

# Dell EMC SC Series Best Practices with VMware vSphere 5.x–6.x

Dell EMC Engineering February 2017

# Revisions

Date	Description	
July 2016	Initial release: combined vSphere 5.x and 6.x best practice documents and added SCOS 7.1 updates	
September 2016	nor revisions and corrections	
October 2016	Changed Disk.AutoremoveOnPDL to reflect current VMware guidance	
January 2017	Updated for vSphere 6.5 changes; added appendix D summarizing all host settings	
February 2017	Updated Linux guest disk timeout recommendations in section 4.7.2	

# Acknowledgements

Author: Darin Schmitz

The information in this publication is provided "as is." Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Copyright © 2017 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be the property of their respective owners. Published in the USA. [2/16/2017] [Best Practices Guide] [2060-M-BP-V]

Dell believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

# Table of contents

Re	visions	5	2
1	Introc	Juction	7
	1.1	Audience	7
	1.2	Prerequisites	7
2	Fibre	Channel switch zoning	8
	2.1	Single initiator multiple target zoning	8
	2.2	WWN zoning	8
	2.3	Port zoning	8
	2.4	Virtual ports	9
3	Host	initiator settings	10
4	Modif	fying queue depth and timeouts	11
	4.1	Host bus adapter queue depth	11
	4.2	Storage driver queue depth and timeouts	12
	4.2.1	Fibre Channel HBAs	12
	4.2.2	Software iSCSI initiator	13
	4.2.3	SAS HBAs	13
	4.3	Adjusting settings for permanent device loss conditions	14
	4.4	Modifying the VMFS queue depth for virtual machines (DSNRO)	14
	4.5	Adaptive queue depth	17
	4.6	Modifying the guest OS queue depth	17
	4.6.1	Windows Server 2008, 2008 R2, or 2012	17
	4.7	Setting operating system disk timeouts	18
	4.7.1	Windows	19
	4.7.2	Linux	19
5	Gues	t virtual SCSI adapter selection	20
	5.1	BusLogic Parallel	20
	5.2	LSI Logic Parallel	20
	5.3	LSI Logic SAS	20
	5.4	VMware Paravirtual	20
6	Марр	ing volumes to an ESXi server	21
	6.1	Basic volume mapping concepts	21
	6.2	Basic SC Series volume mappings	21
	6.3	Multipathed volume concepts	22
	6.4	Multipathed SC Series volumes	23

6.5	Configuring the VMware iSCSI software initiator for a single path	25
6.6	Configuring the VMware iSCSI software initiator for multipathing	26
6.6.1	VMkernel port binding considerations with software iSCSI	26
6.7	iSCSI port multi-VLAN configuration recommendations	28
6.8	Configuring the FCoE software initiator for multipathing	29
6.9	VMware multipathing policies	29
6.9.1	Round robin policy	30
6.9.2	Fixed policy	33
6.9.3	Most recently used policy	34
6.10	Multipathing using a fixed path selection policy	34
6.11	Multipathing using a round robin path selection policy	35
6.12	Asymmetric logical unit access (ALUA)	35
6.13	Unmapping volumes from an ESXi host	36
6.14	Mapping volumes from multiple arrays	37
6.15	Additional multipathing resources	37
Boot	from SAN	38
7.1	Configuring boot from SAN	38
Volun	ne creation and sizing	40
8.1	Volume sizing and the 64 TB limit	40
8.2	Virtual machines per datastore	40
8.3	VMFS partition alignment	41
8.4	VMFS file systems and block sizes	43
8.4.1	VMFS-3	43
8.4.2	VMFS-5	43
8.4.3	VMFS-6	43
Volun	ne mapping layout	44
9.1	Multiple virtual machines per volume	44
9.1.1	Storage of non-virtual machine files	44
9.1.2	Separation of operating system page files	44
9.1.3	Separation of virtual machine swap files	45
9.1.4	Virtual machine placement	45
9.2	One virtual machine per volume	47
Raw	device mapping (RDM)	48
Data	Progression and storage profile selection	49
11.1	On-Demand Data Progression	50
	6.6 6.6.1 6.7 6.8 6.9 6.9.1 6.9.2 6.9.3 6.10 6.11 6.12 6.13 6.14 6.15 Boot 1 7.1 Volun 8.1 8.2 8.3 8.4 8.4.1 8.4.2 8.4.3 Volun 9.1 9.1.1 9.1.2 9.1.3 9.1.4 9.2 Raw o Data	6.6       Configuring the VMware ISCSI software initiator for multipathing.         6.1       VMkernel port binding considerations with software ISCSI         6.7       ISCSI port multi-VLAN configuration recommendations.         6.8       Configuring the FCoE software initiator for multipathing.         6.9       VMware multipathing policies.         6.9.1       Round robin policy.         6.9.2       Fixed policy.         6.10       Multipathing using a fixed path selection policy.         6.11       Multipathing using a fixed path selection policy.         6.12       Asymmetric logical unit access (ALUA).         6.13       Unmapping volumes from an ESXi host.         6.14       Mapping volumes from multiple arrays.         6.15       Additional multipathing resources.         Boot from SAN.

	11.2 Data reduction (compression and deduplication)	51
	11.2.1 Data Reduction Input	52
12	Thin provisioning and virtual disks	53
	12.1 Virtual disk formats	53
	12.1.1 Thick provision lazy zeroed (zeroedthick)	53
	12.1.2 Thick provision eager zeroed (eagerzeroedthick)	53
	12.1.3 Thin provisioned (thin)	54
	12.1.4 Space efficient sparse (SE sparse disks)	54
	12.2 Thin provisioning relationship	54
	12.3 SC Series thin write functionality	54
	12.4 SC Series thin provisioning or VMware thin provisioning	54
	12.5 Windows free space recovery	55
13	Extending VMware volumes	56
	13.1 Increasing the size of VMFS datastores	56
	13.1.1 Expanding an extent in an existing VMFS datastore	56
	13.1.2 Adding a new extent to an existing datastore	57
	13.2 Increasing the size of a virtual machine disk (VMDK) file	57
	13.3 Increasing the size of a raw device mapping (RDM)	57
14	Snapshots (replays) and virtual machine backups	58
	14.1 Backing up virtual machines	58
	14.1.1 Backing up virtual machines to tape or disk	58
	14.1.2 Backing up virtual machines using snapshots	58
	14.2 Recovering virtual machine data from a snapshot	59
	14.2.1 Recovering a file from a virtual disk	60
	14.2.2 Recovering an entire virtual disk	60
	14.2.3 Recovering an entire virtual machine	61
15	Replication and remote recovery	62
	15.1 Synchronous replication	62
	15.2 Asynchronous replication	62
	15.3 Replication considerations with standard replications	63
	15.4 Replication considerations with Live Volumes	63
	15.4.1 Asynchronous Live Volume	63
	15.4.2 Synchronous Live Volume	64
	15.4.3 Live Volume auto failover	64
	15.5 Replication tips and tricks	65

	15.6 Virtual machine recovery at a DR site	65
16	6 VMware storage features	
	16.1 Storage I/O Controls (SIOC)	
	16.1.1 SIOC and volume QoS	67
	16.2 Storage distributed resource scheduler (SDRS)	
	16.3 vStorage APIs for array integration (VAAI)	
	16.3.1 Block zeroing (SCSI WRITE SAME)	
	16.3.2 Full copy (SCSI EXTENDED COPY)	
	16.3.3 Hardware accelerated locking	69
	16.3.4 Dead space reclamation (SCSI UNMAP)	69
	16.3.5 Thin provisioning stun	
	16.4 vStorage APIs for Storage Awareness (VASA)	71
	16.5 Virtual Volumes (VVols)	71
A	Determining the appropriate queue depth for an ESXi host	73
	A.1 Fibre Channel	73
	A.2 iSCSI	74
	A.3 Using esxtop to monitor queue depth	74
В	Deploying vSphere client plug-ins	76
	B.1 SC Series vSphere client plug-in 1.7	76
	B.2 Dell Storage vSphere Web Client Plugin 4.x	76
С	Configuring Dell Storage Manager VMware integrations	
D	Host and cluster settings	
	D.1 Recommended settings	79
	D.2 Optional settings	80
Е	Additional resources	81
	E.1 Technical support and resources	81
	E.2 VMware support	81

# 1 Introduction

This document provides configuration examples, tips, recommended settings, and other storage guidelines for integrating VMware<sup>®</sup> vSphere<sup>®</sup> 5.x and 6.x hosts with the Dell EMC<sup>™</sup> SC Series storage. It also provides answers to many frequently asked questions with regard to how VMware interacts with various SC Series features such as Dynamic Capacity (thin provisioning), Data Progression (automated tiering), and Remote Instant Replay (replication).

#### 1.1 Audience

This highly technical document is intended for storage and server administrators, as well as other information technology professionals interested in learning more about how VMware vSphere integrates with SC Series storage.

#### 1.2 Prerequisites

Understanding the material in this document requires formal training or advanced working knowledge of the following:

- Installation and configuration of VMware vSphere 5.x/6.x
- Configuration and operation of the SC Series
- Operation of Dell Storage Manager (DSM)/Enterprise Manager (EM) software
- Using operating systems such as Microsoft<sup>®</sup> Windows<sup>®</sup> or Linux<sup>®</sup>

For important information about configuring ESXi hosts to use the SAN, refer to the *vSphere Storage Guide* located on <u>VMware vSphere 6.x Documentation</u> or <u>VMware vSphere 5.x Documentation</u>.

**Note:** This document provides general recommendations that may not be applicable to all configurations or needs.

# 2 Fibre Channel switch zoning

Zoning Fibre Channel switches for an ESXi host is similar to zoning any other server connected to the SC Series array. The fundamental points are explained in this section.

## 2.1 Single initiator multiple target zoning

Each Fibre Channel zone created should have a single initiator (HBA port) and multiple targets (SC Series front-end ports). This means that each HBA port requires its own Fibre Channel zone that contains itself and the SC Series front-end ports. Independent zones should be created for each HBA installed in the host.

## 2.2 WWN zoning

When zoning by WWN, the zone only needs to contain the host HBA port and the SC Series front-end primary ports. In most cases, it is not necessary to include the SC Series front-end reserve ports because they are not used for volume mappings. For example, if the host has two HBAs connected to two disjointed fabrics, the Fibre Channel zones would look similar to the configuration shown in Table 1.

Name	WWN	Description
ESX1-HBA1	2100001B32017114	ESX1 HBA port 1
(zone created in fabric 1)	5000D31000036001	Controller1 front-end primary plugged into fabric
	5000D31000036009	Controller2 front-end primary plugged into fabric 1
ESX1-HBA2	210000E08B930AA6	ESX1 HBA port 2
(zone created in fabric 2)	5000D31000036002	Controller1 front-end primary plugged into fabric 2
	5000D3100003600A	Controller2 front-end primary plugged into fabric 2

Table 1 Example of zoning by WWN

#### 2.3 Port zoning

Caution: Due to the supportability of port zoning, WWN zoning is preferred over port zoning.

Port zoning involves creating zones by including the physical ports instead of the WWN. Although this method has security advantages, it creates supportability challenges in the environment.

# 2.4 Virtual ports

If the SC Series array is configured to use virtual port mode, include all of the front-end virtual ports within each fault domain in the zone with each ESXi initiator. See Figure 1.

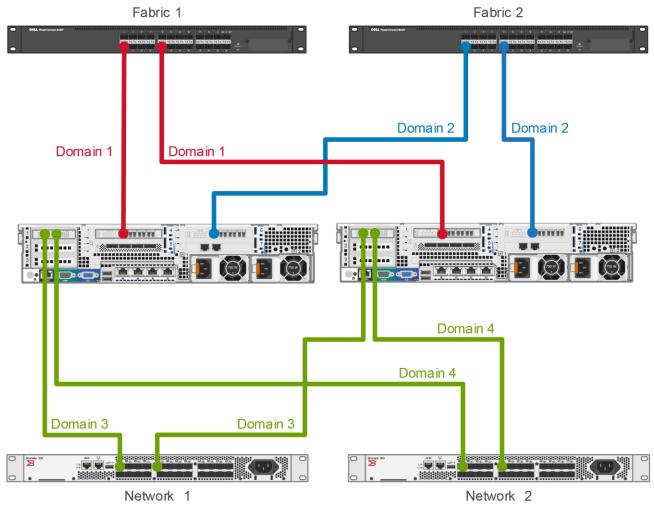


Figure 1 Virtual port domains, Fibre Channel (FC) and iSCSI

# 3 Host initiator settings

Make sure the initiator settings are configured in the ESXi host according to the latest *Dell Storage Center System Manager Administrator's Guide* available on the Knowledge Center at the <u>SC Series Customer Portal</u> (login required).

The current recommended settings are:

- QLogic<sup>®</sup> Fibre Channel card BIOS settings:
  - Set connection options to 1 for point to point only.
  - Set login retry count to 60 attempts.
  - Set port down retry to 60 attempts.
  - Set link down timeout to 30 seconds.
  - Set Execution Throttle to 255 (if available). The ESXi VMkernel driver module and DSNRO effectively control the queue depth. According to the QLogic support article, <u>HBA Execution</u> <u>Throttle And Queue Depth In A VMware Environment</u>, the Execution Throttle variable is ignored by the QLogic driver.
- Emulex Fibre Channel card BIOS settings:
  - Set the Node Time Out field, lpfc\_devloss\_tmo (formerly nodev\_tmo) to 60 seconds.
  - Set topology to 2 for Auto Topology (point to point first).
  - Set queuedepth to 255. A queue depth of 255 allows the ESXi VMkernel driver module and DSNRO to 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. Here is an example script for enabling ARP redirect for hardware iSCSI HBAs:

```
esxcli iscsi physicalnetworkportal param set --option ArpRedirect -v=1 -A vmhba4
```

**Note:** This command replaces the former esxcfg-hwiscsi command found in <u>VMware KB Article 1010309</u>. See the vSphere 5 Documentation Center article, <u>Reference to Replacements for Service Console</u> <u>Commands</u>, for an explanation of the new syntax.

- iSCSI initiator settings for delayed ACK: During periods of high network congestion in some environments, iSCSI transfer latency may exceed acceptable levels. VMware recommends disabling delayed ACK using the steps described in the article, <u>ESX/ESXi hosts might experience read or write</u> performance issues with certain storage arrays, located in the VMware Knowledge Base.
- SAS HBA card BIOS settings: For Dell Storage SCv2000 Series arrays configured with SAS front-end ports, factory default settings within the card BIOS are recommended for SAS ESXi host HBAs.

# 4 Modifying queue depth and timeouts

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 an ESXi host HBA port or iSCSI initiator, and the target is the SC Series 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. While increasing the queue depth can sometimes increase performance, if it is set too high, there is an increased risk of overdriving the storage array.

When data travels between the application and the storage array, there are several places where the queue depth can be set to throttle the number of concurrent disk transactions. The most common places where queue depth can be modified are listed in Table 2.

Area	Setting
The application itself	Default=dependent on application
The virtual SCSI card driver in the guest	Default=32
The virtual machine file system (VMFS) layer (DSNRO)	Default=32
The HBA VMkernel Module driver	Default=64
The HBA BIOS	Default=varies

Table 2 Areas for setting queue depth

The remainder of this section explains how to set the queue depth in each layer.

**Caution:** The appropriate queue depth for a host may vary due to a number of factors. As a best practice, only increase or decrease the queue depth if necessary. See appendix A for information on determining the proper queue depth.

#### 4.1 Host bus adapter queue depth

When configuring the host bus adapter, the Execution Throttle or queue depth should be set to 255 as to not gate the driver module queue depth. Note that in some cases, the Execution Throttle variable in QLogic cards is unavailable because it is outdated and has been deprecated in favor of the queue depth variable set in the driver module. The queue depth variable within the VMkernel driver module loaded for each HBA in the system and DSNRO ultimately regulate the host queue depth.

#### 4.2 Storage driver queue depth and timeouts

The VMkernel driver module ultimately regulates the queue depth for the HBA if it needs to be changed. See appendix A for 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 properly survive an SC Series controller failover. To configure these settings, refer to the section, "Adjust Queue Depth for QLogic, Emulex, and Brocade HBAs", in the document, *vSphere Troubleshooting* located at VMware <u>vSphere 6.x</u> <u>Documentation</u> or <u>vSphere 5.x Documentation</u>.

