Distributed Storage Systems

Storage Virtualization

Slides adapted from Jie Feng, Edward Whalen, Curt White, Albert Lee Kwok Chiu, Sanjay P. Ahuja

Agenda

Today's topics

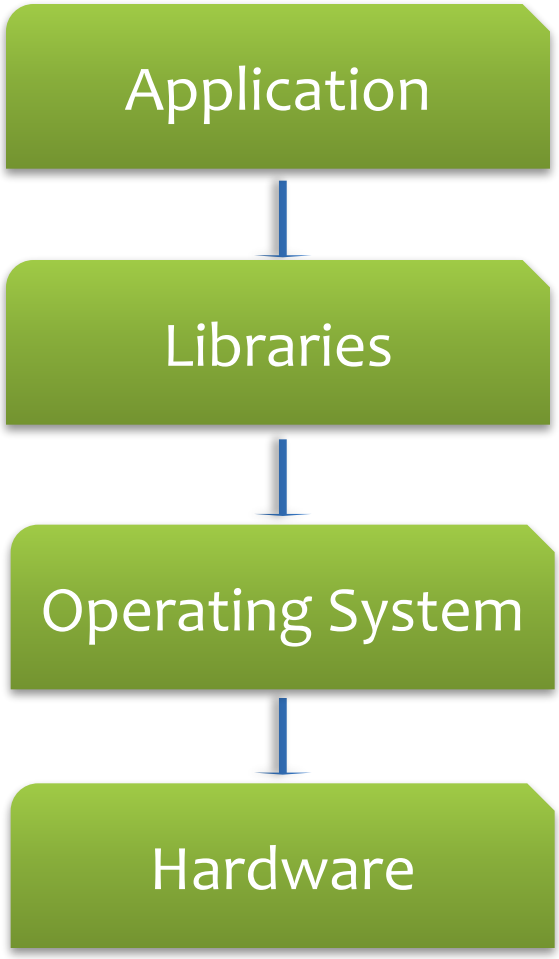
- Storage Virtualization
- Storage Area networks
- Network Attached Storage

Class Structure

	Lecture	Lab
Week 1	Course introduction, networking basics, socket programming	Python sockets
Week 2	RPC, NFS, Practical RPC	Flask, JsonRPC, REST API
Week 3	AFS, reliable storage introduction	ZeroMQ, ProtoBuf
Week 4	Hard drives, RAID levels	RPi stack intro, RPi RAID with ZMQ
Week 5	Finite fields, Reed-Solomon Codes	Kodo intro, RS and RLNC with Kodo
Week 6	Repair problem, RS vs Regenerating codes	RPi simple distributed storage with Kodo RS
Week 7	Regenerating codes, XORBAS	RPi Regenerate lost fragments with RS
Week 8	Hadoop	RPi RLNC, recovery with recode
Week 9	**Storage Virtualization, Network Attached Storage, Storage Area Networks**	**RPi basic HDFS (namenode+datanode, read & write pipeline)**
Week 10	Object Storage	RPi basic S3 API
Week 11	Compression, Delta Encoding	Mini project consultation
Week 12	Data Deduplication	RPi Dedup
Week 13	Fog storage	Mini project consultation
Week 14	Security for Storage Systems and Recap	Mini project consultation

Virtualization

Machine Stack showing virtualization opportunities



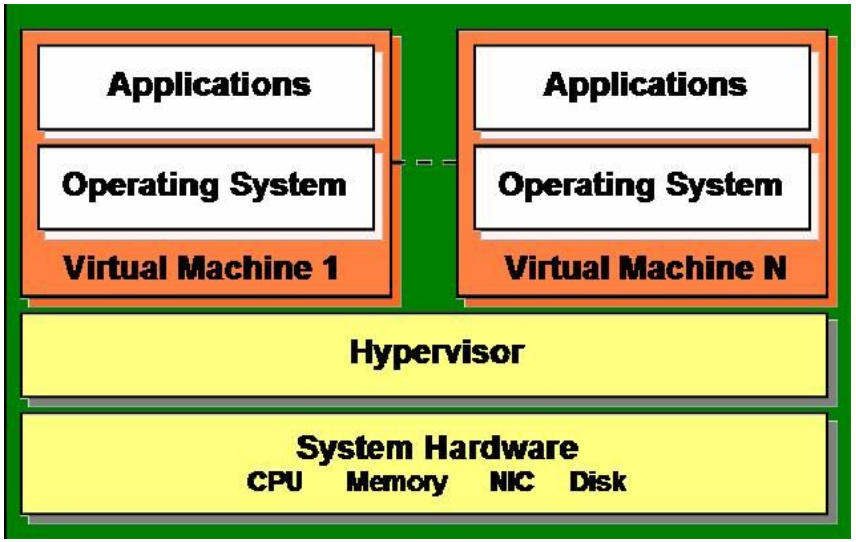
- Creation of a virtual version of hardware using software
- Runs several applications at the same time on a single physical server by hosting each of them inside their own virtual machine
- By running multiple virtual machines simultaneously, a physical server can be utilized efficiently

Primary approaches to virtualization

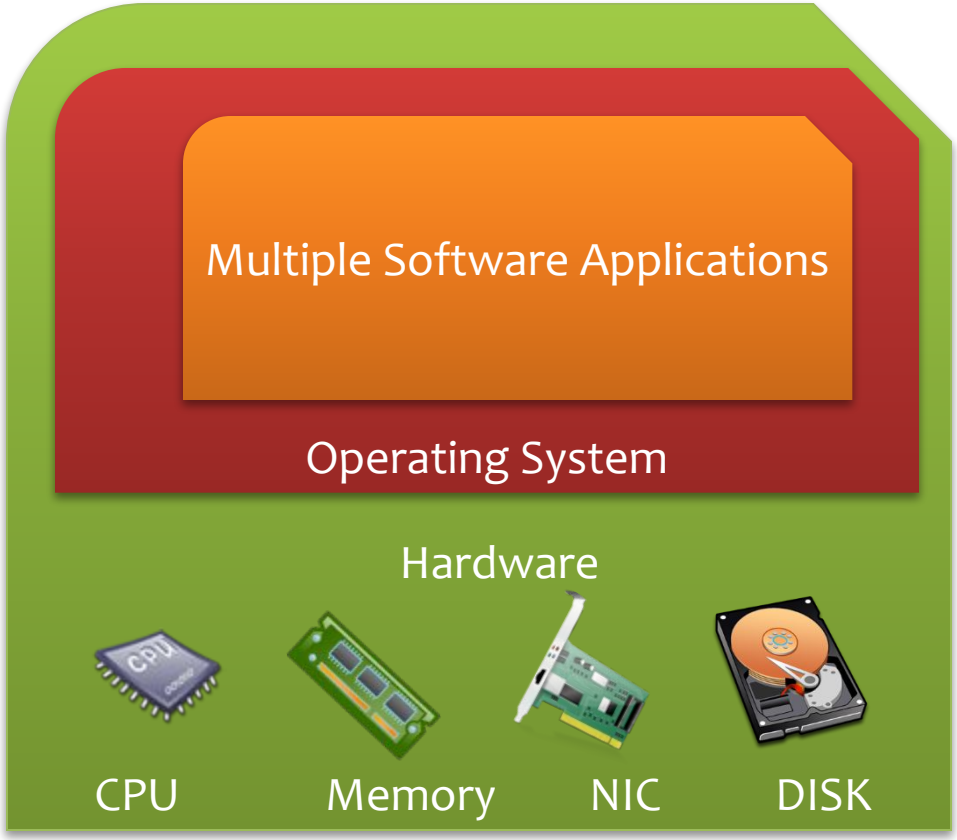
- Platform virtualization Ex : Server
- Resources virtualization Ex : **Storage**, Network

Hypervisor

- Provides support for running multiple operating systems concurrently in virtual servers created within a physical server
- Software responsible for hosting and managing all VMs
- Runs directly on the hardware
- Example: VMWare, Xen, KVM

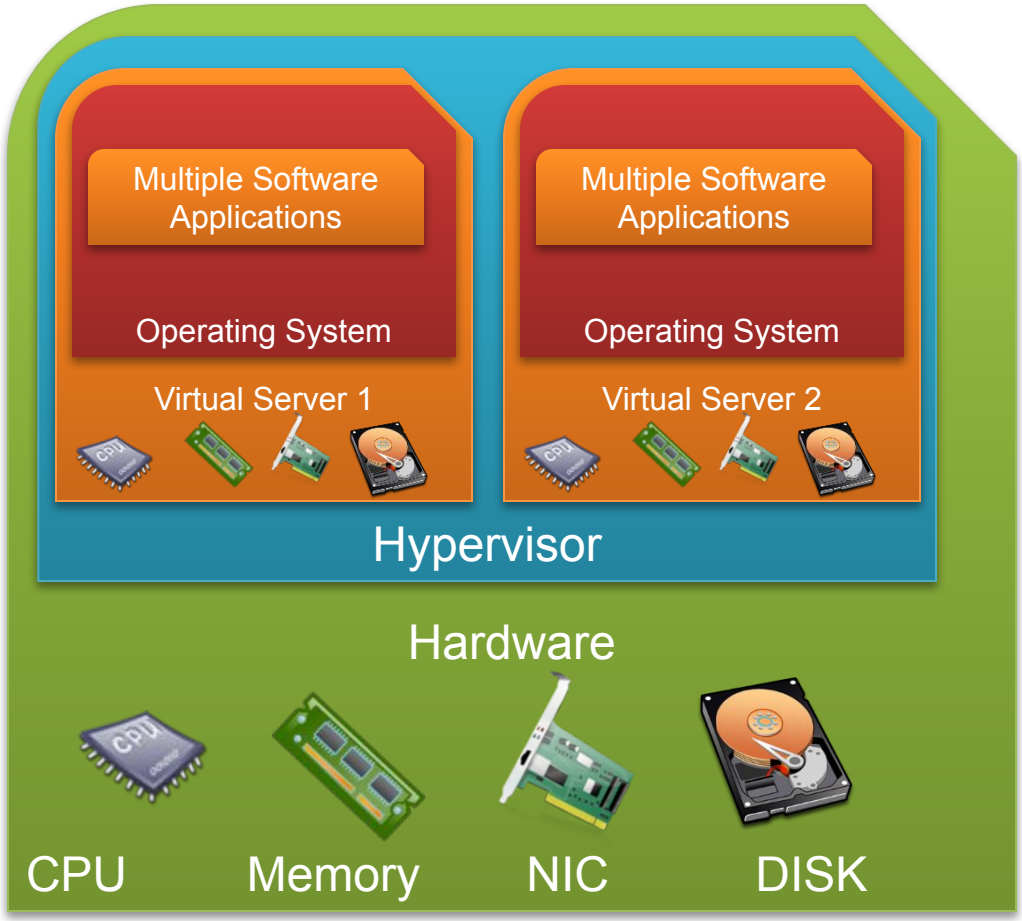


Server without Virtualization



- Only one OS can run at a time within a server
- Under utilization of resources
- Inflexible and costly infrastructure
- Hardware changes require manual effort and access to the physical server

