

# An Introduction to iSCSI

Presenter:

Mel Tsai

[mtsai@eecs.berkeley.edu](mailto:mtsai@eecs.berkeley.edu)

11/13/2002

# Outline

- 1) The basics of SCSI
- 2) Introduction to iSCSI
  - 1) iSCSI vs. DAS/NAS/SAN
- 3) iSCSI Internals
  - Initiators, Targets, Connections, and Sessions
  - SCSI command encapsulation
  - Ordered Delivery, Naming
- 4) TCP/IP offload engines
- 5) Summary: the pros and cons of iSCSI
- 6) Status of iSCSI

# SCSI Basics

- SCSI: **Small Computer System Interface**
  - Derived from the ca. **1979** “Shutgart Associates System Interface” (SASI)
- The SCSI interface is used to attach **hard disk drives**, **CD-ROM drives**, and other peripherals (e.g. **scanners**) to a host machine

# SCSI Basics (cont.)

- **SCSI-1**, early 80's
  - Original specification, disks accessed via a Common Command Set (CCS), now OBSOLETE
- **SCSI-2**, mid-80's
  - Backward compatible with SCSI-1, adds (faster) parallel interfaces & broad support for **non-disk devices** (CDROM, tape, scanners)
- **SCSI-3**, early 90's to present
  - Adds many new standards, including the Fiber Channel Protocol (**FCP**), serial packet protocol, low-voltage differential (LVD) signaling

# Two faces of SCSI

- SCSI refers to **two important entities**:
  - The **physical transport**, i.e. the signaling & cabling for SCSI-compliant devices
  - The higher-level **data transmission protocol & formats**
- (iSCSI uses the SCSI data transmission protocol, **not** the physical transport)

# SCSI Transfer Modes

- There is also an important difference between SCSI standards & SCSI transfer modes
- "Regular" SCSI (SCSI-1)
- Wide SCSI
- Fast SCSI
- Fast Wide SCSI
- Ultra SCSI
- Wide Ultra SCSI
- Ultra2 SCSI
- Wide Ultra2 SCSI
- Ultra3 SCSI
- Ultra160 (Ultra160/m) SCSI
- Ultra160+ SCSI
- Ultra320 SCSI

# SCSI Features

- SCSI has a “**client-server**” architecture model
  - Clients are **initiators**
  - Servers are **targets**
- **Multiple SCSI initiators/targets** on a single physical bus
  - Up to **16 devices** on SCSI-2, prioritized by SCSI “Device ID” number
  - Targets can be further divided into **logical units**, e.g. individual disks in a multi-disk CD-ROM changer

# SCSI Features (cont.)

- Command **queuing** and **reordering**
  - Multiple outstanding commands can be served by the device a better order than issued
  - SCSI disks can offer higher performance in a **multi-user, multi-tasking** environment (e.g. vs. IDE/ATA drives)
- Rich command set:
  - **Standard** commands for formatting, polling, reading, writing, etc.
  - **Specialized** command sets for CD-ROMs, tape drives, scanners

# SCSI Command-Descriptor Blocks

Table 3 — Typical CDB for 12-byte commands

Bit Byte	7	6	5	4	3	2	1	0
0	OPERATION CODE							
1	Reserved			SERVICE ACTION (if required)				
2	(MSB)							
3								
4		LOGICAL BLOCK ADDRESS (if required)						
5								(LSB)
6	(MSB)							
7		TRANSFER LENGTH (if required)						
8		PARAMETER LIST LENGTH (if required)						
9		ALLOCATION LENGTH (if required)						
10		Reserved						
11		CONTROL						

# iSCSI: SCSI over IP

- SCSI is already used **everywhere**
  - In direct-attached storage (**DAS**)
  - In devices connected to network-attached storage (**NAS**)
  - As the FC layer-4 block I/O protocol in **Fibre Channel SANs**
- Why iSCSI?
  - A **low-cost** alternative to FC SANs: remote storage can be accessed via TCP/IP using the **block I/O of SCSI**
  - Better **resource utilization** because more clients can use the pooled storage than possible with FC SANs
  - **Interoperability**: FC SAN equipment is often vendor-specific

