Redbooks
ibm.com/redbooks

IBM FlashSystem V9000 and VMware Best Practices Guide

Rawley Burbridge
Matt Levan
Dusan Tekeljak
James Thompson
Axel Westphal
Karen Orlando

Storage

IBM

Redpaper

International Technical Support Organization

IBM FlashSystem V9000 and VMware Best Practices Guide

December 2015

REDP-5247-00

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (December 2015)

This edition applies to FlashSystem V9000 software version 7.5, and VMware version 6.

This document was created or updated on December 7, 2015.

Contents

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Easy Tier®
FlashCopy®
FlashSystem™
HyperSwap®
IBM®
IBM FlashCore™
IBM FlashSystem®
IBM Spectrum™
IBM Spectrum Control™
IBM Spectrum Protect™
MicroLatency®
Real-time Compression™
Redbooks®
Redpaper™
Redbooks (logo) ®
Storwize®
System Storage®
Tivoli®
Variable Stripe RAID™
XIV®

The following terms are trademarks of other companies:

Inc., and Inc. device are trademarks or registered trademarks of Kenexa, an IBM Company.

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

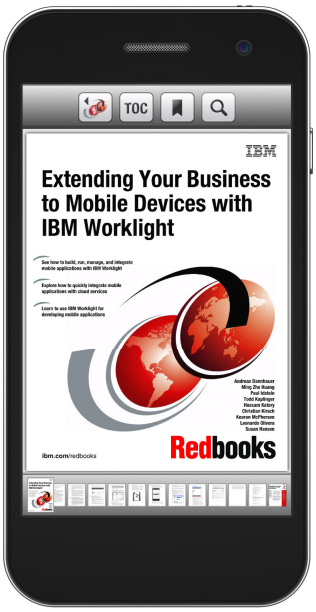
Other company, product, or service names may be trademarks or service marks of others.

Find and read thousands of IBM Redbooks publications

- Search, bookmark, save and organize favorites
- Get up-to-the-minute Redbooks news and announcements
- Link to the latest Redbooks blogs and videos

Get the latest version of the Redbooks Mobile App

iOS **Download Now** Android



Promote your business in an IBM Redbooks publication

Place a Sponsorship Promotion in an IBM® Redbooks® publication, featuring your business or solution with a link to your web site.

Qualified IBM Business Partners may place a full page promotion in the most popular Redbooks publications. Imagine the power of being seen by users who download millions of Redbooks publications each year!

ibm.com/Redbooks
About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK

Preface

This IBM® Redpaper™ publication describes best practices for deploying IBM FlashSystem™ V9000 enterprise storage system in a VMware vSphere environment. It includes guidelines and examples of the latest FlashSystem V9000 hardware and software integrated with VMware version 6 to demonstrate the business benefits of these solutions.

Topics illustrate planning, configuring, operations, and preferred practices that include integration of FlashSystem V9000 with the VMware vCloud suite of applications:

- vCenter Web Client (VWC)
- vStorage APIs for Storage Awareness (VASA)
- vStorage APIs for Array Integration (VAAI)
- vCenter Site Recovery Manager (SRM/SRA)

The authors also describe how to deploy a cloud-based solution with FlashSystem V9000 in an environment with VMware and IBM Spectrum™ Control Base Edition 2.1.1.

The chapters cover the following topics:

- Chapter 1, “Environment overview” on page 1
- Chapter 2, “Planning guidelines for FlashSystem V9000 and VMware” on page 11
- Chapter 3, “VMware configurations for FlashSystem V9000” on page 21
- Chapter 4, “Integrated management” on page 31
- Chapter 5, “VMware and FlashSystem V9000 multi-site guidelines” on page 49
- Chapter 6, “Data reduction considerations” on page 69

This paper is intended for presales consulting engineers, sales engineers, and IBM clients who want to deploy IBM FlashSystem V9000 in virtualized data centers that are based on VMware vSphere.

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose Center.

Rawley Burbridge is an IBM FlashSystem Corporate Solutions Architect who specializes in VMware and FlashSystem solutions. Rawley has been working with VMware and storage technologies for more than 10 years and has spent over two years working exclusively with IBM FlashSystem. He has performed many IT-related roles, including VMware and storage administration, and has written many technical white papers and advised on several IBM Redbooks® publications. Rawley holds a bachelor's degree in Computer Information Systems from Missouri State University.

Matt Levan is an IBM FlashSystem Corporate Solutions Architect. He has had numerous roles in IT and storage technologies over more than 15 years, including storage administration, technical support, and solutions architect. His current role is to help worldwide field technical sellers for IBM flash memory products. His primary responsibilities are to provide technical sales engineering, competition-winning strategies, and cohesive technical sales solutions to IBM internal organizations and IBM clients worldwide. Previously, Matt spent many years at technology-leading companies, such as Novus Consulting Group, VeriSign, EMC, and Innovative Data Solutions.

Dusan Telekjak has more than five years of experience in the virtualization field. He has a background in closely related technologies, including server operating systems, networking, and storage. Before joining the VMware Center of Excellence at IBM in 2012, he worked extensively with Microsoft cloud technologies. His main scope of work for IBM consists of design and integration projects, including various vSphere in Lenovo or IBM Flex Systems implementations. Dusan has a master's degree in engineering from Slovak University of Technology and holds several IT industry certifications, including VCAP-DCD, VCAP-DCA, MCITP, and others. He was honored with the #vExpert2015 award by VMware for his contribution to the community.

James Thompson is a Performance Analyst for IBM Systems and Technology Group. He has worked at IBM for 15 years, supporting the development of IBM storage products. He holds a bachelor's degree in Computer Science from Utah State University.

Axel Westphal is a certified IT Specialist at the IBM EMEA Storage Competence Center (ESCC) in Mainz, Germany. He joined IBM in 1996, working for Global Services as a systems engineer. His areas of expertise include set up and demonstration of IBM System Storage® products and solutions in various environments. He has written several storage white papers and co-authored the IBM Redbook publication titled *IBM System Storage DS8000: Architecture and Implementation*, SG24-8886.

Karen Orlando is a Project Leader at the International Technical Support Organization, Tucson Arizona Center. Karen has more than 25 years in the IT industry, with extensive experience in open systems management, and in information, development, software development, and test of IBM hardware and software for storage. She holds a degree in Business Information Systems from the University of Phoenix and has been a certified Project Management Professional (PMP) since 2005.

Now you can become a published author, too

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time. Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us.

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form:

  ibm.com/redbooks

- Send your comments by email:

  redbooks@us.ibm.com

- Mail your comments:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Environment overview

This chapter introduces the IBM FlashSystem V9000 storage system and VMware virtualization environment. Many of the advanced software features of the FlashSystem V9000 are highlighted in this chapter. The VMware vCloud suite features and functions are described in further detail.

The FlashSystem V9000 integrates with the VMware vCloud Suite to facilitate administration, management, and monitoring. The chapter describes IBM Spectrum Control™ Base Edition, which is a centralized server system that consolidates a range of IBM storage provisioning, automation, and monitoring functions through a unified server platform. It provides a single-server backend location and enables centralized management of IBM storage resources for different virtualization and cloud platforms.

This chapter includes the following sections:

1.1 FlashSystem V9000 basics

The FlashSystem V9000, shown in Figure 1-1, delivers high-capacity and fully integrated management for the enterprise data center. It uses a full-featured and scalable all-flash architecture that performs at up to 2.5 million IOPS with the IBM MicroLatency® module. It is scalable to 19.2 GBps and delivers an effective capacity of up to 2.28 PB. Using its flash-optimized design, FlashSystem V9000 can provide response times of 200 microseconds.

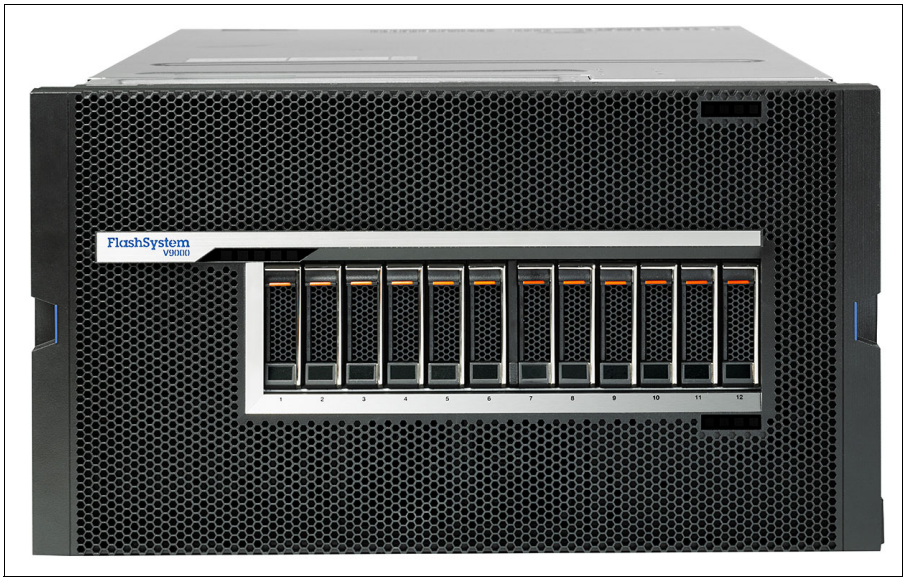


Figure 1-1 IBM FlashSystem V9000

FlashSystem V9000 delivers enterprise-class advanced storage capabilities, including these, among others:

- IBM Real-Time Compression
- IBM EasyTier functions
- Advanced Copy Services
- Data virtualization
- Highly available configurations
- Thin provisioning

Advanced Encryption Standard 256-bit (AES 256) hardware-based encryption adds to the rich feature set.

Scalability

FlashSystem V9000 offers the flexibility to future-proof the purchase of an all-flash solution by using the ability to scale up for increased capacity, scale out for increased performance, or both. It delivers up to 57 TB per building block, scales to four building blocks, and offers up to four additional 57 TB V9000 storage enclosure expansion units for large-scale enterprise storage system capability. Building blocks can be either fixed or scalable. You can combine scalable building blocks to create larger clustered systems in a way that operations are not disrupted.

For more information about FlashSystem V9000 scalability, see Chapter 5 of the IBM Redbooks publication titled *Introducing and Implementing IBM FlashSystem V9000*, SG24-8273.

1.1.1 Benefits of FlashSystem V9000 FlashCore Technology

The IBM FlashCore™ technology used in the FlashSystem V9000 employs several new and patented mechanisms to provide greater capacity and throughput, yet at a lower cost than previously available, as shown in Figure 1-2.

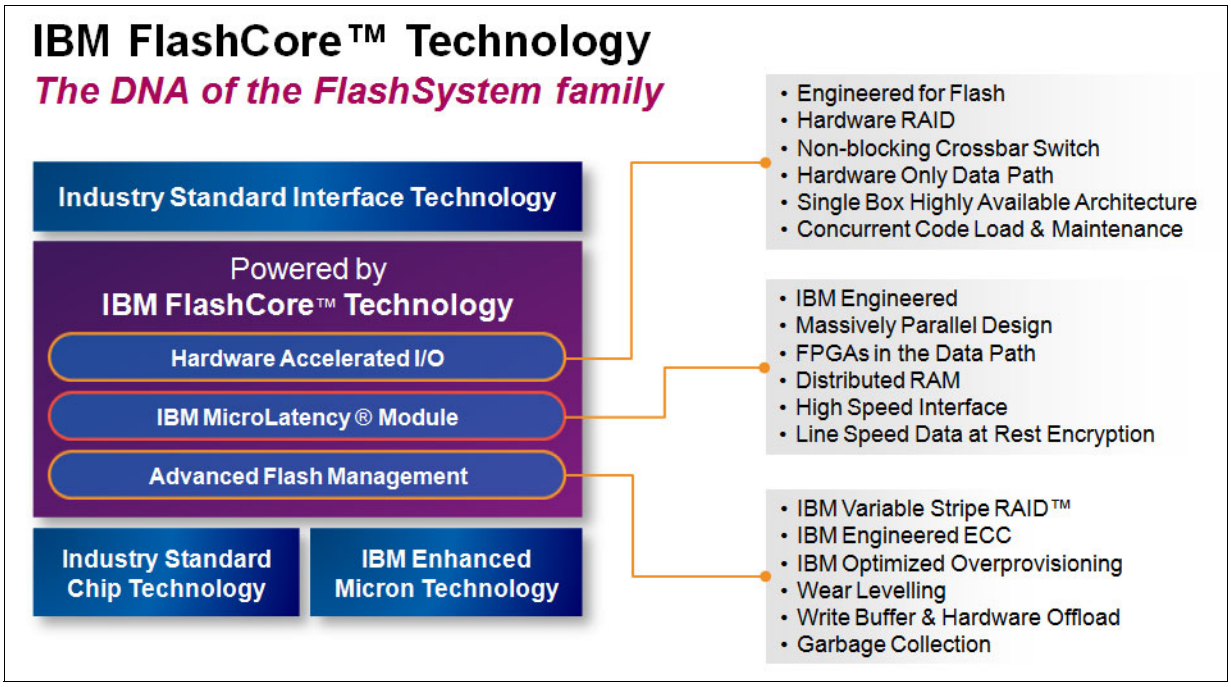


Figure 1-2 IBM FlashCore technology

Hardware Accelerated I/O

FlashSystem V9000 hardware design offers several unique IBM features:

- Hardware RAID
- Non-blocking crossbar switch
- Hardware-only data path
- Single box highly available architecture
- Concurrent code load
- Concurrent maintenance

IBM MicroLatency module

FlashSystem V9000 uses the new 20 nm multilevel cell (MLC) flash card memory chips. IT also uses IBM engineered massively parallel design, field programmable gate arrays (FPGAs) in the data path, distributed RAM, and high-speed interfaces, plus hardware-based data-at-rest encryption.

Advanced flash management

IBM FlashSystem V9000 has unique patented designs to ensure maximum availability. These include IBM Variable Stripe RAID™ (VSR), IBM engineered error correction code (ECC), IBM optimized over-provisioning, advanced wear leveling on the IBM MicroLatency module, write buffer and hardware offload, and garbage collection.

This is all made possible because of the following IBM patented innovations:

- ECC algorithms that correct high bit-error rates
- Variable voltage and read level shifting that help to maximize flash endurance
- Health binning and heat segregation, which continually monitor the health of flash blocks and perform asymmetrical wear leveling and sub-chip tiering

1.1.2 FlashSystem V9000 Tier 1 advanced software features

The advanced software features described in the subsections that follow are supported.

Real-time Compression

The IBM Real-time Compression™ software that is embedded in the FlashSystem V9000 addresses the requirements of primary storage data reduction, including performance with the use of dedicated compression acceleration hardware. It does so by using a purpose-built technology called Real-Time Compression that uses the Random Access Compression Engine (RACE) accelerator cards.

Thin provisioning

In a shared storage environment, thin provisioning is a method for optimizing the use of available storage capacity. Thin provisioning relies on allocating blocks of data only as they are needed, rather than the older method of allocating all blocks when a volume is created. This method eliminates almost all white space, which helps to avoid poor usage rates that occur in traditional storage allocation methods where large pools of storage capacity are allocated to individual hosts but remain unused.

Advanced Copy Services

Advanced Copy Services is a class of storage array and storage device functions that allows various forms of block-level data duplication. By using Advanced Copy Services, you can make mirror images of all or part of your data, eventually, between distant sites. The following Copy Service functions are implemented within a FlashSystem V9000 (IBM FlashCopy® and Image Mode Migration) or between one FlashSystem V9000 and another FlashSystem V9000 (IBM Metro Mirror and IBM Global Mirror):

- FlashCopy is the function for making a point-in-time copy, which can be either a full clone or incremental snapshot of the volume
- Metro Mirror is the remote copy function for synchronous remote replication
- Global Mirror the remote copy function for asynchronous remote replication

Easy Tier

IBM Easy Tier® is a performance function that automatically migrates extents of a volume to or from one class of storage tier to another storage tier. Easy Tier works by monitoring the host I/O activity and latency on the extents of all volumes in a multi-tier storage pool over a 24-hour period. It uses that information to create a migration plan to either move hot data up to the highest performance storage tier or move data that has cooled off to a lower-performance storage tier.

HyperSwap

IBM HyperSwap® is a high availability and disaster recovery function in a single solution. HyperSwap allows each volume to be presented by two I/O groups. The configuration tolerates combinations of control enclosure node and site failures by using the VMware host multipathing driver. The use of FlashCopy helps maintain a golden image during automatic resynchronization. No FlashCopy license is required to use HyperSwap. However, because remote mirroring is used to support HyperSwap capability, a remote mirroring license is a requirement for using HyperSwap.

1.2 VMware overview

In this paper, we focus on integration and preferred practices for FlashSystem V9000 software Version 7.5 with the following VMware vSphere Version 6.0 platform products:

- VMware ESXi 6.0
- VMware vCenter Server 6.0
- VMware vCenter Site Recovery Manager 6.0

VMware, Inc. was founded in 1998 to bring virtual machine (VM) technology to industry-standard computers. For more information, see the company's website:

http://www.vmware.com

The version of VMware vSphere ESXi covered in this paper is 6.0, and we advise checking with VMware for any future release changes.

VMware vSphere hypervisor architecture provides a robust, production-proven, high-performance virtualization layer. It enables multiple virtual machines to share hardware resources with performance that can match (and in some cases exceed) native throughput.

Because of the virtualization, the guest operating system is not aware where the resources, for example CPU and memory, originate. One virtual machine is isolated from the other. This is possible because of the virtualization layer, and it enables the sharing of physical devices with the virtual machines. The resources do not materialize from thin air though, and an associated physical device or pool of resources needs to be available for providing resources to the virtual machines. Later in this paper, we describe virtualization tools that are available and that can help you respond to your business needs, enabling resilience and fast recoveries from a failure.

1.2.1 Terms and concepts

The subsections that follow define terms and concepts that are used throughout this paper.

VMware vSphere ESXi

VMware vSphere ESXi is a lightweight virtualization operating system, also referred to as a *hypervisor*, that enables the deployment of multiple, secure, independent virtual machines on a single physical server.

VMware vCenter Server

VMware vCenter Server provides unified management for the entire virtual infrastructure and enables key vSphere capabilities such as live migration. vCenter Server can manage thousands of virtual machines across multiple locations and streamlines administration with features such as rapid provisioning and automated policy enforcement.

Virtual Machine File System

Virtual Machine File System (VMFS) is a cluster file system optimized for virtual machines. A virtual machine consists of a small set of files that are stored in a virtual machine folder. VMFS is the default file system for physical SCSI disks and partitions. VMFS is supported on a wide range of Fibre Channel and iSCSI SAN storage arrays. VMFS is a cluster file system allowing shared access to allow multiple ESXi hosts to concurrently read and write to the same storage. VMFS can expand dynamically. This allows for an increase in the size of the VMFS without downtime.

Notes:

A VMFS datastore can span across multiple LUNs (extents), but the recommended implementation is to have a one-to-one relationship.

When we refer to a *datastore*, we mean a VMware container that is viewed on the VMware side, but if we refer to a *logical unit number (LUN)*, we mean the storage array volume.

Storage I/O Control

Storage I/O Control (SIOC) protects virtual machines (VMs) from I/O intense VMs that consume a high portion of the overall available I/O resources. This is also called the noisy neighbor problem. The problem comes from the fact that ESXi hosts share LUNs and one host might have multiple VMs accessing the same LUN, whereas another host might have only one virtual machine. Without SIOC, both hosts have the same device queues available. Assuming that one host has two VMs and the other has just one VM accessing the same LUN, it will allow the lonely VM to have twice the storage array queue and cause congestion on the storage array. With SIOC enabled, SIOC will change the device queue length for the single VM and therefore reduce the storage array queue for all VMs to an equal share and throttle the storage queue.

VMware vSphere Storage Distributed Resource Scheduler

VMware vSphere Storage Distributed Resource Scheduler (SDRS) automates load balancing by using storage characteristics to determine the best place for a virtual machine's data to reside, both when it is created and when it is used over time.

VMware vSphere vMotion

VMware vSphere vMotion enables live migration of virtual machines between servers and across virtual switches with no disruption to users or loss of service, eliminating the need to schedule application downtime for planned server maintenance.

VMware vSphere Storage vMotion

VMware vSphere Storage vMotion enables live migration of virtual-machine disks with no disruption to users, eliminating the need to schedule application downtime for planned storage maintenance or storage migrations.

VMware vSphere High Availability

VMware vSphere High Availability (HA) provides cost-effective, automated restart within minutes for all applications if a hardware or operating system failure occurs.

VMware vCenter Site Recovery Manager

VMware vCenter Site Recovery Manager (SRM) is the disaster recovery management product that ensures the simple and reliable disaster protection for all virtualized applications. Site Recovery Manager uses VMware vSphere Replication or storage-based replication to provide centralized management of recovery plans, enable nondisruptive testing, and automate site recovery and migration processes.

It integrates with third-party storage arrays and replication appliances to provide a complete integrated Business Continuity solution. This integration is achieved through a unique Storage Replication Adapter (SRA) that is created by storage array or replication vendors.

vStorage APIs for Array Integration

vStorage APIs for Array Integration (VAAI) is a feature that provides hardware acceleration functions. It enables your host to offload specific virtual machine and storage management operations to compliant storage hardware. With the storage hardware assistance, your host performs these operations faster and consumes less CPU, memory, and storage fabric bandwidth. You can read more about its features in 3.1, “ESXi host offloading with VAAI and FlashSystem V9000” on page 22.

vSphere Storage APIs for Storage Awareness

vSphere Storage APIs for Storage Awareness (VASA) providers communicate with Virtual Center to indicate storage topology, capability and state information which supports policy-based management, operations management and DRS functions. VASA providers help to identify trends in a VM’s storage capacity use for troubleshooting, correlate events on the datastore and LUNS with a VM’s performance characteristics, and monitor health of storage. The policy-based management functions of a VASA provider helps administrators choose the appropriate storage device, and monitors and reports information about existing storage policies. You can read more about its features in 4.5.1, “Register VASA provider with vCenter server” on page 44.

Note: See the “About vSphere Storage” page in the VMware vSphere 6.0 Documentation Center for information about storage-related settings in VMware (vSphere Storage Guide):

http://vmw.re/1Gwf4hg

1.3 IBM Spectrum Control Base Edition

IBM Spectrum Control Base Edition, is a centralized server system that consolidates a range of IBM storage provisioning, automation, and monitoring solutions through a unified server platform. The Base Edition of IBM Spectrum Control is an improved and enhanced version of the IBM Storage Integration Server solution. It provides a single-server backend location and enables centralized management of IBM storage resources for different virtualization and cloud platforms.

The following solution components are included in the IBM Spectrum Control Base package:

- IBM Storage Enhancements for VMware vSphere Web Client (VWC)
- IBM Storage Provider for VMware VASA, v1.0 and v2.0
- IBM Storage Plug-in for VMware vRealize Orchestrator (VCO)
- Storage Management Pack for VMware vRealize Operations Manager (vROps)

When this paper was written, the latest supported version was IBM Spectrum Control Base Edition 2.1.1. See IBM Knowledge Center for the IBM Spectrum Control Base documentation and user guide:

http://www.ibm.com/support/knowledgecenter/STWMS9_2.1.1/scb_2.1.1_kc_welcome.html

Figure 1-3 shows how Spectrum Control Base Edition is integrated in a VMware environment with FlashSystem V9000 storage.

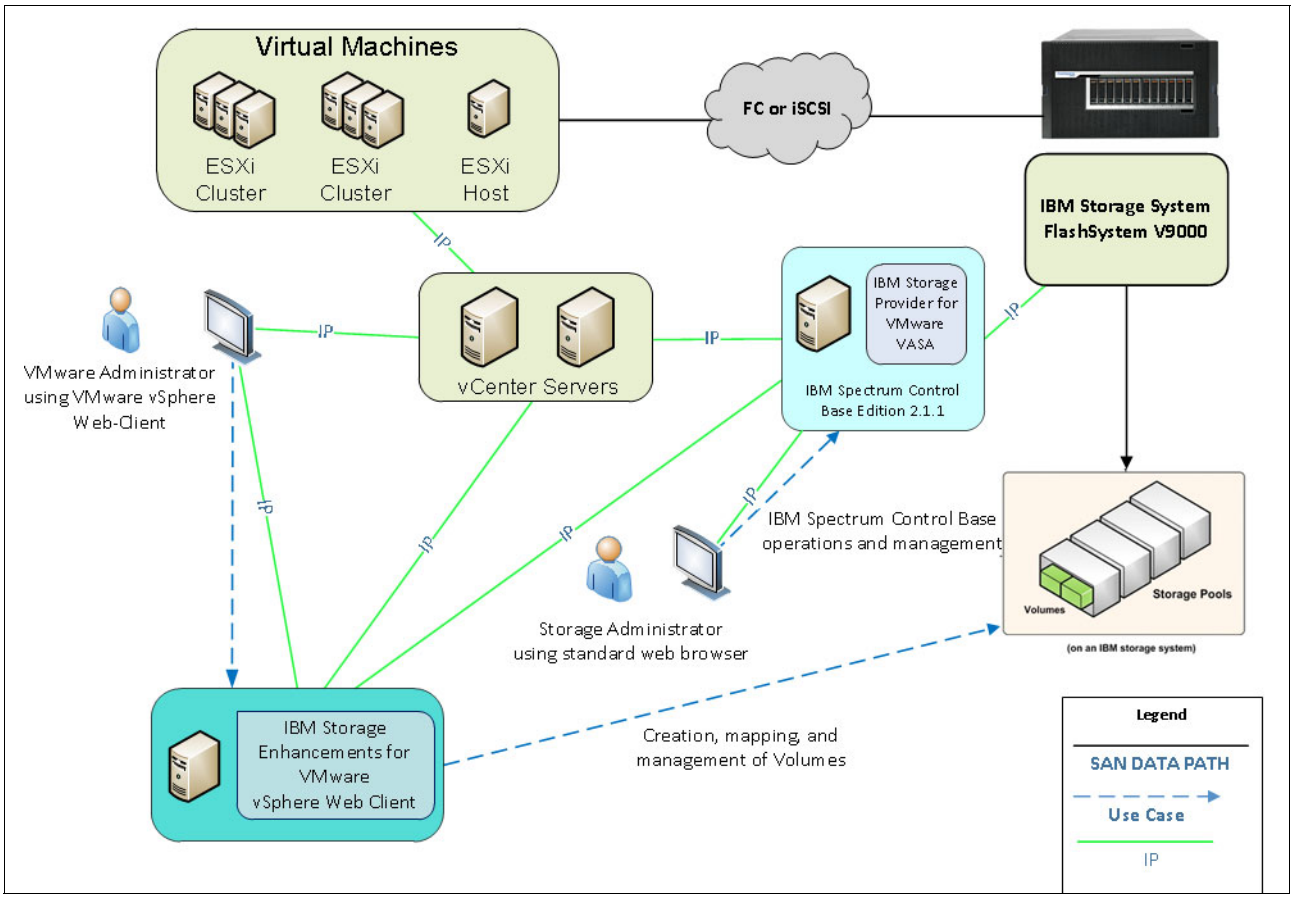


Figure 1-3 IBM Spectrum Control Base used in VMware environment with IBM FlashSystem V9000

For more about the IBM Spectrum Control Base Edition 2.1.1, see the release notes:

http://www.ibm.com/support/knowledgecenter/STWMS9_2.1.1/scb_2.1.1_kc_rn.dita

You can download the latest version of Spectrum Control Base Edition at no charge from the IBM Fix Central web page:

http://www.ibm.com/support/fixcentral/

Spectrum Control Base Edition can be managed through a standard web browser and a graphical user interface (GUI) or through a terminal and a command-line interface (CLI).

Note: At the time of the writing, VMware vRealize Operations Manager, VMware vRealize Orchestrator, and VASA 2.0 are not supported by Version 2.1.1 of the Spectrum Control Base Edition and FlashSystem V9000. They are currently supported only by the IBM XIV® storage array.

