INTRODUCTION TO STORAGE TOPOLOGIES AND APPLICATIONS
SESSION OPT-1051

Storage 101

- **Storage Terms**
  - Host Bus Adapter
  - Disks, JBODs, RAID, Storage Arrays

- **Storage Protocols**
  - SCSI
  - Storage Transport Protocols
    - Fibre Channel (FC), iSCSI, TCP/IP

- **Storage Topologies**
  - DAS, NAS and SAN

- **Storage for Applications**
  - Storage Access Patterns for Multi-Tiered Applications
Storage Terms

- Host Bus Adapter (HBA)
- Storage I/O devices
- Small Computer Serial Interface (SCSI)

- Host Bus Adapter (HBA)
- Storage subsystem
  - Just a bunch of disks
  - Redundant array of independent disks (RAID)
  - Storage arrays
Storage Terms

- Host bus adapter
- Hardware resident on host server
- Connection interface to disk subsystem
- Connection method
  - Copper
  - Optical

Storage Interconnects

- Parallel SCSI copper interconnects
- Optical direct connect
- Fiber channel switch
Storage Terms

I/O Devices—Disk Drives

- Fundamental unit for data storage
- Disk drive types
  - Parallel SCSI
  - Fibre channel
  - Advanced Technology Attachment (ATA) or Integrated Drive Electronics (IDE)
  - Serial ATA (SATA)
Storage Terms

I/O Devices—Disk Drives Components

- The platters are made of aluminum alloy or glass substrate
- Platters are accessed for read and write operations using 2 read/write heads

I/O Devices—JBODs

Server With SCSI

```
  TRGT 0  TRGT 1  TRGT 2  TRGT 3
SCSI—JBOD
```

Server With Fibre Channel Arbitrated Loop (FC-AL)

```
  TRGT 0  TRGT 1  TRGT 2  TRGT 3  TRGT 4  TRGT 5
FC-AL—JBOD
```

Server With Fibre Channel Arbitrated Loop (FC-AL)

```
  TRGT 0  TRGT 1  TRGT 2  TRGT 3  TRGT 4  TRGT 5
```

iSCSI

```
  TRGT 0  TRGT 1  TRGT 2  TRGT 3  TRGT 4  TRGT 5
```

SCSI—JBOD

```
  TRGT 0  TRGT 1  TRGT 2  TRGT 3  TRGT 4  TRGT 5
```

FC-AL—JBOD

```
  TRGT 0  TRGT 1  TRGT 2  TRGT 3  TRGT 4  TRGT 5
```
Storage Terms

I/O Devices—RAID

• Redundant Array of Inexpensive Disks (RAID)
  
  Word coined by researchers at University of California, Berkeley in 1987
  
  A method to inexpensively put together a set of physical hard drives into an array
  
  Provides fault tolerance by mirroring or parity operations
  
  RAID can be performed using hardware or using host based software

• RAID 0
  
  For performance and size, two or more hard disks are concatenated together to form a larger volume

• RAID 1 (mirroring)
  
  For reliability and availability

• RAID 3
  
  For reliability, availability of data using using a dedicated parity drive

• RAID 5
  
  For reliability, availability of both data and parity
Storage Terms

I/O Devices—RAID0: Striping

- Data is segmented and split onto multiple spindles
- I/O benefits
  - Short reads and writes, easily handles multiple simultaneous reads
  - Long reads and writes, single operations split and processed in parallel
- Redundancy—none
- Cost—good (no extra hardware)

I/O Devices—RAID: Mirroring

- Data is duplicated on multiple spindles
- I/O benefits
  - Short and long reads with shorter latency
  - Short and long writes slower due to multiple writes
- Redundancy—good
- Cost—need double the amount of disks
Storage Terms

I/O Devices—RAID 0+1

- Data is striped and duplicated on multiple spindles
- I/O benefits
  - Short and long reads with shorter latency
  - Short and long writes faster as writes are spread across multiple spindles
- Redundancy—good
- Cost—need double the amount of disks

I/O Devices—RAID 1+0

- Data is striped and duplicated on multiple spindles
- I/O benefits
  - Short and long reads with shorter latency
  - Short and long writes faster as writes are spread across multiple spindles
  - Lesser downtime than RAID 0+1
- Redundancy—good
- Cost—need double the amount of disks
Storage Terms

