SNIA

STORAGE NETWORKING INDUSTRY ASSOCIATION

EDUCATION

IP Storage Protocols: iSCSI

John L. Hufferd, Sr. Technical Staff Member, IBM Corporation
Ahmad Zamer – Storage Technology Initiatives Manager, Intel
Will Akin, Principal Storage Architect, ICG/CTO, Intel

Abstract
IP Storage Protocols: iSCSI

- **This session will explain the various parts of iSCSI**
  - Network encapsulations of iSCSI PDUs
  - Session Relationship to SCSI and TCP/IP Connections
  - iSCSI flow from Initiator to Target
  - Error Recovery, Discovery and Security
- **It will also explain Companion Processes**
  - Boot
  - SLP
  - iSNS
- **And the session will describe iSCSI Environments**
  - From the small office, to the High End Enterprise
- **This session is appropriate for end user and developers of iSCSI technologies**

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced without modification
  - The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.

Terms

- iSCSI - Internet SCSI
- NAS - Network Attached Storage
- HBA - Host Bus Adapter
- TOE - TCP/IP Offload Engine
- FC - Fibre Channel
- SAN - Storage Area Network
- iSAN - iSCSI Storage Area Network
- PDU - Protocol Data Unit
- WWN - World Wide Name

Agenda

- Introduction
- iSCSI Features
  - Error handling, Boot, Discovery
- iSCSI usage models
- IP Security
- Q & A

Agenda

- **Introduction**
- iSCSI Features
  - Error Handling, Boot, Discovery
- iSCSI usage models
- IP Security
- Q & A

Small Computer System Interconnect (SCSI)

SCSI BUS

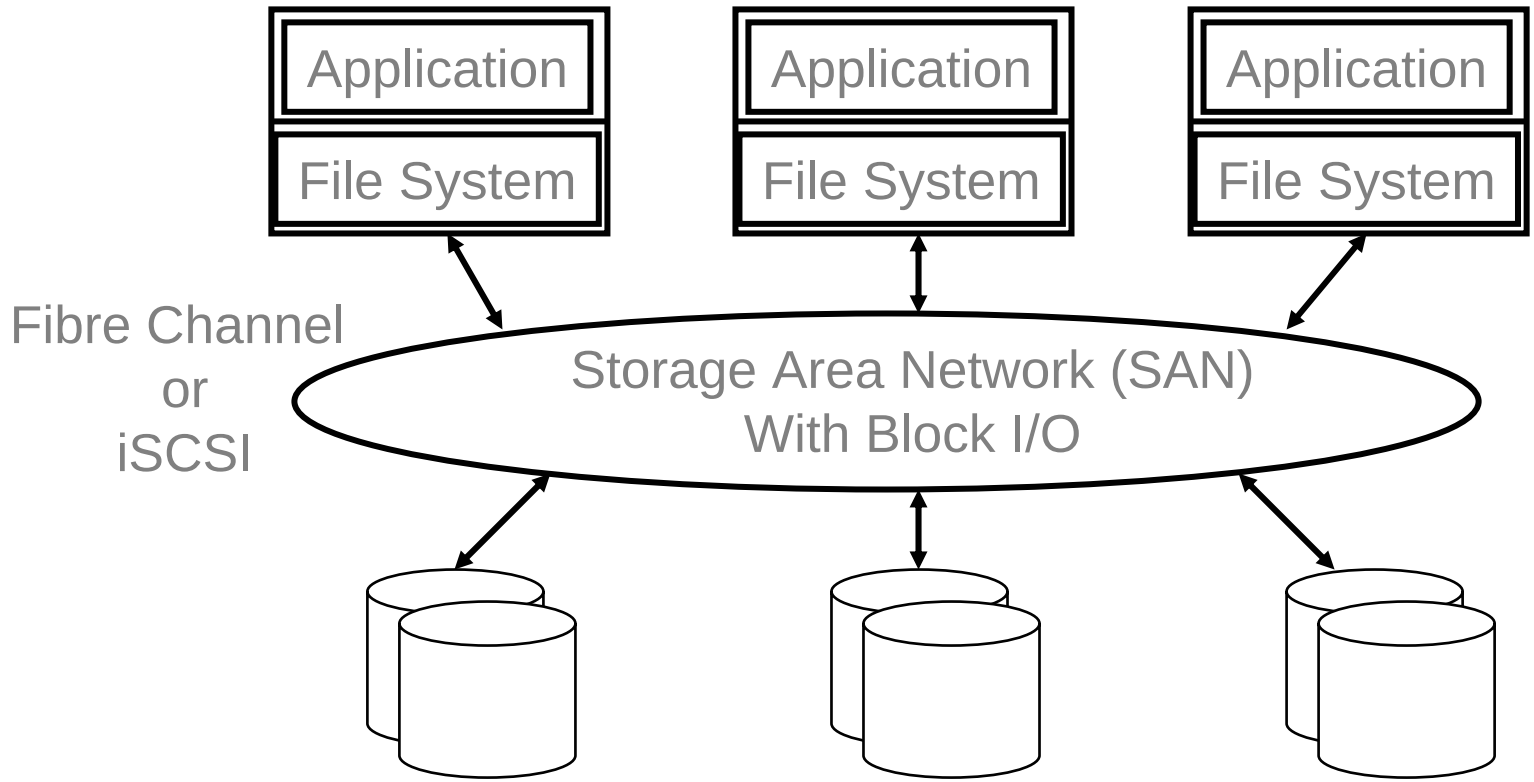
Tapes

Desktop / Server Computer

SCSI Disks

There are 3 main hard drive types available today:
ATA (used mostly in desktop and laptop systems)
SCSI and FC (used in server-class systems)

Systems with SCSI over Networks



Both Fibre Channel and iSCSI can makeup a SAN

Replaces shared bus with switched fabric

iSCSI is:

- **Internet SCSI: i**nternet **S**mall **C**omputer **S**ystem **I**nterconnect
- 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
  - On Ethernet LANs: Copper & Optical
  - On ATM WANs
  - On SONET WANs
  - Wireless
  - Etc.

Data Encapsulation Into Network Packets

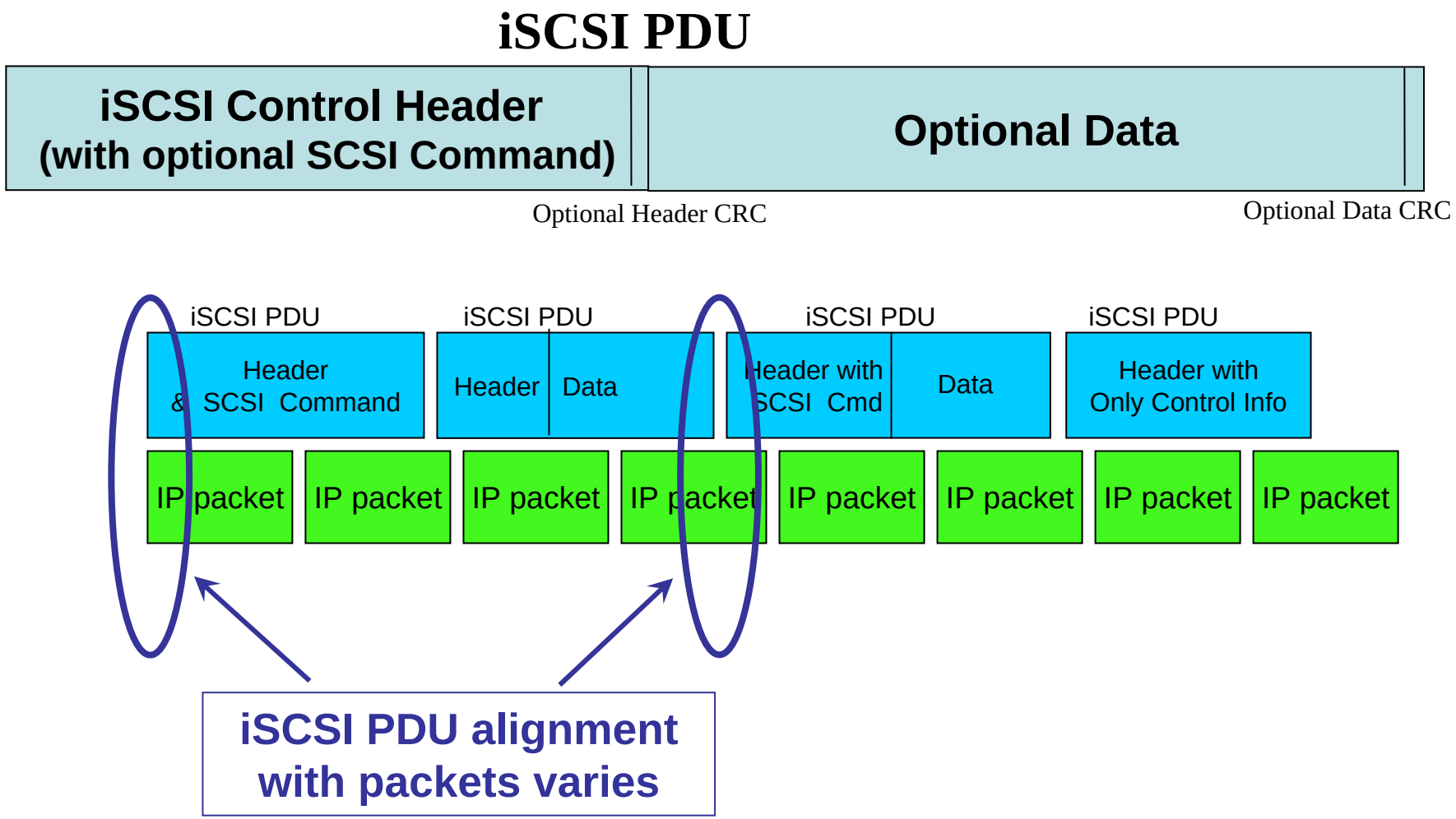
iSCSI Protocol Data Unit (PDU): Provides ordering and control information. Contains iSCSI control info, with optional SCSI Commands &/or Data