# iSCSI vs. SANs

- Wide-area coverage:
  - It is **costly** to extend FC SANs beyond 10 km
- iSCSI does not necessarily replace the FC SAN:
  - iSCSI provides a low-cost method for **extending the reach of FC SANs** by utilizing any existing IP infrastructure
  - Cisco makes an **FC-to-iSCSI router**
  - FC SANs and iSCSI can be **complementary**

# iSCSI vs. NAS

- Primary advantage of iSCSI vs. NAS: **Block I/O**
  - NAS uses file I/O access such as **NFS** or **CIFS**
  - File I/O limits the performance of **datacenter applications**
- iSCSI makes remote storage look like a **local SCSI drive**
  - Simplifies management and setup of clients... all you need is an Ethernet connection
  - Clients can structure data however they like, i.e. native ext3 or NTFS
  - Backup is achieved by copying to the “local” SCSI disk

# iSCSI Deployment

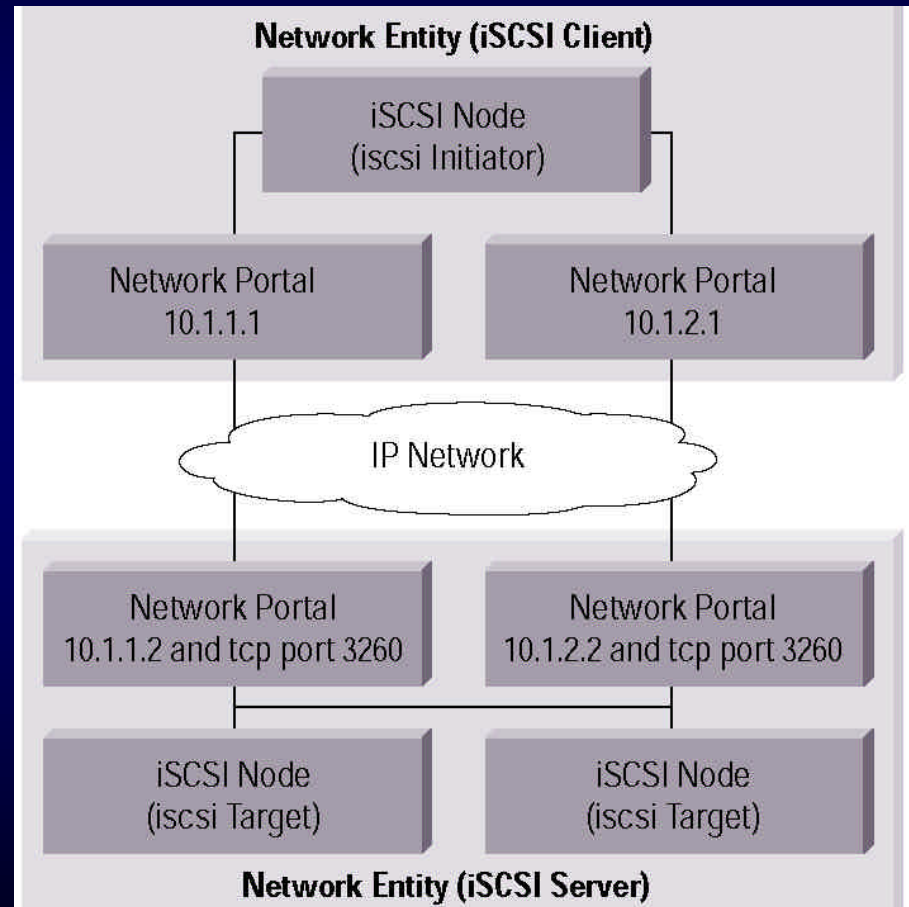
- A major disadvantage of iSCSI and NAS is **increased LAN utilization**
  - Higher latency when the LAN is saturated, especially during backups
- Possible solution:
  - Build a **dedicated IP network** for your iSCSI traffic
  - This is still **cheaper than FC**

