VMware Infrastructure 3

SAN Conceptual and Design Basics

VMware ESX Server can be used in conjunction with a SAN (storage area network), a specialized high-speed network that connects computer systems to high performance storage subsystems. Using ESX Server together with a SAN provides extra storage for consolidation, improves reliability, and helps with disaster recovery.

To use ESX Server effectively with a SAN, you're expected to be familiar with the SAN technology. This white paper offers a brief introduction to some basic SAN concepts, but doesn't aim to be an exhaustive source of information on SANs. If you are an ESX Server administrator planning to set up ESX Server hosts to work with SANs, you should also consult other resources available in print and on the Internet to achieve a working knowledge of SAN concepts. Additional information on how ESX Server interacts with SAN may be found in *SAN Configuration Guide*.

The white paper discusses these topics:

SAN Basics

A SAN is a specialized high-speed network of storage devices and switches connected to computer systems. This white paper refers to the computer systems as servers or hosts.

A SAN presents shared pools of storage devices to multiple servers. Each server can access the storage as if it were directly attached to that server. A SAN supports centralized storage management. SANs make it possible to move data between various storage devices, share data between multiple servers, and back up and restore data rapidly and efficiently. In addition, a properly configured SAN facilitates both disaster recovery and high availability.

The physical components of a SAN can be grouped in a single rack or data center or connected over long distances. This makes a SAN a feasible solution for businesses of any size: the SAN can grow easily with the business it supports.

SAN Component Overview

This section gives an overview of SAN components. The numbers in the text correspond to numbers in Figure 1, "SAN Components," on page 3.

In its simplest form, a SAN consists of one or more servers (**1**) attached to a storage array (**2**) using one or more SAN switches. Each server might host numerous applications that require dedicated storage for applications processing.

The following components, discussed in more detail in "SAN Components" on page 4 are involved:

- **SAN Switches (3)** — SAN switches connect various elements of the SAN. In particular, they might connect hosts to storage arrays. SAN switches also allow administrators to set up path redundancy in the event of a path failure from host server to switch or from storage array to switch.
- **Fabric (4)** — The SAN fabric is the actual network portion of the SAN. When one or more SAN switches are connected, a fabric is created. The FC protocol is used to communicate over the entire network. A SAN can consist of multiple interconnected fabrics. Even a simple SAN often consists of two fabrics for redundancy.

- **Connections: Host Bus Adapters (5) and Storage Processors (6)** — Host servers and storage systems are connected to the SAN fabric through ports in the fabric.
  - A host connects to a fabric port through an HBA.
  - Storage devices connect to fabric ports through their storage processors.

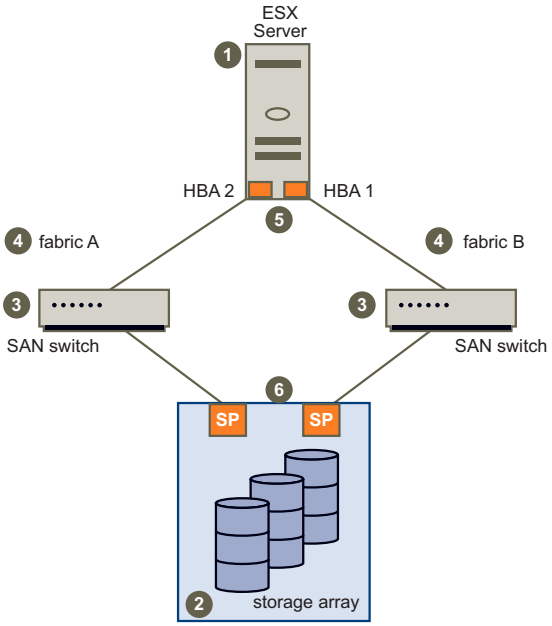

Figure 1. SAN Components

How a SAN Works

The SAN components interact as follows:

1 When a host wants to access a storage device on the SAN, it sends out a block-based access request for the storage device.

