

# IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation

Jon Tate

Angelo Bernasconi

Antonio Rainero

Ole Rasmussen



Storage





International Technical Support Organization

**IBM Storwize V7000, Spectrum Virtualize, HyperSwap,  
and VMware Implementation**

November 2015

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

**First Edition (November 2015)**

This edition applies to version 7.8 of IBM Storwize software, VMware 6.0, and any other product details that are indicated throughout this book.

© Copyright International Business Machines Corporation 2015. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

# Contents

<b>Notices</b> .....	vii
Trademarks .....	viii
<b>IBM Redbooks promotions</b> .....	ix
<b>Preface</b> .....	xi
Authors .....	xii
Now you can become a published author, too! .....	xiii
Comments welcome .....	xiii
Stay connected to IBM Redbooks .....	xiv
<b>Chapter 1. Introduction</b> .....	1
1.1 IBM Storwize V7000 .....	2
1.2 Storwize V7000 HyperSwap function .....	4
1.3 Comparison with Enhanced Stretched Cluster .....	9
1.4 Integration of VMware with Storwize V7000 .....	11
1.4.1 VMware and Storwize V7000 HyperSwap overview .....	12
1.4.2 VMware Metro Storage Cluster (vMSC) .....	13
1.4.3 Benefits of this integrated solution .....	13
1.4.4 Benefits in more detail .....	13
<b>Chapter 2. Hardware and software description</b> .....	15
2.1 Hardware description .....	16
2.2 IBM System Storwize V7000 .....	16
2.3 SAN directors and switches .....	16
2.3.1 SAN384B-2 and SAN768B-2 directors .....	17
2.3.2 SAN24B-5, SAN48B-5, and SAN96B-5 switches .....	17
2.4 FCIP routers .....	19
2.4.1 8 Gbps Extension Blade .....	19
2.4.2 SAN06B-R extension switch .....	20
2.5 Software high availability .....	20
2.6 VMware ESX and VMware ESXi .....	20
2.6.1 VMware vSphere .....	21
2.6.2 vSphere vMotion .....	21
2.6.3 vSphere High Availability .....	21
2.6.4 VMware Distributed Resource Scheduler .....	22
<b>Chapter 3. IBM System Storwize V7000 HyperSwap architecture</b> .....	23
3.1 Storwize V7000 HyperSwap overview .....	24
3.2 Failure domains, sites, and controllers .....	25
3.3 Storwize V7000 active-active Metro Mirror .....	27
3.3.1 Active-active Metro Mirror prerequisites .....	31
3.3.2 Storwize V7000 HyperSwap read operations .....	31
3.3.3 Storwize V7000 HyperSwap write operations .....	33
3.3.4 Storwize V7000 HyperSwap configuration quorum disk .....	34
3.3.5 View management .....	36
3.3.6 Storwize V7000 cluster state and voting .....	37
3.3.7 Quorum disk requirements .....	38
3.3.8 IP Quorum .....	39

3.3.9 Failure scenarios in a HyperSwap configuration . . . . .	40
3.4 Storwize V7000 HyperSwap configurations . . . . .	42
3.4.1 No ISL configuration . . . . .	43
3.4.2 ISL configuration . . . . .	46
3.4.3 FCIP configuration . . . . .	50
3.5 Fibre Channel settings for distance . . . . .	52
<b>Chapter 4. Implementation . . . . .</b>	<b>55</b>
4.1 Test environment . . . . .	56
4.2 IBM Fibre Channel SAN . . . . .	58
4.2.1 Logical switches and virtual fabric configuration . . . . .	59
4.2.2 Zoning configuration . . . . .	62
4.3 Storwize V7000 HyperSwap planning . . . . .	64
4.3.1 Active-active Metro Mirror considerations . . . . .	65
4.3.2 Quorum disk considerations . . . . .	67
4.3.3 IP Quorum . . . . .	69
4.4 Storwize V7000 HyperSwap configuration . . . . .	74
4.4.1 Configuring the HyperSwap system topology using the CLI . . . . .	74
4.4.2 Configuring the HyperSwap system topology using the GUI . . . . .	81
4.4.3 Configuring the HyperSwap volumes using the CLI . . . . .	88
4.4.4 Configuring the HyperSwap volumes using the GUI . . . . .	94
4.4.5 Summary . . . . .	103
<b>Chapter 5. VMware . . . . .</b>	<b>105</b>
5.1 VMware configuration checklist . . . . .	106
5.2 VMware with Storwize V7000 HyperSwap . . . . .	108
5.3 VMware vCenter setup . . . . .	108
5.3.1 Metro vMotion vMSC . . . . .	109
5.4 ESXi host installations . . . . .	109
5.4.1 ESXi host bus adapter requirements . . . . .	110
5.4.2 Initial ESXi verification . . . . .	110
5.4.3 Path selection policies (PSP) and native multipath drivers (NMP) . . . . .	111
5.4.4 Set the maximum number of logical unit numbers . . . . .	113
5.4.5 Set the default path selection policy (PSP) . . . . .	114
5.4.6 Verifying Node ID path in vSphere web client . . . . .	115
5.4.7 Path failover behavior for an invalid path . . . . .	116
5.4.8 VMware Distributed Resource Scheduler . . . . .	116
5.5 Naming conventions . . . . .	119
5.6 VMware vSphere High Availability . . . . .	123
5.6.1 High availability admission control . . . . .	123
5.6.2 High availability heartbeat . . . . .	123
5.6.3 HA advanced settings . . . . .	124
5.6.4 Enhanced All Paths Down detection in vSphere 6 . . . . .	125
5.6.5 Permanent Device Loss (PDL) . . . . .	126
5.6.6 Virtual Machine Component Protection (VMCP) . . . . .	127
5.6.7 Storage failure detection flow . . . . .	130
5.7 VMware vStorage API for Array Integration . . . . .	131
5.8 vCenter Services protection . . . . .	132
5.8.1 vCenter availability solution: Windows Server Failover Clustering . . . . .	133
5.9 VMware recovery planning . . . . .	134
5.9.1 VMware alternatives to minimize the impact of a complete site failure (split-brain scenario) . . . . .	134
5.9.2 Investigating a site failure . . . . .	135

5.10 Design comments . . . . .	137
5.11 Script examples. . . . .	138
5.11.1 PowerShell test script to move VMs 40 times between two ESXi hosts . . . . .	138
5.11.2 PowerShell script to extract data from the entire environment to verify active and preferred paths . . . . .	139
<b>Chapter 6. Storwize V7000 HyperSwap diagnostic and recovery guidelines . . . . .</b>	<b>145</b>
6.1 Solution recovery planning . . . . .	146
6.2 Storwize V7000 recovery planning . . . . .	146
6.3 Storwize V7000 HyperSwap diagnosis and recovery guidelines . . . . .	151
6.3.1 Critical event scenarios and complete site or domain failure . . . . .	152
6.3.2 Storwize V7000 HyperSwap diagnosis guidelines . . . . .	152
6.3.3 Storwize V7000 HyperSwap recovery guidelines . . . . .	160
6.4 Other disaster recovery with HyperSwap . . . . .	171
6.4.1 Continue to use the copy that hosts currently access. . . . .	173
6.4.2 Go back to the up-to-date copy and discard the stale disaster recovery copy . .	174
6.4.3 Disaster recovery with Volume Groups. . . . .	174
6.4.4 Convert to single-copy volumes . . . . .	175
<b>Related publications . . . . .</b>	<b>177</b>
IBM Redbooks . . . . .	177
Other publications . . . . .	178
Online resources . . . . .	178
Help from IBM . . . . .	179





# 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](http://www.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:

AIX®	IBM Spectrum Accelerate™	Redbooks®
DS8000®	IBM Spectrum Archive™	Redbooks (logo)  ®
Easy Tier®	IBM Spectrum Control™	Storwize®
FlashCopy®	IBM Spectrum Protect™	System Storage®
Global Technology Services®	IBM Spectrum Scale™	Tivoli®
HyperSwap®	IBM Spectrum Storage™	XIV®
IBM®	IBM Spectrum Virtualize™	z/OS®
IBM FlashSystem®	Insight™	
IBM Spectrum™	Real-time Compression™	

The following terms are trademarks of other companies:

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.

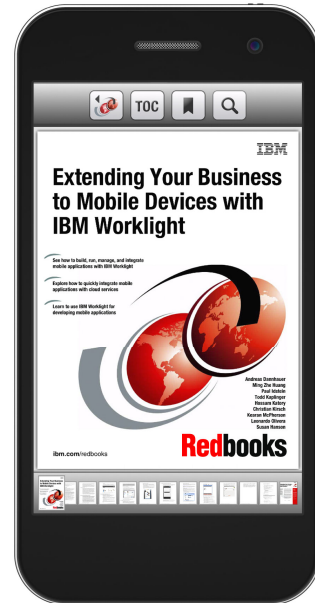
Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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**



## 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](http://ibm.com/Redbooks)  
About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK

# Preface

IBM® Spectrum Virtualize Software Version 7.8 provides software-defined storage capabilities across various platforms, including IBM SAN Volume Controller, IBM Storwize® V7000, Storwize V7000 (Unified), Storwize V5000, Storwize V3700, and Storwize V3500. These offerings help clients reduce the complexities and cost of managing their storage in the following ways:

- ▶ Centralizing management of storage volumes to enable administrators to manage storage volumes from a single point
- ▶ Improving utilization of storage capacity with virtual volumes to enable businesses to tap into previously unused disk capacity
- ▶ Avoiding downtime for backups, maintenance, and upgrades
- ▶ Performing data migration without disruption to applications
- ▶ Enabling all storage devices to be organized into storage pools from which virtual volumes, whether standard, compressed, or thin-provisioned, are created with the characteristics that you want
- ▶ Delivering automation of storage management with SmartCloud Virtual Storage Center, IBM Tivoli® Storage Productivity Center (as applicable by platform), and IBM Tivoli Storage FlashCopy® Manager (as applicable by platform)
- ▶ Increasing the performance efficiency of storage pools with IBM Easy Tier®
- ▶ Restoring data access quickly with near and remote copy capabilities across Fibre Channel (FC), Fibre Channel over Ethernet (FCoE), and IP networks

In this IBM Redbooks® publication, which is aimed at storage administrators and technical professionals, we describe the IBM HyperSwap® capability in IBM Spectrum™ Virtualize Software V7.8. HyperSwap delivers high availability (HA) and disaster recovery (DR) in one solution and reuses capital investments to achieve a range of recovery and management options that are transparent to host operations.

This book describes how you can use HyperSwap with VMware to create an environment that can withstand robust workloads.

## Authors

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



**Jon Tate** is a Project Manager for IBM Storage at the International Technical Support Organization (ITSO), San Jose Center. Before Jon joined the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2/3 support for IBM storage products. Jon has 29 years of experience in storage software and management, services, and support. He is an IBM Certified IT Specialist, an IBM SAN Certified Specialist, and a Project Management Professional (PMP). He also serves as the UK Chairman of the Storage Networking Industry Association (SNIA).



**Angelo Bernasconi** is an Executive Certified, Storage, SAN, and Storage Virtualization IT Specialist. He is a member of IBM Italy TEC. Angelo is a CTS Storage FTSS at IBM System Italy. He has 29 years of experience in the delivery of maintenance, professional services, and solutions for IBM Enterprise clients in z/OS®, and for the last 14 years, he focused on open systems. He holds a degree in electronics. His areas of expertise include storage hardware, storage area networks (SANs), storage virtualization design, solutions, implementation, DR solutions, and data deduplication. Angelo writes extensively about SAN and storage products in IBM Redbooks publications and white papers. He is also a member of the Storage Networking Industry Association (SNIA) Italian committee.



**Antonio Rainero** is a Consulting IT Specialist working for the IBM Global Technology Services® organization in IBM Italy. He joined IBM in 1998 and has more than 15 years of experience in the delivery of storage services for Open Systems and z/OS clients. His areas of expertise include storage systems implementation, SANs, storage virtualization, performance analysis, disaster recovery, and high availability solutions. He has co-authored several IBM Redbooks publications. Antonio holds a degree in Computer Science from University of Udine, Italy.



**Ole Rasmussen** is an IT Specialist working in IBM Strategic Outsourcing in Copenhagen, Denmark. He joined IBM during the transition of a large Danish bank's IT departments to IBM in 2004 - 2005, where he worked as a Technical Architect and Technical Specialist. He has worked in the customer decentralized area since 1990, and his primary focus is on Microsoft Windows and clustering. He has been an evangelist for VMware since 2003 and has participated in the design and implementation of VMware from that time. He achieved VMware Certification VCP 4.1 in late 2011 and VCP 5.X in early 2012.

Many people contributed to this book. In particular, we thank the development and PFE teams in IBM Hursley, UK, especially the following contributors:

John Wilkinson  
Chris Canto  
**IBM Hursley, UK**

Special thanks to the Brocade Communications Systems staff in San Jose, California, for their unparalleled support of this residency in terms of equipment and support in many areas:

Silviano Gaona  
Brian Steffler  
Marcus Thordal  
Jim Baldyga  
**Brocade Communications Systems**

## 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 at:  
[ibm.com/redbooks/residencies.html](http://ibm.com/redbooks/residencies.html)

## Comments welcome

Your comments are important to us!

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

- ▶ Use the online **Contact us** review Redbooks form found at:

[ibm.com/redbooks](http://ibm.com/redbooks)

- ▶ Send your comments in an email to:

[redbooks@us.ibm.com](mailto:redbooks@us.ibm.com)

- ▶ Mail your comments to:

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>





# Introduction

Business continuity and continuous application availability are among the top requirements for many organizations today. Advances in virtualization, storage, and networking make enhanced business continuity possible. Information technology solutions can now be designed to manage both planned and unplanned outages and to take advantage of the flexibility, efficient use of resources, and cost savings that are available from cloud computing.

The IBM Storwize V7000 HyperSwap design offers significant functionality for maintaining business continuity in a VMware environment. With VMware vMotion, you can dynamically migrate applications across data centers without interrupting the applications.

By combining VMware vMotion and Storwize V7000 HyperSwap, IT organizations can implement a robust business continuity, disaster avoidance, and recovery solution for virtualized application environments.

This chapter includes the following sections:

- ▶ IBM Storwize V7000
- ▶ Storwize V7000 HyperSwap function
- ▶ Comparison with Enhanced Stretched Cluster
- ▶ Integration of VMware with Storwize V7000

## 1.1 IBM Storwize V7000

IBM Storwize V7000 is a modular storage system that is built upon the IBM Spectrum Virtualize™ technology that provides many advanced capabilities, such as virtualization of storage area network (SAN)-attached storage devices, replication functions, thin provisioning, automated tiering, and IBM Real-time Compression™.

IBM Storwize V7000 helps improve business application availability and greater resource use. The objective is to manage storage resources in your IT infrastructure and to ensure that they are used to the advantage of your business. These processes take place quickly, efficiently, and in real time and help you avoid increases in administrative costs.

The first generation of IBM Storwize V7000 was released in late 2010. Since then, many enhancements to the Storwize V7000 capabilities were introduced. Now, with the second generation released in 2014, the hardware component was improved substantially to support more advanced processors, more memory, and faster interconnects.

The Storwize V7000 supports attachment to servers through Fibre Channel (FC) protocols and Internet Small Computer System Interface (iSCSI) protocols over IP networks at 1 Gbps and 10 Gbps speeds. These configurations can help reduce costs and simplify server configuration. The Storwize V7000 also supports the Fibre Channel over Ethernet (FCoE) protocol.

The Storwize V7000 combines hardware and software in an integrated, modular solution that is highly scalable. The Storwize V7000 hardware building blocks are *control enclosures* and *expansion enclosures*. Each Storwize V7000 system has one control enclosure and up to four optional control enclosures that contain two redundant node canisters each and disk drives. Control enclosures run the Storwize V7000 software component and provide connectivity to hosts and external storage. Storwize V7000 configurations with multiple control enclosures are often referred to as a Storwize V7000 *clustered* system. Additional storage capacity can be provided by expansion enclosures, up to 20 for each control enclosure, containing disk drives and directly connected to control enclosures by using dedicated serial-attached SCSI (SAS) connections.

The configuration flexibility means that your implementation can start small and grow with your business to manage large storage environments. The scalable architecture and tight integration enable your business to take advantage of the high throughput of solid-state drives (SSDs). This high throughput supports high performance for critical applications.

The Storwize V7000 also includes the IBM System Storage® Easy Tier function, which helps improve performance at a lower cost through the more efficient use of SSDs. The Easy Tier function automatically identifies highly active data within volumes and moves only the active data to an SSD. It targets SSD use to the data that benefits the most, which delivers the maximum benefit even from small amounts of SSD capacity. The Storwize V7000 software helps move critical data to and from SSDs as needed without any disruption to applications.

By using storage virtualization, an organization can implement pools of storage across physically separate disk systems (which might be from different vendors). Storage can then be deployed from these pools. The storage can be migrated between pools without any outage of the attached host systems. Storage virtualization capabilities allow major changes in the storage infrastructure without affecting the applications, which helps your business improve customer service.

**IBM Spectrum Virtualize and Storwize V7000:** The IBM Spectrum Storage™ family name is used to denote the IBM portfolio of software-defined storage offerings as a whole. It is the anchor of our software-defined storage brand and encompasses a full range of solutions to help organizations achieve data without borders.

The portfolio includes these members:

IBM Spectrum Virtualize: Storage virtualization that frees client data from IT boundaries

IBM Spectrum Control™: Simplified control and optimization of storage and data infrastructure

IBM Spectrum Protect™: Single point of administration for data backup and recovery

IBM Spectrum Archive™: Enables easy access to long-term storage of low activity data

IBM Spectrum Scale™: High-performance, scalable storage manages yottabytes of unstructured data

IBM Spectrum Accelerate™: Accelerating the speed of deployment and access to data for new workloads

The SAN Volume Controller and Storwize family of products all run the same software, and that software is what provides all of the features and functions of our products. Until the IBM Spectrum Storage family was announced earlier this year, that software did not have a name for the entire portfolio. With the IBM Spectrum Storage family, the software that powers the SAN Volume Controller and Storwize products now has a name: IBM Spectrum Virtualize.

Figure 1-1 shows an overview of the storage virtualization capabilities.

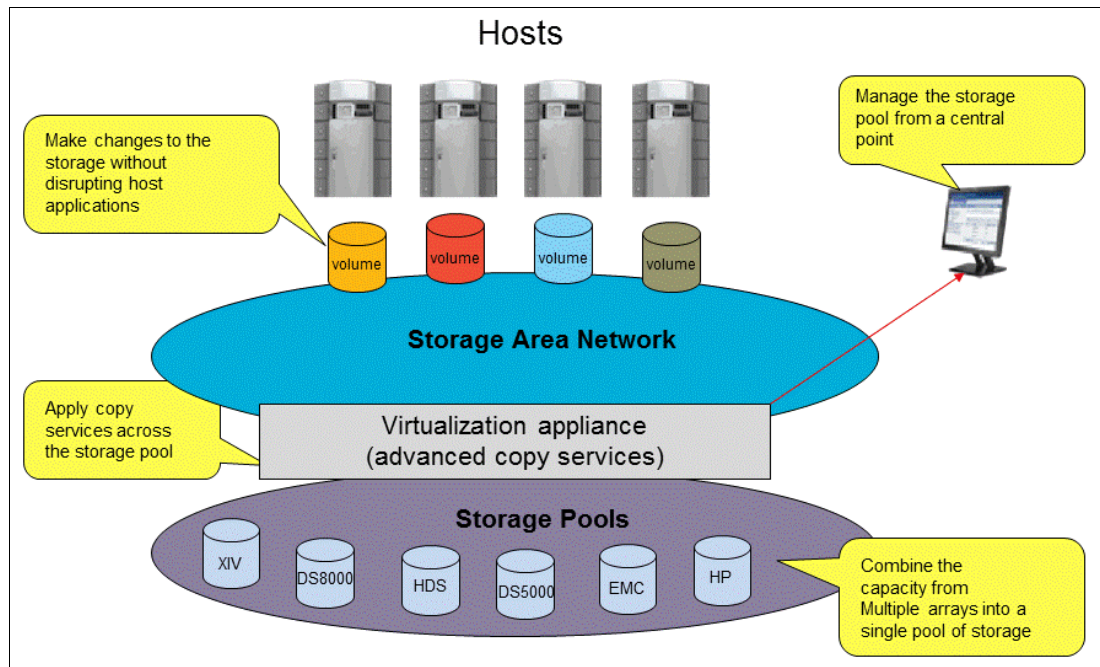


Figure 1-1 Storage virtualization

The Storwize V7000 includes a dynamic data migration function to move data from one storage system to another, including itself, yet maintain access to the data. The Volume Mirroring function stores two copies of a volume on different storage pools. This function helps improve application availability in a failure or during disruptive maintenance to an internal array or disk system.

The Metro Mirror and Global Mirror functions operate between Storwize V7000 systems at different locations. They help create copies of data for use in a catastrophic event at a data center. For even greater flexibility, Metro Mirror and Global Mirror also support replication between Storwize V7000 systems.

The IBM FlashCopy function quickly creates a copy of active data to use for backup or parallel processing activities. With this capability, you can use disk backup copies to recover almost instantly from corrupted data, which significantly speeds up application recovery.

Figure 1-2 shows the IBM Storwize V7000 overview.

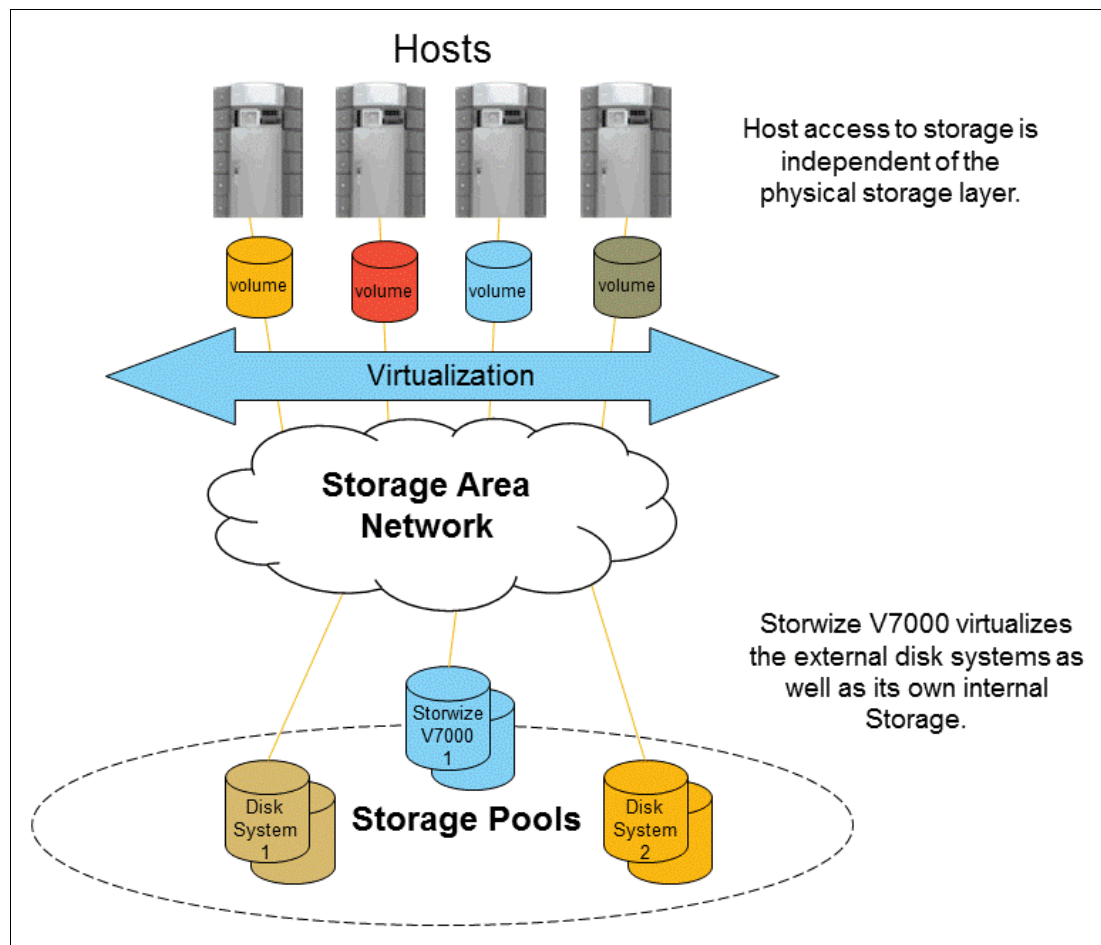


Figure 1-2 Storwize V7000 overview

## 1.2 Storwize V7000 HyperSwap function

*HyperSwap* is the high availability (HA) solution for IBM Storwize V7000 that provides continuous data availability in case of hardware failure, power failure, connectivity failure, or

disasters. The HyperSwap capabilities are also available on other IBM storage technologies, such as IBM Storwize V5000 and IBM FlashSystem® V9000.

Before version 7.5, the only available data protection feature on Storwize V7000 was the Volume Mirroring feature that provides data availability in a failure of internal or external storage. However, the Volume Mirroring feature does not provide any protection in a loss of the Storwize V7000 cluster or control enclosures. To protect data against the complete loss of storage systems, host-based data availability solutions can be implemented. These solutions rely on a combination of storage system and application or operating system capabilities. Usually, they delegate the management of the storage loss events to the host.

The IBM Storwize V7000 Version 7.5 introduced the HyperSwap technology that provides an HA solution, which is transparent to the host, between two locations at up to 300 km (186.4 miles) apart.

A new Metro Mirror capability, the *active-active* Metro Mirror, is used to maintain a fully independent copy of the data 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 the data that is transmitted between sites and to minimize host read and write latency.

The Storwize V7000 HyperSwap configuration requires that at least one control enclosure is implemented in each location. Therefore, a minimum of two control enclosures for each Storwize V7000 cluster are needed to implement the HyperSwap. Configuration with three or four control enclosures is also supported for the HyperSwap.

Figure 1-3 shows a typical Storwize V7000 HyperSwap configuration.

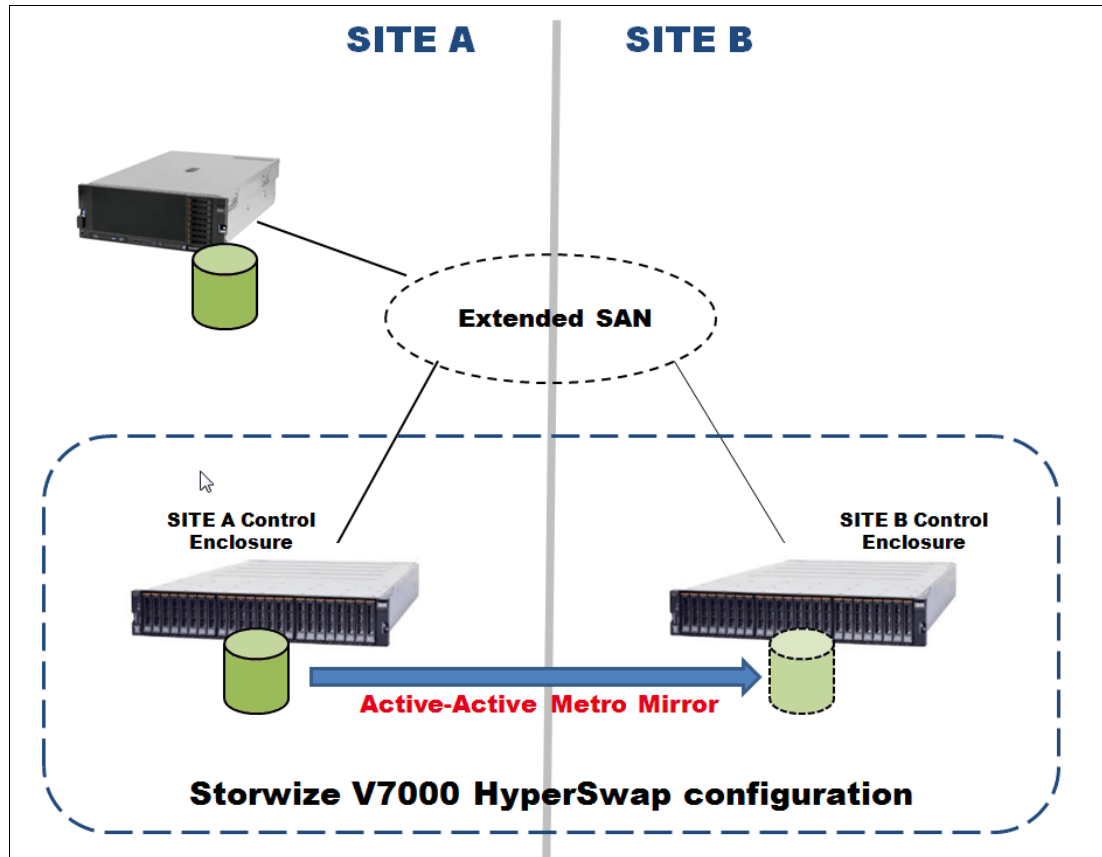


Figure 1-3 Typical Storwize V7000 HyperSwap configuration

With a copy of the data that is stored at each location, HyperSwap configurations can handle different failure scenarios.

Figure 1-4 shows how HyperSwap operates in a storage failure in one location. In this case, after the storage failure was detected in Site A, the HyperSwap function provides access to the data through the copy in the surviving site.

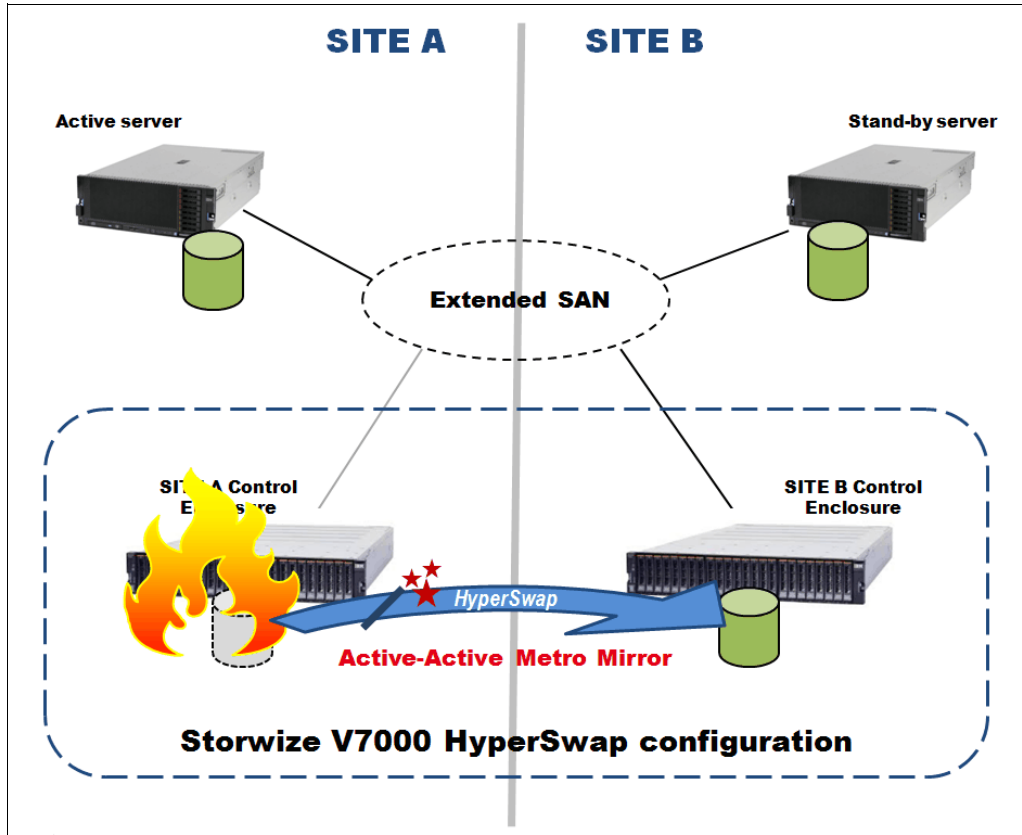


Figure 1-4 HyperSwap in a storage failure scenario

You can lose an entire location, and access to the disks remains available at the alternate location. The use of this behavior requires clustering software at the application and server layer to fail over to a server at the alternate location and resume access to the disks. This scenario is depicted in Figure 1-5. The active-active Metro Mirror feature provides the capability to keep both copies of the storage in synchronization. Therefore, the loss of one location causes no disruption to the alternate location.

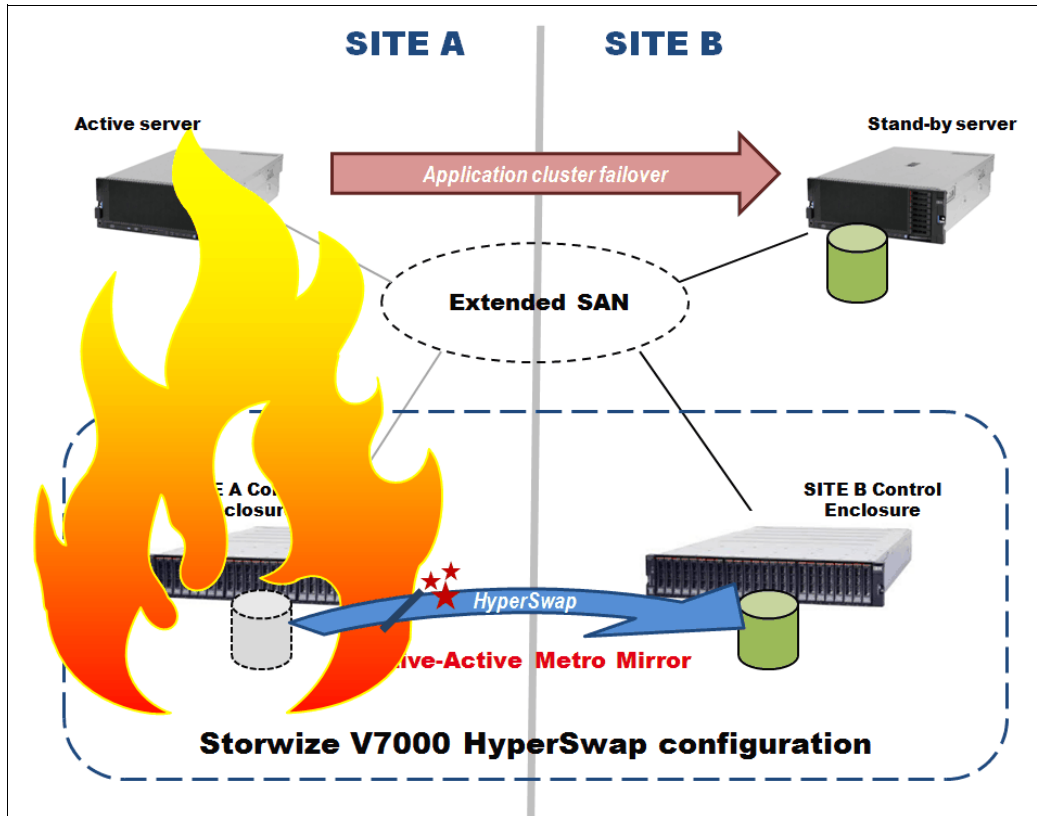


Figure 1-5 HyperSwap in a full-site failure scenario

By combining HyperSwap configurations with the Volume Mirroring and virtualization capabilities, Storwize V7000 provides even greater data availability in a partial site failure, as shown in the configuration that is depicted in Figure 1-6.

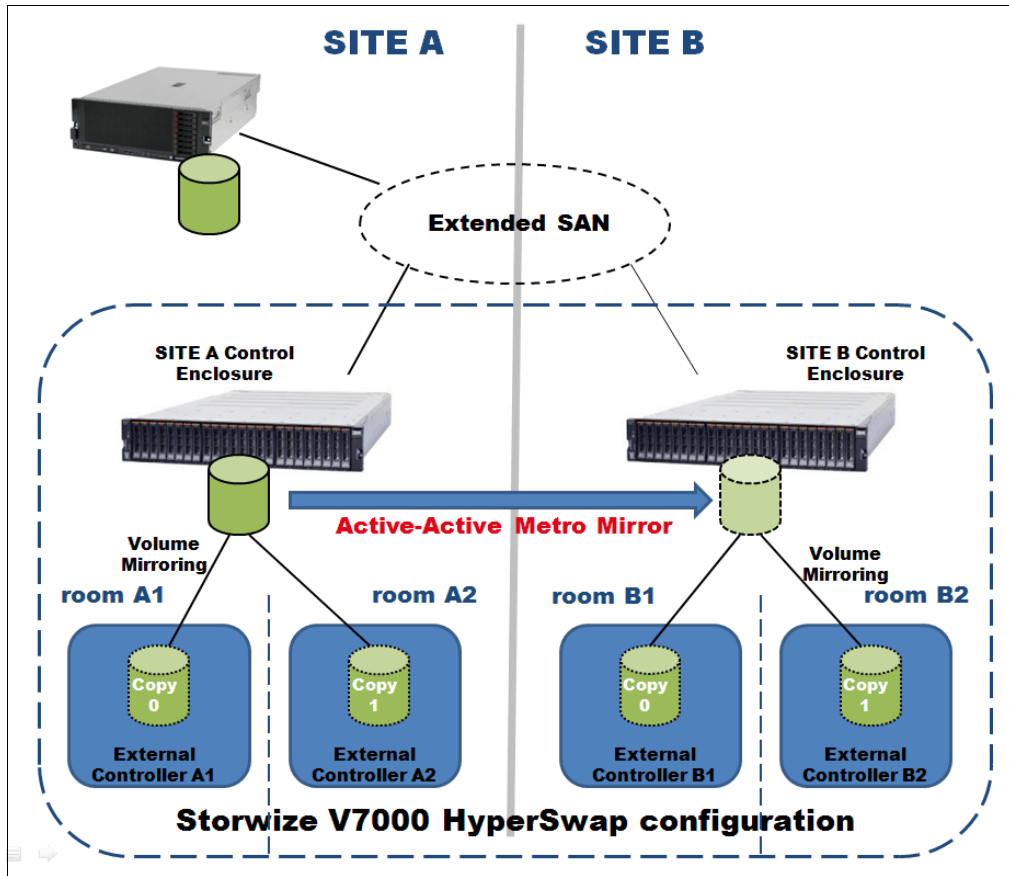


Figure 1-6 HyperSwap and Volume Mirroring combination

In addition to the active-active Metro Mirror feature, the HyperSwap feature also introduced the *site awareness* concept for node canisters, internal and external storage, and hosts. Finally, the HyperSwap *DR feature* allows us to manage rolling disaster scenarios effectively.

As with any clustering solution, avoiding a “split-brain” situation (where the control enclosures are no longer able to communicate with each other) requires a tiebreak. The Storwize V7000 HyperSwap configuration is no exception. Storwize V7000 software uses a tiebreak mechanism that is facilitated through the implementation of a quorum disk. It uses three quorum disks from the managed disks that are attached to the cluster to be used for this purpose. The tiebreak mechanism requires each location to provide one of quorum disks by using either internal or external storage capacity. The third quorum disk must be provided by external storage that is installed in a third location (*quorum site*). An additional external storage device is required to implement a Storwize V7000 HyperSwap configuration. Usually, the management of the quorum disks is apparent to the Storwize V7000 users.

**Note:** IBM Storwize V7000 Version 7.6 introduced the IP quorum feature that allows the use of network-based quorum capabilities. With this feature, the use of external storage for the tiebreak quorum is no longer required.

When the HyperSwap feature is enabled, the Storwize V7000 software can automatically choose quorum disks and place one quorum disk in each of the three sites. Users can still manually select quorum disks in each of the three sites, if they prefer.



With IP Quorum, introduced with version 7.6, the system will select automatically the IP Quorum App to be the active one.

Several requirements must be validated for the Storwize V7000 HyperSwap implementations, specifically for the SAN extension. For more information about HyperSwap prerequisites, see the IBM Storwize V7000 Knowledge Center:

<https://ibm.biz/BdX9Xz>

The Storwize V7000 HyperSwap function is supported in a fully virtualized environment (that is, not involving any internal storage in the HyperSwap configuration). You can use the storage controller of your choice at any of the three locations, and they can be from different vendors. The list of supported hardware and other interoperability information can be retrieved at the following link:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1005252>

**Licensing:** The HyperSwap function requires the Remote Mirroring license. No additional licenses are required for Storwize products that use external storage for the tiebreak quorum disk.

**Note:** The information in this book is based on the Storwize V7000 and VMware environment. However, the Storwize V7000 HyperSwap configuration can be applied to any other operating system and environment. These systems include native Microsoft Cluster, IBM AIX® Power HA, and Linux Cluster. All of the HyperSwap benefits and protection criteria provide data protection and business continuity requirements, regardless of the operating system that your application uses.

### 1.3 Comparison with Enhanced Stretched Cluster

The HyperSwap function is available with IBM Spectrum Virtualize. Most of the concepts and practices that are described in this book apply to devices that can run the HyperSwap function.

Since software version 7.2, an HA function that is called *Enhanced Stretched Cluster* (ESC) is available. Many HyperSwap concepts, such as site awareness and the disaster recovery (DR) feature, are in fact inherited from the ESC function. Nevertheless, important differences between the two solutions exist as summarized in Table 1-1.

Table 1-1 *Enhanced Stretched Cluster and HyperSwap comparison*

	Enhanced Stretched Cluster	HyperSwap
The function is available on these products.	Spectrum Virtualize only.	<ul style="list-style-type: none"> <li>▶ Spectrum Virtualize with two or more I/O Groups</li> <li>▶ Storwize V7000</li> <li>▶ Storwize V5000</li> <li>▶ FlashSystem V9000</li> </ul>
Configuration.	Command-line interface (CLI) or graphical user interface (GUI) on a single system; simple object creation.	Command-line interface (CLI) or graphical user interface (GUI) on a single system; simple object creation.

	<b>Enhanced Stretched Cluster</b>	<b>HyperSwap</b>
The number of sites on which data is stored.	Two.	Two.
Distance between sites.	Up to 300 km (186.4 miles).	Up to 300 km (186.4 miles).
Independent copies, which are maintained, of data.	Two.	Two (four if you use additional Volume Mirroring to two pools in each site).
Technology for host to access multiple copies and automatically fail over.	Standard host multipathing driver.	Standard host multipathing driver.
Cache that is retained if only one site is online?	No.	Yes.
Host-to-storage-system path optimization.	Manual configuration of preferred node for each volume before version 7.5; automatic configuration that is based on host site as HyperSwap from version 7.5.	Automatic configuration based on host site (requires Asymmetric Logical Unit Access (ALUA)/Target Port Group Support (TPGS) support from the multipathing driver).
Synchronization and resynchronization of copies.	Automatic.	Automatic.
Stale consistent data is retained during resynchronization for DR?	No.	Yes.
Scope of failure and resynchronization.	Single volume.	One or more volumes; the scope is user-configurable.
Ability to use FlashCopy with an HA solution.	Yes (although no awareness of the site locality of the data).	Limited: You can use FlashCopy maps with a HyperSwap Volume as a source; avoids sending data across link between sites.
Ability to use Metro Mirror, Global Mirror, or Global Mirror Change Volume with an HA solution.	One remote copy; it can maintain current copies on up to four sites.	No.
Maximum number of highly available volumes.	5,000	1,250
Minimum required paths for each logical unit (LUN) for each host port.	Two.	Four.
Minimum number of I/O Groups.	One I/O Group is supported, but it is not recommended.	Two.
Licensing.	Included in base product.	Requires Remote Mirroring license for volumes. Exact license requirements might vary by product.

The Enhanced Stretched Cluster function uses a *stretched* system topology, and the HyperSwap function uses a hyperswap topology. These topologies both spread the nodes of the system across two sites. Storage that is at a third site acts as a tiebreaking quorum device.

The topologies differ in how the nodes are distributed across the sites:

- ▶ For each I/O Group in the system, the stretched topology has one node on one site, and one node on the other site. The topology works with any number of 1-4 I/O Groups, but because the I/O Group is split into two locations, this topology is only available with Spectrum Virtualize.
- ▶ The hyperswap topology locates both nodes of an I/O Group in the same site, making this technology possible to use with either Storwize or SAN Volume Controller products. Therefore, to get a volume that is resiliently stored on both sites, at least two I/O Groups (or control enclosures) are required.

The stretched topology uses fewer system resources; therefore, a greater number of highly available volumes can be configured. However, during a disaster that makes one site unavailable, the system cache on the nodes of the surviving site will be disabled.

The hyperswap topology uses additional system resources to support a fully independent cache on each site; therefore, full performance is possible even if one site is lost. In certain environments, a hyperswap topology provides better performance than a stretched topology.

Both topologies allow the full configuration of the highly available volumes through a single point of configuration. The Enhanced Stretched Cluster function might be fully configured through either the GUI or the CLI. Starting with version 7.8, the HyperSwap function can be configured through the system CLI or GUI as well.

## 1.4 Integration of VMware with Storwize V7000

Virtualization is now recognized as a key technology for improving the efficiency and cost-effectiveness of a company's IT infrastructure. As a result, critical business applications are moved to virtualized environments. This process creates requirements for higher availability, protection of critical business data, and the ability to fail over and continue to support business operations in a local outage or widespread disaster.

IT departments can now run a secure migration of a live virtualized application and its associated storage between data centers with no downtime or user disruption to users, which means that managers can realize the following benefits:

- ▶ Disaster avoidance and recovery.
- ▶ Load balancing between data centers.
- ▶ Better use of a cloud infrastructure.
- ▶ Optimization of power consumption.
- ▶ The correct level of performance for applications can be maintained.
- ▶ The use of Virtual Machine Component Protections (VMCPs) that are new in vSphere 6.

The solution that is described in this book addresses these needs through the combination of VMware vMotion and Storwize V7000 HyperSwap capabilities.

Continuous access to data is provided by a Storwize V7000 HyperSwap configuration. vMotion provides the VM live migration capability. The combination of VMware vMotion and Storwize V7000 HyperSwap enables the design and implementation of a robust business continuity, disaster avoidance, and recovery solution for virtualized application environments.

### 1.4.1 VMware and Storwize V7000 HyperSwap overview

Figure 1-7 shows a simple overview of the stretched VMware cluster and the VMware stack with Storwize V7000 HyperSwap.

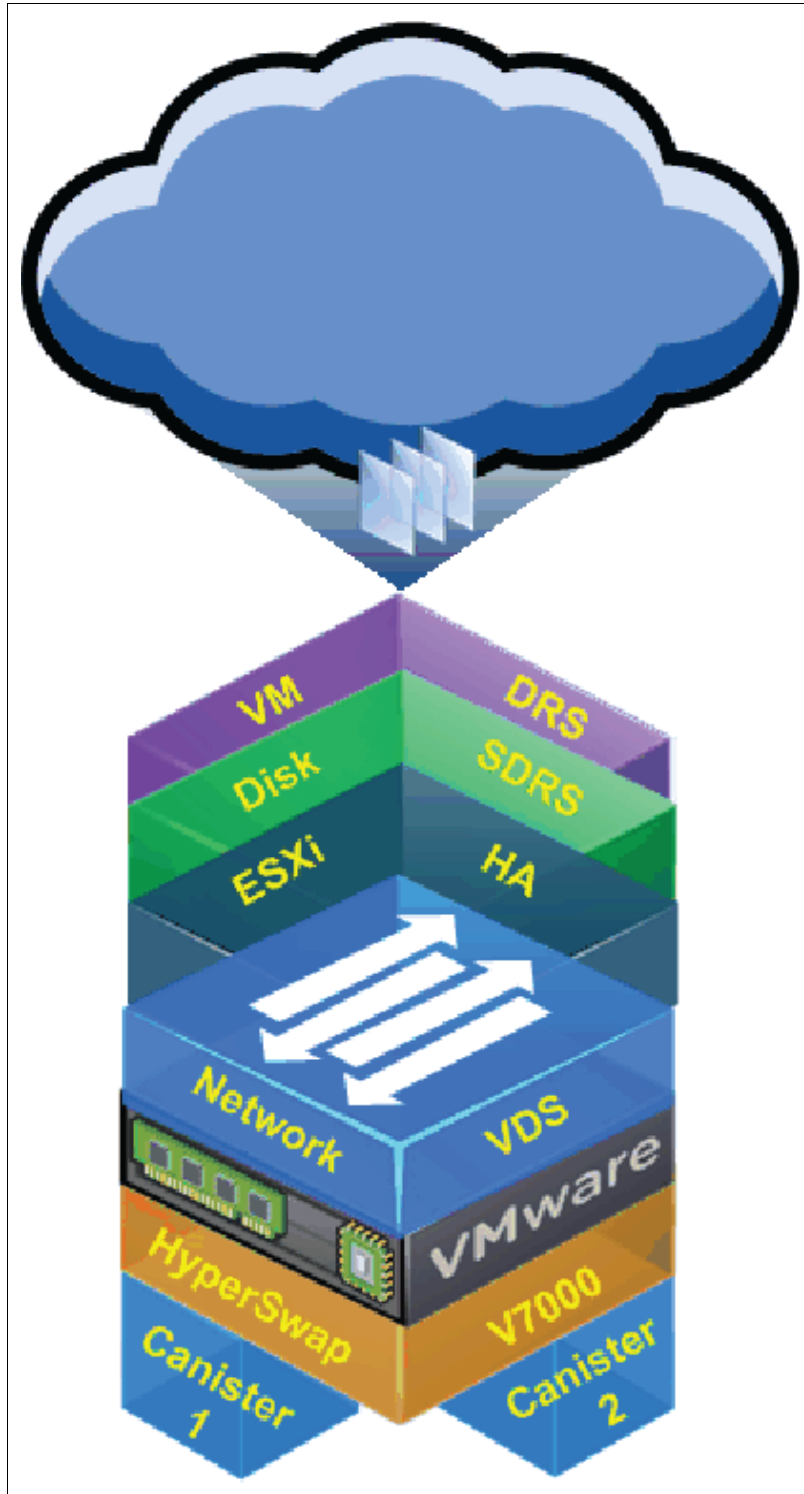


Figure 1-7 VMware and HyperSwap component stack

## 1.4.2 VMware Metro Storage Cluster (vMSC)

For the VMware guide to a Metro Storage Cluster (vMSC) in vSphere 6, see this website:

<https://ibm.biz/BdXCzf>

Storwize V7000 HyperSwap uses the vMSC base.

## 1.4.3 Benefits of this integrated solution

In the uniform ESX Cluster host access configuration, all ESXi hosts connect to storage cluster nodes in all sites, and paths stretch across the distance. This configuration offers certain benefits:

- ▶ Primary benefits:
  - Fully active-active data centers with balanced workloads
  - Disaster and downtime avoidance
- ▶ Secondary benefit
  - Useful during DR when combined with other processes

At the application layer, these tiers benefit from this configuration:

- ▶ Tier 0 applications, such as web servers in server farms
- ▶ Tier 1 - 3 applications benefit from this configuration, but not as much as a single Tier 0.

## 1.4.4 Benefits in more detail

The following benefits are described in more detail:

- ▶ Disaster avoidance

IT managers can use vMotion over distance to migrate applications in preparation for a natural disaster or a planned outage. Rather than recovering after the occurrence of the event, vMotion over distance helps avoid the disaster.

In vSphere 6, VMware delivers Long Distance vMotion, which allows up to 100 ms round-trip time (RTT). VMware calls this Cross Center vMotion. For more information, see this website:

<https://ibm.biz/BdXQYq>

This function is still supported in Storwize V7000 HyperSwap up to 5 ms RTT.

Disaster avoidance is preferable to DR whenever possible. Disaster avoidance augments DR. It provides IT managers with better control over when and how to migrate services.

- ▶ User performance and load balancing between data centers by using Distributed Resource Scheduler (DRS) from VMware.

In typical environments, significant data center capacity is set aside for spikes during peak demand. Backup and DR data centers are often idle. The solution is to relocate virtual server hotspots to underused data centers. This configuration increases the use of compute, network, and storage assets. Current assets are used as “spike insurance.” By moving workloads dynamically, as needed, you can use external cloud resources to handle loads during peak demand periods.

- ▶ Zero maintenance downtime for hardware maintenance

Eliminating downtime during maintenance is a key advantage that vMotion over distance offers. Virtual machines can be relocated to a remote data center during maintenance times so that users can access applications continuously.

- ▶ Zero maintenance downtime for applications

Zero maintenance downtime for applications can be achieved by using the VMware Fault Tolerant (FT) mechanism to secure up to four virtual CPUs in a 100% cloned active/active state. This approach requires special licenses and guaranteed network bandwidth requirements. For more information, see this website:

<https://ibm.biz/BdXQYi>

Also, by using Microsoft Clustering Services (MSCS) with VMware, we can maintain high service-level agreements (SLAs). A matrix of the supported combinations is at the VMware site:

<https://ibm.biz/BdXQZQ>



# Hardware and software description

This chapter describes the necessary hardware to implement an IBM Storwize V7000 HyperSwap configuration. It also briefly describes VMware and several useful features when you implement an IBM Storwize V7000 HyperSwap configuration.

This chapter includes the following sections:

- ▶ Hardware description
- ▶ IBM System Storwize V7000
- ▶ SAN directors and switches
- ▶ FCIP routers
- ▶ Software high availability
- ▶ VMware ESX and VMware ESXi

## 2.1 Hardware description

The following sections concentrate on the necessary hardware when you implement a Storwize V7000 HyperSwap configuration (V7000 HS). All of the products that are mentioned can provide the necessary functionality to implement a HyperSwap configuration. It is up to you to choose the most suitable product for your environment.

Consider these hardware factors when you implement a Storwize V7000 HyperSwap:

- ▶ Distance and latency between data centers
- ▶ Connectivity between data centers
- ▶ Bandwidth of the data that is sent
- ▶ Client's budget
- ▶ Current client infrastructure

All of these considerations can result in different hardware requirements. This section suggests hardware possibilities and provides guidelines for the features to purchase with that hardware.

## 2.2 IBM System Storwize V7000

A HyperSwap configuration requires the use of the IBM Storwize V7000. The controller provides an active-active storage interface that can allow for simple failover and failback capabilities during a site disruption or failure.

To implement an IBM Storwize V7000 HyperSwap, you can use either Storwize V7000 Gen1 or Storwize V7000 Gen2. Storwize V7000 Gen2 is shown in Figure 2-1.

Both of these models offer node capabilities that make them suitable for HyperSwap. Also, depending on the architecture that you want to deploy, you must be running at a minimum level of firmware. Check with your IBM service representative or see Chapter 3, “IBM System Storwize V7000 HyperSwap architecture” on page 23 to ensure that the Storwize V7000 node's model and firmware versions are supported for the environment that you want to implement.

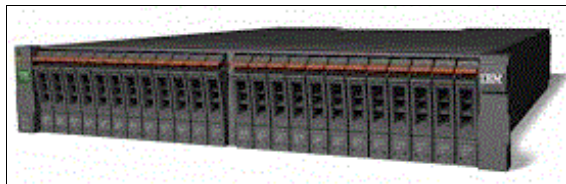


Figure 2-1 Storwize V7000 Gen2

## 2.3 SAN directors and switches

To implement an IBM Storwize V7000 HyperSwap solution, any storage area network (SAN) fabrics can be extended across two data centers, or site or failure domains, depending on the configuration that you choose. How you want to extend this fabric depends on the distance between failure domains. Your choices of architecture are outlined in Chapter 3, “IBM System Storwize V7000 HyperSwap architecture” on page 23.



This section does not address any particular wavelength division multiplexing (WDM) devices or any Ethernet infrastructure options other than Fibre Channel (FC) over IP (FCIP) devices. All of the hardware that is described is compatible with coarse wavelength division multiplexing (CWDM) devices (by using colored small form-factor pluggables (SFPs)), dense wavelength division multiplexing (DWDM) devices, and FCIP routers.

### 2.3.1 SAN384B-2 and SAN768B-2 directors

The IBM System Storage SAN384B-2 and SAN768B-2 directors provide scalable, reliable, and high-performance foundations for virtualized infrastructures. They increase business agility while they provide nonstop access to information and reduce infrastructure and administrative costs. The SAN768B-2 and SAN384B-2 fabric backbones, which are shown in Figure 2-2, have 6 gigabit per second (Gbps) FC capabilities and deliver a new level of scalability and advanced capabilities to this reliable, high-performance technology.

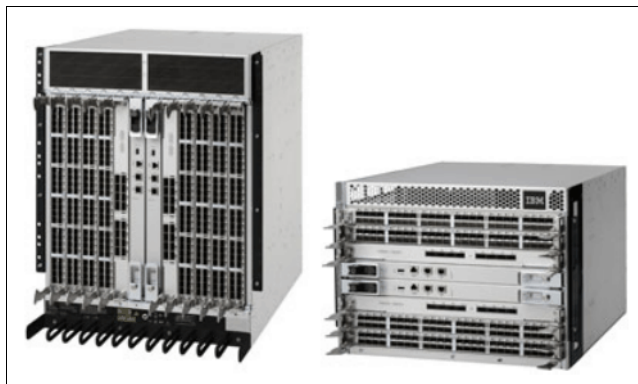


Figure 2-2 SAN768B-2 and SAN384B-2 fabric backbones

Both directors offer 16, 10, 8, 4, and 2 Gbps connections and they can have up to 512 or 256 ports. The enterprise software bundle that you get when you purchase directors includes the Extended Fabrics and Trunking features. The Extended Fabrics feature is essential for implementing a Storwize V7000 HyperSwap solution over 10 km (6.2 miles). The Trunking feature is necessary if multiple links are required to accommodate the bandwidth that is used for SAN traffic.