Server with Virtualization



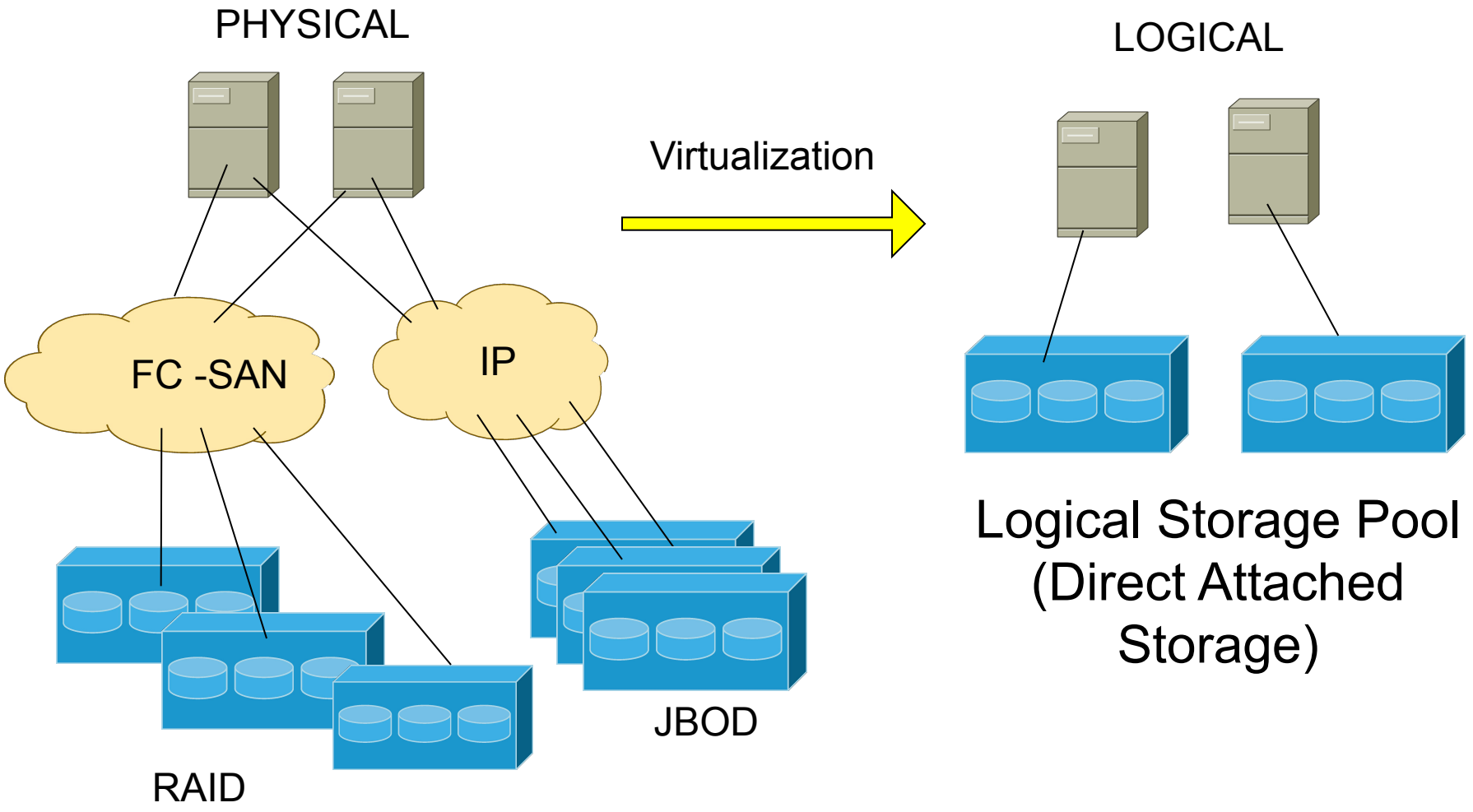
- Can run multiple OS simultaneously
- Each OS can have different hardware configuration
- Efficient utilization of hardware resources
- Each virtual machine is independent
- Save electricity, initial cost to buy servers, space
- Easy to manage and monitor virtual machines centrally

Storage Virtualization

Problem we are facing

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

Storage Virtualization



JBOD (abbreviated from "**Just a Bunch Of Disks**"/"**Just a Bunch Of Drives**") is an architecture using multiple hard drives exposed as individual devices.

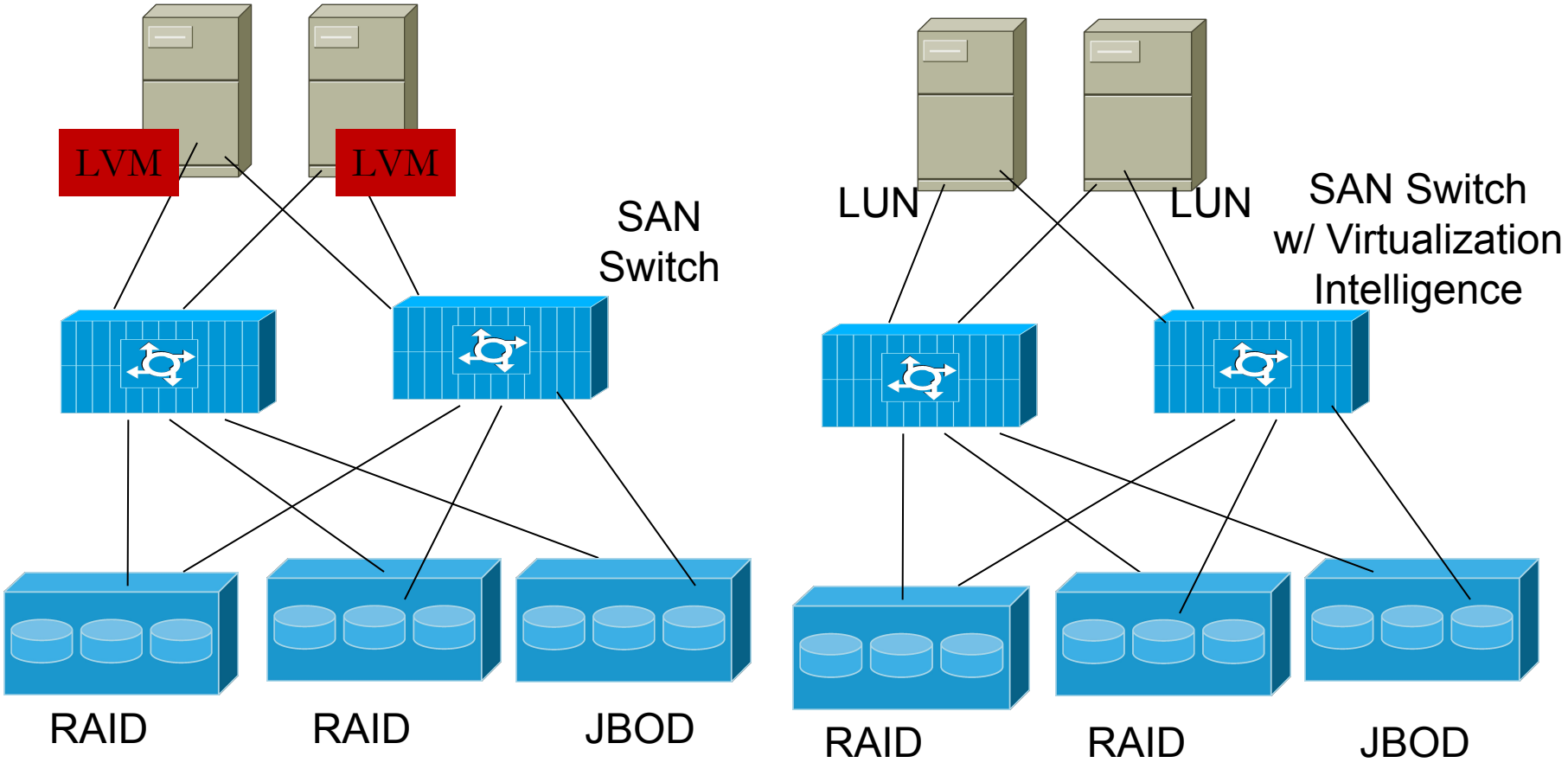
Storage Virtualization

- Hides physical storage from applications on host systems
- Presents a simplified (logical) view of storage resources to the applications
- Allows the application to reference the storage resource by its *common name*
- Actual storage could be on complex, multilayered, multipath storage networks
- RAID is an early example of storage virtualization

Virtualization Intelligence

- **Host-Based**: storage virtualization could be implemented on the host through Logical Volume Management (LVM) which provides the logical view of the storage to the host operating system.
- **Switch-based**: intelligence of storage virtualization could be implemented on the SAN switches. Each server is assigned a Logical Unit Number (LUN) to access the storage resources
  - Pros:
    - Ease of configuration and management
    - Redundancy/high availability
  - Cons:
    - Potential bottleneck on the switch
    - Higher cost

Storage Virtualization



LVM : Logical Volume Manager
LUN : Logical Unit Number
SAN : Storage Area Network

Three Basic Forms of Network Storage

- Direct access storage (DAS)
- Network attached storage (NAS)
- Storage area network (SAN)

- And a number of variations on each (especially the last two)