IBM Storage Provider for VMware VASA

The IBM Storage Provider for VMware VASA improves the ability to monitor and automate storage-related operations on VMware platforms.

From its Spectrum Control host, the Storage Provider for VMware VASA provides a standard interface for any connected VMware vCenter Server using the VMware vSphere APIs for Storage Awareness (VASA). It delivers information about IBM storage topology, capabilities, and state, together with storage events and alerts to vCenter Server in real time.

IBM Storage Enhancements for VMware vSphere Web Client

The IBM Storage Enhancements for the VMware vSphere Web Client plug-in integrates into the VMware vSphere Web Client platform and enables VMware administrators to independently and centrally manage their storage resources on IBM storage systems.

Depending on the IBM storage system in use, VMware administrators can self-provision volumes (LUNs) in selected storage pools that were predefined by the storage administrators. The volumes are mapped to ESXi hosts, clusters, or data centers as logical drives that can be used for storing VMware datastores (virtual machine data containers).

The Storage Enhancements for vSphere Web Client are installed only on the vSphere Web Client server, which allows multiple vCenter servers to use IBM storage resources. Storage pool attachment and detachment operations are performed on the Spectrum Control side, rather than on the vSphere Client side.

IBM Storage Enhancements for VMware vSphere Web Client are automatically deployed and enabled for each and every vCenter server that is registered for vSphere Web Client services on the connected Spectrum Control.

2

Planning guidelines for FlashSystem V9000 and VMware

This chapter describes planning guidelines and considerations when allocating and configuring IBM FlashSystem V9000 storage to an existing VMware environment.

We describe our lab environment, as well as planning topics that discuss VMware maximum considerations, SAN design and setup, and VMware mapping to FlashSystem V9000.

In the topic on multipathing, you'll find descriptions of path-selection plug-ins (PSPs) and guidance for access methods in these topics:

2.1 IBM FlashSystem V9000 planning

This paper is based on the assumption that you are configuring IBM FlashSystem V9000 storage to an existing VMware environment that consists of a storage area network (SAN) with two redundant fabrics. We also assume that you have already done general planning for FlashSystem V9000 and have conducted a Technical Delivery Assessment (TDA) to ensure that the planned solution is valid, and the system is installed and cabled to the SAN.

We show examples based on FlashSystem V9000 scalable systems configured with three 4-port 8 Gb FC adapters per AC2 control enclosure. FlashSystem V9000 is configured by considering the port use-for-performance method that is described in Appendix A in the IBM Redbooks publication titled *Introducing and Implementing IBM FlashSystem V9000*, SG24-8273. This method uses the customer SAN for connecting the AE2 and AC2 components and uses the internal switches only for intra-cluster communication.

2.1.1 General planning

See Chapter 4 in *Introducing and Implementing IBM FlashSystem V9000*, SG24-8273 for general planning guidelines for FlashSystem V9000. This paper focuses mainly on the logical planning and considerations.

2.1.2 Lab environment

As depicted in Figure 2-1 on page 13, our lab environment consists of two FlashSystem V9000s, three IBM System x3850 X5 servers, and SAN switches for the redundant internal and external fabrics. The equipment is connected to the lab ethernet network for management communication. VMware ESXi 6 is installed on the x3850 X5 servers and multiple virtual machines are configured to set up the VMware environment. Separate VMs are configured for the various components: Spectrum Control Base server, vCenter server, Spectrum Protect Snapshot, and some Linux test hosts.

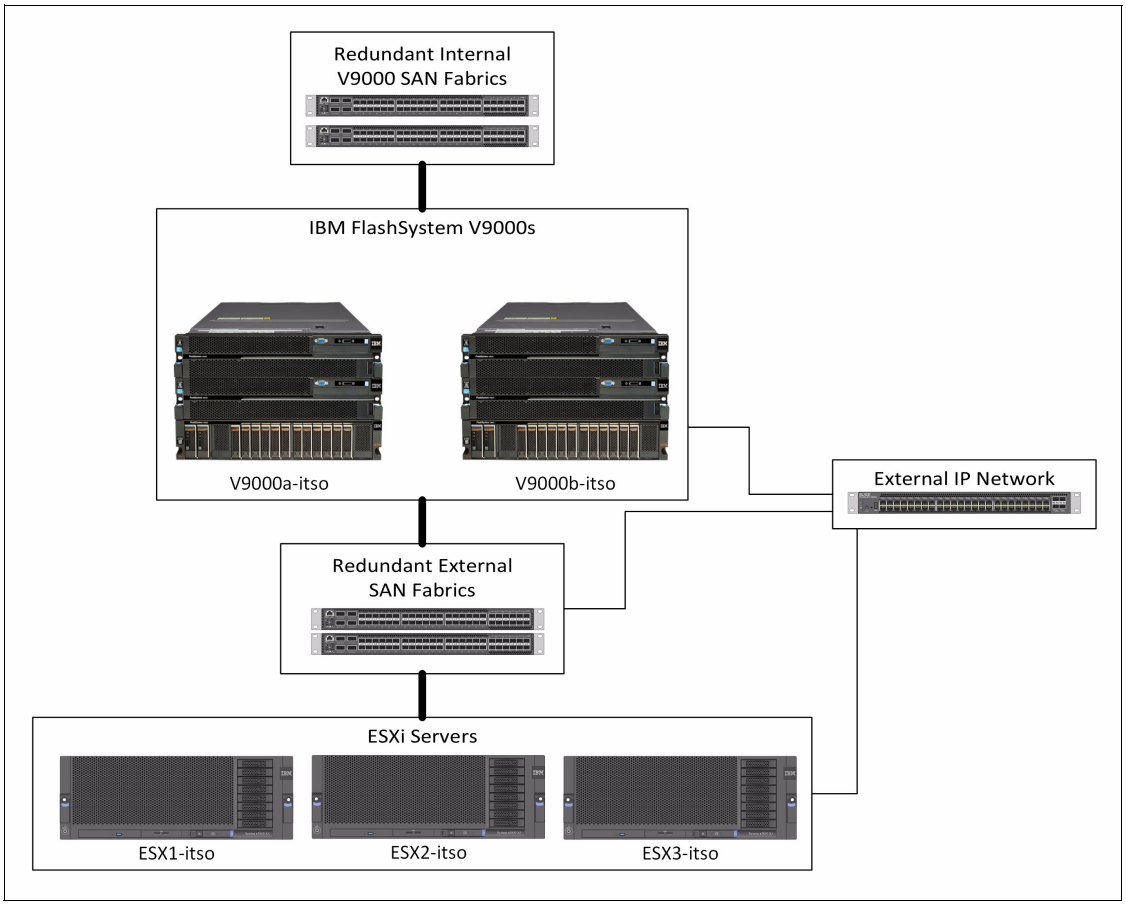


Figure 2-1 Physical Lab environment

2.1.3 VMware maximums

Ensure that you know the VMware maximums for the specific release. For very simple and small environments, you probably do not hit any of the maximums, but the larger your environment gets, the more important they get. Some are related and might not be obvious.

For instance, you have a maximum of 1024 paths per host, and you have 256 LUNs per host. Now, if you have 8 rather than 4 paths, you end up with a maximum of 128 LUNs per host, because you are now limited by the number of paths (1024/8 = 128). Often, all ESXi hosts in a cluster are zoned identically so that all hosts see the same LUNs. This means that 256 is the maximum for not only a host but for the cluster, too.

Also, the *maximum* means that it is a technical maximum or limitation. It does not necessarily mean that you should fully use the maximum. For instance, the maximum number of virtual machines per host is 512. But this does not mean that you can run 512 virtual machines per host. Virtual CPUs per core are 25. You can set up that many virtual CPUs, but they will probably not perform well.

If you have a test environment where, most of the time, no virtual machine is actually doing anything, this looks fine. But in a production environment, it is not a good idea.

The same applies to storage. For example, the maximum size of a single LUN (volume) is now 64 TB, but this does not mean that it is a good idea to use a 64 TB datastore. In most cases, it probably is not. VMware publishes documents listing configuration maximums for each release. Study the configuration maximums when planning your design.

2.1.4 Paths

The number of paths to a storage volume, commonly referred to as a *LUN* (logical unit number), is a composition of physical and logical elements that include the following components:

- Physical number of the following components:
  - FlashSystem V9000 host accessible ports (target ports)
  - Ports per ESXi host (initiator ports)
- Logical number of the following components:
  - FlashSystem V9000 host objects
  - Initiator ports (WWPNs) per FlashSystem V9000 host object
  - FlashSystem V9000 mappings per volume
  - Target ports per initiator ports (zoning)

Calculating the number of paths to a storage volume

It is possible to have a single FlashSystem V9000 host object for each ESXi host or to have multiple host objects up to a maximum of one per ESXi host initiator port.

With single initiator zoning, in our setup it is possible to have up to eight target ports.

In our lab setup, for each FlashSystem V9000 we have 16 FlashSystem V9000 host accessible ports (eight per AC2 control enclosure). Each ESXi host has 4 ports. For each ESXi host and AC2 control enclosure, the control enclosure ports are split between the two fabrics.

Given the physical elements in our lab setup (Figure 2-1 on page 13), there are several different ways that the logical elements can be configured, as shown in Table 2-1. The third row is the recommended way to get four paths per volume for a host system with four ports. The fourth row is not recommended, and the numbers are crossed through to indicate this.

Table 2-1 Calculating the # of paths to a volume based on logical configuration

# of host objects per host	# of zones required	# initiator ports (WWPNs) per host object	# of host object mappings per volume	# of target ports per initiator port	# of paths
1	4	4	1	2	8
4	8	1	2	1	2
2	4	2	1	2	4
~~1~~	~~2~~	~~4~~	~~1~~	~~8~~	~~32~~

Table 2-1 shows how different logical configurations affect the number (#) of paths to a volume.

The calculation is simple arithmetic: (# initiator ports per host object) * (number of host object mappings per volume) * (number of target ports per initiator port) = (number of paths).

For the first row, the ESXi host is configured as a single-host object with four WWPNs in FlashSystem V9000 management GUI. Each volume has one mapping to the host object.

There are four zones, two per fabric, containing one of the initiator ports and two ports from FlashSystem V9000 (one from each control enclosure node).

The second row represents a logical configuration that requires the most storage administrative effort but also allows you to specify exactly between which ports you want volumes to be accessed. Volumes are accessible over only two of the four initiator ports. With multiple volumes you could determine which volumes are accessible over which ports. This could be used to isolate volume traffic to different virtual machines. You could also create additional mappings to have the volumes accessible over all ports and this would increase the number of paths.

The third row represents a typical way to keep the number of paths to four. This scales with hosts with more than two ports. In this case, there would be a host object for every pair of ports. These port pairs would each be connected to a separate fabric. An equal portion of volumes would be mapped to each host object.

The fourth row is not a recommended configuration and shows a high number of paths that would result from lumping all the host ports and all the target ports into one zone.

Note: The goal is to have four redundant paths between the hosts and the volumes. This provides good performance as well as protection from single point of failures.

2.2 SAN design and setup

The SAN is primarily responsible for the flow of data between devices. Managing this device communication is of utmost importance for the effective, efficient, and also secure use of the storage network. Zoning plays a key role in the management of device communication. Zoning is used to specify the devices in the fabric that should be allowed to communicate with each other. If zoning is enforced, devices that are not in the same zone cannot communicate.

In addition, zoning provides protection from disruption in the fabric. Changes in the fabric result in notifications (registered state change notifications (RSCNs)) being sent to switches and devices in the fabric. Zoning puts bounds on the scope of RSCN delivery by limiting their delivery to devices when there is a change within their zone.

Tip: As a preferred practice, create the host zones with a single initiator. Do not group multiple initiators from the same host or additional hosts into the same zone. Use one host initiator port per zone.

2.2.1 Zoning

There must be a single zone for each host port. This zone must contain the host port, and one port from each AC2 control enclosure node that the host will need to access. Although there are two ports from each control enclosure node per SAN fabric in a usual dual-fabric configuration, ensure that the host accesses only one of them. Now, if your ESXi host has four (or more) host bus adapters, it takes a little more planning because eight paths are not an optimum number. You must instead configure your zoning (and FlashSystem V9000 host definitions) as though the single host is two or more separate hosts.

Figure 2-2 shows cabling to two fabrics for two ESXi hosts to a IBM FlashSystem V9000. Port pairs (ports with the same number between AC2 control enclosures) on FlashSystem V9000 are cabled alternately between the two fabrics.

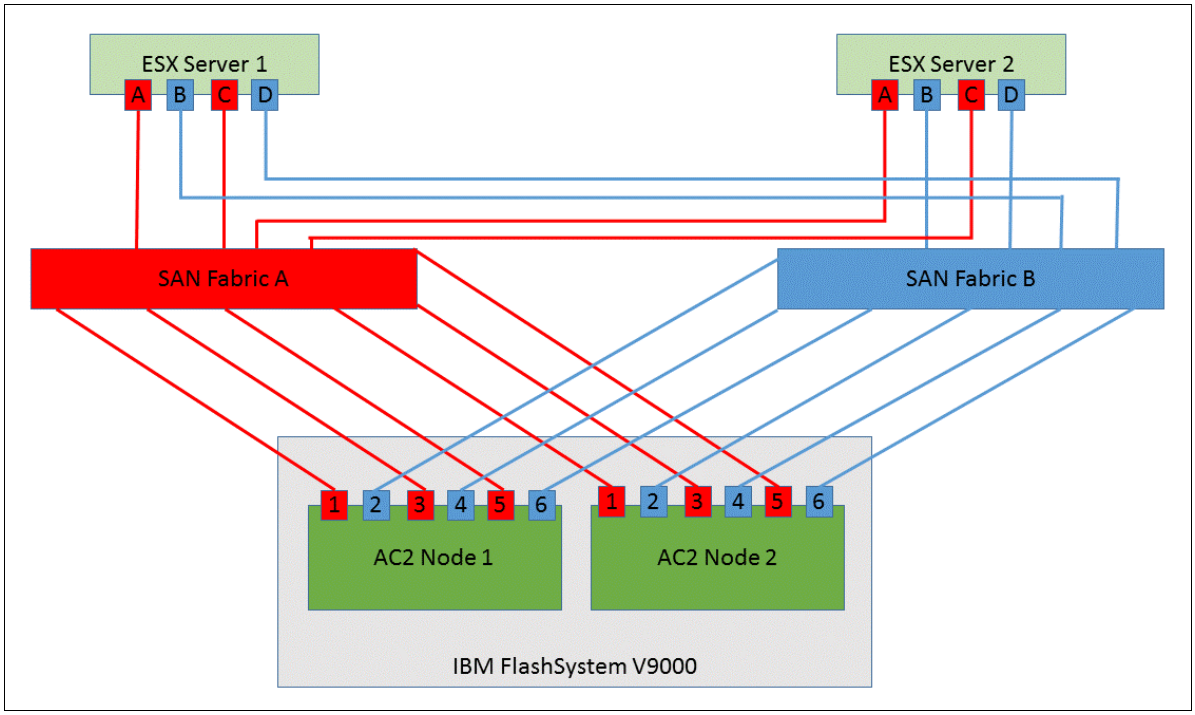


Figure 2-2 SAN cabling for IBM FlashSystem V9000 with 4 ports per ESXi host

When creating zones we want to distribute the expected workload across the IBM FlashSystem v9000 ports. In an environment with many ESXi hosts, low-throughput (IOPS or MB/sec) hosts can share port pairs, and dedicated port pairs can be assigned to high-throughput hosts.

To share port pairs your zoning could look like the red and blue zone examples that follow.

Red zones:

- ESXi Server 1 port A with both AC2 control enclosure nodes port 1s
- ESXi Server one port C with both AC2 control enclosure nodes port 3s
- ESXi Server 2 port A with both AC2 control enclosure nodes port 1s
- ESXi Server 2-port C with both AC2 control enclosure nodes port 3s

Blue zones:

- ESXi Server 1-port B with both AC2 control enclosure nodes port 2s
- ESXi Server 1 port D with both AC2 control enclosure nodes port 4s
- ESXi Server 2-port B with both AC2 control enclosure nodes port 2s
- ESXi Server 2 port D with both AC2 control enclosure nodes port 4s

If you want to have dedicated port pairs for high throughput hosts your zoning could look like the red and blue zone examples that follow.

Red zones:

- ESXi Server 1 port A with both AC2 control enclosure nodes port 1s
- ESXi Server 1-port C with both AC2 control enclosure nodes port 1s
- ESXi Server 2 port A with both AC2 control enclosure nodes port 3s
- ESXi Server 2-port C with both AC2 control enclosure nodes port 3s

Blue zones:

- ESXi Server 1-port B with both AC2 control enclosure nodes port 2s
- ESXi Server 1 port D with both AC2 control enclosure nodes port 2s
- ESXi Server 2-port B with both AC2 control enclosure nodes port 4s
- ESXi Server 2 port D with both AC2 control enclosure nodes port 4s

2.3 VMware mapping to FlashSystem V9000

After the zoning is complete between FlashSystemV9000 and the ESXi hosts you have to create host objects through FlashSystem V9000 management GUI or CLI.

Once the host objects are created, you can use FlashSystemV9000 to configure storage volumes and make assignments, or you can use Spectrum Control Base and the vSphere web plug-in.

Figure 2-3 shows the ESXi host split into two *pseudo-hosts* with 2 ports per host object.

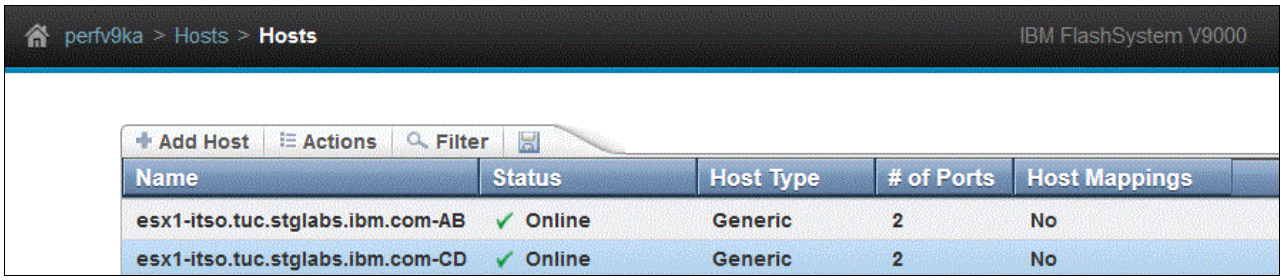


Figure 2-3 ESXi host split into two pseudo-hosts with two ports per host object

A pseudo-host is not a defined function or feature of the management GUI. If you need to define a pseudo-host, you are simply adding another host object. Instead of creating one host object with four WWPNs, you define two hosts with two WWPNs. This is now the reference for the term *pseudo-host*.

During volume assignment, alternate which volume is assigned to one of the *pseudo-hosts* in a round-robin fashion (a pseudo-host is nothing more than another regular host definition in the IBM FlashSystem host configuration. Each pseudo-host contains two unique host WWPNs, one WWPN mapped to each fabric).

Note: Be careful not to share the volume to more than two adapters per host so you do not over-subscribe the number of datapaths per volume, per host.

The following list shows examples of how to map the volumes to hosts:

- Volume1 shared to ESXi_Server1_AB and ESXi_Server2_AB
- Volume2 shared to ESXi_Server1_CD and ESXi_Server2_CD
- Volume3 shared to ESXi_Server1_AB and ESXi_Server2_AB
- Volume4 shared to ESXi_Server1_CD and ESXi_Server2_CD

When creating volumes the default is for FlashSystem V9000 to alternate which AC2 control enclosure node is the preferred node. Multiple volumes are needed to improve aggregate performance and use both AC2 control enclosure node resources in parallel. If benchmarking FlashSystem V9000 you will want to use all the port pairs and have your volumes be a multiple of the ESXi host ports and AC2 control enclosure node ports.

2.4 Multipathing, path selection

There are three general VMwareNative Multipathing plug-ins (NMP) path-selection policies or path-selection plug-ins (PSPs). PSPs are a VMware ESXi host setting that defines a path policy to a logical unit number (LUN). The three PSPs are listed and described in more detail:

- Most Recently Used (MRU)
- Fixed
- Round Robin (RR)

2.4.1 Most Recently Used

During ESXi host boot, or when a LUN will be connected, the first working path will be discovered and is used until this path becomes unavailable. If the active path becomes unavailable, the ESXi host switches to an available path and remains selected until this path fails. It will not return its previous path even if that path becomes available again. MRU is the default PSP for most active/passive storage arrays. MRU has no preferred path even though it is sometimes shown.

2.4.2 Fixed

During the first ESXi host boot, or when a LUN will be connected, the first working path will be discovered and becomes the preferred path. The preferred path can be set and will remain the preferred path from that time on. If the preferred path becomes unavailable, an alternative working path will be selected until the preferred path become available again, then the working path will switch back to the preferred path.

2.4.3 Round Robin

Round Robin path selection policy uses a round robin algorithm to load balance paths across all LUNs when connecting to a storage array. This is the default for VMware starting with ESXi 5.5. It is recommended to use Round Robin for earlier versions, but you will have to explicitly set this. Data can travel only through one path at a time. For active/passive storage arrays, only the paths to the preferred storage array will be used. Whereas for an active/active storage array, all paths will be used for transferring data, assuming that paths to the preferred yes are available.

With ALUA in an active/active storage array, such as FlashSystem V9000, only the optimized paths to the preferred control enclosure node are used for transferring data and Round Robin will cycle only through those optimized paths. You should have half the LUNs preferred by one control enclosure node, and the other half preferred by the other control enclosure node.

Round Robin path switching

By default, the Round Robin policy switches to a different path after 1000 IOPS. Lowering this value can, in some scenarios, drastically increase storage performance in terms of latency, MBps, and IOPS. To increase storage throughput, you can change the value from 1000 to a lower value. Test and benchmark it in your test environment before making adjustments to your production environment.

Example 2-1 shows how to change the default Round Robin policy to switch after 1 IOPS.

Example 2-1 Changing the Round Robin policy to switch after 1 IOPS instead of 1000 IOPS

```
esxcli storage nmp psp roundrobin deviceconfig set --type=iops --iops=1 --device
naa.xxx

# where .xxx matches the first few characters of your naa IDs.

#Use the following command to issue a loop of the previous command for each #LUN
attached to the ESXi server:

for i in `esxcfg-scsidevs -c |awk '{print $1}' | grep naa.xxx`; do esxcli storage
nmp psp roundrobin deviceconfig set --type=iops --iops=1 --device=$i; done

Use the following command to set iops to 1 for all Storwize LUNs connected in the
future. Reboot is required for already connected LUNs.

esxcli storage nmp satp rule add -s "VMW_SATP_ALUA" -V "IBM" -M "2145" -P
"VMW_PSP_RR" -O "iops=1"
```

For more information about adjusting the Round Robin IOPS limit from default 1000 to 1, see the VMware Knowledge Base:

http://kb.vmware.com/kb/2069356

See the IBM Systems and Technology Technical White Paper titled “Drive performance in VMware environments with IBM FlashSystem” for more information about driving performance in VMware environments with IBM FlashSystem:

http://ibm.co/1MXqrvp

2.4.4 ALUA

Asymmetric Logical Unit Access (ALUA) is an access method that allows a storage array to report its port state. An active/active (Asymmetrical active/active) storage array can report what its preferred control enclosure node is, or an active/passive array can report which control enclosure is active and owns the LUN.

With asymmetrical active/active storage arrays (such as FlashSystem V9000) and a round-robin path policy, all paths actively are used for transmitting data if you are accessing LUNs with preferred control enclosure nodes to all the control enclosures. Half of your path goes to the preferred control enclosure node, and the other half go to the not-preferred control enclosure node for half of the LUNs. The other half of the LUNs go to prefer the other control enclosure node.

For FlashSystem V9000, when volumes are created, they are automatically assigned a preferred control enclosure node. The preferred control enclosure node is the one that will destage writes from cache to the flash memory.

With ALUA, the host knows what the optimized paths are and what the non-optimized paths are and uses only the optimized paths for Round Robin cycles.

Note: It is possible to use all paths, both preferred and non-preferred for LUN. To do this run the following command:

```
esxcli storage nmp satp rule add -s "VMW_SATP_ALUA" -V "IBM" -M "2145" -P
"VMW_PSP_RR" -O "useANO=1"
```

If you have multiple LUNs assigned to a ESXi host and the LUNs are evenly split between control enclosure nodes, then this will likely not improve performance, and may have some read hit penalty for certain workloads.

3

VMware configurations for FlashSystem V9000

This chapter describes storage-related configurations in VMware vSphere for IBM FlashSystem V9000. It includes the following sections:

3.1 ESXi host offloading with VAAI and FlashSystem V9000

vStorage APIs for Array Integration (VAAI) is a group of APIs that were introduced with vSphere 4.1. They allow for certain storage operations from the ESXi host to be offloaded to a supported storage array.

VAAI is also known as *hardware acceleration*. One example is the copy of virtual machine files. Without VAAI, the VM files are copied through the host, but with VAAI, the data is copied within the same storage array. VAAI increases the ESXi performance because the SAN fabric is not used, so fewer CPU cycles are needed because the copy does not need to be handled by the host.

There were different implementations of the VAAI block primitives in vSphere 4.1, but starting with vSphere 5.0, all of the primitives have been ratified by T10.

IBM FlashSystem V9000 supports following VAAI primitives (at the time of writing):

- Automatic test and set (ATS)
- Extended Copy (XCOPY)
- Write_Same

3.1.1 Atomic test and set

Atomic Test and Set (ATS) also known as *hardware-assisted locking*, intelligently relegates resource access serialization down to the granularity of the block level during VMware metadata updates. It does this rather than using a mature SCSI2 reservation, which serializes access to the adjacent ESXi hosts with a minimum scope of an entire LUN.

Important: VMware File System (VMFS Version 3 or later) uses ATS in a multi-tenant ESXi cluster that shares capacity within a VMFS datastore by serializing access only to the VMFS metadata associated with the VMDK or file update needed through an on-disk locking mechanism. As a result, the function of ATS is identical whether implemented to grant exclusive access to a VMDK or to another file.

ATS is a standard T10 SCSI command with opcode **`0x89 (COMPARE AND WRITE)`**. The ATS primitive has the following advantages where LUNs are used by multiple applications or processes at one time:

- Significantly reduces SCSI reservation contentions by locking a range of blocks within a LUN rather than issuing a SCSI reservation on the entire LUN
- Enables parallel storage processing
- Reduces latency for multiple ESXi hosts accessing the same LUN during common vSphere operations involving VMFS metadata updates, including these updates:
  - VM, VMDK, or template creation or deletion
  - VM snapshot creation or deletion
  - Virtual machine migration and storage vMotion migration
  - Virtual machine power on or off
  - Writes to thin-provisioned and thick, lazy-zeroed virtual disks
- Increases cluster scalability by greatly extending the number of ESXi hosts and VMs that can viably reside simultaneously on a VMFS datastore

3.1.2 Extended copy

Extended copy (XCOPY), also known as a *hardware-accelerated move*, offloads copy operations from VMware ESXi to the IBM storage system (IBM FlashSystem V9000 in the examples for this paper). This process allows for rapid movement of data when performing copy or move operations within the IBM storage system. It reduces CPU cycles and host bus adapter (HBA) workload of the ESXi host.

Similarly, it reduces the volume of traffic moving through the SAN when deploying a virtual machine. It does so by synchronizing individual VM-level or file system operations, including clone and migration activities, with the physical storage level operations at the granularity of individual blocks on the devices. The potential scope in the context of the storage is both within and across LUNs. This command has the following benefits:

- Expedites copy operations including these tasks:
  - Cloning of virtual machines, including deploying from template
  - Migrating virtual machines from one datastore to another (storage vMotion)
- Minimizes host processing and resource allocation
  - Copies data from one LUN to another without reading and writing through the ESXi host and network
- Reduces SAN traffic

The SCSI opcode for XCOPY is **0x83**.

