

# Storage Networking Industry Association

## Clearing the Confusion: A Primer on Internet Protocol Storage Part II

Ahmad Zamer, Intel Corporation



# Presenters

- **Brice Clark – HP**
- **Gary Orenstein – Nishan**
- **Jeff Martin – SAN Valley**
- **Ahmad Zamer - Intel**



# IP Storage Technologies

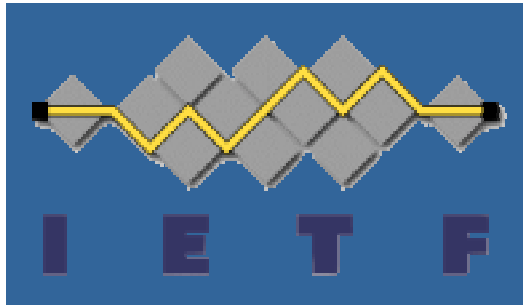


# Introduction

- If IP Storage is going to transform storage networking, what are the underlying technologies to accomplish this ?
- What are the differences ?
- What are the challenges ?
- Which one is right for my environment ?



# IP Storage Standards



Storage Networking Industry  
Association

- **IETF IP Storage (IPS) Working Group**
  - iSCSI
  - FCIP
  - iFCP
  - iSNS
- **Storage Networking Industry Association (SNIA)**
  - SNIA IP Storage Forum



# IP Storage Transports

- iSCSI
  - Internet Small Computer Systems Interface
- FCIP
  - Fibre Channel over TCP/IP
- iFCP
  - iFCP—Internet Fibre Channel Protocol
- iSNS
  - iSNS—Internet Storage Name Service

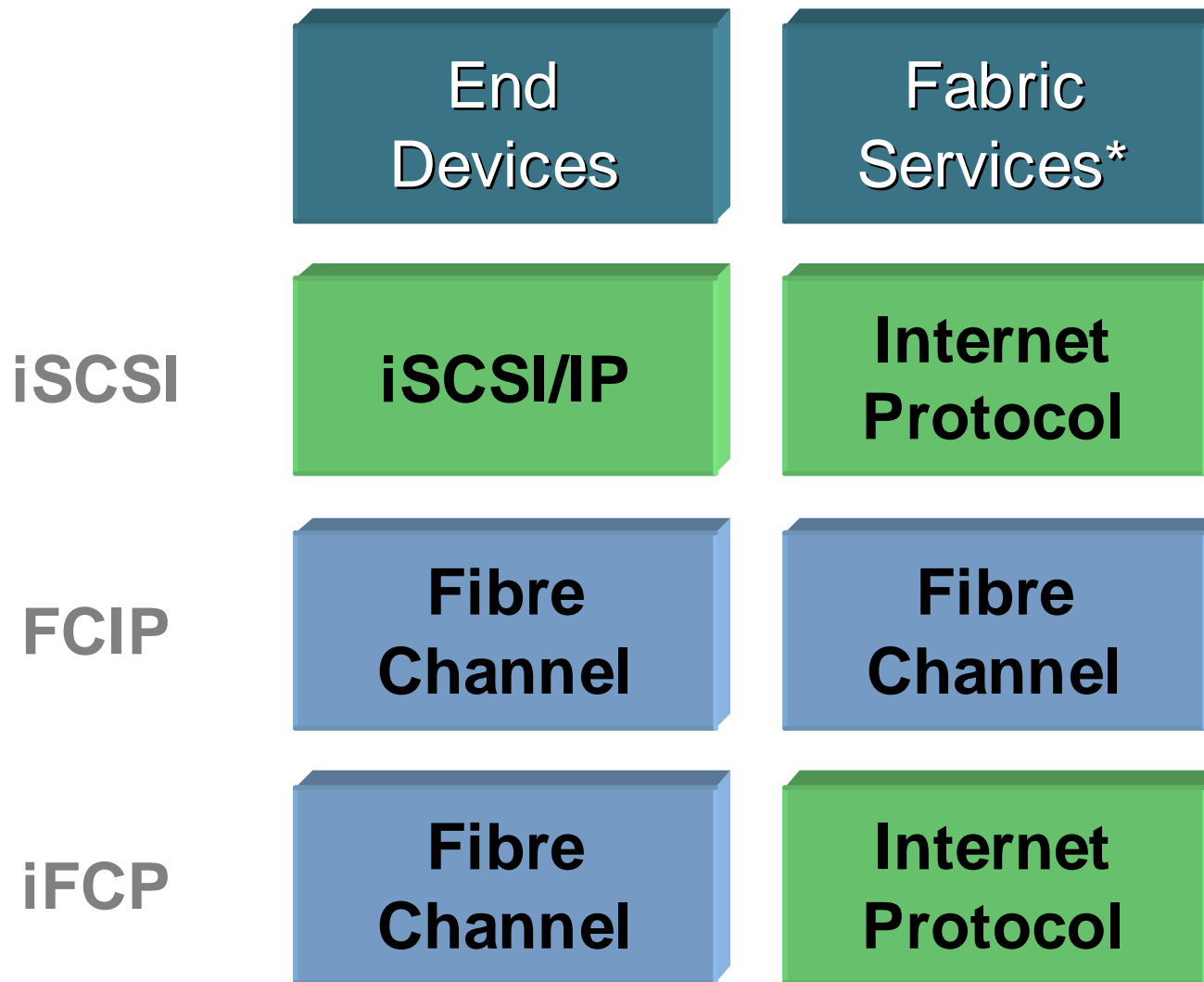


# What are the technologies? (iSCSI, iFCP, FCIP)

- iSCSI
  - iSCSI is a TCP/IP-based protocol for establishing and managing connections between IP-based storage devices, hosts and clients
- FCIP
  - FCIP is a TCP/IP-based tunneling protocol for connecting geographically distributed Fibre Channel SANs transparently to both FC and IP
- iFCP
  - iFCP is a TCP/IP-based protocol for interconnecting Fibre Channel storage devices or Fibre Channel SANs using an IP infrastructure in place of Fibre Channel switching and routing elements



# IP Storage: iSCSI, FCIP, iFCP



\* Fabric Services include routing, device discovery, management, authentication, inter-switch communication



# IP Storage Protocols: iSCSI, iFCP and FCIP

## iSCSI

## iFCP

## FCIP

Devices:

iSCSI/IP

Fibre Channel

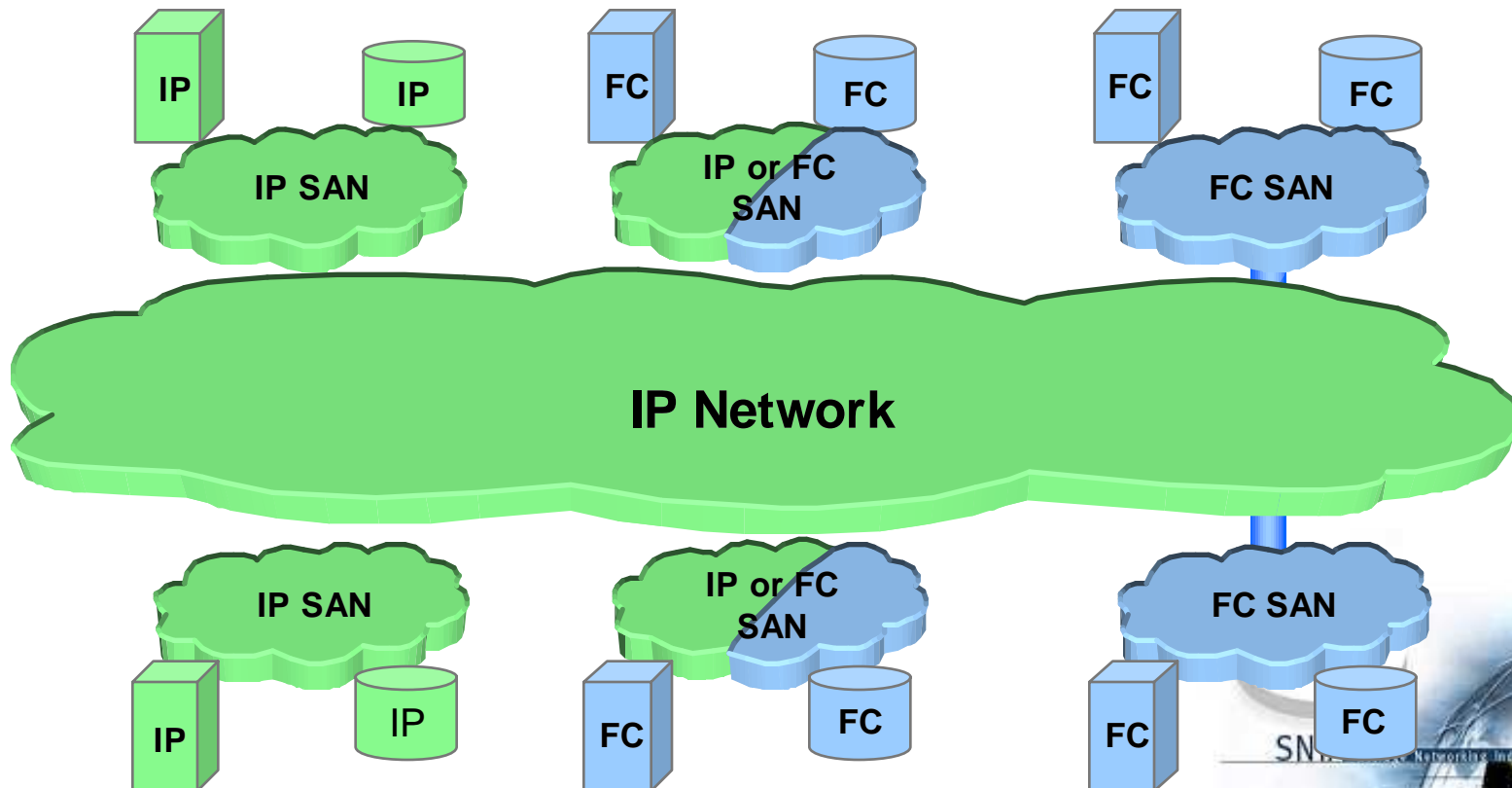
Fibre Channel

Fabric Services:

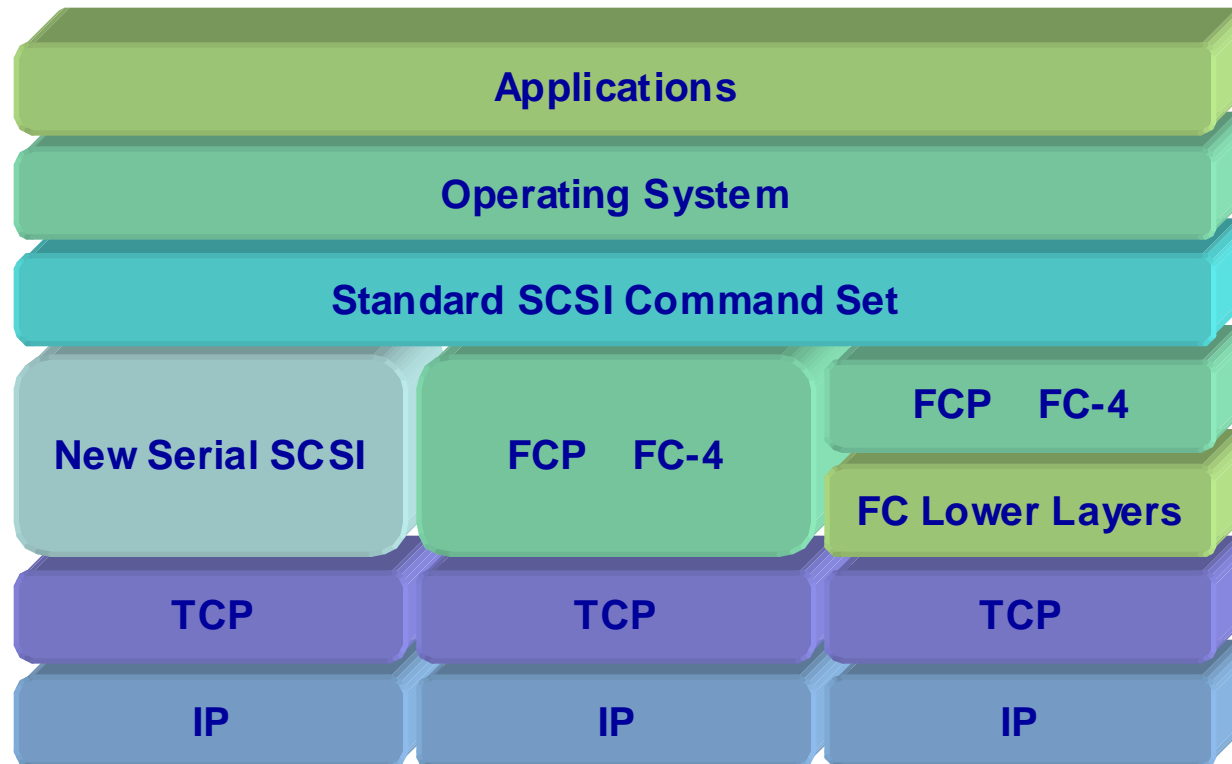
Internet Protocol

Internet Protocol

Fibre Channel



# iSCSI, iFCP and FCIP Protocol Stacks



iSCSI

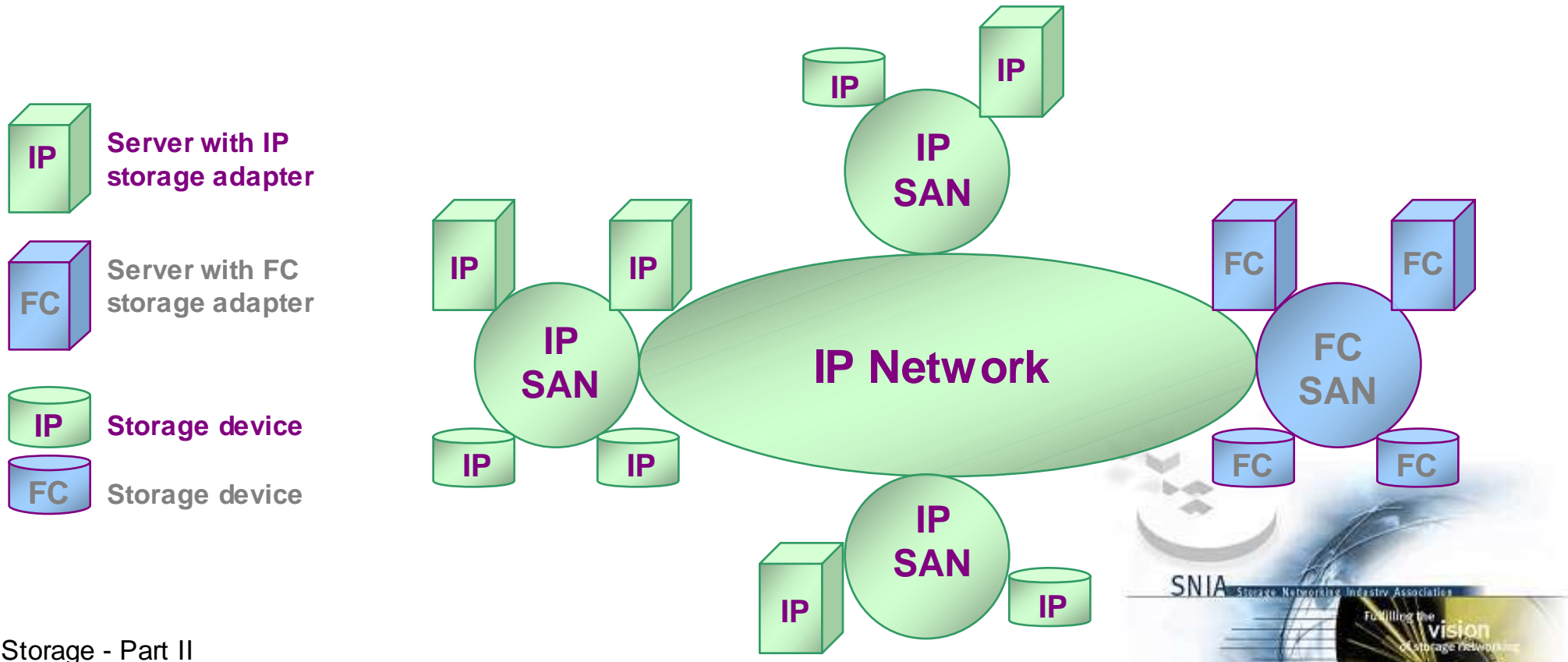
iFCP

FCIP



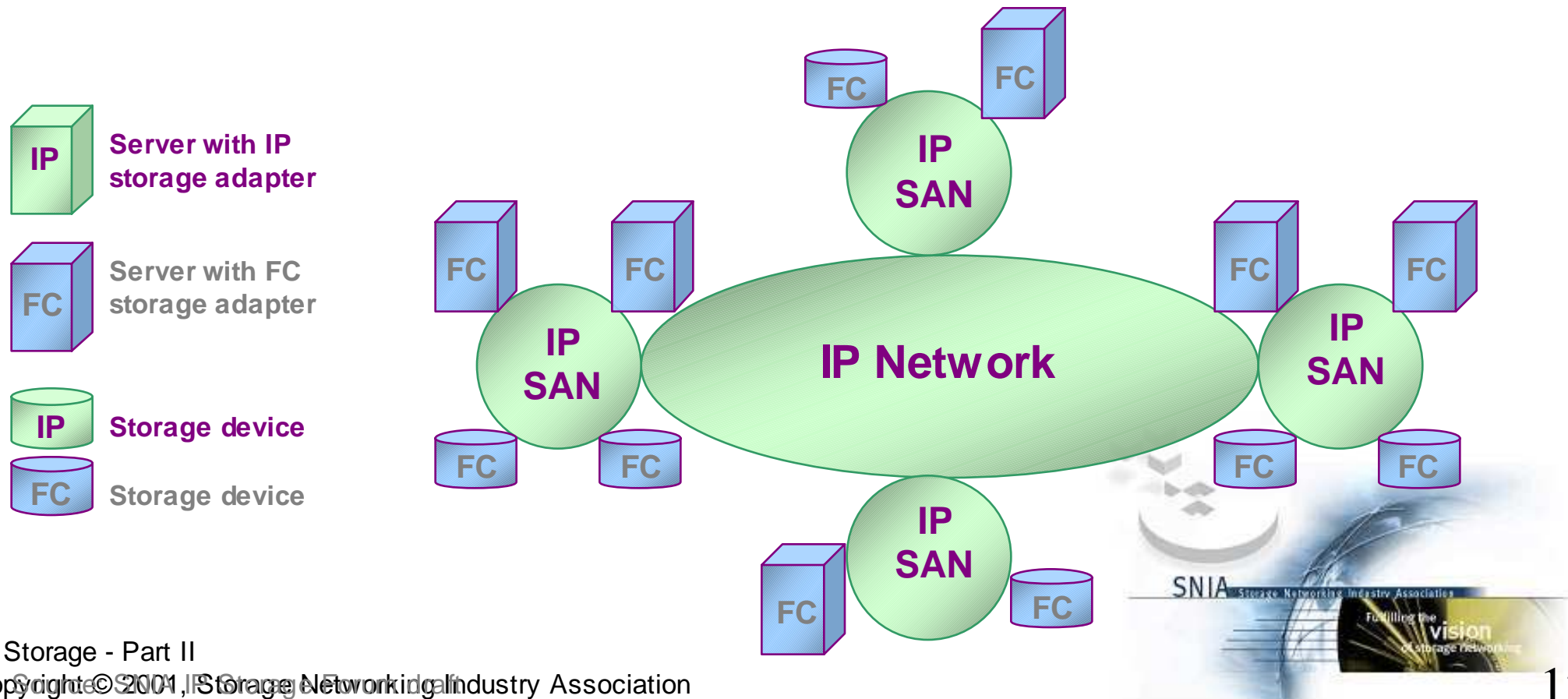
# iSCSI

- A transport protocol for SCSI that operates on top of TCP
- A new mechanism for encapsulating SCSI commands on an IP network
- A protocol for a new generation of storage end nodes that natively use TCP/IP



