Dell Compellent Storage Center

Best Practices with vSphere 5.x

Compellent

Dell Compellent Document Number: 680-041-020

Document revision

Date	Revision	Comments
9/14/2011	1	Initial Release

THIS BEST PRACTICES GUIDE IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

Contents

Tables

General syntax

Table 1. Document syntax

Item	Convention
Menu items, dialog box titles, field names, keys	**Bold**
Mouse click required	Click:
User Input	`Monospace Font`
User typing required	Type:
Website addresses	http://www.compellent.com
Email addresses	info@compellent.com

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.

Overview

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 or vSphere 5.x
- Configuration and operation of the Dell 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 vSphere 5.x integrates with Storage Center.

Introduction

This document will provide configuration examples, tips, recommended settings, and other storage guidelines a user can follow while integrating VMware ESXi 5.x Server hosts with the Storage Center. This document has been written to answer many frequently asked questions with regard to how VMware interacts with the Storage Center's various features such as Dynamic Capacity, Data Progression, and Remote Instant Replay.

Dell Compellent advises customers to read the vSphere Storage Guide, which is publicly available on the vSphere documentation pages to provide additional important information about configuring ESXi hosts 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 individual or business needs.

Fiber Channel Switch Zoning

Zoning fibre channel switches for an ESXi host is done much the same way as any other server connected to the Storage Center. Here are the fundamental points:

Single Initiator Multiple Target Zoning

Each fiber channel zone created should have a single initiator (HBA port) and multiple targets (Storage Center front-end ports). This means that each HBA port needs its own fiber channel zone containing itself and the Storage Center front-end ports. Zoning ESXi hosts by either port (commonly referred to as hard zoning) or WWN (commonly referred to as soft zoning) is acceptable.

Port Zoning

If the Storage Center front-end ports are plugged into switch ports 0, 1, 2, & 3, and the first ESXi 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 ESXi host. If the environment has multiple fabrics, the additional HBA ports in the host should have separate unique zones created in their respective fabrics.

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 similar to this:

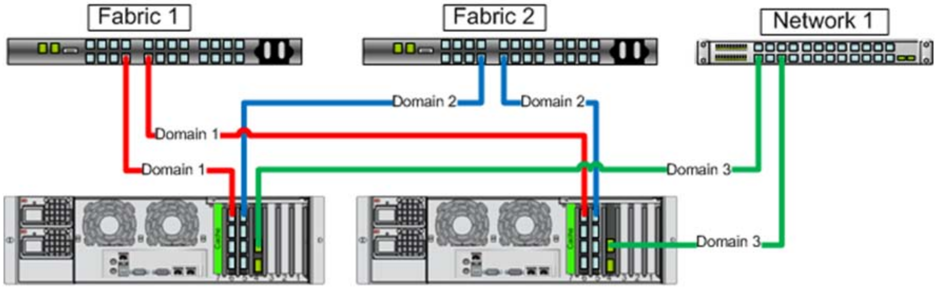
Name: ESX1-HBA1	(Zone created in fabric 1)
WWN: 2100001B32017114	(ESX1 HBA Port 1)
WWN: 5000D31000036001	(Controller1 front-end primary plugged into fabric 1)
WWN: 5000D31000036009	(Controller2 front-end primary plugged into fabric 1)

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 Virtual Port Mode, all of the Front End virtual ports within each Fault Domain should be included in the zone with each ESXi initiator.

Figure 1 - Virtual Port Domains - FC and iSCSI



Host Bus Adapter Settings

Make sure that the HBA BIOS settings are configured in the ESXi host according to the latest "Storage Center User Guide" found on Knowledge Center. At the time of this writing, here are the current 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 ESXi VMkernel driver module and DSNRO can more conveniently control the queue depth

Emulex Fiber Channel Card BIOS Settings

- The Node Time Out field "lpfc_devloss_tmo" (formerly "nodev_tmo") field should be set to 60 seconds.
    - More info: http://kb.vmware.com/kb/1008487
- The "topology" field should be set to 2 for point to point only
- The "queuedepth" field should be set to 255
    - This queue depth can be set to 255 because the ESXi VMkernel driver module and DSNRO can more conveniently control the queue depth

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: "vSphere Storage Guide"
- Enabling ARP redirect for Hardware iSCSI HBAs
    - http://kb.vmware.com/kb/1010309
    - Example: `esxcli iscsi physicalnetworkportal param set --option ArpRedirect=true -A vmhba4`

Modifying Queue Depth in an ESXi Environment

Queue depth is defined as the number of disk transactions that are allowed to be “in flight” between an initiator and a target, where the initiator is typically an ESXi host HBA 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, there is 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	(Default=dependent on application)
• The virtual SCSI card driver in the guest	(Default=32)
• The VMFS layer (DSNRO)	(Default=32)
• The HBA VMkernel Module driver	(Default=32)
• The HBA BIOS	(Default=32)

The following sections explain how the queue depth is set in each of the layers in the event it needs to be changed.

Caution

The appropriate queue depth for a host may vary due to a number of factors, so it is recommended to only increase or decrease the queue depth 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 VMkernel driver module loaded for each HBA in the system and DSNRO ultimately regulate 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 will be 32.

Modifying ESXi Storage Driver Queue Depth and Timeouts

As mentioned in the previous section, the VMkernel 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 ESXi host to survive a Storage Center controller failover properly.

Please refer to the latest VMware documentation for instructions on how to configure these settings:

- VMware document: "vSphere Troubleshooting"
    - Section Title: "Adjust Queue Depth for a QLogic and Emulex HBAs"

Caution

Before executing these commands, please refer to the latest documentation from VMware 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. Find the appropriate driver name for the module that is loaded:
    - a. For QLogic: `esxcli system module list |grep qla`
    - b. For Emulex: `esxcli system module list |grep lpfc`
        - i. Depending on the HBA model, the output could be similar to:
            - 1. QLogic: `qla2xxx`
            - 2. Emulex: `lpfc820`

Note

The steps below contain *example* module names. Actual module names should be acquired from the step above.

2. Set the driver queue depth and timeouts using the esxcli command:
    - a. For QLogic: `esxcli system module parameters set -m qla2xxx -p "ql2xmaxqdepth=255 ql2xloginretrycount=60 qlport_down_retry=60"`
    - b. For Emulex: `esxcli system module parameters set -m lpfc820 -p "lpfc_devloss_tmo=60 lpfc_hba_queue_depth=255"`
3. Reboot the ESXi host for these changes to take effect.
4. To verify the settings, use the following command:
    - a. `esxcli system module parameters list -m=module`

Similarly, for the software iSCSI Initiator:

1. Example of setting the queue depth to 255:
    - a. `esxcli system module parameters set -m iscsi_vmk -p iscsivmk_LunQDepth=255`
2. Reboot the ESXi host for the change to take effect.
3. To verify the settings, use the following command:
    - a. `esxcli system module parameters list -m iscsi_vmk`

Modifying the VMFS Queue Depth for Virtual Machines (DSNRO)

Another setting which controls the queue depth at the virtual machine and per datastore level is located in the ESXi host's advanced settings:

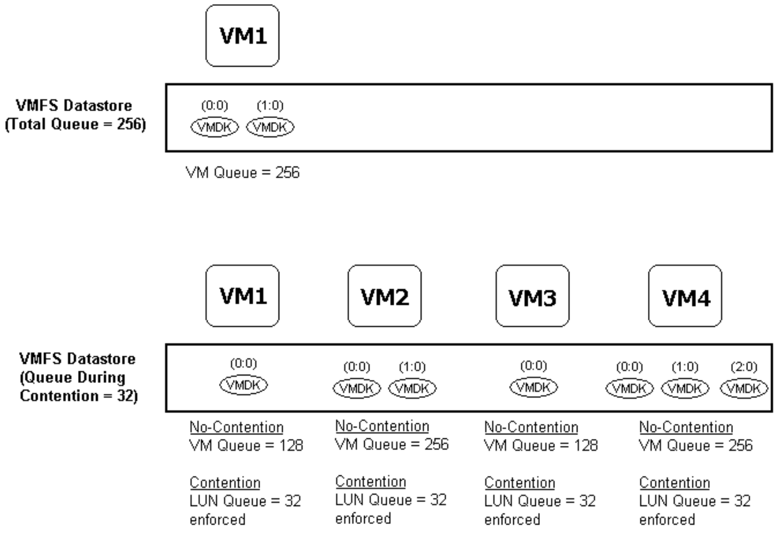
Disk.SchedNumReqOutstanding (Default=32)

This is another value that can be increased or decreased depending on how many virtual machines are to be placed on each datastore or their I/O requirements. Keep in mind, this queue depth limit is only enforced when more than one virtual machine per host 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, during contention 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 the host has one 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.

VMware recommends that if the VMkernel Module driver queue depth is changed, that this variable is set to match, since they will gate each other. (See Appendix A for more information about determining the appropriate queue depth.)

Figure 2: Example queue utilization when the Disk.SchedNumReqOutstanding is set to 32



Note

The Disk.SchedNumReqOutstanding limit does not apply to LUNs mapped as Raw Device Mappings (RDMs). Each RDM will have its own queue.

More information on the Disk.SchedNumReqOutstanding variable can be found in the following documents:

- VMware document: "vSphere Troubleshooting"
  - Section: "Change Maximum Outstanding Disk Requests for 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 this 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 2008/R2

Since the default LSI Logic driver is already at an acceptable version, all that needs to be done is add the following registry keys:

1. Using regedit, add the following keys: (Backup the 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.

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. Download the following driver from the LSI Logic download page:
   - Adapter: LSI20320-R
   - Driver: Windows Server 2003 (32-bit)
   - Version: WHQL 1.20.18 (Dated: 13-JUN-05)
   - 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 the 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.

Note

Please visit VMware’s Knowledge Base for the most current information about setting the queue depth with different vSCSI controllers or operating systems.

Setting Operating System Disk Timeouts

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 document:

- VMware document: “vSphere Storage Guide”
    - Section Title: “Set Timeout on Windows Guest OS”

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

Windows

1. Using the registry editor, modify the following key: (Backup the registry first)
   [HKLM\SYSTEM\CurrentControlSet\Services\Disk]
   "TimeOutValue"=dword:0000003c (x03c hex = 60 seconds in decimal)

2. Reboot the virtual machine.

Linux

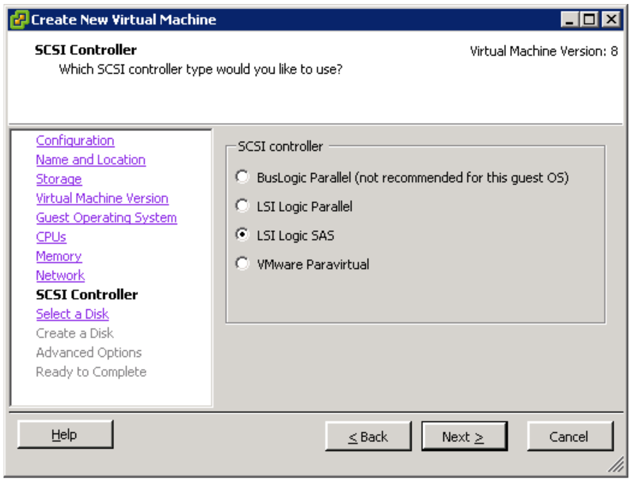
For more information about setting disk timeouts in Linux, please refer to the following VMware Knowledge Base article:

- Increasing the disk timeout values for a Linux virtual machine
    - http://kb.vmware.com/kb/1009465

Guest Virtual SCSI Adapters

When creating a new virtual machine there are four types of virtual SCSI Controllers that can be selected depending on the guest operating system.

