

WHITE PAPER

VMware ESX Server 3.0

iSCSI Best Practices and Configuration Guide



fo3dus

Contents

Introduction.....	1
This Paper	1
iSCSI Explained.....	2
Initiators.....	2
Discovery and Logins.....	2
Authentication	3
Designing the Environment	4
Network Considerations.....	4
Disk Considerations.....	5
Initiator Considerations.....	5
Conclusion	8

Conventions

The following conventions are used in this manual.

Style	Purpose
blue(online only)	Cross references, links
<code>Courier</code>	Commands, filenames, directories, paths, user input
Semi-Bold	Interactive interface objects, keys, buttons
Bold	Items of highlighted interest, terms
<i>italic</i>	Vairables, parameters
<i>italic</i>	Web addresses

The following abbreviations are used in the graphics in this manual.

Abbreviation	Description
VC	VirtualCenter
VM	Virtual Machine
SAN	Storage area network type datastore, shared between managed hosts
SP	Storage processor
OS	Operating System

Introduction

VMware Infrastructure 3 is the industry's first full Infrastructure virtualization suite that empowers enterprises and small businesses alike to transform, manage and optimize their IT systems infrastructure through virtualization. VMware Infrastructure 3 delivers comprehensive virtualization, management, resource optimization, application availability and operational automation capabilities in an integrated offering.

With the release of VMware ESX Server 3, VMware has expanded support for storage options available for the VMFS cluster file system. In addition to Fibre Channel SAN storage, VMware Infrastructure 3 adds support for iSCSI SAN storage. The adoption of iSCSI storage is quickly expanding throughout IT datacenters worldwide. The benefits of an iSCSI storage solution are evident:

Enterprise class storage at a fraction of the price of a Fibre Channel SAN

iSCSI SAN can be integrated into an existing Ethernet infrastructure.

iSCSI storage permits IT departments to deploy their storage solutions using existing resources without the need for additional hardware or IT staff.

The support of iSCSI SAN in a virtual environment will provide IT administrators a low-cost, highly available solution for deploying virtual machines. For the first time, IT departments that could not afford a Fibre Channel SAN will now be able to take advantage of advanced VMware Infrastructure functionality such as redundant storage paths, centralized storage environment, live migration of running virtual machines with VMware VMotion, dynamic allocation and balancing of computing capacity with VMware DRS and automatic re-start of virtual machines affected by server failure with VMware HA.

Customers with existing Fibre Channel SAN storage environments can also benefit from the addition of iSCSI storage to their environment by incorporating a "Tiered" solution. Rather than place all virtual machines on an expensive Fibre Channel SAN where the cost of storage is at a premium, businesses can decide between the Fibre Channel storage or iSCSI platforms. A tiered solution gives businesses a choice to locate virtual machines based on the level of importance and business requirements of the application as well as performance levels required. With new iSCSI support, businesses can now make better overall usage of their existing Fibre Channel SAN investment by classifying virtual machines as either fibre storage (Tier 1) or iSCSI (Tier 2). This type of "Tiered" solution builds upon and increases the flexibility that a VMware virtual infrastructure provides.

This Paper

This paper provides information regarding the configuration and best practices of iSCSI with VMware ESX Server 3. This paper assumes that the audience has an understanding of the basic concepts surrounding virtualization and a VMware ESX Server configuration.

iSCSI Explained

Internet Small Computer Systems Interface (iSCSI - RFC 3720) is a relatively new implementation for SANs. Simply put, iSCSI is a means of transport for the SCSI protocol, encapsulating the suite of SCSI commands within a TCP packet. The commands are encapsulated at the block I/O level, rather than the file I/O level making the storage appear local to the host OS. Since all the commands are at the block level, all SCSI rules still apply. For instance, a SCSI storage device cannot be shared without a filesystem that allows SCSI command sharing. The VMFS cluster allows for multiple hosts to connect to the same storage concurrently.

Now that flexible and inexpensive storage can be connected over a standard Ethernet infrastructure, a world of new possibilities is presented to IT Administrators on how to deploy storage solutions.

Initiators

An iSCSI initiator provides a means for an iSCSI client (the ESX Server) to access storage from the iSCSI target (the storage device). There are two implementations of iSCSI initiators; hardware and software.