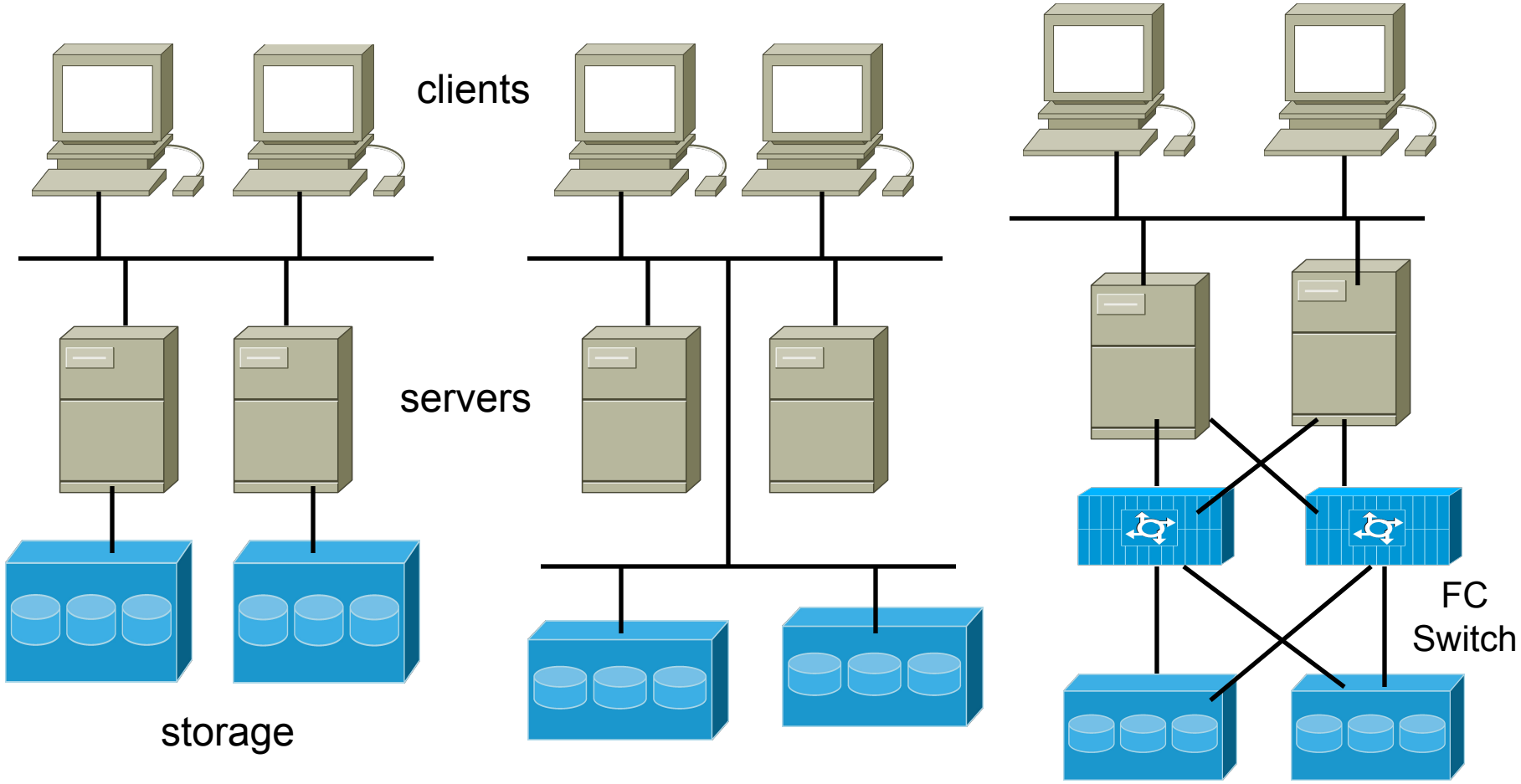
Three Basic Forms of Network Storage

	DAS	NAS	SAN
Storage Type	sectors	shared files	blocks
Data Transmission	IDE/SCSI	TCP/IP, Ethernet	Fibre Channel
Access Mode	clients or servers	clients or servers	servers
Capacity (bytes)	10^9	10^9 - 10^{12}	> 10^{12}
Complexity	Easy	Moderate	Difficult
Management Cost (per GB)	High	Moderate	Low

DAS

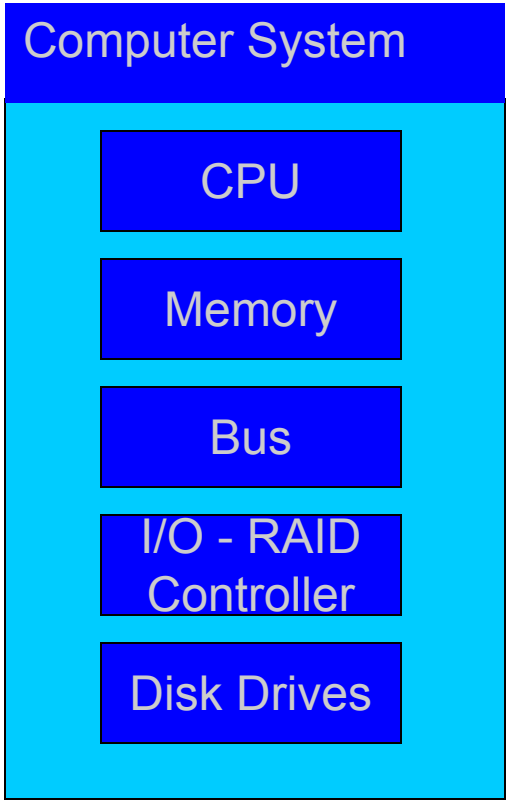
NAS

FC-SAN

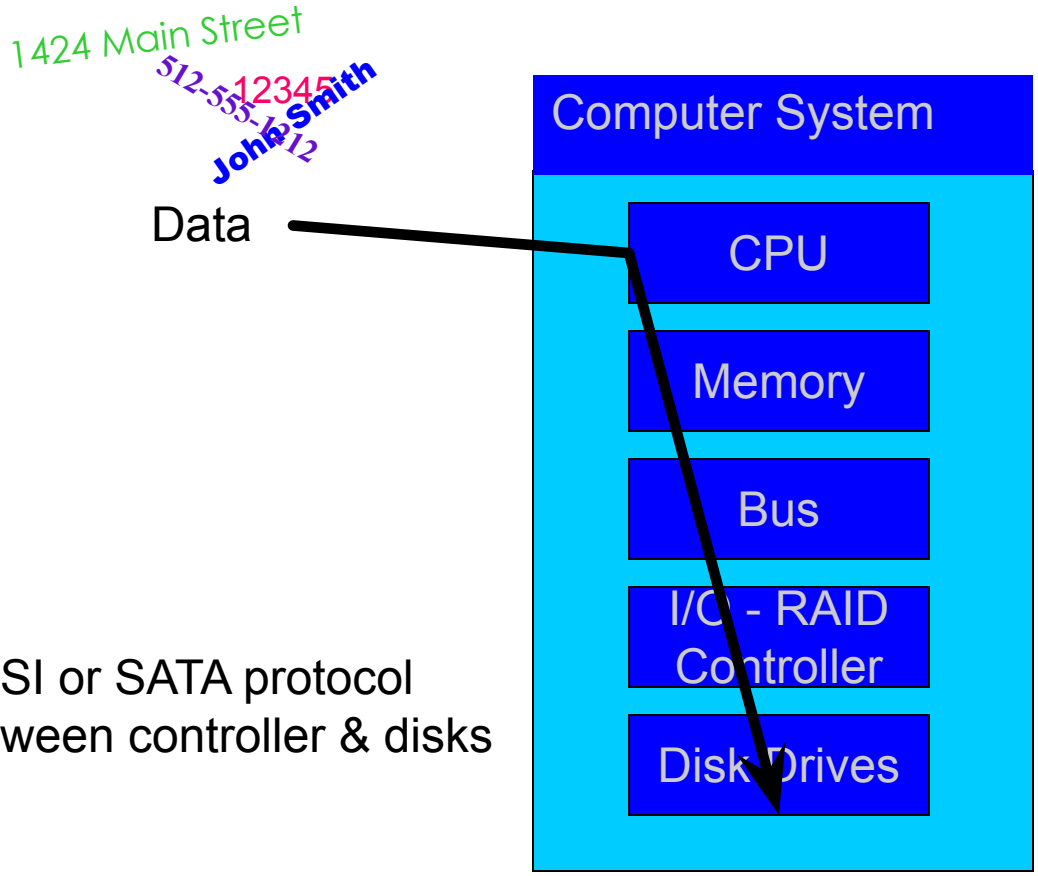


Direct-Attached Storage

DAS: Direct Attached Storage (One Computer)



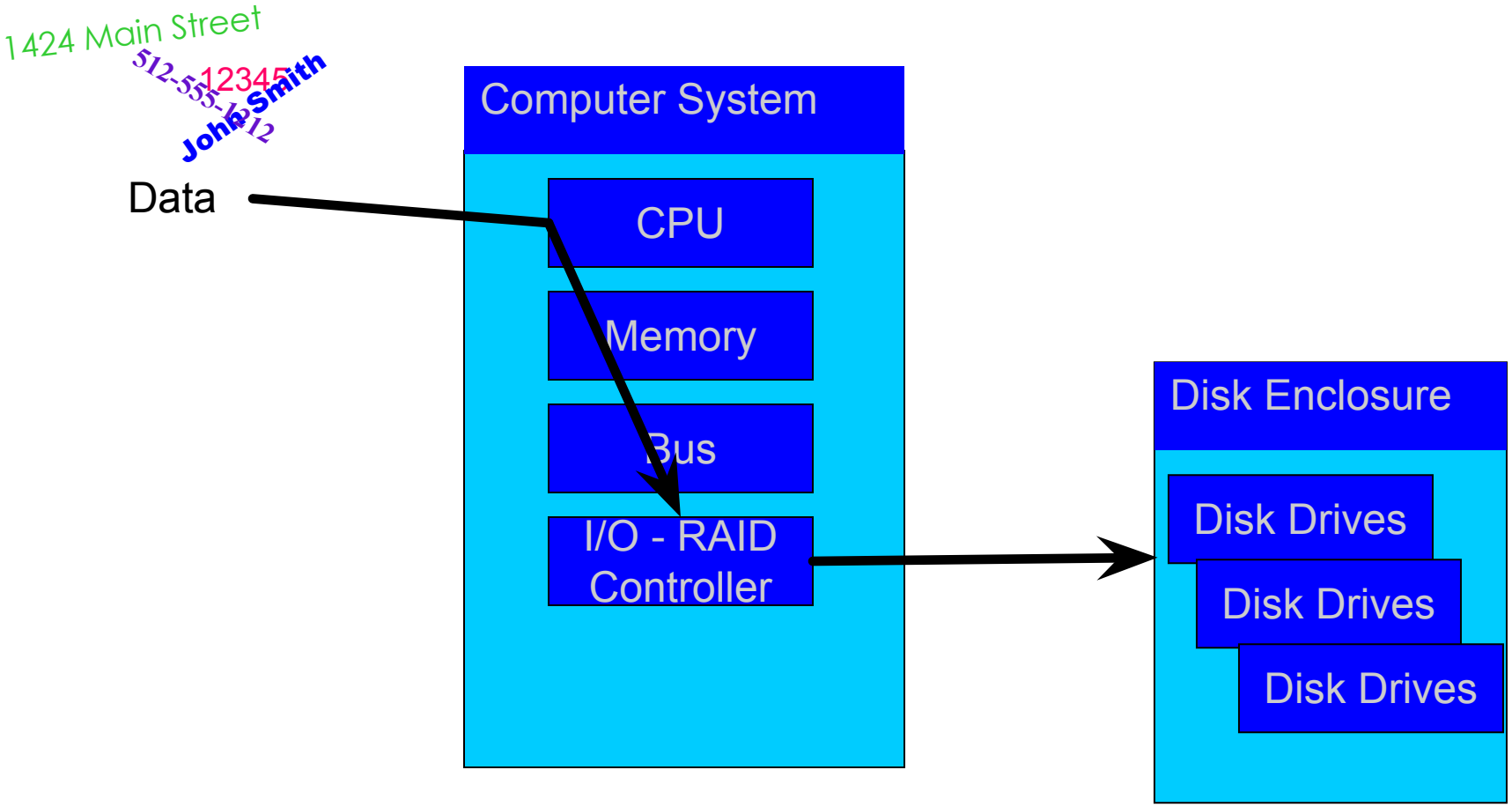
DAS: Direct Attached Storage (One Computer)