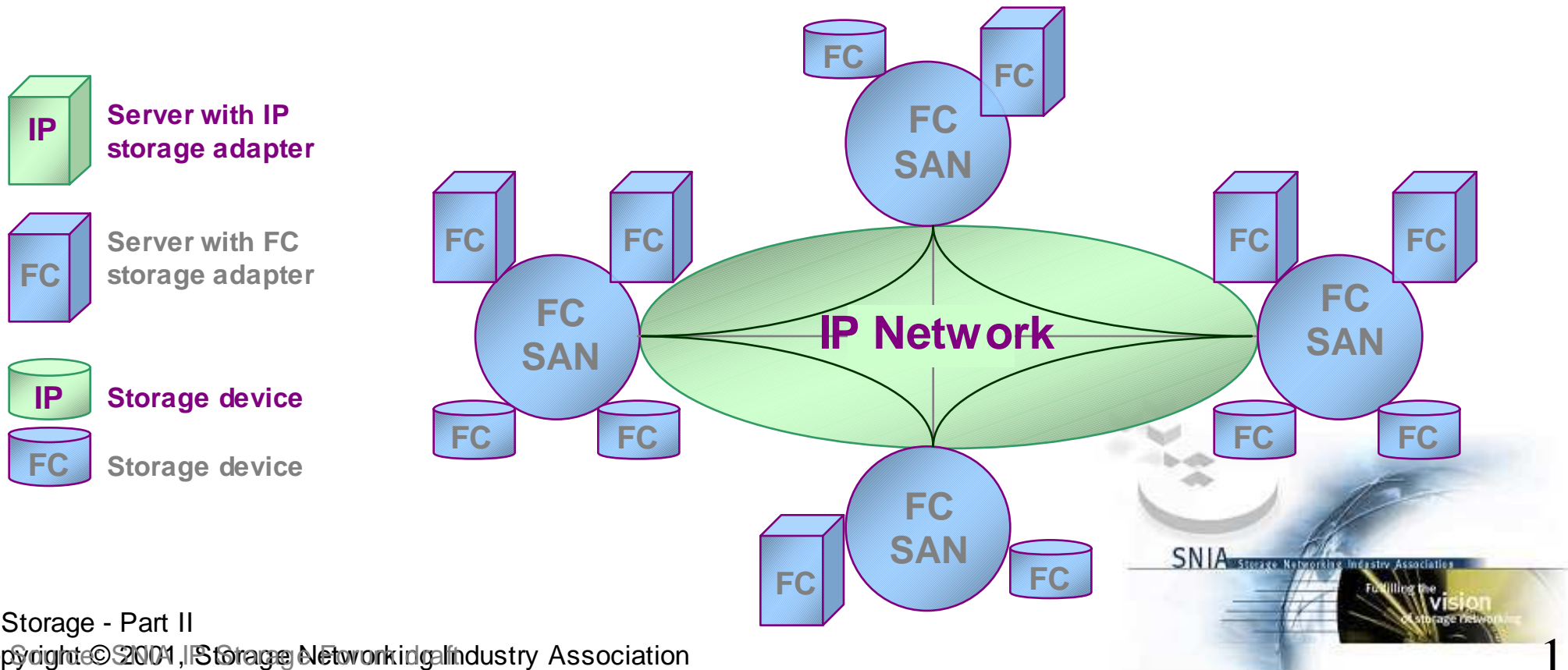
# iFCP

- A gateway-to-gateway protocol for the implementation of a FC fabric in which TCP/IP switching and routing elements replace FC fabric components
- The protocol enables the attachment of existing FC storage products to an IP network

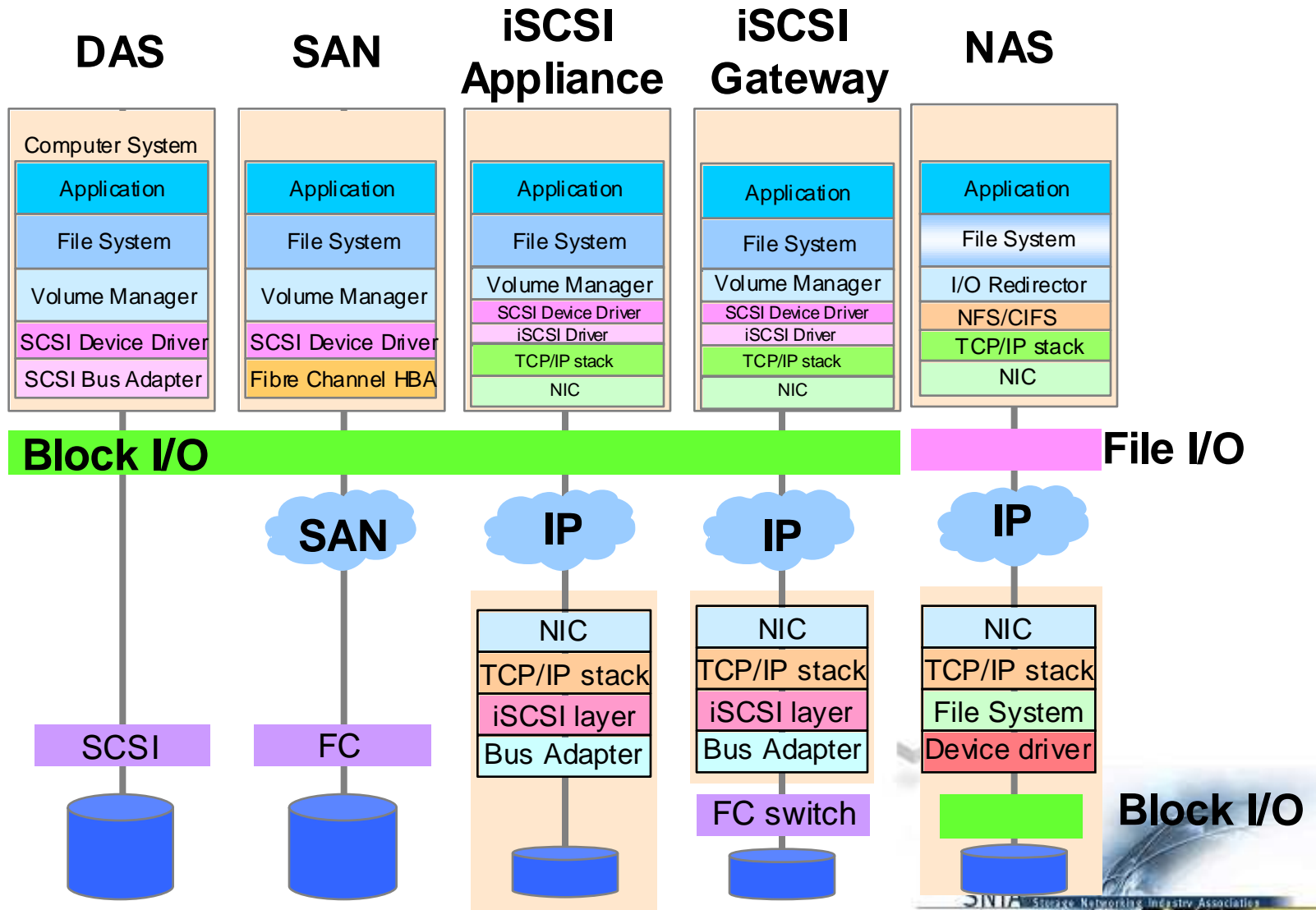


# FCIP

- Relies on IP-based network services to provide the connectivity between the SAN islands over LANs, MANs, or WANs
- Relies upon TCP for congestion control and management and upon both TCP and FC for data error correction and data loss recovery
- FC over TCP/IP treats all classes of FC frames the same as datagrams



# SAN, NAS, iSCSI Comparison



# iSCSI Momentum

**First multi-vendor interoperability demonstration:  
14 months**

**Feb. '00**  
IETF  
iSCSI  
Proposal

**July '00**  
Draft 0 of  
iSCSI  
spec.  
Vendors  
start to  
implement

**Jan. '01**  
SNIA IP  
Storage  
Forum  
Created

**Apr. '01**  
First iSCSI  
interoperability  
demonstration  
at SNW.

**May '01**  
N+1 iSCSI  
interoperability  
demos.

iSCSI router  
wins people's  
choice and best  
enterprise  
networking  
product awards.

**June '01**  
IP Storage  
Forum has  
over 50  
members

**IP Storage Forum:  
Over 55 member  
companies**

**July '01**  
First iSCSI  
plugfest,  
over 20  
vendors  
participate.

SNIA Storage Networking Industry Association

Fulfilling the  
vision  
of Storage Networking

# iFCP



# Presenters

- **Gary Orenstein – Nishan**



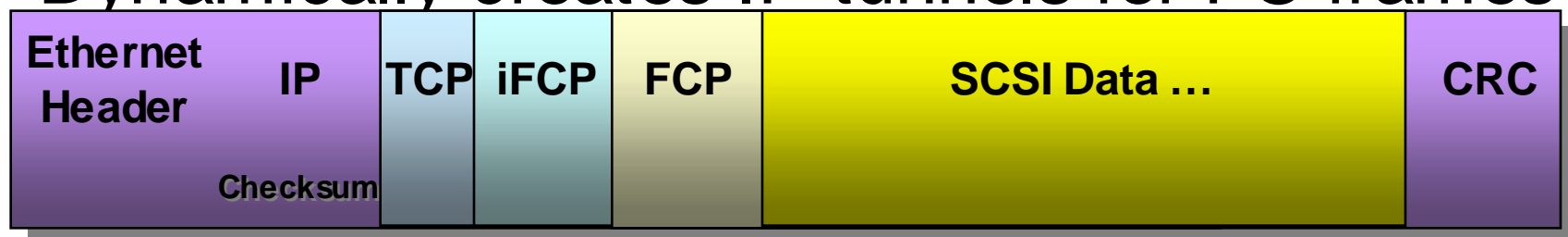
# iFCP

- iFCP is a gateway-to-gateway protocol for the implementation of a fibre channel fabric over a TCP/IP transport
- Traffic between fibre channel devices is routed and switched by TCP/IP network

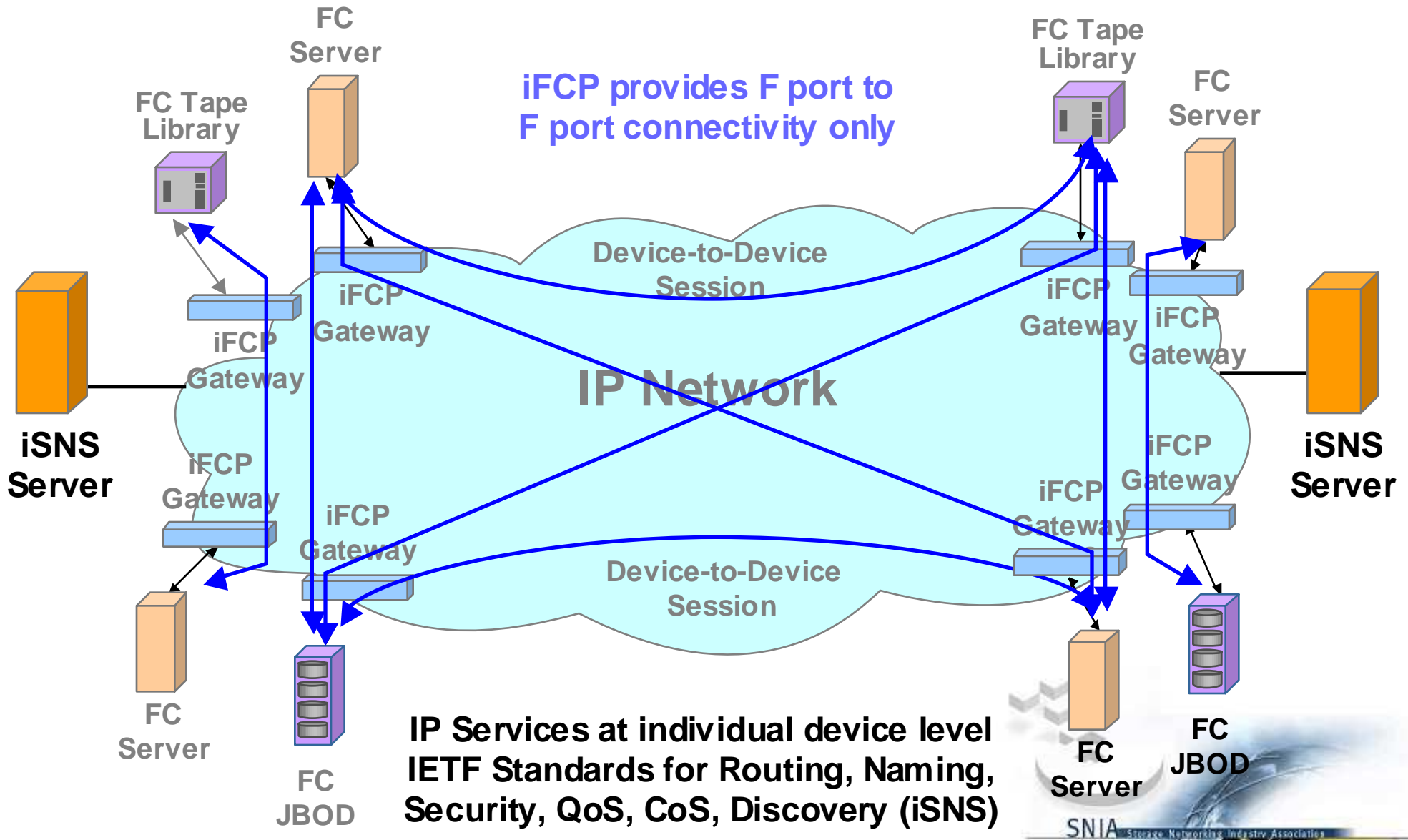


# iFCP

- The iFCP layer maps Fibre Channel frames to a predetermined TCP connection for transport
- FC messaging services and routing services are terminated at the gateways so the fabrics are not merged to one another
- Dynamically creates IP tunnels for FC frames



# iFCP Approach

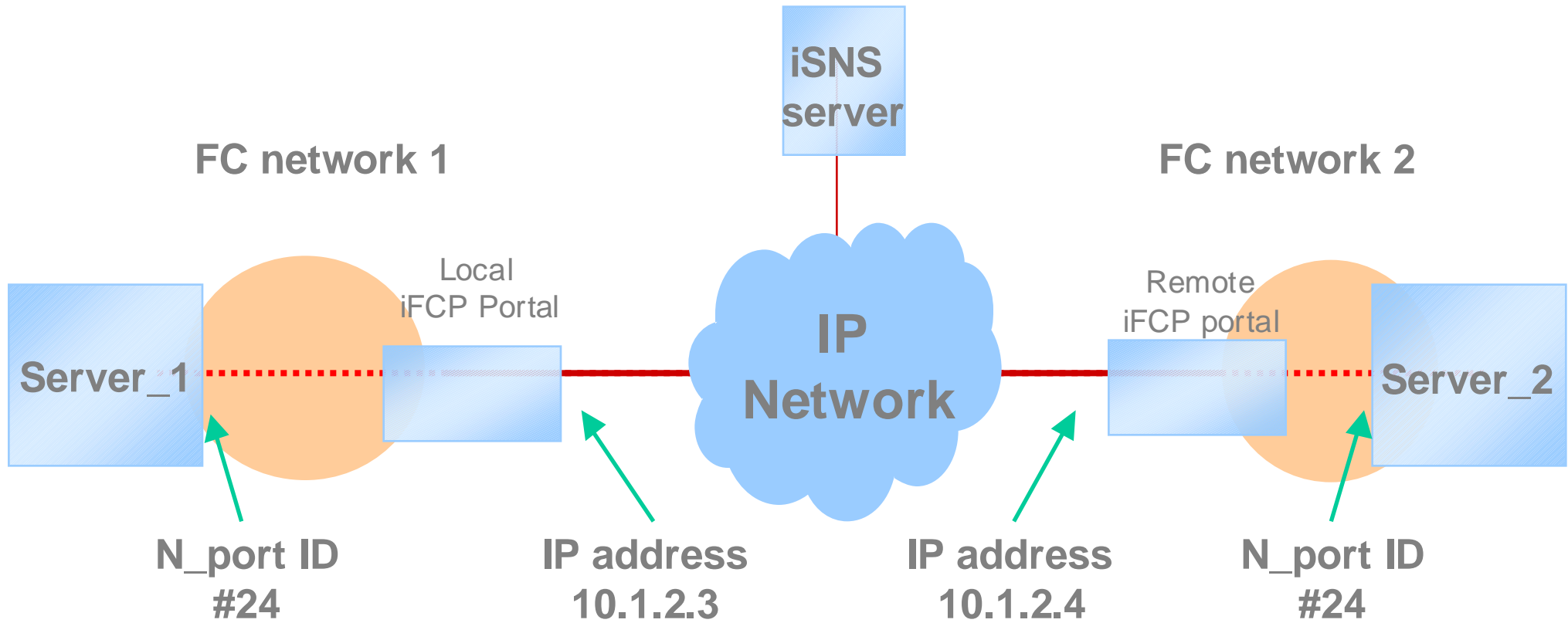


# iSNS

- iSNS (Internet Storage Name Server)
- Provides registration and discovery of SCSI devices and Fibre Channel-based
- In IP-based storage like iSCSI end devices registered with iSNS
- In iFCP, Fibre Channel-based storage end devices register with iSNS by a iFCP gateway