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

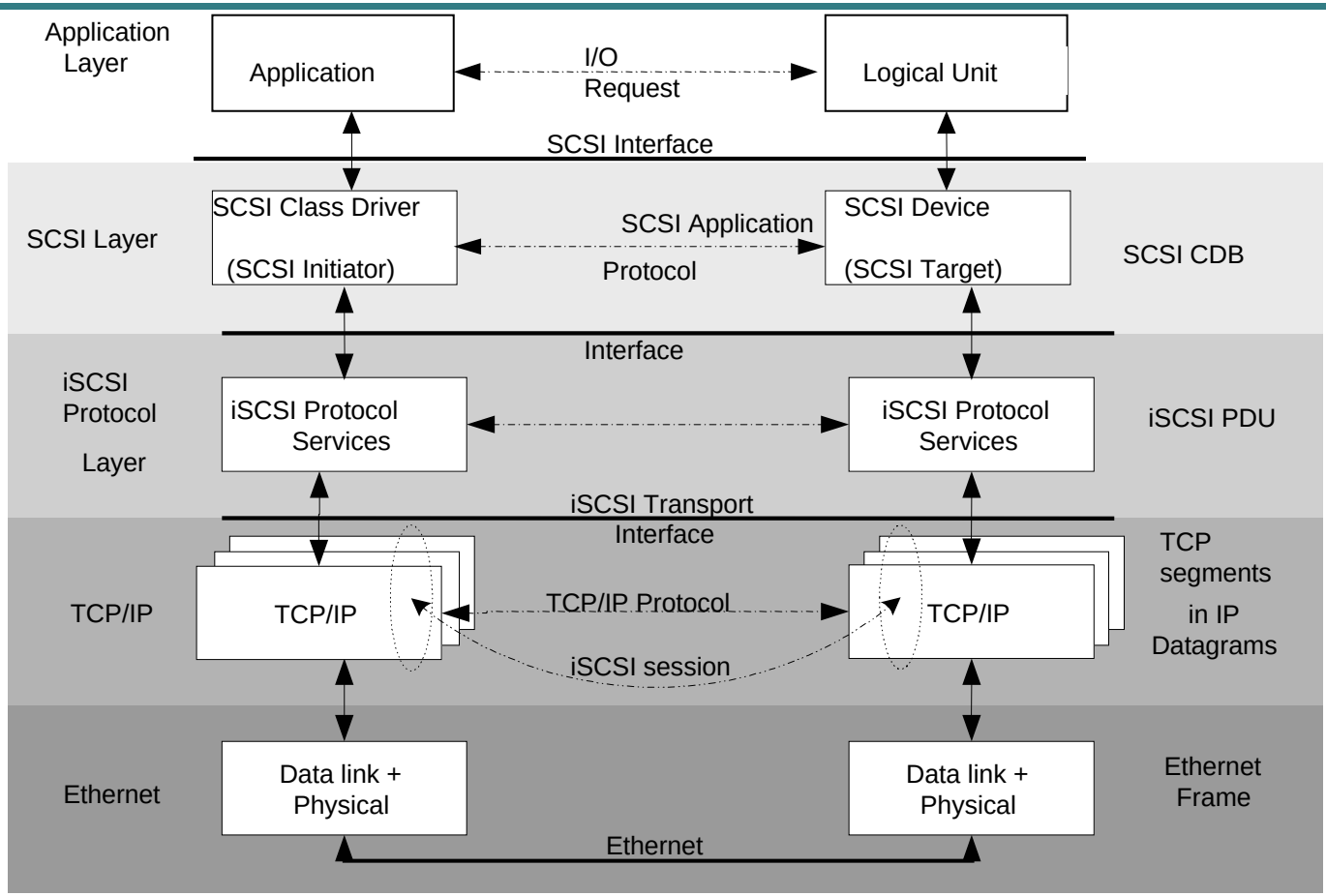
Provides IP "routing" capability so that packet can find its way through the network

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

iSCSI Mapping

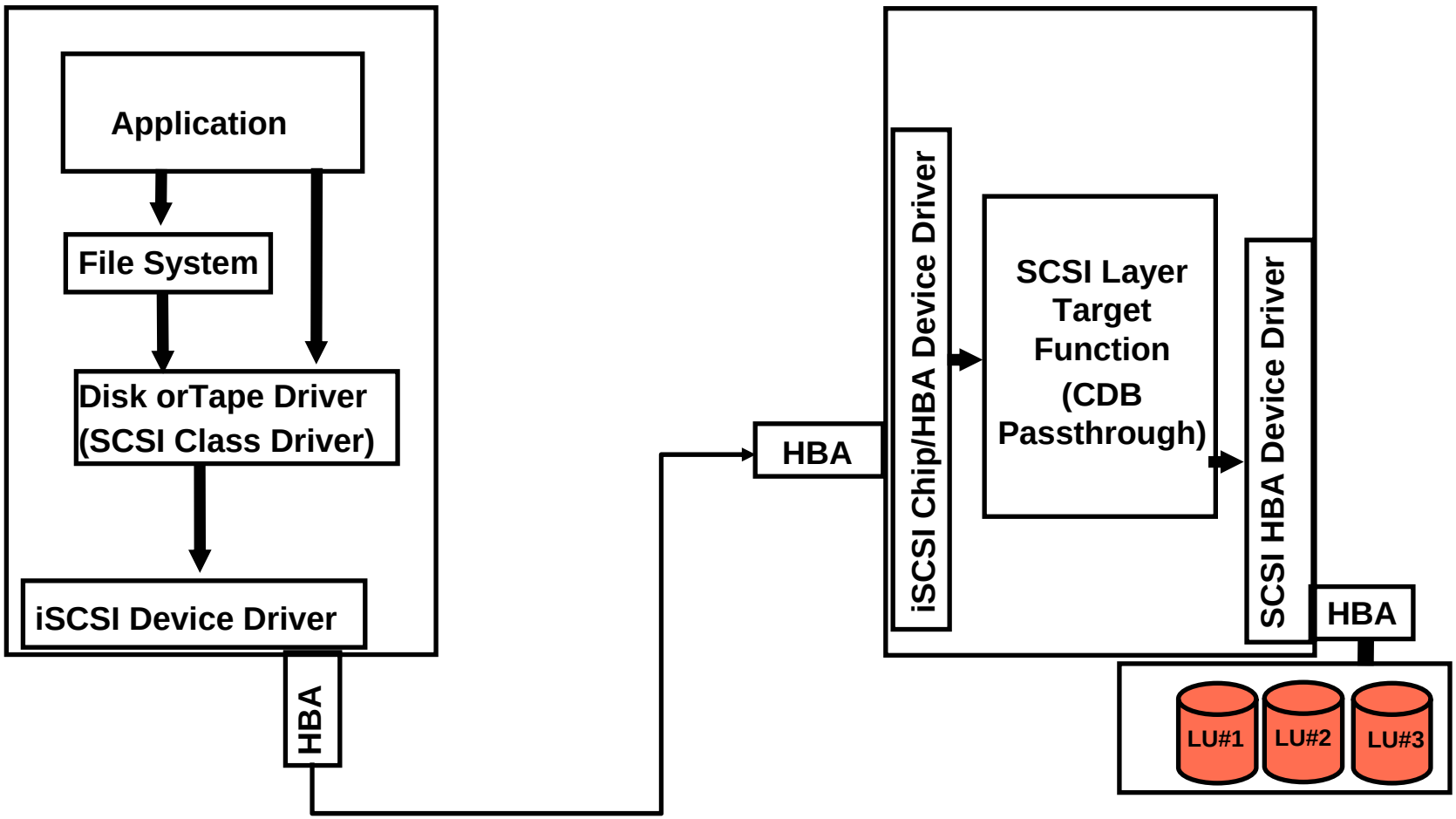


iSCSI - Layered Model



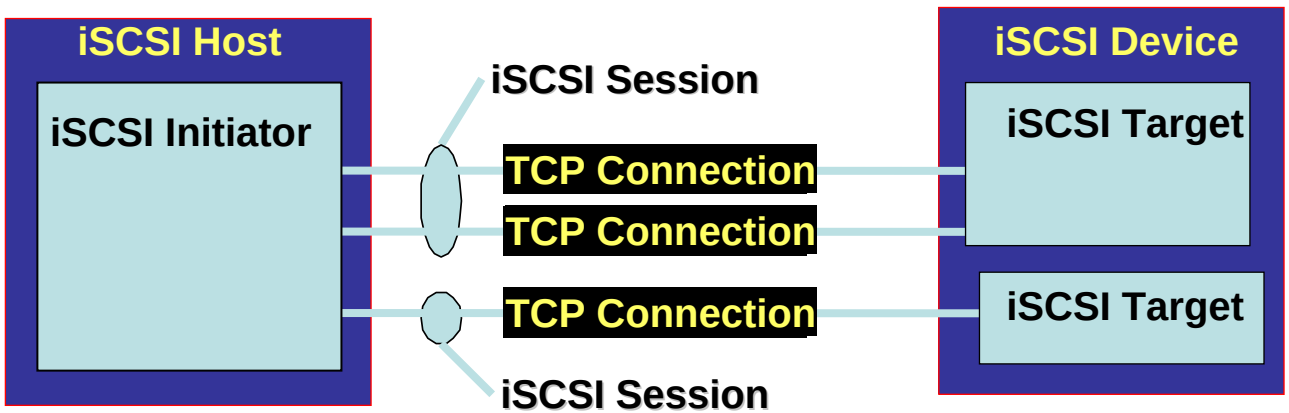
Transparently encapsulates SCSI Command Descriptor Blocks (CDBs)

Application to LU Command Flow

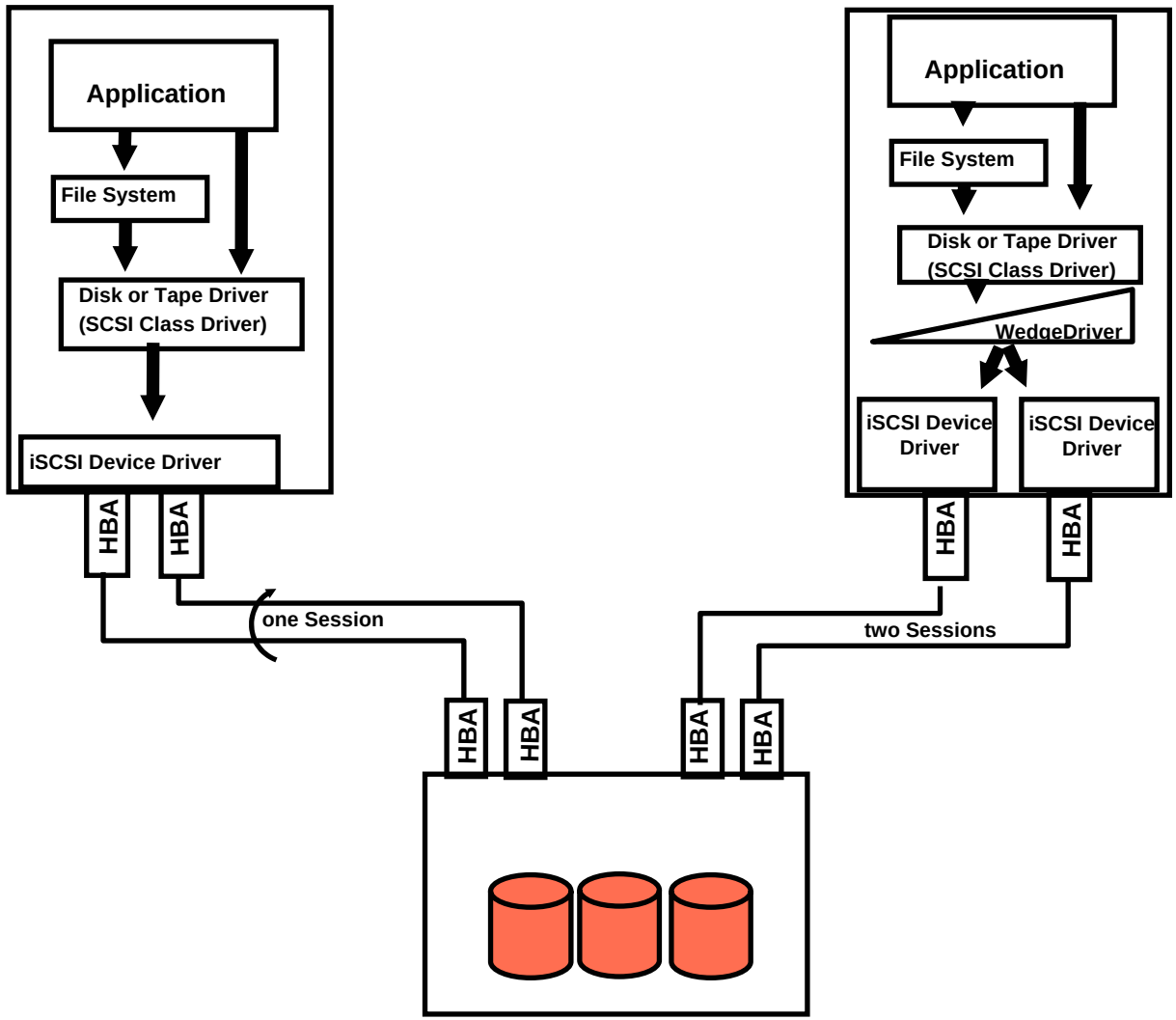


iSCSI Structure

- **iSCSI has the concept of a Session**
  - A Session can be thought of as a SCSI port made up of one or more TCP/IP connections
    - Login phase begins each connection
    - Is started after 1st Login is complete
    - Deliver SCSI commands in order
    - Can recover from lost connections