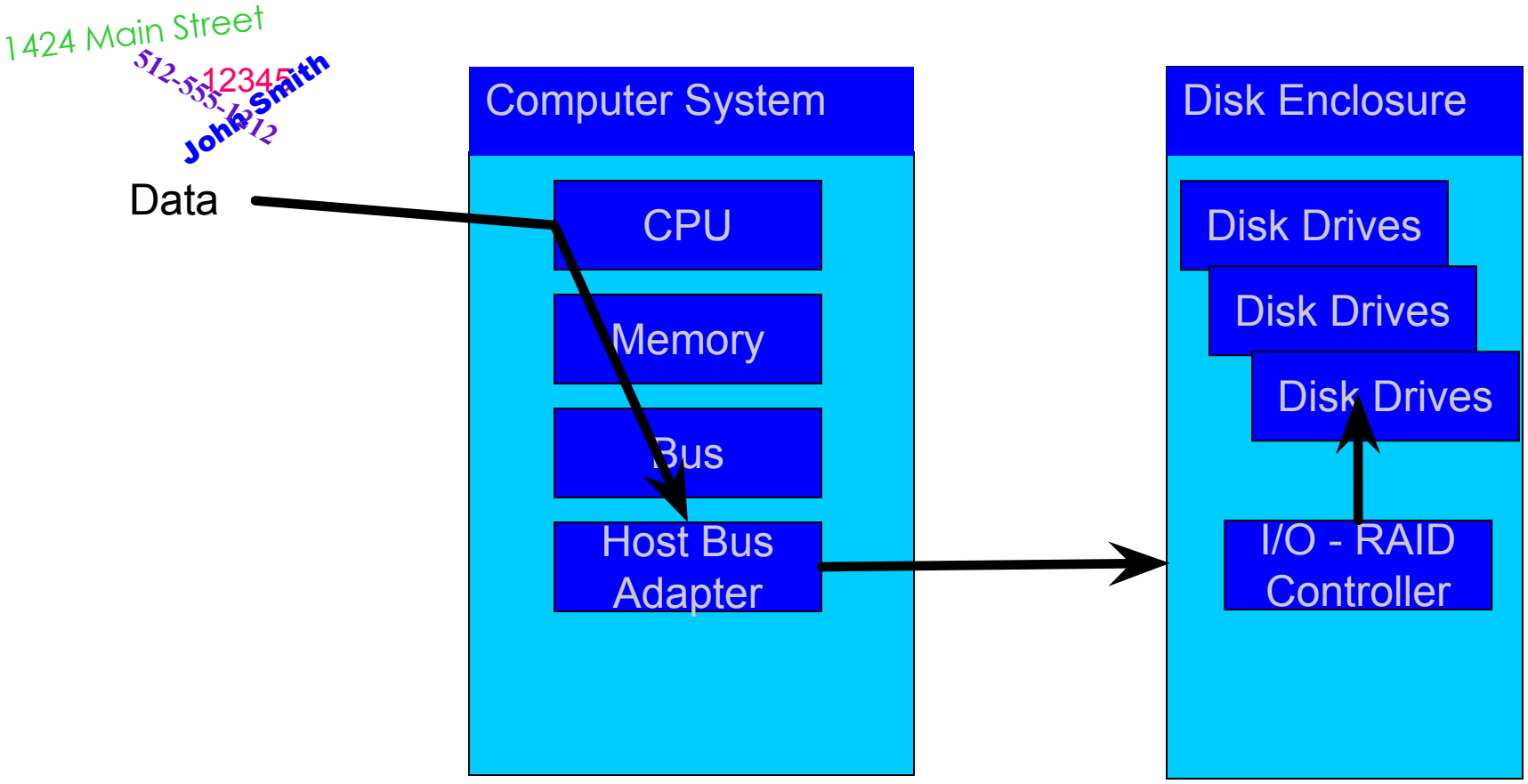
SCSI or SATA protocol between controller & disks

SCSI: Small Computer System Interface

DAS with Internal Controller & External Storage



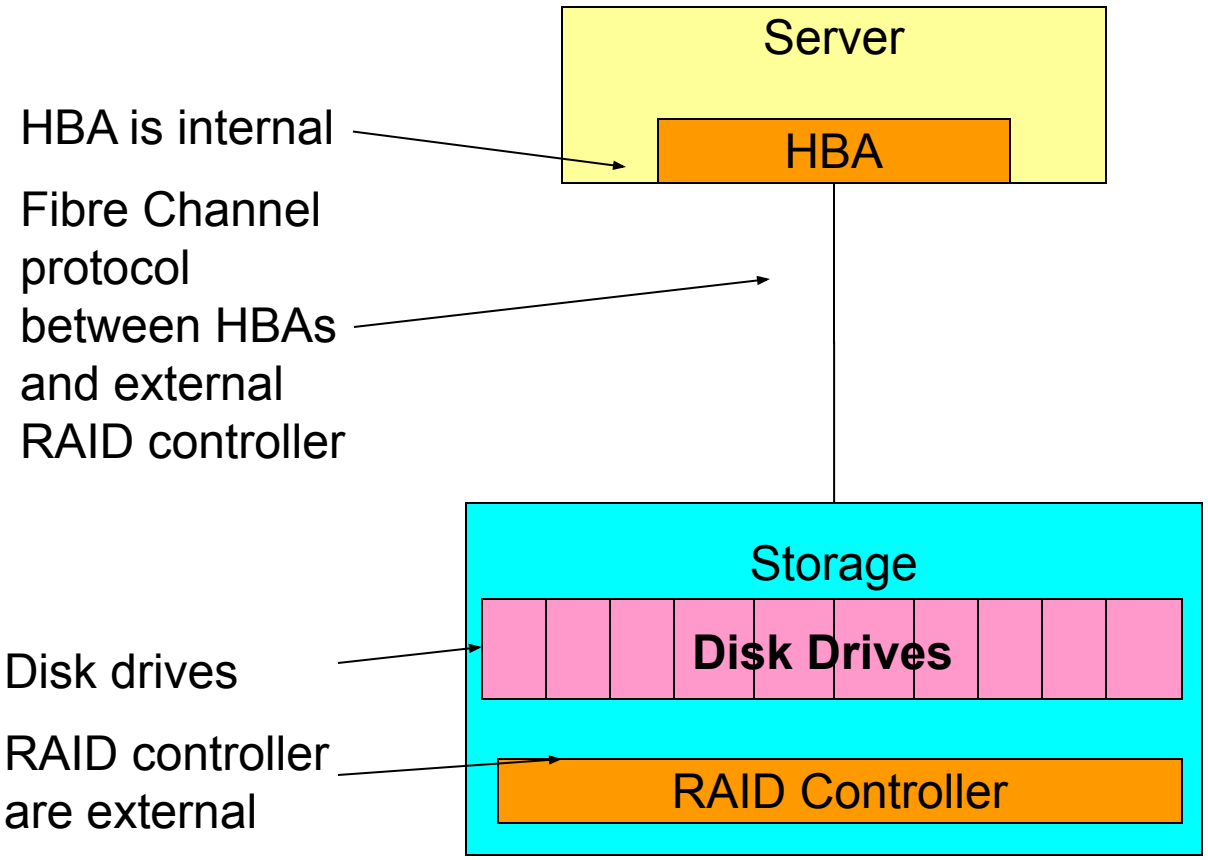
DAS with External Controller & External Storage



I/O Transfer

- RAID Controller
  - Contains the “smarts”
  - Determines how the data will be written (striping, mirroring, RAID 10, RAID 5, ...)
- Host Bus Adapter (HBA)
  - Simply transfers the data to a RAID controller
  - Doesn’t do any RAID or striping calculations
  - “Dumb” for speed
  - Required for external storage

