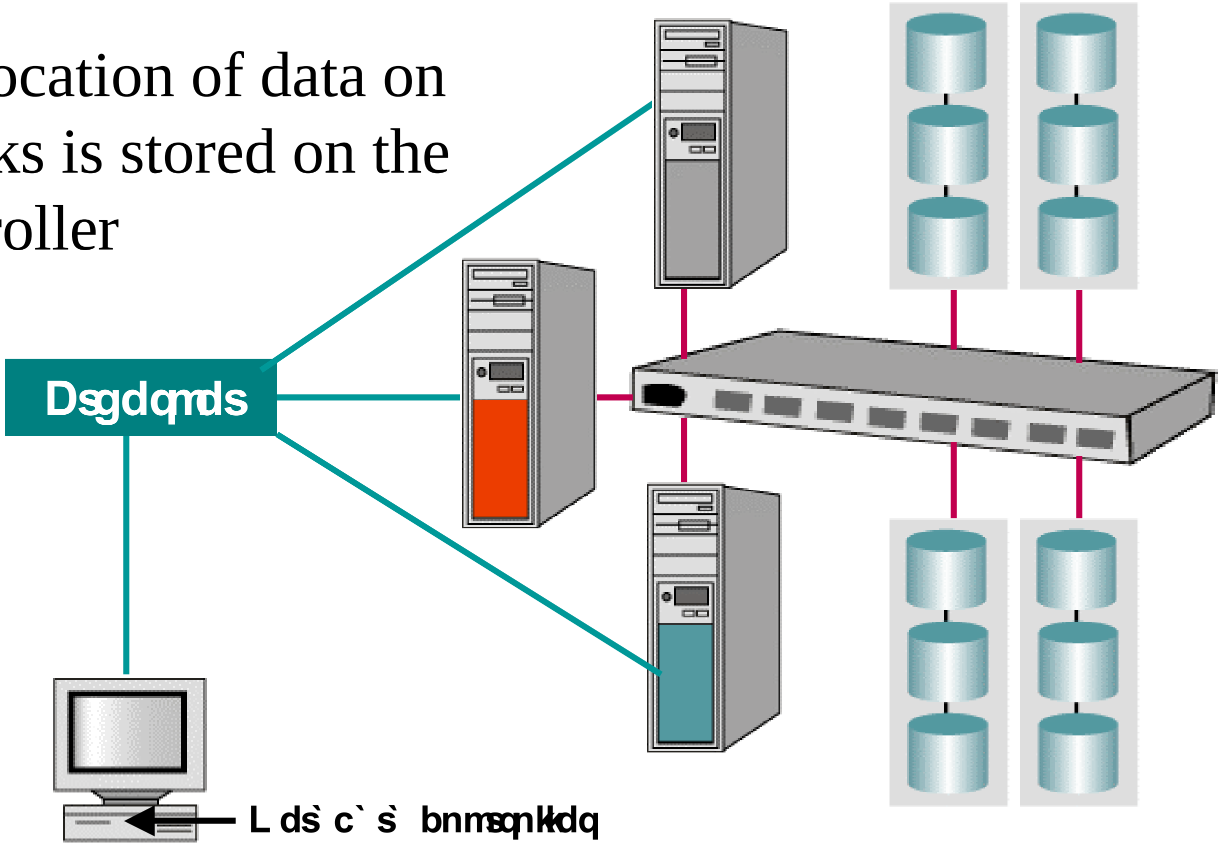
- Get faster computers
- Add more processors
- Add more RAM for caching
- Add faster host adapters
  - Take advantage of future technologies like Infiniband
- Add more computers

