Compellent Storage Center

Best Practices with ESX/ESXi 4.x (vSphere)

This document has been archived and will no longer be maintained or updated. For more information go to the Storage Solutions Technical Documents page on Dell TechCenter or contact support.

Compellent Corporate Office

Compellent Technologies
7625 Smetana Lane
Eden Prairie, Minnesota 55344

www.compellent.com
Contents

Disclaimers ........................................................................................................... 4
General Syntax ..................................................................................................... 4
Conventions .......................................................................................................... 4
Where to Get Help ............................................................................................. 5
Customer Support ............................................................................................ 5
Document Revision ........................................................................................... 5
Overview ...................................................................................................................... 6
Prerequisites ......................................................................................................... 6
Intended audience ................................................................................................ 6
Introduction ........................................................................................................... 6
Fibre Channel Switch Zoning ....................................................................................... 7
  Single Initiator Multiple Target Zoning .............................................................. 7
  Port Zoning ...................................................................................................... 7
  WWN Zoning .................................................................................................. 7
  Virtual Ports .................................................................................................... 7
Host Bus Adapter Settings ........................................................................................... 8
  QLogic Fibre Channel Card BIOS Settings ......................................................... 8
  Emulex Fiber Channel Card BIOS Settings ....................................................... 8
  QLogic iSCSI HBAs ......................................................................................... 8
Modifying Queue Depth in an ESX Environment ......................................................... 9
  Overview ......................................................................................................... 9
  Host Bus Adapter Queue Depth ....................................................................... 9
  Modifying ESX Storage Driver Queue Depth and Timeouts ......................... 10
  Modifying the VMFS Queue Depth for Virtual Machines ............................ 11
  Modifying the Guest OS Queue Depth .......................................................... 12
Setting Operating System Disk Timeouts ................................................................. 14
Guest Virtual SCSI Adapters ..................................................................................... 15
Mapping Volumes to an ESX Server ........................................................................ 16
  Basic Volume Mapping Concepts .................................................................. 16
  Basic Volume Mapping in Storage Center 4.x and earlier .............................. 16
  Basic Volume Mappings in Storage Center 5.x and later .............................. 16
  Multi-Pathed Volume Concepts .................................................................... 17
  Multi-Pathed Volumes in Storage Center 4.x and earlier ............................... 18
  Multi-Pathed Volumes in Storage Center 5.x and later .................................. 18
  Configuring the VMware iSCSI software initiator for a single path .......... 20
  Configuring the VMware iSCSI software initiator for multipathing ........... 21
  VMware Multi-Pathing Policies .................................................................... 21
  Fixed Policy ................................................................................................. 21
  Round Robin ............................................................................................... 22
  Most Recently Used (MRU) ......................................................................... 23
  Multi-Pathing using a Fixed path selection policy ...................................... 23
  Multi-Pathing using a Round Robin path selection policy .......................... 24
  Asymmetric Logical Unit Access (ALUA) ...................................................... 24
  Additional Multi-pathing resources .............................................................. 24
Boot from SAN ...................................................................................................... 25
Configuring boot from SAN .................................................................................. 25
Disclaimers

Information in this document is subject to change without notice. © 2010 Compellent Technologies. All rights reserved.

Reproduction in any manner without the express written permission of Compellent Technologies is strictly prohibited.

Trademarks used in this text are property of Compellent Technologies, or their respective owners.

General Syntax

Table 1: Document syntax

<table>
<thead>
<tr>
<th>Item</th>
<th>Convention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menu items, dialog box titles, field names, keys</td>
<td><strong>Bold</strong></td>
</tr>
<tr>
<td>Mouse click required</td>
<td>Click:</td>
</tr>
<tr>
<td>User Input</td>
<td><strong>Monospace Font</strong></td>
</tr>
<tr>
<td>User typing required</td>
<td><strong>Type:</strong></td>
</tr>
<tr>
<td>Website addresses</td>
<td><a href="http://www.compellent.com">http://www.compellent.com</a></td>
</tr>
<tr>
<td>Email addresses</td>
<td><a href="mailto:info@compellent.com">info@compellent.com</a></td>
</tr>
</tbody>
</table>

Conventions

- **Note**
  Notes are used to convey special information or instructions.

- **Timesaver**
  Timesavers are tips specifically designed to save time or reduce the number of steps.

- **Caution**
  Caution indicates the potential for risk including system or data damage.

- **Warning**
  Warning indicates that failure to follow directions could result in bodily harm.
Where to Get Help

If you have questions or comments contact:

**Customer Support**

Tel 866-EZSTORE (866.397.8673)

[support@compellent.com](mailto:support@compellent.com)

**Document Revision**

<table>
<thead>
<tr>
<th>Date</th>
<th>Revision</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/16/2009</td>
<td>3</td>
<td>Initial Release</td>
</tr>
<tr>
<td>9/30/2009</td>
<td>4</td>
<td>Added section for space recovery</td>
</tr>
<tr>
<td>11/3/2009</td>
<td>5</td>
<td>Modified RDM recommendations</td>
</tr>
<tr>
<td>11/10/2010</td>
<td>6</td>
<td>Add ESX 4.1 and SC 5.x info</td>
</tr>
<tr>
<td>1/4/2011</td>
<td>7</td>
<td>Minor revisions, Add document number</td>
</tr>
</tbody>
</table>
Prerequisites

This document assumes the reader has had formal training or has advanced working knowledge of the following:

- Installation and configuration of VMware vSphere 4.x
- Configuration and operation of the Compellent Storage Center
- Operating systems such as Windows or Linux

Intended audience

This document is highly technical and intended for storage and server administrators, as well as other information technology professionals interested in learning more about how VMware ESX 4.0 integrates with the Compellent Storage Center.

Introduction

This document will provide configuration examples, tips, recommended settings, and other storage guidelines a user can follow while integrating VMware ESX Server with the Compellent Storage Center. This document has been written to answer many frequently asked questions with regard to how VMware interacts with the Compellent Storage Center's various features such as Dynamic Capacity, Data Progression, and Remote Instant Replay.

Compellent advises customers to read the Fiber Channel or iSCSI SAN configuration guides, which are publicly available on the VMware ESX documentation pages to provide additional important information about configuring your ESX servers to use the SAN.

Please note that the information contained within this document is intended only to be general recommendations and may not be applicable to all configurations. There are certain circumstances and environments where the configuration may vary based upon your individual or business needs.
Fibre Channel Switch Zoning

Zoning your fibre channel switch for an ESX server is done much the same way as you would for any other server connected to the Compellent Storage Center. Here are the fundamental points:

Single Initiator Multiple Target Zoning

Each fibre channel zone you create should have a single initiator (HBA port) and multiple targets (Storage Center front-end ports). This means that each HBA port needs its own fibre channel zone containing itself and the Storage Center front-end ports. Zoning your ESX servers by either port number or WWN is acceptable.

Port Zoning

If the Storage Center front-end ports are plugged into switch ports 0, 1, 2, & 3, and the first ESX HBA port is plugged into switch port 10, the resulting zone should contain switch ports 0, 1, 2, 3, & 10.

Repeat this for each of the HBAs in the ESX server. If you have disjoint fabrics, the second HBA port in the host should have its zone created in the second fabric.

In smaller implementations, it is recommended you use port zoning to keep the configuration simple.

WWN Zoning

When zoning by WWN, the zone only needs to contain the host HBA port and the Storage Center front-end “primary” ports. In most cases, it is not necessary to include the Storage Center front-end “reserve” ports because they are not used for volume mappings. For example, if the host has two HBAs connected to two disjoint fabrics, the fiber channel zones would look something like this:

**Name: ESX1-HBA1**  (Zone created in fabric 1)
- WWN: 2100001B32017114 (ESX1 HBA Port 1)
- WWN: 5000D31000036001 (Controller1 front-end primary plugged into fabric 1)
- WWN: 5000D31000036009 (Controller2 front-end primary plugged into fabric 1)

**Name: ESX1-HBA2**  (Zone created in fabric 2)
- WWN: 210000E08B930AA6 (ESX1 HBA Port 2)
- WWN: 5000D31000036002 (Controller1 front-end primary plugged into fabric 2)
- WWN: 5000D3100003600A (Controller2 front-end primary plugged into fabric 2)

Virtual Ports

If the Storage Center is configured to use fiber channel virtual ports, all of the Front End ports within each Fault Domain should be included in the zone with the appropriate ESX HBA.
Host Bus Adapter Settings

Make sure that you configure the HBA BIOS settings in your ESX server according to the latest “Storage Center System Manager User Guide” found on Knowledge Center. At the time of this writing, here are the current Compellent recommendations:

QLogic Fibre Channel Card BIOS Settings

- The “connection options” field should be set to 1 for point to point only
- The “login retry count” field should be set to 60 attempts
- The “port down retry” count field should be set to 60 attempts
- The “link down timeout” field should be set to 30 seconds.
- The “queue depth” (or “Execution Throttle”) field should be set to 255.
  - This queue depth can be set to 255 because the ESX driver module ultimately controls the queue depth of the HBA.

Emulex Fiber Channel Card BIOS Settings

- The “lpfc_devloss_tmo” (formerly “nodev_tmo”) field should be set to 60 seconds.
  - On Windows, it is called node timeout.
- The “topology” field should be set to 1 for point to point only
- The “queuedepth” field should be set to 255
  - This queue depth can be set to 255 because the ESX driver module ultimately controls the queue depth of the HBA.

QLogic iSCSI HBAs

- The “ARP Redirect” must be enabled for controller failover to work properly with hardware iSCSI HBAs.
  - For steps to enable ARP Redirect on the iSCSI adapter consult the following VMware documentation:
    - VMware Document: “iSCSI SAN Configuration Guide”
  - Enabling ARP redirect for Hardware iSCSI HBAs
    - [http://kb.vmware.com/kb/1010309](http://kb.vmware.com/kb/1010309)
Overview

Queue depth is defined as the amount of disk transactions that are allowed to be "in flight" between an initiator and a target, where the initiator is typically a server port and the target is typically the Storage Center front-end port.

Since any given target can have multiple initiators sending it data, the initiator queue depth is generally used to throttle the number of transactions being sent to a target to keep it from becoming "flooded". When this happens, the transactions start to pile up causing higher latencies and degraded performance. That being said, while increasing the queue depth can sometimes increase performance, if it is set too high, you run an increased risk of overdriving the SAN.

As data travels between the application and the storage array, there are several places that the queue depth can be set to throttle the number of concurrent disk transactions.

