



Storage Networking-Evolution to IP?

Ed Chapman
echapman@cisco.com

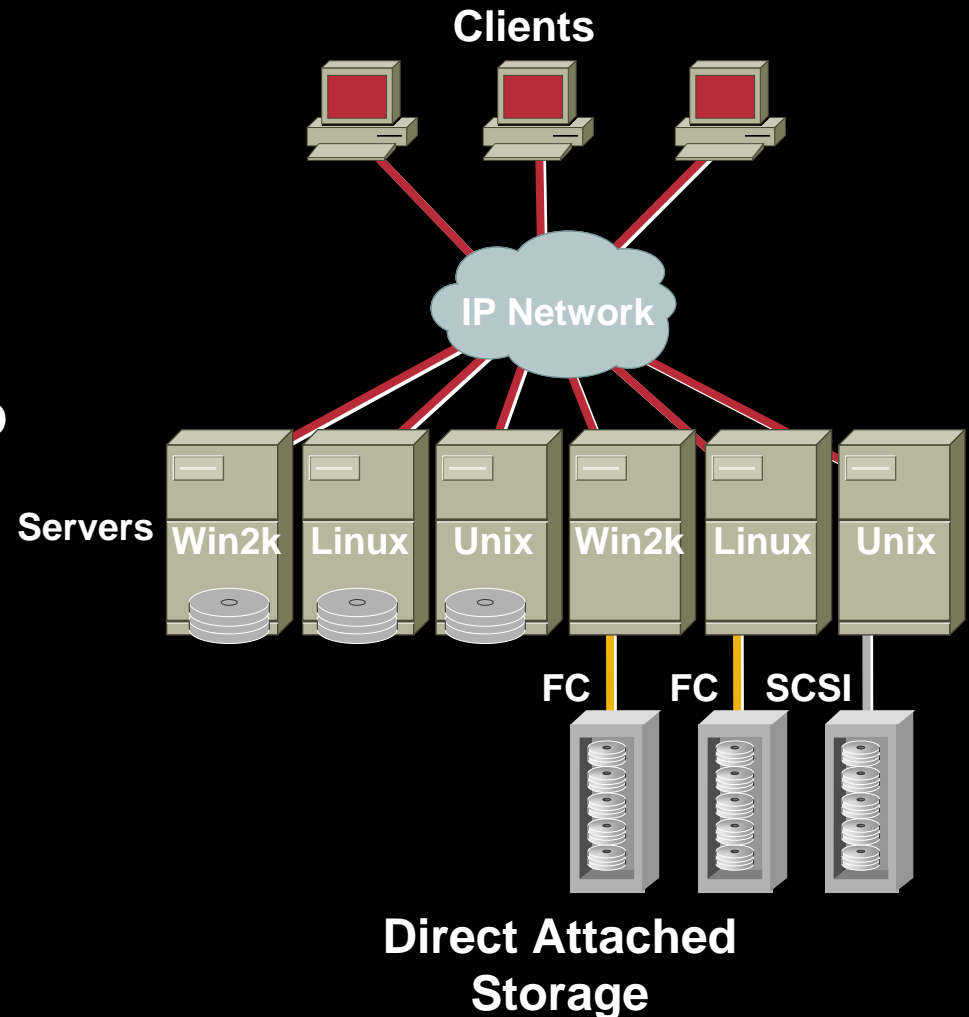
Senior Director
Data Center, Switching and Wireless Business Unit
Cisco Systems, Inc.

Agenda

- **Networking Storage**
- **ISCSI- Why and Where**
- **ISCSI- Design Considerations**
 - Discovery
 - Nic
 - High Availability
 - Security
 - IO
- **IP Storage Networking Looking Forward**
 - Network Boot
 - I-SER, I-Warp
- **Conclusion**

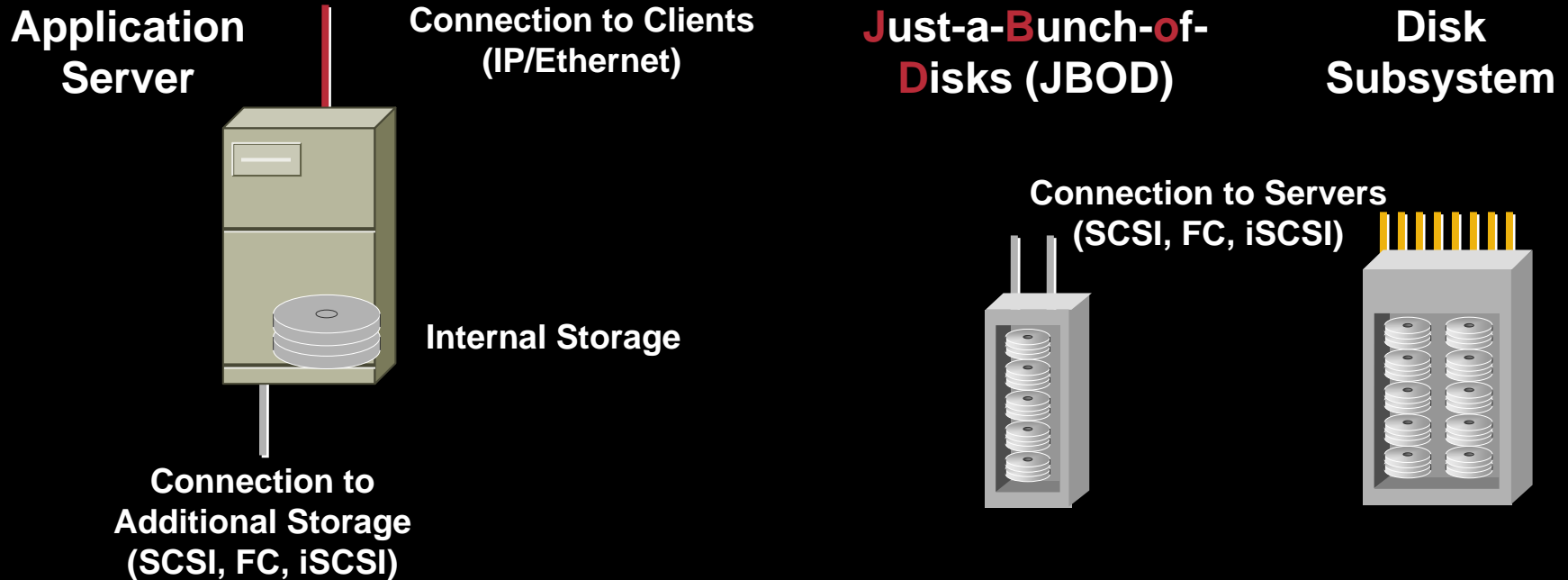
The Typical Storage Environment

- **Direct Attached Storage (DAS)**
- **Storage is captive 'behind' the server**
- **Server CPU must handle user I/O requests, but also:**
 - User database inquiries
 - User file/print serving
 - Data integrity checking
 - Communication with other devices
- **Data access is file system and platform dependant**
- **Costly to scale; complex to manage**



Storage System Components

Cisco.com

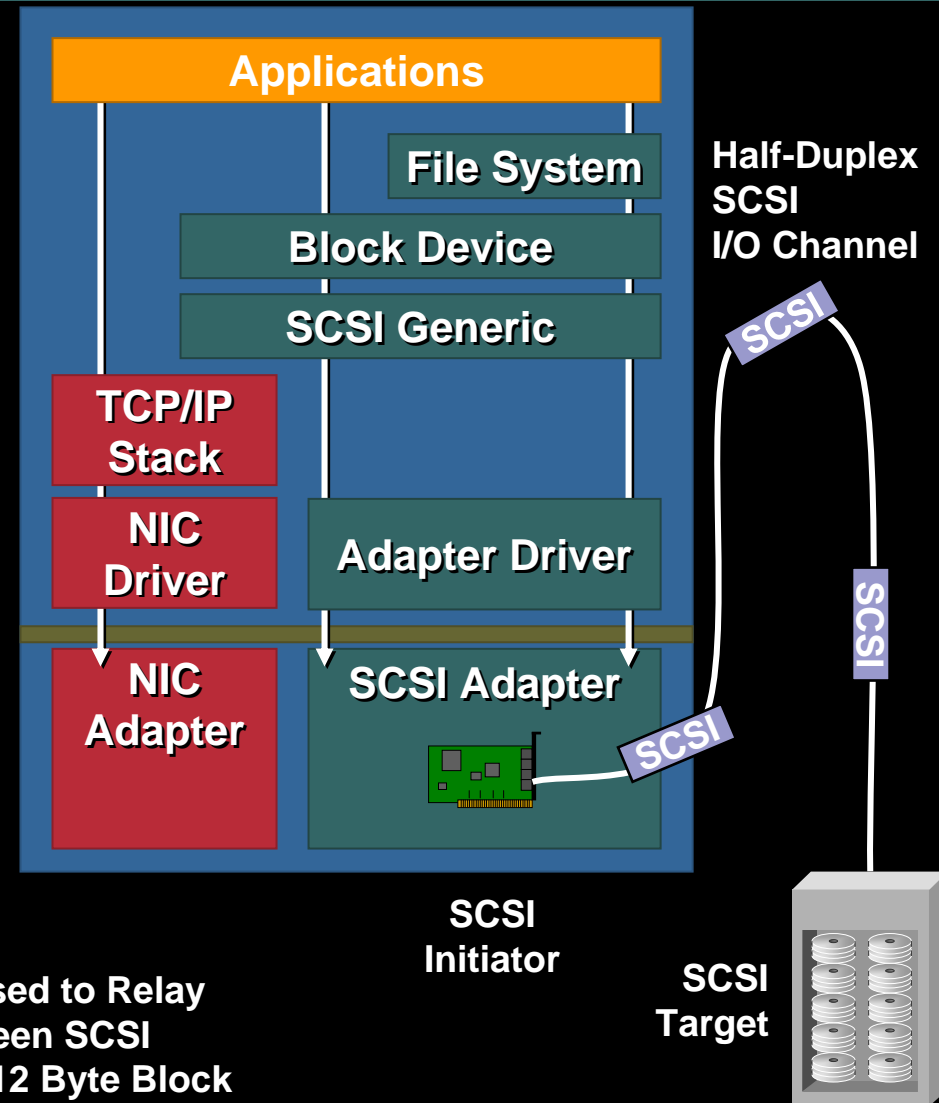


- Commonly have internal storage for host OS
- Direct-attached or network-attached to additional storage
- Front-end IP connection for client communication

- JBOD are simple groups of disks with no data protection (no RAID)
- JBODs are commonly Parallel SCSI or Fibre Channel Loop attached
- Disk Subsystems are complex arrays of disks with many services (RAID)

The SCSI I/O Channel

- SCSI is the protocol used to communicate between Servers and Storage devices
- SCSI I/O Channel provides a half-duplex pipe for SCSI CDBs and Data
- Parallel implementation
 - Bus width: 8, 16 bits
 - Bus speed: 5–80 Mhz
 - Throughput: 5–320 MBps
 - Devices/bus: 2–16 devices
 - Cable length: 1.5m–25m
- A network approach can scale the I/O channel in many areas (length, devices, speed)



SCSI CDB: SCSI Command Descriptor Block Used to Relay SCSI Commands, Parameters, and Status between SCSI Initiators and SCSI Targets; Typically 6, 10, or 12 Byte Block

Networking the I/O Channel

- Same SCSI protocol (SCSI-3) carried over a network transport via **serial** implementation
- Transport must not jeopardize SCSI payload (security, integrity, latency)
- Two primary transports to choose from today: namely IP and Fibre Channel

Note: Infiniband is a potential alternative

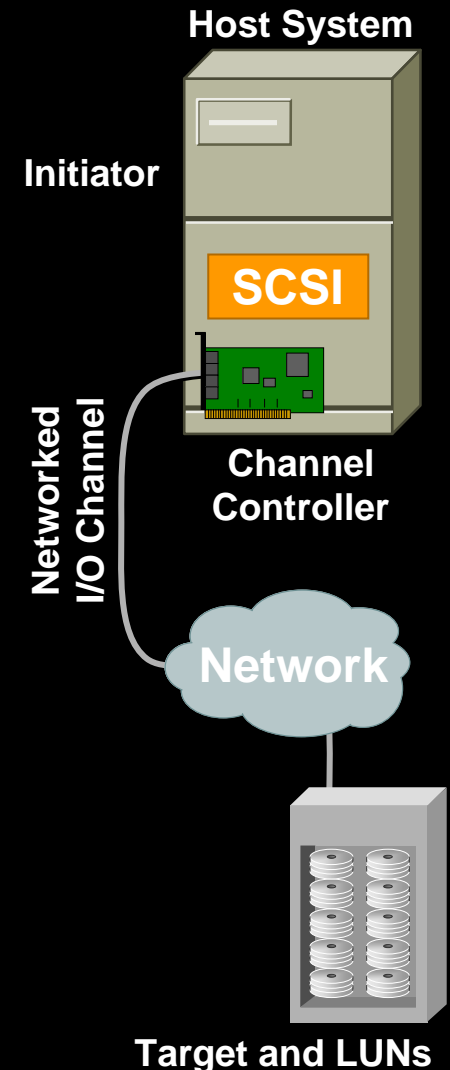
- A networked I/O channel allows for multiple improvements:

