Using EMC Symmetrix Storage in VMware vSphere Environments

Version 8.0

- Layout, Configuration, Management, and Performance
- Replication, Cloning, and Provisioning
- Disaster Restart and Recovery

Cody Hosterman
Drew Tonnesen

EMC²
TECHBOOKS

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

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” EMC CORPORATION 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 EMC software described in this publication requires an applicable software license.

For the most up-to-date regulatory document for your product line, go to the Technical Documentation and Advisories section on EMC Powerlink.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

All other trademarks used herein are the property of their respective owners.

Part number H2529.11

Contents

Preface

Chapter 1 VMware vSphere and EMC Symmetrix

Chapter 2 Management of EMC Symmetrix Arrays

Chapter 3 EMC Virtual Provisioning and VMware vSphere

Chapter 4 Data Placement and Performance in vSphere

Chapter 5 Cloning of vSphere Virtual Machines

Chapter 6 Disaster Protection for VMware vSphere

Figures

Tables

Preface

This TechBook describes how the VMware vSphere platform works with EMC Symmetrix storage systems and software technologies.

As part of an effort to improve and enhance the performance and capabilities of its product lines, EMC periodically releases revisions of its hardware and software. Therefore, some functions described in this document may not be supported by all versions of the software or hardware currently in use. For the most up-to-date information on product features, refer to your product release notes.

Audience

This document is part of the EMC Symmetrix documentation set, and is intended for use by storage administrators, system administrators and VMware administrators.

Readers of this document are expected to be familiar with the following topics:

- EMC Symmetrix system operation
- EMC SRDF, EMC TimeFinder, and EMC Solutions Enabler
- VMware vSphere products

Organization

The document is divided into six chapters with one appendix:

Chapter 1, “VMware vSphere and EMC Symmetrix,” discusses the installation, setup and configuration of VMware vSphere environments with EMC Symmetrix arrays. This chapter also presents best practices when using EMC Symmetrix storage with VMware vSphere platforms.

Chapter 2, “Management of EMC Symmetrix Arrays,” discusses the various ways to manage a VMware environment using EMC technologies.

Chapter 3, “EMC Virtual Provisioning and VMware vSphere,” discusses EMC Virtual Provisioning in detail along with an introduction to vSphere vStorage APIs for Storage Awareness (VASA) and vSphere Thin Provisioning.

Chapter 4, “Data Placement and Performance in vSphere,” addresses EMC and VMware technologies that help with placing data and balancing performance in the VMware environment and on the Symmetrix.

Chapter 5, “Cloning of vSphere Virtual Machines,” presents how EMC TimeFinder and the vStorage APIs for Array Integration (VAAI) can be used in VMware vSphere environments to clone virtual machines. It also discusses how to create TimeFinder copies of SRDF R2 and RecoverPoint images of VMware file systems hosted on Symmetrix volumes and then mount those volumes to the disaster recovery site.

Chapter 6, “Disaster Protection for VMware vSphere,” discusses the use of EMC SRDF in a VMware vSphere environment to provide disaster restart protection for VMware vSphere data.

IMPORTANT

Examples provided in this guide cover methods for performing various VMware vSphere activities using Symmetrix systems and EMC software. These examples were developed for laboratory testing and may need tailoring to suit other operational environments. Any procedures outlined in this guide should be thoroughly tested before implementing in a production environment.

Authors

This TechBook was authored by a team from Partner Engineering based at Hopkinton, Massachusetts and Santa Clara, California.

Cody Hosterman is a Senior Systems Integration Engineer in the Technical Partner Management team in EMC Central Partner Engineering focusing on VMware integration. Cody has been with EMC since 2008. Cody has a Bachelor's degree in Information Sciences and Technology from Penn State University.

Drew Tonnesen is a Consulting Systems Engineer in Symmetrix Systems Engineering focusing on VMware and other virtualization technologies. Before starting in his current position, Drew worked as a Global Solutions Consultant focusing on Oracle Technology. He has worked at EMC since 2006 in various capacities. Drew has over 16 years of experience working in the IT industry. Drew holds a Master's degree from the University of Connecticut.

Related documents

The following documents are available at www.Powerlink.EMC.com:

- *EMC VSI for VMware vSphere: Storage Viewer Product Guide*
- *Using EMC SRDF Adapter for VMware vCenter Site Recovery Manager TechBook*
- *EMC Solutions Enabler Symmetrix TimeFinder Family CLI Product Guide*
- *EMC Symmetrix TimeFinder Product Guide*
- *Solutions Enabler Management Product Guide*
- *Solutions Enabler Controls Product Guide*
- *Symmetrix Remote Data Facility (SRDF) Product Guide*
- *Symmetrix Management Console online help*
- *EMC Support Matrix*
- *Tuning the VMware Native Multipathing Performance of VMware vSphere Hosts Connected to EMC Symmetrix Storage White Paper*
- *Implementing EMC Symmetrix Virtual Provisioning with VMware vSphere White Paper*
- *Implementing VMware's vStorage API for Storage Awareness with Symmetrix Storage Arrays White Paper*
- *Storage Tiering for VMware Environments Deployed on EMC Symmetrix VMAX with Enginuity 5875 White Paper*

Conventions used in this document

EMC uses the following conventions for special notices.

Note: A note presents information that is important, but not hazard-related.

⚠ CAUTION

A caution contains information essential to avoid data loss or damage to the system or equipment.

IMPORTANT

An important notice contains information essential to operation of the software or hardware.

Typographical conventions

EMC uses the following type style conventions in this document:

Normal

Used in running (nonprocedural) text for:

- Names of interface elements (such as names of windows, dialog boxes, buttons, fields, and menus)
- Names of resources, attributes, pools, Boolean expressions, buttons, DQL statements, keywords, clauses, environment variables, functions, utilities
- URLs, pathnames, filenames, directory names, computer names, filenames, links, groups, service keys, file systems, notifications

Bold

Used in running (nonprocedural) text for:

- Names of commands, daemons, options, programs, processes, services, applications, utilities, kernels, notifications, system calls, man pages

Used in procedures for:

- Names of interface elements (such as names of windows, dialog boxes, buttons, fields, and menus)
- What user specifically selects, clicks, presses, or types

Italic

Used in all text (including procedures) for:

- Full titles of publications referenced in text
- Emphasis (for example a new term)
- Variables

`Courier`

Used for:

- System output, such as an error message or script
- URLs, complete paths, filenames, prompts, and syntax when shown outside of running text

`Courier bold`

Used for:

- Specific user input (such as commands)

`Courier italic`

Used in procedures for:

- Variables on command line
- User input variables

< > Angle brackets enclose parameter or variable values supplied by the user

[] Square brackets enclose optional values

| Vertical bar indicates alternate selections - the bar means “or”

{} Braces indicate content that you must specify (that is, x or y or z)

... Ellipses indicate nonessential information omitted from the example

We'd like to hear from you!

Your feedback on our TechBooks is important to us! We want our books to be as helpful and relevant as possible, so please feel free to send us your comments, opinions and thoughts on this or any other TechBook at TechBooks@emc.com.

1

VMware vSphere and EMC Symmetrix

This chapter discusses the configuration and best practices when connecting a VMware virtualization platform to a EMC Symmetrix storage array.

Note: In this chapter, VMware ESX will refer to both ESX and ESXi unless otherwise noted.

Introduction

VMware® ESX®/ESXi™ virtualizes IT assets into a flexible, cost-effective pool of compute, storage, and networking resources. These resources can be then mapped to specific business needs by creating virtual machines.

EMC® Symmetrix® storage arrays are loosely coupled parallel processing machines that handle various workloads from disparate hardware and operating systems simultaneously. When VMware ESX/ESXi is used with EMC Symmetrix storage arrays, it is critical to ensure proper configuration of both the storage array and the ESX/ESXi host to ensure optimal performance and availability.

This chapter addresses the following topics:

- Configuration of EMC Symmetrix arrays when used with VMware virtualization platform.
- Discovering and using EMC Symmetrix devices in VMware.
- Configuring EMC Symmetrix storage arrays and VMware ESX/ESXi for optimal operation.

Detailed information on configuring and using VMware ESX/ESXi in EMC FC, NAS and iSCSI environments can also be found in the *Host Connectivity Guide for VMware ESX*, located at www.Powerlink®.EMC.com. This is the authoritative guide for connecting VMware ESX to EMC Symmetrix storage arrays and should be consulted for the most current information.

Setup of VMware ESX 4 and ESXi 5

VMware and EMC fully support booting the VMware ESX 4 and ESXi 5[1] hosts from EMC Symmetrix storage arrays when using either QLogic or Emulex HBAs or all 10 Gb/s CNAs from Emulex, QLogic, or Brocade.

Booting the VMware ESX from the SAN enables the physical servers to be treated as an appliance, allowing for easier upgrades and maintenance. Furthermore, booting VMware ESX from the SAN can simplify the processes for providing disaster restart protection of the virtual infrastructure. Specific considerations when booting VMware ESX hosts are beyond the scope of this document. The *E-Lab Interoperability Navigator*, available at www.Powerlink.EMC.com, and appropriate VMware documentation, should be consulted for further details.

Regardless of whether the VMware ESX is booted off an EMC Symmetrix storage array or internal disks, the considerations for installing VMware ESX do not change. Readers should consult the *ESX and vCenter Server Installation Guide* available on www.VMware.com for further details.

[1] In this TechBook, unless specified explicitly, minor releases of vSphere™ environments such as vSphere 4.1 are implied in the use of vSphere or ESX 4 and ESXi 5.

Using EMC Symmetrix with VMware ESX 4 and ESXi 5

Fibre channel port settings

VMware ESX 4 and ESXi 5 require SPC-2 compliant SCSI devices. When connecting VMware ESX 4 or ESXi 5 to an EMC Symmetrix storage array the SPC-2 bit on the appropriate Fibre Channel ports needs to be set. The SPC-2 is set by default on DMX™ beginning with Enginuity™ 5773 and on all VMAX™ arrays. At Enginuity level 5876 SPC-3 is implemented which supersedes the SPC-2 bit.[1]

Note: The following bit settings are either no longer required in VMware environments or are on by default: the common serial number, C; SCSI3 or SC3; OS2007; PP for FC-SW; ACLX; EAN; and WWN.

Note: Please consult the *EMC Support Matrix* for an up-to-date listing of port settings.

⚠ CAUTION

The bit settings described previously, and in the *EMC Support Matrix*, are critical for proper operation of VMware virtualization platforms and EMC software. The bit settings can be either set for the Symmetrix FA port or per HBA.

Symmetrix FA connectivity

Each ESXi server in the VMware vSphere™ environment should have at least two physical HBAs, and each HBA should be connected to at least two different front-end ports on different directors of the Symmetrix.

If the Symmetrix in use has only one engine then each HBA should be connected to the odd and even directors within it. Connectivity to the Symmetrix front-end ports should consist of first connecting unique

[1] SPC-3 is limited to VMAX arrays shipped with Enginuity 5876 and VMAX 10K shipped after October 2012. If a VMAX/VMAXe is upgraded to 5876, SPC-2 is still required, though it is on by default on all VMAX/VMAXe arrays.

hosts to port 0 of the front-end directors before connecting additional hosts to port 1 of the same director and processor. Figure 1 displays the connectivity for a single engine VMAX. Connectivity for all VMAX family arrays would be similar.

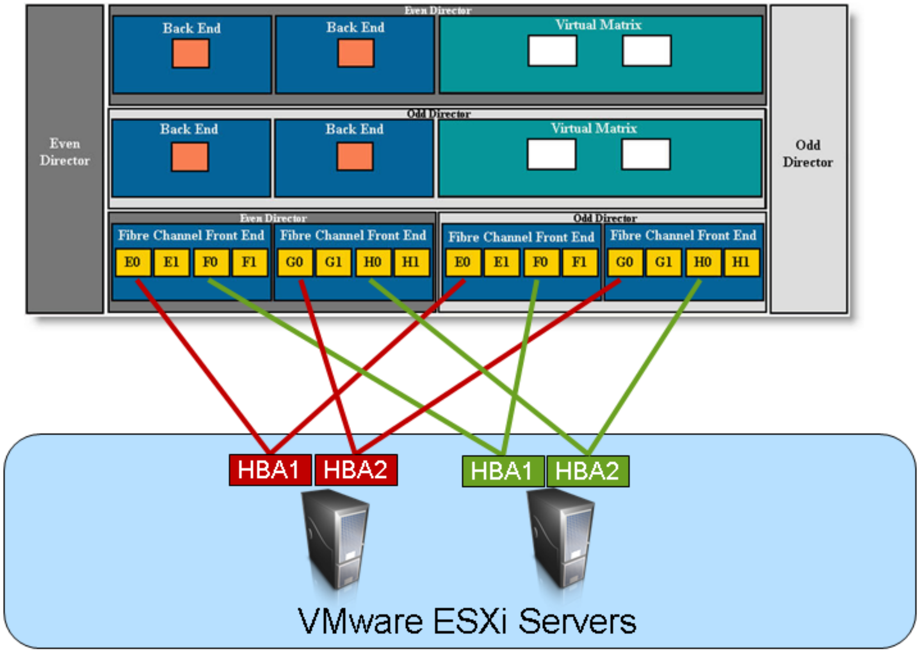


Figure 1 Connecting ESXi servers to a single engine Symmetrix VMAX

If multiple engines are available on the Symmetrix, the HBAs from the VMware ESXi servers in the VMware vSphere environment should be connected to different directors on different engines. In this situation, the first VMware ESXi server would be connected to four different processors on four different directors over two engines instead of four different processors on two directors in one engine Figure 2 demonstrates this connectivity.

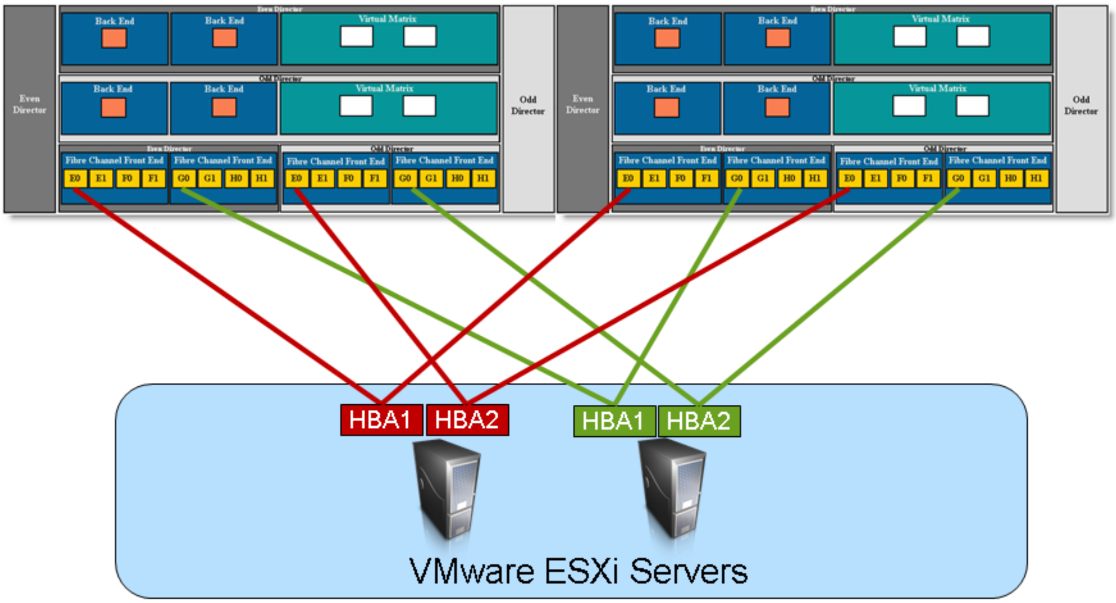


Figure 2 **Connecting ESXi servers to a multi-engine Symmetrix VMAX**

As more engines become available, the connectivity can be scaled as needed. These methodologies for connectivity ensure all front-end directors and processors are utilized, providing maximum potential performance and load balancing for VMware vSphere environments connected to Symmetrix storage arrays.

Driver configuration in VMware ESX 4 and ESXi 5

The drivers provided by VMware as part of the VMware ESX distribution should be utilized when connecting VMware ESX to EMC Symmetrix storage. However, EMC E-Lab™ does perform extensive testing to ensure the BIOS, BootBIOS and the VMware supplied drivers work together properly with EMC storage arrays.

Adding and removing EMC Symmetrix devices to VMware ESX hosts

The addition or removal of EMC Symmetrix devices to and from VMware ESX is a two-step process:

1. Appropriate changes need to be made to the EMC Symmetrix storage array configuration. In addition to LUN masking, this may include creation and assignment of EMC Symmetrix volumes and metavolumes to the Fibre Channel ports utilized by the VMware ESX. The configuration changes can be performed using EMC Solutions Enabler software, EMC Ionix™ ControlCenter®, Symmetrix Management Console (SMC) or Unisphere™ for VMAX from an independent storage management host.

2. The second step of the process forces the VMware kernel to rescan the Fibre Channel bus to detect changes in the environment. This can be achieved by the one of the following three processes:

   - Utilize the graphical user interface (vSphere Client)
   - Utilize the command line utilities
   - Wait for no more than 5 minutes

The process to discover changes to the storage environment using these tools are discussed in the next two subsections.

Using the vSphere Client

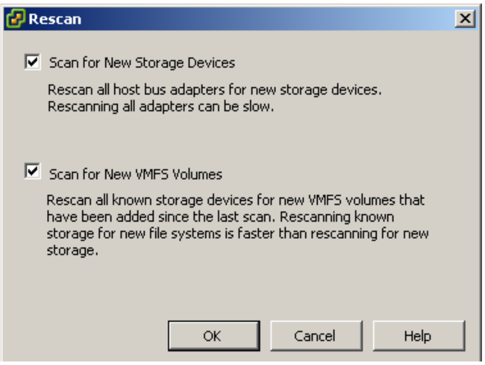
Changes to the storage environment can be detected using the vSphere Client by following the process listed below:

1. In the vCenter inventory, right-click the VMware ESX host (or, preferably, the vCenter cluster or datacenter object to rescan multiple ESX hosts at once) on which you need to detect the changes.

2. In the menu, select **Rescan for Datastores...**

3. Selecting **Rescan for Datastores...** using vSphere Client results in a new window shown in Figure 3 on page 30, providing users with two options: Scan for New Storage Devices and Scan for New VMFS Volumes. The two options allow users to customize the rescan to either detect changes to the storage area network, or to the changes in the VMFS volumes. The process to scan the storage area network is much slower than the process to scan for changes to VMFS volumes. The storage area network should be scanned only if there are known changes to the environment.

4. Click **OK** to initiate the rescan process on the VMware ESX.



Rescan

☑ Scan for New Storage Devices

Rescan all host bus adapters for new storage devices. Rescanning all adapters can be slow.

☑ Scan for New VMFS Volumes

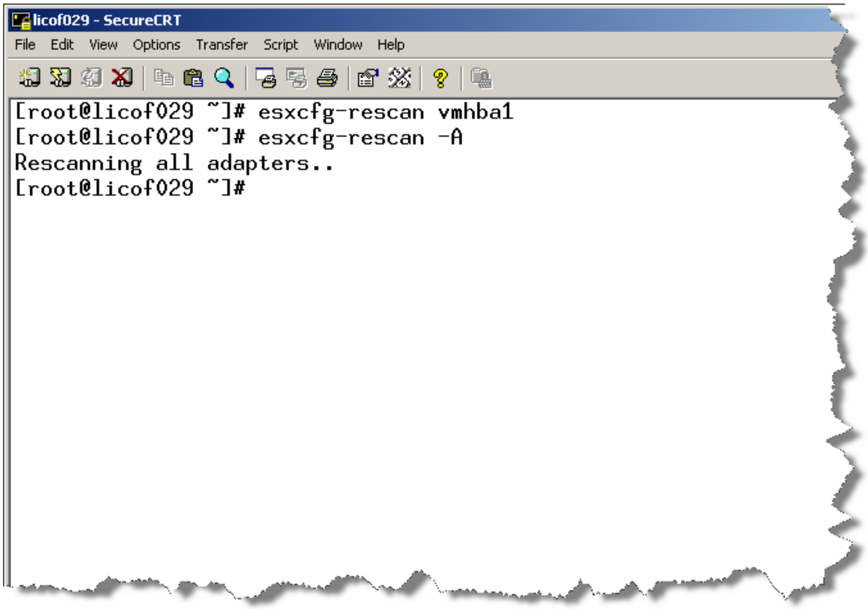
Rescan all known storage devices for new VMFS volumes that have been added since the last scan. Rescanning known storage for new file systems is faster than rescanning for new storage.

Figure 3 Rescanning options in the vSphere Client

Using VMware ESX/ESXi command line utilities

VMware ESX (not ESXi) version 4 uses a service console utility, **esxcfg-rescan**, to detect changes to the storage environment. For ESXi in 4 and 5, the vCLI command vicfg-rescan provides the same functionality. The vCLI utilities take the VMkernel SCSI adapter name (**vmhbax**) as an argument. These utilities should be executed on all relevant VMkernel SCSI adapters if EMC Symmetrix devices are presented to the VMware ESX on multiple paths. One may also pass the 'A' flag to scan all adapters. There is no output from the command unless scanning all adapters. Figure 4 on page 31 displays an example using **esxcfg-rescan**.

Note: The use of the remote CLI or the vSphere client is highly recommended, but in the case of network connectivity issues, ESXi 5 still offers the option of using the CLI on the host (Tech Support Mode must be enabled). The command to scan all adapters is: `esxcli storage core adapter rescan --all`.



```
[root@licof029 ~]# esxcfg-rescan vmhba1
[root@licof029 ~]# esxcfg-rescan -A
Rescanning all adapters..
[root@licof029 ~]#
```

Figure 4 **Using esxcfg-rescan to rescan for changes to the SAN environment**

Creating VMFS volumes on VMware ESX 4 and ESXi 5[1]

VMware Virtual Machine File System (VMFS) volumes created utilizing the vSphere Client are automatically aligned on 64 KB boundaries. Therefore, EMC strongly recommends utilizing the vSphere Client to create and format VMFS volumes.

Note: A detailed description of track and sector alignment in x86 environments is presented in the section "Partition alignment" on page 44.

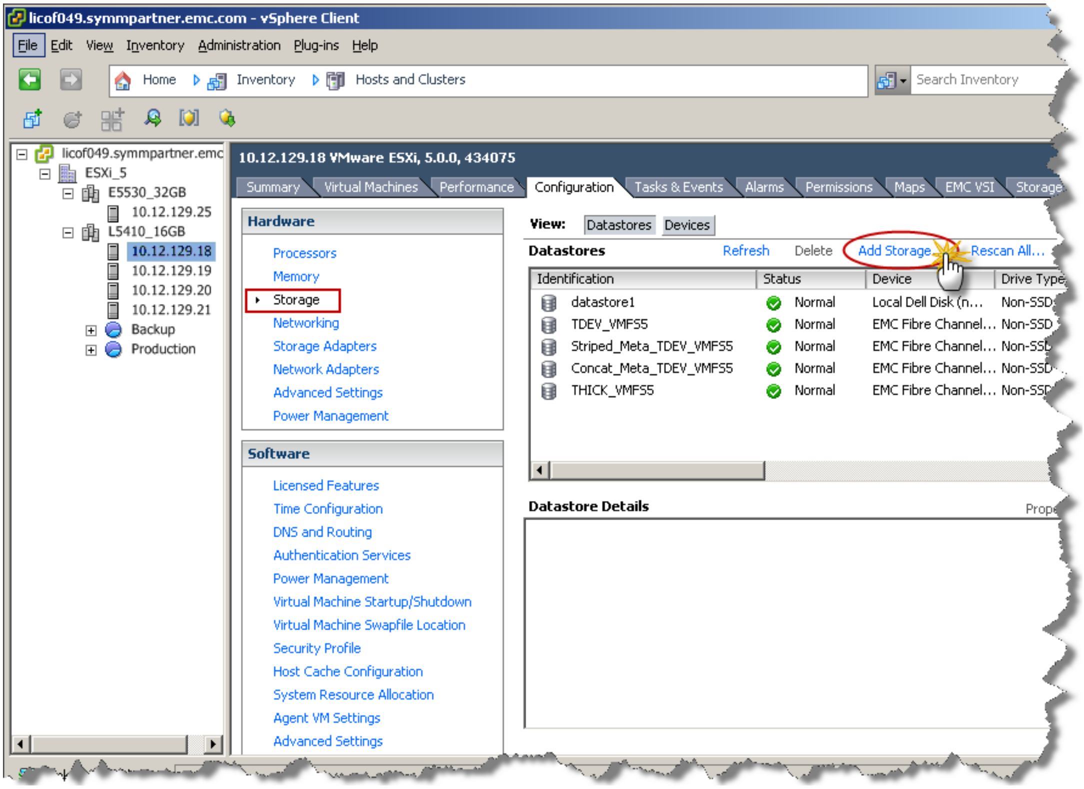
[1] Due to an interoperability issue with a VAAI primitive (see "VMware vSphere vStorage API for Array Integration," in Chapter 5) and Enginuity, customers running an Enginuity level prior to 5875.267.201 with ESXi 5.1 will be unable to create VMFS-5 datastores. This includes attempting to install ESXi in a boot-to-SAN configuration. This issue does not exist in Enginuity 5876. For more information, customers who are running this environment should review EMC Technical Advisory emc289707.

Creating a VMFS datastore using vSphere Client

The vSphere Client offers a single process to create an aligned VMFS. The following section will walk through creating the VMFS datastore in the vSphere Client, noting the differences between ESX 4 and ESXi 5 where necessary.

Note: A datastore in a VMware environment can be either a NFS file system or a VMFS. Therefore the term *datastore* is utilized in the rest of the document. Furthermore, a group of VMware ESX hosts sharing a set of datastores is referred to as a *cluster*. This is distinct from a *datastore cluster*, the details of which can be found in Chapter 4.

The user starts by selecting an ESX host on the left-hand side of the vSphere Client. If shared storage is presented across a cluster of hosts, any host can be selected. A series of tabs are available on the right-hand side of the Client. The **Configuration** tab is chosen and the storage object on the right-hand **Hardware** pane in Figure 5, boxed in red, provides the path to create a new datastore. Selecting this **Storage** object displays all available datastores on the VMware ESX. In addition to the current state information, the pane also provides the options to manage the datastore information and create a new datastore. The wizard to create a new datastore can be launched by selecting the **Add storage** button circled in red in Figure 5 on the top right-hand corner of the storage pane of the Configuration tab.

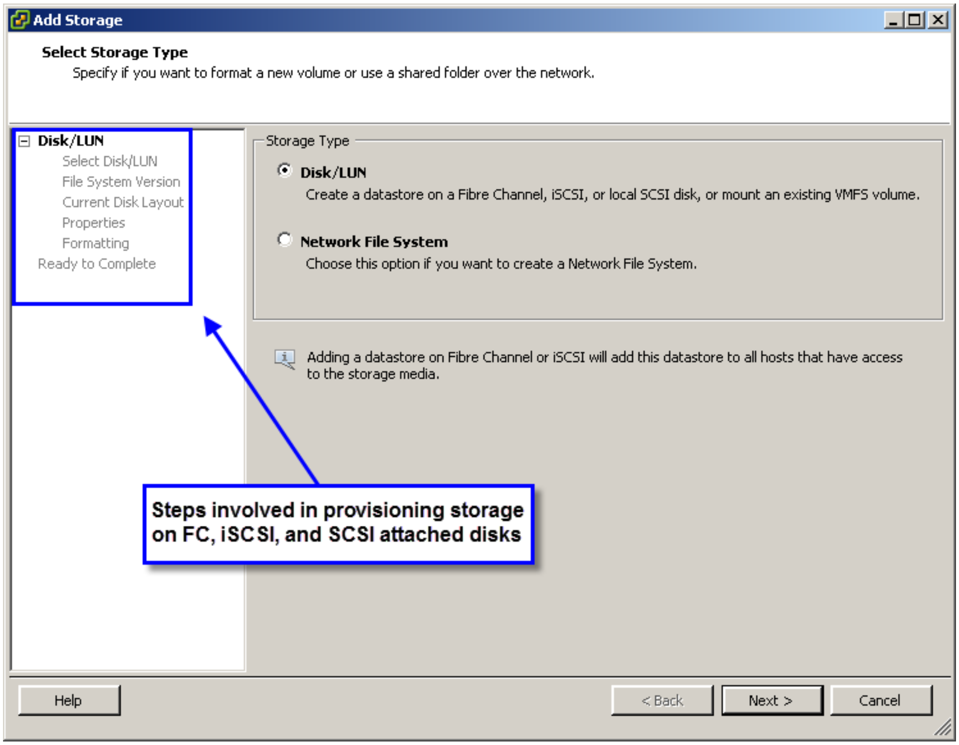


Identification	Status	Device	Drive Type
datastore1	Normal	Local Dell Disk (n...	Non-SSD
TDEV_VMFS5	Normal	EMC Fibre Channel...	Non-SSD
Striped_Meta_TDEV_VMFS5	Normal	EMC Fibre Channel...	Non-SSD
Concat_Meta_TDEV_VMFS5	Normal	EMC Fibre Channel...	Non-SSD
THICK_VMFS5	Normal	EMC Fibre Channel...	Non-SSD

Figure 5 Displaying and managing datastores in the vSphere Client

The **Add Storage** wizard on startup presents a summary of the required steps to provision a new datastore in ESX, as seen in the highlighted box in Figure 6 on page 34.

The **Disk/LUN** option should be selected to provision a datastore on a Fibre Channel, or iSCSI-attached EMC Symmetrix storage array.



Add Storage

Select Storage Type

Specify if you want to format a new volume or use a shared folder over the network.

Storage Type

● **Disk/LUN**
Create a datastore on a Fibre Channel, iSCSI, or local SCSI disk, or mount an existing VMFS volume.

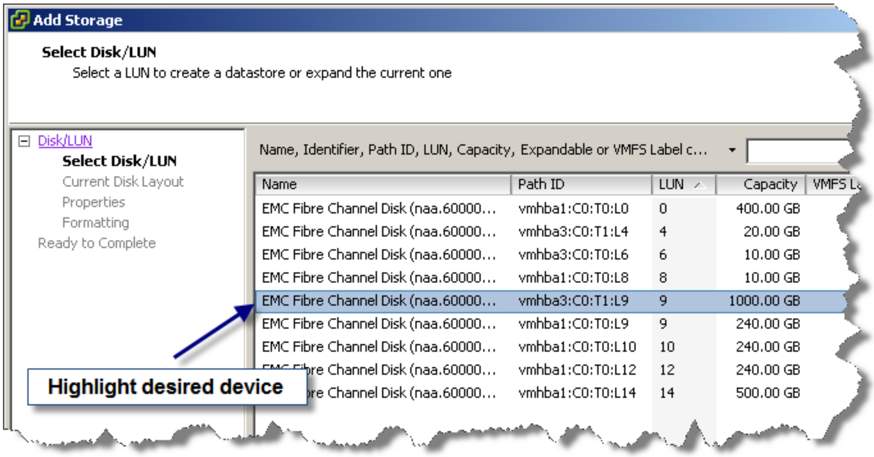
○ **Network File System**
Choose this option if you want to create a Network File System.

Adding a datastore on Fibre Channel or iSCSI will add this datastore to all hosts that have access to the storage media.

Steps involved in provisioning storage on FC, iSCSI, and SCSI attached disks

Figure 6 Provisioning a new datastore in the vSphere Client — Storage Type

Selecting the **Next** button in the wizard presents all viable FC or iSCSI attached devices. The next step in the process involves selecting the appropriate device in the list provided by the wizard and selecting the **Next** button as in Figure 7 on page 35.



Add Storage

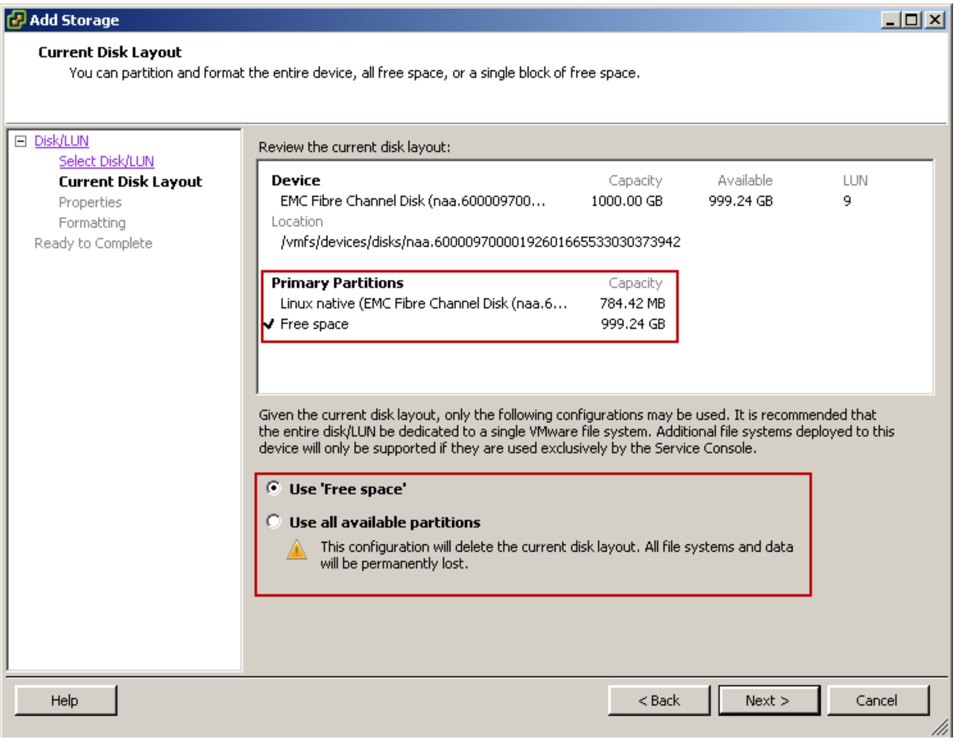
Select Disk/LUN
Select a LUN to create a datastore or expand the current one

Name	Path ID	LUN	Capacity	VMFS L
EMC Fibre Channel Disk (naa.60000...	vmhba1:C0:T0:L0	0	400.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba3:C0:T1:L4	4	20.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba3:C0:T0:L6	6	10.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba1:C0:T0:L8	8	10.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba3:C0:T1:L9	9	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba1:C0:T0:L9	9	240.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba1:C0:T0:L10	10	240.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba1:C0:T0:L12	12	240.00 GB	
re Channel Disk (naa.60000...	vmhba1:C0:T0:L14	14	500.00 GB	

Figure 7 **Provisioning a new datastore in the vSphere Client — Disk/LUN**

It is important to note that devices that have existing VMFS volumes are not presented on this screen[1]. This is independent of whether or not that device contains free space. However, devices with existing non-VMFS formatted partitions but with free space are visible in the wizard. An example of this is exhibited in Figure 8 on page 36.

[1] Devices that have incorrect VMFS signatures will appear in this wizard. Incorrect signatures are usually due to devices that contain replicated VMFS volumes or changes in the storage environment that result in changes to the SCSI personality of devices. For more information on handling these types of volumes please refer to section, "Transitioning disk copies to cloned virtual machines" on page 235.



Current Disk Layout

You can partition and format the entire device, all free space, or a single block of free space.

Review the current disk layout:

Device	Capacity	Available	LUN
EMC Fibre Channel Disk (naa.600009700...	1000.00 GB	999.24 GB	9

Location

/vmfs/devices/disks/naa.60000970000192601665533030373942

Primary Partitions	Capacity
Linux native (EMC Fibre Channel Disk (naa.6...	784.42 MB
✓ Free space	999.24 GB

Given the current disk layout, only the following configurations may be used. It is recommended that the entire disk/LUN be dedicated to a single VMware file system. Additional file systems deployed to this device will only be supported if they are used exclusively by the Service Console.

● **Use 'Free space'**

○ **Use all available partitions**

⚠ This configuration will delete the current disk layout. All file systems and data will be permanently lost.

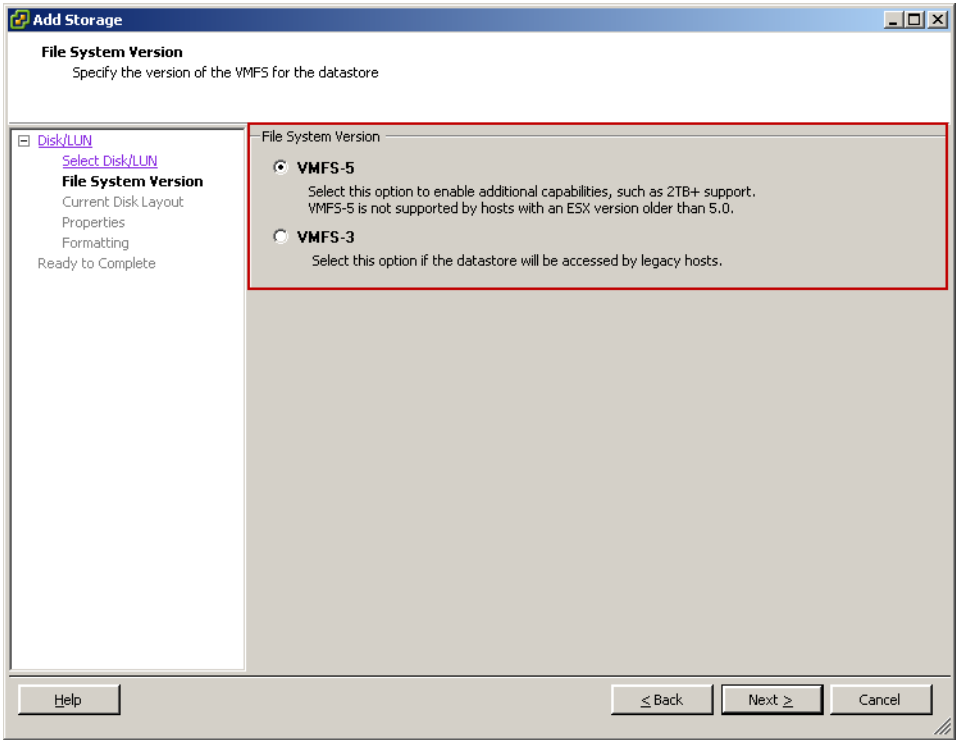
Figure 8 Provisioning a new datastore in the vSphere Client — Disk Layout

---

Note: The vSphere Client allows only one datastore on a device. EMC Symmetrix storage arrays support nondisruptive expansion of storage LUNs. The excess capacity available after expansion can be utilized to expand the existing datastore on the LUN. "Growing VMFS in VMware vSphere using an expanded metavolume" on page 69 focuses on this feature of EMC Symmetrix storage arrays.

In ESXi 5, the next screen will prompt the user for the type of VMFS format, either the newer VMFS-5 format or the older format VMFS-3 if legacy ESX hosts will access the datastore. This screen is shown in Figure 9 on page 37. This screen is not present for ESX 4 hosts since these hosts cannot address VMFS-5 datastores.

Note: The step "File System Version" in the wizard is not available prior to ESXi 5.



Add Storage

File System Version

Specify the version of the VMFS for the datastore

File System Version

● VMFS-5

Select this option to enable additional capabilities, such as 2TB+ support.
VMFS-5 is not supported by hosts with an ESX version older than 5.0.

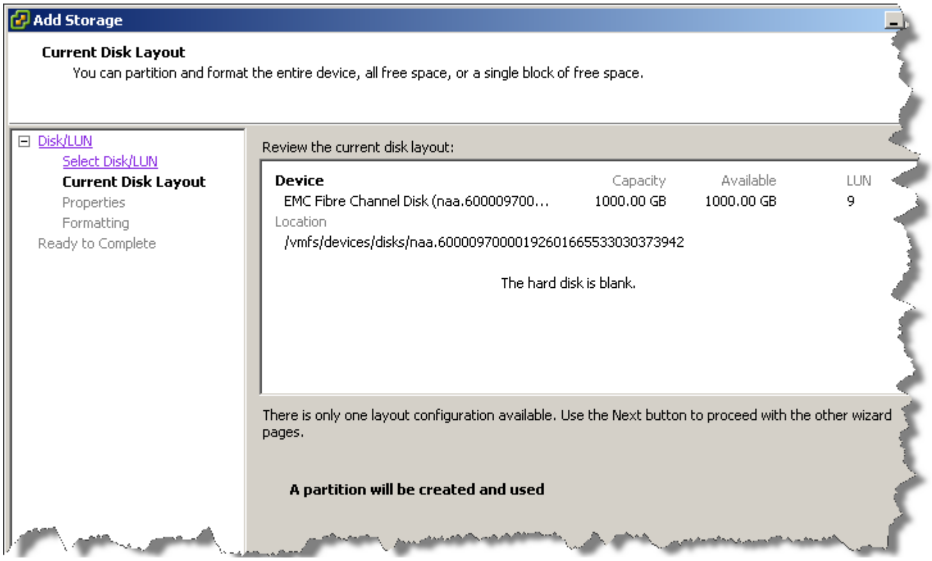
○ VMFS-3

Select this option if the datastore will be accessed by legacy hosts.

Figure 9 Provisioning a new datastore in the vSphere Client — VMFS file system

In ESX 4 and ESXi 5, the user is then presented with either a summary screen or with a screen with two options depending on the configuration of the selected device. If the selected device has no existing partition the wizard presents a summary screen detailing the proposed layout on the selected device.

This is seen in Figure 10 on page 38. For devices with existing partitions (as is the case in the example detailed in Figure 8 on page 36) the wizard will prompt the user with the option of either deleting the existing partition or creating a VMFS volume on the free space available on the device.



Current Disk Layout

You can partition and format the entire device, all free space, or a single block of free space.

Review the current disk layout:

Device	Capacity	Available	LUN
EMC Fibre Channel Disk (naa.600009700...	1000.00 GB	1000.00 GB	9

Location

/vmfs/devices/disks/naa.60000970000192601665533030373942

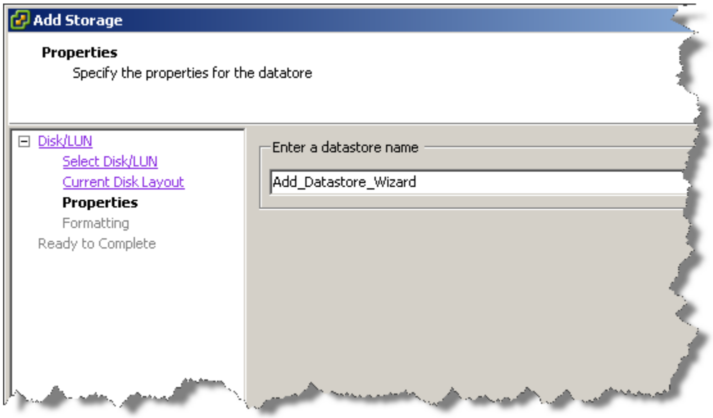
The hard disk is blank.

There is only one layout configuration available. Use the Next button to proceed with the other wizard pages.

A partition will be created and used

Figure 10 **Provisioning a new datastore in the vSphere Client — Disk Layout**

After selecting the appropriate option (if applicable), clicking on the **Next** button on the wizard enables the user to provide a name for the datastore. This is seen in Figure 11.



Properties

Specify the properties for the datatore

Enter a datastore name

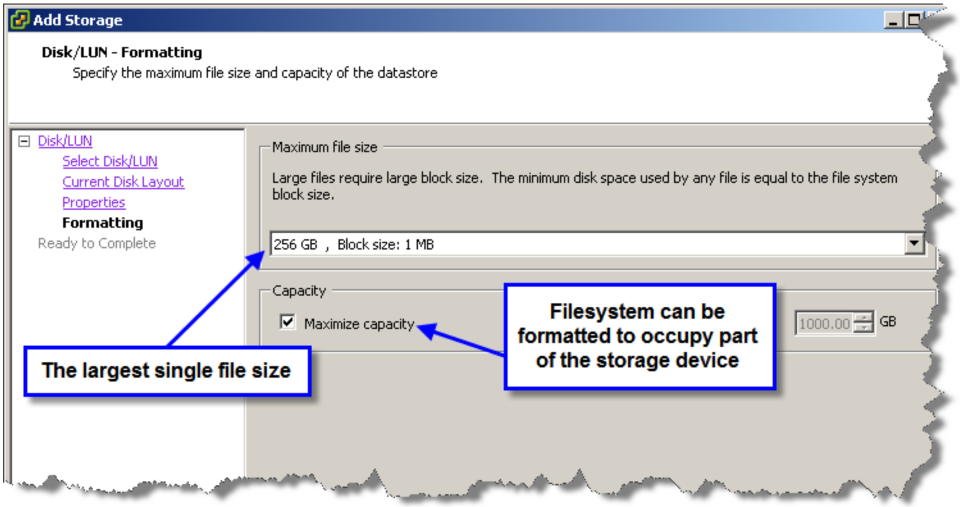
Add_Datastore_Wizard

Figure 11 **Provisioning a new datastore in the vSphere Client — Datastore name**

The final step in the wizard is the selection of options for formatting the device with VMFS. As seen in Figure 12, for ESX 4 the wizard automatically maximizes capacity of the LUN, though it can be customized to a smaller size, and defaults to a block size of 1 MB.

The block size of the VMFS influences the maximum size of a single file on the file system. The default block size (1 MB) should not be changed unless a virtual disk larger than 256 GB has to be created on that file system.

Unlike other file systems, VMFS-3 uses sub-block level allocation for small files. This approach reduces wasted space commonly found in file systems with an average file size smaller than the block size.



Add Storage

Disk/LUN - Formatting

Specify the maximum file size and capacity of the datastore

Maximum file size

Large files require large block size. The minimum disk space used by any file is equal to the file system block size.

256 GB , Block size: 1 MB

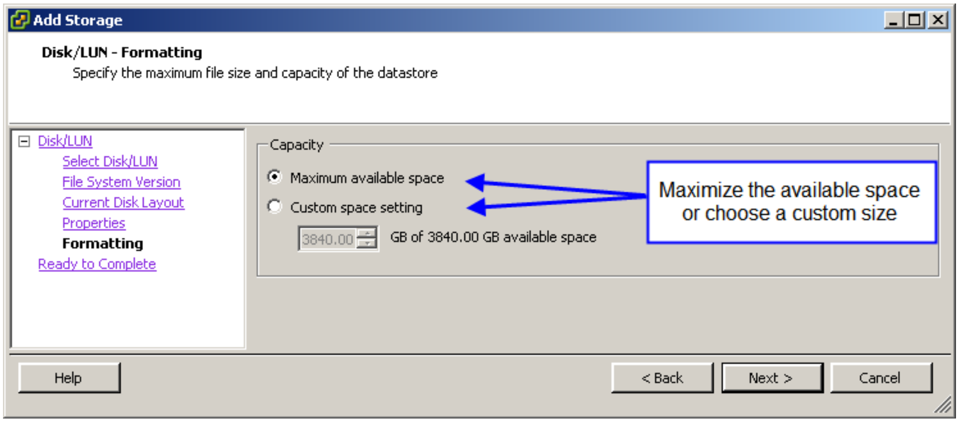
Capacity

☑ Maximize capacity 1000.00 GB

Disk/LUN
Select Disk/LUN
Current Disk Layout
Properties
Formatting
Ready to Complete
The largest single file size
Filesystem can be formatted to occupy part of the storage device

Figure 12 Provisioning a new datastore for ESX 4 — Formatting

For the screen presented in ESXi 5 when selecting a VMFS-5 format, shown in Figure 13, the only option is to choose the capacity. All VMFS-5 datastores are formatted with a 1 MB block size. If the VMFS-3 format is chosen in ESXi 5, the user will be presented with the same options as ESX 4.



Add Storage

Disk/LUN - Formatting

Specify the maximum file size and capacity of the datastore

Capacity

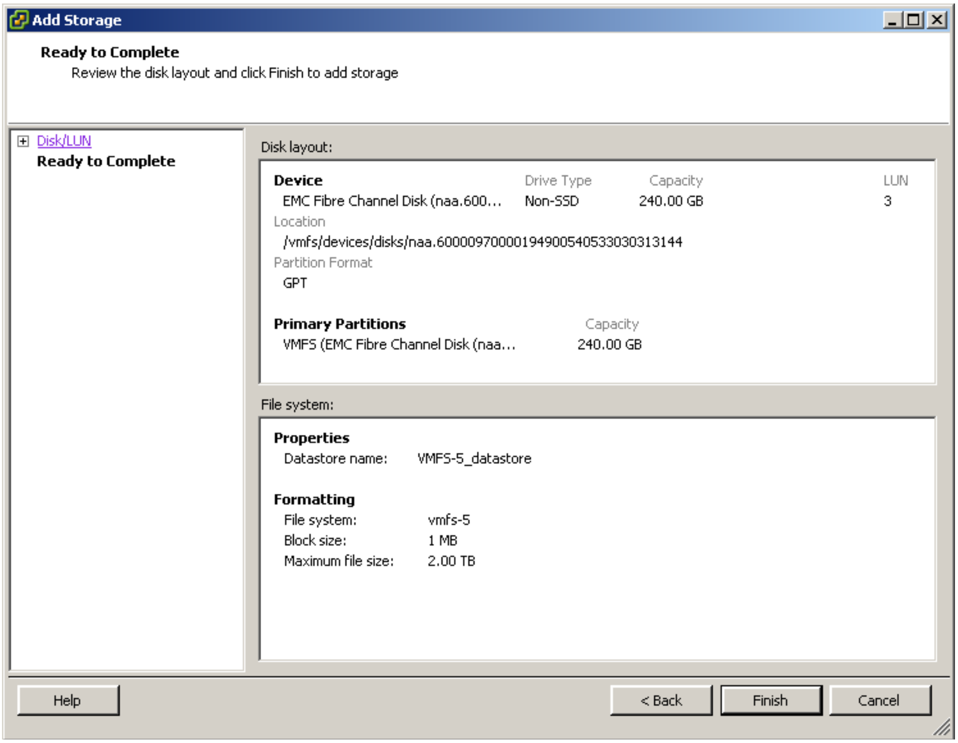
● Maximum available space

○ Custom space setting

3840.00 GB of 3840.00 GB available space

Figure 13 **Provisioning a new datastore for ESXi 5 — Formatting**

Selecting **Next** and **Finish** at the screen in Figure 14 on page 41 results in the creation of a datastore on the selected device.



Add Storage

Ready to Complete

Review the disk layout and click Finish to add storage

Disk/LUN

Ready to Complete

Disk layout:

Device	Drive Type	Capacity	LUN
EMC Fibre Channel Disk (naa.600...	Non-SSD	240.00 GB	3

Location

/vmfs/devices/disks/naa.60000970000194900540533030313144

Partition Format

GPT

Primary Partitions	Capacity
VMFS (EMC Fibre Channel Disk (naa...	240.00 GB

File system:

Properties

Datastore name: VMFS-5_datastore

Formatting

File system: vmfs-5

Block size: 1 MB

Maximum file size: 2.00 TB

Figure 14 Provisioning a new datastore in the vSphere Client — Complete

Creating a VMFS datastore using command line utilities

VMware ESX 4 and ESXi 5 provides a command line utility, **vmkfstools**, to create VMFS.[1] The VMFS volume can be created on either FC or iSCSI attached EMC Symmetrix storage devices by utilizing **fdisk** for ESX 4 and **parted** for ESXi 5. Due to the complexity involved in utilizing command line utilities, VMware and EMC recommends use of the vSphere Client to create a VMware datastore on EMC Symmetrix devices.

[1] The vmkfstools vCLI can be used with ESXi. The vmkfstools vCLI supports most but not all of the options that the vmkfstools service console command supports. See VMware Knowledge Base article 1008194 for more information.

Upgrading VMFS volumes from VMFS-3 to VMFS-5

With the release of vSphere 5 users have the ability to upgrade their existing VMFS-3 datastores to VMFS-5. VMFS-5 has a number of improvements over VMFS-3 such as:

- Support of greater than 2 TB storage devices for each VMFS extent.
- Increased resource limits such as file descriptors.
- Standard 1MB file system block size with support of 2 TB virtual disks.
- Support of greater than 2 TB disk size for RDMs in physical compatibility mode.
- Scalability improvements on storage devices that support hardware acceleration.
- Default use of hardware assisted locking, also called atomic test and set (ATS) locking, on storage devices that support hardware acceleration.
- Online, in-place upgrade process that upgrades existing datastores without disrupting hosts or virtual machines that are currently running.

There are a number of things to consider if you upgrade from VMFS-3 to VMFS-5. With ESXi 5, if you create a new VMFS-5 datastore, the device is formatted with GPT. The GPT format enables you to create datastores larger than 2 TB and up to 64 TB without resorting to the use of physical extent spanning. VMFS-3 datastores continue to use the MBR format for their storage devices. Some things to consider about upgrades:

- For VMFS-3 datastores, the 2 TB limit still applies, even when the storage device has a capacity of more than 2 TB. To be able to use the entire storage space, upgrade a VMFS-3 datastore to VMFS-5. Conversion of the MBR format to GPT happens only after you expand the datastore.
- When you upgrade a VMFS-3 datastore to VMFS-5, any spanned extents will have the GPT format.
- When you upgrade a VMFS-3 datastore, remove from the storage device any partitions that ESXi does not recognize, for example, partitions that use the EXT2 or EXT3 formats. Otherwise, the host cannot format the device with GPT and the upgrade fails.

- You cannot expand a VMFS-3 datastore on devices that have the GPT partition format.

Because of the many considerations when upgrading VMFS, many customers may prefer to create new VMFS-5 datastores and migrate their virtual machines onto them from the VMFS-3 datastores. This can be performed online utilizing Storage vMotion®.

Partition alignment

Modern hard disk systems use the logical block address (LBA) to position the head. This is true for both SCSI and IDE disks. However, older disks systems used a different addressing scheme called CHS (cylinder, head, and sectors) to describe the geometry of the drive. Hard disks using this addressing scheme expect three numbers to position the disk head accurately. Various specifications for IDE and BIOS have evolved over the years to accommodate larger disk storage capacities. These standards provide various combinations for the maximum value for CHS. These range from 1024-65536 for cylinders, 16-255 for heads, and 1-255 sectors per track.

The BIOS of all x86-based computers still supports CHS addressing. The BIOS also provides a mechanism that maps LBA addresses to CHS addresses using the geometry information provided by the disks. Modern operating systems such as Linux do not normally use the mapping information provided by the BIOS to access the disk. However, these operating systems need the geometry information when communicating with the BIOS or with other operating systems that use CHS mapping information, such as DOS or Microsoft Windows.

The first cylinder of all hard disks contains a reserved area called the Master Boot Record (MBR). When an IBM compatible system is booted, the BIOS reads the MBR from the first available disk. The bootstrap loader code found at this location is used to load the operating system. The MBR also contains critical partition table information for four entries describing the location of the primary data partitions on the disk. The partition table structure resembles:

```
struct partition {
char active;     /* 0x80: bootable, 0: not bootable */
char begin[3];   /* CHS for first sector */
char type;
char end[3];     /* CHS for last sector */
int start;      /* 32 bit sector number (counting from 0)
   */
int length;      /* 32 bit number of sectors */
};
```

The reserved space for the MBR can cause alignment issues on devices when using certain operating systems on virtual machines. How to rectify this is addressed in the following section.

Partition alignment for virtual machines using VMFS volumes

EMC Symmetrix storage arrays manage the data on physical disk using a logical construct known as a track. A track can be either 32 KB or 64 KB in size depending on the Symmetrix hardware and the EMC Enginuity code running on the array. A hypervolume can thus be viewed as a collection of 32 KB or 64 KB tracks. A data partition created by VMware ESX consists of all or a subset of the tracks that represent the hypervolume. This is pictorially represented in Figure 15.

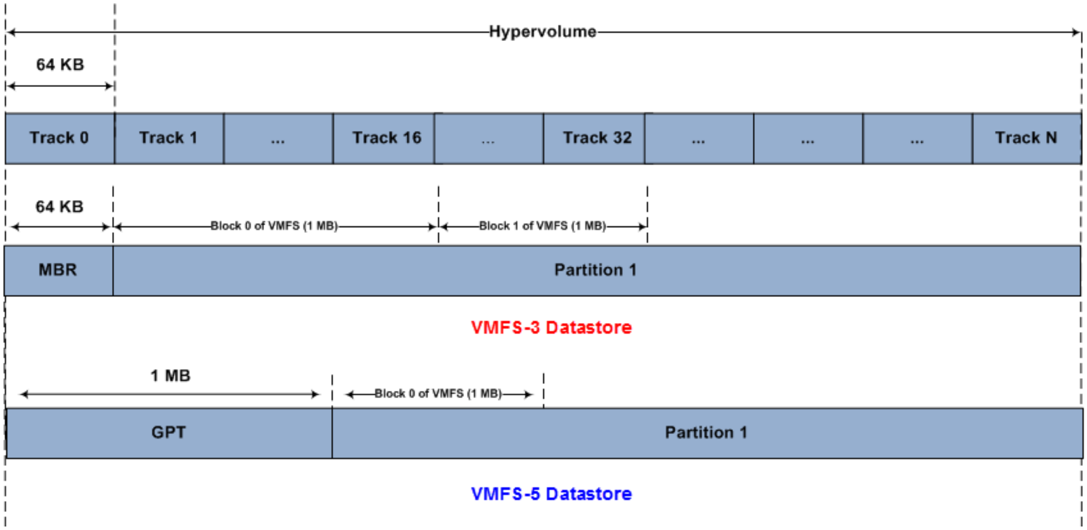


Figure 15 **EMC Symmetrix track mapping of a hypervolume with VMFS**

Figure 15 also shows the layout of the data partition that is created by a VMFS-3 or VMFS-5 datastore when created using the vSphere Client. Figure 15 shows that both VMFS are aligned on a track boundary, despite the fact that VMFS-3 uses a MBR while VMFS-5 employs a GUID partition table or GPT. ESX automatically adjusts the MBR of the VMFS-3 datastore to occupy the first 64 KB of the disk. This is in contrast to VMFS-5 and the GPT since the GPT occupies the first 1 MB by default and does not require adjusting by ESXi. VMware ESX 4 and ESXi 5, by default, create a VMFS on the data partition using a 1 MB block size. Since the block size is a multiple of the track size, file allocations are in even multiples of the track size. Thus, virtual disks created on the partitions normally created by VMware ESX 4 and ESXi 5 are always aligned.

While the VMFS and thus the virtual disks created on that VMFS are aligned, due to CHS addressing explained in "Partition alignment" on page 44, the operating systems of the virtual machines created on those virtual disks are not necessarily aligned.[1] Figure 16 shows a 4 GB virtual disk created on a VMFS-3 which houses a Windows 2003 virtual machine which will be used for this example.

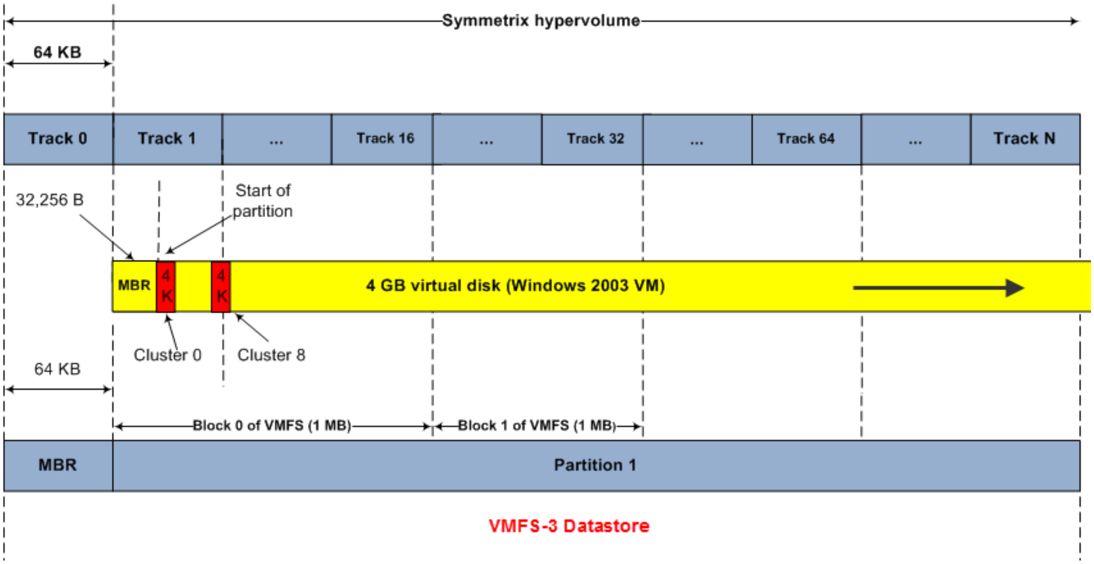


Figure 16 Misalignment in a Windows 2003 virtual machine

As one can see, both the VMFS -3 and virtual disk are aligned to Track 1 of a Symmetrix hypervolume. (This would also be the case with VMFS-5, the only difference being the start of the virtual disk would be at Track 16 as shown in Figure 15 on page 45.) The issue arises when Windows 2003 is installed. Windows takes the first 63 sectors of the virtual disk to write the MBR. Unfortunately that represents 32,256 bytes, just short of 32 KB. Windows begins writing its file system, NTFS, to the partition which starts immediately after the MBR. By default, for this size disk, Windows will write in 4 KB clusters. When the eighth cluster is written, a small part of it will be located in Track 1 but the rest will cross the boundary into Track 2. Again, this is seen in Figure 16 on page 46, with the clusters represented in red. Due to the offset, the track crossing will continue

[1] All Windows 2008, Vista/Windows 7 and higher operating systems are aligned by default.

with cluster 24 being the next one to cross a boundary.

An unaligned partition, as in this case, results in a track crossing and an additional I/O, incurring a penalty on latency and throughput. In other words if the data on cluster 8 is needed the Enginuity operating environment has to allocate cache slots for both track 1 and 2 and populate it with data from the physical medium. The additional I/O (especially if small) can impact system resources significantly. An aligned partition eliminates the need for the additional I/O and results in an overall performance improvement.

Prior experience with misaligned Windows partitions and file systems has shown as much as 20 to 30 percent degradation in performance. Aligning the data partitions on 64 KB boundary results in positive improvements in overall I/O response time experienced by all hosts connected to the shared storage array.

Therefore, EMC recommends aligning the virtual disk on a track boundary to ensure the optimal performance from the storage subsystem.

The alignment process for Microsoft Windows servers has been addressed extensively by Microsoft (for example, see Microsoft support article ID 923076). Readers should follow the process described in these articles to align disks presented to virtual machines running Microsoft Windows operating system. A similar procedure using **fdisk** or **sfdisk** can be used to align disks in virtual machines with Linux as the guest operating system.

The aligned Windows 2003 VM™ ensures optimal use of the EMC Symmetrix storage system since the cluster allocation will line up with the Symmetrix track. Figure 17 on page 48 shows the newly aligned virtual machine. Comparing Figure 16 on page 46 and Figure 17 clearly shows the benefit of aligning the Windows OS to the Symmetrix track.

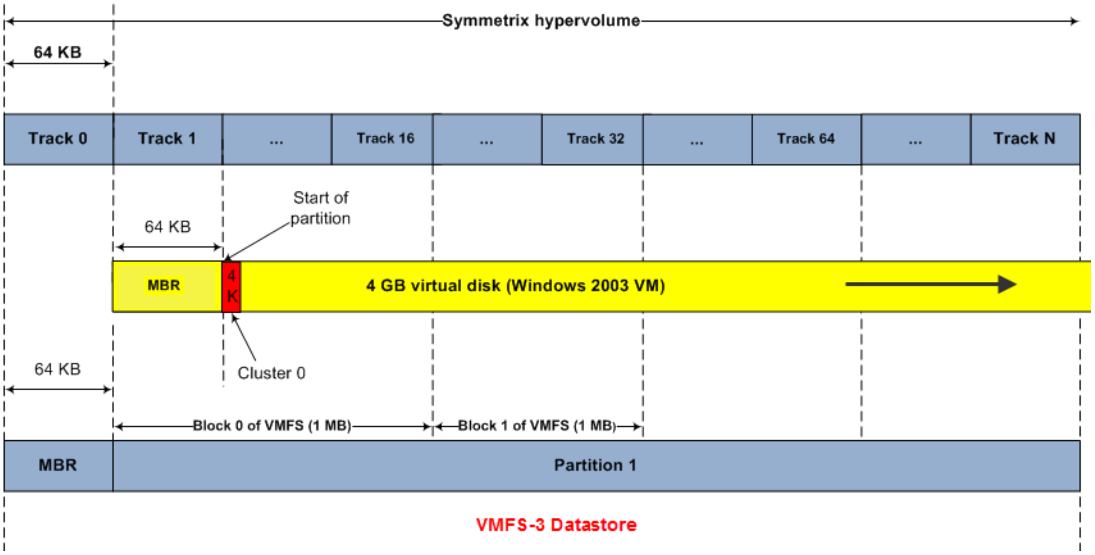


Figure 17 Alignment in a Windows 2003 virtual machine

IMPORTANT

EMC neither recommends nor requires alignment of boot partitions. The partition alignment discussed in this section applies only to volumes containing application data.

Note: The VMware vCenter™ Converter version 5.0 and higher can align partitions on the fly for the user to ensure optimal performance.

Partition alignment for virtual machines using RDM

EMC Symmetrix devices accessed by virtual machines using raw device mapping (RDM) do not contain VMFS volumes. In this configuration, the alignment problem is the same as that seen on physical servers. The process employed for aligning partitions on physical servers needs to be used in the virtual machines.

Mapping VMware devices to EMC Symmetrix devices

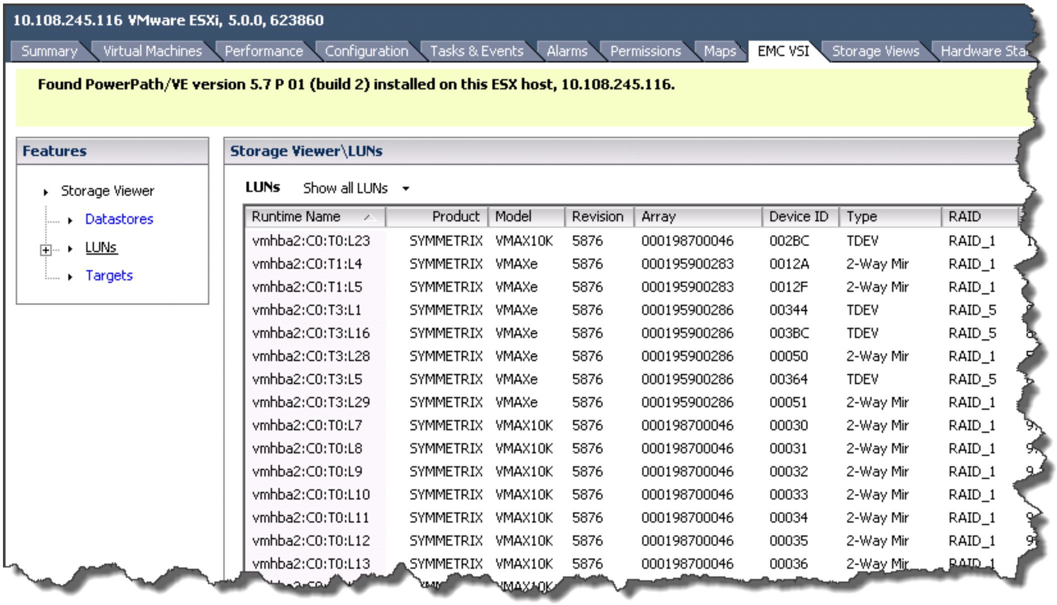
The mapping of the VMware canonical device name to Symmetrix devices is a critical component when using EMC Symmetrix-based storage software. To aid in this, EMC provides a plug-in called EMC Virtual Storage Integrator for VMware vSphere, also known as VSI. This free tool, available to download at www.Powerlink.EMC.com, enables additional capabilities to the vSphere Client, so that users may now view detailed storage-specific information. VSI version 5 fully supports both ESX 4 and ESXi 5. VSI provides simple, storage mapping functionality for various EMC storage-related entities that exist within vSphere Client, including datastores, LUNs, and SCSI targets. It does so through the Storage Viewer feature of VSI. The storage information displayed through VSI provides distinction among the types of storage used, the specific arrays and devices presented, the paths that are used for the storage, and individual characteristics of the existing storage.

Note: EMC Storage Viewer is one of the features of the product EMC Virtual Storage Integrator. EMC Virtual Storage Integrator includes additional features that enable tasks such as path management (Path Management feature), simplified provisioning of Symmetrix storage to VMware vSphere environments (Unified Storage Management feature), and management of SRM environments (SRA Utilities). Further details of the product can be obtained at www.Powerlink.EMC.com.

EMC Storage Viewer provides two main views:

- VMware ESX context view
- Virtual Machine context view.

The LUN subview of the ESX context view, as seen in Figure 18 on page 50, provides detailed information about EMC storage associated with the LUNs visible on the selected ESX host.



10.108.245.116 VMware ESXi, 5.0.0, 623860

Found PowerPath/VE version 5.7 P 01 (build 2) installed on this ESX host, 10.108.245.116.

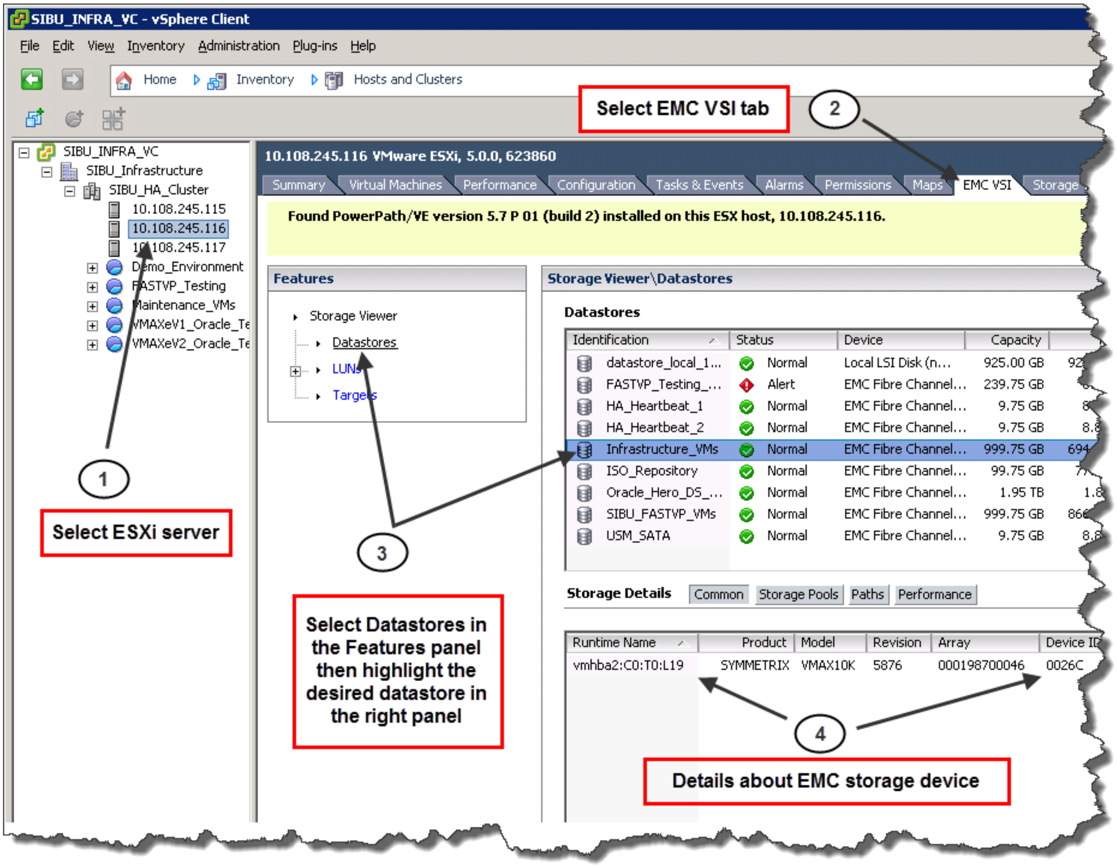
Runtime Name	Product	Model	Revision	Array	Device ID	Type	RAID
vmhba2:C0:T0:L23	SYMMETRIX	VMAX10K	5876	000198700046	002BC	TDEV	RAID_1
vmhba2:C0:T1:L4	SYMMETRIX	VMAXe	5876	000195900283	0012A	2-Way Mir	RAID_1
vmhba2:C0:T1:L5	SYMMETRIX	VMAXe	5876	000195900283	0012F	2-Way Mir	RAID_1
vmhba2:C0:T3:L1	SYMMETRIX	VMAXe	5876	000195900286	00344	TDEV	RAID_5
vmhba2:C0:T3:L16	SYMMETRIX	VMAXe	5876	000195900286	003BC	TDEV	RAID_5
vmhba2:C0:T3:L28	SYMMETRIX	VMAXe	5876	000195900286	00050	2-Way Mir	RAID_1
vmhba2:C0:T3:L5	SYMMETRIX	VMAXe	5876	000195900286	00364	TDEV	RAID_5
vmhba2:C0:T3:L29	SYMMETRIX	VMAXe	5876	000195900286	00051	2-Way Mir	RAID_1
vmhba2:C0:T0:L7	SYMMETRIX	VMAX10K	5876	000198700046	00030	2-Way Mir	RAID_1
vmhba2:C0:T0:L8	SYMMETRIX	VMAX10K	5876	000198700046	00031	2-Way Mir	RAID_1
vmhba2:C0:T0:L9	SYMMETRIX	VMAX10K	5876	000198700046	00032	2-Way Mir	RAID_1
vmhba2:C0:T0:L10	SYMMETRIX	VMAX10K	5876	000198700046	00033	2-Way Mir	RAID_1
vmhba2:C0:T0:L11	SYMMETRIX	VMAX10K	5876	000198700046	00034	2-Way Mir	RAID_1
vmhba2:C0:T0:L12	SYMMETRIX	VMAX10K	5876	000198700046	00035	2-Way Mir	RAID_1
vmhba2:C0:T0:L13	SYMMETRIX	VMAX10K	5876	000198700046	00036	2-Way Mir	RAID_1

Figure 18 LUN subview of the VMware ESX context view of the EMC Storage Viewer feature of VSI

Detailed discussion of EMC VSI is beyond the scope of this document. Interested readers should consult the *EMC VSI for VMware vSphere: Storage Viewer Product Guide,* available at www.Powerlink.EMC.com.

Mapping the VMFS datastore to EMC Symmetrix devices

The mapping of components of VMFS to EMC Symmetrix devices can be an essential piece of information when troubleshooting or configuring/changing underlying storage. EMC VSI, discussed in the previous section, can be used to graphically provide these relationships. Figure 19 shows an example of the information provided by the VSI Storage Viewer feature that simplifies the process of determining the relationship between VMware datastores and EMC Symmetrix devices.



Found PowerPath/VE version 5.7 P 01 (build 2) installed on this ESX host, 10.108.245.116.

Identification	Status	Device	Capacity	
datastore_local_1...	Normal	Local LSI Disk (n...	925.00 GB	92
FASTVP_Testing_...	Alert	EMC Fibre Channel...	239.75 GB	
HA_Heartbeat_1	Normal	EMC Fibre Channel...	9.75 GB	8
HA_Heartbeat_2	Normal	EMC Fibre Channel...	9.75 GB	8.8
Infrastructure_VMs	Normal	EMC Fibre Channel...	999.75 GB	694
ISO_Repository	Normal	EMC Fibre Channel...	99.75 GB	7
Oracle_Hero_DS_...	Normal	EMC Fibre Channel...	1.95 TB	1.8
SIBU_FASTVP_VMs	Normal	EMC Fibre Channel...	999.75 GB	866
USM_SATA	Normal	EMC Fibre Channel...	9.75 GB	8.8

Runtime Name	Product	Model	Revision	Array	Device ID
vmhba2:C0:T0:L19	SYMMETRIX	VMAX10K	5876	000198700046	0026C

Figure 19 **Datastore subview of the VMware ESX context view of the EMC Storage Viewer feature of VSI**

Detailed discussion of the VSI for VMware vSphere: Storage Viewer is beyond the scope of this document. Interested readers should consult the *EMC VSI for VMware vSphere: Storage Viewer Product Guide* for this feature, available at www.Powerlink.EMC.com.

Optimizing VMware and EMC Symmetrix for interoperability

The EMC Symmetrix product line includes the VMAX (40K, 20K, 10K)/VMAXe Virtual Matrix™ and DMX Direct Matrix Architecture family. EMC Symmetrix is a fully redundant, high-availability storage processor providing non-disruptive component replacements and code upgrades. The Symmetrix system features high levels of performance, data integrity, reliability, and availability. Configuring the Symmetrix storage array appropriately for a VMware ESX environment is critical to ensure scalable, high-performance architecture. This section discusses these best practices.

Storage considerations for VMware ESX/ESXi

Physical disk size and data protection

EMC Symmetrix storage arrays offer customers a wide choice of physical drives to meet different workloads. These include high performance 100 GB, 200 GB, or 400 GB Enterprise Flash Drives (EFD) in 3.5" or 2.5"; 300 GB or 600 GB, 10k or 15k rpm Fibre Channel drives; 2 TB and 3 TB 7200 rpm SATA-II drives; and small form-factor 2.5" SAS drives from 146 GB to1 TB which can reduce power consumption by 40 percent and weigh 54 percent less while delivering performance similar to the 10K rpm Fibre Channel drives. Various drive sizes can be intermixed on the same storage array to allow customers with the option of providing different applications with the appropriate service level.

In addition to the different physical drives, EMC also offers various protection levels on an EMC Symmetrix storage array. The storage arrays support RAID 1, RAID 10, RAID 5 and RAID 6 protection types. The RAID protection type can be mixed not only in the same storage array but also on the same physical disks. Furthermore, the storage utilization can be optimized by deploying Virtual Provisioning™ and Fully Automated Storage Tiering for Virtual Pools or FAST™ VP.

EMC Symmetrix VMAX FAST VP automates the identification of data volumes for the purposes relocating application data across different performance/capacity tiers within an array. FAST VP pro-actively monitors workloads at both the LUN and sub-LUN level in order to identify "busy" data that would benefit from being moved to higher performing drives. FAST VP will also identify less "busy" data that could be relocated to higher capacity drives, without existing performance being affected. This promotion/demotion activity is based on policies that associate a storage group to multiple drive technologies, or RAID protection schemes, via thin storage pools, as well as the performance requirements of the application contained within the storage group. Data movement executed during this activity is performed non-disruptively, without affecting business continuity and data availability.

Introduced in Enginuity 5876[1], Symmetrix adds to the drive options through Federated Tiered Storage™, or FTS. FTS allows supported, SAN-attached disk arrays to provide physical disk space for a Symmetrix VMAX array. This permits the user to manage, monitor, migrate, and replicate data residing on both Symmetrix VMAX and other supported arrays using familiar EMC software and Enginuity features. This applies equally to data that already exists on external arrays, as well as to new storage that is being allocated.

The flexibility provided by EMC Symmetrix storage arrays enables customers to provide different service levels to the virtual machines using the same storage array. However, to configure appropriate storage for virtual infrastructure, a knowledge of the anticipated I/O workload is required.

For customers who leverage FAST VP in their environments, using it in VMware vSphere environments can take the guesswork out of much of the disk placement. For example, a virtual disk could be placed in a thin pool that is part of a FAST VP policy, thus allowing a more nuanced placement of the data over time, the result being that the most accessed data would be moved to the fastest disks in the policy, while the least accessed would be placed on the slower, more cost-effective tier.

LUN configuration and size presented to the VMware ESX hosts

The most common configuration of a VMware ESX cluster presents the storage to the virtual machines as flat files in a VMFS. It is, therefore, tempting to present the storage requirement for the VMware ESX hosts as one large LUN. This is even more so the case in vSphere 5 which supports datastores up to 64 TB in size. Using a singular, large LUN, however, can be detrimental to the scalability and performance characteristics of the environment.

[1] To utilize FTS on the VMAX 10K platform requires Enginuity 5876.159.102.

Presenting the storage as one large LUN forces the VMkernel to serially queue I/Os from all of the virtual machines utilizing the LUN. The VMware parameter, **Disk.SchedNumReqOutstanding**, prevents one virtual machine from monopolizing the Fibre Channel queue for the LUN. VMware also has an adaptive queue. ESX 3.5 Update 4 introduced an adaptive queue depth algorithm that is controlled by two parameters: **QFullSampleSize** and **QFullThreshold**. In ESX 3.5 through 5.0 these settings are global to the ESX host. Starting in ESXi 5.1, however, these parameters can be set on a per-device basis. When using Symmetrix storage, EMC recommends leaving the **QFullSampleSize** parameter at the default of 32 and the **QFullThreshold** parameter at the default of 4. Refer to VMware KB article 1008113 for more information.[1] Despite these tuning parameters, a long queue against the LUN results in unnecessary and unpredictable elongation of response time.

This problem can be further exacerbated in configurations that allow multiple VMware ESX hosts to share a single LUN. In this configuration, all VMware ESX hosts utilizing the LUN share the queue provided by the Symmetrix Fibre Channel port. In a large farm with multiple active virtual machines, it is easy to create a very long queue on the EMC Symmetrix storage array front-end port causing unpredictable and sometimes elongated response time. When such an event occurs, the benefits of moderate queuing are lost.

The potential response time elongation and performance degradation can be addressed by presenting a number of smaller LUNs to the VMware ESX cluster. However, this imposes overhead for managing the virtual infrastructure. Furthermore, the limitation of 256 SCSI devices per VMware ESX can impose severe restrictions on the total amount of storage that can be presented.

Table 1 on page 56 compares the advantages and disadvantages of presenting the storage to a VMware ESX farm as a single or multiple LUNs. The table shows that the benefits of presenting storage as multiple LUNs outweigh the disadvantages. EMC recommends presenting the storage for a VMware ESX cluster from EMC Symmetrix storage arrays as striped metavolumes.

[1] Adaptive queue depth parameters should only be altered under the explicit direction of EMC or VMware support. If throttling is required EMC recommends leveraging VMware Storage I/O Control.

The anticipated I/O activity influences the maximum size of the LUN that can be presented to the VMware ESX hosts. The appropriate size for an environment should be determined after a thorough analysis of the performance data collected from existing physical or virtual environments.

Table 1 Comparing approaches for presenting storage to VMware ESX hosts

	Storage as a single LUN	Storage as multiple LUNs
Management	• Easier management. • Storage can be over-provisioned. • One VMFS to manage.	• Small management overhead. • Storage provisioning has to be on demand.
Performance	• Can result in poor response time.	• Multiple queues to storage ensure minimal response times.
Scalability	• Limits number of virtual machines due to response time elongation. • Limits number of I/O-intensive virtual machines.	• Multiple VMFS allow more virtual machines per ESX server. • Response time of limited concern (can optimize).
Functionality	• All virtual machines share one LUN. • Cannot leverage ***all*** available storage functionality.	• Use VMFS when storage functionality not needed. • Enables judicious use of RDMs as needed.

Spanned VMFS

VMFS supports concatenation of multiple SCSI disks to create a single file system. Allocation schemes used in VMFS spread the data across all LUNs supporting the file system thus exploiting all available spindles. Spanned VMFS tend to be most useful when the business requirements on ESX 4 call for a VMFS volume larger than 2 TB. As ESXi 5 supports extent sizes well beyond 2 TB, spanned VMFS is less applicable to this version. The benefits versus the risks of using a spanned VMFS on ESX 4 should be evaluated before deciding on the best possible solution for any VMware environment.

IMPORTANT

In a VMFS-3 or VMFS-5 file system, if any member besides the first extent (the head) of a spanned VMFS volume is unavailable, the datastore will be still available for use, except for the data from the missing extent.

⚠ CAUTION

Although the loss of a physical extent is not of great concern in the EMC Symmetrix storage systems, good change control mechanisms are required to prevent inadvertent loss of access.

Number of VMFS in a VMware environment

Virtualization enables better utilization of IT assets and fortunately the fundamentals for managing information in the virtualized environment are no different from a physical environment. EMC recommends the following best practices for a virtualized infrastructure:

- A VMFS to store virtual machine boot disks. In most modern operating systems, there is minimal I/O to the boot disk. Furthermore, most of the I/O to boot disk tend to be paging activity that is sensitive to response time. By separating the boot disks from application data, the risk of response time elongation due to application-related I/O activity is mitigated.
- Data managers such as Microsoft SQL Server and IBM MQSeries use an active log and recovery data structure that track changes to the data. In case of an unplanned application or operating system disruption, the active log or the recovery data structure is critical to ensure proper recovery and data consistency. Since the recovery structures are a critical component, any virtual machine that supports data managers should be provided a separate VMFS for storing active log files and other structures critical for recovery. Furthermore, if mirrored recovery structures are employed, the copy should be stored in a separate VMFS.
- Application data, including database files, should be stored in a separate VMFS. Furthermore, this file system should not contain any structures that are critical for application or database recovery.
- As discussed in "Physical disk size and data protection" on page 53, VMware ESX serializes and queues all I/Os scheduled for a SCSI target. The average response time from the disk depends on the average queue length and residency in the queue. As the utilization rate of the disks increases, the queue length and hence, the response time increase nonlinearly. Therefore, applications requiring high performance or predictable response time should be provided their own VMFS.

- VMware ESX provides a sophisticated mechanism to provide fair access to the disk subsystem. This is known as Storage I/O Control, or SIOC. This capability allows virtual machines with different service-level requirements to share the same VMFS.

Expanding Symmetrix metavolumes

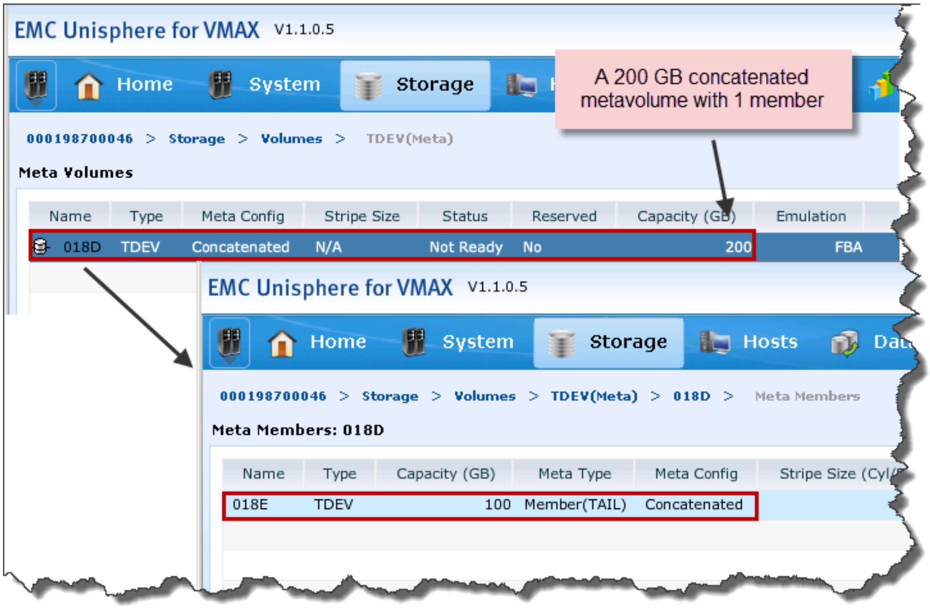
EMC Symmetrix storage arrays offer ways of non-disruptively increasing the size of metavolumes presented to hosts. Concatenated or striped metavolumes can be increased in size without affecting host connectivity. In addition to being comprised out of regular devices, metavolumes can be formed out of thin devices, bringing additional benefits to using metavolumes.

Concatenated metavolume expansion

Concatenated metavolumes are the easiest to expand due to how they are constructed. Concatenated devices are volume sets that are organized with the first byte of data at the beginning of the first device. Addressing continues linearly to the end of the first metamember before any data on the next member is referenced. Thus the addressing mechanism used for a concatenated device ensures the first metamember receives all the data until it is full, and then data is directed to the next member and so on. Therefore, when a new device is added it just needs to be appended to the end of the metavolume without requiring any movement or reorganizing of data.

Figure 20 shows the Symmetrix device information in EMC Unisphere for VMAX (Unisphere) of a concatenated two member (one head, one member (tail)) metavolume 1ED. It is currently 200 GB in size[1].

[1] For more information on Unisphere for VMAX and Solutions Enabler, refer to Chapter 2, “Management of EMC Symmetrix Arrays.”


EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
000198700046 > Storage > Volumes > TDEV(Meta)
EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
Hosts
Dat
000198700046 > Storage > Volumes > TDEV(Meta) > 018D > Meta Members

A 200 GB concatenated metavolume with 1 member

Meta Volumes

Name	Type	Meta Config	Stripe Size	Status	Reserved	Capacity (GB)	Emulation
018D	TDEV	Concatenated	N/A	Not Ready	No	200	FBA

Meta Members: 018D

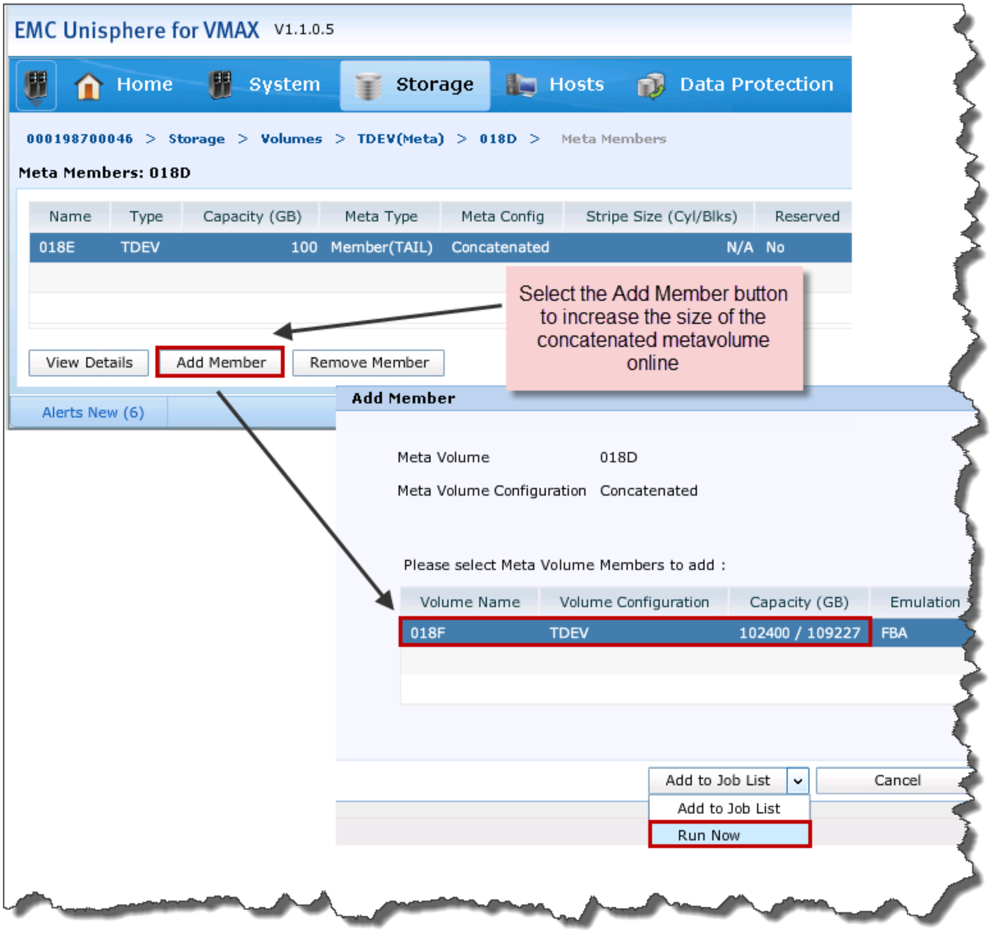
Name	Type	Capacity (GB)	Meta Type	Meta Config	Stripe Size (Cyl
018E	TDEV	100	Member(TAIL)	Concatenated	

Figure 20 **Concatenated metavolume before expansion with Unisphere for VMAX**

With Solutions Enabler, to add additional members to an existing concatenated metavolume, use the following syntax from within the symconfigure command:

```
add dev SymDevName[:SymDevName] to meta SymDevName;
```

In this example, one more member will be added to the metavolume to increase the size from 200 GB to 300 GB. The procedure to perform this using Unisphere is shown in Figure 21.



Meta Members: 018D

Name	Type	Capacity (GB)	Meta Type	Meta Config	Stripe Size (Cyl/Blks)	Reserved
018E	TDEV	100	Member(TAIL)	Concatenated	N/A	No

Select the Add Member button to increase the size of the concatenated metavolume online

Add Member

Meta Volume 018D

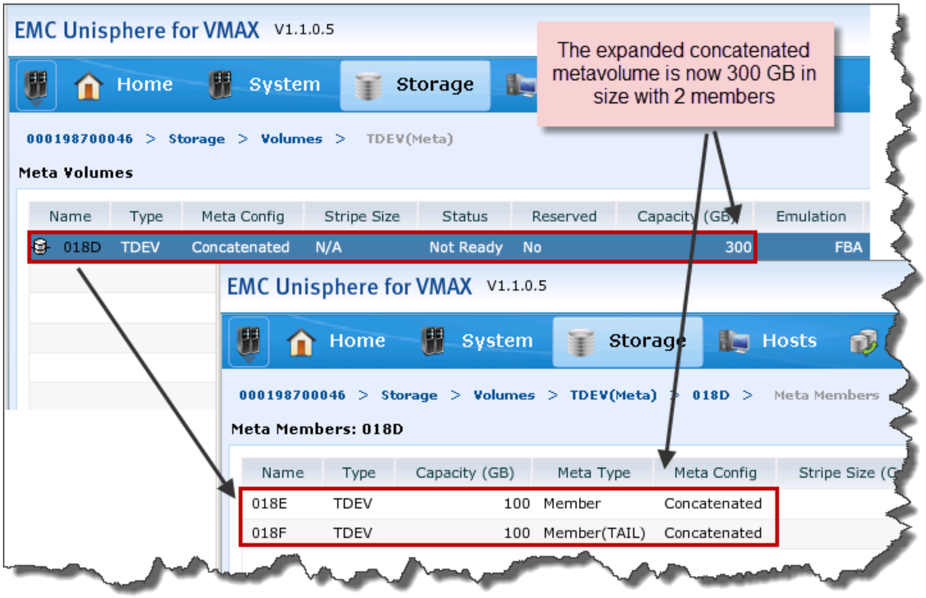
Meta Volume Configuration Concatenated

Please select Meta Volume Members to add :

Volume Name	Volume Configuration	Capacity (GB)	Emulation
018F	TDEV	102400 / 109227	FBA

Figure 21 Expanding a concatenated metavolume using Unisphere for VMAX

Figure 22 shows the metavolume after it has been expanded with Unisphere which is now 300 GB in size.



EMC Unisphere for VMAX V1.1.0.5

000198700046 > Storage > Volumes > TDEV(Meta)

Meta Volumes

Name	Type	Meta Config	Stripe Size	Status	Reserved	Capacity (GB)	Emulation
018D	TDEV	Concatenated	N/A	Not Ready	No	300	FBA

The expanded concatenated metavolume is now 300 GB in size with 2 members

EMC Unisphere for VMAX V1.1.0.5

000198700046 > Storage > Volumes > TDEV(Meta) > 018D > Meta Members

Meta Members: 018D

Name	Type	Capacity (GB)	Meta Type	Meta Config	Stripe Size (G
018E	TDEV	100	Member	Concatenated	
018F	TDEV	100	Member(TAIL)	Concatenated	

Figure 22 Concatenated metavolume after expansion with Unisphere for VMAX

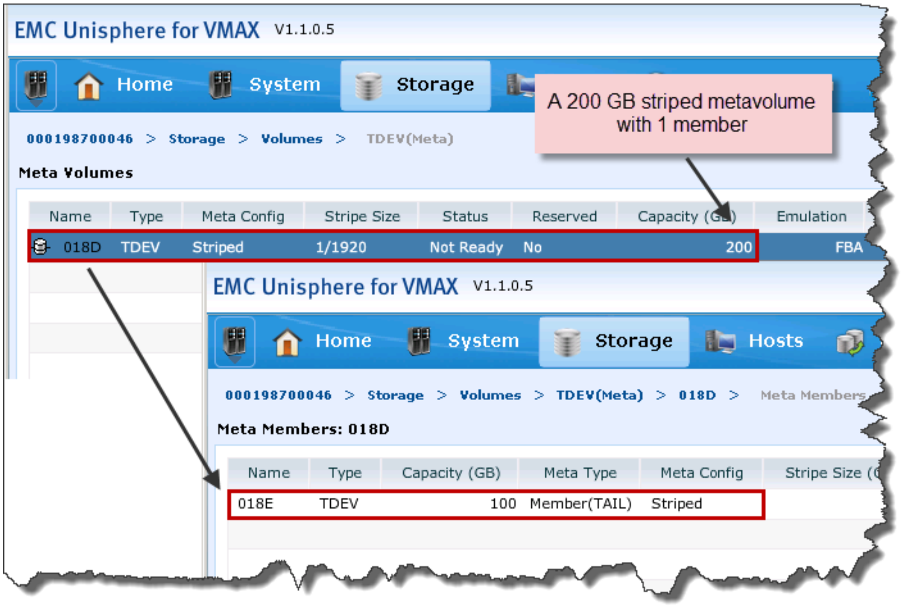
Striped metavolume expansion

Striped metavolumes are a little more complicated to expand due to a more complex configuration. Striped metavolume addressing divides each metamember into a series of stripes. When the stripe on the last member of the metavolume has been addressed or written to, the stripe of the first member (the meta head) is addressed. Thus, when there is write I/O to a striped volume, equal size stripes of data from each participating drive are written alternately to each member of the set. Striping data across the multiple drives in definable cylinder stripes benefits *sequential I/O* by avoiding stacking multiple I/Os to a single spindle and disk director. This scheme allows creation of volumes larger than the single volume limit of the Enginuity operating system, while balancing the I/O activity between the disk devices and the Symmetrix disk directors.

Due to the more complex architecture of a striped metavolume, expanding one requires the use of a BCV[1] metavolume. Since the data is striped over each volume evenly, simply appending a member at the end of a striped volume would result in an unbalanced amount of data striped over the individual members. For this reason it is not allowed.

Instead, the data is copied over to an identical BCV metavolume first, and then any new members are added to the original metavolume. From there, the data is copied off of the BCV metavolume and re-striped evenly over the newly expanded metavolume.

The following example will use thin devices to demonstrate striped metavolume expansion. The process would be the same for thick devices. Figure 23 shows a screenshot from Unisphere for VMAX of a two member striped meta (one head, one member (tail)). It is currently 200 GB in size.


EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
000198700046 > Storage > Volumes > TDEV(Meta)
A 200 GB striped metavolume with 1 member

Meta Volumes

Name	Type	Meta Config	Stripe Size	Status	Reserved	Capacity (GB)	Emulation
018D	TDEV	Striped	1/1920	Not Ready	No	200	FBA

EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
Hosts
000198700046 > Storage > Volumes > TDEV(Meta) > 018D > Meta Members

Meta Members: 018D

Name	Type	Capacity (GB)	Meta Type	Meta Config	Stripe Size ((
018E	TDEV	100	Member(TAIL)	Striped	

Figure 23 **Viewing a striped metavolume before expansion with Unisphere**

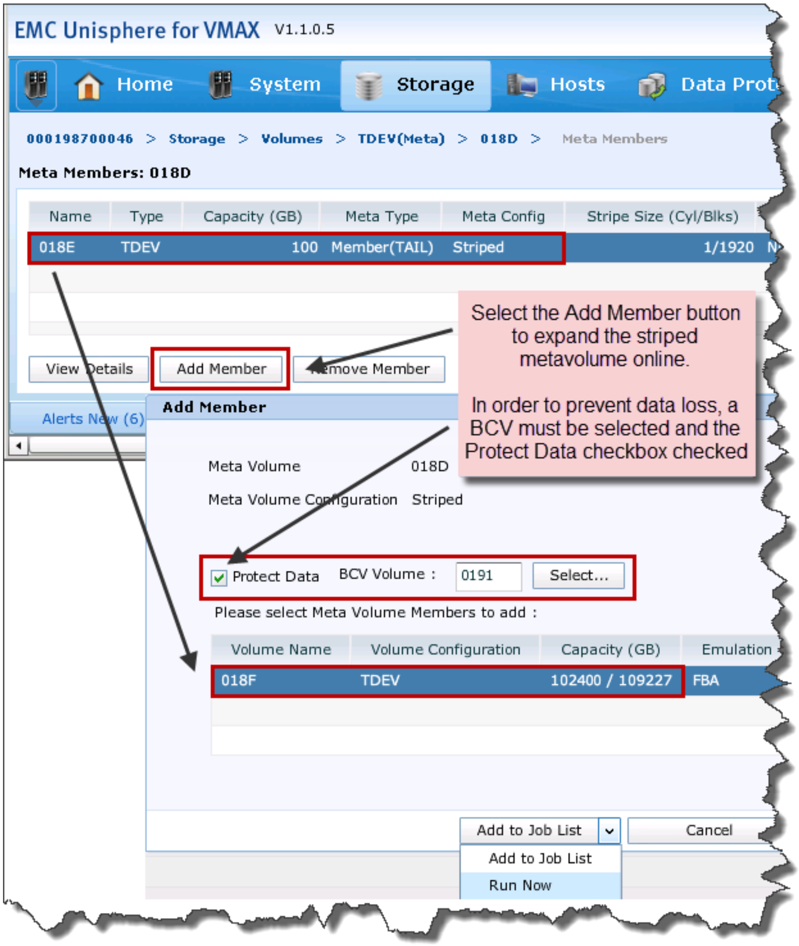
[1] EMC Business Continuous Volume are special devices used with EMC TimeFinder® software. Regardless, expanding a striped metavolume with a BCV **does not** require a TimeFinder license. More information can be found in Chapter 5, “Cloning of vSphere Virtual Machines.”

If using Solutions Enabler to add additional members to an existing striped metavolume, use the following syntax from within the symconfigure command:

```
add dev SymDevName[:SymDevName] to meta SymDev

[,protect_data=[TRUE|FALSE],bcv_meta_head=SymDev];
```

The protect_data option can be either TRUE or FALSE. For an active VMFS volume with stored data, this value should always be set to TRUE. When set to TRUE, the configuration manager automatically creates a protective copy to the BCV meta of the original device striping. Because this occurs automatically, there is no need to perform a BCV establish. When enabling protection with the protect_data option, a BCV meta identical to the existing (original) striped meta has to be specified. The bcv_meta_head value should be set to the device number of a bcv_meta that matches the original metavolume in capacity, stripe count, and stripe size.

In this example, one more member will be added to the striped metavolume to increase the size from 200 GB to 300 GB. This can be seen using Unisphere in Figure 24.



EMC Unisphere for VMAX V1.1.0.5

000198700046 > Storage > Volumes > TDEV(Meta) > 018D > Meta Members

Meta Members: 018D

Name	Type	Capacity (GB)	Meta Type	Meta Config	Stripe Size (Cyl/Blks)
018E	TDEV	100	Member(TAIL)	Striped	1/1920

Add Member

Meta Volume 018D

Meta Volume Configuration Striped

☑ Protect Data BCV Volume : 0191

Please select Meta Volume Members to add :

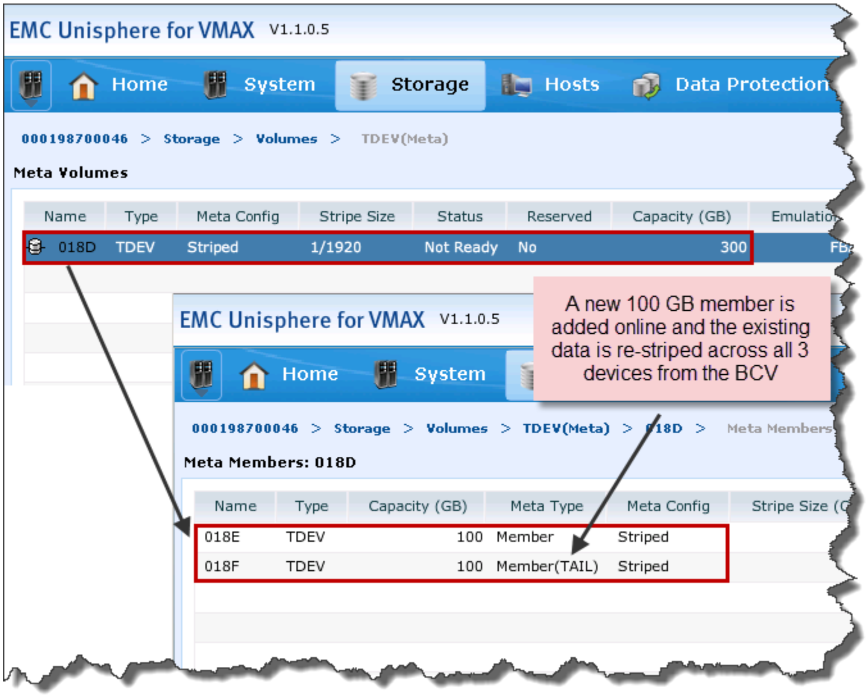
Volume Name	Volume Configuration	Capacity (GB)	Emulation
018F	TDEV	102400 / 109227	FBA

Select the Add Member button to expand the striped metavolume online.

In order to prevent data loss, a BCV must be selected and the Protect Data checkbox checked

Figure 24 Initiating a striped metavolume expansion with Unisphere for VMAX

Figure 25 shows the metavolume after it has been expanded with Unisphere. The data has now been completely re-striped over the newly expanded striped metavolume, which is now 300 GB in size, and includes the newly added member.


EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
Hosts
Data Protection

000198700046 > Storage > Volumes > TDEV(Meta)

Meta Volumes

Name	Type	Meta Config	Stripe Size	Status	Reserved	Capacity (GB)	Emulatio
018D	TDEV	Striped	1/1920	Not Ready	No	300	FB

EMC Unisphere for VMAX V1.1.0.5
Home
System

000198700046 > Storage > Volumes > TDEV(Meta) > 018D > Meta Members

Meta Members: 018D

Name	Type	Capacity (GB)	Meta Type	Meta Config	Stripe Size (G
018E	TDEV	100	Member	Striped	
018F	TDEV	100	Member(TAIL)	Striped	

A new 100 GB member is added online and the existing data is re-striped across all 3 devices from the BCV

Figure 25 Viewing a striped metavolume after expansion with Unisphere for VMAX

Thin meta expansion

The following stipulations refer to all types of thin metavolumes:

- *Members* must be unmasked, unmapped, and unbound before being added to a thin metavolume.
- Only thin devices can be used to create a thin metavolume. Mixing of thin and traditional, non-thin devices to create a metavolume is not allowed.

The following specifically refer to creating and expanding ***concatenated*** thin metavolumes:

- If using Solutions Enabler 7.1 and Enginuity 5874 and earlier, bound thin devices may be converted to concatenated thin metavolumes, but they cannot be mapped and masked to a host. Any data on the thin device is preserved through the conversion, but since the device is unmapped and unmasked, this is an offline operation. Beginning with Solutions Enabler 7.2 and Enginuity 5875, a thin device can be bound and mapped/masked to the host during the conversion, allowing an entirely non-disruptive operation.
- Concatenated thin metavolumes can be expanded non-disruptively regardless of the Solutions Enabler version

The following specifically refer to creating and expanding ***striped*** thin metavolumes:

- For a striped metavolume, all members must have the same device size.
- Only unbound thin devices can be converted to striped thin metavolumes. The striped thin metavolumes must first be formed, and then it can be bound and presented to a host.

If using Solutions Enabler 7.1 and Enginuity 5874 and earlier, bound striped thin metavolumes cannot be expanded online, they must be unbound, unmapped and unmasked. Since this process will cause complete destruction of the data stored on the metavolume, any data on the striped thin metavolume must be migrated off before attempting to expand it in order to prevent this data loss. Beginning with Solutions Enabler 7.2 and Enginuity 5875, a striped thin metavolume can be expanded with no impact to host access or to the data stored on the metavolume with the assistance of a BCV[1].

[1] Previously, BCVs used for thin striped meta expansion had to be thick devices as thin devices with the BCV attribute were not supported. Starting with Enginuity 5875.198 and Solutions Enabler 7.3 thin devices with the BCV attribute are now supported and can now be used for meta expansion. There are a number of restrictions that should be reviewed, however, and which can be found in the EMC Solutions Enabler Symmetrix Array Controls CLI Product Guide on www.Powerlink.EMC.com.

Converting to Symmetrix metavolumes

Symmetrix storage arrays running the Enginuity operating environment 5874 and earlier do not support converting devices online to metavolumes. They first must be unmasked/unmapped before they can be converted. Beginning with Enginuity 5875 and Solutions Enabler 7.2, non-metavolumes can be converted to concatenated[1] metavolumes online without destroying existing data, in other words non-disruptively. Once converted it can then be expanded as needed, again non-disruptively.

[1] Conversion to a striped metavolume from a non-meta device still requires that the device be unmapped first, therefore it is an offline operation. Conversion from a concatenated meta to a striped meta, however, is an online operation.

Growing VMFS in VMware vSphere using an expanded metavolume

This section presents a method to expand a VMFS by utilizing the non-disruptive metavolume expansion capability of the Symmetrix storage system.

vSphere 4 and 5 offer **VMFS Volume Grow,** which allows one to increase the size of a datastore that resides on a VMFS volume without resorting to physical extent spanning. It complements the dynamic LUN expansion capability that exists in Symmetrix storage arrays. If a LUN is increased in size, then **VMFS Volume Grow** enables the VMFS volume to dynamically increase in size as well. Often, virtual machines threaten to outgrow their current VMFS datastores and they either need to be moved to a new, larger datastore or the current datastore needs to be increased in size to satisfy the growing storage requirement. Prior to vSphere, the option for increasing the size of an existing VMFS volume was to extend through a process called spanning. Even if the newly available space was situated upon the LUN which the original VMFS volume resided, the only option was to add an extent. A separate disk partition was created on the additional space and then the new partition could be added to the VMFS volume as an extent. With VMFS Volume Grow the existing partition table is changed and extended to include the additional capacity of the underlying LUN partition. The process of increasing the size of VMFS volumes is integrated into the vSphere Client.

It is important to note there are four options when growing a Virtual Machine File System in vSphere. The following options are available when expanding VMware file systems:

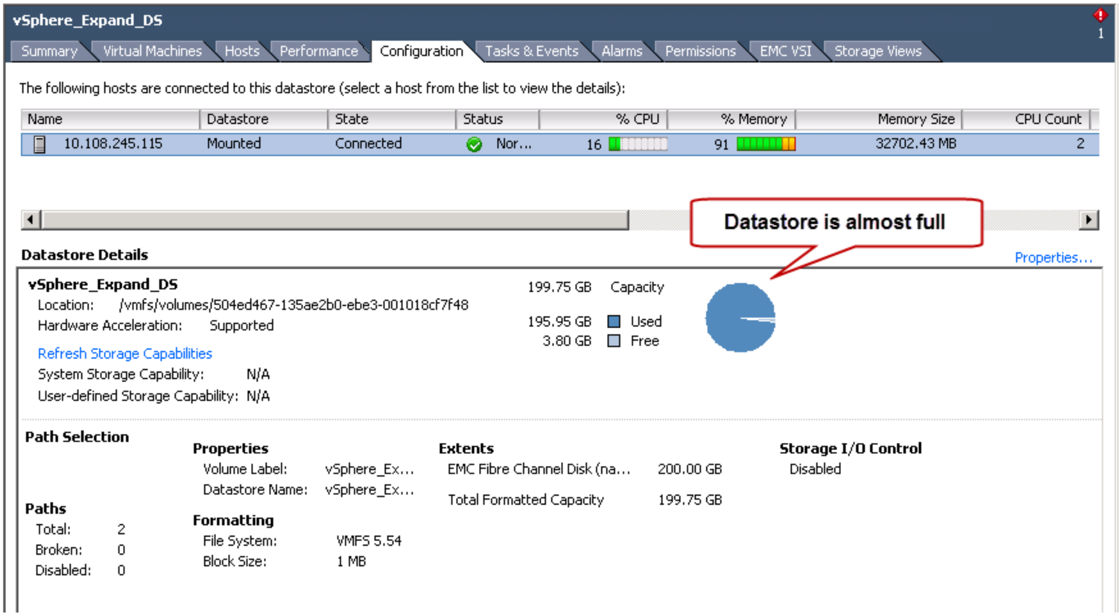
- **Use free space to add new extent—** This options adds the free space on a disk as a new datastore.
- **Use free space to expand existing extent—** This option grows an existing extent to a required capacity.
- **Use free space—** This option deploys an extent in the remaining free space of a disk. This option is available only when adding an extent.

5. **Use all available partitions**— This option dedicates the entire disk to a single datastore extent. This option is available only when adding an extent and when the disk you are formatting is not blank. The disk is reformatted, and the datastores and any data that it contains are erased.

The recommended selection is the “Use free space to expand existing extent” option as using this choice eliminates the need for additional extents and therefore reduces management complexity of affected VMFS.

The process to grow a VMFS by extending a Symmetrix metavolume is listed below:

1. In vCenter Server, as seen in Figure 26, the datastore vSphere_Expand_DS is nearly full. The underlying device has no additional physical storage for VMFS expansion. The datastore consumes 200 GB available on the device. Using **VMFS Volume Grow** and non-disruptive metavolume expansion, this issue can be resolved.



The following hosts are connected to this datastore (select a host from the list to view the details):

Name	Datastore	State	Status	% CPU	% Memory	Memory Size	CPU Count
10.108.245.115	Mounted	Connected	✓ Nor...	16	91	32702.43 MB	2

Datastore is almost full

Datastore Details

vSphere_Expand_DS
Location: /vmfs/volumes/504ed467-135ae2b0-ebe3-001018cf7f48
Hardware Acceleration: Supported
System Storage Capability: N/A
User-defined Storage Capability: N/A

199.75 GB Capacity
195.95 GB Used
3.80 GB Free

Path Selection

Properties
Volume Label: vSphere_Ex...
Datastore Name: vSphere_Ex...

Extents
EMC Fibre Channel Disk (na... 200.00 GB
Total Formatted Capacity 199.75 GB

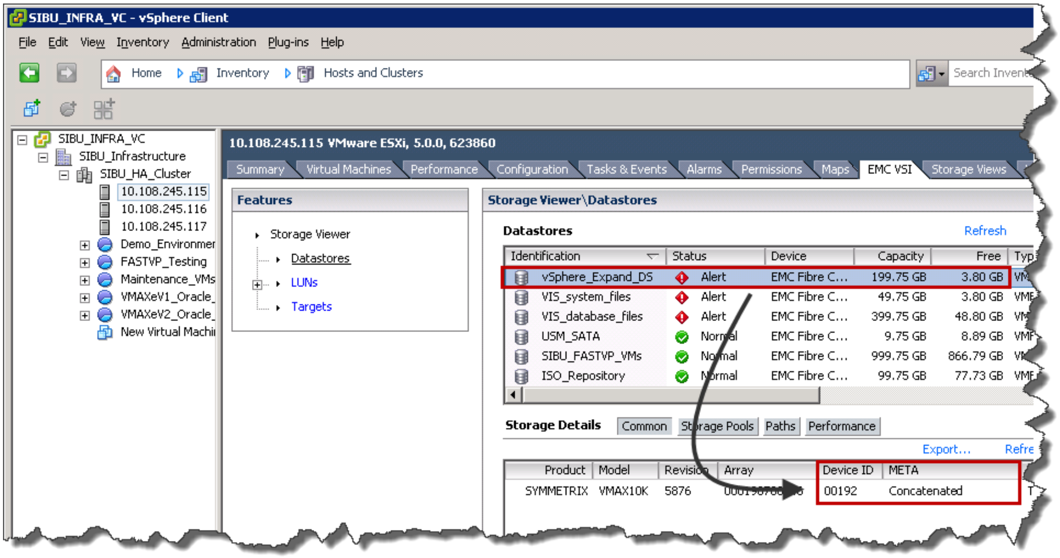
Storage I/O Control
Disabled

Paths
Total: 2
Broken: 0
Disabled: 0

Formatting
File System: VMFS 5.54
Block Size: 1 MB

Figure 26 VMFS datastore before being grown using VMFS Volume Grow

2. The first step in the expansion process is the identification of the metavolume that hosts the VMFS. As shown in Figure 27, EMC VSI can be used to obtain the mapping information.



Datastores

Identification	Status	Device	Capacity	Free	Typ
vSphere_Expand_DS	Alert	EMC Fibre C...	199.75 GB	3.80 GB	VM
VIS_system_files	Alert	EMC Fibre C...	49.75 GB	3.80 GB	VMF
VIS_database_files	Alert	EMC Fibre C...	399.75 GB	48.80 GB	VM
USM_SATA	Normal	EMC Fibre C...	9.75 GB	8.89 GB	VMF
SIBU_FASTVP_VMs	Normal	EMC Fibre C...	999.75 GB	866.79 GB	VM
ISO_Repository	Normal	EMC Fibre C...	99.75 GB	77.73 GB	VMF

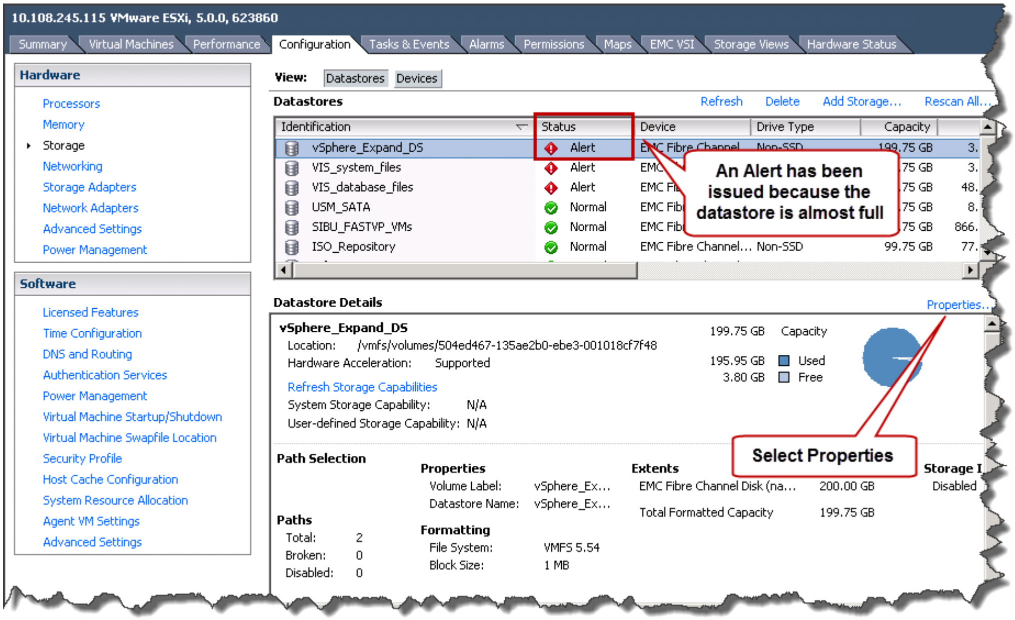
Product	Model	Revision	Array	Device ID	META
SYMMETRIX	VMAX10K	5876	0001987000[illegible]	00192	Concatenated

Figure 27 Identifying the Symmetrix metavolume to be expanded

3. After using one of the methods of metavolume expansion detailed in the previous sections, a rescan needs to be performed on the ESXi host server to see the expanded metavolume's larger size.

   For this example, assume the underlying metavolume at this point has been expanded to provide an additional 100 GB of storage. Once a rescan has been executed, the VMFS datastore can be grown. To do this in vCenter, navigate to the datastores listing on any ESXi host that uses the specified VMFS datastore.

   Select the datastore and click the Properties link, which is shown in Figure 28 on page 72. In this example, the datastore, vSphere_Expand_DS, will be the one grown.



Datastores

Identification	Status	Device	Drive Type	Capacity	
vSphere_Expand_DS	Alert	EMC Fibre Channel	Non-SSD	199.75 GB	3.
VIS_system_files	Alert	EMC		75 GB	3.
VIS_database_files	Alert	EMC Fi		75 GB	48.
USM_SATA	Normal	EMC Fib		75 GB	8.
SIBU_FASTVP_VMs	Normal	EMC Fib		75 GB	866.
ISO_Repository	Normal	EMC Fibre Channel...	Non-SSD	99.75 GB	77.

An Alert has been issued because the datastore is almost full

Datastore Details

vSphere_Expand_DS
Location: /vmfs/volumes/504ed467-135ae2b0-ebe3-001018cf7f48
Hardware Acceleration: Supported

199.75 GB Capacity
195.95 GB ■ Used
3.80 GB □ Free

System Storage Capability: N/A
User-defined Storage Capability: N/A

Path Selection

Paths
Total: 2
Broken: 0
Disabled: 0

Properties
Volume Label: vSphere_Ex...
Datastore Name: vSphere_Ex...

Formatting
File System: VMFS 5.54
Block Size: 1 MB

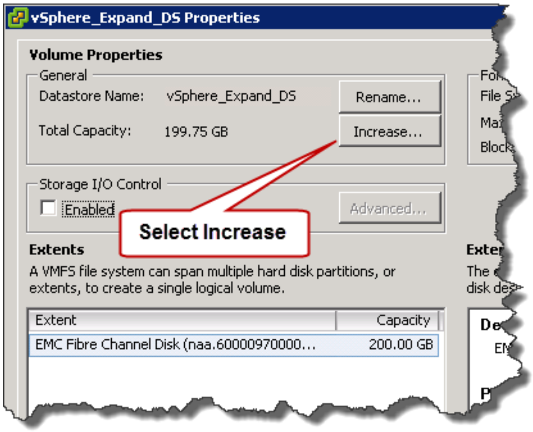
Extents
EMC Fibre Channel Disk (na... 200.00 GB
Total Formatted Capacity 199.75 GB

Storage I
Disabled

Select Properties

Figure 28 VMFS properties in vSphere Client

4. In the Datastore Properties pop-up window, click the Increase button as shown in Figure 29 on page 72.



Volume Properties

General
Datastore Name: vSphere_Expand_DS
Total Capacity: 199.75 GB

Storage I/O Control
□ Enabled

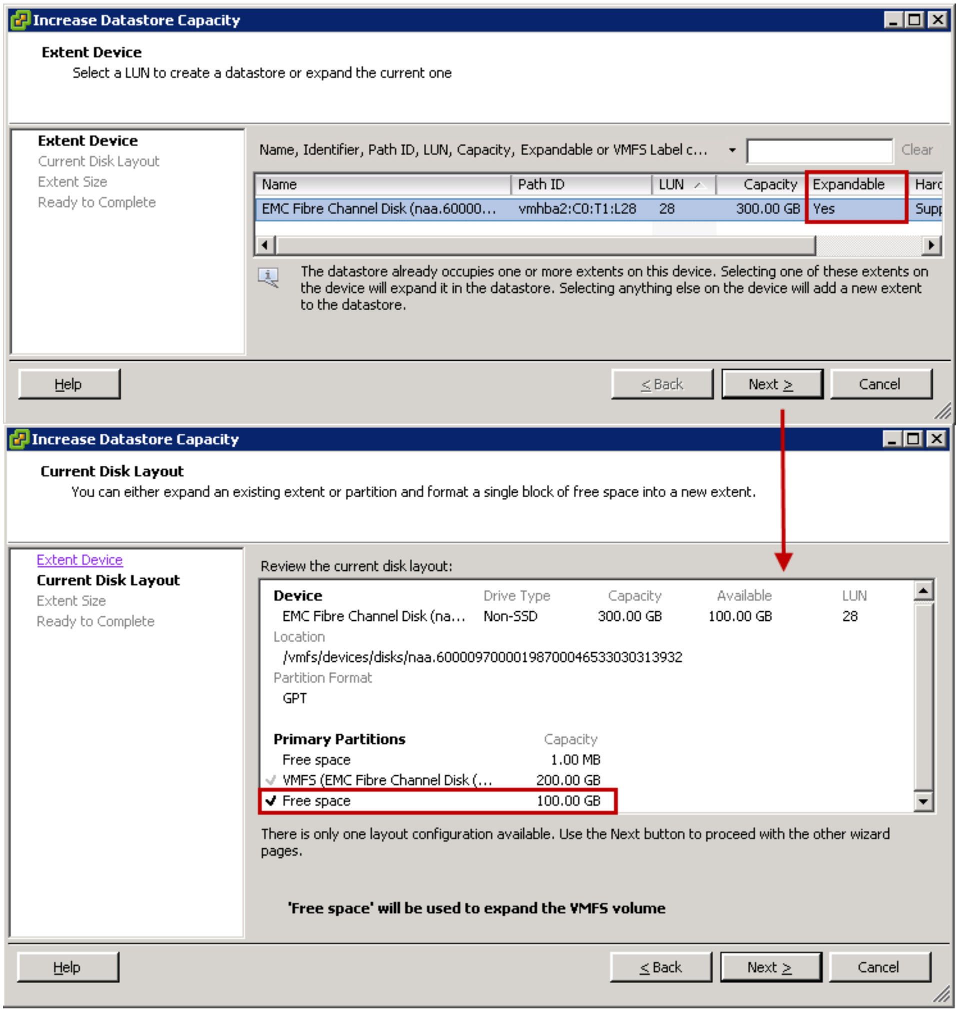
Select Increase

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

Extent	Capacity
EMC Fibre Channel Disk (naa.60000970000...	200.00 GB

Figure 29 Initiating VMFS Volume Grow in vSphere Client

5. In the Increase Datastore Capacity window that appears select the device that it currently resides on. It will have a "Yes" in the Expandable column, as in Figure 30 on page 73. The next window, as seen in Figure 30 on page 73, will confirm the expansion of the VMFS volume; it shows that currently this datastore uses 200 GB and there is 100 GB free.



Extent Device

Select a LUN to create a datastore or expand the current one

Name, Identifier, Path ID, LUN, Capacity, Expandable or VMFS Label c...

Name	Path ID	LUN	Capacity	Expandable	Harc
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L28	28	300.00 GB	Yes	Supp

The datastore already occupies one or more extents on this device. Selecting one of these extents on the device will expand it in the datastore. Selecting anything else on the device will add a new extent to the datastore.

Current Disk Layout

You can either expand an existing extent or partition and format a single block of free space into a new extent.

Review the current disk layout:

Device	Drive Type	Capacity	Available	LUN
EMC Fibre Channel Disk (na...	Non-SSD	300.00 GB	100.00 GB	28

Location

/vmfs/devices/disks/naa.60000970000198700046533030313932

Partition Format

GPT

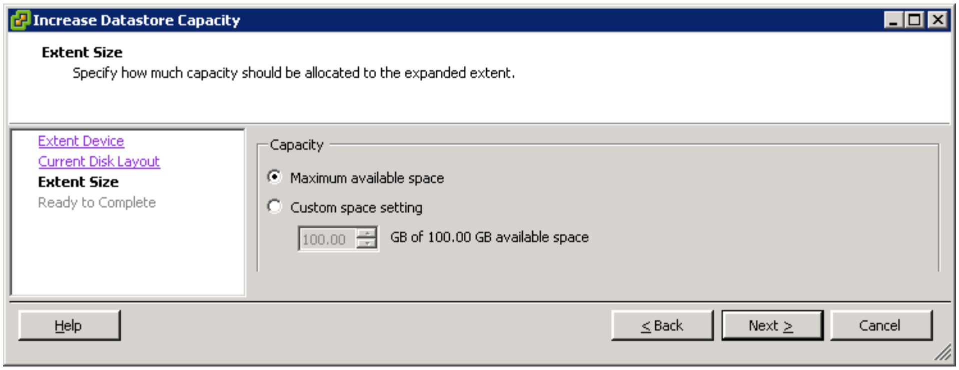
Primary Partitions	Capacity
Free space	1.00 MB
✓ VMFS (EMC Fibre Channel Disk (...	200.00 GB
✓ Free space	100.00 GB

There is only one layout configuration available. Use the Next button to proceed with the other wizard pages.

'Free space' will be used to expand the VMFS volume

Figure 30 Confirming available free space in the VMFS Volume Grow wizard

6. The option to use only part of the newly available space is offered. In Figure 31 the default to maximize all the available space has been chosen and therefore all 100 GB of the newly added space will be merged into the VMFS volume.



Increase Datastore Capacity

Extent Size

Specify how much capacity should be allocated to the expanded extent.

Capacity

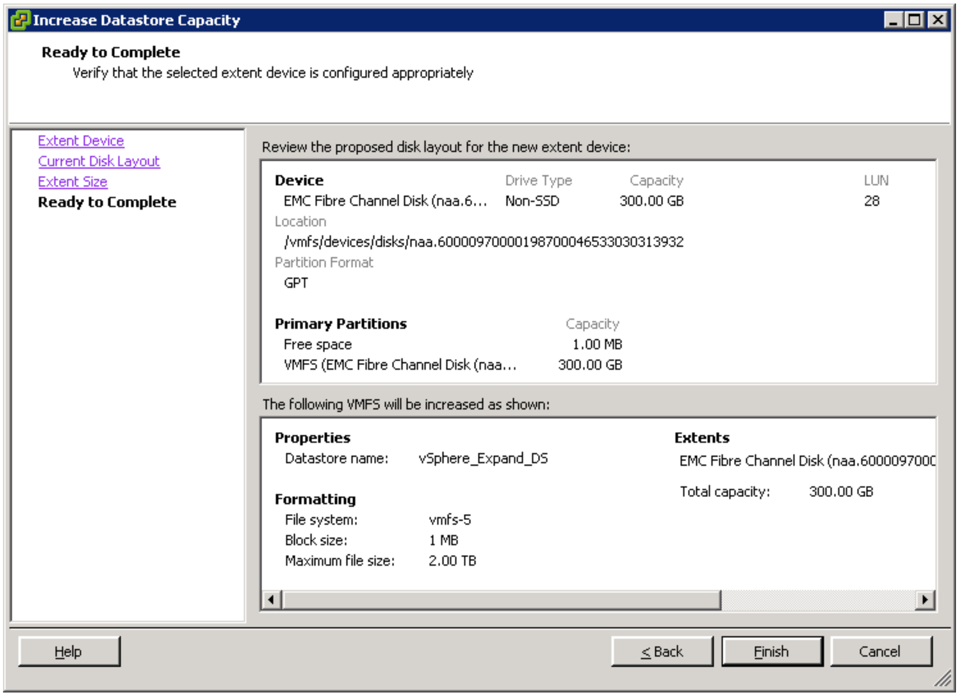
● Maximum available space

○ Custom space setting

100.00 GB of 100.00 GB available space

Figure 31 Choosing how much additional capacity for the VMFS Volume Grow operation

7. A final confirmation screen is presented in Figure 32 on page 75.



Ready to Complete

Verify that the selected extent device is configured appropriately

Review the proposed disk layout for the new extent device:

Device	Drive Type	Capacity	LUN
EMC Fibre Channel Disk (naa.6...	Non-SSD	300.00 GB	28

Location

/vmfs/devices/disks/naa.60000970000198700046533030313932

Partition Format

GPT

Primary Partitions	Capacity
Free space	1.00 MB
VMFS (EMC Fibre Channel Disk (naa...	300.00 GB

The following VMFS will be increased as shown:

Properties

Datastore name: vSphere_Expand_DS

Formatting

File system: vmfs-5

Block size: 1 MB

Maximum file size: 2.00 TB

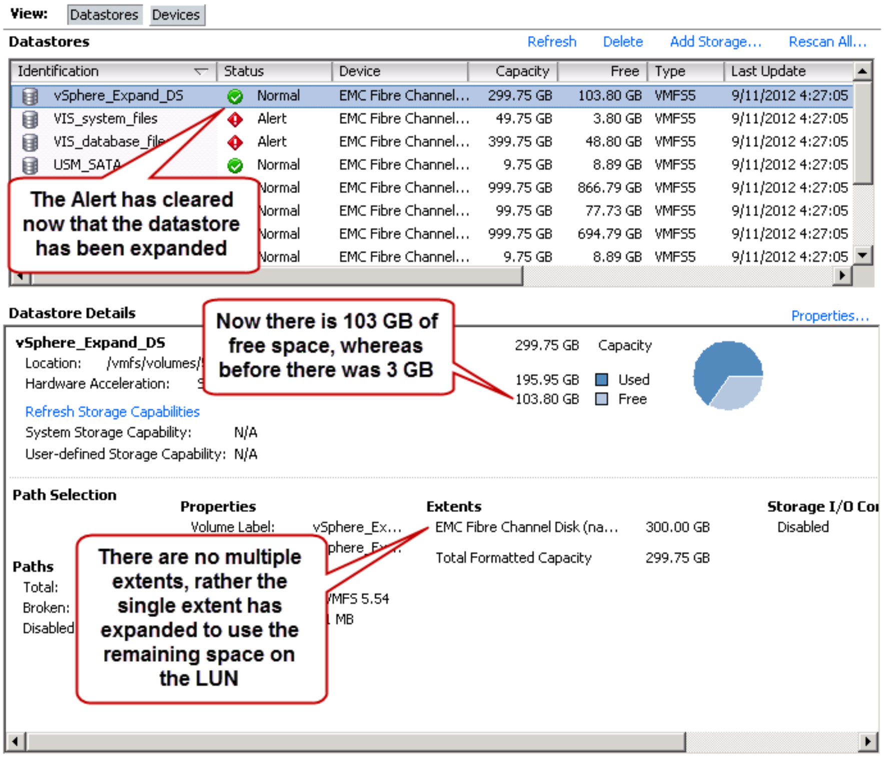
Extents

EMC Fibre Channel Disk (naa.600009700C

Total capacity: 300.00 GB

Figure 32 Increase Datastore Capacity wizard confirmation screen

8. The datastore vSphere_Expand_DS now occupies all 300 GB of the metavolume as shown in Figure 33.



View: Datastores Devices

Datastores Refresh Delete Add Storage... Rescan All...

Identification	Status	Device	Capacity	Free	Type	Last Update
vSphere_Expand_DS	Normal	EMC Fibre Channel...	299.75 GB	103.80 GB	VMFS5	9/11/2012 4:27:05
VIS_system_files	Alert	EMC Fibre Channel...	49.75 GB	3.80 GB	VMFS5	9/11/2012 4:27:05
VIS_database_file	Alert	EMC Fibre Channel...	399.75 GB	48.80 GB	VMFS5	9/11/2012 4:27:05
USM_SATA	Normal	EMC Fibre Channel...	9.75 GB	8.89 GB	VMFS5	9/11/2012 4:27:05
	Normal	EMC Fibre Channel...	999.75 GB	866.79 GB	VMFS5	9/11/2012 4:27:05
	Normal	EMC Fibre Channel...	99.75 GB	77.73 GB	VMFS5	9/11/2012 4:27:05
	Normal	EMC Fibre Channel...	999.75 GB	694.79 GB	VMFS5	9/11/2012 4:27:05
	Normal	EMC Fibre Channel...	9.75 GB	8.89 GB	VMFS5	9/11/2012 4:27:05

The Alert has cleared now that the datastore has been expanded

Datastore Details Properties...

vSphere_Expand_DS
Location: /vmfs/volumes/
Hardware Acceleration: S

Refresh Storage Capabilities
System Storage Capability: N/A
User-defined Storage Capability: N/A

299.75 GB Capacity
195.95 GB ■ Used
103.80 GB □ Free

Now there is 103 GB of free space, whereas before there was 3 GB

Path Selection

Properties
Volume Label: vSphere_Ex...
phere_Ex...
MFS 5.54
1 MB

Extents
EMC Fibre Channel Disk (na... 300.00 GB
Total Formatted Capacity 299.75 GB

Storage I/O Co
Disabled

Paths
Total:
Broken:
Disabled

There are no multiple extents, rather the single extent has expanded to use the remaining space on the LUN

Figure 33 Newly expanded VMFS volume

Path management

Path failover and load balancing in VMware vSphere environments

The ability to dynamically multipath and load balance through the use of the native or third-party storage vendor multipathing software is available in VMware vSphere. The Pluggable Storage Architecture (PSA) is a modular storage construct that allows storage partners (such as EMC with PowerPath®/VE (Virtual Edition)) to write a plug-in to best leverage the unique capabilities of their storage arrays. These modules can interface with the storage array to continuously resolve the best path selection, as well as make use of redundant paths to greatly increase performance and reliability of I/O from the ESX host to storage.

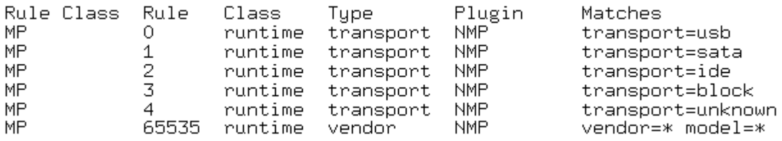
Native Multipathing plug-in

By default, the native multipathing plug-in (NMP) supplied by VMware is used to manage I/O. NMP can be configured to support fixed, most recently used (MRU), and round-robin (RR) path selection polices (PSP).[1]

NMP and other multipathing plug-ins are designed to be able to coexist on an ESX 4 or ESXi 5 host; nevertheless, multiple plug-ins cannot manage the same device simultaneously. To address this, VMware created the concept of claim rules. Claim rules are used to assign storage devices to the proper multipathing plug-in (MPP). When an ESX host boots or performs a rescan, the ESX host discovers all physical paths to the storage devices visible to the host. Using the claim rules defined in the /etc/vmware/esx.conf file, the ESX host determines which multipathing module will be responsible for managing a specific storage device. Claim rules are numbered. For each physical path, the ESX host processes the claim rules starting with the lowest number first. The attributes of the physical path are compared with the path specification in the claim rule. If there is a match, the ESX host assigns the MPP specified in the claim rule to manage the physical path.

[1] EMC does not support using the MRU PSP for NMP with Symmetrix.

This assignment process continues until all physical paths are claimed by an MPP. Figure 34 has a sample claim rules list with only NMP installed.



```
Rule Class  Rule    Class    Type       Plugin      Matches
MP          0       runtime  transport  NMP         transport=usb
MP          1       runtime  transport  NMP         transport=sata
MP          2       runtime  transport  NMP         transport=ide
MP          3       runtime  transport  NMP         transport=block
MP          4       runtime  transport  NMP         transport=unknown
MP          65535   runtime  vendor     NMP         vendor=* model=*
```

Figure 34 Default claim rules with the native multipathing module

The rule set effectively claims all devices for NMP since no other options exist. If changes to the claim rules are needed after installation, devices can be manually unclaimed and the rules can be reloaded without a reboot as long as I/O is not running on the device (for instance, the device can contain a mounted VMFS volume but it cannot contain running virtual machines). It is a best practice to make claim rule changes after installation but before the immediate post-installation reboot. An administrator can choose to modify the claim rules, for instance, in order to have NMP manage EMC or non-EMC devices. It is important to note that after initial installation of a MPP, claim rules do not go into effect until after the vSphere host is rebooted. For instructions on changing claim rules, consult the *PowerPath/VE for VMware ESX and ESXi Installation and Administration Guide* at www.Powerlink.EMC.com or the *VMware vSphere SAN Configuration Guide* available from www.VMware.com.

Managing NMP in vSphere Client

Claim rules and claiming operations must all be done through the CLI, but the ability to choose the NMP multipathing policy can be performed in the vSphere Client itself. By default, Symmetrix devices being managed by NMP will be set to the policy of “Fixed” unless running vSphere 5.1 when they will be set to “Round Robin”. Using a policy of “Fixed” is not optimal and will not take advantage of the ability to have multiple, parallel paths to devices residing on the Symmetrix, leading to under-utilized resources.

Therefore, EMC strongly recommends setting the NMP policy of all Symmetrix devices to “Round Robin” to maximize throughput.[1]

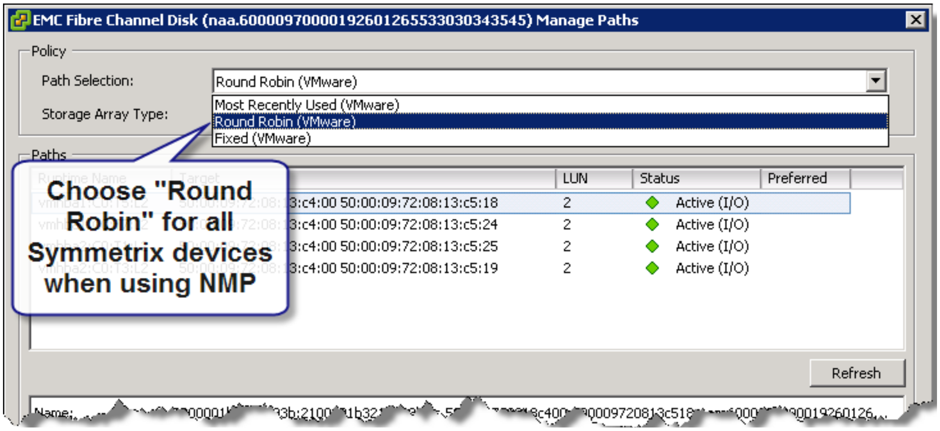
“Round Robin” uses an automatic path selection rotating through all available paths and enabling the distribution of the load across those paths. The simplest way to ensure that all Symmetrix devices are set to “Round Robin” is to change the default policy setting from the

"Fixed" policy to the "Round Robin" policy. This should be performed before presenting any Symmetrix devices to an ESX 4 or an ESXi 5.0 host to ensure that all of the devices use this policy from the beginning.

This can be executed through the use of the service console or through vSphere CLI (recommended) by issuing the command:

```
esxcli nmp satp setdefaultpsp -s VMW_SATP_SYMM  -P
   VMW_PSP_RR
```

From then on all Symmetrix devices will be, by default, set to "Round Robin." Alternatively, each device can be manually changed in the vSphere Client as in Figure 35.



EMC Fibre Channel Disk (naa.60000970000192601265533030343545) Manage Paths

Policy

Path Selection: Round Robin (VMware)

- Most Recently Used (VMware)
- Round Robin (VMware)
- Fixed (VMware)

Storage Array Type:

Paths

Runtime Name	Target	LUN	Status	Preferred
vmhba1:C0:T5:L2	50:00:09:72:08:13:c4:00 50:00:09:72:08:13:c5:18	2	Active (I/O)	
vmhb[illegible]	[illegible]:72:08:13:c4:00 50:00:09:72:08:13:c5:24	2	Active (I/O)	
[illegible]	[illegible]3:c4:00 50:00:09:72:08:13:c5:25	2	Active (I/O)	
vmhba2:C0:T3:L2	50:00:09:72:08:13:c4:00 50:00:09:72:08:13:c5:19	2	Active (I/O)	

Choose "Round Robin" for all Symmetrix devices when using NMP

Figure 35 NMP policy selection in vSphere Client

If using the vSphere Client, individual device assignment can be avoided by using the Path Management feature of EMC Virtual Storage Integrator. The Path Management feature will allow all paths for EMC devices to be changed at once.

[1] If the Symmetrix is running any release prior to Enginuity 5876.85.59, the EMC special devices known as gatekeepers require a "Fixed" policy unless PowerPath/VE is employed or shared storage clustering. See VMware KB article 1037959 for more detail.

This is accomplished by accessing the EMC context menu either at the host or cluster level as shown in Figure 36.

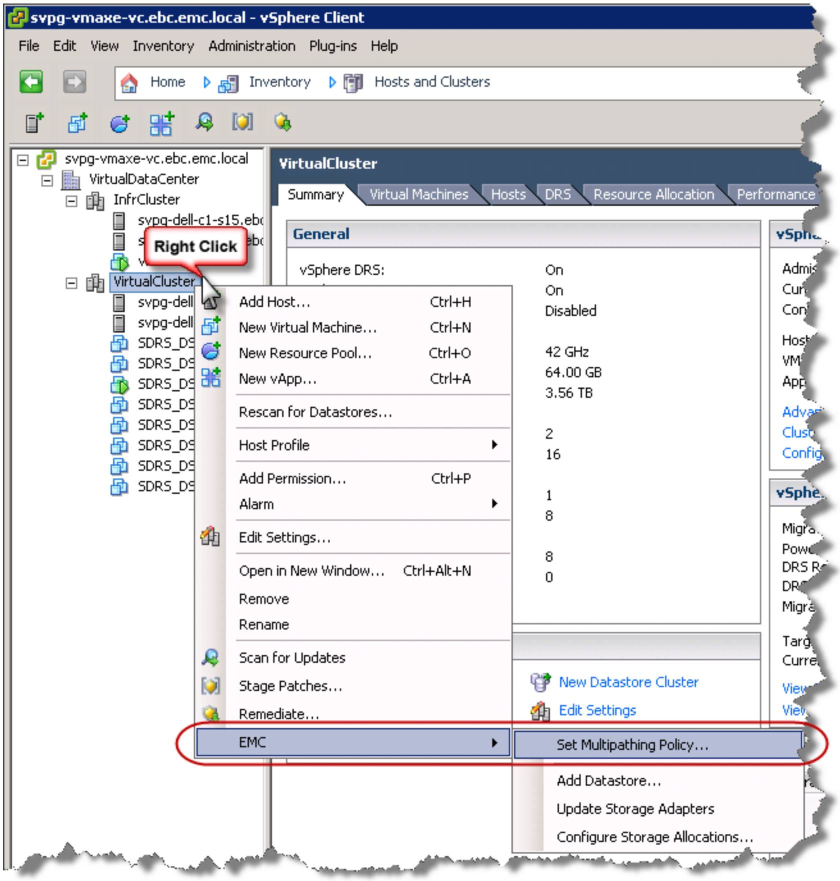
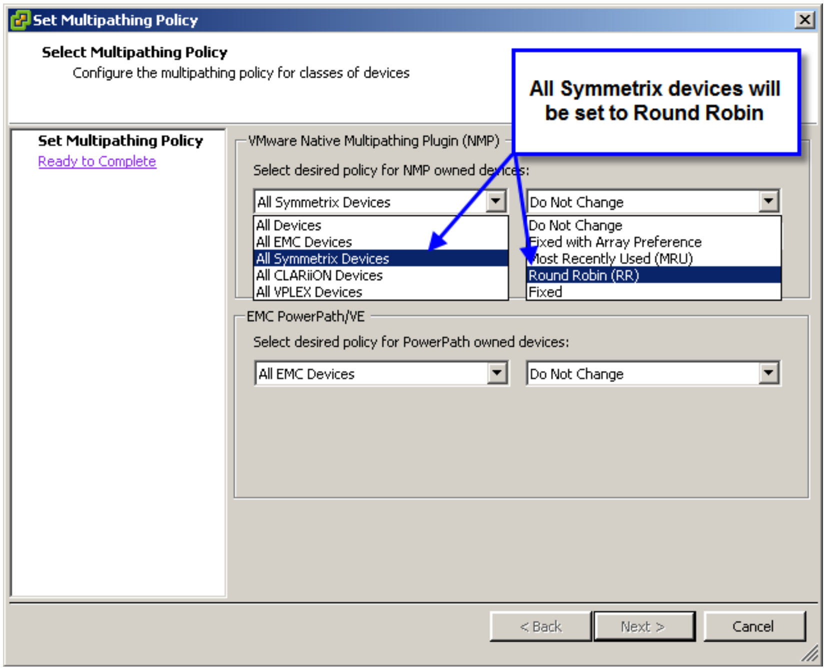


Figure 36 **Path Management feature of EMC Virtual Storage Integrator — Context Menu**

Once accessed, the Multipathing Policy dialog box is presented. In this screen, the user has the option to change the NMP or PowerPath/VE (if PowerPath/VE software and tools are installed) policy. Figure 37 is an example where the NMP policy is being set to Round Robin for all Symmetrix devices.



Set Multipathing Policy

Select Multipathing Policy

Configure the multipathing policy for classes of devices

All Symmetrix devices will be set to Round Robin

VMware Native Multipathing Plugin (NMP)

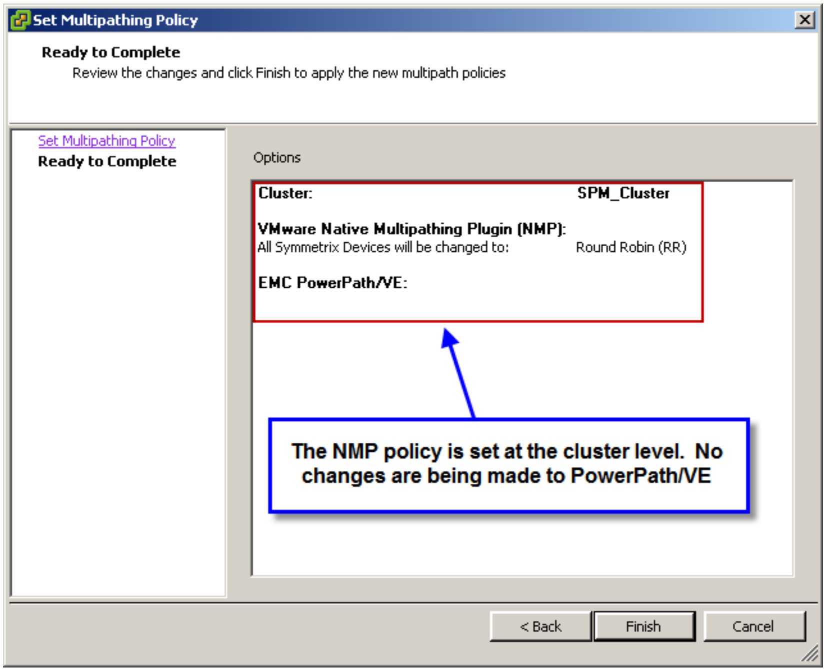
Select desired policy for NMP owned devices: All Symmetrix Devices — Round Robin (RR)

EMC PowerPath/VE

Select desired policy for PowerPath owned devices: All EMC Devices — Do Not Change

Figure 37 NMP multipathing policy for Symmetrix devices — Round Robin

Once selected, the user selects Next where the final screen appears as in Figure 38. It provides a summary of the action(s) to be taken before implementation.



Ready to Complete

Review the changes and click Finish to apply the new multipath policies

Options

Cluster: SPM_Cluster

VMware Native Multipathing Plugin (NMP):

All Symmetrix Devices will be changed to: Round Robin (RR)

EMC PowerPath/VE:

The NMP policy is set at the cluster level. No changes are being made to PowerPath/VE

Figure 38 NMP multipathing policy for Symmetrix devices set the cluster level

Note that the NMP and PowerPath/VE policies can be set differently for each type of EMC array as in Figure 39 on page 84, or set at the global level. For more information on using the Path Management feature of VSI, see "Managing PowerPath/VE in vCenter Server" on page 88.

NMP Round Robin and gatekeepers

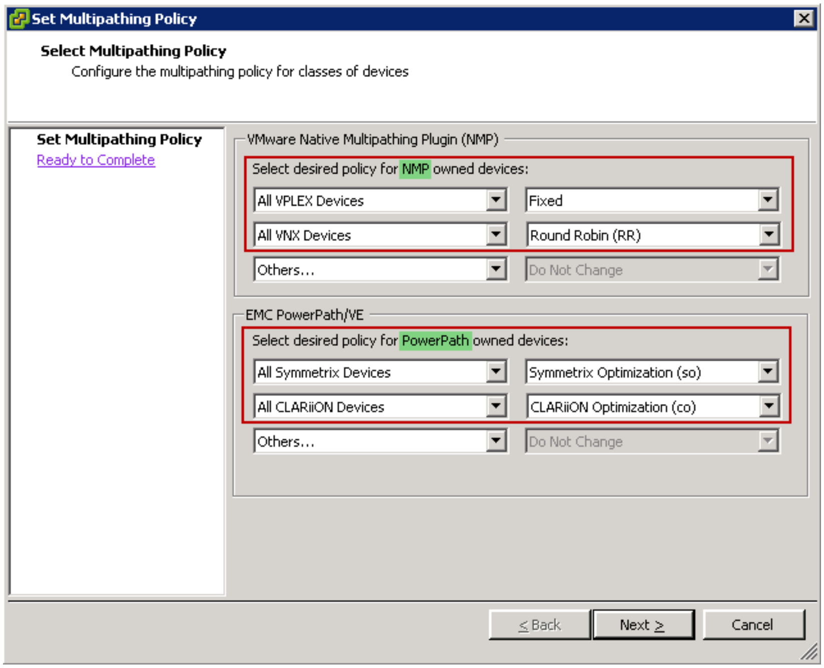
As the use of NMP Round Robin with gatekeepers is only supported with Enginuity 5876.85.59 or higher, one of the following three procedures should be used in conjunction with version 5.4 of the Path Management feature of VSI to address restrictions:

1. Customers running Enginuity 5875 or earlier with vSphere 5.1 should use the Path Management feature to change all Symmetrix paths to Fixed, followed by a second change setting all Symmetrix devices back to Round Robin. By default, vSphere 5.1 uses Round Robin for all Symmetrix devices, including gatekeepers, so

initially all devices will use Round Robin. The use of Round Robin with gatekeepers is not supported at Enginuity 5875 or earlier so the gatekeepers will need to be set to Fixed. Rather than do this manually, the Path Management feature is designed to filter out gatekeepers when setting Round Robin. So the logic is that if all devices are first set to Fixed, on the second change to Round Robin the gatekeepers will be skipped, and thus remain under a Fixed policy, while all the other devices are returned to Round Robin.

2. If you are running Enginuity 5876.85.59 or higher and vSphere 5.0.x or earlier, use the Path Management feature to change all paths to Round Robin since in vSphere 5.0.x and earlier the default NMP PSP is Fixed, but Round Robin is supported. This change to Round Robin will not include gatekeepers, however, as they will be skipped. If desired, each gatekeeper, therefore, will need to be manually changed to Round Robin. This is not required, however, as the support of Round Robin for gatekeepers was added for reasons of ease of use only.

3. If you are running 5876.85.59 or higher and vSphere 5.1 or later, Round Robin is the default NMP PSP for Symmetrix devices and is supported for gatekeepers so no remediation is necessary.



Set Multipathing Policy

Select Multipathing Policy

Configure the multipathing policy for classes of devices

Set Multipathing Policy

Ready to Complete

VMware Native Multipathing Plugin (NMP)

Select desired policy for NMP owned devices:

All VPLEX Devices — Fixed

All VNX Devices — Round Robin (RR)

Others... — Do Not Change

EMC PowerPath/VE

Select desired policy for PowerPath owned devices:

All Symmetrix Devices — Symmetrix Optimization (so)

All CLARiiON Devices — CLARiiON Optimization (co)

Others... — Do Not Change

Figure 39 **Setting multiple policies for different EMC devices**

NMP Round Robin and the I/O operation limit

The NMP Round Robin path selection policy has a parameter known as the "I/O operation limit" which controls the number of I/Os sent down each path before switching to the next path. The default value is 1000, therefore NMP defaults to switching from one path to another after sending 1000 I/Os down any given path. Tuning the Round Robin I/O operation limit parameter can significantly improve the performance of certain workloads, markedly so in sequential workloads. In case of environments which have random and OLTP type 2 workloads in their environments setting the Round Robin parameter to lower numbers still yields the best throughput; however lowering the value does not improve performance as significantly as it does for sequential workloads.[1]

[1] Overall performance difference between a value of 1000 and 1 is within 10%.

For these reasons, EMC recommends that the NMP Round Robin I/O operation limit parameter be set to 1. This ensures the best possible performance regardless of the workload being generated from the vSphere environment.

It is important to note that this setting is per device, it is not a global setting on an ESX/ESXi host. In addition setting it on one host will not propagate the change to other hosts in the cluster. Therefore it must be set for every Symmetrix device on every host. For large environments, it is strongly recommend that this process be scripted. The process to alter this parameter has changed slightly between ESX/ESXi 4 and ESXi 5; instructions for setting them both are shown below. Note that this parameter cannot be altered using the vSphere Client as this is a CLI-only operation.[1]

For ESX/ESXi 4:

- To check the IO Operations limit:

```
esxcli nmp roundrobin getconfig --device=<device
NAA>
```

- To set the IO Operations limit:

```
esxcli nmp roundrobin setconfig --device=<device
NAA> --iops 1 --type iops
```

For ESXi 5:

- To check the IO Operations limit:

```
esxcli storage nmp psp roundrobin deviceconfig get
--device=<device NAA>
```

- To set the IO Operations limit:

```
esxcli storage nmp psp roundrobin deviceconfig set
--device=<device NAA> --iops=1 --type iops
```

PowerPath/VE Multipathing Plug-in and management

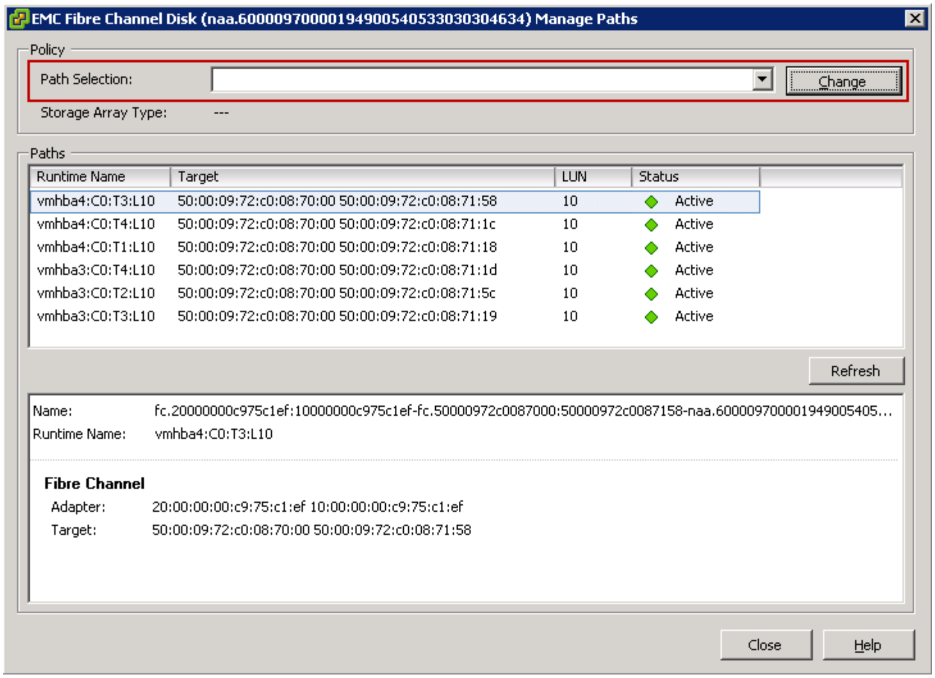
EMC PowerPath/VE delivers PowerPath multipathing features to optimize VMware vSphere environments. PowerPath/VE uses a command set, called rpowermt, to monitor, manage, and configure PowerPath/VE for vSphere.

[1] For more information, reference the *Tuning the VMware Native Multipathing Performance of VMware vSphere Hosts Connected to EMC Symmetrix Storage White Paper*, which can be found at http://www.emc.com and at www.Powerlink.EMC.com.

The syntax, arguments, and options are very similar to the traditional powermt commands used on all other PowerPath multipathing supported operating system platforms. There is one significant difference in that rpowermt is a remote management tool.

While ESX 4 has a service console interface, ESXi 4 and 5 do not.[1] In order to manage an ESXi host, customers have the option to use vCenter Server or vCLI (also referred to as VMware Remote Tools) on a remote server. PowerPath/VE for vSphere uses the rpowermt command line utility for both ESX and ESXi.

PowerPath/VE for vSphere cannot be managed on the ESX host itself. There is also no option for choosing PowerPath within the "Manage Paths" dialog within the vSphere Client as seen in Figure 35 on page 79. Similarly, if a device is currently managed by PowerPath, no policy options are available for selection as seen in Figure 40.



EMC Fibre Channel Disk (naa.60000970000194900540533030304634) Manage Paths

Policy

Path Selection:

Storage Array Type: ---

Paths

Runtime Name	Target	LUN	Status
vmhba4:C0:T3:L10	50:00:09:72:c0:08:70:00 50:00:09:72:c0:08:71:58	10	● Active
vmhba4:C0:T4:L10	50:00:09:72:c0:08:70:00 50:00:09:72:c0:08:71:1c	10	● Active
vmhba4:C0:T1:L10	50:00:09:72:c0:08:70:00 50:00:09:72:c0:08:71:18	10	● Active
vmhba3:C0:T4:L10	50:00:09:72:c0:08:70:00 50:00:09:72:c0:08:71:1d	10	● Active
vmhba3:C0:T2:L10	50:00:09:72:c0:08:70:00 50:00:09:72:c0:08:71:5c	10	● Active
vmhba3:C0:T3:L10	50:00:09:72:c0:08:70:00 50:00:09:72:c0:08:71:19	10	● Active

Name: fc.20000000c975c1ef:10000000c975c1ef-fc.50000972c0087000:50000972c0087158-naa.600009700001949005405...
Runtime Name: vmhba4:C0:T3:L10

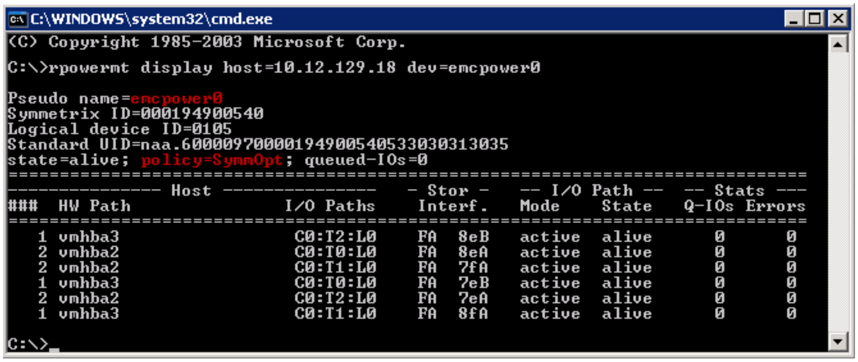
Fibre Channel
Adapter: 20:00:00:00:c9:75:c1:ef 10:00:00:00:c9:75:c1:ef
Target: 50:00:09:72:c0:08:70:00 50:00:09:72:c0:08:71:58

Figure 40 Manage Paths dialog viewing a device under PowerPath ownership

[1] There is a limited CLI available on ESXi if Tech Support mode is enabled.

There is neither a local nor remote GUI for PowerPath on ESX with the exception of the Path Management tools for VSI previously discussed. For functionality not present in VSI, administrators must designate a virtual or a physical machine to manage one or multiple ESX hosts.

When the ESX host is connected to a Symmetrix array, the PowerPath/VE kernel module running on the vSphere host will associate all paths to each device presented from the array and associate a pseudo device name. An example of this is shown in Figure 41, which shows the output of rpowermt display host=x.x.x.x dev=emcpower0. Note in the output that the device has six paths and displays the recommended optimization mode (SymmOpt = Symmetrix optimization).



```
(C) Copyright 1985-2003 Microsoft Corp.

C:\>rpowermt display host=10.12.129.18 dev=emcpower0

Pseudo name=emcpower0
Symmetrix ID=000194900540
Logical device ID=0105
Standard UID=naa.60000970000194900540533030313035
state=alive; policy=SymmOpt; queued-IOs=0
==============================================================================
--------------- Host ---------------   - Stor -   -- I/O Path --  -- Stats ---
###  HW Path               I/O Paths    Interf.   Mode    State   Q-IOs Errors
==============================================================================
   1 vmhba3                 C0:T2:L0    FA  8eB   active  alive      0      0
   2 vmhba2                 C0:T0:L0    FA  8eA   active  alive      0      0
   2 vmhba2                 C0:T1:L0    FA  7fA   active  alive      0      0
   1 vmhba3                 C0:T0:L0    FA  7eB   active  alive      0      0
   2 vmhba2                 C0:T2:L0    FA  7eA   active  alive      0      0
   1 vmhba3                 C0:T1:L0    FA  8fA   active  alive      0      0

C:\>_
```

Figure 41 Output of the rpowermt display command on a Symmetrix VMAX device

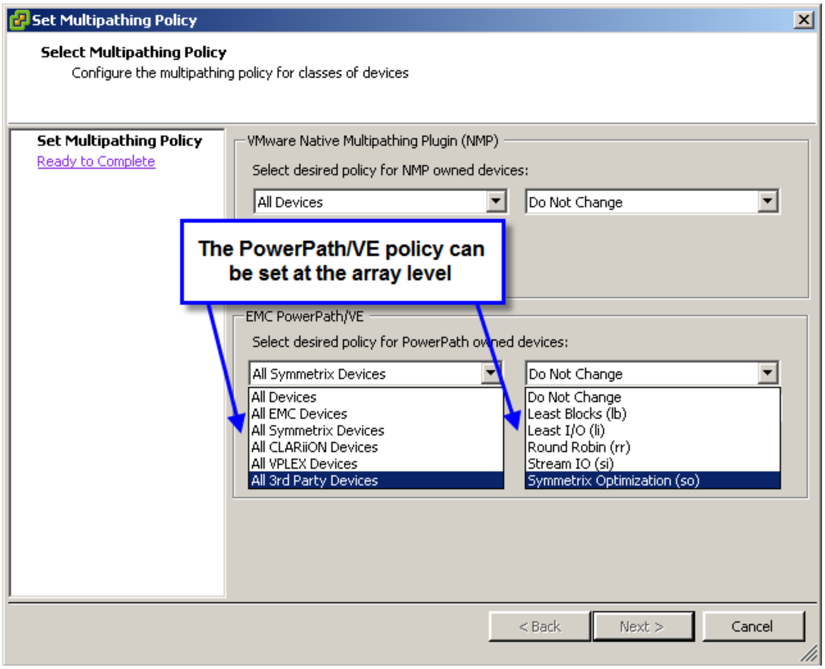
For more information on the rpowermt commands and output, consult the *PowerPath/VE for VMware vSphere Installation and Administration Guide*.

As more Symmetrix directors become available, the connectivity can be scaled as needed. PowerPath/VE supports up to 32 paths to a device[1]. These methodologies for connectivity ensure all front-end directors and processors are utilized, providing maximum potential performance and load balancing for vSphere hosts connected to the Symmetrix storage arrays in combination with PowerPath/VE.

[1] ESX supports 1024 total combined paths to all devices.

Managing PowerPath/VE in vCenter Server

PowerPath/VE for vSphere is managed, monitored, and configured using rpowermt as discussed in the previous section. This CLI-based management is common across all PowerPath platforms, but there is also integration within vCenter through the Path Management feature of EMC Virtual Storage Integrator as mentioned in "Managing NMP in vSphere Client" on page 78. Through VSI, the PowerPath/VE policy can be set at the host or cluster level for EMC devices. The policy can vary for each EMC array type if desired as displayed in Figure 42.



Set Multipathing Policy

Select Multipathing Policy

Configure the multipathing policy for classes of devices

VMware Native Multipathing Plugin (NMP)

Select desired policy for NMP owned devices: All Devices | Do Not Change

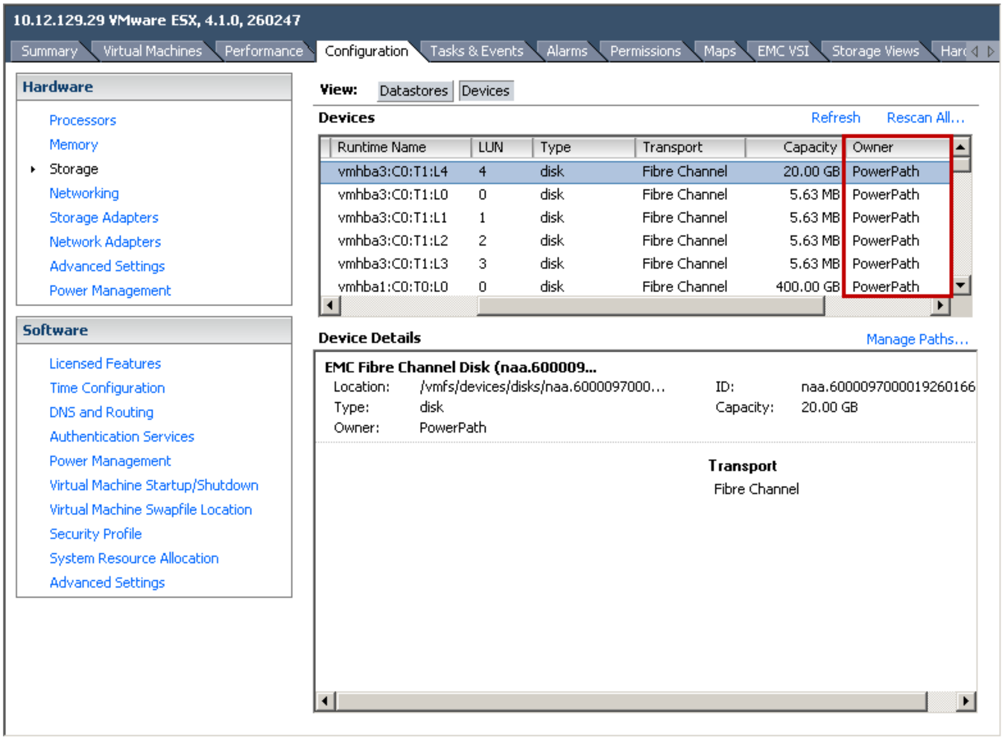
EMC PowerPath/VE

Select desired policy for PowerPath owned devices: All Symmetrix Devices | Do Not Change

The PowerPath/VE policy can be set at the array level

Figure 42 Setting policy for PowerPath/VE in the Path Management feature of VSI

In the vSphere Client one can also see LUN ownership which will display "PowerPath" if it is the owner of the device. An example of this is seen in Figure 43 on page 89, under the Configuration tab of the host and within the Storage Devices list, the owner of the device is shown as being PowerPath.



Devices

Runtime Name	LUN	Type	Transport	Capacity	Owner
vmhba3:C0:T1:L4	4	disk	Fibre Channel	20.00 GB	PowerPath
vmhba3:C0:T1:L0	0	disk	Fibre Channel	5.63 MB	PowerPath
vmhba3:C0:T1:L1	1	disk	Fibre Channel	5.63 MB	PowerPath
vmhba3:C0:T1:L2	2	disk	Fibre Channel	5.63 MB	PowerPath
vmhba3:C0:T1:L3	3	disk	Fibre Channel	5.63 MB	PowerPath
vmhba1:C0:T0:L0	0	disk	Fibre Channel	400.00 GB	PowerPath

EMC Fibre Channel Disk (naa.600009...

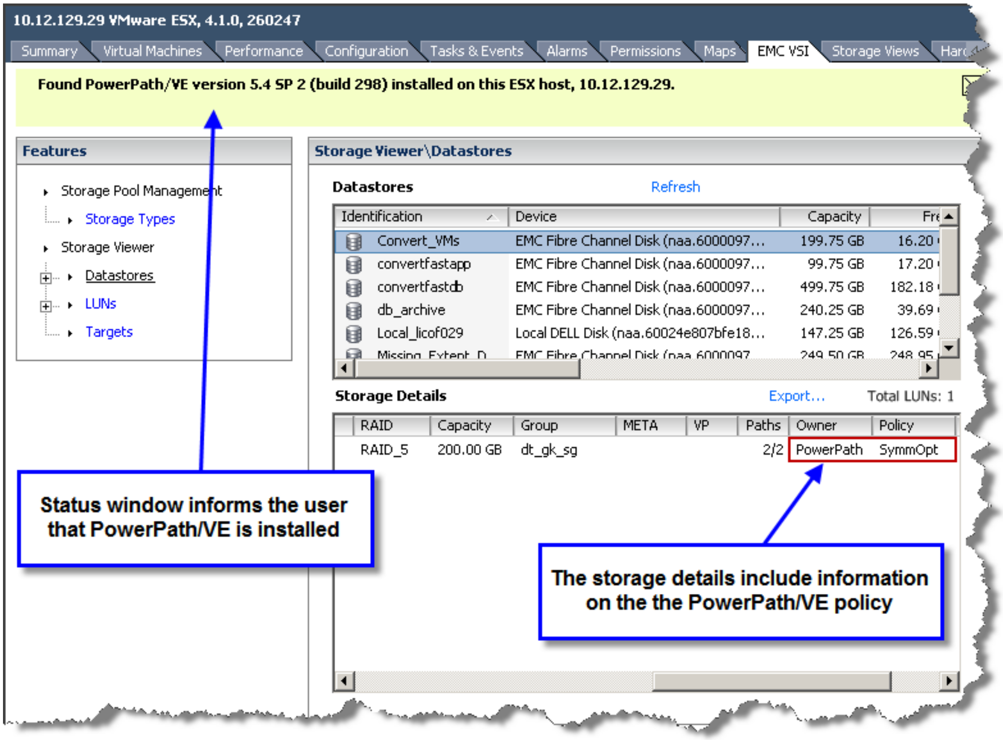
Location: /vmfs/devices/disks/naa.6000097000... ID: naa.6000097000019260166
Type: disk Capacity: 20.00 GB
Owner: PowerPath

Transport
Fibre Channel

Figure 43 Viewing the multipathing plug-in owning a device in vSphere Client

The Storage Viewer feature of VSI can also be used to determine if a device is under PowerPath ownership. By highlighting a particular device, it can be seen in Figure 44 on page 90 that not only the multipath plug-in owner is listed, but also the path management policy that is in effect. So in this particular case, one sees that the device backing the datastore Convert_VMs is managed by the PowerPath policy of SymmOpt.

One of the other features of Storage Viewer is shown at the top of Figure 44. This is a status box that alerts users to important information about the environment. One of these alerts is whether or not PowerPath is installed on the host being accessed. In this screenshot one sees the message indicating it is installed. If PowerPath were installed but not licensed, this would also be indicated in this box.



10.12.129.29 VMware ESX, 4.1.0, 260247

Found PowerPath/VE version 5.4 SP 2 (build 298) installed on this ESX host, 10.12.129.29.

Identification	Device	Capacity	Fre
Convert_VMs	EMC Fibre Channel Disk (naa.6000097…	199.75 GB	16.20
convertfastapp	EMC Fibre Channel Disk (naa.6000097…	99.75 GB	17.20
convertfastdb	EMC Fibre Channel Disk (naa.6000097…	499.75 GB	182.18
db_archive	EMC Fibre Channel Disk (naa.6000097…	240.25 GB	39.69
Local_licof029	Local DELL Disk (naa.60024e807bfe18…	147.25 GB	126.59
Missing_Extent_D	EMC Fibre Channel Disk (naa.6000097	249.50 GB	248.95

RAID	Capacity	Group	META	VP	Paths	Owner	Policy
RAID_5	200.00 GB	dt_gk_sg			2/2	PowerPath	SymmOpt

Status window informs the user that PowerPath/VE is installed

The storage details include information on the the PowerPath/VE policy

Figure 44 **PowerPath ownership displayed in the Storage Viewer feature of VSI**

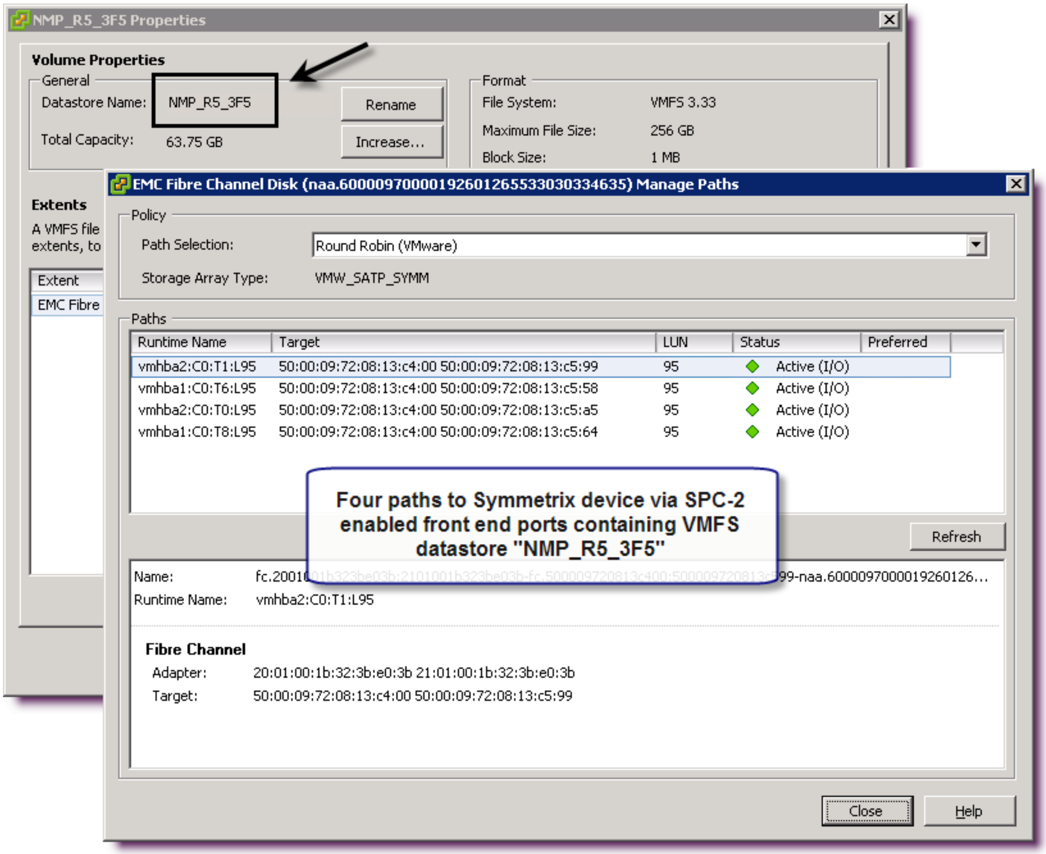
Improving resiliency of VMware ESX 4 and ESXi 5 to SAN failures

VMware ESX uses modified QLogic and Emulex drivers that support multiple targets on every initiator. This functionality can be used to provide greater resiliency in a VMware ESX environment by reducing the impact of storage port failures. Furthermore, presenting a device on multiple paths allows for better load balancing and reduced sensitivity to storage port queuing. This is extremely important in environments that share EMC Symmetrix storage array ports between VMware ESX hosts and other operating systems.