# iSNS Operation



**Problem: Two identical N\_port IDs**

**Solution: Create new ID (based on IP address + N\_port ID) = 2422**



# iFCP Participating Companies

- ADIC
- BakBone
- BMC Software
- Chaparral
- DataCore
- Dell
- EMC
- Emulex
- Eurologic
- Fujitsu
- IBM
- JNI
- Legato
- Nortel
- QLogic
- Quantum
- Seek Systems
- Siemens
- SpectraLogic
- StorageNetworks
- Sun
- Troika
- Xiotech



Source: Nishan Systems website

IP Storage - Part II

Copyright © 2001, Storage Networking Industry Association

# Presenters

- **Jeff Martin – SAN Valley**



# FCIP

**Jeff Martin – SAN Valley**



# Leverage Fibre Channel Applications & Performance

- Preserves existing Fibre Channel infrastructure and investments
  - Fibre Channel widely deployed in live, production environments
  - Interoperable, multi-vendor solutions available
- Extends Fibre Channel applications over regional/global distances
  - High performance, highly reliable, robust storage networking
- Fully supports Fibre Channel Fabric services across the MAN/WAN



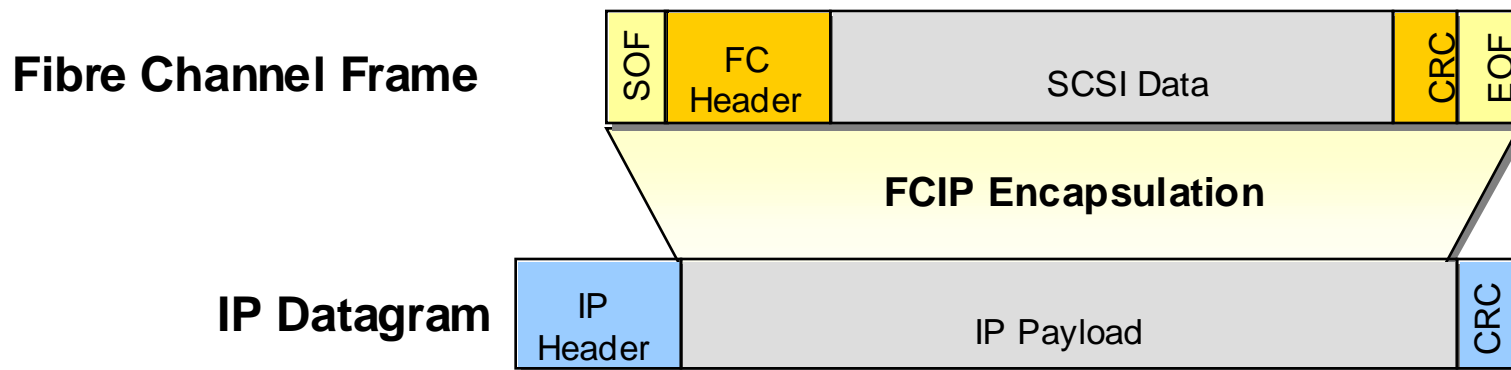
# FCIP – IETF IPS Working Group Draft Standard

- Specifies the encapsulation for Fibre Channel frames being transported by TCP/IP
- Specifies the use of the encapsulation to create a virtual Fibre Channel link that connects Fibre Channel devices and fabric elements
- Specifies the TCP/IP environment for supporting virtual Fibre Channel links and providing the capabilities for tunneling Fibre Channel traffic over an IP-based network
  - includes security, data integrity, congestion and performance

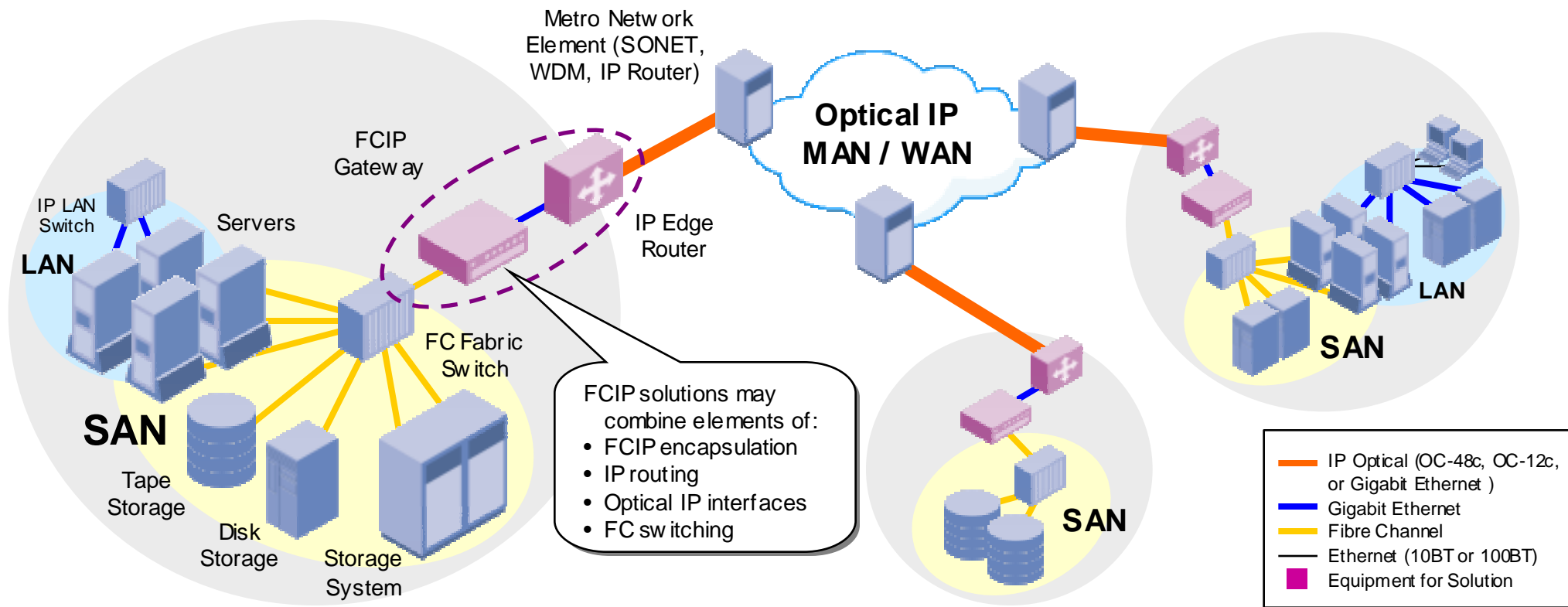


# Fibre Channel Over IP (FCIP) Protocol

- Transparent Operation for Local & Remote SANs
  - Only FCIP Gateway needs to be aware of FCIP encapsulation
  - Appears like FC to the SAN, and IP to the LAN/MAN/WAN network



# Interconnecting Remote SANs Using FCIP

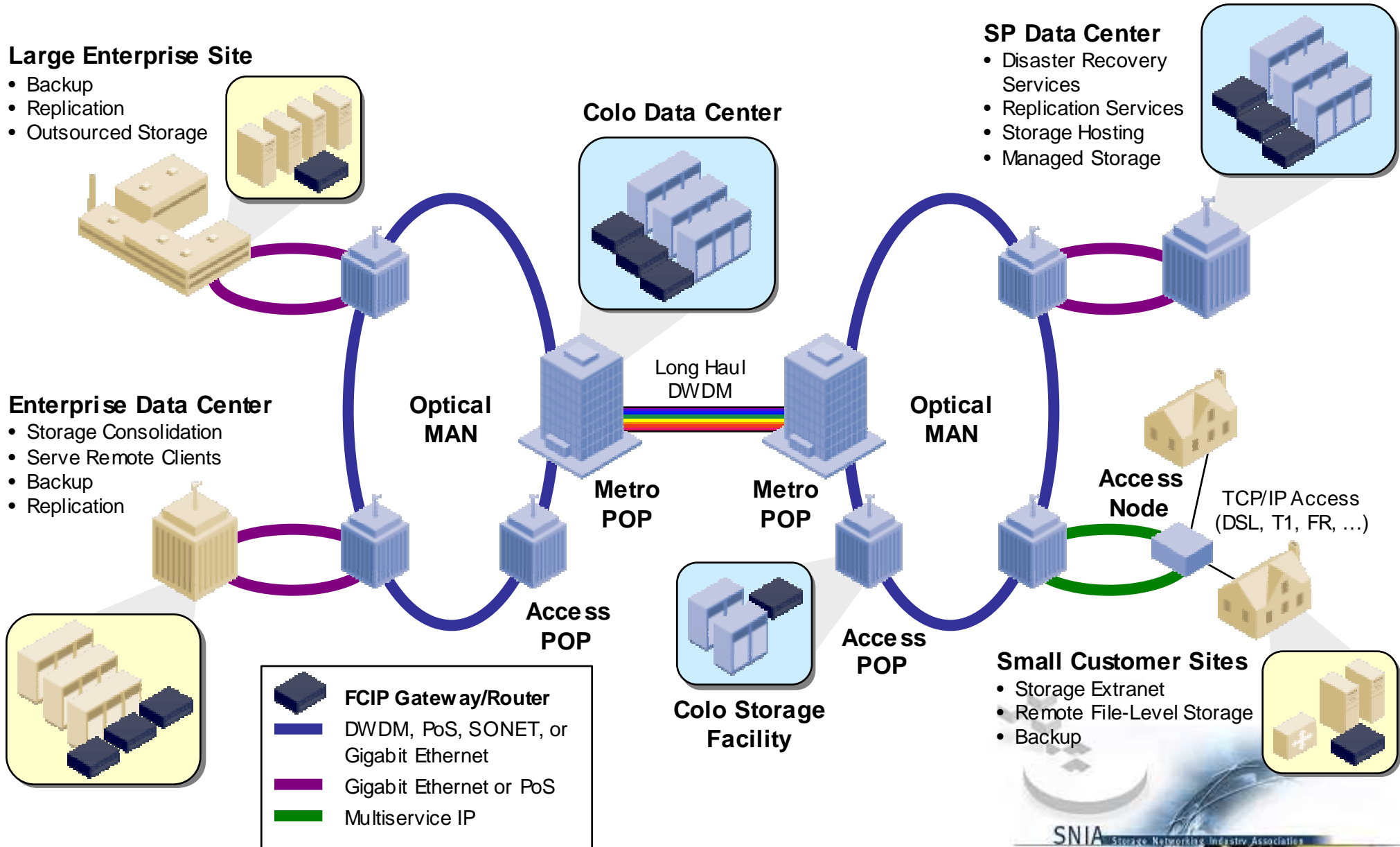


Description	Benefits
<ul style="list-style-type: none"> <li>• FCIP Gateway encapsulates FC frames into IP</li> <li>• Multi-point connectivity over IP MAN/WANS</li> <li>• Can interconnect SANs up to several hundred km (+++)</li> </ul>	<ul style="list-style-type: none"> <li>• Preserves FC infrastructure and investments</li> <li>• Only edge devices need to be aware of FCIP encapsulation</li> <li>• Takes advantage of existing IP/optical networks</li> </ul>

SNIA Storage Networking Industry Association

Fulfilling the vision of Storage Networking

# FCIP Over Service Provider Networks

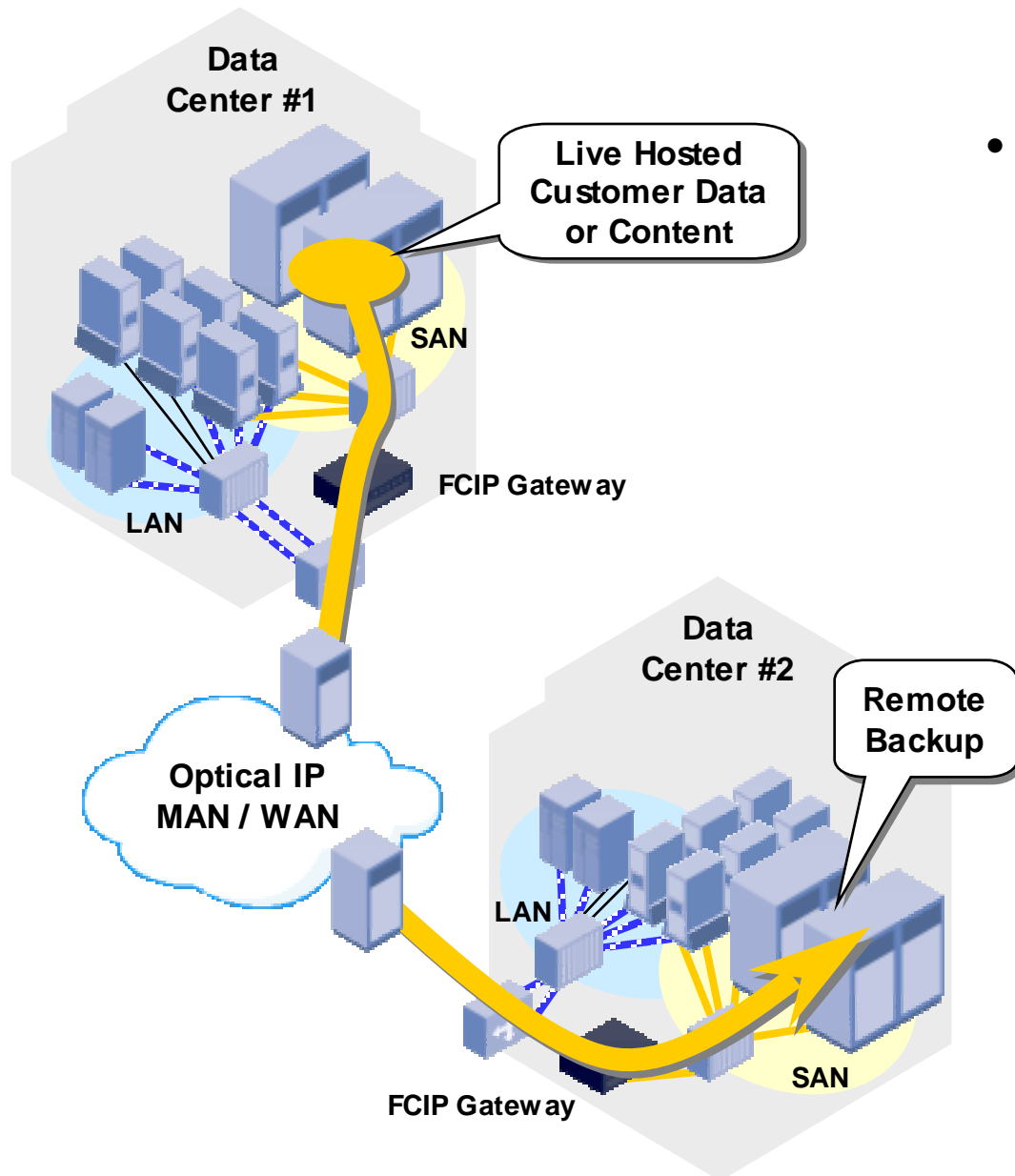


# FCIP Applications

- FCIP enables SAN interconnectivity over long distances
  - Remote backup and restore
  - Data sharing
- At higher link speeds, synchronous applications can be implemented
  - Synchronous disk mirroring
  - Shared storage
  - Data sharing



