EMC²
where information lives®

White Paper

Using EMC Celerra IP Storage with VMware Infrastructure 3 over iSCSI and NFS

Best Practices Planning

Abstract

This white paper provides a detailed overview of the EMC® Celerra® network-attached storage product with VMware Infrastructure 3. VMware Infrastructure 3 is a package that includes ESX Server 3 software along with management products such as VMotion, VirtualCenter 2, Virtual SMP, and resource pools. It also includes a new VMware file system and features such as Distributed Resource Scheduler, High Availability, and Consolidated Backup. This white paper covers the use of Celerra protocols and features in support of the VMware Infrastructure 3 environment.

April 2007

Part Number H2756

Table of Contents

Executive summary

With the release of VMware Infrastructure 3, virtual hardware support was extended to include the use of IP storage devices. This support enables ESX environments to take full advantage of the NFS protocol and IP block storage using iSCSI. This significant improvement provides a method to tie virtualized computing to virtualized storage, offering a dynamic set of capabilities within the data center and resulting in improved performance and system reliability. This white paper describes how best to utilize EMC® Celerra® in a VMware ESX environment.

Introduction

Virtualization has been a familiar term used in technical conversations for many years now. We have grown accustomed to almost all things virtual from virtual memory to virtual networks to virtual storage. In the past several years there has been considerable momentum in the virtual server space with VMware products providing a framework for companies to better partition and utilize their hardware resources. With this architecture, x86-based systems can run independent virtual environments, with different operating systems and resource requirements, all within the same physical server.

Several VMware products offer the ability to virtualize hardware resources, however, the focus in this white paper will be on the ESX Server product due to its support of Celerra IP storage. With VMware Infrastructure 3, Celerra iSCSI LUNs and NFS exported file systems provide the storage devices used by ESX Server to create datastores for virtual machines and virtual disks.

Prior to VMware Infrastructure 3, the only supported storage was either locally-attached or SAN-attached devices. ESX Server 3 introduces two additional options for NFS and iSCSI storage devices, both of which are offered by Celerra Network Server. In addition to the network storage connectivity options, Celerra provides advanced scalability and reliability characteristics, combining some of the benefits of Fibre Channel with the flexibility and ease of use that come with IP storage.

The iSCSI and NFS capabilities provided in ESX Server 3 differ from previous versions of ESX Server, as well as current versions of the VMware GSX and Workstation products. Those products offer no explicit IP storage connectivity and require the guest OS to use one of the virtual machine interfaces as well as a software iSCSI or NFS client to access the network storage system. Those access methods may be suitable in many environments, but direct IP storage to Celerra as provided in ESX Server 3 does provide performance benefits that do not exist when accessing storage through the guest OS.

Audience

Anyone interested in how ESX 3 can be integrated with NFS and iSCSI storage will benefit from this white paper, however, system administrators and architects who are familiar with server virtualization and want to understand how Celerra can be used to support ESX will benefit most. A working knowledge of ESX Server and Celerra is helpful in understanding the concepts presented.

Terminology

Fail-Safe Network (FSN) - A high-availability feature that extends link failover out into the network by providing switch-level redundancy. An FSN appears as a single link with a single MAC address and potentially multiple IP addresses.

iSCSI target - An iSCSI endpoint, identified by a unique iSCSI name, which executes commands issued by the iSCSI initiator.

Link aggregation - A high-availability feature based on the IEEE 802.3ad Link Aggregation Control Protocol (LACP) standard allowing Ethernet ports with similar characteristics to the same switch to combine into a single virtual device/link with a single MAC address and potentially multiple IP addresses.

LUN - For iSCSI on a Celerra Network Server, a logical unit is an iSCSI software feature that processes SCSI commands, such as reading from and writing to storage media. From a iSCSI host perspective, a logical unit appears as a disk device.

NFS - A distributed file system providing transparent access to remote file systems. NFS allows all network systems to share a single copy of a directory.

RAID 1 - RAID method that provides data integrity by mirroring (copying) data onto another disk in the LUN. This RAID type provides the greatest data integrity at the greatest cost in disk space.

RAID 5 - Data is striped across disks in large stripes. Parity information is stored so data can be reconstructed if needed. One disk can fail without data loss. Performance is good for reads but slow for writes.

Celerra Replication Service - A service that produces a read-only, point-in-time copy of a source file system. The service periodically updates the copy, making it consistent with the source file system.

Replication failover - The process that changes the destination file system from read-only to read/write and stops the transmission of replicated data. The source file system, if available, becomes read-only.

SnapSure™ - On a Celerra system, a feature providing read-only point-in-time copies of a file system. They are also referred to as checkpoints.

Virtual network device - A combination of multiple physical devices defined by a single MAC address.

Virtual provisioned LUN - An iSCSI LUN without reserved space on the file system. File system space must be available for allocation whenever data is added to the LUN

Virtual local area network (VLAN) – A group of devices physically residing on different network segments but communicating as if they resided on the same network segment. VLANs are configured by management software and are based on logical versus physical connections for increased administrative flexibility.

Celerra overview

The Celerra NAS product family covers a broad range of configurations and capabilities that scale from the midrange to the high end of networked storage. Although differences exist along the product line, there are some common building blocks. These basic building blocks are combined to fill out a broad, scalable product line with consistent support and configuration options.

A Celerra frame provides n+1 power and cooling redundancy and supports a scalable number of physical disks, depending on the model you select and the needs of your solution. For the purposes of this white paper, there are two additional building blocks that are important to discuss:

- Data Movers
- Control Stations

Data Movers move data back and forth between the LAN and back-end storage (disks). The Control Station is the management station for the system. Celerra is configured and controlled via the Control Station. Figure 1 shows how Celerra works.

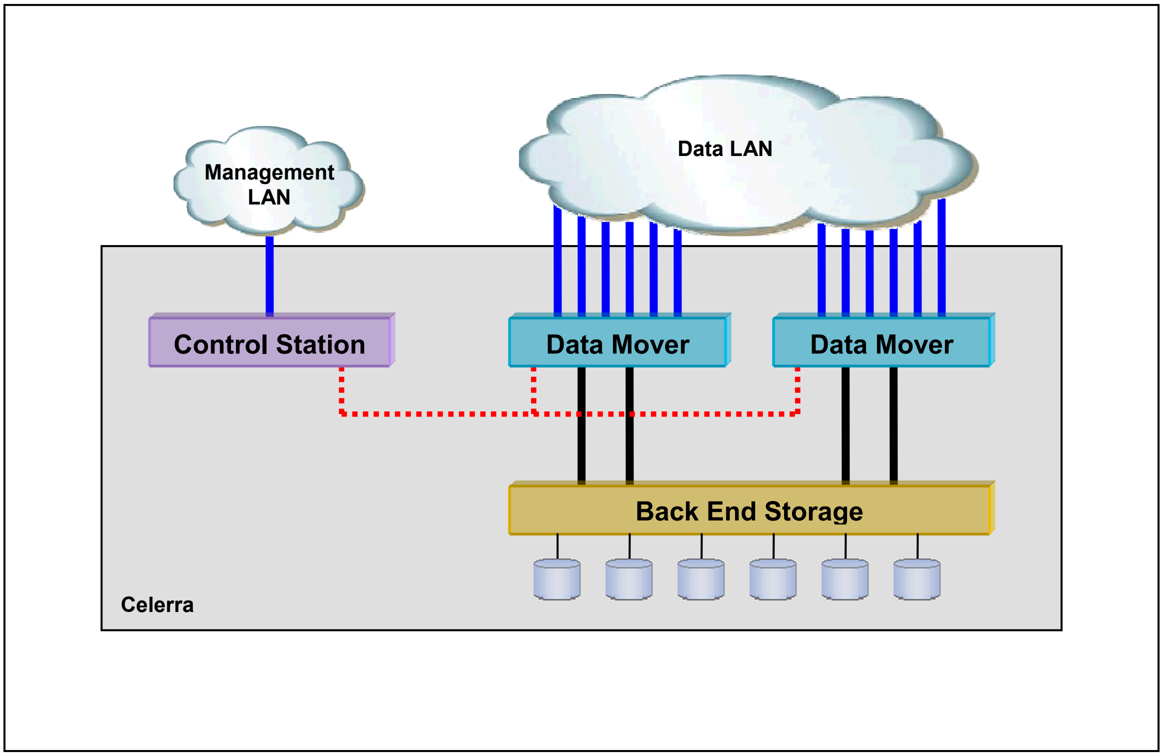


Figure 1. Celerra block diagram

Data Movers

A Celerra has one or more Data Movers installed in its frame. You can think of a Data Mover as an independent server running EMC's optimized NAS operating system, DART (data access in real time). Each Data Mover has its own network ports and network identity, and also has its own connections to back-end storage. So, in many ways, a Data Mover operates as an independent server, bridging the LAN and the back-end storage disk array.

Multiple Data Movers are grouped together as a single system for high availability and ease of use. For high availability, Celerra supports a configuration in which one Data Mover can act as a hot standby for one or more active Data Movers. In the case of a failure of an active Data Mover, the hot standby can be booted quickly to take over the identity and storage of the failed device. For ease of use, all Data Movers in a cabinet are logically grouped together for administration as a single system through the Control Station.

Control Station

The Control Station is the single point of management and control of a Celerra frame. Regardless of the number of Data Movers or disk drives in the system, all administration of the system is done through the Control Station. Control Stations not only provide the interface to configure Data Movers and back-end storage, but also provide heartbeat monitoring of the Data Movers. It is important to note that if a Control Station is inoperable for any reason, the Data Movers continue to operate normally. However, the Celerra architecture provides an option for a redundant Control Station to support continuous management for an increased level of availability.

The Control Station runs a version of the Linux OS that EMC has optimized for Celerra and NAS administration. The diagram in Figure 1 shows a Celerra system with two Data Movers. The Celerra NAS family supports up to 14 Data Movers depending on the product model.

Basics of iSCSI on Celerra

Celerra provides access to block and file data using iSCSI and NFS. These storage protocols are provides as standard TCP/IP network services as depicted in the topology diagram of Figure 2.

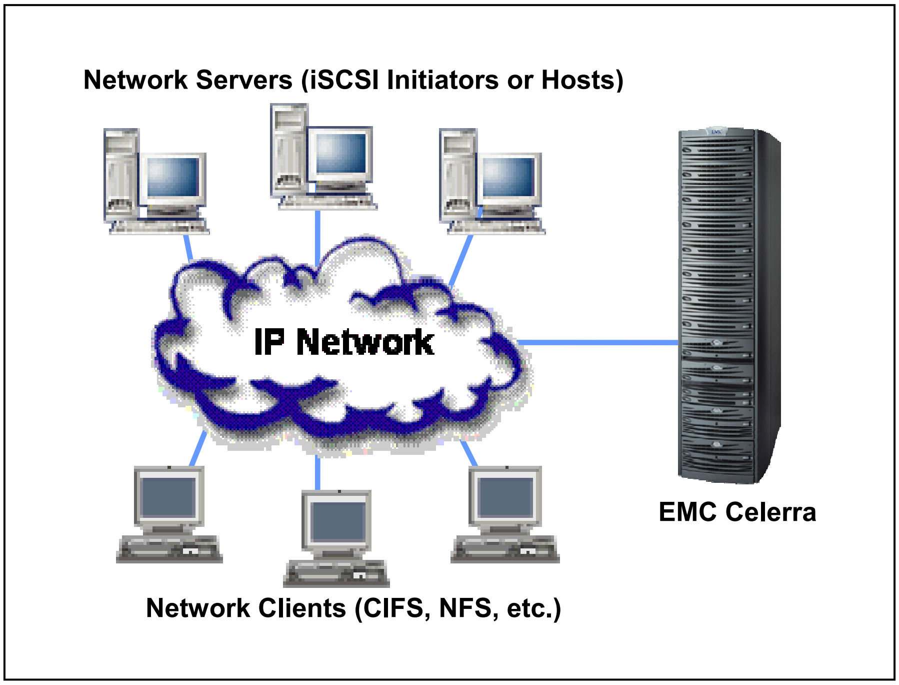

Network Servers (iSCSI Initiators or Hosts)
IP Network
EMC Celerra
Network Clients (CIFS, NFS, etc.)

Figure 2. Celerra iSCSI topology

Native iSCSI

Celerra delivers a native iSCSI solution. This means that when deployed, all the connected devices have TCP/IP interfaces capable of supporting the iSCSI protocol. A simplified view of this topology is shown in Figure 3. No special hardware, software, or additional configuration is required. The administrator needs only to configure the two endpoints – the iSCSI target on the Celerra and the iSCSI host initiator on the server. This is a much simpler and easier option to configure and manage, than that of a typical iSCSI bridging deployment. A bridging solution requires additional hardware that acts as the bridge between the Fibre Channel network connected to the storage array and the TCP/IP network connected to the host initiator. The addition of the bridge adds hardware cost and configuration complexity that is not necessary when deploying iSCSI on Celerra. With a bridging deployment, you are required to configure the iSCSI target on the storage array, the iSCSI host initiator on your server, plus both corresponding connections on the bridge. This added complexity is not required when using Celerra.

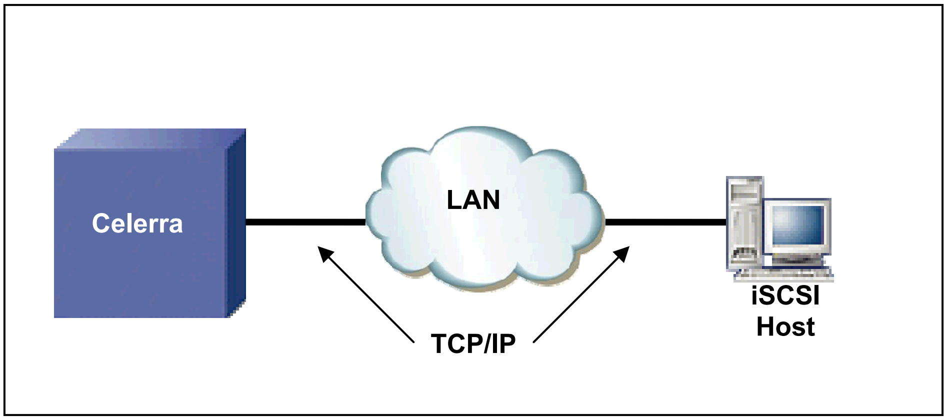


Figure 3. Native iSCSI

NFS

NFS or Network File System is a remote file sharing protocol offered by the Celerra network storage system. Celerra provides support for NFS servers through one or more Data Movers allowing NFS clients to safely access shared storage across the network. Celerra provides added functionality to the NFS service by offering support for snapshot copies as well as integrated replication and backup technologies. The core components providing the NFS service are the Celerra Data Mover or blade server.

VMware Infrastructure 3

Overview

VMware Infrastructure 3 consists of virtualization software that provides server consolidation by allowing several instances of similar and dissimilar operating systems to run as virtual machines on one physical machine. This cost-effective, highly scalable virtual machine platform offers advanced resource management capabilities. VMware Infrastructure 3 minimizes the total cost of ownership (TCO) of computing infrastructure by:

- Increasing resource utilization.
- Decreasing the number of servers and all associated costs.
- Maximizing server manageability.

Figure 4 shows several VMware ESX Servers with virtual machines containing guest operating systems that sit on top of the virtualization layer. It also illustrates the VMware VirtualCenter Server, which provides the management and monitoring of all VMware Infrastructure 3 components. At the upper layer of Figure 4 are the integrated VMware Infrastructure 3 features for Distributed Resource Scheduler, High Availability, and Consolidated Backups.