Figure 45 on page 92 shows how to implement multiple-target zoning in a VMware ESX environment. It can be seen that the VMware ESX has two HBAs, vmhba1 and vmhba2. However, the device hosting the datastore NMP_R5_3F5 can be accessed from four different paths with the following runtime names: vmhba1:C0:T6:L95, vmhba1:C0:T8:L95, vmhba2:C0:T0:L95 and vmhba2:C0:T1:L95. This is achieved by zoning each HBA in the VMware ESX to two EMC Symmetrix storage array ports.

The VMkernel assigns a unique SCSI target number to each storage array port. In Figure 45 on page 92, the VMkernel has assigned SCSI target numbers 6 and 8 to each EMC Symmetrix storage array port into which vmhba1 logs in, and 0 and 1 to each EMC Symmetrix storage array port into which vmhba2 logs in.

Note: This section does not discuss how multiple target zoning can be implemented in the fabric. EMC always recommends one initiator and one target in each zone. Multiple target zoning discussed in this section can be implemented as a collection of zones in which each zone contains the same initiator but different targets.



NMP_R5_3F5 Properties

Volume Properties

General

Datastore Name: NMP_R5_3F5

Total Capacity: 63.75 GB

Rename

Increase...

Format

File System: VMFS 3.33

Maximum File Size: 256 GB

Block Size: 1 MB

Extents

A VMFS file
extents, to

Extent

EMC Fibre

EMC Fibre Channel Disk (naa.60000970000192601265533030334635) Manage Paths

Policy

Path Selection: Round Robin (VMware)

Storage Array Type: VMW_SATP_SYMM

Paths

Runtime Name	Target	LUN	Status	Preferred
vmhba2:C0:T1:L95	50:00:09:72:08:13:c4:00 50:00:09:72:08:13:c5:99	95	◆ Active (I/O)	
vmhba1:C0:T6:L95	50:00:09:72:08:13:c4:00 50:00:09:72:08:13:c5:58	95	◆ Active (I/O)	
vmhba2:C0:T0:L95	50:00:09:72:08:13:c4:00 50:00:09:72:08:13:c5:a5	95	◆ Active (I/O)	
vmhba1:C0:T8:L95	50:00:09:72:08:13:c4:00 50:00:09:72:08:13:c5:64	95	◆ Active (I/O)	

Four paths to Symmetrix device via SPC-2 enabled front end ports containing VMFS datastore "NMP_R5_3F5"

Refresh

Name: fc.2001001b323be03b:2101001b323be03b-fc.500009720813c400:500009720813c599-naa.6000097000019260126...

Runtime Name: vmhba2:C0:T1:L95

Fibre Channel

Adapter: 20:01:00:1b:32:3b:e0:3b 21:01:00:1b:32:3b:e0:3b

Target: 50:00:09:72:08:13:c4:00 50:00:09:72:08:13:c5:99

Close

Help

Figure 45 Increasing resiliency of ESX host to SAN failures

All Paths Down (APD) condition

All paths down or APD, occurs on an ESX host when a storage device is removed in an uncontrolled manner from the host (or the device fails), and the VMkernel core storage stack does not know how long the loss of device access will last. A typical way of getting into APD would be a Fiber Channel switch failure or the removal of the zoning for that host. APD results in a hung host and the need for a hard reset.

Note: Redundancy of your multi-pathing solution, such as PowerPath/VE, can help avoid some situations of APD.

APD in vSphere 4

To alleviate some of the issues arising from APD, a number of advanced settings were added which could be changed by the VMware administrator to mitigate the problem. One of these settings changed the behavior of hostd when it was scanning the SAN, so that it would no longer be blocked, even when it came across a non-responding device (i.e. a device in APD state). This setting was automatically added in ESX 4.0 Update 3 and ESX 4.1 Update 1. Previous versions required customers to manually add the setting.[1]

APD in vSphere 5.0 — Permanent Device Loss (PDL)

In vSphere 5.0, there are changes to the way VMware handles APD. Most importantly, VMware now differentiates between a device to which connectivity is permanently lost, and a device to which connectivity is transiently or temporarily lost. A device which is never coming back is known as a Permanent Device Loss (PDL). The difference between APD and PDL is the following:

- APD is now considered a transient condition, i.e. the device may come back at some time in the future.
- PDL is considered a permanent condition where the device is never coming back.

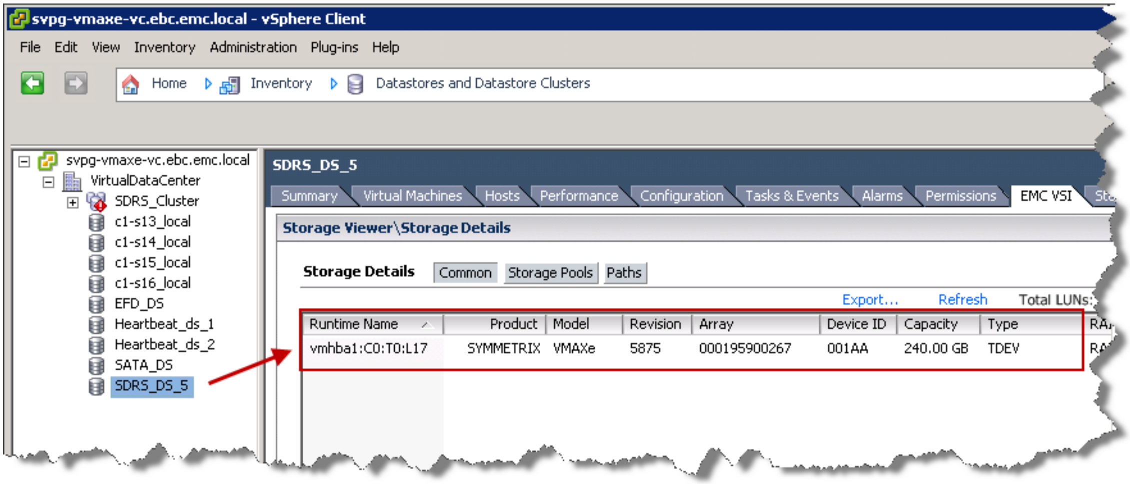
[1] For a more detailed explanation of these settings, please see VMware KB article 1016626.

Avoiding APD/PDL — Device removal

Prior to vSphere 5, there was no clearly defined way to remove a storage device from ESX. With vSphere 5 and within the vSphere Client there are two new storage device management techniques that aid in the removal of the device: mount/unmount a VMFS-3 or VMFS-5 volume and attach/detach a device. In order to remove a datastore and Symmetrix device completely, avoiding APD/PDL, the user needs to first unmount the volume and then detach the device.

Unmounting

In the following example, the datastore SDRS_DS_5 seen in Figure 46, will be unmounted and the device associated with it detached in the vSphere Client. Before proceeding, make sure that there are no VMs on the datastore, that the datastore is not part of a datastore cluster (SDRS), is not used by vSphere HA as a heartbeat datastore, and does not have Storage I/O Control enabled. All these must be rectified before proceeding.

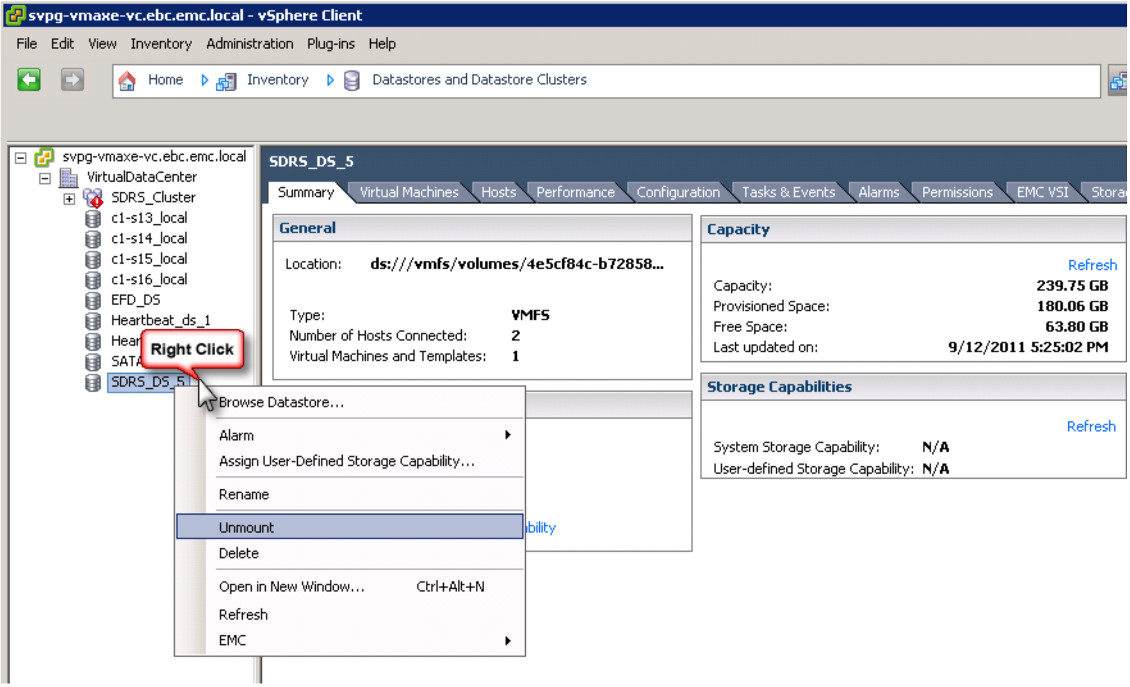


Runtime Name	Product	Model	Revision	Array	Device ID	Capacity	Type	RA
vmhba1:C0:T0:L17	SYMMETRIX	VMAXe	5875	000195900267	001AA	240.00 GB	TDEV	RA

Figure 46 Datastore SDRS_DS_5

SDRS_DS_5 meets all these prerequisites so it can be unmounted. Note the Runtime Name, vmhba1:C0:T0:L17, as it will be needed when the device is detached.

Once the datastore is identified, right-click on the datastore and choose “Unmount” as shown in Figure 47. The unmount may also be accomplished through the CLI. The command to do an unmount is `esxcli storage filesystem unmount` if you prefer to do it from the ESXi shell.



General

Location: ds:///vmfs/volumes/4e5cf84c-b72858...

Type: VMFS
Number of Hosts Connected: 2
Virtual Machines and Templates: 1

Capacity

Capacity: 239.75 GB
Provisioned Space: 180.06 GB
Free Space: 63.80 GB
Last updated on: 9/12/2011 5:25:02 PM

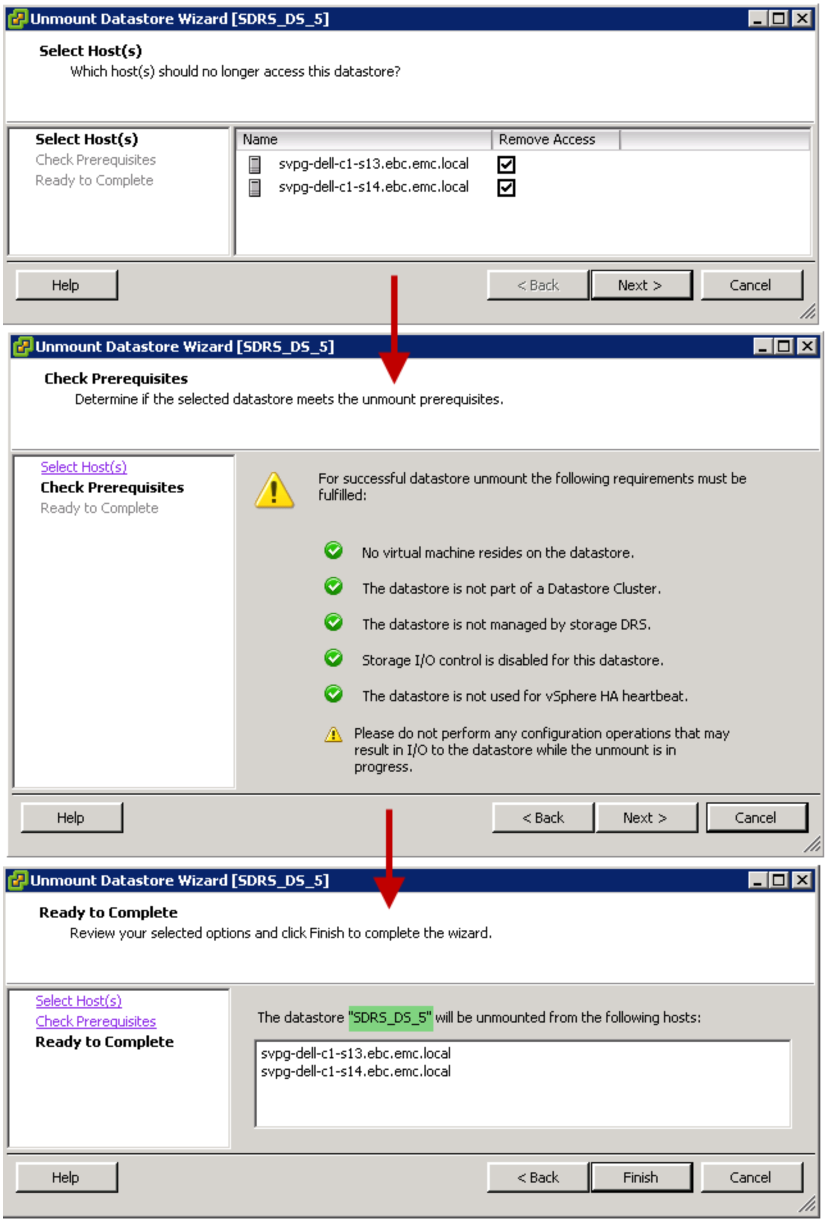
Storage Capabilities

System Storage Capability: N/A
User-defined Storage Capability: N/A

Figure 47 Unmount datastore SDRS_DS_5

A number of screens follow with the unmount wizard, which are ordered together in Figure 48 on page 96:

1. The first screen presents the hosts that have the volume mounted. Check the boxes of the hosts to unmount the volume from all listed hosts.
2. The wizard will then ensure all prerequisites are met before allowing the user to proceed, presenting the results for verification.
3. The final screen is a summary screen where the user finishes the task.



Unmount Datastore Wizard [SDRS_DS_5]

Select Host(s)

Which host(s) should no longer access this datastore?

Select Host(s)
Check Prerequisites
Ready to Complete

Name	Remove Access
svpg-dell-c1-s13.ebc.emc.local	☑
svpg-dell-c1-s14.ebc.emc.local	☑

Unmount Datastore Wizard [SDRS_DS_5]

Check Prerequisites

Determine if the selected datastore meets the unmount prerequisites.

Select Host(s)
Check Prerequisites
Ready to Complete

For successful datastore unmount the following requirements must be fulfilled:

- ✓ No virtual machine resides on the datastore.
- ✓ The datastore is not part of a Datastore Cluster.
- ✓ The datastore is not managed by storage DRS.
- ✓ Storage I/O control is disabled for this datastore.
- ✓ The datastore is not used for vSphere HA heartbeat.
- ⚠ Please do not perform any configuration operations that may result in I/O to the datastore while the unmount is in progress.

Unmount Datastore Wizard [SDRS_DS_5]

Ready to Complete

Review your selected options and click Finish to complete the wizard.

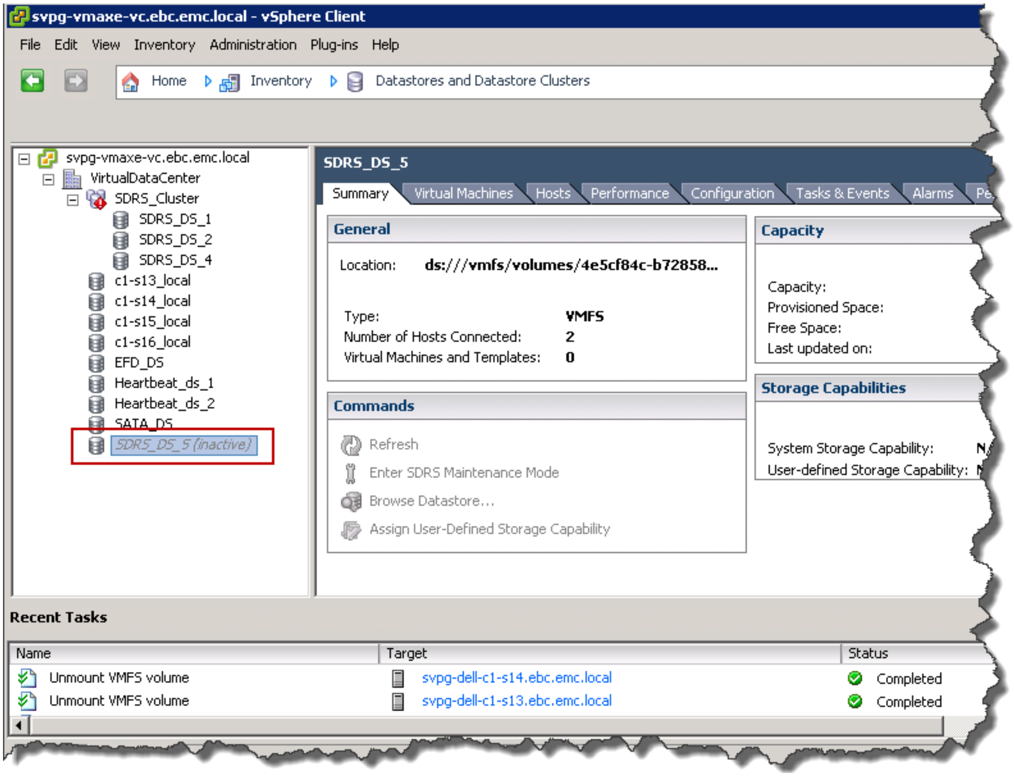
Select Host(s)
Check Prerequisites
Ready to Complete

The datastore "SDRS_DS_5" will be unmounted from the following hosts:

svpg-dell-c1-s13.ebc.emc.local
svpg-dell-c1-s14.ebc.emc.local

Figure 48 **The unmount wizard**

Once the wizard is complete, the result will be an inactive datastore which is grayed out. In Figure 49, the two unmount tasks are shown as completed in the bottom part of the image and the inactive datastore is boxed and highlighted.



General

Location: ds:///vmfs/volumes/4e5cf84c-b72858...

Type: VMFS
Number of Hosts Connected: 2
Virtual Machines and Templates: 0

Capacity

Capacity:
Provisioned Space:
Free Space:
Last updated on:

Storage Capabilities

System Storage Capability: N
User-defined Storage Capability: N

Recent Tasks

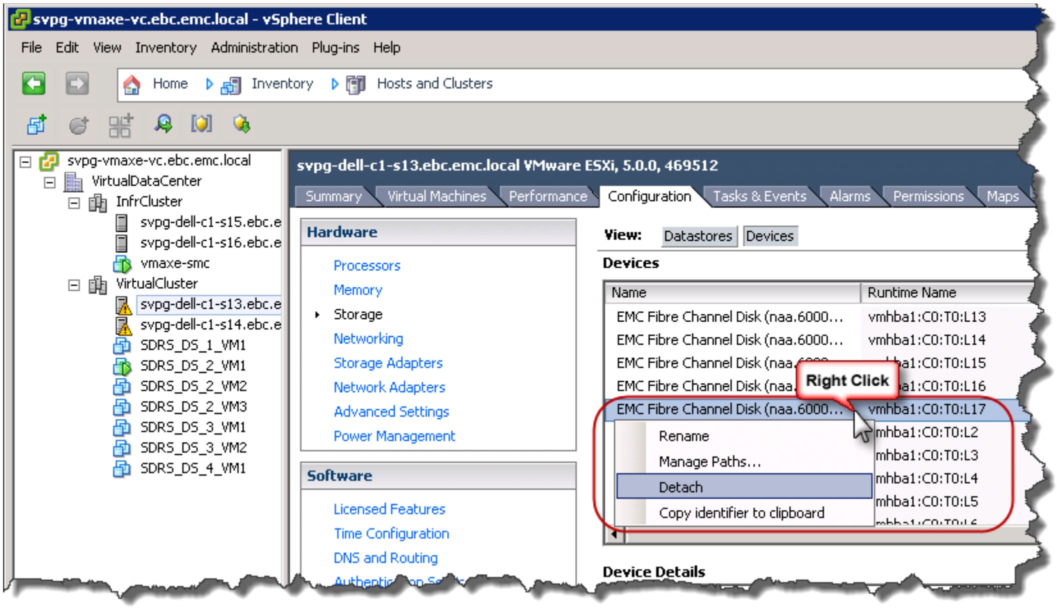
Name	Target	Status
Unmount VMFS volume	svpg-dell-c1-s14.ebc.emc.local	Completed
Unmount VMFS volume	svpg-dell-c1-s13.ebc.emc.local	Completed

Figure 49 Datastore unmount completion

Detaching

The second part of the process to remove the devices involves detaching them in the vSphere Client. Start by recalling the runtime name gathered before the unmount (vmhba1:C0:T0:L17).

Once the device is identified in the Hardware/Storage option of the Configuration tab of one of the hosts, right-click on it and select detach as in Figure 50.

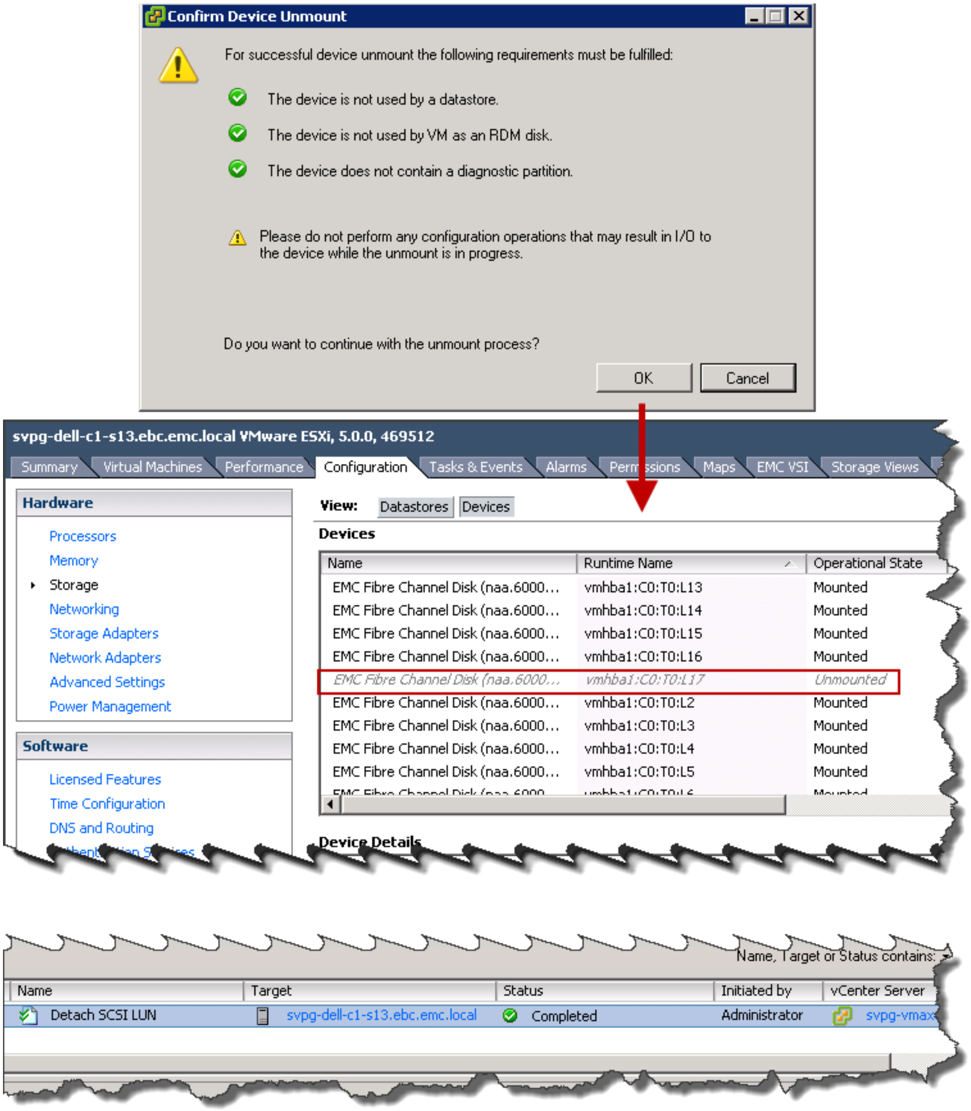


Name	Runtime Name
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L13
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L14
EMC Fibre Channel Disk (naa.6000	a1:C0:T0:L15
EMC Fibre Channel Disk (naa.	a1:C0:T0:L16
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L17
	mhba1:C0:T0:L2
	mhba1:C0:T0:L3
	mhba1:C0:T0:L4
	mhba1:C0:T0:L5

Figure 50 Detaching a device after unmounting the volume

As in unmounting the volume, the wizard, seen in Figure 51 on page 99, will run prerequisites before allowing the detach of the device. Once detached, run a rescan to remove the inactive datastore from the list. The grayed out device, however, will remain in the device list until the LUN is unmasked from the host and a second rescan is run. If you prefer to use CLI, the detach can be done through the ESXi shell with the command `esxcli storage core device set --state=off`.

Note: Unlike the unmounting of a volume, detaching the underlying device must be done on each host of a cluster independently. The wizard has no ability to check other hosts in a cluster.



For successful device unmount the following requirements must be fulfilled:

✓ The device is not used by a datastore.

✓ The device is not used by VM as an RDM disk.

✓ The device does not contain a diagnostic partition.

Please do not perform any configuration operations that may result in I/O to the device while the unmount is in progress.

Do you want to continue with the unmount process?

Devices

Name	Runtime Name	Operational State
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L13	Mounted
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L14	Mounted
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L15	Mounted
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L16	Mounted
EMC Fibre Channel Disk (naa.6000...	*vmhba1:C0:T0:L17*	*Unmounted*
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L2	Mounted
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L3	Mounted
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L4	Mounted
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L5	Mounted
EMC Fibre Channel Disk (naa.6000...	vmhba1:C0:T0:L6	Mounted

Name, Target or Status contains:

Name	Target	Status	Initiated by	vCenter Server
Detach SCSI LUN	svpg-dell-c1-s13.ebc.emc.local	✓ Completed	Administrator	svpg-vmax

Figure 51 Completing the detach wizard

2

Management of EMC Symmetrix Arrays

This chapter discusses the various ways to manage a VMware environment using EMC technologies:

Introduction

Managing and monitoring the Symmetrix storage system that provides a VMware environment with persistent medium to store virtual machine data is a critical task for any VMware system administrator. EMC has numerous options to offer the administrator from virtual appliances running on vSphere to specific Symmetrix system management context commands created just for VMware.

EMC offers two primary complementary methods of managing and monitoring Symmetrix storage arrays:

- Command line offering Solutions Enabler SYMCLI
- The graphical user interface of EMC Unisphere for VMAX which contains both storage management and performance monitoring capabilities.[1]

The following sections will describe the best ways to deploy and use these management applications in a VMware environment.

[1] In version 1.x of Unisphere for VMAX there are some tasks (primarily related to CKD devices and SRDF® Star) that were not ported from Symmetrix Management Console (SMC). If those tasks are critical to the customer environment, they are advised to use SMC, either independently or jointly with Unisphere for VMAX. For details surrounding those tasks, please see the EMC Unisphere for VMAX Release Notes on Powerlink at www.Powerlink.EMC.com.

Solutions Enabler in VMware environments

Solutions Enabler is a specialized library consisting of commands that can be invoked on the command line, or within scripts. These commands can be used to monitor device configuration and status, and perform control operations on devices and data objects within managed storage arrays. SYMCLI resides on a host system to monitor and perform control operations on Symmetrix arrays. SYMCLI commands are invoked from the host operating system command line. The SYMCLI commands are built on top of SYMAPI library functions, which use system calls that generate low-level SCSI commands to specialized devices on the Symmetrix called gatekeepers. Gatekeepers are very small devices carved from disks in the Symmetrix that act as SCSI targets for the SYMCLI commands. To reduce the number of inquiries from the host to the storage arrays, configuration and status information is maintained in a Symmetrix host database file, symapi_db.bin by default. It is known as the Symmetrix configuration database.

Thin gatekeepers and multipathing

Beginning with Enginuity 5876 and Solutions Enabler V7.4, thin devices (TDEV) may be used as gatekeepers. In addition, these thin devices need not be bound to thin pools.

New to Enginuity 5876 and Solutions Enabler V7.4 is support for Native Multipathing (NMP) with gatekeepers. This means that the NMP policy of Round Robin in vSphere 5.x is supported for gatekeeper devices so long as the Enginuity code on the Symmetrix is 5876 or higher. Customers running Enginuity 5875 or lower must still use Fixed Path policy for gatekeepers on any vSphere version.

Prior to Enginuity 5876, PowerPath/VE was the only supported multipathing software with gatekeepers, though it still remains the preferred multipathing software for VMware on Symmetrix.

Note: Gatekeepers may not be shared among multiple virtual machines, regardless of the multipathing software.

Solutions Enabler virtual appliance

Solutions Enabler version 7.1 and higher includes a virtual appliance (SE vApp) that can be deployed as a virtual machine in a VMware vSphere environment. The appliance is preconfigured to provide Solutions Enabler SYMAPI services that are required by EMC applications like VSI or the EMC Symmetrix Remote Data Facility (SRDF) SRA for Site Recovery Manager. It encompasses all the components you need to manage your Symmetrix environment using the **storsrvd** daemon and Solutions Enabler network client access. These include:

- EMC Solutions Enabler (solely intended as a SYMAPI server for Solutions Enabler client access)
- Linux OS (SUSE 11 SP1 JeOS)
- SMI- S Provider

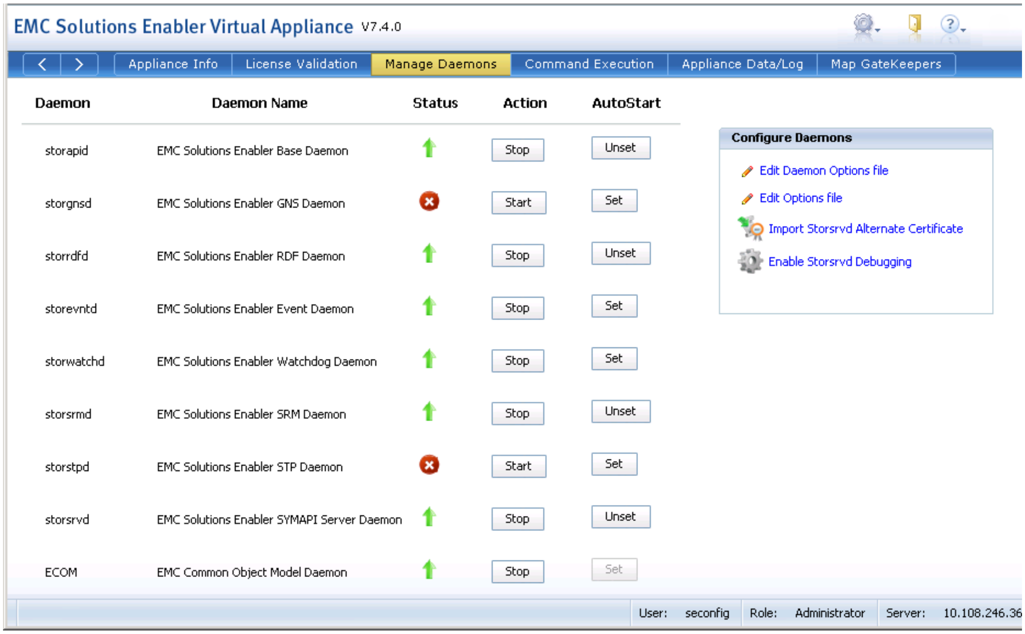
Note: Version 7.4 of the SE vApp includes VASA support.

In addition, the Solutions Enabler virtual appliance includes a browser-based configuration tool, called the Solutions Enabler Virtual Appliance Configuration Manager. This is seen in Figure 52 on page 105. This tool enables you to perform the following configuration tasks:

- Monitor the application status
- Start and stop selected daemons
- Import and export persistent data
- Configure the nethost file (required for client access)
- Discover storage arrays
- Modify options and daemon options
- Add Symmetrix-based and host-based license keys
- Run a limited set of Solutions Enabler CLI commands
- Configure ESX host and gatekeeper devices
- Launch Unisphere for VMAX (available only in Unisphere versions of the appliance console)
- Configure iSCSI initiator and map iSCSI gatekeeper devices
- Configure additional NIC card (optional)

- Download SYMAPI debug logs
- Import CA signed certificate for web browser
- Import Custom certificate for storsrvd daemon
- Check disk usage
- Restart appliance
- Configure symavoid entries
- Load Symmetrix-based eLicenses

Note: Root login is not supported on the virtual appliance.



Daemon	Daemon Name	Status	Action	AutoStart
storapid	EMC Solutions Enabler Base Daemon	↑	Stop	Unset
storgnsd	EMC Solutions Enabler GNS Daemon	✗	Start	Set
storrdfd	EMC Solutions Enabler RDF Daemon	↑	Stop	Unset
storevntd	EMC Solutions Enabler Event Daemon	↑	Stop	Set
storwatchd	EMC Solutions Enabler Watchdog Daemon	↑	Stop	Set
storsrmd	EMC Solutions Enabler SRM Daemon	↑	Stop	Unset
storstpd	EMC Solutions Enabler STP Daemon	✗	Start	Set
storsrvd	EMC Solutions Enabler SYMAPI Server Daemon	↑	Stop	Unset
ECOM	EMC Common Object Model Daemon	↑	Stop	Set

User: seconfig Role: Administrator Server: 10.108.246.36

Figure 52 Solutions Enabler virtual appliance management web interface

The Solutions Enabler virtual appliance simplifies and accelerates the process to provide Solutions Enabler API services in an enterprise environment. It is simple and easy to manage, preconfigured, centralized, and can be configured to be highly available. By leveraging the features provided by the Solutions Enabler virtual appliance, customers can quickly deploy and securely manage the services provided to burgeoning EMC applications that require access to Solutions Enabler API service.

Note: Like any other SYMAPI server, the Solutions Enabler vApp maintains the requirement for direct access to Symmetrix management LUNs known as gatekeepers. A minimum of six gatekeepers (six gatekeepers per Symmetrix array) needs to be presented to the virtual appliance as physical pass-through raw device mappings to allow for management and control of a Symmetrix storage array. Readers should consult the Solutions Enabler version installation guide available on Powerlink.EMC.com for detailed description of the virtual appliance. For more information on gatekeepers, see Primus article emc255976.

Extension of Solutions Enabler to support VMware virtual environments

Starting with Solutions Enabler 7.2, it is possible to performs inquiries into a vSphere environment using SYMCLI commands to identify and resolve storage configuration. The `symvm` command can add devices to virtual machines and display storage configuration information about authorized vSphere environments. `Symvm` works with VMware ESX and VMware ESXi.

In order to access this information, each ESX server must have credentials stored in the authorization database. The `symcfg auth` command is used to set credentials for each ESX server.

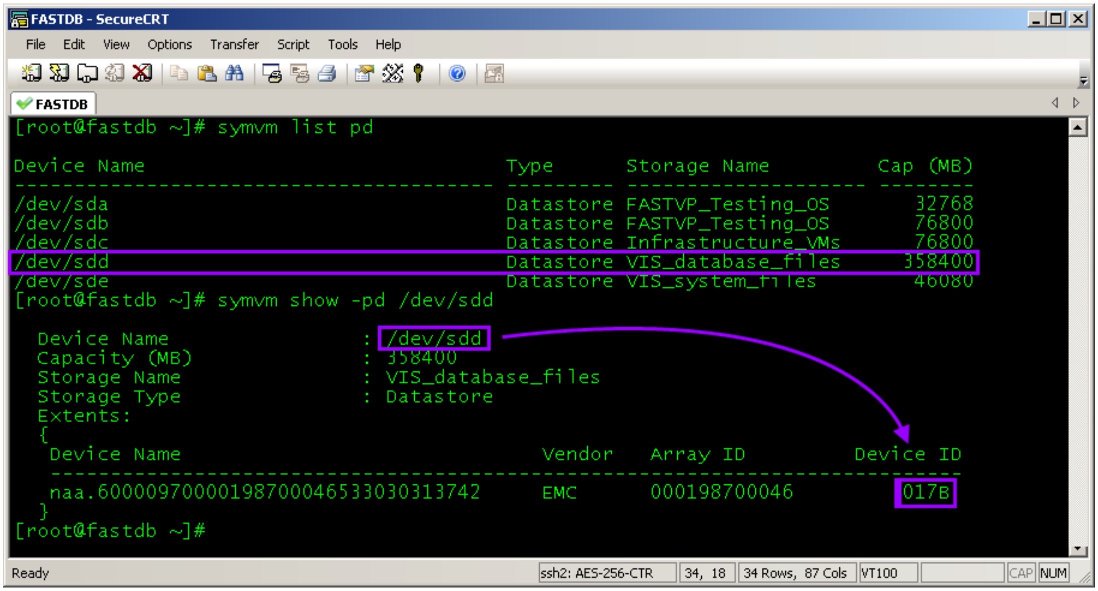
For a ESX host with a FQDN of api105.emc.com the following would be an example of the authorization command:

```
symcfg auth -vmware add -host api105.emc.com -username
   root -password Letmein!
```

Once the authorizations are put in place, Solutions Enabler can be used to query the authorized ESX servers as well as add available devices as raw device mappings to virtual machines. An example of this is shown in Figure 53 on page 107. A short list of functionality available when using `symvm` is as follows:

- Adding available Symmetrix devices as raw device mappings to virtual machines (devices must be mapped and masked to the specified host first)
- List available candidate devices for RDMs
- List a virtual machine’s RDMs

- List operating systems of virtual machines
- List datastores on a given ESX host
- Provide details of a given datastore
- List details of a given virtual machine



```
[root@fastdb ~]# symvm list pd

Device Name                               Type      Storage Name          Cap (MB)
-----------------------------------------  --------- --------------------  --------
/dev/sda                                  Datastore FASTVP_Testing_OS        32768
/dev/sdb                                  Datastore FASTVP_Testing_OS        76800
/dev/sdc                                  Datastore Infrastructure_VMs       76800
/dev/sdd                                  Datastore VIS_database_files      358400
/dev/sde                                  Datastore VIS_system_files         46080
[root@fastdb ~]# symvm show -pd /dev/sdd

    Device Name                 : /dev/sdd
    Capacity (MB)               : 358400
    Storage Name                : VIS_database_files
    Storage Type                : Datastore
    Extents:
    {
      Device Name                               Vendor   Array ID          Device ID
      ---------------------------------------------------------------------------
      naa.60000970000198700046533030313742      EMC      000198700046          017B
    }
[root@fastdb ~]#
```

Figure 53 **Solutions Enabler symvm command**

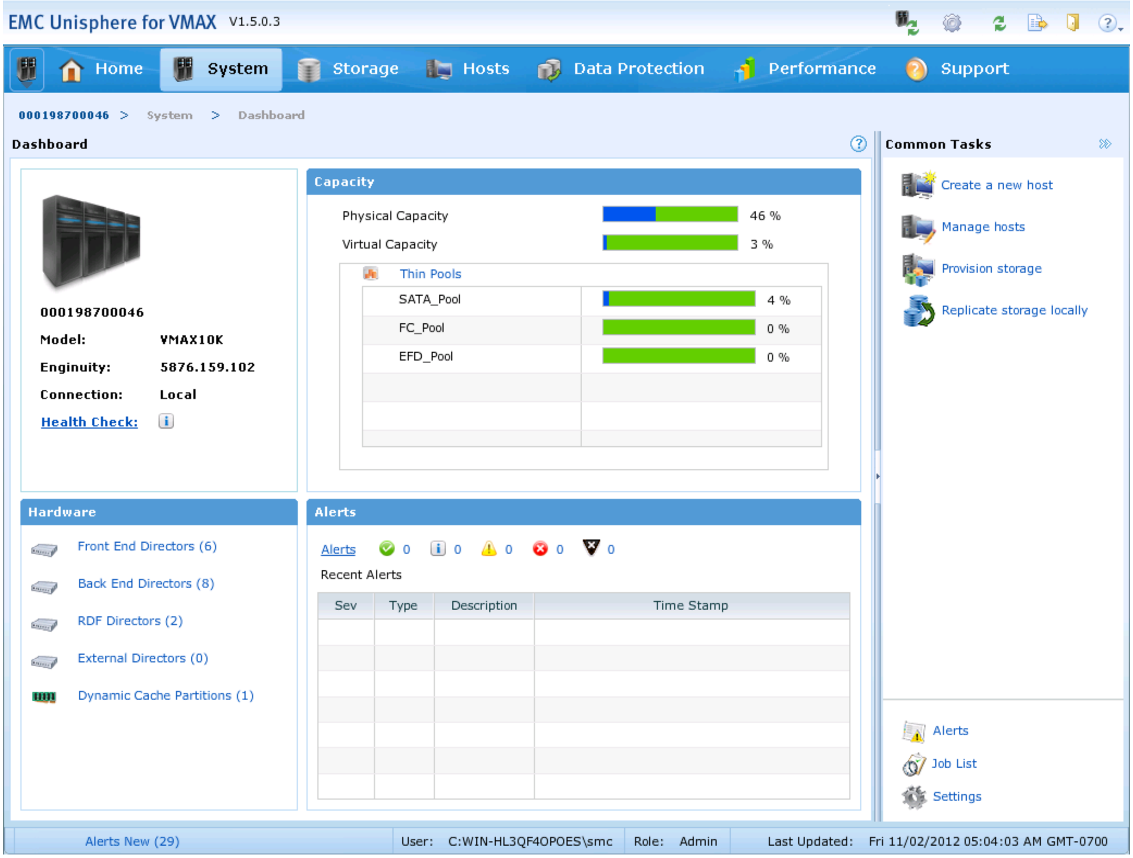
For the complete list of functionality and directions on specific syntax, refer to the *Solutions Enabler Management Product Guide* available on www.Powerlink.EMC.com.

EMC Unisphere for VMAX in VMware environments

Beginning with Enginuity 5876, Symmetrix Management Console has been transformed into EMC Unisphere for VMAX (hitherto known also as Unisphere) which offers big-button navigation and streamlined operations to simplify and reduce the time required to manage a data center. Unisphere for VMAX simplifies storage management under a common framework, incorporating Symmetrix Performance Analyzer which previously required a separate interface. You can use Unisphere to:

- Manage user accounts and roles
- Perform configuration operations (create volumes, mask volumes, set Symmetrix attributes, set volume attributes, set port flags, and create SAVE volume pools)
- Manage volumes (change volume configuration, set volume status, and create/dissolve meta volumes)
- Manage Fully Automated Storage Tiering (FAST, FAST VP)
- Perform and monitor replication operations (TimeFinder/Snap, TimeFinder/Clone, Symmetrix Remote Data Facility (SRDF), Open Replicator for Symmetrix (ORS))
- Manage advanced Symmetrix features, such as:
  - Fully Automated Storage Tiering (FAST)
  - Fully Automated Storage Tiering for virtual pools (FAST VP)
  - Enhanced Virtual LUN Technology
  - Auto-provisioning Groups
  - Virtual Provisioning
  - Federated Live Migration
  - Federated Tiered Storage (FTS)
- Monitor alerts

The Unisphere for VMAX dashboard page is shown in Figure 54.



000198700046

Model: VMAX10K

Enginuity: 5876.159.102

Connection: Local

Health Check:

Physical Capacity	46 %
Virtual Capacity	3 %

Thin Pools

SATA_Pool	4 %
FC_Pool	0 %
EFD_Pool	0 %

Alerts ✓ 0 ⓘ 0 ⚠ 0 ✖ 0 ▼ 0

Recent Alerts

Sev	Type	Description	Time Stamp

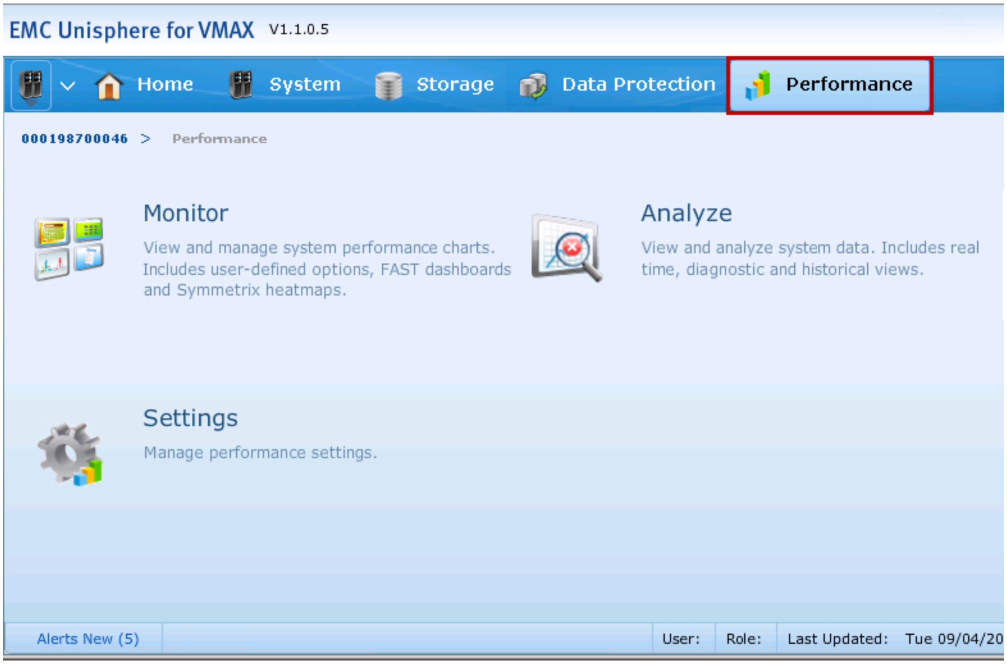
User: C:WIN-HL3QF4OPOES\smc Role: Admin Last Updated: Fri 11/02/2012 05:04:03 AM GMT-0700

Figure 54 Unisphere for VMAX dashboard

In addition, with the Performance monitoring option, Unisphere for VMAX provides tools for performing analysis and historical trending of Symmetrix system performance data. You can use the performance option to:

- Monitor performance and capacity over time
- Drill-down through data to investigate issues
- View graphs detailing system performance
- Set performance thresholds and alerts
- View high frequency metrics in real time

- Perform root cause analysis
- View Symmetrix system heat maps
- Execute scheduled and ongoing reports (queries), and export that data to a file
- Utilize predefined dashboards for many of the system components
- Customize your own dashboard templates


EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
Data Protection
Performance
000198700046 > Performance
Alerts New (5)
User:
Role:
Last Updated: Tue 09/04/20

Monitor

View and manage system performance charts. Includes user-defined options, FAST dashboards and Symmetrix heatmaps.

Analyze

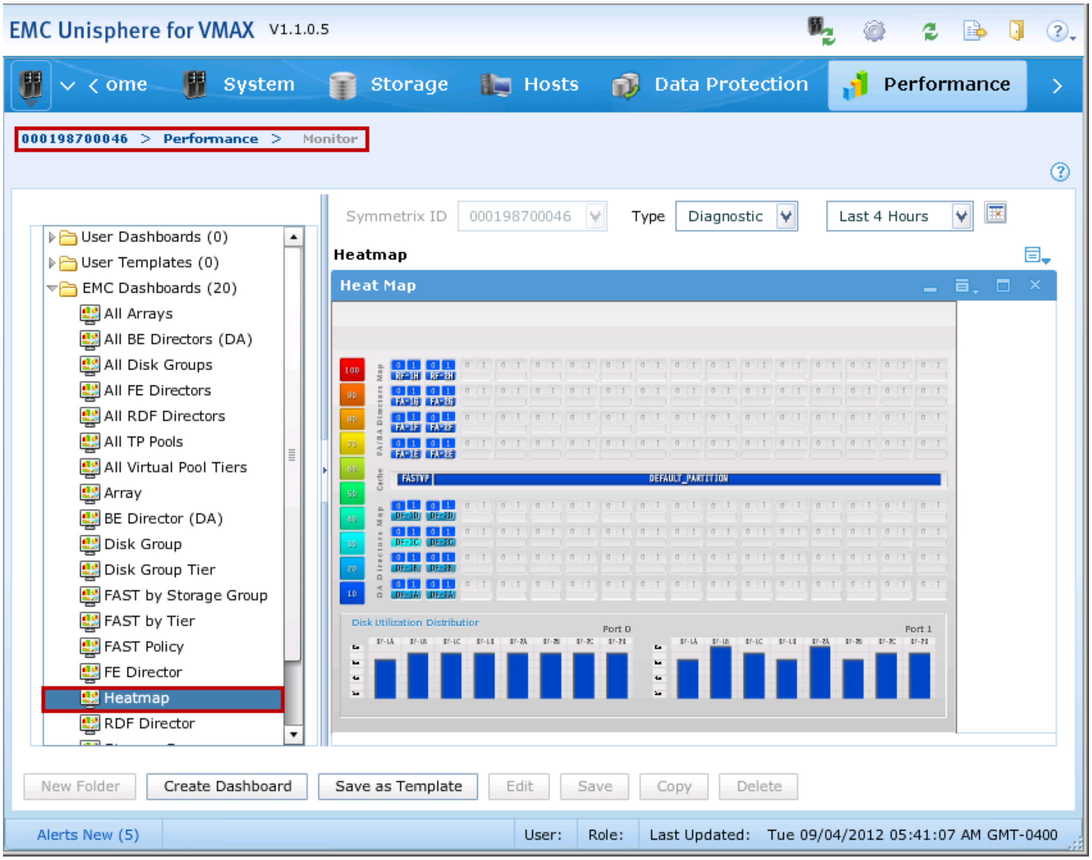
View and analyze system data. Includes real time, diagnostic and historical views.

Settings

Manage performance settings.

Figure 55 Unisphere for VMAX performance monitoring option

There are three main options from the Performance screen seen in Figure 55: Monitor, Analyze, and Settings. Using the Monitor screen the user can review diagnostic, historical, or real-time information about the array. This includes EMC dashboards such as the Heatmap in Figure 56 on page 111.



000198700046 > Performance > Monitor

Symmetrix ID: 000198700046

Type: Diagnostic

Last 4 Hours

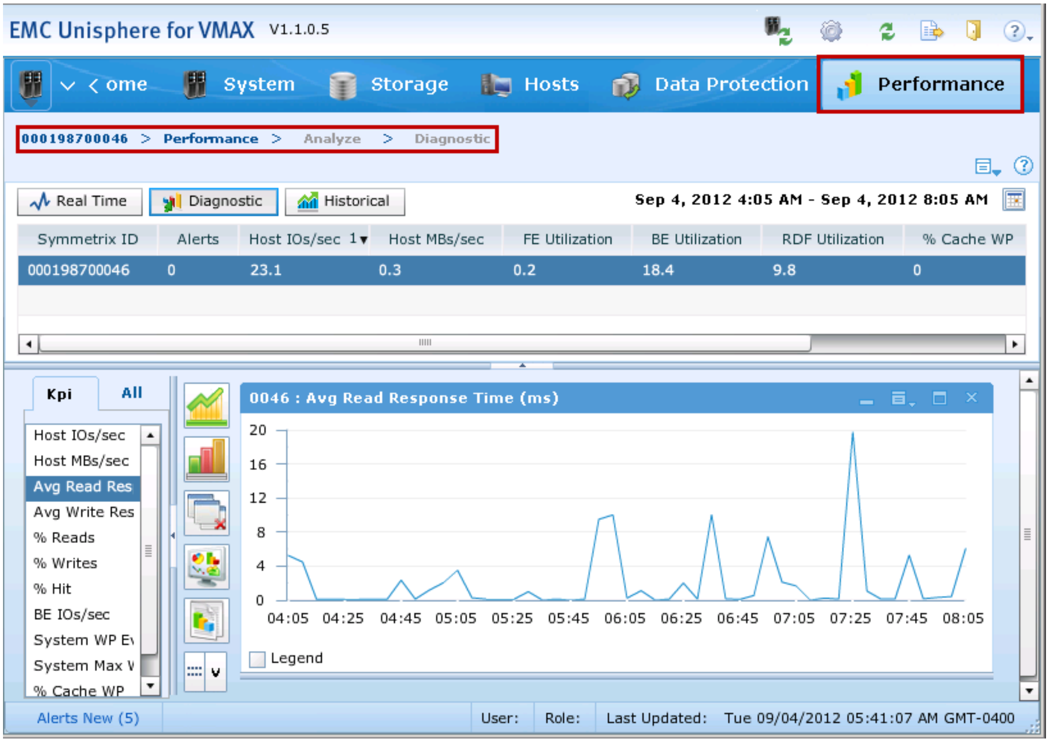
User:

Role:

Last Updated: Tue 09/04/2012 05:41:07 AM GMT-0400

Figure 56 Unisphere Heatmap dashboard

The analyze screen, Figure 57 on page 112, can be used to convert the raw data from performance monitoring into useful graphs to help diagnose issues, view historical data, or even watch real-time activity.



000198700046 > Performance > Analyze > Diagnostic

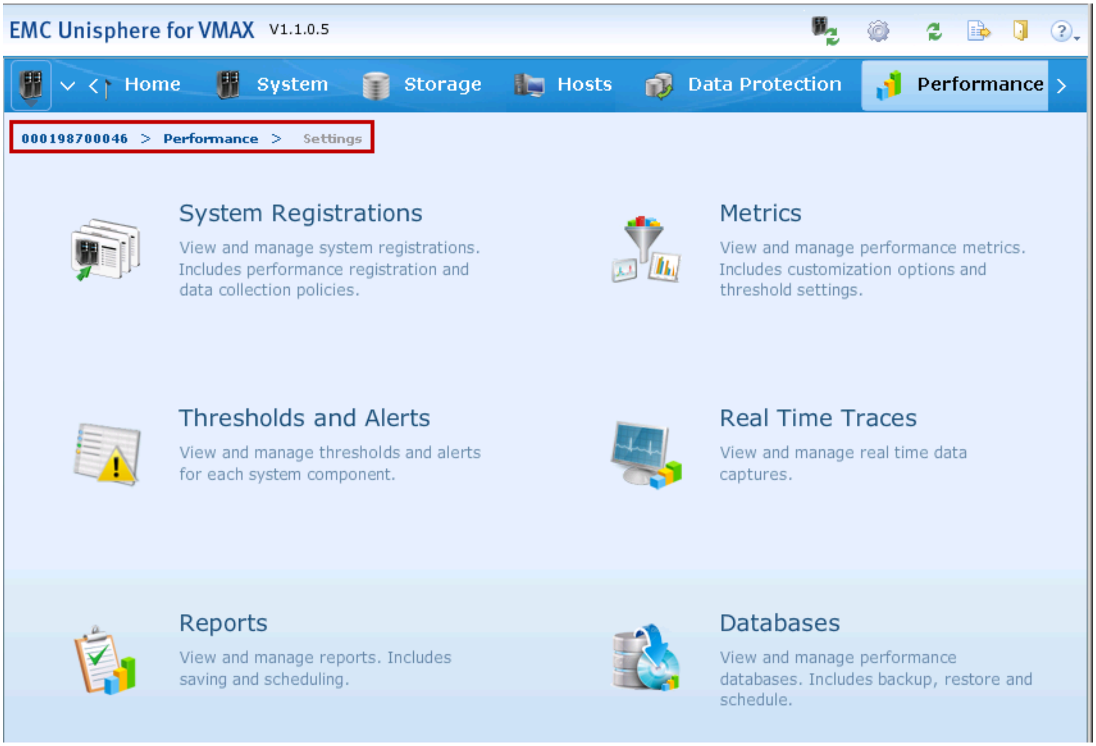
Sep 4, 2012 4:05 AM - Sep 4, 2012 8:05 AM

Symmetrix ID	Alerts	Host IOs/sec 1▼	Host MBs/sec	FE Utilization	BE Utilization	RDF Utilization	% Cache WP
000198700046	0	23.1	0.3	0.2	18.4	9.8	0

User: Role: Last Updated: Tue 09/04/2012 05:41:07 AM GMT-0400

Figure 57 **Unisphere analyze diagnostics**

Finally, the settings screen in Figure 58 on page 113 provides the user the ability to modify the metrics that are collected, run traces and reports, and setup thresholds and alerts.


EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
Hosts
Data Protection
Performance
000198700046 > Performance > Settings

System Registrations

View and manage system registrations. Includes performance registration and data collection policies.

Metrics

View and manage performance metrics. Includes customization options and threshold settings.

Thresholds and Alerts

View and manage thresholds and alerts for each system component.

Real Time Traces

View and manage real time data captures.

Reports

View and manage reports. Includes saving and scheduling.

Databases

View and manage performance databases. Includes backup, restore and schedule.

Figure 58 Settings in the performance monitoring option

Unisphere for VMAX can be run on a number of different kinds of open systems hosts, physical or virtual. Unisphere for VMAX is also available as a virtual appliance for ESX version 4.0 (and later) in the VMware infrastructure. Whether a full install, or a virtual appliance, Unisphere for VMAX maintains the requirement for direct access to Symmetrix management LUNs known as gatekeepers. A minimum of six gatekeepers (six gatekeepers per Symmetrix array) needs to be presented to the virtual appliance as physical pass-through raw device mappings to allow for management and control of a Symmetrix storage array.

Unisphere for VMAX virtual appliance

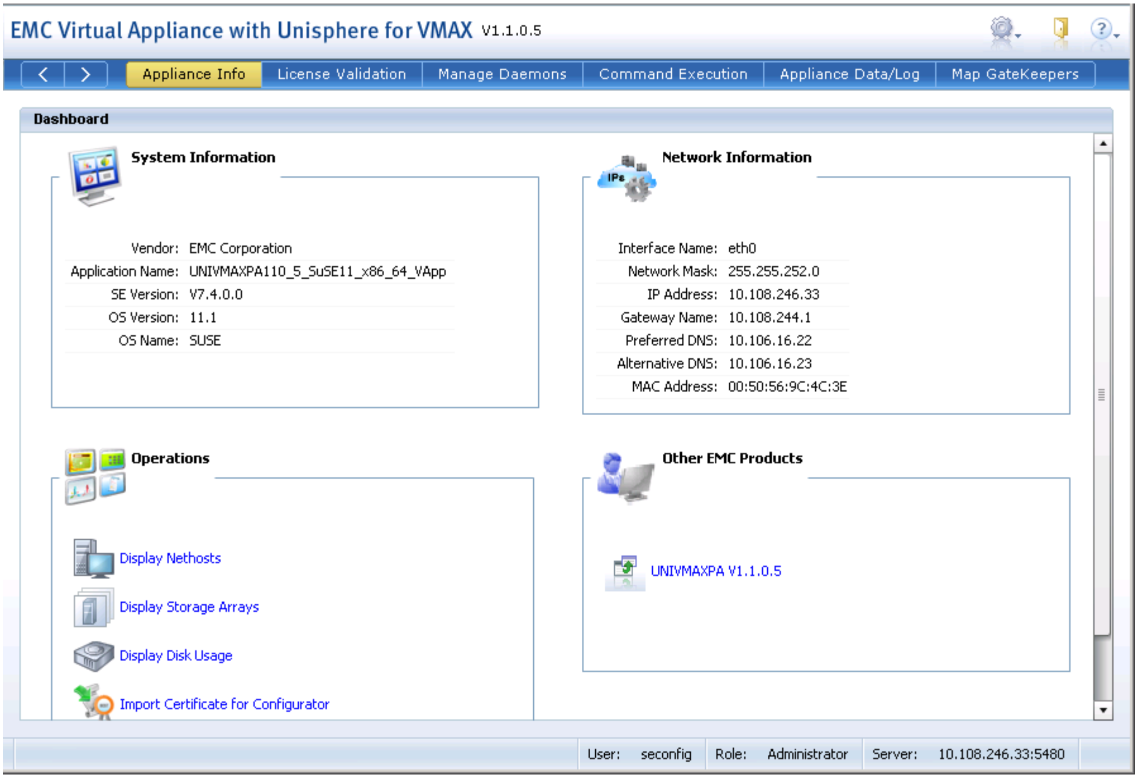
As an alternative to installing Unisphere for VMAX on a separate host configured by the user, Unisphere for VMAX is available in a virtual appliance format. The Unisphere for VMAX virtual appliance (henceforth known as the Unisphere Virtual Appliance or Unisphere vApp) is simple and easy to manage, preconfigured, centralized, and can be configured to be highly available. The Unisphere Virtual Appliance is a virtual machine that provides all the basic components required to manage a Symmetrix environment. This virtual appliance includes:

- EMC Unisphere for VMAX 1.5
- The Performance option
- EMC Solutions Enabler V7.5.0 (solely intended as a SYMAPI server for Solutions Enabler client access
- Linux OS (SUSE 11 64bit SP1)
- SMI-S Provider V4.5.0

In addition, the Unisphere vApp also includes a browser-based configuration tool, called the Solutions Enabler Virtual Appliance Configuration Manager shown in Figure 59 on page 115 (version 1.1). This web interface offers the following configuration tasks not available in SMC or from the virtual appliance directly:

- Launch Unisphere
- Monitor the application status
- Start and stop selected daemons
- Import and export persistent data
- Configure the nethost file (required for client access)
- Discover storage systems
- Modify options and daemon options
- Add host-based license keys
- Run a limited set of Solutions Enabler CLI commands
- Configure ESX host and gatekeeper volumes
- Load Symmetrix-based eLicenses

Readers should consult the EMC Unisphere for VMAX installation guide available on Powerlink.EMC.com for detailed description of the virtual appliance.



Vendor: EMC Corporation
Application Name: UNIVMAXPA110_5_SuSE11_x86_64_VApp
SE Version: V7.4.0.0
OS Version: 11.1
OS Name: SUSE

Interface Name: eth0
Network Mask: 255.255.252.0
IP Address: 10.108.246.33
Gateway Name: 10.108.244.1
Preferred DNS: 10.106.16.22
Alternative DNS: 10.106.16.23
MAC Address: 00:50:56:9C:4C:3E

User: seconfig Role: Administrator Server: 10.108.246.33:5480

Figure 59 **Unisphere Virtual Appliance management web interface**

The introduction of the Unisphere Virtual Appliance simplifies and accelerates the process to deploy Unisphere for VMAX quickly and securely in an enterprise environment.

EMC Unisphere for VMAX virtual environment integration

The ability to connect to and resolve storage information in a vSphere environment was introduced in the 7.3 release of Symmetrix Management Console and is available in Unisphere beginning with version 1.1. This can be accessed from the Hosts menu under "Virtual Servers" as shown in Figure 60 on page 116.

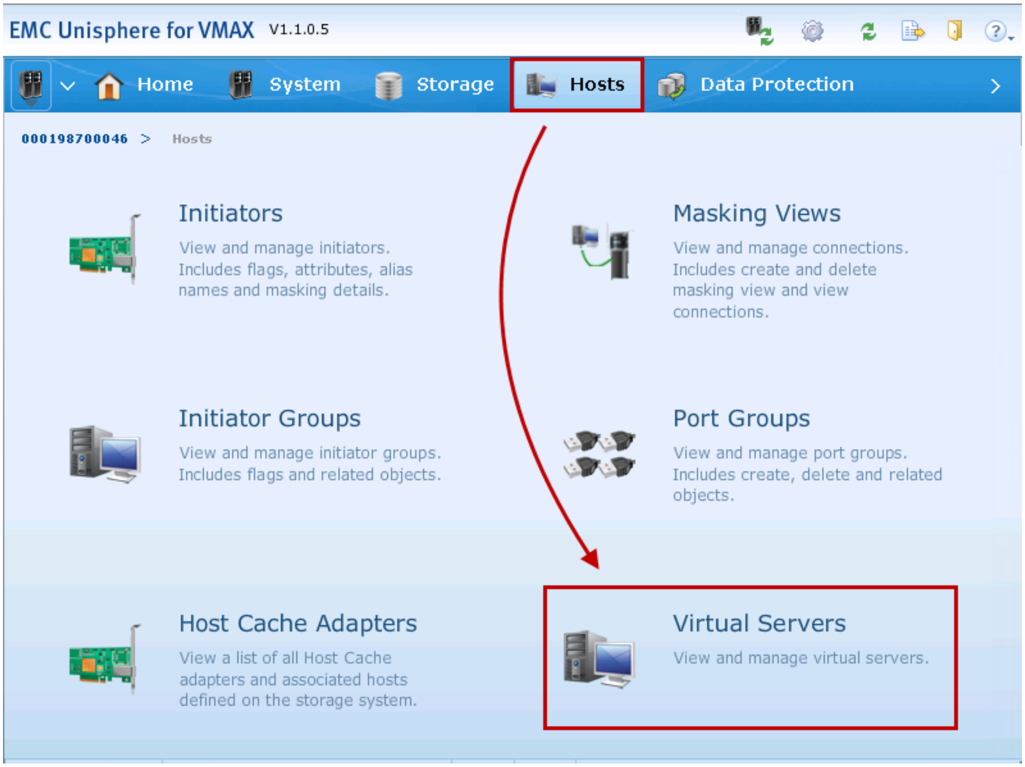
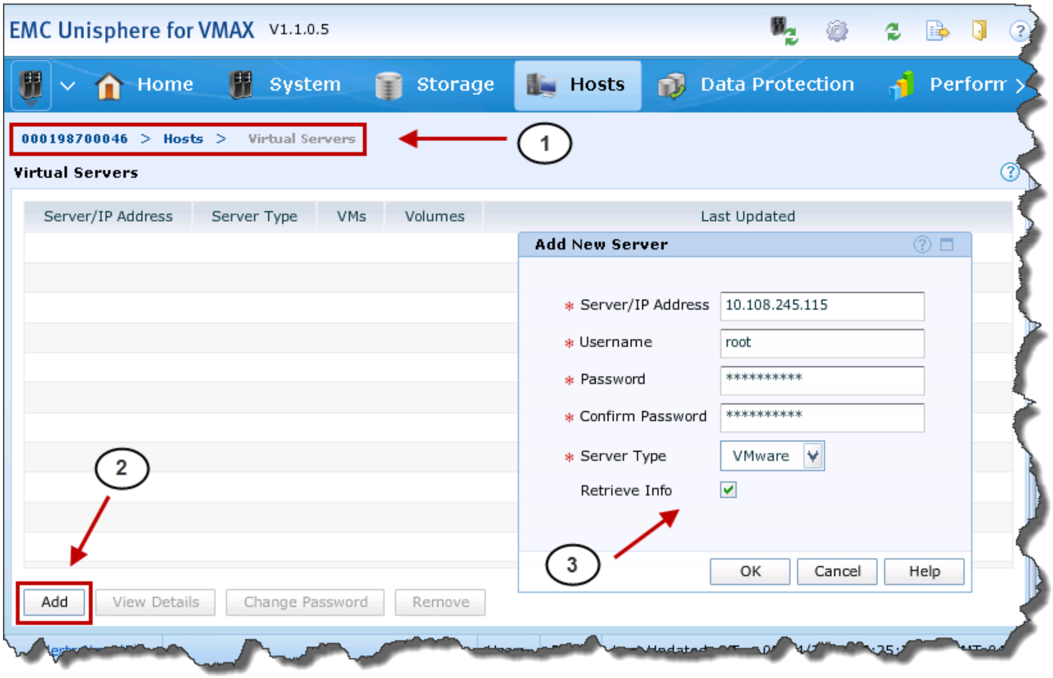

EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
Hosts
Data Protection
000198700046 > Hosts
Initiators
View and manage initiators. Includes flags, attributes, alias names and masking details.
Masking Views
View and manage connections. Includes create and delete masking view and view connections.
Initiator Groups
View and manage initiator groups. Includes flags and related objects.
Port Groups
View and manage port groups. Includes create, delete and related objects.
Host Cache Adapters
View a list of all Host Cache adapters and associated hosts defined on the storage system.
Virtual Servers
View and manage virtual servers.

Figure 60 Unisphere for VMAX VMware integration

In order to connect to various ESX hosts, authorizations must be created to allow authentication from Unisphere. It is important to note that authorizations are required for each ESX host. Figure 61 on page 117 shows the process to register an ESX host. By checking the box next to *Retrieve Info*, Unisphere will gather all information about the storage and virtual machines associated with that ESX host.



000198700046 > Hosts > Virtual Servers

Virtual Servers

Server/IP Address	Server Type	VMs	Volumes	Last Updated

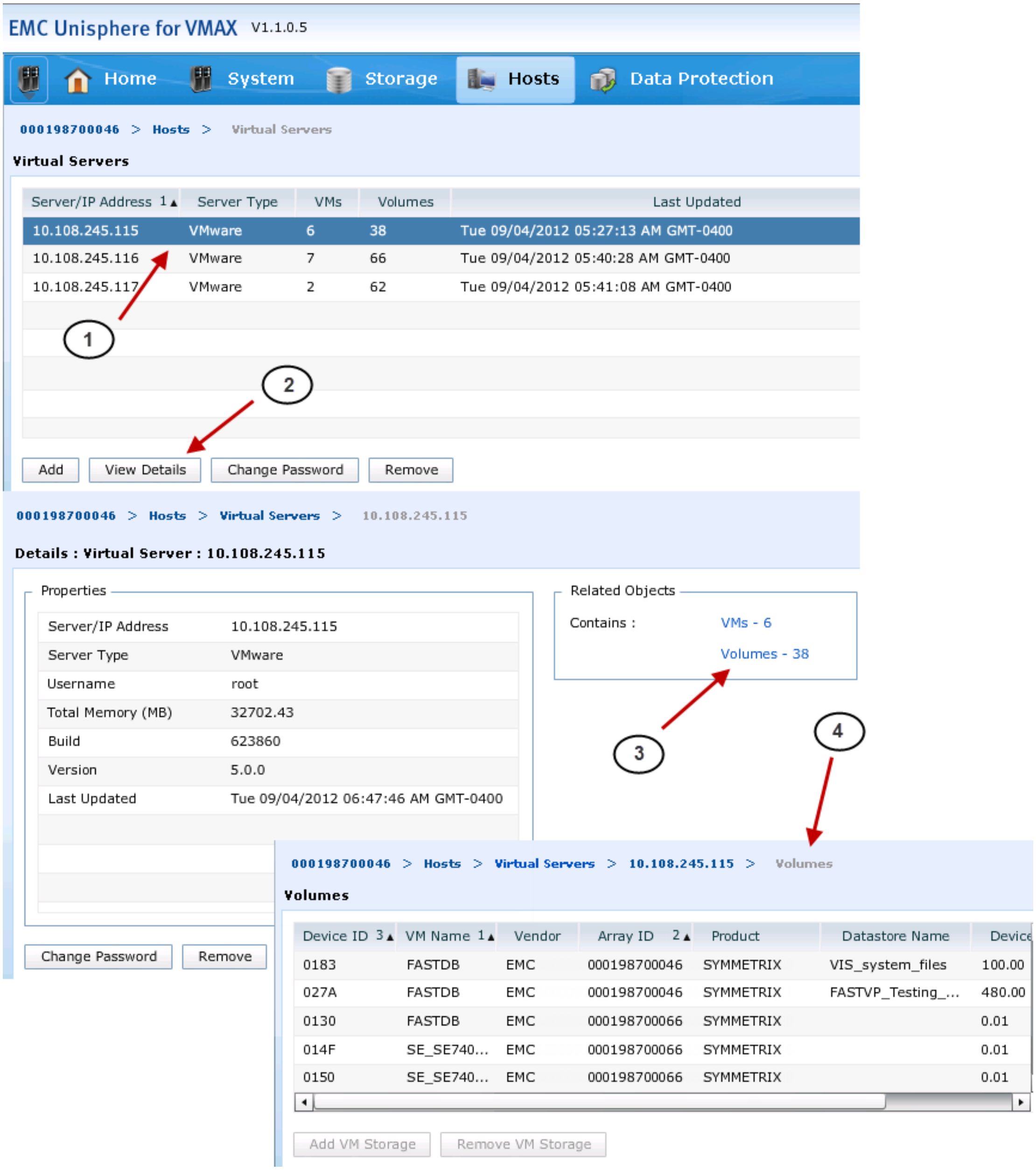
Add New Server

* Server/IP Address	10.108.245.115
* Username	root
* Password	**********
* Confirm Password	**********
* Server Type	VMware
Retrieve Info	☑

Figure 61 Registering an ESXi host with Unisphere for VMAX

Once the ESX host is registered, the storage information can then be viewed from Unisphere.

From the Virtual Servers screen, the user can select one of the registered servers and click on **View Details**. The pop-up box displays information about the ESX server and has two other drill-down related objects, **VMs** and **Volumes**. The navigation through **Volumes** is shown in Figure 62 on page 118.



000198700046 > Hosts > Virtual Servers

Virtual Servers

Server/IP Address 1▲	Server Type	VMs	Volumes	Last Updated
10.108.245.115	VMware	6	38	Tue 09/04/2012 05:27:13 AM GMT-0400
10.108.245.116	VMware	7	66	Tue 09/04/2012 05:40:28 AM GMT-0400
10.108.245.117	VMware	2	62	Tue 09/04/2012 05:41:08 AM GMT-0400

000198700046 > Hosts > Virtual Servers > 10.108.245.115

Details : Virtual Server : 10.108.245.115

Properties

Server/IP Address	10.108.245.115
Server Type	VMware
Username	root
Total Memory (MB)	32702.43
Build	623860
Version	5.0.0
Last Updated	Tue 09/04/2012 06:47:46 AM GMT-0400

Related Objects

Contains : VMs - 6

Volumes - 38

000198700046 > Hosts > Virtual Servers > 10.108.245.115 > Volumes

Volumes

Device ID 3▲	VM Name 1▲	Vendor	Array ID 2▲	Product	Datastore Name	Device
0183	FASTDB	EMC	000198700046	SYMMETRIX	VIS_system_files	100.00
027A	FASTDB	EMC	000198700046	SYMMETRIX	FASTVP_Testing_...	480.00
0130	FASTDB	EMC	000198700066	SYMMETRIX		0.01
014F	SE_SE740...	EMC	000198700066	SYMMETRIX		0.01
0150	SE_SE740...	EMC	000198700066	SYMMETRIX		0.01

Figure 62 Viewing ESX host details with Unisphere for VMAX

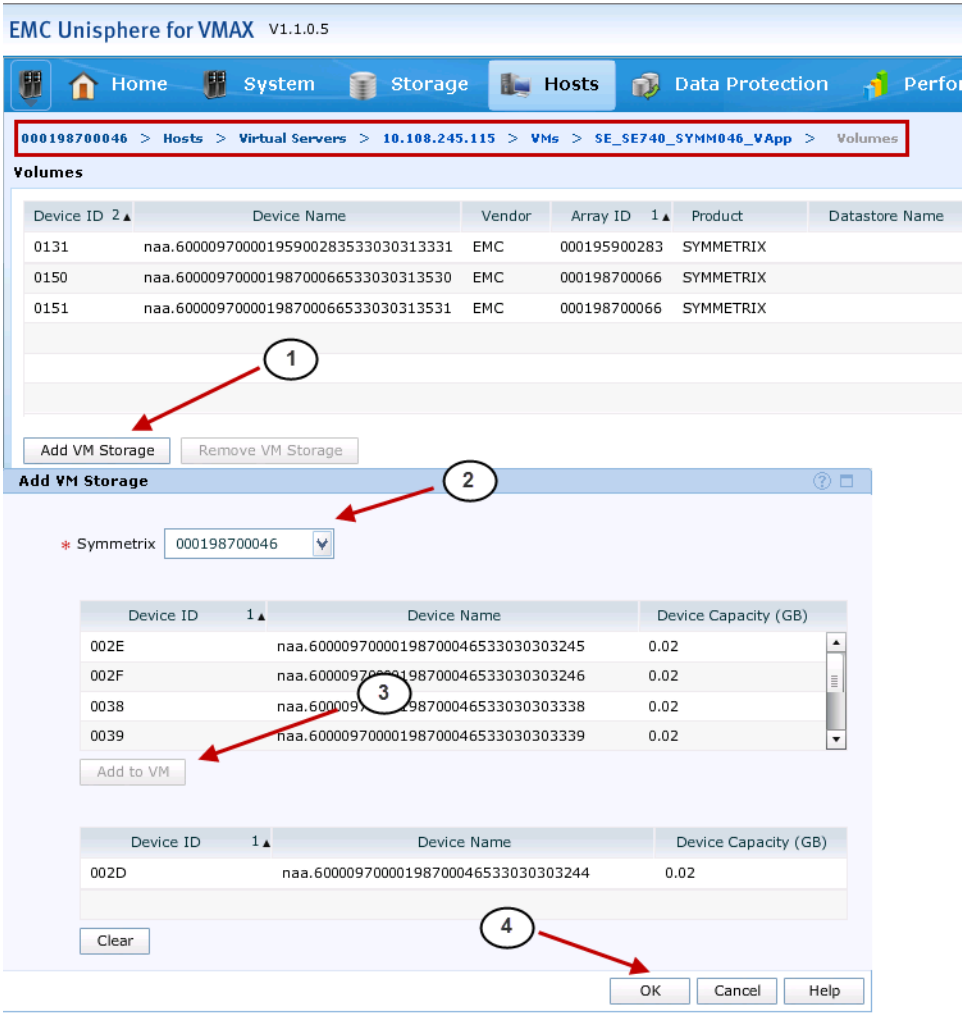
The **Volumes** detail screen has the following columns:

- Device ID
- VM Name
- Device Name
- Vendor
- Array ID
- Product
- Datastore Name
- Device Capacity

The **VMs** detail screen has the following columns:

- VM Name
- VM OS
- VM State
- Num of CPUs
- Total Memory (MB)

From the **VMs** screen, the user can add or remove VM storage. This screen will allow the user to map ESX-presented LUNs as RDMs to the selected VM, the detail of which is provided in Figure 63 on page 120.



000198700046 > Hosts > Virtual Servers > 10.108.245.115 > VMs > SE_SE740_SYMM046_VApp > Volumes

Volumes

Device ID 2▲	Device Name	Vendor	Array ID 1▲	Product	Datastore Name
0131	naa.60000970000195900283533030313331	EMC	000195900283	SYMMETRIX	
0150	naa.60000970000198700066533030313530	EMC	000198700066	SYMMETRIX	
0151	naa.60000970000198700066533030313531	EMC	000198700066	SYMMETRIX	

Add VM Storage

* Symmetrix 000198700046

Device ID 1▲	Device Name	Device Capacity (GB)
002E	naa.60000970000198700046533030303245	0.02
002F	naa.60000970000198700046533030303246	0.02
0038	naa.60000970000198700046533030303338	0.02
0039	naa.60000970000198700046533030303339	0.02

Device ID 1▲	Device Name	Device Capacity (GB)
002D	naa.60000970000198700046533030303244	0.02

Figure 63 Mapping RDMs to a VM in Unisphere for VMAX

Installing and configuring Symmetrix management applications on virtual machines

Even though the use of virtual appliances for Symmetrix management is generally preferred, Symmetrix management applications can be installed on virtual machines running supported operating systems. Symmetrix gatekeepers must be presented as RDMs in physical compatibility mode, with a minimum of six per Symmetrix being mapped. Otherwise, there are no special considerations for installing and using these applications under virtual machines as compared to physical machines.

Installation of Symmetrix management applications in a virtual machine follows the same process as installation on a physical server; however, installation in a virtual machine using a physical medium requires additional steps to present the CD to the virtual machines. After the installation is complete, add appropriate license keys to enable the base functionality and other components. Refer to product installation guides for detailed information.

Note: Sharing gatekeepers across multiple virtual machines is not supported.

vSphere vStorage APIs for Storage Awareness (VASA)

vSphere vStorage APIs for Storage Awareness are vCenter Server-based APIs that enable storage arrays to inform the vCenter Server about their configurations, capabilities, and storage health and events. They allow an administrator to be aware of physical storage topology, capability, and state. EMC supports vSphere vStorage APIs for Storage Awareness via the plug-in, or "provider", downloadable from www.Powerlink.EMC.com.[1]

Profile-Driven Storage

Managing datastores and matching the SLA requirements of virtual machines with the appropriate datastore can be challenging and cumbersome tasks. vSphere 5 introduces Profile-Driven Storage, which enables rapid and intelligent placement of virtual machines based on SLA, availability, performance or other requirements through provided storage capabilities.

Using Profile-Driven Storage, various storage characteristics, typically defined as a tier, can be requested in a virtual machine storage profile. These profiles are used during provisioning, cloning and Storage vMotion to ensure that only those datastores or datastore clusters that are compliant with the virtual machine storage profile are made available. The virtual machine storage profile can also help select a similar type of datastores when creating a Storage DRS datastore cluster. Profile-Driven Storage will reduce the amount of manual administration required for virtual machine placement while improving virtual machine SLA storage compliance.

Profile-Driven Storage delivers these benefits by taking advantage of the following:

- Full integration with vSphere vStorage APIs – Storage Awareness, enabling usage of storage characterization supplied by storage vendors.
- Support for NFS, iSCSI and Fibre Channel (FC) storage, and all storage arrays on the HCL.

[1] For full VASA functionality, the minimum version of this provider that should be used is SMI-S Provider version 4.3.2 for SMI 1.4. It must be installed on a host that has access to the Symmetrix through gatekeepers.

- Enabling the vSphere administrator to tag storage based on customer- or business-specific descriptions.
- Using storage characterizations and/or administrator-defined descriptions to create virtual machine placement rules in the form of storage profiles.
- Providing an easy means to check a virtual machine's compliance with these rules. This ensures that the virtual machine is not deployed or migrated to an incorrect type of storage without the administrator's being informed about it.

VASA features

Below are some of the general features of VASA broken out into their respective management areas:

Operations management

- Identify trends in a VM's storage capacity usage for troubleshooting
- Correlate events on datastore and LUNs with VM's performance characteristics
- Monitor health of storage

Policy-based management

- Choose the right storage in terms of space, performance and SLA requirements
- Monitor and report against existing storage policies
- Create a foundation for manageability in vSphere 5 and expand in future releases

Storage capacity management

- Identify trends in a VM's storage capacity usage
- Identify impact of newly created (or migrated) VMs on storage space

Vendor providers and storage data representation

The vCenter Server communicates with the vendor provider to obtain information that the vendor provider collects from available storage devices. The vCenter Server can then display the storage data in the vSphere Client. The storage vendor first needs to be registered in the vSphere Client. This can be achieved by selecting the “Storage Providers” icon in the home screen of the vSphere Client as in Figure 64.

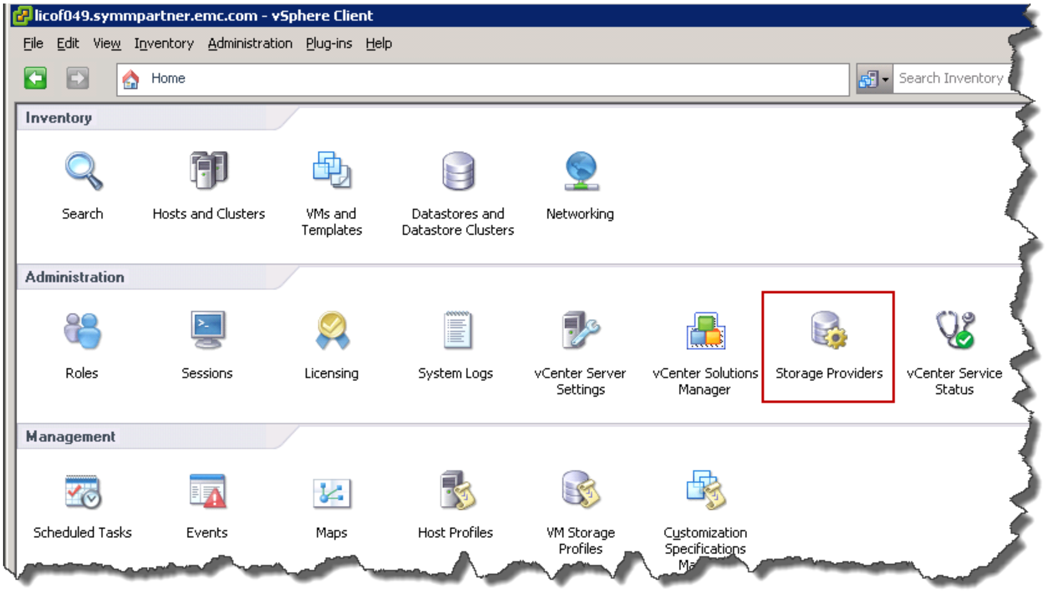


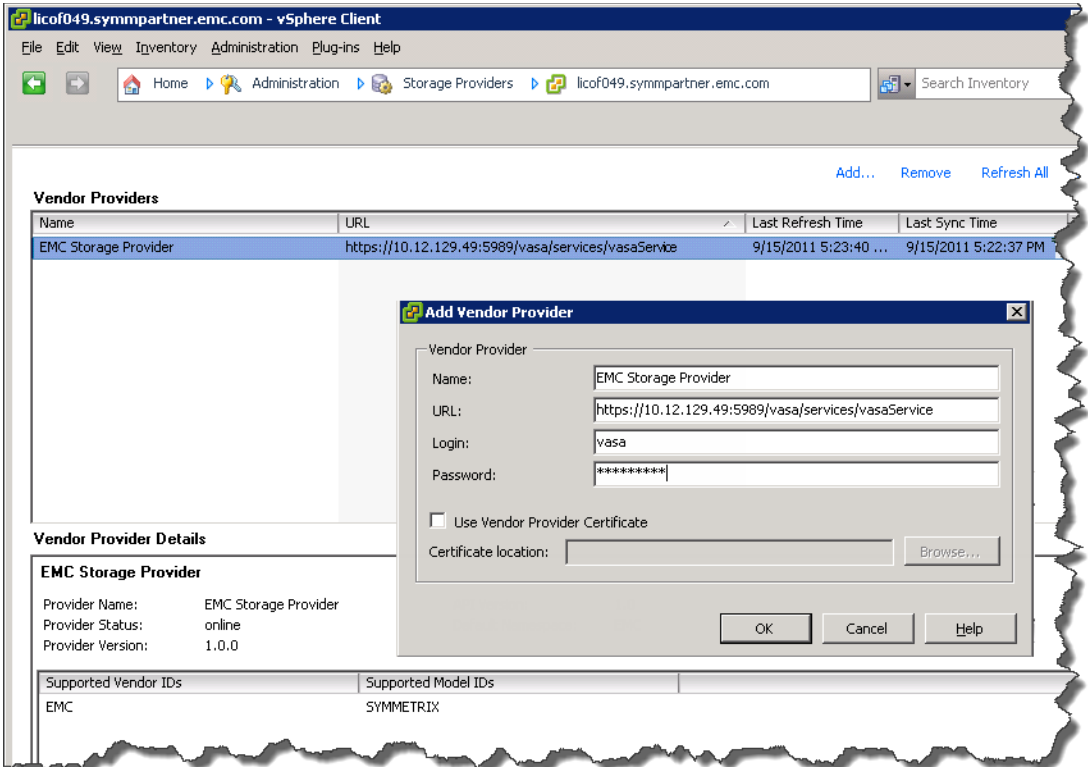
Figure 64 Storage providers in the vSphere Client

The EMC SMI-S Provider can then be registered using the dialog box provided in the vSphere Client. Table 2 lists the values required for each variable in the registration box and whether or not the field is validated. A CIM user with administrator privileges needs to be created before registering the SMI-S Provider. Details on how to this is accomplished can be found in the post-installation tasks of the SMI-S Provider release notes.

Table 2 Provider registration

Add Vendor Provider	
Variable	**Value**
Name	Any descriptive name (not validated)
URL	https://<VASA_host>:5989/vasa/services/vasaService (validated)
Login	Administrative user (validated)
Password	Administrative user password (validated)

In Figure 65 is the registration box with the result of the registration appearing in the background.



