SAN System Design and Deployment Guide

Second Edition

Latest Revision: August 2008

Table of Contents

Preface

This guide, or “cookbook,” describes how to design and deploy virtual infrastructure systems using VMware® Infrastructure 3 with SANs (storage area networks). It describes SAN options supported with VMware Infrastructure 3 and also describes benefits, implications, and disadvantages of various design choices. The guide answers questions related to SAN management, such as how to:

- Manage multiple hosts and clients
- Set up multipathing and failover
- Create cluster-aware virtual infrastructure
- Carry out server and storage consolidation and distribution
- Manage data growth using centralized data pools and virtual volume provisioning

This guide describes various SAN storage system design options and includes the benefits, drawbacks, and ramifications of various solutions. It also provides step-by-step instructions on how to approach the design, implementation, testing, and deployment of SAN storage solutions with VMware Infrastructure, how to monitor and optimize performance, and how to maintain and troubleshoot SAN storage systems in a VMware Infrastructure environment. In addition, Appendix A provides a checklist of SAN system design and implementation. For specific, step-by-step instructions on how to use VMware ESX commands and perform related storage configuration, monitoring, and maintenance operations, please see the VMware *ESX Basic System Administration Guide,* which is available online at www.vmware.com.

The guide is intended primarily for VMware Infrastructure system designers and storage system architects who have at least intermediate-level expertise and experience with VMware products, virtual infrastructure architecture, data storage, and datacenter operations.

Conventions and Abbreviations

This manual uses the style conventions listed in the following table:

Style	Purpose
`Monospace`	Used for commands, filenames, directories, and paths
`Monospace bold`	Used to indicate user input
Bold	Used for these terms: Interface objects, keys, buttons; Items of highlighted interest; glossary terms
Italic	Used for book titles
<name>	Angle brackets and italics indicate variable and parameter names

The graphics in this manual use the following abbreviations:

Abbreviation	Description
VC	VirtualCenter
Database	VirtualCenter database
Host #	VirtualCenter managed hosts
VM #	Virtual machines on a managed host
User #	User with access permissions
Disk #	Storage disk for the managed host
datastore	Storage for the managed host
SAN	Storage area network type datastore shared between managed hosts

Additional Resources and Support

The following technical resources and support are available.

SAN Reference Information

You can find information about SANs in various print magazines and on the Internet. Two Web-based resources are recognized in the SAN industry for their wealth of information. These sites are:

- http://www.searchstorage.com
- http://www.snia.org

Because the industry changes constantly and quickly, you are encouraged to stay abreast of the latest developments by checking these resources frequently.

VMware Technology Network

Use the VMware Technology Network to access related VMware documentation, white papers, and technical information:

- Product Information – http://www.vmware.com/products/
- Technology Information – http://www.vmware.com/vcommunity/technology
- Documentation – http://www.vmware.com/support/pubs
- Knowledge Base – http://www.vmware.com/support/kb
- Discussion Forums – http://www.vmware.com/community
- User Groups – http://www.vmware.com/vcommunity/usergroups.html

Go to http://www.vmtn.net for more information about the VMware Technology Network.

VMware Support and Education Resources

Use online support to submit technical support requests, view your product and contract information, and register your products. Go to:

http://www.vmware.com/support

Customers with appropriate support contracts can use telephone support for the fastest response on priority 1 issues. Go to:

http://www.vmware.com/support/phone_support.html

Support Offerings

Find out how VMware's support offerings can help you meet your business needs. Go to:

http://www.vmware.com/support/services

VMware Education Services

VMware courses offer extensive hands-on labs, case study examples, and course materials designed to be used as on-the-job reference tools. For more information about VMware Education Services, go to:

http://mylearn1.vmware.com/mgrreg/index.cfm

1 Introduction to VMware and SAN Storage Solutions

VMware® Infrastructure allows enterprises and small businesses alike to transform, manage, and optimize their IT systems infrastructure through virtualization. VMware Infrastructure delivers comprehensive virtualization, management, resource optimization, application availability, and operational automation capabilities in an integrated offering.

This chapter provides an overview of virtualization infrastructure operation and the VMware infrastructure architecture. It also summarizes the VMware Infrastructure components and their operation.

Topics included in this chapter are the following:

VMware Virtualization Overview

Virtualization is an abstraction layer that decouples the physical hardware from the operating system of computers to deliver greater IT resource utilization and flexibility. Virtualization allows multiple virtual machines, with heterogeneous operating systems (for example, Windows 2003 Server and Linux) and applications to run in isolation, side-by-side on the same physical machine.

Figure 1-1 provides a logical view of the various components comprising a VMware Infrastructure 3 system.

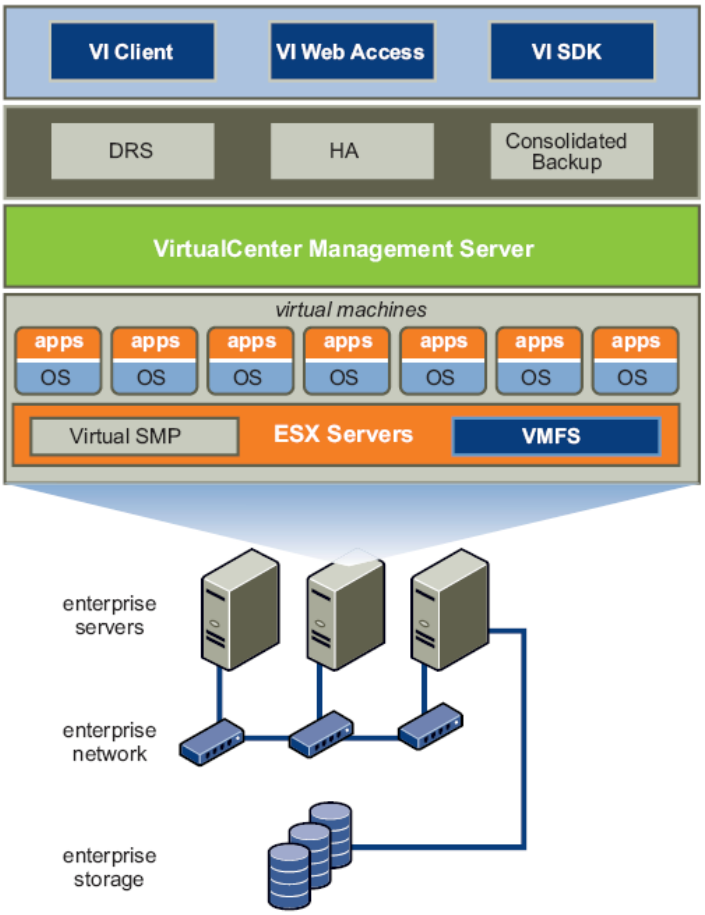


Figure 1-1. VMware Infrastructure

VMware Infrastructure includes the following components as shown in Figure 1-1:

- **VMware ESX**— Production-proven virtualization layer run on physical servers that allows processor, memory, storage, and networking resources to be provisioned to multiple virtual machines.
- **VMware Virtual Machine File System (VMFS)** — High-performance cluster file system for virtual machines.
- **VMware Virtual Symmetric Multi-Processing (SMP)** — Capability that enables a single virtual machine to use multiple physical processors simultaneously.
- **VirtualCenter Management Server** — Central point for configuring, provisioning, and managing virtualized IT infrastructure.
- **VMware Virtual Machine** — Representation of a physical machine by software. A virtual machine has its own set of virtual hardware (for example, RAM, CPU, network adapter, and hard disk storage) upon which an operating system and applications are loaded. The operating system sees a consistent, normalized set of hardware regardless of the actual physical hardware components. VMware virtual machines contain advanced hardware features, such as 64-bit computing and virtual symmetric multiprocessing.

- **Virtual Infrastructure Client (VI Client)** — Interface that allows administrators and users to connect remotely to the VirtualCenter Management Server or individual ESX installations from any Windows PC.
- **Virtual Infrastructure Web Access** — Web interface for virtual machine management and remote consoles access.

Optional components of VMware Infrastructure are the following:

- **VMware VMotion™** — Enables the live migration of running virtual machines from one physical server to another with zero downtime, continuous service availability, and complete transaction integrity.
- **VMware High Availability (HA)** — Provides easy-to-use, cost-effective high availability for applications running in virtual machines. In the event of server failure, affected virtual machines are automatically restarted on other production servers that have spare capacity.
- **VMware Distributed Resource Scheduler (DRS)** — Allocates and balances computing capacity dynamically across collections of hardware resources for virtual machines.
- **VMware Consolidated Backup** — Provides an easy-to-use, centralized facility for agent-free backup of virtual machines that simplifies backup administration and reduces the load on ESX installations.
- **VMware Infrastructure SDK** — Provides a standard interface for VMware and third-party solutions to access VMware Infrastructure.

Physical Topology of the Datacenter

With VMware Infrastructure, IT departments can build a virtual datacenter using their existing industry standard technology and hardware. Users do not need to purchase specialized hardware. In addition, VMware Infrastructure allows users to create a virtual datacenter that is centrally managed by management servers and can be controlled through a wide selection of interfaces.

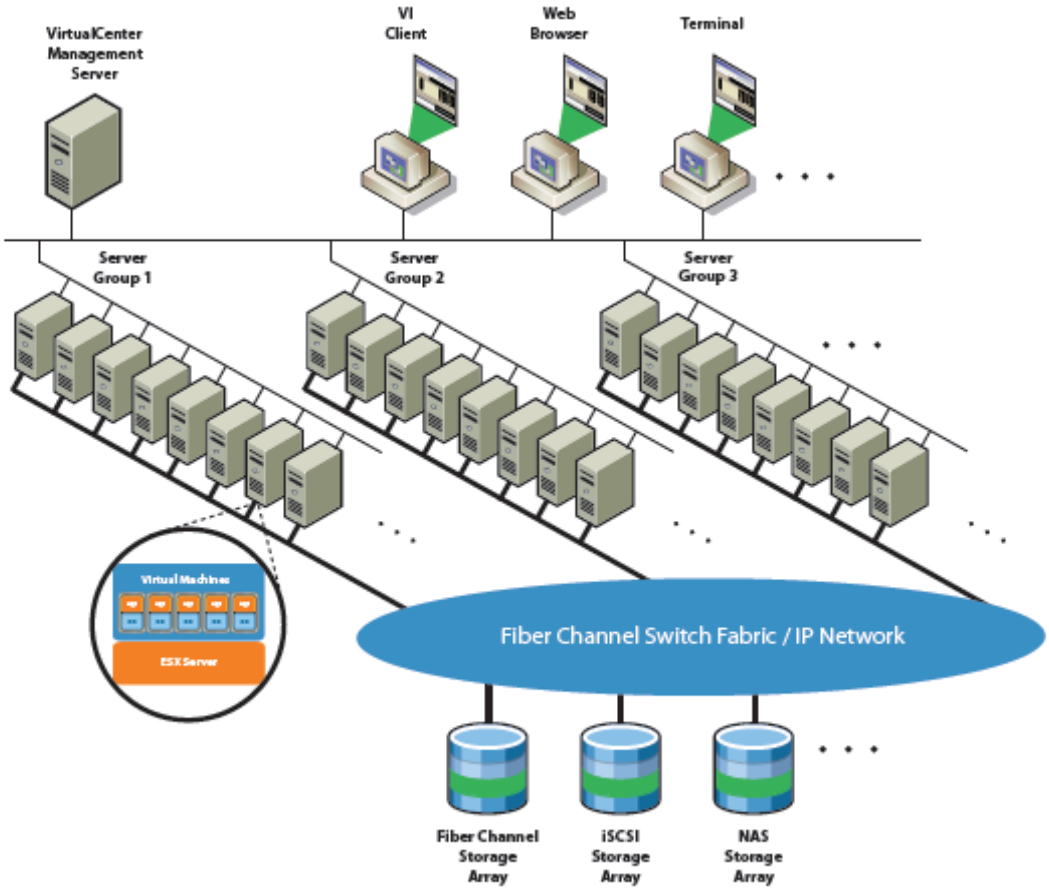


Figure 1-2. VMware Infrastructure Datacenter Physical Building Blocks

As Figure 1-2 shows, a typical VMware Infrastructure datacenter consists of basic physical building blocks such as x86 computing servers, storage networks and arrays, IP networks, a management server, and desktop clients.

Computing Servers

The computing servers are industry-standard x86 servers that run VMware ESX on the “bare metal.” Each computing server is referred to as a standalone **host** in the virtual environment. A number of similarly configured x86 servers can be grouped together with connections to the same network and storage subsystems to provide an aggregate set of resources in the virtual environment, called a **cluster**.

Storage Networks and Arrays

Fibre Channel SAN arrays, iSCSI SAN arrays, and NAS (network-attached storage) arrays are widely used storage technologies supported by VMware Infrastructure to meet different datacenter storage needs. Sharing the storage arrays among groups of servers via SANs allows aggregation of the storage resources and provides more flexibility in provisioning resources to virtual machines.

IP Networks

Each computing server can have multiple gigabit Ethernet network interface cards to provide high bandwidth and reliable networking to the entire datacenter.

Management Server

The VirtualCenter Management Server provides a convenient, single point of control to the datacenter. It runs on Windows 2003 Server to provide many essential datacenter services such as access control, performance monitoring, and configuration. It unifies the resources from the individual computing servers to be shared among virtual machines in the entire datacenter. VirtualCenter Management Server accomplishes this by managing the assignment of virtual machines to the computing servers. VirtualCenter Management Server also manages the assignment of resources to the virtual machines within a given computing server, based on the policies set by the system administrator.

Computing servers continue to function even in the unlikely event that VirtualCenter Management Server becomes unreachable (for example, the network is severed). Computing servers can be managed separately and continue to run their assigned virtual machines based on the latest resource assignments. Once the VirtualCenter Management Server becomes available, it can manage the datacenter as a whole again.

Virtual Datacenter Architecture

VMware Infrastructure virtualizes the entire IT infrastructure including servers, storage, and networks. It aggregates these various resources and presents a simple and uniform set of elements in the virtual environment. With VMware Infrastructure, you can manage IT resources like a shared utility, and provision them dynamically to different business units and projects without worrying about the underlying hardware differences and limitations.

Figure 1-3 shows the configuration and architectural design of a typical VMware Infrastructure deployment.

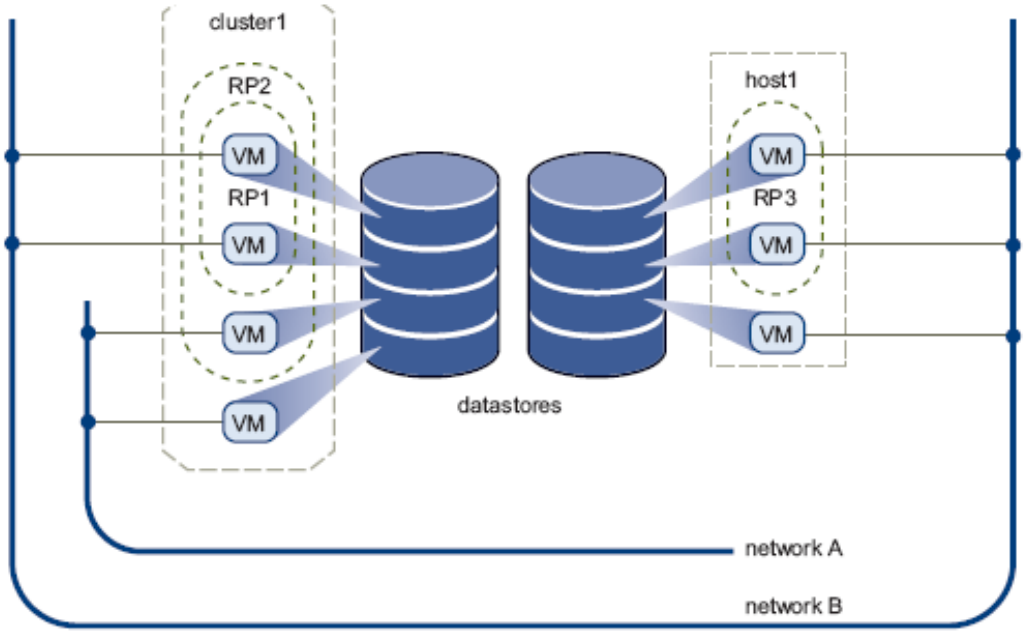


Figure 1-3. Virtual Datacenter Architecture

As shown in Figure 1-3, VMware Infrastructure presents a simple set of virtual elements used to build a virtual datacenter:

- Computing and memory resources called **hosts**, **clusters** and **resource pools**
- Storage resources called **datastores**
- Networking resources called **networks**
- **Virtual machines**

A **host** is the virtual representation of the computing and memory resources of a physical machine running VMware ESX. When one or more physical machines are grouped together to work and be managed as a whole, the aggregate computing and memory resources form a **cluster**. Machines can be dynamically added or removed from a cluster. Computing and memory resources from hosts and clusters can be finely partitioned into a hierarchy of **resource pools**.

Datastores are virtual representations of combinations of underlying physical storage resources in the datacenter. These physical storage resources can come from the local SCSI disks of the server, the Fibre Channel SAN disk arrays, the iSCSI SAN disk arrays, or NAS arrays. Networks in the virtual environment connect virtual machines to each other or to the physical network outside of the virtual datacenter.

Virtual machines are designated to a particular host, a cluster or resource pool, and a datastore when they are created. A virtual machine consumes resources, just like a physical appliance consumes electricity. While in a powered-off, suspended, or idle state, it consumes practically no resources. Once powered on, it consumes resources dynamically, using more as the workload increases and returning resources as the workload decreases.

Provisioning virtual machines is much faster and easier than provisioning physical machines. Once a virtual machine is provisioned, you can install the appropriate operating system and applications unaltered on the virtual machine to handle a particular workload, just as though you were installing them on a physical machine. To make things easier, you can even provision a virtual machine with the operating system and applications already installed and configured.

Resources are provisioned to virtual machines based on the policies set by the system administrator who owns the resources. The policies can reserve a set of resources for a particular virtual machine to guarantee its performance. The policies can also prioritize resources, and set a variable portion of the total resources to each virtual machine. A virtual machine is prevented from powering on (to consume resources) if powering on violates the resource allocation policies. For more information on resource management, see the VMware *Resource Management Guide*.

Hosts, Clusters, and Resource Pools

Clusters and resources pools from hosts provide flexible and dynamic ways to organize the aggregated computing and memory resources in the virtual environment, and link them back to the underlying physical resources.

A **host** represents the aggregate computing and memory resources of a physical x86 server. For example, if a physical x86 server has four dual-core CPUs running at 4GHz each with 32GB of system memory, then the host has 32GHz of computing power and 32GB of memory available for running the virtual machines that are assigned to it.

A **cluster** represents the aggregate computing and memory resources of a group of physical x86 servers sharing the same network and storage arrays. For example, if a group contains eight servers, each server has four dual-core CPUs running at 4GHz each with 32GB of memory. The cluster thus has 256GHz of computing power and 256GB of memory available for running the virtual machines assigned to it.

The virtual resource owners do not need to be concerned with the physical composition (number of servers, quantity and type of CPUs—whether multicore or hyperthreading) of the underlying cluster to provision resources. They simply set up the resource provisioning policies based on the aggregate available resources. VMware Infrastructure automatically assigns the appropriate resources dynamically to the virtual machines within the boundaries of those policies.

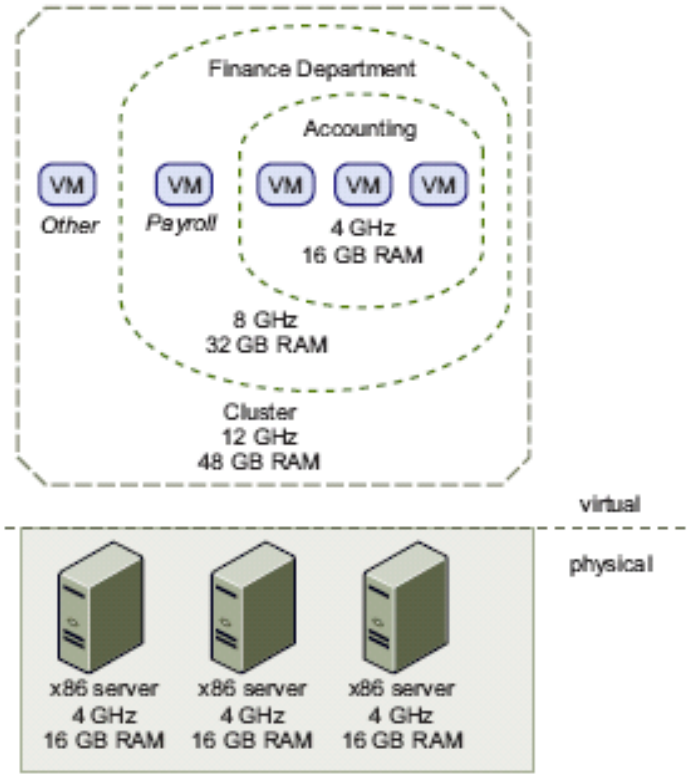


Figure 1-4. Hosts, Clusters, and Resource Pools

Resources pools provide a flexible and dynamic way to divide and organize computing and memory resources from a host or cluster. Any resource pools can be partitioned into smaller resource pools at a fine-grain level to further divide and assign resources to different groups, or to use resources for different purposes.

Figure 1-4 illustrates the concept of resource pools. Three x86 servers with 4GHz computing power and 16GB of memory each are aggregated to form a cluster with 12GHz of computing power and 48GHz of memory. A resource pool ("Finance Department") reserves 8GHz of computing power and 32GB of memory from the cluster, leaving 4GHz of computing power and 16GB of memory for the "Other" virtual machine. From the "Finance Department" resource pool, a smaller resource pool ("Accounting") reserves 4GHz of computing power and 16GB of memory for the virtual machines from the accounting department. That leaves 4GHz and 16GB of memory for the virtual machine called "Payroll."

Resources reserved for individual resource pools can be dynamically changed. Imagine that at the end of the year, Accounting's workload increases, so they want to increase the resource pool "Accounting" from 4GHz of computing power to 6GHz. You can simply make the change to the resource pool dynamically without shutting down the associated virtual machines.

Note that resources reserved for a resource pool or virtual machine are not taken away immediately, but respond dynamically to the demand. For example, if the 4GHz of computing resources reserved for the Accounting department are not being used, the virtual machine "Payroll" can make use of the remaining processing capacity during its peak time. When Accounting again requires the processing capacity,

"Payroll" dynamically gives back resources. As a result, even though resources are reserved for different resource pools, they are not wasted if not used by their owner.

As demonstrated by the example, resource pools can be nested, organized hierarchically, and dynamically reconfigured so that the IT environment matches the company organization. Individual business units can use dedicated infrastructure resources while still benefiting from the efficiency of resource pooling.

VMware VMotion, VMware DRS, and VMware HA

VMware VMotion, VMware DRS, and VMware HA are distributed services that enable efficient and automated resource management and high virtual machine availability.

VMware VMotion

Virtual machines run on and consume resources allocated from individual physical x86 servers through VMware ESX. VMotion enables the migration of running virtual machines from one physical server to another without service interruption, as shown in Figure 1-5. This migration allows virtual machines to move from a heavily loaded server to a lightly loaded one. The effect is a more efficient assignment of resources. Hence, with VMotion, resources can be dynamically reallocated to virtual machines across physical servers.

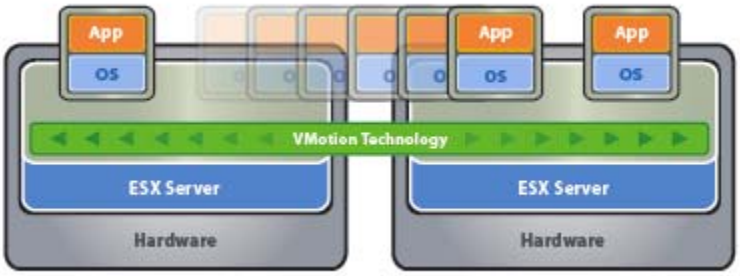


Figure 1-5. VMware VMotion

VMware DRS

Taking the VMotion capability one step further by adding an intelligent scheduler, VMware DRS enables the system administrator to set resource assignment policies that reflect business needs and let VMware DRS do the calculation and automatically handle the details of physical resource assignments. VMware DRS dynamically monitors the workload of the running virtual machines and the resource utilization of the physical servers within a cluster. It checks those results against the resource assignment policies. If there is a potential for violation or improvement, it uses VMotion to dynamically reassign virtual machines to different physical servers, as shown in Figure 1-6, to ensure that the policies are complied with and that resource allocation is optimal.

If a new physical server is made available, VMware DRS automatically redistributes the virtual machines to take advantage of it. Conversely, if a physical server needs to be taken down for any reason, VMware DRS redistributes its virtual machines to other servers automatically.

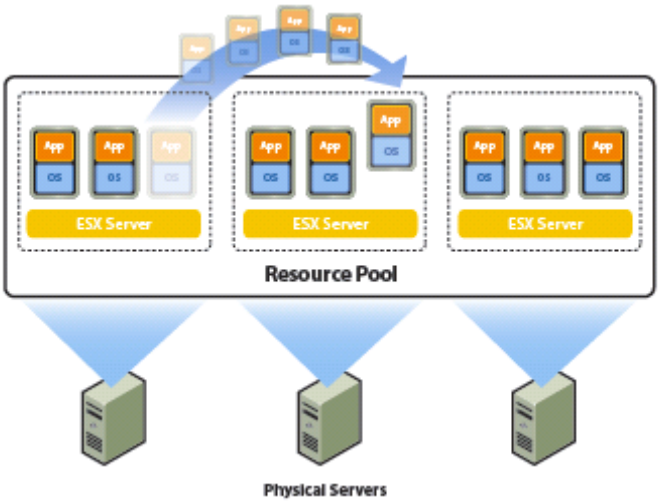


Figure 1-6. VMware DRS

For more information, see the VMware white paper titled "Resource Management with DRS." Also see the VMware *Resource Management Guide*.

VMware HA

VMware HA offers a simple, low-cost, high-availability alternative to application clustering. It enables a quick and automatic restart of virtual machines on a different physical server within a cluster if the hosting server fails. All applications within the virtual machines benefit from high availability, not just one (via application clustering).

VMware HA works by placing an agent on each physical server to maintain a "heartbeat" with the other servers in the cluster. As shown in Figure 1-7, loss of a "heartbeat" from one server automatically initiates the restarting of all affected virtual machines on other servers.

You can set up VMware HA simply by designating the priority order of the virtual machines to be restarted in the cluster. This is much simpler than the setup and configuration effort required for application clustering. Furthermore, even though VMware HA requires a certain amount of non-reserved resources to be maintained at all times to ensure that the remaining live servers can handle the total workload, it does not require doubling the amount of resources, as application clustering does.

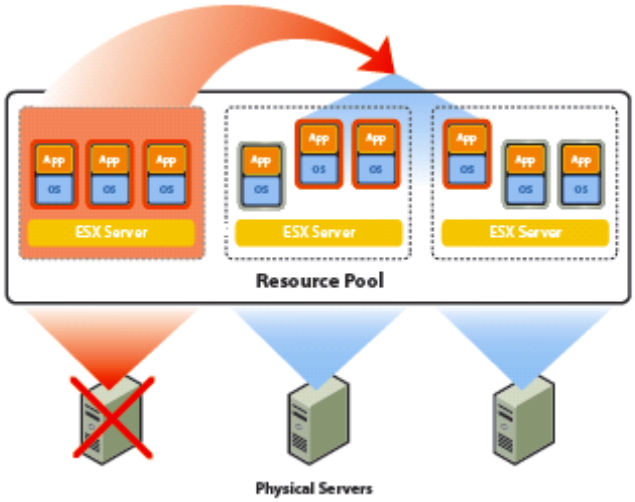


Figure 1-7. VMware HA

For more information, see the VMware white paper titled “Automating High Availability (HA) Services with VMware HA.”

VMware Consolidated Backup

VMware Infrastructure’s storage architecture enables a simple virtual machine backup solution: VMware Consolidated Backup (VCB). VCB provides a centralized facility for agent-less backup of virtual machines. As shown in Figure 1-8, VCB works in conjunction with third-party backup software residing on a separate backup proxy server (not on the server running VMware ESX), but does not require a backup agent running inside the virtual machines. The third-party backup software manages the backup schedule.

For each supported third-party backup application, there is a VCB integration module that is either supplied by the backup software vendor or by VMware. When a backup job is started, the third-party backup application runs a pre-backup script (part of the integration module) to prepare all virtual machines that are part of the current job for backup. VCB then creates a quiesced snapshot of each virtual machine to be protected. When a quiesced snapshot is taken, optional pre-freeze and post-thaw scripts in the virtual machine can be run before and after the snapshot is taken. These scripts can be used to quiesce critical applications running in the virtual machine. On virtual machines running Microsoft Windows operating systems, the operation to create a quiesced snapshot also ensures that the file systems are in a consistent state (file system sync) when the snapshot is being taken. The quiesced snapshots of the virtual machines to be protected are then exposed to the backup proxy server.

Finally, the third-party backup software backs up the files on the mounted snapshot to its backup targets. By taking snapshots of the virtual disks and backing them up at any time, VCB provides a simple, less intrusive and low overhead backup solution for virtual environments. You need not worry about backup windows.

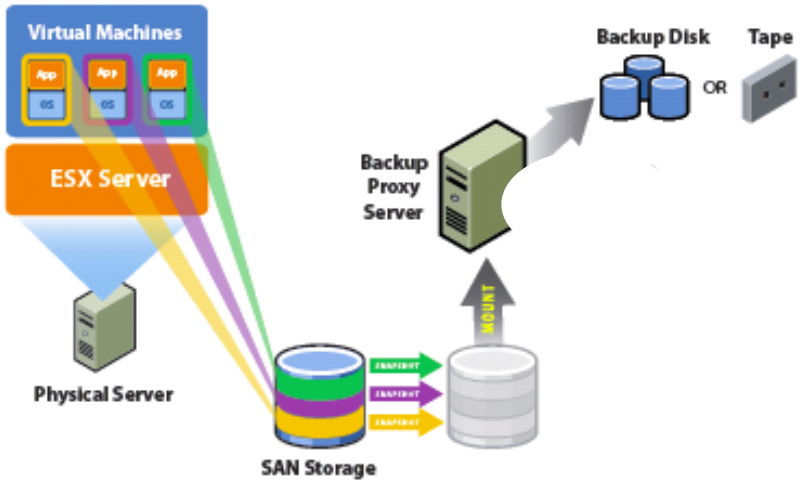


Figure 1-8. How Consolidated Backup Works

For more information, see the VMware white paper titled “Consolidated Backup in VMware Infrastructure 3.”

More About VMware Infrastructure Components

Figure 1-9 provides a high-level overview of the installable components in VMware Infrastructure system configurations.

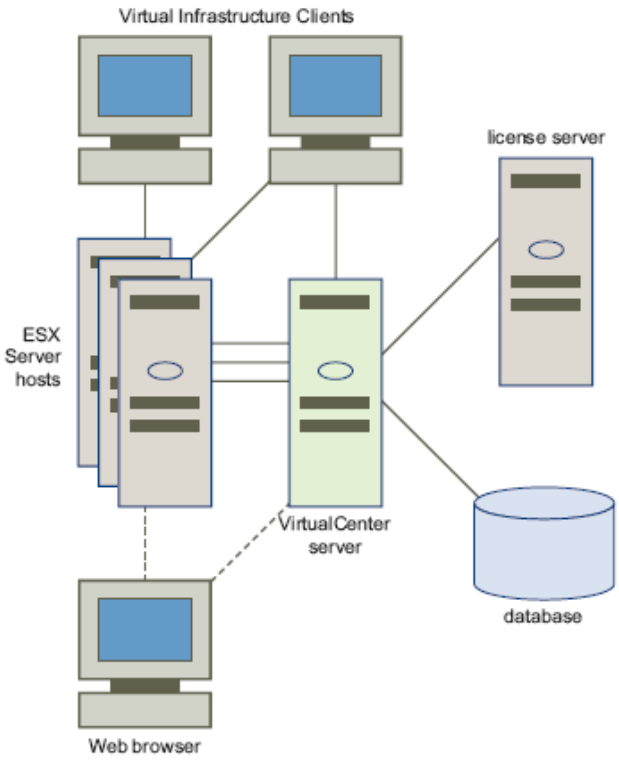


Figure 1-9. VMware Infrastructure Components

The components in this figure are the following:

- **VMware ESX Host** — ESX Server provides a virtualization layer that abstracts the processor, memory, storage, and networking resources of the physical host into multiple virtual machines. Virtual machines are created as a set of configuration and disk files that together perform all the functions of a physical machine.

  Through VMware ESX, you run the virtual machines, install operating systems, run applications, and configure the virtual machines. Configuration includes identifying the virtual machine's resources, such as storage devices.

  The server incorporates a resource manager and service console that provide bootstrapping, management, and other services that manage your virtual machines.

  Each ESX installation includes a Virtual Infrastructure (VI) Client to help you manage your host. If your ESX host is registered with the VirtualCenter Management Server, the VI Client accommodates all VirtualCenter features.

- **VirtualCenter Server** — The VirtualCenter Server installs on a Windows machine as a service. It allows you to centrally manage and direct actions on the virtual machines and the virtual machine hosts. The VirtualCenter Server allows the use of advanced VMware Infrastructure features such as VMware DRS, VMware HA, and VMotion.

  As a Windows service, the VirtualCenter Server runs continuously in the background, performing its monitoring and managing activities even when no VI Clients are connected and even if nobody is logged onto the computer where it resides. It must have network access to all the hosts it manages and be available for network access from any machine on which the VI Client is run.

- **Virtual Infrastructure (VI) Client** — The VI Client installs on a Windows machine, and is the primary method of interaction with virtual infrastructure. The VI Client runs on a machine with network access to the VirtualCenter Server or ESX host. The VI Client has two roles:

  - A console to operate virtual machines.
  - An administration interface into VirtualCenter Servers and ESX hosts. The interface presents different options depending on the type of server to which you are connected.

  The VI Client is the primary interface for creating, managing, and monitoring virtual machines, their resources, and their hosts. The VI Client is installed on a Windows machine that is separate from your ESX or VirtualCenter Server installation. While all VirtualCenter activities are performed by the VirtualCenter Server, you must use the VI Client to monitor, manage, and control the server. A single VirtualCenter Server or ESX installation can support multiple simultaneously-connected VI Clients.

- **Web Browser** — A browser allows you to download the VI Client from the VirtualCenter Server or ESX hosts. When you have appropriate logon credentials, a browser also lets you perform limited management of your VirtualCenter Server and ESX hosts using Virtual Infrastructure Web Access. VI Web Access provides a Web interface through which you can perform basic virtual machine management and configuration, and get console access to virtual machines. It is installed with

VMware ESX. Similar to the VI Client, VI Web Access works directly with an ESX host or through VirtualCenter.

- **VMware Service Console** – A command-line interface to VMware ESX for configuring your ESX hosts. Typically, this tool is used only in conjunction with a VMware technical support representative; VI Client and VI Web Access are the preferred tools for accessing and managing VMware Infrastructure components and virtual machines.
- **License Server** — The license server installs on a Windows system to authorize VirtualCenter Servers and ESX hosts appropriately for your licensing agreement. You cannot interact directly with the license server. Administrators use the VI Client to make changes to software licensing.
- **Virtual Center Database** — The VirtualCenter Server uses a database to organize all the configuration data for the virtual infrastructure environment and provide a persistent storage area for maintaining the status of each virtual machine, host, and user managed in the VirtualCenter environment.

In addition to the components shown in Figure 1-9, VMware Infrastructure also includes the following software components:

- **Datastore** – The storage locations for the virtual machine files specified when the virtual machines were created. Datastores hide the idiosyncrasies of various storage options (such as VMFS volumes on local SCSI disks of the server, the Fibre Channel SAN disk arrays, the iSCSI SAN disk arrays, or NAS arrays) and provide a uniform model for various storage products required by virtual machines.
- **VirtualCenter agent** – Software on each managed host that provides an interface between the VirtualCenter Server and the host agent. It is installed the first time any ESX host is added to the VirtualCenter inventory.
- **Host agent** – Software on each managed host that collects, communicates, and executes the actions received through the VI Client. It is installed as part of the ESX installation.

Chapter 6 provides more information on the operation of VMware Infrastructure software components and on how to use the VI Client to manage VMware Infrastructure using SAN storage.

More About the VMware ESX Architecture

The VMware ESX architecture allows administrators to allocate hardware resources to multiple workloads in fully isolated virtual machine environments. The following figure shows the main components of an ESX host.

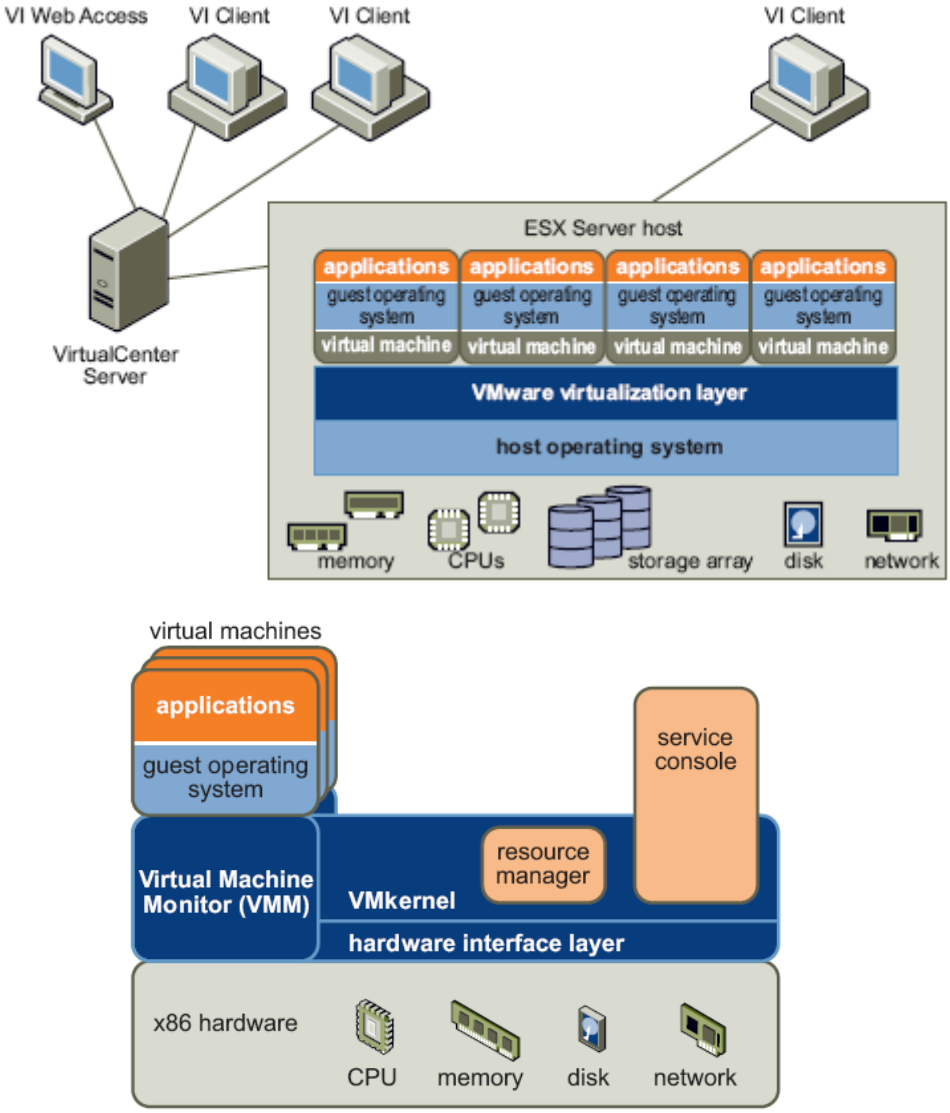


Figure 1-10. VMware ESX Architecture

A VMware ESX system has the following key components:

- **Virtualization Layer** — This layer provides the idealized hardware environment and virtualization of underlying physical resources to the virtual machines. It includes the Virtual Machine Monitor (VMM), which is responsible for virtualization, and VMkernel. VMkernel manages most of the physical resources on the hardware, including memory, physical processors, storage, and networking controllers.

The virtualization layer schedules both the service console running on the ESX host and the virtual machine operating systems. The virtualization layer manages how the operating systems access physical resources. VMkernel needs its own drivers to provide access to the physical devices. VMkernel drivers are modified Linux drivers, even though VMkernel is not a Linux variant.

- **Hardware Interface Components** — The virtual machine communicates with hardware, such as a CPU or disk, using hardware interface components. These components include device drivers, which enable hardware-specific service delivery while hiding hardware differences from other parts of the system.
- **User Interface** — Administrators can view and manage ESX hosts and virtual machines in several ways.
  - A **VI Client** can connect directly to the ESX host. This is appropriate if your environment has only one host.

    A VI Client can also connect to a VirtualCenter Management Server and interact with all ESX hosts managed by that VirtualCenter Server.
  - The **VI Web Access Client** allows you to perform many management tasks using a browser-based interface. The operations that the VI Web Access Client provides are a subset of those available using the VI Client.
  - The **service console** command-line interface is used only rarely. Starting with ESX 3, the VI Client replaces the service console for most interactions. (Commands have also changed from previous versions of VMware ESX).

VMware Virtualization

The VMware virtualization layer is common across VMware desktop products (such as VMware Workstation) and server products (such as VMware ESX). This layer provides a consistent platform for developing, testing, delivering, and supporting application workloads, and is organized as follows:

- Each virtual machine runs its own operating system (the guest operating system) and applications.
- The virtualization layer provides the virtual devices that map to shares of specific physical devices. These devices include virtualized CPU, memory, I/O buses, network interfaces, storage adapters and devices, human interface devices, and BIOS.

CPU, Memory, and Network Virtualization

A VMware virtual machine offers complete hardware virtualization. The guest operating system and applications running on a virtual machine do not need to know about the actual physical resources they are accessing (such as which physical CPU they are running on in a multiprocessor system, or which physical memory is mapped to their pages).

- **CPU Virtualization** — Each virtual machine appears to run on its own CPU (or a set of CPUs), fully isolated from other virtual machines. Registers, the translation look-aside buffer, and other control structures are maintained separately for each virtual machine.

Most instructions are executed directly on the physical CPU, allowing resource-intensive workloads to run at near-native speed. The virtualization layer also safely performs privileged instructions specified by physical CPUs.

- **Memory Virtualization** — A contiguous memory space is visible to each virtual machine even though the allocated physical memory might not be contiguous. Instead, noncontiguous physical pages are remapped and presented to each virtual machine. With unusually memory-intensive loads, server memory becomes overcommitted. In that case, some of the physical memory of a virtual machine might be mapped to shared pages or to pages that are unmapped or swapped out.

  VMware ESX performs this virtual memory management without the information the guest operating system has, and without interfering with the guest operating system's memory management subsystem.

- **Network Virtualization** — The virtualization layer guarantees that each virtual machine is isolated from other virtual machines. Virtual machines can talk to each other only via networking mechanisms similar to those used to connect separate physical machines.

  Isolation allows administrators to build internal firewalls or other network isolation environments, allowing some virtual machines to connect to the outside while others connect only via virtual networks through other virtual machines.

Virtual SCSI and Disk Configuration Options

VMware Infrastructure also provides for virtualization of data storage. In an ESX environment, each virtual machine includes from one to four virtual SCSI HBAs (host bus adapters). These virtual adapters may appear as either BusLogic or LSI Logic SCSI controllers. They are the only types of SCSI controllers that are accessible by a virtual machine.

Each virtual disk accessible by a virtual machine (through one of the virtual SCSI adapters) resides in VMFS or NFS storage volumes, or on a raw disk. From the standpoint of the virtual machine, each virtual disk appears as if it were a SCSI drive connected to a SCSI adapter. Whether the actual physical disk device is being accessed through SCSI, iSCSI, RAID, NFS, or Fibre Channel (FC) controllers is transparent to the guest operating system and to applications running on the virtual machine. Chapter 3, “VMware Virtualization of Storage,” provides more details on the virtual SCSI HBAs, as well as specific disk configuration options using VMFS and raw disk device mapping (RDM).

Software and Hardware Compatibility

In the VMware ESX architecture, the operating system of the virtual machine (the guest operating system) interacts only with the standard, x86-compatible virtual hardware presented by the virtualization layer. This allows VMware products to support any x86-compatible operating system.

In practice, VMware products support a large subset of x86-compatible operating systems that are tested throughout the product development cycle. VMware documents the installation and operation of these guest operating systems and trains its technical personnel in supporting them.

Most applications interact only with the guest operating system, not with the underlying hardware. As a result, you can run applications on the hardware of your choice as long as you install a virtual machine with the operating system the application requires.

2 Storage Area Network Concepts

VMware ESX 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. 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. Using VMware ESX together with a SAN provides extra storage for consolidation, improves reliability, and facilitates the implementation of both disaster recovery and high availability solutions. The physical components of a SAN can be grouped in a single rack or datacenter, or can be connected over long distances. This flexibility makes a SAN a feasible solution for businesses of any size: the SAN can grow easily with the business it supports. SANs include Fibre Channel storage or IP storage. The term FC SAN refers to a SAN using Fibre Channel protocol while the term IP SAN refers to a SAN using an IP-based protocol. When the term SAN is used by itself, this refers to FC or IP based SAN.

To use VMware ESX effectively with a SAN, you need to be familiar with SAN terminology and basic SAN architecture and design. This chapter provides an overview of SAN concepts, shows different SAN configurations that can be used with VMware ESX in VMware Infrastructure solutions, and describes some of the key operations that users can perform with VMware SAN solutions.

Topics included in this chapter are the following:

NOTE: In this chapter, computer systems are referred to as **servers** or **hosts**.

SAN Component Overview

Figure 2-1 provides a basic overview of a SAN configuration. (The numbers in the text below correspond to number labels in the figure.) 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 shown in the figure are also discussed in more detail in "SAN Components" starting on page 25:

- **Fabric (4)** — A configuration of multiple Fibre Channel protocol-based switches connected together is commonly referred to as a FC fabric or FC SAN. A collection of IP networking switches that provides connectivity to iSCSI storage is referred to as iSCSI fabric or iSCSI SAN. The SAN fabric is the actual network portion of the SAN. The connection of one or more SAN switches creates a fabric. For Fibre Channel the fabric can contain between one and 239 switches. (Multiple switches required for redundancy.) Each FC switch is identified by a unique domain ID (from 1 to 239). Fibre Channel protocol is used to communicate over the entire network. A FC SAN or an iSCSI SAN can consist of two separate fabrics for additional redundancy.
- **SAN Switches (3)** — SAN switches connect various elements of the SAN together, such as HBAs, other switches, and storage arrays. FC SAN switches and networking switches provide routing functions. SAN switches also allow administrators to set up path redundancy in the event of a path failure, from a host server to a SAN switch, from a storage array to a SAN switch, or between SAN switches.
- **Connections: Host Bus Adapters (5) and Storage Processors (6)** — Host servers and storage systems are connected to the SAN fabric through ports in the SAN fabric.
  - A host connects to a SAN fabric port through an HBA.
  - Storage devices connect to SAN fabric ports through their storage processors (SPs).
- **SAN Topologies** — Figure 2-1 illustrates a fabric topology. For Fibre Channel, FC SAN topologies include Point-To-Point (a connection of only two nodes that involves an initiator or a host bus adapter connecting directly to a target device), Fibre Channel Arbitrated Loop (FC-AL ring topology consisting of up to 126 devices in the same loop), and Switched Fabric (a connection of initiators and storage devices using a switch for routing).

NOTE: See the VMware *SAN Compatibility Guide* for specific SAN vendor products and configurations supported with VMware Infrastructure.

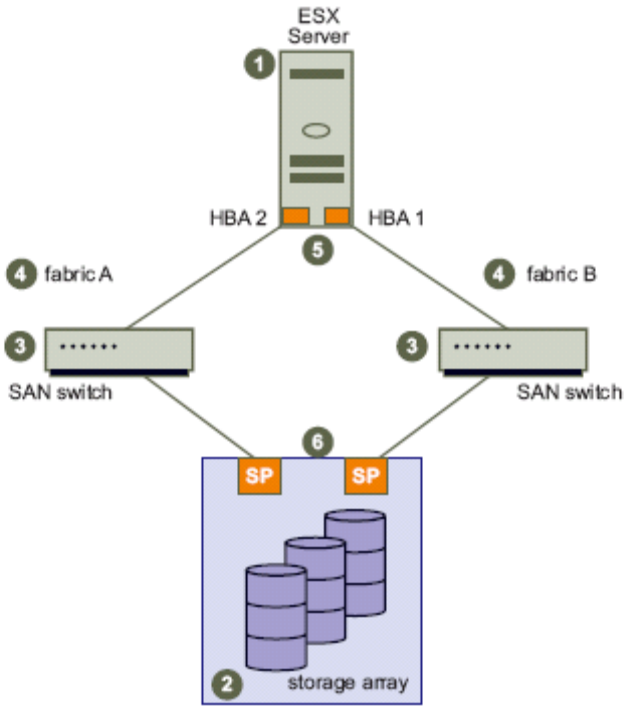


Figure 2-1. FC SAN Components

In this figure, implementing an FC-protocol SAN solution, the ESX host is equipped with a dedicated hardware FC HBA and both SAN switches and storage arrays are FC-based. Multiple FC SAN switches provide multiple paths to make a connection to SAN storage arrays. (See “Multipathing and Path Failover” later in this chapter for more information.)

In an iSCSI SAN solution, ESX hosts may use dedicated iSCSI HBAs or an Ethernet NIC HBA configured to provide software-based iSCSI protocol support. In an iSCSI solution, switching is provided by a typical TCP/IP LAN and the storage arrays support the iSCSI protocol over Ethernet (TCP/IP) connections. (For more information on iSCSI implementation details using VMware Infrastructure, see Appendix B.)

How a SAN Works

SAN components interact as follows when a host computer wants to access information residing in SAN storage:

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 (for FC protocol based storage) or IP packets (for IP storage).The request is accepted by the HBA for that host. Binary data is encoded from eight-bit to ten-bit for serial transmission on optical cable.

3. At the same time, the request is packaged according to the rules of the FC protocol (for FC protocol based storage) or the rules of IP storage protocols (FCIP, iFCP, or iSCSI).
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 routes it to the storage processor, which sends it on to the storage device.

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

SAN Components

The components of a SAN can be grouped as follows:

- Host Components
- Fabric Components
- Storage Components

Figure 2-2 shows the component layers in SAN system configurations.

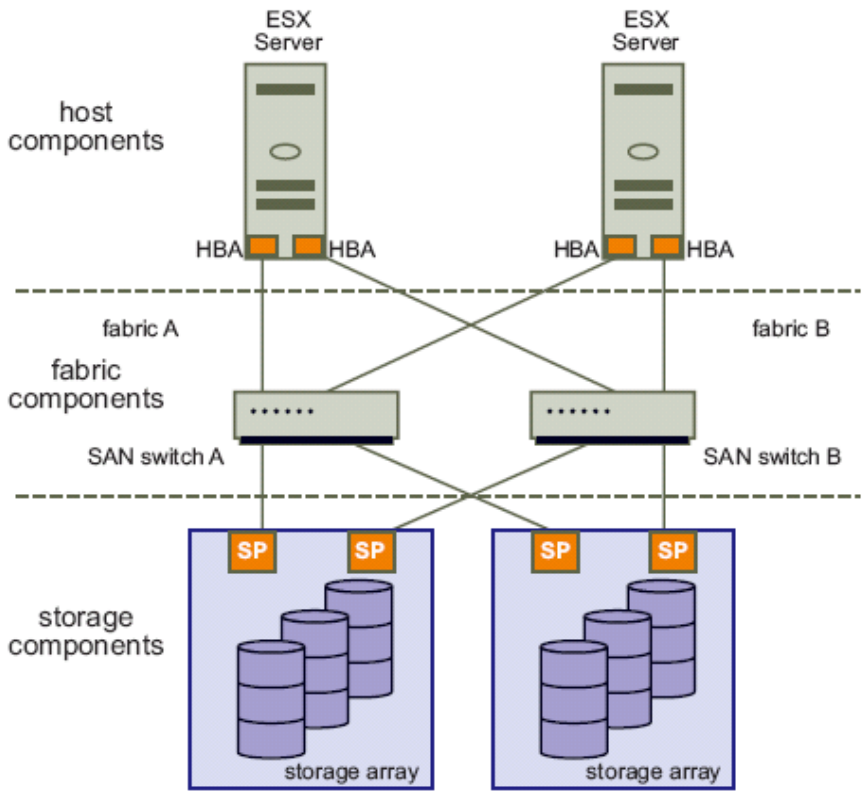


Figure 2-2. SAN Component Layers

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 individual host servers. 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 connected 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.
- **FC Data Routers** — FC Data routers are intelligent bridges between SCSI devices and FC devices in the FC SAN. Servers in the FC SAN can access SCSI disk or tape devices in the FC SAN through the FC data routers in the FC fabric layer.
- **Cables** — SAN cables are usually special fiber optic cables that connect all of the fabric components. The type of SAN cable, the fiber optic signal, and switch licensing determine the maximum distances between SAN components, and contribute to the total bandwidth rating of the SAN.
- **Communications Protocol** — For Fibre Channel storage, FC fabric components communicate using the FC communications protocol. FC is the storage interface protocol used for most 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. For IP storage, IP fabric components communicate using FCIP, iFCP or iSCSI protocol.

Storage Components

The storage components of a SAN are the storage arrays. Storage arrays include the storage processors (SPs), which provide 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 (Redundant Array of Independent Drives) and volume functionality.

Storage Processors

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 SP. In most cases, this is the FC protocol. SPs provide internal access to the drives, which can use either a switch or a bus architecture. In high-end storage systems, drives are normally connected in loops. The 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

Disk arrays are groups of multiple disk devices and are the typical SAN disk storage devices. 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, the array groups several drives 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 volume. A volume is a single unit of storage. Depending on the host system environment, a volume is also known as a logical drive. From a VI Client, a volume looks like any other storage unit available for access.

In advanced storage arrays, RAID groups can have one or more volumes created for access by one or more servers. The ability to create more than one volume 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 volume.

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 features such as snapshots, internal copies, and replication.

- A snapshot is a point-in-time copy of a volume. Snapshots are used as backup sources for the overall backup procedures defined for the storage array.
- Internal copies allow data movement from one volume to another, providing additional copies for testing.

- Replication provides constant synchronization between volumes 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 backup capabilities and processes on a SAN.

- 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, including the following topics:

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:

- **WWN** — The World Wide Node Name is a globally unique identifier for a Fibre Channel HBA. Each FC HBA can have multiple ports, each with its own unique WWPN.
- **WWPN** — This World Wide Port Name is a globally unique identifier for a port on a FC HBA. The FC switches discover the WWPN of a device or host and assign a port address to the device. To view the WWPN using the VI Client, click the host's **Configuration** tab and choose **Storage Adapters**. You can then select the storage adapter for which you want to see the WWPN.