Distance limitations greatly increased

Dedicated bandwidth (not shared)

High # of addressable devices

Bandwidth increase (including link bundling)



Considerations for Networked I/O

- There are several necessary considerations for networking the I/O channel:

Security: Ensuring the appropriate initiators access appropriate targets

Performance: Ensuring the networked channel has adequate bandwidth and flow control

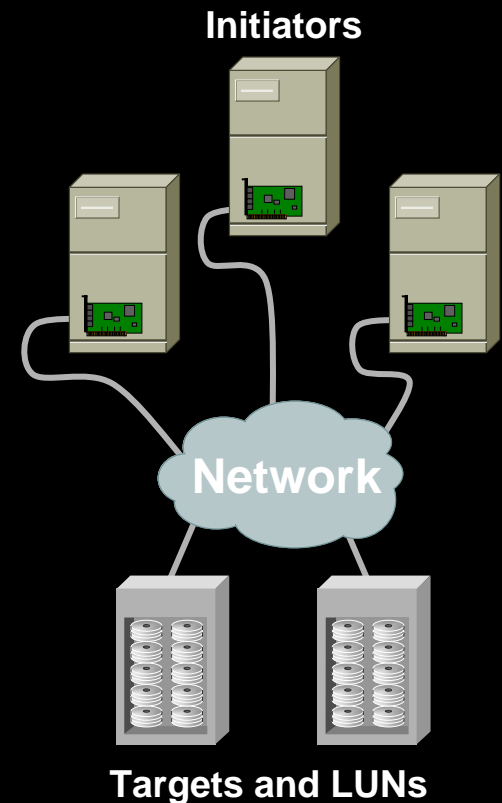
Latency: Ensuring the channel does not experience intolerable latency compromising application integrity

Availability: Ensuring the networked channel can recover from topological faults in a timely manner

Resource Management: Ensuring network resources can be monitored and accounted for

Fault Management: Ensuring the proper tools exist to troubleshoot the chosen network

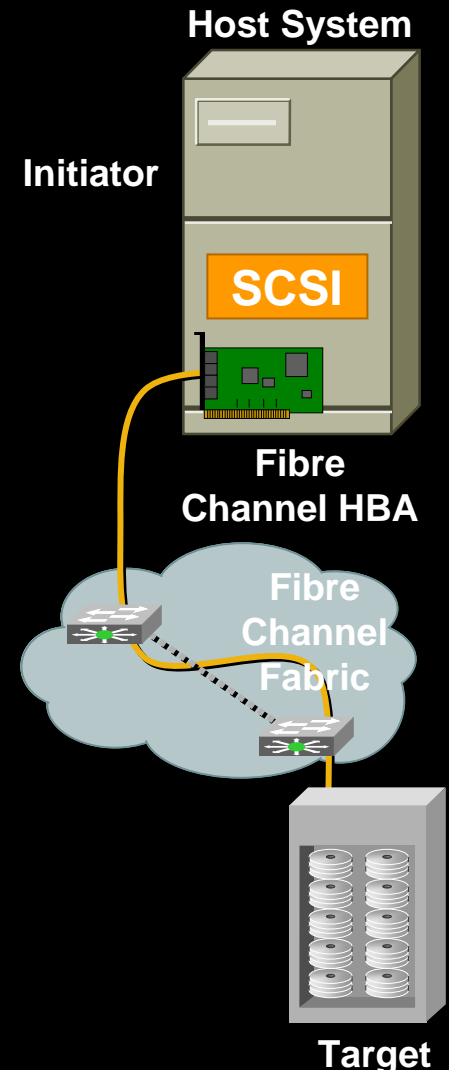
Interoperability: Ensuring the network can be expanded without jeopardizing network stability



Fibre Channel Networking

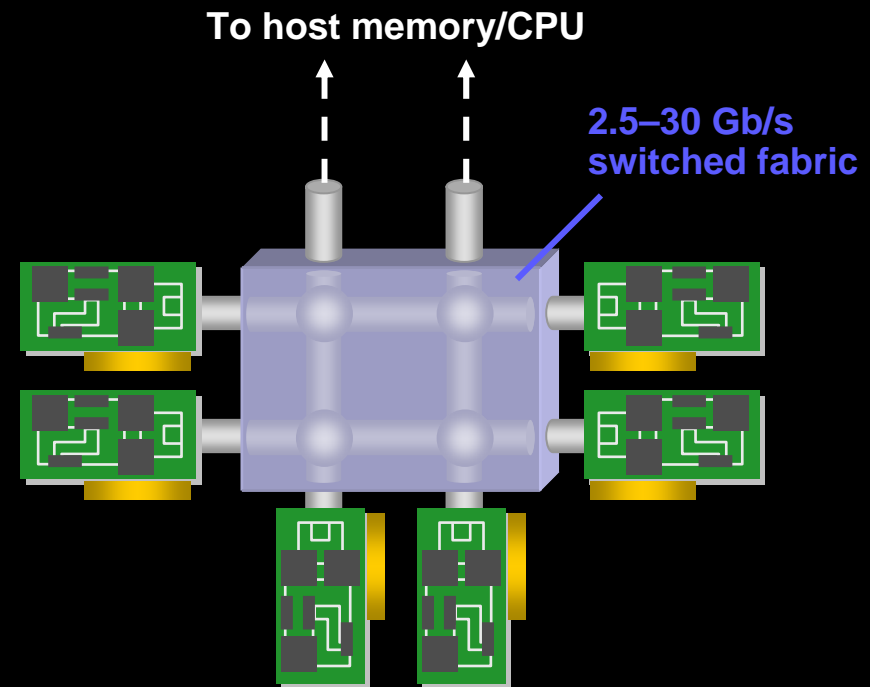
Cisco.com

- **Very common method for networking SCSI**
- **Fibre Channel provides high-speed transport for SCSI payload**
- **Fibre Channel overcomes many shortcomings of DAS including:**
 - Addressing for up to 16 million nodes (24 bits)
 - Loop (shared) and Fabric (switched) transport
 - Speeds of 1/2/4/10 Gbps
 - Support for multiple protocols
- **Combines best attributes of a channel and a network**



InfiniBand

- **InfiniBand features:**
 - 2.5, 10, or 30 Gb/s
 - Switched and routed fabric model increases scalability
 - Low-latency supports IPC
 - Supports “zero-copy” data movement
 - Reduces CPU utilization during I/O operations
- **Aims to create a unified fabric of servers, storage, and peripherals**