Vendor Providers

Name	URL	Last Refresh Time	Last Sync Time
EMC Storage Provider	https://10.12.129.49:5989/vasa/services/vasaService	9/15/2011 5:23:40 ...	9/15/2011 5:22:37 PM

Vendor Provider

Name: EMC Storage Provider

URL: https://10.12.129.49:5989/vasa/services/vasaService

Login: vasa

Password: *********

☐ Use Vendor Provider Certificate

Certificate location:

Vendor Provider Details

EMC Storage Provider

Provider Name:	EMC Storage Provider
Provider Status:	online
Provider Version:	1.0.0

Supported Vendor IDs	Supported Model IDs
EMC	SYMMETRIX

Figure 65 EMC SMI-S Provider registration

Once registered, the EMC SMI-S Provider will supply information to the vCenter. Information that the vendor provider supplies can be divided into the following categories:

- Storage topology — Information about physical storage elements appears on the Storage Views tab. It includes such data as storage arrays, array IDs, etc.

  This type of information can be helpful when you need to track virtual machine-to-storage relationships and configuration, or to identify any changes in physical storage configuration.

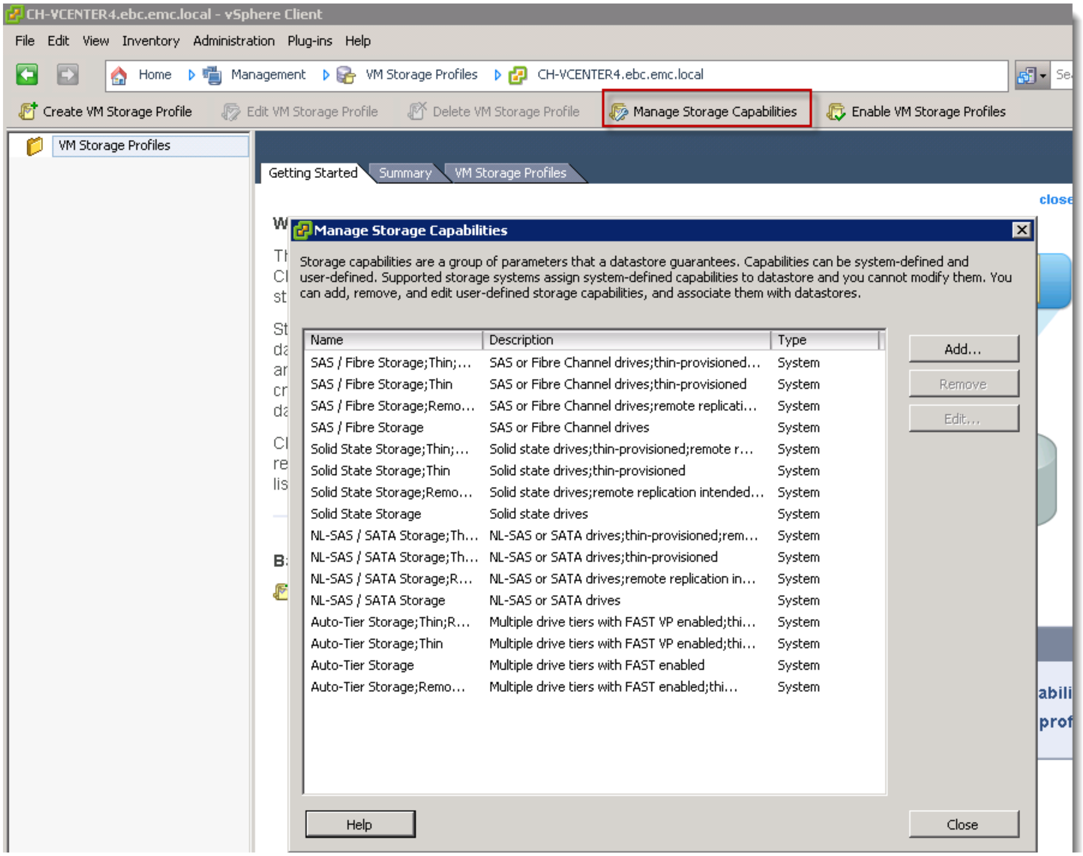
- Storage capabilities — The vendor provider collects and communicates information about physical capabilities and services that the underlying storage offers.

This information can be useful when, for example, you need to properly aggregate storage into tiers, or select the right storage, in terms of space and performance, for a particular virtual machine. Table 3, next, and Figure 66 on page 127 detail the default storage capabilities that accompany the registration of the EMC storage provider.

Table 3 VASA-provided storage capabilities

Name	Description
SAS/Fibre Storage; Thin; Remote Replication	SAS or Fibre Channel drives; thin-provisioned; remote replication intended to provide disaster recovery.
SAS/Fibre Storage; Thin	SAS or Fibre Channel drives; thin-provisioned.
SAS/Fibre Storage; Remote Replication	SAS or Fibre Channel drives; remote replication intended to provide disaster recovery.
SAS/Fibre Storage	SAS or Fibre Channel drives.
Solid State Storage; Thin; Remote Replication	Solid state drives; thin-provisioned; remote replication intended to provide disaster recovery.
Solid State Storage; Thin	Solid state drives; thin-provisioned.
Solid State Storage; Remote Replication	Solid state drives; remote replication intended to provide disaster recovery.
Solid State Storage	Solid state drives.
NL-SAS/SATA Storage; Thin; Remote Replication	Near Line SAS or SATA drives; thin-provisioned; remote replication intended to provide disaster recovery.
NL-SAS/SATA Storage; Thin	Near Line SAS or SATA drives; thin-provisioned.
NL-SAS/SATA Storage; Remote Replication	Near Line SAS or SATA drives; remote replication intended to provide disaster recovery.
NL-SAS/SATA Storage	Near Line SAS or SATA drives.
Auto-Tier[a] Storage; Thin; Remote Replication	Multiple drives tiers with FAST VP enabled; thin-provisioned; remote replication intended to provide disaster recovery.
Auto-Tier Storage; Thin	Multiple drives tiers with FAST VP enabled; thin-provisioned.
Auto-Tier Storage; Remote Replication	Multiple drives tiers with FAST enabled; remote replication intended to provide disaster recovery.
Auto-Tier Storage	Multiple drives tiers with FAST enabled.

a.In addition to FAST VP enabled devices (pinned or unpinned), a device that has extents in more than a single tier (technology) will also be categorized as Auto-Tier.


CH-VCENTER4.ebc.emc.local - vSphere Client
File Edit View Inventory Administration Plug-ins Help
Home
Management
VM Storage Profiles
CH-VCENTER4.ebc.emc.local
Create VM Storage Profile
Edit VM Storage Profile
Delete VM Storage Profile
Manage Storage Capabilities
Enable VM Storage Profiles
VM Storage Profiles
Getting Started
Summary
VM Storage Profiles
close
Manage Storage Capabilities
Add...
Remove
Edit...
Help
Close

Storage capabilities are a group of parameters that a datastore guarantees. Capabilities can be system-defined and user-defined. Supported storage systems assign system-defined capabilities to datastore and you cannot modify them. You can add, remove, and edit user-defined storage capabilities, and associate them with datastores.

Name	Description	Type
SAS / Fibre Storage;Thin;...	SAS or Fibre Channel drives;thin-provisioned...	System
SAS / Fibre Storage;Thin	SAS or Fibre Channel drives;thin-provisioned	System
SAS / Fibre Storage;Remo...	SAS or Fibre Channel drives;remote replicati...	System
SAS / Fibre Storage	SAS or Fibre Channel drives	System
Solid State Storage;Thin;...	Solid state drives;thin-provisioned;remote r...	System
Solid State Storage;Thin	Solid state drives;thin-provisioned	System
Solid State Storage;Remo...	Solid state drives;remote replication intended...	System
Solid State Storage	Solid state drives	System
NL-SAS / SATA Storage;Th...	NL-SAS or SATA drives;thin-provisioned;rem...	System
NL-SAS / SATA Storage;Th...	NL-SAS or SATA drives;thin-provisioned	System
NL-SAS / SATA Storage;R...	NL-SAS or SATA drives;remote replication in...	System
NL-SAS / SATA Storage	NL-SAS or SATA drives	System
Auto-Tier Storage;Thin;R...	Multiple drive tiers with FAST VP enabled;thi...	System
Auto-Tier Storage;Thin	Multiple drive tiers with FAST VP enabled;thi...	System
Auto-Tier Storage	Multiple drive tiers with FAST enabled	System
Auto-Tier Storage;Remo...	Multiple drive tiers with FAST enabled;thi...	System

Figure 66 **Manage storage capabilities**

Note: Depending on the Symmetrix arrays and devices that are presented to the vCenter environment, not all of the capabilities that appear in Figure 66 may be listed. For instance, if the vCenter environment is using VMAXe/10K arrays exclusively, only storage capabilities that are specific to thin devices will be shown.

- Storage status — This category includes reporting about status of various storage entities. It also includes alarms and events for notifying about configuration changes.

This type of information can help one troubleshoot storage connectivity and performance problems. It can also help one correlate array-generated events and alarms to corresponding performance and load changes on the array.

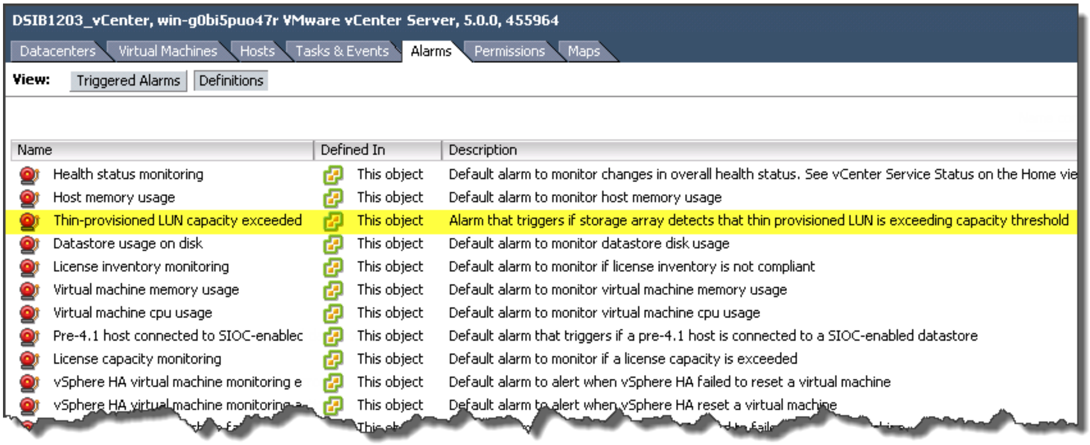
The alarms that are part of the storage status category are particularly useful when using Virtual Provisioning because they provide warnings when the thin LUNs on the VMAX/VMAXe are approaching capacity, something that before had to be monitored solely on the storage side.

Thin-provisioned LUN alarm

Within vSphere 5 there is a new alarm definition for VASA. It is named “Thin-provisioned LUN capacity exceeded” and will be triggered if the thin LUN on the Symmetrix exceeds predefined thresholds. These thresholds are as follows:

- < 65% is green
- >= 65% is yellow
- >= 80% is red

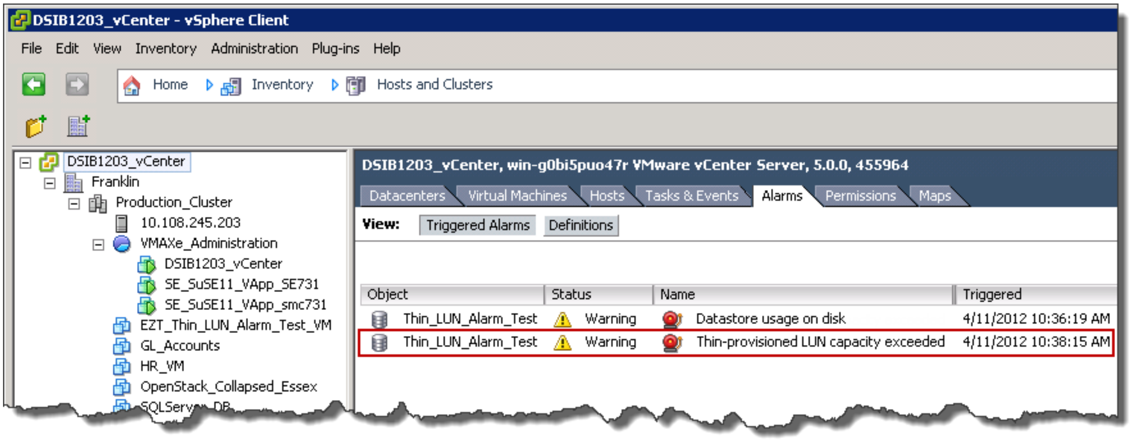
The alarm definition is seen in Figure 67.



Name	Defined In	Description
Health status monitoring	This object	Default alarm to monitor changes in overall health status. See vCenter Service Status on the Home vie
Host memory usage	This object	Default alarm to monitor host memory usage
Thin-provisioned LUN capacity exceeded	This object	Alarm that triggers if storage array detects that thin provisioned LUN is exceeding capacity threshold
Datastore usage on disk	This object	Default alarm to monitor datastore disk usage
License inventory monitoring	This object	Default alarm to monitor if license inventory is not compliant
Virtual machine memory usage	This object	Default alarm to monitor virtual machine memory usage
Virtual machine cpu usage	This object	Default alarm to monitor virtual machine cpu usage
Pre-4.1 host connected to SIOC-enablec	This object	Default alarm that triggers if a pre-4.1 host is connected to a SIOC-enabled datastore
License capacity monitoring	This object	Default alarm to monitor if a license capacity is exceeded
vSphere HA virtual machine monitoring e	This object	Default alarm to alert when vSphere HA failed to reset a virtual machine
vSphere HA virtual machine monitoring a	This object	Default alarm to alert when vSphere HA reset a virtual machine

Figure 67 Thin-provisioned LUN capacity exceeded alarm definition

As a thin LUN below 65% will produce no alarm, a triggered alarm is either a warning (yellow) or an alert (red). A warning alert is seen in Figure 68 on page 129 for datastore “Thin_LUN_Alarm_Test”.

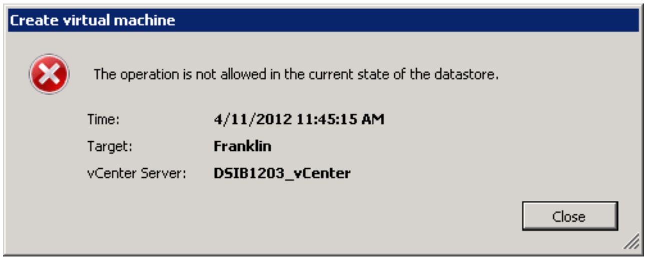


DSIB1203_vCenter, win-g0bi5puo47r VMware vCenter Server, 5.0.0, 455964

Object	Status	Name	Triggered
Thin_LUN_Alarm_Test	⚠ Warning	Datastore usage on disk	4/11/2012 10:36:19 AM
Thin_LUN_Alarm_Test	⚠ Warning	Thin-provisioned LUN capacity exceeded	4/11/2012 10:38:15 AM

Figure 68 Thin-provisioned LUN alarm in vCenter

While the warning alarm is just that and triggers nothing additional, an alert alarm will prevent the user from creating other virtual machines on the datastore using that thin LUN. VMware takes this step to reserve the remaining space for the current virtual machines and avoid any out of space conditions for them. If the user attempts to create a virtual machine on the datastore which has this alert, the error seen in Figure 69 will result.



The operation is not allowed in the current state of the datastore.

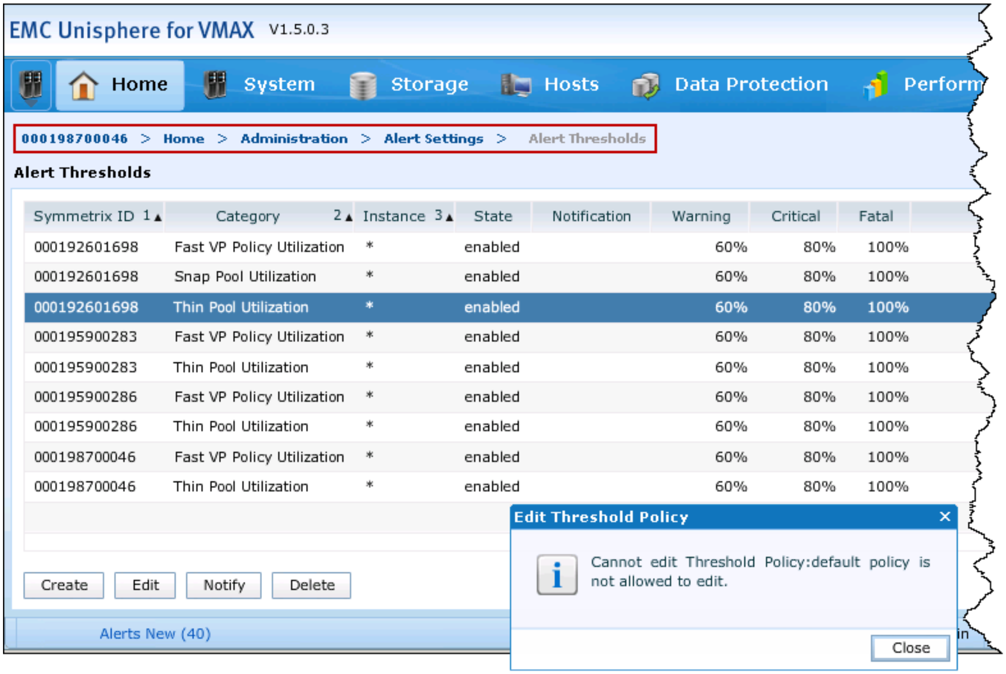
Time:	4/11/2012 11:45:15 AM
Target:	Franklin
vCenter Server:	DSIB1203_vCenter

Figure 69 Error received after alert on thin-provisioned LUN

The user must take corrective action at this point and may choose to increase the size of the thin LUN, have the system administrator perform a space reclamation on the thin LUN, use VMware Dead Space Reclamation (if supported and there is reclaimable space) or use a different datastore altogether for future virtual machines. Once the thin device has enough free space to be beyond the warning and alert thresholds, the alarm will be removed.

In order for vCenter to report an exhaustion of space on the thin-provisioned LUN on the Symmetrix, the SMI-S Provider takes advantage of the events that are automatically recorded by the event daemon on the Symmetrix. These events are recorded regardless of whether the user chooses to be alerted to them in Unisphere. Therefore once the SMI-S Provider is properly installed and registered in vCenter, no additional configuration is required on the Symmetrix or within EMC Unisphere for VMAX.

If the user wishes to see alerts for thin device allocation within Unisphere, the system administrator can configure them through the path *Home > Administration > Alert Settings > Alert Thresholds*. The configuration screen is shown in Figure 70 on page 131 where the default alert for thin device allocation is highlighted. The default alert cannot be altered - attempts to edit it will result in the error in the bottom-right of Figure 70 on page 131. Users can, however, create specific alerts for particular thin pools which will supersede the default alert.


EMC Unisphere for VMAX V1.5.0.3
Home
System
Storage
Hosts
Data Protection
Perform
000198700046 > Home > Administration > Alert Settings > Alert Thresholds
Create
Edit
Notify
Delete
Alerts New (40)
Close

Alert Thresholds

Symmetrix ID 1▲	Category 2▲	Instance 3▲	State	Notification	Warning	Critical	Fatal
000192601698	Fast VP Policy Utilization	*	enabled		60%	80%	100%
000192601698	Snap Pool Utilization	*	enabled		60%	80%	100%
000192601698	Thin Pool Utilization	*	enabled		60%	80%	100%
000195900283	Fast VP Policy Utilization	*	enabled		60%	80%	100%
000195900283	Thin Pool Utilization	*	enabled		60%	80%	100%
000195900286	Fast VP Policy Utilization	*	enabled		60%	80%	100%
000195900286	Thin Pool Utilization	*	enabled		60%	80%	100%
000198700046	Fast VP Policy Utilization	*	enabled		60%	80%	100%
000198700046	Thin Pool Utilization	*	enabled		60%	80%	100%

Edit Threshold Policy

Cannot edit Threshold Policy:default policy is not allowed to edit.

Figure 70 Thin device allocation alert within EMC Unisphere for VMAX

Such alarms and alerts allow the VMware administrator and system administrator to be aware of the state of the storage capacity presented to the VMware environment. The thin LUN alarm is only one of the ways that the vSphere vStorage APIs for Storage Awareness can provide insight for the VMware administrator into the Symmetrix storage.

Managing storage tiers, provisioning, migrating, cloning virtual machines and correct virtual machine placement in vSphere deployments have become more efficient and user friendly with VASA. It removes the need for maintaining complex and tedious spreadsheets and validating compliance manually during every migration or creation of a virtual machine or virtual disk.[1]

[1] For more information on using VASA with EMC Symmetrix, please reference the *Implementing VMware's vStorage API for Storage Awareness with Symmetrix Storage Arrays White Paper*, which can be found at www.EMC.com and on www.Powerlink.EMC.com.

⚠ CAUTION

When using VASA functionality, EMC does not recommend preallocating thin devices once a datastore has been created on the device. Preallocation of the device beyond the predefined thresholds will cause the vCenter alarm to be triggered. This could result in an inability to create virtual machines on the datastore. If preallocation of space within the datastore is required, EMC recommends using eagerzeroedthick virtual machines.

3

EMC Virtual Provisioning and VMware vSphere

This chapter discusses deploying and using EMC Symmetrix Virtual Provisioning in a VMware environment and recommendations and on combined use with vSphere Thin Provisioning:

Introduction

One of the biggest challenges facing storage administrators is provisioning storage for new applications. Administrators typically allocate space based on anticipated future growth of applications. This is done to mitigate recurring operational functions, such as incrementally increasing storage allocations or adding discrete blocks of storage as existing space is consumed. Using this approach results in more physical storage being allocated to the application than needed for a significant amount of time and at a higher cost than necessary. The overprovisioning of physical storage also leads to increased power, cooling, and floor space requirements. Even with the most careful planning, it may be necessary to provision additional storage in the future, which could potentially require an application outage.

A second layer of storage overprovisioning happens when a server and application administrator overallocate storage for their environment. The operating system sees the space as completely allocated but internally only a fraction of the allocated space is used.

EMC Symmetrix Virtual Provisioning (which will be referred to as virtual provisioning throughout the rest of the document) can address both of these issues. Virtual Provisioning allows more storage to be presented to an application than is physically available. More importantly, Virtual Provisioning will allocate physical storage only when the storage is actually written to or when purposively pre-allocated by an administrator. This allows more flexibility in predicting future growth, reduces the initial costs of provisioning storage to an application, and can obviate the inherent waste in over-allocation of space and administrative management of subsequent storage allocations.

EMC Symmetrix VMAX storage arrays continue to provide increased storage utilization and optimization, enhanced capabilities, and greater interoperability and security. The implementation of Virtual Provisioning, generally known in the industry as “thin provisioning,” for these arrays enables organizations to improve ease of use, enhance performance, and increase capacity utilization for certain applications and workloads. It is important to note that the Symmetrix VMAXe/VMAX 10K storage array is designed and configured for a one-hundred percent virtually provisioned environment.

EMC Symmetrix Virtual Provisioning overview

Symmetrix thin devices are logical devices that can be used in many of the same ways that Symmetrix devices have traditionally been used. Unlike standard Symmetrix devices, thin devices do not need to have physical storage completely allocated at the time the device is created and presented to a host. A thin device is not usable until it has been bound to a shared storage pool known as a thin pool. The thin pool is comprised of special devices known as data devices that provide the actual physical storage to support the thin device allocations.

When a write is performed to part of any thin device for which physical storage has not yet been allocated, the Symmetrix allocates physical storage from the thin pool for that portion of the thin device only. The Symmetrix operating environment, Enginuity, satisfies the requirement by providing a unit of physical storage from the thin pool called a thin device extent. This approach reduces the amount of storage that is actually consumed.

The thin device extent is the minimum amount of physical storage that can be reserved at a time for the dedicated use of a thin device. An entire thin device extent is physically allocated to the thin device at the time the thin storage allocation is made as a result of a host write operation. A round-robin mechanism is used to balance the allocation of data device extents across all of the data devices in the thin pool that are enabled and that have remaining unused capacity. The thin device extent size is twelve 64 KB tracks (768 KB). Note that the initial bind of a thin device to a pool causes one thin device extent to be allocated per thin device. If the thin device is a metavolume, then one thin device extent is allocated per metavolume member. So a four-member thin metavolume would cause four extents (3,072 KB) to be allocated when the device is initially bound to a thin pool.

When a read is performed on a thin device, the data being read is retrieved from the appropriate data device in the thin pool. If a read is performed against an unallocated portion of the thin device, zeros are returned.

When more physical data storage is required to service existing or future thin devices, such as when a thin pool is running out of physical space, data devices can be added to existing thin pools online. Users can rebalance extents over the thin pool when any new data devices are added. This feature is discussed in detail in the next section. New thin devices can also be created and bound to existing thin pools.

When data devices are added to a thin pool they can be in an “enabled” or “disabled” state. In order for the data device to be used for thin extent allocation, however, it needs to be “enabled”. Conversely, for it to be removed from the thin pool, it needs to be in a “disabled” state. Active data devices can be disabled, which will cause any allocated extents to be drained and reallocated to the other enabled devices in the pool. They can then be removed from the pool when the drain operation has completed.

Figure 71 depicts the relationships between thin devices and their associated thin pools. There are nine devices associated with thin Pool A and three thin devices associated with thin pool B.

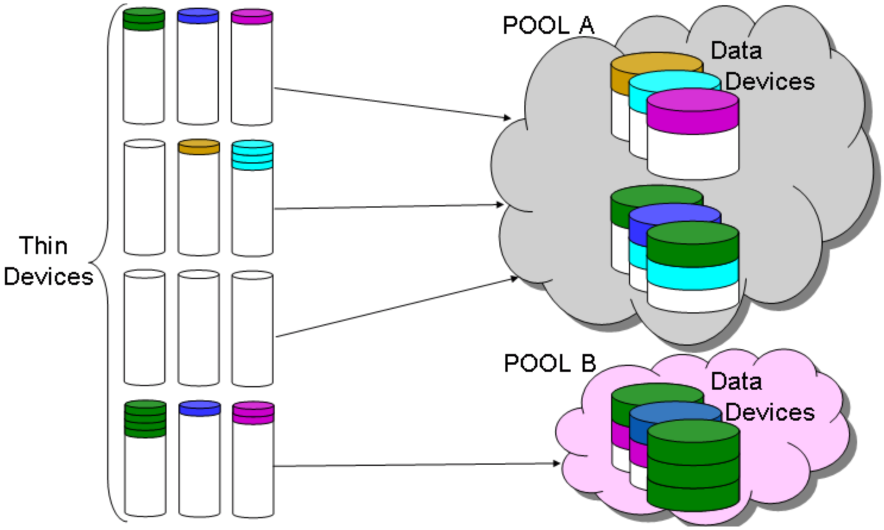


Figure 71 Thin devices and thin pools containing data devices

The way thin extents are allocated across the data devices results in a form of striping in the thin pool. The more data devices that are in the thin pool, the wider the striping is, and therefore the greater the number of devices available to service I/O.

Thin device sizing

The maximum size of a thin device on a Symmetrix VMAX is approximately 240 GB. If a larger size is needed, then a metavolume comprised of multiple thin devices can be created. For environments using 5874 and earlier, it is recommended that thin metavolumes be concatenated rather than striped since concatenated metavolumes support fast expansion capabilities, as new metavolume members can easily be added to the existing concatenated metavolume. This functionality is required when the provisioned thin device has been completely consumed by the host, and further storage allocations are required. Beginning with Enginuity 5875, striped metavolumes can be expanded to larger sizes online. Therefore, EMC recommends that customers utilize striped thin metavolumes with the 5875 Enginuity level and later.

Concatenated metavolumes using standard volumes have the drawback of being limited by the performance of a single metavolume member. In the case of a concatenated metavolume comprised of thin devices, however, each member device is typically spread across the entire underlying thin pool, thus somewhat eliminating that drawback.

However, there are several conditions in which the performance of concatenated thin metavolumes can be limited. Specifically:

- Enginuity allows one outstanding write per thin device per path with Synchronous SRDF. With concatenated metavolumes, this could cause a performance problem by limiting the concurrency of writes. This limit will not affect striped metavolumes in the same way because of the small size of the metavolume stripe (one cylinder or 1,920 blocks).
- Symmetrix Enginuity has a logical volume write-pending limit to prevent one volume from monopolizing writable cache. Because each metavolume member gets a small percentage of cache, a striped metavolume is likely to offer more writable cache to the metavolume.

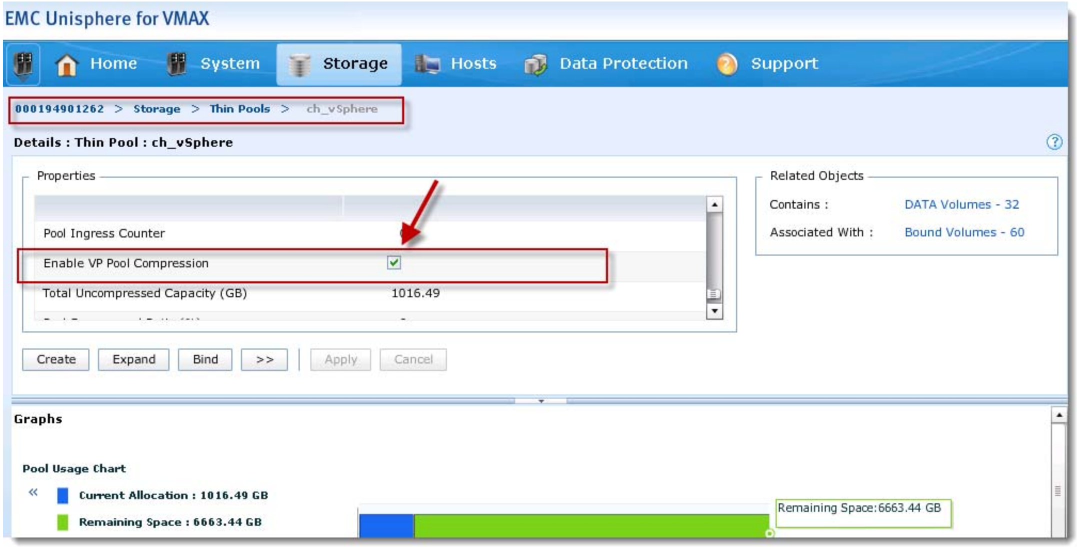
If changes in metavolume configuration is desired, Solutions Enabler 7.3 and 5875.198 and later support converting a concatenated thin metavolume into a striped thin metavolume while protecting the data using a thin or thick BCV device (if thin it must be persistently allocated). Prior to these software versions, protected metavolume conversion was not allowed.

Thin device compression

Beginning with Enginuity release 5876.159.102 and Solutions Enabler 7.5, data within a thin device can be compressed/decompressed to save space.

It is important to note that users should leverage compression only if the hosted application can sustain a slightly elevated response time when reading from, or writing to, compressed data. If the application is sensitive to heightened response times it is recommended to not compress the data. Preferably, data compression should be primarily used for data that is completely inactive and is expected to remain that way for an elongated period of time. If a thin device (one that hosts a VMFS volume or is in use as an RDM) is presently compressed and an imminent I/O workload is anticipated that will require significant decompression to commit new writes, it is suggested to manually decompress the volume first thereby avoiding the decompression latency penalty.

To compress a thin device, the entire bound pool must be enabled for compression. This can be achieved by setting the compression capability on a given pool to “enabled” in Unisphere for VMAX or through the Solutions Enabler CLI. An example of this process with Unisphere can be seen in Figure 72. When compression is no longer desired for allocations in a pool, the compression capability on a pool can be disabled.



000194901262 > Storage > Thin Pools > ch_vSphere

Details : Thin Pool : ch_vSphere

Pool Ingress Counter	
Enable VP Pool Compression	☑
Total Uncompressed Capacity (GB)	1016.49

Contains : DATA Volumes - 32

Associated With : Bound Volumes - 60

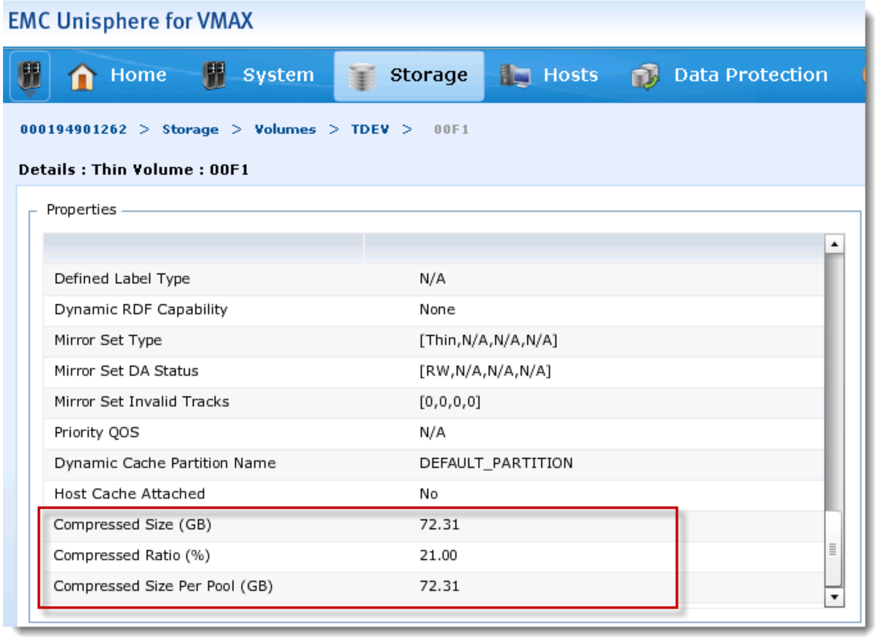
🟦 Current Allocation : 1016.49 GB

🟩 Remaining Space : 6663.44 GB

Remaining Space:6663.44 GB

Figure 72 Enabling thin device compression on a thin pool

Note that once a pool is enabled for compression, a quick background process will run to reserve storage in the pool that will be used to temporarily decompress data (called the Decompressed Read Queue or DRQ). When compression is disabled, the reserved storage in the pool will be returned to the pool only after all compressed allocations are decompressed. Disabling the compression capability on a pool will not automatically cause compressed allocations to be decompressed. This is an action that must be executed by the user. Solutions Enabler or Unisphere will report the current state of the compression capability on a pool, the amount of compressed allocations in the pool, and which thin devices have compressed allocations. Figure 73 on page 140 shows a thin device with a compression ratio of 21% in Unisphere for VMAX.



000194901262 > Storage > Volumes > TDEV > 00F1

Details : Thin Volume : 00F1

Properties

Defined Label Type	N/A
Dynamic RDF Capability	None
Mirror Set Type	[Thin,N/A,N/A,N/A]
Mirror Set DA Status	[RW,N/A,N/A,N/A]
Mirror Set Invalid Tracks	[0,0,0,0]
Priority QOS	N/A
Dynamic Cache Partition Name	DEFAULT_PARTITION
Host Cache Attached	No
Compressed Size (GB)	72.31
Compressed Ratio (%)	21.00
Compressed Size Per Pool (GB)	72.31

Figure 73 Viewing thin device-specific compression information

When a user chooses to compress a device, only allocations that have been written will be compressed. Allocations that are not written to or fully-zeroed will be reclaimed during compression if the tracks are not persistently allocated. In effect, the compression process also performs a zero-space reclamation operation while analyzing for compressible allocations. This implies that after a decompression the allocated tracks will not match the original allocated track count for the TDEV. If this behavior is undesirable, the user can make use of the persistent allocation feature, and then the allocated tracks will be maintained and no compression will take place on persistently allocated tracks. If allocations are persistent and a reclaim of space is in fact desired, then the persistent attribute for existing allocations may be cleared to allow compression and reclamation to occur.

Note: If a thin device has allocations on multiple pools, only the allocations on compression-enabled pools will be compressed.

Note that compression of allocations for a thin device is a low-priority background task. Once the compression request is accepted, the thin device will have a low-priority background task associated with it that performs the compression. While this task is running, no other background task, like allocate or reclaim, can be run against the thin device until it completes or is terminated by the user.

A read of a compressed track will temporarily uncompress the track into the DRQ maintained in the pool. The space in the DRQ is controlled by a least recently used algorithm ensuring that a track can always be decompressed and that the most recently utilized tracks will still be available in an decompressed form. Writes to compressed allocations will always cause decompression.

FAST VP and thin device compression

Over time if the data has not been accessed, compression may occur if the device is under FAST VP management and FAST VP is supporting compression on the association. The FAST engine will be able to treat compression as a sub-tier, for instance demoting from SATA to SATA-Compressed (within the same actual tier), likewise promoting from SATA-compressed to SATA or higher.

Performance considerations

The architecture of Virtual Provisioning creates a naturally striped environment where the thin extents are allocated across all devices in the assigned storage pool. The larger the storage pool for the allocations is, the greater the number of devices that can be leveraged for VMware vSphere I/O.

One of the possible consequences of using a large pool of data devices that is shared with other applications is variability in performance. In other words, possible contention with other applications for physical disk resources may cause inconsistent performance levels. If this variability is not desirable for a particular application, that application could be dedicated to its own thin pool. Symmetrix arrays support up to 512 pools. Pools in this instance include Virtual Provisioning thin pools, SRDF/A Delta Set Extension (DSE) pools, and TimeFinder/Snap pools. If the array is running Enginuity 5875, FAST VP should be enabled to automatically uncover and remediate performance issues by moving the right data to the right performance tier.

When a new data extent is required from the thin pool there is an additional latency introduced to the write I/O while the thin extent location is assigned and formatted. This latency is approximately one millisecond. Thus, for a sequential write stream, a new extent allocation will occur on the first new allocation, and again when the current stripe has been fully written and the writes are moved to a new extent. If the application cannot tolerate the additional latency it is recommended to preallocate some storage to the thin device when the thin device is bound to the thin pool.[1]

[1] If using VASA please see "vSphere vStorage APIs for Storage Awareness (VASA)" in Chapter 2 for a CAUTION concerning preallocation.

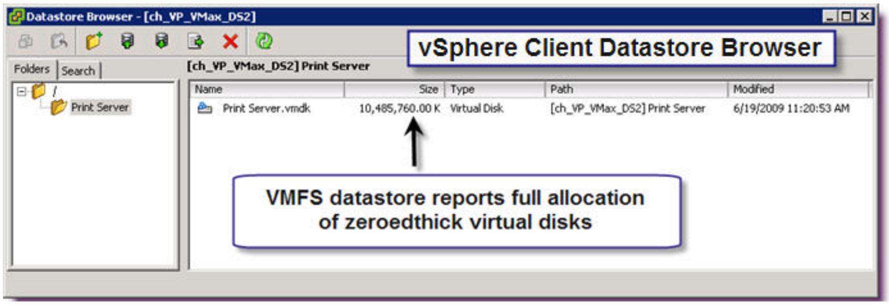
Virtual disk allocation schemes

VMware vSphere offers multiple ways of formatting virtual disks and has integrated these options into VMware vCenter. For new virtual machine creation, only the formats eagerzeroedthick, zeroedthick, and thin are offered as options.

“Zeroedthick” allocation format

When creating, cloning, or converting virtual disks in the vSphere Client, the default option is called “Thick Provision Lazy Zeroed” in vSphere 5 and simply “Thick” in vSphere 4. The “Thick Provision Lazy Zeroed” or “Thick” selection is actually the “zeroedthick” format. In this allocation scheme, the storage required for the virtual disks is reserved in the datastore but the VMware kernel does not initialize all the blocks. The blocks are initialized by the guest operating system as write activities to previously uninitialized blocks are performed. The VMFS will return zeros to the guest operating system if it attempts to read blocks of data that it has not previously written to. This is true even in cases where information from a previous allocation (data “deleted” on the host, but not de-allocated on the thin pool) is available. The VMFS will not present stale data to the guest operating system when the virtual disk is created using the “zeroedthick” format.

Since the VMFS volume will report the virtual disk as fully allocated, the risk of oversubscribing is reduced. This is due to the fact that the oversubscription does not occur on both the VMware layer and the Symmetrix layer. The virtual disks will not require more space on the VMFS volume as their reserved size is static with this allocation mechanism and more space will be needed only if additional virtual disks are added. Therefore, the only free capacity that must be monitored diligently is the free capacity on the thin pool itself. This is shown in Figure 74, which displays a single 10 GB virtual disk that uses the “zeroedthick” allocation method. The datastore browser reports the virtual disk as consuming the full 10 GB.



vSphere Client Datastore Browser

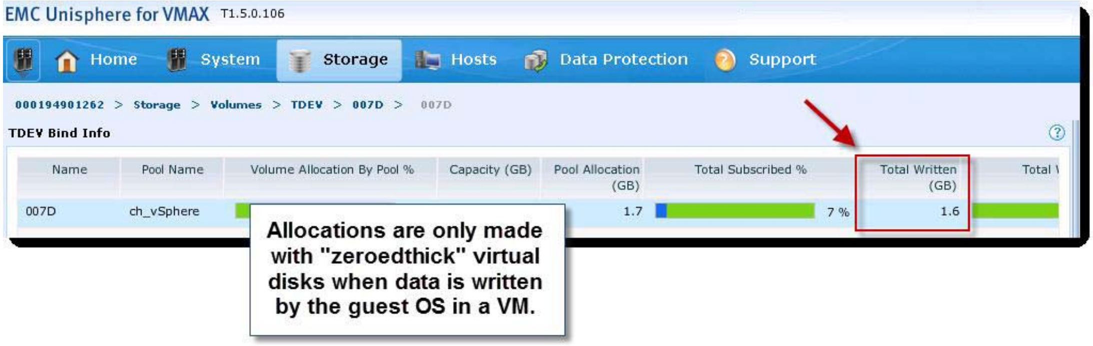
[ch_VP_VMax_DS2] Print Server

Name	Size	Type	Path	Modified
Print Server.vmdk	10,485,760.00 K	Virtual Disk	[ch_VP_VMax_DS2] Print Server	6/19/2009 11:20:53 AM

VMFS datastore reports full allocation of zeroedthick virtual disks

Figure 74 **Zeroedthick virtual disk allocation size as seen in a VMFS datastore browser**

However, since the VMware kernel does not actually initialize unused blocks, the full 10 GB is not consumed on the thin device. In the example, the virtual disk resides on thin device 007D and, as seen in Figure 75, only consumes about 1,600 MB of space on it.


EMC Unisphere for VMAX T1.5.0.106
Home
System
Storage
Hosts
Data Protection
Support
000194901262 > Storage > Volumes > TDEV > 007D > 007D

TDEV Bind Info

Name	Pool Name	Volume Allocation By Pool %	Capacity (GB)	Pool Allocation (GB)	Total Subscribed %	Total Written (GB)	Total \
007D	ch_vSphere			1.7	7 %	1.6	

Allocations are only made with "zeroedthick" virtual disks when data is written by the guest OS in a VM.

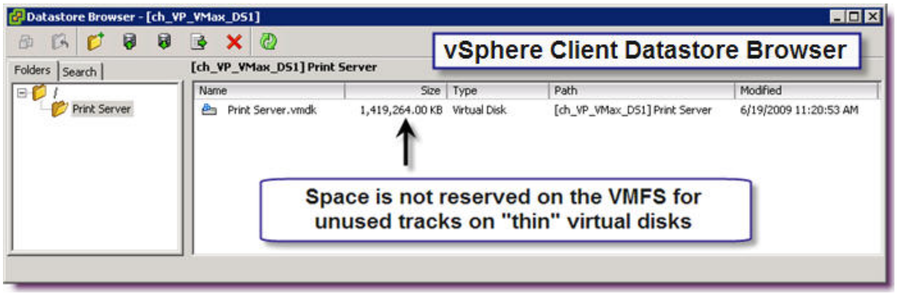
Figure 75 **Zeroedthick virtual disk allocation on Symmetrix thin devices**

“Thin” allocation format

Like zeroedthick, the “thin” allocation mechanism is also Virtual Provisioning-friendly, but, as will be explained in this section, should be used with caution in conjunction with Symmetrix Virtual Provisioning. “Thin” virtual disks increase the efficiency of storage utilization for virtualization environments by using only the amount of underlying storage resources needed for that virtual disk, exactly like zeroedthick. But unlike zeroedthick, thin devices do not reserve space on the VMFS volume-allowing more virtual disks per VMFS. Upon the initial provisioning of the virtual disk, the disk is provided with an allocation equal to one block size worth of storage from the datastore. As that space is filled, additional chunks of storage in multiples of the VMFS block size are allocated for the virtual disk so that the underlying storage demand will grow as its size increases.

It is important to note that this allocation size depends on the block size of the target VMFS. For VMFS-3, the default block size is 1 MB but can also be 2, 4, or 8 MB. Therefore, if the VMFS block size is 8 MB, then the initial allocation of a new “thin” virtual disk will be 8 MB and will be subsequently increased in size by 8 MB chunks. For VMFS-5, the block size is not configurable and therefore is, with one exception, always 1 MB. The sole exception is if the VMFS-5 volume was upgraded from a VMFS-3 volume. In this case, the block size will be inherited from the original VMFS volume and therefore could be 1, 2, 4 or 8 MB.

A single 10 GB virtual disk (with only 1.6 GB of actual data on it) that uses the thin allocation method is shown in Figure 76. The datastore browser reports the virtual disk as consuming only 1.6 GB on the volume.

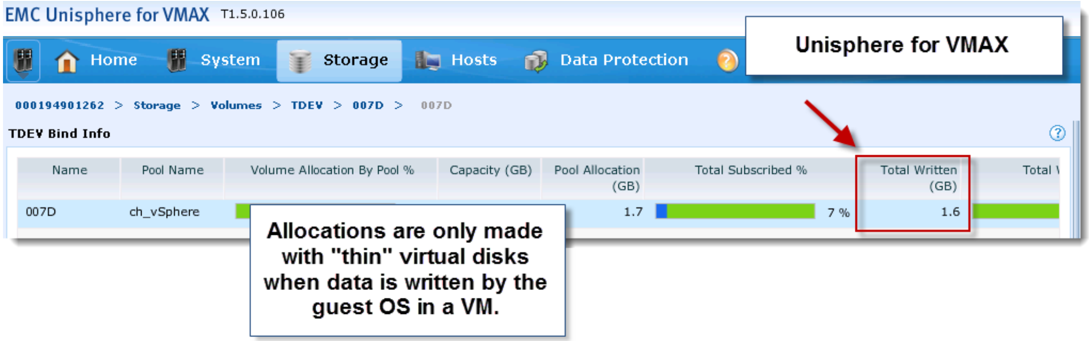

Datastore Browser - [ch_VP_VMax_DS1]
vSphere Client Datastore Browser
Folders
Search
/
Print Server
Space is not reserved on the VMFS for unused tracks on "thin" virtual disks

[ch_VP_VMax_DS1] Print Server

Name	Size	Type	Path	Modified
Print Server.vmdk	1,419,264.00 KB	Virtual Disk	[ch_VP_VMax_DS1] Print Server	6/19/2009 11:20:53 AM

Figure 76 Thin virtual disk allocation size as seen in a VMFS datastore browser

As is the case with the “zeroedthick” allocation format, the VMware kernel does not actually initialize unused blocks for “thin” virtual disks, so the full 10 GB is neither reserved on the VMFS nor consumed on the thin device. The virtual disk presented in Figure 76 on page 145 resides on thin device 007D and, as seen in Figure 77, only consumes about 1,600 MB of space on it.


EMC Unisphere for VMAX T1.5.0.106
Home
System
Storage
Hosts
Data Protection
000194901262 > Storage > Volumes > TDEV > 007D > 007D

Unisphere for VMAX

TDEV Bind Info

Name	Pool Name	Volume Allocation By Pool %	Capacity (GB)	Pool Allocation (GB)	Total Subscribed %	Total Written (GB)	Total \
007D	ch_vSphere			1.7	7 %	1.6	

Allocations are only made with "thin" virtual disks when data is written by the guest OS in a VM.

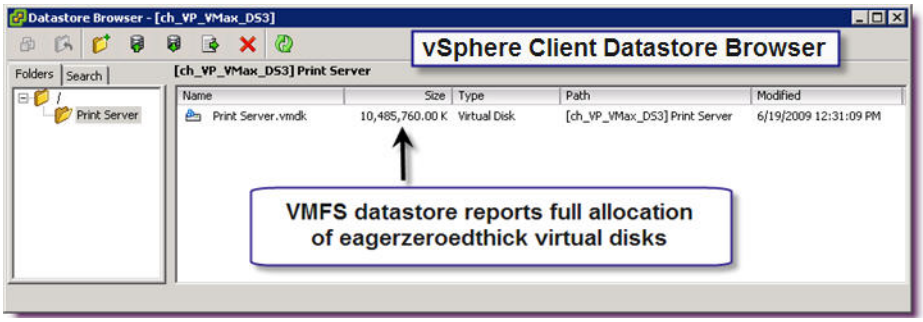
Figure 77 Thin virtual disk allocation on Symmetrix thin devices

“Eagerzeroedthick” allocation format

With the “eagerzeroedthick” allocation mechanism (referred to as “Thick Provision Eager Zeroed” in the vSphere Client in vSphere 5), space required for the virtual disk is completely allocated and written to at creation time. This leads to a full reservation of space on the VMFS datastore and on the underlying Symmetrix device. Accordingly, it takes longer to create disks in this format than to create other types of disks.[1] Because of this behavior, “eagerzeroedthick” format is not ideal for use with virtually provisioned devices.

[1] This creation, however, is considerably quicker when the VAAI primitive, Block Zero, is in use. See““Eagerzeroedthick” with ESX 4.1+ and Enginuity 5875” on page 147 for more detail.

A single 10 GB virtual disk (with only 1.6 GB of actual data on it) that uses the “eagerzeroedthick” allocation method is shown in Figure 78. The datastore browser reports the virtual disk as consuming the entire 10 GB on the volume.



Datastore Browser - [ch_VP_VMax_DS3]

vSphere Client Datastore Browser

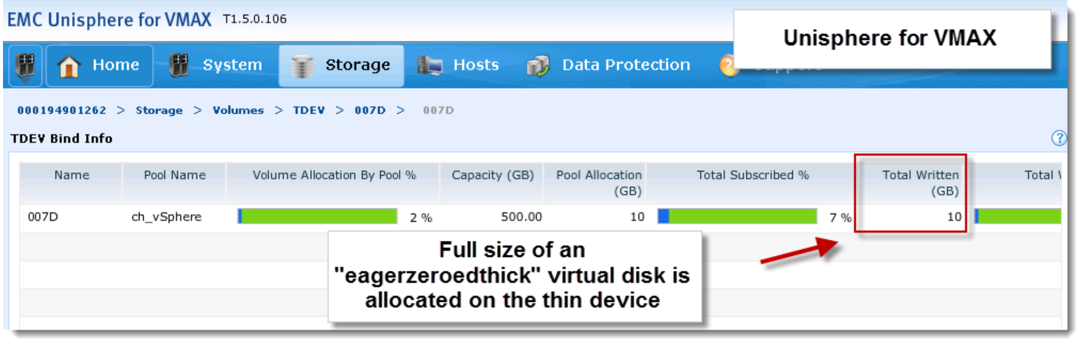
[ch_VP_VMax_DS3] Print Server

Name	Size	Type	Path	Modified
Print Server.vmdk	10,485,760.00 K	Virtual Disk	[ch_VP_VMax_DS3] Print Server	6/19/2009 12:31:09 PM

VMFS datastore reports full allocation of eagerzeroedthick virtual disks

Figure 78 Eagerzeroedthick virtual disk allocation size as seen in a VMFS datastore browser

Unlike with “zeroedthick” and “thin”, the VMware kernel initializes all the unused blocks, so the full 10 GB is reserved on the VMFS and consumed on the thin device. The “eagerzeroedthick” virtual disk resides on thin device 0484 and, as highlighted in Figure 79, consumes 10 GB of space on it.



EMC Unisphere for VMAX T1.5.0.106

Unisphere for VMAX

000194901262 > Storage > Volumes > TDEV > 007D > 007D

TDEV Bind Info

Name	Pool Name	Volume Allocation By Pool %	Capacity (GB)	Pool Allocation (GB)	Total Subscribed %	Total Written (GB)	Total \
007D	ch_vSphere	2 %	500.00	10	7 %	10	

Full size of an "eagerzeroedthick" virtual disk is allocated on the thin device

Figure 79 Eagerzeroedthick virtual disk allocation on Symmetrix thin devices

“Eagerzeroedthick” with ESX 4.1+ and Enginuity 5875

VMware vSphere 4.1/5.x offers a variety of VMware vStorage APIs for Array Integration (VAAI) that provides the capability to offload specific storage operations to EMC Symmetrix VMAX/VMAXe to increase both overall system performance and efficiency. Symmetrix VMAX/VMAXe with Enginuity 5875 supports the following of VMware's new vStorage APIs - Full Copy and Block Zero and Hardware-Assisted Locking.

Without the block zeroing primitive, the virtual disk creation is not complete until the host has completed the zeroing process. Consequently, for a very large disk this can take a long time. The block zeroing primitive, which uses the WRITE SAME SCSI command (0x93), enables the disk array to return the cursor to the requesting service as though the process of writing the zeros has been completed and then finish the job of zeroing out those blocks in the background on the array.

This primitive has a profound effect on Symmetrix Virtual Provisioning when deploying "eagerzeroedthick" virtual disks. When this feature is enabled, combined with Symmetrix Virtual Provisioning, the typical host zeroing that occurs when deploying disks using this allocation mechanism, will be offloaded to the array. Furthermore, the Symmetrix will not write the zeros to disk but instead simply set the track as "Not Written By Host" and allocate the respective tracks on the thin pool. If desired, these allocated, but unwritten, tracks can be de-allocated using Symmetrix Management Console or Solutions Enabler. Refer to the *Solutions Enabler Controls Product Guide* or the Symmetrix Management Console online help for more detail.

Block Zero is enabled by default on both the Symmetrix VMAX family arrays running 5875 Enginuity or later, and on a properly licensed ESX(i), 4.1 or later, host. Block Zero can be disabled on an ESX host if desired through the use of the vSphere Client or command line utilities. Block Zero can be disabled or enabled by altering the setting DataMover.HardwareAcceleratedInit in the ESX host advanced settings under DataMover.

Virtual disk type recommendations

In general, before vSphere 4.1, EMC recommended using "zeroedthick" instead of "thin" virtual disks when using Symmetrix Virtual Provisioning. The reason that "thin on thin" was not always recommended is that using thin provisioning on two separate layers (host and array) increases the risk of out-of-space conditions for virtual machines. vSphere 4.1 (and even more so in vSphere 5) in conjunction with the latest release of Enginuity integrate these layers better than ever. As a result, using thin on thin is now acceptable in far more situations, but it is important to remember that risks still remain in doing so.

For this reason, EMC recommends “zeroedthick” (for better reliability as only the array must be monitored) or “eagerzeroedthick” (for absolute reliability as all space is completely reserved) for mission critical virtual machines.

The relative performance, protection against capacity exhaustion and efficiency between the three types of virtual disks are show in Figure 80 on page 150.

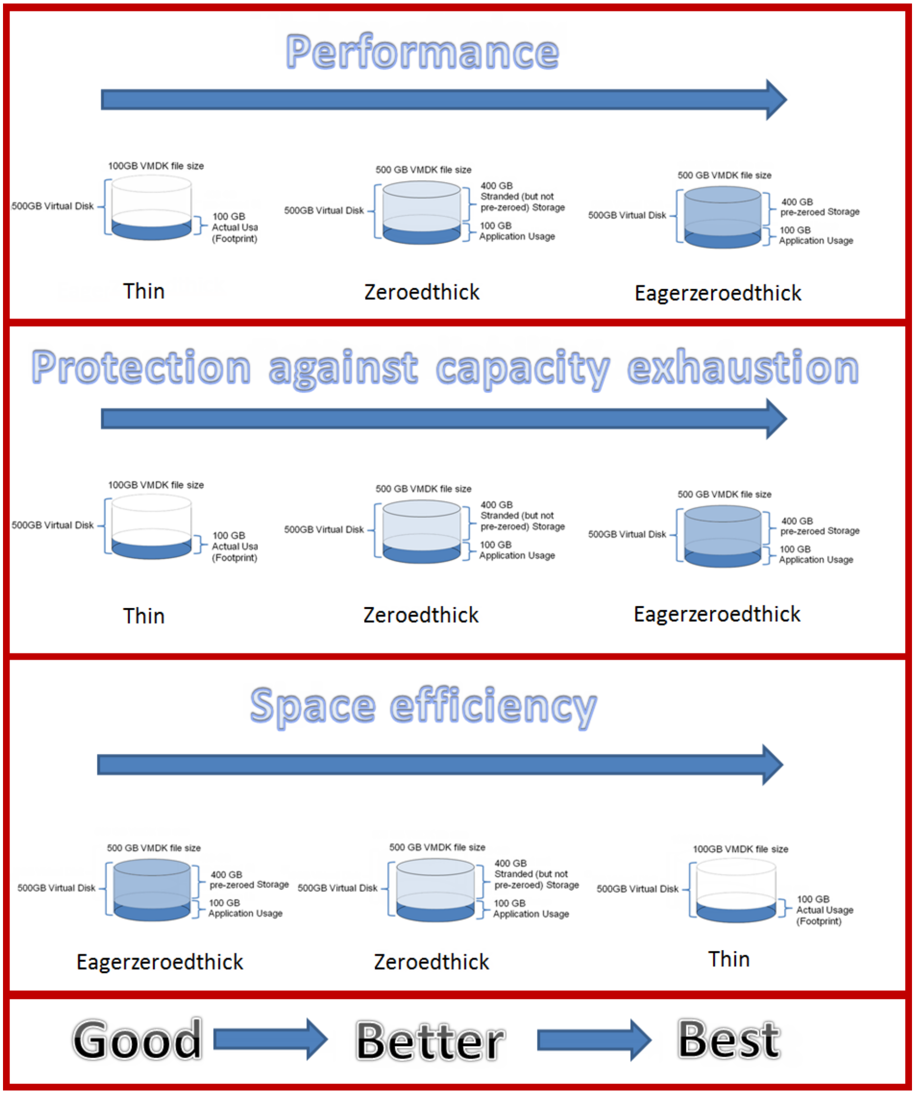


Figure 80 Virtual disk comparison

It is important to remember that using “thin” rather than “zeroedthick” virtual disks does not provide increased space savings on the physical disks. As previously discussed, since “zeroedthick” only writes data written by the guest OS, it does not consume more space on the Symmetrix thin pool than a similarly sized "thin" virtual disk.

The primary difference between "zeroedthick" and "thin" virtual disks is not a question of space, but a question of the quantity and capacity of the Symmetrix thin devices presented to the ESX server. Due to the architecture of Virtual Provisioning, binding more thin devices to a thin pool or increasing the size of them does not take up more physical disk space than fewer or smaller thin devices.

So whether 10 small thin devices are presented to an ESX host, or one large thin device, the storage usage on the Symmetrix is the same. Therefore, packing more virtual machines into smaller or fewer thin devices is not necessarily more space efficient. Prior to ESXi 5, a single-extent VMFS volume was limited to approximately 2 TB and this led customers to use "thin on thin" to keep the number of devices presented to a host minimal as a way to ease management.

"Thin" virtual disks allowed for fewer presented devices and a higher virtual machine density. Now with the larger allowable single-extent VMFS volume size in ESXi 5 (up to 64 TB) the need to add more thin devices is reduced as larger, more appropriately sized thin devices can be presented and used and expanded as necessary. With ESXi 5, customers therefore do not need to make the virtual disks "thin" in order to fit a large number of them on a single volume as it is no longer constrained by the 2 TB minus 512 byte limit.

However, as previously mentioned, VMware and EMC have recently ameliorated the practicality of using thin virtual disks with Symmetrix Virtual Provisioning. This has been achieved through:

1. Refining vCenter and array reporting while also widening the breadth of direct Symmetrix integration.

2. Reducing the slight performance overhead that existed with thin VMDKs. This was caused by zeroing-on-demand and intermittent expansion of the virtual disk as new blocks were written to by the guest OS. It has been respectively significantly diminished with the advent of Block Zero and Hardware-Assisted Locking (ATS).

3. The introduction of Storage DRS in vSphere 5 (SDRS[1]) further reduces the risk of running out of space on a VMFS volume as it can be configured to migrate virtual machines from a datastore when that datastore reaches a user-specified percent-full

threshold. This all but eliminates the risk of running out of space on VMFS volumes, with the only assumption being that there is available capacity elsewhere to move the virtual machines to.

These improvements can make “thin on thin” a much more viable option-especially in vSphere 5. Essentially, it depends on how risk-averse an organization is and the importance/priority of an application running in a virtual machine. If virtual machine density and storage efficiency is valued above the added protection provided by a thick virtual disk, “thin on thin” may be used. If “thin on thin” is used, alerts on the vCenter level and the array level should be configured.

[1] See “VMware Storage Distributed Resource Scheduler (SDRS)” on page 205 for more detail.

Dead Space Reclamation (UNMAP)

Referred to as Dead Space Reclamation by VMware, ESXi 5.0 U1 introduced support for the SCSI UNMAP command (0x42) which issues requests to de-allocate extents on a thin device. Dead Space Reclamation offers the ability to reclaim blocks of a thin-provisioned LUN on the array via a VMware-provided command line utility.

It is important to note that the Symmetrix only supports this functionality with Enginuity 5876.159 and higher. If a customer is running Enginuity 5876.82 or lower, they must therefore upgrade to 5876.159 to leverage this functionality or utilize an alternative, manual process to reclaim dead space[1].

A Dead Space Reclamation should be executed after a virtual disk is deleted or is migrated to a different datastore. As shown in the previous section, when virtual machines were migrated from a datastore, the blocks used by the virtual machine prior to the migration were still being reported as “in use” by the array as shown in Figure 10. With this new VMware Storage API for Array Integration primitive, the Symmetrix will be informed that the blocks are no longer in use, resulting in better reporting of disk space consumption and a much more efficient use of resources.

To confirm that the Enginuity level is correct via the ESXi host and that UNMAP is in fact supported on a LUN, the following command can be issued to query the device:

```
esxcli storage core device vaai status get -d <naa>
```

Example output:

```
naa.<naa>

       VAAI Plugin Name: VMW_VAAI_SYMM

       ATS Status: supported

       Clone Status: supported

       Zero Status: supported

       Delete Status: supported
```

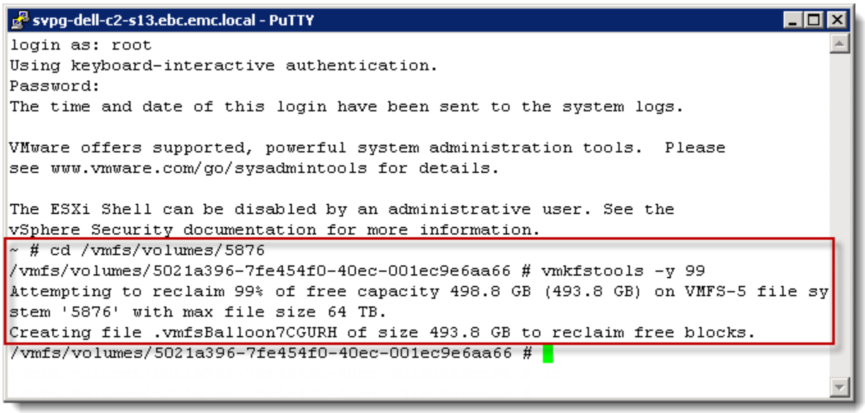
[1] Please refer to the white paper, Implementing Symmetrix Virtual Provisioning with VMware vSphere for details on the process to manually reclaim space in environments that do not support UNMAP.

A device displaying `Delete Status` as supported means that the device is capable of receiving SCSI UNMAP commands or, in other words, the Symmetrix is running the correct level of Enginuity to support UNMAP from an ESXi host.

In order to reclaim space, ESXi 5.0 Update 1 includes an updated version of `vmkfstools` that provides an option (`-y`) to issue the UNMAP operation. This command should be issued against the device after navigating to the root directory of the VMFS volume residing on it:

```
cd /vmfs/volumes/<volume-name>

vmkfstools -y <percentage of deleted block to
reclaim>
```

This command creates temporary files at the top level of the datastore. These files can be as large as the aggregate size of blocks being reclaimed as seen in Figure 81. If the reclaim operation is somehow interrupted, these temporary files may not be deleted automatically and users may need to delete them manually.



```
login as: root
Using keyboard-interactive authentication.
Password:
The time and date of this login have been sent to the system logs.

VMware offers supported, powerful system administration tools.  Please
see www.vmware.com/go/sysadmintools for details.

The ESXi Shell can be disabled by an administrative user. See the
vSphere Security documentation for more information.
~ # cd /vmfs/volumes/5876
/vmfs/volumes/5021a396-7fe454f0-40ec-001ec9e6aa66 # vmkfstools -y 99
Attempting to reclaim 99% of free capacity 498.8 GB (493.8 GB) on VMFS-5 file sy
stem '5876' with max file size 64 TB.
Creating file .vmfsBalloon7CGURH of size 493.8 GB to reclaim free blocks.
/vmfs/volumes/5021a396-7fe454f0-40ec-001ec9e6aa66 #
```

Figure 81 SSH session to ESXi server to reclaim dead space with vmkfstools

There is no practical way of knowing how long a `vmkfstools -y` operation will take to complete. It can run for a few minutes or even for a few hours depending on the size of the datastore, the amount of dead space, and the available resources on the array to execute the operation. Reclamation on the Symmetrix is a low-priority task (in order for it not to negatively affect host I/O performance) so a heavy workload on the device, pool or the array itself can throttle the process and slow the reclamation.

Figure 82 contains some baseline results[1] of using the UNMAP command. Two separate reclaims are shown together, one at 60% and another at 99%. The results demonstrate that the time it takes to reclaim space increases fairly consistently as the amount of space increases. For instance, to reclaim 150 GB at 99% takes 100 seconds and to reclaim 300 GB at 99% takes 202 seconds, almost exactly double the time for double the size. The results for 60% are similar.

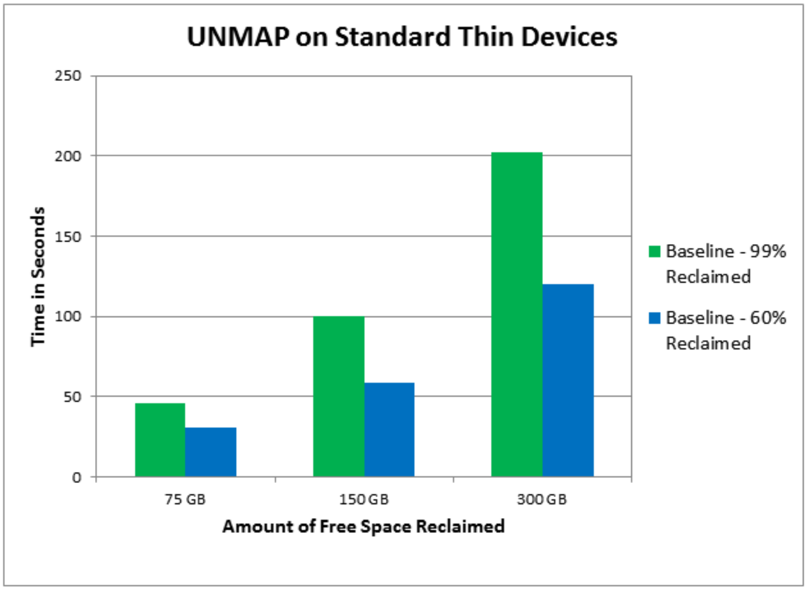


Figure 82 **UNMAP results on standard thin devices**

Note: These results are only meant as examples of reclaim operations and are dependent upon the lab environment under which they were generated. Although every effort is made to mimic the typical activity in a customer environment it is important to understand that results will vary.

[1] Baseline in this case means no other relationships such as TimeFinder/Clone or SRDF were on the devices. Please refer to the following white paper for more UNMAP results: Using VMware vSphere Storage APIs for Array Integration with EMC Symmetrix.

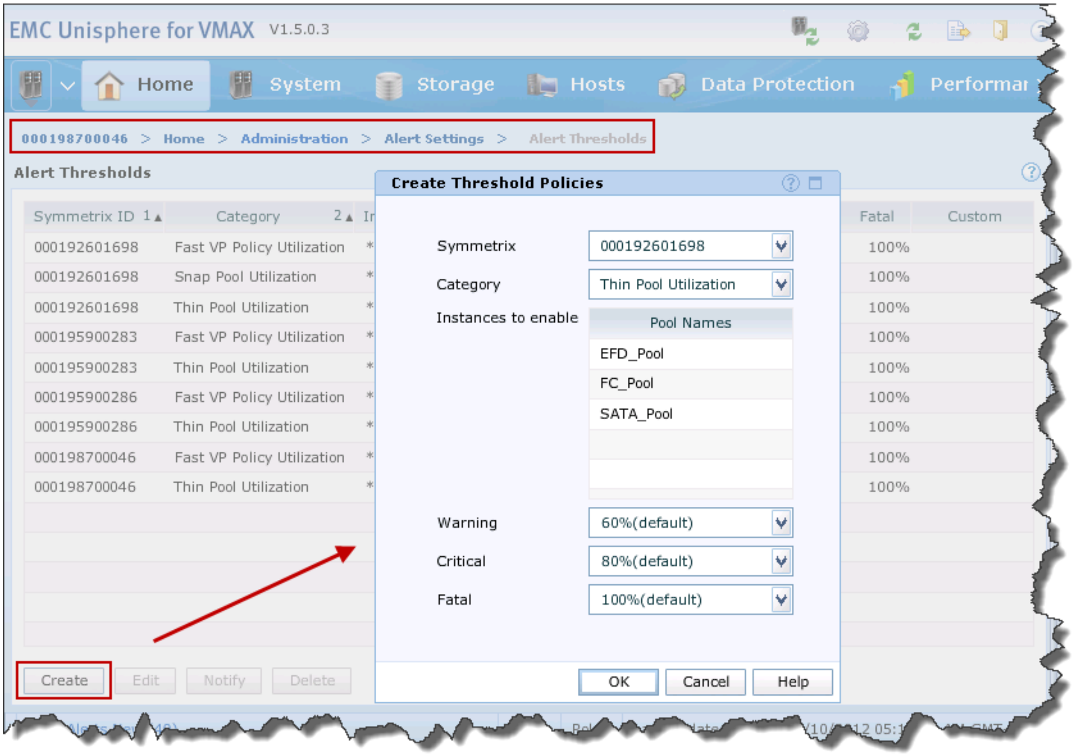
Thin pool management

When storage is provisioned from a thin pool to support multiple thin devices there is usually more “virtual” storage provisioned to hosts than is supported by the underlying data devices. This condition is known as oversubscription. This is one of the main reasons for using Virtual Provisioning. However, there is a possibility that applications using a thin pool may grow rapidly and request more storage capacity from the thin pool than is actually there. This section discusses the steps necessary to avoid out-of-space conditions.

Capacity monitoring

There are several methodologies to monitor the capacity consumption of the Symmetrix thin pools. The Solutions Enabler `symcfg monitor` command can be used to monitor pool utilization, as well as display the current allocations through the `symcfg show` pool command.

There are also event thresholds that can be monitored through the SYMAPI event daemon, thresholds (defaults shown) that can be set with the Unisphere for VMAX as shown in Figure 83 on page 157, and SNMP traps that can be sent for monitoring by EMC Ionix ControlCenter or any data center management product.


EMC Unisphere for VMAX V1.5.0.3
Home
System
Storage
Hosts
Data Protection
Performar
000198700046 > Home > Administration > Alert Settings > Alert Thresholds
Alert Thresholds
Create Threshold Policies
Create
Edit
Notify
Delete
OK
Cancel
Help

Symmetrix ID 1▲	Category 2▲	Ir	Fatal	Custom
000192601698	Fast VP Policy Utilization	*	100%	
000192601698	Snap Pool Utilization	*	100%	
000192601698	Thin Pool Utilization	*	100%	
000195900283	Fast VP Policy Utilization	*	100%	
000195900283	Thin Pool Utilization	*	100%	
000195900286	Fast VP Policy Utilization	*	100%	
000195900286	Thin Pool Utilization	*	100%	
000198700046	Fast VP Policy Utilization	*	100%	
000198700046	Thin Pool Utilization	*	100%	

Symmetrix: 000192601698

Category: Thin Pool Utilization

Instances to enable:

Pool Names
EFD_Pool
FC_Pool
SATA_Pool

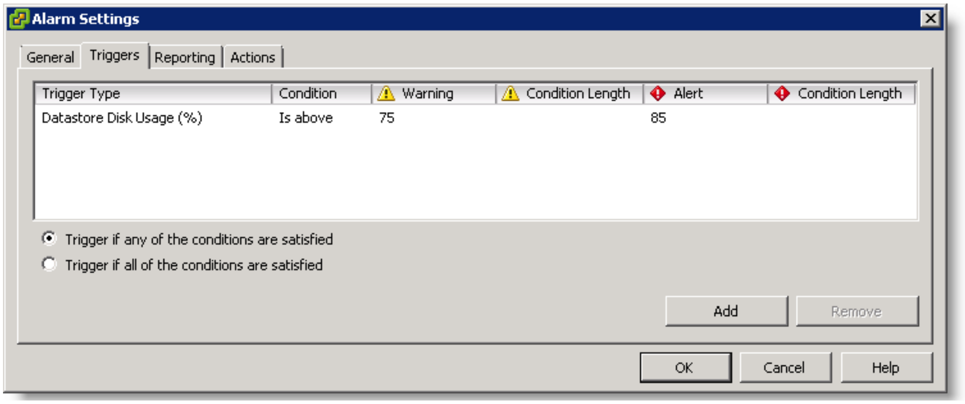
Warning: 60%(default)

Critical: 80%(default)

Fatal: 100%(default)

Figure 83 Setting Pool Utilization Thresholds in Unisphere for VMAX

VMFS volume capacity can be monitored through the use of a vCenter alarm which is shown in Figure 84.



Trigger Type	Condition	⚠ Warning	⚠ Condition Length	◆ Alert	◆ Condition Length
Datastore Disk Usage (%)	Is above	75		85	

● Trigger if any of the conditions are satisfied

○ Trigger if all of the conditions are satisfied

Figure 84 Datastore capacity monitoring in vCenter

System administrators and storage administrators must put processes in place to monitor the capacity for thin pools to make sure that they do not get filled. Thin pools can be dynamically expanded to include more physical capacity without application impact. For DMX arrays running 5773 or VMAX arrays running an Enginuity level earlier than 5874 Service Release 1, these devices should be added in groups long before the thin pool approaches a full condition.

If devices are added individually, hot spots on the disks can be created since much of the write activity is suddenly directed to the newly added data device(s) because other data devices in the pool are full. With the introduction of thin pool write balancing in Enginuity 5874 SR1, the data devices can be added individually without introducing hot spots on the pool, though it is best to add the disks when the activity levels on the pool are low to allow the balancing to take place without conflict.

Viewing thin pool details with the Virtual Storage Integrator

The Storage Viewer feature of Virtual Storage Integrator offers the ability to view the details of thin pools backing Symmetrix thin devices in use as a VMFS volume or RDM in a VMware vSphere environment[1].

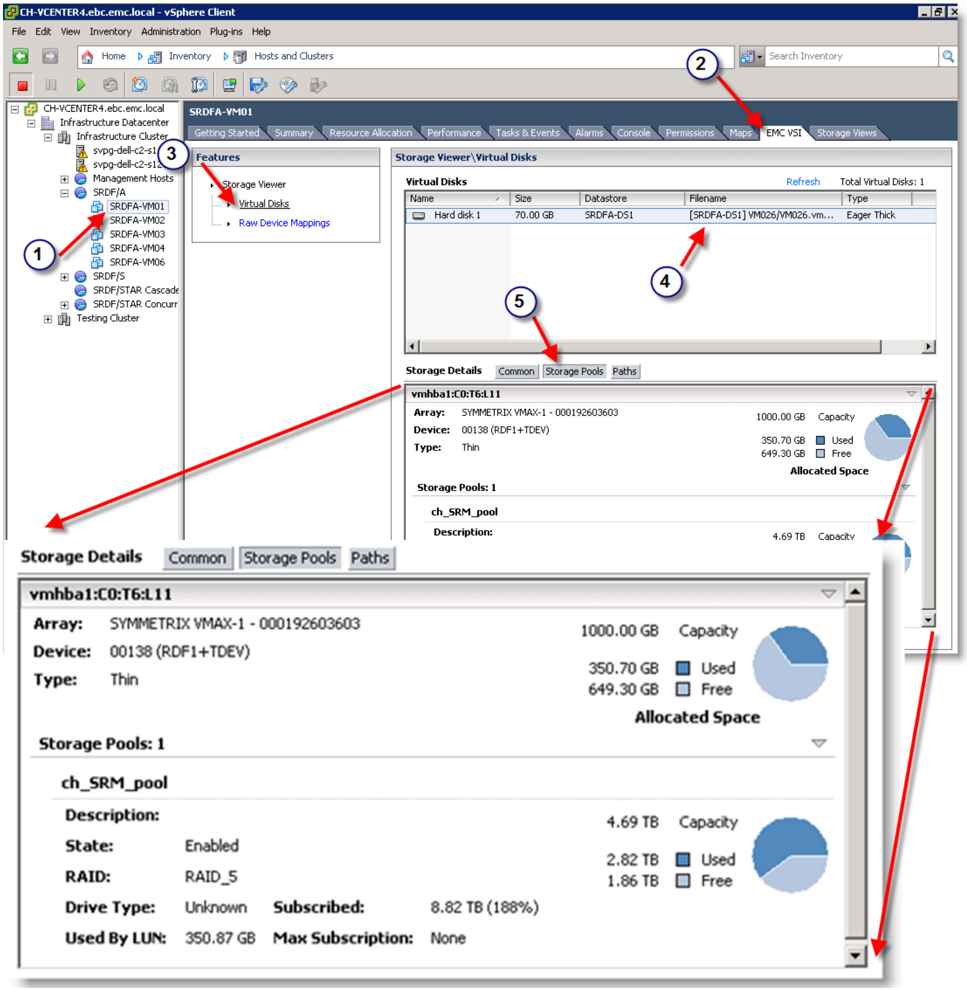
When the Storage Pools button is selected, the view will show the storage pool information for each LUN for that datastore. For each LUN, the following fields are shown:

- Array — Shows the serial number of the array on which the LUN resides. If a friendly name is configured for the array, it will appear in place of the serial number.
- Device — Shows the device name for the LUN.
- Capacity pie chart — Displays a graphic depicting the total capacity of the LUN, along with the free and used space.
- Storage Pools — In most cases a LUN is configured from a single storage pool. However, with FAST VP or after a rebinding operation, portions of a LUN can be moved to other storage pools which can better accommodate its performance needs. For that reason, the number of storage pools is displayed, as well as the details for each storage pool. The details for a particular storage pool can be collapsed by selecting the toggle arrow on the heading for that storage pool. The following details are presented for each storage pool:
  - Name — The heading bar shows the name of the storage pool.
  - Description — Shows the description that the user assigned to the storage pool when it was created.
  - Capacity pie chart — Displays a graphic depicting the total physical capacity of the storage pool, along with the free and used space.
  - State — Shows the state of the storage pool as Enabled or Disabled.
  - RAID — Shows the RAID type of the physical devices backing the thin pool, such as RAID_1, RAID_5, etc.
  - Drive Type — Shows the type of drives used in the storage pool.
  - Used By LUN — Shows the total space consumed by this LUN in the storage pool.

[1] A connection to a remote or local instance of Solutions Enabler with access to the correct array must be configured to allow this functionality. Consult VSI documentation on Powerlink.EMC.com for more information.

- Subscribed — Shows the total capacity of the sum of all of the thin devices bound to the pool and the ratio (as a percentage) of that capacity to the total enabled physical capacity of the thin pool.
- Max subscription — Shows the maximum subscription percentage allowed by the storage pool if configured.

An example of this is shown in Figure 85 on page 161. In this case, a virtual machine is selected in the vCenter inventory and its single virtual disk resides on a VMFS volume on a Symmetrix thin device. The device shown is of type RDF1+TDEV and has allocations on a single pool. This screen can be arrived at in several different ways. Figure 85 on page 161 shows accessing the thin pool details by clicking on a virtual machine. It can also be reached by clicking the EMC VSI tab after selecting a datastore or an ESXi host within the vSphere Client inventory.



Virtual Disks — Total Virtual Disks: 1

Name	Size	Datastore	Filename	Type
Hard disk 1	70.00 GB	SRDFA-DS1	[SRDFA-DS1] VM026/VM026.vm...	Eager Thick

vmhba1:C0:T6:L11

Array: SYMMETRIX VMAX-1 - 000192603603

Device: 00138 (RDF1+TDEV)

Type: Thin

1000.00 GB Capacity

350.70 GB ■ Used

649.30 GB □ Free

Allocated Space

Storage Pools: 1

ch_SRM_pool

Description:

State: Enabled

RAID: RAID_5

Drive Type: Unknown

Subscribed: 8.82 TB (188%)

Used By LUN: 350.87 GB

Max Subscription: None

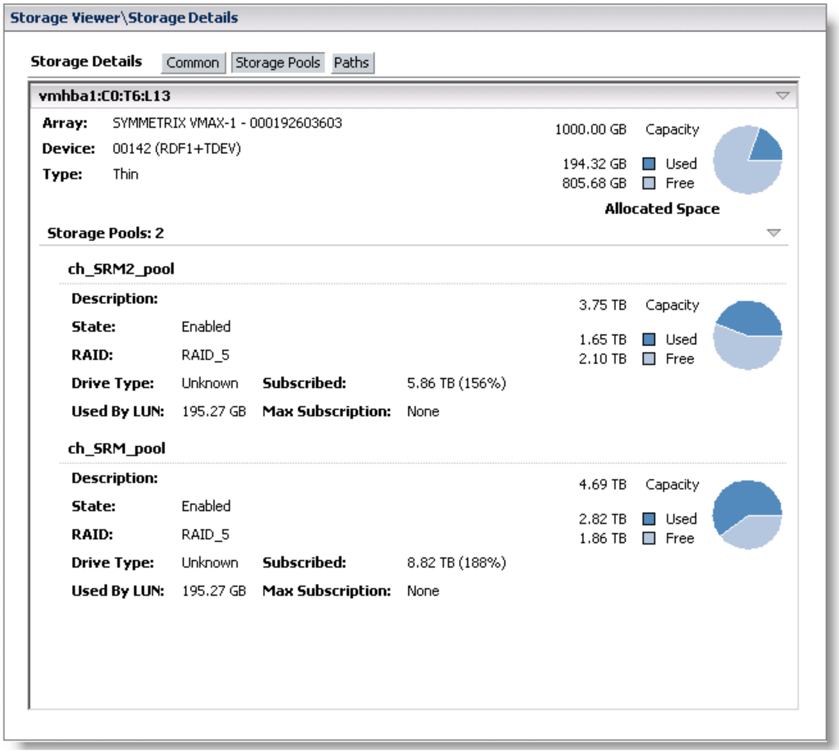
4.69 TB Capacity

2.82 TB ■ Used

1.86 TB □ Free

Figure 85 Viewing thin pool details with the Virtual Storage Integrator

Figure 86 shows a VMFS volume residing on a Symmetrix thin device with allocations on two different pools. For more information refer to the *EMC VSI for VMware vSphere: Storage Viewer Version 5.x Product Guide* on www.Powerlink.EMC.com.



Storage Viewer\Storage Details

Storage Details | Common | Storage Pools | Paths

vmhba1:C0:T6:L13

Array: SYMMETRIX VMAX-1 - 000192603603
Device: 00142 (RDF1+TDEV)
Type: Thin

1000.00 GB Capacity
194.32 GB ■ Used
805.68 GB □ Free
Allocated Space

Storage Pools: 2

ch_SRM2_pool

Description:			
State:	Enabled		
RAID:	RAID_5		
Drive Type:	Unknown	**Subscribed:**	5.86 TB (156%)
Used By LUN:	195.27 GB	**Max Subscription:**	None

3.75 TB Capacity
1.65 TB ■ Used
2.10 TB □ Free

ch_SRM_pool

Description:			
State:	Enabled		
RAID:	RAID_5		
Drive Type:	Unknown	**Subscribed:**	8.82 TB (188%)
Used By LUN:	195.27 GB	**Max Subscription:**	None

4.69 TB Capacity
2.82 TB ■ Used
1.86 TB □ Free

Figure 86 **Viewing pool information for a Symmetrix thin device bound to multiple pools with Virtual Storage Integrator 5**

Exhaustion of oversubscribed pools

A VMFS datastore that is deployed on a Symmetrix thin device can detect only the logical size of the thin device. For example, if the array reports the thin device is 2 TB in size when in reality it is only backed by 1 TB of data devices, the datastore still only considers 2 TB to be the LUN's size as it has no way of knowing the thin pool’s configuration. As the datastore fills up, it cannot determine whether the actual amount of physical space is still sufficient for its needs.

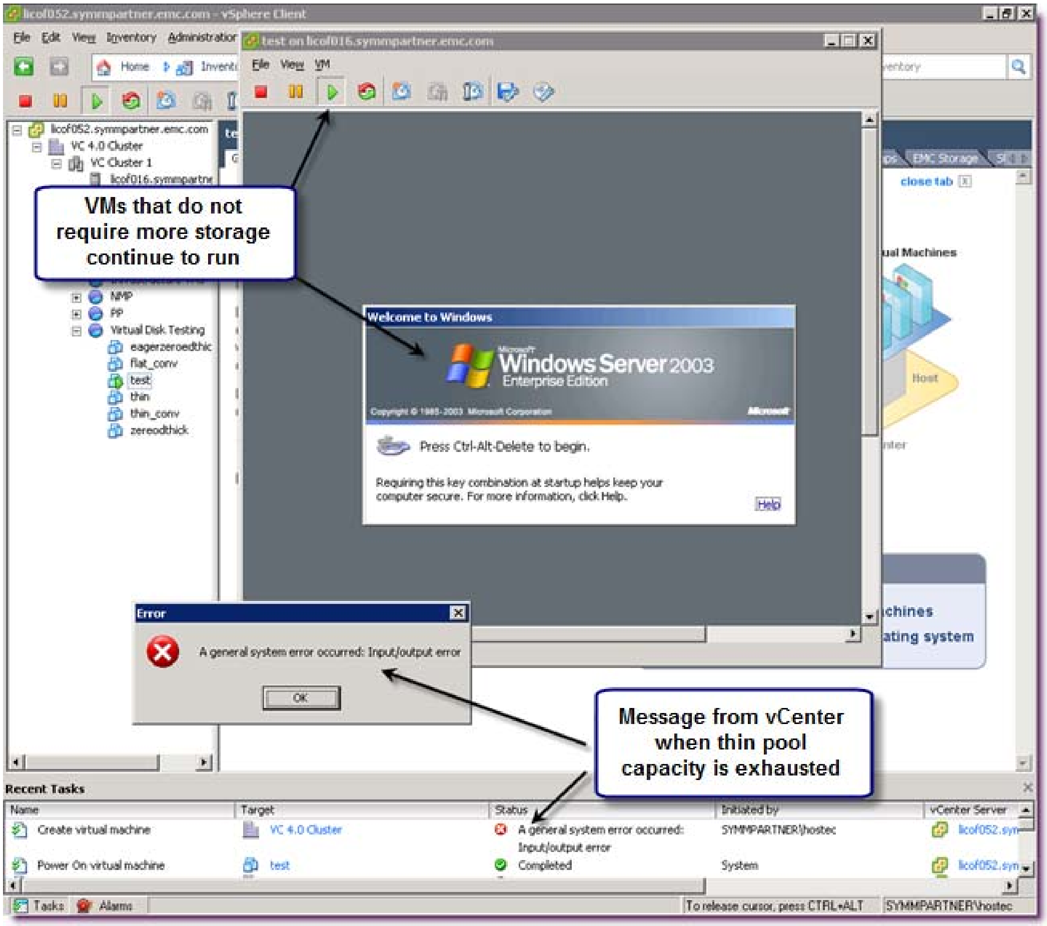
Before ESXi 5, the resulting behavior of running out of space on a thin pool could cause data loss or corruption on the affected virtual machines.

With the introduction of ESXi 5 and Enginuity 5875, a new vStorage API for Thin Provisioning referred to "Out of Space" errors (sometimes called "Stun & Resume") was added. This enables the storage array to notify the ESXi kernel that the thin pool capacity is approaching physical capacity exhaustion or has actually been exhausted.