I/O Devices—RAID 3

- Data protection using ECC (error correction control) parity disk
- I/O benefits
  - Short and long reads at normal speeds
  - Short and long writes slower due to parity calculations
- Redundancy—good
- Cost—more optimal than RAID 0+1 or 1+0

I/O Devices—RAID 5

- Data protection using ECC (error correction control) parity spread over all drives
- I/O benefits
  - Short and long reads at normal speeds
  - Short and long writes slower due to parity calculations
- Redundancy—better than RAID3
- Cost—needs only one extra disk for an entire logical disk
Storage Terms

I/O Devices—Intelligent Storage Arrays

- Host Channel Connections
- Dual-Processor Channel Director Cache
- System Bus
- Dual-Processor Disk Director
- Disks

Host Bus Adapters

Wow, these diagrams are quite complex. It seems like they're showing the relationships between different components in an intelligent storage array. The Host Channel Connections and Dual-Processor Channel Director Cache are connected to the System Bus, which in turn connects to the Dual-Processor Disk Director and Disks.
Storage Protocols

Small Computer System Interface—SCSI

- SCSI is a **STANDARD** that defines an interface between an **initiator** (usually a computer) and a **target** (usually a storage device such as a hard disk).
- **INTERFACE** refers to connectors, cables, electrical signals, optical signals and the command protocol that allow initiators and targets to communicate.
Storage Protocols

Small Computer System Interface—SCSI

- **SCSI command protocol** is the de facto standard that is used extensively in high-performance storage applications
- The **command** part of SCSI can be:
  - Transported over a Fibre Channel storage area network
  - Encapsulated in IP and carried across IP networks

SCSI Commands

- Data transfer
  - READ, WRITE
- Commands used in boot/discovery:
  - REPORT LUNS
  - INQUIRY
  - TEST UNIT READY
Storage Protocols

SCSI Read

1. **Send SCSI Cmd** issued by initiator—the command sent is READ;
2. **SCSI command received** by target;
   
   Data transfers occur during the 'working' phase between initiator and target;
3. **Send command complete** is returned by the target;
4. **Command complete received** by target
Storage Protocols

SCSI Write

1. Send SCSI Cmd issued by initiator—the command sent is WRITE;
2. SCSI command received by target;
3. Target returns TRGT-RDY;
4. Initiator receives TRGT-RDY; Data transfers occur during the ‘working’ phase between initiator and target;
5. Status complete is returned by the target;
6. Status complete received by initiator

Storage Protocols

SCSI Write

<table>
<thead>
<tr>
<th>Initiator</th>
<th>Physical Transport Bus (Copper Wire)</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>Arbitration Selection</td>
<td>Disk</td>
</tr>
<tr>
<td></td>
<td>Write</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16k Data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Command Out</td>
<td></td>
</tr>
</tbody>
</table>

SCSI

Message Out

Data in, 16 Kbytes

Status In

Message In

Time
Storage Protocols

Storage Transport Protocols

• Protocols used to transport SCSI
  Fibre Channel
  FCIP
  iSCSI

Storage Transport Protocols—Fibre Channel

FCP Goes Here (FC-4)

FC-3
FC-2
FC-AL
FC-1
FC-0
Fibre Channel Protocol Hierarchy

- FC offers a Layer 2 service
- FC frames always begin and end with a delimiter (ordered set: SOFi3, SOFn3, EOFn, EOFt...)
- Each frame carries a source identifier and a destination identifier
- An unidirectional flow of related frames between an Initiator and a recipient is called a sequence
- An unidirectional or bi-directional series of sequences between an originator and a responder is called an exchange
Storage Protocol

Fibre Channel Protocol—Essentials

- FCP defines how SCSI-3 commands are transported over a fibre channel network
- Transport involves more than encapsulation
- Individual SCSI commands are mapped to FC sequences
- The set of SCSI commands that form a SCSI transaction are mapped to a FC exchange
- Fibre channel preserves the frame order

---

Storage Protocols

Fibre Channel Protocol—SCSI—FC Mapping

<table>
<thead>
<tr>
<th>SCSI function</th>
<th>FCP equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O operation</td>
<td>Exchange</td>
</tr>
<tr>
<td>Protocol Service Request and Response</td>
<td>Sequence</td>
</tr>
<tr>
<td>Send SCSI Command Request</td>
<td>Unsolicited command IU (FCP_CMND)</td>
</tr>
<tr>
<td>Data delivery request</td>
<td>Data descriptor IU (FCP_XFER_RDY)</td>
</tr>
<tr>
<td>Data delivery action</td>
<td>Solicited data IU (FCP_DATA)</td>
</tr>
<tr>
<td>Send Command Complete Response</td>
<td>Command status IU (FCP_RSP)</td>
</tr>
<tr>
<td>REQ/ACK for Command Complete</td>
<td>Confirmation IU (FCP_CONF)</td>
</tr>
</tbody>
</table>
Storage Protocols

