Storage Area Networks SAN

Shane Healy
Objective/Agenda

• Provide a basic overview of what Storage Area Networks (SAN) are, what the constituent components are, and how these components fit together to deliver high availability data storage.

• Core Components

• High Availability
What is a SAN?

- **SAN**: Storage Area Network
  - A network whose primary purpose is the transfer of data between computer systems and storage elements and among storage elements.
  - The combination of a high-speed transport with Fibre/SCSI protocol results in an efficient means to deploy servers, disk arrays, and tape subsystems and frees all components from the constraints of parallel SCSI architecture.
  - SANs are extremely high-speed, back-end networks dedicated to storage.
  - SANs provide raw storage in the sense that the servers "see" the storage as physical disk drives.
  - Not a replacement for the SCSI protocol. Fibre Channel Protocol encapsulates SCSI with extra packets. Each SCSI packet is within a FCP frame.
SAN Components

• A SAN consists of:
  – a communication infrastructure, which provides physical connections
  – a management layer, which organizes these connections
  – storage elements
  – computer systems

• SAN is made up of:
  – A protocol – Fibre Channel Protocol
  – A switching method – FC switches
  – Storage medium with intelligent controllers
  – Servers for managing the data flow
  – A physical connection scheme – Fabrics
What does a SAN look like?

• A SAN typically looks physically much like this.
• Note that the servers are connected to the storage through fibre channel switches, not directly attached.
Non SAN Environment

Heterogeneous Servers

General Purpose Network
LAN/MAN/WAN

Array Controller with Disks

SCSI

Array Controllers with Disks

DLT Library
What else can a SAN look like?

- **Point-to-Point**
  - Only 2 Devices
  - (Direct Connect)

- **Arbitrated Loop**
  - Up to 126 Devices
  - (Fibre Channel Hubs)

- **Switched Fabric**
  - Up to 16 Million Devices
  - (Fibre Channel Switches)
Fibre Channel Topologies

• Point to Point
  – Direct connection between server and storage via fibre cable(s)
Fibre Channel Loop (arbitrated loop)

- Resembles a ring or FDDI connections
- Usually uses a hub for arbitration

Each Device takes turn to the loop by an arbitration schema
Fabric

- Switch provides addressing to specific LUN (Logical Unit Number or Disk drives)
What makes up a SAN?

- Servers
- Host Bus Adapters (HBA)
- Gigabit Interface Converters (GBIC)
- Cabling
- Switches
- Zoning
- Storage Controllers
- Disk Arrays
- Software
Servers

– They usually have some onboard disks for the operating system, but also have the ability to boot directly from the SAN.

– Typically at present in Intel Servers have a mirrored pair for the OS. On VMS they boot directly from the SAN.

– Multiple OS types are supported e.g. Windows, UNIX, Solaris, VMS etc.

– Blade technologies are being introduced in Intel where a server is essentially a motherboard.

– Currently Intel Primarily use HP servers, however multiple hardware platforms are supported e.g. IBM, DELL etc.
HBAs

- An HBA is a Host Bus Adapter
- It sits in the PCI (or PCIX) bus of a server and converts electrical signals into modulated laser signals.
- HBAs have replaceable GBICs (Gigabit Interface Converter) which contain the laser and receiver.
- The fiber cables plug into the GBIC, one to the transmit side, one to the receive side.
- Servers can (and typically do) have more than one HBA.
**Fibre Channel Cables**

- **Cabling**
  - At Intel, cabling a SAN requires the use of fiber optic cables.
  - We don't use any copper fibre channel due to a 30M Limitation.
  - Fiber optic cable comes in three types
    - 50 micron multi-mode
    - 62.5 micron multi-mode
    - 9 micron single mode
  - We use different cable types to accomplish different objectives, mostly to address distance requirements.