Behavior of ESX(i) without Out of Space error support

If a VMware environment is running a version of ESX(i) prior to version 5, or the Enginuity level on the target array is lower than 5875, the out-of-space error (Stun & Resume) is not supported. In this case, if a thin pool is oversubscribed and has no available space for new extent allocation, different behaviors can be observed depending on the activity that caused the thin pool capacity to be exceeded.

If the capacity was exceeded when a new virtual machine was being deployed using the vCenter wizard, the error message shown in Figure 87 on page 164 is posted.



licof052.symmpartner.emc.com - vSphere Client
File Edit View Inventory Administration
Home Inventory
licof052.symmpartner.emc.com
VC 4.0 Cluster
VC Cluster 1
licof016.symmpartner
NMP
PP
Virtual Disk Testing
eagerzeroedthick
flat_conv
test
thin
thin_conv
zereodthick
test on licof016.symmpartner.emc.com
File View VM
Welcome to Windows
Microsoft Windows Server 2003 Enterprise Edition
Copyright © 1985-2003 Microsoft Corporation
Microsoft
Help
close tab
EMC Storage
ual Machines
Host
nter
chines
ating system
Tasks
Alarms
To release cursor, press CTRL+ALT
SYMMPARTNER\hostec

VMs that do not require more storage continue to run

Press Ctrl-Alt-Delete to begin.

Requiring this key combination at startup helps keep your computer secure. For more information, click Help.

Error

A general system error occurred: Input/output error

OK

Message from vCenter when thin pool capacity is exhausted

Recent Tasks

Name	Target	Status	Initiated by	vCenter Server
Create virtual machine	VC 4.0 Cluster	A general system error occurred: Input/output error	SYMMPARTNER\hostec	licof052.sym
Power On virtual machine	test	Completed	System	licof052.sym

Figure 87 Error message displayed by vSphere Client on thin pool full conditions

The behavior of virtual machines residing on a VMFS volume when the thin pool is exhausted depends on a number of factors including the guest operating system, the application running in the virtual machine, and also the format that was utilized to create the virtual disks. The virtual machines behave no differently from the same configuration running in a physical environment. In general the following comments can be made for VMware vSphere environments:

- Virtual machines configured with virtual disks using the "eagerzeroedthick" format continue to operate without any disruption.

- Virtual machines that do not require additional storage allocations continue to operate normally.
- If any virtual machine in the environment is impacted due to lack of additional storage, other virtual machines continue to operate normally as long as those machines do not require additional storage.
- Some of the virtual machines may need to be restarted after additional storage is added to the thin pool. In this case, if the virtual machine hosts an ACID-compliant application (such as relational databases), the application performs a recovery process to achieve a consistent point in time after a restart.
- The VMkernel continues to be responsive as long as the ESXi server is installed on a device with sufficient free storage for critical processes.

Behavior of ESX(i) with Out of Space error support

With the new vStorage API for Thin Provisioning introduced in ESXi 5, the host integrates with the physical storage and becomes aware of underlying thin devices and their space usage. Using this new level of array thin provisioning integration, an ESXi host can monitor the use of space on thin pools to avoid data loss and/or data unavailability. This integration offers an automated warning that an out-of-space condition on a thin pool has occurred. This in-band out-of-space error is only supported with Enginuity 5875 and later.

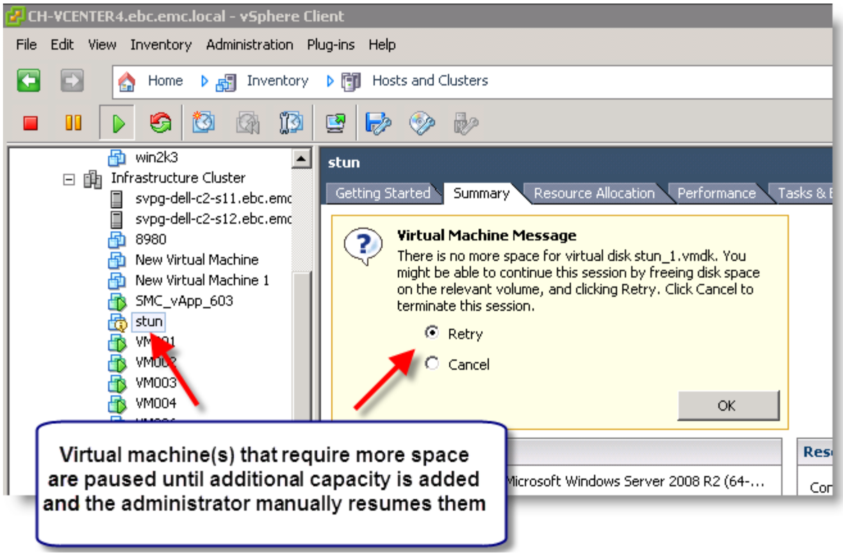
If a thin pool has had its physical capacity exhausted, vCenter issues an out-of-space error. The out-of-space error works in an in-band fashion through the use of SCSI sense codes issued directly from the array. When a thin pool hosting VMFS volumes exhausts its assigned capacity the array returns a tailored response to the ESXi host.[1] This informs the ESXi host that the pool is full and consequently any virtual machines requiring more space will be paused or “stunned” until more capacity is added. Furthermore, any Storage vMotion, Clone or Deploy from Template operations with that device as a target (or any device sharing that pool) will be fail with an out-of-space error.

[1] The response is a device status of CHECK CONDITION and Sense Key = 7h, ASC = 27h, ASCQ=7h

IMPORTANT

The out-of-space error "stun" behavior will prevent data loss or corruption, but, by design it will also cause virtual machines to be temporarily halted. Therefore, any applications running on suspended virtual machines will not be accessible until additional capacity has been added and the virtual machines have been manually resumed. Even though data unavailability might be preferred over experiencing data loss/corruption, this is not an acceptable situation for most organizations in their production environments and should only be thought of as a last resort. It would be best that this situation is never arrived at—prevention of such an event can be achieved through proper and diligent monitoring of thin pool capacity.

Once a virtual machine is paused, additional capacity must then be added to the thin pool to which the thin device is bound. After this has been successfully completed, the VMware administrator must manually resume each paused virtual machine affected by the out-of-space condition. This can be seen in Figure 88. Any virtual machine that does not require additional space (or virtual machines utilizing "eagerzeroedthick" virtual disks) will continue to operate normally.



Virtual Machine Message

There is no more space for virtual disk stun_1.vmdk. You might be able to continue this session by freeing disk space on the relevant volume, and clicking Retry. Click Cancel to terminate this session.

● Retry

○ Cancel

OK

Virtual machine(s) that require more space are paused until additional capacity is added and the administrator manually resumes them

Figure 88 Resuming a paused virtual machine after an out-of-space condition

A more detailed discussion of Symmetrix Virtual Provisioning and its applicability in a VMware environment is beyond the scope of this book. For more detailed information on Symmetrix Virtual Provisioning, please refer to the technical note *Best Practices for Fast, Simple Capacity Allocation with EMC Symmetrix Virtual Provisioning*. For more specific and detailed information on Symmetrix Virtual Provisioning and VMware, refer to the *Implementing EMC Symmetrix Virtual Provisioning with VMware vSphere White Paper* available on Powerlink.EMC.com.

4

Data Placement and Performance in vSphere

This chapter addresses EMC and VMware technologies that help with placing data and improving performance in the VMware environment and on the Symmetrix:

Introduction

Unlike storage arrays of the past, today's EMC Symmetrix arrays contain multiple drive types and protection methodologies. This gives the storage administrator, server administrator, and VMware administrator the challenge of selecting the correct storage configuration, or storage class, for each application being deployed. The trend toward virtualizing the entire environment to optimize IT infrastructures exacerbates the problem by consolidating multiple disparate applications on a small number of large devices.

Given this challenge, it is not uncommon that a single storage type, best suited for the most demanding application, is selected for all virtual machine deployments, effectively assigning all applications, regardless of their performance requirements, to the same high-performance tier. This traditional approach is wasteful since all applications and data are not equally performance-critical to the business. Furthermore, within applications themselves, particularly those reliant upon databases, there is also the opportunity to further diversify the storage make-up.

Both EMC and VMware offer technologies to help ensure that critical applications reside on storage that meets the needs of those applications. These technologies can work independently or, when configured correctly, can complement each other.

This chapter addresses the following technologies:

- EMC Fully Automated Storage Tiering (FAST)
- EMC Virtual LUN
- EMC Host I/O Limit
- vSphere Storage I/O Control
- vSphere Storage Dynamic Resource Scheduler

EMC Fully Automated Storage Tiering (FAST)

Beginning with the release of Enginuity 5874, EMC offers Fully Automated Storage Tiering (FAST) technology. The first incarnation of FAST is known as EMC Symmetrix VMAX Fully Automated Storage Tiering for disk provisioned environments, or FAST DP. With the release of Enginuity 5875, EMC offers the second incarnation of FAST, known as EMC Fully Automated Storage Tiering for Virtual Provisioning environments, or FAST VP.

EMC Symmetrix FAST and FAST VP automate the identification of data volumes for the purposes of relocating application data across different performance/capacity tiers within an array. FAST DP operates on standard Symmetrix devices. Data movements executed between tiers are performed at the full volume level. FAST VP operates on thin devices. As such, data movements executed can be performed at the sub-LUN level, and a single thin device may have extents allocated across multiple thin pools within the array, or on an external array using Federated Tiered Storage. This promotion/demotion of the data across these performance tiers is based on policies that associate a storage group to multiple drive technologies, or RAID protection schemes, by way of thin storage pools, as well as the performance requirements of the application contained within the storage group. Data movement executed during this activity is performed non-disruptively, without affecting business continuity and data availability.

Because FAST DP and FAST VP support different device types—standard and virtual, respectively—they both can operate simultaneously within a single array. Aside from some shared configuration parameters, the management and operation of each are separate.

Federated Tiered Storage (FTS)

Introduced with Enginuity 5876[1], Federated Tiered Storage (FTS) allows LUNs that exist on external arrays to be used to provide physical storage for Symmetrix VMAX arrays. The external LUNs can be used as raw storage space for the creation of Symmetrix devices in the same way internal Symmetrix physical drives are used. These

[1] To utilize FTS on the VMAX 10K platform requires Enginuity release 5876.159.102.

devices are referred to as eDisks. Data on the external LUNs can also be preserved and accessed through Symmetrix devices. This allows the use of Symmetrix Enginuity functionality, such as local replication, remote replication, storage tiering, data management, and data migration with data that resides on external arrays. Beginning with Enginuity release 5876.159.102, FAST VP will now support four tiers if one of those tiers is on a external array using FTS.

FAST benefits

The primary benefits of FAST include:

- Elimination of manually tiering applications when performance objectives change over time.
- Automating the process of identifying data that can benefit from Enterprise Flash Drives or that can be kept on higher capacity, less expensive SATA drives without impacting performance.
- Improving application performance at the same cost, or providing the same application performance at lower cost. Cost is defined as: acquisition (both HW and SW), space/energy, and management expense.
- Optimizing and prioritizing business applications, allowing customers to dynamically allocate resources within a single array.
- Delivering greater flexibility in meeting different price/performance ratios throughout the life-cycle of the information stored.

Due to advances in drive technology, and the need for storage consolidation, the number of drive types supported by Symmetrix arrays has grown significantly. These drives span a range of storage service specializations and cost characteristics that differ greatly.

Several differences exist between the four drive technologies supported by the Symmetrix VMAX Series arrays: Enterprise Flash Drive (EFD), Fibre Channel (FC), Serial Attached SCSI (SAS), and SATA. The primary differences are:

- Response time
- Cost per unit of storage capacity
- Cost per unit of storage request processing

- At one extreme are EFDs, which have a very low response time, and a low cost per IOPS but with a high cost per unit of storage capacity
- At the other extreme are SATA drives, which have a low cost per unit of storage capacity, but high response times and high cost per unit of storage request processing
- Between these two extremes lie Fibre Channel and SAS drives

Based on the nature of the differences that exist between these four drive types, the following observations can be made regarding the most suited workload type for each drive.

- Enterprise Flash Drives: EFDs are more suited for workloads that have a high back-end random read storage request density. Such workloads take advantage of both the low response time provided by the drive, and the low cost per unit of storage request processing without requiring a log of storage capacity.
- SATA drives: SATA drives are suited toward workloads that have a low back-end storage request density.
- Fibre Channel/SAS drives: Fibre Channel and SAS drives are the best drive type for workloads with a back-end storage request density that is not consistently high or low.

This disparity in suitable workloads presents both an opportunity and a challenge for storage administrators.

To the degree it can be arranged for storage workloads to be served by the best suited drive technology, the opportunity exists to improve application performance, reduce hardware acquisition expenses, and reduce operating expenses (including energy costs and space consumption).

The challenge, however, lies in how to realize these benefits without introducing additional administrative overhead and complexity.

The approach taken with FAST is to automate the process of identifying which regions of storage should reside on a given drive technology, and to automatically and non-disruptively move storage between tiers to optimize storage resource usage accordingly. This also needs to be done while taking into account optional constraints on tier capacity usage that may be imposed on specific groups of storage devices.

FAST managed objects

There are three main elements related to the use of both FAST and FAST VP on Symmetrix VMAX. These are shown in Figure 89 on page 174 and are:

- Storage tier — A shared resource with common technologies.
- FAST policy — Manages a set of tier usage rules that provide guidelines for data placement and movement across Symmetrix tiers to achieve service levels and for one or more storage groups.
- Storage group — A logical grouping of devices for common management.

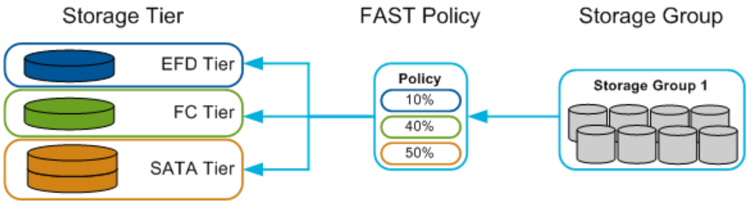


Figure 89 **FAST managed objects**

Each of the three managed objects can be created and managed by using either Unisphere for VMAX (Unisphere) or the Solutions Enabler Command Line Interface (SYMCLI).

FAST VP components

There are two components of FAST VP — the FAST controller and the Symmetrix microcode or Enginuity.

The FAST controller is a service that runs on the Symmetrix VMAX/VMAXe service processor. The Symmetrix microcode is a part of the Enginuity operating environment that controls components within the array. When FAST VP is active, both components participate in the execution of two algorithms, the intelligent tiering algorithm and the allocation compliance algorithm, to determine appropriate data placement.

The intelligent tiering algorithm uses performance data collected by Enginuity, as well as supporting calculations performed by the FAST controller, to issue data movement requests to the Virtual LUN (VLUN) VP data movement engine.

The allocation compliance algorithm enforces the upper limits of storage capacity that can be used in each tier by a given storage group by also issuing data movement requests to satisfy the capacity compliance.

Performance time windows can be defined to specify when the FAST VP controller should collect performance data, upon which analysis is performed to determine the appropriate tier for devices. By default, this will occur 24 hours a day. Defined data movement windows determine when to execute the data movements necessary to move data between tiers. Data movements performed by the microcode are achieved by moving allocated extents between tiers. The size of data movement can be as small as 768 KB, representing a single allocated thin device extent, but more typically will be an entire extent group, which is 7,680 KB in size.

FAST VP has two modes of operation: Automatic or Off. When operating in Automatic mode, data analysis and data movements will occur continuously during the defined data movement windows. In Off mode, performance statistics will continue to be collected, but no data analysis or data movements will take place.

Figure 90 on page 176 shows the FAST controller operation.

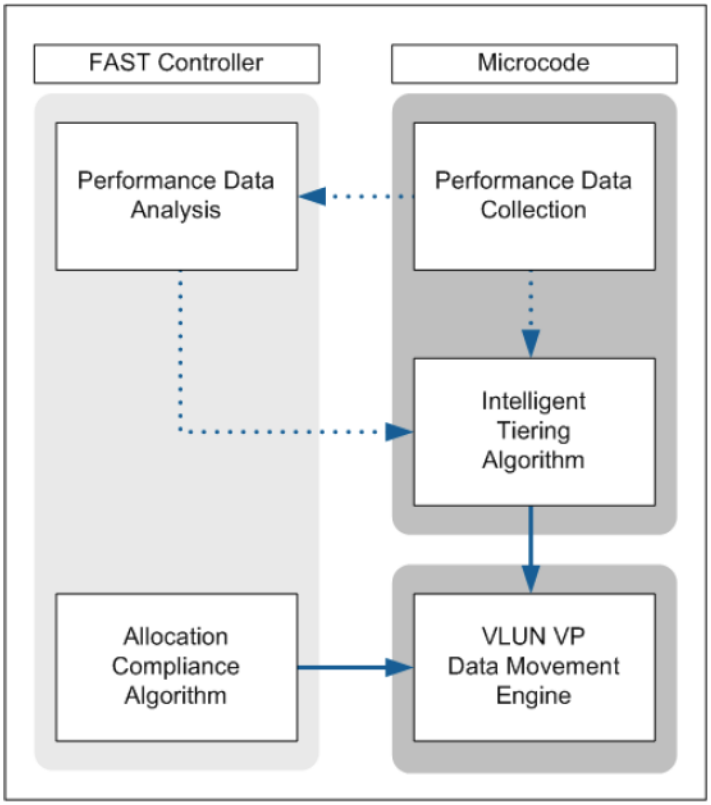

FAST Controller
Microcode
Performance Data Analysis
Performance Data Collection
Intelligent Tiering Algorithm
Allocation Compliance Algorithm
VLUN VP Data Movement Engine

Figure 90 FAST VP components

Note: For more information on FAST VP specifically please see the technical note FAST VP for EMC Symmetrix VMAX Theory and Best Practices for Planning and Performance available at http://support.EMC.com.

FAST VP allocation by FAST Policy

A new feature for FAST VP in 5876 is the ability for a device to allocate new extents from any thin pool participating in the FAST VP Policy. When this feature is enabled, FAST VP will attempt to allocate new extents in the most appropriate tier, based upon performance metrics. If those performance metrics are unavailable it will default to allocating in the pool to which the device is bound. If, however, the chosen pool is full, regardless of performance metrics, then FAST VP will allocate from one of the other thin pools in the policy. As long as there is space available in one of the thin pools, new extent allocations will be successful.

This new feature is enabled at the Symmetrix array level and applies to all devices managed by FAST VP. The feature cannot, therefore, be applied to some FAST VP policies and not others. By default it is disabled and any new allocations will come from the pool to which the device is bound.

Note: A pinned device is one that is not considered to have performance metrics available and therefore new allocations will be done in the pool to which the device is bound.

FAST VP SRDF coordination

The use of FAST VP with SRDF devices is fully supported; however FAST VP operates within a single array, and therefore will only impact the RDF devices on that array. There is no coordination of the data movement between RDF pairs. Each device's extents will move according to the manner in which they are accessed on that array, source or target.

For instance, an R1 device will typically be subject to a read/write workload, while the R2 will only experience the writes that are propagated across the link from the R1. Because the reads to the R1 are not propagated to the R2, FAST VP will make its decisions based solely on the writes and therefore the R2 data may not be moved to the same tier as the R1.

To rectify this problem, EMC introduced FAST VP SRDF coordination in 5876. FAST VP SRDF coordination allows the R1 performance metrics to be transmitted across the link and used by the FAST VP engine on the R2 array to make promotion and demotion decisions. Both arrays involved with replication must be at Enginuity 5876 to take advantage of this feature.

FAST VP SRDF coordination is enabled or disabled at the storage group that is associated with the FAST VP policy. The default state is disabled. Note that only the R1 site must have coordination enabled. Enabling coordination on the R2 has no impact unless an SRDF swap is performed.

Note: FAST VP SRDF coordination is supported for single and concurrent SRDF pairings (R1 and R11 devices) in any mode of operation: Synchronous, asynchronous, or adaptive copy. FAST VP SRDF coordination is not supported for SRDF/Star, SRDF/EDP, or Cascaded SRDF including R21 and R22 devices.

FAST VP storage group reassociate

Beginning with Solutions Enabler V7.4, released with Enginuity 5876, a storage group that is under a FAST VP policy, may be reassociated to a new FAST VP policy non-disruptively. With a reassociation all current attributes set on the association propagate automatically to the new association. For example, if SRDF coordination is set on the original policy, it is automatically set on the new policy upon reassociation.

FAST VP space reclamation and UNMAP with VMware

Space reclamation may be run against a thin device under FAST VP control. However, during the space reclamation process, the sub-LUN performance metrics are not updated, and no data movements are performed. Furthermore, if FAST VP is moving extents on the device, the reclaim request will fail. It is a best practice, therefore, to pin the device before issuing a reclaim.

Space reclamation may also be achieved with the SCSI command UNMAP. Beginning with Enginuity 5876.159.102 and VMware vSphere 5.0 U1, EMC offers Thin Provisioning Block Space Reclamation, better known as UNMAP. This feature enables the reclamation of blocks of thin-provisioned LUNs by informing the array that specific blocks are obsolete. Currently UNMAP does not occur automatically - it is performed at the VMFS level as a manual process using `vmkfstools`; however this process is fully supported on thin devices under FAST VP control.

FAST VP Compression

Enginuity release 5876.159.102 introduces compression capabilities for Virtually Provisioned environments (VP compression) that provide customers with 2:1 or greater data compression for very infrequently accessed data. EMC defines infrequently accessed data as that which is not accessed within month-end, quarter-end, or year-end activities, nor accessed as part of full backup processing. Data within a thin pool may be compressed manually for an individual device or group of devices, via Solutions Enabler. Alternatively, inactive data may be compressed automatically for thin devices that are managed by FAST VP.

FAST VP Compression is enabled at the system level by setting the time to compress parameter to a value between 40 and 400 days. Data that is seen to be inactive for a period of time longer than this parameter will be considered eligible for automatic compression. The default value for the time to compress is “never”.

Automated FAST VP Compression

FAST VP Compression automates VP Compression for thin devices that are managed by FAST VP. This automated compression takes place at the sub-LUN level. Compression decisions are based on the activity level of individual extent groups (120 tracks), similar to promotion and demotion decisions. However, the actual compression itself is performed per thin device extent (12 tracks).

Extent groups that have been inactive on disk for a user-defined period of time, known as the time to compress, will be considered eligible for compression. The time to compress parameter can be set between 40 and 400 days. The default time to compress is “never”, meaning no data will be compressed.

Note: It is important to understand that uncompressing data is an extremely expensive task both in terms of time and resources. It is therefore best to set the time to compress parameter to a value after which the data is not expected to be needed. For example, if a customer runs quarterly reports on financial data it would not be recommended to set the time to compress to something less than 90 days. The reason for this is the following: If only the current month's financial data is accessed except during quarterly reporting, there is a very real chance that a portion of a quarter's data will already be compressed at the time of report running. In order for the reports to gain access to the compressed data it will need to be uncompressed, requiring both thin pool space and time along with processing power.The impact would be significant and should therefore be avoided. It is very likely in such environments that year-end reports are also required. In that event, it may be best to set the time to compress parameter to greater than a year.

In order for thin device extent to be compressed automatically by FAST VP, all of the following criteria must be met:

- The extent must belong to an extent group that has been inactive for a period equal to or greater than the time to compress
- The extent must be located in a thin pool that has been enabled for compression
- The extent must compress by at least 50%

FAST VP Compression and policy compliance

For the purposes of calculating a storage group's compliance with a policy, FAST VP does not take into account any compression that has occurred on any of the devices within the group. Only the logical, allocated capacity consumed by the devices is considered. For example, if 50 GB of data belonging to a storage group was demoted to the SATA tier and subsequently compressed to only consume 20 GB, the compliance calculation for the storage group would be based on 50 GB consumed within the SATA tier, not the 20 GB actually occupied.

FAST VP Compression and SRDF coordination

If FAST VP SRDF coordination is enabled for a storage group associated with a FAST VP policy, performance metrics collected for an R1 device are transferred to the R2, allowing the R1 workload to influence promotion and demotion decisions on the R2 data.

As a part of the performance metrics collected, the length of time the data has been inactive is also transferred. This period of inactivity on R1 data can then influence compression decisions on the corresponding R2 data.

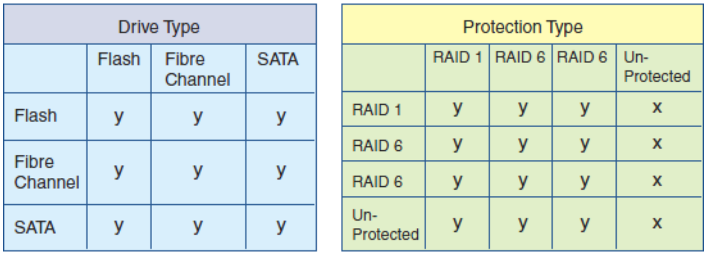
FAST management

Management and operation of FAST are provided by SMC or Unisphere as well as the Solutions Enabler Command Line Interface (SYMCLI). Also, detailed performance trending, forecasting, alerts, and resource utilization are provided through Symmetrix Performance Analyzer (SPA) or the Performance monitoring option of Unisphere. EMC Ionix ControlCenter provides the capability for advanced reporting and analysis to be used for charge-back and capacity planning.

EMC Virtual LUN migration

This feature offers system administrators the ability to transparently migrate host visible LUNs from differing tiers of storage that are available in the Symmetrix VMAX. The storage tiers can represent differing hardware capability as well as differing tiers of protection. The LUNs can be migrated to either unallocated space (also referred to as unconfigured space) or to configured space, which is defined as existing Symmetrix LUNs that are not currently assigned to a server—existing, not-ready volumes—within the same subsystem. The data on the original source LUNs is cleared using instant VTOC once the migration has been deemed successful. The migration does not require swap or DRV space, and is non-disruptive to the attached hosts or other internal Symmetrix applications such as TimeFinder and SRDF.

Figure 91 shows the valid combinations of drive types and protection types that are available for migration.



	Drive Type		
	Flash	Fibre Channel	SATA
Flash	y	y	y
Fibre Channel	y	y	y
SATA	y	y	y

	Protection Type			
	RAID 1	RAID 6	RAID 6	Un-Protected
RAID 1	y	y	y	x
RAID 6	y	y	y	x
RAID 6	y	y	y	x
Un-Protected	y	y	y	x

Figure 91 Virtual LUN eligibility tables

The device migration is completely transparent to the host on which an application is running since the operation is executed against the Symmetrix device. Thus, the target and LUN number are not changed and applications are uninterrupted. Furthermore, in SRDF environments, the migration does not require customers to re-establish their disaster recovery protection after the migration.

The Virtual LUN feature leverages the virtual RAID architecture introduced in Enginuity 5874, which abstracts device protection from its logical representation to a server. This powerful approach allows a device to have more simultaneous protection types such as BCVs, SRDF, Concurrent SRDF, and spares. It also enables seamless transition from one protection type to another while servers and their associated applications and Symmetrix software are accessing the device.

The Virtual LUN feature offers customers the ability to effectively utilize SATA storage—a much cheaper, yet reliable, form of storage. It also facilitates fluid movement of data across the various storage tiers present within the subsystem—the realization of true *tiered storage in the box*.

Thus, Symmetrix VMAX becomes the first enterprise storage subsystem to offer a comprehensive "tiered storage in the box", ILM capability that complements the customer's tiering initiatives. Customers can now achieve varied cost/performance profiles by moving lower priority application data to less expensive storage, or conversely, moving higher priority or critical application data to higher performing storage as their needs dictate.

Specific use cases for customer applications enable the moving of data volumes transparently from tier to tier based on changing performance (moving to faster or slower disks) or availability requirements (changing RAID protection on the array). This migration can be performed transparently without interrupting those applications or host systems utilizing the array volumes and with only a minimal impact to performance during the migration.

VLUN VP

Beginning with Enginuity 5875, the Virtual LUN feature supports "thin-to-thin" migrations. Known as VLUN VP (VLUN), "thin-to-thin" mobility enables users to meet tiered storage requirements by migrating thin FBA LUNs between virtual pools in the same array. Virtual LUN VP mobility gives administrators the option to "re-tier" a thin volume or set of thin volumes by moving them between thin pools in a given FAST configuration. This manual "override" option helps FAST users respond rapidly to changing performance requirements or unexpected events.

Virtual LUN VP migrations are session-based—each session may contain multiple devices to be migrated at the same time. There may also be multiple concurrent migration sessions. At the time of submission, a migration session name is specified. This session name is subsequently used for monitoring and managing the migration.

While an entire thin device will be specified for migration, only thin device extents that are allocated will be relocated. Thin device extents that have been allocated, but not written to (for example, pre-allocated tracks), will be relocated but will not cause any actual data to be copied. New extent allocations that occur as a result of a host write to the thin device during the migration will be satisfied from the migration target pool.

The advances in VLUN enable customers to move Symmetrix thin devices from one thin pool to another thin pool on the same Symmetrix without disrupting user applications and with minimal impact to host I/O. Beginning with 5876, VLUN VP users have the option to move only part of the thin device from one source pool to one destination pool. Users may move whole or part[1] of thin devices between thin pools to:

- Change the disk media on which the thin devices are stored
- Change the thin device RAID protection level
- Consolidate thin device that is/was managed by FAST VP to a single thin pool

[1] If a thin device has extents in more than one thin pool, the extents in each thin pool can be moved independently to the same or different thin pools.There is no way, however, to relocate part of a thin device when all its extents reside in a single thin pool.

- Move all extents from a thin device that are in one thin pool to another thin pool

In a VMware environment a customer may have any number of use cases for VLUN. For instance, if a customer elects to not use FAST VP, manual tiering is achievable through VLUN. A heavily used datastore residing on SATA drives may require the ability to provide improved performance. That thin device underlying the datastore could be manually moved to EFD. Conversely a datastore residing on FC that houses data needing to be archived could be moved to a SATA device which, although a less-performing disk tier, has a much smaller cost per GB.

In a FAST VP environment, customers may wish to circumvent the automatic process in cases where they know all the data on a thin device has changed its function. Take, for instance, the previous example of archiving. A datastore that contains information that has now been designated as archive could be removed from FAST VP control then migrated over to the thin pool that is comprised of the disk technology most suited for archived data, SATA.

Note: When using VLUN VP with FAST VP, the destination pool must be part of the FAST VP policy.

Note: The *Storage Tiering for VMware Environments Deployed on EMC Symmetrix VMAX with Enginuity 5876 White Paper* on www.EMC.com provides detailed examples and a specific use case on using FAST VP in a VMware environment.

VLUN migration with Unisphere for VMAX example

A VLUN migration can be performed using the **symmigrate** command, with EMC Solutions Enabler 7.0 and higher; or with the Migrate Storage Locally wizard available in the Symmetrix Management Console; or the VLUN Migration wizard in Unisphere for VMAX.

The following example uses Unisphere for VMAX to migrate a thin LUN.

Within Unisphere, the VLUN Migration option as seen in Figure 92 on page 185, appears in the extended menu in a number of places.

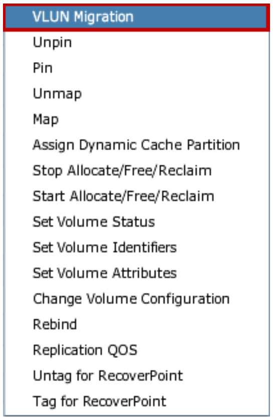
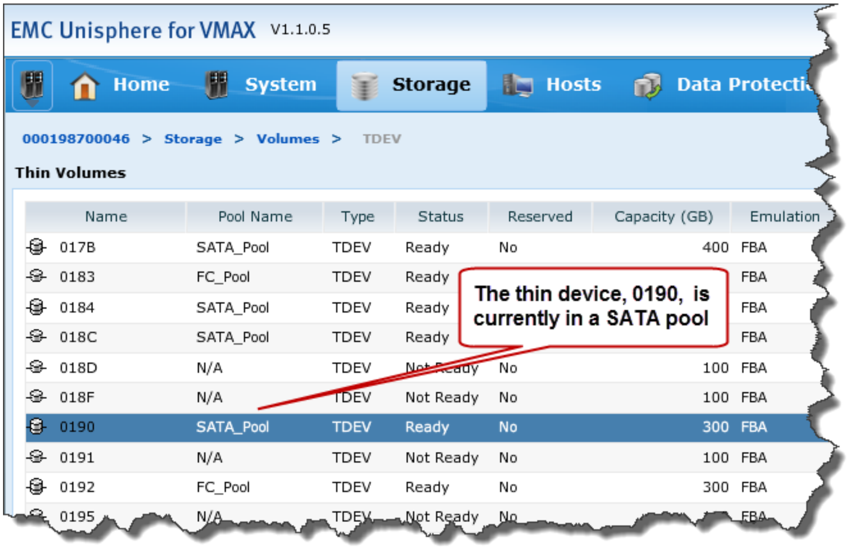


Figure 92 VLUN Migration option in Unisphere for VMAX menu

The user, for instance, may find it easier to access it from the Storage Groups section of Unisphere as the LUNs are grouped together. However, the menu is available from the Volumes section also. This location will be used in this example with the assumption the LUN has not been presented to a host.

1. Start by navigating the Unisphere menu to Storage > Volumes > TDEV and highlighting the thin device to be migrated as in Figure 93 on page 186.



EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
Hosts
Data Protecti
000198700046 > Storage > Volumes > TDEV

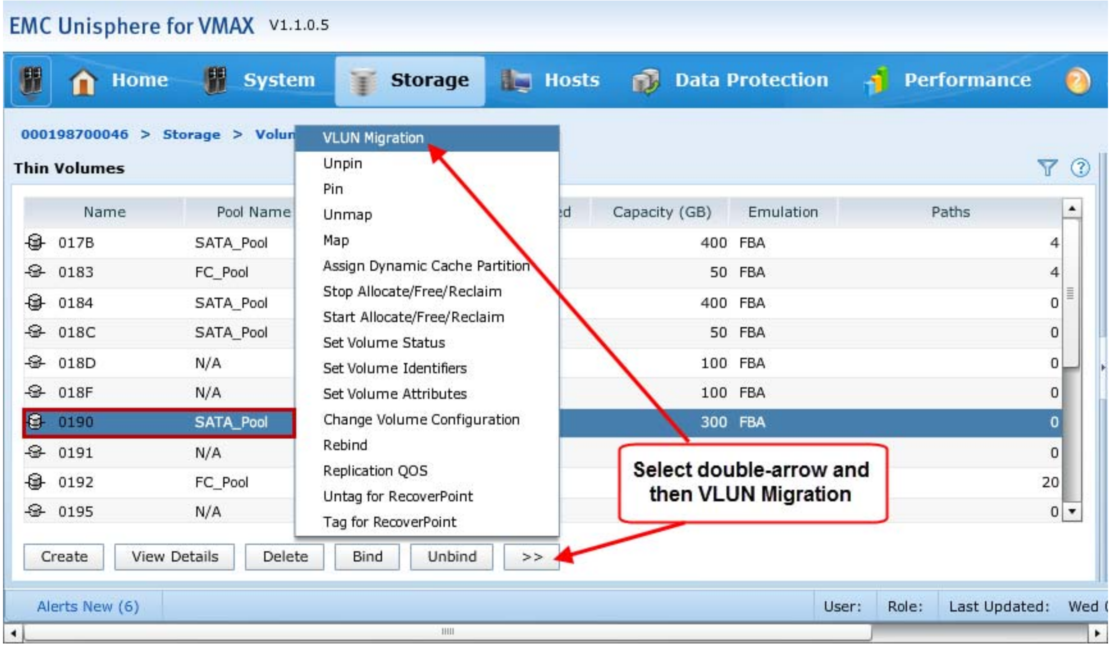
Thin Volumes

Name	Pool Name	Type	Status	Reserved	Capacity (GB)	Emulation
017B	SATA_Pool	TDEV	Ready	No	400	FBA
0183	FC_Pool	TDEV	Ready			FBA
0184	SATA_Pool	TDEV	Ready			FBA
018C	SATA_Pool	TDEV	Ready			FBA
018D	N/A	TDEV	Not Ready	No	100	FBA
018F	N/A	TDEV	Not Ready	No	100	FBA
0190	SATA_Pool	TDEV	Ready	No	300	FBA
0191	N/A	TDEV	Not Ready	No	100	FBA
0192	FC_Pool	TDEV	Ready	No	300	FBA
0195	N/A	TDEV	Not Ready	No		FBA

The thin device, 0190, is currently in a SATA pool

Figure 93 Selecting thin device in Unisphere for VMAX

2. Now select the double-arrow at the bottom of the screen and select *VLUN Migration* from the menu as shown in Figure 94.



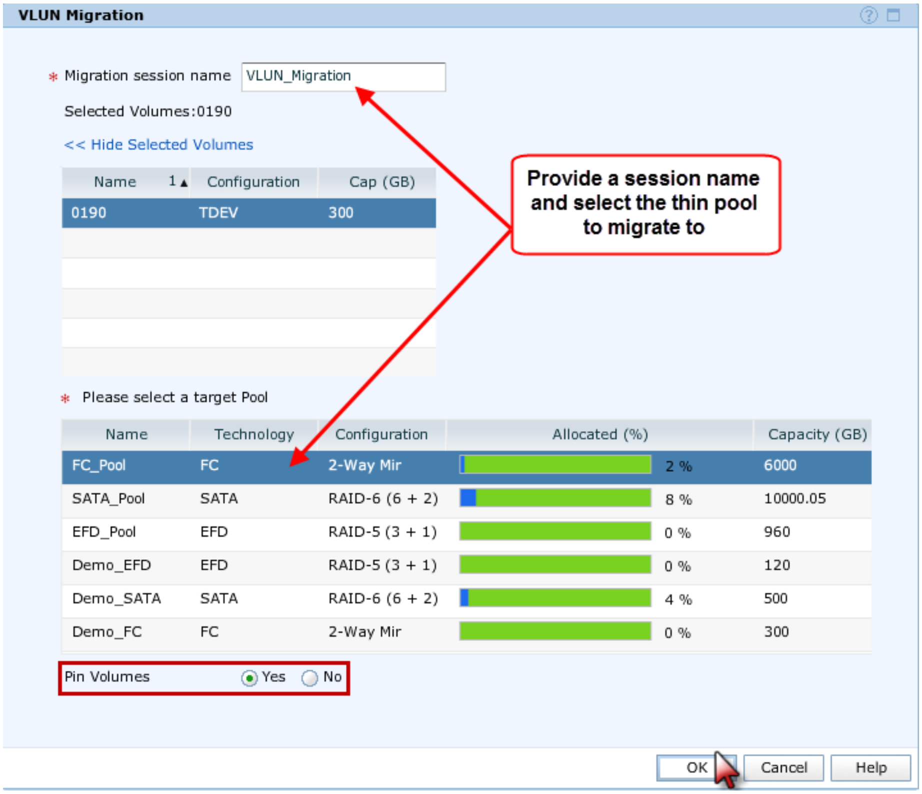
Name	Pool Name	Capacity (GB)	Emulation	Paths
017B	SATA_Pool	400	FBA	4
0183	FC_Pool	50	FBA	4
0184	SATA_Pool	400	FBA	0
018C	SATA_Pool	50	FBA	0
018D	N/A	100	FBA	0
018F	N/A	100	FBA	0
0190	SATA_Pool	300	FBA	0
0191	N/A			0
0192	FC_Pool			20
0195	N/A			0

Select double-arrow and then VLUN Migration

Figure 94 Selecting thin device and VLUN Migration menu in Unisphere for VMAX

3. Once the wizard is launched, enter the following information as shown in Figure 95 on page 189:

   - A session name (no spaces). The name associated with this migration can be used to interrogate the progress of the migration. It can be anything.
   - Select the pool to migrate to, in this case FC_Pool. Note that the wizard will allow you to select the same pool the thin device is currently bound to. This is because if the device has extents in other pools, the user may wish to bring them all back to the original pool.
   - Using the radio buttons, select whether to Pin the volumes. This is specifically for FAST VP. Note that during a VLUN Migration, FAST VP will not move extents, regardless if the device is pinned or not; however upon termination of the migration session, it can move extents so it is important to pin the device to prevent that.
   - Select "OK" to start the migration.



VLUN Migration

* Migration session name: VLUN_Migration

Selected Volumes:0190

<< Hide Selected Volumes

Name 1▲	Configuration	Cap (GB)
0190	TDEV	300

* Please select a target Pool

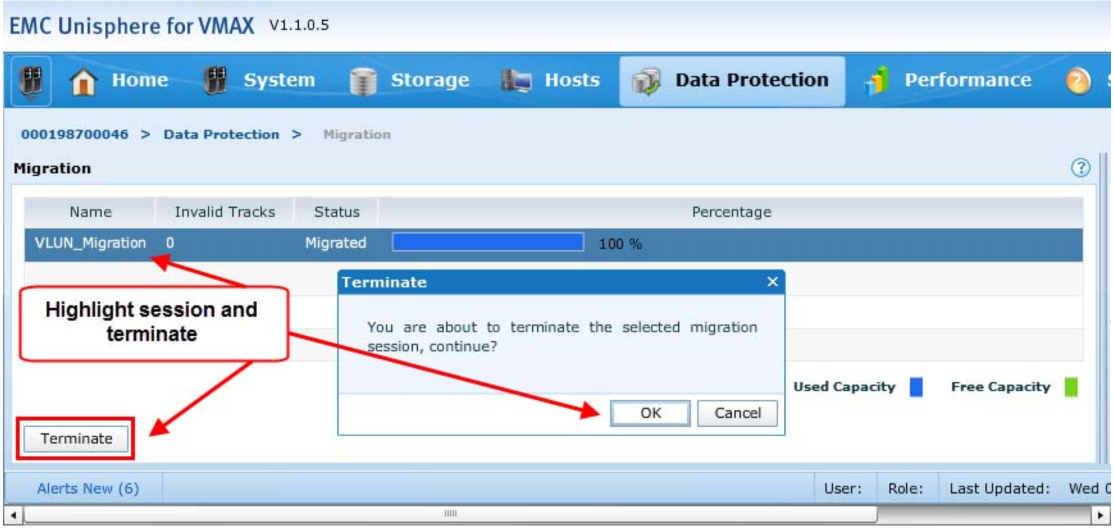
Name	Technology	Configuration	Allocated (%)	Capacity (GB)
FC_Pool	FC	2-Way Mir	2 %	6000
SATA_Pool	SATA	RAID-6 (6 + 2)	8 %	10000.05
EFD_Pool	EFD	RAID-5 (3 + 1)	0 %	960
Demo_EFD	EFD	RAID-5 (3 + 1)	0 %	120
Demo_SATA	SATA	RAID-6 (6 + 2)	4 %	500
Demo_FC	FC	2-Way Mir	0 %	300

Pin Volumes ● Yes ○ No

Provide a session name and select the thin pool to migrate to

Figure 95 Inputing information for the VLUN migration in Unisphere for VMAX

4. After the session is submitted, navigate to Data Protection > Migration to see the active session. The status will show "SyncInProg" while the migration is active. Once the status shows "Migrated" as in Figure 96 on page 190 it can be terminated. Note that the session will not terminate on its own.



Name	Invalid Tracks	Status	Percentage
VLUN_Migration	0	Migrated	100 %

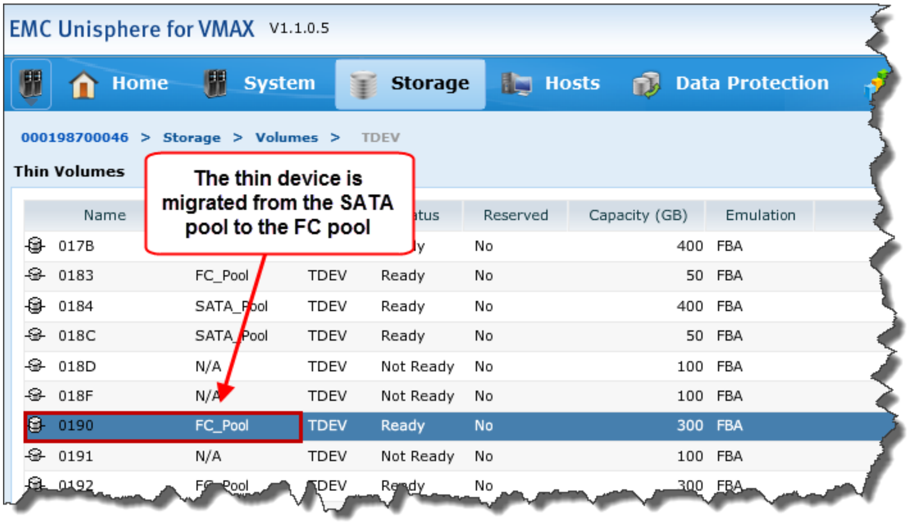
Terminate

You are about to terminate the selected migration session, continue?

Highlight session and terminate

Figure 96 **Terminating a VLUN Migration session in Unisphere for VMAX**

5. Returning to Storage > Volumes > TDEV the thin device now shows the new thin pool, FC_Pool, in Figure 97 on page 191.


EMC Unisphere for VMAX V1.1.0.5
Home
System
Storage
Hosts
Data Protection
000198700046 > Storage > Volumes > TDEV

Thin Volumes

The thin device is migrated from the SATA pool to the FC pool

Name			tus	Reserved	Capacity (GB)	Emulation
017B			y	No	400	FBA
0183	FC_Pool	TDEV	Ready	No	50	FBA
0184	SATA_Pool	TDEV	Ready	No	400	FBA
018C	SATA_Pool	TDEV	Ready	No	50	FBA
018D	N/A	TDEV	Not Ready	No	100	FBA
018F	N/A	TDEV	Not Ready	No	100	FBA
0190	FC_Pool	TDEV	Ready	No	300	FBA
0191	N/A	TDEV	Not Ready	No	100	FBA
0192	FC_Pool	TDEV	Ready	No	300	FBA

Figure 97 Migrated thin device in new pool as seen in Unisphere for VMAX

EMC Host I/O Limit

Beginning with Enginuity release 5876.159.102, EMC provides support for the setting of the Host I/O Limits. This feature allows users to place limits on the front-end bandwidth and IOPS consumed by applications on Symmetrix systems.

Limits are set on a per-storage-group basis. As users build masking views with these storage groups, limits for maximum front-end IOPS or MB/s are evenly distributed across the directors within the associated masking view. The bandwidth and IOPS are monitored by the Symmetrix system to ensure that they do not exceed the user-specified maximum.

The benefits of implementing Host I/O Limit are:

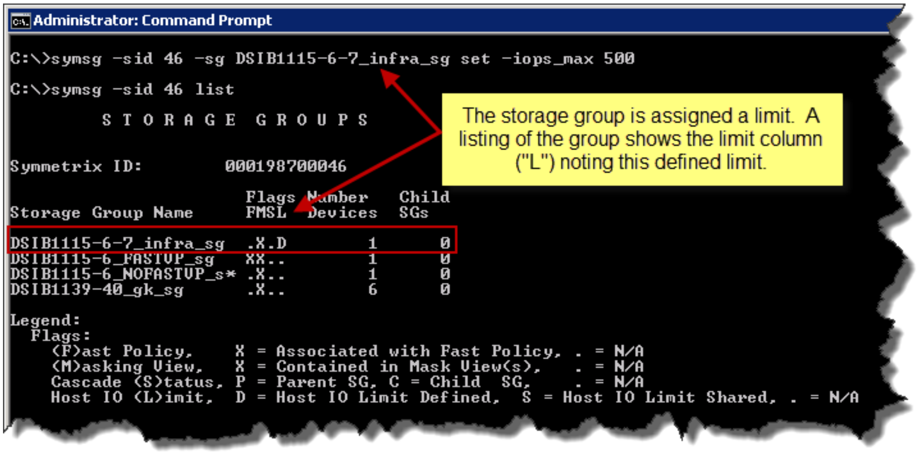
- Ensures that applications cannot exceed their set limit, reducing the potential of impacting other applications
- Enforces performance ceilings on a per-storage-group basis
- Ensures applications do not exceed their allocated share, thus reducing the potential of impacting other applications
- Provides greater levels of control on performance allocation in multi-tenant environments
- Enables predictability needed to service more customers on the same array
- Simplifies quality-of-service management by presenting controls in industry-standard terms of IOPS and throughput
- Provides the flexibility of setting either IOPS or throughput or both, based on application characteristics and business needs
- Manages expectations of application administrators with regard to performance and allows customers to provide incentives for their users to upgrade their performance service levels

Host I/O Limit requires both Enginuity 5876.159.102 and Solutions Enabler V7.5.

To configure the Host I/O Limits feature on a set of devices, a Host I/O Limit is added to a storage group. A provisioning view is then created using that storage group. This causes the Host I/O Limit defined on the storage group to be applied to the devices in the storage group for the ports defined in the port group. The Host I/O Limit for a storage group can be either active or inactive.

Note: In most cases, the total Host I/O Limit may only be achieved with proper host load balancing between directors (by multipath software on the hosts, such as PowerPath/VE.

The Host I/O Limit can be set using the SYMCLI command `symsg` or Unisphere V1.5. Figure 98 demonstrates the use of SYMCLI to set the Host I/O Limit.

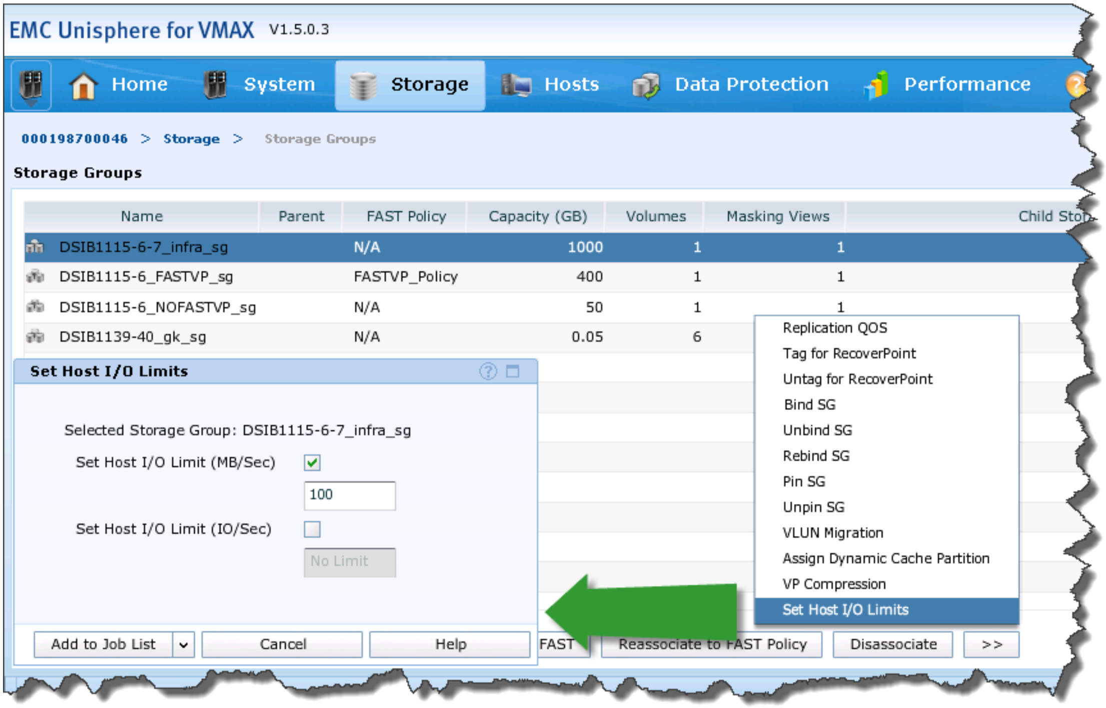


```
C:\>symsg -sid 46 -sg DSIB1115-6-7_infra_sg set -iops_max 500

C:\>symsg -sid 46 list

         S T O R A G E   G R O U P S


Symmetrix ID:          000198700046

                       Flags  Number   Child
Storage Group Name     FMSL   Devices  SGs

DSIB1115-6-7_infra_sg  .X.D         1      0
DSIB1115-6_FASTVP_sg   XX..         1      0
DSIB1115-6_NOFASTVP_s* .X..         1      0
DSIB1139-40_gk_sg      .X..         6      0

Legend:
  Flags:
    (F)ast Policy,    X = Associated with Fast Policy, . = N/A
    (M)asking View,   X = Contained in Mask View(s),   . = N/A
    Cascade (S)tatus, P = Parent SG, C = Child  SG,    . = N/A
    Host IO (L)imit,  D = Host IO Limit Defined,  S = Host IO Limit Shared, . = N/A
```

The storage group is assigned a limit. A listing of the group shows the limit column ("L") noting this defined limit.

Figure 98 **Setting the Host I/O Limit through SYMCLI V7.5**

Using Unisphere, navigate to the storage group and by selecting it and then clicking on the double-arrow on the bottom of the screen, the option to set the Host I/O Limit is available. In this case, seen in Figure 99 on page 194, the check box is selected for MB/Sec. Run the job to complete the change.


EMC Unisphere for VMAX V1.5.0.3
Home
System
Storage
Hosts
Data Protection
Performance
000198700046 > Storage > Storage Groups
Storage Groups

Name	Parent	FAST Policy	Capacity (GB)	Volumes	Masking Views	Child Sto
DSIB1115-6-7_infra_sg		N/A	1000	1	1	
DSIB1115-6_FASTVP_sg		FASTVP_Policy	400	1	1	
DSIB1115-6_NOFASTVP_sg		N/A	50	1	1	
DSIB1139-40_gk_sg		N/A	0.05	6		

Set Host I/O Limits

Selected Storage Group: DSIB1115-6-7_infra_sg

Set Host I/O Limit (MB/Sec) ☑ 100

Set Host I/O Limit (IO/Sec) □ No Limit

Figure 99 **Setting the Host I/O Limit through Unisphere V1.5**

For more information about Host I/O Limit please refer to the Technical Note *Host I/O Limits for Symmetrix VMAX Family Arrays* on Powerlink.

vSphere Storage I/O Control

Storage I/O Control is a capability in vSphere 4.1 and higher that allows cluster-wide storage I/O prioritization. Storage I/O Control extends the constructs of shares and limits to handle storage I/O resources – in other words I/O DRS. The I/O control is enabled on the datastore and the prioritization is set at the virtual machine (VM) level. VMware claims that this allows better workload consolidation and helps reduce extra costs associated with over-provisioning. The idea is that you can take many different applications, with disparate I/O requirements, and store them on the same datastore.

If the datastore (that is the underlying LUN) begins to provide degraded performance in terms of I/O response, VMware will automatically adjust a VM's I/O share based upon user-defined levels. This ensures that those VMs running more important applications on a particular datastore get preference over VMs that run less critical applications against that datastore.

When one enables Storage I/O Control on a datastore, ESX begins to monitor the device latency that hosts observe when communicating with that datastore. When device latency exceeds a threshold (currently defaults to 30 milliseconds), the datastore is considered to be congested and each virtual machine that accesses that datastore is allocated I/O resources in proportion to their shares as assigned by the VMware administrator.

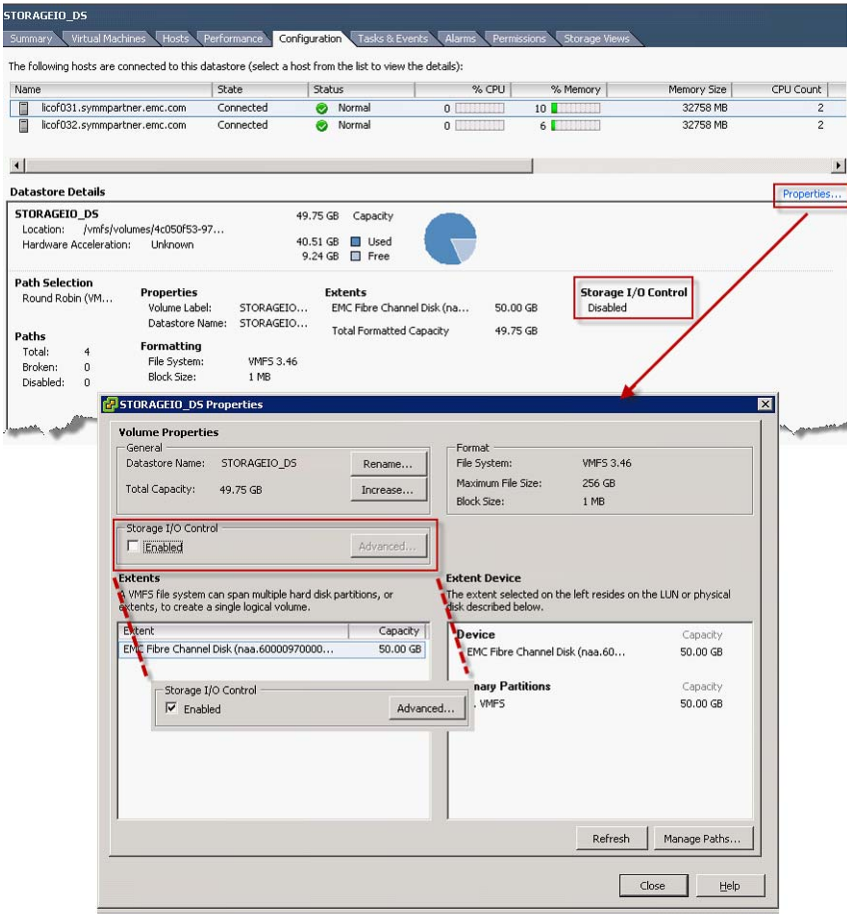
Note: Storage I/O Control is fully supported to be activated on volumes also under FAST VP control.

Configuring Storage I/O Control is a two-step process as it is disabled by default:

1. Enable Storage I/O Control on the datastore.
2. Set the number of storage I/O shares and upper limit of I/O operations per second (IOPS) allowed for each VM. By default, all virtual machine shares are set to Normal (1000) with unlimited IOPS.

Enable Storage I/O Control

The first step is to enable Storage I/O Control on the datastore. Figure 100 on page 196 illustrates how the datastore STORAGEIO_DS is enabled for Storage I/O Control. Simply select the Properties link from the Configuration tab of the datastore and one will be presented with the Properties box where there is a check box to enable the Storage I/O Control feature.



The following hosts are connected to this datastore (select a host from the list to view the details):

Name	State	Status	% CPU	% Memory	Memory Size	CPU Count
licof031.symmpartner.emc.com	Connected	Normal	0	10	32758 MB	2
licof032.symmpartner.emc.com	Connected	Normal	0	6	32758 MB	2

STORAGEIO_DS
Location: /vmfs/volumes/4c050f53-97...
Hardware Acceleration: Unknown

49.75 GB Capacity
40.51 GB Used
9.24 GB Free

Path Selection
Round Robin (VM...

Paths
Total: 4
Broken: 0
Disabled: 0

Properties
Volume Label: STORAGEIO...
Datastore Name: STORAGEIO...

Formatting
File System: VMFS 3.46
Block Size: 1 MB

Extents
EMC Fibre Channel Disk (na... 50.00 GB
Total Formatted Capacity 49.75 GB

Storage I/O Control
Disabled

Volume Properties

General
Datastore Name: STORAGEIO_DS
Total Capacity: 49.75 GB

Format
File System: VMFS 3.46
Maximum File Size: 256 GB
Block Size: 1 MB

Storage I/O Control
□ Enabled

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

Extent	Capacity
EMC Fibre Channel Disk (naa.60000970000...	50.00 GB

Storage I/O Control
☑ Enabled

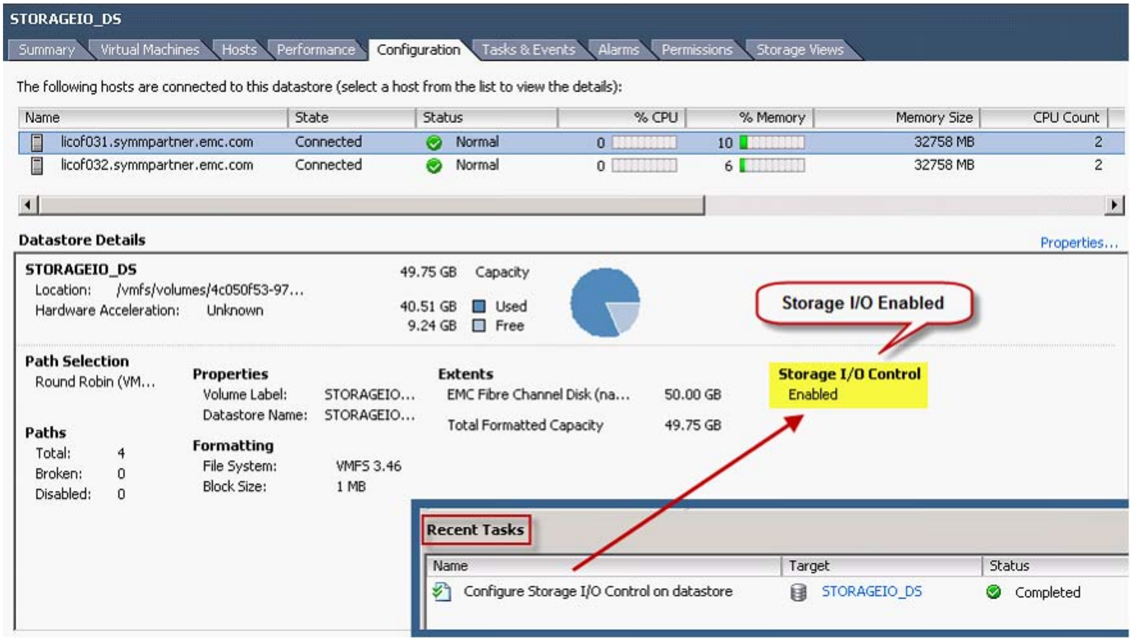
Extent Device
The extent selected on the left resides on the LUN or physical disk described below.

Device	Capacity
EMC Fibre Channel Disk (naa.60...	50.00 GB

nary Partitions	Capacity
. VMFS	50.00 GB

Figure 100 Enabling Storage I/O Control on a datastore

The configuration change will show up as a Recent Task and once complete the Configuration page will show the feature as enabled as seen in Figure 101.



The following hosts are connected to this datastore (select a host from the list to view the details):

Name	State	Status	% CPU	% Memory	Memory Size	CPU Count
licof031.symmpartner.emc.com	Connected	Normal	0	10	32758 MB	2
licof032.symmpartner.emc.com	Connected	Normal	0	6	32758 MB	2

Datastore Details

STORAGEIO_DS
Location: /vmfs/volumes/4c050f53-97...
Hardware Acceleration: Unknown

49.75 GB Capacity
40.51 GB Used
9.24 GB Free

Path Selection
Round Robin (VM...

Paths
Total: 4
Broken: 0
Disabled: 0

Properties
Volume Label: STORAGEIO...
Datastore Name: STORAGEIO...

Formatting
File System: VMFS 3.46
Block Size: 1 MB

Extents
EMC Fibre Channel Disk (na... 50.00 GB
Total Formatted Capacity 49.75 GB

Storage I/O Control
Enabled

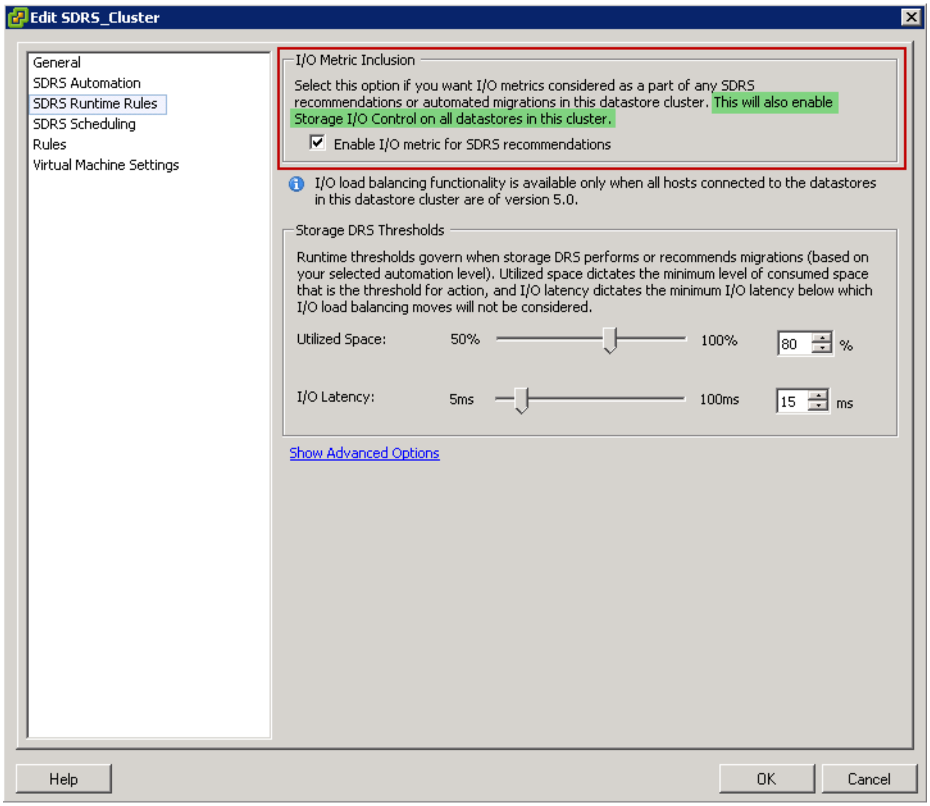
Recent Tasks

Name	Target	Status
Configure Storage I/O Control on datastore	STORAGEIO_DS	Completed

Figure 101 Storage I/O Control enabled on datastore

To enable Storage I/O Control on all datastores of a Datastore Cluster simply check the box to enable the I/O metric for SDRS recommendations as in Figure 102 on page 198.[1]

[1] I/O metrics with SDRS should not be utilized when the underlying Symmetrix storage is managed by FAST. See "Using SDRS with EMC FAST (DP and VP)" on page 209 for more detail.



Edit SDRS_Cluster

- General
- SDRS Automation
- SDRS Runtime Rules
- SDRS Scheduling
- Rules
- Virtual Machine Settings

I/O Metric Inclusion

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

☑ Enable I/O metric for SDRS recommendations

I/O load balancing functionality is available only when all hosts connected to the datastores in this datastore cluster are of version 5.0.

Storage DRS Thresholds

Runtime thresholds govern when storage DRS performs or recommends migrations (based on your selected automation level). Utilized space dictates the minimum level of consumed space that is the threshold for action, and I/O latency dictates the minimum I/O latency below which I/O load balancing moves will not be considered.

Utilized Space: 50% 100% 80 %

I/O Latency: 5ms 100ms 15 ms

Figure 102 **Enabling Storage I/O Control on all datastores of a datastore cluster**

Storage I/O Control resource shares and limits

The second step in the setup of Storage I/O Control is to allocate the number of storage I/O shares and upper limit of I/O operations per second (IOPS) allowed for each virtual machine. When storage I/O congestion is detected for a datastore, the I/O workloads of the virtual machines accessing that datastore are adjusted according to the proportion of virtual machine shares each VM has been allocated.

Shares

Storage I/O shares are similar to those used for memory and CPU resource allocation. They represent the relative priority of a virtual machine with regard to the distribution of storage I/O resources. Under resource contention, VMs with higher share values have greater access to the storage array, which typically results in higher throughput and lower latency. There are three default values for shares: Low (500), Normal (1000), or High (2000) as well as a fourth option, Custom. Choosing Custom will allow a user-defined amount to be set.

IOPS

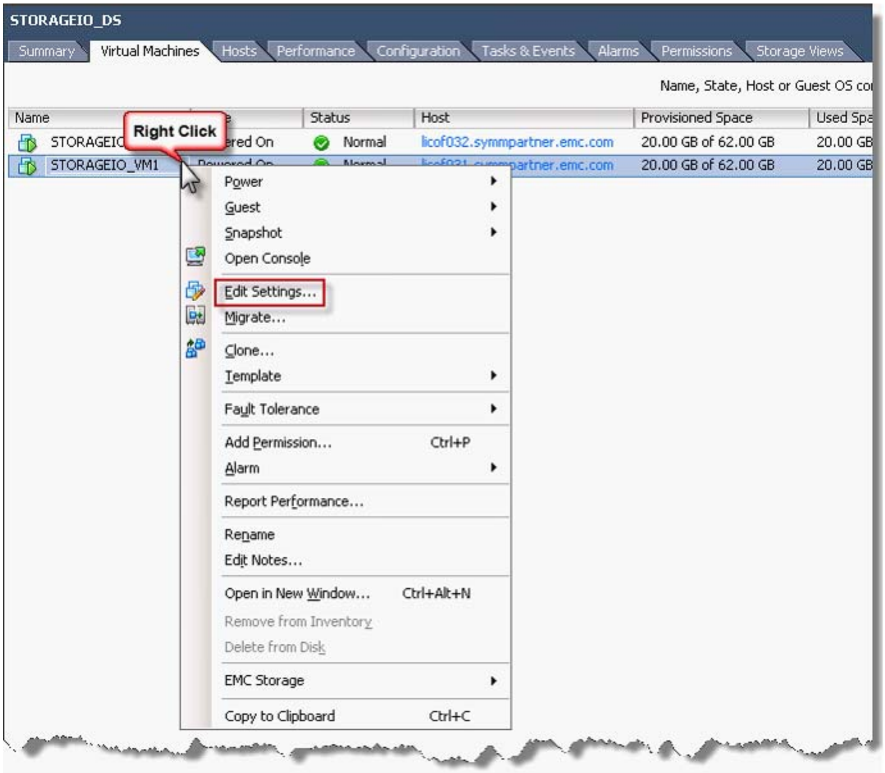
In addition to setting the shares, one can limit the IOPS that are permitted for a VM.[1] By default, the IOPS are always unlimited as they have no direct correlation with the shares setting. If one prefers to limit based upon MB per second, it will be necessary to convert to IOPS using the typical I/O size for that VM. For example, to restrict a backup application with 64KB IOs to 10 MB per second a limit of 160 IOPS (10240000/64000) should be set.

One should allocate storage I/O resources to virtual machines based on priority by assigning a relative amount of shares to the VM. Unless virtual machine workloads are very similar, shares do not necessarily dictate allocation in terms of I/O operations or MBs per second. Higher shares allow a virtual machine to keep more concurrent I/O operations pending at the storage device or datastore compared to a virtual machine with lower shares. Therefore, two virtual machines might experience different throughput based on their workloads.

Setting Storage I/O Control resource shares and limits

The screen to edit shares and IOPS can be accessed from a number of places within the vSphere Client. Perhaps the easiest is to select the Virtual Machines tab of the datastore on which Storage I/O Control has been enabled. In Figure 103 on page 200 there are two VMs listed for the datastore STORAGEIO_DS: STORAGEIO_VM1 and STORAGEIO_VM2. To set the appropriate SIOC policy, the settings of the VM need to be changed. This can be achieved by right-clicking on the chosen VM and select **Edit Settings**.

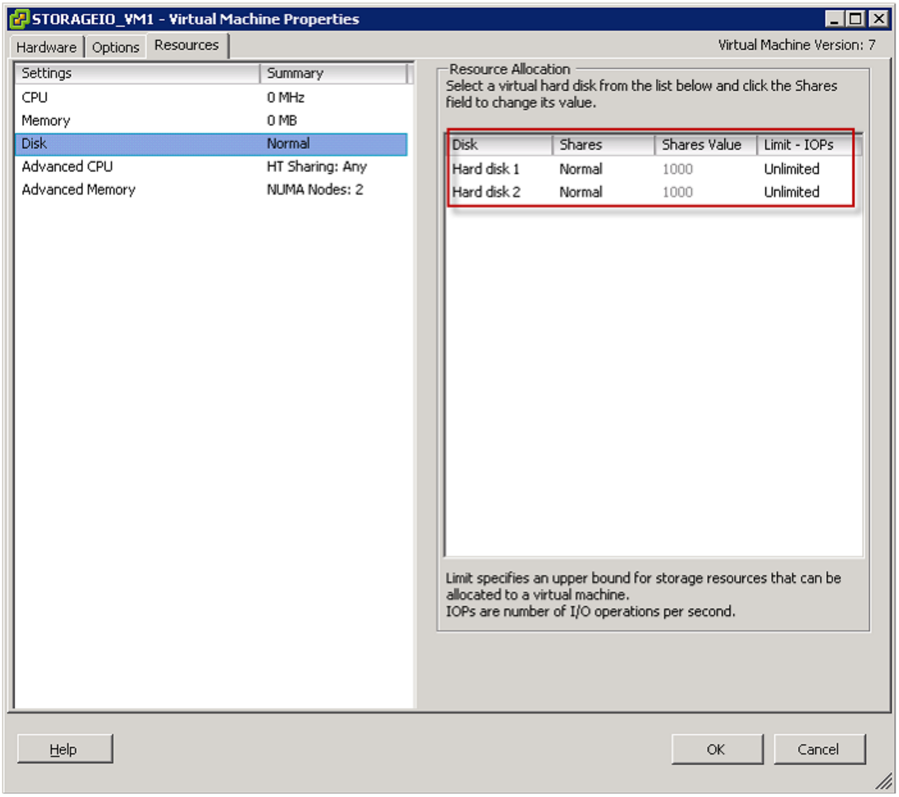
[1] For information on how to limit IOPS at the storage level see “EMC Host I/O Limit” on page 192.



Name	e	Status	Host	Provisioned Space	Used Spa
STORAGEIO	ered On	Normal	licof032.symmpartner.emc.com	20.00 GB of 62.00 GB	20.00 GB
STORAGEIO_VM1	Powered On	Normal	licof031.symmpartner.emc.com	20.00 GB of 62.00 GB	20.00 GB

Figure 103 Edit settings for VM for SIOC

Within the VM Properties screen shown in Figure 104 on page 201, select the **Resources** tab. One will notice that this tab is also where the CPU and memory resources can be allocated in a DRS cluster. Choose the Disk Setting in the left-hand panel and notice that for each hard disk on the right the default of Normal for Shares and Unlimited for IOPS is present.



Virtual Machine Version: 7

Settings	Summary
CPU	0 MHz
Memory	0 MB
Disk	Normal
Advanced CPU	HT Sharing: Any
Advanced Memory	NUMA Nodes: 2

Resource Allocation

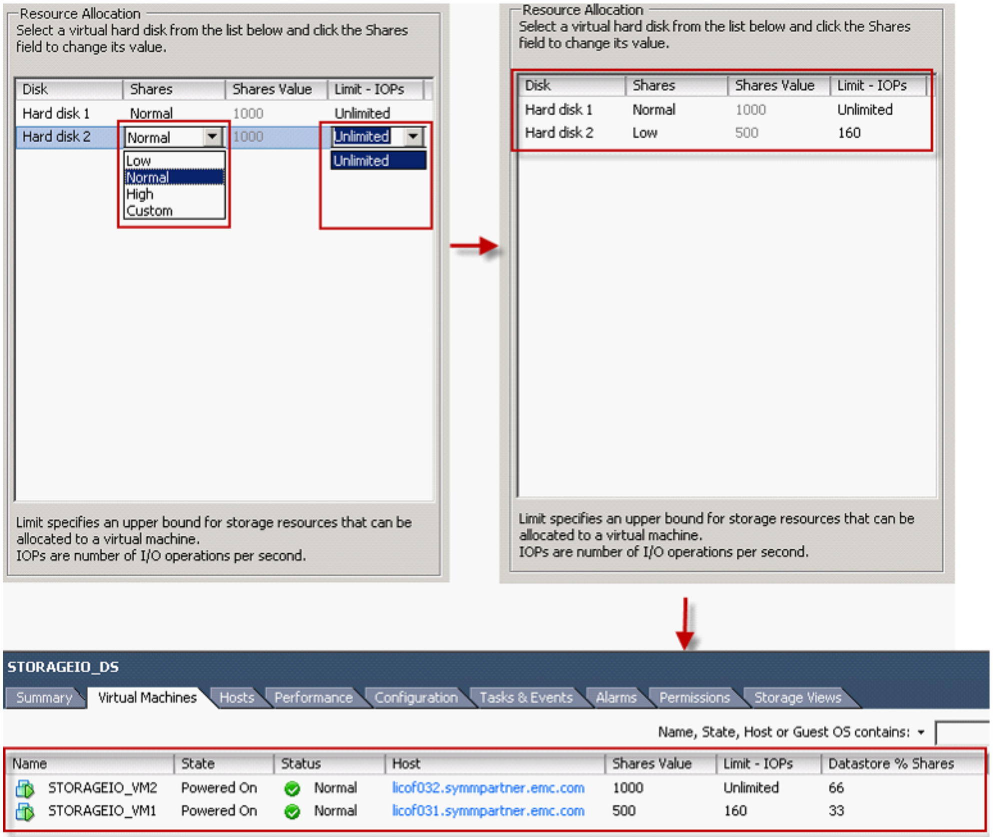
Select a virtual hard disk from the list below and click the Shares field to change its value.

Disk	Shares	Shares Value	Limit - IOPs
Hard disk 1	Normal	1000	Unlimited
Hard disk 2	Normal	1000	Unlimited

Limit specifies an upper bound for storage resources that can be allocated to a virtual machine.
IOPs are number of I/O operations per second.

Figure 104 Default VM disk resource allocation

Using the drop-down boxes within the Shares and IOPS columns, one can change the values. Figure 105 on page 202 shows that the shares for Hard disk 2 are changed to Low, while the IOPS value has been set to 160. Selecting the OK button at the bottom of the VM Properties screen commits the changes. The new values are reflected in the **Virtual Machines** tab, as shown in Figure 105 on page 202.



Resource Allocation

Select a virtual hard disk from the list below and click the Shares field to change its value.

Disk	Shares	Shares Value	Limit - IOPs
Hard disk 1	Normal	1000	Unlimited
Hard disk 2	Normal	1000	Unlimited

Limit specifies an upper bound for storage resources that can be allocated to a virtual machine.
IOPs are number of I/O operations per second.

Resource Allocation

Select a virtual hard disk from the list below and click the Shares field to change its value.

Disk	Shares	Shares Value	Limit - IOPs
Hard disk 1	Normal	1000	Unlimited
Hard disk 2	Low	500	160

Limit specifies an upper bound for storage resources that can be allocated to a virtual machine.
IOPs are number of I/O operations per second.

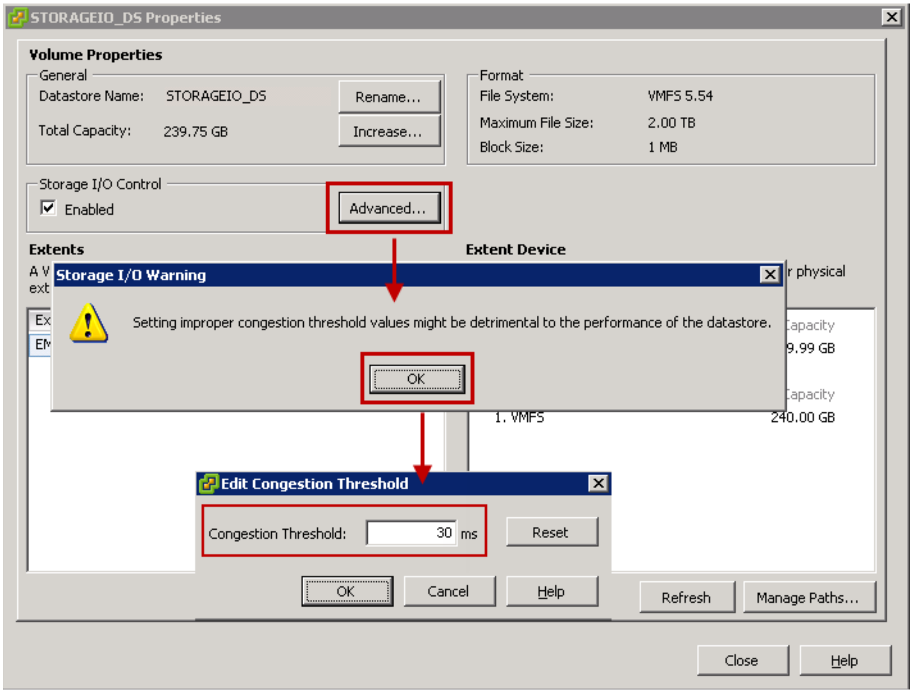
Name	State	Status	Host	Shares Value	Limit - IOPs	Datastore % Shares
STORAGEIO_VM2	Powered On	Normal	licof032.symmpartner.emc.com	1000	Unlimited	66
STORAGEIO_VM1	Powered On	Normal	licof031.symmpartner.emc.com	500	160	33

Figure 105 Changing VM resource allocation

Changing Congestion Threshold for Storage I/O Control

Although it is possible to change the Congestion Threshold, i.e. the value at which VMware begins implementing the shares and IOPS settings for a VM, in general it is not recommended. The current default value is 30 ms. There may be certain cases that necessitate it, however. If, for instance, the datastore is backed by an EFD LUN and the response time requirement for an application is lower than 30 milliseconds; or conversely if the datastore is backed by a SATA LUN and the response time requirement is more lax. In these cases it might be beneficial to alter the threshold.

In Figure 106 the location to change the Congestion Threshold is shown. Following the same process as when enabling Storage I/O Control ("Enable Storage I/O Control" on page 196), enter into the Properties of the datastore. Select the Advanced button next to the Enabled check box. Immediately, the user is presented with a warning message about changing the setting. Select OK. The Congestion Threshold value can then be manually changed.



STORAGEIO_DS Properties

Volume Properties

General

Datastore Name: STORAGEIO_DS

Total Capacity: 239.75 GB

Format

File System: VMFS 5.54

Maximum File Size: 2.00 TB

Block Size: 1 MB

Storage I/O Control

☑ Enabled

Storage I/O Warning

Setting improper congestion threshold values might be detrimental to the performance of the datastore.

Edit Congestion Threshold

Congestion Threshold: 30 ms

Figure 106 Changing the congestion threshold

Storage I/O Control requirements

The following requirements and limitations apply to Storage I/O Control:

- Datastores that are Storage I/O Control-enabled must be managed by a single vCenter Server system.
- Storage I/O Control is supported on Fibre Channel-connected, iSCSI-connected, and NFS-connected storage (only in vSphere 5). Raw Device Mapping (RDM) is not supported.

- Storage I/O Control does not support datastores with multiple extents.
- Congestion Threshold has a limiting range of 10 to 100 milliseconds.

Note: EMC does not recommend making changes to the Congestion Threshold.

Spreading I/O demands

Storage I/O Control was designed to help alleviate many of the performance issues that can arise when many different types of virtual machines share the same VMFS volume on a large SCSI disk presented to the VMware ESX hosts. This technique of using a single large disk allows optimal use of the storage capacity.

However, this approach can result in performance issues if the I/O demands of the virtual machines cannot be met by the large SCSI disk hosting the VMFS, regardless if SIOC is being used. In order to prevent these performance issues from developing, no matter the size of the LUN, EMC recommends using Symmetrix Virtual Provisioning, which will spread the I/O over a large pool of disks. A detailed discussion of Virtual Provisioning can be found in "EMC Symmetrix Virtual Provisioning overview" on page 135.

VMware Storage Distributed Resource Scheduler (SDRS)

Datastore clusters

With vSphere 5, a new VMware vCenter object called a datastore cluster is introduced. A datastore cluster is what Storage DRS acts upon. A datastore cluster is a collection of datastores with shared resources and a shared management interface. Datastore clusters are to datastores what clusters are to hosts. When you create a datastore cluster, you can use vSphere Storage DRS to manage storage resources.

When a datastore cluster is created, Storage DRS can manage the storage resources comparably to how vSphere DRS manages compute resources in a cluster. As with a cluster of hosts, a datastore cluster is used to aggregate storage resources, enabling smart and rapid placement of new virtual machines and virtual disk drives as well as load balancing of existing workloads. It is important to note that Storage DRS does not have to be enabled on a datastore cluster, when Storage DRS is not enabled, a datastore cluster is essentially just a folder to group datastores.

Storage DRS provides initial placement and ongoing balancing recommendations to datastores in a Storage DRS-enabled datastore cluster. Initial placement occurs when Storage DRS selects a datastore within a datastore cluster on which to place a virtual machine disk. This happens when the virtual machine is being created or cloned, when a virtual machine disk is being migrated to another datastore cluster, or when you add a disk to an existing virtual machine. Initial placement recommendations are made in accordance with space constraints and with respect to the goals of space and I/O load balancing.

These goals aim to minimize the risk of over-provisioning one datastore, storage I/O bottlenecks, and performance impact on virtual machines.

Storage DRS is invoked at the configured frequency (by default, every eight hours) or when one or more datastores in a datastore cluster exceeds the user-configurable space utilization or I/O latency thresholds. When Storage DRS is invoked, it checks each datastore's space utilization and I/O latency values against the threshold.

For I/O latency, Storage DRS uses the 90th percentile I/O latency measured over the course of a day to compare against the threshold. Storage DRS applies the datastore utilization reporting mechanism of VMware vCenter Server, to make recommendations whenever the configured utilized space threshold is exceeded. Storage DRS will calculate all possible moves, to balance the load accordingly while considering the cost and the benefit of the migration. Storage DRS considers moving virtual machines that are powered off or powered on for space balancing. Storage DRS includes powered-off virtual machines with snapshots in these considerations.

Affinity rules and maintenance mode

Storage DRS affinity rules enable control over which virtual disks should or should not be placed on the same datastore within a datastore cluster. By default, a virtual machine's virtual disks are kept together on the same datastore. Storage DRS offers three types of affinity rules:

- **VMDK Anti-Affinity** —Virtual disks of a given virtual machine with multiple virtual disks are always placed on different datastores.
- **VMDK Affinity**—Virtual disks are always kept together on the same datastore.
- **VM Anti-Affinity**—Two specified virtual machines, including associated disks, are always placed on different datastores.

In addition, Storage DRS offers Datastore Maintenance Mode, which automatically evacuates all registered virtual machines and virtual disk drives from the selected datastore to the remaining datastores in the datastore cluster. This is shown in Figure 107 on page 207.

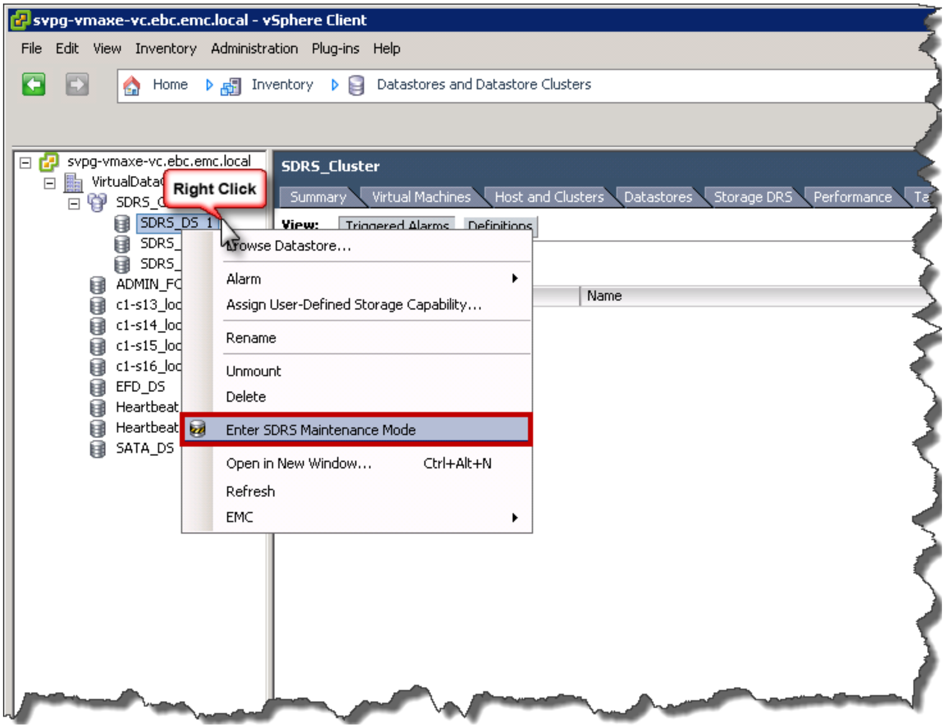


Figure 107 Putting a datastore in maintenance mode in SDRS

Datastore cluster requirements

Datastores and hosts that are associated with a datastore cluster must meet certain requirements to use datastore cluster features successfully. EMC strongly encourages using the vSphere vStorage APIs for Storage Awareness (VASA) integration within vSphere to ensure the environment adheres to the guidelines outlined below. For detailed information on implementing VASA with Symmetrix storage please see "vSphere vStorage APIs for Storage Awareness (VASA)" on page 122.

Follow these guidelines when you create a datastore cluster:

- Datastore clusters should contain similar or interchangeable datastores.