Figure 3: vSCSI adapter selection



Create New Virtual Machine

SCSI Controller

Which SCSI controller type would you like to use?

Virtual Machine Version: 8

SCSI controller

○ BusLogic Parallel (not recommended for this guest OS)
○ LSI Logic Parallel
● LSI Logic SAS
○ VMware Paravirtual

BusLogic Parallel

This vSCSI controller is used for certain older operating systems. Due to this controller's queue depth limitations, its use is not recommended unless it is the only option available for that particular 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.

LSI Logic SAS

This vSCSI controller is available for virtual machines with hardware version 7 & 8, 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. Some operating system vendors are gradually withdrawing support for SCSI in favor of SAS, thus making the LSI Logic SAS controller a good choice for future compatibility.

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 Virtual Machine Administration Guide" guide, in a section titled, "About Paravirtualized SCSI Controllers".

Mapping Volumes to an ESXi Server

Basic Volume Mapping Concepts

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

For example:
There are three ESXi hosts named ESX1, ESX2, and ESX3.
A new volume is created named "LUN10-vm-storage".

This volume must be mapped to each of the ESXi hosts as the same LUN:

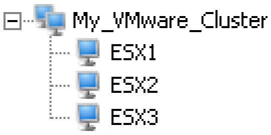
Volume: "LUN10-vm-storage" → Mapped to ESX1 -as- LUN 10
Volume: "LUN10-vm-storage" → Mapped to ESX2 -as- LUN 10
Volume: "LUN10-vm-storage" → Mapped to ESX3 -as- LUN 10

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 ESXi host.

Basic Volume Mappings in Storage Center 5.x and later

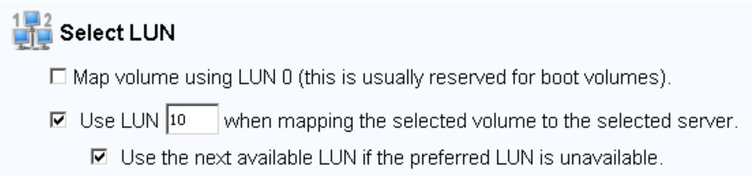
However in Storage Center versions 5.x and higher, the mapping process is greatly automated by creating a server cluster object. This will allow the volume to be mapped to multiple ESXi hosts at the same time, automatically keeping the LUN numbering consistent for all the paths.



As an added benefit, when a new ESXi host is placed into the server cluster, all of the existing volume mappings assigned to the cluster object 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 ESXi host is as simple as adding it to the cluster object.

Similarly, if the host is removed 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 they are removed. 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, the system can auto select the LUN number, or a preferred LUN number can be manually specified from the advanced settings screen in the mapping wizard.



Select LUN

☐ Map volume using LUN 0 (this is usually reserved for boot volumes).

☑ Use LUN 10 when mapping the selected volume to the selected server.

☑ Use the next available LUN if the preferred LUN is unavailable.

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

Timesaver

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

Multi-Pathed Volume Concepts

If there is an ESXi host (or hosts) that have multiple ports, whether they are FC, iSCSI, or Ethernet, ESXi has built-in functionality to provide native multi-pathing of volumes over fiber channel, FCoE, 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

Note

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 ESXi initiator in each server will need to be mapped to one front end port, while the second ESXi 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 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 ESXi, a single volume cannot be mapped 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

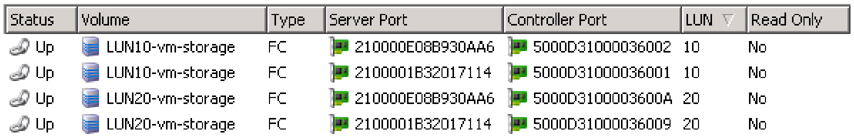
In Storage Center versions 4.x and earlier, when mapping a volume to a host, if it has multiple HBAs assigned in the server object, the mapping wizard will allow the selection of both paths. When the volume is presented down multiple paths using the same LUN number, ESXi's native multi-pathing module will automatically detect and use both paths with either Fixed or Round Robin policies.

Note

Before beginning, with certain versions of Storage Center, multi-pathing may need to be enabled 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 ESXi host should look when finished.

Figure 4: Example of Multi-pathing mappings for ESX1

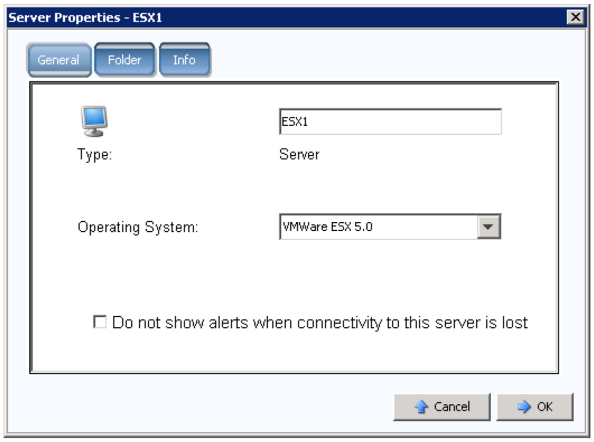


Status	Volume	Type	Server Port	Controller Port	LUN ▽	Read Only
Up	LUN10-vm-storage	FC	210000E08B930AA6	5000D31000036002	10	No
Up	LUN10-vm-storage	FC	2100001B32017114	5000D31000036001	10	No
Up	LUN20-vm-storage	FC	210000E08B930AA6	5000D3100003600A	20	No
Up	LUN20-vm-storage	FC	2100001B32017114	5000D31000036009	20	No

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. Many of the common mapping errors can be prevented by simply 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.

Figure 5: Server Operating System Selection



Server Properties - ESX1

ESX1

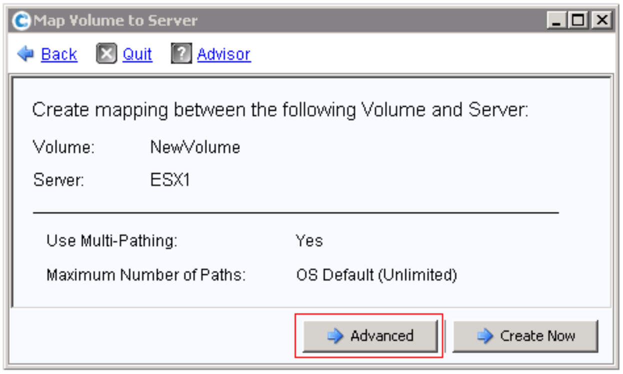
Type: Server

Operating System: VMWare ESX 5.0

☐ Do not show alerts when connectivity to this server is lost

Multipathing to an ESXi host is automatic if the server object has multiple HBA's or iSCSI initiator ports assigned to it. In other words, the advanced options will have to be used if the server does not need a volume multipathed.

Figure 6: Advanced Server Mapping Options



Create mapping between the following Volume and Server:

Volume: NewVolume

Server: ESX1

Use Multi-Pathing: Yes

Maximum Number of Paths: OS Default (Unlimited)

Table 2. Advanced Mapping Options for an ESXi 5.x host

Function	**Description**
Select LUN	This option is to manually specify the LUN. If this box is not checked, the system will automatically assign the next available LUN.
Restrict Mapping paths	This option is used when a volume only needs to be mapped to a specific HBA in the ESXi host.
Map to Controller	By default, the system will automatically select which controller the volume should be mapped. To force a particular controller to handle the I/O, use this option.
Configure Multipathing	This option designates how many of the Storage Center FE ports that the system 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 ESXi to see the Storage Center via the ESXi software initiator.

From within the VMware vSphere Client:

1. Enable the "Software iSCSI Client" within the ESXi firewall (located in the "Security Profile" of the ESXi host)
2. Add a "VMkernel port" to a virtual switch assigned to the physical NIC for iSCSI (See figure below)
3. From the Storage Adapters configuration screen, click Add…
4. Select "Add Software iSCSI Adapter", then click OK
5. From within the Storage Adapters, highlight the iSCSI Software Adapter (i.e. vmhba33), click "Properties"
6. Under the "Dynamic Discovery" tab, add all of the Storage Center iSCSI IP addresses that are assigned to the iSCSI cards in the Storage Center controller(s), or just the iSCSI Control Port IP address.
7. Rescan the iSCSI Initiator.

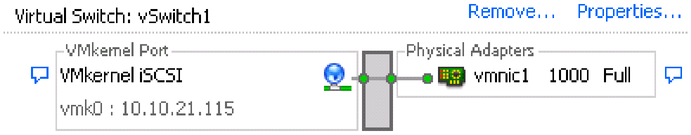
From Within the Storage Center GUI:

8. Create a server object for the ESXi host using the IP Address previously specified for the VMkernel in step 2 above
9. Map a volume to the ESXi host

From within the VMware vSphere Client:

10. Navigate to the Storage Adapters section, and rescan the iSCSI HBA for new LUN's.

Figure 7: Configuring the VMkernel port



Virtual Switch: vSwitch1

VMkernel Port: VMkernel iSCSI
vmk0 : 10.10.21.115

Physical Adapters: vmnic1 1000 Full

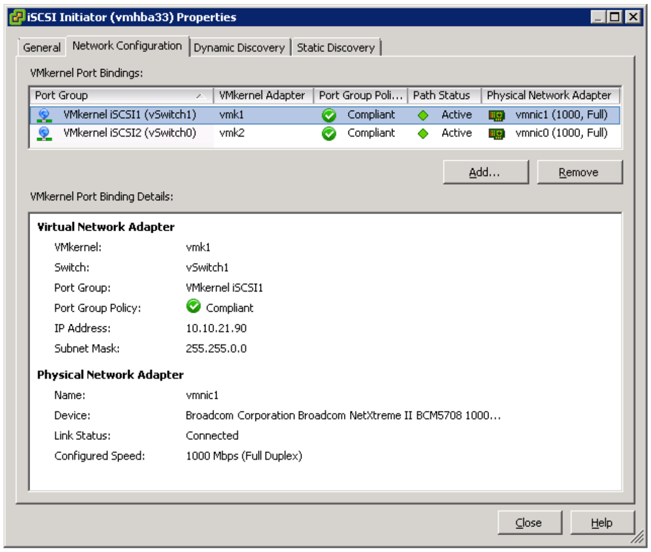
Configuring the VMware iSCSI software initiator for multipathing

In ESXi 5.x, the software initiator can be completely configured via the vSphere client. Instructions on how to configure this can be found in the following document:

- VMware document: "vSphere Storage Guide"
    - Section Title: "Configuring Software iSCSI Adapter"
    - Subsection: "Multiple Network Adapters in iSCSI Configuration"

After following the VMware instructions on how to configure the software iSCSI initiator to use both NICs for multipathing (See figure below), the ESXi host can then be added to the Storage Center.