- **Port_ID or Port Address** — Within the FC SAN, each port has a unique port ID that serves as the FC address for the port. This ID 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.
- **iSCSI Qualified Name (iqn)** – a globally unique identifier for an initiator or a target node (not ports). It is UTF-8 encoding with human readable format of up to 233 bytes. This address is not used for routing. Optionally there is an extended version called Extended Unique Identifier (eui).

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

A 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 volume on a storage array through more than one path. Having more than one path from a host to a volume is called **multipathing**.

By default, VMware ESX systems use only one path from the host to a given volume at any time. If the path actively being used by the VMware ESX system fails, the server selects another of the available paths. The process of detecting a failed path by the built-in ESX multipathing mechanism and switching to another path is called **path failover**. A path fails if any of the components along the path fails, which may include the HBA, cable, switch port, or storage processor. This method of server-based multipathing may take up to a minute to complete, depending on the recovery mechanism used by the SAN components (that is, the SAN array hardware components).

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 volumes simultaneously through all the SPs 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 volume. The other SP acts as backup for the volume and may be actively servicing other volume 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.

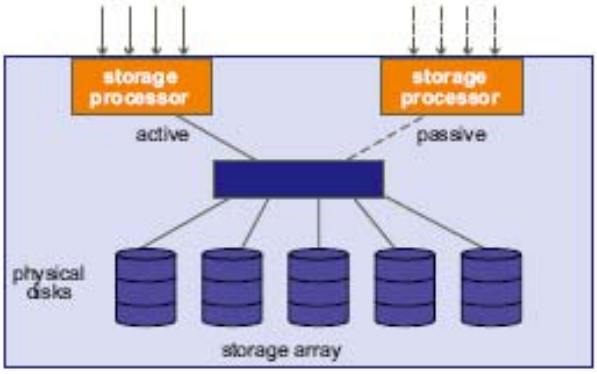


Figure 2-3. Active/Passive Storage Array

Using active/passive arrays with a fixed path policy can potentially lead to path thrashing. See "Understanding Path Thrashing" on page 182. In Figure 2-3, one storage processor is active while the other is passive. Data arrives through the active array only.

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. The main issues with zoning that you need to consider are the following:

- Soft versus hard zoning. For more information, go to:

  http://www.snia.org/education/dictionary/)
- Zone security
- Zone size and merging issues

When a SAN is configured using zoning, the devices outside a zone are not visible to the devices inside the zone. When there is one HBA or initiator to a single storage processor port or target zone, it is commonly referred to as single zone. This type of single zoning protects devices within a zone from fabric notifications, such as Registered State Change Notification (RSCN) changes from other zones. In addition, SAN traffic within each zone is isolated from the other zones. Thus, using single zone is a common industry practice.

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 volumes. You can use zoning in several ways.

- **Zoning for security and isolation** — You can manage zones defined for testing independently within the SAN so they do not interfere with the activity going on in the production zones. Similarly, you can 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 you have 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. Different vendors might refer to LUN masking as selective storage presentation, access control, or partitioning.

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.

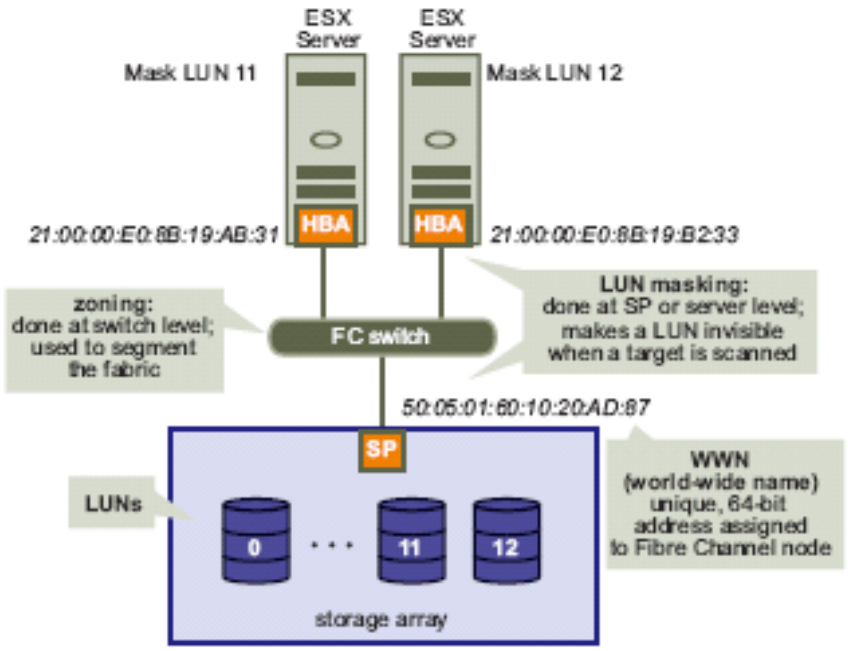


Figure 2-4. LUN Zoning and Masking

A volume has slightly different behavior, depending on the type of host that is accessing it. Usually, the host type assignment deals with operating-system-specific features or issues. ESX systems are typically configured with a host type of Linux for volume access. See Chapter 6, "Managing ESX Systems That Use SAN Storage" and the VMware Knowledge Base for more information.

IP Storage

Storage Area Network encompasses Fibre Channel (FC) storage, using FC protocol, or other protocols commonly referred to as Internet Protocol (IP) storage. Various IP storage protocols are Fibre Channel tunneling in an IP Protocol (FCIP), Internet Fibre Channel Protocol (iFCP), and SCSI encapsulated over the internet (iSCSI) Protocol.

Below are the advantages to using IP storage:

- Providing global access to an existing IP infrastructure
- Existing IP network assumes that administration skills are existing and does not require much additional training to IT staff
- The protocol is suitable for LAN, MAN and WAN (one network for the entire enterprise deployment)
- IP protocols are routed protocol so can be scalable
- IP protocols can be combined with FC SAN for penetration across multiple organizations (connects to FC via iSCSI gateway)
- Leverage existing security benefits inherent in IP protocols

The following table shows the differences in IP storage protocols.

Protocol/App	FCIP	iFCP	iSCSI
Protocol	"Tunnels" FC frames over IP. Merge fabrics together	Network Address Translation (NAT) to transport FC frames over IP	Transport serial SCSI-3 over TCP/IP. Device-device connection.
Application	Provides an extension to fabrics	Provides an extension to fabrics	Host to Target connectivity

FCIP bridges FC SANs together using Inter Switch Links (ISL) over long geographical distance. This allows SAN fabrics to merge together. This protocol is based on the FC BackBone (FC-BB) standard. Unlike FCIP, iFCP does not merge fabrics. Instead it performs a network address translation (NAT) functions to route FC frames. The NAT function performs the iFCP gateway switch function. iSCSI is a pure IP solution that encapsulates serial SCSI-3 data in IP packets. Similar with FCIP and iFCP, flow control and reliability are managed by the TCP layer. There are additional iSCSI security benefits such as in firewalls, intrusion detection systems (IDS), virtual private networks (VPN), encryption and authentication.

More Information on SANs

You can find information about SAN in print and on the Internet. A number of Web-based resources are recognized in the SAN industry for the wealth of information they provide. These sites are:

- http://www.fibrechannel.org/
- http://www.searchstorage.com
- http://www.snia.org
- http://www.t11.org/index.html

Because the industry is always changing, you are encouraged to stay abreast of the latest developments by checking these resources frequently.

3 VMware Virtualization of Storage

VMware Infrastructure enables enterprise-class storage performance, functionality, and availability without adding complexity to the user applications and guest operating systems. To satisfy the demands of business-critical applications in an enterprise environment, and to do so effectively and efficiently, virtual infrastructure must make optimal use of both server and storage resources. The VMware Infrastructure architecture, combined with the range of resource allocation, management, and optimization tools that VMware provides, make that job easier. It provides flexibility in scaling systems to meet changing business demands and helps to deliver high availability, backup, and disaster recovery solutions that are vital to businesses.

The previous chapter provided background on SAN systems and design. This chapter builds on that knowledge, providing an overview of the VMware storage architecture and describing how VMware Infrastructure can take advantage of SAN storage in implementing VMware virtualization solutions. When looking at storage, customers face many challenges in picking the right mix of features, performance, and price. Besides cost, the most common criteria by which customers need to evaluate storage solutions are reliability, availability, and scalability (also referred to as RAS). This chapter describes the various storage options available, and helps you choose and implement the solution that best meets your needs.

Topics included in this chapter are the following:

Storage Concepts and Terminology

To use VMware Infrastructure and VMware ESX effectively with a SAN or any other types of data storage system, you must have a working knowledge of some essential VMware Infrastructure, VMware ESX and storage concepts. Here is a summary:

- **Datastore** — This is a formatted logical container, analogous to a file system on a logical **volume**. The datastore holds virtual machine files and can exist on different types of physical storage including SCSI, iSCSI, Fibre Channel SAN, or NFS. Datastores can be of the two types: VMFS-based or NFS-based (version 3).
- **Disk** or **drive** — These terms refer to a physical disk.
- **Disk partition** — This is a part of a hard disk that is reserved for a specific purpose. In the context of ESX storage, disk partitions on various physical storage devices can be reserved and formatted as datastores.
- **Extent** — In the context of ESX systems, an extent is a **logical volume** on a physical storage device that can be dynamically added to an existing VMFS-based datastore. The datastore can stretch over multiple extents, yet appear as a single volume analogous to a **spanned volume**.
- **Failover path** — The redundant physical path that the ESX system can use when communicating with its networked storage. The ESX system uses the failover path if any component responsible for transferring storage data fails.
- **Fibre Channel (FC)** — A high-speed data transmitting technology is used by ESX systems to transport SCSI traffic from virtual machines to storage devices on a SAN. The Fibre Channel Protocol (FCP) is a packetized protocol used to transmit SCSI serially over a high-speed network consisting of routing appliances (called **switches)** that are connected together by optical cables.
- **iSCSI (Internet SCSI)** — This packages SCSI storage traffic into TCP so it can travel through IP networks, instead of requiring a specialized FC network. With an iSCSI connection, your ESX system (initiator) communicates with a remote storage device (target) as it would do with a local hard disk.
- **LUN (logical unit number)** — The logical unit or identification number for a storage volume. This document refers to logical storage locations as volumes rather than LUNs to avoid confusion.
- **Multipathing** — A technique that lets you use more than one physical path, or an element on this path, for transferring data between the ESX system and its remote storage. The redundant use of physical paths or elements, such as adapters, helps ensure uninterrupted traffic between the ESX system and storage devices.
- **NAS (network-attached storage)** — A specialized storage device that connects to a network and can provide file access services to ESX systems. ESX systems use the NFS protocol to communicate with NAS servers.
- **NFS (network file system)** — A file sharing protocol that VMware ESX supports to communicate with a NAS device. (VMware ESX supports NFS version 3.)
- **Partition** — A divided section of a volume that is formatted.
- **Raw device** — A logical volume used by a virtual machine that is not formatted with VMFS.

- **Raw device mapping (RDM)** — A special file in a VMFS volume that acts as a proxy for a raw device and maps a logical volume directly to a virtual machine.
- **Spanned volume** — A single volume that uses space from one or more logical volumes using a process of concatenation.
- **Storage device** — A physical disk or storage array that can be either internal or located outside of your system, and can be connected to the system either directly or through an adapter.
- **Virtual disk** — In an ESX environment, this is a partition of a volume that has been formatted with a file system or is a volume that has not been formatted as a VMFS volume. If the virtual disk is not formatted as a VMFS volume, then it is a RDM volume.
- **VMFS (VMware File System)** — A high-performance, cluster file system that provides storage virtualization optimized for virtual machines.
- **Volume** — This term refers to an allocation of storage. The volume size can be less than or more than a physical disk drive. An allocation of storage from a RAID set is known as a volume or a logical volume.

NOTE: For complete descriptions of VMware Infrastructure and other storage acronyms and terms, see the glossary at the end of this guide.

LUNs, Virtual Disks, and Storage Volumes

As Figure 3-1 below illustrates, in RAID storage arrays such as those used in SAN systems, a volume is a logical storage unit that typically spans multiple physical disk drives. To avoid confusion between physical and logical storage addressing, this document uses the term **volume** instead of **LUN** to describe a storage allocation. This storage allocation, or volume B as shown in Figure 3-1, can be formatted using VMFS-3 or left unformatted for RDM mode storage access. Any volume can further be divided into multiple partitions. Each partition or RDM volume that is presented to ESX host is identified as a virtual disk. Each virtual disk (VMFS-3 or RDM) can either store a virtual machine operating system boot image or serve as storage for virtual machine data. When a virtual disk contains an operating system boot image, it is referred to as a **virtual machine**.

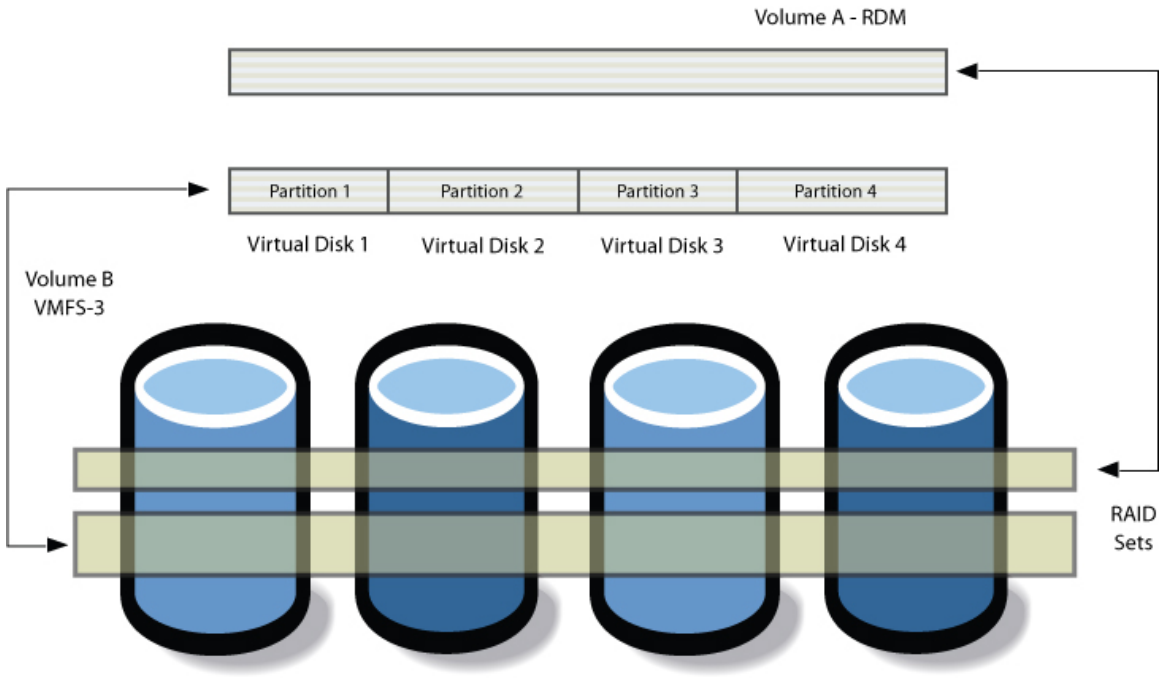


Figure 3-1. Volumes Spanning Physical Disks in a RAID Storage Array

When multiple partitions are created in a volume, each is formatted as VMFS-3. Successive partitions to be created in the same volume are also formatted as VMFS-3. A unique LUN is given to each volume from the RAID vendor's array management software. The LUN is then presented to a physical host, such as an ESX host.

It is important to differentiate between a volume and a **LUN**. In Figure 3-2 below, here are two volumes, A and B. The RAID array management software gives volume A the unique LUN of 6 and gives volume B a unique LUN of 8. Both LUNs are then presented to an ESX host so the host now has read/write access to these two volumes. Suppose these volumes A and B are replicated to a remote data site. The replication process creates two new volumes, C and D, which are exact copies of volumes A and B. The new volumes are presented to the same ESX host with two different LUN ID numbers, 20 and 21.

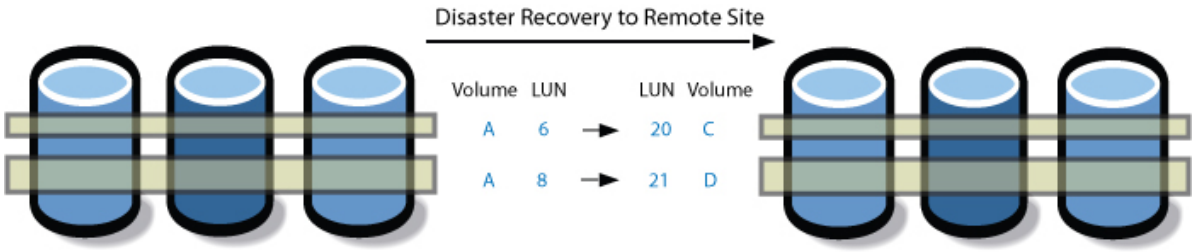


Figure 3-2. LUN Addressing of Storage Array Volumes

As part of the data replication schema, only storage volumes referenced by the new LUN IDs 20 and 21 can be active (with read/write access), while storage volumes accessed with LUN IDs 6 and 8 are now in read-only mode. At some point in the

future, the two new volumes C and D, with LUN IDs 20 and 21, might revert to read-only mode. In this case, read/write access is given back to volumes A and B with LUN IDs 6 and 8. The associations between the volumes can be broken, depending on a user's action. For example, if the user decides to break the synchronization between replicas, the association between the four volumes A, B, C and D with LUN IDs 6, 8, 20 and 21, respectively, is broken. In that case, all volumes have read/write access. Thus, it is important to recognize the difference in meaning between volumes and LUNs. Using the terms interchangeably in data replication may confuse users trying to distinguish data residing on the original volume from data stored in the replica volume.

Addressing IT Storage Challenges

This section describes different storage system solutions and compares their specific features, capabilities, advantages, and disadvantages. Specific business and application requirements usually drive customers' decisions to use specific technologies. Here is a brief summary of different solutions available for virtualization within a SAN environment:

- Traditional SCSI or Direct-Attach Storage (DAS)
  - Limited to the number of available PCI buses per server
  - Physical device limitation (distance and number of devices) per PCI bus (per HBA)
  - Devices limited to use by a single server
- Network-Attached Storage (NAS)
  - TCP/IP used to service file I/O requests from network clients
- Storage Area Network (SAN)
  - Fibre Channel attached storage
  - IP storage (FCIP, IFCP, iSCSI)

The following table describes the interface and data transfer features of the different solutions, and the performance benefits of each solution:

Technology	Application	Transfers	Interface	Performance
Fibre Channel	Datacenter	Block access of data/volume	FC HBA	Typically high (due to dedicated network)
NAS	Small and medium-sized businesses (SMB)	File (no direct volume access)	Network adapter	Typically medium (depends on integrity of LAN)
iSCSI	Small and medium-sized businesses (SMB)	Block access of data/volume	iSCSI HBA	Typically medium (depends on integrity of LAN)
DAS	Branch office	Block access	SCSI HBA	Typically high (due to dedicated bus)

For a branch office application, DAS provides a simple-to-manage solution that yields high performance and can be deployed in little time (within a day). SCSI protocol is a proven technology that is the key mechanism for delivering and managing storage.

Because DAS uses SCSI as the underlying technology, this solution is highly reliable and quick to deploy.

The disadvantages of DAS are inherent in its design:

- Address ranges limit the number of physical devices you can connect
- Parallel bus technologies limit the length of the cables used to connect SCSI devices together
- Sharing storage is not allowed with SCSI devices.

These technology limitations become critical when your business needs to expand. Business growth also generally means that your applications need access to more data. So, your storage configuration needs to be able to scale up (more devices, more servers, and more applications). DAS solutions can scale only to the maximum number of addresses allowed by SCSI, which is 15 devices per SCSI bus.

For small to medium-size businesses (SMBs), the most economical and efficient storage solution to use is typically NAS. The key benefits of NAS are allowing multiple servers to access the same data storage array, thus reducing overall IT infrastructure costs, and ease of remote management. NAS uses the Network File System (NFS) protocol to manage data and provides a mechanism to transfer data across an LAN or WAN infrastructure. Because many server applications can share the same NAS array, contention on the same storage volume can affect performance. The failure of one storage volume can affect multiple applications at the same time. In addition, LAN congestion can limit NAS performance during backups. These potential bottlenecks apply particularly to IP storage. Because IP storage is part of the LAN and WAN infrastructure, limitations in areas such as network routing apply.

NAS is currently being used extensively across a wide range of businesses in different industries — the deciding factor in using NAS versus FC or iSCSI is not related to the type of business an organization is in, but rather the characteristics of the applications the organization runs. NAS is generally used for file sharing and Tier II type applications, while FC is more commonly used for higher-end Tier I applications like large Oracle databases, high I/O applications, and OLTP.

For mission-critical applications such as database applications, Fibre Channel (FC) protocol is the technology of choice. FC is the protocol used for SANs. FC fabric is easy to scale and maintain (at a price) and is fault-tolerant. VMware ESX provides an enterprise-grade operating system tested extensively on many FC storage arrays in both VMware Quality Assurance laboratories as well as at OEM partner test facilities. With FC technology, VMware ESX can provide end-to-end fault-tolerance including application clustering and redundant HBA paths (allowing FC fabric to surviving FC fabric disruptions such as ISL failures and providing redundant paths to storage array controllers).

When choosing a storage solution, customers look for system features and capabilities that can meet the virtualization infrastructure requirements for their specific environment. The following table lists some specific “pain points” and feature requirements that customers might have when choosing a storage solution, and describes how specific SAN storage and VMware Infrastructure capabilities can address those requirements.

Customer “Pain Points”	SAN Solution	VMware Infrastructure 3 Solutions
No centralized management capabilities	Server consolidation provides opportunity for storage consolidation on SAN	Centralized management of ESX hosts and virtual machines using VirtualCenter
Increased I/O loads because of increasing amounts of data	Multipathing, new server and storage deployments	Built-in VMware Infrastructure 3 multipathing, VirtualCenter, DRS
Risk of hardware failure	Multipathing and failover	Built-in VMware Infrastructure 3 multipathing; automatic failover; high availability
Application failure	Application redundancy and clustering	VMware HA, MSCS
Volume security	Volume protection	Virtual SCSI volumes, VMFS
Backup strategies and cost	LAN-free backup	VCB
Data growth management issues	Storage consolidation	VMFS (hot-add, spanning); RDM

Reliability, Availability, and Scalability

Besides required features, performance, and cost, the criteria that typically drive customer choices are the reliability, availability, and scalability of specific storage solutions. SAN solutions are specifically designed to meet these additional criteria and satisfy the requirements of mission-critical business applications. The datacenter, virtualization infrastructure, and storage systems built to run these applications typically handle large volumes of important information and data, must operate reliably and continually, and must also be able to grow to meet increasing business volume, peak traffic, and an expanding number of programs, applications, and users. The key capabilities that SAN solutions provide to meet these requirements include:

- Storage clustering, data sharing, disaster planning, and flexibility of storage planning (central versus distributed)
- Ease of connectivity
- Storage consolidation
- LAN-free backup
- Server-less backup - Network Data Management Protocol (NDMP), disk to tape
- Ease of scalability
  - Storage and server expansion
  - Bandwidth on demand
  - Load balancing

VMware Infrastructure 3 and SAN Solution Support

The capability of the SAN storage solution is only one part of the systems designed to provide enterprise virtualization infrastructure. VMware Infrastructure 3 provides specific features to help deliver reliability, availability, and scalability (RAS) of enterprise systems using SAN storage.

Reliability

The traditional definition of reliability in a SAN means that the system must be fault tolerant during fabric disruptions such as port login and logout anomalies, FC switch failures, or other conditions that causes a RSCN storm. VMware ESX is well suited for error recovery, and guards against I/O subsystem malfunctions that may impact the underlying applications. Because virtual machines are protected from SAN errors by SCSI emulation, the applications they run are also protected from any failure of the physical SAN components.

Reliability in SAN	VMware Infrastructure 3 Solutions
Fabric disruptions	Automatic failover path detection hides complexity of SAN multipathing
Data integrity and performance	VMFS-3 (rescan logic, auto-discovery, hiding SAN errors, distributed journal for faster crash recovery)

Availability

Availability generally refers to the accessibility of a system or application to perform work or perform tasks when requested. For SAN storage, availability means that data must be readily available in the shortest possible time after a SAN error condition. Thus, redundancy is a key factor in providing highly available I/O subsystems. VMware ESX has a built-in multipathing algorithm that automatically detects an error condition and chooses an alternative path to continue servicing data or application requests.

Availability in SAN	VMware Infrastructure 3 Solutions
Link failures	HBA multipathing auto-detects an alternate path
Storage port failures	Storage port multipathing auto-detects alternate storage ports
Dynamic load performance	VMware DRS
Fault-tolerance and disaster recovery	VMware HA
Storage clustering	MSCS support (within local storage; for more information, see the VMware *Setup for Microsoft Cluster Service* documentation available at http://www.vmware.com/support/pubs)
Higher bandwidth	4GFC support
LAN-free backup	VMware Consolidated Backup (VCB)

Scalability

SAN scalability in traditional terms means the ability to grow your storage infrastructure with minimal or no disruption to underlying data services. Similarly, with VMware ESX, growing your virtualization infrastructure means being able to add more virtual machines as workloads increase. Adding virtual machines with VMware ESX is simplified by the use of a template deployment. Adding virtual machines or more storage to existing virtualization infrastructure requires only two simple steps: presenting new volumes to ESX hosts and rescanning the ESX hosts to detect new volumes. With VMware ESX, you can easily scale storage infrastructure to accommodate increased storage workloads.

Scalability in SAN	VMware Infrastructure 3 Solutions
Server expansion	VMware template deployment
Storage expansion	▪ VMFS spanning (32 max) ▪ Rescan to 256 volumes (auto-detect) ▪ Volume hot-add to virtual machines
Storage I/O bandwidth on demand	Fixed policy load-balancing
Heterogeneous environment	Extensive QA testing for heterogeneous support

New VMware Infrastructure Storage Features and Enhancements

This section highlights the major new storage features and enhancements provided by VMware Infrastructure 3. This section also describes differences between VMware Infrastructure 3 and previous versions in the way specific storage features work.

What's New for SAN Deployment in VMware Infrastructure 3?

The following list summarizes new storage features and enhancements added by VMware Infrastructure 3:

- Enhanced support for array-based data replication technologies through the functionality of new logical volume manager (LVM) tools.
- VMFS-3
- NAS and iSCSI support is the first for VMware Infrastructure 3. NFS version 3 is also supported. In addition, the ESX host kernel has a built-in TCP/IP stack optimized for IP storage.
- VMware DRS and VMware HA
- Improved SCSI emulation drivers
- FC-AL support with HBA multipathing
- Heterogeneous array support and 4GFC HBA support

- Storage VMotion
- Node Port ID Virtualization (NPIV)

The following table summarizes the features that VMware Infrastructure 3 provides for each type of storage:

Storage Solution	*HBA Failover*	*SP Failover*	*MSCS Cluster*	*VMotion*	*RDM*	*Boot from SAN*	*VMware HA / DRS*
Fibre Channel	√	√	√	√	√	√	√
NAS	√	√	**No**	√	**No**	√	√
iSCSI (HW)	√	√	**No**	√	√	√	√
iSCSI (SW)	√	√	**No**	√	√	**No**	√

VMFS-3 Enhancements

This section describes changes between VMware Infrastructure 2.X and VMware Infrastructure 3 pertaining to SAN storage support. Understanding these changes helps when you need to modify or update existing infrastructure. The new and updated storage features in VMware Infrastructure 3 provide more built-in support for RAS (reliability, availability, and scalability). Improvements allow an existing virtual infrastructure to grow with higher demand and to service increasing SAN storage workloads. Unlike VMFS-2 that stores virtual machine logs, configuration files (`.vmx` extension), and core files on local storage, virtual machines in VMFS-3 volumes can have all associated files located in directories residing on SAN storage. SAN storage enables the use of large number of files and large data blocks. VMFS-3 is designed to scale better than VMFS-2.

VMFS-3 provides a distributed journaling file system. A journaling file system is a fault-resilient file system that ensures data integrity because all updates to directories and bitmaps are constantly written to a serial log on the disk before the original disk log is updated. In the event of a system failure, a full journaling file system ensures that all the data on the disk has been restored to its pre-crash configuration. VMFS-3 also recovers unsaved data and stores it in the location where it would have gone if the computer had not crashed. Journaling is thus an important feature for mission-critical applications. Other benefits of distributed journaling file system are:

- Provides exclusive repository of virtual machines and virtual machine state
- Provides better organization through directories,
- Stores a large number of files, to host more virtual machines
- Uses stronger consistency mechanisms
- Provides crash recovery and testing of metadata update code in I/O paths
- Provides the ability to hot-add storage volumes

The VMFS-3 Logical Volume Manager (LVM) eliminates the need for various disk modes (public versus shared) required in the older VMware Infrastructure 2 releases. With the VMware Infrastructure 3 LVM, volumes are treated as a dynamic pool of resources.

Benefits of the LVM are:

- It consolidates multiple physical disks into a single logical device.
- Volume availability is not compromised due to missing disks.
- It provides automatic resignaturing for volumes hosted on SAN array snapshots.

The limits of VMFS-3 are:

- Volume size: 2TB per physical extent (PE) , 64TB total
- Maximum number of files: approximately 3840 (soft limit), while the maximum number of files per directory is approximately 220 (soft limit)
- Single access mode: public
- Maximum file size: 768GB
- Single block size: 1MB, 2MB, 4MB, or 8MB

VMFS-3 Performance Improvements

Some of the key changes made to VMware Infrastructure 3 to improve performance are the following:

- Reduced I/O-to-disk for metadata operations
- Less contention on global resources
- Less disruption due to SCSI reservations
- Faster virtual machine management operations

VMFS-3 Scalability

Changes made to VMware Infrastructure 3 that improve infrastructure scalability include:

- Increased number of file system objects that can be handled without compromising performance
- Increased fairness across multiple virtual machines hosted on the same volume

Storage VMotion

According to IDC reports, a few key traditional challenges remain today regarding SAN implementation issues and costs:

- Reducing down time during any SAN disruptions (either planned or unplanned) such as GBIC errors, “path down” conditions due to cable failures, and switch zone merge conflicts can cause loss of productivity, or even worse, affect or lose business data.
- When businesses grow, there is more information to store and the increased storage demands require that an IT shop maintain or increase performance when deploying additional RAID sets.
- During any disaster recovery (DR), ease-of-data-migration, while keeping costs down, remains a challenge in moving data effectively from a disaster area.

Ultimately the evolution of technologies and market environment dictates how customer chooses a design and deploy it for production.

What are the technologies available today?

- Backup and Recovery provides a simple solution from an installation and deployment standpoint. However it may not meet your Recovery Point Objective (RPO), defined as the point in time to which data must be restored, or your Recover Time Objective (RTO), which is defined as the boundary of time and service level with which a process must be accomplished.
- Another solution is data replication, but this solution normally requires duplicate hardware across different geographical locations (either across the street or across cities). In addition, there are additional management costs in setup, maintenance, and troubleshooting.
- For mission-critical applications that have strict RPOs and RTOs (for example, providing recovery within 60 seconds), application clustering provides a solution. However this solution has the highest cost and is typically very complex to build and maintain, as there are various levels of software and hardware compatibility issues to resolve.
- Either existing or current DR solutions being developed take time to implement due to technical challenges, company politics, business resource constraints, and other reasons.

The solution that VMware offers today is Storage VMotion (available with VMware Infrastructure 3). Storage VMotion allows IT administrators to minimize service disruption due to planned storage downtime that would previously be incurred for rebalancing or retiring storage arrays. Storage VMotion simplifies array migration and upgrade tasks and reduces I/O bottlenecks by moving virtual machines to the best available storage resource in your environment. Migrations using Storage VMotion must be administered through the Remote Command Line Interface (Remote CLI).

Storage VMotion is a vendor-agnostic technology that can migrate data across storage tiers. (Tier 1 being the highest cost, highly available, and highest performance storage, Tier 2 having lower RPO and RTO requirements, and Tier 3 being the most cost-effective.)

The key use cases for Storage VMotion are the following:

1. Growing storage requires that the same attributes are maintained in terms of storage performance. With Storage VMotion, data from an active/passive array can be migrated or upgraded to active/active array (for example, moving from storage tier 2 to storage tier 1). This new active/active storage array can provide a better multipathing strategy.
2. Storage VMotion can move data from one RAID set to another newly created RAID set or sets that have better performance, such as those with faster disk drives, higher cache size, and better cache-hit ratio algorithms.
3. Because Storage VMotion utilizes Network File Copy services, data can be moved between protocols such as FC, NFS or iSCSI or SCSI.

Storage VMotion is a very cost-effective and easy-to-use RTO solution, in particular, in operations with planned downtime such as site upgrade and service.

Node Port ID Virtualization (NPIV)

What is N_Port ID Virtualization or NPIV? In the simplest terms, it is allowing multiple N_Port IDs to share a single physical N_Port or an HBA port. With NPIV, a single physical port can function as multiple initiators, each having its own N_Port ID and a unique World Wide Port Name. With unique WWPNs, a single physical port can now belong to many different zones.

Storage or SAN management software applications and operating systems can leverage this NPIV-unique capability to provide fabric segmentation on a finer granularity – at the application level. The operating system is associated or is tracked using WWPN via NPIV. This translates to a single virtual machine being tracked by SAN management applications. For example, the virtual machine's I/O can be monitored per zone on an array port basis. During a VMotion operation, all NPIV resources (such as the WWNN and WWPN of each virtual machine) are also moved from source to destination).

HP was the first vendor to offer an NPIV-based FC interconnect option for blade servers with HP BladeSystem c-Class, which provides a use case for NPIV operating in a production environment.

VMware Storage Architecture

It is important for SAN administrators to understand how VMware Infrastructure's storage components work for effective and efficient management of systems. Storage architects must understand VMware Infrastructure's storage components and architecture so they can best integrate applications and optimize performance. This knowledge also serves as a foundation for troubleshooting storage-related problems when they occur.

Storage Architecture Overview

VMware Infrastructure Storage Architecture serves to provide layers of abstraction that hide and manage the complexity of and differences between physical storage subsystems, and present simple standard storage elements to the virtual environment (see Figure 3-3). To the applications and guest operating systems inside each virtual machine, storage is presented simply as SCSI disks connected to a virtual BusLogic or LSI SCSI HBA.

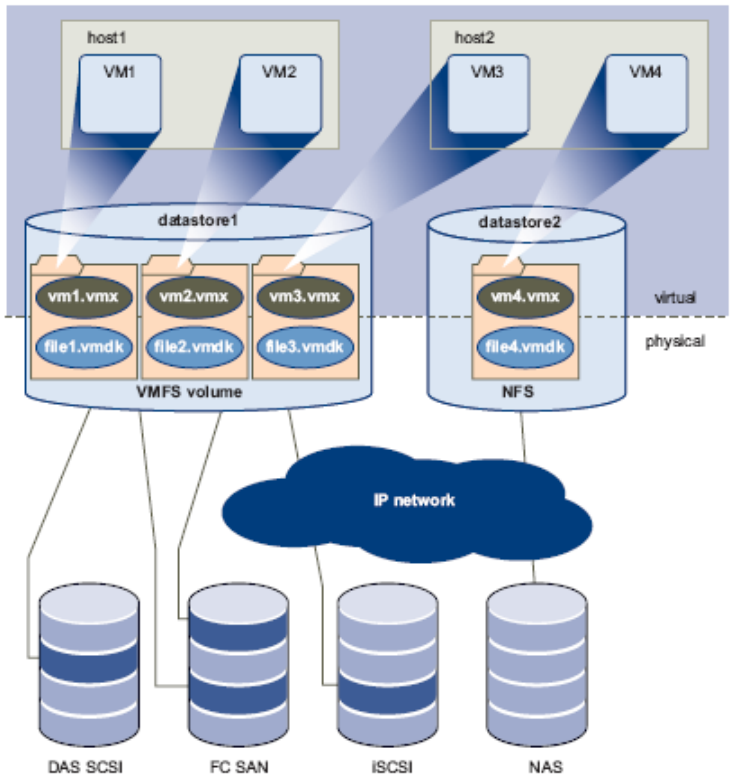


Figure 3-3. VMware Infrastructure Storage Architecture

The virtual SCSI disks inside the virtual machines are provisioned from datastore elements in the datacenter. A datastore is like a storage appliance that serves up storage space for virtual disks inside the virtual machines, and stores the virtual machine definitions themselves. As shown in Figure 3-3, a virtual machine is stored as a set of files in its own directory in the datastore. A virtual disk inside each virtual machine is located on one or more volumes on physical storage and is treated as either a VMFS volume or an RDM volume. A virtual disk can be easily manipulated (copied, moved, back-up, and so on) just like a file. A user can also hot-add virtual disks to a virtual machine without powering it down.

The datastore provides a simple model to allocate storage space for the virtual machines without exposing them to the complexity of the variety of physical storage technologies available, such as Fibre Channel SAN, iSCSI SAN, Direct Attached Storage, and NAS.

A datastore is physically just a VMFS file system volume or an NFS-mounted directory. Each datastore can span multiple physical storage subsystems. As shown in Figure 3-3, a single VMFS volume can contain one or more smaller volumes from a direct-attached SCSI disk array on a physical server, a Fibre Channel SAN disk farm, or iSCSI SAN disk farm. New volumes added to any of the physical storage subsystems are automatically discovered and made available. They can be added to extend a previously created datastore without powering down physical servers or storage subsystems. Conversely, if any of the volumes within a datastore fails or becomes unavailable, only those virtual machines that reside in that volume are affected. All other virtual machines residing in other volumes continue to function as normal.

File System Formats

Datastores that you use can have the following file system formats:

- **VMFS** — VMware ESX deploys this type of file system on local SCSI disks, iSCSI volumes, or Fibre Channel volumes, creating one directory for each virtual machine. VMFS is a clustered file system that can be accessed simultaneously by multiple ESX systems.

  NOTE: ESX 3 supports only VMFS version 3 (VMFS-3); if you are using a VMFS-2 datastore, the datastore will be read-only. For information on upgrading your VMFS-2 datastores, see "Upgrading Datastores" on page 113. VMFS-3 is not backward compatible with versions of VMware ESX earlier than ESX 3.

- **Raw Device Mapping (RDM)** — RDM allows support of existing file systems on a volume. Instead of using the VMFS-based datastore, your virtual machines can have direct access to raw devices using RDM as a proxy. For more information on RDM, see "Raw Device Mapping" on page 49.

- **NFS** — VMware ESX can use a designated NFS volume located on an NFS server. (VMware ESX supports NFS version 3.) VMware ESX mounts the NFS volume, creating one directory for each virtual machine. From the viewpoint of the user on a client computer, the mounted files are indistinguishable from local files.

This document focuses on the first two file system types: VMFS and RDM.

VMFS

VMFS is a clustered file system that leverages shared storage to allow multiple physical servers to read and write to the same storage simultaneously. VMFS provides on-disk distributed locking to ensure that the same virtual machine is not powered on by multiple servers at the same time. If a physical server fails, the on-disk lock for each virtual machine can be released so that virtual machines can be restarted on other physical servers.

VMFS also features enterprise-class crash consistency and recovery mechanisms, such as distributed journaling, crash-consistent virtual machine I/O paths, and machine state snapshots. These mechanisms can aide quick root-cause analysis and recovery from virtual machine, physical server, and storage subsystem failures.

Raw Device Mapping

VMFS also supports Raw Device Mapping (RDM). RDM provides a mechanism for a virtual machine to have direct access to a volume on the physical storage subsystem (with Fibre Channel or iSCSI only). As an example, RDM can be used to support the following two applications:

- SAN array snapshot or other layered applications that run in the virtual machines. RDM improves the scalability of backup offloading systems, using features inherent to the SAN.

- Any use of Microsoft Clustering Services (MSCS) that span physical servers, including virtual-to-virtual clusters and physical-to-virtual clusters. Cluster data and quorum disks should be configured as RDMs rather than as individual files on a shared VMFS.

For more information, see the VMware *Setup for Microsoft Cluster Service* documentation available at:

http://www.vmware.com/support/pubs

An RDM can be thought of as providing a symbolic link from a VMFS volume to a raw volume (see Figure 3-4). The mapping makes volumes appear as files in a VMFS volume. The mapping file—not the raw volume—is referenced in the virtual machine configuration.

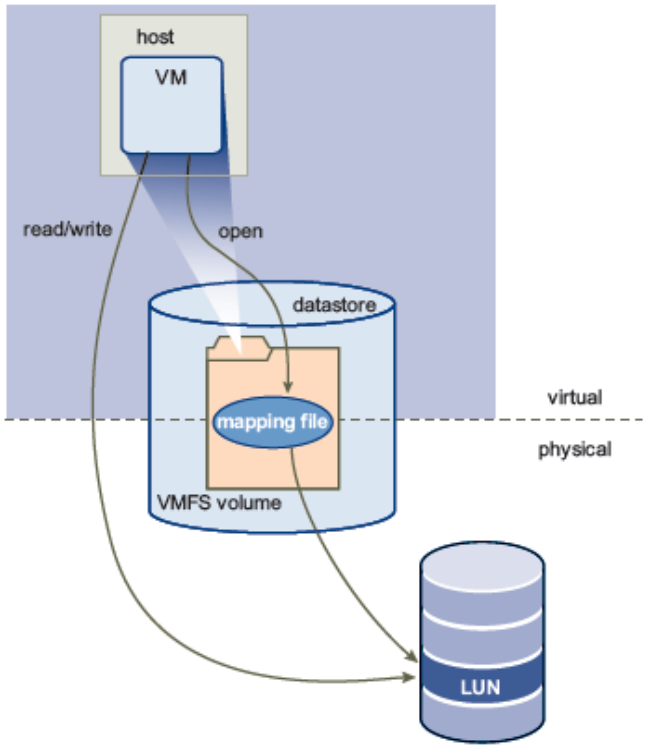


Figure 3-4. VMware Raw Device Mapping

When a volume is opened for access, VMFS resolves the RDM file to the correct physical device and performs appropriate access checking and locking before accessing the volume. Thereafter, reads and writes go directly to the raw volume rather than going through the mapping file.

NOTE: For more details about RDM operation with VMware Infrastructure, see "More about Raw Device Mapping" later in this chapter. Also, see the section "Data Access: VMFS or RDM" in Chapter 4 for considerations, benefits, and limitations on using RDM with VMware Infrastructure.

VMware ESX Storage Components

This section provides a more detailed technical description of internal ESX components and their operation. Figure 3-5 provides a more detailed view of the ESX architecture and specific components that perform VMware storage operations.

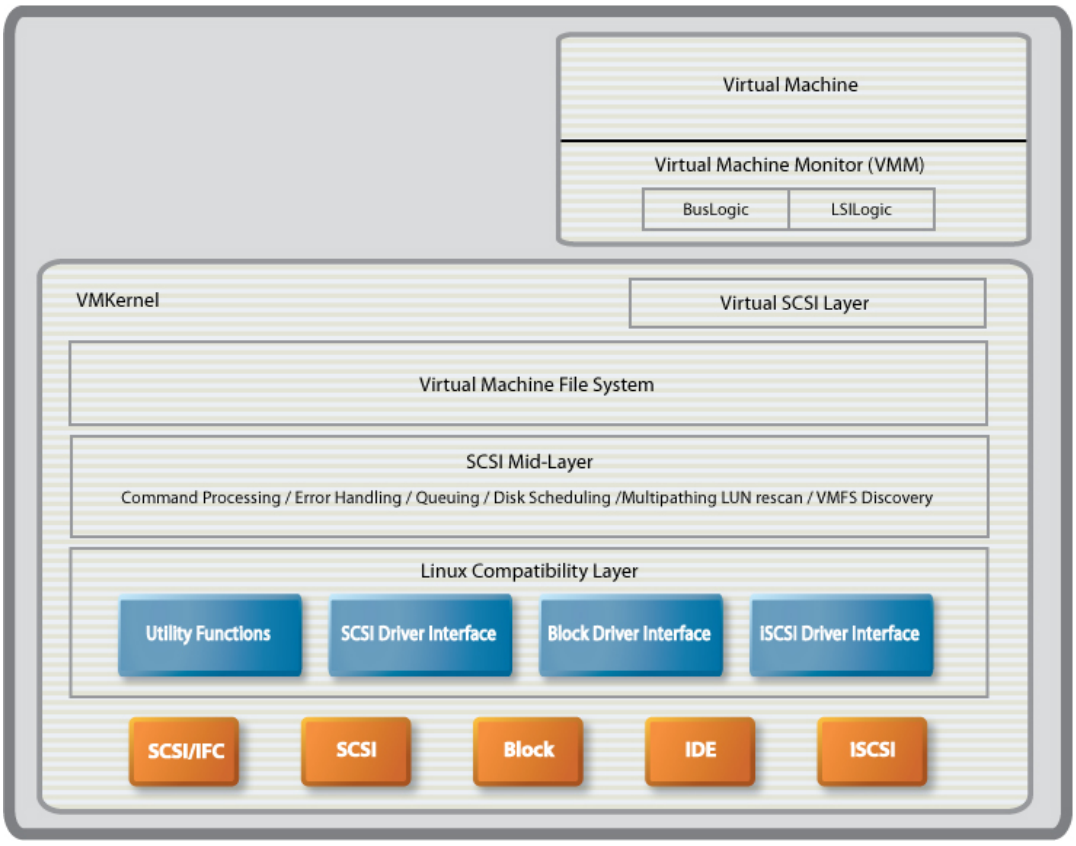


Figure 3-5. Storage Architecture Components

The key components shown in this figure of the storage architecture are the following:

- Virtual Machine Monitor (VMM)
- Virtual SCSI Layer
- VMware File System (VMFS)
- SCSI Mid-Layer
- HBA Device Drivers

Virtual Machine Monitor

The Virtual Machine Monitor (VMM) module's primary responsibility is to monitor a virtual machine's activities at all levels (CPU, memory, I/O, and other guest operating system functions and interactions with VMkernel). The VMM module contains a layer that emulates SCSI devices within a virtual machine. A virtual

machine operating system does not have direct access to Fibre Channel devices because VMware Infrastructure virtualizes storage and presents only a SCSI interface to the operating system. Thus, from any type of virtual machine (regardless of operating system), applications only access storage subsystems only via a SCSI driver. Virtual machines can use either BusLogic or LSI Logic SCSI drivers. These SCSI drivers enable the use of virtual SCSI HBAs within a virtual machine.

Within a Windows virtual machine, under the Windows control panel display for **Computer Management > Device Manager > SCSI and RAID Controllers**, there are listings for **BusLogic** or **LSI Logic** drivers. **BusLogic** indicates that BusLogic BT-958 emulation is being used. BT-958 is a SCSI-3 protocol providing Ultra SCSI (Fast-40) transfer rates of 40MB per second. The driver emulation supports the capability of "SCSI Configured AutoMatically," also known as SCAM, which allows SCSI devices to be configured with an ID number automatically, so you do not have to assign IDs manually.

LSI Logic indicates that the LSI53C1030 Ultra-320 SCSI emulation is being used. In addition to the benefits of supporting Ultra320 technology (including low voltage differential, SCSI domain validation in SPI-4 specification, PCI-X compliant, and better cyclical redundancy check), the LSI53C1030 emulation also provides TolerANT technology benefits—primarily better signal noise tolerance. Other benefits include the use of active negation on SCSI drivers and input signal filtering on SCSI receivers to improve data integrity. All of these design benefits are well suited for applications located on a SAN environment that might have to endure fabric problems and other changes (such as cabling failures, fabric merges, or zone conflicts that would cause SAN disruptions). Another important key benefit of LSI53C1030 is the underlying support of Fusion Message Passing Technology (commonly known as Fusion-MPT) architecture. Providing an efficiency mechanism to host processors, Fusion-MPT architecture enables I/O controllers to send multiple reply messages in a single interrupt to the host processor to reduce context switching. This results in transfer rates up to 100,000 Ultra320 SCSI IOPS, with minimal system overhead or device intervention.

Virtual SCSI HBAs allow virtual machines access to logical SCSI devices, just as a physical HBA allows access to physical storage devices. However, in contrast to a physical HBA, the virtual SCSI HBA does not allow storage administrators (such as SAN administrators) access to the physical machine.

In an ESX environment, each virtual machine includes from one to four virtual SCSI HBAs. These virtual adapters may appear as either BusLogic or LSI Logic SCSI controllers. These two types are the only SCSI controllers accessible by virtual machines.

Virtual SCSI Layer

The virtual SCSI layer's primary responsibility is to manage SCSI commands and intercommunication between the VMM, the VMFS, and SCSI mid-layer below. All SCSI commands from virtual machines must go through the virtual SCSI layer. Also, I/O abort and reset operations are managed at this layer. From here, the virtual SCSI layer passes I/O or SCSI commands from virtual machines to lower layers, either via VMFS or RDM (which supports two modes: pass-through and non-pass-through). In RDM pass-through mode, all SCSI commands are allowed to pass through without traps.

The VMware File System

VMFS is proprietary to VMware ESX and is optimized for storing and accessing large files. The use of large block sizes keeps virtual machine disk performance close to that of native SCSI disks.

The size of VMFS-3 metadata on a VMFS-3 volume with on-disk version 3.31 or prior will be no more than 1200MB. A more exact calculation of VMFS metadata storage space requirements needs to take into consideration factors such as the size of LVM metadata, VMFS major version, and size of VMFS system files. You may also want to contact the VMware PSO organization to assist you with storage planning and determining more exact VMFS metadata requirements for your environment.

VMFS is well suited for SAN storage because of the built-in logic for rescan that detects changes in LUNs automatically. Another key benefit of VMFS is that it further hides the complexity of storage on SAN by hiding SAN errors from virtual machines.

The most unique feature of VMFS is that, as a clustered file system, it leverages shared storage to allow multiple physical servers to read and write to the same storage simultaneously. VMFS provides on-disk distributed locking (using volume SCSI-2 reservations) to ensure that the same virtual machine is not powered on by multiple servers at the same time. If a physical server fails, the on-disk lock for each virtual machine can be released so that virtual machines can be restarted on other physical servers. VMFS also features enterprise-class crash consistency and recovery mechanisms, such as distributed journaling, crash-consistent virtual machine I/O path, and machine state snapshots. These mechanisms can aid quick root-cause analysis and recovery from virtual machine, physical server, and storage subsystem failures.

In a simple configuration, the virtual machines' disks are stored as files within a VMFS. When guest operating systems issue SCSI commands to their virtual disks, the virtualization layer translates these commands to VMFS file operations. ESX systems also use VMFS to store virtual machine files. To minimize disk I/O overhead, VMFS has been optimized to run multiple virtual machines as one workload.

VMFS is first configured as part of the ESX installation. When you create a new VMFS-3 volume, it must be 1.1 GB or larger. Details on VMFS configuration are provided in the VMware *Installation and Upgrade Guide* as well as the *Server Configuration Guide*.

A VMFS volume can be extended over 32 physical storage extents, including SAN volumes and local storage. This allows pooling of storage and flexibility in creating the storage volumes necessary for your virtual machine. With the new ESX 3 LVM, you can extend a volume while virtual machines are running on the volume. This lets you add new space to your VMFS volumes as your virtual machine needs it.

SCSI Mid-Layer

The SCSI mid-layer is the most important layer in VMkernel for storage activities, managing physical HBAs on ESX hosts, queuing requests, and handling SCSI errors. In addition, this layer contains automatic rescan logic that detects changes to LUN mapping assigned to an ESX host. Path management such as automatic path selection, path collapsing, failover and failback to specific volumes are also handled in the SCSI mid-layer.

The SCSI mid-layer gathers information from HBAs, switches, and storage port processors to identify path structures between the ESX host and the physical volume on storage arrays. During a rescan, VMware ESX looks for device information such as the network address authority (NAA) identifier, and serial number. VMware ESX identifies all available paths to a storage array and collapses it to one single active path (regardless of how many paths are available). All other available paths are marked as standby. Path change detection is automatic. Depending on the storage device response to the TEST_UNIT_READY SCSI command, VMware ESX marks the path as on, active, standby, or dead.

During boot up or a rescan operation, VMware ESX automatically assigns a path policy of Fixed for all active/active storage array types. With a Fixed path policy, the preferred path is selected if that path is in the on state. For active/active storage array types, VMware ESX performs a path failover only if a SCSI I/O request fails with a FC driver status of NO_CONNECT, which indicates a loss of FC connectivity. Commands that fail with check conditions are returned to the guest operating system. When a path failover is completed, VMware ESX issues the command to the next path that is in the on state.

For active/passive storage array types, VMware ESX automatically assigns a path policy of MRU (Most Recently Used). A device response to TEST_UNIT_READY of NO_CONNECT and specific SCSI check conditions triggers VMware ESX to test all available paths to see if they are in the on state.

NOTE: For active/passive storage arrays that are not on the VMware SAN Compatibility list, manually changing an active/passive array to use the MRU policy is not sufficient to make the array be fully interoperable with VMware ESX. Any new storage arrays must be approved by VMware and be listed in the VMware *SAN Compatibility Guide*.

VMware ESX multipathing software does not actively signal virtual machines to abort I/O requests. If the multipathing mechanism detects that the current path is no longer operational, VMware ESX initiates a process to activate another path to the volume and re-issues the virtual machine I/O request to the new path (instead of immediately returning the I/O failure to the virtual machine). There can be some delay in completing the I/O request for the virtual machine. This is the case if the process of making another path operational involves issuing SCSI command to the standby storage processor on an active/passive array. During this process of path failover, I/O requests to the individual volume are queued. If a virtual machine is issuing synchronous I/O requests to the volume at the same time, the virtual machine appears to be stalled temporarily. If the virtual machine is not issuing synchronous I/O to this volume, it continues to run. Thus, it is recommended that you set the virtual machine **Disk TimeOutValue** setting to at least 60 seconds to allow SCSI devices and path selection time to stabilize during a physical path disruption.

Host Bus Adapter Device Drivers

The only means by which virtual machines can access storage on a SAN is through a FC HBA. VMware provides modified standard Linux HBA device drivers to work with the VMware SCSI mid-layer. VMware's modified HBA drivers are loaded automatically during ESX installation. There is a tight interoperability relationship between FC HBAs and SAN storage arrays. Therefore SAN components such as HBAs and storage arrays must be certified by VMware or at an OEM partner site. Test programs are

designed to check compatibility and interoperability between VMware ESX HBA device drivers and SAN equipment under test with different load and stress conditions. Before deploying any storage components on ESX hosts, you should review the VMware-supported storage components listed in the VMware *SAN Hardware Compatibility Guide* and the *I/O Compatibility Guide*. Device drivers not included on these lists are not supported.