**Caution:** Before executing the following commands, refer to the latest documentation from VMware for the latest information.

#### 4.2.1 Fibre Channel HBAs

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

- 1. Locate the appropriate driver name for the module that is loaded:
  - For QLogic, enter:

esxcli system module list |grep ql

- For Emulex, enter:

esxcli system module list |grep lpfc

Depending on the HBA model, the output could be similar to the following:

- QLogic: qla2xxx or qlnativefc
- Emulex: 1pfc820

**Note:** The following steps contain example module names. Actual module names should be acquired when completing step 1.

- 2. Set the driver queue depth and timeouts using the esxcli command:
  - For QLogic, enter:

esxcli system module parameters set -m qlnativefc -p "ql2xmaxqdepth=255 ql2xloginretrycount=60 qlport down retry=60"

- For Emulex, enter:

esxcli system module parameters set -m lpfc820 -p "lpfc\_devloss\_tmo=60 lpfc\_lun\_queue\_depth=255"

**Note:** In certain multipathing configurations, the qlport\_down\_retry value may be set lower to decrease failover times between paths if one of the paths fails.

- 3. Reboot the ESXi host for these changes to take effect.
- 4. To verify the settings, use the following command:

```
esxcli system module parameters list -m=module(i.e. -m=qla2xxx)
```

#### 4.2.2 Software iSCSI initiator

Similarly, for the software iSCSI initiator, complete the following steps:

1. Set the queue depth to 255 (example shown):

esxcli system module parameters set -m iscsi vmk -p iscsivmk LunQDepth=255

- 2. Increase the login timeout to 60 seconds (example shown):
  - Determine the iSCSI adapter name:

esxcli iscsi adapter list

- Set the login timeout parameter:

esxcli iscsi adapter param set -A vmhba37 -k LoginTimeout -v 60

- 3. Reboot the ESXi host for the change to take effect.
- 4. To verify the settings, use the following commands:

```
esxcli system module parameters list -m iscsi_vmk
esxcli iscsi adapter param get -A vmhba37
```

**Note:** In earlier versions of ESXi 5.x, the option to set the login timeout parameter is not available. To enable the login timeout parameter, it may require applying the patch as described in <u>VMware KB article 2007680</u>.

#### 4.2.3 SAS HBAs

For the mpt3sas driver module, use all of the default driver module settings.

## 4.3 Adjusting settings for permanent device loss conditions

When using vSphere High Availability (HA), it is a best practice to modify the following host settings in dealing with permanent device loss (PDL) conditions.

- 1. From within the host advanced system settings, make the following modifications:
  - VMkernel.Boot.terminateVMOnPDL = Yes/True (Default = No/False, Reboot required)

This setting automatically terminates a virtual machine that reside on a datastore in a PDL condition. For example, if a storage administrator accidentally removes a path mapping to a single host causing a PDL condition, this allows HA to terminate that virtual machine, and restart it on a different host within the cluster.

- vSphere 5.5: Disk.AutoremoveOnPDL = 0 (non-default advanced setting)
- vSphere 6.0+: Disk.AutoremoveOnPDL = 1 (default advanced setting)

This setting prevents a disk device entering a PDL state from being automatically removed from a host, which prevents a device that has inadvertently removed from being treated as a new device.

- 2. From within the HA cluster advanced options, add the following configuration parameter:
  - das.maskCleanShutdownEnabled = True (Default setting = True)

This setting instructs the fault domain manager (FDM) to presume a virtual machine should be restarted when its home datastore is not accessible.

For more information about these settings, refer to the following:

- <u>vSphere Availability guide</u>
- PDL Conditions and High Availability
- PDL AutoRemove feature in vSphere 5.5

## 4.4 Modifying the VMFS queue depth for virtual machines (DSNRO)

Disk scheduled number requests outstanding (DSNRO) is an advanced setting within each ESXi host that controls the queue depth at the datastore level.

DSNRO is a value that can be increased or decreased depending on how many virtual machines are to be placed on each datastore, or based on the VM I/O requirements. This queue depth limit is only enforced when more than one virtual machine per host is active on that datastore. For example, if the value is set to the default, the first virtual machine active on a datastore will have its queue depth limited only by the queue depth set in the VMkernel driver module. When a second, third, or fourth virtual machine is added to the datastore, during contention the limit will be enforced to the maximum queue depth of 32, or as set by the esxcli storage core device command.

**Note:** Previous to ESXi 5.5, DSNRO was a global setting named Disk.SchedNumReqOutstanding that applied to all datastores. Modifying DSNRO for an individual datastore was not possible.

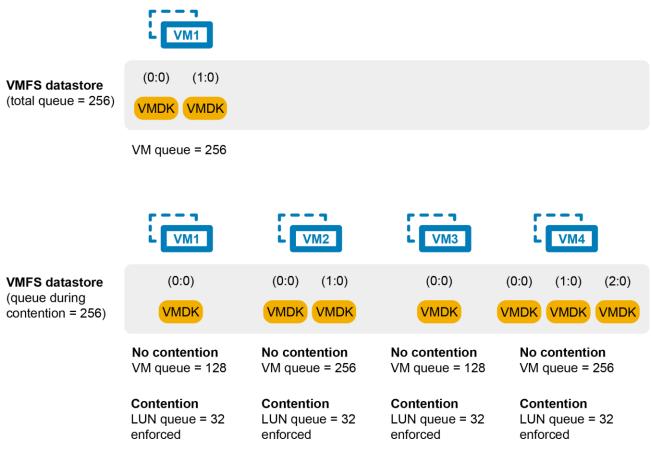


Figure 2 Example queue utilization with the DSNRO set to 32

**Note:** The DSNRO limit does not apply to volumes mapped as raw device mappings (RDMs). Each RDM will have its own queue.

The DSNRO setting can be modified on a per-datastore basis using the command line:

esxcli storage core device set -d <naa.dev> -O <value of 1-256>

**Note:** This allows fine-tuning of DSNRO on a per-volume basis, however the drawback is that it must be set on each datastore (and each host) if the desired queue depth is greater than 32.

To globally set the DSNRO, the easiest method is to use a scripting tool such as the VMware PowerCLI utility. The following example script demonstrates setting the DSNRO to 64 globally across all datastores, on each host. It is important to remember that any volume added after this script is executed would need to be changed manually or by re-running the script.

#### Example PowerCLI Script (dsnro.ps1):

```
#Connect to vCenter. Change server, user, and password credentials below
Connect-VIServer -Server 'vCenter Server IP or FQDN' -User
'administrator@vsphere.local' -Password 'thepassword'
#Change this variable to the desired dsnro (Default=32 Max=256)
dsnro = 64
#Retrieve a list of ALL ESXi hosts from the vCenter server
$esxhosts = get-vmhost
#Cycle through each host retrieving all storage devices associated with that
host
foreach ($hostname in $esxhosts) {
  $esxcli = get-esxcli -vmhost $hostname
  $hostdevices = $esxcli.storage.core.device.list()
  #Cycle through all storage devices on that host changing Compellent volumes
DSNRO to the new value
  foreach ($cmlvolume in $hostdevices) {
      if ($cmlvolume.Vendor -like "COMPELNT")
      {
      $esxcli.storage.core.device.set($null, $cmlvolume.Device, $null, $null,
$null, $null, $dsnro, $null)
      }
    }
}
```

As a best practice for modifying any host settings, run tests to determine the impact of changing this variable beforehand. Once successful tests have been completed, as a precautionary measure, it is recommended only to execute this script during a maintenance window. Since DSNRO helps to ensure fairness across the virtual machines residing on a datastore, modifying this value could lead to individual virtual machines monopolizing disk resources.

More information on the DSNRO can be found in the following documentation:

- "Change Maximum Outstanding Disk Requests in the vSphere Web Client" in vSphere Troubleshooting located at <u>VMware vSphere 6.x documentation</u> or <u>vSphere 5.x documentation</u>.
- <u>Setting the Maximum Outstanding Disk Requests for virtual machines</u> in the VMware Knowledge Base.

## 4.5 Adaptive queue depth

At times of high congestion, VMware has an adaptive queue depth algorithm that can be enabled with the QFullSampleSize and QFullThreshold variables. These variables aid in relieving the congestion by dynamically reducing and increasing the logical unit number (LUN) queue depth. Due to the architecture of SC Series storage, enabling these settings is not recommended unless under the guidance of Dell Copilot Support or Dell Support. For more information, see <u>Controlling LUN queue depth throttling in VMware</u> <u>ESX/ESXi</u> in the VMware Knowledge Base.

## 4.6 Modifying the guest OS queue depth

The queue depth can also be set within the guest operating system if needed. 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. Examples are shown in section 4.6.1. Additional vSCSI adapters can be added to a virtual machine to increase the total queue depth available to the virtual machine.

#### 4.6.1 Windows Server 2008, 2008 R2, or 2012

To adjust the queue depth for the LSI<sup>®</sup> Logic drivers, the task involves adding one of the following registry keys:

- 1. Back up the registry.
- 2. Use regedit to add the following keys:
  - a. For LSI Logic Parallel (LSI\_SCSI):

Windows Registry Editor Version 5.00

```
[HKLM\SYSTEM\CurrentControlSet\Services\LSI_SCSI\Parameters\Device]
"DriverParameter"="MaximumTargetQueueDepth=128;"
; The semicolon is required at the end of the queue depth value
"MaximumTargetQueueDepth"=dword:00000080
. 20 her is argued to 122 in decimal
```

; 80 hex is equal to 128 in decimal

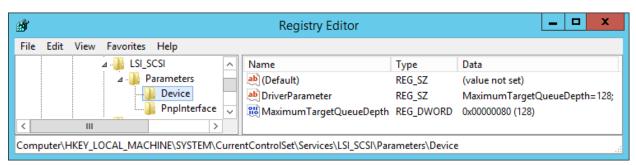


Figure 3 Registry settings for the LSI Logic Parallel vSCSI adapter

#### b. For LSI Logic SAS (LSI\_SAS):

Windows Registry Editor Version 5.00

[HKLM\SYSTEM\CurrentControlSet\Services\LSI\_SAS\Parameters\Device] "DriverParameter"="MaximumTargetQueueDepth=128;" ; The semicolon is required at the end of the queue depth value "MaximumTargetQueueDepth"=dword:00000080 ; 80 hex is equal to 128 in decimal

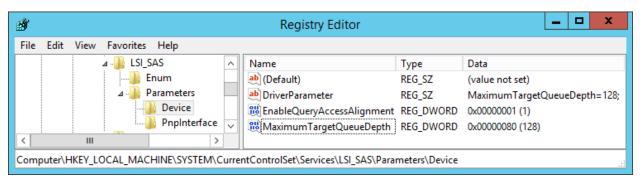


Figure 4 Registry setting for the LSI Logic SAS vSCSI Adapter

c. For VMware Paravirtual SCSI (PVSCSI), the Paravirtual adapter is different from the LSI vSCSI adapters in that its queue depth can be adjusted higher: up to 256 for devices and 1024 for the adapter.

For more information, see the VMware KB article, <u>Large-scale workloads with intensive I/O</u> patterns might require queue depths significantly greater than Paravirtual SCSI default values.

3. Reboot the virtual machine.

**Note:** See the <u>VMware Knowledge Base</u> for the most current information about setting the queue depth with different vSCSI controllers or operating systems.

#### 4.7 Setting operating system disk timeouts

For each operating system running within a virtual machine, if VMware tools are not installed, the disk timeouts must also be set to 60 seconds for the operating system to handle storage controller failovers properly.

Examples of setting the operating-system timeouts can be found in the section, "Set Timeout on Windows Guest OS" in the *vSphere Storage Guide* at <u>VMware vSphere 6.x documentation</u> or <u>vSphere 5.x</u> <u>documentation</u>.

The general steps to set the disk timeout within Windows and Linux are listed in the following sections.

#### 4.7.1 Windows

- 1. Back up the registry.
- 2. Using the registry editor, modify the following key.

```
Windows Registry Editor Version 5.00
```

[HKLM\SYSTEM\CurrentControlSet\Services\Disk]

"TimeOutValue"=dword:000003c

; 3c in hex is equal to 60 seconds in decimal

<u>څ</u>	Registry Edito	r	_ <b>D</b> X
File Edit View Favorites Help			
▲     disk       □     Enum       □     Dnscache       ▷     Onscache       ▷     DPS       ▷     DsmSvc       □     DXGKrnl       ▷     Eaphost       ▷     EFS	<ul> <li>Name</li> <li>(Default)</li> <li>AutoRunAlwaysDisable</li> <li>DisplayName</li> <li>ErrorControl</li> <li>Group</li> <li>ImagePath</li> <li>Owners</li> <li>Start</li> <li>TimeoutValue</li> <li>Type</li> </ul>	REG_SZ REG_DWORD REG_SZ	Data (value not set) Brother RemovableDisk(U) @disk.inf,%disk_ServiceDesc%;Disk Driver 0x00000001 (1) System32\drivers\disk.sys disk.inf 0x0000000 (0) 0x0000000 (0) 0x00000001 (1)

Figure 5 Registry key to set the Windows disk timeout

3. Reboot the virtual machine.

**Note:** This registry value is automatically set when installing VMware Tools. For more information, see <u>Inconsistent Windows virtual machine performance when disks are located on SAN datastores</u> in the VMware Knowledge Base.

#### 4.7.2 Linux

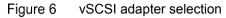
For more information about setting disk timeouts in Linux, refer to the VMware Knowledge Base article, Increasing the disk timeout values for a Linux 2.6 virtual machine.

**Caution:** For RHEL 7.x virtual machines, verify the disk timeouts are correctly set to 60 seconds. For more information, see the Red Hat knowledge base (login required), article 1578643: <u>The udev rules for SCSI timeout</u> are missing in open-vm-tools package in Red Hat Enterprise Linux 7.

# Guest virtual SCSI adapter selection

When creating a new virtual machine, there are four types of virtual SCSI (vSCSI) controllers to choose from. Based on the operating system selected, vSphere will automatically recommend and select a SCSI controller that is best suited for that particular operating system. The best practice is to follow the built-in recommendation. The nuances of each adapter are described in the following subsections.

🗕 🛃 New SCSI controller	LSI Logic SAS
SCSI Bus Sharing	None
Change Type	LSI Logic SAS 🗸 🗸
	BusLogic Parallel
	VMware Paravirtual
	LSI Logic SAS
	LSI Logic Parallel



## 5.1 BusLogic Parallel

5

This vSCSI controller is used for older operating systems. Due to the queue depth limitations of this controller, it is not recommended unless it is the only option available for that particular operating system. This is because certain versions of Windows issue only enough I/O to fill a queue depth of one.

## 5.2 LSI Logic Parallel

Because this vSCSI adapter is supported by many operating system versions, it is a good overall choice and is recommended for virtual machines that do not support the LSI Logic SAS adapter.

## 5.3 LSI Logic SAS

This vSCSI controller is available for virtual machines with hardware versions 7 and later. It also has similar performance characteristics of the LSI Logic Parallel. This adapter adds support for SCSI-3 reservations, which are required for Microsoft Cluster Services (MSCS). Some operating system vendors are gradually withdrawing support for SCSI in favor of SAS, making the LSI Logic SAS controller a good choice for future compatibility.

## 5.4 VMware Paravirtual

This vSCSI controller is a high-performance adapter that can result in greater throughput and lower CPU utilization. More information about the usage and limitations of this adapter can be found in the section, "About VMware Paravirtual SCSI Controllers", in the *vSphere Virtual Machine Administration Guide* located at <u>VMware vSphere 6.x documentation</u> or <u>vSphere 5.x documentation</u>.

# 6 Mapping volumes to an ESXi server

Within the SC Series, mapping is the process of presenting a volume to a host. The following subsections describe basic concepts on how vSphere treats different scenarios.

## 6.1 Basic volume mapping concepts

When sharing volumes between ESXi hosts for such tasks as vMotion, HA, and DRS, for consistency it is recommended that each volume is mapped to clustered ESXi hosts using the same LUN.

For example:

- There are three ESXi hosts named ESXi1, ESXi2, and ESXi3.
- 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 ESXi1 -as- LUN 10

Volume: "LUN10-vm-storage" → Mapped to ESXi2 -as- LUN 10

Volume: "LUN10-vm-storage" → Mapped to ESXi3 -as- LUN 10

#### 6.2 Basic SC Series volume mappings

In SCOS versions 5.x and later, the mapping process is automated by creating a server cluster object. This allows the volume to be mapped to multiple ESXi hosts at the same time, automatically keeping the LUN numbering consistent for all the paths.