For more information, see this web page about IBM System Storage SAN384B-2 and SAN768B-2 directors:

<http://www.ibm.com/systems/networking/switches/san/b-type/san768b-2/index.html>

### 2.3.2 SAN24B-5, SAN48B-5, and SAN96B-5 switches

IBM System Storage offers a wide range of FC switches to suit various client data center needs and budgets. The IBM System Storage SAN24B-5, SAN48B-5, and SAN80B-4 are designed to support highly virtualized environments while also maintaining excellent cost-performance ratios.

#### **SAN24B-5 switch**

The IBM System Networking SAN24B-5 switch can be configured with 12 or 24 active ports. It offers 2, 4, 8, and 16 Gbps speeds in a 1U form factor. This switch is suited for smaller environments and for environments where a small performance switch is needed for Storwize V7000 HyperSwap solution node traffic.

The IBM System Networking SAN24B-5 switch is shown in Figure 2-3.



Figure 2-3 SAN24B-5 switch

When you implement a Storwize V7000 HyperSwap solution over 10 km (6.2 miles) with the SAN24B-5, you must purchase the Extended Fabrics feature to allow the switch to extend the distance of the links.

For more information about the IBM System Networking SAN24B-5 switch, see this web page:

<http://www.ibm.com/systems/networking/switches/san/b-type/san24b-5/index.html>

### **SAN48B-5 switch**

The IBM System Storage SAN48B-5 switch (Figure 2-4) can be configured with 24, 32, or 48 active ports. It offers 2, 4, 8, 10, and 16 Gbps speeds in a 1U form factor. The performance, reliability, and price of this switch make it a suitable candidate for an edge switch in large to mid-sized environments.



Figure 2-4 SAN48B-5 switch

When you implement a Storwize V7000 HyperSwap solution over 10 km (6.2 miles) with the SAN48B-5, you must purchase the Extended Fabrics feature to allow the switch to extend the distance of links.

For more information about the IBM System Storage SAN48B-5 switch, see this web page:

<http://www.ibm.com/systems/networking/switches/san/b-type/san48b-5/index.html>

### **SAN96B-5 switch**

The IBM System Storage SAN96B-5 switch, which is shown in Figure 2-5, offers 1, 2, 4, and 8 Gbps speeds. High availability (HA) features make this candidate ideal for a core switch in medium-sized environments and an edge switch in larger enterprise environments.



Figure 2-5 SAN96B-5 switch

The Extended Fabrics feature is enabled on the SAN96B-5 by default, which makes it useful for a Storwize V7000 HyperSwap solution over 10 km (6.2 miles). Because this switch is suited to larger environments, you might need to purchase the Trunking Activation license to ensure sufficient bandwidth between failure domains.

For more information about the IBM System Storage SAN96B-5 switch, see this web page:

<http://www.ibm.com/systems/networking/switches/san/b-type/san96b-5/>

## 2.4 FCIP routers

When you implement a Storwize V7000 HyperSwap solution over long distances, it is not always possible or feasible to extend SAN fabrics by using direct FC connectivity or WDM. Either the distance between the two failure domains is over 10 km (6.2 miles) or it is too expensive to lay cable or hire dark fiber service.

Many dual data center environments already have Internet Protocol (IP) connections between data centers. This configuration allows you to use FCIP technologies to enable the SAN fabric to extend across data centers while you use the existing infrastructure. When you implement a Storwize V7000 HyperSwap solution with Fibre Channel over IP (FCIP), minimum bandwidth requirements must be met to support the solutions. For more information, see Chapter 3, “IBM System Storwize V7000 HyperSwap architecture” on page 23.

### 2.4.1 8 Gbps Extension Blade

The 8 Gbps Extension Blade is an FCIP blade (Figure 2-6) that can be placed into both the SAN384B-2 and SAN768B-2 SAN directors. This blade uses 8 Gbps Fibre Channel, FCIP, and 10-Gigabit Ethernet (GbE) technology to enable fast, reliable, and cost-effective remote data replication, backup, and migration with existing Ethernet infrastructures.

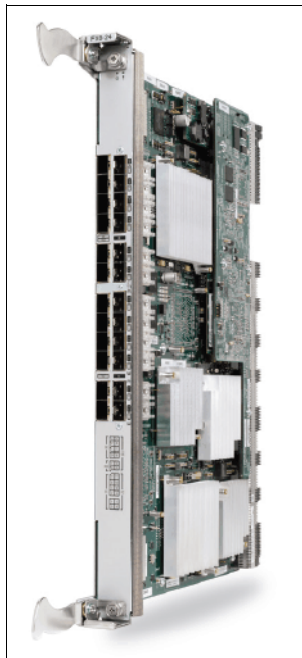


Figure 2-6 8 Gbps Extension Blade

The 8 Gbps Extension Blade has twelve 8 Gbps FC ports and ten 1 GbE Ethernet ports, by default. With the 8 Gbps Extension Blade 10 GbE Activation feature on the SAN384B-2 and SAN768B-2 directors, you can have two 10 GbE ports or ten 1 GbE Ethernet ports and one 10 GbE port on the blade. When you order this blade, you must also order the 8 Gbps Advanced Extension Activation feature on the SAN384B-2 and SAN 768B-2 directors.

## 2.4.2 SAN06B-R extension switch

The IBM System Storage SAN06B-R extension switch that is shown in Figure 2-7 optimizes backup, replication, and data migration over a range of distances by using both FC and FCIP networking technologies.



Figure 2-7 SAN06B-R extension switch

The SAN06B-R extension switch provides up to sixteen 8 Gbps FC ports and six 1 GbE ports to enable FCIP routing. To enable FCIP routing on the switch (the R06 Trunking Activation feature), you must also order the R06 8 Gbps Advanced Extension feature or the R06 Enterprise Package.

For information about the IBM System Storage SAN06B-R extension switch, see this page:

<http://www.ibm.com/systems/networking/switches/san/b-type/san06b-r/index.html>

## 2.5 Software high availability

When you implement a solution, such as an IBM Storwize V7000 HyperSwap, ensure that you provide availability for the application layer of the environment and the infrastructure layer. This availability maximizes the benefit that can be derived from both the storage infrastructure and the host operating systems and applications.

Many different software stacks can achieve host availability for applications. This book focuses on VMware and the features that the VMware ESXi and vSphere platforms provide. This section outlines VMware ESXi and vSphere and other features that are useful when you implement a stretched cluster solution.

## 2.6 VMware ESX and VMware ESXi

VMware ESX and VMware ESXi are hypervisors to abstract processor, memory, storage, and networking resources into multiple VMs that can run unmodified operating systems and applications. VMware ESX and VMware ESXi are designed to reduce server sprawl by running applications on virtual machines (VMs) that are made up of fewer physical servers.

VMware ESX and VMware ESXi hosts can be organized into clusters. This configuration allows ESX to provide flexibility in terms of the VMs that run on the physical infrastructure.

## 2.6.1 VMware vSphere

VMware vSphere is the management software suite that is used to manage the VMs inside an ESX or ESXi host. When you allocate resources, such as memory, storage, networking, or processors, to a VM, a vSphere vCenter server manages how these resources are allocated and maintained. The vCenter component of the vSphere software suite can manage single ESX or ESXi hosts and clusters of hosts.

VMware vSphere offers several features that allow for mobility of the VMs between ESXi hosts and storage. These features can add to the availability of the VMs that run in a cluster.

## 2.6.2 vSphere vMotion

vMotion is a technology that is designed to combat planned downtime. vMotion is used to move VMs between host and data stores to allow scheduled maintenance procedures to proceed without affecting VM availability or performance. vMotion is included in the Enterprise and Enterprise Plus versions of VMware vSphere.

### vSphere Host vMotion

Host vMotion eliminates the need to schedule application downtime for planned server maintenance. Host vMotion eliminates the need to schedule application downtime through the live migration of VMs across servers with no disruption to users or loss of service. This process is managed from a vCenter server, which maintains client or application access to a VM while it moves between physical servers.

In an IBM Storwize V7000 HyperSwap configuration, this feature is useful for moving VMs between two failure domains. You might need to move VMs to balance loads across failure domains or because a failure domain needs an outage for maintenance.

For more information, see the VMware vSphere vMotion web page:

<http://www.vmware.com>

### vSphere Storage vMotion

Storage vMotion eliminates the need to schedule application downtime because of planned storage maintenance or during storage migrations. Storage vMotion eliminates the need to schedule application downtime by enabling the live migration of VM disks with no disruption to users or loss of service. The vCenter server manages the copy of data from one datastore to another data store. With vStorage application programming interfaces (APIs) for Array Integration (VAAI), this process can be offloaded to the storage subsystem, which saves resources on both the vCenter host and the data network.

In a Storwize V7000 HyperSwap solution, this feature is useful for moving a VM's VM disk (VMDK) file between two storage subsystems. You might move this file to ensure that it is on the same failure domain as the VM or to remove a storage device that is becoming obsolete or is undergoing maintenance.

## 2.6.3 vSphere High Availability

vSphere High Availability (HA) provides cost-effective, automated restarts of applications within minutes of hardware or operating system failures. With the addition of the Fault Domain Manager, VMware HA is reliable in operation and easily scalable in its ability to protect VMs. VMware HA can provide increased uptime.

For more information about vSphere HA, see this web page:

<http://www.vmware.com>

## **2.6.4 VMware Distributed Resource Scheduler**

VMware Distributed Resource Scheduler (DRS) dynamically balances computing capacity across a collection of hardware resources that are aggregated into logical resource pools. VMware DRS continuously monitors the use across resource pools and intelligently allocates available resources among the VMs that are based on predefined rules that reflect business needs and changing priorities. When a VM experiences an increased load, VMware DRS automatically allocates more resources by redistributing VMs among the physical servers in the resource pool.

VMware DRS migrates and allocates resources by using a set of user-defined rules and policies. You can use these rules and policies to make critical or high-performing VMs.



# IBM System Storwize V7000 HyperSwap architecture

This chapter focuses on the IBM System Storwize V7000 architecture as it applies to the HyperSwap configuration. It is based on the assumption that you have a base understanding of the general Storwize V7000 architecture.

This chapter includes the following sections:

- ▶ Storwize V7000 HyperSwap overview
- ▶ Failure domains, sites, and controllers
- ▶ Storwize V7000 active-active Metro Mirror
- ▶ Storwize V7000 HyperSwap configurations
- ▶ Fibre Channel settings for distance

## 3.1 Storwize V7000 HyperSwap overview

The HyperSwap high availability (HA) function in the IBM System Storwize V7000 software allows business continuity in a hardware failure, power failure, connectivity failure, or disasters, such as fire or flooding. It is available on the IBM SAN Volume Controller, Storwize V7000, Storwize V7000 Unified only for the Block Data protocol, and Storwize V5000 products.

It provides highly available volumes that are accessible through two sites at up to 300 km (186.4 miles) 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 completes. The HyperSwap function automatically optimizes itself to minimize the data that is transmitted between sites and to minimize host read and write latency.

If the nodes or storage at either site go offline, leaving an online and accessible up-to-date copy, the HyperSwap function automatically fails over access to the online copy. The HyperSwap function also automatically resynchronizes the two copies when possible.

The HyperSwap function builds on two existing technologies in the product:

- ▶ Nondisruptive Volume Move (NDVM) function that was introduced in version 6.4 of the SAN Volume Controller software
- ▶ Remote Copy features that include Metro Mirror, Global Mirror, and Global Mirror with Change Volumes

In a standard Storwize V7000 implementation or Enhanced Stretched Cluster (ESC), each volume must be accessed by the hosts only through the caching I/O Group. The caching I/O Group is the I/O Group that owns the volume. In a Storwize V7000 HyperSwap implementation or SAN Volume Controller HyperSwap implementation, each volume must be accessed by the hosts that use all of the I/O Groups in the clustered Storwize V7000 or SAN Volume Controller cluster. This design requires at least eight paths for each volume from each host.

The HyperSwap function works with the standard multipathing drivers that are available on a wide variety of host types, with no additional required host support to access the highly available volume. Where multipathing drivers support Asymmetric Logical Unit Access (ALUA), the storage system informs the multipathing driver of the nodes that are closest to it, and the nodes to use to minimize I/O latency. You only need to tell the storage system the site to which a host is connected, and it will configure host pathing optimally.

The key benefit of a Storwize V7000 HyperSwap, compared to a disaster recovery (DR) solution that is based on Metro Mirror or Global Mirror, is that Storwize V7000 HyperSwap offers fast nondisruptive failover in small-scale outages. For example, if a single storage device is affected, Storwize V7000 fails over internally with minimal delay. If a fabric element fails, or a Storwize V7000 node canister fails, a host can fail over on the alternate node canister in the same I/O Group. Or, a host can switch the internal Metro Mirror direction dynamically to use the Storwize V7000 node canister that is in the second I/O Group.

One of the benefits of the Storwize V7000 HyperSwap configuration, and a key advantage over several alternatives, is that the failover uses the same multipathing driver that is used with conventional Storwize V7000 deployments. This flexibility offers a wide interoperability matrix, matching the hosts and controllers that are already deployed in the client data centers. Many of the multipathing drivers are also the default, standard multipathing drivers for the operating system (OS).



Another benefit is that Storwize V7000 HyperSwap always has an automatic quorum to act as a tiebreak. Therefore, no external management software or human intervention is ever required to perform a failover. Storwize V7000 HyperSwap uses the same quorum architecture that is used by ESC.

Remember these key points about the Storwize V7000 HyperSwap configuration:

- ▶ Use of HyperSwap is optional. Existing standard configurations are still supported. However, we encourage clients to use the new feature for its benefits.
- ▶ For the Storwize V7000 HyperSwap configuration, a clustered solution with more than one I/O Group is required and a topology attribute was introduced. The topology value for HyperSwap is **hyperswap**. Configure it by using the **chsystem** command, or view it by using the **lssystem** command.
- ▶ The topology value that is set to **hyperswap** enables the site for the new host site-awareness features and DR capability. We describe the host site-awareness feature in 3.2, “Failure domains, sites, and controllers” on page 25.
- ▶ You can convert an existing standard Storwize V7000 configuration to a HyperSwap configuration nondisruptively any time after you upgrade to version 7.5.

**Note:** The Storwize V7000 HyperSwap configuration is a two-site, active-active site only high-availability (HA) or business continuity (BC) solution. A three-site solution is not possible at the time of writing this book.

## 3.2 Failure domains, sites, and controllers

In a Storwize V7000 configuration, the term *failure domain* is used to identify components of the Storwize V7000 that are contained within a boundary so that any failure that occurs (such as a power failure, fire, or flood) is contained within that boundary. Failure domain is also referred to as the *failure site*. The failure therefore cannot propagate or affect components that are outside of that boundary. The components that make up a Storwize V7000 HyperSwap configuration must span three independent failure domains. Two failure domains contain Storwize V7000 I/O Groups and if you are virtualizing external storage, the storage controllers that contain customer data. The third failure domain contains a storage controller where the active quorum disk is located.

Failure domains are typically areas or rooms in the data center, buildings on the same campus, or even buildings in different towns. Different kinds of failure domains protect against different types of failure conditions:

- ▶ If each failure domain is an area with a separate electrical power source within the same data center, the Storwize V7000 in a HyperSwap configuration can maintain availability if any single power source fails.
- ▶ If each site is a different building, the Storwize V7000 in a HyperSwap configuration can maintain availability if a loss of any single building occurs (for example, a power failure or fire).

Ideally, each of the three failure domains that are used for the HyperSwap configuration is in a separate building and powered by a separate power source. Although this configuration offers the highest level of protection against all possible failure and disaster conditions, it is not always possible. Compromise is often required.

If a third building is not available, place the failure domain that contains the active quorum disk in the same building as one of the other two failure domains. When this configuration is used, the following rules apply:

- ▶ Each failure domain must be powered by an independent power supply or uninterruptible power supply (UPS).
- ▶ The storage controller that is used for the active quorum disk must be separate from the storage controller that is used for the customer data.
- ▶ Each failure domain must be on independent and isolated storage area networks (SANs) (separate cabling and switches).
- ▶ All cabling (power, SAN, and Internet Protocol (IP)) from one failure domain must not be physically routed through another failure domain.
- ▶ Each failure domain must be placed in a separate fire compartment.
- ▶ Storwize V7000 HyperSwap solution can be deployed in different ways:
  - One I/O Group for each site
  - Two I/O Groups for each site
  - Two I/O Groups in one site and one I/O Group in the other site
  - Three I/O Groups in one site and one I/O Group in the other site

The adopted solution can vary, depending on the client, workload, and availability requirements.

**Remember:** The key prerequisite for failure domains is that each I/O Group must be placed in a separate failure domain.

Version 7.5 introduced a *host site-awareness* concept for Storwize V7000 node canisters and external storage controllers. The following characteristics apply:

- ▶ Site awareness can be used only when the topology value is set to **hyperswap**.
- ▶ If the topology value is set to **hyperswap**, the DR feature is enabled.
- ▶ The site object is now added, and the valid sites are 1, 2, and 3. You can set a name for each site, if you prefer.
- ▶ The default names for the sites are Site1, Site2, and Site3. The I/O Group of the Storwize V7000 is in sites 1 and 2. Site3 is the optional third site for a quorum tiebreaker disk.
- ▶ A Site field is added to node canisters and controllers. You can set it by using these Storwize V7000 CLI commands: **addnode**, **chnode**, and **chcontroller**, or using specific GUI Wizard as shown in Chapter 4, “Implementation” on page 55. The nodes and controller must have sites set in advance before you set the topology to **hyperswap**, and must have a site assigned. When using the GUI Wizard, all the required commands will be automatically executed by the GUI.
- ▶ A site field is added to the host. You can set the site value only by using these Storwize V7000 commands: **mkhost** and **chhost**, or by using the GUI. You must set the sites for the host in advance before you set the topology value to **hyperswap**, and the host must have an assigned site.
- ▶ You can view the site fields by using GUI or the following commands:
  - **l snode**
  - **l shost**
  - **l scontroller**
  - **l smdisk**
  - **l smdiskgrp**

- ▶ The nodes, canister, and hosts can be assigned only to sites 1 or 2. Hosts cannot be assigned to site 3.
- ▶ Optionally, you can specify the site for each controller. The default for a controller is for its site to be undefined. Controllers can be assigned to sites 1, 2, or 3, or they can be set to **undefined** again.
- ▶ The clustered Storwize V7000 that is used for the HyperSwap configuration and with its I/O Group spread among site1 and site2 must be configured with *layer replication*. You can set layer replication only by using the Storwize V7000 CLI **chsystem** command. You can see the layer replication by using the **lssystem** command. (No GUI is allowed in this release.)
- ▶ A managed disk (MDisk) derives its site value from the following component:
  - The control enclosure, which is also referred to as an I/O Group, that hosts the disk drives that make up each array or MDisk. In a multiple expansion enclosure, the MDisk derives its site from the control enclosure to which the expansion enclosures are connected.
  - The controller to which the MDisk is associated at that time. Certain back-end storage devices are presented to Storwize V7000 as multiple controller objects, and an MDisk might be associated with any of them from time to time. Ensure that all of these controller objects are specified with the same site so that any MDisks that are associated with that controller are associated with a well-defined single site.
- ▶ The site for a controller can be changed when the DR feature is disabled. It can also be changed if the controller has no MDisks or image mode MDisks. The site for a controller cannot be changed when the DR feature is enabled if the controller uses MDisks or image mode MDisks.
- ▶ The site property for a controller adjusts the I/O routing and error reporting for connectivity between the nodes and associated MDisks. These changes are effective for any MDisk controller with a defined site, even if the DR feature is disabled.
- ▶ The site property for a host adjusts the I/O routing and error reporting for connectivity between hosts and the nodes in the same site. These changes are effective only at SAN login time. Therefore, any changes potentially will require a host reboot or Fibre Channel (FC) host bus adapter (HBA) rescan, depending on the OS that is used.

### 3.3 Storwize V7000 active-active Metro Mirror

With the Storwize V7000 HyperSwap solution, a new type of volume is introduced that is called the *HyperSwap Volume*. These HyperSwap Volumes consist of a Master Volume and a Master Change Volume (CV) in one system site, and an Auxiliary Volume and an Auxiliary CV in the other system site. An active-active Metro Mirror relationship exists between the two sites. As with a regular Metro Mirror relationship, the active-active relationship attempts to keep the Master Volume and Auxiliary Volume synchronized. The relationship uses the CVs as journaling volumes during any resynchronization process.

The HyperSwap Volume always uses the unique identifier (UID) of the Master Volume. The HyperSwap Volume is assigned to the host by mapping only the Master Volume even though access to the Auxiliary Volume is guaranteed by the HyperSwap function.

Figure 3-1 shows how the HyperSwap Volume works.

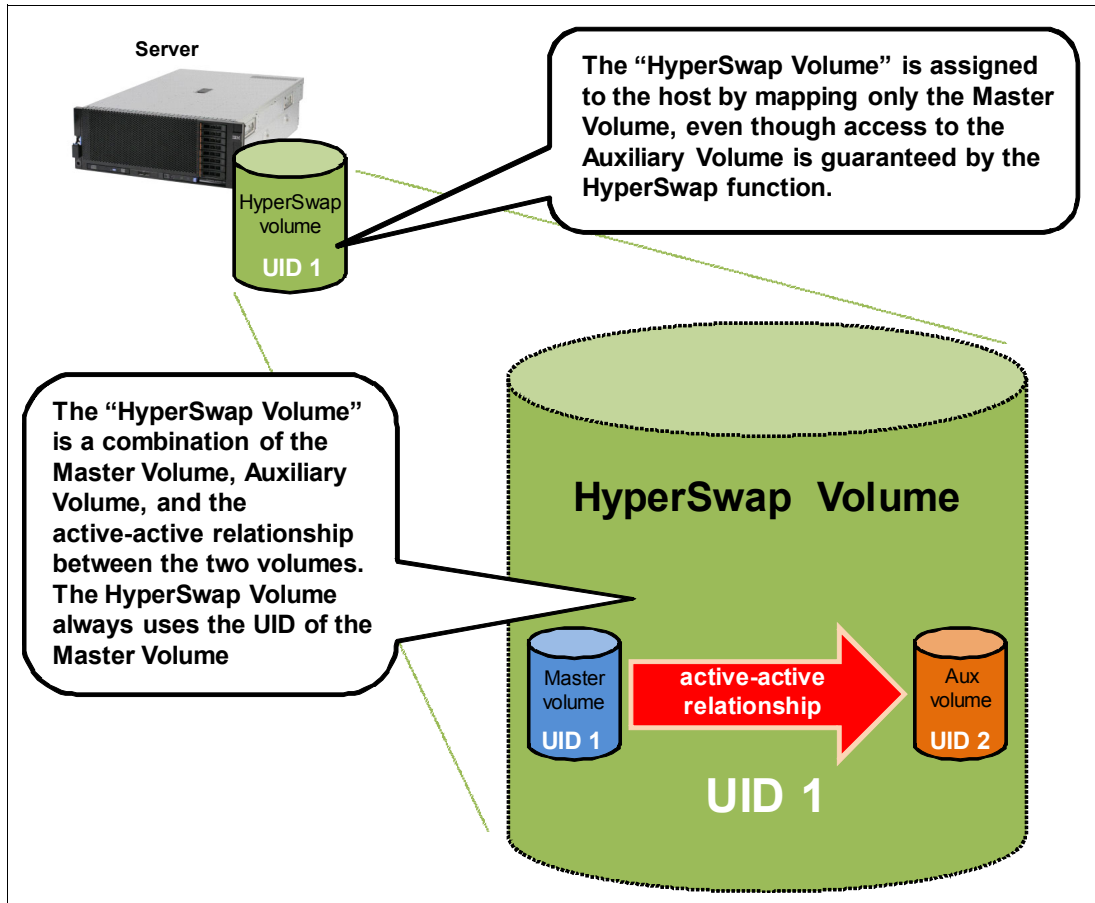


Figure 3-1 HyperSwap Volume

The active-active Metro Mirror replication workload will traverse the SAN by using the node-to-node communication on the Private SAN.

Master and Auxiliary Volumes also have a specific role of Primary or Secondary. Master or Auxiliary Volumes will be Primary or Secondary based on the Metro Mirror active-active relationship direction.

The role and the direction of each Metro Mirror relationship can be checked with the following Storwize V7000 CLI command as shown and detailed in Example 3-1.

*Example 3-1 lsrcrelationship command and details*

```

IBM_Storwize:ITSO_V7K_HyperSwap:superuser>lsrcrelationship
id name      master_cluster_id master_cluster_name master_vdisk_id
master_vdisk_name      aux_cluster_id  aux_cluster_name  aux_vdisk_id
aux_vdisk_name      primary consistency_group_id consistency_group_name state
bg_copy_priority progress copy_type      cycling_mode freeze_time
42 Rel_ESX_A 0000010021E001E0 ITSO_V7K_HyperSwap 42
HyperSwap_Volume_ESX_A_M 0000010021E001E0 ITSO_V7K_HyperSwap 43
HyperSwap_Volume_ESX_A_A master 0 CG_ESX_AtoB
consistent_synchronized 50 activeactive
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>lsrcrelationship Rel_ESX_A
id 42
name Rel_ESX_A
master_cluster_id 0000010021E001E0
master_cluster_name ITSO_V7K_HyperSwap
master_vdisk_id 42
master_vdisk_name HyperSwap_Volume_ESX_A_M
aux_cluster_id 0000010021E001E0
aux_cluster_name ITSO_V7K_HyperSwap
aux_vdisk_id 43
aux_vdisk_name HyperSwap_Volume_ESX_A_A
primary master
consistency_group_id 0
consistency_group_name CG_ESX_AtoB
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type activeactive
cycling_mode
cycle_period_seconds 300
master_change_vdisk_id 44
master_change_vdisk_name HyperSwap_Volume_ESX_A_Mcv
aux_change_vdisk_id 45
aux_change_vdisk_name HyperSwap_Volume_ESX_A_Acv

```

Or with GUI as shown in Figure 3-2 on page 29.

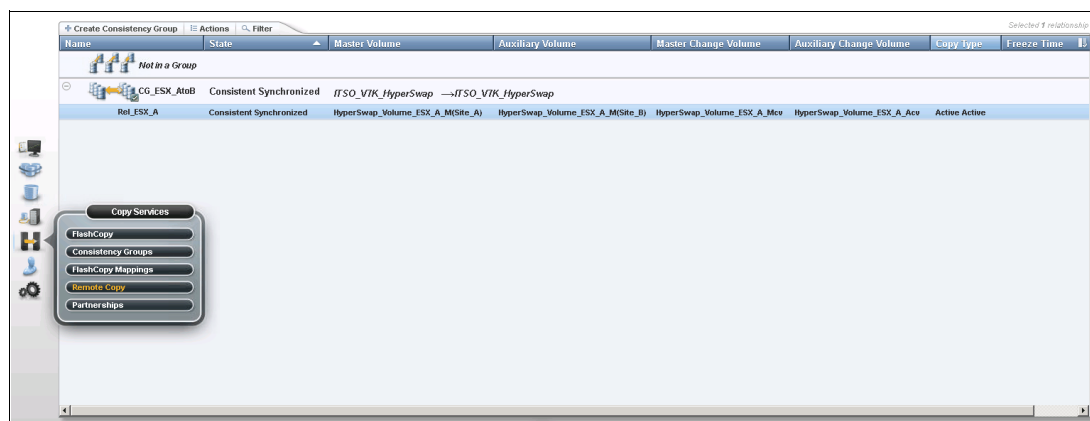


Figure 3-2 Metro Mirror Active-Active relationship

One site is considered the Primary for each HyperSwap Volume Group (active-active Consistency Group) or stand-alone HyperSwap Volume (active-active relationship). This site is dynamically chosen according to the site that receives more data (more than 75% of write I/Os) for the Volume Group or volume. This choice can change after a period of 20 minutes of sustained I/O that is submitted to nodes on the non-Primary site.

The Primary site processes all reads and writes, so reads to the non-Primary site have an increase in latency (of 1x the round-trip time (RTT) between the sites) while the data is retrieved over the Private SAN. This increase in latency causes an increase in Primary-to-non-Primary bandwidth use.

Writes to the non-Primary site likewise have an increase in latency of 1x the RTT between sites (for a total of 2x the RTT) over writes to the Primary site.

The Master and Auxiliary Volumes of an active-active relationship must be placed in different storage pools, and they must belong to different cache I/O Groups. The Master Volume must belong to an I/O Group in the Primary site, and the Auxiliary Volume must belong to an I/O Group in the Secondary site. For the associated CV, you must follow the same rules.

The Master CV must be in the same I/O Group as the Master Volume, and it is recommended that it is in the same pool as the Master Volume. A similar restriction applies to the Auxiliary CV and the Auxiliary Volume.

MDisks that are hosted in the Storwize V7000 control enclosure or expansion enclosure are assigned a site based on the I/O Group of the enclosure.

For external virtualized storage, the storage controller must have the same site attribute as the I/O Group that is in the same site.

With the Storwize V7000 HyperSwap function, you can group multiple Metro Mirror volumes relationships for HA. This capability is important where an application spans many volumes and requires data consistency across all of those volumes.

The use of Volume Groups to control the synchronization and failover across many volumes in an application guarantees that all volume copies on a site have data from the same point in time. Therefore, DR can use that site's volume copies. It also guarantees that at least one site has an up-to-date copy of every volume in the Volume Group. And it further guarantees that the other site, if it does not have an up-to-date copy of every volume, has a consistent copy of every volume for an out-of-date point-in-time.

The HyperSwap function automatically controls synchronization and resynchronization of volume copies. Just before resynchronizing data to a volume copy, that copy usually contains consistent but stale (out-of-date) data. The storage system automatically retains that stale consistent data during the resynchronization process by using CV technology.

If a problem occurs at the site with the online copy before the resynchronization completes, therefore taking that copy offline, you can manually enable read and write access to the stale, consistent, and older copy of data, allowing the use of this data for DR. You typically use this option if you know that the offline copy will remain offline for an extended period, and the stale, consistent but older data is useful enough to keep your business running.

Normally, with DR solutions that provide business continuity with stale data, after the problem is resolved and you are restoring access to the offline copy, you can choose. Either revert to that now-online copy, which before the disaster held the latest copy of data, or continue to work on the stale data that was used during the disaster. With either choice, the other copy is resynchronized to match the chosen copy.

### 3.3.1 Active-active Metro Mirror prerequisites

The three quorum disk candidates keep the status of the replicated volumes. An allocation of bitmap memory is required to enable Metro Mirror and FlashCopy for CVs. You can allocate memory by using the **chiogrp** commands:

- ▶ `chiogrp -feature remote -size memory_size io_group_name | io_group_id`
- ▶ `chiogrp -feature flash -size memory_size io_group_name | io_group_id`

The default value is set to 20 MB. You need to adjust this value to base it on the number of your replicated volumes and amount of space, if you will be notified that you are running out of space.

You can also check the allocated space with the following Storwize V7000 CLI command as shown in Example 3-2.

*Example 3-2 lsiogrp command example*

---

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>lsiogrp 0
id 0
name io_grp0_SITE_A
node_count 2
vdisk_count 4
host_count 3
flash_copy_total_memory 20.0MB
flash_copy_free_memory 19.6MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 19.8MB
mirroring_total_memory 20.0MB
mirroring_free_memory 20.0MB
raid_total_memory 40.0MB
raid_free_memory 39.5MB
maintenance no
compression_active no
accessible_vdisk_count 5
compression_supported yes
max_enclosures 21
encryption_supported yes
flash_copy_maximum_memory 2048.0MB
site_id 1
site_name ITSO_SITE_A
```

---

### 3.3.2 Storwize V7000 HyperSwap read operations

In the Storwize V7000 HyperSwap configuration, when the workload and the configuration are stable, the read operation goes straight to the local copy in accordance with the host, node canister, controller, or MDisk site-awareness attribute, as shown in Figure 3-3.

- Read on primary goes straight to local copy

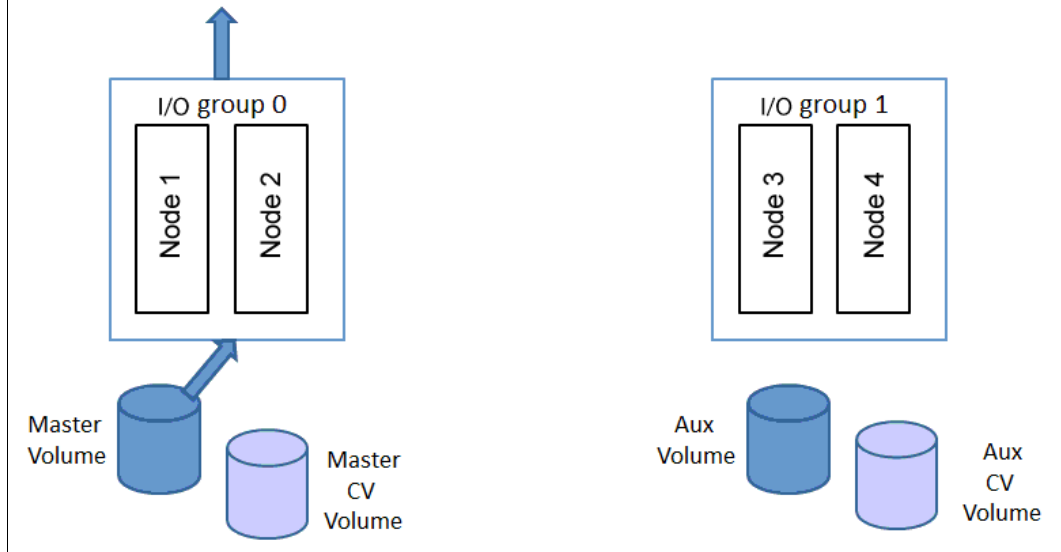


Figure 3-3 Read to local copy

Under certain circumstances, when an ESX Server virtual machine (VM) in site 1 is moved to the ESX Server in site 2 for example, the read operation goes straight to the remote I/O Group and the Secondary copy due to the new Host Site Awareness feature.

**Note:** At the time of writing this book, the read data is satisfied by the remote I/O Group, traversing the Private SAN, as shown in Figure 3-4.

- Read on secondary goes to remote I/O group, satisfied from remote storage

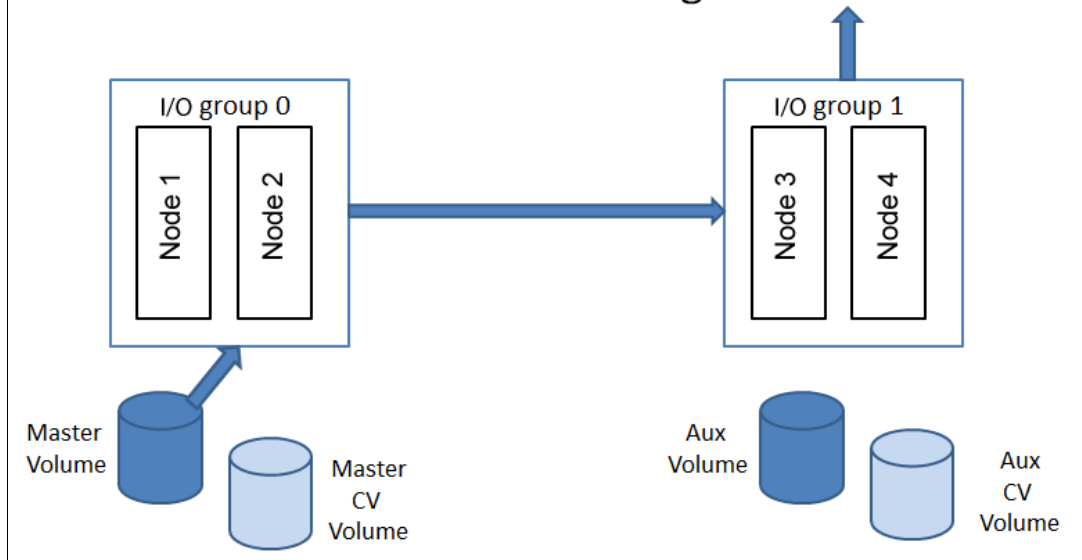


Figure 3-4 Read from Secondary



This scenario changes in a 10 - 20 minute time frame of sustained I/O to the non-Primary site. After that time, the active-active Metro Mirror will switch the direction and the volumes' roles dynamically to mitigate the latency impact.

### 3.3.3 Storwize V7000 HyperSwap write operations

Since version 7.5 and the introduction of Host Site Awareness, write operations are managed by the node with the same host site affinity. This new attribute makes the configuration easier and faster because a multipathing software round-robin is used to optimize all of the available paths that belong to the nodes with the same host site affinity. Therefore, scenarios and behavior can differ depending on whether you are writing on the Primary site (volume) or on the Secondary site (volume).

When hosts are writing on the Primary site, and the host, node, controller, or MDisk site awareness is satisfied, the write I/Os go straight to the Primary site volume and I/O Group, and they are replicated to the Secondary site volume and I/O Group as shown in Figure 3-5.

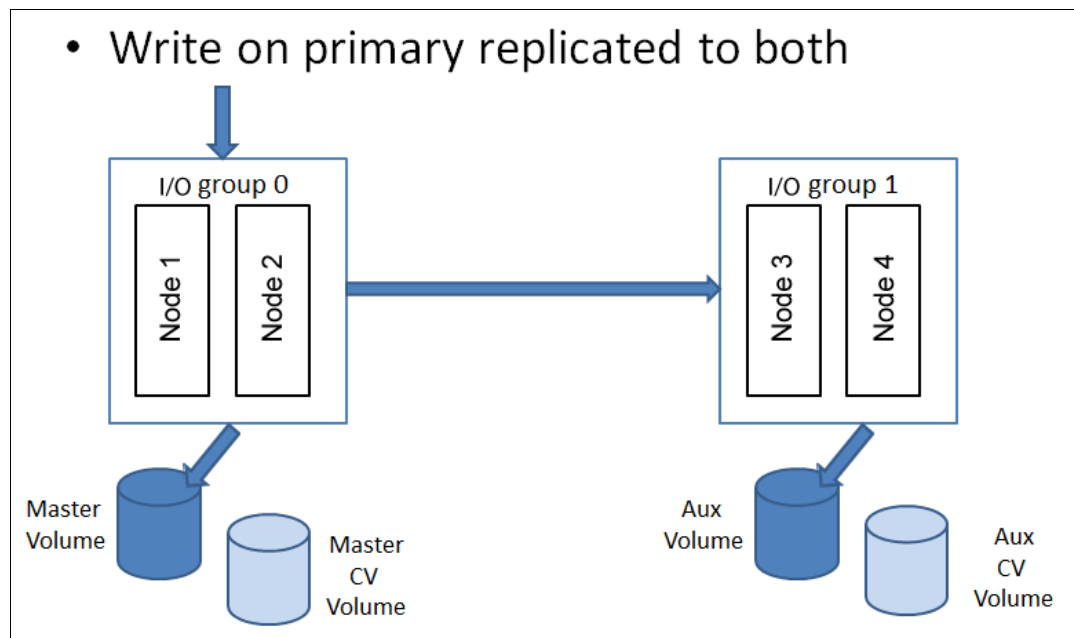


Figure 3-5 Write on local volume

Under certain circumstances, for example, when an ESX Server VM in site 1 is moved to the ESX Server in site 2, the write operation goes straight to the remote I/O Group and Secondary copy due to the new Host Site Awareness feature (Figure 3-6).

- Write on secondary forwarded to primary, then replicated to both

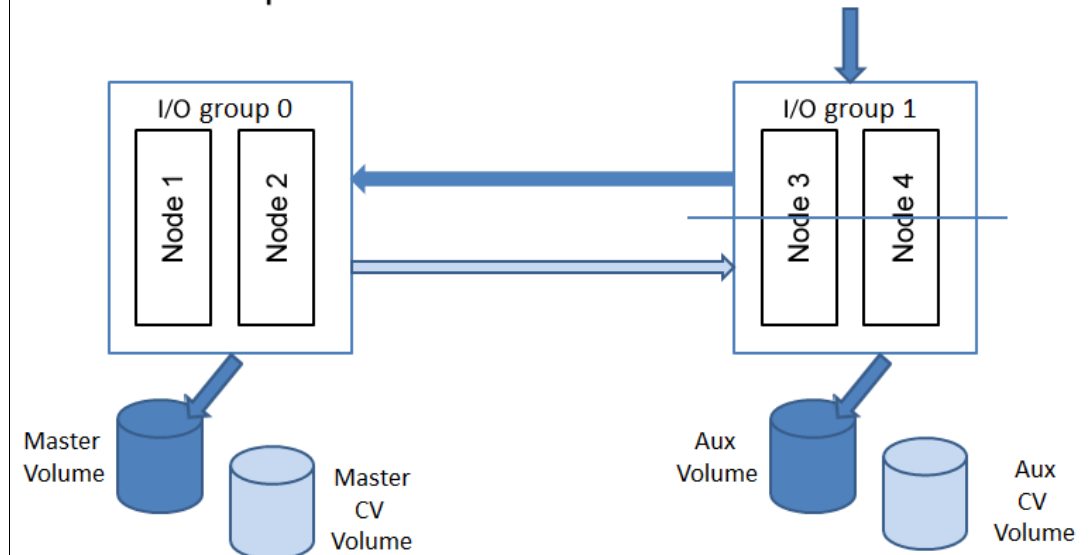


Figure 3-6 Write on Secondary

In this scenario, a write to I/O Group 1 needs to be applied to both copies, but the replication code cannot handle that task on I/O Group 0 (because I/O Group 0 currently holds the Primary copy). The write data is initially transferred from the host into a data buffer on a node in I/O Group 1. The node in I/O Group 1 sends the write, both metadata and customer data, to a node in I/O Group 0.

On the node in I/O Group 0, the write is largely handled as though it were written directly to that I/O Group by a host. The replication code applies the write to the I/O Group 0 cache, and replicates it to an I/O Group 1 node, to apply to the cache there.

This means that writes to the secondary site will have an increased latency, and will use additional bandwidth between the sites. However, sustained writes mainly to the secondary site will switch the direction of replication, removing this impact.

### 3.3.4 Storwize V7000 HyperSwap configuration quorum disk

The quorum disk fulfills two functions for cluster reliability:

- ▶ Acts as a tiebreaker in split-brain scenarios
- ▶ Saves critical configuration metadata

The Storwize V7000 quorum algorithm distinguishes between the active quorum disk and quorum disk candidates. Three quorum disk candidates exist. At any time, only one of these candidates acts as the active quorum disk. The other two are reserved to become active if the current active quorum disk fails. All three quorum disks store configuration metadata, but only the active quorum disk acts as the tiebreaker for split-brain scenarios.

The tie-break mechanism requires each location to provide one of the quorum disks by using either internal or external storage capacity. The third quorum disk must be provided by an external storage device that is installed in a third location (*quorum site*). For this reason, an additional external storage device is required to implement a Storwize V7000 HyperSwap configuration.

**Requirement:** A quorum disk must be placed in each of the three failure domains. Set the quorum disk in the third failure domain as the active quorum disk.

When the hyperswap topology and automatic quorum disk selection are enabled, three quorum disks total are created in sites 1, 2, and 3.

If a site has no suitable MDisks, fewer than three quorum disks are automatically created.

If you control the quorum by using the **chquorum** command, the choice of quorum disks must also follow the one-disk-for-each-site rule. If you used the **chquorum** command to manually assign quorum disks and configure the topology as hyperswap, the controller ignores any quorum disk that is not assigned to a site. Storwize V7000 chooses only quorum disks that are configured to site 3 as the active quorum disk and chooses only quorum disks that are configured to site 1 or site 2 as stand-by quorum disks.

If you are not virtualizing any external storage in site 1 and site 2, the quorum disk for those sites is chosen in the same way as in any regular Storwize V7000 implementation as Disk Drive. In Example 3-3, object ID 8 and object ID 15 are referred to as Disk Drive. In our example, they are *spare* drives, but the Disk Drive can be either *member*.

Example 3-3 *Isquorum and Isdrive commands*

```

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>Isquorum
quorum_index status id name controller_id controller_name active
object_type override site_id site_name
0 online 8 no drive
yes 1 ITS0_SITE_A
1 online 15 no drive
yes 2 ITS0_SITE_B
2 online 2 mdisk0_V7K_HS_Q 0 ITS0_V7K_Q_N1 yes mdisk
yes 3 ITS0_SITE_Q
3 online yes device no
ITS0-2.englab.brocade.com/10.18.228.171