The most common places to control queue depth are:

- The application itself
- The virtual SCSI card driver in the guest
- The VMFS layer
- The HBA driver
- The HBA BIOS

The following sections explain how the queue depth is set in each of the layers in the event you need to change it.

Caution

The appropriate queue depth for a server may vary due to a number of factors, so it is recommended that you increase or decrease the queue depth only if necessary. See Appendix A for more info on determining the proper queue depth.

Host Bus Adapter Queue Depth

When configuring the host bus adapter for the first time, as mentioned previously, the queue depth should be set to 255. This is because the driver module loaded for each HBA in the system ultimately regulates the HBA's queue depth. For example, if the HBA BIOS is set to 255 and the driver module is set to 32, the maximum queue depth for that card or port is going to be 32.
Modifying ESX Storage Driver Queue Depth and Timeouts

As mentioned in the previous section, the HBA driver module ultimately regulates the queue depth for the HBA if it needs to be changed. (See Appendix A for more information about determining the appropriate queue depth.)

In addition to setting the queue depth in the driver module, the disk timeouts must also be set within the same command. These timeouts need to be set in order for the ESX host to survive a Storage Center controller failover properly.

Please refer to the latest documentation for instructions on how to configure these settings located on VMware’s web site:

- VMware document: “Fibre Channel SAN Configuration Guide”
  - Section Title: “Adjust Queue Depth for a QLogic HBA”
  - Section Title: “Adjust Queue Depth for an Emulex HBA”
- VMware document: “iSCSI SAN Configuration Guide”
  - Section Title: “Setting Maximum Queue Depth for Software iSCSI”

Before executing these commands, please refer to the latest documentation from VMware listed above for any last minute additions or changes.

For each of these adapters, the method to set the driver queue depth and timeouts uses the following general steps:

1) First, find the appropriate driver name for the module that is loaded:
   a. `vmkload_mod -l |grep “qla\|lpf”`
      i. Depending on the HBA model, it could be similar to:
         1. QLogic: `qla2xxx`
         2. Emulex: `lpfcdd_7xx`

2) Next, set the driver queue depth and timeouts using the `esxcfg-module` command:
   a. `esxcfg-module -s "param=value param2=value..." <driver_name>`
   b. Where:
      i. QLogic Parameters: “ql2xmaxqdepth=255
         ql2xloginretrycount=60
         qlport_down_retry=60”
      ii. Emulex Parameters: ”lpfc_devloss_tmo=60
         lpfc_hba_queue_depth=255”
   iii. Driver_name: As found in Step 1 (e.g. qla2xxx)

3) Next, you need to update the boot config:
   a. Both: # `esxcfg-boot -b`

4) Finally, you need to reboot the ESX host for these changes to take effect.

Similarly, for the software iSCSI Initiator:

1) Example of setting the queue depth to 128:
   a. `esxcfg-module -s iscsivmk_LunQDepth=128 iscsi_vmk`

2) Reboot the ESX host for the change to take effect.
Modifying the VMFS Queue Depth for Virtual Machines

Another setting which controls the queue depth at the virtual machine level is located in the ESX server’s advanced settings:

**Disk.SchedNumReqOutstanding** (Default=32)

This value can be increased or decreased depending on how many virtual machines are to be placed on each datastore. Keep in mind, this queue depth limit is only enforced when more than one virtual machine is active on that datastore.

For example, if left at default, the first virtual machine active on a datastore will have its queue depth limited only by the queue depth of the storage adapter. When a second, third, or fourth virtual machine is added to the datastore, the limit will be enforced to the maximum 32 queue depth or as set by the Disk.SchedNumReqOutstanding variable.

It is important to remember that this is a **global setting**, so it applies to ALL VMFS datastores with more than one virtual machine active on them. So if you have a datastore with 2 virtual machines, and another datastore with 8 virtual machines, each of the virtual machines will have a maximum queue depth of 32 enforced by default.

We recommend keeping this variable set at the default value of 32 unless your virtual machines have higher than normal performance requirements. (See Appendix A for more information about determining the appropriate queue depth.)

The Disk.SchedNumReqOutstanding limit does not apply to LUNs mapped as Raw Device Mappings (RDMs).
More information on the Disk.SchedNumReqOutstanding variable can be found in the following documents:

- VMware document: “Fibre Channel SAN Configuration Guide”
  - Section Title: “Equalize Disk Access Between Virtual Machines”
- VMware document: “iSCSI SAN Configuration Guide”
  - Section Title: “Equalize Disk Access Between Virtual Machines”

Modifying the Guest OS Queue Depth

The queue depth can also be set within the guest operating system if needed. By default, the Windows operating systems have a default queue depth of 32 set for each vSCSI controller, but can be increased up to 128 if necessary.

The method to adjust the queue depth varies between operating systems, but here are two examples.

**Windows Server 2003 (32 bit)**

The default LSI Logic driver (SYMMPI) is an older LSI driver that must be updated to get the queue depth higher than 32.

1) First, download the following driver from the LSI Logic download page:
   a. Adapter: LSI20320-R
   c. Version: WHQL 1.20.18 (Dated: 13-JUN-05)
   d. Filename: LSI_U320_W2003_IT_MID1011438.zip
2) Update the current “LSI Logic PCI-X Ultra320 SCSI HBA” driver to the newer WHQL driver version 1.20.18.
3) Using regedit, add the following keys: (Backup your registry first)

   [HKLM\SYSTEM\CurrentControlSet\Services\symmpi\Parameters\Device]
   "DriverParameter"="MaximumTargetQueueDepth=128;" (semicolon required)
   "MaximumTargetQueueDepth"=dword:00000080  (80 hex = 128 decimal)
4) Reboot the virtual machine.

**Windows Server 2008**

Since the default LSI Logic driver is already at an acceptable version, all you need to do is add the following registry keys.

1) Using regedit, add the following keys: (Backup your registry first)

   **For LSI Logic Parallel: (LSI_SCSI)**

   [HKLM\SYSTEM\CurrentControlSet\Services\LSI_SCSI\Parameters\Device]
   "DriverParameter"="MaximumTargetQueueDepth=128;" (semicolon required)
   "MaximumTargetQueueDepth"=dword:00000080  (80 hex = 128 decimal)

   **For LSI Logic SAS: (LSI_SAS)**

   [HKLM\SYSTEM\CurrentControlSet\Services\LSI_SAS\Parameters\Device]
   "DriverParameter"="MaximumTargetQueueDepth=128;" (semicolon required)
   "MaximumTargetQueueDepth"=dword:00000080  (80 hex = 128 decimal)
2) Reboot the virtual machine.

*Note*

*Please visit VMware’s Knowledge Base for the most current information about setting the queue depth with different vSCSI controllers or operating systems.*
For each operating system running within a virtual machine, the disk timeouts must also be set so the operating system can handle storage controller failovers properly.

Examples of how to set the operating system timeouts can be found in the following VMware documents:

- VMware document: "Fiber Channel SAN Configuration Guide"
  - Section Title: "Set Operating System Timeout"
- VMware document: "iSCSI SAN Configuration Guide"
  - Section Title: "Set Operating System Timeout"

Here are the general steps to setting the disk timeout within Windows and Linux:

**Windows**

1) Using the registry editor, modify the following key: (Backup your registry first)

[HKLM\SYSTEM\CurrentControlSet\Services\Disk]
"TimeOutValue"=dword:0000003c   (3c hex = 60 seconds in decimal)

2) Reboot the virtual machine.

**Linux**

For more information about setting disk timeouts in Linux, please refer to the following VMware Knowledge Base article:

- Increasing the disk timeout values for a Linux virtual machine
  - [http://kb.vmware.com/kb/1009465](http://kb.vmware.com/kb/1009465)
Guest Virtual SCSI Adapters

When creating a new virtual machine there are four types of virtual SCSI Controllers you can select depending on the guest operating system selection.

**BusLogic Parallel**
This vSCSI controller is used for certain older operating systems. Due to this controller’s queue depth limitations, it is not recommended you select it unless that is the only option available to your operating system. This is because when using certain versions of Windows, the OS issues only enough I/O to fill a queue depth of one.

**LSI Logic Parallel**
Since this vSCSI adapter is supported by many operating system versions, and is a good overall choice, it is recommended for virtual machines with hardware version 4. By default its queue depth is set to 32, but can be increased up to 128 if needed.

**LSI Logic SAS**
This vSCSI controller is available for virtual machines with hardware version 7, and has similar performance characteristics of the LSI Logic Parallel. This adapter is required for MSCS Clustering in Windows Server 2008 because SCSI3 reservations are needed.

**VMware Paravirtual**
This vSCSI controller is a high-performance adapter that can result in greater throughput and lower CPU utilization. Due to feature limitations when using this adapter, we recommend against using it unless the virtual machine has very specific performance needs. More information about the limitations of this adapter can be found in the “vSphere Basic System Administration” guide, in a section titled, “About Paravirtualized SCSI Adapters”.
Mapping Volumes to an ESX Server

Basic Volume Mapping Concepts

When sharing volumes between ESX hosts for such tasks as VMotion, HA, and DRS, it is important that each volume is mapped to each ESX server using the same Logical Unit Number (LUN).

*For example:*
You have three ESX servers named ESX1, ESX2, and ESX3.
You create a new volume named "LUN10-vm-storage".

This volume must be mapped to each of the ESX servers as the same LUN:

- Volume: "LUN10-vm-storage" → Mapped to ESX1 -as- LUN 10
- Volume: "LUN10-vm-storage" → Mapped to ESX2 -as- LUN 10
- Volume: "LUN10-vm-storage" → Mapped to ESX3 -as- LUN 10

Basic Volume Mapping in Storage Center 4.x and earlier

In Storage Center versions 4.x and earlier, each mapping must be created separately, each time specifying the same LUN, for each individual ESX host.

Basic Volume Mappings in Storage Center 5.x and later

However in Storage Center versions 5.x and higher, the mapping process is greatly automated by creating a server cluster object. This will allow you to map a volume to multiple ESX hosts at the same time, automatically keeping the LUN numbering consistent for all the paths.

As an added benefit, when a new ESX host is placed into the server cluster, all of the existing volume mappings assigned to the server cluster will be applied to the new host. This means that if the cluster has 100 volumes mapped to it, presenting all of them to a newly created ESX host is as simple as adding it to the cluster object.

Similarly, if you remove a host from the server cluster, the cluster mappings will also be removed, so it is important that those volumes are not being used by the host when you remove it.

Only volumes that are mapped to an individual host, such as the boot volume, will remain once a host is removed from the server cluster.
Also in Storage Center versions 5.x and higher, you can let the system auto select
the LUN number, or you can manually specify a preferred LUN number from the
advanced settings screen in the mapping wizard.

