Big Data Processing Technologies

Chentao Wu
Associate Professor
Dept. of Computer Science and Engineering
wuct@cs.sjtu.edu.cn

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Schedule

- lec1: Introduction on big data and cloud computing
- lec2: Introduction on data storage
- lec3: Data reliability (Replication/Archive/EC)
- lec4: Data consistency problem
- lec5: Block storage and file storage
- lec6: Object-based storage
- lec7: Distributed file system
- lec8: Metadata management

Collaborators

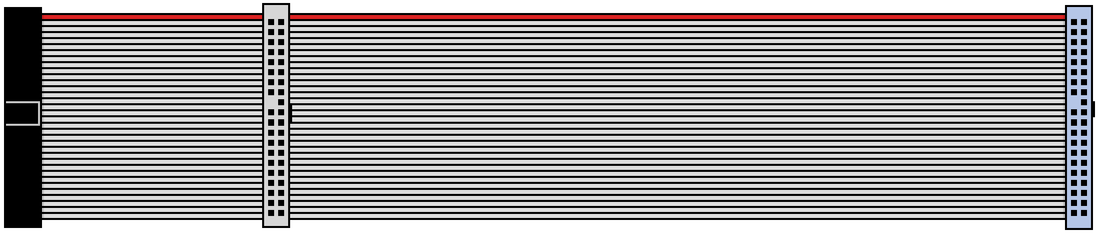
阿里云
aliyun.com

1 Interfaces of Storage Devices

ATA/IDE Interface

- **AT Attachment (ATA)**, is an interface standard for the connection of storage devices such as hard disk drives, floppy disk drives, and optical disc drives in computers. The standard is maintained by the X3/INCITS committee.

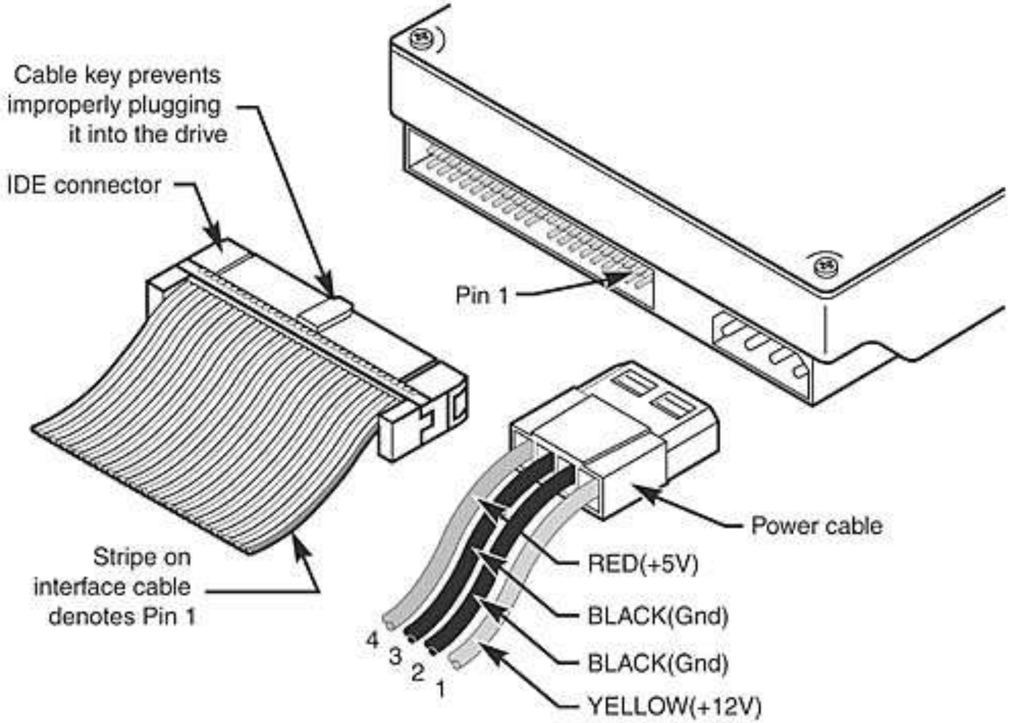
- **Parallel ATA** developed by Western Digital
- **Also called "IDE"**
  - **Integrated Device Electronics**



ATA I/O Connector

- The ATA interface connector is normally a 40-pin header-type connector with pins spaced 0.1 inches apart and generally keyed to prevent the possibility of installing it upside down.

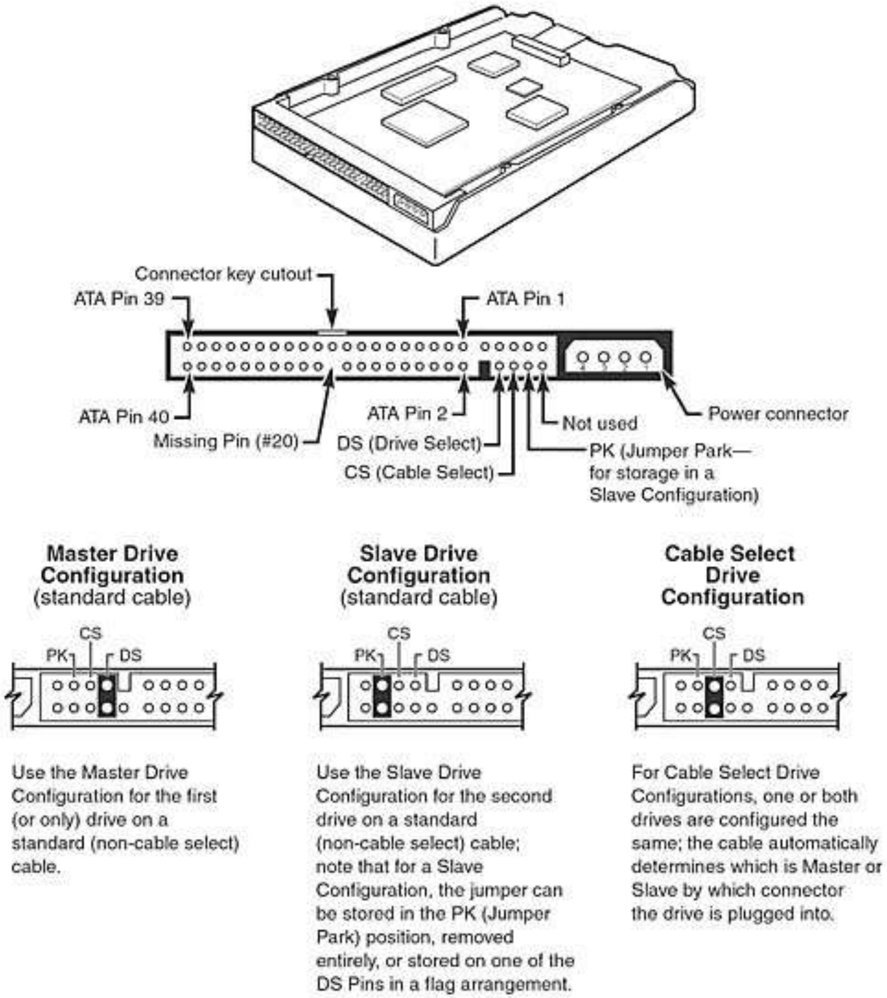
- Plugging in an IDE cable backward usually won't cause any permanent damage, however, it can lock up the system and prevent it from running at all.


Cable key prevents improperly plugging it into the drive
IDE connector
Pin 1
Power cable
RED(+5V)
BLACK(Gnd)
BLACK(Gnd)
YELLOW(+12V)
Stripe on interface cable denotes Pin 1
4
3
2
1

Dual Drive Configurations

- Most IDE drives can be configured with three settings.

- The diagram illustrates the settings of master, slave, and cable select


Connector key cutout
ATA Pin 39
ATA Pin 1
ATA Pin 40
Missing Pin (#20)
ATA Pin 2
DS (Drive Select)
CS (Cable Select)
Not used
PK (Jumper Park—for storage in a Slave Configuration)
Power connector
Master Drive Configuration (standard cable)
CS
PK
DS
Slave Drive Configuration (standard cable)
CS
PK
DS
Cable Select Drive Configuration
CS
PK
DS

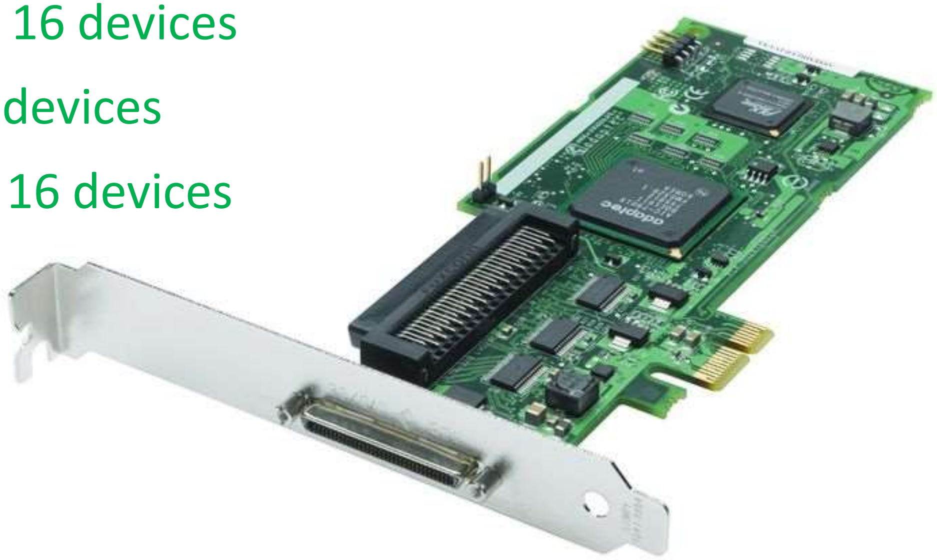
Use the Master Drive Configuration for the first (or only) drive on a standard (non-cable select) cable.

Use the Slave Drive Configuration for the second drive on a standard (non-cable select) cable; note that for a Slave Configuration, the jumper can be stored in the PK (Jumper Park) position, removed entirely, or stored on one of the DS Pins in a flag arrangement.

For Cable Select Drive Configurations, one or both drives are configured the same; the cable automatically determines which is Master or Slave by which connector the drive is plugged into.

Small Computer System Interface (SCSI)

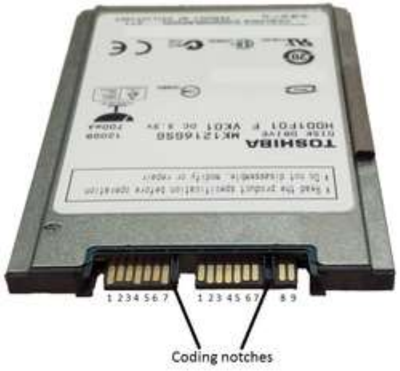
- **SCSI** refers to the types of cables and ports used to connect certain types of hard drives, optical drives, scanners, and other peripheral devices to a computer.
- **Fast SCSI:** 10 MBps; connects 8 devices
- **Fast Wide SCSI:** 20 MBps; connects 16 devices
- **Ultra Wide SCSI:** 40 MBps; connects 16 devices
- **Ultra3 SCSI:** 160 MBps; connects 16 devices
- **Ultra-640 SCSI:** 640 MBps; connects 16 devices



Serial ATA (SATA)

- **Serial ATA (SATA)** is a computer bus interface that connects host bus adapters to mass storage devices such as hard disk drives, optical drives, and solid-state drives.

- **Compared to PATA/IDE**
    - reduced cable size and cost
        - seven conductors instead of 40 or 80
    - native hot swapping
    - faster data transfer
        - through higher signaling rates
        - through an I/O queuing protocol

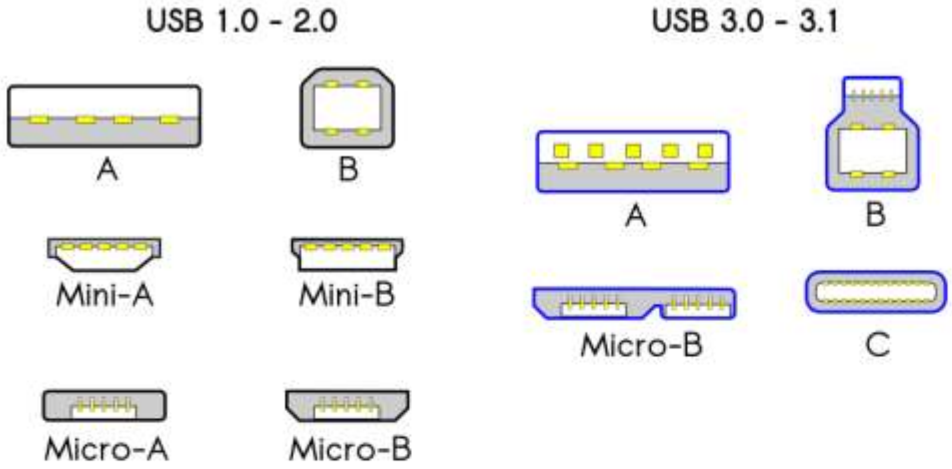


Serial Attached SCSI (SAS)

- **Serial Attached SCSI (SAS)** is a point-to-point serial protocol that moves data to and from computer storage devices such as hard drives and tape drives.
- SAS replaces the older Parallel SCSI bus technology.

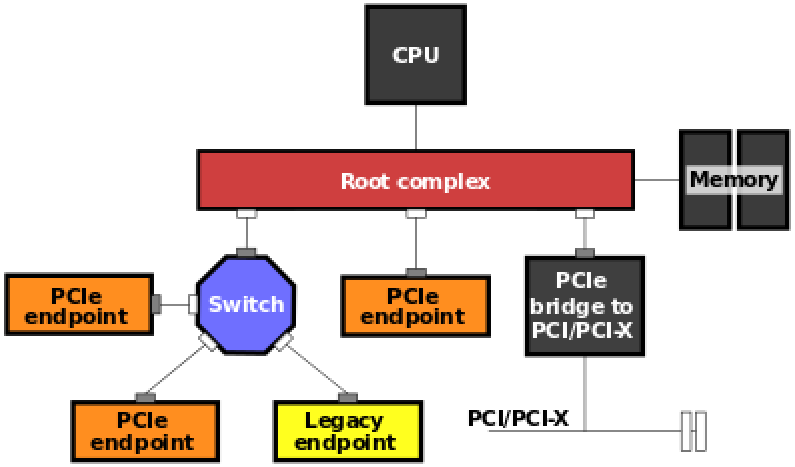
USB

- **Universal Serial Bus (USB)**, is an industry standard initially developed in the mid-1990s that defines the cables, connectors and communications protocols used in a bus for connection, communication, and power supply between computers and electronic devices.

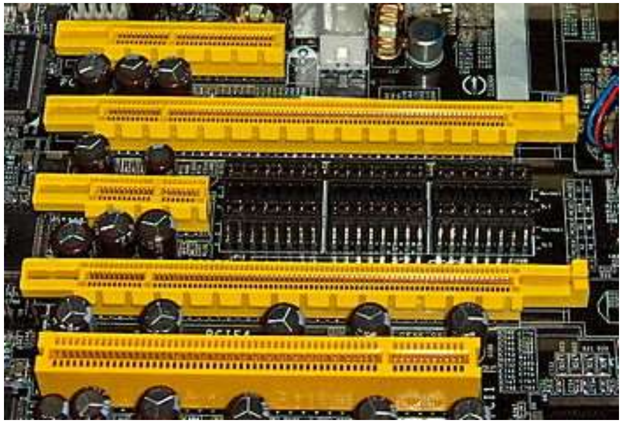


PCI Express (PCIe)

- **PCI Express** (Peripheral Component Interconnect Express) is a high-speed serial computer expansion bus standard, designed to replace the older PCI, PCI-X, and AGP bus standards.



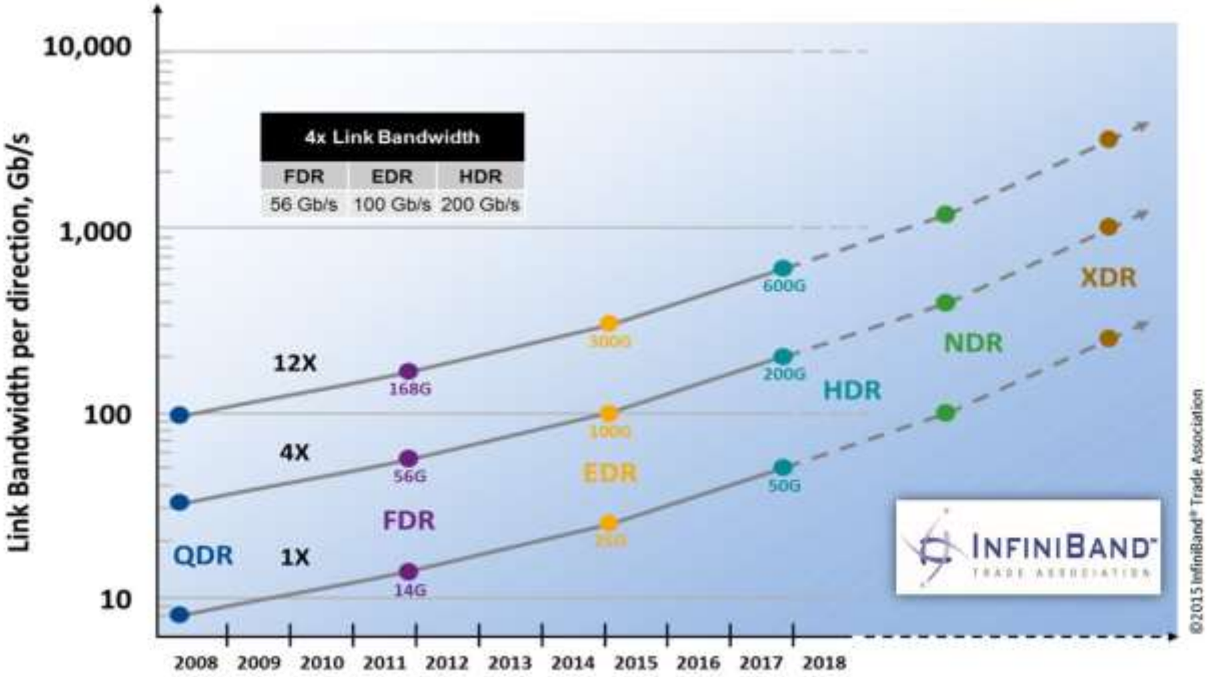
- Intel NVMe SSD with PCIe



- PCI Express 4/16/1/16
- Typical PCI

Infiniband (IB)

- **InfiniBand (IB)** is a computer-networking communications standard used in high-performance computing that features very high throughput and very low latency.
  - Support RDMA



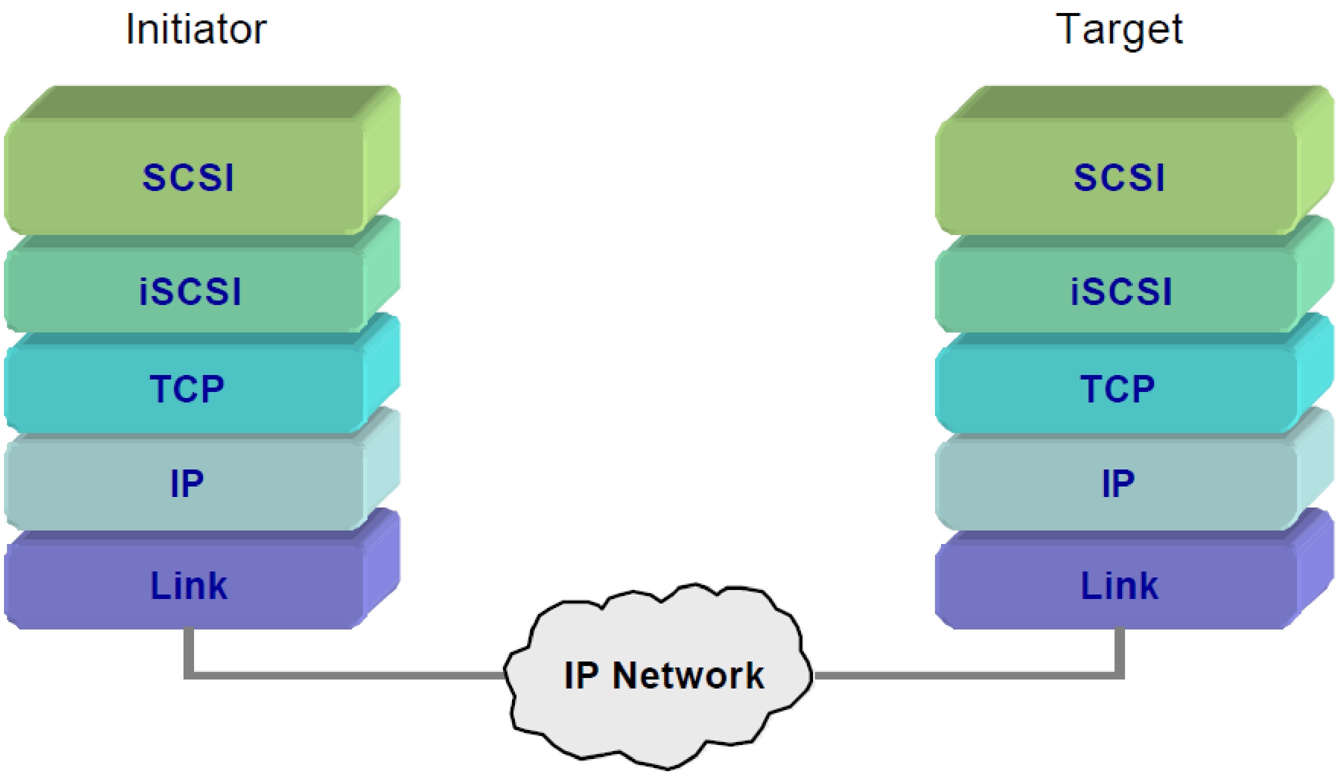
4x Link Bandwidth		
FDR	EDR	HDR
56 Gb/s	100 Gb/s	200 Gb/s

iSCSI (Internet SCSI)(1)

- Why iSCSI?
  - **Storage Area Networks (SANs)** based on serial gigabit transports overcome the distance, performance, scalability and availability restrictions of parallel SCSI implementations.
- What is iSCSI?
  - **Internet SCSI (iSCSI) protocol**
  - **Defined by the IP Storage work group of the IETF**
  - **IETF RFC 3720**