Multiple Connections Between Hosts and Storage Controllers



iSCSI Integrity

SNIA
Storage Networking Industry Association
EDUCATION

- **iSCSI adds Cyclic Redundancy Check (CRC)**
  - CRC-32C - A 32 bit check word algorithm
  - End to End Checking
  - In addition to TCP/IP Checksums
  - In addition to Ethernet Link level CRCs
- **CRC "check word" is called a "Digest"**
- **iSCSI Digests for iSCSI Headers and Data**
  - Header Digest is optional to use (MUST implement)
    - Insures correct operation and data placement
  - Data Digest is optional to use (MUST implement)
    - Insures data is unmodified through-out network path

iSCSI Message Types

Called Protocol Data Units (PDUs)

- Initiator to Target
  - NOP-out
  - SCSI Command
    - Encapsulates a SCSI CDB
  - SCSI Task Mgmt Cmd
  - Login Command
  - Text Command
    - Including SendTargets
      - Used in iSCSI Discovery
  - SCSI data
    - Output Data for Writes
  - Logout Command
- Target to Initiator
  - NOP-in
  - SCSI Response
    - Can contain status
  - SCSI Task Mgmt Rsp
  - Login Response
  - Text Response
  - SCSI data
    - Input Data from Reads
  - Logout Response
    - **Ready to Transfer**
    - **R2T**
    - **Async Event**

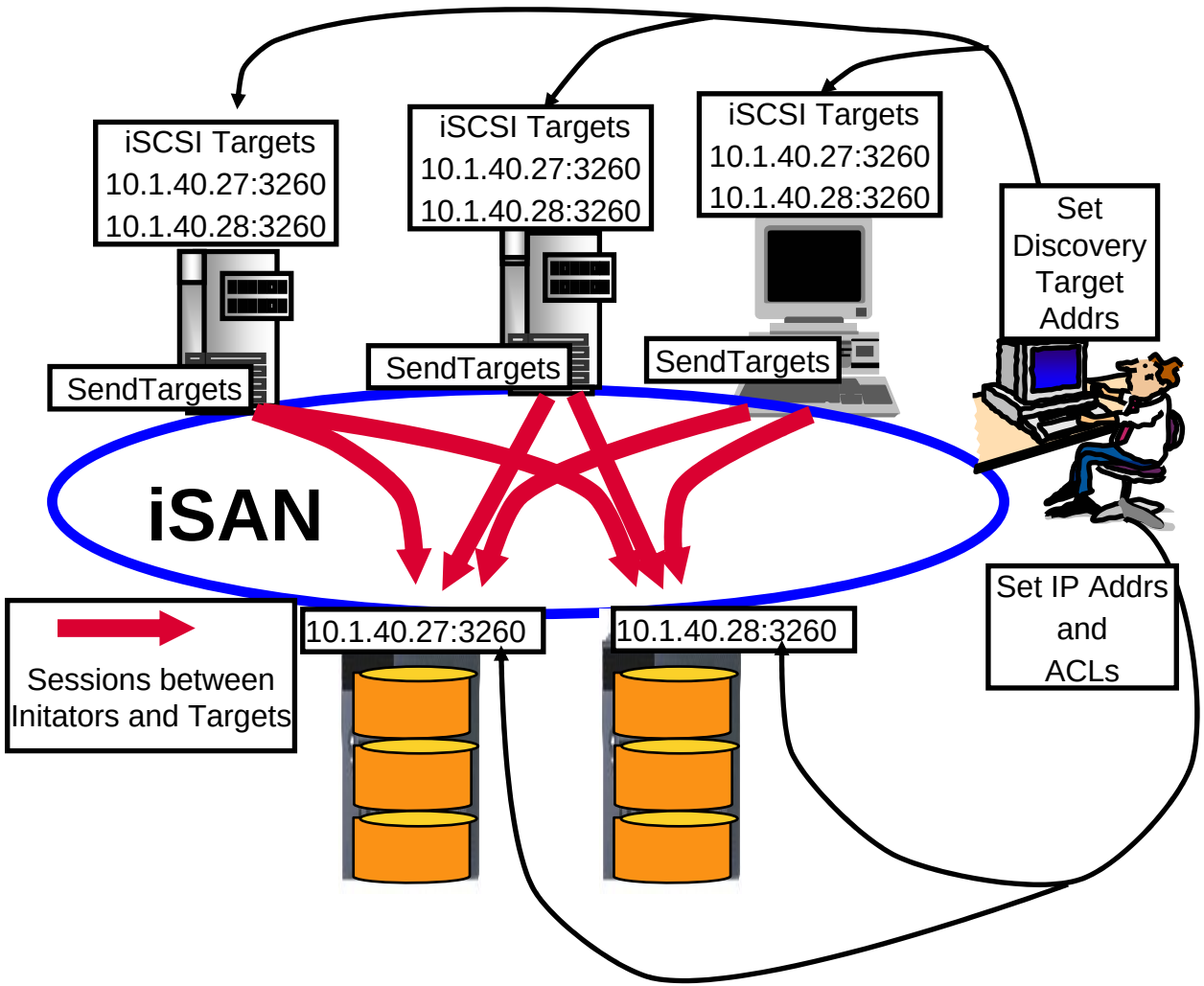
Agenda

- Introduction
- **iSCSI Features**
  - Error Handling, Boot, Discovery
- iSCSI usage models
- IP Security
- Q & A

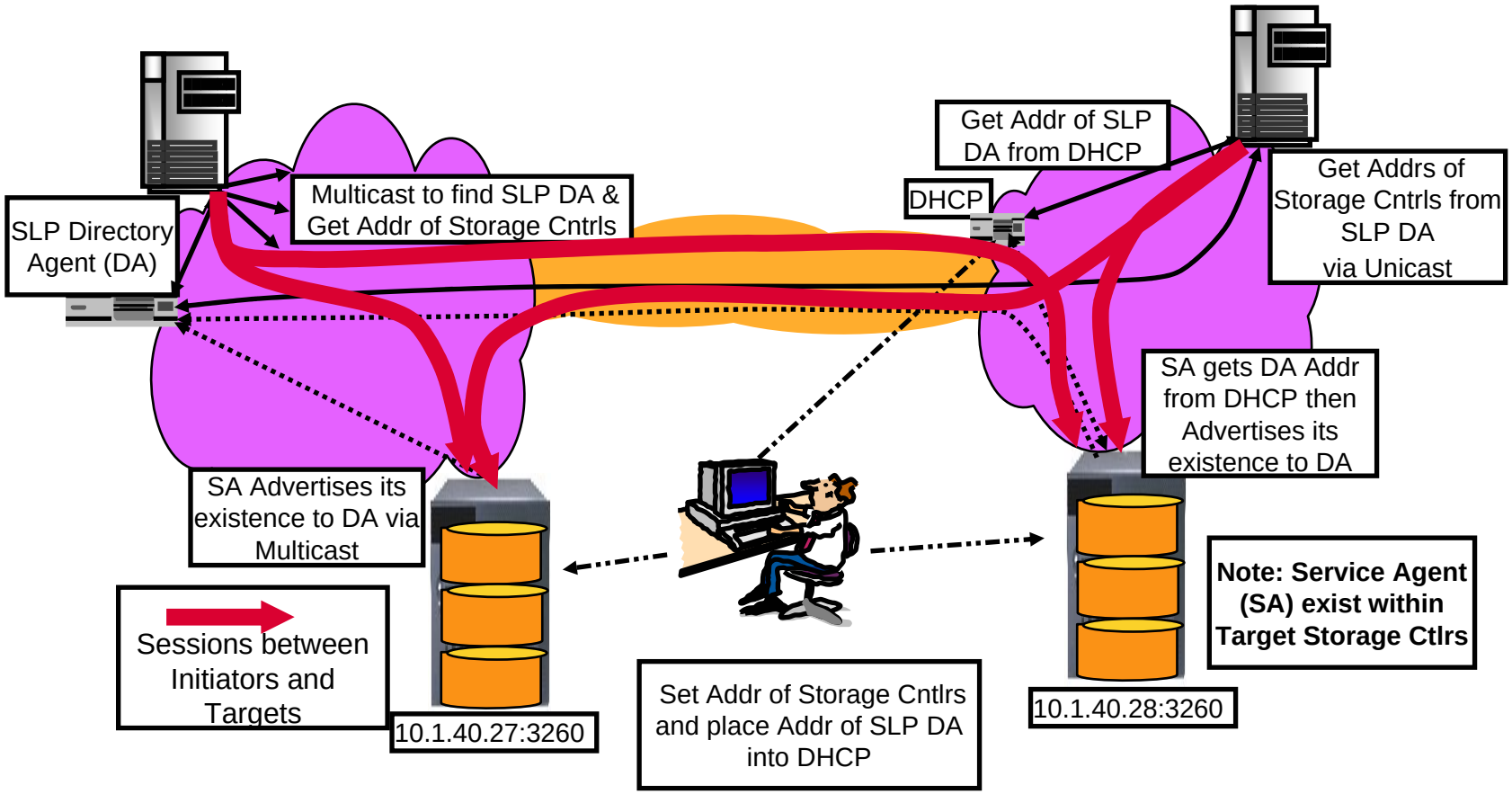
iSCSI Error Handling

- **ErrorRecoveryLevel = 0**
  - When iSCSI detects errors it will bring down the TCP connection and restart it
  - iSCSI will let the SCSI layer retry the operation
- **ErrorRecoveryLevel = 1**
    - Detected errors (Header or Data) causes PDUs to be discarded
  - iSCSI will retransmit discarded commands
  - iSCSI will retransmit discarded data
- **ErrorRecoveryLevel = 2**
    - Caused by loss of the TCP/IP connection
  - Connection & Allegiance reestablishment
  - Uses ErrorRecoveryLevel 1 to recover lost PDUs