Figure 8: Binding Multiple VMkernel ports to the Software iSCSI Initiator



iSCSI Initiator (vmhba33) Properties

VMkernel Port Bindings:

Port Group	VMkernel Adapter	Port Group Poli...	Path Status	Physical Network Adapter
VMkernel iSCSI1 (vSwitch1)	vmk1	✓ Compliant	◆ Active	vmnic1 (1000, Full)
VMkernel iSCSI2 (vSwitch0)	vmk2	✓ Compliant	◆ Active	vmnic0 (1000, Full)

VMkernel Port Binding Details:

Virtual Network Adapter

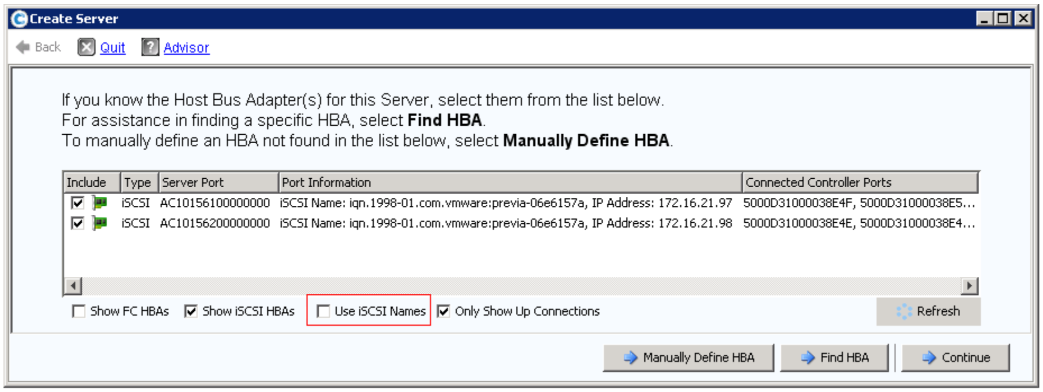
VMkernel:	vmk1
Switch:	vSwitch1
Port Group:	VMkernel iSCSI1
Port Group Policy:	✓ Compliant
IP Address:	10.10.21.90
Subnet Mask:	255.255.0.0

Physical Network Adapter

Name:	vmnic1
Device:	Broadcom Corporation Broadcom NetXtreme II BCM5708 1000...
Link Status:	Connected
Configured Speed:	1000 Mbps (Full Duplex)

There are two methods to adding the ESXi Software Initiator to the Storage Center. If the initiator is added by IP address, each initiator port/IP can be added and configured independently, thus allowing volumes to be mapped through a single path to the initiator (if needed).

Figure 9: Adding iSCSI HBAs by IP address to a server (without using iSCSI Names)



Create Server

If you know the Host Bus Adapter(s) for this Server, select them from the list below.
For assistance in finding a specific HBA, select **Find HBA**.
To manually define an HBA not found in the list below, select **Manually Define HBA**.

Include	Type	Server Port	Port Information	Connected Controller Ports
☑	iSCSI	AC10156100000000	iSCSI Name: iqn.1998-01.com.vmware:previa-06e6157a, IP Address: 172.16.21.97	5000D31000038E4F, 5000D31000038E5...
☑	iSCSI	AC10156200000000	iSCSI Name: iqn.1998-01.com.vmware:previa-06e6157a, IP Address: 172.16.21.98	5000D31000038E4E, 5000D31000038E4...

☐ Show FC HBAs ☑ Show iSCSI HBAs ☐ Use iSCSI Names ☑ Only Show Up Connections

When the initiator is added to the server object using iSCSI names, both initiator ports are managed as a single object, thus all volumes mapped to it are automatically mapped down all paths. Keep in mind, this method will still use multipathing for each of the volumes; however the ability to map a volume down a single path will be lost.

If unsure which option to pick, the iSCSI initiator ports should be added to the server object by IP address, so that there is flexibility to map volumes independently to each initiator in the future.

VMware Multi-Pathing Policies

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

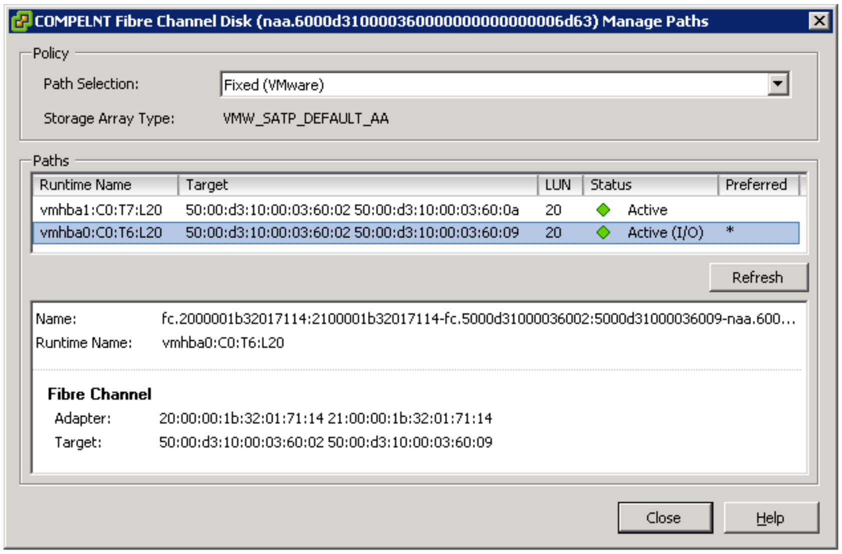
Fixed Policy

If the Fixed policy is used, it will give the greatest control over the flow of storage traffic. However, one must be careful to evenly distribute the load across all host HBAs, Front-End Ports, fabrics, 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 figure below)
HBA1 loses connectivity; HBA2 takes over its connections.
HBA1 resumes connectivity; HBA2 will fail its connections back to HBA1.

Figure 10: Example of a datastore path selection policy set to Fixed



COMPELNT Fibre Channel Disk (naa.6000d31000036000000000000000006d63) Manage Paths

Policy

Path Selection: Fixed (VMware)

Storage Array Type: VMW_SATP_DEFAULT_AA

Paths

Runtime Name	Target	LUN	Status	Preferred
vmhba1:C0:T7:L20	50:00:d3:10:00:03:60:02 50:00:d3:10:00:03:60:0a	20	◆ Active	
vmhba0:C0:T6:L20	50:00:d3:10:00:03:60:02 50:00:d3:10:00:03:60:09	20	◆ Active (I/O)	*

Name: fc.2000001b32017114:2100001b32017114-fc.5000d31000036002:5000d31000036009-naa.600...
Runtime Name: vmhba0:C0:T6:L20

Fibre Channel
Adapter: 20:00:00:1b:32:01:71:14 21:00:00:1b:32:01:71:14
Target: 50:00:d3:10:00:03:60:02 50:00:d3:10:00:03:60:09

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 for the volumes in bursts evenly and sequentially across paths in an alternating fashion.

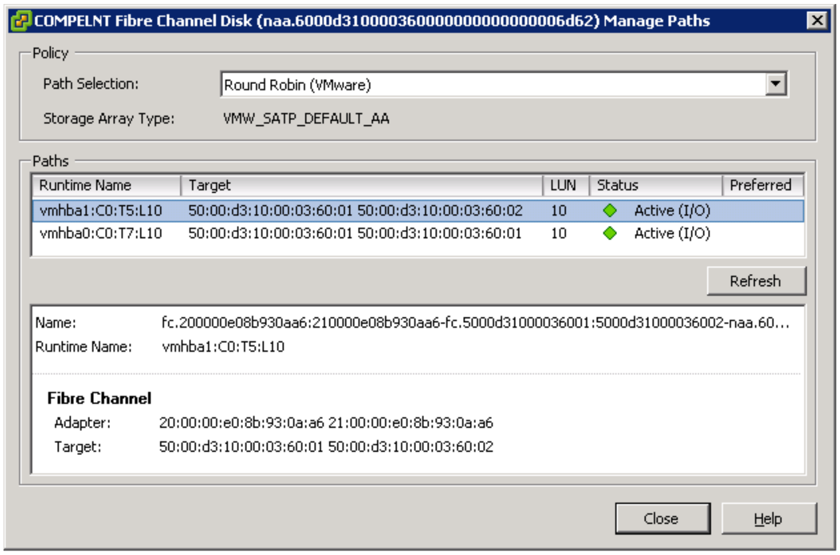
Using round robin will reduce the management headaches of manually balancing the storage load across all storage paths as with a fixed policy; however there are certain situations where using round robin does not make sense. For instance, it is generally not considered best practice to enable round robin 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 round robin is enabled for one or more datastores/LUNs, care must be taken 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 figure 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.

Figure 11: Example of a datastore path selection policy set to Round Robin



COMPELNT Fibre Channel Disk (naa.6000d31000036000000000000000006d62) Manage Paths

Policy

Path Selection: Round Robin (VMware)

Storage Array Type: VMW_SATP_DEFAULT_AA

Paths

Runtime Name	Target	LUN	Status	Preferred
vmhba1:C0:T5:L10	50:00:d3:10:00:03:60:01 50:00:d3:10:00:03:60:02	10	◆ Active (I/O)	
vmhba0:C0:T7:L10	50:00:d3:10:00:03:60:01 50:00:d3:10:00:03:60:01	10	◆ Active (I/O)	

Name: fc.200000e08b930aa6:210000e08b930aa6-fc.5000d31000036001:5000d31000036002-naa.60...
Runtime Name: vmhba1:C0:T5:L10

Fibre Channel
Adapter: 20:00:00:e0:8b:93:0a:a6 21:00:00:e0:8b:93:0a:a6
Target: 50:00:d3:10:00:03:60:01 50:00:d3:10:00:03:60:02

Note

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 storage nmp satp set -P VMW_PSP_RR -s VMW_SATP_DEFAULT_AA
```

If you would like to change all existing paths from Fixed to Round Robin, here is an example VMware PowerCLI PowerShell command to complete the task.