Scaling in Capacity

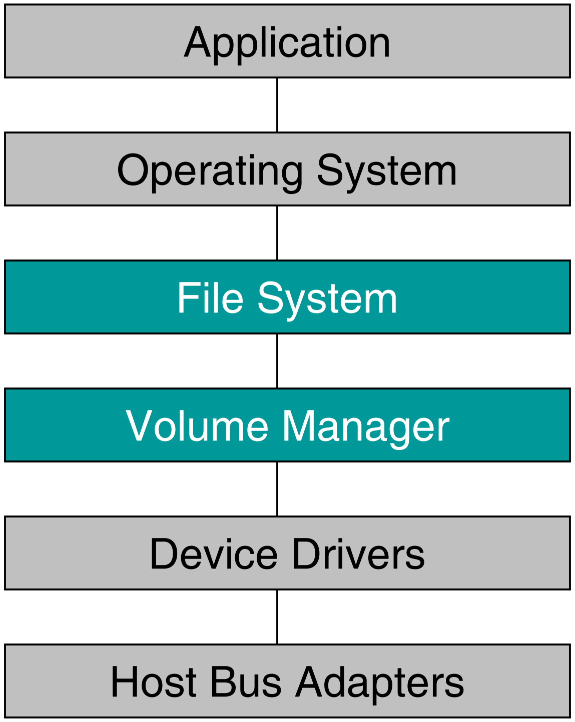
- Just add more disks of any type
- Use older model disks for scratch space or non-mission-critical operations
- Use JBODs instead of RAID or expensive disk sub-systems
  - 1 Terabyte of fault tolerant disk costs as little as $20,000

4) Out-of-Band (Asymmetrical) Disk Virtualization

The physical location of data on virtualized disks is stored on the metadata controller



Host Intervention



Out-of-Band Virtualization Requires Intervention on Host

They intercept I/O path and redirect in one of two places:

1) Volume Manager

2) File System

Redirection via Volume Manager

- Virtualized disk appears like local hard disks when, in fact, the data could be stored anywhere on the SAN.
- Disk I/O requests involve communication over Ethernet with the meta-data server, followed by direct reads and writes over fibre channel.

Volume Level Out-of-Band vs. In-Band Virtualization

Out-Of-Band Disk Virtualization (via Volume Manager)	**In-Band Disk Virtualization**
Requires special software at host	Transparent; no software required
Partition at host	Partition at disk controller
No in-band bottleneck	Caching and fast hardware make up for in-band bottleneck
Meta data controller is single point of failure	No single point of failure

5) Interception via File System (“SAN File System”)

- Host computer maps a network drive volume on the meta data server
- Meta data consists of:
  - File system information
  - Physical location of data on disks
- File system meta data read/written over Ethernet
- Actual content data sent/received over fibre channel

Benefits of SAN File Systems

- All the benefits of a network file system
  - File sharing between multiple hosts
  - Centralized rights and permissions
- Without the performance hit
  - Movement of data over high speed fibre channel link instead of Ethernet.
  - Less I/O processing
  - File system I/O processing on dedicated meta data server.
- Enables “serverless” backup without special software.

Shortcomings of Out-Of-Band Virtualization

- Partitioning occurs in software at the host computer
  - Not very secure; little protection against human error
  - New versions of driver and OS could require extensive quality assurance
- Involves either 3rd party volume manager or 3rd party file system.
  - Does not use native OS tools
- Meta data controller is a point of failure and possible I/O bottleneck

Future of SAN File Systems

- Distributed meta data instead of central meta data controller
  - All of the benefits of a high performance shared file system
  - Allows for more reliable SAN partitioning
- Development efforts underway
- Scalable, SAN-attached, NAS Devices
  - Possibly packaged as blades on an enterprise switch.

Today’s Recommendation

- In-Band Virtualization
  - Requires nothing special on host computers
  - Very secure (LUN masking occurs at disk controller)
  - I/O bottleneck alleviated by high-performance hardware and sophisticated caching
- SAN File Systems
  - For specific applications that need the bandwidth and that are not vulnerable to user error.
  - Enforce policies with in-band virtualization.