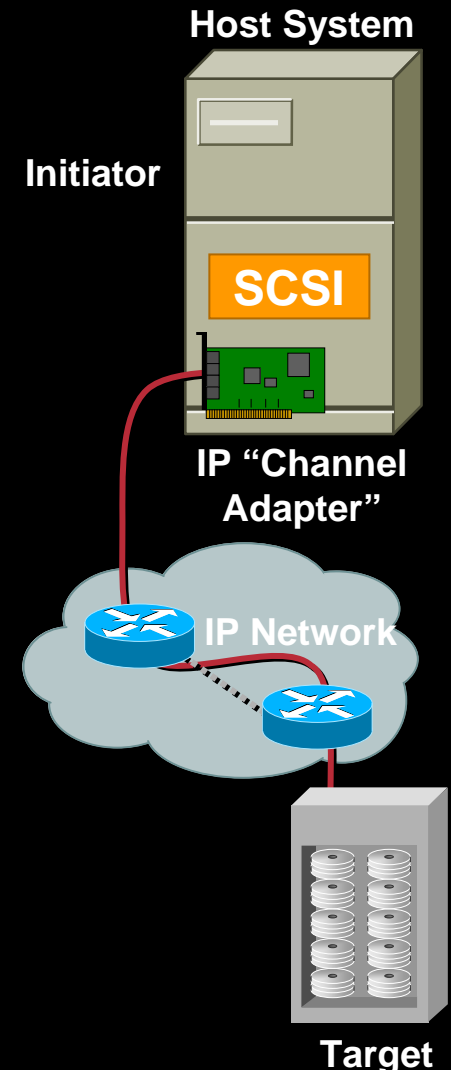


InfiniBand Fabric

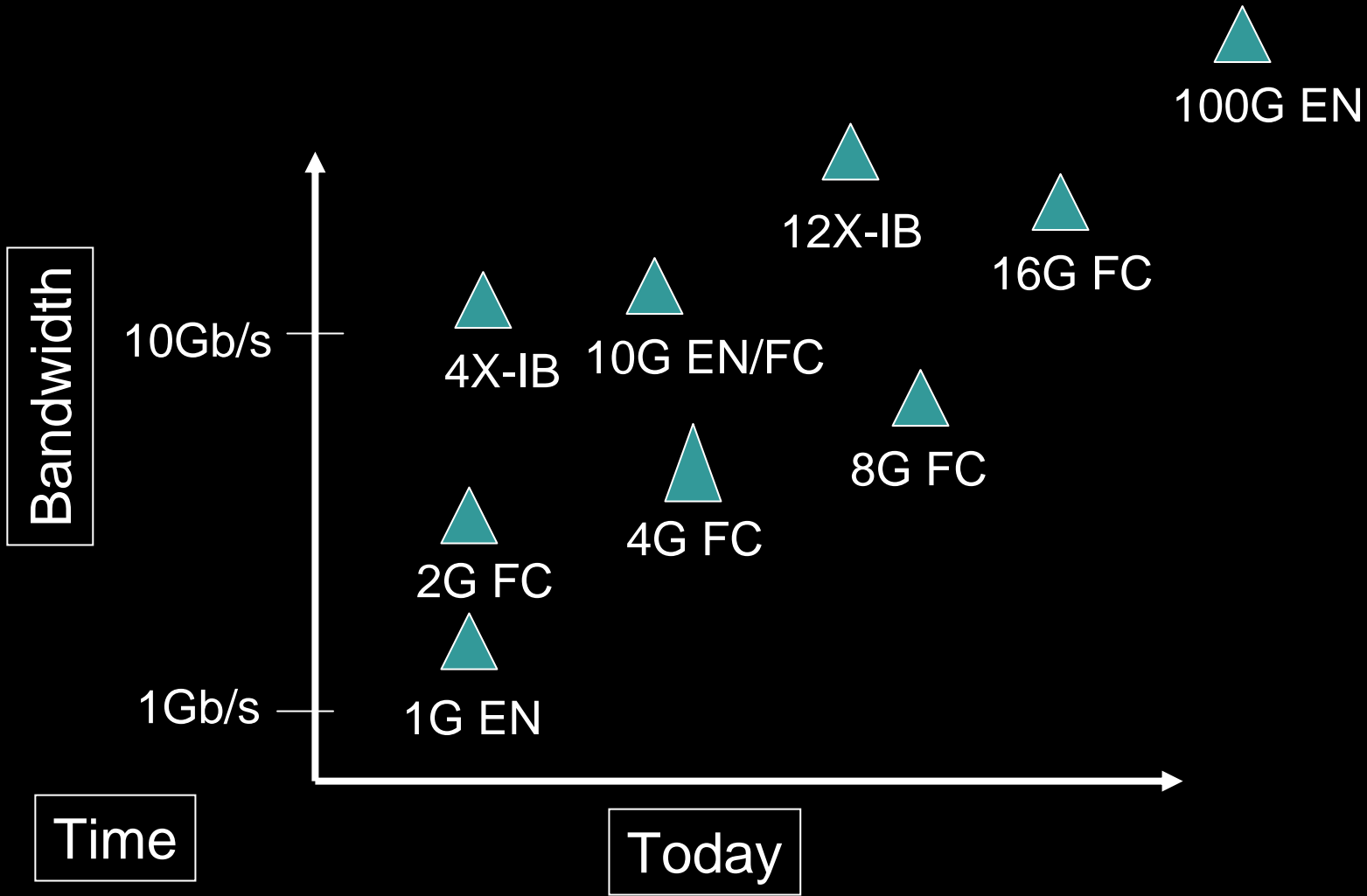
IP—An Alternate I/O Transport

Cisco.com

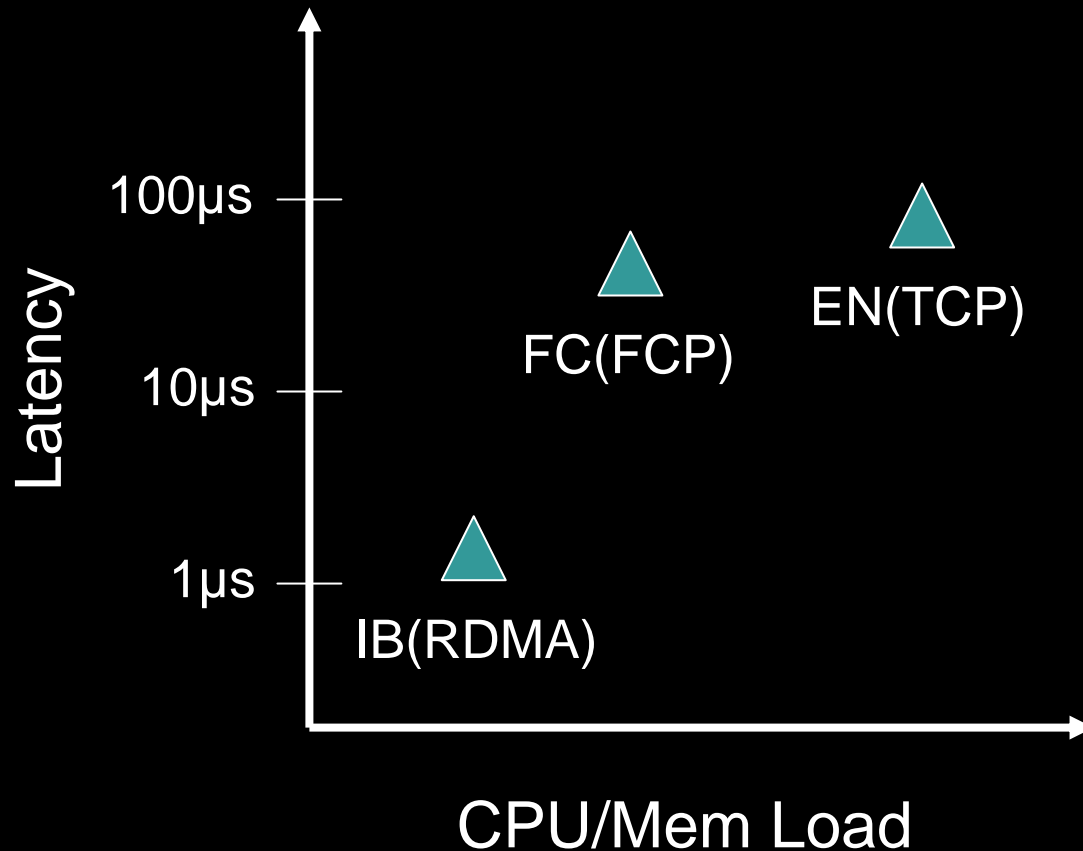
- **Viable transport for I/O traffic**
- **Not necessarily for long-haul I/O only**
- **Similar characteristics to Fibre Channel:**
 - Addressing for close to 4 billion nodes (IPv4)
 - Primarily a switched transport (with routing)
 - Ethernet speeds of 1/10 Gbps or various WAN speeds
 - Support for multiple high level protocols
- **Cost and manageability advantages with IP**
- **IP knowledge base widespread in industry**



IO roadmap



Latencies



- Latencies to application memory in microsecond range for RDMA transfers

- Compare processor memory loads for TCP, FCP and RDMA based protocols

Ethernet and IP – Dynamic Duo?

Cisco.com

Years

1994

(Token Bus)



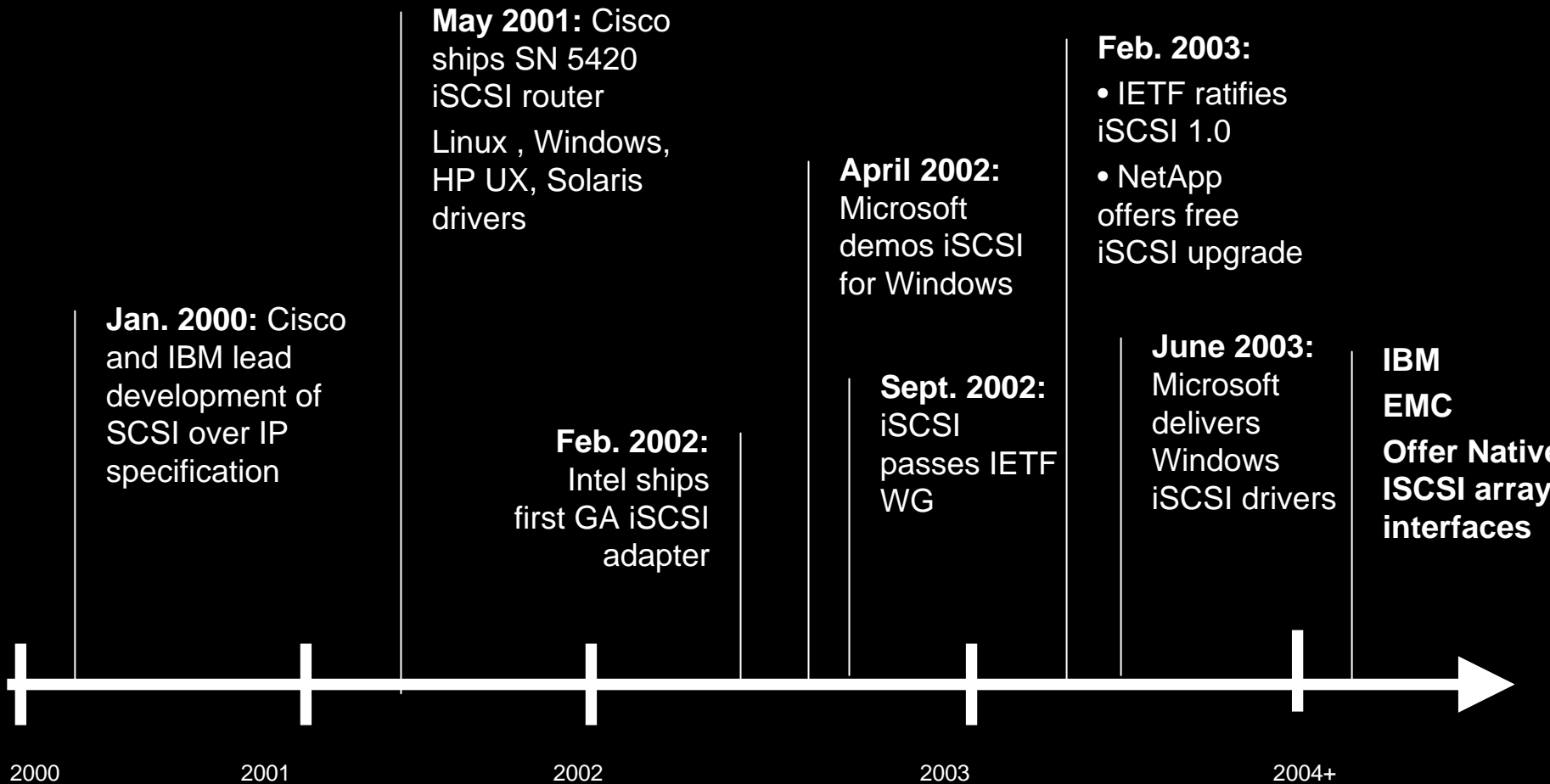
2005

n
posts
her

Agenda

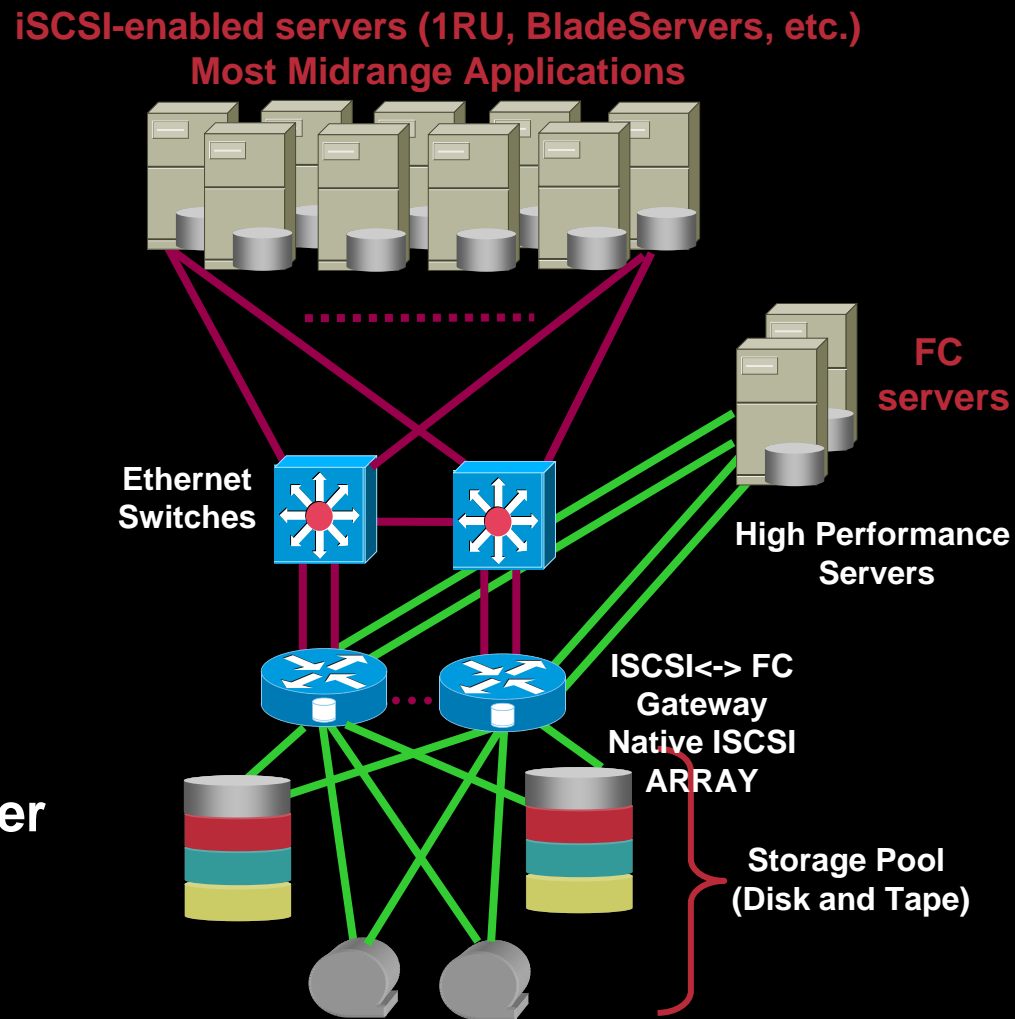
- Networking Storage
- **ISCSI- Why and Where**
- **ISCSI- Design Considerations**
 - Discovery
 - Nic
 - High Availability
 - Security
 - IO
- **IP Networking Looking Forward**
 - Network Boot
 - I-SER, I-Warp
- **Conclusion**