Figure 7 Example of a server cluster object

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 are 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 are removed; 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.

In the mapping wizard, the system can auto select the LUN number, or a preferred LUN number can be manually specified from the advanced settings screen as shown in Figure 8.

Select LUN	
Use next available LUN	
LUN to use when mapping to Volume	10
Use the next available LUN if specified LUN is unavaila Map volume using LUN 0 (this is usually reserved for boot)	

Figure 8 Manually specifying a LUN in the advanced settings screen

This advanced option allows administrators who already have a LUN numbering scheme to continue using it. However, if a LUN is not manually specified, the system automatically selects a LUN for each volume incrementally starting at LUN 1.

**Timesaver:** When naming volumes from within the SC Series GUI, it may be helpful to specify the LUN number as part of the volume name. This helps to quickly identify which volumes are mapped using each LUN.

#### 6.3 Multipathed volume concepts

If there are ESXi hosts that have multiple initiator ports, ESXi has built-in functionality to provide native multipathing of volumes over Fibre Channel, Fibre Channel over Ethernet (FCoE), hardware iSCSI, or software iSCSI.

Building on the previous example, here is an example of multipathing mappings:

Volume: "LUN10-vm-storage" → Mapped to ESXi1/HBA1 -as- LUN 10

Volume: "LUN10-vm-storage" → Mapped to ESXi1/HBA2 -as- LUN 10

Volume: "LUN10-vm-storage" → Mapped to ESXi2/HBA1 -as- LUN 10

Volume: "LUN10-vm-storage" → Mapped to ESXi2/HBA2 -as- LUN 10

Volume: "LUN10-vm-storage" → Mapped to ESXi3/HBA1 -as- LUN 10

Volume: "LUN10-vm-storage" → Mapped to ESXi3/HBA2 -as- LUN 10

**Note:** With older versions of ESXi, if the LUN number is not consistent between multiple hosts or multiple HBAs, 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 needs to be mapped to one front end port, while the second ESXi initiator is mapped to the other front end port in that same controller. For example:

"LUN10-vm-storage" → Controller1/Port1 → FC-Switch-1 → Mapped to ESXi1/HBA1 as LUN 10

"LUN10-vm-storage" → Controller1/Port2 → FC-Switch-2 → Mapped to ESXi1/HBA2 as LUN 10

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

"LUN20-vm-storage" → Controller2/Port1 → FC-Switch-1 → Mapped to ESXi1/HBA1 as LUN 20

"LUN20-vm-storage"  $\rightarrow$  Controller2/Port2  $\rightarrow$  FC-Switch-2  $\rightarrow$  Mapped to ESXi1/HBA2 as LUN 20

This means that when configuring multipathing 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.

#### 6.4 Multipathed SC Series volumes

When multipathing SC Series volumes, much of the process is automated. Many of the common mapping errors can be prevented by simply selecting the correct operating system in the server properties screen. Based on the OS selected, SC Series correctly maps volumes by applying a set of rules to the server that are unique to each operating system.

强 Edit Settings [SC1:	Edit Settings [SC12] [ESXi1]			
Name	ESXI			
Operating System	VMware ESXi 6.0			
Notes				
	<b>*</b>			
Server Folder	E 🙀 Servers			
	Alert On Lost Connectivity			
? Help	🗷 Cancel 🔷 OK			

Figure 9 Server operating system selection

Multipathing to an ESXi host is automatic when the server object has more than one HBA or iSCSI initiator ports assigned to it. In other words, the advanced options must be used if the server does not need a volume multipathed.

🚍 Map Yolume to Server	×
The following volume and server will be mapped	
Volume 闄 New Volume 1	
Server 🖳 ESXi1	
Advanced Options	
🥐 Help 🔶 🔶 Finish	

Figure 10 Advanced server mapping options

Function	Description					
Select LUN	Uncheck the <b>Use next available LUN</b> option to manually specify the LUN. If this box is not checked, the system automatically assigns the next available LUN.					
Restrict Mapping Paths	Use this option when a volume only needs to be mapped to a specific HBA in the ESXi host.					
Configure Multipathing	This option designates how many of the SC Series front-end ports that the system will allow the volume to be mapped through. For example, if each controller has four front-end ports, selecting <b>unlimited</b> maps the volume through all four. Selecting two only uses two of the four front-end ports. The system automatically selects the two front-end ports with the fewest mappings.					
Configure Volume Use	Read-only mappings are not recognized by VMFS, so this option should not be used.					

Table 3 Advanced mapping options for an ESXi host

## 6.5 Configuring the VMware iSCSI software initiator for a single path

Although it is not recommended, for instances where iSCSI multipathing cannot be configured, the steps required for a single-path iSCSI configuration are as follows.

From within the VMware vSphere Client:

- 1. In the ESXi host Security Profile > ESXi firewall, enable the Software iSCSI Client.
- 2. Add a VMkernel port to a virtual switch assigned to the physical NIC for iSCSI (see Figure 11).

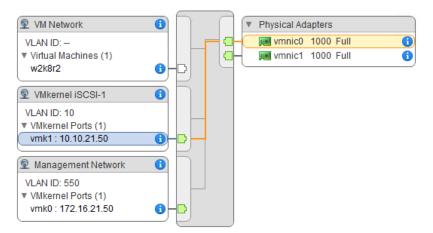


Figure 11 Configuring the VMkernel port

- 3. From the **Storage Adapters** configuration screen, click the plus sign to add a new adapter, select **Software iSCSI Adapter**, and click OK.
- 4. Within the Storage Adapters area, highlight the iSCSI software adapter (such as vmhba38).
- 5. To add the SC Series front-end ports, select Targets, then select the Dynamic Discovery tab. Click Add... and enter the iSCSI control port IP address, labeled as the well-known IP address for the fault domain. If virtual ports are not enabled for iSCSI front-end ports, each of the iSCSI adapter IP addresses for the system must be added manually.
- 6. Rescan the iSCSI initiator.
- 7. From DSM/EM:
  - a. Create a server object for the ESXi host.
  - b. Map a volume to the ESXi host.
- 8. From within the VMware vSphere client, navigate to the **Storage Adapters** section and rescan the iSCSI HBA for new volumes.

## 6.6 Configuring the VMware iSCSI software initiator for multipathing

To configure the VMware iSCSI software initiator for multipathing, see sections "Configuring Software iSCSI Adapter" and "Multiple Network Adapters in iSCSI Configuration" in the *vSphere Storage Guide* at <u>VMware</u> <u>vSphere 6.x documentation</u> or <u>vSphere 5.x documentation</u>.

For users with previous experience configuring iSCSI for multipathing, here are a few key points for a successful configuration:

- Verify that there is one VMkernel interface for each physical NIC to be used for storage traffic, and that they follow the virtual switch port binding recommendations listed in section 6.6.1.
- Adjust the failover order on each VMkernel interface for a 1:1 VMkernel to physical NIC ratio.
- Add both VMkernel interfaces to the iSCSI software adapter network port binding (see Figure 12). If the prerequisite failover orders have not been set, the vSphere client will not allow the operation.
- Rescan the iSCSI software adapter for new volumes.
- From within the DSM/EM client, create the server object for the host.

iSCSI Software Adapter										
Svmhba38 iSCSI Online iqn.1998-01.com.vmware:tssrv248-2b03cc0b										
Adapter Details										
Properties Devices Paths Targets Network Port Binding Advanced Options										
Port Group	VMkernel Ad	Port Group Policy	Path Status	Physical Network Adapter						
VMkernel iSCSI-1 (vSwitch0)	👥 vmk1	Compliant	🔷 Not used	衬 vmnic0 (1 Gbit/s, Full)						

Figure 12 Binding multiple VMkernel ports to the software iSCSI initiator

#### 6.6.1 VMkernel port binding considerations with software iSCSI

When using the software iSCSI initiator provided within an ESXi host, careful consideration needs to be given to the appropriate virtual switch configurations adhering to VMware iSCSI port-binding requirements. The number of fault domains, subnets, and IP configuration need to be carefully examined to avoid configuration errors. In the case where the iSCSI front-end ports are configured in a single subnet and single fault domain, the vmnics must reside within a single vSwitch in order for VMkernel port binding to be used as required by VMware.

**Note:** The most common environment in which a single fault domain is utilized for the iSCSI network is a redundant configuration with a single VLAN.

For example, with subnet 192.168.0.x and subnet mask 255.255.255.0, the vSwitch would look similar to Figure 13.

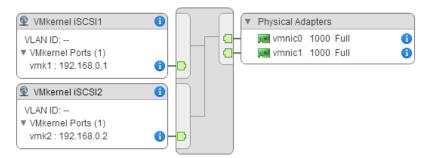


Figure 13 iSCSI ports and NICs on a single vSwitch/single subnet (VMkernel port binding allowed)

**Note:** Using port binding is not recommended when the VMkernel ports are on different networks as shown in Figure 14, because it may cause very long rescan times and other storage management problems. See <u>Considerations for using software iSCSI port binding in ESX/ESXi</u> in the VMware Knowledge Base.

In configurations using multiple fault domains, with the iSCSI front-end ports configured in a multiple subnet and multiple fault domain configuration, the VMkernel ports and vmnics must reside in separate vSwitches. Due to the network separation, VMkernel port binding must **not** be used.

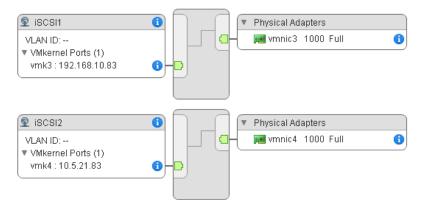
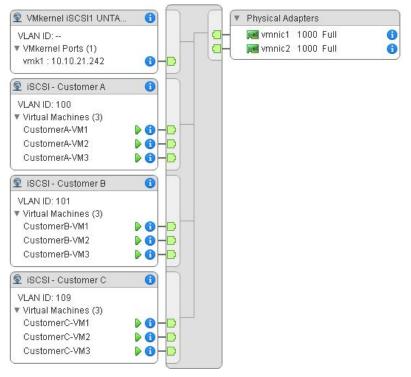


Figure 14 iSCSI VMkernel ports and NICs on multiple vSwitches/multiple subnets (port binding must not be used)

## 6.7 iSCSI port multi-VLAN configuration recommendations

For SC Series systems with iSCSI capabilities, SCOS 6.5 and later has multiple-VLAN support for diverse network configurations and multi-tenant environments. This allows iSCSI storage I/O to be separated between vSwitches for customers using software iSCSI initiators within the guest, or to isolate certain ESXi host-cluster VMkernel ports to their own dedicated iSCSI storage VLAN.





Since iSCSI traffic is not encrypted in its plain form, it is a best practice to isolate that traffic for security purposes. A common misconception is that the Challenge-Handshake Authentication Protocol (CHAP) encrypts iSCSI traffic, but it simply provides authentication for the connection to prevent unauthorized access.

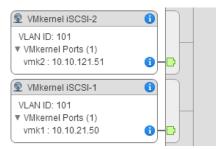


Figure 16 Isolating the ESXi software iSCSI initiator using VLANs

**Note:** The configuration of iSCSI VLANs is new functionality that can only be configured through the DSM/EM client.

•	Domain 100		Show	🔍 Edit Settir	ngs 📜 Edit ISCS	61 10 Ports 🎾 A	dd Ports to Fault Dom	nain 🍺 Remove Ports from Fault Domain 🚏	Create VLAN Copy 🐋 Conf	igure NAT Port Forwa	arding 🚉 Configure CHA
SC22 Volumes Servers TSSRV251 Fault Domains Fault Domains Domain 100	Index Transport Type ISCSI Transport Mode Well Known IP Address Gateway Subnet Mask	5 ISCSI Virtual Port 100.0.0.10 100.0.0.1 255.255.255.0	)				MTU Fault Domain Type VLAN ID Class of Service Pri	9000 (Jumbo) Virtual 100 ority 0			
Domain 101	iSCSI Name	iqn.2002-03.c	om.comp	pellent:5000d	131000ec7c31.Fa	aultDomain5					
Pault Domain 1	Physical Ports										
isks	Name	Controller	Slot	Slot Port	IP Address	Subnet Mask		ISCSI Name	Purpose	Status	Status Messag
torage Types eplay Profiles	JE 5000D31000EC7C32	SN 60540	1	2	100.0.0.3	255.255.255.0	0 100.0.0.1 ign.2	2002-03.com.compellent:5000d31000ec7c32	Front End	Up	
epidy Profiles	J 5000D31000EC7C34	SN 60540	1	1	100.0.2	255.255.255.0	0 100.0.0.1 iqn.2	002-03.com.compellent:5000d31000ec7c34	Front End	Up	
	JE 5000D31000EC7C36	SN 60541	1	2	100.0.0.5	255.255.255.0	0 100.0.0.1 iqn.2	002-03.com.compellent:5000d31000ec7c36	Front End	Up	
	J 5000D31000EC7C38	SN 60541	1	1	100.0.0.4	255.255.255.0	0 100.0.0.1 ign.2	2002-03.com.compellent:5000d31000ec7c38	Front End	Up	
	Virtual Ports										
	Name		Control	ler	Par	rent	Preferred Par	ent			
	5000D31000EC7C33	SN 6054	0		5000D31000E0	C7C32	5000D31000EC7C32				
	= 5000D31000EC7C35	SN 6054	0		5000D31000E0	C7C34	5000D31000EC7C34				
	5000D31000EC7C37	SN 6054			5000D31000E0		5000D31000EC7C36				

Figure 17 Configuration of VLANs within the Dell Storage Manager (Enterprise Manager) client

The DSM client allows the creation of new VLANs, configuration of iSCSI port IP addresses, and setting of the well-known IP address (formerly known as the control port) for the new iSCSI fault domains.

Note: For added flexibility, Jumbo Frames can be enabled on a per-fault-domain basis.

#### 6.8 Configuring the FCoE software initiator for multipathing

When using the ESXi software FCoE, volumes are assigned to the VMW\_SATP\_LOCAL by default. In some instances, this will cause paths to go unclaimed. To remedy this situation, it may be necessary to adjust the claim rules on each ESXi host so that FCoE volumes are correctly assigned to the VMW\_SATP\_DEFAULT\_AA.

For SCOS versions 6.5 and earlier:

esxcli storage nmp satp rule add -R fcoe -s VMW SATP DEFAULT AA

For SCOS versions 6.6 and later (see section 6.12 for more information):

esxcli storage nmp satp rule add -R fcoe -s VMW SATP ALUA

#### 6.9 VMware multipathing policies

When configuring the path-selection policy of each datastore or volume, administrators have the choice of round robin, fixed, or most recently used. The default path selection policy for the SC Series system depends on SCOS version and the associated SATP module detected, but round robin is highly recommended for ease of management if there are no technical reasons against using it (such as with Microsoft failover clusters or Dell Live Volume).

#### 6.9.1 Round robin policy

The round robin path selection policy uses automatic path selection and load balancing to rotate I/O through all 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 datastore volume in bursts evenly and sequentially across paths in an alternating fashion.

Using round robin reduces 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 a best practice to enable round robin between an iSCSI path and Fibre Channel path, nor enabling it to balance the load between an 8 Gb FC and a 16 Gb FC path. If round robin is enabled for one or more datastores, care must be taken to ensure all the paths included are identical in type, speed, and have the same queue depth settings.

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

- 1. Load is distributed evenly between HBA1 and HBA2
- 2. HBA1 loses connectivity; HBA2 will assume all I/O load
- 3. HBA1 resumes connectivity; load is distributed evenly again between both

Aultipathing Details									
Device: COMPELNT Fib	re Channel Disk (naa.6000d3	100002b9000000	000003e187)						
Multipathing Policies	Policies								
▶ Path Selection Po	licy Round Robin (VMwa								
Storage Array Type	Array Type Policy VMW_SATP_DEFAULT_AA								
Paths									
Owner Plugin	NMP								
	Refresh Enable	Refresh Enable Disable							
	Runtime Name	Status	Target	LUN	Preferred				
	vmhba2:C0:T2:L1	<ul> <li>Active (I/O)</li> </ul>	50:00:d3:10:00:02:b9:02 50:00:d3	1					
	vmhba1:C0:T7:L1	<ul> <li>Active (I/O)</li> </ul>	50:00:d3:10:00:02:b9:02 50:00:d3	1					
	vmhba2:C0:T5:L1	<ul> <li>Active (I/O)</li> </ul>	50:00:d3:10:00:02:b9:02 50:00:d3	1					
	vmhba1:C0:T5:L1	<ul> <li>Active (I/O)</li> </ul>	50:00:d3:10:00:02:b9:02 50:00:d3	1					

Figure 18 Example of a datastore path selection policy set to round robin

#### 6.9.1.1 Setting round robin to be the default path selection policy

The round robin path selection policy (PSP) can be set to the default using 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 may have to be set manually.