Fibre Channel Protocol—Read

- Initiator (Server)
- Fibre Channel
- Fibre Channel
- SCSI Initiator
- Physical Transport
- Fibre Channel Arbitration Selection
- Fibre Channel
- Target Disk
- Data in 16 K Bytes
- Status In
- Message In
- Command Out
- Message Out
- FC Frame
- 8 Frames \( \leq 16\) Bytes
- Data in, 16 Kbytes

Fibre Channel Protocol—Write

- Initiator (Server)
- Fibre Channel
- Fibre Channel
- SCSI Initiator
- Physical Transport
- Fibre Channel Arbitration Selection
- Fibre Channel
- Target Disk
- Data in 16 K Bytes
- Status In
- Message In
- Command Out
- Message Out
- FC Frame
- 8 Frames \( \leq 16\) Bytes
- Data in, 16 Kbytes

Storage Protocols

Storage Protocols—iSCSI

- SCSI data converted into an iSCSI by adding a special header
- iSCSI data is encapsulated into an IP packet
- IP is the transport protocol

![Diagram showing IP, TCP, iSCSI, and SCSI layers]

---

Storage Protocols

Storage Protocols—iSCSI

<table>
<thead>
<tr>
<th>SCSI Applications (File Systems, Databases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCSI Block Commands</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>SCSI Commands, Data, and Status</td>
</tr>
<tr>
<td>Parallel SCSI interfaces</td>
</tr>
<tr>
<td>TCP</td>
</tr>
</tbody>
</table>
Storage Protocols

Storage Protocols—iSCSI—SCSI Relationship

- **SCSI device**
  - IP host
  - iSCSI node initiator
  - FC storage
  - iSCSI node target

- **SCSI port**
  - iSCSI initiator port
  - iSCSI target port

- **Network portal**
  - Any network interface with TCP/IP
  - Provides physical IP network access

---

Storage Protocols

iSCSI Reads

- **Initiator Server**
  - Read
  - TCP/IP
  - IP Packets
  - Message Out
  - Command Out

- **Physical Transport**
  - iSCSI Arbitration Selection

- **Target Disk**
  - FC Frame
  - Command Out
  - Data in, 16 Kbytes
  - 8 Frames
  - 16K Bytes

- **Status in**
  - Message in

---

Cisco.com
Storage Protocols

**iSCSI Writes**

- Initiator Server
- TCP/IP
- IP Packets
- Message Out
- Command Out
- IP Packets
- iSCSI Router
- Command Out
- FC Frame
- Target Disk

**Physical Transport**

- iSCSI
- Arbitration Selection
- FC
- SCSI

**SCSI**

- Data in, 16 Kbytes
- Message Out
- 8 Frames
- 16K Bytes
- Status in
- Message in

---

**Storage Protocols—FCIP**

- FCIP is a mechanism that allows SAN islands to be interconnected over IP networks
- Each interconnection is called a FCIP link and can contain one (1) or more TCP connection(s)
- Each end of a FCIP link is associated to a Virtual E_Port (VE_Port)
- VE_Ports communicate between themselves just like normally interconnected E_Ports by using SW_ILS: ELP, ESC, BF, RCF, FSPF, etc.
- The result (when all goes well…) is a fully merged fibre channel fabric
Storage Protocols

Storage Protocols—FCIP

- The FCIP Link carries encapsulated fibre channel traffic between Link End Points (LEPs) over an IP network by using TCP on port 3225
- The result is a virtual ISL between VE_Ports
- FC frame not changed

![Diagram showing FCIP packets and protocols]

- FC = Fibre Channel
- F = F_Port
- E = E_Port
- VE = Virtual E_Port
- GE = Gigabit Ethernet
Storage Protocols

Storage Protocols—FCIP

ANSI: FC-BB-2 Essentials

- Defines a slightly complex model;
- FC-BB-2 covers the FC portion of this model (FC entity and above);
- Cisco’s FCIP implementation for IPS-8 closely follows this model

Note: DE within Each FCIP_LEP Means FCIP_DE

Storage Protocols—FCIP

IETF: Fibre Channel over TCP/IP Essentials (FCIP)