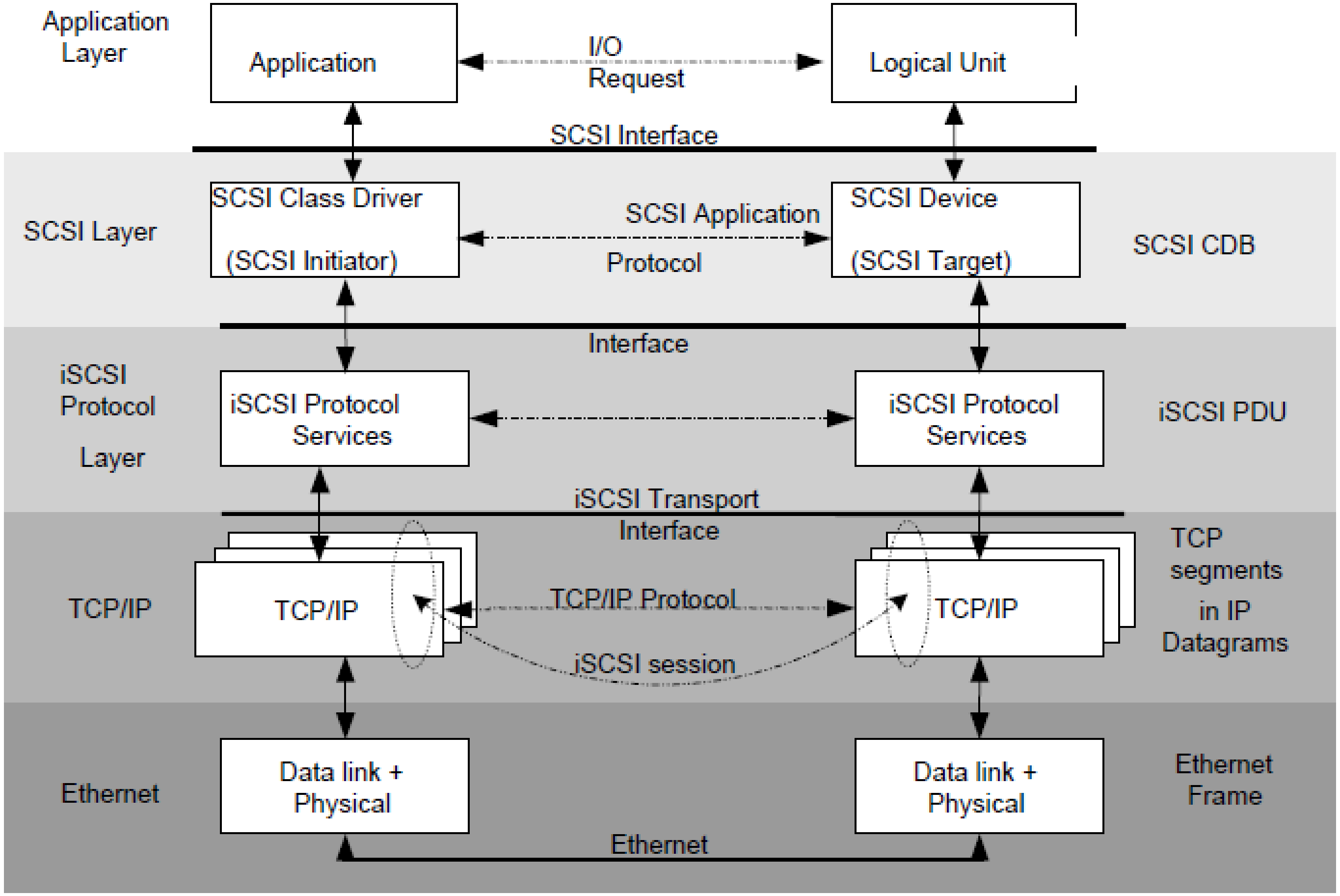
iSCSI (Internet SCSI) (2)

- **iSCSI Protocol Layering Model**



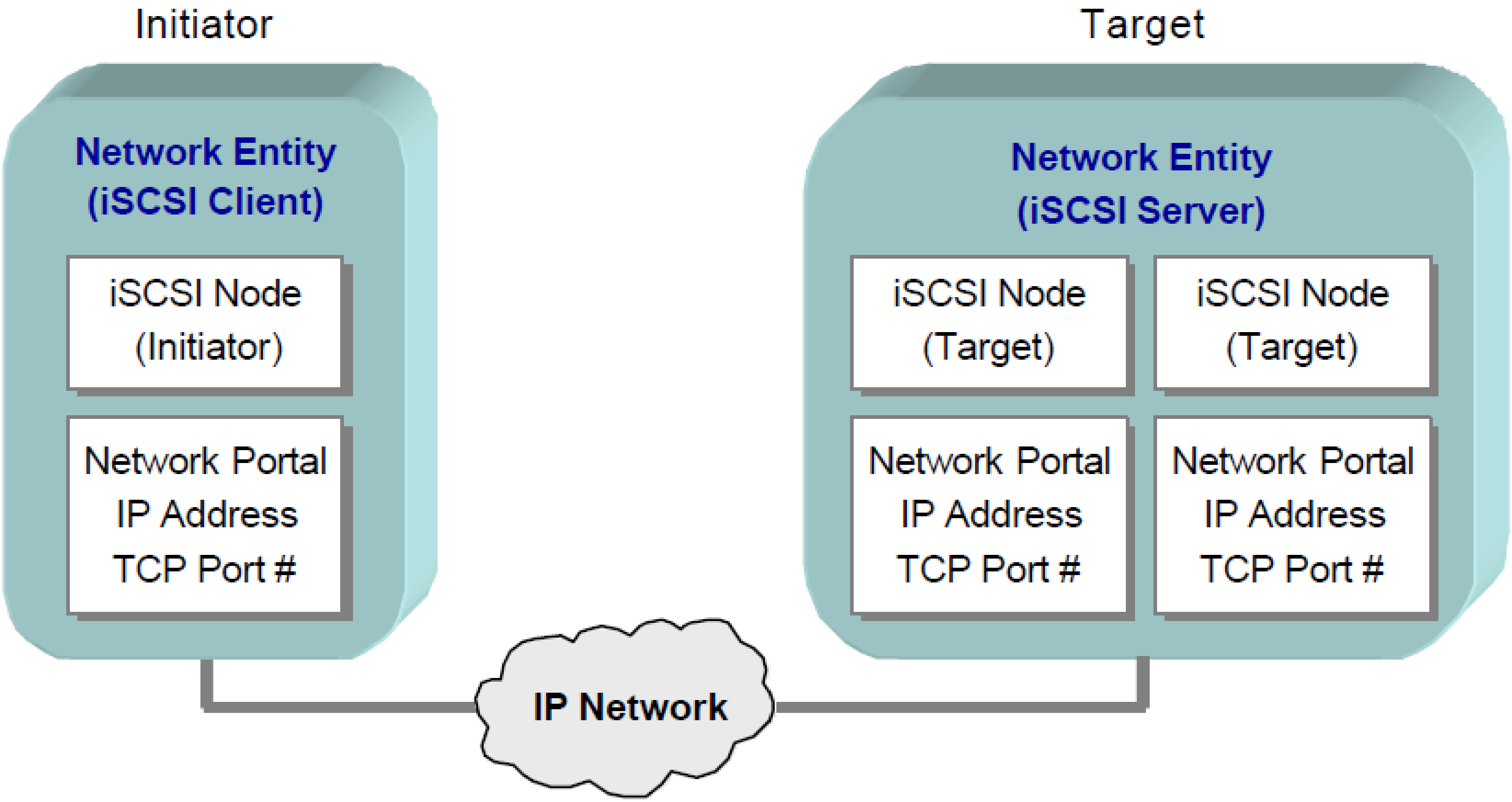
iSCSI (Internet SCSI) (3)

- **Encapsulates SCSI Command Descriptor Blocks (CDBs)**



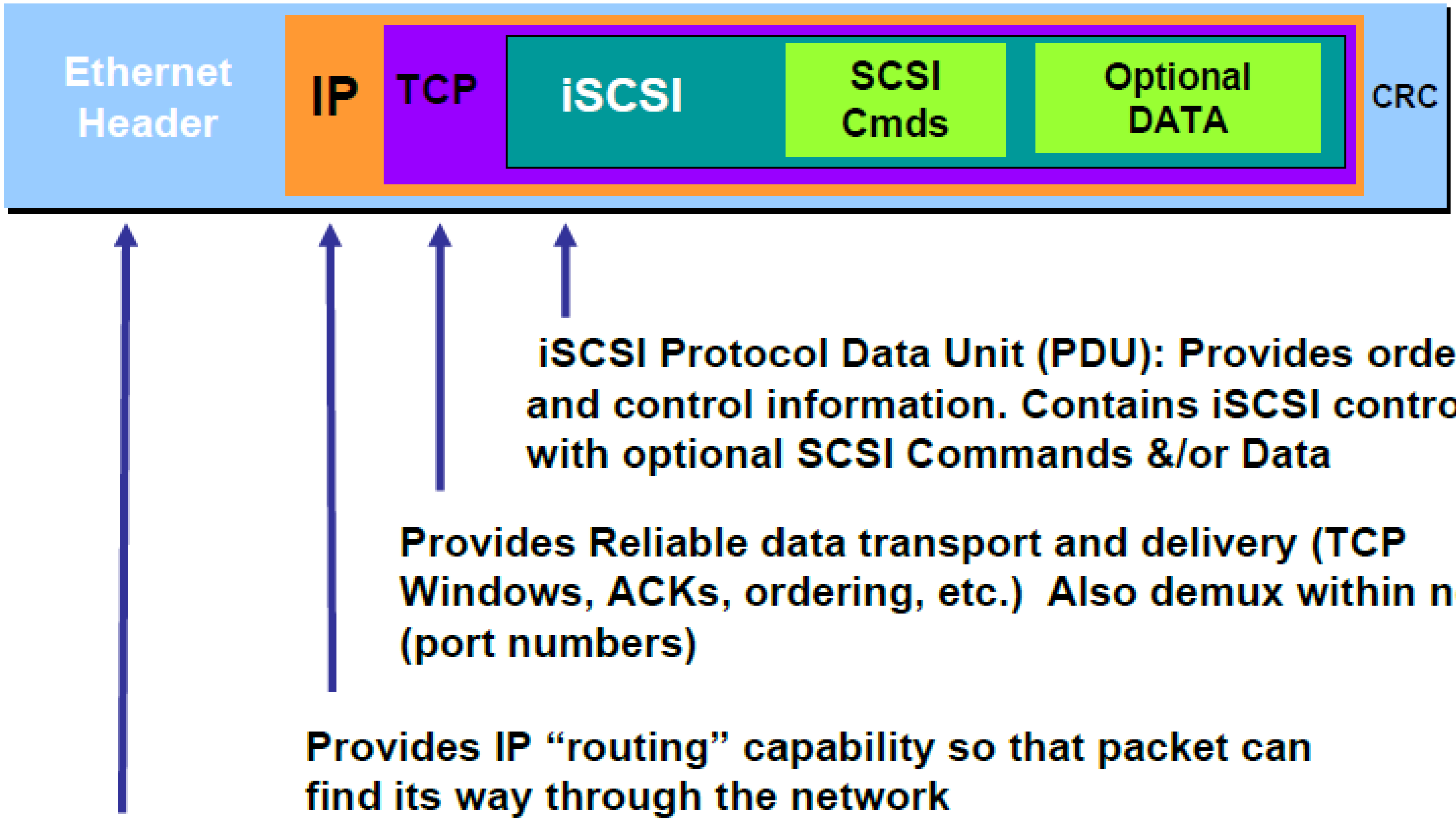
iSCSI (Internet SCSI) (4)

- **iSCSI Protocol – Highest Level**



iSCSI (Internet SCSI) (5)

- **Data Encapsulation**



iSCSI Protocol Data Unit (PDU): Provides ordering and control information. Contains iSCSI control info, with optional SCSI Commands &/or Data

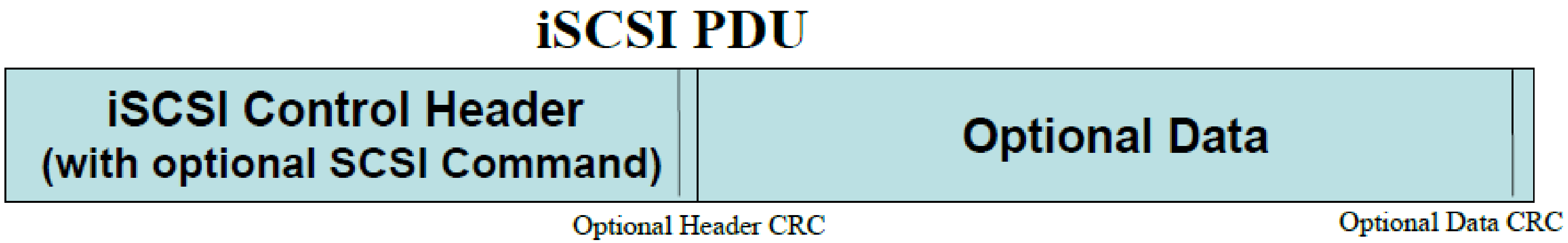
Provides Reliable data transport and delivery (TCP Windows, ACKs, ordering, etc.) Also demux within node (port numbers)

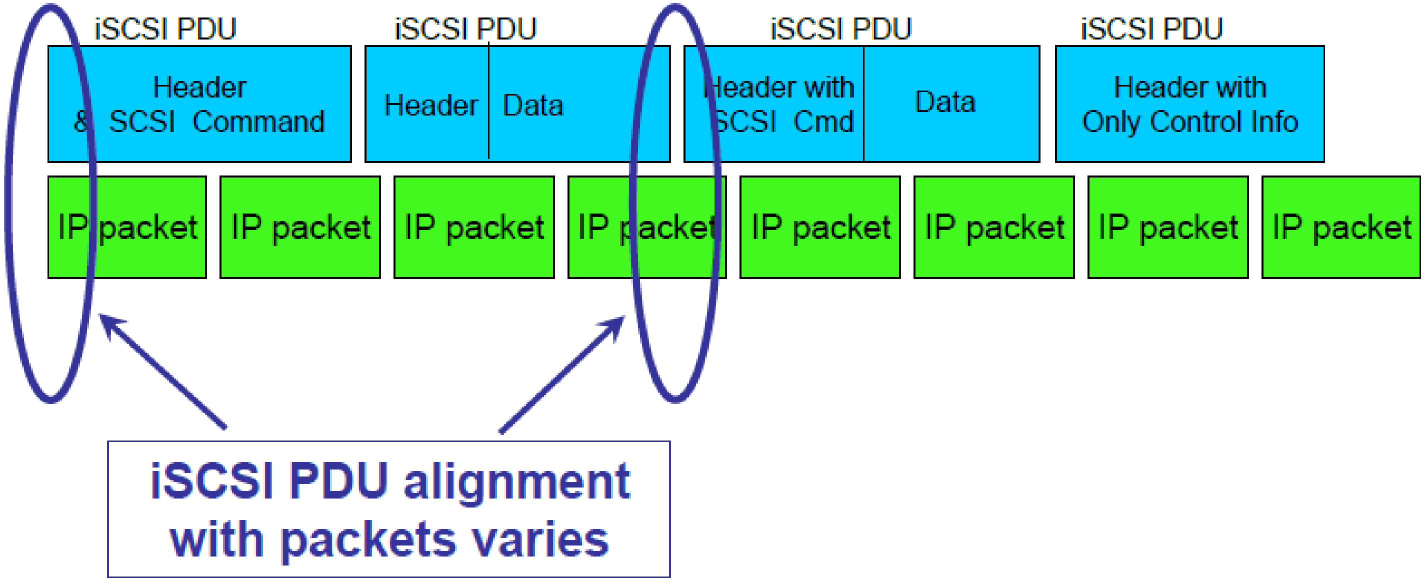
Provides IP "routing" capability so that packet can find its way through the network

Provides physical network capability (Cat 5, MAC, etc.)

iSCSI (Internet SCSI) (6)

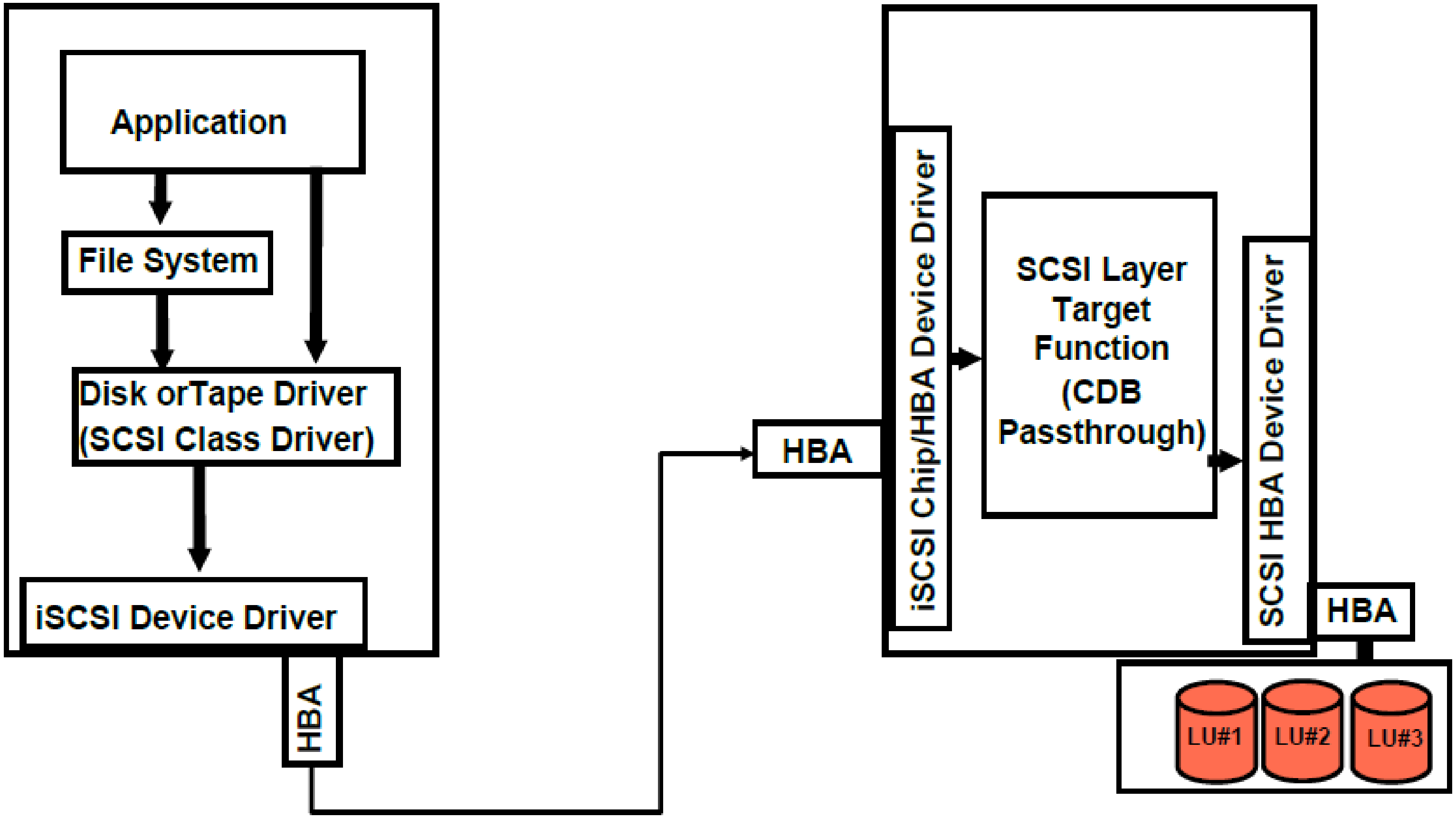
- **iSCSI Protocol Data Unit (PDU)**





iSCSI Command Flow

- **From application to Logical Unit (LU)**

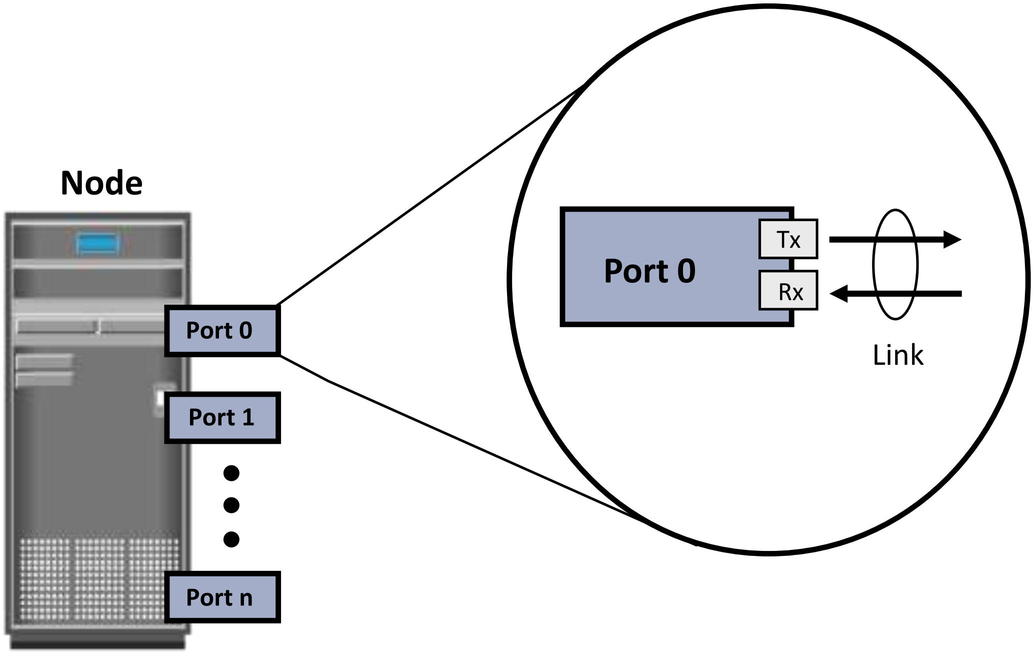


FC (Fiber Channel)

- Fiber Channel, or FC, is a high-speed network technology (commonly running at 1, 2, 4, 8, 16, 32, and 128 gigabit per second rates) primarily used to connect computer data storage to servers.
- Fibre Channel is mainly used in Storage Area Networks (SAN) in commercial data centers.