For SCOS 6.5 and earlier:

esxcli storage nmp satp set -P VMW PSP RR -s VMW SATP DEFAULT AA

For SCOS 6.6 and later (see section 6.12 for more information):

esxcli storage nmp satp set -P VMW PSP RR -s VMW SATP ALUA

**Note:** The default path selection policy that is applied to the VMW\_SATP\_ALUA module will also be applied to the protocol endpoints presented to that ESXi host for virtual volumes.

If an entire cluster needs the pathing policy switched from fixed to round robin, use the following VMware PowerCLI PowerShell command:

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

Similarly, where the ALUA SATP has set all volume policies to MRU, use the following command:

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

**Caution:** The round robin path selection policy may be unsupported for use with Microsoft Clustering Services dependent on the version of ESXi. Check the current support status.

In addition, the following PowerCLI script can be used to change the default PSP on all the ESXi hosts within a vCenter instance.

#### Example PowerCLI script

```
# This script will change the default PSP for a SATP
# Use "Connect-VIServer vcenter-host.your.domain" to connect before executing
script
# Valid SATP options for SC Series arrays are:
# SCOS 6.5 and earlier = "VMW_SATP_DEFAULT_AA"
# SCOS 6.6 and later = "VMW_SATP_ALUA"
$satp = "VMW_SATP_ALUA"
# Valid PSP options for each SATP are:
# "VMW_PSP_RR" (Recommended)
# "VMW_PSP_FIXED"
```

```
$defaultpsp = "VMW_PSP_RR"
$esxHosts = Get-VMHost
foreach ($esx in $esxHosts) {
   Write-Host "Setting $esx SATP $satp to $defaultpsp as default"
   $esxcli = Get-EsxCli -VMHost $esx
   $esxcli.storage.nmp.satp.set($null,$defaultpsp,$satp)
}
Write-Host "Script completed"
```

#### 6.9.1.2 Conditions for path changes

When using the round robin path selection policy, the native multipathing plug-in (NMP) has custom properties that can be tuned to optimize traffic flow between each of the paths. By default, the round robin PSP will switch paths every 1,000 I/Os or 10 megabytes, whichever comes first. Dependent on a number of factors (including average block size, queue depth, and the read-to-write ratios) Dell internal lab testing has shown that modifying the IOPS value can have a noticeable performance impact based on the I/O patterns. Because the optimum tuning value for each datastore is dependent on the data patterns issued by the guests, thorough testing is recommended to prevent negative performance. When beginning testing, a good starting point is to set the IOPS to 3 as shown in the following.

To set the values for a particular datastore, use the following steps:

1. Find the datastore device naa identifier.

```
~ # esxcli storage nmp device list
naa.6000d3100002b90000000000016d88
Device Display Name: rmvol1
Storage Array Type: VMW_SATP_DEFAULT_AA
Storage Array Type Device Config: SATP VMW_SATP_DEFAULT_AA does not
support device configuration.
Path Selection Policy: VMW_PSP_RR
Path Selection Policy Device Config:
{policy=rr,iops=1000,bytes=10485760,useANO=0; lastPathIndex=0:
NumIOsPending=0,numBytesPending=0}
Path Selection Policy Device Custom Config:
Working Paths: vmhba2:C0:T3:L11, vmhba3:C0:T6:L11, vmhba2:C0:T7:L11,
vmhba3:C0:T3:L11
Is Local SAS Device: false
Is Boot USB Device: false
```

2. If not set previously, modify the disk device to use the round robin PSP.

~ # esxcli storage nmp device set --device=naa.xxx --psp=VMW PSP RR

3. Set the IOPS of the device to 3.

 $\sim$  # esxcli storage nmp psp roundrobin device config set --device naa.xxx -- type=iops --iops=3 Note: Monitor the vSphere host processor because changing these variables can affect its performance.

If desired, to automatically set the IOPS path change condition for all volumes mapped to an ESXi host, create a claim rule with the following command:

```
esxcli storage nmp satp rule add -s "VMW_SATP_ALUA" -V "COMPELNT" -P "VMW PSP RR" -O "iops=3"
```

To remove the claim rule, issue the following command:

```
esxcli storage nmp satp rule remove -s "VMW_SATP_ALUA" -V "COMPELNT" -P
"VMW PSP RR" -O "iops=3"
```

A reboot may be required for this setting to take effect.

For more information and command examples for adjusting these settings, refer to the article, <u>Adjusting</u> <u>Round Robin IOPS limit from default 1000 to 1</u> in the VMware Knowledge Base.

#### 6.9.2 Fixed policy

If the fixed policy is used, it gives the greatest control over the flow of storage traffic. However, use caution to evenly distribute the load across all host HBAs, front-end ports, fabrics, and SC Series controllers.

When using the fixed policy, if a path fails, all of the datastores using it as their preferred path will fail over to the secondary path. When service resumes, the datastores will resume I/O on their preferred path.

Here is an example of using the fixed policy:

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

Mul	tipathing Details								
De	vice: COMPELNT Fib	re Channel Disk	(naa.6000d3	100002b90000000	000003e187)				
N	Iultipathing Policies	Edit Multipathing							
	Path Selection Pol	licy Fixed (V							
	Storage Array Type	Policy VMW_S							
Paths									
	Owner Plugin	NMP	NMP						
	<ul> <li>Paths</li> </ul>								
		Refresh	Enable	Disable					
		Runtime Name		Status	Target	LUN	Preferred		
		vmhba2:C0:T	2:L1	<ul> <li>Active</li> </ul>	50:00:d3:10:00:02:b9:02 50:00:d3	1			
		vmhba1:C0:T	7:L1	<ul> <li>Active (I/O)</li> </ul>	50:00:d3:10:00:02:b9:02 50:00:d3	1	*		
		vmhba2:C0:T	5:L1	<ul> <li>Active</li> </ul>	50:00:d3:10:00:02:b9:02 50:00:d3	1			
		vmhba1:C0:T	5:L1	<ul> <li>Active</li> </ul>	50:00:d3:10:00:02:b9:02 50:00:d3	1			

Figure 19 Example of a datastore path selection policy set to fixed with a preferred path

#### 6.9.3 Most recently used policy

The most recently used (MRU) PSP is generally used with active/passive arrays (to prevent path thrashing), and is therefore not recommended for use with the SC Series because a volume can only be active on one controller at a time.

#### 6.10 Multipathing using a fixed path selection policy

Keep in mind that with a fixed policy, only the preferred path actively transfers data. To distribute the I/O loads for multiple datastores over multiple HBAs, the preferred path must be set for each datastore consistently between each host. Here are some bad and good examples:

Example 1 (bad example):

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 example):

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 distribute the load between both HBAs.

Although the fixed multipathing policy gives more fine-tuned control over which path transfers the data for each datastore, you must manually validate that all paths have proportional amounts of traffic to each ESXi host.

## 6.11 Multipathing using a round robin path selection policy

When using round robin, although it must be manually defined for each volume and host (or set to the default), it will provide path failure protection and will remove some of the guesswork of distributing load between paths manually, when compared to using a fixed policy. To reiterate from previous sections, when using round robin, be sure 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) Storage traffic is evenly distributed between all HBAs.

#### 6.12 Asymmetric logical unit access (ALUA)

The ALUA protocol was designed for arrays that VMware classifies as asymmetrical storage systems. Historically, since the SC Series is considered an active-active storage system where all the paths are active at all times (unless a path failed), ALUA was not necessary.

However, with the introduction of the SAS protocol to Dell Storage SCv2000 Series arrays, ALUA became necessary for implementation into the front end target ports for controller failover purposes. For example, for the path to the controller where the volume is active, ALUA will set the path state as active-optimized, and for the path to the secondary controller, ALUA will set the path state to standby. During a controller failover, the ALUA path state will change such that the secondary controller will become the active-optimized path.

It is important to note that for the other protocols supported by SC Series such as Fibre Channel, iSCSI, and FCoE, the pathing architecture remains mostly unchanged. Consistent with the former architecture, all of the volume paths to the controller that the volume is active on will be labeled active-optimized. Because the SC Series storage does not proxy volume traffic between controllers, the paths to the second controller will be set to the standby state, and only change during a controller failover.

**Note:** Because of the ALUA protocol addition, the vSphere storage array type plug-in (SATP) module settings, such as configuring round robin as the default PSP, will need to be changed on each host when upgrading to SCOS 6.6 and later.

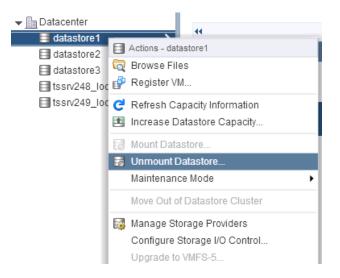
For more detailed information about ALUA in a vSphere environment, see the VMware vSphere Blog post, <u>Configuration Settings for ALUA Devices</u>.

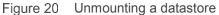
## 6.13 Unmapping volumes from an ESXi host

Within ESXi, VMware has added the ability to gracefully remove volumes before unmapping them to prevent an all-paths down (APD) state.

Before attempting this procedure, consult the section, "Performing Planned Storage Device Removal" in the *vSphere Storage Guide* at <u>VMware vSphere 6.x documentation</u> or <u>vSphere 5.x documentation</u>.

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





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

Storage Devices										
💀 💷   🗟 🖬 🛤 🥥 🔘 🗉 🔜	₽.		Q Filter	Q Filter 🔹						
Name	Туре	Capacity	Operational State	Hardware Acceleration	Drive Type	Transport				
Local USB Direct-Access (mpx.vmhba32:	disk	0.00 B	Attached	Not supported	HDD	Block Adapter				
COMPELNT Fibre Channel Disk (naa.600	disk	1.00 TB	Attached	Supported	HDD	Fibre Channel				
Local ATA Disk (t10.ATAMZ2D5EA1	disk	93.16 GB	Attached	Unknown	Flash	Block Adapter				
Local USB CD-ROM (mpx.vmhba32:C0:T0	cdrom		Attached	Not supported	HDD	Block Adapter				

Figure 21 Detaching a storage device

- 4. Repeat step 1 through step 3 for each host the volume is presented.
- 5. Within the DSM/EM client, unmap the volume.
- 6. 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 SC Series vSphere client plug-in. More information on how to obtain the plug-in is in appendix B.

## 6.14 Mapping volumes from multiple arrays

When mapping volumes from separate arrays to the same ESXi cluster, attention should be paid to the LUN when mapping volumes from each system. For administrative simplicity, it is recommended to use separate numbering schemes when mapping volumes from each of the arrays. For example, array A might map volumes using LUNs 1 through 50, while array B might use LUNs 51-100. This helps avoid numbering conflicts if a volume is ever promoted to a Live Volume between the two arrays in the future.

## 6.15 Additional multipathing resources

For more information on multipathing, see the section, "Understanding Multipathing and Failover" in the vSphere Storage Guide at VMware vSphere 6.x documentation or vSphere 5.x documentation.

# 7 Boot from SAN

Booting ESXi hosts from SAN yields both advantages and disadvantages. In some cases, such as with blade servers that do not have internal disk drives, booting from SAN may be the only option. However, many ESXi hosts can have internal mirrored drives, providing the flexibility of choice. The benefits of booting from SAN are obvious — it alleviates the need for internal drives and allows the ability to take snapshots (replays) of the boot volume.

However, there are also benefits to booting from local disks and having just the virtual machines located on SAN resources. Because it takes a short period of time to freshly load and patch an ESXi host, booting from local disks gives them the advantage of staying online, in the event that maintenance needs to be performed on 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.

The decision to boot from SAN depends on many business-related factors including cost, recoverability, and configuration needs. Therefore, Dell does not offer a specific recommendation.

# 7.1 Configuring boot from SAN

When deciding to boot ESXi hosts from SAN, a few best practices 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 be added to the boot volume.

To use the advanced mapping screen in DSM/EM, it must be enabled through the **Preferences** menu in the SC Series settings.

🚆 Edit Settings [SC12]			3
General	Storage Center User F	Preferences	
License	Volume		
	Volume Size	500 GB 👻	
Preferences		Show Block Size Option	
Storage	Base Volume Name	New Volume	
Alerts and Logs	Read Cache	Enabled	
Phone Home	Write Cache	C Enabled	
Automated Reports		Allow Cache Selection	
	Compression	Enabled	
Secure Console		Allow Compression Selection	
SMTP Server	Replay		
SNMP Server	Replay Profiles Minimum Replay Interval	Daily 5	Change
Time	Server		
IP Filtering	Allow Advanced Mapping Operating Cystem	VIMWare ESXI'S.5	
Users and User Groups	Storage	WWWARE EDVI 2.2	
Password Configuration	Storage Profile	Recommended (All Tiers)	
		Allow Storage Profile Selection	
	Storage Type	Assigned - Redundant - 2 MB	
		Allow Storage Type Selection	
		Only use Standard 2MB Datapage Size	
		Only use Standard Redundant Storage	
? Help		🗶 Cancel 🔺 Apply	🛧 ок

Figure 22 Enabling advanced mapping

Once advanced mapping is enabled, there are a few options that need to be changed to force mapping down a single path for the initial installation, as shown in Figure 23 and the settings that follow.

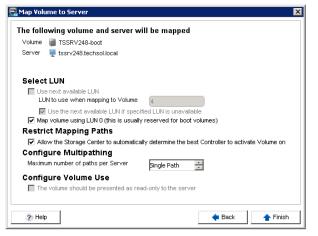


Figure 23 Advanced mapping screen for configuring boot from SAN

#### Map volume using LUN 0: Checked

#### Maximum number of paths allowed: Single-path

Once the ESXi host is running correctly, the second path can then be added to the boot volume by modifying the mapping. To do this, right-click the mapping, select **Modify Mapping**, and select the following:

#### Maximum number of paths allowed: Unlimited

Once the additional path has been added, the HBAs on the ESXi host can be rescanned.

# 8 Volume creation and sizing

Administrators are tasked with complex decisions such as determining the best volume size, number of virtual machines per datastore, and file system versions for their environment. This section provides guidance for making those decisions.

### 8.1 Volume sizing and the 64 TB limit

Although the maximum size of a volume that can be presented to ESXi is 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 prudent to start with datastore sizes in the 500–750 GB range. This is based on the consideration that a 750 GB datastore will accommodate approximately 15 x 40 GB virtual disks, leaving a small amount of overhead for virtual machine configuration files, logs, snapshots (replays), and memory swap, keeping the datastore performing adequately.

The largest single extent 64 TB VMFS-5 volume size is  $(64 \times 1,024 \times 1,024 \times 1,024 \times 1,024 = 70,368,744,177,664)$  bytes. For any volumes larger than this size, VMware will simply not consume the additional space.

**Note:** These sizing recommendations are provided to limit the number of virtual machines on each datastore and to keep performance manageable, not for actual capacity reasons. If there is a virtual machine that requires a large virtual disk, creating a large datastore to accommodate the VM is acceptable.

**Note:** Within certain versions of ESXi 5.x, hosts may have maximum addressable storage limits of 4 TB, 8 TB, or 60 TB due to the VMFS heap size. Read the following VMware KB article for more information: http://kb.vmware.com/kb/1004424.

## 8.2 Virtual machines per datastore

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

The reason behind keeping a limited number of virtual machines and/or VMDK files per datastore is due to potential I/O contention, queue depth contention, or other conflicts that may degrade system performance. For the same reason, it is advisable to limit datastore sizes to 500–750 GB. This helps limit the total number of virtual machines that can be placed on each datastore.