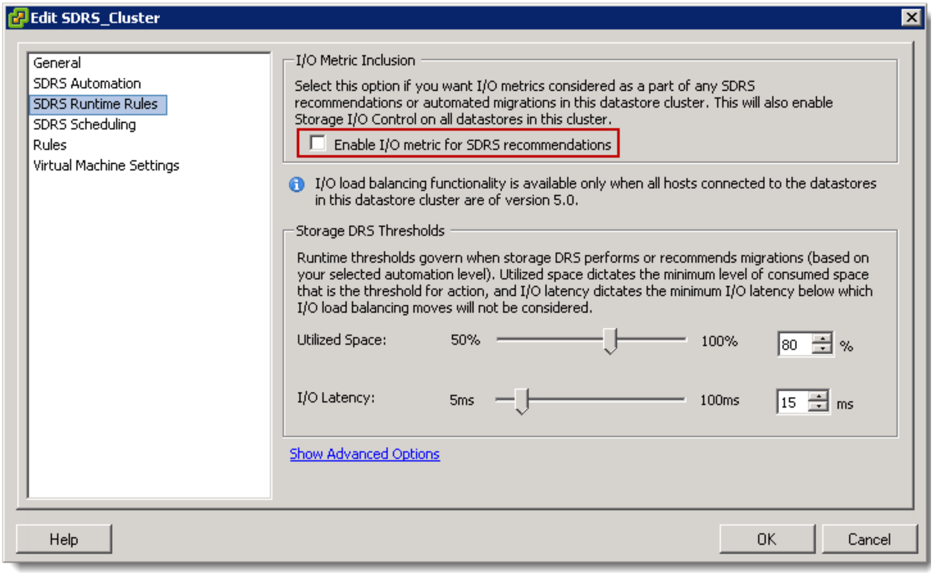
- A datastore cluster can contain a mix of datastores with different sizes and I/O capacities, and can be from different arrays and vendors. Nevertheless, EMC does not recommend mixing datastores backed by devices that have different properties, i.e. different RAID types or disk technologies, unless the devices are part of a FAST VP policy.
- NFS and VMFS datastores cannot be combined in the same datastore cluster.
- If VMFS volumes are protected by TimeFinder, do not mix and match protected and non-protected devices in a datastore cluster. If SDRS is enabled, only manual mode is supported with datastores protected by TimeFinder.
- All hosts attached to the datastores in a datastore cluster must be ESXi 5 and later. If datastores in the datastore cluster are connected to ESX/ESXi 4.x and earlier hosts, Storage DRS does not run.
- Datastores shared across multiple datacenters cannot be included in a datastore cluster.
- As a best practice, do not include datastores that have hardware acceleration enabled in the same datastore cluster as datastores that do not have hardware acceleration enabled. Datastores in a datastore cluster must be homogeneous to guarantee consistent hardware acceleration-supported behavior.
- Replicated datastores cannot be combined with non-replicated datastores in the same Storage DRS enabled datastore cluster. If SDRS is enabled, only manual mode is supported with replicated datastores.
  - Do not mix and match replication modes and solutions within a datastore cluster. For example, do not have SRDF/A and SRDF/S protected devices within the same datastore cluster or mix EMC RecoverPoint and SRDF devices.
  - To maintain proper consistency, if a device is contained in a consistency group and that device is part of a datastore cluster, do not add devices into that datastore cluster that are not protected by that consistency group. This will make sure that a VM that has consistency dependencies on other VMs does not move out of the consistency group due to a Storage DRS move. In general, a datastore cluster should have a one to one relationship with a consistency group.

Site Recovery Manager Interoperability with Storage DRS and Storage vMotion

Due to some specific and limited cases where recoverability can be compromised during storage movement, VMware vCenter Site Recovery Manager 5.x is not supported for use with Storage vMotion and is not supported for use with the Storage Distributed Resource Scheduler (SDRS) including the use of datastore clusters.

Using SDRS with EMC FAST (DP and VP)

When EMC FAST is used in conjunction with SDRS, only capacity-based SDRS is recommended and not storage I/O load balancing. To use this configuration, simply uncheck the box labeled "Enable I/O metric for SDRS recommendations" when creating the datastore cluster or after creation by editing it. This is seen in Figure 108. Unchecking this box will ensure that only capacity-based SDRS is in use.



I/O Metric Inclusion

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

☐ Enable I/O metric for SDRS recommendations

I/O load balancing functionality is available only when all hosts connected to the datastores in this datastore cluster are of version 5.0.

Storage DRS Thresholds

Runtime thresholds govern when storage DRS performs or recommends migrations (based on your selected automation level). Utilized space dictates the minimum level of consumed space that is the threshold for action, and I/O latency dictates the minimum I/O latency below which I/O load balancing moves will not be considered.

Utilized Space: 50% 100% 80 %

I/O Latency: 5ms 100ms 15 ms

Show Advanced Options

Figure 108 **Disabling performance-based SDRS**

Unlike FAST DP, which operates on thick devices at the whole device level, FAST VP operates on thin devices at the far more granular extent level. Because FAST VP is actively managing the data on disks, knowing the performance requirements of a VM (on a datastore under FAST VP control) is important before a VM is migrated from one datastore to another. This is because the exact thin pool distribution of the VM's data may not be the same as it was before the move (though it could return to it at some point assuming the access patterns and FAST VP policy do not change).

Therefore, if the VM houses performance sensitive applications, EMC advises not using SDRS with FAST VP for that VM. Preventing movement could be accomplished by setting up a rule or at the very least using Manual Mode (no automation) for SDRS.

SDRS and EMC Host I/O Limit

It is important to take note when EMC Host I/O Limits are set on storage groups that contain devices that are part of a datastore cluster (as part of a SDRS implementation). When SDRS is utilized all devices that are in that datastore cluster should have the same Host I/O Limit. Moreover, the devices should come from the same EMC array. This recommendation is given because the Host I/O Limit throttles I/O to those devices. This may be done to limit the ability of applications on those devices from impacting other devices on the array or perhaps to simplify QoS management. Whatever the reason, if a datastore cluster contains devices - whether from the same or a different array - that do not have a Host I/O limit, there is always the possibility in the course of its balancing that SDRS will relocate virtual machines on those limited devices to non-limited devices. Such a change might alter the desired QoS or permit the applications on the virtual machines to exceed the desired throughput. It is therefore prudent to have device homogeneity when using EMC Host I/O Limits in conjunction with SDRS.

5

Cloning of vSphere Virtual Machines

This chapter discusses the use of EMC technology to clone VMware virtual machines.

Introduction

VMware ESX virtualizes IT assets into a flexible, cost-effective pool of compute, storage, and networking resources. These resources can be then mapped to specific business needs by creating virtual machines. VMware ESX provides several utilities to manage the environment. This includes utilities to clone, back up and restore virtual machines. All these utilities use host CPU resources to perform the functions. Furthermore, the utilities cannot operate on data not residing in the VMware environment.

All businesses have line-of-business operations that interact and operate with data on a disparate set of applications and operating systems. These enterprises can benefit by leveraging technology offered by storage array vendors to provide alternative methodologies to protect and replicate the data. The same storage technology can also be used for presenting various organizations in the business with a point-in-time view of their data without any disruption or impact to the production workload.

VMware ESX can be used in conjunction with SAN-attached EMC Symmetrix storage arrays and the advanced storage functionality they offer. The configuration of virtualized servers when used in conjunction with EMC Symmetrix storage array functionality is not very different from the setup used if the applications were running on a physical server. However, it is critical to ensure proper configuration of both the storage array and of VMware ESX so applications in the virtual environment can exploit storage array functionality. This chapter will include the use of EMC TimeFinder with VMware ESX to clone virtual machines and their data. It will also include the benefits of offloading software-based VMware cloning operations to the array that can be achieved using “Full Copy”, which is one of the primitives available through VMware vStorage APIs for Array Integration, or VAAI.

EMC TimeFinder overview

The TimeFinder family of products are Symmetrix local replication solutions designed to non-disruptively create point-in-time copies of critical data. You can configure backup sessions, initiate copies, and terminate TimeFinder operations from mainframe and open systems controlling hosts using EMC Symmetrix host-based control software.

The TimeFinder local replication solutions include TimeFinder/Clone, TimeFinder/Snap, and TimeFinder VP Snap.[1] TimeFinder/Clone creates full-device and extent-level point-in-time copies. TimeFinder/Snap creates pointer-based logical copies that consume less storage space on physical drives. TimeFinder VP Snap provides the efficiency of Snap technology with improved cache utilization and simplified pool management.

Each solution guarantees high data availability. The source device is always available to production applications. The target device becomes read/write enabled as soon as you initiate the point-in-time copy. Host applications can therefore immediately access the point-in-time image of critical data from the target device while TimeFinder copies data in the background. Detailed description of these products is available in the *EMC Symmetrix TimeFinder Product Guide* on www.Powerlink.EMC.com.

TimeFinder products run on the EMC Symmetrix storage array. However, the management of the functionality is performed using EMC Solutions Enabler software (with appropriate add-on modules), Symmetrix Management Console (SMC), Unisphere for VMAX (Unisphere), or EMC Ionix ControlCenter.

TimeFinder VP Snap

VP Snap leverages TimeFinder/Clone technology to create space-efficient snaps for thin devices by allowing multiple sessions to share capacity allocations within a thin pool. VP Snap provides the

[1] Prior to Enginuity 5874, TimeFinder/Mirror was also an option for local replication; however native TimeFinder/Mirror is no longer supported on Symmetrix VMAX Series arrays and Enginuity levels 5874 and higher. TimeFinder/Mirror scripts may still be used with Symmetrix VMAX Series arrays by running TimeFinder/Clone in emulation mode. Note that emulation mode uses TimeFinder/Clone and not the TimeFinder/Mirror infrastructure.

efficiency of Snap technology with improved cache utilization and simplified pool management. With VP Snap, tracks can be stored in the same thin pool as the source, or in another pool of your choice.

VP Snap sessions copy data from the source device to the target device only if triggered by a host I/O. Read I/Os to protected tracks on the target device do not result in data being copied.

Figure 109 shows several VP Snap sessions sharing allocations within a thin pool.

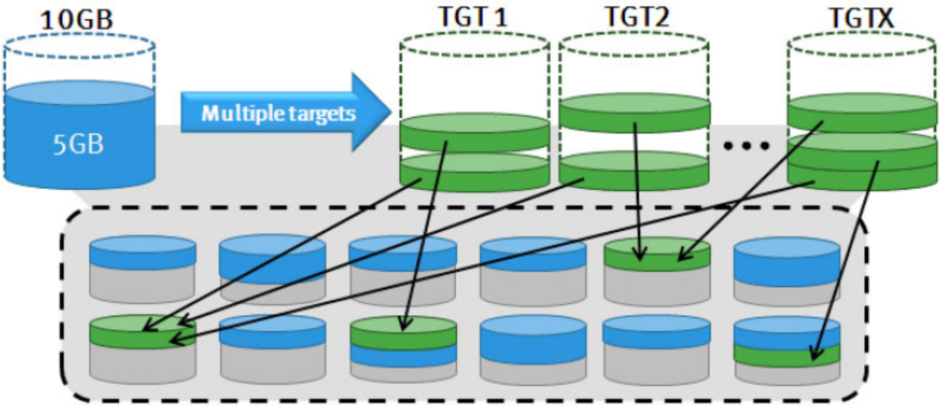


Figure 109 VP Snap sessions sharing allocations within a thin pool

Note: VP Snap source and target devices are optimized by FAST VP, but the shared target allocations are not moved.

There are certain restrictions when using VP Snap:

- The -vse attribute may only be applied during the creation of the session.
- Both the source device and the target device must be thin devices.
- All of the VP Snap target devices for a particular source device must be bound to the same thin pool.
- Once created, VP Snap sessions cannot be changed to any other mode.
- Copying from a source device to a larger target device is not supported with VP Snap.

- VP Snap sessions cannot be combined with TimeFinder/Clone nocopy sessions on the same source device.
- If a FAST VP target extent is part of a VP Snap session, the shared tracks cannot be moved between tiers.

TimeFinder and VMFS

The TimeFinder family of products operates on Symmetrix devices. Using TimeFinder on a Symmetrix device containing a VMFS requires that all extents of the file system be replicated. If proper planning and procedures are not followed, this requirement forces replication of all data that is present on a VMFS, including virtual disk images that are not needed. Therefore, EMC recommends separation of virtual machines that require use of storage-array-based replication on one or more VMFS volumes. This does not completely eliminate replication of unneeded data, but minimizes the storage overhead. The storage overhead can be eliminated by the use of RDMs on virtual machines that require storage-array-based replication.

VMware ESX allows creation of VMFS on partitions of the physical devices. Furthermore, VMFS supports spanning of the file systems across partitions, so it is possible, for example, to create a VMFS with part of the file system on one partition on one disk and another part on a different partition on a different disk. However, such designs complicate the management of the VMware storage environment by introducing unnecessary intricacies. If use of EMC TimeFinder technology (or EMC SRDF) with VMFS is desired, EMC strongly recommends creating only one partition per physical disk, and one device for each VMFS datastore.

Each TimeFinder software product mentioned in the earlier paragraphs has different performance and availability characteristics; however, functionally all variants of TimeFinder software essentially are being used to create a copy. Therefore, each individual product will not be covered in detail herein. Where appropriate, mention will be made of any important differences in the process.

Furthermore, with the introduction of the Symmetrix VMAX/VMAXe Enginuity 5875 code, EMC offers new storage integrations with VMware vSphere 4.1 and 5 enabling customers to dramatically improve efficiency in their VMware environment. VMware vSphere 4.1 and 5 include VMware vStorage APIs for Array Integration (VAAI) that permit the offloading of specific storage operations to EMC Symmetrix VMAX/VMAXe to increase both overall system performance and efficiency. One of these primitives, Full Copy is tremendously helpful in reducing the time to create VMware clones. This feature delivers hardware-accelerated copying of data by performing all duplication and migration operations on the array. Using VAAI customers can achieve considerably faster data movement through VMware Storage vMotion, as well as virtual machine creation, deployment from templates, and virtual machine cloning.

Copying virtual machines after shutdown

Ideally, virtual machines should be shut down before the metadata and virtual disks associated with the virtual machines are copied. Copying virtual machines after shut down ensures a clean copy of the data that can be used for back up or quick startup of the cloned virtual machine.

Using TimeFinder/Clone with cold virtual machines

TimeFinder/Clone provides the flexibility of copying any standard device on the Symmetrix storage array to another standard device of equal or larger size in the storage array.[1] TimeFinder/Clone, by removing the requirement of having devices with a special BCV attribute, offers customers greater flexibility and functionality to clone production data.

Note: For the types of supported devices with TimeFinder/Clone see the *EMC Symmetrix TimeFinder Product Guide* on www.Powerlink.EMC.com.

When the clone relationship between the source and target devices is created, the target devices are presented as not ready (NR) to any host that is accessing the volumes. It is for this reason that EMC recommends unmounting any VMFS datastores that are currently on those devices before creating the clone. The creation operation makes the VMFS on the target unavailable to the VMware ESX hosts. Therefore, the recommendation of unmounting the datastores before performing an establish operation applies directly to the virtual machines that are impacted by the absence of those datastores. The target devices in a TimeFinder/Clone operation can be accessed by VMware ESX as soon as the clone relationship is activated. However, accessing the data on the target volume before the copy process completes causes a small performance impact since the data is copied from the source device before being presented to the requesting host. Similarly, writes to tracks on the source device not copied to the target device experiences a small performance impact.

[1] The full set of features of TimeFinder/Clone are not supported when going from a smaller source to a larger target volume. For more information on these restrictions refer to the *EMC Solutions Enabler Symmetrix TimeFinder Family CLI Product Guide*.

Note: TimeFinder/Clone technology also offers the option of copying the data as soon as the TimeFinder/Clone relationship is created (by using the keyword –precopy instead of –copy). This option mitigates the performance impact that is experienced when the TimeFinder/Clone pairs are activated.

Organizations use a cloned virtual machine image for different purposes. For example, a cloned virtual machine may be configured for reporting activities during the day and for backups in the night. VMware ESX does not allow virtual machines to power on if any of the virtual disk devices configured on it are not available (this includes reasons such as the storage device is not presented to the ESX or the device is in the "Not Ready" state). When clone targets are in a created state, the VMFS volume on the clone target is unavailable to the VMware ESX, and any virtual machine using that VMFS volume cannot be powered on. This restriction must be considered when designing virtual environments that re-provision cloned virtual machines.

Virtual machines running on VMware ESX can be powered on with RDMs (in either virtual or physical compatibility mode) that map to devices that are in a "not ready" state. However, the VMFS holding the configuration file, the metadata information about the virtual machine, and the virtual disk mapping files have to be available for the power-on operation.

Copying cold virtual machines on VMFS using TimeFinder/Clone

TimeFinder/Clone devices are managed as an atomic unit by creating a device or composite group. Composite groups are used if the virtual machines that are being cloned use VMFS volumes from multiple Symmetrix storage arrays. SMC, Unisphere or Solutions Enabler commands can be used to create the device or composite group, and manage the copying process.

The next few paragraphs present the steps required to clone a group of virtual machines utilizing EMC TimeFinder/Clone technology:

1. Identify the device number of the Symmetrix volumes used by the VMFS by utilizing EMC Virtual Storage Integrator for VMware vSphere (EMC VSI) which maps physical extents of a VMFS to the Symmetrix device number. A device group containing the member(s) of the VMFS should be created. This can be done with the following commands:

```
symdg create <group_name>
```

```
symld -g <group_name> -sid <Symmetrix SN> add dev <dev
#>
```

where **<dev #>** are the devices identified by EMC VSI.

2. The devices that hold the copy of the data need to be added or associated with the group. Note that TimeFinder/Clone can be used when devices with the BCV attributes are used as a target. If the target devices have the BCV attribute turned on, the devices should be associated with the device group using the command

   ```
   symbcv -g <group_name> associate dev <dev #>
   ```

   When STD devices are used as the target, the devices need to be added to the devices using the **symld** command along with the `-tgt` switch[1]

   ```
   symld -g <group_name> -sid <Symmetrix SN> -tgt add dev
   <dev #>
   ```

3. The next step in the process of copying virtual machines using EMC TimeFinder/Clone is the creation of TimeFinder/Clone pairs. The command to perform this activity depends on the attributes of the target devices. If the target devices have the BCV attribute, the command

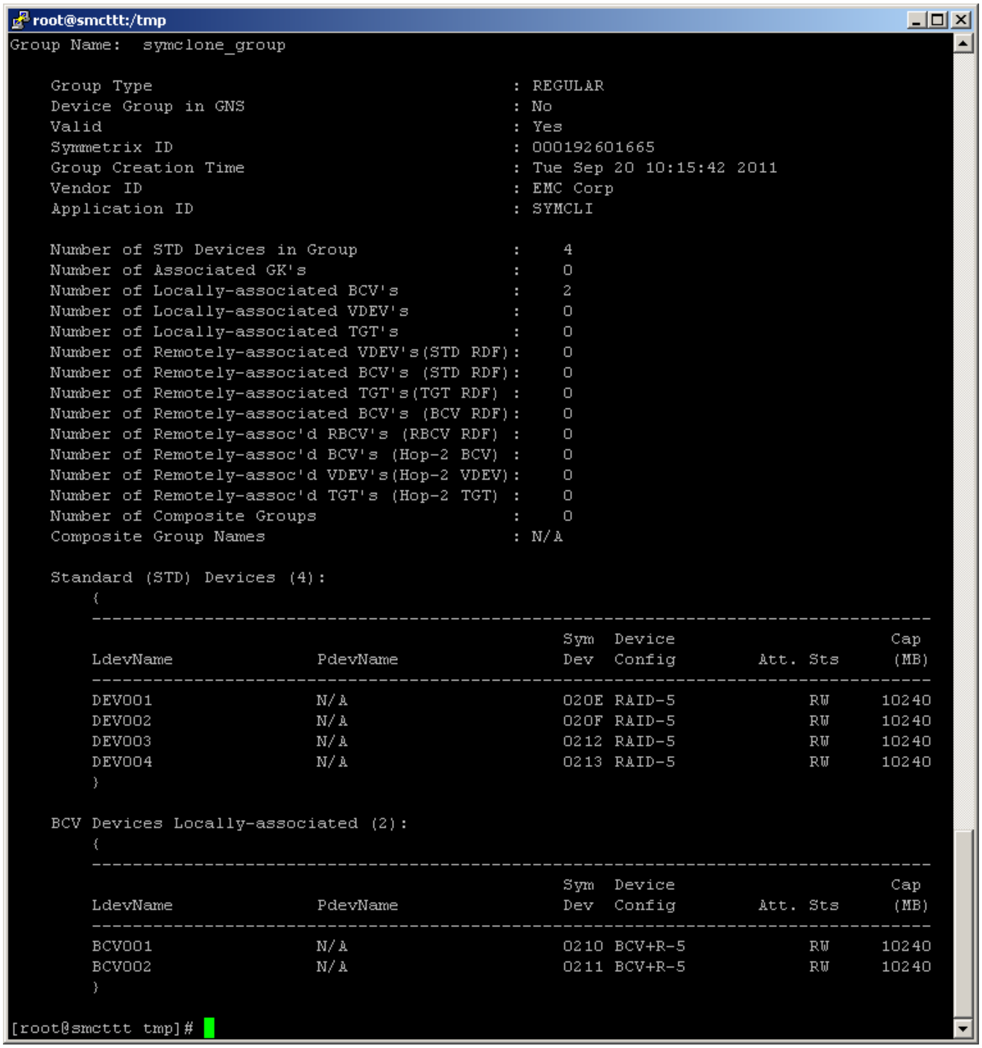
   ```
   symclone -g <group_name> -differential -copy -noprompt
      create
   ```

   automatically creates a TimeFinder/Clone relationship between the source and target devices. If STD devices are used, the `-tgt` switch is required to the above command.

   Note: The –differential option is required if incremental reestablish or restore of data is desired. Starting with Solutions Enabler version 7.0 the -differential option became the default.

[1] Solutions Enabler version 6.3 and later introduced a new logical device type, TGT. Standard devices that would be used as a target device can be added to a device group (or composite group) by using the symld command with the -tgt keyword. When standard devices are added using this option, it is not necessary to explicitly define the relationship between the source and target device when the TimeFinder/Clone pairs are created.

Figure 110 shows a devices group that consists of two standard devices and two BCV devices.



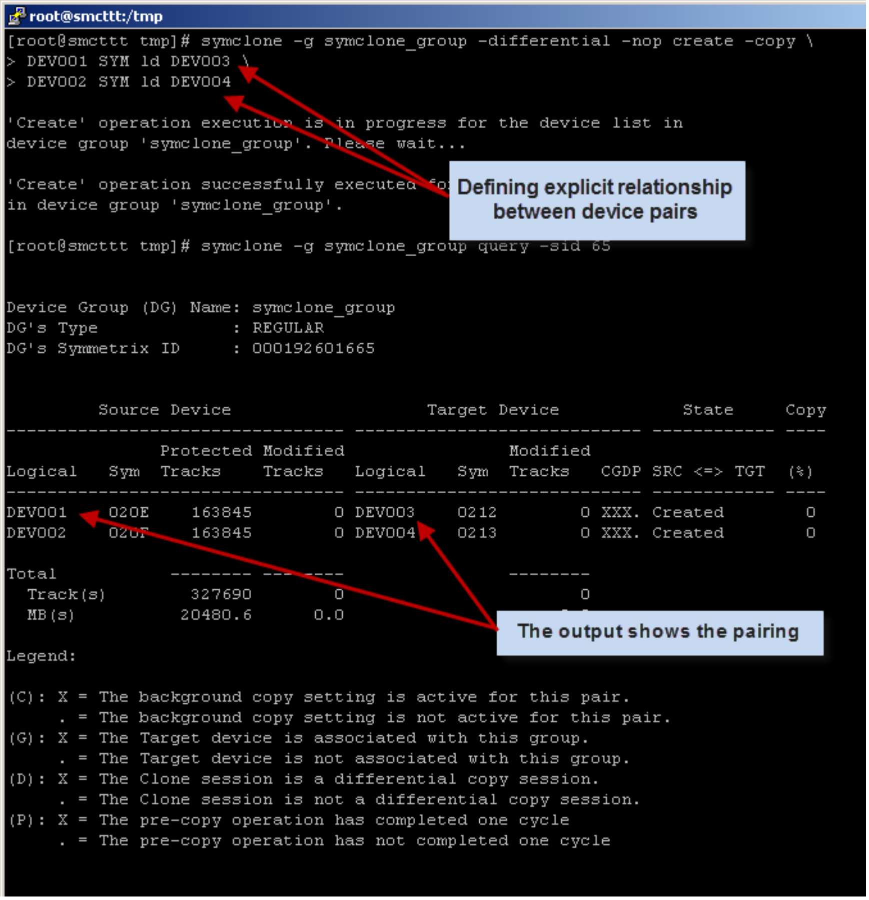
root@smcttt:/tmp

```
Group Name:  symclone_group

    Group Type                                   : REGULAR
    Device Group in GNS                          : No
    Valid                                        : Yes
    Symmetrix ID                                 : 000192601665
    Group Creation Time                          : Tue Sep 20 10:15:42 2011
    Vendor ID                                    : EMC Corp
    Application ID                               : SYMCLI

    Number of STD Devices in Group               :    4
    Number of Associated GK's                    :    0
    Number of Locally-associated BCV's           :    2
    Number of Locally-associated VDEV's          :    0
    Number of Locally-associated TGT's           :    0
    Number of Remotely-associated VDEV's(STD RDF):    0
    Number of Remotely-associated BCV's (STD RDF):    0
    Number of Remotely-associated TGT's(TGT RDF) :    0
    Number of Remotely-associated BCV's (BCV RDF):    0
    Number of Remotely-assoc'd RBCV's (RBCV RDF) :    0
    Number of Remotely-assoc'd BCV's (Hop-2 BCV) :    0
    Number of Remotely-assoc'd VDEV's(Hop-2 VDEV):    0
    Number of Remotely-assoc'd TGT's (Hop-2 TGT) :    0
    Number of Composite Groups                   :    0
    Composite Group Names                        : N/A

    Standard (STD) Devices (4):
        {
        ----------------------------------------------------------------------------------
                                                      Sym  Device                     Cap
        LdevName              PdevName                Dev  Config        Att. Sts     (MB)
        ----------------------------------------------------------------------------------
        DEV001                N/A                     020E RAID-5             RW     10240
        DEV002                N/A                     020F RAID-5             RW     10240
        DEV003                N/A                     0212 RAID-5             RW     10240
        DEV004                N/A                     0213 RAID-5             RW     10240
        }

    BCV Devices Locally-associated (2):
        {
        ----------------------------------------------------------------------------------
                                                      Sym  Device                     Cap
        LdevName              PdevName                Dev  Config        Att. Sts     (MB)
        ----------------------------------------------------------------------------------
        BCV001                N/A                     0210 BCV+R-5            RW     10240
        BCV002                N/A                     0211 BCV+R-5            RW     10240
        }

[root@smcttt tmp]#
```

Figure 110 Displaying the contents of a device group

The Symmetrix devices, 20E and 20F, contain the VMFS that needs to be copied. An explicit relationship between device 20E and 20F, and 212 and 213 can be created as shown in Figure 111 on page 221. When the TimeFinder/Clone pairs are created no data is copied (see note on page 218 for an exception to this rule).

However, the target devices are changed to a not ready (NR) state.



```
[root@smcttt tmp]# symclone -g symclone_group -differential -nop create -copy \
> DEV001 SYM ld DEV003 \
> DEV002 SYM ld DEV004

'Create' operation execution is in progress for the device list in
device group 'symclone_group'. Please wait...

'Create' operation successfully executed fo
in device group 'symclone_group'.

[root@smcttt tmp]# symclone -g symclone_group query -sid 65


Device Group (DG) Name: symclone_group
DG's Type             : REGULAR
DG's Symmetrix ID     : 000192601665

         Source Device                   Target Device            State      Copy
--------------------------------- ---------------------------- ------------ ----
               Protected Modified                Modified
Logical   Sym  Tracks    Tracks   Logical   Sym  Tracks   CGDP SRC <=> TGT  (%)
--------------------------------- ---------------------------- ------------ ----
DEV001    020E    163845        0 DEV003    0212        0 XXX. Created        0
DEV002    020F    163845        0 DEV004    0213        0 XXX. Created        0

Total           -------- --------                --------
  Track(s)        327690        0                       0
  MB(s)          20480.6      0.0

Legend:

(C): X = The background copy setting is active for this pair.
     . = The background copy setting is not active for this pair.
(G): X = The Target device is associated with this group.
     . = The Target device is not associated with this group.
(D): X = The Clone session is a differential copy session.
     . = The Clone session is not a differential copy session.
(P): X = The pre-copy operation has completed one cycle
     . = The pre-copy operation has not completed one cycle
```

Defining explicit relationship between device pairs

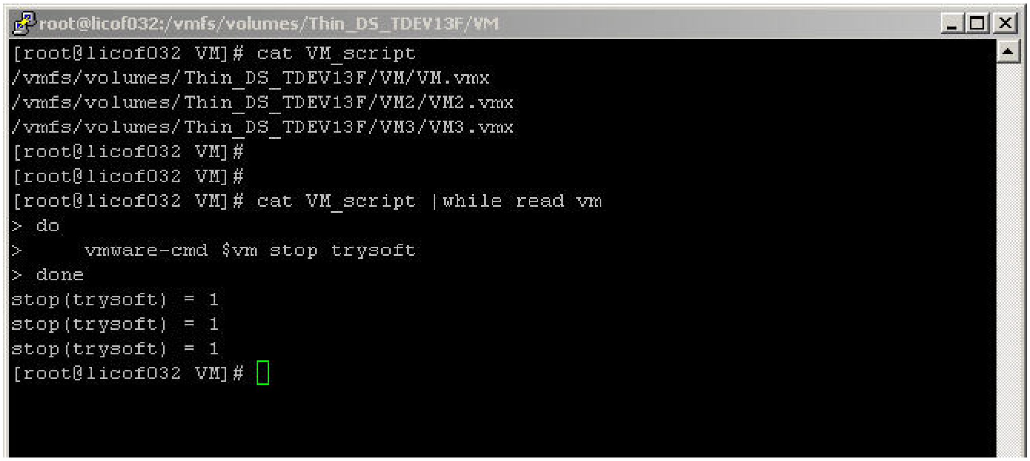
The output shows the pairing

Figure 111 Creating a TimeFinder/Clone relationship between standard devices

4. After the creation is run, the virtual machines can be brought down to make a “cold” copy of the data. The virtual machines can be either shut down using the vSphere Client, ESX VMware remote CLI, or scripts written using Perl or PowerShell. The command line utility, `vmware-cmd`, on ESX 4, may be the most appropriate tool if a number of virtual machines need to be powered down before the BCVs are split from the standard devices. On ESXi 5 the CLI `vim-cmd` can be used by first determining the VMs on the host, their current states, and then powering them off. The commands are:

```
vim-cmd vmsvc/getallvms
vim-cmd vmsvc/power.getstat VMID
vim-cmd vmsvc/power.off VMID
```

Figure 112 presents a sample script to shut down a group of virtual machines on an ESX 4 host using the command line utility `vmware-cmd`. In the script, the file VM_script contains a list of the virtual machines that are being cloned by EMC TimeFinder and need to be shut down.



```
[root@licof032 VM]# cat VM_script
/vmfs/volumes/Thin_DS_TDEV13F/VM/VM.vmx
/vmfs/volumes/Thin_DS_TDEV13F/VM2/VM2.vmx
/vmfs/volumes/Thin_DS_TDEV13F/VM3/VM3.vmx
[root@licof032 VM]#
[root@licof032 VM]#
[root@licof032 VM]# cat VM_script |while read vm
> do
>     vmware-cmd $vm stop trysoft
> done
stop(trysoft) = 1
stop(trysoft) = 1
stop(trysoft) = 1
[root@licof032 VM]#
```

Figure 112 Using command line utilities to shut down virtual machines

The status of the virtual machines can be checked by executing the command, `vmware-cmd <configuration_file> getstate`. When the virtual machine is shut down the command returns the string, **getstate() = off**. Once all of the virtual machines accessing the VMFS have been shut down, the pairs can be activated.

5. With the machines down, the TimeFinder/Clone pairs can be activated. The command,

```
symclone -g <group_name> -noprompt activate
```

activates the TimeFinder/Clone pairs. The state of the target devices is changed to "read-write." Any VMware ESX with access to the target devices is presented with a point-in-time view of the source device at the moment of activation. The Symmetrix storage array copies the data from the source device to the target device in the background.

6. The virtual machines accessing the VMFS on the source devices can be powered on and made available to the users.

The steps discussed above are shown in Figure 113.

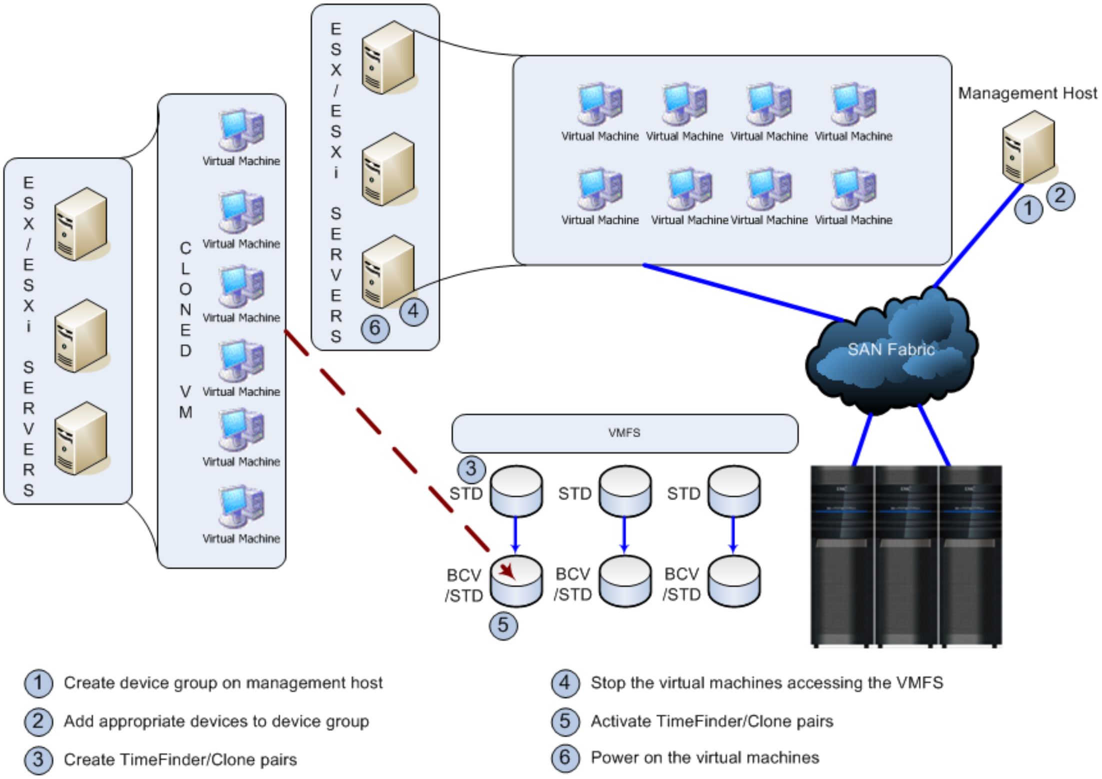

ESX/ESXi SERVERS
CLONED VM
Virtual Machine
ESX/ESXi SERVERS
Management Host
SAN Fabric
VMFS
STD
BCV /STD
1
2
3
4
5
6
1 Create device group on management host
2 Add appropriate devices to device group
3 Create TimeFinder/Clone pairs
4 Stop the virtual machines accessing the VMFS
5 Activate TimeFinder/Clone pairs
6 Power on the virtual machines

Figure 113 Copying virtual machine data using EMC TimeFinder/Clone

Copying cold virtual machines with RDMs using TimeFinder/Clone

The first step in copying virtual machines accessing disks as raw device mappings (RDM) is identifying the Symmetrix device numbers associated with the virtual machine. This can be most easily done by using EMC VSI.

Once the Symmetrix devices used by the virtual machines are identified, the process to create a copy of a virtual machine that is utilizing raw device mapping (RDM) is identical to the one presented in "Copying cold virtual machines on VMFS using

TimeFinder/Clone" on page 218.

Using TimeFinder/Snap with cold virtual machines

TimeFinder/Snap[1] enables users to create what appears to be a complete copy of their data while consuming only a fraction of the disk space required by the original copy. This is achieved by use of **virtual devices** as the target of the process. Virtual devices are constructs inside the Symmetrix storage array with minimal physical storage associated with them. Therefore, the virtual devices are normally presented in a not-ready state to any host accessing it.

When the TimeFinder/Snap relationship is created, the Symmetrix generates the metadata to associate the specified virtual devices with the source devices. The virtual devices continue to be presented in a not-ready (NR) to any host that is accessing the volumes. Therefore, the same considerations discussed in "Using TimeFinder/Clone with cold virtual machines" on page 217 when using TimeFinder/Clone technology in a VMware ESX environment apply. The virtual devices in a TimeFinder/Snap operation can be accessed by VMware ESX as soon as the snap relationship is activated. However, unlike TimeFinder/Clone technology, there is no copying of data after the session is activated. The virtual devices tracks are populated with the address of the source device tracks. When VMware ESX accesses the virtual devices, the data is fetched from the source device and presented to the requestor. The data changed by either the hosts accessing the source device or target device is stored in a temporary save area. The amount of data saved depends on the write activity on the source devices, target devices, and the duration for which the TimeFinder/Snap session remains activated.

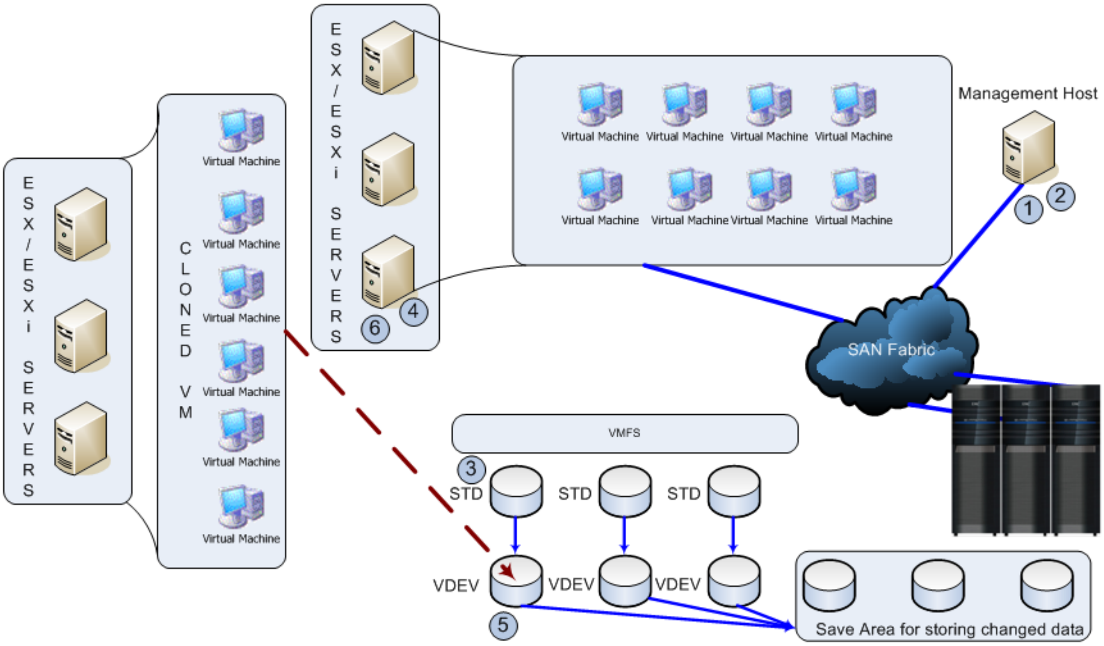
Copying cold virtual machines on VMFS using TimeFinder/Snap

TimeFinder/Snap devices are managed together using devices or composite groups. Solutions Enabler commands are executed to create SYMCLI groups for TimeFinder/Snap operations. If the VMFS

[1] TimeFinder/Snap is not supported on the Symmetrix VMAXe/VMAX 10K platform running Enginuity 5875 and higher. For space-saving capabilities on the VMAXe/VMAX 10K platform, TimeFinder VP Snap should be used. See the Simple Support Matrix on www.Powerlink.EMC.com for additional information.

spans more than one Symmetrix or the virtual machines are presented with storage from multiple VMFS volumes spanning multiple Symmetrix devices, a composite group should be used.

Figure 114 depicts the necessary steps to make a copy of powered off virtual machines using TimeFinder/Snap technology.



1 Create device group on management host

2 Add appropriate devices and associate BCVs to device group

3 Create TimeFinder/Snap pairs

4 Stop the virtual machines accessing the VMFS

5 Activate TimeFinder/Snap pairs

6 Power on the virtual machines

Figure 114 Copying inactive VMFS with TimeFinder/Snap

1. The device (or composite) groups containing the source devices and the target virtual devices should be created first. The commands, **symdg** and **symld**, shown on page 218 should be used for this step.

2. The TimeFinder/Snap pairs can be created using the command:

```
symsnap -g <group_name> create -noprompt
```

The command creates the protection bitmaps to use when the TimeFinder/Snap pairs are activated. No data is copied or moved at this time.

3. Once the create operation is completed, the virtual machines accessing the source VMFS must be shut down to make a cold copy of the virtual machines. The virtual machines can be either shut down using the vSphere Client or the remote VMware CLI.

4. With the virtual machines in a powered-off state, the TimeFinder/Snap copy can now be activated:

   ```
   symsnap -g <group_name> -noprompt activate
   ```

   The activate keyword presents a *virtual* copy of the VMFS to the target hosts for further processing. If using STD devices, the `-tgt` switch is needed.

5. The virtual machines using the source VMFS volumes can now be powered on. The same tools that were utilized to shut down the virtual machines can be used to power them on.

Copying cold virtual machines with RDMs using TimeFinder/Snap

The process for using TimeFinder/Snap technology to copy virtual machines that access disks as raw device mapping (RDM) is no different from that discussed for TimeFinder/Mirror or TimeFinder/Clone. The first step in copying virtual machines accessing disks as raw device mappings (RDM) is identifying the Symmetrix device numbers associated with the virtual machine. This can be most easily done by using EMC VSI.

Once the Symmetrix devices used by the virtual machines are identified, the process to create a copy of a virtual machine that is utilizing raw device mapping (RDM) is identical to the one presented in "Copying cold virtual machines on VMFS using TimeFinder/Snap" on page 225.

Using TimeFinder/VP Snap with cold virtual machines

TimeFinder/VP Snap, although using the TimeFinder/Clone technology, enables users to create space-efficient snaps with multiple sessions sharing allocations in the thin pool. This is achieved as VP Snap sessions copy data from the source device to the target device only if triggered by a host write operation. Read I/Os to protected tracks on the target device do not result in data being copied.

For a single activated VP Snap session on a source device, the target represents a single point-in-time copy of the source. Copied data resides on allocations in the thin pool. When there is a second VP Snap session from the same source device to a different target, the allocations can be shared.

When the VP Snap relationship between the source and target devices is created, the target devices are presented as not ready (NR) to any host that is accessing the volumes. It is for this reason that EMC recommends unmounting any VMFS datastores that are currently on those devices before creating the clone. The creation operation makes the VMFS on the target unavailable to the VMware ESX hosts. Therefore, the recommendation of unmounting the datastores before performing an establish operation applies directly to the virtual machines that is impacted by the absence of those datastores. The target devices in a TimeFinder/VP Snap operation can be accessed by VMware ESX as soon as the VP Snap relationship

Copying cold virtual machines on VMFS using TimeFinder/VP Snap

TimeFinder/VP Snap devices are managed as an atomic unit by creating a device or composite group. Composite groups are used if the virtual machines that are being cloned use VMFS volumes from multiple Symmetrix storage arrays. Unisphere or Solutions Enabler commands can be used to create the device or composite group[1], and manage the copying process.

The next few paragraphs present the steps required to clone a group of virtual machines utilizing EMC TimeFinder/VP Snap technology:

1. Identify the device number of the Symmetrix volumes used by the VMFS by utilizing EMC Virtual Storage Integrator for VMware vSphere (EMC VSI) which maps physical extents of a VMFS to the Symmetrix device number. A device group containing the member(s) of the VMFS should be created. This can be done with the following commands:

   ```
   symdg create <group_name>
   symld -g <group_name> -sid <Symmetrix SN> add dev <dev
   #>
   ```

   where **<dev #>** are the devices identified by EMC VSI.

[1] Composite groups are not yet supported in Unisphere

2. The devices that hold the copy of the data need to be added or associated with the group. Note that TimeFinder/VP Snap can be used when devices with the BCV attributes are used as a target. If the target devices have the BCV attribute turned on, the devices should be associated with the device group using the command

   ```
   symbcv -g <group_name> associate dev <dev #>
   ```

   When STD devices are used as the target, the devices need to be added to the devices using the **symld** command along with the `-tgt` switch

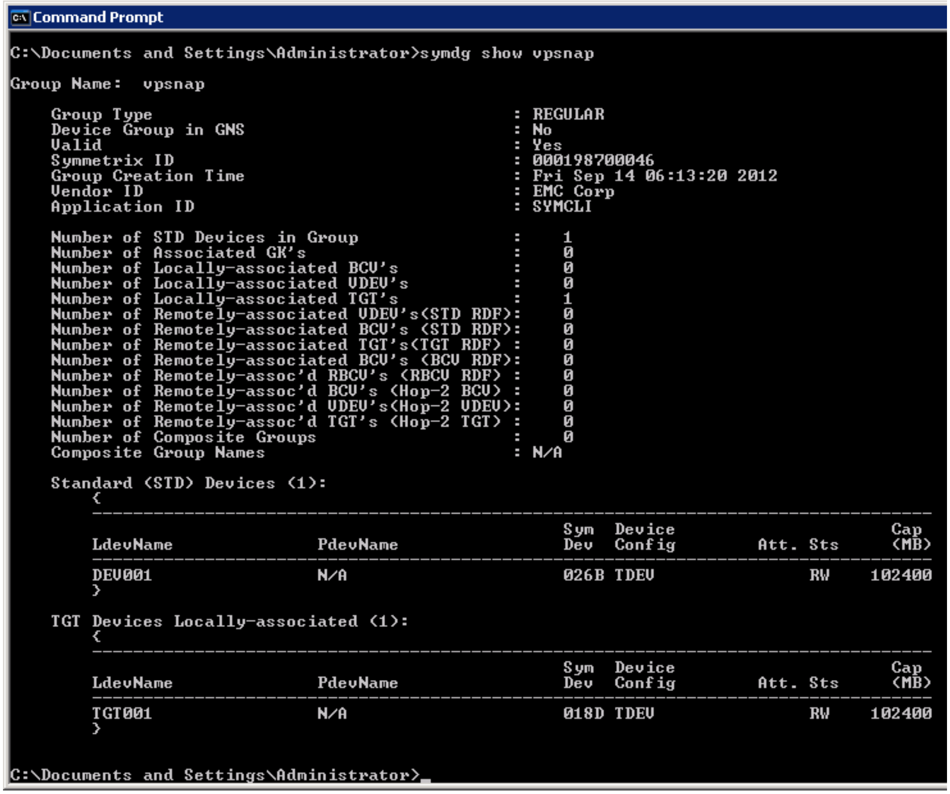
   ```
   symld -g <group_name> -sid <Symmetrix SN> -tgt add dev
   <dev #>
   ```

3. The next step in the process of copying virtual machines using EMC TimeFinder/VP Snap is the creation of TimeFinder/VP Snap pairs. The command will essentially be the same as that for TimeFinder/Clone except the VP Snap switch is used `-vse.` The command to perform the `create` depends on the attributes of the target devices. If the target devices have the BCV attribute, the command

   ```
   symclone -g <group_name> -vse -noprompt create
   ```

   automatically creates a TimeFinder/VP Snap relationship between the source and target devices. If STD devices are used, the `-tgt` switch is required to the above command.

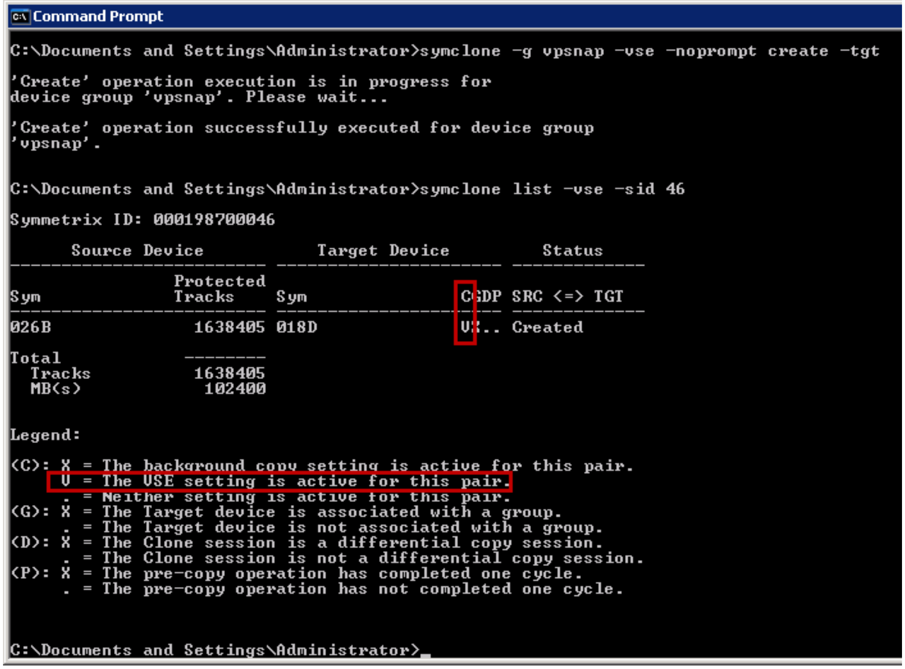
Figure 115 shows a devices group that consists of one standard devices and one target device.



```
C:\Documents and Settings\Administrator>symdg show vpsnap

Group Name:  vpsnap

    Group Type                                   : REGULAR
    Device Group in GNS                          : No
    Valid                                        : Yes
    Symmetrix ID                                 : 000198700046
    Group Creation Time                          : Fri Sep 14 06:13:20 2012
    Vendor ID                                    : EMC Corp
    Application ID                               : SYMCLI

    Number of STD Devices in Group               :    1
    Number of Associated GK's                    :    0
    Number of Locally-associated BCV's           :    0
    Number of Locally-associated VDEV's          :    0
    Number of Locally-associated TGT's           :    1
    Number of Remotely-associated VDEV's(STD RDF):    0
    Number of Remotely-associated BCV's (STD RDF):    0
    Number of Remotely-associated TGT's(TGT RDF) :    0
    Number of Remotely-associated BCV's (BCV RDF):    0
    Number of Remotely-assoc'd RBCV's (RBCV RDF) :    0
    Number of Remotely-assoc'd BCV's (Hop-2 BCV) :    0
    Number of Remotely-assoc'd VDEV's(Hop-2 VDEV):    0
    Number of Remotely-assoc'd TGT's (Hop-2 TGT) :    0
    Number of Composite Groups                   :    0
    Composite Group Names                        : N/A

    Standard (STD) Devices (1):
        {
        ----------------------------------------------------------------------------------
                                                      Sym  Device                    Cap
        LdevName              PdevName                Dev  Config        Att. Sts    (MB)
        ----------------------------------------------------------------------------------
        DEV001                N/A                     026B TDEV               RW   102400
        }

    TGT Devices Locally-associated (1):
        {
        ----------------------------------------------------------------------------------
                                                      Sym  Device                    Cap
        LdevName              PdevName                Dev  Config        Att. Sts    (MB)
        ----------------------------------------------------------------------------------
        TGT001                N/A                     018D TDEV               RW   102400
        }


C:\Documents and Settings\Administrator>_
```

Figure 115 Displaying the contents of a device group

The Symmetrix device 26B contains the VMFS that needs to be copied. In Figure 116 the `create` is shown along with a query of the newly created relationship with the VP Snap identifier highlighted in the red boxes. Note the target device is changed to a not ready (NR) state.



Command Prompt

```
C:\Documents and Settings\Administrator>symclone -g vpsnap -vse -noprompt create -tgt

'Create' operation execution is in progress for
device group 'vpsnap'. Please wait...

'Create' operation successfully executed for device group
'vpsnap'.


C:\Documents and Settings\Administrator>symclone list -vse -sid 46

Symmetrix ID: 000198700046

      Source Device            Target Device          Status
------------------------- ---------------------- ------------
                Protected
Sym              Tracks    Sym                  CGDP SRC <=> TGT
------------------------- ---------------------- ------------
026B              1638405 018D                  VX.. Created

Total             ---------
  Tracks          1638405
  MB(s)            102400


Legend:

(C): X = The background copy setting is active for this pair.
     V = The VSE setting is active for this pair.
     . = Neither setting is active for this pair.
(G): X = The Target device is associated with a group.
     . = The Target device is not associated with a group.
(D): X = The Clone session is a differential copy session.
     . = The Clone session is not a differential copy session.
(P): X = The pre-copy operation has completed one cycle.
     . = The pre-copy operation has not completed one cycle.


C:\Documents and Settings\Administrator>_
```

Figure 116 Creating a TimeFinder/VP Snap relationship between standard devices

4. The virtual machines utilizing the VMFS can be shut down as soon as the TimeFinder/VP Snap create process completes successfully. This can be performed using different graphical interface or command line tools.

5. The TimeFinder/Clone pairs can be activated once the virtual machines have been shut down. The command,

   ```
   symclone -g <group_name> -noprompt activate
   ```

   activates the TimeFinder/VP Snap pair. The state of the target devices is changed to “read-write”. Once again, the `-tgt` switch is needed for STD devices. Any VMware ESX with access to the target devices is presented with a point-in-time view of the source device at the moment of activation. The Symmetrix storage array will only use the disk space in the thin pool if writes are performed on the VP Snap source or target devices. Reads do not trigger new writes. As mentioned, if more VP Snap sessions are added to the same source device, data is copied to the targets based on whether the source data is new with respect to the point-in-time of each copy. When data is copied to more than one target, only a single shared copy resides in the thin pool.

6. The virtual machines accessing the VMFS on the source devices can be powered on and made available to the users.

Figure 117 depicts the necessary steps to make a copy of powered off virtual machines using TimeFinder/Snap technology.

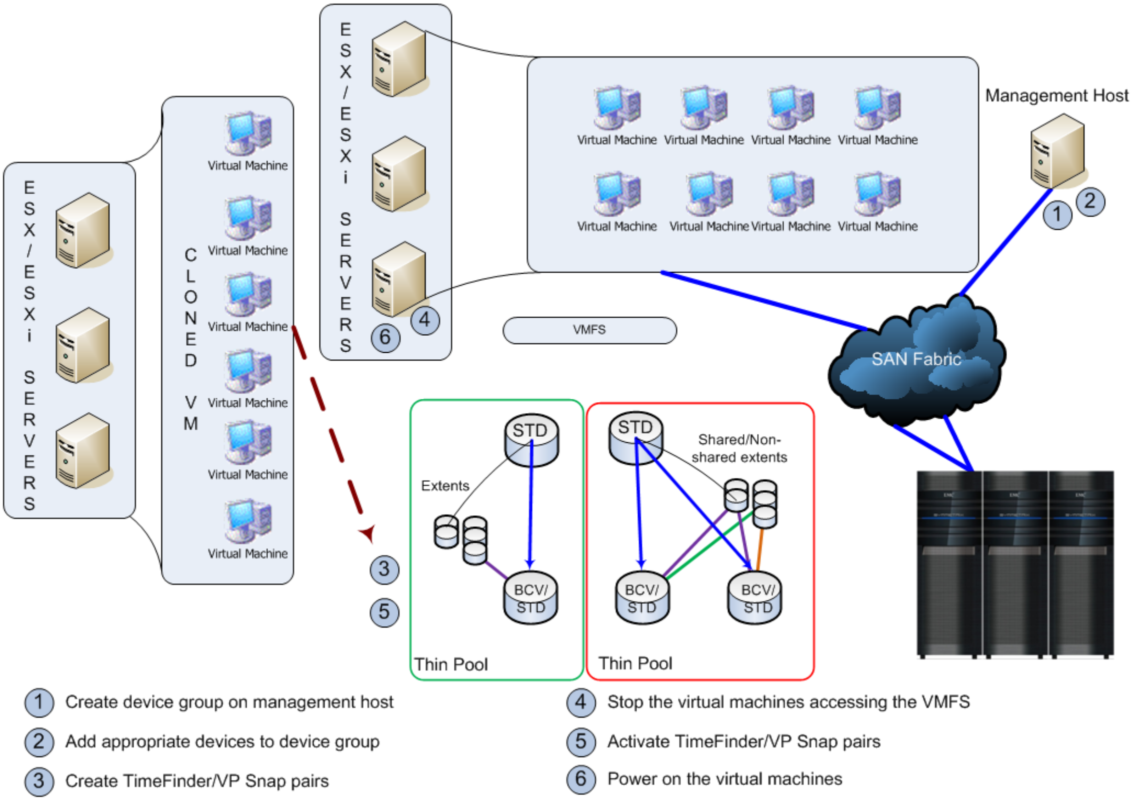

ESX/ESXi SERVERS
CLONED VM
Virtual Machine
ESX/ESXi SERVERS
Management Host
VMFS
SAN Fabric
Extents
STD
BCV/ STD
Thin Pool
Shared/Non-shared extents
1
2
3
4
5
6
1 Create device group on management host
2 Add appropriate devices to device group
3 Create TimeFinder/VP Snap pairs
4 Stop the virtual machines accessing the VMFS
5 Activate TimeFinder/VP Snap pairs
6 Power on the virtual machines

Figure 117 Copying virtual machine data using EMC TimeFinder/VP Snap

Copying cold virtual machines with RDMs using TimeFinder/VP Snap

The first step in copying virtual machines accessing disks as raw device mappings (RDM) is identifying the Symmetrix device numbers associated with the virtual machine. This can be most easily done by using EMC VSI.

Once the Symmetrix devices used by the virtual machines are identified, the process to create a copy of a virtual machine that is utilizing raw device mapping (RDM) is identical to the one presented in "Copying cold virtual machines on VMFS using TimeFinder/VP Snap" on page 227.

Copying running virtual machines using EMC Consistency technology

"Copying virtual machines after shutdown" on page 217 discussed use of the TimeFinder family of products to clone virtual machines that have been shut down. Although this is the ideal way to obtain a copy of the data, it is impractical in most production environments. For these environments, EMC Consistency technology can be leveraged to create a copy of the virtual machines while it is servicing applications and users. TimeFinder/CG enables a group of live virtual machines to be copied in an instant when used in conjunction with storage array based copying software such as TimeFinder/Clone. The image created in this way is a dependent-write consistent data state and can be utilized as a restartable copy of the virtual machine.

Virtual machines running modern operating systems such as Microsoft Windows and database management systems enforce the principle of dependent-write I/O. That is, no dependent write is issued until the predecessor write it is dependent upon has completed. For example, Microsoft Windows does not update the contents of a file on a NT file system (NTFS) until an appropriate entry in the file system journal is made. This technique enables the operating system to quickly bring NTFS to a consistent state when recovering from an unplanned outage such as a power failure.

Note: Microsoft Window NT file system is a journaled file system and not a logged file system. When recovering from an unplanned outage the contents of the journal may not be sufficient to recover the file system. A full check of the file system using **chkdsk** is needed for these situations.

Using the EMC Consistency technology option during the virtual machine copying process also creates a copy with a dependent-write consistent data state.

The process of creating a consistent copy of a running virtual machine does not differ from the process of creating a copy of a cold virtual machine save for one switch in the `activate` command. Simply follow the same steps as outlined in this chapter but do not shutdown the virtual machines. At the point of activation, add the switch `-consistent` to the command as follows:

TimeFinder/Clone

For BCV devices: `symclone -g <group_name> -noprompt -consistent activate`

or

For STD devices: `symclone -g <group_name> -noprompt -tgt -consistent activate`

TimeFinder/Snap

For BCV devices: `symsnap -g <group_name> -noprompt -consistent activate`

or

For STD devices: `symsnap -g <group_name> -noprompt -tgt -consistent activate`

TimeFinder/VP Snap

For BCV devices: `symclone -g <group_name> -noprompt -consistent activate`

or

For STD devices: `symclone -g <group_name> -noprompt -tgt -consistent activate`

Transitioning disk copies to cloned virtual machines

This section discusses how to use the copy of the data to create cloned virtual machines. The cloned virtual machines can be deployed to support ancillary business processes such as development and testing. The methodology deployed in creating the copy of the data influences the type of supporting business operations that it can support.

Cloning virtual machines on VMFS in vSphere environments

VMware ESX 4/ESXi 5 assigns a unique signature to all VMFS volumes when they are formatted with VMFS. Furthermore, the VMFS label is also stored on the device.

Since storage array technologies create exact replicas of the source volumes, all information including the unique signature and label is replicated. If a copy of a VMFS volume is presented to any VMware ESX host or cluster group, the VMware ESX, by default, automatically masks the copy. The device holding the copy is determined by comparing the signature stored on the device with the computed signature. BCVs, for example, have a different unique ID from the standard device it is associated with it. Therefore, the computed signature for a BCV device always differs from the one stored on it. This enables VMware ESX to always identify the copy correctly.

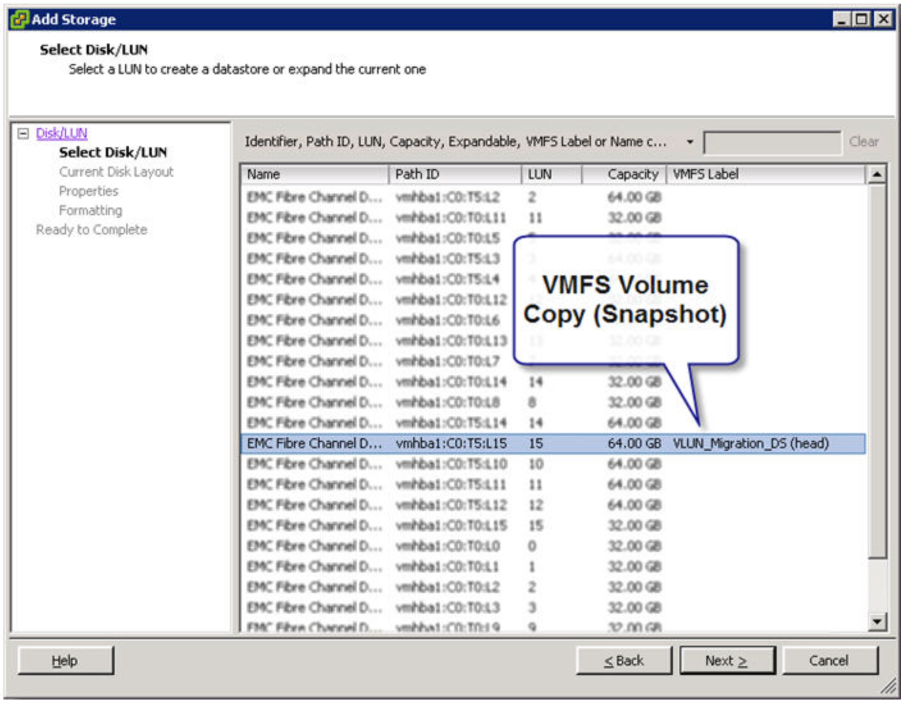
VMware vSphere has the ability to individually resignature and/or mount VMFS volume copies through the use of the vSphere Client or with the CLI utility `esxcfg-volume` (`vicfg-volume` for ESXi). Since it is volume-specific, it allows for much greater control in the handling of snapshots. This feature is very useful when creating and managing volume copies created by local replication products such as TimeFinder/Snap or TimeFinder/Clone or remote replication technologies such as Symmetrix Remote Data Facility (SRDF).

While either the CLI or the GUI can be used to perform these operations, the GUI (vSphere Client) is always recommended. Nevertheless, there are a few features available in the CLI that cannot be performed in vSphere Client. The use and implications of both options will be discussed in the following sections.

Cloning vSphere virtual machines using the vSphere Client

In the vSphere Client, the Add Storage wizard is the function to resignature VMFS volume copies or mount cloned ones. The Add Storage wizard now displays the VMFS label next to storage devices that contain a VMFS volume copy. Therefore, if a device is not mounted, and shows up in the wizard and has a label associated with it, you can make the assumption that it is indeed a snapshot.

Once a replicated VMFS datastore is presented to a host, it will appear in the Add Storage wizard with the name of the original datastore in the VMFS label column shown in Figure 118.



Select Disk/LUN

Select a LUN to create a datastore or expand the current one

Identifier, Path ID, LUN, Capacity, Expandable, VMFS Label or Name c...

Name	Path ID	LUN	Capacity	VMFS Label
EMC Fibre Channel D...	vmhba1:C0:T5:L2	2	64.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L11	11	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L5	5	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T5:L3	3	64.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T5:L4	4		
EMC Fibre Channel D...	vmhba1:C0:T0:L12	12	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L6			
EMC Fibre Channel D...	vmhba1:C0:T0:L13	13	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L7	7	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L14	14	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L8	8	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T5:L14	14	64.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T5:L15	15	64.00 GB	VLUN_Migration_DS (head)
EMC Fibre Channel D...	vmhba1:C0:T5:L10	10	64.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T5:L11	11	64.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T5:L12	12	64.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L15	15	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L0	0	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L1	1	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L2	2	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L3	3	32.00 GB	
EMC Fibre Channel D...	vmhba1:C0:T0:L9	9	32.00 GB	

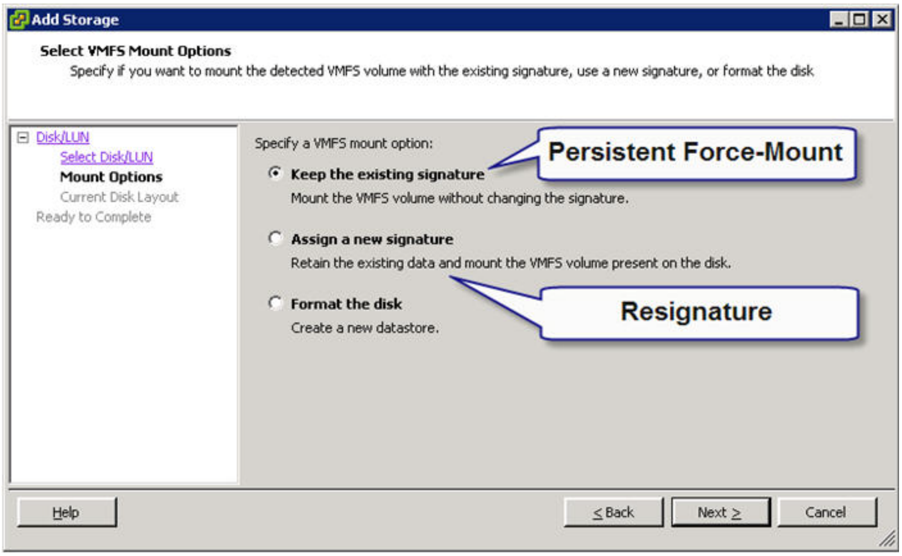
Figure 118 **Mounting a VMFS volume copy using the Add Storage wizard in vSphere Client**

Once the device with the replicated VMFS volume is selected the wizard offers three choices of how to handle presenting it to the ESX as a datastore which is displayed in Figure 119 on page 237.

- **Keep the existing signature** — This option initiates a persistent force-mount. This will cause the VMFS volume copy to be persistently mounted on the ESX host server (in other words it will remain an accessible datastore over reboots of the ESX host).[1] If the original VMFS volume is already presented to the same ESX host the mounting of the volume copy will fail. The ability to non-persistently mount a VMFS volume copy is a feature only available through the CLI. In addition, CLI utility has to be used to force-mount a VMFS datastore copy while retaining the existing signature on multiple ESX servers.

Note: If the device holding the copy of VMware datastore is going to be re-provisioned for other activities, it is important to unmount the persistently mounted VMware datastore before the device is reused. A snapshot datastore can be unmounted by using the vSphere Client.

- **Assign a new signature** — This option will resignature the VMFS volume copy entirely and will therefore allow it to be presented to the same host as the original VMFS volume.
- **Format the disk** — This option wipes out any data on the device and creates an entirely new VMFS volume with a new signature.



Add Storage

Select VMFS Mount Options

Specify if you want to mount the detected VMFS volume with the existing signature, use a new signature, or format the disk

Specify a VMFS mount option:

● **Keep the existing signature**
Mount the VMFS volume without changing the signature.

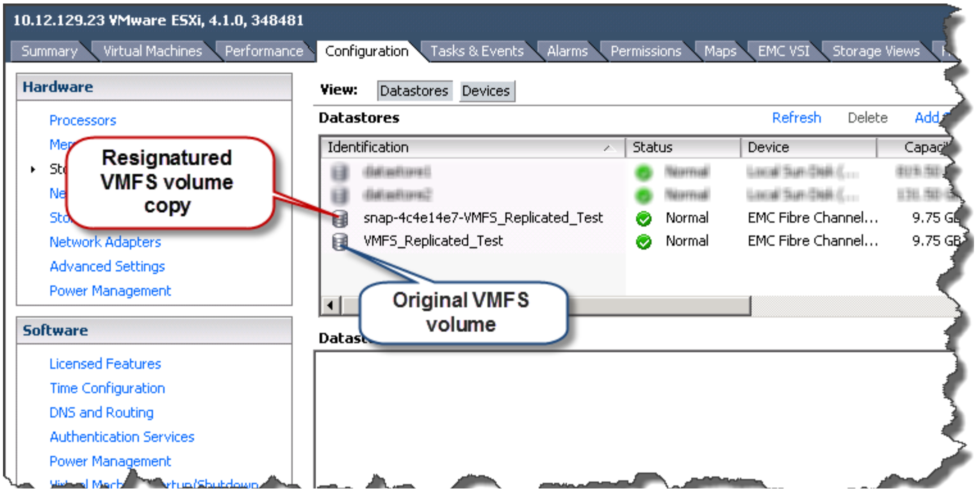
○ **Assign a new signature**
Retain the existing data and mount the VMFS volume present on the disk.

○ **Format the disk**
Create a new datastore.

Figure 119 VMFS volume copy mounting options in vSphere Client

[1] This is true so long as the original device is not presented to the ESX host, in which case the mount of the VMFS volume copy will fail.

If the volume has been resignatured, it will appear in the Datastores list, as in Figure 120, with “snap-XXXXXXXX-” appended to the beginning of the original VMFS volume name. Any virtual machines residing on resignatured datastore will have to be manually registered with an ESX to be recognized.



Identification	Status	Device	Capac
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
snap-4c4e14e7-VMFS_Replicated_Test	Normal	EMC Fibre Channel...	9.75 GE
VMFS_Replicated_Test	Normal	EMC Fibre Channel...	9.75 GB

Resignatured VMFS volume copy

Original VMFS volume

Figure 120 Original VMFS volume and resignatured VMFS volume copy

Using the command line to mount VMFS volume copies

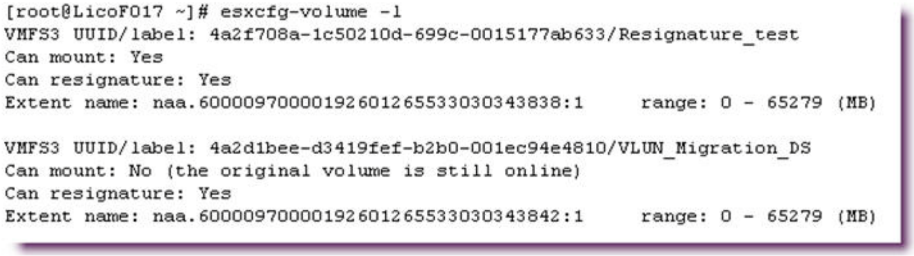
In the ESX service console CLI, the `esxcfg-volume` (or vicfg-volume with the VMware remote CLI tools) command is the function used to resignature VMFS volume copies or mount cloned ones. A feature that is unique to `esxcfg-volume` and not available in vSphere Client is the ability to non-persistently force-mount a volume copy (meaning it will not be mounted after a reboot of the ESX host).

`esxcfg-volume` comes with the following parameter options:

```
-l | --list

-m | --mount <vmfs uuid|label>

-u | --umount <vmfs uuid|label>

-r | --resignature <vmfs uuid|label>

-M | --persistent-mount <vmfs uuid|label
```

The command `esxcfg-volume -l` will list all of the volumes that have been detected as snapshots or replicas. If a volume copy is found, it displays the UUID, whether or not it can be mounted (decided by the presence of the original VMFS volume), whether or not it can be resignatured, and its extent name.

Figure 121 shows the list command returns two VMFS volume copies. The first volume copy can be mounted due to the fact that the original VMFS volume is not present on that ESX host. The second discovered volume copy cannot be mounted because the original volume is still online.

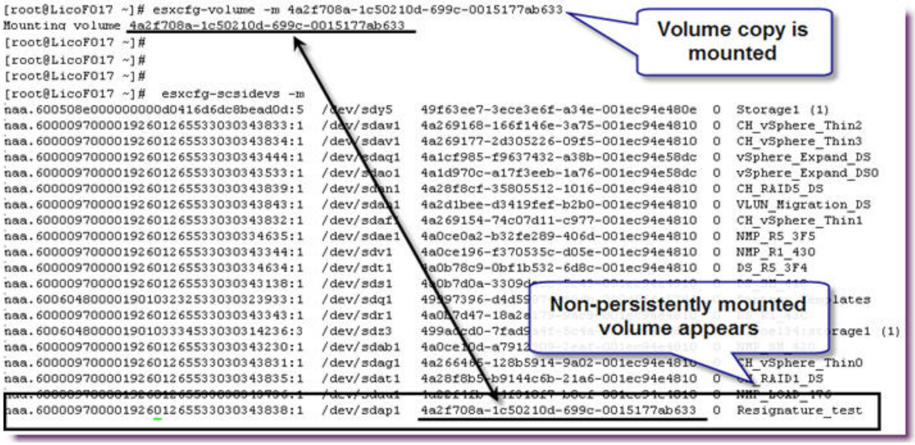


```
[root@LicoF017 ~]# esxcfg-volume -l
VMFS3 UUID/label: 4a2f708a-1c50210d-699c-0015177ab633/Resignature_test
Can mount: Yes
Can resignature: Yes
Extent name: naa.60000970000192601265533030343838:1     range: 0 - 65279 (MB)

VMFS3 UUID/label: 4a2d1bee-d3419fef-b2b0-001ec94e4810/VLUN_Migration_DS
Can mount: No (the original volume is still online)
Can resignature: Yes
Extent name: naa.60000970000192601265533030343842:1     range: 0 - 65279 (MB)
```

Figure 121 Listing available VMFS volume copies using the esxcfg-volume CLI command

In this example the first discovered volume copy “4a2f708a-1c50210d-699c-0015177ab633/Resignature_test” will be non-persistently force-mounted using the command `esxcfg-volume -m`. Since it does not have its original VMFS volume present it can be mounted without the need to resignature.

In Figure 122 on page 240, this volume is non-persistently mounted through the CLI and then `esxcfg-scsidevs -m` is executed to list the VMFS volumes and the newly mounted volume is listed. Since it is non-persistent, it will disappear after the next reboot of the ESX host.



```
[root@LicoF017 ~]# esxcfg-volume -m 4a2f708a-1c50210d-699c-0015177ab633
Mounting volume 4a2f708a-1c50210d-699c-0015177ab633
[root@LicoF017 ~]#
[root@LicoF017 ~]#
[root@LicoF017 ~]#
[root@LicoF017 ~]#  esxcfg-scsidevs -m
naa.600508e000000000d0416d6dc8bead0d:5   /dev/sdy5    49f63ee7-3ece3e6f-a34e-001ec94e480e  0  Storage1 (1)
naa.60000970000192601265533030343833:1   /dev/sdaw1   4a269168-166f146e-3a75-001ec94e4810  0  CH_vSphere_Thin2
naa.60000970000192601265533030343834:1   /dev/sdav1   4a269177-2d305226-09f5-001ec94e4810  0  CH_vSphere_Thin3
naa.60000970000192601265533030343444:1   /dev/sdaq1   4a1cf985-f9637432-a38b-001ec94e58dc  0  vSphere_Expand_DS
naa.60000970000192601265533030343533:1   /dev/sdao1   4a1d970c-a17f3eeb-1a76-001ec94e58dc  0  vSphere_Expand_DS0
naa.60000970000192601265533030343839:1   /dev/sdan1   4a28f8cf-35805512-1016-001ec94e4810  0  CH_RAID5_DS
naa.60000970000192601265533030343843:1   /dev/sdab1   4a2d1bee-d3419fef-b2b0-001ec94e4810  0  VLUN_Migration_DS
naa.60000970000192601265533030343832:1   /dev/sdaf1   4a269154-74c07d11-c977-001ec94e4810  0  CH_vSphere_Thin1
naa.60000970000192601265533030334635:1   /dev/sdae1   4a0ce0a2-b32fe289-406d-001ec94e4810  0  NMP_R5_3F5
naa.60000970000192601265533030343344:1   /dev/sdv1    4a0ce196-f370535c-d05e-001ec94e4810  0  NMP_R1_430
naa.60000970000192601265533030334634:1   /dev/sdt1    4a0b78c9-0bf1b532-6d8c-001ec94e4810  0  DS_R5_3F4
naa.60000970000192601265533030343138:1   /dev/sds1    4a0b7d0a-3309d[illegible]
naa.60060480000190103232533030323933:1   /dev/sdq1    49997396-d4d59[illegible]lates
naa.60000970000192601265533030343343:1   /dev/sdr1    4a0b7d47-18a2e[illegible]
naa.60060480000190103334533030314236:3   /dev/sdz3    499a0cd0-7fad9[illegible]rage1 (1)
naa.60000970000192601265533030343230:1   /dev/sdab1   4a0ce10d-a7912[illegible]
naa.60000970000192601265533030343831:1   /dev/sdag1   4a266465-128b5914-9a02-001ec94e4810  0  CH_vSphere_Thin0
naa.60000970000192601265533030343835:1   /dev/sdat1   4a28f8b5-b9144c6b-21a6-001ec94e4810  0  CH_RAID1_DS
[illegible]
naa.60000970000192601265533030343838:1   /dev/sdap1   4a2f708a-1c50210d-699c-0015177ab633  0  Resignature_test
```

Figure 122 Non-persistently force-mounting a VMFS volume copy using the esxcfg-volume CLI command

Otherwise, if the cloned volume is never going to be presented to this host again it can be mounted persistently, which is the equivalent of using the vSphere Client Add Storage wizard to do it, by using the `esxcfg-volume -M` command.

Cloning virtual machines using RDM in VMware vSphere environments

The configuration file located in the VMFS volumes can be used to clone virtual machines provided with storage using RDM. However, in a vSphere environment it is easier to use copies of the configuration files on the target VMware ESX.

When a RDM is generated, a file is created on a VMFS that points to the physical device that is mapped. The file that provides the mapping also includes the unique ID and LUN number of the device it is mapping. The configuration file for the virtual machine using the RDM contains an entry that includes the label of the VMFS holding the RDM and its name. If the VMFS holding the information for the virtual machines is replicated and presented on the target VMware ESX, the virtual disks that provide the mapping are also available in addition to the configuration files. However, the mapping file cannot be used on the target VMware ESX since the cloned virtual machines need to be provided with access to the devices holding the copy of the data. Therefore, EMC recommends using a copy of the source virtual machine's configuration file instead of replicating the VMFS. The following steps clone virtual machines using RDMs in a vSphere environment:

1. On the target VMware ESX create a directory on a datastore (VMFS or NAS storage) that holds the files related to the cloned virtual machine. A VMFS on internal disk, un-replicated SAN-attached disk or NAS-attached storage should be used for storing the files for the cloned virtual disk. This step has to be performed once.
2. Copy the configuration file for the source virtual machine to the directory created in step 1. The command line utility **scp** can be used for this purpose. This step has to be repeated only if the configuration of the source virtual machine changes.
3. Register the cloned virtual machine using the vCenter client or the VMware remote CLI. This step does not need to be repeated.
4. Generate RDMs on the target VMware ESX in the directory created in step 1. The RDMs should be configured to address the target devices.
5. The cloned virtual machine can be powered on using either the VMware vSphere Client or the VMware remote CLI.

Note: The process listed in this section assumes the source virtual machine does not have storage presented as a mix of a virtual disk on a VMFS and RDM. The process to clone virtual machines with a mix of RDMs and virtual disks is complex and beyond the scope of this document. Readers are requested to contact the authors at cody.hosterman@emc.com or drew.tonnesen@emc.com if such requirements arise.

Cloning virtual machines using vStorage APIs for Array Integration (VAAI)

With a minimum vSphere 4.1 release and a Symmetrix VMAX/VMAXe with a minimum of Enginuity 5875, users can take advantage of vStorage APIs that offload certain functions in vSphere to the array, such as cloning. When a customer desires to clone a single or a few virtual machines, or move a virtual machine from one datastore to another, using TimeFinder adds unnecessary complexity. Using VAAI in these cases, in essence, does what TimeFinder does, just on a smaller scale. The following section will cover VAAI in more detail.

VMware vSphere vStorage API for Array Integration

Storage APIs for Array Integration is an API for storage partners to leverage that permits certain functions to be delegated to the storage array, thus greatly enhancing the performance of those functions. This API is fully supported by EMC Symmetrix VMAX running Enginuity 5875 or later, including the most current release Enginuity 5876.[1] Beginning with the vSphere 4.1 release, this array offload capability supports three primitives: hardware-accelerated Full Copy, hardware-accelerated Block Zero, and hardware-assisted locking. Beginning with vSphere 5.0 U1 and Enginuity version 5876.159.102 the UNMAP primitive is supported.

The primitive that will be covered herein is hardware-accelerated **Full Copy** as it is this functionality that enables the offloading of VMware clones to the array.[2]

Note: The VAAI primitives are supported with Federated Tiered Storage (FTS) both in external provisioning mode and encapsulation mode with a single restriction - the Thin Provisioning Block Space Reclamation (UNMAP) primitive is not supported with encapsulation mode.

Hardware-accelerated Full Copy

The time it takes to deploy or migrate a virtual machine will be greatly reduced by use of the **Full Copy** primitive, as the process for data migration is entirely executed on the storage array and not on the ESX server. The host simply initiates the process and reports on the progress of the operation on the array. This decreases overall traffic on the ESX server. In addition to deploying new virtual machines from a template or through cloning, **Full Copy** is also

[1] Customers using or intending to use VAAI with Enginuity 5875 should refer to the following EMC Technical Advisory for additional patching information: ETA emc263675: Symmetrix VMAX: VMware vStorage API for Array Integration (VAAI). This ETA will provide the properly patched version of Enginuity 5875 to ensure the best performance of the VAAI primitives and to block the UNMAP primitive.

[2] For more information on other Storage APIs please refer to the EMC whitepaper Using VMware vSphere Storage APIs for Array Integration with EMC Symmetrix found on EMC.com: http://www.emc.com/collateral/hardware/white-papers/h8115-vmware-vstorage-vmax-wp.pdf

utilized when doing a Storage vMotion. When a virtual machine is migrated between datastores on the same array the live copy is performed entirely on the array.

Not only does Full Copy save time, but it also saves server CPU cycles, memory, IP and SAN network bandwidth, and storage front-end controller I/O. This is due to the fact that the host is relieved of the normal function it would serve of having to read the data from the array and then write the data back down to the array. That activity requires CPU cycles and memory as well as significant network bandwidth. Figure 123 provides a graphical representation of Full Copy.

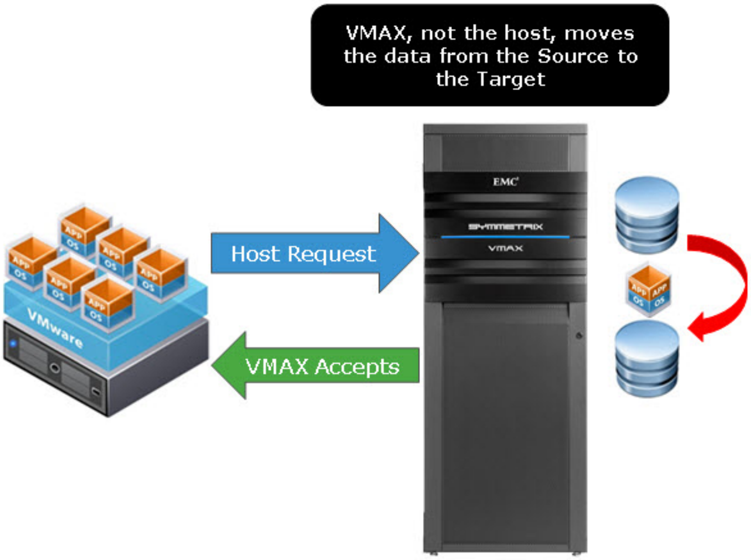
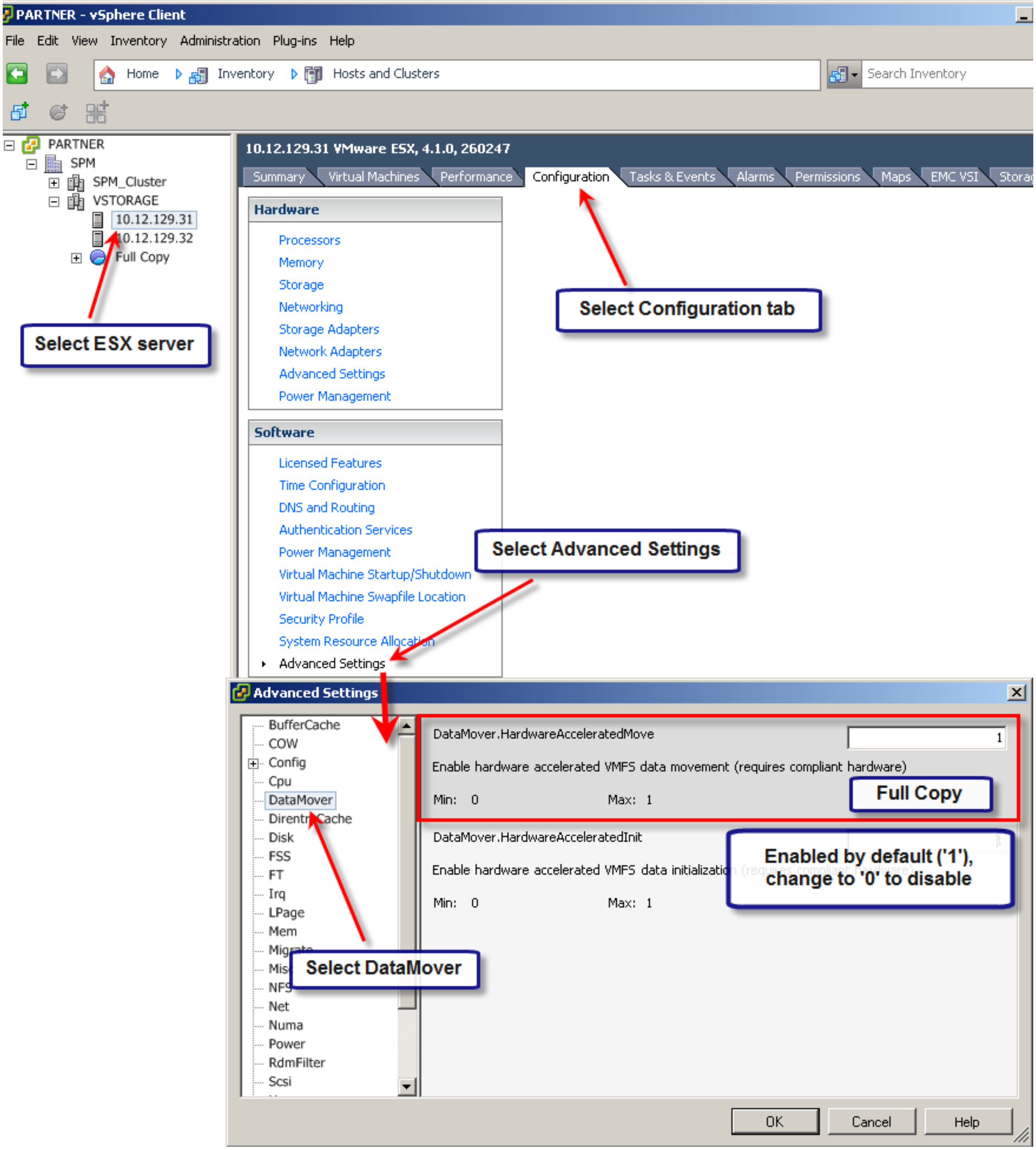

VMAX, not the host, moves the data from the Source to the Target
Host Request
VMAX Accepts
VMware
EMC
SYMMETRIX
VMAX
APP
OS

Figure 123 **Hardware-accelerated Fully Copy**

Enabling the vStorage APIs for Array Integration

The VAAIs are enabled by default on both the Symmetrix VMAX/VMAXe running 5875 Enginuity or later, and on the 4.1 ESX and ESXi 5 servers (properly licensed) and should not require any user intervention.[1] The primitive, however, can be disabled through the ESX server if desired. Using the vSphere Client, Full Copy can be disabled or enabled by altering the setting, *DataMover.HardwareAcceleratedMove,* in the ESX server advanced settings under DataMover as shown in Figure 124 on page 245.