```
Get-Cluster InsertClusterNameHere | Get-VMHost | Get-ScsiLun | where
{$_.Vendor -eq "COMPELNT" –and $_.Multipathpolicy -eq "Fixed"} | Set-
ScsiLun -Multipathpolicy RoundRobin
```

Caution

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 used with 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, the preferred path must be set for each datastore consistently between each host. 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, one must manually validate that all paths have proportional amounts of traffic to each ESXi host.

Multi-Pathing using a Round Robin path selection policy

If the decision is made to use round robin, it must be manually defined for each LUN and host (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 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 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: "vSphere Storage Guide"
    - Section: "Understanding Multipathing and Failover"

Unmapping Volumes from an ESXi host

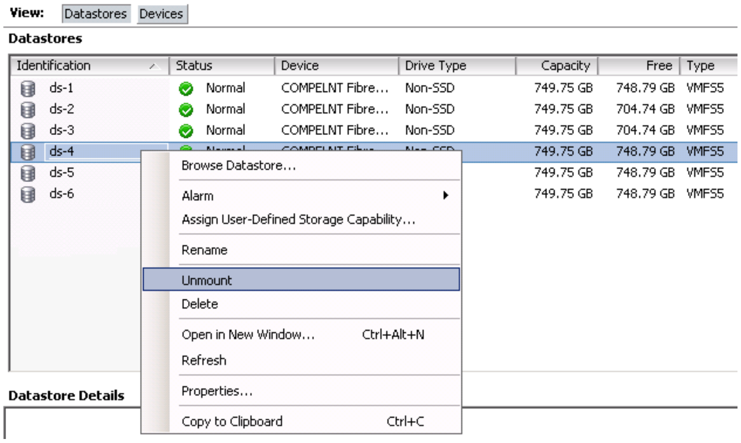
Within ESXi 5.x, VMware has added the ability to gracefully remove volumes before unmapping them to prevent an All Paths Down (APD) state.

Before attempting this procedure, please consult the latest VMware documentation for any last minute changes:

- VMware Document: "vSphere Storage Guide"
    - Section Title: "Planned Device Removal"

1. Make note of the volume's naa identifier. This will be referenced later.
2. From the datastore view, right click on the datastore and select "Unmount".

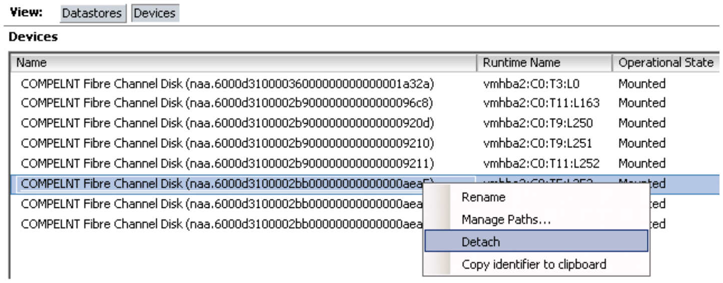
Figure 12: Unmounting a datastore



Identification	Status	Device	Drive Type	Capacity	Free	Type
ds-1	Normal	COMPELNT Fibre…	Non-SSD	749.75 GB	748.79 GB	VMFS5
ds-2	Normal	COMPELNT Fibre…	Non-SSD	749.75 GB	704.74 GB	VMFS5
ds-3	Normal	COMPELNT Fibre…	Non-SSD	749.75 GB	704.74 GB	VMFS5
ds-4				749.75 GB	748.79 GB	VMFS5
ds-5				749.75 GB	748.79 GB	VMFS5
ds-6				749.75 GB	748.79 GB	VMFS5

3. Once the datastore has been successfully unmounted, then select "Detach" on the disk device.

Figure 13: Detaching a datastore



Name	Runtime Name	Operational State
COMPELNT Fibre Channel Disk (naa.6000d31000036000000000000001a32a)	vmhba2:C0:T3:L0	Mounted
COMPELNT Fibre Channel Disk (naa.6000d3100002b90000000000000096c8)	vmhba2:C0:T11:L163	Mounted
COMPELNT Fibre Channel Disk (naa.6000d3100002b9000000000000000920d)	vmhba2:C0:T9:L250	Mounted
COMPELNT Fibre Channel Disk (naa.6000d3100002b9000000000000009210)	vmhba2:C0:T9:L251	Mounted
COMPELNT Fibre Channel Disk (naa.6000d3100002b9000000000000009211)	vmhba2:C0:T11:L252	Mounted
COMPELNT Fibre Channel Disk (naa.6000d3100002bb00000000000000aea		
COMPELNT Fibre Channel Disk (naa.6000d3100002bb00000000000000aea		ted
COMPELNT Fibre Channel Disk (naa.6000d3100002bb00000000000000aea		ted

4. From within the Storage Center GUI, unmap the volume.
5. From within the vSphere client, rescan the adapters to ensure that the disk has been removed.

Note

Graceful removal of volumes from an ESXi host is done automatically when using the Dell Compellent vSphere Client Plug-in.

Boot from SAN

There is an ongoing discussion about whether or not to boot ESXi hosts from SAN. In some cases, such as with blade servers that do not have internal disk drives, booting from SAN may be the only option, but a lot of ESXi hosts can have internal mirrored drives giving the flexibility to choose. The benefits of booting from SAN are obvious. It alleviates the need for internal drives and allows 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 ESXi host, booting from local disks gives them the advantage of staying online if for some reason maintenance needs to be done to the fibre channel switches, ethernet switches, or the array itself. 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 ESXi, when booting from SAN RDMs couldn't be used, however since 3.x this behavior has changed. When booting from SAN with ESX(i) 3.x, 4.x, or 5.x, RDMs can also be utilized.

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

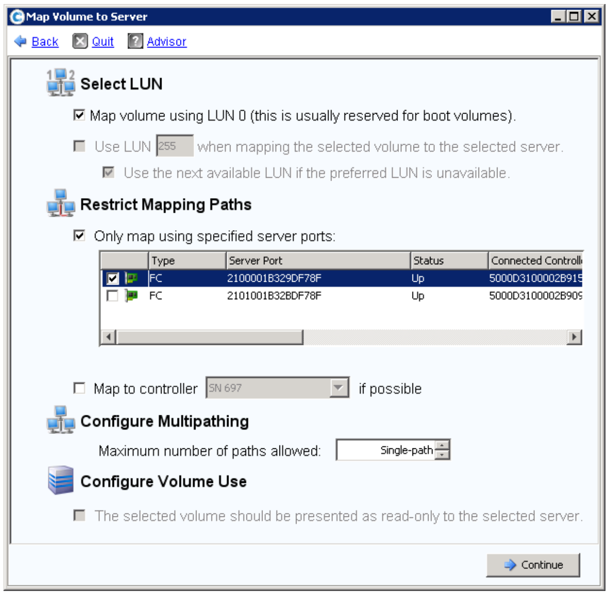
When deciding to boot ESXi hosts from SAN, there are a few best practices that need consideration.

When mapping the boot volume to the ESXi host for the initial install, the boot volume should only be mapped down a single path to a single HBA. Once ESXi has been loaded and multipath modules are operating correctly, the second path can then be added to the boot volume.

In Storage Center 4.x and earlier, when initially mapping the boot volume to the ESXi host, the mapping wizard will allow the selection of the individual paths, one must be sure to specify LUN 0. Once the ESXi host is up and running, use the same procedure to add the second path, making sure to rescan the host HBAs once the path has been added.

In Storage Center 5.x and later, entering the advanced mapping screen is needed to select a few options to force mapping down a single path.

Figure 14 - Advanced mapping screen for configuring boot from SAN



Select LUN

☑ Map volume using LUN 0 (this is usually reserved for boot volumes).

□ Use LUN 255 when mapping the selected volume to the selected server.

☑ Use the next available LUN if the preferred LUN is unavailable.

Restrict Mapping Paths

☑ Only map using specified server ports:

	Type	Server Port	Status	Connected Controll
☑	FC	2100001B329DF78F	Up	5000D3100002B915
□	FC	2101001B32BDF78F	Up	5000D3100002B909

□ Map to controller SN 697 if possible

Configure Multipathing

Maximum number of paths allowed: Single-path

Configure Volume Use

□ The selected volume should be presented as read-only to the selected server.

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

Once the ESXi host is up and is running correctly, the second path can then be added 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, the HBAs on the ESXi host can be rescanned.

Volume Creation and Sizing

Volume Sizing and the 64 TB Limit

Although the maximum size of a LUN that can be presented to ESXi 5.x has been increased to 64 TB, the general recommendation is to start with smaller and more manageable initial datastore sizes and expand them as needed. Remember that a datastore can easily be expanded to a larger size later, so it is a prudent idea to start with datastore sizes in the 500GB – 750GB range. Considering that 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, this will usually keep the datastore performing adequately.

The largest single extent 64TB VMFS-5 volume size is (64 * 1024 * 1024 * 1024 * 1024) bytes or 70368744177664 bytes. For any volumes larger than this size, VMware will simply not consume the additional space.

Note

The sizing recommendations are more about limiting the number of virtual machines on each datastore to keep performance manageable, than actual capacity reasons. If there is a virtual machine that requires a large disk, there is nothing wrong with creating a large datastore for it.

Virtual Machines per Datastore

Although there are no steadfast rules for how many virtual machines should placed on a datastore, due to the scalability enhancements of VMFS-5, a good conservative approach is to place anywhere between 15-25 virtual machines on each.

The reasoning behind keeping a limited number or Virtual Machines and/or VMDK files per datastore is due to potential I/O contention, queue depth contention, or SCSI reservation conflicts that may degrade system performance. That is also the reasoning behind creating 500GB – 750GB datastores, because this helps limit the number of virtual machines placed 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 can be 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 25 or more.

Since the appropriate number of virtual machines that can be put onto each datastore is subjective and dependent on the environment, a good recommendation is to start with 15 virtual machines, and increase/decrease the number of virtual machines on each datastore as needed. Moving virtual machines between datastores can even be done non-disruptively if licensed to use VMware’s Storage vMotion.

The 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 set limits thus increasing disk latency. 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 there are 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.)

A second common indicator that a datastore has too many virtual machines placed on it would be the frequent occurrence of “SCSI Reservation Conflicts” as seen when monitoring with esxtop. New to ESXi 5.x is a field in the Disk Device screen for Reserve Stats (RESVSTATS). That said, when monitoring it is normal to see a few reservations (RESV/s) entries and even a few conflicts (CONS/s) from time to time, but when you notice conflicts (CONS/s) happening very frequently on a particular volume, it may be time to move some of the virtual machines to a different datastore. The VAAI Hardware Assisted Locking primitive will help to alleviate the performance degradation caused by these SCSI-2 reservations, so if your datastores are impacted, is recommended to upgrade to a version of Storage Center firmware that supports this primitive. (See section on VAAI)

Note

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 the 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 testing is performed to determine the impact that an aligned partition may have on particular applications because all workloads are different.

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

Note

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.

Figure 15: 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

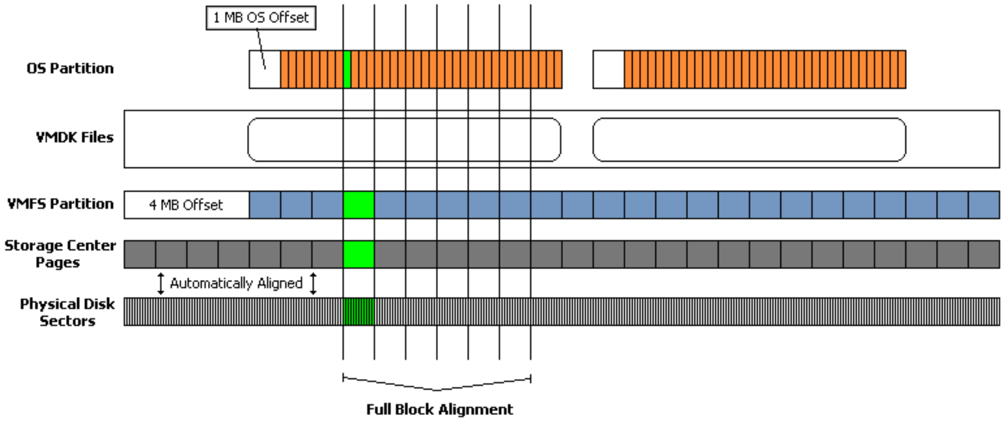
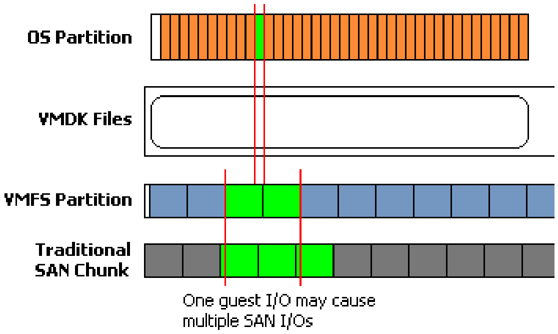


Figure 16: This is an example of an unaligned partition in a traditional SAN where performance can be improved by alignment



VMFS file systems and block sizes

Within ESXi 5.x, the choice can be made to use either the VMFS-3 or VMFS-5 file system formats which affect the block size of the datastore.

VMFS-3

With VMFS-3, choosing a block size for a datastore determines the maximum size of a VMDK file that can be placed on it. In other words, the block size should be chose based on the largest virtual disk planned to be put on the datastore. Choose this datastore file system version if backwards compatibility with ESX 4.x hosts is needed.

Table 3. VMFS Block Size chart

Block Size	Maximum VMDK Size
1 MB	256 GB
2 MB	512 GB
4 MB	1024 GB
8 MB	2048 GB – 512B

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

One should also consider future growth of the virtual machine disks, and the future upgrade to VMFS-5 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. Also remember that if a VMFS-3 datastore with a 2 MB block size is upgraded to VMFS-5, the block size remains at 2 MB which can cause VAAI to be disabled and the host to use the default DataMover instead.

Note

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 datastores. Please consult the vStorage APIs for Array Integration FAQ for more information.

VMFS-5

With VMFS-5 the only avaliable block size is 1 MB allowing for up to a 64 TB datastore, and up to a 2TB-512B VMDK. This format also allows the VAAI Dead Space Reclamation primitive (SCSI UNMAP) to reclaim storage after a VMDK is deleted (See section on VAAI). Keep in mind, this file system version is not backwards compatible with ESX 4.x hosts, so there are special considerations when using or migrating to this format.

For more information please consult the following references:

- Dell Compellent Document: "VMFS-3 to VMFS-5 Upgrade Guide"
  - Available from Knowledge Center

Where possible, our recommendation is to create new VMFS-5 datastores, and migrate virtual machines to them using storage vMotion.

Caution

Before upgrading a VMFS-3 datastore to VMFS-5, it is recommended that a Replay be taken of the datastore for protection against any possible loss of data.

LUN Mapping Layout