FC Node Ports

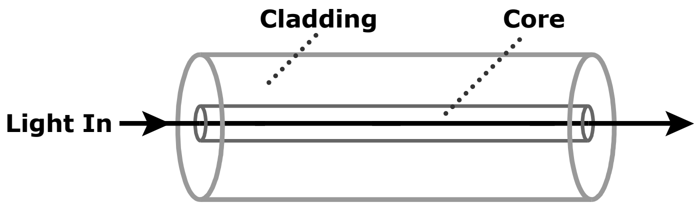
- Provide physical interface for communicating with other nodes
- Exist on
  - HBA (Host Bus Adapter) in server
  - Front-end adapters in storage
- Each port has a transmit (Tx) link and a receive (Rx) link



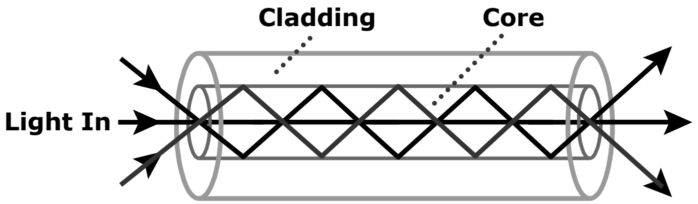
FC Cables

- Implementation uses
  - Copper cables for short distance
  - Optical fiber cables for long distance
- Two types of optical cables: single-mode and multimode

Single-mode	Multimode
Carries single beam of light	Can carry multiple beams of light simultaneously
Distance up to 10km	Used for short distance (Modal dispersion weakens signal strength after certain distance)



Single-mode Fiber



Multimode Fiber

FC Connectors

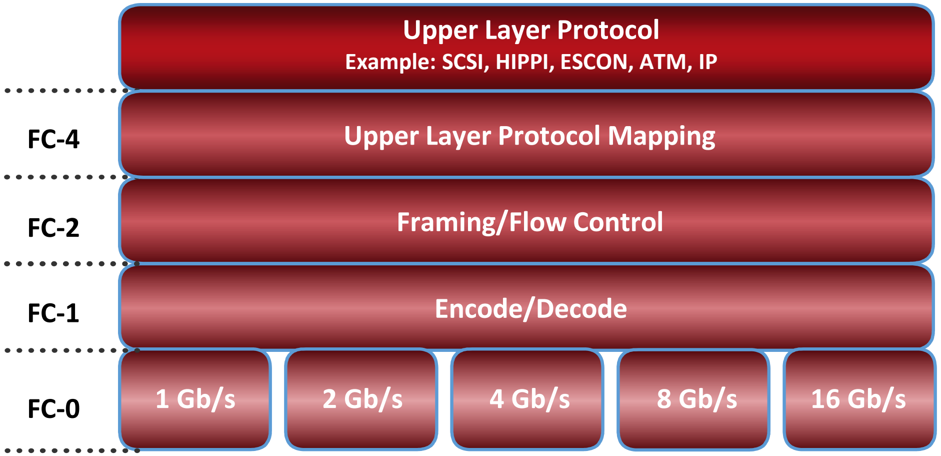
- Attached at the end of a cable
- Enable swift connection and disconnection of the cable to and from a port
- Commonly used connectors for fiber optic cables are:
  - Standard Connector (SC)
    - Duplex connectors
  - Lucent Connector (LC)
    - Duplex connectors
  - Straight Tip (ST)
    - Patch panel connectors
    - Simplex connectors

Standard Connector

Lucent Connector

Straight Tip Connector

Fibre Channel Protocol Stack



Upper Layer Protocol
Example: SCSI, HIPPI, ESCON, ATM, IP

FC-4 Upper Layer Protocol Mapping

FC-2 Framing/Flow Control

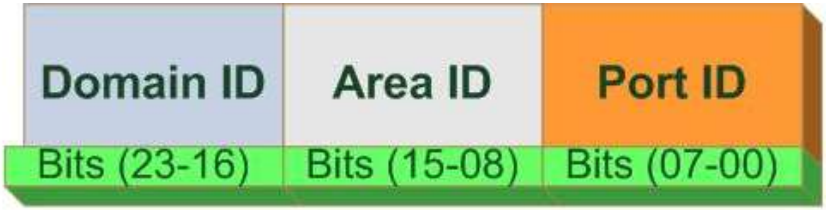
FC-1 Encode/Decode

FC-0 1 Gb/s 2 Gb/s 4 Gb/s 8 Gb/s 16 Gb/s

FC Layer	Function	Features Specified by FC Layer
FC-4	Mapping interface	Mapping upper layer protocol (e.g. SCSI) to lower FC layers
FC-3	Common services	Not implemented
FC-2	Routing, flow control	Frame structure, FC addressing, flow control
FC-1	Encode/decode	8b/10b or 64b/66b encoding, bit and frame synchronization
FC-0	Physical layer	Media, cables, connector

FC Addressing in Switched Fabric

- FC Address is assigned to nodes during fabric login
  - Used for communication between nodes within FC SAN
- Address format



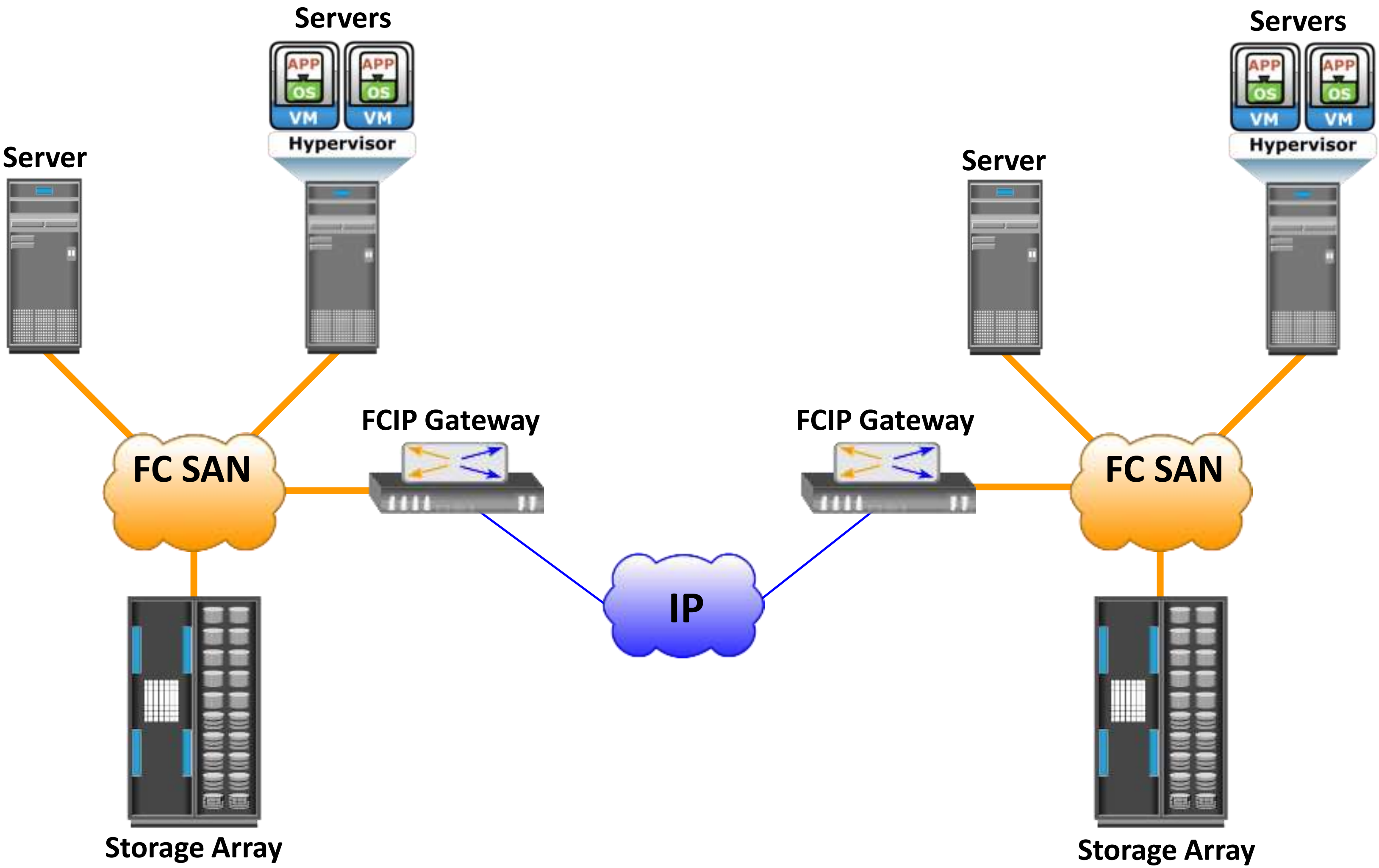
Domain ID	Area ID	Port ID
Bits (23-16)	Bits (15-08)	Bits (07-00)

- Domain ID is a unique number provided to each switch in the fabric
  - 239 addresses are available for domain ID
- Maximum possible number of node ports in a switched fabric:
  - 239 domains X 256 areas X 256 ports = 15,663,104

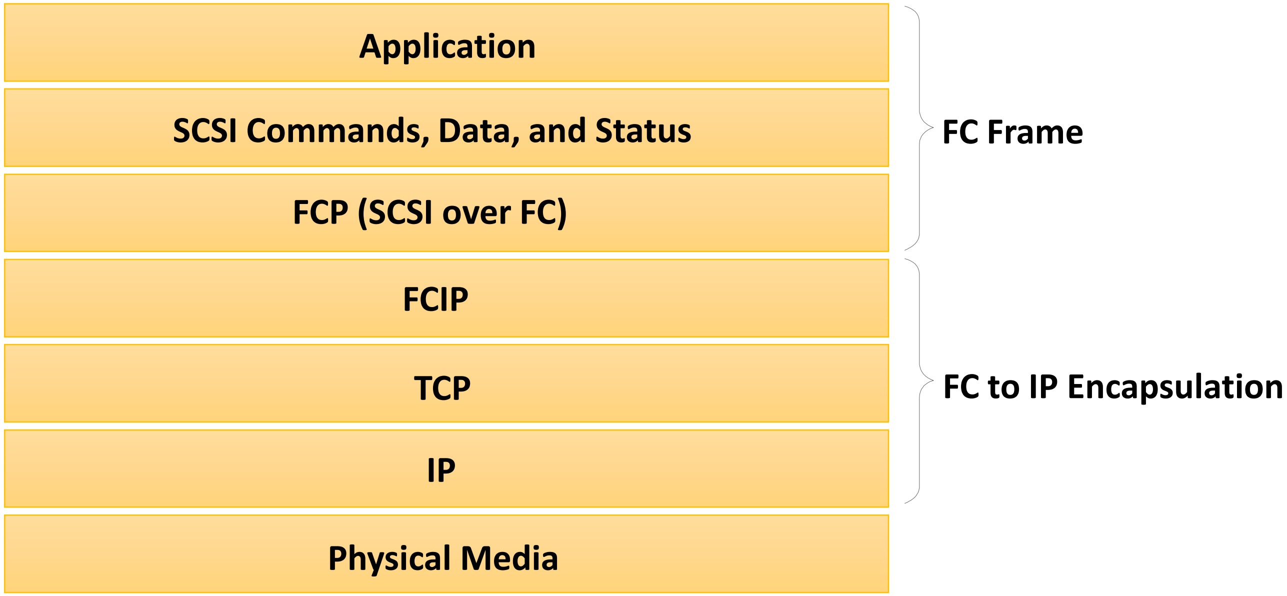
FCIP (IP SAN Protocol)

- IP-based protocol that is used to connect distributed FC SAN islands
- Creates virtual FC links over existing IP network that is used to transport FC data between different FC SANs
- Encapsulates FC frames onto IP packet
- Provides disaster recovery solution

FCIP Topology



FCIP Protocol Stack



Application

SCSI Commands, Data, and Status

FCP (SCSI over FC)

FCIP

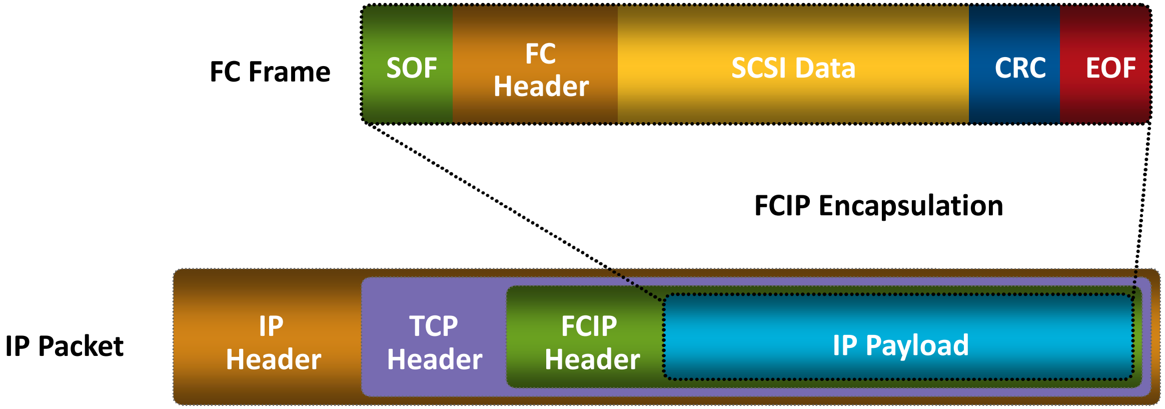
TCP

IP

Physical Media

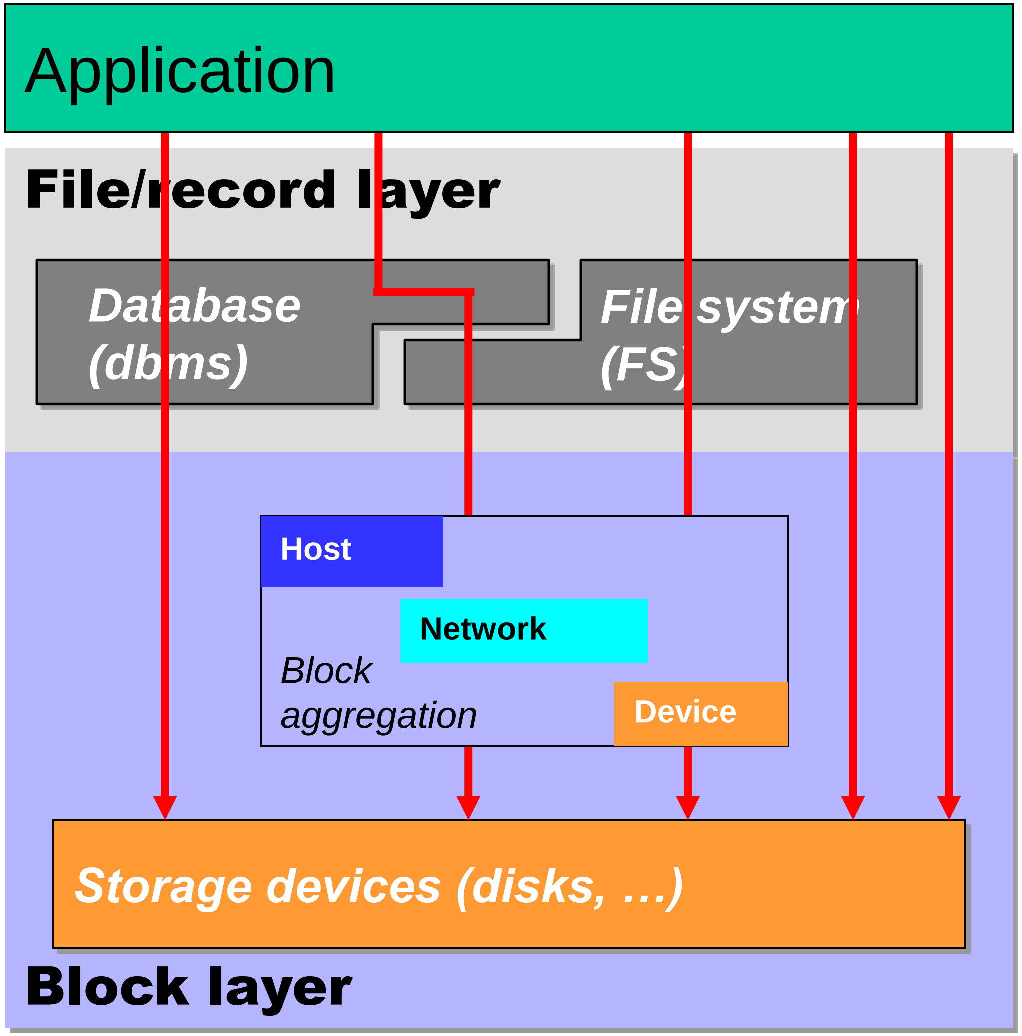
FC Frame

FC to IP Encapsulation

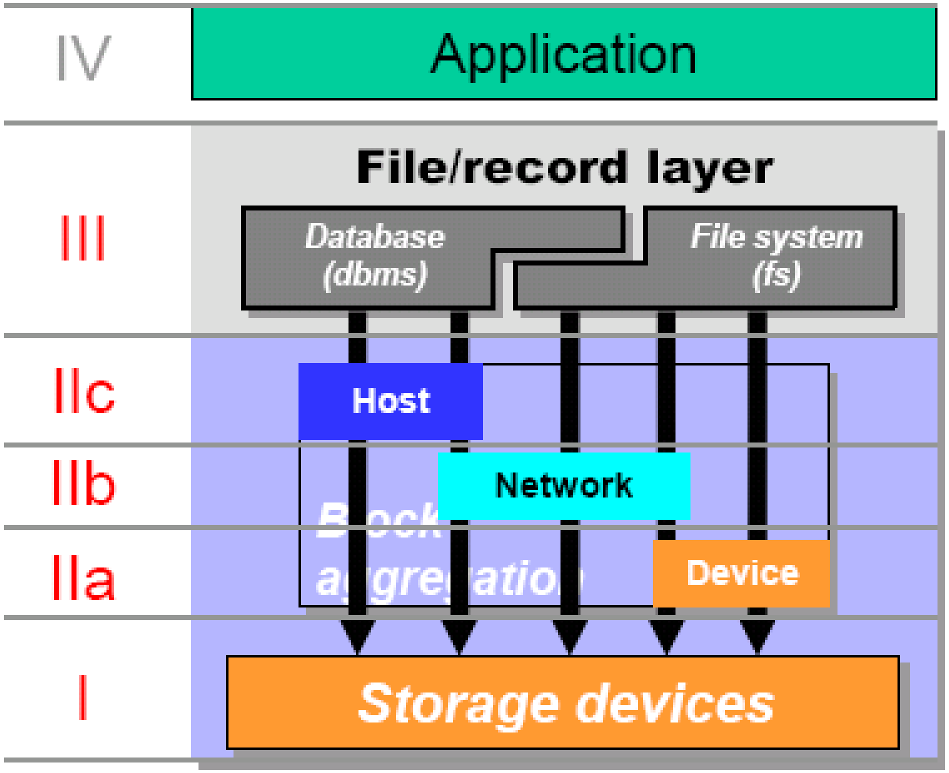


2 Block Storage

The SNIA shared storage model (1)



The SNIA shared storage model (2)



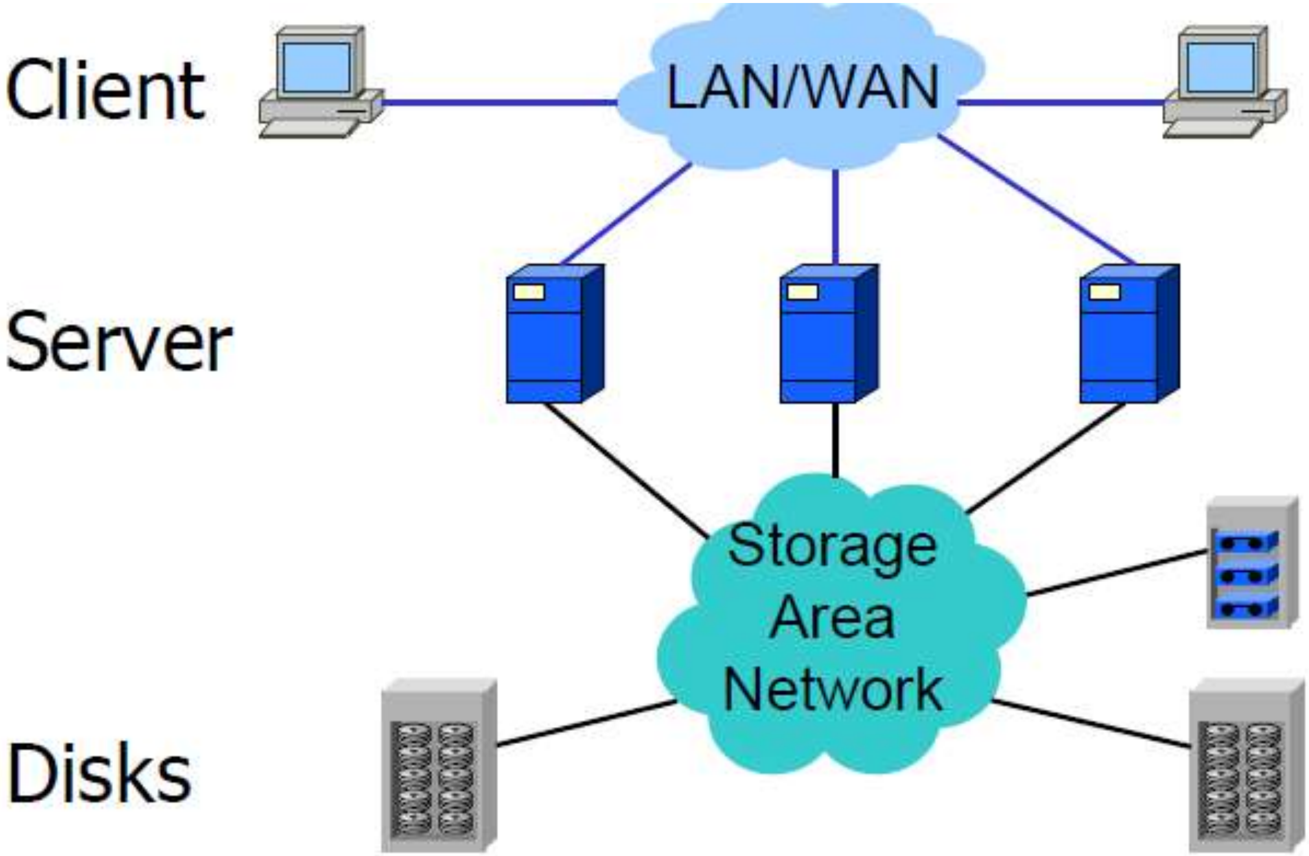
The layers are as follows:

- IV. Application
- III. File/record layer
  - IIIb. Database
  - IIIa. File system
- II. Block aggregation layer, with three function-placements:
  - IIc. Host
  - IIb. Network
  - IIa. Device
- I. Storage devices