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>Isdrive
id status error_sequence_number use tech_type capacity mdisk_id mdisk_name
member_id enclosure_id slot_id node_id node_name auto_manage
0 online 3 member sas_hdd 558.4GB 0 mdisk_1A 2
1 3 inactive
1 online 5 member sas_hdd 558.4GB 0 mdisk_1A 4
1 5 inactive
.
data removed for brevity
.
8 online spare sas_hdd 558.4GB
1 8 inactive
.
data removed for brevity
.
15 online spare sas_hdd 558.4GB
2 8 inactive

```

If a quorum disk is not configured at each site, you might be restricted when, or if, a Tier 3 (T3) recovery procedure is possible. You might be restricted about how resilient the cluster is

if a site failure occurs. Without access to a quorum disk, Storwize V7000 cannot continue I/O operations when one copy of a mirrored volume goes offline.

In Example 3-3 on page 35 it is possible to also see the IP Quorum object (only for version 7.6 and above).

Figure 3-7 on page 36 shows the Quorum disk and the IP Quorum configuration from the GUI.

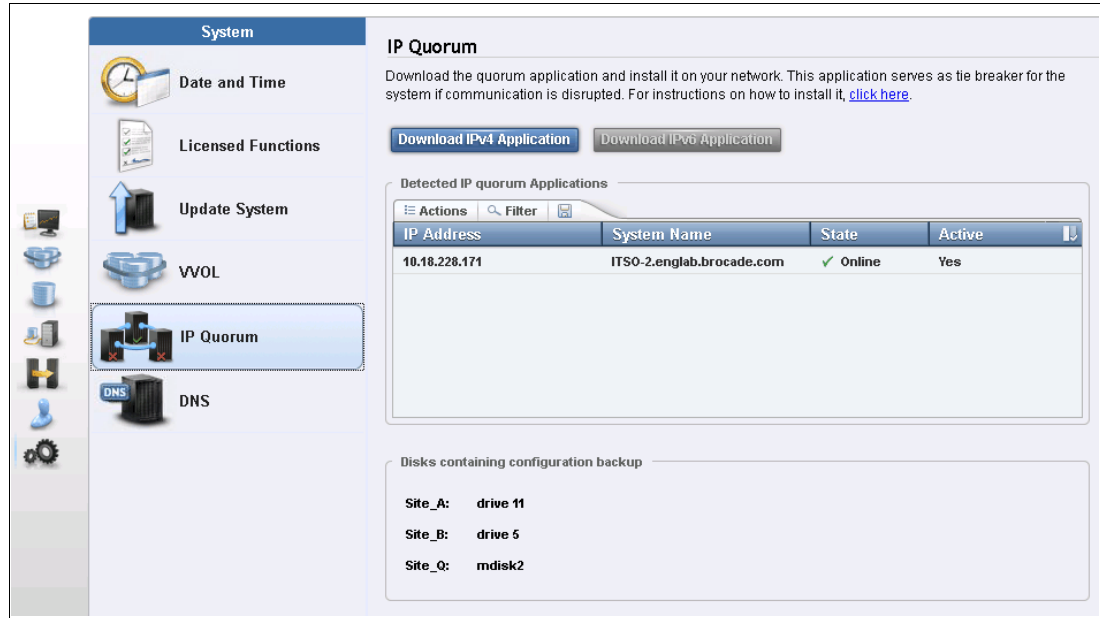


Figure 3-7 Quorum disk and IP Quorum example

### 3.3.5 View management

All nodes with the same cluster ID are configured to form a cluster. The subset of those nodes that operate as a cluster is determined by connectivity and the history of the cluster operation.

Asymmetric network faults (for example, where node 1 can see node 2 and node 2 can see node 3 but node 1 cannot see node 3) might result in situations where nodes are not mutually fully connected.

To keep the behavior of the cluster well-defined under such error scenarios, nodes are grouped into mutually fully connected sets that are called *views*. At most, one view continues to operate as the cluster at any time.

The first step in establishing a cluster operation is therefore to establish a set of views for all of the nodes.

In a Storwize V7000 HyperSwap configuration, where node 1 and node 3 make up I/O Group 0 in site1, and node 2 and node 4 make up I/O Group 1 in site2, four potential asymmetric views are possible across the Peripheral Component Interconnect (PCI) and FC I/O fabric as shown in Example 3-4.

*Example 3-4 View example*

---

- Node 1 - Node 3
- Node 2 - Node 4
- Node 2 - Node 3 - Node 4

The algorithm that is used to generate the views attempts to generate the greatest view possible out of all nodes, which is followed by the greatest view possible out of all remaining nodes, which is followed again by the greatest view possible out of all remaining nodes, and so on, until all nodes are in a view. Ordinarily, the view generation algorithm will stabilize quickly after a debounce period of about 1 second.

After the views are generated, the node in each view with the lowest node unique ID becomes the boss node for the view and proceeds to determine (from the quorum rules) whether the nodes in that view can operate as the cluster.

When multiple solutions exist to the view assignment, which are otherwise equally good, such as examples 3 and 4 in Example 3-4, views and asymmetric failures will be detected and tolerated, by selecting a subset of nodes with symmetric connectivity. Certain choices of the subset of nodes might be suboptimal. Those views will be reduced by 1 to a symmetric view, such as views 1 and 2.

In this scenario with multiple equal symmetric views, a quorum race will start and the winning node will determine the winning site and I/O Group. Other nodes will be shut down.

### 3.3.6 Storwize V7000 cluster state and voting

The cluster state information on the active quorum disk is used to decide which Storwize V7000 I/O Groups survive if exactly half of the I/O Groups in the cluster fail at the same time. Each node has one vote, and the quorum disk has one-half vote for determining cluster quorum.

The Storwize V7000 cluster manager implements a dynamic quorum. Following a loss of nodes, if the cluster can continue operation, it dynamically alters the voting set that defines the nodes that must be present to allow more node failures to be tolerated. In this way, the voting set is continually updated to match the set of nodes that are present. This process enables servicing of the cluster.

The cluster manager determines the dynamic quorum from the current voting set and a quorum disk, if available. If nodes are added to a cluster, they get added to the voting set. When nodes are removed, they are also removed from the voting set. Over time, the voting set, and the nodes in the cluster, can completely change. The process of updating the voting set for dynamic quorum is automatic and concurrent.

The cluster can migrate onto a separate set of nodes from the set where it started. Within a Storwize V7000 cluster, the quorum is defined in one of the following ways:

- ▶ Since version 7.2 and now in a Storwize V7000 7.5 HyperSwap, the system continues to maintain the voting set with a dynamic quorum as it did for previous versions. But to provide greater resiliency in planned or unplanned failures of nodes, the voting rules are changed.
- ▶ In particular, all of the voting set nodes of a site, plus the quorum disk, are enough to achieve a quorum, even if that voting set of nodes is less than half of the nodes in the system.

A human vote, by using the **overridequorum** command, is also enough to establish a quorum in this case.

To prevent unwanted behavior by the cluster, if no quorum disk exists, the voting rules require that more nodes are present than the largest site's voting set.

Consider these examples:

- ▶ If a two-I/O Group four-node system has one node down for service, one site has two nodes and the other site has one node.
- ▶ If the intersite link fails, either site can establish a quorum by using the quorum disk. Alternatively, you can use the **overridequorum** command to force a DR feature invocation, even when the site has only one node.
- ▶ As a further example, if an eight-node cluster has one node down for service, and a failure causes a loss of connectivity to the quorum disk and several nodes, five nodes are necessary to continue the cluster operation.

Figure 3-8 summarizes the behavior of the Storwize V7000 HyperSwap as a result of failures that affected the site or failure domains.

Failure Domain 1 Node 1	Failure Domain 1 Node 2	Failure Domain 2 Node 1	Failure Domain 2 Node 2	Failure Domain 3 Quorum disk or IP Quorum	Cluster Status
Operational	Operational	Operational	Operational	Operational	Operational, optimal
Failed	Operational	Operational	Operational	Operational	Operational, Write cache disabled in IO group 1
Operational	Failed	Operational	Operational	Operational	Operational, Write cache disabled in IO group 1
Operational	Operational	Failed	Operational	Operational	Operational, Write cache disabled in IO group 2
Operational	Operational	Operational	Failed	Operational	Operational, Write cache disabled in IO group 2
Failed	Operational	Operational	Failed	Operational	It depends by the sequence of the events. If the two failures happen at the same time, the cluster will go in split brain. If the failure happen at different time the cluster will be operational.
Operational	Operational	Operational	Operational	Failed	Operational, Active Quorum disk moved. If IP Quorum fails, lowest Node_id Node selected as Quorum
Operational, Link to Failure Domain 2 has failed, Split Brain	Operational, Link to Failure Domain 2 has failed, Split Brain	Operational, Link to Failure Domain 1 has failed, Split Brain	Operational, Link to Failure Domain 1 has failed, Split Brain	Operational	The I/O group that access the active quorum disk or IP Quorum first remains active and the partner goes offline. If this is the beginning of a rolling disaster and then the I/O group who won the Quorum race goes offline too, then the surviving site can be restored with overridequorum
Operational, Link to Failure Domain 2 has failed, Split Brain	Operational	Operational, Link to Failure Domain 1 has failed, Split Brain	Operational	Operational	This is an Asymmetric failure and the Cluster status will be determined following the Cluster View as explained in CH 3.3.5.
Operational	Operational	Failed	Failed	Failed	Stopped, then the surviving site can be restored with overridequorum command.
Failed	Failed	Operational	Operational	Failed	Stopped, then the surviving site can be restored with overridequorum command.

Figure 3-8 Storwize V7000 HyperSwap behavior

### 3.3.7 Quorum disk requirements

The storage controller that provides the quorum disk in a Storwize V7000 HyperSwap configuration in the third site must be supported as an *extended quorum disk*. Storage controllers that provide extended quorum support, at time of writing, are listed on the Storwize V7000 Support Portal web page:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009559>

**Requirement:** Quorum disk storage controllers must be FC-attached or Fibre Channel over IP (FCIP)-attached. They must be able to provide less than 80 ms response times and have a guaranteed bandwidth of greater than 2 MBps.

**Important:** The quorum disk candidate requirements for the Storwize V7000 HyperSwap configuration are listed:

- ▶ The Storwize V7000 HyperSwap configuration requires three quorum disk candidates. One quorum disk candidate must be placed in each of the three failure domains.
- ▶ The active quorum disk must be assigned to a failure domain or to site 3.
- ▶ Dynamic quorum selection must be disabled by using the **chquorum** command.
- ▶ Quorum disk candidates and the active quorum disk assignment must be performed manually by using the **chquorum** command.

**Note:** At the time of writing this book, the quorum interoperability matrix for Storwize V7000 was available at this website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009559>

This matrix refers to the supported quorum storage controller as *extended quorum* disk support for SAN Volume Controller. This statement is also considered valid for the Storwize V7000 HyperSwap configuration.

### 3.3.8 IP Quorum

In an Enhanced Stretched Cluster configuration or HyperSwap configuration, you must use a third, independent site to house quorum devices. To use a quorum disk as the quorum device, this third site must use Fibre Channel connectivity together with an external storage system. Sometimes, Fibre Channel connectivity is not possible. In a local environment, no extra hardware or networking, such as Fibre Channel or SAS-attached storage, is required beyond what is normally always provisioned within a system.

Starting with Spectrum Virtualize version 7.6 it is possible to use an IP-based quorum application as the quorum device for the third site, no Fibre Channel connectivity is used. Java applications are run on hosts at the third site. However, there are strict requirements on the IP network.

- ▶ Up to five IP quorum can be deployed, and in an Enhanced Stretched Cluster it is suggested to configure at least two IP quorum App, and one of those has to be at a third independent site.
- ▶ All IP quorum applications must be reconfigured and redeployed to hosts when certain aspects of the system configuration change. These aspects include adding or removing a node from the system, or when node service IP addresses are changed.
- ▶ For stable quorum resolution, an IP network must provide the following requirements:
  - Connectivity from the hosts to the service IP addresses of all nodes. If IP quorum is configured incorrectly, the network must also deal with possible security implications of exposing the service IP addresses, because this connectivity can also be used to access the service GUI.
  - Port 1260 is used by IP quorum applications to communicate from the hosts to all nodes.
  - The maximum round-trip delay must not exceed 80 ms, which means 40 ms in each direction.
  - A minimum bandwidth of 2 MBps is ensured for node-to-quorum traffic.

- As a native OS or in a virtual machine (no need for dedicated server/VM).
- Red Hat Enterprise Linux 6.5/7; SUSE Linux Enterprise Server 11m3/12; IBM Java 7.1/8
  - Use the IBM SCORE process for others.
- App must be able to create files (.LCK, .LOG) in its working directory.

Even with IP quorum applications at the third site, quorum disks at site one and site two are required, because they are used to store metadata.

To provide quorum resolution, use the `mkquorumapp` command or GUI to generate a Java application that is copied from the system and run on a host at a third site.

### 3.3.9 Failure scenarios in a HyperSwap configuration

Figure 3-9 on page 41 illustrates several failure scenarios in a Storwize V7000 HyperSwap configuration. The three failure scenarios are described:

- ▶ Power off FC Switch SAN768B-A1 in failure domain 1: As long as each I/O Group node canister is connected with a dedicated FC port to each FC Switch, SAN768B-A2 takes over the load and no side effects are expected.
- ▶ Power off Storwize V7000 node canister 1 in failure domain 1: The Storwize V7000 node canister 2 takes over the load and continues processing host I/O. The write cache for this I/O Group is disabled to avoid data loss in case Storwize V7000 node canister 2 also fails.
- ▶ Power off failure domain 1: I/O operations can continue from failure domain 2 and the active-active Metro Mirror direction is switched immediately.

Figure 3-9 illustrates several failure scenarios in a Storwize V7000 HyperSwap configuration. The blue lines represent local I/O traffic and the green lines represent I/O traffic between failure domains.

**This note applies only if you are virtualizing the external storage controller:** Power off this storage system. I/O operations can continue from failure domain 2 by using the storage system that is virtualized in failure domain 2 and if the active-active Metro Mirror direction is switched immediately.



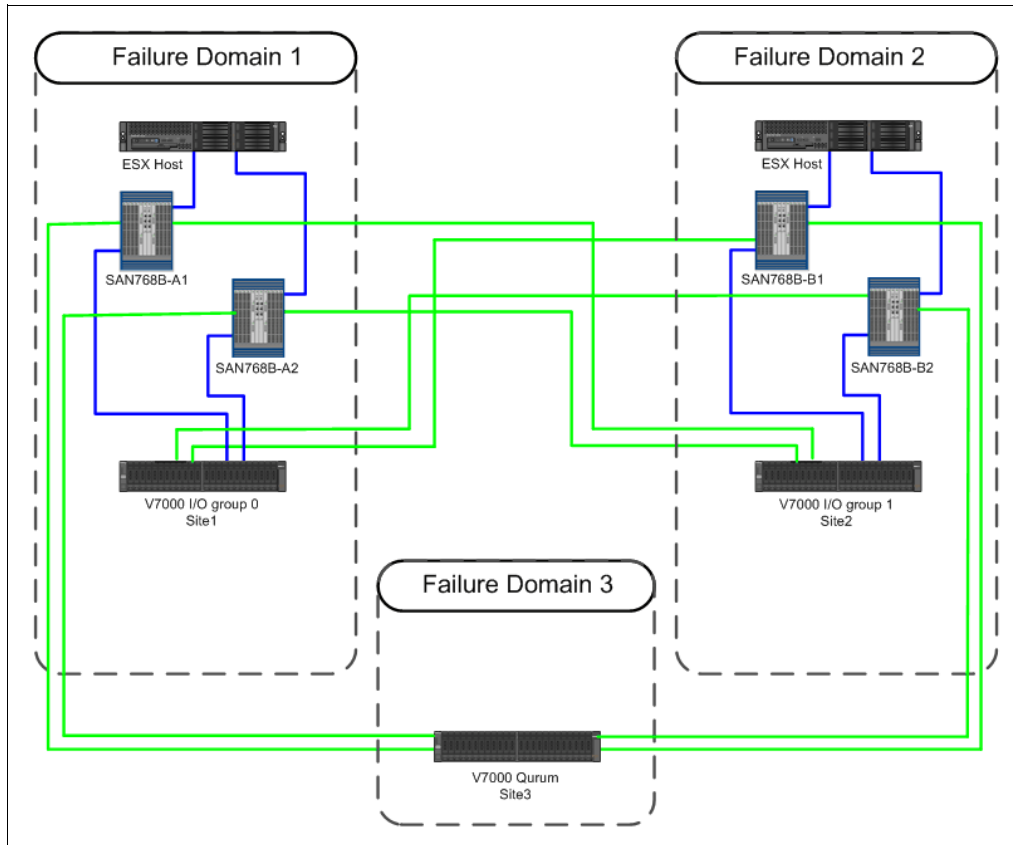


Figure 3-9 Storwize V7000 HyperSwap configuration

As Table 3-1 shows, Storwize V7000 HyperSwap can handle every kind of single failure automatically without affecting applications.

Table 3-1 Failure scenarios

Failure scenario	Storwize V7000 HyperSwap cluster behavior	Server and application impact
Single switch failure.	The system continues to operate by using an alternate path in the same failure domain to the same node.	None.
Single data storage failure.	The system continues to operate by using the secondary data copy and the replication direction is switched.	None.
Single quorum storage failure or IP Quorum	The system continues to operate on the same data copy.	None.
Failure of either failure domain 1 or 2, which contains the Storwize V7000 I/O Group.	The system continues to operate on the remaining failure domain that contains the Storwize V7000 I/O Group.	The servers without high availability (HA) functions in the failed site stop. The servers in the other site continue to operate. The servers with HA software functions are restarted from the HA software. The same disks are seen with the same UIDs in the surviving failure domain. Storwize V7000 cache in the surviving site will still be available.

Failure scenario	Storwize V7000 HyperSwap cluster behavior	Server and application impact
Failure of failure domain 3, which contains the active quorum disk.	The system continues to operate on both failure domains 1 and 2. Storwize V7000 selects another active quorum disk.	None.
Access loss between failure domains 1 and 2, which contain the Storwize V7000 I/O Groups.	The system continues to operate the failure domain with the Storwize V7000 I/O Group, which wins the quorum race. The cluster continues with operation, while the I/O Group node canister in the other failure domain stops.	The servers without HA functions in the failed site stop. The servers in the other site continue to operate. The servers with HA software functions are restarted from the HA software. The same disks are seen with the same UIDs in the surviving failure domain. Storwize V7000 cache in the surviving site will still be available.
Access loss between failure domains 1 and 2, which contain the Storwize V7000 I/O Groups because of a rolling disaster. One site is down, and the other site is still working. Later, the working site also goes down because of the rolling disaster.	The system continues to operate the failure domain with the Storwize V7000 I/O Group, which wins the quorum race. The cluster continues with operation, while the I/O Group node canister in the other failure domain stops. Later, the “winning” Storwize V7000 I/O Group is down too because of the rolling disaster. All Storwize V7000 I/O Groups are down.	The system can restart in the frozen surviving site by using the Storwize V7000 DR feature, which is manually triggered by the <code>overridequorum</code> command. The servers with HA software functions are restarted from the HA software. The same disks are seen with the same UIDs in the surviving failure domain. Storwize V7000 cache in the surviving site will still be available. Several recovery actions must occur to restore the failed site.

### 3.4 Storwize V7000 HyperSwap configurations

The Storwize V7000 node canisters of an HyperSwap configuration must connect to each other by FC or FCIP links. These links provide paths for node-to-node communication and for host access to controller nodes. HyperSwap supports three approaches for node-to-node intracluster communication between failure domains:

- ▶ Attach each Storwize V7000 node canister to the FC switches directly in the local and the remote failure domains. Therefore, all node-to-node traffic can occur without traversing inter-switch links (ISLs). This approach is referred to as *HyperSwap No ISL configuration*.
- ▶ Attach each Storwize V7000 node canister only to local FC switches and configure ISLs between failure domains for node-to-node traffic. This approach is referred to as *HyperSwap ISL configuration*.
- ▶ Attach each Storwize V7000 node canister only to local FC switches and configure FCIP between failure domains for node-to-node traffic. This approach is referred to as *HyperSwap FCIP configuration*.

Each of these HyperSwap configurations, along with their associated attributes, is described in the following sections to assist with the selection of the appropriate configuration to meet your requirements:

- ▶ No ISL configuration
- ▶ ISL configuration
- ▶ FCIP configuration

The maximum distance between failure domains without ISLs is limited to 40 km (24.8 miles). This limitation is to ensure that any burst in I/O traffic that can occur does not use all of the buffer-to-buffer (BB) credits. The link speed is also limited by the cable length between nodes.

Table 3-2 lists the supported distances for each of the Storwize V7000 HyperSwap configurations with their associated versions and port speed requirements.

*Table 3-2 Supported distances*

Configuration	Storwize V7000 version	Maximum length	Maximum link speed
No ISL	5.1 or later	< 10 km (6.2 miles)	8 Gbps
No ISL	6.3 or later	< 20 km (12.4 miles)	4 Gbps
No ISL	6.3 or later	< 40 km (24.8 miles)	2 Gbps
ISL	6.3 or later	< 300 km (186.4 miles)	2, 4, 8, or 16 Gbps
FCIP	6.4 or later	< 300 km (186.4 miles)	2, 4, 8, or 16 Gbps

### 3.4.1 No ISL configuration

This configuration is similar to a standard Storwize V7000 environment. The main difference is that I/O Groups are distributed across two failure domains. Figure 3-10 on page 44 illustrates the No ISL configuration. Failure domain 1 and failure domain 2 contain the Storwize V7000 node canister and expansion, with customer data. Failure domain 3 contains the storage subsystem that provides the active quorum disk.

#### Advantages

The No ISL configuration offers these advantages:

- ▶ The HA solution is distributed across two independent data centers.
- ▶ The configuration is similar to a standard Storwize V7000.
- ▶ Hardware effort is limited. WDM devices can be used but they are not required.

#### Requirements

The No ISL configuration has these requirements:

- ▶ Four dedicated fiber links per I/O Group (each I/O Group has two nodes so eight links are required) between failure domains.
- ▶ ISLs are not used between Storwize V7000 node canisters.
- ▶ Passive WDM devices can be used between failure domains.
- ▶ Active or passive WDM can be used between failure domains.
- ▶ Long wave small form-factor pluggables (SFPs) are required to reach 10 km (6.2 miles) without WDM.
- ▶ The supported distance is up to 40 km (24.8 miles) with WDM.
- ▶ Two independent fiber links between site 1 and site 2 must be configured with WDM connections.
- ▶ A third failure domain is required for quorum disk placement.
- ▶ The quorum disk storage system must be attached through FC.

Figure 3-10 illustrates a high level overview of the Storwize V7000 HyperSwap No ISL configuration.

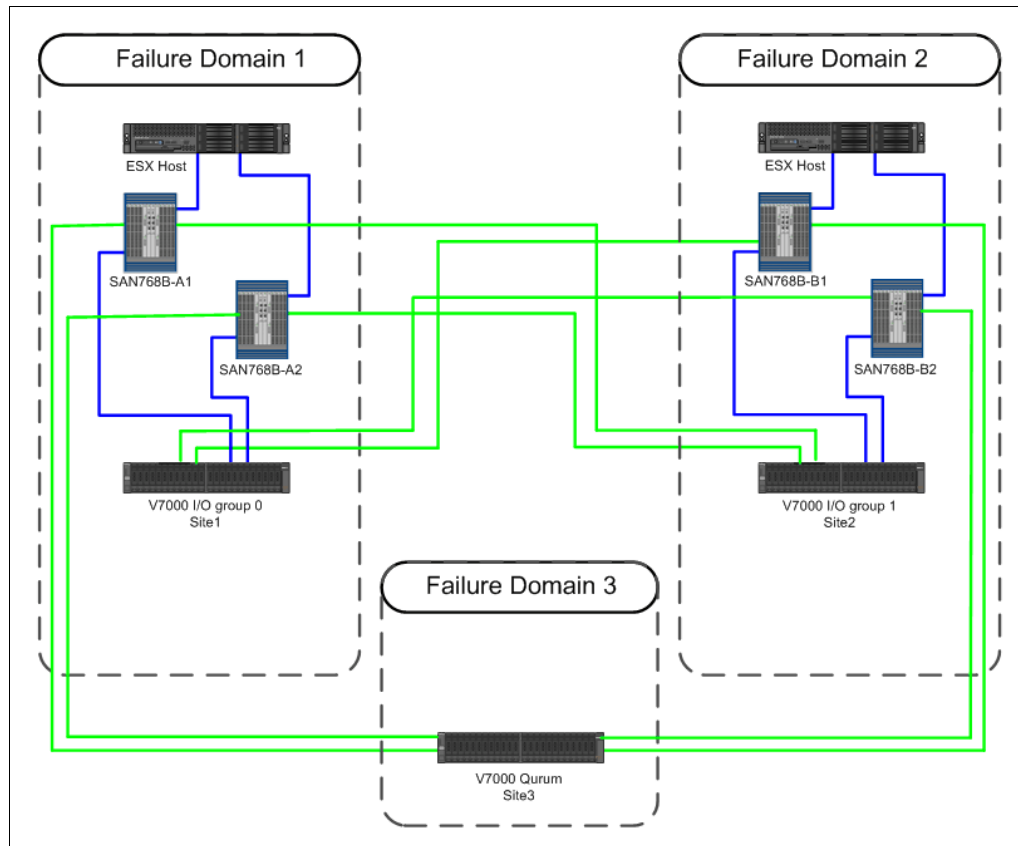


Figure 3-10 Storwize V7000 HyperSwap: No ISL configuration

### Zoning requirements

Zoning requirements for the Storwize V7000 HyperSwap No ISL configuration are the same as the zoning requirements for a standard configuration:

- ▶ Servers access only the Storwize V7000 node canister. No direct access exists from servers to back-end storage.
- ▶ A separate zone is configured for node-to-node traffic.
- ▶ Port masking can be used in addition to zoning to segregate node-to-node traffic on specific Storwize V7000 node canister FC ports. Port masking is enabled with the **chsystem -localfcportmask** command.
- ▶ The Storwize V7000 node canisters of the same I/O Group communicate across the controller enclosure internal PCI and FC I/O fabric.
- ▶ If you virtualize external storage, the zones must not contain multiple back-end disk systems. Each StorWize V7000 node canister must use dedicated FC ports for the external storage controller workload.

Figure 3-11 illustrates the Storwize V7000 HyperSwap no ISL configuration with passive WDM connections between failure domains 1 and 2.

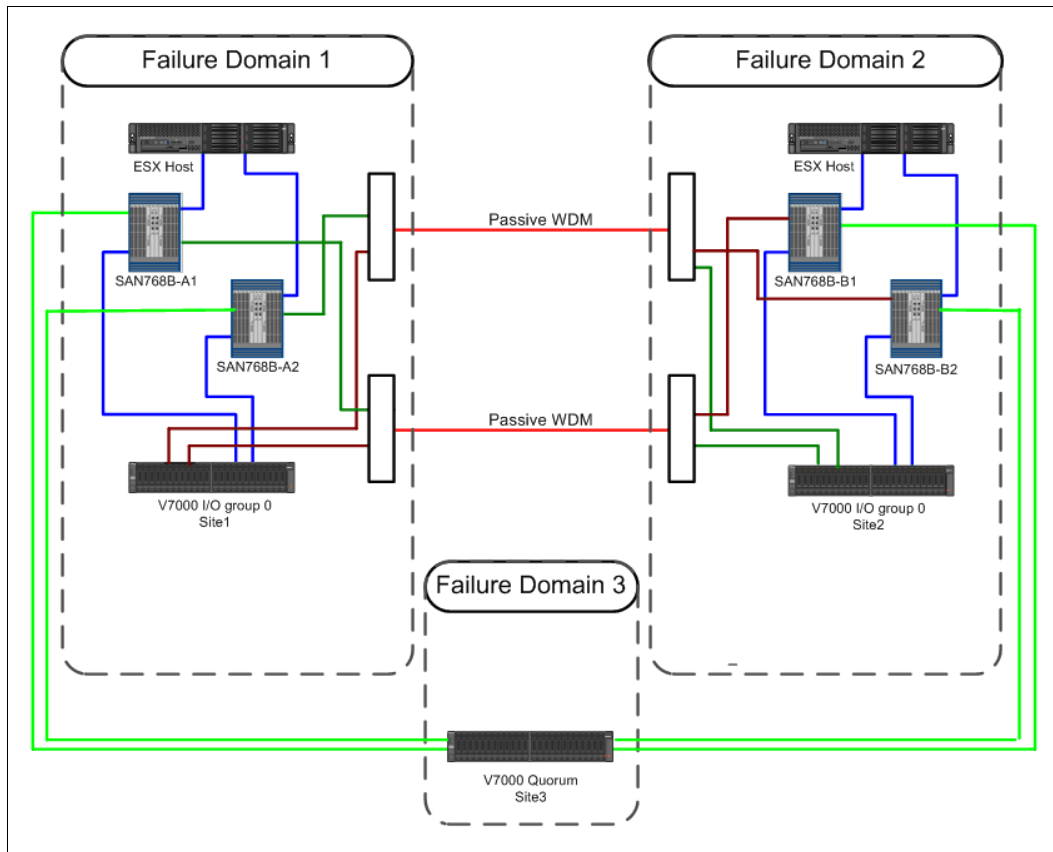


Figure 3-11 Storwize V7000 HyperSwap: No ISL configuration (with WDM)

### Preferred practices for Storwize V7000 Fibre Channel ports

Storwize V7000 Gen2 can have up to four 8 Gbps 4 Port Fibre Channel HBA feature for each node canister that is installed. Storwize V7000 Gen1 can have only one 8 Gbps 4 Port Fibre Channel HBA feature for each node canister that is installed.

The following preferred practices for Storwize V7000 Gen2 are listed:

- ▶ Dedicate two FC ports for node-to-remote site node communication. Attach these ports directly to the switch at the remote site.
- ▶ Use the other FC ports for host and storage attachment, possibly separating the storage workload from that host. Attach these ports to the switch at the local site.
- ▶ Access the third site quorum device by using the same ports that are dedicated for the storage workload.
- ▶ On Storwize V7000, configure the `localfcportmask` value to the port that is dedicated to the node-to-node connectivity. Use switch zoning to ensure that the node-to-remote site node communication uses the two FC ports that are dedicated to this purpose.

## 3.4.2 ISL configuration

This configuration is considered ideal. The Storwize V7000 I/O Groups are distributed across two failure domains, and node-to-node communication between failure domains is performed over ISLs.

The use of ISLs increases the supported distance for Storwize V7000 HyperSwap configurations to 300 km (186.4 miles). Although the maximum supported distance is 300 km (186.4 miles), instances occur where host-dependent I/Os must traverse the long-distance links multiple times. Therefore, the associated performance degradation might exceed acceptable levels. To mitigate this possibility, limit the distance between failure domains to 150 km (93.2 miles).

**Guideline:** Limiting the distance between failure domains to 150 km (93.2 miles) minimizes the risk of encountering elevated response times.

### Advantages

The ISL configuration offers these advantages:

- ▶ ISLs enable longer distances greater than 40 km (24.8 miles) between failure domains.
- ▶ Active and passive WDM devices can be used between failure domains.
- ▶ The supported distance is up to 300 km (186.4 miles) with WDM.

### Requirements

The ISL configuration has these requirements:

- ▶ Four required, dedicated FC ports, with two ports for each node canister for each I/O Group between failure domains.
- ▶ The use of ISLs for node-to-node communication requires configuring two separate SANs:
  - One SAN is dedicated for Storwize V7000 node-to-node communication. This SAN is referred to as the *Private* SAN.
  - One SAN is dedicated for host and storage controller attachment. This SAN is referred to as the *Public* SAN.
  - Each SAN must have at least one ISL for redundancy, and the bandwidth that is supplied by the ISL must be sized correctly.

- ▶ In Storwize V7000 HyperSwap, the minimum bandwidth for node-to-node communication between the sites is the peak write throughput from all hosts in both sites. This bandwidth is sufficient only if all volumes are accessed from hosts in one site. The HyperSwap cluster needs additional bandwidth between the sites in these scenarios:
  - a. A HyperSwap Volume is accessed concurrently by hosts in different sites.
  - b. The two copies of a HyperSwap Volume in different sites are being synchronized (initially or after an outage).
  - c. A host loses all paths to an I/O Group in its local site (and therefore accesses an I/O Group in the remote site). This scenario can happen because both nodes of the local I/O Group are offline, or because of SAN failures.

Scenario c requires additional inter-site bandwidth in the Public SAN. Scenarios a and b require additional inter-site bandwidth in the Private SAN. Scenarios b and c are the results of multiple failures or large outages. Scenario a can happen during normal operation.

To ensure the flawless operation of the HyperSwap cluster under all workload conditions, the bandwidth for node-to-node communication between the sites needs to be the sum of these numbers: for volumes that are accessed from hosts in the same site, the peak write throughput, and in addition for volumes that are accessed concurrently from hosts in different sites, the peak read and twice the peak write throughput.

- ▶ A third failure domain is required for quorum disk placement.
- ▶ Storage controllers that contain quorum disks must be attached through FC.
- ▶ A guaranteed minimum bandwidth of 2 MB is required for node-to-quorum traffic.
- ▶ No more than one ISL hop is supported for connectivity between failure domains.

**Tip:** Private and Public SANs can be implemented by using any of the following approaches:

- ▶ Dedicated FC switches for each SAN
- ▶ Switch partitioning features
- ▶ Virtual or logical fabrics

Figure 3-12 on page 48 illustrates the Storwize V7000 HyperSwap with ISL configuration between failure domains 1 and 2.

## Zoning requirements

The Storwize V7000 HyperSwap ISL configuration requires Private and Public SANs. The two SANs must be configured according to the following rules:

- ▶ Two ports of each Storwize V7000 node canister are attached to the Private SANs.
- ▶ Other ports (based with the Storwize V7000 configuration that was purchased) are attached to the Public SANs.
- ▶ A single trunk between switches is required for the Private SAN.
- ▶ Hosts and storage systems are attached to fabrics of the Public SANs.
- ▶ If you virtualize external storage, the zones must not contain multiple back-end disk systems. Each StorWize V7000 node canister must use dedicated FC ports for the external storage controller workload.
- ▶ Port masking must be used in addition to zoning to segregate node-to-node traffic on specific Storwize V7000 node canister FC ports. Port masking is enabled with the `chsystem -localfcportmask` command.

- ▶ Failure domain 3 (the quorum disk) must be attached to the Public SAN.
- ▶ ISLs that belong to the Private SANs must not be shared with other traffic, and they must not be oversubscribed.

For more information, see the following IBM Support web page:

<https://ibm.biz/BdsvBk>

Figure 3-12 illustrates the Storwize V7000 HyperSwap ISL configuration. The Private and Public SANs are represented as logical switches on each of the four physical switches.

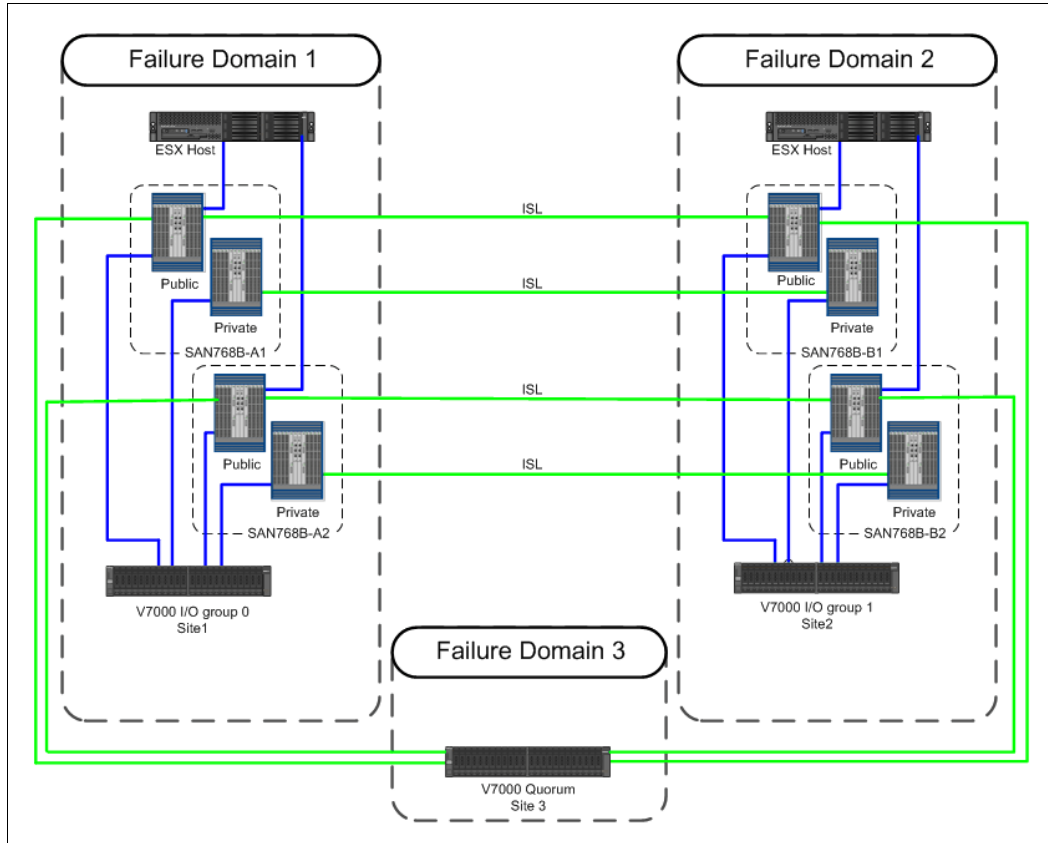


Figure 3-12 Storwize V7000 HyperSwap: ISL configuration



Figure 3-13 illustrates the Storwize V7000 HyperSwap ISL configuration with active or passive WDM between failure domains 1 and 2. The Private and Public SANs are represented as logical switches on each of the four physical switches.

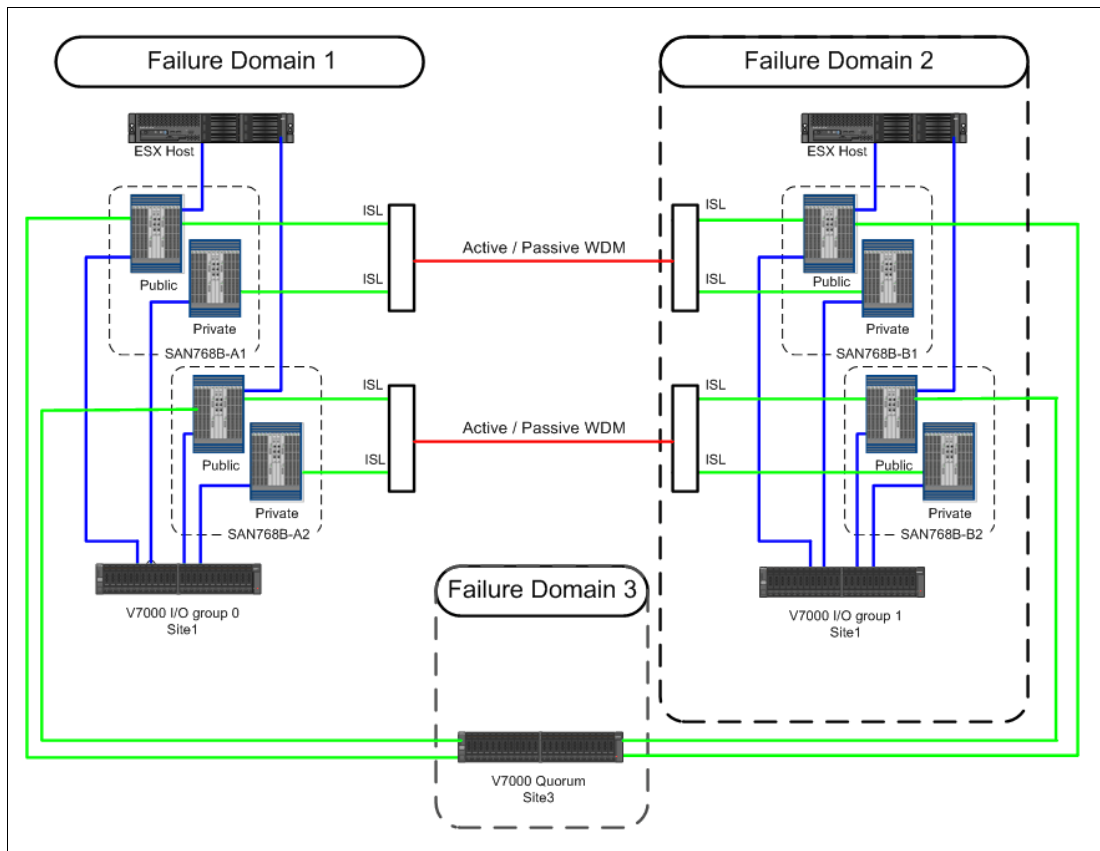


Figure 3-13 Storwize V7000 HyperSwap: ISL configuration (with WDM)

### Preferred practices for Storwize V7000 Fibre Channel ports

Storwize V7000 Gen2 can have up to 4 x 8 Gbps 4 Port Fibre Channel HBA feature for each node canister that is installed. Storwize V7000 Gen1 can have only 1 x 8 Gbps 4 Port Fibre Channel HBA feature for each node canister that is installed.

The preferred practices for Storwize V7000 Gen2 are listed:

- ▶ Dedicate two FC ports for node-to-remote site node communication. Attach these ports directly to the switch at each local site.
- ▶ Use the other FC ports for host and storage attachment, possibly separating the storage workload from that host. Attach these ports to the switch at the local site.
- ▶ Access the third site quorum device by using the same ports that are dedicated for the storage workload.
- ▶ On Storwize V7000, configure the `localfcportmask` value to the port that is dedicated to the node-to-node connectivity. Use switch zoning to ensure that the node-to-remote site node communication uses the two FC ports that are dedicated to this purpose.

### 3.4.3 FCIP configuration

In this configuration, FCIP links are used between failure domains. This configuration is a variation of the ISL configuration that was described, so many of the same requirements apply.

#### Advantage

The FCIP configuration uses existing IP networks for extended distance connectivity.

#### Requirements

The FCIP configuration has these requirements:

- ▶ It requires at least two FCIP tunnels between failure domains.
- ▶ It requires four dedicated FC ports, with two ports for each node canister for each I/O Group between failure domains.
- ▶ The use of ISLs for node-to-node communication requires the configuration of two separate SANs:
  - One SAN is dedicated for Storwize V7000 node-to-node communication. This SAN is referred to as the *Private* SAN.
  - One SAN is dedicated for host and storage controller attachment. This SAN is referred to as the *Public* SAN.
  - Each SAN must have at least one ISL for redundancy, and the bandwidth that is supplied by the ISL must be sized correctly.
- ▶ In Storwize V7000 HyperSwap, the minimum bandwidth for node-to-node communication between the sites is the peak write throughput from all hosts in both sites. This bandwidth is sufficient only if all volumes are accessed from hosts in one site. The HyperSwap cluster needs additional bandwidth between the sites in these scenarios:
  - a. A HyperSwap volume is accessed concurrently by hosts in different sites.
  - b. The two copies of a HyperSwap volume in different sites are being synchronized (initially or after an outage).
  - c. A host loses all paths to an I/O Group in its local site (and therefore accesses an I/O Group in the remote site). That scenario can happen because both nodes of the local I/O Group are offline, or because of SAN failures.

Scenario c on page 47 requires additional inter-site bandwidth in the Public SAN.

Scenarios a on page 47 and b on page 47 require additional inter-site bandwidth in the Private SAN. Scenarios b on page 47 and c on page 47 are the effects of multiple failures or large outages. Scenario a on page 47 can happen during normal operation.

To ensure the flawless operation of the HyperSwap cluster under all workload conditions, the bandwidth for node-to-node communication between the sites needs be the sum of these numbers: For volumes that are accessed from hosts in the same site, the peak write throughput, and in addition for volumes that are accessed concurrently from hosts in different sites, the peak read and twice the peak write throughput.

- ▶ A third failure domain is required for quorum disk placement.
- ▶ Storage controllers that contain quorum disks must be attached to the FC.
- ▶ A guaranteed minimum bandwidth of 2 MB is required for node-to-quorum traffic.
- ▶ Failure domain 3 (quorum disk) must be either FC or attached to FCIP. If it is attached to FCIP, the response time to the quorum disk cannot exceed 80 ms.

- ▶ Storage controllers that contain quorum disks must be either FC or attached to FCIP.
- ▶ No more than one ISL hop is supported for connectivity between failure domains.

**Tip:** Private and Public SANs can be implemented by using any of the following approaches:

- ▶ Dedicated FC switches for each SAN
- ▶ Switch partitioning features
- ▶ Virtual or logical fabrics

## Zoning requirements

The Storwize V7000 HyperSwap FCIP configuration requires Private and Public SANs. The two SANs must be configured according to the following rules:

- ▶ Two ports of each Storwize V7000 node canister are attached to the Private SANs.
- ▶ Other ports (based on the Storwize V7000 configuration that was purchased) are attached to the Public SANs.
- ▶ A single trunk between switches is required for the Private SAN.
- ▶ Hosts and storage systems are attached to fabrics of the Public SANs.
- ▶ If you virtualize external storage, the zones must not contain multiple back-end disk systems. Each StorWize V7000 node canister needs to use dedicated FC ports for the external storage controller workload.
- ▶ Port masking must be used in addition to zoning to segregate node-to-node traffic on specific Storwize V7000 node canister FC ports. Port masking is enabled with the `chsystem -localfcportmask` command.
- ▶ Failure domain 3 (the quorum disk) must be attached to the Public SAN.
- ▶ ISLs that belong to the Private SANs must not be shared with other traffic, and they must not be oversubscribed.

For more information, see the following IBM Support web page:

<https://ibm.biz/BdsvBk>

Figure 3-14 illustrates the Storwize V7000 HyperSwap FCIP configuration.

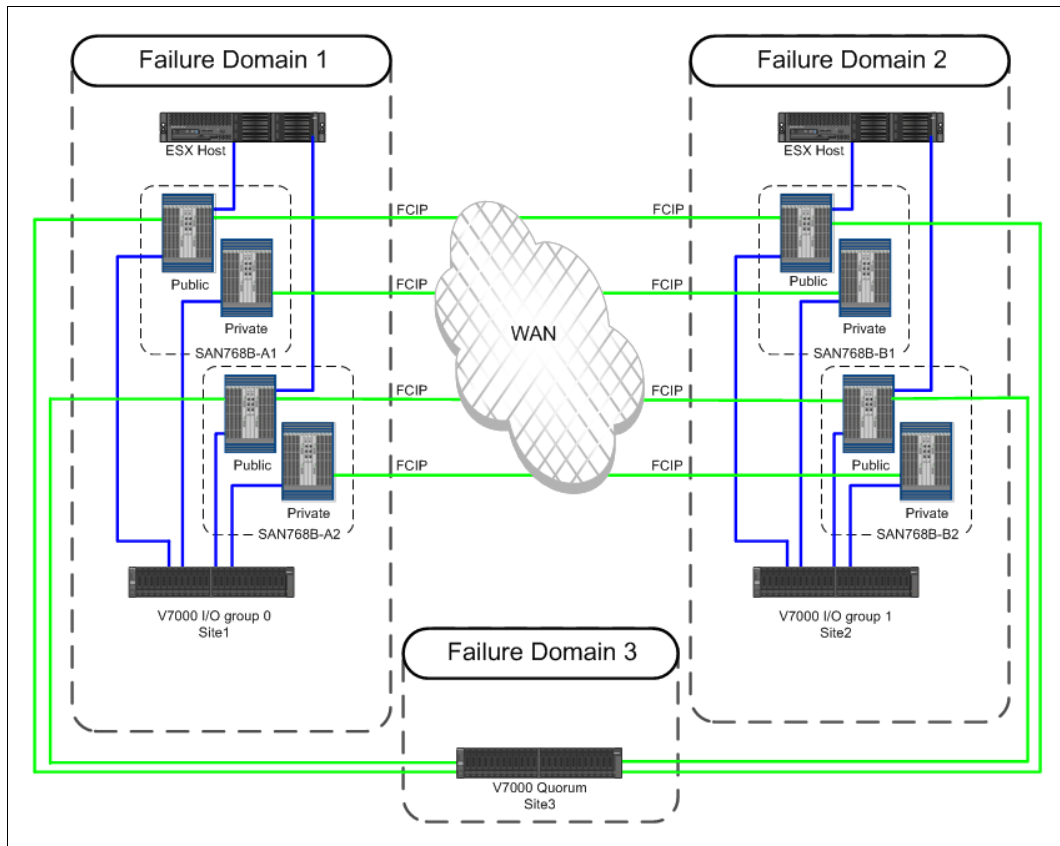


Figure 3-14 FCIP configuration

### Preferred practices for Storwize V7000 Fibre Channel ports

The Storwize V7000 Gen2 can have up to four 8 Gbps 4 Port Fibre Channel HBA feature for each node canister that is installed. The Storwize V7000 Gen1 can have only one 8 Gbps 4 Port Fibre Channel HBA feature for each node canister that is installed.

The preferred practices for Storwize V7000 Gen2 are listed:

- ▶ Dedicate two FC ports for node-to-remote site node communication. Attach these ports directly to the switch at each local site.
- ▶ Use the other FC ports for host and storage attachment, possibly separating the storage workload from that host. Attach these ports to the switch at the local site.
- ▶ The third site quorum device must be accessed by using the same ports that are dedicated for the storage workload.
- ▶ On Storwize V7000, configure the `localfcportmask` value to the port that is dedicated to the node-to-node connectivity. Use switch zoning to ensure that the node-to-remote site node communication uses the two FC ports that are dedicated to this purpose.

## 3.5 Fibre Channel settings for distance

Usage of long wave (LW) SFPs is an appropriate method to overcome long distances. Active and passive dense wavelength division multiplexing (DWDM) and coarse wavelength division multiplexing (CWDM) technology are supported.

Passive WDM devices are not capable of changing wavelengths by themselves. Colored SFPs are required and must be supported by the switch vendor.

Active WDM devices can change wavelengths by themselves. All active WDM components that are already supported by Spectrum Virtualize Metro Mirror are also supported by Storwize V7000 HyperSwap configurations.

Buffer credits, which are also called *buffer-to-buffer* (BB) credits, are used for FC flow control. They represent the number of frames that a port can store. Each time that a port transmits a frame, that port's BB credit is decremented by one. For each R\_RDY that is received, that port's BB credit is incremented by one. If the BB credit is zero, the corresponding port cannot transmit until an R\_RDY is received back.

Therefore, buffer-to-buffer credits are necessary to have multiple FC frames in flight (Figure 3-15). An appropriate number of buffer-to-buffer credits are required for optimal performance. The number of buffer credits to achieve maximum performance over a certain distance depends on the speed of the link.

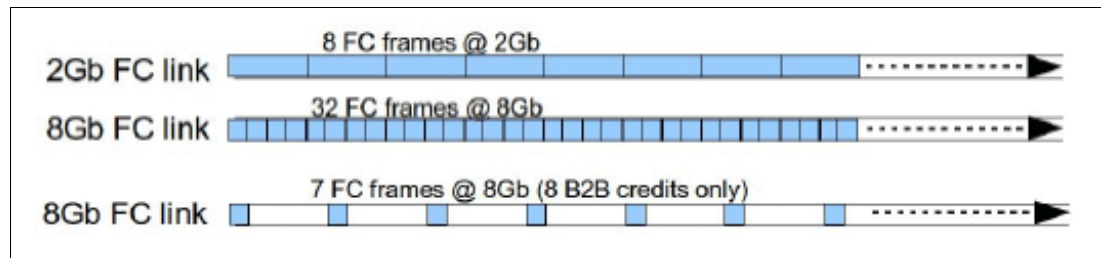


Figure 3-15 Buffer credits in flight

The calculation assumes that the other end of the link starts transmitting the R\_RDY acknowledgment frame in the same moment that the last bit of the incoming frame arrives at the receiving end. This scenario is not true. The following guidelines provide the minimum numbers. The performance drops dramatically if insufficient buffer credits exist for the link distance and link speed. Table 3-3 illustrates the relationship between BB credits and distance.

Table 3-3 Buffer-to-buffer credits

FC link speed	BB credits for 10 km (6.2 miles)
1 Gbps	5
2 Gbps	10
4 Gbps	20
8 Gbps	40
16 Gbps	80

The number of buffer-to-buffer credits that are provided by an Storwize V7000 FC HBA is limited. An HBA port provides 41 buffer credits, which are sufficient for a 10 km (6.2 miles) distance at 8 Gbps. These numbers are determined by the HBA's hardware, and they cannot be changed.

The Storwize V7000 Gen1 and Gen2 still supply 41 buffer credits for an 8 Gbps FC port and 80 buffer credits for a 16 Gbps FC port. At the time of writing this book, IBM has a statement of direction to release a 4 port 16 Gb FC adapter, and for each port, 41 buffer credits will be available.

FC switches have default settings for the BB credits. (In IBM b-type/Brocade switches, the default is 8 BB credits for each port.) Although the Storwize V7000 HBAs provide 41 BB credits, the switches stay at the default value. Therefore, you must adjust the switch BB credits manually. For Brocade switches, the port buffer credits can be changed by using the **portcfgportbuffers** command.



# Implementation

This chapter explains how this solution is implemented. It includes the following sections:

- ▶ Test environment
- ▶ IBM Fibre Channel SAN
- ▶ Storwize V7000 HyperSwap planning
- ▶ Storwize V7000 HyperSwap configuration

## 4.1 Test environment

The solution that is described in this book was tested in a lab by using dedicated resources. Figure 4-1 shows the environment that was implemented.

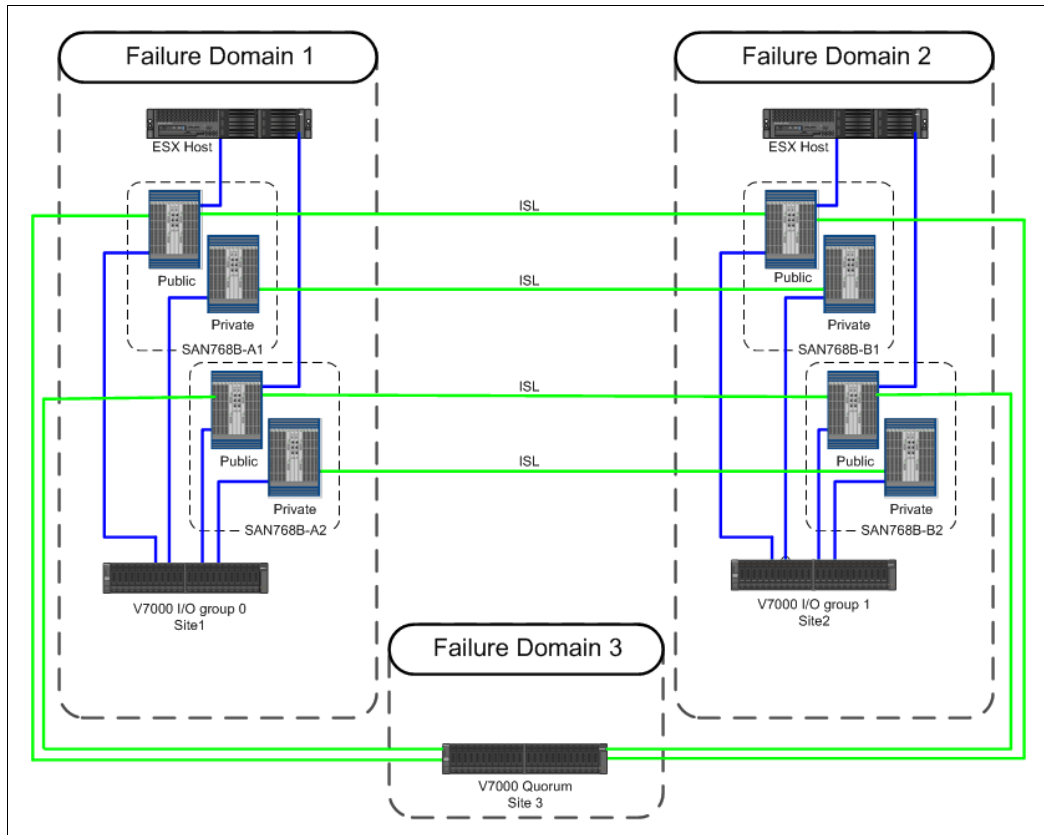


Figure 4-1 IBM SAN and Storwize V7000 HyperSwap cluster with VMware

This chapter describes the setup of the storage components that were used for the lab environment.



Figure 4-2 shows our laboratory storage area network (SAN) design from the storage and SAN components' point of view.

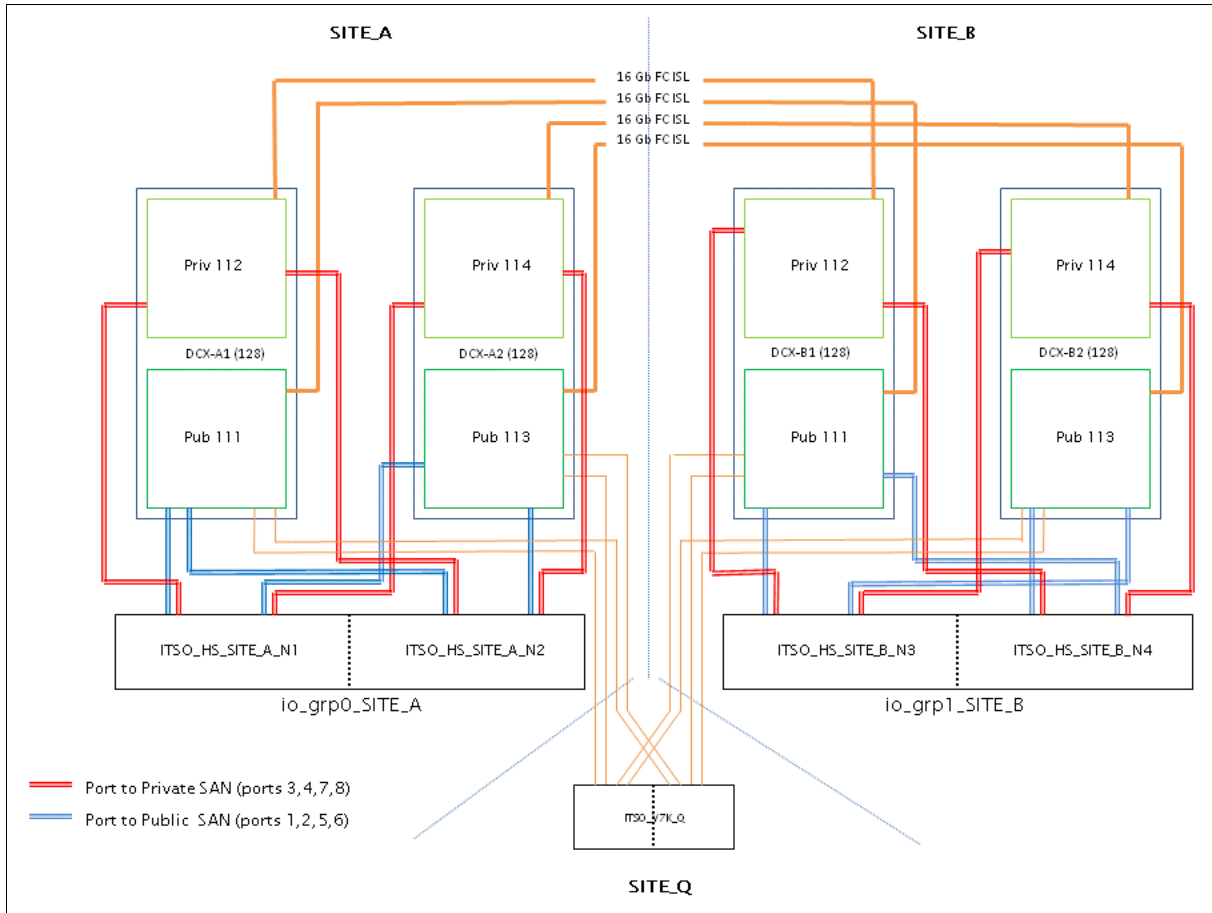


Figure 4-2 SAN design from storage and SAN components' point of view

Figure 4-3 shows our laboratory SAN design from the host, SAN component, and Storwize V7000 points' of view.

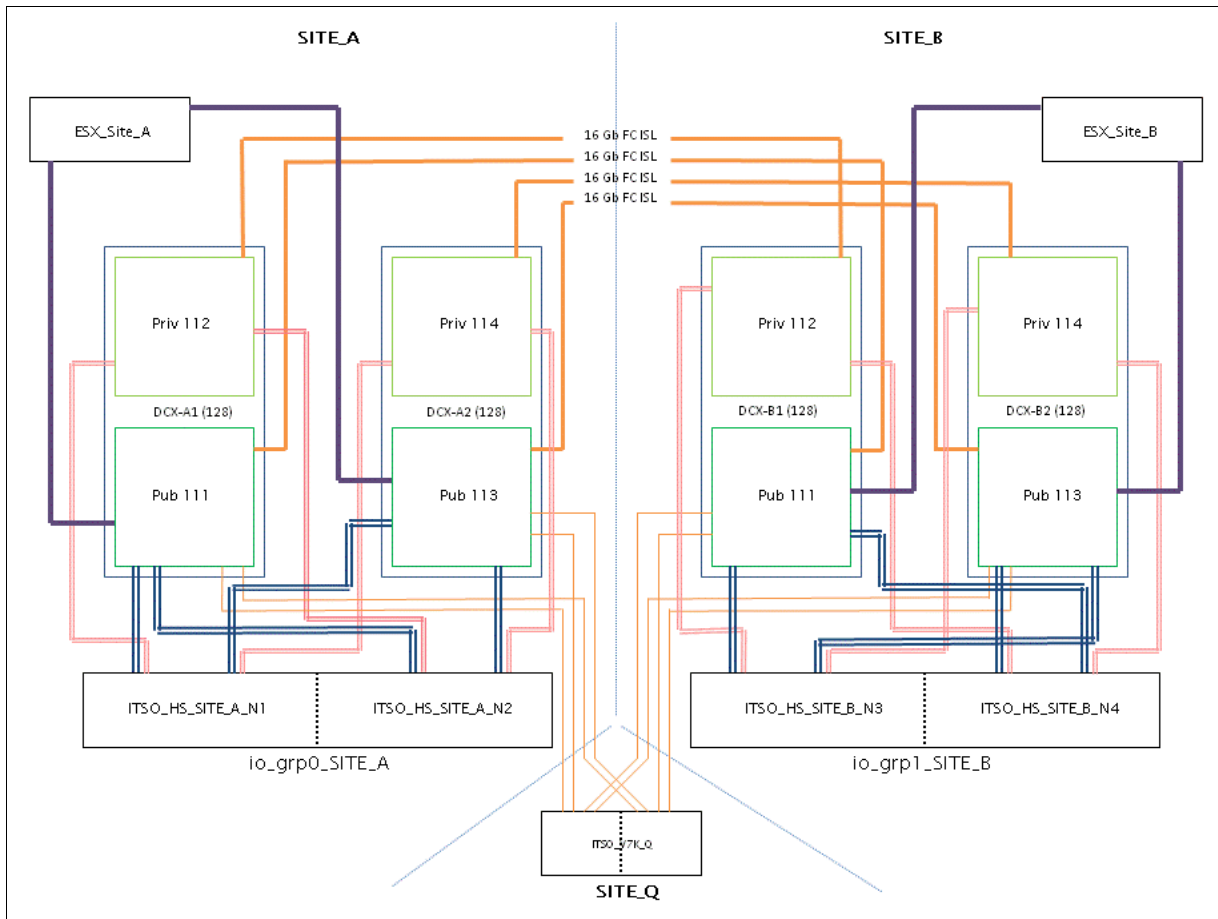


Figure 4-3 SAN design from host, Storwize V7000, and SAN components' point of view

## 4.2 IBM Fibre Channel SAN

The following section is based on the assumption that you are familiar with general Fibre Channel (FC) SAN design and technologies. Typically, the SAN design has servers and storage that connect into dual, redundant fabrics. The lab configuration has a redundant fabric design that uses two IBM SAN768B-2 chassis at each data center site. Each IBM SAN768B-2 is equipped with an IBM FC 16 Gbps 48-port blade in Slot 8.

The Storwize V7000 HyperSwap system in the lab was implemented in an inter-switch link (ISL) configuration that requires two types of SAN:

- ▶ **Public SAN:** In this type, server hosts, storage, and Storwize V7000 nodes connect. Data storage traffic traverses the Public SAN.
- ▶ **Private SAN:** Only the Storwize V7000 node canisters connect into the Private SAN, which is used for cluster and replication traffic.

The Virtual Fabric feature of the IBM SAN768B-2 was used to implement segregated Private and Public SANs on the same chassis. The following sections describe how the SAN was implemented.

## 4.2.1 Logical switches and virtual fabric configuration

A total of four virtual fabrics, with two logical switches each, were created.

Figure 4-4 shows a picture of the SAN port topology in Site\_A.

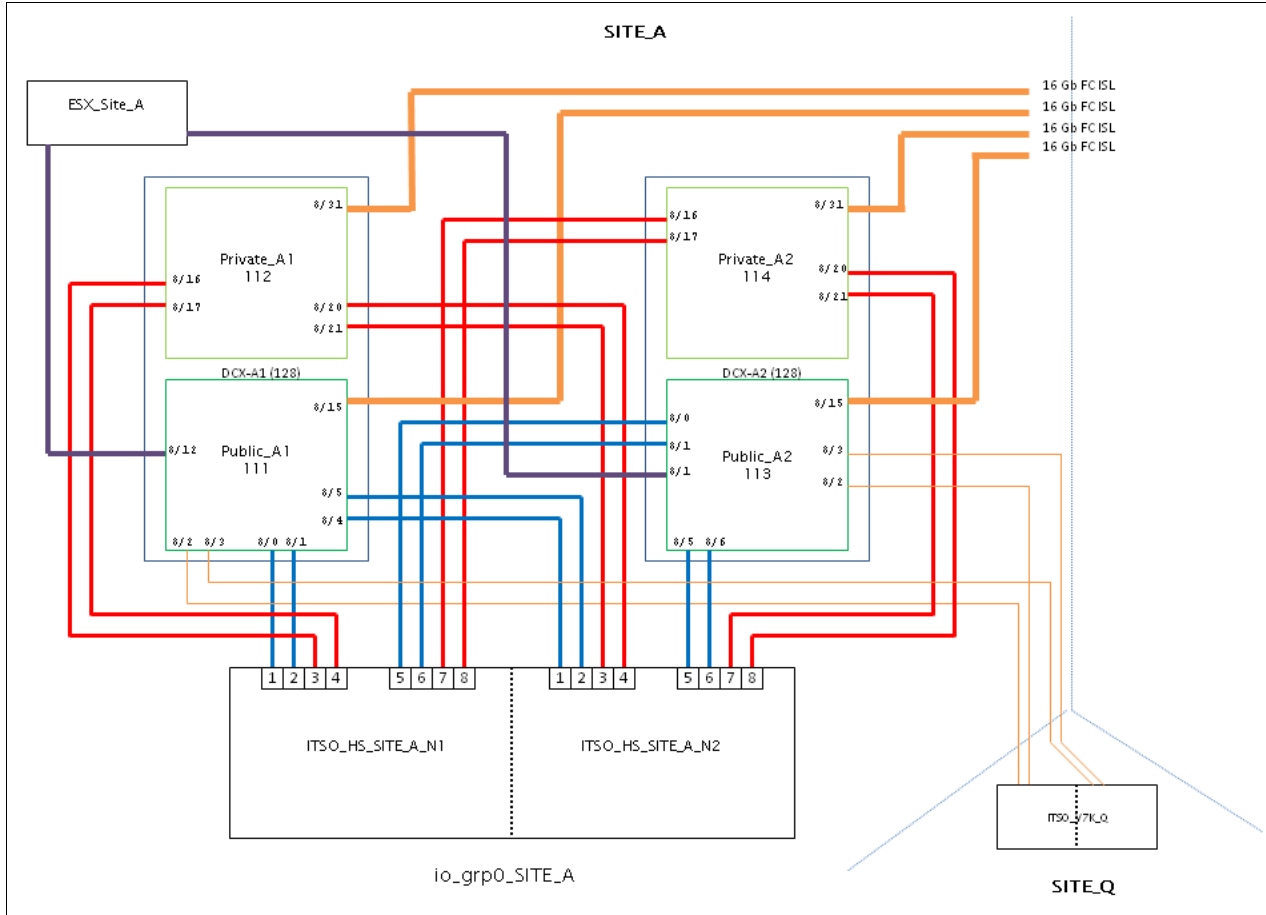


Figure 4-4 Site\_A: SAN port topology

Figure 4-5 shows a picture of the SAN port topology in Site\_B.

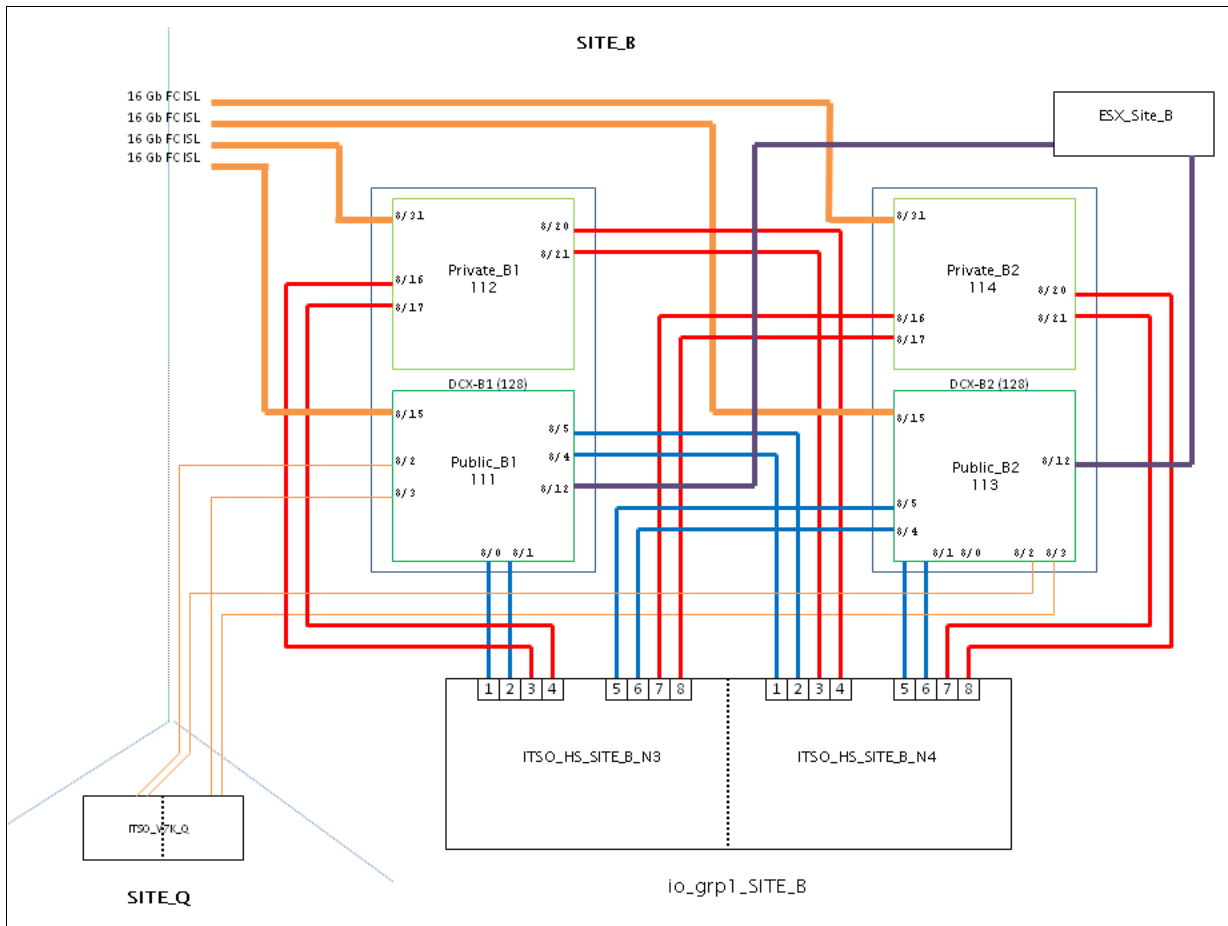


Figure 4-5 Site\_B: SAN port topology

Table 4-1 reports the virtual fabric design for Public and Private SANs.

Table 4-1 Virtual fabric layout

Virtual fabric name	Physical switch	Logical switch	Logical switch number	Ports
Fabric-Public-1	SAN768B-2_A1	Public_A1	111	8/0 - 8/15
	SAN768B-2_B1	Public_B1		8/0 - 8/15
Fabric-Public-2	SAN768B-2_A2	Public_A2	113	8/0 - 8/15
	SAN768B-2_B2	Public_B2		8/0 - 8/15
Fabric-Private-1	SAN768B-2_A1	Private_A1	112	8/16 - 8/31
	SAN768B-2_B1	Private_B1		8/16 - 8/31
Fabric-Private-2	SAN768B-2_A2	Private_A2	114	8/16 - 8/31
	SAN768B-2_B2	Private_B2		8/16 - 8/31

Ports 8/15 and 8/31 were used as ISLs for the Public and Private SANs.

Example 4-1 shows how to create a logical switch by using the Fabric Operating System (FOS) commands.

*Example 4-1 Creating the Public\_A1 logical switch*

---

```
SAN768B-2_A1:FID128:admin> lscfg --create 111  
About to create switch with fid=111. Please wait...  
Logical Switch with FID (111) has been successfully created.
```

Logical Switch has been created with default configurations. Please configure the Logical Switch with appropriate switch and protocol settings before activating the Logical Switch.