This advanced option will allow administrators who already have a LUN numbering
scheme to continue doing so, but if a LUN is not manually specified, the system will
auto select a LUN for each volume incrementally starting at LUN 1.

When naming volumes from within the Compellent GUI, it may be helpful to
specify the LUN number as part of the volume name. This will help you
quickly identify which volume is mapped using each LUN.

Multi-Pathed Volume Concepts

If you have an ESX server (or servers) that have multiple ports, whether it is FC,
iSCSI, or ethernet, ESX has built in functionality to provide native multi-pathing of
volumes over fiber channel, hardware iSCSI, or software iSCSI. Please note that
even when multi-pathing, the LUN must still remain consistent between paths.

Building on the example from above, here is an example of multi-pathing mappings:

Volume: "LUN10-vm-storage" → Mapped to ESX1/HBA1 -as- LUN 10
Volume: "LUN10-vm-storage" → Mapped to ESX1/HBA2 -as- LUN 10
Volume: "LUN10-vm-storage" → Mapped to ESX2/HBA1 -as- LUN 10
Volume: "LUN10-vm-storage" → Mapped to ESX2/HBA2 -as- LUN 10
Volume: "LUN10-vm-storage" → Mapped to ESX3/HBA1 -as- LUN 10
Volume: "LUN10-vm-storage" → Mapped to ESX3/HBA2 -as- LUN 10

If the LUN number does not remain consistent between multiple hosts or
multiple HBA's, VMFS datastores may not be visible to all nodes,
preventing use of VMotion, HA, DRS, or FT.

Keep in mind that when a volume uses multiple paths, the first ESX initiator in each
server will need to be mapped to one front end port, while the second ESX initiator
will be mapped to the other front end port in that same controller. For example:

"LUN10-vm-storage" → Controller1/PrimaryPort1 → FC-Switch-1 → Mapped to ESX1/HBA1 as LUN 10
"LUN10-vm-storage" → Controller1/PrimaryPort2 → FC-Switch-2 → Mapped to ESX1/HBA2 as LUN 10

Likewise, if different volume is active on the second Compellent controller, it may be
mapped such as:

"LUN20-vm-storage" → Controller2/PrimaryPort1 → FC-Switch-1 → Mapped to ESX1/HBA1 as LUN 20
"LUN20-vm-storage" → Controller2/PrimaryPort2 → FC-Switch-2 → Mapped to ESX1/HBA2 as LUN 20
This means that when configuring multi-pathing in ESX, you cannot map a single volume to both controllers at the same time, because a volume can only be active on one controller at a time.

**Multi-Pathed Volumes in Storage Center 4.x and earlier**

In Storage Center versions 4.x and earlier, when mapping a volume to a host, if it has multiple HBAs assigned in the server object, the mapping wizard will allow you to select both paths. When the volume is presented down multiple paths using the same LUN number, ESX’s native multi-pathing module will automatically detect and use both paths.

*Note: Before beginning, with certain versions of Storage Center, you may need to enable multi-pathing from within the Storage Center GUI. From within the system properties, under the "mapping" section, check the box labeled, "Allow volumes to be mapped to multiple fault domains", then click OK.*

Below is an example of how two volumes mapped from two separate controllers to a single ESX host should look when finished.

**Screenshot: Example of Multi-pathing mappings for ESX1**

<table>
<thead>
<tr>
<th>Status</th>
<th>Volume</th>
<th>Type</th>
<th>Server Port</th>
<th>Controller Port</th>
<th>LUN</th>
<th>Read Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up</td>
<td>LUN10-vm-storage</td>
<td>FC</td>
<td>210000003B930A46</td>
<td>5000D01000000002</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>Up</td>
<td>LUN10-vm-storage</td>
<td>FC</td>
<td>210000003B2017114</td>
<td>5000D01000000001</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>Up</td>
<td>LUN20-vm-storage</td>
<td>FC</td>
<td>210000003B930A46</td>
<td>5000D01000000002</td>
<td>20</td>
<td>No</td>
</tr>
<tr>
<td>Up</td>
<td>LUN20-vm-storage</td>
<td>FC</td>
<td>210000003B2017114</td>
<td>5000D01000000009</td>
<td>20</td>
<td>No</td>
</tr>
</tbody>
</table>

**Multi-Pathed Volumes in Storage Center 5.x and later**

When multi-pathing volumes in Storage Center versions 5.x and later, much of the process is automated. You can prevent many of the common mapping errors simply by selecting the operating system in the server properties screen. Based on the OS selected, it will apply a set of rules to the server, unique to each operating system, to correctly map volumes.

Multipathing to an ESX host is automatic if the server object has multiple HBA’s or iSCSI initiator ports assigned to it. In other words, you will have to use the advanced options if you don’t want to multipath a volume.
From the advanced mapping page, here are some of the options to mapping to an ESX 4.x host.

Select LUN: This option is to manually specify the LUN. If you do not check this box, it will automatically pick a LUN for you.

Restrict Mapping paths: This option is used when you need to only map a volume to a specific HBA in the ESX host.

Map to Controller: By default, the system will automatically select which controller the volume should be mapped. If you would like to force a particular controller to handle the I/O, use this option to do so.

Configure Multipathing: This option designates how many of the Storage Center FE ports you will allow the volume to be mapped through. For example if each controller has 4 Front End ports, selecting unlimited will map the volume through all 4, whereas selecting 2 will only use 2 of the 4 front end ports. The system will automatically select the 2 front end ports based on which already have the fewest mappings.
Configuring the VMware iSCSI software initiator for a single path

Mapping volumes via VMware's iSCSI initiator follows the same rules for LUN numbering as with fibre channel, but there are a few extra steps required for ESX to see the Compellent via the ESX software initiator.

From within the VMware vSphere Client:
1) Enable the "Software iSCSI Client" within the ESX firewall (located in the "Security Profile" of the ESX server)
2) Add a "VMKernel port" to a virtual switch assigned to the physical NIC you want to use for iSCSI (See screenshot below)
3) From within the Storage Adapters, highlight the iSCSI Software Adapter, click "Properties", then on the general tab, click "Configure" to set the status to "Enabled".
4) Under the “Dynamic Discovery” tab, add all of the Compellent iSCSI IP addresses that are assigned to the Compellent iSCSI cards in your controller(s), or just the iSCSI Control Port IP address.
5) Rescan the iSCSI Initiator.

From Within the Compellent GUI:
6) Create a server object for the ESX server using the IP Address you specified for the VMKernel in step 2 above
7) Map a volume to the ESX server

From within the VMware vSphere Client:
8) Navigate to the Storage Adapters section, and rescan the iSCSI HBA for new LUN’s.

Screenshot: Configuring the VMKernel port
A new feature in ESX 4.x is the ability to enable multipathing to storage using the VMware iSCSI software initiator. Instructions on how to configure this can be found in the following document:

- VMware document: “iSCSI SAN Configuration Guide”
  - Section Title: “Setting Up Software iSCSI Initiators”
    - Subsection: “Set Up Multipathing for Software iSCSI”

After following the instructions on how to configure the software iSCSI initiator to use both NICs for multipathing, you can then add the ESX host to the Storage Center.

Since the ESX software iSCSI adapter appears differently to the Storage Center than other software iSCSI initiators, there is an additional step to making sure both paths are added correctly.

**Screenshot: Adding iSCSI HBAs to a server without using iSCSI Names**

When adding the iSCSI Initiators to the server object in the Compellent GUI, it is recommended that you **uncheck** the “Use iSCSI Names” so that each initiator can be added and configured independently.

If the ESX Software Initiator is added with its iSCSI name, it will still multipath the volumes mapped through it, however you will lose the ability to map a volume to a single path if desired.

**VMware Multi-Pathing Policies**

When configuring the path selection policy of each datastore or LUN, you have the option to set it to Fixed, Round Robin, or Most Recently Used. The default path selection policy for the Storage Center is set to Fixed, but you can use Round Robin as well.

**Fixed Policy**

If you use the fixed policy, it will give you the greatest control over the flow of storage traffic. However, you must be careful to evenly distribute the load across all host HBAs, Front-End Ports, and Storage Center controllers.
When using the fixed policy, if a path fails, all of the LUNs using it as their preferred path will fail over to the secondary path. When service resumes, the LUNs will resume I/O on their preferred path.

*Fixed Example: (See screenshot below)*

HBA1 loses connectivity; HBA2 takes over its connections. HBA1 resumes connectivity; HBA2 will fail its connections back to HBA1.

*Screenshot: Example of a datastore path selection policy set to Fixed*

---

**Round Robin**

The round robin path selection policy uses automatic path selection and load balancing to rotate I/O through all available paths. It is important to note that round robin load balancing does not aggregate the storage link bandwidth; it merely distributes the load across adapters.

Using round robin will reduce the management headaches of manually balancing the storage load across all storage paths as you would with a fixed policy; however there are certain situations where using round robin does not make sense.

For instance, it is generally not considered good practice to enable round robin between an iSCSI path and fiber channel path, nor enabling it to balance the load between a 2GB FC and a 4GB FC path.

If you chose to enable round robin for one or more datastores/LUNs, you should be careful to ensure all the paths included are identical in type, speed, and have the same queue depth setting.

Here is an example of what happens during a path failure using round robin.

*Round Robin Example: (See screenshot below)*

Load is distributed evenly between HBA1 and HBA2. HBA1 loses connectivity; HBA2 will assume all I/O load. HBA1 resumes connectivity; load is distributed evenly again between both.
The round robin path selection policy (PSP) can be set to the default with the following command. After setting round robin as the default and rebooting, any new volumes mapped will acquire this policy, however, mappings that already existed beforehand will have to be set manually.

```
# esxcli nmp satp setdefaultpsp --psp VMW_PSP_RR --satp VMW_SATP_DEFAULT_AA
```

The round robin path selection policy should not be used for volumes belonging to guests running Microsoft Clustering Services.

**Most Recently Used (MRU)**

The Most Recently Used path selection policy is generally reserved for Active/Passive arrays (to prevent path thrashing), and is therefore not needed with the Storage Center because a volume is only active on one controller at a time.