3.1.3 WRITE_SAME

Block Zeroing, Write_Same (Zero), or *hardware-accelerated initialization* use the **WRITE_SAME 0x93** SCSI command to issue a chain of identical write transactions to the storage system, thus almost entirely eliminating server processor and memory use by eliminating the need for the host to execute repetitive identical write transactions. It also reduces the volume of host HBA and SAN traffic when performing repetitive block-level write operations within virtual machine disks to the IBM storage system.

Similarly, it allows the IBM storage system to minimize internal bandwidth consumption. For example, when provisioning a VMDK file with the `eagerzeroedthick` specification, the Zero Block's primitive issues a single **WRITE_SAME** command that replicates zeroes across the capacity range that is represented by the difference between the VMDK's provisioned capacity and the capacity consumed by actual data. The alternative requires the ESXi host to issue individual writes to fill the VMDK file with zeros.

The scope of the Zero Block's primitive is the VMDK creation within a VMFS datastore. Therefore, the scope of the primitive is generally within a single LUN on the storage subsystem, but it can potentially span LUNs backing multi-extent datastores.

Block Zeroing offers the following benefits:

- Offloads initial formatting of Eager Zero Thick (EZT) VMDKs to the storage array
- Assigns zeros to large areas of storage without writing zeros from the ESXi host
- Speeds up creation of new virtual machines
- Reduces elapsed time, server workload, and network workload

Note: In the case of thin-provisioned volumes, FlashSystem V9000 further augments this benefit by flagging the capacity as having been “zeroed” in metadata without the requirement to physically write zeros to the cache and the disk. This implies even faster provisioning of the `eagerzeroed` VMDKs. See 6.2.1, “Using FlashSystem V9000 Real-time Compression with VMware” on page 78 for details about how to reclaim space with compressed volumes.

3.1.4 FlashSystem V9000 and VAAI

To verify which VAAI operations are supported by your storage device, issue a command as shown in Example 3-1.

Example 3-1 Verify device VAAI support where “naa.xxx” stands for device identifier

```
[root@ESX1-ITSO:~] esxcli storage core device vaai status get -d naa.xxx
naa.xxx
   VAAI Plugin Name:
   ATS Status: supported
   Clone Status: supported
   Zero Status: supported
   Delete Status: unsupported
```

You can verify and change your VAAI settings in host **Advanced Settings** (Figure 3-1). A value of **1** means that the feature is enabled. If the setting is host-wide, it will be enabled if connected storage supports it.

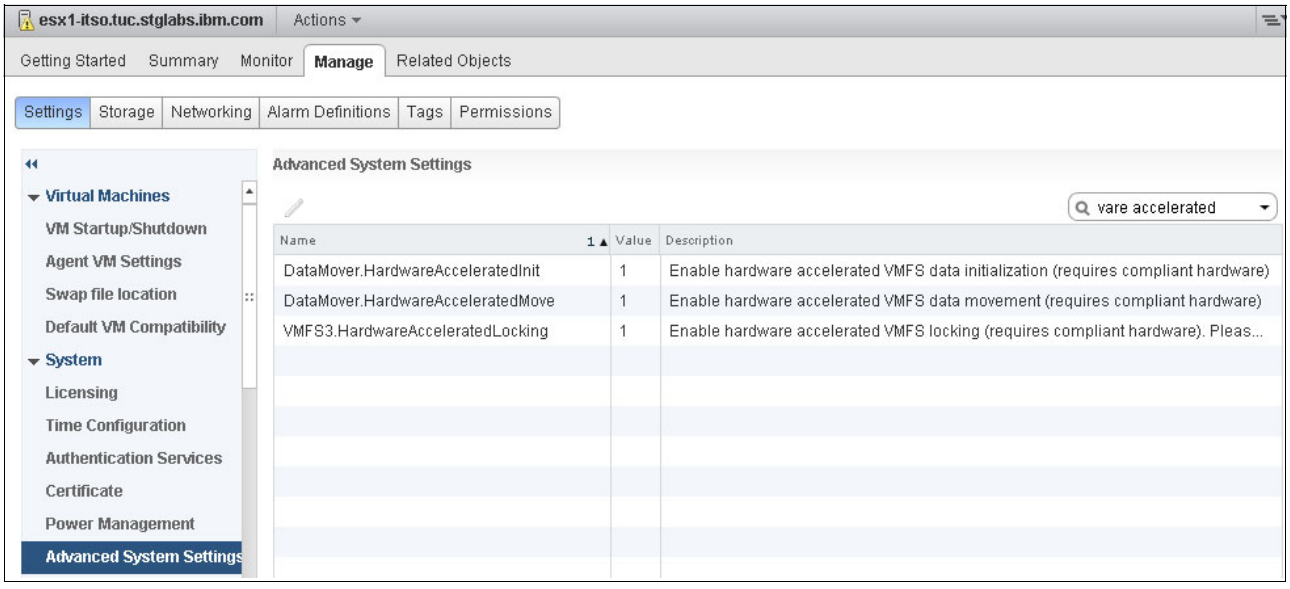


Figure 3-1 VAAI settings

Table 3-1 lists the VAAI settings and parameter descriptions.

Table 3-1 Parameters description

Parameter name	Description
`DataMover.HardwareAcceleratedInit`	Zero Blocks/Write Same
`DataMover.HardwareAcceleratedMove`	Clone Blocks/XCOPY
`VMFS3.HardwareAcceleratedLocking`	Atomic Test & Set (ATS)

It is advisable to keep all VAAI operations enabled when using FlashSystem V9000 so that as much work as possible is offloaded to storage.

If you are offloading Data Mover (XCOPY, WRITE_SAME) operations to storage, there is no effective way to set the I/O limitation for those tasks. Therefore, consider scheduling the big migrations or provisioning outside of business hours, because another system running from backend storage can be affected by this procedure. This is especially important if you are using FlashSystem V9000 to manage external storage, because the external storage might not be as powerful as FlashSystem V9000 itself.

Hardware offloading is not used if following occurs:

- The source and destination VMFS datastores have different block sizes. (This can happen when using exiting VMFS3 datastores.)
- The source disk is RDM and the destination is non-RDM (regular VMDK).
- The source VMDK type is `eagerzeroedthick` but the destination VMDK type is `thin`.
- The source or destination VMDK is in a sparse or hosted format.
- The source virtual machine has a snapshot.
- The logical address and transfer length in the requested operation are not aligned to the minimum alignment required by the storage device. (All datastores created with the vSphere Web Client are aligned automatically.)
- The VMFS has multiple LUNs or extents, and they are on different arrays.
- Hardware cloning between arrays, even within the same VMFS datastore, does not work. (This is not the case if arrays are managed by FlashSystem V9000 using external virtualization.)

Notes:

You may decide to increase a block size, which will be processed by storage globally. Although we do not recommend changing default values, you might notice a small improvement in the performance during Data Mover operations (typically around 10%).

The default value is 4096 KB, and the maximum value is 16384 KB, as this example shows:

```
# esxcfg-advcfg -s 16384 /DataMover/MaxHWTransferSize
```

This change is global and will affect all your VAAI enabled storage devices connected to the ESXi, therefore it can have unpredictable impact on different storage arrays.

Important: When this paper was written, VMware vSphere 5.5 Update 2 and vSphere 6.0 introduced a new method for a datastore heartbeat using Atomic Test and Set for heartbeat I/O. Due to the low timeout value for heartbeat I/O using ATS, this can lead to host disconnects and application outages if delays of eight seconds or longer are experienced in completing individual heartbeat I/Os on backend storage systems or the SAN infrastructure.

To roll back to the traditional heartbeat method, issue the following command on each ESXi connected to IBM FlashSystem V9000 storage or other IBM Storwize® based arrays:

```
esxcli system settings advanced set -i 0 -o /VMFS3/useATSForHBOnVMFS5
```

This reversion of VMFS heartbeat activity is preferred, rather than globally disabling VAAI or ATS when using applicable storage systems.

For more details and updates regarding this issue, follow these web pages:

Host Disconnects Using VMware vSphere 5.5.0 Update 2 and vSphere 6.0

http://www.ibm.com/support/docview.wss?uid=ssg1S1005201

Enabling or disabling VAAI ATS heartbeat

http://kb.vmware.com/kb/2113956

3.2 Storage I/O Control

Storage I/O Control (SIOC) is used to control the I/O use of a virtual machine and to gradually enforce the predefined I/O share levels.

With vSphere 5.1, a new automated method was introduced to detect latency congestion thresholds for SIOC. It is achieved by injecting I/O to device during idle periods to detect maximum throughput level of the device and sets its value to latency at 90% of the maximum throughput (see Figure 3-2 on page 27). This is easiest way for you to assure that you will get the best performance from your device without guessing which threshold is optimal for your storage. You can also use the old method by manually specifying the value yourself.

The lowest value that you can set is 5 ms. Typically, you will not reach this value with FlashSystem V9000, because it runs IOPS in microseconds. However, if the specified latency is reached, SIOC acts to reduce latency to acceptable levels.

SIOC has the following limitations and requirements:

- SIOC-enabled datastores can be managed only when attached to hosts that are managed by the same vCenter Server.

  For optimal effectiveness, use dedicated array pools for SIOC managed hosts, not polls shared with other systems.
- Raw Device Mapping (RDM) disks are not supported.
- SIOC does not support datastores that have more than one volume (multiple extents).
- SIOC is only available with the vSphere Enterprise Plus edition.

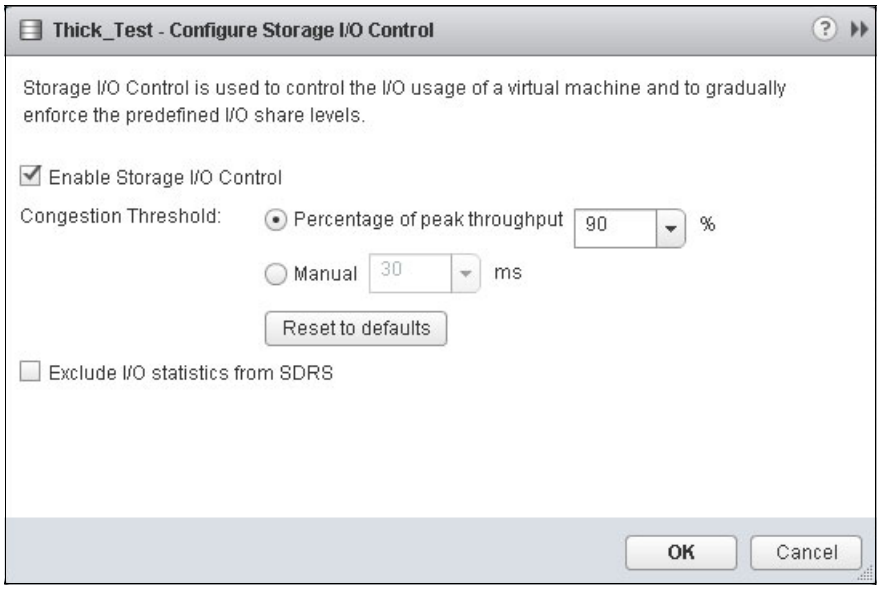


Figure 3-2 SIOC settings

Tips:

Using automatic congestion threshold detection is advisable in environments with the Easy Tier function enabled, because the injector feature of SOIC can reduce its effectiveness.

Setting a fixed value too low is not recommended, and careful planning for needs is essential.

3.3 Storage DRS

Storage DRS (SDRS) consists of vSphere Enterprise plus features that can be used to balance storage space and load between datastores by using Storage vMotion to migrate VMs. Depending on your environment, you can automate these tasks or decide to be notified and implement actions yourself.

As in Storage IO Control, the minimum latency threshold that you can set to trigger possible virtual machine migration or recommendations to balance load is 5 ms. This is highly unlikely for FlashSystem V9000, because its response times are typically in microseconds unless there is a big resource contention.

You can see the SDRS latency threshold setting in Figure 3-3.

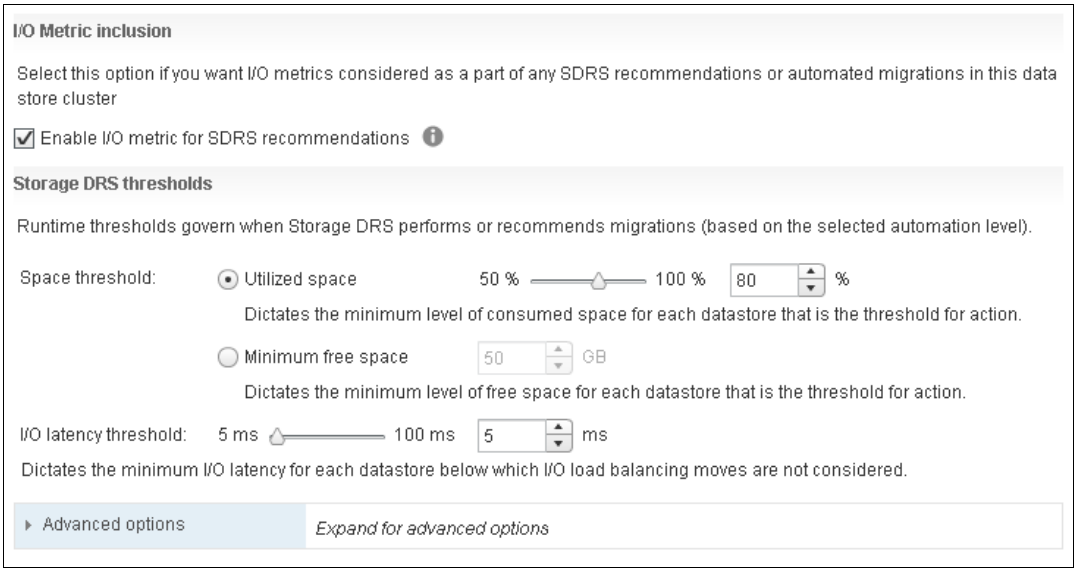


Figure 3-3 SDRS settings

Tips:

Try to group datastores into SDRS clusters with similar backing storage, but avoid mixing traditional spin drives and flash disk datastores in one cluster.

In Easy Tier environments, consider turning off **I/O metric for SDRS recommendation** or set it to high value. Alternatively, you can set **I/O balance automation level** to **manual**.

3.4 Marking a device as flash

ESXi does not automatically recognize that volumes provisioned from FlashSystem V9000 are flash-based, but you may decide to have the volumes recognized as flash-based. For example, some guest operating systems can identify virtual disks that reside on flash-based datastores.

You can do this by marking your device as *Flash* from the VMware Web Client under the host **Storage Devices** tab. Select the device, and then click the **F** icon, as shown in Figure 3-4 on page 29.

Important: The device must not be in use. Therefore, if you have already created your datastore, you need to unmount it first.

Repeat this task for each host that is connected to the volume.

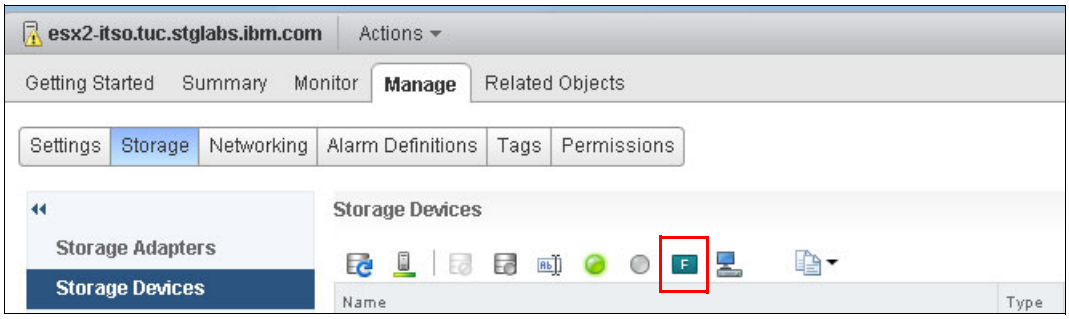


Figure 3-4 Mark as Flash

4

Integrated management

Many integration points exist in VMware, including vCenter, vRealize Orchestrator, management capabilities in both traditional vSphere client and the vSphere Web Client, or VMware vSphere APIs for Storage Awareness (VASA).

This chapter provides information about how to use the IBM Storage Enhancements for the VMware vSphere Web Client plug-in to create and manage IBM FlashSystem V9000 volumes in pools that have been attached to the IBM Spectrum Control Base server.

This chapter covers the following topics:

4.1 Introduction

IBM FlashSystem V9000 storage addresses business performance and capacity requirements, but a viable independent software vendor integration that can be used by a software-defined environment (SDE) remains a challenge. The IBM Spectrum Control Base server addresses this requirement with automation, elasticity, capabilities of storage as a service, and operations management for storage management.

4.2 IBM Spectrum Control Base Edition server

Spectrum Control Base Edition is a centralized server system that consolidates a range of IBM storage provisioning, automation, and monitoring solutions through a unified server platform. The following solution components are included in the IBM Spectrum Control Base package:

- IBM Storage Enhancements for VMware vSphere Web Client (vWC)
- IBM Storage Provider for VMware VASA, v1.0 and v.2.0
- IBM Storage Plug-in for VMware vCenter Orchestrator (vCO)
- Storage Management Pack for VMware vCenter Operations Manager (vCOps)

For more information, see the IBM Spectrum Control Base Edition release notes and user guide:

http://ibm.co/1W9L1mJ

You can download the latest version of Spectrum Control Base Edition at no charge from the IBM Fix Central web page:

http://www.ibm.com/support/fixcentral/

IBM Spectrum Control Base Edition can be managed through a standard web browser and a graphical user interface (GUI) or through terminal and a command-line interface (CLI).

Note: At the time of the writing, VMware vCenter Operations Manager and VMware vRealize Orchestrator are not supported with version 2.1.1 of the Spectrum Control Base Edition. They are currently supported only by the IBM XIV storage array. However, VASA 2.0 Virtual Volumes (VVols) are currently supported with FlashSystem V9000.

After Spectrum Control Base Edition is installed, different tasks are required before the server can become fully operational. These tasks are described in the topics that follow.

4.2.1 Register FlashSystem V9000 with the Spectrum Control Base server

All IBM storage systems that provide storage resources to your VMware platforms must be defined as storage arrays on IBM Spectrum Control Base Edition. To access the storage array management options, go to the Storage Systems pane of the Spectrum Control GUI.

Spectrum Control Base Server uses a single login account to access all of the IBM storage arrays. It is advisable to create a unique storage administrator account for traceability in the environment.

For this setup example, we created the user `ibmscb`, as shown in Figure 4-1.

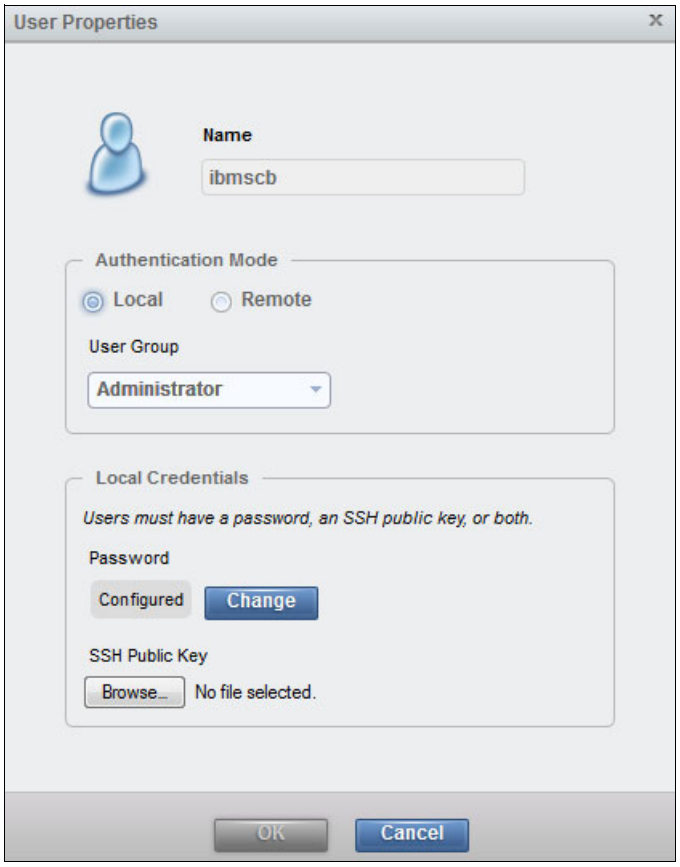


Figure 4-1 FlashSystem V9000 user for Spectrum Control Base server

Entering the storage array credentials

Next, we configure the Spectrum Control Base with the `ibmscb` user (see Figure 4-1). The storage array credentials are used to connect to the IBM storage system or systems, which your VMware platforms use for storage provisioning.

The IBM Spectrum Control Base GUI provides an intuitive, easy-to-use, browser-based interface for managing IBM storage resources. Simply follow these steps:

1. Log in to the Spectrum Control Base GUI.
2. Click the **Settings** button in the upper-left corner of the pane, and select **Storage credentials** from the Settings menu.

3. The Storage Credentials dialog window (see Figure 4-2) shows the currently defined storage array user name, or you can specify a new storage array user name.

   Enter the user name and password that was defined in FlashSystem V9000, and click **Apply.**

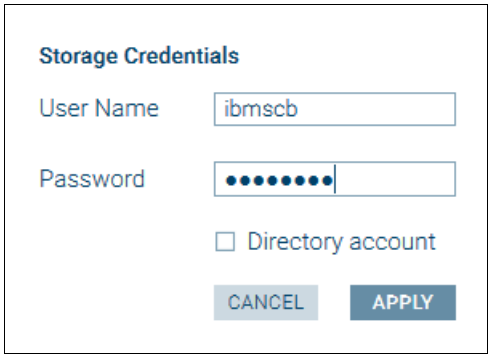


Figure 4-2 FlashSystem V9000 user name for Spectrum Control Base Server

4. The next step is to add FlashSystem V9000 as a storage array. Spectrum Control Base will use the credentials that you have set to access FlashSystem V9000.

   In the Storage Systems pane, click the **Add** button. When the Add New Array dialog window shown in Figure 4-3 opens, enter the IP and host name of FlashSystem V9000 and specify the type as **FlashSystem**.

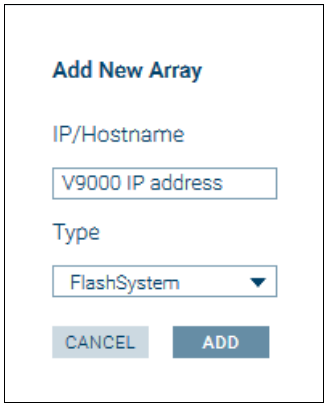


Figure 4-3 Adding FlashSystem V9000 as new storage array

The newly added FlashSystem V9000 is displayed in the Storage Systems pane, as shown in Figure 4-4.

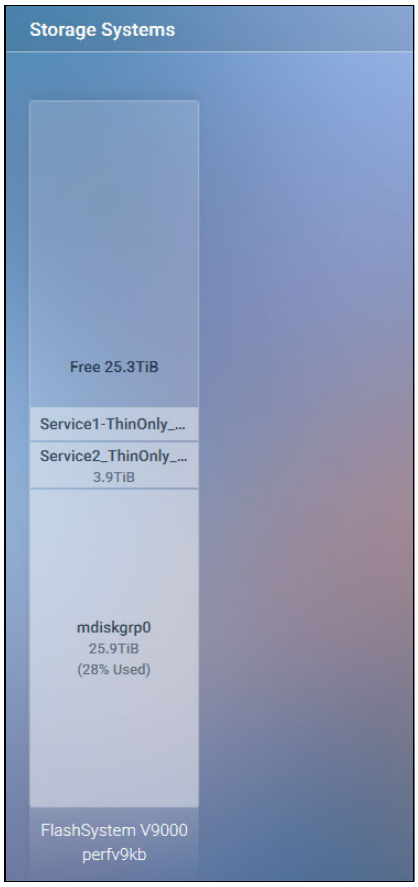


Figure 4-4 FlashSystem V9000 in the Storage Systems pane of the Spectrum Control Base GUI

4.2.2 Managing integration with the vSphere web client

Before you can use the IBM Storage Enhancements for VMware inside the vSphere Web Client, you must define the vCenter servers on the Spectrum Control Base server. Specify the vCenter servers for which you want to provide storage resources. Then you can attach storage services to each vCenter server you have defined in Spectrum Control Base. Figure 4-5 shows the Add New vCenter Server dialog box.

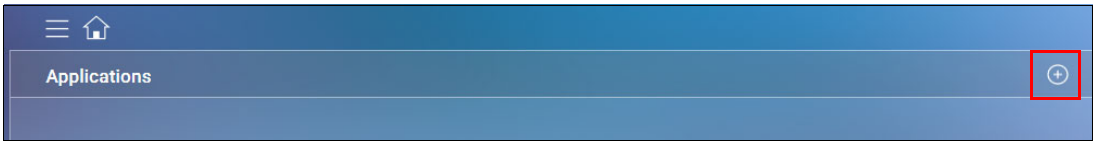


Figure 4-5 Main menu to add the vCenter server

Complete these steps to define the vCenter servers on the Spectrum Control Base server:

1. Click the **Add** (plus sign) icon in the Spectrum Control Base Applications pane.
2. When the Add New vCenter Server dialog window opens (see Figure 4-6), enter the IP address or host name of the vCenter server, as well as the user name and password for logging in to that vCenter server.

3. If the provided IP address and login credentials are valid, the vCenter server will be displayed in the Applications pane.

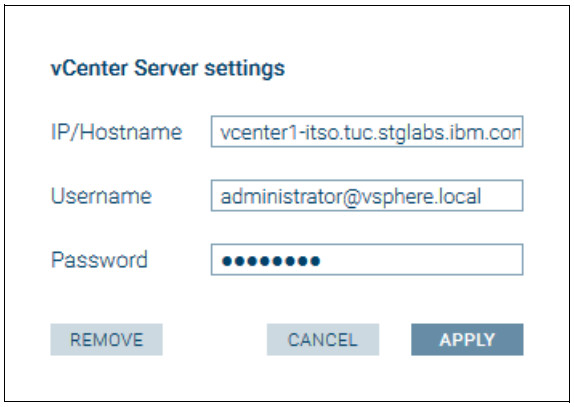


Figure 4-6 Adding the vCenter Server to Spectrum Control Base

4.2.3 Managing storage spaces and services

After defining FlashSystem V9000 in the Spectrum Control Base, you must add virtual storage elements: spaces and services. Spectrum Control administers storage at the virtual level, using spaces and services. This simplifies storage provisioning and enables users to allocate their own storage resources to suit requirements of a specific VM.

Storage spaces are added and managed by using the Manage Spaces option in the Settings menu. You can also add a new space from the drop-down menu of the Default Space tab in the Spaces/Services pane. Storage services contain one or more physical storage pools. In addition to storage capacity, a service has a set of capabilities that define the storage quality, such as thin or thick provisioning, compression, snapshots, and encryption.

Note: A Spectrum Control Base *Service* must be attached to a Spectrum Control Base *Resource*, which can be either a pool or a child pool in FlashSystem V9000.

Storage services are added and configured in the Spaces/Services pane of the Spectrum Control GUI, as shown in Figure 4-7.