Another HBA component that requires testing to be certified with storage arrays is the boot BIOS available from Emulex or Qlogic. Boot BIOS versions are usually not listed separately, but are listed as supported HBA models in the VMware *I/O Compatibility Guide*. Using the boot BIOS functionality, ESX hosts can be booted from SAN.

VMware Infrastructure Storage Operations

This section reviews VMware storage components and provides additional details of VMware Infrastructure operations using these storage components. In the most common configuration, a virtual machine uses a virtual hard disk to store its operating system, program files, and other data associated with its activities. A virtual disk is a large physical file that can be copied, moved, archived, and backed up as easily as any other file.

Virtual disk files reside on specially formatted volumes called datastores. A datastore can be deployed on the host machine's internal direct-attached storage devices or on networked storage devices. A networked storage device represents an external shared storage device or array that is located outside of your system and is typically accessed over a network through an adapter.

Storing virtual disks and other essential elements of your virtual machine on a single datastore shared between physical hosts lets you:

- Use such features as VMware DRS (Distributed Resource Scheduling) and VMware HA (High Availability) options.
- Use VMotion to move running virtual machines from one ESX system to another without service interruption.
- Use VMware Consolidated Backup to perform backups more efficiently.
- Have better protection from planned or unplanned server outages.
- Have more control over load balancing.

VMware ESX lets you access a variety of physical storage devices (both internal and external), configure and format them, and use them for your storage needs.

Datastores and File Systems

ESX virtual machines store their virtual disk files on specially formatted logical containers, or **datastores**, which can exist on different types of physical storage devices. A datastore can use disk space on one physical device or several physical devices.

The datastore management process starts with storage space that your storage administrator preallocates for ESX systems on different storage devices. The storage

space is presented to your ESX system as volumes with logical unit numbers or, in the case of a network-attached storage, as NFS volumes.

Using the VI Client, you can create datastores by accessing and formatting available volumes or by mounting the NFS volumes. After you create the datastores, you can use them to store virtual machine files. When needed, you can modify the datastores, for example, to add, rename, or remove extents in the datastores.

Types of Storage

Datastores can reside on a variety of storage devices. You can deploy a datastore on your system's direct-attached storage device or on a networked storage device. VMware ESX supports the following types of storage devices:

- **Local** — Stores files locally on an internal or external SCSI device.
- **Fibre Channel** — Stores files remotely on a SAN. Requires FC adapters.
- **iSCSI (hardware initiated)** — Stores files on remote iSCSI storage devices. Files are accessed over TCP/IP network using hardware-based iSCSI HBAs.
- **iSCSI (software initiated)** — Stores files on remote iSCSI storage devices. Files are accessed over TCP/IP network using software-based iSCSI code in the VMkernel. Requires a standard network adapter for network connectivity.
- **Network File System (NFS)** — Stores files on remote file servers. Files are accessed over TCP/IP network using the NFS protocol. Requires a standard network adapter for network connectivity. VMware ESX supports NFS version 3.

You use the VI Client to access storage devices mapped to your ESX system and deploy datastores on them. For more information, see "Chapter 6, Managing VMware Infrastructure 3 with SAN."

Available Disk Configurations

Virtual machines can be configured with multiple virtual SCSI drives, although the guest operating system may place limitations on the total number of SCSI drives allowed. Although all SCSI devices are presented as SCSI targets, there are three physical implementation alternatives:

- A `.vmdk` file stored on a VMFS volume. See "Storage Architecture Overview" on page 47.
- Device mapping to a volume. Device mappings can be to SAN volumes, local SCSI, or iSCSI volumes. See "Raw Device Mapping" on page 49.
- Local SCSI device passed through directly to the virtual machine (for example, a local tape drive).

From the standpoint of the virtual machine, each virtual disk appears as if it were a SCSI drive connected to a SCSI adapter. Whether the actual physical disk device is being accessed through SCSI, iSCSI, RAID, NFS, or FC controllers is transparent to the guest operating system and to applications running on the virtual machine.

How Virtual Machines Access Storage

When a virtual machine accesses a datastore, it issues SCSI commands to the virtual disk. Because datastores can exist on various types of physical storage, these commands are packetized into other forms, depending on the protocol the ESX system uses to connect to the associated storage device. VMware ESX supports FC, iSCSI, and NFS protocols.

Figure 3-6 shows five virtual machines (each using a different type of storage) to illustrate the differences between each type.

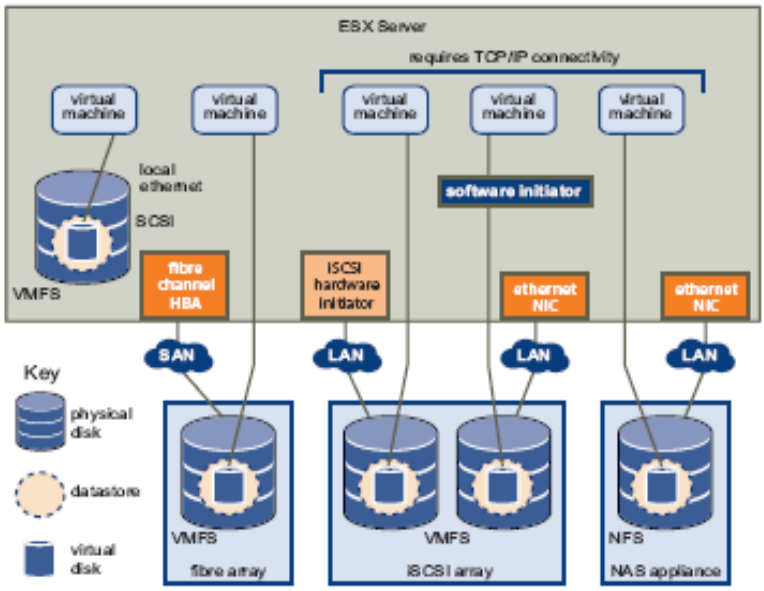


Figure 3-6. Types of Storage

You configure individual virtual machines to access the virtual disks on the physical storage devices. Virtual machines access data using VMFS or RDM.

- **VMFS** — In a simple configuration, the virtual machines' disks are stored as `.vmdk` files within an ESX VMFS. When guest operating systems issue SCSI commands to their virtual disks, the virtualization layer translates these commands to VMFS file operations.

  In a default setup, the virtual machine always goes through VMFS when it accesses a file, be it on a SAN or a host's local hard drives. See "Storage Architecture Overview" on page 47.

- **RDM** — A mapping file inside the VMFS acts as a proxy to give the guest operating system access to the raw device. See "Raw Device Mapping" on page 49.

  RDM is recommended when a virtual machine must interact with a real disk on the SAN. This is the case, for example, when you make disk array snapshots or, more rarely, if you have a large amount of data that you do not want to move onto a virtual disk. It is also required for Microsoft Cluster Service setup. See the VMware document *Setup for Microsoft Cluster Service* for more information.

Sharing a VMFS across ESX Hosts

VMFS is designed for concurrent access and enforces the appropriate controls for access from multiple ESX hosts and virtual machines. VMFS can

- **Coordinate access to virtual disk files** — VMware ESX uses file-level locks, which are managed by the VMFS distributed lock manager.
- **Coordinate access to VMFS internal file system information (metadata)** — VMware ESX uses SCSI reservations on the entire volume. See "Metadata Updates," below.

NOTE: SCSI reservations are not held during metadata updates to the VMFS volume. VMware ESX uses short-lived SCSI reservations as part of its distributed locking protocol.

The fact that virtual machines share a common VMFS makes it more difficult to characterize peak-access periods or optimize performance. You need to plan virtual machine storage access for peak periods, but different applications might have different peak-access periods. Increasing the number of virtual machines that share a VMFS increases the potential for performance degradation due to I/O contention.

VMware recommends that you load balance virtual machines and applications over the combined collection of servers, CPU, and storage resources in your datacenter. That means you should run a mix of virtual machines on each server in your datacenter so that not all servers experience high demand at the same time.

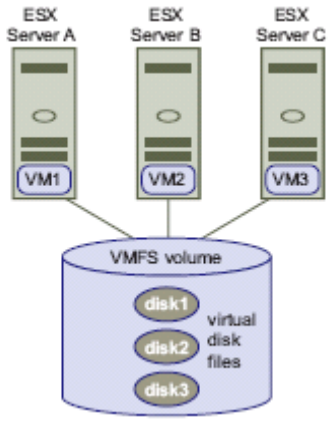


Figure 3-7. Accessing Virtual Disk Files

Metadata Updates

A VMFS holds a collection of files, directories, symbolic links, RDMs, and other data elements, and also stores the corresponding metadata for these objects. Metadata is accessed each time the attributes of a file are accessed or modified. These operations include, but are not limited to:

- Creating, growing, or locking a file.
- Changing a file's attributes.
- Powering a virtual machine on or off.

Access Control on ESX Hosts

Access control allows you to limit the number of ESX hosts (or other hosts) that can see a volume. Access control can be useful to:

- Reduce the number of volumes presented to an ESX system.
- Prevent non-ESX systems from seeing ESX volumes and from possibly destroying VMFS volumes.

For more information on LUN masking operations, see "Masking Volumes Using Disk.MaskLUN" on page 118.

The LUN masking option provided in ESX hosts is useful in masking LUNs that are meant to be hidden from hosts or in masking LUNs in a SAN management array that are not readily available. Suppose, for example, that a volume with LUN 9 was originally mapped and recognized by an ESX host. This volume was then chosen to store critical data. After finishing the deployment of virtual machines from this volume, an ESX administrator could mask LUN 9 so that no one could accidentally destroy the datastore located on the volume associated with LUN 9. To simplify operations, masking this LUN or preventing access to this volume from the ESX host does not require a SAN administrator to change anything on the storage management agent.

More about Raw Device Mapping

RDM files contain metadata used to manage and redirect disk accesses to the physical device. RDM provides the advantages of direct access to a physical device while keeping some advantages of a virtual disk in the VMFS file system. In effect, it merges VMFS manageability with raw device access.

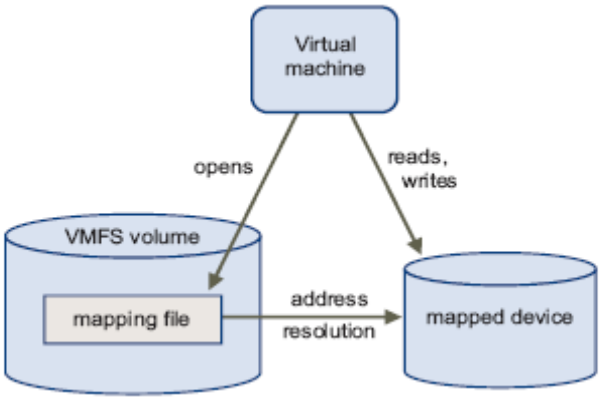


Figure 3-8. Raw Device Mapping Redirects Data Transfers

While VMFS is recommended for most virtual disk storage, sometimes you need raw disks. The most common use is as data drives for Microsoft Cluster Service (MSCS) configurations using clusters between virtual machines or between physical and virtual machines.

NOTE: For more information on MSCS configurations supported with VMware Infrastructure, see the VMware *Setup for Microsoft Cluster Service* documentation available at http://www.vmware.com/support/pubs.

Think of an RDM as a symbolic link from a VMFS volume to a raw volume. The mapping makes volumes appear as files in a VMFS volume. The mapping file—not the raw volume—is referenced in the virtual machine configuration. The mapping file contains a reference to the raw volume. Using RDMs, you can:

- Use VMotion to migrate virtual machines using raw volumes.
- Add raw volumes to virtual machines using the VI Client.
- Use file system features such as distributed file locking, permissions, and naming.

Two compatibility modes are available for RDMs:

- Virtual compatibility mode allows a mapping to act exactly like a virtual disk file, including the use of storage array snapshots.
- Physical compatibility mode allows direct access of the SCSI device, for those applications needing lower level control.

VMware VMotion, VMware DRS, and VMware HA are all supported in both RDM physical and virtual compatibility modes.

RDM Characteristics

An RDM file is a special file in a VMFS volume that manages metadata for its mapped device. The mapping file is presented to the management software as an ordinary disk file, available for the usual file system operations. To the virtual machine, the storage virtualization layer presents the mapped device as a virtual SCSI device. Key contents of the metadata in the mapping file include the location of the mapped device (name resolution) and the locking state of the mapped device.

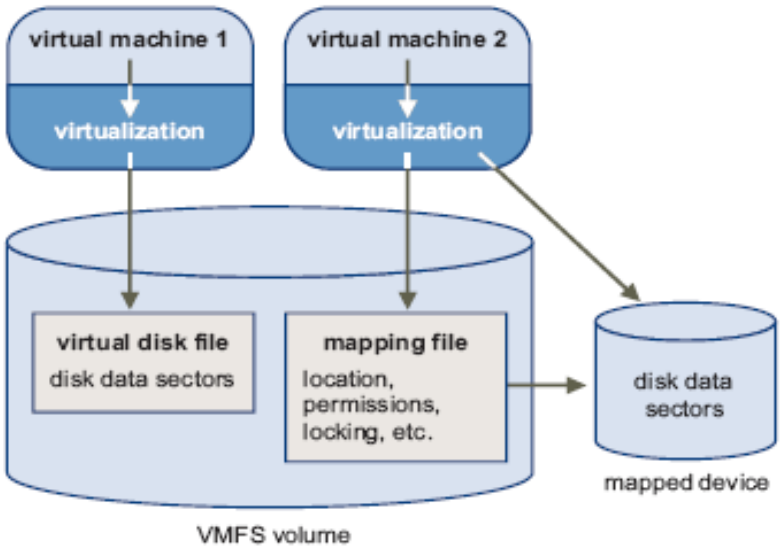


Figure 3-9. Mapping File Metadata

Virtual and Physical Compatibility Modes

Virtual mode for a mapping specifies full virtualization of the mapped device. It appears to the guest operating system exactly the same as a virtual disk file in a VMFS volume. The real hardware characteristics are hidden. Virtual mode allows customers using raw disks to realize the benefits of VMFS, such as advanced file locking for data protection and snapshots for streamlining development processes. Virtual mode is also more portable across storage hardware than physical mode, presenting the same behavior as a virtual disk file.

Physical mode for a raw device mapping specifies minimal SCSI virtualization of the mapped device, allowing the greatest flexibility for SAN management software. In physical mode, VMkernel passes all SCSI commands to the device, with one exception: the REPORT LUN command is virtualized, so that VMkernel can isolate the volume for the owning virtual machine. Otherwise, all physical characteristics of the underlying hardware are exposed. Physical mode is useful to run SAN management agents or other SCSI target-based software in the virtual machine. Physical mode also allows virtual to physical clustering for cost-effective high availability.

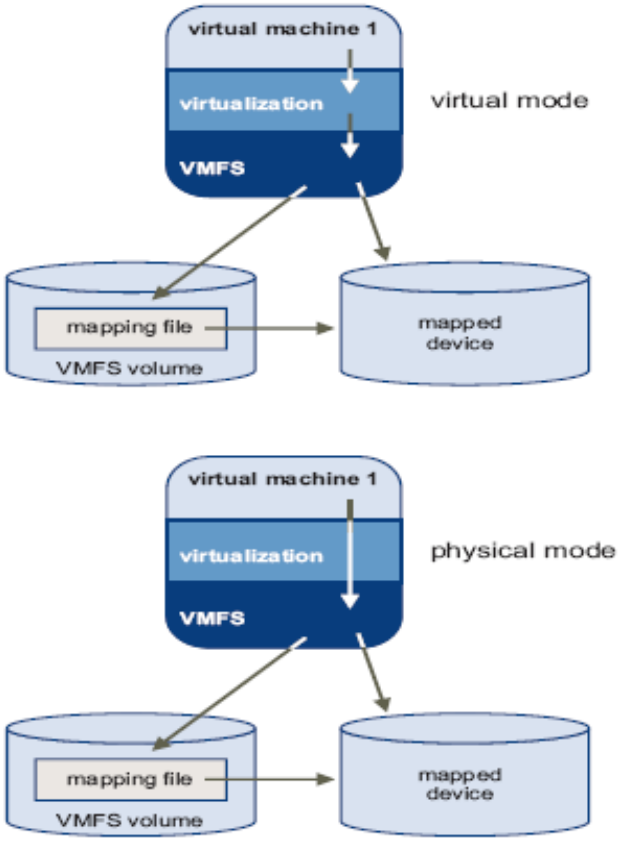


Figure 3-10. Virtual and Physical Compatibility Modes

Dynamic Name Resolution

Raw device mapping lets you give a permanent name to a device by referring to the name of the mapping file in the `/vmfs` subtree.

The example in Figure 3-11 shows three volumes. Volume 1 is accessed by its device name, which is relative to the first visible volume. Volume 2 is a mapped device, managed by a mapping file on volume 3. The mapping file is accessed by its path name in the `/vmfs` subtree, which is fixed.

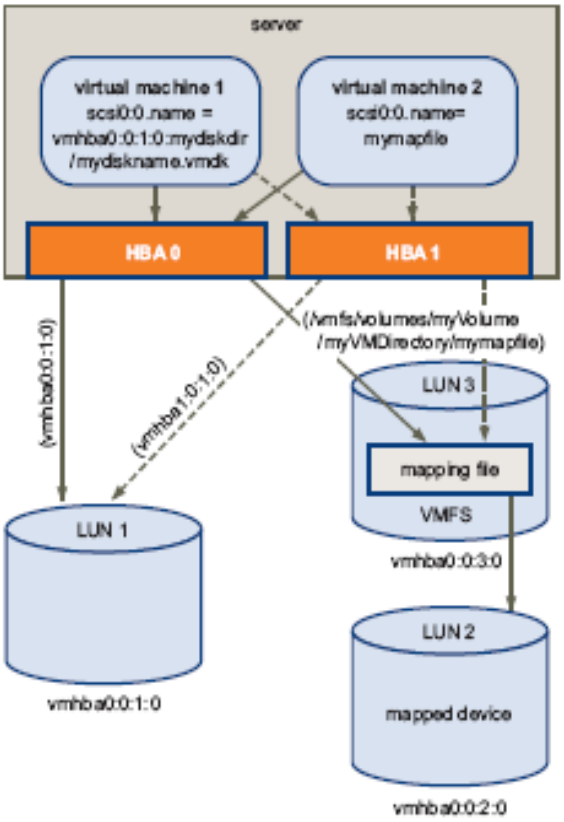


Figure 3-11. Example of Name Resolution

All mapped volumes with LUN 1, 2, and 3 are uniquely identified by VMFS, and the identification is stored in its internal data structures. Any change in the SCSI path, such as an FC switch failure or the addition of a new HBA, has the potential to change the vmhba device name, because the name includes the path designation (initiator, target, and LUN). Dynamic name resolution compensates for these changes by adjusting the data structures to retarget volumes to their new device names.

Raw Device Mapping with Virtual Machine Clusters

VMware recommends the use of RDM with virtual machine clusters that need to access the same raw volume for failover scenarios. The setup is similar to that of a virtual machine cluster that accesses the same virtual disk file, but an RDM file replaces the virtual disk file.

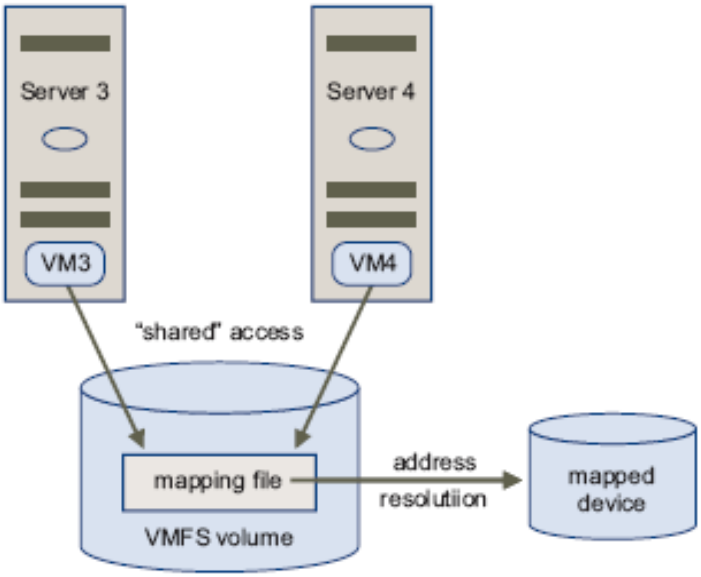


Figure 3-12. Access from Clustered Virtual Machines

For more information on configuring clustering, see the VMware *VirtualCenter Virtual Machine Clustering* manual.

How Virtual Machines Access Data on a SAN

A virtual machine interacts with a SAN as follows:

1. When the guest operating system in a virtual machine needs to read or write to SCSI disk, it issues SCSI commands to the virtual disk.
2. Device drivers in the virtual machine's operating system communicate with the virtual SCSI controllers. VMware ESX supports two types of virtual SCSI controllers: BusLogic and LSI Logic.
3. The virtual SCSI controller forwards the command to VMkernel.
4. VMkernel performs the following operations:
   - Locates the file in the VMFS volume that corresponds to the guest virtual machine disk.
   - Maps the requests for the blocks on the virtual disk to blocks on the appropriate physical device.
   - Sends the modified I/O request from the device driver in the VMkernel to the physical HBA (host HBA).
5. The host HBA performs the following operations:
   - Converts the request from its binary data form to the optical form required for transmission on the fiber optic cable.
   - Packages the request according to the rules of the FC protocol.
   - Transmits the request to the SAN.
6. Depending on which port the HBA uses to connect to the fabric, one of the SAN switches receives the request and routes it to the storage device that the host wants to access.

   From the host's perspective, this storage device appears to be a specific disk, but it might be a logical device that corresponds to a physical device on the SAN. The switch must determine which physical device has been made available to the host for its targeted logical device.

Volume Display and Rescan

A SAN is dynamic, so the volumes that are available to a certain host can change based on a number of factors including:

- New volumes created on the SAN storage arrays
- Changes to LUN masking
- Changes in SAN connectivity or other aspects of the SAN

VMkernel discovers volumes when it boots; and those volumes may then be viewed in the VI Client. If changes are made to the LUN identification of volumes, you must rescan to see those changes. During a rescan operation, VMware ESX automatically assigns a path policy of Fixed for all active/active storage array types. For active/passive storage array types, VMware ESX automatically assigns a path policy of MRU (Most Recently Used).

NOTE: Rescans can be performed to locate new storage device and VMFS volume targets or go to existing targets. See information on performing rescans in “Performing a Rescan of Available SAN Storage Devices” on page 116. Also see the VMware *Server Configuration Guide*.

The best time to rescan ESX hosts is when there is a minimal amount of I/O traffic on the incoming and outgoing SAN fabric ports between an ESX host and the array storage port processors. (The levels of I/O traffic vary by environment.) To determine a minimum and maximum level of I/O traffic for your environment, you need to first establish a record or baseline of I/O activity for your environment. Do this by recording I/O traffic patterns in the SAN fabric ports (for example, using a command such as `portperfshow` for Brocade switches). Once you have determined that I/O traffic has dropped to 20 percent of available port bandwidth, for example, by measuring traffic on the SAN fabric port where a HBA from an ESX host is connected, you can rescan the ESX host with minimal interruptions to running virtual machines.

Zoning and VMware ESX

Zoning provides access control in the SAN topology and defines the HBAs that can connect to specific storage processors or SPs. (See “Zoning” on page 31 for a description of SAN zoning features.) When a SAN is configured using zoning, the devices outside a zone are not visible to the devices inside the zone. Zoning also has the following effects:

- Reduces the number of targets and LUNs presented to an ESX host.
- Controls and isolates paths within a fabric.
- Can prevent non-ESX systems from seeing a particular storage system and from possibly destroying ESX VMFS data.
- Can be used to separate different environments (for example, a test from a production environment).

VMware recommends you use zoning with care. If you have a large deployment, you might decide to create separate zones for different company operations (for example to separate accounting from human resources). However, creating too many small zones (for example, zones including very small numbers of either hosts or virtual machines) may also not be the best strategy. Too many small zones

- Can lead to longer times for SAN fabrics to merge.
- May make infrastructure more prone to SAN administrator errors.
- Exceed the maximum size that a single FC SAN switch can hold in its cache memory.
- Create more chances for zone conflicts to occur during SAN fabric merging.

NOTE: For other zoning best practices, check with the specific vendor of the storage array you plan to use.

Third-Party Management Applications

Most SAN hardware is packaged with SAN management software. This software typically runs on the storage array or on a single server, independent of the servers that use the SAN for storage. This third-party management software can be used for a number of tasks:

- Managing storage arrays, including volume creation, array cache management, LUN mapping, and volume security.
- Setting up replication, checkpoints, snapshots, and mirroring.

If you decide to run the SAN management software inside a virtual machine, you reap the benefits of running an application on a virtual machine (failover using VMotion, failover using VMware HA, and so on). Because of the additional level of indirection, however, the management software might not be able to see the SAN. This can be resolved by using an RDM. See "Managing Raw Device Mappings" on page 107 for more information.

NOTE: Whether or not a virtual machine can run management software successfully depends on the specific storage array you are using.

Using ESX Boot from SAN

When you have SAN storage configured with an ESX host, you can place the ESX boot image on one of the volumes on the SAN. You may want to use ESX boot from SAN in the following situations:

- When you do not want to handle maintenance of local storage.
- When you need easy cloning of service consoles.
- In diskless hardware configurations, such as on some Blade systems.

You should not use boot from SAN in the following situations:

- When you are using Microsoft Cluster Service.
- When there is a risk of I/O contention between the service console and VMkernel.

NOTE: With ESX 2.5, you could not use boot from SAN together with RDM. With ESX 3, this restriction has been removed.

How ESX Boot from SAN Works

When you have set up your system to use boot from SAN, the boot image is not stored on the ESX system's local disk, but instead on a SAN volume.

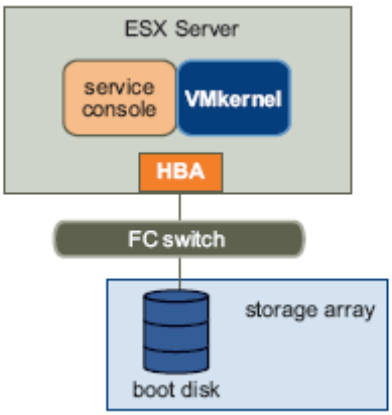


Figure 3-13. How ESX Boot from SAN Works

On a system set up to boot from SAN:

- The HBA BIOS must designate the FC card as the boot controller. See “Setting Up the FC HBA for Boot from SAN” in the VMware *SAN Configuration Guide* for specific instructions.
- The FC card must be configured to initiate a primitive connection to the target boot LUN.

Benefits of ESX Boot from SAN

In a boot from SAN environment, the operating system is installed on one or more volumes in the SAN array. The servers are then informed about the boot image location. When the servers are started, they boot from the volumes on the SAN array.

NOTE: When you use boot from SAN in conjunction with a VMware ESX system, each server must have its own boot volume.

Booting from a SAN provides numerous benefits including:

- **Cheaper servers** — Servers can be more dense and run cooler without internal storage.
- **Easier server replacement** — You can replace servers and have the new server point to the old boot location.
- **Less wasted space**.
- **Easier backup processes** — You can back up the system boot images in the SAN as part of the overall SAN backup procedures.
- **Improved management** — Creating and managing the operating system image is easier and more efficient.

Systems must meet specific criteria to support booting from SAN. See “ESX Boot from SAN Requirements” on page 93 for more information on setting up the boot from SAN option. Also see the VMware *SAN Configuration Guide* for specific installation instructions and tasks to set up the ESX boot from SAN option.

Frequently Asked Questions

Below are some commonly asked questions involving VMware Infrastructure, ESX configuration, and SAN storage. The answers to these questions can help you with deployment strategies and troubleshooting.

Do HBA drivers retry failed commands?

In general, HBA drivers do not retry failed commands. There can be specific circumstances, such as when a driver is attempting to detect FC port failure, under which an HBA driver does retry a command. But it depends on the type of HBA and the specific version of the driver.

What is ESX SCSI I/O timeout?

VMware ESX does not have a specific timeout time for I/O operations issued by virtual machines. The virtual machine itself controls the timeout. ESX-generated I/O requests, such as file system metadata, have a 40-second timeout. Any synchronous VMkernel internal SCSI command, such as a TUR or START_UNIT, has a 40-second timeout.

What happens during a rescan?

VMware ESX issues an INQUIRY to each possible LUN on each possible target on the adapter to determine if a volume is present.

Does SCSI I/O timeout differ for RDM and VMFS?

No.

Does VMware ESX rely on the HBA’s port down, link down, and loop down timers when determining failover actions, or does ESX keep track of an internal counter based on the notification from the HBA that the link or target just went down?

VMware ESX relies on the FC driver timers such as port down and link down.

Are disk.maxLUN values maintained by target?

The configuration value `/proc/vmware/config/Disk/MaxLUN` is a per-target value. It limits the highest LUN on each target for which VMware ESX will probe during a rescan. The total number of volumes usable by VMware ESX is 256.

Does VMware ESX use SCSI-3 reservation?

VMware ESX does not use persistent reservations.

When two ESX hosts have access to the same disk or VMFS partition, or when a metadata change is initiated by one of the ESX hosts, is the volume reserved (locked) so that no other change can be performed during this operation? If I attempted to change metadata in the same vdisk during that time, would I see a reservation conflict?

The volume is reserved, but not for the entire duration of the metadata change. It is reserved long enough to make sure that the subsequent metadata change is atomic across all servers. Disks are locked exclusively for a host, so you cannot attempt a metadata change to the same disk from two different machines at the same time.

What situations can cause a possible reservation conflict? What if I change the label of a VMFS partition?

Reservation conflicts occur when you extend, shrink, create, or destroy files on VMFS volumes from multiple machines at a rapid rate.

How does VMware ESX handle I/O incompletes? If I send WRITE commands followed by lots of data, and I do not get status back, should I wait for the SCSI layer to abort it?

VMware ESX does not abort a command issued by a virtual machine. It is up to the virtual machine to time out the request and issue an abort command. Windows virtual machines typically wait 30 seconds before timing out a request.

Under what conditions does VMware ESX decide to failover or retry the same path?

VMware ESX does not take any proactive steps to fail over a path if the corresponding physical link fluctuates between on and off. VMware ESX fails over a path on the loss of FC connectivity or the detection of a passive path.

How does VMware ESX identify a path?

VMware ESX uses the serial number and the volume number or LUN to identify alternate paths. It actually does an INQUIRY for the VPROD device ID first (page 0x83). If the device does not support a device ID, it issues an INQUIRY for the serial number (page 0x80).

How does the host client determine that two or more target device on a SAN fabric are really just multiple paths to the same volume? Is this based on serial number?

The volume number or LUN of the path and the unique ID extracted from the SCSI INQUIRY must match in order for VMware ESX to collapse paths.

How does the host client determine which paths are active and which are passive?

How VMware ESX determines if a path is active or passive depends on the specific SAN array in use. Typically, but not always, ESX issues a SCSI TEST_UNIT_READY command on the path and interprets the response from the array to determine if the path is active or passive. SCSI path state evaluation is done whenever an FC event occurs.

How does the host client prioritize active and passive paths? Are I/O operations load balanced among the active paths?

I/O operations are not dynamically balanced. A manual intervention is needed to assign a path to each volume.

How does the host client use the WWPN/WWNN and the fabric-assigned routing addresses (S_ID and D_ID) of a target volume? Is there a mechanism for binding this information to logical devices exported to the applications running on the host?

The FC driver binds WWPN/WWNN to an HBA number rather than to a volume ID or LUN. This information is not exported to virtual machine applications.

How are device drivers loaded?

They are loaded according to PCI slot assignment. The board with the lowest device number, then the lowest function number, is loaded first. The function number distinguishes the individual ports on a physical board. The `/proc/pci` file lists the boards and their locations on the PCI bus.

4 Planning for VMware Infrastructure 3 with SAN

When SAN administrators configure a SAN array for use by an ESX host and its virtual machines, some aspects of configuration and setup are different than with other types of storage. The key considerations when planning a VMware Infrastructure installation with SAN include:

- Which SAN hardware should you select?
- How should you provision LUNs to the virtual infrastructure (volume size versus number of volumes, and using VMFS versus RDM)?
- Which virtual machine deployment methods should you use (cloning and patching guest operating systems)?
- What storage multipathing and failover options should you use (active/passive versus active/active)?
- Will VMware ESX boot from SAN?

This chapter describes the factors to consider (for both VMware Infrastructure and SAN storage administrators) when using VMware ESX with a SAN array. It also provides information on choosing from different SAN options when configuring ESX hosts to work with SAN storage.

Topics included in this chapter are the following:

Considerations for VMware ESX System Designs

The types of server hosts you deploy and the amount of storage space that virtual machines require determine the level of service the infrastructure can provide and how well the environment can scale to higher service demands as your business grows. The following is a list of factors you need to consider when building your infrastructure to scale in response to workload changes:

- What types of SAN configuration or topologies do you need?
  - Do you want to use single fabric or dual fabric? VMware Infrastructure 3 supports both.
  - How many paths to each volume are needed? It is highly recommended that at least two paths be provided to each volume for redundancy.
  - Is there enough bandwidth for your SAN? VMware Infrastructure 3 supports both 2GFC and 4GFC.
  - What types of array do you need? VMware Infrastructure 3 supports active/passive, active/active, FC-AL, and direct-connect storage arrays. It is very important that you get the latest information on arrays from the hardware compatibility list posted on VMware.com.
- How many virtual machines can I install per ESX host? This determines the type of server (CPU, memory, and so on).
- How big is each virtual machine's operating system and its data disks? This determines the storage capacity now (that is, which storage array model to use, how much disk space to buy now, and how much disk space to buy in six months). For **each** virtual machine, you can roughly estimate storage requirements using the following calculations:
  - (Size of virtual machine) + (size of suspend/resume space for virtual machine)) + (size of RAM for virtual machine) + (100MB for log files per virtual machine) is the minimum space needed for each virtual machine.

    NOTE: Size of suspend/resume snapshots of running virtual machines is equal to the size of the virtual machine.
  - For example, assuming a 15GB virtual machine with 1GB virtual RAM, the calculation result is:

    15GB (size of virtual machine) + 15GB (space for suspend/resume) + 1GB (virtual RAM) + 100MB

    The total recommended storage requirement is approximately 31.1GB. You should also plan extra storage capacity to accommodate disk-based snapshots according to vendor recommendations.
- What sorts of applications are planned for the virtual machines? Having this information helps determine the network adapter and FC bandwidth requirements.
- What rate of growth (business, data, and bandwidth) do you expect for your environment? This determines how to build your VMware and ESX infrastructure to allow room for growth while keeping disruption to a minimum.

VMware ESX with SAN Design Basics

Support for QLogic and Emulex FC HBAs allows an ESX host system to be connected to a SAN array. The virtual machines can then be stored on the SAN array volumes and can also use the SAN array volumes to store application data. Using ESX with a SAN improves flexibility, efficiency, and reliability. It also supports centralized management as well as failover and load balancing technologies.

Using a SAN with VMware ESX allows you to improve your environment's failure resilience:

- You can store data redundantly and configure multiple FC fabrics, eliminating a single point of failure.
- Site Recovery Manager can extend disaster recovery (DR) capabilities provided the storage array replication software is integrated. See information on VMware.com for solution compatibility.
- ESX host systems provide multipathing by default and automatically support it for every virtual machine. See "Multipathing and Path Failover" on page 29.
- Using ESX host systems extends failure resistance to the server. When you use SAN storage, all applications can instantly be restarted after host failure. See "Designing for Server Failure" on page 82.

Using VMware ESX with a SAN makes high availability and automatic load balancing affordable for more applications than if dedicated hardware were used to provide standby services.

- Because shared central storage is available, building virtual machine clusters that use MSCS becomes possible. See "Using Cluster Services" on page 83.
- If virtual machines are used as standby systems for existing physical servers, shared storage is essential and a SAN is the best solution.
- You can use the VMware VMotion capabilities to migrate virtual machines seamlessly from one host to another.
- You can use VMware HA in conjunction with a SAN for a cold-standby solution that guarantees an immediate, automatic response.
- You can use VMware DRS to automatically migrate virtual machines from one host to another for load balancing. Because storage is on a SAN array, applications continue running seamlessly.
- If you use VMware DRS clusters, you can put an ESX host into maintenance mode to have the system migrate all running virtual machines to other ESX hosts. You can then perform upgrades or other maintenance operations.
- The transportability and encapsulation of VMware virtual machines complements the shared nature of SAN storage. When virtual machines are located on SAN-based storage, it becomes possible to shut down a virtual machine on one server and power it up on another server—or to suspend it on one server and resume operation on another server on the same network—in a matter of minutes. This allows you to migrate computing resources while maintaining consistent, shared access.

Use Cases for SAN Shared Storage

Using VMware ESX in conjunction with SAN is particularly useful for the following tasks:

- **Maintenance with zero downtime** — When performing maintenance, you can use VMware DRS or VMotion to migrate virtual machines to other servers. If using shared storage, you can perform maintenance without interruption to the user.
- **Load balancing** — You can use VMotion explicitly or use VMware DRS to migrate virtual machines to other hosts for load balancing. If using shared storage, you can perform load balancing without interruption to the user.
- **Storage consolidation and simplification of storage layout** — Consolidating storage resources has administrative and utilization benefits in a virtual infrastructure. You can start by reserving a large volume and then allocating portions to virtual machines as needed. Volume reservation and creation from the storage device needs to happen only once.
- **Disaster recovery** — Having all data stored on a SAN can greatly facilitate remote storage of data backups. In addition, you can restart virtual machines on remote ESX hosts for recovery if one site is compromised.

Additional SAN Configuration Resources

In addition to this document, a number of other resources can help you configure your ESX host system in conjunction with a SAN.

- **VMware I/O Compatibility Guide** — Lists the currently approved HBAs, HBA drivers, and driver versions. See http://www.vmware.com/support/pubs/.
- **VMware SAN Compatibility Guide** — Lists currently approved storage arrays. Get the latest information from http://www.vmware.com/support/pubs/.
- **VMware Release Notes** — Provides information about known issues and workarounds. For the latest release notes, go to:

  http://www.vmware.com/support/pubs
- **VMware Knowledge Base** — Has information on common issues and workarounds. See http://www.vmware.com/kb.

Also, be sure to use your storage array vendor's documentation to answer most setup questions. Your storage array vendor might also offer documentation on using the storage array in an ESX environment.

VMware ESX, VMFS, and SAN Storage Choices

This section discusses available ESX host, VMFS, and SAN storage choices and provides advice on how to make them.

Creating and Growing VMFS

VMFS can be deployed on a variety of SCSI-based storage devices, including Fibre Channel and iSCSI SAN equipment. A virtual disk stored on VMFS always appears to the virtual machine as a mounted SCSI device. The virtual disk hides a physical storage layer from the virtual machine's operating system. This allows you to run even operating systems not certified for SAN inside the virtual machine.

For the operating system inside the virtual machine, VMFS preserves the guest operating system's file system semantics, which ensure correct application behavior and data integrity for applications running in virtual machines.

You can set up VMFS-based datastores in advance on any storage device that your ESX host discovers. Select a larger volume (2TB maximum) if you plan to create multiple virtual machines on it. You can then add virtual machines dynamically without having to request additional disk space.

However, if more space is needed, you can increase the VMFS datastore size by adding extents at any time—up to 64TB. Each VMFS extent has a maximum size of 2TB.

Considerations When Creating a VMFS

You need to plan how to set up storage for your ESX host systems before you format storage devices with VMFS. It is recommended to have one VMFS partition per datastore in most configurations. You can, however, decide to use one large VMFS datastore or one that expands across multiple LUN extents. VMware ESX lets you have up to 256 LUNs per system, with the minimum volume size of 1.2GB.

You might want fewer, larger VMFS volumes for the following reasons:

- More flexibility to create virtual machines without going back to the storage administrator for more space.
- Simpler to resize virtual disks, create storage array snapshots, and so on.
- Fewer VMFS-based datastores to manage.

You might want more, smaller storage volumes, each with a separate VMFS datastore, for the following reasons:

- Less contention on each VMFS due to locking and SCSI reservation issues.
- Less wasted storage space.
- Different applications might need different RAID characteristics.
- More flexibility, as the multipathing policy and disk shares are set per volume.
- Use of Microsoft Cluster Service requires that each cluster disk resource is in its own LUN (RDM type is required for MSCS in VMware ESX environment).

- Different backup policies and disk-based snapshots can be applied on an individual LUN basis.

You might decide to configure some of your servers to use fewer, larger VMFS datastores and other servers to use more, smaller VMFS datastores.

Choosing Fewer, Larger Volumes or More, Smaller Volumes

During ESX installation, you are prompted to create partitions for your system. You need to plan how to set up storage for your ESX host systems before you install. You can choose one of these approaches:

- Many volumes with one VMFS datastore on each LUN.
- Many volumes with one VMFS datastore spanning more than one LUN.
- Fewer larger volumes with one VMFS datastore on each LUN.
- Fewer larger volumes with one VMFS datastore spanning more than one LUN.

For VMware Infrastructure 3, it is recommended that you can have at most 16 VMFS extents per volume. You can, however, decide to use one large volume or multiple small volumes depending on I/O characteristics and your requirements.

Making Volume Decisions

When the storage characterization for a virtual machine is not available, there is often no simple answer when you need to decide on the volume size and number of LUNs to use. You can use a predictive or an adaptive approach for making the decision.

Predictive Scheme

In the predictive scheme, you:

- Create several volumes with different storage characteristics.
- Build a VMFS datastore in each volume (label each datastore according to its characteristics).
- Locate each application in the appropriate RAID for its requirements.
- Use disk shares to distinguish high-priority from low-priority virtual machines. Note that disk shares are relevant only within a given ESX host. The shares assigned to virtual machines on one ESX host have no effect on virtual machines on other ESX hosts.

Adaptive Scheme

In the adaptive scheme, you:

- Create a large volume (RAID 1+0 or RAID 5), with write caching enabled.
- Build a VMFS datastore on that LUN.
- Place four or five virtual disks on the VMFS datastore.
- Run the applications and see whether or not disk performance is acceptable.

- If performance is acceptable, you can place additional virtual disks on the VMFS. If it is not, you create a new, larger volume, possibly with a different RAID level, and repeat the process. You can use cold migration so you do not lose virtual machines when recreating the volume.

Special Volume Configuration Tips

- Each volume should have the right RAID level and storage characteristics for the applications in virtual machines that use the volume.
- If multiple virtual machines access the same datastore, use disk shares to prioritize virtual machines.

Data Access: VMFS or RDM

By default, a virtual disk is created in a VMFS volume during virtual machine creation. When guest operating systems issue SCSI commands to their virtual disks, the virtualization layer translates these commands to VMFS file operations.

An alternative to VMFS is using RDMs. As described earlier, RDMs are implemented using special files stored in a VMFS volume that act as a proxy for a raw device. Using an RDM maintains many of the same advantages as creating a virtual disk in the VMFS but gains the advantage of benefits similar to those of direct access to a physical device.

Benefits of RDM Implementation in VMware ESX

Raw device mapping provides a number of benefits as listed below.

- **User-Friendly Persistent Names** — RDM provides a user-friendly name for a mapped device. When you use a mapping, you do not need to refer to the device by its device name. Instead, you refer to it by the name of the mapping file. For example:

  ```
  /vmfs/volumes/myVolume/myVMDirectory/myRawDisk.vmdk
  ```

- **Dynamic Name Resolution** — RDM stores unique identification information for each mapped device. The VMFS file system resolves each mapping to its current SCSI device, regardless of changes in the physical configuration of the server due to adapter hardware changes, path changes, device relocation, and so forth.
- **Distributed File Locking** — RDM makes it possible to use VMFS distributed locking for raw SCSI devices. Distributed locking on a raw device mapping makes it safe to use a shared raw volume without losing data when two virtual machines on different servers try to access the same LUN.
- **File Permissions** — RDM makes it possible to set up file permissions. The permissions of the mapping file are enforced at file open time to protect the mapped volume.
- **File System Operations** — RDM makes it possible to use file system utilities to work with a mapped volume, using the mapping file as a proxy. Most operations that are valid for an ordinary file can be applied to the mapping file and are redirected to operate on the mapped device.

- **Snapshots** — RDM makes it possible to use virtual machine storage array snapshots on a mapped volume.

  NOTE: Snapshots are not available when raw device mapping is used in physical compatibility mode. See “Virtual and Physical Compatibility Modes” on page 61.

- **VMotion** — RDM lets you migrate a virtual machine using VMotion. When you use RDM, the mapping file acts as a proxy to allow VirtualCenter to migrate the virtual machine using the same mechanism that exists for migrating virtual disk files. See Figure 4-1.

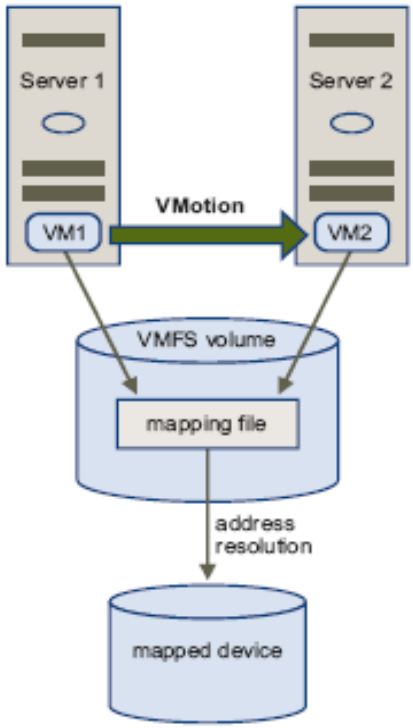


Figure 4-1. VMotion of a Virtual Machine Using an RDM

- **SAN Management Agents** — RDM makes it possible to run some SAN management agents inside a virtual machine. Similarly, any software that needs to access a device using hardware-specific SCSI commands can be run inside a virtual machine. This kind of software is called “SCSI target-based software.”

  NOTE: When you use SAN management agents, you need to select physical compatibility mode for the mapping file.

See Chapter 6 for more information on viewing and configuring datastores and managing RDMs using the VI Client.

VMware works with vendors of storage management software to ensure that their software functions correctly in environments that include VMware ESX. Some of these applications are:

- SAN management software
- Storage resource management (SRM) software
- Storage array snapshot software

- Replication software

Such software uses physical compatibility mode for RDMs so that the software can access SCSI devices directly.

Various management products are best run centrally (not on the ESX host), while others run well in the service console or in the virtual machines themselves. VMware does not certify or provide a compatibility matrix for these types of applications. To find out whether a SAN management application is supported in an ESX environment, contact the SAN management software provider.

Limitations of RDM in VMware ESX

When planning to use RDM, consider the following:

- **RDM is not available for devices that do not support the export of serial numbers** —RDM (in the current implementation) uses a SCSI serial number to identify the mapped device. Thus these devices (also known as block devices that connect directly to the `cciss` device driver or a tape device) cannot be used in RDMs.
- **RDM is available with VMFS-2 and VMFS-3 volumes only** — RDM requires the VMFS-2 or VMFS-3 format. In VMware ESX 3, the VMFS-2 file system is read-only. You need to upgrade the file system to VMFS-3 to be able to use the files it stores.
- **RDM does not allow use of VMware snapshots in physical compatibility mode** — The term **snapshot** here applies to the ESX host feature and not the snapshot feature in storage array data replication technologies. If you are using RDM in physical compatibility mode, you cannot use a snapshot with the disk. Physical compatibility mode allows the virtual machine to manage its own snapshot or mirroring operations.

  For more information on compatibility modes, see “Virtual and Physical Compatibility Modes” on page 61. For the support of snapshots or similar data replication features inherent in storage arrays, contact the specific array vendor for support.
- **No partition mapping** — RDM requires the mapped device to be a whole volume presented from a storage array. Mapping to a partition is not supported.
- **Using RDM to deploy LUNs** — This can require many more LUNs than is used in the typical shared VMFS configuration. The maximum number of LUNs supported by VMware ESX 3.x is 256.

Sharing Diagnostic Partitions

VMware ESX hosts collect debugging data in the form of a core dump, similar to most other operating systems. The location of this core dump can be specified as local storage, on a SAN volume, or on a dedicated partition. If your ESX host has a local disk, that disk is most appropriately used for the diagnostic partition, rather than using remote storage for it. That way, if you have an issue with remote storage that causes a core dump, you can use the core dump created in local storage to help you resolve the issue.

However, for diskless servers that boot from SAN, multiple ESX host systems can share one diagnostic partition on a SAN volume. If more than one ESX host system is using a volume as a diagnostic partition, that LUN for this volume must be zoned so that all the servers can access it.

Each ESX host requires a minimum of 100MB of storage space, so the size of the volume determines how many servers can share it. Each ESX host is mapped to a diagnostic slot.

If there is only one diagnostic slot on the storage device, then all ESX hosts sharing that device also map to the same slot, which can create problems.

For example, suppose you have configured 16 ESX hosts in your environment. If you have allocated enough memory for 16 slots, it is unlikely that core dumps will be mapped to the same location on the diagnostic partition, even if two ESX hosts perform a core dump at the same time.

Path Management and Failover

VMware ESX supports multipathing to maintain a constant connection between the server machine and the storage device in case of the failure of an HBA, switch, SP, or FC cable. Multipathing support does not require specific failover drivers or software.

To support path switching, the server typically has two or more HBAs available from which the storage array can be reached using one or more switches. Alternatively, the setup could include one HBA and two storage processors so that the HBA can use a different path to reach the disk array.

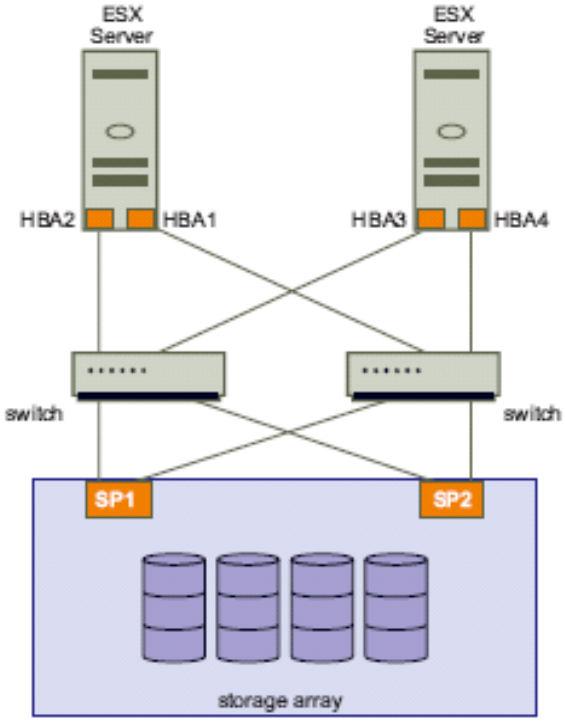


Figure 4-2. Multipathing and Failover

In Figure 4-2, multiple paths connect each server with the storage device. For example, if HBA1 or the link between HBA1 and the FC switch fails, HBA2 takes over and provides the connection between the server and the switch. The process of one HBA taking over for another is called **HBA failover**.

Similarly, if SP1 fails or the links between SP1 and the switches breaks, SP2 takes over and provides the connection between the switch and the storage device. This process is called **SP failover**. VMware ESX supports both HBA and SP failover with its multipathing capability.

You can choose a multipathing policy for your system, either Fixed or Most Recently Used. If the policy is Fixed, you can specify a preferred path. Each volume that is visible to the ESX host can have its own path policy. See "Viewing the Current Multipathing State" on page 119 for information on viewing the current multipathing state and on setting the multipathing policy.

NOTE: Virtual machine I/O might be delayed for at most 60 seconds while failover takes place, particularly on an active/passive array. This delay is necessary to allow the SAN fabric to stabilize its configuration after topology changes or other fabric events.

In the case of an active/passive array with a Fixed path policy, path thrashing may be a problem. See "Understanding Path Thrashing" on page 182.

Choosing to Boot ESX Systems from SAN

Rather than having ESX systems boot from their own local storage, you can set them up to boot up from a boot image stored on SAN. Before you consider how to set up your system for boot from SAN, you need to decide whether it makes sense for your environment. See "Using ESX Boot from SAN" in the previous chapter for more information on booting ESX systems from SAN.

You might want to use boot from SAN in the following situations:

- When you do not want to handle maintenance of local storage.
- When you need easy cloning of service consoles.
- In diskless hardware configurations, such as on some blade systems.

You should not use boot from SAN in the following situations:

- When you are using Microsoft Cluster Service with ESX Server 3.5 or older releases. VMware Infrastructure 3.5 Update 1 lifted this restriction (details provided in http://www.vmware.com/pdf/vi3_35/esx_3/vi3_35_25_u1_mscs.pdf)
- When there is a risk of I/O contention between the service console and VMkernel.
- SAN vendor does not support boot from SAN.

NOTE: With VMware ESX 2.5, you could not use boot from SAN together with RDM. With VMware ESX 3, this restriction has been removed.

Choosing Virtual Machine Locations

When you are working on optimizing performance for your virtual machines, storage location is an important factor. There is always a trade-off between expensive storage that offers high performance and high availability, and storage with lower cost and lower performance. Storage can be divided into different tiers depending on a number of factors:

- **High Tier** — Offers high performance and high availability. Might offer built-in snapshots to facilitate backups and point-in-time (PiT) restorations. Supports replication, full SP redundancy, and fibre drives. Uses high cost spindles.
- **Mid Tier** — Offers mid-range performance, lower availability, some SP redundancy, and SCSI drives. Might offer snapshots. Uses medium cost spindles.
- **Lower Tier** — Offers low performance; little internal storage redundancy. Uses low-end SCSI drives or SATA (serial low-cost spindles).

Not all applications need to be on the highest performance, most available storage—at least not throughout their entire life cycle.

NOTE: If you need some of the functionality of the high tier, such as snapshots, but do not want to pay for it, you might be able to achieve some of the high-performance characteristics in software. For example, you can create snapshots in software.

When you decide where to place a virtual machine, ask yourself these questions:

- How critical is the virtual machine?
- What are its performance and availability requirements?
- What are its point-in-time (PiT) restoration requirements?
- What are its backup requirements?
- What are its replication requirements?

A virtual machine might change tiers throughout its life cycle due to changes in criticality or changes in technology that push higher tier features to a lower tier. Criticality is relative, and might change for a variety of reasons, including changes in the organization, operational processes, regulatory requirements and disaster recovery planning.

Designing for Server Failure

The RAID architecture of SAN storage inherently protects you from failure at the physical disk level. A dual fabric, with duplication of all fabric components, protects the SAN from most fabric failures. The final step in making your whole environment failure resistant is to protect against server failure. This section briefly discusses ESX system failover options.

Using VMware HA

VMware HA allows you to organize virtual machines into failover groups. When a host fails, all its virtual machines are immediately started on different hosts. VMware HA requires SAN shared storage. When a virtual machine is restored on a different host,

it loses its memory state but its disk state is exactly as it was when the host failed (crash-consistent failover). See the *Resource Management Guide* for detailed information.