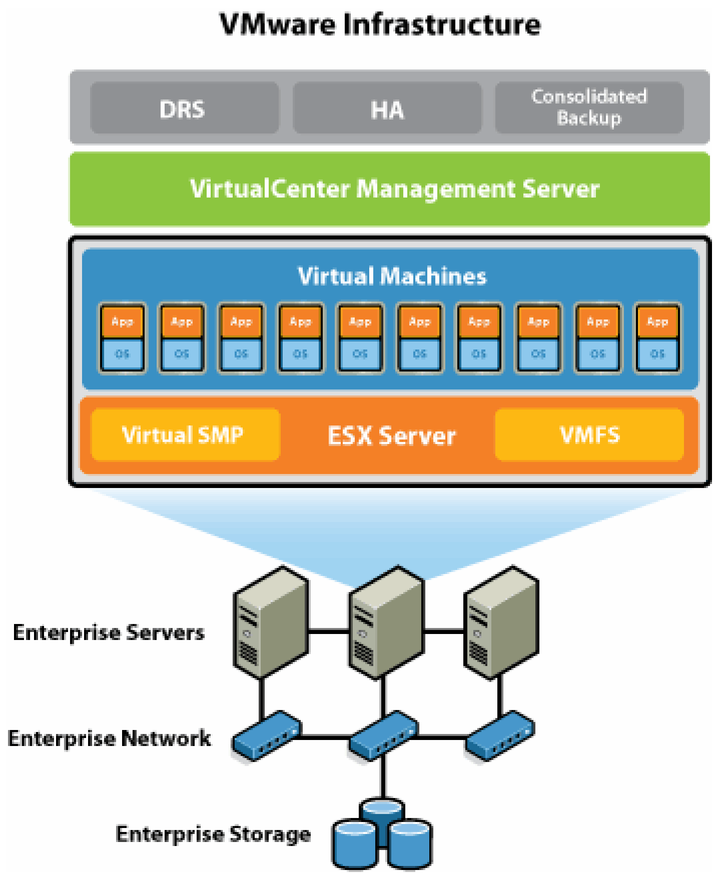


Figure 4. Architecture of VMware Infrastructure 3

ESX Server runs directly on the host hardware, allowing virtual machines to run on top of the virtualization layer provided by VMware. The combination of an operating system and applications is referred to as a virtual machine. ESX 3 servers are managed using the VirtualCenter Management Interface. VirtualCenter 2.0 allows you to manage a number of ESX Servers, as well as to perform operations such as VMotion.

Terminology

Cluster — A cluster within VirtualCenter 2.0 that is a collection of ESX Server hosts and associated virtual machines that share resources and a management interface.

Datastore — A file system, either VMFS or NFS, that serves as a virtual representation of an underlying pool of physical storage resources. These physical storage resources can be comprised of SCSI disks from a local server, Fibre Channel SAN disk arrays, iSCSI SAN disk arrays, or network-attached storage arrays.

Distributed Resource Scheduler (DRS) — Intelligently allocates and balances computing capacity dynamically across collections of hardware resources for virtual machines

ESX Server — A production-proven virtualization layer run on physical servers that abstract processor, memory, storage, and networking resources to be provisioned to multiple virtual machines

Guest operating system — An operating system that runs within a virtual machine.

ISO image — A CD image that can be downloaded and burnt on a CD-ROM or mounted as a loopback device.

Mapping file — A VMFS file containing metadata used to map and manage a raw device.

Raw device mapping (RDM) — Raw device mapping includes a combination of a pointer, which is a .vmdk file that resides on a VMFS volume, and a physical raw device that the .vmdk file points to. In physical compatibility mode it allows SCSI commands to be passed directly to the device.

Service Console (COS) — The modified Linux kernel that serves as the management interface to the ESX Server. It provides ssh, HTTP, and VirtualCenter access to the ESX host. Do not confuse with VMkernel.

Templates —A means to import virtual machines and store them as templates that can be deployed at a later time to create new virtual machines.

VirtualCenter Management Server — The central point for configuring, provisioning, and managing virtualized IT infrastructure

Virtual Infrastructure Client (VI Client) — An interface that allows administrators and users to connect remotely to the VirtualCenter Management Server or individual ESX Server installations from any Windows PC

Virtual machine — A virtualized x86 PC on which a guest operating system and an associated application run.

Virtual machine configuration file — A file containing a virtual machine configuration that is created by the Configuration Wizard or the Configuration Editor. The VMware ESX Server uses this file to identify and run a specific virtual machine. It usually has a .vmx extension.

Virtual Machine File System (VMFS) — A VMware proprietary file system installed onto data stores and used by ESX to house virtual machines.

VMkernel — A kernel that controls the server hardware and schedules virtual machine computations and I/O operations.

VMotion — The VMotion feature provides the ability to migrate a running virtual machine from one physical ESX Server to another.

VMware ESX Server features

VMware VMotion

VMotion technology provides the ability to migrate a running virtual machine from one physical ESX Server to another—without application service interruption—allowing for fast reconfiguration and optimization of resources without impacting users. With VMotion, VMware allows administrators to move virtual machine partitions from machine to machine on the fly, as real workloads run in the partitions. This allows administrators to do hardware maintenance without interrupting applications and users. It also allows the administrator to do dynamic load balancing to maintain high utilization and performance.

ESX Server version 2.0.1 was the first platform to support VMotion. System administrators can use VMotion, a systems management and provisioning product that works through VMware VirtualCenter, to quickly provision and reprovision servers with any number of virtual machines.

Distributed Resource Scheduler and VMware High Availability

The VMware Distributed Resource Scheduler (DRS) feature improves resource allocation across all hosts by collecting resource (such as CPU and memory) usage information for all hosts and virtual machines in the cluster and generating recommendations for virtual machine placement. These recommendations can be applied automatically or manually. Depending on the configured DRS automation level, DRS can display or automatically implement recommendations. The result is a self-managing, highly optimized, highly efficient computer cluster with built-in resource and load balancing.

VMware High Availability (HA) detects ESX Server hardware machine failures and automatically restarts virtual machines and their resident applications and services on alternate ESX Server hardware, enabling servers to recover more rapidly and deliver a higher level of availability. Using VMware HA and DRS together combines automatic failover with load balancing. This combination results in a fast rebalancing of virtual machines after HA has moved virtual machines to different hosts.

VMware clustering

The ESX Server can be clustered at a virtual machine level within a single ESX Server (referred to as an *in-the-box-cluster*) or among two or more ESX Servers (an *outside-the-box-cluster*). The cluster setup within a box is useful for providing high availability when software or administrative errors are the likely causes of failure. Users who want a higher level of protection in the event of hardware failures, as well as software/logical failures, benefit from clustering outside the box.

ESX Server 3 network concepts

The ESX Server 3 architecture requires multiple network interfaces in order to support the connectivity requirements of the management console, VMkernel, and virtual machines.

The primary network interface types are:

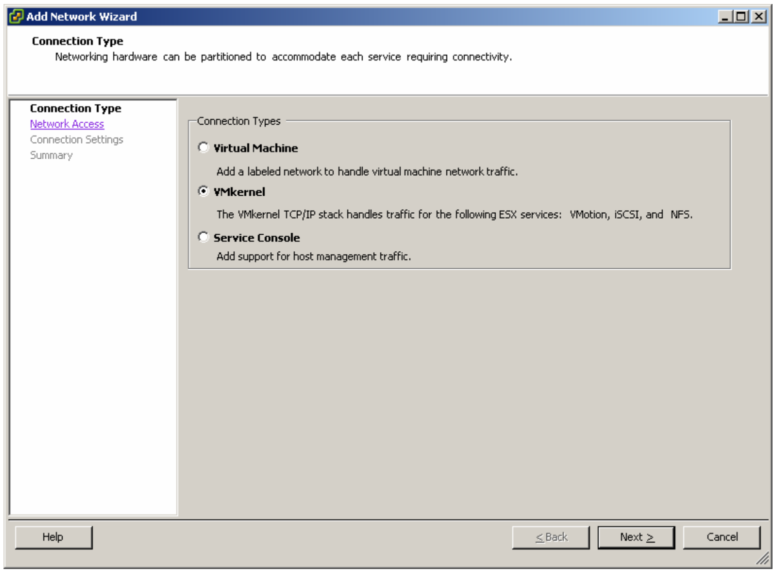
- **Virtual machine** - Connections will be made available to the guest operating systems.
- **Service Console** - Interface provides ssh, HTTP, and VirtualCenter access to the ESX host.
- **VMkernel** – This interface is used for all IP storage traffic including NFS mounts, iSCSI sessions, and VMotion.

Best practice - ESX Server 3 allows VMkernel to share a network connection with the Service Console. It can also be used to transport the state information of virtual machines when using VMotion. If the ESX host is configured for VMotion or DRS, it is best to create a separate interface for those services.

In order to use IP storage the ESX host requires the configuration of one or more physical network interfaces for the VMkernel. This interface requires a unique IP address and should be dedicated to serving the storage needs of the virtual machines.

At a functional level the VMkernel manages the IP storage interfaces, including those used for iSCSI and NFS access to Celerra. When configuring ESX Server for IP storage with Celerra, the VMkernel network interfaces are configured to access one or more Data Mover iSCSI targets or NFS servers.

The first step in the process is to configure the VMkernel interface through the network configuration wizard of the VI client, as shown in Figure 5.



Add Network Wizard

Connection Type

Networking hardware can be partitioned to accommodate each service requiring connectivity.

Connection Type
Network Access
Connection Settings
Summary

Connection Types

○ **Virtual Machine**

Add a labeled network to handle virtual machine network traffic.

● **VMkernel**

The VMkernel TCP/IP stack handles traffic for the following ESX services: VMotion, iSCSI, and NFS.

○ **Service Console**

Add support for host management traffic.

Figure 5. VMkernel Network Configuration Wizard

Best practice – Use one or more 1 Gb network interfaces for the VMkernel.

Since the VMkernel interface is in effect the SCSI bus for IP storage, it is a recommended practice that the network traffic from Celerra be segmented through a private LAN using either a virtual LAN or a dedicated IP SAN switch. Based upon the throughput requirements for the virtual machines, you may need to configure more interfaces for additional network paths to the Celerra Data Mover.

Storage Access Path High Availability options

The ESX hosts offer several advanced networking options to improve the service level of the VMkernel IP storage interface. NIC teaming in ESX Server 3 provides options for load balancing and redundancy.

The options for NIC teaming include:

- Mac Based –default in ESX Server 2.5
- Port Based –default in ESX Server 3
- IP Based –optional in either version

Consider the use of NIC teams with IP or port hashing to distribute the load across multiple network interfaces on the ESX Server host. When multiple Celerra iSCSI targets or NFS file systems are provisioned for ESX Server, NIC teaming can be combined with Celerra advanced network functionality to route the sessions across multiple Data Mover network interfaces for NFS and iSCSI sessions. The NIC teaming feature will allow for the use of multiple NICs on the VMkernel switch. Logical network connections configured for link aggregation on Celerra provide a session-based load-balancing solution that can improve throughput to the virtual machines. Figure 6 provides a topological diagram of how EtherChannel can be used with ESX and Celerra.

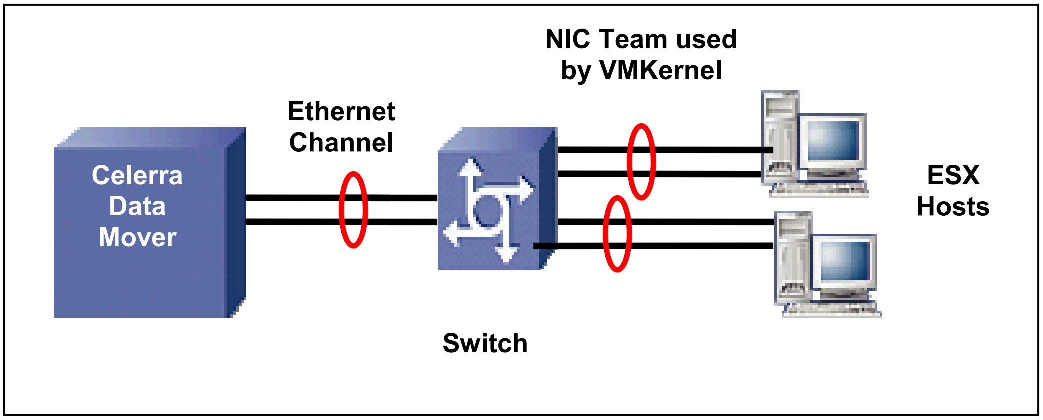


Figure 6. Celerra networking with EtherChannel

VLANs

VLANs allow for secure separation and QoS of the Ethernet packets carrying ESX Server I/O. Their use is a recommended configuration practice for ESX Server. VMware Infrastructure 3 supports various VLAN options with 802.1Q, including standard port-based VLAN tagging. The VLAN tag can be applied to the virtual switch network interface in order to limit the scope of network traffic on that Virtual Machine Network Interface.

With the use of virtual group tagging a port assignment of 4095 is applied. This requires the host to provide the 802.1Q tag to the network interface of the virtual machine. ESX Server provides support for 802.1D and 802.1Q using up to 4,094 ports. The Celerra network stack also provides support for 802.1Q VLAN tagging. A virtual network interface (VLAN) on Celerra is a logical association that can be assigned to either a physical or a logical network device such as an aggregated link or Fail-Safe Network interface. VLAN on Celerra occurs above the Layer 2 hardware address so that the Data Mover can support multiple physical interfaces for high availability and load balancing.

Since both Celerra and ESX Server 3 support 802.1Q, an environment in which the storage network has been segmented with VLANs is completely supported.

Best practice - Segment network traffic for VMkernel IP storage through a private LAN using either a VLAN or an independent IP SAN switch.

External Switch Tagging

External Switch Tagging (EST) is the default configuration for network ports on the ESX host. It is similar to physical network configuration with the VLAN tag being transparent to the individual physical server. This option is applied at the port and does not require any special configuration from the ESX Server. The tag is appended when a packet arrives at a switch port and stripped away when a packet leaves a switch port. Since there is a one-to-one relationship between the NIC and the physical switch, EST does have the impact of limiting the number of VLANs that can be used within the system.

The inclusion of a reference to jumbo frames is provided here as an informational messages. Since the Celerra Data Mover supports jumbo frames it would seem like a natural choice for the ESX Server IP VMkernel interfaces; however, there is currently no support for jumbo frames within the VMkernel network stack. Had that been available it would have been a recommended practice to enable jumbo frames within the IP storage network.

It may be possible to achieve larger frame sizes through the use of an iSCSI HBA device such as the QLogic 4050 series. Both Celerra and the QLA 4052 support Jumbo Ethernet Frame sizes up to 9000 bytes.

Celerra storage provisioning for ESX Server 3

With the support for IP storage in ESX Server 3, EMC Celerra has become a preferred platform for supporting the datastores used for virtual machines. Celerra NFS and iSCSI support provide two options for configuring network storage for the ESX host systems. Celerra with CLARiiON® or Symmetrix® back-end arrays provide the ESX hosts with highly available, RAID-protected storage.

The use of Celerra as a storage system does not preclude the use of Fibre Channel devices. ESX Server supports the use of Fibre Channel devices along with iSCSI and NFS. When configured with VMotion ESX enables placement of the virtual machines on different storage volumes based upon the needs of the application. This further offers a method to relocate or migrate a virtual machine among the tiers of storage.

Storage is configured for the ESX hosts through the Celerra Management interface. Celerra provides striped (RAID 1) and parity (RAID 3/RAID 5) options for performance and protection of the devices used to create ESX volumes. Celerra Automatic Volume Manager (AVM) runs an internal algorithm that identifies the optimal location of the disks that make up the file system. Storage administrators are only required to select the storage pool type and desired capacity in order to establish a file system that can be presented to ESX for use as an NFS datastore.