[1] Refer to VMware documentation for required licensing to enable VAAI features. The UNMAP primitive requires a minimum of vSphere 5.0 U1 and Enginuity 5876.159.102.



DataMover.HardwareAcceleratedMove 1

Enable hardware accelerated VMFS data movement (requires compliant hardware)

Min: 0 Max: 1

DataMover.HardwareAcceleratedInit

Enable hardware accelerated VMFS data initializatio

Min: 0 Max: 1

Select ESX server

Select Configuration tab

Select Advanced Settings

Full Copy

Enabled by default ('1'), change to '0' to disable

Select DataMover

Figure 124 Enabling/disabling hardware-accelerated Full Copy in ESX

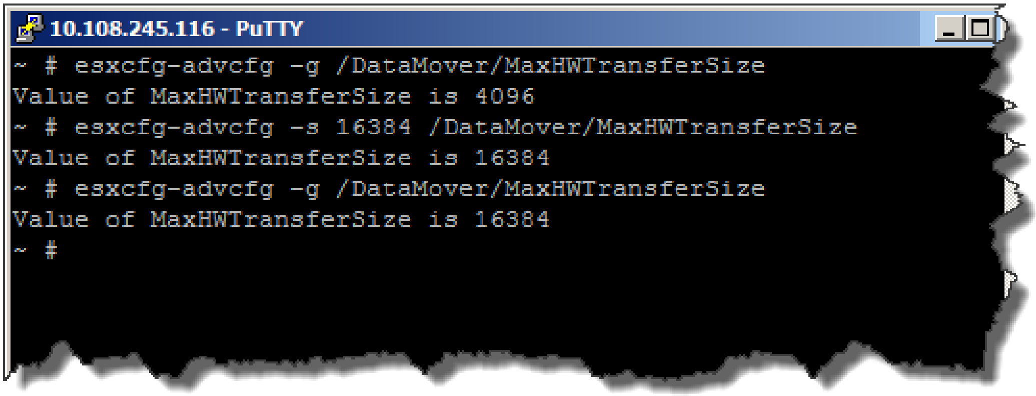
Full Copy Tuning

By default, when the Full Copy primitive is supported, VMware informs the storage array to copy data in 4 MB chunks. For vSphere environments that utilize many different arrays, both Symmetrix and non-Symmetrix platforms, the default value should not be altered. In all-Symmetrix environments, however, customers are advised to take advantage of VMware's ability to send larger chunks to the array for copy. The default copy size can be incremented to a maximum value of 16 MB and should be set at that value for Symmetrix-only environments to take advantage of the array's ability to move larger chunks of storage at once.

The following commands show how to query the current copy (or transfer) size and then how to alter it.

```
# esxcfg-advcfg -g /DataMover/MaxHWTransferSize

Value of MaxHWTransferSize is 4096

# esxcfg-advcfg -s 16384 /DataMover/MaxHWTransferSize

Value of MaxHWTransferSize is 16384
```

Note that this change is per ESX(i) server and can only be done through the CLI (no GUI interface) as in Figure 125:



```
~ # esxcfg-advcfg -g /DataMover/MaxHWTransferSize
Value of MaxHWTransferSize is 4096
~ # esxcfg-advcfg -s 16384 /DataMover/MaxHWTransferSize
Value of MaxHWTransferSize is 16384
~ # esxcfg-advcfg -g /DataMover/MaxHWTransferSize
Value of MaxHWTransferSize is 16384
~ #
```

Figure 125 Changing the default copy size for Full Copy

Note: The value of this parameter does not guarantee VMware will only send that size to the array, rather it simply means it is the largest copy size it may request; however it is likely that the majority of the copy sizes will be at the maximum value for any task that can utilize Full Copy.

Full Copy use cases

This section provides Full Copy use cases.

Hardware-accelerated Full Copy

The primary use cases that show the benefit of hardware-accelerated Full Copy are related to deploying virtual machines - whether from a template or executing a hot or cold clone. With Full Copy the creation of any of these virtual machines will be offloaded to the array. In addition to deploying new virtual machines, Full Copy is also utilized by Storage vMotion operations, significantly reducing the time to move a virtual machine to a new datastore.

Following are a number of use cases for Full Copy concerned with the deployment of virtual machines - some from a template and others from cloning - and Storage vMotion. Each use case test is run on thin metadevices with hardware-accelerated Full Copy disabled and enabled. For the Full Copy enabled results both the 4 MB and 16 MB maximum copy sizes are displayed.

Use case configuration

In general, the Full Copy use cases were conducted using the same virtual machine, or a clone of that virtual machine converted to a template. This virtual machine was created using the vSphere Client, taking all standard defaults and with the following characteristics:

- 40 GB virtual disk
- Windows Server 2008 R2 operating system (no VMware Tools)

The virtual machine was then filled with data equal to 25 GB of the 40 GB. This configuration might be thought of as a worst case scenario for customers since most virtual machines that are used for cloning, even with the OS and all applications deployed on it, do not approach 25 GB of actual data.

Note: All results presented in graphs of Full Copy functionality are dependent upon the lab environment under which they were generated.While the majority of the time hardware acceleration will outperform a software copy, there are many factors that can impact how fast both software clones and Full Copy clones are created. For software clones the amount of CPU and memory on the ESXi server as well as the network will play a role. For Full Copy the number of engines and directors and other hardware components on the Symmetrix will impact the cloning time. The Symmetrix also intelligently balances workloads to ensure the highest priority to mission critical tasks. In both cases, the existing load on the environment will most certainly make a difference. Therefore results will vary.

Use Case 1: Deploying a virtual machine from a template

The first use case for Full Copy is the deployment of a virtual machine from a template. This is probably the most common type of virtual machine duplication for customers. A customer begins by creating a virtual machine, then installs the operating system and following this, loads the applications. The virtual machine is then customized so that when users are presented with the cloned virtual machine, they are able to enter in their personalized information.When an administrator completes the virtual machine base configuration, it is powered down and then converted to a template.

As using thin striped metavolumes is a Symmetrix best practice for VMware environments, these clones were created on a datastore backed by a thin striped metavolume.

The following graph in Figure 126 shows the time difference between a clone created using Full Copy and one using software copy.

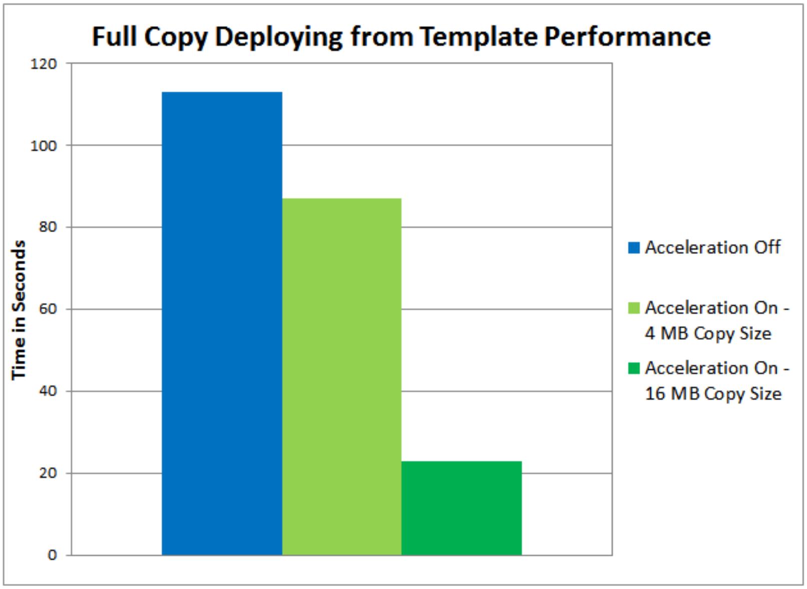


Figure 126 Deploying virtual machines from template

Use Case 2: Cloning hot and cold virtual machines

The second use case involves creating clones from existing virtual machines, both running (hot) and not running (cold). In addition, a third type of test was run in which a hot clone was sustaining heavy read I/O while it was being cloned. In each instance, utilizing the Full Copy functionality resulted in a significant performance benefit as visible in Figure 127, Figure 128 on page 251 and Figure 129 on page 252.

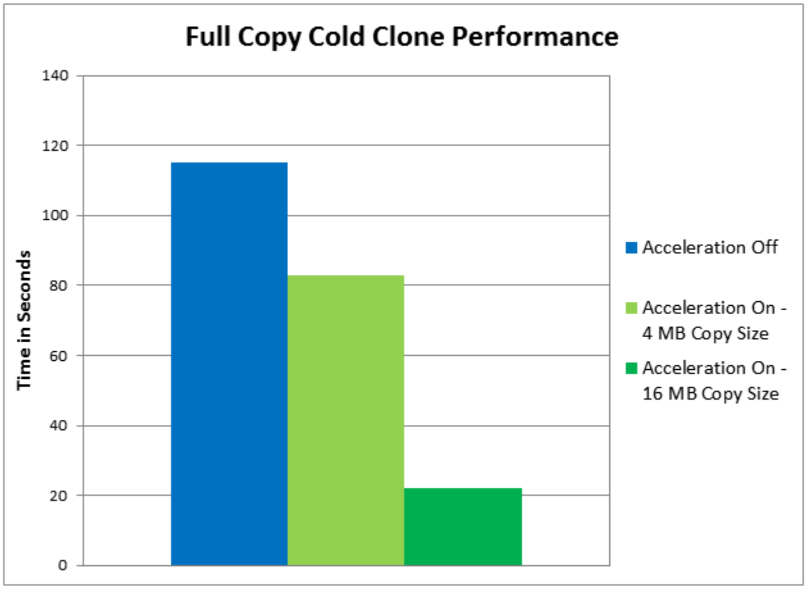


Figure 127 Performance of cold clones using hardware-accelerated Full Copy

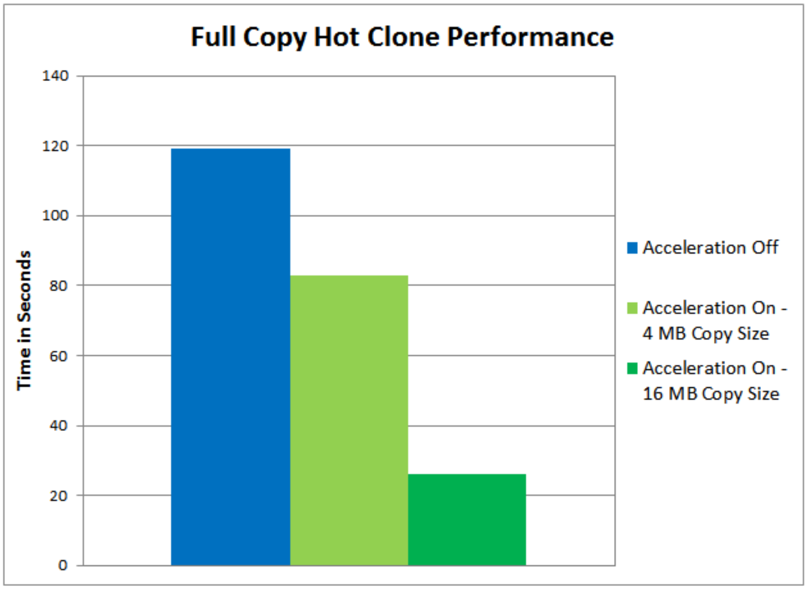


Figure 128 Performance of hot clones using hardware-accelerated Full Copy

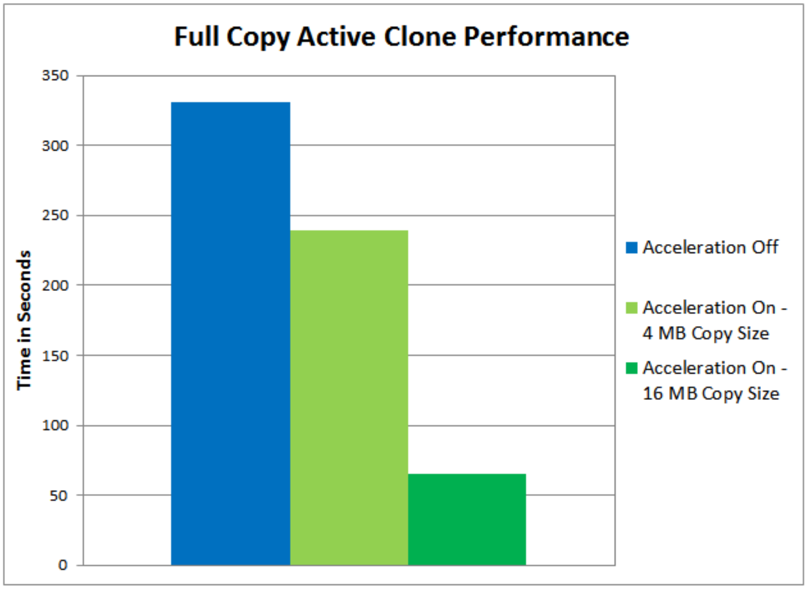


Figure 129 Cloning running VMs under load using Full Copy

Use Case 3: Creating simultaneous multiple clones

The third use case that was tested was to deploy four clones simultaneously from the same source virtual machine. This particular test can only be conducted against a cold virtual machine.

As seen in Figure 130 on page 253, the results again demonstrate the benefit of offloading as compared to software copying.

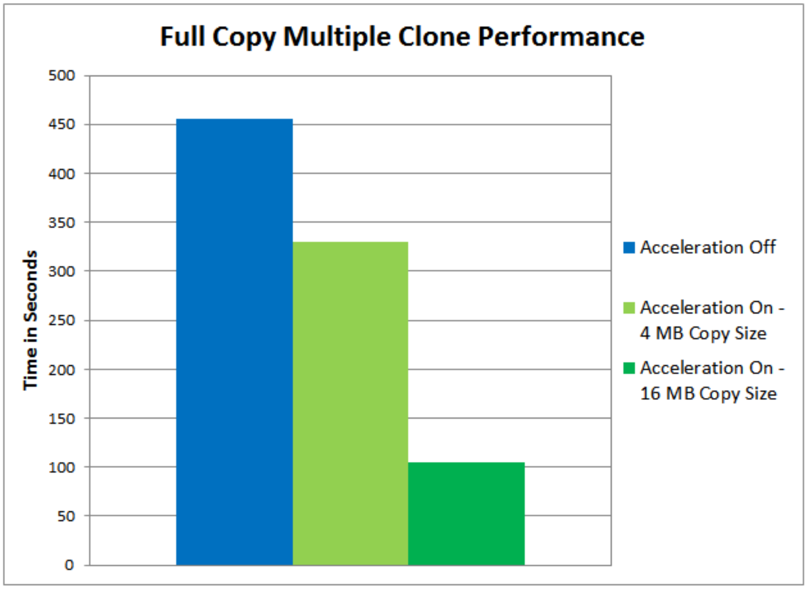


Figure 130 Cloning multiple virtual machines simultaneously with Full Copy

Use Case 4: Storage vMotion

The final use case is one that demonstrates a great benefit for customers who require datastore relocation of their virtual machine, but at the same time desire to reduce the impact to their live applications running on that virtual machine. This use case then is for Storage vMotion. With Full Copy enabled, the process of moving a virtual machine from one datastore to another datastore is offloaded. As mentioned previously, software copying requires CPU, memory, and network bandwidth. The resources it takes to software copy, therefore, might negatively impact the applications running on that virtual machine that is being moved. By utilizing Full Copy this is avoided and additionally as seen in Figure 131 on page 254, Full Copy is far quicker in moving that virtual machine.

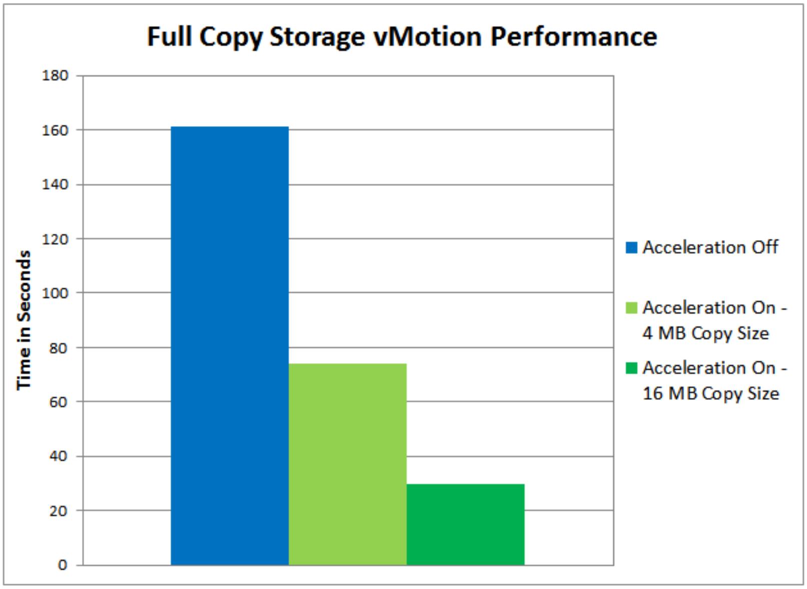


Figure 131 Storage vMotion using Full Copy

IMPORTANT

As can be seen from the graphs, utilizing the 16 MB copy size improves performance by a factor of four over the default 4 MB. Therefore if the VMware environment is only utilizing Symmetrix storage, adjust the parameter accordingly as detailed in "Full Copy Tuning".

Server resources — Impact to CPU and memory

While the relative time benefit of using Full Copy is apparent in the presented use cases, what of the CPU and memory utilization on the ESXi host? In this particular test environment and in these particular use cases, the CPU and memory utilization of the ESXi host did not differ significantly between performing a software clone or a Full Copy clone. That is not to say, however, this will hold true of all environments. The ESXi host used in this testing was not constrained by CPU or memory resources in any way as they were both plentiful; and the ESXi host was not under any other workload than the testing.

In a customer environment where resources may be limited and workloads on the ESXi host many and varied, the impact of running a software clone instead of a Full Copy clone may be more significant.

Caveats for using hardware-accelerated Full Copy[1]

The following are some general caveats that EMC provides when using this feature:

- Limit the number of simultaneous clones using Full Copy to three or four. This is not a strict limitation, but EMC believes this will ensure the best performance when offloading the copy process.
- A Symmetrix metadevice that has SAN Copy™, TimeFinder/Clone, TimeFinder/Snap or ChangeTracker sessions and certain RecoverPoint sessions will not support hardware-accelerated Full Copy. Any cloning or Storage vMotion operation run on datastores backed by these volumes will automatically be diverted to the default VMware software copy. Note that the vSphere Client has no knowledge of these sessions and as such the "Hardware Accelerated" column in the vSphere Client will still indicate "Supported" for these devices or datastores.
- Full copy is not supported for use with Open Replicator. VMware will revert to software copy in these cases.
- Although SRDF is supported with Full Copy, certain RDF operations, such as an RDF failover, will be blocked until the Symmetrix has completed copying the data from a clone or Storage vMotion.
- Using Full Copy on SRDF devices that are in a consistency group can render the group into a "sync in progress" mode until the copy completes on the Symmetrix. If this is undesirable state, customers are advised to disable Full Copy through the GUI or CLI while cloning to these devices.

[1] Beginning with Enginuity 5876.159.102, Full Copy (XCOPY) can be disabled at the system level if the caveats are particularly problematic for a customer. A special E-pack is required for GA Enginuity 5875 and 5876. If this change is required, please contact EMC Customer Support as there is no capability for customers to disable XCOPY even if at the correct Enginuity version.

- Full Copy will not be used if there is a metavolume reconfiguration taking place on the device backing the source or target datastore. VMware will revert to traditional writes in these cases. Conversely if a Full Copy was recently executed and a metavolume reconfiguration is attempted on the source or target metavolume backing the datastore, it will fail until the background copy is complete.

Choosing a virtual machine cloning methodology

The replication techniques described in the previous sections each have pros and cons with respect to their applicability to solve a given business problem. The following matrix in Table 4 provides a comparison of the different replication methods to use and the differing attributes of those methods.

Note: VMAXe/VMAX 10K does not support TF/Snap.

Table 4 **Comparison of storage array-based virtual machine cloning technologies**

	TF/Snap	TF/VP Snap	TF/Clone	VAAI Full Copy
Maximum number of copies	128 [b]	32	Incremental: 16 Non-inc: Unlimited	Unlimited [a]
Number of simultaneous copies	128 [b]	32	16	Unlimited [a]
Production impact	Minimal [c]	Minimal	Minimal [c]	Minimal
Scripting	Required	Required	Required	None
VM clone needed a long time	Not Recommended	Not Recommended	Recommended	Recommended
High write usage to VM clone	Not Recommended	Not Recommended	Recommended	Not Recommended
Restore capability	Yes	Yes	Yes	No
Clone a single VM	No	No	No	Yes

a. If the maximum number of extent copy tasks is reached on the storage system, cloning will revert to software.

b. The DMX array needs to be running Enginuity version 5772 or later for this feature. Older releases of code support a maximum of 15 simultaneous TF/Snap sessions.

c. Enhancements in Enginuity version 5772 or later minimize impact of TF/Snap and TF/Clone operations. Production volumes experience Copy on First Write and Copy on Access impact with older releases of Enginuity code.

Table 5 shows examples of the choices a VMware or storage administrator might make for cloning virtual machines based on the matrix presented in Table 4 on page 257.

Table 5 **Virtual machine cloning requirements and solutions**

System Requirements	Replication Choice
The application on the source volumes is performance sensitive, and the slightest degradation causes responsiveness of the system to miss SLAs	TimeFinder/Clone
Space and economy are a concern. Multiple copies are needed and retained only for a short time, with performance not critical	TimeFinder/Snap or TimeFinder/VP Snap
More than two simultaneous copies need to be made and the copies are to be used by development teams for a long time	TimeFinder/Clone
A single VM in the environment needs to be copied	VAAI Full Copy

Cloning at disaster protection site with vSphere 5

Many EMC customers employ a disaster recovery site for their VMware environments. Many also wish to take advantage of this hardware if it is idle, utilizing it as perhaps a test or development environment. They do not, however, wish to compromise their acceptable data loss, or Recovery Point Objective (RPO). In order to accomplish this, TimeFinder software can be used in conjunction with SRDF or RecoverPoint Continuous Remote Replication (CRR) to create point-in-time copies of the production data on the disaster recovery site. This is accomplished with no downtime and risk to the production data.

This section explains how to use TimeFinder copies of the remotely replicated production volumes and present them to a VMware environment without impacting the RPO of the production environment.

Note: This section does not address the cloning of devices used as RDMs at the production site.

Note: The procedures included in this section do not address solutions that leverage VMware's vCenter Site Recovery Manager. For information on using EMC Symmetrix with SRM, consult the *Using EMC SRDF Adapter for VMware vCenter Site Recovery Manager 5 TechBook*.

Using TimeFinder copies

Creating copies of volumes with TimeFinder software is covered extensively in "Cloning of vSphere Virtual Machines" on page 211 and, therefore, it will not be repeated herein. In particular, as this section is concerned with production environments, "Copying running virtual machines using EMC Consistency technology" on page 233 should be consulted. That section details the use of EMC Consistency technology which ensures that the point-in-time copy of the volumes is in a dependent-write consistent data state and the virtual machine in the datastore are restartable.

In addition, the assumption is made that the system administrator has mapped and masked all devices to the ESXi host(s) involved in this process. This means that both the R2 and RecoverPoint remote devices and any of their potential BCVs, clones, or snaps are seen by the ESXi host. If additional information is required on this procedure, see "Adding and removing EMC Symmetrix devices to VMware ESX hosts" on page 28.

The examples in this chapter are based upon a production and disaster recovery site that each have a Symmetrix VMAXe running SRDF between them. Because of this, only TimeFinder/Mirror and TimeFinder/Clone are used as TimeFinder/Snap is not currently supported on the VMAXe/VMAX 10K. If the environments are running on a VMAX, there is no reason TimeFinder/Snap cannot be used instead of Clone with SRDF. Differences in how to use TimeFinder with RecoverPoint will be addressed next.

ESXi and duplicate extents

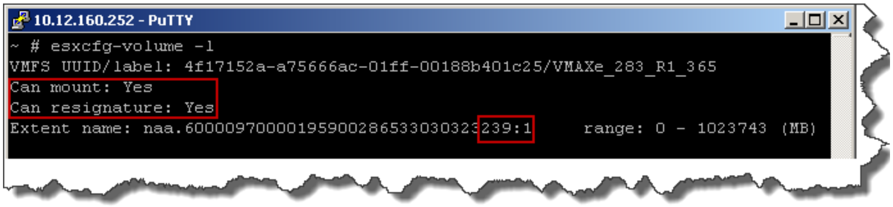
In vSphere 5, mounting a copy of a remote production volume requires some thought and planning. The issue that one encounters is that when the R2 is cloned, the VMFS volume signature is copied over. The signature is generated in part from the WWN of the device. Since the R2 has a different WWN than the R1 device, ESXi sees the signature as invalid. It will, therefore, require the R2 device to be resignatured when mounting.[1]

The same problem also exists when cloning the R2 to another device using TimeFinder software. The clone or BCV device has a different WWN (just like the R2) and therefore putting the R1 signature on that volume will also result in ESXi requiring a resignature. What complicates the situation, however, is now ESXi sees not just the R2 but also the copy of the R2 and recognizes that they have the same signature. When ESXi is presented with duplicates, by default it will not allow the user to mount either of the VMFS volumes.

Using the CLI, one can see if duplicates exist. Figure 132 on page 261 shows an example of what ESXi will report when only the R2, which is identified in the figure by the last three digits of the network address authority number (naa), or 239:1, is masked to the host.

[1] The user can force mount the volume with the same signature if desired. The only exception to this is when using VMware SRM.

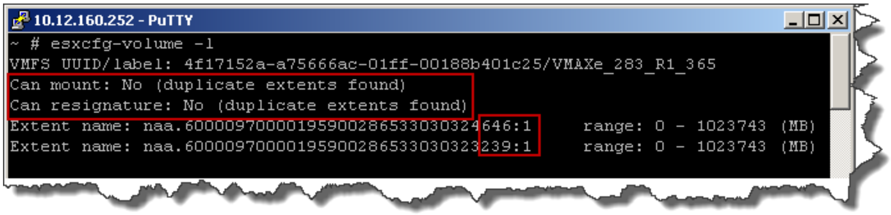
Using the command `esxcfg-volume -l`, ESXi displays only the R2 that is masked to the host and reports that it can both be mounted and resignatured. ESXi has recognized that the signature is not valid for the device because the WWN is different than the R1.



```
~ # esxcfg-volume -l
VMFS UUID/label: 4f17152a-a75666ac-01ff-00188b401c25/VMAXe_283_R1_365
Can mount: Yes
Can resignature: Yes
Extent name: naa.60000970000195900286533030323239:1     range: 0 - 1023743 (MB)
```

Figure 132 An R2 device on the remote ESXi host which can be resignatured

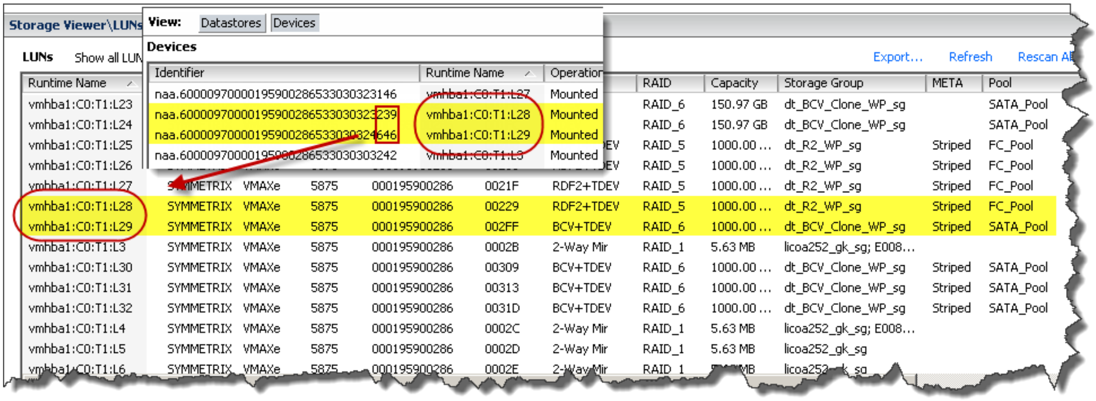
Now if a masked BCV, with the naa ending 646 as seen in Figure 133, of the R2 in the previous Figure 132 is established, and the same command is run on the ESXi host, the results are much different, as seen in Figure 133. ESXi recognizes that the devices have the same VMFS volume signature. When presented with this situation, ESXi will not allow either device to be mounted or resignatured.



```
~ # esxcfg-volume -l
VMFS UUID/label: 4f17152a-a75666ac-01ff-00188b401c25/VMAXe_283_R1_365
Can mount: No (duplicate extents found)
Can resignature: No (duplicate extents found)
Extent name: naa.60000970000195900286533030324646:1     range: 0 - 1023743 (MB)
Extent name: naa.60000970000195900286533030323239:1     range: 0 - 1023743 (MB)
```

Figure 133 An R2 device and its copy on the remote ESXi host that cannot be resignatured

If an attempt is made in the vSphere Client to resignature and mount this datastore, it will be prevented. In the vSphere Client, using the EMC Virtual Storage Integrator (VSI) Storage Viewer feature shown in Figure 134, both the R2 (239, L28) and the BCV (646, L29) are visible.



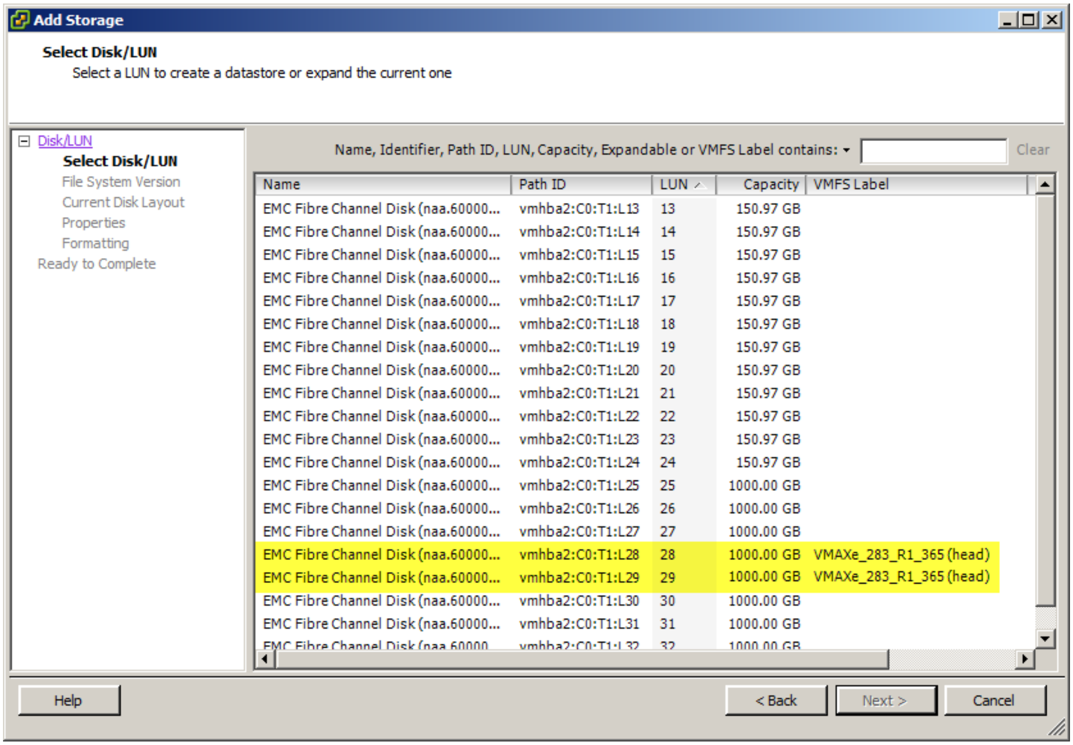
Devices

Identifier	Runtime Name	Operation
naa.60000970000195900286533030323146	vmhba1:C0:T1:L27	Mounted
naa.60000970000195900286533030323239	vmhba1:C0:T1:L28	Mounted
naa.60000970000195900286533030324646	vmhba1:C0:T1:L29	Mounted
naa.60000970000195900286533030303242	vmhba1:C0:T1:L3	Mounted

Runtime Name							RAID	Capacity	Storage Group	META	Pool
vmhba1:C0:T1:L23							RAID_6	150.97 GB	dt_BCV_Clone_WP_sg		SATA_Pool
vmhba1:C0:T1:L24							RAID_6	150.97 GB	dt_BCV_Clone_WP_sg		SATA_Pool
vmhba1:C0:T1:L25							RAID_5	1000.00 ...	dt_R2_WP_sg	Striped	FC_Pool
vmhba1:C0:T1:L26							RAID_5	1000.00 ...	dt_R2_WP_sg	Striped	FC_Pool
vmhba1:C0:T1:L27	SYMMETRIX	VMAXe	5875	000195900286	0021F	RDF2+TDEV	RAID_5	1000.00 ...	dt_R2_WP_sg	Striped	FC_Pool
vmhba1:C0:T1:L28	SYMMETRIX	VMAXe	5875	000195900286	00229	RDF2+TDEV	RAID_5	1000.00 ...	dt_R2_WP_sg	Striped	FC_Pool
vmhba1:C0:T1:L29	SYMMETRIX	VMAXe	5875	000195900286	002FF	BCV+TDEV	RAID_6	1000.00 ...	dt_BCV_Clone_WP_sg	Striped	SATA_Pool
vmhba1:C0:T1:L3	SYMMETRIX	VMAXe	5875	000195900286	0002B	2-Way Mir	RAID_1	5.63 MB	licoa252_gk_sg; E008...		
vmhba1:C0:T1:L30	SYMMETRIX	VMAXe	5875	000195900286	00309	BCV+TDEV	RAID_6	1000.00 ...	dt_BCV_Clone_WP_sg	Striped	SATA_Pool
vmhba1:C0:T1:L31	SYMMETRIX	VMAXe	5875	000195900286	00313	BCV+TDEV	RAID_6	1000.00 ...	dt_BCV_Clone_WP_sg	Striped	SATA_Pool
vmhba1:C0:T1:L32	SYMMETRIX	VMAXe	5875	000195900286	0031D	BCV+TDEV	RAID_6	1000.00 ...	dt_BCV_Clone_WP_sg	Striped	SATA_Pool
vmhba1:C0:T1:L4	SYMMETRIX	VMAXe	5875	000195900286	0002C	2-Way Mir	RAID_1	5.63 MB	licoa252_gk_sg; E008...		
vmhba1:C0:T1:L5	SYMMETRIX	VMAXe	5875	000195900286	0002D	2-Way Mir	RAID_1	5.63 MB	licoa252_gk_sg		
vmhba1:C0:T1:L6	SYMMETRIX	VMAXe	5875	000195900286	0002E	2-Way Mir	RAID_1	[illegible] MB	licoa252_gk_sg		

Figure 134 Duplicate extents in VSI

Using the add storage wizard, the user is presented with a list of devices to use for mounting or creating a new datastore. In Figure 135 on page 263 both the R2 and BCV are displayed along with the VMFS label which is the same for both, VMAXe_238_R1_365.



Select Disk/LUN

Select a LUN to create a datastore or expand the current one

Name, Identifier, Path ID, LUN, Capacity, Expandable or VMFS Label contains:

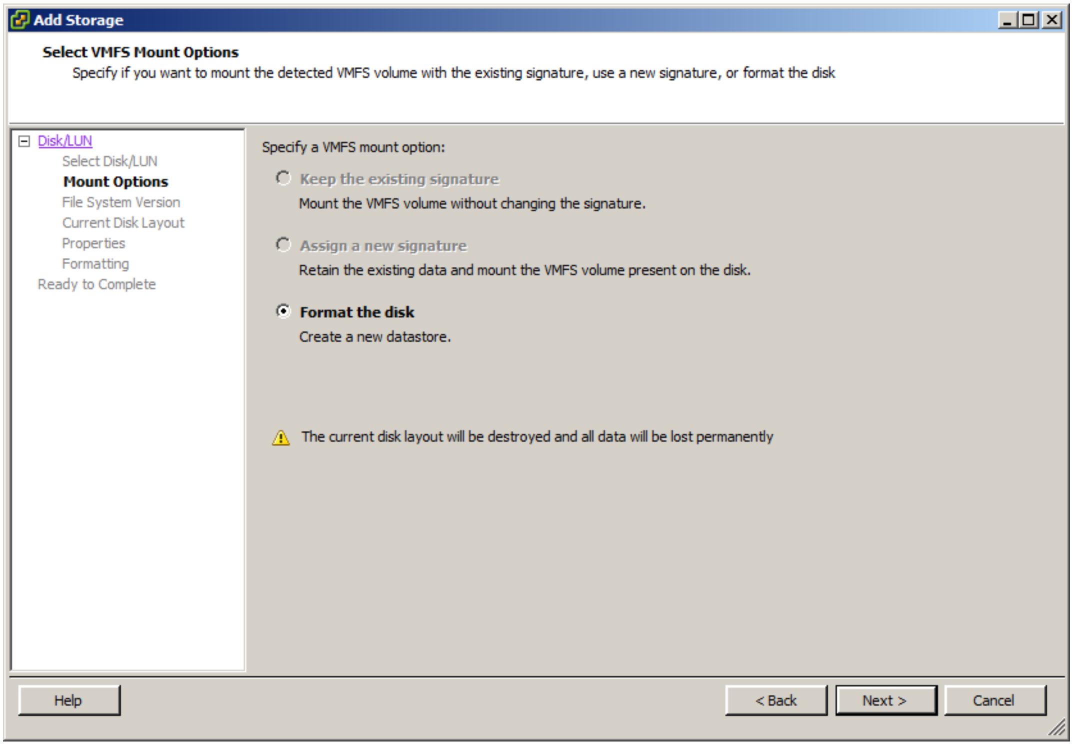
Name	Path ID	LUN	Capacity	VMFS Label
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L13	13	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L14	14	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L15	15	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L16	16	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L17	17	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L18	18	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L19	19	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L20	20	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L21	21	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L22	22	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L23	23	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L24	24	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L25	25	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L26	26	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L27	27	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L28	28	1000.00 GB	VMAXe_283_R1_365 (head)
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L29	29	1000.00 GB	VMAXe_283_R1_365 (head)
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L30	30	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L31	31	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L32	32	1000.00 GB	

Figure 135 Attempt to mount duplicate extents

If the user proceeds, the only option presented will be to format the disk and thus the user would lose the datastore on the BCV or clone. This is shown in Figure 136 on page 264.

IMPORTANT

If the R2 device is chosen by accident during this process, the format will fail as the R2 is a write disabled device.



Add Storage

Select VMFS Mount Options

Specify if you want to mount the detected VMFS volume with the existing signature, use a new signature, or format the disk

Specify a VMFS mount option:

○ Keep the existing signature
Mount the VMFS volume without changing the signature.

○ Assign a new signature
Retain the existing data and mount the VMFS volume present on the disk.

● Format the disk
Create a new datastore.

⚠ The current disk layout will be destroyed and all data will be lost permanently

Figure 136 Attempt to mount duplicate extents permits format only

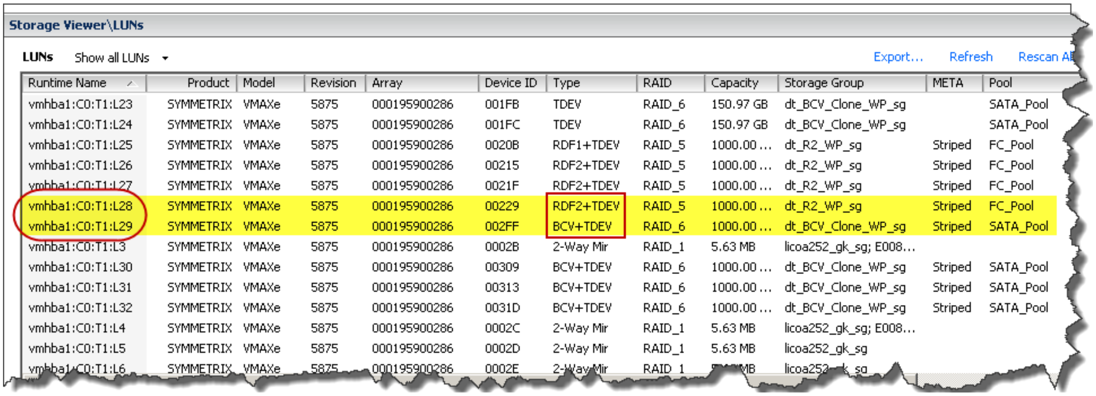
Fortunately, this situation can be avoided. There are a few methodologies that can be employed but only two are recommended and presented. These will be addressed in the following sections.

Detaching the remote volume

One of the benefits that comes with vSphere 5 is that the user is able to “detach” a device from the ESXi host. This was discussed in “Avoiding APD/PDL — Device removal” on page 94 in relation to the All Paths Down condition. In a situation where duplicate extents exist, if one device is detached from the ESXi host, ESXi will then see a single snapshot VMFS volume and revert to the behavior of being able to resignature and mount the device. The next section will walk through an example of this.

Detaching a device and resignaturing the copy

For consistency this example will use the same devices as presented earlier. Reviewing the VSI screen, Figure 137, there is an R2 and a BCV masked to the ESXi host.



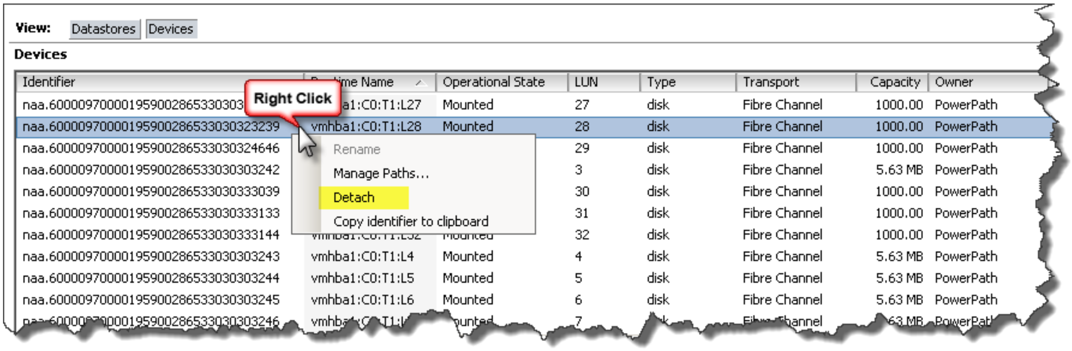
Storage Viewer\LUNs

LUNs Show all LUNs

Runtime Name	Product	Model	Revision	Array	Device ID	Type	RAID	Capacity	Storage Group	META	Pool
vmhba1:C0:T1:L23	SYMMETRIX	VMAXe	5875	000195900286	001FB	TDEV	RAID_6	150.97 GB	dt_BCV_Clone_WP_sg		SATA_Pool
vmhba1:C0:T1:L24	SYMMETRIX	VMAXe	5875	000195900286	001FC	TDEV	RAID_6	150.97 GB	dt_BCV_Clone_WP_sg		SATA_Pool
vmhba1:C0:T1:L25	SYMMETRIX	VMAXe	5875	000195900286	0020B	RDF1+TDEV	RAID_5	1000.00 ...	dt_R2_WP_sg	Striped	FC_Pool
vmhba1:C0:T1:L26	SYMMETRIX	VMAXe	5875	000195900286	00215	RDF2+TDEV	RAID_5	1000.00 ...	dt_R2_WP_sg	Striped	FC_Pool
vmhba1:C0:T1:L27	SYMMETRIX	VMAXe	5875	000195900286	0021F	RDF2+TDEV	RAID_5	1000.00 ...	dt_R2_WP_sg	Striped	FC_Pool
vmhba1:C0:T1:L28	SYMMETRIX	VMAXe	5875	000195900286	00229	RDF2+TDEV	RAID_5	1000.00 ...	dt_R2_WP_sg	Striped	FC_Pool
vmhba1:C0:T1:L29	SYMMETRIX	VMAXe	5875	000195900286	002FF	BCV+TDEV	RAID_6	1000.00 ...	dt_BCV_Clone_WP_sg	Striped	SATA_Pool
vmhba1:C0:T1:L3	SYMMETRIX	VMAXe	5875	000195900286	0002B	2-Way Mir	RAID_1	5.63 MB	licoa252_gk_sg; E008...		
vmhba1:C0:T1:L30	SYMMETRIX	VMAXe	5875	000195900286	00309	BCV+TDEV	RAID_6	1000.00 ...	dt_BCV_Clone_WP_sg	Striped	SATA_Pool
vmhba1:C0:T1:L31	SYMMETRIX	VMAXe	5875	000195900286	00313	BCV+TDEV	RAID_6	1000.00 ...	dt_BCV_Clone_WP_sg	Striped	SATA_Pool
vmhba1:C0:T1:L32	SYMMETRIX	VMAXe	5875	000195900286	0031D	BCV+TDEV	RAID_6	1000.00 ...	dt_BCV_Clone_WP_sg	Striped	SATA_Pool
vmhba1:C0:T1:L4	SYMMETRIX	VMAXe	5875	000195900286	0002C	2-Way Mir	RAID_1	5.63 MB	licoa252_gk_sg; E008...		
vmhba1:C0:T1:L5	SYMMETRIX	VMAXe	5875	000195900286	0002D	2-Way Mir	RAID_1	5.63 MB	licoa252_gk_sg		
vmhba1:C0:T1:L6	SYMMETRIX	VMAXe	5875	000195900286	0002E	2-Way Mir	RAID_1	[illegible] MB	licoa252_gk_sg		

Figure 137 **VSI displaying two duplicate volumes presented to a single ESXi host**

At this point the BCV is established and split from the R2 device. The establish and split is done first to reduce the amount of time the device is detached. Using the vSphere 5 functionality, the R2 can now be detached. Select the R2 device in the devices window in the vSphere Client. Right-click on it and select “Detach” as in Figure 138 on page 266.



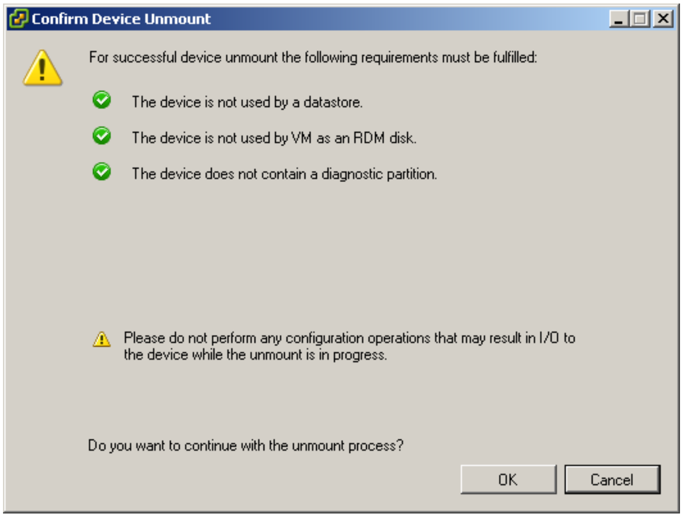
Devices

Identifier	Runtime Name	Operational State	LUN	Type	Transport	Capacity	Owner
naa.600009700001959002865330303	a1:C0:T1:L27	Mounted	27	disk	Fibre Channel	1000.00	PowerPath
naa.60000970000195900286533030323239	vmhba1:C0:T1:L28	Mounted	28	disk	Fibre Channel	1000.00	PowerPath
naa.60000970000195900286533030324646			29	disk	Fibre Channel	1000.00	PowerPath
naa.60000970000195900286533030303242			3	disk	Fibre Channel	5.63 MB	PowerPath
naa.60000970000195900286533030333039			30	disk	Fibre Channel	1000.00	PowerPath
naa.60000970000195900286533030333133			31	disk	Fibre Channel	1000.00	PowerPath
naa.60000970000195900286533030333144	vmhba1:C0:T1:L32	Mounted	32	disk	Fibre Channel	1000.00	PowerPath
naa.60000970000195900286533030303243	vmhba1:C0:T1:L4	Mounted	4	disk	Fibre Channel	5.63 MB	PowerPath
naa.60000970000195900286533030303244	vmhba1:C0:T1:L5	Mounted	5	disk	Fibre Channel	5.63 MB	PowerPath
naa.60000970000195900286533030303245	vmhba1:C0:T1:L6	Mounted	6	disk	Fibre Channel	5.63 MB	PowerPath
naa.60000970000195900286533030303246	vmhba	ounted	7		Fibre Channel	63 MB	PowerPat

Figure 138 **Detaching the R2 device**

VMware will run through three prerequisites, seen in Figure 139 to be sure that the device can be detached before allowing the user to confirm.

Note: Detaching the device is a VMware function and has no impact on the Symmetrix mapping or masking of the device to the ESXi host.



For successful device unmount the following requirements must be fulfilled:

✓ The device is not used by a datastore.

✓ The device is not used by VM as an RDM disk.

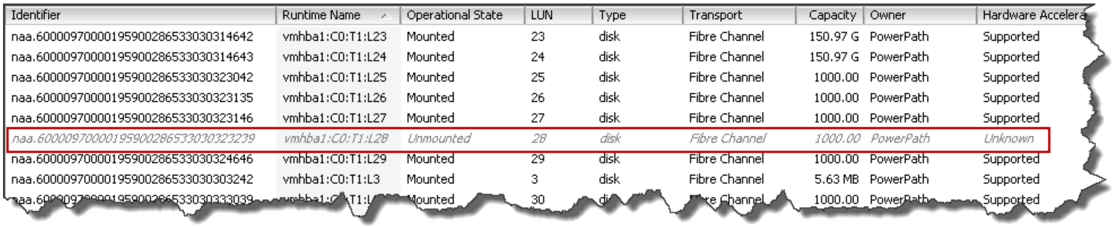
✓ The device does not contain a diagnostic partition.

⚠ Please do not perform any configuration operations that may result in I/O to the device while the unmount is in progress.

Do you want to continue with the unmount process?

Figure 139 **Confirming the detaching of the R2**

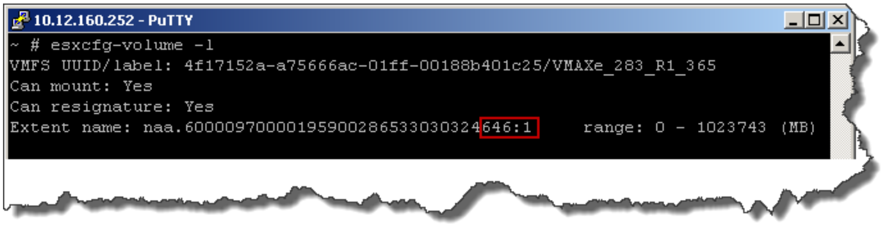
Once confirmed, the device is detached. The device does not disappear from the devices screen, but rather grays out. This is shown in Figure 140.



Identifier	Runtime Name	Operational State	LUN	Type	Transport	Capacity	Owner	Hardware Accelera
naa.60000970000195900286533030314642	vmhba1:C0:T1:L23	Mounted	23	disk	Fibre Channel	150.97 G	PowerPath	Supported
naa.60000970000195900286533030314643	vmhba1:C0:T1:L24	Mounted	24	disk	Fibre Channel	150.97 G	PowerPath	Supported
naa.60000970000195900286533030323042	vmhba1:C0:T1:L25	Mounted	25	disk	Fibre Channel	1000.00	PowerPath	Supported
naa.60000970000195900286533030323135	vmhba1:C0:T1:L26	Mounted	26	disk	Fibre Channel	1000.00	PowerPath	Supported
naa.60000970000195900286533030323146	vmhba1:C0:T1:L27	Mounted	27	disk	Fibre Channel	1000.00	PowerPath	Supported
naa.60000970000195900286533030323239	vmhba1:C0:T1:L28	Unmounted	28	disk	Fibre Channel	1000.00	PowerPath	Unknown
naa.60000970000195900286533030324646	vmhba1:C0:T1:L29	Mounted	29	disk	Fibre Channel	1000.00	PowerPath	Supported
naa.60000970000195900286533030303242	vmhba1:C0:T1:L3	Mounted	3	disk	Fibre Channel	5.63 MB	PowerPath	Supported
naa.60000970000195900286533030333039	vmhba1:C0:T1:L30	Mounted	30	disk	Fibre Channel	1000.00	PowerPath	Supported

Figure 140 Post-detach of device now grayed out

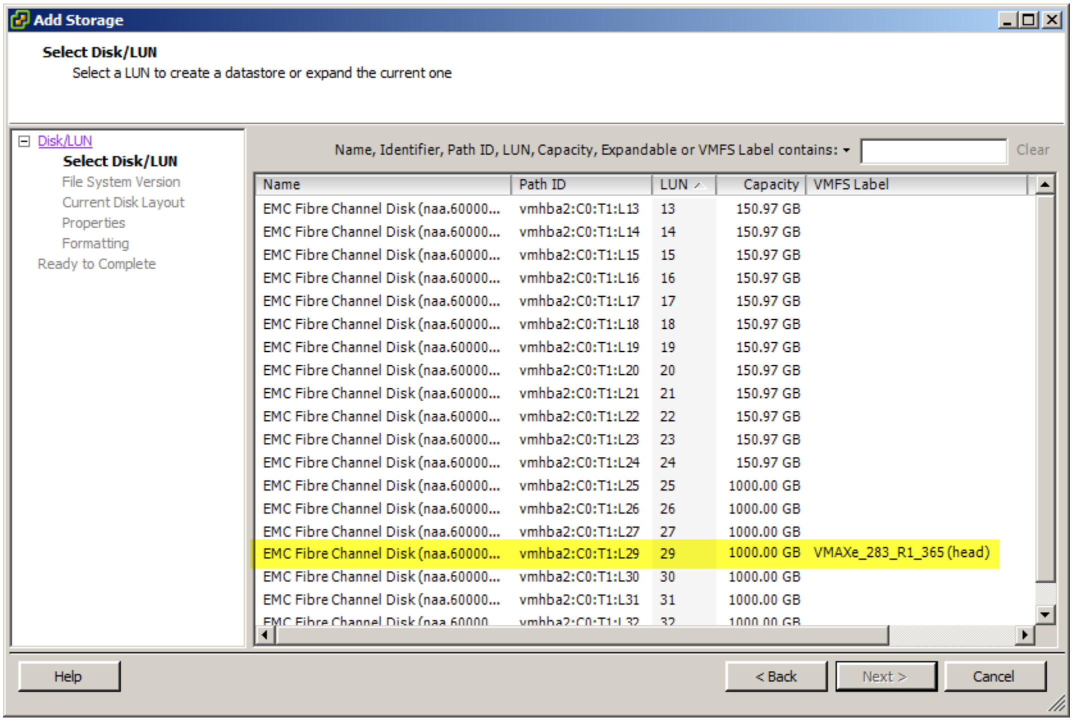
Now that the device has been detached, when `esxcfg-volume -l` is run, only the BCV is shown, and it is available for resignaturing and mounting as seen in Figure 141.



```
~ # esxcfg-volume -l
VMFS UUID/label: 4f17152a-a75666ac-01ff-00188b401c25/VMAXe_283_R1_365
Can mount: Yes
Can resignature: Yes
Extent name: naa.60000970000195900286533030324646:1     range: 0 - 1023743 (MB)
```

Figure 141 Esxcfg-volume -l showing only the BCV after the R2 device is detached

Subsequently when using the vSphere client, the R2 no longer appears, just the BCV device as in the add storage wizard in Figure 142.



Select Disk/LUN

Select a LUN to create a datastore or expand the current one

Disk/LUN
- **Select Disk/LUN**
- File System Version
- Current Disk Layout
- Properties
- Formatting

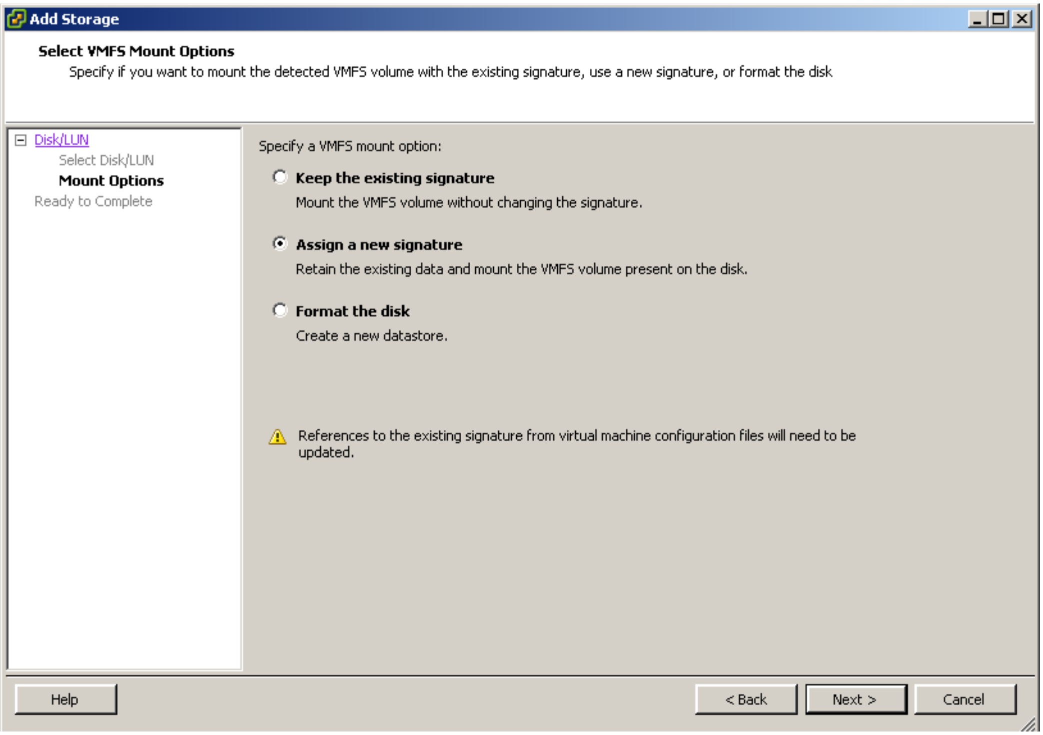
Ready to Complete

Name, Identifier, Path ID, LUN, Capacity, Expandable or VMFS Label contains:

Name	Path ID	LUN	Capacity	VMFS Label
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L13	13	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L14	14	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L15	15	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L16	16	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L17	17	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L18	18	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L19	19	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L20	20	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L21	21	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L22	22	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L23	23	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L24	24	150.97 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L25	25	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L26	26	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L27	27	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L29	29	1000.00 GB	VMAXe_283_R1_365 (head)
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L30	30	1000.00 GB	
EMC Fibre Channel Disk (naa.60000...	vmhba2:C0:T1:L31	31	1000.00 GB	
EMC Fibre Channel Disk (naa.60000	vmhba2:C0:T1:L32	32	1000.00 GB	

Figure 142 Mounting the BCV after R2 detach

Proceeding through the wizard, the user can now choose among all three options presented in Figure 143 on page 269: keeping the signature, assigning a new signature, or formatting the disk.



Add Storage

Select VMFS Mount Options

Specify if you want to mount the detected VMFS volume with the existing signature, use a new signature, or format the disk

Specify a VMFS mount option:

○ **Keep the existing signature**
Mount the VMFS volume without changing the signature.

● **Assign a new signature**
Retain the existing data and mount the VMFS volume present on the disk.

○ **Format the disk**
Create a new datastore.

⚠ References to the existing signature from virtual machine configuration files will need to be updated.

Figure 143 Choosing how to mount the BCV

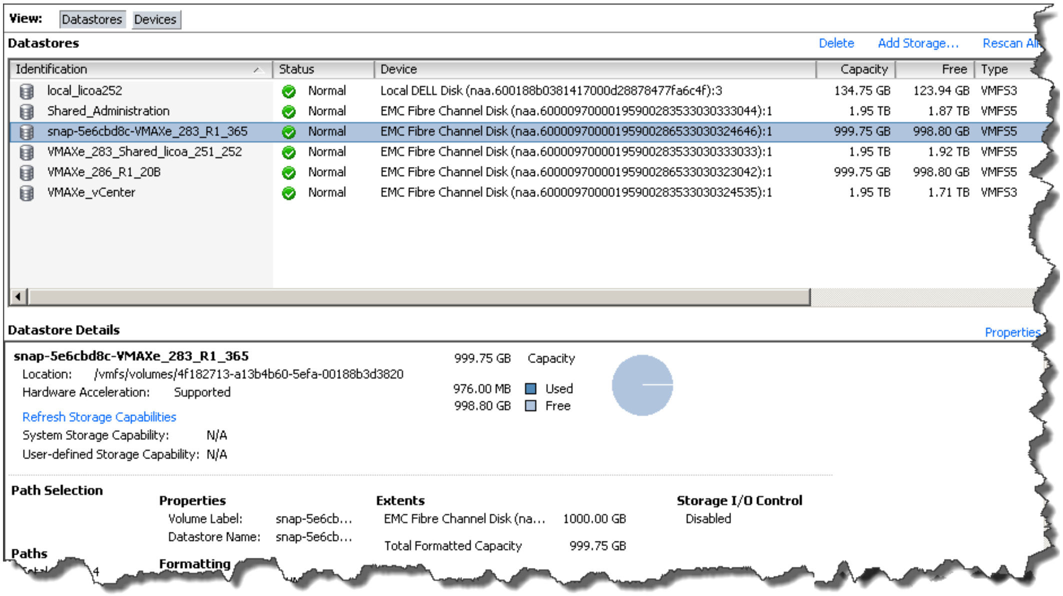
The best practice is to always assign a new signature to the volume to prevent any problems when the R2 is re-attached.

⚠ WARNING

If the user chooses NOT to resignature the volume and mounts it with the same signature, the R2 cannot be re-attached. The user will receive an error that states the operation is not allowed in the current state. This is because ESXi recognizes the volumes have the same signature and since one is already mounted, a second volume cannot, in a sense, occupy the same space. The resolution to this is to unmount the datastore and detach the underlying BCV before attempting to re-attach the R2.

When a new signature is created, VMware automatically assigns a prefix to the datastore that is “snap-xxxxxxxx-”. The “xxxxxxxx” part of the name is system generated. In this example, once the datastore is mounted, it appears as snap-5e6cbd8c-VMAXe_283_R1_365 as in

Figure 144.



Datastores

Identification	Status	Device	Capacity	Free	Type
local_licoa252	✓ Normal	Local DELL Disk (naa.600188b0381417000d28878477fa6c4f):3	134.75 GB	123.94 GB	VMFS3
Shared_Administration	✓ Normal	EMC Fibre Channel Disk (naa.60000970000195900283533030333044):1	1.95 TB	1.87 TB	VMFS5
snap-5e6cbd8c-VMAXe_283_R1_365	✓ Normal	EMC Fibre Channel Disk (naa.60000970000195900286533030324646):1	999.75 GB	998.80 GB	VMFS5
VMAXe_283_Shared_licoa_251_252	✓ Normal	EMC Fibre Channel Disk (naa.60000970000195900283533030333033):1	1.95 TB	1.92 TB	VMFS5
VMAXe_286_R1_20B	✓ Normal	EMC Fibre Channel Disk (naa.60000970000195900286533030323042):1	999.75 GB	998.80 GB	VMFS5
VMAXe_vCenter	✓ Normal	EMC Fibre Channel Disk (naa.60000970000195900283533030324535):1	1.95 TB	1.71 TB	VMFS3

Datastore Details

snap-5e6cbd8c-VMAXe_283_R1_365
Location: /vmfs/volumes/4f182713-a13b4b60-5efa-00188b3d3820
Hardware Acceleration: Supported

System Storage Capability: N/A
User-defined Storage Capability: N/A

999.75 GB Capacity
976.00 MB ■ Used
998.80 GB □ Free

Path Selection

Properties
Volume Label: snap-5e6cb...
Datastore Name: snap-5e6cb...

Extents
EMC Fibre Channel Disk (na... 1000.00 GB
Total Formatted Capacity 999.75 GB

Storage I/O Control
Disabled

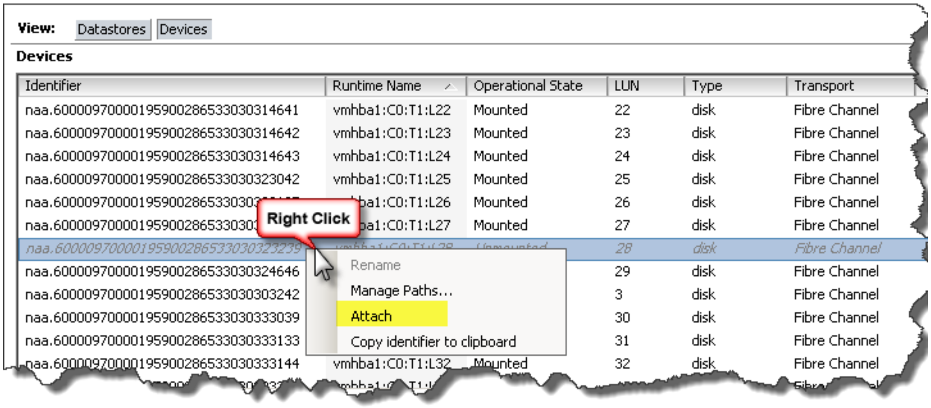
Paths
4

Formatting

Figure 144 **Post-mount of BCV with resignature**

With the completion of the resignature and mount, the R2 can now be re-attached. This is a best practice to ensure in the event of the production site failure it would be available immediately.

To do this, return to the devices screen and locate the grayed-out R2 device. Right-click on the device and select "Attach" from the menu as in Figure 145 on page 271. If there are multiple detached devices, use the VSI Storage Viewer to correlate the runtime names with the correct devices.



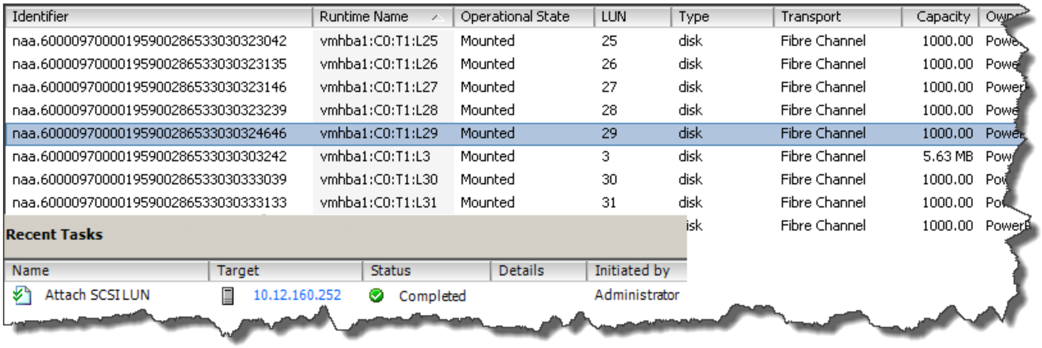
View: Datastores | Devices

Devices

Identifier	Runtime Name	Operational State	LUN	Type	Transport
naa.60000970000195900286533030314641	vmhba1:C0:T1:L22	Mounted	22	disk	Fibre Channel
naa.60000970000195900286533030314642	vmhba1:C0:T1:L23	Mounted	23	disk	Fibre Channel
naa.60000970000195900286533030314643	vmhba1:C0:T1:L24	Mounted	24	disk	Fibre Channel
naa.60000970000195900286533030323042	vmhba1:C0:T1:L25	Mounted	25	disk	Fibre Channel
naa.6000097000019590028653303	ba1:C0:T1:L26	Mounted	26	disk	Fibre Channel
naa.600009700001959002865330303	ba1:C0:T1:L27	Mounted	27	disk	Fibre Channel
naa.60000970000195900286533030323239	*vmhba1:C0:T1:L28*	*Unmounted*	*28*	*disk*	*Fibre Channel*
naa.60000970000195900286533030324646			29	disk	Fibre Channel
naa.60000970000195900286533030303242			3	disk	Fibre Channel
naa.60000970000195900286533030333039			30	disk	Fibre Channel
naa.60000970000195900286533030333133			31	disk	Fibre Channel
naa.60000970000195900286533030333144	vmhba1:C0:T1:L32	Mounted	32	disk	Fibre Channel

Figure 145 Re-attach R2 after mounting BCV

Unlike detaching, re-attaching does all the validation in the background. A task will appear indicating that the device is attaching as in Figure 146. Once the task completes the device will return to black from gray.



Identifier	Runtime Name	Operational State	LUN	Type	Transport	Capacity	Own
naa.60000970000195900286533030323042	vmhba1:C0:T1:L25	Mounted	25	disk	Fibre Channel	1000.00	Powe
naa.60000970000195900286533030323135	vmhba1:C0:T1:L26	Mounted	26	disk	Fibre Channel	1000.00	Powe
naa.60000970000195900286533030323146	vmhba1:C0:T1:L27	Mounted	27	disk	Fibre Channel	1000.00	Power
naa.60000970000195900286533030323239	vmhba1:C0:T1:L28	Mounted	28	disk	Fibre Channel	1000.00	Powe
naa.60000970000195900286533030324646	vmhba1:C0:T1:L29	Mounted	29	disk	Fibre Channel	1000.00	Powe
naa.60000970000195900286533030303242	vmhba1:C0:T1:L3	Mounted	3	disk	Fibre Channel	5.63 MB	Pow
naa.60000970000195900286533030333039	vmhba1:C0:T1:L30	Mounted	30	disk	Fibre Channel	1000.00	Po
naa.60000970000195900286533030333133	vmhba1:C0:T1:L31	Mounted	31	disk	Fibre Channel	1000.00	Po
				isk	Fibre Channel	1000.00	Power

Recent Tasks

Name	Target	Status	Details	Initiated by
Attach SCSI LUN	10.12.160.252	✓ Completed		Administrator

Figure 146 Successful re-attachment of the R2

Note: It is possible to unmask the R2 from the remote ESXi hosts instead of detaching the device, and then remasking it once the resignature and mounting is complete, rescanning the HBAs each time. This is not a recommended best practice, however, and can lead to complications.

LVM.enableResignature parameter

The second methodology for mounting TimeFinder copies of remote devices involves changing an advance parameter on the ESXi host. This parameter is /LVM/EnableResignature and controls whether ESXi will automatically resignature snapshot LUNs (e.g. TimeFinder copies). By default this parameter is set to ‘0’ so that resignaturing must proceed as outlined in the previous section. If this parameter is set to 1, all snapshot LUNs will be resignatured when the ESXi host rescans the HBAs. Since in vSphere 5 this will be done automatically when there is some VMFS management operation, once the parameter is set it is inevitable that any of the VMFS volumes that can be resignatured, will be resignatured. It is therefore essential when setting this parameter to be aware of all volumes that the ESXi host in question can see.

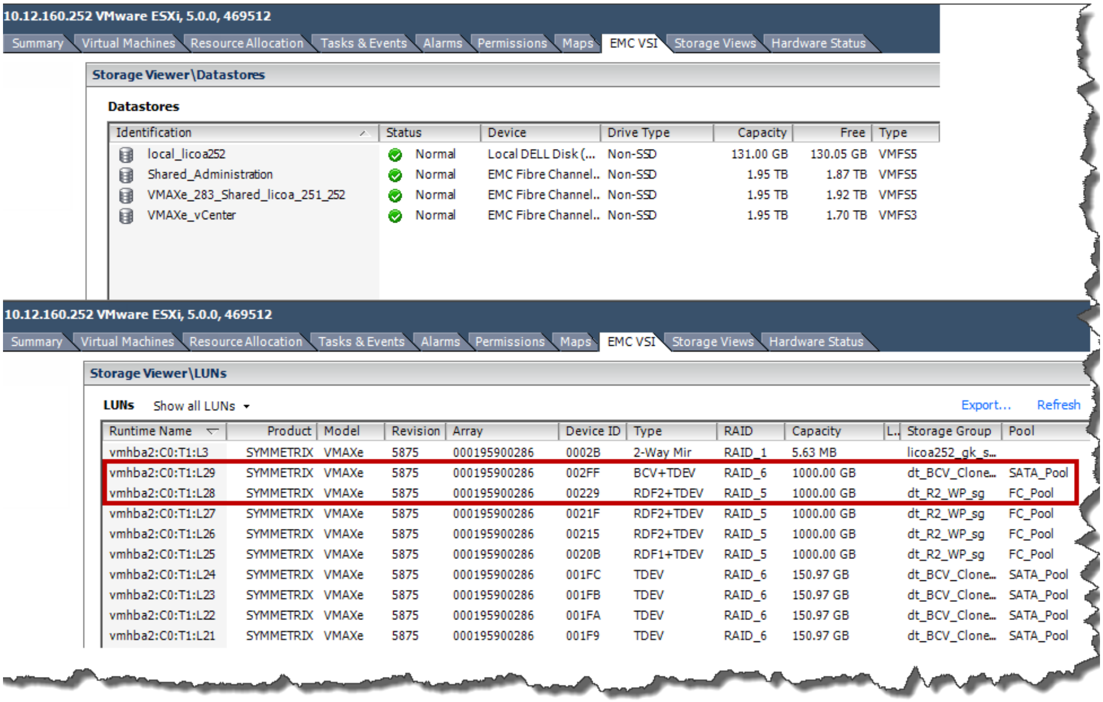
Note: If there are multiple ESXi hosts in a cluster that all have access to the same devices, even if only one hosts has the parameter set the resignature and mounting will be done.

⚠ WARNING

Setting the LVM.enableResignature flag on ESX hosts is a host-wide operation and, if set, all snapshot LUNs that can be resignatured are resignatured during the subsequent host rescan. If snapshot volumes are forcefully mounted to ESX hosts on the recovery site, these LUNs are resignatured as part of a host rescan during a test failover operation. Accordingly, all of the virtual machines on these volumes become inaccessible. To prevent outages, ensure that no forcefully-mounted snapshot LUNs are visible to ESX hosts on the recovery site.

LVM.enableResignature example

To demonstrate how this parameter works, the LUNs presented previously will be used. As a reminder, the two devices used in the example are the R2 device, 229, and the BCV device, 2FF, both seen here in Figure 147.



Storage Viewer\Datastores

Datastores

Identification	Status	Device	Drive Type	Capacity	Free	Type
local_licoa252	Normal	Local DELL Disk (...	Non-SSD	131.00 GB	130.05 GB	VMFS5
Shared_Administration	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.87 TB	VMFS5
VMAXe_283_Shared_licoa_251_252	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.92 TB	VMFS5
VMAXe_vCenter	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.70 TB	VMFS3

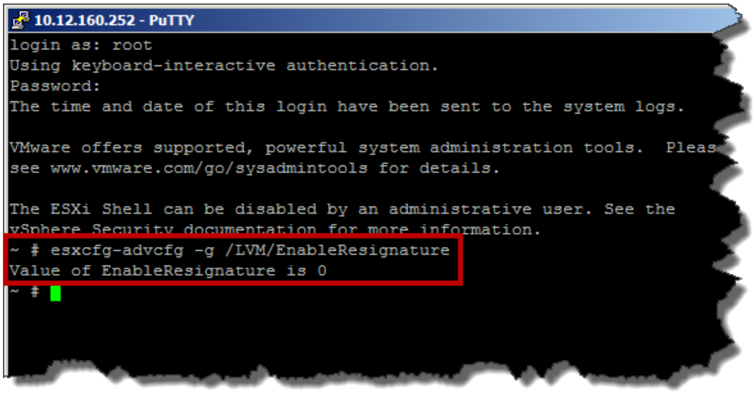
Storage Viewer\LUNs

LUNs Show all LUNs

Runtime Name	Product	Model	Revision	Array	Device ID	Type	RAID	Capacity	L..	Storage Group	Pool
vmhba2:C0:T1:L3	SYMMETRIX	VMAXe	5875	000195900286	0002B	2-Way Mir	RAID_1	5.63 MB		licoa252_gk_s...	
vmhba2:C0:T1:L29	SYMMETRIX	VMAXe	5875	000195900286	002FF	BCV+TDEV	RAID_6	1000.00 GB		dt_BCV_Clone...	SATA_Pool
vmhba2:C0:T1:L28	SYMMETRIX	VMAXe	5875	000195900286	00229	RDF2+TDEV	RAID_5	1000.00 GB		dt_R2_WP_sg	FC_Pool
vmhba2:C0:T1:L27	SYMMETRIX	VMAXe	5875	000195900286	0021F	RDF2+TDEV	RAID_5	1000.00 GB		dt_R2_WP_sg	FC_Pool
vmhba2:C0:T1:L26	SYMMETRIX	VMAXe	5875	000195900286	00215	RDF2+TDEV	RAID_5	1000.00 GB		dt_R2_WP_sg	FC_Pool
vmhba2:C0:T1:L25	SYMMETRIX	VMAXe	5875	000195900286	0020B	RDF1+TDEV	RAID_5	1000.00 GB		dt_R2_WP_sg	FC_Pool
vmhba2:C0:T1:L24	SYMMETRIX	VMAXe	5875	000195900286	001FC	TDEV	RAID_6	150.97 GB		dt_BCV_Clone...	SATA_Pool
vmhba2:C0:T1:L23	SYMMETRIX	VMAXe	5875	000195900286	001FB	TDEV	RAID_6	150.97 GB		dt_BCV_Clone...	SATA_Pool
vmhba2:C0:T1:L22	SYMMETRIX	VMAXe	5875	000195900286	001FA	TDEV	RAID_6	150.97 GB		dt_BCV_Clone...	SATA_Pool
vmhba2:C0:T1:L21	SYMMETRIX	VMAXe	5875	000195900286	001F9	TDEV	RAID_6	150.97 GB		dt_BCV_Clone...	SATA_Pool

Figure 147 LMV.enableResignature example devices

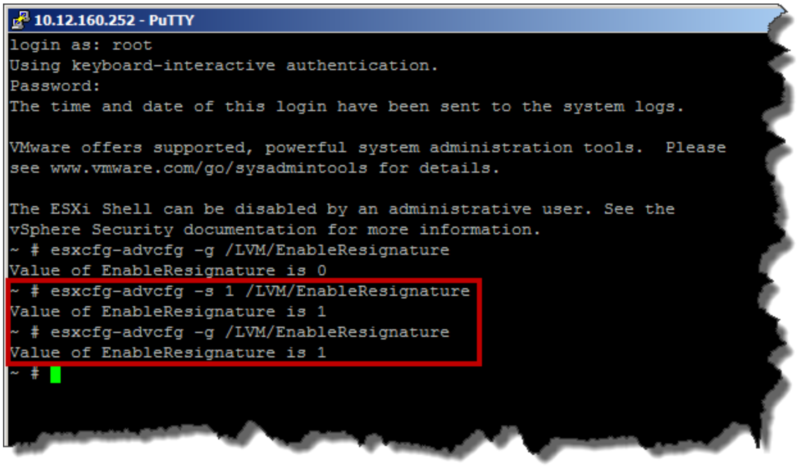
First, the parameter needs to be set. As this parameter is not available through the GUI client, the CLI will need to be used. Through local or remote CLI issue the command to check the current value of the parameter which should be ‘0’ as in Figure 148.



```
login as: root
Using keyboard-interactive authentication.
Password:
The time and date of this login have been sent to the system logs.

VMware offers supported, powerful system administration tools.  Pleas
see www.vmware.com/go/sysadmintools for details.

The ESXi Shell can be disabled by an administrative user. See the
vSphere Security documentation for more information.
~ # esxcfg-advcfg -g /LVM/EnableResignature
Value of EnableResignature is 0
~ #
```

Figure 148 **Default LVM.EnableResignature setting**

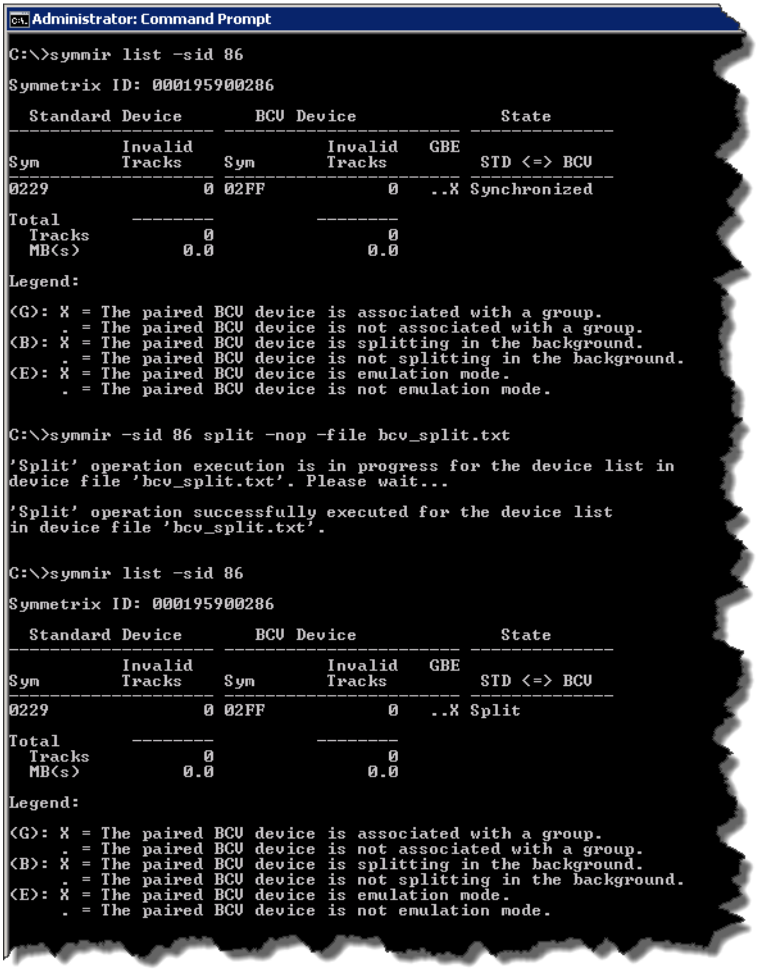
Now, change the value to ‘1’, which activates the setting. The command is `esxcfg-advcfg -s 1/LVM/EnableResignature` and seen in Figure 149.



```
login as: root
Using keyboard-interactive authentication.
Password:
The time and date of this login have been sent to the system logs.

VMware offers supported, powerful system administration tools.  Please
see www.vmware.com/go/sysadmintools for details.

The ESXi Shell can be disabled by an administrative user. See the
vSphere Security documentation for more information.
~ # esxcfg-advcfg -g /LVM/EnableResignature
Value of EnableResignature is 0
~ # esxcfg-advcfg -s 1 /LVM/EnableResignature
Value of EnableResignature is 1
~ # esxcfg-advcfg -g /LVM/EnableResignature
Value of EnableResignature is 1
~ #
```

Figure 149 **Active LVM.EnableResignature setting**

With the setting active, the BCV can be split and once the ESXi host does a rescan, the BCV with the VMFS volume from the R2 will automatically be resignatured and mounted with the prefix "snap-xxxxxxxx-". Figure 150 is a capture of the split of the BCV from the R2 device. Recall that the BCV is already masked to the ESXi host.

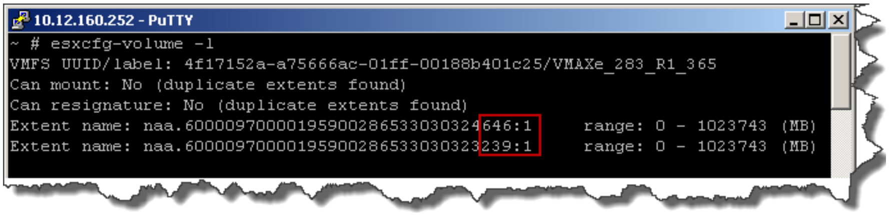


Administrator: Command Prompt

```
C:\>symmir list -sid 86

Symmetrix ID: 000195900286

  Standard Device       BCV Device                State
--------------------  -----------------------   --------------
           Invalid              Invalid   GBE
Sym        Tracks      Sym      Tracks         STD <=> BCV
--------------------  -----------------------   --------------
0229                 0 02FF              0    ..X Synchronized

Total            --------            --------
  Tracks                0                   0
  MB(s)               0.0                 0.0

Legend:

(G): X = The paired BCV device is associated with a group.
     . = The paired BCV device is not associated with a group.
(B): X = The paired BCV device is splitting in the background.
     . = The paired BCV device is not splitting in the background.
(E): X = The paired BCV device is emulation mode.
     . = The paired BCV device is not emulation mode.


C:\>symmir -sid 86 split -nop -file bcv_split.txt

'Split' operation execution is in progress for the device list in
device file 'bcv_split.txt'. Please wait...

'Split' operation successfully executed for the device list
in device file 'bcv_split.txt'.


C:\>symmir list -sid 86

Symmetrix ID: 000195900286

  Standard Device       BCV Device                State
--------------------  -----------------------   --------------
           Invalid              Invalid   GBE
Sym        Tracks      Sym      Tracks         STD <=> BCV
--------------------  -----------------------   --------------
0229                 0 02FF              0    ..X Split

Total            --------            --------
  Tracks                0                   0
  MB(s)               0.0                 0.0

Legend:

(G): X = The paired BCV device is associated with a group.
     . = The paired BCV device is not associated with a group.
(B): X = The paired BCV device is splitting in the background.
     . = The paired BCV device is not splitting in the background.
(E): X = The paired BCV device is emulation mode.
     . = The paired BCV device is not emulation mode.
```

Figure 150 Splitting the BCV from the R2

Once the BCV is split, if the `esxcfg-volume -l` command is run on the ESXi host, the host will report that there are duplicate extents as in Figure 151.

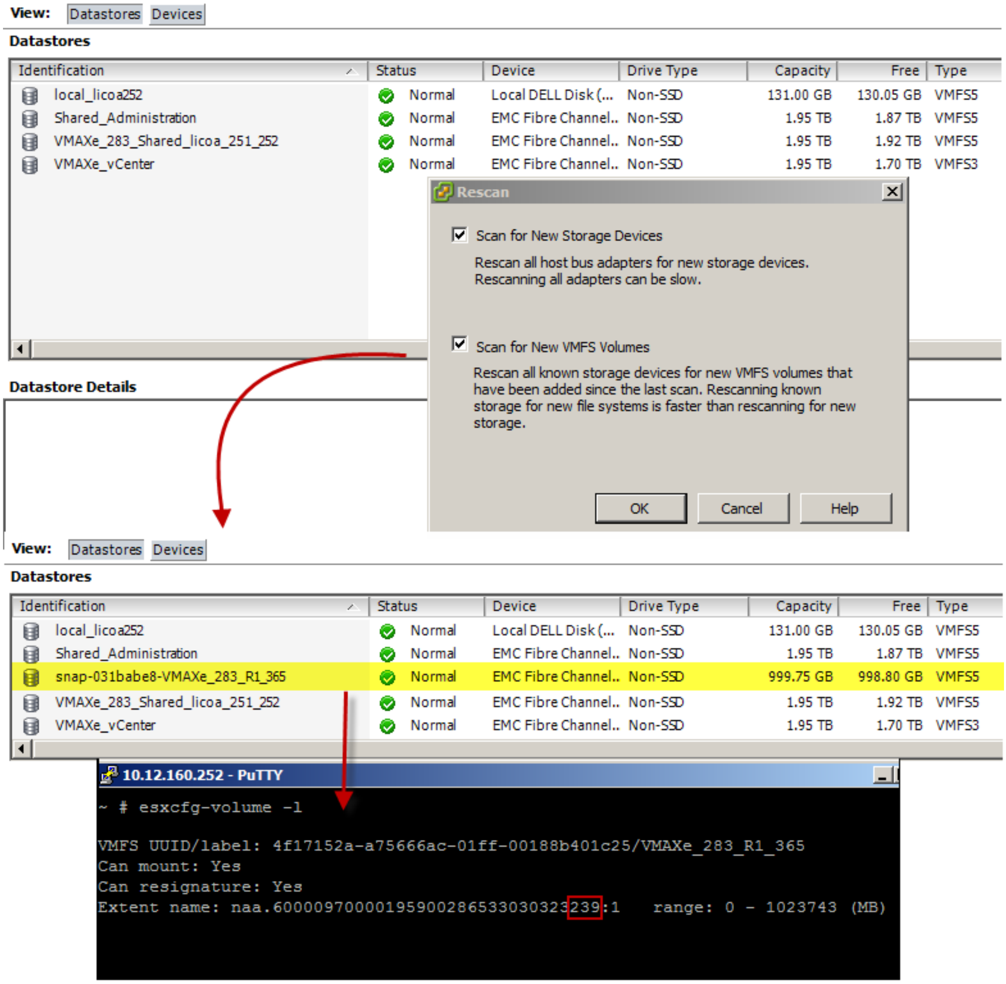


```
~ # esxcfg-volume -l
VMFS UUID/label: 4f17152a-a75666ac-01ff-00188b401c25/VMAXe_283_R1_365
Can mount: No (duplicate extents found)
Can resignature: No (duplicate extents found)
Extent name: naa.60000970000195900286533030324646:1     range: 0 - 1023743 (MB)
Extent name: naa.60000970000195900286533030323239:1     range: 0 - 1023743 (MB)
```

Figure 151 Duplicate extents post-BCV split before resignaturing

If no VMFS volume management commands have been run, which by default forces an auto-rescan[1], the user should execute a rescan either from the CLI or vSphere Client. Once the rescan completes the datastore will be mounted with the system-generated name, and ESXi will no longer report duplicate extents. This is all demonstrated in Figure 152 on page 277.