Typical Block Devices

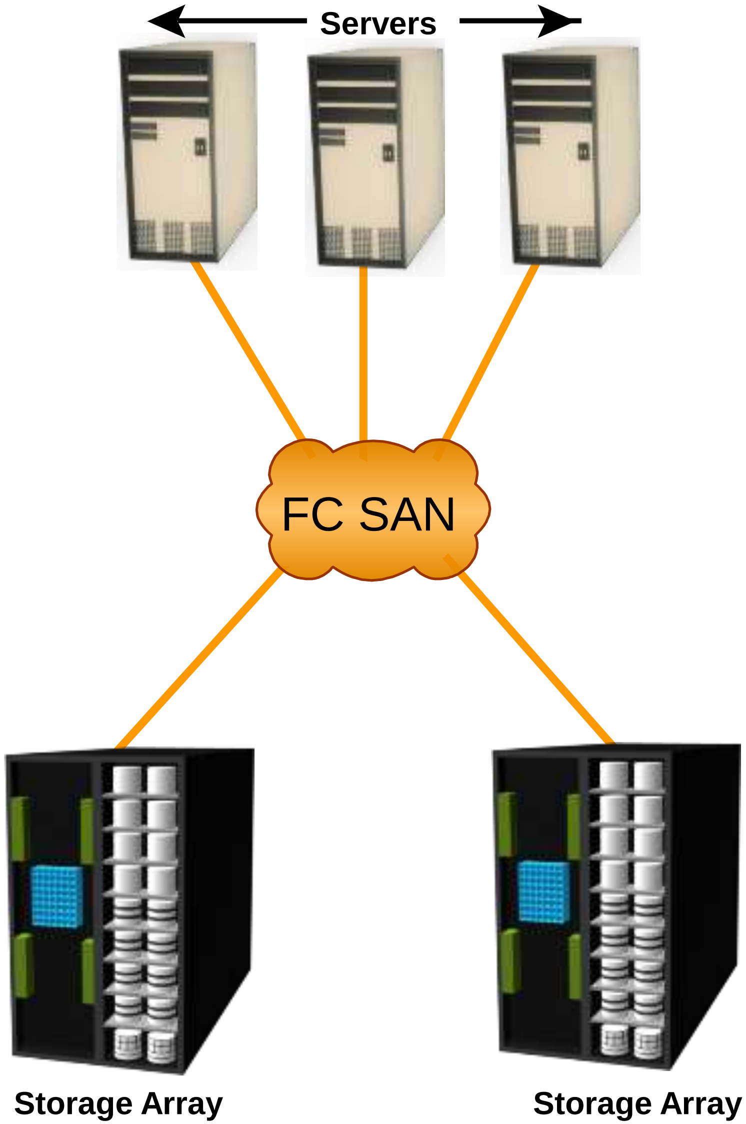
- Hard Disk Drives (HDDs)
- Solid State Drives (SSDs)
- Storage Arrays (RAID)
- Storage Area Network (SAN)
  - Dedicated high speed network of servers and shared storage devices

Storage Area Network (SAN)



Features of a SAN

- Provide block level data access
- Resource Consolidation
  - Centralized storage and management
- Scalability
  - Theoretical limit: Appx. 15 million devices
- Secure Access



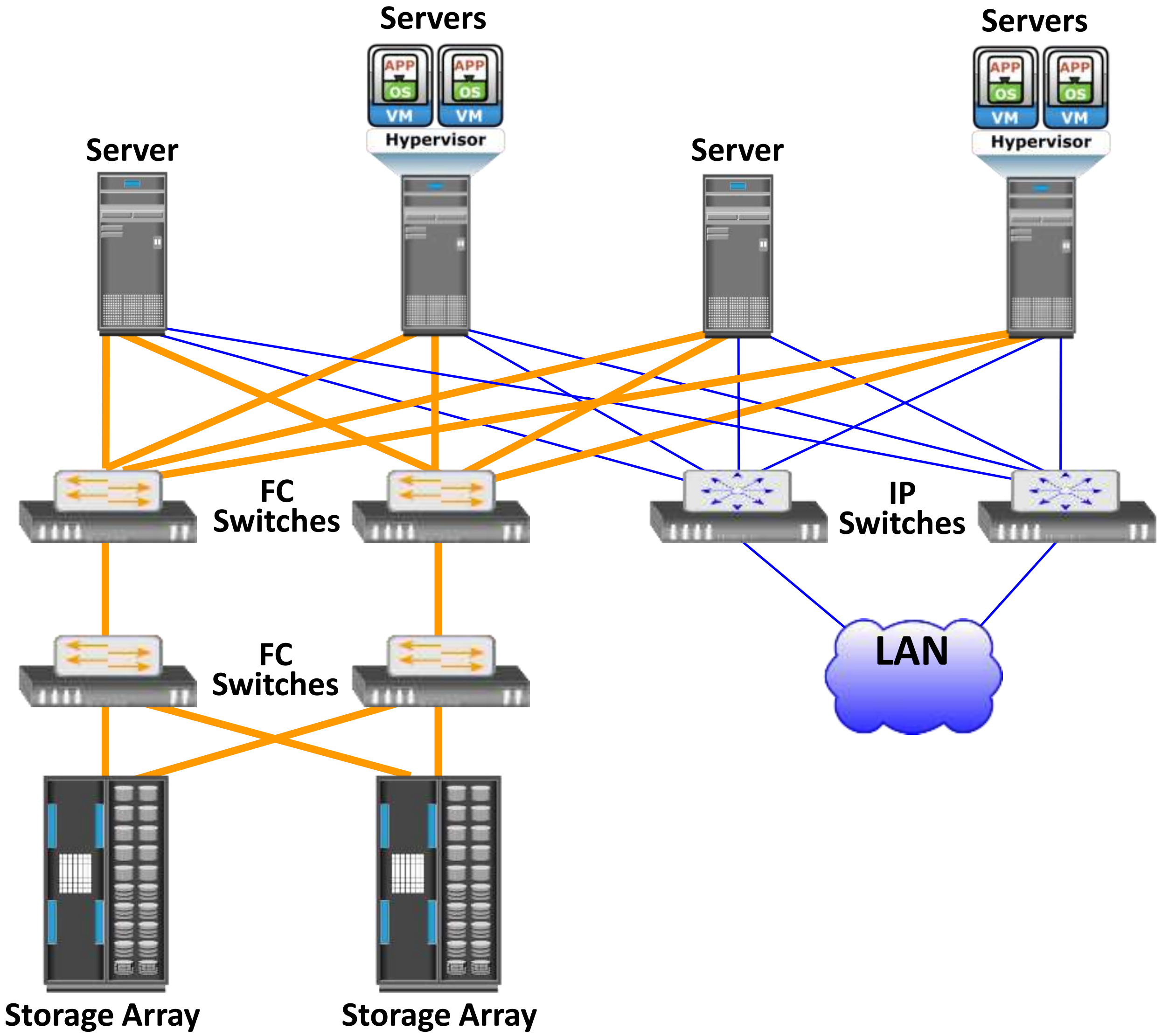
Types of SANs in Data Center

- Storage Area Network (SAN)
- IP SAN
- FC SAN
- FCoE SAN
- Infiniband SAN??

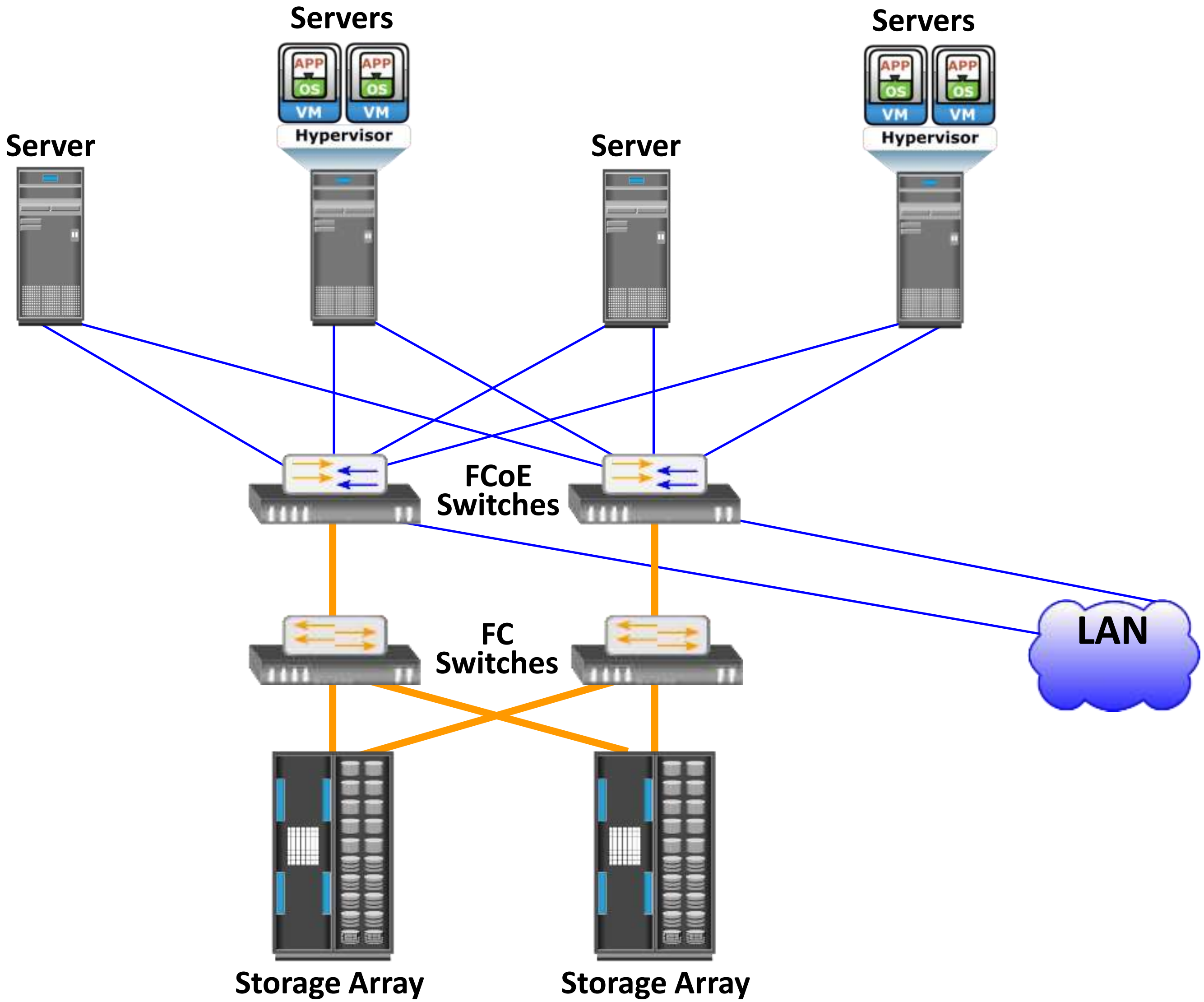
Drivers for FCoE

- FCoE is a protocol that transports FC data over Ethernet network (Converged Enhanced Ethernet)
- FCoE is being positioned as a storage networking option because:
  - Enables consolidation of FC SAN traffic and Ethernet traffic onto a common Ethernet infrastructure
  - Reduces the number of adapters, switch ports, and cables
  - Reduces cost and eases data center management
  - Reduces power and cooling cost, and floor space

Data Center Infrastructure – Before Using FCoE



Data Center Infrastructure – After Using FCoE

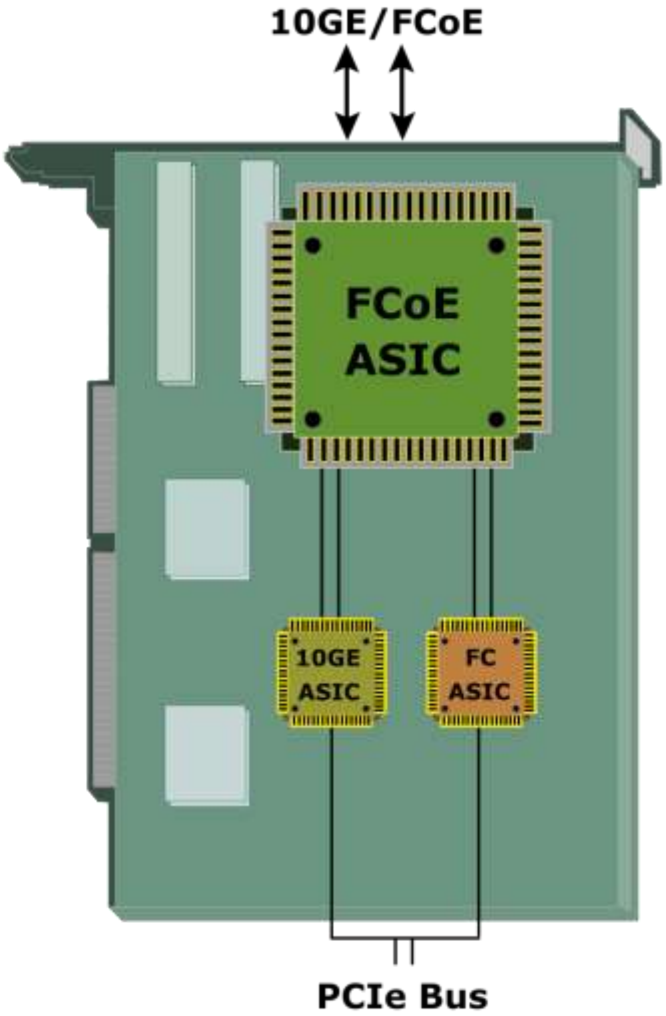


Components of an FCoE Network

- Converged Network Adapter (CNA)
- Cable
- FCoE switch

Converged Network Adapter (CNA)

- Provides functionality of both – a standard NIC and an FC HBA
  - Eliminates the need to deploy separate adapters and cables for FC and Ethernet communications
- Contains separate modules for 10 Gigabit Ethernet, FC, and FCoE ASICs
  - FCoE ASIC encapsulates FC frames into Ethernet frames



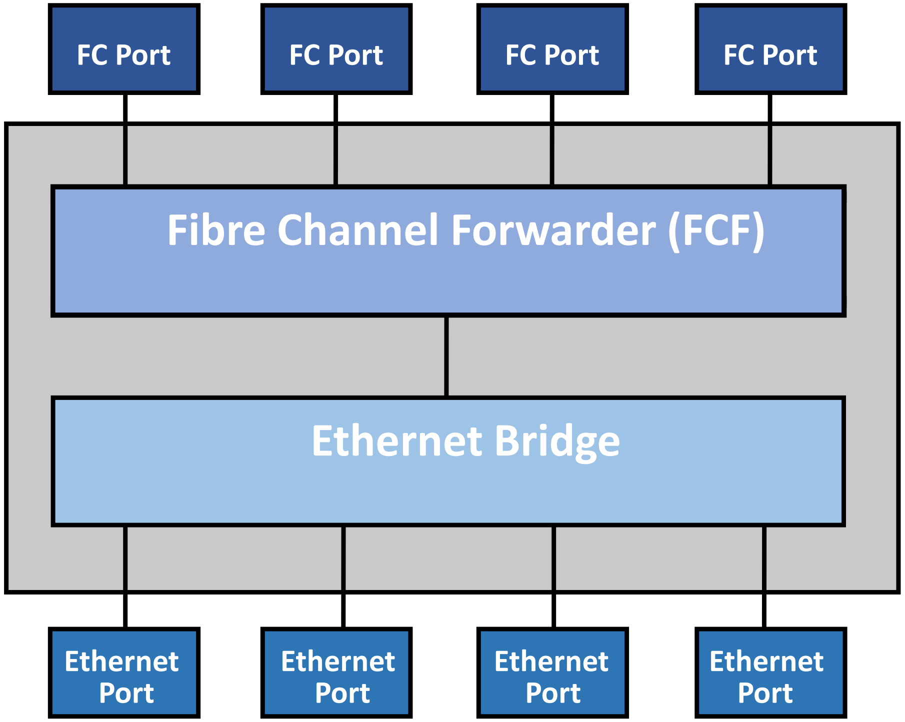
Cable

- Two options are available for FCoE cabling
  - Copper based Twinax cable
  - Standard fiber optical cable

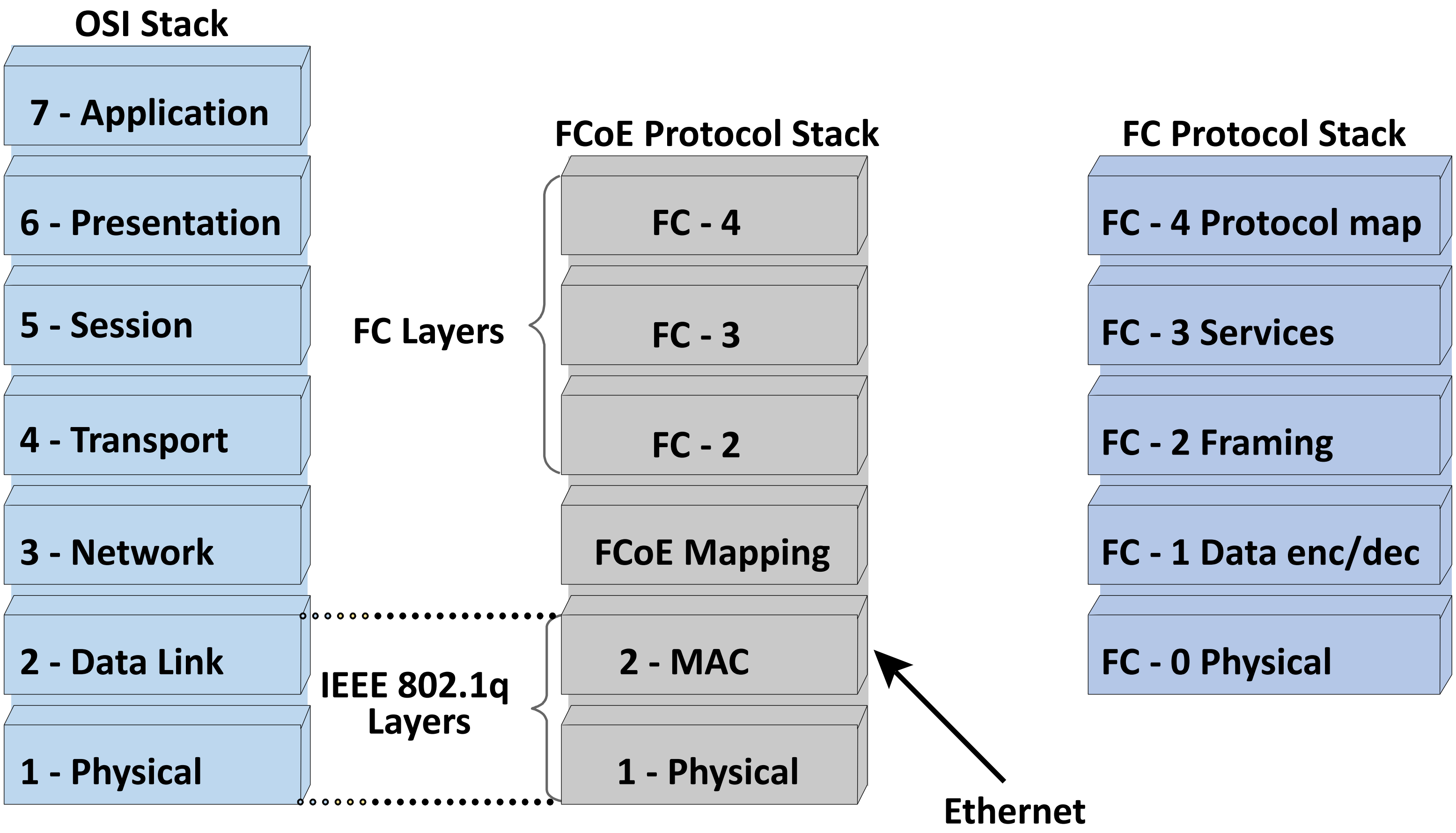
Twinax Cable	Fiber Optical Cable
Suitable for shorter distances (up to 10 meters)	Can run over longer distances
Requires less power and are less expensive than fiber optical cable	Relatively more expensive than Twinax cables
Uses Small Form Factor Pluggable Plus (SFP+) connector	Uses Small Form Factor Pluggable Plus (SFP+) connector

FCoE Switch

- Provides both Ethernet and FC switch functionalities
- Consists of FCF, Ethernet bridge, and set of CEE ports and FC ports (optional)
  - FCF encapsulates and de-encapsulates FC frames
- Forwards frames based on Ethertype



FCoE Frame Mapping

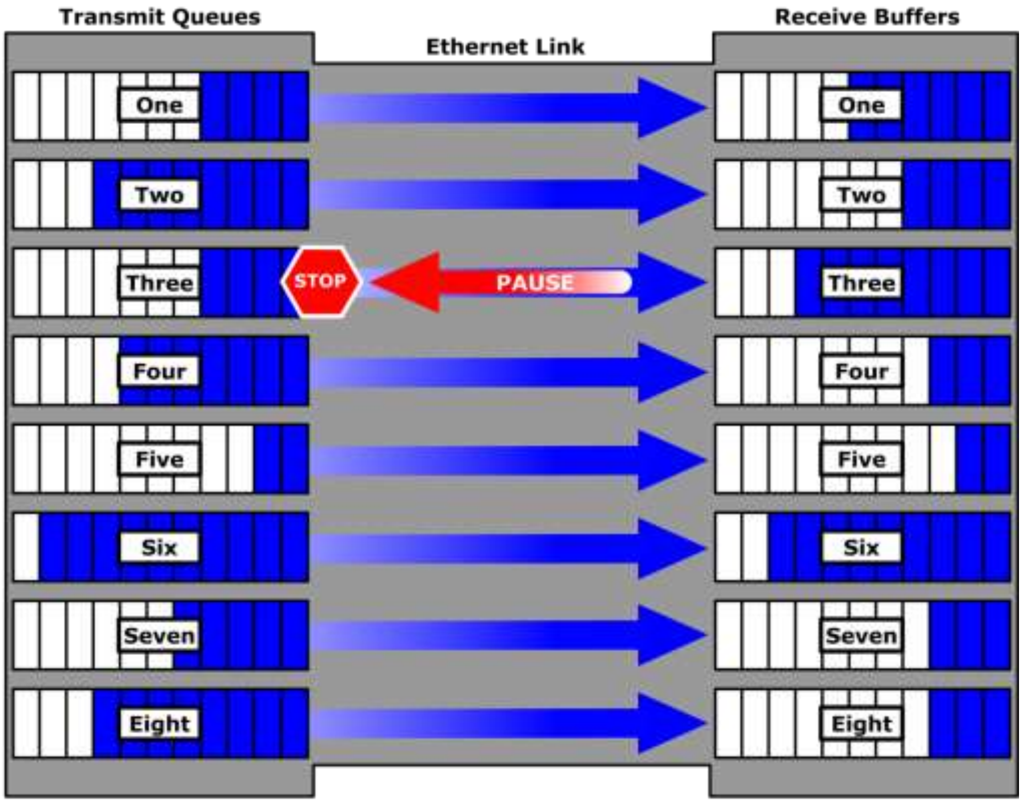


Converged Enhanced Ethernet

- Provides lossless Ethernet
- Lossless Ethernet requires following functionalities:
  - Priority-based flow control (PFC)
  - Enhanced transmission selection (ETS)
  - Congestion notification (CN)
  - Data center bridging exchange protocol(DCBX)