2 SCSI commands are encapsulated into FC packets. The request is accepted by the HBA for that host and is converted from its binary data form to the optical form required for transmission on the fiber optic cable.

3 At the same time, the request is packaged according to the rules of the FC protocol.

4 The HBA transmits the request to the SAN.

5 Depending on which port is used by the HBA to connect to the fabric, one of the SAN switches receives the request and sends it to the storage processor, which sends it on to the storage device.

The remaining sections of this white paper provide additional information about the components of the SAN and how they interoperate. These sections also present general information on configuration options and design considerations.

SAN Components

The components of an FC SAN can be grouped as follows and are discussed below:

Figure 2 shows the SAN component layers.

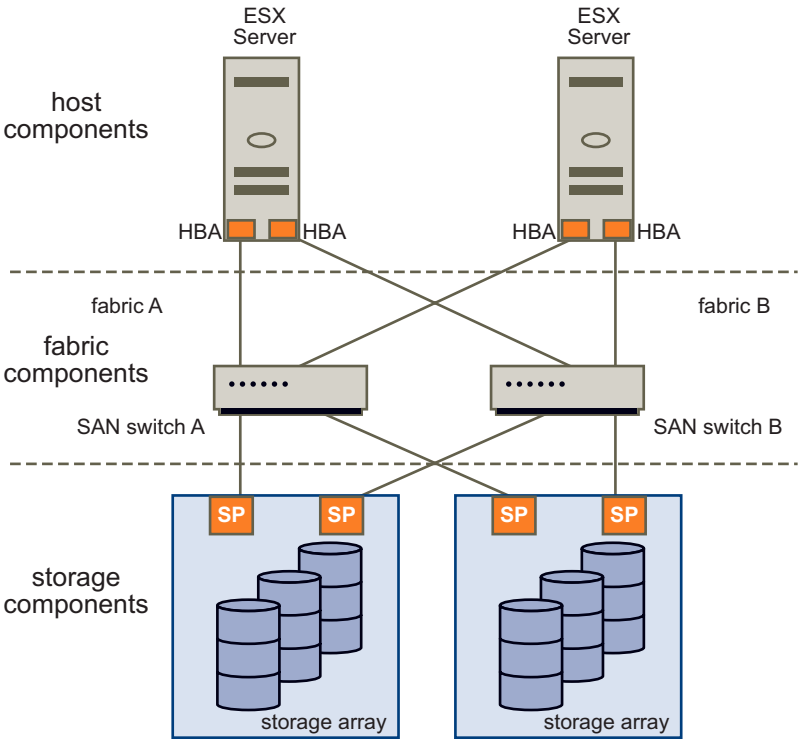

Figure 2. SAN Component Layers

Host Components

The host components of a SAN consist of the servers themselves and the components that enable the servers to be physically connected to the SAN.

- **HBAs** are located in the servers, along with a component that performs digital-to-optical signal conversion. Each host connects to the fabric ports through its HBAs.
- **HBA drivers** running on the servers enable the servers' operating systems to communicate with the HBA.

Fabric Components

All hosts connect to the storage devices on the SAN through the SAN fabric. The network portion of the SAN consists of the following fabric components:

- **SAN Switches** — SAN switches can connect to servers, storage devices, and other switches, and thus provide the connection points for the SAN fabric. The type of SAN switch, its design features, and its port capacity all contribute to its overall capacity, performance, and fault tolerance. The number of switches, types of switches, and manner in which the switches are interconnected define the fabric topology.
  - For smaller SANs, the standard SAN switches (called modular switches) can typically support 16 or 24 ports (though some 32-port modular switches are becoming available). Sometimes modular switches are interconnected to create a fault-tolerant fabric.
  - For larger SAN fabrics, director-class switches provide a larger port capacity (64 to 128 ports per switch) and built-in fault tolerance.
- **Data Routers** — Data routers are intelligent bridges between SCSI devices and FC devices in the SAN. Servers in the SAN can access SCSI disk or tape devices in the SAN through the data routers in the fabric layer.
- **Cables** — SAN cables are usually special fiber optic cables that are used to connect all of the fabric components. The type of SAN cable and the fiber optic signal determine the maximum distances between SAN components and contribute to the total bandwidth rating of the SAN.
- **Communications Protocol** — Fabric components communicate using the FC communications protocol. FC is the storage interface protocol used for most of today's SANs. FC was developed as a protocol for transferring data between two ports on a serial I/O bus cable at high speeds. FC supports point-to-point, arbitrated loop, and switched fabric topologies. Switched fabric topology is the basis for most current SANs.