```
SAN768B-2_A1:FID128:admin> lscfg --config 111 -slot 8 -port 0-15  
This operation requires that the affected ports be disabled.  
Would you like to continue [y/n]?: y  
Making this configuration change. Please wait...  
Configuration change successful.  
Please enable your ports/switch when you are ready to continue.
```

```
SAN768B-2_A1:FID128:admin> setcontext 111  
Please change passwords for switch default accounts now.  
Use Control-C to exit or press 'Enter' key to proceed.
```

Password was not changed. Will prompt again at next login until password is changed.

```
switch_111:FID111:admin> switchname Public_A1  
Done.
```

---

Example 4-2 shows the switch configuration for the logical switch Public\_A1. All of the other logical switches that were created have similar configurations.

*Example 4-2 Public\_A1 logical switch configuration*

---

```
SAN768B-2_A1:FID128:admin> setcontext 111  
Public_A1:FID111:admin> switchenable  
Public_A1:FID111:admin> switchshow  
switchName:      Public_A1  
switchType:      121.3  
switchState:     Online  
switchMode:      Native  
switchRole:      Principal  
switchDomain:    1  
switchId:        fffc01  
switchWwn:       10:00:00:05:33:b5:3e:01  
zoning:          OFF  
switchBeacon:    OFF  
FC Router:       OFF  
Allow XISL Use:  ON  
LS Attributes:   [FID: 111, Base Switch: No, Default Switch: No, Address Mode 0]
```

Index	Slot	Port	Address	Media	Speed	State	Proto	
192	8	0	010000	id	N8	Online	FC F-Port	50:05:07:68:0b:21:21:a8
193	8	1	010100	id	N8	Online	FC F-Port	50:05:07:68:0b:22:21:a8
194	8	2	010200	id	N8	Online	FC F-Port	50:05:07:68:02:10:00:ef
195	8	3	010300	id	N8	Online	FC F-Port	50:05:07:68:02:10:00:f0

```

196 8 4 010413 id N8 Online FC F-Port 50:05:07:68:0b:21:21:a9
197 8 5 010513 id N8 Online FC F-Port 50:05:07:68:0b:22:21:a9
198 8 6 010613 id N8 Online FC F-Port 50:05:07:68:0c:23:00:00
199 8 7 010713 id N8 Online FC F-Port 50:05:07:68:0c:24:00:00
200 8 8 010813 id N8 Online FC F-Port 50:05:07:68:02:10:2b:6c
201 8 9 010913 id N8 Online FC F-Port 50:05:07:68:02:10:2b:6d
202 8 10 010a13 id N8 Online FC F-Port 50:05:07:68:02:10:05:a8
203 8 11 010b13 id N8 Online FC F-Port 50:05:07:68:02:10:05:a9
204 8 12 010c13 id N8 Online FC F-Port 10:00:8c:7c:ff:09:6f:81
205 8 13 010d13 id N8 Online FC F-Port 10:00:00:05:1e:c7:6b:91
206 8 14 010e13 id N8 No_Light FC
207 8 15 010f13 id N16 Online FC E-Port 10:00:00:05:1e:90:16:e9
"Public_B1" (downstream)(Trunk master)
Public_A1:FID111:admin>

```

---

For more information about creating virtual fabrics, see *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116.

## 4.2.2 Zoning configuration

In this section, the zoning configurations for both Public and Private SANs are described.

### Private SAN

In a Storwize V7000 HyperSwap configuration, the Private SAN is used both for internal cluster and replication traffic. One zone with all of the ports connected to the logical switches was created in each Private SAN.

Example 4-3 shows the effective configuration for the fabric Fabric-Private-1.

*Example 4-3 Fabric-Private-1 effective configuration*

---

```

Private_A1:FID112:admin> cfgshow
Defined configuration:
  cfg:  ITS0_PRI_F1
        Z_V7K_HS
zone:  Z_V7K_HS
        V7K_HS_IOGA_C1_P3; V7K_HS_IOGA_C1_P4; V7K_HS_IOGA_C2_P3;
        V7K_HS_IOGA_C2_P4; V7K_HS_IOGB_C1_P3; V7K_HS_IOGB_C1_P4;
        V7K_HS_IOGB_C2_P3; V7K_HS_IOGB_C2_P4
.
multiple lines omitted
.
Effective configuration:
  cfg:  ITS0_PRI_F1
zone:  Z_V7K_HS
        50:05:07:68:0b:23:21:a8
        50:05:07:68:0b:24:21:a8
        50:05:07:68:0b:23:21:a9
        50:05:07:68:0b:24:21:a9
        50:05:07:68:0b:23:21:7a
        50:05:07:68:0b:24:21:7a
        50:05:07:68:0b:23:21:7b
        50:05:07:68:0b:24:21:7b

```

---

## Public SAN

In a Storwize V7000 HyperSwap configuration, the Public SAN is used for hosts, external storage, and quorum communication.

Because our lab used no external storage controllers to provide storage capacity, only two types of zones were defined:

- ▶ Quorum zone. This zone contains all of the Storwize V7000 ports that connect to the fabric (on both logical switches) and all of the ports of the external storage controller (that in this case is still a Storwize V7000) that connect to the fabric.
- ▶ Host zones. These zones contain at least one port for each Storwize V7000 node canister and only one host port (single initiator zone).

Example 4-4 shows the effective configuration for the fabric Fabric-Public-1.

### Example 4-4 Fabric-Public-1 effective configuration

---

```
Public_A1:FID111:admin> cfgshow
Defined configuration:
cfg:   ITS0_PUB_F1
      Z_V7K_HS_QUORUM;Z_VMW55A_P1_V7K_HS; Z_VMW55A_P2_V7K_HS;
      Z_VMW6B_P1_V7K_HS; Z_VMW6B_P2_V7K_HS;Z_VMW6NA_P1_V7K_HS;
      Z_VMW6NA_P2_V7K_HS; Z_VMW6NB_P1_V7K_HS; Z_VMW6NB_P2_V7K_HS
zone:  Z_V7K_HS_QUORUM
      V7K_HS_IOGA_C1_P1; V7K_HS_IOGA_C1_P2; V7K_QUORUM_C1_P1;
      V7K_QUORUM_C2_P1; V7K_HS_IOGA_C2_P1; V7K_HS_IOGA_C2_P2;
      V7K_HS_IOGB_C1_P1; V7K_HS_IOGB_C1_P2; V7K_QUORUM_C1_P2;
      V7K_QUORUM_C2_P2; V7K_HS_IOGB_C2_P1; V7K_HS_IOGB_C2_P2
zone:  Z_VMW55A_P1_V7K_HS
      V7K_HS_IOGA_C1_P1; V7K_HS_IOGA_C2_P1; V7K_HS_IOGB_C1_P1;
      V7K_HS_IOGB_C2_P1; VMW55_SITEA_P1
zone:  Z_VMW55A_P2_V7K_HS
      V7K_HS_IOGA_C1_P2; V7K_HS_IOGA_C2_P2; V7K_HS_IOGB_C1_P2;
      V7K_HS_IOGB_C2_P2; VMW55_SITEA_P2
.
multiple lines omitted
.
```

```
Effective configuration:
cfg:   ITS0_PUB_F1
zone:  Z_V7K_HS_QUORUM
      50:05:07:68:0b:21:21:a8
      50:05:07:68:0b:22:21:a8
      50:05:07:68:02:10:00:ef
      50:05:07:68:02:10:00:f0
      50:05:07:68:0b:21:21:a9
      50:05:07:68:0b:22:21:a9
      50:05:07:68:0b:21:21:7a
      50:05:07:68:0b:22:21:7a
      50:05:07:68:02:20:00:ef
      50:05:07:68:02:20:00:f0
      50:05:07:68:0b:21:21:7b
      50:05:07:68:0b:22:21:7b
zone:  Z_VMW55A_P1_V7K_HS
      50:05:07:68:0b:21:21:a8
      50:05:07:68:0b:21:21:a9
```

```
50:05:07:68:0b:21:21:7a
50:05:07:68:0b:21:21:7b
10:00:8c:7c:ff:09:6f:81
zone: Z_VMW55A_P2_V7K_HS
50:05:07:68:0b:22:21:a8
50:05:07:68:0b:22:21:a9
50:05:07:68:0b:22:21:7a
50:05:07:68:0b:22:21:7b
10:00:00:05:1e:c7:6b:91
.
multiple lines omitted
.
```

---

To simplify the host and external storage zoning, the Storwize V7000 local port masking feature was used as described in “Local port masking setting” on page 79.

## 4.3 Storwize V7000 HyperSwap planning

In this section, general planning guidelines are described.

Before you implement any Storwize V7000 solution, it is important to perform a capacity planning exercise in terms of both physical resources and connectivity requirements. A good starting point might be to collect as much workload information, as possible. The workload analysis helps to you to size the number of I/O Groups (control enclosures) and the type and number of disk drives correctly.

For this activity, you can use a storage modeling tool, such as Disk Magic. With HyperSwap configurations, it is also important to understand the workload distribution between the sites because the workload distribution affects the connectivity requirements. Underestimating the connectivity resources can lead to poorly performing solutions.

For new Storwize V7000 implementations, ensure that the attached devices, host, and levels of the I/O stack components (host bus adapter (HBA) firmware, device drivers, and multipathing), are supported. Use the System Storage Interoperation Center (SSIC):

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

Check the V7.8.x Supported Hardware List, Device Driver, Firmware, and Recommended Software Levels for IBM Storwize V7000:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009559>

Also, check the latest code level that is available at this website:

<http://www.ibm.com/support/docview.wss?rs=591&uid=ssg1S1003705>

Finally, check the full list of supported extended quorum devices:

<https://ibm.biz/BdX9Xe>

The IBM Storwize V7000 in the lab environment is running version 7.8.0.0. The third site, the quorum storage, is on a Storwize V7000 that is also running version 7.6.1.5.

This book does not cover the physical installation or the initial configuration. This book is based on the assumption that you are familiar with the major concepts of Storwize V7000 systems, such as node canisters, I/O Groups, managed disks (MDisks), and quorum disks.



Also, see the IBM Storwize V7000 Knowledge Center:

<https://ibm.biz/BdsvBk>

The following sections list general planning considerations before you implement the HyperSwap function.

### 4.3.1 Active-active Metro Mirror considerations

The HyperSwap function is based on the new Metro Mirror active-active option. This new Metro Mirror option introduced the capability to define synchronous replication between volumes that are defined in two *different* I/O Groups (control enclosures) of the same Storwize V7000 cluster.

Before the active-active option, the intra-cluster replication only occurred among volumes in the *same* I/O Groups. In addition, an active-active Metro Mirror relationship uses thin-provisioned Change Volumes (CVs): one CV for the source volume and one CV for the target volume, which acted as the journaling volume during the resynchronization process. These additional copies guarantee that a consistent copy of the data is maintained in a failure during a resynchronization process.

To establish active-active Metro Mirror between two volumes, a remote copy relationship must be defined between the source volume (*Master Volume*) and the target volume (*Auxiliary Volume*).

When a relationship is first created, the Master Volume is always assigned the role of the Primary, and the Auxiliary Volume is assigned the role of the Secondary. These roles can be reversed at any stage, for instance, following a HyperSwap operation, with the Auxiliary becoming the Primary, and the Master becoming the Secondary. The primary attribute of a remote copy relationship identifies the volume that currently acts as source of the replication.

Example 4-5 reports the output of the `lsrcrelationship` command that shows the involved volume and the primary attribute.

**Note:** Because this example is based on the command-line interface (CLI), the “volume” is referred to as the “vdisk”.

#### Example 4-5 `lsrcrelationship` command

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsrcrelationship 0
id 0
name Rel_ESX_A
master_cluster_id 00000100216001E0
master_cluster_name ITS0_V7K_HyperSwap
master_vdisk_id 0
master_vdisk_name V_ESX_A_Master
aux_cluster_id 00000100216001E0
aux_cluster_name ITS0_V7K_HyperSwap
aux_vdisk_id 5
aux_vdisk_name V_ESX_A_Aux
primary master
consistency_group_id 0
consistency_group_name CG_ESX_AtoB
state consistent_synchronized
bg_copy_priority 50
progress
```

```
freeze_time
status online
sync
copy_type activeactive
cycling_mode
cycle_period_seconds 300
master_change_vdisk_id 3
master_change_vdisk_name V_ESX_A_Master_CV
aux_change_vdisk_id 6
aux_change_vdisk_name V_ESX_A_Aux_CV
```

---

The HyperSwap function enforces rules when you define the active-active relationships:

- ▶ Master and Auxiliary Volumes must belong to different I/O Groups with different site definitions.
- ▶ Master and Auxiliary Volumes must be placed in different storage pools with different site definitions.
- ▶ A Master Volume and an Auxiliary Volume must be managed by an I/O Group with the same site definition as the storage pool that provides the capacity for the volume.
- ▶ Storage controllers that are defined on site 3 (*quorum site*) cannot provide capacity for the volumes to be defined in an active-active relationship.
- ▶ The Master CV must be defined in the same I/O Group as the Master Volume and must use capacity from a storage pool in the same site.
- ▶ The Auxiliary CV must be defined in the same I/O Group as the Auxiliary Volume and must use capacity from a storage pool in the same site.
- ▶ The Master CV must be the same size as the Master Volume.
- ▶ The Auxiliary CV must be the same size as the Auxiliary Volume.
- ▶ An active-active relationship cannot be created if a Master Volume is mapped to a host with no site definition.

Other general remote copy restrictions, which are not specific to active-active relationships, also apply:

- ▶ Master and Auxiliary Volumes in a Metro Mirror relationship must be the same size.
- ▶ Volumes in a Metro Mirror relationship cannot be expanded or shrunk.
- ▶ Volumes in a Metro Mirror relationship cannot be moved between I/O Groups.

HyperSwap is a two-site active-active solution, which means that no restrictions exist for using the same I/O Group to manage both Master and Auxiliary Volumes. We recommend that you spread the volumes that are managed by an I/O Group evenly in both nodes.

With the Storwize V7000 HyperSwap function, you can group multiple active-active Metro Mirror relationships together for high availability (HA) by creating *Consistency Groups*. The use of Consistency Groups is important where an application spans many volumes and requires that the data is consistent across more volumes. All of the relationships that belong to a Consistency Group must be consistent in terms of Master and Auxiliary site definition, even though this rule is not a strict requirement. In fact, when a relationship is added to a non-empty Consistency Group, it sets the Master and Auxiliary roles according to the replication direction of the Consistency Group.

As described in Chapter 3, “IBM System Storwize V7000 HyperSwap architecture” on page 23 in the current HyperSwap implementation, the read and write operations are always routed to the Primary copy. Therefore, hosts that access the Secondary copy will experience

an increased latency in the I/O operations. As a mitigation of this behavior, if sustained workload (that is, more than 75% of I/O operations for at least 20 minutes) is running over Secondary volumes, the HyperSwap function will switch the direction of the active-active relationships, swapping the Secondary volume to Primary and vice versa.

HyperSwap Volumes in Consistency Groups all switch direction together. So, the direction that a set of active-active relationships in a Consistency Group will replicate will depend on which of the two sites has the most host I/O across all HyperSwap Volumes.

Defining an effective distribution of the Master and Auxiliary Volumes according to the real workload distribution is important in a HyperSwap configuration. For instance, consider a uniform VMware environment where a datastore with multiple virtual machines (VMs) is accessed from ESX servers in both sites.

**Note:** The distribution of VMs among the ESX servers also determines the workload distribution between the Primary and the Secondary copy of the HyperSwap Volume that contains the datastore.

In this scenario, the sustained workload might run on both sites in different time frames, which will lead the HyperSwap function to swap the active-active relationship back and forth to adjust the direction of the relationship according to the workload. To avoid this thrashing behavior, ensure that a datastore is only used for VMs primarily running on a single site and define the Master and Auxiliary Volumes.

## 4.3.2 Quorum disk considerations

The quorum disk fulfills two functions for cluster reliability:

- ▶ Acts as a tiebreaker in split-brain scenarios
- ▶ Saves critical configuration metadata

Starting with version 7.6, the quorum disk acting as a tiebreaker can be replaced with an IP quorum device, as described in 4.3.3, “IP Quorum” on page 69. The Storwize V7000 HyperSwap solution not using the IP Quorum feature requires an external controller to provide the active quorum disk. In this section, standard disk quorum considerations are discussed.

The Storwize V7000 quorum algorithm distinguishes between the active quorum disk and quorum disk candidates. Three quorum disk candidates exist. At any time, only one of these candidates acts as the active quorum disk. The other two candidates are reserved to become active if the current active quorum disk fails. All three quorum disks are used to store configuration metadata, but only the active quorum disk acts as the tiebreaker for split-brain scenarios.

The quorum disks assignment can be automatic (default) or manual. With the automatic assignment, the quorum disks are chosen by the Storwize V7000 code with an internal selection algorithm. The user can manually select the quorum disks and override the automatic selection.

Specific requirements apply to the quorum disks' placement when the hyperswap topology is enabled. The HyperSwap enforces that a quorum disk is placed in each of the three failure domains. Storwize V7000 chooses only quorum disks that are configured to site 3 as the active quorum disk and chooses only quorum disks that are configured to site 1 or site 2 as candidate quorum disks. Any MDisk that is not assigned to a site is ignored in the quorum disk selection.

If a site has no suitable MDisks, fewer than three quorum disks are automatically created. For example, if the Storwize V7000 can select only two quorum disks, only two are used. However, enabling HyperSwap configuration with fewer than one quorum disk for each site is not recommended because it can seriously affect the operation in case of split-brain and rolling disaster scenarios.

If you are not virtualizing any external storage in site 1 and site 2, the quorum disks for those sites will be chosen as in any regular Storwize V7000 implementation as internal disk drives. The site 3 quorum disk recommendation is to use a dedicated 1 GB MDisk that is assigned to a storage pool that contains only the quorum disk.

If you are virtualizing external storage in site 1 and site 2, the general recommendation is to create dedicated MDisks to be used as quorum disks:

1. Create three 1 GB MDisks (one for each site).
2. Create three storage pools (one for each site), which contain only the 1 GB disk each.
3. Manually configure the quorum disks by using the 1 GB disks. Set the MDisk in site 3 as the active quorum disk.

To modify the quorum disk assignment, use either the CLI or graphical user interface (GUI).

The CLI output in Example 4-6 shows that the Storwize V7000 cluster initially automatically assigns the quorum disks.

*Example 4-6 lsquorum command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsquorum
quorum_index status id name controller_id controller_name active
object_type override site_id site_name
0 online 8 no drive
no 1 ITS0_SITE_A
1 online 15 yes drive
no 2 ITS0_SITE_B
2 online 2 mdisk0_V7K_HS_Q 0 ITS0_V7K_Q_N1 no mdisk
no 3 ITS0_SITE_Q
```

---

To change from automatic selection to manual selection, run the commands that are shown in Example 4-7.

*Example 4-7 chquorum command to assign the quorum disks manually*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chquorum -override yes -mdisk 2 2
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chquorum -override yes -drive 15 1
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chquorum -override yes -drive 8 0
```

---

Finally, to set the active quorum disk, run the command that is shown in Example 4-8.

*Example 4-8 chquorum command to change the active quorum disks*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chquorum -active 2
```

---

After that process is complete, when you run the `lsquorum` command, you get output as shown in Example 4-9.

Example 4-9 `lsquorum` command

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsquorum
quorum_index status id name controller_id controller_name active
object_type override site_id site_name
0 online 8 no drive
yes 1 ITS0_SITE_A
1 online 15 no drive
yes 2 ITS0_SITE_B
2 online 2 mdisk0_V7K_HS_Q 0 ITS0_V7K_Q_N1 yes mdisk
yes 3 ITS0_SITE_Q
```

---

The output shows that the `ITS0_V7K_Q_N1` controller in power domain 3, site 3, is now the active quorum disk.

The storage controller that provides the quorum disk in the Storwize V7000 HyperSwap configuration in the third site must be supported as an *extended quorum disk*. The storage controllers that provide extended quorum support are listed on the Storwize V7000 Support Portal web page:

<https://ibm.biz/BdX9Xe>

### 4.3.3 IP Quorum

Storwize V7000 version 7.6 introduced the IP Quorum feature that eliminates requirement for Fibre Channel networking and disk storage at third site. This feature deploys a Java application to a third site that act as a tie-breaker in split-brain scenarios. The Java application runs on a standard server and needs only standard network connectivity to be exploited. See Chapter 3, “IBM System Storwize V7000 HyperSwap architecture” on page 23 for further details on IP quorum requirements.

To implement the IP Quorum function, the following action must be performed.

1. Create the quorum application. This can be accomplished either using the CLI or the GUI.
  - a. Using the CLI, the command `mkquorumapp` must be used, as shown in Example 4-10.

Example 4-10 *The `mkquorumapp` command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkquorumapp
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>
```

---

The quorum application is created with the `ip_quorum.jar` name and it is available in the `/dumps` directory, as shown in the Example 4-11:

Example 4-11 *The `lsdumps` command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsdumps
id filename
0 reinst.7836494-1.trc
1 snap.single.7836494-2.141021.123724.tgz
2 snap.single.7836494-1.141021.123839.tgz
.
multiple lines omitted
.
```

---

## 25 ip\_quorum.jar

To download the quorum application, you can use either the CLI or the GUI. With the CLI the *pscp* tool must be used, as shown in the Example 4-12:

*Example 4-12 pscp command to download the quorum app*

```
pscp -unsafe -load V7K_HS admin@V7K_ip:/dumps/ip_quorum.jar local_directory
```

If you prefer to use the GUI, click **Settings** → **Support** page, as shown in Figure 4-6. If the page is not displaying a list of individual log files, click **Show full log listing**.

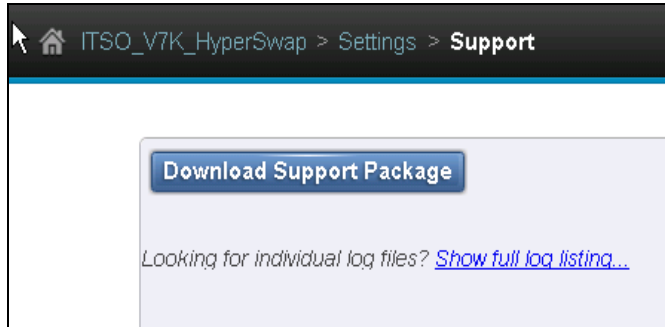


Figure 4-6 Download Support Package window

Next, right-click the row for the *ip\_quorum.jar* file and choose **Download**, as shown in Figure 4-7.

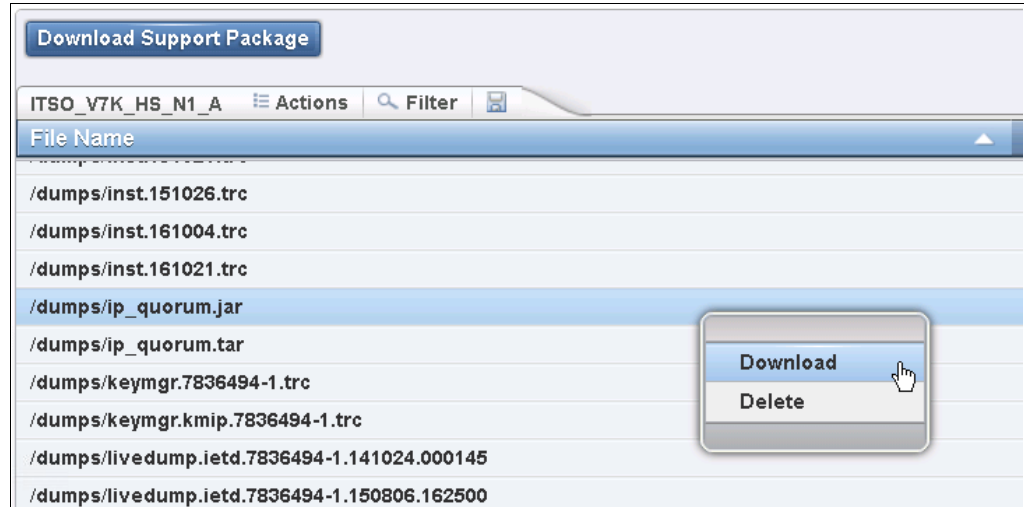


Figure 4-7 Quorum application download

The file is downloaded to your local workstation.

- b. To create and download the quorum application through the GUI, go in **Settings** → **System** menu and select the IP quorum on the left panel, as shown in Figure 4-8 on page 71.



Figure 4-8 Select the IP Quorum

**GUI support:** The GUI support for the IP quorum has been introduced with Storwize V7000 version 7.7

In the IP Quorum menu click on **Download IPv4 Application** (or **Download IPv6 Application** if you are using IPv6 networks) to start the quorum application creation, as shown in Figure 4-9.

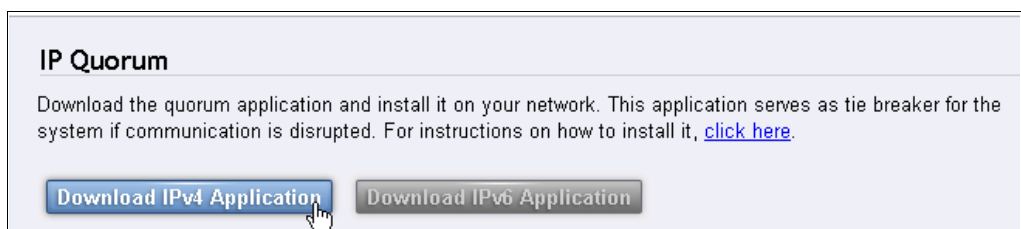


Figure 4-9 IP quorum download section

As soon as the quorum application is created, the download is initiated to the local workstation.

2. Install the quorum application. To install the quorum application, first transfer the application to a directory on the host that is to run the IP quorum application. Then verify the IP connectivity using the **ping** command from the host to service address of each node of the Storwize V7000 cluster. Finally, on the host, use the command **java -jar ip\_quorum.jar** to initialize the IP quorum application. As soon as the quorum application is initialized, two files are created in the directory containing the quorum application, as shown in Example 4-13.

#### Example 4-13 The IP quorum files

```
[root@oc2244547338 IPQ]# ls -la ip*
-rw-r--r--. 1 root root 52413 Oct 25 10:27 ip_quorum.jar
-rw-r--r--. 1 root root  2089 Oct 25 11:39 ip_quorum.log
-rw-r--r--. 1 root root    0 Oct 25 11:39 ip_quorum.log.lck
```

The **ip\_quorum.log** contains the application initialization and the heartbeating information and can be used for troubleshooting in case of issues. The **ip\_quorum.log.lck** file is created once the quorum application is started. A sample of the **ip\_quorum.log** content is reported in Example 4-14.

#### Example 4-14 ip\_quorum.log content

```
2016-10-25 11:38:26:170 Quorum CONFIG: === IP quorum ===
2016-10-25 11:38:26:176 Quorum CONFIG: Name set to null.
2016-10-25 11:38:28:783 Quorum FINE: Node 10.18.228.57:1,260 (0, 0)
2016-10-25 11:38:28:798 Quorum FINE: Node 10.18.228.58:1,260 (0, 0)
2016-10-25 11:38:28:798 Quorum FINE: Node 10.18.228.55:1,260 (0, 0)
2016-10-25 11:38:28:798 Quorum FINE: Node 10.18.228.56:1,260 (0, 0)
2016-10-25 11:38:28:798 Quorum CONFIG: Successfully parsed the configuration,
found 4 nodes.
2016-10-25 11:38:28:798 10.18.228.57 [9] INFO: Trying to open socket
2016-10-25 11:38:28:798 10.18.228.58 [10] INFO: Trying to open socket
2016-10-25 11:38:28:798 10.18.228.55 [11] INFO: Trying to open socket
2016-10-25 11:38:28:798 10.18.228.56 [12] INFO: Trying to open socket
2016-10-25 11:38:29:205 10.18.228.55 [11] INFO: Creating UID
2016-10-25 11:38:29:205 10.18.228.57 [9] INFO: Waiting for UID
2016-10-25 11:38:29:205 10.18.228.56 [12] INFO: Waiting for UID
2016-10-25 11:38:29:205 10.18.228.55 [11] FINE: <Msg [protocol=1, sequence=0,
command=CREATE_UID_REQUEST, length=0]
2016-10-25 11:38:29:220 10.18.228.58 [10] INFO: Waiting for UID
2016-10-25 11:38:29:486 10.18.228.55 [11] FINE: >Msg [protocol=1, sequence=0,
command=CREATE_UID_RESPONSE, length=16] Data [quorumUid=4,
clusterUid=1100071567886]
2016-10-25 11:38:29:486 10.18.228.55 [11] FINE: <Msg [protocol=2, sequence=1,
command=CONNECT_REQUEST, length=88] Data [quorumUid=4,
clusterUid=1100071567886, generationId=0, shortLeaseExtension=true,
infoText=ITS0-1.englab.brocade.com/10.18.228.170]
2016-10-25 11:38:29:486 10.18.228.58 [10] INFO: *Connecting
2016-10-25 11:38:29:486 10.18.228.58 [10] FINE: <Msg [protocol=2, sequence=0,
command=CONNECT_REQUEST, length=88] Data [quorumUid=4,
clusterUid=1100071567886, generationId=0, shortLeaseExtension=true,
infoText=ITS0-1.englab.brocade.com/10.18.228.170]
2016-10-25 11:38:29:486 10.18.228.56 [12] INFO: *Connecting
2016-10-25 11:38:29:502 10.18.228.56 [12] FINE: <Msg [protocol=2, sequence=0,
command=CONNECT_REQUEST, length=88] Data [quorumUid=4,
clusterUid=1100071567886, generationId=0, shortLeaseExtension=true,
infoText=ITS0-1.englab.brocade.com/10.18.228.170]
2016-10-25 11:38:29:502 10.18.228.57 [9] INFO: *Connecting
2016-10-25 11:38:29:502 10.18.228.57 [9] FINE: <Msg [protocol=2, sequence=0,
command=CONNECT_REQUEST, length=88] Data [quorumUid=4,
clusterUid=1100071567886, generationId=0, shortLeaseExtension=true,
infoText=ITS0-1.englab.brocade.com/10.18.228.170]
2016-10-25 11:38:29:705 10.18.228.55 [11] FINE: >Msg [protocol=1, sequence=1,
command=CONNECT_RESPONSE, length=4] Data [result=SUCCESS]
2016-10-25 11:38:29:705 10.18.228.55 [11] INFO: Connected to 10.18.228.55
```



```

2016-10-25 11:38:29:736 10.18.228.58 [10] FINE: >Msg [protocol=1, sequence=0,
command=CONNECT_RESPONSE, length=4] Data [result=SUCCESS]
2016-10-25 11:38:29:736 10.18.228.58 [10] INFO: Connected to 10.18.228.58
2016-10-25 11:38:29:752 10.18.228.56 [12] FINE: >Msg [protocol=1, sequence=0,
command=CONNECT_RESPONSE, length=4] Data [result=SUCCESS]
2016-10-25 11:38:29:752 10.18.228.56 [12] INFO: Connected to 10.18.228.56
2016-10-25 11:38:29:752 10.18.228.57 [9] FINE: >Msg [protocol=1, sequence=0,
command=CONNECT_RESPONSE, length=4] Data [result=SUCCESS]
2016-10-25 11:38:29:752 10.18.228.57 [9] INFO: Connected to 10.18.228.57
2016-10-25 11:38:36:316 10.18.228.57 [9] FINE: <Msg [protocol=1, sequence=1,
command=HEARTBEAT_REQUEST, length=0]
2016-10-25 11:38:36:316 10.18.228.55 [11] FINE: <Msg [protocol=1, sequence=2,
command=HEARTBEAT_REQUEST, length=0]
2016-10-25 11:38:36:316 10.18.228.58 [10] FINE: <Msg [protocol=1, sequence=1,
command=HEARTBEAT_REQUEST, length=0]
2016-10-25 11:38:36:316 10.18.228.56 [12] FINE: <Msg [protocol=1, sequence=1,
command=HEARTBEAT_REQUEST, length=0]
2016-10-25 11:38:36:316 10.18.228.57 [9] FINE: >Msg [protocol=1, sequence=1,
command=HEARTBEAT_RESPONSE, length=0]
2016-10-25 11:38:36:316 10.18.228.55 [11] FINE: >Msg [protocol=1, sequence=2,
command=HEARTBEAT_RESPONSE, length=0]
2016-10-25 11:38:36:316 10.18.228.58 [10] FINE: >Msg [protocol=1, sequence=1,
command=HEARTBEAT_RESPONSE, length=0]
2016-10-25 11:38:36:316 10.18.228.56 [12] FINE: >Msg [protocol=1, sequence=1,
command=HEARTBEAT_RESPONSE, length=0]

```

---

3. Checking the IP quorum status. Once the quorum application is started, the new quorum is automatically added to the Storwize V7000 system. To check the new quorum status through the CLI use the `1squorum` command, as shown in Example 4-15.

*Example 4-15 Isquorum output*

```

quorum_index status id name controller_id controller_name active object_type
override site_id site_name
0         online 11
                no      drive      no
1         Site_A
2         online 5
                no      drive      no
3         Site_B
3         online
                yes   device    no
ITS0-1.englab.brocade.com/10.18.228.170

```

---

In the GUI, the IP quorum status can be checked on the **Settings** → **System** → **IP Quorum** menu. The **Detected IP quorum Applications** section displays the quorum application status, as shown in Figure 4-10 on page 74.

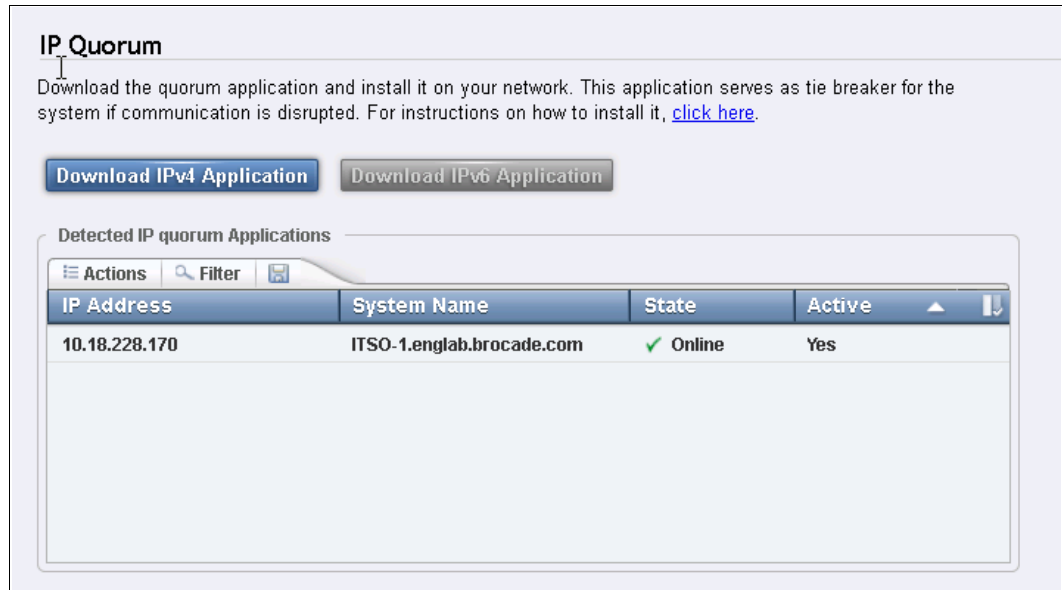


Figure 4-10 Detected IP quorum Applications

## 4.4 Storwize V7000 HyperSwap configuration

The following sections describe how the HyperSwap configuration in the lab environment was implemented by using the CLI and the GUI. The steps that are described are generally valid to implement any Storwize V7000 HyperSwap solutions.

### 4.4.1 Configuring the HyperSwap system topology using the CLI

Before you enable the HyperSwap function, you must set several Storwize V7000 object values.

#### Site names

The *site* attribute corresponds to a physical location that houses the physical objects of the system. In a client installation, the site attribute might correspond to a true separate office, a different data center building, or simply different rooms or racked areas of a single data center that was planned for internal redundancy.

Four predefined site IDs are available for different purposes, as shown in Table 4-2.

Table 4-2 Site definitions

Site ID	Default site name	Objects that can be in site	Purpose
None	It has no name, and it cannot be renamed.	Hosts, nodes, and controllers	The default site for objects when they are not assigned to a specific site. The hyperswap topology requires objects to be in a site other than 0.
1	Site 1	Hosts, nodes, and controllers	The first of two sites between which to perform HA. Site 1 has no implied preferences compared to site 2.
2	Site 2	Hosts, nodes, and controllers	The second of two sites between which to perform HA. Site 2 has no implied preferences compared to site 1.
3	Site 3	Controllers	A third site that provides quorum abilities to act as a tiebreak between sites 1 and 2 when connectivity is lost.

Sites 1, 2, and 3 can be renamed from the default name by using the **chsite** command, as shown in Example 4-16.

Example 4-16 chsite command

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chsite -name ITS0_SITE_A 1
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chsite -name ITS0_SITE_B 2
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chsite -name ITS0_SITE_Q 3
```

---

To list the sites, you can use the **lssite** command, as shown in Example 4-17.

Example 4-17 lssite command

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lssite
id site_name
1 ITS0_SITE_A
2 ITS0_SITE_B
3 ITS0_SITE_Q
```

---

## Node canisters

With a HyperSwap system topology, all node canisters of a control enclosure must belong to the same site. You need to assign one control enclosure to each of sites 1 and 2. So, to configure HyperSwap Volumes in a Storwize V7000, you need a system with at least two control enclosures. For SAN Volume Controller, at least four nodes are required.

For larger systems that use the HyperSwap system topology, if most volumes are configured as HyperSwap Volumes, it is preferable to put two control enclosures on each site, instead of one control enclosure on one site and two control enclosures on the other site. This way, the site with only one control enclosure does not become a bottleneck.

If you are not using the HyperSwap function on every volume, an asymmetric configuration of two or even three control enclosures on the site that contains the volumes that are not HyperSwap Volumes is possible, with only one control enclosure on the other site.

Also, in a configuration where all of the volumes use the HyperSwap function, it is possible to configure the system with more control enclosures on one site than on the other site, but beware that the site with fewer nodes might become a bottleneck.

When a Storwize V7000 system is created, the initial node canister will be assigned to site 0 (no site) by default, and it must be reconfigured manually to either site 1 or site 2.

To set the site attribute to a node canister, you can use the command **chnodecanister**, as shown in Example 4-18.

*Example 4-18 chnodecanister command*

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>chnodecanister -site ITSO_SITE_A 8
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>chnodecanister -site ITSO_SITE_A 9
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>chnodecanister -site ITSO_SITE_B 3
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>chnodecanister -site ITSO_SITE_B 4
```

**Note:** Changing the site attribute for a node canister is only allowed when the system topology is set to standard.

The current site definition for the node canisters can be checked by using the **lnodecanister** command, as shown in Example 4-19.

*Example 4-19 lnodecanister command*

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>lnodecanister
id name                UPS_serial_number WWNN                status IO_group_id
IO_group_name config_node UPS_unique_id hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number site_id
site_name
8 ITSO_HS_SITE_A_N1    500507680B0021A8 online 0
io_grp0_SITE_A yes    400
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsohssitean1 01-1 1
1 7836494 1 ITSO_SITE_A
9 ITSO_HS_SITE_A_N2    500507680B0021A9 online 0
io_grp0_SITE_A no    400
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsohssitean2 01-2 1
2 7836494 1 ITSO_SITE_A
3 ITSO_HS_SITE_B_N3    500507680B00217A online 1
io_grp1_SITE_B no    500
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsohssitebn3 02-1 2
1 7836640 2 ITSO_SITE_B
4 ITSO_HS_SITE_B_N4    500507680B00217B online 1
io_grp1_SITE_B no    500
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsohssitebn4 02-2 2
2 7836640 2 ITSO_SITE_B
```

Each I/O Group must have sufficient bitmap capacity that is defined by using the **chiogrp** command for the HyperSwap Volumes in addition to the necessary bitmap capacity requirements of other FlashCopy and Global Mirror or Metro Mirror objects.

For each volume in the HyperSwap configuration, for every 8 GB logical capacity, rounded up to the next highest 8 GB, you will need 4 KB remote bitmap memory and 8 KB flash bitmap memory, both of which are defined in both I/O Groups of the HyperSwap Volume. For example, for 1,024 volumes of 100 GB each, the bitmap requirements are 104 MB for remote and 51 MB for flash. To set the bitmap memory, you can use the **chiogrp** command, as shown in Example 4-20.

*Example 4-20 chiogrp command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chiogrp -feature remote -size 51
io_grp0_SITE_A
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chiogrp -feature flash -size 104
io_grp0_SITE_B
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chiogrp -feature remote -size 51
io_grp0_SITE_A
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chiogrp -feature flash -size 104
io_grp0_SITE_B
```

---

## Hosts

Also, host objects have a site attribute. For the existing hosts, the site can be configured by using the **chhost** command, as shown in Example 4-21.

*Example 4-21 chhost command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chhost -site ITS0_SITE_A 3
```

---

You can also specify the site attribute when the host is created by using the **mkhost** command, as shown in Example 4-22.

*Example 4-22 mkhost command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkhost -fcwwpn 100000051EC76BA1 -site
ITS0_SITE_A -name test2
Host, id [4], successfully created
```

---

You can check the current site definition for the host objects by using the **lshost** command, as shown in Example 4-23.

*Example 4-23 lshost command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lshost
id name                port_count iogrp_count status site_id site_name
0 ESX_HS_6.00_Site_A   2          4          online 1      ITS0_SITE_A
1 ESX_ESC_6.00_Site_B 2          4          online 2      ITS0_SITE_B
2 ESX_M5_6.00_Site_B  2          4          online 2      ITS0_SITE_B
3 test1                1          4          online 1      ITS0_SITE_A
4 test2                1          4          online 2      ITS0_SITE_B
```

---

**Note:** The system will configure host multipathing dynamically so that hosts in site 1 will preferentially send I/O to node canisters in site 1, and similarly for site 2. So, for optimum performance, all of the worldwide port names (WWPNs) that are associated with this host object must be on that site. For clustered host systems that are attached to both sites, you need to define a host object for each site to optimize the I/O for each physical server in the clustered host system.

When HyperSwap Volumes are mapped to a host by using `mkvdi skhostmap`, the host must be assigned to either site 1 or site 2. Assigning a HyperSwap Volume to a host with no site attribute is not allowed.

By default, host objects are associated with all I/O Groups (control enclosures). If you use the `-iogrp` parameter for the `mkhost` command to override this association, you will need to ensure that hosts that access HyperSwap Volumes are associated with at least the I/O Groups in which the Master and Auxiliary Volumes of the HyperSwap Volumes are cached. If the host is not associated with this I/O Group, the host cannot access HyperSwap Volumes through both sites.

## Controllers

For Storwize V7000, an internal storage array that was created by using the `mkarray` command obtains its site information from the node canisters in the control enclosure it is attached to. For virtualized external storage, you must tell the system the location of each storage controller. After the storage controllers are detected, you can set the site attribute by using the `chcontroller` command, as shown in Example 4-24.

*Example 4-24 chcontroller command*

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chcontroller -site ITS0_SITE_Q 1
```

**Note:** Changing the site attribute for a controller is only allowed when the system topology is set to standard.

You can check the current site definition for the controllers by using the `lscontroller` command, as shown in Example 4-25.

*Example 4-25 lscontroller command*

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lscontroller
id controller_name ctrl_s/n          vendor_id          product_id_low
product_id_high      site_id site_name
0 ITS0_V7K_Q_N1      2076              IBM                2145
3 ITS0_SITE_Q
1 ITS0_V7K_Q_N2      2076              IBM                2145
3 ITS0_SITE_Q
```

**Note:** Certain types of controllers, for instance, the Storwize family products, present themselves with multiple worldwide node names (WWNNs) to Storwize V7000. In this case, multiple controllers' objects (one for each WWNN) that refer to the same physical device are detected in the Storwize V7000 cluster. While this situation is not an issue for the Storwize V7000 cluster, you must be careful to ensure that the site definition is consistent among the controllers that refer to the same physical device.

Controllers must be assigned to site 1 or site 2 if they have any managed disks (MDisks) and the system is set to use the hyperswap topology. MDisks can be assigned to storage pools only if they are allocated from a storage controller with a well-defined site that matches that of the storage pool.

No check occurs to ensure that you set the sites on all controllers when you change the system to use the hyperswap topology, but the system will not let you create HyperSwap Volumes unless all of the MDisks in each storage pool have the site set up correctly. So, we recommend that you set up the controller sites before you change the topology.

**Note:** A storage pool inherits the site attribute from the MDisks that are contained in the storage pool. A storage pool that contains MDisks with different site attributes has no site definition.

## Topology

After all of the site settings are complete, you can change the system topology to hyperswap. Use the command **chsystem**, as shown in Example 4-26.

*Example 4-26 chsystem command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chsystem -topology hyperswap
```

---

To check the current topology setting, you can use the **lssystem** command, as shown in Example 4-27.

*Example 4-27 lssystem command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lssystem
id 00000100216001E0
name ITS0_V7K_HyperSwap
location local
.
multiple lines omitted
.
topology hyperswap
topology_status dual_site
rc_auth_method none
vdisk_protection_time 15
vdisk_protection_enabled no
product_name IBM Storwize V7000
odx off
```

---

**Note:** The system topology cannot be reverted to standard if any HyperSwap Volumes are defined.

## Local port masking setting

Before version 7.1, all of the ports of a Storwize V7000 were allowed to carry any type of traffic, that is: host, storage, cluster, and replication traffic. Therefore, the only way to restrict the port usage was through SAN zoning. Storwize V7000 version 7.1 introduced the *local and remote port masking* feature so that you can enable node-to-node communication, and remote copy traffic, only on specific ports. This feature greatly simplifies the SAN zoning activity.

The active-active Metro Mirror function uses only the node-to-node enabled ports to send and receive replication traffic. By default, the local and remote port masking is disabled. Therefore, all of the ports are node-to-node and replication is enabled. For the HyperSwap configuration, we strongly recommend that you dedicate ports for the node-to-node communication by enabling the local port masking feature. In the lab configuration, where the Public and Private SANs are used, the node-to-node communication must be allowed in the *Private SANs* only.

Example 4-28 shows how to set the local port masking for a Storwize V7000 system by using the **chsystem** command.





**Note:** The available bandwidth for synchronization is expressed as a percentage of the total available bandwidth between the sites by using the parameter **backgroundcopyrate**. In Example 4-30, we set the available bandwidth to 50% of 4,000 Mbps, which is 2,000 Mbps (250 MBps).

The system will attempt to synchronize at this rate where possible if any active-active relationships require synchronization (including resynchronization after a copy is offline for a period). This behavior is true no matter how much new host write data is submitted that requires replication between sites. Therefore, be careful not to configure the synchronization rate so high that this synchronization bandwidth consumption affects the amount of bandwidth that is needed for host writes.

Additionally, in a Storwize V7000 HyperSwap configuration, the active-active Metro Mirror traffic uses the node-to-node communication ports. Because this communication is used for the cluster internode traffic and heartbeating also, oversizing the synchronization bandwidth might cause cluster issues, such as node asserts or split brain.

The second attribute that affects the synchronization bandwidth is the *relationship bandwidth limit*. This parameter sets the maximum rate for each active-active relationship to synchronize. The default setting is 25 MBps, and the valid range is 1 - 1,000 MBps. (This unit of measure is *megabytes*, not the megabits of the partnership configuration). The relationship bandwidth limit is a system-wide parameter that you can modify by using the **chsystem** command, as shown in Example 4-31.

*Example 4-31 chsystem command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chsystem -relationshipbandwidthlimit 50
```

---

The relationship bandwidth limit setting is shown in the output of the **lssystem** command, as shown in Example 4-32.

*Example 4-32 lssystem command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lssystem
id 00000100216001E0
name ITS0_V7K_HyperSwap
location local
.
multiple lines omitted
.
relationship_bandwidth_limit 50
tier ssd
tier_capacity 0.00MB
tier_free_capacity 0.00MB
.
multiple lines omitted
.
```

---

## 4.4.2 Configuring the HyperSwap system topology using the GUI

Spectrum Virtualize version 7.6 introduced GUI support for the non standard topology systems (that is Enhanced Stretched Cluster and HyperSwap topologies).

To set up the HyperSwap configuration, go the **Monitoring** → **System** window, select **Monitoring** → **System Topology**, as shown in Figure 4-11.



Figure 4-11 Modify System Topology

The Modify System Topology wizard opens, as depicted in Figure 4-12 on page 82.

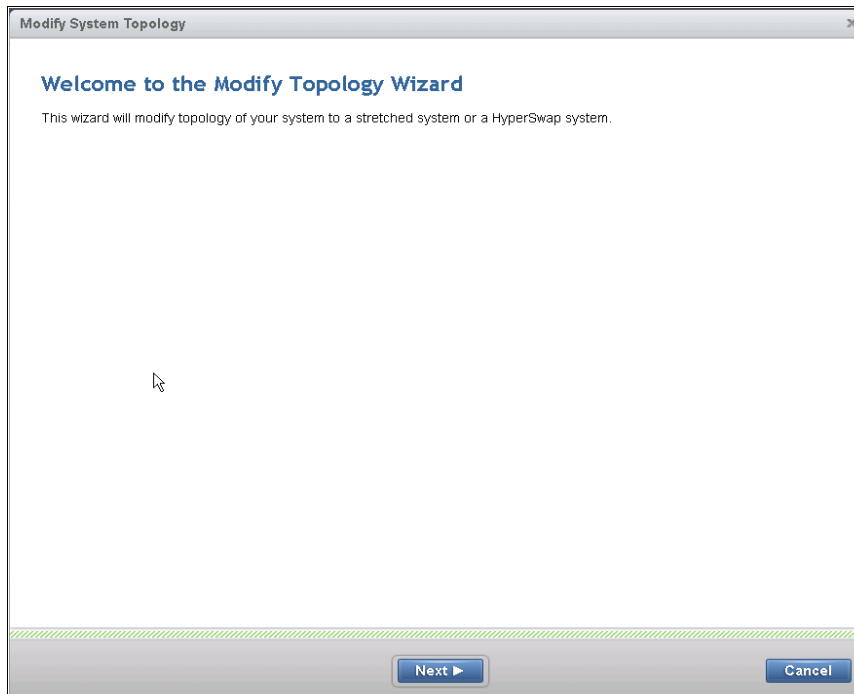


Figure 4-12 The Modify System Topology main windows

Click **Next** to start the configuration. In the *Assign Site Names* window you can specify the site names, as shown in Figure 4-13 on page 83.

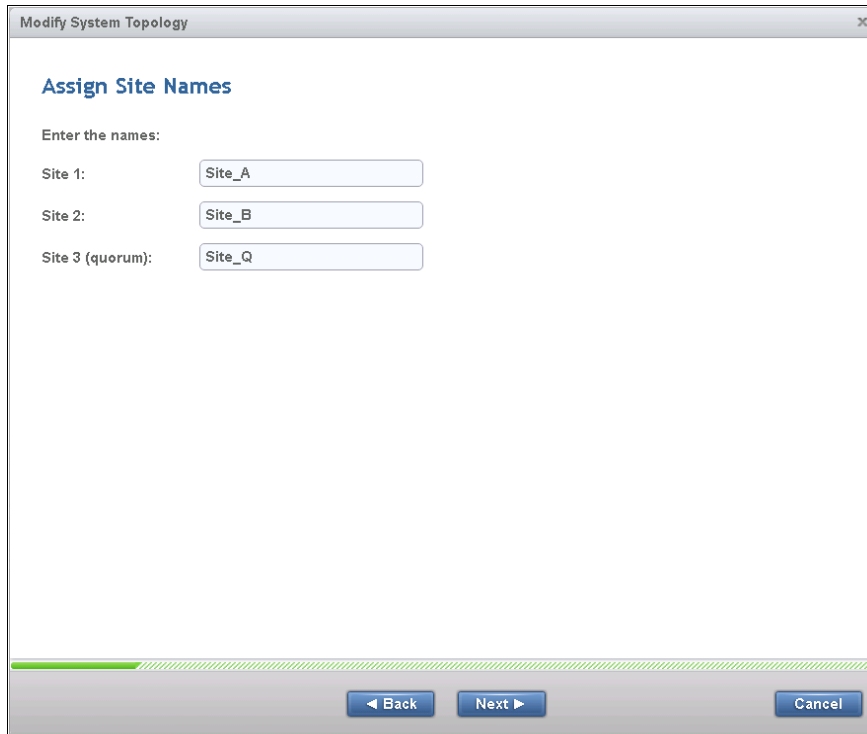


Figure 4-13 The Assign Site Names window

Click **Next** to go to the *Assign Nodes* window. Check if the node site assignment is consistent with the physical location, or eventually use the swap button to change the node site, as shown in Figure 4-14 on page 84.