Priority-Based Flow Control (PFC)

- Creates eight virtual links on a single physical link
- Uses PAUSE capability of Ethernet for each virtual link
    - A virtual link can be paused and restarted independently
    - PAUSE mechanism is based on user priorities or classes of service

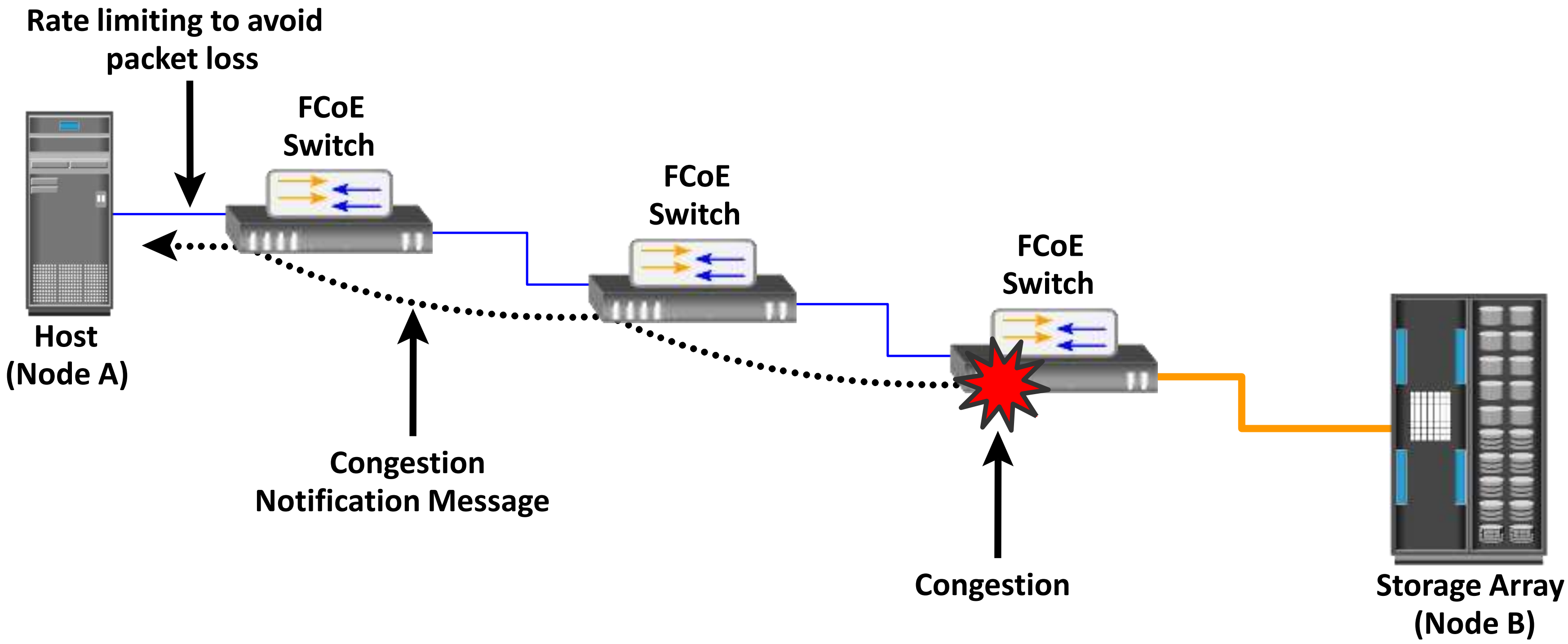


Enhanced Transmission Selection (ETS)

- Allocates bandwidth to different traffic classes such as LAN, SAN, and Inter Process Communication (IPC)
- Provides available bandwidth to other classes of traffic when a particular class of traffic does not use its allocated bandwidth

Congestion Notification (CN)

- Provides a mechanism for detecting congestion and notifying the source
  - Enables a switch to send a signal to other ports that need to stop or slow down their transmissions

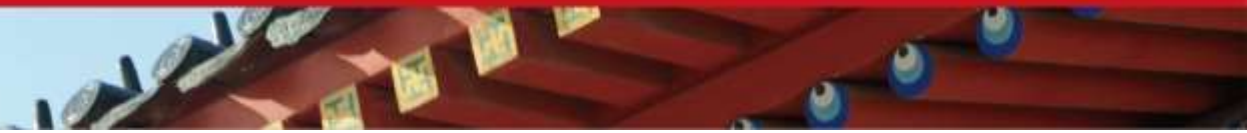


Data Center Bridging Exchange Protocol (DCBX)

- Enables Convergence Enhanced Ethernet (CEE) devices to convey and configure their features with other CEE devices in the network
  - Allows a switch to distribute configuration values to attached adapters
- Ensures consistent configuration across network

3

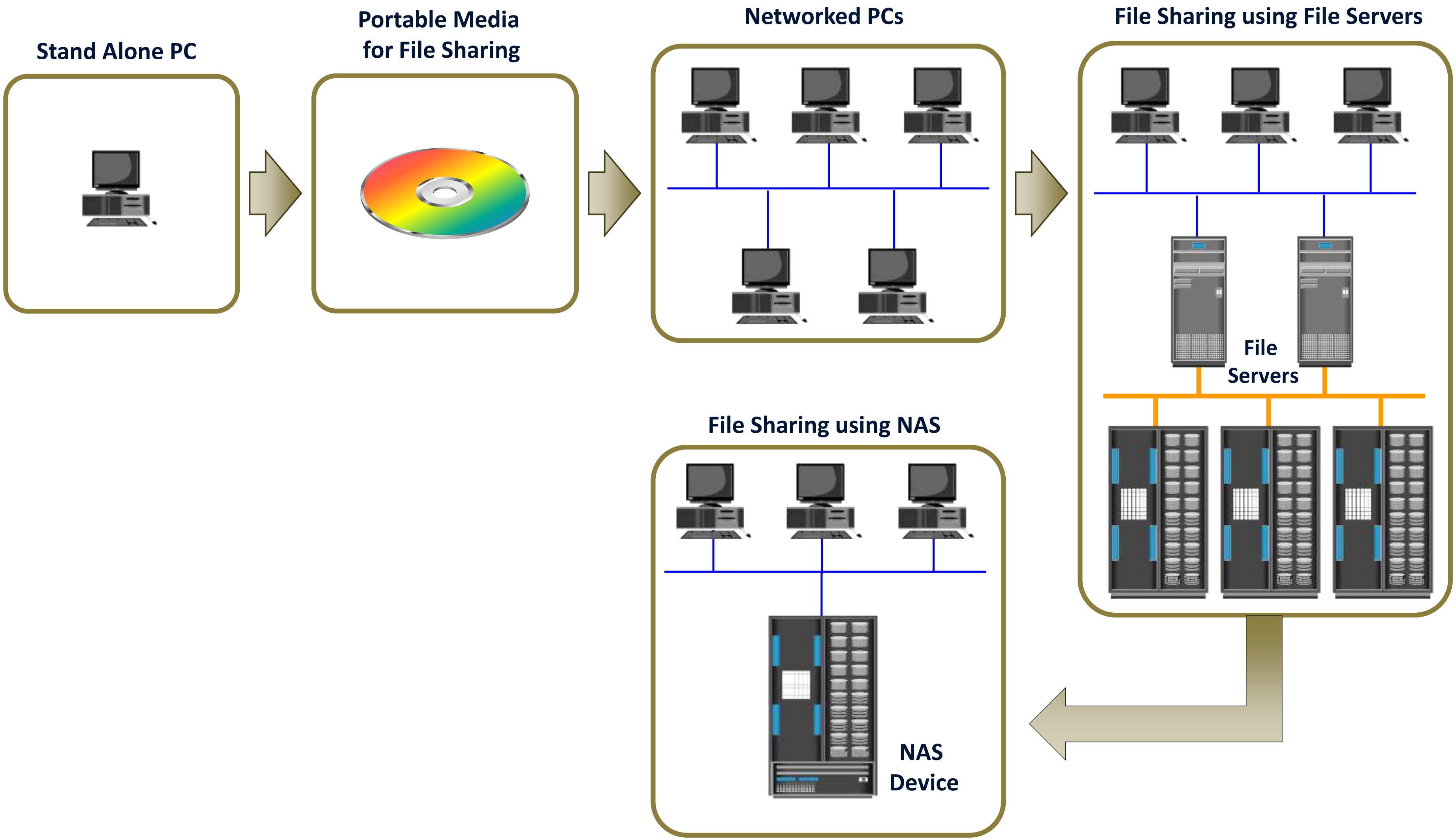
File Storage



File Sharing Environment

- File sharing enables users to share files with other users
- Creator or owner of a file determines the type of access to be given to other users
- File sharing environment ensures data integrity when multiple users access a shared file at the same time
- Examples of file sharing methods:
  - File Transfer Protocol (FTP)
  - Distributed File System (DFS)
  - Network File System (NFS) and Common Internet File System (CIFS)
  - Peer-to-Peer (P2P)

File Sharing Technology Evolution



Stand Alone PC

Portable Media for File Sharing

Networked PCs

File Sharing using File Servers

File Servers

File Sharing using NAS

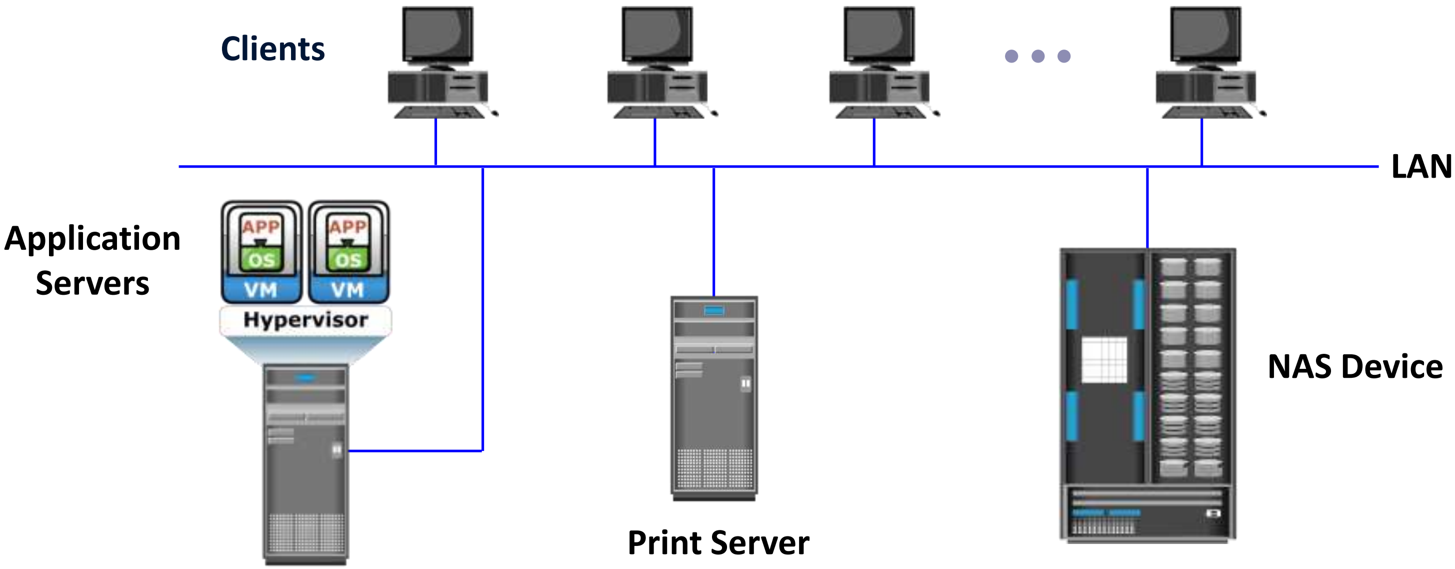
NAS Device

What is NAS (Network-Attached Storage)?

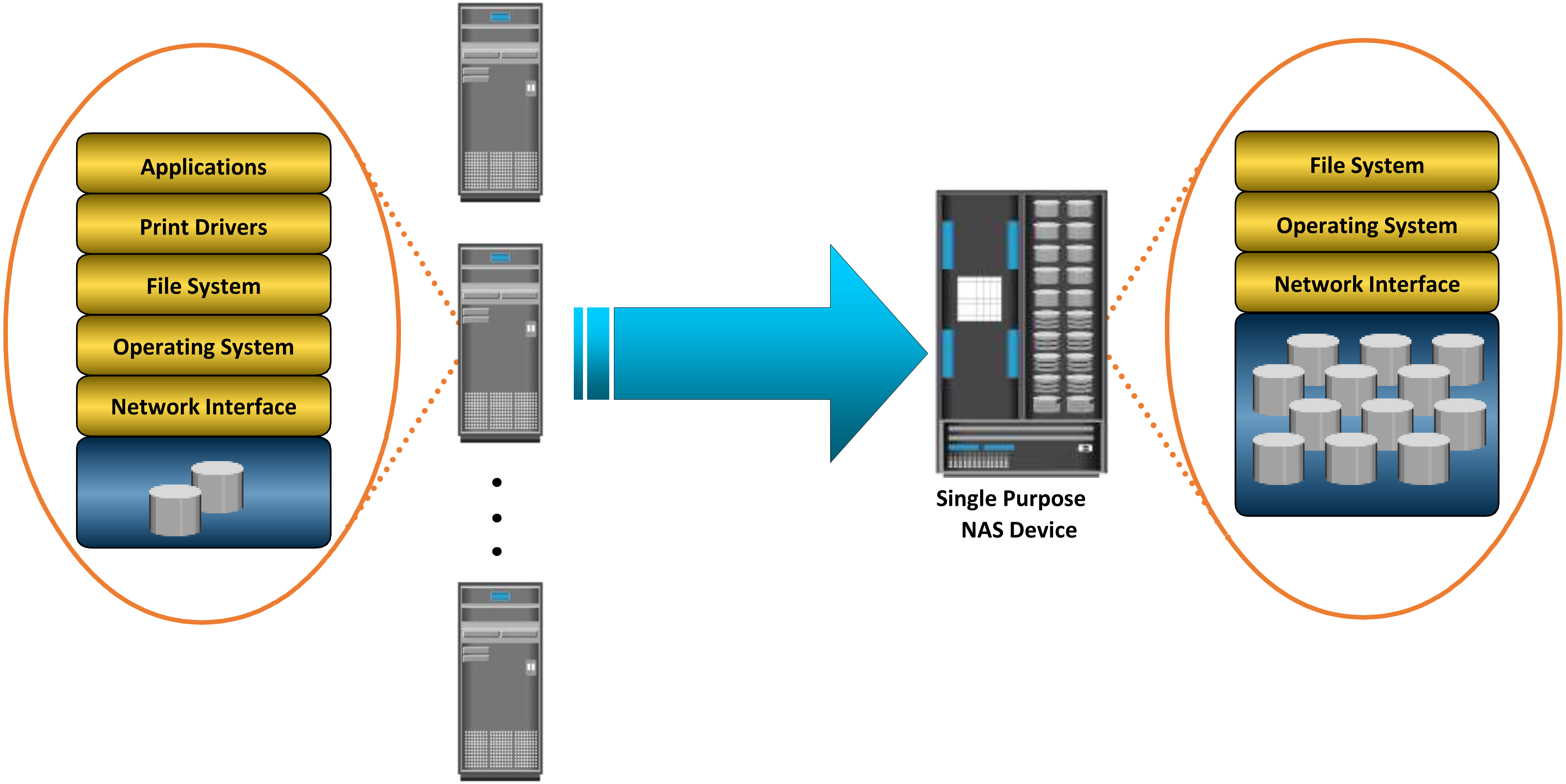
NAS

It is an IP-based, dedicated, high-performance file sharing and storage device.

- Enables NAS clients to share files over IP network
- Uses specialized operating system that is optimized for file I/O
- Enables both UNIX and Windows users to share data



General Purpose Servers Vs. NAS Devices



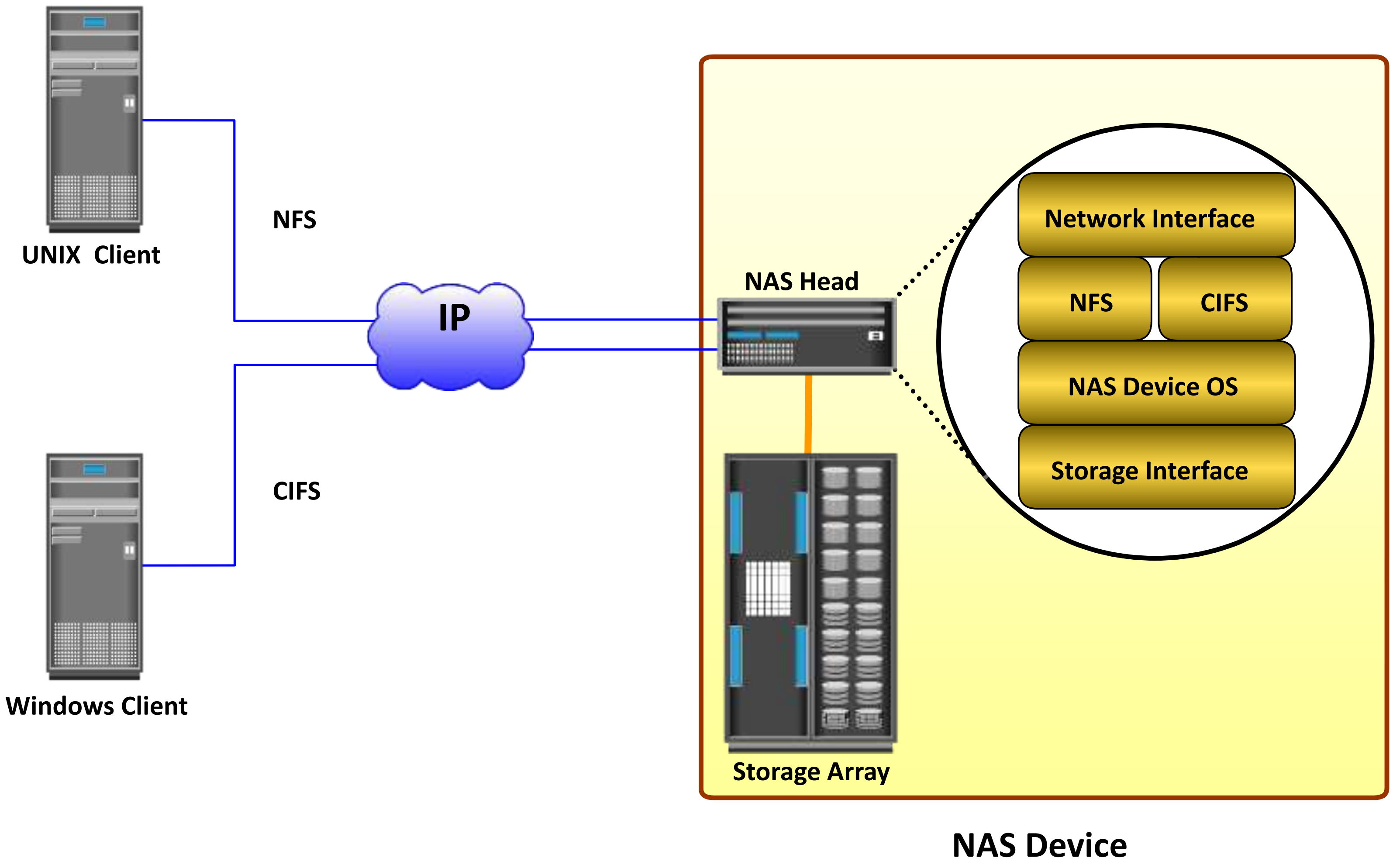
Single Purpose
NAS Device

General Purpose Servers
(Windows or UNIX)

Benefits of NAS

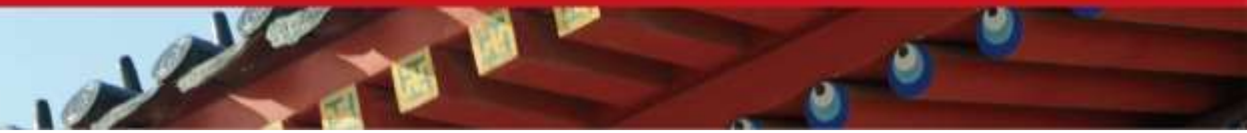
- Improved efficiency
- Improved flexibility
- Centralized storage
- Simplified management
- Scalability
- High availability – through native clustering and replication
- Security – authentication, authorization, and file locking in conjunction with industry-standard security
- Low cost
- Ease of deployment

Components of NAS



NAS File Sharing Protocols

- Two common NAS file sharing protocols are:
  - Common Internet File System (CIFS)
  - Network File System (NFS)



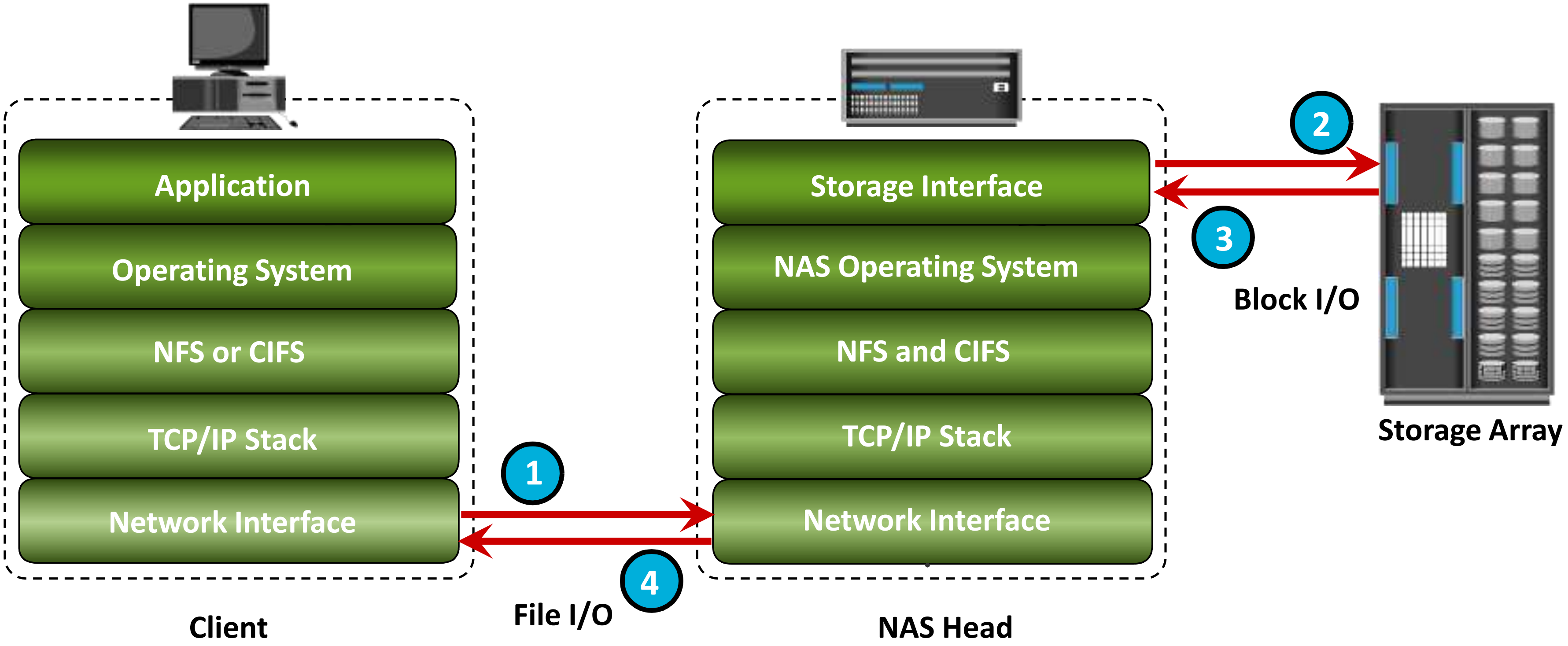
Common Internet File System (CIFS)