Storage Components

The storage components of a SAN are the storage arrays. Storage arrays include storage processors (SPs). The SPs are the front end of the storage array. SPs communicate with the disk array (which includes all the disks in the storage array) and provide the RAID/LUN functionality.

Storage Processors

SPs provide front-side host attachments to the storage devices from the servers, either directly or through a switch. The server HBAs must conform to the protocol supported by the storage processor. In most cases, this is the FC protocol.

Storage processors provide internal access to the drives, which can be using a switch or bus architecture. In high-end storage systems, drives are normally connected in loops. This back-end loop technology employed by the SP provides several benefits:

- High-speed access to the drives
- Ability to add more drives to the loop
- Redundant access to a single drive from multiple loops (when drives are dual-ported and attached to two loops)

Storage Devices

Data is stored on disk arrays or tape devices (or both).

Disk arrays are groups of multiple disk devices and are the typical SAN disk storage device. They can vary greatly in design, capacity, performance, and other features.

Storage arrays rarely provide hosts direct access to individual drives. Instead, the storage array uses RAID (Redundant Array of Independent Drives) technology to group a set of drives. RAID uses independent drives to provide capacity, performance, and redundancy. Using specialized algorithms, several drives are grouped to provide common pooled storage. These RAID algorithms, commonly known as RAID levels, define the characteristics of the particular grouping.

In simple systems that provide RAID capability, a RAID group is equivalent to a single LUN. A LUN is a single unit of storage. Depending on the host system environment, a LUN is also known as a volume or a logical drive. From a VI Client, a LUN looks like any other storage unit available for access.

In advanced storage arrays, RAID groups can have one or more LUNs created for access by one or more servers. The ability to create more than one LUN from a single RAID group provides fine granularity to the storage creation process. You are not limited to the total capacity of the entire RAID group for a single LUN.

Note A SAN administrator must be familiar with the different RAID levels and understand how to manage them. Discussion of those topics is beyond the scope of this document.

Most storage arrays provide additional data protection and replication features such as snapshots, internal copies, and remote mirroring.

- A snapshot is a point-in-time copy of a LUN. Snapshots are used as backup sources for the overall backup procedures defined for the storage array.
- Internal copies allow data movement from one LUN to another for an additional copy for testing.
- Remote mirroring provides constant synchronization between LUNs on one storage array and a second, independent (usually remote) storage array for disaster recovery.

Tape Storage Devices

Tape storage devices are part of the SAN backup capabilities and processes.

- Smaller SANs might use high-capacity tape drives. These tape drives vary in their transfer rates and storage capacities. A high-capacity tape drive might exist as a standalone drive, or it might be part of a tape library.
- Typically, a large SAN, or a SAN with critical backup requirements, is configured with one or more tape libraries. A tape library consolidates one or more tape drives into a single enclosure. Tapes can be inserted and removed from the tape drives in the library automatically with a robotic arm. Many tape libraries offer large storage capacities—sometimes into the petabyte (PB) range.

Understanding SAN Interactions

The previous section's primary focus was the components of a SAN. This section discusses how SAN components interact:

- "SAN Ports and Port Naming" on page 7
- "Multipathing and Path Failover" on page 7
- "Active/Active and Active/Passive Disk Arrays" on page 8
- "Zoning" on page 9

SAN Ports and Port Naming

- In the context of this document, a port is the connection from a device into the SAN. Each node in the SAN— each host, storage device, and fabric component (router or switch)—has one or more ports that connect it to the SAN. Ports can be identified in a number of ways:
- **WWPN** — World Wide Port Name. A globally unique identifier for a port which allows certain applications to access the port. The FC switches discover the WWPN of a device or host and assign a port address to the device.
- **Port_ID** (or port address) — Within the SAN, each port has a unique port ID that serves as the FC address for the port. This enables routing of data through the SAN to that port. The FC switches assign the port ID when the device logs into the fabric. The port ID is valid only while the device is logged on.