The software initiator is simply a driver that interacts with the node OS' TCP/IP stack to contact the iSCSI target via an existing Ethernet card. This adds a significant amount of workload to the node's CPU(s) because the iSCSI protocol needs to be unpackaged and read by the host CPU, resulting in a performance decrease under any type of significant I/O load. Implementations of a software initiator should be restricted to areas where performance is not a requirement, such as a development environment. No additional software is needed to configure a software initiator on VMware ESX Server.

A hardware initiator is an adapter card (commonly referred to as a Host Bus Adapter, or HBA) that implements connectivity from the iSCSI client to the iSCSI target but does so in a more efficient approach. Rather than utilizing the host OS' CPU cycles to process the iSCSI protocol, the host offloads the traffic to the HBA where it is processed. This results in a much more efficient approach for iSCSI connectivity by saving precious CPU cycles for the OS and applications. In an iSCSI SAN, a hardware initiator can meet the performance levels that a production environment requires.

Discovery and Logins

Discovery allows an initiator to find the iSCSI target(s) to which it has access. This requires a minimum of user configuration. Several methods of discovery are available:

- A static list of targets defined at the initiator
- A query to the iSCSI host on a specified TCP port
- A Storage Name Server (SNS)
- Service Location Protocol (SLP)

A login will occur once an iSCSI target is discovered and a TCP/IP connection gets created. This connection provides that means for iSCSI data transfer, the negotiation of parameters and (optionally) authentication. By default, iSCSI hosts and nodes communicate across TCP port 3260. After completion of the login phase, the initiator can begin sending SCSI commands to the target.

Authentication

Challenge Handshake Authentication Protocol (CHAP) is an authentication method that is commonly used in iSCSI implementations. The goal of CHAP is to authenticate the iSCSI client with the target and to prevent any illegitimate access to the target's storage. During the CHAP authentication process, the target sends its ID along with a random key to the initiator. The initiator replies with a hash value containing the same random key, the initiator's ID and a CHAPsecret, or password. If the CHAPsecret is correct, the target grants access to the initiator.

Note: ESX Server does not support Kerberos, Secure Remote Protocol (SRP), or public key authentication methods for iSCSI. Additionally, it does not support IPsec authentication and encryption.

Designing the Environment

In order to provide the highest level of services for the iSCSI environment, proper design will improve performance and provide redundancy in the event of a hardware failure. The following are suggestions for implementing an iSCSI solution with a VMware ESX Server environment.

Network Considerations

To ensure the best results for an iSCSI implementation on VMware ESX Server, the proper network architecture is essential to a healthy and performing environment. There are several different ways of approaching network architecture and fine-tuning it to meet the demands of a virtual environment.

Below are recommendations for configuring the iSCSI network.

- Consider creating a physically or logically segmented network for iSCSI traffic, separating the existing LAN traffic from the storage traffic. Inherently, Fibre Channel SANs have an added security advantage over iSCSI SANs because they are based on a physically isolated fabric. In a typical network every server, workstation, printer, etc. communicates using an Ethernet-based network, the same network that the iSCSI nodes and hosts reside. This opens iSCSI traffic to a wide array of security vulnerabilities that do not exist in the deployment of a Fibre Channel SAN.
- Use Gigabit network adapters for iSCSI initiators. A 100Mbit connection becomes saturated and does not scale for multiple VMs. Also, ensure that the iSCSI interface is set to Full Duplex or configured to negotiate at Full Duplex.
- For production environments, use a hardware iSCSI initiator. By offloading the iSCSI protocol to the hardware-based initiator, not only is disk performance improved, CPU cycles become freed up allowing more resources for the VMs.
- In an environment where creating a separate network for isolating the iSCSI environment is not feasible, network administrators should configure Quality of service (QoS) on their network switches to prioritize iSCSI traffic.
- Use network authentication when a private network infrastructure is not available for iSCSI traffic.
- Use static IP addresses for initiators. If DHCP must be used and the storage is on a public LAN, be sure CHAP authentication is implemented.
- Network administrators should consider the benefits of using “Jumbo frames”, especially when using a software initiator (for the initial release of VMware ESX Server 3, VMware supports the QLogic 4010 which does not support jumbo frames). A standard Ethernet frame size is 1500 bytes. This has been the standard from 10Mbit to 100Mbit to 1Gbit network infrastructures and has remained the same over time to keep consistent interoperability with legacy network devices throughout an environment.