The art to virtual-machine placement revolves 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. For example, in some cases, the appropriate number for high-I/O-load virtual machines may be less than five, 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 15–25 virtual machines, and increase or decrease the number of virtual machines on each datastore as needed. Moving virtual machines between datastores can even be done non-disruptively when licensed to use the VMware Storage vMotion feature.

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 and increasing disk latency. Remember that if the driver module is set to a queue depth of 256, 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 x 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.

In rare situations where vStorage APIs for array integration (VAAI) hardware-assisted locking is unavailable or disabled, a far less common indicator that a datastore has too many virtual machines placed on it would be the frequent occurrence of SCSI reservation conflicts. Within esxtop, there is a field in the Disk Device screen for reserve stats (RESVSTATS). When monitoring datastores, it is normal to see a few reservation (RESV/s) entries and even a few conflicts (CONS/s) from time to time. However, when CONS/s happen 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. If datastores are impacted, it is recommended to upgrade to a version of SCOS that supports this VAAI primitive. See section 16.3 on VAAI for more information.

**Note:** Many resources discuss VMware infrastructure design and sizing. The general rule discussed previously may vary based on the needs of the environment.

# 8.3 VMFS partition alignment

Partition alignment is a performance-tuning technique used with traditional SANs to align the guest operating system partitions and VMFS partitions to the physical spinning 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 SC Series automatically aligns its 512K, 2M, or 4M pages to the physical sector boundaries of spinning drives. Since the largest percentage of performance gains are seen from aligning the SC Series pages to the physical disks, the remaining areas that can be aligned and tuned have a minor effect on performance.

As found in Dell EMC internal lab testing, any performance gains achieved by manually aligning partitions are usually not substantial enough  $(\pm 1\%)$  to justify the extra effort. However, because all workloads are different, before deciding whether or not to align VMFS partitions, it is recommended that testing is performed to determine the impact of an aligned partition on particular applications.

To manually align the VMFS block boundaries to the SC Series page boundaries for performance testing, the recommended offset when creating a new datastore is 8,192 (4 MB).

**Note:** Using the SC Series vSphere client plug-in to create new datastores will automatically align them to the recommended offset. More information on how to obtain the plug-in is in appendix B.

Figure 24 shows an example of a fully aligned partition in the SC Series where one guest I/O will only access necessary physical disk sectors.

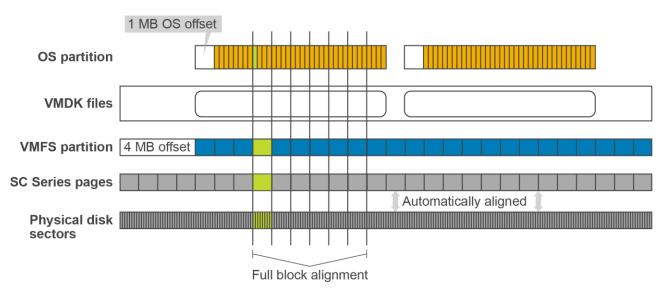


Figure 24 Fully aligned partition in SC Series

Figure 25 shows an example of an unaligned partition in a traditional SAN where performance can be improved by alignment.

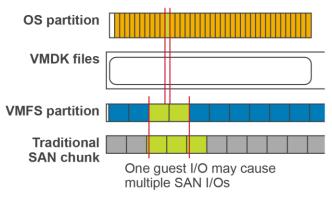


Figure 25 Unaligned partition in a traditional SAN

## 8.4 VMFS file systems and block sizes

Within ESXi, it is recommended to use VMFS-5, which is the recommended file system format selectable from the vSphere client when creating a new datastore.

#### 8.4.1 VMFS-3

If there are any remaining VMFS-3 datastores in the environment, it is recommended they be retired and virtual machines be migrated to the latest-version VMFS-5 formatted datastore. As a best practice, create new VMFS-5 datastores, and migrate virtual machines to them using Storage vMotion when possible.

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

#### 8.4.2 VMFS-5

With VMFS-5, the default block size is 1 MB, and allows for up to a 64 TB datastore with up to a 62 TB VMDK. This format is required for functionality such as the VAAI space reclamation primitive (SCSI UNMAP) to reclaim storage after a VMDK is deleted. See section 16.3 for more information on VAAI.

#### 8.4.3 VMFS-6

VMFS-6 shares many of the same configuration maximums as VMFS-5, however VMFS-6 allows for automatic space reclamation (UNMAP) for thin provisioned volumes. Since there is no upgrade path from VMFS-5 to VMFS-6, migration techniques such as Storage vMotion must be used to migrate virtual machines to newly created VMFS-6 datastores to take advantage of the new features.

# 9 Volume mapping layout

In addition to volume sizing, another important factor to consider is the placement of files and virtual machine data.

## 9.1 Multiple virtual machines per volume

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 results in higher consolidation ratios. When deciding how to lay out 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 the layout of the virtual machines, there are some basic concepts that should be considered.

#### 9.1.1 Storage of non-virtual machine files

One general recommendation is to create a content library or a datastore for administrative items. This can be used to store virtual machine templates, ISO images, virtual floppies, and/or scripts.

#### 9.1.2 Separation of operating system page files

In situations where memory swapping cannot be avoided, one technique to consider with virtual machine placement is separating the operating system paging files or swap files onto a separate datastore.

There are two main reasons for separating operating system page files onto their own datastore. First, because page files can generate a lot of disk activity when the memory in the virtual machine or ESXi host runs low, this practice could keep volume snapshots (replays) smaller. Second, if replicating those volumes, this practice will conserve bandwidth by not replicating the operating system page file data.

Depending on the memory-swap conditions unique to each environment, separating page files may or may not make a significant reduction in snapshot sizes. A good way to determine whether separating page files 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 page files to determine the actual impact.

If separating page files will make an impact in reducing snapshot sizes, the general recommendation is to create pairs of volumes for each datastore containing virtual machines. If a volume is created to contain 10 virtual machines, then a second volume should be created to store the operating system page files for those 10 machines. For example:

**Create one datastore for virtual machines:** This will usually contain the virtual machine disk (VMDK) files, configuration files, and logs for the virtual machines.

**Create one paired datastore for the corresponding virtual machine page files:** This should contain virtual machine page files. Using Windows as an example, create a virtual disk (P:) on this volume large enough to store the Windows paging file for each virtual machine. This volume can be sized considerably smaller than the main datastore since it only needs enough space to store page files.

A question often asked is should all of the operating system page files be placed on a single datastore. This is not a good practice for the following two reasons.

First, the page file datastore can also experience contention from queue depth utilization or disk I/O. 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 storage performance decrease.

Second, that datastore could become a single point of failure. Operating systems are usually not tolerant of unexpected disk-drive removal. If an administrator were to accidentally unmap the page file volume, the number of virtual machines within the failure domain would be isolated to a subset of the virtual machines instead of all the virtual machines.

#### 9.1.3 Separation of virtual machine swap files

Each virtual machine also has a memory swap file located in its home directory that is used by the ESXi host when the VMware Tools balloon driver is unable to reclaim enough memory. In other words, the virtual machine swap (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 volume. While doing this may help to reduce snapshot sizes and preserve replication bandwidth, it should only be done under the guidance of VMware support.

#### 9.1.4 Virtual machine placement

This example technique gives a great deal of flexibility when building out the storage architecture in the environment, while keeping with the basic concepts discussed above. The following example layout meets most virtual infrastructure needs because it adds the flexibility of being able to add RDMs 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 (drives C, D, and E)

LUN11 – Page file (Paired with LUN10) for OS paging files [If necessary]

LUN12 - LUN19 - Reserved LUNs for virtual machine RDMs for machines in this group

LUN20 - OS/DATA (drives C, D, and E)

LUN21 - Page file (Paired with LUN20) for OS paging files [If desired]

LUN22 - LUN29 - Reserved LUNs for virtual machine RDMs for machines in this group

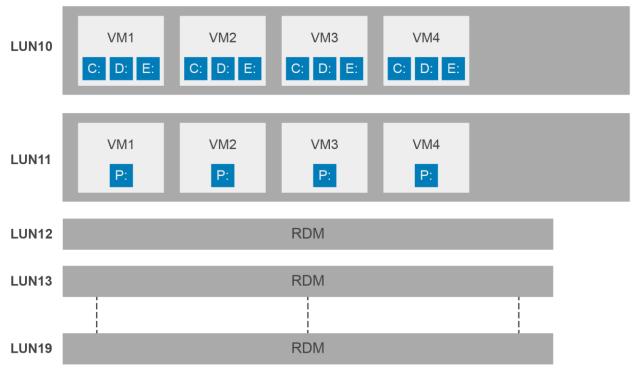


Figure 26 Virtual machine placement (with RDMs)

**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 improves organization as the clusters grow larger.

**Note:** Many factors may influence architecting storage with respect to the placement of virtual machines. The method shown above is merely a suggestion; business needs may dictate different alternatives.

# 9.2 One virtual machine per volume

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. Using a 1:1 virtual machine-to-datastore ratio should be done by exception, not by rule.

Advantages of creating one volume per virtual machine include:

- Granularity in replication: Since the SC Series replicates at the volume level, if there is one virtual machine per volume, administrators can choose which virtual machine to replicate.
- Reduced I/O contention: A single volume is dedicated to a single virtual machine.
- Flexibility with volume mappings: Since a path can be individually assigned to each volume, 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.
- Simplified backup/restore of an entire virtual machine: If a VM needs to be restored, an administrator can unmap or remap a snapshot in its place.

Disadvantages of creating one volume per virtual machine include:

- Maximum virtual machines equal to disk device maximums in the ESXi cluster. For example, if the HBA has a maximum limit of 512 volumes that can be mapped to the ESXi host, since each logical unit number can be used only once when mapping across multiple ESXi hosts, it would have a 512 virtual machine limit (assuming that no extra LUNs would be needed for recoveries).
- Increased administrative overhead: Managing a volume for each virtual machine and all the corresponding mappings may be challenging.

# 10 Raw device mapping (RDM)

A raw device mapping (RDM) is used to map a particular volume 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 RDMs, but in most cases, using the VMFS datastores is recommended over using RDMs.

Advantages of using RDMs include:

- Virtual mode RDMs (vRDMs) up to 62 TB in size and physical mode RDMs (pRDMs) up to 64 TB in size can be mapped directly to a guest.
- A clustered resource (such as Microsoft Cluster Services) can be created with options including:
  - 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.
- Physical machines can be converted to virtual machines more easily because a physical machine volume can be mapped as an RDM.
- When a VM has special disk performance needs:
  - There may be a slight disk performance increase when using an RDM compared to a VMFS virtual disk due to the lack of contention, no VMFS write penalties, and better queue depth utilization.
  - There can be independent disk queues per RDM.
  - Certain types of SAN software can be used, such as Dell Replay Manager or the Windows free space recovery feature. See the Knowledge Center at the <u>SC Series Customer Portal</u> (login required) for more information about these features.
- A different storage profile can be assigned to each volume. For example, if a database server has its database and logs separated onto different volumes, each can have a separate storage profile.
- A different snapshot schedule (replay profile) can be used with each volume. For example, a database and its transaction logs may have different snapshot (replay) intervals and retention periods for expiration.
- Virtual mode RDMs support vSphere snapshots.

Disadvantages of using RDMs include:

- This may add administrative overhead due to the number of mappings.
- There are a limited number of volumes that can be mapped to an ESXi host. If every virtual machine used RDMs for drives, the cluster would have a relatively small maximum number of drives.
- Physical mode RDMs cannot be used in conjunction with ESXi snapshots. While VMware snapshots are not available for physical mode RDMs, SC Series snapshots (replays) can still be used to recover data

**Note:** All of the previous RDM-related tasks can be automated using the Dell SC Series vSphere Client Plug-in. More information on how to obtain the plug-in is in appendix B.

# 11 Data Progression and storage profile selection

Data Progression migrates inactive data to the lower tier of inexpensive storage while keeping the most active data on the highest tier of fast storage, as shown in Figure 27. This works to the advantage of VMware datastores because multiple virtual machines are usually kept on a single volume.

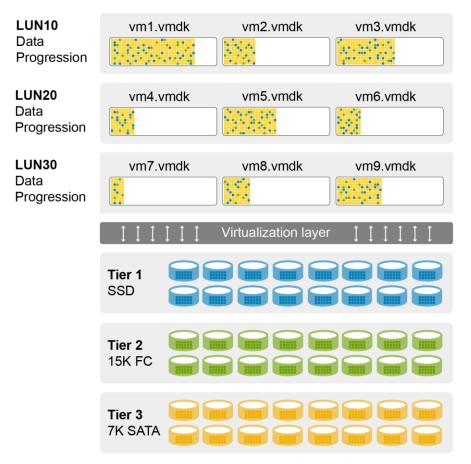


Figure 27 Data Progression and thin provisioning working together

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

However, if there is a business case where particular virtual machines would require different RAID types, some decisions on how Data Progression is configured for the volumes must be made.

The following is an advanced example of virtual machine RAID groupings in which forcing particular volumes into different profiles is desired.

LUN0 - Boot volume for ESXi

-- Data Progression: Recommended (All Tiers)

LUN1 - Templates/ISO/General Storage

-- Data Progression: Recommended (All Tiers)

LUN10 - OS/DATA (Server group1 - High performance - 4 VM's - drives C, D, and E)

-- High Priority (Tier 1)

LUN20 - OS/DATA (Server group2 - Low performance - 15 VM's – drives C, D, and E)

-- Data Progression: Low Priority (Tier 3)

LUN30 - OS/DATA (Server group 3 - Application grouping - 5 VM's - drives C, D, and E)

-- Data Progression: Recommended (All Tiers)

As stated at the beginning of this section, unless there is specific business need that requires a particular virtual machine or application to be pinned into a specific storage type, tier, or RAID level, it is recommended to keep the configuration simple. In most cases, the **Recommended** Data Progression setting can be used to automatically classify and migrate data based upon usage.

**Note:** As a Data Progression best practice, create a snapshot schedule (replay profile) for each volume that takes one daily snapshot (replay) at a minimum and does not expire for 25 hours or more. This will have a dramatic effect on Data Progression behavior and can increase the overall system performance.

## 11.1 On-Demand Data Progression

With the release of SCOS 6.4 and Dell all-flash arrays, On-Demand Data Progression (ODDP) was introduced not only to enhance flash performance capabilities, but also to add new tiering possibilities with spinning media. In essence, this new capability allows Data Progression to run outside of its normal cycle to move data according to rules defined in the storage profile. For example, with an all-flash system, when ODDP is triggered by a snapshot, it can migrate data from the write-intensive SSD drives to the read-intensive SSD drives to free up space for new incoming writes.

Within a VMware environment, this new functionality may change the storage profile strategies previously implemented in the environment. With this in mind, it is recommended that administrators review all of the new storage profile selections, such as the ones shown in Figure 28, to see if improvements can be made to their existing tiering strategies.

C SC22
🚊 🚽 🤯 Storage
🕀 🕡 Volumes
🖮 🤖 Replay Profiles
🖨 🎬 Storage Profiles
Low Priority (Tier 3)
Write Intensive (Tier 1)
Flash Only with Progression (Tier 1 to Tier 2)
Low Priority with Progression (Tier 3 to Tier 2)

Figure 28 New storage profile available with Storage Center OS 6.4

For detailed information about On-Demand Data Progression, read the *Storage Center System Manager Administrator's Guide* available on Knowledge Center at the <u>SC Series Customer Portal</u> (login required).

## 11.2 Data reduction (compression and deduplication)

Released with SCOS 7.0, the data reduction feature takes advantage of both compression and deduplication for reducing the data footprint. As part of its daily processing cycle, Data Progression deduplicates and compresses the data within volumes that have the data reduction feature enabled. Depending on the data reduction policy set on a particular volume, the data can either be compressed, or deduplicated and then compressed for maximum savings.

Storage	
Data Reduction	Profile

Deduplication with Compression	¥
None	
Compression	
Deduplication with Compression	

Figure 29 Data reduction options within volume settings

**Note:** Depending on factors such as array hardware and application workloads, enabling data reduction may have a performance tradeoff in exchange for the capacity savings.

The advanced options available in the volume settings offer control over the data reduction input, which consists of pages an administrator deems eligible for data reduction. The choice between inaccessible snapshot pages and all snapshot pages is effectively a tradeoff between maximum savings and maximum performance.

