

IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware

Jon Tate

Angelo Bernasconi

Antonio Rainero

Ole Rasmussen



Storage



International Technical Support Organization

**IBM Spectrum Virtualize and SAN Volume Controller
Enhanced Stretched Cluster with VMware**

October 2015

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

Second Edition (October 2015)

This edition applies to IBM System Storage SAN Volume Controller Version 7.8 and VMware Version 6.0.

© 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

Notices	vii
Trademarks	viii
IBM Redbooks promotions	ix
Summary of changes	xi
October 2015, Second Edition	xi
Preface	xiii
Authors	xiii
Now you can become a published author, too!	xv
Comments welcome	xv
Stay connected to IBM Redbooks	xv
Chapter 1. Introduction	1
1.1 IBM Spectrum Virtualize	2
1.2 Enhanced Stretched Cluster option	4
1.3 Integration of Spectrum Virtualize, SAN Volume Controller, Layer 2 IP Network, Storage Networking infrastructure, and VMware	6
1.3.1 Application mobility over distance	7
1.3.2 Benefits of this integrated solution	7
1.3.3 Benefits in more detail	8
1.3.4 When to use VMware stretched clusters	9
1.3.5 When not to use VMware stretched clusters	9
1.3.6 IBM, VMware, and Layer 2 IP network solution	10
1.4 Open Data Center Interoperable Network	10
Chapter 2. Hardware and software descriptions	13
2.1 Hardware description	14
2.2 IBM System Storage Spectrum Virtualize and SAN Volume Controller	14
2.3 SAN directors and switches	14
2.3.1 SAN384B-2 and SAN768B-2 directors	15
2.3.2 SAN24B-5, SAN48B-5, and SAN96B-5 switches	15
2.4 FCIP routers	17
2.4.1 8 Gbps Extension Blade	17
2.4.2 SAN06B-R extension switch	18
2.5 Ethernet switches and routers	19
2.5.1 IBM System Networking switches	19
2.5.2 Brocade IP routers and Layer 4 - 7 application delivery controllers	20
2.6 Software high availability	23
2.7 VMware ESX and VMware ESXi	23
2.7.1 VMware vSphere	23
2.7.2 vSphere vMotion	23
2.7.3 vSphere High Availability	24
2.7.4 VMware vCenter Site Recovery Manager	24
2.7.5 VMware Distributed Resource Scheduler	25
Chapter 3. Enhanced Stretched Cluster architecture	27
3.1 Enhanced Stretched Cluster overview	28

3.2	Failure domains, sites, and controllers	29
3.3	Spectrum Virtualize volume mirroring	31
3.3.1	Volume mirroring prerequisites	31
3.3.2	Read operations	31
3.3.3	Write operations	32
3.3.4	Quorum disk	33
3.3.5	Cluster state and voting	34
3.3.6	Quorum disk requirements	35
3.3.7	IP Quorum.	36
3.3.8	Failure scenarios in an Enhanced Stretched Cluster configuration	36
3.4	Spectrum Virtualize enhanced stretched cluster configurations	38
3.4.1	No ISL configuration	39
3.4.2	ISL configuration	42
3.4.3	FCIP configuration	46
3.5	Fibre Channel settings for distance.	48
3.6	Spectrum Virtualize I/O operations on mirrored volumes	49
3.6.1	Preferred node	49
3.6.2	Primary copy	50
3.7	Enhanced stretched cluster three-site DR configuration.	51
	Chapter 4. Implementation	53
4.1	Test environment	54
4.2	ADX: Application Delivery Controller.	56
4.2.1	VIP and real server configuration	58
4.2.2	Global Server Load Balancing (GSLB) configuration	58
4.2.3	Application Resource Broker (ARB) server installation.	60
4.2.4	ADX registration in the ARB plug-in	62
4.2.5	Enable VM mobility in the ARB plug-in in vCenter	65
4.2.6	Additional references	67
4.3	IP networking configuration.	67
4.3.1	Layer 2 switch configuration	71
4.3.2	IP Core (MLXe) configuration	73
4.3.3	Data Center Interconnect (Brocade CER series) configuration	78
4.4	IBM Fibre Channel SAN	80
4.4.1	Creating the logical switches.	82
4.4.2	Creating FCIP tunnels.	91
4.5	Spectrum Virtualize with an Enhanced Stretched Cluster.	94
4.6	Volume mirroring.	95
4.7	Read operations	96
4.8	Write operations	96
4.9	Quorum disk	97
4.9.1	Quorum disk requirements and placement.	98
4.9.2	Automatic quorum disk selection	98
4.10	IP Quorum.	100
4.11	Enhanced Stretched Cluster configuration	105
4.11.1	Using the CLI	105
4.11.2	Using the GUI	110
4.12	Storage allocation to the Spectrum Virtualize	115
4.13	Volume allocation	116
	Chapter 5. VMware environment	119
5.1	VMware configuration checklist.	120
5.2	VMware and Enhanced Stretched Cluster	121

5.3 VMware vCenter setup	121
5.3.1 Metro vMotion	122
5.4 ESXi host installations.	122
5.4.1 ESXi host HBA requirements	122
5.4.2 Initial ESXi verification	123
5.4.3 Path selection policies (PSPs) and Native Multipath Plugins (NMPs)	124
5.4.4 Set default PSP.	125
5.4.5 Verifying Node ID path in vSphere web client.	126
5.4.6 Path failover behavior for an invalid path	127
5.5 VMware Distributed Resource Scheduler (DRS)	127
5.6 Naming conventions	130
5.7 VMware high availability	132
5.7.1 HA admission control	133
5.7.2 HA Heartbeat	133
5.7.3 HA advanced settings.	134
5.7.4 All Paths Down detection enhanced in vSphere 6	135
5.7.5 Permanent Device Loss (PDL)	136
5.7.6 Virtual Machine Component Protection (VMCP).	137
5.7.7 Storage failure detection flow	140
5.8 VMware vStorage API for Array Integration	140
5.9 Protecting vCenter Services	141
5.9.1 vCenter Availability Based Windows Server Failover Clustering (WSFC)	142
5.10 VMware recovery planning	143
5.10.1 VMware alternatives to minimize the impact of a complete site failure (“split brain” scenario).	144
5.10.2 Investigating a site failure	144
5.11 Design comments	146
5.12 Script examples.	146
5.12.1 PowerShell test script to move VMs 40 times between two ESXi hosts	147
5.12.2 PowerShell script to extract data from the entire environment and verify active and preferred paths	148
Chapter 6. Enhanced Stretched Cluster diagnostic and recovery guidelines	153
6.1 Solution recovery planning	154
6.2 Recovery planning	154
6.3 Enhanced Stretched Cluster diagnosis and recovery guidelines	160
6.3.1 Critical event scenarios and complete site or domain failure	160
6.3.2 Diagnosis guidelines.	161
6.3.3 Recovery guidelines	167
Related publications	177
IBM Redbooks	177
VMware online resources.	178
Other publications	178
Websites	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™	Redbooks (logo)  ®
DS8000®	IBM Spectrum Storage™	Storwize®
Easy Tier®	IBM Spectrum Virtualize™	System Storage®
FlashCopy®	Insight™	XIV®
Global Technology Services®	PowerHA®	z/OS®
HyperSwap®	Real-time Compression™	
IBM®	Redbooks®	

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.

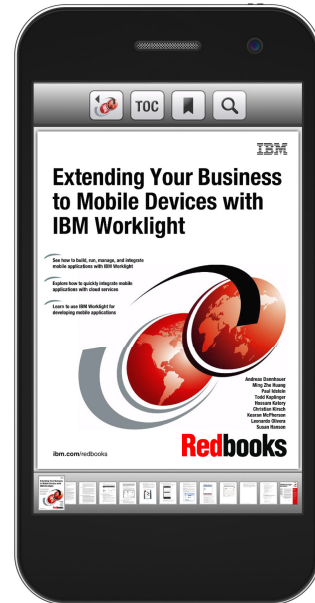
Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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
About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition might also include minor corrections and editorial changes that are not identified.

Summary of Changes
for SG24-8211-01
for IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware
as created or updated on December 19, 2016.

October 2015, Second Edition

This revision includes the following new and changed information.

New information

- ▶ Host site awareness
- ▶ Changes to VMware path selection

Preface

This IBM® Redbooks® publication describes the IBM storage area network (SAN) and IBM Spectrum™ Virtualize, and SAN Volume Controller Enhanced Stretched Cluster configuration when combined with VMware. It describes guidelines, settings, and implementation steps necessary to achieve a satisfactory implementation.

Business continuity and continuous availability of applications 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 cloud computing offers.

The IBM Enhanced Stretched Cluster design offers significant functions for maintaining business continuity in a VMware environment. You can dynamically move applications across data centers without interruption to those applications.

The live application mobility across data centers relies on these products and technologies:

- ▶ IBM Spectrum Virtualize™ and SAN Volume Controller Enhanced Stretched Cluster Solution
- ▶ VMware Metro vMotion for live migration of virtual machines
- ▶ A Layer 2 IP Network and storage networking infrastructure for high-performance traffic management
- ▶ Data center interconnection

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 joining 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, and 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.



Angelo Bernasconi is an Executive Certified, Storage, SAN, and Storage Virtualization IT Specialist. He is a member of IBM Italy TEC and 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 customers in z/OS®, and for the last 14 years he has focused on open systems. He holds a degree in electronics and his areas of expertise include storage hardware, SANs, storage virtualization design, solutions, implementation, DR solutions, and data deduplication. Angelo writes extensively about SAN and storage products. He is also member of the Storage Networking Industry Association SNIA Italian committee.



Antonio Rainero is a Consulting IT Specialist working for the 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 books for IBM. 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 advocate 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.

There are many people who contributed to this book and the previous edition. In particular, we thank the development and PFE teams in IBM Hursley, UK.

Chris Canto
Dave Carr
Robin Findlay
Carlos Fuente
Katja Gebuhr
Lucy Harris
Geoff Lane
Andrew Martin
Paul Merrison
Steve Randle
Bill Scales
Matt Smith
Barry Whyte
IBM Hursley, UK

Chris Saul
Bill Wiegand
IBM US

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

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

- ▶ Send your comments in an email to:

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 Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster (ESC) design offers significant functionality for maintaining business continuity in a VMware environment. You can dynamically migrate applications across data centers without interrupting the applications.

The live application mobility across data centers relies on these elements:

- ▶ IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster configuration
- ▶ VMware Metro vMotion for live migration of virtual machines
- ▶ A Layer 2 IP Network and storage networking infrastructure for high-performance traffic management
- ▶ Data center interconnection

This chapter includes the following sections:

- ▶ IBM Spectrum Virtualize
- ▶ Enhanced Stretched Cluster option
- ▶ Integration of Spectrum Virtualize, SAN Volume Controller, Layer 2 IP Network, Storage Networking infrastructure, and VMware
- ▶ Open Data Center Interoperable Network

1.1 IBM Spectrum Virtualize

IBM Spectrum Virtualize and SAN Volume Controller is a storage virtualization system that enables a single point of control for storage resources. It 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, while avoiding increases in administrative costs.

IBM Spectrum Virtualize and SAN Volume Controller: 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 the 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 speed of deployment and access to data for new workloads

The SAN Volume Controller and IBM Storwize® products all run the same software, and that software is what provides all of the features and functions of these products. Until the IBM Spectrum Storage family was announced earlier this year, that software did not have a name per se. With the IBM Spectrum Storage family, the software that powers the SAN Volume Controller and Storwize products now has the name “IBM Spectrum Virtualize”.

IBM Spectrum Virtualize and SAN Volume Controller 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. Spectrum Virtualize also supports Fibre Channel over Ethernet (FCoE) protocol.

IBM Spectrum Virtualize and SAN Volume Controller combines hardware and software in an integrated, modular solution that is highly scalable. An I/O group is formed by combining a redundant pair of storage engines that are based on IBM System x server technology. Highly available I/O groups are the basic configuration element of a Spectrum Virtualize cluster.

The configuration flexibility means that your implementation can start small and grow with your business to manage very 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 configuration supports high performance for critical applications.

IBM Spectrum Virtualize also includes the IBM System Storage® Easy Tier® function, which helps improve performance at lower cost through 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, and helps move critical data to and from SSDs as needed without disruption to applications.

It helps increase the amount of storage capacity that is available to host applications by pooling the capacity from multiple disk systems within the SAN. In addition, it combines various IBM technologies that include thin provisioning, automated tiering, storage virtualization, IBM Real-time Compression™, clustering, replication, multiprotocol support, and an updated graphical user interface (GUI). Together, these technologies enable a IBM Spectrum Virtualize and SAN Volume Controller to deliver exceptional storage efficiency.

Because this configuration hides the physical characteristics of storage from host systems, it also helps applications continue to run without disruption while you change your storage infrastructure. This helps your business improve customer service.

Figure 1-1 shows an overview of IBM Spectrum Virtualize and SAN Volume Controller.

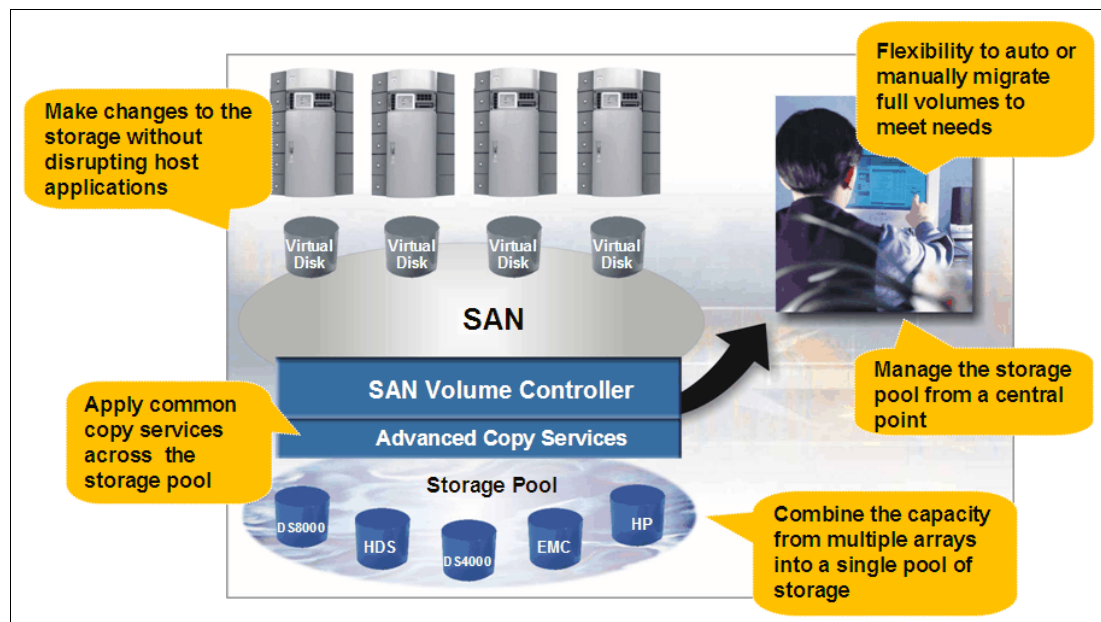


Figure 1-1 IBM Spectrum Virtualize and SAN Volume Controller

IBM Spectrum Virtualize and SAN Volume Controller includes a dynamic data migration function to move data from one storage system to another, yet maintain access to the data. The Volume Mirroring function stores two copies of a volume on different storage systems. This configuration helps improve application availability during a failure or disruptive maintenance to an array or disk system. The controller's stretched cluster configuration automatically uses whichever copy of the data remains available.

With the stretched cluster, administrators can apply a single set of advanced network-based replication services that operate in a consistent manner. This set is applied regardless of the type of storage that is being used. The Metro Mirror and Global Mirror functions operate between systems at different locations. They help create copies of data for use during a catastrophic event at a data center. For even greater flexibility, Metro Mirror and Global Mirror also support replication between SAN Volume Controller systems and IBM Storwize V7000 Unified systems.

The IBM FlashCopy® function stretched cluster quickly creates a copy of active data that can be used for backup purposes or for parallel processing activities. This capability enables disk

backup copies to be used to recover almost instantly from corrupted data, which significantly speeds application recovery.

Figure 1-2 shows the SAN Volume Controller structure and components.

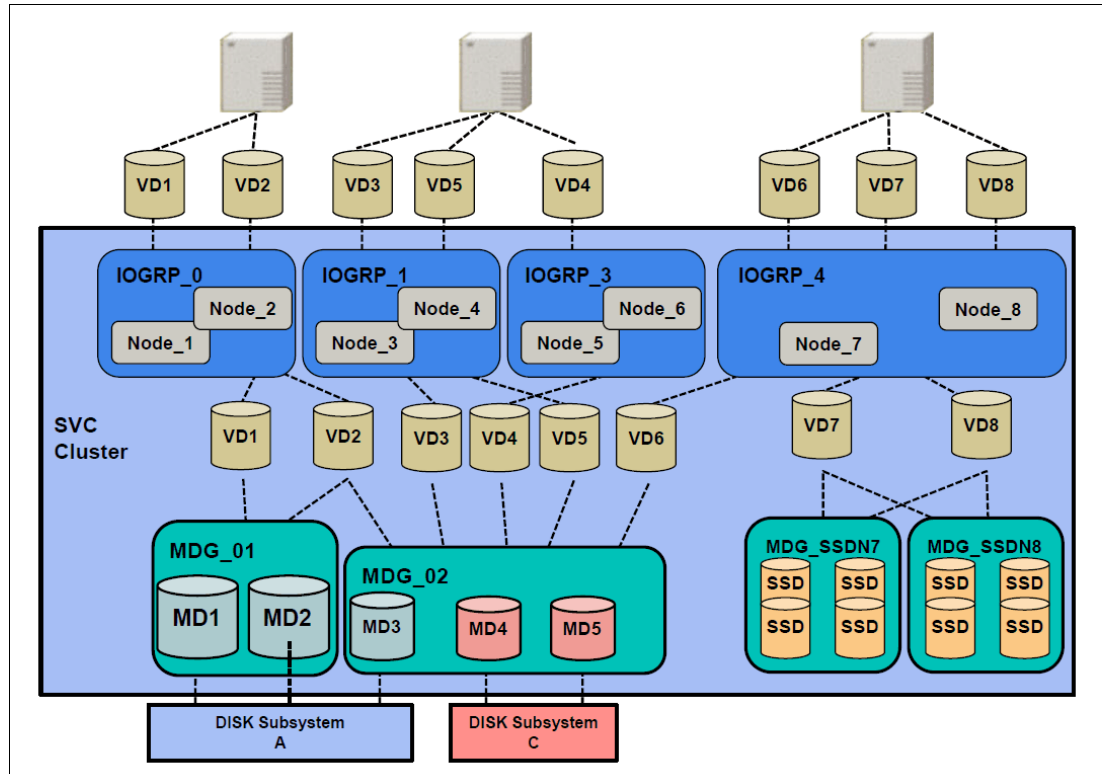


Figure 1-2 SAN Volume Controller structure

1.2 Enhanced Stretched Cluster option

When it was first introduced, all nodes were supposed to be installed in the same physical location with a maximum supported distance of 100 meters between the nodes within an I/O group. Software version 5.1 introduced support for the stretched cluster configuration where nodes within an I/O group can be separated by a distance of up to 10 km by using specific Long Wave SFPs. This distance limitation came from the restriction that all communication between node ports could not traverse inter-switch links (ISLs) which in fact limited the maximum supported distance between the nodes physical locations. Starting with software version 6.3, which was released in October 2011, the ISL restriction was removed, which allowed the distance between the nodes physical locations to be extended to 300 km.

Combining stretched cluster configurations with the Volume Mirroring capability provides a continuous availability platform whereby host access is maintained during the loss of any single location.

With the stretched cluster configurations, the two nodes in an I/O group are separated by the distance between two locations. If a copy of the volume is stored at each location, you can lose either the SAN or power at one location, and access to the disks remains available at the alternate location. Using 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 provides the capability to keep both copies of the storage in synchronization, while the

cache is mirrored between both nodes. Therefore, the loss of one location causes no disruption to the alternate location.

Software version 7.2 introduced the *Enhanced Stretched Cluster* feature (ESC) that further improved the stretched cluster configurations introducing the *site awareness* concept for nodes and external storage, and the *DR feature* that allows to manage effectively rolling disaster scenarios.

Finally, with Spectrum Virtualize version 7.5, the site awareness concept has been extended to hosts allowing more efficiency for host IO traffic through the SAN and easier host path management.

With Spectrum Virtualize version 7.6, the IP quorum concept was introduced using a CLI only configuration option, and in version 7.7 configuration options have been incorporated into the GUI.

Two important concepts behind the stretched cluster configurations are the quorum disk and volume mirroring.

As with any clustering solution, avoiding a “split brain” situation (where nodes are no longer able to communicate with each other) requires a tie break. Spectrum Virtualize is no exception and uses a tie-break 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 and the management of the quorum disks is the responsibility of the users.

The Enhanced Stretched Cluster design can automatically choose quorum disks and place one 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, the system will select automatically the IP Quorum App to be the active one.

Volume mirroring was introduced in software version 4.3. This feature allows a single volume to have two physical copies of the data on two independent managed disk (MDisk) groups, such as storage pools or storage controllers.

This feature provides these capabilities:

- ▶ A way to change the extent size of a volume
- ▶ Another way to migrate between storage controllers or to split off a copy of a volume for development or test purposes.
- ▶ A way to increase redundancy and reliability of lower-cost storage controllers
- ▶ A temporary mechanism to add a second copy to a set of volumes to run disruptive maintenance to a storage controller without any loss of access to servers and applications

Another capability that Volume Mirroring provides is the ability to split the cluster, yet maintain access to clustered servers and applications.

For example, imagine that you have two servers that act as a cluster for an application. These two servers are in different rooms and power domains, and they are attached to different fabrics. You also have two storage controllers, one in each room. You want to mirror data between the controllers and, at the same time, provide access to users when you lose power or access to disks within one of the rooms. You can now do this by implementing the ESC configuration.

A number of requirements must be validated for the Enhanced Stretched Cluster implementations, in particular regarding SAN extension. For more information about ESC prerequisites, see the IBM SAN Volume Controller Knowledge Center at:

<https://ibm.biz/BdXuvS>

You can use the storage controller of your choice at any of the three locations, and they can be from different vendors. This is possible by using the base Spectrum Virtualize virtualization license, which is available at no additional charge. The list of supported hardware and other interoperability information can be retrieved at the following link:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1005253>

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

1.3 Integration of Spectrum Virtualize, SAN Volume Controller, Layer 2 IP Network, Storage Networking infrastructure, and VMware

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 being moved to virtualized environments. This process creates requirements for higher availability, protection of critical business data, and the ability to fail over and continue supporting business operations in a local outage or a widespread disaster.

VMware vMotion is a feature of VMware's ESXi servers that allows the live migration of virtual machines from one ESXi server to another with no application downtime. Typically, vMotion is used to keep IT environments up and running, which provides unprecedented flexibility and availability to meet the increasing demands for data.

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. 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
- ▶ Maintaining the correct level of performance for applications

vMotion over distance, spanning data centers or geographical boundaries, requires a specialized infrastructure or environment with these three key capabilities:

- ▶ Data synchronization between data centers to allow servers, regardless of their locations, to always have access to that data
- ▶ A network infrastructure that provides high performance, high reliability, and correct Layer 2 extension capabilities to connect the data centers
- ▶ IP traffic management of client network access to the application server site

The solution described in this book addresses these needs through the combination of VMware Metro vMotion with the ESC capabilities. This combination runs over a Layer 2 IP network and storage infrastructure.

Continuous access to data is provided by an ESC configuration and Volume Mirroring capability.

The Layer 2 IP Network and storage networking infrastructure provides a reliable and high-performance, end-to-end solution with network adapters, edge, aggregation, and core switching. It also offers high-performance application delivery controllers. This combination provides a flexible infrastructure that results in simplification, reduced costs, higher resource use, and, most important, data protection and resiliency.

The combination of VMware Metro vMotion, ESC, and the Layer 2 IP networking infrastructure enables the design and implementation of a robust business continuity, disaster avoidance, and recovery solution for virtualized application environments.

1.3.1 Application mobility over distance

VMware vMotion uses the VMware clustered file system, the Virtual Machine File System (VMFS), to enable access to virtual storage. The underlying storage that is used by the VMFS data store is one or more volumes supplied by SAN Volume Controller that are accessible by all of the vSphere hosts.

During vMotion, the active memory and precise running state of a virtual machine are rapidly transmitted over a high-speed network from one physical server to another. Access to the virtual machine's disk storage is instantly switched to the new physical host. The virtual machine retains its network identity and connections after the vMotion operation, which ensures a seamless migration process. Using the powerful features of Metro vMotion to migrate VMs over an extended distance creates a new paradigm for business continuity. This enables newer data center functions, such as disaster avoidance, data center load balancing, and data center resource (power and cooling) optimization.

For more information about VMware practices, see the *VMware Metro Storage Cluster: (vMSC)* white paper at:

<http://ibm.biz/Bdx4gq>

1.3.2 Benefits of this integrated solution

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

- ▶ Primary benefits:
 - Fully active-active data centers with balanced workloads
 - Disaster and downtime avoidance
- ▶ Secondary benefit:
 - Can be used during disaster recovery 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 can benefit from it, but not as much as a single Tier 0

Some VMware products can be used to help protect against loss of a data center. You must check whether they are supported in your environment.

- ▶ VMware vCenter Site recovery manager 5.x:
<http://ibm.biz/Bdx4gv>
- ▶ VMware vCenter Server Heartbeat:
<http://ibm.biz/Bdx4ga>

Figure 1-3 shows the VMware vMotion configuration.

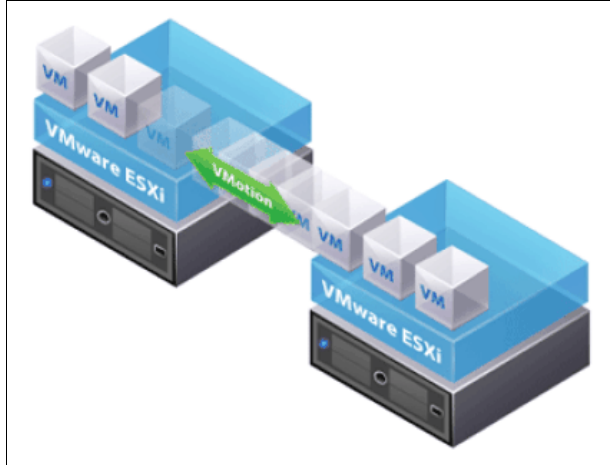


Figure 1-3 VMware vMotion configuration

1.3.3 Benefits in more detail

VMWare provides these benefits:

- ▶ Disaster avoidance

vMotion over distance allows IT managers 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.

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

- ▶ User performance and load balancing between data centers

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

- ▶ Optimization to decrease power costs

This helps your business take advantage of energy price volatility between data center regions and time of use. The dynamic move of VMs to data centers with cheaper energy reduces costs by using vSphere Distributed Power Management (DPM).

- ▶ Zero maintenance downtime

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 have continuous access to applications.

- ▶ Disaster recovery test

A disaster recovery test can be run on live applications without affecting the business. You can also use this process to completely test the functions of the disaster recovery site with an actual user load.

Figure 1-4 shows a VMware and ESC.

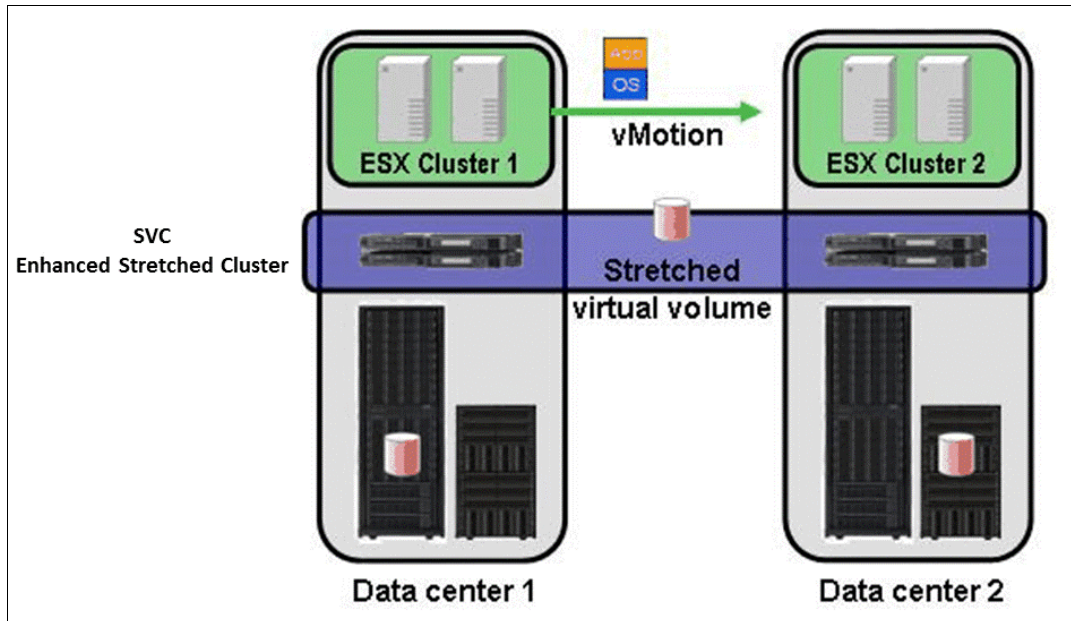


Figure 1-4 ESC and VMware

1.3.4 When to use VMware stretched clusters

Use VMware stretched clusters in these situations:

- ▶ When there is a requirement for intersite nondisruptive mobility of workloads between active-active data centers
- ▶ When there are proximate data centers with high-speed, low-latency links that can give less than 10 ms round-trip time (RTT)
- ▶ When you are enabling multisite load balancing
- ▶ When you are increasing availability of workloads through partial or complete site subsystem failures (that is, to recover from a total network, storage, or host chassis failure at a site)

1.3.5 When not to use VMware stretched clusters

Generally, avoid using VMware Stretched clusters in these situations:

- ▶ When orchestrated and complex reactive recovery is required
- ▶ When the distance between sites is long (100 km or more)
- ▶ When there are highly customized environments with rapid changes in configuration

- ▶ When environments require consistent, repeatable, and testable recovery time objectives
- ▶ If there are environments where both data centers might simultaneously be hit with a common disaster
- ▶ When disaster recovery compliance must be shown through audit trails and repeatable processes

For more information, see VMware's guide titled *Stretched Clusters and VMware vCenter Site Recovery Manager* at:

<http://ibm.biz/Bdx4gy>

1.3.6 IBM, VMware, and Layer 2 IP network solution

Building the correct storage and network infrastructure to enable data and application mobility requires a data center infrastructure that can provide both optimal storage extension capabilities and advanced network functionality. Design a comprehensive solution that includes storage, network, and server infrastructure, and implement it to facilitate the movement of applications across data centers. These products from IBM and VMware have been validated in the joint solution:

- ▶ VMware vSphere 5 with Enterprise Plus licensing to enable Metro vMotion
- ▶ ESC configuration to ensure the availability of access to the storage in both data centers
- ▶ Layer 2 IP network switch, as described in Chapter 2, "Hardware and software descriptions" on page 13
- ▶ IBM SAN FC switch and director family products for FC connectivity and FC/IP connectivity, as described in Chapter 2, "Hardware and software descriptions" on page 13

1.4 Open Data Center Interoperable Network

IBM experts took the lead to define the issues related to agile data center networks and to promote preferred practices by creating a roadmap that is based on the Open Data Center Interoperable Network industry standard, which is known as ODIN.