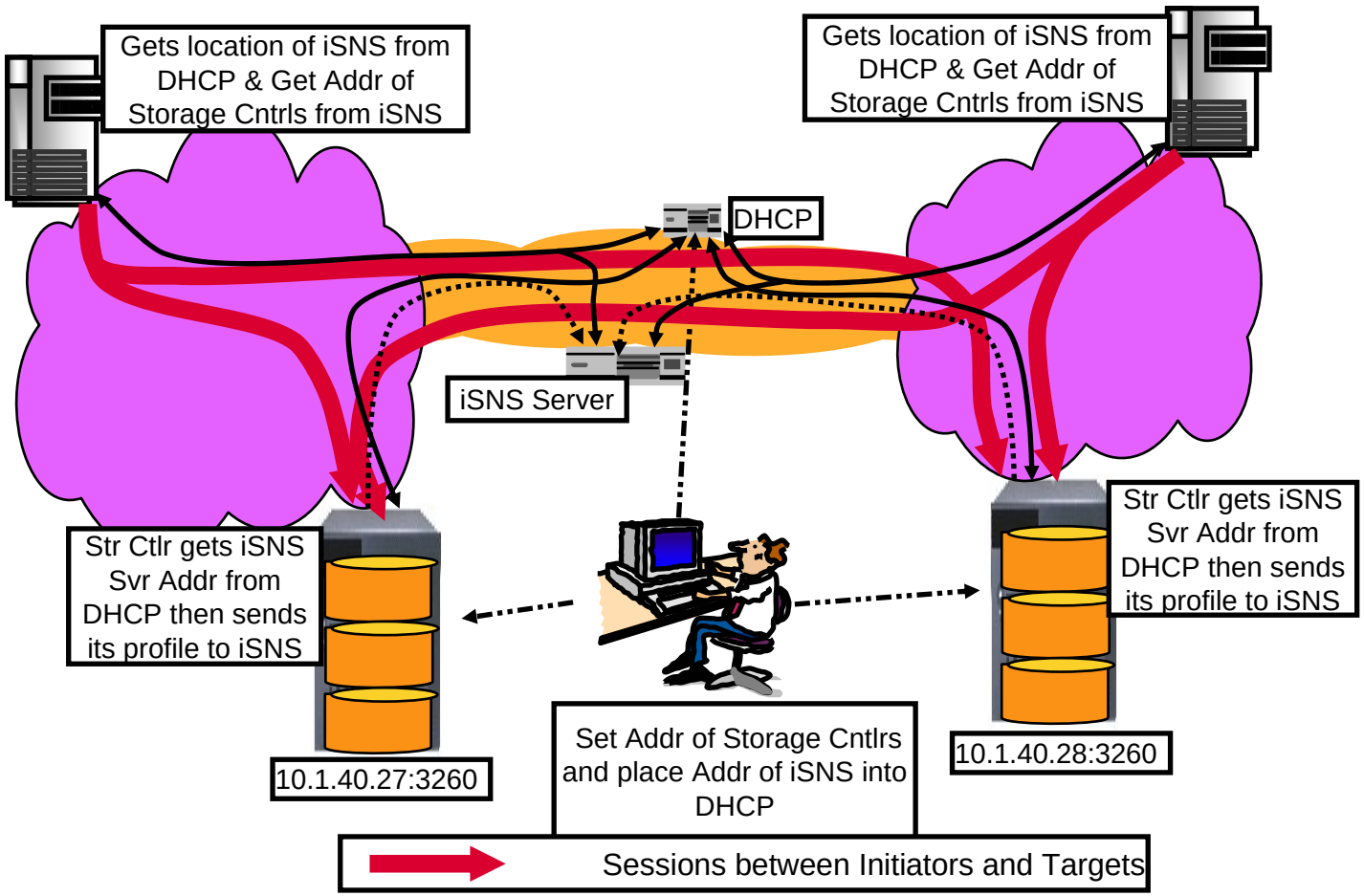
Discovery via SendTargets



Discovery via SLP



Discovery via iSNS



iSCSI Relocation

- **After attempting to Login at specified location:**
  - The specified Target may signal a relocation when Normal Login attempted
    - Temporary relocation
    - Permanent relocation
- **Relocation used for:**
  - Corrections between Discovery DB updates
  - Admin or automatic Hardware disablement
    - for Service
    - Because of HW problems
  - For load balancing

iSCSI Boot

SNIA
Storage Networking Industry Association
EDUCATION

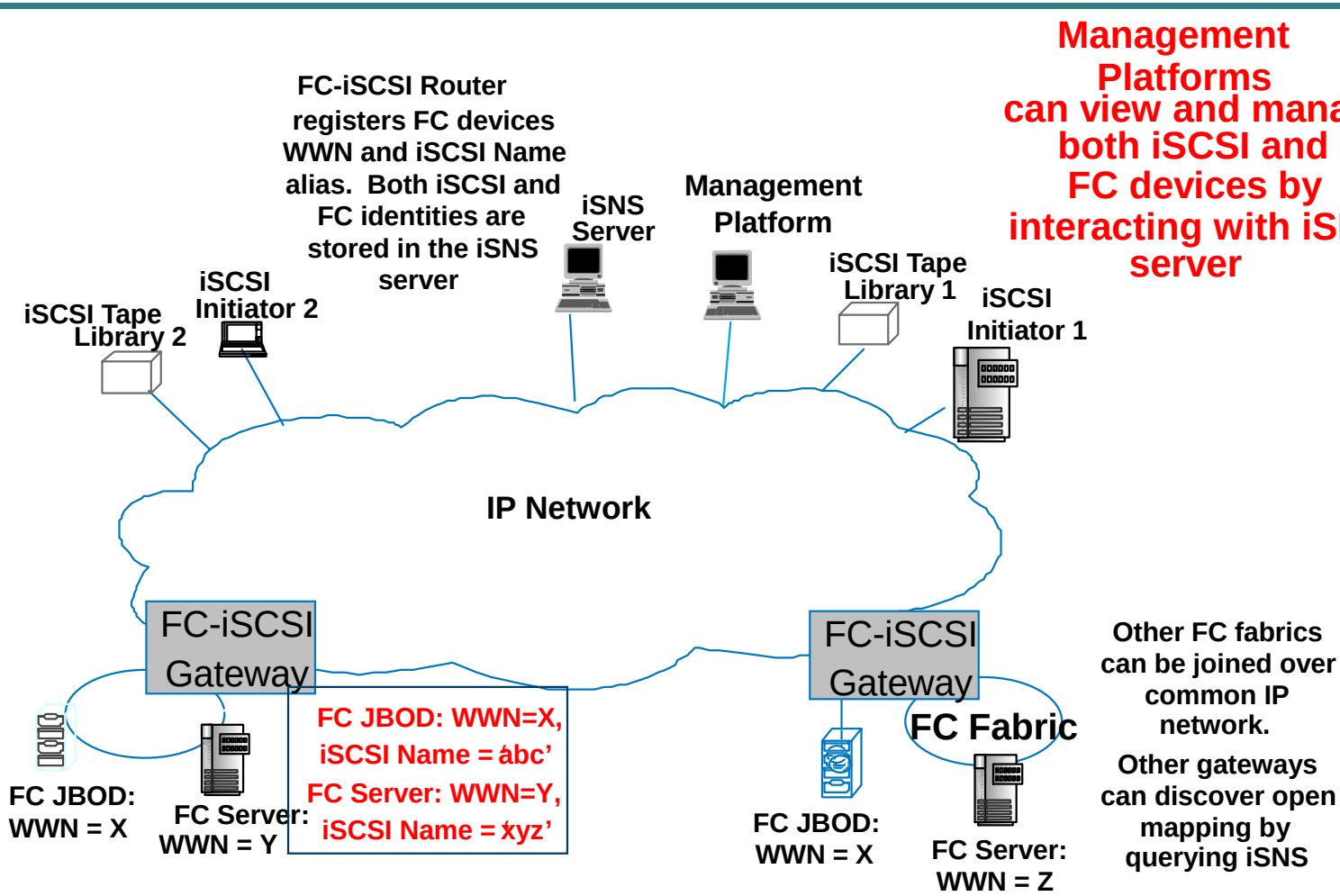
- **Static configuration information for Boot**
  - Admin sets authorized iSCSI Target Node Name and iSCSI Address, Optional LUN
    - Default LUN is 0
- **Dynamic configuration via use of DHCP, SLP, iSNS**
  - DHCP can be used by Host to get an IP address
  - DHCP can hold the iSCSI Boot Service Option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

Agenda

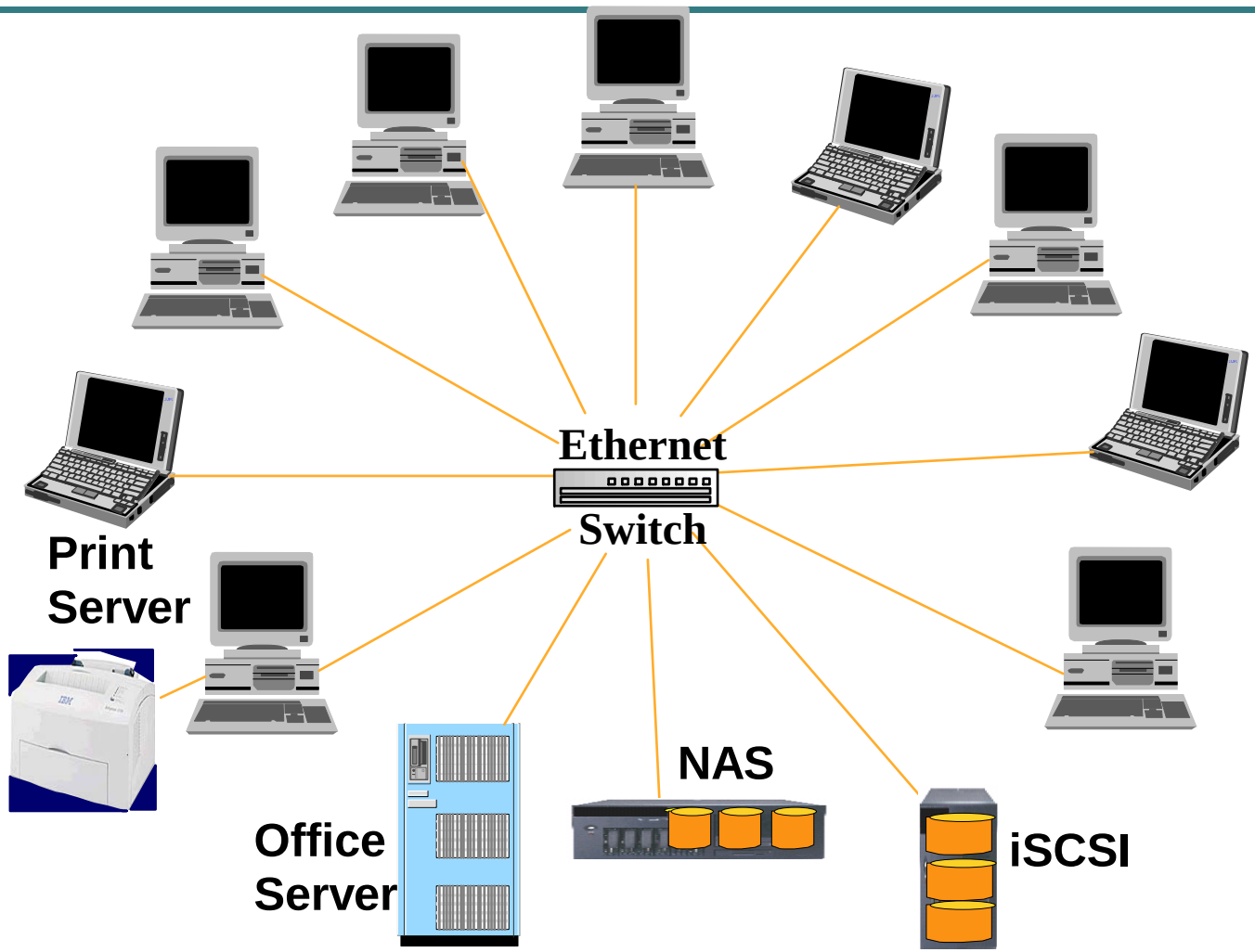
- Introduction
- iSCSI Features
  - Boot, Discovery, Error Handling
- **iSCSI usage models**
- IP Security
- Q & A