NOTE: You must be licensed to use VMware HA.

Using Cluster Services

Server clustering is a method of tying two or more servers together using a high-speed network connection so that the group of servers functions as a single, logical server. If one of the servers fails, then the other servers in the cluster continue operating, picking up the operations performed by the failed server.

VMware tests Microsoft Cluster Service in conjunction with ESX systems, but other cluster solutions might also work. Different configuration options are available for achieving failover with clustering:

- **Cluster in a box** — Two virtual machines on one host act as failover servers for each other. When one virtual machine fails, the other takes over. (This does not protect against host failures. It is most commonly done during testing of the clustered application.)
- **Cluster across boxes** — For a virtual machine on an ESX host, there is a matching virtual machine on another ESX host.
- **Physical to virtual clustering (N+1 clustering)** — A virtual machine on an ESX host acts as a failover server for a physical server. Because virtual machines running on a single host can act as failover servers for numerous physical servers, this clustering method provides a cost-effective N+1 solution.

See the VMware document, *Setup for Microsoft Cluster Service*, for more information.

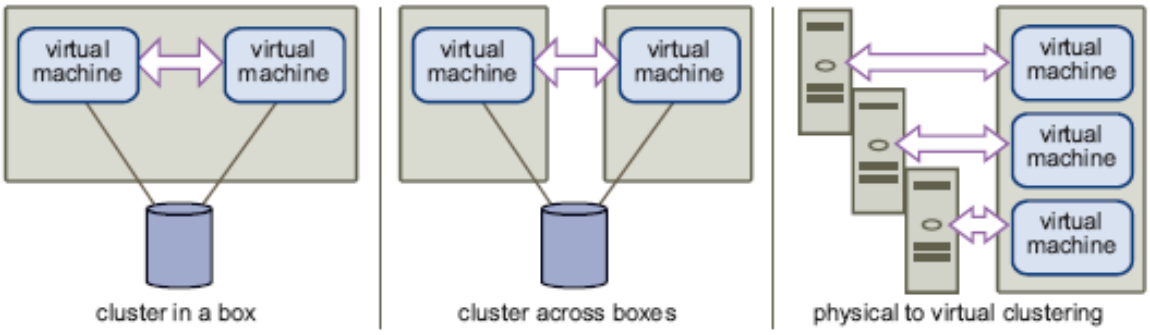


Figure 4-3. Clustering Using a Clustering Service

Server Failover and Storage Considerations

For each type of server failover, you must consider storage issues:

- Approaches to server failover work only if each server has access to the same storage. Because multiple servers require a lot of disk space, and because failover for the storage array complements failover for the server, SANs are usually employed in conjunction with server failover.
- When you design a SAN to work in conjunction with server failover, all volumes that are used by the clustered virtual machines must be seen by all ESX hosts. This is counterintuitive for SAN administrators, but is appropriate when using virtual machines.

Note that just because a volume is accessible to a host, all virtual machines on that host do not necessarily have access to all data on that volume. A virtual machine can access only the virtual disks for which it was configured. In case of a configuration error, virtual disks are locked when the virtual machine boots so no corruption occurs.

When you are using ESX boot from SAN, each boot volume should, as a rule, be seen only by the ESX host system that is booting from that volume. An exception is when you are trying to recover from a crash by pointing a second ESX host system to the same volume. In this case, the SAN volume in question is not really a boot from SAN volume. No ESX system is booting from it because it is corrupted. The SAN volume is a regular non-boot volume that is made visible to an ESX system.

Optimizing Resource Utilization

VMware Infrastructure allows you to optimize resource allocation by migrating virtual machines from over-utilized hosts to under-utilized hosts. There are two options:

- Migrate virtual machines manually using VMotion.
- Migrate virtual machines automatically using VMware DRS.

You can use VMotion or DRS only if the virtual disks are located on shared storage accessible to multiple servers. In most cases, SAN storage is used. For additional information on VMotion, see the *Virtual Machine Management Guide*. For additional information on DRS, see the *Resource Management Guide*.

VMotion

VMotion technology enables intelligent workload management. VMotion allows administrators to manually migrate virtual machines to different hosts. Administrators can migrate a running virtual machine to a different physical server connected to the same SAN, without service interruption. VMotion makes it possible to:

- Perform zero-downtime maintenance by moving virtual machines around so the underlying hardware and storage can be serviced without disrupting user sessions.
- Continuously balance workloads across the datacenter to most effectively use resources in response to changing business demands.

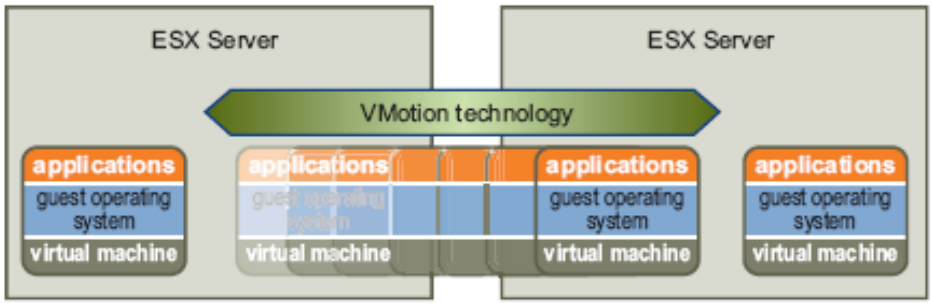


Figure 4-4. Migration with VMotion

VMware DRS

VMware DRS helps improve resource allocation across all hosts and resource pools.

DRS collects resource use information for all hosts and virtual machines in a VMware cluster and provides recommendations (or migrates virtual machines) in one of two situations:

- Initial placement — When you first power on a virtual machine in the cluster, DRS either places the virtual machine or makes a recommendation.
- Load balancing — DRS tries to improve resource use across the cluster by either performing automatic migrations of virtual machines (VMotion) or providing recommendation for virtual machine migrations.

For detailed information, see the VMware *Resource Management Guide*.

SAN System Design Choices

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 switches and storage arrays must provide timely and reliable access for all competing applications.

This section discusses some general SAN design basics. Topics included here are the following:

Determining 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 volumes and capacity of each volume)
- Redundancy level (RAID level)
- Response times (average time per I/O)
- Overall processing priority

Capacity planning services from VMware can provide exact data regarding your current infrastructure. See http://www.vmware.com/products/capacity_planner/ for more details.

Identifying 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.

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. Volumes 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 volumes. You must provide sufficient RAID-group resources for each set of volumes.
- 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 time.

Caching

Though ESX 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 volume 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. Include 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 volumes.

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, 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 volume failure allows applications to survive storage access faults. Mirroring can accomplish that protection.

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

- **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.

Site Recovery Manager (SRM) makes disaster recovery easier because you do not have to recreate all the virtual machines on the remote site when a failure occurs. Disk-based replication is integrated with SRM to provide a seamless failover from a replicated VMware Infrastructure environment.

5 Installing VMware Infrastructure 3 with SAN

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.

Topics included in this chapter are the following:

NOTE: This chapter provides an overview and high-level description of installation steps and procedures. For step-by-step installation instructions of VMware Infrastructure components, see the *VMware Installation and Upgrade Guide*, available at http://www.vmware.com.

SAN Compatibility 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)

Check your vendor's documentation to ensure both your SAN hardware and software is up-to-date and meets all requirements necessary to work with VMware Infrastructure and ESX hosts.

SAN Configuration and Setup

When you are ready to set up the SAN, complete these tasks:

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.
5. Document the SAN installation and all operational procedures.

Installation and Setup Overview

This section gives an overview of the installation and setup steps, with pointers to relevant information provided in VMware documentation, in particular, the VMware *Installation and Upgrade*, *Server Configuration*, and *SAN Configuration* guides.

Table 5-1. Installation and Setup Steps

Step	Description	Reference Documentation
1	Design your SAN if it is not already configured. Most existing SANs require only minor modification to work with ESX systems.	"VMware ESX with SAN Design Basics" on page 73. "VMware ESX, VMFS, and SAN Storage Choices" on page 75.
2	Check that all SAN components meet requirements.	"SAN Compatibility Requirements" on page 89. Also see the VMware ESX 3 Storage/SAN Compatibility Guide.
3	SAN Considerations	SAN connections are generally made through a switched fabric topology (FC-SW) although point-to-point topologies are also supported. In a few cases, direct attached storage connections (that is, connections without switches) are supported but that support is limited to certain vendor devices, notably those from EMC and IBM. **NOTE:** VMware strongly recommends single-initiator zoning in a switched fabric topology.
4	Set up the HBAs for the ESX hosts.	For special requirements that apply only to boot from SAN, see the previous section, "ESX Boot from SAN Requirements" on page 93. See also Chapter 6, "Using Boot from SAN with ESX Systems" in the VMware *SAN Configuration Guide*.
5	Perform any necessary storage array modification.	For an overview, see "Setting Up SAN Storage Devices with VMware ESX" in the VMware *SAN Configuration Guide*. Most vendors have vendor-specific documentation for setting up a SAN to work with VMware ESX.

Step	Description	Reference Documentation
6	Install VMware ESX on the hosts you have connected to the SAN and for which you have set up the HBAs.	VMware ESX *Installation and Upgrade Guide*.
7	Create virtual machines.	*Virtual Machine Management Guide*.
8	Set up your system for VMware HA failover or for using Microsoft Clustering Services. This step is optional.	*VMware Resource Management Guide* for ESX 3 and VirtualCenter 2. Also see the VMware *Setup for Microsoft Cluster Service* document.
9	Upgrade or modify your environment as needed.	Chapter 6, "Managing VMware Infrastructure with SAN." Search the VMware knowledge base articles for machine-specific information and late-breaking news.

VMware ESX Configuration and Setup

In preparation for configuring your SAN and setting up your ESX system to use SAN storage, review the following requirements and recommendations:

- **Hardware and Firmware** — Only a limited number of SAN storage hardware and firmware combinations are supported in conjunction with ESX systems. For an up-to-date list, see the *SAN Compatibility Guide* for ESX 3.5 at http://www.vmware.com/support/pubs/vi_pages/vi_pubs_35.html
- **Diagnostic Partition** — Unless you are using diskless servers, do not set up the diagnostic partition on a SAN volume.

  In the case of diskless servers that boot from SAN, a shared diagnostic partition is appropriate. See "Sharing Diagnostic Partitions" on page 79 for additional information on that special case.
- **Raw Device Mappings** (RDMs) — A SAN is likely to contain a large numbers of LUNs, some of which may be managed or replicated by the SAN storage hardware. In that case, use of RDMs can maintain independence of these LUNs, yet allow access to the raw devices from ESX systems. For more information on RDMs, see the VMware *Server Configuration Guide*.
- **Multipathing** — Multipathing provides protection against single points of failure in the SAN by managing redundant paths from an ESX host to any particular LUN and providing path failover and load distribution. If more than one ESX host is sharing access to a LUN, the LUN should be presented to all ESX hosts across additional redundant paths.
- **Queue Size** — Make sure the BusLogic or LSI Logic driver in the guest operating system specifies a queue depth that matches VMware recommendations. You can set the queue depth for the physical HBA during system setup or maintenance. For supported driver revisions and queue depth recommendations, see the VMware *SAN Compatibility Guide* as above.

- **SCSI Timeout** — On virtual machines running Microsoft Windows, consider increasing the value of the SCSI TimeoutValue parameter to allow Windows to better tolerate delayed I/O resulting from unanticipated path failover. See "Setting the HBA Timeout for Failover" on page 147.

FC HBA Setup

During FC HBA setup, consider the following points:

- **HBA Default Settings** — FC HBAs work correctly with the default configuration settings. Follow the configuration guidelines given by your storage array vendor.

  NOTE: For best results, use the same model of HBA and firmware within the same ESX host, if multiple HBAs are present. In addition, having both Emulex and QLogic HBAs in the same server mapped to the same FC target is not supported.

- **Setting the Timeout for Failover** — The timeout value used for detecting when a path fails is set in the HBA driver. Setting the timeout to 30 seconds is recommended to ensure optimal performance. To edit and/or determine the timeout value, follow the instructions in "Setting the HBA Timeout for Failover" on page 147.

- **Dedicated Adapter for Tape Drives** — For best results, use a dedicated SCSI adapter for any tape drives that you are connecting to an ESX system.

Setting Volume Access for VMware ESX

When setting volume allocations, note the following points:

- **Storage Provisioning via LUN Masking** — To ensure that an ESX system recognizes any VMFS volumes at startup, be sure to provision or mask all LUNs to the appropriate HBAs before using the ESX system in a SAN environment.

  NOTE: Provisioning all LUNs to all ESX HBAs at the same time is recommended. HBA failover works only if all HBAs see the same LUNs.

- **VMotion and VMware DRS** — When using VirtualCenter and VMotion or DRS, make sure that the LUNs for associated virtual machines are mapped to their respective ESX hosts. This is required to migrate virtual machines from one ESX host to another.

- **Active/Passive Array Considerations** — When performing virtual machine migrations across ESX hosts attached to active/passive SAN storage devices, make sure that all ESX hosts have consistent paths to the same active storage processors for the LUNs. Not doing so can cause path thrashing when a VMotion or DRS related migration occurs. See "Understanding Path Thrashing" on page 182.

VMware does not support path failover for storage arrays not listed in the VMware *SAN Compatibility Guide*. In those cases, you must connect the server to a single active port on the storage array.

Raw Device Mapping Considerations

- Use RDM to access a virtual machine disk if you want to use some of the hardware snapshot functions of the disk array, or if you want to access a disk from both a virtual machine and a physical machine in a cold-standby host configuration for data volumes.
- Use RDM for the shared disks in a Microsoft Cluster Service setup. See the VMware document “Setup for Microsoft Cluster Service” for details.

VMFS Volume Sizing Considerations

- Allocate a large volume for use by multiple virtual machines and set it up as a VMFS volume. You can then create or delete virtual machines dynamically without having to request additional disk space each time you add a virtual machine.

See Chapter 6, “Managing VMware Infrastructure with SAN” for additional recommendations. Also see “Common Problems and Troubleshooting” in Chapter 10 for troubleshooting information and remedies to common problems.

ESX Boot from SAN Requirements

When you have SAN storage configured with your ESX system, you can place the ESX boot image on one of the volumes on the SAN. This configuration has various advantages; however, systems must meet specific criteria, as described in this section. See “Using ESX Boot from SAN” on page 66 for more information on the benefits of using the boot from SAN option. Also see the VMware *SAN Configuration Guide* for specific installation instructions and tasks to set up the ESX boot from SAN option.

In addition to the general ESX with SAN configuration tasks, you must also complete the following tasks to enable your ESX host to boot from SAN.

1. Ensure the configuration settings meet the basic boot from SAN requirements. See “ESX Boot from SAN Requirements” in Table 5-1.
2. Prepare the hardware elements. This includes your HBA, network devices, and storage system. Refer to the product documentation for each device. Also see "Setting up the FC HBA for Boot from SAN" in the VMware *SAN Configuration Guide*.
3. Configure LUN masking on your SAN to ensure that each ESX host has a dedicated LUN for the boot partitions. The boot volume must be dedicated to a single server.
4. Choose the location for the diagnostic partition. Diagnostic partitions can be put on the same volume as the boot partition. Core dumps are stored in diagnostic partitions. See “Sharing Diagnostic Partitions” on page 79.

The VMware *SAN Configuration Guide* provides additional instructions on installation and other tasks you need to complete before you can successfully boot your ESX host from SAN.

The following table summarizes considerations and requirements to enable ESX systems to boot from SAN.

Table 5-2. Boot from SAN Requirements

Requirement	Description
ESX system requirements	ESX 3.0 or later is recommended. When you use an ESX 3 system, RDMs are supported in conjunction with boot from SAN. For an ESX 2.5.x system, RDMs are not supported in conjunction with boot from SAN.
HBA requirements	The BIOS for your HBA card must be enabled and correctly configured to allow booting from a SAN device. The HBA should be plugged into the lowest PCI bus and slot number. This allows the drivers to detect the HBA quickly because the drivers scan the HBAs in ascending PCI bus and slot numbers. **NOTE**: For specific HBA driver and version information, see the *ESX I/O Compatibility Guide*.
Boot LUN considerations	When you boot from an active/passive storage array, the storage processor whose WWN is specified in the BIOS configuration of the HBA must be active. If that storage processor is passive, the HBA cannot support the boot process. To facilitate BIOS configuration, use LUN masking to ensure that the boot LUN can be seen only by its corresponding ESX host.
Hardware specific considerations	Some hardware specific considerations apply. For example, if you are running an IBM eServer BladeCenter and use boot from SAN, you must disable IDE drives on the blades. For additional hardware-specific considerations, check the VMware knowledge base articles and see the VMware *SAN Compatibility Guide*.

VMware ESX with SAN Restrictions

The following restrictions apply when you use VMware ESX with a SAN:

- VMware ESX does not support FC connected tape devices. These devices can, however, be managed by the VMware Consolidated Backup proxy server, which is discussed in the VMware *Virtual Machine Backup Guide*.
- You cannot use virtual machine logical volume manager (LVM) software to mirror virtual disks. Dynamic disks in a Microsoft Windows virtual machine are an exception, but they also require special considerations.

6 Managing VMware Infrastructure 3 with SAN

VMware Infrastructure management includes the tasks you must perform to configure, manage, and maintain the operation of ESX hosts and virtual machines. This chapter focuses on management operations pertaining to VMware Infrastructure configurations that use SAN storage.

The following sections in this chapter describe operations specific to managing VMware Infrastructure SAN storage.

For more information on using the VI Client and performing operations to manage ESX hosts and virtual machines, see the VMware *Basic System Administration* guide and *Server Configuration Guide*.

VMware Infrastructure Component Overview

To run your VMware Infrastructure environment, you need the following items:

- **VMware ESX** — The virtualization platform used to create the virtual machines as a set of configuration and disk files that together perform all the functions of a physical machine.

  Through VMware ESX, you run the virtual machines, install operating systems, run applications, and configure the virtual machines. Configuration includes identifying the virtual machine's resources, such as storage devices.

  The server incorporates a resource manager and service console that provides bootstrapping, management, and other services that manage your virtual machines.

Each ESX host has a VI Client available for your management use. If your ESX host is registered with the VirtualCenter Management Server, a VI Client that accommodates the VirtualCenter features is available.

For complete information on installing VMware ESX, see the *Installation and Upgrade Guide*. For complete information on configuring VMware ESX, see the *Server Configuration Guide*.

- **VirtualCenter** — A service that acts as a central administrator for VMware ESX hosts that are connected on a network. VirtualCenter directs actions on the virtual machines and the virtual machine hosts (ESX installations). The VirtualCenter Management Server (VirtualCenter Server) provides the working core of VirtualCenter.

  The VirtualCenter Server is a single Windows service and is installed to run automatically. As a Windows service, the VirtualCenter Server runs continuously in the background, performing its monitoring and managing activities even when no VI Clients are connected and even if nobody is logged on to the computer where it resides. It must have network access to all the hosts it manages and be available for network access from any machine where the VI Client is run.

- **VirtualCenter database** — A persistent storage area for maintaining status of each virtual machine, host, and user managed in the VirtualCenter environment. The VirtualCenter database can be remote or local to the VirtualCenter Server machine.

  The database is installed and configured during VirtualCenter installation. If you are accessing your ESX host directly through a VI Client, and not through a VirtualCenter Server and associated VI Client, you do not use a VirtualCenter database.

- **Datastore** — The storage locations for virtual machine files specified when creating virtual machines. Datastores hide the idiosyncrasies of various storage options (such as VMFS volumes on local SCSI disks of the server, the Fibre Channel SAN disk arrays, the iSCSI SAN disk arrays, or network-attached storage (NAS) arrays) and provide a uniform model for various storage products required by virtual machines.

- **VirtualCenter agent** — On each managed host, software (vpxd) that provides the interface between the VirtualCenter Server and host agents (hostd). It is installed the first time any ESX host is added to the VirtualCenter inventory.

- **Host agent** — On each managed host, software that collects, communicates, and executes the actions received through the VI Client. It is installed as part of the ESX installation.

- **VirtualCenter license server** — Server that stores software licenses required for most operations in VirtualCenter and VMware ESX, such as powering on a virtual machine.

  VirtualCenter and VMware ESX support two modes of licensing: license server-based and host-based. In host-based licensing mode, the license files are stored on individual ESX hosts. In license server-based licensing mode, licenses are stored on a license server, which makes these licenses available to one or more hosts. You can run a mixed environment employing both host-based and license server-based licensing.

VirtualCenter and features that require VirtualCenter, such as VMotion, must be licensed in license server-based mode. ESX-specific features can be licensed in either license server-based or host-based mode.

See the *Installation and Upgrade Guide* for information on setting up and configuring licensing.

The figure below illustrates the components of a VMware Infrastructure configuration with a VirtualCenter Server.

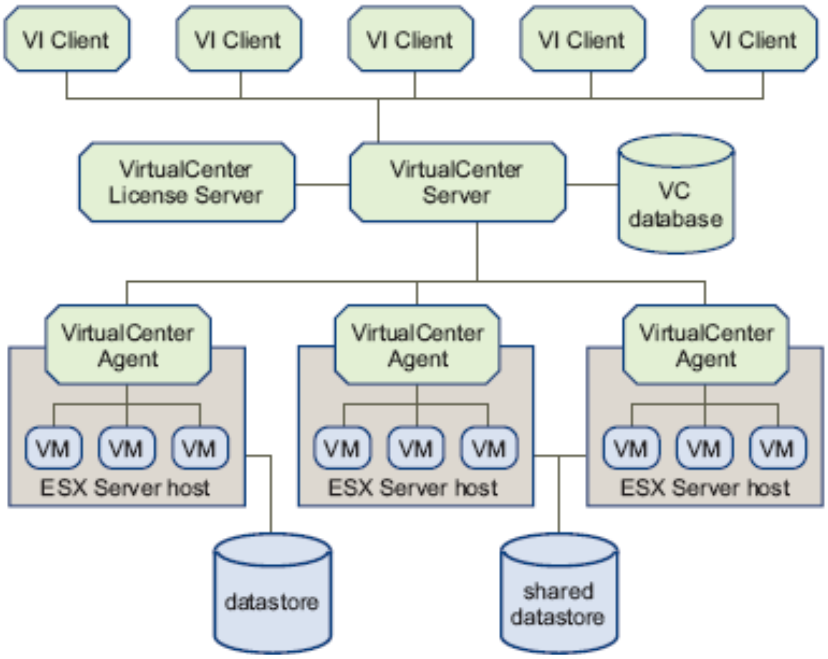


Figure 6-1. VMware Infrastructure Components with a VirtualCenter Server

VMware Infrastructure User Interface Options

Whether connecting directly to VMware ESX or through a VirtualCenter Server, user interface options for performing infrastructure management tasks include the following:

- **Virtual Infrastructure (VI) Client** — The VI Client is a required component and provides the primary interface for creating, managing, and monitoring virtual machines, their resources, and hosts. It also provides console access to virtual machines.

  VI Client is installed on a Windows machine separate from your ESX or VirtualCenter Server installation. While all VirtualCenter activities are performed by the VirtualCenter Server, you must use the VI Client to monitor, manage, and control the server. A single VirtualCenter Server or ESX host can support multiple, simultaneously connected VI Clients.

  The VI Client provides the user interface to both the VirtualCenter Server and ESX hosts. The VI Client runs on a machine with network access to the VirtualCenter Server or ESX host. The interface displays slightly different options depending on which type of server you are connected to.

- **Virtual Infrastructure (VI) Web Access** —Web interface through which you can perform basic virtual machine management and configuration, and get console access to virtual machines. It is installed with your ESX host. Similar to the VI Client, VI Web Access works directly with an ESX host or through VirtualCenter. See the VMware *Web Access Administrator's Guide* for additional information.
- **VMware Service Console** — Command-line interface to VMware ESX for configuring your ESX hosts. Typically, this is used in conjunction with support provided by a VMware technical support representative.

VI Client Overview

There are two primary methods for managing your virtual machines using VI Client:

- Directly through an ESX standalone host that can manage only those virtual machines and the related resources installed on it.
- Through a VirtualCenter Server that manages multiple virtual machines and their resources distributed over many ESX hosts.

The VI Client adapts to the server it is connected to. When the VI Client is connected to a VirtualCenter Server, the VI Client displays all the options available to the VMware Infrastructure environment, based on the licensing you have configured and the permissions of the user. When the VI Client is connected to an ESX host, the VI Client displays only the options appropriate to single host management.

The VI Client is used to log on to either a VirtualCenter Server or an ESX host. Each server supports multiple VI Client logons. The VI Client can be installed on any machine that has network access to the VirtualCenter Server or an ESX host.

By default, administrators are allowed to log on to a VirtualCenter Server. Administrators here are defined to be either:

- Members of the local Administrators group if the VirtualCenter Server is not a domain controller.
- Members of the domain Administrators group if the VirtualCenter Server is a domain controller.

The default VI Client layout is a single window with a menu bar, a navigation bar, a toolbar, a status bar, a panel section, and pop-up menus.

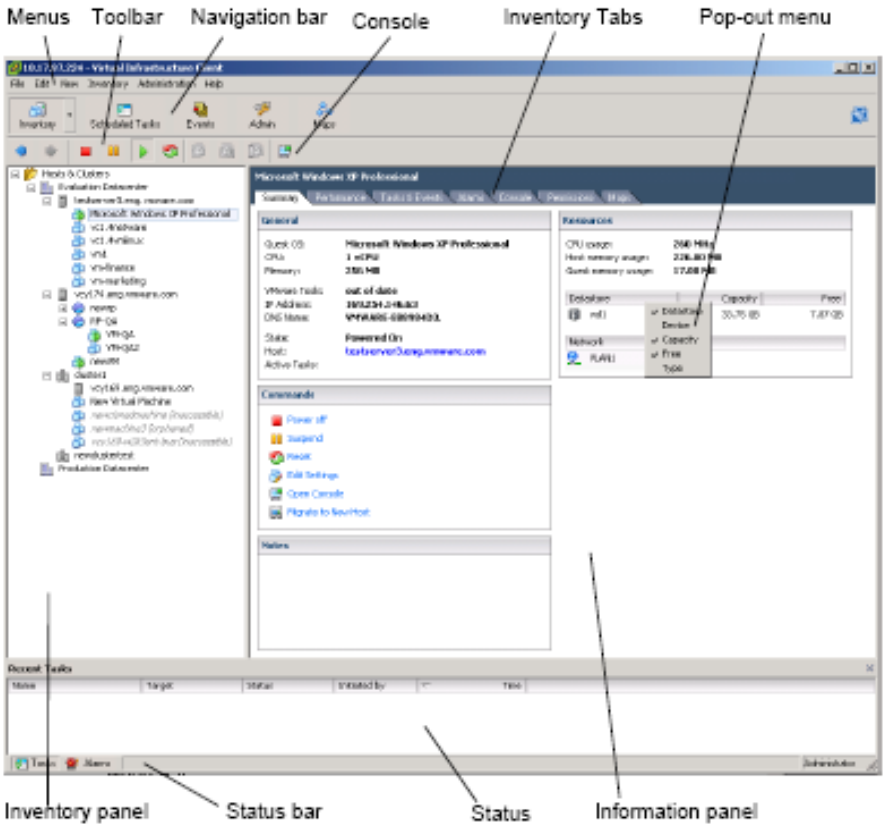


Figure 6-2. VI Client Layout

Managed Infrastructure Computing Resources

VirtualCenter monitors and manages various components (including hosts and virtual machines) of your virtual and physical infrastructure—potentially hundreds of virtual machines and other objects. The names of specific infrastructure components in your environment can be changed to reflect their business location or function. For example, they can be named after company departments or locations or functions. The managed components are:

- **Virtual Machines and Templates** — A virtualized x86 personal computer environment in which a guest operating system and associated application software can run. Multiple virtual machines can operate on the same managed host machine concurrently. Templates are virtual machines that are not allowed to be powered on but are used instead to create multiple instances of the same virtual machines design.
- **Hosts** — The primary component upon which all virtual machines reside. If the VI Client is connected to a VirtualCenter Server, many hosts can be managed from the same point. If the Virtual Infrastructure Client is connected to an ESX system, there can be only one host.

  NOTE: When VirtualCenter refers to a host, this means the physical machine on which the virtual machines are running. All virtual machines within the VMware

Infrastructure environment are physically on ESX hosts. The term **host** in this document means the ESX host that has virtual machines on it.

- **Resource pools** — A structure that allows delegation of control over the resources of a host. Resource pools are used to compartmentalize CPU and memory resources in a cluster. You can create multiple resource pools as direct children of a host or cluster, and configure them. You can then delegate control over them to other individuals or organizations. The managed resources are CPU and memory from a host or cluster. Virtual machines execute in, and draw their resources from, resource pools.
- **Clusters** — A collection of ESX hosts with shared resources and a shared management interface. When you add a host to a cluster, the host's resources become part of the cluster's resources. The cluster manages the CPU and memory resources of all hosts. For more information, see the *Resource Management Guide*.
- **Datastores** — Virtual representations of combinations of underlying physical storage resources in the datacenter. These physical storage resources can come from the local SCSI disk of the server, the FC SAN disk arrays, the iSCSI SAN disk arrays, or NAS arrays.
- **Networks** — Networks that connect virtual machines to each other in the virtual environment or to the physical network outside. Networks also connect VMkernel to VMotion and IP storage networks and the service console to the management network.
- **Folders** — Containers used to group objects and organize them into hierarchies. This not only is convenient but also provides a natural structure upon which to apply permissions. Folders are created for the following object types:
  - Datacenters
  - Virtual machines (which include templates)
  - Compute resources (which include hosts and clusters)

  The datacenter folders form a hierarchy directly under the root node and allow users to group their datacenters in any convenient way. Within each datacenter are one hierarchy of folders with virtual machines and/or templates and one hierarchy of folders with hosts and clusters.
- **Datacenters** — Unlike a folder, which is used to organize a specific object type, a datacenter is an aggregation of all the different types of objects needed to do work in virtual infrastructure: hosts, virtual machines, networks, and datastores.

  Within a datacenter there are four separate categories of objects:
  - Virtual machines (and templates)
  - Hosts (and clusters)
  - Networks
  - Datastores

  Because it is often not possible to put these objects into a hierarchy, objects in these categories are provided in flat lists.

  Datacenters act as the namespace boundary for these objects. You cannot have two objects (for example, two hosts) with the same name in the same

datacenter, but you can have two objects with the same name in different datacenters. Because of the namespace property, VMotion is permitted between any two compatible hosts within a datacenter, but even powered-off virtual machines cannot be moved between hosts in different datacenters. Moving an entire host between two datacenters is permitted.

Additional VMware Infrastructure 3 Functionality

Additional VirtualCenter features include:

- **VMotion** — A feature that enables you to move running virtual machines from one ESX host to another without service interruption. It requires licensing on both the source and target host. The VirtualCenter Server centrally coordinates all VMotion activities.
- **VMware HA** — A feature that enables a cluster with high availability. If a host goes down, all virtual machines that were on the host are promptly restarted on different hosts.

  When you enable the cluster for high availability, you specify the number of hosts you would like to be able to recover. If you specify the allowed number of host failures as 1, VMware HA maintains enough capacity across the cluster to tolerate the failure of one host.

  All running virtual machines on that host can be restarted on remaining hosts. By default, you cannot power on a virtual machine if doing so violates required failover capacity. See the *Resource Management Guide* for more information.
- **VMware DRS** — A feature that helps improve resource allocation across all hosts and resource pools. VMware DRS collects resource usage information for all hosts and virtual machines in the cluster and gives recommendations (or migrates virtual machines) in one of two situations:
  - **Initial placement** — When you first power on a virtual machine in the cluster, DRS either places the virtual machine or makes a recommendation.
  - **Load balancing** — DRS tries to improve resource utilization across the cluster by performing automatic migrations of virtual machines (VMotion) or by providing a recommendation for virtual machine migrations.
- **VMware Infrastructure SDK package** — APIs for managing virtual infrastructure and documentation describing those APIs. The SDK also includes the VirtualCenter Web Service interface, Web Services Description Language (WSDL), and example files. This is available through an external link. To download the SDK package, go to http://www.vmware.com/support/developer.

Accessing and Managing Virtual Disk Files

Typically, you use VI Client to perform a variety of operations on your virtual machines. Direct manipulation of your virtual disk files on VMFS is possible through the ESX service console and VMware SDKs, although using the VI Client is the preferred method.

From the service console, you can view and manipulate files in the `/vmfs/volumes` directory in mounted VMFS volumes with ordinary file commands, such as `ls` and `cp`.

Although mounted VMFS volumes might appear similar to any other file system, such as ext3, VMFS is primarily intended to store large files, such as disk images with the size of up to 2TB. You can use `ftp`, `scp`, and `cp` commands for copying files to and from a VMFS volume as long as the host file system supports these large files.

Additional file operations are enabled through the `vmkfstools` command. This command supports the creation of a VMFS on a SCSI disk and is used for the following:

- Creating, extending, and deleting disk images.
- Importing, exporting, and renaming disk images.
- Setting and querying properties of disk images.
- Creating and extending a VMFS file system.

The vmkfstools Commands

The vmkfstools (virtual machine kernel files system tools) commands provide additional functions that are useful when you need to create files of a particular block size, and when you need to import files from and export files to the service console's file system. In addition, vmkfstools is designed to work with large files, overcoming the 2GB limit of some standard file utilities.

NOTE: For a list of supported vmkfstools commands, see the VMware *Server Configuration Guide*.

Managing Storage in a VMware SAN Infrastructure

The VI Client displays detailed information on available datastores, storage devices the datastores use, and configured adapters.

Creating and Managing Datastores

Datastores are created and managed the ESX through the VI Client interface in one of two ways:

- **They are discovered when a host is added to the inventory** – When you add an ESX host to the Virtual Center inventory, the VI Client displays any datastores recognized by the host.
- **They are created on an available storage device (LUN)** – You can use the VI Client *Add Storage* interface to create and configure a new datastore. For more information, see “Managing Raw Device Mappings” on page 107.

Viewing Datastores

You can view a list of available datastores and analyze their properties. To display datastores:

1. Select the host for which you want to see the storage devices and click the **Configuration** tab.
2. In the Hardware panel, choose **Storage (SCSI, SAN, and NFS)**.

   The list of datastores (volumes) appears in the Storage panel. For each datastore, the Storage section shows summary information, including:

   - The target storage device where the datastore is located. See “Understanding Storage Device Naming” on page 106.
   - The type of file system the datastore uses—for example, VMFS, Raw Device Mapping (RDM), or NFS. (See “File System Formats” on page 49.)
   - The total capacity, including the used and available space.
3. To view additional details about the specific datastore, select the datastore from the list. The Details section shows the following information:

   - The location of the datastore.
   - The individual extents the datastore spans and their capacity. An extent is a VMFS-formatted partition (a piece of a volume). For example, vmhba 0:0:14 is a volume, and vmhba 0:0:14:1 is a partition. One VMFS volume can have multiple extents.

     NOTE: The abbreviation vmhba refers to the physical HBA (SCSI, FC, network adapter, or iSCSI HBA) on the ESX system, not to the SCSI controller used by the virtual machines.
   - The paths used to access the storage device.

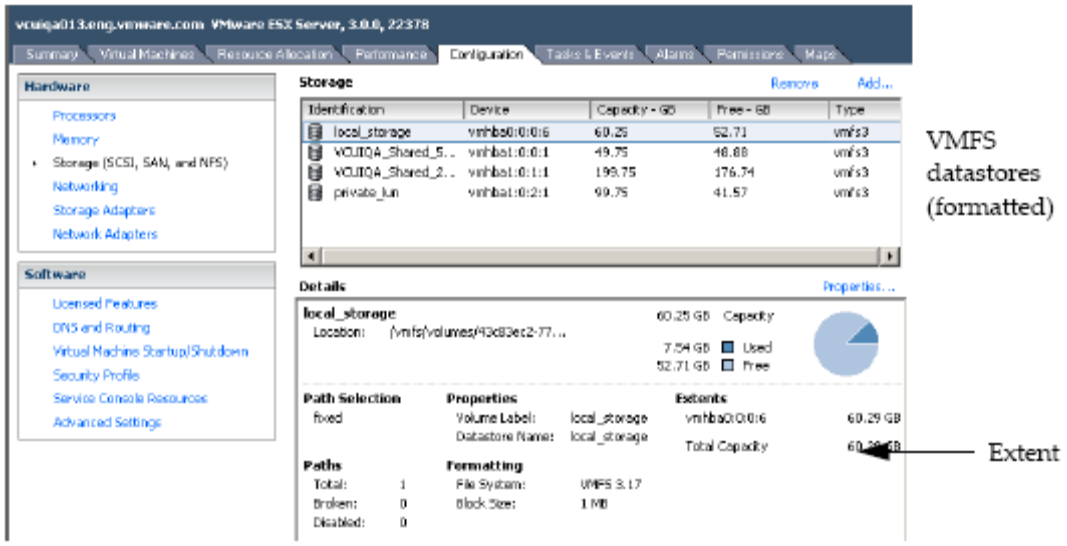


Figure 6-3. Datastore Information

You can edit or remove any of the existing datastores. When you edit a datastore, you can change its label, add extents, or modify paths for storage devices. You can also upgrade the datastore.

It is also possible to browse a datastore to view a graphical representation of the files and folders in the datastore as shown in Figure 6-4.

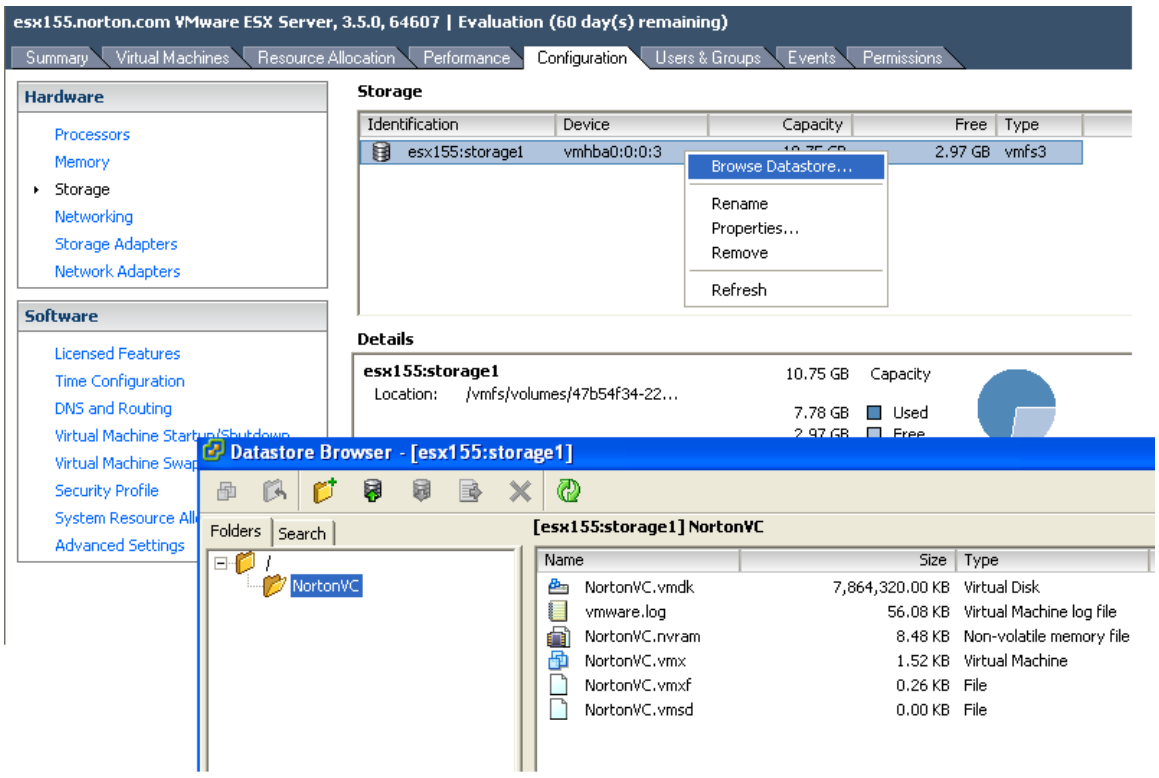


Figure 6-4. Browsing a datastore

Viewing Storage Adapters

The VI Client displays any storage adapters available to your system. To display storage adapters, on the host **Configuration** tab, click the **Storage Adapters** link in the Hardware panel.

You can view the following information about the storage adapters:

- Existing storage adapters.
- Type of storage adapter, such as Fibre Channel SCSI or iSCSI.
- Details for each adapter, such as the storage device it connects to and its target ID.

To view the configuration properties for a specific adapter:

1. Select the host for which you want to see the HBAs and click the **Configuration** tab.

   You can view a list of all storage devices from the Summary tab, but you cannot see details or manage a device from there.

2. In the Hardware panel, choose **Storage Adapters**.

   The list of storage adapters appears. You can select each adapter for additional information.

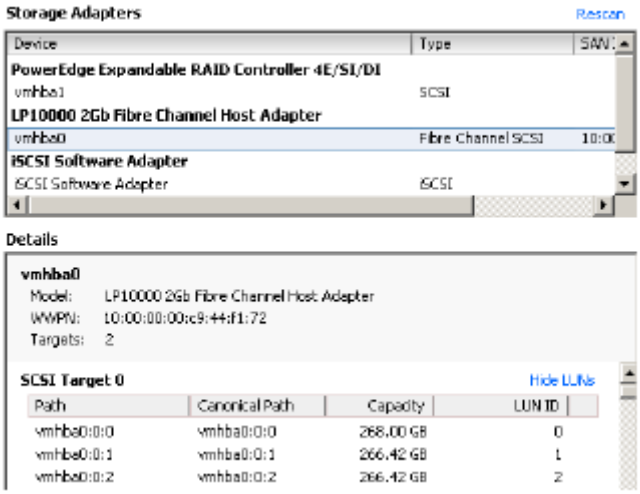


Figure 6-5. Host Bus Adapter information

The Details view provides information about the number of volumes the adapter connects to and the paths it uses. If you want to change the path's configuration and/or properties, select this path from the list, right-click the path, and click **Manage Paths** to bring up the Manage Paths Wizard. For information on managing paths, see "Managing Multiple Paths for Fibre Channel" on page 119.

Understanding Storage Device Naming Conventions

In the VI Client, the name of a storage device or volume appears as a sequence of three or four numbers, separated by colons, such as vmhba1:1:3:1. The name has the following meaning:

<SCSI HBA>:*<SCSI target>*:*<SCSI LUN>*:*<disk partition>*

NOTE: The abbreviation vmhba refers to different physical HBAs on the ESX system. It can also refer to the software iSCSI initiator (vmhba40) that VMware ESX implements using the VMkernel network stack.

The sequence of numbers in an ESX device name may change but still refer to the same physical device. For example, vmhba1:2:3 represents SCSI HBA 1, attached to SCSI target 2, on SCSI LUN 3. When the ESX system is rebooted, the device name for LUN 3 could change to vmhba1:1:3. The numbers have the following meaning:

- The first number, the HBA, changes when an outage on the FC or iSCSI network occurs. In this case, the ESX system has to use a different HBA to access the storage device.
- The second number, the SCSI target, changes in case of any modifications in the mappings of the FC or iSCSI targets visible to the ESX host.
- The fourth number indicates a partition on a disk or volume. When a datastore occupies the entire disk or volume, the fourth number is not present.

The vmhba1:1:3:1 example refers to the first partition on SCSI volume with LUN 3, SCSI target 1, which is accessed through HBA 1.

Resolving Issues with LUNs That Are Not Visible

If the display (or output) of storage devices differs from what you expect, check the following:

- **Cable connectivity** — If you do not see a SCSI target, the problem could be cable connectivity or zoning. Check the cables first.
- **Zoning** — Zoning limits access to specific storage array ports, increases security, and decreases traffic over the network. See your specific storage vendor's documentation for zoning capabilities and requirements. Use the accompanying SAN switch software to configure and manage zoning.
- **LUN masking** — If an ESX host sees a particular storage array port but not the expected LUNs behind that port, it might be that LUN masking has not been set up properly.

  For boot from SAN, ensure that each ESX host sees only required LUNs. In particular, do not allow any ESX host to see any boot LUN other than its own. Use disk array software to make sure the ESX host can see only the LUNs that it is supposed to see.

  Ensure that the Disk.MaxLUN and Disk.MaskLUN settings allow you to view the LUN you expect to see. See "Changing the Number of LUNs Scanned Using Disk.MaxLUN" on page 117.
- **Storage processor** — If a disk array has more than one storage processor, make sure that the SAN switch has a connection to the SP that owns the volumes

you want to access. On some disk arrays, only one SP is active and the other SP is passive until there is a failure. If you are connected to the wrong SP (the one with the passive path) you might not see the expected LUNs, or you might see the LUNs but get errors when trying to access them.

- **Volume or volume resignature** — If you used array-based data replication to make a clone or a snapshot of existing volumes and ESX host configurations, rescans might not detect volume or ESX changes because volume resignature options are not set correctly. VMFS volume resignaturing allows you to make a hardware snapshot of a volume (that is either configured as VMFS or a RDM volume) and access that snapshot from an ESX system. It involves resignaturing the volume UUID and creating a new volume label. You can control resignaturing as follows:
  - Use the LVM.EnableResignature option to turn auto-resignaturing on or off (the default is off).

  NOTE: As a rule, a volume should appear with the same LUN ID to all hosts that access the same volume.

To mount both the original and snapshot volumes on the same ESX host:

1. In the VI Client, select the host in the inventory panel.
2. Click the **Configuration** tab and click **Advanced Settings**.
3. Perform the following tasks repeatedly, as needed:
   a) Create the array-based snapshot.
   b) Make the snapshot from the storage array visible to VMware ESX.
   c) Select **LVM** in the left panel; then set the LVM.EnableResignature option to **1**.

   NOTE: Changing LVM.EnableResignature is a global change that affects all LUNs mapped to an ESX host.
4. Rescan the LUN.

   After rescan, the volume appears as /vmfs/volumes/snap-*DIGIT*-<*old-label*>

   NOTE: Any virtual machines on this new snapshot volume are not auto-discovered. You have to manually register the virtual machines. If the .vmx file for any of the virtual machines or the .vmsd file for virtual machine snapshots contains /vmfs/volumes/<*label or UUID*>/ paths, you must change these items to reflect the resignatured volume path.

If necessary, set the LVM.EnableResignature option to **0** after resignaturing is complete. For more information, see "Understanding Resignaturing Options" in Chapter 10.

Managing Raw Device Mappings

The tools available to manage RDMs include the VMware Virtual Infrastructure Client, the vmkfstools utility, and ordinary file system utilities used in the service console.

Using the VI Client, you can configure hardware, add disks, and create a raw device mapping. You can use the VI Client to import and export virtual disks, including

mapped raw devices, and to manage paths. You can also perform common file system operations, such as renaming and setting permissions.

Creating a Raw Device Mapping

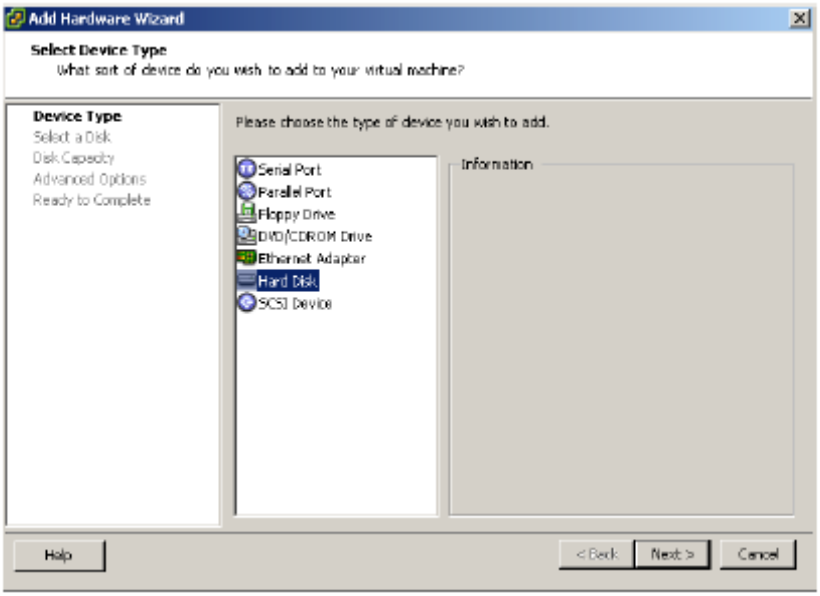
Use the VI Client to create an RDM.

To create an RDM:

1. Log on as administrator or as the owner of the virtual machine to which the mapped disk will belong.
2. Select the virtual machine from the inventory panel.
3. On the Summary tab, click the **Edit Settings** link.

   The Virtual Machine Properties dialog box opens.
4. Click **Add**.

   The Add Hardware Wizard opens.



5. Choose **Hard Disk** as the type of device you want to add and click **Next**.
6. In the Select a Disk panel, select **Mapped SAN volume**.
7. Choose the volume from the list of available volumes.
8. Select **Store with Virtual Machine**.
9. Choose **Physical** or **Virtual** for the Compatibility mode.
10. Specify the SCSI identifier in the Specify Advanced Options panel.

    Typically, you do not need to change the default settings.
11. In the Ready to Complete panel, review your options and click **Finish** to add the disk.

Configuring Datastores in a VMware SAN Infrastructure

An ESX system uses datastores to store all files associated with its virtual machines. The datastore is a logical storage unit recognized by an ESX host that can utilize disk space on one physical disk device or one disk partition, or can span several physical devices. The datastore can exist on different types of physical devices including SCSI, iSCSI, Fibre Channel SAN, or NFS.

For SAN storage, VMware ESX supports FC adapters, which allow an ESX system to be connected to a SAN and see the disk arrays on the SAN. This section shows you operations to configure and manage FC storage:

For additional information:

- For more information on datastores, see “Datastores and File Systems” on page 55.
- For information on configuring SANs, see the VMware *SAN Configuration Guide*.
- For information on supported SAN storage devices for VMware ESX, see the VMware *SAN Compatibility Guide*.
- For information about multipathing for FC HBAs and how to manage paths, see “Managing Multiple Paths for Fibre Channel” on page 119.

Figure 6-6 depicts virtual machines using FC storage.

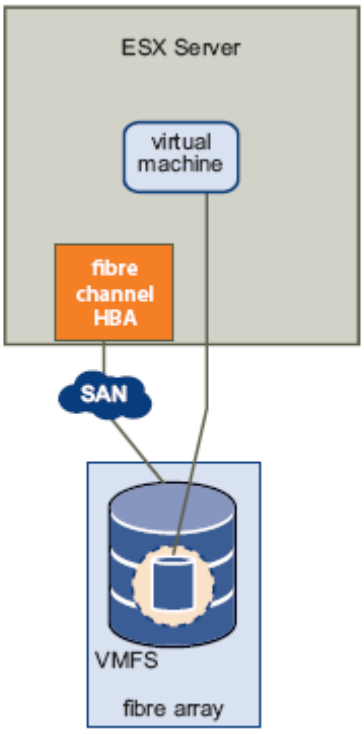


Figure 6-6. Fibre Channel Storage

In this configuration, VMware ESX connects to SAN storage using an FC adapter. The adapter connects to SAN fabric consisting of FC switches and storage arrays, which then present volumes from physical disks to your ESX system. You can access the volumes and create a datastore that you use for your ESX storage needs. The datastore uses the VMFS format. After you create a VMFS-based datastore, you can modify it. See the following sections for more information.

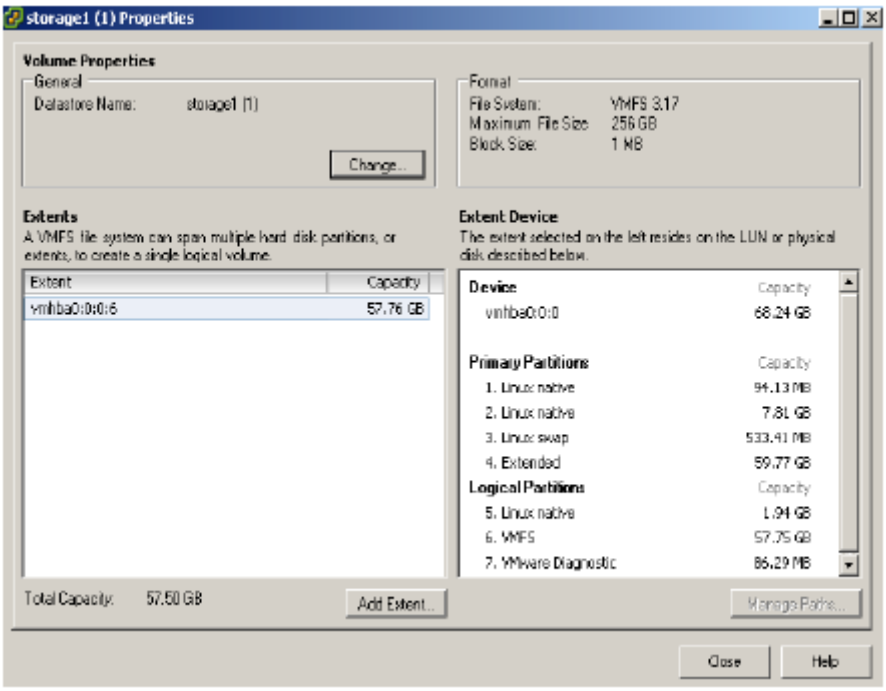
Changing the Names of Datastores

You can change the name of an existing VMFS-based datastore.

To edit the name of the datastore:

1. Log into the VI Client, and select a server from the inventory panel.
2. Click the **Configuration** tab, and click **Storage (SCSI, SAN, and NFS)**.
3. Select the datastore whose name you want to edit, and click the **Properties** link.

   The Volume Properties dialog box appears.



4. Under General, click **Change**.

   The Properties dialog box opens.
5. Enter the new datastore name, and click **OK**.

Adding Extents to Datastores

You can expand a datastore that uses the VMFS format by attaching a hard disk partition as an extent. The datastore can span over 32 physical storage extents.