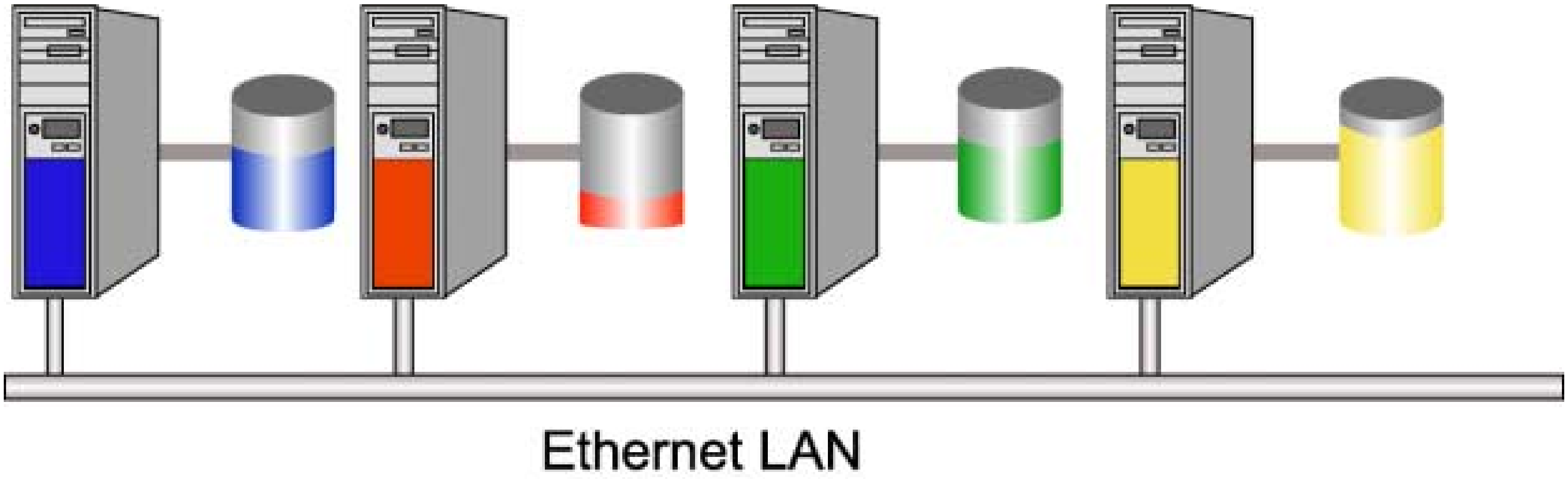
Chapter 4

Features and Benefits of Intelligent Disk Systems

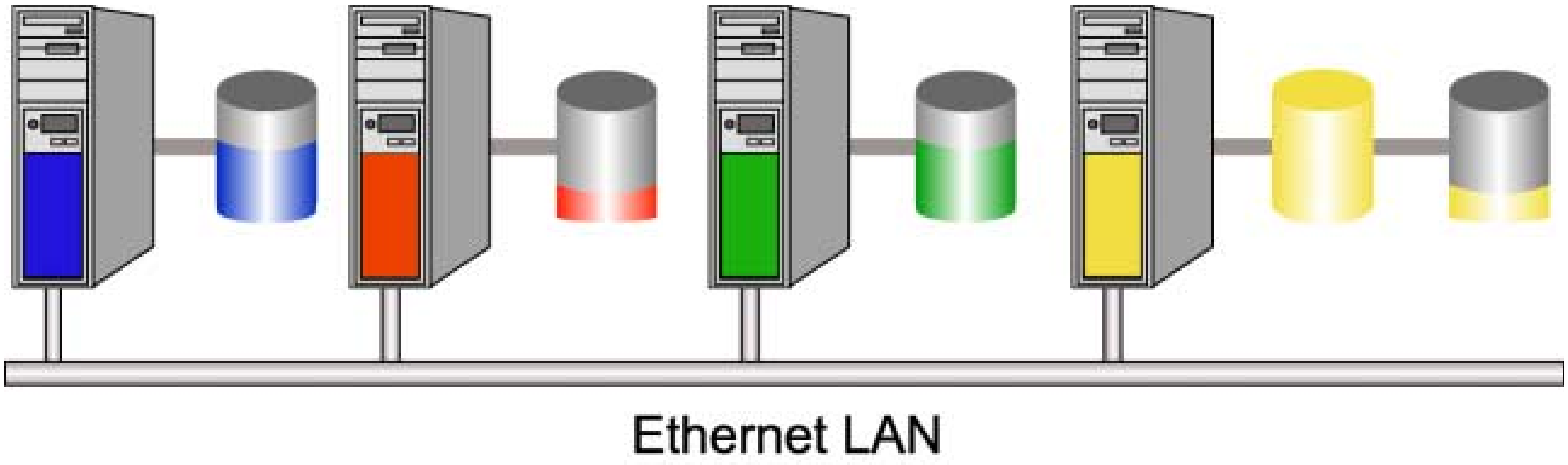
Storage on Demand

- Enabled by large array or disk virtualization system
  - Can be done with RAID controllers, but beware of maxing out I/O and cache with too many hosts.
- Allocate storage on demand
- De-allocate storage when no longer needed
- Great for testing and staging