### 11.2.1 Data Reduction Input

Within the advanced volume settings, the **Data Reduction Input** setting can be specified to control which pages are eligible for data reduction.

Edit Advance	ed Volume Settings		X
Storage Allow Snapshots to coalesce into active Snapshot Import to lowest tier Data Reduction Input Limits	All Snapshot Pages	~	
Space Consumption Limit	All Snapshot Pages	~	
Operating System OpenVMS Unique Disk ID Related View Volume Maximums	7		
Maximum Volume Count	100		Current 1
Maximum Configured Volume Space	500 TB	~	Current 500 GB
? Help	[		Cancel 🔶 OK

Figure 30 Advanced volume settings showing the Data Reduction Input selection

**All Snapshot Pages:** All frozen pages in the lowest tier of the system that are part of a snapshot (replay), are eligible for data reduction. In the case of a single-tiered system, such as a replication target system, all frozen pages are eligible. This setting yields the highest reductions, however there is a performance tradeoff of higher latency due to any accessed pages being decompressed as they are accessed.

**Inaccessible Snapshot Pages:** Only frozen inaccessible pages are eligible for data reduction. These are pages kept only as part of a snapshot for recovery purposes, and are not accessible by the host. This setting yields the lowest reductions, however it has the best performance since no host-accessible pages are compressed or deduplicated.

**Caution:** When initially enabling compression on a system, the Data Progression total running time may be extended and could result in the cycle continuing into regular business hours. It is recommended that administrators adjust the Data Progression settings within the system settings to limit the maximum run time, or limit the number of volumes to be compressed each day.

For more information about the data reduction feature, read the white paper, <u>Dell Storage Center SCOS 7.0</u> <u>Data Reduction with Deduplication and Compression</u>.

# 12 Thin provisioning and virtual disks

Dell SC Series thin provisioning allows less storage to be consumed for virtual machines, thereby saving upfront storage costs. This section describes the relationship that this feature has with virtual machine storage.

# 12.1 Virtual disk formats

In ESXi, VMFS can store virtual disks using one of the four different formats described in the following sections.

🔁 New Virtual Machine		? >>
1 Select creation type	Virtual Hardware VM Options	SDRS Rules
<ul> <li>1a Select a creation type</li> <li>2 Edit settings</li> </ul>	👻 🛄 New Hard disk	40 GB v
<ul> <li>2a Select a name and folder</li> </ul>	Maximum Size	1,022.79 GB
✓ 2b Select a compute resource	VM storage policy	Datastore Default
✓ 2c Select storage	Location Disk Provisioning	Store with the virtual machine   Thick provision lazy zeroed
<ul> <li>2d Select compatibility</li> <li>2e Select a guest OS</li> </ul>	Disk Tonoioning	Thick provision eager zeroed Thin provision
2f Customize hardware	Shares	Normal v 1000
3 Ready to complete	Limit - IOPs	Unlimited
	Virtual flash read cache	0 GB - Advanced
	Virtual Device Node	New SCSI controll   • SCSI(0:0)   •
	Disk Mode	Dependent 🔹 🛈
	▶ 🛃 New SCSI controller	LSI Logic SAS
	▶ 属 New Network	VM Network 🔍 🗹 Connect
	New device:	Select 🗸 🗸
		Compatibility: ESXi 6.0 and later (VM version 11)
		Back Next Finish Cancel

Figure 31 Virtual disk format selection

### 12.1.1 Thick provision lazy zeroed (zeroedthick)

Only a small amount of disk space is used within the SC Series at virtual disk creation time, and new blocks are only allocated on the SC Series 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 default, on-first-write style of block zeroing 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.

### 12.1.2 Thick provision eager zeroed (eagerzeroedthick)

Space required for the virtual disk is fully allocated at creation time. Unlike the zeroedthick format, all of the data blocks within the virtual disk are zeroed out during creation. Disks in this format may take longer to create than other types of disks because all of the blocks must be zeroed out before the disks can be used. When using VAAI, the time it takes to create an eagerzeroedthick disk is greatly reduced. This format is generally reserved for Microsoft clusters and the highest-I/O-workload virtual machines because it does not suffer from operational write penalties like the zeroedthick or thin formats.

#### 12.1.3 Thin provisioned (thin)

The logical space required for the virtual disk is not allocated during creation, but it is allocated on demand during the 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.

### 12.1.4 Space efficient sparse (SE sparse disks)

SE sparse disks are a new virtual disk type introduced for use with VMware Horizon View Composer. These virtual disks are used with linked clones for improved performance, space reclamation, and a new 4 KB grain size.

### 12.2 Thin provisioning relationship

The following points describe how each virtual disk format affects SC Series thin provisioning:

- Thick provision lazy zeroed: Virtual disks are thin provisioned by the SC Series.
- Thick provision eager zeroed: Virtual disks are thin provisioned by the SC Series (see section 12.3).
- Thin provisioned: Virtual disks are thin provisioned by the SC Series. There are no additional storage savings while using this format because the array already uses its thin provisioning (see section 12.4).
- SE sparse (available with VMware Horizon View virtual desktops): Virtual disks are thin provisioned by the SC Series. Space reclamation within the VMDK is available.

As a best practice, use the default virtual disk format of thick provision lazy zeroed 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 on-first-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.

# 12.3 SC Series thin write functionality

Most versions of SCOS have the ability to detect incoming sequential zeroes while being written and 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.

### 12.4 SC Series thin provisioning or VMware thin provisioning

A common question is whether to use array-based thin provisioning or the VMware thin provisioned VMDK format. Since SC Series 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 thin provisioning is needed, 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 Storage Manager can also be used to alert for low-space conditions.

## 12.5 Windows free space recovery

One of the nuances of the Windows NTFS file system with Windows Server<sup>®</sup> 2012 and prior, is that gradually over time, the actual usage of the file system can grow apart from what the SC Series reports as being allocated. In this example, 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, thin provisioning (Dynamic Capacity) has assigned those blocks to that volume, so the SC Series 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 SC Series to determine if an allocated block is actually still in use by the OS. However, the DSM/EM server agent contains the necessary functionality to recover this free space from machines running earlier versions of Windows that do not possess SCSI UNMAP capabilities. 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 that blocks kept as part of a snapshot (replay) cannot be freed until that snapshot 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 an RDM 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 SC Series that can be freed. The disk must be an NTFS basic disk (either MBR or GPT).
- The virtual disk cannot be a VMDK or an RDM set to virtual compatibility mode (vRDM).
  - VMware does not provide the necessary APIs 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 drive C (VMDK) and a drive D (RDMP), Windows free space recovery will only be able to reclaim space for the drive D.
  - The restriction against using vRDMs 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, an alternative method of backing up the pRDMs is needed. For example, the Dell
    Storage PowerShell SDK installation provides an example PowerShell script that can be used to
    back up physical mode RDMs as part of the pre-execution steps of the backup job.
- The free space recovery agent also works with volumes mapped directly to the virtual machine through the Microsoft software iSCSI initiator. Volumes mapped to the virtual machine through the Microsoft iSCSI initiator interact with the SAN directly, and consequently, space recovery works as intended.

For more information on Windows free space recovery and compatible versions of Windows, consult the *Dell Storage Manager Administrator's Guide* on the Knowledge Center at the <u>SC Series Customer Portal</u> (login required).

# 13 Extending VMware volumes

Within an ESXi host, there are three ways to extend or grow storage. This section provides the general steps required. Additional information can be found in the *vSphere Storage Guide*, section "Increasing VMFS Datastore Capacity", and in the *vSphere Virtual Machine Administration Guide*, section "Virtual Disk Configuration" located within the <u>VMware vSphere 6.x documentation</u> or <u>vSphere 5.x documentation</u>.

### 13.1 Increasing the size of VMFS datastores

Two methods can be used to increase the size of VMFS datastores: Expanding an existing extent (recommended), or adding a new extent (not recommended).

#### 13.1.1 Expanding an extent in an existing VMFS datastore

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



Figure 32 Growing Datastore3 using the adjacent 100 GB (in green)

In Figure 32, extending the space at the end of an SC Series volume can be done using Dell Storage Manager or the vSphere client plug-ins. After the volume has been extended and the host HBA has been rescanned, click **Increase...** to edit the properties of the datastore to grow and follow the Increase Datastore Capacity wizard instructions.

Be sure to select the **Expandable** volume, otherwise a VMFS extent will be added to the datastore (see section 13.1.2 on VMFS extents).

当 datastore4 - Increase Datastore Capacity								
1 Select Device						Q Filter	•	
2 Specify Configuration	Name	LUN	Capacity	Hardware Accel	Drive Type	Expandable		
3 Ready To Complete	COMPELNT Fibre Channel Disk (naa.6000d310	4	2.00 TB	Supported	HDD	Yes		

Figure 33 Extending a 2 TB datastore by 1 TB

**Note:** As an alternative to extending a datastore volume when a virtual machine needs additional disk space, create a new datastore volume and migrate that virtual machine. This helps to keep volume sizes manageable and helps 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 SC Series vSphere Client Plug-in. More information on how to obtain the plug-in is in appendix B.

#### 13.1.2 Adding a new extent to an existing datastore

This legacy functionality was used with previous versions of vSphere to concatenate multiple volumes to create VMFS-3 datastores larger than 2 TB. Since the maximum datastore size with VMFS-5 has been increased to 64 TB, the use of extents is no longer necessary.

**Caution:** Due to the complexities of coordinating snapshots (replays) and recoveries of datastores that are spanned across multiple SC Series volumes, the use of VMFS extents is highly discouraged. However, if the use of extents is absolutely necessary, snapshots of those volumes should be taken using the consistent snapshot schedule (replay profile) functionality.

### 13.2 Increasing the size of a virtual machine disk (VMDK) file

Hot extending a SCSI virtual disk is available from within the vSphere client when editing the settings of a virtual machine.

📴 webserver01 - Edit Setting	]\$
Virtual Hardware VM Option	as SDRS Rules vApp Options
▶ 🔲 CPU	1 🔹 🕑
▶ 🏧 Memory	4096 <b>v</b> MB <b>v</b>
👻 🚍 *Hard disk 1	50 🔶 GB 🖵
Maximum Size	2.00 TB

Figure 34 Growing a virtual disk from the virtual machine properties screen

After growing the virtual disk from the vSphere client, an administrator must log into the virtual machine, rescan for new disks, and then extend the file system.

**Caution:** Microsoft does not support extending the system partition (drive C) of a machine in certain versions of Windows.

## 13.3 Increasing the size of 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 within the DSM client, rescan disks from Windows disk management, and then use DISKPART or the Storage MMC console to extend the drive.

A physical mode RDM whose placeholder VMDK resides on a VMFS-5 datastore can be extended up to the 64 TB limit. However, depending on the version, vRDMs may have a limit anywhere between 2 TB to 62 TB.

**Caution:** Similar to datastore volumes, it is also very important not to extend an RDM volume with its pointer file residing on a VMFS-3 datastore past the 2047 GB (1.99 TB) limit.

# 14 Snapshots (replays) and virtual machine backups

Backup and recovery are important to any virtualized infrastructure. This section discusses several common techniques to improve the robustness of virtualized environments such as using snapshots (replays) and virtual machine backups

### 14.1 Backing up virtual machines

The key to any good backup strategy is to test the backup and verify the results. There are many ways to back up virtual machines, but depending on business needs, each solution is usually unique to each environment. Testing and verification may prove that one solution works better than another does, so it is best to test a few different options.

Since the subject of virtual machine backup is vast, this section only covers a few basics. For more information about virtual machine backup strategies, consult VMware documentation.

#### 14.1.1 Backing up virtual machines to tape or disk

Perhaps the most common methods of backing up virtual machines to tape are using third-party backup software or using the backup client software installed within the guest. Options include:

Backup software using vStorage APIs for Data Protection (VADP): This is a successor to VMware Consolidated Backup and provides an integrated method to back up virtual machines for backup solutions such as Dell NetVault<sup>™</sup> Backup for VMware. For more information, see <u>VMware vSphere Storage APIs</u> - <u>Data Protection (VADP) FAQ</u> in the VMware Knowledge Base.

**vSphere Data Protection:** This is a software package designed to allow operational backup and recovery of virtual machines within a vSphere environment. For more information, see <u>vSphere Data Protection</u>.

**Backup client loaded within the guest:** Using this method, traditional 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.

#### 14.1.2 Backing up virtual machines using snapshots

The options for backing up virtual machines using SC Series snapshots are described as follows:

**Snapshots scheduled from within the DSM/EM client:** From within the DSM/EM client, a snapshot schedule (replay profile) can be created to schedule snapshots of virtual machine volumes. In most cases, using snapshots to back up virtual machines is sufficient to perform a standard recovery. It is important to remember that snapshots 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 is recovered 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.

**Snapshots taken using Dell Replay Manager software:** Since virtual machines running transactional databases are more sensitive to crash-consistent data, Dell has developed its Replay Manger software to

utilize the Microsoft VSS framework for taking snapshots of VSS-aware applications. This software package will ensure that the application data is in a consistent state before executing the snapshot.

**Snapshots taken using Dell scripting tools**: For applications that need a custom method for taking consistent snapshots of the data, Dell has developed two scripting tools:

- Dell Storage PowerShell SDK: This scripting tool also allows scripting for many of the same storage tasks using the Microsoft PowerShell scripting language.
- Dell Compellent Command Utility (CompCU): A Java-based scripting tool that allows scripting for many of the SC Series tasks (such as taking snapshots).

A good example of using one of these scripting utilities is writing a script to take a snapshot of an Oracle database after it is put into hot backup mode.

**Snapshots used for SC Series replication and VMware Site Recovery Manager:** Replicating snapshots to a disaster recovery site not only ensures an off-site backup, but when using Site Recovery Manager, it provides an automated recovery of the virtual infrastructure in the event a disaster is declared.

## 14.2 Recovering virtual machine data from a snapshot

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

The basic steps are as follows:

- 1. Within the DSM/EM client, select the snapshot to recover from and select **Create Volume From Replay**.
- Continue through the local recovery wizard to create the view volume, and map it to the ESXi host designated to recover the data. Be sure to map the recovery view volume using a LUN that is not already in use.
- 3. Select a host or cluster, select **Storage**, and select **Rescan Storage** to detect the new volumes.
- 4. With the host or cluster still selected, select **Storage**, select **New Datastore** to open the wizard, and take the following steps:
  - a. Location: Verify the cluster or host and click Next.
  - b. Type: Select VMFS and click Next.
  - c. Datastore name: Enter a name for the recovery datastore.
    - i. Select a host that has visibility to the new volume.
    - ii. Select the LUN for the view volume that was just mapped to the host and click Next.
  - d. Mount option: Select from three options:
    - > Assign a New Signature: This option regenerates the datastore signature so that it can be accessed by the host (select this option if you are unsure of which option to use).
    - > Keep the Existing Signature: This option should only be used if the original datastore is not present on the host.
    - > Format the disk: This option formats the view volume and creates a new datastore from it.
  - e. Ready to complete: Verify all selections and select Finish.

Once the datastore has been resignatured, the snap datastore is accessible.

Datacenter
datastore1
datastore2
datastore3
datastore4
snap-4b51effb-datastore1
tssrv248\_local
tssrv249\_local

Figure 35 The datastores tab showing the snapshot datastore

The recovery datastore is now designated with *snap-xxxxxx-originalname*. From this tab, the datastore can be browsed to perform the recovery using one of the methods listed below.

**Note:** All of the above tasks can be automated by using the recovery functionality in the Dell SC Series vSphere Client Plug-ins. More information on how to obtain the plug-in is in appendix B.

#### 14.2.1 Recovering a file from a virtual disk

To recover a file from within a virtual disk located on this snap datastore:

- 1. Select Edit Settings for the virtual machine.
- 2. Choose Add an Existing Hard Disk.
- 3. Select Use an Existing Virtual Disk.
- 4. Browse to select the virtual disk to recover from, and browse the snap datastore for the virtual disk containing the data to be recovered.
- 5. From within the operating system, assign a drive letter to the virtual disk and recover, copy, or move the file back to its original location.
- 6. After completing the file recovery, remove the recovered virtual disk from the virtual machine before unmapping or deleting the view volume.

#### 14.2.2 Recovering an entire virtual disk

To recover an entire virtual disk from the snap datastore:

- 1. Browse to the virtual disk to be recovered.
- 2. Right-click the virtual disk and select **Move to**.
- 3. Using the wizard, browse to the destination datastore and folder, and click OK.