**Multi-Pathing using a Fixed path selection policy**

Keep in mind with a fixed policy, only the preferred path will actively transfer data. To distribute the I/O loads for multiple datastores over multiple HBA’s, you can do this by setting the preferred path for each datastore. Here are some examples:

**Example 1: (Bad)**

Volume: "LUN10-vm-storage" → Mapped to ESX1/HBA1 -as- LUN 10 (Active/Preferred)
Volume: "LUN10-vm-storage" → Mapped to ESX1/HBA2 -as- LUN 10 (Standby)

Volume: "LUN20-vm-storage" → Mapped to ESX1/HBA1 -as- LUN 20 (Active/Preferred)
Volume: "LUN20-vm-storage" → Mapped to ESX1/HBA2 -as- LUN 20 (Standby)

This example would cause all I/O for both volumes to be transferred over HBA1.
Example 2: (Good)
Volume: "LUN10-vm-storage" → Mapped to ESX1/HBA1 -as- LUN 10 (Active/Preferred)
Volume: "LUN10-vm-storage" → Mapped to ESX1/HBA2 -as- LUN 10 (Standby)
Volume: "LUN20-vm-storage" → Mapped to ESX1/HBA1 -as- LUN 20 (Standby)
Volume: "LUN20-vm-storage" → Mapped to ESX1/HBA2 -as- LUN 20 (Active/Preferred)

This example sets the preferred path to more evenly distribute the load between both HBAs.

Although the fixed multi-pathing policy gives greater control over which path transfers the data for each datastore, you must manually validate that all paths have proportional amounts of traffic on each ESX host.

Multi-Pathing using a Round Robin path selection policy

If you decide to use round robin, it must be manually defined for each LUN (or set to the default), but will provide both path failure protection, and remove some of the guesswork of distributing load between paths manually as you would with a fixed policy. To reiterate from previous sections in this document, be sure when using round robin that the paths are of the same type, speed, and have the same queue depth setting.

Example 1:
Volume: "LUN10-vm-storage" → Mapped to ESX1/HBA1 -as- LUN 10 (Active)
Volume: "LUN10-vm-storage" → Mapped to ESX1/HBA2 -as- LUN 10 (Active)
Volume: "LUN20-vm-storage" → Mapped to ESX1/HBA1 -as- LUN 20 (Active)
Volume: "LUN20-vm-storage" → Mapped to ESX1/HBA2 -as- LUN 20 (Active)

Asymmetric Logical Unit Access (ALUA)

The ALUA protocol was designed for more traditional arrays that VMware classifies as Asymmetrical Storage Systems. Since the Storage Center is considered an Active-active storage system where all the paths are active at all times (unless a path fails), ALUA is not necessary.

Additional Multi-pathing resources

- VMware Document: "Fiber Channel SAN Configuration Guide"
- VMware Document: "iSCSI SAN Configuration Guide"
There is an ongoing discussion about whether or not to boot your ESX servers from SAN. In some cases, such as with blade servers that do not have internal disk drives, booting from SAN is the only option, but a lot of ESX servers can have internal mirrored drives giving you the flexibility to choose.

The benefits of booting from SAN are obvious. It alleviates the need for internal drives and allows you the ability to take replays of the boot volume.

However there are also benefits to booting from Local disks and having the virtual machines located on SAN resources. Since it only takes about 15-30 minutes to freshly load and patch an ESX server, booting from local disks gives them the advantage of staying online if for some reason you need to do maintenance to your fibre channel switches, ethernet switches, or the controllers themselves. The other clear advantage of booting from local disks is being able to use the VMware iSCSI software initiator instead of iSCSI HBAs or fibre channel cards.

In previous versions of ESX, if you booted from SAN you couldn't use RDM's, however since 3.x this behavior has changed. If you decide to boot from SAN with ESX 3.x or 4.x you can also utilize RDM's.

Since the decision to boot from SAN depends on many business related factors including cost, recoverability, and configuration needs, we have no specific recommendation.

**Configuring boot from SAN**

If you decide to boot your ESX server from SAN, there are a few best practices you need to consider.

When mapping the boot volume to the ESX server for the initial install, the boot volume should only be mapped down a single path to a single HBA. Once ESX has been loaded and operating correctly, you can then add the second path for the boot volume.

In Storage Center 4.x and earlier, when initially mapping the boot volume to the ESX host, the mapping wizard will allow you to select the individual paths, making sure to specify LUN 0. You then use the same procedure to add the second path once the host is up and running, being sure to rescan the host HBAs once the path has been added.
In Storage Center 5.x and later, you will need to enter the advanced mapping screen to select a few options to force mapping down a single path.

- Check: Map volume using LUN 0
- Check: Only map using specified server ports
  - Select the HBA that is selected to boot from within the HBA BIOS
- Maximum number of paths allowed: Single-path

Once the ESX host is up and is running correctly, you can then add the second path to the boot volume by modifying the mapping. To do this, right click on the mapping and select, “Modify Mapping”.

- Uncheck: Only map using specified server ports
- Maximum number of paths allowed: Unlimited

Once the 2nd path has been added, you can then rescan the HBAs on the ESX host.
Volume Sizing and the 2 TB Limit

Although the maximum size of a LUN that can be presented to ESX is 2 TB\(^1\), the general recommendation is to create your datastore sized in the 500GB – 750GB range. A 750 GB datastore will accommodate approximately 15 40GB virtual disks, leaving a small amount of overhead for virtual machine configuration files, logs, snapshots, and memory swap.

\(^1\) According to VMware, the maximum size of a single LUN that can be presented to an ESX server is 2 TB (minus 512 B). Because this size is just short of a full 2 TB, the maximum volume size that you can specify in the Storage Center GUI is either 2047 GB or 1.99 TB.

If you create or extend a VMFS volume beyond the 2047GB/1.99TB limit, that volume will become inaccessible by the ESX host. If this happens, the most likely scenario will result in recovering data from a replay or tape.

Virtual Machines per Datastore

Although there are no steadfast rules for how many virtual machines you should place on a datastore, the general consensus in the VMware community is to place anywhere between 10-20 virtual machines on each.

The reasoning behind keeping a limited number of Virtual Machines and/or VMDK files per datastore is due to potential I/O contention, queue depth contention, or SCSI reservation errors that may degrade system performance. That is also the reasoning behind creating 500GB – 750GB datastores, because this helps limit the number of virtual machines you place on each.

The art to virtual machine placement revolves highly around analyzing the typical disk I/O patterns for each of the virtual machines and placing them accordingly. In other words, the “sweet spot” of how many virtual machines you put on each datastore is greatly influenced by the disk load of each. For example, in some cases the appropriate number for high I/O load virtual machines may be less than 5, while the number of virtual machines with low I/O disk requirements may be up to 20.

Since the appropriate number of virtual machines you can put onto each datastore is subjective and dependent on your environment, a good recommendation is to start with 10 virtual machines, and increase/decrease the number of virtual machines on each datastore as needed.

The most common indicator that a datastore has too many virtual machines placed on it would be the frequent occurrence of “SCSI Reservation Errors” in the vmkwarning log file. That said, it is normal to see a few of these entries in the log from time to time, but when you notice them happening very frequently, it may be time to move some of the virtual machines to a new datastore of their own. Moving
virtual machines between datastores can even be done non-disruptively if you are licensed to use VMware’s Storage vMotion.

The second most common indicator that the datastore has too many virtual machines placed on it is if the queue depth of the datastore is regularly exceeding the limit set in the driver module. Remember that if the driver module is set to a 256 queue depth, the maximum queue depth of each datastore is also 256. This means that if you have 16 virtual machines on a datastore all heavily driving a 32 queue depth (16 * 32 = 512), they are essentially overdriving the disk queues by double, and the resulting high latency will most likely degrade performance. (See Appendix A for more information on determining if the queue depth of a datastore is being correctly utilized.)

There are many resources available that discuss VMware infrastructure design and sizing, so this should only be used as a general rule of thumb, and may vary based upon the needs of your environment.

VMFS Partition Alignment

Partition alignment is a performance tuning technique used with traditional SANs to align the guest operating system and VMFS partitions to the physical media, in turn reducing the number of disk transactions it takes to process an I/O.

Due to how Dynamic Block Architecture virtualizes the blocks, manual partition alignment is generally not necessary. This is because the Storage Center automatically aligns its 512K, 2M, or 4M pages to the physical sector boundaries of the drives. Since the largest percentage of performance gains are seen from aligning the Storage Center pages to the physical disk, the remaining areas that can be aligned and tuned have a minor effect on performance.

Based on internal lab testing, we have found that any performance gains achieved by manually aligning partitions are usually not substantial enough (±1%) to justify the extra effort. However, before deciding whether or not to align VMFS partitions, it is recommended that you perform testing to determine the impact that an aligned partition may have on your particular application because all workloads are different.

To manually align the VMFS block boundaries to the Storage Center page boundaries for your own performance testing, the recommended offset when creating a new datastore is 8192 (or 4 MB).

Using the Compellent Storage Center Integrations for VMware vSphere (Client Plug-in) to create new datastores will automatically align them to the recommended 4 MB offset.

Please consult the following VMware documentation for more info:
- Document: “Recommendations for Aligning VMFS Partitions”
This is an example of a fully aligned partition in the Storage Center, where one guest I/O will only access necessary physical disk sectors:

![Fully aligned partition diagram]

This is an example of an unaligned partition in a traditional SAN where performance can be improved by alignment:

![Unaligned partition diagram]

**VMFS block sizes**

Choosing a block size for a datastore determines the maximum size of a VMDK file that can be placed on it.

<table>
<thead>
<tr>
<th>Block Size</th>
<th>Maximum VMDK Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 MB</td>
<td>256 GB</td>
</tr>
<tr>
<td>2 MB</td>
<td>512 GB</td>
</tr>
<tr>
<td>4 MB</td>
<td>1024 GB</td>
</tr>
<tr>
<td>8 MB</td>
<td>2048 GB</td>
</tr>
</tbody>
</table>

In other words, you should choose your block size based on the largest virtual disk you plan to put on the datastore.

The default block size is 1 MB, so if you need your virtual disks to be sized greater than 256 GB, you will need to increase this value. For example, if the largest virtual disk you need to place on a datastore is 200 GB, then a 1 MB block size should be
sufficient, and similarly, if you have a virtual machine that will require a 400 GB virtual disk, then the 2 MB block size should be sufficient.

You should also consider future growth of the virtual machine disks when choosing the block size. If a virtual machine resides on a datastore formatted with a 1 MB block size, and in the future it needs one of its virtual disks extended beyond 256 GB, the virtual machine would have to be relocated to a different datastore with a larger block size. This is because a datastore must be re-formatted to change the block size.

*Since certain VAAI offload operations require that the source and destination datastores have the same VMFS block size, it is worth considering a standard block size for all of your datastores. Please consult the vStorage APIs for Array Integration FAQ for more information.*
Multiple Virtual Machines per LUN

One of the most common techniques in virtualization is to place more than one virtual machine on each volume. This allows for the encapsulation of virtual machines, and thus higher consolidation ratios.