In-depth information on SAN ports can be found at **http://www.snia.org**, the Web site of the Storage Networking Industry Association.

Multipathing and Path Failover

An FC path describes a route:

- From a specific HBA port in the host,
- Through the switches in the fabric, and
- Into a specific storage port on the storage array.

A given host might be able to access a LUN on a storage array through more than one path. Having more than one path from a host to a LUN is called **multipathing**.

By default, VMware ESX Server systems use only one path from the host to a given LUN at any time. If the path actively being used by the VMware ESX Server system fails, the server selects another of the available paths. The process of detecting a failed path and switching to another is called path failover. A path fails if any of the components along the path—HBA, cable, switch port, or storage processor— fails.

Active/Active and Active/Passive Disk Arrays

It is useful to distinguish between active/active and active/passive disk arrays:

- An **active/active disk array** allows access to the LUNs simultaneously through all the storage processors that are available without significant performance degradation. All the paths are active at all times (unless a path fails).
- In an **active/passive disk array**, one SP is actively servicing a given LUN. The other SP acts as backup for the LUN and may be actively servicing other LUN I/O. I/O can be sent only to an active processor. If the primary storage processor fails, one of the secondary storage processors becomes active, either automatically or through administrator intervention.

Note	Using active/passive arrays with path policy Fixed can potentially lead to path thrashing. See ESX Server SAN Configuration Guide for more information on resolving path thrashing.

In Figure 3, one storage processor is active, the other is passive. Data arrives through the active array only.

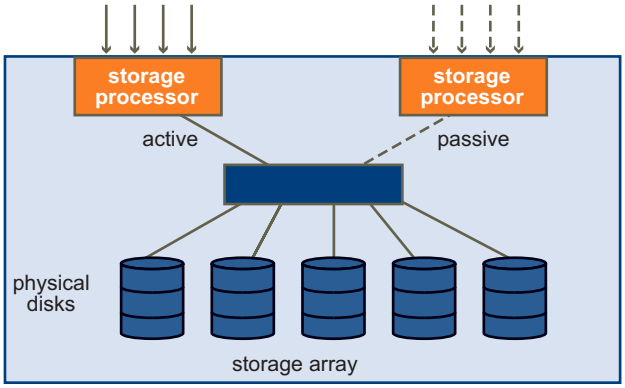

Figure 3. Active/passive Storage Array

Zoning

Zoning provides access control in the SAN topology; it defines which HBAs can connect to which SPs. You can have multiple ports to the same SP in different zones to reduce the number of presented paths.

When a SAN is configured using zoning, the devices outside a zone are not visible to the devices inside the zone. In addition, SAN traffic within each zone is isolated from the other zones.

Within a complex SAN environment, SAN switches provide zoning. Zoning defines and configures the necessary security and access rights for the entire SAN.

Typically, zones are created for each group of servers that access a shared group of storage devices and LUNs. You can use zoning in several ways. Here are some examples:

- **Zoning for security and isolation** — You can manage zones defined for testing independently within the SAN so they don't interfere with the activity going on in the production zones. Similarly, you could set up different zones for different departments.
- **Zoning for shared services** — Another use of zones is to allow common server access for backups. SAN designs often have a backup server with tape services that require SAN-wide access to host servers individually for backup and recovery processes. These backup servers need to be able to access the servers they back up. A SAN zone might be defined for the backup server to access a particular host to perform a backup or recovery process. The zone is then redefined for access to another host when the backup server is ready to perform backup or recovery processes on that host.
- **Multiple storage arrays** — Zones are also useful when there are multiple storage arrays. Through the use of separate zones, each storage array is managed separately from the others, with no concern for access conflicts between servers.