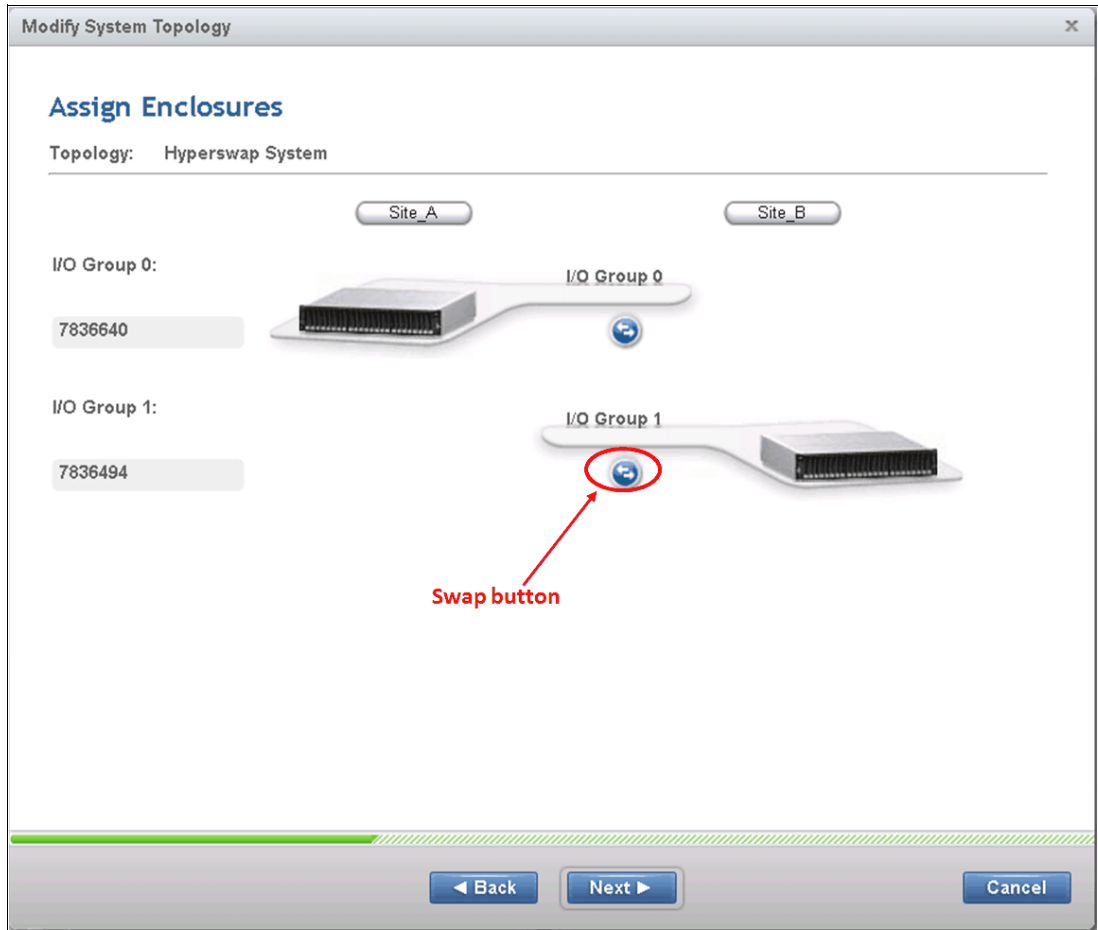


Figure 4-14 Assign Nodes window

Click **Next** to go to *Assign Hosts to a Site* window. In this panel the lists of the existing Hosts is presented. To change the host site, select the host, right-click and select **Modify Site**, as shown in Figure 4-15 on page 85.

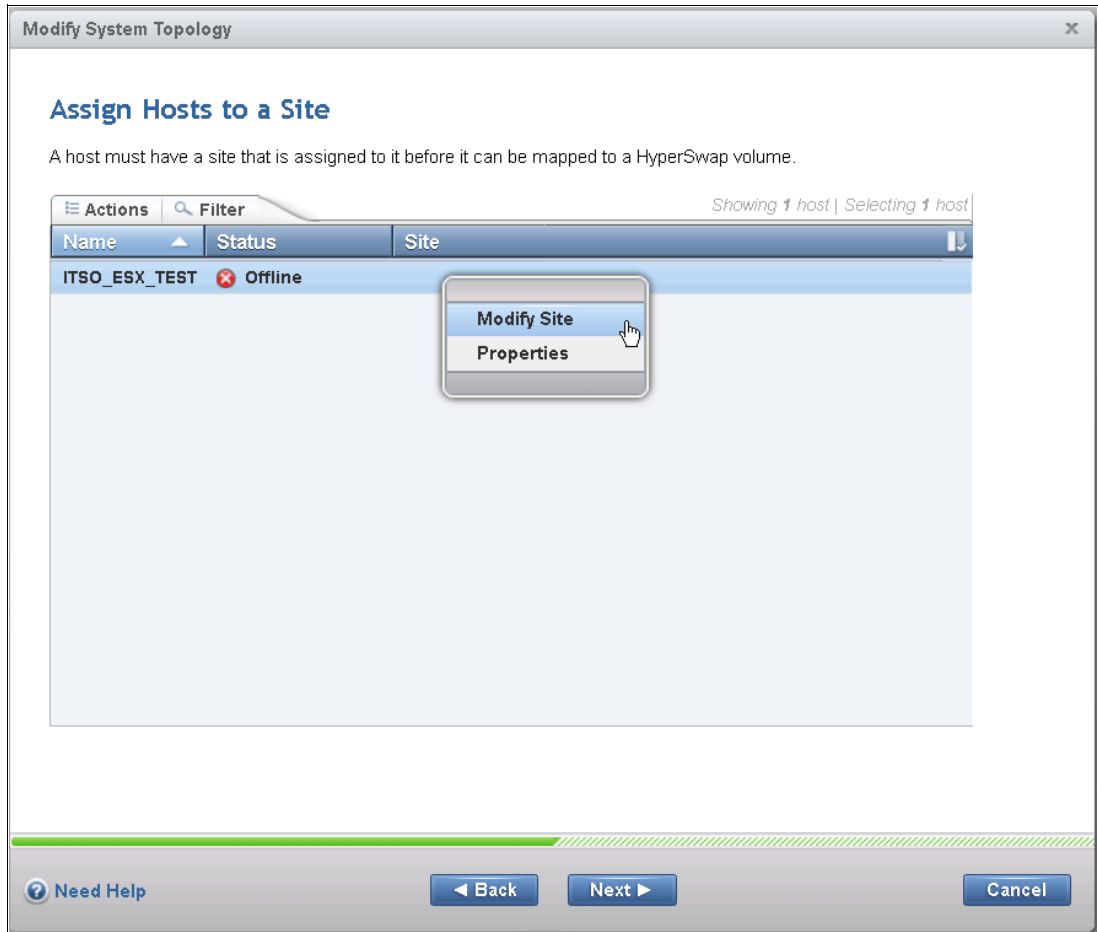


Figure 4-15 The Assign Hosts to a Site window

The site selection window opens. Select the host site from the drop down menu, as shown in Figure 4-16.

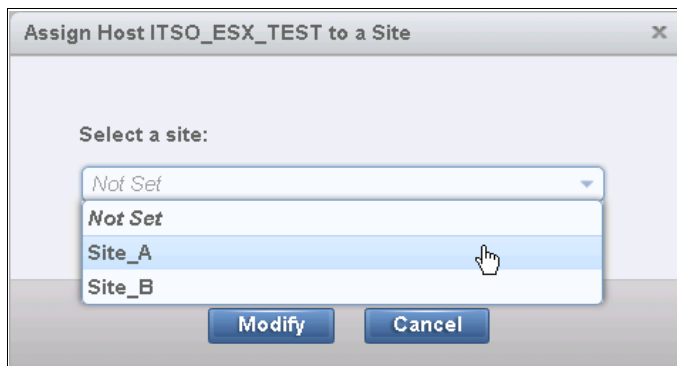


Figure 4-16 Site selection menu for hosts

When the site settings for all the hosts is completed, click **Next**. If there is any external controller connected to the V7000, the *Assign External Storage Systems to Sites* window opens. To change the controller site, select the controller, right-click and select **Modify Site**, as shown in Figure 4-17.

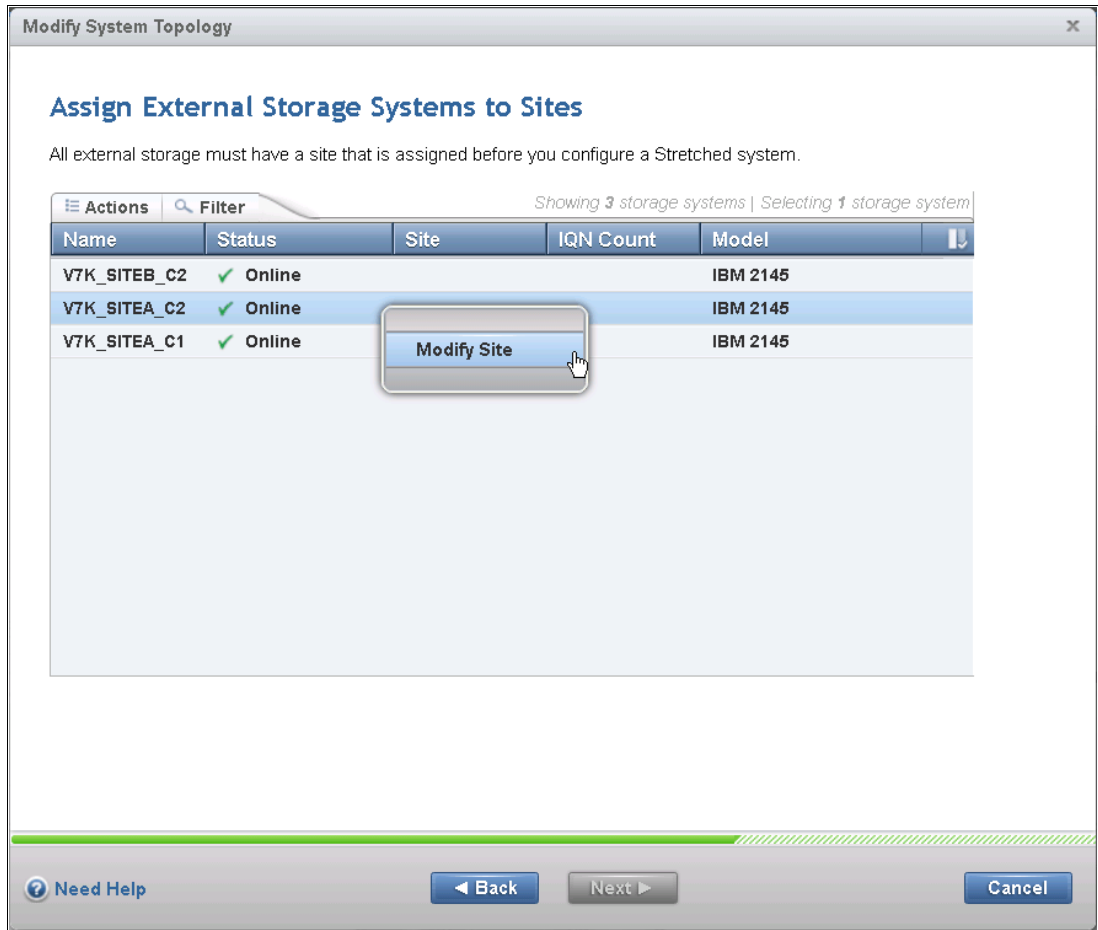


Figure 4-17 The Assign External Storage Systems to Sites window

The site selection window opens. Select the controller site from the drop down menu, as shown in Figure 4-18 on page 86.

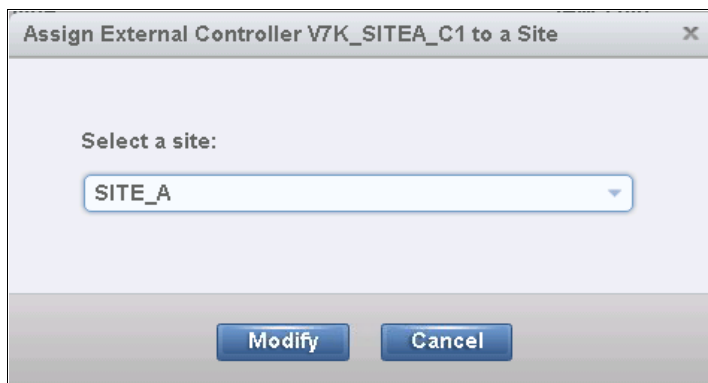


Figure 4-18 Site selection menu for controllers

When the site settings for all the external storage is completed, click **Next**. The *Set Bandwidth Between Sites* window opens, as shown in Figure 4-20 on page 87. In this window set the bandwidth and the background copy rate for the Active-Active partnership.

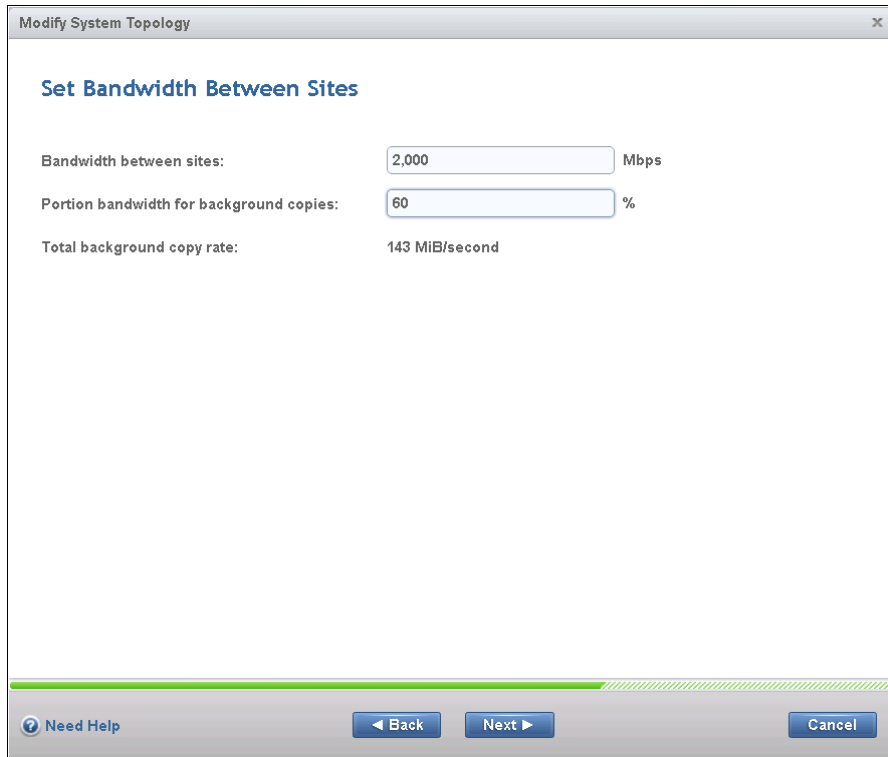


Figure 4-19 The Set Bandwidth Between Sites window

When the bandwidth settings is completed, click **Next**. The Summary window opens, as shown in Figure 4-20 on page 87. In this window all the configuration settings are summarized. Click **Finish** to create the configuration.

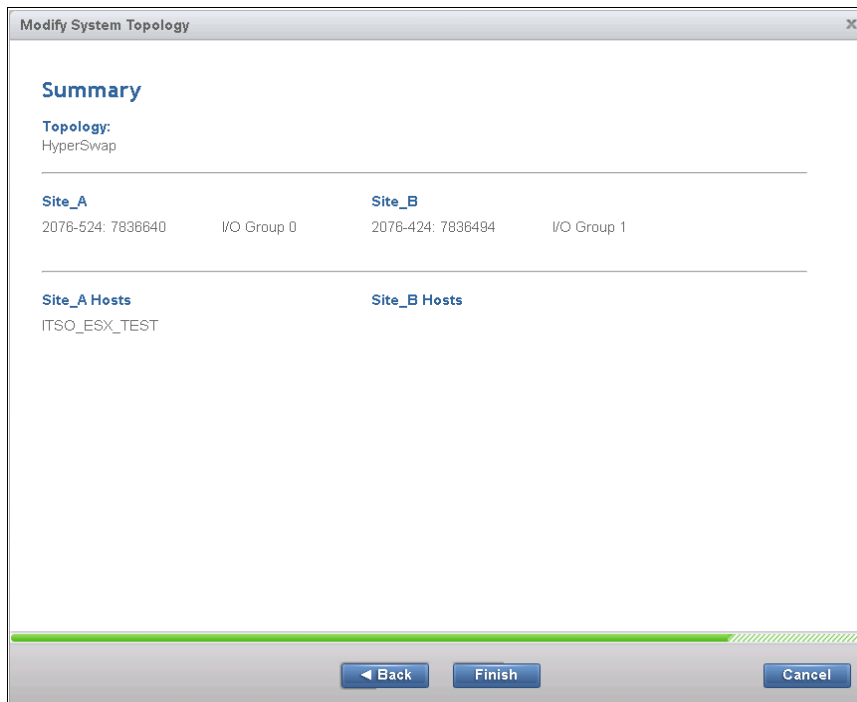


Figure 4-20 The Modify System Topology summary window.

When the HyperSwap topology is enabled, the GUI shows the new topology as depicted in Figure 4-21.

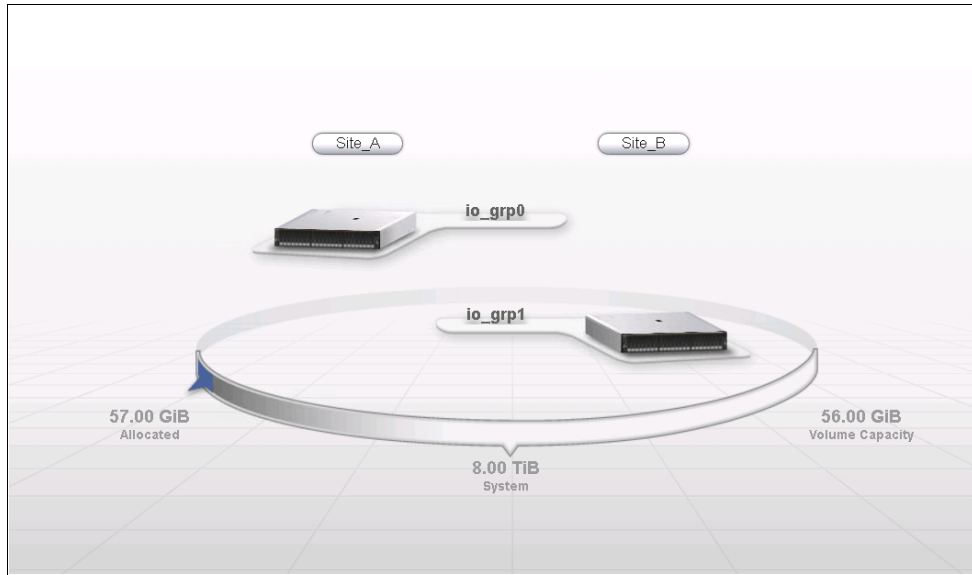


Figure 4-21 GUI that shows the system in the HyperSwap topology

### 4.4.3 Configuring the HyperSwap volumes using the CLI

In the following section, we describe the steps to manage HyperSwap volumes using the CLI in a Storwize V7000 HyperSwap configuration.

#### HyperSwap Volumes creation

Storwize V7000 version 7.6 introduced the `mkvolume` CLI command to create the HyperSwap volumes. This command automatically creates all the components of a HyperSwap volume, that is:

- ▶ Master and Auxiliary volumes
- ▶ Master and Auxiliary Change volumes
- ▶ FlashCopy mappings for Master and Auxiliary volumes to their change volumes
- ▶ Active-Active Metro Mirror relationship

The syntax of the `mkvolume` command is shown in Example 4-33.

Example 4-33 `mkvolume` command

```
>>- mkvolume -- --+-----+-- -- size -- disk_size ---->
                '- name -- name -'

>-- --+-----+-- ----->
        '- -unit --- b ---'
            +- kb -+
            +- mb -+
            +- gb -+
            +- tb -+
            '- pb -'

>--+-----+-- ----->
        '- -iogrp --- iogroup_id ---+-'
```



```

        '- iogroup_name -'

>-- -pool --+- storage_pool_id ----+-- -- ----->
        '- storage_pool_name -'

>--+-----+-- --+-----+----->
        '- -cache --+- none -----+'      '-+- -thin -----+-'
            +- readonly --+                '- -compressed -'
            '- readwrite -'

>--+-----+-- ----->
        '- -buffersize --+- buffer_size -----+-'
            '- buffer_percentage% -'

>--+-----+-- ----->
        '- -warning --+- warning_capacity ----+-'
            '- warning_percentage% -'

>--+-----+-- --+-----+----->
        '- -noautoexpand -'      '- -grainsize --+- 32 --+-'
                                   +- 64 --+
                                   +- 128 +-
                                   '- 256 -'

>--+-----+-- -----><
        '- -udid -- udid -'

```

Example 4-34 shows a **mkvolume** sample.

*Example 4-34 mkvolume command example*

---

```

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkvolume -name V_ESX_A_HyperSwap -pool
1:0 -size 200 -unit gb
Volume, id [0], successfully created

```

---

The **mkvolume** command in Example 4-34 on page 89 creates four volumes, four FlashCopy mappings and an Active-Active Metro Mirror relationship, as shown in Example 4-35. The volume ID of a HyperSwap volume corresponds to the volume ID of the Master volume.

*Example 4-35 Creation example*

---

```

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsvdisk -filtervalue
volume_name=V_ESX_A_HyperSwap
id name IO_group_id IO_group_name status mdisk_grp_id mdisk_grp_name capacity
type FC_id FC_name RC_id RC_name vdisk_UID fc_map_count copy_count
fast_write_state se_copy_count RC_change compressed_copy_count parent_mdisk_grp_id
parent_mdisk_grp_name formatting encrypt volume_id volume_name function
0 V_ESX_A_HyperSwap 0 io_grp0 online 1 Pool_SITE_A 200.00GB striped many many 0
rcrcl0 6005076400878007800000000000066 2 1 not_empty 0 no 0 1 Pool_SITE_A yes no
0 V_ESX_A_HyperSwap master
1 vdisk1 1 io_grp1 offline 0 Pool_SITE_B 200.00GB striped many many 0 rcrcl0
6005076400878007800000000000067 2 1 not_empty 0 no 0 0 Pool_SITE_B yes no 0
V_ESX_A_HyperSwap aux
2 vdisk2 0 io_grp0 online 1 Pool_SITE_A 200.00GB striped many many 0 rcrcl0
6005076400878007800000000000068 2 1 not_empty 1 yes 0 1 Pool_SITE_A no no 0
V_ESX_A_HyperSwap master_change

```

```
3 vdisk3 1 io_grp1 online 0 Pool_SITE_B 200.00GB striped many many 0 rcre10
60050764008780078000000000000069 2 1 empty 1 yes 0 0 Pool_SITE_B no no 0
V_ESX_A_HyperSwap aux_change
```

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsrcrelationship -filtervalue
master_vdisk_name=V_ESX_A_HyperSwap
id name master_cluster_id master_cluster_name master_vdisk_id master_vdisk_name
aux_cluster_id aux_cluster_name aux_vdisk_id aux_vdisk_name primary
consistency_group_id consistency_group_name state
bg_copy_priority progress copy_type cycling_mode freeze_time
0 rcre10 000001002160020E ITS0_V7K_HyperSwap 0 V_ESX_A_HyperSwap
000001002160020E ITS0_V7K_HyperSwap 1 vdisk1 master consistent_synchronized 50
activeactive
```

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsfcmap
id name source_vdisk_id source_vdisk_name target_vdisk_id target_vdisk_name
group_id group_name status progress copy_rate clean_progress incremental
partner_FC_id partner_FC_name restoring start_time rc_controlled
0 fcmap0 0 V_ESX_A_HyperSwap 2 vdisk2 idle_or_copied 0 0 100 off 1 fcmap1 no yes
1 fcmap1 2 vdisk2 0 V_ESX_A_HyperSwap idle_or_copied 0 50 100 off 0 fcmap0 no yes
2 fcmap2 1 vdisk1 3 vdisk3 idle_or_copied 0 0 100 off 3 fcmap3 no yes
3 fcmap3 3 vdisk3 1 vdisk1 idle_or_copied 0 50 100 off 2 fcmap2 no yes
```

As shown in Example 4-35 on page 89 three more attributes have been added to the vdisk information related to the HyperSwap volumes: *volume\_id*, the *volume\_name* and *function*. The *volume\_id* and *volume\_name* attributes refer respectively to the ID and the name of the HyperSwap volume. The *function* attribute refers to the specific role of the vdisk in reference to the HyperSwap volume (master, aux, master\_change, aux\_change).

### HyperSwap Volumes deletion

To delete a HyperSwap volume, the **rmvolume** command can be used. This command removes all the HyperSwap volume related objects. A sample of the **rmvolume** command is reported in Example 4-36.

*Example 4-36 rmvolume command sample*

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>rmvolume V_ESX_A_HyperSwap
```

### HyperSwap Volumes legacy mode creation using the CLI

The **mkvolume** command greatly simplifies the HyperSwap volumes deployment, creating multiple Storwize V7000 objects with a single command. However, there are some cases where the **mkvolume** command cannot be used. For instance, when a volume is initially created as a non HyperSwap volume and later it must be changed to a HyperSwap volume. In this case, most of the HyperSwap volume objects must be manually created (with the exception of the Master Volume). Also Storwize V7000 versions prior to 7.6 do not support the **mkvolume** command.

In the following sections the CLI commands for the creation of all the HyperSwap volume objects are described.

## Master Volume

The *Master Volume* in an active-active relationship is the volume that holds the application data. The Master Volume can be an existing volume or a new volume. To create a Master Volume, you can use the `mkvdisk` command, as shown in Example 4-37.

### Example 4-37 `mkvdisk` command

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkvdisk -name V_ESX_A_Master -iogrp
io_grp0_SITE_A -mdiskgrp Pool_SITE_A -size 200 -unit gb -accessiogrp 0:1
Virtual Disk, id [0], successfully created
```

---

The Master Volume holds the initial copy of the data, which is then replicated to the other site. For a completely new Master Volume, it does not matter which site this volume is created on.

**Important:** The site of the storage pool that provides the capacity for the volume must match the site of the caching I/O Group. This requirement also applies if you use an existing volume.

The `-accessiogrp` parameter is important, because it allows the Master Volume to be accessed on both sites. Specify the caching I/O Groups that you will use for the Auxiliary Volume and the caching I/O Groups that you specified for the Master Volume.

If you are using an existing Master Volume, it normally has access only through its own I/O Group. To allow access to the Master Volume through both sites, you will need to add access to the Master Volume through the Auxiliary Volumes I/O Group, too. Use the `addvdiskaccess` command as shown in Example 4-38.

### Example 4-38 `addvdiskaccess` command

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>addvdiskaccess -iogrp 1 V_ESX_A_Master1
```

---

This part of the process is not policed, but it must be completed for the HyperSwap Volume to provide HA through nodes on both sites. This step is performed for the Master Volume only.

## Auxiliary Volume

Next, whether or not an existing or new Master Volume is used, we need an Auxiliary Volume to create the active-active relationship. This Auxiliary Volume must be the same size as the Master Volume. However, the Auxiliary Volume uses storage from the other site, and the Auxiliary Volume is in an I/O Group on the other site. Similar to the Master Volume, you can use the `mkvdisk` command to create the Auxiliary Volume, as shown in Example 4-39.

### Example 4-39 `mkvdisk` command

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkvdisk -name V_ESX_A_Aux -iogrp
io_grp1_SITE_B -mdiskgrp Pool_SITE_B -size 200 -unit gb
Virtual Disk, id [1], successfully created
```

---

An Auxiliary Volume must not be mapped to any hosts. You do not need to allow access to an Auxiliary Volume to the Master I/O caching group.

You can use an Auxiliary Volume of a different provisioning type than the Master Volume. For example, you can mix a fully allocated volume with a thin-provisioned volume, or you can mix compressed and non-compressed thin-provisioned volumes. However, we do not recommend this configuration.

Normally, the Master and Auxiliary Volumes need to be on storage that performs similarly. If similarly performing storage is not possible, write performance will be dictated by the slower of the two types of storage, and read performance will be the read performance of the volume that currently acts as the Master of the HyperSwap Volume.

### **Change Volumes**

Two thin-provisioned volumes are also required to act as CVs for an active-active relationship. These CVs must be the same logical size as the Master Volume. One CV, the *Master CV*, is created in the same I/O Group as the Master Volume and in a storage pool in the same site (although not necessarily the same storage pool as the Master Volume). The *Auxiliary CV* is created in the Auxiliary Volume's I/O Group and in a storage pool in the same site.

Example 4-40 shows the `mkvdisk` command to create the CVs.

#### *Example 4-40 mkvdisk command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkvdisk -name V_ESX_A_Master_CV -size
200 -unit gb -iogrp io_grp0_SITE_A -mdiskgrp Pool_SITE_A -rsize 0% -autoexpand
Virtual Disk, id [3], successfully created
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkvdisk -name V_ESX_A_Aux_CV -size 200
-unit gb -iogrp io_grp0_SITE_B -mdiskgrp Pool_SITE_B -rsize 0% -autoexpand
Virtual Disk, id [4], successfully created
```

---

A CV must not be mapped to any hosts.

The CVs are used to enable automatic resynchronization after a link outage or other fault causes replication to stop. They are used to store differences between the copies while resynchronizing the active-active relationship, and they normally require only enough storage performance to satisfy the resynchronization rate. If access is enabled to the stale copy during resynchronization, a portion of host reads and writes will be serviced by the CV storage, but this servicing will decrease toward zero within a short period of time. Therefore, CVs can be stored on storage with lower performance than the Master and Auxiliary Volume storage performance without significantly affecting host I/O performance.

Typically, the CVs will consume capacity that is equal to the initially specified `rsize`. During resynchronization, the CV at the stale copy will grow because it retains the necessary data to revert to the stale image. It will grow to consume the same amount of storage as the quantity of changes between the two copies. Therefore, if a stale copy needs 20% of its data changed to be synchronized with the up-to-date copy, its CV will grow to consume 20% of its logical size. After resynchronization, the CV will automatically shrink back to the initially specified `rsize`.

**Note:** Use the `-autoexpand` option when you create the CVs. This option will prevent the CV from going offline if the capacity that is used during the resynchronization exceeds the specified `rsize`.

### **Active-active Metro Mirror relationship**

The next step is to create the active-active Metro Mirror relationship. Active-active relationships are a special relationship that can be used only in a HyperSwap configuration. To create an active-active Metro Mirror relationship, you can use the `mkrcrelationship` command, as shown in Example 4-41.

#### *Example 4-41 mkrcrelationship command with the active-active option*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkrcrelationship -master V_ESX_A_Master
-aux V_ESX_A_Aux -cluster ITS0_V7K_HyperSwap -activeactive -name Re1_ESX_A
```

---

RC Relationship, id [0], successfully created

---

If the Master Volume is not written to yet, you can add the `-sync` parameter in the `mkrcrelationship` command to avoid the initial synchronization process.

**Note:** Do not use the `-nofmtdisk` parameter of the `mkvdisk` command to disable the quick initialization of fully allocated volume data for HyperSwap Volumes. If it is necessary, ensure that the `-sync` parameter of the `mkrcrelationship` command is omitted so that the system fully synchronizes the two copies, even if neither copy was written to.

Now, the Auxiliary Volume goes offline. From now on, it will only be accessed internally by the HyperSwap function. The Master Volume remains online.

To complete the active-active relationship, the two CVs must be added. Example 4-42 shows how to add the CVs to the active-active relationship by using the `chrcrelationship` command.

*Example 4-42 chrcrelationship command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chrcrelationship -masterchange
V_ESX_A_Master_CV Re1_ESX_A
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chrcrelationship -auxchange
V_ESX_A_Aux_CV Re1_ESX_A
```

---

At this point, the active-active relationship starts to replicate automatically. If the relationship was created without the `-sync` flag, it will perform a full synchronization of the existing data from the Master Volume to the Auxiliary Volume. This initial synchronization process does not use the CVs.

## Consistency Groups

In HyperSwap configurations where the application spans multiple volumes, you are required to group all of the related active-active relationships in *Consistency Groups*. This grouping allows all volumes for a specific application to fail over together, ensuring that at least one site has an up-to-date copy of every volume for the application.

To create a Consistency Group, you can use the `mkrcconsistgrp` command, as shown in Example 4-43.

*Example 4-43 mkrcconsistgrp command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>mkrcconsistgrp -name CG_ESX_A
RC Consistency Group, id [2], successfully created
```

---

To add an existing relationship to a Consistency Group, you can use the `chrcrelationship` command, as shown in Example 4-44.

*Example 4-44 chrcrelationship command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>chrcrelationship -consistgrp CG_ESX_A
Re1_ESX_A
```

---

Or, when the relationship is created, use the `-consistgrp` parameter directly in the `mkrcrelationship` command (Example 4-41 on page 92).

The following restrictions apply when you add a relationship to a Consistency Group:

- ▶ The active-active relationship's state attribute must match the active-active Consistency Group's state attribute.
- ▶ If the state is not `consistent_synchronized`, the site of the volume that acts as the Primary copy of the active-active relationship must be the same as the site of the volumes that act as the Primary copies of the relationships in the active-active Consistency Group.
- ▶ If the state is `consistent_synchronized`, and the site of the Primary volume of the active-active relationship is not the same as the Primary site of the Consistency Group, the relationship direction is switched as the relationship is added so that the Primary site matches.
- ▶ If the site of the Master Volume of the active-active relationship does not match the site of the Master Volumes of the relationships in the Consistency Group, the roles of the Master and Auxiliary Volumes in the active-active relationship are swapped. Host access will continue to be provided through the same volume ID and host maps, which will now be the Auxiliary Volume of the relationship. The relationship ID will be retained even though this ID will now match the Auxiliary Volume ID. The Master and Auxiliary roles will be restored if the relationship is removed from the Consistency Group.

## FlashCopy

You can use FlashCopy to take point-in-time copies of HyperSwap Volumes to use for testing, cloning, and backup.

A FlashCopy map with a HyperSwap Volume as its source might not cross sites. A FlashCopy mapping where the target volume is on site 1 must use the volume of the HyperSwap Volume on site 1 as its source, and likewise for site 2. It is not possible for a FlashCopy map with a HyperSwap Volume as its source to copy data between sites.

These two FlashCopy maps can both be used independently to take point-in-time copies of the HyperSwap Volume on the two sites. Although the system provides no coordination of these maps, you can always create a FlashCopy Consistency Group that contains both FlashCopy mappings to run the FlashCopy simultaneously.

To trigger the FlashCopy map, the copy of the HyperSwap Volume on the same site as the FlashCopy target volume must be one of the following copies:

- ▶ A Primary copy of an active-active relationship in any state
- ▶ A Secondary copy of an active-active relationship in the `consistent_synchronized` state.

If access was enabled to an old but consistent copy of the HyperSwap Volume, a FlashCopy map can only be triggered on the site that contains that copy.

A FlashCopy map cannot be created with a HyperSwap Volume as its target. For this reason, a FlashCopy reverse function is not allowed with HyperSwap Volumes. If necessary, delete the active-active relationship to convert the HyperSwap Volume to a regular volume before you create and trigger the FlashCopy map.

**Note:** IBM Tivoli Storage FlashCopy Manager does not support HyperSwap Volumes.

### 4.4.4 Configuring the HyperSwap volumes using the GUI

Storwize V7000 version 7.6 introduced GUI support for the HyperSwap volume management. In this section we describe how to create and delete a HyperSwap volume through the GUI.

## HyperSwap Volume creation

To create a HyperSwap volume, go to **Monitoring** → **System** and click on **Create Volumes**, as shown in Figure 4-22.

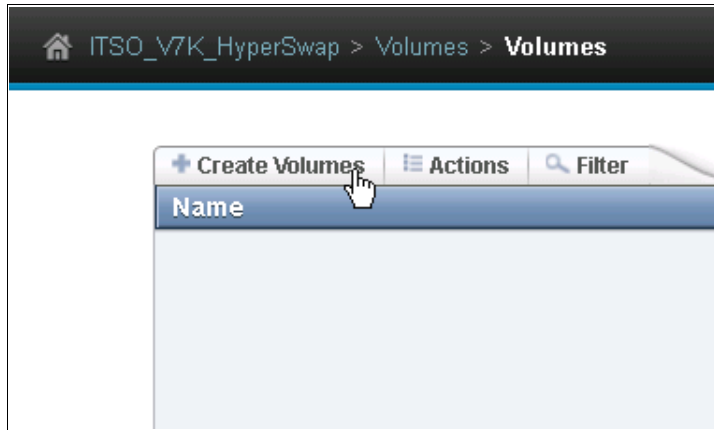


Figure 4-22 Create Volumes

The Create Volumes wizard opens. According to the system configuration, the wizard offers some volume presets. In a HyperSwap configuration, two presets are available, the Basic preset and the HyperSwap preset. Click on **HyperSwap** to create a HyperSwap volume, as shown in Figure 4-23 on page 95.



Figure 4-23 Create Volumes wizard

The Create Volume window expands showing the volume attributes, as shown in Figure 4-24.

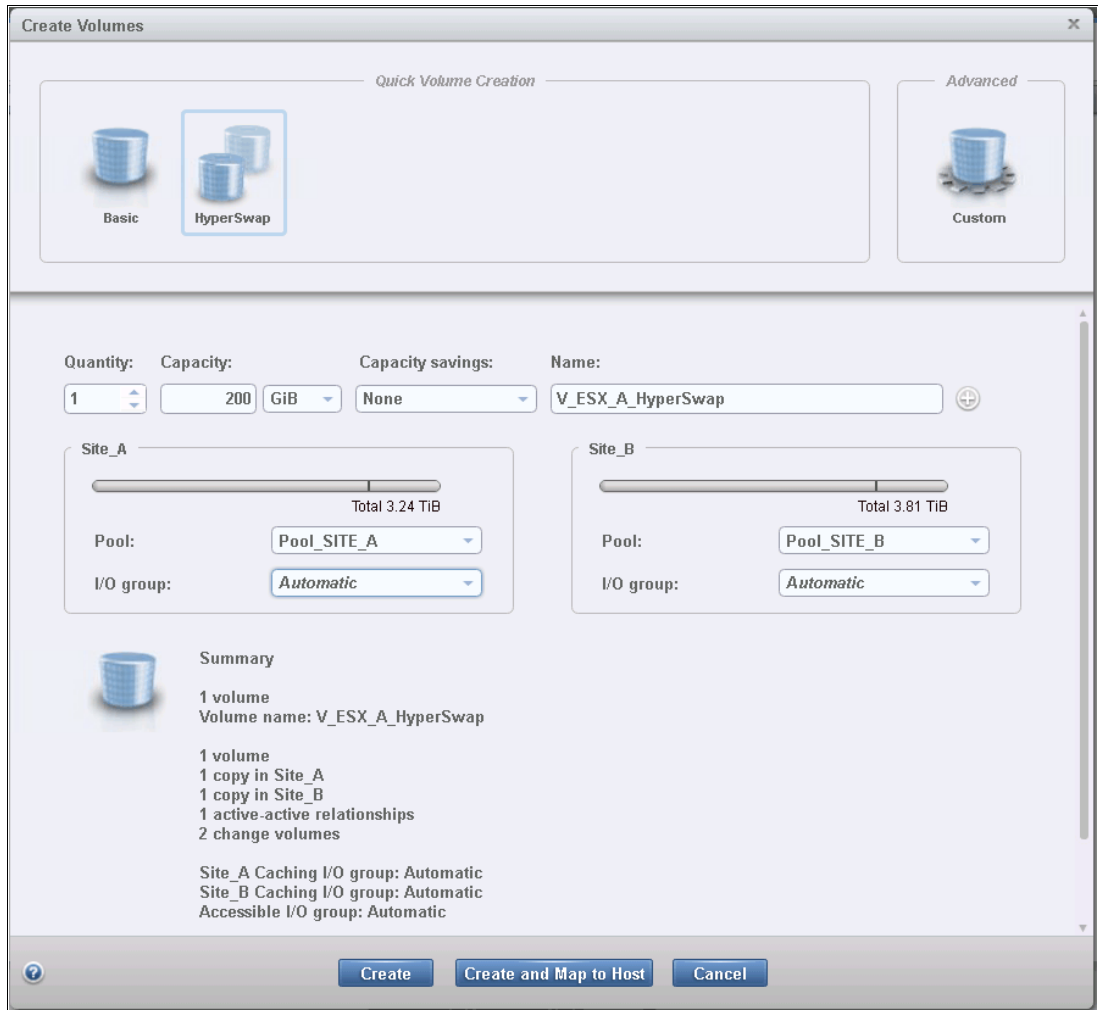


Figure 4-24 Create Volume wizard attributes box

Fill up the volume information and click **Create** to create the volume. Optionally click on **Create and Map to Host** to create the volume and start the host mapping wizard. The results of the volume creation is presented in Figure 4-25 on page 97.



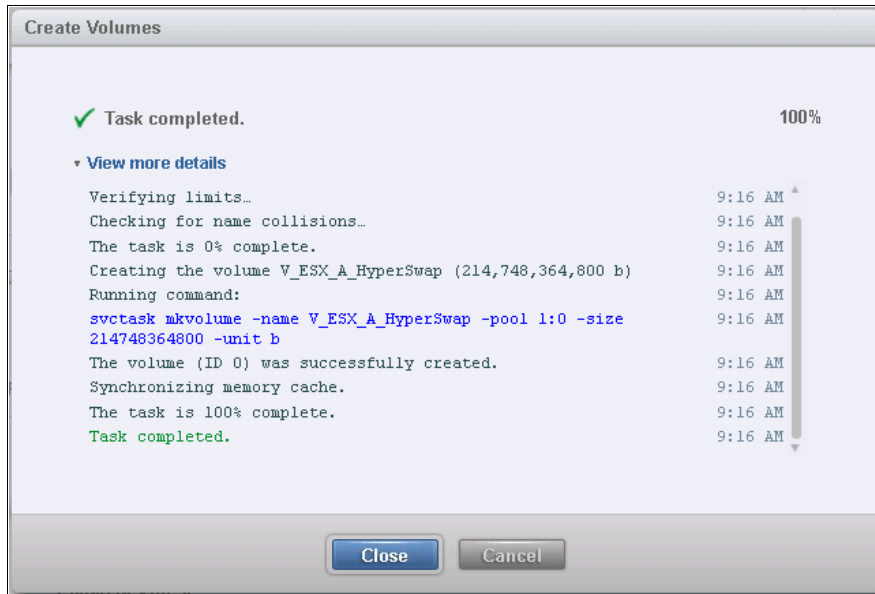


Figure 4-25 Create Volumes wizard results

**Naming convention:** When a HyperSwap volume is created with the `mkvolume` command, only the Master volume is identified with the HyperSwap volume name. Auxiliary and Change volumes will use standard names (that is `vdiskxx`).

**Volume distribution:** When creating multiple HyperSwap volumes in sequence, the `mkvolume` command distributes the volumes' preferred node evenly across the I/O group nodes. The preferred node for Change volumes are chosen accordingly to the Master and Auxiliary volumes; that is the Master Change volume will have the same preferred node as the Master volume and Auxiliary Change volume will have the same preferred node as the Auxiliary volume.

After the volume creation, the volume menu shows the created volumes, as shown in Figure 4-26 on page 98.



Figure 4-26 HyperSwap volume sample

**Change Volumes:** In the Volumes menu only the Master and Auxiliary volumes are reported. The Change volumes are purposely hidden because they are not available for the normal usage.

In the Copy Services menu the Active-Active relationship as well as the FlashCopy mappings to change volumes are reported, as shown in Figure 4-27 and Figure 4-28 on page 99.

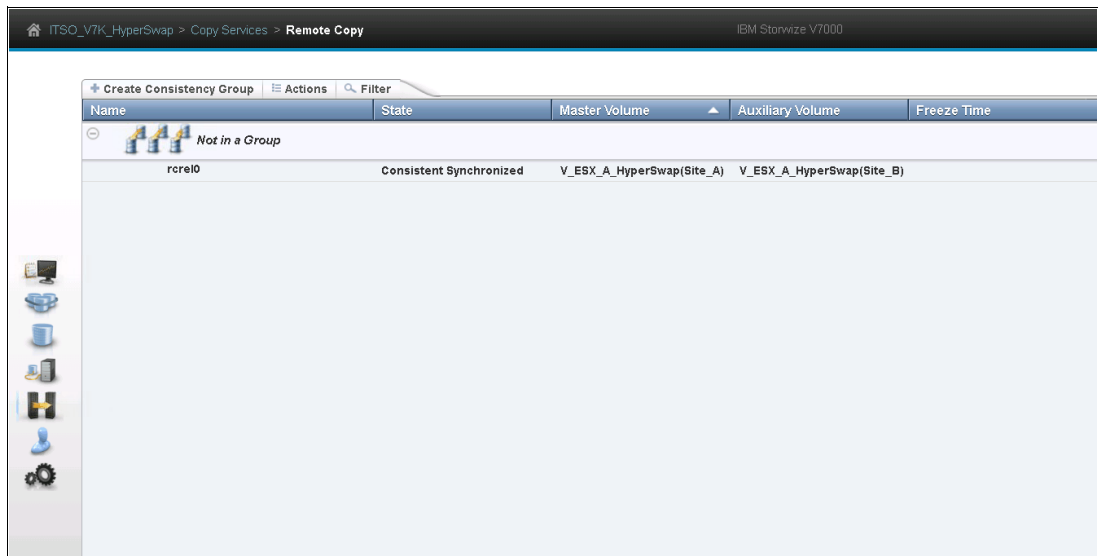


Figure 4-27 HyperSwap volume Active-Active relationship

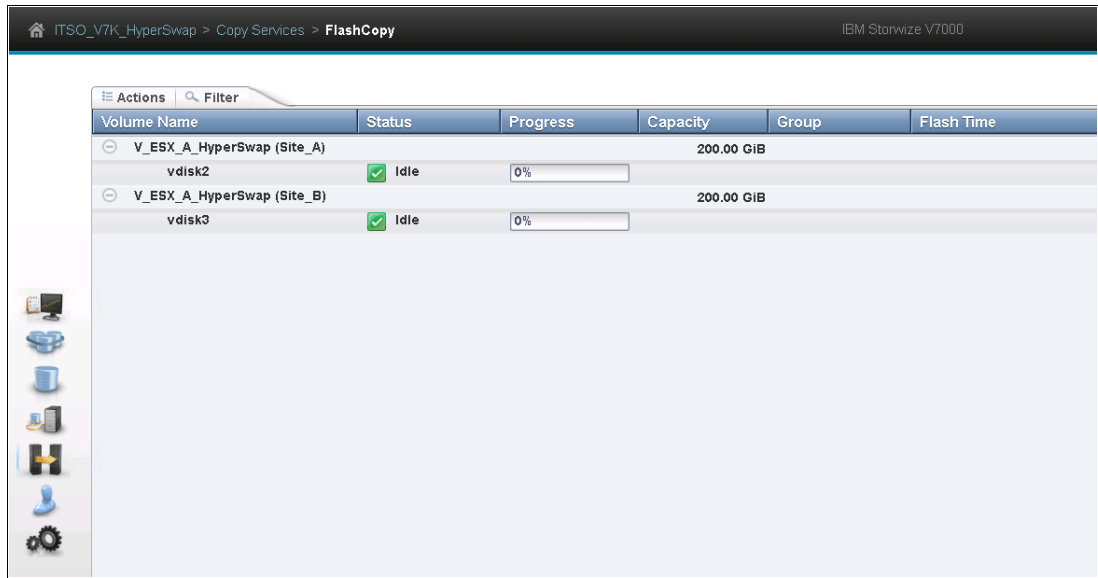


Figure 4-28 HyperSwap volume Change volume mappings

## Consistency Groups

To create a consistency group go to **Copy Services** → **Remote Copy** as shown in Figure 4-29 on page 99.

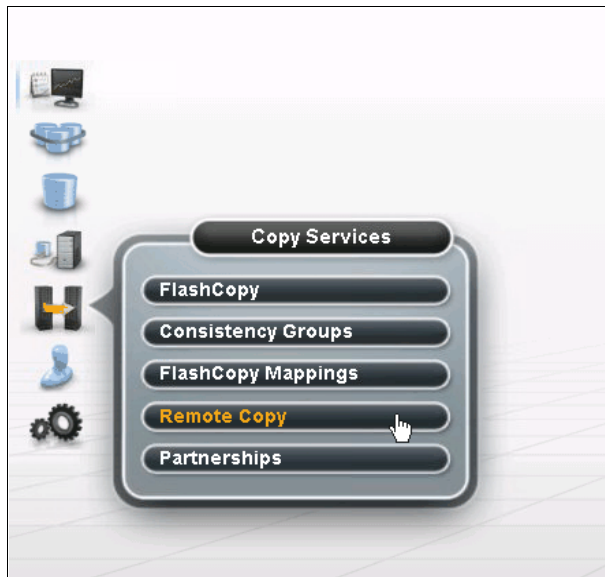


Figure 4-29 Remote Copy menu

On the Copy Services windows, click on **Create Consistency Group**, as shown in Figure 4-30 on page 100.

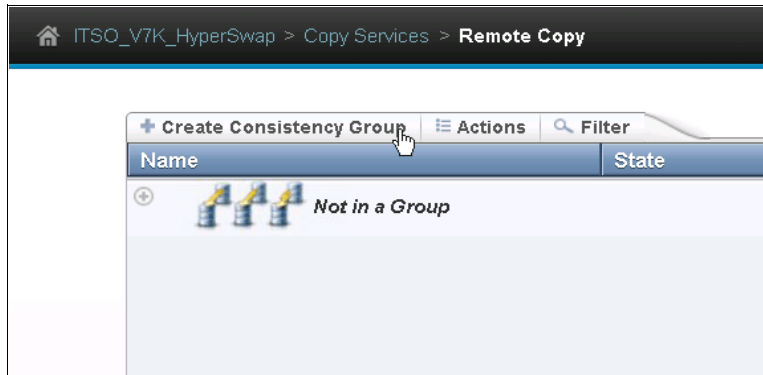


Figure 4-30 Create Consistency Group

The Create Consistency Group wizard opens (see Figure 4-31).

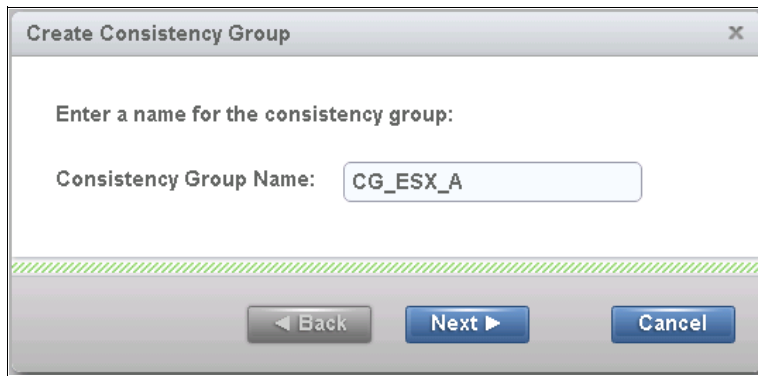


Figure 4-31 Select the consistency group name

Select a consistency group name and click **Next**. On the following box check **On this system** as location for the auxiliary volumes, as shown in Figure 4-32. Click **Next** to proceed.

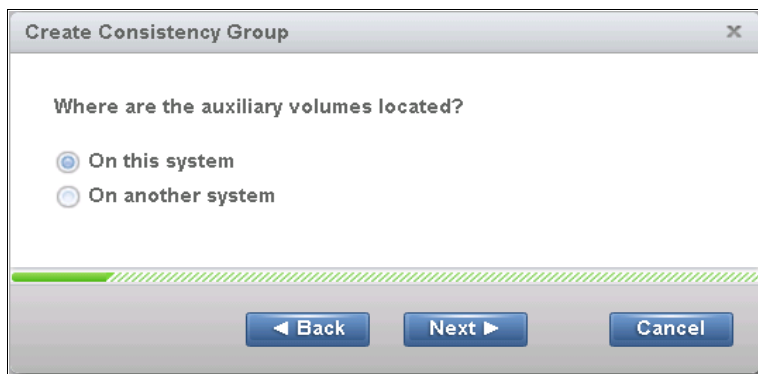


Figure 4-32 Select the auxiliary volume location

Finally, check on **No, create an empty consistency group** to create an empty consistency group, as shown in Figure 4-33 on page 101. Click **Finish** to create the consistency group.

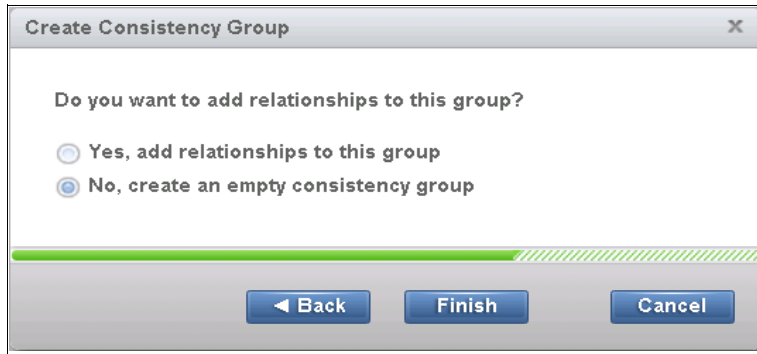


Figure 4-33 Create an empty consistency group

Once the consistency group is created, to move the relationships, expand the **Not in a Group** section and select the relationships to be moved. Then right-click on the selected relationships and select **Add to Consistency Group**, as shown in Figure 4-34.

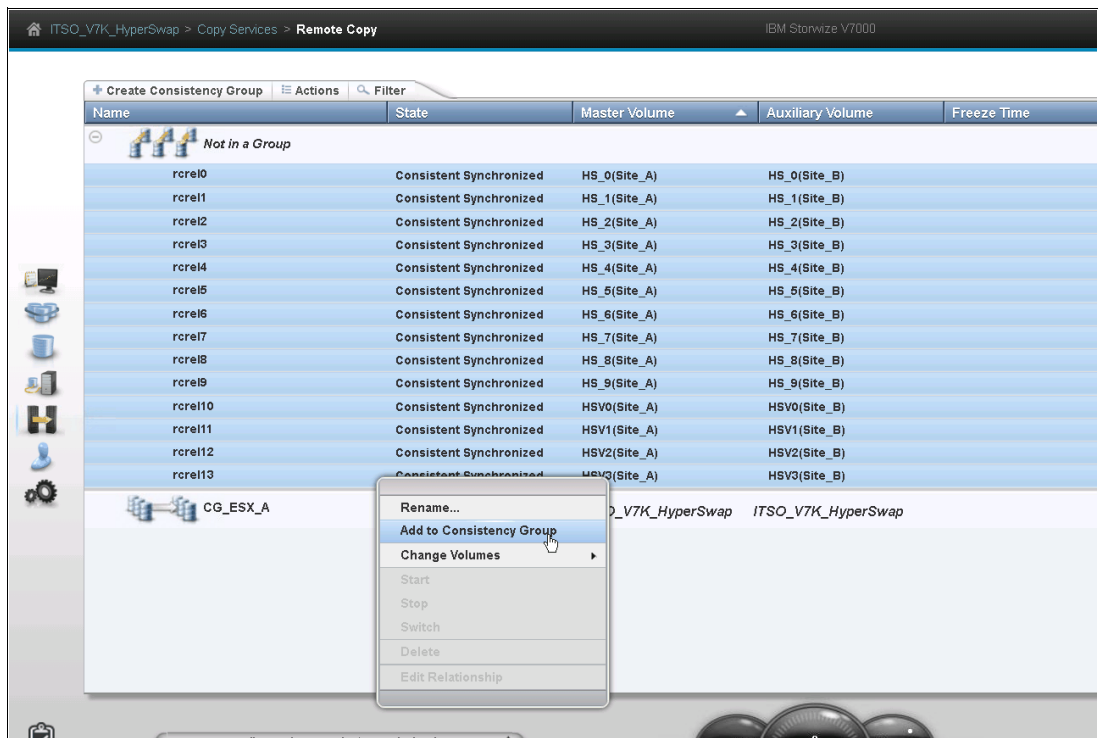


Figure 4-34 Add to Consistency Group

Select the consistency group where the relationships have to be moved and click **Add Relationships to Consistency Group** (see Figure 4-35 on page 102).

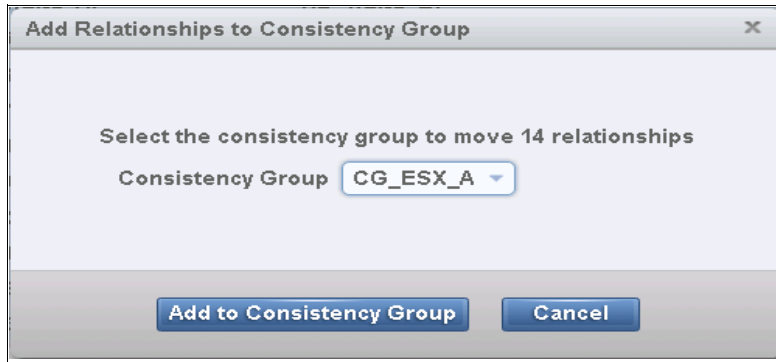


Figure 4-35 Select the consistency group

## HyperSwap Volume deletion

To delete a HyperSwap volume, select the volume from the Volumes menu and then go to **Actions** → **Delete**, as shown in Figure 4-36.

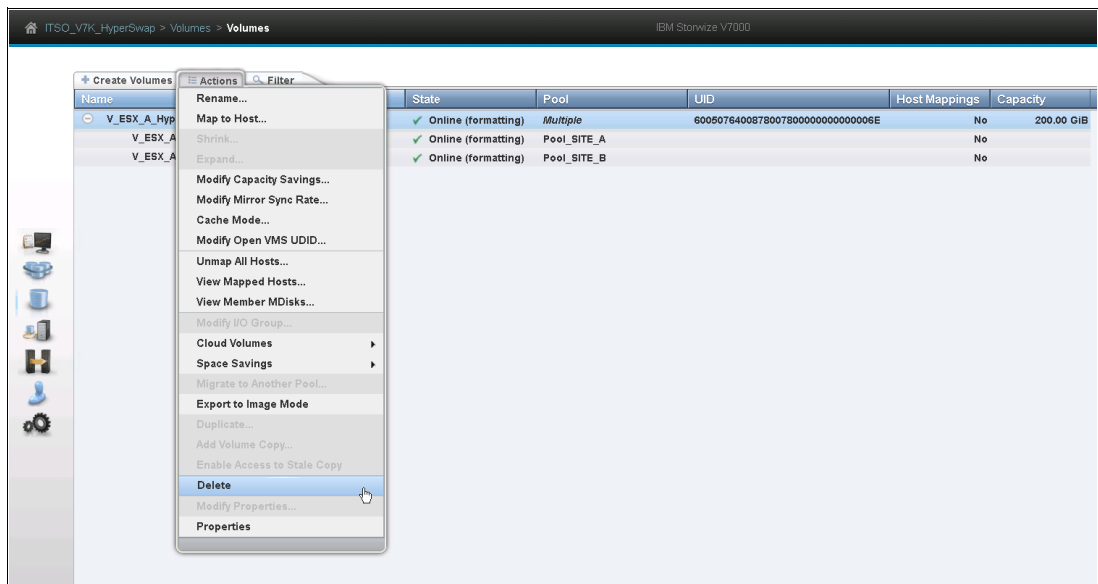


Figure 4-36 Delete a HyperSwap volume

Confirm the number of volumes to be deleted and click **Delete**, as shown in Figure 4-37 on page 103.

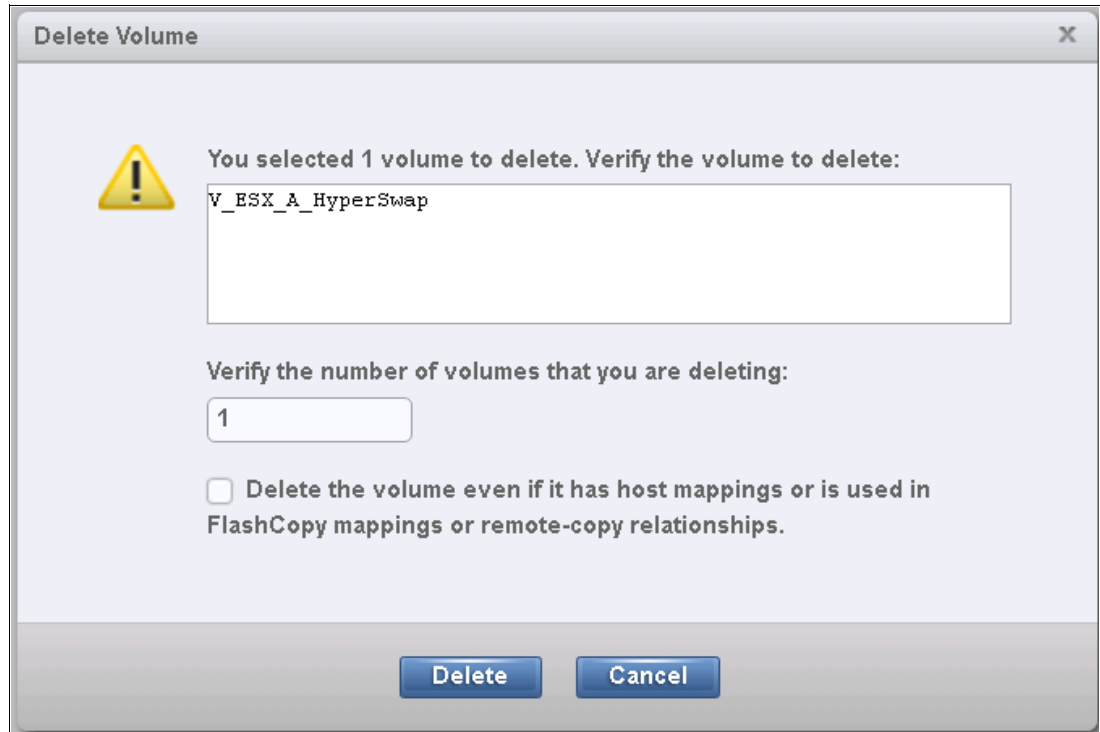


Figure 4-37 Delete volume confirmation panel

This action will remove all the HyperSwap related objects.

## 4.4.5 Summary

A summary of the implementation tasks is described.

### Planning

Specific planning activities must be performed before you implement a Storwize V7000 solution:

- ▶ Bandwidth requirements assessment. This task is important for sizing the node-to-node communication and the SAN extension resources (ISLs and Fibre Channel over IP (FCIP) tunnels). Analyzing the workload to be implemented in the HyperSwap configuration is essential for this assessment.
- ▶ Quorum configuration and placement. The HyperSwap configuration requires a tiebreaker to handle split brain conditions. Either an external storage disk or an IP quorum device can be used for this tiebreaking function. If an external quorum disk is used, all of the requirements in terms of placement and SAN connectivity must be fulfilled. If the IP quorum feature is used, all of the network requirements in terms of TCP ports and IP addressing must be fulfilled.
- ▶ Master and Auxiliary placement. You must define the Master and Auxiliary Volume disks (VDisks) according to the host site definition. It is important to avoid unplanned HyperSwap Volume switching. In clustered environments where HyperSwap Volumes are accessed from the server in both sites, an assessment of the workload might help to identify better placement for Master and Auxiliary Volumes.

## Implementation

Starting with V7000 version 7.6 most of the configuration tasks for a HyperSwap configuration can be performed using the GUI wizard. Furthermore, to simplify the HyperSwap volumes management the **mkvolume** and **rmvolume** commands have been introduced. Nevertheless it is worth reviewing the detailed implementation steps and the related CLI commands.

1. Set the site names (**chs site** command).
2. Set the site attribute for the node canisters (**chnodecanister** command).
3. Set the site attribute for the external controllers (**chcontroller** command).
4. Set the site attribute for the existing hosts (**chhost** command).
5. Check the quorum assignment (**lsquorum** command) and change the assignment, if needed (**chquorum** command).
6. Set the hyperswap topology (**chsystem** command).
7. Create the Master and Auxiliary Volumes (**mkvdisk** command).
8. Create the Master and Auxiliary CVs (**mkvdisk** command).
9. Create the active-active Metro Mirror relationship (**mkrcrelationship** command).
10. Add the CV to the active-active Metro Mirror relationship (**chrcrelationship** command).
11. Create the Consistency Groups, if needed (**mkrcconsistgrp** command).
12. Add the active-active Metro Mirror relationships to the consistency groups (**chrcrelationship** command)

**Note:** For steps 7-10 can be used the **mkvolume** command instead





# VMware

This chapter addresses the steps to create a VMware environment. It includes the following sections:

- ▶ VMware configuration checklist
- ▶ VMware with Storwize V7000 HyperSwap
- ▶ VMware vCenter setup
- ▶ ESXi host installations
- ▶ Naming conventions
- ▶ VMware vSphere High Availability
- ▶ VMware vStorage API for Array Integration
- ▶ vCenter Services protection
- ▶ VMware recovery planning
- ▶ Design comments
- ▶ Script examples

## 5.1 VMware configuration checklist

The following items are required to gain the full benefit of the vSphere Metro Storage Cluster (vMSC) environment. This high-level list includes the major tasks that you must complete. The detail and expertise that are required to complete these tasks are beyond the intended scope of this book. Links are provided to find assistance on various topics.

**Tip:** VMware Communities are a good source of information:

<http://communities.VMware.com/welcome>

**Note:** The VMware Product Interoperability Matrixes are available on the VMware website:

[http://partnerweb.VMware.com/comp\\_guide2/sim/interop\\_matrix.php?](http://partnerweb.VMware.com/comp_guide2/sim/interop_matrix.php?)

You must complete the following items:

1. Create naming conventions (5.5, “Naming conventions” on page 119):
  - Data center-wide naming VMware ESXi
  - Storage Distributed Resources Scheduler (SDRS) data stores and pools data center affinity
  - VMware vSphere Distributed Resource Scheduler (DRS) vMotion pools data center affinity
2. Set up *all* hardware, and create a detailed inventory list:
  - Follow the VMware Compatibility Guide:  
<http://ibm.biz/BdxrmT>
  - Create an inventory list with details that cover the entire installation.
  - Mark the IBM SAN Volume Controller node names carefully and create associations in vSphere so that you know which SAN Volume Controller nodes are hosted in each data center.
3. Build ESXhosts (5.4, “ESXi host installations” on page 109):
  - Build two ESXhosts in each data center for maximum resiliency.
  - Patch and update to the latest VMware patch level.
  - Follow VMware vSphere High Availability (HA) deployment:  
<https://ibm.biz/BdXuQD>

4. Create one VM to host vCenter that is protected by vCenter (5.3, “VMware vCenter setup” on page 108):
  - Update and patch vCenter.
  - Build a stretched ESXi cluster between two data centers (5.6.3, “HA advanced settings” on page 124).
  - Optional: Implement I/O control on storage.
  - Optional: Implement VMware vSphere Distributed Switches (VDSs).
5. Build an SDRS pool:
  - Create at least two pools to match data center affinity.
  - Differentiate between Mirrored and Non-Mirrored logical unit numbers (LUNs) if both Mirrored and Non-Mirrored LUNs are used.
  - Set the SDRS pool to manual in the beginning and monitor it before you automate it.
6. Enable DRS (5.4.8, “VMware Distributed Resource Scheduler” on page 116):
  - Create affinity rules to the ESXi host in each data center.
  - Create affinity rules to VMs, if necessary.
  - Create VM to ESXi affinity rules.
  - Set DRS to partial or automatic if the rules are trusted 100%.

## 5.2 VMware with Storwize V7000 HyperSwap

Figure 5-1 shows an overview of the VMware in an IBM Storwize V7000 HyperSwap environment solution. It shows how the read/write operation is performed from an ESXi host perspective. Read/write operations are always performed by the node canister with the same defined Host Site Awareness attribute and that currently has the Primary attribute on the volume.

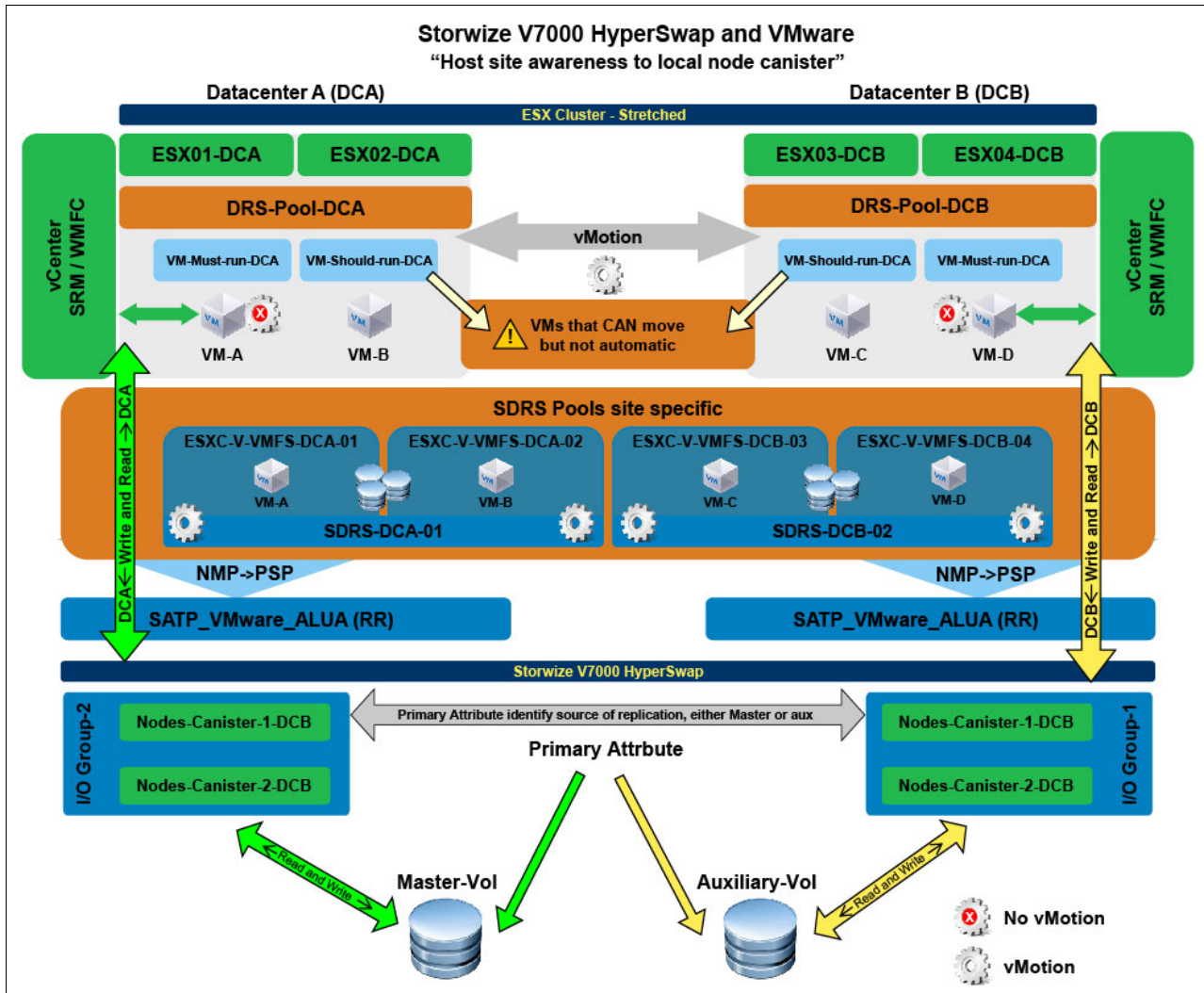


Figure 5-1 VMware stretched cluster with Storwize V7000 HyperSwap

## 5.3 VMware vCenter setup

You must implement vCenter configuration options as part of the Storwize V7000 HyperSwap configuration.

In a Storwize V7000 HyperSwap solution, vCenter can span the two sites without problems. However, ensure that connectivity and startup are set to **Automatic with host**. That way, in a total failure, the vCenter tries to start automatically, with the other vital virtual machines (VMs), such as domain controllers, Domain Name System (DNS), and Dynamic Host Configuration Protocol (DHCP), if used.

Create affinity rules to keep these VMware components on the same Primary site.

**Clarification:** The Storwize V7000 HyperSwap example implementation uses the vCenter 6.0.0 virtual appliance Tiny Model.

### 5.3.1 Metro vMotion vMSC

We recommend that you use the enhanced version of vMotion, which is called Metro vMotion, in a HyperSwap environment. Metro vMotion raises the allowed latency value 5 - 10 milliseconds (ms) round-trip time (RTT). This increase is required when failure domains are separated by a distance of more than 300 km (186.4 miles). The Enterprise Plus License is required for Metro vMotion.

In our Storwize V7000 HyperSwap solution, the maximum distance is 100 km (62 miles), which is an RTT of maximum 5 ms, which is not, in vMotion terms, a Metro vMotion. But, this distance is supported.

The VMware guide to a vMSC in vSphere 6 is available at this website:

<https://ibm.biz/BdXCzf>

vSphere new Long Distance vMotion allows up to 100 ms RTT. vSphere Long Distance vMotion is also usable in this scenario, but we need to keep the RTT under 5 ms due to our storage-related requirements.

## 5.4 ESXi host installations

This chapter does not go into detail about the installation and setup of an ESXi host. It focuses on the design and implementation that relate to a specific ESXi configuration with Storwize V7000 HyperSwap.

**Important:** Adhere to all VMware best practice configuration guidelines for the installation of ESXi hosts.

The best way to ensure standardization across ESXi hosts is to create an ESXi pre-build image. This image helps to ensure that all settings are the same between ESXi hosts. This consistency is critical to the reliable operation of the cluster. You can create this image by using VMware Image Builder or a custom scripted installation and configuration. The standardization of the ESXi hosts safeguards against potential mismatches in configurations.

**Important:** Due to the host affinity awareness in Storwize V7000 HyperSwap, the ESXi host needs to reboot after Storwize V7000 HyperSwap finds the optimized paths. For more information, see Chapter 4, "Implementation" on page 55.

## 5.4.1 ESXi host bus adapter requirements

The host bus adapters (HBAs) for the ESXi hosts must comply with these requirements:

- ▶ ESXi hosts require a minimum of two HBAs. The HBAs must be the same type and speed.
- ▶ The HBAs must be listed in the VMware Compatibility Guide:

<http://ibm.biz/BdxrmT>

- ▶ The HBA firmware levels must be current and supported according to the relevant hardware compatibility guides:

- VMware Compatibility Guide:

<http://ibm.biz/BdxrmT>

- IBM System Storage Interoperation Center (SSIC):

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

## 5.4.2 Initial ESXi verification

Check the latency RTT between ESXi hosts to ensure that the latency RTT does not exceed the maximum supported time of 10 ms in a Metro vMotion setup or 5 ms in a non-Metro vMotion setup. To run the latency test, use this command:

```
Vmkping <IP of remote ESXhost, vMotion Network>
```

Perform a 500-ping test, pipe it to a file for later comparison, and include the date in the file name:

```
vkping-test01-06112013.txt
```

Example 5-1 is from a `vmkernel` ping test between two hosts in different data centers.

*Example 5-1* `vmkping <hostname> -c 500`

---

```
vmkping ESXi-02-dcb.ibmse.local -c 500 > vkping-test01-06112013.txt
```

```
ping ESXi-02-dcb.ibmse.local (10.17.86.182): 56 data bytes
```

```
64 bytes from 10.17.86.182: icmp_seq=4 ttl=64 time=0.141 ms
```

---

Verify that the ping times that are returned are consistent and repeatable.

**Tip:** Keep a record of the ping times for future reference. This record can assist with future troubleshooting, if required.

If quality of service (QoS) is enabled on the physical network switches, the QoS settings must be validated. The validation ensures that adequate bandwidth is available so that the RTT is not affected by other traffic on the network.

Example 5-2 shows the `esxcli storage san fc list` command to verify that the adapters are online and functional.

*Example 5-2 esxcli storage san fc list command*

---

```
# esxcli storage san fc list
OutPut: Adapter: vmhba2
       Port ID: 0A8F00
       Node Name: 20:00:00:24:ff:07:50:ab
       Port Name: 21:00:00:24:ff:07:50:ab
       Speed: 4 Gbps
       Port Type: NPort
       Port State: ONLINE

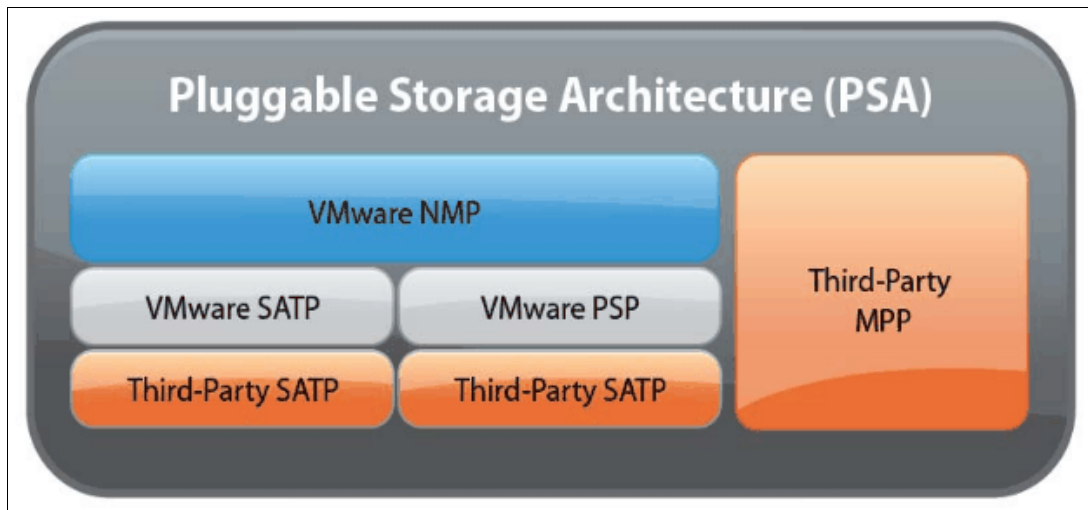
       Adapter: vmhba3
       Port ID: 1E8F00
       Node Name: 20:00:00:24:ff:07:52:98
       Port Name: 21:00:00:24:ff:07:52:98
       Speed: 4 Gbps
       Port Type: NPort
       Port State: ONLINE
```

---

### 5.4.3 Path selection policies (PSP) and native multipath drivers (NMP)

The VMware overall Pluggable Storage Architecture (PSA) with the Native Multipath Plugin (NMP) will work with third-party drivers and the VMware built-in driver. In this solution, we use the VMware native drivers and the NMP layer.

Figure 5-2 shows the PSA architecture.



*Figure 5-2 PSA plug-in architecture*

For optimal performance, you must configure the ESXi paths so that the active paths access the SAN Volume Controller nodes that are local, which means that the nodes are in the same failure domain as the ESXi server.

With Storwize V7000 HyperSwap, Host Site Awareness allows us to use Round-Robin (RR), which is the path selection policy (PSP) `VMW_PSP_RR`. However, we need to verify that the host uses the correct node, which is the expected node.

Example 5-3 lists the disk and devices that are visible to the host.

*Example 5-3 esxcli storage nmp device list*

---

```
naa.600507680183053ef80000000000095
  Device Display Name: IBM Fibre Channel Disk
(naa.600507680183053ef80000000000095)
  Storage Array Type: VMW_SATP_ALUA
  Storage Array Type Device Config: {implicit_support=on;explicit_support=off;
explicit_allow=on;alua_followover=on;{TPG_id=0,TPG_state=A0}{TPG_id=1,TPG_state=AN
0}}
  Path Selection Policy: VMW_PSP_RR
  Path Selection Policy Device Config:
{policy=rr,iops=1000,bytes=10485760,useAN0=0; lastPathIndex=1:
NumIOsPending=0,numBytesPending=0}
  Path Selection Policy Device Custom Config:
  Working Paths: vmhba3:C0:T0:L3, vmhba4:C0:T0:L3
  Is Local SAS Device: false
  Is Boot USB Device: false
```

---

**Note:** Example 5-3 shows that the device is visible as VMW\_SATP\_ALUA. Ensure that this device is the default by using VMW\_PSP\_RR.

Storwize V7000 7.5 HyperSwap uses VMW\_SATP\_ALUA with VMW\_PSP\_RR.

## Verifying the path selection policy

You can verify the current PSP for each LUN by using the command that is shown in Example 5-4.

*Example 5-4 Verifying the PSP*

---

```
esxcli storage nmp device list | grep "Path Selection Policy:"
OutPut: (One for each Path active)
Path Selection Policy: VMW_PSP_RR
  Path Selection Policy: VMW_PSP_RR
  Path Selection Policy: VMW_PSP_RR
  Path Selection Policy: VMW_PSP_FIXED
```

---

For more information about how to obtain the LUN path information from the ESXi hosts, see the VMware Knowledge Base article, *Obtaining LUN path information for ESX or ESXi hosts*, 1003973:

<http://ibm.biz/BdxriP>



When the SATP\_ALUA\_RR PSP is enabled without the optimized path, you see I/O that passes through all of the paths as shown in Figure 5-3.

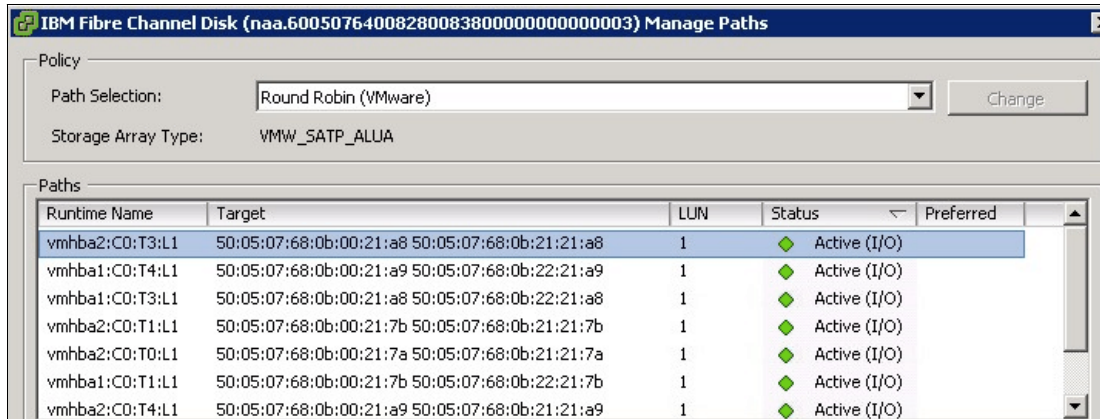


Figure 5-3 Active I/O on all paths before host affinity awareness is enabled and active

After you enable the host affinity awareness on Storwize V7000 HyperSwap, you must reboot the ESXi host. Now, the I/O uses the optimized paths only, as shown in Figure 5-4.

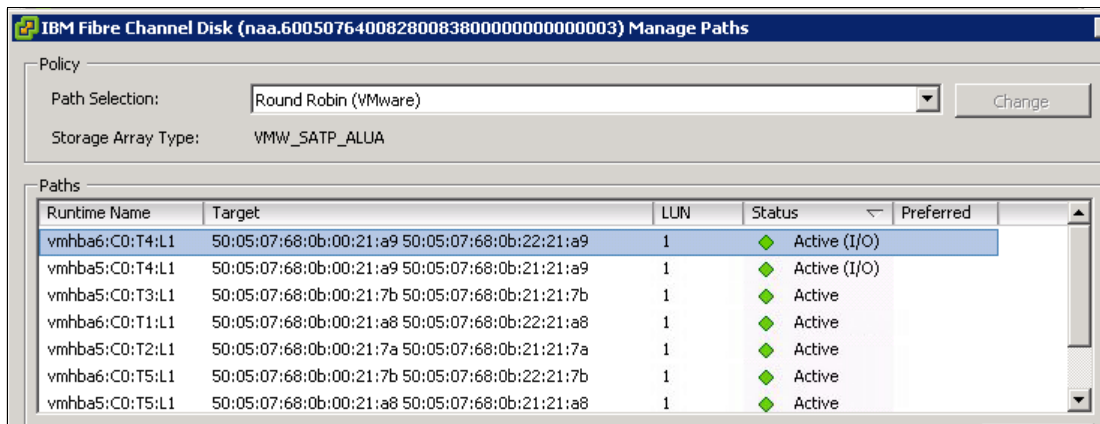


Figure 5-4 Active I/O on two paths after host affinity awareness is enabled

#### 5.4.4 Set the maximum number of logical unit numbers

In Storwize V7000 HyperSwap, we used two node canisters at each site. This configuration affects the number of possible paths from the ESXhost. For more information about the configuration maximums, see this VMware document:

<https://www.vmware.com/pdf/vsphere6/r60/vsphere-60-configuration-maximums.pdf>

The two parameters that are shown in Figure 5-5 are listed under the advanced settings. We set the VMkernel.Boot.storageMaxDevices setting to 128.

VMkernel.Boot.storageMaxDevices	128	Maximum number of supported SCSI devices
VMkernel.Boot.storageMaxPaths	1024	Maximum number of supported SCSI paths

Figure 5-5 VMkernel.Boot.storageMaxDevices setting

**Important:** In Storwize V7000 HyperSwap, we use eight LUN paths from each vSphere ESXi host. Therefore, we must reduce the number of LUNs to 128 because vSphere 6 has a maximum of 1,024 active paths at any time on all HBAs. Two node canisters are at each site in our environment.

## 5.4.5 Set the default path selection policy (PSP)

Set the PSP for the entire ESXi host, which requires that the ESXhost is restarted. From the ESXi shell console, list the available vendors and levels, as shown in Example 5-5.

**Note:** The default for VMW\_SATP\_ALUA is VMW\_PSP\_MRU, which needs to be changed.

*Example 5-5 Listing vendors and levels before settings default*

---

```
esxcli storage nmp satp list
```

Name	Default PSP	DStorwize V7000 HyperSwaption
<b>VMW_SATP_ALUA</b>	<b>VMW_PSP_MRU</b>	<b>Supports non-specific arrays that use the ALUA protocol</b>
VMW_SATP_MSA	VMW_PSP_MRU	Placeholder (plugin not loaded)
VMW_SATP_DEFAULT_AP	VMW_PSP_MRU	Placeholder (plugin not loaded)
VMW_SATP_SVC	VMW_PSP_FIXED	Placeholder (plugin not loaded)
VMW_SATP_EQL	VMW_PSP_FIXED	Placeholder (plugin not loaded)
VMW_SATP_INV	VMW_PSP_FIXED	Placeholder (plugin not loaded)
VMW_SATP_EVA	VMW_PSP_FIXED	Placeholder (plugin not loaded)
VMW_SATP_ALUA_CX	VMW_PSP_RR	Placeholder (plugin not loaded)
VMW_SATP_SYMM	VMW_PSP_RR	Placeholder (plugin not loaded)
VMW_SATP_CX	VMW_PSP_MRU	Placeholder (plugin not loaded)
VMW_SATP_LSI	VMW_PSP_MRU	Placeholder (plugin not loaded)
VMW_SATP_DEFAULT_AA	VMW_PSP_FIXED	Supports non-specific active/active arrays
VMW_SATP_LOCAL	VMW_PSP_FIXED	Supports direct attached devices

---

Example 5-6 shows SATP\_ALUA.

*Example 5-6 List SATP*

---

```
esxcli storage nmp satp list | grep SATP_ALUA
```

<b>VMW_SATP_ALUA</b>	<b>VMW_PSP_MRU</b>	<b>Supports non-specific arrays that use the ALUA protocol</b>
VMW_SATP_ALUA_CX	<b>VMW_PSP_RR</b>	Placeholder (plugin not loaded)

---

For more information, see the VMware Knowledge Base article, *Multipathing policies*:

<https://ibm.biz/BdXCnk>

To simplify adding disks in the future, we need to set the new default value on boot, as shown in Example 5-7.

*Example 5-7 Change default to VMW\_PSP\_RR*

---

```
esxcli storage nmp satp set --default-psp=VMW_PSP_RR --satp=VMW_SATP_ALUA
```

*Default PSP for VMW\_SATP\_ALUA is now VMW\_PSP\_RR*

---

You must reboot ESXhost for this change to take effect.

## 5.4.6 Verifying Node ID path in vSphere web client

First, create the Storwize V7000 HyperSwap node canister table. Table 5-1 shows an example.

Even with Host Site Awareness, it is important to create this node canister table initially to verify that all of the components are the expected components and that you are using the correct policy regarding the node canisters.

Table 5-1 Node canister list table example

Node canister ID	Data center
50:05:07:68:01:10:B1:3F	Data Center A (DCA) (HBA3)
50:05:07:68:01:10:27:E2	DCA (HBA4)
50:05:07:68:01:40:B0:C6	DCB (HBA3)
50:05:07:68:01:40:37:E5	DCB (HBA4)

Table 5-2 shows the ESX01-DCA data stores that map to local Storwize V7000 node canisters.

Table 5-2 ESXi-DCA Storwize V7000 HyperSwap map example

ESXhost	Data store	SAN Volume Controller ID	Policy	Preferred state
ESX01-DCA	ESXC_00_VMFS_V_DCA_01	50:05:07:68:01:10:B1:3F	RR	Not marked
ESX01-DCA	ESXC_01_VMFS_V_DCB_01	50:05:07:68:01:10:27:E2	RR	Not marked
ESX01-DCA	ESXC_00_VMFS_V_DCA_02	50:05:07:68:01:10:B1:3F	RR	Not marked
ESX01-DCA	ESXC_01_VMFS_V_DCB_02	50:05:07:68:01:10:27:E2	RR	Not marked

Table 5-3 shows ESX02-DCB to data stores map example.

Table 5-3 ESXi-DCB example

ESXhost	Data store	SAN Volume Controller ID	Policy	Preferred state
ESX02-DCB	ESXC_00_VMFS_V_DCA_01	50:05:07:68:01:40:B0:C6	RR	Not marked
ESX02-DCB	ESXC_01_VMFS_V_DCB_01	50:05:07:68:01:40:B0:C6	RR	Not marked
ESX02-DCB	ESXC_00_VMFS_V_DCA_02	50:05:07:68:01:40:37:E5	RR	Not marked
ESX02-DCB	ESXC_01_VMFS_V_DCB_02	50:05:07:68:01:40:37:E5	RR	Not marked

## Ensure that the path selection policy is set to RR (VMware)

Verify that the target information matches the expected ID of the host awareness node according to the inventory plan. In this case, 50:05:07:68:01:10:B1:3F is in Data Center A (DCA) and 50:05:07:68:01:40:B0:C6 is in Data Center B (DCB).

**Guideline:** In Storwize V7000 HyperSwap, the optimized path is monitored for the load behavior. For more information, see Chapter 3, “IBM System Storwize V7000 HyperSwap architecture” on page 23.

For instructions to verify the active and preferred paths by using a script, see 5.11.2, “PowerShell script to extract data from the entire environment to verify active and preferred paths” on page 139. These instructions are still valid to use for verification, even with the Host Site Awareness.

### 5.4.7 Path failover behavior for an invalid path

If an active path fails, the ESXi PSP randomly selects an alternative path that is determined initially by the handshake of ESXi and the Storwize V7000 HyperSwap node canister, which is part of Host Site Awareness.

Storwize V7000 HyperSwap determines the optimized path to the local node, and the ESXi host will use RR to use one of the two paths to the local node.

The ESXi host has a site-specific path to a local node, and all read/writes go through the path, even if the Primary copy (Master) Storwize V7000 HyperSwap is at another site on another node. The node canister that is not on an optimized path is used only in the case of a total site failure. When the optimized path returns, the ESXi host resumes the use of the optimized path.

### 5.4.8 VMware Distributed Resource Scheduler

It is important to use the VMware vSphere Distributed Resource Scheduler (DRS) in a Storwize V7000 HyperSwap setup because, in normal operations, you do *not* want VMs to move to the other site. You want VMs to move to the other site only in the case of a site failure, or intentionally.

Before you use DRS, you must create an ESXi cluster and enable DRS in the menus.

**DRS Blog Guide:** The DRS Blog Guide is at the following website:

<https://ibm.biz/BdXLXP>

This blog is one of many blogs that provide good information about how DRS works.

**Important:** DRS rules, along with accurate and meaningful naming standards, are the most important operational considerations when you manage a Storwize V7000 HyperSwap.

DRS mode can be set to *automatic* under normal conditions, but only if the appropriate rules, which are shown in Table 5-4, are always in place. Be aware again of the high availability (HA) settings to ignore these rules in an HA failure.

Table 5-4 DRS rules matrix

DRS-Rules	ESX-DRS-Host	Storwize V7000 HyperSwap
VM-Should-Run-In-DCA	ESX-Host-In-DCA	VMs in Data Center A (DCA) that potentially can be vMotion to Data Center B
VM-Should-Run-In-DCB	ESX-Host-In-DCB	VMs in Data Center B (DCB) that potentially can be vMotion to Data Center A
VM-Must-Run-DCA	ESX-Host-In-DCA	No vMotion, stick to Data Center A
VM-Must-Run-DCB	ESX-Host-In-DCB	No vMotion, stick to Data Center B

The Should-Run rules apply to VMs that can be moved to the alternate site to manage pre-disaster situations.

The Must-Run rules apply to VMs that must *never* be moved to the alternate site. These rules are used for VMs, such as a domain controller or vCenter Primary or Secondary, dedicated vReplicator Servers, or a Microsoft Cluster Service (MSCS) cluster that runs with virtual nodes.

**Important:** All VMs must be in a rule when they are running a Storwize V7000 HyperSwap. Create the rules when the VMs are running, because you cannot create empty groups.

Since vSphere 5.5, the VM to VM anti-affinity rules changed so that whenever HA restarts a VM, the rules will respect this setting.

Figure 5-6 shows the creation of a DRS group's rule.

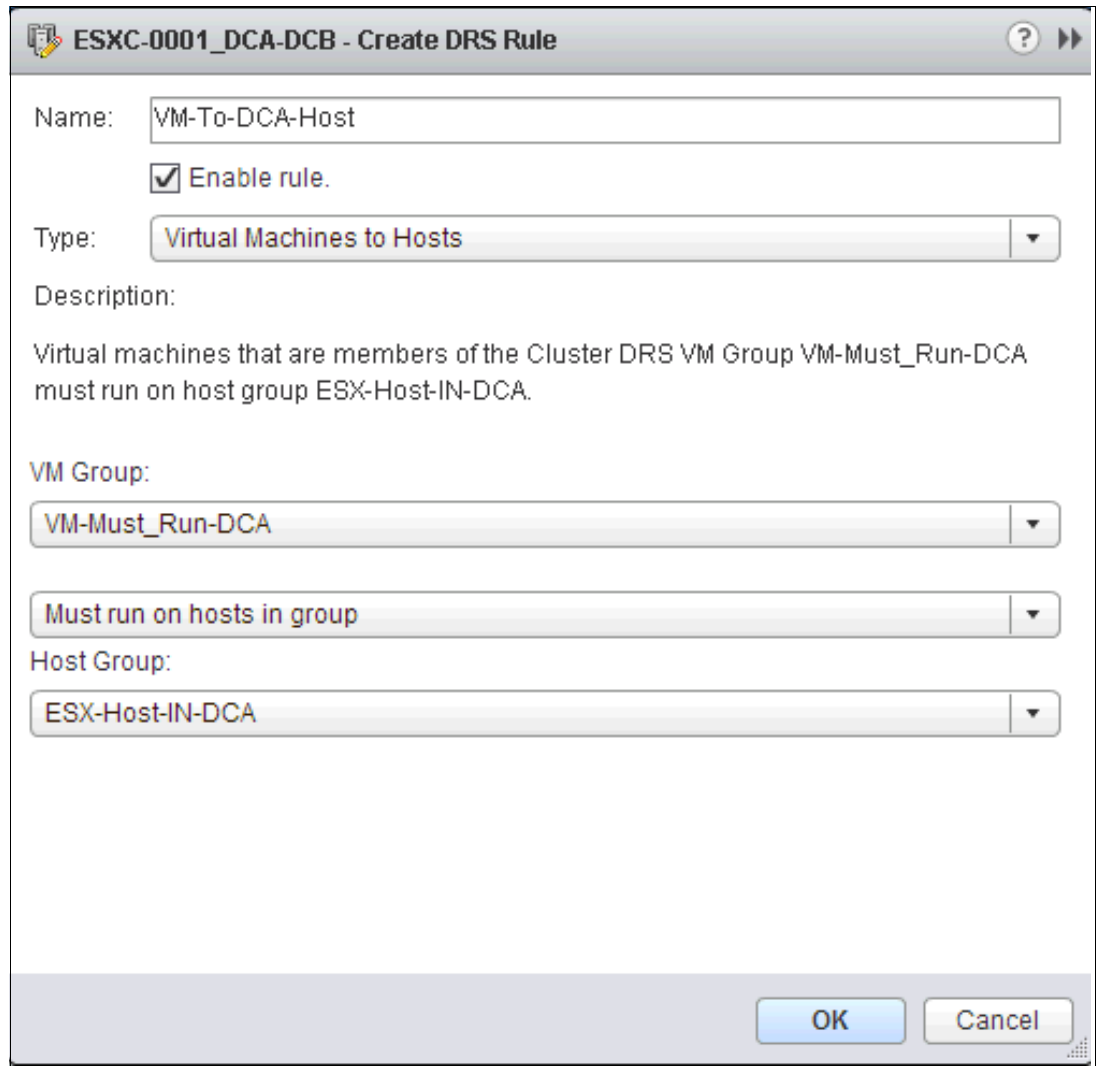
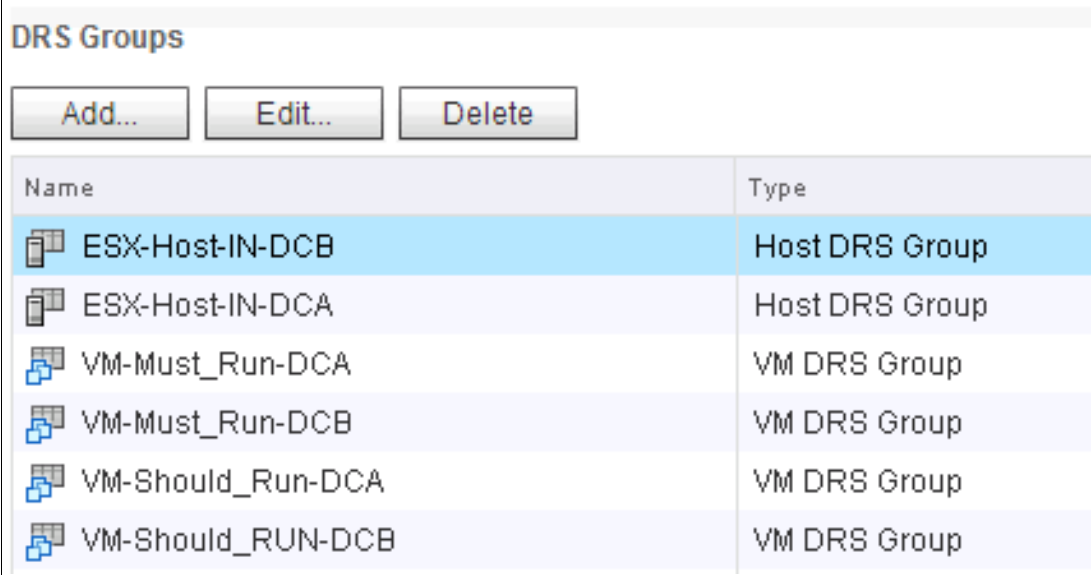


Figure 5-6 Creating DRS group rules

Figure 5-7 shows an example of DRS Groups rules that are implemented in vCenter.



Name	Type
ESX-Host-IN-DCB	Host DRS Group
ESX-Host-IN-DCA	Host DRS Group
VM-Must_Run-DCA	VM DRS Group
VM-Must_Run-DCB	VM DRS Group
VM-Should_Run-DCA	VM DRS Group
VM-Should_RUN-DCB	VM DRS Group

Figure 5-7 DRS VM rules

**Important:** Systems commonly encounter Critical Events because of missing and outdated guidelines in these rules.

Align your VMs with the corresponding groups with the datastore where they are stored.

When rules are active for a VM, the VM can be manually moved by overriding the rules by using the migrate option.

## 5.5 Naming conventions

The use of a strict and well-thought-out naming convention is critical to the reliable operation of the environment. In combination with Storwize V7000 HyperSwap, a meaningful naming convention helps you to know where resources are running and to identify a resource by only its name.

**Consideration:** Implementing and maintaining a meaningful naming convention is the most important disaster prevention option that is available that requires no software to control. It provides administrators with the ability to visually determine whether VMs and datastores are running at the correct site.

Several examples of naming standards are shown:

▶ ESX Cluster names:

- ESX;C;####;\_DCS
- ESXC-0001\_DCA-DCB

▶ ESX host names:

- ESXi-##-DC.<DNS-ZONE>
- ESXi-01-DCA.DNS-zoneA.com

▶ VMs:

- VM-<DCA-####>.<domainname.XXX>

▶ Datastores:

- ESX\_<Cluster##>\_<DiskType>\_<MirrorType>\_<DC><#LUN\_ID>\_<OWNER>
- <Cluster> {just Unique} prefer a number
- <DiskType> [VMFS/NFS/RDM]
- <MirrorType>

Several mirror types are listed:

- M = Metro Mirrored Disk (for synchronous disaster recovery (DR))
- V = Volume Mirrored Disk (for business continuity)
- G = Global Mirror Disk (for asynchronous DR)
- H = HyperSwap (business continuity for Storwize V7000 and SAN Volume Controller (two sites only))
- N = Not Mirrored

- <DC> The preferred data center that holds the Primary disk copy of the LUN
- <LUN\_ID> Optional: The unique Small Computer System Interface (SCSI) ID that is assigned by storage [0 - 255]
- <OWNER> Optional: If the client wants dedicated LUNs to belong to certain applications, see the APP-ID or the name of the VM that owns that LUN.

▶ SDRS-Pools:

- SDRS-<Datacenter><####>
- For example: SDRS-DCA-001 (a pool of datastores in Data Center A (DCA))

Consider the following naming examples:

- ▶ ESXC\_01\_VMFS\_H\_DCA\_01: A HyperSwap Volume mirrored LUN
- ▶ ESXC\_01\_VMFS\_H\_DCB\_02: A volume mirrored LUN, which is preferred from DCB
- ▶ ESXC\_01\_VMFS\_V\_DCB\_03\_VM-DCA\_01: With a dedicated VM as owner



Data centers and clusters must be clearly named. Figure 5-8 is an example where we can clearly identify where the cluster is used and what the cluster is used for.

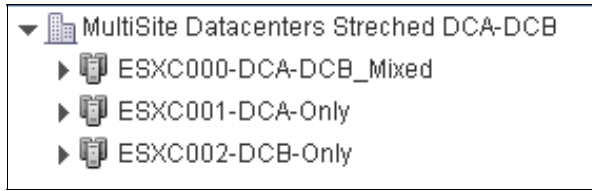


Figure 5-8 Data center and cluster naming example

Figure 5-9 shows a standard datastore naming example.

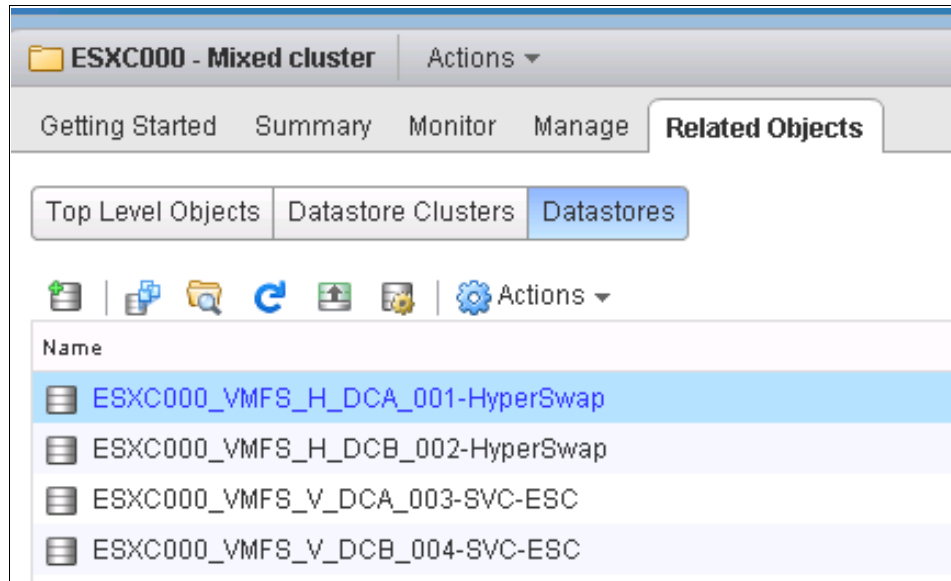


Figure 5-9 Datastore naming example

In addition to the actual names, we also suggest that you use a naming convention for folders that easily identifies the folder contents, as shown in Figure 5-10.

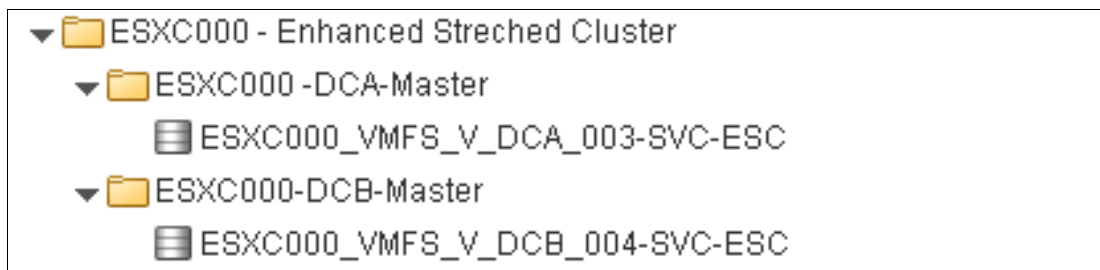


Figure 5-10 Use of folders in addition to naming standards

For a complete view of an environment where we use both Enhanced Stretched Cluster (ESC) volumes and HyperSwap Volumes in the same cluster, see Figure 5-11.

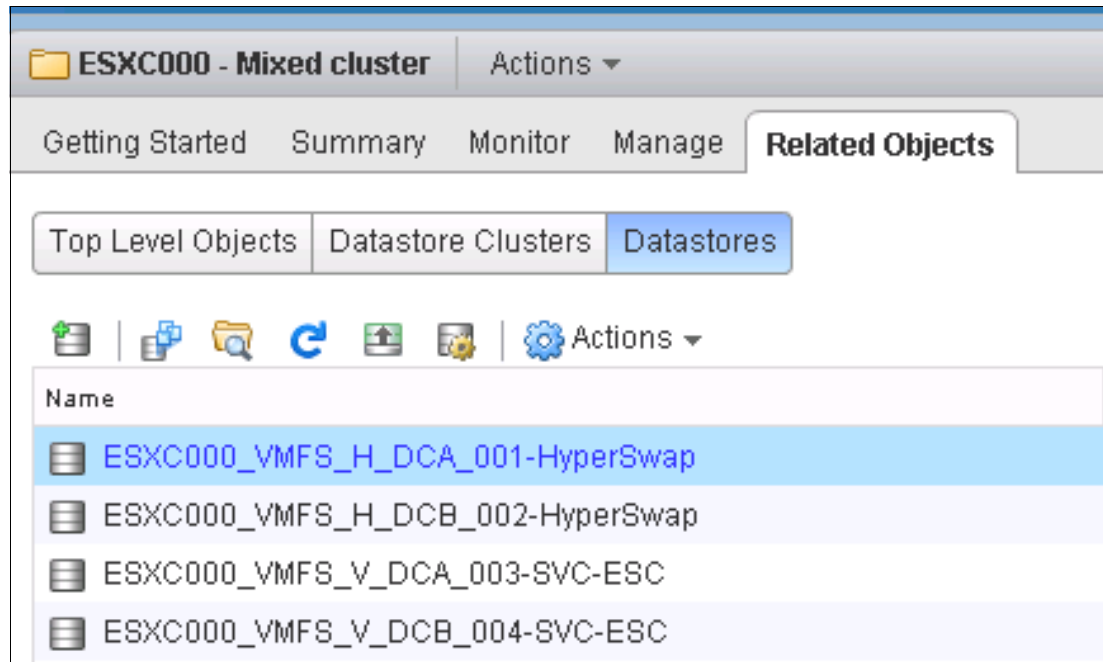


Figure 5-11 Mixed folder view example

When we open the folders, we keep the cluster ID as part of the folder name so that we keep unique names, as shown in Figure 5-12.

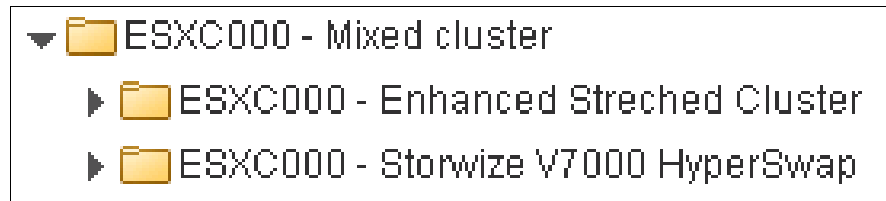


Figure 5-12 Complete folder view

Figure 5-13 is an example of how easily we can view only volumes on a certain site by using the search function if we know the naming standards. This example shows a search for all datastores with the term “DCA” in the name. Because we implemented naming standards, we can use the search function to easily search for all datastores with DCA in the name.

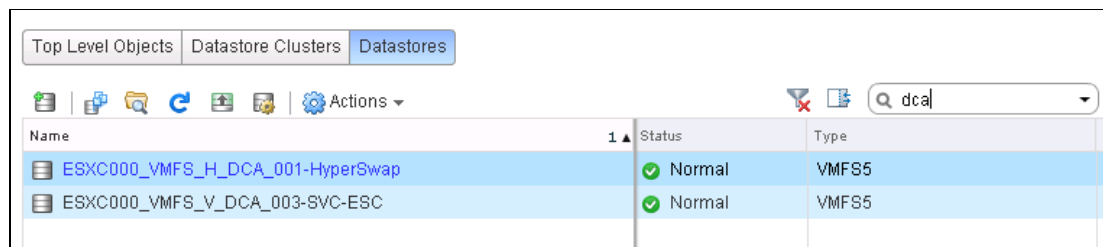


Figure 5-13 Search for datastores by reference

## 5.6 VMware vSphere High Availability

Setting up a redundancy network for VMware vSphere High Availability (HA) is critical between the ESXi hosts on the cluster. The instructions are beyond the intended scope of this book. For information about the setup and configuration of VMware vSphere HA, see *vSphere High-Availability Deployment Best Practices*:

<http://ibm.biz/BdxrmN>

### 5.6.1 High availability admission control

In a VMware Storwize V7000 HyperSwap environment, ensure that each site can absorb the workload from the alternate site in a failure. Ensure that you reserve resources at each site, which are referred to as *Admission Control*.

For the vMSC environment, set the Admission Control policy to 50%. This amount varies according to your environment and needs. This setting can be changed on behalf of other resource controls or priorities in the cluster. The resource control is important in case of failure and disaster prevention scenarios, where the VMs can move to the partner ESXi host in the other data center.

Figure 5-14 shows setting the HA admission control policy settings (reserved failover CPU capacity and the reserved failover memory capacity) to 50%.

Host Monitoring	<input checked="" type="checkbox"/> Enable host monitoring
Admission Control	
Admission Control Status	Admission control will prevent powering on VMs that violate availability constraints <input checked="" type="checkbox"/> Enable admission control
Policy	Specify the type of the policy that admission control should enforce. <input type="radio"/> Host failures cluster tolerates: 1 <input checked="" type="radio"/> Percentage of cluster resources reserved as failover spare capacity: Reserved failover CPU capacity: 50 % CPU Reserved failover Memory capacity: 50 % Memory

Figure 5-14 HA admission control settings

### 5.6.2 High availability heartbeat

*Heartbeat* is a method for detecting possible downtime of an ESXi host to enable recovery actions that are based on the defined policies. The feature that is called Fault Domain Manager (FDM) is implemented in vSphere 5, which is completely rewritten HA code.

FDM operates at an Internet Protocol (IP) level, not at the DNS level. In addition, vCenter is now a component of FDM. Before FDM, HA automatically selected an isolation state, but now FDM and vCenter interact in the selection process.

When an ESXi host is isolated from the other ESXi host, you need a rule for what to do with the VMs on the host, if an isolation state occurs. Generally, set the policy to **Power off, then failover** in the cluster HA setup if a host enters the isolated state.

The VM options for the host isolation response are shown in Figure 5-15.

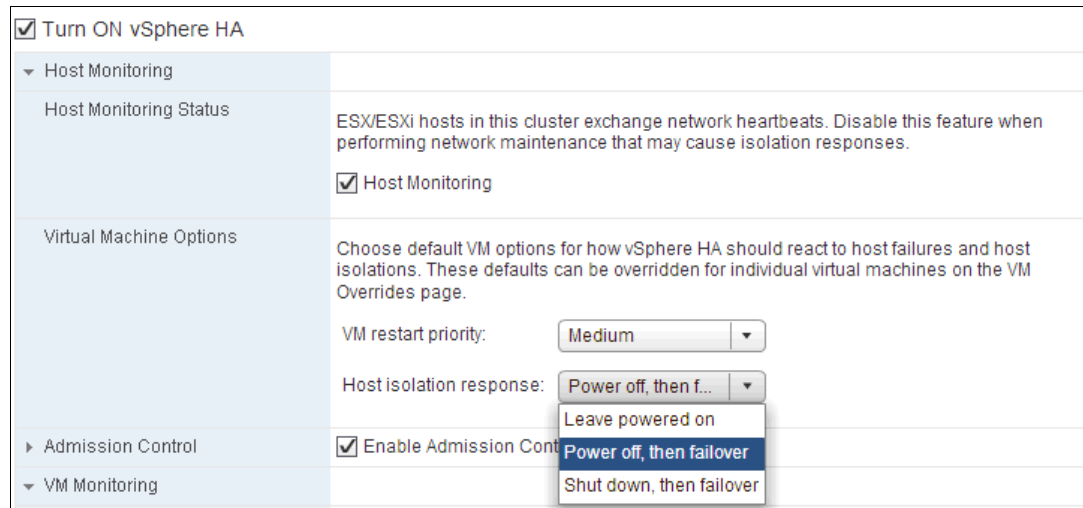


Figure 5-15 Virtual machine isolation response

**Important:** Ensure that you review the following list because HA was changed.

If the host is isolated because the redundancy management network is down, the following two heartbeat mechanisms are available:

- ▶ Networking heartbeat

Primary control: This mechanism checks the basic network for isolation of an ESXi host. Ensure that at least two interfaces are defined with isolation addresses. For more information, see 5.6.3, “HA advanced settings” on page 124.

- ▶ HA datastore heartbeat

Secondary control: VMware allows vCenter to find the best possible datastores for control. You can manually set the datastore that you think is the best datastore and where the ESXi host has the most connections, but we suggest that you use the default.

As Figure 5-16 shows, the best selection policy is **Automatically select datastores accessible from the host** for the heartbeat.

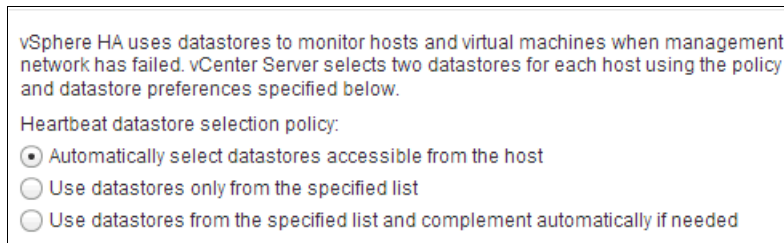


Figure 5-16 Datastore selection for heartbeat

### 5.6.3 HA advanced settings

You need to consider several other important HA settings. Your environment might require additional or different settings, but the following settings were applicable to our example environment.

Table 5-5 describes the advanced setting that must be applied.

**Remember:** This list is not a comprehensive list of the advanced settings. The settings that are listed in this chapter are critical to our example Storwize V7000 HyperSwap implementation.

For more information, see the VMware Knowledge Base article titled *Advanced configuration options for VMware High Availability in vSphere 5.x*, 2033250:

<https://ibm.biz/BdRxV8>

Table 5-5 HA advanced setting

HA string	HA value	Brief explanation
das.maskCleanShutdownEnabled	TRUE	Since version 5.1, this option is set to TRUE, by default. The TRUE setting allows HA to restart VMs that were powered off while the Permanent Device Loss (PDL) condition was in progress.

## 5.6.4 Enhanced All Paths Down detection in vSphere 6

Since vSphere 5.0 Update 1, vSphere uses SCSI Sense Codes to determine whether a VM is on a datastore that is in an All Paths Down (APD) state.

**Important:** This mechanism is part of the process to secure the ESXi host isolation handling of the VMs. *Do not disable this mechanism.*

See the VMware article, *Handling Transient APD Conditions*:

<https://ibm.biz/BdDwq7>

The APD timeout default is 140 seconds, which is enough time to cover most connection losses.

Figure 5-17 shows an example of the datastore event in vCenter for this situation.

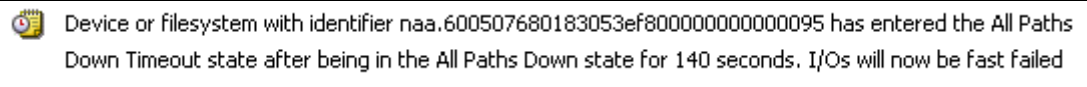


Figure 5-17 APD event

New in vSphere 6.0 is a view of the volumes that failed due to APD or PDL.

Figure 5-18 shows the APD or PDL status under cluster monitoring.

	Unhealthy Hosts	Datastores under APD or PDL	Datastore Cluster	Failure
Summary	10.18.228.61	ESXC000_002_VMFS_H...		APD Detected
Heartbeat	10.18.228.61	ESXC000_001_VMFS_H...		APD Detected
Configuration Issues				
<b>Datastores under APD or PDL</b>				

Figure 5-18 APD or PDL datastore status page

For more information about PDL and ADL states, see the VMware documentation for *Working with Permanent Device Loss*:

<http://ibm.biz/Bdx4k7>

## 5.6.5 Permanent Device Loss (PDL)

From vSphere 5.5 and higher, the advanced system setting to autoremove paths to a disk that is in PDL, Disk.AutoremoveOnPDL, is set to 1, by default. This setting is *not* recommended for use in a Storwize V7000 HyperSwap environment because Storwize V7000 HyperSwap expects the disks to be automatically visible soon afterward. By removing the disks, an operator needs to execute a Storwize V7000 HyperSwap manually to the disks to get them back, which can lead to a misunderstanding of where the disks are and lead someone to think that the disks are gone.

**Note:** In Advanced System Settings, set the Disk.AutoremoveOnPDL value to 0 (zero).

Follow these steps:

1. To access the settings, navigate to the vSphere web client and click the ESXhost. Select **Manage** → **Settings** → **Advanced System Settings**, as shown in Figure 5-19.
2. In the filter box, click the **Edit** icon (pencil). Select **DISK:Autoremove** from the list box. Figure 5-19 shows how to search for the AutoremoveOnPDL setting.

Name	Value	Description
Disk.AutoremoveOnPDL	0	Autoremove paths to a disk that is in PDL (Per...

Figure 5-19 Advanced System Settings for PDL

Edit the advanced option for AutoremoveOnPDL to 0, as shown in Figure 5-20.

AutoremoveOnPDL:

Autoremove paths to a disk that is in PDL (Permanent Device Loss)

OK Cancel

Figure 5-20 Edit the AutoremoveOnPDL setting

Ensure that the settings are the same on all ESXi hosts in your cluster.

**Consideration:** VMs that are not running any I/O operations might not be disabled correctly. If this problem occurs, you can stop the VM manually from the console by using the `vmfstools` command.

VMs that are running on multiple datastores with a 50% PDL will not be stopped.

Also, pending I/Os or a raw device disk can potentially lock the disk. If you are removing a LUN, the PDL is effective and waits for all I/O to be released.

To ensure that PDL is working as intended after you change the settings, test it by “zoning out” one disk to one of the ESXi hosts. This process triggers the automatic PDL, so the VMs are powered off from the host and restarted on one of the other ESXi hosts.

## 5.6.6 Virtual Machine Component Protection (VMCP)

This new feature in vSphere 6 enables VMware to react to failures from either an APD or PDL state. This feature is configured for the cluster, but it can be overridden by the individual VM setting.

This feature is one of the most important features to set when you work with storage-related failures and actions, and to automate HA on VMs.

**Important:** Ensure that you verify the setting on the VMs. This feature might not be enabled on existing VMs if the setting in the cluster was made before the VMs were created.

For more information, see the *VM Component Protection (VMCP)* blog from VMware:  
<https://ibm.biz/BdXukg>

By using the blog, we created the workflow and recovery timelines that are shown in Figure 5-21.

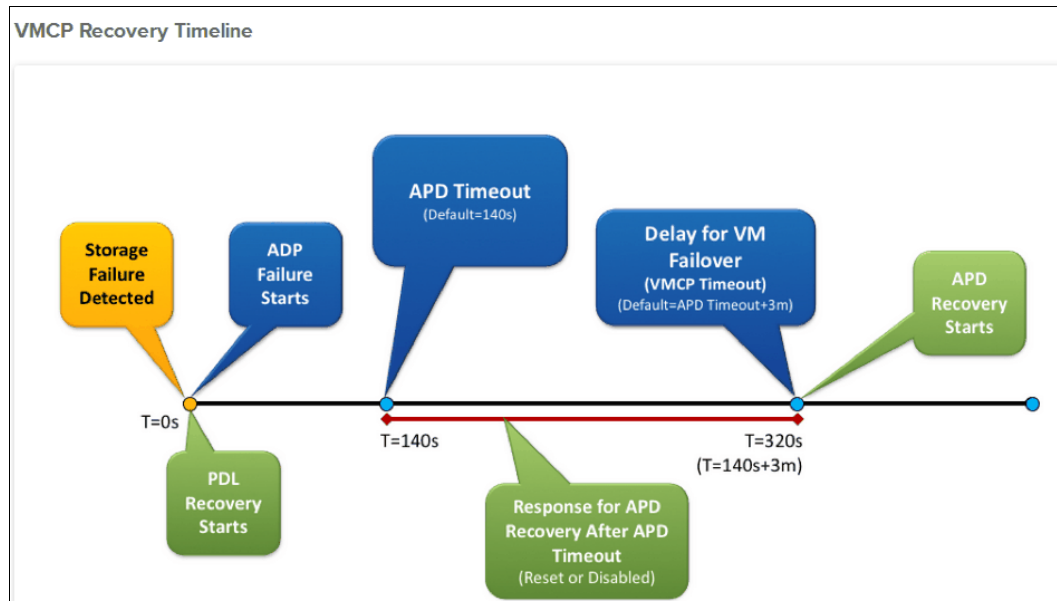


Figure 5-21 VMCP recovery timeline

**Note:** The VMCP timeout can be adjusted, for example:

VMCP timeout = [VMCP default 180 seconds + APD TimeOut.140 seconds default]

So, in this case, a 320-second default timeout occurred before VMCP was activated.

**Tip:** You must use the Web-Client, and not the Vi-Client because not all options are enabled in Vi-Client.

Be aware of the DRS groups and affinities, because the DRS groups and affinities can prevent a VM from being restarted. Also, if a VM has more disks that are spread over multiple datastores, the VM will not be restarted if not all of the datastores are under the APD or PDL state.

HA has an option to work with DRS affinity rules. Depending on your design and goals, you must consider that option. Figure 5-22 shows the vSphere HA and DRS rules settings.

VM/Host Rules	vSphere HA Rule Settings	
VM Overrides	vSphere HA can enforce VM/Host rules when restarting virtual machines.	
Host Options	VM anti-affinity rules	Ignore rules
Profiles	VM to Host affinity rules	Ignore rules

Figure 5-22 HA ignore rules



**Important:** VMs that were created before you enabled VMCP do not have the APD/PDL setting enabled, so these settings need to be enabled manually for each VM.

Figure 5-23 shows the VM that is disabled for APD/PDL.

The screenshot shows a configuration window for a VM. The settings are as follows:

- Automation level: Use Cluster Settings
- VM restart priority: Use Cluster Settings
- Response for Host Isolation: Use Cluster Settings
- Response for Datastore with Permanent Device Loss (PDL): Disabled
- Response for Datastore with All Paths Down (APD): Disabled
- Delay for VM failover for APD: 3 minutes
- Response for APD recovery after APD timeout: Disabled
- VM Monitoring: Use Cluster Settings
- VM monitoring sensitivity: --

Below these settings is a section for 'Relevant Cluster Settings':

▼ Relevant Cluster Settings	
▶ vSphere DRS	Fully Automated
▶ vSphere HA	Expand for details

At the bottom right are 'OK' and 'Cancel' buttons.

Figure 5-23 VM disabled for APD and PDL responses

When you set the cluster to use VMCP, ensure that you select the correct sensitivity. The setting can range from low to high, or from a response time of 180 seconds down to a response time of 30 seconds after an APD occurs. You can also choose a customized setting, but remember that using more than one setting in your solution, in a site failure or many failing datastores, can confuse the operator.

**Tip:** Keep settings uniform across all VMs in the same cluster.

VMCP monitoring sensitivity is key to get operations back up and running. In a split-brain scenario, you need an aggressive setting to allow HA to decide and to force only one instance of the VM.

Figure 5-24 shows the VM monitoring sensitivity settings in HA as high.

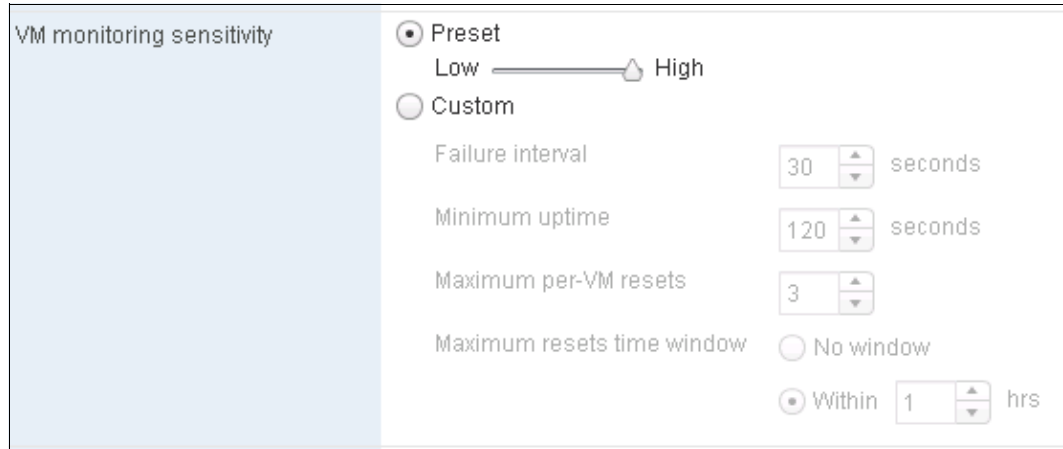


Figure 5-24 VM monitoring sensitivity setting is high

### 5.6.7 Storage failure detection flow

Figure 5-25 illustrates the entire process in a storage failure detection scenario.

For more information, see the *VM Component Protection (VMCP)* blog:

<https://ibm.biz/BdXukg>

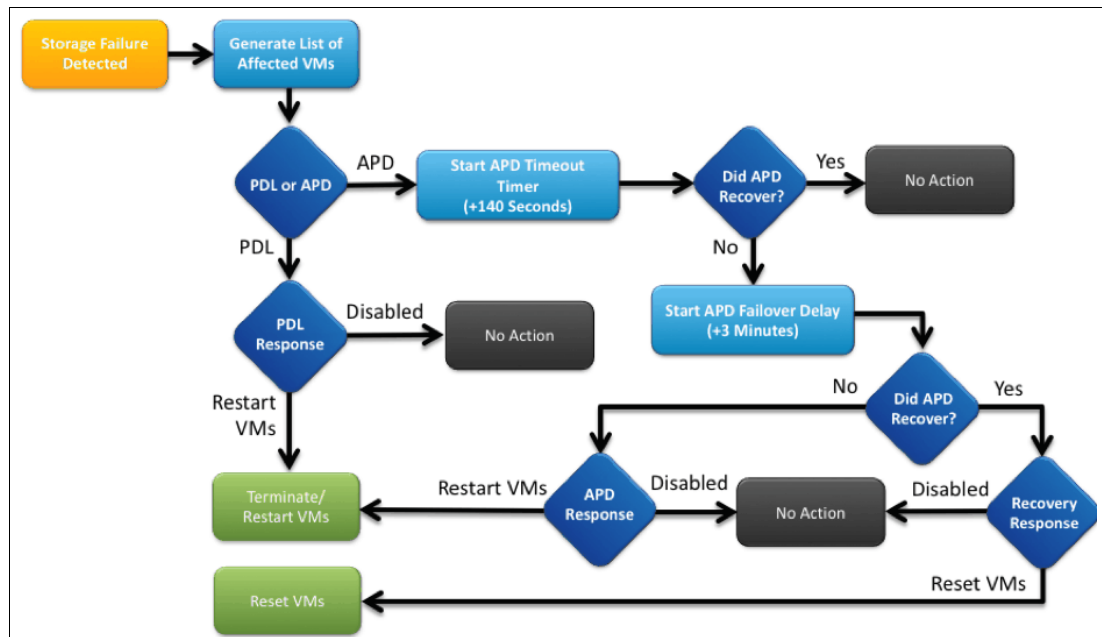


Figure 5-25 Storage failure detection flow

## 5.7 VMware vStorage API for Array Integration

VMware vStorage application programming interface (API) for Array Integration (VAAI) is supported if it is listed on the hardware compatibility list and software version 7.5 supports VAAI.

With VMware vSphere 6.x, you do not need to install a plug-in to support VAAI if the underlying storage controller supports VAAI.

You can use several commands to check the VAAI status. Example 5-8 shows the `esxcli storage core device vaa1 status get` command.

*Example 5-8 Checking the VAAI status*

---