When deciding how to layout your VMFS volumes and virtual disks, as discussed earlier, it should reflect the performance needs as well as application and backup needs of the guest operating systems.

Regardless of how you decide to layout your virtual machines, here are some basic concepts you should consider:

Storage of non-virtual machine files
As a general recommendation, you should create one or more VMFS datastores for administrative items. You can use these to store all of your virtual machine templates, ISO images, virtual floppies, and/or scripts.

Separation of the operating system pagefiles
One technique to consider with virtual machine placement is separating the operating system pagefile/swap files onto a separate datastore.

There are two main reasons for separating operating system pagefiles onto their own volume/datastore.

- Since pagefiles can generate a lot of disk activity when the memory in the virtual machine or ESX host runs low, it could keep volume replays smaller.
- If you are replicating those volumes, it will conserve bandwidth by not replicating the operating system pagefile data

Depending on the memory swap conditions unique to each environment, separating pagefiles may or may not make a significant reduction in replay sizes. A good way to determine whether or not separating pagefiles will make a difference in your environment is to use the vSphere client performance charts to monitor Swap/Balloon usage of the ESX host. If these numbers are high, you should consider testing the separation of pagefiles to determine the actual impact.

If you decide that separating pagefiles will make an impact in reducing replay sizes, the general recommendation is to create “pairs” of volumes for each datastore containing virtual machines. If you create a volume that will contain 10 virtual machines, you should to create a second volume to store the operating system pagefiles for those 10 machines.
For example:

- Create one datastore for Virtual Machines
  - This will usually contain the virtual disks (vmdk files), configuration files, and logs for your virtual machines.
- Create one “paired” datastore for the corresponding virtual machine pagefiles
  - This should contain virtual machine pagefiles. Using Windows as an example, you would create a 2GB - 16GB virtual disk (P:) on this volume to store the Windows paging file for each virtual machine.
  - This volume can be sized considerably smaller than the “main datastore” as it only needs enough space to store pagefiles.

Often the question is asked whether or not it is a good idea to place all of the operating system pagefiles on a single datastore. Generally speaking, this is not a very good practice for a couple of reasons.

First, the pagefile datastore can also experience contention from queue depth utilization or disk I/O; so too many vmdk files during a sudden memory swapping event could decrease performance even further. For example, if a node in the ESX HA cluster fails, and the effected virtual machines are consolidated on the remaining hosts. The sudden reduction in overall memory could cause a sudden increase in paging activity that could overload the datastore causing a performance decrease.

Second, it becomes a matter of how many eggs you want in each basket. Operating systems are usually not tolerant of disk drives being unexpectedly removed. If an administrator were to accidentally unmap the pagefile volume, the number of virtual machines affected would be isolated to a handful of virtual machines instead of all the virtual machines.

**Separation of the virtual machine swap files**

Each virtual machine also has a memory swap file located in its home directory which is used by the ESX host when the VMware Tools balloon driver was unable to reclaim enough memory. In other words, the vswp file is generally only used as a last resort by the ESX host to reclaim memory. VMware recommends keeping the vswp files located in the virtual machine home directories, however if needed, it is also possible to relocate the .vswp file to a dedicated LUN. Doing this can also help to reduce replay sizes and preserve replication bandwidth, but should only be done under the guidance of VMware support.
Virtual Machine Placement

This example technique will give you a great deal of flexibility when building out your storage architecture, while keeping with the basic concepts discussed above. The example layout below will meet most virtual infrastructure needs, because it adds the flexibility of being able to add RDM's to virtual machines later if needed. The key to this technique is reserving LUN numbers in the middle of the LUN sequence to help better organize your virtual machines.

An example of this technique is as follows:

- **LUN0**: Boot LUN for ESX (When booting from SAN)
- **LUN1**: Templates/ISO/General Storage
- **LUN10**: OS/DATA (C:/D:/E: Drives)
- **LUN11**: Pagefile (Paired with LUN10) for VM pagefiles (P: Drives)
- **LUN12-LUN19**: Reserved LUNs for virtual machine RDM's for machines in this group
- **LUN20**: OS/DATA (C:/D:/E: Drives)
- **LUN21**: Pagefile (Paired with LUN20) for VM pagefiles (P: Drives)
- **LUN22-LUN29**: Reserved LUNs for virtual machine RDM's for machines in this group

Virtual Machine Placement (With RDMs)

To help organize the LUN layout for your ESX clusters, some administrators prefer to store their layout in a spreadsheet. Not only does this help to design your LUN layout in advance, but it also helps you keep things straight as the clusters grow larger.

There are many factors that may influence architecting storage with respect to the placement of virtual machines. The method shown above is merely a suggestion, as your business needs may dictate different alternatives.
One Virtual Machine per LUN

Although creating one volume for each virtual machine is not a very common technique, there are both advantages and disadvantages that will be discussed below. Keep in mind that deciding to use this technique should be based on factors unique to your business, and may not be appropriate for all circumstances.

Advantages

- Granularity in replication
  - Since the Storage Center replicates at the volume level, if you have one virtual machine per volume, you can pick and choose which virtual machine to replicate.
- There is no I/O contention as a single LUN is dedicated to a virtual machine.
- Flexibility with volume mappings.
  - Since a path can be individually assigned to each LUN, this could allow a virtual machine a specific path to a controller.
- Statistical Reporting
  - You will be able to monitor storage usage and performance for an individual virtual machine.
- Backup/Restore of an entire virtual machine is simplified
  - If a VM needs to be restored, you can just unmap/remap a replay in its place.

Disadvantages

- You will have a maximum of 256 virtual machines in your ESX cluster.
  - The HBA has a maximum limit of 256 LUNs that can be mapped to the ESX server, and since we can only use each LUN number once when mapping across multiple ESX servers, it would essentially have a 256 virtual machine limit.
- Increased administrative overhead
  - Managing a LUN for each virtual machine and all the corresponding mappings may get challenging.
Raw Device Mappings (RDM's) are used to map a particular LUN directly to a virtual machine. When an RDM, set to physical compatibility mode, is mapped to a virtual machine, the operating system writes directly to the volume bypassing the VMFS file system.

There are several distinct advantages and disadvantages to using RDM's, but in most cases, using the VMFS datastores will meet most virtual machines needs.

Advantages of RDM's:
- Ability to create a clustered resource (i.e. Microsoft Cluster Services)
  - Virtual Machine to Virtual Machine
  - Virtual Machine to Physical Machine
- The volume can be remapped to another physical server in the event of a disaster or recovery.
- Ability to convert physical machines to virtual machines more easily
  - Physical machine volume can be mapped as an RDM.
- Can be used when a VM has special disk performance needs
  - There may be a slight disk performance increase when using an RDM versus a VMFS virtual disk due to the lack of contention, no VMFS write penalties, and better queue depth utilization.
- The ability to use certain types of SAN software
  - For example, the Storage Center's Space Recovery feature or Replay Manager.
    - More information about these features can be found in the Compellent Knowledge Center.
- The ability to assign a different data progression profile to each volume.
  - For example, if you have a database VM where the database and logs are separated onto different volumes, each can have a separate data progression profile.
- The ability to adding a different replay profile to each volume.
  - For example, a database and its transaction logs may have different replay intervals and retention periods for expiration.

Disadvantages of RDM's:
- Added administrative overhead due to the number of mappings
- There are a limited number of LUNs that can be mapped to an ESX server
  - If every virtual machine used RDM's for drives, you would have a maximum number of 255 drives across the cluster.
- Physical mode RDMs cannot use ESX snapshots
  - While ESX snapshots are not available for physical mode RDMs, Compellent Replays can still be used to recover data.
Just like a physical server attached to the Storage Center, Data Progression will migrate inactive data to the lower tier inexpensive storage while keeping the most active data on the highest tier fast storage. This works to the advantage of VMware because multiple virtual machines are usually kept on a single volume.

However, if you do encounter the business case where particular virtual machines would require different RAID types, some decisions on how you configure Data Progression on volumes must be made.

Here are some advanced examples of virtual machine RAID groupings:

**Example 1: Separating virtual machines based on RAID type**

- **LUN0** - Boot LUN for ESX  
  - Data Progression: Recommended (All Tiers)
- **LUN1** - Templates/ISO/General Storage  
  - Data Progression: Recommended (All Tiers)
- **LUN2** - OS/DATA (Server group1 - High performance - 4 VM's - C:/D:/E: Drives)  
  - High Priority (Tier 1)
- **LUN3** - Pagefile (Paired with LUN2) for VM pagefiles  
  - Data Progression: Recommended (All Tiers)
- **LUN4** - LUN8 - Reserved LUNs for virtual machine RDM's for machines in this group  
  - RAID types vary based on needs of the VM they are mapped to
- **LUN10** - OS/DATA (Server group2 - Low performance - 15 VM's - C:/D:/E: Drives)  
  - Data Progression: Low Priority (Tier 3)
- **LUN11** - Pagefile (Paired with LUN10) for VM pagefiles  
  - Data Progression: Recommended (All Tiers)
- **LUN12** - LUN19 - Reserved LUNs for virtual machine RDM's for machines in this group  
  - RAID types vary based on needs of the VM they are mapped to
- **LUN20** - OS/DATA (Server group 3 - Application grouping - 5 VM's - C:/D:/E: Drives)  
  - Data Progression: Recommended (All Tiers)
- **LUN21** - Pagefile (Paired with LUN20) for VM pagefiles  
  - Data Progression: Recommended (All Tiers)
- **LUN22** - LUN29 - Reserved LUNs for virtual machine RDM's for machines in this group  
  - RAID types vary based on needs of the VM they are mapped to

Like previously mentioned at the beginning of this section, unless you have a specific business need that requires a particular virtual machine or application to have a specific RAID type, our recommendation is to keep the configuration simple. In most cases, you can use the Data Progression “Recommended” setting, and let it sort out the virtual machine data automatically by usage.

**A note about Data Progression Best Practices:**  You should create a replay schedule for each volume that (at a minimum) takes one daily replay that doesn’t expire for 25 hours or more. This will have a dramatic effect on Data Progression behavior, which will increase the overall system performance.
Introduction

Compellent’s thin provisioning feature named “Dynamic Capacity” allows less storage to be consumed for virtual machines thus saving storage costs. The following section describes the relationship that this feature has with virtual machine storage.

Virtual Disk Formats

In ESX 4.x, VMFS can create virtual disks using one of three different formats.