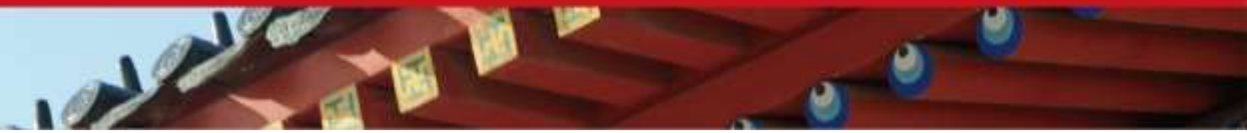
- Client-server application protocol
    - An open variation of the Server Message Block (SMB) protocol
- Enables clients to access files that are on a server over TCP/IP
- Stateful Protocol
    - Maintains connection information regarding every connected client
    - Can automatically restore connections and reopen files that were open prior to interruption

Network File System (NFS)

- Client-server application protocol
- Enables clients to access files that are on a server
- Uses Remote Procedure Call (RPC) mechanism to provide access to remote file system
- Currently, three versions of NFS are in use:
  - NFS v2 is stateless and uses UDP as transport layer protocol
  - NFS v3 is stateless and uses UDP or optionally TCP as transport layer protocol
  - NFS v4 is stateful and uses TCP as transport layer protocol

NAS I/O Operation

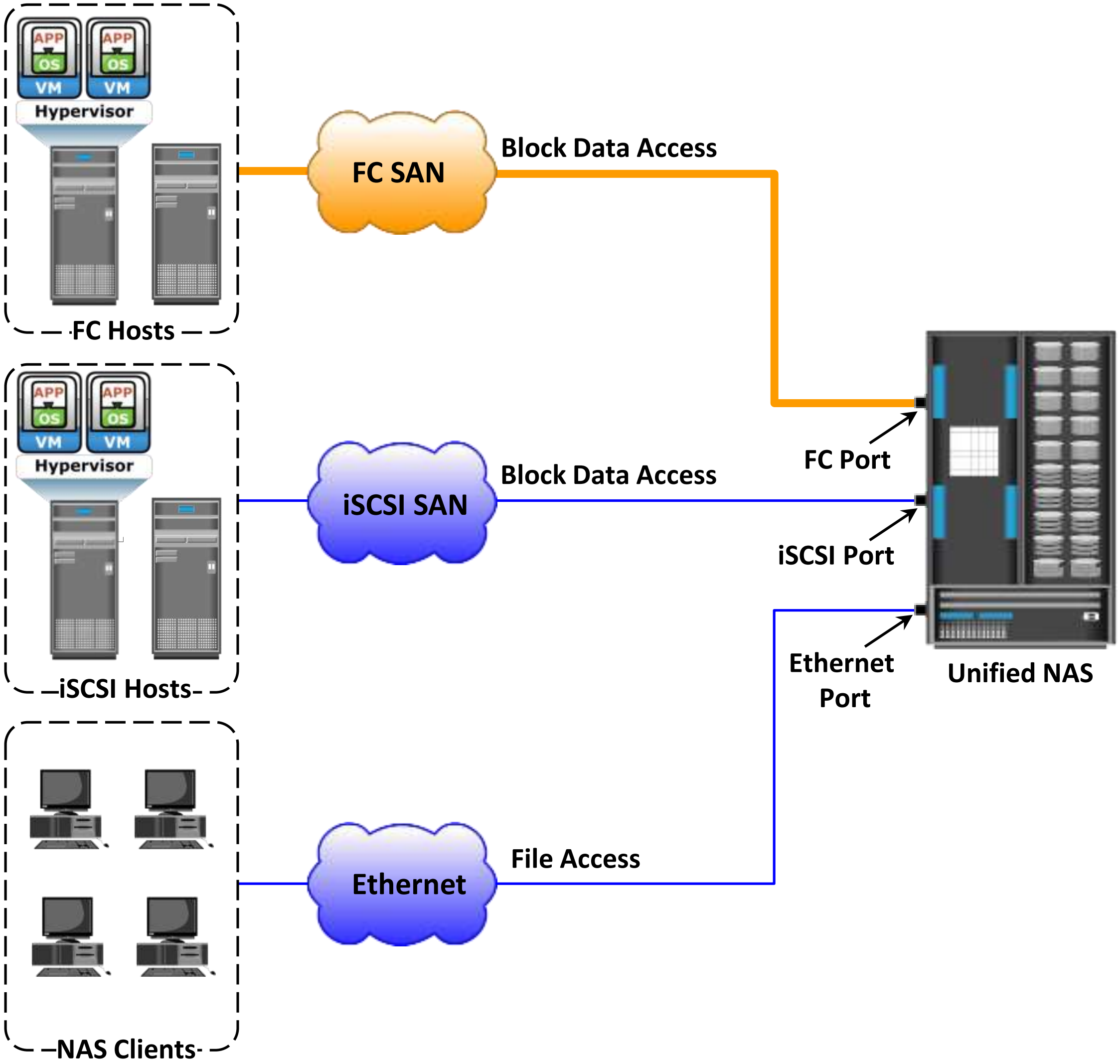




NAS Implementation – Unified NAS

- Consolidates NAS-based (file-level) and SAN-based (block-level) access on a single storage platform
- Supports both CIFS and NFS protocols for file access and iSCSI and FC protocols for block level access
- Provides unified management for both NAS head and storage

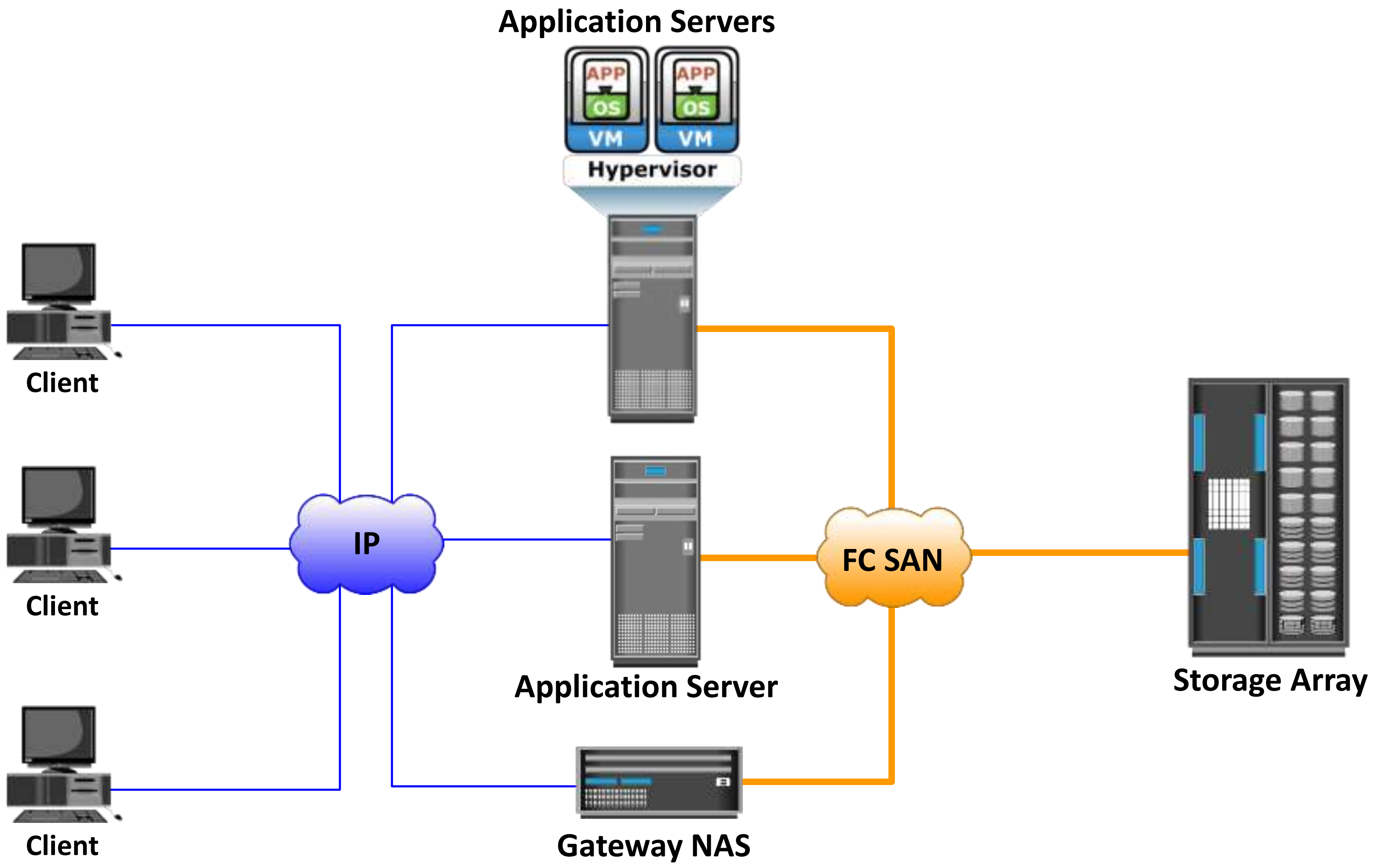
Unified NAS Connectivity



NAS Implementation – Gateway NAS

- Uses external and independently-managed storage
  - NAS heads access SAN-attached or direct-attached storage arrays
- NAS heads share storage with other application servers that perform block I/O
- Requires separate management of NAS head and storage

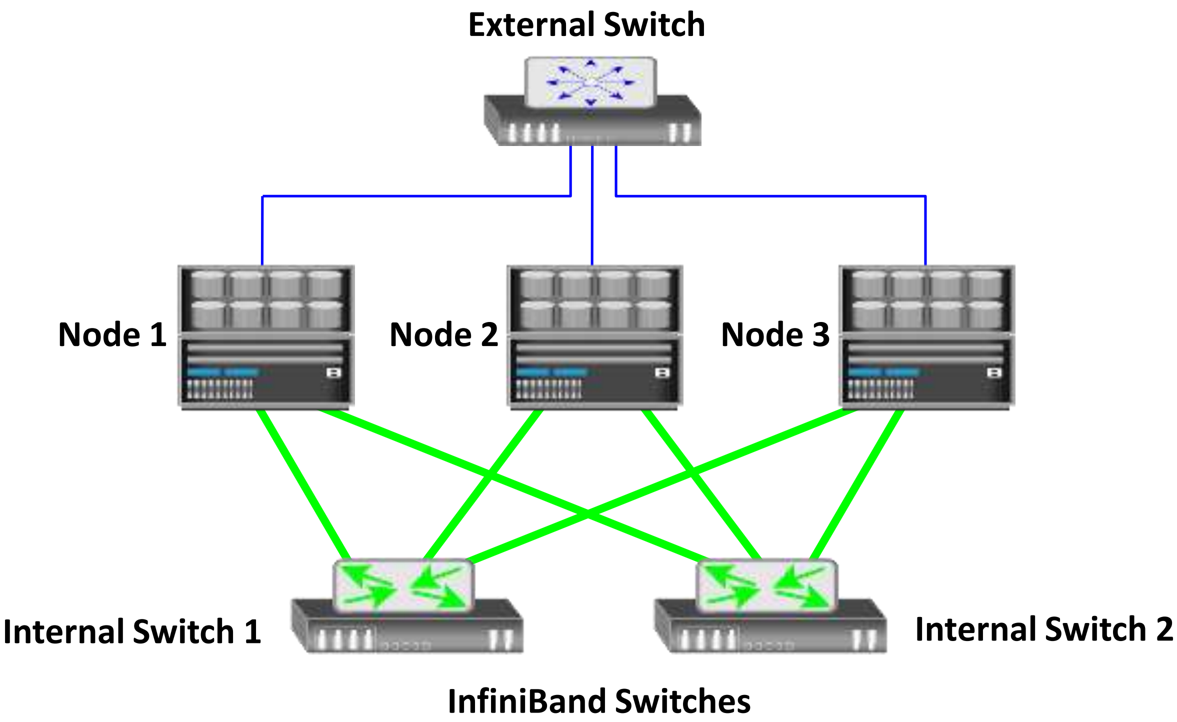
Gateway NAS Connectivity



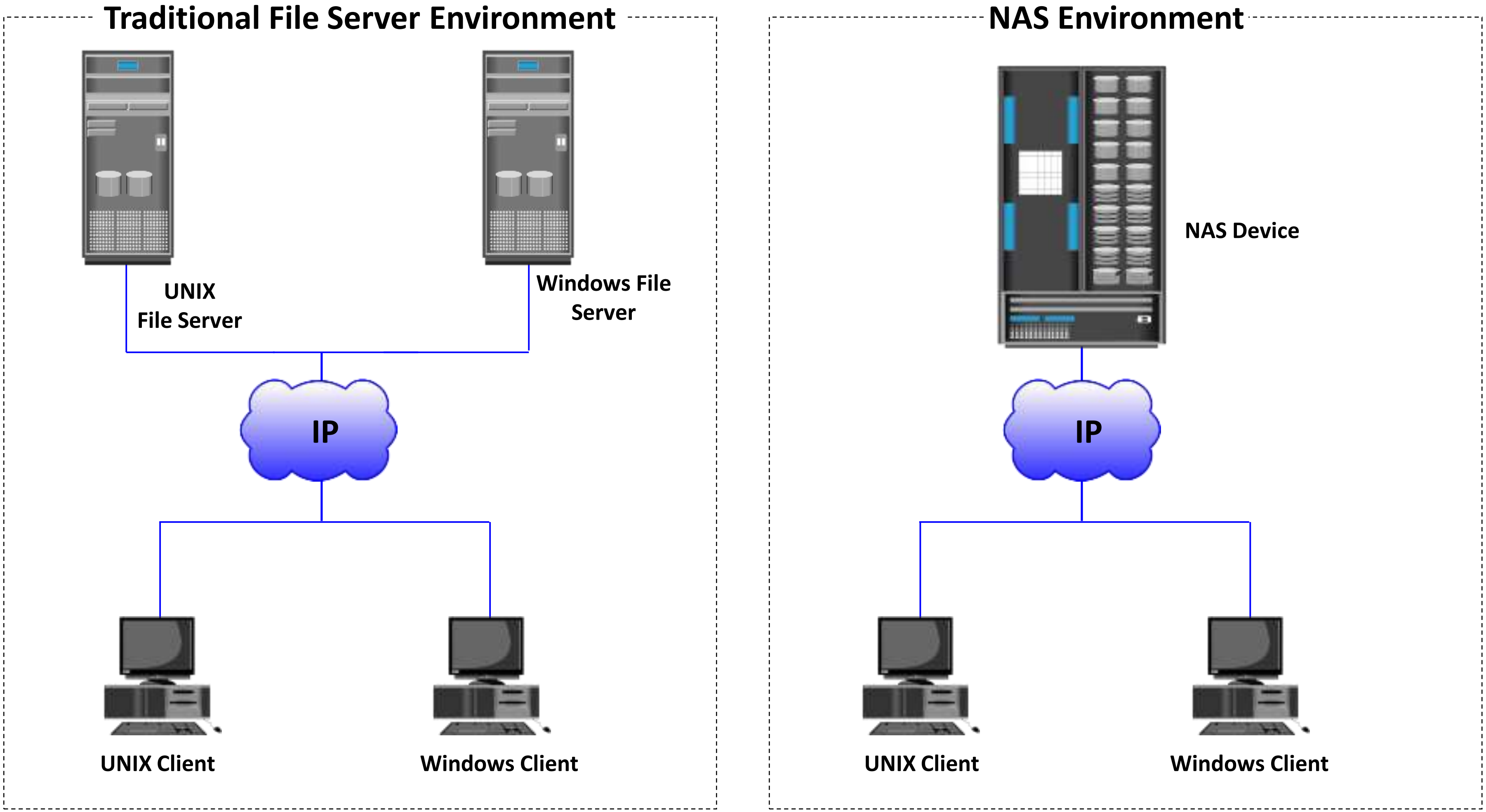
NAS Implementation – Scale-out NAS

- Pools multiple nodes together in a cluster that works as a single NAS device
  - Pool is managed centrally
- Scales performance and/or capacity with addition of nodes to the pool non-disruptively
- Creates a single file system that runs on all nodes in the cluster
  - Clients, connected to any node, can access entire file system
  - File system grows dynamically as nodes are added
- Stripes data across all nodes in a pool along with mirror or parity protection

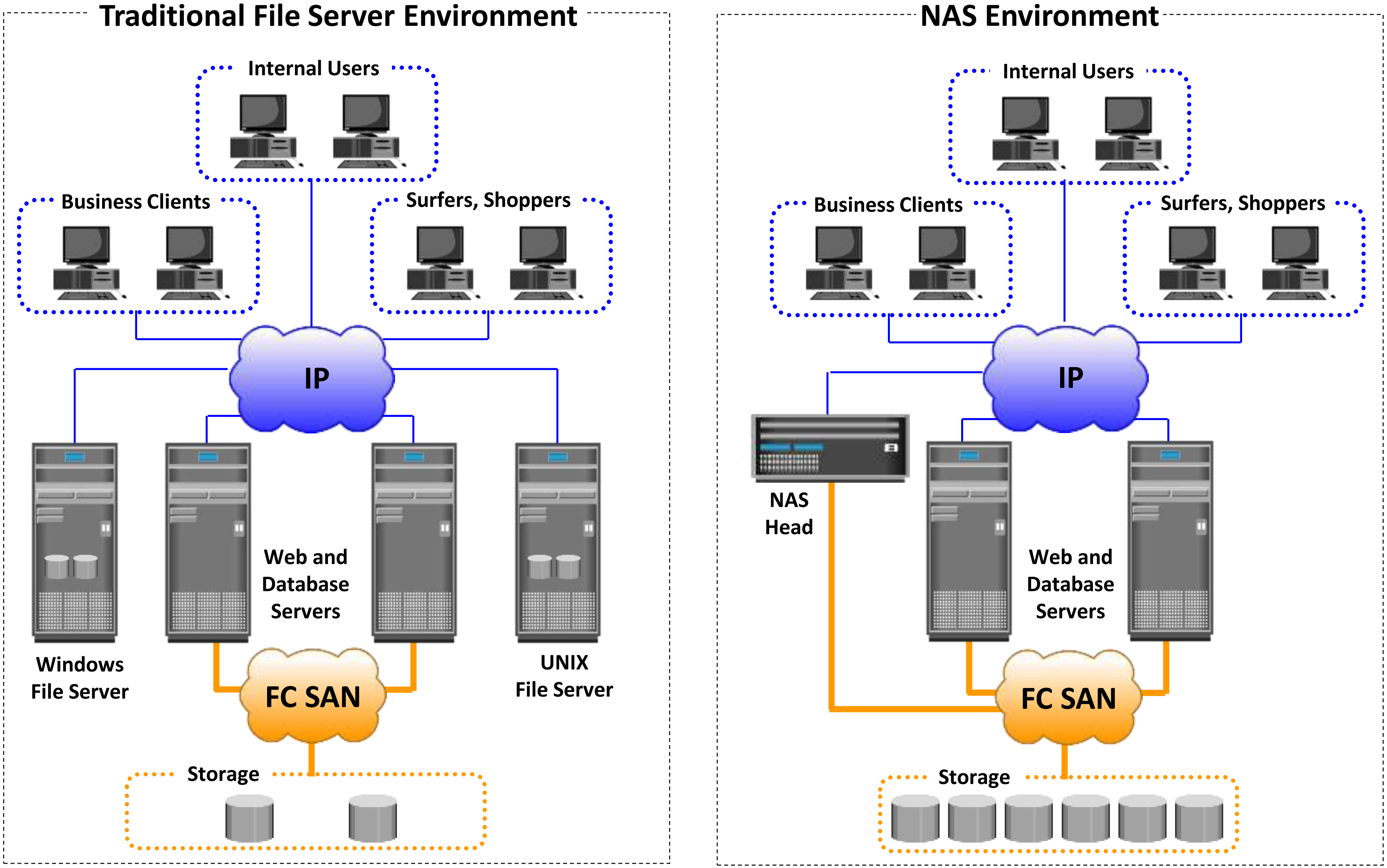
Scale-out NAS Connectivity



NAS Use Case 1 – Server Consolidation with NAS



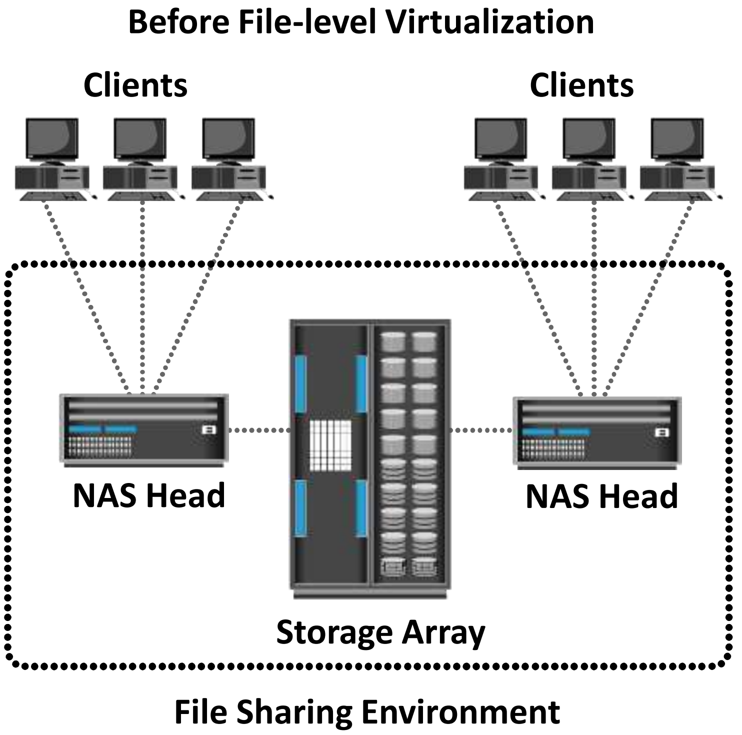
NAS Use Case 2 – Storage Consolidation with NAS