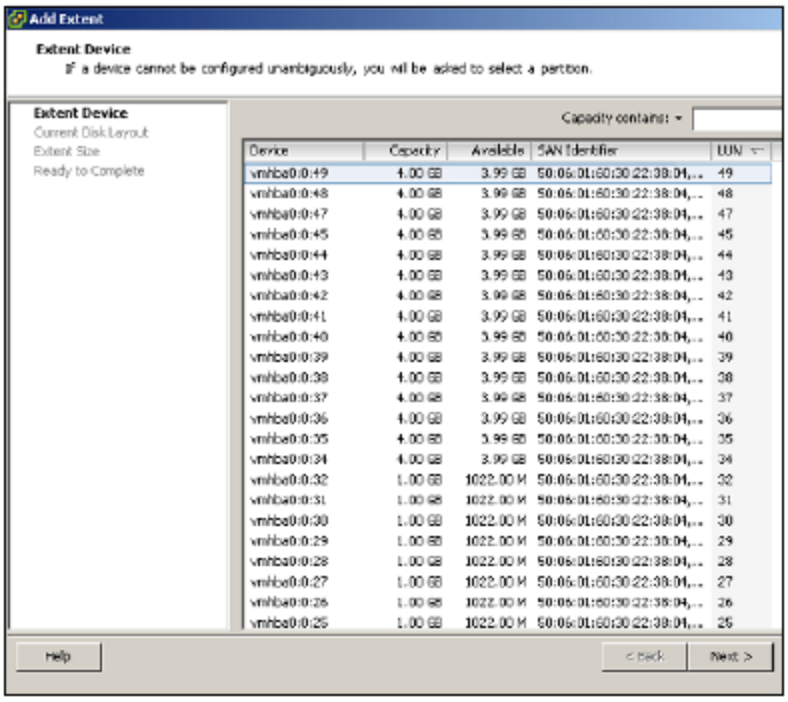
You can dynamically add the new extents to the datastore when you need to create new virtual machines on this datastore, or when the virtual machines running on this datastore require more space.

To add one or more extents to the datastore:

1. Log into the VI Client, and select a server from the inventory panel.
2. Click the **Configuration** tab, and click **Storage (SCSI, SAN, and NFS)**.
3. Select the datastore you want to expand, and click the **Properties** link.

   The Volume Properties dialog box appears.
4. Under Extents, click **Add Extent**.

   The Add Extent Wizard opens.

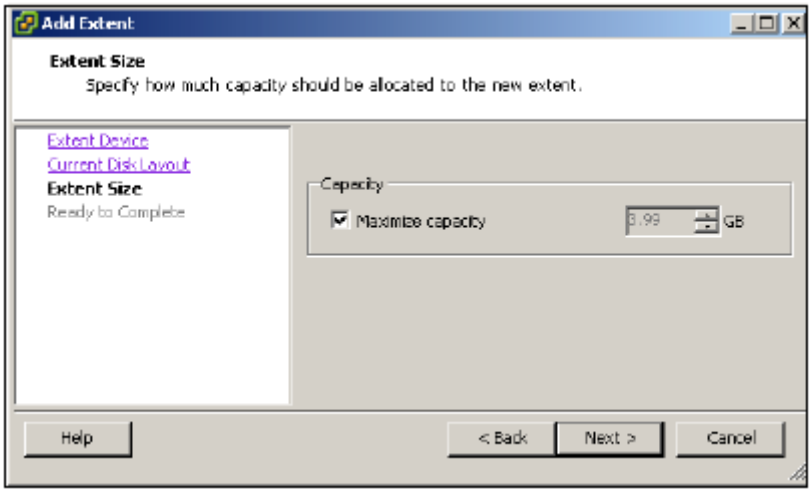


5. Select the disk you want to add as the new extent and click **Next**.

   The Current Disk Layout page appears.
6. Review the current layout of the disk you are using for the extent to make sure the disk does not contain any important information.

   NOTE: If a disk or partition you add was formatted previously, it will be reformatted and lose the file systems and any data it contained.
7. Click **Next**.

   The Extent Size page appears.



8. Set the capacity for the extent.

   By default, the entire free space on the storage device is available.

9. Click **Next**.

   The Ready to Complete page appears.

10. Review the proposed extent layout and the new configuration of your datastore; then click **Finish**.

NOTE: Adding extents is supported for SCSI, FC, and iSCSI storage volumes only.

Removing Existing Datastores

Using the VI Client, you can remove a datastore that you do not use, so it will not be used as storage for virtual machines.

CAUTION: Removing a datastore from the ESX system breaks the connection between the system and the storage device that holds the datastore, and stops all functions of that storage device. You cannot remove a datastore if it holds virtual disks of a currently running virtual machine. To remove the datastore from the ESX host, you must mask or remove the LUN from the SAN array, and then rescan from the VI Client.

To remove a datastore:

1. Power down all virtual machines that use the datastore you want to remove.
2. Log into VI Client, and select a server from the inventory panel.
3. Click the **Configuration** tab, and click **Storage (SCSI, SAN, and NFS)** to display all storage devices.
4. Select the datastore you want to remove, and click **Remove**.
5. Confirm that you want to remove the datastore.
6. From the storage array management software, unmap the datastore (or the LUN) from ESX hosts that were previously using the datastore.
7. Click **Rescan** to update the view of available storage options.

Editing Existing VMFS Datastores

Datastores that use the VMFS format are deployed on SCSI-based storage devices. After you create a VMFS-based datastore, you can modify it. See the following sections for more information:

- "VMFS Versions" on page 113
- "Upgrading Datastores" on page 113

VMFS Versions

VMware ESX offers the following versions of VMFS:

- **VMFS-2** — This file system is created with VMware ESX 2.x.
- **VMFS-3** — This file system is created with VMware ESX 3. VMFS-3 enhancements include multi-directory support. A virtual machine must reside on a VMFS-3 file system before an ESX 3 host can power it on.

Access of the two VMFS datastore versions is different, based on the ESX host version. The following table provides a summary of access operations available to VMware ESX 2.x and 3.x.

Table 6-1. Host Access to VMFS File Systems

Host	VMFS-2 Datastore	VMFS-3 Datastore
ESX 2.x host	Read/Write (runs virtual machines)	No access
ESX 3 host	Read only (copies virtual machines)	Read/Write (runs virtual machines)

Upgrading Datastores

As described in the previous section, ESX 3 includes a new file system, VMFS version 3 (VMFS-3). If your datastore was formatted with VMFS-2, you can read files stored on VMFS-2, but you are not able to use them. To use the files, upgrade VMFS-2 to VMFS-3.

When you are upgrading VMFS-2 to VMFS-3, the ESX file-locking mechanism ensures that no remote ESX host or local process is accessing the VMFS volume being converted. VMware ESX preserves all files on the datastore.

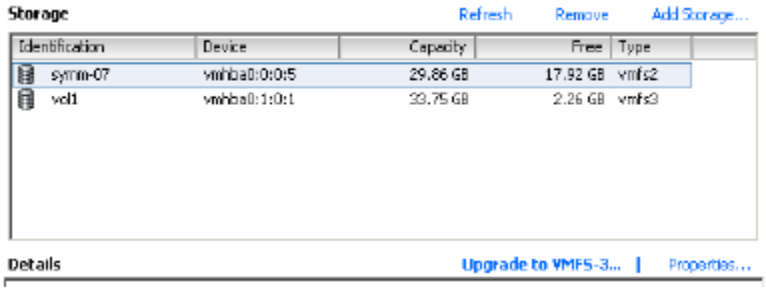
As a precaution, before using the upgrade option, consider the following:

- Commit or discard any changes to virtual disks in the VMFS-2 volume you want to upgrade.
- Back up the VMFS-2 volume you want to upgrade.
- Be sure no powered-on virtual machines are using this VMFS-2 volume.
- Be sure no other ESX host is accessing this VMFS-2 volume.
- Be sure this VMFS-2 volume is not mounted on any other ESX host.

CAUTION: The VMFS-2 to VMFS-3 conversion is a one-way process. After converting the VMFS-based datastore to VMFS-3, you cannot revert back to VMFS-2.

To upgrade the VMFS-2 to VMFS-3:

1. Log into the VI Client, and select a server from the inventory panel.
2. Click the **Configuration** tab, and click **Storage (SCSI, SAN, and NFS)**.
3. Click the datastore that uses the VMFS-2 format.



4. Click **Upgrade to VMFS-3**.

Adding SAN Storage Devices to VMware ESX

Creating Datastores on SAN Devices

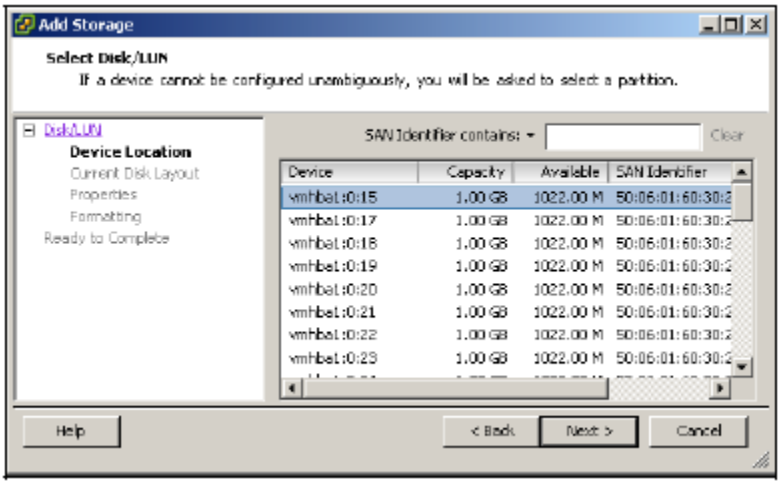
When you create a datastore on an FC storage device, the Add Storage Wizard guides you through the configuration. Before creating a new datastore on an FC device, rescan an FC adapter to discover any newly added LUNs. For more information, see “Performing a Rescan” on page 116.

To create a datastore on a SAN device:

1. Log into the VMware VI Client, and select a server from the inventory panel.
2. Click the **Configuration** tab, and click **Storage (SCSI, SAN, and NFS)** under hardware.
3. Click the **Add Storage** link.

   The Select Storage Type page appears.
4. Select the disk/volume storage type and click **Next**.

   The Select Disk/LUN page appears.



5. Select the FC device you want to use for your datastore and click **Next**.

   The Current Disk Layout page appears.

6. Look over the current disk layout and click **Next**.

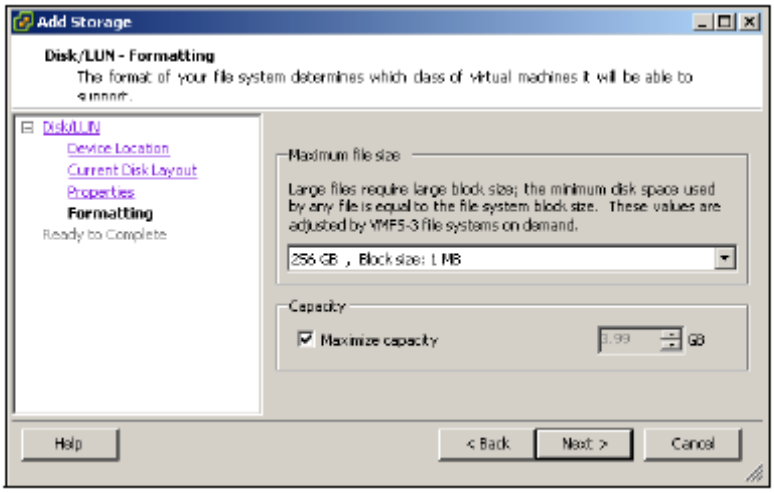
   The Disk/LUN–Properties page appears.

7. Enter a datastore name.

   The datastore name appears in the VI Client and must be unique within the current VMware Infrastructure instance.

8. Click **Next**.

   The Disk/LUN–Formatting page appears.



9. If you need to, adjust the file system values and capacity you use for the datastore.

   By default, the entire free space available on the storage device is offered to you.

10. Click **Next**.

The Ready to Complete page appears.

11. Review the datastore information, and click **Finish**.

    This process creates the datastore on an FC disk for the ESX host.

12. Perform a rescan.

For advanced configuration using multipathing, masking, and zoning, see the VMware *SAN Configuration Guide*.

Performing a Rescan of Available SAN Storage Devices

If a new LUN becomes accessible through the host bus adapter, then an ESX rescan registers this new storage device for use. If an existing LUN is no longer used and has been removed from the SAN, then VMware ESX removes it from the configuration.

Consider performing a rescan when:

- Any changes are made to the LUNs available to your ESX system.
- Any changes are made to storage host bus adapters.
- New datastores are created.
- Existing datastores are edited or removed.

In addition, you should also perform a rescan each time you:

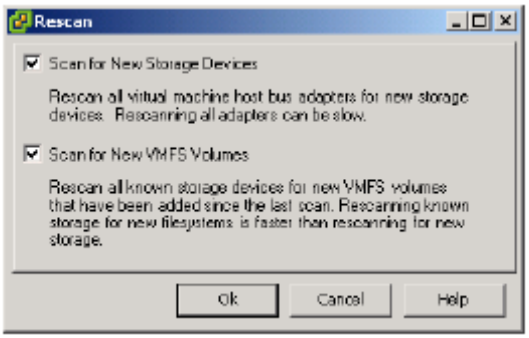
- Zone a new disk array on the SAN to an ESX host.
- Create new LUNs on a SAN disk array.
- Change the LUN masking on an ESX host disk array.

CAUTION: Do not rescan when a path is out of service. As stated above, if one path fails, the other takes over, so your system continues to be fully functional. If, however, you rescan at a time when a path is not available, the ESX host removes the path from its list of available paths to the device. The ESX host will not use the path until the next time a rescan is performed while the path has been restored to active service.

To perform a rescan:

1. In the VI Client, select a host, and click the **Configuration** tab.
2. In the Hardware panel, choose **Storage Adapters**, and click **Rescan** above the Storage Adapters panel.

   The Rescan dialog box opens.



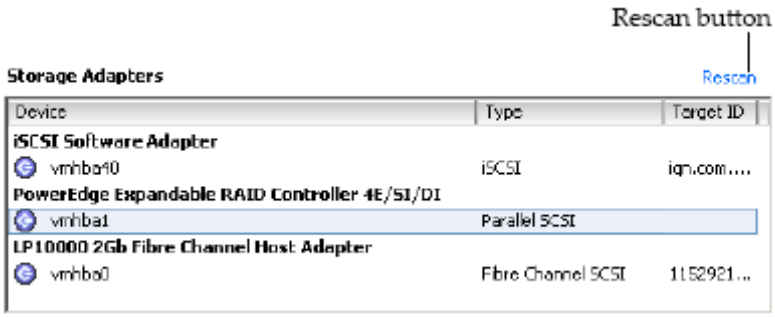
3. To discover new disks or LUNs, select **Scan for New Storage Devices**, and click **OK**.

   If new LUNs are discovered, they appear in the disk/LUN list.

4. To discover new datastores, select **Scan for New VMFS Volumes**, and click **OK**.

   If new datastores or VMFS volumes are discovered, they appear in the datastore list.

NOTE: From the Storage Adapter display, you can also select an individual adapter and click **Rescan** to rescan just that adapter.



Advanced LUN Configuration Options

This section discusses a number of advanced configuration options:

- "Changing the Number of LUNs Scanned Using Disk.MaxLUN" on page 117
- "Masking Volumes Using Disk.MaskLUN" on page 118
- "Changing Sparse LUN Support Using DiskSupportSparseLUN" on page 119

Changing the Number of LUNs Scanned Using Disk.MaxLUN

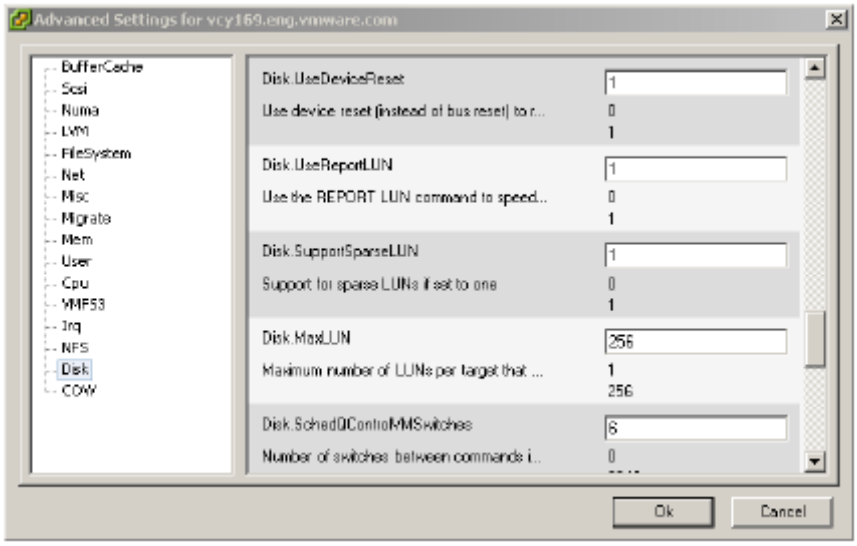
By default, the VMkernel scans for LUN 0 to LUN 255 for every target (a total of 256 volumes). You can change the Disk.MaxLun parameter to change this number. This change might improve LUN discovery speed.

NOTE: You cannot discover volumes with LUN ID numbers higher than 255. Reducing the value can shorten both rescan time and boot time. The time to rescan LUNs depends on several factors, including the type of storage array and whether

sparse LUN support is enabled. See "Changing Sparse LUN Support Using DiskSupportSparseLUN" on page 119.

To change the value of Disk.MaxLUN:

1. In the VI Client's inventory panel, select the host, click the **Configuration** tab, and click **Advanced Settings**.
2. In the dialog box that appears, select **Disk**.
3. Scroll down to **Disk.MaxLUN**, change the existing value to the desired value, and click **OK**.



Masking Volumes Using Disk.MaskLUN

The Disk.MaskLUN parameter allows you to mask specific volumes on specific HBAs. Masked volumes are not touched or accessible by the VMkernel, even during initial scanning. This is the equivalent of host-based LUN masking.

Use this option when you want to prevent the ESX system from accessing some FC volumes, but do not want to use the FC switch or FC device volume masking mechanisms.

To change the value of Disk.MaskLUN:

1. In the VI Client's inventory panel, select the host, click the **Configuration** tab, and click **Advanced Settings**.
2. In the dialog box that appears, select **Disk**.
3. Scroll down to Disk.MaskLUN and change the existing value to the desired value.

   For example, if you want if you want to mask out volume IDs 1 to 4, 6, and 8 to 25, enter the following:

   vmhba1:1:1-4, 6, 8-25
4. Click **OK**.

CAUTION: If a target, LUN, or vmhba number changes because of a server or SAN reconfiguration, the incorrect LUN may be masked or exposed.

Changing Sparse LUN Support Using DiskSupportSparseLUN

By default, the VMkernel is configured to support sparse LUNs—that is, a case where some LUNs in the range 0 to n-1 are not present, but LUN n is present.

If all LUNs are sequential, you can change the Disk.SupportSparseLUN parameter. This change decreases the time needed to scan for LUNs.

Managing Multiple Paths for Fibre Channel LUNs

VMware ESX supports multipathing to maintain a constant connection between the server machine and the storage device in case of the failure of an HBA, switch, SP, or cable. Multipathing support does not require specific failover drivers. To support path switching, the server typically has two or more HBAs available, from which the storage array can be reached using one or more switches. Alternatively, the setup can include one HBA and two SPs so that the HBA can use a different path to reach the disk array.

For more information on how multipathing works, see "Multipathing and Path Failover" on page 29. Also see the VMware *SAN Configuration Guide*. For information on managing paths, see the following sections.

- "Viewing the Current Multipathing State" on page 119
- "Active Paths" on page 121
- "Setting Multipathing Policies for SAN Devices" on page 121
- "Disabling and Enabling Paths" on page 123
- "Setting the Preferred Path (Fixed Path Policy Only)" on page 124
- "Managing Paths for Raw Device Mappings" on page 125

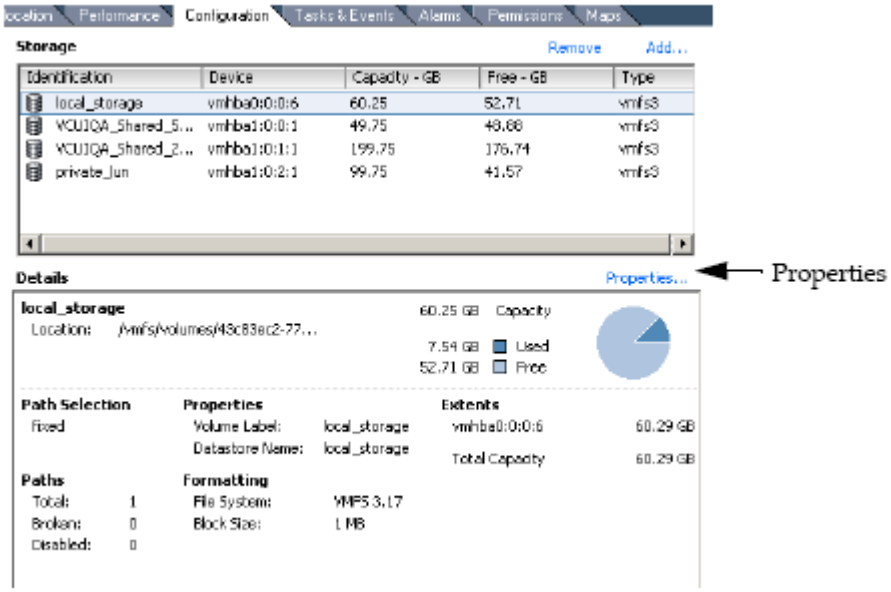
Viewing the Current Multipathing State

Use the VI Client to view the current multipathing state.

To view the current multipathing state:

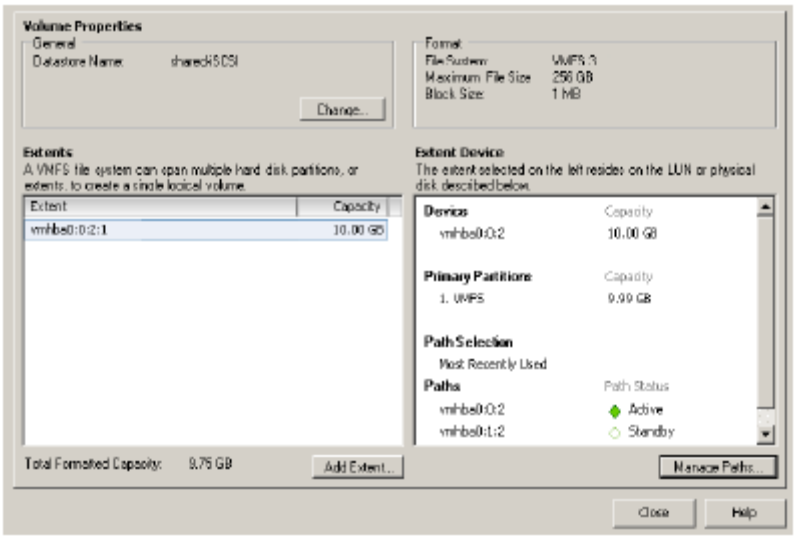
1. Log into the VI Client, and select a server from the inventory panel.
2. Click the **Configuration** tab, and click **Storage (SCSI, SAN, and NFS)** under hardware.
3. From the list of configured datastores, select the datastore whose paths you want to view or configure.

   Information about the datastore appears in the Details panel.



4. To view additional information, or to change the multipathing policy, click the **Properties** link located above the Details panel.

The Volume Properties dialog box for this datastore opens.



The Extent Device panel displays information about the extent, the path selection algorithm, the available paths, and the active path.

5. In the Extents panel, select the extent for which you want to view or change information.

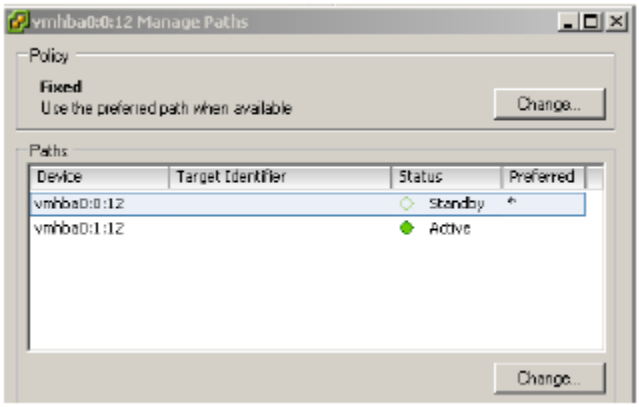
The Extent Device panel in the dialog box includes information on the status of each path to the storage device. The following path status might be displayed:

- **Active** — The path is working and is the current path being used for transferring data.
- **Disabled** — The path has been disabled and no data can be transferred.

- **Standby** — The path is working but is not currently being used for transferring data.
- **Dead** — The software cannot connect to the disk through this path.

6. Click **Manage Paths** to open the Manage Paths Wizard.

   If you are using the Fixed path policy, you can see which path is the preferred path. The preferred path is marked with an asterisk (*) in the fourth column.



You can use the Manage Paths Wizard to enable or disable your paths, set multipathing policy, and specify the preferred path.

Active Paths

VMware ESX does not perform I/O load balancing across paths for a given storage device. At any one time, only a single path is used to issue I/O to storage device. This path is known as the active path.

- If the path policy of a storage device is set to Fixed, VMware ESX selects the path marked Preferred as the active path.
- If the preferred path is disabled or unavailable, the ESX system uses an alternate working path as the active path.
- If the path policy of a storage device is set to most recently used, the ESX host selects an active path to the storage device that prevents path thrashing. The preferred path designation is not considered.

In some SAN terminology, the term **active** means any path that is available for issuing I/O to a volume. From the ESX host's point of view, the term **active** means the one and only path that the ESX host is using to issue I/O to a volume.

Setting Multipathing Policies for SAN Devices

The following multipathing policies are currently supported:

- **Fixed** — The ESX host always uses the preferred path to the disk when that path is available. If it cannot access the disk through the preferred path, then it tries the alternate paths. Fixed is the default policy for active/active storage devices.
- **Most Recently Used** — The ESX host uses the most recent path to the disk, until this path becomes unavailable. That is, the ESX host does not automatically

revert back to the preferred path. Most Recently Used is the default policy for active/passive storage devices and is required for those devices.

- **Round Robin** – This option is available as an experimental feature in ESX 3.5 and is available to configurations that are mapped to active/active storage array types, as described earlier. VMware ESX will use each available path, in turn, to perform load distribution of host based I/O operations.

The ESX host automatically sets the multipathing policy according to the make and model of the array it detects. If the detected array is not supported, it is treated as active/active. For a list of supported arrays, see the SAN Compatibility Guide.

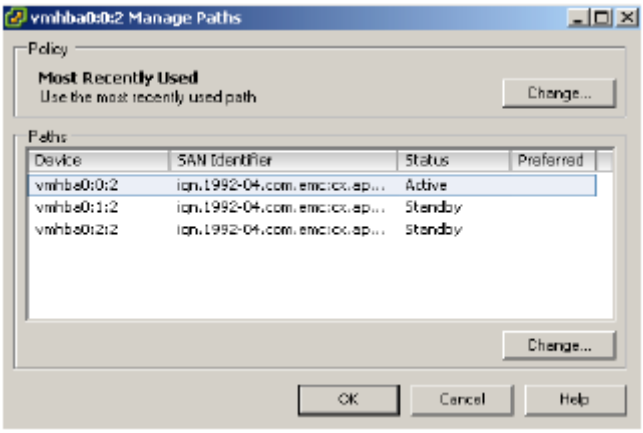
NOTE: Manually changing Most Recently Used to Fixed is not recommended. The system sets this policy for those arrays that require it.

To set the multipathing policy:

1. In the VI Client's inventory panel, select the host, and click the **Configuration** tab.
2. In the Hardware panel, select **Storage**.
3. Select the datastore for which you want to change the multipathing policy, and click **Properties** in the Details panel.
4. In the Extent panel, select the device for which you want to make the change, and click **Manage Paths** in the Extent Device panel on the right.

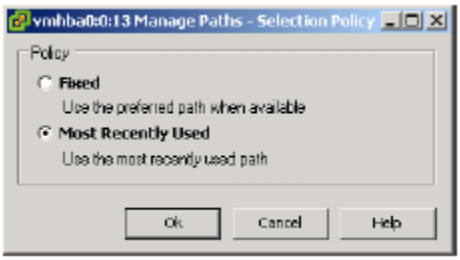
   NOTE: If you are managing paths for RDMs, see "Managing Paths for Raw Device Mappings" on page 125.

   The Manage Paths dialog box shows the list of different paths to the disk, with the multipathing policy for the disk and the connection status for each path. It also shows the preferred path to the disk.



5. Under **Policy**, click **Change**.

   The Selection Policy page opens.



6. Select the desired multipathing policy in the dialog box that appears:
   - Fixed
   - Most Recently Used
   - Round Robin
7. Click **OK**, and click **Close** to save your settings and return to the Configuration page.

Disabling and Enabling Paths

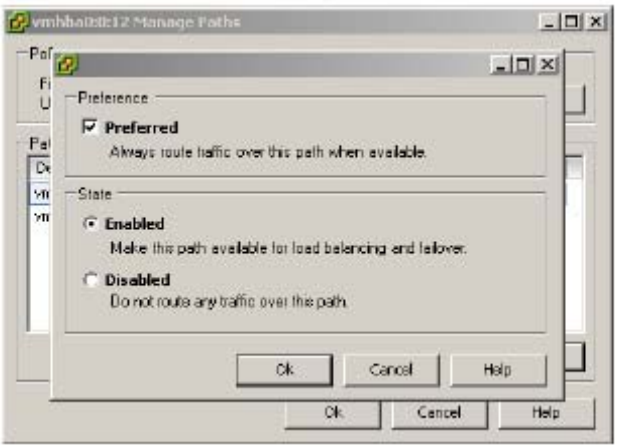
If you need to temporarily disable paths for maintenance or any other reasons, you can do so using the VI Client.

To disable a path:

1. In the VI Client's inventory panel, select the host, and click the **Configuration** tab.
2. In the Hardware panel, select **Storage**.
3. Select the datastore for a path you want to disable, and click **Properties** in the Details panel.
4. In the Extent panel, select the device for which you want to make the change, and click **Manage Paths** in the Extent Device panel on the right.

   NOTE: If you are managing paths for RDMs, see "Managing Paths for Raw Device Mappings" on page 125.

   The Manage Paths dialog box appears.
5. Under Paths, select the path you want to disable, and click **Change**.



6. Select the **Disable** radio button to disable the path.
7. Click **OK** twice to save your changes and exit the dialog boxes.

To enable a path:

The procedure for enabling a path is the same as for disabling (see "To disable a path" on page 123), except you select the **Enable** radio button.

Setting the Preferred Path (Fixed Path Policy Only)

If you set path policy to Fixed, the server always uses the preferred path when it is available. ESX systems can also be configured to load balance traffic across multiple HBAs to multiple volumes with active/active arrays. To do this, assign preferred paths to your LUNs so that your HBAs are being used evenly. For example, if you have two volumes (A and B) and two HBAs (X and Y), you can set HBA X to be the preferred path for volume A, and HBA Y as the preferred path for volume B, thus maximizing utilization of your HBAs. Path policy must be set to Fixed for this case.

To set the preferred path:

1. In the VI Client's Inventory panel, select the host, and click the **Configuration** tab.
2. In the Hardware panel, select **Storage**.
3. Select the datastore for a path you want to change, and click **Properties** in the Details panel.
4. In the Extent panel, select the device for which you want to make the change, and click **Manage Paths** in the Extent Device panel on the right.

   The Manage Paths dialog box appears.

   NOTE: If you are managing paths for RDMs, see "Managing Paths for Raw Device Mappings" on page 125.
5. Under Paths, select the path you want to make the preferred path, and click **Change**.
6. In the Preference panel, click **Preferred**.

   If Preferred is not an option, make sure that the path policy is Fixed.

7. Click **OK** twice to save your settings and exit the dialog boxes.

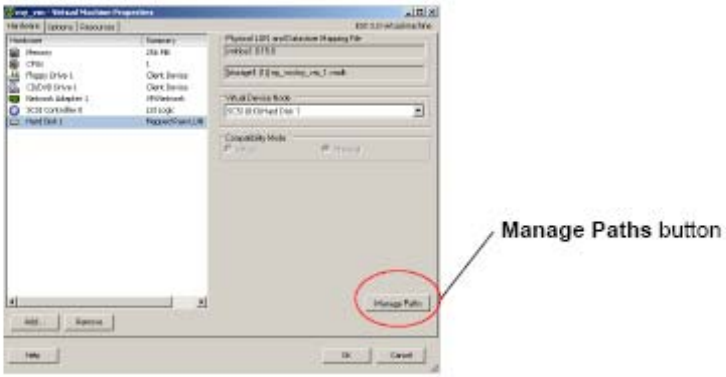
Managing Paths for Raw Device Mappings

You use the Manage Paths Wizard to manage paths for RDMs and mapped raw LUNs.

To manage paths:

1. Log on as administrator or as the owner of the virtual machine to which the mapped disk belongs.
2. Select the virtual machine from the inventory panel.
3. On the Summary tab, click the **Edit Settings** link.

   The Virtual Machine Properties dialog box opens.
4. On the Hardware tab, select **Hard Disk,** and click **Manage Paths**.



   The Manage Paths Wizard opens.
5. Use the Manage Paths Wizard to enable or disable your paths, set multipathing policy, and specify the preferred path.

   NOTE: Refer to the sections "To set the multipathing policy" on page 122, "To disable a path" on page 123, "To enable a path" on page 124, and "To set the preferred path" on page 124 for more information on specific path management operations.

7 Growing VMware Infrastructure and Storage Space

Inevitably, as your business and virtual infrastructure grows, you will also outgrow the original processing and storage capacity you planned for your VMware Infrastructure. You need the flexibility to quickly add capacity, and rearrange or reallocate space based on current or future business requirements. This chapter provides more information on the considerations for growing VMware Infrastructure and storage, including the various options available, and the implications of those options. These options include:

- Growing your storage capacity — Adding extents to datastores, adding volumes to ESX hosts, and expanding storage by spanning VMFS (maximum of 32).
- Using templates to deploy new virtual machines — Creating and using operating system templates for future ESX and virtual machine cloning.
- Providing bandwidth on-demand for load balancing.
- Adding new CPU and memory resources to your virtual machine — Allocating new CPUs and memory resources for existing or new virtual machines.
- Adding more servers to existing VMware Infrastructure — Scaling your infrastructure to allow storage and server expansion.

The specific topics covered in this chapter are the following:

VMware Infrastructure Expansion Basics

As your business operations expand, you will likely need to increase application capacity (storage, CPU, or memory) to expand your virtual infrastructure. When expanding your infrastructure, you need to consider the following:

- How to expand storage. Consider adding more storage or more virtual machines to your existing servers before adding more ESX hosts.
- Creating and using operating system templates for future ESX and virtual machine cloning.
- Providing bandwidth on-demand for load balancing.
- How to allocate new CPUs and memory resources for existing or new virtual machines.
- Adding more servers to existing VMware Infrastructure and scaling your infrastructure to allow storage and server expansion.

Expanding your virtual infrastructure means that you need to access more storage, share data with multiple applications, increase virtual machine capability to service data requests faster, and add more work load (that is, more virtual machines) to service increased customer demands.

As your virtual infrastructure grows, so does the number of virtual infrastructure components you need to manage. With more components comes more complexity. Managing this complexity as you scale your virtual infrastructure requires advanced planning and knowledge of how each of the different components of the architecture scale. There are also many questions to answer. How do I plan for adding networks? Where do I put my templates for rapid provisioning? How many volumes do I need to hold my virtual machines? These are just some of the questions you might have as you consider growing your virtual infrastructure.

With VMware Infrastructure, growing your environment using the right methodology can minimize performance impact to mission critical applications and, at the same time, expand your capability to service more applications, virtual machines, and, ultimately, more customers. VMware Infrastructure helps you meet another goal, which is expanding your infrastructure with the least amount of time and effort.

When designing your virtual infrastructure, plan for expansion and scalability to anticipate future and long-term business growth. Too often, short-term goals drive architectural decisions. You might plan your infrastructure capacity and number of hosts based on current requirements to consolidate all of your existing servers. However, by the time the original consolidation is finished, you might end up needing more physical hosts to accommodate new workloads that have come about since the time you started. To prevent this problem from occurring, you might want to plan additional capacity in the design of your virtualization environment to begin with, anticipating growth of up to 40 percent of new servers.

Similarly, you might start off with a great design, with internal storage and only a few network cards in each host. However, new projects that are not forecasted can creep into your requirements and put an increased burden on your environment. For example, suppose you have a new line of business from an acquisition that needs three new networks. Maybe you have been using dedicated network cards on a small

form factor server. Now there is no room to put new network cards and you are not set up for VLAN tagging. How do you cope with this new requirement?

Customers typically like the idea of breaking down their datacenter environment into logical units. Do not get too granular, though. Creating a datacenter for each line of business might organize your servers nicely and provide for easy lines of delineation (and prevent political battles over ownership). However, it might not be the most efficient architecture as you continue to scale. You might end up with over- and under-utilized datacenters. Perhaps a more suitable strategy is to leverage resource pools for your organization and keep datacenter utilization at a higher level.

As a designer, think about what you want out of your datacenter, but have never had, and design around that. Virtualization frees up time from low-value tasks such as provisioning, patching, and failover configuration, so you can have more time for high-value tasks such as change management, trend analysis, and workload profiling.

Growing Your Storage Capacity

Some customers start off with internal storage in their initial virtual infrastructure designs. While usable initially, this simply does not scale. Moving to shared storage later can be time consuming and have considerable impact on your virtual machine availability. By including shared storage in your original design, you can save money in your hardware budget in the long run. Think about investing early in shared storage for your environment.

What type of shared storage is best? (Note that some ESX features are not available with every type of shared storage.) While VMware ESX supports NAS storage, you cannot use VMware HA or VCB with this type of storage. VCB is available only with FC storage. iSCSI typically costs considerably less than FC, so unless you are running I/O-intensive workloads, iSCSI can also provide a good solution.

You do not have to use the same storage type throughout your environment. For example, lower cost iSCSI or NAS might be your choice for storing templates or archived virtual machines. Think of your virtual machine use and design your storage solutions appropriately.

As your business grows, you can easily expand your storage capacity using the following three options:

1. Expand your existing volumes zoned to ESX hosts by using array software management agent to add extents to existing volumes and rescan.

   NOTE: After using array software management agent to add an extent, users need to rescan their ESX hosts.

2. Map new storage volumes to ESX hosts and rescan.

3. With existing volumes already assigned to an ESX host, span an existing volume to concatenate it with other volumes to make a larger single volume.

When adding more storage to your environment, consider disk parameters to keep them compatible. Choose disk characteristics that are close to the existing disk structures already assigned to ESX hosts or your specific application running inside a virtual machine to minimize performance impact. For example, consider using the same RAID level, seek-time, access-time, spindle speed, and MTBF.

The best time to rescan ESX hosts is when there is minimal amount of I/O traffic on the incoming and outgoing SAN fabric ports between an ESX host and the array storage port processors. (The level of I/O traffic varies for different user environments.) To determine a minimum and maximum level of I/O traffic for your environment, you need to first establish a record of I/O activity for your environment. Record I/O traffic patterns in the SAN fabric ports (for example, using a command such as `portperfshow`, for Brocade switches). Once you have determined that I/O traffic has dropped to 20 percent of available port bandwidth, for example, by measuring traffic on the SAN fabric port where a HBA from an ESX host is connected, you should rescan the ESX host to minimize interruptions to the running virtual machines.

With VMware Infrastructure 3, you can add a virtual volume while virtual machines are still running. You can add new space to your virtual machine as it needs more space. From the virtual machine (using VirtualCenter), just select **Edit Properties** and add a storage disk to your virtual machine.

Adding Extents to Datastores

When you add extents from the array management software, you must rescan the ESX host to detect the additional storage. Also, check with your SAN array vendor to see if adding extents using array software is supported with VMware ESX.

Adding Volumes to ESX Hosts

Adding new physical volumes to an ESX host is as simple as mapping new volumes to the ESX Server HBAs and rescanning the ESX host. (Be aware of LUN limits and of disrupting the environment during a rescan. For more information on the maximum number of LUNs per system, see the VMware *SAN Hardware Compatibility Guide*.) After adding a new volume, you have the option to add it to an existing VMFS volume using the **Add extent** option. For RDM storage expansion, you can use available array management software.

Storage Expansion – VMFS Spanning

Volume spanning for ESX hosts is the least favorable option for SAN administrators, for two reasons. First, since we cannot determine if all files related to one virtual machine are all on a single physical extent, we also cannot know whether the loss of an extent will result in the loss of the virtual machine or its data. Second, a VMFS partition being used by a guest operating system, or serving as a data disk with pending I/O, cannot be spanned while the virtual machine is in a powered-on state. Volumes are concatenated and not striped, so there is no improvement to performance. In ESX 2.5.x and older versions, breaking the spanned partition also breaks the VMFS and data can be lost if one of the spanned volumes goes down for any reason. These problems are resolved in VMware Infrastructure 3. If you are using ESX 2.5.x or older versions, plan carefully if you need to remove one of the volumes being spanned.

A VMFS volume can be spanned to a maximum of 32 physical storage extents, including SAN volumes and local storage. Since each volume can have a block size up to 8MB (which allows for a maximum volume size of 2TB), 32 physical storage extents provide a maximum volume size of 64TB. It is recommended that you span SAN volumes together, or span local storage together, if choosing span options. In

other words, do not span SAN volumes together with local storage, because SCSI and disk types behind SAN storage arrays have different disk parameters. In addition, other hosts are not able to access the local extents of shared volumes. Spanning allows pooling of storage and flexibility in creating the storage volume necessary for your virtual machine.

Using Templates to Deploy New Virtual Machines

Using templates in virtual infrastructure provides significant savings in the time required to deploy a virtual machine. All virtual machines are built exactly the same, so administrators can reduce the amount of time spent creating virtual machines and spend their valuable time on higher-value tasks.

There are some challenges in learning to use templates. For example, where do you put the template store? With internal disks, you are stuck copying virtual machines across the network. You could leverage technology such as NAS or iSCSI to move templates, but you are still copying them across the network. If your templates are stored on FC storage, the VMware Virtual Infrastructure Server tells the storage processor in the SAN to do a block copy in order to transport and deploy the template. This step eliminates LAN network traffic and ultimately provides the fastest way to deploy a template.

Managing Storage Bandwidth

With active/active SAN storage arrays, you can manually load balance the I/O traffic on a per-volume or virtual machine basis. For disk-intensive I/O applications, consider moving the virtual machines hosting these applications (residing on a single volume or multiple volumes) to use a dedicated SAN path. In general, if you observe that more than 95 percent of the fabric port bandwidth is used, you should move I/O traffic from disk I/O-intensive applications to a dedicated SAN path or paths. You can determine the amount of bandwidth available by using the `portperfshow` command (on a Brocade switch). Then, using VirtualCenter or the options for the `esxcfg-mpath` command, you can move I/O traffic for a specific volume to an available path that carries less traffic.

Adding New CPU and Memory Resources to Virtual Machines

One of the biggest issues in scaling virtual infrastructure is how to size a virtual machine appropriately. If you oversize virtual machines, you have fewer resources available to service other workloads in your environment. Unlike physical servers, virtual machines are dynamic. You can easily change the resources allocated to a virtual machine, and you do not even have to visit your datacenter.

Be conservative with the amount of memory you give to a virtual machine. Transparent page sharing used by ESX hosts greatly reduces the amount of physical memory a virtual machine uses.

One of the biggest factors limiting the ability to scale is the use of SMP for virtual machines that do not actually leverage SMP. Not all applications are multithreaded, so they can take advantage of multiple processors. In fact, few applications benefit from using SMP configuration. By giving a second, third, or fourth processor to applications that are not capable of using them, you actually decrease the performance of the virtual machine and prevent other virtual machines from performing to their full capacity. Make sure you do not set false performance ceilings when you scale your environment. Check with your independent software vendor (ISV) to verify that your application is multithreaded before assigning additional CPU resources to virtual machines.

CPU Tuning

CPU tuning for virtual machine support and specifying multiprocessing resources to run virtual machines are fairly complicated operations. Although CPU tuning is relatively easy to do, in some situations increasing CPU utilization can actually result in decreased system and application performance.

Before trying out techniques for tuning the CPU to improve performance, you need to understand virtual machine operation on individual processors. Starting with version 2.1 of VMware ESX, VMware introduced VSMP or Virtual SMP™ (symmetric multiprocessing). In version 3 of VMware ESX, SMP support was extended to allow four-CPU implementations within a virtual machine. VSMP was designed to broaden the reach of virtual machines to include more production environment applications. Some larger applications, such as Microsoft SQL Server, Oracle RDBMS, and Microsoft Exchange, can take advantage of multiple processors (although more with regard to threading than actual raw throughput). In practice, however, most applications do not scale linearly across multiple processors. In fact, most applications available are not multithreaded and so do not gain any increased performance from adding more processors.

VSMP was designed to allow multiple threads (one per virtual CPU) to be scheduled across one or more physical CPUs. VSMP requires that two virtual CPUs are scheduled on two separate physical CPUs (PCPUs) or, in the case of hyperthreading on the server, logical CPUs (LCPUs). Since both processors have to be scheduled at the same time and kept in sync, invariably one of the processors runs idle and wastes time in the operating system and in the underlying VMkernel. (Processors that are scheduled to run, but are idle, are accounted for with the CPU %WAIT parameter.)

To evaluate situations in which CPU tuning can improve virtual machine performance with systems providing SMP and VSMP support, consider the following key points on determining optimal use of CPUs:

- If you see unbalanced %WAIT times in your normal application workloads, you should adjust your virtual machine so that it uses only a single virtual CPU.
- If you see balanced %WAIT times and the numbers are relatively low, you have used VSMP appropriately in your environment.

Using SMP or VSMP for every virtual machine is not the answer to solving every performance issue, and can actually raise performance issues. If you have SMP-aware applications, like databases or other applications that can use multiple processing, then using VSMP can help your performance. Before deploying in a production environment, you might want to perform tests with different applications

in a lab configuration to select the best number of processors used for specific applications.

Many of the support requests that VMware receives concern performance optimization for specific applications. VMware has found that a majority of performance problems are self-inflicted, with problems caused by misconfiguration or less-than-optimal configuration settings for the particular mix of virtual machines, host processors, and applications deployed in the environment. To avoid problems and optimize application performance and CPU utilization in VMware Infrastructure 3, business functions using the resource pools feature. Resource pools, which are essentially CPU, memory, and disk space resource silos, can help provide guaranteed levels of performance for virtual machines and applications. Resource pools extend the concepts of shares, limits, and reservations from an individually-managed virtual machine to a group of virtual machines. The following section describes shares, limits, and reservations as applied to resource pools and to individual virtual machines.

Resource Pools, Shares, Reservations, and Limits

Resource pools group virtual machines together so that **shares**, **reservations**, or **limits** for CPU or memory accessed by a resource pool can be set for all its virtual machines collectively. You can also set shares, reservations, and limits on individual virtual machines.

Shares determine the priority of virtual machines for access to CPU resources only when there are not enough resources to meet demand. (The same principle holds true for memory access.) To guarantee or limit resources for a virtual machine or a group of virtual machines, when there is no resource contention, you must use reservations and limits.

- **Reservations** specify the amount of a resource guaranteed to a virtual machine.
- **Limits** specify the most resources that a virtual machine can use, regardless of what is available.

For more details on how to allocate appropriate CPU and memory resources for your environment, see “Lab 3804: Performance Troubleshooting” (and the corresponding lab manual) available from http://www.vmware.com/vmtn/vmworld/. Also see the VMware *Resource Management Guide*.

Adding More Servers to Existing VMware Infrastructure

When you investigate server provisioning for your virtual infrastructure, you might find that you can reuse a lot of your existing. When you look for servers to reuse, choose those servers that can handle larger amounts of memory. If it is going to take a fork-lift upgrade to add memory to a server, it's probably better to choose a different server to which you can add memory more easily. Also think about servers that will allow you to add more network adapters. Older servers are limited in their network adapter capacity. They also have limited expansion space for new cards (network adapters and FC cards). Just because you have freed up a server through consolidation does not mean it is the best server to use for consolidation.

When buying new hardware, choose hardware that has 64-bit extensions in the processor, such as the Intel EM64T or AMD Opteron processors. Not only do these processors support larger memory, but also they provide faster access to the memory and can run newer 64-bit applications. Consider the expansion capabilities of any new server hardware you buy. Many VMware customers choose the larger-form factor servers (4RU), instead of the dense-form factor servers (2RU), just to get the extra expansion slots in the larger chassis. The larger chassis usually include more DIMM slots, so you use less expensive DIMMs to get the same amount of memory as servers with a denser form factor.

Consider buying newer dual- and quad-core processors. You get a performance boost plus save money on VMware licenses, since VMware charges by the socket.

8 High Availability, Backup, and Disaster Recovery

An important aspect of designing an enterprise-class datacenter for VMware Infrastructure is providing the appropriate level of high availability, backup, and disaster recovery services needed to handle demands and be able to respond to disasters and system failures.

This chapter describes the different disasters and failures that customers face, and the high availability, backup strategies and technologies used to recover from these situations. Customers need to deal with both planned and unplanned disasters or failures. A planned disaster or failure, such as downtime resulting from hardware upgrades or major infrastructure change, is disruptive but easy to recover from). An unplanned disaster, such as hardware failure, can result in lost data.

The following topics are covered in this chapter:

Overview

High availability (HA) means maintaining service or access to data as much of the time as is humanly and technologically possible. Discussion of numbers in the range of 99.99 percent (four 9s) to 99.999 percent (five 9s) is common in systems providing HA. Ultimately, customers do not care if their environment has achieved 99.999 percent availability if a critical application goes down during that 0.001percent time window. Thus, it is crucial to have a strategy that not only provides 99.999 percent availability but also is capable of servicing customers whose applications are unfortunately in that 0.001 percent time frame. The goal in those cases is to bring service back up as soon as possible after a failure occurs.

Business continuity (BC) is another term commonly used in discussions of enterprise IT and business operations. BC is not the same as disaster recovery, as BC includes disaster recovery strategies as part of planning disaster recovery for IT infrastructure. Specific to IT infrastructure, disaster recovery and BC planning strives to maintain access to data continuously, with no downtime. Disaster recovery with no downtime is achievable only by using redundant storage components and redundant SAN infrastructure mirroring and clustering technologies. It is outside the scope of this chapter to discuss all facets of BC that encompass every aspect of keeping a business operational during a disaster.

This chapter discusses disaster recovery for IT infrastructure, but not other non-IT aspects of recovery and business continuity planning. It describes a strategic paradigm for planning and implementing HA solutions for disaster recovery that leverage both existing VMware and other industry technologies. Using all available technologies in delivering HA solutions provides the best chance to achieve 100 percent customer satisfaction.

At the most fundamental level, backup strategies preserve data. Data is the most important asset in almost all business sectors and technologies. The second most important requirement for businesses is continuous access to data. You can have slow access to data, but you cannot lose data. Planning and understanding what VMware supports regarding backup technologies can save you time, money, and legal consequences (in the event you actually lose essential data).

Instead of looking immediately at backup strategies, start by getting a high-level overview of what exactly disaster recovery means and where backup strategies can be applied. Disaster recovery can be divided into two main categories: planned and unplanned. Backing up is an important part of planned disaster recovery. A planned disaster is a disaster that you can predict will happen in the future although, the majority of time, you do not know exactly when it will occur. Instances when you can more accurately predict events can be characterized as planned disasters, such as when you need to perform hardware upgrades or change the infrastructure for better manageability.

A hardware failure is an example of an unplanned disaster; you must have a plan to recover from this type of disaster. A failure or disaster can have a narrowly focused or localized effect, or it can affect an entire site. Thus, you need one plan for a localized disaster (for example, an HBA failure or a cable failure) and another plan for a site disaster (such as a building fire, flood, earthquake, or other environmental disaster).

Consider the following technologies when building your Virtual Infrastructure and designing a planned disaster recovery strategy:

- VMware VMotion
- Cloning in VMware Infrastructure
- Snapshots and suspend in VMware Infrastructure
- RAID technologies
- Data replication technologies
- Backup strategies
- VMware VCB solutions
- SAN extensions
- VMware DRS

Consider the following technologies when building your Virtual Infrastructure for unplanned disaster recovery:

- VMware multipathing
- VMware HA
- Data replication technologies
- SAN extensions

Some data replication technologies and SAN extensions are listed as options for both planned and unplanned disaster recovery. Planning and implementing these options is described in later sections of this chapter.

Planned Disaster Recovery Options

This section describes options to consider for planned disaster recovery:

- VMotion
- Cloning in VMware Infrastructure
- Snapshot and suspend/resume in VMware Infrastructure
- RAID technologies
- Industry replication technologies
- Industry backup applications
- VMware Consolidated Backup (VCB)
- Industry SAN extension technologies
- VMware DRS

Planned DR Options with VMware VMotion

VMotion technology enables intelligent workload management. VMotion allows administrators to migrate virtual machines manually to different hosts.

Administrators can migrate a running virtual machine to a different physical server connected to the same SAN, without service interruption. VMotion makes it possible to:

- Perform zero-downtime maintenance by moving virtual machines, so the underlying hardware and storage can be serviced without user disruption.
- Continuously balance workloads across the datacenter to use resources most effectively in response to changing business demands.

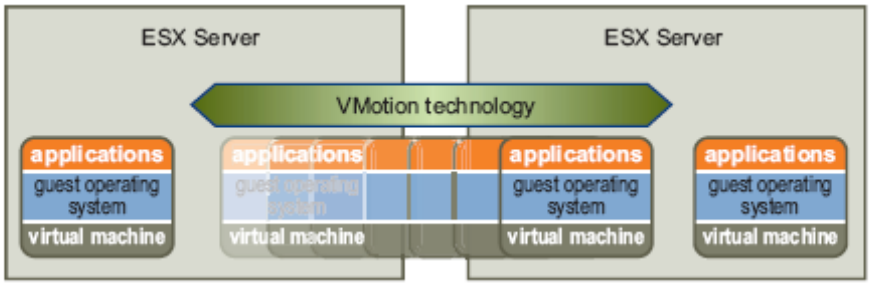


Figure 8-1. Migration with VMotion

Because VMotion operations are not automatic, virtual machine migration between servers is considered a planned disaster recovery. When you plan to upgrade your server environment or need to replace a server component, VMotion can migrate your working environment for continuous application access.

Planned DR Options with Cloning in VMware Infrastructure

Creating templates (and cloning virtual machines) for future virtual machine deployment is considered a planned disaster recovery solution. Using templates is a viable and cost-effective option, since templates are also generally included as part of any planning and strategy to grow your virtual infrastructure using VMware Infrastructure.

Having a planned disaster solution means that if a disaster strikes, you are ready to recovery quickly. You can deploy virtual machines quickly so that application downtime is minimized or eliminated. Using templates provides the fastest and most efficient method for recreating and deploying virtual machines.

Viruses can quickly bring down an operating system. If you do not have templates to create and deploy virtual machines, you can spend hours troubleshooting to determine which viruses attacked your environment, locating the viruses, and removing them. The hours lost in downtime and troubleshooting can translate to millions of dollars that your business cannot afford. Instead of spending hours on recovery, you can use templates to deploy new operating systems and virtual machines in minutes.