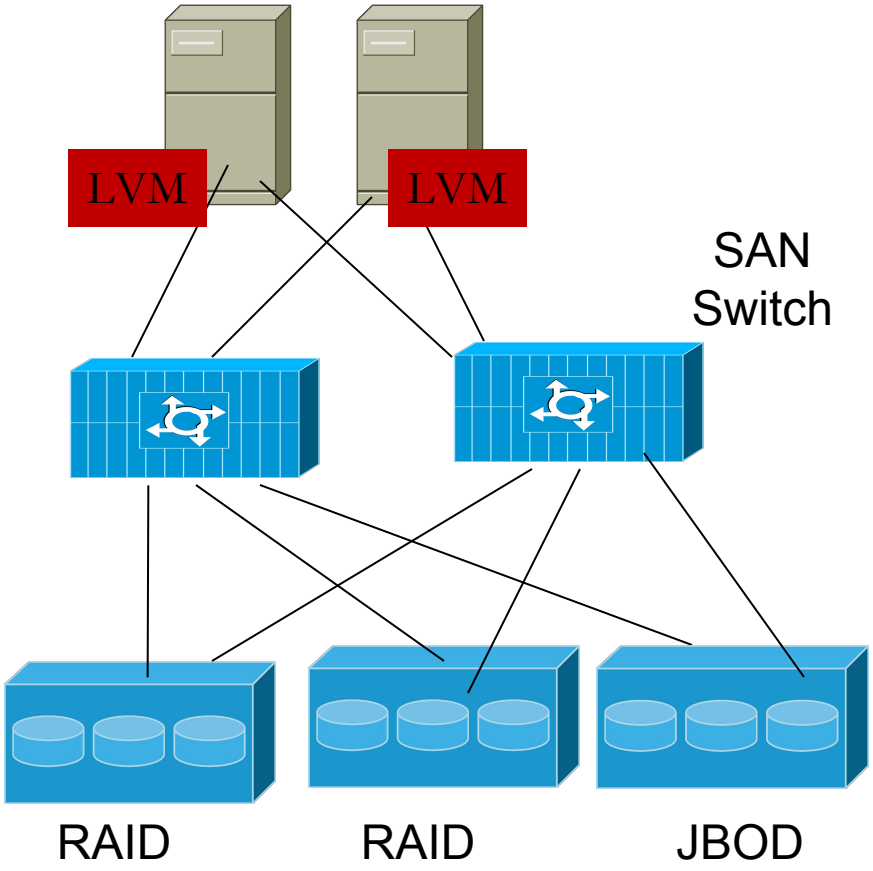
DAS over Fibre Channel

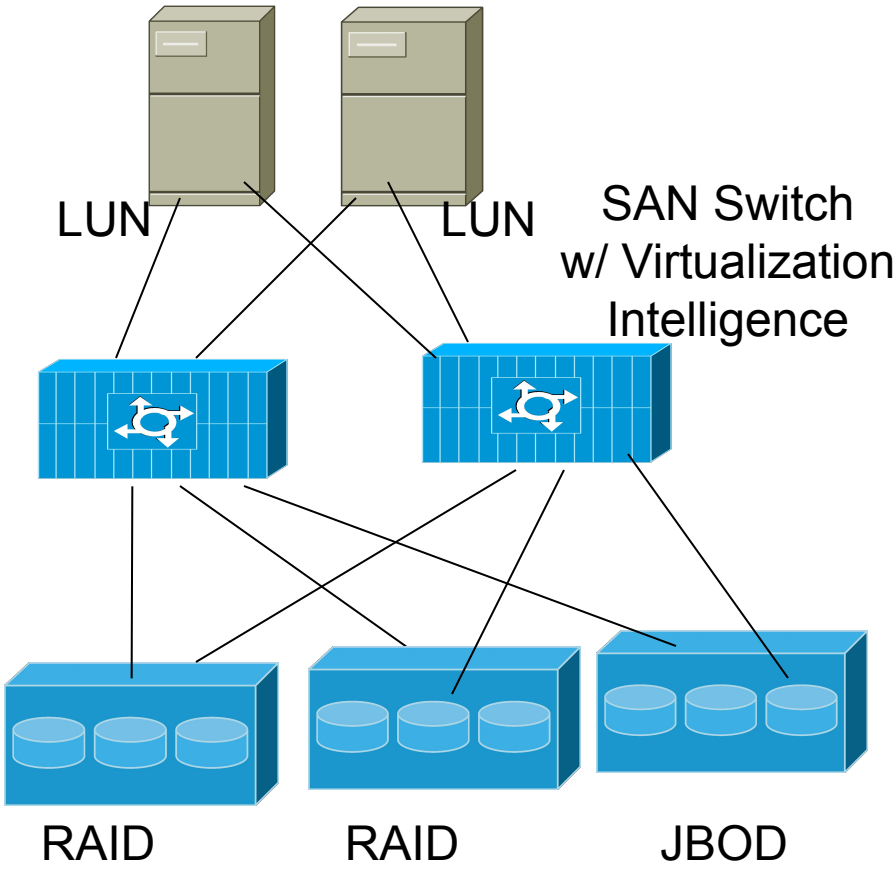


External SAN Array

Storage Area Network

Storage Virtualization





LVM : Logical Volume Manager
LUN : Logical Unit Number
SAN : Storage Area Network

Storage Area Network (SAN)

- A Storage Area Network (SAN) is a specialized, dedicated high speed network joining servers and storage, including disks, disk arrays, tapes
- Storage (data store) is separated from the processors (and separated processing)
- High capacity, high availability, high scalability, ease of configuration, ease of reconfiguration

- **Fibre Channel** is the de facto SAN networking architecture, although other network standards could be used

Fibre Channel (FC)

- Fiber Channel is well established in the open systems environment as the underlying architecture of the SAN
- FC has with 5 independent layers
  - The physical layers are 0 to 2.
  - These lower layers carry physical attributes of the network and transport the data created by the higher level protocols, such as SCSI, TCP/IP, or FICON
- Fibre Channel utilizes specialized
  - Switches
  - Host Bus Adapters
  - RAID controllers
  - Cables

Fibre Channel (FC)

- How does it work?
  - Serial interface
  - Data is transferred across a single piece of medium at the fastest speed supported
  - No complex signaling required
  - SCSI on top of Fibre Channel (No re-inventing the wheel)
  - Immediate OS support
- What's with the funny name?
  - History: developed to only support fiber optic cabling
  - When copper cabling support was added, ISO decided not to rename the technology
  - ISO changed to the French spelling to reduce association with fiber optics only medium

Fibre Channel (FC)

- Hot-pluggable - Devices can be removed or added at will with no ill effects to data communications
- Provides a data link layer above the physical interconnect, analogous to Ethernet
- Sophisticated error detection at the frame level
- Data is checked and resent if necessary
- Up to 127 devices (vs SCSI: 15)
- Up to 10 km of cabling (vs 3-15 ft. for SCSI)

- Combines the characteristics of
  - networks (large address space, scalability) and
  - I/O channels (high speed, low latency, hardware error detection)

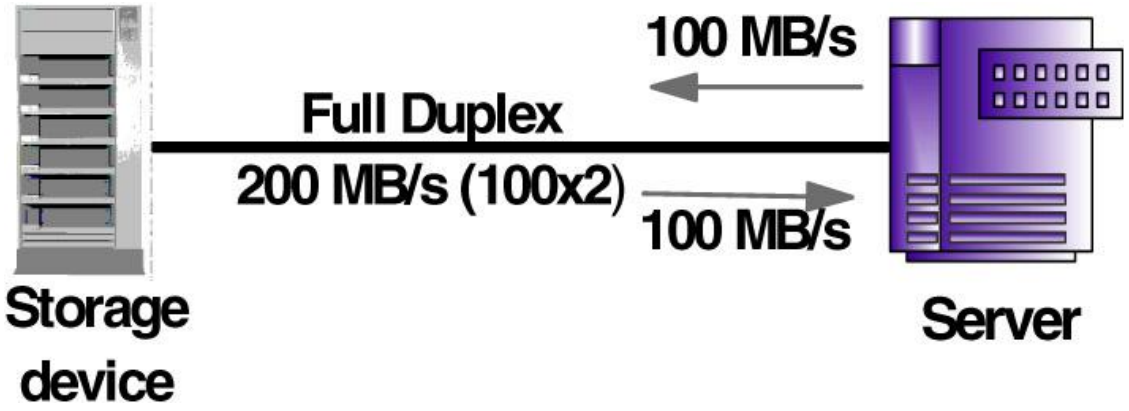
SAN Benefits

- **Scalability:**
  - Fibre Channel networks allow # of nodes to increase without loss of performance
  - More switches added → switching capacity grows
- **High Performance**
  - Fibre Channel fabrics provide a switched 100 MB/s full duplex interconnect
- **Storage Management**
  - SAN-attached storage allows uniform management of the entire investment in storage
- **Decoupling Servers and Storage**
  - Servers can be upgraded while leaving storage in place
  - Storage can be added at will and dynamically allocated to servers without downtime

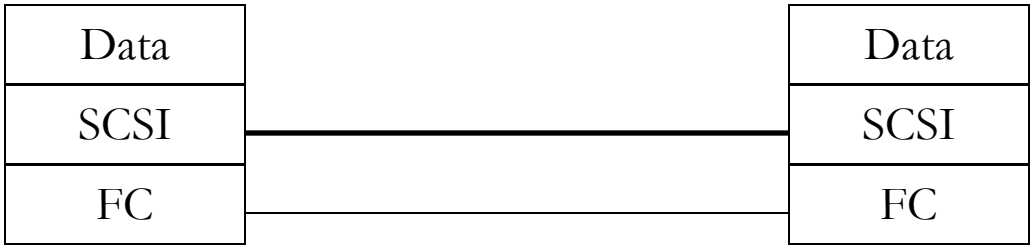
SAN Topologies

- Fibre Channel based networks support three types of topologies:
  - Point-to-point
  - Loop (arbitrated) – shared media
  - Switched

FC - Point-to-Point



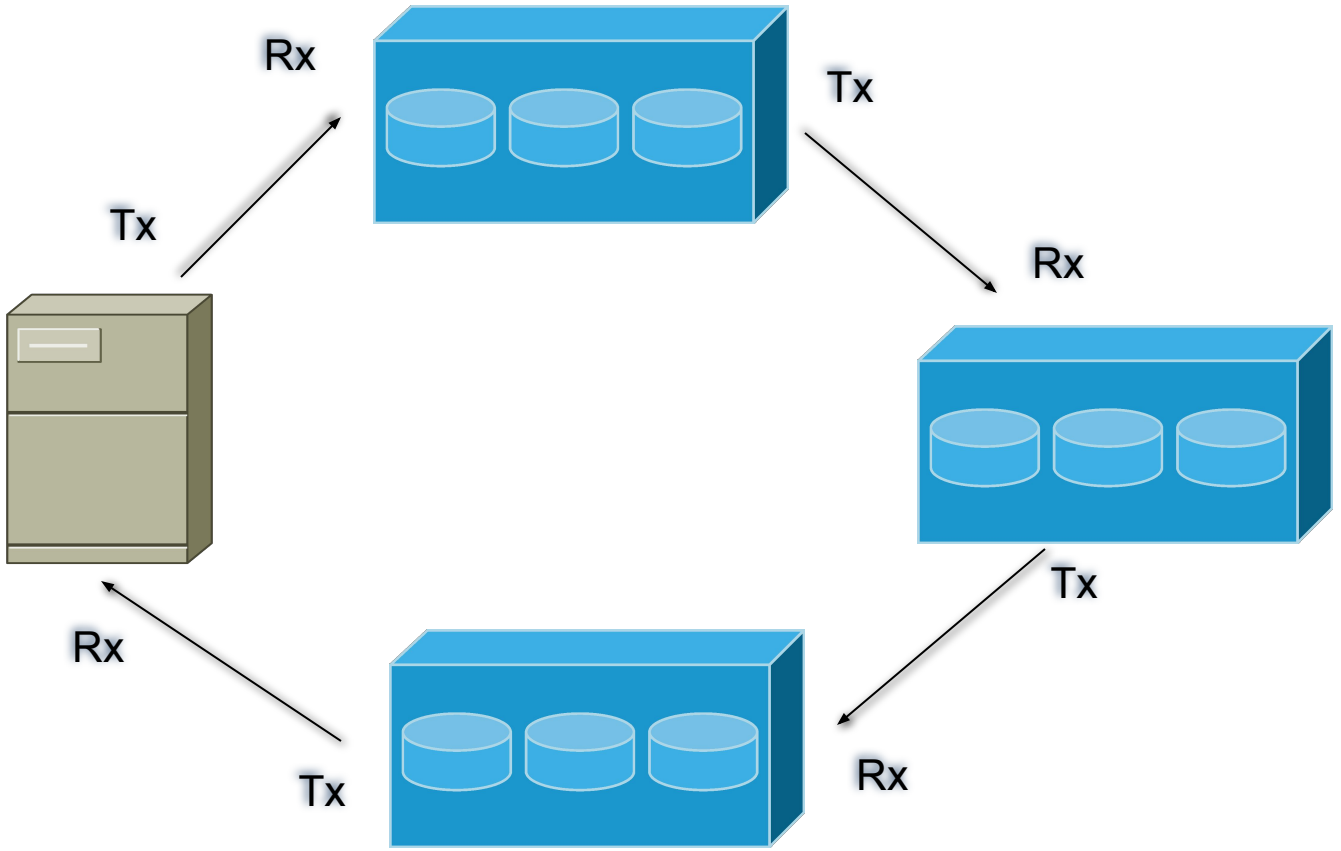
- The point-to-point topology is the easiest Fibre Channel configuration to implement, and it is also the easiest to administer.
- The distance between nodes can be up to 10 km