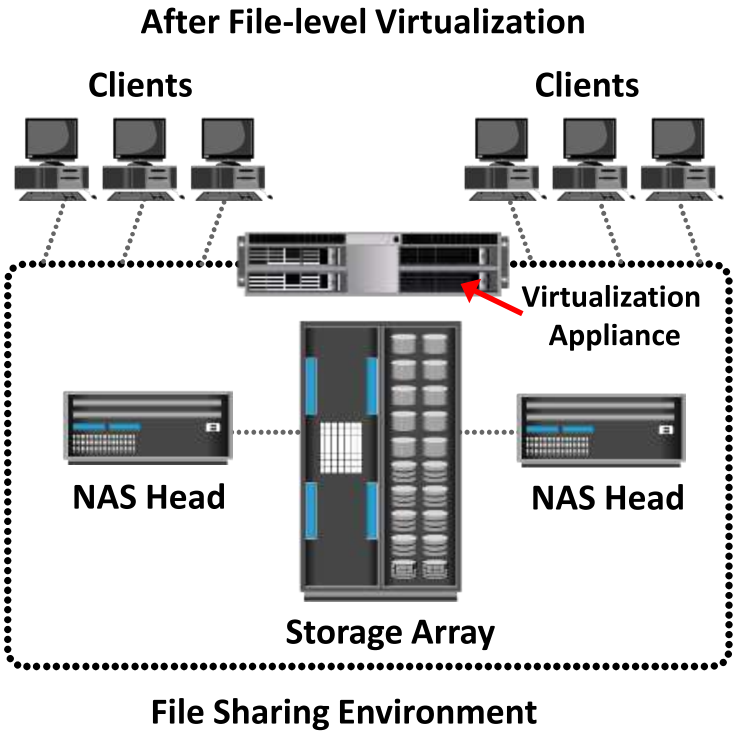


File-level Virtualization

- Eliminates dependency between data accessed at the file-level and the location where the files are physically stored
- Enables users to use a logical path, rather than a physical path, to access files
- Uses global namespace that maps logical path of file resources to their physical path
- Provides non-disruptive file mobility across file servers or NAS devices

Comparison: Before and After File-level Virtualization





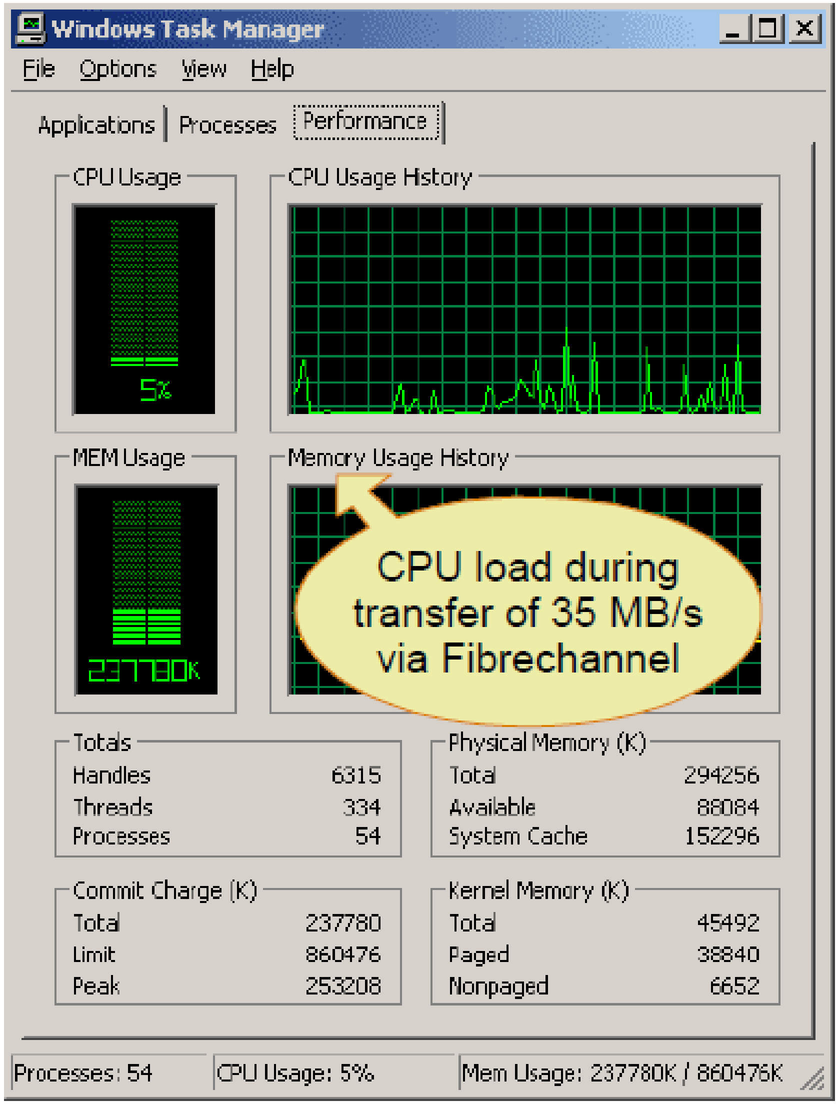
- Dependency between client access and file location
- Underutilized storage resources
- Downtime is caused by data migrations

- Break dependencies between client access and file location
- Storage utilization is optimized
- Non-disruptive migrations

Contents

4 SAN vs. NAS

FC vs TCP/IP (FC SAN vs. IP SAN)



Totals	
Handles	6315
Threads	334
Processes	54

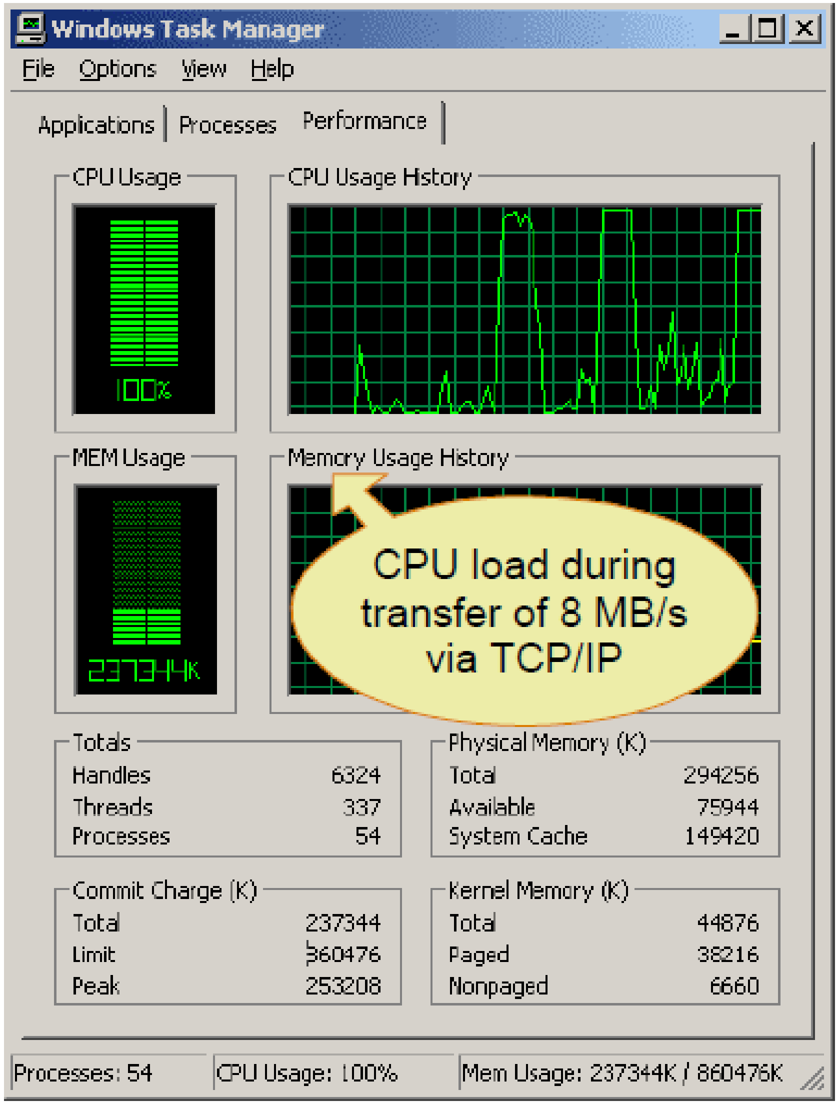
Physical Memory (K)	
Total	294256
Available	88084
System Cache	152296

Commit Charge (K)	
Total	237780
Limit	860476
Peak	253208

Kernel Memory (K)	
Total	45492
Paged	38840
Nonpaged	6652

Processes: 54 | CPU Usage: 5% | Mem Usage: 237780K / 860476K

CPU load during transfer of 35 MB/s via Fibrechannel



Totals	
Handles	6324
Threads	337
Processes	54

Physical Memory (K)	
Total	294256
Available	75944
System Cache	149420

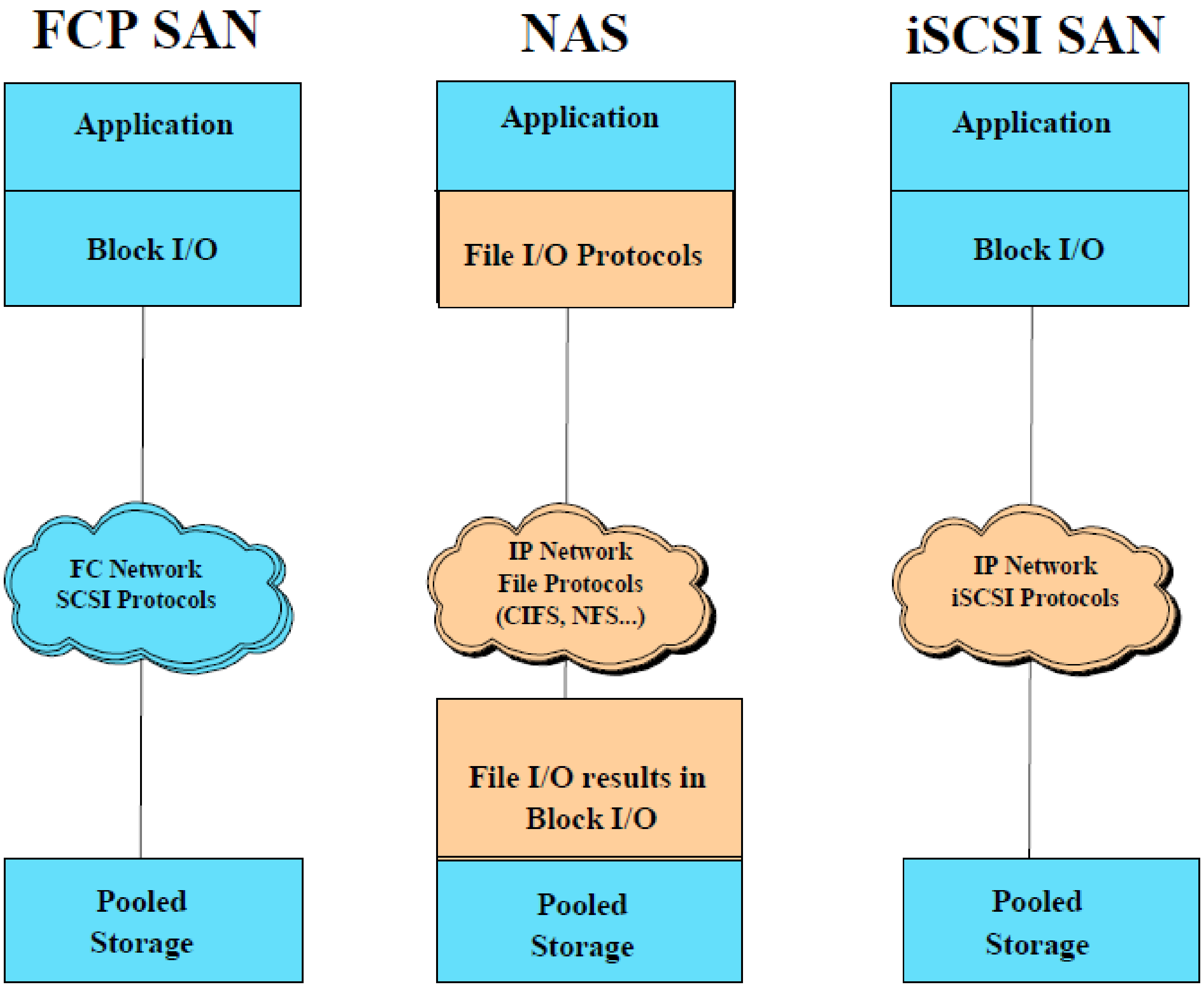
Commit Charge (K)	
Total	237344
Limit	360476
Peak	253208

Kernel Memory (K)	
Total	44876
Paged	38216
Nonpaged	6660

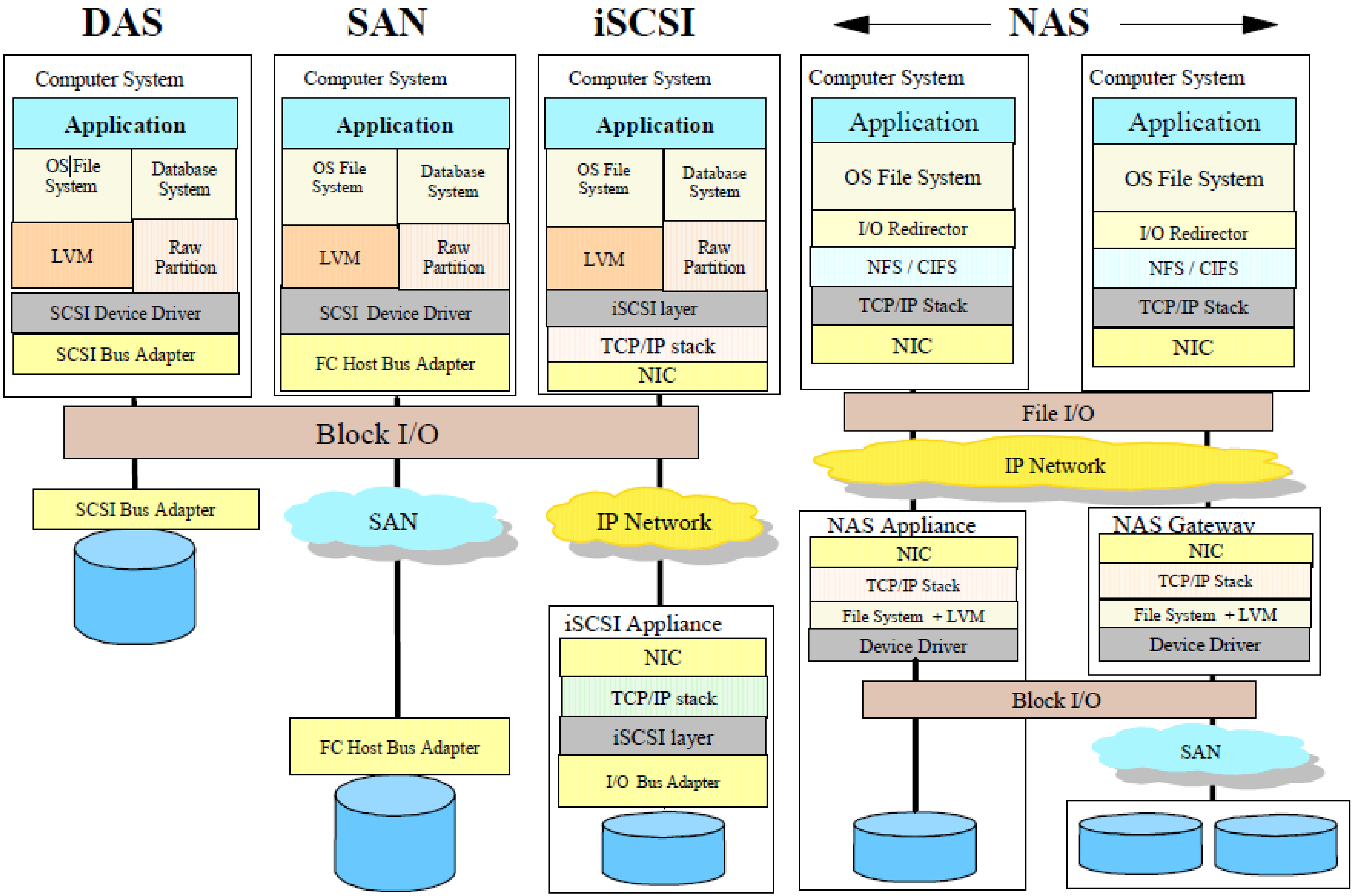
Processes: 54 | CPU Usage: 100% | Mem Usage: 237344K / 860476K

CPU load during transfer of 8 MB/s via TCP/IP

Application Protocol Support



Transporting Application Data



5 Cloud NAS & SAN

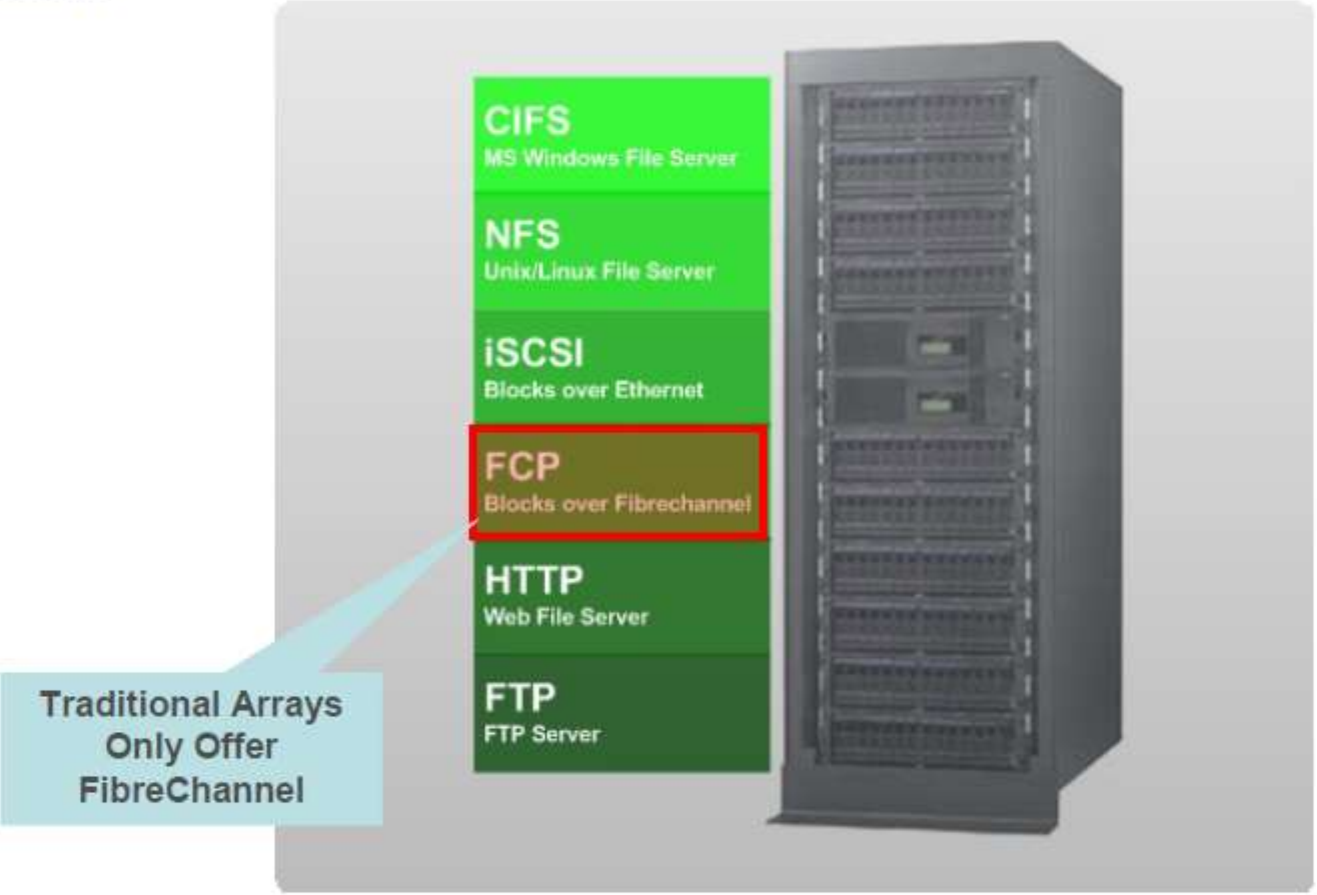
Cloud Storage Clients

- Characteristics
  - Hybrid: Web+Local (App)
  - RESTful HTTP
  - Disconnected Operations
  - Local Caching
  - Data Synchronization
  - Data as Objects with Metadata
- Examples
  - Mac & iPhone: Apple iDisk/iCloud
  - Windows: Microsoft Live Sync
  - Linux: Ubuntu One
  - Google Docs
  - Social apps