<table>
<thead>
<tr>
<th>Optical Core Size</th>
<th>50 micron</th>
<th>62.5 micron</th>
<th>9 micron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport Mode</td>
<td>Multi-mode</td>
<td>Multi-mode</td>
<td>single-mode</td>
</tr>
<tr>
<td>Wavelength</td>
<td>FC short wave</td>
<td>FC short wave</td>
<td>FC long wave</td>
</tr>
<tr>
<td>Distance Limit</td>
<td>500 meters</td>
<td>175 meters</td>
<td>10,000 meters</td>
</tr>
<tr>
<td>Color of Cover</td>
<td>orange</td>
<td>orange</td>
<td>yellow</td>
</tr>
</tbody>
</table>
• **Multimode fiber**
  - Is optical fiber that is designed to carry multiple light rays or modes concurrently, each at a slightly different reflection angle within the optical fiber core.
  - Multimode fiber transmission is used for relatively short distances because the modes tend to disperse over longer lengths (this is called *modal dispersion*).

• **Single mode fiber**
  - Is optical fiber that is designed for the transmission of a single ray or mode of light as a carrier and is used for long-distance signal transmission.
  - Single mode fiber has a much smaller core than multimode fiber.
**Fibre Channel Cables**

- **50 Micron**
  - Short Wave/Multi-mode

- **62.5 Micron**
  - Short Wave/Multi-mode

- **9 Micron**
  - Long Wave/Single-mode

**Orange Protective Cover**

Cladding is a reflective material that causes the light waves to be confined within the core material.
Switches

- Fibre Channel switches are very similar to network switches.
  - They enable attached SAN devices to communicate on a virtual circuit at maximum speed.
  - Unlike a hub or a fibre channel loop, switches do not share bandwidth between all connected devices.

- We currently use Brocade switches at Intel.
- Brocade Silkworm series 2400, 2800, 3800, 3900, and 12000
- Transport speeds of either 1 GB/sec or 2 GB/sec
- Future implementations will support 10 GB/sec.
- Each switch can either stand alone or become part of a "fabric," or a group of switches forming a single virtual switch.
Switched Fabrics

- Switches can be joined together to form a "fabric," or a group of switches forming a single virtual switch.
- Fabrics can be configured to support many security features such as zoning and fabric logins.
- Switches are combined by simply plugging fiber cables into them, connecting the switches together. They then autosense the fabric topology and assume their role in it.
Switching (Physical)

- **Physical Topology**
  - The physical topology can be described as actual hardware components in a fabric and the Fibre Channel cabling that interconnects them.

- **Design Considerations and Topology of a fabric are:**
  - Number of switches in the fabric
  - Number of hops between any two switches
  - Number of ports per switch
  - Number of ISLs (Inter Switch Links) between switches
  - Physical distance between any two switches

- The number of tiers in the fabric is based on the number of switches that are traversed between the farthest two points in the fabric.

- As the number of Tiers increases, the distance that a fabric management message must travel to reach every switch in the fabric increases. It affects the time it takes to propagate and complete a fabric reconfiguration event (e.g., adding a new switch), or zone set propagation event.
Switching (Logical)

- **Logical Topology**
- In contrast to a fabric's physical topology, a logical topology is concerned with where the Fibre Channel elements are attached around the fabric and the relationships (zoning) that define how these elements will be used together.
- When describing a fabric's logical topology, it is in terms of its number of logical tiers.
- The number of logical tiers in the topology is based on the number of switches traversed between a server and the storage zoned to it.
- Since servers and storage can be located anywhere in the fabric, there may be several different logical tier relationships in the same fabric.
- Below demonstrates an example of multiple logical tier relationships in the same fabric.
Zoning

- **Zoning** is a method of providing access control in a Fibre Channel *fabric* consisting of a single Fibre Channel switch or cascaded switches.
- Within the fabric, the SAN administrator can group server HBA ports (initiators) with storage ports (disk or tape drives) into zones.
- The individual switch ports or device World Wide Name(s) are referred to as zone *members*.
- A zone effectively partitions a Fibre Channel fabric into a subset of logical entities, within which each participant is able to connect and communicate only with other members of the same zone.
The storage array is really the core component of today’s storage area network.
- It contains many, many disks
- Grouped into RAID sets
- Assigned to hosts on the SAN
- They are typically the size of a refrigerator, but can be smaller or larger, depending on capacity. Storage arrays range from very simple JBOD (Just a Bunch Of Disks) to sophisticated RAID controlled smart storage that is aware of and communicates with the connected hosts.
RAID