- FCIP follows the model proposed in FC-BB-2
- FCIP covers the lower portion of this model (FCIP entity and below)
- Cisco’s FCIP implementation for IPS-8 follows this model

Note: DE within Each FCIP_LEP Means FCIP_DE
Storage Topologies

Direct Attached Storage

- Direct Attached Storage (DAS)
- Storage is captive ‘behind’ the server, limited mobility
- Limited scalability due to limited devices
- No storage sharing possible
- Costly to scale; complex to manage
- 80% of storage market is direct-attached storage
Storage Topologies

Network Attached Storage

- Storage is accessed at a file level via NFS or CIFS
- Storage is accessed over an IP network
- Storage devices can be shared between servers—files can be shared between users
- TCP can be tuned to optimize for storage transport

**NAS = Optimized for File I/O**

Storage Area Networks

*Separation of Storage from the Server*

- Storage is accessed at a block-level via SCSI
- High performance interconnect providing high I/O throughput
- Lower TCO relative to direct attached storage, storage can be shared
- Limited vendor interoperability and data access
- Complex management mainly due to multi-vendor interoperability concerns
Storage for Applications

Storage Design Criteria

- Multi-tier definition and general architecture
- Data characteristics
- Data access patterns
- Storage design
Multi-Tiered Applications

- **Presentation tier**
  Web server farms, application GUIs, SMTP gateways

- **Middle tier**
  Application load balancing, messaging, mail multiplexers

- **Storage tier**
  Databases, mail stores

Presentation Tier—Data Characteristics

- **WEB servers, GUIs**
  Data characteristics
  - Small HTML, SHTML and XML files
  - File system based
  - Unrelated data sets
  - Access patterns
    - Concurrent access
    - Random access
Storage for Applications

Presentation Tier—Storage Design

- **Web servers**
  - A few internal disks to contain data files
  - Spread data files across disks manually, if necessary
  - Normally no RAID features but could implement RAID 0 to improve performance

- **GUIs**
  - Normally run on local workstations/PCs

Storage for Applications

Presentation Tier

- **Storage characteristics**
  - Small files
  - File system based
  - Unrelated data

- **Access patterns**
  - Concurrent
  - Random

- **Storage design**
  - No RAID
  - Local disks
  - Manual spreading of data
Storage for Applications

Middle Tier

- Application servers
- Messaging servers
- Load balancers

Storage for Applications

Middle Tier—Data Characteristics

- OLTP manager
  - Transitional data
  - Small files residing on file systems
  - Unrelated data sets
  - Persistent request handling
- Messaging software
  - Transitional data
  - Unrelated data sets
  - Logs generated—sequential in access
Storage for Applications

Middle Tier—Data Access Patterns

• Messaging software
  Random access
  Logs—sequential in access
• OLTP managers
  Random access
  Logs—sequential in access

Middle Tier—Storage Design

Design for Random Access

• Spread data over multiple disks using
• RAID 0 + 1 technology
  RAID 0
  Stripe width to be multiples of FS block size—8KB
  Do not exceed the combined bandwidth of the HBA, for Eg.
  HBA bandwidth—100 mb/s
  Random access time per disk—15 mb/s
  RAID 0 Stripe—6 to 7 disks
Storage for Applications

Middle Tier—Storage Design (Cont.)

Design for Random Access

• Isolate mutually exclusive data sets
• Isolate logs and data files
• Analyse different software’s storage needs
• For Eg. MQ Series needs a separate/var/mqm file system, make sure this file system is resident on RAID 0 device
• Messaging systems may need a cached storage device, use a cached controller such as a Sun T300 or Compaq storage

Presentation Tier

• Storage characteristics
  Small files
  File system based
  Unrelated data
• Access patterns
  Concurrent
  Random
• Priorities
  1. Performance
  2. Manageability
  3. Reliability
• Storage design
  No RAID
  Local disks
  Manual spreading of data

Middle Tier

• Storage characteristics
  Transitional data
  Small files
  File system based
  Unrelated data sets
• Access patterns
  Random access
  Sequential if logging
• Priorities
  1. Performance
  2. Reliability
  3. Manageability
• Storage design
  Raid 1: Small stripe width
  Minimum number of disks
  Cached disk subsystems
  Isolate files with concurrent access
Storage for Applications