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 the 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 it is decided to layout the virtual machines; here are some basic concepts that need consideration:

Storage of non-virtual machine files

As a general recommendation, one or more VMFS datastores should be created for administrative items. This datastore can be used to store 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 datastore.

- Since pagefiles can generate a lot if disk activity when the memory in the virtual machine or ESXi host runs low, it could keep volume replays smaller.
- If 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 is to use the vSphere client performance charts to monitor Swap/Balloon usage of the ESXi host. If these numbers are high, consider testing the separation of pagefiles to determine the actual impact.

If decided 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 a volume is created that will contain 10 virtual machines, then a second a second volume should be created 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 the virtual machines.
- Create one “paired” datastore for the corresponding virtual machine pagefiles
- This should contain virtual machine pagefiles. Using Windows as an example, one 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 ESXi 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 that datastore becoming a single point of failure. 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 ESXi 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 ESXi 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

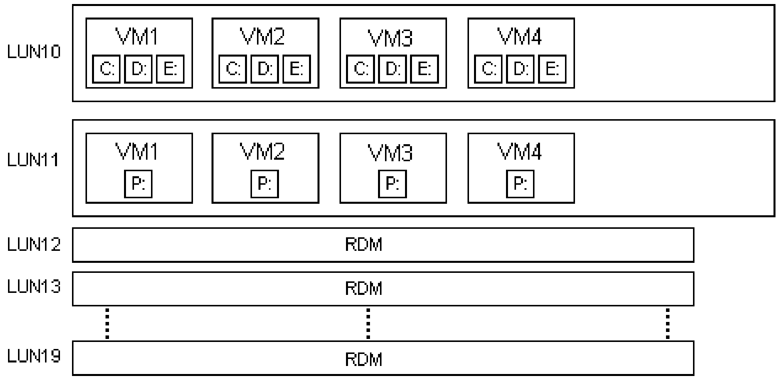
This example technique will give a great deal of flexibility when building out the storage architecture in the environment, while keeping with the basic concepts discussed above. The example layout below will meet most virtual infrastructure needs, because it adds the flexibility of being able to add 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 the virtual machines.

An example of this technique is as follows:
LUN0 - Boot LUN for ESXi (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

Figure 17: Virtual Machine Placement (With RDMs)



Timesaver

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

Note

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 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 business related factors, and may not be appropriate for all circumstances.

Advantages

- Granularity in replication
  - Since the Storage Center replicates at the volume level, if there is one virtual machine per volume, administrators can choose which virtual machine to replicate
- I/O contention is reduced because a single LUN is dedicated to a single 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
  - Storage usage and performance can be monitored for an individual virtual machine
- Backup/Restore of an entire virtual machine is simplified
  - If a VM needs to be restored, an administrator can just unmap/remap a replay in its place

Disadvantages

- There will be a maximum of 256 virtual machines in the ESXi cluster
  - The HBA has a maximum limit of 256 LUNs that can be mapped to the ESXi host, and since we can only use each LUN number once when mapping across multiple ESXi hosts, 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)

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:

- Virtual and Physical mode RDMs up to 64 TB in size can be mapped directly to a guest
  - A pRDM supports up to 64 TB max file size while a vRDM supports 2TB-512B max file size
- 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
  - Independent queue per RDM
- 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 Dell Compellent Knowledge Center
- The ability to assign a different data progression profile to each volume.
  - For example, if a database server has its 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
- Virtual mode RDMs support vSphere snapshots

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 ESXi host
  - If every virtual machine used RDM's for drives, the cluster would have a maximum number of 255 drives
- Physical mode RDMs cannot be used in conjunction with ESXi snapshots
  - While ESXi snapshots are not available for physical mode RDMs, Storage Center Replays can still be used to recover data

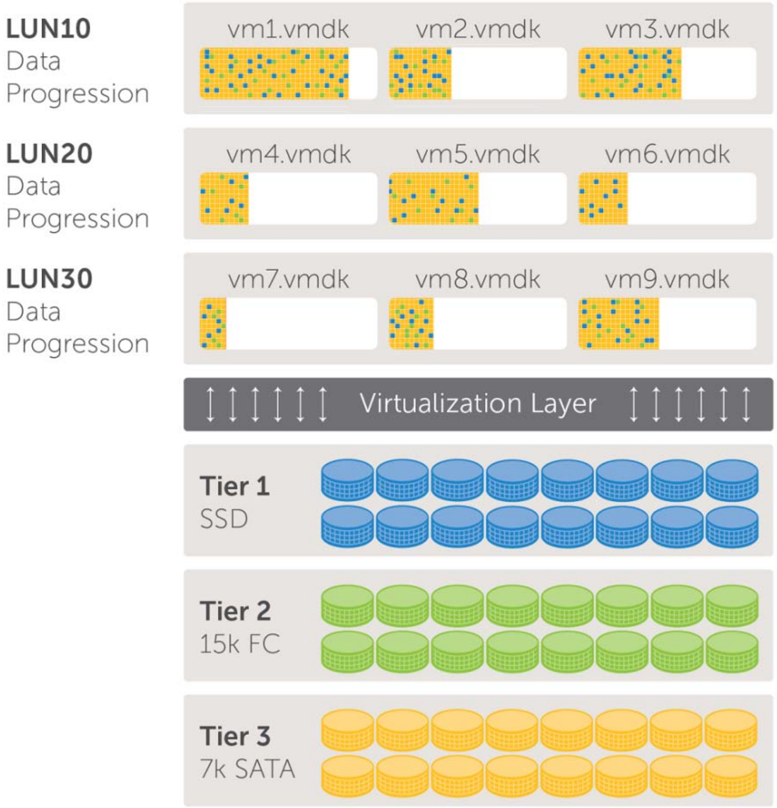
Note

All of the above RDM related tasks can be automated by using the Dell Compellent vSphere Client Plug-in. This can be downloaded from the Dell Compellent Knowledge Center.

Data Progression and RAID types

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.

Figure 18: Data Progression and thin provisioning working together



When using data progression, virtual machines that have completely different storage needs such as tier or RAID level, can be placed on the same datastore. This gives the administrator the ability to sort virtual machines by business purpose rather than disk performance characteristics.

However, if a business case is encountered where particular virtual machines would require different RAID types; some decisions on how Data Progression is configured for the volumes must be made. Below is an advanced example of possible situations where forcing particular volumes into different profiles is desired.

Here is an example of virtual machine RAID groupings:

LUN0 - Boot LUN for ESXi
-- 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 - LUN9 - 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 there is specific business needs that require a particular virtual machine or application to have a specific RAID type, our recommendation is to keep the configuration simple. In most cases, the Data Progression “Recommended” setting can be used to sort out the virtual machine data automatically by usage.

Note

A note about Data Progression Best Practices: A replay schedule for each volume should be created 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.

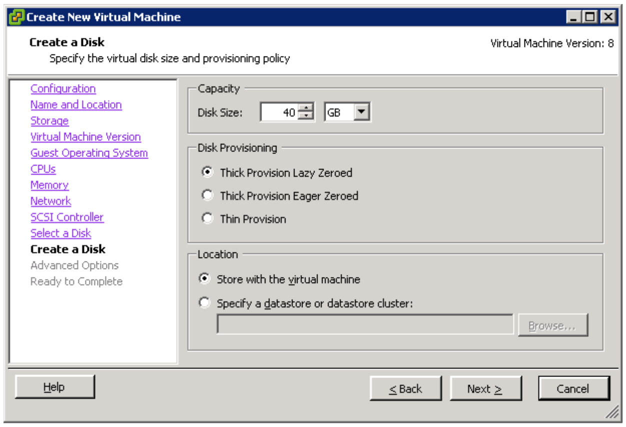
Thin Provisioning and VMDK Files

Dell Compellent's thin provisioning 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 ESXi 5.x, VMFS can create virtual disks using one of three different formats.

Figure 19: Virtual disk format selection



Create a Disk
Specify the virtual disk size and provisioning policy

Virtual Machine Version: 8

Capacity
Disk Size: 40 GB

Disk Provisioning
● Thick Provision Lazy Zeroed
○ Thick Provision Eager Zeroed
○ Thin Provision

Location
● Store with the virtual machine
○ Specify a datastore or datastore cluster:

Thick Provision Lazy Zeroed

(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, ESXi will first zero out the block, to ensure the secure 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.

Thick Provision Eager Zeroed

(a.k.a. "eagerzeroedthick")

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. (Note that when using VAAI, the time it takes to create an eagerzeroedthick disk is greatly reduced.) This format is generally used for Microsoft clusters, and the highest I/O workload virtual machines because it does not suffer from operational write penalties as the zeroedthick or thin formats.

Thin Provisioned

(a.k.a. "thin")

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.

Thin Provisioning Relationship

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

- Thick Provision Lazy Zeroed (zeroedthick)
  - Virtual disks will be thin provisioned by the Storage Center
- Thick Provision Eager Zeroed (eagerzeroedthick)
  - Virtual disks will be thin provisioned by the Storage Center (See the "Storage Center Thin Write Functionality" section below.)
- Thin Provisioned (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)

We recommend sticking with the default virtual disk format of Thick Provision Lazy Zeroed (zeroedthick) unless there are specific needs to pre-allocate virtual disk storage such as Microsoft clustering, VMware Fault Tolerance(FT), or for virtual machines that may be impacted by the thin or zeroedthick write penalties. If the application is sensitive to VMFS write penalties, it is recommended to test eagerzeroedthick virtual disks to determine the actual performance impact.

Storage Center Thin Write Functionality

Certain versions of Storage Center 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 Thick Provision Eager Zeroed (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 provisioned VMDK format. Since the Storage Center uses thin provisioning on all volumes by default, it is not necessary to use VMware' thin provisioning because there are no additional storage savings by doing so.

However, if VMware's thin provisioning is needed for whatever reason, one must pay careful attention not to accidentally overrun the storage allocated. To prevent any unfavorable situations, the built-in vSphere datastore threshold alerting capabilities should be used to warn against running out of space on a datastore. The threshold alerting capabilities of Dell Compellent Enterprise Manager can also be used to alert on low space conditions.

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, pretend there is a 20 GB data volume, where 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 "Dell 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 (pRDM).
    - 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 (vRDM).
    - 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 ESXi host snapshots.
        - This means that if software that uses VMware snapshots is needed, there will have be an alternative method of backing up the physical mode RDMs applied. 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 "Dell Compellent Enterprise Manager User Guide".

Extending VMware Volumes

Within an ESXi host, there are three ways to extend or grow storage. The general steps are listed below, but additional information can be found the following documentation pages:

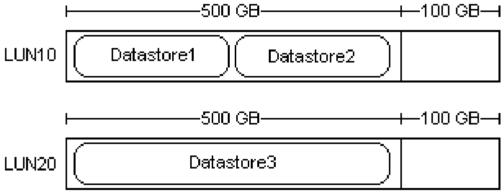
- VMware document: "vSphere Storage Guide"
  - Subsection: "Increase VMFS Datastores"
- VMware document: "vSphere Virtual Machine Administration Guide"
  - Subsection: "Change the Virtual Disk Configuration in the vSphere Client"

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.

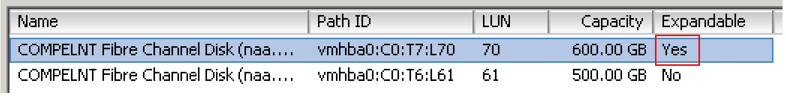
Figure 20: Datastore2 and Datastore3 can be grown by 100GB, but Datastore1 cannot



To extend the space at the end of a Storage Center volume as shown above, it can be done from the Storage Center GUI. After the volume has been extended and the hosts HBA has been rescanned, the properties of the datastore can be edited 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 the unintended action will actually add a VMFS "extent" to the datastore (see section below on VMFS extents).

Figure 21: Screenshot from the wizard after extending a 500GB datastore by 100GB



Name	Path ID	LUN	Capacity	Expandable
COMPELNT Fibre Channel Disk (naa.…	vmhba0:C0:T7:L70	70	600.00 GB	Yes
COMPELNT Fibre Channel Disk (naa.…	vmhba0:C0:T6:L61	61	500.00 GB	No

Warning

If a VMFS-3 volume (or pRDM residing on a VMFS-3 datastore) is extended beyond the 2 TB limits, that volume will become inaccessible by the ESXi host. If this happens, the most likely scenario will result in recovering data from a replay or tape.

Note

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.

Note

All of the above tasks can be automated by using the Dell Compellent vSphere Client Plug-in. This can be downloaded from the Dell Compellent Knowledge Center.

Adding a new extent to an existing datastore

This functionality is used to concatenate multiple Storage Center volumes into a single datastore. This feature is used to create VMFS-3 datastores larger than 2 TB. If a datastore larger than 2 TB is needed, it is highly recommended that VMFS-5 is used instead.

Warning

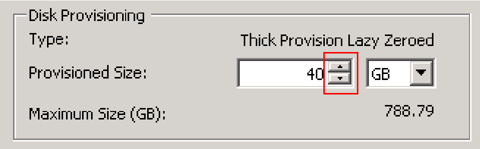
Due to the complexities of coordinating replays and recoveries of datastores that are spanned across multiple Storage Center volumes, the use of VMFS extents is highly discouraged. However, if the use of extents is needed, Replays of those volumes should be taken using the Consistent Replay Profile functionality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