The practical, cost-effective evolution of data center networks needs to be based on open industry standards. This can be a challenging and often confusing proposition because there are so many different emerging network architectures, both standard and proprietary. The ODIN materials created by IBM developers address issues such as virtualization and virtual machine (VM) migration across flat Layer 2 networks, lossless Ethernet, software-defined networking (SDN), the OpenFlow communications protocol, and extended-distance WAN connectivity (including low latency).

ODIN provides the following benefits:

- ▶ Customer choice and forward-thinking designs for data center networks, including vendor-neutral requests for quote (RFQs)
- ▶ Lower total cost of ownership (TCO) by enabling a multivendor network (for more information, see the Gartner Group analysis, "Debunking the myth of the single-vendor network", 17 November 2010)
- ▶ Avoiding confusion in the marketplace between proprietary and vendor-neutral solutions
- ▶ Providing guidelines, relative maturity, and interpretation of networking standards

These IBM developed solutions are based on open industry standards:

- ▶ Fibre Channel and FC-IP storage technologies
- ▶ Lossless Ethernet
- ▶ Flat, Layer 2 networks
- ▶ Distance extension options that use dark fiber Wavelength Division Multiplexing (WDM), and Virtual Private LAN Service (VPLS) or Multiprotocol Label Switching (MPLS)

For more information, see:

<http://www.odin.com/ibm/>

Figure 1-5 shows a typical ODIN solution.

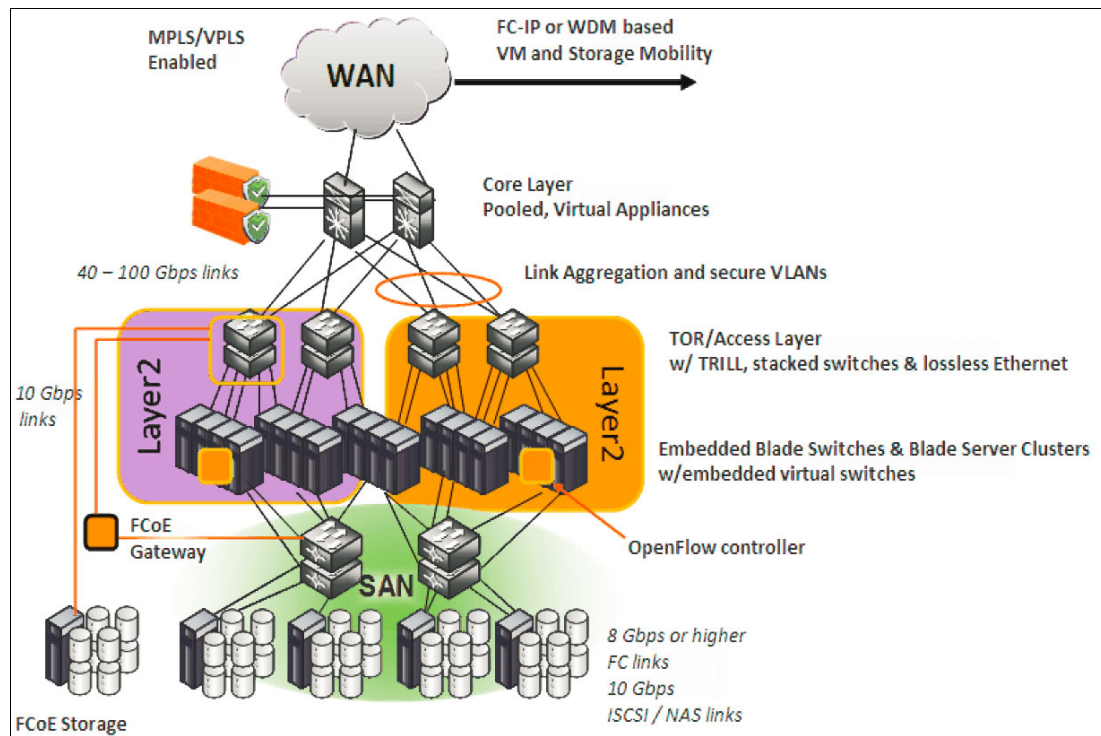


Figure 1-5 ODIN

ODIN has these major targets:

- ▶ Support customer choice and forward-thinking network design choices, including vendor-neutral RFQs
- ▶ Lowering TCO, as explained in the Gartner Group analysis, “Debunking the myth of the single-vendor network,” 17 November 2010
- ▶ Avoid confusion in the marketplace between proprietary and vendor-neutral solutions
- ▶ Provide guidelines, relative maturity, and interpretation of the many networking standards
- ▶ Give clients a single, trusted source for understanding standards-based networking requirements and a voice in what those requirements will look like in the future

The ESC solution complies with ODIN.



Hardware and software descriptions

This chapter describes hardware that is needed to implement an IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster. It also briefly describes VMware and some of the features that are useful when you are implementing an Enhanced Stretched Cluster.

This chapter includes the following sections:

- ▶ Hardware description
- ▶ IBM System Storage Spectrum Virtualize and SAN Volume Controller
- ▶ SAN directors and switches
- ▶ FCIP routers
- ▶ Ethernet switches and routers
- ▶ Software high availability
- ▶ VMware ESX and VMware ESXi

2.1 Hardware description

The following sections concentrate on the hardware that you need when you implement an Enhanced Stretched Cluster (ESC). All of the products that are mentioned can provide the functions that are necessary to implement an ESC. It is up to you to choose the most suitable product for your environment.

Consider these hardware factors when you are implementing ESC:

- ▶ Distance and latency between data centers
- ▶ Connectivity between data centers
- ▶ Bandwidth of sent data
- ▶ Customer budget
- ▶ Current customer infrastructure

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

2.2 IBM System Storage Spectrum Virtualize and SAN Volume Controller

As described in Chapter 1, “Introduction” on page 1, an ESC configuration requires the use of the IBM Spectrum Virtualize and SAN Volume Controller. The controller provides an active-active storage interface that can allow for simple fail over and fail back capabilities during a site disruption or failure.

IBM Storwize V7000 *cannot* be configured in an ESC configuration.

To implement an ESC over 4 km, use either the 2145-CF8, 2145-CG8, or 2145-DH8 hardware models of the controllers (which is shown in Figure 2-1). Use these models because they have increased node capabilities. 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, “Enhanced Stretched Cluster architecture” on page 27 to ensure that the SAN Volume Controller node’s model and firmware version are supported for what you want to implement.



Figure 2-1 2145-DH8 nodes

2.3 SAN directors and switches

To implement an ESC solution, any 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, “Enhanced Stretched Cluster architecture” on page 27.

This section does not address any particular Wavelength Division Multiplexing (WDM) devices or any Ethernet infrastructure options other than Fibre Channel 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, or 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 providing nonstop access to information and reducing 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) Fibre Channel capabilities and deliver a new level of scalability and advanced capabilities to this reliable, high-performance technology.

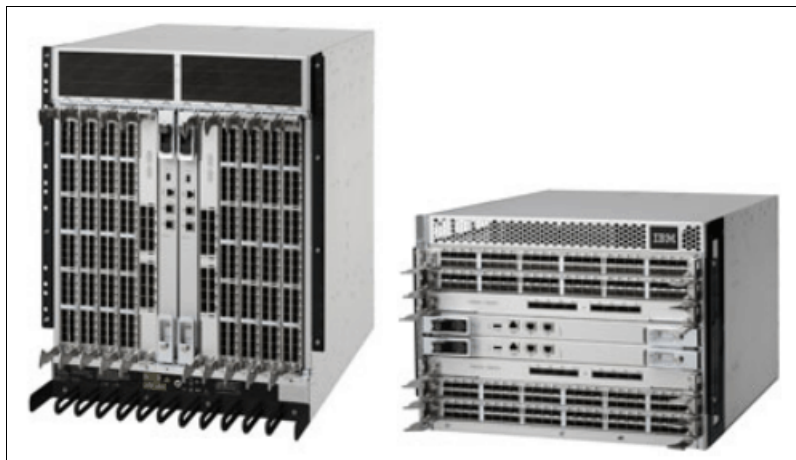


Figure 2-2 SAN768B-2 and SAN384B-2 fabric backbones

Both directors are capable of 16, 10, 8, 4, and 2 Gbps connections, and 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 an ESC solution over 10 km. The Trunking feature is necessary if multiple links are required to accommodate the bandwidth that is used for SAN traffic.

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

<http://www.ibm.com/systems/storage/san/b-type/>

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

IBM System Storage offers a wide range of Fibre Channel 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 is able to be configured with 12 or 24 active ports. It is capable of 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 SAN Volume Controller node traffic.

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



Figure 2-3 SAN24B-5 switch

When you are implementing ESC solutions over 10 km with the SAN24B-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 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 is capable of 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 are implementing ESC solutions over 10 km 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, shown in Figure 2-5, is capable of 1, 2, 4, and 8 Gbps speeds. High availability features make this an ideal candidate for a core switch in medium-sized environments and an edge switch in larger enterprise environments.



Figure 2-5 SAN96B-5 switch

The Extended Fabric feature is enabled on the SAN96B-5 by default, which makes it useful for ESC configurations over 10 km. 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 are implementing an ESC over long distances, it is not always possible or feasible to extend SAN fabrics by using direct Fibre Channel connectivity or WDM. Either the distance between the two failure domains is over 10 km or it is too expensive to lay cable or hire dark fiber service.

Many dual data center environments already have IP connections between data centers. This configuration allows FCIP technologies to be used to enable the SAN fabric to extend across data centers while using the existing infrastructure. When you are implementing an ESC with FCIP, minimum bandwidth requirements must be met to support the solutions. For more information, see Chapter 3, “Enhanced Stretched Cluster architecture” on page 27.

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 GbE technology to enable fast, reliable, and cost-effective remote data replication, backup, and migration with existing Ethernet infrastructures.

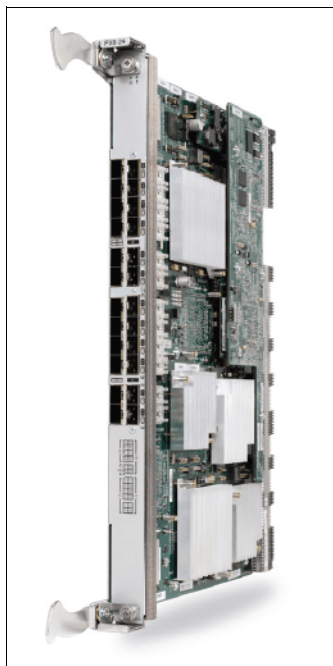


Figure 2-6 8 Gbps Extension Blade

The 8 Gbps Extension Blade has twelve 8 Gbps Fibre Channel 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 on the SAN384B-2 and SAN 768B-2 directors feature.

2.4.2 SAN06B-R extension switch

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

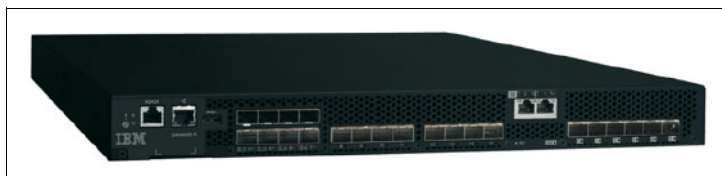


Figure 2-7 SAN06B-R extension switch

The SAN06B-R extension switch provides up to sixteen 8 Gbps Fibre Channel 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 Ethernet switches and routers

To support vMotion over long distances, a scalable and robust IP network must be available to the vSphere hosts for data connectivity and to the SAN FCIP Extension devices for storage traffic over FCIP. Layer 2 extension between the data centers is also required to enable vMotion support. This configuration can be accomplished readily with standards-compliant Multiprotocol Label Switching (MPLS), Virtual Private LAN Service (VPLS), or Virtual Leased Line (VLL) technology.

2.5.1 IBM System Networking switches

The following networking switches can be used in an ESC.

IBM System Networking RackSwitch G8124E

The IBM RackSwitch G8124E is a 10-Gigabit Ethernet switch that is specifically designed for the data center. It provides a virtualized, cooler, easier network solution. Designed with top performance in mind, the G8124E provides line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. It also provides large data center grade buffers to keep traffic moving.

Figure 2-8 shows the IBM RackSwitch G8124E.



Figure 2-8 IBM RackSwitch G8124E

The G8124E offers twenty-four 10-Gigabit Ethernet ports in a high-density, 1U footprint, which makes it ideal for data center top of rack switching.

IBM System Networking RackSwitch G8264 and RackSwitch G8264T

Designed with top performance in mind, the IBM RackSwitch G8264 and IBM RackSwitch G8264T are ideal for today's big data, cloud, and optimized workloads. Both are enterprise class and full-featured data center switches that deliver line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. The RackSwitch G8264 and G824T are ideal for latency-sensitive applications and to support IBM Virtual Fabric. These characteristics help you reduce the number of I/O adapters for a single dual-port 10 Gb adapter, which reduces cost and complexity.

Figure 2-9 shows the IBM RackSwitch G8264T.

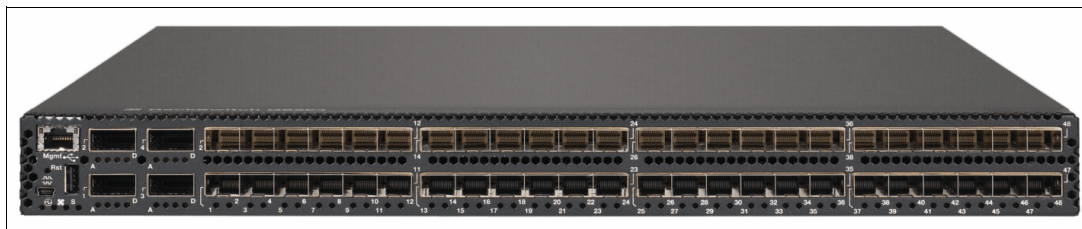


Figure 2-9 IBM RackSwitch G8264T

The RackSwitch G8264 supports up to $48 \times 1/10$ Gb SFP+ ports and 4×40 Gb QSFP+ ports. The RackSwitch G8264T supports up to 48×10 G Base-T ports and 4×40 Gb QSFP+ ports. This number of ports makes these switches ideal for consolidating many racks of infrastructure into a single pair of redundant switches. They provide high-speed inter-switch links with 40 Gb link speeds.

2.5.2 Brocade IP routers and Layer 4 - 7 application delivery controllers

The following routers and application delivery controllers can be used in an ESC.

Brocade MLX Series routers

The Brocade MLX Series of high-performance routers provides a rich set of high-performance IPv4, IPv6, and MPLS capabilities. They also have advanced Layer 2 switching capabilities. The Brocade MLX Series routers provide a high-density, highly available, and scalable IP network aggregation in the data center core. They enable Layer 2 data center extension by using MPLS, VPLS, and VLL.

Figure 2-10 shows the Brocade MLX Series of high-performance routers.



Figure 2-10 Brocade MLX Series routers

The Brocade MLX Series routers are available in 4-, 8-, 16-, and 32-slot options. They are designed with a fully distributed, non-blocking architecture with up to 15.36 Tbps fabric capacity, providing packet forwarding rates of approximately 5 billion packets per second. The 32-slot chassis supports up to 1,536 1 GbE, 256 10 GbE, and 32 100 GbE wire-speed ports.

Brocade NetIron CES2000

The Brocade NetIron CES 2000 Series offers compact 1U, multiservice edge and aggregation switches that combine powerful capabilities with high performance and availability. The switches provide a broad set of advanced Layer 2, IPv4, IPv6, and MPLS capabilities in the same device.

The Brocade NetIron CES 2000 Series is shown in Figure 2-11.



Figure 2-11 Brocade Netron CES2000

The Brocade Netron CES 2000 Series switches are available in 24-port and 48-port 1 GbE configurations. Both have two 10 GbE uplinks in both Hybrid Fiber (HF) and RJ45 versions to suit various deployment needs.

Brocade ADX Series L4-L7 application delivery controllers

The Brocade ADX Series, shown in Figure 2-12, enables high-speed application delivery by using a purpose-built architecture that is designed with high core density and embedded application accelerators. With the support of advanced traffic management, an open application scripting engine, and extensible application programming interfaces (APIs), the Brocade ADX Series of application delivery switches optimizes service delivery.



Figure 2-12 Brocade ADX Series L4-L7 Application Deliver Controllers

The Brocade ADX Series provides Global Server Load Balancing (GSLB). GSLB provides host and application health checks and directs new client connections to the correct data center location after a VM has been moved. It works in tandem with the Brocade Application Resource Broker.

The Application Resource Broker is a plug-in for vSphere that enables automated provisioning of virtual machines (VMs). It simplifies the management of application resources and ensures optimal application performance by dynamically adding and removing application resources within globally distributed data centers. Application Resource Broker provides these capabilities through real-time monitoring of application resource responsiveness, traffic load information, and infrastructure capacity information from server infrastructures.

The Brocade ADX Series is available in 1 RU devices with up to 24 1-GbE ports and two 10-GbE uplinks. It is also available as a 4-slot or 8-slot chassis with the largest supporting up to 48 1-GbE ports and 16 10-GbE ports.

2.6 Software high availability

When you are implementing a solution such as an ESC, make sure to 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 VMware's ESXi and vSphere platforms provide. This section outlines VMware ESXi and vSphere and some other features that are useful when you are implementing a stretched cluster solution.

2.7 VMware ESX and VMware ESXi

VMware ESX and VMware ESXi are hypervisors that allow you 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 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 what virtual machines are running on what physical infrastructure.

2.7.1 VMware vSphere

VMware vSphere is the management software suite that is used to manage the virtual machines inside an ESX or ESXi host. When you are allocating resources such as memory, storage, networking, or processors to a virtual machine, 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 has several features that allow for mobility of VMs between ESXi hosts and storage. These features can add to the availability of the VMs that are running in a cluster.

2.7.2 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. It 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. It does so through live migration of virtual machines 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 is moving between physical servers.

In an ESC 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 requires an outage for maintenance.

For more information, see VMware's 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. It does so by enabling live migration of virtual machine disks with no disruption to users or loss of service. The vCenter server manages the migration of data from one data store to another. With vStorage 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 an ESC solution, this feature is useful for moving a VM's virtual machine 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.

2.7.3 vSphere High Availability

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

For more information about vSphere High Availability, see this web page:

<http://www.vmware.com>

2.7.4 VMware vCenter Site Recovery Manager

Site Recovery Manager integrates with VMware vSphere, VMware vCenter Server, and underlying storage replication products to automate end-to-end recovery processes for virtual applications. It provides a simple interface for setting up recovery plans that are coordinated across all infrastructure layers. Recovery plans can be tested non-disruptively as frequently as required to ensure that the plan meets availability objectives. During a domain failover or migration, Site Recovery Manager automates both the fail over and fail back processes. It ensures fast and highly predictable recovery point objectives (RPOs) and recovery time objectives (RTOs).

For Version 5 and later, see the VMware Compatibility Guide at:

http://partnerweb.vmware.com/comp_guide2/search.php?deviceCategory=sra

2.7.5 VMware Distributed Resource Scheduler

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

VMware DRS allocates resources by using a set of user-defined rules and policies. These rules and policies can be used to make critical or high-performing VMs high-priority to ensure that particular VMs never run on the same storage or host, or to save on power and cooling costs by powering off ESXi servers that are not currently needed.

For more information, see VMware's Distributed Resource Scheduler, Distributed Power Management web page at:

<http://www.vmware.com>



Enhanced Stretched Cluster architecture

This chapter focuses on the IBM Spectrum Virtualize and SAN Volume Controller architecture as it applies to the Enhanced Stretched Cluster configuration. It is based on the assumption that you have a general understanding of the IBM Spectrum Virtualize and SAN Volume Controller architecture.

This chapter includes the following sections:

- ▶ Enhanced Stretched Cluster overview
- ▶ Failure domains, sites, and controllers
- ▶ Spectrum Virtualize volume mirroring
- ▶ Spectrum Virtualize enhanced stretched cluster configurations
- ▶ Fibre Channel settings for distance
- ▶ Spectrum Virtualize I/O operations on mirrored volumes

3.1 Enhanced Stretched Cluster overview

In a standard IBM Spectrum Virtualize and SAN Volume Controller configuration, all nodes are physically located within the same rack. Version 5.1 and later provide support for stretched cluster configurations, where nodes within an I/O group can be physically separated from one another by up to 10 km. This capability allows nodes to be placed in separate failure domains or sites, which provides protection against failures that affect a single failure domain or site. That initial support included the restriction that all communication between SAN Volume Controller node ports cannot traverse inter-switch links (ISLs). This limited the maximum supported distance between failure domains or sites. Starting with software version 6.3, the ISL restriction was removed, which allowed the distance between failure domains to be extended to 300 km. Additionally, in software version 6.3, the maximum supported distance for non-ISL configurations was extended to 40 km.

The Enhanced Stretched Cluster configuration provides a continuous availability platform, whereby host access is maintained during the loss of any single failure domain. This availability is accomplished through the inherent active/active architecture of IBM Spectrum Virtualize and SAN Volume Controller along with the use of volume mirroring. During a failure, the SAN Volume Controller nodes and associated mirror copy of the data remain online and available to service all host IO.

The existing stretched cluster configuration has two locations, each with one node from each I/O group pair. The quorum disks are usually held in a third location. This solution is quite successful in the field for many IBM clients.

It allows SAN Volume Controller hardware to be geographically distributed, resembling how IBM Metro Mirror is deployed. Stretched cluster deployments can span as short a distance as two racks in a data center, two buildings in a campus, across a city, or as far as 100 km or potentially more.

The key benefit of a stretched cluster, compared to Metro Mirror, is that it allows for fast non-disruptive failover during small scale outages. For example, if only a single storage device is affected, SAN Volume Controller fails over internally with minimal delay. If there is a failure in a fabric element, or SAN Volume Controller node, a host can fail over to another SAN Volume Controller node and continue performing I/O.

One of the attractions of stretched cluster, and a key advantage over some of the alternatives, is that the failover uses the same multi-pathing driver as is used with conventional “unstretched” deployments. This gives a wide interoperability matrix, matching the hosts and controllers that clients have already deployed in their data centers. Many of the multi-pathing drivers are also the default/standard multi-pathing drivers for the OS in question.

Another aspect is that the system always has an automatic quorum to act as a tie-break. This means that no external management software or human intervention is ever required to perform a failover.

Enhanced Stretched Cluster (ESC) was released in Version 7.2, which kept all of the previous stretched cluster functions and benefits, and added high availability (HA), failure and recovery capabilities, and bandwidth savings.

Remember these key points about the ESC:

- ▶ Use of ESC is optional. Existing stretched cluster configurations are still supported. However, using the new feature is recommended for its benefits.
- ▶ For the cluster topology that was introduced, a cluster is either *standard* or *stretched*. Configure it by using the `chsystem` command, or view it by using the `lssystem` command.

- ▶ Topology set to **stretched** enables site awareness features and disaster recovery (DR) capability.
- ▶ It is possible to convert an existing stretched cluster to an ESC to change the topology non-disruptively any time after upgrade to Version 7.2.
- ▶ The ESC must be configured in advance to use the disaster recovery features.
- ▶ Use of solid-state drives (SSDs) with SAN Volume Controller nodes is not supported in an ESC deployment.

With Spectrum Virtualize Version 7.5, host site awareness for hosts is now supported. This new host object property is better described in 3.2, “Failure domains, sites, and controllers”.

3.2 Failure domains, sites, and controllers

In an ESC configuration, the term *failure domain* is used to identify components of the cluster that are contained within a boundary. In this configuration, any failure that occurs (such as power failure, fire, or flood) is contained within that boundary. Failure domain is also referred to as *failure site*. The failure cannot propagate or affect components that are outside of that boundary. The components that comprise an ESC configuration must span three independent failure domains. Two failure domains contain SAN Volume Controller nodes and 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 SAN Volume Controller can maintain availability if any single power source fails.
- ▶ If each site is a different building, the SAN Volume Controller can maintain availability if a loss of any single building were to occur (for example, power failure or fire).

Ideally, each of the three failure domains that are used for the ESC 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. Some 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 SANs (separate cabling and switches).
- ▶ All cabling (power, SAN, and IP) from one failure domain must not be physically routed through another failure domain.
- ▶ Each failure domain must be placed in separate fire compartments.

Remember: The key prerequisite for failure domains is that each node from an I/O group must be placed in a separate failure domain. Each I/O group within the SAN Volume Controller cluster must adhere to this rule.

Version 7.2 introduced a *site awareness* concept for nodes and controller. Version 7.5 introduced the *site awareness* concept for hosts too.

- ▶ Site awareness can be used only when topology is set to **stretched**.
- ▶ Topology set to **stretched** also means that the DR feature is enabled.
- ▶ 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. Sites 1 and 2 are where the two halves of the ESC are located. Site 3 is the optional third site for a quorum tie-breaker disk.
- ▶ A Site field is added to nodes and controllers. You can only set it by using these Spectrum Virtualize command-line interface commands (no GUI allowed in this initial release): **addnode**, **chnode**, and **chcontroller**. The nodes and controller must have sites set in advance, before you set Topology to **stretched**, and must have a site assigned.
- ▶ A Site field is added to host. You can set it only by using these Spectrum Virtualize command-line interface commands (no GUI available in this release): **mkhost** and **chhost**. The host must have sites set in advance, before you set Topology to **stretched**, and must have a site assigned.
- ▶ You can view site fields by using **lnode**, **lshost**, **lscontroller**, **lsmdisk**, or **lsmdiskgrp** commands.
- ▶ Nodes and Hosts can be assigned only to sites 1 or 2. Nodes and Hosts cannot be assigned to site 3.
- ▶ You can specify the site for each controller. The default for a controller is for its site to be undefined. This is the default for pre-existing controllers for upgrades to Version 7.2. Controllers can be assigned to sites 1, 2, or 3, or can be set to **undefined** again.
- ▶ An MDisk derives its site value from the controller that it is associated with it at that time. Some back-end storage devices are presented as multiple controller objects, and an MDisk might be associated with any of them from time to time. Make sure that all such controller objects have the same site specified to ensure that any MDisks 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 managed (or image mode) MDisks. The site for a controller cannot be changed when the DR feature is enabled if the controller uses managed (or image mode) MDisks.
- ▶ The site property for a controller adjusts the I/O routing and error reporting for connectivity between nodes and the associated MDisks. These changes are effective for any MDisk controller that has a site defined, 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, which means that any changes potentially require a Host reboot or FC HBA rescan, depending on the operating system used.

3.3 Spectrum Virtualize volume mirroring

The ESC configuration uses the volume mirroring function, which allows the creation of one volume with two copies of MDisk extents. If they are placed in different MDisk groups, the two data copies allow volume mirroring to eliminate impact to volume availability if one or more MDisks fails. The synchronization between both copies is incremental and is started automatically. A mirrored volume has the same functions and behavior as a standard volume.

In the software stack, volume mirroring is below the cache and copy services. Therefore, IBM FlashCopy and IBM Metro Mirror or Global Mirror do not recognize that a volume is mirrored. All operations that can be run on non-mirrored volumes can also be run on mirrored volumes. These operations include migration and expand or shrink operations.

As with nonmirrored volumes, each mirrored volume is owned by the preferred node within the I/O group. Therefore, the mirrored volume goes offline if the I/O group goes offline.

I/O traffic is routed to minimize I/O data flows. Data payloads are transferred only the minimum number of times. I/O protocol control messages do flow across the link, but these are small in comparison to the data payload.

Read operations can be run from either node in the I/O group. However, all read operations run to the local site copy if both copies are in sync.

3.3.1 Volume mirroring prerequisites

The three quorum disk candidates keep the status of the mirrored volume. The last status and the definition of primary and secondary volume copy (for read operations) are saved to the quorum disk. Therefore, an active quorum disk is required for volume mirroring. To ensure data consistency, Spectrum Virtualize disables mirrored volumes if access to all quorum disk candidates is lost. Quorum disk availability is critical for ESC configurations. Additionally, the allocation of bitmap memory is required before you enable volume mirroring. You can allocate memory by using the **chiogrp** command:

```
chiogrp -feature mirror -size memory_size io_group_name | io_group_id
```

The volume mirroring grain size is fixed at 256 KB. At this setting, one bit of the synchronization bitmap represents 256 KB of virtual capacity. A bitmap memory space of 1 MB is required for each 2 TB of mirrored volume capacity.

3.3.2 Read operations

Spectrum Virtualize volume mirroring functions implement a read algorithm with one copy that is designated as the primary for all read operations. Spectrum Virtualize reads the data from the primary copy and does not automatically distribute the read requests across both copies. The first copy that is created becomes the primary by default. You can change this setting by using the **chvdisk** command:

```
chvdisk -primary copyid vdiskname
```

Starting with software version 7.2 and then with version 7.5, if an ESC solution is implemented, the primary copy concepts are overridden. Accordingly, read operations run locally with site attributes assigned to each SAN Volume Controller node, controller, and host.

3.3.3 Write operations

Write operations are run on both mirror copies. The storage controller with the lowest performance determines the response time between the SAN Volume Controller and the storage controller back-end. The SAN Volume Controller cache can hide high back-end response times from the host up to a certain level.

If a back-end write fails or a copy goes offline, a bitmap file is used to track out-of-sync grains. As soon as the missing copy is back online, Spectrum Virtualize evaluates the changed bitmap and automatically resynchronizes both copies. The resynchronization process has a similar performance impact on the system as a FlashCopy background copy or volume migration. The resync bandwidth can be controlled with the `chvolume -syncrate` command. Volume access is not affected by the resynchronization process, and is run concurrently with host I/O. Setting the sync rate to 100% regains synchronization sooner after a site loss is recovered, but if it is used concurrently with regular production payloads, it can affect performance. Therefore, using `-syncrate` must be planned carefully, or the resynchronization must be scheduled for an optimal time frame.