# iSCSI Competitors:

- Fibre Channel over IP (**FCIP**)
  - A “dumb” point-to-point IP tunnel between FC SANs
- Internet Fiber Channel Protocol (**iFCP**)
  - Slightly smarter than FCIP... Encapsulation FC traffic in IP, maps individual FC devices to IP addresses

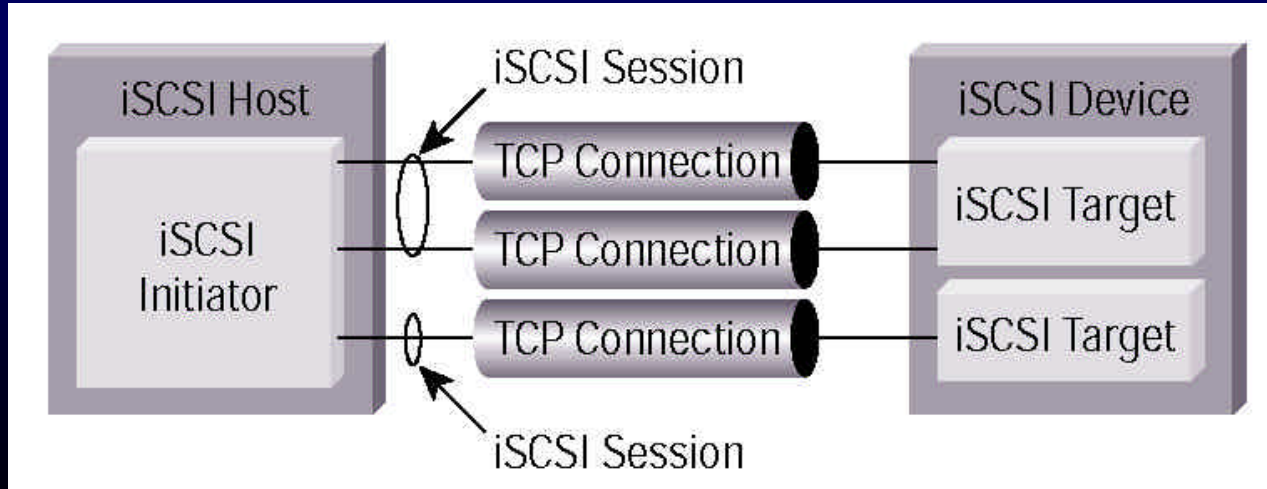
# iSCSI Basic Operation

- Terminology:
  - Network Entities
  - iSCSI Nodes
    - Initiators
    - Targets
  - Network Portals



# iSCSI Sessions

- Only **one iSCSI session** can exist between an initiator/target pair
  - Normal sessions
  - Discovery sessions
  - Session IDs (SSIDs)
- **Multiple** parallel TCP/IP connections can exist in a session (CIDs)



# An iSCSI Session

- An iSCSI initiator **logs in** to an iSCSI target after establishing a TCP connection
  - Various methods of client authentication
- After client authentication, a session is initialized
  - Via a driver on the client, the session encapsulates local SCSI commands into **remote iSCSI commands** for the target
  - This is the “full-feature phase”
  - Block I/O data can be transferred securely, e.g. via IPsec
- Once finished, the session is terminated (logout & shutdown) by either the initiator or target