If a VMDK file is being moved back to its original location, remember that the virtual machine must be powered off to overwrite the virtual disk. 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 can be re-added to the virtual machine, and then Storage vMotion can be used to move the virtual disk back to the original datastore while the VM is powered on.

#### 14.2.3 Recovering an entire virtual machine

To recover an entire virtual machine from the snap datastore:

- 1. Browse to the virtual machine configuration file (\*.vmx).
- 2. Right-click the file and select Register VM.
- 3. Use 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 one of the virtual machines onto an isolated network or private vSwitch.

**Note:** Once the virtual machine has been recovered, it can be migrated back to the original datastore using Storage vMotion, if the original datastore has enough space, or the original virtual machine was deleted.

# 15 Replication and remote recovery

SC Series replication in coordination with the vSphere line of products can provide a robust disaster recovery solution. Because each replication method affects recovery differently, choosing the correct method to meet business requirements is important. This section provides a brief summary of the different options.

### 15.1 Synchronous replication

In a synchronous replication, the data is replicated in real-time to the destination, and 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 because they need to be highly available with low latencies. High latencies across the link will slow down access times on the source volume.

For SCOS versions prior to 6.3, the downside of synchronous replication was that snapshots (replays) on the source volume were not replicated to the destination, and any disruption to the link would force the entire volume to be re-replicated from scratch. In versions 6.3 and later, the synchronous replication engine was completely re-written to remedy these limitations.

In SCOS 6.3 and later, in addition to snapshots being replicated, two synchronous replication modes were introduced, high availability and high consistency, to control how the source volume behaves when the destination volume becomes unavailable.

**High availability mode:** This accepts and journals writes to the source volume when the destination volume is unavailable (or when latency is too high) to avoid interrupting service. However, if writes are accepted to the source volume, the destination volume data becomes stale.

**High consistency mode:** This prevents writes to the source volume when the destination volume is unavailable to guarantee that the volumes remain identical. However, the source volume cannot be modified during this time, which can interrupt operations.

Keep in mind that synchronous replication does not make both the source and destination volumes read/writeable. That functionality is inherent within the SC Series Live Volume feature.

## 15.2 Asynchronous replication

In an asynchronous replication, the I/O needs only to be committed and acknowledged to the source system, and 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:

**Frozen snapshot:** The snapshot schedule dictates how often data is sent to the destination. When each snapshot (replay) is taken, the SC Series determines which blocks have changed since the last snapshot (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 snapshot:** With this method, the data is transferred in near real-time to the destination, potentially requiring more bandwidth than if the system were just replicating the snapshots. 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. Another benefit of asynchronous replication is that snapshots are transferred to the destination volume, allowing for checkpoints at the source system and destination system.

### 15.3 Replication considerations with standard replications

One key consideration about SC Series replication is that when a volume is replicated either synchronously or asynchronously, the replication only flows in one direction. Changes made to the destination volume are not replicated back to the source. That is why it is extremely important not to map the replication destination volume directly to a host, but instead to create a read-writable view volume.

Block changes are not replicated bi-directionally 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 with a standard replication, but is possible with Live Volume.

There are some best practices for replication and remote recovery to consider. ESXi host hardware is needed at the DR site in which to map replicated volumes in the event the source ESXi cluster becomes inoperable. Also, preparations should be made to replicate all of the Virtual Center resources to the DR site as well. This includes vCenter, DSM, and other management software hosts.

### 15.4 Replication considerations with Live Volumes

Live Volume is a feature that allows a volume to be accessed from two disparate SC Series systems, allowing for the migration of data, increased uptime, and planned maintenance. Although this technology enables new functionality such as long distance vMotion, there are some caveats described in the following sections that administrators need to be aware of when choosing to use asynchronous or synchronous Live Volume.

**Caution:** LUN conflicts should be considered when designing Live Volumes into the environment's architecture. When a Live Volume is created, it attempts to keep LUN assignments identical between sites. For example, if a VMFS volume is mapped as LUN 10 at the primary site, it attempts to map the destination Live Volume as LUN 10. In some instances, the secondary site may already have existing LUNs that conflict, and cause the Live Volume to not operate as expected.

#### 15.4.1 Asynchronous Live Volume

When a Live Volume replication is created, the volume becomes read-writable from both the main system and secondary system. This allows vMotion of virtual machines over distance. However, it is important that the VMware 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 (dependent on vSphere support agreement level), and the virtual machine IP networks Layer 2 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.

Due to the asynchronous nature of the replication and how it proxies data, it is recommended that asynchronous (async) Live Volume only be used for disaster avoidance or planned migration. Live Volume is most practically used for planned maintenance operations such as migrating workloads between racks or data centers. Because of the nature of the proxied I/O, any disruption to the link or primary Live Volume causes the secondary Live Volume datastore to become unavailable as well. If for any reason the primary Live Volume from the last known good snapshot. The DSM replication disaster recovery wizard is designed to assist in this type of recovery.

#### 15.4.2 Synchronous Live Volume

SCOS 6.5 introduced synchronous capabilities to Live Volume using the newly revamped sync replication engine added in a previous release. This offers synchronous (sync) Live Volume as a new disaster recovery option to save time and effort recovering virtual machines. Since the disk signatures of the datastores remain the same between sites, during a recovery, volumes do not need to be resignatured, nor do virtual machines need to be removed and re-added to inventory. When using Sync Live Volume without Auto Failover enabled, the disaster must still be declared through DSM/EM to bring the secondary Live Volumes online. This saves most of the laborious recovery steps previously required, and virtual machines can simply be powered on after the volume comes online and the host has been rescanned.

One of the key considerations to using sync Live Volume is that round-trip link latency is now an important factor for application performance. In relation to this, the high consistency mode is the default option to ensure data integrity, meaning that low latencies are especially important. Sync Live Volume is most practically used within a data center or a campus environment where transactions need to be guaranteed to both storage systems, and round trip link latency can be kept low. Although round trip latencies between sites have no hard-set limitation, it is generally limited by the tolerances for each application or VMware support tolerances as noted previously.

For more information about Live Volume, consult the paper, <u>Storage Center Synchronous Replication and</u> <u>Live Volume Solutions Guide</u>.

#### 15.4.3 Live Volume auto failover

When using SCOS 6.7 or later, Live Volume auto failover may be used for high availability and stretched cluster configurations such as the vSphere Metro Storage Cluster (vMSC) solution certified by VMware. When using DSM as a tiebreaker at an independent third site, this allows multiple SC Series arrays to be configured to survive an entire array failure for increased storage availability scenarios.

For more information about the Live Volume vMSC solution, consult the VMware Knowledge Base article, <u>Implementing vSphere Metro Storage Cluster using Dell Storage Live Volume (2144158)</u>, and the Dell paper, <u>Storage Center Synchronous Replication and Live Volume Solutions Guide</u>.

## 15.5 Replication tips and tricks

Since replicated volumes often 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 100 virtual machines in the ESXi cluster, and only eight of them need to be replicated to the DR site, a special replicated volume should be created to place the eight virtual machines.

As mentioned previously, if separation proves to reduce snapshot sizes, operating system page files should be kept on a separate volume that is not replicated. That keeps replication and snapshot sizes smaller because the data in the page file changes frequently and it is generally not needed for a system restore.

To set replication priorities, an administrator can take advantage of the SC Series QoS to prioritize replication bandwidth of certain volumes. For example, if there is a 100 Mb connection between sites, two QoS definitions can be created such that the critical volume would get 80 Mb of the bandwidth, and the lower priority volume would get 20 Mb of the bandwidth.

# 15.6 Virtual machine recovery at a DR site

When recovering virtual machines at the disaster recovery site, the same general steps as outlined in the section 14.2 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 the DSM/EM software. These features allow an administrator to pre-define the recovery settings with specifics such as the appropriate hosts, mappings, LUN numbers, and host HBAs. After the recovery has been predefined, a recovery at the secondary site is mostly automated.

**Caution:** It is extremely important that the destination volume, usually denoted by *Repl of*, is never 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 to 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. For more information, see the Copilot services technical alert (CSTA), "Mapping Replicated Volumes at a DR Site", which is available on the Knowledge Center at the <u>SC Series Customer Portal</u> (login required).

# 16 VMware storage features

The vSphere platform has several features that correspond with SC Series features. This section details considerations that should be made with regard to features such as Storage I/O Controls (SIOC), storage distributed resource scheduler (SDRS), and VAAI.

# 16.1 Storage I/O Controls (SIOC)

SIOC is a feature that was introduced in ESX/ESXi 4.1 to help VMware administrators regulate storage performance and provide fairness across hosts sharing a volume. Because SC Series uses a shared pool of disks which can have differing performance characteristics, practice caution when using this feature. Also, due to how Data Progression migrates portions of volumes into different storage tiers and RAID levels at the block level, this could ultimately affect the latency of the volume, and trigger the resource scheduler at inappropriate times. It may not make sense to use SIOC unless pinning particular volumes into specific tiers of disk.

In one example that illustrates this scenario (shown in Figure 36), a datastore contains multiple virtual disks, and each virtual disk has different portions of blocks in tier 1 and tier 3. If VM1 begins to read a large amount 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 disks.

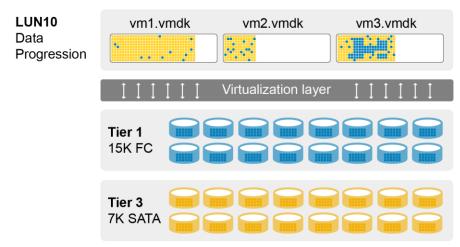


Figure 36 Multiple virtual disks with blocks residing in multiple tiers

As shown in Figure 37, the default setting for the congestion threshold is 30 milliseconds of latency or 90 percent of peak throughput.

datastore2 - Configure Storage I/O Control								
Storage I/O Control is used to control the I/O usage of a virtual machine and to gradually enforce the predefined I/O share levels.								
🗹 Enable Storage I/O Co	ntrol							
Congestion Threshold:	<ul> <li>Percentage of peak throughput 90 - %</li> </ul>							
	Manual 30 - ms							
	Reset to defaults							
Exclude I/O statistics fr	rom SDRS							
	ОК Са	incel						

Figure 37 Setting the congestion threshold

As a best practice, leave this at the default value unless under the guidance of VMware, Dell Copilot Support, or Dell Support.

For more information on SIOC, refer to the *vSphere Resource Management Guide* located at VMware <u>vSphere 6.x documentation</u> or <u>vSphere 5.x documentation</u>, or the article, <u>Unmanaged I/O workload detected</u> <u>on shared datastore running Storage I/O Control (SIOC) for congestion management</u> in the VMware Knowledge Base.

#### 16.1.1 SIOC and volume QoS

Similar to VMware SIOC, SCOS 7.0 introduces the volume QoS feature, which allows array administrators to limit bandwidth, IOPS, and prioritize storage workloads at the array level. While SIOC allows vSphere administrators control over storage within the vSphere cluster, volume QoS can balance all array workloads external to vSphere to prevent a *noisy neighbor effect*. Volume QoS allows all workloads within the array to be balanced for fairness.

Before implementing volume QoS, refer to the guide, <u>Volume Quality of Service Best Practices with Dell SC</u> <u>Series Storage</u>.

# 16.2 Storage distributed resource scheduler (SDRS)

SDRS is a feature introduced in ESXi 5.0 that automatically load balances virtual machines within a datastore cluster based on capacity and/or performance. When creating SDRS datastore clusters with SC Series storage, it is important to remember a few guidelines:

- Group datastores with similar disk characteristics, such as replicated, non-replicated, storage profile, application, or performance.
- Use SDRS for initial placement based on capacity. When placing virtual machines on the datastore cluster, it places them based on which datastore has the most space available.
- Set the automation level to manual mode. This means that SDRS will make recommendations about moving virtual machines, but not move them automatically. As a best practice, run SDRS 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, the entire virtual machine could be 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 snapshot (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 snapshots are being taken, the appropriate volume that contains the good snapshot of the virtual machine may be difficult to locate.
  - If the SCOS version in use does not support VAAI, the move process could be slow (without full copy) or could leave storage allocated on its originating datastore (without dead space reclamation). See the section 16.3 on VAAI for more information.
- Storage DRS could produce incorrect recommendations for virtual machine placement when I/O
  metric inclusion is enabled on an SC Series system using Data Progression. This is because when a
  datastore is inactive, the SIOC injector performs 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 recommendations.

#### 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 data store cluster

Enable I/O metric for SDRS recommendations

Figure 38 Disabling I/O metric inclusion

To reiterate, use SDRS only for capacity recommendations, set it to manual automation mode, and disable the I/O metric inclusion. 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.

## 16.3 vStorage APIs for array integration (VAAI)

ESXi 5.x introduced five primitives that the ESXi host can use to offload specific storage functionality to the array. These are all enabled by default in vSphere 6.x because they are based on new T10 standardized SCSI-3 commands.

**Caution:** Using VAAI requires multiple software version prerequisites from both VMware and Dell. Because certain primitives are only available with specific versions of SCOS, consult the <u>VMware Compatibility Guide</u> for the latest VAAI-certified versions.

#### 16.3.1 Block zeroing (SCSI WRITE SAME)

Traditionally, when an ESXi host would create an eagerzeroedthick virtual disk, it would transmit all of the zeros over the SAN to the array, which consumed bandwidth, processor, and disk resources. 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 sends a few WRITE SAME commands asking the array to write 40 GBs worth of zeros, and have the array respond when finished. In the case of the SC Series, since the array does not write zeros due to its thin write feature, creating an eagerzeroedthick virtual disk only takes seconds instead of minutes.

### 16.3.2 Full copy (SCSI EXTENDED COPY)

This primitive is used to offload to the array the copying or movement of blocks. It does this with the SCSI EXTENDED COPY command that allows the ESXi host to instruct the SC Series on which blocks it needs copied or moved, leaving the heavy lifting to the array. This has a 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 SC Series which blocks need to be copied/moved and the array takes care of the operations.

#### 16.3.3 Hardware accelerated locking

This primitive is intended to add scalability to the VMFS-5 file system by removing the need for SCSI-2 reservations. With VMFS, when an ESXi host needs exclusive access to a volume to update metadata, traditionally it will set a SCSI-2 non-persistent reservation on the volume to guarantee it had exclusive rights. Typically, during this operation, VMware documents small performance degradation to other virtual machines accessing this volume at the same time from different hosts. With the new hardware accelerated locking primitive, it uses a new SCSI-3 method called atomic test and 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 and increase the future scalability of VMFS.

#### 16.3.4 Dead space reclamation (SCSI UNMAP)

This primitive allows enhanced integration with arrays offering thin provisioning. In non-VAAI-capable versions of SCOS, when a VMDK was deleted, the array had no idea that the blocks were no longer in use, meaning they could not be freed back into the pagepool to be re-used. This caused the *high water mark* effect in which blocks of data that the operating system no longer needed could still be consumed by the array. 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 snapshot, they will not return to the pagepool until that snapshot has expired. Second, when using VMFS-5 and earlier, dead space reclamation does not work within the VMDK, and only works at the VMFS level. This means that if a large file is deleted from within a VMDK, the space would not be returned to the pagepool unless the VMDK itself was deleted.

**Note:** In most patch levels of ESXi, the dead space reclamation primitive must be invoked manually. See the article, <u>Using esxcli in vSphere 5.5 and 6.0 to reclaim VMFS deleted blocks on thin-provisioned LUNs</u> in the VMware Knowledge Base.

With vSphere 6.5, the default space reclamation behavior is different. When using VMFS-6 formatted datastores, space reclamation is enabled by default on all datastores, and the reclamation process is automatically invoked by the ESXi hosts. It is important for administrators to know that the automatic space reclamation process operates asynchronously at low priority, and is therefore not immediate. In addition, certain data pages will not be freed back into the Storage Center page pool, until after the daily Data Progression cycle has completed.

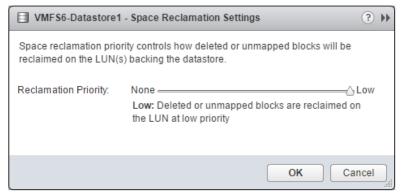


Figure 39 VMFS-6 Space Reclamation Settings

The space reclamation priority can be viewed or changed per datastore with the following commands:

```
esxcli storage vmfs reclaim config get -l DatastoreName
Reclaim Granularity: 1048576 Bytes
Reclaim Priority: low
```

esxcli storage vmfs reclaim config set -l DatastoreName -p high

#### 16.3.5 Thin provisioning stun

This primitive allows the SC Series 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 that are requesting additional pages until additional storage is added to remedy the situation.