The choice of which storage and RAID algorithm you choose is largely based upon the throughput requirements of your applications or virtual machines. RAID 5 provides the most efficient use of disk space with good performance to satisfy the requirements of your applications. RAID 1 provides the best performance at the cost of additional disk capacity to mirror all of the data in the file system. In testing performed within EMC labs, RAID 5 was chosen for both virtual machine boot disk images as well as virtual disk storage used for application data. Understanding the application and storage requirements within the computing environment will help to identify the appropriate RAID configuration for your servers. EMC NAS specialists are trained to translate those requirements into the correct storage configuration.

ESX provides the tools to create a datastore from a Celerra NFS file system export or an iSCSI LUN. A user-assigned label is required to identify the datastore.

The virtual disks are assigned to a virtual machine and are managed by the guest operating system just like a standard SCSI device. For example, a virtual disk would be assigned to a virtual machine running on an ESX host. In order to make the device useful, a guest operating system would be installed on one of the disks. The format of the virtual disk is determined by the guest OS or install program. One potential configuration would be to present an NFS file system from Celerra to an ESX host. The ESX would use the NFS file system to create an NFS datastore and VMDK file for a newly defined virtual machine. In the case of a Windows guest, the VMDK would be formatted as an NTFS file system. Additional virtual disks used for applications could be provisioned from one or more Celerra file systems and formatted as NTFS by the Windows guest OS.

Naming of storage objects

An area deserving careful consideration when establishing an ESX environment with Celerra is in the naming of the storage objects. Providing descriptive names for the file systems, exports, iSCSI targets, and the datastores in ESX can establish valuable information for ongoing administration and troubleshooting of the environment. Prudent use of labels including the storage system from which they are configured will be helpful in management and troubleshooting when maintenance is required.

Best practice – Incorporating identifying elements (for example, IP addresses or NFS server names) into your datastore definition as well as annotating with the name of the Celerra being used will ease future management and configuration tasks.

Celerra NFS configuration for ESX Server 3

As mentioned in the “Introduction” section, ESX 3 provides an option to create a datastore from an NFS file system. Celerra NFS exported file systems and feature set complement those provided by ESX, providing an extremely flexible and distinctly reliable environment.

VMware Infrastructure 3 management utilities are used to configure and mount NFS file systems from Celerra. The VI client is also used to assign a datastore name to the export. The datastore name is the key reference that is used to manage the datastore within the ESX environment.

The NFS datastore is viewed as a pool of space used to support virtual disks. One or more virtual disks are created within the datastore and assigned to virtual machines. Each virtual machine will use the primary virtual disk to install the guest operating system and boot information. The additional disks are used to support application and user data.

NFS datastores offer support for the use of ESX virtual disks, virtual machine configuration files, snapshots, disk extension, VMotion, and Disaster Recovery Services.

Additionally, Celerra provides support for replication, local snapshots, virtual provisioning, NDMP backups, and virtual provisioning of the file systems used for ESX.

In order to access the NFS exported file system the ESX host must have a VMkernel defined with network access to the Celerra. By default the ESX host will mount the file system using the root user account. The file system must include root access for the interface that will mount the file system. A good practice would be to limit the export to only the VMkernel interfaces.

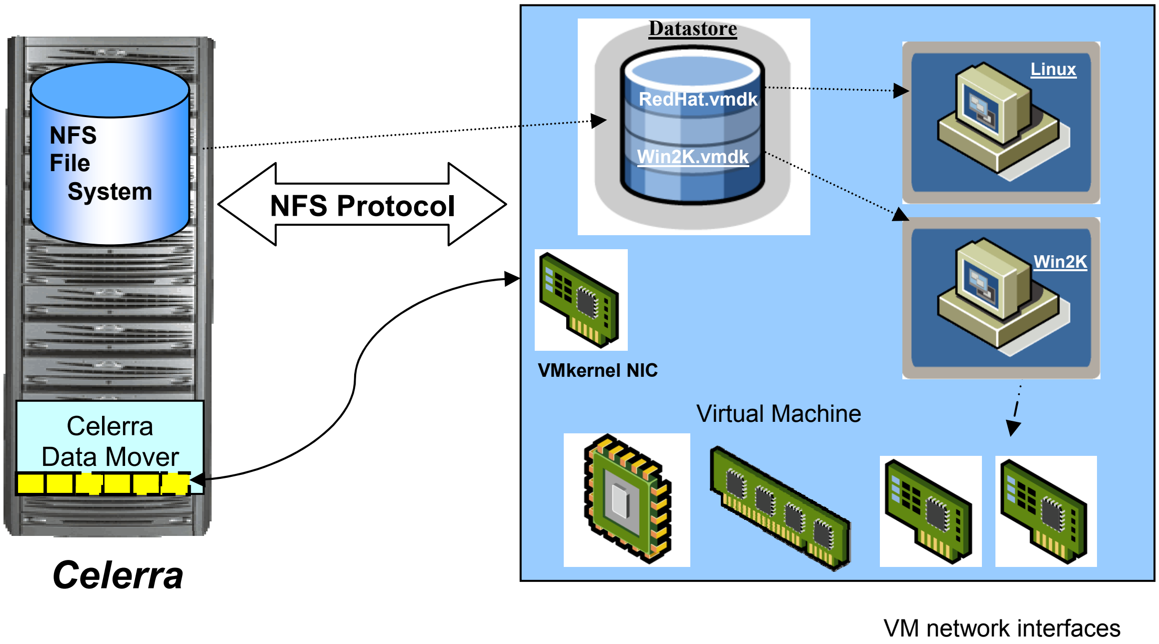


Figure 7. ESX NFS architecture in VMware Infrastructure 3

Figure 7 provides a rough diagram of the ESX NFS architecture in VMware Infrastructure 3 with a focus on the interaction of the storage devices and the ESX host. ESX creates a different folder for each VM in the environment. All virtual objects including virtual disks and ESX snap logs are stored in the file system. In Figure 7 there are several network interfaces. The VMkernel interface is the one that will be of considerable importance with network storage since all of the I/O for the ESX virtual disks is carried across the VMkernel NIC.

The virtual machine case can use one of the virtual machine network interfaces to access other NFS file system exports, CIFS shares, or iSCSI targets from Celerra.

Added Celerra NFS benefits with ESX Server

The use of NFS is also a valuable option with ESX for storage of common files such as virtual machine templates and clone images.

Potential customer use cases include the following:

- Sharing network library elements such as ISO images
- Storing virtual machine templates and clones

Users may choose to store commonly used virtual machine images as templates to be used again in the future, reducing the time required to set up a new virtual machine. NFS facilitates this time-saving method by enabling all ESX hosts within an environment access to the same file system. A user placing their virtual machine templates into a file system shared in this way could eliminate the need to copy the templates into numerous disparate file systems throughout the network.

Best practice – Create a NFS datastore to hold template and ISO images. Mount the NFS datastore on each ESX Server to provide a central location for these files.

Business continuity (SnapSure)

The use of Celerra checkpoints and Celerra Replicator™ assist in the ongoing management of the file system. Depending on the state of the guest OS and applications, snapshots provide a crash-consistent image of the file systems that contain operating system and application virtual disks.

Celerra snapshots can be used to create point-in-time images of the file systems containing VMDK files for OS and application data. The snapshot image can be used for near-line restores of virtual disk objects including ESX snapshots and individual virtual machines. The snapshot file system provides a secondary image of all ESX file objects and can be integrated into NDMP backup processes or copied to a second Celerra for disaster recovery purposes.

NFS security

Celerra offers several options for limiting access to data, including:

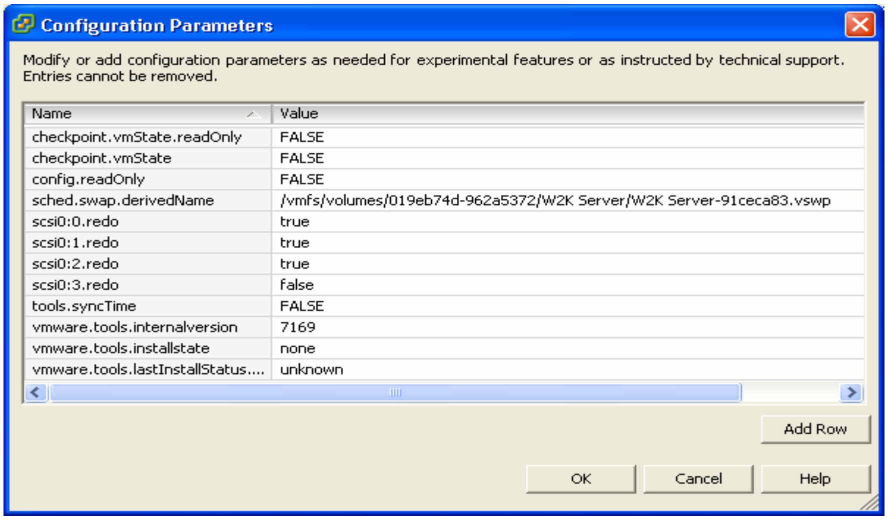
- Separation of traffic through the use of VLANs
- Limiting access to certain host IP addresses
- Limiting access by certain user accounts
- ESX provides the use of a non-root account called a delegate to mount the NFS file system. Consider using this access method to avoid exporting the file system with root access.

Since ESX Server will be mounting file systems that contain virtual operating system and application disks, it is good practice to limit access to the storage.

Additional considerations

By default the NFS client within ESX will be limited to eight NFS mounts. This should be sufficient in most cases but there is an option to extend the number of file systems to a maximum of 32 through the advanced settings tab of the ESX VirtualCenter interface.

When a virtual machine is created, ESX creates a swap file that is stored in the virtual machine folder with the virtual disk. The swap file is used to swap resources when the ESX host becomes busy. In order to limit the impact of swapping to a network device VMware recommends storing the file on local disk. This is accomplished by modifying the configuration parameters in the advanced settings of the VM as illustrated in Figure 8. Adding a line for sched.swap.dir that includes the location on local disk will accomplish the task. Documentation on ESX resource management provides additional information on setting the value.



Configuration Parameters

Modify or add configuration parameters as needed for experimental features or as instructed by technical support. Entries cannot be removed.

Name	Value
checkpoint.vmState.readOnly	FALSE
checkpoint.vmState	FALSE
config.readOnly	FALSE
sched.swap.derivedName	/vmfs/volumes/019eb74d-962a5372/W2K Server/W2K Server-91ceca83.vswp
scsi0:0.redo	true
scsi0:1.redo	true
scsi0:2.redo	true
scsi0:3.redo	false
tools.syncTime	FALSE
vmware.tools.internalversion	7169
vmware.tools.installstate	none
vmware.tools.lastInstallStatus....	unknown

Figure 8. ESX NFS Configuration Parameters

NOTE – While this configuration option will provide performance improvements, it negates the ability to use the NFS datastore for VMotion.

Celerra iSCSI configuration for ESX Server 3

VMware Infrastructure 3 introduced native iSCSI client support through a VMkernel TCP/IP stack and software initiator. With this IP storage option, the ESX iSCSI software initiator can be configured to access up to 255 iSCSI LUNs from one or more Celerra iSCSI targets.

iSCSI LUN configuration

The configuration of iSCSI first requires that the ESX host have a network connection configured for IP storage and also have the iSCSI service enabled. This ESX VMkernel storage interface configuration used for iSCSI was covered in the ESX network concepts section of the paper.

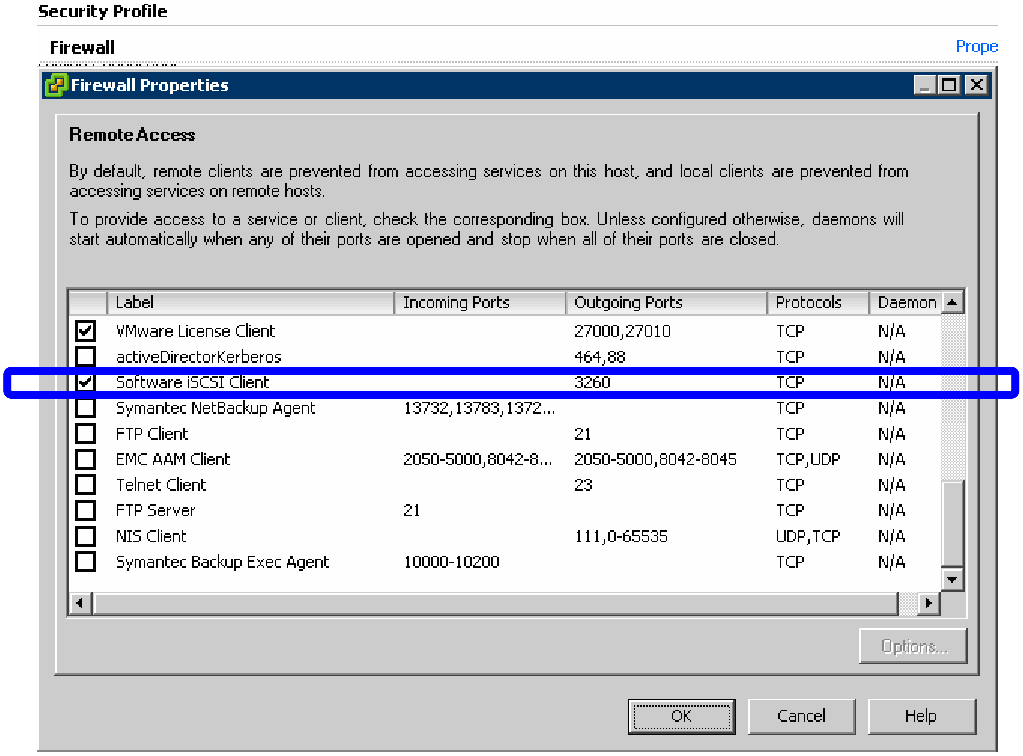
To establish a session, the following steps must be completed (Figure 9):

1. Enable the iSCSI client in the security profile, firewall properties interface of the VI client.

2. Configure the iSCSI software initiator on the ESX Server host.

3. Configure iSCSI LUNs and mask them to the IQN of the software initiator defined for this ESX Sever host. If using a hardware device you will need to identify the IQN of that device for masking.

Prior to configuration of the ESX host for iSCSI, you must ensure that the firewall has been modified to enable the iSCSI client on the host. This allows the client to establish sessions with the iSCSI target on Celerra.



Security Profile

Firewall

Firewall Properties

Remote Access

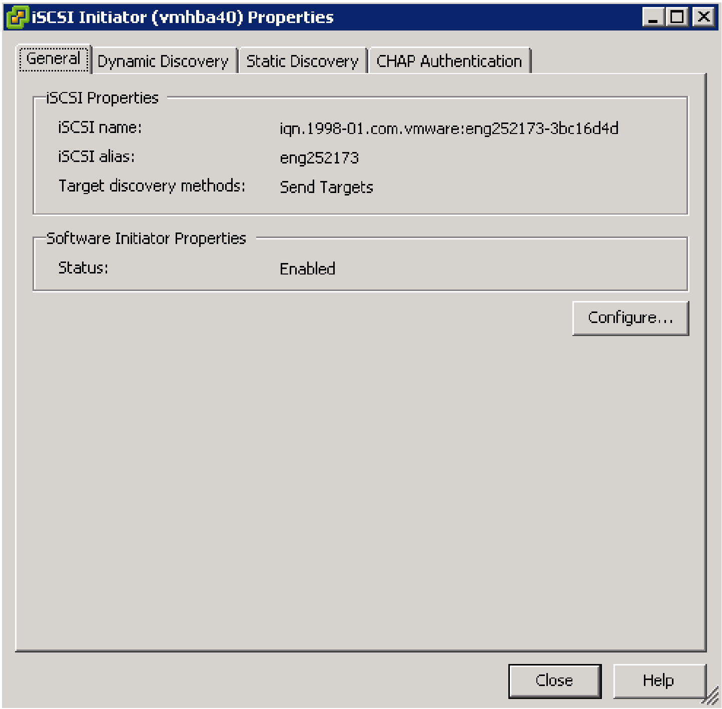
By default, remote clients are prevented from accessing services on this host, and local clients are prevented from accessing services on remote hosts.