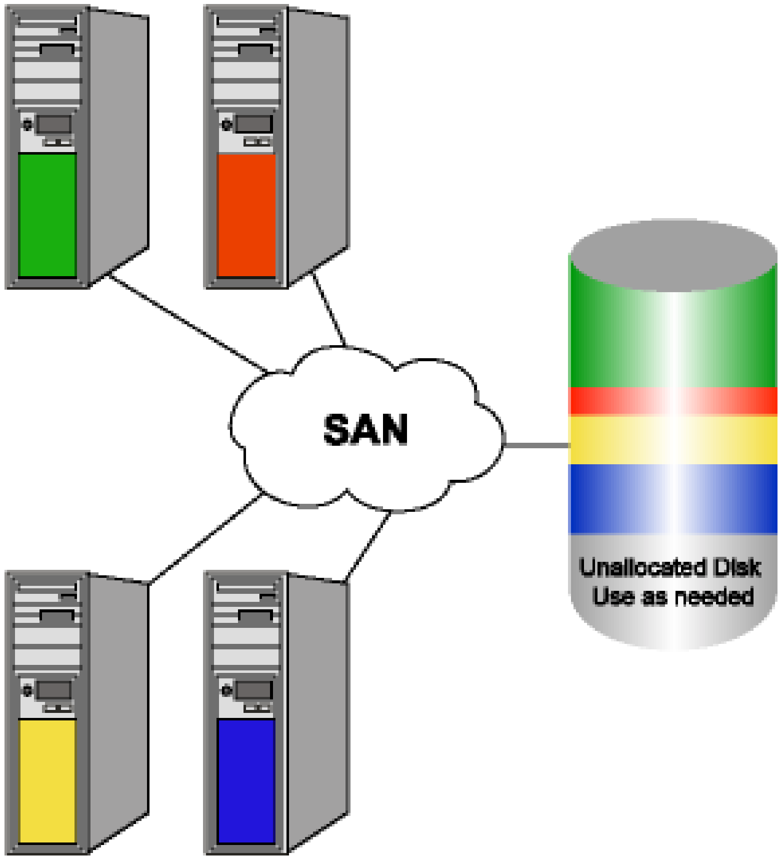
Problem: Distributed Storage



Adding Storage the Old-Fashioned Way



Allocating Storage on Demand



Advantages

- Creates a shared storage “pool”
  - Dynamically allocate storage to servers as needed
  - To add storage, simply add to the “pool”
- Economizes on physical space
- Centralizes disk management
- Assures balanced storage utilization

Snap Shots and Broken Mirrors

- Snap Shots = instant virtual copy of disk volume
  - Common feature to volume management software, but can be done in hardware.
  - Copy on write with logical block mappings
- Broken Mirrors - scheduled break of mirrored volumes.
- Benefits
  - Testing and Staging
  - Midday copies
  - Backup bandwidth optimization

Remote Hardware Mirroring

- Originally unique to EMC
  - Now available from lower cost vendors
- Synchronous - real time copy done over fibre channel, SCSI, or channel extender.
- Asynchronous - works over IP and with slower connections.

Disk Virtualizers v. RAID Controllers

- RAID controllers are relatively simple devices modified to work on fibre channel SANs.
- Virtualizers use disk more intelligently.
  - Better performance, especially with multiple hosts
  - Smarter caching
  - Sometimes no caching required at all
  - Features vary tremendously
    - Snap shots, mirroring, remote mirroring
    - Performance tuning
    - Expansion capabilities

Chapter 5

SAN vs. NAS

What is NAS?

- NAS devices are actually file server appliances.
- Just like any server where files are stored, except optimized just for storage:
  - Easy to configure for storage.
  - Nothing else to configure. Just does storage.
  - Might have some cool bells and whistles.
    - Performance enhancements
    - Volume management tools

Apples v. Oranges

- SAN is a way to plug things in.
- NAS is a thing you plug in.

- SAN is a way to network your SCSI storage devices.
- NAS is a file server that has to store its data on disk somewhere.

- What common problems do they claim to solve?

What Do SAN and NAS Have in Common?

- Both types of vendors want to sell you hard disks at a premium:
  - SAN Vendors - Package SANs to be large, centralized disk systems with tons of proprietary hardware and software.
  - NAS Vendors - Claim to be attaching storage to your network, when really they are selling you file servers. They too want to sell you a large, centralized disk system.
  - Who do you buy your large, centralized disk system from?