A Brief History of iSCSI

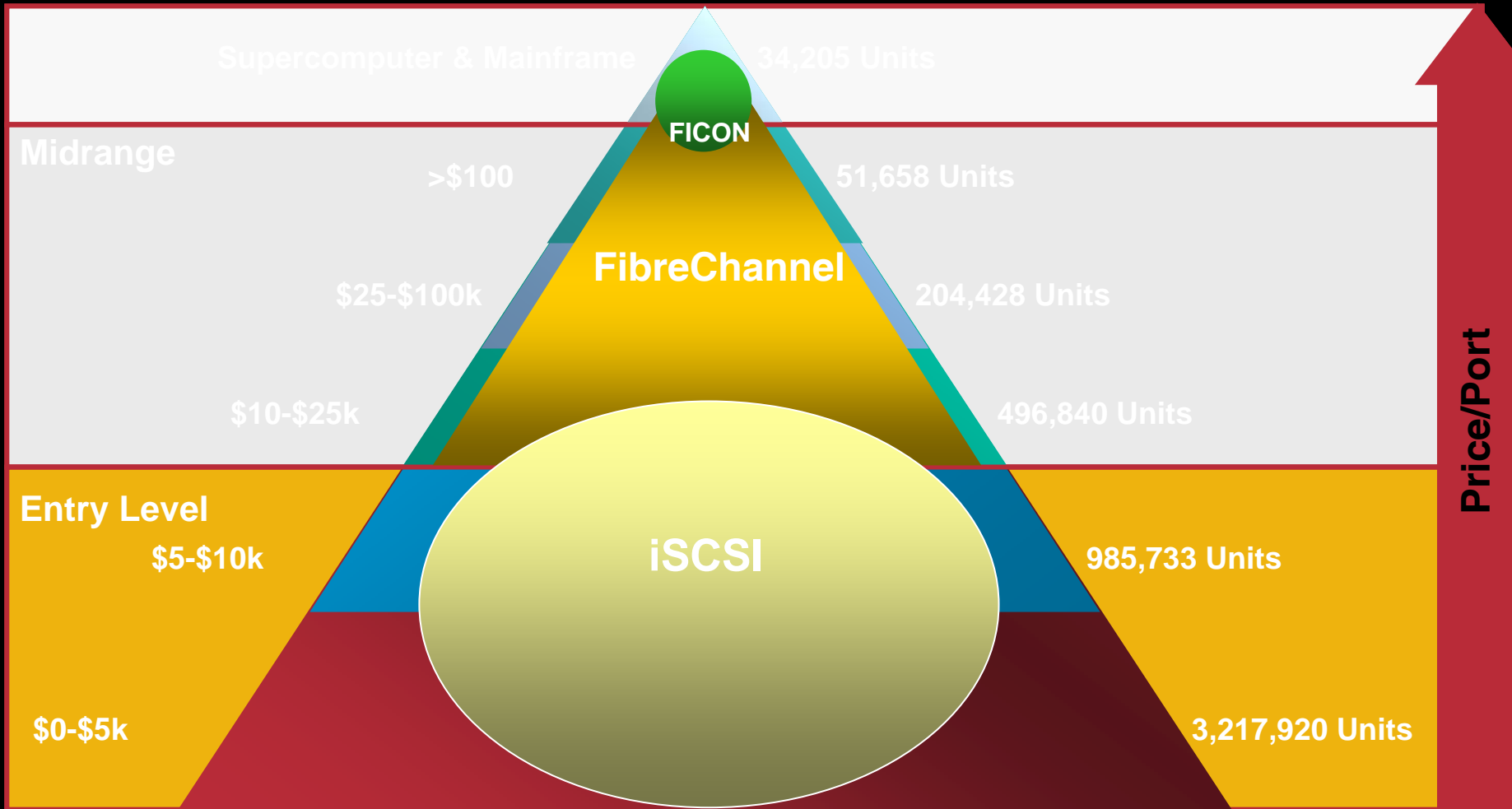


The ISCSI SAN Overview

- **ISCSI SAN Environments**
 - Enterprise Department
 - “Stranded Servers”
 - Small-Medium Business
- **Midrange Applications**
 - Email – Exchange, Notes
 - Database – SQL, Oracle
 - Financials – Great Plains, Lawson, Oracle
 - Web Servers – IIS
 - File Servers
 - Print Servers
 - Customer Developed
- **Most ISCSI servers cost under \$10,000**



Server Market Pricing



Agenda

- Networking Storage
- ISCSI- Why and Where
- **ISCSI- Design Considerations**
 - Discovery
 - Nic
 - High Availability
 - Security
 - IO
- **IP Storage Networking Looking Forward**
 - Network Boot
 - I-SER, I-Warp
- **Conclusion**

- **Small networks**

- **Static configuration, initiators and targets**
 - **'SendTargets' command makes configuration easier**

- **Medium-sized networks**

- **Service Location Protocol (SLP multicast discovery)**

- **Large-sized networks**

- **iSNS (Internet storage name service)**
 - **Includes soft zone domains**
 - **Includes database for ongoing management**

iSCSI Login Sequence (No Authentication)

Cisco.com

Initiator (PC1) with
iSCSI Driver

Single TCP Session

Target
(SCSI Router)

TCP port 3260
(listen)

Discovery:
Contact Target
and Negotiate
Security and
Session
Parameters

Establish TCP Session (SYN, SYN/ACK, ACK sequence)

0x03 iSCSI Login Command

SessionType = discovery; InitiatorName=iqn.1987-05.com.acme.....PC1

0x23 iSCSI Login Response (success)

Auth=none; HeaderDigest=none; DataDigest=none; ...

0x03 iSCSI Login Command

*SessionType = discovery; InitiatorName=iqn.1987-05.com.acme.....PC1
DataPDULength= ... ; MaxBurstSize= ... ; ...*

0x23 iSCSI Login Response (success)

DataPDULength= ... ; MaxBurstSize= ... ; ...

0x04 iSCSI Text Command

SendTargets=All

0x24 iSCSI Text Response

TargetName=iqn.1987-05.com.acme.....betty

0x03 iSCSI Login Command

*SessionType=normal; InitiatorName=iqn.1987-05.com.acme.....PC1;
TargetName=iqn.1987-05.com.acme.....betty*

0x23 iSCSI Login Response (success)

Auth=none; HeaderDigest=none; DataDigest=none; ...

0x03 iSCSI Login Command

*SessionType=normal; InitiatorName=iqn.1987-05.com.acme.....PC1;
TargetName=iqn.1987-05.com.acme.....betty*

0x23 iSCSI Login Response (success)

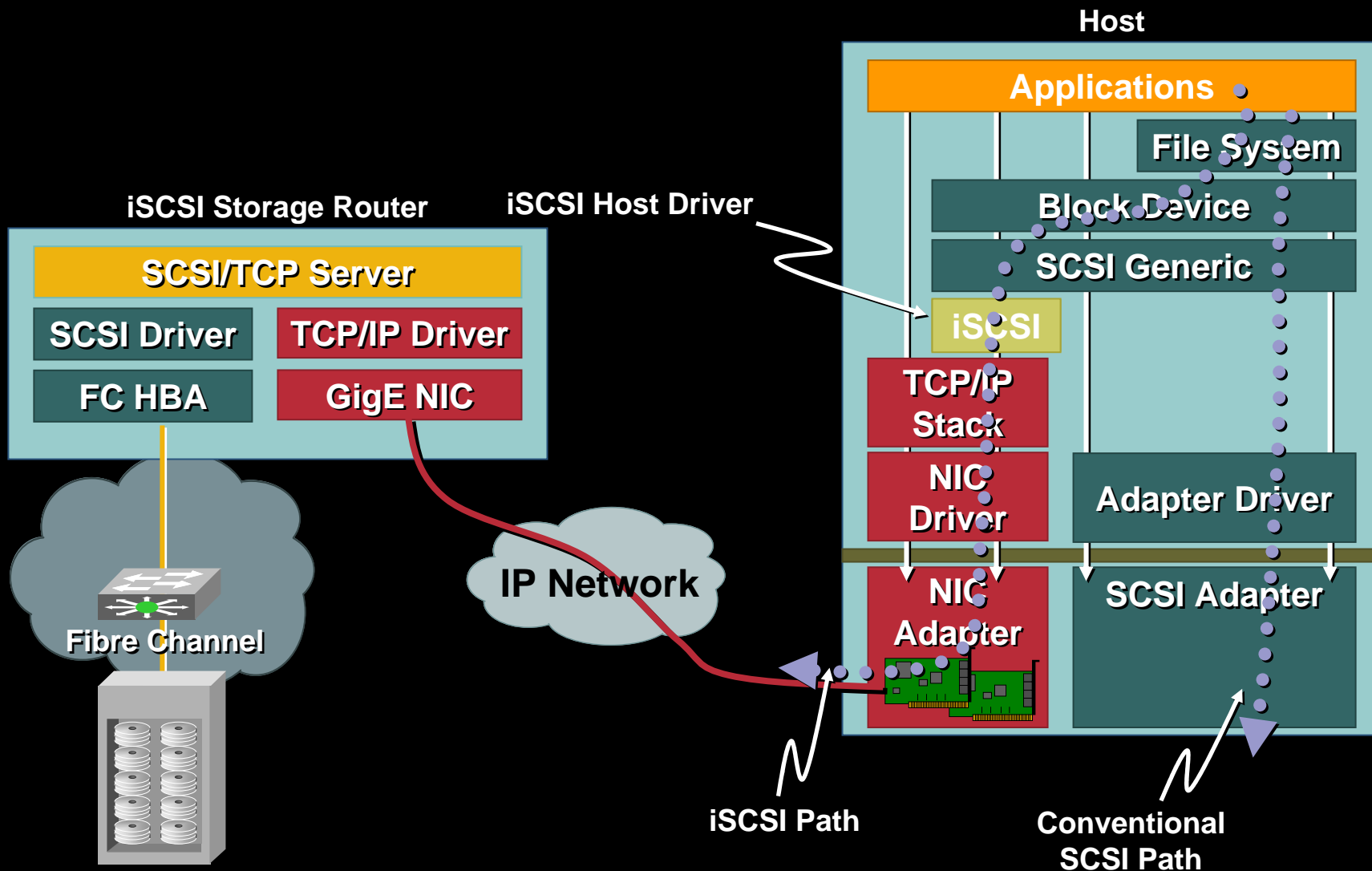
DataPDULength= ... ; MaxBurstSize= ... ; TargetAlias=betty; ...

Block Device
Has Already
Initialized onto
the Fibre
Channel Fabric

Discovery:
Solicit Available
Targets

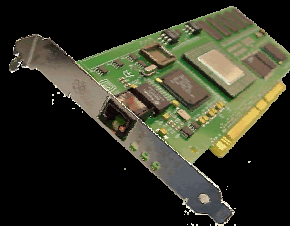
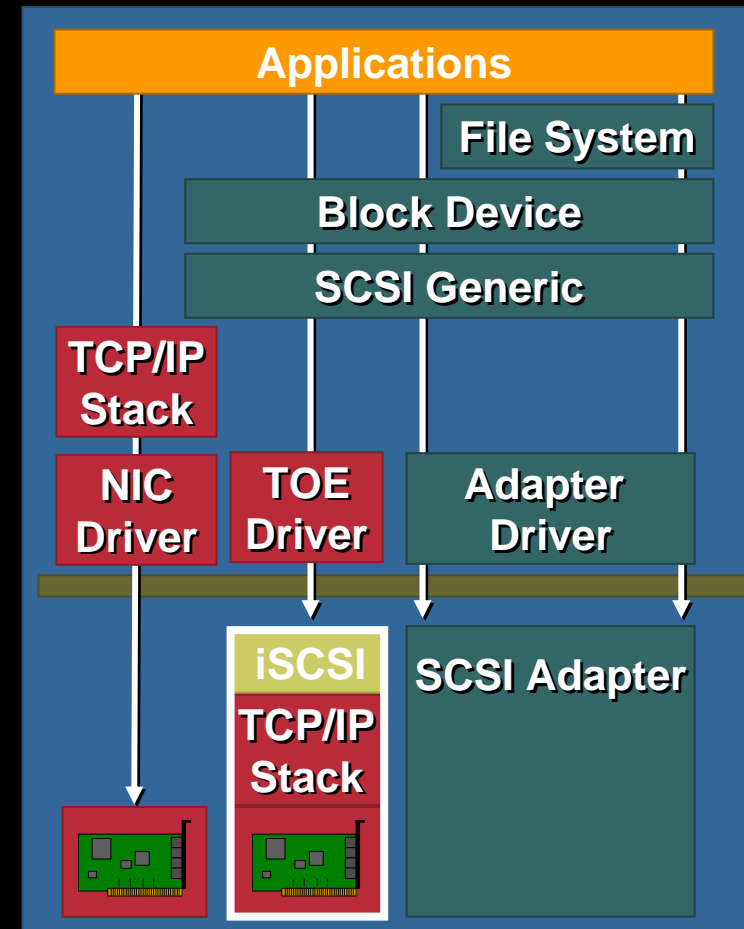
Normal Login—
Login to each
Target and
Negotiate
Security and
Session
Parameters

iSCSI Architecture—Software Driver



iSCSI + TCP Offload Engine (TOE)

- Hardware implementation of iSCSI within NIC
- Offloads TCP and iSCSI processing into hardware
- Relieves host CPU from iSCSI and TCP processing
- Two forms of offload:
 - Partial offload: Data-path only
 - Full offload: Control and Data-path
- Wire-rate iSCSI performance