Planned DR Options with Snapshots in VMware Infrastructure

Snapshot software allows an administrator to make an instantaneous copy of any single virtual machine. Snapshot software is available from several different sources:

- Software included with VMware ESX allows you to create basic snapshots of virtual machines.
- Third-party backup software might allow for more comprehensive backup procedures and additional configuration options than those available in VMware ESX. For example, some third-party software packages allow you to set backup and recovery policies according to specific user requirements on memory or data storage.

Administrators make snapshots for a variety of reasons, including:

- Backup
- Disaster recovery
- Availability of multiple configurations, versions, or both
- Forensics (looking at a snapshot to find the cause of problems while your system is running)
- Data mining (looking at a copy of performance data to find ways to reduce load on production systems)

If you decide to use SAN snapshots to back up your data, consider the following points:

- Some vendors support snapshots for both VMFS and RDMs. If both are supported, you can make either a snapshot of the whole virtual machine file system for a host, or snapshots for the individual virtual machines (one per disk).
- Some vendors support snapshots only for setups using RDM. If only RDM is supported, you can make snapshots of individual virtual machines. See your storage vendor's documentation for additional information.

Planned DR Options with Existing RAID Technologies

RAID technologies can be thought of as planned DR solutions within an array. This solution is not viable if the entire IT center goes down, however, affecting the entire disk array unit. To maximize data availability, consider appropriate RAID levels that can still maintain data availability when a disk drive or drives fail.

Planned DR Options with Industry Replication Technologies

Replication technologies, such as mirroring a volume, duplicating or cloning a volume, or creating a volume snapshot at the SAN management application layer, can be used to backup data generated from VMware Infrastructure. Whichever technology you use, you need more storage space to store these new volumes. Please check with your OEM array vendor if a specific data replication application has been tested and qualified with VMware Infrastructure. White papers, jointly produced

by the OEM vendor and VMware, list the specific steps of how to use this technology for your environment. Please check www.vmware.com for a list of available white papers and other documentation on VMware products.

Planned DR Options with Industry Backup Applications

If you are using third-party backup software, you must first make sure that software is supported with ESX hosts. See the *Backup Compatibility Guide* at http://www.vmware.com/pdf/esx_backup_guide.pdf for more information.

Using third-party software has the advantage of a uniform environment. However, you have to consider that the additional cost of the third-party snapshot software can become higher as your SAN grows.

If you decide to use snapshots to back up your data, you must consider the following points:

- Some vendors support snapshots for both VMFS and RDMs. If both are supported, you can make either a snapshot of the whole virtual machine file system for a host, or snapshots for the individual virtual machines (one per disk).
- Some vendors support snapshots only for a setup using RDM. If only RDM is supported, you can make snapshots of individual virtual machines.

See your storage vendor's documentation for additional information.

NOTE: ESX systems also include a VCB component, which is discussed in detail in the VMware *Virtual Machine Backup Guide*.

Backups in a SAN Environment

Within the SAN environment, backups have two goals. The first goal is to archive online data to offline media. This process is repeated periodically for all online data on a time schedule. The second goal is to provide access to offline data for recovery from a problem. For example, database recovery often requires retrieval of archived log files that are not currently online.

Scheduling a backup depends on a number of factors:

- Identification of critical applications that require more frequent backup cycles within a given period of time.
- Recovery point and recovery time goals. Consider how precise your recovery point needs to be, and how long you are willing to wait for it.
- The rate of change (RoC) associated with the data. For example, if you are using synchronous/asynchronous replication, the RoC affects the amount of bandwidth required between the primary and secondary storage devices.
- Overall impact on SAN environment, storage performance (while backing up), and other applications.
- Identification of peak traffic periods on the SAN (backups scheduled during those peak periods can slow both the applications and the backup itself).
- Time to schedule all backups within the datacenter.
- Time it takes to back up an individual application.

- Resource availability for archiving data—usually offline media access (tape).

Include a recovery-time objective for each application when you design your backup strategy. That is, consider the time and resources necessary to re-provision the data. For example, if a scheduled backup stores so much data that recovery requires a considerable amount of time, you need to re-examine the scheduled. Perhaps you should perform the backup more frequently, so that less data is backed up at a time and the recovery time decreases.

If a particular application requires recovery within a certain time frame, the backup process needs to provide a time schedule and specific data processing to meet this requirement. Fast recovery can require the use of recovery volumes that reside on online storage to minimize or eliminate the need to access slow, offline media for missing data components.

Choosing Your Backup Solution

When choosing your backup solution, consider that a backup can be one or all of these:

- Crash consistent
- File system consistent
- Application consistent

VMware offers a file-system-consistent backup. In most cases, a file-system-consistent backup allows you to recover completely from failure. However, if your applications require synchronization across file systems or with a database, the VMware solution might not provide enough consistency. In these cases, you should investigate a third-party backup solution to see whether it better suits your needs.

Array-Based Replication Software

SAN administrators usually use specialized array-based software for backup, disaster recovery, data mining, forensics, and configuration test. ESX administrators might be used to working with tools included with the ESX host for performing the same operations. When you use an ESX system in conjunction with a SAN, you need to decide whether array-based or host-based tools are more suitable for your situation.

Array-Based (Third-Party) Solution

When considering an array-based solution, be aware of the following points:

- If you use the array-based solution, pass-through RDM (not VMFS) is usually the appropriate choice.
- Array-based solutions usually result in more comprehensive statistics. With RDM, data always go along the same path, resulting in easier performance management.
- Security tends to be more manageable when you use RDM and an array-based solution because with RDM, virtual machines more closely resemble physical machines.

File-Based (VMware) Solution

When considering a file-based solution using VMware tools and VMFS, be aware of the following points:

- Using VMware tools and VMFS is better for provisioning: one large volume is allocated and multiple `.vmdk` files can be placed on that volume. With RDM, a new volume is required for each virtual machine.
- Snapshots and replication (within a datacenter) is included with your ESX host at no extra cost. The file-based solution is therefore more cost effective than the array-based solution.
- For most ESX administrators, using VMware ESX tools is easier.
- ESX administrators who use the file-based solution are more independent from the SAN administrator.

Performing Backups with VMware VCB

VMware Consolidated Backup addresses most of the problems you encounter when performing traditional backups. VCB allows you to:

- Reduce the load on your ESX systems by moving the backup tasks to one or more dedicated backup proxies.
- Avoid congesting and overloading the datacenter network infrastructure by enabling LAN-free backup.
- Eliminate the need for a backup window by moving to a snapshot-based backup approach.
- Simplify backup administration by making the deployment of backup agents optional in each virtual machine you back up.
- Back up virtual machines that are powered off.

For more information on setting up and using VCB, see the VMware *Virtual Machine Backup Guide*.

Planned DR Options with Industry SAN-Extension Technologies

There are existing technologies that can bridge two SAN fabrics in different geographical locations. This capability provides long-distance support beyond the distance limitation of Fibre Channel (using extended long wave cable and wavelength division multiplexing to provide connection between 100 km to 200 km). For disaster recovery, this capability provides a perfect opportunity for you to move data from one location to another for backup purposes.

One example of this technology is FCIP, Fibre Channel over IP, which provides an expansion to the existing SAN fabric (see Figure 8-2). When two SAN fabrics in two different geographical locations are connected via FCIP protocol, they become or merge into one large SAN fabric. FCIP protocol provides a tunneling technology to bridge FC protocols from separate SAN fabrics together over an existing WAN network. Because the SAN fabric is now merged between two physical locations, any changes in one location, such as fabric zones or changes to ISLs, propagates to the

other. It might not be viable to keep this connection over the long term. For the short term, however, it provides an excellent opportunity to implement planned disaster recovery.

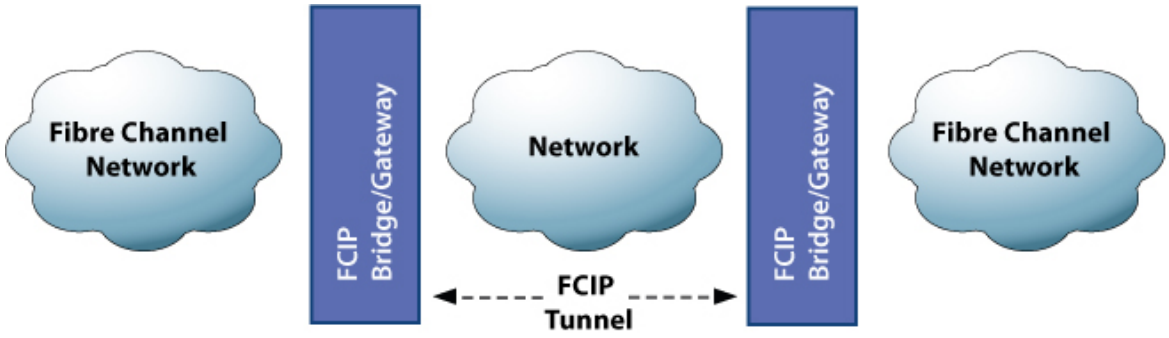


Figure 8-2. FCIP Protocol Connecting SAN Fabrics in Different Geographic Locations

You can implement a FCIP connection as part of a planned disaster recovery solution to connect two SAN fabrics in two separate geographical locations. (Refer to www.searchstorage.com for information on available vendors deploying this technology.) Because the two SAN fabrics are merged into one single fabric, you can map initiators from one SAN fabric to targets on a different geographical location. Once you have defined mapping of initiators to targets, you can copy sensitive data to your new targets for backup, or use data replication technologies belonging to a new array to which you did not have access before.

NOTE: The rate of change for application data is a factor in deciding the bandwidth requirements of the WAN connecting the two fabrics. The latency that can be tolerated is another factor in the choice of the WAN.

Another technology, similar to FCIP, is iFCP. iFCP is another protocol to bridge two SAN fabrics from two geographical locations (see Figure 8-3). Unlike FCIP, that merges two fabrics together, iFCP provides network address translation to route FC protocol frames from one location to another via an existing WAN network. (Refer to www.searchstorage.com for information on vendors deploying this technology.) When the two geographical dispersed SAN fabrics are connected, you can access more storage devices onto which you can move critical data for safekeeping. After the move of critical data has been completed, you can disconnect the long distance connections to minimize any extra bandwidth charges.

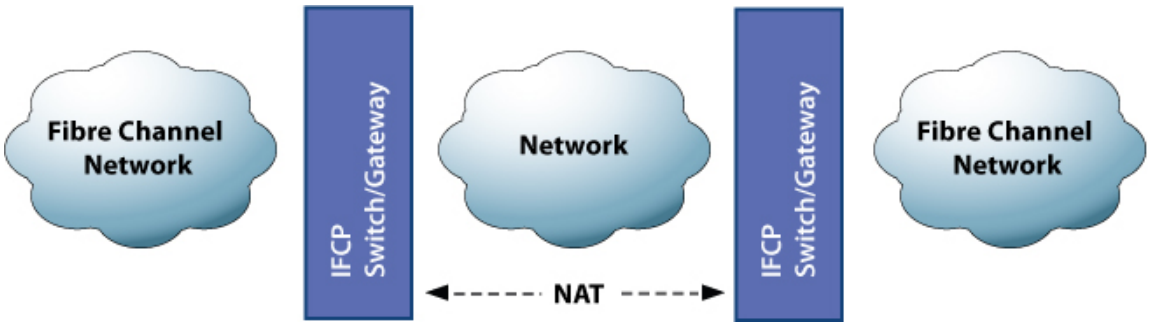


Figure 8-3. iFCP Protocol Connecting SAN Fabrics in Different Geographic Locations

Planned DR Options with VMware DRS

Because VMware DRS redistributes its virtual machines to other servers automatically, you can use this technology to recover from CPU, memory, and entire physical server failures.

Unplanned Disaster Recovery Options

This section describes options to consider for unplanned disaster recovery:

- VMware multipathing
- VMware HA
- Industry replication technologies
- SAN extensions

Unplanned DR Options with VMware Multipathing

VMware ESX supports multipathing to maintain a constant connection between the server machine and the storage device in case of the failure of an HBA, switch, SP, or FC cable. Multipathing support does not require specific failover drivers. For more information on how multipathing works, see “Multipathing and Path Failover” on page 29. Also see the VMware *SAN Configuration Guide*.

You can choose a multipathing policy of either Fixed or Most Recently Used for your system. If the policy is Fixed, you can specify a preferred path. Each volume that is visible to the ESX host can have its own path policy. See “Viewing the Current Multipathing State” on page 119 for information on viewing the current multipathing state and on setting the multipathing policy.

NOTE: Virtual machine I/O might be delayed for at most sixty seconds while failover takes place, particularly on an active/passive array. This delay is necessary to allow the SAN fabric to stabilize its configuration after topology changes or other fabric events. In the case of an active/passive array with the path policy Fixed, path thrashing may be a problem. See “Resolving Path Thrashing Problems” on page 182.

Unplanned DR Options with VMware HA

VMware HA allows you to organize virtual machines into failover groups. When a host fails, all its virtual machines are immediately started on different hosts. High availability requires SAN shared storage. When a virtual machine is restored on a different host, it loses its memory state but its disk state is exactly as it was when the host failed (crash-consistent failover). See the VMware *Resource Management Guide* for detailed information.

NOTE: You must be licensed to use VMware HA, which is offered as a feature in VMware Infrastructure 3 Enterprise, or can be purchased as a separate add on to VMware Infrastructure 3.

Unplanned DR Options with Industry Replication Technologies

In addition to using data replication for planned down times, you can also think of it as an unplanned disaster recovery option. This is possible because some OEM vendors support the concept of physical array cluster failover. Consider the situation in which an array in site A belongs to a cluster of arrays in site B location. Both sites maintain a heart beat so, as with any traditional cluster implementation, when one site goes down, the remote site automatically becomes active.

NOTE: Please check with your OEM array vendor if a specific cluster failover solution has been tested and qualified with VMware Infrastructure. White papers, jointly produced by the OEM vendor and VMware, list the specific steps of how to use this technology for your environment.

Unplanned DR Options with SAN Extensions

As with FCIP and iFCP, you can think of a permanent connection a design to implement unplanned disaster recovery. However, you must plan to provide a cluster of initiators and targets that are located in both geographical locations. This option might not be economically feasible, but nonetheless viable to consider for short-term connection if you foresee a disaster approaching. For example, if a hurricane were coming to your area, establish this connection and plan your redundancies of initiators and targets for the merged fabric. After the hurricane passes, you can disconnect the two fabrics, saving your data and establishing a short-term plan for when a major environmental disaster is approaching.

Considering High Availability Options for VMware Infrastructure

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 SPs, 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 volumes. 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, 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 volume failure allows applications to survive storage access faults. Mirroring can accomplish that protection.

  Mirroring designates a second non-addressable volume that captures all write operations to the primary volume. Mirroring provides fault tolerance at the volume level. Volume mirroring can be implemented at the server, SAN switch, or storage array level.
- **Duplication of SAN Environment** — For extreme HA 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 might share operational workloads, or the second SAN environment might be a failover-only site.

Using Cluster Services

Server clustering is a method of tying two or more servers together using a high-speed network connection so that the group of servers functions as a single, logical server. If one of the servers fails, the other servers in the cluster continue operating, picking up the operations performed by the failed server. For more information on configuring and using clustering, see "Using Cluster Services" on page 83. Also see the VMware document, *Setup for Microsoft Cluster Service*, for more information

Designing for Server Failure

The RAID architecture of SAN storage inherently protects you from failure at the physical disk level. A dual fabric, with duplication of all fabric components, protects the SAN from most fabric failures. The final step in making your whole environment failure resistant is to protect against server failure. This section briefly discusses ESX system failover options.

Server Failover and Storage Considerations

For each type of server failover, you must consider storage issues:

- Approaches to server failover work only if each server has access to the same storage. Because multiple servers require a lot of disk space, and because failover for the storage array complements failover for the server, SANs are usually employed in conjunction with server failover.
- When you design a SAN to work in conjunction with server failover, all volumes that are used by the clustered virtual machines must be seen by all ESX hosts. This is counterintuitive for SAN administrators, but is appropriate when using virtual machines.

Note that just because a volume is accessible to a host, all virtual machines on that host do not necessarily have access to all data on that volume. A virtual machine can access only the virtual disks for which it was configured. In case of a configuration error, virtual disks are locked when the virtual machine boots so no corruption occurs.

NOTE: When you are using ESX boot from SAN, each boot volume or LUN should, as a rule, be seen only by the ESX system that is booting from that volume. An exception is when you are trying to recover from a crash by pointing a second ESX system to the same volume. In this case, the SAN volume in question is not really a boot from SAN volume. No ESX system is booting from it because it is corrupted. The SAN volume is a regular non-boot volume that is made visible to an ESX system.

Planning for Disaster Recovery

If a site fails for any reason, you might need to recover the failed applications and data immediately from a remote site. The SAN must provide access to the data from an alternate server to start the data recovery process. The SAN might handle the site data synchronization.

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

Failover

Path failover refers to situations when the active path to a volume is changed from one path to another, usually because of a SAN component failure along the current path. A server usually has one or two HBAs, and each HBA sees one or two storage processors on a given SAN array. You can determine the active path— the path currently used by the server—by looking at the volume's properties.

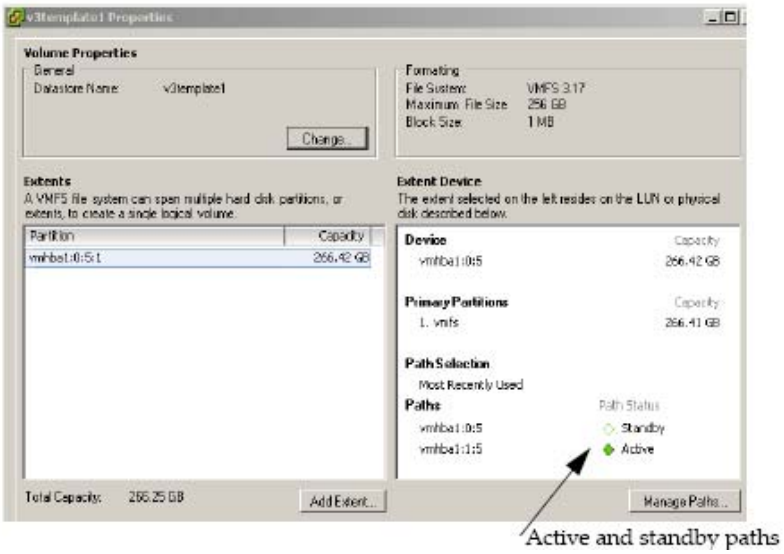


Figure 8-4. Active and Standby Paths

When an FC cable is pulled, I/O might pause for 30 to 60 seconds until the FC driver determines that the link is down, and until failover has occurred. As a result, the virtual machines (with their virtual disks installed on SAN storage) can appear unresponsive. If you attempt to display the host, its storage devices, or its adapter, the operation might appear to hang. After failover is complete, I/O resumes normally.

In case of disastrous events that include multiple breakages, all connections to SAN storage devices might be lost. If none of the connections to the storage device is working, some virtual machines might encounter I/O errors on their virtual SCSI disks.

Setting the HBA Timeout for Failover

The timeout value for I/O retry operations is usually set in the HBA BIOS driver. (You might also want to change the operating system timeout, as discussed in “Setting Operating System Timeout” on page 148.)

VMware recommends that you set the timeout value to 30 seconds:

- For QLogic HBAs, the timeout value is 2*n + 5 seconds, where n is the value of the PortDownRetryCount parameter of the BIOS of the QLogic card. You can change the path-failure detection time by changing the value of the module parameter qlport_down_retry (whose default value comes from the BIOS setting). The recommended setting for this parameter is 14.
- For Emulex HBAs, you can modify the path-failure detection time by changing the value of the module parameters lpfc_linkdown_tmo (the default is 30) and lpfc_nodedev_tmo (the default is 30). The sum of the values of these two parameters determines the path-failure detection time. The recommended setting for each is the default.

To change these parameters, you must pass an extra option, such as qlport_down_retry or lpfc_linkdown_tmo, to the driver. The following section explains how you can pass these options to the driver.

Setting Device Driver Options for SCSI Controllers

To set device driver options for QLogic, Emulex, or other SCSI card drivers:

1. Back up the file `/etc/vmware/esx.conf` and open it for editing. The file includes a section for each SCSI device, as in the following example:

```
/device/002:02.0/class = "0c0400"
/device/002:02.0/devID = "2312"
/device/002:02.0/irq = "19"
/device/002:02.0/name = "Company Name Corp CNQ31xx/32xx (rev 02)"
/device/002:02.0/options = ""
/device/002:02.0/owner = "vmkernel"
/device/002:02.0/subsysDevID = "028c"
/device/002:02.0/subsysVendor = "2910"
/device/002:02.0/vendor = "2038"
/device/002:02.0/vmkname = "vmhba0"
```

2. Find the options line right under the name line, and modify it as appropriate.
3. Repeat these steps for every SCSI adapter that is controlled by the same driver if needed.

Alternatively, you can perform the following steps to set device driver options:

1. Back up the file `/etc/vmware/esx.conf`
2. Run the following command:

```
esxcfg-module –s <whatever option supported from the device
driver> = <numeric value> <driver name>.o
```

   For example:

```
esxcfg-module –s <vendor>_cmd_timeout=20 <FC driver>.o
```

   This example changed the command timeout to 20 seconds for a FC driver such as qla2300_7xx.o or lpfcdd_732.o

3. Run the command:

```
esxcfg-boot –b
```

4. Reboot the ESX host.

Setting Operating System Timeout

You might want to increase the standard disktimeout value so that a Windows guest operating system is not extensively disrupted during failover. You can set the operating system timeout for Windows 2000 and Windows Server 2003 guest operating systems using the registry.

1. Back up your Windows registry.
2. Select **Start** > **Run**, type **regedit.exe**, and click **OK**.
3. In the left panel hierarchy view, double-click **HKEY_LOCAL_MACHINE**, then **System**, then **CurrentControlSet**, then **Services**, and finally **Disk**.
4. Select **TimeOutValue** and set the data value to **x03c** (hexadecimal) or **60** (decimal).

After you have made this change, Windows waits at least 60 seconds for delayed disk operations to complete before it generates errors.

5. Click **OK** to exit the Registry Editor.

VMware Infrastructure Backup and Recovery

Backup, restoration, and disaster recovery are among the most critical processes of datacenter management. VMware ESX and VMware Infrastructure provide many different methods, each suitable for a specific environment, to perform backup and restore tasks.

There is no best method other than the one that satisfies the need of the present configuration and that it is fully tested and supported by VMware. Each method has advantages and disadvantages that are outlined in this section.

Backup Concepts

The following concepts are essential for understanding backup procedures:

- **Differential backup** — Backs up only those files that have changed since the last full backup.
- **File-level backup** — A type of backup that is defined at the level of files and folders.
- **Full backup** — Backs up all selected files.
- **Full virtual machine backup** — Backs up all files that comprise the entire virtual machine. These files include disk image `.vmdk` files, `.vmx` files, and so on.
- **Image-level (volume-level) backup** — Backs up an entire storage volume.
- **Incremental backup** — Backs up only those files that have changed since the last backup, whether it is a full or incremental backup.
- **Quiescing** — A process of bringing the on-disk data of a physical or virtual computer into a state suitable for backups. This process might include operations such as flushing dirty buffers from the operating system's in-memory cache to disk, or other higher-level, application-specific tasks.
- **VCB proxy** — In the context of VMware Consolidated Backup, VCB proxy is a physical machine running Microsoft Windows 2003, VCB, and third-party backup software that is used to perform LAN-free, file-level and image-level virtual machine backups.

Backup Components

When you perform a backup, the following three components of backup software are generally involved in the process:

- **Backup client (backup agent)** – A program that scans virtual machine file systems and transfers data to be backed up to a backup server. During restore operations, the backup client writes the data into the file systems.
- **Backup server** – A program that writes the data, pushed by the backup client, to a backup medium such as a robotic tape library. During the restore operation,

the backup server reads the data from the backup medium and pushes it to the backup client.

- **Scheduler** – A program that allows you to schedule regular automatic backup jobs and coordinate their execution. Backups can be scheduled at periodic intervals, or individual files can be automatically backed up immediately after they have been updated.

Depending on where you run each of the components, you can choose different approaches as described in the next section.

Backup Approaches

Each of the backup software components can be run in a virtual machine, on the service console, or on a VCB proxy running Microsoft Windows 2003. While the location of the scheduler is not important, the locations of the backup server and backup client are important.

Depending on where you want to run each component, you can choose from one of the following approaches:

- **Traditional backup approach** – You deploy a backup client to every system that requires backup services. You can then regularly perform automatic backups. With this approach, several methodologies exist. Choose a method that best suits your needs and requirements.

  For more information, see the section, “Using Traditional Backup Methods.”

- **VMware Consolidated Backup** – VCB enables offloaded backup for virtual machines running on ESX hosts. This approach lets you use the virtual machine snapshot technology and SAN-based data transfer in conjunction with traditional file-based backup software.

  When running VCB, you can back up virtual machine contents from a centralized Microsoft Windows 2003 backup proxy rather than directly from the ESX system. Utilizing a backup proxy reduces the load on the ESX system, allowing it to run more virtual machines.

For more information on VCB, see “VMware Consolidated Backup” on page 14.

Using Traditional Backup Methods

With the traditional backup methods, you deploy a backup agent on each host whose data needs to be secured. Backups are then conducted regularly in an automated way.

The backup agent scans the file system for changes during periods of low system use and sends the changed information across the network to a backup server. That server then writes the data to a backup medium, such as a robotic tape library.

You can back up your service console and virtual machines using traditional methods. Keep in mind the following:

- To be able to capture the data in its consistent state, perform backups at the times of the lowest activity on the network and when your computer resources are mostly idle. While performing backups, you might need to take critical applications off line.

- Make sure that network bandwidth between the servers you are backing up and the backup server is sufficient.
- With a large number of servers, both physical and virtual, allocate enough resources to manage backup software on each host. Remember that managing agents in every virtual machine is time consuming.

What to Back Up

Within the ESX environment, you need to back up the following major items regularly:

- **Virtual machine contents** — Because virtual machines are used frequently, critical information stored in their disk files constantly changes. As with physical machines, virtual machine data needs to be backed up periodically to prevent corruption and loss due to human or technical errors.

  The virtual machine data you back up includes such content as virtual disks, raw device mappings (RDM), and configuration files. The two backup images you need to be concerned with are:

  - Data disk image belonging to each virtual machine (VMFS partitions containing application data and other data)
  - Boot disk image of each virtual machine (the VMFS operating system partition)
- **VMware ESX service console** — VMware ESX command-line management interface that provides ESX management tools and a command prompt for more direct management of VMware ESX. It also keeps track of all the virtual machines on the server and their configurations.

  NOTE: In earlier releases, the service console was the main interface to the ESX host. With ESX 3 and later, the VI Client has priority, although you still might use the service console to perform some advanced administration operations.

  During its lifetime, the service console does not experience any major changes other than periodic upgrades. In case of a failure, you can easily recover the state of your service console by reinstalling VMware ESX. Therefore, although you might consider backing up the service console, you do not need to back it up as frequently as the virtual machines and their data.

Generally, use the following backup schedule for your virtual machines:

- At the image level, perform backups periodically for Windows, and nightly for Linux. For example, back up a boot disk image of a Windows virtual machine once a week.
- At the file level, perform backups once a day. For example, back up files on drives D, E, and so on every night.

Backing Up Virtual Machines

Depending on your needs and available resources, you might choose one of the traditional methods for backing up your virtual machines. Traditional backup methods do not use VCB.

Traditional backup methods offer the following options:

- Run backup clients from within a virtual machine performing file-level or image-level backups. As long as you are backing up over the network, no compatibility guide is needed.
- Run backup clients from the VMware ESX service console, backing up virtual machines in their entirety.
- **NOTE**: Virtual machine files for ESX 2.5.1 and later releases have the `.dsk` extension. For ESX 2.5 or older, the virtual machine files have the `.vmdk` extension.
- Back up virtual machine data by running a backup server within a virtual machine that is connected to a tape drive or other SCSI-based backup media attached to the physical system.

If you decide to use SAN snapshots to back up your data, consider the following points:

- Some vendors support snapshots for both VMFS and RDMs. If both are supported, you can make either a snapshot of the whole virtual machine file system for a host, or snapshots for the individual virtual machines (one per disk).
- Some vendors support snapshots only for a setup using RDM. If only RDM is supported, you can make snapshots of individual virtual machines. See your storage vendor's documentation for additional information.

In a modern datacenter environment, it has become increasingly difficult to apply the traditional approach to your backup processes. Using it might cause a number of problems, some of which are described in the previous section on using traditional backup methods. To avoid many of the problems, consider using VMware Consolidated Backup.

VMware Backup Solution Planning and Implementation

To back up and restore successfully in an ESX environment, system administrators should be aware of specific guidelines. Two primary areas for consideration are physical disk configuration and storage location of ESX software and virtual machines. Although VMware supports ESX operation in a boot-from-SAN environment, it is recommended that local storage be used for VMware ESX installations. Local storage should be of a RAID array (RAID 1, for example). This ensures data redundancy and quick recovery in case of failure. It also takes less time to recover a local boot image if there is a problem than to troubleshoot a boot-from-SAN issue. Boot disk images and data disks for all virtual machines should be located on a RAID array on SAN. See Figure 8-5 below.

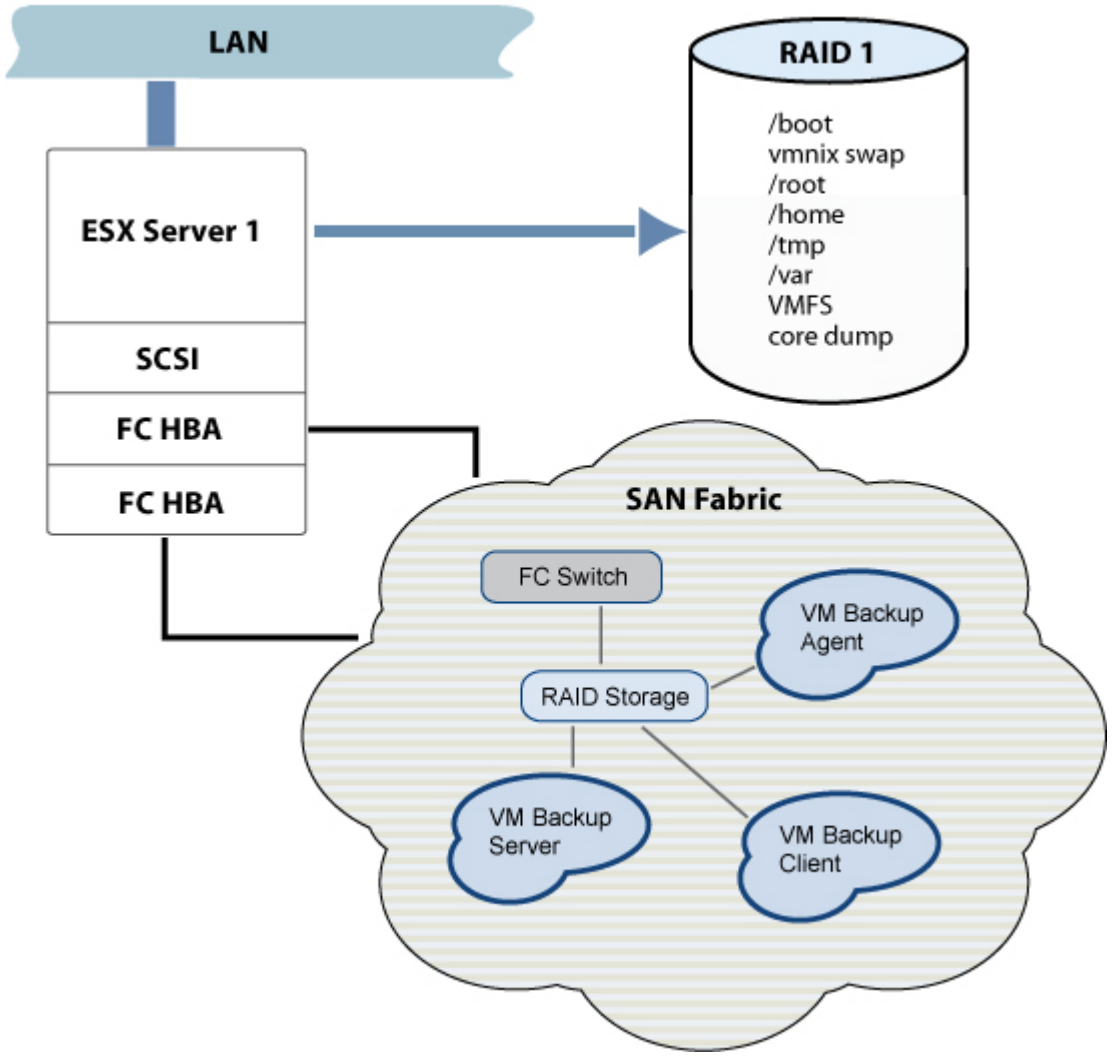


Figure 8-5. ESX Data Redundancy Setup Using RAID

The recommended disk layout for a virtual machine (VM) is shown in Figure 8-8. VMware recommends that you keep the operating system of a virtual machine separate from other data. This means using a dedicated VMFS partition just for storing an operating system. This VMFS partition can be from a shared volume or can be the entire partition of a unique volume. Keep all application data and other data in one or more separate VMFS partitions or RDM volumes. With this structure, the VMFS partition with the boot disk image is kept to an ideal minimum size. The smaller VMFS operating system partition means it takes less time for cloning, exporting, and backup. It is also useful to have separate backup policies for the VMFS operating system partition and the VMFS data disks.

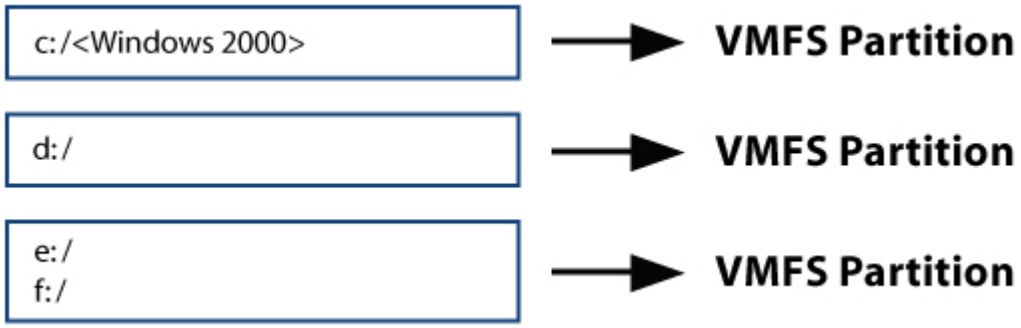


Figure 8-6. VMFS Partitioning to Separate OS and Virtual Machine Data Files

Shared LAN and SAN Impact on Backup and Recovery Strategies

The topology diagram (Figure 8-7) illustrates by physical connections how ESX hosts may co-exist within a LAN and SAN. The example shows three ESX Server hosts. ESX Server host 1 has one local SCSI HBA. The SCSI HBA is not connected to any other SCSI devices but the external DAS SCSI tape device shown. For ESX 2.5.x or older, the SCSI HBA is in shared mode to do tape backup and restore operations. In addition, ESX host 1 has two FC connections to SAN. ESX hosts 2 and 3 are both connected to the same LAN and SAN.

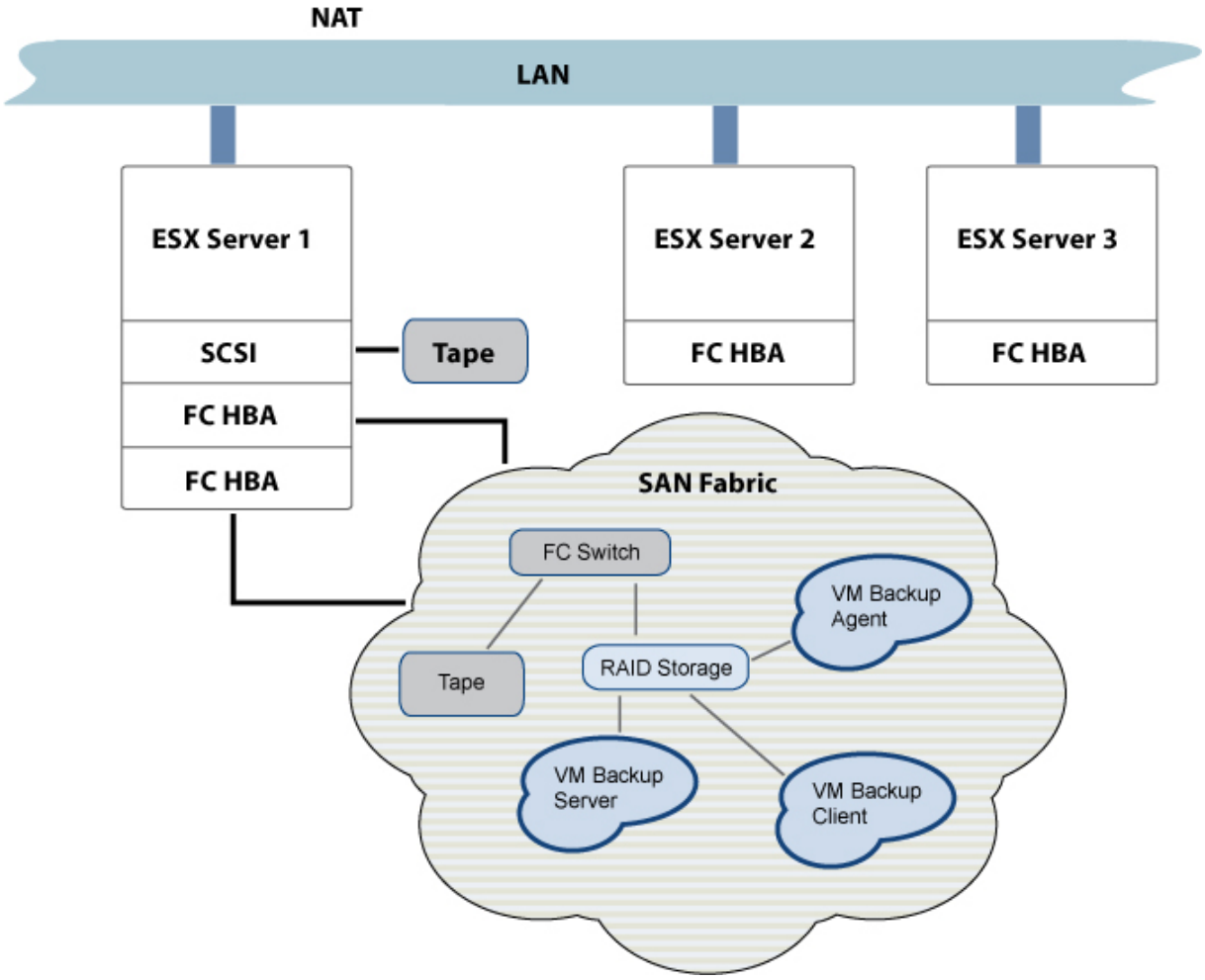


Figure 8-7. Shared LAN and SAN Topology

Backup and restore operations can be initiated from any ESX host to a LAN-based backup server using NAS, or to SAN-based backup targets via a VCB solution. Virtual machines owned by any ESX host can reside in the same RAID array on SAN. The FC tape shown in Figure 8-9 is not controlled by any ESX host. Direct connections of FC tape devices to VMware ESX are not supported.

Backup and restore operations in this diagram illustrate the possible data flow between the following:

- DAS ◄─►NAS (LAN backup)
- NAS ◄─► SAN (LAN backup to SAN)
- DAS ◄─► SAN (LAN-free between SCSI tape and RAID)
- SAN ◄─► SAN (LAN-free between FC tape and RAID). In ESX environments, this solution is possible only with VMware Consolidated Backup.

The architectural diagram (Figure 8-8) illustrates the logical data flow between VMware ESX and a DAS device or a device located on NAS or SAN. Note that all virtual machines are physically located on a RAID array on SAN.

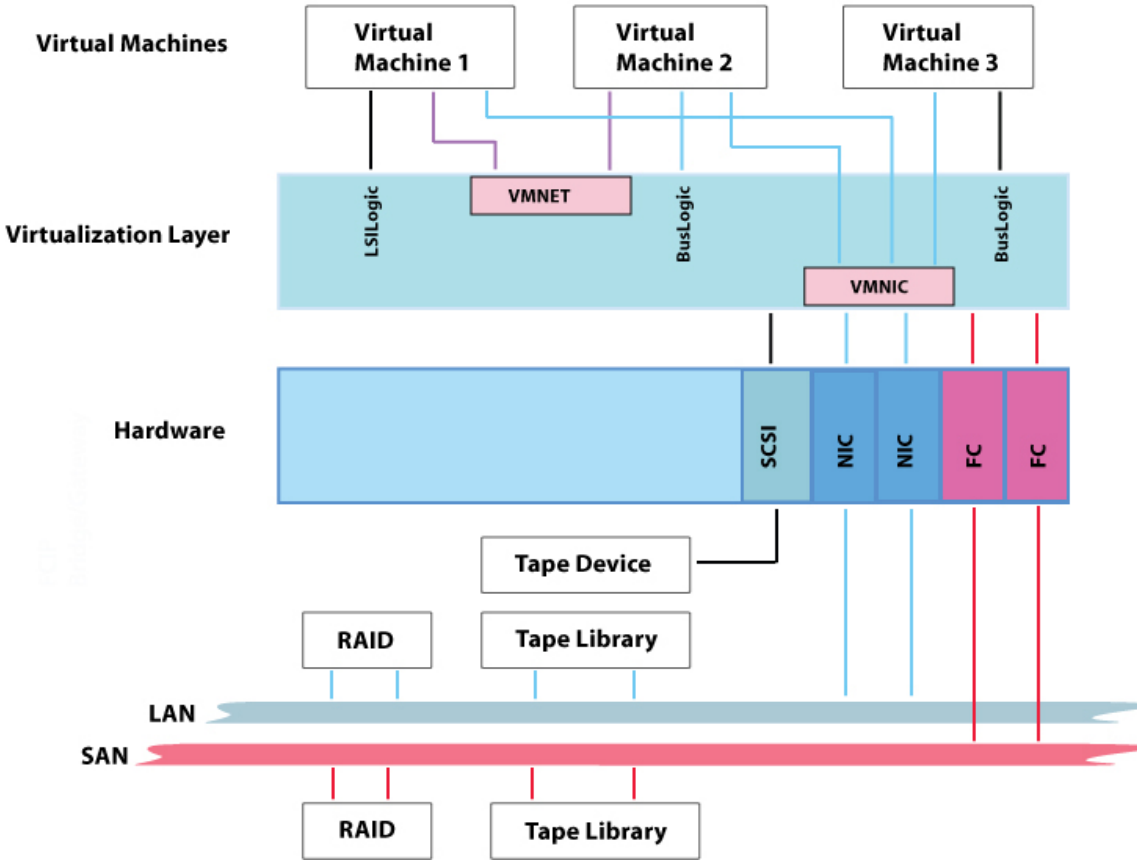


Figure 8-8. Data Flow between VMware ESX and Storage Devices

VMware ESX provides a virtualization layer to manage all resources of an Intel platform. As shown in the diagram, every network adapter operation between a virtual machine and the destination node on LAN is through the VMware virtual NIC VMNIC. FC and SCSI operation between the virtual machine and the destination node on SAN is emulated as SCSI within the virtualization layer. Thus in the diagram, only a SCSI link is shown to connect from virtual machines to the virtualization layer. The FC to SCSI emulation from each virtual machine is either a BusLogic or a LSI Logic SCSI device driver. For all tape backup devices to be assigned to a virtual machine, use an LSI Logic device driver.

NOTE: Figure 8-10 shows a more unique configuration not shown in the topology diagram of Figure 8-9. Also note in this figure that VMNET is shown to represent an internal network configuration that virtual machines can use exclusively, without a LAN connection. VMNET, *x* vmnics (virtual NICs), and vmnets (virtual networks), used in ESX 2.x architecture, is replaced by using vSwitches in VMware Infrastructure 3.

From the example, VM 1 can serve as a backup server while VM 2 serves as a backup client. Communication between the two servers is virtualized within VMware ESX. The major advantage of using this VMNET is that traffic is localized within an ESX host and SAN. This implementation using VMNET offloads any traffic from LAN.

Backup Policy Schedules and Priority

You need to back up three major components regularly:

- The data disk image belonging to each virtual machine (VMFS partitions or RDM volumes containing application data and other data)
- The boot disk image of each virtual machine (the VMFS operating system partition)
- The VMware ESX service console

You need to set a policy for each component mentioned. Probably, data disks assigned to each virtual machine will change frequently. A daily incremental backup is thus recommended. The boot disk image of a virtual machine might not change as frequently and should be backed up at least once a week. Changes to the VMware ESX service console are minimal, so the backup policy used is at the discretion of the system administrator. Table 8-1 lists the recommended backup policy priorities.

Table 8-1. Backup Policy Prioritization

What to Back Up? (In Order of Priority)	Backup Policy Recommended	Archive Policy
Data disk image of virtual machines	Daily incremental backup, real-time data replication	Full backup weekly
Boot disk image of virtual machines (VMFS operating system partitions)	Weekly differential backup	Full backup weekly

Table 8-2 lists the options, in order of recommendation, of what components of VMware ESX to back up. Each column heading lists what to back up, and the row cells describe backup options and information on choosing the destination of backups. Each backup method is described further in Table 8-3.

Table 8-2. Backup Methods

Back Up Data Disks from Virtual Machines	Back Up VMFS Operating System Partitions of Virtual Machines	Back Up ESX Service Console
Use VCB solution.	Use VCB solution.	Use a virtual machine as a backup server.
Use a virtual machine as a backup client.	Use data replication techniques within a RAID array.	Use a virtual machine as a backup server.
Use a virtual machine as a backup server.	Use a virtual machine as a backup server.	Use the ESX service console as a backup client.
Use data replication techniques within a RAID array.		Use the ESX service console as a backup server.

Back Up Data Disks from Virtual Machines	Back Up VMFS Operating System Partitions of Virtual Machines	Back Up ESX Service Console
Requires Shutting Down the Virtual Machine		
Use `vmkfstools -e` to export `*.dsk` files to a secondary location.	Use ESX service console as a backup client. Use ESX service console as a backup server. Use `vmkfstools -e` option to export `*.dsk` files to a secondary location.	

In Table 8-3, below, each option is further refined to allow more options in choosing the backup destination. Options 1 to 9 (from the first column in Table 8-2) are the most commonly configured because they are easy to set up, cost the least, and take advantage of the usually already-configured backup server on LAN.

Table 8-3. Backup Configuration Options

#	Backup Initiator	Backup Target	I/O Traffic Protocol	Device Sharing* (Dedicated and/or Shared)	VMware Support (Yes/ No)
1	Virtual machine (backup client)	Backup server on LAN	IP and SCSI or FC	N/A	Yes
2	" "	Virtual machine backup server	FC	N/A	Yes
3	" "	Local SCSI disk	SCSI	N/A	Yes
4	" "	External SCSI disk	SCSI	N/A	Yes
5	" "	Local SCSI tape	SCSI	N/A	Yes
6	" "	External SCSI tape	SCSI	N/A	Yes
7	" "	FC RAID array	FC	N/A	Yes
8	" "	FC direct-connect tape	FC	N/A	Yes
9	" "	FC tape on SAN	FC	N/A	Yes
10	Console operating system (backup client)	Backup server on LAN	IP and SCSI or FC	N/A	Yes
11	" "	Virtual machine backup server	FC	N/A	Yes
12	" "	Local SCSI disk	SCSI	N/A	Yes
13	" "	External SCSI disk	SCSI	N/A	Yes
14	" "	Local SCSI tape	SCSI	N/A	Yes
15	" "	External SCSI tape	SCSI	N/A	Yes

#	Backup Initiator	Backup Target	I/O Traffic Protocol	Device Sharing* (Dedicated and/or Shared)	VMware Support (Yes/ No)
16	" "	FC RAID array	FC	N/A	Yes
17	" "	FC direct-connect tape	FC	N/A	Yes
18	" "	FC tape on SAN	FC	N/A	Yes
19	Virtual machine (backup server)	Backup server on LAN	IP and SCSI or FC	N/A	Yes
20	" "	Virtual machine backup server	FC	Dedicated / Shared	Yes (dedicated) / No (shared)
21	" "	Local SCSI disk	SCSI	N/A	Yes
22	" "	External SCSI disk	SCSI	N/A	Yes
23	" "	Local SCSI tape	SCSI	Dedicated / Shared	Yes (dedicated) / No (shared)
24	" "	External SCSI tape	SCSI	Dedicated / Shared	Yes (dedicated) / No (shared)
25	" "	FC RAID array	FC	N/A	No
26	" "	FC direct-connect tape	FC	Dedicated / Shared	Yes (dedicated) / No (shared)
27	" "	FC tape on SAN	FC	Dedicated / Shared	Yes (dedicated) / No (shared)

*Device sharing applies to tape library in ESX 2.5.x releases. **Dedicated** means that all tape drives within a tape library are dedicated to either the VMware ESX service console or the virtual machine. **Shared** means that tape drives within a tape library are shared between VMware ESX service console and a virtual machine. In other words, the ESX service console may own one SCSI drive while the virtual machine owns another from the same tape library.

Backup Options Advantages and Disadvantages

Table 8-4, below, lists the major advantages and disadvantages of the backup and restore implementations detailed in Table 8-2.

Table 8-4. Backup Option Advantages and Disadvantages

#	Backup Method	Advantages	Disadvantages
1	**Virtual machine (backup client)**	Since the virtual machine is a backup client, file-level backup and restore behaves the same way as on any other physical machine with a backup client installed.	Backs up only specific files within a virtual machine and not the entire virtual machine. This approach does not take advantage of the suspend/resume and snapshot capabilities of VMware ESX. LAN traffic increases as data on the virtual machine grows.
2		Running the backup server software within a virtual machine offloads work from service console. Backup traffic is LAN-free.	Requires one virtual machine dedicated to be on, 24 x 7.
3, 4, 5, 6		Does not require NAS. Easy to setup. Backup traffic is LAN-free.	The SCSI HBA and SCSI tape device must be tested and supported by VMware.
7		Fast and reliable hardware. Backup traffic is LAN-free.	Expensive and more difficult to manage a SAN.
8/9		Fast and reliable hardware. Backup traffic is LAN-free.	Expensive and more difficult to manage a SAN. The FC HBAs and FC tape libraries must be tested and supported by VMware.
10	**Console operating system (backup client)**	No need for DAS as backup target is NAS. This offloads work from the service console.	You need to set up a NAS environment.
11		Running the backup server software within a virtual machine offloads work from service console. Backup traffic is LAN-free.	Requires one virtual machine dedicated to be on, 24 x 7.
12, 13, 14, 15		This is a simple way to back up the service console. It is almost identical to the procedure for backing up any Linux machine with a local tape drive. Does not require NAS. Easy to set up. Backup traffic is LAN-free.	The tape drive must be dedicated for service console use only. This puts more workload on the service console than is recommended. The SCSI HBA and SCSI tape device must be tested and supported by VMware.
16		Fast and reliable hardware. Backup traffic is LAN-free.	Expensive and more difficult to manage a SAN. The FC HBAs and FC RAID array must be tested and supported by VMware.
17,18		Fast and reliable hardware. Backup traffic is LAN-free.	Expensive and more difficult to manage a SAN. The FC HBAs and FC RAID array must be tested and supported by VMware.

#	Backup Method	Advantages	Disadvantages
19	**Virtual machine (backup server)**	If NAS is already implemented, then this solution is quick to set up.	Requires one virtual machine dedicated to be on, 24x7. If the owning ESX host needs to be serviced, then this virtual machine needs to be migrated to another ESX host (possible with VMotion).
21, 22, 23, 24		Running the backup server software within a virtual machine offloads work from service console. Backup traffic is LAN-free.	Requires one virtual machine dedicated to be on, 24 x 7.
25		Fast and reliable hardware. Backup traffic is LAN-free.	Expensive and more difficult to manage a SAN. The FC HBAs and FC RAID array must be tested and supported by VMware.
26/27		Fast and reliable hardware. Backup traffic is LAN-free.	Expensive and more difficult to manage a SAN. The FC HBAs and FC tape libraries must be tested and supported by VMware.

How to Choose the Best Option

The previous section demonstrates that there is no single way of implementing backup and restore for an ESX environment. Choosing the best option for your environment depends on how the existing configuration is set up and on future backup and restore needs. Use Table 8-5 as a general guideline to addressing some specific situations.

Table 8-5. General Backup Selection Guidelines

Situation	Recommendation
There is an existing server on LAN dedicated as a backup server.	Install backup client software on the ESX service console and on the virtual machine to do backup and restore to and from NAS.
LAN bandwidth is low to medium and is available for backup traffic.	" "
There is no server on LAN dedicated as a backup server.	Install DAS devices with ESX hosts.
LAN bandwidth is limited and backup traffic is restricted at night use only.	" "
VMware supports the specific HBA and DAS devices you have.	" "
Backup and restore is needed for large-scale operations and FC backup is enabled through existing SAN backup infrastructure.	Use VMware Consolidated Backup (VCB) strategy.
Backup and restore is needed for large-scale operations of servers running OLTP applications.	Install backup client software on the ESX service console and on the virtual machine to back up and restore to and from RAID arrays on SAN.

Situation	Recommendation
LAN bandwidth is not available for backup traffic.	Install backup client software on the ESX service console and on the virtual machine to back up and restore to and from RAID arrays on SAN.
VMware supports the specific FC HBA and FC RAID array you have.	" "
Backup and restore is needed for large-scale operations of servers running OLTP applications.	Install backup server software on a virtual machine. Install the backup client on the ESX service console. All backup and restore traffic is LAN-free using either an FC RAID array or an FC tape library.
LAN bandwidth is not available for backup traffic.	" "
VMware supports the specific FC HBA and FC tape library you have.	" "

Implementation Order

Table 8-6, lists, in order of recommendation, what to do for backup and restore of each component in ESX environments. Refer to the corresponding implement step number in the next section for a step-by-step guide.

Table 8-6. Implementation Order

Back Up Data Disks from Virtual Machines	Implement Step #
Use VCB.	See the *VCB Guide*.
Use a virtual machine as a backup client.	1
Use a virtual machine as a backup server.	2
Use data replication techniques within a RAID array.	5
Use `vmkfstools -e` to export `*.dsk` files to a secondary location. (This requires you to shut down the virtual machine.)	6
Back Up VMFS Operating System Partitions of Virtual Machines	**Implement Step #**
Use VCB.	See the *VCB Guide*.
Use data replication techniques within a RAID array.	5
Use a virtual machine as a backup server.	2
Use the ESX service console as a backup client. (This requires you to shut down the virtual machine.)	3
Use the ESX service console as a backup server. (This requires you to shut down the virtual machine.)	4
Use `vmkfstools -e` to export `*.dsk` files to a second location. (This requires you to shut down the virtual machine.)	6
Back Up the ESX Service Console	**Implement Step #**
Use a virtual machine as a backup server.	2
Use the ESX service console as a backup client.	3
Use the ESX service console as a backup server.	4

Backup Solution Implementation Steps

You can apply the same backup options to virtual machines, as if they were physical machines. For a low-cost solution, virtual machines can be used as backup clients, which require minimal setup time. The advantage is that the virtual machine backup clients can use an existing backup server dedicated on SAN for data backup. If a backup server is not readily available, the virtual machine itself can be set up as a backup server, to provide service to virtual machines on the same physical ESX host or neighboring physical hosts acting as clients.