```
esxcli storage core device vaa1 status get
VAAI Plugin Name:
ATS Status: supported
Clone Status: supported
Zero Status: supported
Delete Status: unsupported
```

---

To determine whether VAAI is enabled, issue the three commands and check whether the default interval value is set to 1, which means that VAAI is enabled. See Example 5-9.

*Example 5-9 Checking whether VAAI is enabled*

---

```
esxcli system settings advanced list -o /DataMover/HardwareAcceleratedMove
Path: /DataMover/HardwareAcceleratedMove
  Type: integer
  Int Value: 1
  Default Int Value: 1
  Min Value: 0
  Max Value: 1
  String Value:
  Default String Value:
  Valid Characters:
DStorwize V7000 HyperSwapription: Enable hardware accelerated VMFS data movement
(requires compliant hardware)
```

```
esxcli system settings advanced list -o /VMFS3/HardwareAcceleratedLocking
Path: /VMFS3/HardwareAcceleratedLocking
Type: integer
  Int Value: 1
  Default Int Value: 1
  Min Value: 0
  Max Value: 1
  String Value:
  Default String Value:
  Valid Characters:
DStorwize V7000 HyperSwapription: Enable hardware accelerated VMFS locking
(requires compliant hardware)
```

```
esxcli system settings advanced list -o /DataMover/HardwareAcceleratedInit
Path: /DataMover/HardwareAcceleratedInit
  Type: integer
  Int Value: 1
```

**Default Int Value: 1**

Min Value: 0

Max Value: 1

String Value:

Default String Value:

Valid Characters:

DStorwize V7000 HyperSwapription: Enable hardware accelerated VMFS data initialization (requires compliant hardware)

## 5.8 vCenter Services protection

Secure the VMs from hardware failure by running vCenter as a VM on the Primary data center. VMware has a list of possible solutions to secure this important component. This VMware Knowledge Base article, *Supported vCenter Server high availability options*, 1024051, provides an overview of the options:

<https://ibm.biz/BdXL4t>

For more information, see the *VMware vCenter Server 6.0 Availability Guide*:

<https://ibm.biz/BdXLte>

Figure 5-26 shows the vCenter supported HA solutions from VMware.

	Supported High Availability Solutions					
	vSphere HA	vSphere FT	WSFC/MSCS for VCDB	WSFC/MSCS for vCenter Server	vCenter Server Heartbeat	vCenter Server Watchdog
4.x	Yes <sup>1</sup>	No	No	No	Yes <sup>5</sup>	No
5.0	Yes <sup>1</sup>	No	No	No	Yes <sup>5</sup>	No
5.1	Yes <sup>1</sup>	No	No	No	Yes <sup>5</sup>	No
5.5	Yes <sup>1</sup>	No	Yes <sup>4</sup>	Yes <sup>7</sup>	Yes <sup>5</sup>	No
6.0	Yes <sup>1</sup>	Yes <sup>2</sup>	Yes	Yes <sup>6</sup>	No	Yes <sup>3</sup>

Figure 5-26 vCenter failover solution support matrix

Based on the size of the solution, VMware has a table of how to scale the solution, as shown in Figure 5-27.

Size	vCPU	vRAM (GB)	Hosts (Max)	VMs (Max)
Tiny	2	8	20	400
Small	4	16	150	3k
Medium	8	24	300	6k
Large	16	32	1k	10k

Figure 5-27 vCenter scaling

**Note:** It is not supported to use heartbeat by using vCenter 6.0. For medium to large solutions, you need to implement Windows Server Failover Clustering (WSFC), which is described next.

For small solutions, use the vCenter appliance and enable VMware vSphere Fault Tolerance (FT) as a good alternative.

### 5.8.1 vCenter availability solution: Windows Server Failover Clustering

You can also use a vCenter availability-based solution, such as Microsoft Windows Server Failover Clustering (WSFC). *The Setup for Failover Clustering and Microsoft Cluster Service* VMware guide at the following link shows how to build a cluster by using Microsoft Cluster technologies to provide the best failover solution, and to protect the vCenter components:

<https://ibm.biz/BdXLtJ>

Figure 5-28 shows the VMware best practices guide to build a cluster by using Microsoft Cluster.

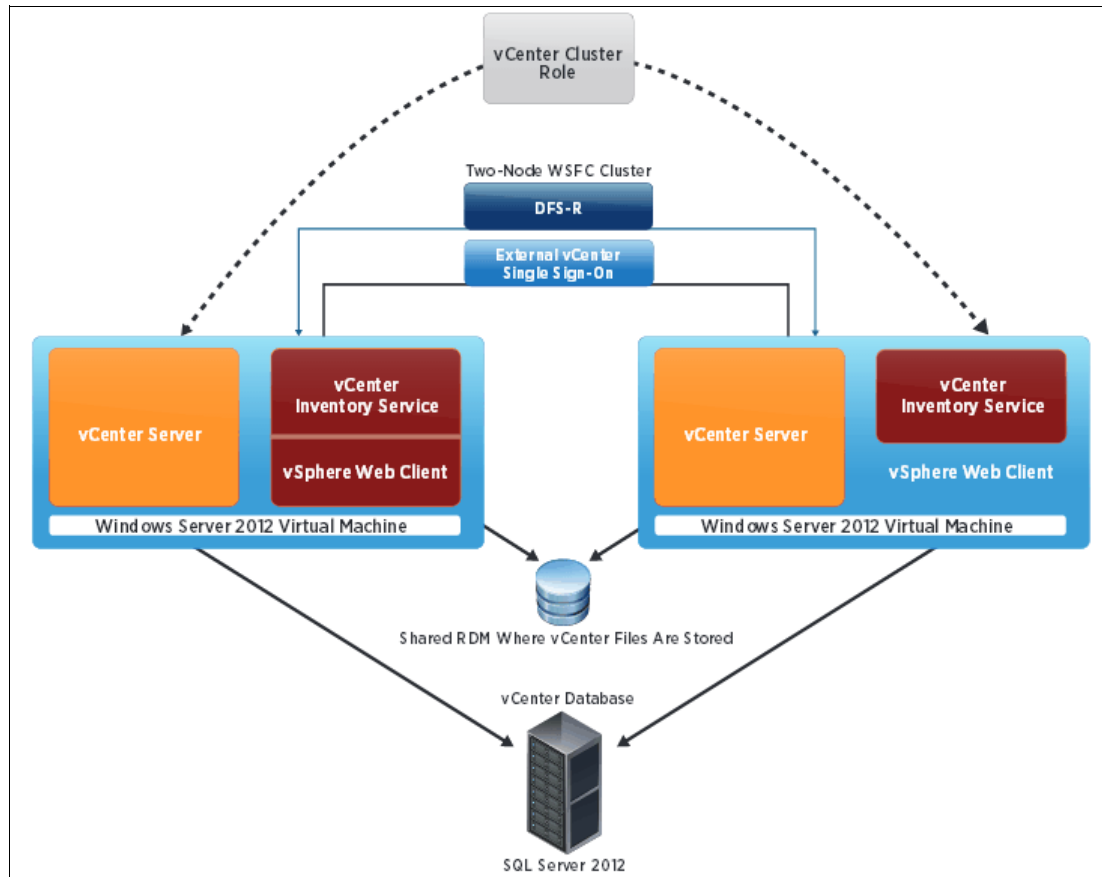


Figure 5-28 WSFC Microsoft Cluster running VMware vCenter

## 5.9 VMware recovery planning

If the implementation guidelines are followed and the inventory database is up-to-date, recovery depends on the situation.

The VMware environment can be documented in several ways. You can use PowerShell or even a product, such as RVTools, to extract all of the vital data from the vCenter database to comma-separated values (CSV) files. You can use these files when you are planning for recovery. You can also use these files to detect connections, and even missing relationships, in the virtual infrastructure.

Recovery is more than merely getting things started. It is getting them started in a way that the infrastructure can start to function. Therefore, categorization of VMs is important. At a minimum, complete these tasks:

- ▶ Extract data from vCenter (by using RVtools or another tool), and save this data to separate media. Schedule and save this data to separate media for extraction at least twice a week.
- ▶ Categorize the entire environment and the VMs in visible folders, and ensure that restart priorities are clearly understood and documented.
- ▶ Create and enforce naming standards. For more information, see 5.5, “Naming conventions” on page 119.

All of these tasks are in addition to normal, day-to-day planning, such as backup operations. Floor space management is also important so that you know the physical location of servers to start, and in the correct order.

Print the basic Internet Protocol (IP) plan for reference or save it to a device that does not require a network to be available. In a worst case scenario, you might not have anything to work from except that piece of paper and several closed servers on the floor.

### 5.9.1 VMware alternatives to minimize the impact of a complete site failure (split-brain scenario)

To help bring back vital VMs after a site failure, consider including vCenter Failover Method as Storwize V7000 HyperSwap and vCenter site Recovery Manager in your plan.

For more information, see vCenter Failover Method as Storwize V7000 HyperSwap in 5.8, “vCenter Services protection” on page 132.

#### **vCenter site Recovery Manager**

vCenter Site Recovery Manager (SRM) is a software suite that takes care of an entire site failure scenario. These products can be used in different scenarios and become a part of any recovery plan. SRM interacts with vReplicator, if used.

A split brain is controlled by VMware HA, where the master ESX host is the controlling part in each ESX Cluster. For this reason, use VMCP. With VMCP in aggressive mode, the split brain will be solved because the VMware master node will ensure that only one instance of the VM is running.

For more information, see 5.6.6, “Virtual Machine Component Protection (VMCP)” on page 127.

## 5.9.2 Investigating a site failure

Use these suggestions to investigate a site failure:

- ▶ Because a site failure is a critical situation, you must first determine the root cause and identify the nature of the ongoing situation.
- ▶ VMware offers many Knowledge Base (KB) articles about investigating HA failures. Because a HA failure is always a critical situation, you must always create an VMware support case as the first step when you start to investigate a failure. For several good links to guides that describe various scenarios and methods, see *Troubleshooting VMware High Availability (HA) in VMware vSphere*, 1001596:  
<https://ibm.biz/BdXCgD>
- ▶ For information to re-create your ESX Cluster if it is broken, see *Recreating a VMware High Availability Cluster in vSphere*, 1003715:  
<https://ibm.biz/BdXCg4>
- ▶ A SYSLOG tool that collects and analyzes the data, such as the license-based IBM VMware Log Insight™, can help with the analysis of the situation.
- ▶ Storwize V7000 HyperSwap recovers from site failures and ensures that you still have one site that can run the environment.
- ▶ Create a prioritized plan.

**Prioritized plan:** The following tasks are recommended:

- ▶ Create a VMware support case.
  - ▶ Ensure that the ESXi management cluster is running first, if applicable.
  - ▶ Ensure that vCenter is operational. If vRecovery is running, use vRecovery to activate the copy on the remaining site.
  - ▶ Start VMs, such as data centers, DNS, and the Windows Internet Name Service (WINS) server.
- ▶ Ensure that storage is running at the site, as expected.
  - ▶ Investigate whether any of the datastores are offline, inactive, or not functional for any reason before you start any VMs.

Depending on the situation, you need to look at the vCenter log files and the ESX log files. For a link to all VMware products and descriptions and where to find corresponding log files to each product, see *Location of log files for VMware products*, 1021806:

<https://ibm.biz/BdXCgy>

Investigate these logs:

- ▶ The Windows version of the VMware vCenter Server 6.0 logs are in the %ALLUSERSPROFILE%\VMWare\vCenterServer\logs folder.
- ▶ The VMware vCenter Server Appliance 6.0 logs are in the /var/log/vmware/ folder.
- ▶ vCenter log file locations are shown in Figure 5-29.

vCenter Server	vCenter Server Appliance	Description
vmware-vpx\vpxd.log	vpxd/vpxd.log	The main vCenter Serverlog
vmware-vpx\vpxd-profiler.log	vpxd/vpxd-profiler.log	Profile metrics for operations performed in vCenter Server
vmware-vpx\vpxd-alert.log	vpxd/vpxd-alert.log	Non-fatal information logged about the vpxd process
perfcharts\stats.log	perfcharts/stats.log	VMware Performance Charts
eam\eam.log	eam/eam.log	VMware ESX Agent Manager
invsvc	invsvc	VMware Inventory Service
netdump	netdumper	VMware vSphere ESXi Dump Collector
vapi	vapi	VMware vAPI Endpoint
vmdird	vmdird	VMware Directory Service daemon
vmsyslogcollector	syslog	vSphere Syslog Collector
vmware-sps\sps.log	vmware-sps/sps.log	VMware vSphere Profile-Driven Storage Service
vpostgres	vpostgres	vFabric Postgres database service
vsphere-client	vsphere-client	VMware vSphere Web Client
vws	vws	VMware System and Hardware Health Manager
workflow	workflow	VMware vCenter Workflow Manger
sso	sso	VMware Single Sign-On

Figure 5-29 vCenter log files

- ▶ For more information about vCenter log files, see *Location of VMware vCenter Server 6.0 log files*, 2110014:

<https://ibm.biz/BdXCggs>

- ▶ Monitor the ESXi host log files, either from Secure Shell (SSH) directly or through vCenter, if vCenter is still running. You can find the host log files by following the link under /var/log:
  - /scratch/log/vmkernel.log
  - cd/scratch/log/vmkernel.log
- ▶ Look for the SCSI Sense Codes in /var/log/vmkernel.log and review several of them. If the vmkernel.log file was compressed, look into the log with the **zcat** command rather than the **cat** command. See Example 5-10.

*Example 5-10 cat /var/log/vmkernel.log | grep "Valid sense data:"*

```
cat /var/log/vmkernel.log | grep "Valid sense data:"
zcat vmkernel.1.gz | grep "H:0x0 D:0x2 P:0x0 Valid sense data: 0x5 0x25 0x0"
satp_alua_issueCommandOnPath:665: Path "vmhba4:C0:T0:L1" (PERM LOSS) command
0xa3 failed with status Device is permanently unavailable. H:0x0 D:0x2 P:0x0
Valid sense data: 0x5 0x25 0x0.
```



In an APD situation, you cannot recover automatically. The situation needs to be resolved at the storage array fabric layer to restore connectivity to the host.

Optionally, all ESXi hosts that were affected by the APD might require a reboot. Table 5-6 shows the sense code for permanent device loss (PDL).

APD or PDL in vSphere 6 is optimized to react on these failures, as described for Storwize V7000 HyperSwap in 5.6.4, “Enhanced All Paths Down detection in vSphere 6” on page 125.

Table 5-6 Sense Code for PDL

Valid sense data	Message	
H:0x0 D:0x2 P:0x0 Valid sense data: 0x5 0x25 0x0	LOGICAL UNIT NOT SUPPORTED	Seen in ALUA-SATP when PDL is current
H:0x0 D:0x2 P:0x0 Valid sense data: 0x4 0x4c 0x0	LOGICAL UNIT FAILED	Not seen
H:0x0 D:0x2 P:0x0 Valid sense data: 0x4 0x3e 0x3	LOGICAL UNIT FAILED SELF-TEST	Not seen
HH:0x0 D:0x2 P:0x0 Valid sense data: 0x4 0x3e 0x1	LOGICAL UNIT FAILURE	Not seen

## 5.10 Design comments

When you create a Storwize V7000 HyperSwap with a VMware stretched cluster, you combine options to prevent a disaster and allow access to the data stores across the clusters. With this configuration, you can access failover or vMotion instantly, without rezoning your SAN disk or switching to the mirrored copy of the disk.

A VMware stretched cluster with a Storwize V7000 HyperSwap is managed best by implementing rules that, under normal operation, bind the VMs to each data center that is part of a Storwize V7000 HyperSwap. Use affinity rules to secure it, and use vMotion to prevent disasters and to balance loads.

This solution is uniform, which means that all disks are visible to all ESXhosts within the same ESX Cluster.

Due to the nature of Storwize V7000 HyperSwap, where it potentially can swap the Primary site due to a changed load, the best way to prevent a swap is for the DRS groups to be in alignment with the site awareness.

Use vMSC in enterprise solutions, where distance is the key factor and the actual calculation is measured in ms. From the VMware perspective, the solution is accepted up to 10 ms round-trip time (RTT). However, in this example, Storwize V7000 HyperSwap is based on a 3 ms solution, for a maximum distance of 300 km (186.4 miles).

The only limit is the response times to the disks that are at the remote site. These response times can be controlled through the preferred node path. The key to success is to keep these paths under control and to monitor and validate the affinity rules during daily operation.

Even though VMware in vSphere 6.0 expanded vMotion capabilities by using Long Distance vMotion up to 100 ms, this solution cannot support more than a 5 ms delay due to the 300 km (186.4 miles) distance of the fiber optic cable.

## 5.11 Script examples

The following script examples can help test VM mobility and check the preferred path policy.

Because we plan to operate on a large scale, we need to be able to detect and organize the disks quickly. The fastest way to detect and organize the disks quickly is to use vSphere PowerCLI or Microsoft PowerShell scripts.

Both scripts must be modified to suit your environment. Also, ensure that someone who is familiar with PowerShell implements the scripts.

**Reminder:** You use these scripts at your own risk. But, because they do *not* change anything, they are basically harmless.

### 5.11.1 PowerShell test script to move VMs 40 times between two ESXi hosts

It is important to ensure that the infrastructure is stable by testing the capability to move VMs between the ESXhosts in the clusters. Use this sample script to automate the process.

The script requires a PowerShell working environment with installed VMware automation tools. The script can then be run from the PowerCLI shell after you copy the entire script into a file.

**Tip:** Look for #[Change], which indicates where you must change the script to match the names in your environment.

Example 5-11 shows the vMotion test script.

*Example 5-11 vMotion test script*

```
##### vMotion test script #####
## powerShell Script vMotions Tester
#####
function migrateVM
{ param ($VMn, $dest)
  get-vm -name $VMn | move-vm -Destination (get-vmhost $dest) }
#[Change]name here to your testing VM in vCenter
$vm = "Your-test-vm"
#[Change] the names here of your two ESXhost @ each site.
$dest1 = "esxi-01-dca"
$dest2 = "esxi-02-dcb"
$cnt = 0
$NumberOfvMotions = 40 (will be 40, because we start from 0)

#[Change] the name here to your vCenter:
Connect-VIServer -server "vCenterDCA"

## Perform 40 vMotions between 2 ESXhosts in each datacenter
do {
#
# Get VM information and its current location
#
  $vmname = get-vmhost -VM $vm -ErrorAction SilentlyContinue
```

```

    if ( $vmname.Name -eq $dest1 ) { $mdest = $dest2 }
    else { $mdest = $dest1 }
    #
    # Migrate VM to Destination
    #
    write-output "Performing Migration # +$cnt+ To: $mdest"
    migrateVM $vm $mdest
    $cnt++
# If fast vMotion, you can lower this A bit , 300 = 5 min.
    start-sleep -s 300
    }
while ($cnt -lt $NumberOfVMotions)

```

---

## 5.11.2 PowerShell script to extract data from the entire environment to verify active and preferred paths

This script extracts the preferred path and active state on datastores into CSV files to verify the actual preferred path settings on all HBAs and datastores. The script does not modify or set anything. It merely writes the settings to a CSV file.

**Note:** Look for #[Change], which indicates where you must change the script to match the names in your environment.

Example 5-12 shows the script.

*Example 5-12 Script to verify active and preferred paths*

---

```

# Copy & Paste the entire Text into a Text file
#[Change] the name here: "IBMSE-VC-VCS.IBMSE.LOCAL" to the name of your Virtual
Center.
#[FILEPATH]
# The Script collects all valuable Storage and Path info
# Author: Ole Rasmussen, IBM Denmark for ITSO-Redbooks
# Version 2.0 Storwize V7000 HyperSwap Version with ALUA
# ##### after 7.5 we don't set Preferred Path, we rely on Host Site awareness.
# # Step 2: Run this Script and Verify that ALL paths are set to Fixed.
# Step 3: Then Use the OutCome as Role model for upcoming Luns, copy or rename the
ReportOnLunPath.csv to Master.CSV
#
##OBS remember that the name of the DATAStore itself points to where the Vdisk
Primary Copy is, NOT where the path should be
##
## In a Storwize V7000 HyperSwap cluster where Latency is more than 0.5 ms, or
latency is visible to performance in the cluster.
## ESXC_00_VMFS_V_DCA_01 --primary Copy on DCA
## ESX host resides on DCB --> (Via name in host)
## i.e esxi-02-dcb.ibmse.local --> Points to use SVC Node:
50:05:07:68:01:40:B1:3F in DCB
## On Path
fc.20008c7cff0ad700:10008c7cff0ad700-fc.500507680100b13f:500507680140b13f-naa.6005
076801840542d80000000000000000

### Load VMware Library .
Add-Pssnapin VMware.VimAutomation.Core -Erroraction SilentlyContinue

```

```
Add-Pssnapin VMware.Vumautomation -Erroraction Silentlycontinue
```

```
# Force To Load VMware PowerShell Plugin  
[Reflection.Assembly]::Loadwithpartialname("VMware.Vim")
```

```
# On Error Continue  
$Erroractionpreference = "Continue"  
#####  
#[CHANGE] Edit the folder names & report-Folder will be target for a Folder  
named Date & time of execution,  
$reportDirectory = "C:\Redbook Scripts\Report-Folder"  
$TimeStamp = Get-Date -UFormat %Y-%m-%d-%H%M  
$Folder = $reportDirectory+"\\"+$TimeStamp  
mkdir $folder| Out-Null  
$reportDirectory = $folder ## Actual including a timestamp for running the  
report.
```

```
##### [CHANGE] Connect to Virtual Center  
$vi = Connect-VIServer "IBMSE-VC-VCS.IBMSE.LOCAL" -ErrorAction:Stop  
if (!$?) {  
    Write-host -BackgroundColor DarkYellow -ForegroundColor DarkRed  
    "Could not connect to Virtualcenter"  
    $noError = $false  
    Break  
}
```

```
##### Report Array Unit's
```

```
$ReportLUN = @()
```

```
$NotFixedDisks = @()  
$MruDisks = @()  
$ALLdisks= @()  
$FixedDisks = @()
```

```
### Get All Esxhost  
$ESXhosts = Get-VMhost | where {$_.State -ne "Maintenance" -and $_.State -eq  
"Connected" } ## Only host not in MT mode , change -EQ if only in MT
```

```
##### Export of Different Raw data of Disk PSP Settings
```

```
#  
# If you don't want them, add # in front  
#####  
$FixedDisks= $esxhosts | Get-ScsiLun -LunType "disk" | where {$_.MultipathPolicy  
-eq "Fixed"}  
$NotFixedDisks= $ESXhosts | Get-ScsiLun -LunType "disk" | where  
{$_.MultipathPolicy -eq "RoundRobin"}  
$MruDisks = $ESXhosts | Get-ScsiLun -LunType "disk" | where {$_.MultipathPolicy  
-eq "MostRecentlyUsed"}  
#$ALLdisks = $ESXhosts | Get-ScsiLun -LunType "disk"
```

```

$DatastoreGather = @() ## To Fetch the Name of the Datastore for later compare
with ID: naa.XXXXXXXXXXX

#
# Use Datastore view to get All Datastores in the Cluster
### OBS OBS## If you want to get all Disk and not only VMFS, change below line to
this:
#$dsView = Get-Datastore | get-view
$dsView = Get-Datastore | where {$_ .Type -eq "VMFS"} | get-view ## ONLY VMFS
datastores, not RAW Device,
$DatastoreGather = @()
$DataCounter = 0
$DatastoresTotal = $Dsview.Length

ForEach ($DS in $Dsview)
{
    $DataCounter++
    Write-Progress -Activity " " -Status "Find ALL Datastores on VC" -Id 1
-PercentComplete (100*$DataCounter/$DatastoresTotal)
    $DatastoreObject = "" | Select-Object Datastore, canonicalName, OverAllStatus
    $Datastoreobject.canonicalName = $DS.Info.Vmfs.extent[0].Diskname
    $Datastoreobject.Datastore = $DS.Info.Vmfs.name
    $Datastoreobject.OverallStatus = $DS.OverallStatus
    $DatastoreGather += $DatastoreObject
}

# Get all ESXhost in Virtual Center, and Not those in Maintenance mode, and only
Connected.
#
#

$NumberOfESXhosts = $ESXhosts.Length
$ESXhostCounter = 0

foreach($esx in $ESXhosts){
    $ESXhostCounter++
    Write-Progress -Activity " " -Status "ESXhost [#of $NumberOfESXhosts] Activity
progress ..." -Id 2 -PercentComplete (100*$ESXhostCounter/$NumberOfESXhosts)
    ## Only Getting Datastores of type DISK and No local disk (!
    $luns = Get-ScsiLun -VMhost $esx | Where-Object {$_ .luntype -eq "disk" -and
!$_ .IsLocal }
    $LUNsTotal = $LUNs.Length

    $LUNsCounter = 0

    ForEach ($LUN in $LUNs) {
        $lunsCounter++
        Write-Progress -Activity " " -Status "Lun on host [$LUNsTotal] -->
Activity Progress ..." -Id 3 -PercentComplete (100*$LunsCounter/$LUNsTotal)
        $lunPath = Get-ScsiLunPath -ScsiLun $lun
        $LUNID = $LUN.Id
    }
}

```

```

    $lunPathCounter = 0
    $LunpathTotal = $lunPath.length

    foreach ($Path in $lunPath) {
        $lunPathCounter++
        Write-Progress -Activity " " -Status "Path's on host
[$LunpathTotal] --> Activity Progress ..." -Id 4 -PercentComplete
(100*$Lunpathcounter/$LunpathTotal)
        $LUNInfo = "" | Select-Object ESXhost, LunCanonName, Datastore,
Datacenter, SVCNodeID, Policy, Prefer, ActiveState, VMHBAname, LUNPath, LUNID

        $LUNInfo.ESXhost = $esx.Name
        $LUNInfo.LunCanonName= $Path.ScsiCanonicalName
        $LUNInfo.Datastore = ($Datastoregather | where {$_ .canonicalName
-eq $LUNInfo.LunCanonName}).Datastore
        #if($esx.Name -clike "dcb") { Write-host "DCB"}
        #LUNInfo.Datacenter =

        $LUNInfo.SVCNodeID = $Path.SanId
        $LUNInfo.Policy = $path.ScsiLun.MultipathPolicy
        $LUNInfo.Prefer = $Path.Preferred
        $LUNInfo.ActiveState = $Path.State
        $LUNInfo.VMHBAname = $Path.ScsiLun.RuntimeName
        $LUNInfo.LUNPath = $Path.LunPath
        $LUNInfo.LUNID = $LUNID

        $ReportLUN += $LUNInfo
    }
} ## End LUN Loop
###

} ##End ## ESXhosts Loop

```

**Write-host** -ForegroundColor DarkYellow -BackgroundColor Black "Completed all collection of Data: "

```

##### rename Target CSV file #####
##[FILEPATH]
#Change the name of the File: the Delimiter if not an #";"
$reportCount = 5
if ($reportLun) {$reportLun| Export-Csv -Path
"$reportDirectory\ReportOnLunPath.csv" -Delimiter "," }
if ($FixedDisks) { $FixedDisks| Export-Csv -Path
"$reportDirectory\Fixed-Disks.csv" -Delimiter ","}
if( $NotFixedDisks) {$NotFixedDisks| Export-Csv -Path
"$reportDirectory\NotFixed-Disks.csv" -Delimiter ","}
if ($MruDisks) {$MruDisks| Export-Csv -Path "$reportDirectory\MRU-Disks.csv"
-Delimiter ","}
if ($ALLdisks ) {$ALLdisks| Export-Csv -Path
"$reportDirectory\ALLFixed-Disks.csv" -Delimiter ","}

```

---









## Storwize V7000 HyperSwap diagnostic and recovery guidelines

This chapter addresses IBM Storwize V7000 HyperSwap diagnostic and recovery guidelines. These features help you understand what is happening in your Storwize V7000 HyperSwap environment after a critical event. This knowledge is crucial when you decide to alleviate the situation. You might decide to wait until the failure in one of the two sites is fixed or to declare a disaster and start the recovery action.

All of the operations that are described are guidelines to help you in a critical event or a rolling disaster. Several of the operations are specific to our lab environment and several of the operations are common with every Storwize V7000 HyperSwap installation.

Before you start a recovery action after a disaster is declared, it is important that you are familiar with all of the recovery steps that you will follow. We strongly suggest testing and documenting a recovery plan that includes all of the tasks that you must perform, according to the design and configuration of your environment. It is also best to execute the recovery action with IBM Support engaged.

This chapter includes the following sections:

- ▶ Solution recovery planning
- ▶ Storwize V7000 recovery planning
- ▶ Storwize V7000 HyperSwap diagnosis and recovery guidelines
- ▶ Other disaster recovery with HyperSwap

## 6.1 Solution recovery planning

In the context of the Storwize V7000 HyperSwap environment, solution recovery planning is more application-oriented. Therefore, any plan must be made with the client application's owner. In every IT environment, when a business continuity or disaster recovery (DR) solution is designed, incorporate a solution recovery plan into the process.

It is imperative to identify high-priority applications that are critical to the nature of the business. Then, create a plan to recover those applications, in tandem with the other elements that are described in this chapter.

## 6.2 Storwize V7000 recovery planning

To achieve the most benefit from the Storwize V7000 HyperSwap configuration, post-installation planning must include several important steps. These steps ensure that your infrastructure can be recovered with either the same or a different configuration in one of the surviving sites with minimal impact for the client applications. Correct planning and configuration backup also help to minimize possible downtime.

You can categorize the recovery in the following ways:

- ▶ Recover a fully redundant Storwize V7000 HyperSwap configuration in the surviving site without HyperSwap.
- ▶ Recover a fully redundant Storwize V7000 HyperSwap configuration in the surviving site with HyperSwap implemented in the same site or on a remote site.
- ▶ Recover according to one of these scenarios, with a fallback chance on the original recovered site after the critical event.

Regardless of which scenario you face, apply the following guidelines.

To plan the Storwize V7000 HyperSwap configuration, complete these steps:

1. Collect a detailed Storwize V7000 HyperSwap configuration. To do so, run a daily Storwize V7000 HyperSwap configuration backup with the command-line interface (CLI) commands that are shown in Example 6-1.

*Example 6-1 Saving the Storwize V7000 configuration*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>svcconfig backup
.....
.....
CMMVC6155I SVCCONFIG processing completed successfully
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsdumps
id filename
0 reinst.7836494-1.trc
1 svc.config.cron.bak_7836494-2
.
.lines removed for brevity
.
40 svc.config.backup.xml_7836494-1
```

---

2. Save the .xml file that is produced in a safe place, as shown in Example 6-2.

*Example 6-2 Copying the configuration*

---

```
C:\Program Files\PuTTY>pscp -load V7K_HyperSwap
admin@10.17.89.251:/tmp/SVC.config.backup.xml_7836494-1
c:\temp\configbackup.xml
configbackup.xml | 97 kB | 97.2 kB/s | ETA: 00:00:00 | 100%
```

---

3. Save the output of the CLI commands that is shown in Example 6-3 in .txt format.

*Example 6-3 List of Storwize V7000 commands to issue*

---

```
lssystem
lssite
lsnodecanister
lsnodecanister <nodes name>
lsnodecanisterhw <nodes name>
lsiogrp
lsiogrp <iogrps name>
lscontroller
lscontroller <controllers name>
lsmdiskgrp
lsmdiskgrp <mdiskgrps name>
lsmdisk
lsquorum
lsquorum <quorum id>
lsvdisk
lshost
lshost <host name>
lshostvdiskmap
lsrcrelationship
lsrcconsistgrp
```

---

From the output of these commands and the .xml file, you have a complete picture of the Storwize V7000 HyperSwap infrastructure. Remember the Storwize V7000 HyperSwap ports' worldwide node names (WWNNs), so that you can reuse them during the recovery operation that is described in 6.3.3, "Storwize V7000 HyperSwap recovery guidelines" on page 160.

Example 6-4, which is contained in the .xml file, shows what you need to re-create a Storwize V7000 HyperSwap environment after a critical event.

*Example 6-4 XML configuration file example*

---

```
<xml
  label="Configuration Back-up"
  version="750"
  file_version="1.206.9.169"
  timestamp="2015/08/12 13:20:30 PDT" >

  <!-- cluster section -->

  <object type="cluster" >
    <property name="id" value="00000100216001E0" />
    <property name="name" value="ITS0_V7K_HyperSwap" />

  </object >
  .
  many lines omitted for brevity
  .
  <!-- controller section -->

  <object type="controller" >
    <property name="id" value="0" />
    <property name="controller_name" value="ITS0_V7K_Q_N1" />
    <property name="WWNN" value="5005076802000EF" />
    <property name="mdisk_link_count" value="2" />
    <property name="max_mdisk_link_count" value="2" />
    <property name="degraded" value="no" />
    <property name="vendor_id" value="IBM" />
    <property name="product_id_low" value="2145" />
    <property name="product_id_high" value="" />
    <property name="product_revision" value="0000" />
    <property name="ctrl_s/n" value="2076" />
    <property name="allow_quorum" value="yes" />
    <property name="fabric_type" value="fc" />
    <property name="site_id" value="3" />
    <property name="site_name" value="ITS0_SITE_Q" />
    <property name="WWPN" value="50050768021000EF" />
    <property name="path_count" value="0" />
    <property name="max_path_count" value="0" />
    <property name="WWPN" value="50050768022000EF" />
    <property name="path_count" value="0" />
    <property name="max_path_count" value="0" />
  </object >
  <object type="controller" >
    <property name="id" value="1" />
    <property name="controller_name" value="ITS0_V7K_Q_N2" />
    <property name="WWNN" value="5005076802000F0" />
    <property name="mdisk_link_count" value="2" />
```

```

    <property name="max_mdisk_link_count" value="2" />
    <property name="degraded" value="no" />
    <property name="vendor_id" value="IBM      " />
    <property name="product_id_low" value="2145  " />
    <property name="product_id_high" value="      " />
    <property name="product_revision" value="0000" />
    <property name="ctrl_s/n" value="2076      " />
    <property name="allow_quorum" value="yes" />
    <property name="fabric_type" value="fc" />
    <property name="site_id" value="3" />
    <property name="site_name" value="ITS0_SITE_Q" />
    <property name="WWPN" value="50050768021000F0" />
    <property name="path_count" value="8" />
    <property name="max_path_count" value="8" />
    <property name="WWPN" value="50050768022000F0" />
    <property name="path_count" value="8" />
    <property name="max_path_count" value="8" />
  </object >

```

*many lines omitted for brevity*

---

You can also get this information from the .txt command output that is shown in Example 6-5.

*Example 6-5 Isnodecanister example output command*

---

```

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsnodecanister ITS0_HS_SITE_A_N1
id 8
name ITS0_HS_SITE_A_N1
UPS_serial_number
WWNN 500507680B0021A8
status online
IO_group_id 0
IO_group_name io_grp0_SITE_A
partner_node_id 9
partner_node_name ITS0_HS_SITE_A_N2
config_node yes
UPS_unique_id
port_id 500507680B2121A8
port_status active
port_speed 4Gb
port_id 500507680B2221A8
port_status active
port_speed 4Gb
port_id 500507680B2321A8
port_status active
port_speed 2Gb
port_id 500507680B2421A8
port_status active
port_speed 2Gb
hardware 400
iscsi_name iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsohssitean1
iscsi_alias
failover_active no
failover_name ITS0_HS_SITE_A_N2
failover_iscsi_name iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsohssitean2
failover_iscsi_alias

```

```
panel_name 01-1
enclosure_id 1
canister_id 1
enclosure_serial_number 7836494
service_IP_address 10.18.228.55
service_gateway 10.18.228.1
service_subnet_mask 255.255.255.0
service_IP_address_6
service_gateway_6
service_prefix_6
service_IP_mode static
service_IP_mode_6
site_id 1
site_name ITS0_SITE_A
identify_LED off
product_mtm 2076-424
code_level 7.5.0.2 (build 115.51.1507081154000)
```

---

For more information about backing up your configuration, see the IBM Storwize V7000 Knowledge Center:

[https://www.ibm.com/support/knowledgecenter/ST3FR7\\_7.8.0/com.ibm.storwize.v7000.780.doc/svc\\_configbackupovr\\_1e4imh.html](https://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/svc_configbackupovr_1e4imh.html)

and

[https://www.ibm.com/support/knowledgecenter/ST3FR7\\_7.8.0/com.ibm.storwize.v7000.780.doc/svc\\_clustconfbackuptsk\\_1e4k69.html](https://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/svc_clustconfbackuptsk_1e4k69.html)

4. Create an up-to-date, high-level copy of your configuration that describes all elements and connections.
5. Create a standard labeling schema and naming convention for your Fibre Channel (FC) or Ethernet (ETH) cabling, and ensure that it is fully documented.
6. Back up your storage area network (SAN) zoning by using your FC switch CLI or graphical user interface (GUI).

The essential zoning configuration data, domain ID, zoning, alias, configuration, and zone set can be saved in a .txt file by using the output from the CLI commands. You can also use the appropriate utility to back up the entire configuration.

The following IBM b-type/Brocade FC switch or director commands are helpful to collect the essential zoning configuration data:

- switchshow
- fabricshow
- cfgshow

During the implementation, use WWNN zoning. During the recovery phase after a critical event, reuse the same domain ID and same port number that were used in the failing site, if possible. Zoning is propagated on each switch because of the SAN extension with inter-switch link (ISL). For more information, see 6.3.3, “Storwize V7000 HyperSwap recovery guidelines” on page 160.

For more information about how to back up your FC switch or director zoning configuration, see your switch vendor’s documentation.

7. Back up your back-end storage subsystems configuration.

In your Storwize V7000 HyperSwap implementation, potentially you can also virtualize the external storage controller.

If you virtualized the external storage controller, back up your storage subsystem configuration. This way, if a critical event occurs, you can re-create the same environment when you reestablish your V7000 HyperSwap infrastructure in a different site with new storage subsystems.

For more information, see 6.3.3, “Storwize V7000 HyperSwap recovery guidelines” on page 160. Back up your storage subsystem in one of the following ways:

- For the IBM DS8000® storage subsystem, save the output of the DS8000 CLI commands in .txt format, as shown in Example 6-6.

*Example 6-6 DS8000 commands*

---

```
lsarraysite -l
lsarray -l
lsrank -l
lsextpool -l
lsfbvol -l
lshostconnect -l
lsvolgrp -l
showvolgrp -lunmap <SVC vg_name>
```

---

- For the IBM XIV® Storage System, save the output of the XCLI commands in .txt format, as shown in Example 6-7.

*Example 6-7 XIV subsystem commands*

---

```
host_list
host_list_ports
mapping_list
vol_mapping_list
pool_list
vol_list
```

---

- For IBM Storwize V7000, collect the configuration files and the output report as described in 6.2, “Storwize V7000 recovery planning” on page 146.
- For any other supported storage vendor’s products, see their documentation for Storwize V7000 MDisk configuration and mapping.

## 6.3 Storwize V7000 HyperSwap diagnosis and recovery guidelines

This section provides guidelines for diagnosing a critical event in one of the two sites where the Storwize V7000 HyperSwap is implemented. With these guidelines, you can determine the extent of any damage, what is still running, what can be recovered, and with what impact on the performance.

### 6.3.1 Critical event scenarios and complete site or domain failure

Many critical event scenarios might occur in a Storwize V7000 HyperSwap environment. Certain events can be handled by using standard (*business as usual*) recovery procedures. This section addresses all of the required operations to recover from a *complete site failure*.

Certain parts of the recovery depend on the environment design. This section shows the actions to diagnose the situation and, later, to recover the Storwize V7000 HyperSwap. Most of the steps are basic and they can be used in every environment and configuration, though.

**Important:** Because of the importance of the success of the recovery action, we suggest that you do not improvise this action and perform all steps under the direction of IBM Support.

The following list includes several scenarios that you might face and their required recovery actions:

- ▶ Back-end storage box failure in one failure domain: Only if you are virtualizing the external storage controller. Use business as usual recovery because of Storwize V7000 HyperSwap active-active Metro Mirror.
- ▶ Partial SAN failure in one failure domain: Use business as usual recovery because of SAN resilience.
- ▶ Total SAN failure in one failure domain: Use business as usual recovery because of SAN resilience, but pay attention to the performance impact. You need to act to minimize the impact on applications.
- ▶ Storwize V7000 HyperSwap node canister failure in one failure domain: Use business as usual recovery because of Storwize V7000 high availability (HA).
- ▶ Complete site failure in one failure domain: For example, a rolling disaster that first affects the connectivity between the sites and later destroys one entire site and the connectivity to Site3, where the active quorum disk exists. This recovery is covered in the scope of 6.3.2, “Storwize V7000 HyperSwap diagnosis guidelines” on page 152. All of the required operations are described, starting with 6.3.2, “Storwize V7000 HyperSwap diagnosis guidelines” on page 152.

### 6.3.2 Storwize V7000 HyperSwap diagnosis guidelines

This section provides guidelines about how to diagnose a critical event in one of the two sites where the Storwize V7000 HyperSwap configuration is implemented.



The Storwize V7000 HyperSwap configuration that is used in the examples is shown in Figure 6-1.

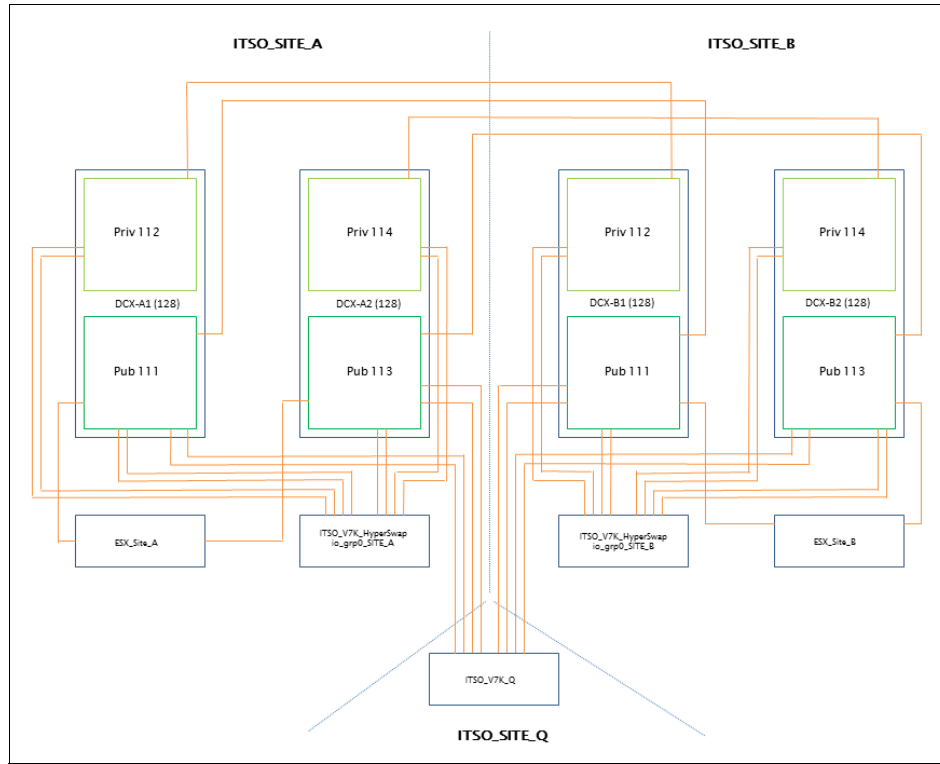


Figure 6-1 Environment diagram

The configuration that is implemented must be consistent with one of the supported configurations that are described in 3.4, “Storwize V7000 HyperSwap configurations” on page 42.