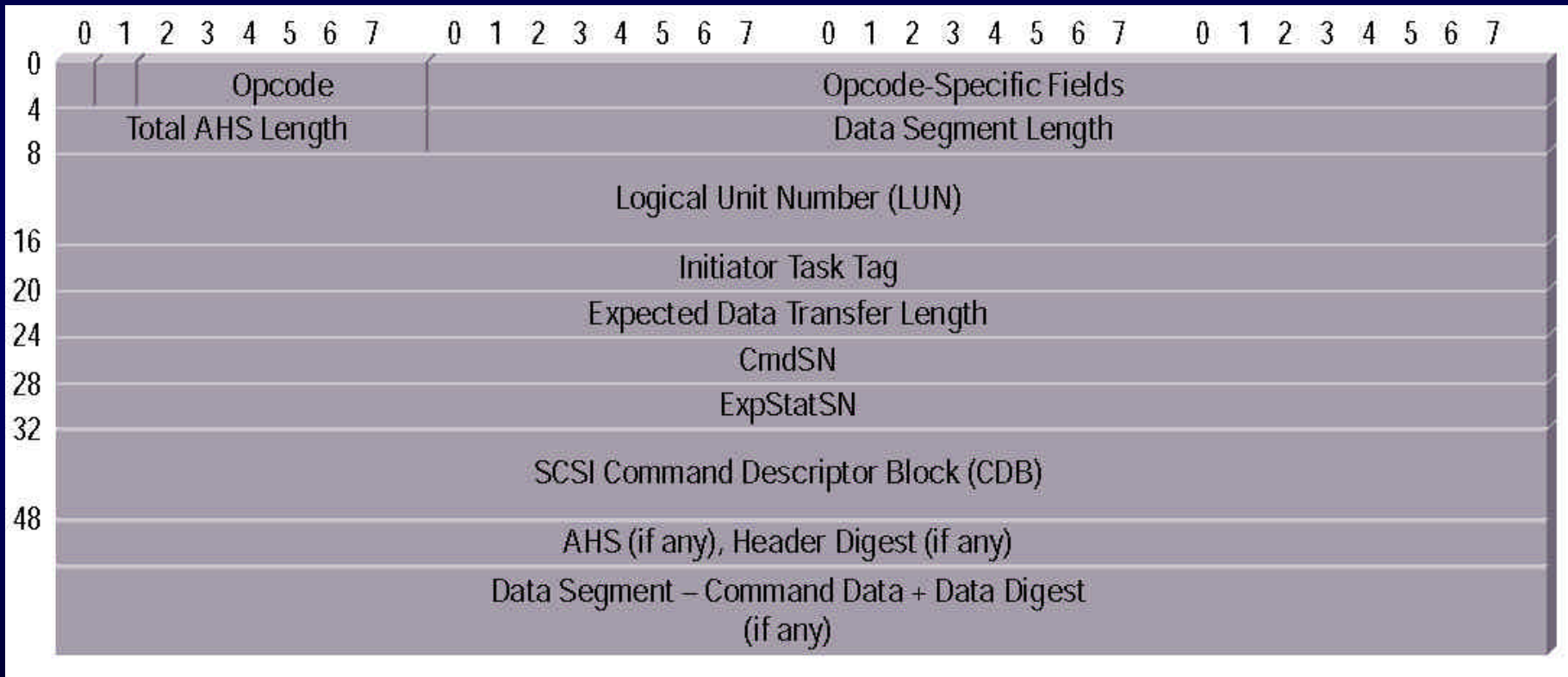
# Some goals and requirements of iSCSI

- Does **not** require modification of TCP/IP infrastructure
- An initiator can attach to **multiple** network portals (IP addresses) on a single target
- iSCSI sessions can operate over a **single TCP/IP connection** and use TCP/IP connections **conservatively**
- iSCSI should support all **SCSI-3 command sets**
  - New feature of SCSI-3: device-to-device copy

# iSCSI Internals

- iSCSI wraps a local SCSI command into an iSCSI protocol data unit (PDU) **request**
  - PDUs **wrap** the SCSI command descriptor blocks (CDBs)
  - The CDB and other info is placed in the PDU's **Basic Header Segment** (BHS)
- SCSI responses and status info from the target are returned as iSCSI PDU **responses**

# An iSCSI PDU



# In-order delivery of messages

- Sequence numbers, similar to TCP/IP
  - Commands
    - **Command sequence number** (maintained by **initiator**)
    - **Expected** command sequence number (maintained by **target**)
    - **Maximum** command sequence number (maintained by **target**)
    - “Immediate delivery” commands
  - Status
    - Status sequence number (maintained by **target**)
    - Expected status sequence number (maintained by **initiator**)
  - Data
    - Data sequence numbers (**reading** from target)
    - Request-to-transfer sequence numbers (**writing** to target)