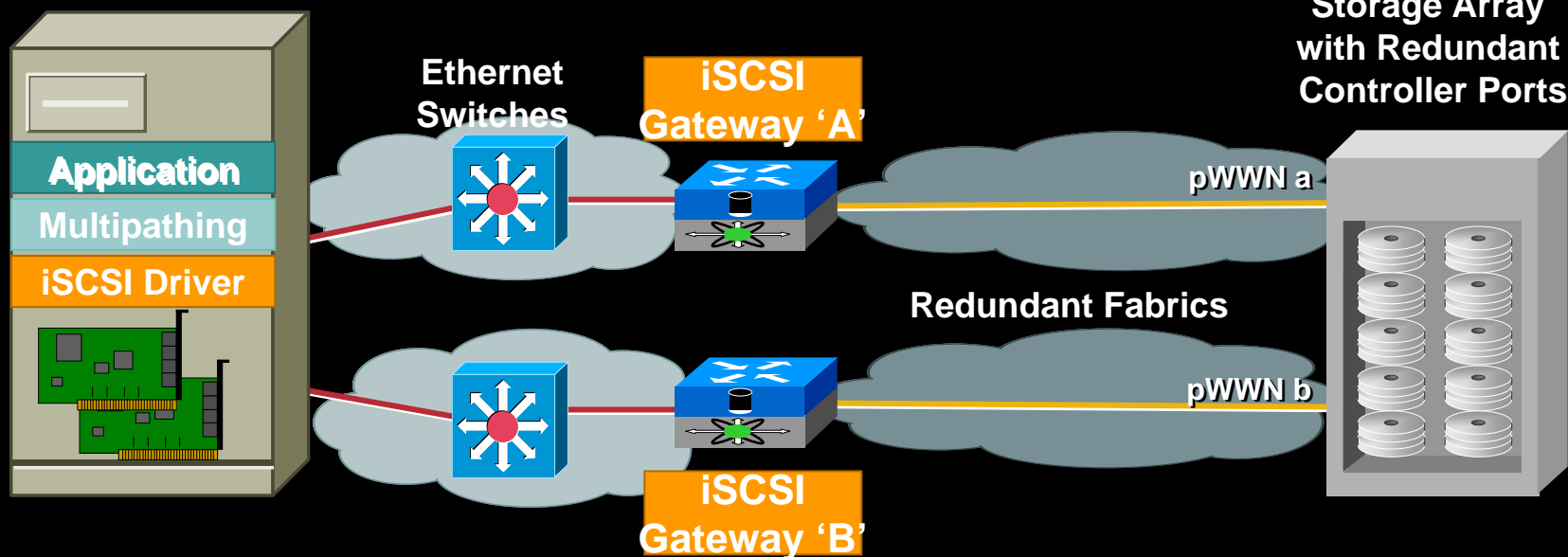


iSCSI HA Design—Multipathing

Cisco.com

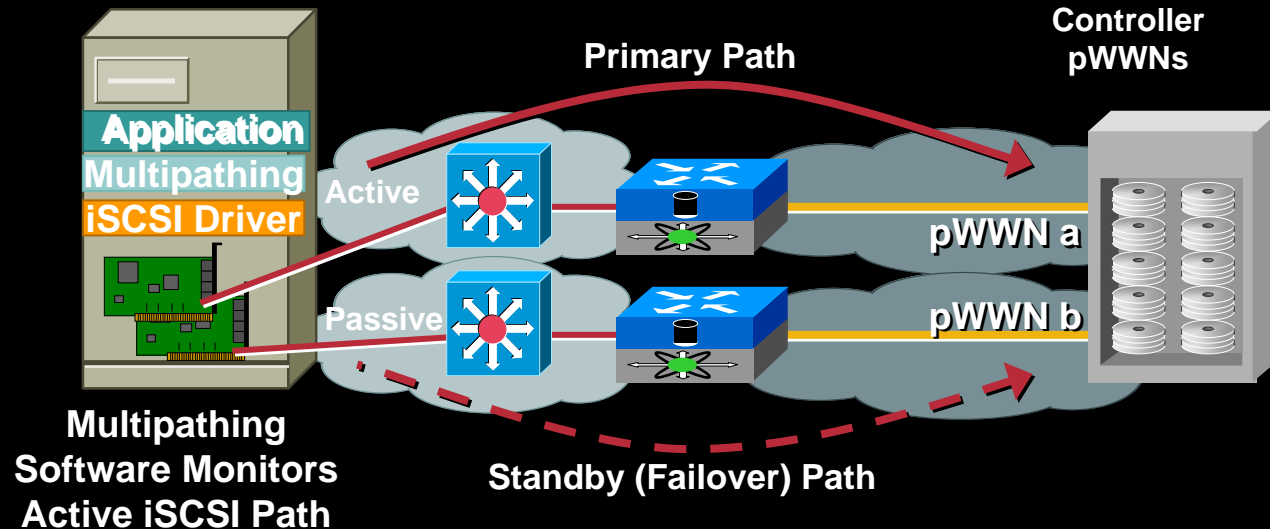
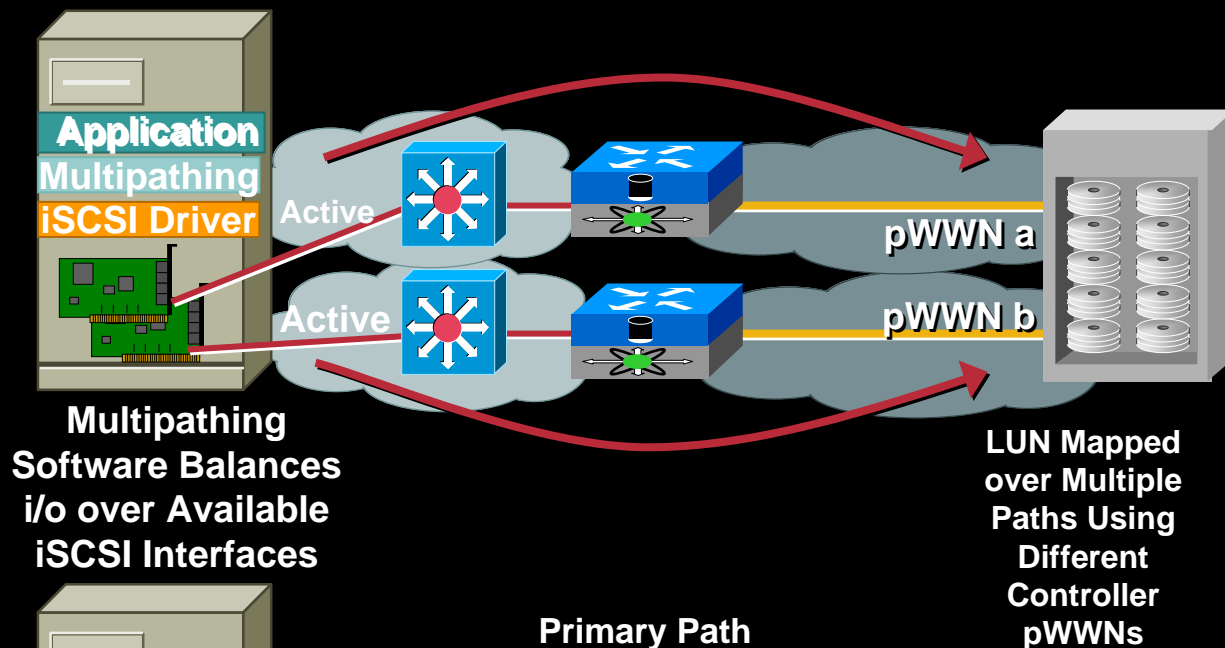
- **Multipathing**: multiple hardware paths to a single drive (LUN)
- A totally redundant single-host system
- A similar configuration for clustered servers
- Multipathing includes EMC-Powerpath, Veritas-DMP, and Compaq-SecurePath, etc...

Host with
Multiple(iSCSI) NICs
and Multipathing
Software Installed



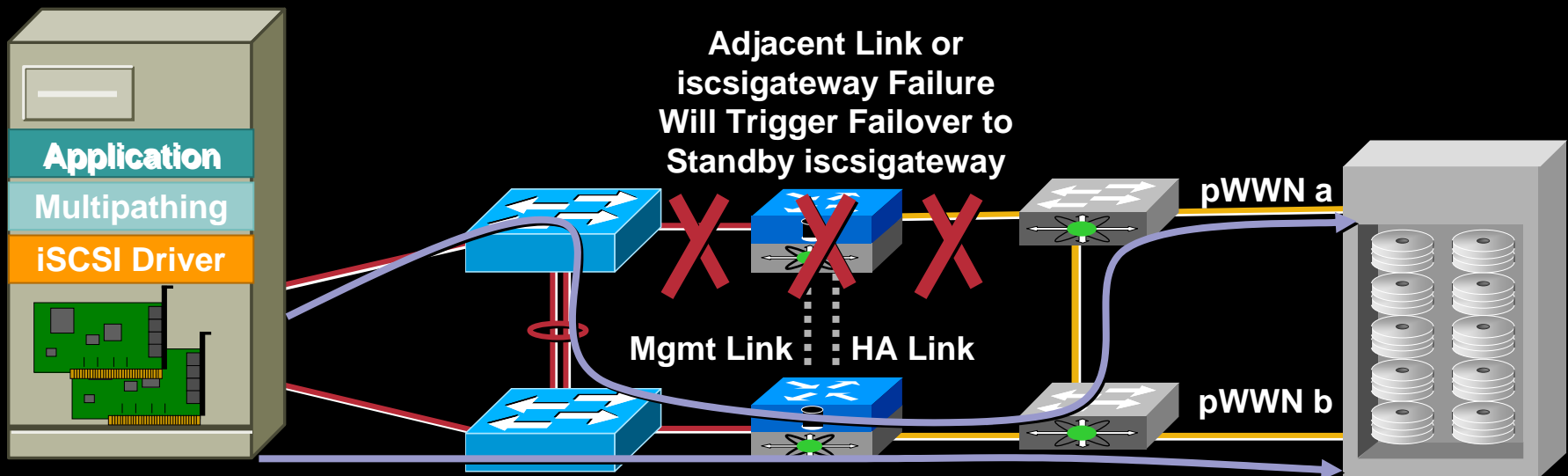
iSCSI HA Multipathing Variations

- **Active/Active:** balanced i/o over both paths (implementation specific)
- **Active/Passive:** i/o over primary path—switches to standby path upon failure



HA Combinations: Multipathing + iSCSI Clustering

- Multipathing and iSCSI Clustering can be used together or in isolation
 - iscsigateway defined for each path
 - iscsigateway will failover when it **recognizes** a failure
 - Multipathing will failover for other failures through timeouts