For more information about VAAI primitives, see sections, "Array Thin Provisioning and VMFS Datastores" and "Storage Hardware Acceleration" in the *vSphere Storage Guide* at VMware <u>vSphere 6.x documentation</u> or <u>vSphere 5.x documentation</u>.

# 16.4 vStorage APIs for Storage Awareness (VASA)

The vStorage APIs for Storage Awareness are a set of protocols, routines, and tools encapsulated into an API that enable vCenter to detect the built-in capabilities of storage arrays. Properties of a volume such as its RAID level, performance capabilities, whether or not it is replicated, and if Data Progression is enabled can all be obtained by vCenter. In turn, all of these capabilities are used by vCenter to enable profile-driven storage, allowing the administrator to determine if a virtual machine meets the organization's compliance policies.

VASA 2.0 is also required for VMware Virtual Volumes (VVols). VVols uses the VASA provider to execute the out-of-band tasks required for the volume-to-container relationship management.

Vcsite11.techsol.local Actio	ns 🔻								
Getting Started Summary Mor	Getting Started Summary Monitor Manage Related Objects								
Settings         Scheduled Tasks         Storage Providers         Alarm Definitions         Tags         Permissions         Sessions									
Storage Providers									
🕂 🔯 📃 🗙		Group by: Storage provider	•						
Storage Provider/Storage System	Status	URL	VASA API Version						
👻 Dell Storage VASA Provider	Online	https://172.16.21.175:3034/vasa-provider/vasa2/vasa-version.xml	2.0						
SC 22 (1/1 online)									

Figure 40 Adding a VASA 2.0 storage provider to vCenter

The SC Series VASA 2.0 provider is a built-in component of the DSM Data Collector, and supersedes the VASA 1.0 provider built into the CITV appliance.

## 16.5 Virtual Volumes (VVols)

VMware Virtual Volumes is a new approach available with vSphere 6.x to managing storage that delivers more granular storage capabilities to the VMDK level. To accomplish this, VMware introduces new concepts such as storage containers, protocol endpoints, and storage providers as part of a new framework for delivering software defined storage and storage policy-based management capabilities to traditional arrays.

For more detailed information on VMware Virtual Volumes, read the VMware document, <u>What's New:</u> <u>vSphere Virtual Volumes</u>.

With these new foundational changes in mind, there are additional best practices that need to be followed.

- VMware vCenter Server and the DSM Data Collector servers need to be made highly available. Any disruption in services may affect availability of virtual machines stored within virtual volumes.
- When using vSphere HA, restart priority should be modified so that vCenter and DSM are brought up first.
- vCenter and DSM virtual machines must be stored on traditional datastores residing on SAN storage.
   Do not store or migrate these virtual machines onto a VVol datastore.
- Take regular array snapshots to protect vCenter and the VASA provider within DSM.
- Ensure proper NTP time synchronization between vCenter, ESXi, DSM, and the arrays.

**Caution:** When utilizing Virtual Volumes, the VASA 2.0 provider contained within DSM becomes a critical piece of infrastructure that needs to be protected from outages and loss of data like any other business-critical enterprise application. Without the VASA provider active, vCenter will not be able to bind or unbind virtual volumes to or from the protocol endpoints. In addition, if the VASA provider is lost completely, recovery of virtual disks from within the storage container will become be an arduous task.

For more information about the requirements for configuring SC Series storage and DSM for Virtual Volumes, read the following documents before beginning the implementation:

- Dell Storage Manager 2016 R1 Administrators Guide, available on the Knowledge Center at the <u>SC</u> <u>Series customer portal</u> (login required) or the <u>Dell Support</u> site.
- Dell SC Series Virtual Volumes Best Practices

# 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 disks to handle the incoming requests). On the other hand, if set improperly, the ESXi hosts could overdrive the controller front-end ports or the back-end disks and potentially make the performance worse.

The general guideline is to set the queue depth high enough to achieve an acceptable number of IOPS from the back-end disks, but to not set it too high and allow an ESXi host to flood the front or back end of the array. The following sections provide a few basic pointers.

## A.1 Fibre Channel

With 2 Gb SC Series front-end (FE) ports:

- Each 2 Gb FE port has a maximum queue depth of 256 and care must be taken to not overdrive it.
- It is best to leave the ESXi queue depths set to default and only increase them if absolutely necessary.
- The recommended settings for controllers with 2 Gb FE ports are:
  - HBA BIOS = 255 (if available)
  - Driver module = 32 (default) (HBA Queue depth is regulated by the driver module)
  - DSNRO= 32 (default)
  - Guest vSCSI controller = 32 (default)

With 4/8/16 Gb SC Series front-end ports:

- Each 4/8/16 Gb front-end port can accept many more (~1900+) outstanding I/Os.
- Since each FE port can accept more outstanding I/Os, the ESXi queue depths can be set more flexibly to accommodate guest I/O. Keep in mind, the queue depth may need to be decreased if the front-end ports become saturated, the back-end disks become maxed out, or the latencies become too high.
- The recommended settings for controllers with 4/8/16 Gb FE ports are:
  - HBA BIOS = 255 (if available)
  - Driver module = 255
  - DSNRO = 32 (default) (increase or decrease as necessary)
  - Guest vSCSI controller = 32 (default) (increase or decrease as necessary)

# A.2 iSCSI

With 1 Gb SC Series front-end ports, leave the queue depth set to default and only increase if necessary.

With 10 Gb SC Series front-end ports, use the following settings:

- HBA BIOS (if using hardware iSCSI) = 255
- Driver Module (iscsi\_vmk) = 255
- DSNRO = 32 (default) (increase or decrease as necessary)
- Guest vSCSI controller = 32 (default) (increase or decrease as necessary)

# A.3 Using esxtop to monitor queue depth

The best way to determine the appropriate queue depth is by using the esxtop utility. This utility can be executed from the ESXi Shell through SSH (command: esxtop), or vCLI 5.x or the vMA virtual appliance (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. Use the following steps to navigate to the screen:

- 1. From the command line, enter one of these two options:
  - # esxtop
  - resxtop.sh --server esxserver.domain.local
- 2. Press **u** to open the Disk Device screen.
- 3. Type **L 36** and press Enter to expand the devices field. This will expand the disk devices, allowing volumes to be identified by naa identifier.
- 4. Press **f** and choose the fields to monitor:
  - 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. Type **s 2** and press Enter to set the refresh time to 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, as shown in Figure 41.

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

Figure 41 esxtop with a queue depth of 32

If the LOAD is consistently greater than 1.00 on one or more volumes, the latencies are still acceptable, the back-end disks have available IOPS, and increasing the queue depth may make sense. However, if the

LOAD is consistently less than 1.00 on a majority of the volumes, and the performance and latencies are acceptable, then there is usually no need to adjust the queue depth.

In Figure 41, the device queue depth is set to 32. Note that three of the four volumes consistently have a LOAD above 1.00. If the back-end disks are not maxed out, it may make sense to increase the queue depth and the DSNRO setting.

As shown in Figure 42, increasing the queue depth from the previous example has increased the total IOPs from 6700 to 7350, but the average device latency (DAVG/cmd) increased from 18 ms to 68 ms. That means the latency more than tripled for a mere 9-percent performance gain. In this case, it may not make sense to increase the queue depth because latencies became too high. In a system that is sized with the appropriate number of drives to handle the application load, the average response time (latency) should remain below 20 ms.

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

Figure 42 Queue depth increased to 255

For more information about the disk statistics in esxtop, consult the esxtop man page or the *vSphere Monitoring and Performance Guide* on VMware <u>vSphere 6.x documentation</u> or <u>vSphere 5.x documentation</u>.

**Note:** Although the per-LUN queue depth maximum is 256, the per-adapter maximum within ESXi is 4096. By increasing the per LUN queue depth from 64 to 128, it can take fewer LUNs to saturate a port queue. For example, 4096/64=64 LUNs, but 4096/128=32 LUNs.

# B Deploying vSphere client plug-ins

With SC Series storage, multiple vSphere client plug-ins are available to aid in administration and management of the arrays. Depending on the version, these plug-ins allow administrators to provision datastores, take snapshots, create replications, and even report usage and performance statistics directly from within the vSphere client. While all new plug-in features and functionality are being integrated into the web client plug-in, the traditional client plug-in for environments still using vSphere 5.x contains the necessary functionality to perform most daily storage administration tasks.

## B.1 SC Series vSphere client plug-in 1.7

This client plug-in is designed for use with the traditional VMware vSphere client for Windows. The administrators guide and executable can be downloaded from <u>Knowledge Center</u> (login required) by searching for "Storage Center vSphere Client Plugin 1.7 Download".

Vesite Litechsol.local - vSphere C File Edit View Inventory Administra Home Market Inventory Inventory Inventory				ଟ 🗸 Search Invent	ory Q
<b>a</b> a <b>a</b>					
Area is intervaled     Works is intervaled     Works intervaled	essi6.kechsol.local VMvare ESX Summary Virtual Machines Pro View General Usage Satistics Concectivity Info Volume Regions Volume Regions Commands Refresh Add Datastore Refresh Provision VMs	nformance Configuration Tasks & Concellent Volume Lando smd51 smd52 demovms Storage Usage Total volume space consumed: Total avait space consumed:	500.00 ( 750.00 ( 750.00 ( 750.00 ( 2.00 1) 1.61 TB 2.01 TB	B         7         SN 697         [sc12.techsol.loca           BB         3         SN 697         [sc12.techsol.loca           BB         4         SN 697         [sc12.techsol.loca           BB         4         SN 697         [sc12.techsol.loca	a) 1
🚰 Tasks 💇 Alarms					

Figure 43 Screenshot of datastore usage statistics

# B.2 Dell Storage vSphere Web Client Plugin 4.x

The web client plug-in is contained within the Dell Storage Integration for VMware (Compellent Integration Tools for VMware/CITV 4.x) virtual appliance. Once the appliance is deployed and given an IP address, the web plug-in can be registered with vCenter, and administrators can provide their DSM login credentials to manage all of their arrays.

The CITV 4.x virtual appliance can be downloaded from the Dell Support site at: <a href="http://www.dell.com/support/home/us/en/19/Drivers/Drivers/DriversDetails?driverId=RH38M">http://www.dell.com/support/home/us/en/19/Drivers/Drivers/DriversDetails?driverId=RH38M</a>

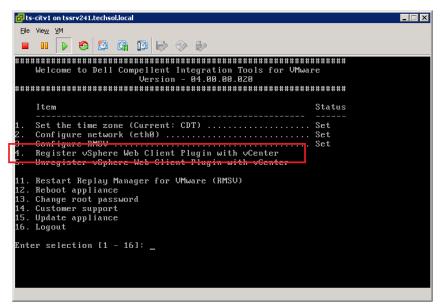


Figure 44 CITV 4.x virtual appliance configuration screen

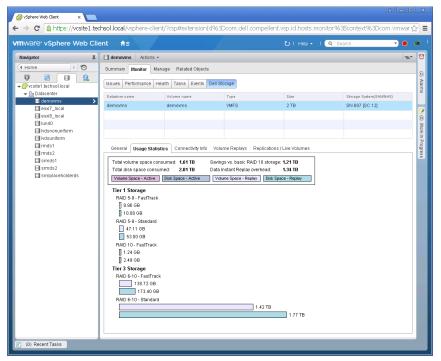


Figure 45 SC Series web client plug-in usage statistics screen

# Configuring Dell Storage Manager VMware integrations

With DSM/EM, the Data Collector can be configured to gather storage statistics and perform basic storage administration functions with vCenter. This functionality is also used to configure and enable VVols.

To add vCenter credentials into the Data Collector, enter the servers viewer screen, right-click **Servers**, and select **Register Server**.

Servers	≶ C ♦ ♦	Sum	mary 🥐
<mark>&gt;=</mark> Se <sup></sup>	Create Folder	,	Add Windows Server Agent
	C Update Information on Servers Rescan for Volumes on Servers Space Recovery History		Server VCenter Server

Figure 46 Adding a vCenter Server

С

🗬 Register Server	X		
Host or IP Address	vcenter.mycompany.com		
User Name	Domain\vCenterAdministratorAccount		
User Password	*****		
HTTPS	Enabled		
Auto Manage Storage Centers	Enabled		
Auto Manage Virtual Machines On Storage Centers	🔲 Enabled		
Folder	····· 🖻 Servers		
? Help	Cancel		

Figure 47 Registering vCenter server credentials

After entering vCenter credentials, administrators can see aggregate storage statistics and provision VMFS datastores and RDMs. For example, when creating a new volume, selecting an ESXi host automatically gives the option to format it with VMFS. Similarly, when creating a new volume to be assigned to a virtual machine, DSM can automatically add the volume as a pRDM.

# D Host and cluster settings

This section summarizes the host settings for each of the storage protocols. At the time of this writing, these settings were valid for vSphere 6.x running on SCOS 6.6 and later. For each setting, see the referenced section number to determine if the setting is applicable for your environment.

### D.1 Recommended settings

Apply these settings to each of the ESXi hosts:

• Fibre Channel: Driver Module settings for queue depth and controller failover (section 4.2.1)

```
esxcli system module parameters set -m qlnativefc -p "ql2xmaxqdepth=255 ql2xloginretrycount=60 qlport down retry=60"
```

• Fibre Channel and iSCSI: Setting round robin as the default path selection policy (section 6.9.1.1)

esxcli storage nmp satp set -P VMW\_PSP\_RR -s VMW\_SATP\_ALUA

• FCoE: Setting the ALUA module default for FCoE volumes

esxcli storage nmp satp rule add -R fcoe -s VMW\_SATP\_ALUA

iSCSI: Setting ARP redirect for hardware iSCSI adapters (section 3)

```
esxcli isc<br/>si physical<br/>networkportal param set --option ArpRedirect -v=1 -A\ensuremath{\mathsf{vmhbaXX}}
```

iSCSI: Software iSCSI Queue Depth (section 4.2.2)

esxcli system module parameters set -m iscsi vmk -p iscsivmk LunQDepth=255

iSCSI: Software iSCSI login timeout (section 4.2.2)

esxcli iscsi adapter param set -A vmhbaXX -k LoginTimeout -v 60

• iSCSI: Disable Delayed ACK (section 3)

vmkiscsi-tool -W -a delayed ack=0 -j vmhbaXX

• All protocols: HA Cluster Settings (section 4.3)

esxcli system settings kernel set -s terminateVMonPDL -v TRUE

• All protocols: HA Cluster Settings (section 4.3)

esxcli system settings advanced set -o "/Disk/AutoremoveOnPDL" -i 1

• All protocols: Advanced options HA Cluster setting (section 4.3)

das.maskCleanShutdownEnabled = True

# D.2 Optional settings

These settings are optional, but are listed because they solve certain corner-case performance issues. Settings in this section should be tested to ascertain if the settings improve function.

• All protocols: DSNRO per datastore (section 4.4)

esxcli storage core device set -d <naa.dev> -O <value of 1-256>

• All protocols: Conditions for round robin path changes, individual datastore (section 6.9.1.2)

esxcli storage nmp psp roundrobin deviceconfig set --device naa.XXX -type=iops --iops=3

• All protocols: Conditions for round robin path changes, all datastores on host (section 6.9.1.2)

esxcli storage nmp satp rule add -s "VMW\_SATP\_ALUA" -V "COMPELNT" -P "VMW PSP RR" -O "iops=3"

# E Additional resources

## E.1 Technical support and resources

<u>Dell.com/support</u> is focused on meeting customer needs with proven services and support.

<u>Dell TechCenter</u> is an online technical community where IT professionals have access to numerous resources for Dell software, hardware and services.

<u>Storage Solutions Technical Documents</u> on Dell TechCenter provide expertise that helps to ensure customer success on Dell Storage platforms.

## E.2 VMware support

For VMware support, see the following resources:

- VMware vSphere 6.x Documentation
- VMware vSphere 5.x Documentation
- <u>VMware Knowledge Base</u>
- VMware.com
- Education and training
- Online documentation
- <u>VMware communities</u>