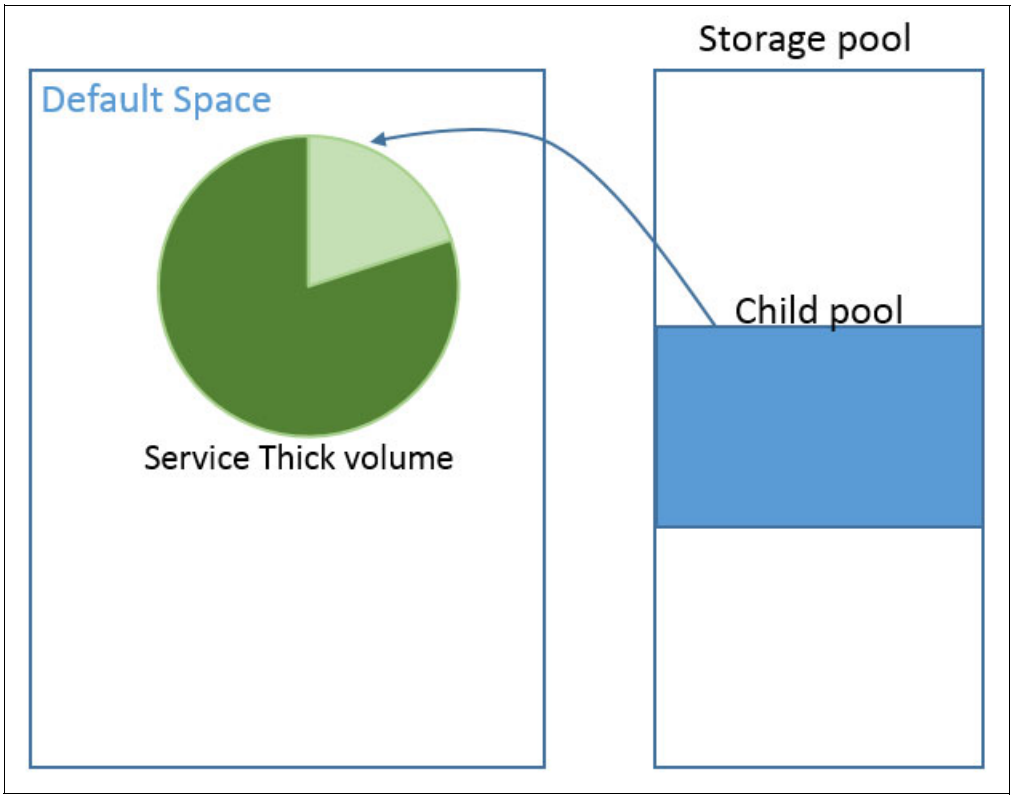


Figure 4-7 Service in default space attached to a storage child pool

Storage services contain one or more physical storage pools. In addition to storage capacity, a service has a set of capabilities, defining the storage quality, such as thin/thick provisioning, compression, snapshots and encryption.

Storage services are added and configured via the Spaces and Services pane of the Spectrum Control GUI, as shown in Figure 4-8.

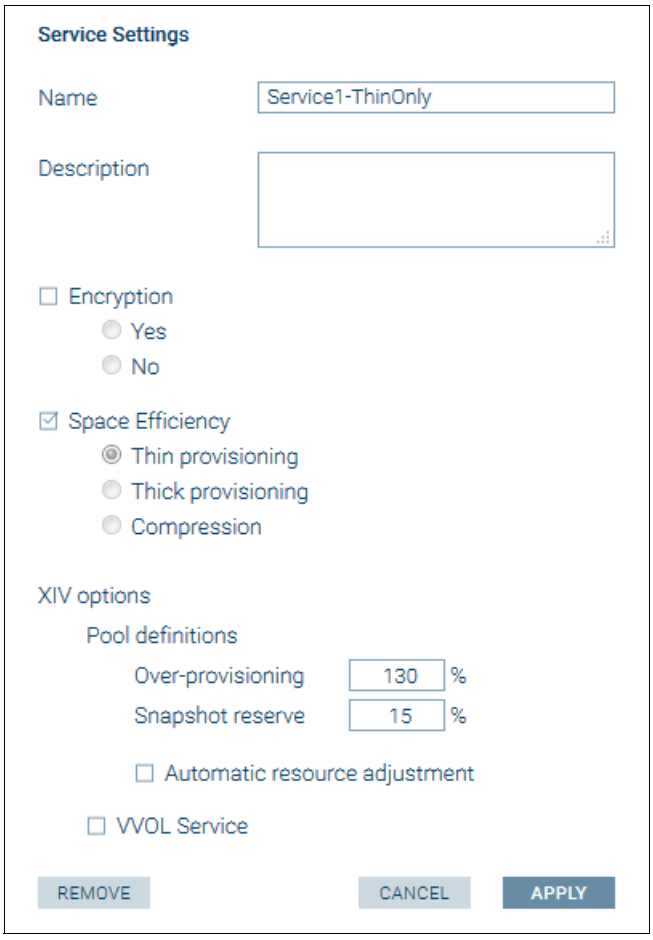


Figure 4-8 Adding a service in Spectrum Control Base Edition

Table 4-1 describes the main fields.

Table 4-1 Storage services main fields

Parameter	Description and values
Encryption	Enables or disables encryption for the service. If disabled, you can attach any storage resource (encrypted or not) to the service. For the FlashSystem V9000 it can be enabled.
Space efficiency	Enables storage space efficiency features for the service. When enabled, you can configure the service to be thick- or thin-provisioned or make it use IBM Real-time Compression. If it is disabled, you can attach any storage resource (thin, thick, compressed or not) to the service.
Over-provisioning	Percentage of over-provisioned storage space on the service, defining the ratio between logical and physical storage capacity. For example, when configured to 100, it sets a 1:1 ratio between the two values, but a value of 200 defines the logical capacity (soft size) to be twice the physical capacity.

Parameter	Description and values
Snapshot reserve	Percentage of storage space on the service reserved for snapshots.

Note: A Spectrum Control Base *Service* must be attached to a Spectrum Control Base *Resource* which can be either a pool or a child pool in the FlashSystem V9000.

There are often cases where you want to sub-divide a storage pool (or managed disk group) but maintain a larger number of mdisks in that pool. Child pools are logically similar to storage pools, but allow you to specify one or more sub divided child pools. Quotas and warnings can be set independently per child pool in the FlashSystem V9000. This allows to manage applications differently from the storage perspective.

Also, you must attach the vCenter server to any storage service that you want to use for volume management operations on the vSphere Web Client side. Log in to Spectrum Control Base and follow these steps:

1. In the Applications pane, click the vCenter server to which you want to attach one or more services.

2. In the Services pane, select the storage space from which you want to choose storage services. The services available on the selected storage space are immediately displayed. Figure 4-9 on page 39 shows a service for thin-provisioned volumes and one for fully allocated volumes.

3. Click a service that you want to attach to the vCenter server. The service color changes to green to indicate the successful attachment, as illustrated in Figure 4-9 on page 39.

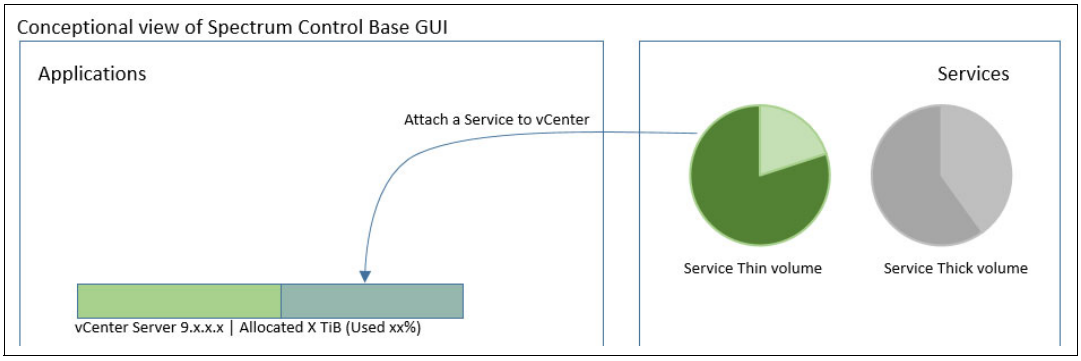


Figure 4-9 Attaching a service to a vCenter server in the Spectrum Control Base GUI

4.3 Provisioning FlashSystem V9000 volumes using VMware

The IBM *Storage Enhancements* for the VMware vSphere Web Client plug-in is used to create and manage FlashSystem V9000 volumes in storage pools that have been attached to the Spectrum Control Base server. As Figure 4-10 shows, the plug-in enables you to create new FlashSystem V9000 volumes directly from the vCenter Web Client.

Through the vSphere Web Client plug-in, select **ITSO1** (1) → **Actions** (2) → **All IBM Storage Enhancements for VMware vSphere Web Client Actions** (3) → **Create New Volume** (4). The numbers indicate the sequence as indicated by the numbers shown in Figure 4-10.

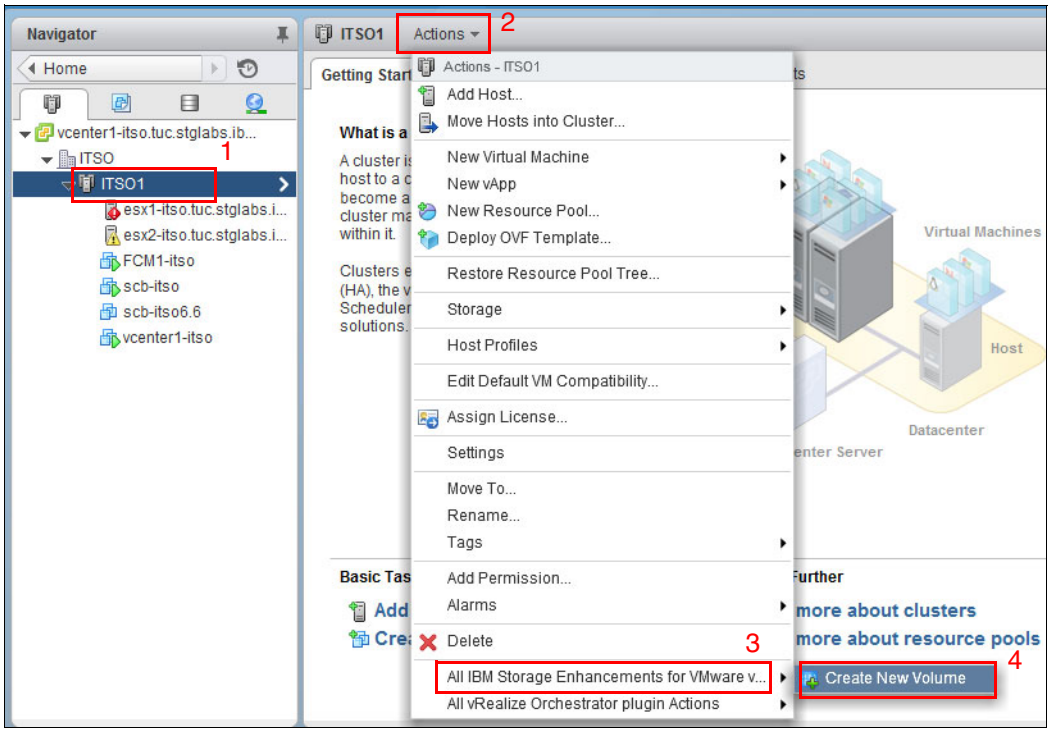


Figure 4-10 Creating a volume with the VMware vSphere Web Client plug-in

This action opens the Create New Volume window (Figure 4-11). In the Create New Volume window, the Host Mappings drop-down menu shows that the ITSO cluster is the selected host that will be mapped to the newly created volumes. In this example, two 500 GB volumes will be created and mapped to the ITSO cluster.

The volume names are `ITSO_vol_1` and `ITSO_vol_2`, and they are created by *Service1* as thin-provisioned volumes in a child pool on FlashSystem V9000. The number in the brackets sequentially increases by one, as highlighted in Figure 4-11.

Additional volume properties can be selected when the volumes are created by using the Storage Enhancements for VMware vSphere Web Client:

- Enable Thin Provisioning
- Enable Data Compression
- Enable Virtual Disk Mirroring

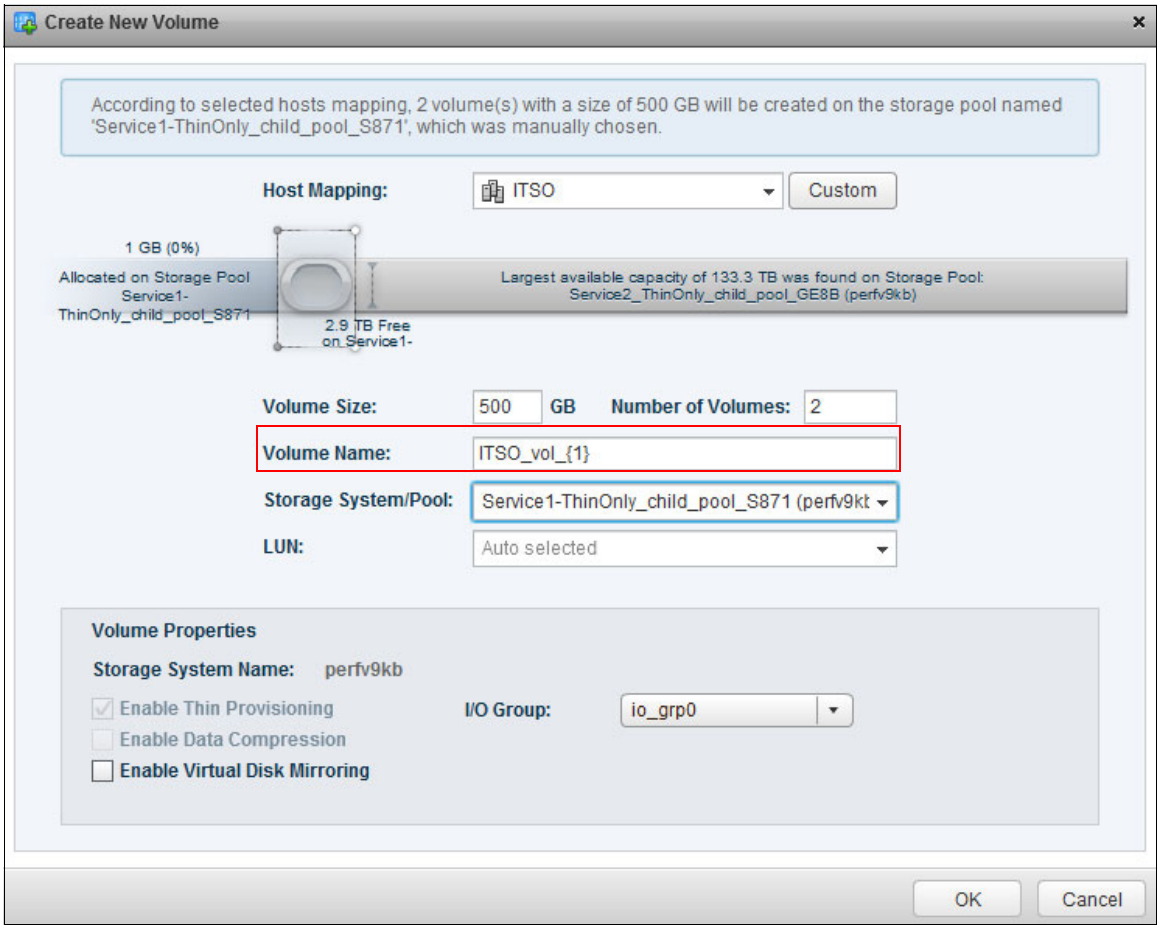


Figure 4-11 Create new volumes using VMware

Figure 4-12 shows the new FlashSystem V9000 volumes created and mapped directly from the vSphere Web Client, without the need for the VMware administrator to access FlashSystem V9000 GUI or command-line interface directly. This view lists the storage array where the volumes are located. The names and size of the volumes are also in this view.

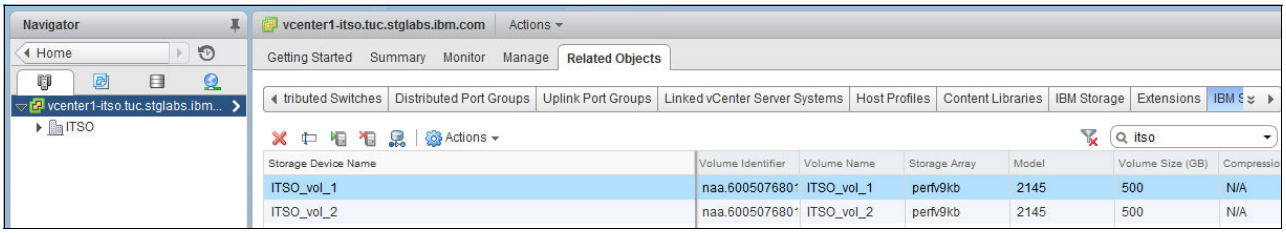


Figure 4-12 FlashSystem V9000 volumes created and mapped from the vSphere Client

The result of the volume created by the VMware vSphere Web Client plug-in is also visible in FlashSystem V9000 GUI, as shown in Figure 4-13. This particular view of the GUI lists the two volumes that were defined on FlashSystem V9000 by using the vSphere Web Client plug-in.

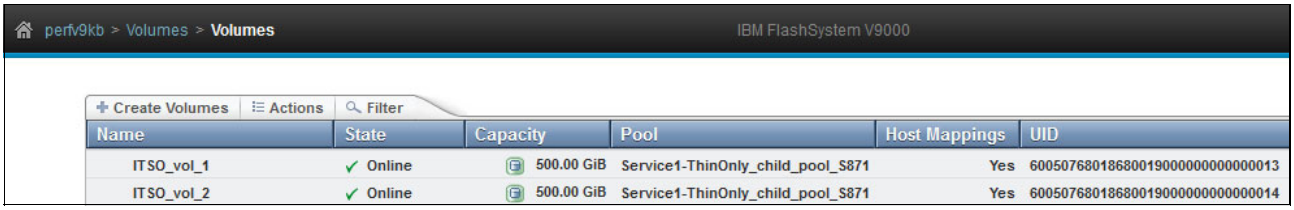


Figure 4-13 FlashSystem V9000 GUI listing of the volumes

4.4 Expanding a volume using VMware

The IBM FlashSystem V9000 supports dynamically expanding the size of a virtual disk, and VMFS volumes can be extended while virtual machines are running. First, you must extend the volume on FlashSystem V9000, and then you can extend the VMFS volume.

1. Through the vSphere Web Client, make the following selections to list all attached volumes, as shown in Figure 4-14: **ITSO** (1) → **Related Objects** (2) → **IBM Storage Volumes** (3).
2. Right-click the volume that will be increased in size, and select the action **Extend** (2).

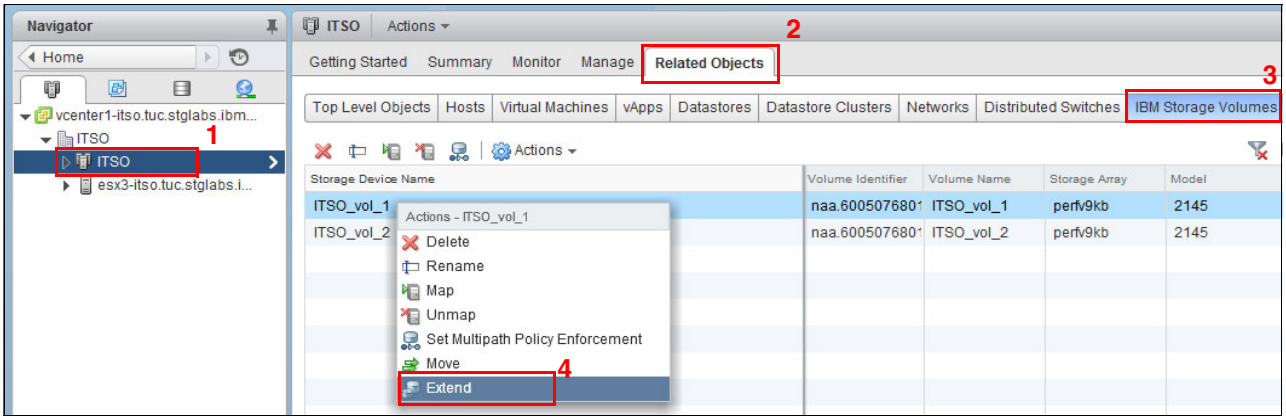


Figure 4-14 Extending a volume using the vSphere Web Client

Figure 4-15 shows the wizard that allows you to specify a new size for FlashSystem V9000 volume. In this example, the volume is increased from 500 GiB to 700 GiB.

Note: A vol*ume*, which is defined to be in a FlashCopy, Metro Mirror, or Global Mirror mapping on FlashSystem V9000, cannot be expanded. Therefore, the FlashCopy, Metro Mirror, or Global Mirror on that volume must be deleted before the volume can be expanded.

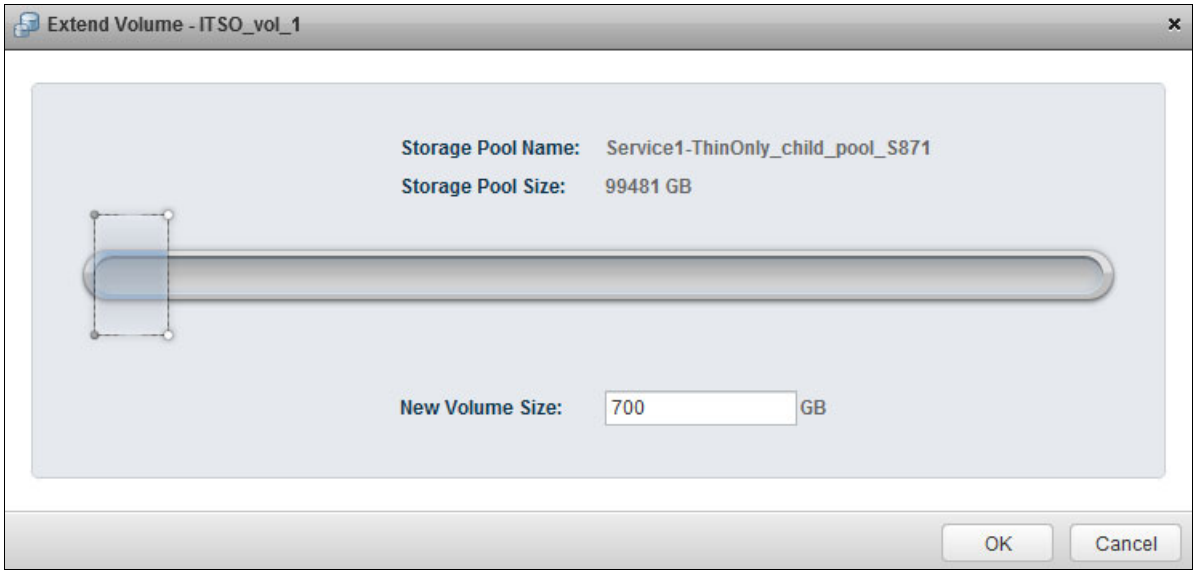


Figure 4-15 Extending the volume size with the vSphere Web Client

4.5 Additional volume functions using VMware

In addition to creating and expanding FlashSystem V9000 volumes with the vSphere Web Client plug-in, the following functions are available, as shown in Figure 4-16 on page 44:

- Delete

  This action will delete the volume from FlashSystem V9000. The action will fail, if the volume still contains a datastore or is a raw-mapped LUN.

- Rename

  Renaming a FlashSystem V9000 volume is a logical action that does not have any physical effect on the volume or its logical connections. Renaming a volume also changes its displayed name in the vSphere environment.

- Map

  This action maps the volume to another ESXi server of the vSphere cluster

- Unmap

  This action unmaps a volume from the ESXi server.

  Note: A volume (LUN) must remain mapped to at least one ESXi host. Otherwise, you cannot view the volume or perform any actions on it from vSphere Web Client.

- Set Multipath Enforcement

  Generally, you do not need to change the default multipathing settings your host uses for a specific storage device. However, if you want to make any changes, you can use this action to modify a path selection policy and specify the preferred path for the Fixed policy. You can also use this dialog window to change multipathing for SCSI-based protocol endpoints.

- Move

  This action can be used to move a volume from one storage pool to another storage pool while the volume is still in use. This action is useful for maintenance or migration purposes.

- Extend

  See 4.4, “Expanding a volume using VMware” on page 42.

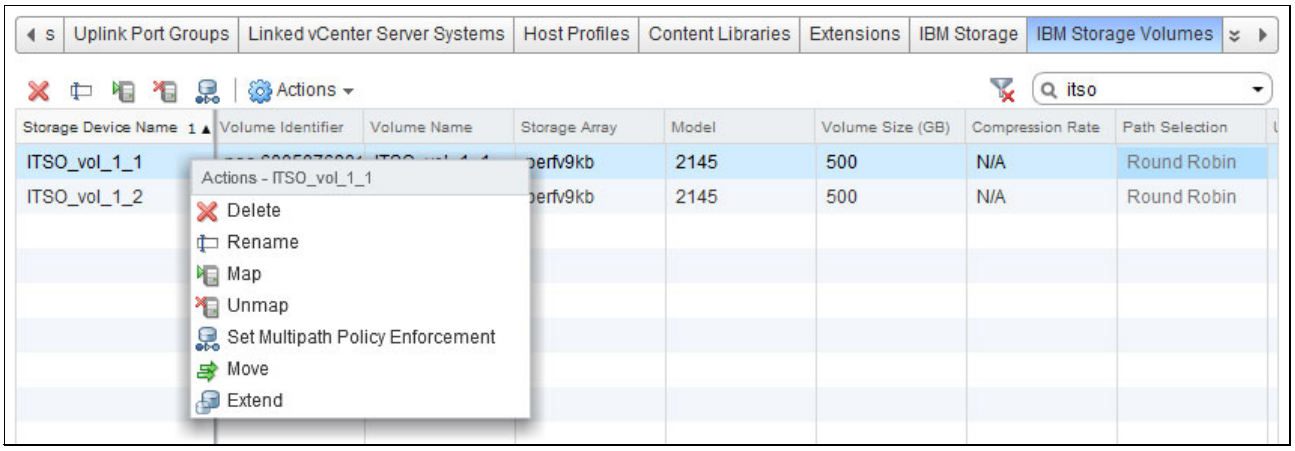


Figure 4-16 Volume action with the vSphere Web Client

4.5.1 Register VASA provider with vCenter server

The Spectrum Control Base server provides a unified VASA provider for IBM Storage systems in your environment.

To enable VASA integration with your vCenter, you need to specify a VASA user in your Spectrum Control Base. This user name is used by vCenter to communicate with Spectrum Control Base server.

1. Log in to the Spectrum Control Base GUI. Click the **Settings** button in the upper-left corner of the pane, and select **VASA credentials** from the Settings menu.

   The VASA Credentials dialog window is shown in Figure 4-17. It presents the currently defined VASA user name, or you can specify a new one.

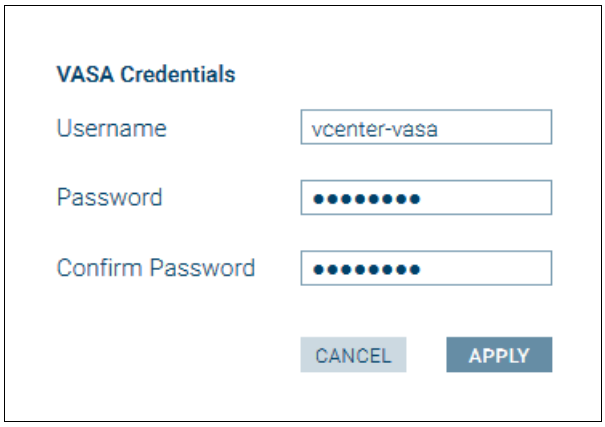


Figure 4-17 Specify VASA credentials in Spectrum Control Base

2. Enter the user name and password, and then click **Apply.**

Add a VASA provider

The next step is to add VASA provider in vCenter.

Note: At the time of writing IBM FlashSystem V9000 supports VASA1 provider of IBM Spectrum Control Base, which is listening on the following interface:

`https://<Spectrum Control IP address>:8443/services/vasa1`

Define the VASA storage provider in vSphere Web Client:

1. Click the vCenter server that you want to use.
2. In the Manage tab, click **Storage Providers**.
3. Click **+** (plus sign) to add the new storage provider (see Figure 4-18).