The write behavior for the mirrored copies can cause difficulties when there is a loss of a failure domain, so it must be considered. Beginning with software version 6.2, it provides the `-mirrorwritepriority` volume attribute to prioritize between strict data redundancy (redundancy) and best performance (latency) for mirrored volumes. The `-mirrorwritepriority` attribute can be changed by using the `chvdisk` command:

```
chvdisk -mirrorwritepriority latency|redundancy
```

Important: At the time of writing, each node is responsible for running write operations on the back-end storage subsystem that has same site affinity. The data traverses the node-to-node connectivity only once for cache synchronization for all kinds of volumes. For compressed volumes, the data traverses the node-to-node connectivity a second time for the destage process.

It is important to have the correct size of the node-to-node connections to avoid creating a bottleneck. Consult an IBM specialist for assistance with the correct sizing.

Even if `mirrorwritepriority` is set to `latency` or `redundancy`, normally, cache pages are discarded after both writes are complete.

Remember: The default setting for `-mirrorwritepriority` is `latency`.

If a controller is slow in responding and the SAN Volume Controller starts error recovery (at around 5 seconds), that error recovery completes at 30 seconds without completing the I/O successfully. If that happens, the copy is marked stale, and it continues with only one copy, and the destage completes. This is with the goal of completing the host I/O at no later than 30 seconds.

If controller performance is poor, it might fail to respond to I/O for a long time. If that occurs, even the `redundant` operation marks it stale and proceeds with one copy.

Basically, the difference between `latency` and `redundancy` is how long the system waits and the probability of such an event.

In software version 7.2, the timings are adjusted for `latency` mode so that the overall time is about 40 seconds. This removes many of the causes of out-of-sync volumes conditions and achieves an overall timing guarantee that is in line with other error scenarios for ESCs. This means that the timeliness of `latency` can be obtained with less risk of out-of-sync volumes.

There are many possible causes of error recovery in the storage subsystem. IBM Spectrum Virtualize and SAN Volume Controller attempts to allow such error recovery to complete where it is expected to be quick. However, when error recovery takes longer to complete, with volume mirroring configured, Spectrum Virtualize fails the MDisk that is performing error recovery, and continues system operation with the surviving storage subsystem that is providing a timely response. This means that the failed storage subsystem cannot be used to provide a redundant copy if the surviving storage subsystem fails, including if its site fails.

The Spectrum Virtualize `mirrorwritepriority` setting provides two choices for how to balance timely I/O handling versus maintaining maximum redundancy:

- ▶ In **latency** mode, Spectrum Virtualize fails a controller that is performing error recovery and continues with a surviving controller to try to maintain an I/O timeout of 40 seconds when ESC is enabled.
- ▶ In **redundancy** mode, Spectrum Virtualize allows a failing controller more time and keeps the two redundant copies operating until it determines that a controller completely failed. This can be up to 3 minutes.

Tip: In a Spectrum Virtualize ESC configuration with the same kind of controllers, **latency** is the preferred setting for `-mirrorwritepriority`. This is because any potential error in the back-end controller has less impact on host performance.

Up until software version 7.4, whatever the configuration was, the node used by the host for the write operation was the preferred node. Since software version 7.5, with the introduction of host site awareness, write operations are managed by the node having the same host site affinity. This new attribute makes the configuration easier and faster because a multipathing software round robin fashion is used to optimize all the available paths belonging to the nodes having the same host site affinity.

3.3.4 Quorum disk

The quorum disk fulfills two functions for cluster reliability:

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

The quorum algorithm distinguishes between the active quorum disk and quorum disk candidates. There are three quorum disk candidates. 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 tiebreaker for split brain scenarios.

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.

If the DR feature is disabled, the quorum selection algorithm operates as software version 7.1 and previous versions do.

When the DR feature is enabled and automatic quorum disk selection is also enabled, three quorum disks are created, one in each site, in sites 1, 2, and 3.

If a site has no suitable MDisk, fewer than three quorum disks are automatically created. For example, if it can create only two quorum disks, only two are used.

If you are controlling the quorum by using the **chquorum** command, the choice of quorum disks must also follow the one-disk-per-site rule. If you used **chquorum** to manually assign quorum disks and configure the topology as stretched, the controller ignores any quorum disk that is not assigned to a site. Spectrum Virtualize 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 2 as stand-by quorum disks.

If you do not have a quorum disk configured at each site, that might restrict when, or if, a T3 recovery procedure is possible and how resilient the cluster is if a site failure occurs. Without access to a quorum disk, it cannot continue I/O operations when one copy of a mirrored volume goes offline.

Note: For stretched clusters implemented with the DR feature enabled, it is best to manually configure quorum devices to track which MDisk was chosen and to select the MDisks that you want to be your quorum disks.

3.3.5 Cluster state and voting

The cluster state information on the active quorum disk is used to decide which SAN Volume Controller nodes survive if exactly half the nodes 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 cluster manager implements a dynamic quorum. This means that, following a loss of nodes, if the cluster is able to continue operation, it dynamically alters the voting set that defines which nodes 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 is 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 cluster, the quorum is defined in the following way. Since software version 7.2, and now in a Spectrum Virtualize 7.5 ESC, the system continues to maintain the voting set with a dynamic quorum as it did for previous versions. But to provide greater resiliency in the case of 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, through the use of the **overridequorum** command, is also enough to establish a quorum in this case.

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

Consider these examples:

- ▶ Consider an example where 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 of these sites 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 just one node.

- ▶ As a further example, if there is an eight-node cluster with one node down for service, and a failure causes connectivity loss to the quorum disk and some nodes, five nodes are necessary to continue cluster operation.

Figure 3-1 summarizes the behavior of the cluster as a result of failures that affected the site or failure domains.

Failure Domain 1 Node 1	Failure Domain 2 Node 2	Failure Domain 3 Quorum disk or IP Quorum	Cluster Status
Operational	Operational	Operational	Operational, optimal
Failed	Operational	Operational	Operational, Write cache disabled
Operational	Failed	Operational	Operational, Write cache disabled
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	The node that accesses the active quorum disk first remains active and the partner node goes offline. If this is the beginning of a rolling disaster and the node who win the Quorum race goes offline too, then the surviving site can be restored with <code>overridequorum</code> command.
Operational	Failed	Failed	Stopped, then the surviving site can be restored with <code>overridequorum</code> command.
Failed	Operational	Failed	Stopped, then the surviving site can be restored with <code>overridequorum</code> command.

Figure 3-1 Stretched cluster behavior

3.3.6 Quorum disk requirements

The storage controller that provides the quorum disk in an ESC configuration in the third site must be supported as an *extended quorum disk*. Storage controllers that provide extended quorum support are listed on the 2145 Support Portal web page:

<https://ibm.biz/Bdsvxb>

Quorum disk storage controllers must be Fibre Channel or FCIP-attached. They must be able to provide less than 80 ms response times, with a guaranteed bandwidth of greater than 2 MB.

Important: These are the quorum disk candidate requirements for a stretched cluster configuration:

- ▶ The ESC 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 done manually by using the `chquorum` command.

3.3.7 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.

3.3.8 Failure scenarios in an Enhanced Stretched Cluster configuration

Figure 3-2 illustrates several failure scenarios in a split I/O group cluster. The blue lines represent local I/O traffic and the green lines represent I/O traffic between failure domains.

- ▶ Power off FC Switch SAN768B-A1 in Failure Domain 1: FC Switch SAN768B-A2 takes over the load and routes I/O to Spectrum Virtualize Node 1 and Spectrum Virtualize Node 2.

- ▶ Power off Spectrum Virtualize Node 1 in Failure Domain 1: Spectrum Virtualize Node 2 takes over the load and continues processing host I/O. Spectrum Virtualize Node 2 changes the cache mode to write-through to avoid data loss in case Spectrum Virtualize Node 2 also fails.
- ▶ Power off Storage System V7000-A: Spectrum Virtualize waits a short time (15 – 30 seconds), pauses volume copies on Storage System V7000-A, and then continues I/O operations by using the remaining volume copies on Storage System V7000-B.
- ▶ Power off Failure Domain 1: I/O operations can continue from Failure Domain 2.

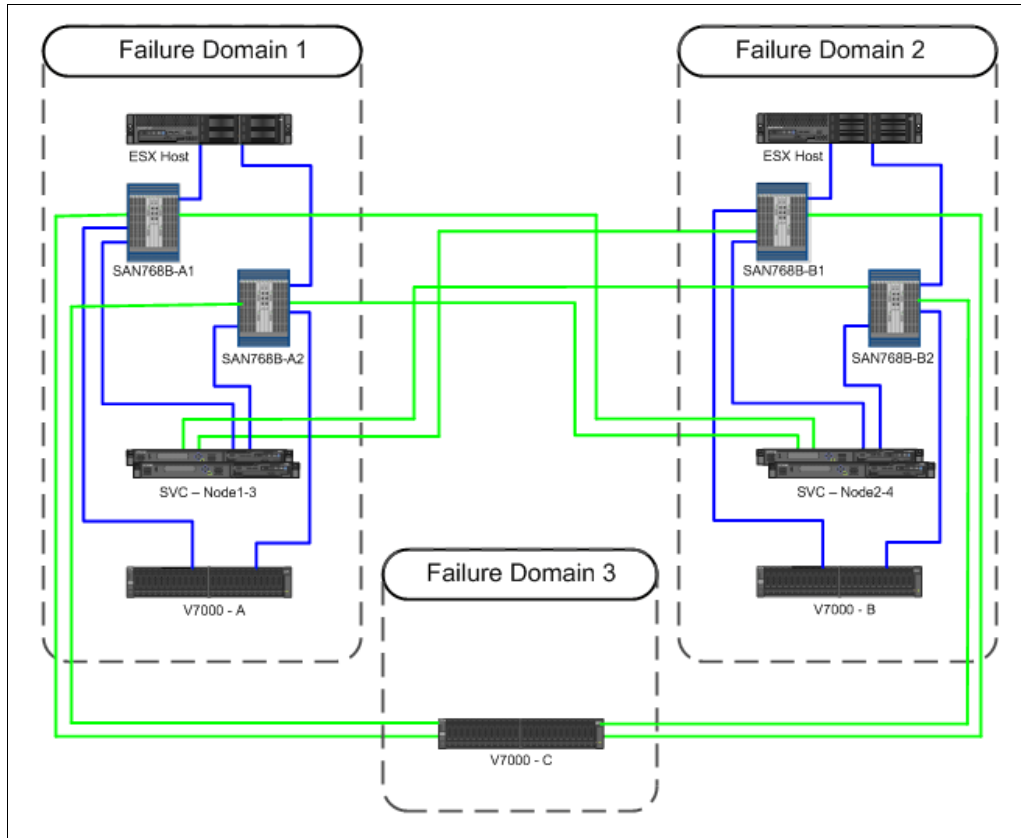


Figure 3-2 Spectrum Virtualize stretched cluster configuration

As Table 3-1 shows, Spectrum Virtualize can handle every kind of single failure automatically without affecting applications.

Table 3-1 Failure scenarios

Failure scenario	Spectrum Virtualize stretched cluster behavior	Server and application impact
Single switch failure.	System continues to operate by using an alternate path in the same failure domain to the same node.	None.
Single data storage failure.	System continues to operate by using the secondary data copy.	None.

Failure scenario	Spectrum Virtualize stretched cluster behavior	Server and application impact
Single quorum storage failure, or IP Quorum failure.	System continues to operate on the same data copy.	None.
Failure of either Failure Domain 1 or 2 (which contain the Spectrum Virtualize nodes).	System continues to operate on the remaining failure domain that contains Spectrum Virtualize nodes.	Servers without HA functions in the failed site stop. Servers in the other site continue to operate. 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. Spectrum Virtualize cache is disabled, which might degrade performance.
Failure of Failure Domain 3 (which contains the active quorum disk).	System continues to operate on both Failure Domains 1 and 2. Spectrum Virtualize selects another quorum disk.	None.
Access loss between Failure Domains 1 and 2 (which contains the Spectrum Virtualize nodes).	System continues to operate the failure domain with Spectrum Virtualize node, which wins the quorum race. The cluster continues with operation, while the node in the other failure domain stops.	Servers without HA functions in the failed site stop. Servers in the other site continue to operate. 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. The Spectrum Virtualize cache is disabled, which might degrade performance.
Access loss between Failure Domains 1 and 2 (which contain the Spectrum Virtualize nodes) because of a rolling disaster. One site is down, and the other is still working. Later, the working site also goes down because of the rolling disaster.	System continues to operate the failure domain with Spectrum Virtualize node, which wins the quorum race. The cluster continues with operation, while the node in the other failure domain stops. Later, the "winning" Spectrum Virtualize Spectrum Virtualize is down, too, because of the rolling disaster. All Spectrum Virtualizes are down.	The system can restart in the surviving site by using the Spectrum Virtualize 7.2 DR feature manually triggered by the new overridequorum command. 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. The Spectrum Virtualize cache is disabled. Some recovery actions must take place to restore the failed site.

3.4 Spectrum Virtualize enhanced stretched cluster configurations

The Spectrum Virtualize nodes of an enhanced stretched cluster configuration must be connected to each other by Fibre Channel or Fibre Channel over IP (FCIP) links. These links provide paths for node-to-node communication and for host access to controller nodes. Enhanced stretched cluster supports three different approaches for node-to-node intracluster communication between failure domains:

- ▶ Attach each Spectrum Virtualize node to the Fibre Channel switches directly in the local and the remote failure domain. Therefore, all node-to-node traffic can be done without traversing ISLs. This approach is referred to as enhanced stretched cluster *No ISL configuration*.

- ▶ Attach each Spectrum Virtualize node only to local Fibre Channel switches and configure ISLs between failure domains for node-to-node traffic. This approach is referred to as enhanced stretched cluster *ISL configuration*.
- ▶ Attach each Spectrum Virtualize node only to local Fibre Channel switches and configure FCIP between failure domains for node-to-node traffic. Support for FCIP was introduced in SAN Volume Controller Version 6.4. This approach is referred to as enhanced stretched cluster *FCIP configuration*.

Each of these enhanced stretched cluster configurations, along with their associated attributes, is described in the sections that follow 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. This limitation is to ensure that any burst in I/O traffic that can occur does not use all of the buffer-to-buffer credits. The link speed is also limited by the cable length between nodes. Table 3-2 lists the supported distances for each of the Spectrum Virtualize enhanced stretched cluster configurations, along with their associated versions and port speed requirements.

Table 3-2 Supported distances

Configuration	Spectrum virtualize version	Maximum length	Maximum link speed
No ISL	5.1 or later	< 10 km	8 Gbps
No ISL	6.3 or later	< 20 km	4 Gbps
No ISL	6.3 or later	< 40 km	2 Gbps
ISL	6.3 or later	< 300 km	2, 4, 8, 16 Gbps
FCIP	6.4 or later	< 300 km	2, 4, 8, 16 Gbps

3.4.1 No ISL configuration

This configuration is similar to a standard Spectrum Virtualize environment. The main difference is that nodes are distributed across two failure domains. Figure 3-3 on page 40 illustrates the No ISL configuration. Failure Domain 1 and Failure Domain 2 contain the Spectrum Virtualize nodes, along with customer data. Failure Domain 3 contains the storage subsystem that provides the active quorum disk.

Advantages

The No ISL configuration has these advantages:

- ▶ The HA solution is distributed across two independent data centers.
- ▶ The configuration is similar to a standard Spectrum Virtualize cluster.
- ▶ Hardware effort is limited. Wavelength Division Multiplexing (WDM) devices can be used but are not required.

Requirements

The No ISL configuration has these requirements:

- ▶ Four dedicated fiber links per I/O group between failure domains.

- ▶ ISLs are not used between SAN Volume Controller nodes.
- ▶ Passive WDM devices can be used between failure domains (up to SAN Volume Controller version 6.2).
- ▶ Active or passive WDM can be used between failure domains (Spectrum Virtualize version 6.3 and later).
- ▶ Long wave small form-factor pluggables (SFPs) are required to reach 10 km without WDM.
- ▶ The supported distance is up to 40 km 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.
- ▶ Quorum disk storage system must be attached to the Fibre Channel.

Figure 3-3 illustrates the Spectrum Virtualize enhanced stretched cluster No ISL configuration. Failure Domain 1 and Failure Domain 2 contain the Spectrum Virtualize nodes, along with customer data. Failure Domain 3 contains the storage subsystem that provides the active quorum disk.

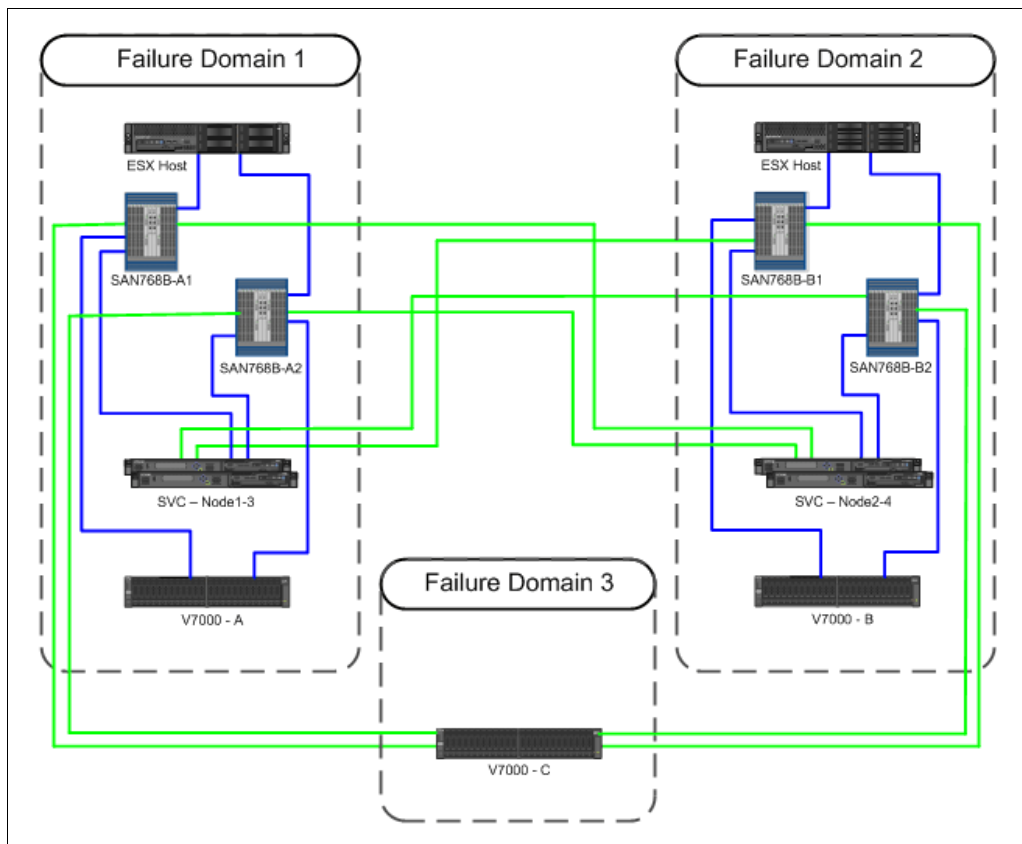


Figure 3-3 Spectrum Virtualize enhanced stretched cluster: No ISL configuration

Zoning requirements

Zoning requirements for the Spectrum Virtualize enhanced stretched cluster No ISL configuration are the same as with a standard configuration;

- ▶ Servers access only Spectrum Virtualize nodes. There is no direct access 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 SAN Volume Controller node FC ports. Port masking is enabled by issuing the `chsystem -localfcportmask` command.
- ▶ Spectrum Virtualize nodes of the same I/O group do not communicate by using ISLs.
- ▶ Zones should not contain multiple back-end disk systems.

Figure 3-4 illustrates the Spectrum Virtualize enhanced stretched cluster No ISL configuration with passive WDM connections between Failure Domains 1 and 2.

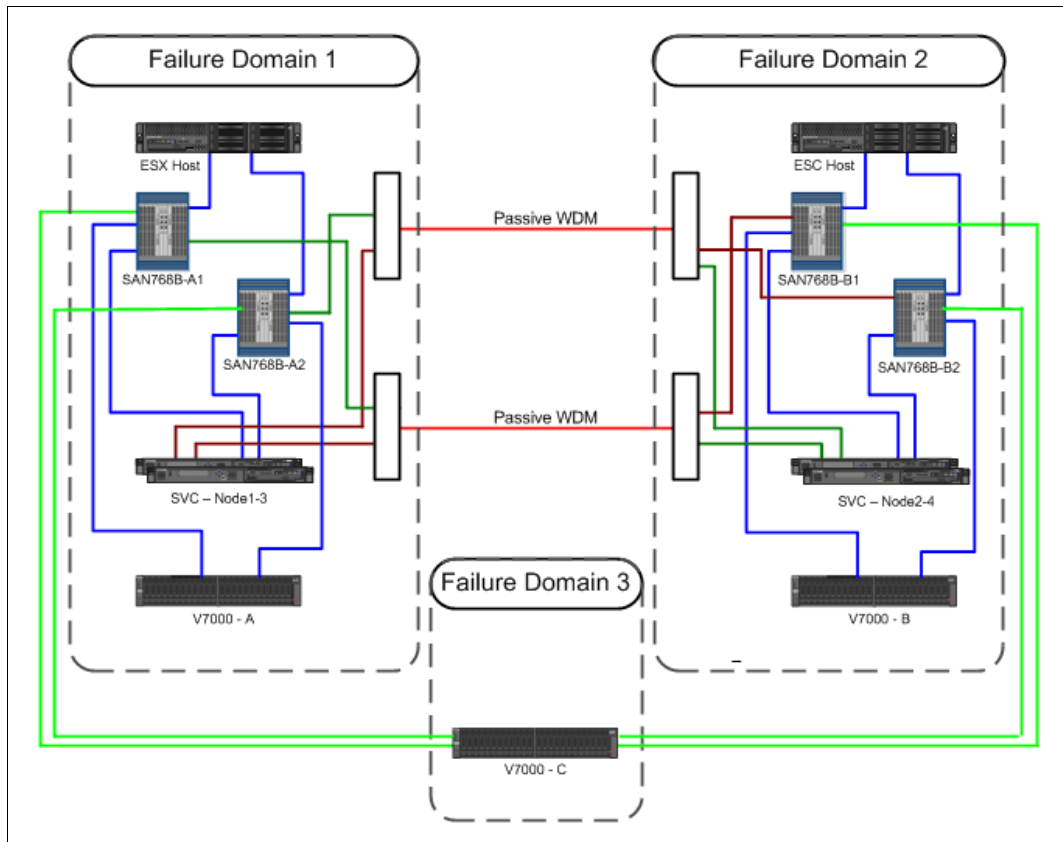


Figure 3-4 Spectrum Virtualize Enhanced Stretched Cluster: No ISL configuration (with WDM)

Best practices for Spectrum Virtualize Fibre Channel ports

Spectrum Virtualize 2145-CG8 Nodes can have AHA7 additional 4x8 Gbps Fibre Channel HBA feature installed and new Spectrum Virtualize 2145-DH8 Node can have up to 12x8 Gbps Fibre Channel ports plus 2x16 Gbps Fibre Channel ports.

The following are the recommended practices:

- ▶ Dedicate two FC ports for node-to-remote site node communication. Attach these ports directly to the switch at the remote site.
- ▶ Dedicate two FC ports for Metro Mirror or Global Mirror if you are going to implement more than a Spectrum Virtualize enhanced stretched cluster that is extended to a DR solution with a remote Spectrum Virtualize cluster.

- ▶ 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 should be accessed using the same ports dedicated for the storage workload.
- ▶ On Spectrum Virtualize Version 7.1.0 and later, configure the **partnerportmask** to the two ports that are dedicated to Metro or Global Mirror, and configure the **localfcportmask** to the port 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.
- ▶ For Spectrum Virtualize Version 6.4.1, interim fixes use switch zoning to restrict the use of each port.
- ▶ Starting with Version 7.2.0, you can assign sites to nodes and controllers, and enable enhanced stretch cluster for improved DR capabilities and improved traffic routing.

3.4.2 ISL configuration

This configuration is similar to a standard Spectrum Virtualize configuration. The differences are that the nodes are distributed across two failure domains, and node-to-node communication between failure domains is performed over ISLs. Spectrum Virtualize support for ISLs was introduced in Version 6.3.

The use of ISLs increases the supported distance for Spectrum Virtualize enhanced stretched cluster configurations to 300 km. Although the maximum supported distance is 300 km, there are instances where host-dependent I/Os must traverse the long-distance links multiple times. Because of this, the associated performance degradation might exceed acceptable levels. To mitigate this possibility, limit the distance between failure domains to 150 km.

Guideline: Limiting the distance between failure domains to 150 km minimizes the risk of encountering elevated response times.

Advantages

The ISL configuration has these advantages:

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

Requirements

The ISL configuration has these requirements:

- ▶ Requires four dedicated FC ports, with two ports for each node for each I/O group between failure domains.
- ▶ Using ISLs for node-to-node communication requires configuring two separate SANs:
 - One SAN is dedicated for Spectrum Virtualize node-to-node communication. This SAN is referred to as the *private* SAN.
 - One SAN is dedicated for host and storage controller attachment. This 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 correctly sized. For private SANs, the bandwidth must be able to sustain the data rate that is generated by all write operations because of cache

synchronization. For public SANs, the bandwidth must be able to sustain all of the read operations that are redirected to the local site controller, in case there is a controller or node failure or an application is moving between the sites, before reconfiguring the preferred node properties for those specific volumes. In addition, they must be able to sustain the other ISL traffic for this SAN that also uses the ISL. It is suggested to contact you Spectrum Virtualize IBM IT Specialist to size the correct bandwidth.

- ▶ A third failure domain is required for quorum disk placement.
- ▶ Storage controllers that contain quorum disks must be attached to the Fibre Channel.
- ▶ 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 Fibre Channel switches for each SAN
- ▶ Switch partitioning features
- ▶ Virtual or logical fabrics

Figure 3-5 on page 44 illustrates the ISL configuration. Failure Domain 1 and Failure Domain 2 contain the Spectrum Virtualize nodes, along with customer data. Failure Domain 3 contains the storage subsystem that provides the active quorum disk.

Zoning requirements

The Spectrum Virtualize enhanced stretched cluster ISL configuration requires private and public SANs. The two SANs must be configured according to the following rules:

- ▶ Two ports of each Spectrum Virtualize node are attached to the private SANs.
- ▶ Other ports are configured to the public SAN in accordance with the configuration implemented.
- ▶ A single trunk between switches is required for the private SAN.
- ▶ Hosts and storage systems are attached to fabrics of the public SANs.
- ▶ Links that are used for Spectrum Virtualize Metro Mirror or Global Mirror must be attached to the public SANs or to a dedicated geographical SAN by using dedicated SAN Volume Controller node FC ports. If you are still using CG8 nodes, provide each node with the additional AHA7 4x8 Gbps Fibre Channel host bus adapter (HBA) port feature.
- ▶ Port masking can be used in addition to zoning to segregate node-to-node traffic on specific Spectrum Virtualize node 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 must not be over-subscribed.

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

<https://ibm.biz/Bdsvxb>

Figure 3-5 illustrates the Spectrum Virtualize enhanced stretched cluster ISL configuration. The private and public SANs are represented as logical switches on each of the four physical switches.

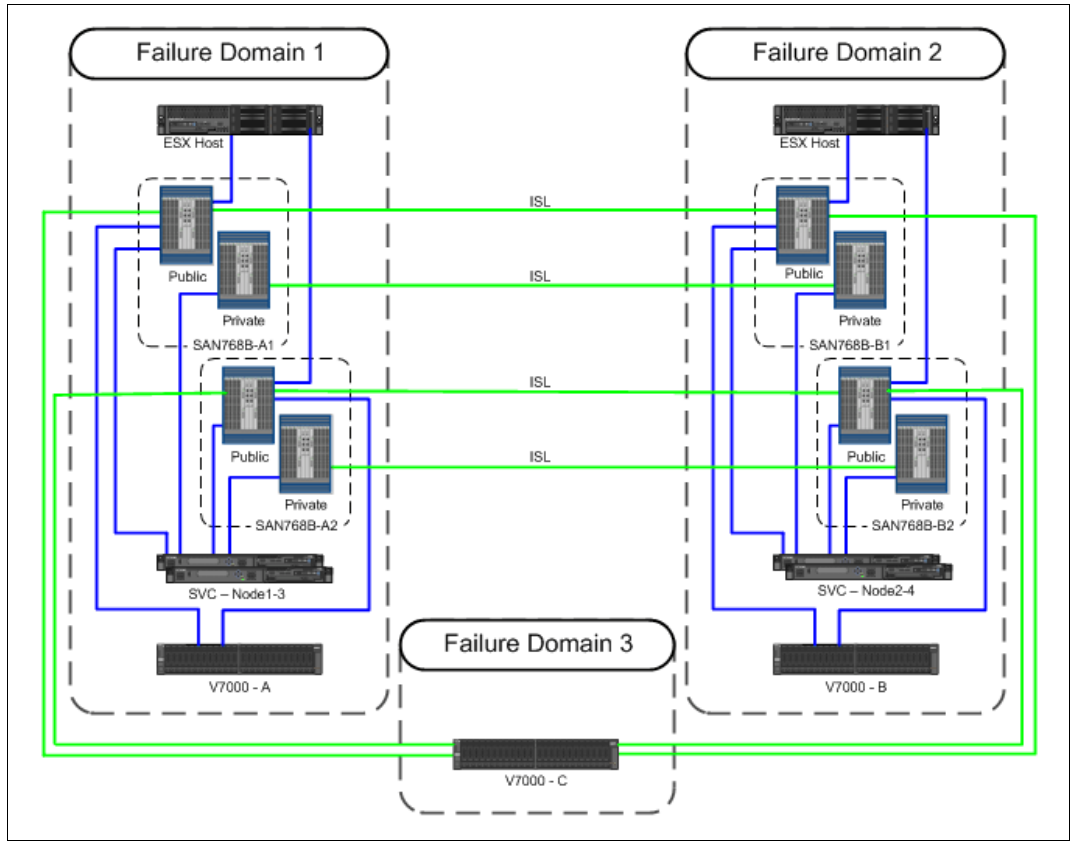


Figure 3-5 Spectrum Virtualize enhanced stretched cluster: ISL configuration

Figure 3-6 illustrates the Spectrum Virtualize enhanced stretched cluster *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.