[1] This behavior can be disabled with an advanced setting if desired.



Datastores

Identification	Status	Device	Drive Type	Capacity	Free	Type
local_licoa252	Normal	Local DELL Disk (...	Non-SSD	131.00 GB	130.05 GB	VMFS5
Shared_Administration	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.87 TB	VMFS5
VMAXe_283_Shared_licoa_251_252	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.92 TB	VMFS5
VMAXe_vCenter	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.70 TB	VMFS3

Rescan

☑ Scan for New Storage Devices

Rescan all host bus adapters for new storage devices. Rescanning all adapters can be slow.

☑ Scan for New VMFS Volumes

Rescan all known storage devices for new VMFS volumes that have been added since the last scan. Rescanning known storage for new file systems is faster than rescanning for new storage.

Datastore Details

Datastores

Identification	Status	Device	Drive Type	Capacity	Free	Type
local_licoa252	Normal	Local DELL Disk (...	Non-SSD	131.00 GB	130.05 GB	VMFS5
Shared_Administration	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.87 TB	VMFS5
snap-031babe8-VMAXe_283_R1_365	Normal	EMC Fibre Channel..	Non-SSD	999.75 GB	998.80 GB	VMFS5
VMAXe_283_Shared_licoa_251_252	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.92 TB	VMFS5
VMAXe_vCenter	Normal	EMC Fibre Channel..	Non-SSD	1.95 TB	1.70 TB	VMFS3

10.12.160.252 - PuTTY

```
~ # esxcfg-volume -l

VMFS UUID/label: 4f17152a-a75666ac-01ff-00188b401c25/VMAXe_283_R1_365
Can mount: Yes
Can resignature: Yes
Extent name: naa.60000970000195900286533030323239:1    range: 0 - 1023743 (MB)
```

Figure 152 **Automated VMFS volume mounting after setting LVM.EnableResignature parameter**

IMPORTANT

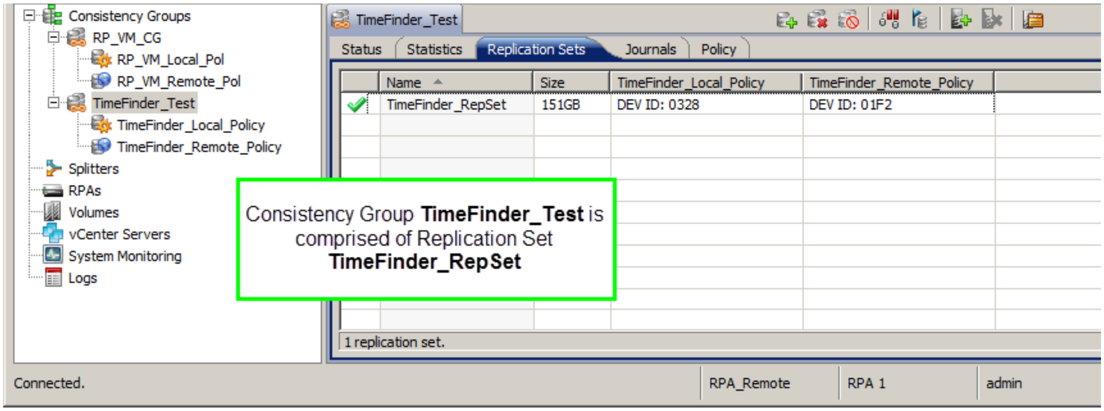
As there can be unintended consequences with this methodology, extreme care should be taken.

RecoverPoint

Creating TimeFinder copies with RecoverPoint CRR and presenting them to a VMware environment has a special requirement that differs from creating TimeFinder copies with SRDF. This involves making a RecoverPoint image accessible on the remote site before activating or splitting a TimeFinder copy. The following section will detail that requirement.

Enabling image access

RecoverPoint CRR uses the concept of consistency groups to maintain write-order consistency. A consistency group is an object that is comprised of paired volumes called replication sets.[1] A replication set is a relationship between two devices or volumes, one local and one remote. Single or multiple replication sets can exist in a consistency group. The consistency group guarantees write-order consistency over multiple volumes. The consistency group, whether made up of a single or multiple volumes, represents the "image" that is enabled or disabled. Figure 153 shows the setup used in this particular test from the RecoverPoint Management Application GUI interface. There is a single replication set being replicated which will subsequently be enabled as the "image".



	Name	Size	TimeFinder_Local_Policy	TimeFinder_Remote_Policy
✓	TimeFinder_RepSet	151GB	DEV ID: 0328	DEV ID: 01F2

1 replication set.

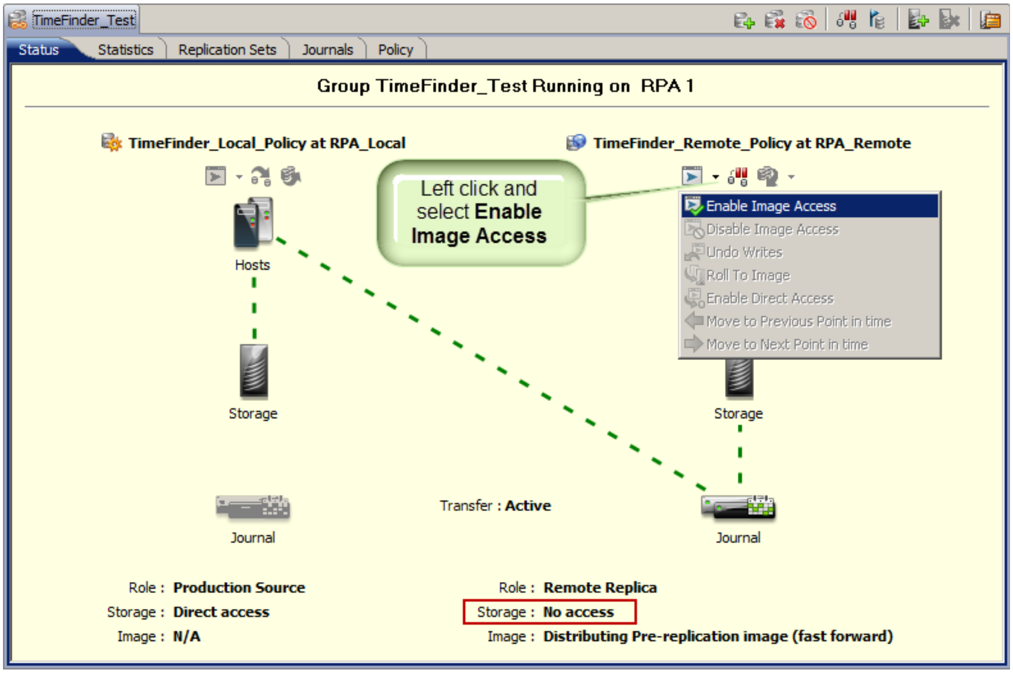
Connected. | RPA_Remote | RPA 1 | admin

Consistency Group **TimeFinder_Test** is comprised of Replication Set **TimeFinder_RepSet**

Figure 153 RecoverPoint consistency group and replication set

[1] There are additional components of consistency groups which are not mentioned as they are not essential to this discussion.

When **creating** a TimeFinder copy of the remote RecoverPoint volume, there are no special requirements. Whether using TimeFinder/Clone or TimeFinder/Mirror the TimeFinder session can be created or established. While TimeFinder copies can be created off the RecoverPoint image while it is in a “disabled” state, they should not be **activated** or **split** until the RecoverPoint image is set to “enabled” for image access. To enable image access using the RecoverPoint Management Application, left -click on the down arrow as demonstrated in Figure 154 and select “Enable Image Access” from the drop-down box.



Group TimeFinder_Test Running on RPA 1

Left click and select **Enable Image Access**

Transfer : **Active**

Role : **Production Source**
Storage : **Direct access**
Image : **N/A**

Role : **Remote Replica**
Storage : **No access**
Image : **Distributing Pre-replication image (fast forward)**

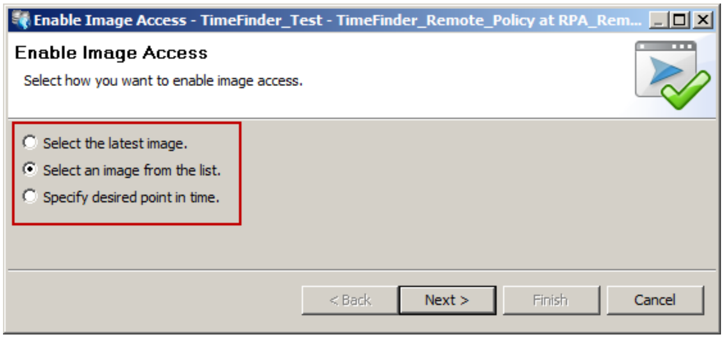
Figure 154 Enable Image Access in RecoverPoint

Selecting this brings up a dialog that provides three different options in relation to the image, seen in Figure 155 on page 280:

a. Select the latest image.

b. Select an image from the list.

c. Specify desired point in time.



Enable Image Access - TimeFinder_Test - TimeFinder_Remote_Policy at RPA_Rem...

Enable Image Access

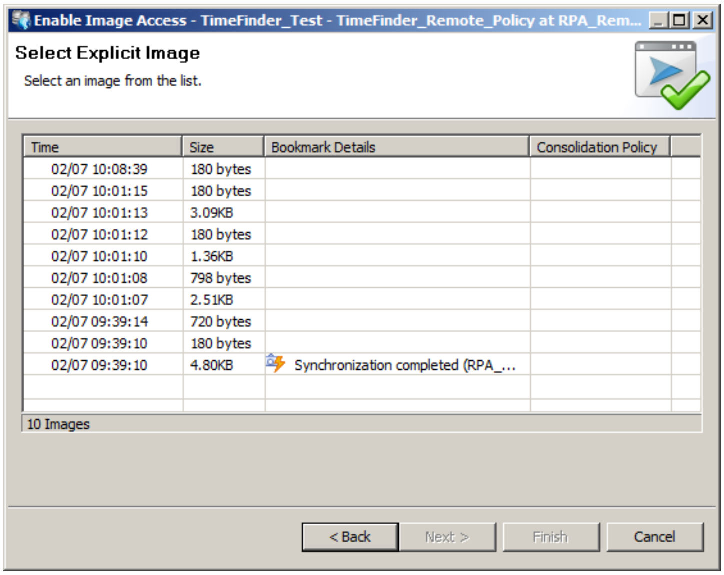
Select how you want to enable image access.

○ Select the latest image.
● Select an image from the list.
○ Specify desired point in time.

Figure 155 **Image access options**

Though any of the options will produce a viable image for TimeFinder, typically it will be sufficient to use the latest image.

If, however, as in Figure 155, the option to select from a list is chosen, the user will receive a list of images to choose from, seen in Figure 156.



Enable Image Access - TimeFinder_Test - TimeFinder_Remote_Policy at RPA_Rem...

Select Explicit Image

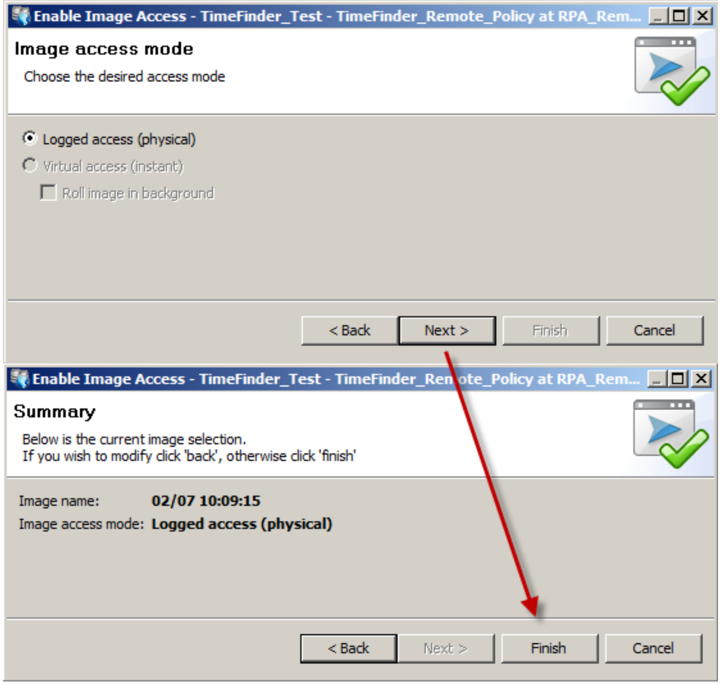
Select an image from the list.

Time	Size	Bookmark Details	Consolidation Policy
02/07 10:08:39	180 bytes		
02/07 10:01:15	180 bytes		
02/07 10:01:13	3.09KB		
02/07 10:01:12	180 bytes		
02/07 10:01:10	1.36KB		
02/07 10:01:08	798 bytes		
02/07 10:01:07	2.51KB		
02/07 09:39:14	720 bytes		
02/07 09:39:10	180 bytes		
02/07 09:39:10	4.80KB	Synchronization completed (RPA_...	

10 Images

Figure 156 **Selecting a specific RecoverPoint image**

Selecting the latest image will walk the user through two additional screens before enabling access. These are displayed together in Figure 157.



Enable Image Access - TimeFinder_Test - TimeFinder_Remote_Policy at RPA_Rem...

Image access mode

Choose the desired access mode

● Logged access (physical)

○ Virtual access (instant)

□ Roll image in background

Enable Image Access - TimeFinder_Test - TimeFinder_Remote_Policy at RPA_Rem...

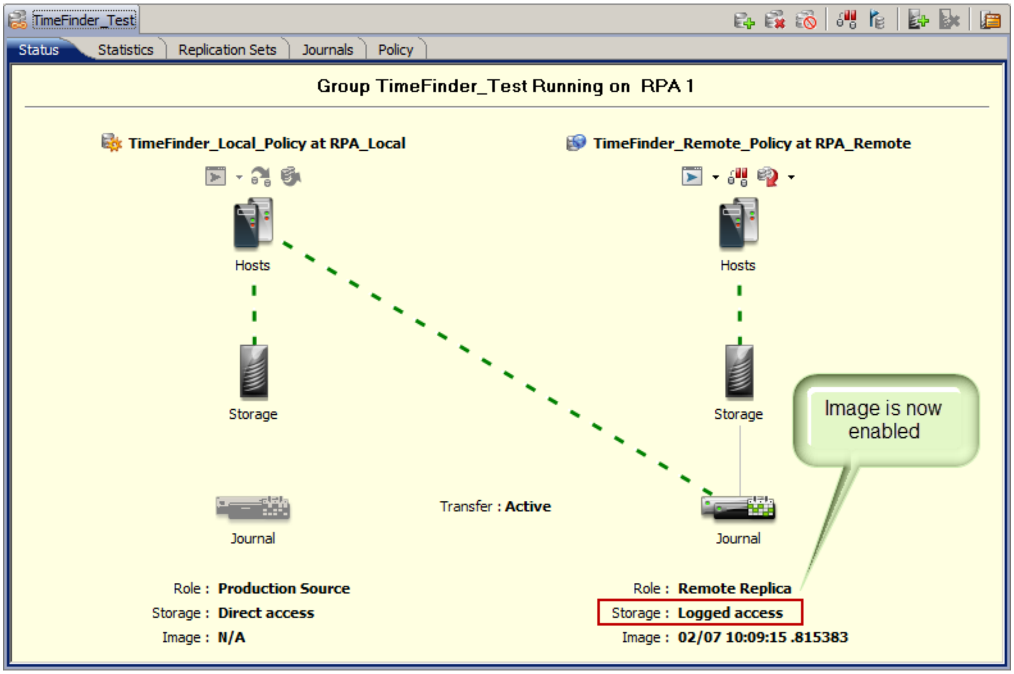
Summary

Below is the current image selection.
If you wish to modify click 'back', otherwise click 'finish'

Image name: **02/07 10:09:15**
Image access mode: **Logged access (physical)**

Figure 157 Completing image access enablement in RecoverPoint

Once “Finish” is clicked, the management GUI in Figure 158 will show that the image is now enabled.



Group TimeFinder_Test Running on RPA 1

TimeFinder_Local_Policy at RPA_Local

TimeFinder_Remote_Policy at RPA_Remote

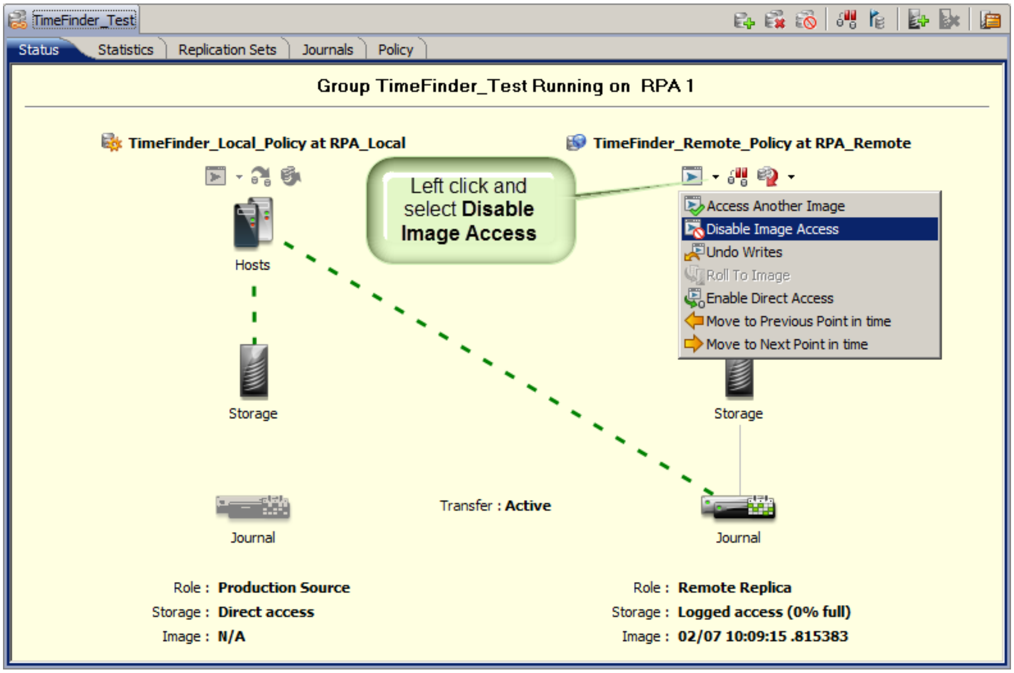
Transfer : **Active**

Role : **Production Source**
Storage : **Direct access**
Image : **N/A**

Role : **Remote Replica**
Storage : **Logged access**
Image : **02/07 10:09:15 .815383**

Figure 158 RecoverPoint image enabled

Once the image is enabled, the user can activate or split the TimeFinder copy. As image access is enabled at the consistency group level, it ensures that the TimeFinder copy is write-order consistent when it is activated or split. As soon as the TimeFinder copy is activated, the RecoverPoint image can be disabled in much the same way as it was enabled. Figure 159 on page 283 provides the GUI example of disabling the image.



Group TimeFinder_Test Running on RPA 1

Left click and select **Disable Image Access**

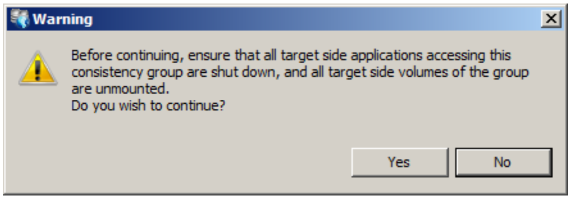
Transfer : **Active**

Role : **Production Source**
Storage : **Direct access**
Image : **N/A**

Role : **Remote Replica**
Storage : **Logged access (0% full)**
Image : **02/07 10:09:15 .815383**

Figure 159 **Disabling image access in RecoverPoint**

Before disabling the image, RecoverPoint will provide a warning dialog seen in Figure 160 to ensure that the image is not being actively used. RecoverPoint will not fail if there is active IO so it is important to check before disabling.



Before continuing, ensure that all target side applications accessing this consistency group are shut down, and all target side volumes of the group are unmounted.
Do you wish to continue?

Figure 160 **Warning dialog upon image disabling in RecoverPoint**

The TimeFinder copy can now be mounted and resignatured. Note that because the RecoverPoint image is enabled read/write at one point in the process, unlike an SRDF R2, EMC does not recommend the second methodology presented in "LVM.enableResignature parameter" on page 272 of changing the LVM.EnableResignature parameter. Though the image may only be set read/write for a short period of time, there is a chance that a rescan of an ESXi host could be run during that time and the enabled image would be resignatured and mounted. If the image was then disabled after that it would cause problems with the ESXi host. For this reason EMC recommends using the best practice of the first methodology of detaching the RecoverPoint device found in "Detaching the remote volume" on page 265.

Note: The use of Enginuity Consistency Assist (ECA) does not change the requirement for RecoverPoint image access before activating the TimeFinder copy.

6

Disaster Protection for VMware vSphere

This chapter discusses the use of EMC SRDF technology to provide disaster restart protection for VMware Infrastructure.

Introduction

The VMware virtualization platform virtualizes the x86-based physical infrastructure into a pool of resources. Virtual machines are presented with a virtual hardware environment independent of the underlying physical hardware. This enables organizations to leverage disparate physical hardware in the environment and provide low total cost of ownership. The virtualization of the physical hardware can also be used to create disaster recovery and business continuity solutions that would have been impractical otherwise.

These solutions normally involve a combination of virtual environments at one or more geographically separated data centers and EMC remote replication technology. One example of such architecture has physical servers running various business applications in their primary data center while the secondary data center has limited number of virtualized physical servers. During normal operations, the physical servers in the secondary data center are used for supporting workloads such as QA and testing. In case of a disruption in services at the primary data center, the physical servers in the secondary data center run the business applications in a virtualized environment. The focus of this chapter is a discussion of these types of solutions.

The purpose of this chapter is to discuss the following:

- VMware vCenter Site Recovery Manager
- EMC SRDF solutions
- Mounting and resignaturing of replicated VMFS volumes

SRDF Storage Replication Adapter for VMware Site Recovery Manager

VMware vCenter Site Recovery Manager (SRM) delivers advanced capabilities for disaster recovery management, nondisruptive disaster recovery testing, and automated failover. VMware vCenter SRM can manage the failover from production data centers to disaster recovery sites, as well as the failover between two sites with active workloads. SRM solves two main issues: disaster recovery automation and simple, repeatable testing of disaster recovery scenarios and processes.

VMware vCenter Site Recovery Manager helps to significantly reduce the operational complexity of a traditional DR solution by streamlining the entire process. This includes the initial setup of the solution, the maintenance of the solution and finally the execution of the solution. Without SRM, the initial configuration and maintenance complexity can be a huge demand on resources. Like any physical environment, the manual recovery of a virtual environment can be complex and error-prone — leading to unwelcome, additional downtime. SRM avoids this through the automation and simplicity provided by its highly customizable recovery plans and virtual machine protection groups.

Tested, proven, and documented procedures are imperative for a disaster recovery solution. VMware vCenter Site Recovery Manager offers the ability to create operational procedures that allow for periodic, automated, repeatable and, most importantly, non-disruptive execution of disaster recovery tests. This helps to avoid/reduce costly downtime to business critical application caused by traditional disaster recovery testing while helping to ensure that actual disaster recoveries execute as effortlessly as possible.

Site Recovery Manager 5 new workflow features

VMware vCenter Site Recovery Manager 5 includes six workflow capabilities, each discussed in this section:

- "Test Recovery"
- "Cleanup"
- "Planned migration"
- "Reprotection"

- "Failback"
- "Disaster recovery event"

Test Recovery

In order to assure confidence in a recovery plan that has been configured for failover it is important to test that recovery plan to make sure all parts are working as expected. The test recovery operation provided by VMware SRM allows administrators to create recovery plans and realistically test them without having to execute an actual failover. VMware SRM in conjunction with a SRA, allows recovery plans to be executed in test mode and recovers the virtual machines on the recovery site in an isolated environment where they cannot interfere with production applications. The virtual machines are not recovered on the remote site R2 device but instead on a local copy of that R2 device ensuring that replication and remote protection is not affected during the test.

Cleanup

The cleanup operation simply reverts the environment after a test recovery operation has been executed. Once the test recovery is complete and all aspects of the plan have been approved and verified the administrator can revert the environment to its previous state quickly and easily by initiating the cleanup operation workflow.

Planned migration

Prior to Site Recovery Manager 5, workflow capabilities included both execution and testing of a recovery plan. With version 5, VMware has introduced a new workflow designed to deliver migration from a protected site to a recovery site through execution of a planned migration workflow. Planned migration ensures an orderly and pretested transition from a protected site to a recovery site while minimizing the risk of data loss. The migration runs in a highly controlled fashion. It will halt the workflow if an error occurs, which differs from a traditional disaster-event recovery plan.

With Site Recovery Manager 5, all recovery plans, whether they are for migration or recovery, run as part of a planned workflow that ensures that systems are properly shut down and that data is synchronized with the recovery site prior to migration of the workloads. This ensures that systems are properly quiesced and that all data changes have been completely replicated prior to starting the virtual machines at the recovery site. If, however, an error is encountered during the recovery plan execution, planned migration will stop the workflow, providing an opportunity to fix the problem that caused the error before attempting to continue.

Reprotection

After a recovery plan or planned migration has run, there are often cases where the environment must continue to be protected against failure, to ensure its resilience or to meet objectives for disaster recovery. With Site Recovery Manager 5, reprotection is a new extension to recovery plans for use only with array-based replication. It enables the environment at the recovery site to establish synchronized replication and protection of the environment back to the original protected site.

After failover to the recovery site, selecting to reprotect the environment will establish synchronization and attempt to replicate data between the protection groups now running at the recovery site and the previously protected, primary site. This capability to reprotect an environment ensures that environments are protected against failure even after a site recovery scenario. It also enables automated failback to a primary site following a migration or failover.

Failback

An automated failback workflow can be run to return the entire environment to the primary site from the secondary site. This will happen after reprotection has ensured that data replication and synchronization have been established to the original site.

Failback will run the same workflow that was used to migrate the environment to the protected site. It will guarantee that the critical systems encapsulated by the recovery plan are returned to their original environment. The workflow will execute only if reprotection has successfully completed.

Failback ensures the following:

- All virtual machines that were initially migrated to the recovery site will be moved back to the primary site.
- Environments that require that disaster recovery testing be done with live environments with genuine migrations can be returned to their initial site.
- Simplified recovery processes will enable a return to standard operations after a failure.

- Failover can be done in case of disaster or in case of planned migration.

Disaster recovery event

As previously noted, all recovery plans in Site Recovery Manager 5 now include an initial attempt to synchronize data between the protection and recovery sites, even during a disaster recovery scenario.

During a disaster recovery event, an initial attempt will be made to shut down the protection group’s virtual machines and establish a final synchronization between sites. This is designed to ensure that virtual machines are static and quiescent before running the recovery plan, to minimize data loss where possible during a disaster. If the protected site no longer is available, the recovery plan will continue to execute and will run to completion even if errors are encountered.

This new attribute minimizes the possibility of data loss while still enabling disaster recovery to continue, balancing the requirement for virtual machine consistency with the ability to achieve aggressive recovery-point objectives.

SRDF Storage Replication Adapter

VMware vCenter Site Recovery Manager leverages storage array-based replication such as EMC Symmetrix Remote Data Facility (SRDF) to protect virtual machines in VMware vCenter vSphere environments. The interaction between VMware vCenter Site Recovery Manager and the storage array replication is managed through a well-defined set of specifications. The VMware-defined specifications are implemented by the storage array vendor as a lightweight application referred to as the storage replication adapter or SRA.

EMC SRDF Adapter is an SRA that enables VMware vCenter Site Recovery Manager to interact with an EMC Symmetrix storage environment. Each version of SRM requires a certain version of the SRDF SRA. See Table 6 on page 291 for details.

Table 6 **SRDF SRA interoperability**

SRM version	SRA version	Solutions Enabler version	Enginuity version
4.0	2.x	7.01 and later	5671, 5771, 5773, 5874, 5875
4.1	2.x	7.01 and later	5671, 5771, 5773, 5874, 5875
5.0	5.0.1.0	7.3.1 and later	5671, 5771, 5773, 5874, 5875, 5876
5.1	5.1.0.0	7.5 and later	5874, 5875, 5876[a]

[a] Note that DMX support is dropped in the SRDF SRA version 5.1

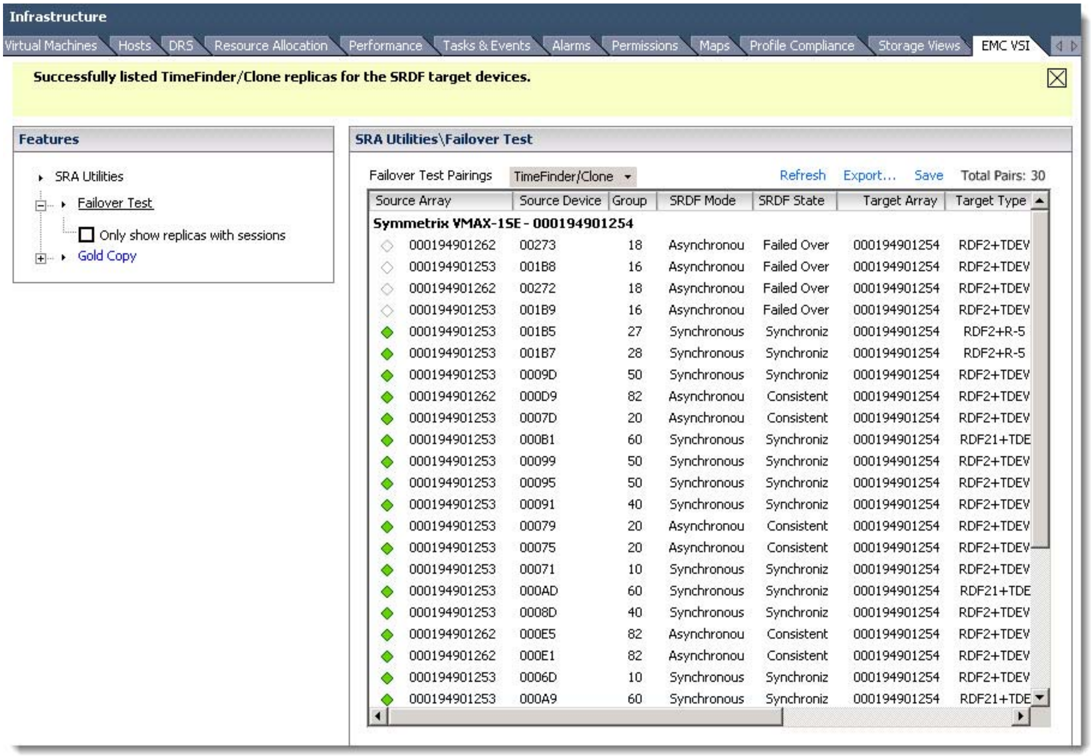
The EMC SRDF Adapter for VMware vCenter Site Recovery Manager supports a variety of SRDF modes and each version has expanded support. Table 7 shows the supported SRDF modes for each SRDF SRA version.

Table 7 **Supported SRDF Modes**

SRDF SRA Version	Supported 2-site SRDF	Supported 3-site SRDF
2.x	Asynchronous, Synchronous	N/A
5.0.x	Asynchronous, Synchronous	Concurrent SRDF/Star, Cascaded SRDF/Star
5.1.x	Asynchronous, Synchronous	Concurrent SRDF/Star, Cascaded SRDF/Star, Concurrent SRDF, Cascaded SRDF

The EMC SRDF Adapter for VMware vCenter Site Recovery Manager, through the use of consistency groups, can also address configurations in which the data for a group of related virtual machines is spread across multiple RDF groups.

The EMC Virtual Storage Integrator (VSI) provides a vSphere Client integrated GUI interface to complement the SRM feature set and to configure and customize the adapter's behavior to suit any environment. Figure 161 displays an example screen of the VSI that allows for configuration of test recovery TimeFinder pairings.



Successfully listed TimeFinder/Clone replicas for the SRDF target devices.

□ Only show replicas with sessions

SRA Utilities\Failover Test

Failover Test Pairings: TimeFinder/Clone

Total Pairs: 30

Symmetrix VMAX-1SE - 000194901254

Source Array	Source Device	Group	SRDF Mode	SRDF State	Target Array	Target Type
◇ 000194901262	00273	18	Asynchronou	Failed Over	000194901254	RDF2+TDEV
◇ 000194901253	001B8	16	Asynchronou	Failed Over	000194901254	RDF2+TDEV
◇ 000194901262	00272	18	Asynchronou	Failed Over	000194901254	RDF2+TDEV
◇ 000194901253	001B9	16	Asynchronou	Failed Over	000194901254	RDF2+TDEV
◆ 000194901253	001B5	27	Synchronous	Synchroniz	000194901254	RDF2+R-5
◆ 000194901253	001B7	28	Synchronous	Synchroniz	000194901254	RDF2+R-5
◆ 000194901253	0009D	50	Synchronous	Synchroniz	000194901254	RDF2+TDEV
◆ 000194901262	000D9	82	Asynchronou	Consistent	000194901254	RDF2+TDEV
◆ 000194901253	0007D	20	Asynchronou	Consistent	000194901254	RDF2+TDEV
◆ 000194901253	000B1	60	Synchronous	Synchroniz	000194901254	RDF21+TDE
◆ 000194901253	00099	50	Synchronous	Synchroniz	000194901254	RDF2+TDEV
◆ 000194901253	00095	50	Synchronous	Synchroniz	000194901254	RDF2+TDEV
◆ 000194901253	00091	40	Synchronous	Synchroniz	000194901254	RDF2+TDEV
◆ 000194901253	00079	20	Asynchronou	Consistent	000194901254	RDF2+TDEV
◆ 000194901253	00075	20	Asynchronou	Consistent	000194901254	RDF2+TDEV
◆ 000194901253	00071	10	Synchronous	Synchroniz	000194901254	RDF2+TDEV
◆ 000194901253	000AD	60	Synchronous	Synchroniz	000194901254	RDF21+TDE
◆ 000194901253	0008D	40	Synchronous	Synchroniz	000194901254	RDF2+TDEV
◆ 000194901262	000E5	82	Asynchronou	Consistent	000194901254	RDF2+TDEV
◆ 000194901262	000E1	82	Asynchronou	Consistent	000194901254	RDF2+TDEV
◆ 000194901253	0006D	10	Synchronous	Synchroniz	000194901254	RDF2+TDEV
◆ 000194901253	000A9	60	Synchronous	Synchroniz	000194901254	RDF21+TDE

Figure 161 **EMC Virtual Storage Integrator Symmetrix SRA Utilities**

Table 8 on page 293 shows the supported methods of test failover with each release of the SRDF SRA. This includes test failover with TimeFinder replicas or without TimeFinder. The latter method is a mode that splits the RDF link temporarily and recovers the test environment directly from the R2 devices.

Table 8 **Supported test failover modes**

SRDF SRA Version	TimeFinder support for test failover operations[a]	Test failover without TimeFinder support (direct off of R2 devices)
2.x	Mirror, Clone, Snap	Supported with two site SRDF
5.0.x	Mirror, Clone, Snap	Supported with two site SRDF[b]
5.1.x	Clone, VP Snap, Snap	Supported with two site SRDF and three site SRDF

a.The same TimeFinder methods are supported for gold copy creation during recovery operations

b.The 5.0.0.7 release of the SRDF SRA does not support test failover without TimeFinder; this was introduced in the 5.0.1.0 release.

For more detailed information on using SRM with the EMC SRDF SRA as well as the Virtual Storage Integrator interface, see the *Using EMC SRDF Adapter for VMware vCenter Site Recovery Manager TechBook* on Powerlink.EMC.com.

Business continuity solutions for VMware vSphere

Business continuity solutions for a production environment with the VMware virtualization platform fortunately is not too complicated. In addition to, or in place of, a tape-based disaster recovery solution, EMC SRDF can be used as the mechanism to replicate data from the production data center to the remote data center. The copy of the data in the remote data center can be presented to a VMware ESXi cluster. The vSphere environment at the remote data center thus provides a business continuity solution.

Recoverable vs. restartable copies of data

The Symmetrix-based replication technologies can generate a restartable or recoverable copy of the data. The difference between the two types of copies can be confusing; a clear understanding of the differences between the two is critical to ensure that the recovery goals for a vSphere environment can be met.

Recoverable disk copies

A recoverable copy of the data is one in which the application (if it supports it) can apply logs and roll the data forward to a point in time after the copy was created. The recoverable copy is most relevant in the database realm where database administrators use it frequently to create backup copies of database. In the event of a failure to the database, the ability to recover the database not only to a point in time when the last backup was taken, but also to roll-forward subsequent transactions up to the point of failure, is critical to most business applications. Without that capability, in an event of a failure, there will be an unacceptable loss of all transactions that occurred since the last backup.

Creating recoverable images of applications running inside virtual machines using EMC replication technology requires that the application or the virtual machine be shut down when it is copied. A recoverable copy of an application can also be created if the application supports a mechanism to suspend writes when the copy of the data is created. Most database vendors provide functionality in their RDBMS engine to suspend writes. This functionality has to be invoked inside the virtual machine when EMC technology is deployed to ensure a recoverable copy of the data is generated on the target devices.

Restartable disk copies

If a copy of a running virtual machine is created using EMC Consistency technology[1] without any action inside the virtual machines, the copy is normally a restartable image of the virtual machine. This means that when the data is used on cloned virtual machines, the operating system and/or the application goes into crash recovery. The exact implications of crash recovery in a virtual machine depend on the application that the machine supports:

- If the source virtual machine is a file server or runs an application that uses flat files, the operating system performs a file-system check and fixes any inconsistencies in the file system. Modern file systems such as Microsoft NTFS use journals to accelerate the process
- When the virtual machine runs any database or application with a log-based recovery mechanism, the application uses the transaction logs to bring the database or application to a point of consistency. The process deployed varies depending on the database or application, and is beyond the scope of this document.

Most applications and databases cannot perform roll-forward recovery from a restartable copy of the data. Therefore, in most cases a restartable copy of data created from a virtual machine that is running a database engine is inappropriate for performing backups. However, applications that use flat files or virtual machines that act as file servers can be backed up from a restartable copy of the data. This is possible since none of the file systems provide a roll-forward logging mechanism that enables recovery.

SRDF/S

Synchronous SRDF (SRDF/S) is a method of replicating production data changes from locations less than 200 km apart. Synchronous replication takes writes that are inbound to the source Symmetrix and copies them to the target Symmetrix. The resources of the storage arrays are exclusively used for the copy. The write operation from the virtual machine is not acknowledged back to the host until both

[1] See “Copying running virtual machines using EMC Consistency technology” on page 233 for more detail.

Symmetrix arrays have a copy of the data in their cache. Readers should consult Powerlink.EMC.com for further information about SRDF/S.

Figure 162 is a schematic representation of the business continuity solution that integrates a VMware environment and SRDF technology. The solution shows two virtual machines accessing devices on the Symmetrix storage arrays as RDMs.

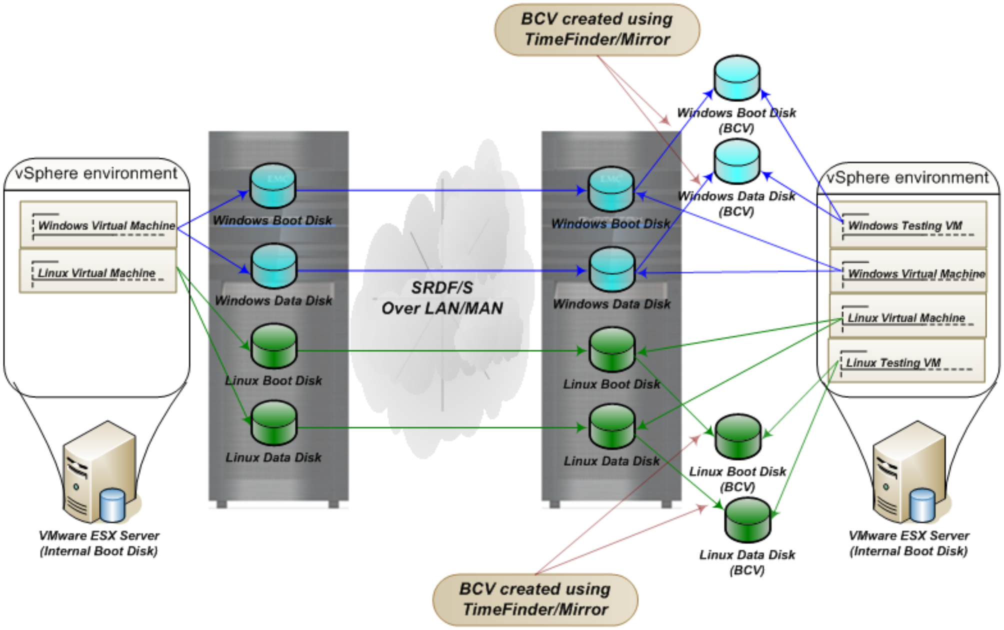


Figure 162 Business continuity solution using SRDF/S in a VMware environment using RDM

An equivalent solution utilizing the VMFS is depicted in Figure 163 on page 297. The proposed solution provides an excellent opportunity to consolidate the vSphere environment at the remote site. It is possible to run virtual machines on any VMware ESX host in the cluster. This capability also allows the consolidation of the production VMware ESX hosts to fewer VMware ESX hosts at the recovery site. However, by doing so, there is a potential for duplicate virtual machine IDs when multiple virtual machines are consolidated in the remote site. If this occurs, the virtual machine IDs can be easily changed at the remote site.

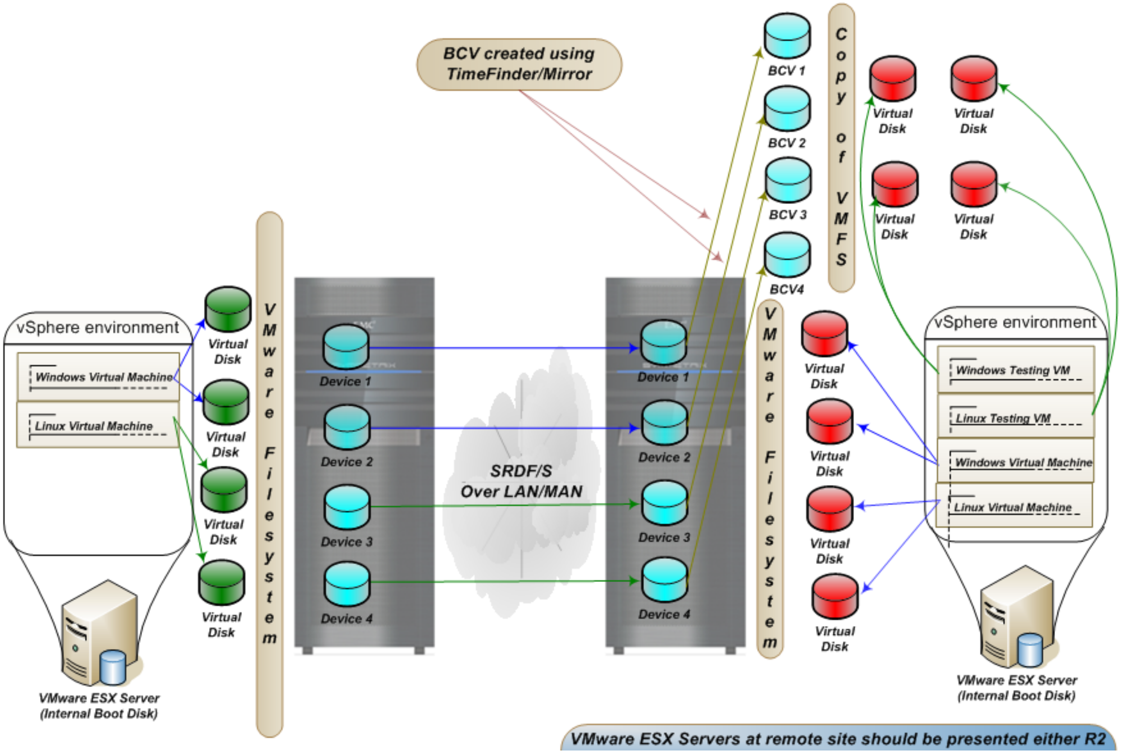


Figure 163 **Business continuity solution using SRDF/S in a VMware environment with VMFS**

SRDF/A

SRDF/A, or asynchronous SRDF, is a method of replicating production data changes from one Symmetrix to another using delta set technology. Delta sets are the collection of changed blocks grouped together by a time interval that can be configured at the source site. The default time interval is 30 seconds. The delta sets are then transmitted from the source site to the target site in the order they were created. SRDF/A preserves the dependent-write consistency of the database at all times at the remote site. Further details about SRDF/A can be obtained on Powerlink.EMC.com. The SRDF/A TechBook can also be consulted for further information.

The distance between the source and target Symmetrix is unlimited and there is no host impact. Writes are acknowledged immediately when they hit the cache of the source Symmetrix. Figure 164 shows the SRDF/A process as applied to a VMware environment using RDM.

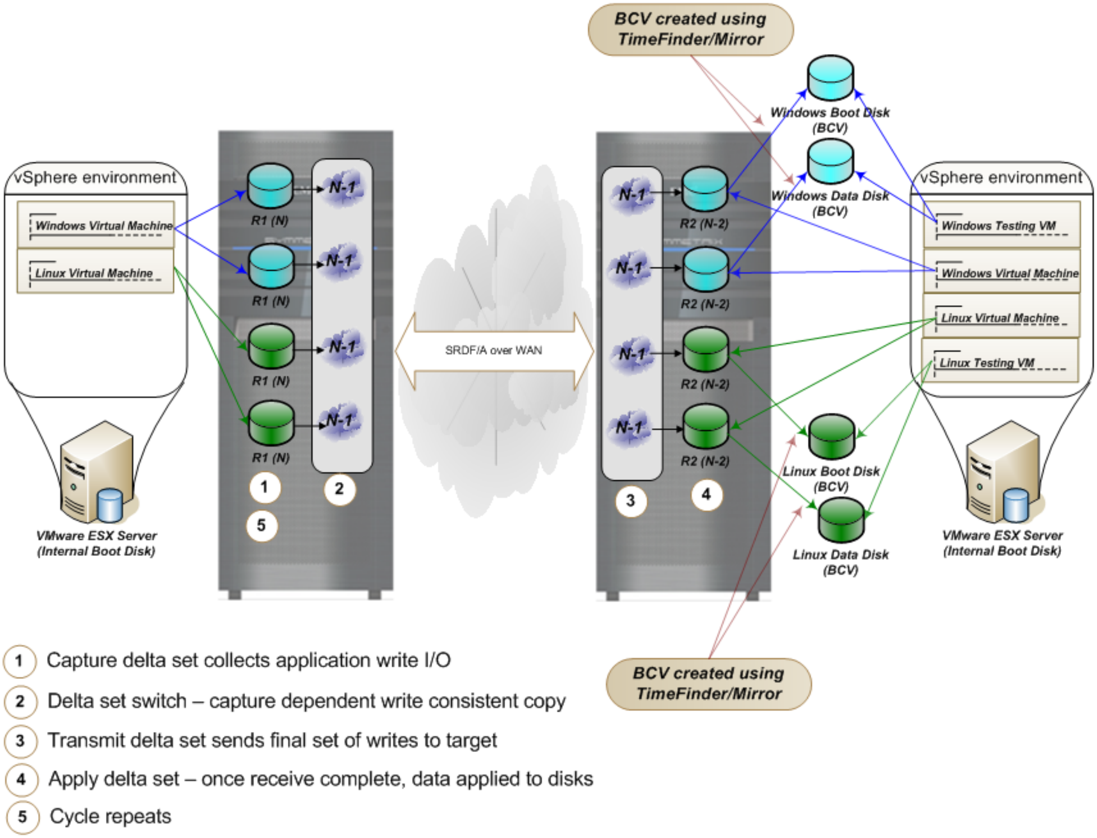

BCV created using TimeFinder/Mirror
Windows Boot Disk (BCV)
Windows Data Disk (BCV)
vSphere environment
Windows Virtual Machine
Linux Virtual Machine
R1 (N)
R1 (N)
R1 (N)
R1 (N)
N-1
N-1
N-1
N-1
SRDF/A over WAN
R2 (N-2)
R2 (N-2)
R2 (N-2)
R2 (N-2)
vSphere environment
Windows Testing VM
Windows Virtual Machine
Linux Virtual Machine
Linux Testing VM
Linux Boot Disk (BCV)
Linux Data Disk (BCV)
VMware ESX Server (Internal Boot Disk)
VMware ESX Server (Internal Boot Disk)
1
2
5
3
4
BCV created using TimeFinder/Mirror
1 Capture delta set collects application write I/O
2 Delta set switch – capture dependent write consistent copy
3 Transmit delta set sends final set of writes to target
4 Apply delta set – once receive complete, data applied to disks
5 Cycle repeats

Figure 164 Business continuity solution using SRDF/A in a VMware environment with RDM

A similar process, as shown in Figure 165, can also be used to replicate disks containing VMFS.

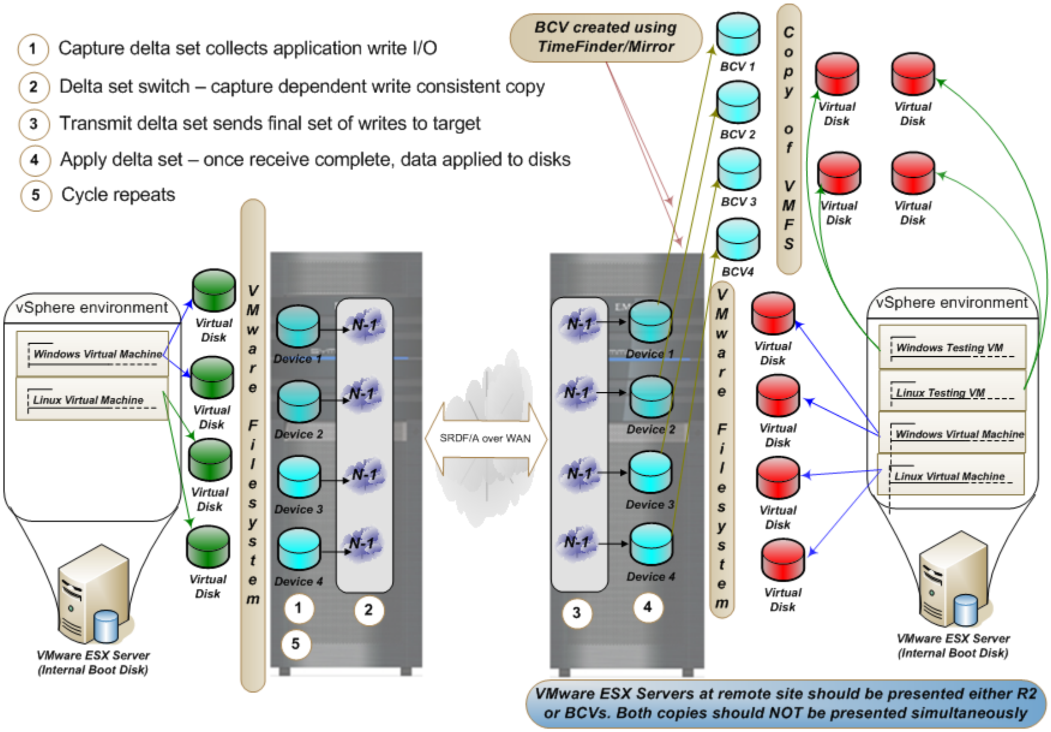

1 Capture delta set collects application write I/O
2 Delta set switch – capture dependent write consistent copy
3 Transmit delta set sends final set of writes to target
4 Apply delta set – once receive complete, data applied to disks
5 Cycle repeats
BCV created using TimeFinder/Mirror
BCV 1
BCV 2
BCV 3
BCV4
Copy of VMFS
Virtual Disk
vSphere environment
Windows Virtual Machine
Linux Virtual Machine
Windows Testing VM
Linux Testing VM
VMware File system
Device 1
Device 2
Device 3
Device 4
N-1
SRDF/A over WAN
VMware ESX Server (Internal Boot Disk)
1
2
3
4
5
VMware ESX Servers at remote site should be presented either R2 or BCVs. Both copies should NOT be presented simultaneously

Figure 165 Business continuity solution using SRDF/A in a VMware environment with VMFS

Before the asynchronous mode of SRDF can be established, initial copy of the production data has to occur. In other words, a baseline full copy of all the volumes that are going to participate in the asynchronous replication must be executed first. This is usually accomplished using the adaptive-copy mode of SRDF.

SRDF Automated Replication

SRDF Automated Replication, or SRDF/AR, is a continuous movement of dependent-write consistent data to a remote site using SRDF adaptive copy mode and TimeFinder consistent split technology. TimeFinder copies are used to create a dependent-write consistent point-in-time image of the data to be replicated. The BCVs also have a R1 personality, which means that SRDF in adaptive-copy mode can be used to replicate the data from the BCVs to the target site.

Since the BCVs are not changing, replication completes in a finite length of time. The time required for replication depends on the size of the network pipe between the two locations and the amount of data to copy. On the remote Symmetrix, another BCV copy of the data is made using data on the R2s. This is necessary because of the behavior of SRDF adaptive mode. SRDF adaptive copy mode does not maintain write-order when moving data. The copy of data on the R2s is inconsistent until all of the tracks owed from the BCVs are copied. Without a BCV at the remote site, if a disaster were to occur while the R2s were synchronizing with BCV/R1s, no valid copy of the data would be available at the DR site. The BCV copy of the data in the remote Symmetrix is commonly called the gold copy of the data. The whole process then repeats.

With SRDF/AR, there is no host impact. Writes are acknowledged immediately when they hit the cache of the source Symmetrix. A solution using SRDF/AR to protect a VMware environment is depicted in Figure 166 on page 301.

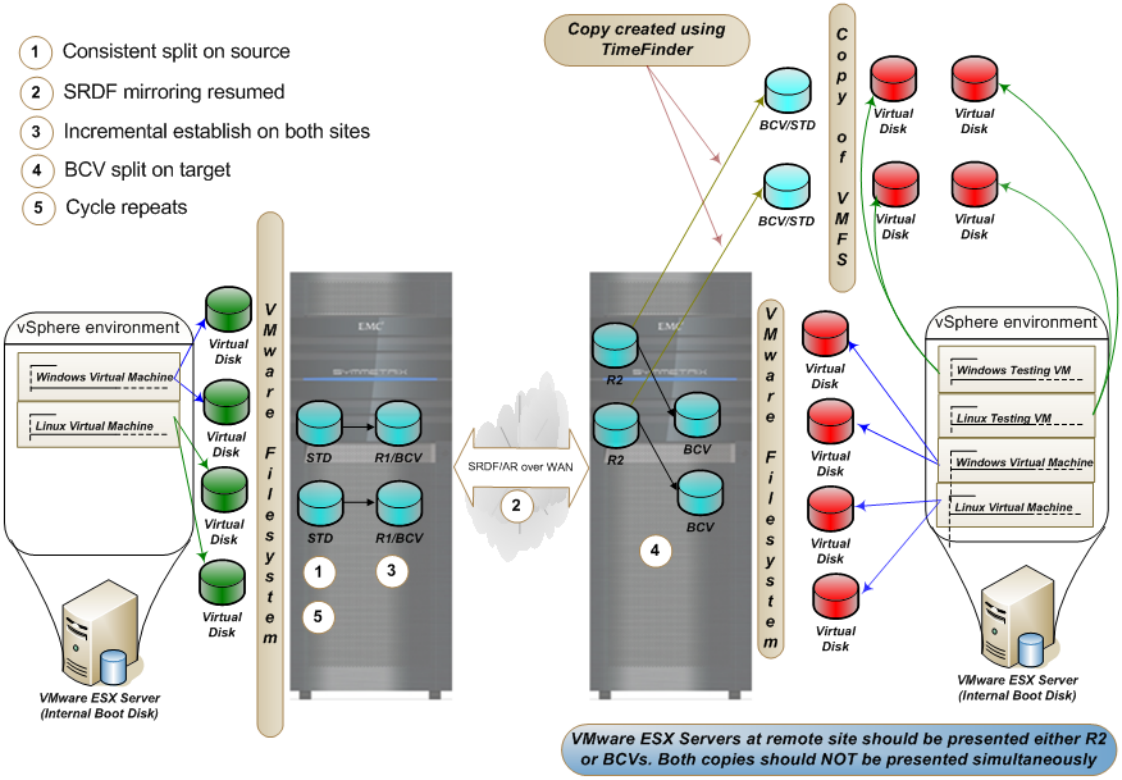


Figure 166 Business continuity solution using SRDF/AR in a VMware environment with VMFS

Cycle times for SRDF/AR are usually in the minutes-to-hours range. The RPO is double the cycle time in a worst-case scenario. This may be a good fit for customers who do not require RPOs measured in seconds but also cannot afford to have RPOs measured in days.

An added benefit of having a longer cycle time is the increased probability of a track being updated more than once in an hour interval than in, for instance, a 30-second interval. Since this multi-changed track will only be sent once in that hour, rather than multiple times if the interval were 30 seconds, there are reduced bandwidth requirements for the SRDF/AR solution.

A variation of the SRDF/AR solution presented in Figure 166 is the SRDF/AR multi-hop solution. The variation enables long-distance replication with zero seconds of data loss. This is achieved through the use of a bunker Symmetrix. Production data is replicated synchronously to the bunker Symmetrix, which is within 200 km of the production Symmetrix. The distance between the production and bunker Symmetrix is far enough to avoid potential disasters at the primary site, but is also close enough to operate SRDF in synchronous mode. The data on the R2s at the bunker site is replicated to another Symmetrix located thousands of miles away using the process shown in Figure 166 on page 301.

Readers should consult the *Symmetrix Remote Data Facility (SRDF) Product Guide* on www.Powerlink.EMC.com for more detailed discussions on SRDF/AR technology.

SRDF/Star overview

The SRDF/Star disaster recovery solution provides advanced multi-site business continuity protection for enterprise environments. It combines the power of Symmetrix Remote Data Facility (SRDF) synchronous and asynchronous replication, enabling the most advanced three-site business continuance solution available today.

SRDF/Star enables concurrent SRDF/S and SRDF/A operations from the same source volumes with the ability to incrementally establish an SRDF/A session between the two remote sites in the event of a primary site outage — a capability only available through SRDF/Star software.

This capability takes the promise of concurrent synchronous and asynchronous operations (from the same source device) to its logical conclusion. SRDF/Star allows you to quickly re-establish protection between the two remote sites in the event of a primary site failure, and then just as quickly restore the primary site when conditions permit.

With SRDF/Star, enterprises can quickly resynchronize the SRDF/S and SRDF/A copies by replicating only the differences between the sessions-allowing for much faster resumption of protected services after a source site failure.

Concurrent SRDF

Concurrent SRDF allows the same source data to be copied concurrently to Symmetrix arrays at two remote locations. As Figure 167 shows, the capability of a concurrent R1 device to have one of its links synchronous and the other asynchronous is supported as an SRDF/Star topology. Additionally, SRDF/Star allows the reconfiguration between concurrent and cascaded modes dynamically.

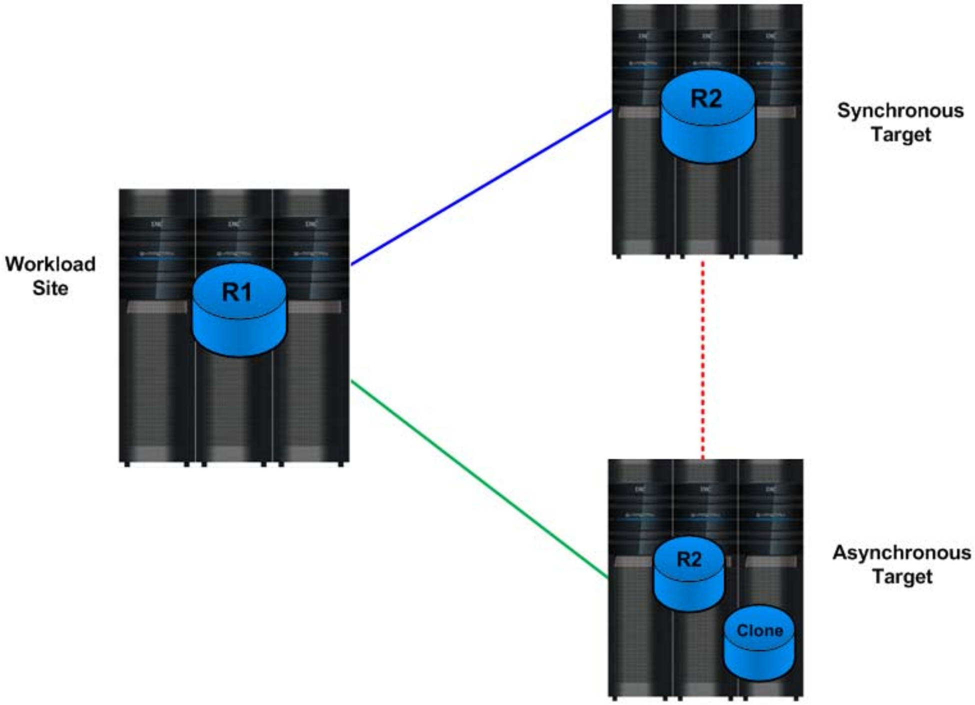


Figure 167 Concurrent SRDF

Cascaded SRDF

Introduced with Enginuity 5773, Cascaded SRDF allows a device to be both a synchronous target (R2) and an asynchronous source (R1) creating an R21 device type. SRDF/Star supports the cascaded topology and allows the dynamic reconfiguration between cascaded and concurrent modes. See Figure 168 for a representation of the configuration.

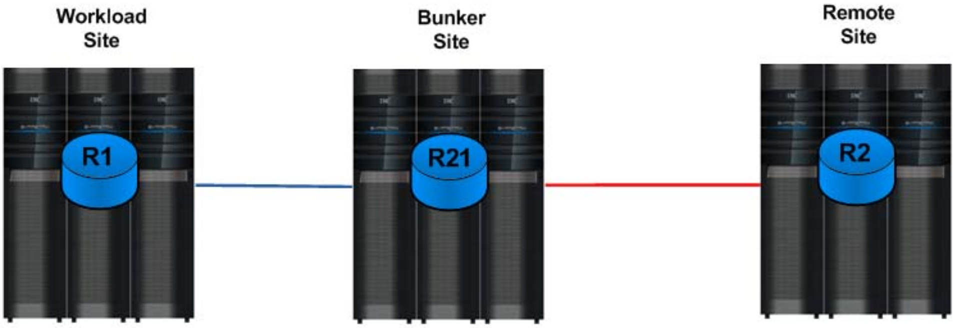


Figure 168 Cascaded SRDF

Extended Distance Protection

The SRDF/Extended Distance Protection (EDP) functionality is a licensed SRDF feature that offers a long distance disaster recovery (DR) solution.

Figure 169 on page 305 shows an SRDF/EDP configuration. The SRDF/EDP is a three-site configuration that requires Enginuity version 5874 or higher running on Symmetrix B and Enginuity version 5773 or higher running on Symmetrix A and Symmetrix C.

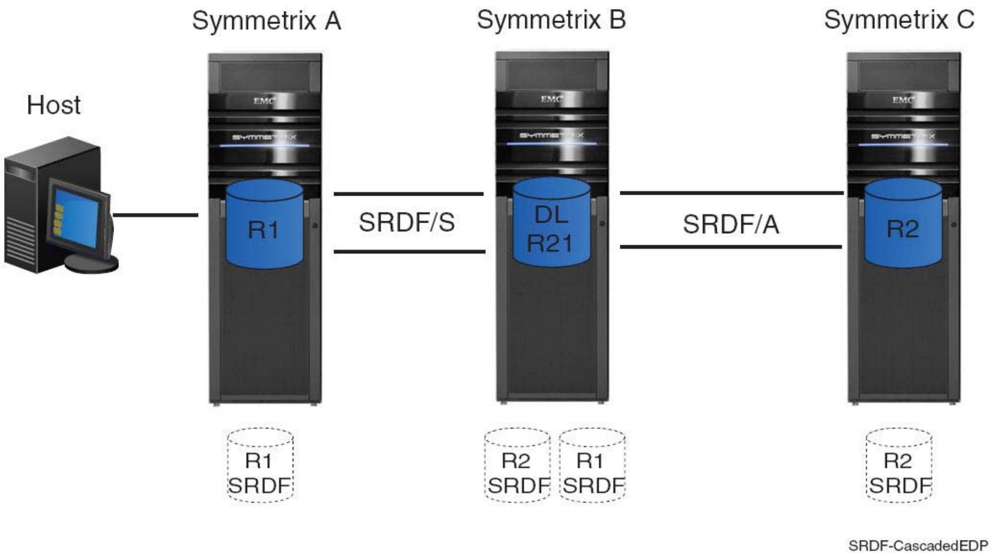


Figure 169 SRDF/EDP basic configuration

This is achieved through a Cascaded SRDF setup, where a Symmetrix VMAX system at a secondary site uses DL R21 devices to capture only the differential data that would be owed to the tertiary site in the event of a primary site failure.

It is the data on the diskless R21 devices that helps these configurations achieve a zero Recovery Point Objective (RPO).

Diskless R21 devices (DL R21) operate like R21 volumes in Cascaded SRDF except that DL R21 devices do not store full copies of data and have no local mirrors (RAID groups) configured to them. Unlike ordinary R21 volumes, DL R21 devices are not Symmetrix logical volumes. The DL R21 devices must be preconfigured within the Symmetrix system prior to being placed in an SRDF relationship.

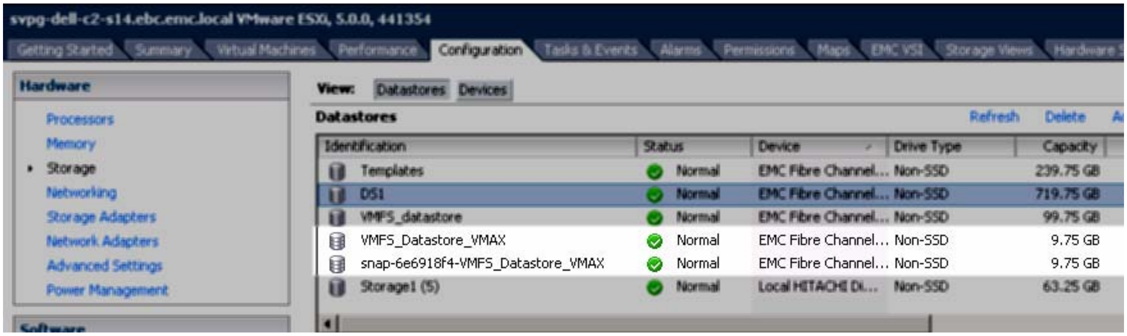
Configuring remote site virtual machines on replicated VMFS volumes

The process of creating matching virtual machines at the remote site is similar to the process that was presented for cloning virtual machines in VMware environments in Chapter 5, "Cloning of vSphere Virtual Machines."

Inherent in ESX 4 and 5 is the ability to handle replicas of source volumes. Since storage array technologies create exact replicas of the source volumes, all information including the unique VMFS signature (and label, if applicable) is replicated. If a copy of a VMFS volume is presented to any VMware ESX 4 or 5 cluster, the VMware ESX automatically masks the copy. The device that holds the copy is determined by comparing the signature stored on the device with the computed signature. R2s, for example, have different unique IDs from the R1 devices with which it is associated. Therefore, the computed signature for a R2 device differs from the one stored on it. This enables the VMware ESX host to always identify the copy correctly.

vSphere 4 and 5 include the ability to individually resignature and/or mount VMFS volume copies through the use of the vSphere Client or with the CLI utility vicfg-volume. The resignature and force-mount are volume-specific, which allows for greater granularity in the handling of snapshots. This feature is very useful when creating and managing volume copies created by local replication products such as TimeFinder, and in this particular case remote replication technologies such as Symmetrix Remote Data Facility (SRDF).

Figure 170 shows an example of this. In the screenshot, there are two highlighted volumes, VMFS_Datastore_VMAX and snap-6e6918f4-VMFS_Datastore_VMAX. These volumes are hosted on two separate devices, one device is a copy of the other created with the TimeFinder software.



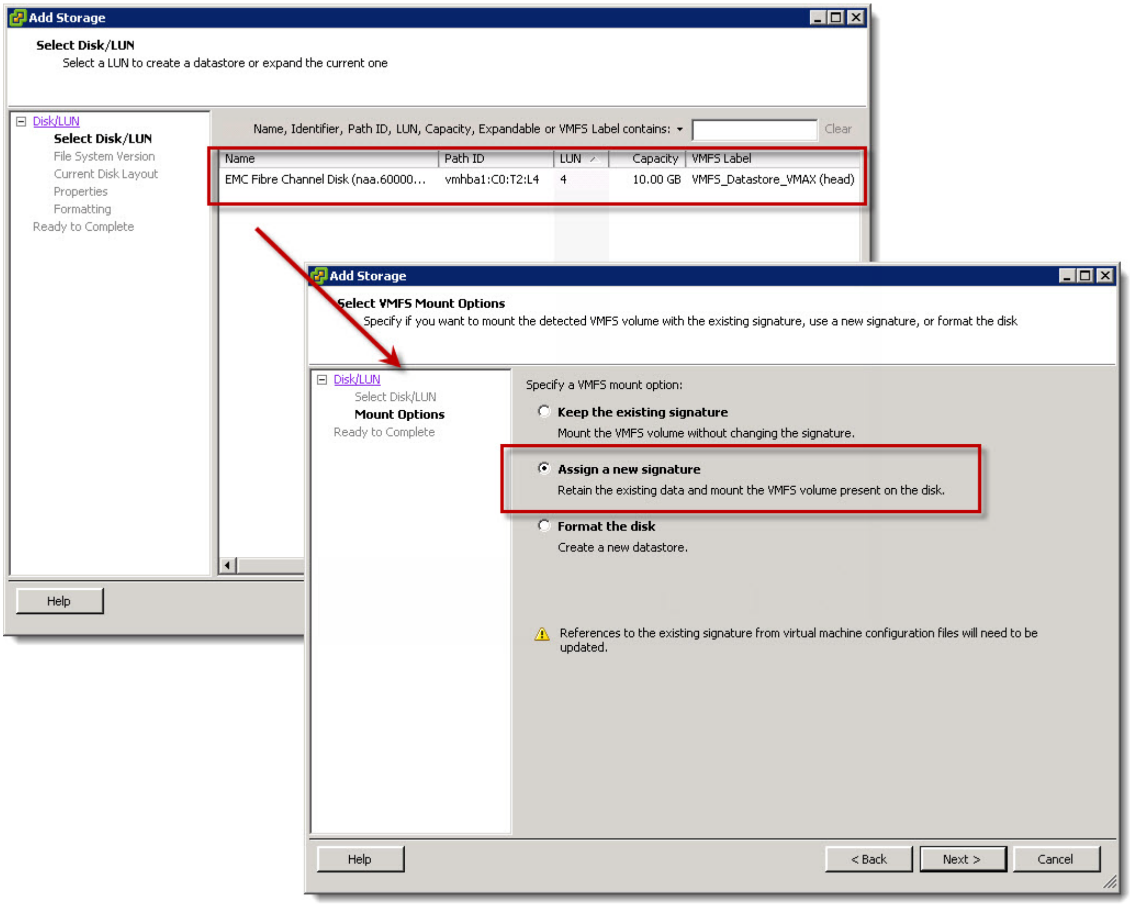
svpg-dell-c2-s14.ebc.emc.local VMware ESXi, 5.0.0, 441354

Datastores

Identification	Status	Device	Drive Type	Capacity
Templates	✓ Normal	EMC Fibre Channel...	Non-SSD	239.75 GB
DS1	✓ Normal	EMC Fibre Channel...	Non-SSD	719.75 GB
VMFS_datastore	✓ Normal	EMC Fibre Channel...	Non-SSD	99.75 GB
VMFS_Datastore_VMAX	✓ Normal	EMC Fibre Channel...	Non-SSD	9.75 GB
snap-6e6918f4-VMFS_Datastore_VMAX	✓ Normal	EMC Fibre Channel...	Non-SSD	9.75 GB
Storage1 (5)	✓ Normal	Local HITACHI Di...	Non-SSD	63.25 GB

Figure 170 Individually resignature a volume

The vSphere Client was used to resignature the volume with the name snap-6e6918f4-VMFS_Datastore_VMAX. The Add Storage wizard in Figure 171 demonstrates how the volume was resignatured.



Select Disk/LUN
Select a LUN to create a datastore or expand the current one

Name, Identifier, Path ID, LUN, Capacity, Expandable or VMFS Label contains:

Name	Path ID	LUN	Capacity	VMFS Label
EMC Fibre Channel Disk (naa.60000...	vmhba1:C0:T2:L4	4	10.00 GB	VMFS_Datastore_VMAX (head)

Select VMFS Mount Options
Specify if you want to mount the detected VMFS volume with the existing signature, use a new signature, or format the disk

Specify a VMFS mount option:

○ **Keep the existing signature**
Mount the VMFS volume without changing the signature.

● **Assign a new signature**
Retain the existing data and mount the VMFS volume present on the disk.

○ **Format the disk**
Create a new datastore.

⚠ References to the existing signature from virtual machine configuration files will need to be updated.

Figure 171 Add Storage wizard used to resignature a volume

Creating virtual machines at the remote site

The following steps discuss the process of creating virtual machines at the remote site for ESXi 5 (a very similar process is used in ESX 4).

1. The first step to create virtual machines at the remote site is to enable access to the R2 devices for the VMware ESX cluster at the remote data center.

2. The vCenter Server does not allow duplication of objects in a *Datacenter* object. If the same vCenter is used to manage the VMware ESX at the production and remote site, the servers should be added to different *Datacenter* constructs in vCenter Server.

3. The SCSI bus should be scanned after ensuring the R2 devices are split from the R1 devices after the device pairs are in a synchronized or a consistent state. The scanning of the SCSI bus can be done either using the service console or the vSphere Client. The devices that hold the copy of the VMFS are displayed on the VMware ESX cluster at the remote site.

4. In vSphere 4 and 5 environments, when a virtual machine is created all files related to the virtual machine are stored in a directory on a datastore. This includes the configuration file and, by default, the virtual disks associated with a virtual machine. Thus, the configuration files automatically get replicated with the virtual disks when storage array based copying technologies are leveraged.

   The registration of the virtual machines from the target device can be performed using the vSphere CLI tools or the vSphere Client. The re-registration of cloned virtual machines is not required when the configuration information of the production virtual machine changes. The changes are automatically propagated and used when needed.

   As recommended in step 2 , if the VMware ESX hosts at the remote site are added to a separate *Datacenter* construct, the names of the virtual machines at the remote *Datacenter* matches those of the production *Datacenter.*

5. The virtual machines can be started on the VMware ESX hosts at the remote site without any modification if the following requirements are met:

   - The target VMware ESX hosts have the same virtual network switch configuration — i.e., the name and number of virtual switches should be duplicated from the source VMware ESX cluster.
   - All VMFS used by the source virtual machines are replicated. Furthermore, the VMFS labels should be unique on the target VMware ESX hosts.

- The minimum memory and processor resource reservation requirements of all cloned virtual machines can be supported on the target VMware ESX host(s). For example, if ten source virtual machines, each with a memory resource reservation of 256 MB, need to be cloned, the target VMware ESX cluster should have at least 2.5 GB of physical RAM allocated to the VMkernel.
- Virtual devices such as CD-ROM and floppy drives are attached to physical hardware, or are started in a disconnected state when the virtual machines are powered on.

6. The cloned virtual machines can be powered on the vSphere Client or through command line utilities when required.

Configuring remote site virtual machines with replicated RDMs

When a RDM is generated, a virtual disk is created on a VMFS pointing to the physical device that is mapped. The virtual disk that provides the mapping also includes the unique ID of the device it is mapping. The configuration file for the virtual machine using the RDM contains an entry that includes the label of the VMFS holding the RDM and the name of the RDM. If the VMFS holding the information for the virtual machines is replicated and presented on the VMware ESX host at the remote site, the virtual disks that provide the mapping are also available in addition to the configuration files. However, the mapping file cannot be used on the VMware ESX host at the remote site since they point to non-existent devices. Therefore, EMC recommends using a copy of the source virtual machine’s configuration file instead of replicating the VMFS.

The following steps create copies of production virtual machines using RDMs at the remote site:

1. On the VMware ESX host cluster at the remote site create a directory on a datastore to hold the files related to the cloned virtual machine. A VMFS on internal disk, unreplicated SAN-attached disk, or NAS-attached storage should be used for storing the files for the cloned virtual disk. This step has to be performed only once.
2. Copy the configuration file for the source virtual machine to the directory created in step 1. This step has to be repeated only if the configuration of the source virtual machine changes.

3. Register the cloned virtual machine using the vSphere Client or VMware remote tools. This step does not need to be repeated.
4. Generate RDMs on the target VMware ESXi host in the directory created in step 1 above. The RDMs should be configured to address the R2.
5. The virtual machine at the remote site can be powered on using either the vSphere Client or the VMware remote tools when needed.

Note: The process listed in this section assumes the source virtual machine does not have a virtual disk on a VMFS. The process to clone virtual machines with a mix of RDMs and virtual disks is complex and beyond the scope of this document. Readers are requested to contact the authors at cody.hosterman@emc.com or drew.tonnesen@emc.com, if such requirements arise.

Write disabled devices in a VMware environment

In order to prevent unwanted and invalid changes to target SRDF devices, EMC sets these devices (e.g. R2) to write-disabled upon RDF pair creation. These devices remain write-disabled until the pair is split or failed over.

In most VMware environments, the read/write enabled R1 devices are presented to one cluster while the write-disabled R2 devices are presented to another. The presence of a multitude of write-disabled devices that contain unmounted VMFS volume copies (referred to as unresolved VMFS volumes) pose issues for certain VMware operations. These operations are:

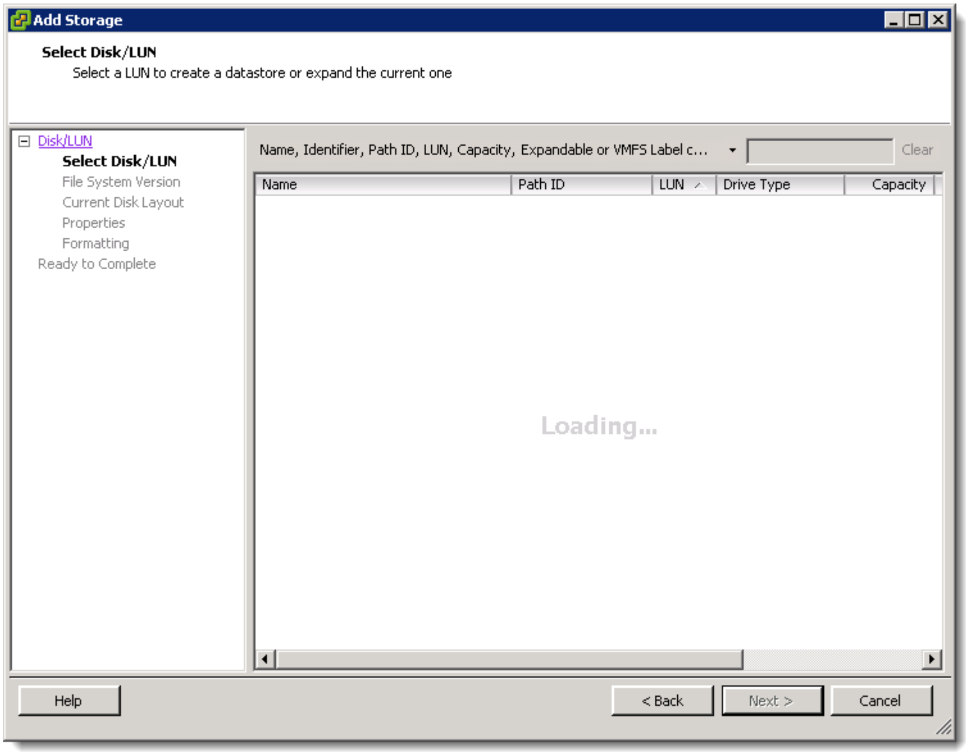
- HBA storage rescan
- The “Add Storage Wizard”

Both of these operations query the devices and the presence of these write-disabled devices can slow them down due to the ESXi hostd process retrying operations on these devices until a retry limit or timeout is reached.

IMPORTANT

The HBA storage rescan issue has been resolved in vSphere 4 only at vSphere 4.1 U3 and later and in vSphere 5 only at vSphere 5.1 and later.

In the "Add Storage Wizard" the slow scan manifests itself in the second screen when ESXi is attempting to display available devices. The screen will say "Loading" until it completes reading through these write-disabled devices which can take more than 25 seconds per device as seen in Figure 172. Normally ESXi would spend in the order of milliseconds reading a device.



Select Disk/LUN

Select a LUN to create a datastore or expand the current one

Figure 172 Device query screen in the "Add Storage Wizard"

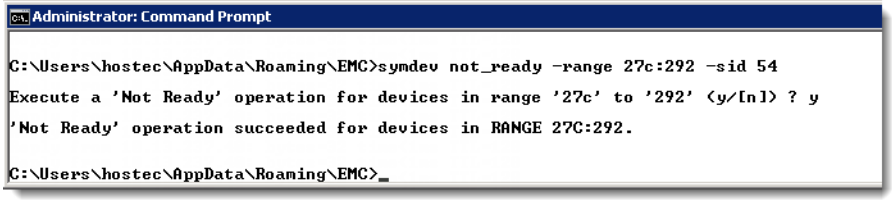
During the device query operation of the wizard, the hostd process examines all of the devices that do not fail the initial SCSI open() call such as the SRDF R2 volumes. Even though write-disabled, these devices are still have a status of Ready and therefore succeed the call. The ESXi host therefore can detect that they host a VMFS volume (since they are only write-disabled, not read/write disabled). Furthermore, ESXi detects that the volumes have an invalid signature. After the ESXi host detects this situation it attempts to update metadata on the unresolved VMFS volume. This attempt fails because the SRDF R2 device is protected and write-disabled by the active SRDF session. Nevertheless, ESXi will attempt this metadata update numerous times before excluding it as a candidate and moving on. These retries greatly increase the time to discover all of the candidate devices in the “Add Storage Wizard”.

Note: Invalid signatures are caused when a VMFS volume is exactly copied from an source device to an target device. Since the WWN and therefore the Network Address Authority (NAA) is different on the source and target device, the VMFS signature, which is based on the device NAA, is consequently invalid on that device.

In addition to the increased scan times, the vmkernel.log will contain messages similar to the following:

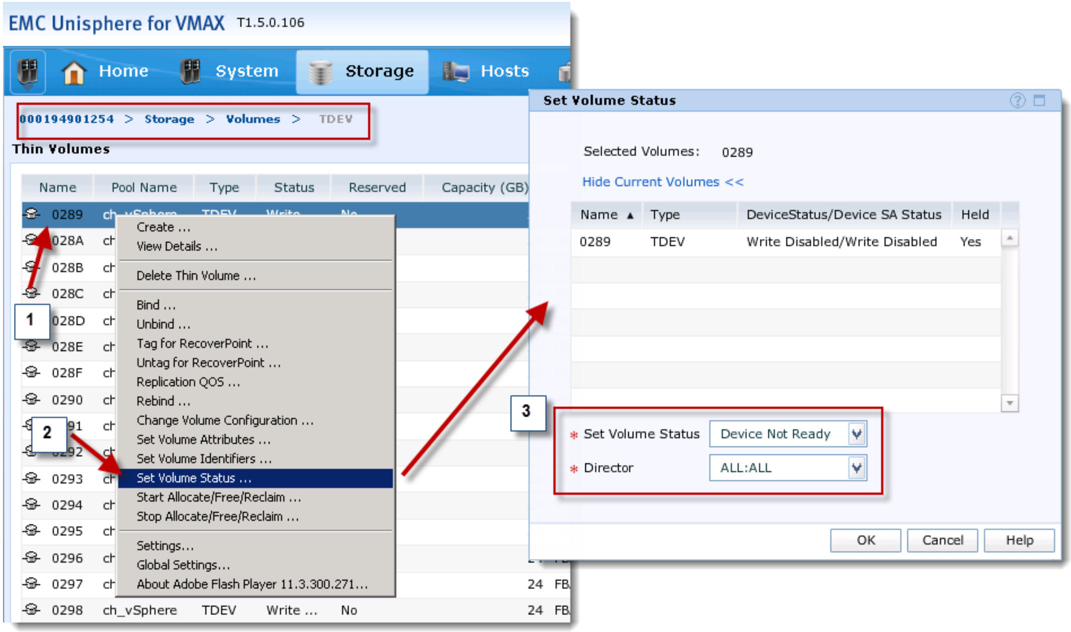
```
Jul 20 10:05:25 VMkernel: 44:00:29:48.149
cpu46:4142)ScsiDeviceIO: 1672: Command 0x2a to
device "naa.60060480000290xxxxxx533030384641"
failed H:0x0 D:0x2 P:0x3 Possible sense data: 0x7
0x27 0x0.
```

In order to avoid this retry situation and therefore reducing scan times, users can use SYMCLI (Figure 173) or Unisphere for VMAX (Figure 174 on page 313) to set the write disabled devices to a state of Not Ready. When a device is Not Ready, ESXi fails the SCSI open() call and immediately skips to the next device. Setting devices to a status of Not Ready will therefore remediate the “Add Storage Wizard” slowdown.



```
C:\Users\hostec\AppData\Roaming\EMC>symdev not_ready -range 27c:292 -sid 54
Execute a 'Not Ready' operation for devices in range '27c' to '292' (y/[n]) ? y
'Not Ready' operation succeeded for devices in RANGE 27C:292.

C:\Users\hostec\AppData\Roaming\EMC>_
```

Figure 173 Setting devices not ready with Solutions Enabler CLI


EMC Unisphere for VMAX T1.5.0.106
Home
System
Storage
Hosts
1
2
3
Create ...
View Details ...
Delete Thin Volume ...
Bind ...
Unbind ...
Tag for RecoverPoint ...
Untag for RecoverPoint ...
Replication QOS ...
Rebind ...
Change Volume Configuration ...
Set Volume Attributes ...
Set Volume Identifiers ...
Set Volume Status ...
Start Allocate/Free/Reclaim ...
Stop Allocate/Free/Reclaim ...
Settings...
Global Settings...
About Adobe Flash Player 11.3.300.271...
Hide Current Volumes <<
OK
Cancel
Help

000194901254 > Storage > Volumes > TDEV

Thin Volumes

Name	Pool Name	Type	Status	Reserved	Capacity (GB)
0289	ch_vSphere	TDEV	Write ...	No	
028A	ch				
028B	ch				
028C	ch				
028D	ch				
028E	ch				
028F	ch				
0290	ch				
0291	ch				
0292	ch				
0293	ch				
0294	ch				
0295	ch				
0296	ch				
0297	ch				24 FB
0298	ch_vSphere	TDEV	Write ...	No	24 FB

Set Volume Status

Selected Volumes: 0289

Name ▲	Type	DeviceStatus/Device SA Status	Held
0289	TDEV	Write Disabled/Write Disabled	Yes

* Set Volume Status: Device Not Ready

* Director: ALL:ALL

Figure 174 Setting devices not ready with Unisphere for VMAX

The other operation that is subject to long rescan times is the HBA storage rescan. Fortunately, setting the devices to Not Ready will also prevent issues with this operation, in all versions of ESX(i).