Now lets look at the various environments where iSCSI is appropriate

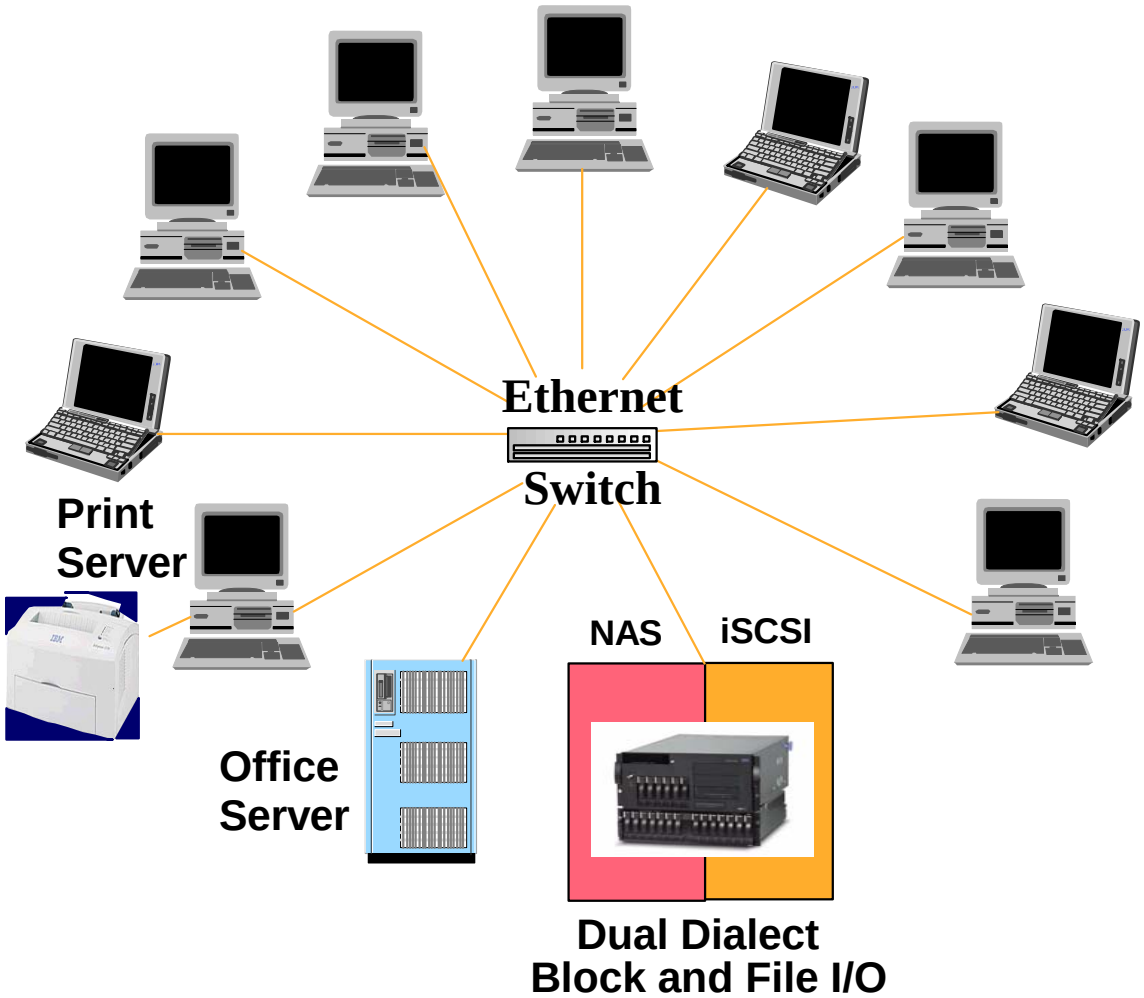
Combining of FC and iSCSI



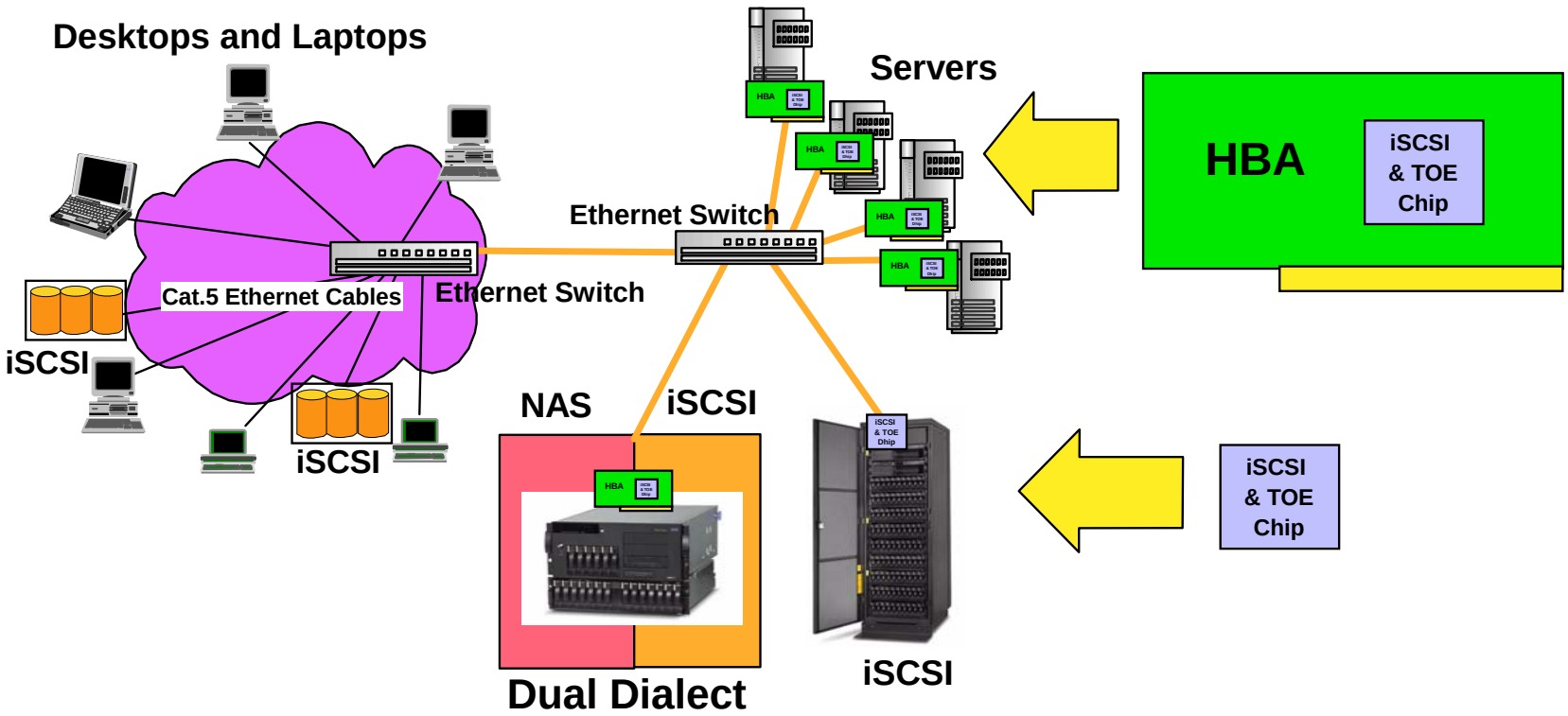
Small Office Interconnect



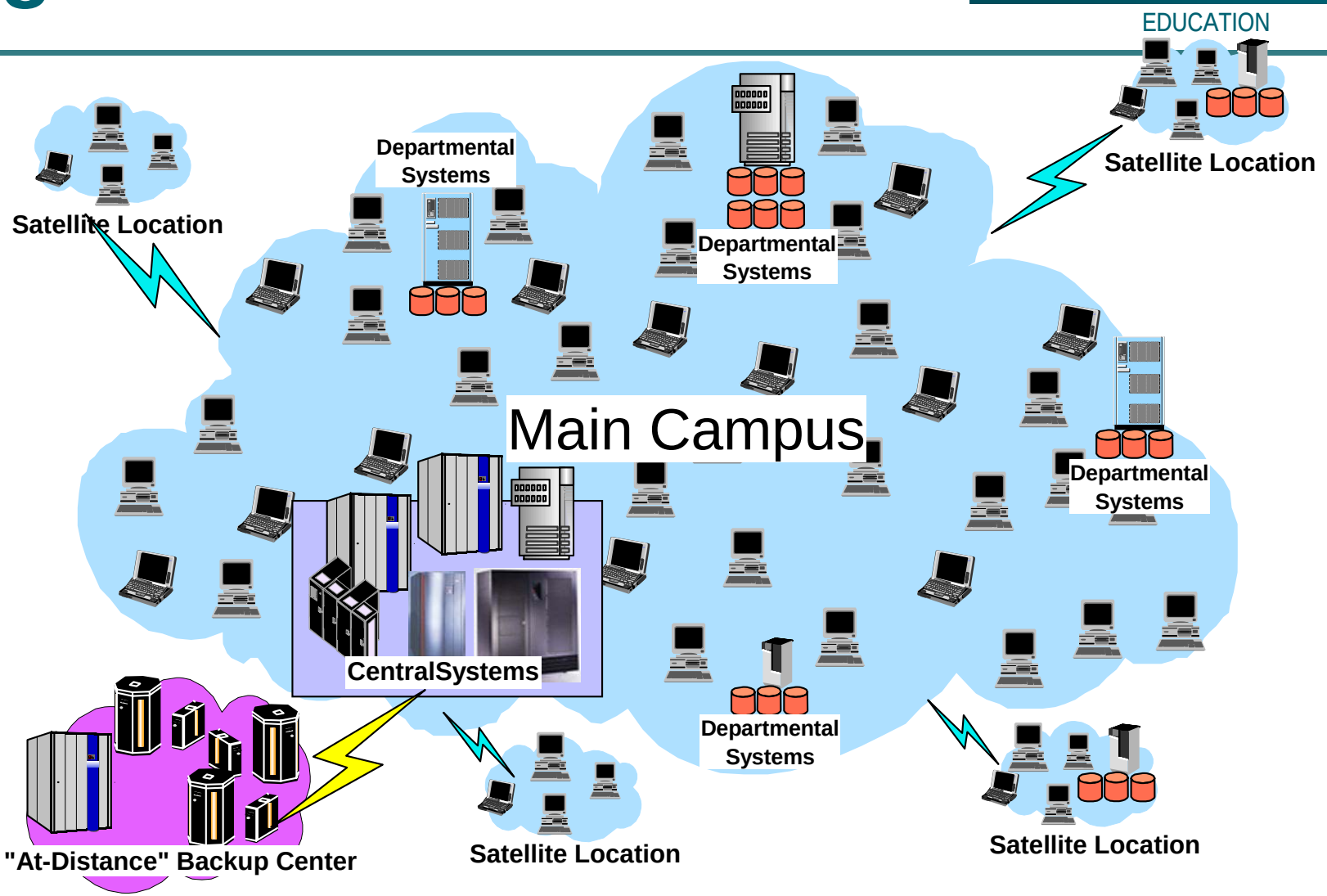
IP Storage Combo -- NAS & iSCSI



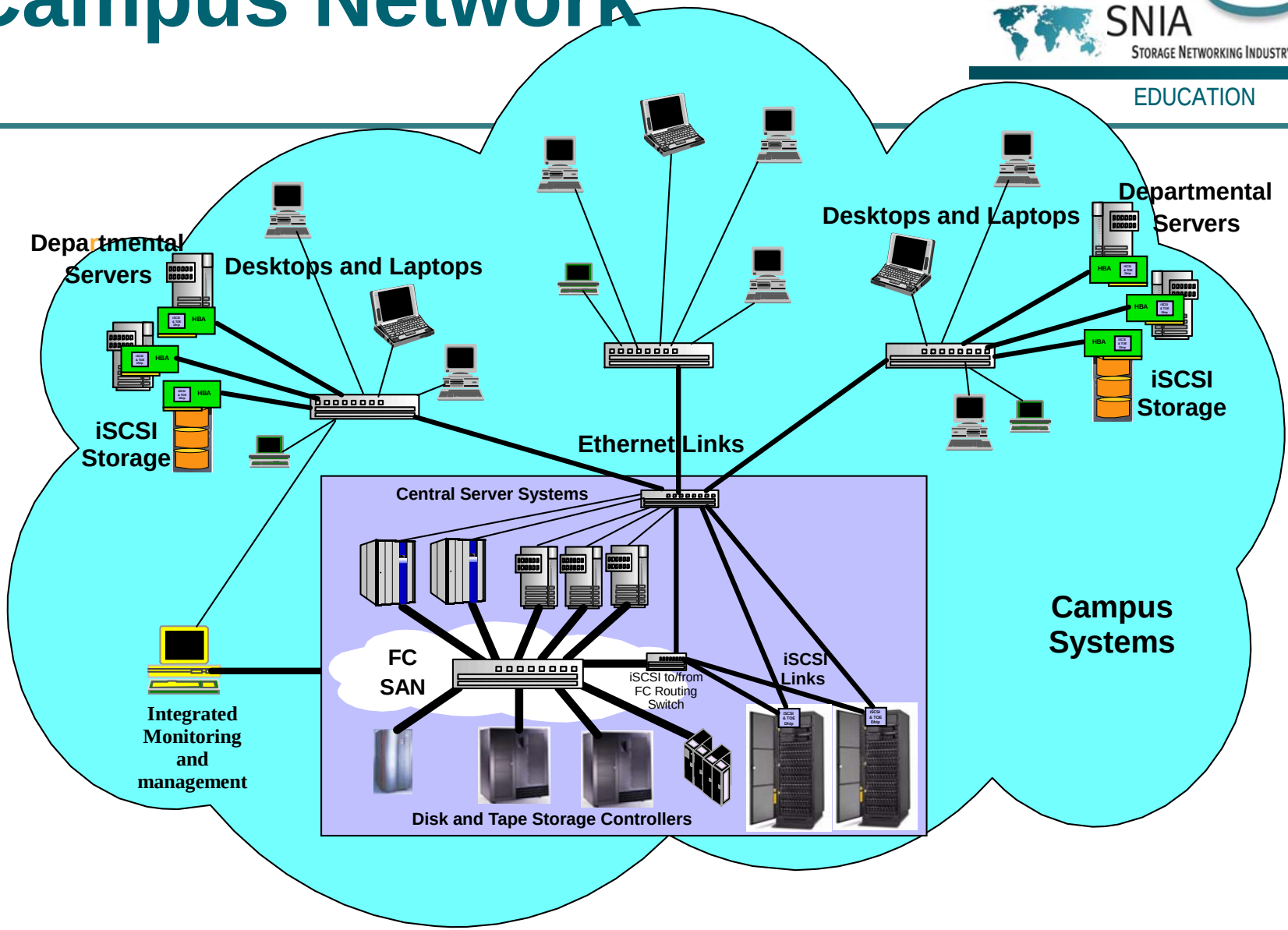
Midrange Environment



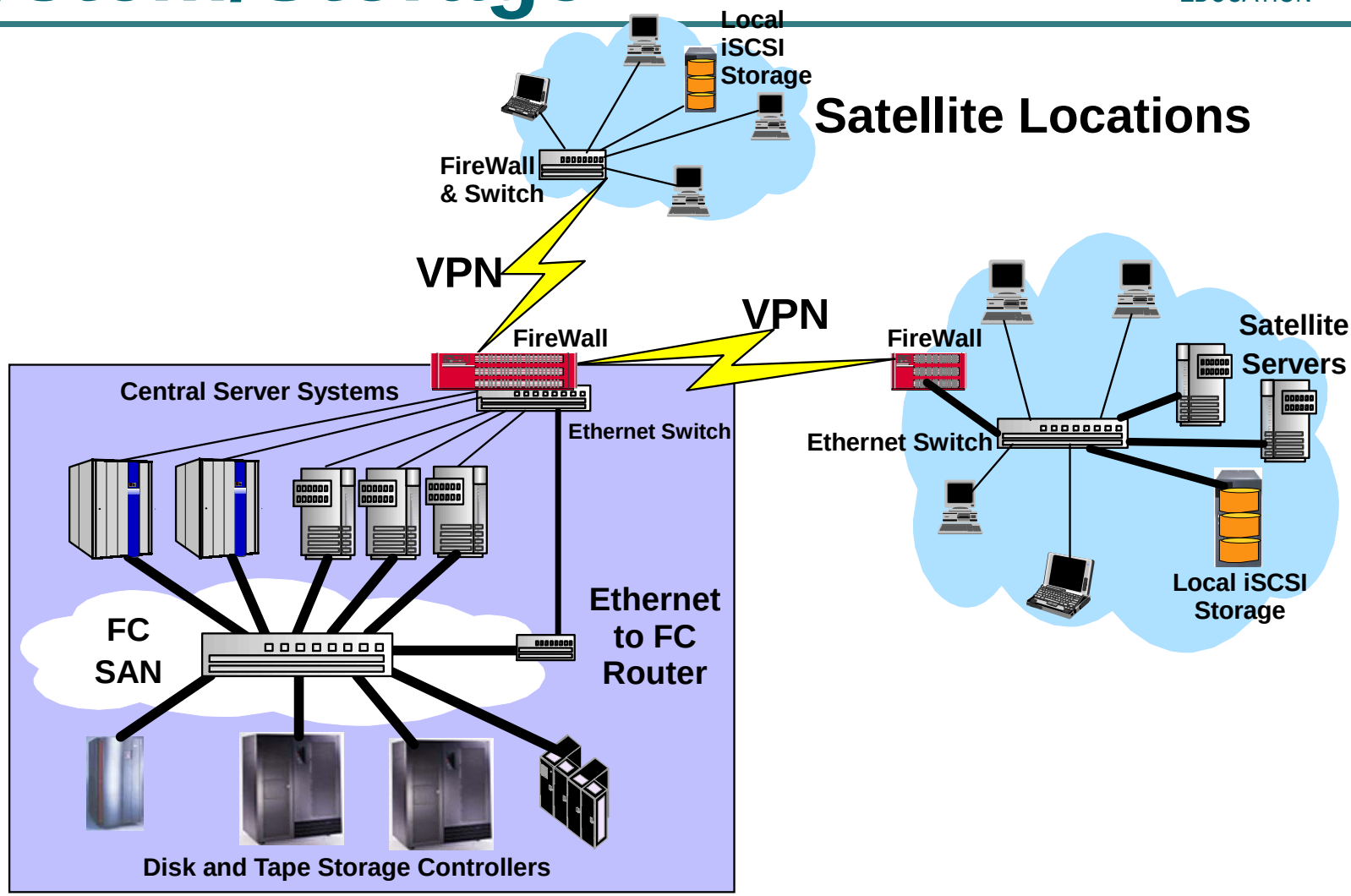
High-End Environment



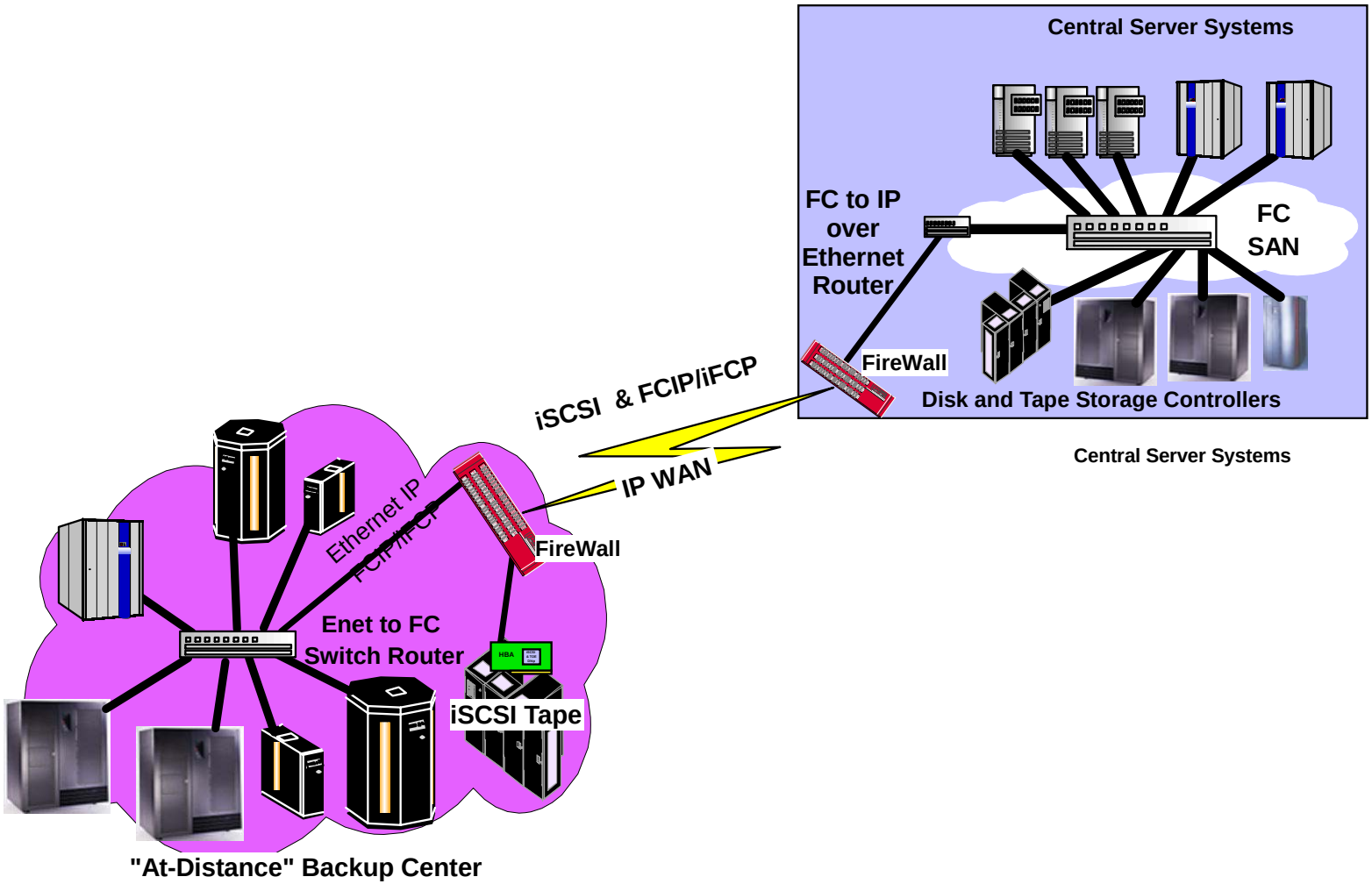
Campus Network



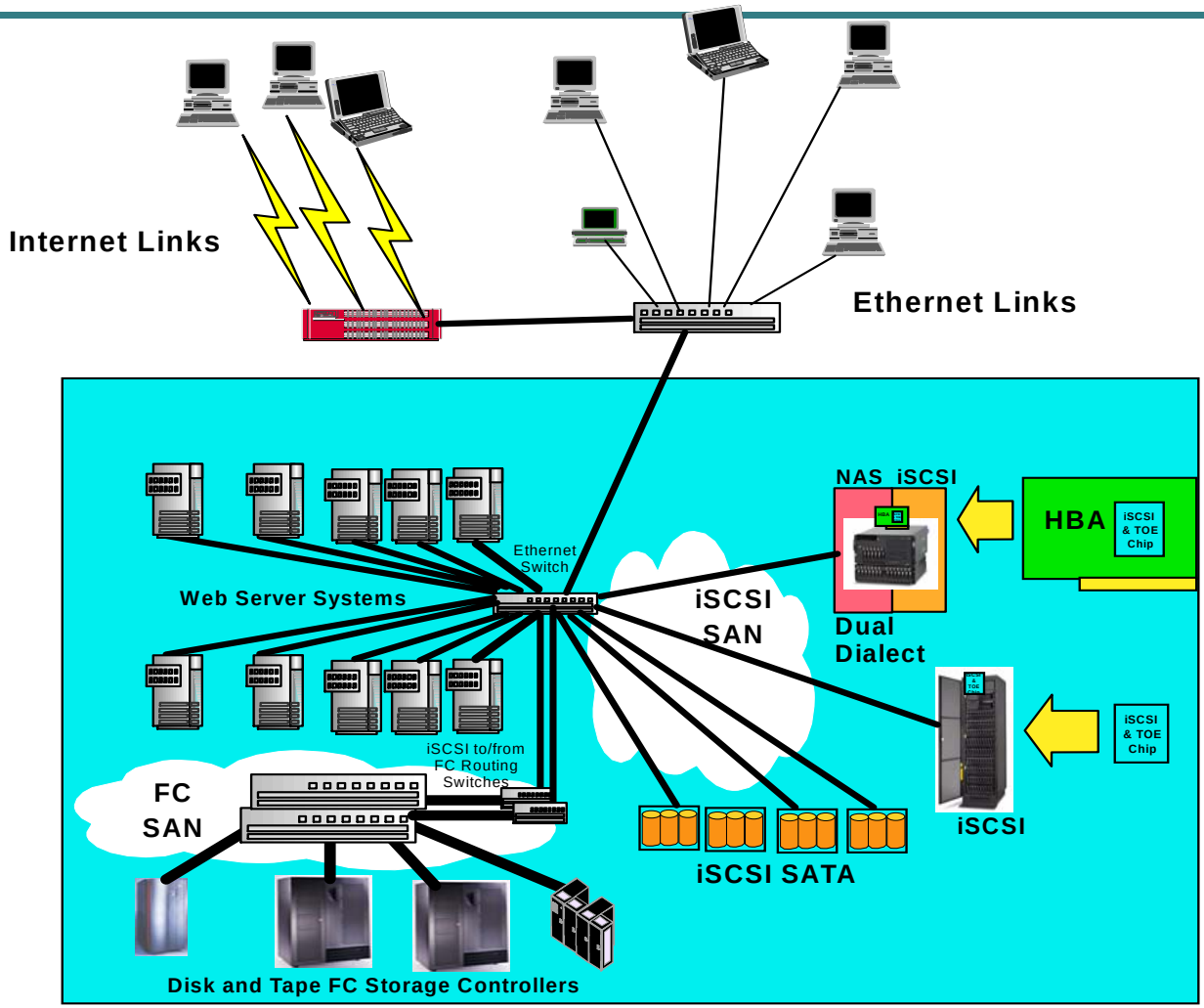
Satellite and Central System/Storage



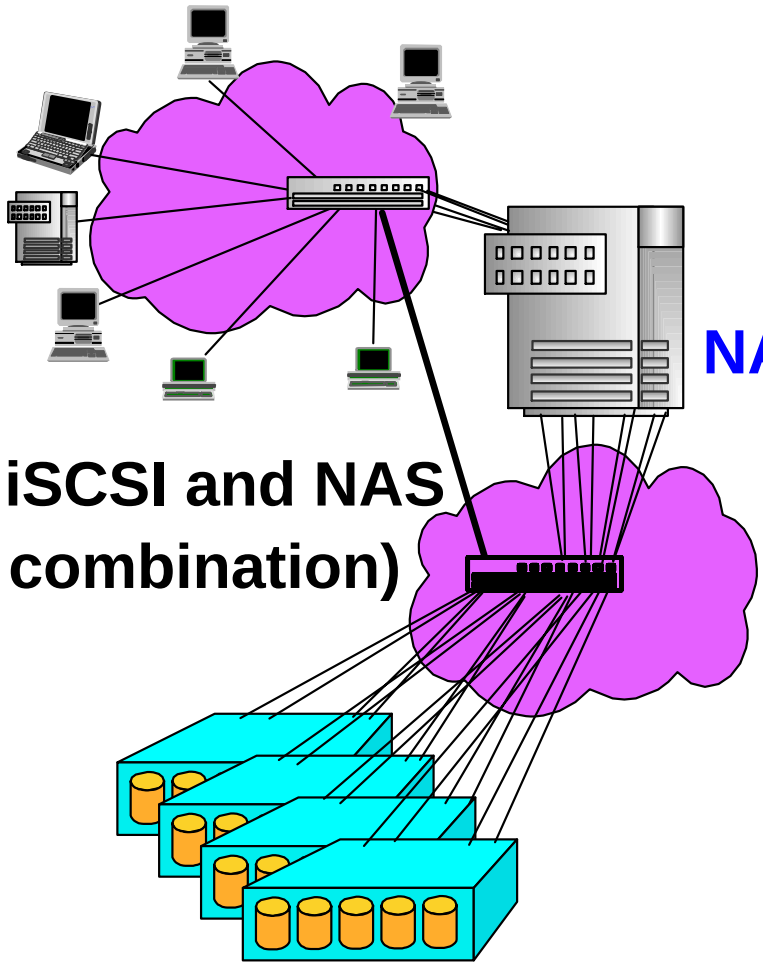
At-Distance



Web Server Installation



Peaceful Co-existence iSAN & NAS



NAS iSCSI-Gateway

Supports both iSCSI and NAS (a Dual Dialect combination)

iSCSI Raid Ctlrs

Agenda

- Introduction
- iSCSI Features
  - Boot, Discovery, Error Handling
- iSCSI usage models
- **IP Security**
- Q & A

Security Properties

- **Connection Authentication:** Who are you? Prove it!
  - Mutual Authentication: Initiator to Target AND vice-versa
- **Packet Integrity:** Has this data been tampered with?