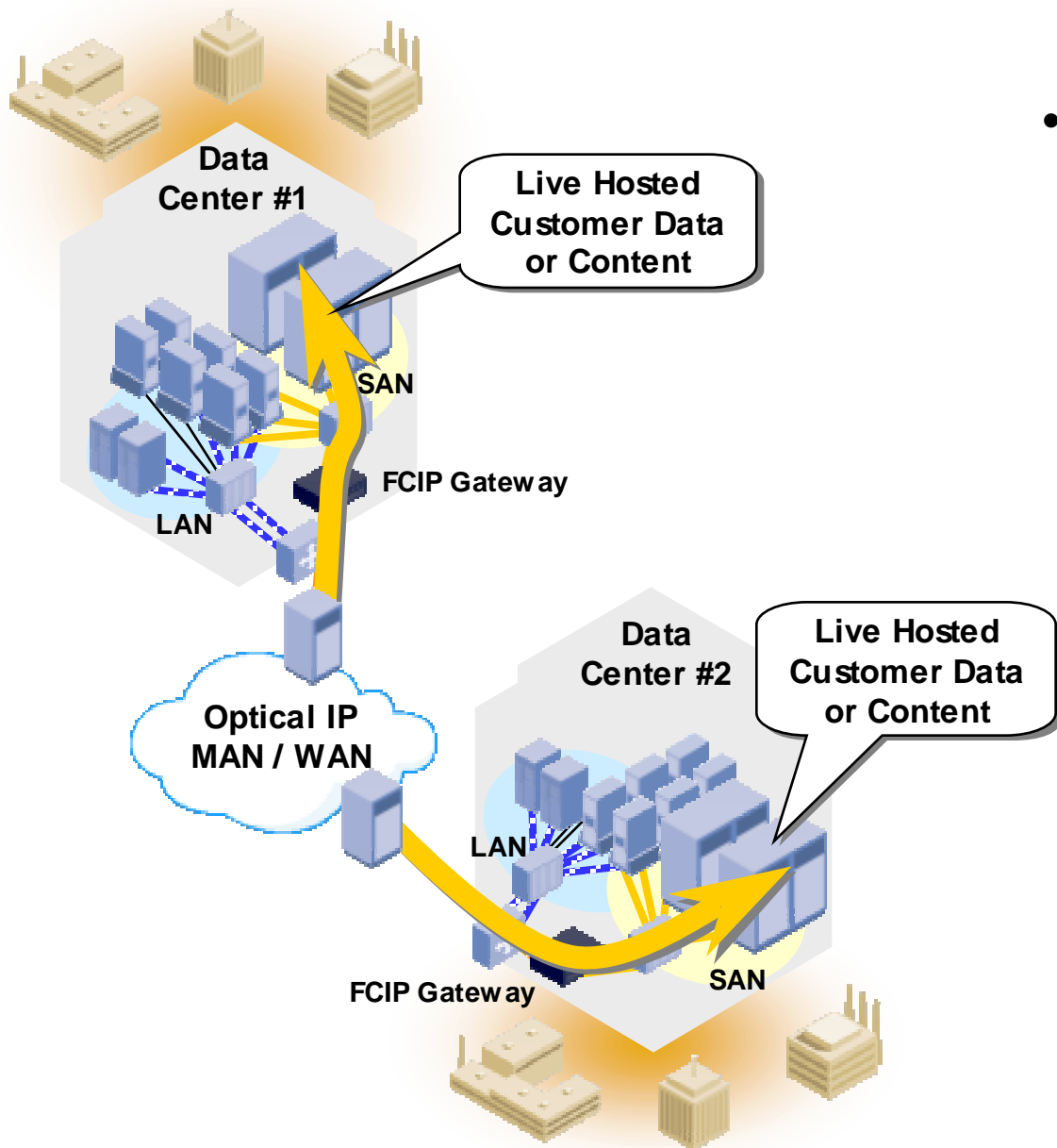
# Data Backup & Restore



- **Application description:**
  - Hosted customer data or content is backed up at remote data center
  - Data centers may be across the metro or in separate geographical regions
  - In case of data loss, backup is accessible directly over the MAN/WAN



# Data Replication

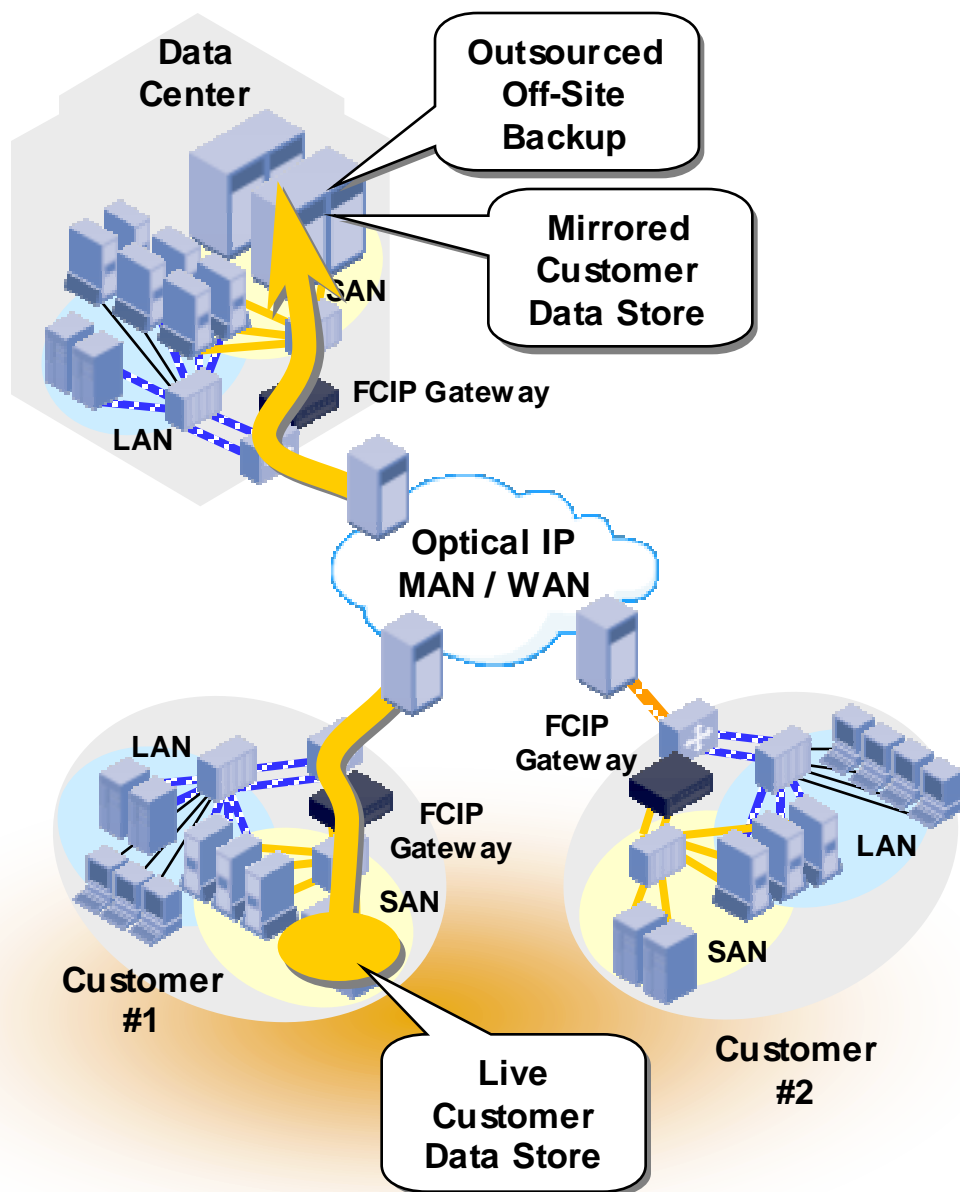


- **Application Description:**

- Hosted customer data or content is kept continuously synchronized across the network
- Each data store contains live, production data
- Can be used to deliver media content to distribution/cache centers near users



# Outsourced Data Backup & Restore or Data Replication

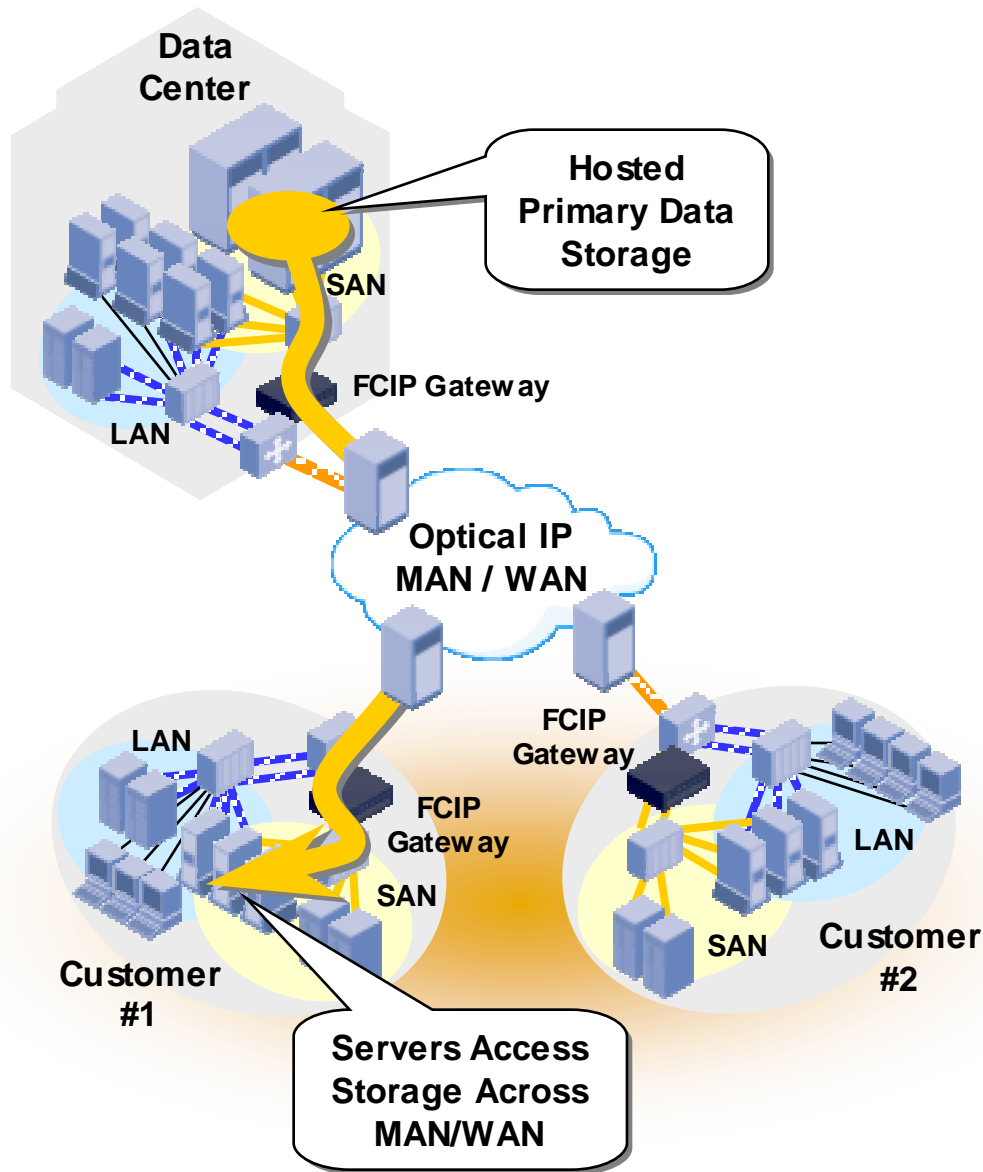


- **Application Description:**

- Customer data is backed up or replicated in a remote data center
- Rapidly-changing or volatile data can be continuously replicated at the data center
- In case of loss at customer site, data is accessible directly over the MAN/WAN



# Outsourced Storage Hosting



- **Application Description:**
  - Customer data is hosted at the service provider data center
  - Servers at customer site(s) efficiently access hosted storage over the network



# FCIP Participating Companies

- Adaptec
- Aristos Logic
- Avaya
- Brocade
- Cisco
- CNT
- Compaq
- Crossroads
- EMC
- Emulex
- Entrada
- Gadzoox
- Eurologic
- FalconStor
- Hitachi Data System
- IBM
- Intel
- JNI
- Legato
- Lucent
- Maranti Networks
- NetConvergence
- Netreon
- Nishan Systems
- Overland Data
- Pirus Networks
- QLogic
- Quantum
- Rhapsody Networks
- SAN Valley
- StorageTek
- StoreAge
- Tek-Tool
- Tivoli
- Tokyo Electron
- Troika Networks
- Vixel



Source: Network World, 12/25/00

# Storage Networking Industry Association

## Clearing the Confusion: A Primer on Internet Protocol Storage Part III

Ahmad Zamer - Intel Corporation



# Overview

- Introduction
- Small Computer System Interface
- Fibre Channel
- Networked Storage
- Benefits of IP Storage
- IP Storage technologies
- **iSCSI**
- **Conclusions**



# iSCSI



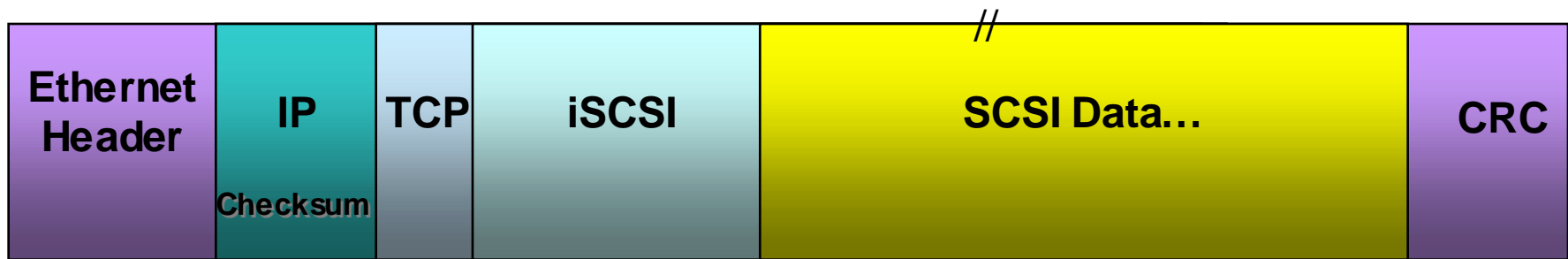
# Presenters

- **Brice Clark – HP**
- **Gary Orenstein – Nishan**
- **Ahmad Zamer - Intel**

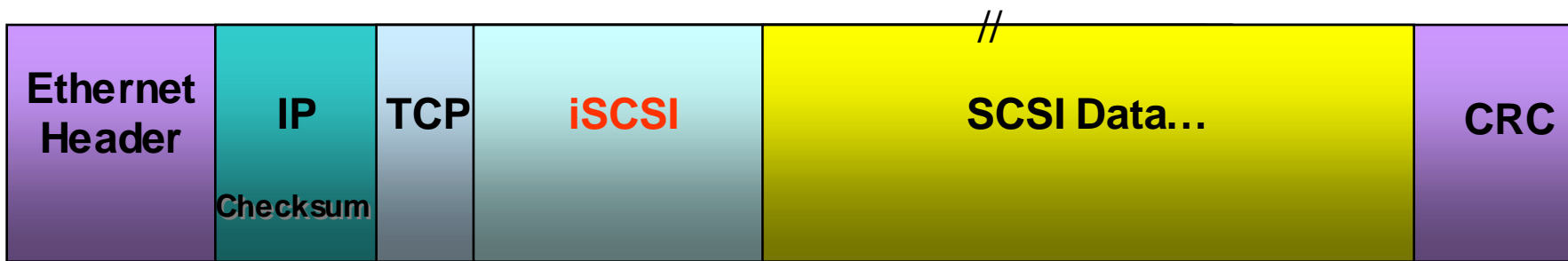
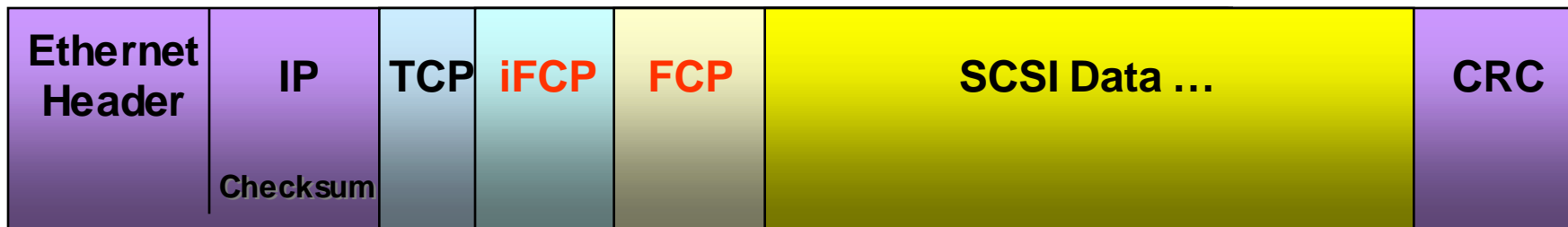
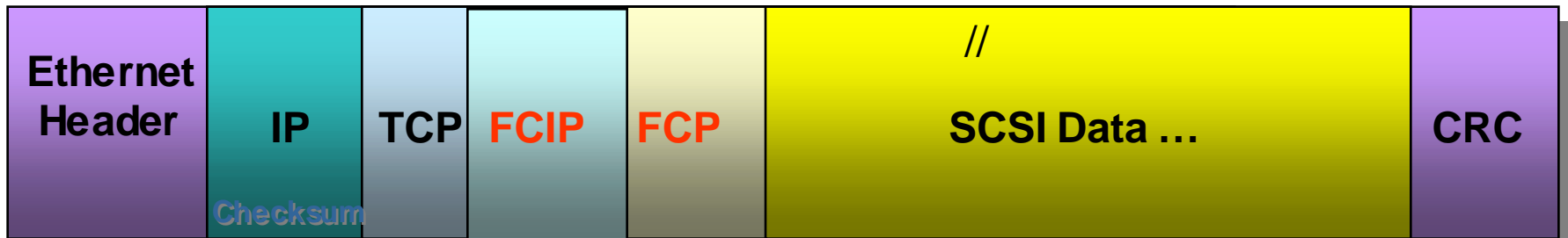


# iSCSI

- **iSCSI is a SCSI transport protocol for mapping of block-oriented storage data over TCP/IP networks**
- **The iSCSI protocol enables universal access to storage devices and Storage Area Networks (SANs) over standard TCP/IP networks**



# iSCSI, iFCP, FCiP



# What is iSCSI ? -

- iSCSI is a Transport for SCSI Commands
  - iSCSI is an End to End protocol
  - iSCSI can be implemented on Desktops, Laptops and Servers
  - iSCSI can be implemented with current TCP/IP Stacks
  - iSCSI can be implemented completely in a HBA
  - iSCSI has the concept of Human readable SCSI Device (Node) naming
- iSCSI Transport includes Security as a base concept
  - Authentication (at the Node Level)
  - Enabled for IPSec and other Security Techniques
- iSCSI defines Discovery as a basic element
- iSCSI define process for remote Boot, as a basic element
- iSCSI defines MIB standards as a basic element