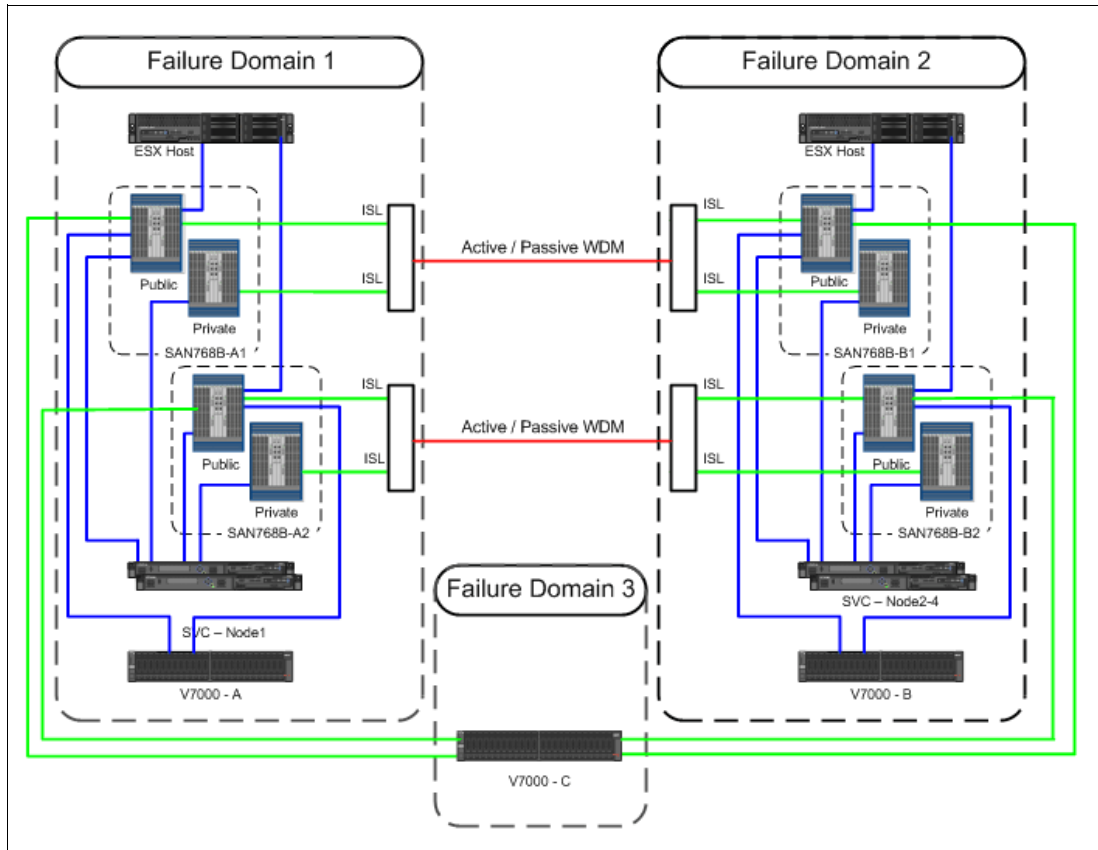


Figure 3-6 Spectrum Virtualize enhanced stretched cluster: ISL configuration (with WDM)

Best practices for Spectrum Virtualize Fibre Channel ports

Spectrum Virtualize 2145-CG8 Nodes can have AHA7 additional 4x8 Gbps Fibre Channel HBA feature installed, and new Spectrum Virtualize 2145-DH8 Nodes can have up to 12x8 Gbps Fibre Channel ports plus 2x16 Gbps Fibre Channel ports.

The following are the best practices:

- ▶ Dedicate two FC ports for node-to-remote site node communication. Attach these directly to the switch at the local site.
- ▶ Dedicate two FC ports for Metro Mirror or Global Mirror if you are going to implement more than a Spectrum Virtualize enhanced stretched cluster that is extended to a DR solution with a remote Spectrum Virtualize cluster.
- ▶ Use the other FC ports for host and storage attachment to separate the storage workload from that host. Attach these to the switch at the local site.
- ▶ The third site quorum device should be accessed using the same ports dedicated for the storage workload.
- ▶ On Spectrum Virtualize Version 7.1.0 and later, configure **partnerportmask** to the two ports that are dedicated to Metro or Global Mirror, and configure **localfcportmask** to the port 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.
- ▶ For Spectrum Virtualize Version 6.4.1, interim fixes use switch zoning to restrict the use of each port.

- ▶ Starting with Version 7.2.0, you can assign sites to nodes and controllers and enable enhanced stretch cluster for improved DR capabilities and improved traffic routing.

3.4.3 FCIP configuration

In this configuration, FCIP links are used between failure domains. Spectrum Virtualize support for FCIP was introduced in Version 6.4. This configuration is a variation of the ISL configuration described previously, so many of the same requirements apply.

Advantages

FCIP configuration has this advantage:

- ▶ Uses existing IP networks for extended distance connectivity.

Requirements

FCIP configuration has these requirements:

- ▶ Requires at least two FCIP tunnels between failure domains.
- ▶ Using ISLs for node-to-node communication requires configuring two separate SANs:
 - One SAN is dedicated for Spectrum Virtualize node-to-node communication. This SAN is referred 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 correctly sized:
 - For private SANs, the bandwidth must be able to sustain the data rate that is generated by all the write operations because of cache synchronization.
 - For public SANs, the bandwidth must be able to sustain all of the read operations that are redirected to the local site controller in case of controller or node failure, or an application moving between the sites before reconfiguring the preferred node properties for those specific volumes. In addition, the public SAN must be able to sustain the other ISL traffic for this SAN that also uses the ISL.
- ▶ A third failure domain is required for quorum disk placement.
- ▶ Failure Domain 3 (quorum disk) must be either Fibre Channel 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 Fibre Channel or attached to FCIP.
- ▶ A 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.

Zoning requirements

The Spectrum Virtualize enhanced stretched cluster FCIP configuration requires private and public SANs. The two SANs must be configured according to the following rules:

- ▶ Two ports of each Spectrum Virtualize node are attached to the private SANs.
- ▶ Other ports are configured to the public SAN in accordance with the configuration implemented.
- ▶ A single trunk between switches is required for the private SAN.
- ▶ Hosts and storage systems are attached to fabrics of the public SANs.

- ▶ Port masking can be used in addition to zoning to segregate node-to-node traffic on specific Spectrum Virtualize node FC ports. Port masking is enabled with the `chsystem -localfcportmask` command.
- ▶ Links that are used for Spectrum Virtualize Metro Mirror or Global Mirror must be attached to the public SANs or to a dedicated geographical SAN by using dedicated FC ports. If you are still using 2145-CG8 nodes, provide each node with the additional AHA7 4x8 Gbps Fibre Channel HBA port feature.
- ▶ Failure Domain 3 (quorum disk) must be attached to the public SAN.
- ▶ ISLs that belong to the private SANs must not be shared with other traffic and must not be over-subscribed.

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

<https://ibm.biz/Bdsvxb>

Figure 3-7 illustrates the Spectrum Virtualize enhanced stretched cluster FCIP configuration.

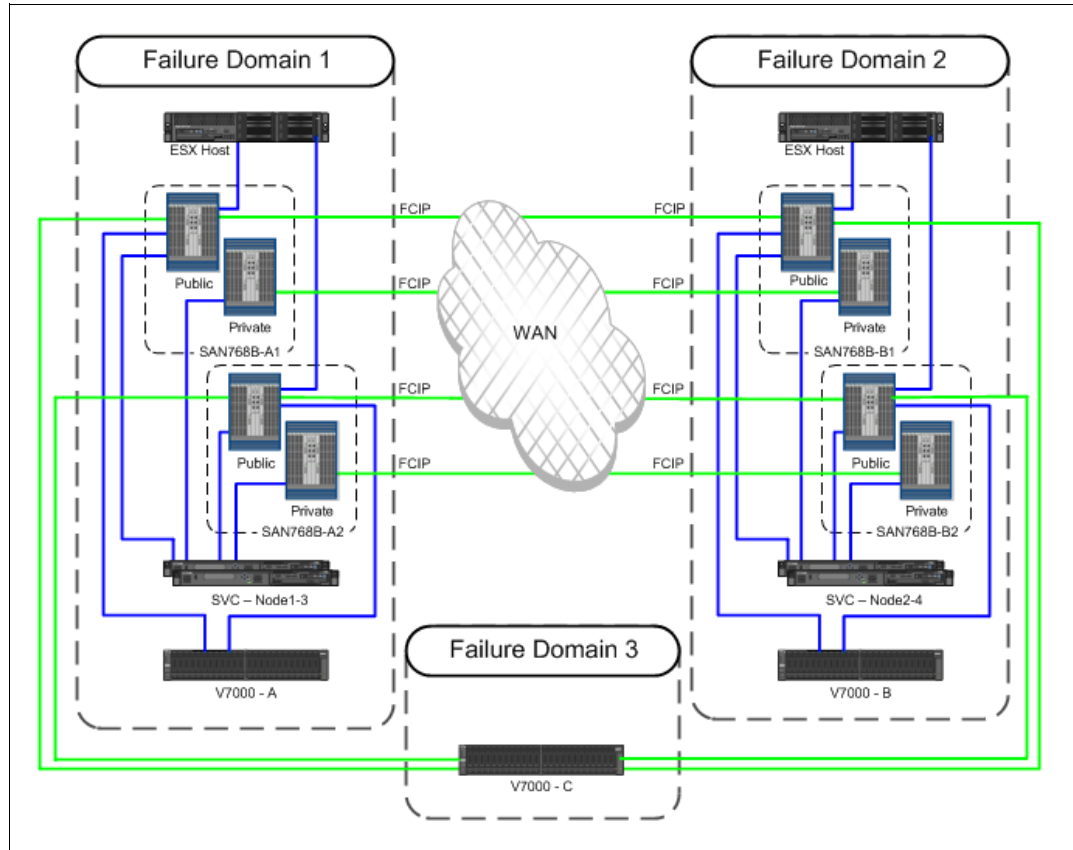


Figure 3-7 FCIP configuration

Best practices for Spectrum Virtualize Fibre Channel ports

Spectrum Virtualize 2145-CG8 Nodes can have AHA7 additional 4x8 Gbps Fibre Channel HBA feature installed, and new Spectrum Virtualize 2145-DH8 Nodes can have up to 12x8 Gbps Fibre Channel ports plus 2x16 Gbps Fibre Channel ports.

The following procedures are the current leading practices:

- ▶ Dedicate two FC ports for node-to-remote site node communication. Attach these directly to the switch at the local site.

- ▶ Dedicate two FC ports for Metro Mirror or Global Mirror if you are going to implement more than a SAN Volume Controller enhanced stretched cluster that is extended to a DR solution with a remote SAN Volume Controller cluster.
- ▶ Use the other FC ports for host and storage attachment to separate the storage workload from that host. Attach these to the switch at the local site.
- ▶ The third site quorum device should be accessed using the same ports that are dedicated for the storage workload.
- ▶ On Spectrum Virtualize Version 7.1.0 and later, configure `partnerportmask` to the two ports that are dedicated to Metro or Global Mirror, and configure `localfcportmask` to the port 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 for this purpose.
- ▶ For Spectrum Virtualize Version 6.4.1, interim fixes use switch zoning to restrict the use of each port.
- ▶ Starting with Version 7.2.0, you can assign sites to nodes and controllers and enable enhanced stretch cluster for improved DR capabilities and improved traffic routing.

3.5 Fibre Channel settings for distance

Usage of Longwave (LW) SFPs is an appropriate method to overcome long distances. Starting with Spectrum Virtualize 6.3, active and passive dense wavelength division multiplexing (DWDM) and coarse wavelength division multiplexing (CWDM) technologies 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 Spectrum Virtualize stretched cluster configurations.

Buffer credits, also called *buffer-to-buffer* (BB) credits, are used for Fibre Channel 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.

Thus buffer-to-buffer credits are necessary to have multiple Fibre Channel frames in flight (Figure 3-8). 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.

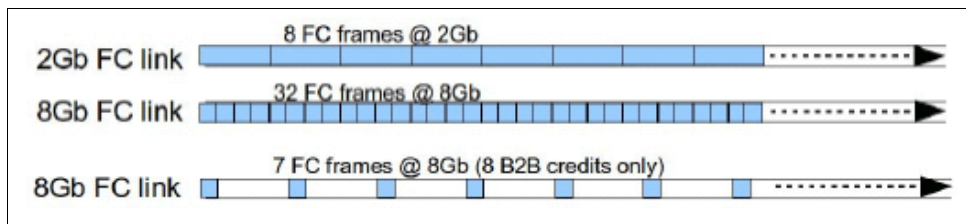


Figure 3-8 Buffer credits in flight

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 is not true. The guidelines that follow give the minimum numbers. The performance drops dramatically if there are not enough buffer credits 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	Distance with eight credits
1 Gbit/s	5	16 km
2 Gbit/s	10	8 km
4 Gbit/s	20	4 km
8 Gbit/s	40	2 km
16 Gbit/s	80	1 km

The number of buffer-to-buffer credits that are provided by a Spectrum Virtualize Fibre Channel HBA is limited. An HBA of a model 2145-CG8 node provides 41 buffer credits, which are sufficient for a 10 km distance at 8 Gbit/s. The HBAs in all earlier Spectrum Virtualize models provide only eight buffer credits. This is enough for only 4 km with a 4 Gbit/s link speed. These numbers are determined by the HBA's hardware and cannot be changed.

The new 2145-DH8 none still supply 41 buffer credits for 8 Gbit/s FC port and 80 buffer credits for 16 Gbit/s FC port. At the time of writing, 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 Brocade switches: Eight BB credits for each port). Although the Spectrum Virtualize HBAs provide 41 BB credits, the switches stay at the default value. Therefore, you must adjust the switch BB credits manually. For IBM b-type/Brocade switches, the port buffer credits can be changed by using the `portcfgportbuffers` command.

3.6 Spectrum Virtualize I/O operations on mirrored volumes

There are two architectural behaviors with the Spectrum Virtualize that must be considered for mirrored volumes in an Enhanced Stretched Cluster configuration:

- ▶ Preferred node
- ▶ Primary copy

3.6.1 Preferred node

Before version 7.5, the general recommendation was to have the volume assignments based on the local to local policy. This policy means that if a host is in Power Domain 1, Site 1, the preferred node must be in Power Domain 1, Site 1 as well. With the introduction of the host site awareness in version 7.5, the host I/Os are normally performed by a node with the same site definition as the host. The preferred node definition is no longer taken in account. For this reason, no particular recommendation must be followed for the preferred node definition, other than distributing them evenly across all the nodes in the cluster.

Additionally, for hosts with multipath drivers that support a Spectrum Virtualize, the host directs all read/write activity to the node with the same site affinity.

3.6.2 Primary copy

Before version 7.2, the primary copy of the volume mirror was used for all read operations. When volumes are initially created, the first copy that is created becomes the primary. The primary copy can be changed after initial volume creation without interrupting I/O processing. Starting with Spectrum Virtualize 7.2, the concept of *primary copy* remains. However, it is no longer used by read operations because Spectrum Virtualize read operations from its back-end storage are routed from the local site controller with node and controller site properties. The *secondary copy* is read from only if the primary copy, the one with same site properties, is not available. This process is independent of which node the host is reading from. Because of this behavior, place the primary copy of the mirror in the same failure domain as the host that is accessing the mirrored volume. Set the host to access the mirrored volume through the node that is in the same failure domain or site as the host.

Figure 3-9 illustrates the data flow for a read operation. The green line represents an optimal configuration where each host and SAN Volume Controller node reads from its local site copy. The red line is a non-optimal configuration where the controller failure causes the host in Failure Domain 2 to access a volume that is in Failure Domain 1.

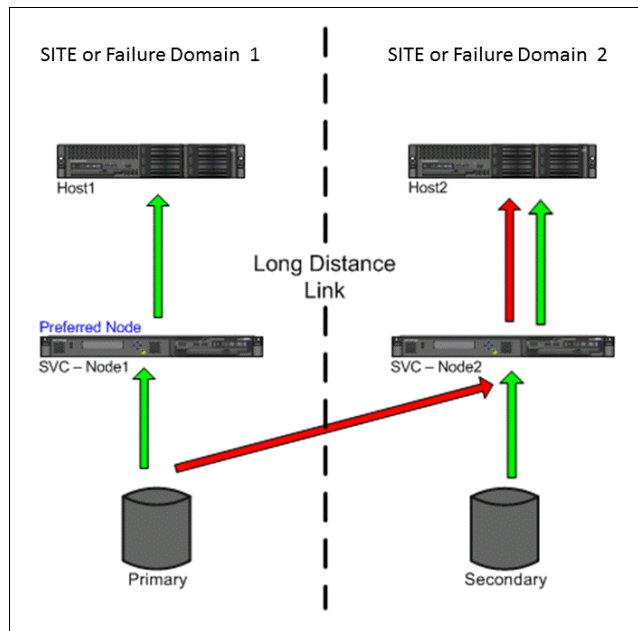


Figure 3-9 Read operation

Write operations

All writes are mirrored between nodes in the I/O groups, so they must traverse the long-distance links. The writes are then destaged to both copies of the volume mirror from each node accordingly, with the **Node** and **Controller** site attributes.

This process is independent of which node the host is writing to. For optimal performance, an extra transfer across the long-distance links can be avoided by ensuring that the host writes are occurring to the node that is in the same site or failure domain as the host.

Figure 3-10 illustrates the data flow for write operations. The green line represents an optimal configuration where writes are occurring on the node that is local to their respective failure domains. The red line represents a non-optimal configuration where writes are occurring on the node that is remote to their respective failure domain. This might happen when the local

Spectrum Virtualize node related to the same site where the host is located is offline after failure or for maintenance.

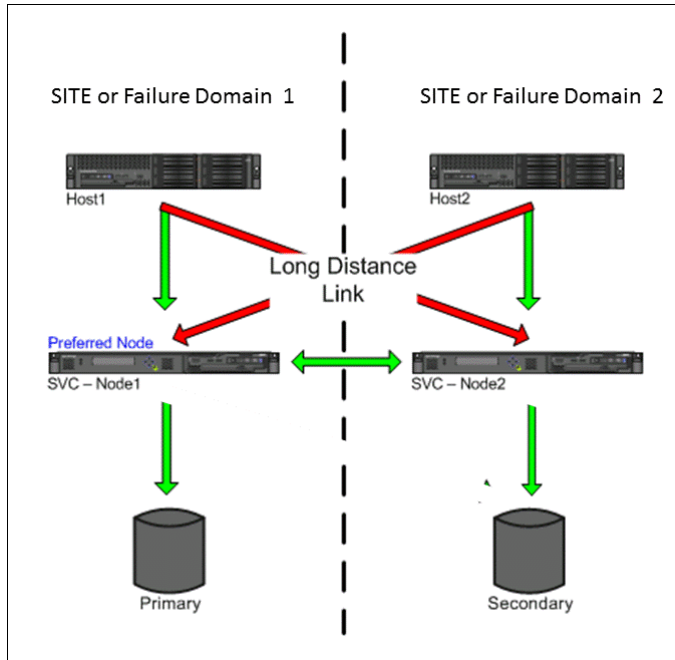


Figure 3-10 Write operation

Guideline: Have the host access the mirrored volume through the Spectrum Virtualize node that is in the same failure domain or site as the host.

3.7 Enhanced stretched cluster three-site DR configuration

Spectrum Virtualize in its enhanced stretched cluster configuration is a high availability solution with a business continuity solution, from a storage point of view.

With Spectrum Virtualize Version 7.2 and the introduction of the DR feature, the enhanced stretched cluster is used. It is a solid and valid solution for data mobility and a DR solution that gives data consistency even during a rolling disaster.

Enhanced stretched cluster is still a two-site solution.

In those scenarios where a three-site solution or even four-site solution is required, the Spectrum Virtualize enhanced stretched cluster configuration can be extended by adding Metro Mirror or Global Mirror to create a partnership with a remote Spectrum Virtualize cluster up to 80 ms round-trip latency distance or 250 ms round-trip latency distance. The distance depends on which connectivity protocol and speed are used (IP 1 Gb or 10 Gb, or Fibre Channel) to create the cluster partnership.

The remote Spectrum Virtualize cluster can also be an enhanced stretched cluster configuration. In that case, you automatically have a four-site HA, business continuity, and DR solution.

Figure 3-11 shows the high-level design of a Spectrum Virtualize enhanced stretched cluster with Metro or Global Mirror.

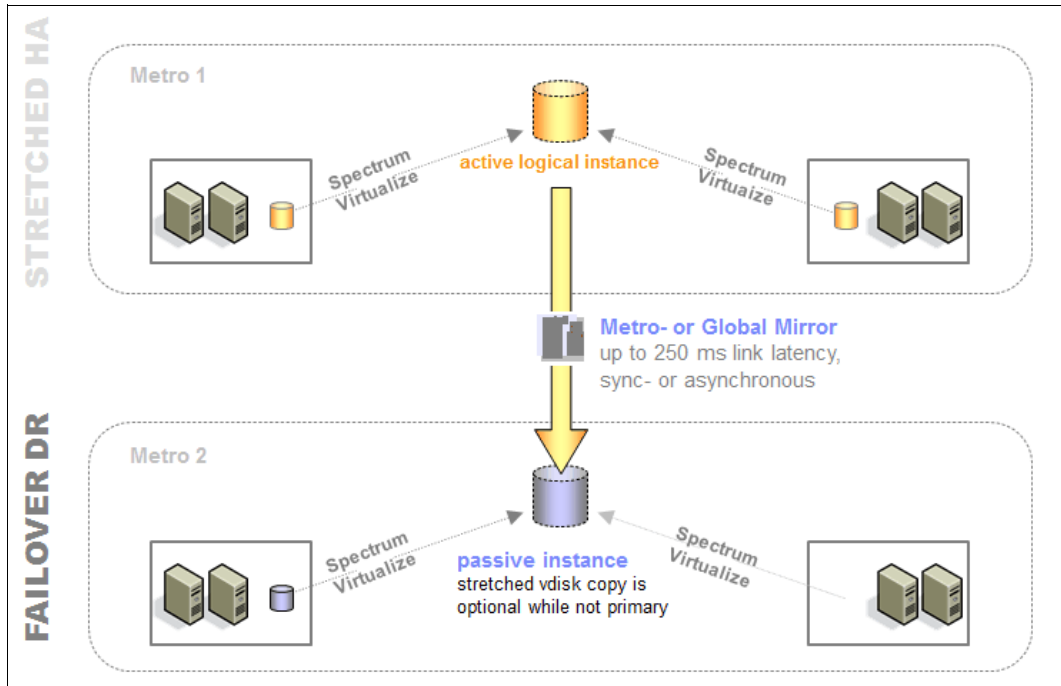


Figure 3-11 Spectrum Virtualize enhanced stretched cluster plus Metro or Global Mirror



Implementation

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

- ▶ Test environment
- ▶ ADX: Application Delivery Controller
- ▶ IP networking configuration
- ▶ IBM Fibre Channel SAN
- ▶ Spectrum Virtualize with an Enhanced Stretched Cluster
- ▶ Volume mirroring
- ▶ Read operations
- ▶ Write operations
- ▶ Quorum disk
- ▶ Quorum disk requirements and placement
- ▶ Automatic quorum disk selection
- ▶ Using the GUI
- ▶ Volume allocation

4.1 Test environment

Figure 4-1 shows the environment that you will implement.

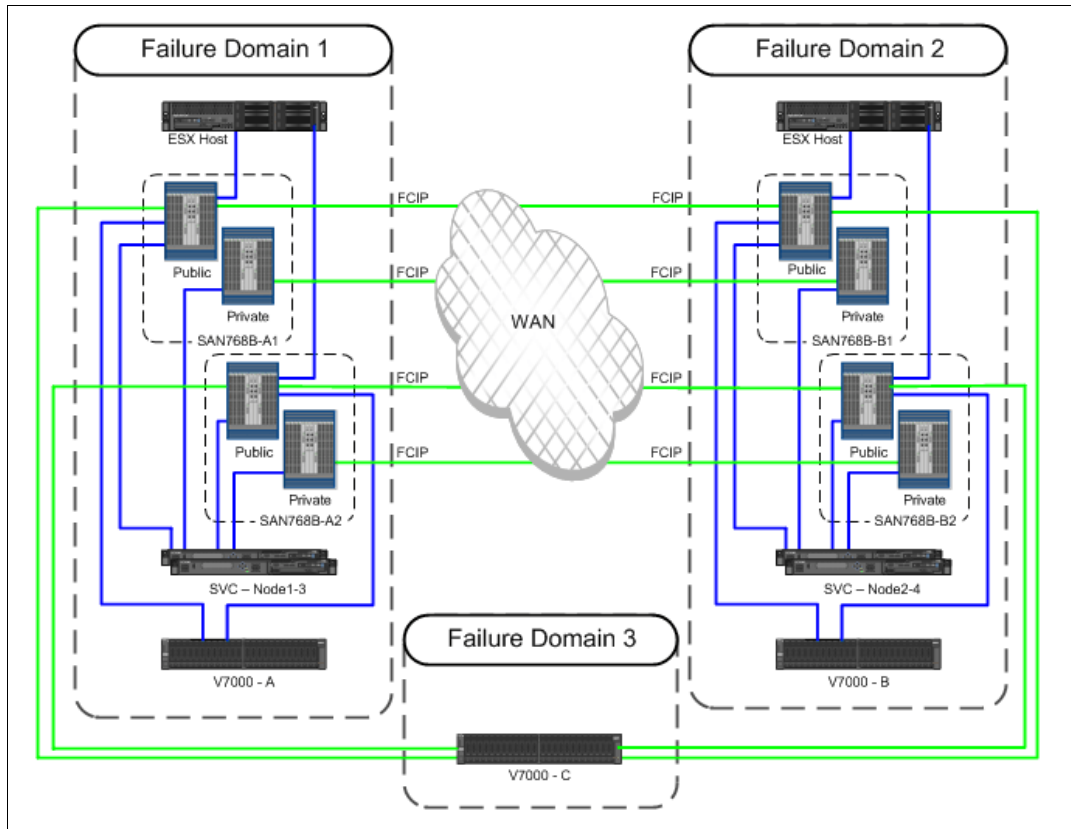


Figure 4-1 IBM SAN and stretched cluster with VMware

This chapter describes the products that you must configure to create this solution. Where an IBM product does not exist, the best available product in the marketplace is selected.

Figure 4-2 shows the example laboratory SAN design from the storage and SAN components point of view.

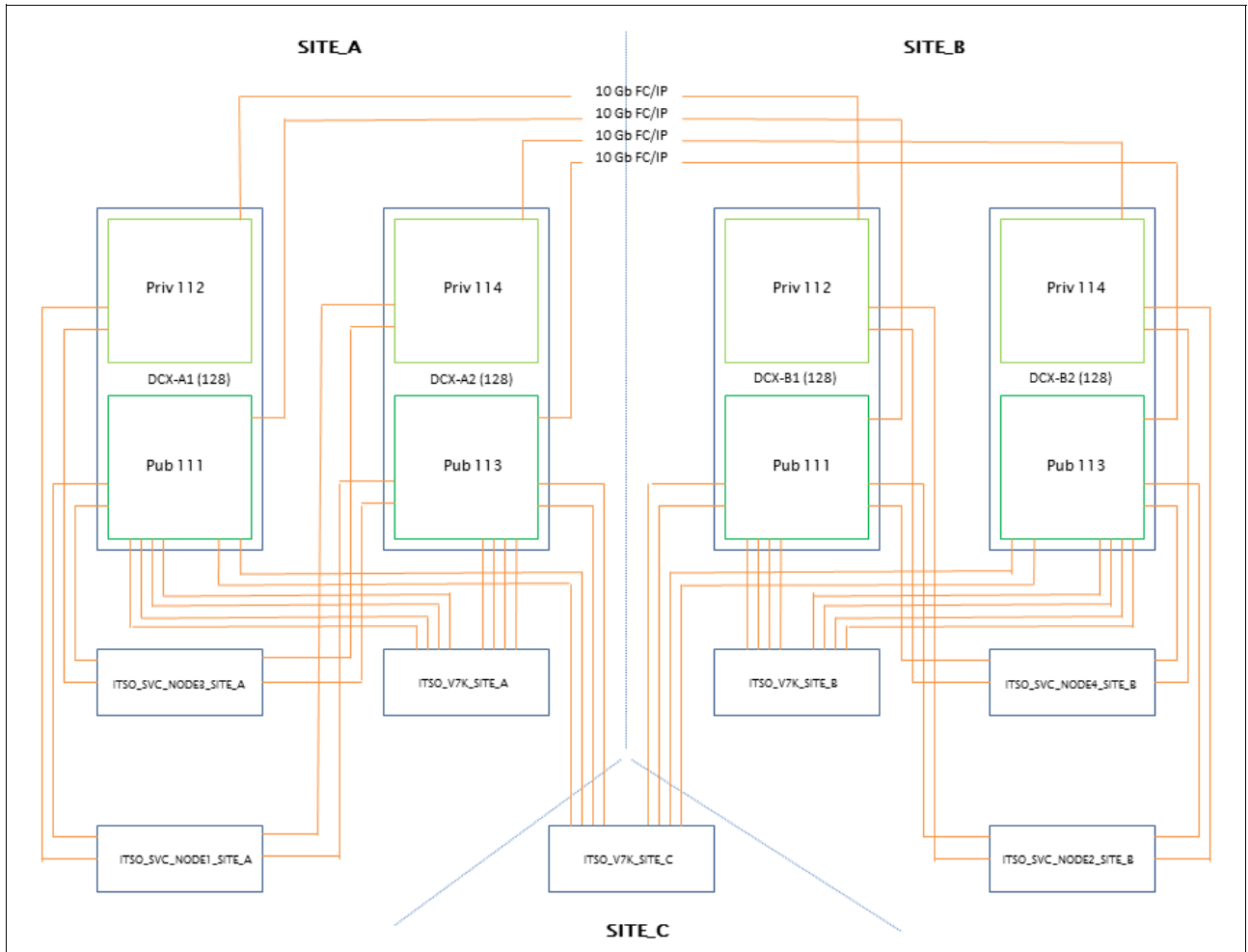


Figure 4-2 SAN design from storage and SAN component point of view

Figure 4-3 shows the example laboratory SAN design from host, SAN component, IBM Spectrum Virtualize, and Spectrum Virtualize points of view.