  - Cryptographic Packet by Packet authentication & integrity check, not just checksum or CRC
  - Anti-Replay to prevent regeneration attack
- **Privacy:** Encryption of the Data
- **Authorization:** What are you allowed to do?
  - iSCSI: Who can connect to which Target
  - LUN masking & mapping handled by SCSI, not iSCSI
- **iSCSI Security Features:** Must be implemented but are
  - Optional to use
  - Subject to negotiation

iSCSI Security Considerations

- **Connection Authentication is iSCSI way to determine trustworthiness via**
  - CHAP -- Challenge Handshake Authentication Protocol with strong secrets is required
    - Can't use passwords
    - Stronger than basic CHAP when specification is followed
  - SRP -- Secure Remote Password
  - Kerberos -- A Third Party Authentication protocol
  - SPKM-1,SPKM-2 -- Simple Public Key Mechanism

- **Connection Security may be used with or without IPsec's Packet Security:**
  - Packet Authentication
    - Origin assurance
    - Anti-Reply protection
  - Privacy
    - Encryption

SNIA

STORAGE NETWORKING INDUSTRY ASSOCIATION

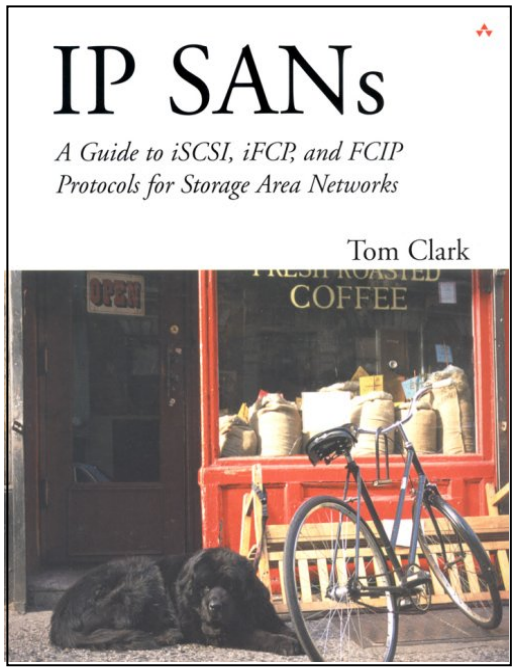
EDUCATION

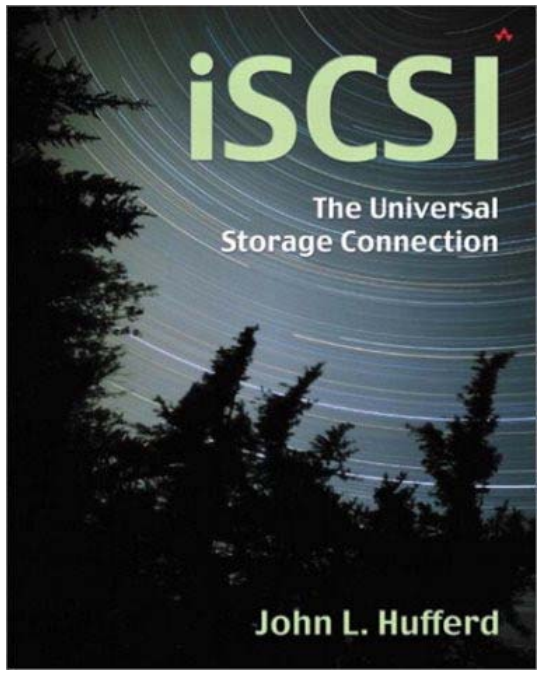
Conclusions

iSCSI is the Network Storage Alternative

- The performance on 1Gb Ethernet networks are “Good Enough” for many applications
- Host systems can use the cost effective software iSCSI Initiators
- Host system can use the low overhead of HW iSCSI HBA for Initiators
- With Ethernet networks moving to 10Gb, most storage networking needs can be handled by iSCSI
- iSCSI is not just a Low-End protocol but will also apply to the High End environments.

iSCSI References





Both Books

Published by Addison-Wesley

Available in Book Stores

and Amazon.com

Volume purchases available

The detail specification can be found at

http://www.ietf.org/rfc/rfc3720.txt?number=3720

Q&A / Feedback

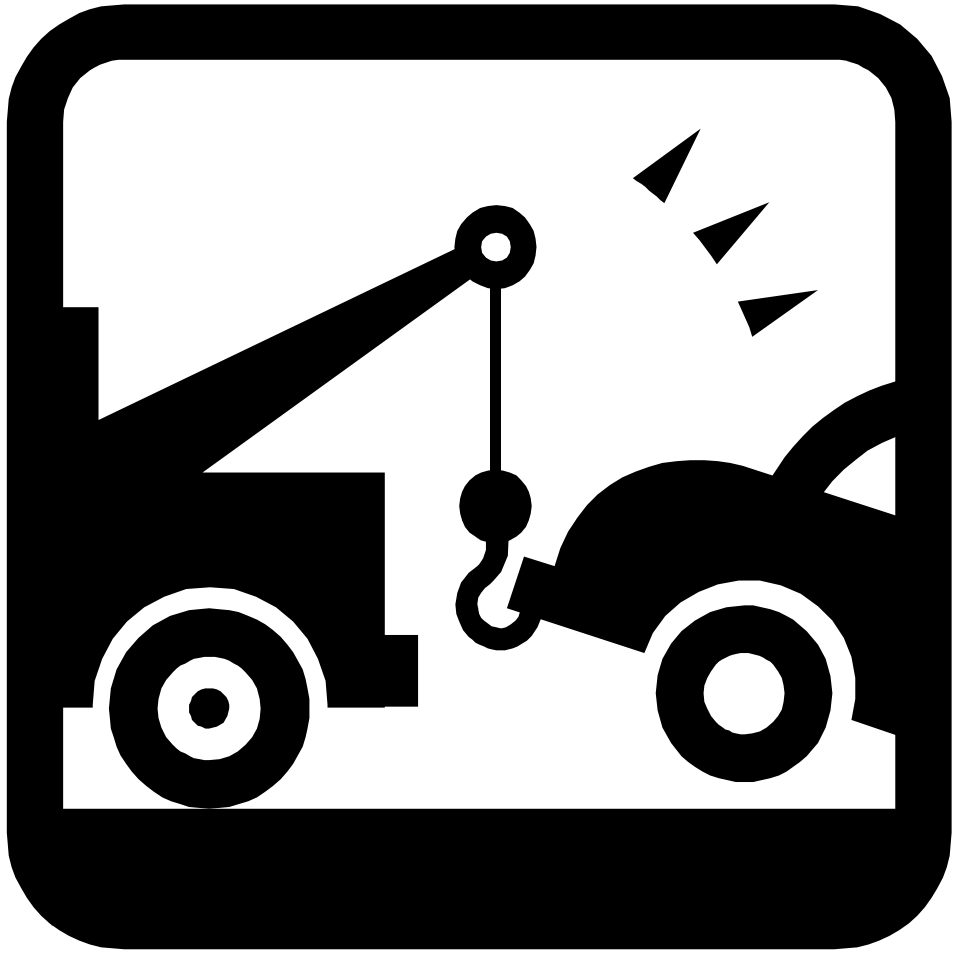
- Please send any questions or comments on this presentation to SNIA: track-networking@snia.org

Many thanks to the following Group and individuals for their contributions to this tutorial.