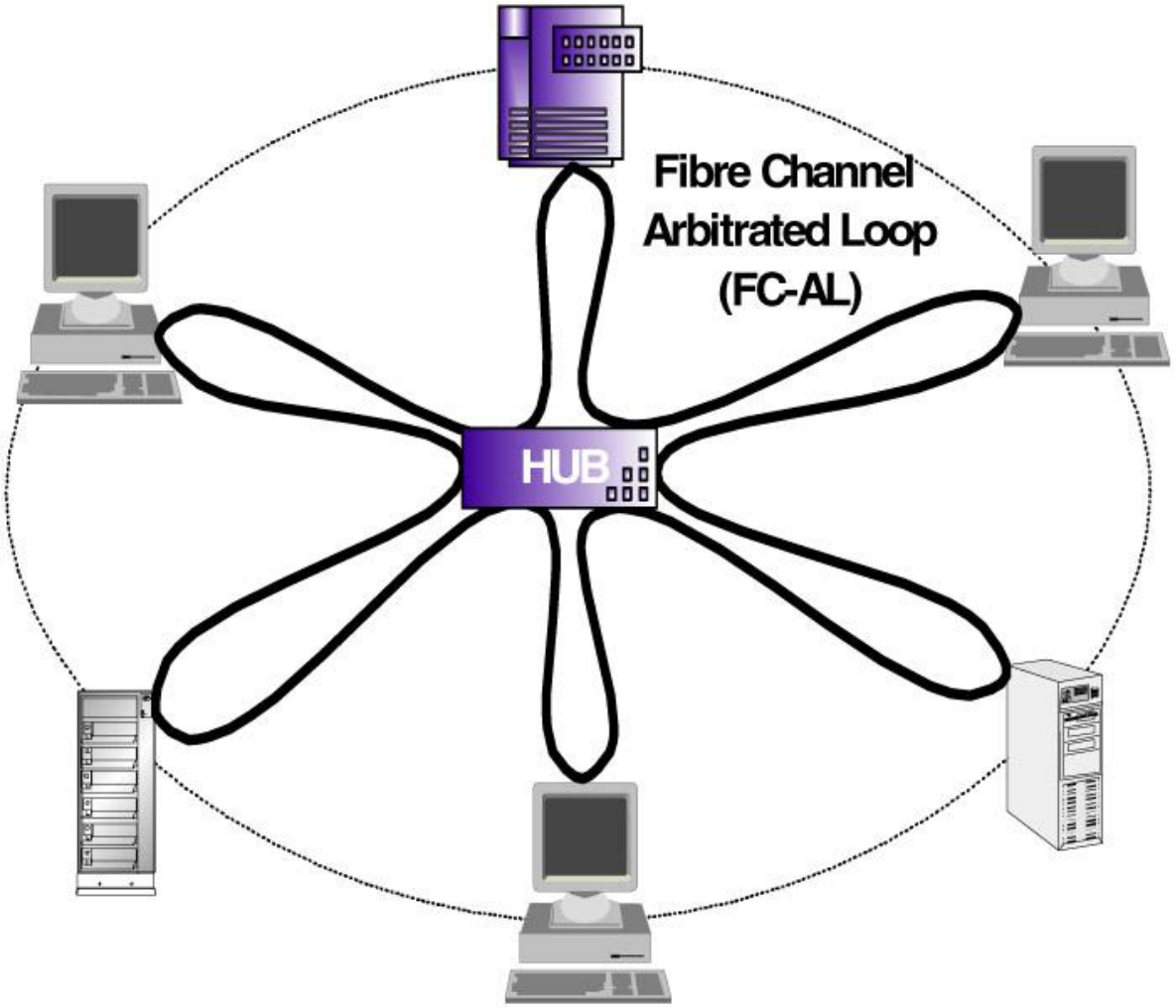
Arbitrated Loop

- Shared Media Transport
  - Similar in concept to shared Ethernet
- Not common for FC-based SAN
- Commonly used for JBOD (Just a Bunch of Disks)
- An arbitration protocol determines who can access the media

Arbitrated Loop (Daisy Chain)



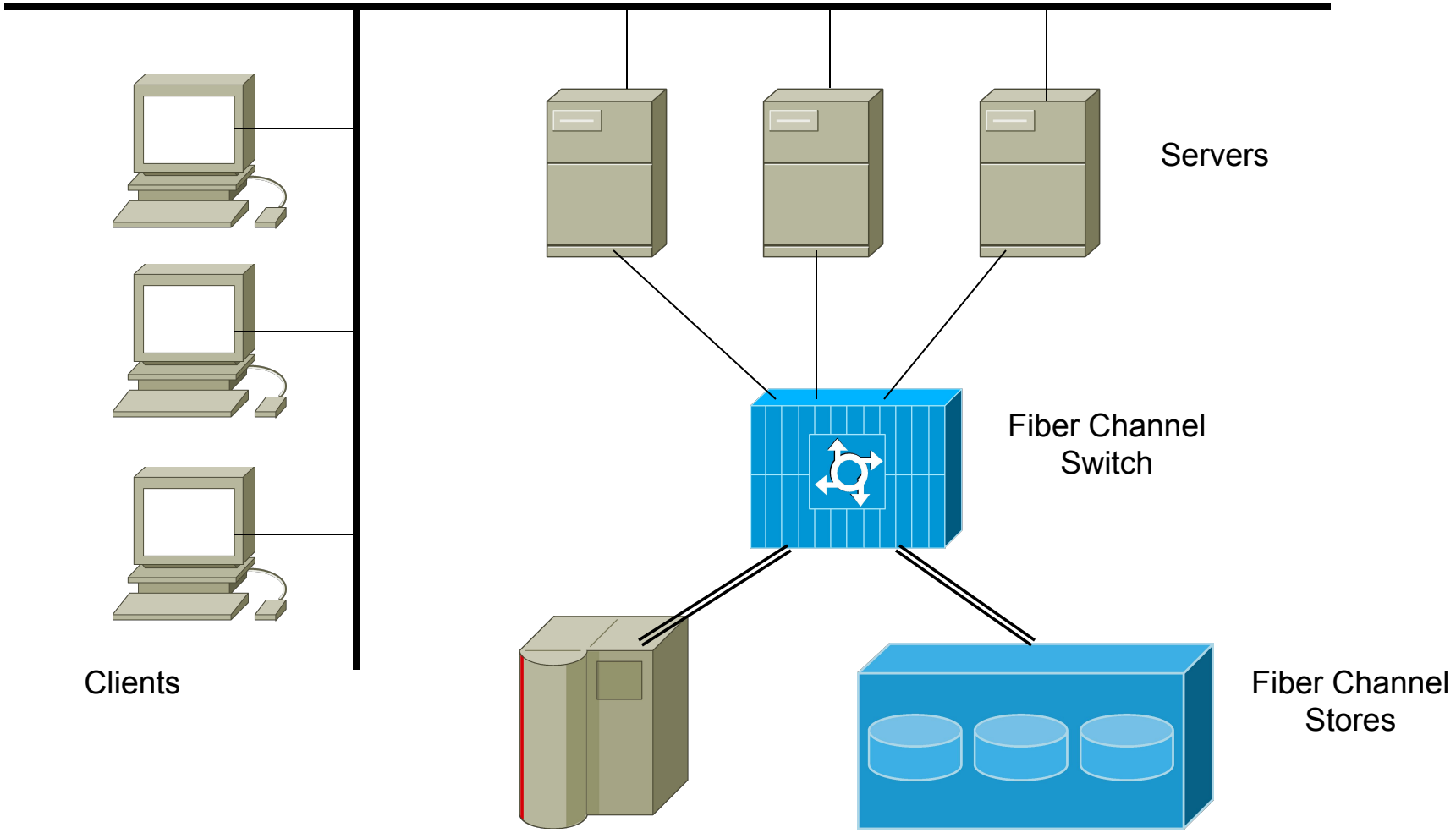
FC – Arbitrated Loop (FC Hub)



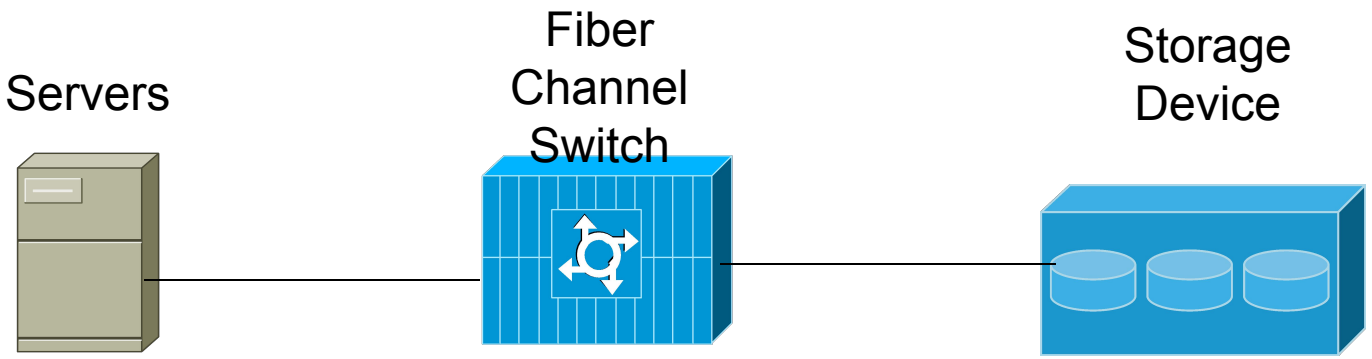
Switched FC SAN

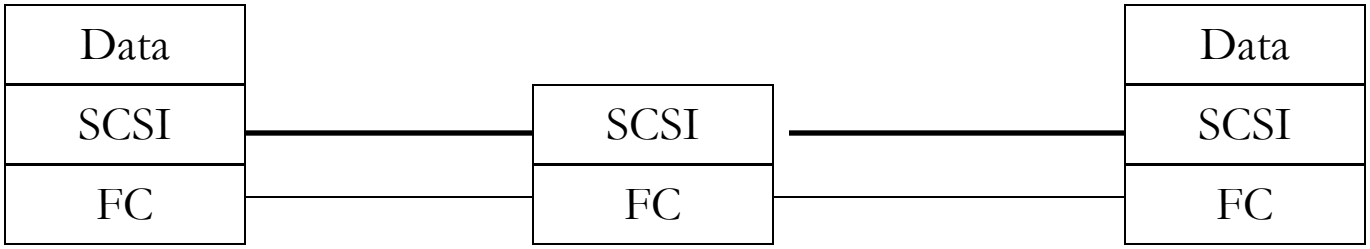
- Fibre Channel-switches function similar to traditional network switches
- Switches provide increased bandwidth, scalable performance, an increased # of devices, increased redundancy
- FC-switches vary in the number of ports and media types they support
- Multiple FC switches can be connected to form a *switch fabric* capable of supporting a large number of host servers and storage subsystems