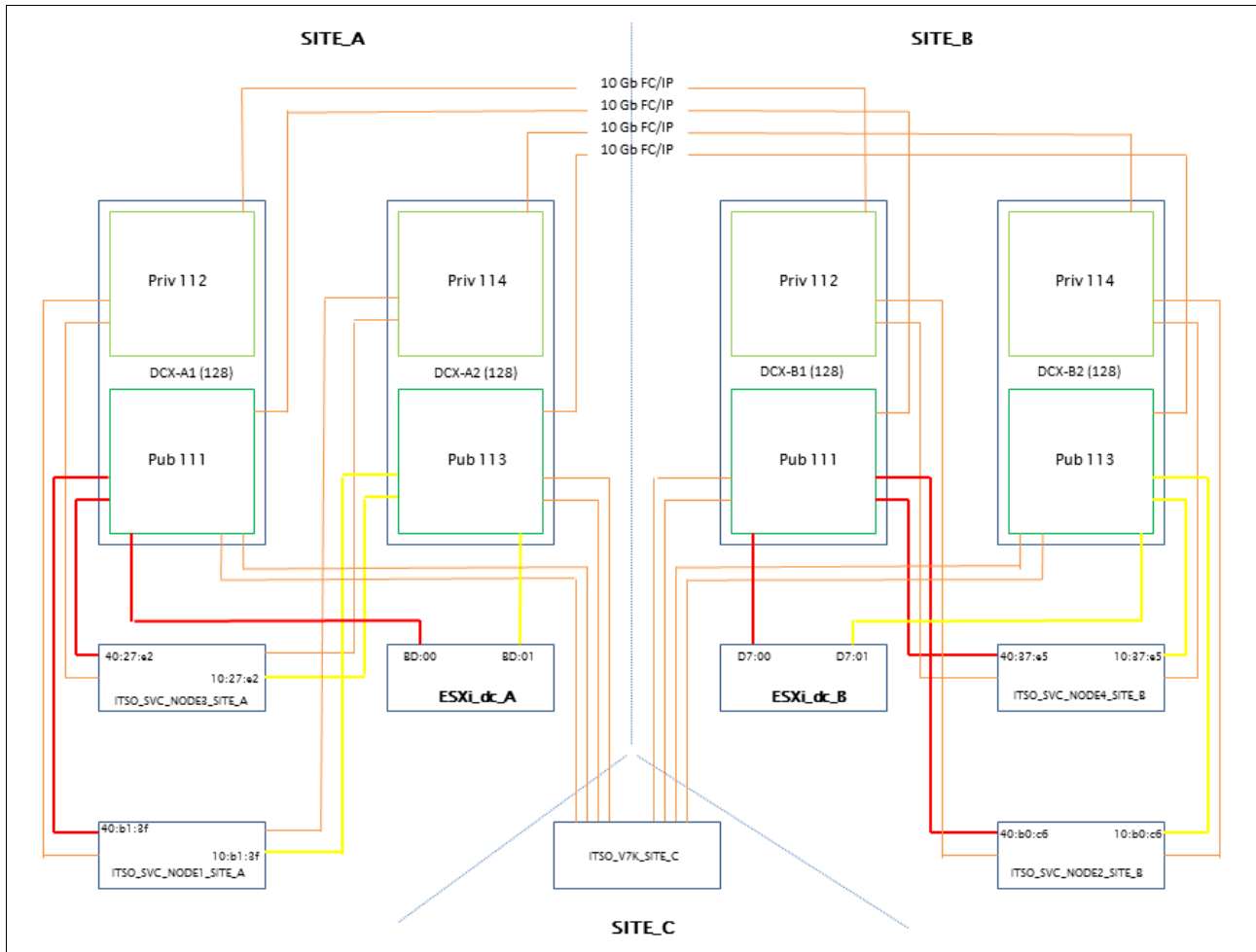


Figure 4-3 SAN design from host, Spectrum Virtualize, and SAN component point of view

4.2 ADX: Application Delivery Controller

The Brocade Application Delivery Controller (ADC) provides virtual machine (VM) aware application delivery. It uses Global Server Load Balancing (GSLB) technology along with network address translation (NAT) and the Application Resource Broker (ARB) plug-in. These products offer seamless access for clients who connect to VMs that migrate between data centers. The configuration example uses Brocade ADX Series Version 12.4.0.

For each VM or group of VMs (also known as *real servers*) in a data center, a virtual IP (VIP) address is created for client access. Each ADX in each data center has a different VIP for the same set of real servers. The setup example has a VM (real server) ITSO_VM_1 with an IP address of 192.168.201.101.

On the ADX in Data Center A, create a VIP of 177.20.0.88. This VIP is associated with ITSO_VM_1.

On the ADX in Data Center B, create a VIP of 178.20.0.88 that you also associate with ITSO_VM_1. For Data Center B ADX configuration, you also disable GSLB recognition of the associated ITSO_VM_1 real server configuration. Therefore, in the GSLB view, the only real server that is online is the one seen in Data Center A.

Figure 4-4 shows what the VIP configuration to real server looks like.

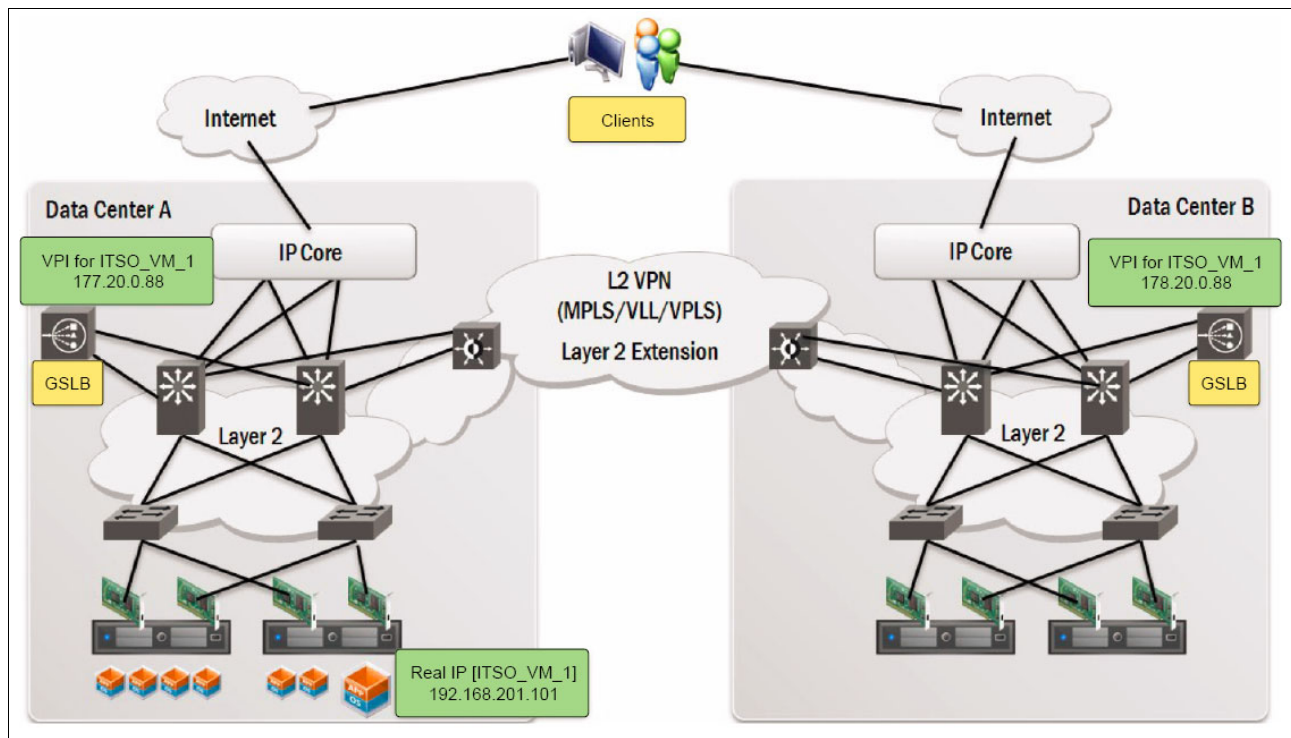


Figure 4-4 VIP and real server configuration for ITSO_VM_1

GSLB enables an ADX to add intelligence to authoritative Domain Name System (DNS) servers by serving as a proxy to the VMs and providing optimal IP addresses to the querying clients. As a DNS proxy, the GSLB ADX evaluates the IP addresses that are in the DNS replies from the authoritative DNS server, for which the ADX is a proxy. It then places the “best” host address for the client at the top of the DNS response.

In this solution, the best host address is the VIP for the data center where the VM is active. For example, if ITSO_VM_1 is active in Data Center A, GSLB directs clients to VIP 177.20.0.88, which used NAT on the back end to connect to ITSO_VM_1.

If ITSO_VM_1 does a vMotion migration to Data Center B, the ARB vCenter plug-in automatically detects the move. It updates the GSLB configuration on the ADXs in both data centers to direct clients to access the VIP in Data Center B, 178.20.0.88.

Consideration: Existing client connections to the real servers (VMs) are not affected with this configuration, and client traffic might traverse the WAN interconnects for some time. There might be other methods, such as route injection, that can be used to redirect client requests more immediately. Also, upstream DNS caching might affect the time before new client requests are directed to the correct data center. To help mitigate this limitation, the Brocade ADX uses a DNS record time to live (TTL) value of 10 seconds.

You must configure the following items in this order:

1. VIP address assignment that outside clients use and the corresponding real server configuration behind the VIP
2. GSLB
3. ARB server installation

4. ADX registration in ARB plug-in
5. VM mobility enablement in the ARB plug-in in vCenter

4.2.1 VIP and real server configuration

For the ADX in Data Center A, create a VIP of 177.20.0.88 for the VM ITSO_VM_1 at a real IP address of 192.168.201.101 (see Example 4-1). In this example, select port 50 to bind for ITSO_VM_1.

Example 4-1 Data Center A - Brocade ADX: VIP and real server configuration

```
telnet@DC1-SLB1-ADX(config)#server real itso_vm_1 192.168.201.101
telnet@DC1-SLB1-ADX(config-rs-itso_vm_1)#source-nat
telnet@DC1-SLB1-ADX(config-rs-itso_vm_1)#port 50
telnet@DC1-SLB1-ADX(config-rs-itso_vm_1)#exit
telnet@DC1-SLB1-ADX(config)#server virtual dca-itso_vm_1 177.20.0.88
telnet@DC1-SLB1-ADX(config-vs-dca-itso_vm_1)#port 50
telnet@DC1-SLB1-ADX(config-vs-dca-itso_vm_1)#bind 50 itso_vm_1 50
```

For the ADX in Data Center B, create a VIP of 178.20.0.88 for the same VM, ITSO_VM_1 (Example 4-2). However, because the ITSO_VM_1 is active in Data Center A, also run the **gslb-disable** command on the real server configuration in Data Center B. This command forces GSLB to see only that the real server (VM) is online in Data Center A and to direct requests to that VIP.

Example 4-2 Data Center B - Brocade ADX: VIP and real server configuration

```
telnet@DC2-SLB1-ADX(config)#server real itso_vm_1 192.168.201.101
telnet@DC2-SLB1-ADX(config-rs-itso_vm_1)#source-nat
telnet@DC2-SLB1-ADX(config-rs-itso_vm_1)#port 50
telnet@DC2-SLB1-ADX(config-rs-itso_vm_1)#gslb-disable
telnet@DC2-SLB1-ADX(config-rs-itso_vm_1)#exit
telnet@DC2-SLB1-ADX(config)#server virtual dcb-itso_vm_1 178.20.0.88
telnet@DC2-SLB1-ADX(config-vs-dcb-itso_vm_1)#port 50
telnet@DC2-SLB1-ADX(config-vs-dcb-itso_vm_1)#bind 50 itso_vm_1 50
```

4.2.2 Global Server Load Balancing (GSLB) configuration

The Brocade ADX can act either as a proxy for local or remote DNS servers or be populated with host information to use for responses. Configure the ADX in Data Center A to respond to DNS requests without needing to query another local or remote DNS server for the `itso.com` domain.

To do so, configure a VIP with the **dns-proxy** command for DNS clients to access. Additionally, define a list of hosts and corresponding IP addresses for the `itso.com` domain, as shown in Example 4-3.

Example 4-3 Configuration of Brocade ADX in Data Center A as a DNS server

```
telnet@DC1-SLB1-ADX(config)#server virtual dns-proxy 177.20.0.250
telnet@DC1-SLB1-ADX(config-dns-proxy)#port dns
telnet@DC1-SLB1-ADX(config-dns-proxy)#exit
telnet@DC1-SLB1-ADX(config)#gslb dns zone itso.com
telnet@DC1-SLB1-ADX(config-gslb-dns-itso.com)#host-info itso_vm_1 50
```

```
telnet@DC1-SLB1-ADX(config-gslb-dns-itso.com)#host-info itso_vm_1 ip-list
177.20.0.88
telnet@DC1-SLB1-ADX(config-gslb-dns-itso.com)#host-info itso_vm_1 ip-list
178.20.0.88
```

Next, configure the ADX in Data Center A for GSLB (Example 4-4).

Example 4-4 ADX in Data Center A: GSLB configuration

```
telnet@DC1-SLB1-ADX(config)#gslb protocol
telnet@DC1-SLB1-ADX(config)#gslb site DataCenterA
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterA)#weight 50
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterA)#si DC1-SLB1-ADX 192.168.1.2
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterA)#gslb site DataCenterB
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterB)#weight 50
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterB)#si DC2-SLB1-ADX 192.168.1.3
```

Configure the ADX in Data Center B for GSLB (Example 4-5).

Example 4-5 Brocade ADX in Data Center B: GSLB configuration

```
telnet@DC2-SLB1-ADX(config)#gslb protocol
telnet@DC2-SLB1-ADX(config)#gslb site DataCenterA
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterA)#weight 50
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterA)#si DC1-SLB1-ADX 192.168.1.2
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterA)#gslb site DataCenterB
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterB)#weight 50
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterB)#si DC2-SLB1-ADX 192.168.1.3
```

Use the **show gslb dns detail** command to view the status of your configuration.

Example 4-6 shows the *itso.com* zone with two VIPs. Although there is one active binding for each VIP, only VIP 177.20.0.88, which corresponds to Data Center A, is active. It is active because the real server definition for *ITSO_VM_1* behind VIP 178.20.0.88 is disabled.

Example 4-6 Checking the GSLB status by using the show gslb dns detail command

```
telnet@DC1-SLB1-ADX(config)#show gslb dns detail

ZONE: itso.com

ZONE: itso.com
HOST: itso_vm_1:
(Global GSLB policy)
GSLB affinity group: global

Flashback   DNS resp.
delay       selection
(x100us)    counters
TCP APP     Count (%)

* 177.20.0.88 : cfg v-ip  ACTIVE N-AM    0  0  ---
Active Bindings: 1
site: DataCenter1, weight: 50, SI: DC1-SLB1-ADX (192.168.1.2)
session util: 0%, avail. sessions: 7999944
preference: 128

* 178.20.0.88 : cfg v-ip  DOWN  N-AM    -- --  ---
Active Bindings: 1
site: DataCenter2, weight: 50, SI: DC2-SLB1-ADX (192.168.1.3)
```

```
session util: 0%, avail. sessions: 31999728
preference: 128
```

```
telnet@DC1-SLB1-ADX(config)#
```

4.2.3 Application Resource Broker (ARB) server installation

The ARB 2.0 application runs on a server that communicates with vCenter as a plug-in. The ARB 2.0 application requires the following server (physical or virtual) for installation:

- ▶ Processor: Minimum two cores and 2 GHz
- ▶ Memory: Minimum 2 GB RAM
- ▶ Microsoft Windows Server 2003, Windows Server 2008 R2, or Red Hat Enterprise Linux (RHEL) 6.0

The ARB 2.0 plug-in is compatible with VMware vCenter 4.1 or later. Each VM that ARB uses requires VM Tools to be installed. The ARB 2.0 plug-in requires ADX v12.4 or later.

The installation of ARB on a server is fairly straightforward. The ARB server must be accessible to the vCenter server where you want to install the ARB plug-in. Have the fully qualified domain name (FQDN), IP address, and login credentials of this vCenter server ready to refer to before you install ARB.

Figure 4-5 shows the ARB prerequisites.

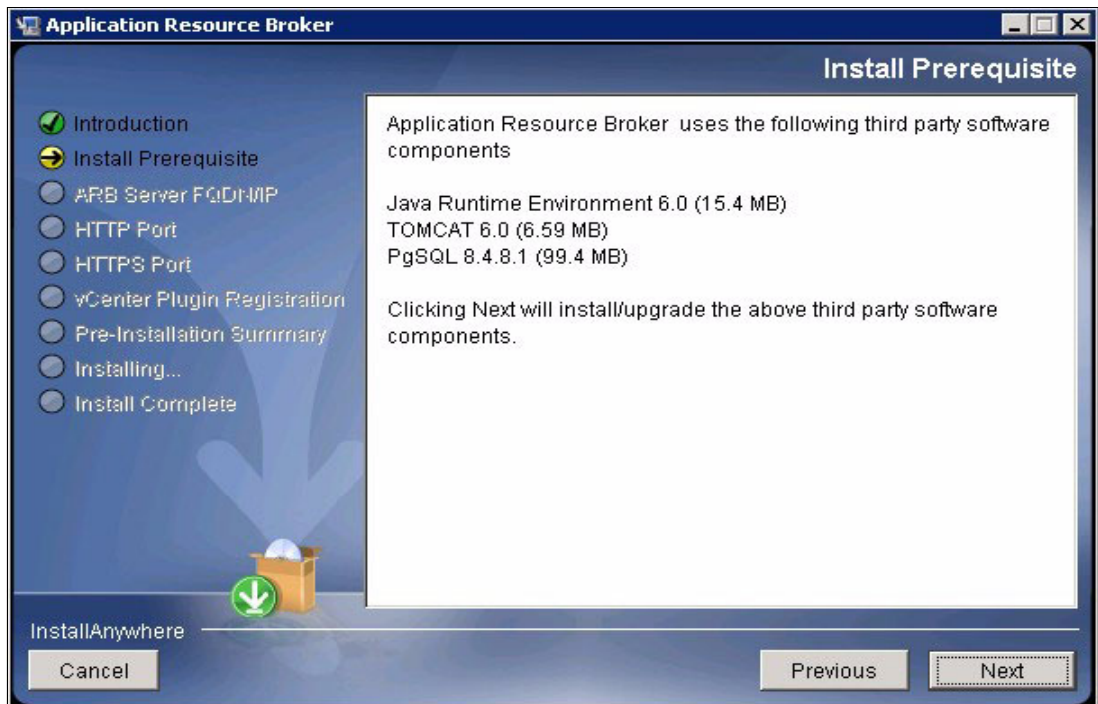


Figure 4-5 ARB installation prerequisites

Figure 4-6 shows ARB Server IP address settings.

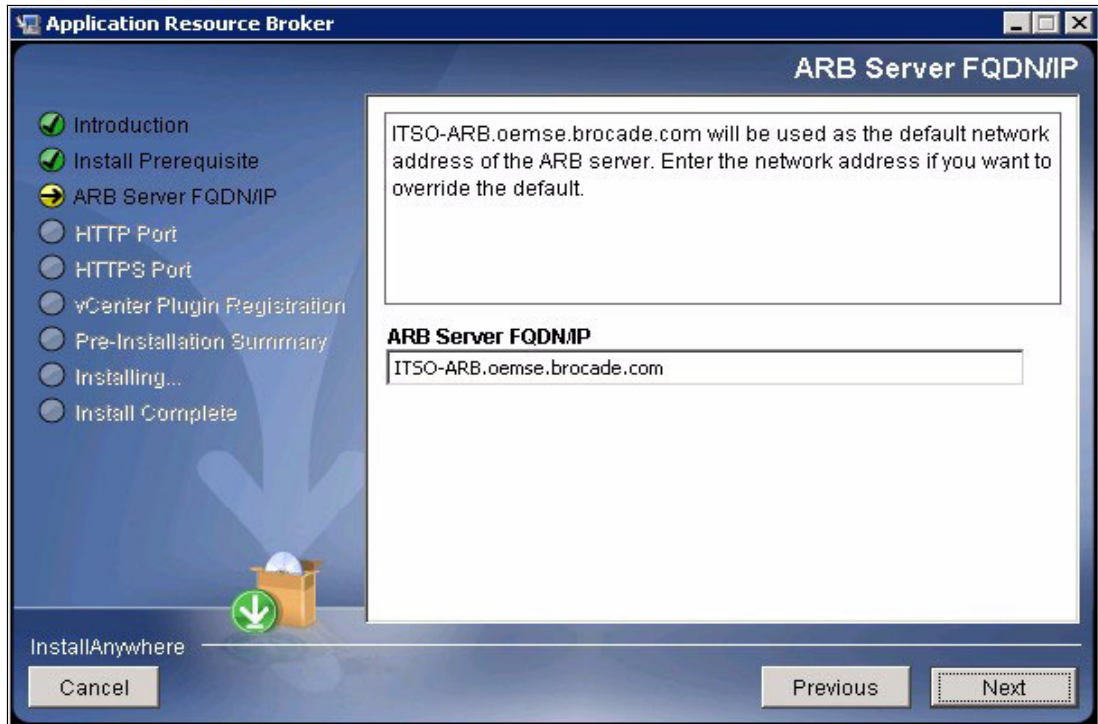


Figure 4-6 ARB Server FQDN/IP address settings

Figure 4-7 shows the ARB vCenter server settings.

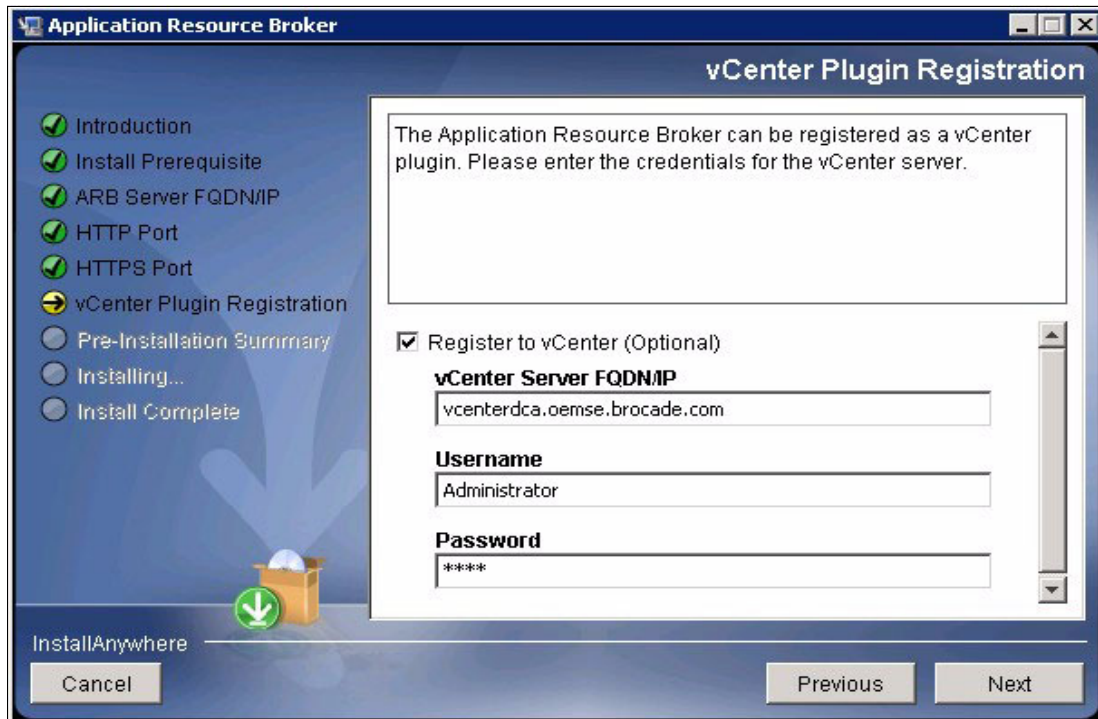


Figure 4-7 ARB vCenter server settings

In Figure 4-8, the ARB installation is complete.

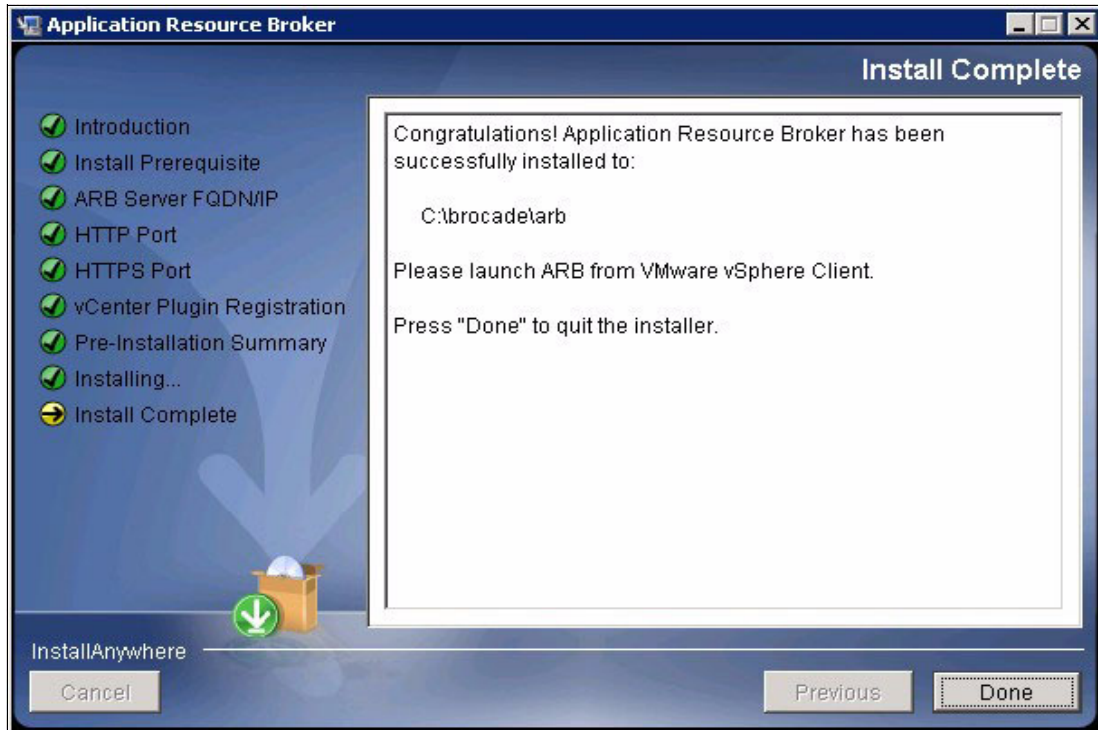


Figure 4-8 ARB installation complete

After you complete the ARB installation, connect to your vCenter server by using the vCenter client, and verify that it was installed correctly by clicking **Plug-ins** → **Manage Plug-ins**. In the Status column, make sure that it says Enabled, as shown in Figure 4-9.

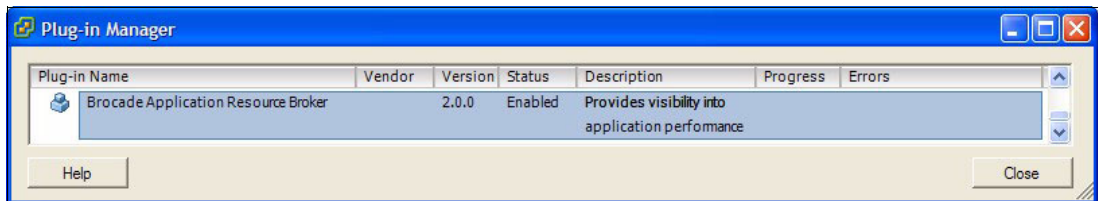


Figure 4-9 Viewing the Brocade ARB plug-in in vCenter

4.2.4 ADX registration in the ARB plug-in

You can access the ARB plug-in interface in the vCenter client by clicking the cluster level and then the Application Resource Broker tab. Then click the ADX Devices tab within the ARB window to register your ADX devices as shown in Figure 4-10.

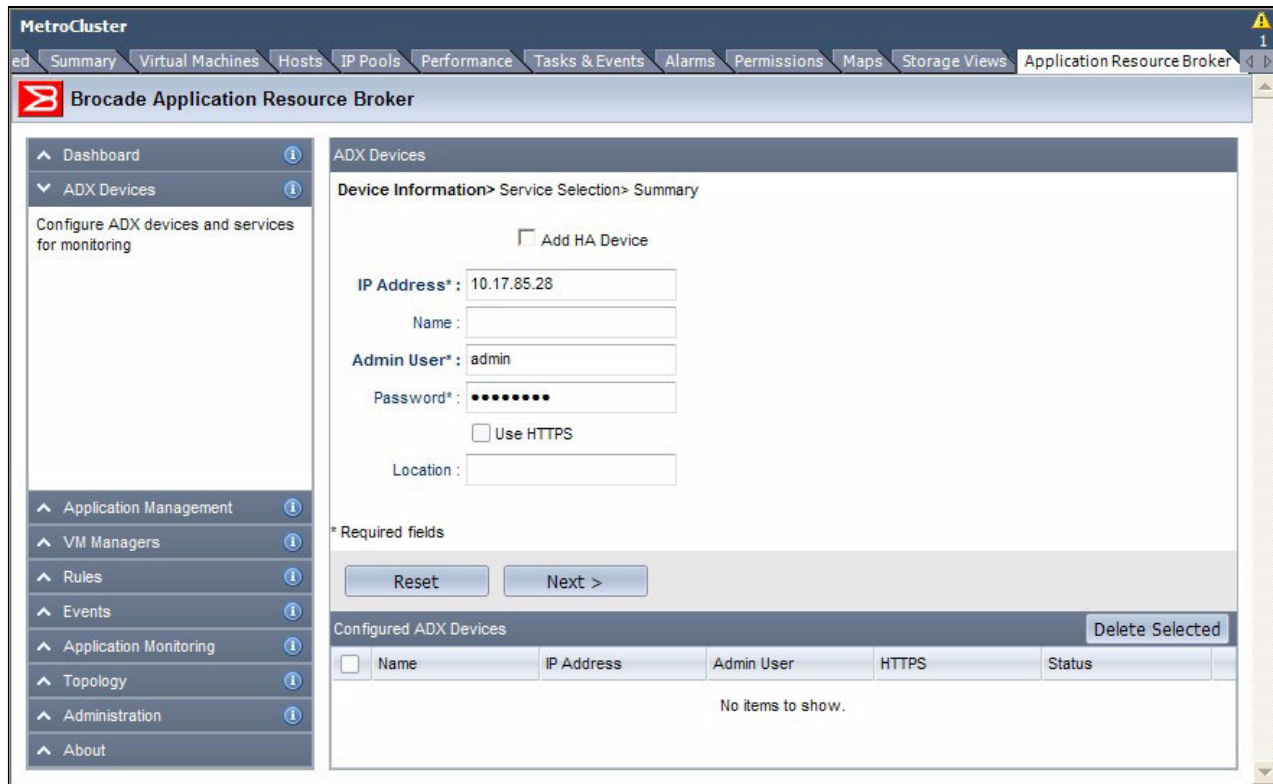


Figure 4-10 Brocade ARB window in vCenter client: ADX Devices tab

Enter the IP management address of your ADX device, along with the HTTP user name and password (defaults: admin, password).