# iSCSI Device Naming

- All iSCSI initiators/targets get a globally-unique **World-Wide Unique Identifier (WWUI)**
- The WWUI is nice because:
  - **Multiple** iSCSI targets can exist behind a **single IP address**
  - The **same target** can exist behind **multiple IP addresses**
  - Initiators and targets can be referred to **symbolically**, independent of their location
- A “complete” **iSCSI address** is made up of:
  - An IP address
  - A TCP/IP port (typically 3260)
  - The WWUI

# The iSCSI Killer: TCP/IP Overhead

- TCP/IP connections are **expensive at high data rates**
  - Connection establishment & teardown
  - Out-of-order **packet reassembly**
  - Error detection, **packet retransmission**
  - Expensive **memory copying** between protocol layers
- Transferring just 32 KB of data via TCP/IP can involve over **30 transactions** between the NIC and CPU (20 data packets, 10 ACKs)
- Often-cited “rule” for TCP/IP overhead on a server:
  - You need a dedicated **1 GHz processor** for **1 Gbit of TCP/IP traffic**, and a 10 GHz processor for 10 GbE

# Solution: TCP offload engines (TOEs)

- Implement the layer-4 TCP/IP stack with a **separate** CPU, NPU, or ASIC
  - Now you can **present the session layer (5)** to the host
- TOEs can be integrated into **standard Ethernet cards, iSCSI host adapters**, or other iSCSI equipment
- TOEs will become an **absolute requirement** at speeds above 1 Gbps due to server memory bandwidth limitations
  - Even **PCI-X bandwidth** cannot withstand 10 GbE without modification
- On the bright-side: Other applications (besides SANs) will soon need TOEs!

# Summary: iSCSI Pros

1. iSCSI transforms directed-attached disks to **network-attached, block I/O devices**
2. Existing SCSI RAID devices, tape libraries, etc. can be easily **migrated** to your low-cost Ethernet iSCSI SAN with easy management
3. If the iSCSI network is separate from the primary LAN, you can get **LAN-free** and **server-free** backup
4. Block I/O is better than File I/O for **datacenter** apps
5. **High security**
6. Potentially **very fast**... FC will be 2 Gbps tomorrow, while iSCSI could be 10 GbE today
7. **Interoperability** vs. FC
8. No **distance limitations**

# Summary: iSCSI Cons

1. A standard 1 Gbps FC network is **fundamentally faster** than iSCSI at 1 Gbps due to protocol overhead
2. ASIC- or NPU-based **TOEs** are required for high performance
3. You still **need a separate iSCSI IP network** to achieve LAN-free backup
4. Others?

# Where is iSCSI today?

- The iSCSI standard(s) are **not yet finished**
- There are several existing iSCSI products on the market, but it has a while to go
  - **60 corporations** are on the iSCSI working group of the SNIA.
  - Several shipping **iSCSI host bus adaptors**
  - Cisco makes an **iSCSI to FC gateway**
- Cisco has written a public domain **Linux iSCSI kernel driver**
  - A simple kernel module that implements an iSCSI initiator
  - Turns a **remote iSCSI target** (by specifying its IP address) into **/dev/sda** or **/dev/sdb**, etc.

# References & Sources

- The PC Guide, [www.pcguides.com](http://www.pcguides.com)
- SCSI Primary Commands - 2 (SPC-2), ISO/IEC 14776-312
- Cisco Whitepaper: “Introduction to iSCSI”
- “IP Storage Networking: IBM NAS and iSCSI Solutions,” IBM Redbook
- 10 Gigabit Ethernet Alliance: Introduction to TCP offload engine
- “The ins and outs of interconnects,” [nwfusion.com](http://nwfusion.com)
- “Inside iSCSI: Low-Cost Storage Networking”, [extremetech.com](http://extremetech.com)