To provide access to a service or client, check the corresponding box. Unless configured otherwise, daemons will start automatically when any of their ports are opened and stop when all of their ports are closed.

	Label	Incoming Ports	Outgoing Ports	Protocols	Daemon
☑	VMware License Client		27000,27010	TCP	N/A
□	activeDirectorKerberos		464,88	TCP	N/A
☑	Software iSCSI Client		3260	TCP	N/A
□	Symantec NetBackup Agent	13732,13783,1372...		TCP	N/A
□	FTP Client		21	TCP	N/A
□	EMC AAM Client	2050-5000,8042-8...	2050-5000,8042-8045	TCP,UDP	N/A
□	Telnet Client		23	TCP	N/A
□	FTP Server	21		TCP	N/A
□	NIS Client		111,0-65535	UDP,TCP	N/A
□	Symantec Backup Exec Agent	10000-10200		TCP	N/A

Figure 9. ESX Firewall Properties page

The IQN can be identified in the iSCSI Initiator Properties page (Figure 10) of the iSCSI HBA. The default device name for the software initiator is vmhba40. You could also obtain the IQN name by issuing the vmkiscsi-iname command from the Service Console. The management interface is used to enable the iSCSI service as well as to define the network portal being used to access the Celerra iSCSI target.



iSCSI Properties

iSCSI name: iqn.1998-01.com.vmware:eng252173-3bc16d4d

iSCSI alias: eng252173

Target discovery methods: Send Targets

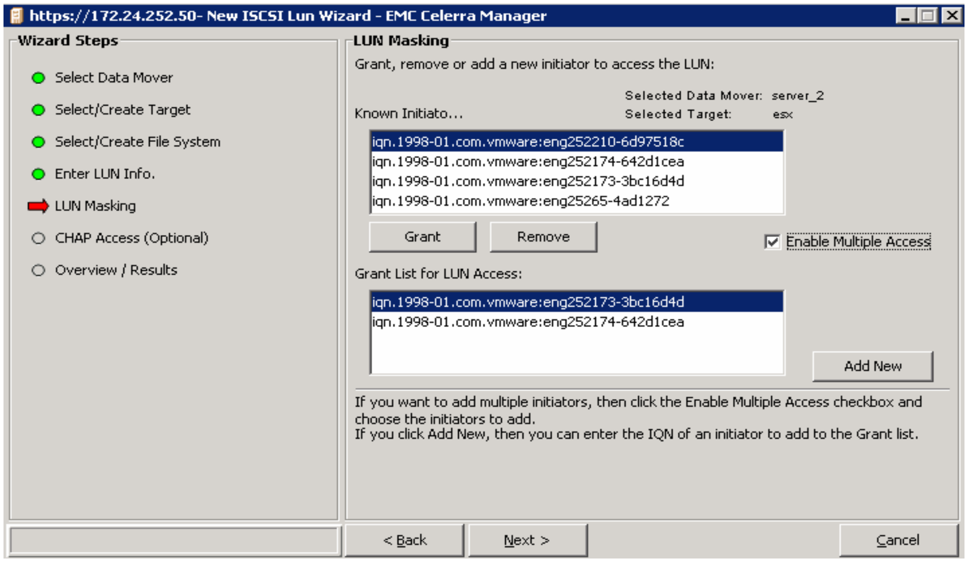
Software Initiator Properties

Status: Enabled

Figure 10. iSCSI Initiator Properties page

Once the client has been configured you would need to ensure that the Celerra has been configured with at least one iSCSI target and LUN.

The Celerra Management iSCSI Wizard can be used to configure an iSCSI LUN. Knowing the IQN of the ESX Server software or hardware initiator will allow you to mask the LUN to the host for further configuration. LUNs are provisioned through Celerra Manager from the Celerra file system and masked to the iSCSI Qualified Name (IQN) of the ESX Server host iSCSI software initiator. As with NFS the VMkernel network interface is used to establish the iSCSI session with the Celerra iSCSI target.



LUN Masking

Grant, remove or add a new initiator to access the LUN:

Selected Data Mover: server_2

Selected Target: esx

Known Initiato...

iqn.1998-01.com.vmware:eng252210-6d97518c
iqn.1998-01.com.vmware:eng252174-642d1cea
iqn.1998-01.com.vmware:eng252173-3bc16d4d
iqn.1998-01.com.vmware:eng25265-4ad1272

☑ Enable Multiple Access

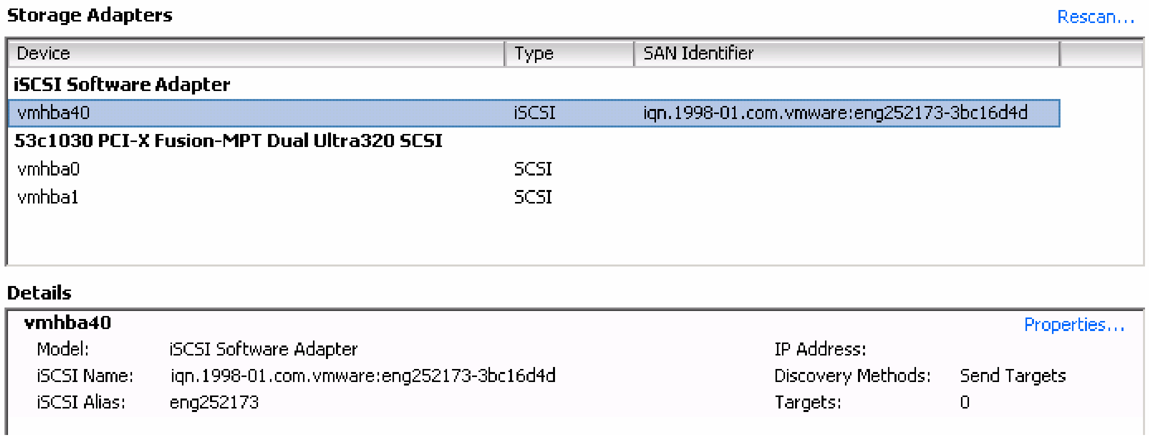
Grant List for LUN Access:

iqn.1998-01.com.vmware:eng252173-3bc16d4d
iqn.1998-01.com.vmware:eng252174-642d1cea

If you want to add multiple initiators, then click the Enable Multiple Access checkbox and choose the initiators to add.
If you click Add New, then you can enter the IQN of an initiator to add to the Grant list.

Figure 11. Celerra Manager New iSCSI Lun Wizard – LUN mask assignment

After the configuration steps have been completed return to the Storage Adapters interface (Figure 12) and scan the iSCSI bus to identify the LUNs that have been configured for this ESX Server host.



Device	Type	SAN Identifier
iSCSI Software Adapter		
vmhba40	iSCSI	iqn.1998-01.com.vmware:eng252173-3bc16d4d
53c1030 PCI-X Fusion-MPT Dual Ultra320 SCSI		
vmhba0	SCSI	
vmhba1	SCSI	

vmhba40

Model: iSCSI Software Adapter
iSCSI Name: iqn.1998-01.com.vmware:eng252173-3bc16d4d
iSCSI Alias: eng252173
IP Address:
Discovery Methods: Send Targets
Targets: 0

Figure 12. Storage Adapters interface

LUN configuration options

There are three methods of leveraging iSCSI LUNs in an ESX environment: Virtual Machine File System (VMFS), raw device mapping (RDM), and using a software initiator (generally the Microsoft software initiator) within the virtual guest OS. Each of these methods is best used in specific use case examples discussed below.

Virtual Machine File System

The Virtual Machine File System (VMFS) is the default disk configuration option for ESX Server. The VI client formats an iSCSI LUN as VMFS3 which is used to create a datastore for virtual disks, virtual machine configuration files, snapshot log files, and file system metadata. It possesses very similar functionality to the NFS datastore discussed in the previous section.

Ideal customer use cases for VMFS:

- Storage for general virtual machines without IO-bound workloads like Application Servers, and Active Directory domain controllers
- Virtual machines that don’t require application-level IO consistency using Native ESX snapshots
- virtual machines that are ideally backed up using VMware Consolidated Backup

VMFS is an excellent choice along with NFS where various virtual machine LUNs don’t require specific I/O performance envelopes, or where ease of provisioning and use are paramount.

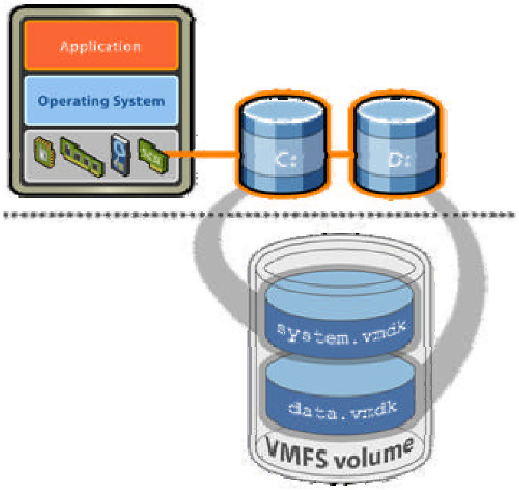
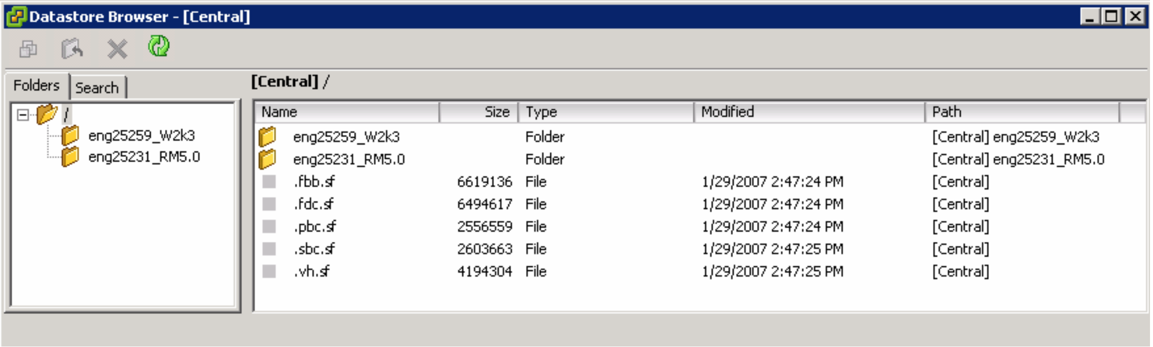


Figure 13. High-level representation of VMFS use with Celerra iSCSI and ESX Server 3

When Celerra iSCSI LUNs are used with an ESX Server host, advanced Celerra features such as Virtual Provisioned LUNs and Dynamic LUN extension are fully supported. One key difference between NFS and VMFS is that the VMFS file system contains ESX metadata that can be used for support purposes. Metadata files are identified by the .sf extension illustrated in Figure 14.



Name	Size	Type	Modified	Path
eng25259_W2k3		Folder		[Central] eng25259_W2k3
eng25231_RM5.0		Folder		[Central] eng25231_RM5.0
.fbb.sf	6619136	File	1/29/2007 2:47:24 PM	[Central]
.fdc.sf	6494617	File	1/29/2007 2:47:24 PM	[Central]
.pbc.sf	2556559	File	1/29/2007 2:47:24 PM	[Central]
.sbc.sf	2603663	File	1/29/2007 2:47:25 PM	[Central]
.vh.sf	4194304	File	1/29/2007 2:47:25 PM	[Central]

Figure 14. View of the VMFS datastore containing two virtual machines

Raw device mapping disk

Raw device mapping (RDM) is a mapping file in a VMFS volume that acts as a proxy for a raw physical device. The RDM contains metadata used to manage and redirect disk accesses to the physical device. The file gives you advantages of direct access to physical devices while keeping some advantages of a virtual disk in VMFS. As a result, it merges VMFS manageability with a raw device access.

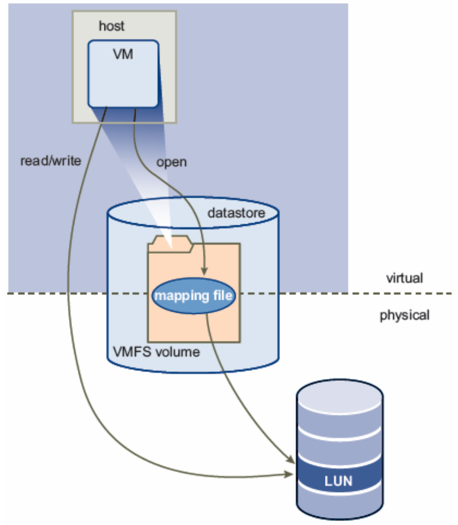


Figure 15. RDM device uses a separate mapping file

The RDM is a device that allows for SCSI commands to be passed from the guest OS to the SCSI target. The benefit is that management utilities that interface to storage platform-based replication tools can be used to create snaps and replicate iSCSI devices at the array level.

The limitation is that, unlike VMFS which can be used to create multiple virtual disks within the file system, RDM can be configured only as a single virtual disk.

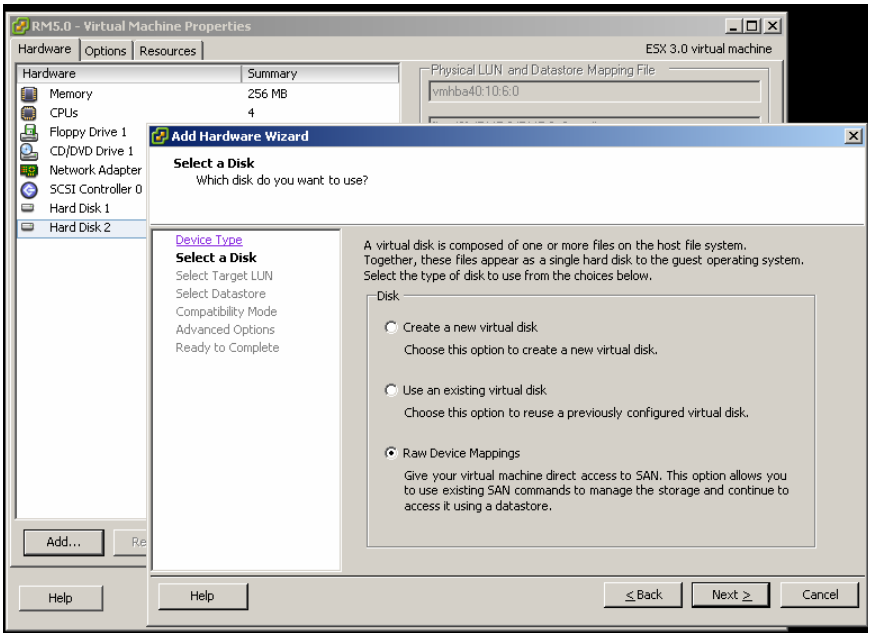
RDM disks are ideal when a LUN has specific performance needs. That is, since the RDM has a one-to-one virtual disk to iSCSI LUN model, it isn't shared by multiple virtual machines like VMFS and NFS. In addition application-integrated (examples include Microsoft Exchange 2003 and 2007 VSS and SQL Server VDI APIs are used) replication is required.