Tip: HTTPS can be used if it is configured.

When you are finished, click **Next**.

Select the configured virtual servers that you want ARB to monitor, as shown in the two panes in Figure 4-11:

- ▶ VIP services not currently imported
- ▶ VIP services already imported

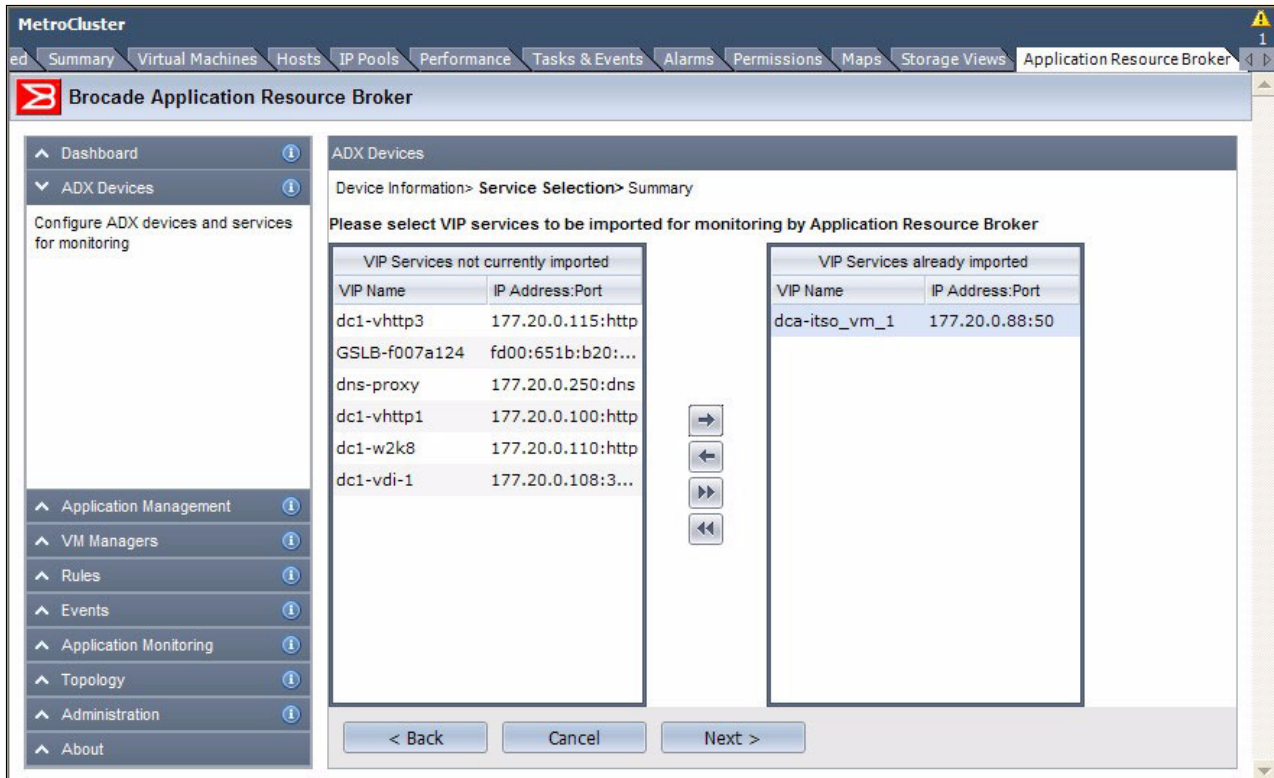


Figure 4-11 Brocade ARB configuration: Selecting the virtual servers for monitoring

On this ADX, select **dca-itso_vm_1** as the VIP that you want monitored. The next window is a summary window for confirmation. Follow the same ADX registration steps for the ADXs in Data Center A and Data Center B. When you are finished, the ADX Devices page will look like Figure 4-12.

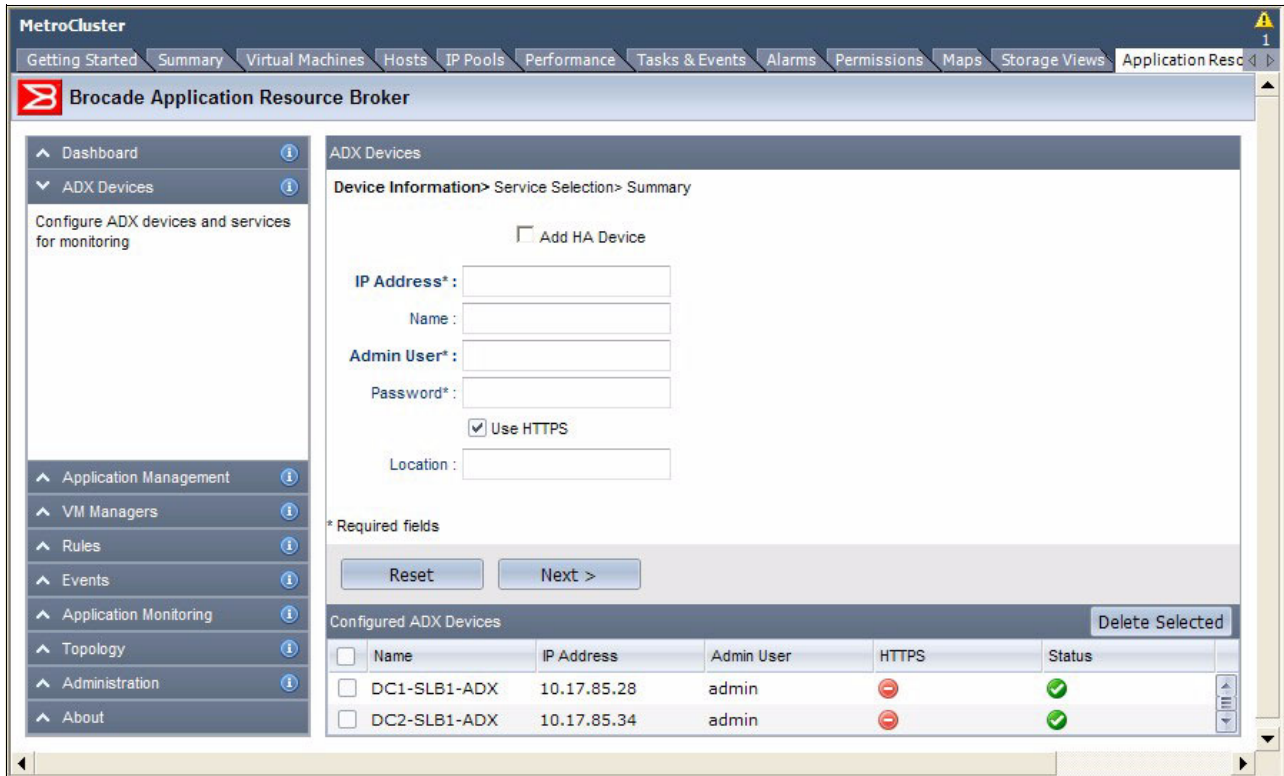


Figure 4-12 Brocade ARB configuration: Both ADXs are now registered

4.2.5 Enable VM mobility in the ARB plug-in in vCenter

To enable VM mobility monitoring for the ITSO_VM_1 VIP, click the Application Management tab in the ARB plug-in window as shown in Figure 4-13.

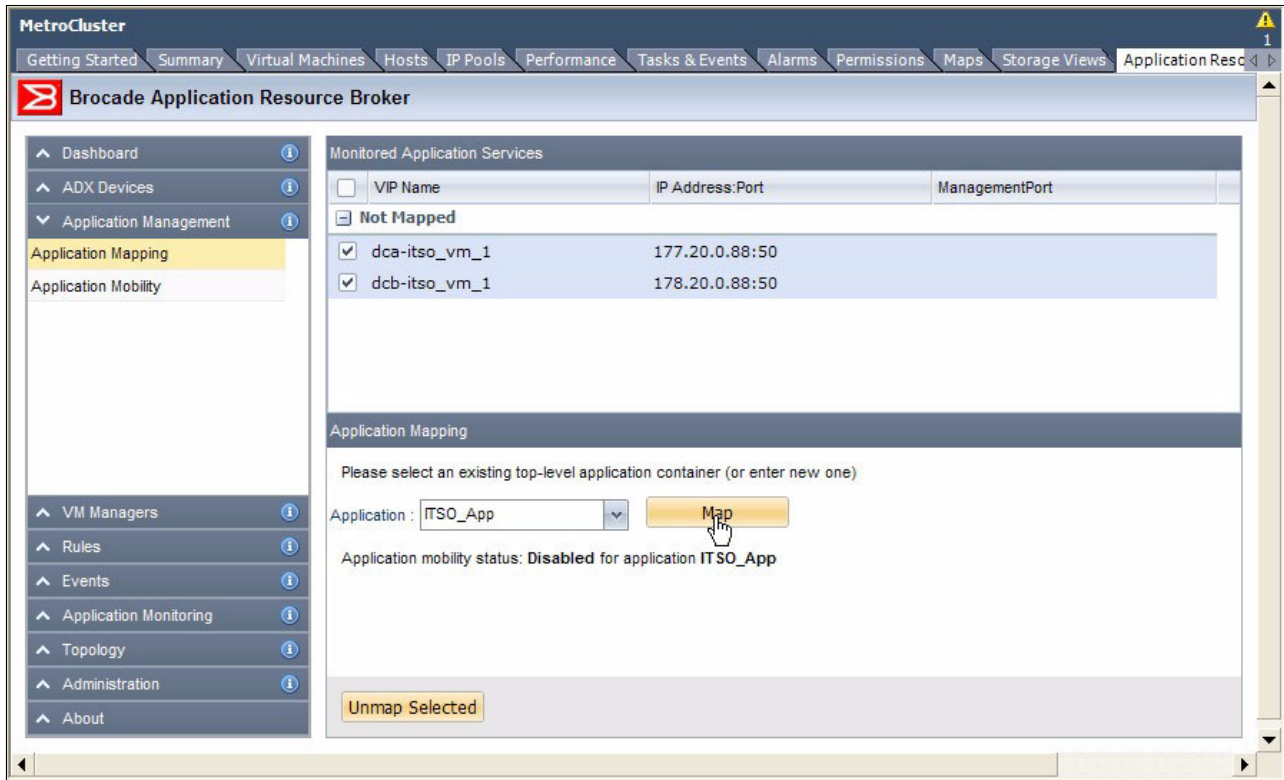


Figure 4-13 Brocade ARB configuration: Registering VIPs to a monitored application

The two virtual servers that you configured on the ADXs in Data Center A and Data Center B, **dca-itso_vm_1** and **dcb-itso_vm_1**, are displayed as **Not Mapped**. Select both of these VIPs to map to the same application because they have the same real server (VM) definition, which is ITSO_VM_1. Map both of them to ITSO_App, and then click **Map**.

After you finish this step, both virtual servers are mapped under ITSO_App, but vMotion Mobility is disabled. Right-click the application, and select **Enable mobility**, as shown in Figure 4-14.

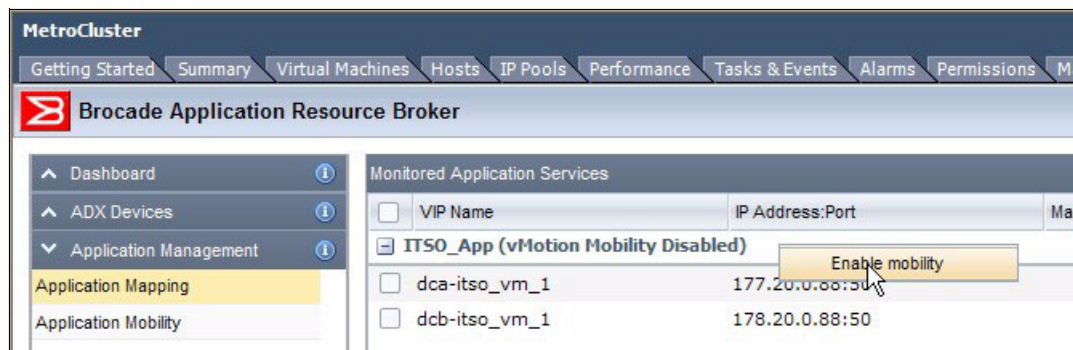


Figure 4-14 Brocade ARB configuration: Enabling VM mobility for a mapped application

Finally, under the Application Mobility tab shown in Figure 4-15, check your configuration, and check which VIP is active and which VIP is backup.

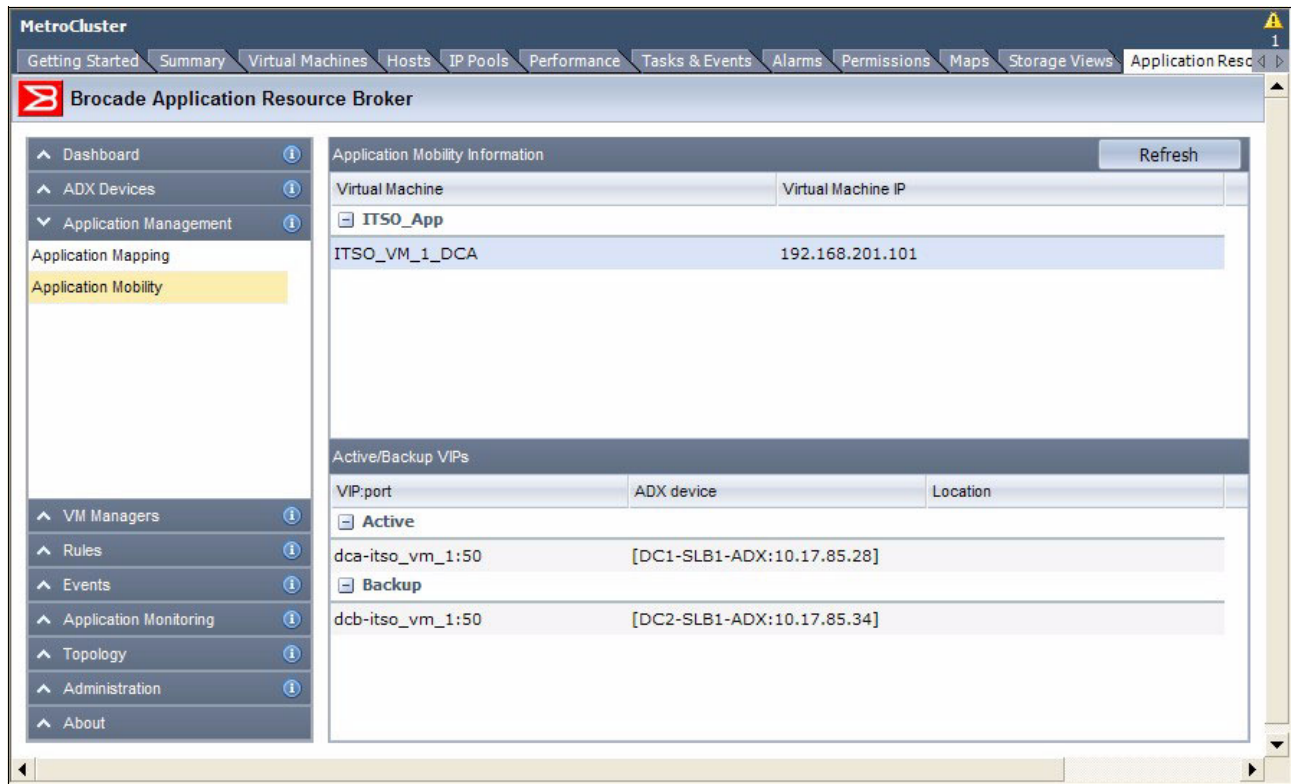


Figure 4-15 Brocade ARB: Checking the Active/Backup VIP for a particular VM

4.2.6 Additional references

For more information, see the following manuals:

- ▶ *Brocade ADX Administration Guide, Supporting v12.4.00*
- ▶ *Brocade ADX Server Load Balancing Guide, Supporting v12.4.00*
- ▶ *Brocade ADX Global Server Load Balancing Guide, Supporting v12.4.00*
- ▶ *Brocade ADX Application Resource Broker Administrator's Guide for v2.0.0*

4.3 IP networking configuration

There are several IP networking components that are used in this solution. End-to-end, data center network design is beyond the scope of this book. However, it covers the configuration of the switches as seen in the context of the setup.

Figure 4-16 shows the IP networking areas.

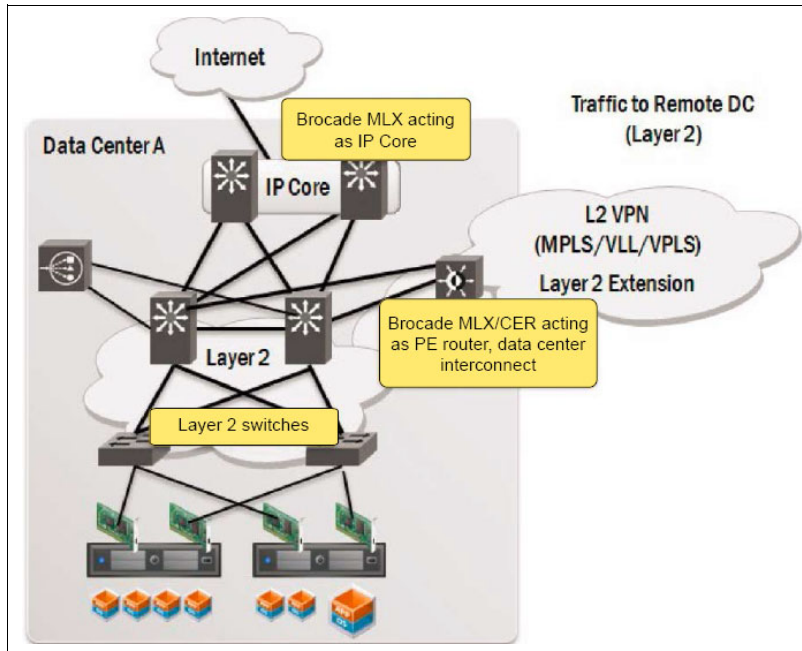


Figure 4-16 High-level IP network architecture

These IP networking components are covered:

- ▶ Layer 2 switches: The ESXi hosts connect directly to Layer 2 switches. The Layer 2 network can be any standards-compliant switch, such as the IBM RackSwitch family of switches or the Brocade VDX series of switches. In a 3-tier network design, there is usually an *edge* made up of top-of-rack switches that connect to a higher density aggregation switch over Layer 2, running a type of Spanning Tree Protocol. With higher-density, top of rack switches, scalable logical switch capabilities, such as stacking or virtual switch clustering, and with passive patch paneling in each rack to a higher density end-of-row switch, the data center network might be collapsed into a flat Layer 2 edge that connects directly to the IP core over either Layer 2 or Layer 3.
- ▶ IP core: The Layer 2 network connects to the IP core, which then connects to the WAN edge and Internet. The configuration example connects from the Layer 2 switches to the IP core over Layer 2. The IP core is made up of Brocade MLXe devices that provide aggregation and routing capabilities among the various subnetworks.
- ▶ Brocade MLXe and CER acting as the data center interconnect or Provider Edge (PE) router: For the Layer 2 extension, the example uses standards-compliant L2 VPN technology that uses Multiprotocol Label Switching (MPLS), Virtual Private LAN Service (VPLS), and Virtual Leased Line (VLL).

Figure 4-17 shows the actual lab configuration.

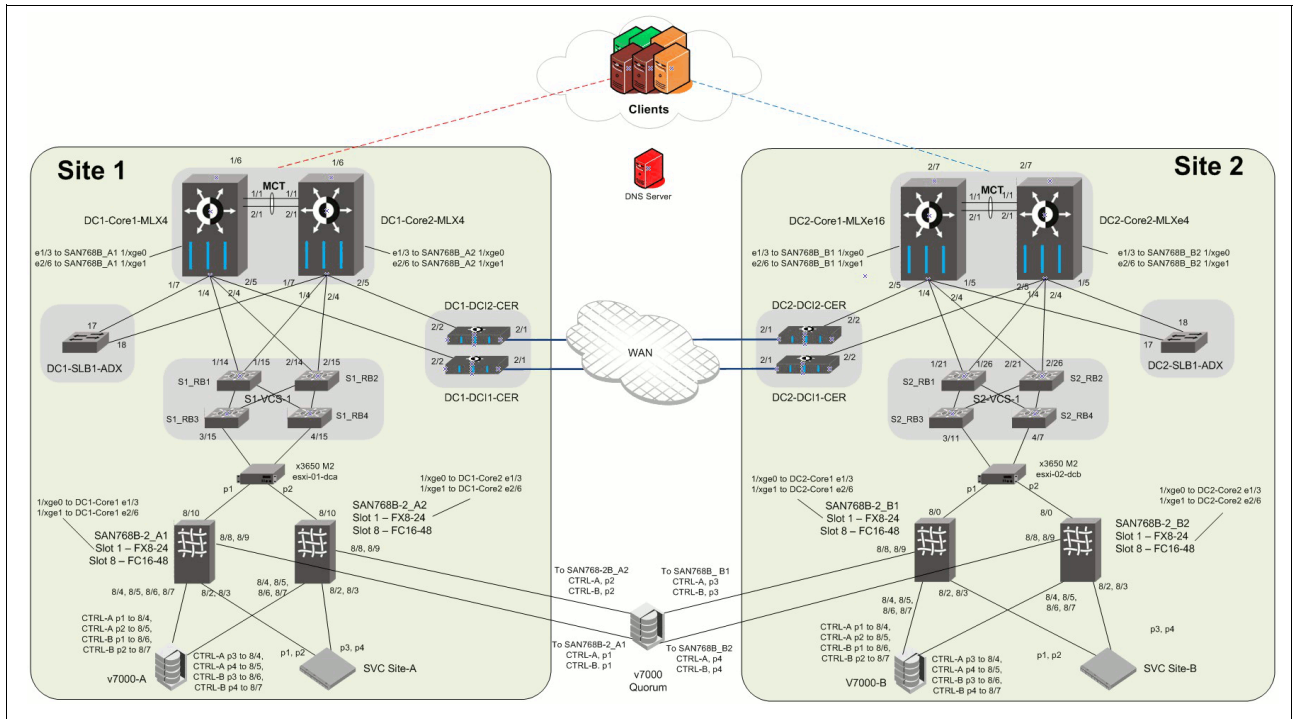


Figure 4-17 Stretched cluster lab architecture

A closer view of just the Data Center A (Site 1) connection is shown in Figure 4-18.

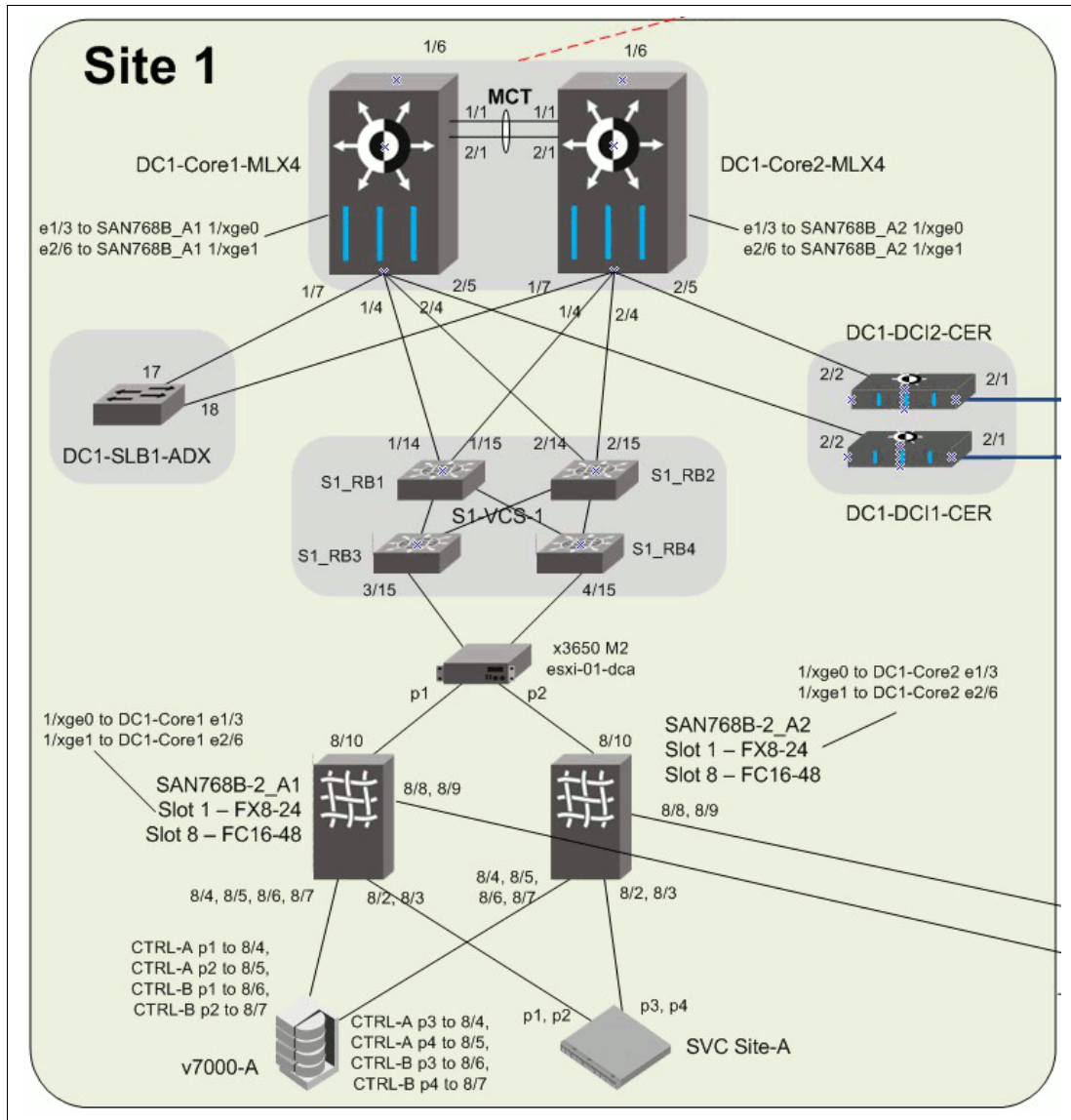


Figure 4-18 Stretched cluster: Data Center A topology

The lab configuration example consists of these components:

- ▶ An IBM x3650 server that is the VMware ESXi host, esxi-01-dca1. This host has two 10 GbE ports and two 16 Gbps FC ports.
- ▶ The 16 Gbps FC ports that are connected to two separate IBM SAN768B-2 chassis, which form two separate, air-gapped FC SAN fabrics.
- ▶ The two 10 GbE ports are connected to two separate 10 GbE VDX switches. There are a total of four VDX switches in a Layer 2 network. These four VDX switches are clustered together by using VCS fabric technology. This configuration makes them look like a single logical switch to other network entities.
- ▶ VDX switches are then connected over Layer 2 to two MLXe routers that act as the IP core. The MLXe routers provide routing between the various subnetworks and aggregate the various connections.
- ▶ The Brocade ADX GSLB/SLB is connected to the MLXe.

- ▶ The data center interconnect links, which are Brocade CES routers, are connected to the MLXs. The Brocade CES routers provide the Layer 2 VPN extension capabilities by using VLL, acting as PE routers.
- ▶ The Brocade MLXe IP Core is also connected to simulated clients that might come in from the Internet, outside the data center.
- ▶ Also connected to the Brocade MLXe are two 10 GbE FCIP connections from each IBM SAN768B chassis. IBM SAN768B-A1 has two links that connect to DC1-Core1-MLX4. One link forms the Private SAN FCIP tunnel, and the second link forms the Public SAN FCIP tunnel. Similarly, IBM SAN768B-A2 has two links that are connected to DC1-Core2-MLX4.

4.3.1 Layer 2 switch configuration

The Layer 2 switch has the following configuration in this example.

Layer 2 loop prevention

Within the Layer 2 network, loop prevention is required. Typically a type of Spanning Tree protocol is used, usually either PVRSTP, MSTP, or RSTP depending on the environment. Consider the following guidelines when you select Spanning Tree configuration:

- ▶ Configure the Bridge Priority to ensure correct Root Bridge placement, which is typically at one of the Aggregation switches closest to the core.
- ▶ Configure switch-to-switch links as point-to-point, and edge ports that connect to the ESXi hosts as edge ports.
- ▶ Although the ESXi vSwitches do not run xSTP, leave xSTP enabled on the edge ports in case something is mis-wired in the future.
- ▶ Consider enabling BPDU guard on edge ports that face the ESXi hosts to prevent loops. Remember that this configuration cuts off access to all VMs behind that port if triggered.

VLAN configuration

VLANs must also be created for the various networks that vSphere and VMs use. As outlined in the Design Chapter, four VLANs must be defined on all switches in the Layer 2 network. This configuration is shown in Example 4-7.

Example 4-7 Defining VLANs on a switch

```

S1_RB4(config)# int vlan 700
S1_RB4(config-vlan-700)# description Management
S1_RB4(config-vlan-700)# int vlan 701
S1_RB4(config-vlan-701)# description VM_Traffic
S1_RB4(config-vlan-701)# int vlan 702
S1_RB4(config-vlan-702)# description vMotion
S1_RB4(config-vlan-702)# int vlan 703
S1_RB4(config-vlan-703)# description Fault_Tolerant

```

The edge ports on the switch that is connected to the ESXi host and all traditional Ethernet links between switches must also be configured as VLAN trunk ports that allow VLANs 700-703. A configuration example is shown in Example 4-8.

Example 4-8 Configuring switch to ESXi host and switch-to-switch links to carry VLANs

```

S1_RB4(config)# int ten 4/0/15
S1_RB4(conf-if-te-4/0/15)# switchport

```

```
S1_RB4(config-if-te-4/0/15)# switchport mode trunk  
S1_RB4(config-if-te-4/0/15)# switchport trunk allowed vlan add 700-703
```

Link Aggregation Port Group (LAG) configuration

In the example configuration, the ESXi uses the “route based on originating virtual port” load balancing. Therefore, a LAG does not need to be configured between the two switches that are connected to S1_RB3 and S1_RB4.

However, create a LAG between S1_RB1 and S1_RB2 to the two MLXs that act as the IP Core. This configuration is possible because your four Layer 2 switches are in a single logical cluster, S1-VCS-1, using Virtual Cluster Switching (VCS) fabric technology. The two MLXs are also clustered together by using Brocade Multi-Chassis Trunking (MCT) technology.