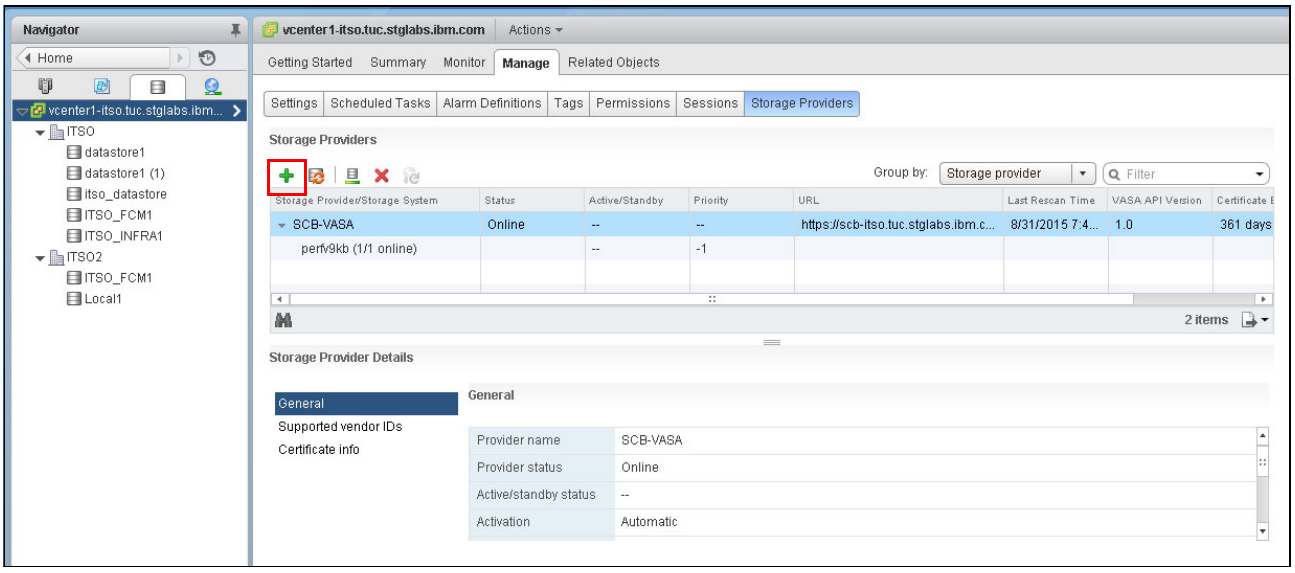


Figure 4-18 Storage providers menu

4. Complete the fields in the dialog window, as shown in Figure 4-19, and click **OK**.

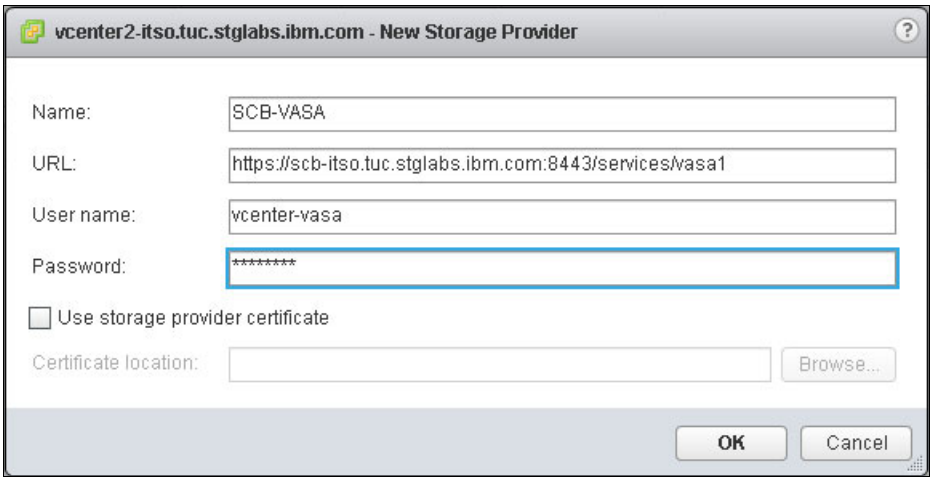


Figure 4-19 New Storage Provider in vCenter

5. Accept the certificate warning if you agree.

The benefit of using a VASA provider in vCenter is the ability to see storage capabilities of your datastores directly from the VMware Web Client, as shown in Figure 4-20.

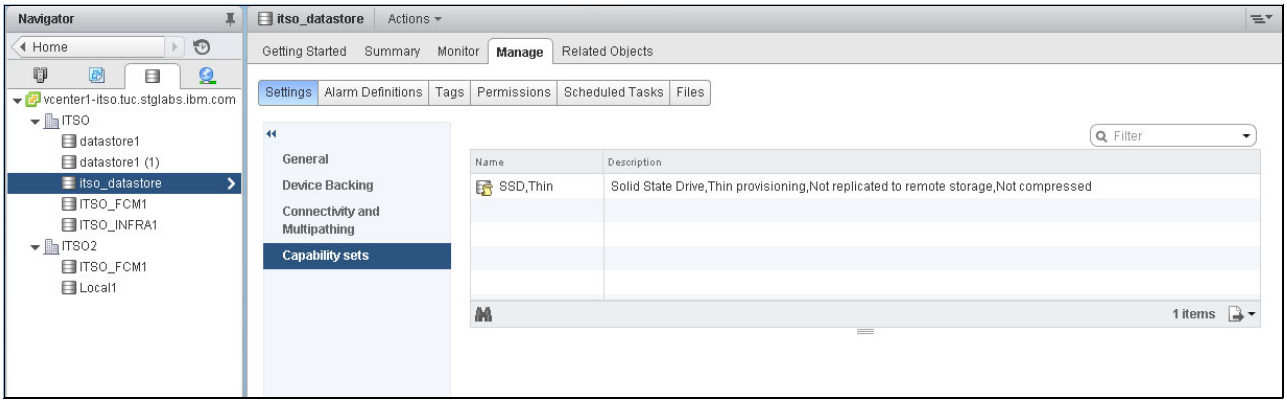


Figure 4-20 Storage capabilities of datastores view from Web Client

You can also use this information with Profile Driven Storage for defining *Storage policies*. This will simplify choosing the correct storage when provisioning virtual machines.

When you have just one single type of storage and all LUNs are the same, Profile-Driven Storage is probably not much of an improvement. Its advantage is realized when you have different types of storage LUNs or different storage systems. A simplified example is when some LUNs are mirrored and some are not and some are fast and some are slow. A catalog of different levels, such as gold, silver, and bronze, is another example.

Therefore, when you provision a virtual machine (VM), you can choose a policy, such as gold. VMware will identify for you which datastores are compatible or incompatible with the policy.

With Profile-Driven Storage, you can also verify whether virtual machines are stored in the correct type of storage. An example is a virtual machine that was migrated to storage that does not match the assigned policy, VMware highlights whether the virtual machine is compliant or noncompliant with the storage policy.

Define a new VM storage policy

Starting with vSphere 6.0, storage DRS also considers these profiles before making any recommendations or taking action. You can also manually add storage capabilities (tags) and then assign this user-defined storage capability to a datastore.

Follow these steps to define a new VM storage policy:

1. In the VMware Web Client home window, click the **VM Storage Policies** icon (see Figure 4-21).

Figure 4-21 VM Storage Policies icon

2. Click **Create a New VM Storage Policy**.
3. Select **vCenter server**, specify Name and Description of the policy, and click **Next** (see Figure 4-22).

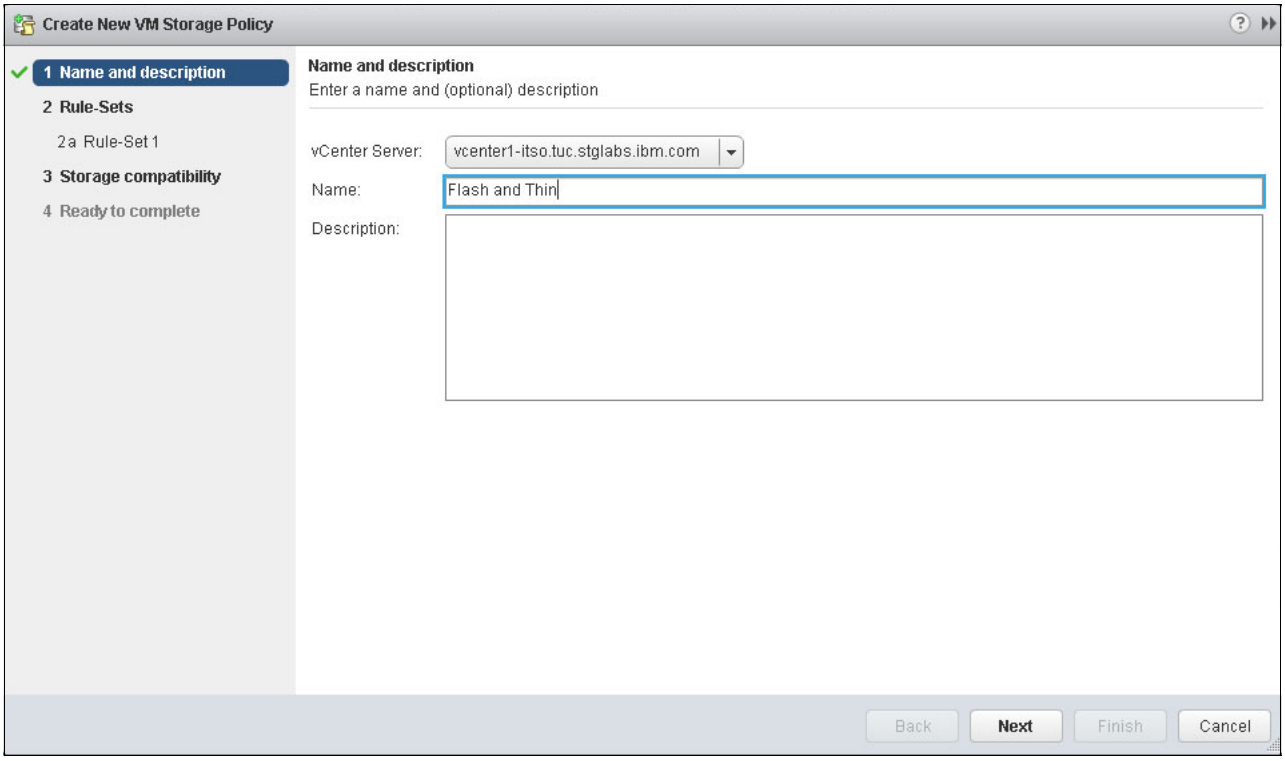


Figure 4-22 Storage Policy Name

4. When defining a rule set in the “Rules-based on data services” drop-down menu, select **ibm.hsg.vasa.VASA10** (see Figure 4-23).

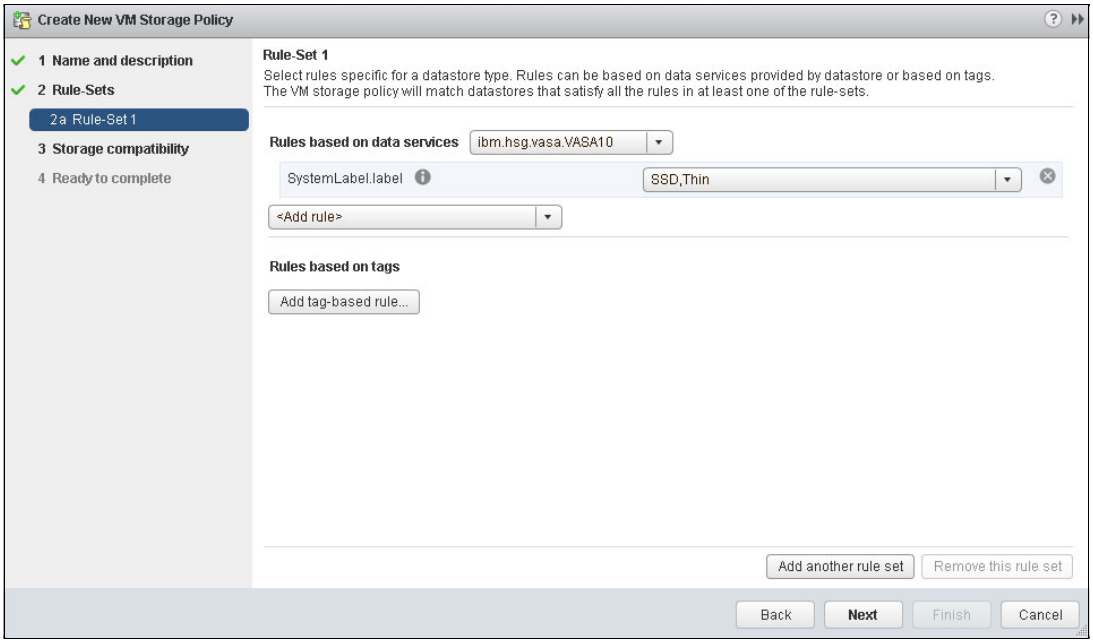


Figure 4-23 Specify a rule for a the new storage policy

5. Click **Next**. Alternatively, you can add your own rule or add another rule set. Now you can review all compatible storage that is based on the specified rule sets (see Figure 4-24).

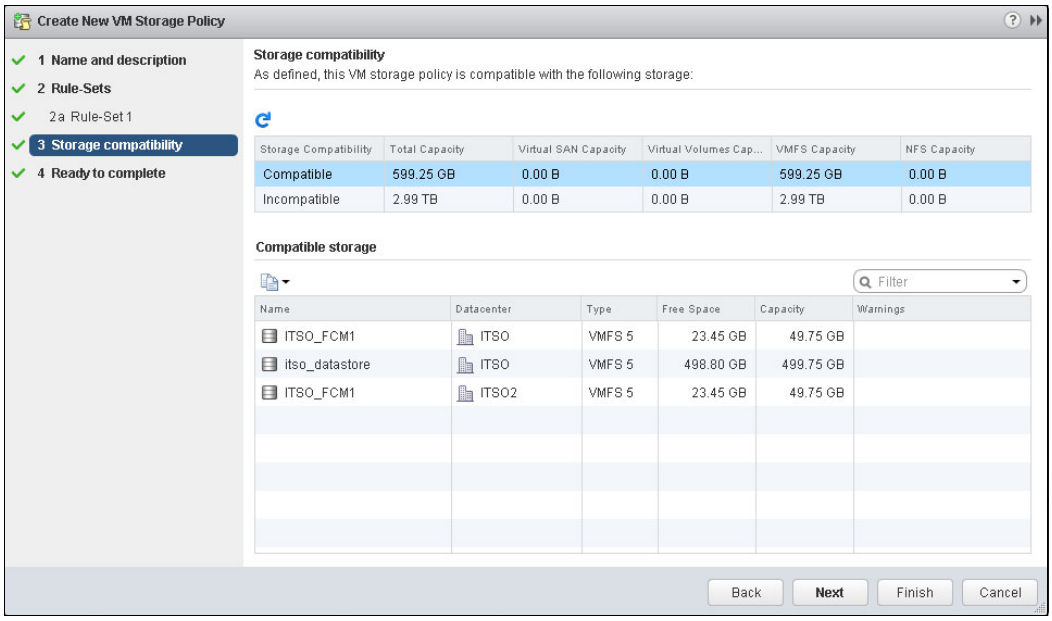


Figure 4-24 Compatible storage based on the rule sets

6. Click **Finish** to create your new policy.

5

VMware and FlashSystem V9000 multi-site guidelines

The term *disaster recovery* is normally used in reference to a large and significant and disruptive event, such as an earthquake. But it can also be relevant for small events, such as a virus that has spread in the network.

Companies prepare by implementing *business continuity* solutions to maintain and restore operations if a disruption or disaster occurs, but they do not always test those solutions regularly. The ability to recover the systems needs to be tested regularly to make sure that procedures work, rather than waiting until a disruption happens. Flaws might be detected each time you test, because perfection is impossible to achieve when the environment changes every day.

In this chapter, we describe some of the solutions that can help you prepare your environment to recover from a disruption:

- Copy Services
- VMware Site Recovery Manager (SRM) and Storage Replication Adapter (SRA)
- HyperSwap overview
- IBM Spectrum Protect™ Snapshot for VMware

This chapter includes the following sections:

5.1 Replication overview

IBM Replication Family Services provide the functions of storage arrays and storage devices, which allows various forms of block-level data duplication. You can make mirror images of part or all of your data between two sites. This is advantageous in disaster recovery scenarios with the capabilities of copying data from production environments to another site for resilience.

The following copy services (relationships) are supported by IBM FlashSystem V9000:

- FlashCopy, for point-in-time copy
- Metro Mirror, for synchronous remote copy
- Global Mirror, for asynchronous remote copy
- Global Mirror with Change Volumes, for asynchronous remote copy for a low-bandwidth connection

A FlashSystem V9000 system partnership can be created between a FlashSystem V9000 system and another FlashSystem V9000, a FlashSystem V840, an IBM SAN Volume Controller (SVC) system, or an IBM Storwize V7000 system operating in the replication layer. For more information about these services, see Chapter 3, “Advanced software functions,” in the IBM Redbooks publication titled *Introducing and Implementing IBM FlashSystem V9000*, SG24-8273.

Note: All of these services are supported with VMware Site Recovery Manager when using IBM Storwize Family Storage Replication Adapter (SRA).

5.1.1 FlashCopy

FlashCopy is known as a *point-in-time copy*. It makes a copy of the blocks from a source volume and duplicates them to the target volumes.

When you initiate a FlashCopy operation, a FlashCopy relationship is created between a source volume and target volume. A FlashCopy relationship is a mapping of the FlashCopy source volume and a FlashCopy target volume. This mapping allows a point-in-time copy of that source volume to be copied to the associated target volume. If it is a persistent FlashCopy, the FlashCopy relationship exists between this volume pair from the time that you initiate a FlashCopy operation until the storage unit copies all data from the source volume to the target volume or you delete the FlashCopy relationship.

5.1.2 Metro Mirror

Metro Mirror is a type of remote copying that creates a synchronous copy of data from a primary volume to a secondary volume. A secondary volume can be located either on the same system or on another system. The maximum distance allowed between systems in Metro Mirror relationships is 300 km.

Tip: Using Metro Mirror over long distances (intersite) can negatively affect latency of FlashSystem V9000. All writes are synchronous, so that also applies if it is used with a slower system, such as IBM Storwize V7000. For best performance, consider using Metro Mirror relationships only between systems with similar performance.

5.1.3 Global Mirror

The Global Mirror function provides an asynchronous copy process. When a host writes to the primary volume, confirmation of I/O completion is received before the write operation has completed for the copy on the secondary volume. The maximum acceptable distance between systems in Global Mirror relationships is 25.000 km or 250 ms latency.

Global Mirror change volumes

Global Mirror *change volumes* are copies of data from a primary volume or secondary volume that are used in Global Mirror relationships. Using change volumes lowers bandwidth requirements by addressing only the average throughput, not the peak.

Remote copy consistency groups

You can group Metro Mirror or Global Mirror relationships into a *consistency group* so that they can be updated at the same time. A command is then simultaneously applied to all of the relationships in the consistency group.

5.1.4 VMware Site Recovery Manager

VMware vCenter Site Recovery Manager (SRM) is well-known in the virtualization world for providing simple, affordable, and reliable business continuity and disaster recovery management.

Tip: Ideally, use small volumes (LUNs) for faster synchronization between the primary and the secondary sites.

Using SRM with IBM FlashSystem V9000 can help you protect your virtual environment.

SRM automates the failover processes and the ability to test failover processes or disaster recovery without an impact on the live environment. This helps you meet your recovery time objectives (RTOs).

VMware vCenter Site Recovery Manager supports two forms of replication:

- Array-based replication (ABR), where the storage system manages the virtual machine replication with the following attributes:
  - Requires compatible storage like FlashSystem V9000
  - Connection to the storage array is performed with Storage Replication Adapter (SRA)
- Host-based replication, which is known as vSphere Replication (VR), where the ESXi is managing the virtual machine replication with the following attributes:
  - Does not depend on storage array compatibility
  - Increased network efficiency by replicating only the most recent data in the changed disk areas
  - Minimum RPO = 15 minutes (or 5 minutes if using virtual SAN)

Figure 5-1 shows an overview of the VMware Site Recovery Manager.

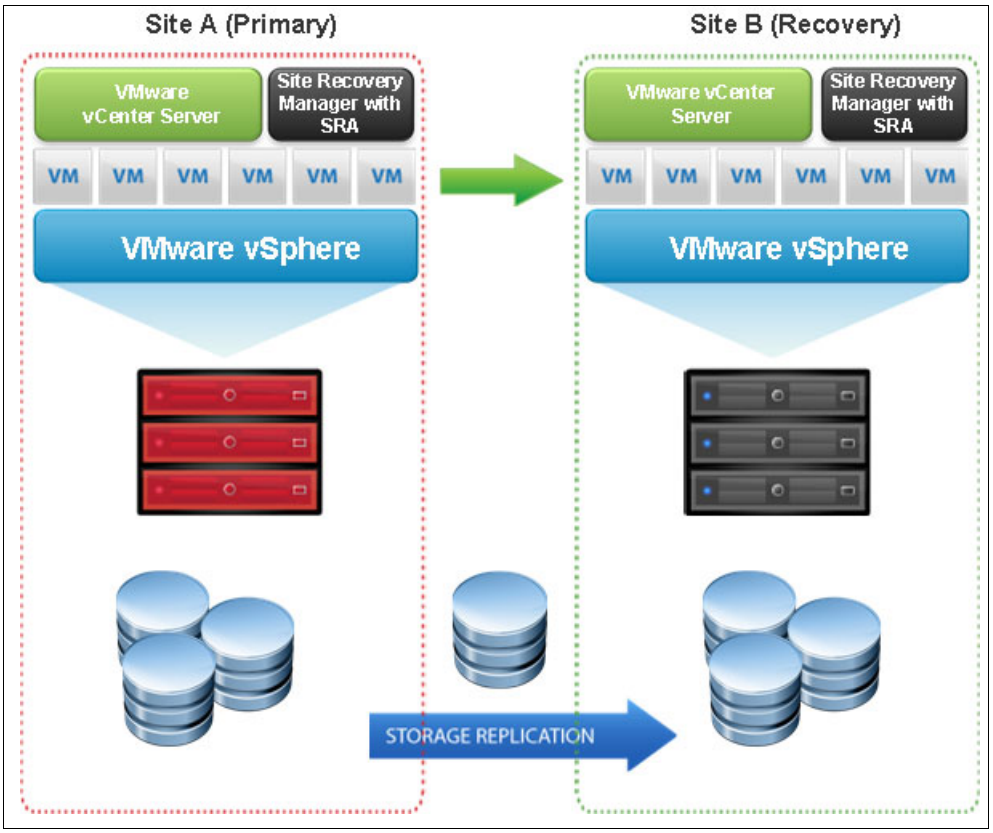


Figure 5-1 VMware Site Recovery Manager

VMware vCenter Site Recovery Manager requires one vCenter server in each site with the respective licenses. Also, if you are using it with IBM FlashSystem V9000, it will require you to use a Storage Replication Adapter (SRA), which is described in the following topic, 5.1.5, “Storage Replication Adapter” on page 52.

To learn more about SRM, see the VMware Site Recovery Manager documentation:

https://www.vmware.com/support/pubs/srm_pubs.html

5.1.5 Storage Replication Adapter

IBM Storwize Family Storage Replication Adapter (SRA) is a storage vendor plug-in developed by IBM. It is required for the correct functioning of VMware vCenter Site Recovery Manager (SRM).

The adapter is used to enable management of Advanced Copy Services on IBM FlashSystem V9000, such as Metro Mirror and Global Mirror (including changed volumes).

The combination of SRM and SRA enables the automated failover of virtual machines from one location to another, connected by either Metro Mirror or Global Mirror technology.

By using the IBM Storwize Family Replication Adapter, VMware administrators can automate the failover of a FlashSystem V9000 at the primary SRM site to a compatible system, such as another FlashSystem V9000, V840, IBM SAN Volume Controller, or IBM Storwize V7000 system at a recovery (secondary) SRM site.

Immediately upon failover, the ESXi servers at the secondary SRM site initiate the replicated datastores on the mirrored volumes of the secondary storage system. When the primary site is back online, perform failback from the recovery site to the primary site by clicking **Reprotect** in the SRM.

For details, see the IBM Storwize Family Storage Replication Adapter Documentation in IBM Knowledge Center:

http://ibm.co/1LNgyCX

Basic Storage Replication Adapter configuration

After you download the SRA plug-in from IBM Fix Central, install it on both sites and specify basic settings:

1. Run **IBMSVCSRAUtil**, as shown in Figure 5-2.

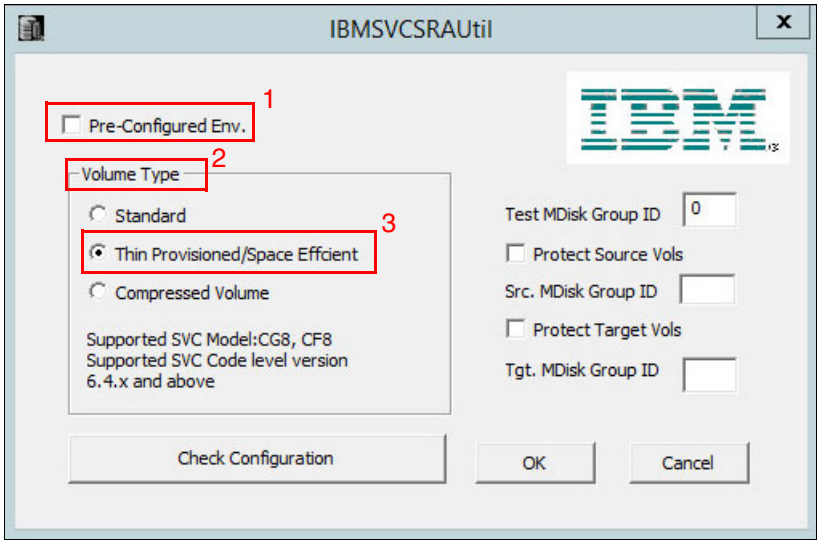


Figure 5-2 IBMSVCSRAUtil interface

   If you want to let SRM handle all tasks, including test recovery, clear the Pre-Configured Env. field and check box (1). This setting needs to be consistent on both the Primary and Recovery site and will require administrator permissions on IBM FlashSystem V9000.

2. In the Volume Type field (2), specify the type of FlashCopy copy volumes created for test recovery purposes. For this example, we select **Thin Provisioned/Space Efficient** (3).
3. Click **OK** to confirm, and then close configuration utility.
4. To create array pairs in SRM, run VMware Web Client, navigate to **Site Recovery** → **Array Based Replication** and click **Add**, as highlighted in Figure 5-3.

   In this example, we assume that your SRM site pairs are already configured.