Cloud block Storage → Unified Storage

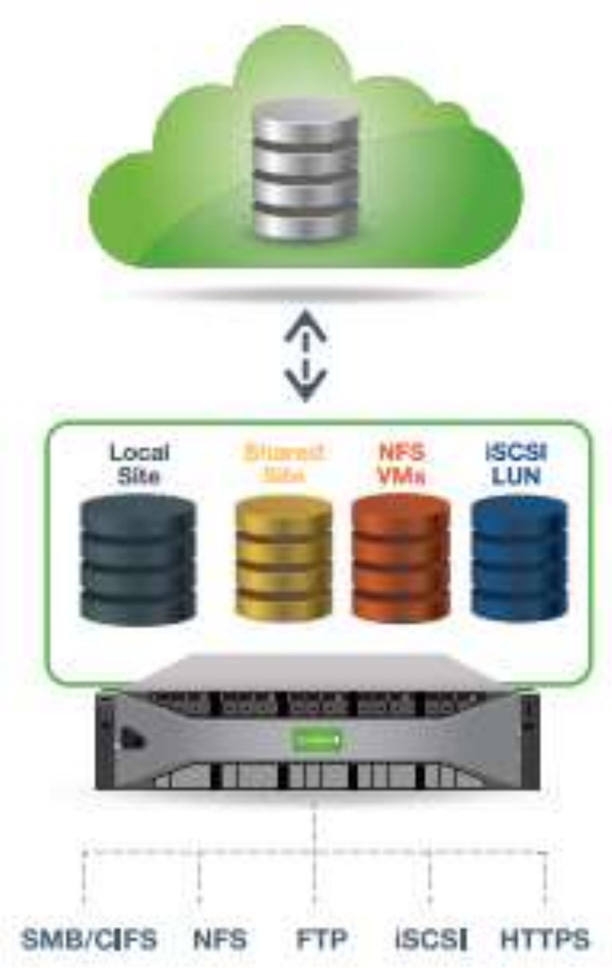
Multi-Protocol

- NAS
- IP SAN
- FC SAN
- FCoE



Traditional Arrays Only Offer FibreChannel

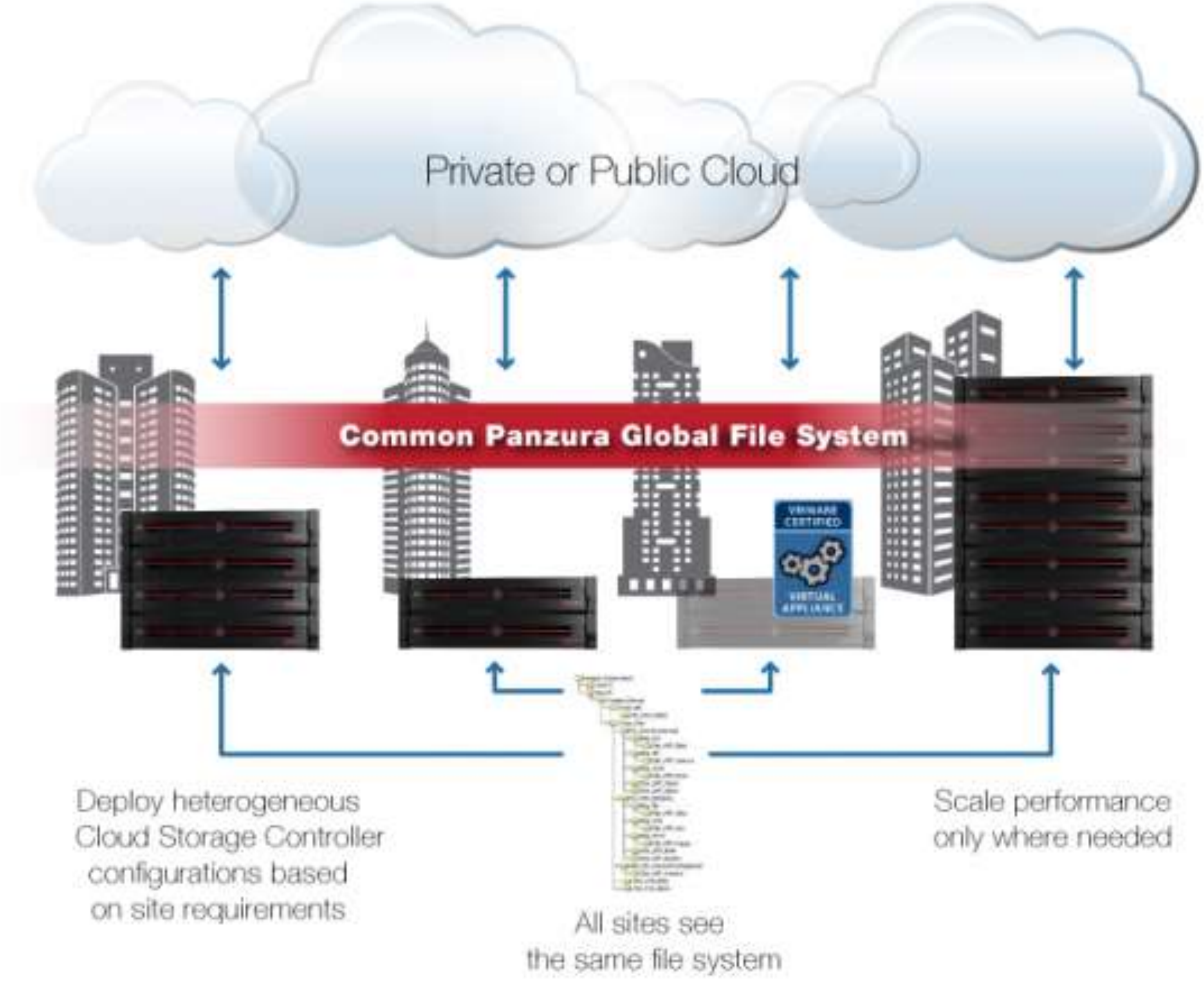
Cloud NAS Architecture



- Azure StorSimple
- Nasuni
- Panzura

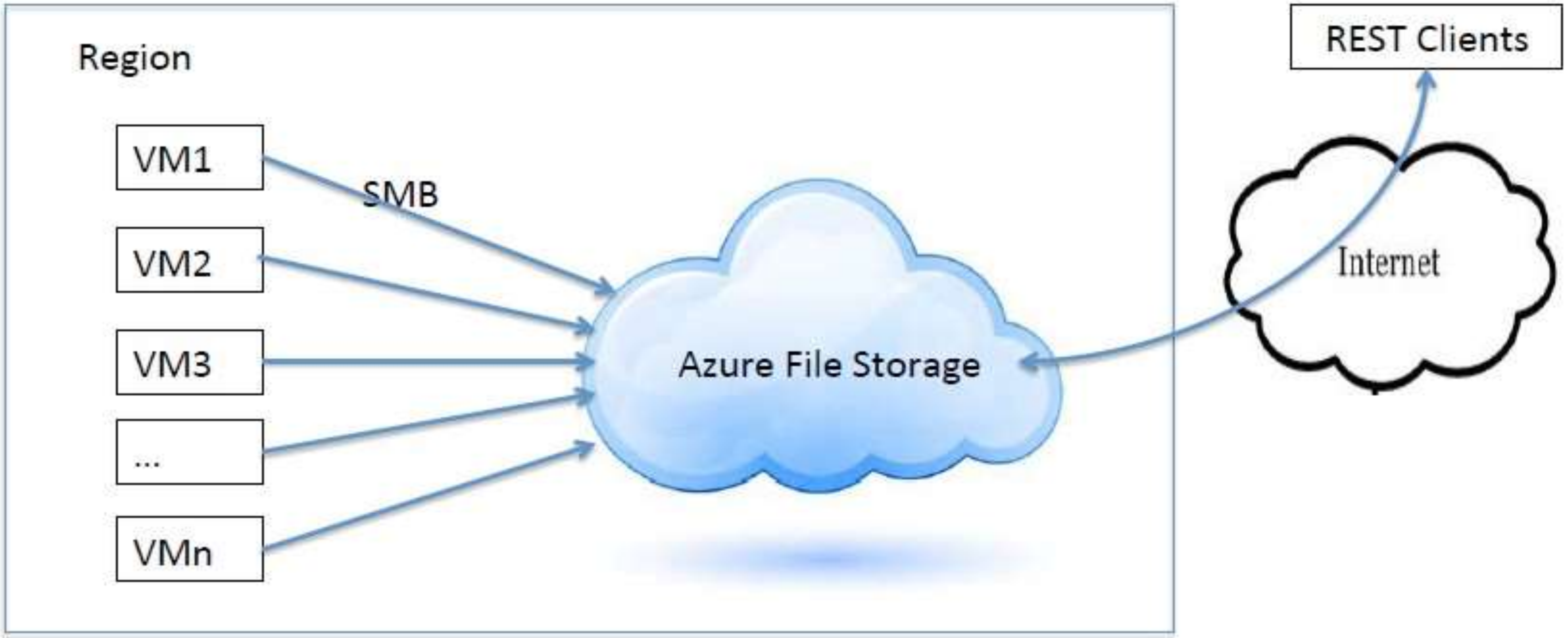
- Global management on namespace
- Lock management
- Privilege management
- Cache Optimization
- Deduplication

Cloud NAS Architecture → Distributed FS

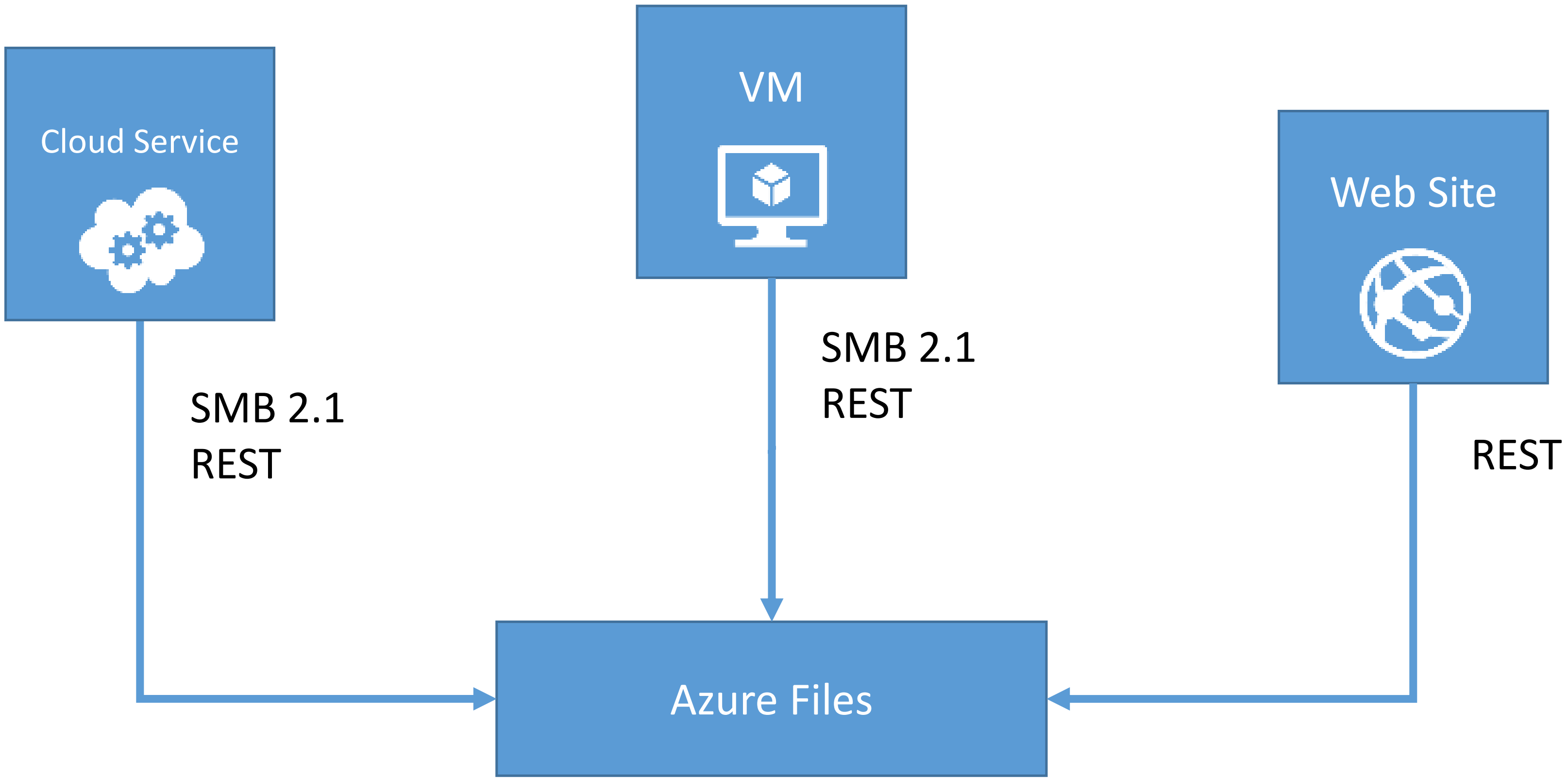

Private or Public Cloud
Common Panzura Global File System
VMWARE CERTIFIED
VIRTUAL APPLIANCE
Deploy heterogeneous Cloud Storage Controller configurations based on site requirements
All sites see the same file system
Scale performance only where needed

Microsoft Azure File Storage (1)

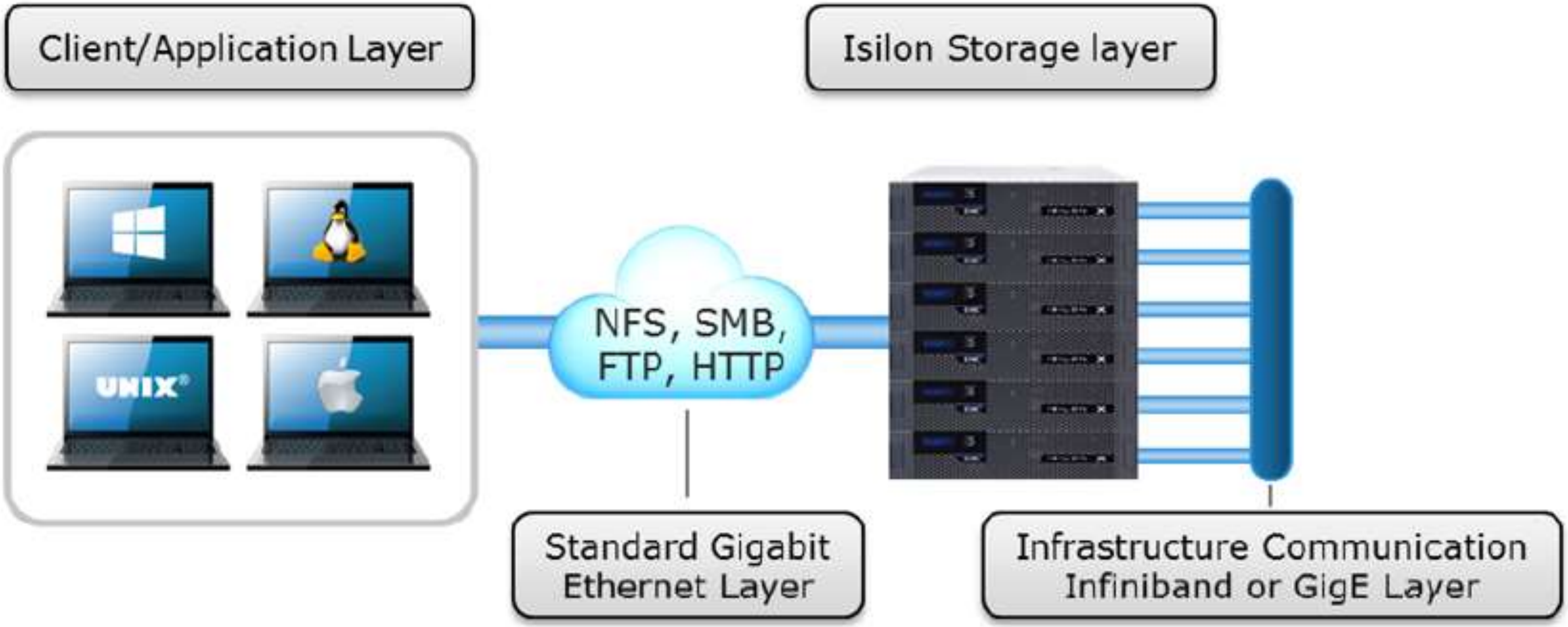
- Support SMB/RESTful Protocols
- File share for a VM region



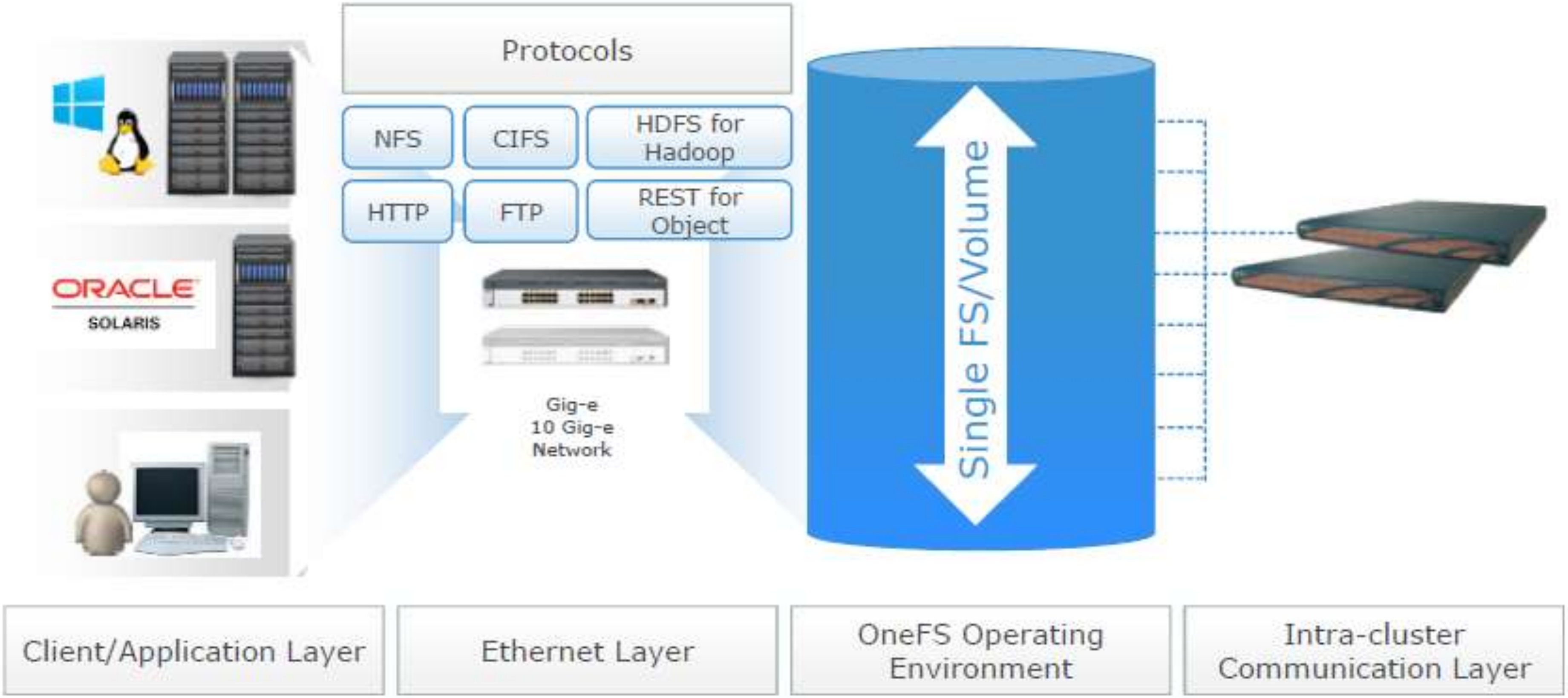
Microsoft Azure File Storage (2)



EMC Isilon


Client/Application Layer
Isilon Storage layer
UNIX®
NFS, SMB,
FTP, HTTP
Standard Gigabit
Ethernet Layer
Infrastructure Communication
Infiniband or GigE Layer

EMC Isilon Scale Out NAS



Client/Application Layer

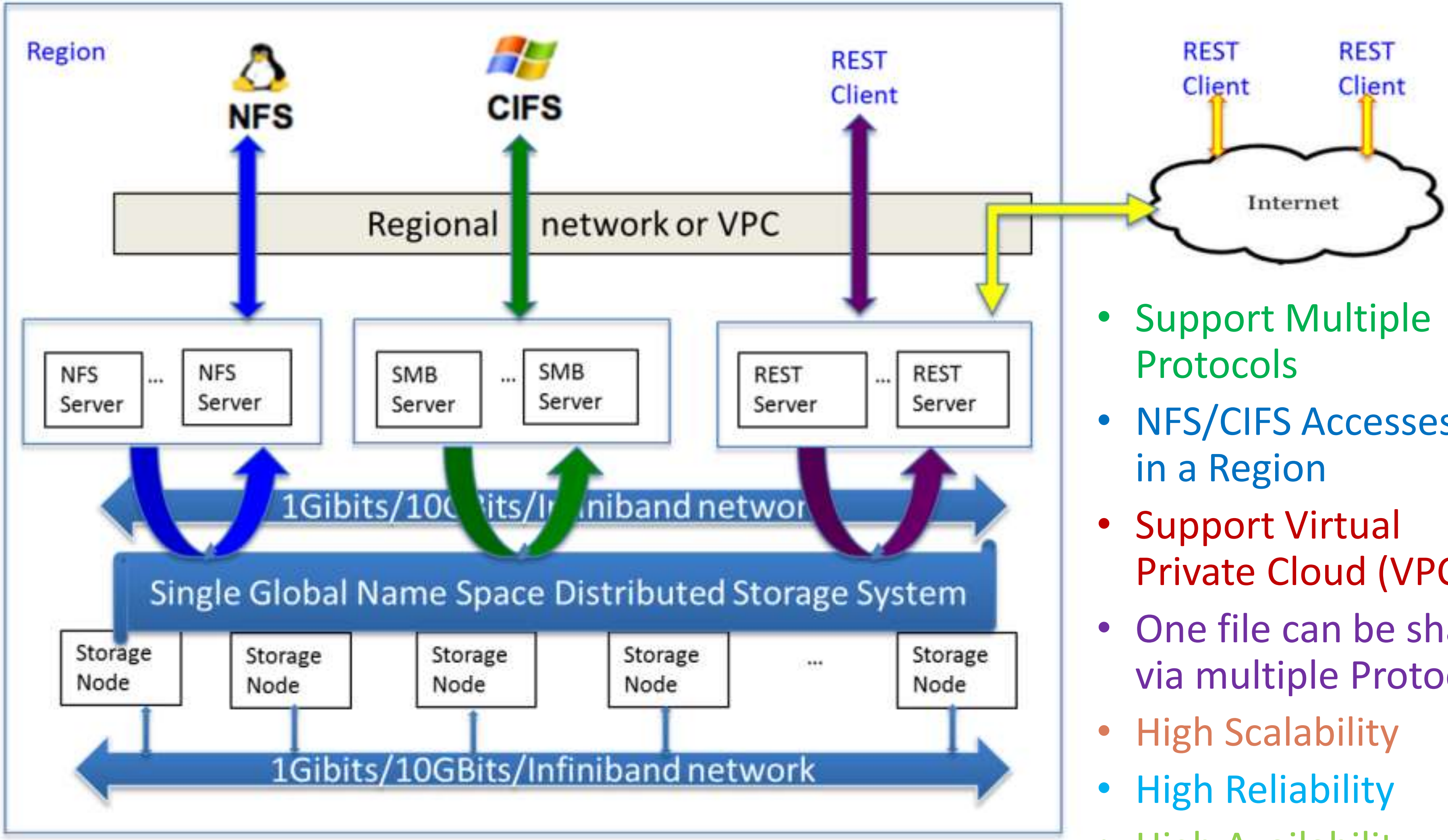
Ethernet Layer

OneFS Operating Environment

Intra-cluster Communication Layer

Ali NAS

阿里云
aliyun.com



- Support Multiple Protocols
- NFS/CIFS Accesses are in a Region
- Support Virtual Private Cloud (VPC)
- One file can be shared via multiple Protocols
- High Scalability
- High Reliability
- High Availability

Thank you!