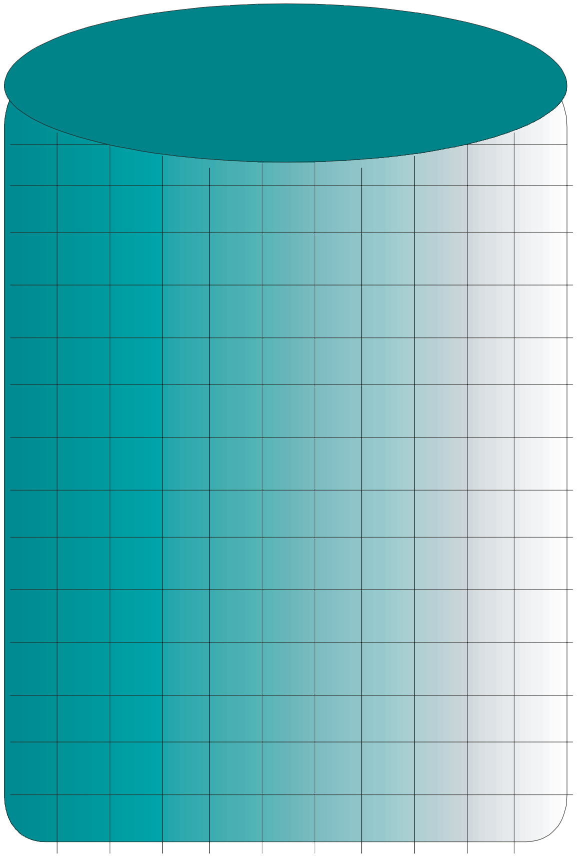
NAS v. SAN

- NAS
  - Add a new device, when you exceed the capacity of your NAS server.
  - Device is easy to configure, so this is not a problem.
- SAN
  - Allocate more disk storage to existing devices when you need more capacity.
  - Disk allocation is easy and on demand, so this is not a problem.

Logical Blocks

The logical block is the smallest unit of measurement for data storage. Storage devices (disks, tape, etc.) are represented as a bunch of logical blocks.

Logical blocks are organized into files by the file system.



SAN Applications v. NAS Applications

- **SAN Applications** – Delivery and virtualization of logical blocks.

- **NAS Applications** – Delivery and virtualization of files.

SAN v. NAS = Files v. Blocks

- Blocks
  - Good: High performance, easy to manipulate
  - Bad: No data sharing. Device sharing only.
- Files
  - Good: Reliable Data Sharing, Nice way to package data.
  - Bad: Network file systems (NFS & CIFS) are slow
    - Network is a bottleneck.
    - I/O processing is a bottleneck.
    - No matter how fast the NAS server, there is a still a bottleneck on the machine that accesses it.

SAN v. NAS = External v. Internal Block Virtualization

- Many leading NAS devices have sophisticated volume management technology built in. How does this compare to similar SAN solutions?

- **NAS** - Internal snapshots means tighter integration with software. Nice controlled environment.
- **SAN** - External snapshots means another machine can have direct access to the snapshot without an I/O bottleneck.
  - Maybe economies of scale purchasing this technology centrally.
  - Block virtualization on platforms that do not offer sophisticated volume management.

Shortcomings of NAS

- Backup is still difficult, but improving
- Multiple devices = multiple mount points
- Really only addresses file services
  - CIFS and NFS are not suitable for high speed application and database storage.
  - No matter how fast the file server, a bottleneck will exist on the client computers.

- BUT - New network file system technology is on the horizon, promising low overhead and low latencies.

SAN & NAS Convergence

- Some NAS devices can access their raw storage needs from a SAN.
  - Why not expand NAS devices by allocating disk over a SAN.
  - Why not backup NAS devices with LAN Free or Serverless backup?
  - Do not confuse this with NAS devices that use fibre channel, but they are a good start.
- SAN file systems are beginning to be offered simple, turnkey appliances.

SAN <> Fibre Channel

- SAN v. NAS is not about fibre channel v. Ethernet!!!
- SANs can be built with a variety of data transports:
  - Fibre Channel
  - ESCON
  - Ethernet (more on this later)
  - Infiniband (in the future)

Conclusions and

Questions & Answers