The following sections list steps in implementing various VMware backup solutions.

How to Use a Virtual Machine as a Backup Client

The easiest method to back up data is use the virtual machine as a client. A large number of backup software applications have been tested and certified by VMware. Using one of these applications, you can start backing up as soon as software installation is complete.

1. Install the backup client software of your choice on each virtual machine.
2. Refer to your product documentation for information on scheduling backup operations. See the previous table for information on recommended backup policies.

How to Use a Virtual Machine as a Backup Server

Just as you can build a backup server using a physical server, you can also create a backup server using a virtual machine. This method takes a bit more time than building a backup server using a physical server, because some hardware setup is required. In addition, the choice and support of tape backup hardware is limited using this method. Use this option only if a SAN backup server is not available.

To use a SCSI DAS tape unit:

1. Power down the ESX host.
2. Install a SCSI HBA that is supported by VMware. See the VMware *I/O Adapter Compatibility Guide* at www.vmware.com/support.
3. Connect the tape unit to the SCSI HBA.
4. Power up the ESX host.
5. Use the MUI (for V2.5.x), or VirtualCenter (for VMware Infrastructure 3), to share this new SCSI HBA on the ESX host.
6. Using the MUI or VirtualCenter, assign the tape unit to the virtual machine that will be the backup server.
7. Use the MUI or VirtualCenter to change the virtual machine to use the LSI Logic device driver.
8. Install the backup server software of your choice on a virtual machine.
9. Refer to your product documentation for information to configure your virtual machine as a backup server.

To use an FC tape unit:

1. Power down the ESX host.
2. Install an FC HBA that is supported by VMware. See the VMware *I/O Adapter Compatibility Guide* at www.vmware.com/support.
3. Connect the FC HBA to a SAN.
4. Connect the FC tape unit to the same SAN.
5. Power up the ESX host.
6. Create the appropriate SAN zone to allow ESX access to the FC tape unit.
7. Use the MUI (for v2.5.x), or VirtualCenter (for VMware Infrastructure 3) to share this new FC HBA on the ESX host.
8. Use the MUI or VirtualCenter, select **Options** > **Advanced Settings** to change SCSI.PassthroughLocking from its default value of **1** to **0**. Changing this value to zero allows sharing of tape drives among different virtual machines.
9. Use the MUI or VirtualCenter to change the virtual machine to use the LSI Logic device driver.
10. Use the MUI or VirtualCenter to assign the tape unit to the virtual machine that will be the backup server.

How to Use the ESX Service Console as a Backup Client

1. Install the backup client software of your choice on the ESX service console.
2. Refer to your product documentation for information on scheduling a backup operation from a Linux backup client. See Table 8-1 for backup policy recommendations.

How to Use the ESX Service Console as a Backup Server

To use a SCSI DAS tape unit:

1. Power down the ESX host.
2. Install a SCSI HBA that is supported by VMware. See the VMware *I/O Adapter Compatibility Guide* at www.vmware.com/support.
3. Connect the tape unit to the SCSI HBA.
4. Power up the ESX host.
5. Use the MUI (for v2.5.x), or VirtualCenter (for VMware Infrastructure 3) to share this new SCSI HBA on the ESX host.
6. Install the backup server software of your choice on the ESX service console.
7. Refer to your product documentation for information on configuring your ESX host as a Linux backup server.

To use an FC tape unit:

1. Power down the ESX host.
2. Install an FC HBA that is supported by VMware. See the VMware *I/O Adapter Compatibility Guide* at www.vmware.com/support.
3. Connect the FC HBA to a SAN.
4. Connect the FC tape unit to the same SAN.
5. Power up the ESX host.
6. Create an appropriate SAN zone to allow the ESX host access to the FC tape unit.
7. Use the MUI (for v2.5.x) or VirtualCenter (for VMware Infrastructure 3) to share this new FC HBA on the ESX host.
8. Refer to your product documentation for information on configuring your ESX host as a Linux backup server.

How to Use Data Replication Techniques Within a RAID Array

Contact your array vendor support for details.

How to Use the vmkfstools –i Option to Export Virtual Machine Files

The key advantage to using export and import commands in moving disk image files (with either `*.dsk` or `*.vmdk` file extension) is that doing so divides large files into smaller, 2GB size files. These files can be located anywhere in storage (either on the same volume or in different volumes). During import, the command re-assembles the 2GB files into the original size for their use.

NOTE: For ESX 2.x, virtual machine files have the `.dsk` file extension; For ESX 2.5.1 and later releases, the virtual machine files have the `.vmdk` extension. For more information, see the VMware *ESX Administration Guide* available at:

www.vmware.com/support.

9 Optimization and Performance Tuning

Besides ensuring reliability, designers of VMware Infrastructure solutions need to make sure that the ESX hosts, virtual machines, and applications they are currently managing provide sufficient performance for current workloads, plus be able to scale to handle peak demands as well as grow to handle future demand and expansion.

This chapter provides information on factors that affect performance of various components in VMware Infrastructure systems.

Topics covered in this chapter are the following:

Introduction to Performance Optimization and Tuning

Ultimately, performance impact to a customer is experienced at the application level. Customers do not usually care what is "underneath the hood," and thus the burden of managing and providing system performance is placed squarely on virtual infrastructure administrators on all levels (from administrators managing VMware Infrastructure, to SAN administrators managing the SAN fabric, to administrators managing SAN storage arrays).

Managing VMware Infrastructure performance takes the form of meeting the demands of current workloads as well as ensuring the ability of applications to scale and handle larger workloads and peak demands. To ensure adequate performance, administrators must also be able to monitor and measure system performance. Although VMware VirtualCenter provides monitoring capabilities and the ability to

configure and manage performance parameter settings, administrators must rely on existing SAN management tools and system knowledge to fine-tune all the components in their environments.

When tuning VMware Infrastructure for performance, you must understand the entire system's operation and architectural design, from end-to-end, so you can match components such as I/O per second (IOPS) and queuing. “End-to-end” refers to all operations and components of the entire system, from the handling of an application request to the underlying support software (in this case, ESX virtualization), the physical HBA, the SAN fabric components, the disk array controllers, and finally the disk drive. Without an insight into each of the supporting components, you might be making “blind” changes, or changes to one area of the system without considering all others. Such changes not only yield no overall system performance improvement, but can in fact make things worse. For example, simply changing the queue depth of an HBA can flood the I/O subsystem if disk drives are not handling requests fast enough, and performance can actually get worse.

Tuning Your Virtual Machines

A virtual machine is practically no different than a physical server that hosts an operating system and runs applications. Tuning or updating a physical server (changing CPUs, memory, or updating device drivers), is similar to tuning or updating a virtual machine environment. You can change the following settings in a virtual machine to improve performance:

- Choose **vmxlsilogic** instead of **vmxbuslogic** as the LSI Logic SCSI driver. The LSI Logic driver provides more SAN-friendly operation and handles SCSI errors better.
- Use VirtualCenter to allocate a larger number of shares, as needed.
- Use VirtualCenter to assign more memory to a virtual machine or dedicate more CPUs to specific virtual machines that need more processing power.
- Run virtual machines on SMP-only servers if you know that your applications are multithreaded and can take advantage of multiple processors.
- Equalize disk access between virtual machines.

  You can adjust the setting of the maximum number of outstanding disk requests with the Disk.SchedNumReqOutstanding parameter in the VI Client. When two or more virtual machines are accessing the same volume, this parameter controls the number of outstanding requests that each virtual machine can issue to the volume. Adjusting the limit can help equalize disk access between virtual machines. This limit is not applicable when only one virtual machine is active on a volume. In that case, the bandwidth is limited by the queue depth of the storage adapter.

  Be careful when changing the SchedNumReqOutstanding parameter above the default value of 16 if you have not first done a benchmark analysis of your environment. You might want to experiment with different value settings by increasing SchedNumReqOutstanding by four. For example, change the default value of 16 to 20, then to 24, and on up to 28, as a conservative approach. Increasing the queue depth of each virtual machine means that you are loading more requests into the queue of the HBA device driver. Increasing the value of the SchedNumReqOutstanding setting eventually floods your HBA driver queue.

Thus, when you increase the value of the SchedNumRegOutstanding parameter setting, you also have to increase the size of the HBA driver queue. The key is to find the optimal number for SchedNumReqOutstanding that best fits your specific HBA driver queue setting. That optimization can be determined only if you measure your HBA driver queue-handling capacity. You might want to contact VMware Professional Services personnel to help you find this optimal number for your environment.

To set the number of outstanding disk requests:

1. In the VI Client, select the host in the inventory panel.
2. Click the **Configuration** tab and click **Advanced Settings**.
3. Click **Disk** in the left panel and scroll down to the Disk.SchedNumReqOutstanding parameter.
4. Change the parameter value to the desired number and click **OK**.
5. Reboot the server.

VMware ESX Sizing Considerations

The size of your infrastructure—the number of virtual machines sharing a volume, the volume size, and the number of physical servers sharing the same volume—affects performance. Here are some general guidelines to consider when designing your VMware Infrastructure:

- Sharing of physical volume: See KB articles 1240, 1267, 1268, 1269 and http://www.vmware.com/resources/techresources/1059
- Reduce the number of SCSI reservations. Operations that require getting a file lock or a metadata lock in VMFS result in short-lived SCSI reservations. SCSI reservations lock an entire volume. Excessive SCSI reservations by a server can degrade the performance of other servers accessing the same VMFS.

  Examples of operations that require getting file locks or metadata locks include:

  — Virtual machine power on

  — VMotion

  — Virtual machines running with virtual disk snapshots

  — File operations, such as open or close from the service console requiring opening files or performing metadata updates. (See “Metadata Updates” on page 58.)

  Performance can degrade if such operations are performed frequently on multiple servers accessing the same VMFS. For instance, you should not run a large number of virtual machines from multiple servers that are using virtual disk snapshots on the same VMFS. Similarly, you should limit the number of VMFS file operations that are executed from the service console when many virtual machines are running on the VMFS.
- No more than 16 physical hosts should share the same volume.
- Limit concurrent VMotion operations to a maximum of four.

NOTE: Contact VMware Professional Services if you want help with sizing your ESX system and choosing configuration options for your specific environment.

Managing ESX Performance Guarantees

VMware Infrastructure allows you to optimize resource allocation by migrating virtual machines from over-utilized hosts to under-utilized hosts. VMware Infrastructure's unique distributed resource scheduling (DRS) capabilities provide can also help optimize available resources. You have two options for migrating virtual machines:

- Migrate virtual machines manually using VMotion.
- Migrate virtual machines automatically using VMware DRS.

You can use VMotion or DRS only if the virtual disks are located on shared storage accessible to multiple servers. In most cases, SAN storage is used. For additional information on VMotion, see the VMware *Virtual Machine Management Guide*. For additional information on DRS, see the VMware *Resource Management Guide*.

VMotion

VMotion technology enables intelligent workload management. VMotion allows administrators to manually migrate virtual machines to different hosts. Administrators can migrate a running virtual machine to a different physical server connected to the same SAN, without service interruption. VMotion makes it possible to do the following:

- Perform zero-downtime maintenance by moving virtual machines around so the underlying hardware and storage can be serviced without disrupting user sessions.
- Continuously balance workloads across the datacenter to most effectively use resources in response to changing business demands.

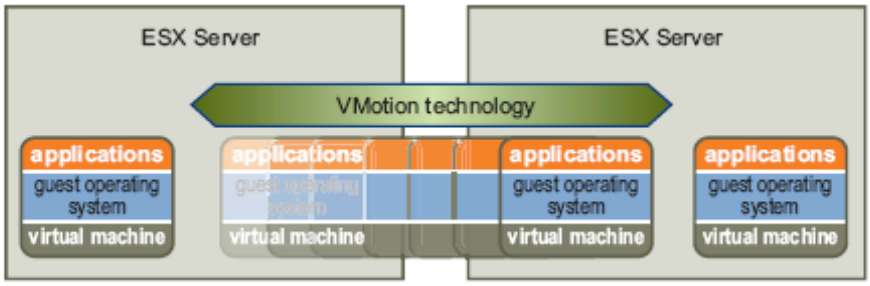


Figure 9-1. Migration with VMotion

VMware DRS

VMware DRS helps improve resource allocation across all hosts and resource pools. (VMware Infrastructure 3 supports resource allocation across up to 16 hosts, compared with 32 in VMware Infrastructure version 3.0.1.)

VMware DRS collects resource usage information for all hosts and virtual machines in a VMware cluster and provides recommendations (or automatically migrates virtual machines) in one of two situations:

- **Initial placement** — When you first power on a virtual machine in the cluster, DRS either places the virtual machine or makes a recommendation.
- **Load balancing** — DRS tries to improve resource utilization across the cluster by either performing automatic migrations of virtual machines (VMotion) or providing recommendation for virtual machine migrations.

For detailed information on DRS configuration and operation, see the VMware *Resource Management Guide*.

Optimizing HBA Driver Queues

VMware ESX uses two Fibre Channel drivers: Emulex and Qlogic. Both drivers support changing the HBA queue depth to improve performance. For Emulex, the default driver queue setting is 30. For Qlogic driver version 6.04, the default is 16, and for version 6.07 and 7.x (in VMware Infrastructure 3), the default is 32. Although Emulex has a maximum of 128 queue tags (compared to 255 for Qlogic), increasing the queue depth in either driver to above 65 has little effect. This is because the maximum parallel execution or queue depth of SCSI operations is 64.

Increasing the HBA driver queue depth eventually floods the queue on the attached disk array. Disk array queue response time is dependent on a number of factors, such as disk controller cache size, write-enable caching (on or off), de-staging algorithms, disk drive rotational latency, disk drive transfer time, disk drive types, and RAID types. You cannot control all of these factors, but you can:

- **Increase the disk controller cache size**. Larger is better, but increasing size has an associated cost. For any given cache, the ideal scenario is that requested data from an application is present when needed. This is referred to as a 100 percent cache hit ratio. In the real world, achieving this ration is not possible and thus the size of the cache and the algorithm for determining disk de-staging play vital roles in maximizing the cache hit success ratio. Since the cache provides a finite amount of space, the cache is eventually filled with “dirtybits” (data modified by the CPU, but not yet written back to storage). Under intense computational processing from applications, service request times from the disk array increase. To decrease or minimize this time, you usually need to buy more disk controller cache.
- **Enable write-caching** (at the HBA level or at the disk array controller level) to improve performance. Keep in mind that, because volatile RAM is used, data integrity issues become an issue if power is lost or system errors occur. Having a backup power supply for these situations is crucial.
- **Optimize disk rotational latency** for best performance by selecting the highest performance disk drive types, such as FC, and use the same disk drive types within the same RAID set. Do not create RAID sets of mixed disk drives or drives with different spindle speeds. Keep in mind that FC disks cost more and deliver less space capacity, but they provide the highest IOPS. You should research and select disk drive capabilities to match your application's IOPS demand requirements.

- **Select the correct RAID types** for a big impact on application performance. RAID 5 is best for read-intensive applications and RAID 1-0 is best for write-intensive applications. Check with your array vendor specification on any specific proprietary RAID implementations that you can adjust to match your application demands. Ideally, you should assign different RAID set volumes to a virtual machine to get the best performance. For example, your virtual machine boot image might be stored on a volume from a RAID 5 disk (where drive c:\ is on RAID 5 volume). To define separate data disks, these disks could be assigned to different RAID sets. For example, you could assign drive d:\ to another RAID 5 volume, and assign drive e:\ and drive g:\ to other RAID 1-0 volumes.
- **Change the queue depth for the HBA driver** to effectively increase the number of outstanding I/O operations that can be placed in the queue for servicing.

  You can adjust the queue depth for a HBA adapter with the following procedure. You should change the queue depth only if you notice unsatisfactory performance. Unsatisfactory performance should be quantified by storage performance experimentation and measurements. Changing the queue depth, without understanding the end-to-end effect, may make your environment's performance worse. Contact your VMware Professional Services personnel for details.

NOTE: Contact VMware Professional Services if you want help with HBA and storage configuration selection and performance tuning.

To set maximum queue depth:

1. Back up the file `/etc/vmware/esx.conf`
2. Run the followings commands:

   ```
   esxcfg-module –s <vendor>_lun_queue_depth=16 <FC driver>.o
   esxcfg-boot –b
   ```

3. Reboot the ESX host.

I/O Load Balancing Using Multipathing

The VMware ESX VMkernel has a built-in proprietary implementation of multipathing that establishes and manages SAN fabric path connections. Once paths are detected, VMkernel establishes which path is active, in standby, or no longer connected (a dead path). When multipathing detects a dead path, it provides failover to alternate paths. Because of this multipathing mechanism, you can leverage all available paths to manually load balance I/O workload.

By default, VMware ESX only uses one path, regardless of available alternate paths. For active/active type arrays, you must manually assign a LUN to a specific path. For active/passive type arrays, VMware ESX automatically picks one path for I/O workloads; you cannot load balance to an alternate path. This is a limitation of the active/passive array types as only one path can be active at one time.

For best performance, each virtual machine should reside on its own LUN. In this way, the volume can be specified a preferred path to the SAN fabric. Before doing this, however, you must determine which virtual machine or volume (if multiple virtual machines are sharing this one volume or LUN) has the most intensive I/O workload. Note that the level of I/O traffic varies for different applications at different times of the day. To determine the maximum level of I/O traffic for a virtual machine (or for a specific volume), monitor actual performance at intervals of a few hours apart (or a shorter time duration) during a typical day. Monitor performance first by measuring the IOPS generated by the application. Second, measure the bandwidth availability between the associated volume and the SAN fabric path.

NOTE: Consider the advantages and disadvantages of configuring a single virtual machine to reside on its own volume as outlined in Chapter 4, “How Virtual Machines Access Storage” on page 57.

To determine the level of I/O traffic that exists in your environment, you need to establish a record of I/O activity. You can do this by recording I/O traffic patterns in the SAN fabric ports (for example, by using commands such as `portperfshow` for Brocade switches). Once you determine that I/O traffic (for example, the SAN fabric port where a HBA from an ESX host is connected), is taking more than 90 percent of available port bandwidth, you should assign an alternate path to offload I/O traffic to less busy path. If available bandwidth is 10 percent or less, you could run out of bandwidth during unusual or unexpected higher I/O peaks or your application may flood the driver queue.

Balancing loads among available paths also improves performance. You can set up your virtual machines to use different paths by changing the preferred path for the different HBAs. This is possible only for active/active storage processors (SP) and requires that you have path policy set to Fixed.

If a path fails, the surviving paths carry all the traffic. Path failover should occur within a minute (or within two minutes for a very busy fabric), as the fabric might converge with a new topology to try to restore service. A delay is necessary to allow the SAN fabric to stabilize its configuration after topology changes or other fabric events.

SAN Fabric Considerations for Performance

The SAN fabric can be optimized by considering the following design options:

- **Minimize the number of switch hops** to minimize the cost of routing between initiator and target ports. Although you may not have direct control, you can request your SAN administrator to assign you a SAN fabric route that has the minimal amount of hops. You can also minimize switch hops by manually changing the routing and changing the ISLs cost settings for routing between your ESX HBA ports and the storage processor ports. For example, with Brocade switches, check with your SAN administrators on using commands such as uRouteConfig and linkCost.

- **Enable trunking on SAN fabric switches** connected to ESX hosts to leverage the higher aggregate bandwidth of available ISLs in the same trunk.
- **Assign a unique zone** between the ESX initiator to the storage controller processor ports. That is, for each WWN (each initiator HBA), regardless of number of ports, there should be one unique zone. This zoning strategy isolates an initiator from RSCN (Register State Change Notification) disruptions.

Disk Array Considerations for Performance

Disk I/O subsystems remain the I/O bottleneck in the majority of SAN systems. Thus, it is critical that you plan ahead to develop the best strategy to balance cost against the highest performance possible for your application demands. When choosing disk arrays, consider the following three primary design options.

1. **Array types, such as active/active versus active/passive.** Active/active array types provide the best performance because you can load balance the array on all available storage processor ports, giving the highest IOPS and bandwidth to server applications. In addition, active/active arrays normally include larger, upgradeable cache options and disk sizes.
2. **Disk caching size**:
   a) Disk controller cache size — More is better, but at a cost. Since caching provides a finite amount of space for a special locality to increase or maximize the hit cache ratio, the finite cache size will eventually be filled with dirty bits under intense computational requirements from applications. To minimize service request time from disk arrays, buy more disk controller cache.
   b) Write caching — You can improve performance by enabling write-caching at the disk array controller level. However, keep in mind that because volatile RAM is used for caching, data integrity may become an issue if there is a power loss or system errors occur. Thus, having a backup power supply is crucial in these situations.
3. **RAID**. Various RAID technologies not only help you plan your organization's virtual infrastructure solution to provide disaster recovery, but also help to optimize disk performance by maximizing the number of IOPS the array sets can generate or use to service application demand. Consider the following factors to help determine the best RAID level for you environment:
   a) Selecting the correct RAID types has a big impact on application performance. RAID 5 is best for read-intensive applications and RAID 1-0 is best for write-intensive applications. Check with your array vendor specification on any specific proprietary RAID implementations that you can adjust to match your application demands. Ideally, you should assign different RAID sets of volumes to a virtual machine to get the best performance. For example, your virtual machine boot image maybe be stored on a volume from a RAID 5 disk (drive c:\ is on RAID 5 volume). To define separate data disks, the disks could be assigned to different RAID sets. For example, you could assign drive d:\ to another RAID 5 volume, and assign drive e:\ and drive g:\ to RAID 1-0 volumes.

b) Do not create RAID sets that include mixed disk drive types or drives with different spindle speeds. Keep in mind that FC disks generally cost more and deliver less space capacity, but provide the highest IOPS. You should research and select disk drive capabilities to match your application's IOPS demand requirements.

Storage Performance Best Practice Summary

This section provides a summary of VMware recommendations for storage configurations and practices to achieve optimal performance:

- Make sure that I/O operations are not queuing up in VMkernel by checking the number of queued commands reported by `esxtop`. Since queued commands provide instantaneous statistics, you need to monitor these numbers over a period of time to see if you are hitting the queue limit. To determine the number of queued commands, look for the QUED counter in `esxtop`, the storage resource screen. If you are hitting queuing limits, adjust the queue depths. See VMware KB article 1267.
- To optimize storage array performance, spread I/O loads over the available paths to the storage (across multiple HBAs and SPs).
- For active/active arrays with fixed failover policy, designate active paths to each logical unit of storage. Doing so allows for the best possible use of your bandwidth to the disk array.
- Avoid operations that excessively open or close files on the VMFS, a distributed file system, or partition as these operations tend to be expensive. If possible, access a file, do everything that needs to be done with the file, and close it, instead of repeatedly opening and closing files to perform a series of incremental operations. As an example, running the `watch tail file_on_vmfs_partition` command in the service console can adversely affect storage performance.

- Use the most effective virtual disk modes. VMware ESX supports the following virtual disk modes: independent persistent, independent nonpersistent, and snapshot. These modes have the following characteristics:

  1. **Independent persistent mode** — Changes are immediately and permanently written to the disk, so they have high performance.
  2. **Independent nonpersistent mode** —Changes to the disk are discarded when you power off or revert to the snapshot. In this mode, disk-write operations are appended to a redo log. When a virtual machine reads from the disk, it first checks the redo log (by looking at a directory of disk blocks contained in the redo log) and, if the redo log is present, reads that information. Otherwise, the read goes to the base disk for the virtual machine. These redo logs, which track the changes in a virtual machine’s file system and allow you to commit changes or revert to a prior point in time, can incur a performance penalty.

3. **Snapshot mode** — A snapshot captures the entire state of the virtual machine at the time you take the snapshot. This includes the memory and disk states as well as the virtual machine settings. When you revert to a snapshot, you return all these items to the state they were in at the time you took the snapshot.

- Ensure that heavily used virtual machines do not all access the same VMFS volume concurrently. Assign each virtual machine to its own dedicated volume, with each volume having a unique LUN.
- Avoid operations that require excessive file locks or metadata locks, such as those used to dynamically grow `.vmdk` files or those that manipulate file permissions.
- Configure the maximum queue depth for HBA cards to optimize performance. See VMware KB article 1267, available on the VMware Web site. Quantify performance by experimentation and measuring storage performance.
- Increase the virtual machines' maximum outstanding disk requests, if needed. See VMware KB article 1268 available on the VMware Web site.
- For iSCSI/NFS, make sure that multiple input Ethernet links are not funneled into too few output links, which can result in an oversubscribed link. Oversubscription is a possibility any time a number of links that are transmitting near capacity are switched to a smaller number of links. Recovering from dropped network packets results in significant performance degradation. Not only do you spend time determining that data was dropped, but also retransmission of dropped packets uses network bandwidth that could otherwise be used for current transactions.
- Applications or systems that write a lot of data to storage, such as data acquisition or transaction logging systems, should not share ISLs to a storage device. These types of applications perform best with multiple connections to storage devices.
- Performance design for a storage network must take into account the physical constraints of the network. Using VLANs or VPNs does not provide a suitable solution to the problem of link oversubscription in shared configurations. VLANs and other virtual partitioning of a network provide a way of logically designing a network, but do not change the physical capabilities of links and trunks between switches.
- If you are working with systems with heavy disk I/O loads, you might need to assign separate storage processors to individual systems to handle the amount of traffic bound for the storage.
- Ensure adequate CPU resources, particularly for software-initiated iSCSI and NAS.
- Protect your service console's root file system from becoming full.
- Guest storage drivers typically set I/O size at 64K, by default. If applications issue I/O operations that are larger than 64K, these operations are split into 64K chunks. Changing the registry settings to issue larger block sizes can enhance performance. See VMware KB article 9645697.
- VMware ESX emulates either a BusLogic or LSI Logic SCSI adapter, which is likely to be different from the physical adapter installed on the server. The specifics of the implementation of the SCSI driver loaded into the guest operating system can affect disk I/O throughput. For example, the depth of the queue for

outstanding commands in a driver can significantly affect disk performance. A queue depth that is too small limits the disk bandwidth that can be pushed through the virtual machine. For BusLogic adapters, VMware provides a custom BusLogic driver for Windows guest operating systems that is recommended for applications requiring high performance. The BusLogic driver is part of VMware Tools and can be installed when you install VMware Tools. The guest driver queues can also be tuned. See the driver-specific documentation for more information on how to do this. The driver queue depth can also be set for some VMkernel drivers. For example, the default queue depth of the QLogic driver is 16. Specifying larger queue depths can yield higher performance. You can also adjust the number of outstanding disk requests per virtual machine in the VMkernel through the ESX Management User Interface. Setting this parameter can help equalize disk bandwidth across virtual machines. See VMware KB article 1268. Quantify performance by experimentation and measuring storage performance.

- VMware ESX supports raw device mapping (RDM), which allows management and access of raw SCSI disks or LUNs as VMFS files. An RDM is a special file on a VMFS volume that acts as a proxy for a raw device. The RDM file contains metadata used to manage and redirect disk accesses to the physical device. Virtual disks are recommended for most virtual disk storage. Raw disks might be needed in some cases.

- VMware ESX supports multiple disk formats. Their description and performance implications are as follows:

  1. **Thick format** — A thick disk has all space allocated at creation time. It has optimal I/O latencies during usage that might raise security issues, as thick-formatted disks can contain stale data as it exists on the physical media on creation.
  2. **Eager-zeroed format** — An eager-zeroed thick disk has all space allocated and is also zeroed out at creation time. Eager-zeroed disks require longer creation times, but provide optimal performance and better security.
  3. **Lazy-zeroed format** — This disk zeroes out blocks on the first write, but provides the same performance as thick and eager-zeroed disks on subsequent writes.

4. **Thin format** — Space required for thin-provisioned virtual disk is allocated and zeroed on demand, instead of upon creation. There is a higher I/O penalty during the first write to unwritten file blocks, but it provides the same performance as thick and eager-zeroed disks on subsequent writes.

Virtual machine disks configured through the VirtualCenter interface are of type lazy-zeroed. The other disk formats can be created from the console command line using `vmkfstools`. For more details, see the `vmkfstools` man page.

10 Common Problems and Troubleshooting

Troubleshooting SAN and storage subsystems is both a science and an art. The science of troubleshooting relies on understanding the components of your SAN or storage subsystems and obtaining a working knowledge of component specifications and limitations. Using your experience to troubleshoot a problem, and more specifically, identify where in the overall system to focus your investigation first, is the art.

The troubleshooting approach you use should be that of collecting data (similar to a science experiment) to validate a hypothesis. Having a working knowledge of Fibre Channel, SCSI protocol, and FC fabric switch commands is crucial in collecting information about SAN subsystem behavior. When you change the configuration or a parameter in a SAN fabric or a storage subsystem, you can measure or record the change in SAN behavior or performance. Using this logical and methodical approach to make changes most often results in resolving a problem.

This chapter provides information on how to troubleshoot and resolve issues in systems using VMware Infrastructure with SAN. It also lists common problems that system designers and administrators of these systems may encounter, along with potential solutions.

The topics covered in this chapter include the following:

Documenting Your Infrastructure Configuration

It is extremely helpful to have a record of your SAN fabric infrastructure architecture and component configuration. The following information is crucial to help isolate system problems:

1. A topology diagram showing all physical, network, and device connections. This diagram should include physical connection ports and a reference to the physical location of each component.
2. A list of IP addresses for all FC fabric switches. You can use this information to log into fabric switches which provide information that helps identify the relationship to components. For example, it can tell you which server is connected to which FC fabric switch at which port, and which fabric switch port is mapped to which fabric storage ports.
3. A list of IP addresses for logging into the service console of each ESX host. You can access the service console to retrieve storage component information (such as HBA type, PCI slot assignment, and HBA BIOS revision), as well as to view storage subsystem event logs.
4. A list of IP addresses for logging into each disk array type or array management application. The disk array management agent can provide information about LUN mapping, disk array firmware revisions, and volume sharing data.

Avoiding Problems

This section provides guidelines to help you avoid problems in VMware Infrastructure deployments with SAN:

- Carefully determine the number of virtual machines to place on a single volume in your environment. In a lightly I/O loaded environment (with 20 to 40 percent I/O bandwidth use), you might consider placing up to 80 virtual machines per volume. In a moderate I/O load environment (40 to 60 percent I/O bandwidth use), up to 40 virtual machines per volume is a conservative number. In a heavy I/O load environment (above 60 percent I/O bandwidth use), 20 virtual machines per volume is a good maximum number to consider. If your I/O bandwidth usage is 95 percent or higher, you should consider adding more HBAs to the servers that are hosting virtual machines and also consider adding inter-switch-links (ISLs) to your SAN infrastructure for the purpose of establishing ISL trunking.
- Do not change any multipath policy that the ESX management system sets for you by default (neither Fixed nor MRU). In particular, working with an active/passive array and setting the path policy to Fixed can lead to path thrashing.
- Do not change disk queuing (or queuing in general) unless you have performed a quantitative analysis of improvements you could achieve by setting new queue depths.
- For active/active arrays, check whether I/O load balancing using multipathing was set, since that may cause a bottleneck at a specific path.

- Determine if your application is CPU-, memory-, or I/O-intensive, to properly assign resources to meet application demands. See Chapter 7, "Growing your Storage Capacity," for more information on this topic.
- Verify from your FC fabric switch that all ESX hosts are zoned properly to a storage processor port and that this zone is saved and enabled as an active zone.
- Verify where each virtual machine is located on a single or multiple data disks and that these disk volumes are still accessible from the ESX host.
- Ensure that the FC HBAs are installed in the correct slots in the ESX host, based on slot and bus speed.

Troubleshooting Basics and Methodology

Some of the main keys to successful troubleshooting are keeping accurate documentation of your system architecture and configuration and, when a problem or issue arises, following a careful, methodical approach to isolate and remedy the situation, or find a workaround.

When a problem occurs, you can typically use the following steps to quickly identify the problem. You might spend fifteen to thirty minutes performing the steps, but if you cannot locate the problem and need to contact VMware support services, you will have already collected important information that lets VMware personnel help you resolve the problem:

1. Capture as much information as possible that describes the symptoms of the problem. Information that is crucial includes:
   a) The date or time the problem first occurred (or when you first noticed it).
   b) The impact of the problem on your application.
   c) A list of events that lead up to the problem.
   d) Whether the problem is recurring or a one-time event.
2. Check that the ESX host on which the problem occurs has access to all available mapped LUNs assigned from the storage array management interface.
   a) From your array management, verify that the LUNs are mapped to the ESX host.
   b) From the ESX host having the problem, verify that all LUNs are still recognized after a rescan.
3. Log into the FC fabric switches that are part of VMware Infrastructure and verify that all connections are still active and healthy.
4. Determine if your issue or problem is a known problem for which a solution or workaround has already been identified.
   a) Visit VMware knowledge base resources at http://kb.vmware.com
   b) Review the problems listed in Table 10-1 for commonly known SAN-related issues with ESX and VMware Infrastructure.

Common Problems and Solutions

This section lists the issues that are most frequently encountered. It either explains how to resolve them or points to the location where the issue is discussed. Some common problems that system administrators might encounter include the following:

1. LUNs not detected by the ESX host.
2. SCSI reservation conflict messages.
3. VirtualCenter freezes.
4. `Disk Partition not recognized` errors.
5. None of the paths to LUN `x:y:z` are working.
6. Service console freezes.

The following table lists solutions to some other common problems and issues:

Table 10-1. Common Issues and Solutions

Issue	Solution
A LUN is not visible in the VI Client.	See "Resolving Issues with LUNs That Are Not Visible" on page 106.
You want to understand how path failover is performed or change how path failover is performed.	The VI Client allows you to perform these actions. See "Managing Multiple Paths for Fibre Channel" on page 119.
You want to view or change the current multipathing policy or preferred path, or disable or enable a path.	The VI Client allows you to perform these actions. See "Managing Multiple Paths for Fibre Channel" on page 119.
You need to increase the Windows disk timeout to avoid disruption during failover.	See "Setting Operating System Timeout" on page 148.
You need to customize driver options for the QLogic or Emulex HBA.	See "Setting Device Driver Options for SCSI Controllers" on page 148.
The server is unable to access a LUN, or access is slow.	Path thrashing might be the problem. See "Resolving Path Thrashing Problems" on page 182.
Access is slow.	If you have many LUNs/VMFS volumes and all of them are VMFS-3, unload the VMFS-2 driver by typing at a command line prompt: `vmkload_mod -u vmfs2` You should see a significant increase in the speed of certain management operations, such as refreshing datastores and rescanning storage adapters.
You have added a new LUN or a new path to storage and want to see it in the VI Client.	You have to rescan. See "Performing a Rescan" on page 116.

Understanding Path Thrashing

In all arrays, the storage processors (SPs) are like independent computers that have access to some shared storage. Algorithms determine how concurrent access is handled.

- For active/passive arrays, only one SP can access all the sectors on the storage that make up a given volume at a time. The ownership is passed around between the SPs. The reason is that storage arrays use caches and SP A must not write something to disk that invalidates SP B's cache. Because the SP has to flush the cache when it is done with its operation, moving ownership takes a little time. During that time, neither SP can process I/O to the volume.
- For active/active arrays, the algorithms allow more fine-grained access to the storage and synchronize caches. Access can happen concurrently through any SP without extra time required.

Arrays with auto volume transfer (AVT) features are active/passive arrays that attempt to look like active/active arrays by passing the ownership of the volume to the various SPs as I/O arrives. This approach works well in a clustering setup, but if many ESX systems access the same volume concurrently through different SPs, the result is path thrashing.

Consider how path selection works during failover:

- On an active/active array, the system starts sending I/O down the new path.
- On an active/passive array, the ESX system checks all standby paths. The SP at the end of the path that is currently under consideration sends information to the system on whether or not it currently owns the volume.
    - If the ESX system finds an SP that owns the volume, that path is selected and I/O is sent down that path.
    - If the ESX host cannot find such a path, the ESX host picks one of the paths and sends the SP (at the other end of the path) a command to move the volume ownership to this SP.

Path thrashing can occur as a result of this path choice. If Server A can reach a LUN through only one SP, and Server B can reach the same LUN only through a different SP, they both continuously cause the ownership of the LUN to move between the two SPs, flipping ownership of the volume back and forth. Because the system moves the ownership quickly, the storage array cannot process any I/O (or can process only very little). As a result, any servers that depend on the volume start timing out I/O.

Resolving Path Thrashing Problems

If your server is unable to access a LUN, or access is slow, you might have a problem with path thrashing (sometimes called LUN thrashing). In path thrashing, two hosts access the volume via different storage processors (SPs), and, as a result, the volume is never actually available.

This problem usually occurs in conjunction with certain SAN configurations and under these circumstances:

- You are working with an active/passive array.
- Path policy is set to Fixed.
- Two hosts access the volume using opposite path order. For example, Host A is set up to access the lower-numbered LUN via SP A. Host B is set up to access the lower-numbered LUN via SP B.

Path thrashing can also occur if Host A lost a certain path and can use only paths to SP A while Host B lost other paths and can use only paths to SP B.

This problem could also occur on a direct connect array (such as AX100) with HBA failover on one or more nodes.

You typically do not see path thrashing with other operating systems:

- No other common operating system uses shared LUNs for more than two servers (that setup is typically reserved for clustering).
- For clustering, only one server is issuing I/O at a time. Path thrashing does not become a problem.

In contrast, multiple ESX systems may be issuing I/O to the same volume concurrently.

To resolve path thrashing:

- Ensure that all hosts sharing the same set of volumes on those active/passive arrays access the same storage processor simultaneously.
- Correct any cabling inconsistencies between different ESX hosts and SAN targets so that all HBAs see the same targets in the same order.
- Make sure the path policy is set to Most Recently Used (the default).

Resolving Issues with Offline VMFS Volumes on Arrays

On some arrays, it is not possible to display the volume with the same volume ID across hosts. As a result, the ESX system incorrectly detects the volume as a snapshot and places the volume offline.

Examples of storage arrays for which the same volume ID might not be visible for a given volume across hosts are Clariion AX100 and few IBM TotalStorage Enterprise Storage Systems (previously Shark Storage systems).

NOTE: If you use Clariion AX100 with Navisphere Express, you cannot configure the same volume ID across storage groups. You must use a version of Navisphere software that has more comprehensive management capabilities.

For IBM TotalStorage 8000, you need to recreate these volumes.

To resolve issues with invisible LUNs on certain arrays:

1. In the VI Client, select the host in the inventory panel.
2. Click the **Configuration** tab and click **Advanced Settings**.
3. Select **LVM** in the left panel and set LVM.DisallowSnapshotLUN to **0** in the right panel.
4. Rescan all LUNs.

After the rescan, all LUNs are available. You might need to rename the volume label.

Understanding Resignaturing Options

This section discusses how the EnableResignature and DisallowSnapshotLUN options interact, and explains the three states that result from changing these options.

State 1 — EnableResignature=no, DisallowSnapshotLUN=yes

This is the default state. In this state:

- You cannot bring snapshots of VMFS volumes by the array into the ESX host.
- Volumes formatted with VMFS must have the same ID for each ESX host.

State 1 is the safest state, but:

- It can cause problems with VMFS on some arrays (like IBM TotalStorage 8000 arrays and the EMC AX100) which do not always present the same volumes with the same ID everywhere unless you take special care.
- You lose the ability to snapshot a VMFS volume and bring it into an ESX system. To do that, change the LMV.EnableResignature setting to 1.

State 2 — EnableResignature=yes

DisallowSnapshotLUN is irrelevant in this state. In this state:

- You can bring snapshots of VMFS volumes into the same or different servers
- VMFS volumes containing volumes from IBM TotalStorage 8000 or AX100 that are not presented with the same volume numbers to all servers effectively lose the ability to use the virtual machines stored on that VMFS volume. Avoid this situation at all costs.

State 3 —- EnableResignature=no, DisallowSnapshotLUN=no

With these settings, snapshots should not be exposed to the ESX host. This is similar to ESX 2.x behavior.

If you have an IBM TotalStorage 8000 or AX100 that cannot be configured to present the same volume numbers to all servers for some reason, you need this setting to allow all ESX systems to use the same volumes for features like VMotion, VMware DRS and VMware HA.

When changing these settings, consider the following issues:

- If you create snapshots of a VMFS volume one or more times and dynamically bring one or more of those snapshots into an ESX host, only the first copy is usable. The usable copy is probably the primary copy. After reboot, you cannot determine which volume (the source or one of the snapshots) is usable. This non-deterministic behavior is dangerous.
- If you create a snapshot of a spanned VMFS volume, an ESX host might reassemble the volume from fragments that belong to different snapshots. This can corrupt your file system.

Resolving Performance Issues

For more information on resolving performance problems, see the following topics:

- "Managing ESX Performance Guarantees" on page 169
- "Optimizing HBA Driver Queues" on page 170
- "I/O Load Balancing Using Multipathing" on page 171
- "SAN Fabric Considerations for Performance" on page 172
- "Disk Array Considerations for Performance" on page 173

Also see the VMware Global Support Services Web site at www.vmware.com/support to access a range of support resources and to submit technical support requests.

A SAN Design Summary

The following table provides a summary list of VMware and SAN design questions and includes summary information on evaluating choices.

Design Option	Description
Which servers do you plan to deploy?	See the VMware *System Compatibility Guide*.
Which HBAs do you plan to deploy?	See the VMware *I/O Compatibility Guide*.
Which storage arrays do you plan to deploy?	See the VMware *SAN Hardware Compatibility List*.
Do you need NFS, iSCSI or FC storage?	See the VMware *SAN Hardware Compatibility List*.
Do you have redundancy planned for your SAN?	Redundancy should be at HBA, ISL, FC switch, and storage port processor levels.
How many virtual machines will you deploy?	Consider application demands and bandwidth availability.
Do you have enough zone space?	The larger the zone the longer zone propagation. Check `cfgsize` for zone space availability.
Do you have zone security?	Zone security eliminates HBA spoofing.
Do you have a strategy for disaster recovery?	Design ways to recover your data.
Do you need to back up your data?	Find a location to safeguard your data.
Do you need SAN extension to connect remote sites?	There are ways for you to connect remote SAN sites.
What type of SAN arrays do you need?	Consider the differences between active/passive and active/active arrays.
What type of HBAs do you need?	Consider multiports and speed of transfer.
What kind of RAID levels do you plan to use?	Consider different RAID levels for different application demands.
Is your application CPU, Memory or I/O intensive?	Adjust resources according to your application demands.
Should you use many small volumes or a few large volumes?	Consider advantages and disadvantages to both options.
Do you have the latest HBA BIOS?	Check your HBA vendor web site.
Do you have the latest SAN array controller firmware?	See the VMware *SAN Hardware Compatibility List.*
Do you have the latest SAN array disk firmware?	Check your SAN array Web site.

Design Option	Description
Do you have the latest VMware releases?	Check VMware downloads for latest patches.
Do you have the proper VMware licenses?	Check VMware downloads to redeem your licenses.
Do you have a license server?	Redeem the appropriate license for your server or host base.

B iSCSI SAN Support in VMware Infrastructure

With iSCSI SAN support in VMware ESX 3, you can configure ESX systems to use an IP network to access remote storage. This appendix provides a summary of recommendations and information about using VMware ESX with iSCSI. For detailed configuration procedures, see the following VMware product documentation:

- *ESX 3 Configuration Guide* (ESX 3.5 and VirtualCenter 2.5)
- *iSCSI SAN Configuration Guide* (ESX 3.5, ESX 3i version 3.5, VirtualCenter 2.5)

This appendix covers the following topics:

- Overview of iSCSI storage, including hardware and software iSCSI initiators, network, and storage arrays.
- Configuration of both hardware and software iSCSI initiators, a description of iSCSI initiator and target naming requirements, storage resource discovery methods, and iSCSI SAN operations, including multipathing, path switching and failover.
- iSCSI networking guidelines, including iSCSI SAN security.
- iSCSI configuration limits for both hardware and software iSCSI initiators with VMware ESX.
- iSCSI initiator configuration, including syntax and a description of vmkping, esxcfg –swiscsi, and esxcfg –hwiscsi commands for configuring both hardware and software iSCSI initiators for VMware Infrastructure.

iSCSI Storage Overview

When using VMware ESX with iSCSI, SCSI storage commands that a virtual machine issues to its virtual disk are converted into TCP/IP protocol packets. These packets are transmitted to a remote device, or target, that stores the virtual disk. To the virtual machine, the device appears as a locally attached SCSI drive.

To access remote targets, your ESX host uses iSCSI initiators. Initiators transport SCSI requests and responses between the ESX host and the target storage device on the IP network.

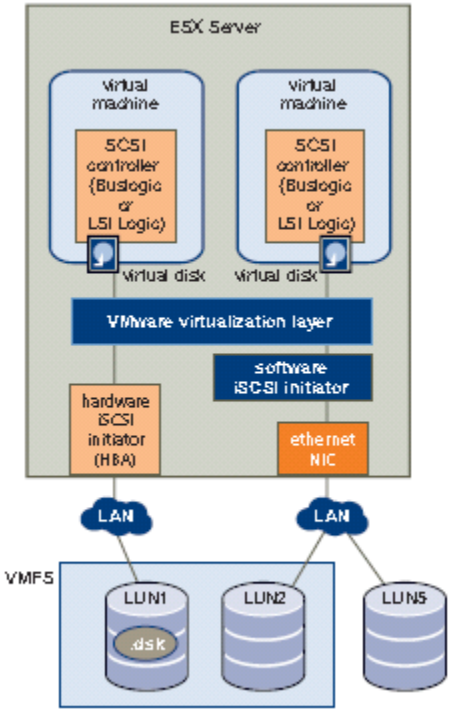


Figure B-1. iSCSI SAN Storage Virtualization

VMware ESX supports both hardware-based and software-based iSCSI initiators:

- **Hardware iSCSI initiator** – A third-party host bus adapter (HBA) with iSCSI over TCP/IP capability. This specialized iSCSI adapter is responsible for all iSCSI processing and management.
- **Software iSCSI initiator** – Code built into VMkernel that lets your ESX host connect to iSCSI storage devices through standard network adapters. The software initiator handles the iSCSI processing while communicating with the network adapter through the network stack. With the software iSCSI initiator, you can use the iSCSI protocol to access iSCSI storage without purchasing specialized hardware.

VMware Infrastructure that uses iSCSI supports nearly all the same features as other storage and Fibre Channel-based SANs. The following table lists specific VMware Infrastructure feature supported by hardware and software iSCSI initiators.

Table B-1. iSCSI Support for VMware Infrastructure Features

Support	VMFS	RDM	VMotion	HA	DRS	VCB	Boot VM	Boot from SAN	MSCS Clustering	MultiPath
iSCSI (HW)	✓	✓	✓	✓	✓	✓	✓	✓	No	✓
iSCSI (SW)	✓	✓	✓	✓	✓	✓	✓	No	No	✓

NOTES:

1. Clustering refers to the use of Microsoft Cluster Services (Windows 2003 and 2000) in a disk configuration shared between two virtual machines or a virtual machine and a physical system. Application-level clustering using MSCS on virtual machines is currently not certified with iSCSI and VMware ESX.
2. Boot from SAN is supported only for hardware iSCSI initiators.
3. Software-initiated iSCSI is supported fully in ESX 3.0 and later releases. Hardware-initiated iSCSI is supported in experimental mode only in ESX 3.0. (Hardware-initiated iSCSI is supported fully in ESX 3.0.1 and later with iSCSI arrays that have been qualified for use with supported hardware initiators.)
4. VMware Consolidated Backup 1.1 supports virtual machines backup and recovery directly from iSCSI storage. When used to back up virtual machines residing on a storage device accessed over a network connection, VMware Consolidated Backup can run in a virtual machine. For more details, see the *Virtual Machine Backup Guide* (ESX 3.5, ESX 3i version 3.5, VirtualCenter 2.5).

Configuring iSCSI Initiators

VMware Infrastructure supports connections to iSCSI arrays using either the software iSCSI initiator in VMkernel or a hardware initiator iSCSI HBA. See the *I/O Compatibility Guide For ESX 3.5 and ESX 3i* for a list of hardware initiators that you can use with VMware ESX. For a list of supported iSCSI storage devices with ESX 3.x, refer to the following guides:

- *Storage/SAN Compatibility Guide for ESX 3.5 and ESX 3i*
- *Storage/SAN Compatibility Guide for ESX 3.0.x*

iSCSI Storage – Hardware Initiator

The hardware iSCSI initiator uses a supported physical iSCSI host bus adapter (HBA). iSCSI HBAs provide connectivity to SANs over Ethernet and TCP/IP network infrastructure. When you use an iSCSI hardware initiator, the hardware initiator does not appear in the ESX host network configuration. Instead, it appears as a storage adapter in the Virtual Center storage configuration display. All disk I/O traffic flows through the hardware HBA. A hardware iSCSI initiator is a TCP/IP HBA, but is optimized for iSCSI traffic.

NOTE: If you are using a routed network, make sure that your router is capable of passing iSCSI traffic.

Configuring Hardware iSCSI Initiators and Storage

When VMware ESX communicates with iSCSI storage through hardware initiators, it uses a specialized third-party adapter that can access iSCSI storage over TCP/IP. This iSCSI adapter handles all iSCSI processing and management for the ESX host.

1. Install the hardware iSCSI HBA.
2. Configure the hardware iSCSI initiator.

   Before you begin configuring the hardware iSCSI initiator, make sure that the iSCSI HBA was installed successfully and appears in the list of storage adapters available for configuration.

   Also, configure the hardware iSCSI initiator before you set up the datastore that resides on an iSCSI storage device. If the initiator is installed, you can view its properties. While you configure the hardware iSCSI initiator, set up your initiator's iSCSI name, IP address, and discovery addresses. VMware also recommends that you set up CHAP parameters.
3. After you configure the hardware iSCSI initiator, perform a rescan so that all LUNs that the initiator can access appear in the list of available storage devices.
4. Configure the datastore.

NOTE: If you remove a target added by dynamic discovery, the target may be returned to the list the next time a rescan happens, the HBA is reset, or the system is rebooted.

iSCSI Storage – Software Initiator

Configuring the software iSCSI initiator requires a VMkernel port and a Service Console port. Unlike other VMkernel services, iSCSI has a service console component, so networks that are used to reach iSCSI targets must be accessible to both service console and VMkernel TCP/IP stacks. The software iSCSI initiator works with the ESX networking stack, which is implemented in VMkernel. The software iSCSI initiator works with a daemon that runs in the service console. The iSCSI daemon initiates the session and then handles login and authentication. After a connection is established, I/O between storage devices and virtual machines on the ESX host is handled by VMkernel. The software iSCSI initiator network and vswif (virtual switch interface) from the service console network must be on the same subnet. The service console can have more than one network port.

NOTE: Make sure that the VMkernel port TCP/IP address can ping the iSCSI storage array. The vmkping utility allows you to verify the VMkernel networking configuration. (For more information, see the Command Line Utilities section at the end of this appendix.)

Configuring Software iSCSI Initiators and Storage

With the software-based iSCSI implementation, you use a standard network adapter (NIC) to connect an ESX host to a remote iSCSI target on the IP network. The ESX software iSCSI initiator built into VMkernel provides this connection by communicating with the network adapter through the network stack.

1. Create a VMkernel port to handle iSCSI networking.

2. Configure a service console connection for software iSCSI communication.
3. Open a firewall port by enabling the iSCSI software client service.
4. Configure the software iSCSI initiator.
5. Rescan for new iSCSI LUNs.
6. Configure the datastore.

NOTE: When configuring the software iSCSI initiator through the VI Client, the user interface guides the user to use the same virtual switch for both the service console and the VMkernel connection to the iSCSI target.

After you install your iSCSI initiators and storage, you might need to modify your storage system configuration to ensure that it works properly with your ESX host. See the supported storage vendor's documentation for any additional configuration requirements or recommendations.

iSCSI Initiator and Target Naming Requirements

All iSCSI initiators and targets are assigned unique and permanent iSCSI names and addresses for network access. The iSCSI name provides a correct identification of a particular iSCSI device, an initiator or a target, regardless of its physical location. When you configure your iSCSI initiators, make sure they have properly formatted names. The initiators can use one of the following formats:

- **IQN (iSCSI qualified name)** – Can be up to 255 characters long and has the following format:

  `iqn.<year-mo>.<reversed_domain_name>:<unique_name>`

  where `<year-mo>` represents the year and month that your domain name was registered, `<reversed_domain_name>` is the official reversed domain name, and `<unique_name>` is any name you assign, for example, the name of your ESX host.
- **EUI (extended unique identifier)** – Represents the EUI prefix followed by the 16-character name. The name includes 24 bits for the initiator's company name assigned by the IEEE and 40 bits for a unique ID, such as a serial number.

Storage Resource Discovery Methods

To determine which storage resources on a network are available for access, VMware ESX provides two discovery methods:

- **Dynamic discovery** – Also known as Send Targets discovery. Each time the initiator contacts a specific iSCSI server, it sends the Send Targets request to the server. The server responds by providing a list of available targets to the initiator.
- **Static Discovery** –Using this method, the initiator detects the targets it must contact and uses their IP addresses and domain names to communicate with them. The initiator does not need to perform any discovery.

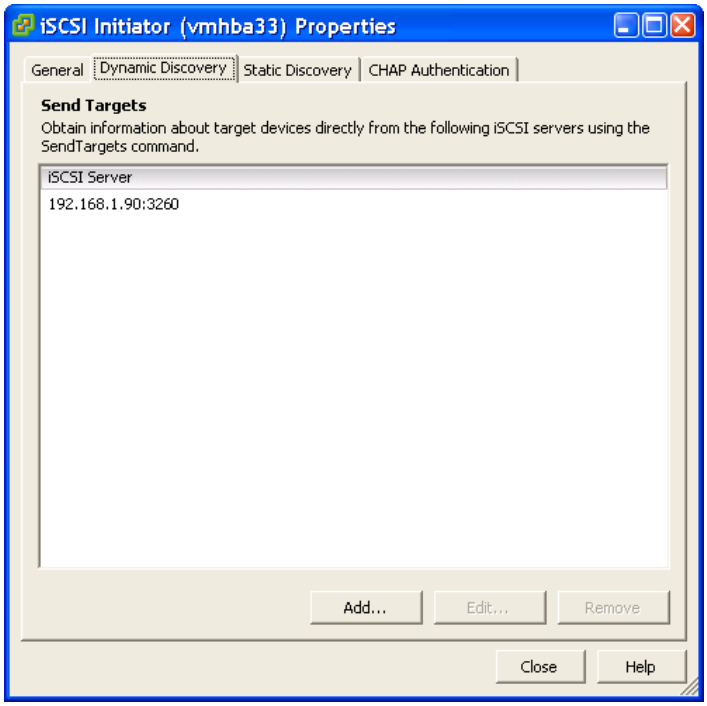


Figure B-2. iSCSI Initiator Discovery Methods

NOTE: The static discovery method is available only when the iSCSI storage is accessed through hardware iSCSI initiators.