**Thick**
(a.k.a. “zeroedthick”) [Default]

Only a small amount of disk space is used within the Storage Center at virtual disk creation time, and new blocks are only allocated on the Storage Center during write operations. However, before any new data is written to the virtual disk, ESX will first zero out the block, to ensure the integrity of the write. This zeroing of the block before the write induces extra I/O and an additional amount of write latency which could potentially affect applications that are sensitive to disk latency or performance.

**Thin Provisioned**

This virtual disk format is used when you select the option labeled “Allocate and commit space on demand”. The Logical space required for the virtual disk is not allocated during creation, but it is allocated on demand during first write issued to the block. Just like thick disks, this format will also zero out the block before writing data inducing extra I/O and an additional amount of write latency.
Eagerzeroedthick

This virtual disk format is used when you select the option labeled “Support clustering features such as Fault Tolerance”. Space required for the virtual disk is fully allocated at creation time. Unlike with the zeroedthick format, all of the data blocks within the virtual disk are zeroed out during creation. Disks in this format might take much longer to create than other types of disks because all of the blocks must be zeroed out before it can be used. This format is generally used for Microsoft clusters, and the highest I/O workload virtual machines because it does not suffer from the same write penalties as the zeroedthick or thin formats.

Thin Provisioning Relationship

The following points describe how each virtual disk format affects Storage Center’s thin provisioning.

- **Zeroedthick**
  - Virtual disks will be thin provisioned by the Storage Center

- **Thin**
  - Virtual disks will be thin provisioned by the Storage Center
  - There are no additional storage savings while using this format because the array already uses its thin provisioning (see below)

- **Eagerzeroedthick**
  - Depending on storage center version, this format may or may not pre-allocate storage for the virtual disk at creation time.
  - If you create a 20GB virtual disk in this format, the Storage Center will normally consume 20GB, with one exception. (See the “Storage Center Thin Write Functionality” section below.)

We recommend sticking with the default virtual disk format (zeroedthick) unless you have a specific need to pre-allocate virtual disk storage such as Microsoft clustering, VMware Fault Tolerance, or for virtual machines that may be impacted by the thin or zeroedthick write penalties.

Storage Center Thin Write Functionality

Certain versions of Storage Center (4.2.3+) have the ability to detect incoming sequential zeros while being written, track them, but not actually write the “zeroed page” to the physical disks.

When creating virtual disks on these versions of firmware, all virtual disk formats will be thin provisioned at the array level, including eagerzeroedthick.

Storage Center Thin Provisioning or VMware Thin Provisioning?

A common question is whether or not to use array based thin provisioning or VMware’s thin provisioning. Since the Storage Center uses thin provisioning on all volumes by default, it is not necessary to use VMware’s thin provisioning because there are no additional storage savings by doing so.

However, if you need to use VMware’s thin provisioning for whatever reason, you must pay careful attention not to accidentally overrun the storage allocated. To prevent any unfavorable situations, you should use the built-in vSphere datastore threshold alerting capabilities, to warn you before running out of space on a datastore.
Windows Free Space Recovery

One of the nuances of the Windows NTFS file system is that gradually over time, the actual usage of the file system can grow apart from what the Storage Center reports as being allocated. For example, if you have a 20 GB data volume and Windows writes 15 GB worth of files, followed by deleting 10 GB worth of those files. Although Windows reports only 5 GB of disk space in-use, Dynamic Capacity has allocated those blocks to that volume, so the Storage Center will still report 15 GB of data being used. This is because when Windows deletes a file, it merely removes the entry in the file allocation table, and there are no built-in mechanisms for the Storage Center to determine if an allocated block is actually still in use by the OS.

However, the “Compellent Enterprise Manager Server Agent” contains the necessary functionality to recover this free space from Windows machines. It does this by comparing the Windows file allocation table to the list of blocks allocated to the volume, and then returning those free blocks into the storage pool to be used elsewhere in the system. It is important to note though, blocks which are kept as part of a replay, cannot be freed until that replay is expired.

The free space recovery functionality can only be used in Windows virtual machines under the following circumstances:

- The virtual disk needs to be mapped as a Raw Device Mapping set to “physical” compatibility mode (RDMP).
  - This allows the free space recovery agent to perform a SCSI query of the physical LBAs in-use, and then correlate them to the blocks allocated on the Storage Center that can be freed.
  - The disk must be an NTFS basic disk (either MBR or GPT)
- The virtual disk cannot be a VMDK, or a Raw Device Mapping set to “virtual” compatibility mode (RDM).
  - This is because VMware does not provide the necessary API’s for the free space recovery agent to correlate the virtual LBAs to the actual physical LBAs needed to perform the space recovery.
  - If a virtual machine has a C: drive (VMDK) and a D: drive (RDMP), Windows free space recovery will only be able to reclaim space for the D: drive.
  - The restriction against using “virtual” mode RDMs for space recovery also implies that these disks cannot participate in ESX server snapshots.
    - This means, that if you intend to use VMware Consolidated Backup, you will have to apply an alternative method of backing up the physical mode RDMs. For example, the “Storage Center Command Set for Windows PowerShell” installation provides an example PowerShell script, which can be used to backup physical mode RDMs as part of the pre-execution steps of the backup job.
- The free space recovery agent will also work with volumes mapped directly to the virtual machine via the Microsoft Software iSCSI initiator.
  - Volumes mapped to the virtual machine through the Microsoft iSCSI initiator interact with the SAN directly, and thus, space recovery works as intended.

For more information on Windows free space recovery, please consult the “Compellent Enterprise Manager User Guide”.

© Compellent Technologies

Document 680-041-017

Page 39
Extending VMware Volumes

Within an ESX server, there are three ways in which you can extend or grow storage. The general steps are listed below, but if you need additional information, please consult the following documentation pages:

- VMware document: “ESX Configuration Guide”
  - Subsection: “Increase VMFS Datastores”
- VMware document: “vSphere Basic System Administration”
  - Subsection: “Change the Virtual Disk Configuration”

Growing VMFS Datastores

**Grow an extent in an existing VMFS datastore**

This functionality is used to grow an existing extent in a VMFS datastore, but can only be done if there is adjacent free capacity.

*Datastore2 and Datastore3 can be grown by 100GB, but Datastore1 cannot.*

To extend the space at the end of a Storage Center volume as shown above, you can do so from the Compellent GUI. After the volume has been extended, and the hosts HBA has been rescanned, you can then edit the properties of the datastore to grow it by clicking on the “Increase...” button, and then follow through the “Increase Datastore Capacity” wizard.

Be careful to select the volume that is “Expandable” otherwise you will be adding a VMFS “extent” to the datastore (see section below on VMFS extents).

**Screenshot from the wizard after extending a 500GB datastore by 100GB.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Path ID</th>
<th>LUN</th>
<th>Capacity</th>
<th>Expandable</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPELNT Fibre Channel Disk</td>
<td>vntha0:C0.T7.L70</td>
<td>70</td>
<td>600.00 GB</td>
<td>Yes</td>
</tr>
<tr>
<td>COMPELNT Fibre Channel Disk</td>
<td>vntha0:C0.T6.L61</td>
<td>61</td>
<td>500.00 GB</td>
<td>No</td>
</tr>
</tbody>
</table>

If you extend a VMFS volume (or RDM) beyond the 2047GB/1.99TB limit, that volume will become inaccessible by the ESX host. If this happens, the most likely scenario will result in recovering data from a replay or tape.
As an alternative to extending a datastore volume when a virtual machine needs additional disk space, consider creating a new datastore volume and migrating that virtual machine. This will help to keep volume sizes manageable, as well as help to keep any single datastore from being overloaded due to I/O contention.

**Relevant VMware Knowledge Base articles:**
- “Unable to grow or expand a VMFS volume or datastore”
  - [http://kb.vmware.com/kb/1017662](http://kb.vmware.com/kb/1017662)

### Adding a new extent to an existing datastore

This functionality is used to grow a datastore larger than 2 TB. Since each datastore can have up to 32 extents (each ~2 TB), this allows the maximum datastore size of up to 64 TB.

**Caution**

*Due to the complexities of coordinating replays and recoveries of datastores that are spanned across multiple Storage Center volumes, the use of VMFS extents is highly discouraged. However, if the use of extents is needed, Replays of those volumes should be taken using the Consistent Replay Profile functionality available in Storage Center versions 5.x and later.*

### Growing Virtual Disks and Raw Device Mappings

#### Extending a virtual disk (vmdk file)

Hot extending a virtual disk is available from within the vSphere client when editing the settings of a virtual machine (or by using vmkfstools from the ESX CLI).

*Screenshot: Growing a virtual disk from the virtual machine properties screen*

<table>
<thead>
<tr>
<th>Disk Provisioning</th>
<th>Type:</th>
<th>Thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provisioned Size:</td>
<td>8 GB</td>
<td></td>
</tr>
<tr>
<td>Maximum Size (GB):</td>
<td>256.00</td>
<td></td>
</tr>
</tbody>
</table>

For Windows machines: After growing the virtual disk from the vSphere client, you must log into the virtual machine, rescan disks from Windows disk management, and then use DISKPART to extend the drive.

**Caution**

*Microsoft does not support extending the system partition (C: drive) of a machine.*

#### Extending a Raw Device Mapping (RDM)

To extend a raw device mapping, you follow the same basic procedure as with a physical server. First extend the RDM volume from the Storage Center GUI, rescan disks from Windows disk management, and then use DISKPART to extend the drive.

Just as with VMFS datastore volumes, it is also very important not to extend an RDM volume past the 2047GB/1.99TB limit.
Replays and Virtual Machine Backups

Backing up virtual machines

The key to any good backup strategy is not only testing the backup, but also verifying the results. There are many ways to back up virtual machines, but depending on your business needs, each solution is usually unique to each environment. Through testing and verification, you may find that one solution works better in your environment than another, so it is best to test a few different options.

Since the subject of backing up virtual machines is so vast, this section will only cover a few basics. If you need more information about virtual machine backup strategies, an excellent resource is the “Virtual Machine Backup Guide” found on VMware’s documentation pages. Depending on the version of ESX you are using, this guide is usually found with the “VMware Consolidated Backup” documentation.

Backing up virtual machines to tape

Perhaps the most common methods of backing up virtual machines to tape are using backup client software installed within the guest, within the service console, or on a VMware Consolidated Backup proxy server.

- **Backup client loaded within the guest**
  - Using this method, backup software is loaded within the guest operating system, and the data is backed up over the network to a backup host containing the tape drive. Depending on the software used, it usually only performs file level backups, but in some cases, it can include additional capabilities for application level backups.