# What is iSCSI ? – cont.

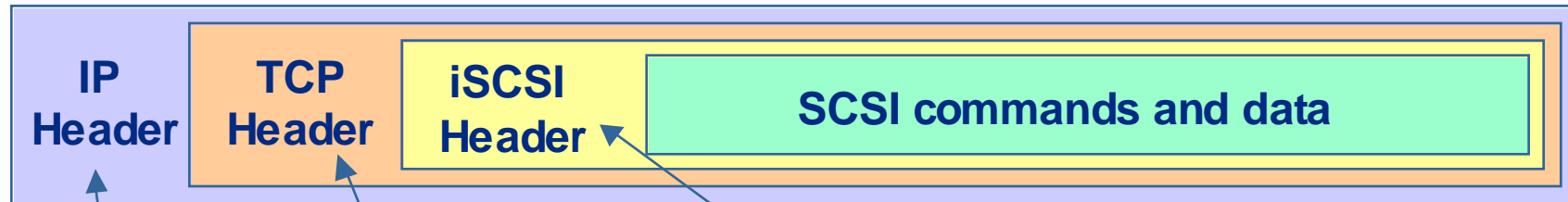
- A Framework for IP Based Storage
  - A frame work for creating standards in the IPS
- iSCSI (version 8)
  - The encapsulation protocol
- iSNS Internet Storage Name Service
  - A std. framework for discovering storage devices.
- A Standard for Bootstrapping Clients
  - Allows clients to boot from iSCSI devices
    - SW vrsn Requires a pre-boot execution environments (PXE).
    - HW vrsn should boot like an normal SCSI HBA

<http://www.ietf.org/html.charters/ips-charter.html>



# iSCSI – Cont.

- **iSCSI (Internet SCSI) specifies a way to “encapsulate” SCSI commands in a TCP/IP network connection:**



Contain “routing” information  
So that the message can find its  
Way through the network

Provides information necessary to  
guarantee delivery

Explains how to extract  
SCSI commands and data



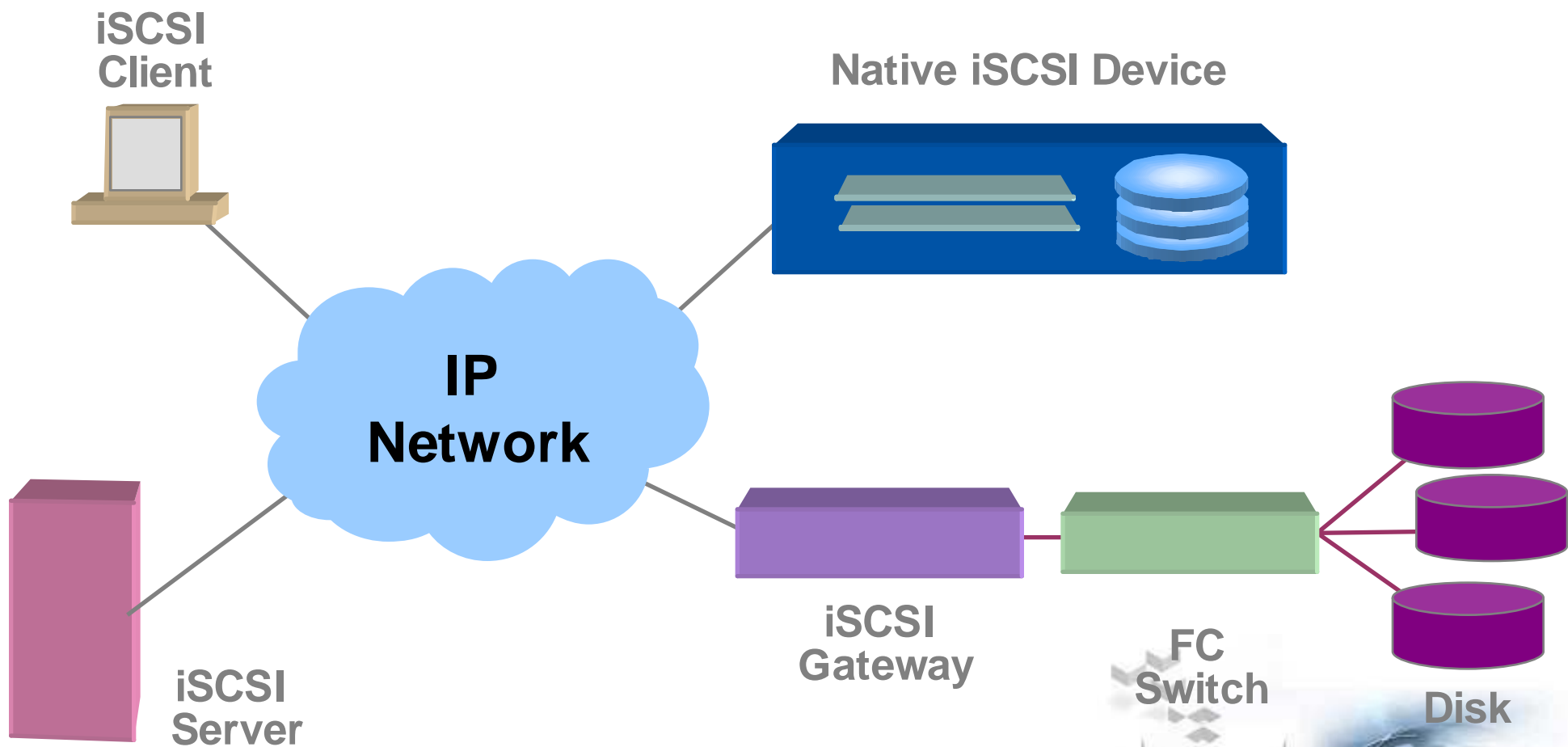
# iSCSI - TCP Packet



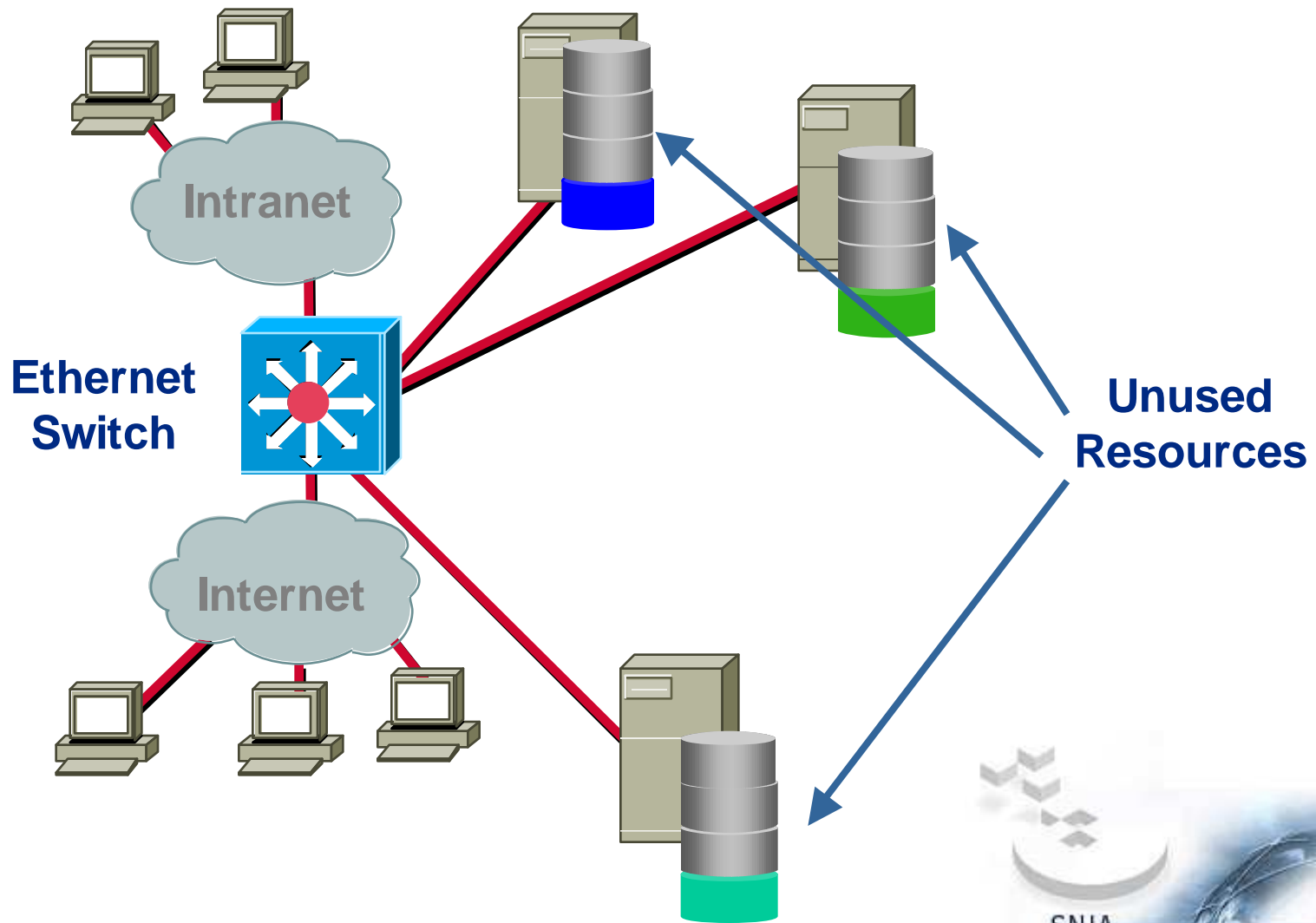
- **Ethernet frame requires additional CPU processing**
- **Headers must be stripped**
- **Packets ordered**
- **Data copied into memory buffers**
- **CRC checked**



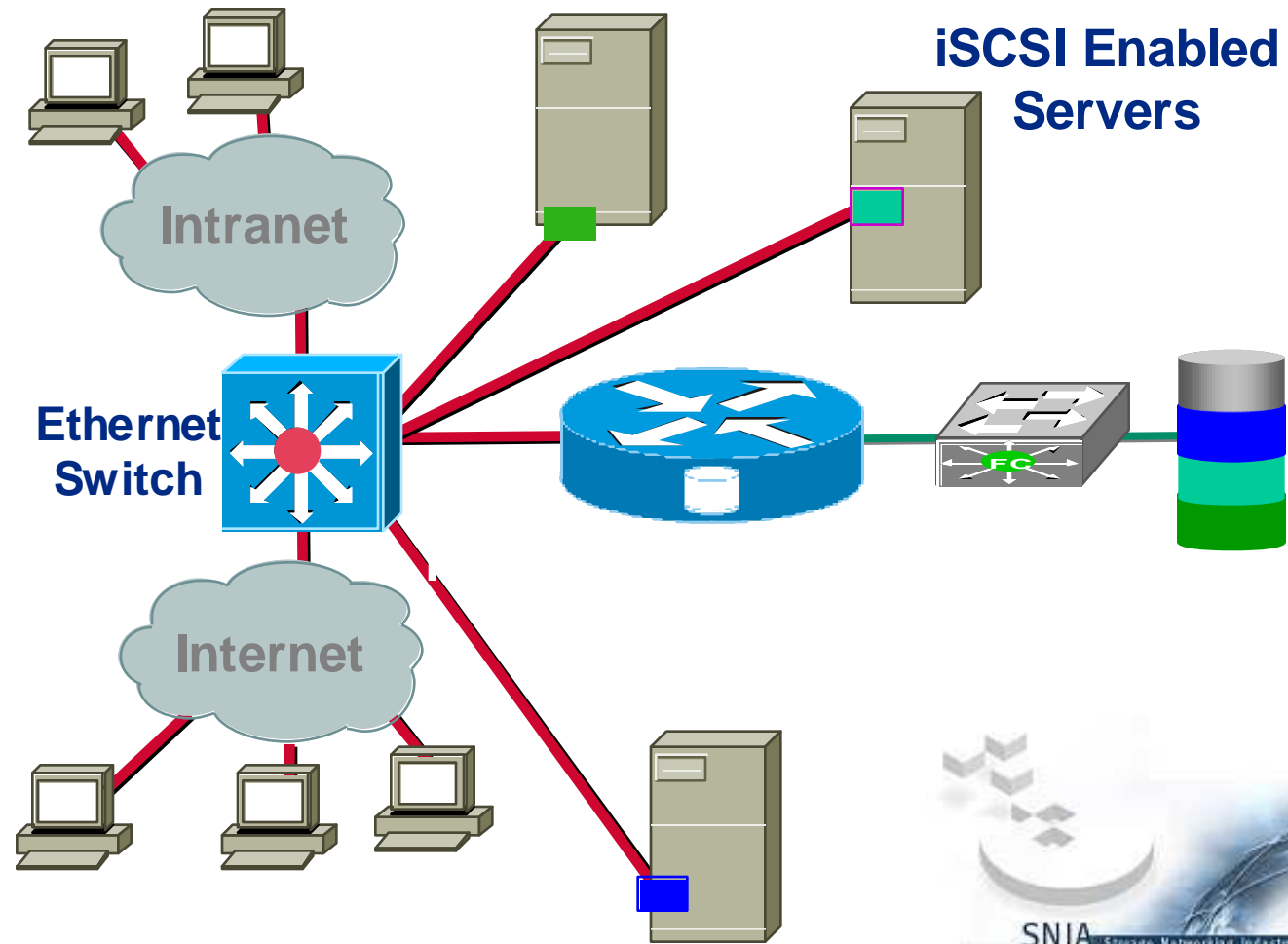
# iSCSI Implementations



# iSCSI for Storage Consolidation



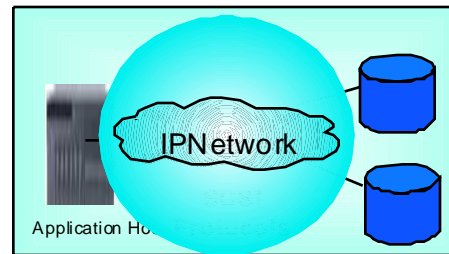
# iSCSI for Storage Consolidation



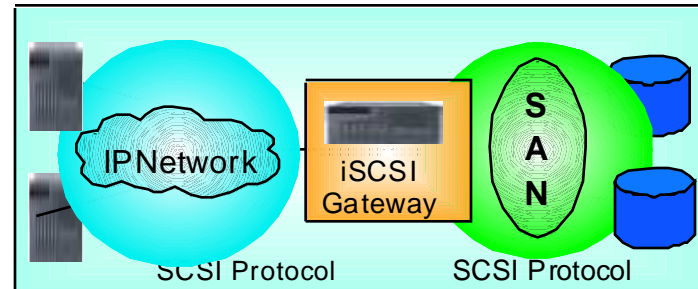
# iSCSI Deployment

Same HW Configurations as NAS  
Workgroup, Departmental, & Enterprise  
(Appliances and Gateways)  
GAs throughout 2001 & 2002

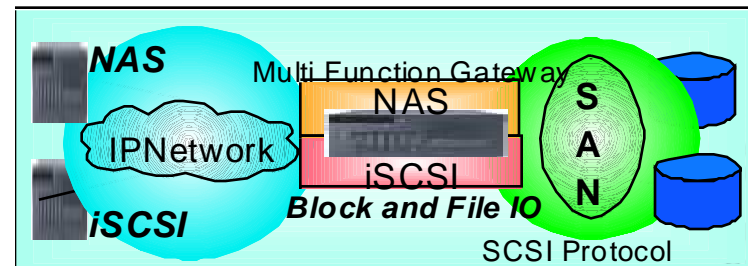
Independent  
iSCSI  
Deployment



Extending  
the SAN



In  
Combination  
with NAS

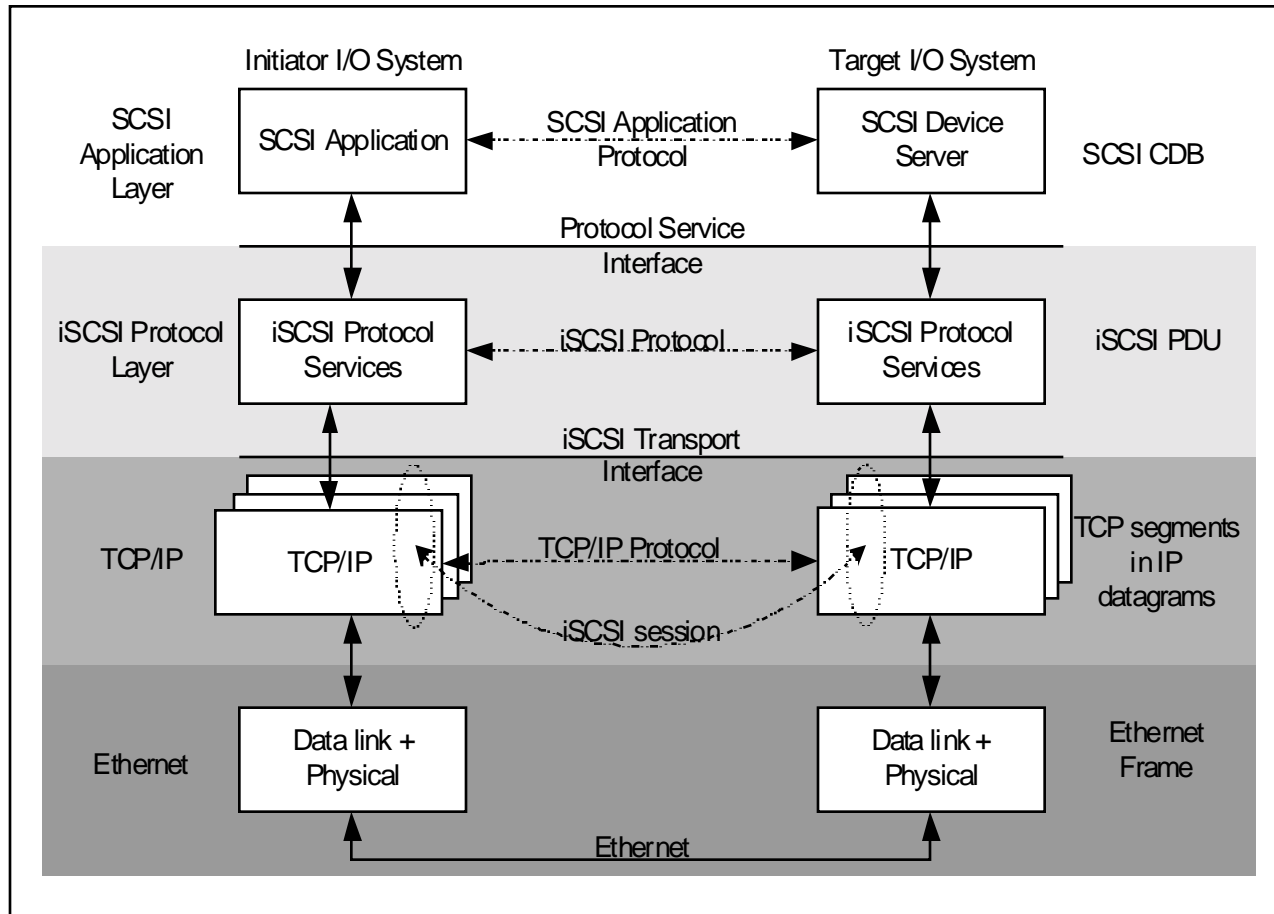


# iSCSI Architecture

- **Overview**
  - **Architectural Model**
  - **Features Beyond // SCSI**
  - **Issues Beyond // SCSI**