Ideal customer use cases for RDM are as follows:

- Exchange and SQL Server log and database LUNs (boot and binaries can be stored using VMFS or NFS)
- Cases where application-integrated replication is required

RDMs allow more control of configuration and layout – however, control and performance tuning mean that RDM implies marginally higher complexity than VMFS, and virtual machines using RDMs cannot be backed up using the VMware Consolidated Backup method. All other VMware Infrastructure 3.0 advanced features (VMotion, HA, DRS) are supported with RDMs and work as well with generic VMFS.

The use of the device provides the guest OS with the ability to pass SCSI commands directly to the storage device presented by the ESX Server kernel. In the case of Celerra, this option is available only if the device is an iSCSI LUN, not if it is an NFS datastore. One of the motivations for using this device type is when Celerra-level iSCSI replication is required. In order to create this device using the VI client, you must select the virtual machine that you are going to configure the LUN for and select the edit settings tab. After selecting the hardware tab you will be able to add a new device by selecting the disk option from the management interface. The key point in this process is to ensure that the new device is created in *physical compatibility mode.* Figure 16 illustrates the configuration screen for an RDM device.



Hardware	Summary
Memory	256 MB
CPUs	4
Floppy Drive 1	
CD/DVD Drive 1	
Network Adapter	
SCSI Controller 0	
Hard Disk 1	
Hard Disk 2	

Physical LUN and Datastore Mapping File: vmhba40:10:6:0

Select a Disk

Which disk do you want to use?

A virtual disk is composed of one or more files on the host file system. Together, these files appear as a single hard disk to the guest operating system. Select the type of disk to use from the choices below.

Disk

○ Create a new virtual disk

Choose this option to create a new virtual disk.

○ Use an existing virtual disk

Choose this option to reuse a previously configured virtual disk.

● Raw Device Mappings

Give your virtual machine direct access to SAN. This option allows you to use existing SAN commands to manage the storage and continue to access it using a datastore.

Figure 16. RDM creation with VirtualCenter Management interface

The primary difference between RDM and VMFS is the number of devices that are supported by each. While VMFS provides a pool of storage that can be used to create many virtual disks and virtual machines, the RDM device is presented directly to the VM as a single disk device. The RDM device can not be subdivided.

Software initiator

The iSCSI software initiator provides a suitable interface when running over Gigabit Ethernet. The use of NIC teaming and load balancing with Celerra provides sufficient throughput for the virtual machines running within ESX.

The iSCSI software initiator is not currently supported for booting the ESX operating environment over iSCSI. If you are interested in booting ESX Server from the storage system you will need to install an iSCSI HBA device. The QLogic 4050 series HBA is the only device qualified to boot ESX from Celerra iSCSI storage at this time. When configuring ESX Server for iSCSI boot, you will need to create and mask the iSCSI LUNs to the IQN name of the QLogic adapter. The NetBIOS settings will also need to be configured in order to establish the iSCSI LUN on Celerra as the boot device when installing ESX Server software.

Best practice – Booting ESX from Celerra requires an iSCSI HBA device. The QLogic 4050 series are the only HBAs qualified to boot from iSCSI storage at this time.

The process of adding storage requires that the ESX Server host establish a session with the storage device. Load balancing or rebalancing can be achieved with the cold migration tools within the Virtual Interface.

Maximum configurations

The following list shows supported features and maximum configurations allowed by ESX Server 3.0 at this time:

- A maximum of 254 LUNs
- A maximum of 128 VMware VMFS 3 volumes
- A maximum size of 64 TB per VMware VMFS 3 volume
- A maximum of eight targets for iSCSI
- Clustering not supported for iSCSI

Microsoft iSCSI Software Initiator within a virtual machine

Using the Microsoft Software Initiator within a guest OS is the same as using it on a physical server. The Microsoft Software Initiator is layered upon the network stack (in this case the guest OS virtual network), handling all iSCSI tasks in software.

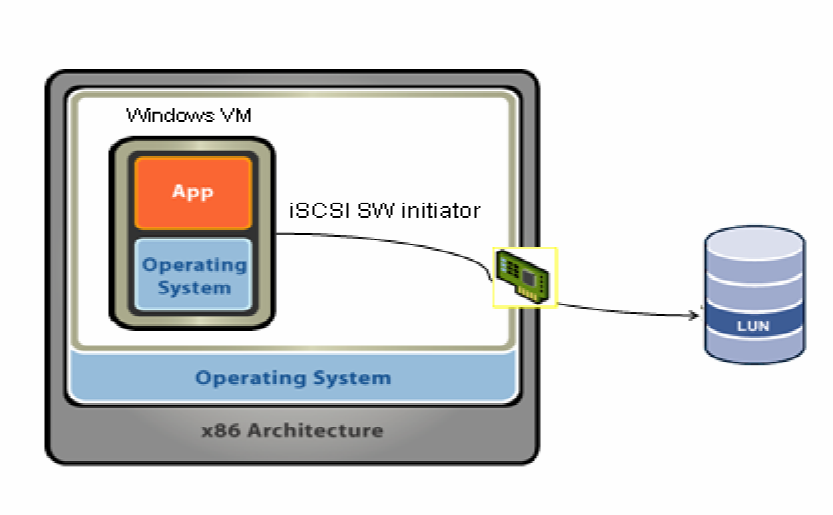


Figure 17. Microsoft iSCSI Software Initiator access to a Celerra iSCSI target LUN

This use case is similar to the RDM case in that the iSCSI LUN has a one-to-one mapping with the LUN configuration on the physical storage and layout, and the performance envelope can be controlled specifically. However, the primary rationale for this configuration is that this is the only storage model of all three that can support automated handling of the LUN addition and removal within the guest OS. All other methods (VMFS, NFS, RDMS) require manual configuration in VirtualCenter to configure storage. This capability is critical for mount hosts (for application-consistent replicas) where LUNs will be automatically and dynamically added and removed.

Ideal customer use cases for using the Microsoft iSCSI Software Initiator within the virtual machine:

- Mount host for LUN replicas taken using VSS and VDI application interfaces (for streamed backup)
- Dynamic presentation of SQL Server databases for test and development use

Figure 18 shows how these approaches can be used together:

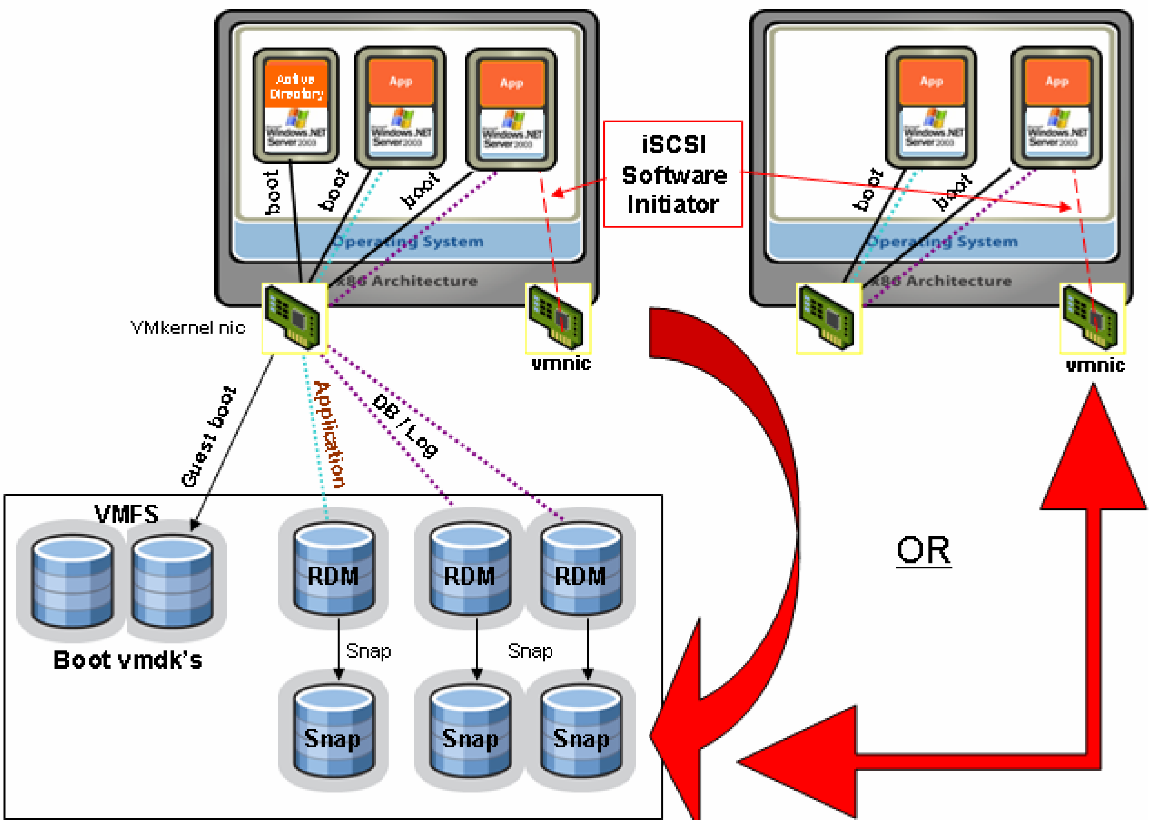
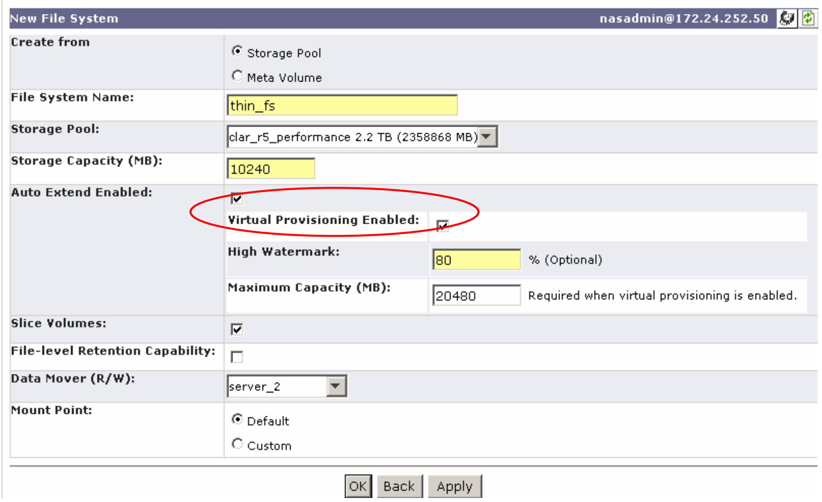


Figure 18. Virtual Machine interface and software initiator for iSCSI snapshot access

Virtual provisioning

With virtual, or thin, provisioning provided through Celerra file systems and iSCSI LUNs, storage is consumed in a more pragmatic manner. Virtual provisioning allows for the creation of storage devices that do not preallocate backing storage capacity for virtual disk space until the virtual machine application generates some data to the virtual disk. The virtual provisioning model avoids the need to overprovision disks based upon expected growth. Storage devices still represent and support the upper size limits to the host that is accessing them, but in most cases the actual disk usage falls well below the apparent allocated size. The benefit is that like virtual resources in the ESX Server architecture, storage is presented as a set of virtual devices that share from a pool of disk resources. Disk consumption increases based upon the needs of the virtual machines in the ESX environment. As a way to address future growth, Celerra monitors the available space and can be configured to automatically extend the backing file system size as the amount of free space decreases.

The Celerra Manager provides the interface from which to define the virtual file system for the NFS server. Selecting **Virtual Provisioning Enabled** when defining a new file system will create a file system that consumes only the initial requested capacity (10 GB in Figure 19). The ESX host will recognize the maximum size of 20 GB and transparently allocate space as the capacity usage grows.



New File System — nasadmin@172.24.252.50

Create from	● Storage Pool ○ Meta Volume
File System Name:	thin_fs
Storage Pool:	clar_r5_performance 2.2 TB (2358868 MB)
Storage Capacity (MB):	10240
Auto Extend Enabled:	☑
Virtual Provisioning Enabled:	☑
High Watermark:	80 % (Optional)
Maximum Capacity (MB):	20480 Required when virtual provisioning is enabled.
Slice Volumes:	☑
File-level Retention Capability:	□
Data Mover (R/W):	server_2
Mount Point:	● Default ○ Custom

Figure 19. Celerra virtual provisioned file system creation interface

ESX Server also provides an option to create thin provisioned virtual disks and file systems. Virtual disks created from an NFS datastore are thinly provisioned by default. VMFS datastores (iSCSI) default to thick or fully allocated virtual disks.

This means that the NFS file space is being consumed only when an application or user operating within the guest OS requests to write to a virtual disk. The disk request is translated into an allocation request within the file NFS system. The benefit of this option is that it preserves the amount of storage space used by the virtual disk. It does not, however, address the issue of overallocation of the file system that is used as the NFS datastore. The Celerra virtually provisioned file system improves on this concept by allocating only the disk space it needs to satisfy the needs of all of the virtual disks in the datastore.

If we view a file system as a one-dimensional expression as illustrated in the top figure in Figure 20, then we see that the nonvirtual or allocated file system has a fixed length, in this case 40 GB. That space is reserved for the ESX Server NFS datastore and will be available to all of the virtual storage devices. In this case we have several disks varying in size from 5 GB to 15 GB. If they are thin provisioned, they may not be allocating the full complement of their defined space, so we have unused space. This space may eventually be used for other virtual disk or snaps but for the moment it is not used.

In the bottom figure of Figure 20, a Celerra virtual file system is being used. In this case not only is ESX Server preserving disk resources, but the file system is as well. It has a fixed capacity but is only using the amount of space that is being requested by the virtual machines in ESX Server. The remaining disk storage is free for other file systems. So the storage which had been allocated in the example above is free to be used shared amongst thin provisioned file systems on Celerra.

Celerra virtual provisioning further extends the thin provisioning benefits by allocating only a portion of the disk space and automatically expanding the backing space as the usage requirements increase. This means that Celerra file systems used for ESX Server will consume disk space in a much more efficient manner than with “dense” provisioning.

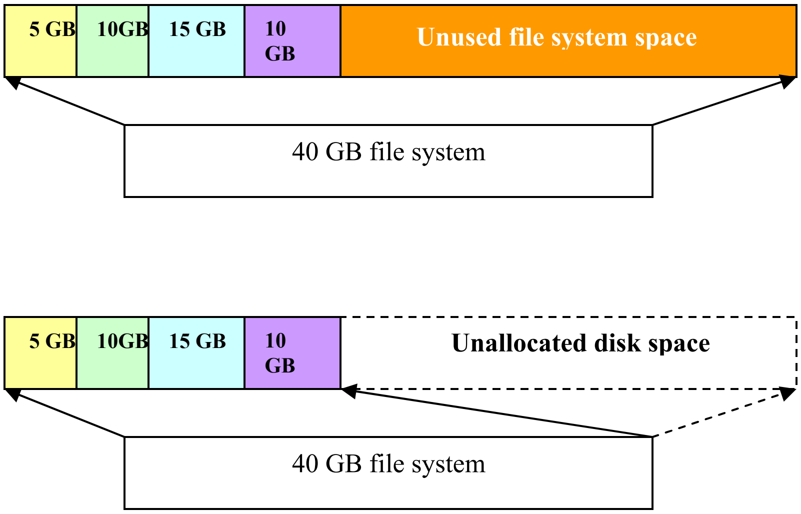
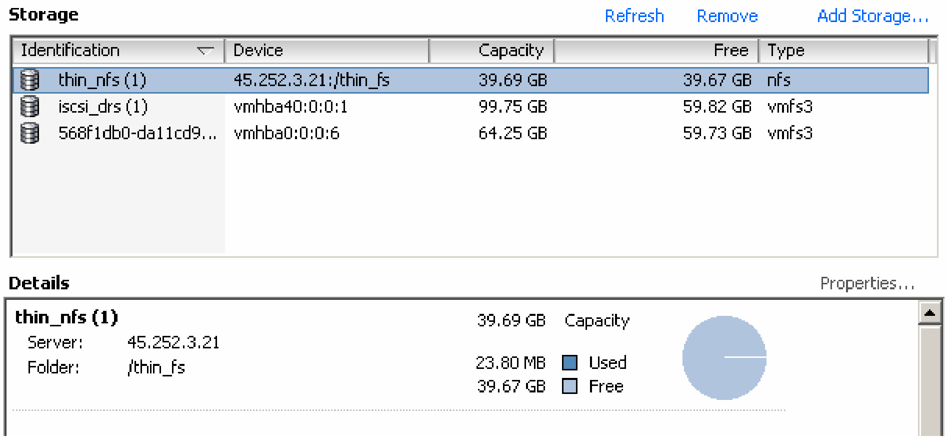


Figure 20. Storage consumption benefits of Celerra virtual provisioned file systems

Figure 21 represents a view of an ESX datastore that was created from a thinly provisioned Celerra file system with an initial size of 20 GB. The properties illustrate two important facts about this device. The first is that the virtually provisioned device has only consumed 2 MB of space within the file system. This space is used for metadata associated with the virtual disk. The second is that the ESX is able to identify the upper limit of a virtual device for both NFS and iSCSI LUNs. While we have defined a 20 GB file system using a virtual file system, ESX identifies that the file system is capable of growing to a maximum size of 40 GB and reports that fact in the properties of the datastore as seen in Figure 21.