- **Backup client loaded within the ESX service console**
  - Certain backup software clients have the ability to be loaded within the ESX service console to perform backups at the host level. These backup clients are usually capable of backing up the entire virtual machine, or even backing up files within the virtual machine. Before considering this option, it is best to check the VMware compatibility lists to find the approved backup software vendors.

- **Backup client loaded onto a VMware Consolidated Backup (VCB) Proxy**
  - Using VMware Consolidated Backup allows you to offload the backup to a SAN attached proxy host. This host has the virtual machine volumes mapped to it so that the backup software can access the SAN directly to backup virtual machine data to disk or tape. Using VCB allows you to back up the entire virtual machine, or even the files within the virtual machine. Again, only certain backup software vendors provide plug-in modules for VCB, so it is best to check VMware’s compatibility lists for approved vendors.
Back up virtual machines using Replays

There are several options for backing up virtual machines using Storage Center Replays.

- **Replays scheduled from within the Storage Center GUI**
  - From within the Storage Center GUI, you can create a replay profile to schedule replays of virtual machine volumes. In most cases, using replays to back up virtual machines is sufficient to perform a standard recovery. It is important to remember that replays can only capture data that has been written to disk, and therefore the virtual machine data is preserved in what is called a ‘crash consistent’ state. In other words, when recovering the virtual machine, the data recovered will be as if the virtual machine had simply lost power. Most modern journaling file systems such as NTFS or EXT3 are designed to recover from such states.

- **Replays taken via Compellent’s Replay Manager Software**
  - Since virtual machines running transactional databases are more sensitive to crash consistent data, Compellent has developed its Replay Manager software to utilize Microsoft’s VSS framework for taking replays of Microsoft Exchange and SQL databases. This is a software agent that is loaded within the guest to ensure that the database is in a consistent state before executing the replay.

- **Replays taken via Compellent’s scripting tools**
  - For applications that need a custom method for taking consistent replays of the data, Compellent has developed two scripting tools:
    - **Compellent Command Utility (CompCU)** – This is a java based scripting tool that allows you to script many of the Storage Center’s tasks (such as taking replays).
    - **Storage Center Command Set for Windows PowerShell** – This scripting tool will also allow you to script many of the same storage tasks using Microsoft’s PowerShell scripting language.
  - A good example of using one of these scripting utilities is writing a script to take a replay of an Oracle database after it is put into hot backup mode.
Recovering Virtual Machine Data from a Replay

When recovering a VMFS datastore from a replay, you can recover an entire virtual machine, an individual virtual disk, or files within a virtual disk.

The basic steps are as follows:

1) From the Storage Center GUI, select the replay you wish to use and then choose:
   - Local Recovery or Create Volume from Replay
2) Continue through the local recovery wizard to create the view volume, and map it to the ESX host you wish to recover the data.
   a. Be sure to map the recovery view volume using a LUN which is not already in use.
3) Rescan the HBAs from the “Storage Adapter” section to detect the new LUN
4) From the vSphere client configuration tab,
   a. Select “Storage”
   b. Click “Add Storage…”
   c. Select “Disk/LUN” and then click “Next”
   d. Select the LUN for the view volume you just mapped to the host and then click “Next”.
   e. You are presented with three options:
      i. Keep the Existing Signature – This option should only be used if the original datastore is not present on the host.
      ii. Assign a New Signature – This option will regenerate the datastore signature so that it can be accessed by the host.
         1. Select this option if you are unsure of which option to use.
      iii. Format the disk – This option will format the view volume, and create a new datastore from it.
   f. Finish through the wizard verifying all selections.
5) Once the datastore has been resignatured, the snap datastore will be accessible:

Screenshot: The storage configuration tab showing snapshot datastore

<table>
<thead>
<tr>
<th>Datastore Identification</th>
<th>Status</th>
<th>Device</th>
<th>Capacity</th>
<th>Free</th>
<th>Type</th>
<th>Last Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUN0-templates.iso</td>
<td>Normal</td>
<td>COMPELNT Fibre...</td>
<td>299.75 GB</td>
<td>269.67 GB</td>
<td>vmdk</td>
<td>6/24/2009 10:12:01 AM</td>
</tr>
<tr>
<td>LUN0-vm-storage</td>
<td>Normal</td>
<td>COMPELNT Fibre...</td>
<td>499.75 GB</td>
<td>499.20 GB</td>
<td>vmdk</td>
<td>6/24/2009 10:15:43 AM</td>
</tr>
<tr>
<td>snap-3f5c3b6f-LUN0-vm-storage</td>
<td>Normal</td>
<td>COMPELNT Fibre...</td>
<td>499.75 GB</td>
<td>499.20 GB</td>
<td>vmdk</td>
<td>6/24/2009 10:15:43 AM</td>
</tr>
<tr>
<td>Storage1</td>
<td>Normal</td>
<td>COMPELNT Fibre...</td>
<td>13.50 GB</td>
<td>11.56 GB</td>
<td>vmdk</td>
<td>6/24/2009 10:15:43 AM</td>
</tr>
</tbody>
</table>

6) The recovery datastore is now designated with “snap-xxxxxxxxx-originalname”
7) From here you can browse the datastore to perform the recovery via one of the methods listed below.

**Recovering a file from a virtual disk**

To recover a file from within a virtual disk located on this snap datastore, simply “Add” a new virtual disk to the virtual machine, and then select “Use an existing virtual disk”. Browse to select the virtual disk to recover from, and add it to the virtual machine. You should now be able to assign a drive letter to the virtual disk, and recover/copy/move the file back to its original location.
After you have completed recovering the file, it is important that you remove the recovered virtual disk from the virtual machine before unmapping or deleting the view volume.

**Recovering an entire virtual disk**

To recover an entire virtual disk from the snap datastore, browse to the virtual disk you wish to recover, right click, and select "Move to". Following through the wizard, browse to the destination datastore and folder, then click "Move".

If you are moving a vmdk file back to its original location, remember that you must power off the virtual machine to overwrite the virtual disk. Also, depending on the size of the virtual disk, this operation may take anywhere between several minutes to several hours to finish. When moving the virtual disk from the vSphere client datastore browser, the time required to move this virtual disk is greatly increased due to the fact that the virtual disk being moved is automatically converted to the eagerzeroedthick format regardless of the original format type.

If you want to preserve the original VMDK format during the copy, you can specify either the "-d thin" or "-d zeroedthick" options when using vmkfstools through the ESX CLI. In addition to preserving the original format, this may also reduce the time required to copy the VMDK, since vmkfstools will not have to write the "white space" (zeros) associated with the eagerzeroedthick format.

**Recovering an entire virtual machine**

To recover an entire virtual machine from the snap datastore, browse to the virtual machine configuration file (*.vmx), right click, then select add to inventory. Follow through the wizard to add the virtual machine into inventory.

To prevent network name or IP address conflicts when powering on the newly recovered virtual machine, it is a good idea to power off, or place one of the virtual machines onto an isolated network or private vSwitch.

If virtual center detects a duplicate UUID, you may be prompted with the following virtual machine message:

**Screenshot: Virtual Machine Question prompting for appropriate UUID action**

The selections behave as follows:

- **I moved it** – This option will keep the configuration file UUIDs and the MAC addresses of the virtual machine ethernet adapters.
• **I copied it** – This option will **regenerate** the configuration file UUIDs and the MAC addresses of the virtual machine ethernet adapters.

If you do not know which option to chose, you should select “I copied it”, which will regenerate a new MAC address to prevent conflicts on the network.
Replication Overview

Storage Center replication in coordination with the vSphere 4.x line of products can provide a robust disaster recovery solution. Since each different replication method effects recovery a little differently, choosing the correct one to meet your business requirements is important. Here is a brief summary of the different options.

- **Synchronous**
  - The data is replicated real-time to the destination. In a synchronous replication, an I/O must be committed on both systems before an acknowledgment is sent back to the host. This limits the type of links that can be used, since they need to be highly available with low latencies. High latencies across the link will slow down access times on the source volume.
  - The downside to this replication method is that replays on the source volume are not replicated to the destination, and any disruption to the link will force the entire volume to be re-replicated from scratch.
  - Keep in mind that synchronous replication does not make both the source and destination volumes read/writeable.

- **Asynchronous**
  - In an asynchronous replication, the I/O needs only be committed and acknowledged to the source system, so the data can be transferred to the destination in a non-concurrent timeframe. There are two different methods to determine when data is transferred to the destination:
    - **By replay schedule** – The replay schedule dictates how often data is sent to the destination. When each replay is taken, the Storage Center determines which blocks have changed since the last replay (the delta changes), and then transfers them to the destination. Depending on the rate of change and the bandwidth, it is entirely possible for the replications to “fall behind”, so it is important to monitor them to verify that your recovery point objective (RPO) can be met.
    - **Replicating the active replay** – With this method, the data is transferred “near real-time” to the destination, usually requiring more bandwidth than if you were just replicating the replays. As each block of data is written on the source volume, it is committed, acknowledged to the host, and then transferred to the destination “as fast as it can”. Keep in mind that the replications can still fall behind if the rate of change exceeds available bandwidth.
  - Asynchronous replications usually have less stringent bandwidth requirements making them the most common replication method.
The benefit of an asynchronous replication is that the replays are transferred to the destination volume, allowing for “check-points” at the source system as well as the destination system.

Replication Considerations with Standard Replications

One thing to keep in mind about the Storage Center replication is that when you replicate a volume either synchronously or asynchronously, the replication only “flows” in one direction. In other words, any changes made to the destination volume will not be replicated back to the source. That is why it is extremely important not to map the replication’s destination volume directly to a host instead of creating a read-writable “view volume”.

Since block changes are not replicated bidirectionally with standard replication, this means that you will not be able to VMotion virtual machines between your source controllers (your main site) and your destination controller (your DR site). That being said, there are a few best practices to replication and remote recovery that you should consider.

- You will need compatible ESX server hardware at your DR site to map your replicated volumes to in the event your source ESX cluster becomes inoperable.
- You should make preparations to have all of your Virtual Center resources replicated to the DR site as well.
- To keep your replication sizes smaller, you should separate the operating system pagefiles onto their own non-replicated volume.

Replication Considerations with Live Volume Replications

When a replication is converted to a Live Volume replication, the volume will become read-writable from both the main system and secondary system. This will allow VMotion of virtual machines over distance; however it is important that VMware’s Long Distance VMotion best practices are followed. This means that the VMotion network between ESX hosts must be gigabit or greater, round trip latency must be 5 milliseconds or less, and the virtual machine IP networks must be “stretched” between data centers. In addition, the storage replication must also be high bandwidth and low latency to ensure Live Volumes can be kept in sync. The amount of bandwidth required to keep Live Volumes in sync highly depends on the environment, so it is best you test to determine the bandwidth requirements for your implementation.