# iSCSI - a Layered Model



- Replaces shared bus with switched fabric
- Transparently encapsulates SCSI CDBs
- **Unlimited target and initiator connectivity**

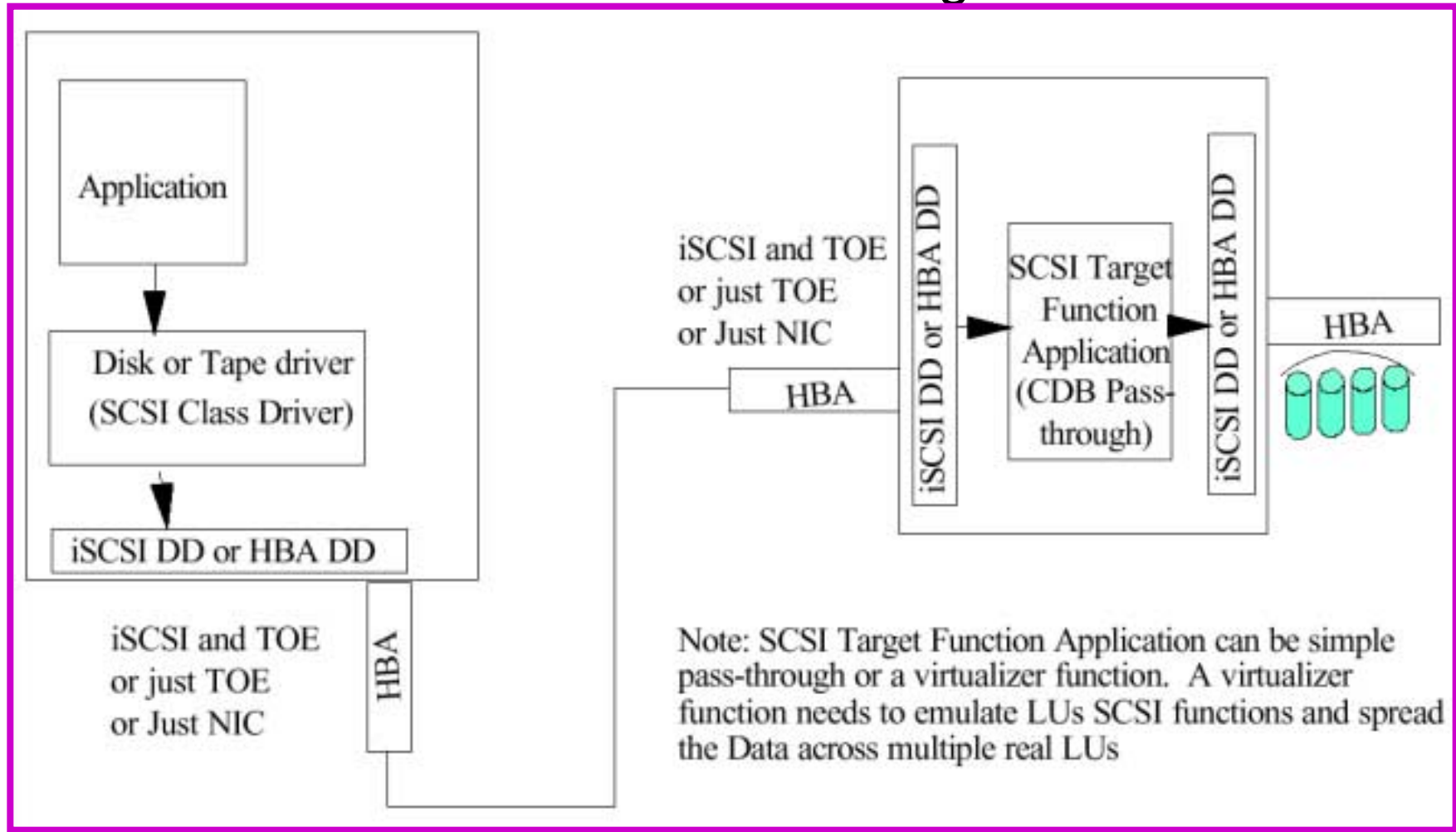
## Newly Raised Issues

- Security
- Establishing Connections
- Naming/discovery
- Command Ordering
- Congestion Control
- QoS



# iSCSI & SCSI Layering

## SCSI Host connection to SCSI target via iSCSI



# iSCSI Structure

- iSCSI has the concept of a Session
  - A session maybe made up of one or more TCP/IP connections
  - The Session can be thought of as a SCSI Port
  - The Session is started after Login is complete



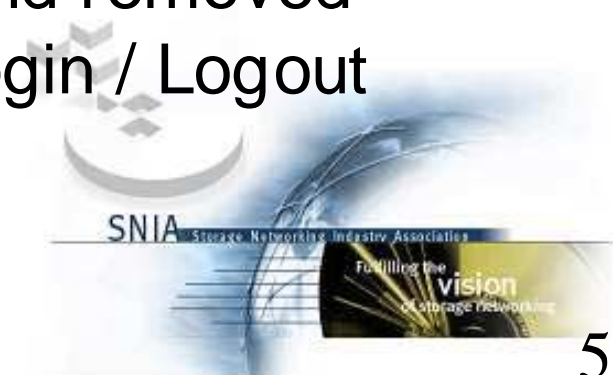
# iSCSI Structure – cont.

- Login begins with the first connection
  - Text Strings are exchanged that define the parameters of the session
    - **Each Side (Initiator and Target) negotiate what supported options**
    - **Forms of Security**
    - **Amount of unsolicited buffer**
    - **Types of Data Delivery**
      - Solicited
      - unsolicited
      - immediate
  - Target Node Names, and Initiator Node Names (for Authentication)
  - Alias can be accepted for admin reporting
  - After exchange of Text Strings, Full Session Mode can carry SCSI CDBs/Data
- Session may end with Logout, or I/O error causing dropped connection
  - But often connection can be reestablished and Session continued



# iSCSI Key Points

- Sessions:
  - iSCSI Session = a group of TCP connections linking an initiator with a target ( i.e. can be one or more connections)
    - Note: A TCP connection that is part of an iSCSI session will only be used to carry iSCSI traffic
  - The iSCSI initiator and target use this session for communicating iSCSI commands, control messages, parameters, and data to each other
  - TCP connections can be added and removed from a session using the iSCSI Login / Logout commands



# iSCSI Key Points (cont.)

- Connection Allegiance:
  - For SCSI commands that require data and/or parameter transfer, the (optional) data phase and status phase must be sent over the same TCP connection that was used to deliver the SCSI command during the command phase
  - Consecutive commands that are part of a SCSI task (i.e. a linked set of commands) MAY use different connections within the session
  - Connection allegiance is strictly per-command and not per-task



# iSCSI Key Points (cont.)

- Tasks:
  - A linked set of SCSI commands
  - One and only one SCSI command at a time can be outstanding within any given iSCSI task
- Initiator Task Tags (ITT) and Target Tags:
  - Initiator tags for all pending commands must be unique initiator-wide
  - SCSI Data packets are matched to their corresponding SCSI commands by using Tags that are specified in the protocol



# iSCSI Key Points (cont.)

- Solicited or Unsolicited Messages:
  - Initiator to Target
    - User data or command parameters will be sent as either solicited data or unsolicited data
    - Solicited data is sent in response to Ready To Transfer (R2T) PDUs
    - Unsolicited data can be part of an iSCSI command PDU ("immediate data") or an iSCSI data PDU
    - The maximum size of an individual data PDU or the immediate-part of the initial unsolicited burst MAY be negotiated at login
  - Target to Initiator
    - Targets operate in either solicited (R2T) data mode or unsolicited (non R2T) data mode

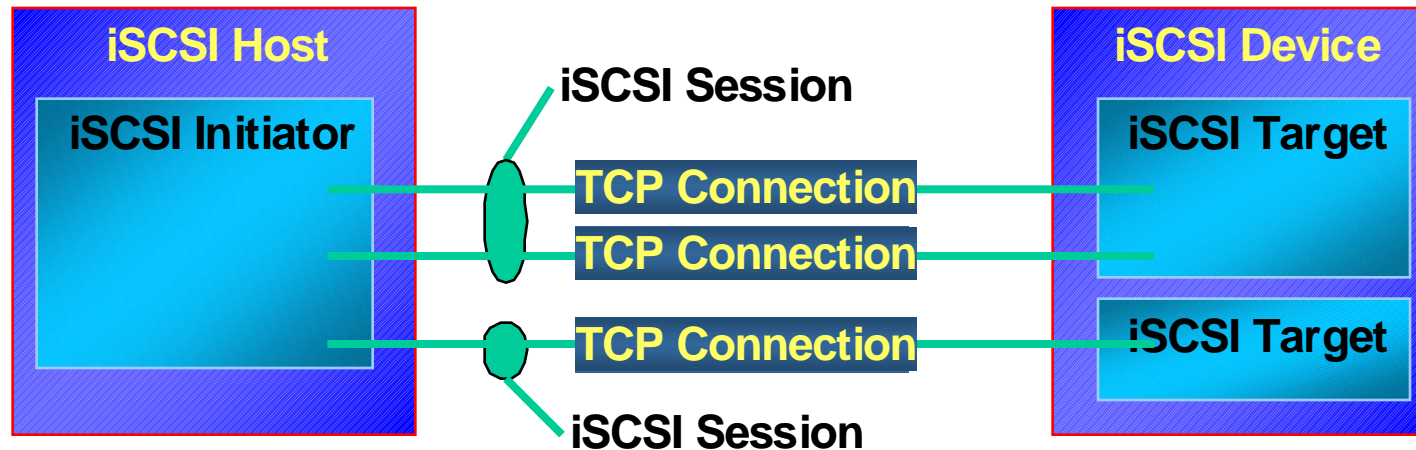


# iSCSI Key Points (cont.b)

- Numbering:
  - iSCSI uses Command and Status Numbering
    - Command numbering
      - session wide and is used for ordered command delivery over multiple connections with in a session. It can also be used as a mechanism for command flow control over a session
    - Status numbering
      - per connection and is used to enable recovery in case of connection failure
  - Fields in the iSCSI PDUs communicate the reference numbers between the initiator and target. During periods when traffic on a connection is unidirectional, iSCSI NOP-message PDUs may be utilized to synchronize the command and status ordering counters of the target and initiator



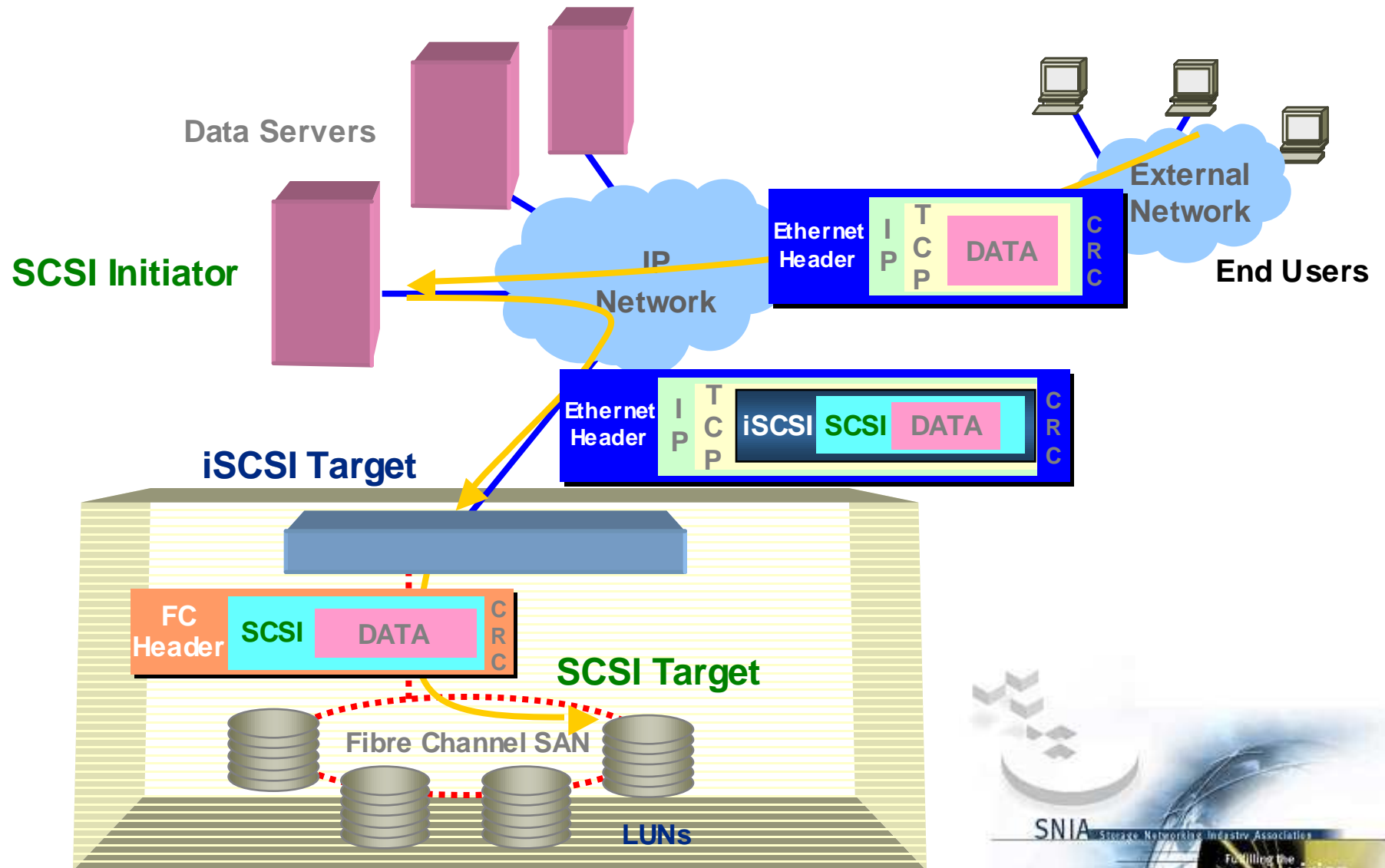
# iSCSI Sessions



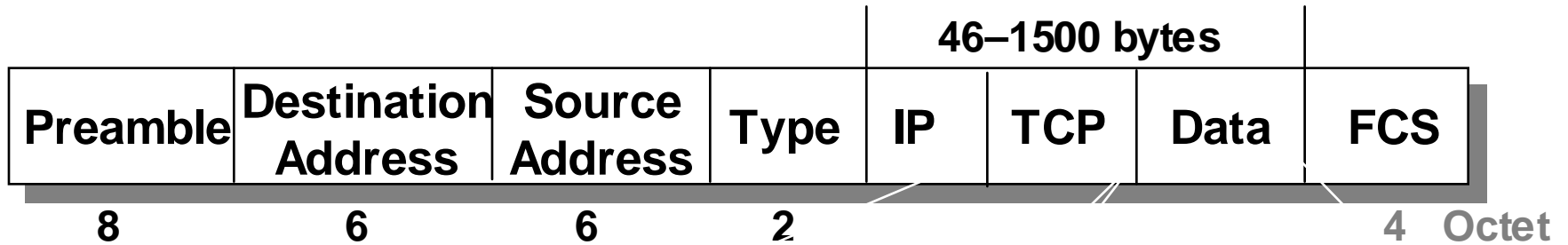
- Session between initiator and target
  - One or more TCP connections per session
  - Login phase begins each connection
- Deliver SCSI commands in order
- Recover from lost connections