A “jumbo frame” is a frame that is 9216 bytes in size; six times that of the default standard. By enabling jumbo frames for an iSCSI implementation, the flow of traffic improves, specifically under busy network conditions. Sending fewer frames generates fewer interrupts via the iSCSI initiator, freeing up valuable CPU cycles. Another benefit to using jumbo packets involves less header inspection, also freeing up additional CPU cycles. Each of the packets passed in a session contain the same sized header regardless of the frame size (approximately 100 bytes). Therefore, proportionally, when sending large amounts of

data there will be less processing needed in a jumbo frame setup than that of a standard configuration. While these savings may seem insignificant, when processing millions of packets the results of fewer headers inspection and less overall traffic prove beneficial in a busy virtual environment.

Two important caveats to keep in mind for configuring jumbo frames:

1. Ensure the initiator (NIC or HBA) supports jumbo frames.
2. A switch port configured for jumbo frames can only pass traffic to other ports configured in the same fashion.

Disk Considerations

Below are recommendations for configuring the iSCSI storage.

- iSCSI vendors offer a variety of RAID implementations to meet the performance and redundancy needs of different applications. Each LUN should have the right RAID level and storage characteristic for the specific applications running in the virtual machines allocated to the LUN.
- If a VM needs more space than can be allocated from the existing VMFS, avoid extending VMFS volumes. Instead, create a new LUN with a new VMFS volume. An extended volume will not balance data across both physical participants, resulting in hot and cold spots the target.
- Enable read/write cache on the iSCSI target.
- Where possible, dedicate disk/RAID groups to LUNs that will host VMFS volumes. Remember that multiple OS' will be requesting I/O from the disk/RAID group simultaneously.

Initiator Considerations

By losing connectivity to the remote iSCSI storage, virtual machines will freeze and/or the ESX Server host could experience a kernel panic. By enabling redundant iSCSI initiators and switches, the possibility of a service outage by a hardware failure reduces greatly. As an exercise, diagram the proposed environment attempt to identify any single points of failure (SPOF). Below is a sample diagram of a fully redundant iSCSI implementation.

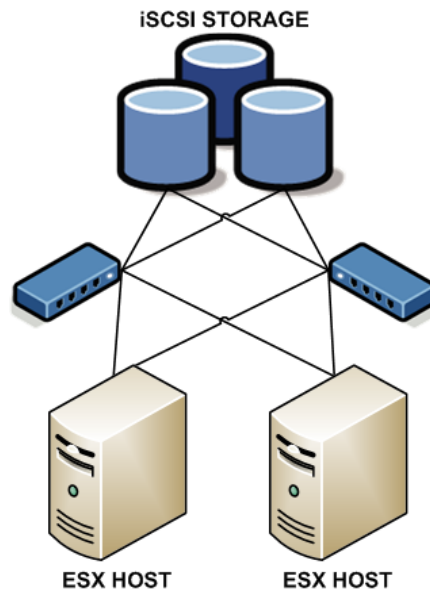


Figure 1 - A topology providing redundant paths to iSCSI storage

Hardware initiator specific

- To enable path failover for any given LUN, each initiator must have access to the LUN(s). By defining a “Fixed” failover path policy, the storage LUN uses the same “preferred” path whenever available. If offline, the ESX Server host will choose the next available path to use and fail back to the “preferred” path once it becomes available again. Alternatively, the Most Recently Used (MRU) failover policy continues to use the same path until a service interruption. With “MRU”, there is no fail back action. Failover policies are configured from either the Virtual Infrastructure (VI) Client or the ESX Server command line.

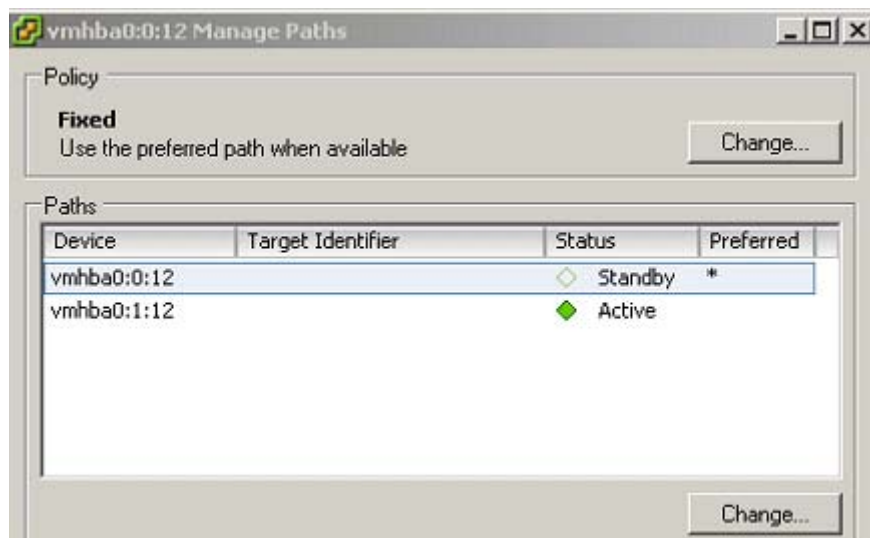


Figure 2 – Managing the failover path policy from the Virtual Infrastructure Client