For a detailed configuration procedure, see the "Configuring Storage" section of the chapter describing ESX storage in the *ESX 3 Configuration Guide* (ESX 3.5 and VirtualCenter 2.5) and the *iSCSI SAN Configuration Guide* (ESX 3.5, ESX 3i version 3.5, VirtualCenter 2.5).

Removing a Target LUN Without Rebooting

A target LUN for an ESX host is unrepresented after any existing active session terminates. If the LUN still appears in the VI client, there are two ways to remove it:

1. If, for any reason, there is a network break or disconnect, all existing sessions will terminate. When an ESX host tries to establish a new session with the LUN, it fails and a subsequent rescan removes the LUN from the ESX list of target LUNs.
2. If this is the only ESX host accessing the iSCSI LUN, you can also disable the LUN from the storage array, which would terminate the existing session.

NOTE: Disabling the software iSCSI initiator removes existing target LUNs. An ESX host reboot is required. The software iSCSI initiator can be disabled either through the VI client or using the esxcfg-swiscsi command utility.

Multipathing and Path Failover

The SAN uses multipathing techniques to transfer data between an ESX host and storage devices. Multipathing allows more than one physical path from the ESX host to a LUN on a storage system. If a path, or any component along the path—HBA or NIC, cable, switch or switch port, or storage processor—fails, the server selects another of the available paths to establish a connection. The process of detecting a failed path and switching to another available route is called path failover.

Path Switching with iSCSI Software Initiators

With software iSCSI, you can connect a single virtual VMKernel iSCSI network switch to multiple physical Ethernet adapters by using the VMware Infrastructure NIC teaming feature. NIC teaming provides network redundancy and rudimentary load balancing capabilities for iSCSI connections between ESX hosts and storage systems. Similar to the SCSI multipath capabilities, NIC teaming for iSCSI provides failover if connections or ports on the ESX host fail.

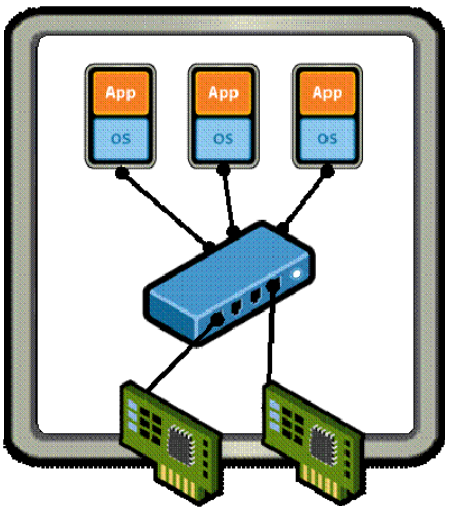


Figure B-3. NIC Teaming Provides Failover Functionality

Software iSCSI initiators establish only one connection to each target. Therefore, storage systems with a single target that contains multiple LUNs have all LUN traffic routed through that one connection. In a system that has two targets, with one LUN each, two connections are established between the ESX host and the two available volumes. For example, when aggregating storage traffic from multiple connections on an ESX host equipped with multiple iSCSI HBAs, traffic for one target can be set to a specific HBA, while traffic for another target uses a different HBA. For more information, see the “Multipathing” section of the *iSCSI SAN Configuration Guide*.

Currently, VMware ESX provides active/passive multipath capability. NIC teaming paths do not appear as multiple paths to storage in ESX host configuration displays, however. NIC teaming is handled entirely by the network layer and must be configured and monitored separately from ESX SCSI storage multipath configuration.

NOTE: ESX 3.0.x has a restriction that if the Ethernet adapters are teamed and one adapter fails, the other one takes over. Both adapters must be connected to the same physical switch and be on the same subnet (both NICs and iSCSI storage ports). ESX 3.5 does not have this restriction.

Path Switching with Hardware iSCSI Initiators

To support path switching, ESX hosts typically have two or more HBAs available from which storage systems can be reached, using one or more switches. Alternatively, an ESX host setup might include one HBA and two storage processors so that the HBA can use a different path to reach the storage system.

If an ESX host is equipped with multiple hardware iSCSI initiators, multipathing and failover is supported if the iSCSI storage array is listed as supported in the *SAN/Storage Compatibility Guide for ESX* and shown as supporting HBA or target failover.

NOTE: If storage arrays expose only single target, multiple portals, VMware ESX detects only one path because hardware iSCSI does not provide portal support.

Array-Based iSCSI Failover

Some iSCSI storage systems manage path use of their ports automatically (transparent to ESX hosts). When using one of these storage systems, VMware ESX does not detect multiple ports to the storage, and thus cannot choose the storage port it connects to. These systems have a single virtual port address that ESX hosts use for initial communication. During this time, the storage system can redirect ESX hosts to communicate with another port on the storage system. The iSCSI initiators in VMware ESX obey this reconnection request and connect with the different port. The storage system uses this technique to spread the load across available ports.

If an ESX host loses connection to one of these ports, it automatically attempts to reconnect with the virtual port of the storage system and should be redirected to an active, usable port. This reconnection and redirection happens quickly and generally does not disrupt running virtual machines. These storage systems can also request that iSCSI initiators reconnect to the system and change the storage port they are connected to. This allows the most effective use of multiple ports. The active path can be displayed through the VI Client or through the esxcfg-mpath command utility.

iSCSI Networking Guidelines

VMware ESX 3.5 supports multiple physical switches in two general configurations:

- iSCSI traffic is on its own dedicated subnet. The switches are trunked or interconnected using the ISL protocol with all switches in the same broadcast domain.
- You can set up a iSCSI virtual switch with two physical NIC uplink adapters and have each uplink adapter physically cabled to a separate physical switch on the same subnet (as long as VMkernel networking can reach the target array using either of the two NICs.

The figure below displays a software iSCSI configuration where each ESX NIC port is connected to a different physical switch in order to provide for path redundancy. The two switches are linked together. The IP configuration is all within the same subnet.

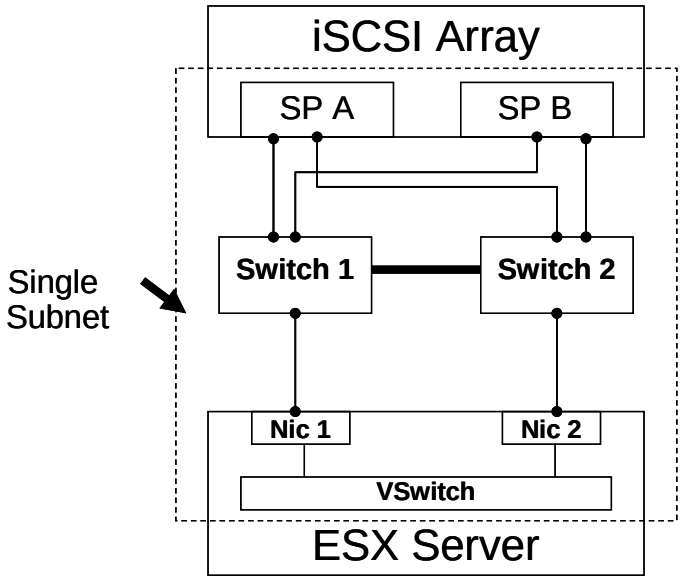


Figure B-4 Single Subnet Configuration

In this configuration, the physical switches and NICs are in different subnets. Two VMkernel ports are assigned addresses on separate subnets. The routing table takes care of directing the traffic to the appropriate address.

The figure below displays a software iSCSI configuration where each ESX NIC port is connected to a different physical switch to provide for path redundancy. The two switches are not linked together. In this configuration, the NIC ports can be in different IP subnets.

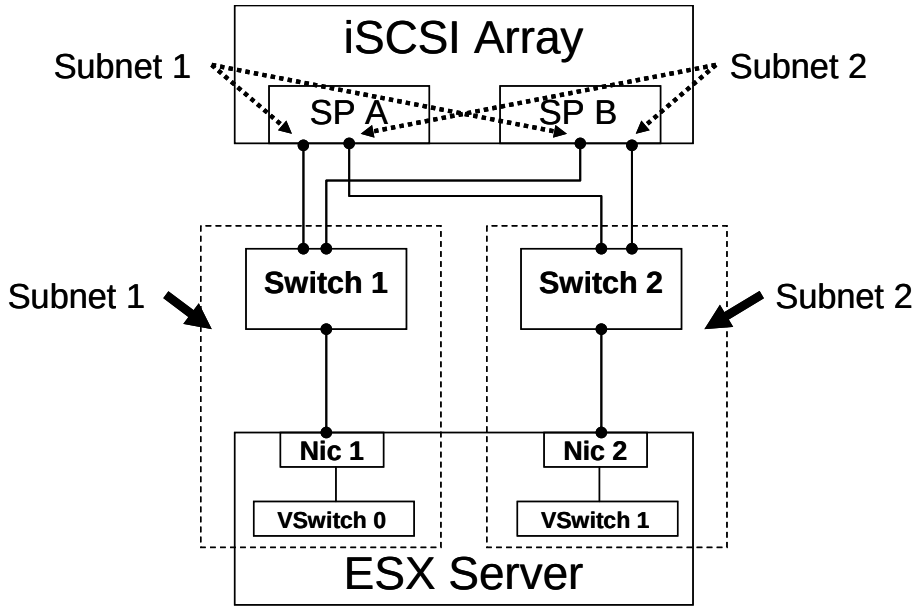


Figure B-5. Multiple Subnet Configuration

The following are some additional general design guidelines for software-based iSCSI initiators:

- All iSCSI servers must be either reachable by the default gateway or be in the same broadcast domain as the associated vSwitches.
- iSCSI with Jumbo Frames is not supported with ESX 3.0.x and 3.5.†
- The iSCSI software initiator is not available over 10GigE network adapters in ESX 3.5. †
- iSCSI storage system designs depend on the number of NIC ports available on the ESX host. Consider requirements for NIC teaming, VMotion, and port groups when designing your system.
- If possible, dedicate specific NICs to route iSCSI traffic. Segregate iSCSI traffic from general network traffic, either through separated physical networks or VLANs.
- For redundancy, configure the software iSCSI initiator to use at least two NICs.
- In multi-NIC scenarios, use teaming with:
  - "Virtual Source Port ID" setting – if all your iSCSI targets share the same IP address.
  - "IP Hash" setting – for all other scenarios, including a scenario in which you have multiple targets.
- If using multiple physical switches, observe the following recommendations, requirements, and restrictions:
  - Make sure the NICs are in the same broadcast domain (Figure B-4).
  - Do not use IP hash-based teaming policy across multiple physical switches.

- Enable link aggregation on the switch ports for IP hash-based teaming; many switches, however, do not allow link aggregation of ports spread across multiple physical switches.
- Configure physical switch Link Aggregation to be static and unconditional; there is no support for PAgP or LACP negotiation.

_____

† Feature support might change in future releases. Check new ESX releases and release notes for the latest support information.

Securing iSCSI SANs

Because iSCSI SAN technology uses IP networks to connect to remote targets, you must ensure the security of network connections. The IP protocol itself does not protect the data that it transports. The IP protocol cannot verify the legitimacy of initiators that access targets on the network. Take specific measures to guarantee storage security across IP networks.

To secure iSCSI devices from unwanted intrusion, require that an ESX host or initiator is authenticated by the iSCSI device or target whenever the host attempts to access data on the target LUN. The goal of authentication is to verify that the initiator has the right to access a target, a right granted when you configure authentication. You have two choices for authentication when setting up iSCSI access on ESX hosts:

- **Challenge Handshake Authentication Protocol (CHAP)** – You can configure the iSCSI SAN to use CHAP authentication. In CHAP authentication, when an initiator contacts an iSCSI target, the target sends a predefined ID value and a random value or key to the initiator. The initiator then creates a one-way hash value and sends it to the target. The hash contains three elements: a predefined ID value, the random value that the target sends, and a private value (or CHAP secret, that the initiator and target share. When the target receives the hash from the initiator, it creates its own hash value by using the same elements and compares it to the initiator's hash. If the results match, the target authenticates the initiator.

  The following display shows setup of the CHAP protocol for ESX hosts.

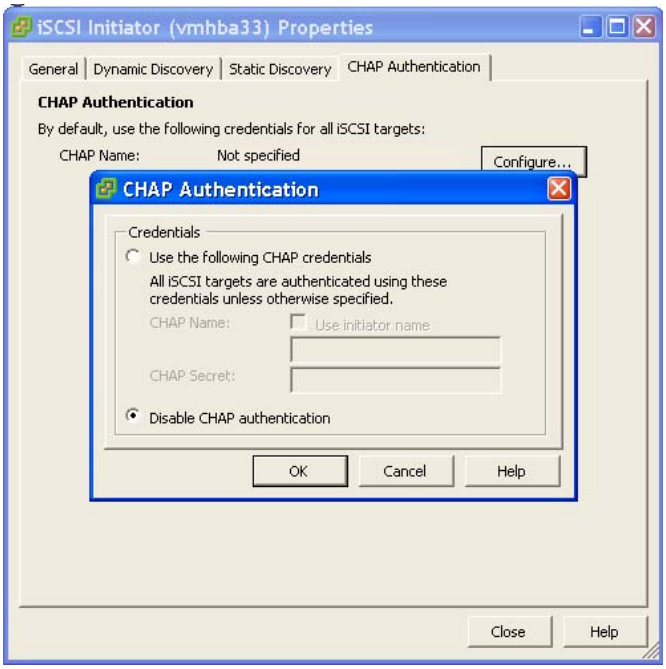


Figure B-6. CHAP Setup for iSCSI Initiators

The following list provides some additional details about enabling the CHAP protocol for VMware ESX operations:

- VMware ESX 3 supports only one-way CHAP authentication for iSCSI initiators. It does not support bi-directional CHAP. In one-way CHAP authentication, the target authenticates the initiator, but the initiator does not authenticate the target. The initiator has only one set of credentials, and all iSCSI targets use them.
- VMware ESX 3 supports CHAP authentication at the HBA level only. It does not support per-target CHAP authentication, which enables you to configure different credentials for each target for greater target security.
- The VI Client does not impose a minimum or maximum length for the CHAP secret you enter. However, some iSCSI storage devices require that the secret exceed a minimum number of characters, or they may have limitations on the specific characters you can enter as part of the secret. To determine specific iSCSI storage device requirements, check the device manufacturer's documentation.
- If you previously enabled CHAP and then disabled it, existing sessions remain active until you reboot your ESX host or until the storage system forces a logout. After that, you are no longer able to connect to those targets that require CHAP.

- **Disabled** – You can configure the iSCSI Initiator to use no authentication. Communication between the initiator and target are still authenticated in a rudimentary way, because iSCSI target devices are typically set up to communicate with specific initiators only.
  - Disable authentication only if you are willing to risk an attack to the iSCSI SAN.

VMware ESX 3 does not support Kerberos, Secure Remote Protocol (SRP), or public-key authentication methods for iSCSI Initiators. Additionally, iSCSI initiators do not support IPsec authentication and encryption.

Protecting an iSCSI SAN

When you plan your iSCSI storage configuration, take measures to improve the overall security of the iSCSI SAN. Your iSCSI configuration is only as secure as your IP network, so by enforcing good security standards when you set up your network, you help safeguard your iSCSI storage. Here are some specific suggestions and recommendations for protecting iSCSI storage:

- **Protect transmitted data** – A primary security risk in using iSCSI SANs is that an attacker might sniff transmitted storage data. VMware recommends that you take additional measures to prevent attackers from easily seeing iSCSI data. Neither the hardware iSCSI HBA nor the ESX host iSCSI initiator encrypts the data they transmit to and from the targets, which makes the data more vulnerable to sniffing attacks.
  - Allowing virtual machines to share virtual switches and VLANs used in an iSCSI storage configuration potentially exposes iSCSI traffic to misuse by a virtual machine attacker. To help ensure that intruders cannot intercept iSCSI transmissions, make sure that none of your virtual machines can see the iSCSI storage network. If you use a hardware iSCSI HBA, you can accomplish this by making sure that the iSCSI HBA and ESX physical network adapter are not inadvertently connected outside the host by virtue of sharing a switch or by some other means. If you configure iSCSI directly through the ESX host, you can provide this protection by configuring iSCSI storage through a different virtual switch than the one used by virtual machines.
  - If you configure iSCSI directly through the host rather than through a hardware adapter, create two network connections for the service console within the virtual network setup. Configure the first service console network connection on its own virtual switch and use it exclusively for management connectivity. Configure the second service console network connection so that it shares the virtual switch you use for iSCSI connectivity. The second service console network connection supports iSCSI activities only. Do not use it for any management activities or other management tool communication. To enforce a degree of separation between iSCSI and the service console on the shared virtual switch, configure them on different VLANs.
  - In addition to protecting iSCSI storage by giving it a dedicated virtual switch, VMware recommends that you also configure an iSCSI SAN on its own VLAN or physical LAN. Placing your iSCSI configuration on a separate LAN or VLAN ensures that no devices other than the iSCSI adapter have visibility into transmissions within the iSCSI SAN. Separate physical LANs also provide this security without having to take shared bandwidth into account. Do not configure the default gateway for the service console on the virtual switch you use for iSCSI connectivity. Instead, configure it on the virtual switch you use for management tool connectivity.
- **Secure iSCSI ports** – When you deploy iSCSI devices, ESX hosts do not open any ports that listen for network connections. This measure reduces the chances that an intruder can break into an ESX host through spare ports and gain control

over the host. Therefore, running iSCSI does not present any additional security risk at the ESX host end of the connection.

Any iSCSI target device must have one or more open TCP ports used to listen for iSCSI connections. If any security vulnerabilities exist in the iSCSI device software, your data can be at risk through no fault of an ESX host. To lower this risk, install all security patches that your storage equipment manufacturer recommends and limit the devices connected to the iSCSI network.

iSCSI Configuration Limits

The limits set for use of iSCSI arrays for VMware ESX 3.5 are the following:

Parameter	Initiator Type Used	Limit
Number of HBAs	Software	1
	Hardware	1 dual port or 2 single ports
Maximum number of targets	Both software and hardware initiator	64
Number of LUNs	Both software and hardware initiator	254
Number of paths to storage	Software	4
	Hardware	8

Only one software-based iSCSI initiator is supported per ESX 3 host and the maximum number of physical NICS is the same as the virtual switch maximum.

There is no benefit in using more than two physical NICs (in active-standby mode) in a NIC team unless there are multiple iSCSI targets. If there is only one target, all data goes through the same NIC. The standby is idle, in case the connection fails. Any other physical NICs present are also idle.

For the most up-to-date information on iSCSI configuration limits, see the Configuration Maximums for VMware Infrastructure 3 document available from the VMware web site.

Running a Third-Party iSCSI initiator in the Virtual Machine

Software iSCSI initiation from guest operating systems requires more specific system management, is less flexible, and has reduced mobility and portability. The prerequisites, conditions, and tradeoffs for running a third-party iSCSI initiator directly in the guest operating system are the following:

- The "portability" of virtual machines is compromised because it contains explicit registry or driver information about the storage device it is utilizing.
- You cannot boot the virtual machine unless you use PXE or some equivalent method to start the software iSCSI initiator.
- You cannot use VMotion on virtual machines because the ESX host cannot confirm the presence of the boot disk on the remote server.

- VMware DRS is not supported because VMotion is required.
- You must have a VMFS or NFS datastore to hold the virtual machine's metadata and swap files.
- You cannot perform a snapshot, suspend, or resume operation on virtual machines since this requires the current disk state to be captured, which cannot be done without a VMware Infrastructure datastore.
- You cannot clone virtual machines from a template.
- Handle all failover and MPIO-like functionality within virtual machines, since failover relies on the virtual machine functionality and is not guest-independent.
- You cannot utilize the VCB framework because you are not able to quiesce virtual machines and perform a snapshot.

iSCSI Initiator Configuration

There are three commands you can use to configure both software iSCSI initiators and hardware iSCSI initiators for VMware Infrastructure. These commands are vmkping, esxcfg –swiscsi, and esxcfg –hwiscsi. The syntax and a description of options available with these commands follow:

esxcfg-swiscsi Utility

Use the esxcfg-swiscsi utility to enable or disable software iSCSI initiator operation on ESX hosts.

Syntax:

`esxcfg-swiscsi [-e][-d][-h][-q][-s] <vmkernel SCSI adapter name>`

Option	Description:
-e	Enables software iSCSI.
-d	Disables software iSCSI. If you are using iSCSI volumes, do not use this option.
-q	Checks if software iSCSI is currently enabled (on) or disabled (off).
-s	Scans for disks available through the software iSCSI interface.
-h	Displays help information.

esxcfg-hwiscsi Utility

Use the esxcfg-hwiscsi utility to configure supported parameters for hardware iSCSI initiator operation.

Syntax:

`/sbin/esxcfg-hwiscsi [-l] [-a allow|deny] [-h] <vmkernel SCSI adapter name>`

Option	Description:
-l	Lists current configuration settings.
-a	Allows or disallows ARP redirection on adapter.
-h	Displays help information.

vmkping Utility

Use the vmkping utility to verify the VMkernel networking configuration.

Syntax:

`vmkping [options] [host | IP address]`

Option	Description:
-D	VMkernel TCP stack debug mode.
-c <*count*>	Sets packet count.
-i <*interval*>	Sets interval.
-s <*size*>	Sets send size.

Glossary

This section provides a description of commonly used acronyms and terms for SANs. Also see the common VMware Glossary for a description and explanation of other more general VMware Infrastructure terms.

A

access control

The access rights for an object. See also **permission**.

active/active disk array

A disk storage array with two dual controllers (or more) configured so both controller nodes to disk process I/O and provide a standby capability for the other. In VMware Infrastructure, 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).

active/passive disk array

A disk storage array with two dual controllers configured so that one controller is active while the other is idle, in standby mode, ready to take over I/O activity should the active primary controller fail or be taken offline. In VMware Infrastructure using an active/passive disk array, only one SP is actively servicing a given LUN. The other SP acts as backup for the LUN and might 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.

active path

In VMware Infrastructure, the path that the ESX host is using to issue I/O to a LUN. Note that in some SAN terminology, the term **active** refers to any path that is available for issuing I/O to a LUN.

adaptive scheme

An approach to defining the size and number and LUNs required for a virtual machine in which you pick and build a particular size LUN volume, then test applications using the LUN volume to determine if performance is acceptable. See “Making LUN Decisions” in the VMware *SANs Configuration Guide* for more information. See also **predictive scheme**.

availability

The general accessibility of a system or application to perform work or perform tasks when requested. See also **high availability**.

B

backup

The copying of data so that it can be restored in the event of a system data loss event, such as after a system failure or a disaster (called **disaster recovery**), or after data have been accidentally deleted or corrupted. Different strategies and types of backups exist, such as differential and file-level, that determine what and how data is backed up and recovered.

backup agent

The software residing on the virtual machine, ESX host, or another server (depending on backup method used) that is responsible for managing and scheduling the backup of data.

boot from SAN

An ESX configuration in which the ESX host boots from a volume configured on the SAN rather than from the server itself and locally attached storage.

C

cache

In general terms, refers to faster temporary memory storage that duplicates data stored elsewhere. VMware Infrastructure and VMware ESX use both read-ahead and write cache to optimize retrieval and write operations to disk storage.

clone

A duplicate of a virtual machine. When a clone is created, VirtualCenter provides an option for customizing the guest operating system of that virtual machine. Hosted products distinguish between full clones and linked clones. Also used as a verb, meaning to make a copy of a virtual machine.

cluster

A server group in a virtual environment. Clusters enable high availability solutions by providing alternate processing resources in case of failures.

compatibility mode

The virtualization type used for SCSI device access (physical or virtual).

controller cache

The cache memory used to store collections of data read from a disk (read-cache) or to collect data that will be written to disk (write-cache).

CPU virtualization

The pooling of physical CPU resources and processors so that virtual processing resources can be allocated to individual virtual machines

D

DAS

Direct attached storage, in which the SCSI device is directly connected to a server.

datacenter

In VMware terminology, a required structure under which hosts and their associated virtual machines are added to the VirtualCenter Server. VirtualCenter Server supports multiple datacenters. A host can be managed under only one datacenter.

data integrity

The assurance that the accuracy or correctness of data is maintained during any operation, such as transfer, storage, and retrieval.

data routers

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.

data replication

The data storage and backup strategy in which data is copied from one location to another, for the purpose of creating a duplicate copy. The replicated data location can be on the same array or on a different array at a remote location.

datastore

A virtual representation of combinations of underlying physical storage resources in the datacenter. A datastore is the storage location (for example, a physical disk, a RAID, or a SAN) for virtual machine files.

datastore extent

In the context of ESX operations, a datastore extent is a logical volume on a physical storage device that can be dynamically added to an existing VMFS-based datastore. The datastore can stretch over multiple extents, yet appear as a single volume analogous to a spanned volume.

diagnostic partition

A dedicated partition on local storage or a SAN volume that provides a location for debugging data, in the form of a core dump, that ESX hosts collect.

disaster recovery

The ability of a business to recover from critical data, hardware, and software failures and to restart or restore operations in the event of a natural or human-caused disaster.

Disk.MaxLUN

A configurable parameter specifying the maximum number of LUNs that VMkernel scans. (Changing this number might improve LUN discovery speed of rescans.) By default, VMkernel scans for LUN 0 to LUN 255 for every target (a total of 256 volumes).

disk partition

A part of hard disk that is reserved for a specific purpose. In the context of ESX storage, disk partitions on various physical storage devices can be reserved and formatted as datastores.

Disk.SupportSparseLUN

A parameter that can decrease the time needed for LUN scans or rescans if all LUNs in a specified range are not present or configured (that is, the range of configured LUNs is not contiguous). In those cases, changing this setting can decrease the time needed to scan for LUNs.

distributed file locking

An ESX locking method on a raw device mapping that makes it safe to use a shared raw volume without losing data when two virtual machines on different servers try to access the same volume. Raw device mapping makes it possible to use VMFS distributed locking for raw SCSI devices.

DRS

See **VMware DRS**.

dynamic name resolution

All raw device mapped volumes are uniquely identified by VMFS, and the identification is stored in its internal data structures. Any change in the SCSI path, such as an FC switch failure or the addition of a new host bus adapter, has the potential to change the vmhba device name, because the name includes the path designation (initiator, target, and LUN). Dynamic name resolution compensates for these changes by adjusting the data structures to retarget volumes to their new device names.

E

extent

Generally, a contiguous area of storage reserved for a file in a computer file system. In the context of ESX operations, an extent is a logical volume on a physical storage device that can be dynamically added to an existing VMFS-based datastore. The datastore can stretch over multiple extents, yet appear as a single volume analogous to a spanned volume.

F

fabric

An FC network topology, used in many SANs, in which devices pass data to each other through interconnecting switches. Fabrics are typically divided into zones. Also called **switched fabric** or **Fibre Channel fabric**. See also **FC (Fibre Channel)**.

fabric disruption

A glitch or failure in configuration of multiple FC switches to provide path between SAN storage and SAN storage array controllers due to events such as cabling problems, fabric merges, zone conflicts, and ISL failures.

failover

The capability of computer hosts, servers, and networks to switch over automatically to a redundant or standby system in the event of failure or abnormal termination of the previously active system. Failover normally occurs automatically. (Compare it with switchover, which generally involves human intervention.)

failover path

By default, at any given time, a VMware ESX system uses only one path from the host to a specific volume. If the path actively being used by the ESX system fails, the server selects another of the available paths, which is the failover path. The process of detecting a failed path by the built-in ESX multipathing mechanism and switching to another path is called **path failover**. A path fails if any of the components along the path fails, which can include the HBA, cable, switch port, or storage processor. This method of server-based multipathing can take up to a minute to complete, depending on the recovery mechanism used by the SAN components (that is, the SAN array hardware components).

fault

A data object containing information about an exceptional condition encountered by an operation.

fault tolerance

The graceful degradation and continued operation of a system despite the failure of some components.

fibre channel (FC)

An ANSI-standard, gigabit-speed network technology used to build storage area networks and to transmit data. FC components include HBAs, switches, and cabling.

Fibre Channel packets

The encapsulation of SCSI commands sent between an ESX HBA and the SAN fabric to request access to SAN storage devices.

Fibre Channel protocol

A protocol by which ESX hosts communicate with SAN storage processors to access storage devices.

fixed path policy

An ESX host setting in which the server always uses the preferred path, when available, to access storage.

G

guest operating system

An operating system that runs inside a virtual machine.

H

HA

See **high availability**.

HBA

A host bus adapter, a device that connects one or more peripheral units to a computer and manages data storage and I/O processing (often for FC, IDE, or SCSI interfaces). An HBA can be physical (attached to a host) or virtual (part of a virtual machine).

HBA device driver

In VMware Infrastructure, a modified standard Linux device driver in the ESX SCSI mid-layer that provides a way for virtual machines to access storage on a SAN through an FC HBA.

HBA failover

The process of one HBA taking over for another, in the event of an HBA failure or a failure in a path connection between a server and associated storage device in a SAN.

HBA timeout

A value set for an HBA driver that determines the timeout in detecting when a path to access storage fails.

high availability (HA)

A computer system design and implementation ensuring operational continuity of a system. See also **VMware HA**.

host

A physical computer capable of running virtual machines. Also called the “host machine” or “host computer.” In VMware Converter, the host is the physical computer on which the VMware Converter software is installed.

host agent

Software that, when installed on a virtual machine host, performs actions on behalf of a remote client.

host bus adapter (HBA)

A device that connects one or more peripheral units to a computer and manages data storage and I/O processing (often for FC, IDE, or SCSI interfaces). An HBA can be physical (attached to a host) or virtual (part of a virtual machine).

hyperthreading

Intel's implementation of the simultaneous multi-threading technology using the Pentium 4 microprocessor architecture.

I

initiator

The host-side endpoint that requests data transfers from SCSI targets. The initiator is the HBA.

IP storage

In VMware ESX 3, any form of storage that uses TCP/IP network communication as its foundation. Both NFS and iSCSI storage can be used as virtual machine datastores. NFS can also be used for direct mounting of .ISO files for presentation to virtual machines as CD-ROM discs.

iSCSI

Internet SCSI, a network protocol that allows the use of the SCSI protocol over TCP/IP networks.

J

journaling

A technique to prevent file system corruption, particularly in the updating of data, by maintaining journal logs that track changes made to data written to storage devices.

L

license server

A server that stores and allocates licenses.

load balancing

The balancing of traffic or load among multiple computers, processes, storage, or other resources in order to optimize resource utilization and decrease processing or lag time.

LUN

Logical unit number, an identifier for a disk volume in a storage array.

LUN device naming

In VMware Infrastructure, a sequence of three or four numbers, separated by colons, that is displayed in the VI Client to represent a LUN device or volume. For example, vmhba1:2:3 represents SCSI LUN3, attached to SCSI target 2, on SCSI HBA 1.

LUN masking

The selective presentation, access control, and partitioning of storage, allowing an administrator to configure storage so different servers or hosts only see certain LUNs or LUN volumes. Masking capabilities for each disk array are vendor-specific, as are the tools for managing LUN masking.

LUN rescan

ESX operation used to update information about HBAs, storage disks, LUNs, datastores, and zones available to virtual machines. Consider rescanning whenever any changes are made affecting storage or LUNs available to an ESX host, such as when changes are made to storage adapters, or existing datastores are edited or removed.

LUN resignature

See **volume resignaturing**.

LUN zoning

Defining a shared group of storage devices and LUN volumes and configuring the group for the necessary security and access rights. You can zone storage security and management, isolating the storage used in different environments or used by different departments or organizations. Zoning is also useful for grouping shared services for operations such as backup and recovery.

LVM

Logical volume manager.

M

mapping file

A VMFS file containing metadata used to map and manage a raw device. A synonym for raw device mapping.

mapping

An abbreviation for a raw device mapping.

mapped device

A raw device managed by a mapping file.

memory virtualization

The pooling of memory so that virtual memory resources can be allocated to individual virtual machines, in same cases, allocating memory greater than that available on an individual physical server or host.

metadata

Structured data that describes characteristics of other stored data entities to help in their identification, discovery, and management. In VMware Infrastructure, metadata stores information about entities such VMFS files, directories, symbolic links, and RDMs. Metadata is accessed each time the attributes of a file are accessed or modified.

metadata file

A mapping file.

Microsoft Clustering Services (MSCS)

Software that distributes data among the nodes of the cluster. If one node fails, other nodes provide failover support for applications such as databases, file servers, and mail servers.

migration

The process of moving a virtual machine between hosts. Unless you use, you must power off the virtual machine when you migrate it. See also **migration with VMotion** and **migration with VMware Converter**.

migration with VMotion

The process of moving a virtual machine that is powered on and has met selected requirements, including the activation of VMotion on both the source and target hosts. When you migrate a virtual machine using VMotion, the operations of the virtual machine can continue without interruption. See also **migration with VMware Converter**.

migration with VMware Converter

The process of moving a virtual machine that is powered off from a local or remote host, while reconfiguring the file format, if necessary, to accommodate the destination machine. See also **migration with VMotion**.

mirroring

The keeping or replication of exact copies of a set of data, typically as a way of providing access to multiple sources of the same information or providing a backup of important data.

multicore

A microprocessor that combines two or more independent processors into a single package, providing some level of parallel processing of tasks.

multipathing

In general, a fault tolerance technique that provides more than one physical path between associated storage devices through buses, controllers, switches, and bridge devices. In VMware Infrastructure, multipathing is having more than one path from a host to a LUN, so a given host is able to access a LUN on a storage array through more than one path. You can set preferences, or path policy, on how the current working or active paths are chosen.

A simple example is an FC disk connected to two FC controllers on the same computer or a disk connected to FC ports. Should one controller, port, or switch fail, the operating system can route I/O through the remaining controller transparently to the application, with no changes visible to the applications, other than perhaps incremental latency.

multipathing policy

For environments providing multipathing, you can choose a policy for your system, either Fixed or Most Recently Used. If the policy is Fixed, you can specify a preferred path. Each LUN (disk) that is visible to the ESX host can have its own path policy.

N

NAS

Network-attached storage.

NAT

Network address translation, a type of network connection that enables you to connect your virtual machines to an external network when you have only one IP network address and that address is used by the host computer. If you use NAT, your virtual machine does not have its own IP address on the external network. Instead, a separate private network is set up on the host computer. Your virtual machine gets an address on that network from the VMware virtual DHCP server. The VMware NAT device passes network data between one or more virtual machines and the external network. It identifies incoming data packets intended for each virtual machine and sends them to the correct destination.

network virtualization

A virtualization layer provided by VMware ESX to give virtual machines network connectivity similar to that provided by physical machines or servers, also providing isolation between individual virtual machines.

network adapter

An expansion card that provides a dedicated connection between a computer and a network.

network adapter teaming

The association of multiple network adapters with a single virtual switch to form a team. Such teams can provide passive failover and share traffic loads between members of physical and virtual networks.

NFS

Network file system, a protocol created by Sun Microsystems for accessing and sharing file systems across a computer network.

P

path failover

The process of detecting a failed path by the built-in ESX multipathing mechanism and switching to another. By default, at any given time, an ESX system uses only one path from the host to a specific volume. If the path actively being used by the ESX system fails, the server selects another of the available paths, which is the **failover path**. A path fails if any of the components along the path fails, which can include the HBA, cable, switch port, or storage processor. This method of server-based multipathing may take up to a minute to complete depending on the recovery mechanism used by the SAN components (that is, the SAN array hardware components).

path management

The management of different routes or paths to maintain a constant connection between server machines and storage devices in case of the failure of an HBA, switch, SP, or FC cable. Path management lets you balance loads among available paths to improve performance.

To support path switching, the server typically has two or more HBAs available from which the storage array can be reached, using one or more switches. Alternatively, the setup can include one HBA and two storage processors so that the HBA can use a different path to reach the disk array. You can set up your ESX host to use different paths to different LUNs by changing the preferred path for the different HBAs.

path policy

Method by which VMware ESX chooses a path to access LUN volumes in a multipathing SAN environment. At any given time, an ESX system uses only one path from the host to a specific LUN volume. VMware provides settings known as path policy (Fixed and Most Recently Used) that control and determine how active paths are chosen, and how VMware ESX uses alternate paths to SAN storage. Each LUN volume visible to ESX hosts can have its own path policy.

path redundancy

See **multipathing**.

path thrashing

Also called **LUN thrashing**. A situation that sometimes occurs in multipath SAN environments to limit access to LUN volumes or lower I/O throughput. SPs are like independent computers that control access to shared storage. When two ESX hosts attempt to access the same volume via different SPs, cache and volume locking with certain types of disk arrays and settings can prevent either host from effectively accessing and completing write operations to the volume. For more information, see "Understanding Path Thrashing" on page 182.

PCI bus

Peripheral Component Interconnect bus, the standard for most server computer motherboards, which provides a way to attach peripheral devices to the computer, through cards such as HBAs and network adapters.

permission

A data object consisting of an authorization role, a user or group name, and a managed entity reference. A permission allows a specified user to access the entity (such as a virtual machine) with any of the privileges pertaining to the role.

physical disk

In hosted products, a hard disk in a virtual machine that is mapped to a physical disk drive or partition on the host machine. A virtual machine's disk can be stored as a file on the host file system or on a local hard disk. When a virtual machine is configured to use a physical disk, VirtualCenter directly accesses the local disk or partition as a raw device (not as a file on a file system). See also **virtual disk**.

Policy

A formal set of guidelines that control the capabilities of a virtual machine. Policies are set in the policy editor.

Port

See **SAN port**.

port group

A construct for configuring virtual network options, such as bandwidth limitations and VLAN tagging policies for each member port. Virtual networks connected to the same port group share network policy configuration. See also **virtual network**.

predictive scheme

AN approach to defining the size and number and LUNs required for a virtual machine, in which you create several volumes with different storage characteristics and match applications to the appropriate storage based on their requirements. See "Making LUN Decisions" in the VMware *SANs Configuration Guide* for more information. See also **adaptive scheme**.

preferred path

An approach to defining the size and number and LUNs required for a virtual machine, in which you define and build several different-sized LUN volumes and test applications with each one to determine which size works best. See "Making LUN Decisions" in the VMware *SANs Configuration Guide* for more information. See also **adaptive scheme**.

provisioning

The creation of a logical volume.

proxy server

In general, a computer that acts as an intermediary for other computers to provide a network service. In VMware Infrastructure, proxy servers can provide backup services for virtual machines, reducing the overhead on ESX hosts in providing those services.

Q

quiescing

Computer activity or processing reaching a quiet or steady and unchanging state, which is the state in which you would create a backup or snapshot of a virtual machine.

R

RAID

Redundant array of independent (or inexpensive) disks. See also **raw disk** and **physical disk**.

RAS

Reliability, availability, and scalability—the criteria by which customers commonly evaluate infrastructure solutions for purchase and deployment.

raw device

Any SCSI device accessed by a mapping file.

raw device mapping (RDM)

A mechanism that enables a virtual machine to have direct access to a volume on the physical storage subsystem (FC or iSCSI only). At the same time, the virtual machine has access to the disk using a mapping file in the VMFS name space.

raw disk

A disk volume accessed by a virtual machine as an alternative to a virtual disk file. It can be accessed through a mapping file. See also **physical disk**.

raw volume

A logical disk located in a SAN or provided by other storage devices.

redundancy

In general, the provisioning of additional or duplicate systems, equipment, or components that function in case a currently operating component or system fails. In storage systems, redundancy means providing additional switches, HBAs, and storage processors – in effect, creating redundant access paths to storage in the event of individual component failures.

redundant I/O paths

Provisioning to provide redundant I/O paths from ESX hosts to storage arrays that can be switched in the event of a port, device, cable, or path failure. Provides fault tolerance 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 volumes.

replication

Use of redundant resources to improve reliability, fault tolerance, or performance. In storage and VMware Infrastructure, replication refers to additional data protection and features such as snapshots, internal copies, and remote mirroring of storage resources in the event of system or component failure, and recovery from other planned and unplanned disasters and downtime.

rescan

A built-in ESX and VMFS functionality that detects changes in LUNs and LUN mapping automatically. During a rescan, VMware ESX also identifies all available paths to a storage array and collapses it to one single active path (regardless of how many paths are available). All other available paths are marked as standby.

reservation

In storage operations, a method that VMFS uses to provide on-disk distributed locking to ensure that the same virtual machine is not powered on by multiple servers at the same time. It is also used to coordinate access to VMFS metadata.

resignaturing

A VMFS capability that allows you to make a hardware snapshot of a volume (that is configured as either a VMFS or a RDM volume) and access that snapshot from an ESX system. This involves resignaturing the volume UUID and creating a new volume label.

As a rule, a volume should appear with the same volume ID or LUN to all hosts that access the same volume. To mount both the original and snapshot volumes on the same ESX host, you can turn on auto-resignaturing using the LVM.EnableResignature parameter.

resource pool

A division of computing resources used to manage allocations between virtual machines.

resume

To return a virtual machine to operation from its suspended state. When you resume a suspended virtual machine, all applications are in the same state they were when the virtual machine was suspended. See also **suspend**.

RSCN

Registered state change notification. A network switch function that sends out notifications of fabric changes, such as targets joining or leaving the fabric, to specified nodes. The initiator registers with the SAN fabric for receiving notifications of any changes to the domain server

S

SAN

Storage area network. A large-capacity network storage device that can be shared among multiple ESX hosts. A SAN is required for VMotion.

SAN fabric

The configuration of multiple FC switches connected together, this is the actual network portion of the SAN. The fabric can contain between one and 239 switches. (Multiple switches provide redundancy.) Each FC switch is identified by a unique domain ID (from 1 to 239). FC protocol is used to communicate over the entire network. A SAN can consist of two separate fabrics for additional redundancy.

SAN management agents

Management software run inside a virtual machine that can manage storage on raw devices and issue hardware-specific SCSI commands to access data.

SAN port

In VMware Infrastructure, 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 using a WWPN, which provides a globally unique identifier for a port or a port ID (or port address) that serves as the FC address for the port. This enables routing of data through the SAN to that port. See also **WWPN**.

SAN switches

Devices that connect various elements of the SAN together, such as HBAs, other switches, and storage arrays. Similar to networking switches, SAN switches provide a routing function. SAN switches also allow administrators to set up path redundancy in the event of a path failure, from a host server to a SAN switch, from a storage array to a SAN switch, or between SAN switches.

SAN zoning

A logical grouping of storage that serves to provide access control, separate environments (for example, staging versus production), and define the HBAs that can connect to specific storage processors or SPs in the same group or zone. Zoning also has the effect of controlling and isolating paths within a fabric and preventing non-ESX systems from seeing a particular storage system and possibly destroying ESX VMFS data.

SATA

Serial advanced technology attachment, also called **Serial ATA**. A standard, based on serial signaling technology, for connecting computers and hard drives.

SCSI

Small computer system interface. One of several standards available for physically connecting and transferring data between computers and peripheral devices and storage.

shared storage

A feature of VMFS that allows multiple physical servers to share, read, and write to the same storage simultaneously. In a simple configuration, the virtual machines' disks are stored as files within a VMFS. When guest operating systems issue SCSI commands to their virtual disks, the virtualization layer translates these commands to VMFS file operations. ESX systems also use VMFS to store virtual machine files. To minimize disk I/O overhead, VMFS has been optimized to run multiple virtual machines as one workload.

share

The allocation of resources, including CPU, memory, network, and storage, to ESX hosts and virtual machines, typically based on virtual machine prioritization and load requirements.

SMB

Small to medium-size business.

SMP

Symmetric multiprocessor computer architecture, in which two or more identical processors are connected to a single, shared, main memory.

Snapshot

A reproduction of the virtual machine at a single point in time, including the state of the data on all the virtual machine's disks and the virtual machine's power state (on, off, or suspended). You can take a snapshot of a virtual machine at any time and go to any snapshot at any time. You can take a snapshot when a virtual machine is powered on, powered off, or suspended. You can configure a virtual machine to exclude specified disks from snapshots.

spanned volume

In VMware ESX, VMFS volumes or datastores that span multiple extents (logical volumes on one or more physical storage devices), yet still appear as a single volume. A VMFS volume can be spanned to a maximum of 32 physical storage extents including SAN volumes and local storage. Each volume can have a block size up to 8MB, which allows for a maximum volume size of up to 2TB, so 32 physical storage extents provide a maximum volume size of up to 64TB.SP

SP failover

In a multipathing environment, in which ESX hosts have two or more HBAs available and a storage array can be reached via alternate switches, SP failover describes the process of changing paths when one of the FC switches fails or the links to one of the disk array storage processors fails.

storage adapter

Also called HBA. The types of storage adapters that can be configured for SAN are FC SCSI and iSCSI.

storage or disk array

Disk arrays are groups of multiple physical disk storage devices. For FC SAN arrays, iSCSI SAN arrays, and NAS arrays, the array storage devices are typically all configured with some level of RAID, which can be used to provide storage for VMware Infrastructure and ESX hosts.

storage port

A port on a storage array that is part of a particular path by which the ESX host can access SAN storage from a specific HBA port in the host and fabric switches.

storage port redundancy

A configuration in which the ESX host is attached to multiple storage ports, so ESX can failover and recover from individual storage port and switch failures.

storage port failover

In a multipathing environment, the process of failing over to an alternate path using a different storage path to a disk array if the storage port in the current active path fails.

storage processor (SP)

A component of SAN systems that processes HBA requests routed through an FC switch, handles the RAID/volume functionality of the disk array, and manages access and other operations performed on the disk array.

storage virtualizer

A system that abstracts and aggregates physical storage on behalf of a storage-using client.

suspend

To save the current state of a running virtual machine. To return a suspended virtual machine to operation, use the resume feature. See also **resume**.

switched fabric

A network topology in which devices such as storage disks connect with each other via switches. Switched fabric networks are different from typical switched networks, such as Ethernet networks, in that switched fabric networks more naturally support redundant paths between devices, and thus failover and increased scalability.

T

template

A master image of a virtual machine. This typically includes a specified operating system and a configuration that provides virtual counterparts to hardware components. A template can include an installed guest operating system and a set of applications. Setting a virtual machine as a template protects any linked clones or snapshots that depend on the template from being disabled inadvertently. VirtualCenter uses templates to create new virtual machines. See also **snapshot**.

timeout

The maximum time before specific operations such as path detection, I/O operation completion, and HBA failover are determined to have failed,

V

VCB

VMware Consolidated Backup. A VMware backup solution that provides a centralized facility for agentless backup of virtual machines. VCB works in conjunction with a third-party backup agent residing on a separate backup proxy server (not on the ESX system) that manages backup schedules.

VCB proxy

A separate server on which a third-party backup agent is installed. This agent manages the schedule for backing up virtual machines. VCB integration modules provide pre-backup scripts run before backup jobs are started for supported third-party backup applications.

virtual disk

A file or set of files that appears as a physical disk drive to a guest operating system. These files can be on the host machine or on a remote file system. See also **physical disk**.

virtual clustering

A clustering configuration in which a virtual machine on an ESX host acts as a failover server for a physical server. Because virtual machines running on a single host can act as failover servers for numerous physical servers, this clustering method provides a cost-effective N+1 solution.

virtual disk snapshot

A point-in-time copy of a LUN volume. Snapshots provide backup sources for the overall backup procedures defined for storage arrays.

virtual infrastructure

In general terms, the abstraction of computer resources that hides the physical characteristics of actual machines from systems, applications, and end users that interact with these new virtual resources. Virtualization can make a single physical resource (such as a server, operating system, application, or storage device) appear as multiple logical resources or make physical resources (such as storage devices and servers) appear as a single logical resource.

virtualization layer

Software such as VMware ESX and other VMware desktop virtualization solutions that abstract the processor, memory, storage, and networking resources of a physical host computer into multiple virtual machines. Virtual machines are created as a set of configuration and disk files that together perform all the functions of a physical computer.

virtual machine (VM)

A virtualized x86 PC environment in which a guest operating system and associated application software can run. Multiple virtual machines can operate on the same host system concurrently.

virtual machine cloning

The replication of virtual machines with similar or identical characteristics which, in VMware Infrastructure, are created using templates that define configuration options, resource allocation, and other attributes.

virtual machine cluster

A configuration of virtual machines accessing the same virtual disk file or using a mapping file to access raw device storage.

virtual SCSI

The virtualization of SCSI HBAs that provides access to virtual disks as if they were a physical SCSI drive connected to a physical SCSI adapter or HBA. Each virtual disk accessible by a virtual machine (through one of the virtual SCSI adapters) resides in VMFS or NFS storage volumes, or on a raw disk. Whether the actual physical disk device is being accessed through SCSI, iSCSI, RAID, NFS, or FC controllers is transparent to the guest operating system and to applications running on the virtual machine.

virtual SMP

The technology that enables a virtual machine to do symmetric multiprocessing. Virtual SMP enables you to assign two virtual processors to a virtual machine on any host machine that has at least two logical processors.

virtual machine file system (VMFS)

A file system optimized for storing virtual machines. One VMFS partition is supported per SCSI storage device or volume. Each version of VMware ESX uses a corresponding version of VMFS. For example, VMFS3 was introduced with ESX 3.

VMFS partition

A part of a VMFS volume that provides data (application and other data) to individual virtual machines. With VMware Infrastructure 3, you can have a maximum of 16 partitions per volume.

VMFS volume

A logical unit of VMFS storage that can use disk space on one or more physical storage devices or disks. Multiple virtual machines can share the same VMFS volume.

VMkernel

In VMware ESX, a high-performance operating system that occupies the virtualization layer and manages most of the physical resources on the hardware, including memory, physical processors, storage, and networking controllers.

vmkfstools

Virtual machine kernel files system tools, providing additional commands that are useful, for example, when you need to create files of a particular block size and to import files from and export files to the service console's file system.

The vmkfstools commands are designed to work with large files, overcoming the 2GB limit of some standard file utilities. For a list of supported vmkfstools commands, see the VMware *Server Configuration Guide*.

VMware Consolidated Backup (VCB)

A VMware backup solution that provides a centralized facility for agentless backup of virtual machines. VCB works in conjunction with a third-party backup agent residing on a separate backup proxy server (not on the ESX system) that manages backup schedules.

VMware Distributed Resource Scheduler (DRS)

In VirtualCenter and VMotion, a feature that intelligently and continuously balances virtual machine workloads across your ESX hosts. VMware DRS detects when virtual machine activity saturates an ESX host and triggers automated VMotion live migrations, moving running virtual machines to other ESX nodes so that all resource commitments are met.

VMware ESX

Virtualization layer software that runs on physical servers to abstract processor, memory, storage, and networking resources to be provisioned to multiple virtual machines running mission-critical enterprise applications.

VMware High Availability (VMware HA)

An optional feature that supports distributed availability services in an environment that includes VMware ESX and VirtualCenter. If you have configured DRS, and one of the hosts managed by VirtualCenter Server goes down, all virtual machines on that host are immediately restarted on another host. See also **VMware Distributed Resource Scheduler** and **VMware VirtualCenter**.

VMware Infrastructure 3

A software suite, including VMware ESX and VirtualCenter, that virtualizes servers, storage, and networking, and enables multiple unmodified operating systems and their applications to run independently in virtual machines while sharing physical resources. The suite delivers comprehensive virtualization, management, resource optimization, application availability, and operational automation capabilities. See also VMware VirtualCenter.

VMware Infrastructure SDK

A developer’s kit, providing a standard interface for VMware and third-party solutions to access VMware Infrastructure.

VMware Service Console

An interface to a virtual machine that provides access to one or more virtual machines on the local host or a remote host running VirtualCenter. You can view the virtual machine's display to run programs within it or to modify guest operating system settings. In addition, you can change the virtual machine's configuration, install the guest operating system, or run the virtual machine in full screen mode.

VMware VirtualCenter Management Server

A software solution for deploying and managing virtual machines across a datacenter. With VirtualCenter, datacenters can instantly provision servers, globally manage resources, and eliminate scheduled downtime for hardware maintenance.

VMware VirtualCenter database

A persistent storage area for maintaining the status of each virtual machine and user that is managed in the VirtualCenter environment. Located on the same machine as the VirtualCenter Server. See also **VirtualCenter**.

VMware VirtualCenter license server

VMware Infrastructure software installed on a Windows system that authorizes VirtualCenter Servers and ESX hosts appropriately for your licensing agreement. There is no direct interaction with the license server. Administrators make changes to software licenses using the VI Client.

VMware VirtualCenter Management Server

The central point in VMware Infrastructure 3 for configuring, provisioning, and managing virtualized IT infrastructure.

VMware Virtual Infrastructure (VI) Client

A Windows-based user interface tool that allows administrators and users to connect remotely to the VirtualCenter Management Server or individual ESX installations from any Windows PC.

VMware virtual machine monitor (VMM)

Module whose primary responsibility is to monitor a virtual machine's activities at all levels (CPU, memory, I/O, and other guest operating system functions and interactions with VMkernel). Specific to storage, the VMM module contains a layer that emulates SCSI devices within a virtual machine.

VMware VMotion

A feature that enables you to move running virtual machines from one ESX system to another without interrupting service. It requires licensing on both the source and target hosts. VMotion is activated by the VirtualCenter agent, and VirtualCenter Server centrally coordinates all VMotion activities. See also **migration with VMotion**.

VMware Virtual Infrastructure Web Access

A Web interface in VMware Infrastructure 3 that provides virtual machine management and remote consoles access similar to, but not providing all of the capability of, the Virtual Infrastructure (VI) Client.

volume

An allocation of storage. The volume size can be less than or more than the size of a physical disk drive. An allocation of storage from a RAID set is a volume or a logical volume.

volume resignaturing

The modification of volume and device metadata so that VMware Infrastructure 3 can access both the original volume and its replicated volume simultaneously. Volume resignaturing allows you to make a hardware snapshot of a volume (that is either configured as VMFS or a RDM volume) and access that snapshot from an ESX system. Doing this involves resignaturing the volume UUID and creating a new volume label.

W

WWN

World Wide Name. A globally unique identifier for an FC initiator or a storage controller.

WWPN

World Wide Port Name. A unique identifier for a port that allows certain applications to access the port. The FC switches discover the WWPN of a device or host and assign a port address to that device. To view the WWPN using a VI Client, click the host's **Configuration** tab and choose **Storage Adapters**. You can then select the storage adapter for which you want to see the WWPN.

Z

zoning

See **SAN zoning** and **LUN zoning**.

Index

O

P

R

S

T

V

W

Z