# iSCSI Encapsulation



# iSCSI Packet



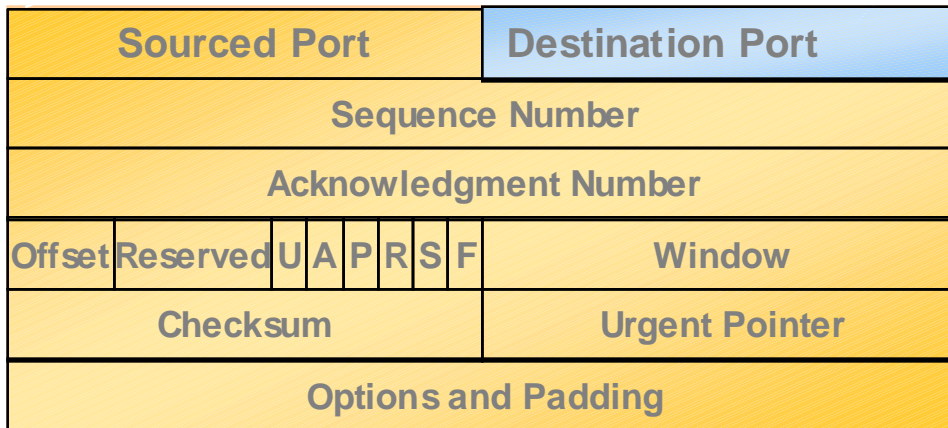
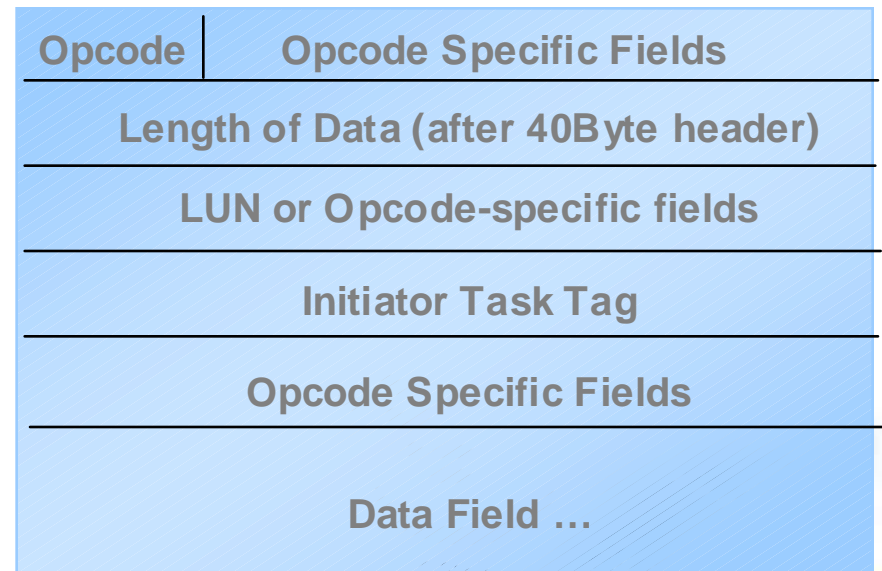
## Well-known

### Ports:

21 FTP  
23 Telnet  
25 SMTP

5003 iSCSI

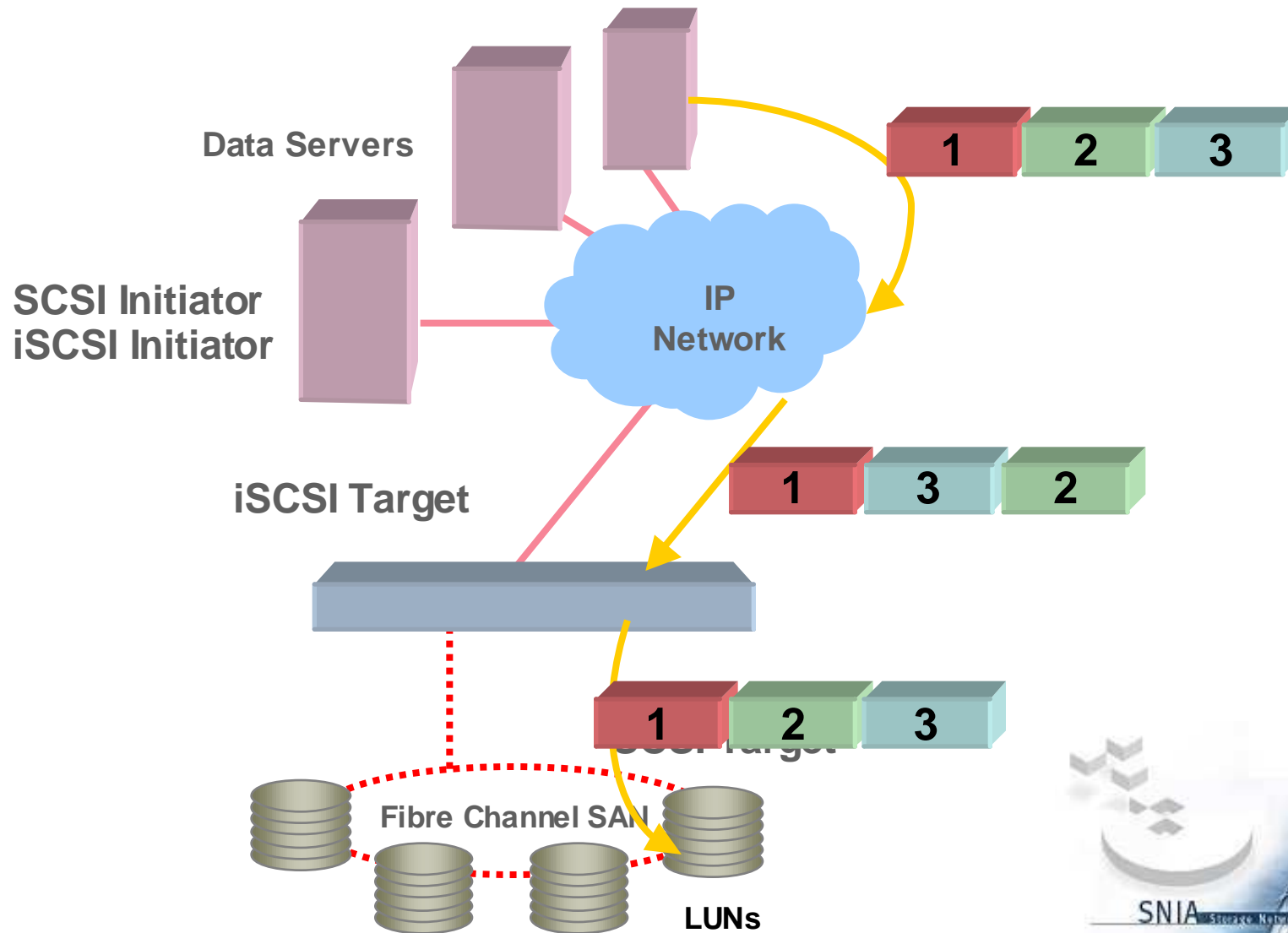
## iSCSI Encapsulated



## TCP Header

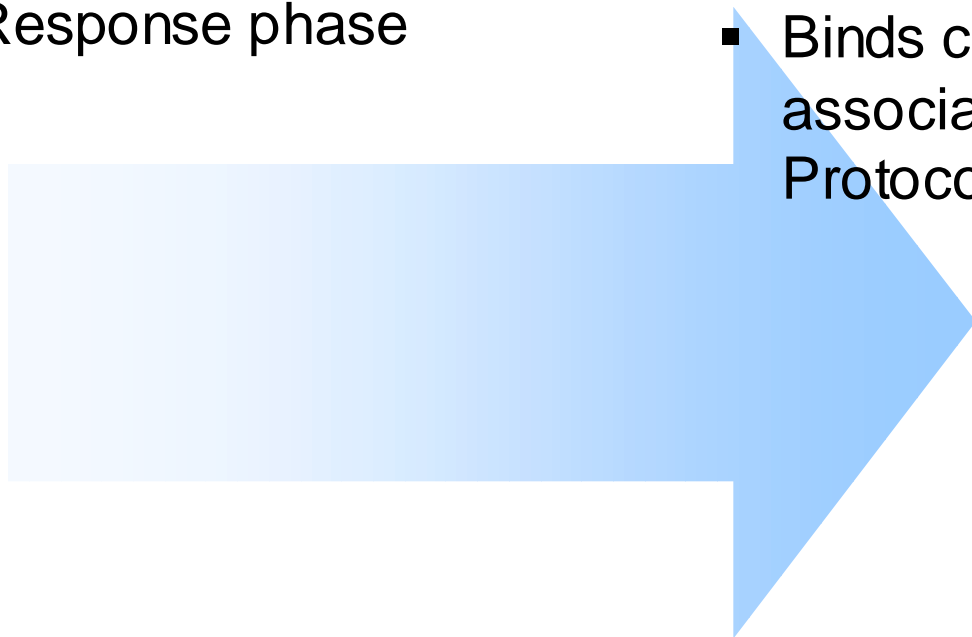


# iSCSI Packet Order

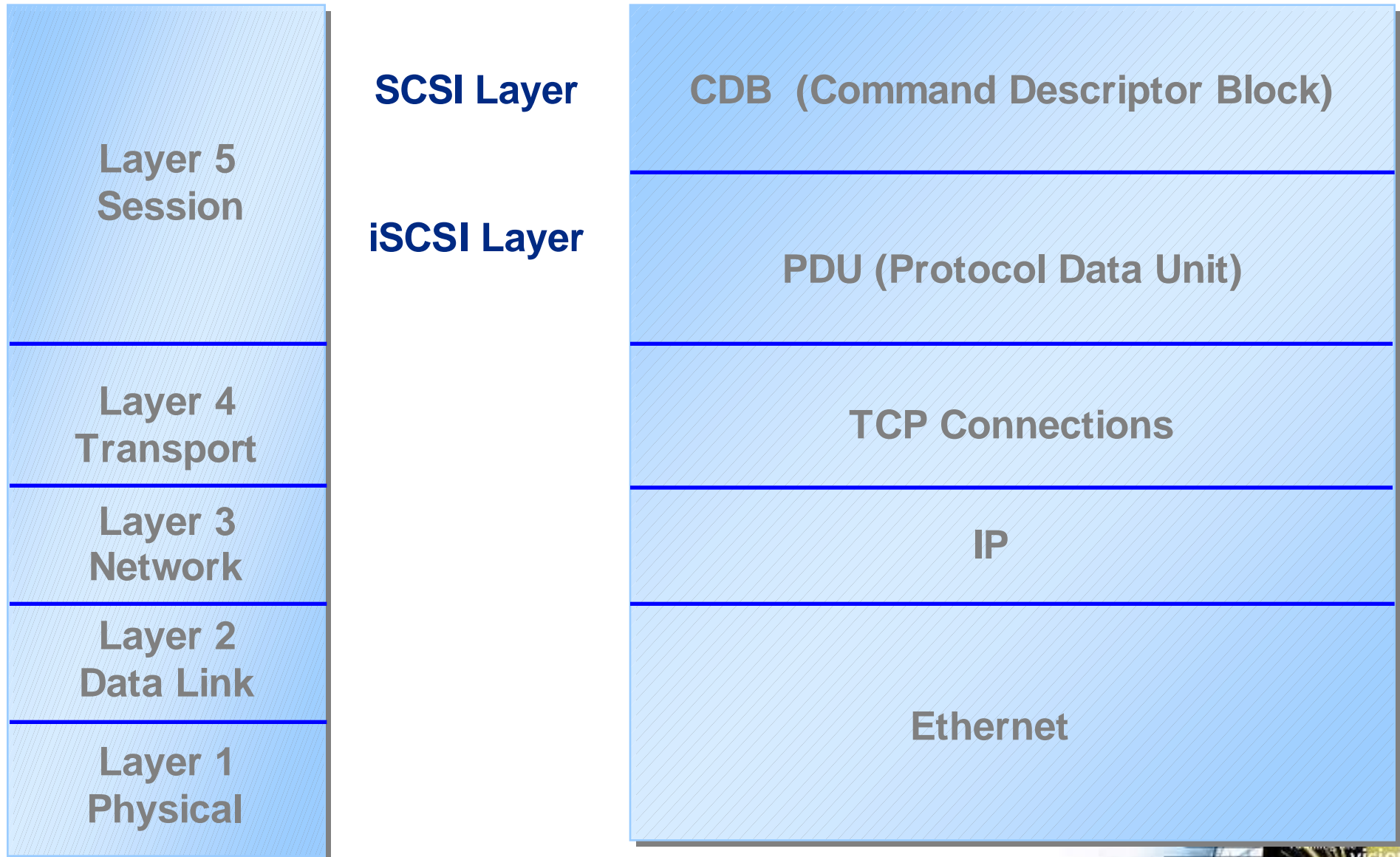


# iSCSI Commands

- SCSI Commands
  - Command phase
  - Optional data phase
  - Response phase
- iSCSI Commands
  - Binds command phase with associated data into iSCSI Protocol Data Unit (PDU)



# iSCSI Commands

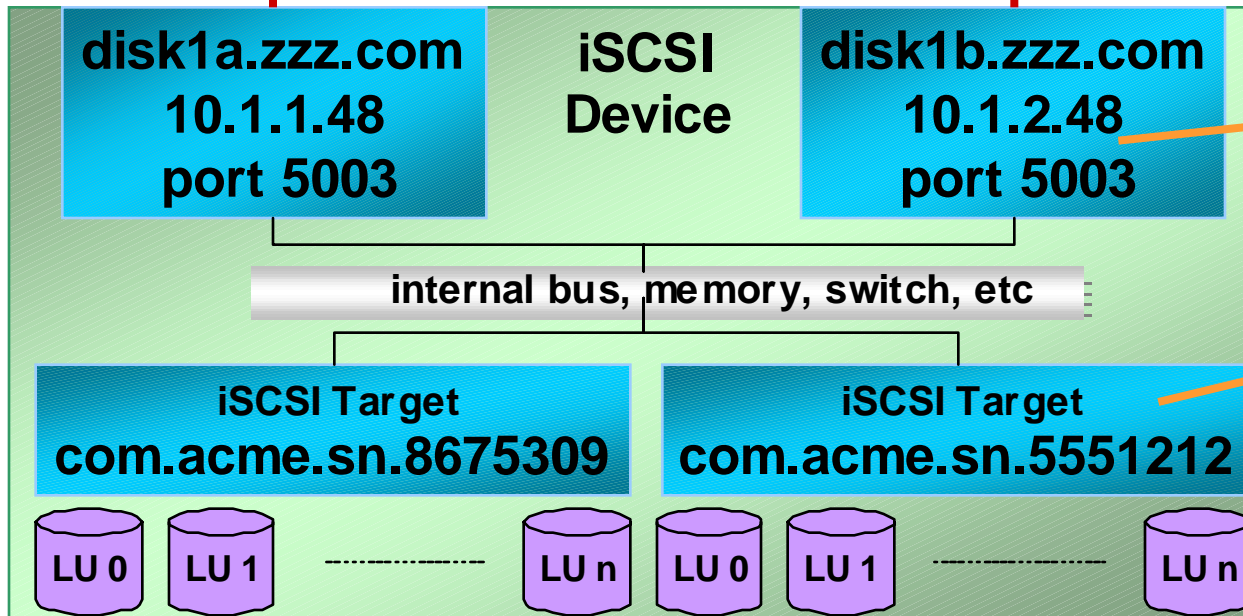
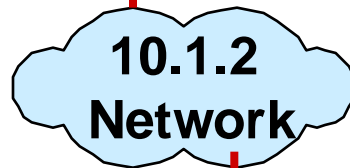
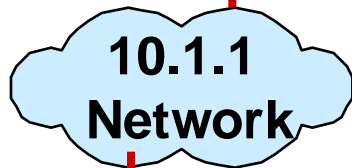


# iSCSI Names and Addresses



The iSCSI Name names the initiator, not the port

This initiator has two addresses



An iSCSI “port” is an IP Address + TCP Port.

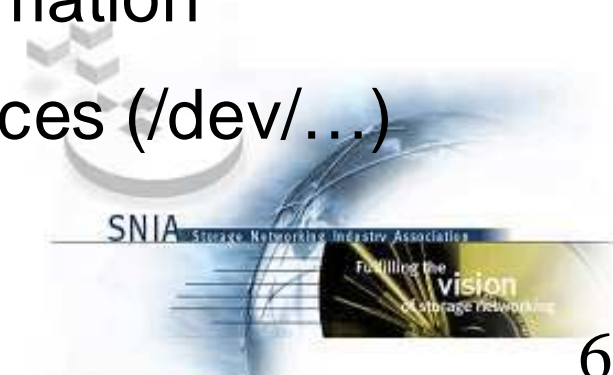
The iSCSI Name names the target, independent of the port on which it is accessed

Each iSCSI Target has its own logical units

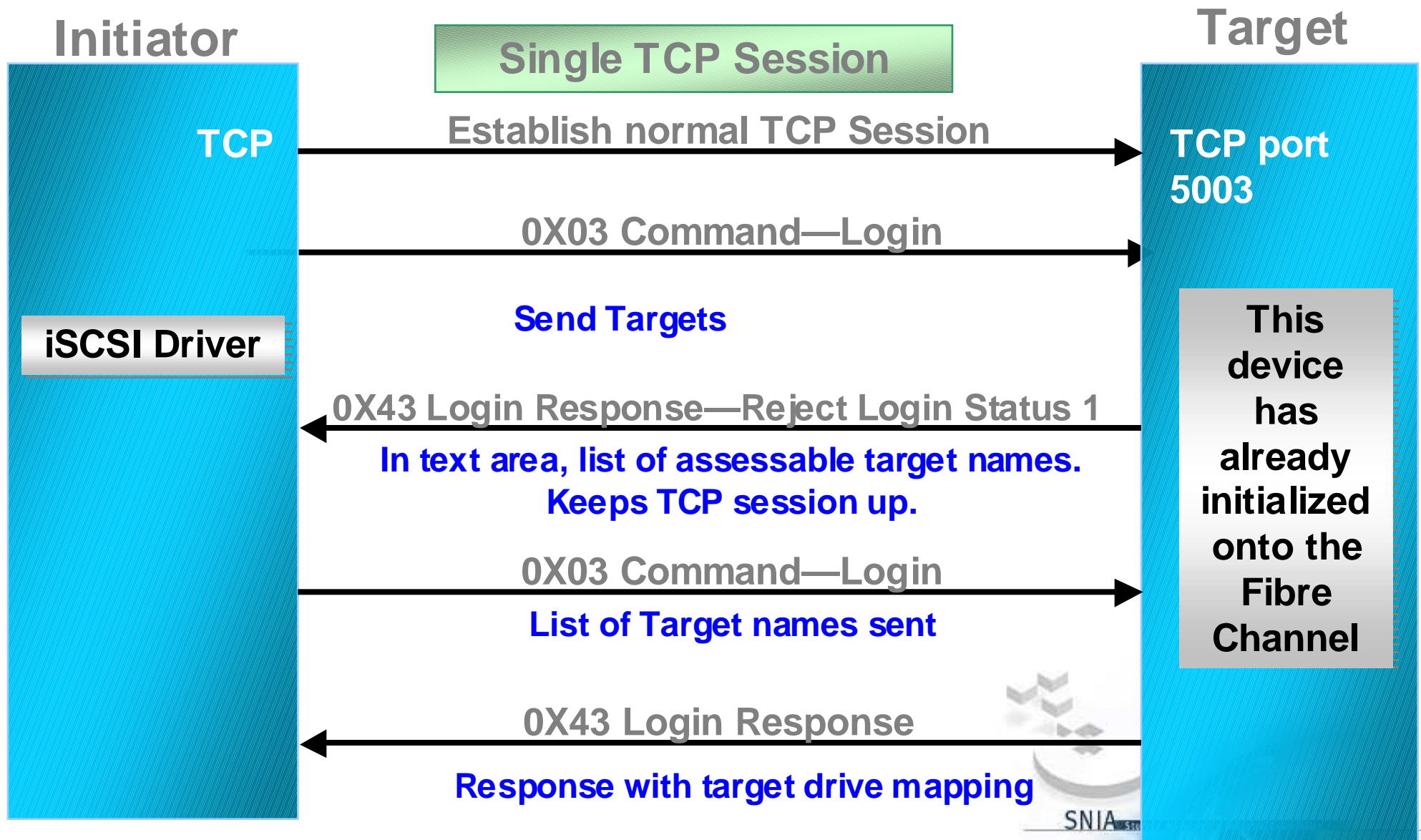


# iSCSI Storage Device Discovery Process

- 1) Host driver requests available iSCSI targets from the SCSI router
- 2) SCSI router sends available iSCSI target names to host
- 3) Host logs into iSCSI targets that were received
- 4) SCSI router accepts the login and sends target identifiers to Host (numbers)
- 5) Host queries targets for device information
- 6) Targets respond with device information
- 7) Host creates table of internal devices (/dev/...)