**Note:** Figure 6-1 shows Quorum Disk on ITS0\_Site\_Q. In our example later in this chapter the IP Quorum has been implemented.

### Up-and-running scenario analysis

In this scenario, all components are running and all of the guidelines were applied when we implemented the solution, as shown in the CLI command output in Example 6-8.

#### Example 6-8 Running example

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lssystem
id 00000100214001E0
name ITS0_V7K_HyperSwap
.
lines omitted for brevity
.
code_level 7.8.0.0 (build 133.14.1610192015000)
.
lines omitted for brevity
.
layer replication
.
lines omitted for brevity
```



```

2 Quorum Pool SITE_Q online 1 0 0 1024 0
0.00MB 0.00MB 0.00MB 0 80 auto
balanced no 0.00MB 0.00MB
0.00MB 2 Quorum Pool SITE_Q 0
0.00MB parent no none 3 ITS0_SITE_Q
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsmdisk
id name status mode mdisk_grp_id mdisk_grp_name capacity
ctrl_LUN# controller_name UID
tier encrypt site_id site_name
0 mdisk_1A online array 0 Pool_SITE_A 3.8TB
enterprise no 1 ITS0_SITE_A
1 mdisk_1B online array 1 Pool_SITE_B 3.8TB
enterprise no 2 ITS0_SITE_B
2 mdisk0_V7K_HS_Q online managed 2 Quorum Pool SITE_Q 1.0GB
0000000000000000 ITS0_V7K_Q_N1
60050768028a000268000000000000000000000000000000000000000000000000 enterprise no
3 ITS0_SITE_Q
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsvdisk
id name IO_group_id IO_group_name status mdisk_grp_id
mdisk_grp_name capacity type FC_id FC_name RC_id RC_name vdisk_UID
fc_map_count copy_count fast_write_state se_copy_count RC_change
compressed_copy_count parent_mdisk_grp_id parent_mdisk_grp_name formatting encrypt
volume_id volume_name function
0 HyperSwap_Volume_ESX_B_M 0 io_grp0 online 1 Pool0
2.00GB striped many many 0 rcrc10 600507640087800780000000000000005C 2
1 not_empty 0 no 0 1
Pool0 yes no 0 HyperSwap_Volume_ESX_B_M master
1 HyperSwap_Volume_ESX_B_A 1 io_grp1 offline 0 Pool1
2.00GB striped many many 0 rcrc10 600507640087800780000000000000005D 2
1 not_empty 0 no 0 0
Pool1 yes no 0 HyperSwap_Volume_ESX_B_M aux
2 HyperSwap_Volume_ESX_B_Mcv 0 io_grp0 online 1 Pool0
2.00GB striped many many 0 rcrc10 600507640087800780000000000000005E 2
1 empty 1 yes 0 1
Pool0 no no 0 HyperSwap_Volume_ESX_B_M
master_change
3 HyperSwap_Volume_ESX_B_Acv 1 io_grp1 online 0 Pool1
2.00GB striped many many 0 rcrc10 600507640087800780000000000000005F 2
1 not_empty 1 yes 0 0
Pool1 no no 0 HyperSwap_Volume_ESX_B_M
aux_change
42 HyperSwap_Volume_ESX_A_M 1 io_grp1 online 0 Pool1
2.00GB striped many many 42 Rel_ESX_A 6005076400878007800000000000000034 2
1 empty 0 no 0 0
Pool1 no no 42 HyperSwap_Volume_ESX_A_M master
43 HyperSwap_Volume_ESX_A_A 0 io_grp0 offline 1 Pool0
2.00GB striped many many 42 Rel_ESX_A 6005076400878007800000000000000035 2
1 empty 0 no 0 1
Pool0 no no 42 HyperSwap_Volume_ESX_A_M aux
44 HyperSwap_Volume_ESX_A_Mcv 1 io_grp1 online 0 Pool1
2.00GB striped many many 42 Rel_ESX_A 6005076400878007800000000000000036 2
1 empty 1 yes 0 0
Pool1 no no 42 HyperSwap_Volume_ESX_A_M
master_change

```

```

45 HyperSwap_Volume_ESX_A_Acv 0          io_grp0      online 1          Pool0
2.00GB striped many many 42 Re1_ESX_A 60050764008780078000000000000037 2
1          empty          1          yes          0          1
Pool0          no          no          42          HyperSwap_Volume_ESX_A_M
aux_change
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsquorum
quorum_index status id name controller_id controller_name active object_type
override site_id site_name
0          online 11          no          drive          no
1          Site_A
1          online 5          no          drive          no
2          Site_B
3          online          yes          device          no
ITS0-1.englab.brocade.com/10.18.228.170

```

---

From the Storwize V7000 CLI command output that is shown in Example 6-8 on page 153, you can see these characteristics of the configuration:

- ▶ The Storwize V7000 HyperSwap system is accessible through the CLI, and it is in the hyperswap topology.
- ▶ The Storwize V7000 node canisters are online, and one of them is the configuration node.
- ▶ The I/O Groups are in the correct state.
- ▶ The managed disk (MDisk) groups are online.
- ▶ The MDisks are online.
- ▶ The volumes are online.
- ▶ The three quorum disks and IP Quorum are in the correct states.

Now, check the active-active Metro Mirror volume relationship and Consistency Groups' status by running a CLI command as shown in Example 6-9.

*Example 6-9 Active-active Metro Mirror relationship status check*

---

```

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsrcrelationship
id name          master_cluster_id master_cluster_name master_vdisk_id
master_vdisk_name aux_cluster_id aux_cluster_name aux_vdisk_id
aux_vdisk_name primary consistency_group_id consistency_group_name state
bg_copy_priority progress copy_type cycling_mode freeze_time
0 Re1_ESX_B 0000010021E001E0 ITS0_V7K_HyperSwap 0
HyperSwap_Volume_ESX_B_M 0000010021E001E0 ITS0_V7K_HyperSwap 1
HyperSwap_Volume_ESX_B_A master 0 CG_ESX_BtoA
consistent_synchronized 50 activeactive
42 Re1_ESX_A 0000010021E001E0 ITS0_V7K_HyperSwap 42
HyperSwap_Volume_ESX_A_M 0000010021E001E0 ITS0_V7K_HyperSwap 43
HyperSwap_Volume_ESX_A_A master 0 CG_ESX_AtoB
consistent_synchronized 50 activeactive
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsrcconsistgrp
id name          master_cluster_id master_cluster_name aux_cluster_id
aux_cluster_name primary state relationship_count copy_type
cycling_mode freeze_time
0 CG_ESX_AtoB 0000010021E001E0 ITS0_V7K_HyperSwap 0000010021E001E0
ITS0_V7K_HyperSwap master consistent_synchronized 1 activeactive
1 CG_ESX_BtoA 0000010021E001E0 ITS0_V7K_HyperSwap 0000010021E001E0
ITS0_V7K_HyperSwap master consistent_synchronized 1 activeactive

```

---

From the Storwize V7000 CLI command output in Example 6-9, you can see that the Metro Mirror active-active relationship is in the `consistent_synchronized` state.

If you need to check several volumes, you can create a customized script directly from the Storwize V7000 command shell. Useful scripts are at the Storwize Scripting Wiki on IBM developerWorks:

<http://ibm.co/1hdCYkA>

### Critical event scenario analysis

In this scenario, the Storwize V7000 HyperSwap environment experienced a critical event that caused the complete loss of Site1 (ITSO\_SITE\_A).

Follow these steps to get a complete view of any damage and to gather enough information about key elements to determine your next recovery actions:

1. Is Storwize V7000 system management available through the GUI or CLI?
  - Yes: Go to Step 2.
  - No: Try to fix the problem by following standard troubleshooting procedures. For more information, see the *IBM Storwize V7000 Troubleshooting, Recovery and Maintenance Guide*, GC27-2291, at the following link:  
[https://www.ibm.com/support/knowledgecenter/ST3FR7\\_7.8.0/com.ibm.storwize.v7000.780.doc/svc\\_webtroubleshooting\\_21pbmm.html](https://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/svc_webtroubleshooting_21pbmm.html)
2. Can you log in to the Storwize V7000 system?
  - Yes: Storwize V7000 system is online; continue with Step 3.
  - No: Storwize V7000 system is offline or has connection problems. Follow these steps:
    - i. Check your connections, cabling, and control enclosure front and back panels for any event messages or indications.
    - ii. Verify the Storwize V7000 system status by using the Service Assistant Tool menu. For more information, see the *IBM Storwize V7000 Recovery and Maintenance Guide*, GC27-2291.
3. Bring a part of the Storwize V7000 system online for further diagnostic tests:
  - a. By using a browser, connect to one of the Storwize V7000 node canister's service Internet Protocol (IP) addresses:  
`https://<service_ip_add>/service/`
  - b. Log in with your Storwize V7000 cluster GUI password. You are redirected to the Service Assistant Tool menu that is shown in Figure 6-2.

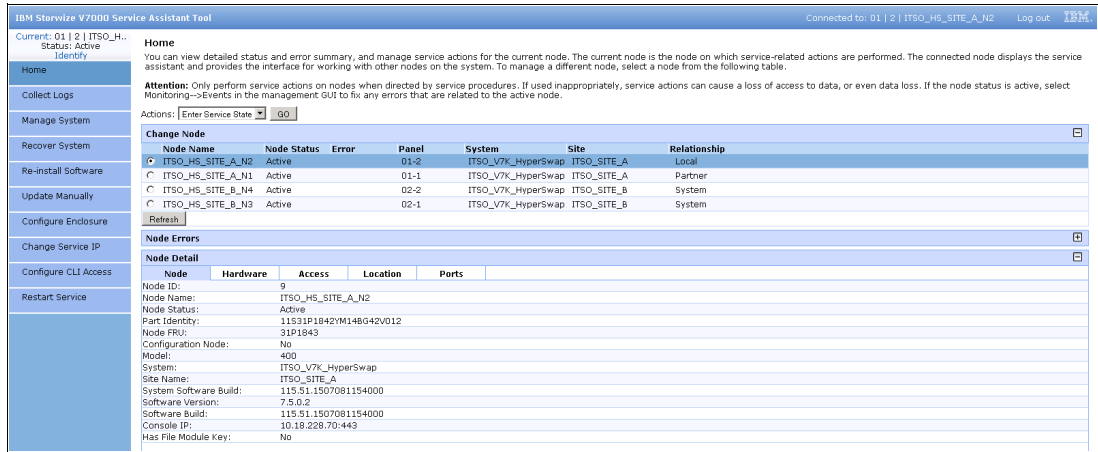


Figure 6-2 Service Assistant Tool menu

- c. After the login, from the menu that is shown in Figure 6-2 on page 158, you can try to bring at least a part of the Storwize V7000 clustered system online for further diagnostic tests. For more information about the Service Assistant menu, see the “Troubleshooting” section in the Storwize V7000 IBM Knowledge Center at the following link:

[https://www.ibm.com/support/knowledgecenter/ST3FR7\\_7.8.0/com.ibm.storwize.v7000.780.doc/svc\\_webtroubleshooting\\_21pbmm.html](https://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/svc_webtroubleshooting_21pbmm.html)

4. If the Storwize V7000 system management is available, run these checks:
  - a. Check the status by running the Storwize V7000 CLI command that is shown in Example 6-10.

*Example 6-10 Issystem example*

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>lssystem
```

- b. Check the status of the node canister and active-active Metro Mirror relationship as shown in Example 6-11.

*Example 6-11 Node canister and active-active Metro Mirror relationship status*

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>l snodecanister
id name                UPS_serial_number WWNN                status IO_group_id
IO_group_name config_node UPS_unique_id hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number
site_id site_name
1 ITSO_V7K_HS_N1_B      500507680B00217A online 0
io_grp0 no 500
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn1b 01-1
1 1 7836640 2 Site_B
2 ITSO_V7K_HS_N2_B      500507680B00217B online 0
io_grp0 yes 500
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn2b 01-2
1 2 7836640 2 Site_B
3 ITSO_V7K_HS_N1_A      500507680B0021A8 offline 1
io_grp1 no 400
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn1a 02-1
2 1 7836494 1 Site_A
```

```

4  ITS0_V7K_HS_N2_A                    500507680B0021A9 offline 1
io_grp1      no                          400
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn2a      02-2
2            2            7836494            1            Site_A
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsrcrelationship
id name      master_cluster_id master_cluster_name master_vdisk_id
master_vdisk_name      aux_cluster_id  aux_cluster_name  aux_vdisk_id
aux_vdisk_name      primary consistency_group_id consistency_group_name
state              bg_copy_priority progress copy_type  cycling_mode
freeze_time
0  Re1_ESX_B 0000010021E001E0 ITS0_V7K_HyperSwap 0
HyperSwap_Volume_ESX_B_M 0000010021E001E0 ITS0_V7K_HyperSwap 1
HyperSwap_Volume_ESX_B_A master 1 CG_ESX_BtoA
consistent_copying 50 100 activeactive
42 Re1_ESX_A 0000010021E001E0 ITS0_V7K_HyperSwap 42
HyperSwap_Volume_ESX_A_M 0000010021E001E0 ITS0_V7K_HyperSwap 43
HyperSwap_Volume_ESX_A_A aux 0 CG_ESX_AtoB
consistent_copying 50 100 activeactive
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsrcconsistgrp
id name      master_cluster_id master_cluster_name aux_cluster_id
aux_cluster_name primary state relationship_count copy_type
cycling_mode freeze_time
0  CG_ESX_AtoB 0000010021E001E0 ITS0_V7K_HyperSwap 0000010021E001E0
ITS0_V7K_HyperSwap aux consistent_copying 1
activeactive
1  CG_ESX_BtoA 0000010021E001E0 ITS0_V7K_HyperSwap 0000010021E001E0
ITS0_V7K_HyperSwap master consistent_copying 1
activeactive

```

---

Observe the following statuses in Example 6-11 on page 158:

- The *config node* role is on ITS0\_HS\_SITE\_A\_N2, but it can change in certain cases where the lost node was the config node.
- Node ITS0\_HS\_SITE\_A\_N1 and ITS0\_HS\_SITE\_A\_N2 are online.
- Node ITS0\_HS\_SITE\_A\_N3 and ITS0\_HS\_SITE\_A\_N4 are offline.
- Active-active Metro Mirror relationship Re1\_ESX\_A and Re1\_ESX\_B and the Consistency Groups that they belong to, CG\_ESX\_AtoB and CG\_ESX\_BtoA, are in **consistent\_copying** status.
- Consistency Group **CG\_ESX\_AtoB** has changed its Primary Copy from **master** to **aux**.
- During this event, the system lost 50% of the Storwize V7000 HyperSwap resources, but it is still up and running with 50% of resources.

If the critical event was not a rolling disaster and stopped with the loss of Site1 (Site\_A), the host applications that were running on Site2 can still run on this site, and the applications that were running on Site1 will start to work with its HyperSwap Secondary copy on Site2.

Later, Site1 can be recovered or rebuilt without any effect on the production systems in the same or another location.

In certain scenarios or in a rolling disaster, connectivity to Site3 can also be lost. And, the site that apparently survived will be lost, too (for example, Site2). These losses occur because the

critical event was triggered from this site (Site2), and the first site that went offline was frozen by the DR feature and set as offline.

Assuming that Site2 wins the quorum race and that the critical event was a rolling disaster that also affected Site3 (where the active quorum exists) or the IP Quorum is no longer reachable, the Storwize V7000 is stopped and needs to be recovered from the only site that is still physically available (in this example, Site1). But, that site was frozen at the time of the first critical event.

Therefore, if the impact of the failure is more serious and you are forced to declare a disaster, you must make a more strategic decision, as addressed in 6.3.3, “Storwize V7000 HyperSwap recovery guidelines” on page 160.

### 6.3.3 Storwize V7000 HyperSwap recovery guidelines

This section explores recovery scenarios. Regardless of the scenario, the common starting point is the complete loss of Site1 or Site2, which is caused by a critical event.

After an initial analysis phase of the event, a strategic decision must be made:

- ▶ Wait until the lost site is restored.
- ▶ Start a recovery procedure by using the Storwize V7000 CLI `overridequorum` command.

If recovery times are too long and you cannot wait for the lost site to be recovered, you must take the appropriate recovery actions.

#### What you need to know to recover your Storwize V7000 HyperSwap configuration

If you cannot recover the site in a reasonable time, you must take recovery actions. Consider these questions to determine the correct recovery action:

- ▶ Where do you want to recover to? In the same site or in a new site?
- ▶ Is it a temporary or permanent recovery?
- ▶ If it is a temporary recovery, do you need to plan a failback scenario?
- ▶ Does the recovery action address performance issues or business continuity issues?

You almost certainly need extra storage space, an extra Storwize V7000 Controller and Expansion enclosure, and extra SAN components. Consider these questions about the extra components:

- ▶ Do you plan to use the new Storwize V7000 Controller and Expansion enclosure that are supplied by IBM?
- ▶ Do you plan to reuse another, existing Storwize V7000 Controller and Expansion enclosure, which might be used for non-business-critical applications at the moment, such as a test environment?
- ▶ Do you plan to use new FC SAN switches or directors?
- ▶ Do you plan to reconfigure FC SAN switches or directors to host the newly acquired Storwize V7000 Controller, Expansion enclosure, and storage?
- ▶ Do you plan to use new back-end storage subsystems?

The answers to these questions direct the recovery strategy, investment of resources, and monetary requirements. These steps must be part of a recovery plan to create a minimal impact on applications and, therefore, service levels.



**Tip:** If you must recover your Storwize V7000 HyperSwap infrastructure, involve IBM Support as early as possible.

### **Recovery guidelines for the example configuration**

These recovery guidelines are based on the assumption that you answered the questions and decided to recover a fully redundant configuration in the same surviving site, starting with the **overridequorum** command to restart the frozen site at the moment of the first critical event. This work involves ordering and installing a new Storwize V7000 Controller and Expansion enclosure, and new FC SAN devices before you begin the following steps.

This recovery action is based on a decision to recover the Storwize V7000 HyperSwap infrastructure at the same performance characteristics as before the critical event. However, the solution has limited business continuity because the Storwize V7000 HyperSwap is recovered at only one site.

Ensure that you have a tested recovery plan in place, and always engage IBM Level 3 Support at the earliest possible time if you need to initiate a recovery of any sort.

**Note:** The failure domains are still represented in Figure 6-3 as the original configuration even if all of the hardware is installed in a single site only because the Storwize V7000 HyperSwap configuration still has a three-site definition.

Figure 6-3 shows the new recovery configuration.

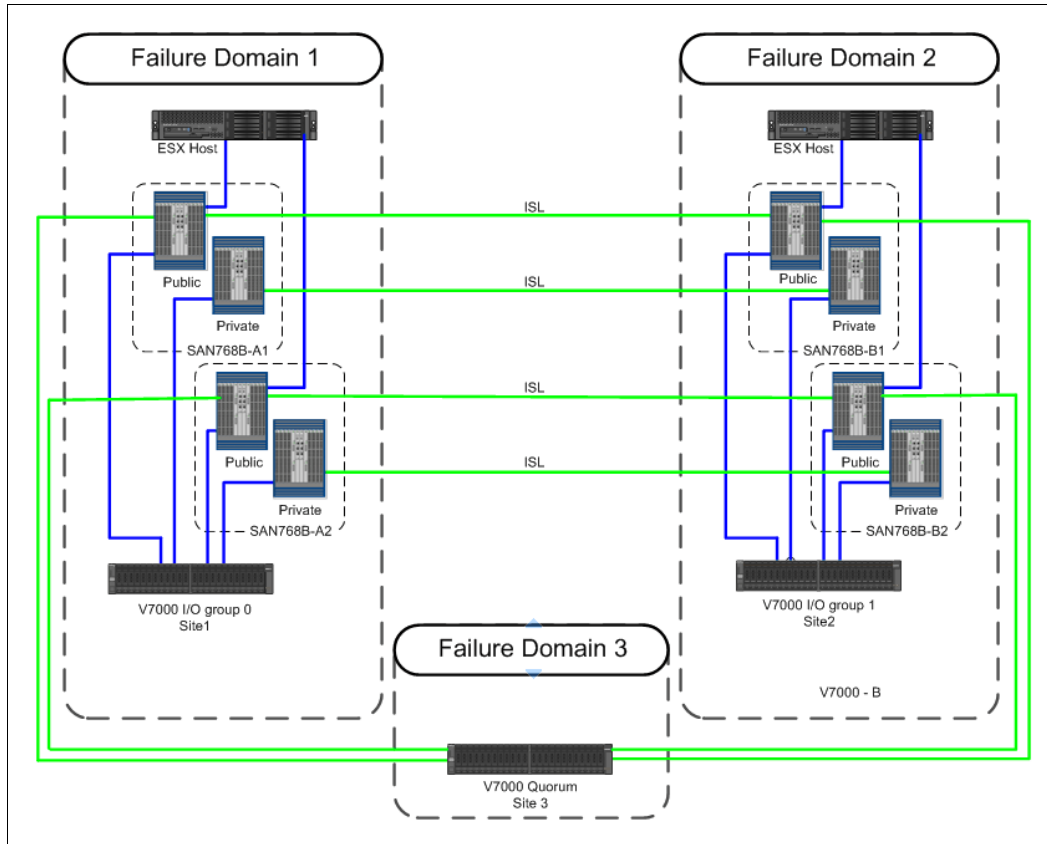


Figure 6-3 New recovery configuration in surviving site

**Note:** Example 6-11 on page 158 shows Quorum Disk on Site\_3. In our example later on this chapter the IP Quorum has been implemented.

### Summary of major steps

The configuration is recovered exactly as it was, even if it is recovered in the same site. You can make it easier in the future to implement this configuration over distance when a new site is provided by completing the following major steps:

1. Ensure that the active quorum disk is available and visible from the site that is not to be moved.
1. Disconnect the ISL links between the failure domains.
2. Uninstall and reinstall devices that you plan to reuse, or install all of the new devices in the new sites.
3. Reconnect the ISL links between the failure domains.

### Steps to restore your Storwize V7000 HyperSwap configuration in the same site

Complete the following steps to restore your Storwize V7000 HyperSwap configuration as it was before the critical event in the same site. The steps are finished after you install the new devices.

**Important:** Before you perform any recovery action, you must ensure that the previous environment or site cannot come back to life with a device or node canister that still has earlier configuration data. This situation will cause serious problems in the environment or site that you are working on. Take any appropriate action to ensure that they cannot come back to life again (link disconnection, power down, and so on).

Follow these steps:

1. Only if you are virtualizing the external storage controller, restore your back-end storage subsystem configuration as it was, starting from your backup. Logical unit number (LUN) masking can be performed in advance because the Storwize V7000 node's WWNN is already known.
2. Restore your internal storage configuration as it was, starting from your backup.
3. Restore your SAN configuration exactly as it was before the critical event. You can restore your SAN configuration by configuring the new switches with the same domain ID as before and connecting them to the surviving switches. The WWPN zoning is then automatically propagated to the new switches.
4. If possible, connect the Storwize V7000 node canister to the same FC switch ports as before the critical event and the external storage controller, if used, as well.

**Note:** If you are virtualizing the external storage controller, Storwize V7000 node canister-to-storage zoning must be reconfigured to be able to see the new storage subsystem's WWNN. Previous WWNNs can be removed, but with care.

5. Do not connect the Storwize V7000 node canister FC ports yet. Wait until directed to do so by the Storwize V7000 node canister WWNN change procedure.

**Note:** All of the recovery action is executed by using the Storwize V7000 CLI. At the time of writing this book, no support is available in the Storwize V7000 HyperSwap GUI to execute the required recovery action.

6. Connect to one of your Storwize V7000 node canisters that is in the frozen site. You are going to recover that node with the Storwize V7000 CLI by using its service IP address and running the `sainfo lsservicenodes` command, as shown in Example 6-12.

*Example 6-12 sainfo lsservicenodes command*

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>sainfo lsservicenodes
panel_name cluster_id      cluster_name      node_id node_name
relation node_status error_data
02-1      0000010021E001E0 ITSO_V7K_HyperSwap 3      ITSO_V7K_HS_N1_A local
Starting  551 01-1 01-2
02-2      0000010021E001E0 ITSO_V7K_HyperSwap 4      ITSO_V7K_HS_N2_A partner
Starting  551 01-1 01-2
```

As Example 6-12 shows, the two node canisters are in the **Starting** state. They show a **551** error code waiting for the 01-2 and 01-1 missing resources that are the node canister `panel_name` that relate to the missing node canisters in the rolling disaster.

7. Run the **overridequorum** command as shown in Example 6-13. No return message is expected, just the prompt, and you will lose the connection in a few seconds.

*Example 6-13 overridequorum command*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>satask overridequorum -force
```

---

Now, a new Storwize V7000 HyperSwap cluster is created with only the available resources online and with the previous resources offline. These resources are still in the Storwize V7000 HyperSwap configuration files and in the quorum disk that is used to re-create the cluster.

8. Run the **lssystem** and **lsnode** commands from the Storwize V7000 CLI. Use your regular Storwize V7000 cluster management IP address to show the new Storwize V7000 HyperSwap cluster that was created and the online and offline resources, as shown in Example 6-14. The new cluster that was created has the same management IP address as the previous cluster.

*Example 6-14 New cluster and the online and offline resources*

---

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lssystem
```

```
id 00001002160020E
```

```
name ITS0_V7K_HyperSwap
```

```
location local
```

```
.
```

```
multiple lines omitted
```

```
.
```

```
code_level 7.8.0.2 (build 133.14.1610192015000)
```

```
console_IP 10.18.228.70:443
```

```
.
```

```
multiple lines omitted
```

```
.
```

```
topology hyperswap
```

```
topology_status recovered_site_1
```

```
rc_auth_method none
```

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>svcinfolcluster
```

```
id name location partnership bandwidth id_alias
```

```
00001002160020E ITS0_V7K_HyperSwap local
```

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lspathcanister
```

```
id name UPS_serial_number WWNN status IO_group_id
```

```
IO_group_name config_node UPS_unique_id hardware iscsi_name
```

```
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number site_id
```

```
site_name
```

```
1 ITS0_V7K_HS_N1_B 500507680B00217A offline 0
```

```
io_grp0 no 500
```

```
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn1b 01-1 1
```

```
1 7836640 2 Site_B
```

```
2 ITS0_V7K_HS_N2_B 500507680B00217B offline 0
```

```
io_grp0 no 500
```

```
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn2b 01-2 1
```

```
2 7836640 2 Site_B
```

```
3 ITS0_V7K_HS_N1_A 500507680B0021A8 online 1
```

```
io_grp1 yes 400
```

```
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn1a 02-1 2
```

```
1 7836494 1 Site_A
```

```

4  ITS0_V7K_HS_N2_A          500507680B0021A9 online 1
io_grp1      no              400
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn2a 02-2 2
2            7836494          1      Site_A

```

Example 6-14 on page 164 shows that the new cluster ID was created with the same management IP address. The two Storwize V7000 node canisters that were in a Starting state with error code 551 are now online. The Storwize V7000 cluster topology status is now **recovered\_site\_1** and its topology is still **hyperswap**.

9. Remove the two offline Storwize V7000 node canisters with the `rmnodecanister -force` command.
10. Now, you will see several 1700 and 1895 errors as shown in Figure 6-4.

Error Code	Last Time Stamp	Status	Description	Object Type	Object ID	Object Name
1700	10/25/16 10:34:12 AM	Alert	Unrecovered Remote Copy relationship	io_grp	0	io_grp0
1895	10/25/16 10:34:12 AM	Alert	Unrecovered FlashCopy mappings	io_grp	0	io_grp0

Figure 6-4 1700 and 1895 errors

Follow the maintenance procedure to fix those errors. The procedure guides you through detailed steps. You are requested to note the active-active Metro Mirror relationship that will be removed because it relates to a Master or Auxiliary Volume that no longer exists. The volumes need to be re-created later. Figure 6-5, Figure 6-6 on page 166, and Figure 6-7 on page 166 show examples of the maintenance procedure.

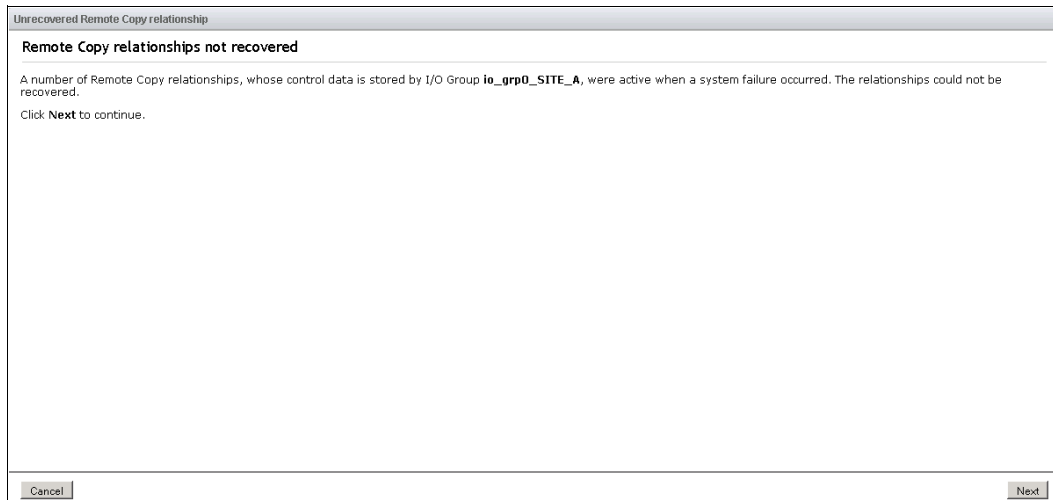


Figure 6-5 Maintenance procedure example

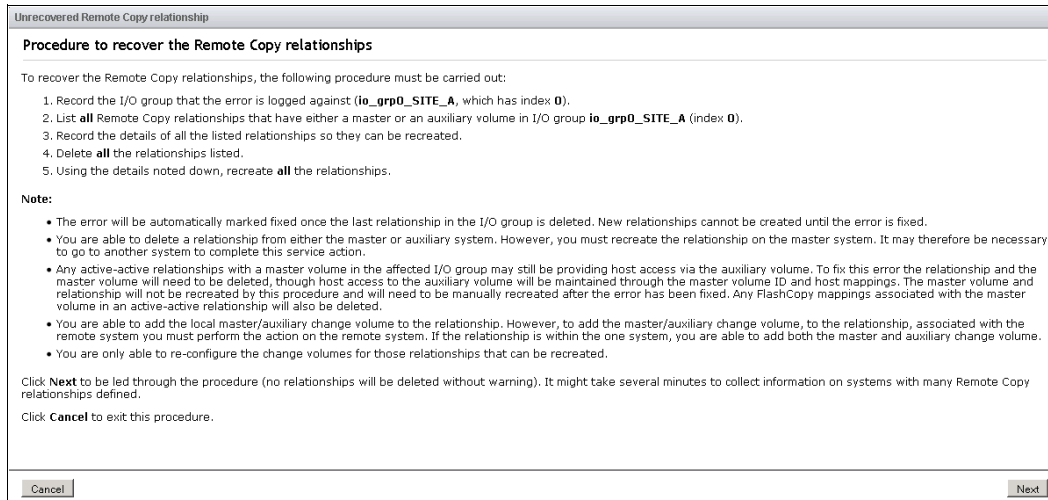


Figure 6-6 Maintenance procedure example

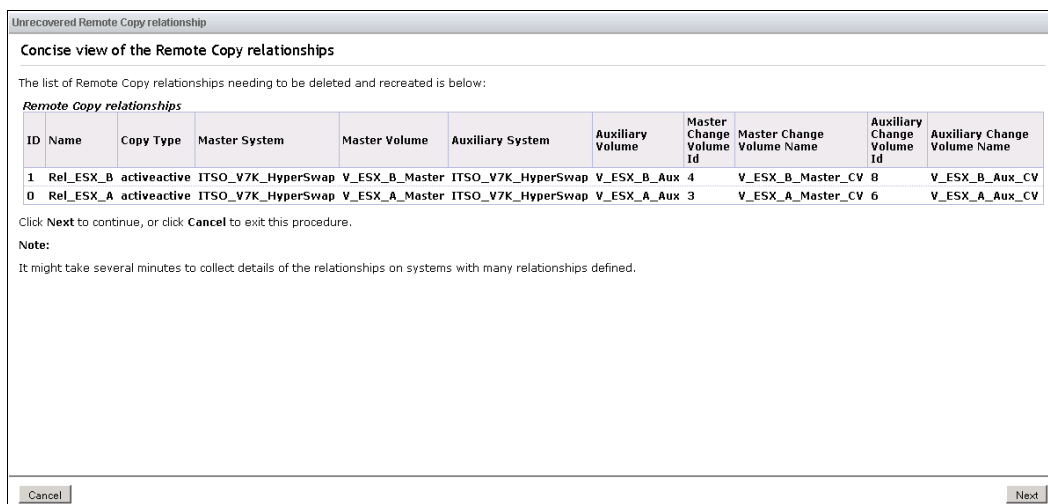


Figure 6-7 Maintenance procedure example

When the errors are fixed, the active-active Metro Mirror relationship will be removed. The Master Volumes that belong to the destroyed site will be removed automatically by the maintenance procedure. The **rmvdisk -force -keepaux** command is executed automatically as part of the maintenance procedure. The new **-keepaux** flag was added to the command in version 7.6 to support HyperSwap. The maintenance procedure executes the following three steps:

- a. Remove the Master Volume.
  - b. Switch the Auxiliary Volume unique identifier (UID) and name to the Master UID and name.
  - c. Remove the active-active Metro Mirror relationship.
11. You still have offline Auxiliary Volumes that originally belonged to the destroyed site. Remove the volumes by using the **rmvdisk** commands.
  12. Remove the storage pool that went offline and that refers to the destroyed site. First, identify which pools are offline, and then remove them with the **rmmdiskgrp** command. They will be re-created later.

13. Power on the newly supplied Storwize V7000 Controller and Expansion enclosure if you did not power on the Storwize V7000 Controller and Expansion enclosure. Set its service IP address, but leave the FC cable disconnected.

Change the new node canister's WWNN to match the WWNN that you recorded earlier in 6.2, "Storwize V7000 recovery planning" on page 146 by using the Service Assistant as shown in Figure 6-8.

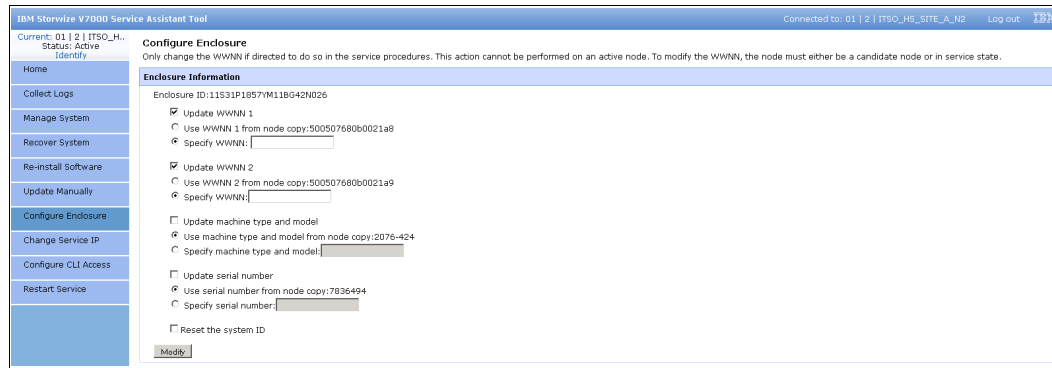


Figure 6-8 Node canister WWNN change

14. Connect the node to the same FC switch ports that it was connected to before the critical event.

**Important:** *This step is the key point of the recovery procedure.* By connecting the new Storwize V7000 node canister to the same SAN ports and by reusing the same WWNN, you avoid rebooting, rediscovering, and reconfiguring. You avoid creating any negative effect from the host's point of view because the lost disk resources and paths are restored.

**Important:** Do *not* connect the new node canister to different ports at the switch or director. Using different ports causes FC port IDs to change, which can affect the hosts' access to volumes or cause problems with adding the new node canister back into the clustered system.

If you cannot connect the Storwize V7000 node canister to the same FC SAN ports as before, complete these steps:

- ▶ Restart the system.
- ▶ Rediscover or reconfigure your host to see the lost disk resources.
- ▶ Restore the paths.

15. Issue the Storwize V7000 `lscontrolenclosurecandidate` CLI command, the `svcinfo lsnodecandidate` command, or the `sainfo lsservicenodes` CLI command as shown in Example 6-15. Verify that the new control enclosure or node canister is in the Candidate state, which means that it is ready to be added to the Storwize V7000 cluster.

*Example 6-15 Verifying candidate node canister*

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>sainfo lsservicenodes
panel_name cluster_id cluster_name node_id node_name relation node_status
error_data
7836494-1                                local   Candidate
7836494-2                                partner Candidate
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>s svcinfo lsnodecandidate
id panel_name UPS_serial_number UPS_unique_id hardware
500507680B0021A8 7836494-1                    500507680B0021A8 400
```

16. Add the control enclosure to the clustered system, and ensure that it is added back to the correct I/O Group with Storwize V7000 CLI commands as shown in Example 6-16.

*Example 6-16 Adding a node*

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>svctask addnode -wwnodename
500507680B00217A -iogrp 0 -site 2
Node, id [5], successfully added
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>lsnodecanister
id name UPS_serial_number WWNN status IO_group_id
IO_group_name config_node UPS_unique_id hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number site_id
site_name
5 node1 500507680B00217A adding 0
io_grp0 no 500
iqn.1986-03.com.ibm:2145.itsov7khyperswap.node1 7836640-1
1 7836640 2 Site_B
6 node2 500507680B00217B adding 0
io_grp0 no 500
iqn.1986-03.com.ibm:2145.itsov7khyperswap.node2 7836640-2
2 7836640 2 Site_B
3 ITSO_V7K_HS_N1_A 500507680B0021A8 online 1
io_grp1 yes 400
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn1a 02-1 2
1 7836494 1 Site_A
4 ITSO_V7K_HS_N2_A 500507680B0021A9 online 1
io_grp1 no 400
iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsov7khsn2a 02-2 2
2 7836494 1 Site_A
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>lsnodecanister
```

.  
*many lines omitted*  
.

```
IBM_Storwize:ITSO_V7K_HyperSwap:superuser>lsnodecanister
id name UPS_serial_number WWNN status IO_group_id
IO_group_name config_node UPS_unique_id hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number site_id
site_name
5 node1 500507680B00217A online 0
io_grp0 no 500
iqn.1986-03.com.ibm:2145.itsov7khyperswap.node1 01-1 1
1 7836640 2 Site_B
6 node2 500507680B00217B online 0
io_grp0 no 500
```





The active-active Metro Mirror relationships that originally were going from Site1 to Site2 will be started. They will be in the `inconsistent_copying` status until they get `consistent_synchronized` status. Those relationships will keep their Primary copy on Site1 and their Secondary copy on Site2.

It will be different for the relationships that originally were going from Site2 to Site1 with the Primary copy on Site2 and the Secondary copy on Site1. The original Master Volume was recovered in step 10 on page 165 by fixing the errors by using the maintenance procedure. The Primary copy is in Site1 now and the Master Volume for those relationships is in Site1 also.

You need to reverse the relationship direction to make the direction from Site2 to Site1, and you must also switch the Master Volume location from Site1 to Site2. Step 22 describes how to reverse those relationships.\*\*\*

20. You need to reverse all of the active-active Metro Mirror relationships that go from Site1 to Site2 with the Primary Volume in Site1 and the Secondary Volume in Site2, to make them go from Site2 to Site1 and with the Primary Volume in Site2 and the Secondary Volume in Site1. Follow these steps:

- i. All of the involved relationships must be in the `consistent_synchronized` status.
- ii. No applications can have access to the Master or Aux volume that is involved in this procedure. Because this change is to the volume state, be careful with this process. Do not perform this process if any data or state from the volume is cached in host systems. Ideally, shut down host systems that use this volume before you perform these steps.

**Important:** If you run these commands *without these precautions*, you will almost certainly cause issues to the applications and corrupt the stale copy.

**Important:** Do *not* remove the active-active Metro Mirror relationship for the volumes that are going to be removed, or the next step will fail.

- iii. Remove the Master Volumes now in Site1 that need to be switched by using the Storwize V7000 CLI command:

```
svctask rmvdisk -force -keepaux <volume name>
```

The `-keepaux` flag retains accessibility of the Auxiliary copy through the Master Volume's host mappings if the volume that is being deleted is the Master Volume of an active-active relationship.

This step will remove all of the active-active Metro Mirror relationships that relate to the volumes that are going to be switched automatically.

- iv. Re-create all of the Master, Aux, and CV volumes that were previously hosted by the I/O Group in Site1 and need to be switched.
- v. Re-create and start all of the active-active Metro Mirror relationships and Consistency Groups that relate to the volumes that need to be switched.

21. Consider running the quorum-assigning procedure to verify and eventually reassign the three quorum disks according to your new back-end storage subsystem by using the Storwize V7000 CLI `chquorum` command.

22. Check the quorum disk location with the command that is shown in Example 6-18.

*Example 6-18 Isquorum command*

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lquorum
```

quorum_index	status	id	name	controller_id	controller_name	active	object_type
0	online	11				no	drive
1			Site_A				
1	online	5				no	drive
2			Site_B				
3	online				yes		device

ITS0-1.englab.brocade.com/10.18.228.170

**Note:** Example 6-18 shows 2 Quorum Disk and the IP Quorum as active Quorum. This reflect our lab. configuration. If you have implemented the Quorm Disk you will see it instead of the IP Quorum.

All of your volumes are now accessible from your host's point of view. The recovery action is terminated, and the Storwize V7000 HyperSwap environment is active again.

All of these operations are guidelines to help you in a critical event or a disaster. Several of these steps are specific to our lab environment, and several steps are common with every Storwize V7000 HyperSwap environment.

Ensure that you tested an established recovery plan. Always engage IBM L3 Support at the earliest possible time if you need to initiate a recovery of any sort.

## 6.4 Other disaster recovery with HyperSwap

Under certain circumstances, other events can affect the Storwize V7000 HyperSwap environment. In these circumstances, the HyperSwap function will automatically use both copies to provide continuous host access to data, providing that both copies are up-to-date. If one copy is up-to-date, the other copy is stale, and the up-to-date copy goes offline, the system cannot automatically use the remaining copy to provide HA to the volume.

However, the user can choose to enable access to that stale copy, which instructs the system to return the state of that volume to the point in time of that stale copy.

Because this change is a step-by-step change to the volume state, be careful with this process. Do not perform this process if any data or state from the volume is cached in host systems. Ideally, shut down the host systems that use the volume before you perform these steps. By running these commands without these precautions, you will almost certainly crash your applications and might corrupt the stale copy.

We show an example where a problem occurred with the Primary copy during a resynchronization between sites as shown in Example 6-19.

*Example 6-19 Isrcrelationship example*

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsrcrelationship 0
id 0
name rcrel0
master_cluster_id 0000020076805032
master_cluster_name dizzy
master_vdisk_id 0
master_vdisk_name vdisk0
aux_cluster_id 0000020076805032
aux_cluster_name dizzy
```

```

aux_vdisk_id 6
aux_vdisk_name vdisk6
primary master
consistency_group_id 0
consistency_group_name rccstgrp0
state consistent_copying
bg_copy_priority 50
progress 81
freeze_time 2015/08/11/12/16/47
status online
sync out_of_sync
copy_type active_active
cycle_period_seconds 300
cycling_mode
master_change_vdisk_id 1
master_change_vdisk_name vdisk1
aux_change_vdisk_id 7
aux_change_vdisk_name vdisk7

```

---

As shown in Example 6-19 on page 171, the HyperSwap Volumes are still resynchronizing. The `consistent_copying` state of the volume shows a resynchronization where the Secondary copy contains a stale image, and the value that is contained in the `freeze_time` field shows when that image dates from. The progress value increases toward 100 as the resynchronization process continues. See the `progress 83` line in Example 6-20.

Now, the site of the Primary copy goes offline as shown in Example 6-20.

*Example 6-20 `lsrrelationship out_of_sync` example*

---

```

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsrrelationship 0

```

```

.
many lines omitted for brevity

```

```

state consistent_copying
bg_copy_priority 50
progress 83
freeze_time 2015/03/04/12/16/47
status primary_offline
sync out_of_sync
many lines omitted for brevity

```

```

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsvdisk
id name   IO_group_id IO_group_name status  mdisk_grp_id mdisk_grp_name capacity
type     FC_id FC_name RC_id RC_name vdisk_UID                                fc_map_count
copy_count fast_write_state se_copy_count RC_change compressed_copy_count
0 vdisk0 0          io_grp0  offline 0          mdiskgrp0
250.00MB striped many many 0    rcre10 6005076801DA0140C800000000000000 2
1          empty          0          no          0

```

---

With the only up-to-date copy of the volume offline, the active-active relationship cannot switch direction to keep the HyperSwap Volume online, so the Master Volume is now offline as shown in Example 6-20.

At this point, you need to look at the `freeze_time` value. If data from that date is not useful, for example, it is from too long ago, or before a recent vital update, it might be best to wait until

the offline up-to-date copy of the volume can be brought back online. However, if the stale data is useful, and it is likely that the up-to-date copy of the volume will remain offline for an extended period of time, you can choose to enable access to the stale copy of the volume. Before you run this command, ensure that no data or state from this volume is cached on host systems:

```
stoprcrelationship -access rcrc10
```

At this point, the data that is presented to hosts from this volume immediately changes to that data that is stored on the stale copy. One way to think of this process is that the volume was consistently rolled back to the point in time that is denoted by the `freeze_time` value.

The volume continues to be readable and writable at this point. You can start your business applications again, and continue from this stale image.

Replication is paused, even if the up-to-date copy becomes online again. The previously stale image, which is now being accessed by hosts, and the previously up-to-date copy, which contains changes that are not present on the previously stale image, are now divergent copies. The two copies were the same at the `freeze_time` point in time, but then different writes were applied to each copy. Either copy might be the copy that the user wants to keep long term.

So, the system allows the user to choose which copy is more useful to the user. This choice will be made based on how much data was missing on the stale copy compared to the up-to-date copy, and how much progress was made on the stale copy since access was enabled to it.

The first step is to determine which copy has the stale copy that is accessible to hosts. This copy will be either the Master or Auxiliary copy, and it is visible under the “Primary” attribute of the active-active relationship.

Next, consider the copy that you want to retain.

### 6.4.1 Continue to use the copy that hosts currently access

In this scenario, we continue to use the copy that hosts are currently accessing. We discard the old “up-to-date copy”.

In this scenario, we consider that the DR that used the stale copy was successful. And, you made useful business progress by using that copy, and you value that work more than any data that the old up-to-date copy has that the copy that you are using does not. Or, maybe so little difference exists between the two copies that you choose to continue to use what you have. Switching back to the up-to-date copy is a disruptive operation for hosts, and you do not want that impact.

So, you keep the stale copy and discard the up-to-date copy. Use this command:

```
startrcrelationship -primary <current_primary> -force <relationship>
```

The `<current_primary>` is the current primary value of the active-active relationship, and the value will be `master` or `aux`. After you decide and lose the ability to use the copy that is not the Primary, the `-force` flag instructs the system that you are aware that this operation cannot be reversed.

You do not need to quiesce host I/O or take any further action. This command resumes the HyperSwap replication, and it copies across any regions that are different between the two copies to resynchronize as fast as possible. Both copies keep a bitmap of volume regions at a

256 KB granularity, which is used to record writes to that copy that are not yet replicated to the other copy.

On this resynchronization, we use both sets of information to undo writes that were only applied to the old up-to-date copy, and also to copy across additional writes that were made to the stale copy during the DR. Because the DR only happened because the copies were resynchronizing before the up-to-date copy went offline, all differences from that interrupted resynchronization process will be reverted on the old up-to-date copy now, also.

The active-active relationship goes into an `inconsistent_copying` state, and while copying continues, the progress increases toward 100. At that point, the relationship goes into a `consistent_synchronized` state, which shows that both copies are up-to-date, and HA is restored.

## 6.4.2 Go back to the up-to-date copy and discard the stale disaster recovery copy

In this scenario, we go back to the up-to-date copy and discard the stale copy that was used for DR.

We describe the actions if you want to go back to the up-to-date copy, which is the copy that held the latest data before the DR. Maybe the stale data turned out to not hold useful data, or the outage of the up-to-date copy was shorter than you expected.

This scenario differs from the last scenario because the image that is visible by hosts will change again. Just as enabling access to the stale copy required hosts to have no cached data from the volume (and ideally they must be fully shut down), the same requirement is true for reverting to the up-to-date copy. So, before you proceed, ensure that no hosts will be affected by the data changes and that no hosts have stale data that they might corrupt the up-to-date copy with.

Use this command:

```
starttrcrelationship -primary <current_secondary> -force <relationship>
```

The `<current_secondary>` copy is the copy other than the current primary value of the active-active relationship, and it will be `master` or `aux`. If the primary field says `master`, use `aux` for this value, and vice versa. As before, you cannot get back to the other set of data after you run this command because of the `-force` flag.

The image that is visible to hosts instantly reverts to the up-to-date copy. As soon as you run this command, bring your hosts back online, and start to use this volume again.

As with the other scenario, the active-active relationship will be in an `inconsistent_copying` state while resynchronization occurs. Again, the resynchronization uses the bitmaps of writes to each copy to accelerate this resynchronization process. After the copies are fully synchronized, the relationship goes back to a `consistent_synchronized` state while HA is restored for the volume.

## 6.4.3 Disaster recovery with Volume Groups

All of the descriptions in 6.4.1, “Continue to use the copy that hosts currently access” on page 173 and 6.4.2, “Go back to the up-to-date copy and discard the stale disaster recovery copy” on page 174 about enabling access to a stale copy of a HyperSwap Volume also apply

to HyperSwap Volume Groups, that is, multiple HyperSwap Volumes where the active-active relationships are contained in a single Consistency Group.

If during resynchronization, any of the up-to-date copies of volumes in a Volume Group are offline or unavailable (typically, all are offline in a disaster), you can choose to enable access to the stale copy of every volume in the Volume Group. Because the HyperSwap function links replication and failover across HyperSwap Volumes in a Volume Group, it guarantees that during resynchronization, all copies on one site have a stale consistent copy of data that was captured at an identical point in time. This stale consistent copy of data is ideal for DR.

The Consistency Group (**consistgrp**) versions of the commands are shown:

- ▶ Use `stoprcconsistgrp -access <consistency_group>` to gain access to the stale copies. Then, use one of the following commands:
  - `startrcconsistgrp -primary <current_primary> -force <consistency_group>` to retain the stale DR copies currently visible to hosts and to resume HyperSwap replication.
  - `startrcconsistgrp -primary <current_secondary> -force <consistency_group>` to revert to the previous up-to-date copy.

#### 6.4.4 Convert to single-copy volumes

We convert to single-copy volumes and retain access through the Auxiliary Volume. In this scenario, you need to retain the Auxiliary Volume, for example, because you want to reverse your active-active Metro Mirror relationship without stopping the server applications. Or, you want to switch the site to which your Master Volume belongs. This scenario is already used in step 22 on page 150.

Run this command:

```
mvdisk -keepaux <mastervdisk>
```

You can run this command with host I/O running. This command deletes the Master Volume's storage and replaces it with the Auxiliary Volume's storage, therefore preserving the Master Volume ID, the Master Volume host maps, and the Auxiliary Volume storage. This command also deletes the active-active relationship. Finally, delete the Change Volumes (CVs), which are not deleted as part of the previous step.

This scenario allows a cleanup of failed Master storage without affecting host I/O access, potentially as part of replacing the Master Volume's storage.





# Related publications

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

## IBM Redbooks

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

- ▶ *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116
- ▶ *Implementing the IBM Storwize V7000 V6.3*, SG24-7938
- ▶ *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933
- ▶ *Implementing the IBM SAN Volume Controller and FlashSystem 820*, SG24-8172
- ▶ *Implementing IBM FlashSystem 900*, SG24-8271
- ▶ *Introducing and Implementing IBM FlashSystem V9000*, SG24-8273
- ▶ *IBM FlashSystem A9000 and IBM FlashSystem A9000R Architecture, Implementation, and Usage*, SG24-8345
- ▶ *VersaStack Solution by Cisco and IBM with Oracle RAC, IBM FlashSystem V9000, and IBM Spectrum Protect*, SG24-8364
- ▶ *IBM FlashSystem V9000 in a VersaStack Environment*, REDP-5264
- ▶ *VersaStack Solution by Cisco and IBM with SQL, Spectrum Control, and Spectrum Protect*, SG24-8301
- ▶ *VersaStack Solution by Cisco and IBM with IBM DB2, IBM Spectrum Control, and IBM Spectrum Protect*, SG24-8302
- ▶ *iSCSI Implementation and Best Practices on IBM Storwize*, SG24-8327

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

<http://www.redbooks.ibm.com/>

The following is a list of useful Redbooks domains related to this book:

IBM Storage Networking Redbooks:

<http://www.redbooks.ibm.com/Redbooks.nsf/domains/san?Open>

IBM Flash storage Redbooks

<http://www.redbooks.ibm.com/Redbooks.nsf/domains/flash?Open>

IBM Software Defined Storage Redbooks

<http://www.redbooks.ibm.com/Redbooks.nsf/domains/sds?Open>

IBM Disk storage Redbooks

<http://www.redbooks.ibm.com/Redbooks.nsf/domains/disk?Open>

IBM Storage Solutions Redbooks

<http://www.redbooks.ibm.com/Redbooks.nsf/domains/storagesolutions?Open>

IBM Tape storage Redbooks

<http://www.redbooks.ibm.com/Redbooks.nsf/domains/tape?Open>

## Other publications

These publications are also relevant as further information sources:

- ▶ *IBM System Storage Master Console: Installation and User's Guide*, GC30-4090
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: CIM Agent Developers Reference*, SC26-7545
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Command-Line Interface User's Guide*, SC26-7544
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Configuration Guide*, SC26-7543
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Host Attachment Guide*, SC26-7563
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Installation Guide*, SC26-7541
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Planning Guide*, GA22-1052
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Service Guide*, SC26-7542
- ▶ *IBM Storwize V7000 Troubleshooting, Recovery and Maintenance Guide*, GC27-2291:  
<https://ibm.biz/BdX9wM>

## Online resources

These websites are also relevant as further information sources:

- ▶ IBM System Storage home page:  
<http://www.ibm.com/systems/storage/>
- ▶ IBM System Storage Interoperation Center (SSIC):  
<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>
- ▶ IBM Storwize V7000 Knowledge Center:  
<https://ibm.biz/BdX9r6>
- ▶ The following website provides additional VMware resources:  
<http://www.vmware.com/support/pubs/>
- ▶ DRS Blog Guide:  
<https://ibm.biz/BdXLP>
- ▶ *vSphere High-Availability Deployment Best Practices*:  
<http://ibm.biz/BdxrmN>

- ▶ *Advanced configuration options for VMware High Availability in vSphere 5.x*, 2033250:  
<https://ibm.biz/BdRxV8>
- ▶ *VMware Handling Transient APD Conditions* guide:  
<https://ibm.biz/BdDwq7>
- ▶ *Working with Permanent Device Loss*:  
<http://ibm.biz/Bdx4k7>
- ▶ *VM Component Protection (VMCP)* blog from VMware:  
<https://ibm.biz/BdXukg>
- ▶ VMware Knowledge Base article, *Supported vCenter Server high availability options*, 1024051:  
<https://ibm.biz/BdXL4t>
- ▶ *VMware vCenter Server 6.0 Availability Guide*:  
<https://ibm.biz/BdXLte>
- ▶ *The Setup for Failover Clustering and Microsoft Cluster Service* VMware guide:  
<https://ibm.biz/BdXLtJ>
- ▶ *Recreating a VMware High Availability Cluster in vSphere*, 1003715:  
<https://ibm.biz/BdXCg4>
- ▶ *Troubleshooting VMware High Availability (HA) in VMware vSphere*, 1001596:  
<https://ibm.biz/BdXCgD>
- ▶ *Location of log files for VMware products*, 1021806:  
<https://ibm.biz/BdXCgy>
- ▶ *Location of VMware vCenter Server 6.0 log files*, 2110014:  
<https://ibm.biz/BdXCgs>

## Help from IBM

IBM Support and downloads

[ibm.com/support](https://ibm.com/support)

IBM Global Services

[ibm.com/services](https://ibm.com/services)











SG24-8317-00

ISBN 0738441147

Printed in U.S.A.

Get connected