Storage

Identification	Device	Capacity	Free	Type
thin_nfs (1)	45.252.3.21:/thin_fs	39.69 GB	39.67 GB	nfs
iscsi_drs (1)	vmhba40:0:0:1	99.75 GB	59.82 GB	vmfs3
568f1db0-da11cd9...	vmhba0:0:0:6	64.25 GB	59.73 GB	vmfs3

Details

thin_nfs (1)
Server: 45.252.3.21
Folder: /thin_fs

39.69 GB Capacity
23.80 MB ■ Used
39.67 GB □ Free

Figure 21. Space consumption of NFS datastore

After adding the virtually provisioned disk, and formatting the 20 GB virtual hard disk as an NTFS file system through the Windows guest OS, a total of 23 MB of space has been consumed.

ISCSI virtual provisioning

The default allocation option for virtual disks created in a VMFS datastore is to create them as thick or fully allocated. With Celerra thin provisioned file systems and iSCSI LUN, disk space is preserved in a similar fashion as was described previously. Disk consumption occurs only when the guest OS or application issue a SCSI write request to the LUN. When creating a thin virtual disk with ESX in a virtual provisioned iSCSI LUN there is very little space consumed even after the guest has formatted the virtual disk (about 4 percent of a 1 GB sparse iSCSI LUN).

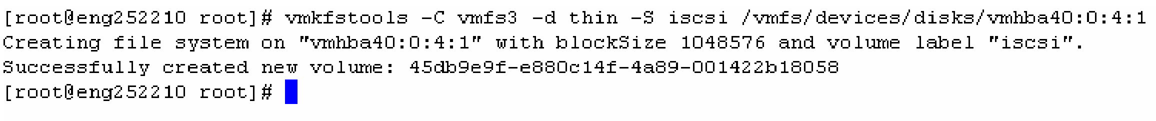
Thin provisioned LUNs initially reserve no space within the Celerra file system. This Celerra feature combined with the thin provisioning of ESX provide a very effective way to preserve disk resources. As additional virtual machines are added and disk space becomes full, Celerra will use the auto-extension feature to increase the size of the file system.

When using iSCSI, virtual provisioning provides other options that are tied to the thin provisioning option on ESX. The iSCSI LUN can be configured as virtually provisioned by specifying the –vp option.

Using the vmkfstools command you can create a thin disk device within the thin or virtual iSCSI LUN.

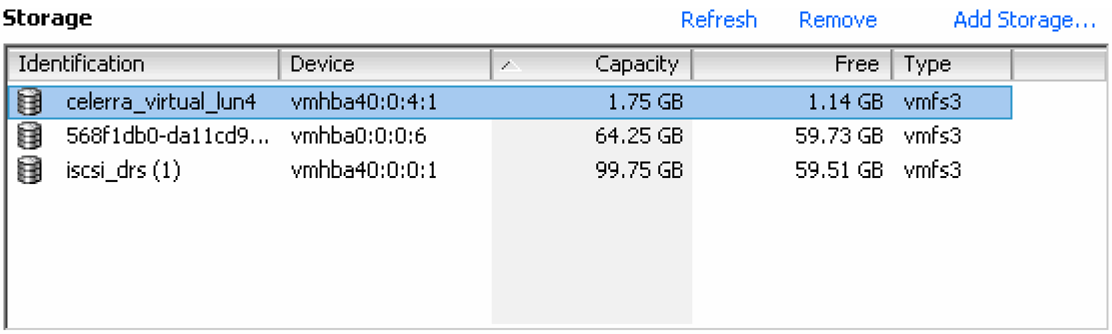
The VMware vmkfstools program is used to create and manipulate virtual disks, file systems, logical volumes, and physical storage devices on the ESX host. It can be used to create thin VMFS datastores and virtual disks.

The virtual provisioned iSCSI LUN can be used to create a thin provisioned VMFS file system as seen in the following screenshot. The VMFS datastore will use only the required blocks when allocated for new space requested by the virtual machine.



```
[root@eng252210 root]# vmkfstools -C vmfs3 -d thin -S iscsi /vmfs/devices/disks/vmhba40:0:4:1
Creating file system on "vmhba40:0:4:1" with blockSize 1048576 and volume label "iscsi".
Successfully created new volume: 45db9e9f-e880c14f-4a89-001422b18058
[root@eng252210 root]#
```

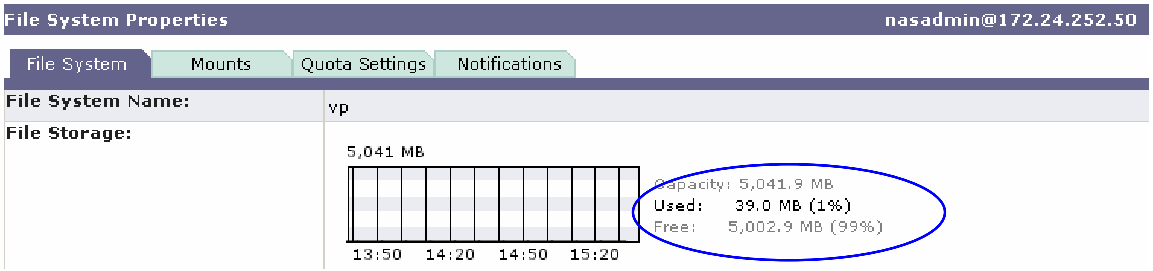
ESX thin provisioned devices can be created within the datastore created on the iSCSI LUN. The virtual device is therefore limited in the amount of blocks consumed on the storage system. ESX and the host both interpret the storage as fully allocated and actual block use is incremented only as new blocks are written to the virtual disks.



Identification	Device	Capacity	Free	Type
celerra_virtual_lun4	vmhba40:0:4:1	1.75 GB	1.14 GB	vmfs3
568f1db0-da11cd9...	vmhba0:0:0:6	64.25 GB	59.73 GB	vmfs3
iscsi_drs (1)	vmhba40:0:0:1	99.75 GB	59.51 GB	vmfs3

Figure 22. Celerra iSCSI thin provisioned LUN used for VMFS

When completed and formatted the amount of consumed space is very minimal as illustrated in the Celerra Manager File System Properties screen in Figure 23. This interface is used to provide the amount of space consumed within a file system. In this case the 5 GB LUN has allocated only 39 MB of space after being formatted as an NTFS file system on a Windows host.



File System Name: vp

File Storage:

Capacity: 5,041.9 MB
Used: 39.0 MB (1%)
Free: 5,002.9 MB (99%)

Figure 23. File System Properties screen

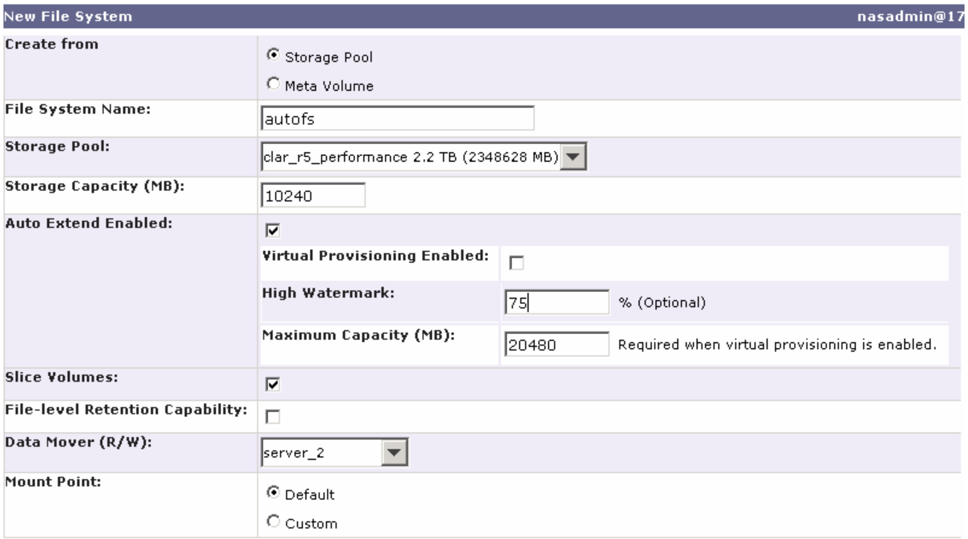
Partition alignment

VirtualCenter 2.0 or later automatically aligns VMFS partitions along the proper boundaries. This provides a performance improvement in heavy I/O environments. It is a good practice to also align the disk partitions for the virtual disks within the guest OS. Use the fdisk or partition alignment tools such as diskpart on Windows to ensure that the starting disk sector is aligned to 64k boundaries

File system extension

As the VMware Infrastructure environment expands, so too may the storage devices that support the datastores and virtual disks. Celerra file system and LUN extension provide the ability to extend the storage devices in order to meet the growth needs of the ESX.

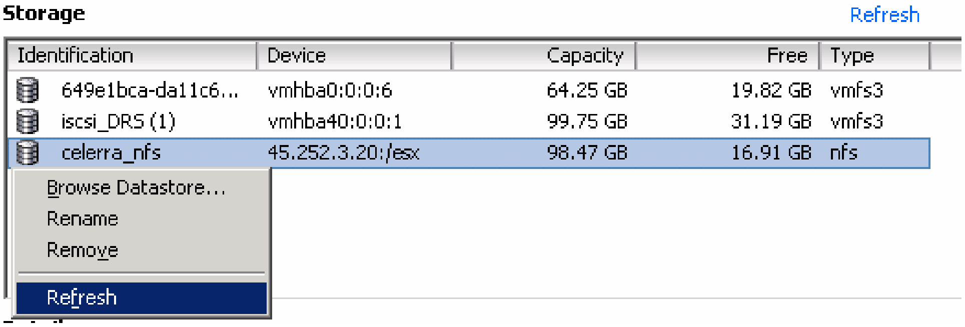
Celerra virtual file systems that were introduced in the previous section can be configured to automatically extend when the usage reaches a predefined threshold. The threshold is based upon a percentage of the virtual file system's size with the default setting of 90%. This value may need to be adjusted downward if the ESX environment is expected to incur substantial writes over a short period of time.



Create from	● Storage Pool ○ Meta Volume
File System Name:	autofs
Storage Pool:	clar_r5_performance 2.2 TB (2348628 MB)
Storage Capacity (MB):	10240
Auto Extend Enabled:	☑
Virtual Provisioning Enabled:	□
High Watermark:	75 % (Optional)
Maximum Capacity (MB):	20480 Required when virtual provisioning is enabled.
Slice Volumes:	☑
File-level Retention Capability:	□
Data Mover (R/W):	server_2
Mount Point:	● Default ○ Custom

Figure 24. Celerra automatic file system extension configuration

The default version of the Celerra file system that is fully allocated can be extended through the Celerra Management interface. Extension of a file system completes in a matter of a few seconds. An NFS file system extension is reflected in ESX by selecting the datastore and refreshing its properties, as shown in Figure 25.



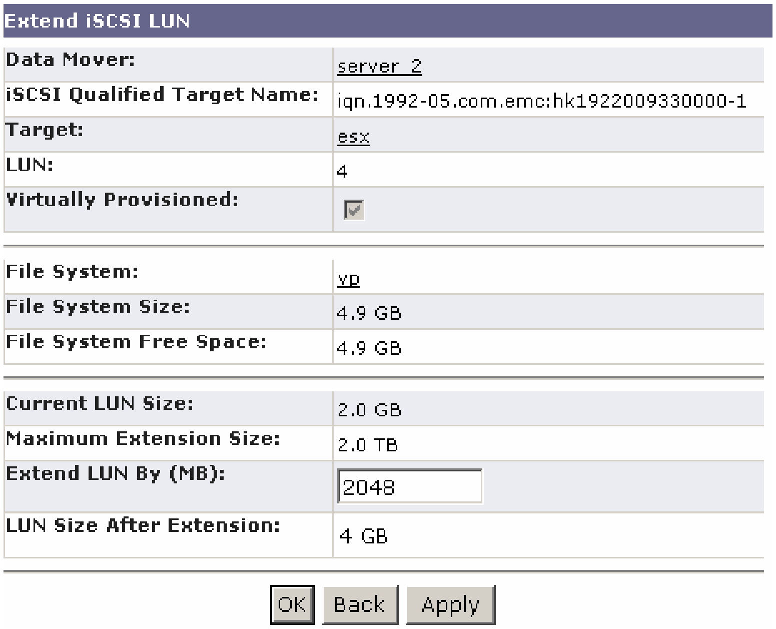
Storage

Identification	Device	Capacity	Free	Type
649e1bca-da11c6...	vmhba0:0:0:6	64.25 GB	19.82 GB	vmfs3
iscsi_DRS (1)	vmhba40:0:0:1	99.75 GB	31.19 GB	vmfs3
celerra_nfs	45.252.3.20:/esx	98.47 GB	16.91 GB	nfs

Figure 25. NFS datastore refresh after file system extension

iSCSI LUN extension

The ESX datastore vmfs size can be increased by either adding additional iSCSI LUNs or extending the size of the existing LUN used for the datastore. The LUN extension is accomplished through Celerra Manager. In Figure 26 an existing 2 GB iSCSI LUN is being used to provide a datastore used as a single 2 GB host virtual disk. The LUN is being doubled in size to provide additional space to the host. The Celerra file system being used for this datastore has 5 GB of space available for extension of the LUN. The LUN will be doubled in size so that we can add another device to a Windows client.