• Short for Redundant Array of Independent (or Inexpensive) Disks, a category of disk drives that employ two or more drives in combination for fault tolerance and performance. RAID disks are then divided into LUN’s (Logical Unit numbers) and presented to servers as partitions or raw space. There are number of different RAID levels:

  - **Disk**
    - No protection
    - JBOD
  - **RAID-1: Mirroring**
    - Some performance gain by splitting read operations
    - Protection against single disk failure
    - Minimum performance hit during failure
  - **RAID 1/0: Striped Mirrors**
    - Performance of stripes combined with split read operations
    - Protection against single disk failure
    - Minimum performance hit in failure mode
  - **RAID-3: Striped Elements**
    - Each data element striped across disks - parity kept on the last disk in the RAID group
    - Extremely fast read access from the disk
    - Used for streaming media
    - Parity protection against single disk failure
    - Performance penalty during failure

  - **RAID-5: Striping with Parity**
    - Performance of striping
    - Protection from single disk failure
    - Parity distributed across member drives within the RAID Group
    - Write performance penalty
    - Performance impact if a disk fails in RAID Group

  - **Hot Spare**
    - Takes the place of failed disk within a RAID group
    - Must have equal or greater capacity than the disk it replaces
    - When failing disk is replaced, the hot spare restores the data to the replacement disk and returns to the hot spare pool
**RAID LEVEL 1: Mirroring & Duplexing**

- A Mirroring
- B Mirroring
- C Mirroring
- D Mirroring

**RAID LEVEL 3: Parallel Transfer with Parity**

- A0 Stripe 0
- B0 Stripe 1
- C0 Stripe 2
- D0 Stripe 3

- A1 Parity Generation
- A2
- A3

- A1
- B1
- C1
- D1

- A2
- B2
- C2
- D2

- A3
- B3
- C3
- D3

- A Parity
- B Parity
- C Parity
- D Parity

**RAID LEVEL 5: Independent Data Disks with Distributed Parity Blocks**

- A0
- A1
- A2
- A3

- B0
- B1
- B2
- B4

- C0
- C1
- C2
- C4

- D0
- D1
- D2
- D3

- E0
- E1
- E2
- E3

**RAID LEVEL 0+1: High Data Transfer Performance**

- A Striping
- B Striping
- C Striping
- D Striping

- E Mirroring
- F Mirroring
- G Mirroring
- H Mirroring

- A
- B
- C
- D
Software

- This is the glue that holds all the pieces together and makes the system work. There is software installed on:
  - The server
    - It is made up of drivers and software that acts as a link in the operating system to create the illusion that the HBA is just another SCSI card connected to a dedicated disk.
  - The switches
    - The switches all have software that enables them to route requests between hosts and storage.
  - The storage frame
    - The storage frame can have sophisticated software that is used to create RAID groups, Luns (virtual disks), and assign storage to hosts.
Server Software

- **HBA drivers**
  - used by the operating system to communicate with the HBA installed in the PCI slot. It makes the HBA look to the system like a SCSI card and enables the out-of-the-box operating system software to issue SCSI commands over the fibre channel network.

- **Agent software**
  - an optional component that not all SAN vendors require. It is usually a small piece of server code installed on the server to communicate through the fibre channel network or through the Ethernet network with the storage frame. It enables the storage frame to know about the server so that it can assign storage and manage resources belonging to the server.