LUN Masking

LUN masking is commonly used for permission management. LUN masking is also referred to as selective storage presentation, access control, and partitioning, depending on the vendor.

LUN masking is performed at the SP or server level; it makes a LUN invisible when a target is scanned. The administrator configures the disk array so each server or group

of servers can see only certain LUNs. Masking capabilities for each disk array are vendor specific, as are the tools for managing LUN masking.

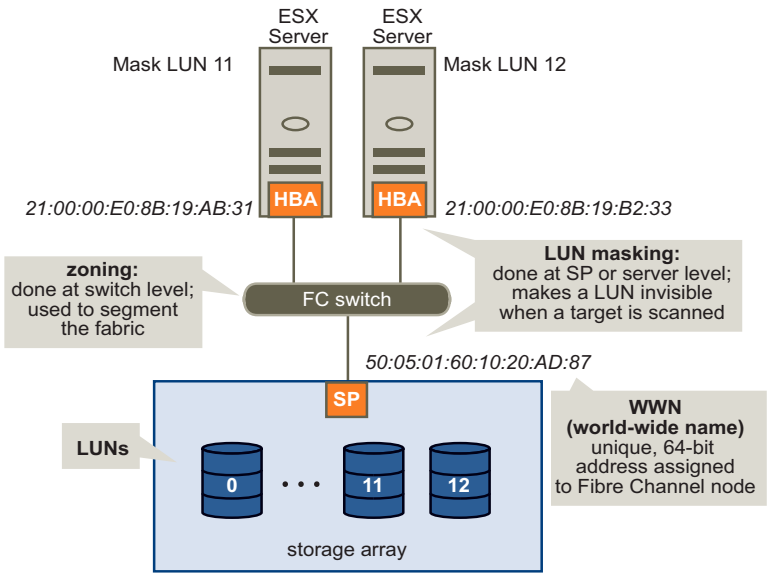

Figure 4. LUN Zoning and Masking

SAN Installation Considerations

Installing a SAN requires careful attention to details and an overall plan that addresses all the hardware, software, storage, and applications issues and their interactions as all the pieces are integrated.

Requirements

To integrate all components of the SAN, you must meet the vendor's hardware and software compatibility requirements, including the following:

- HBA (firmware version, driver version, and patch list)
- Switch (firmware)
- Storage (firmware, host personality firmware, and patch list)

SAN Setup

When you're ready to set up the SAN, complete these tasks.

To prepare the SAN

1 Assemble and cable together all hardware components and install the corresponding software.

   a Check the versions.

   b Set up the HBA.

   c Set up the storage array.

2 Change any configuration settings that might be required.

3 Test the integration.

   During integration testing, test all the operational processes for the SAN environment. These include normal production processing, failure mode testing, backup functions, and so forth.

4 Establish a baseline of performance for each component and for the entire SAN.

   Each baseline provides a measurement metric for future changes and tuning. See *ESX Server SAN Configuration Guide* for additional information.

5 Document the SAN installation and all operational procedures.

SAN Design Basics

When designing a SAN for multiple applications and servers, you must balance the performance, reliability, and capacity attributes of the SAN. Each application demands resources and access to storage provided by the SAN. The SAN's switches and storage arrays must provide timely and reliable access for all competing applications.

This section discusses some SAN design basics. It does not focus on SAN design for ESX Server hosts, but on the following general topics of interest:

- "Defining Application Needs" on page 11
- "Configuring the Storage Array" on page 12
- "Considering High Availability" on page 13
- "Planning for Disaster Recovery" on page 13

Defining Application Needs

The SAN must support fast response times consistently for each application even though the requirements made by applications vary over peak periods for both I/O per second and bandwidth (in megabytes per second).