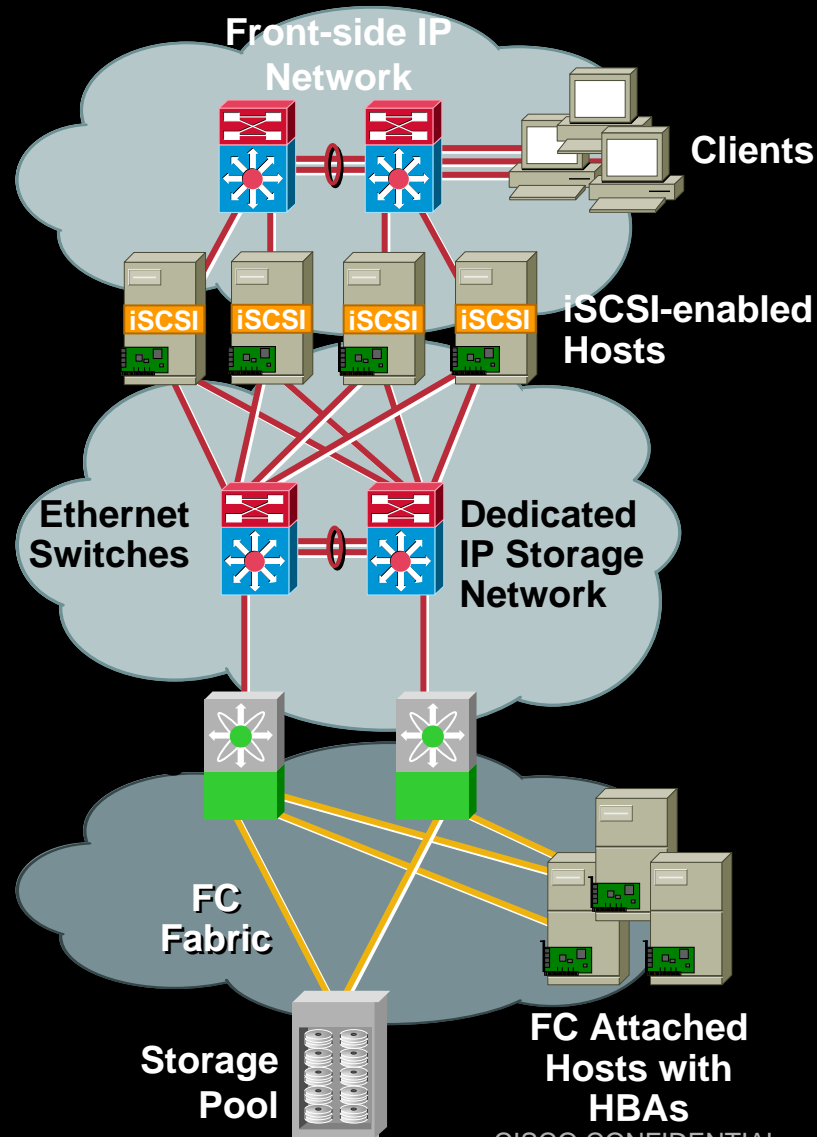
Security for IP Storage

- **Many services exist within IP to secure IP storage traffic, many of which are not available in Fibre Channel**
- **IPSec:**
 - FCIP: use IPSec hardware encryption to encrypt FCIP tunnels across the WAN/MAN
 - iSCSI standard calls for IPSec support; requires hardware acceleration in client; very vendor dependent
- **VLANs:**
 - Use VLANs to isolate storage traffic within LAN, consider Private VLANs
- **Access Control Lists (ACLs)**
 - Use IP and VLAN ACLs to isolate within a LAN; similar to zoning
 - Use storage-router based ACLs to restrict access to storage
- **Authentication:**
 - Use RADIUS/TACACS+ to authenticate iSCSI initiators, 802.1x for switch ports
- **Firewalls:**
 - Firewalls can be used due to static TCP ports

iSCSI Design: Dedicated IP Storage Network

Cisco.com

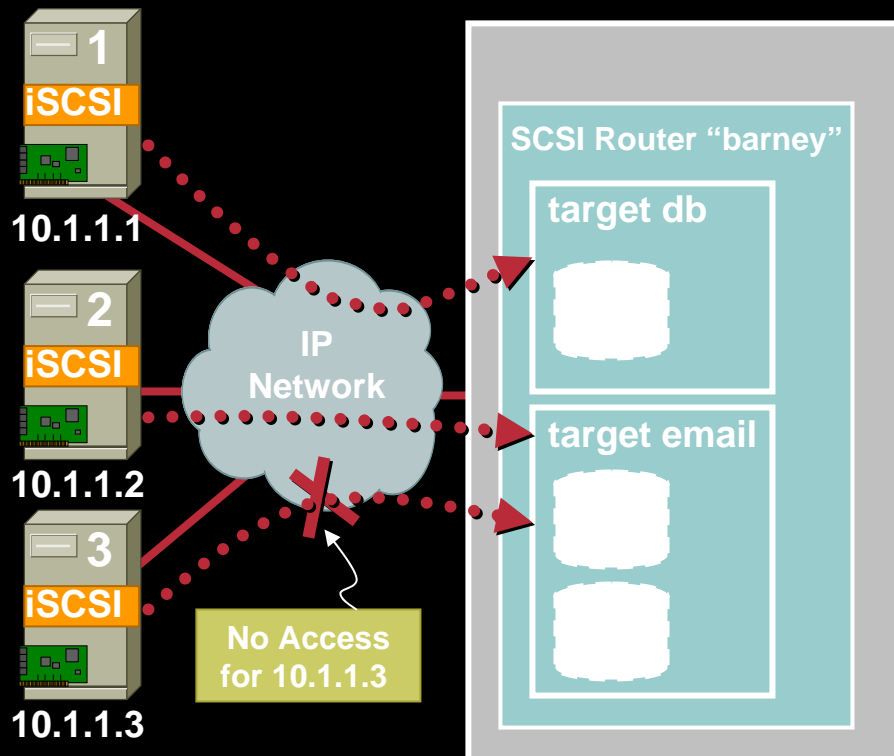
- Isolate IP storage network behind application hosts
- Minimized potential for bandwidth contention
- Can use a VLAN of existing Ethernet network
- Recommend use of dedicated Ethernet interfaces on host for attachment to storage network



iSCSI Access Control

- Access lists restrict access at the target level to specific hosts (initiators)

Hosts will only be presented with permitted targets during discovery/login



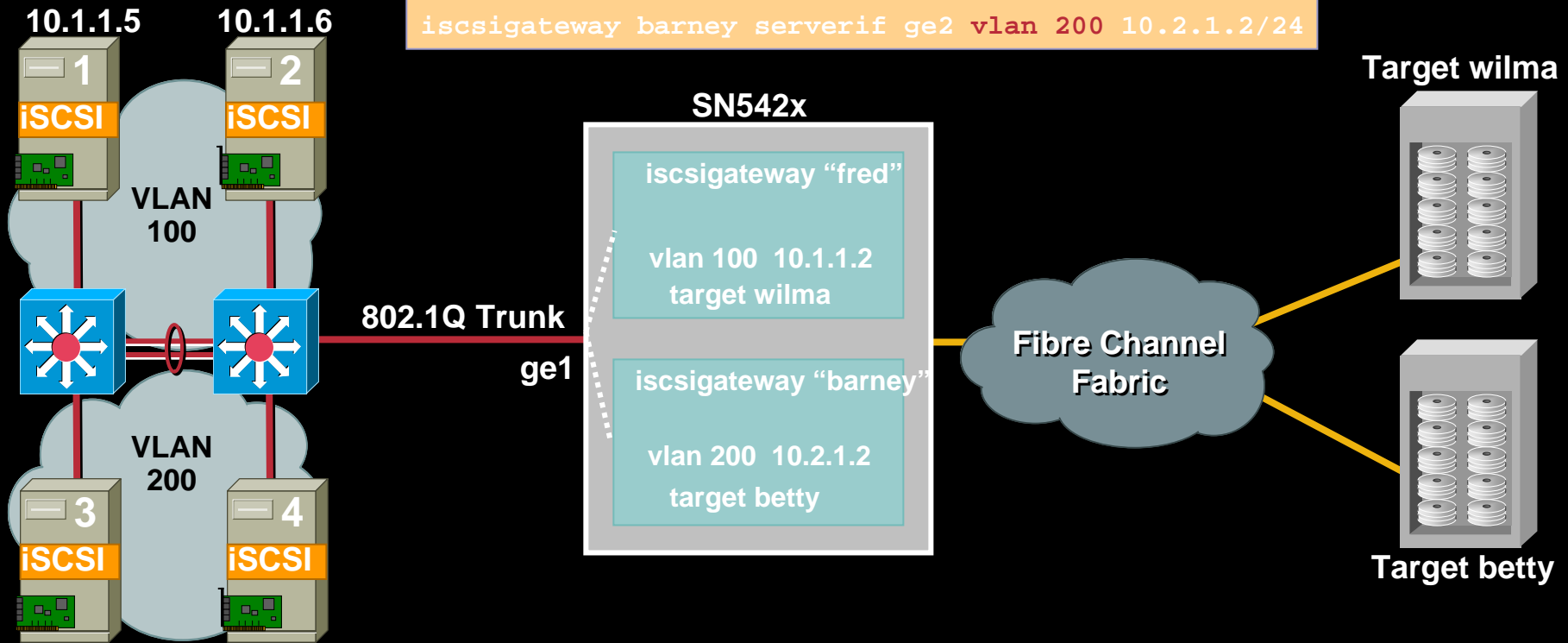
```
!
! ACCESSLIST
!
accesslist server1
accesslist server1 10.1.1.1/255.255.255.255
accesslist server2
accesslist server2 10.1.1.2/255.255.255.255
..
!
! iscsigatew
!
..
iscsigateway barney target db accesslist "server1"
..
iscsigateway barney target email accesslist "server2"
..
!
```

The code block shows the configuration for access lists and their application to targets. The first part shows the definition of two access lists: 'server1' for 10.1.1.1 and 'server2' for 10.1.1.2. The second part shows the application of these access lists to the targets in the router 'barney'. The 'server1' access list is applied to the 'target db', and the 'server2' access list is applied to the 'target email'. Red dashed boxes and arrows highlight the mapping between the access list definitions and their application to the targets.

Restricting Access with VLANs

Cisco.com

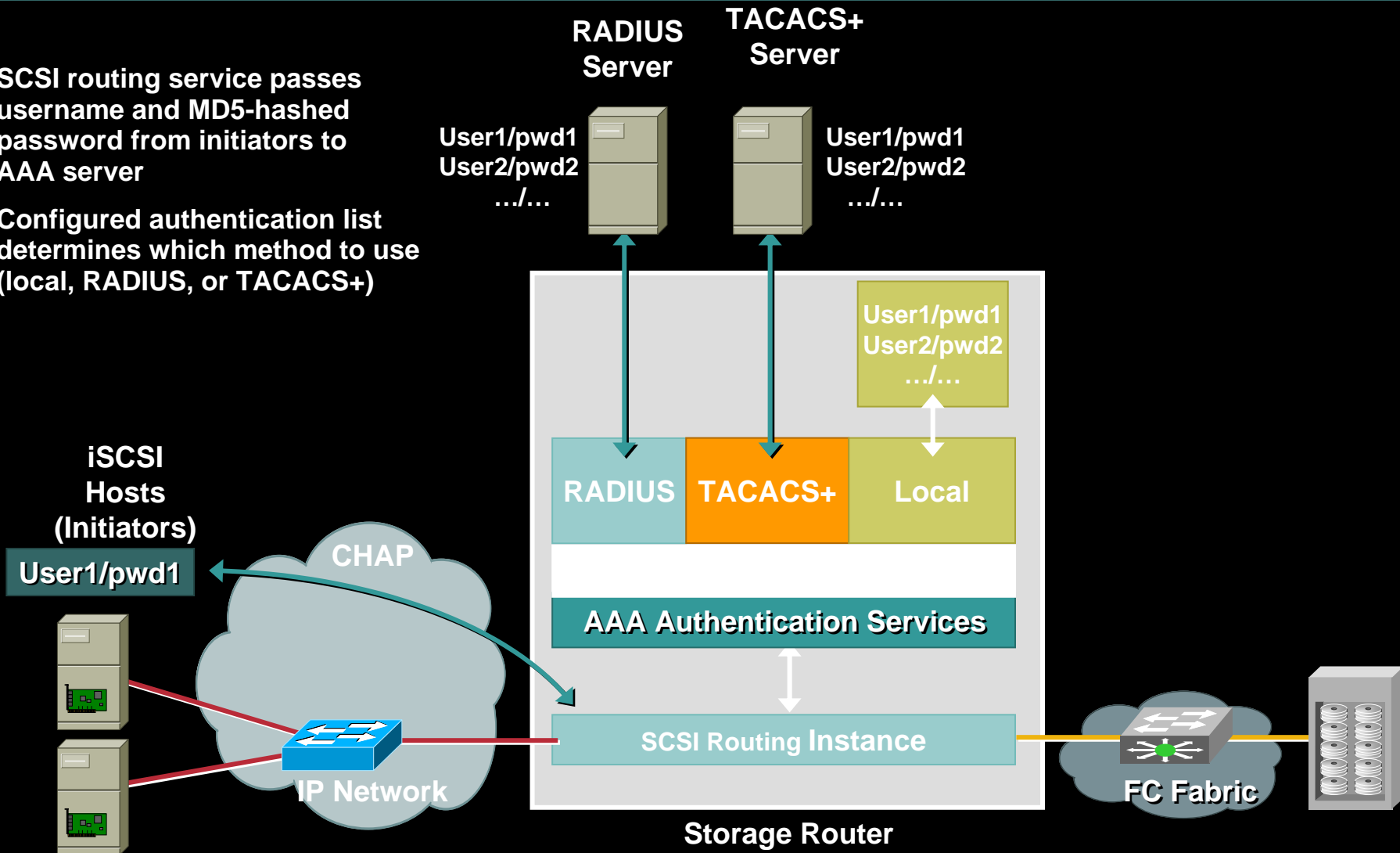
```
iscsigateway fred serverif ge2 vlan 100 10.1.1.2/24  
iscsigateway barney serverif ge2 vlan 200 10.2.1.2/24
```



- Servers 1 and 2 see only iscsigateway "fred" and defined targets "wilma"
- Servers 3 and 4 see only iscsigateway "barney" and defined targets "betty"
- Use access lists to restrict access further

iSCSI Authentication (Cont.)