FC – Switched SAN

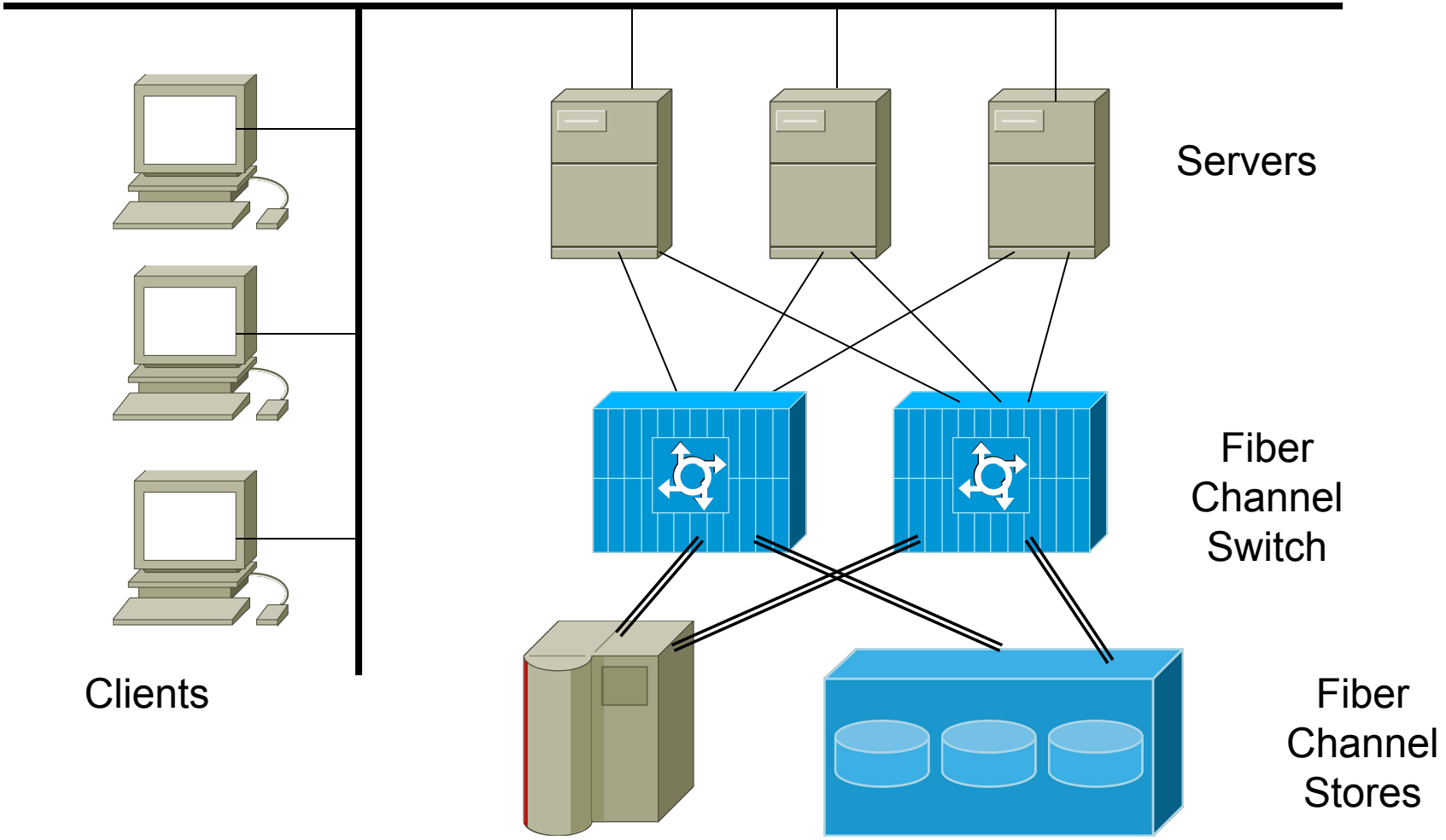


Data Access over Switched SAN



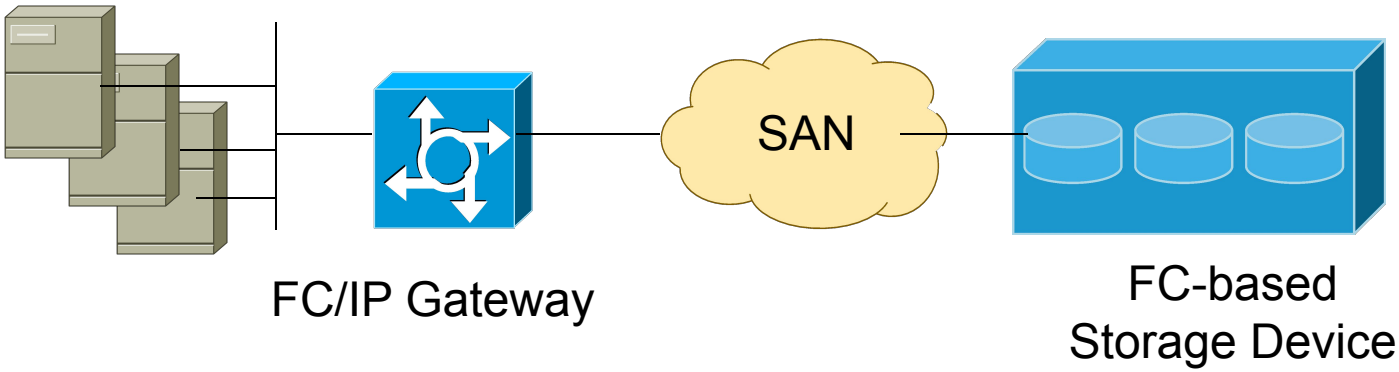


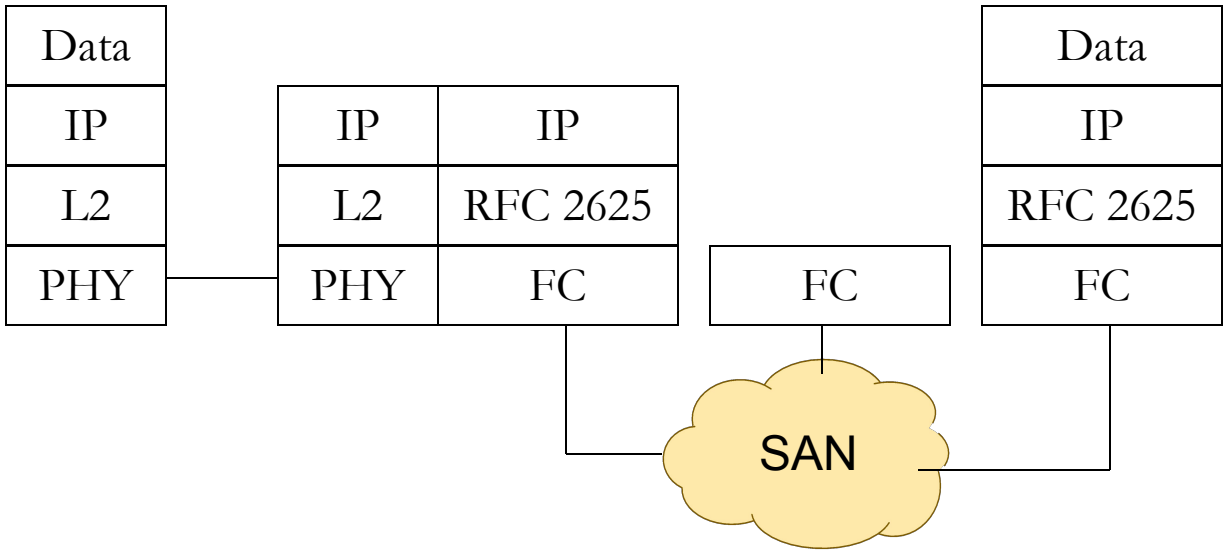
FC - Storage Area Network (redundant architecture)



IP over FC (RFC 2625)

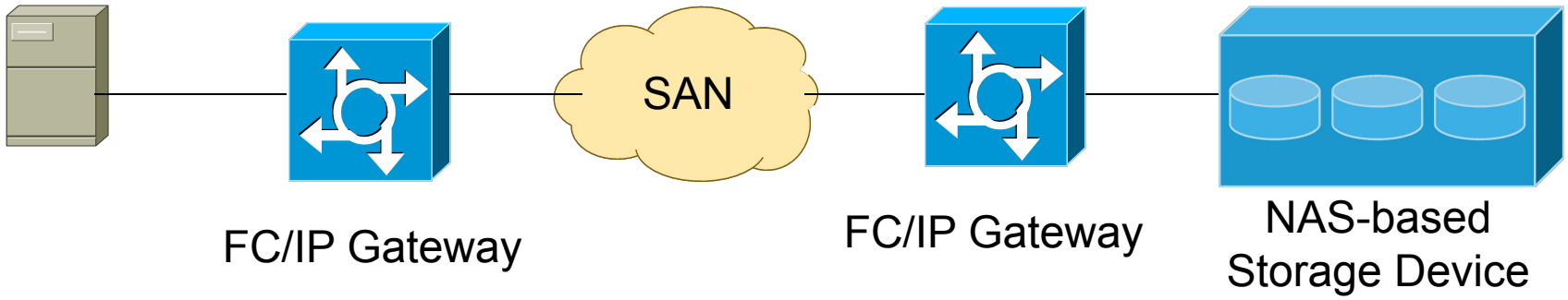
App-1: accessing SAN from IP-based servers