For more information, please consult the “Compellent Storage Center Best Practices for Live Volume” guide available on Knowledge Center.
Replication Tips and Tricks

- Since replicated volumes can contain more than one virtual machine, it is recommended that you sort your virtual machines into specific replicated and non-replicated volumes. For example, if you have 30 virtual machines in your ESX cluster, and only 8 of them need to be replicated to your DR site, create a special “Replicated” volume to place those 8 virtual machines on.
- As mentioned previously, keep operating system pagefiles on a separate volume that you will not replicate. That will keep replication and replay sizes smaller because the data in the pagefile changes frequently and it is generally not needed for a system restore.
- As an alternative to setting replication priorities, you can also take advantage of the Storage Center QOS to prioritize replication bandwidth of certain volumes. For example, if you have a 100 Mb pipe between sites, you could create two QOS definitions such that the “mission critical” volume would get 80 Mb of the bandwidth, and the lower priority volume would get 20 Mb of the bandwidth.

Virtual Machine Recovery at a DR site

When recovering virtual machines at the disaster recovery site, you should follow the same general steps as outlined in the previous section titled “Recovering Virtual Machine Data from a Replay”.

If you have a significant number of volumes you need mapped up to perform a recovery, you can save time during the recovery process by using the “Replication Recovery” functionality within Compellent’s Enterprise Manager Software. These features will allow you to pre-define your recovery with things such as the appropriate hosts, mappings, LUN numbers, and host HBAs. After the recovery has been predefined, a recovery at the secondary site is greatly automated.

It is extremely important that the destination volume, usually denoted by “Repl of”, never gets directly mapped to an ESX host while data is actively being replicated. Doing so will inevitably cause data integrity issues in the destination volume, requiring the entire volume be re-replicated from scratch. The safest recovery method is to always restore the virtual machine from a local recovery or “view volume” as shown in previous sections. Please see the Copilot Services Technical Alert titled, “Mapping Replicated Volumes at a DR Site” available on Compellent Knowledge Center for more info.
Appendix A - Determining the appropriate queue depth for an ESX host

Adjusting the queue depth on your ESX hosts is a very complicated subject. On one hand, increasing it can remove bottlenecks and help to improve performance (as long as you have enough back end spindles to handle the incoming requests). Yet on the other hand, if set improperly, the ESX hosts could overdrive the controller front-end ports or the back end spindles, and potentially make the performance worse.

The general rule of thumb is to set the queue depth high enough so that you achieve an acceptable number of IOPS from the back end spindles, while at the same time, not setting it too high allowing an ESX host to flood the front or back end of the array.

Here are a few basic pointers:

- **Fiber Channel**
  - 2 GBPS Storage Center Front-End Ports
    - Each 2 GBPS FE port has a max queue depth of 256 so you must be careful not to overdrive it
    - It is generally best to leave the ESX queue depths set to default and only increase if absolutely necessary
    - Recommended settings for controllers with 2 GBPS FE ports
      - HBA BIOS = 255
      - HBA Queue depth is actually regulated by the driver module
      - Driver module = 32 (Default)
      - Disk.SchedNumReqOutstanding = 32 (Default)
      - Guest OS LSI Logic = 32 (Default)
  - 4/8 GBPS Storage Center Front-End Ports
    - Each 4/8 GBPS front-end port has a max queue depth of ~1900, so it can accept more outstanding I/Os
    - Since each FE port can accept more outstanding I/Os, the ESX queue depths can be set higher. Keep in mind, the queue depth may need to be decreased if the front-end ports become saturated, the back end spindles become maxed out, or the latencies become too high.
    - Recommended settings for controllers with 4/8 GBPS FE ports
      - HBA BIOS = 255
      - Driver module = 255
      - Disk.SchedNumReqOutstanding = 32 (Default)
      - Guest OS LSI Logic = 32 (Default)
        - Increase/decrease as necessary
• iSCSI
  o Software iSCSI
    ▪ Leave the queue depth set to default and only increase if necessary
  o Hardware iSCSI
    ▪ Leave the queue depth set to default and only increase if necessary

The best way to determine if you have the appropriate queue depth set is by using the esxtop utility. This utility can be executed from one of the following locations:

• ESX Service console
  o Command: esxtop

• Remote CLI package (RCLI) or the vMA Virtual Appliance
  o Command: resxtop.sh

When opening the esxtop utility, the best place to monitor queue depth and performance is from the “Disk Device” screen. Here is how to navigate to that screen:

• From the command line type either:
  o # esxtop
  o # resxtop.sh --server esxserver.domain.local
    ▪ Enter appropriate login credentials
• Enter the “Disk Device” screen by pressing “u”
• Expand the “devices” field by pressing “L 36 <enter>” (Capital “L”)
  o This will expand the disk devices so that you can identify the LUNs
• Chose the “Fields” you wish to monitor by pressing “f”:
  o Press “b” to uncheck the ID field (not needed)
  o OPTIONALLY: (Depending on what you want to monitor)
    ▪ Check or uncheck “h” for overall Latency
    ▪ Check “i” for read latency
    ▪ Check “j” for write latency
  o Press <enter> to return to the monitoring screen
• Set the refresh time by pressing “s 2 <enter>”. (Refresh every 2 seconds)

The quick and easy way to see if your queue depth is set correctly is to monitor the queue depth section in coordination with the latency section.

Screenshot: Screenshot of esxtop with a queue depth of 32 (Edited to fit screen)

<table>
<thead>
<tr>
<th>DEVICE</th>
<th>DQLEN</th>
<th>ACTV</th>
<th>QUED</th>
<th>%USD</th>
<th>LOAD</th>
<th>CMDS/s</th>
<th>DAVG/cmd</th>
<th>XAVG/cmd</th>
<th>GAVG/cmd</th>
<th>QAVG/cmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsa.600728a</td>
<td>32</td>
<td>32</td>
<td>0</td>
<td>100</td>
<td>1.00</td>
<td>1690.27</td>
<td>19.00</td>
<td>0.00</td>
<td>19.00</td>
<td>0.00</td>
</tr>
<tr>
<td>nsa.600728b</td>
<td>32</td>
<td>32</td>
<td>16</td>
<td>100</td>
<td>1.50</td>
<td>1675.51</td>
<td>18.98</td>
<td>9.34</td>
<td>20.32</td>
<td>9.33</td>
</tr>
<tr>
<td>nsa.600728c</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>100</td>
<td>2.00</td>
<td>1691.26</td>
<td>18.94</td>
<td>18.70</td>
<td>37.55</td>
<td>18.70</td>
</tr>
<tr>
<td>nsa.600728d</td>
<td>32</td>
<td>32</td>
<td>96</td>
<td>100</td>
<td>4.00</td>
<td>1674.50</td>
<td>19.03</td>
<td>56.93</td>
<td>75.96</td>
<td>56.93</td>
</tr>
</tbody>
</table>

Generally speaking, if the LOAD is consistently greater than 1.00 on one or more LUNs, the latencies are still acceptable, and the back end spindles have available IOPS, then increasing the queue depth may make sense.

However, if the LOAD is consistently less than 1.00 on a majority of the LUNs, and the performance and latencies are acceptable, then there is usually no need to adjust the queue depth.
In the screenshot above, the device queue depth is set to 32. As you can see, three of the four LUNs consistently have a LOAD above 1.00. If the back end spindles are not maxed out, it may make sense to increase the queue depth.

**Screenshot: The queue depth increased to 255 (Edited to fit screen)**

<table>
<thead>
<tr>
<th>DEVICE</th>
<th>DQLEN</th>
<th>ACTV</th>
<th>QUED</th>
<th>USD</th>
<th>LOAD</th>
<th>CMDs/s</th>
<th>DAVG/cmd</th>
<th>KAVG/cmd</th>
<th>GAvg/cmd</th>
<th>QAVG/cmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>hbaa.60C720a</td>
<td>255 120</td>
<td>0</td>
<td>50</td>
<td>0.50</td>
<td>1647.55</td>
<td>67.65</td>
<td>0.00</td>
<td>57.66</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>hbaa.60C720b</td>
<td>255 120</td>
<td>0</td>
<td>50</td>
<td>0.50</td>
<td>1852.50</td>
<td>67.35</td>
<td>0.00</td>
<td>57.35</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>hbaa.60C720c</td>
<td>255 120</td>
<td>0</td>
<td>50</td>
<td>0.50</td>
<td>1837.16</td>
<td>68.45</td>
<td>0.00</td>
<td>58.45</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>hbaa.60C720d</td>
<td>255 120</td>
<td>0</td>
<td>50</td>
<td>0.50</td>
<td>1813.40</td>
<td>67.23</td>
<td>0.00</td>
<td>57.28</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

As you can see, by increasing the queue depth from the previous example, the total IOPS increased from 6700 to 7350 (109%), but the average device latency (DAVG/cmd) increased from 18ms to 68ms (377%). That means the latency over tripled for a mere 9% performance gain. In this case, it may not make sense to increase the queue depth because latencies became too high.

For more information about the disk statistics in esxtop, consult the esxtop man page, or the VMware document: “vSphere Resource Management Guide – Appendix A”.

**Appendix B - Configuring Enterprise Manager VMware Integrations**

For customers that are running Enterprise Manager Versions 5.3.x or higher, the Enterprise Manager Data Collector can be configured to gather storage statistics and perform basic storage administration functions with vCenter.

To add vCenter credentials into Enterprise Manager, enter the **Servers Viewer** screen, and then right click on **Servers** then select **Register Server**.

![Register Server](image)

After entering vCenter credentials, you will now be able to see aggregate storage statistics, as well as being able to automatically provision VMFS datastores and RDMs.
For example, when creating a new volume, if you select the ESX host, it will automatically give you the option to format it with VMFS.

Similarly, when creating a new volume to be assigned to a virtual machine, Enterprise Manager can automatically add the volume as an RDMP.
Conclusion

Hopefully this document has answered many of the questions you have encountered or will encounter while implementing VMware vSphere with your Compellent Storage Center.

More information

If you would like more information, please review the following web sites:

- Compellent
  - Compellent Training: [http://www.compellent.com/services/training.aspx](http://www.compellent.com/services/training.aspx)

- VMware
  - VMware Infrastructure 4 Online Documentation: [http://pubs.vmware.com/vsp40](http://pubs.vmware.com/vsp40)
  - VMware Communities: [http://communities.vmware.com](http://communities.vmware.com)