A properly designed SAN must provide sufficient resources to process all I/O requests from all applications. Designing an optimal SAN environment is therefore neither

simple nor quick. The first step in designing an optimal SAN is to define the storage requirements for each application in terms of:

- I/O performance (I/O per second)
- Bandwidth (megabytes per second)
- Capacity (number of LUNs and capacity of each LUN)
- Redundancy level (RAID-level)
- Response times (average time per I/O)
- Overall processing priority

Configuring the Storage Array

Storage array design involves mapping the defined storage requirements to the resources of the storage array using these guidelines:

- Each RAID group provides a specific level of I/O performance, capacity, and redundancy. LUNs are assigned to RAID groups based on these requirements.
- If a particular RAID group cannot provide the required I/O performance, capacity, and response times, you must define an additional RAID group for the next set of LUNs. You must provide sufficient RAID-group resources for each set of LUNs.
- The storage arrays need to distribute the RAID groups across all internal channels and access paths. This results in load balancing of all I/O requests to meet performance requirements of I/O operations per second and response times.

Peak Period Activity

Base the SAN design on peak-period activity and consider the nature of the I/O within each peak period. You may find that additional storage array resource capacity is required to accommodate instantaneous peaks.

For example, a peak period may occur during noontime processing, characterized by several peaking I/O sessions requiring twice or even four times the average for the entire peak period. Without additional resources, I/O demands that exceed the capacity of a storage array result in delayed response times.

Caching

Though ESX Server systems benefit from write cache, the cache could be saturated with sufficiently intense I/O. Saturation reduces the cache's effectiveness.

Because the cache is often allocated from a global pool, it should be allocated only if it will be effective.

- A read-ahead cache may be effective for sequential I/O, such as during certain types of backup activities, and for template repositories.
- A read cache is often ineffective when applied to a VMFS-based LUN because multiple virtual machines are accessed concurrently. Because data access is random, the read cache hit rate is often too low to justify allocating a read cache.

- A read cache is often unnecessary when the application and operating system cache data are within the virtual machine's memory. In that case, the read cache caches data objects that the application or operating system already cache.

Considering High Availability

Production systems must not have a single point of failure. Make sure that redundancy is built into the design at all levels. Build in additional switches, HBAs, and storage processors, creating, in effect, a redundant access path.

- **Redundant SAN Components** — Redundant SAN hardware components including HBAs, SAN switches, and storage array access ports, are required. In some cases, multiple storage arrays are part of a fault-tolerant SAN design.
- **Redundant I/O Paths** — I/O paths from the server to the storage array must be redundant and dynamically switchable in the event of a port, device, cable, or path failure.
- **I/O Configuration** — The key to providing fault tolerance is within the configuration of each server's I/O system.

  With multiple HBAs, the I/O system can issue I/O across all of the HBAs to the assigned LUNs. Failures can have the following results:

  - If an HBA, cable, or SAN switch port fails, the path is no longer available and an alternate path is required.
  - If a failure occurs in the primary path between the SAN switch and the storage array, then an alternate path at that level is required.
  - If a SAN switch fails, the entire path from server to storage array is disabled, so a second fabric with a complete alternate path is required.
- **Mirroring** — Protection against LUN failure allows applications to survive storage access faults. Mirroring can accomplish that protection.

  Mirroring designates a second non-addressable LUN that captures all write operations to the primary LUN. Mirroring provides fault tolerance at the LUN level. LUN mirroring can be implemented at the server, SAN switch, or storage array level.

  Note Usually RAIDS don't mirror.

- **Duplication of SAN Environment** — For extremely high availability requirements, SAN environments may be duplicated to provide disaster recovery on a per-site basis. The SAN environment must be duplicated at different physical locations. The two resultant SAN environments may share operational workloads or the second SAN environment may be a failover-only site.

Planning for Disaster Recovery

If a site fails for any reason, you may need to immediately recover the failed applications and data from a remote site. The SAN must provide access to the data from

an alternate server to start the data recovery process. The SAN may handle the site data synchronization.

ESX Server makes disaster recovery easier because you do not have to reinstall an operating system on a different physical machine. Just restore the virtual machine image and continue what you were doing.

VMware, Inc. 3145 Porter Drive Palo Alto, CA 94304 www.vmware.com