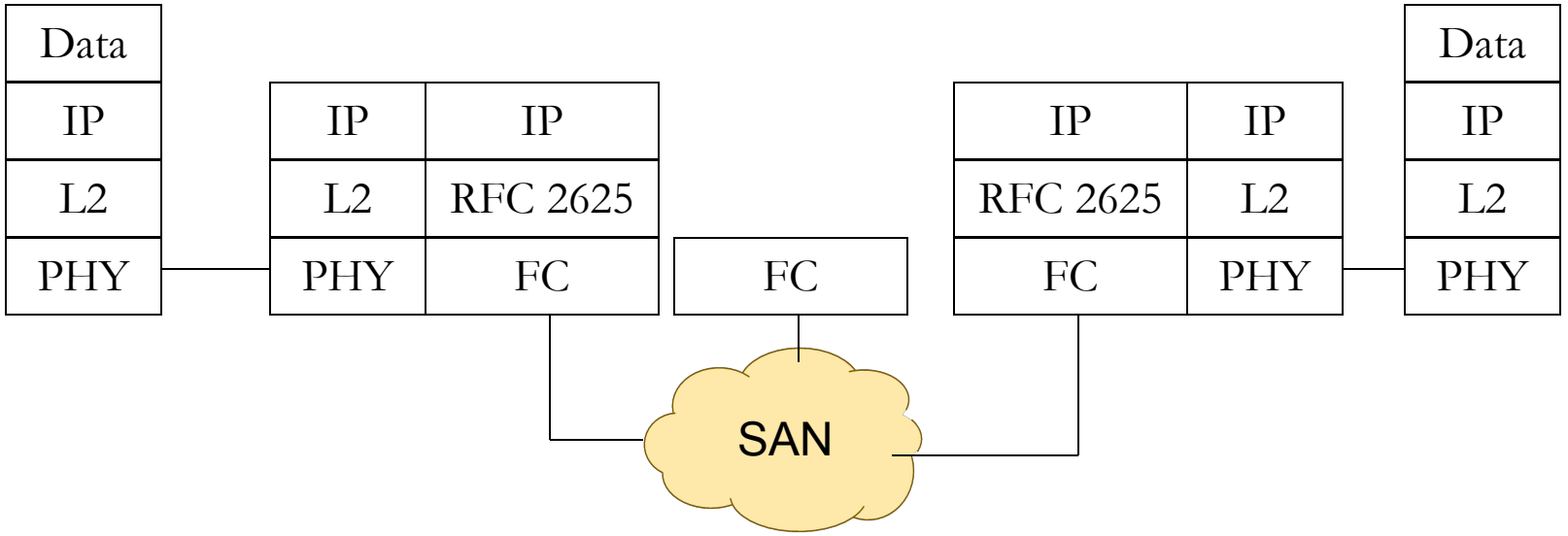




IP over FC (RFC 2625)

(App-2: interworking SAN & NAS)





Network-Attached Storage

Network Attached Storage (NAS)

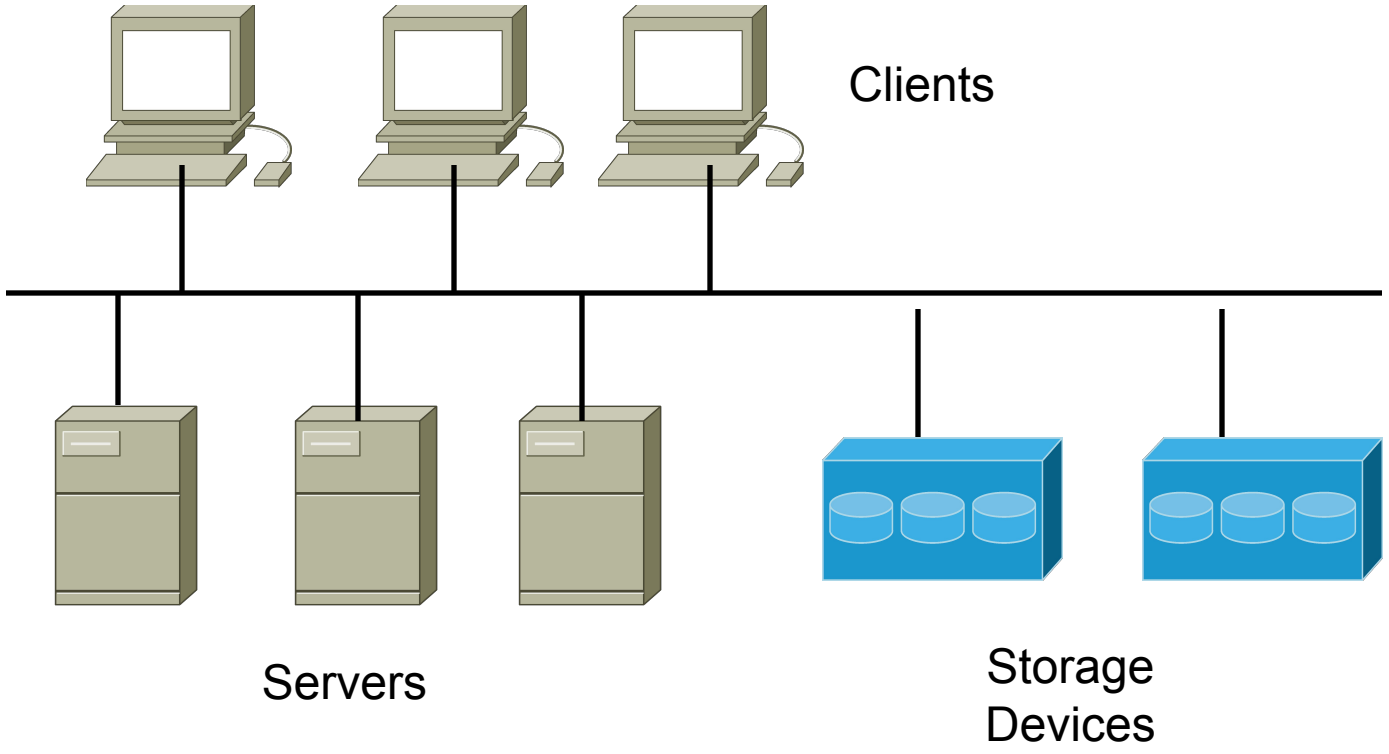
- A dedicated storage device
- Operates in a client/server mode
- NAS is connected to the file server via LAN

- **Protocol:** NFS (or CIFS) over an IP Network
  - Network File System (NFS) – UNIX/Linux
  - Common Internet File System (CIFS) – Windows Remote file system (drives) mounted on the local system (drives)
    - evolved from Microsoft NetBIOS, NetBIOS over TCP/IP (NBT), and Server Message Block (SMB)
  - SAMBA: SMB on Linux (Making Linux a Windows File Server)

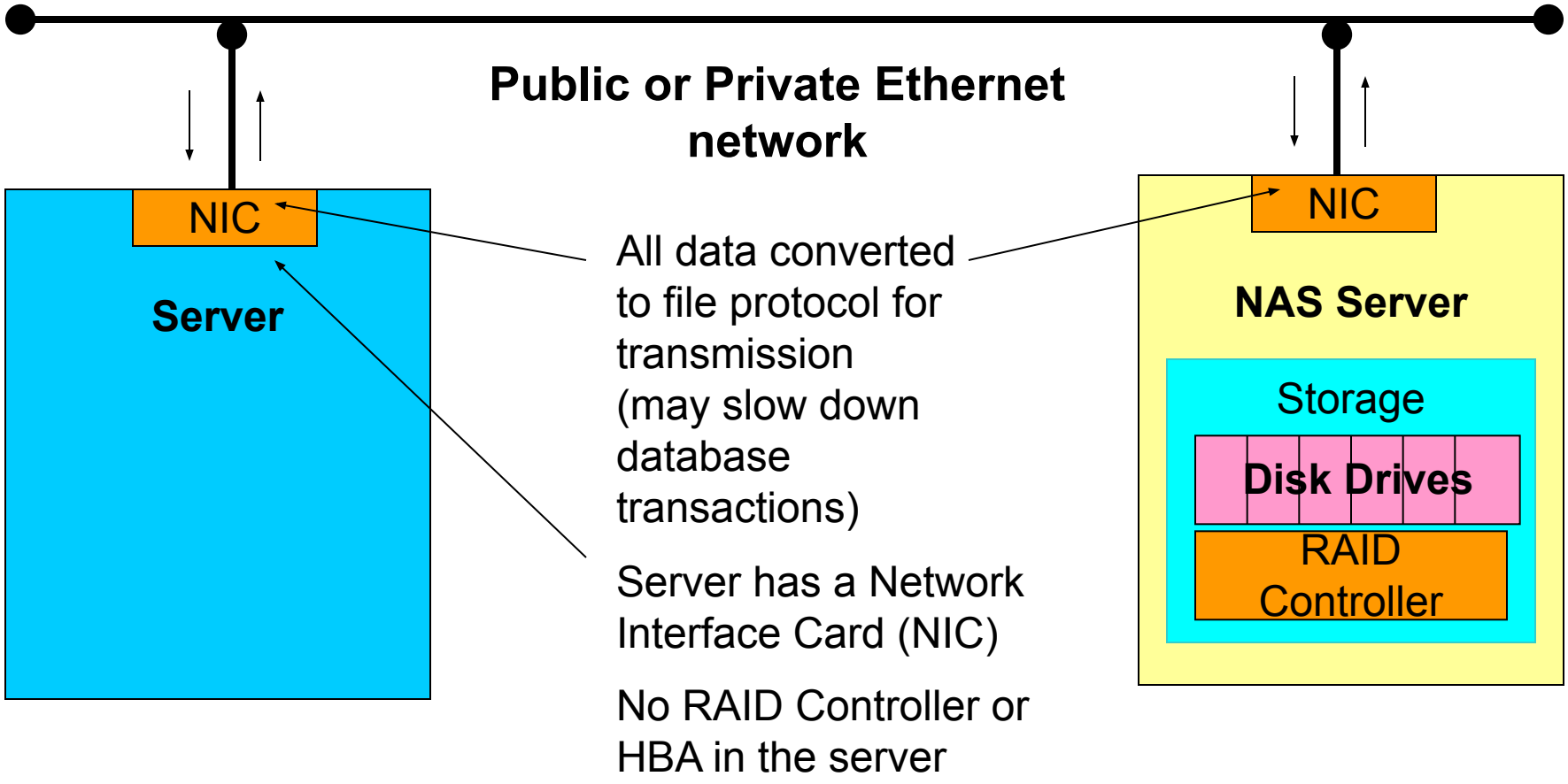
- Disadvantage: Speed and Latency

Network Attached Storage (NAS)

- Specialized storage device or group of storage devices providing centralized fault-tolerant data storage for a network
- Utilizes a TCP/IP network to “share” data
- Storage “Appliances” utilize a stripped-down OS that optimizes file protocol performance



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

iSCSI: Internet SCSI

- An alternate form of networked storage
- Like NAS, utilizes a TCP/IP network
- Encapsulates native SCSI commands in TCP/IP packets
- Supported in Windows 2003 Server and Linux
- TCP/IP Offload Engines (TOEs) on NICs speed up packet encapsulation
- Cisco and IBM co-authored original iSCSI protocol draft
- iSCSI Protocol is a standard maintained by the IETF
  - IP Storage (IPS) Working Group
  - RFC 3720

NAS vs. SAN ?

- Traditionally:
  - NAS is used for low-volume access to a large amount of storage by many users
  - SAN is the solution for terabytes (10^{12}) of storage and multiple, simultaneous access to files, such as streaming audio/video.
- The lines are becoming blurred between the two technologies now
- SAN-versus-NAS debate continues
- Both technologies complement each other