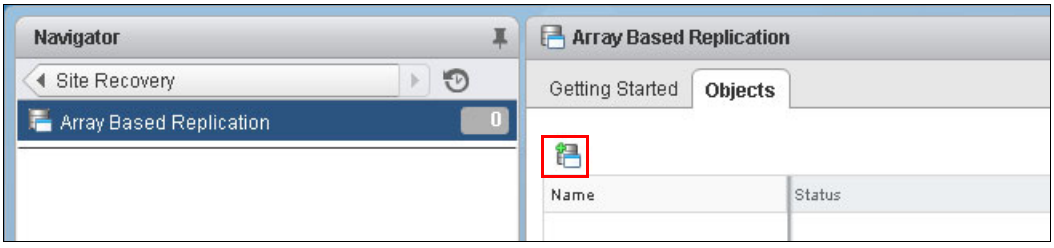


Figure 5-3 Array Based Replication, unconfigured

5. Click **Add a pair of array managers**, and then click **Next** (Figure 5-4).

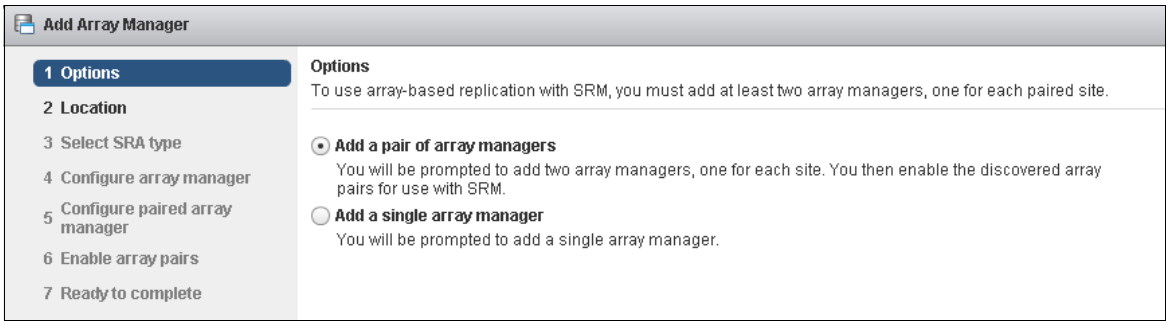


Figure 5-4 Add a pair of array managers

6. If you have more site pairs defined in SRM, select the one where you want to create the array pair, and then click **Next** (Figure 5-5).

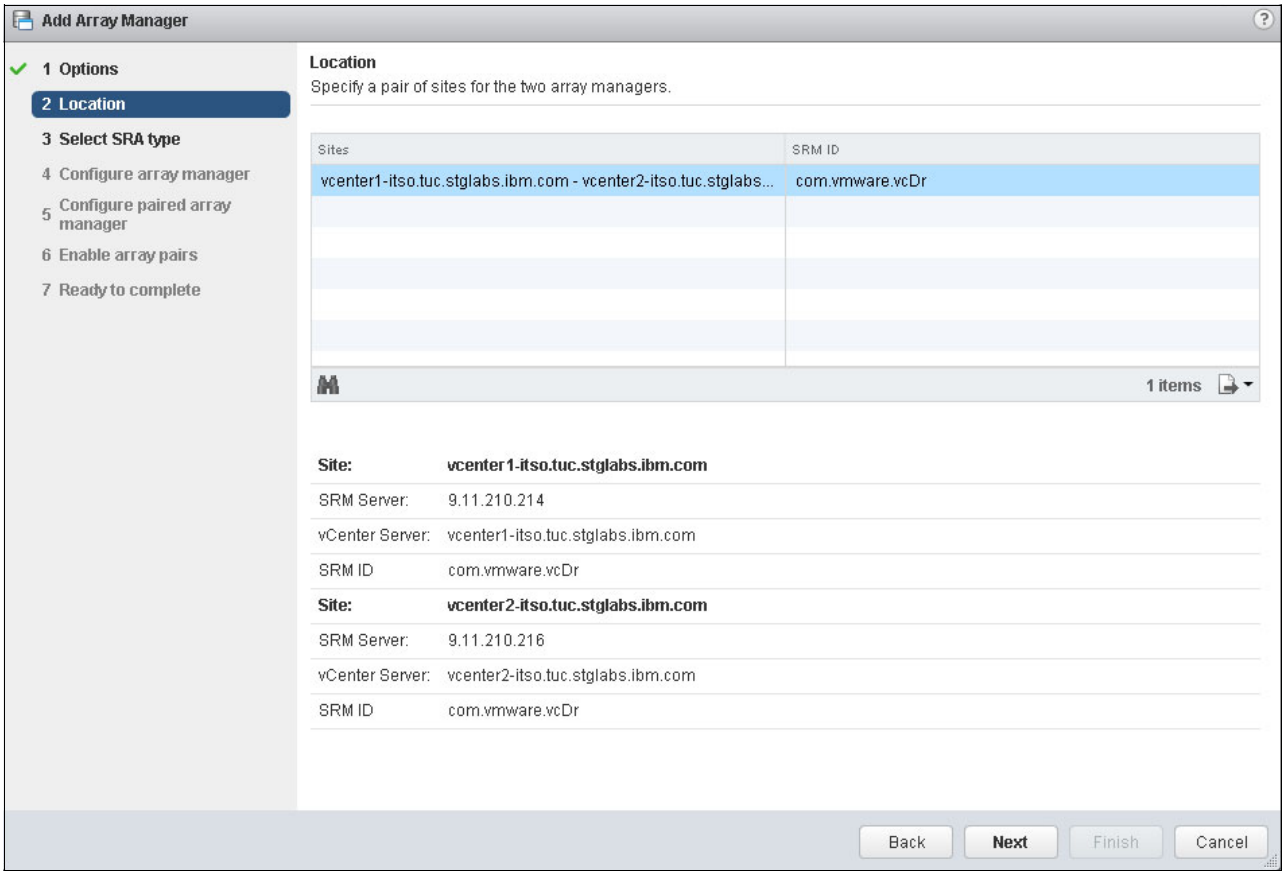


Figure 5-5 Specify the sites

7. Select **IBM Storwize Family Storage Replication Adapter** from the SRA type list, and click **Next** (Figure 5-6).

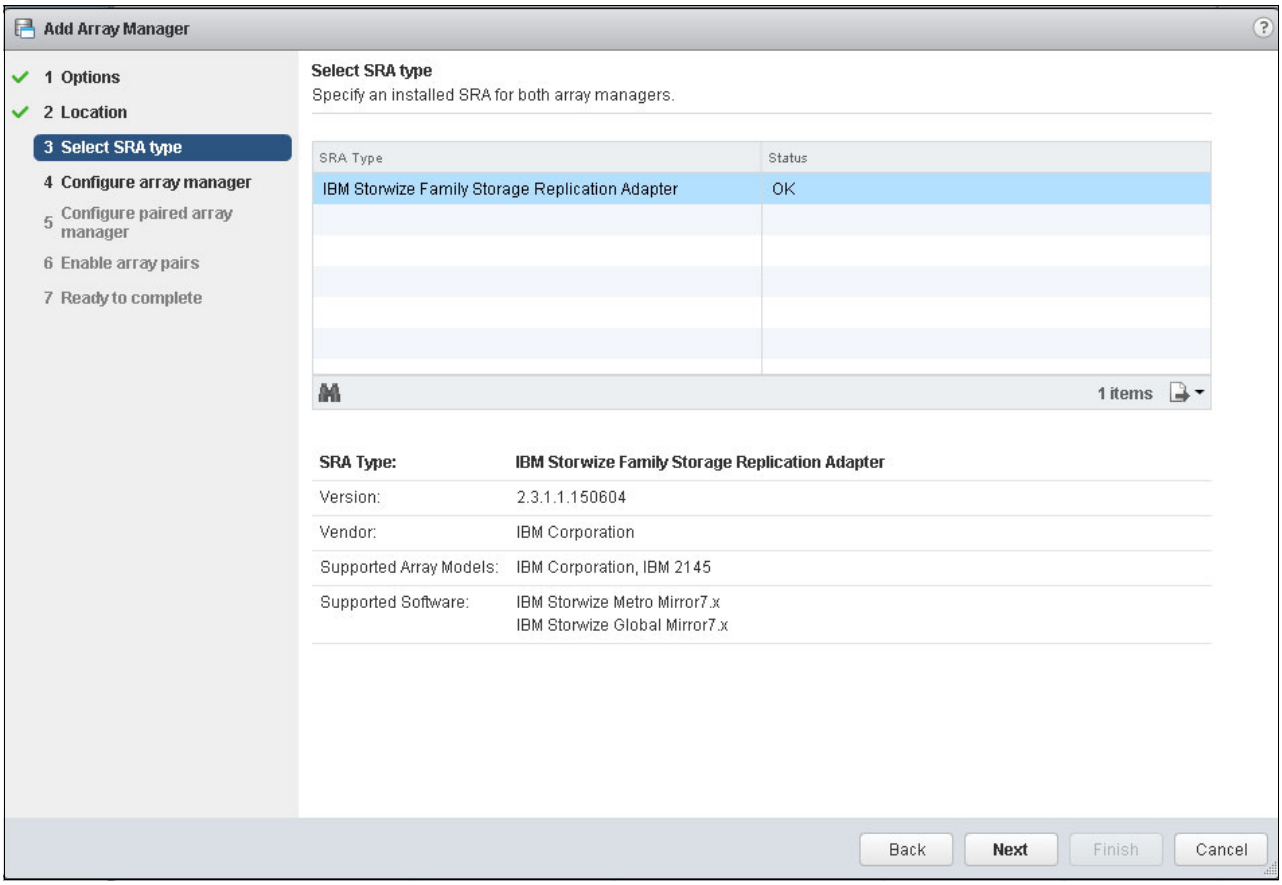


Figure 5-6 Select SRA type

8. Complete your settings for your IBM FlashSystem V9000 on Protected site (Figure 5-7).

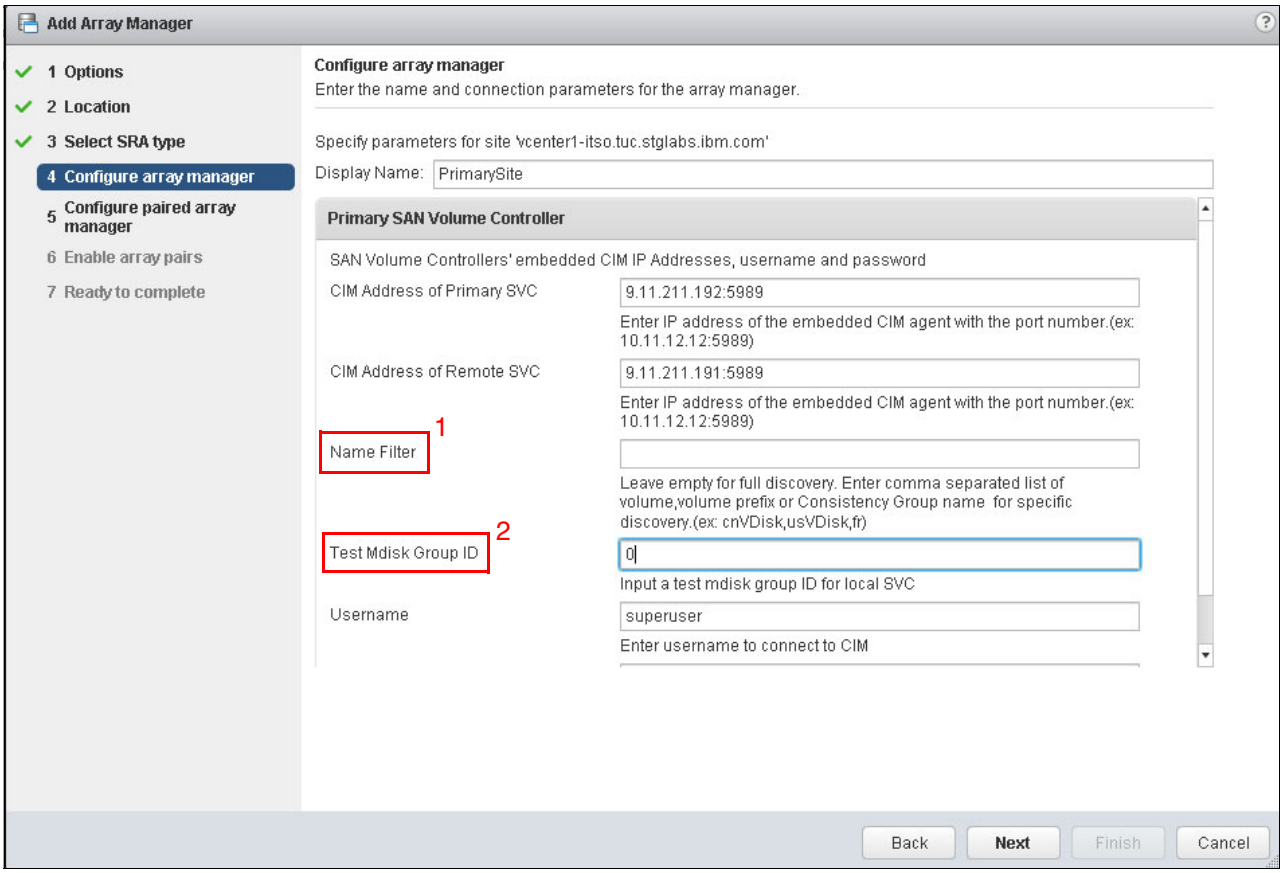


Figure 5-7 Configure Protected Site

If you want to discover only specific remote copy relationships, you can specify a discovery prefix in the Name Filter field (1).

9. In the **Test MDisk Group ID** field (2), specify the MDisk group for FlashCopy volumes that was created during test recovery (this is usually 0 for IBM FlashSystem V9000, because there is only one MDisk group in standard configuration).

10. Click **Next** to configure the recovery site.

The CIM address of the primary SVC (Figure 5-7) is FlashSystem V9000 on the protected site, and the CIM address of the remote SVC is FlashSystem V9000 on the recovery site. Therefore, you will be configure settings for the recovery site (Paired Array Manager), and IP addresses will be switched, as shown on Figure 5-8 on page 57.

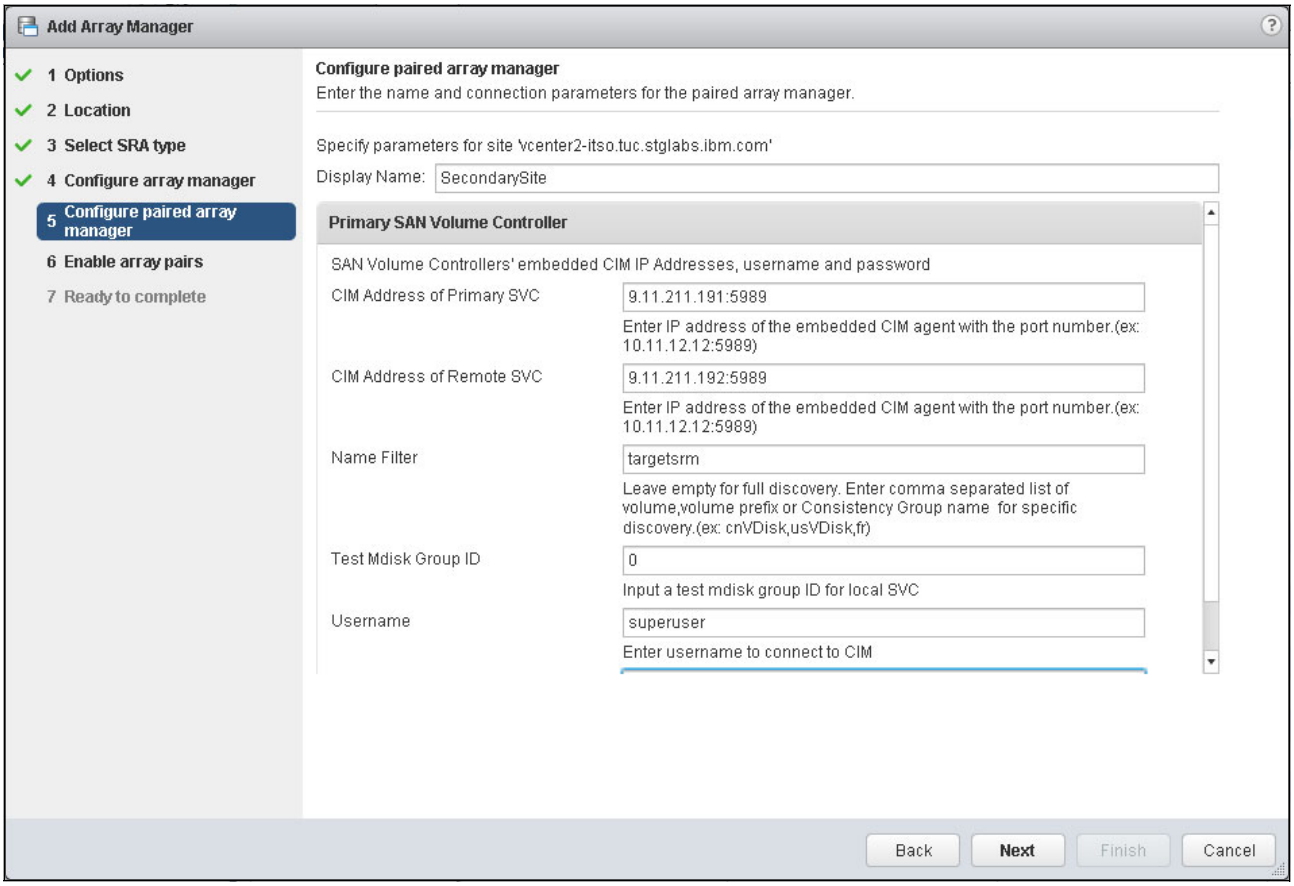


Figure 5-8 Configure the recovery site

11. After you enter the Recovery Site settings, click **Next** and **Enable array pairs** (see Figure 5-9).

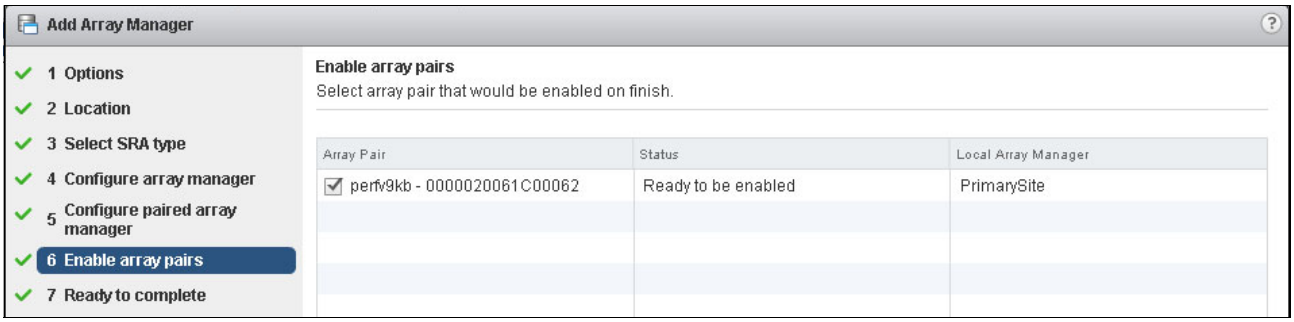


Figure 5-9 Enable array pairs

12. Select the array pair, and then click **Next** to show Ready to complete page where you can review your settings and click **Finish** (Ready to complete) if you agree (see Figure 5-10).

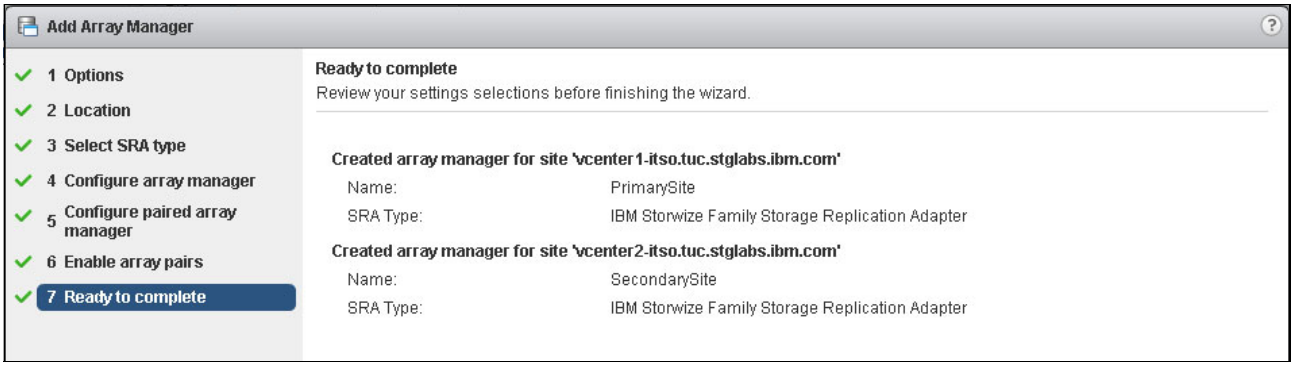


Figure 5-10 Review array pair settings

Your array pairs will be shown in Array Based Replication list, as shown in Figure 5-11.

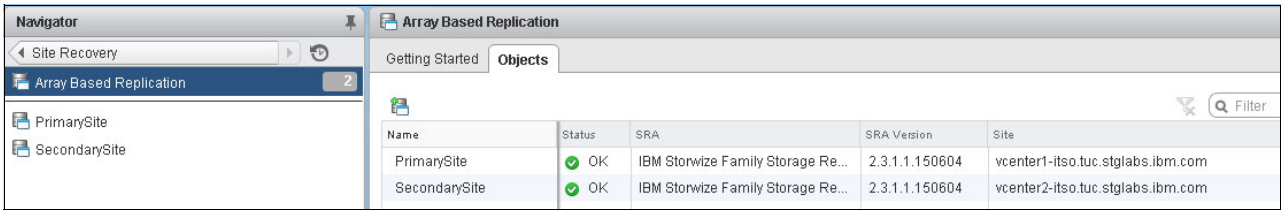


Figure 5-11 Array Based Replication - configured

You can also verify detected relationships by clicking a specific array manager and the **Manage** tab, as shown on Figure 5-12.

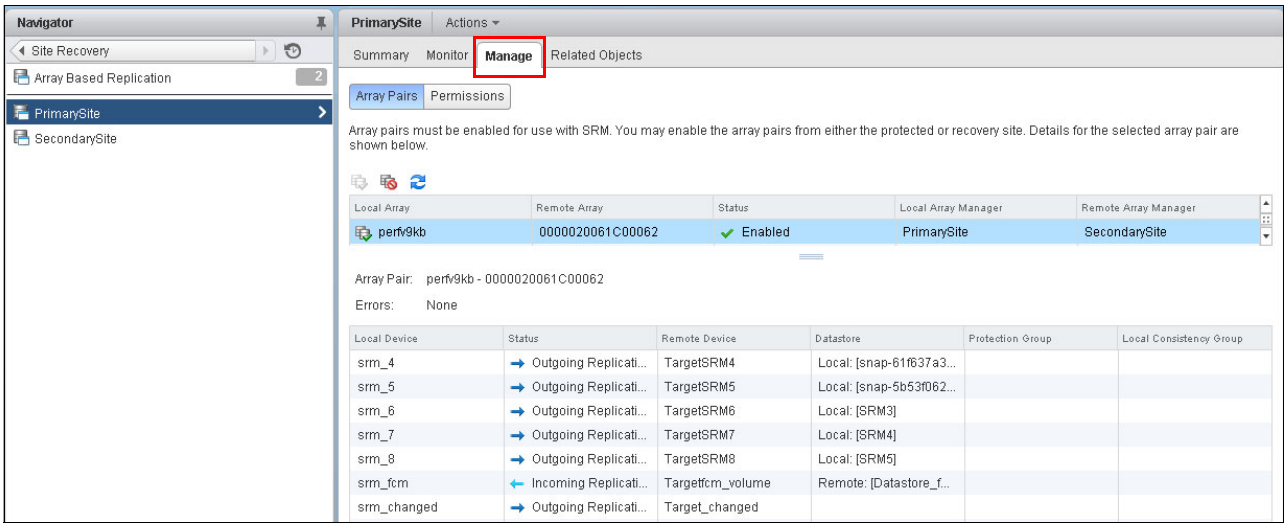


Figure 5-12 Detected mirror relationships

Tip: It is good practice to specify consistency groups in IBM FlashSystem V9000 for your remote copy relationships.

When you are creating protection groups in SRM, try to maintain it within one consistency group. This will guarantee storage consistency (not file system) across multiple volumes.

Creating protection groups spanning across multiple consistency groups is not recommended, especially for Global Mirror relationships

5.2 HyperSwap overview

The HyperSwap high availability function in FlashSystem V9000 software allows business continuity if there is a hardware failure, power failure, connectivity failure, or disasters such as fire or flooding.

The HyperSwap function provides highly available volumes that are accessible through two sites at up to 300 km apart. A fully independent copy of the data is maintained at each site. When data is written by hosts at either site, both copies are synchronously updated before the write operation is completed. The HyperSwap function will automatically optimize itself to minimize data transmitted between sties and to minimize host read and write latency.

The HyperSwap capability allows a volume to be presented by two FlashSystem V9000 I/O groups. The I/O groups can reside in the same data center, but the more common deployment is to have the two I/O groups in different data center locations. The configuration of HyperSwap tolerates combinations of FlashSystem V9000 controller failures, as well as complete failures of sites that contain I/O groups.

The HyperSwap function contains the following configuration attributes:

- FlashSystem V9000 control enclosures are spread across two sites, with storage at a third site acting as a tie-breaking quorum device.
- Both control enclosures of an I/O group are in the same site, which adds resiliency for acquiring storage volumes on both sites, at least two I/O groups are required.
- Additional system resources are used to support a fully independent cache on each site. This allows full performance even if one site is lost.

Hosts, FlashSystem V9000 control enclosures, and FlashSystem V9000 storage enclosures are in one of two failure domains or sites, and volumes are visible as single objects across both sites (I/O groups).

Note: At the time of publication, the two I/O groups in a HyperSwap configuration must reside within a single FlashSystem V9000 cluster. For more detailed information about HyperSwap, see the Redbooks publication titled *Introducing and Implementing IBM FlashSystem V9000*, SG24-8273.

Figure 5-13 shows an overview of how the HyperSwap function works:

- Each primary volume (denoted by the “p” in the volume name) has a secondary volume (denoted by the “s” in the volume name) on the opposite I/O group.
- The secondary volumes are not mapped to the hosts.
- The dual write to the secondary volumes is handled by FlashSystem V9000 HyperSwap function and is transparent to the hosts.

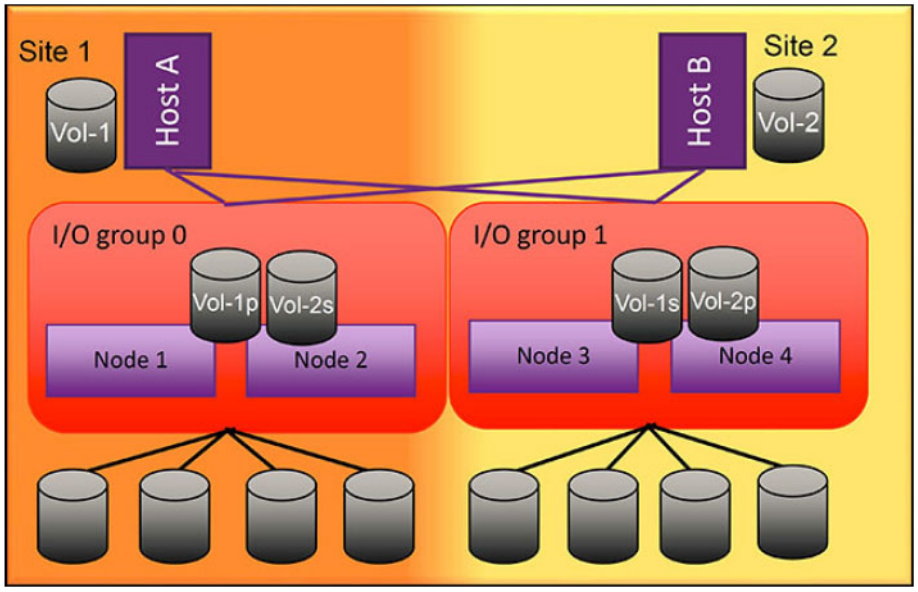


Figure 5-13 HyperSwap overview

5.2.1 HyperSwap with VMware vSphere Metro Storage Cluster

The HyperSwap function on FlashSystem V9000 can be used as a supported VMware vSphere Metro Storage Cluster (vMSC) storage device. The supported use cases of a FlashSystem V9000 using HyperSwap and VMware vSphere benefit from the following attributes:

- A FlashSystem V9000 cluster with HyperSwap presents an accessible VMware VMFS volume to vSphere hosts at two separate data center locations, separated by a distance of up to 300 km.
- A single stretched vSphere cluster that uses VMware High Availability (HA) and Dynamic Resource Scheduler (DRS) functions with hosts at two separate data center locations, separated by a distance of up to 300 km.
- VMware vMotion between vSphere hosts at two separate data center locations are separated by a distance of up to 300 km.
- VMware HA automatically fails over virtual machines between data centers due to server, storage, or site failure.