To create a port-channel group on the switches in S1-VCS-1, place all four ports in the same port channel. Example 4-9 shows placing the ports on Switch S1_RB1.

Example 4-9 Switch S1_RB1: Placing ports connected to the core MLXs in channel-group 2

```
S1_RB1(config)# int ten 1/0/14  
S1_RB1(config-if-te-1/0/14)# channel-group 2 mode active type standard  
S1_RB1(config-if-te-1/0/14)# int ten 1/0/15  
S1_RB1(config-if-te-1/0/15)# channel-group 2 mode active type standard
```

Example 4-10 shows placing the ports on Switch S1_RB2.

Example 4-10 Switch S1_RB2: Placing ports connected to the core MLXs in channel-group 2

```
S1_RB1(config)# int ten 2/0/14  
S1_RB1(config-if-te-2/0/14)# channel-group 2 mode active type standard  
S1_RB1(config-if-te-2/0/14)# int ten 2/0/15  
S1_RB1(config-if-te-2/0/15)# channel-group 2 mode active type standard
```

Now that the port channels are created on both switches, specify the interface definitions at the port-channel level, as shown in the following examples. Example 4-11 shows Switch S1_RB1.

Example 4-11 Switch S1_RB1: Interface Port-Channel 2 configuration

```
S1_RB1(config)# int port-channel 2  
S1_RB1(config-Port-channel-2)# switchport  
S1_RB1(config-Port-channel-2)# switchport mode trunk  
S1_RB1(config-Port-channel-2)# switchport trunk allowed vlan add 701-703
```

Example 4-12 shows the configuration of Switch S1_RB2.

Example 4-12 Switch S1_RB2: Interface Port-Channel 2 configuration

```
S1_RB2(config)# int port-channel 2  
S1_RB2(config-Port-channel-2)# switchport  
S1_RB2(config-Port-channel-2)# switchport mode trunk  
S1_RB2(config-Port-channel-2)# switchport trunk allowed vlan add 701-703
```

Other considerations

The maximum transmission unit (MTU) size can be increased on all Layer 2 ports to support jumbo frames. This configuration can help with traffic efficiency when you are using IP-based storage or vMotion traffic. Configuration of jumbo frames can vary from switch to switch, from being a global to an interface-level command.

Remember: Make sure that the MTU size on all interfaces within the Layer 2 domain is the same, including for the ESXi host configuration. Otherwise, fragmentation might occur.

Keep these basic considerations in mind when you are setting up your network:

- ▶ Harden the switch by using role-based access control (or RBAC, for example, a RADIUS or TACACS+ server), user accounts, disabling Telnet, and restricting login to certain VLANs or IP ranges.
- ▶ Configure SNMP, Syslog, or sFlow monitoring servers.
- ▶ Configure a Network Time Protocol (NTP) server.
- ▶ Enable Link Layer Discovery Protocol (LLDP) and Cisco Discovery Protocol (CDP), or extra network visibility.
- ▶ Configure quality of service for specific traffic classes, or trusting the 802.1 p settings that are sent by ESXi.
- ▶ Configure security and access control lists.

Additional references

For more information about the Brocade VDX, see these manuals:

- ▶ *Network OS Administrator's Guide, Supporting v2.1.1*
- ▶ *Network OS Command Reference Manual, Supporting v2.1.1*

4.3.2 IP Core (MLXe) configuration

Keep the following considerations in mind when you are configuring IP Core (MLXe):

- ▶ VLAN configuration
- ▶ VRRPE: Layer 3 gateway configuration
- ▶ MCT Trunking configuration
- ▶ Other considerations

VLAN configuration

Start by defining the VLANs that the MLXe carries traffic over. These are the same four VLANs, 700 – 703, defined in the Layer 2 switch network. However, the MLXe also carries the following FCIP traffic:

- ▶ VLAN 705 - Fabric A - Private SAN traffic
- ▶ VLAN 706 - Fabric A - Public SAN traffic
- ▶ VLAN 707 - Fabric B - Private SAN traffic
- ▶ VLAN 708 - Fabric B - Public SAN traffic

Within the lab topology, there are three different types of network traffic access patterns for traffic that passes through the MLXe.

Example 4-13 shows the different network traffic types and interfaces applicable on DC1-Core1-MLXe4:

- ▶ Traffic that needs access to only the “internal” network. These are ports that are connected to the Layer 2 network (eth 1/4, eth 2/4), the MCT link between the two MLXes (eth 1/1, eth 2/1), and the data center interconnect link to the CES (eth 2/5). Traffic types that fall into this category are the management, vMotion, and fault-tolerant traffic. Although the management traffic might need to be routed to a different subnetwork (which is why you create a virtual interface in that VLAN), the vMotion and fault-tolerant traffic should never need to be routed.
- ▶ Traffic that needs access to the “internal” network, the external client traffic (eth 1/6), and the ADX GSLB/SLB (eth 1/7). This is the VM traffic that might also need to be routed.
- ▶ FCIP traffic that requires access to directly connected IBM SAN768B-2 (eth 1/3, eth 2/6), the MCT link between the two MLXs (eth 1/1, eth 2/1), and the data center interconnect link to the CES (eth 2/5).

An example of creating these VLANs and tagging the appropriate interfaces is shown in Example 4-13.

Example 4-13 Creating VLANs and virtual routing interfaces on DC1-Core1-MLXe4

```
telnet@DC1-Core1-MLXe(config)#vlan 700 name I-MGMT
telnet@DC1-Core1-MLXe(config-vlan-700)#tag eth 1/1 eth 1/4 eth 2/1 eth 2/4 to 2/5
telnet@DC1-Core1-MLXe(config-vlan-700)#router-interface ve 70

telnet@DC1-Core1-MLXe(config-vlan-700)#vlan 701 name I-VM_Traffic
telnet@DC1-Core1-MLXe(config-vlan-701)#tag ethe 1/1 eth 1/4 eth 1/6 to 1/7 eth 2/1
eth 2/4 to 2/5
telnet@DC1-Core1-MLXe(config-vlan-701)#router-interface ve 71

telnet@DC1-Core1-MLXe(config-vlan-701)#vlan 702 name I-vMotion
telnet@DC1-Core1-MLXe(config-vlan-702)#tag eth 1/1 eth 1/4 eth 2/1 eth 2/4 to 2/5

telnet@DC1-Core1-MLXe(config-vlan-702)#vlan 703 name I-Fault_Tolerant
telnet@DC1-Core1-MLXe(config-vlan-703)#tag ethe 1/1 eth 1/4 eth 1/6 to 1/7 eth 2/1
eth 2/4 to 2/5

telnet@DC1-Core1-MLXe(config-vlan-703)#vlan 705 name I-FCIP-Priv-FabA
telnet@DC1-Core1-MLXe(config-vlan-705)#untag eth 1/3
telnet@DC1-Core1-MLXe(config-vlan-705)#tag eth 1/1 eth 2/1 eth 2/5

telnet@DC1-Core1-MLXe(config-vlan-705)#vlan 706 name I-FCIP-Pub-FabA
telnet@DC1-Core1-MLXe(config-vlan-706)#untag eth 2/6
telnet@DC1-Core1-MLXe(config-vlan-706)#tag ethe 1/1 ethe 2/1 ethe 2/5

telnet@DC1-Core1-MLXe(config-vlan-706)#vlan 707 name I-FCIP-Priv-FabB
telnet@DC1-Core1-MLXe(config-vlan-707)#tag ethe 1/1 ethe 2/1 ethe 2/5

telnet@DC1-Core1-MLXe(config-vlan-707)#vlan 708 name I-FCIP-Pub-FabB
telnet@DC1-Core1-MLXe(config-vlan-708)#tag ethe 1/1 ethe 2/1 ethe 2/5
```

The connection to the data center interconnects, DC1-DCI1-CER and DC1-DCI2-CER, are simply a Layer 2 link. In this configuration, the MLXes are acting as the Customer Edge routers, whereas the CER is configured with MPLS and VLL acting as the Provider Edge routers.

VRRPE: Layer 3 gateway configuration

Each VLAN that requires routing has a virtual router interface created. The Management VLAN has ve 70 created and the VM Traffic VLAN ve 71 within those VLANs. These virtual interfaces must be configured with an IP address to route traffic. The example uses Virtual Router Redundancy Protocol-Extended (VRRPE) to provide active-active Layer 3 gateways for your VMs on each of those virtual routing interfaces.

A VRRP instance (VRID) is created with a single VIP that VMs use as their gateway address. Within the VRID, one or more router interfaces are also configured as the actual paths that traffic passes through. With VRRPE, all router interfaces within a VRID can route traffic instead of having to be switched through a single designated Master interface in the case of regular VRRP.

Example 4-14 shows how to enable VRRPE on a router.

Example 4-14 Enabling the VRRPE protocol on a router

```
telnet@DC1-Core1-MLXe(config)#router vrrp-extended
telnet@DC1-Core1-MLXe(config-vrrpe-router)#exit
telnet@DC1-Core1-MLXe(config)#
```

An IP address must be configured for interface ve 70, in this case 192.168.200.250/24. Next, configure VRRPE and select an arbitrary VRID value of 70. In the context of VRRPE, all VRRPE interfaces are designated as backup interfaces. There is no one Master interface. The VIP of the VRID is also defined, in this case 192.168.200.1. Finally, enable short-path-forwarding, which allows any VRRPE interface to route traffic instead of having to switch to a designated Master interface. Finally, activate the VRRPE configuration.

Example 4-15 shows how to configure VRRPE on interface ve 70 on DC1-Core1-MLXe.

Example 4-15 Configuring VRRPE on DC1-Core1-MLXe4, interface ve 70

```
telnet@DC1-Core1-MLXe(config)#interface ve 70
telnet@DC1-Core1-MLXe(config-vif-70)#port-name I-Internal-Mgmt
telnet@DC1-Core1-MLXe(config-vif-70)#ip address 192.168.200.250/24
telnet@DC1-Core1-MLXe(config-vif-70)#ip vrrp-extended vrid 70
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#backup
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#advertise backup
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#ip-address 192.168.200.1
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#short-path-forwarding
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#activate
```

A similar configuration can be done on DC1-Core2-MLXe4, and the two routers in Data Center B, DC2-Core1-MLXe16 and DC2-Core2-MLXe4. A different/unique IP address for each virtual interface on those routers must be selected, although the VRRPE VRID and VIP remain the same.

Example 4-16 shows how to configure DC1-Core2-MLXe4.

Example 4-16 Configuring VRRPE on DC1-Core2-MLXe4, interface ve 70

```
telnet@DC1-Core2-MLXe(config)#interface ve 70
telnet@DC1-Core2-MLXe(config-vif-70)#port-name I-Internal-Mgmt
telnet@DC1-Core2-MLXe(config-vif-70)#ip address 192.168.200.251/24
telnet@DC1-Core2-MLXe(config-vif-70)#ip vrrp-extended vrid 70
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#backup
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#advertise backup
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#ip-address 192.168.200.1
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#short-path-forwarding
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#activate
```

A third example configuration of DC2-Core1-MLXe16 is shown in Example 4-17.

Example 4-17 Configuring VRRPE on DC2-Core1-MLXe16, interface ve 70

```
telnet@DC1-Core2-MLXe(config)#interface ve 70
telnet@DC1-Core2-MLXe(config-vif-70)#port-name I-Internal-Mgmt
telnet@DC1-Core2-MLXe(config-vif-70)#ip address 192.168.200.252/24
telnet@DC1-Core2-MLXe(config-vif-70)#ip vrrp-extended vrid 70
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#backup
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#advertise backup
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#ip-address 192.168.200.1
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#short-path-forwarding
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#activate
```

MCT Trunking configuration

The two MLXes are also configured together in an MCT cluster. First, create a LAG of two ports between DC1-Core1-MLX4 and DC1-Core2-MLX4 over ports eth 1/1 and eth 2/1 to use as the MCT Inter-Chassis Link (ICL). Example 4-18 shows an example of a static LAG configuration from the DC1-Core1-MLX4 chassis. Create a similar configuration on the DC1-Core2-MLX4.

Example 4-18 MLXe: Creating a static LAG

```
telnet@DC1-Core1-MLXe(config)#lag "ICL" static id 1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#ports eth 1/1 eth 2/1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#primary-port 1/1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#deploy
telnet@DC1-Core1-MLXe(config-lag-ICL1)#port-name "ICL" ethernet 1/1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#port-name "ICL" ethernet 2/1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#int eth 1/1
telnet@DC1-Core1-MLXe(config-if-e10000-1/1)#enable
```

After the LAG is created and up between the two MLXes, configure a VLAN along with a virtual router interface to carry MCT cluster communication traffic, as shown in Example 4-19.

Example 4-19 Creating a VLAN for MCT cluster communication traffic

```
telnet@DC1-Core1-MLXe(config)#vlan 4090 name MCT_SESSION_VLAN
telnet@DC1-Core1-MLXe(config-vlan-4090)#tag ethe 1/1
telnet@DC1-Core1-MLXe(config-vlan-4090)#router-interface ve 1
```

Tip: Only the 1/1 must be tagged because it is the primary port in the LAG.

Next, configure the virtual routing interface with an IP address that the two MLXes use for MCT cluster communication, as shown in Example 4-20.

Example 4-20 Assigning an IP address to a virtual routing interface

```
telnet@DC1-Core1-MLXe(config)#interface ve 1
telnet@DC1-Core1-MLXe(config-vif-1)#port-name MCT-Peer
telnet@DC1-Core1-MLXe(config-vif-1)#ip address 1.1.1.1/24
```

Finally, configure the cluster as shown in Example 4-21. More information about each step can be found in the *Brocade MLXe Configuration Guide*.

Example 4-21 Completing the MCT cluster configuration

```
telnet@DC1-Core1-MLXe(config)#cluster MCT_CLUSTER 1
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#rbridge-id 10
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#session-vlan 4090
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#member-vlan 100 to 999
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#icl ICL ethernet 1/1
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#peer 1.1.1.2 rbridge-id 20 icl
ICL
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#deploy
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#client DC1-SLB1-ADX
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-DC1-SLB1-ADX)#rbridge-id
100
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-DC1-SLB1-ADX)#client-inter
face ethernet 1/7
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-DC1-SLB1-ADX)#deploy
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-DC1-SLB1-ADX)#exit
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#client VCS1
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-VCS1)#rbridge-id 200
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-VCS1)#client-interface
ethernet 1/4
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-VCS1)#deploy
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-VCS1)#exit
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#client CER_DCI
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-CER_DCI)#rbridge-id 300
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-CER_DCI)#client-interface
ethernet 2/5
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-CER_DCI)#deploy
```

Other considerations

The Brocade MLXe routers are high-performance routers that are capable of handling the entire Internet routing table, with advanced MPLS, VPLS, and VLL features, and other high-end service provider-grade capabilities. However, it is beyond the scope of this book to address IP network design, routing protocols, and so on.

Additional references

For more information about the Brocade MLXe, see the *Brocade MLX Series and NetIron Family Configuration Guide, Supporting r05.3.00* or the product web page:

<http://www.brocade.com/products/all/routers/product-details/netiron-mlx-series/index.page>

4.3.3 Data Center Interconnect (Brocade CER series) configuration

The Layer 2 network must be extended from Data Center A to Data Center B to support VM mobility. This configuration can be done by using a Layer 2 VPN technology that uses standards-based MPLS, VPLS, or VLL technology. This is typically provided by the service provider.

However, in some situations, it is beneficial to extend the Layer 2 VPN deeper into the data center. For example, two data centers that are connected point-to-point over dark fiber and MPLS can be used for advanced QoS control or other purposes. The Brocade MLXe chassis and Brocade NetIron CER 1 RU switches both support MPLS, VPLS, and VLL capabilities.

It is beyond the scope of this book to address this technology, because it is complex. However, the configurations between two of the CER interconnects in the lab setup are shown for reference in Example 4-22 and Example 4-23 on page 79. Example 4-22 shows the DC1-DC11-CER configuration.

Example 4-22 DC1-DC11-CER configuration

```
router ospf
  area 0
  !
  interface loopback 1
    ip ospf area 0
    ip address 80.80.80.10/32
  !
  interface ethernet 2/1
    enable
    ip ospf area 0
    ip ospf network point-to-point
    ip address 200.10.10.1/30
  !
  interface ethernet 2/2
    enable
  !
router mpls

  mpls-interface e1/1
    ldp-enable

  mpls-interface e2/1
    ldp-enable

  vll DCIv1700 700
    vll-peer 90.90.90.10
    vlan 700
    tagged e 2/2

  vll DCIv1701 701
    vll-peer 90.90.90.10
    vlan 701
    tagged e 2/2

  vll DCIv1702 702
    vll-peer 90.90.90.10
    vlan 702
```



```

tagged e 2/2

vll DCIv1703 703
vll-peer 90.90.90.10
vlan 703
tagged e 2/2

vll DCIv1705 705
vll-peer 90.90.90.10
vlan 705
tagged e 2/2

vll DCIv1706 706
vll-peer 90.90.90.10
vlan 706
tagged e 2/2

vll DCIv1707 707
vll-peer 90.90.90.10
vlan 707
tagged e 2/2

vll DCIv1708 708
vll-peer 90.90.90.10
vlan 708
tagged e 2/2

```

Example 4-23 shows the configuration of SC2-DCI1-CER in the lab environment.

Example 4-23 DC2-DCI1-CER configuration

```

router ospf
area 0
!
interface loopback 1
ip ospf area 0
ip address 90.90.90.10/32
!
interface ethernet 2/1
enable
ip ospf area 0
ip ospf network point-to-point
ip address 200.10.10.2/30
!
interface ethernet 2/2
enable
!
router mpls

mpls-interface e1/1
ldp-enable

mpls-interface e2/1
ldp-enable

vll DCIv1700 700

```

```
vll-peer 80.80.80.10
vlan 700
tagged e 2/2
```

```
vll DCIv1701 701
vll-peer 80.80.80.10
vlan 701
tagged e 2/2
```

```
vll DCIv1702 702
vll-peer 80.80.80.10
vlan 702
tagged e 2/2
```

```
vll DCIv1703 703
vll-peer 80.80.80.10
vlan 703
tagged e 2/2
```

```
vll DCIv1705 705
vll-peer 80.80.80.10
vlan 705
tagged e 2/2
```

```
vll DCIv1706 706
vll-peer 80.80.80.10
vlan 706
tagged e 2/2
```

```
vll DCIv1707 707
vll-peer 80.80.80.10
vlan 707
tagged e 2/2
```

```
vll DCIv1708 708
vll-peer 80.80.80.10
vlan 708
tagged e 2/2
```

For more information about the Brocade CER, see *Brocade MLX Series and NetIron Family Configuration Guide, Supporting r05.3.00*.

4.4 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 are connected 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 these components:

- ▶ IBM FC 8 Gbps Fibre Channel over IP (FCIP) Extension blade in Slot 1
- ▶ IBM FC 16 Gbps 48-port blade in Slot 8

Enhanced Stretched Cluster (ESC) requires two types of SAN:

- ▶ **Public SAN:** In this type, server hosts, storage, and Spectrum Virtualize nodes connect. Data storage traffic traverses the public SAN.
- ▶ **Private SAN:** Only the Spectrum Virtualize nodes connect into the private SAN, which is used for cluster communication.

Each IBM SAN768B-2 is split into two logical chassis to implement segregated private and public SANs on the same chassis. Figure 4-19 shows the various public and private SAN connections, with each in a different color.

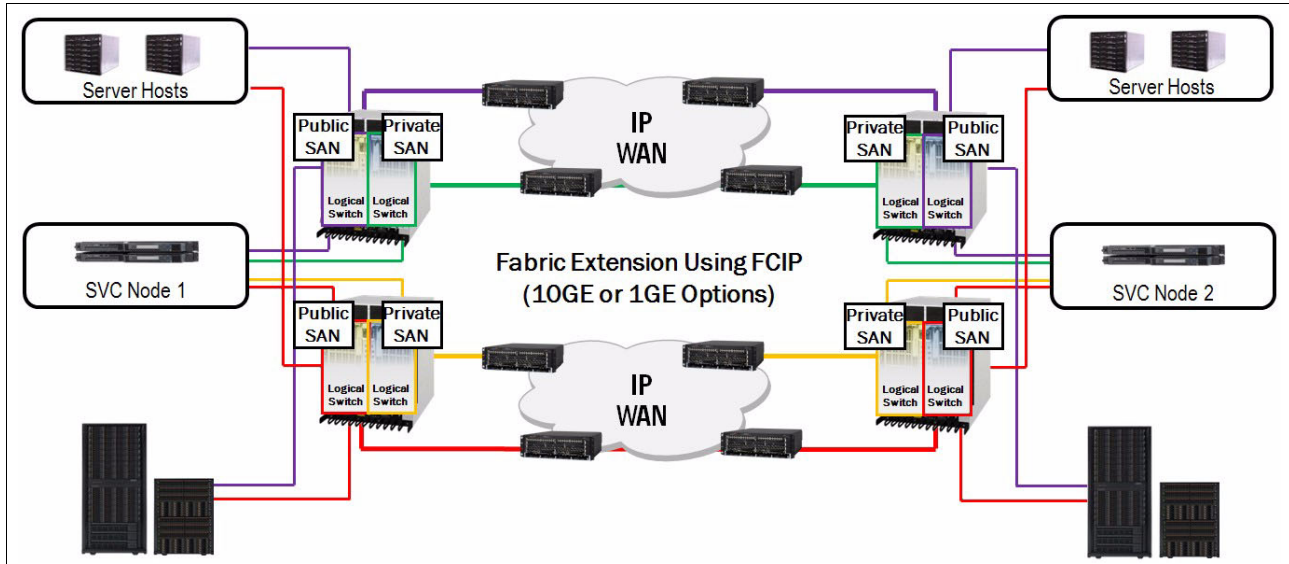


Figure 4-19 Public and private SAN connections at each data center site

Figure 4-20 shows the actual lab environment.

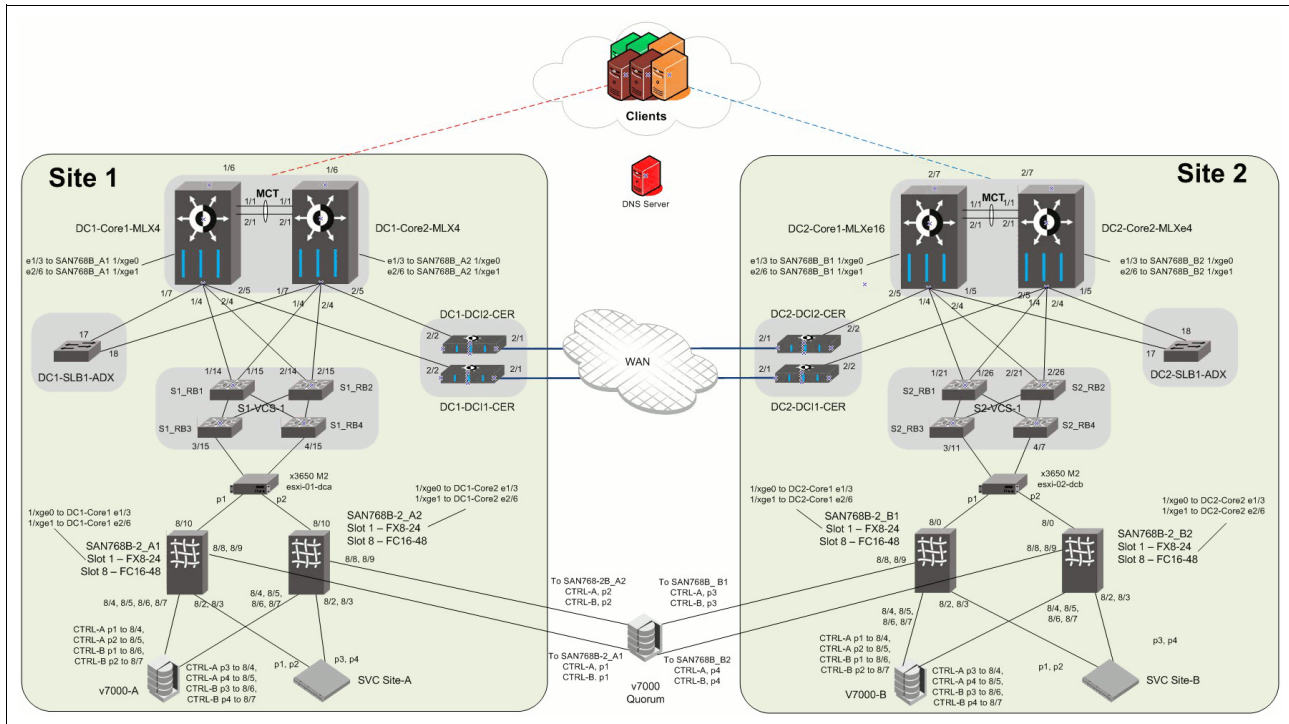


Figure 4-20 Stretched cluster lab topology

4.4.1 Creating the logical switches

Configure a total of four fabrics, each with two different logical switches.

Figure 4-21 shows a closer look at the SAN port topology in Data Center A.

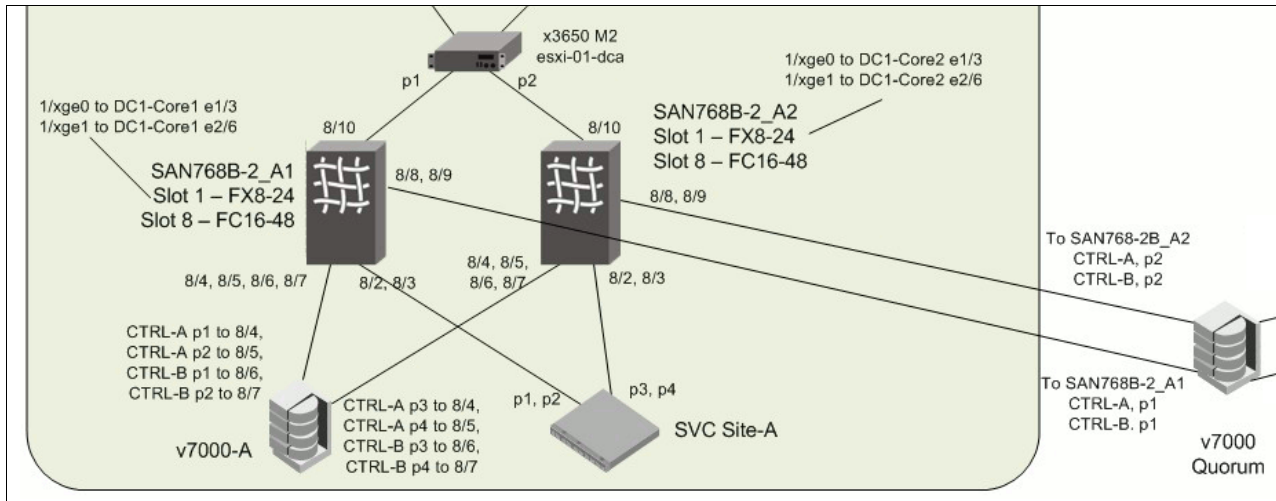


Figure 4-21 Data Center A: SAN port topology

Figure 4-22 shows a closer look at the SAN port topology in Data Center B.

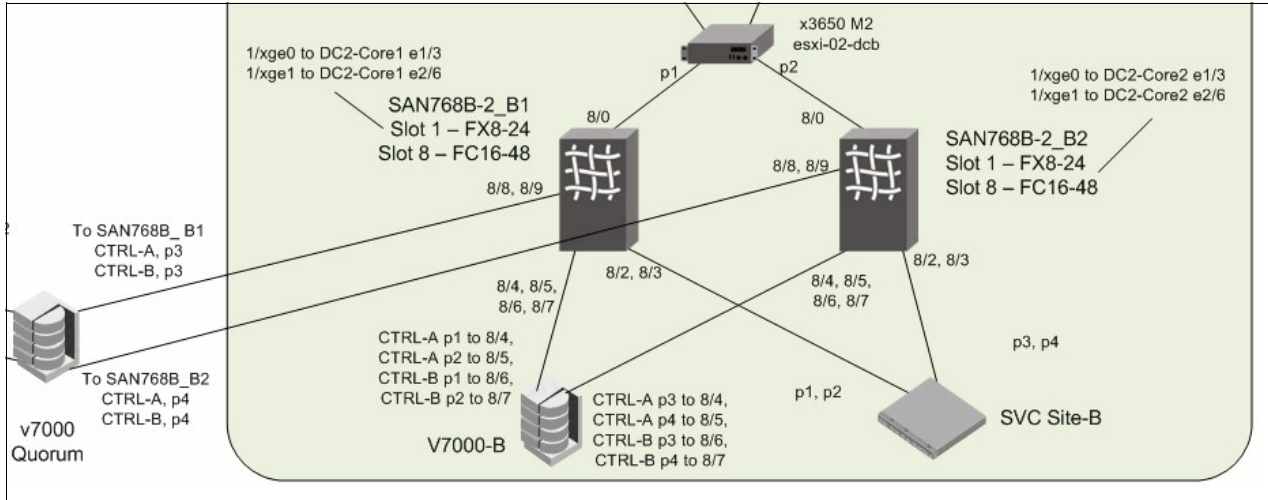


Figure 4-22 Data Center B: SAN port topology

Next, create logical switches on the SAN768B-2s and map them according to Table 4-1.

Table 4-1 Fabric to physical switch to logical switch mappings

Fabric name	Physical switch	Logical switch	LS #	Ports
Fabric-Public-1	SAN768B-2_A1	Public_A1	111	1/12, 8/2, 8/4, 8/5, 8/6, 8/7, 8/8, 8/9, 8/10
	SAN768B-2_B1	Public_B1		1/12, 8/0, 8/1, 8/2, 8/4, 8/5, 8/6, 8/7, 8/8, 8/9
Fabric-Public-2	SAN768B-2_A2	Public_A2	113	1/12, 8/2, 8/4, 8/5, 8/6, 8/7, 8/8, 8/9, 8/10
	SAN768B-2_B2	Public_B2		1/12, 8/0, 8/1, 8/2, 8/4, 8/5, 8/6, 8/7, 8/8, 8/9
Fabric-Private-1	SAN768B-2_A1	Private_A1	112	1/22, 8/3
	SAN768B-2_B1	Private_B1		1/22, 8/3
Fabric-Private-2	SAN768B-2_A2	Private_A2	114	1/22, 8/3
	SAN768B-2_B2	Private_B2		1/22, 8/3

1/12 corresponds to one of the FCIP tunnels on 1/xge0, whereas 1/22 corresponds to one of the FCIP tunnels on 1/xge1.