- SCSI routing service passes username and MD5-hashed password from initiators to AAA server
- Configured authentication list determines which method to use (local, RADIUS, or TACACS+)



ISCSI I/O Profile

- **Windows NT/2000**—
“*perfmon*”

Write to .csv for
later analysis

Select disk bytes read/written
per second

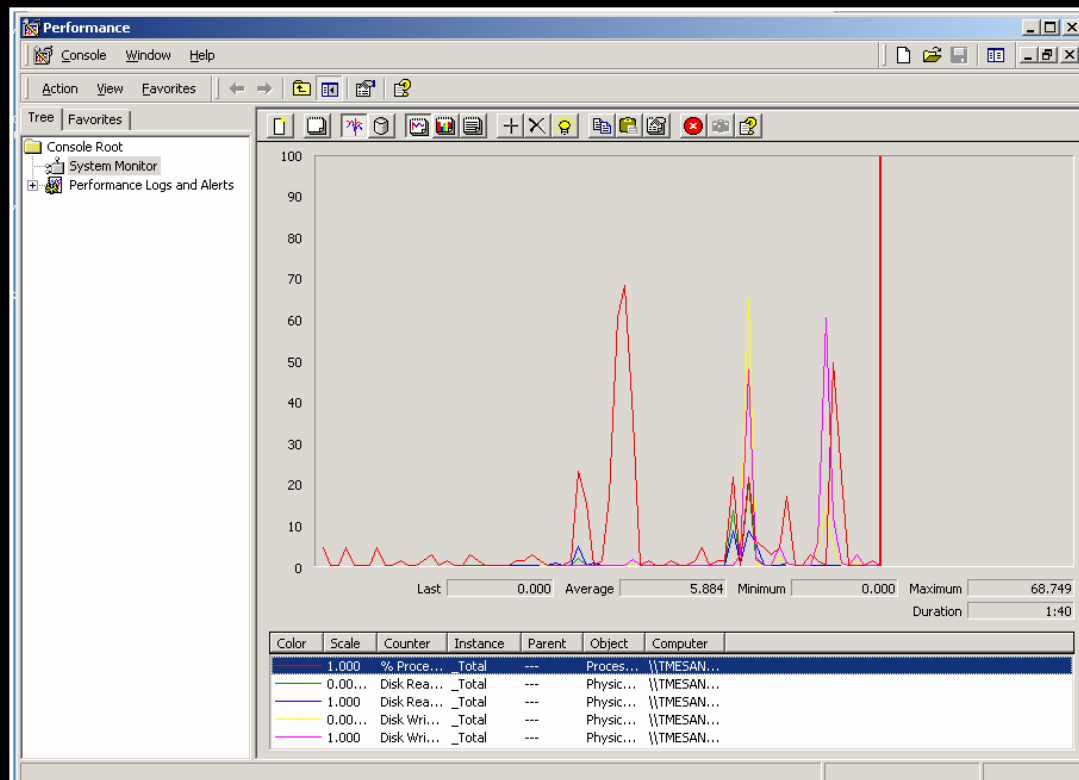
Monitor CPU load

- **Linux and UNIX**—“*iostat*”

-D option shows
I/O per sec

Use scripting to
monitor over time

Monitor CPU load
with “*ps*” or “*top*”



```
user@host% iostat -D /dev/dsk/c0t0d0s7
          sd15          sd16          sd36          sd75
rps wps util  rps wps util  rps wps util  rps wps util
  4  1  1.3    0  0  0.1    0  0  0.0    2  0  0.8
```

Microsoft Exchange Observations

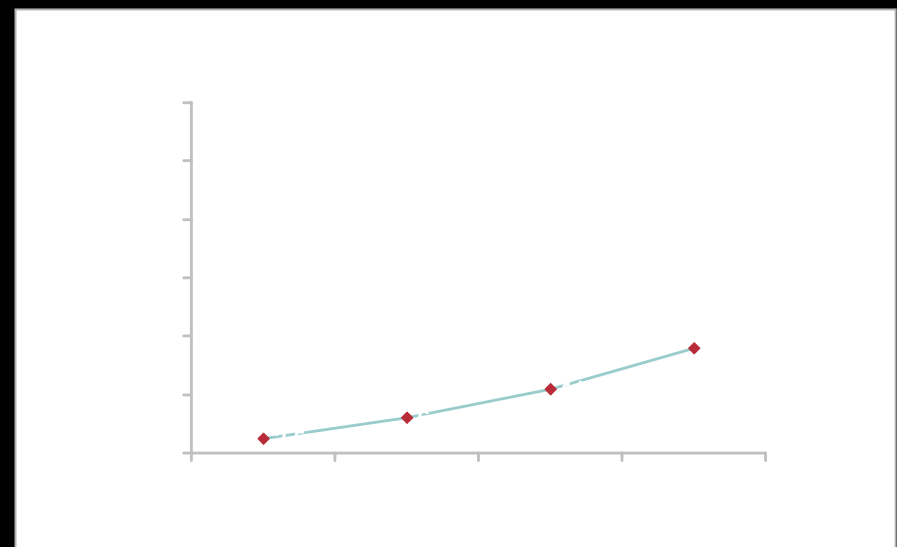
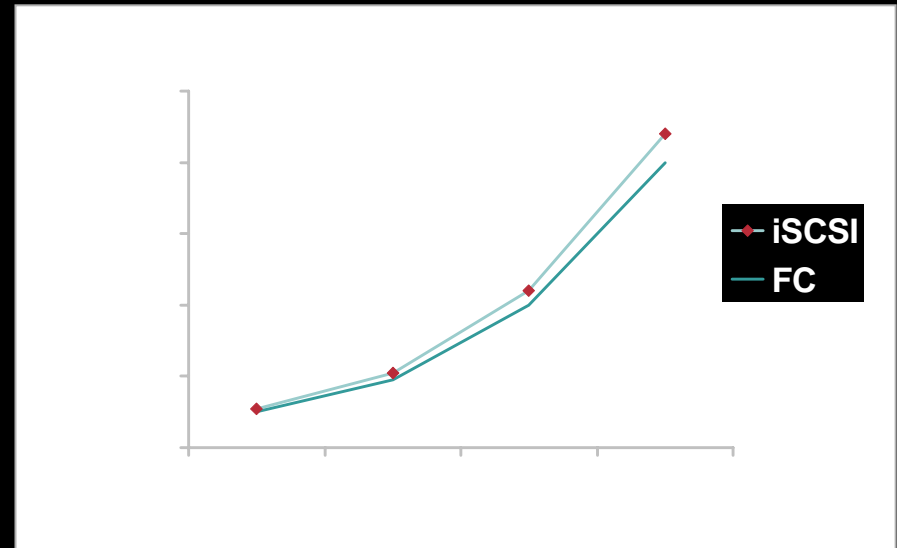
Cisco.com

- Microsoft “loadsim” generator
- Moderate I/O rate independent of storage size

Negligible difference between FC HBA and iSCSI software driver performance

TOE will reduce CPU%

- Server becomes scaling point rather than storage



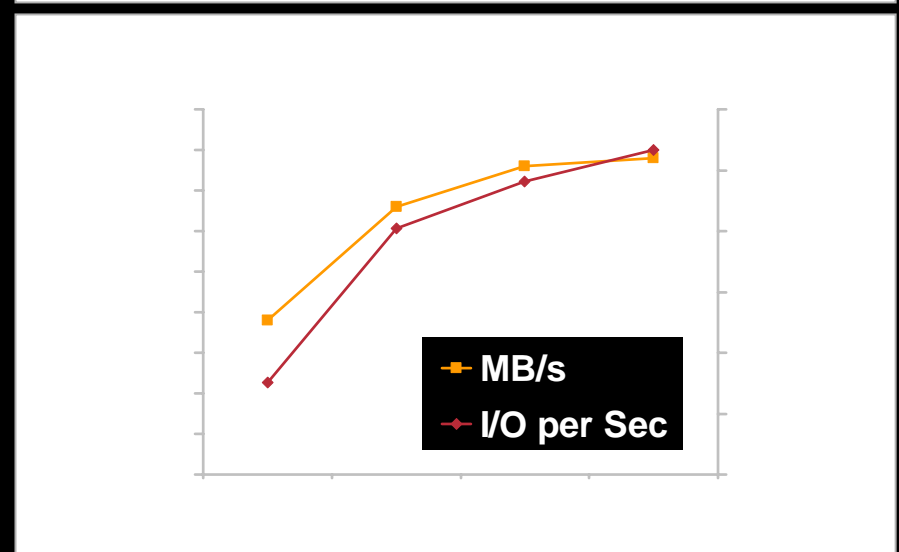
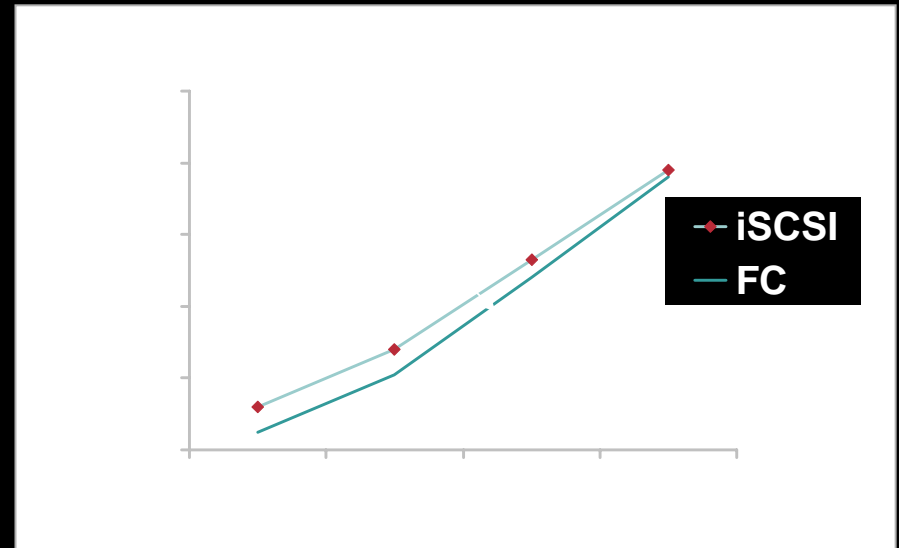
Microsoft SQL Server Observations

- **Moderate I/O rate**

**100/1000 NIC with iSCSI
adequate for most
apps...but dependent upon
app**

**Negligible difference
between FC HBA and iSCSI
software driver performance**

**Use TOE if CPU %
is a concern**



Agenda

- **Networking Storage**
- **ISCSI- Why and Where**
- **ISCSI- Design Considerations**
 - Discovery
 - Nic
 - High Availability
 - Security
 - IO
- **IP Storage Networking Looking Forward**
 - Network Boot
 - I-SER, I-Warp
- **Conclusion**