VMware vSphere host multipathing ensures that running virtual machines continue to operate during various failure scenarios. Table 5-1 outlines the tested and supported failure scenarios when using FlashSystem V9000 HyperSwap and VMware vSphere Metro Storage Cluster (vMSC).

Table 5-1 FlashSystem V9000 HyperSwap and VMware vMSC supported failure scenarios

Failure scenario	HyperSwap behavior	VMware HA impact
Path failure: FlashSystem V9000 Back-End (BE) Port	Single path failure between FlashSystem V9000 control enclosure and flash enclosure. No impact on HyperSwap.	No impact.
Path failure: FlashSystem V9000 Front-End (FE) Port	Single path failure between FlashSystem V9000 control enclosure and vSphere host. vSphere host uses alternate paths.	No impact.
BE flash enclosure failure at site-1	FlashSystem V9000 continues to operate from the volume copy at site 2. When the flash enclosure at site 1 is available HyperSwap will synchronize the copies.	No impact.
BE flash enclosure failure at site-2	Same behavior as failure at site 1	No impact.
FlashSystem V9000 control enclosure failure	FlashSystem V9000 continues to provide access to all volumes through the other control enclosures.	No impact.
Complete site 1 failure (The failure includes all vSphere hosts and FlashSystem V9000 controllers at site-1)	FlashSystem V9000 continues to provide access to all volumes through the control enclosures at site 2. When the control enclosures at site 1 are restored, the volume copies will be synchronized.	Virtual machines running on vSphere hosts at the failed site are impacted. VMware HA automatically restarts them on vSphere hosts at site 2.
Complete site 2 failure	Same behavior as a failure of site 1.	Same behavior as a failure of site 1.
Multiple vSphere host failures Power Off	No impact.	VMware HA automatically restarts the virtual machines on available ESXi hosts in the VMware HA cluster.
Multiple vSphere host failures, network disconnect	No impact.	VMware HA continues to use the datastore heartbeat to exchange cluster heartbeats. No impact.
FlashSystem V9000 inter-site link failure, vSphere cluster management network failure	FlashSystem V9000 active quorum is used to prevent a split-brain scenario by coordinating one I/O group to remain servicing I/O to the volumes, the other I/O group goes offline	vSphere hosts continue to access volumes through the remaining I/O group. No impact.
Active FlashSystem V9000 quorum disk failure	No impact to volume access. A secondary quorum disk is assigned upon failure of the active quorum.	No impact.

Failure scenario	HyperSwap behavior	VMware HA impact
vSphere host isolation	No impact.	HA event dependent upon isolation response rules configured for the vSphere cluster. VMs can be left running, or rules can dictate for VMs to shut down and restart on other hosts in the cluster.
vCenter server failure	No impact.	No impact to running VMs or VMware HA. VMware DRS function is affected until vCenter access is restored.

5.3 IBM Spectrum Protect Snapshot for VMware

IBM Spectrum Protect Snapshot for VMware (formerly known as IBM Tivoli® Storage FlashCopy Manager for VMware) is a data management solution that can be used to streamline storage management in a VMware vSphere environment. You can back up and restore virtual machines and VMware datastores.

Spectrum Protect Snapshot for VMware combines with the VMware vSphere API and the snapshot capabilities of storage devices to protect your environment. You can create nondisruptive off-host backups for VMware virtual machines in a vSphere environment. This off-host approach facilitates faster backup operations.

By using Spectrum Protect Snapshot for VMware with IBM Spectrum Protect for Virtual Environments (formerly known as IBM Tivoli Storage Manager), you can also offload and store VMware image backups on IBM Spectrum Protect server storage for long-term retention.

Figure 5-14 shows a high-level overview of Spectrum Protect Snapshot.

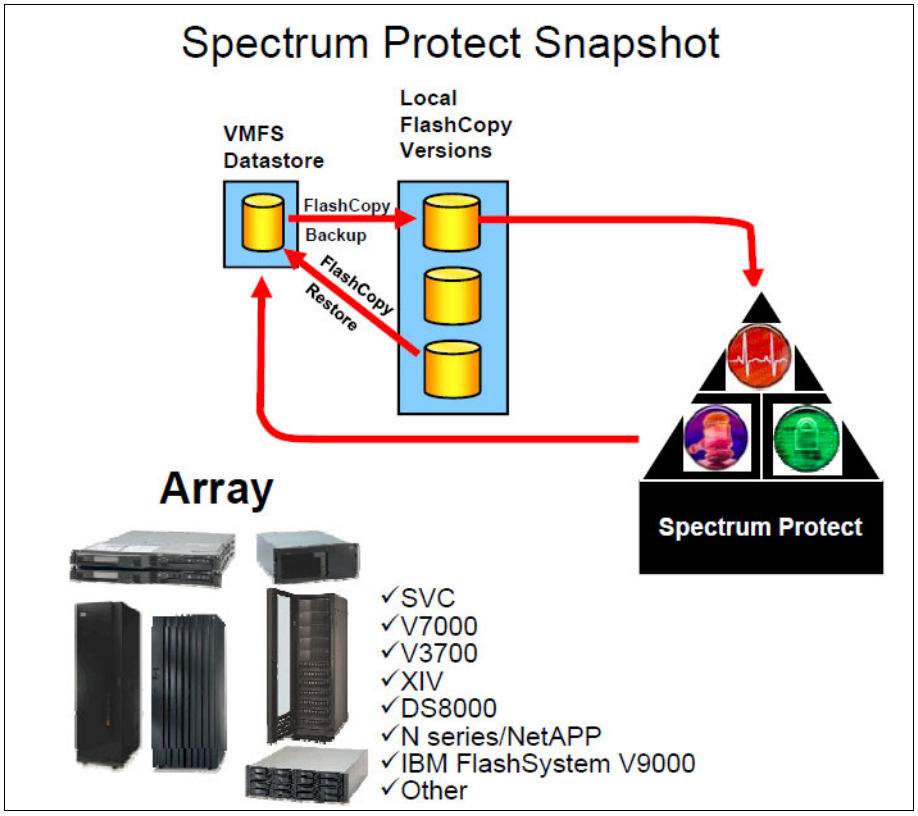


Figure 5-14 Overview of Spectrum Protect Snapshot

Before installing Spectrum Control Protect Snapshot for VMware, review the preinstall checklist to verify that all of the prerequisites are fulfilled. This document provides comprehensive information about the environment that is required and helps you complete the following tasks:

- Prepare for implementation by completing the Pre-installation Checklist before implementation starts to avoid problems:

  `http://ibm.com/support/docview.wss?uid=swg21965151&aid=1`
- Identify and involve all responsible organizations that are required for implementation.
- Verify correct release levels of additional software.
- Get the most current instructions for installation.

Spectrum Protect Snapshot is an off-host backup method, because you install the backup application on a dedicated virtual machine or a physical Linux system: the *vStorage backup server*. Currently, Red Hat Enterprise Linux or SUSE® Linux Enterprise is supported. For more information about the supported Linux release levels, see the preinstallation checklist..

All backup or restore backup operations are started from the vStorage backup server by using the graphical user interface or the command line. Through the VMware vStorage API and FlashCopy capabilities of FlashSystem V9000, you can create file-level, guest-level, and file system-consistent backups.

In Figure 5-15, the vStorage backup server is running as a virtual machine. SAN connectivity to the storage system is not required for the vStorage backup server. FlashSystem V9000 volumes can be attached to the ESXi server using either Fibre Channel or iSCSI. Network-attached storage (NAS) is not supported with FlashSystem V9000.

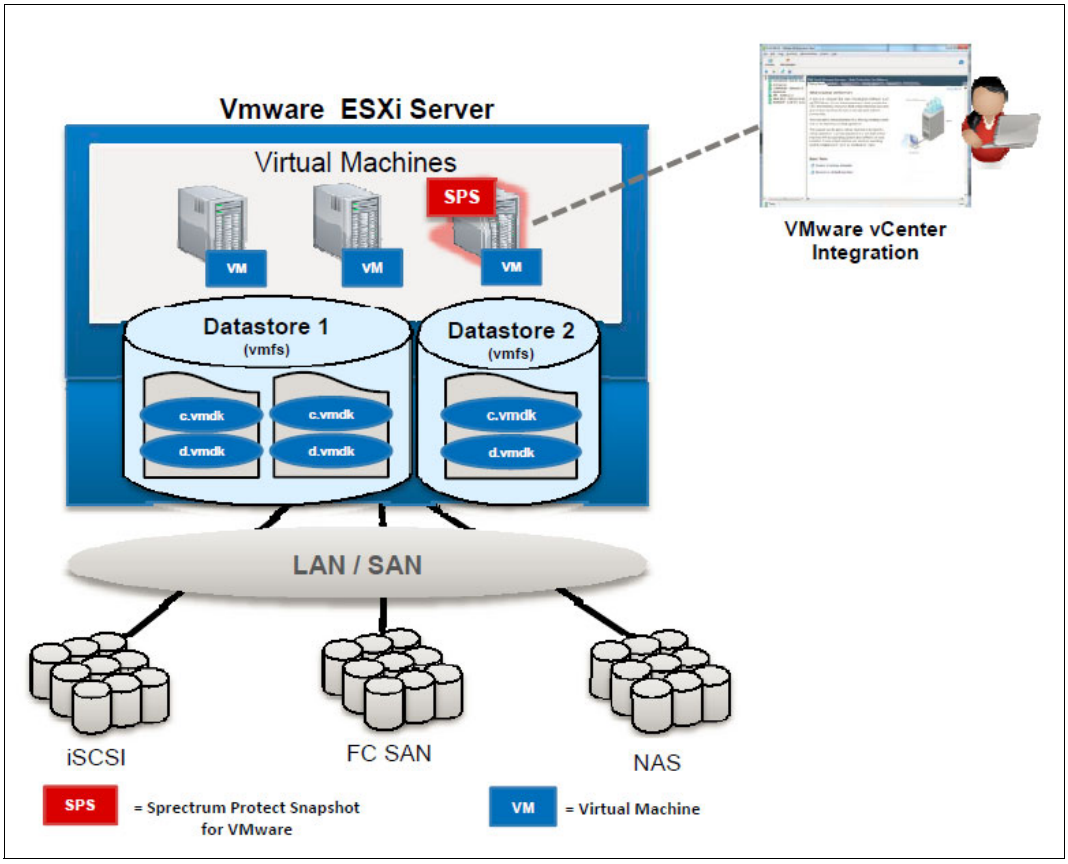


Figure 5-15 Integration of Spectrum Protect Snapshot in the vSphere environment

When you back virtual machine with Spectrum Protect Snapshot for VMware, the following backup tasks are performed, as shown in Figure 5-16 on page 65:

1. Spectrum Protect Snapshot Identifies the logical unit numbers (LUNs) that are associated with the virtual machines. A LUN is a unique identifier for a FlashSystem V9000 volume. In the example in Figure 5-16 on page 65, Datastore 1 contains two virtual machines, that will be used for the backup on disk.
2. For each virtual machine, that will be included in the backup, a VMware snapshot is started through the VMware vSphere API.
3. For each LUN of Datastore 1, a FlashCopy mapping is created on FlashSystem V9000. Each FlashCopy mapping will be added to a *consistency group* and then the FlashCopy process will be started. The target volumes can be thin-provisioned or fully allocated.

   A consistency group is a group of FlashCopy mappings. A consistency group manages the consistency of dependent writes by creating a consistent FlashCopy across multiple volumes.
4. The VMware snapshots will be deleted for the backed up virtual machines in Datastore 1.

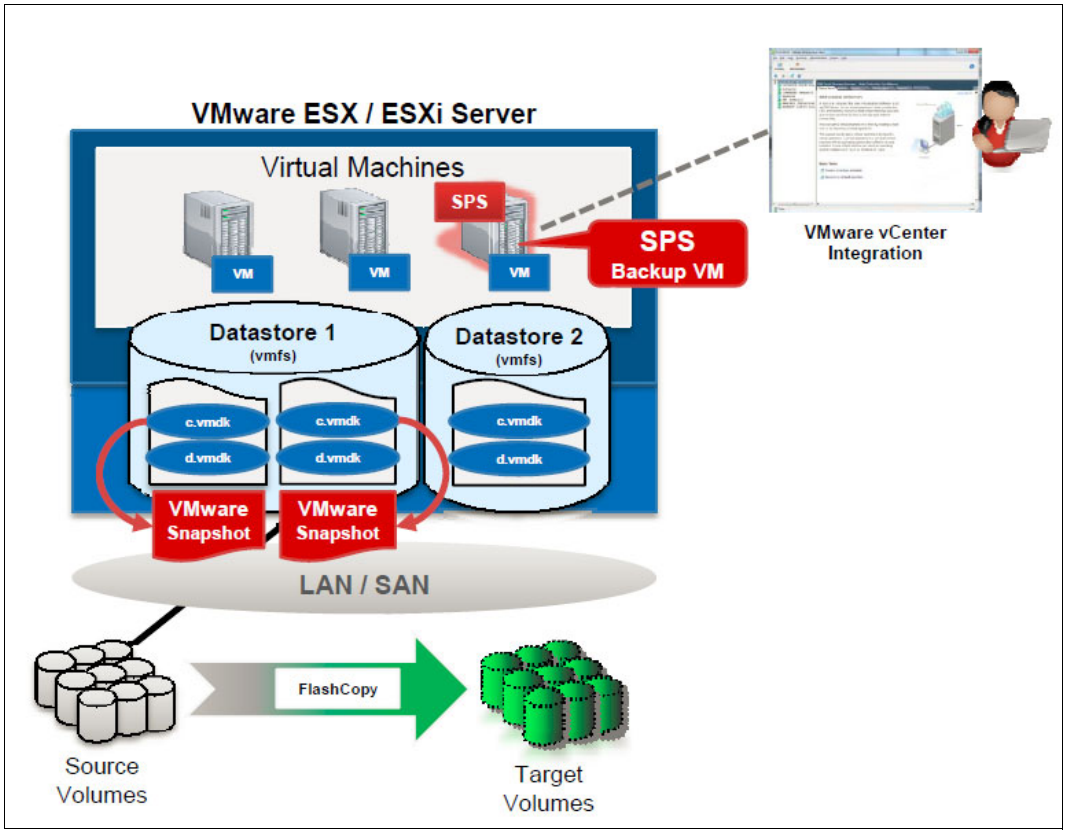


Figure 5-16 Backup process for virtual machines with Spectrum Protect Snapshot for VMware

Spectrum Protect Snapshot for VMware can be integrated with Spectrum Protect for Virtual Environments, formerly known as Tivoli Storage Manager, for Virtual Environments to offload VMware image backups to Spectrum Protect server storage and have a long-term data retention.

The following tasks are completed by Spectrum Protect Snapshot for VMware, when you send FlashCopy backups to Spectrum Protect:

- The target volumes, which contain the backup of Datastore 1, are attached to the *auxiliary* ESXi server. This server is used to temporarily mount a FlashCopy backup when required.
- The virtual machines are registered on the auxiliary ESXi server.
- The backup to Spectrum Protect is performed.
- The virtual machines are unregistered, and the datastore is detached from the auxiliary ESXi server.

Spectrum Protect for Virtual Environments uses the data mover node to send the snapshots to Spectrum Protect. This movement minimizes the impact on resources available to the virtual machines in the vCenter. In addition, multiple data mover nodes can be used so that the Spectrum Protect backup workload can be distributed.

Figure 5-17 illustrates the steps and process.

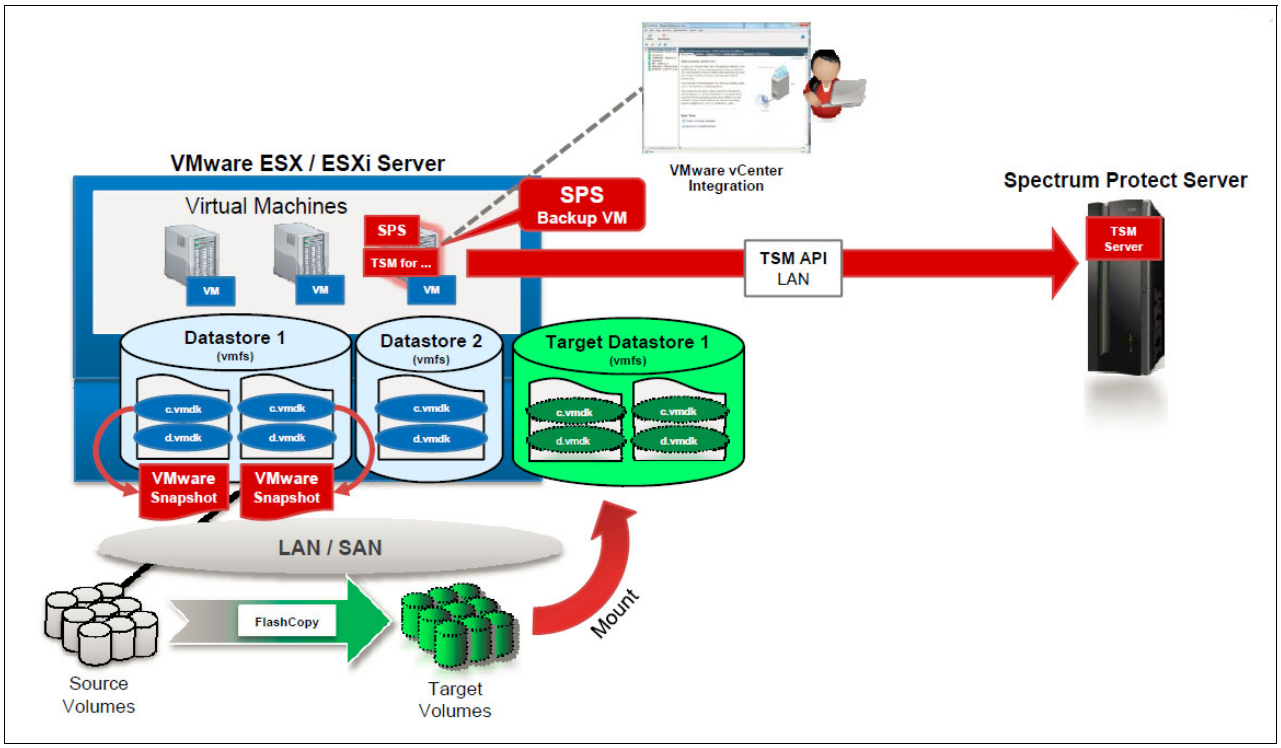


Figure 5-17 Offload of the virtual machine backup to Spectrum Protect Server

The virtual machines or datastores can be restored from the backup on disk by performing a reverse FlashCopy on FlashSystem V9000 or directly from Spectrum Protect.

5.3.1 Unsupported virtual disk types

The following types of virtual disks do not support VMware snapshot operations, so they cannot be used for backup and restore operations with Spectrum Protect Snapshot for VMware:

- Raw device mapped volumes created in physical compatibility mode (pRDM)
- iSCSI disks attached directly to the virtual machine

If you have datastored on these types of disks, it is advisable to use in-guest agents to protect the data on these disks. In-guest data protection solutions require the deployment of backup-and-restore software in the guest machine. With the in-guest backup method, the virtual machine is treated as a physical system. You install the backup application on the guest operating system and start the backup operation on the virtual machine.

Note: At the time of writing, VMware virtual volume (VVols) datastores and their associated virtual machines cannot be backed up with Spectrum Protect Snapshot for VMware.

5.3.2 Integration with VMware vCenter Site Recovery Manager

Spectrum Protect Snapshot for VMware can protect datastores and virtual machines in VMware environments at both the primary site and secondary sites where VMware Site Recovery Manager is installed. The array-based replication can be Metro Mirror or Global Mirror with Change Volumes. Consider the following general guidelines for the configuration:

- Avoid protecting a virtual machine with SRM and Spectrum Protect Snapshot for VMware. If the virtual machine is restored by Spectrum Protect Snapshot on the secondary site, the SRM recovery plan will become invalid for this virtual machine.
- In an SRM environment, it is best to install Spectrum Protect Snapshot in a virtual machine and protect that virtual machine with SRM. This ensures that the Spectrum Protect Snapshot for VMware application, repository, and database are automatically replicated to the secondary site.
- The configuration of Spectrum Protect Snapshot for mirroring is the same with or without SRM.

For more information about the IBM Spectrum Protect Snapshot for VMware integration with SRM, see the IBM Spectrum Protect Snapshot web page in the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/SS36V9_4.1.3/fcm.common/welcome.html

6

Data reduction considerations

IBM FlashSystem V9000 includes advanced features for storage efficiency, such as thin provisioning and IBM Real-time Compression, which can be used to reduce the physical storage capacity required by virtualized workloads.

Thin provisioning and Real-time Compression increase the effective usable capacity of FlashSystem V9000 beyond the physical capacity. This helps with the economics of flash and can help to provide flash capacity for less than the cost of traditional disk capacity.

VMware vSphere also includes a form of thin provisioning that allows administrators to use thin-provisioned virtual machine disks.

In the following topics in this chapter, we examine the use of these functions:

6.1 Thin provisioning

Thin provisioning is a method for optimizing the use of available physical storage by allocating physical blocks of data on demand as a volume consumes it, versus the traditional method of allocating all physical blocks for the volume upon creation. This method helps you avoid the poor usage rates that can occur with the traditional storage allocation method, where large pools of storage capacity are allocated to individual servers but a large portion of it remains unused.

6.1.1 FlashSystem V9000 thin provisioning

Thin provisioning is included on FlashSystem V9000 and can be seamlessly implemented with VMware. Thin-provisioned volumes can be created and provisioned to VMware and used as datastores, or existing volumes can be nondisruptively converted to thin volumes with volume mirroring.

A feature of FlashSystem V9000 thin provisioning is the ability to present volumes to hosts with more capacity than is physically available in the storage pool. An example of this is when a storage system contains 5000 GB of usable storage capacity, but the storage administrator mapped volumes of 500 GB each to 15 hosts as shown in Figure 6-1. In this example, the storage administrator makes 7500 GB of storage space visible to the hosts, even though the storage system has only 5000 GB of usable space.

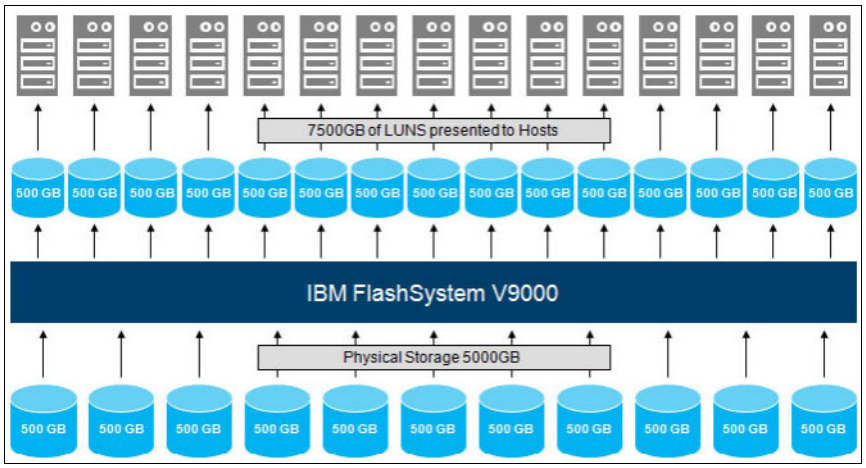


Figure 6-1 Concept of thin provisioning

FlashSystem V9000 volumes can be configured as *thin-provisioned* or *fully allocated*. thin-provisioned volumes are created with real and virtual capacities.

Real capacity defines how much disk space is allocated to a volume. *Virtual capacity* is the capacity of the volume that is reported to other IBM FlashSystem V9000 components (such as FlashCopy or remote copy) and to the hosts.

As Figure 6-2 shows, you can create a volume with real capacity of only 100 GB but virtual capacity of 1 TB. The actual space physical flash capacity that is used by the volume on FlashSystem V9000 will be 100 GB but hosts will see a 1 TB volume. The default for real capacity is 2% of the configured virtual capacity.

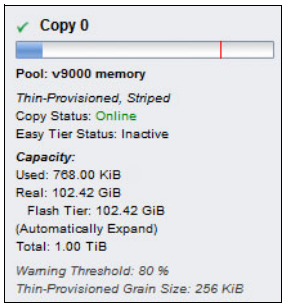


Figure 6-2 Example of a thin-provisioned volume

Thin-provisioned volumes are available in two operating modes:

- *Autoexpand* mode
- *Non-autoexpand* mode

The operating mode controls whether real capacity is automatically added to a thin-provisioned volume when it is required (autoexpand) or whether storage administrator intervention is needed to increase real capacity available to a thin-provisioned volume (non-autoexpand). A volume that is created without autoexpand mode and has zero contingency capacity goes offline when the real capacity is used and must expand.

6.1.2 VMware vStorage thin provisioning

VMware vStorage thin provisioning is a mechanism that allows virtual machines on VMware vSphere hosts to allocate the storage capacity needed for virtual machine disks, but only consume space that is needed for new write operations. VMware vStorage thin provisioning helps to increase storage use by eliminating the space that is often provisioned for virtual machines, but not consumed, and by allowing VMware administrators to over-commit storage provisioned to virtual machines versus what is available in physical VMware datastores as shown in Figure 6-3.

VMware vStorage thin provisioning operates at the virtual machine disk (VMDK) level, meaning over-commitment is on a per VMware datastores level. Blocks in the VMDK file are allocated as they are written. The VMware virtual machine file system (VMFS) driver manages the allocation and consumption within the VMware datastores.

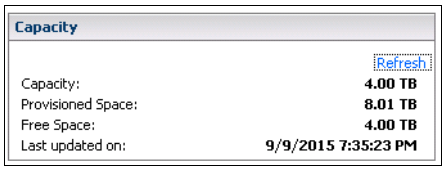


Figure 6-3 Over-allocated VMware datastore

Note: When the capacity of a VMware datastore in use reaches 100%, any thin-provisioned VMDKs that need additional capacity allocated are paused, and the virtual machine goes offline until additional capacity is available.

6.1.3 Using FlashSystem V9000 thin provisioning with VMware

Whether implementing thin provisioning on FlashSystem V9000 or VMware vSphere storage, use is increased by over-allocating the physical capacity available. However, each approach presents differences in administration and monitoring when implemented.

Configuring a FlashSystem V9000 thin-provisioned volume

FlashSystem V9000 GUI simplifies the process of creating a thin-provisioned volume through a wizard-driven process that is fast and efficient, as shown in Figure 6-4.