Figure 22: Growing a virtual disk from the virtual machine properties screen



Disk Provisioning

Type: Thick Provision Lazy Zeroed

Provisioned Size: 40 GB

Maximum Size (GB): 788.79

For Windows machines: After growing the virtual disk from the vSphere client, an administrator must log into the virtual machine, rescan for new disks, and then use DISKPART or the disk management MMC console to extend the drive.

Warning

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, 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.

A physical mode RDM that resides on a VMFS-5 datastore can be extended up to the 64 TB limit, however vRDMs still have a 2TB-512B limit regardless of VMFS version.

Warning

Just as with datastore volumes, it is also very important not to extend an RDM volume residing on a VMFS-3 datastore 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 business needs, each solution is usually unique to each environment. Through testing and verification, it may be found that one solution works better 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 more information is needed about virtual machine backup strategies, please consult VMware's documentation pages.

Backing up virtual machines to tape or disk

Perhaps the most common methods of backing up virtual machines to tape are using backup client software installed within the guest, or by using a third party backup software.

- **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.
- **VMware Data Recovery**
  - This is a software package designed for smaller environments (usually less than 100 VMs per appliance) that allows operational backup and recovery of virtual machines to a datastore.
    - http://www.vmware.com/products/data-recovery/overview.html
- **Third party backup software using vStorage APIs for Data Protection (VADP)**
  - The vStorage APIs for Data Protection are the successor to VMware Consolidated Backup, and provide backup vendors an integrated method to backup virtual machines.
    - http://kb.vmware.com/kb/1021175

Backing 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, a replay profile can be created 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 Dell Compellent's Replay Manager Software**
  - Since virtual machines running transactional databases are more sensitive to crash consistent data, Dell Compellent has developed its Replay Manger 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 Dell Compellent's scripting tools**
    - For applications that need a custom method for taking consistent replays of the data, Dell Compellent has developed two scripting tools:
        - **Dell Compellent Command Utility (CompCU) -** This is a java based scripting tool that allows scripting for many of the Storage Center's tasks (such as taking replays).
        - **Storage Center Command Set for Windows PowerShell** – This scripting tool will also allow scripting for 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.
- **Replays used for Storage Center Replication and VMware Site Recovery Manager**
    - Replicating replays to a disaster recovery site not only ensures an off-site backup, but in addition when using Site Recovery Manager, provides an automated recovery of the virtual infrastructure in the event a disaster is declared.

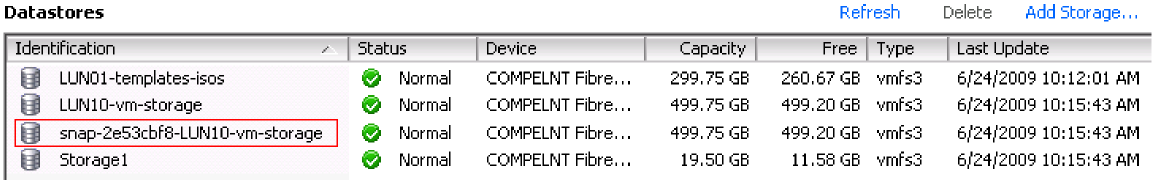
Recovering Virtual Machine Data from a Replay

When recovering a VMFS datastore from a replay, an admin 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 to recover from 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 ESXi host designated 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, highlight an ESXi host, then select the 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 that was just mapped to the host and then click "Next".
    e. Three options are presented:
        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. (Select this option if 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:

Figure 23: The storage configuration tab showing snapshot datastore



Identification	Status	Device	Capacity	Free	Type	Last Update
LUN01-templates-isos	Normal	COMPELNT Fibre...	299.75 GB	260.67 GB	vmfs3	6/24/2009 10:12:01 AM
LUN10-vm-storage	Normal	COMPELNT Fibre...	499.75 GB	499.20 GB	vmfs3	6/24/2009 10:15:43 AM
snap-2e53cbf8-LUN10-vm-storage	Normal	COMPELNT Fibre...	499.75 GB	499.20 GB	vmfs3	6/24/2009 10:15:43 AM
Storage1	Normal	COMPELNT Fibre...	19.50 GB	11.58 GB	vmfs3	6/24/2009 10:15:43 AM

6. The recovery datastore is now designated with “snap-xxxxxxxx-originalname”
7. From here the datastore can be browsed to perform the recovery via one of the methods listed below.

Note

All of the above tasks can be automated by using the recovery functionality in the Dell Compellent vSphere Client Plug-in. This can be downloaded from the Dell Compellent Knowledge Center.

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. Now a drive letter can be assigned to the virtual disk, and recover/copy/move the file back to its original location.

After completing the file recovery, it is important that the recovered virtual disk be removed 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 to be recovered, right click, and select “Move to”. Following through the wizard, browse to the destination datastore and folder, then click “Move”. If a vmdk file is being moving back to its original location, remember that the virtual machine must be powered off 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.

Note

Alternatively, the old VMDK can be removed from the virtual machine, the recovered virtual disk re-added to the virtual machine, and then use Storage vMotion to move the virtual disk back to the original datastore while the VM is powered on.

Recovering an entire virtual machine

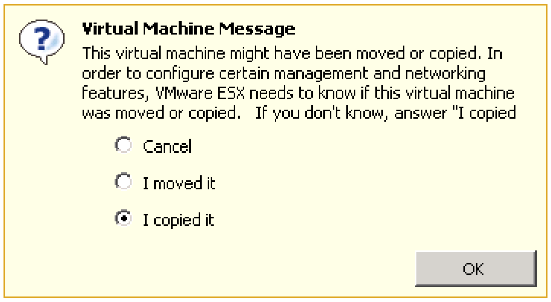
To recover an entire virtual machine from the snap datastore, browse to the virtual machine configuration file (*.vmx), right click, then select add to inventory. Follow through the wizard to add the virtual machine into inventory.

Caution

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 the one of the virtual machines onto an isolated network or private vSwitch.

If virtual center detects a duplicate UUID, it will prompt with the following virtual machine message:

Figure 24: Virtual Machine Question prompting for appropriate UUID action



Virtual Machine Message

This virtual machine might have been moved or copied. In order to configure certain management and networking features, VMware ESX needs to know if this virtual machine was moved or copied. If you don't know, answer "I copied

○ Cancel

○ I moved it

● I copied it

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 unsure of which option to chose, select "I copied it", this will regenerate a new MAC address to prevent conflicts on the network.

Note

Once the virtual machine has been recovered, it can be migrated back to the original datastore using Storage vMotion, provided that the original virtual machine files have been removed to free enough space.

Replication and Remote Recovery

Storage Center replication in coordination with the vSphere 5.x line of products can provide a robust disaster recovery solution. Since each different replication method effects recovery a little differently, choosing the correct method to meet 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 the 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 the system 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 more flexible bandwidth requirements making this 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 a volume is replicated 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".

Block changes are not replicated bidirectionally with standard replication, this means that the ability to vMotion virtual machines between source controllers (main site), and destination controllers (DR site), is not possible. That being said, there are a few best practices to replication and remote recovery that should be considered.

- Compatible ESXi host hardware is needed at the DR site in which to map replicated volumes in the event the source ESXi cluster becomes inoperable.
- Preparations should be made to have all of the Virtual Center resources replicated to the DR site as well.
- To keep replication sizes smaller, the operating system pagefiles should be separated 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 ESXi hosts must be gigabit or greater, round trip latency must be 10 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 testing is recommended to determine the bandwidth requirements for the 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 virtual machines are sorted into specific replicated and non-replicated volumes. For example, if there are 30 virtual machines in the ESXi cluster, and only 8 of them need to be replicated to the 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 is not replicated. 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, an administrator can also take advantage of the Storage Center QOS to prioritize replication bandwidth of certain volumes. For example, if there is a 100 Mb pipe between sites, two QOS definitions can be created 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, the same general steps as outlined in the previous section titled "Recovering Virtual Machine Data from a Replay" should be followed.

Timesaver

If the environment has a significant number of volumes to recover, time can be saved during the recovery process by using the "Replication Recovery" functionality within Dell Compellent's Enterprise Manager Software. These features will allow an admin to pre-define the recovery settings 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.

Caution

It is extremely important that the destination volume, usually denoted by "Repl of", never gets directly mapped to an ESXi 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 Dell Compellent Knowledge Center for more info.

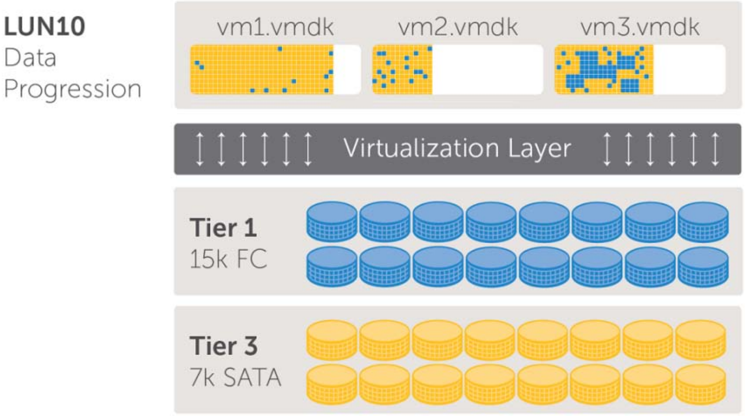
VMware Storage Features

Storage I/O Controls (SIOC)

SIOC is a feature that was added in ESX(i) 4.1 to help VMware administrators regulate storage performance and provide fairness across hosts sharing a LUN. Due to factors such as Data Progression and the fact that Storage Center uses a shared pool of disk spindles, it is recommended that extreme caution is exercised when using this feature. Due to how Data Progression migrates portions of volumes into different storage tiers and RAID levels, this could ultimately affect the latency of the volume, and trigger the resource scheduler at inappropriate times.

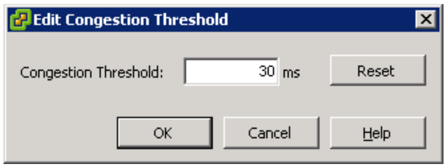
For example, if a datastore contains multiple virtual disks (see figure below), where each virtual disk may have different portions of blocks in tier 1 and tier 3. If VM1 begins to read a lot of archived data from vm1.vmdk residing on tier 3 disks, thereby increasing the latency of the datastore above the congestion threshold, the scheduler could activate and throttle vm3.vmdk although most of its blocks reside on a separate tier of spindles.

Figure 25: Multiple virtual disks with blocks residing in multiple tiers



The default setting for the congestion threshold is 30 milliseconds of latency. It is recommended that this value is left at default unless testing under the guidance of VMware or Dell Compellent Copilot Support.

Figure 1-26: Setting the congestion threshold


Edit Congestion Threshold

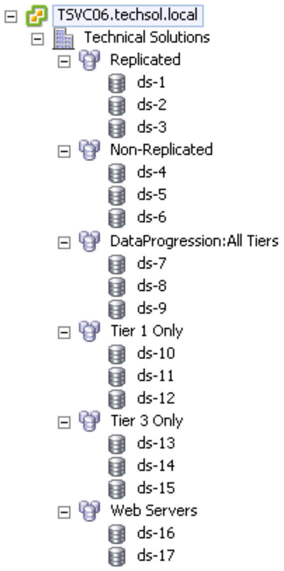
Congestion Threshold: 30 ms

For more information please refer to the following documentation:

- VMware document: "vSphere Resource Management Guide"
- VMware Knowledge base article: "External I/O workload detected on shared datastore running Storage I/O Control (SIOC) for congestion management"
    - http://kb.vmware.com/kb/1020651