# iSCSI Sequence



# Discovery & Configuration

- Static configuration
  - Admin sets authorized iSCSI Target Node Name, & iSCSI Address
    - Can receive a "Target has Moved" Redirect Response
  - Admin set only iSCSI Address of Connoical Port (iscsi) of Target
    - Use Send – Targets to get list of known targets
      - Returned iSCSI Target Node Names
      - iSCSI Addresses for each
      - Alias for each iSCSI Target Node Name
    - Can be used to have single admin location for simple networks



# Discovery & Configuration

- Dynamic configuration via Initiator use of SLP to find Targets
  - iSCSI controllers can register iSCSI Node Name, iSCSI Address, Alias
  - Initiators can query to get information
  - Admin tells Targets what Initiators are permitted



# Discovery & Configuration

- The iSNS maybe used to contain all the information and Domains
  - Initiators can use SLP to locate iSNS
  - Initiators can use iSNS to tell it all the SCSI Devices it is authorized too
  - Can only query for Devices within own domain
  - Can send async state change notifications (uses Heartbeats etc.)
  - Can hold iSCSI Node Names, iSCSI Addresses, Aliases, Public Keys, etc.



# SLP General Model

- Registered Info
  - iSCSI Node name
  - TCP/IP address
  - Alias



- User Agent

- Directory Agent

- Server Agent

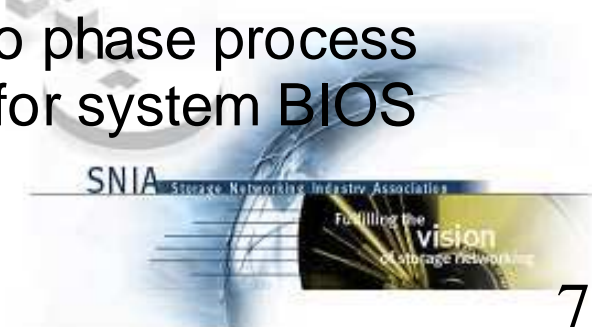
- Advertises info to DA
- May optionally carry info if DA available

SNIA Storage Networking Industry Association

Full vision available

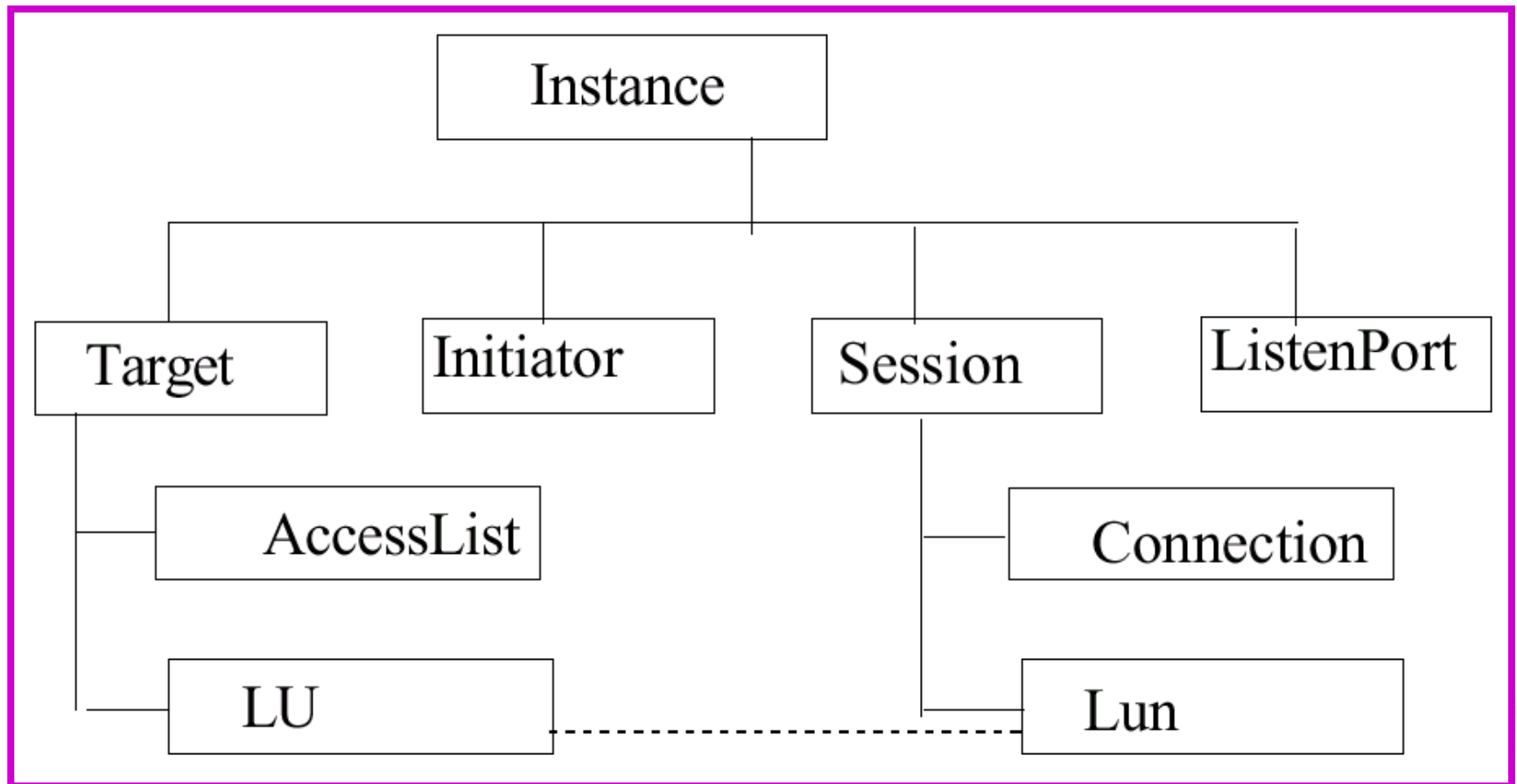
# iSCSI Boot

- Static configuration information for Boot
  - Admin sets authorized iSCSI Node Name and iSCSI Address, Optional LUN
    - Default LUN is 0
- Dynamic configuration via use of DHCP or SLP
  - DHCP be used by Host to get an IP address
  - DHCP can hold the iSCSI Boot Service Option (Admin Set)
    - May contain all that is needed to reach the Boot device
    - May only contain iSCSI Target Node Name, then use SLP to resolve to iSCSI address
  - W/O DHCP information, SLP can be used to find Boot "Service"
- The Boot load process
  - The Admin. or DHCP or SLP can enable the access
  - BootP is also possible as part of a SW two phase process
  - HW HBA can act as a normal SCSI HBA for system BIOS use



# iSCSI MIB Structure

- Handles multiple targets
- Tracks LUs and LUNs in use



# iSCSI Security Considerations

- A basic level of end to end data integrity can be reasonably handled by TCP using the standard checksum
- iSCSI will provide other integrity options leveraging IPsec
  - Security Models
    - No Security
    - Authentication between Initiator and Target
    - ✓ Authentication and Data Integrity (n-bit CRC)
    - Authentication and Data Encryption
- Security is primarily a WAN issue



# iSCSI Security Levels

- 0: None – ok in controlled environments
- 1: Initiator and target authentication
  - Prevents unauthorized access
- 2: Digests for header and data integrity
  - Prevents against man-in-middle, insertion, modification and deletion
- 3: Encryption (IPSEC)
  - Prevents against eavesdropping



# Ordering & Numbering

- Unlike // SCSI, iSCSI PDUs may
  - Arrive out of order (by taking different routes)
  - Not arrive at all
- iSCSI requires
  - Command numbering
    - Ordered delivery over multiple connections
  - Status numbering
    - Detection of a failed connections
  - Data sequencing
    - Detection of missing data PDUs



# Error Handling & Recovery

- // SCSI errors incur costly recovery:
  - Aborted commands; target, bus and host resets
  - OK, because bus errors are infrequent
- iSCSI errors will be more frequent
  - Link failures
  - TCP failures
  - Bad “middle box” (firewall, NAT, router)
  - Does the Internet have a “reset” option??



# IETF Solution

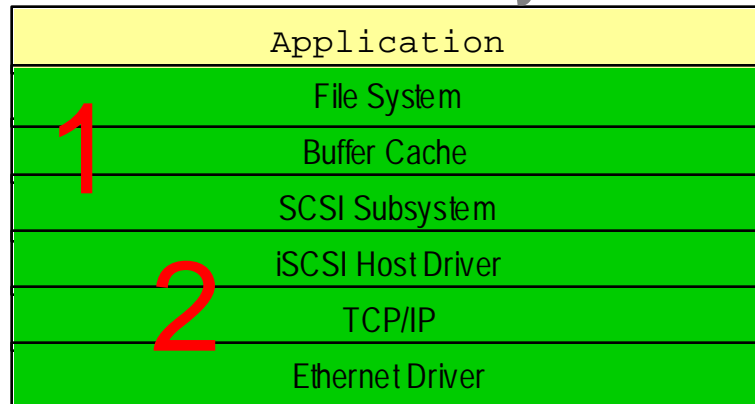
- Need mechanisms below the SCSI layer
  - Separate digests for header and data
    - 64 bit CRC and checksum are not enough
  - Retransmission of data
    - Data PDUs missing or fail integrity check
  - Reestablished TCP connections
- SCSI mechanisms take over when these fail



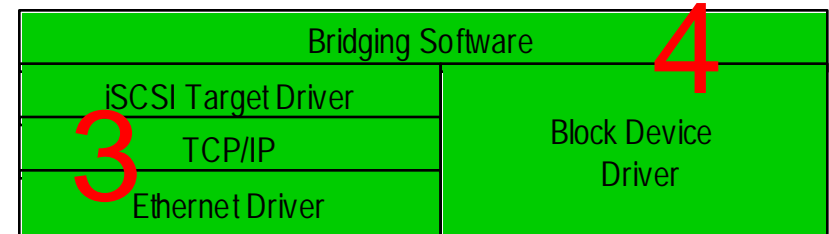
# Challenge 1 - TCP Overhead

Consider a SCSI WRITE command. How many times do you think the data is copied before eventually reaching the target HBA?

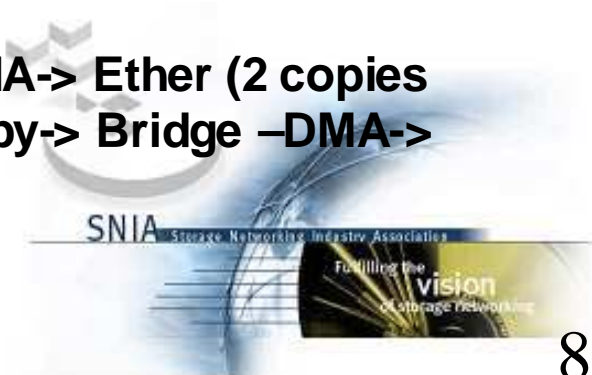
## Linux Host System



## Linux Target System



Application –copy-> Buffer Cache –copy-> TCP/IP –DMA-> Ether (2 copies 1 DMA) Ether –DMA-> Ring Buffer –copy-> TCP/IP –copy-> Bridge –DMA-> HBA (2 copies 2 DMA)



# TCP Overhead (2)

- TCP Processing
  - Every TCP connection that is part of an iSCSI session has processing overhead potential
    - Connection setup / teardown
    - TCP state machine:
      - Acknowledge, Timeout, Retransmission
      - Window management
      - Congestion Control
    - TCP segmentation
    - IP fragmentation
    - Checksum calculations
- Partial or Complete TCP Offload mechanisms are assumed to be required to make iSCSI performance comparable to FC



# Challenge #2 – Framing

- Message Boundaries (The Framing - HW-Issue)
  - iSCSI messages have no alignment relationship with TCP segments
  - And TCP does not have a “built in mechanism” for signaling message boundaries.
    - IETF considered leverage the urgent pointer for some time
  - So how can an iSCSI adapter determine where a message begins and ends??
    - By reading the length field in the iSCSI header
      - Determines where in byte stream current message ends and next begins
      - NIC must stay “in sync” with beginning of byte stream
      - Works well in a perfect world (Maybe a SAN or LAN ????)
    - In a MAN/WAN we have issues
      - IP Frags leading to out-of-order packet delivery and/or packet loss
      - Any “middle box” may fragment an IP packet until, sending each along potentially different routes



# Framing (2)

- Message Boundaries Continued
  - THE SCENARIO:
    - An iSCSI header is not received when expected because the TCP segment that it was part of was delivered out of order
  - THE ISSUE:
    - The receiver does not know where to put the trailing data packets until the packet with the header arrives
  - The different options?
    - Drop all packets until the header arrives
      - They will be retransmitted
    - Buffer packets until the header arrives. Then “re-assemble.”
      - On a 1Gbit WAN link, 16MB of buffer memory is required per TCP connection
      - On a 10 Gbit WAN link, 125MB of buffer memory required per TCP connection



# Framing (3)

- Message Boundaries Continued
  - THE BAD NEWS:
    - Dropping packets greatly impacts performance and significantly increases network congestion
    - Local buffering is expensive and NIC logic is complex



# Framing (3)

- Message Boundaries Continued
  - THE GOOD NEWS:
    - Rare occurrence on a SAN – where performance is most critical.
    - The IETF is aware of this issue and working diligently
    - Stream Markers Proposal
      - Regularly spaced markers give offset to next header allowing faster synchronization
      - Significantly reduces buffering requirements
    - WARP Proposal
      - Self-describing RDMA “chunks” contain address and offset of destination address
      - No need to wait for header, we already know where the data goes
    - SCTP (10GIG)
      - WARP was actually derived from SCTP
      - Contained general RDMA and framing mechanisms
      - Why not just use SCTP instead of TCP??
    - Who’s doing SCTP offload??



# Networking Overhead

- Software iSCSI *can* achieve GbE wire speed
  - but at 100% CPU
- Traditional TCP stacks are expensive
  - multiple memory copies
  - too many interrupts
  - checksums calculations
- We need TCP offload engines (TOE)



# TOE

- The challenge rests on the TOE vendor
  - Interrupt host on command boundaries
  - Offer zero-copy from NIC to app
  - Eliminate TCP reassembly buffer
    - Provides *true* zero-copy
    - Requires RDMA or synchronization
- Proposed IETF solutions for framing
  - WARP - an RDMA mechanism
  - Markers – a synchronization mechanism



# What's Next for iSCSI

- CRC
- SLP (Service Location Protocol)
- Authentication
- Encryption



# iSCSI Message Types

## ■ Initiator to Target

- NOP-out
- SCSI Command (encapsulates a SCSI CDB)
- SCSI Task Management Command
- Login Command
- Text Command
- SCSI data (Write)
- Logout Command

## ■ Target to Initiator

- NOP-in
- SCSI Response (can contain status)
- SCSI Task Management Response
- Login Response
- Text Response
- SCSI data (Read)
- Logout Response
- Ready to Transfer (R2T)
- Async Event



# iSCSI Participating Companies

- Adaptec
- Agilent Technologies
- Aristos Logic Corporation
- Brocade
- Cereva Networks
- Cisco
- CNT Corporation
- Emulex Corporation
- Eurologic Systems
- FalconStor Software
- Hitachi Data Systems
- Hewlett Packard
- IBM
- Intel Corporation
- Legato
- LSI Logic Corporation
- Lucent
- NetConvergence
- Nishan Systems
- Pirus Networks
- Qlogic
- Rhapsody Networks
- SAN Valley Systems
- SANcastle
- San-Stor
- Spectra Logic
- StoneFly Networks
- StoreAge Networking
- Sun Microsystems
- Troika Networks



Source: SNIA iSCSI Group Members press release, 5/10/20

# Conclusions



# Conclusions

- IP-based storage will proliferate
- Benefits are strong
- Significant players
- Clear need
- Standards will be established
- Work with industry leaders



# Acknowledgements

- Information extracted from material and presentations provided by SNIA, Adaptec, Brocade, Cisco Systems, Dragon Slayer Consulting, Emulex, HP, IBM, Intel, NetConvergence, Nishan and SAN Valley.



# Additional Information

- **Storage Networking Industry Association** <http://www.snia.org>
- **Internet Engineering Task Force** <http://www.ietf.org>
- **Fibre Channel Industry Association** <http://www.fibrechannel.org>
- **ANSI** <http://www.ansi.org>
- **SCSI Trade Association** <http://www.scsita.org>
- **Fibre Channel Alliance** <http://www.fibrealliance.com>
- **Host and Target Mode Drivers** <http://www.sourceforge.net/projects/intel-iscsi>