New Applications-Network Boot

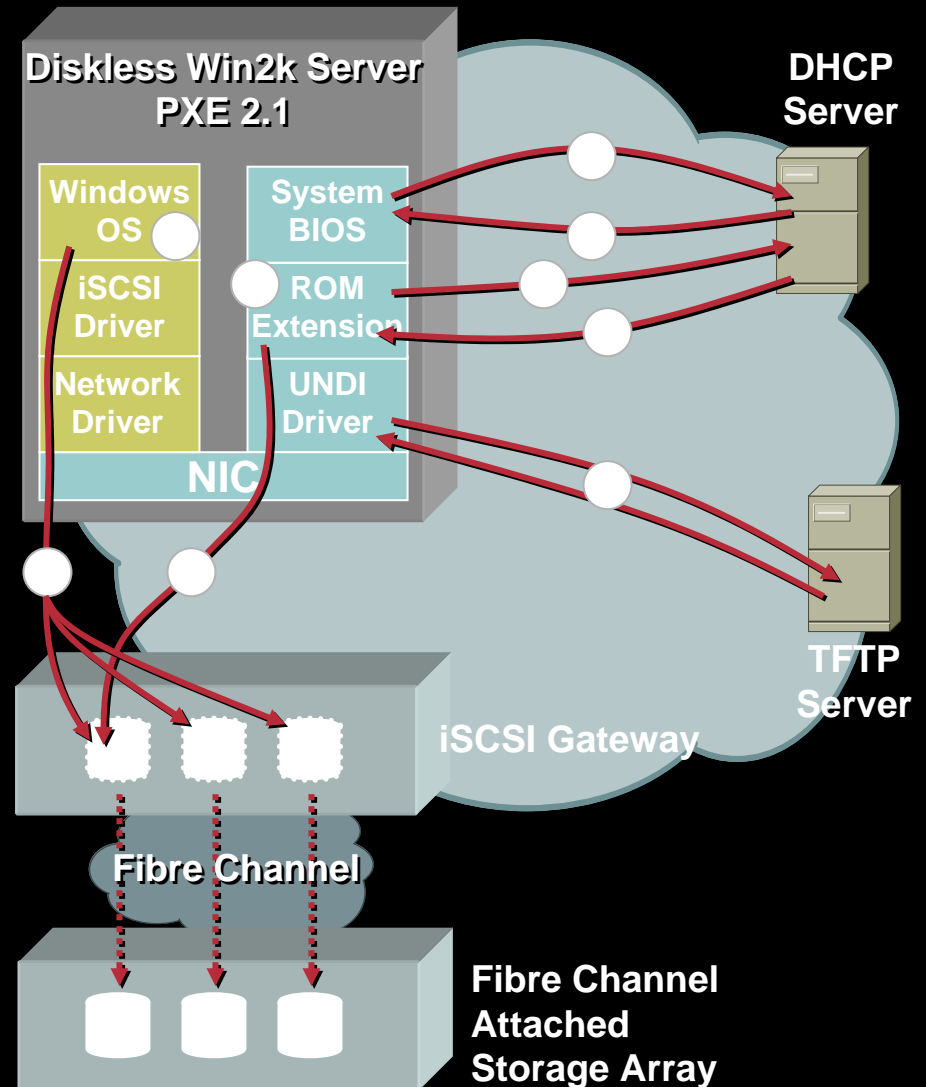
- Typical iSCSI client (e.g. host server) loads in the following order:
 1. Operating system (e.g. Windows 2000)
 2. Network
 3. iSCSI client driver
- How can you load the OS over iSCSI?

- Network boot uses PXE (Pre-Boot Execution Environment) capability present in many server BIOS and NICs (part of Intel's Wired for Management (WfM) spec)

Network and iSCSI client bootstrapped first allowing subsequent load of OS

Network PXE Boot—MS Windows Boot Sequence

1. BIOS sends DHCP request
2. DHCP server returns:
 - Server's IP address and g/way
 - TFTP server address and ROM extension filename
 - iSCSI server, target, and LUN
3. BIOS uses TFTP to fetch and execute "inbp.com" file
4. ROM extension sends DHCP request for "iSCSI Boot String"
5. DHCP server returns iSCSI server, iSCSI target, and LUN
6. ROM extension intercepts INT13 disk r/w and redirects to iSCSI server
7. BIOS reads C: drive (through "inbp.com" to load OS (Windows)
8. BIOS executes Windows OS and loads networks and iSCSI drivers
9. Windows uses iSCSI driver to access drives (normal operation)



Application of Network Boot

Cisco.com

- **Scaling of 1RU Servers and Blade Servers**

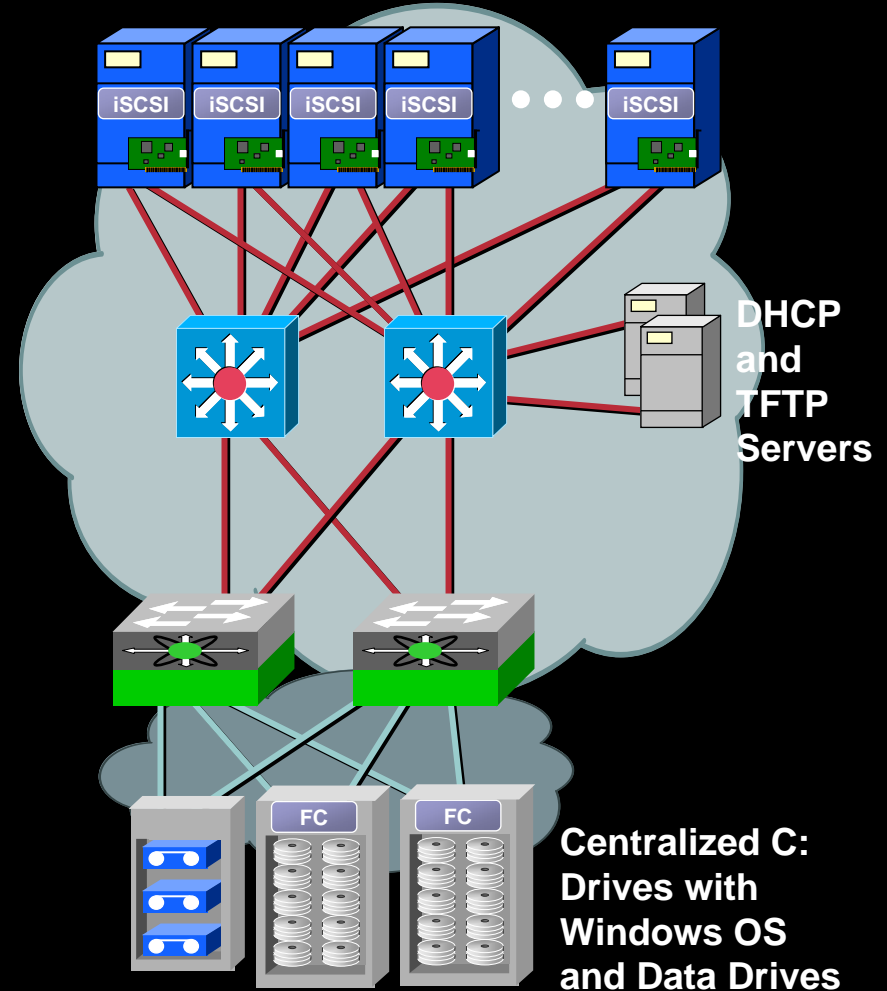
Replicated C: Drives
on storage arrays

- **Easy swap in/out
of servers**

Server expansion

Server failure

1RU Servers, Blade Servers, Clusters, etc...



ISER and IWARP

- **ISER (iscsi RDMA)**

iSER is a new IETF standard extension to iSCSI that includes support for multiple RDMA-based transports including InfiniBand and Ethernet RDMA.

Benefit:

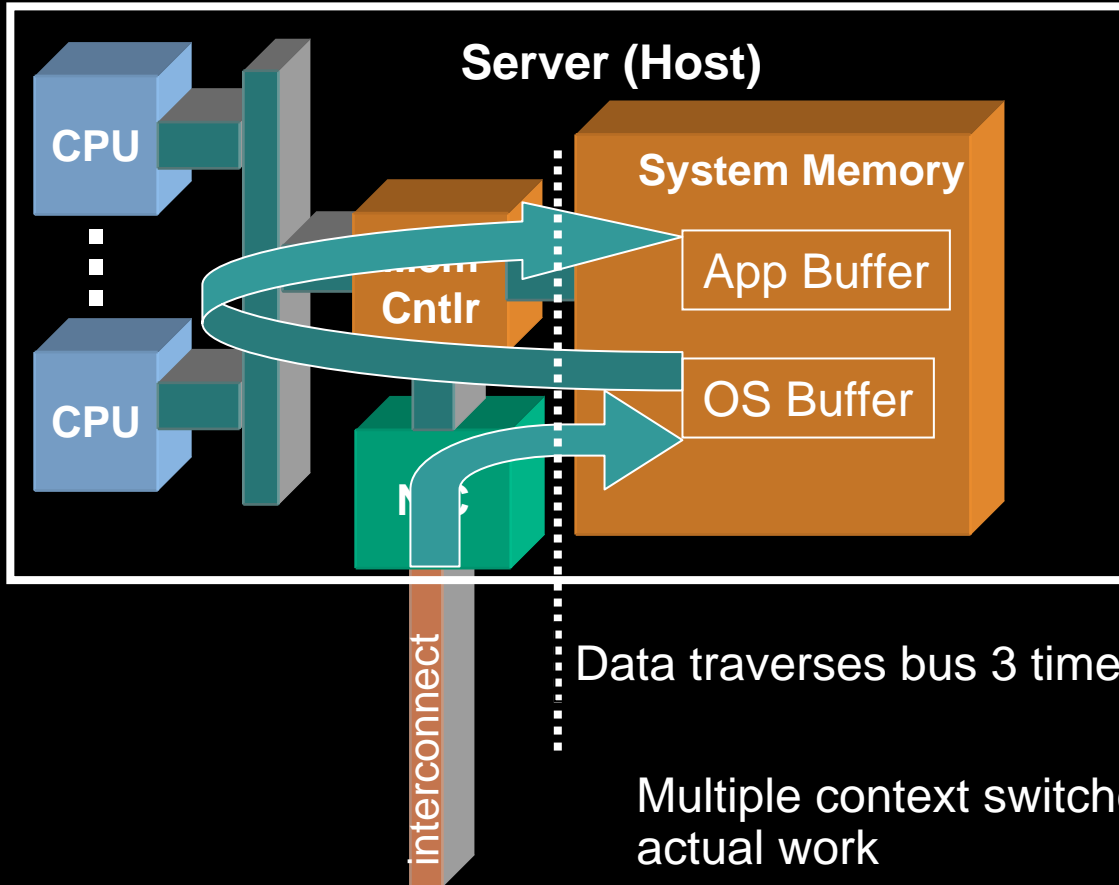
- **iSER brings significantly greater performance to iSCSI**
- **address storage using independent of whether the base network is an ethernet/IP network or an IB network**

- **iWARP enables RDMA to run over Ethernet networks via TCP/IP**

Benefit:

- **iWARP reduces latency and CPU cycles when applications running on separate machines share data directly into their peers' memories over a network.**

Copy on Receive

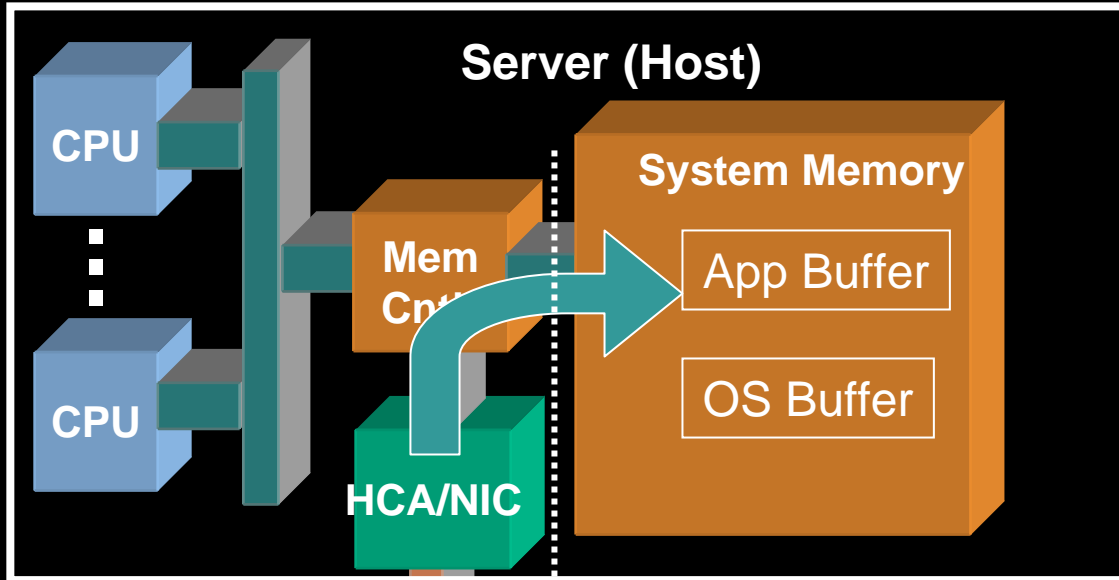


Data traverses bus 3 times

Multiple context switches robs CPU cycles from actual work

Memory bandwidth and per packet interrupts limit max throughput

With RDMA and OS Bypass



Data traverses bus once, saving CPU and memory cycles

Secure Memory – Memory transfers with no CPU overhead

PCI-X becomes the bottleneck for network data transmission

Agenda

- **Networking Storage**
- **ISCSI- Why and Where**
- **ISCSI- Design Considerations**
 - Discovery
 - Nic
 - High Availability
 - Security
 - IO
- **IP Storage Networking Looking Forward**
 - Network Boot
 - I-SER, I-Warp
- **Conclusion**

Conclusion

- **ISCSI works today for mid range applications**
- **ISCSI can be designed with HA and Security**
- **New protocols will extend iSCSI to higher end applications and performance**