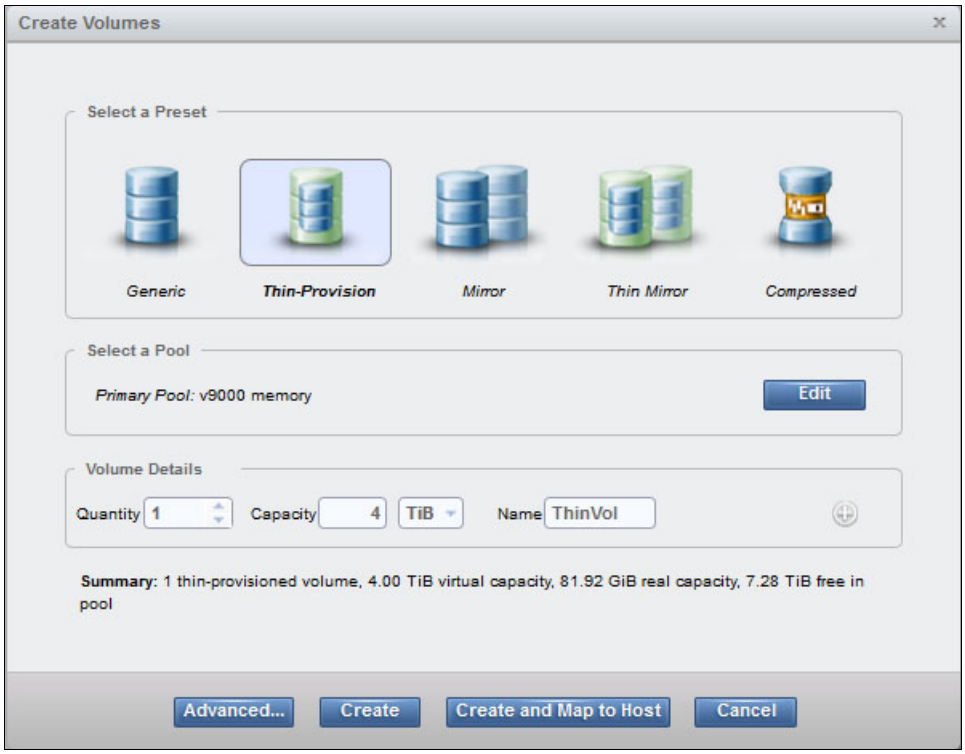


Figure 6-4 Configuring a thin-provision volume

The wizard provides the opportunity to edit the Advanced settings for the volume. Setting options include Real Capacity, Autoexpand Mode, Warning Threshold, and Thin-provisioned Grain Size, as shown in Figure 6-5.

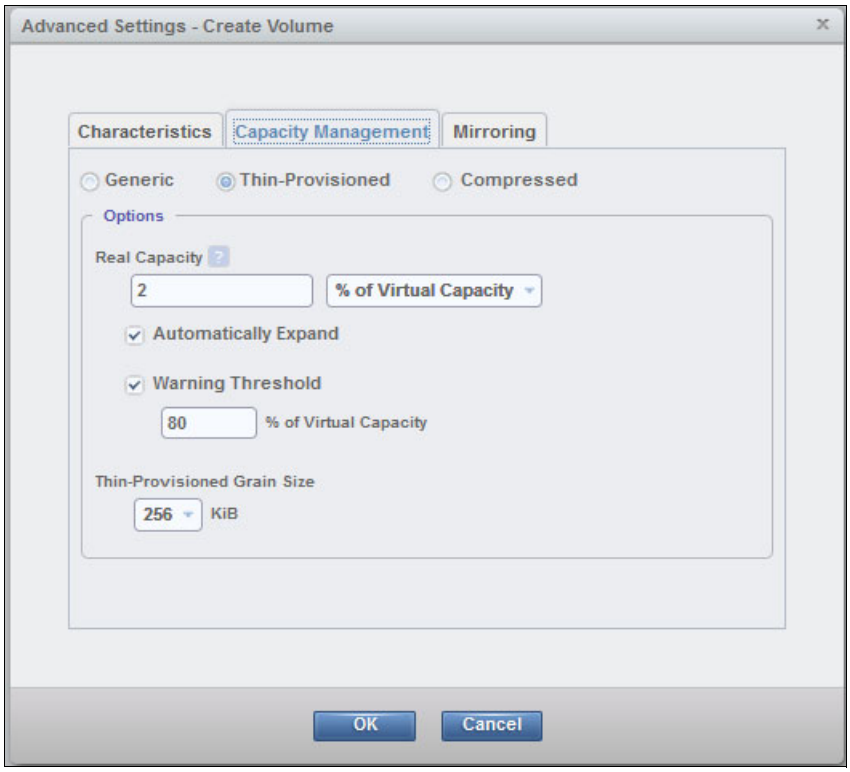


Figure 6-5 Advanced settings of a thin-provisioned volume

Advanced settings, such as Real Capacity, Autoexpand Mode, and Warning Threshold, are dependent upon the environment. For example, an organization that is agile and can quickly add physical storage capacity may choose to set the Warning Threshold to a value higher than an organization that takes more time to make changes.

Thin Provisioned Grain Size is a thin-provisioned volume attribute that affects performance and storage savings. A smaller grain size results in more efficient thin-provisioning, but a larger grain size maximizes performance. The best approach for VMware vSphere environments is to use the default 256 KiB grain size.

VMware VMDK type and FlashSystem V9000 thin provisioning

A VMware virtual machine can be created with three different types of VMDKs, as shown in Figure 6-6 on page 74:

- Thick provision lazy zeroed

  With this default VMDK type, the provisioned capacity for the VMDK is allocated on the VMware VMFS datastore but not zeroed out. When the virtual machine issues a new write operation to the VMDK, the host must first zero out the space.

- Thick provision eager zeroed

  With an eager-zeroed thick VMDK, all capacity is allocated on the VMFS datastore and the pre-zeroed when it is created.

- Thin provision

  A thin-provisioned VMDK consumes capacity on the VMFS datastore, because it is required by the virtual machine. New capacity must be allocated and zeroed before available to the guest.

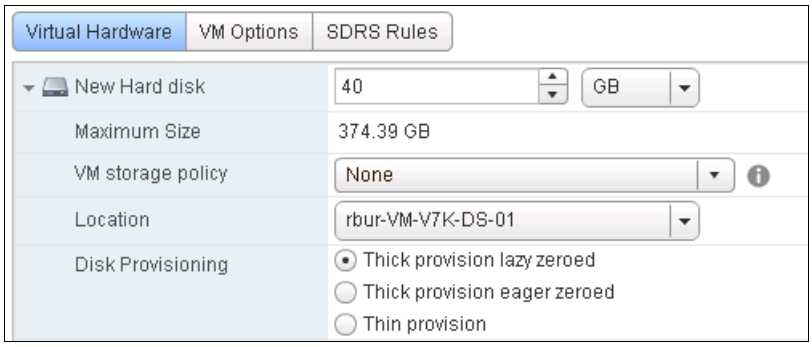


Figure 6-6 VMware VMDK types

FlashSystem V9000 has a zero write-host-detect feature that detects when servers are writing zeros to thin-provisioned volumes and ignores them rather than filling up the space on the thin volume. This means regardless of the VMDK type used on a FlashSystem V9000 thin-provisioned volume, only actual capacity used by the virtual machine is stored and consumes space, any zeros are discarded and not stored.

Tip: The optimal VMDK choice to use with FlashSystem V9000 is *thick provision eager zeroed*. Although fully allocated and pre-zeroed, this disk type will consume no more capacity on a FlashSystem V9000 thin volume, and it maximizes performance and lowers overhead because space does not have to be zeroed on demand.

Thin provisioning and managing stale data

Thin-provisioned volumes maximize storage use by only consuming physical storage space as it is needed. Over time and as data is generated, thin-provisioned volumes grow and the amount of consumed capacity increases. If that data on the thin-provisioned volume is moved or deleted (for example, if a VMware Storage vMotion operation or a virtual machine being deleted), the space on the thin-provisioned volume remains fully used. The behavior can also occur as data within virtual machines VMDKs is created and then deleted.

An example of the stale data phenomenon follows.

Figure 6-7 shows the relationship between the data on a file system such as VMFS and the underlying thin-provisioned storage volume.

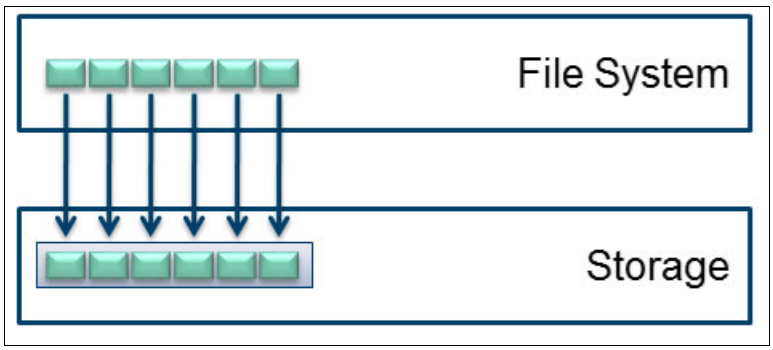


Figure 6-7 Thin provisioning storage mapping

If data was deleted from the file system, as in Figure 6-8, the data is no longer present in the file system. However, the capacity is still consumed within the thin-provisioned volume.

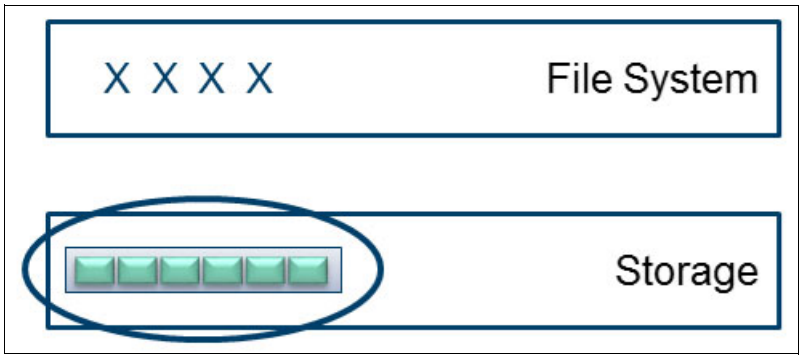


Figure 6-8 Stale data present on thin-provisioned volume after deleting data

When using thin-provisioned volumes it is important to understand that this behavior exists and for how to address it. FlashSystem V9000 includes a feature called zero detect, which provides clients with the ability to reclaim unused allocated disk space (zeros) by mirroring the thin-provisioned volume to another volume copy.

Follow this procedure to reclaim this capacity:

1. Capacity within the file systems must be zeroed out so that zero detect can identify free space.

   In a VMware vSphere environment, there can be multiple file systems. For example, a Microsoft Windows virtual machine might use the Microsoft New Technology file system (NTFS) within a VMDK, and that VMDK might be on a VMware VMFS datastore. Either or both of these layers might contain stale data.

   Microsoft has released a utility, at no change, that can be used to zero out the Microsoft Windows file system:

   `https://technet.microsoft.com/en-us/sysinternals/bb897443.aspx`

   The utility is command-line based and can be executed within a Windows virtual machine to zero out any unused capacity.

Figure 6-9 displays the command-line use for the `SDelete` utility.

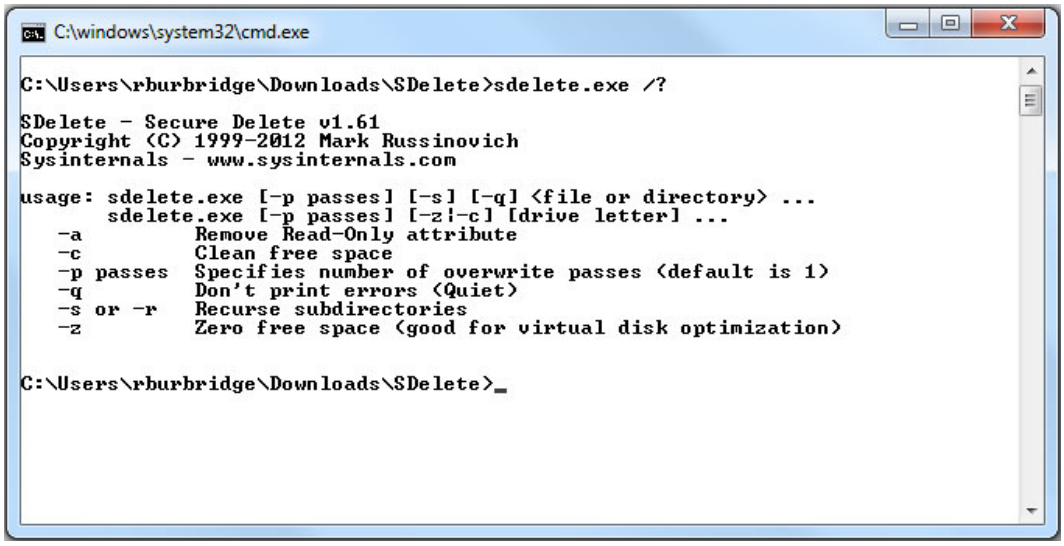


Figure 6-9 SDelete example

Note: Versions of Windows starting with Windows Server 2008 R2 include an automatic reclamation feature that uses SCSI UNMAP. This feature is not currently supported with FlashSystem V9000.

VMware vSphere has implemented reclaim functions by integrating with the SCSI-UNMAP operation, which is not currently supported by FlashSystem V9000.

Without SCSI-UNMAP, the method for pruning VMFS of any stale data is to ensure that any free capacity is zeroed out. This can be accomplished by creating a new virtual machine with an eager zeroed thick VMDK that consumes about 90% of the free space on the VMFS datastore. Zeros will be written for the entire capacity of the VMDK, which will clear the VMware VMFS of stale pointers.

For example, a VMFS datastore is 2.00 TB, and it is thought to contain stale data that is consuming capacity on a FlashSystem V9000 thin-provisioned volume:

- Total Capacity: 2.00 TB
- VMFS Consumed Capacity: 1.5 TB
- VMFS Free Capacity: 500 GB

To prune the stale data a VMDK should be created that consumes 450 GB (90% of the available 500 GB). After the VMDK is created, the virtual machine can be immediately deleted.

2. Space must be reclaimed on FlashSystem V9000 volumes.

   Volumes on FlashSystem V9000 can be seamlessly converted to thin, thick, or compressed. The feature which allows this is called volume mirroring. FlashSystem V9000 can maintain two copies of data for a single volume. While synchronizing the copies, such as a source thin-volume copy to a target thin-volume copy, the zero detect feature recognizes zero-based data within the source copy and does not copy it the target.

To create a thin-provisioned volume copy, complete the following steps:

a. Add the target thin-provisioned copy by right-clicking an existing volume in FlashSystem V9000 GUI and selecting **Volume Copy Actions** → **Add Mirrored Copy**, as shown in Figure 6-10.

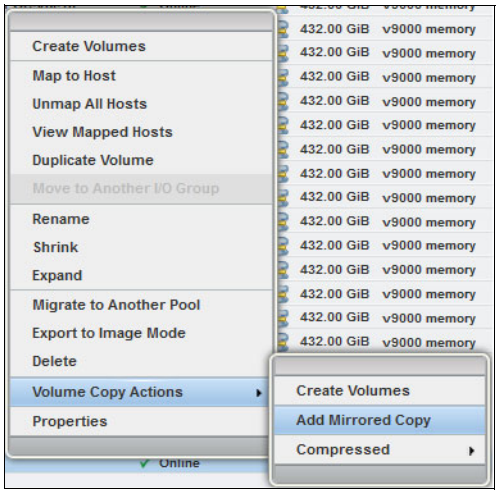


Figure 6-10 Adding a mirrored copy

b. Select the destination Storage Pool for the new thin-provisioned copy, and click **Add Copy** as shown in Figure 6-11.

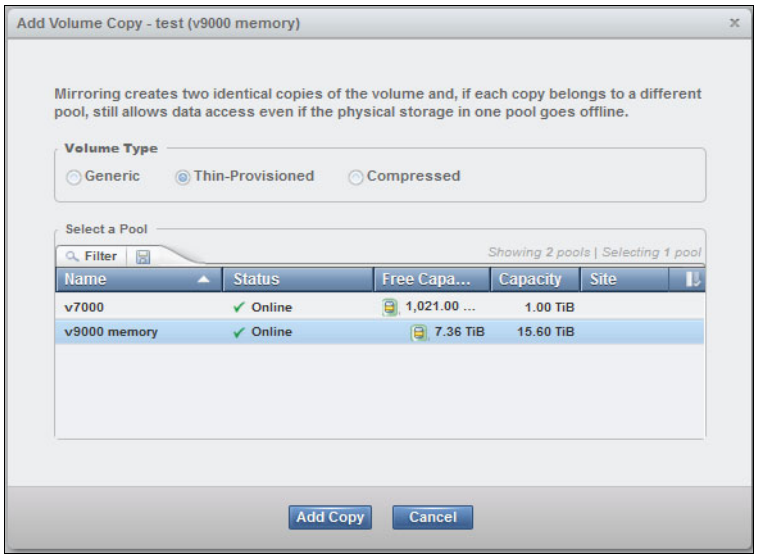


Figure 6-11 Selecting storage pool for the mirrored copy

c. Wait for synchronization of the copy to complete.

d. Remove the original source volume copy.

The necessity and frequency for using these procedures varies, based on the frequency of changes which occur in an environment.

6.2 Real-time Compression

IBM Real-time Compression software is embedded in FlashSystem V9000 addresses the requirements of primary storage data reduction, including performance. It does so by using purpose-built technology called Real-time Compression that uses the Random Access Compression Engine (RACE) engine.

It offers the following benefits:

- Compression for active primary data
- Compression for replicated or mirrored data
- No changes to the existing environment are required
- Overall savings in operational expenses
- Disk space savings are immediate

The space reduction with Real-time Compression occurs when the host writes the data so the reduction is real-time and inline. This process is unlike other compression solutions in which some or all of the reduction is realized only after a post-process compression batch job is run.

The license for compression is included in FlashSystem V9000 base license for internal storage. External storage capacity can be licensed for compression on a capacity basis, per terabyte of virtual data.

FlashSystem V9000 compressed volumes are by default also thin-provisioned, so capacity from the storage pool is consumed as needed versus up-front. Therefore, compressed volumes have the same advantages as thin-provisioned, with the added benefit of more space saved, because the consumed capacity is also compressed.

6.2.1 Using FlashSystem V9000 Real-time Compression with VMware

Compressed volumes are similar to thin-provisioned volumes in that capacity is provided as needed, however, with compressed volumes, the capacity being consumed has been compressed.

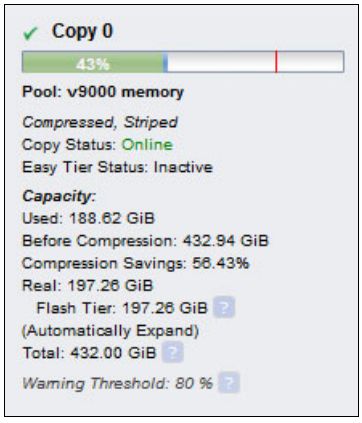


Figure 6-12 Compressed volume copy details

Figure 6-12 on page 78 displays the detailed information of a compressed volume copy and the following descriptions provide an overview of the items found in the capacity section:

Used Indicates the amount of capacity that has been used by the copy of the volume

Before Compression Indicates the amount of capacity of a compressed volume if it were not compressed

Compression Savings Indicates the amount of capacity that is saved by using compression

Real Indicates the capacity to be allocated to each copy of the volume

Total Indicates the capacity of the volume that is available to hosts

Creating a compressed volume uses the same GUI wizard that is used for creating a thin-provisioned volume. The advanced settings are also the same with the exception of the *Grain Size* setting, which is not available with a compressed volume because it uses a 32 KiB compressed chunk size.

Real-time Compression provides benefits to a variety of data types, many of which can be found running in a VMware environment. Table 6-1 shows typical compression results.

Table 6-1 Typical compression results.

Data type	Typical compression
Databases	50-80%
Server/Desktop Virtualization	40-75%
Collaboration Data	20-75%
Engineering Data	50-80%
E-mail	30-80%

IBM Compresstimator is a utility that can be used to estimate the compression rate that is achievable for existing data, including VMware. For more information, see the Comprestimator Utility web page:

http://www.software.ibm.com/webapp/set2/sas/f/comprestimator/home.html

Compressed volume guidelines for VMware

FlashSystem V9000 maximizes compression performance by incorporating hardware that is dedicated for Real-time Compression. Each FlashSystem V9000 control enclosure uses an 8-core CPU and 32 GB of memory for compression thread management and two Intel Quick Assist compression acceleration cards, based on the Coletto Creek chipset, for compression thread execution. This dedicated hardware improves performance and scalability for FlashSystem V9000 Real-time Compression.

Each FlashSystem V9000 volume corresponds to a compression thread, and a given thread is assigned and executes on one of the available compression accelerator cards. To effectively use the compression performance and scalability available on FlashSystem V9000, it is recommended to ensure a sufficient number of FlashSystem V9000 volumes are created and used for VMware.

It is advisable to provision at least 8 FlashSystem V9000 compressed volumes use them for VMware. In general, the more parallelism that can be added for the workload, the better the performance will be for FlashSystem V9000 compressed volumes.

VMware VMDK type and FlashSystem V9000 compressed volumes

The zero write-host-detect feature, which saves space to thin-provisioned volumes by ignoring zeros written from hosts, also benefits compressed volumes. Therefore, only actual capacity used by the virtual machine is stored and consumes space. Any zeros are discarded and not stored. As was the guidance for thin provisioning, it is best to use the thick provision eager zeroed thick VMDK type with FlashSystem V9000 compressed volumes.

An additional benefit of using the eager zeroed thick VMDK type with compressed volumes is that any time contiguous zeros are detected, as in the case of provisioning or cloning a new eager zeroed thick VMDK, Real-time Compression is able to automatically reclaim unused space. This is described further in the following section.

Real-time Compression and managing stale data

Compressed volumes have characteristics that are similar to thin-provisioned volumes, but one area in which they are different is in the management of stale data. The previous section, “Thin provisioning and managing stale data” on page 74, outlined the problem of stale data and how to address it with thin-provisioned volumes.

Although stale data can accumulate within a compressed volume as data is generated and deleted or migrated off, the space can be reclaimed without the need to mirror to a second volume copy. Similar to the procedure with thin-provisioned volumes, the first step for reclaiming storage space on FlashSystem V9000 is to ensure that the capacity within the file systems is zeroed out.

As previously indicated, there can be multiple file system layers where stale data is present. For example, Microsoft Windows NTFS within a virtual machine might contain stale data, and, in turn, the VMware VMFS that stores that virtual machine might contain more stale data.

The previously mentioned Microsoft Windows utility, SDelete, may be used to zero out Windows based file systems.

The method for pruning VMFS of any stale data is to ensure that any free capacity is zeroed out. This can be accomplished by creating a new virtual machine with an eager zeroed thick VMDK that consumes about 90% of the free space on the VMFS datastore. Zeros are written for the entire capacity of the VMDK, which clears the VMware VMFS of pointers to stale data, as the following example explains:

1. A 4 TB VMware datastore was filled with virtual machines that consumed 2.36 TB, as shown in Figure 6-13.

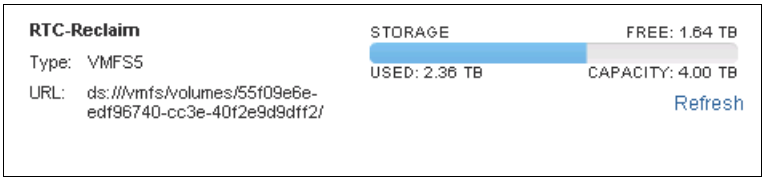


Figure 6-13 Initial VMware datastore capacity use

2. Space use after compression on FlashSystem V9000 compressed volume was 381.05 GB (354.88 GiB), as shown in Figure 6-14.

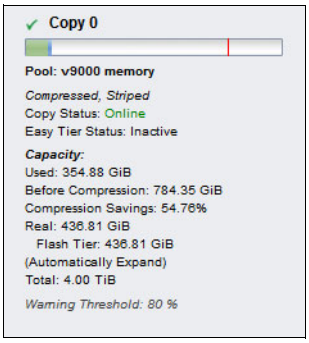


Figure 6-14 Initial FlashSystem V9000 volume capacity use

3. After deleting the virtual machines from the VMware datastore, the datastore space use was reduced to 24.97 GB as shown in Figure 6-15.

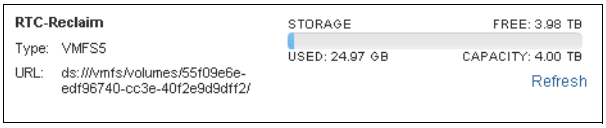


Figure 6-15 VMware datastore capacity use after removing virtual machines

However, FlashSystem V9000 compressed volume continues to recognize that the previously deleted capacity is consumed. An example of this can be observed in Figure 6-16.

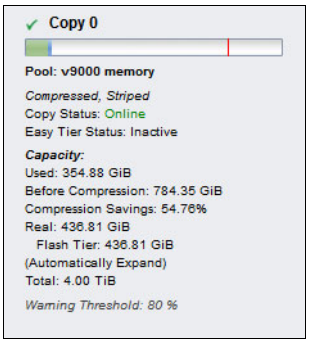


Figure 6-16 FlashSystem V9000 volume capacity use after deleting virtual machines

4. A virtual machine was created with an eager zeroed thick VMDK that consumed about 90% of the free space on the VMware datastore. This resulted in most of the stale data pointers being erased with zeros.

   Figure 6-17 displays the settings that were used. After the VMDK creation is complete, it can be immediately deleted.

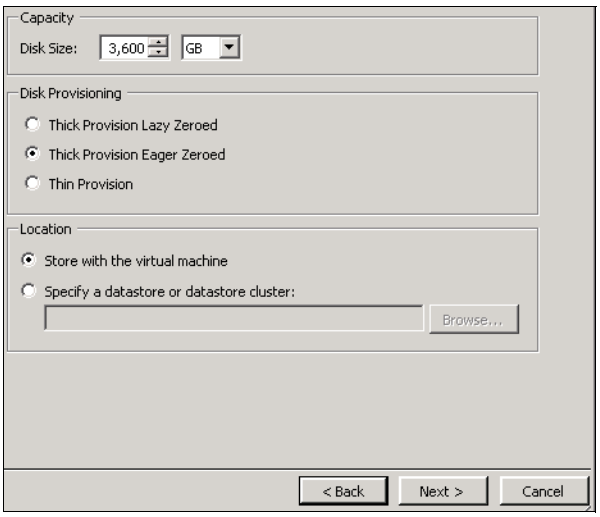


Figure 6-17 Creating eager zeroed thick VMDK

5. The Real-time Compression RACE engine runs continuous routine maintenance on compressed volumes. After zeroing out the VMware VMFS datastore, the RACE engine removes any unnecessary compressed chunks and resizes the compressed volume.

   Figure 6-18 displays the capacity use for the compressed volume after running creating and deleting the eager zeroed thick VMDK.

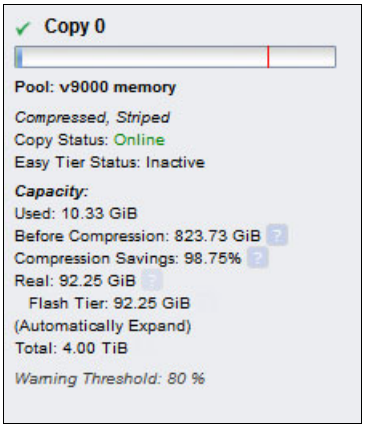


Figure 6-18 Compressed volume use after creating VMDK

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document (some publications referenced in this list might be available in softcopy only):

- *Introducing and Implementing IBM FlashSystem V9000*, SG24-8273
- *IBM SAN Solution Design Best Practices for VMware vSphere ESXi*, SG24-8158
- *IBM FlashSystem V9000 Product Guide*, TIPS1281

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM FlashSystem V9000, IBM Knowledge Center

  https://ibm.biz/fs_V9000_kc
- IBM Spectrum Control Base Edition version 2.1.1, IBM Knowledge Center

  http://www.ibm.com/support/knowledgecenter/STWMS9_2.1.1/
- VMware documentation

  http://www.vmware.com/support/pubs/
- VMware technical papers

  http://www.vmware.com/resources/techresources/
- IBM FlashSystem Ecosystem solutions web page

  http://www.ibm.com/systems/storage/flash/ecosystem/isv.html
- IBM System Storage Interoperation Center (SSIC)

  https://www.ibm.com/systems/support/storage/ssic/interoperability.wss

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5247-00

ISBN 0738454621

Printed in U.S.A.