Storage Tier

• Databases servers
• File servers
• Mail stores

Storage Tier—Data Characteristics

Relational Databases

• Large files, sometimes a few hundreds of them
• Ordered and related sets of data
• Layered data sets, eg. data files and their corresponding indices
• Permanent and definitely relevant data
Storage for Applications

Storage Tier—Data Characteristics

File Servers
- e.g. mail servers, development servers, NFS filers, etc.
- Medium to large files
- Permanent data
- Usually unrelated data sets

Relational Databases
- Online Transaction Processing Databases (OLTP)
  - Random access
  - Sometimes real time updates
  - Small but constant updates
  - e.g. E-commerce applications
  - Banking and financial applications
### Storage for Applications

#### Storage Tier—Data Characteristics

#### Relational Databases
- **Decision Support Systems (DSS)**
  - Large sequential reads 80-90% of the time
  - Large batch updates
- **Flat file servers**
  - Access based on applications usage
  - Normally small to medium files

#### Storage Tier—Storage Design

#### Relational Databases
- **OLTP databases**
  - RAID 0 + 1 spread across multiple disks
  - Preferably on a cached controller storage subsystem
  - RAID 0 designs
    - Stripe width to be multiples of db_block_size
    - Eg. db_block_size = 4k
      - stripe width = 8k
Storage for Applications

Storage Tier—Storage Design

OLTP Databases (Cont.)

- Mutually exclusive data on mutually exclusive data volumes
  - Data and its index on separate volumes
  - Logs, temp tables on separate volumes
  - Data being accessed concurrently from the application needs to be separated as far as possible

Relational Databases

- Decision Support Systems (DSS)
  - RAID 0 + 1 technology spread across many disks
  - RAID 0 design
    - Stripe width usually large like 256K or 512K
    - Stripe width multiples of \( \text{db\_block\_size} \times \text{db\_file\_multiblock\_read\_count} \)
  - for Oracle, similar parameters are available for other RDBMs
Storage for Applications

Storage Tier—Storage Design

DSS (Cont.)

- Stripe across as many HBAs/disks as the bandwidth requirements
  - Eg. bandwidth req. = 1 GB/sec sequential scan
  - #HBAs in stripe = 1 GB/throughput per HBA
  - #Disk/HBA = throughput of HBA/throughput of disks
  - Dedicate separate spindles for the one or two huge tables
  - Isolate mutually exclusive data and its indices
  - Isolate logs, temp tables onto separate disks

Flat File Storages

- RAID 0 + 1 technology
  - Stripe width depends on the application using storage
  - For Eg. mail servers—small stripe width to accommodate small files of mails, typically 4K
  - Categorize data based on functionality
  - Eg. different engineering groups have different sets of disks
Storage for Applications

Presentation Tier
- Storage characteristics
  - Small files
  - File system based
  - Unrelated data
- Access patterns
  - Concurrent
  - Random
- Priorities
  - 1. Performance
  - 2. Manageability
  - 3. Reliability
- Storage design
  - No RAID
  - Local disks
  - Manual spreading of data

Middle Tier
- Storage characteristics
  - Transitional data
  - Small files
  - File system based
  - Unrelated data sets
- Access patterns
  - Random access
  - Sequential if logging
- Priorities
  - 1. Performance
  - 2. Reliability
  - 3. Manageability
- Storage design
  - Raid 1: Small stripe width
  - Minimum number of disks
  - Cached disk subsystems
  - Isolate files with concurrent access

Storage Tier
- Storage characteristics
  - Large data files
  - Related, ordered data assets
  - File system and raw files
  - Permanent data
- Access patterns
  - Mixed mode access
- Priorities
  - 1. Reliability
  - 2. Management
  - 3. Performance
- Storage design
  - RAID 1+0/0+1
  - Stripe width as per app
  - Overall disk throughput not to exceed HBA throughput
  - Cached disk subsystems
  - Isolate files with concurrent access
  - Follow app guidelines for storage layout

Storage for Applications

Enterprise Storage
- Primary Site
- Database/File Storage Servers
- Application Servers/Middleware Servers
- JBODs

Single Storage Management Resource
- Disaster Recovery Site
- ISP 1
- ISP 2
- Internet

Users

Database/Files

Web Servers
Complete Your Online Session Evaluation!

**WHAT:** Complete an online session evaluation and your name will be entered into a daily drawing

**WHY:** Win fabulous prizes! Give us your feedback!

**WHERE:** Go to the Internet stations located throughout the Convention Center

**HOW:** Winners will be posted on the onsite Networkers Website; four winners per day