- **Multipath software**
  - such as Secure Path (HP) or PowerPath (EMC) is used to coordinate communication between two separate HBAs installed in the server. This allows a redundant path between the server and the storage and increases availability.
Switch Software

• Software running on the switches is usually known as firmware.
  – Responsible for coordinating all the traffic going on between servers and storage, between the switches themselves, and managing security.
  – Our Brocade switches have both Telnet and Web Server components and can be managed remotely through a TCP/IP network connection and a web browser or Telnet window.
  – Switches also have an available API (application programming interface) which enables the switches to be controlled programmatically and monitored.
Storage Software

- Storage frames can be as simple as a connected JBOD unit with nothing but a group of disks...
- ...or as complicated as a million dollar storage array with sophisticated SCSI or Fibre Channel controllers and disks
- Most of what is deployed at intel has a high level of intelligence and sophistication.
- For example, Clariion storage arrays ship with a built in web server component that can be used to manage the entire frame or even an entire enterprise of storage.
- Behind the web server is a fully redundant set of fibre channel based RAID controllers that enable disks to be combined into RAID sets, carved into virtual disks, and assigned to hosts (known to the array through agent software).
- On HP storage typically a SAN Appliance is used to run the management software and provides the same tasks.
Management Software

• The last piece of the puzzle.
  – Means different things to different people, but it encompasses the entire range of software to manage a SAN
  – simple monitoring system health
  – provisioning (creating and assigning storage)
  – tracking and trending storage consumption
  – Today there are few, if any, all-encompassing tools for managing the entire Intel enterprise SAN, but these tools are beginning to emerge.
High Availability

• Robust SAN designs include at least two data paths between any server and disk storage subsystem, such as RAID/Controllers etc.
• By using redundant data paths in conjunction with FC director switches, a true redundant environment can be achieved.
• Additionally, the any-to-any connectivity of SAN devices gives the flexibility and portability for data requirements in dynamic environments.
• It can provide redundancy in SAN’s and still achieve data availability even in the event of multiple component failures.
• Storage frames and servers typically include dual redundant devices, e.g. Power supplies, HBA’s, Cables, Controllers etc.
• Implementing High Availability supports non-disruptive upgrade (NDU) of storage system software. While a server is running normally.
• If a failure occurs e.g. a controller failover, software such as Powerpath or SecurePath attempts to fail over (transfer) the I/Os to a different controller or Switch port.
• Failover Software (Powerpath/SecurePath) does not protect against single HBA failures. To protect against such failures dual HBA’s are required.
High Availability Setup

Fabric A

Fabric B

ISL

ISL
High Availability HBA Failure
High Availability Controller Failure

Embedded software performs internal failover on storage backplane

Fabric A

Fabric B

ISL’s
High Availability Switch Failure
So What else can a SAN do?

- **Clustering!**  
  - MS Cluster server requires a shared disk. SANs enable the sharing of disks between multiple servers.

- **Snapshots**  
  - Take an instant snapshot of your data for use as a point-in-time copy.

- **Clones**  
  - Create a bit-for-bit copy of your data that can be assigned to other servers.

- **Faster backups**  
  - Connecting a large tape drive array enables backups over fibre channel instead of Ethernet  
  - You can use snapshots and clones to get even more sophisticated.

- **Data mining**  
  - By creating a copy of your data, you can mine it for information without affecting the performance of the live data.

- **Bunkering**  
  - Create a synchronous or asynchronous copy of all your server’s data in a remote location in case of disaster.

- **Fast failover**  
  - Having servers separated from their storage enables the LUNs (disk drives) to be reassigned to a new server very quickly.
SAN vs. Network Attached Storage (NAS)

- **SAN** - block level I/O
- **NAS** (Network Attached Storage) - file level I/O
- NAS could use SAN for storage
- SAN accommodates file servers, web servers, databases, application servers, etc.
- SAN uses SCSI over Fibre Channel today, separating storage traffic from network traffic.
- NAS operates like a shared file server, storage is accessed across Ethernet, usually sharing bandwidth with applications.
- SAN integrates backup within its high speed storage network.
- NAS backs up data across Ethernet although they can use SCSI or Fibre.
- Operating System dependant
File I/O Server vs. Block I/O Server Architecture

Clients or Servers

IP Network

File I/O Processing

Server (Custom O/S)

Disk Arrays

NAS Server

File I/O Processing

Server (P100 O/S)

SAN Storage

SAN Server

Block I/O (Fibre Channel)
Network Attached Storage

File transfer via the server operating system
Questions?