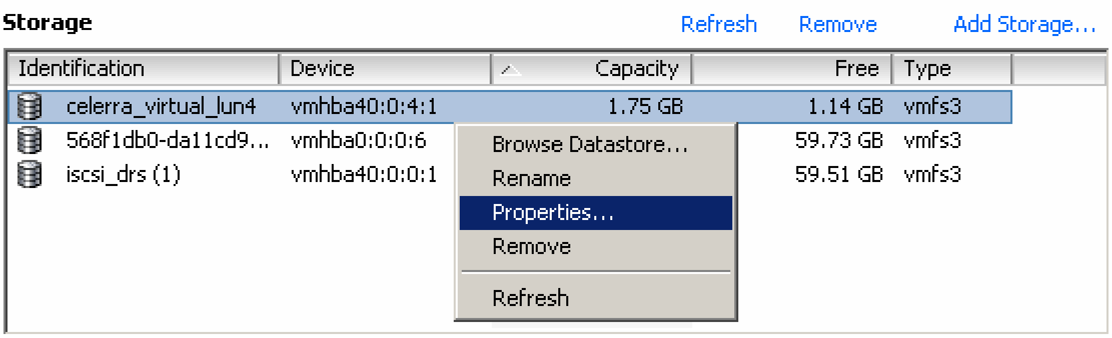


Extend iSCSI LUN

Data Mover:	server_2
iSCSI Qualified Target Name:	iqn.1992-05.com.emc:hk1922009330000-1
Target:	esx
LUN:	4
Virtually Provisioned:	☑
File System:	vp
File System Size:	4.9 GB
File System Free Space:	4.9 GB
Current LUN Size:	2.0 GB
Maximum Extension Size:	2.0 TB
Extend LUN By (MB):	2048
LUN Size After Extension:	4 GB

Figure 26. Celerra iSCSI LUN extension window

The extended iSCSI LUN is added as an additional extent or partition within the ESX Server. After the LUN is extended, VirtualCenter is used to rescan the bus identifying the changes to the iSCSI client. It is further used to add the additional capacity as another extent within the iSCSI datastore.

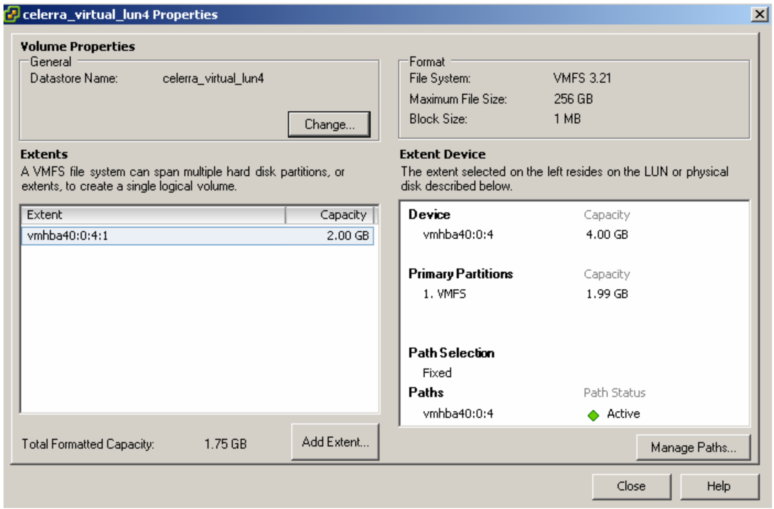


Storage

Identification	Device	Capacity	Free	Type
celerra_virtual_lun4	vmhba40:0:4:1	1.75 GB	1.14 GB	vmfs3
568f1db0-da11cd9...	vmhba0:0:0:6		59.73 GB	vmfs3
iscsi_drs (1)	vmhba40:0:0:1		59.51 GB	vmfs3

Figure 27. iSCSI LUN attributes

The Volume Properties page in Figure 28 shows the existing device properties as well as the additional capacity of the iSCSI LUN that was extended. If the LUN does not appear in this list you may need to rescan the iSCSI bus.



celerra_virtual_lun4 Properties

Volume Properties

General

Datastore Name: celerra_virtual_lun4

Format

File System: VMFS 3.21

Maximum File Size: 256 GB

Block Size: 1 MB

Extents

A VMFS file system can span multiple hard disk partitions, or extents, to create a single logical volume.

Extent	Capacity
vmhba40:0:4:1	2.00 GB

Total Formatted Capacity: 1.75 GB

Extent Device

The extent selected on the left resides on the LUN or physical disk described below.

Device	Capacity
vmhba40:0:4	4.00 GB

Primary Partitions	Capacity
1. VMFS	1.99 GB

Path Selection

Fixed

Paths	Path Status
vmhba40:0:4	◆ Active

Figure 28. Volume Properties page

If the guest OS is accessing an RDM, changes to the iSCSI LUN are managed with the host-based SCSI management utilities. SCSI inquiry commands that are initiated by Windows Disk Manager, for example, will be passed to the Celerra iSCSI target, and result in an additional partition being identified in the management interface. This bypasses the VI client management steps required for identification and configuration when provisioning additional space through VMFS.

Note - Use caution when extending the iSCSI LUN which is running a guest OS. The LUN extension process may affect running systems as the underlying device is modified. It may require that you shut down the guest OS prior to extending the device within the VI environment.

NFS vs. iSCSI

The choice of storage protocol used for ESX datastores is largely driven by preference. While there are some distinctions between the iSCSI and NFS highlighted in table 1, both protocols satisfy the core storage requirements for the VMware Infrastructure 3 feature set.

Table 1 compares the supported ESX feature set for each protocol. There is not intent to represent benefits from a performance comparison, although VMware has executed a test suite that achieved similar performance between the two protocols.[1]

Table 1. ESX feature set comparison

Feature	**NFS**	**iSCSI**
Virtual machine boot	Yes	Yes
Virtual disk support	Yes (default)	Yes
LUN extension	Yes	Yes
Replication	Yes	Yes through Replication Manager
Replication type	Crash-consistent	Application-consistent
Raw device	No	Yes
Security	UNIX_Auth	CHAP

A short list of differences is listed below:

- One obvious difference is in the support of RDM, which can be integrated with device management software to create application-consistent images of the virtual disks. RDM devices are only supported on iSCSI devices.
- Security
  - iSCSI also supports exclusive access that can be protected with CHAP authentication.
  - While NFS exported file systems can be secured, that security is applied at the network level through access-level protocols such as VLAN tagging.
- LUN extension
  - NFS provides a much easier method to extend LUNs or datastores by merely increasing the file system size.
  - LUN extension with iSCSI requires management intervention though VirtualCenter to append the LUN partition to the existing iSCSI datastore.

[1] See the VMworld presentation on IP Storage performance comparison at http://www.vmware.com.

VMware Distributed Resource Scheduler

With the introduction of VMware Infrastructure 3, VMware extends the evolution of virtual infrastructure and virtual machines that began with the first VMware ESX Server release. VMware Infrastructure 3 also introduces a revolutionary new set of infrastructure-wide services for resource optimization, high availability, and data protection built on the VMware platform. These new services deliver capabilities that previously required complex or expensive solutions to implement using only physical machines. Use of these services also provides significantly higher hardware utilization and better alignment of IT resources with business goals and priorities. In the past, companies have had to assemble a patchwork of operating system or software application-specific solutions to obtain the same benefits.

VMware Distribution Resource Scheduler (DRS) dynamically allocates and balances computing capacity across the logical resource pools defined for VMware Infrastructure. VMware DRS continuously monitors utilization across the resource pools and intelligently allocates available resources among virtual machines based on resource allocation rules that reflect business needs and priorities. Virtual machines operating within a resource pool are not tied to the particular physical server on which they are running at any given point in time. When a virtual machine experiences increased load, DRS first evaluates its priority against the established resource allocation rules and then, if justified, allocates additional resources by redistributing virtual machines among the physical servers. VMware VMotion executes the live migration of the virtual machine to a different server with complete transparency to end users.

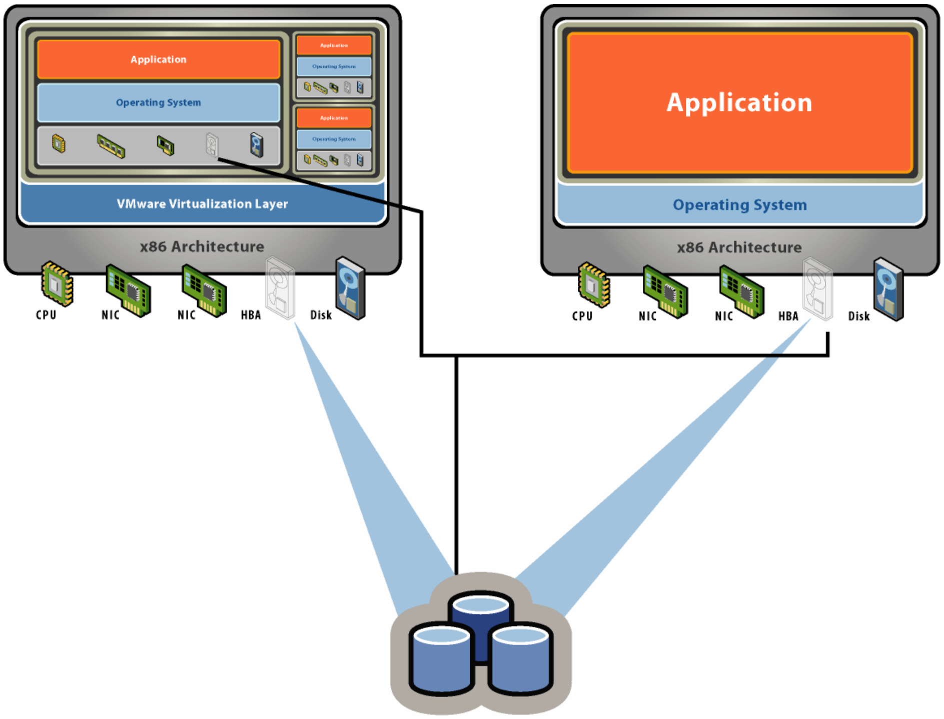


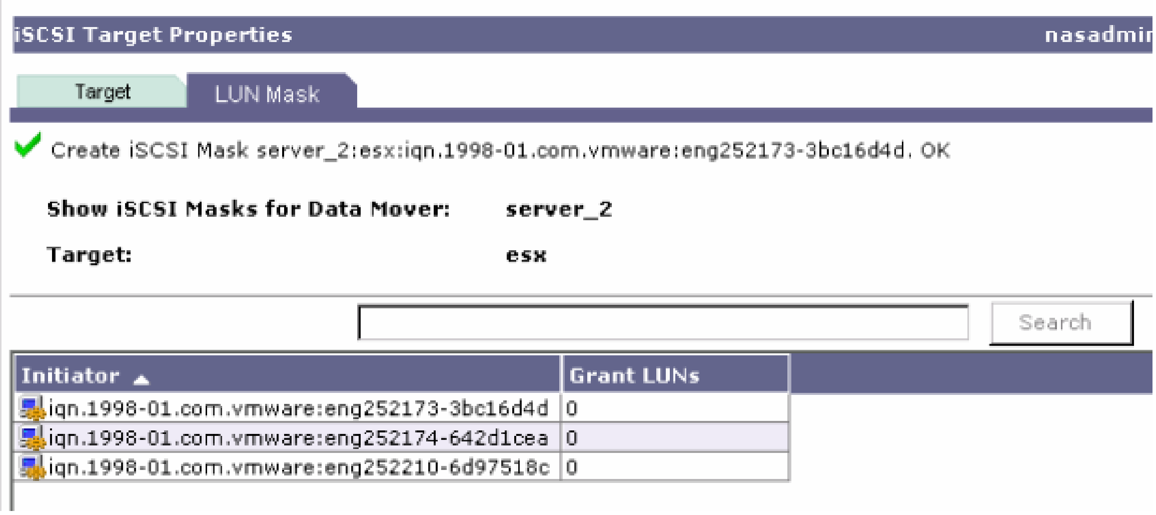
Figure 29. VMware Distributed Resource Scheduler diagram

For DRS to work properly there are configuration requirements that must be met. DRS relies on VMotion to migrate or relocate virtual machines. When a DRS policy triggers the relocation of a running virtual machine, the VMotion engine executes precondition checks to ensure the target host has the necessary resources. If the requirements of the host cannot be met the migration will not take place.

Some of the requirements that are checked are the identifiers ascribed to each network interface, the CPU of the target host, and the storage devices that the host is connected to. When the source host is using Celerra IP storage for the virtual machine, the target host must also be connected to and accessing the storage. For NFS, this is a defined datastore that has been mounted from the Celerra.

> **Best practice -** Identify a naming convention and use it on all of the objects that are defined to the host. Descriptive names that include the Celerra system, file system, and protocol will help to address configuration and troubleshooting issues when DRS and VMotion are employed in the environment

For VMotion and DRS to work properly with iSCSI, the LUNs that provide storage for the virtual machine boot and application disks must be visible to all ESX hosts. Celerra provides this service through the Celerra iSCSI LUN Mask tab of the iSCSI Target Properties page. Figure 30 illustrates a mapping for a single LUN that is being accessed by three separate ESX hosts.



✓ Create iSCSI Mask server_2:esx:iqn.1998-01.com.vmware:eng252173-3bc16d4d. OK

Show iSCSI Masks for Data Mover: server_2

Target: esx

Initiator ▲	Grant LUNs
iqn.1998-01.com.vmware:eng252173-3bc16d4d	0
iqn.1998-01.com.vmware:eng252174-642d1cea	0
iqn.1998-01.com.vmware:eng252210-6d97518c	0

Figure 30. iSCSI Target Properties page

VMotion requirements

Source and destination ESX Servers must have the following:

- Visibility to all SAN LUNs (either FC or iSCSI) and NAS devices used by virtual machines
- Gigabit Ethernet backplane
- Access to the same physical networks
- Consistently labeled virtual switch port groups
- Compatible CPUs

Replication

Celerra provides a framework that can be used with ESX for both local and remote replication. The solution is IP-based and establishes an asynchronous copy of the ESX storage device that can be used to restart the ESX environment locally or at a remote location when control of the ESX datastore is failed over to the remote Celerra.

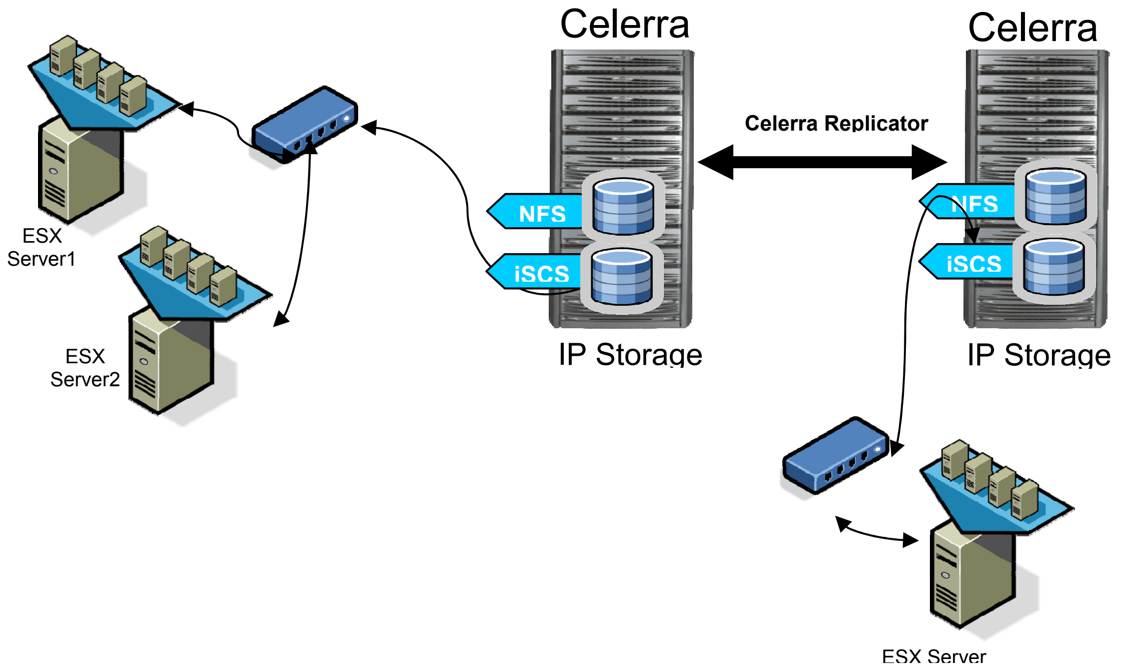


Figure 31. Celerra replication framework

Simple local and remote replication of ESX components are supported for both NFS and iSCSI, with Celerra. NFS uses volume replication based on the Celerra IP Replicator feature that replicates only changed file system data for efficient WAN use. Periodic updates between the primary and secondary Celerra provide crash-consistent NFS datastore volumes.