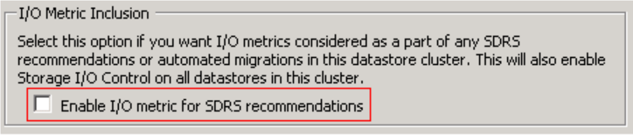
Storage Distributed Resource Scheduler (SDRS)

Storage DRS is a new feature in ESXi 5.0 that will automatically load balance virtual machines within a datastore cluster based on capacity and/or performance. When creating SDRS datastore clusters with Storage Center, it is important to remember a few guidelines.



- Group datastores with like disk characteristics. For example: replicated, non-replicated, storage profile, application, or performance
- Use Storage DRS for initial placement based on capacity. When placing virtual machines on the datastore cluster, it will place them based on which datastore has the most space available.
- Set the Automation Level to Manual Mode. This means that Storage DRS will make recommendations about moving virtual machines, but will not move them automatically. It is suggested that SDRS be run in manual mode to examine all recommendations before applying them. There are a few items to keep in mind before applying the recommendations:
    - Virtual machines that are moved between datastores automatically, will essentially have their data progression history reset. Depending on the storage profile settings, this could mean that the entire virtual machine gets moved back to Tier 1 RAID 10 if the volume was set to use the "Recommended" storage profile.
    - When a virtual machine moves between datastores, its actual location at the time the replay was taken may make the virtual machine harder to find during a recovery. For example, if a virtual machine moves twice in one week while daily replays are being taken, the appropriate volume that contains the good replay of the virtual machine may be difficult to locate.
    - If the Storage Center version in use does not support VAAI, the move process could be slow (without Full Copy) or leave storage allocated on its originating datastore (without Dead Space Reclamation). See the section on VAAI for more information.
- Storage DRS could produce incorrect recommendations for virtual machine placement when I/O metric inclusion is enabled on a Storage Center system using Data Progression. This is because when a datastore is inactive, the SIOC "injector" will perform random read tests to determine latency statistics of the datastore. With Data Progression enabled, the blocks that SIOC reads to determine datastore performance, could potentially reside on SSD, 15K, or even 7K drives. This could ultimately skew the latency results and decrease the effectiveness of the SRDS recommendation.

Figure 27: Disabling I/O Metric Inclusion



I/O Metric Inclusion

Select this option if you want I/O metrics considered as a part of any SDRS recommendations or automated migrations in this datastore cluster. This will also enable Storage I/O Control on all datastores in this cluster.

□ Enable I/O metric for SDRS recommendations

To reiterate, it is our recommendation that Storage DRS be only used for Capacity recommendations, set to Manual Automation mode, and the I/O Metric Inclusion be disabled. This will allow administrators to take advantage of SDRS capacity placement recommendations, while still allowing Data Progression to manage the performance of the data at the block level.

vStorage APIs for Array Integration (VAAI)

Within ESXi 5.0, there are now five primitives that the ESXi host can use to offload specific storage functionality to the array. These are all enabled by default in ESXi 5.x because they are now all based on new standardized SCSI-3 commands.

Block Zeroing (SCSI WRITE SAME)

Traditionally, when an ESXi host would create an Eager Zeroed Thick virtual disk, it would transmit all of the zeros over the SAN to the array, which consumed bandwidth, CPU, and disk resources to transmit the zeros across the wire. The Block Zeroing primitive uses the SCSI WRITE SAME command to offload the heavy lifting to the array. For example, when the ESXi host needs a 40 GB VMDK to be zeroed out, it can simply ask the array to write 40 GB worth of zeros, and have the array respond when finished. In the case of the Storage Center, since the array does not write zeros due to its Thin Write feature, this means that creating an Eager Zeroed Thick virtual disk only takes seconds instead of minutes.

Full Copy (SCSI EXTENDED COPY)

This primitive is used to offload to the array the copying or movement of blocks. It does this by taking advantage of a SCSI command called EXTENDED COPY that allows the ESXi host to instruct the Storage Center on which blocks it needs copied or moved, leaving the heavy lifting to the array itself. This has a large performance benefit to virtual machine cloning and storage vMotions because the ESXi host no longer has to send that data over the wire. The ESXi host simply tells the Storage Center which blocks need to be copied/moved and the array replies back when it is finished.

Hardware Accelerated Locking (ATS)

This primitive is intended to add scalability to the VMFS file system by removing the need for SCSI-2 reservations. With VMFS, when an ESXi host needs exclusive access to a LUN to update metadata, traditionally it will set a SCSI-2 non-persistent reservation on the LUN to guarantee it has exclusive rights. Typically during this operation, VMware has documented small performance degradation to other virtual machines accessing this LUN at the same time from different hosts. With the new hardware accelerated locking primitive, it uses a new SCSI-3 method called Atomic Test & Set (ATS) that greatly minimizes any storage performance degradation during the operation. This primitive was added to migrate away from the SCSI-2 operations to SCSI-3 to increase the future scalability of VMFS.

Dead Space Reclamation (SCSI UNMAP)

This primitive allows enhanced integration with arrays offering thin provisioning. Traditionally, in the Storage Center, when a VMDK was deleted, the array had no idea that the blocks were no longer in use, and thus, they couldn’t be freed back into the pagepool to be re-used. This caused an effect dubbed “the high water mark”, where blocks of data could still be consumed by the array that the operating system no longer needed. The Dead Space Reclamation primitive uses the SCSI UNMAP command to allow the ESXi host to instruct the array when specific blocks are no longer in use, so that the array can return them to the pool to be re-used. This means that if a VMDK is moved or deleted, those blocks will be returned to the pagepool, with two notable exceptions. First, if those blocks are frozen as part of a replay, they will not return to the pagepool until that replay has expired. Secondly, dead space reclamation doesn’t work within the VMDK, and thus only works at the VMFS level. This means that if a large file is deleted from within a VMDK the space wouldn’t be returned to the pagepool unless the VMDK itself was deleted.

Thin Provisioning Stun

This primitive allows the Storage Center to send a special SCSI sense code back to the ESXi host when there is an Out of Space (OOS) condition, allowing ESXi to pause only the virtual machines which are requesting additional pages until additional storage is added to remedy the situation.

Warning

At the time of this writing, the only primitive that has been certified for use with the Storage Center is the Block Zeroing primitive. Until the remaining VAAI primitives have been completed and fully certified with Storage Center 6.x, it is recommended that an admin disable the uncertified VAAI primitives as a precautionary measure.

For more information please see the following resources:

- Dell Compellent Knowledge Center: “Copilot Support Technical Advisory (CSTA)”
    - Title: “CSTA - Disabling Space Reclamation in ESXi v5.0”
- VMware document: “vSphere Storage Guide”
    - Subsection: “Disable Space Reclamation”

Conclusion

This document addresses many of the best practices scenarios an administrator may encounter while implementing or upgrading VMware vSphere with the Dell Compellent Storage Center.

More information

If more information is needed, please review the following web sites:

- Dell Compellent
  - General Web Site
    - http://www.compellent.com/
  - Knowledge Center
    - http://kc.compellent.com
  - Training
    - http://www.compellent.com/Support-Services/Services/Training.aspx

- VMware
  - General Web Site
    - http://www.vmware.com/
  - VMware Education and Training
    - http://mylearn1.vmware.com/mgrreg/index.cfm
  - vSphere Online Documentation
    - http://www.vmware.com/support/pubs/vsphere-esxi-vcenter-server-pubs.html
  - VMware Communities
    - http://communities.vmware.com

Getting Help

Contacting Copilot Support

For customers in the United States:

- Telephone: 866-EZ-STORE (866-397-8673)
- E-mail support@compellent.com

Additional Copilot Support contact information for other countries can be found at:

- http://www.compellent.com/Support-Services/Support/Copilot-Contact.aspx

Appendixes

Appendix A - Determining the appropriate queue depth for an ESXi host

Adjusting the queue depth on ESXi hosts is a very complicated subject. On one hand, increasing it can remove bottlenecks and help to improve performance (as long as there are enough back end spindles to handle the incoming requests). Yet on the other hand, if set improperly, the ESXi 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 to achieve an acceptable number of IOPS from the back end spindles, while at the same time, not setting it too high allowing an ESXi host to flood the front or back end of the array.

Here are a few basic pointers:

- Fiber Channel
    - 2 Gb Storage Center Front-End Ports
        - Each 2 Gb FE port has a max queue depth of 256 so care must be taken to not overdrive it
        - It is generally best to leave the ESXi queue depths set to default and only increase if absolutely necessary
        - Recommended settings for controllers with 2 Gb FE ports
        - HBA BIOS = 255
        - HBA Queue depth is actually regulated by the driver module
        - Driver module = 32 (Default)
        - Disk.SchedNumReqOutstanding = 32 (Default)
        - Guest vSCSI controller = 32 (Default)
    - 4/8 Gb Storage Center Front-End Ports
        - Each 4/8 Gb 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 ESXi 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 Gb FE ports
        - HBA BIOS = 255
        - Driver module = 255
        - Disk.SchedNumReqOutstanding = 32 (Default)
        - Increase/decrease as necessary
        - Guest vSCSI controller = 32 (Default)
        - Increase/decrease as necessary

- iSCSI
    - Software iSCSI
        - Leave the queue depth set to default and only increase if necessary
    - Hardware iSCSI
        - Leave the queue depth set to default and only increase if necessary

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

- ESXi Shell via SSH
    - Command: esxtop

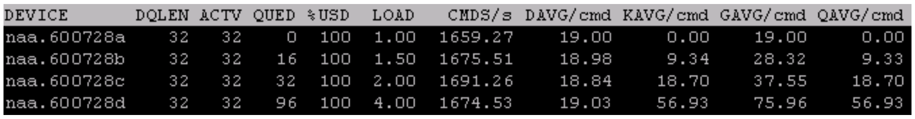
- vCLI 5.x or the vMA Virtual Appliance
  - o Command: resxtop (or resxtop.sh)

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:

1. From the command line type either:
   a. # esxtop
   b. # resxtop.sh --server esxserver.domain.local
      i. Enter appropriate login credentials
2. Enter the “Disk Device” screen by pressing “u”
3. Expand the “devices” field by pressing “L 36 <enter>” (Capital “L”)
   a. This will expand the disk devices so that LUNs can be identified by naa identifier
4. Chose the “Fields” to monitor by pressing “f”:
   a. Press “b” to uncheck the ID field (not needed)
   b. OPTIONALLY: (Depending on preference)
      i. Check or uncheck “i” for overall Latency
      ii. Check “j” for read latency
      iii. Check “k” for write latency
   c. Press <enter> to return to the monitoring screen
5. Set the refresh time by pressing “s 2 <enter>”. (Refresh every 2 seconds)

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

Figure A-1: esxtop with a queue depth of 32 (Edited to fit screen)

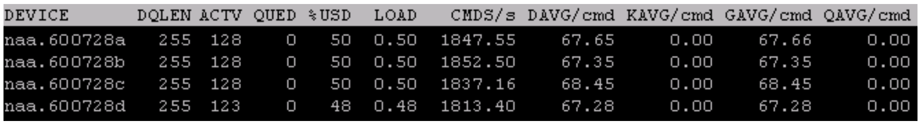