- Many iSCSI targets utilize more than one storage processor (SP) to reach higher levels of performance and to provide redundancy in the case of a failure. Depending on the

vendor, the target will either use an active/active or active/passive method of accessing a LUN. When using an active/active iSCSI target, use “Fixed Mode” failover policy. With an active/passive target, use “MRU”.

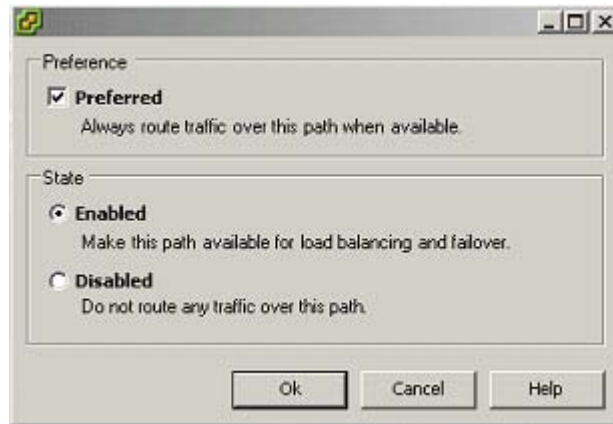


Figure 3 - Setting a fixed path policy from the Virtual Infrastructure client

- Manually spread I/O across all available initiators and SP's. This helps to balance I/O traffic on both the ESX Server host and the iSCSI target. In the event of a failure, the alternate initiator and/or the alternate SP will assume ownership of the LUN.
- Disconnect the HBAs during ESX Server installation when you install an ESX Server host to an existing SAN.
- Set maximum queue depths to prevent bottlenecks at the iSCSI initiator during high I/O periods. Find the appropriate device name in `/etc/vmware/esx.conf` (this example uses the syntax for the QLogic 4010 HBA):

```
/device/001:02:0/vmknname = "vmhba1"
```

...and enter the following:

```
/device/001:02:0/options = "ql4xmaxqdepth=nn"
```

Software initiator specific

- When using a software initiator, use a dedicated virtual switch to lower chances of having network traffic intercepted by potential attackers during transmission. This configuration will physically segment vm network traffic and iSCSI traffic.
- There can only be one software initiator configured in an ESX Server host. When configuring a virtual switch that will provide iSCSI connectivity, bind multiple network connections to the switch to provide load balancing and redundancy.

Conclusion

iSCSI is an excellent fit for many VMware ESX Server environments. When implemented correctly, IT staffs can cut costs and save IT resources while surrendering little to no functionality. Implementing the iSCSI environment correctly through careful design and planning can provide a high level of uptime while minimizing costs associated with a virtual infrastructure deployment.

About the Author

Rob Daly is a Senior Systems Engineer for Foedus. He is a VMware Certified Professional (VCP) and has worked in the IT industry for 9 years. Over the last four years, Rob has concentrated his focus on virtual infrastructure, having designed and implemented a wide variety of VMware ESX solutions for companies throughout the US, several of those within the Fortune 100 realm.



VMware, Inc. 3145 Porter Drive Palo Alto CA 94304 USA Tel 650-475-5000 Fax 650-475-5001 www.vmware.com
© 2006 VMware, Inc. All rights reserved. Protected by one or more of U.S. Patent Nos. 6,397,242, 6,496,847, 6,704,925, 6,711,672, 6,725,289, 6,735,601, 6,785,886, 6,789,156 and 6,795,966; patents pending. VMware, the VMware "boxes" logo and design, Virtual SMP and VMotion are registered trademarks or trademarks of VMware, Inc. in the United States and/or other jurisdictions. Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation. Linux is a registered trademark of Linus Torvalds. All other marks and names mentioned herein may be trademarks of their respective companies.



fo \equiv dus