To use the NFS datastore on the secondary Celerra, the file system must be set for write access. This is to satisfy the SCSI reservation commands issued to the virtual disks when the guest OS is started. The options for setting the NFS datastore to read-write are to:

- Fail over the file system from primary to secondary
- Disassociate the replication relationship between the primary and secondary file systems

iSCSI replication can be managed by EMC Replication Manager (RM) to provide an application-consistent image of the virtual machine. RM offers application consistency for Windows virtual machines through VSS (Exchange and SQL Server) and VDI (SQL Server) integration with the Celerra iSCSI LUNs.

In order to use RM with iSCSI the LUNs presented to ESX must be configured as RDM or as LUNs presented to virtual guest OS machines via the Microsoft iSCSI Software Initiator.

The following are requirements for replication of virtual disks when using iSCSI:

- The device must be a RDM or presented to the guest OS via the Microsoft iSCSI Software Initiator.
- If RDM is used, the device must be configured in physical compatibility mode.
- The RM host must have Microsoft iSCSI Software Initiator installed and an iSCSI session established with the Celerra iSCSI target. The session's purpose is to associate the iSCSI initiator IQN with the iSCSI target IQN.

The EMC Replication Manager Server relies on addressing information from the Celerra iSCSI target in order to create the snap copy of the iSCSI LUN. For that reason the RM Server running in the Windows virtual machine needs an active iSCSI session established with the Celerra iSCSI target providing the LUNs to the ESX host. The RM server also requires that the Microsoft iSCSI Software Initiator be installed. Note: It does not require the LUNs be masked to the Windows iSCSI initiator. RM uses the iSCSI session to an inquiry to the target for validation purposes.

If the snaps are being promoted (read/write) to another virtual machine through the Windows virtual machine's software initiator in the guest OS, the VM will need to have an active session with the iSCSI target.

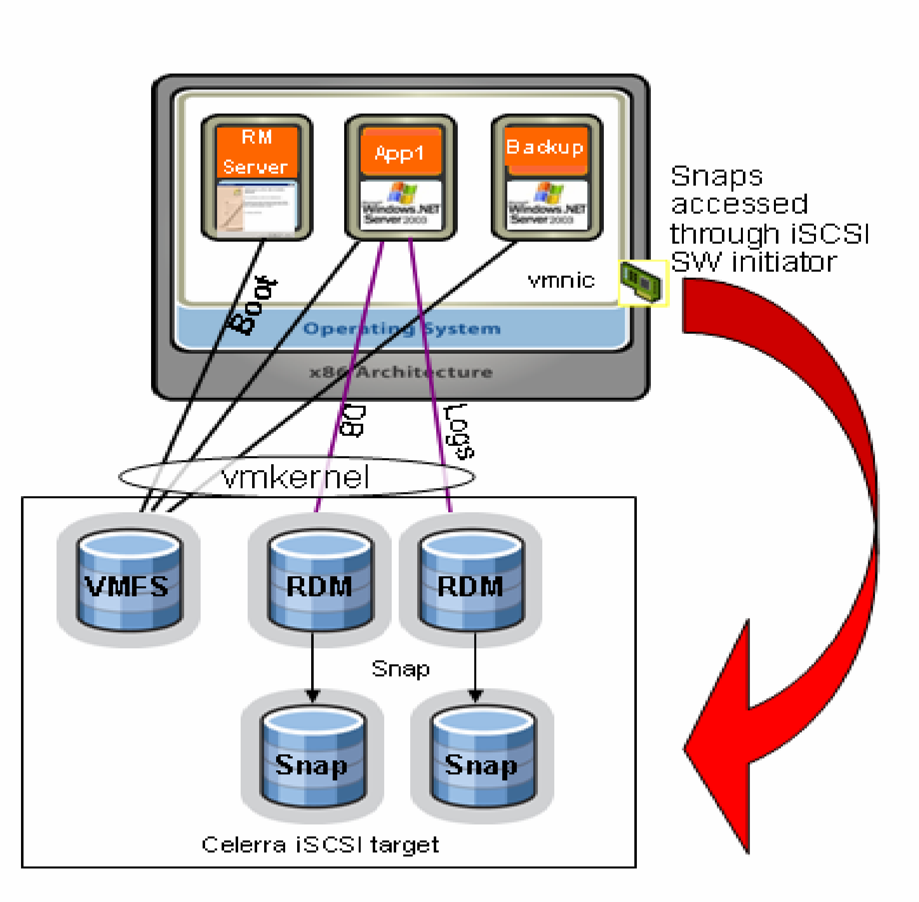


Figure 32. Replication Manager promotes iSCSI snap to virtual sachine using software iSCSI initiator

Conclusion

Leveraging the full capabilities of VMware ESX Server including VMotion, HA, and DRS, requires shared storage. As virtualization is used for production applications, storage performance and total system availability becomes critical.

The EMC Celerra platform is a high performance, high availability IP storage solution – ideal for supporting ESX Server deployments.

This white paper has covered many of the ESX and Celerra considerations. We have seen that the capability to use an NFS exported file system provides a significant benefit in the flexibility of ESX. Due to the open nature of NFS, multiple ESX hosts can access the same repository for files or folders containing virtual disks. The combination of NFS support in ESX and Celerra storage provide an extremely flexible and reliable environment. Since the virtual disks represent file objects stored within the Celerra file system, the management and mobility features of Celerra offer unique options for replicating and migrating virtual machines within the networked environment.

This white paper covered the use of Celera iSCSI LUNs to support the creation of VMFS-3 datastores. Celerra iSCSI LUNs can be:

- Used to create ESX VMware file systems used as datastores
- Directly accessed by the VMkernel as RDM disks
- Accessed through the iSCSI software initiator in the guest OS

The Celerra iSCSI LUNs are used to support operating system or application disks. In the case of VMware RDM disks and guest OSs accessing the disks via the Microsoft iSCSI Software Initiator, EMC's Replication Manger can be used to create an application-consistent image of one or more disks. The RM software can be run within the ESX environment from a Windows 2003 virtual machine and can be used to automatically and dynamically present disk replicas to other machines – physical or virtual. This is ideal for streaming backup of the application-consistent replicas or for rapid creation of multiple test and development copies of databases.

This paper also covered the use of Celerra NFS and iSCSI LUNs with advanced VMware Infrastructure features such as VMotion, DRS, and High Availability. Having these very valuable technologies fully supported with Celerra for both iSCSI and NFS provides an excellent framework to meet business needs as well as to define flexible policies for system and storage management.

Aided by the flexibility and advanced functionality of the Celerra products, VMware Infrastructure 3 and Celerra network storage platforms provide a very useful set of tools to both establish and support your virtual computing environment.

References

There is a considerable amount of additional information available on the VMware website and discussion boards as well as numerous other websites focused on hardware virtualization. Please see http://www.vmtn.com/vmtn for additional resources and user guides.

Appendix: User tips and tools

The esxcfg-nics command provides information about the defined network interfaces for the ESX host. This is a useful command in identifying the network interface status and speed.

```
/usr/bin/esxcfg-nics -l
Name    PCI      Driver  Link Speed     Duplex Description
vmnic0  06:07.00 e1000   Up   1000Mbps  Full   Intel Corporation 8254NXX Gigabit Ethernet Controller
vmnic1  07:08.00 e1000   Up   1000Mbps  Full   Intel Corporation 8254NXX Gigabit Ethernet Controller
vmnic2  0a:04.00 e1000   Up   10Mbps    Half   Intel Corporation 82540EM Gigabit Ethernet Controller (LOM)
vmnic3  0a:04.01 e1000   Up   1000Mbps  Full   Intel Corporation 82540EM Gigabit Ethernet Controller (LOM)
vmnic4  0a:06.00 e1000   Up   1000Mbps  Full   Intel Corporation 82540EM Gigabit Ethernet Controller (LOM)
vmnic5  0a:06.01 e1000   Up   1000Mbps  Full   Intel Corporation 82540EM Gigabit Ethernet Controller (LOM)
```

esxcfg-vmknic returns information about the VMkernel network interfaces. As mentioned in the text of this paper, this interface is of critical importance when using network storage from Celerra.

```
/usr/bin/esxcfg-vmknic -list
Port Group   IP Address     Netmask        Broadcast       MAC Address        MTU   Enabled
VMkernel     172.24.252.68  255.255.255.0  172.24.252.255  00:50:56:62:d3:43  1514     true
VMotion      192.168.1.3    255.255.255.0  192.168.1.255   00:50:56:67:77:72  1514     true
```

esxcfg-vmhbadevs

This command is particularly useful when using iSCSI. The default hba address for the software iSCSI initiator is vmhba40. Executing the l allows you to map the hba back to the storage devices that were created on Celerra.

```
# esxcfg-vmhbadevs | grep vmhba40
```

Also, the –m option will allow you to map the device ID to the /vmfs volume or device file.

The esxcfg-route command is used to specify the default route for the ESX VMkernel interface. Example usage would be esxcfg-route 192.168.1.1 for a class C network 192.168.1.0 using port 1 as the default next-hop router port.

esxcfg-mpath –a will also list the hba information on the device. Again, hba40 is the default device for iSCSI. This is evidenced by the IQN name associated with the hba identifier.

```
# esxcfg-mpath -a

vmhba1 210000e08b1a764e 11:2.0

vmhba2 210000e08b1943ea 11:3.0

vmhba3 210100e08b3943ea 11:3.1

vmhba40 iqn.1998-01.com.vmware:eng25265-4ad12727 sw
```

One of the key options is the use of the esxcfg-nas command. This provides you with the ability to both configure and validate the NFS connections from the ESX host to the Data Mover.

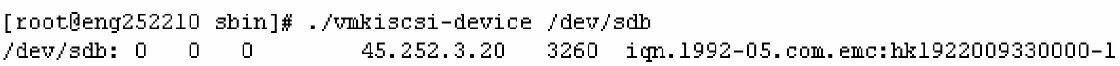
```
#   esxcfg-nas -l

Celerra NFS is /esx from 172.24.252.87 mounted

ESX_template is /u03 from 172.24.252.87 mounted
```

esxcfg-rescan will rescan the SCSI bus by specifying the HBA device that you want to scan for changes to the device graph of the ESX host.

```
# esxcfg-rescan vmhba40

Rescanning vmhba40...done.

On scsi4, removing: 10:0 10:2 10:3 10:4 10:6 12:0 12:1 12:2 13:0.

On scsi4, adding: 10:0 10:2 10:3 10:4 10:6 12:0 12:1 12:2 13:0.
```

The esxcfg-swiscsi command provides the interaction with the iSCSI service. It allows you to enable, and disable the iSCSI software service on the ESX host as well as rescan the iSCSI bus for device changes.

If you want to identify the iSCSI device from the command line of the ESX issue, the following command with the iSCSI device you are trying to map back to the Celerra iSCSI target.



```
[root@eng252210 sbin]# ./vmkiscsi-device /dev/sdb
/dev/sdb: 0   0   0        45.252.3.20   3260  iqn.1992-05.com.emc:hk1922009330000-1
```

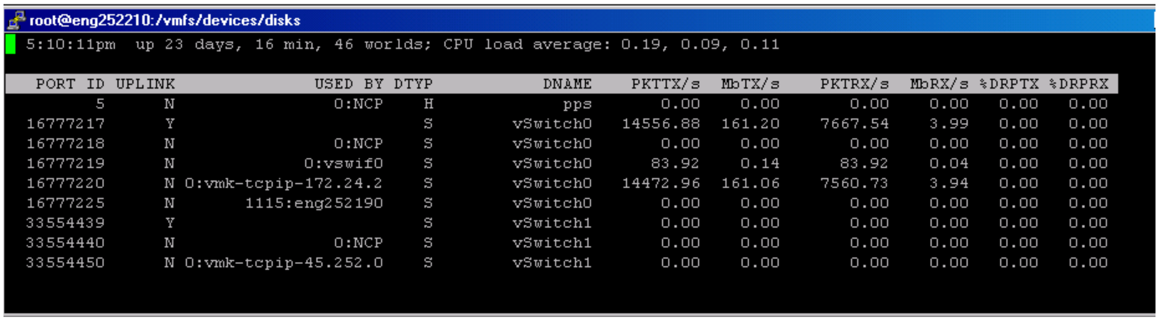
Verify that you are connected to the proper target via command line with the vmkiscsi-ls command. As you can see from the output this is based on the Cisco iSCSI driver, so if you are familiar with that driver interface on Linux you should be familiar with the other options for listing devices (-l) and configuration (-c).

```
[root@eng252210 sbin]# ./vmkiscsi-ls
*******************************************************************************
       Cisco iSCSI Driver Version ... 3.4.2 (16-Feb-2004 )
*******************************************************************************
TARGET NAME             : iqn.1992-05.com.emc:hk1922009330000-1
TARGET ALIAS            : esx
HOST NO                 : 1
BUS NO                  : 0
TARGET ID               : 0
TARGET ADDRESS          : 45.252.3.20:3260
SESSION STATUS          : ESTABLISHED AT Mon Jan 29 14:32:03 2007
NO. OF PORTALS          : 2
PORTAL ADDRESS 1        : 45.252.3.20:3260,1
PORTAL ADDRESS 2        : 172.24.252.60:3260,1
SESSION ID              : ISID 00023d000001 TSID 06
*******************************************************************************
```

Tips for ESX with Celerra

1. The VMkernel TCP/IP stack routing table determines packet flow boundaries:
   a. Put IP Storage and VMotion on separate subnets for isolation.
   b. Traffic will go through the same virtual switch if they are in the same subnet.
2. Use vmkping to check connectivity between the VMkernel and the Celerra network interface. Ping will use the console interface that is not the VMkernel TCP/IP stack.

3. Monitoring the network through esxtop is a useful method to determine the amount of resources, that is, network packets, being transmitted across the VMkernel interfaces.



```
 5:10:11pm  up 23 days, 16 min, 46 worlds; CPU load average: 0.19, 0.09, 0.11

   PORT ID UPLINK                 USED BY DTYP          DNAME    PKTTX/s   MbTX/s    PKTRX/s   MbRX/s %DRPTX %DRPRX
         5      N                   0:NCP    H            pps       0.00     0.00       0.00     0.00   0.00   0.00
  16777217      Y                            S       vSwitch0   14556.88   161.20    7667.54     3.99   0.00   0.00
  16777218      N                   0:NCP    S       vSwitch0       0.00     0.00       0.00     0.00   0.00   0.00
  16777219      N                0:vswif0    S       vSwitch0      83.92     0.14      83.92     0.04   0.00   0.00
  16777220      N 0:vmk-tcpip-172.24.2       S       vSwitch0   14472.96   161.06    7560.73     3.94   0.00   0.00
  16777225      N          1115:eng252190    S       vSwitch0       0.00     0.00       0.00     0.00   0.00   0.00
  33554439      Y                            S       vSwitch1       0.00     0.00       0.00     0.00   0.00   0.00
  33554440      N                   0:NCP    S       vSwitch1       0.00     0.00       0.00     0.00   0.00   0.00
  33554450      N 0:vmk-tcpip-45.252.0       S       vSwitch1       0.00     0.00       0.00     0.00   0.00   0.00
```

Figure 33. Using esxtop to view network interface statistics