```
DEVICE       DQLEN ACTV QUED %USD  LOAD   CMDS/s DAVG/cmd KAVG/cmd GAVG/cmd QAVG/cmd
naa.600728a     32   32    0  100  1.00  1659.27    19.00     0.00    19.00     0.00
naa.600728b     32   32   16  100  1.50  1675.51    18.98     9.34    28.32     9.33
naa.600728c     32   32   32  100  2.00  1691.26    18.84    18.70    37.55    18.70
naa.600728d     32   32   96  100  4.00  1674.53    19.03    56.93    75.96    56.93
```

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 figure above, the device queue depth is set to 32. Please note that 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, as well as increase the Disk.SchedNumReqOutstanding setting.

Figure A-2: queue depth increased to 255 (Edited to fit screen)



```
DEVICE       DQLEN ACTV QUED %USD  LOAD   CMDS/s DAVG/cmd KAVG/cmd GAVG/cmd QAVG/cmd
naa.600728a    255  128    0   50  0.50  1847.55    67.65     0.00    67.66     0.00
naa.600728b    255  128    0   50  0.50  1852.50    67.35     0.00    67.35     0.00
naa.600728c    255  128    0   50  0.50  1837.16    68.45     0.00    68.45     0.00
naa.600728d    255  123    0   48  0.48  1813.40    67.28     0.00    67.28     0.00
```

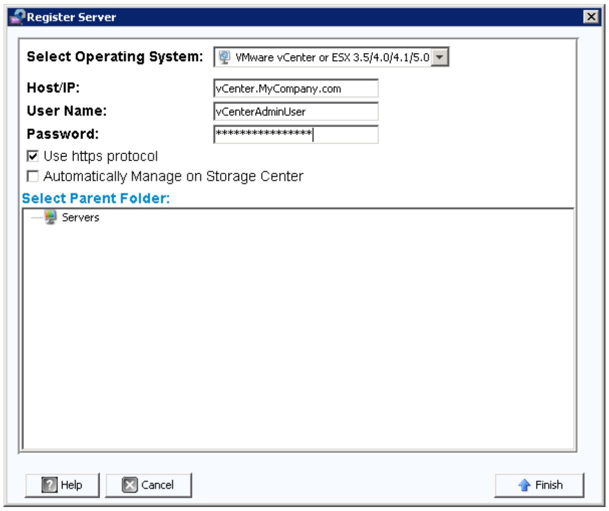
Please note that by increasing the queue depth from the previous example, the total IOPS increased from 6700 to 7350, but the average device latency (DAVG/cmd) increased from 18ms to 68ms. That means the latency over tripled for a mere 9% performance gain. In this case, it may not make sense to increase 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 Monitoring and Performance Guide".

Appendix B - Configuring Enterprise Manager VMware Integrations

For customers that are running Enterprise Manager Versions 5.5.3 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

Select Operating System: VMware vCenter or ESX 3.5/4.0/4.1/5.0

Host/IP: vCenter.MyCompany.com

User Name: vCenterAdminUser

Password: ****************

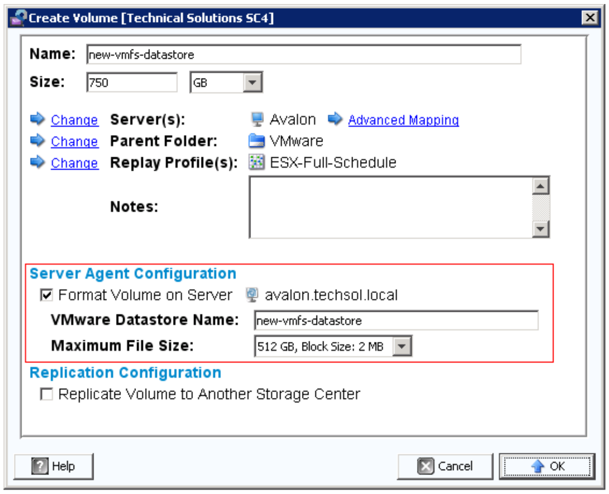
☑ Use https protocol

□ Automatically Manage on Storage Center

Select Parent Folder:

Servers

After entering vCenter credentials, administrators can see aggregate storage statistics, as well as being able to automatically provision VMFS datastores and RDMs. For example, when creating a new volume, by selecting an ESXi host, it will automatically give the option to format it with VMFS.



Create Volume [Technical Solutions SC4]

Name: new-vmfs-datastore

Size: 750 GB

Change Server(s): Avalon Advanced Mapping

Change Parent Folder: VMware

Change Replay Profile(s): ESX-Full-Schedule

Notes:

Server Agent Configuration

☑ Format Volume on Server avalon.techsol.local

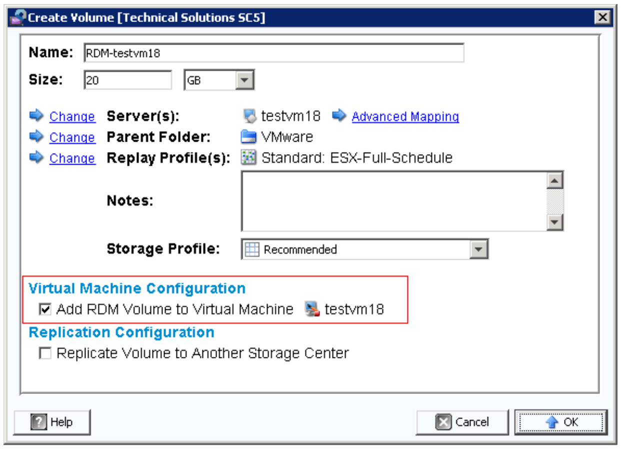
VMware Datastore Name: new-vmfs-datastore

Maximum File Size: 512 GB, Block Size: 2 MB

Replication Configuration

□ Replicate Volume to Another Storage Center

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



Create Volume [Technical Solutions SC5]

Name: RDM-testvm18

Size: 20 GB

Server(s): testvm18

Parent Folder: VMware

Replay Profile(s): Standard: ESX-Full-Schedule

Notes:

Storage Profile: Recommended

Virtual Machine Configuration

☑ Add RDM Volume to Virtual Machine testvm18

Replication Configuration

☐ Replicate Volume to Another Storage Center