SNIA Education Committee

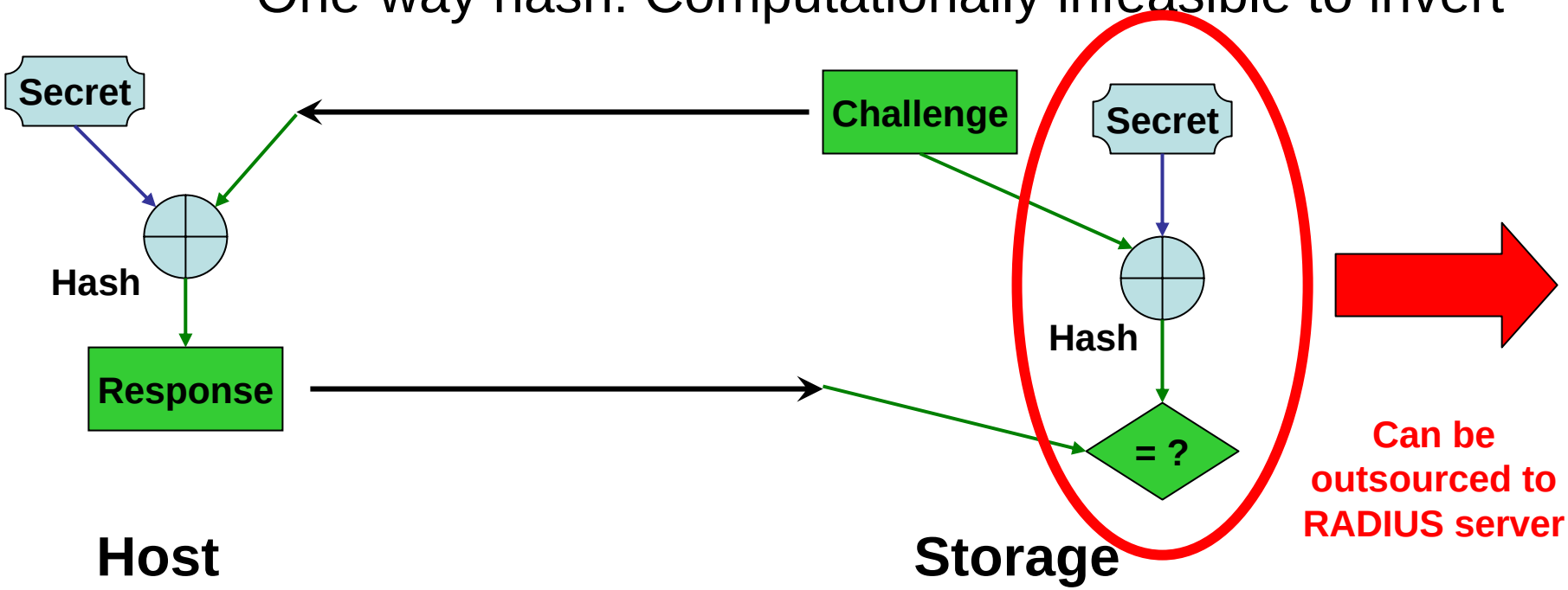
Members of the SNIA IP Storage Forum
David Black
David Dale
John Hufferd
Peter Hunt
Howard Goldstein
Gary Orenstein
Ahmad Zamer

Appendix

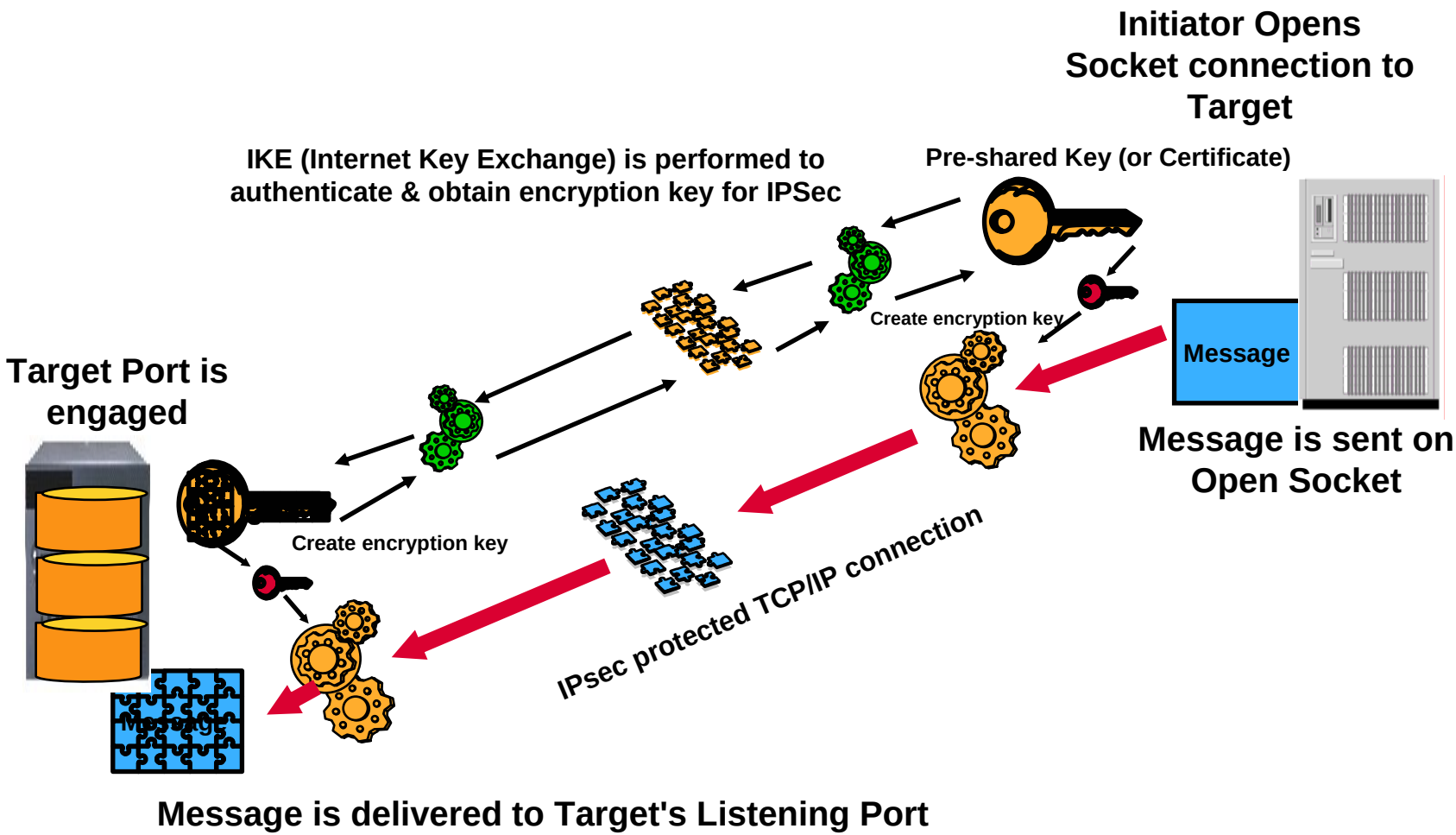


CHAP Authentication Protocol

- Based on shared secret, random challenge
  - Uses a secure (one-way) hash, usually MD5
  - One-way hash: Computationally infeasible to invert



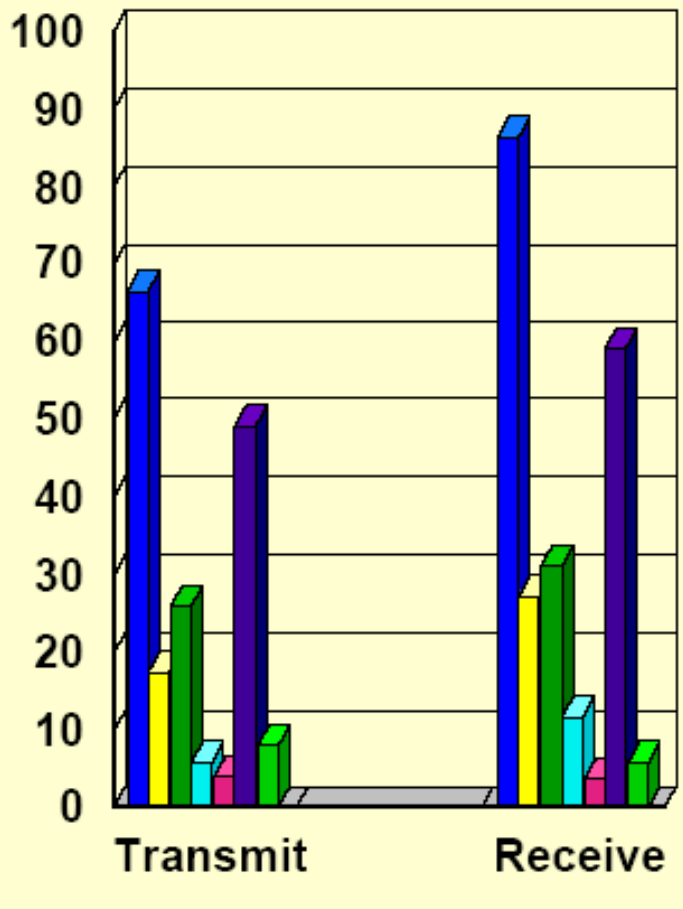
iSCSI with IPsec



NAS vrs iSCSI (on the Storage Controller)

SNIA
Storage Networking Industry Association
EDUCATION

Percent CPU Overhead



With unmodified TCP/IP, iSCSI is 1/3 the overhead of NFS

With all TCP/IP copy overhead offloaded (0 Copy) iSCSI was 1/12 the overhead of NFS

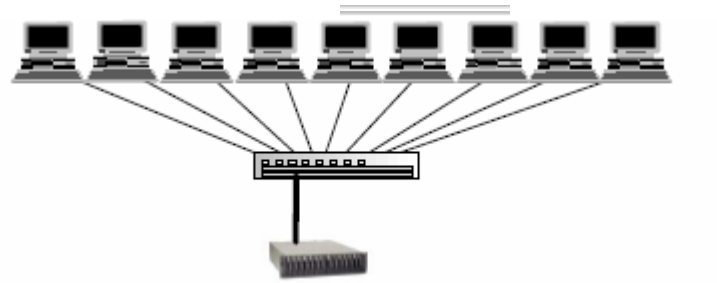
- NFS
- SCSI over GE-TCP/IP
- % of NFS cpu used by SCSI over GE-TCP/IP
- SCSI over GE-TCP/IP 1 copy
- SCSI over GE-TCP/IP 0 copy
- NFS (with 0 TCP/IP data copies) est.
- % of NFS cpu (with 0 TCP/IP data copies) used by SCSI over GE-TCP/IP (0 copies)

Goal: Use NAS for Sharing Files, and iSCSI for everything else

Spreading vrs Centralizing the File System Overhead

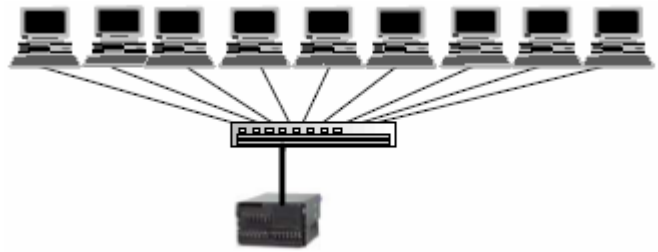
SNIA
Storage Networking Industry Association
EDUCATION

Block I/O (including iSCSI) spread the File System overhead across all the Clients



Block I/O (including iSCSI) Storage Controllers just store the I/O blocks where the Client File System requests (perhaps with Virtualizing LUN Mapping)

NAS Clients move the File System overhead to the NAS server



NAS Servers centralizes the File System functions (and overhead) for all its clients into the NAS Server

Plus the NAS Server still must map the resultant Blocks onto the Storage (perhaps with Virtualizing LUN Mapping)

The non TCP/IP Server side overhead can be 12- 16 times higher in NAS Servers than Block I/O (iSCSI) Storage Controllers

Therefore use NAS for File Sharing and iSCSI for other IP Storage Requirements