Two examples of creating a logical switch by using the CLI and IBM Network Advisor follow. For more information about creating virtual fabrics, see *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116.

Example 4-24 shows creating the Public_A1 on SAN768B-2_A1 logical switch.

Example 4-24 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 1 -port 12  
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> lscfg --config 111 -slot 8 -port 2  
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> lscfg --config 111 -slot 8 -port 4-10  
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.
```

```
switch_111:FID111:admin>  
switch_111:FID111:admin> setcontext 128  
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.

```
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.

```
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
12	1	12	01fcc0	--	--	Offline	VE Disabled
194	8	2	01cf40	id	N16	In_Sync	FC Disabled
196	8	4	01cec0	id	N16	In_Sync	FC Disabled
197	8	5	01ce80	id	N16	In_Sync	FC Disabled
198	8	6	01ce40	id	N16	In_Sync	FC Disabled
199	8	7	01ce00	id	N16	In_Sync	FC Disabled
200	8	8	01cdc0	id	N16	In_Sync	FC Disabled
201	8	9	01cd80	id	N16	In_Sync	FC Disabled
202	8	10	01cd40	id	N16	In_Sync	FC Disabled

```
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
12	1	12	01fcc0	--	--	Offline	VE
194	8	2	01cf40	id	N8	Online	FC F-Port
			50:05:07:68:01:40:b1:3f				
196	8	4	01cec0	id	N8	Online	FC F-Port
			50:05:07:68:02:10:00:ef				
197	8	5	01ce80	id	N8	Online	FC F-Port
			50:05:07:68:02:20:00:ef				
198	8	6	01ce40	id	N8	Online	FC F-Port
			50:05:07:68:02:10:00:f0				
199	8	7	01ce00	id	N8	Online	FC F-Port
			50:05:07:68:02:20:00:f0				
200	8	8	01cdc0	id	N8	Online	FC F-Port
			50:05:07:68:02:10:05:a8				
201	8	9	01cd80	id	N8	Online	FC F-Port
			50:05:07:68:02:10:05:a9				
202	8	10	010000	id	N16	Online	FC F-Port
			10:00:8c:7c:ff:0a:d7:00				

```
Public_A1:FID111:admin>
```

The next example shows creating the Public_B1 on SAN768B-2_B1 logical switch by using the IBM Network Advisor Software management tool.

First, find the Chassis Group that represents the discovered physical switches for SAN768B-2_B1. Right-click the switch, and select **Configuration** → **Logical Switches**, as shown in Figure 4-23.

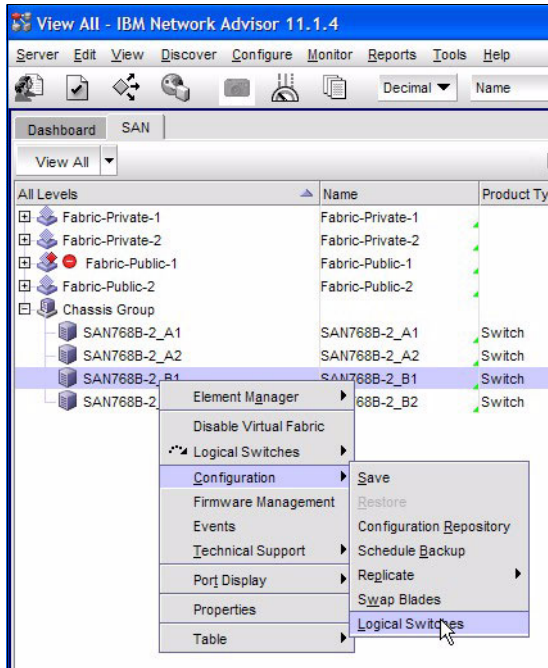


Figure 4-23 Entering the Logical Switches configuration menu

Clarification: In Figure 4-23, most of the logical switches are already created and the fabrics discovered. However, Fabric-Public-1 was re-created as an example. Logical switch Public_A1 in Data Center A was already created. It is a single-switch fabric that is named Fabric-Public-1A.

In the Logical Switches window, make sure that the chassis that you selected is SAN768B-2_B1. Then select **Undiscovered Logical Switches** in the right window, and click **New Switch**, as shown in Figure 4-24.

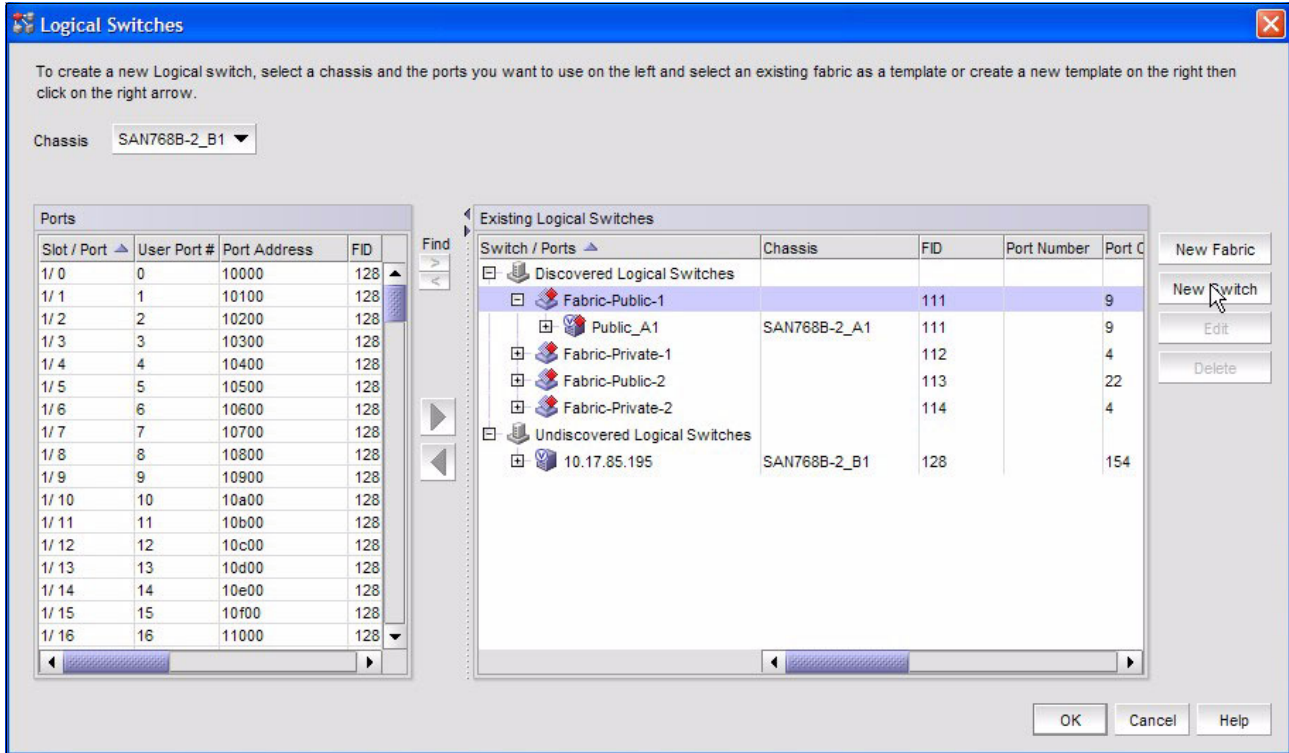


Figure 4-24 Creating a new logical switch on SAN768B-2_B1

In the New Logical Switch window, set the **Logical Fabric ID** to 111. This configuration is chassis-local. It does not have to be the same as the Public_A1 switch that it eventually merges with. However, set it the same for consistency. This process is shown in Figure 4-25.

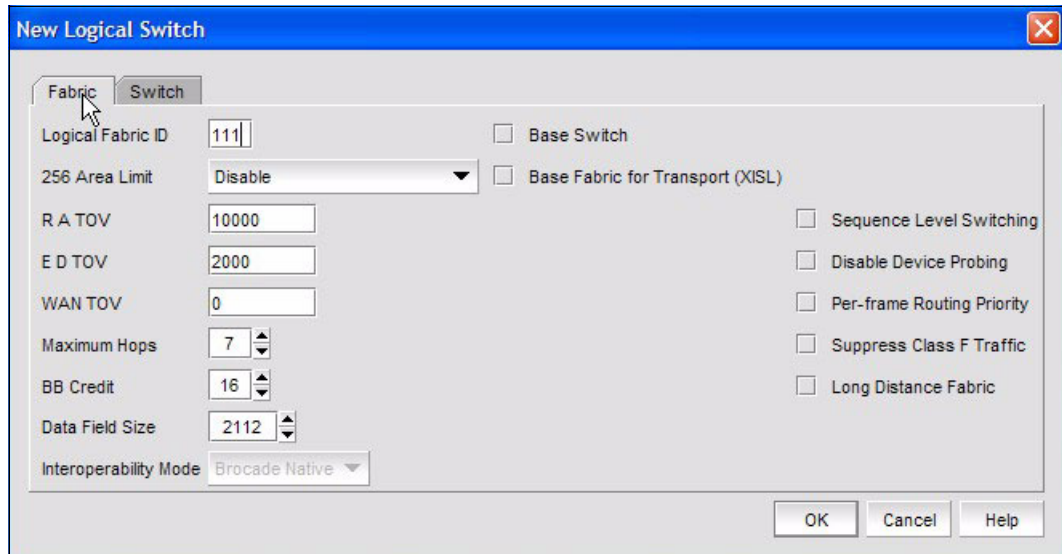


Figure 4-25 Setting the Logical Fabric ID

Next, click the Switch tab and provide a switch name of Public_B1, as shown in Figure 4-26.

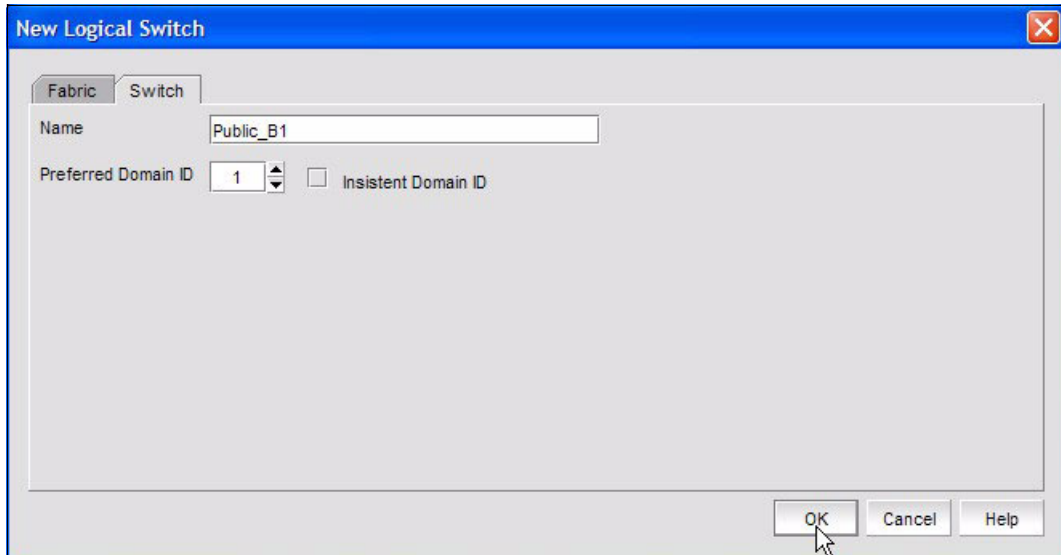


Figure 4-26 Setting the switch name for Public_B1

Now that the new logical switch construct is created, move the ports from **SAN768B-2_B1** to **Public_B1**, as shown in Figure 4-27.

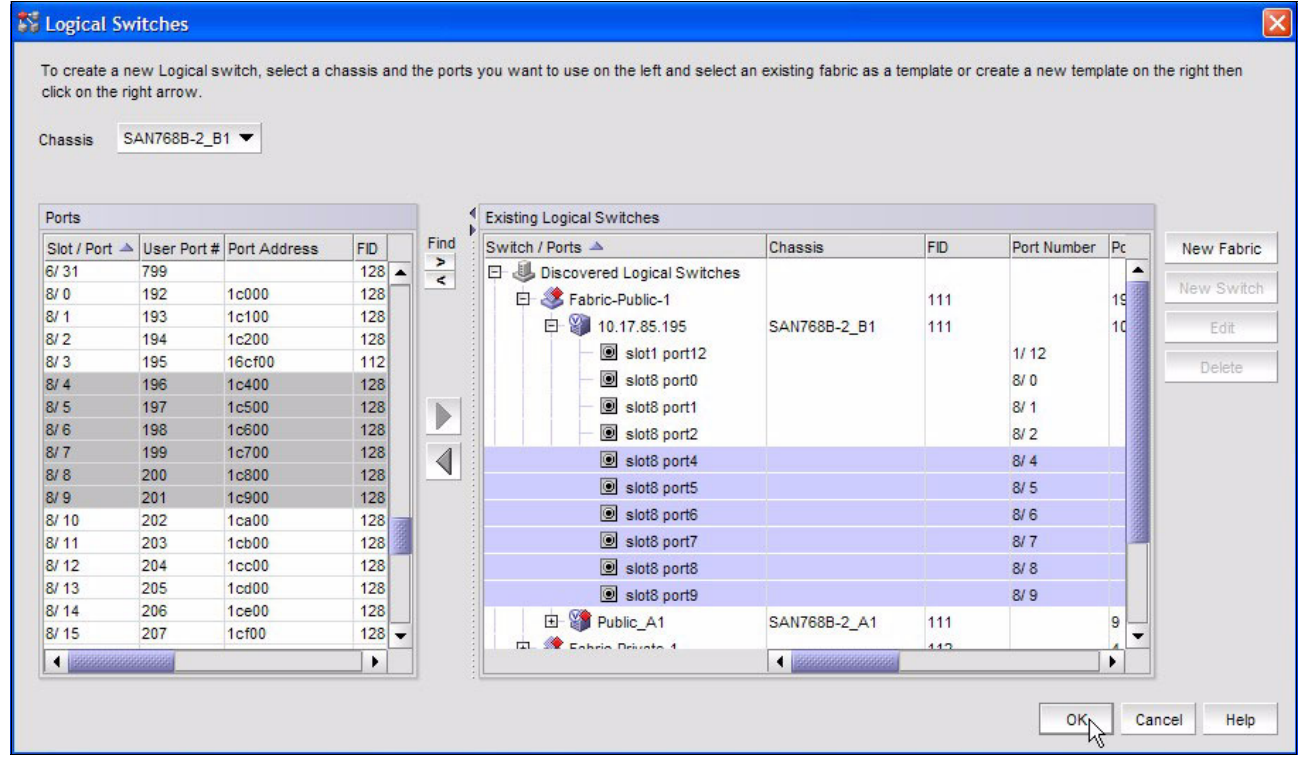


Figure 4-27 Selecting the ports from SAN768B-2_B1 to move to logical switch Public_B1

Review the information in the confirmation window, and then click **Start** to begin the process. Figure 4-28 shows the message that confirms the creation.

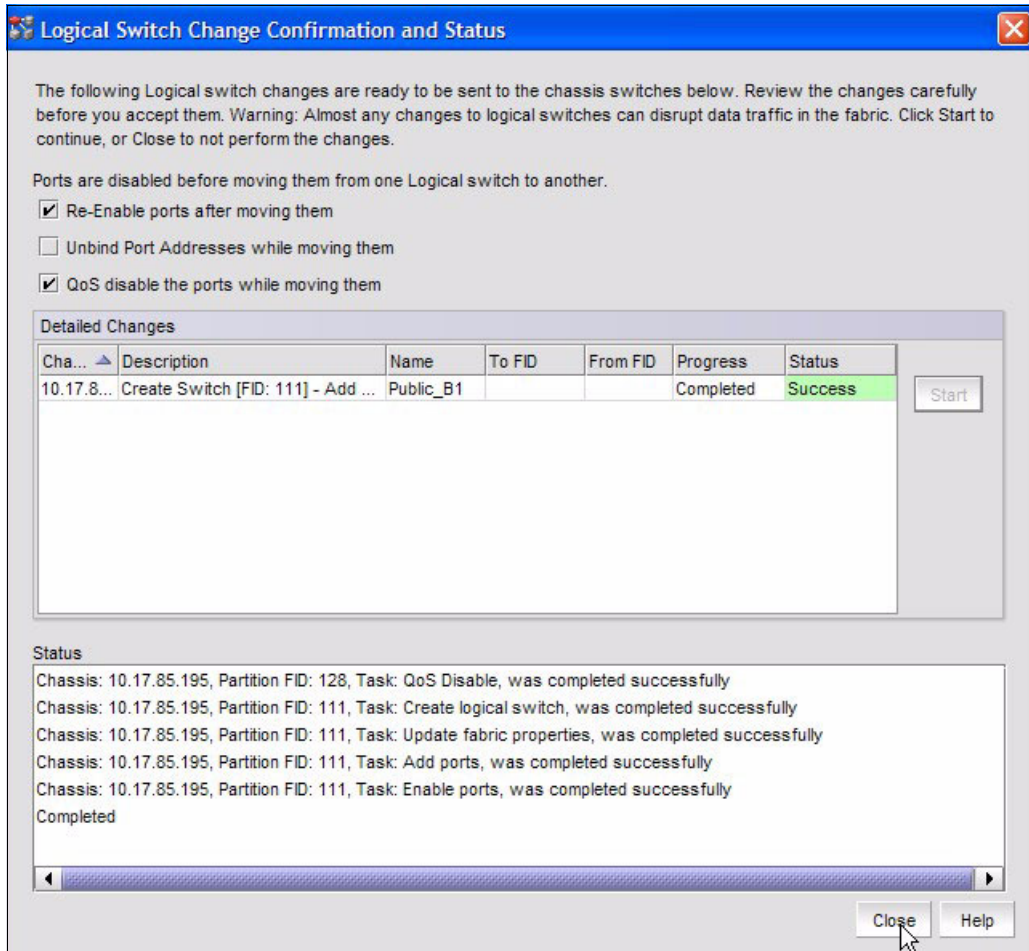


Figure 4-28 Successfully creating Public_B1

The new switch has not been discovered yet. Go to the **Discovery** menu and start a discovery on **SAN768B-2_B1** to add the new logical switch as a monitored switch in the fabric. The IP address for **SAN768B-2_B1** is a previously discovered IP address, so select it as shown in Figure 4-29.

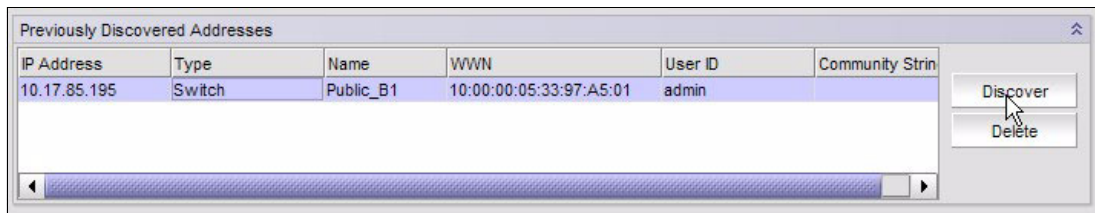


Figure 4-29 Rediscovering SAN768B-2_B1 to add Public_B1

In the Fabric Discovery window, name the new fabric Fabric-Public-1B, for now, as shown in Figure 4-30.

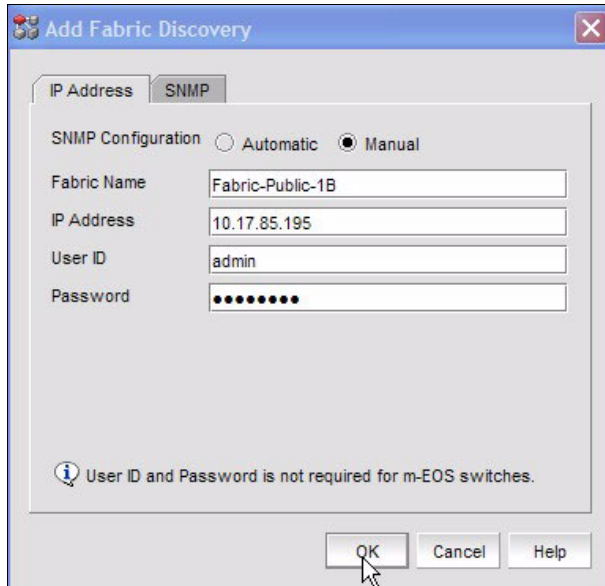


Figure 4-30 Discovering Fabric-Public-1B

Choose to discover and monitor only Public_B1, not the base SAN768B-2_B1 switch, as shown in Figure 4-31.

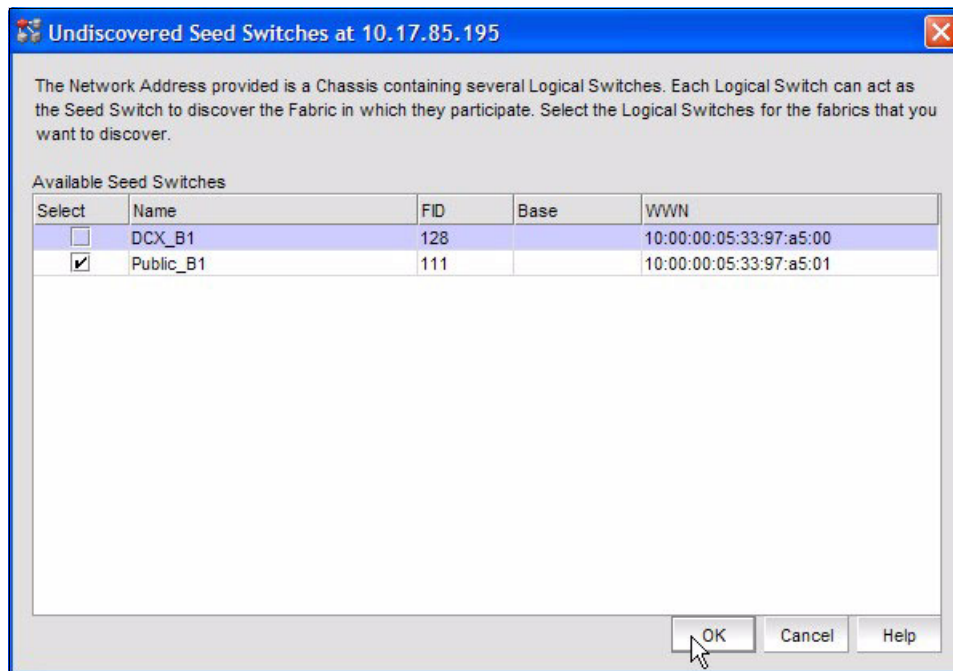


Figure 4-31 Selecting to monitor only Public_B1

Fabric-Public-1B with Public_B1 is now visible in the SAN view, as shown in Figure 4-32.

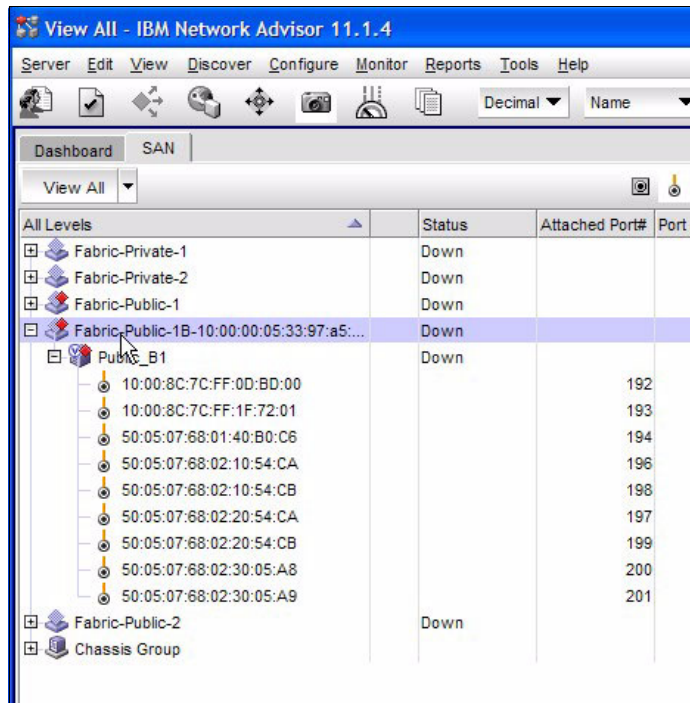


Figure 4-32 Fabric-Public-1B successfully discovered

4.4.2 Creating FCIP tunnels

This section explains how to create the FCIP tunnel connectivity through the command-line interface (CLI). For more information, see *IBM System Storage b-type Multiprotocol Routing: An Introduction and Implementation*, SG24-7544

An FCIP *tunnel* or FCIP *trunk* is a single logical ISL, or VE_port. On the IBM FC 8 Gbps FCIP Extension blade, an FCIP tunnel can have one or more circuits. A circuit is an FCIP connection between two unique IP addresses.

Complete these steps to create an FCIP tunnel manually:

1. Create an IP interface on the physical Ethernet port. This process is done in the default switch context.
2. Validate connectivity between the SAN768B-2 chassis IP interfaces by using **ping**.
3. Within the logical switch that the VE_port belongs to, create an FCIP tunnel from the IP interface that was created in the default switch context.

First, create the IP interface on Public_A1 by using interface 1/xge1, which VE_port 1/12 can use. Create an IP interface with an IP address of 192.168.76.10, netmask 255.255.255.0, with an MTU of 1500 bytes as shown in Example 4-25.

Example 4-25 Creating the IP interface on Public_A1

```
SAN768B-2_A1:FID128:admin> portcfg ipif 1/xge1 create 192.168.76.10 255.255.255.0 1500
```

Operation Succeeded

Create a similar interface but with an IP address of 192.168.76.20/24 on Public_B1, as shown in Example 4-26.

Example 4-26 Creating the IP interface on Public_B1

```
SAN768B-2_B1:FID128:admin> portcfg ipif 1/xge1 create 192.168.76.20 255.255.255.0
1500
Operation Succeeded
```

Run validation tests from Public_A1 to make sure that connectivity through these interfaces works as shown in Example 4-27.

Example 4-27 Validating IP connectivity between Public_A1 and Public_B1

```
SAN768B-2_A1:FID128:admin> portcmd --ping 1/xge1 -s 192.168.76.10 -d 192.168.76.20
Pinging 192.168.76.20 from ip interface 192.168.76.10 on 1/xge1 with 64 bytes of
data
Reply from 192.168.76.20: bytes=64 rtt=2ms ttl=20
Reply from 192.168.76.20: bytes=64 rtt=0ms ttl=20
Reply from 192.168.76.20: bytes=64 rtt=0ms ttl=20
Reply from 192.168.76.20: bytes=64 rtt=0ms ttl=20

Ping Statistics for 192.168.76.20:
    Packets: Sent = 4, Received = 4, Loss = 0 ( 0 percent loss)
    Min RTT = 0ms, Max RTT = 2ms Average = 0ms

SAN768B-2_A1:FID128:admin> portcmd --tracert 1/xge1 -s 192.168.76.10 -d
192.168.76.20
Tracert to 192.168.76.20 from IP interface 192.168.76.10 on 1/xge1, 30 hops max
 1 192.168.76.20 0 ms 0 ms 0 ms
Tracert complete.
```

Now that basic IP connectivity is established, change your logical switch context to 111, where VE_port 1/12 is, to create the FCIP tunnel. Set a minimum bandwidth of 622 Kbps and turn on compression with standard settings, as shown in Example 4-28.

Tip: FastWrite is not needed because IBM Spectrum Virtualize uses a different algorithm to improve transfer over distance. IPSec is also supported and can be turned on if needed.

Example 4-28 Creating the FCIP tunnel on Public_A1

```
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.
```

```
Public_A1:FID111:admin> portcfg fciptunnel 1/12 create 192.168.76.20 192.168.76.10  
-b 622000 -B 1000000 -c 1  
Operation Succeeded
```

```
Public_A1:FID111:admin> portcfgshow fciptunnel 1/12
```

```
-----  
Tunnel ID: 1/12  
Tunnel Description:  
Compression: On (Standard)  
Fastwrite: Off  
Tape Acceleration: Off  
TPerf Option: Off  
IPSec: Disabled  
QoS Percentages: High 50%, Med 30%, Low 20%  
Remote WWN: Not Configured  
Local WWN: 10:00:00:05:33:b5:3e:01  
Flags: 0x00000000  
FICON: Off
```

```
Public_A1:FID111:admin> portcfgshow fcipcircuit 1/12
```

```
-----  
Circuit ID: 1/12.0  
Circuit Num: 0  
Admin Status: Enabled  
Connection Type: Default  
Remote IP: 192.168.76.20  
Local IP: 192.168.76.10  
Metric: 0  
Min Comm Rt: 622000  
Max Comm Rt: 1000000  
SACK: On  
Min Retrans Time: 100  
Max Retransmits: 8  
Keepalive Timeout: 10000  
Path MTU Disc: 0  
VLAN ID: (Not Configured)  
L2CoS: (VLAN Not Configured)  
DSCP: F: 0 H: 0 M: 0 L: 0
```

```
Flags: 0x00000000
Public_A1:FID111:admin>
```

Finally, configure an FCIP tunnel on Public_B1 with the same settings, and verify that the tunnel was established, as seen in Example 4-29.

Example 4-29 Creating an FCIP tunnel on Public_B1 and verifying that it is established

```
SAN768B-2_B1: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.

Public_B1:FID111:admin> portcfg fcip tunnel 1/12 create 192.168.76.10 192.168.76.20
-b 622000 -B 1000000 -c 1
Operation Succeeded

Public_B1:FID111:admin> portshow fcipcircuit all
-----
Tunnel Circuit OpStatus Flags      Uptime  TxMBps  RxMBps ConnCnt CommRt  Met
-----
1/12   0 1/xge1  Up      ---4--s  1m27s   0.00   0.00   1    622/1000  0
-----
Flags: circuit: s=sack v=VLAN Tagged x=crossport 4=IPv4 6=IPv6
        L=Listener I=Initiator

Public_B1:FID111:admin>
```

4.5 Spectrum Virtualize with an Enhanced Stretched Cluster

The IBM Spectrum Virtualize code that is used is based on version 7.8.0.0, and the back-end storage that is used in the example is a Storwize V7000 running version 7.6.1.5. The third site, the quorum storage, is on a Storwize V7000 running version 7.6.1.5

Be sure to check the full list of supported extended quorum devices.

For the version 7.8.x Supported Hardware List, Device Driver, Firmware, and Recommended Software Levels for Spectrum Virtualize, see:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009558>

Figure 4-33 shows the components that are used for the Spectrum Virtualize stretched cluster with SAN connectivity.

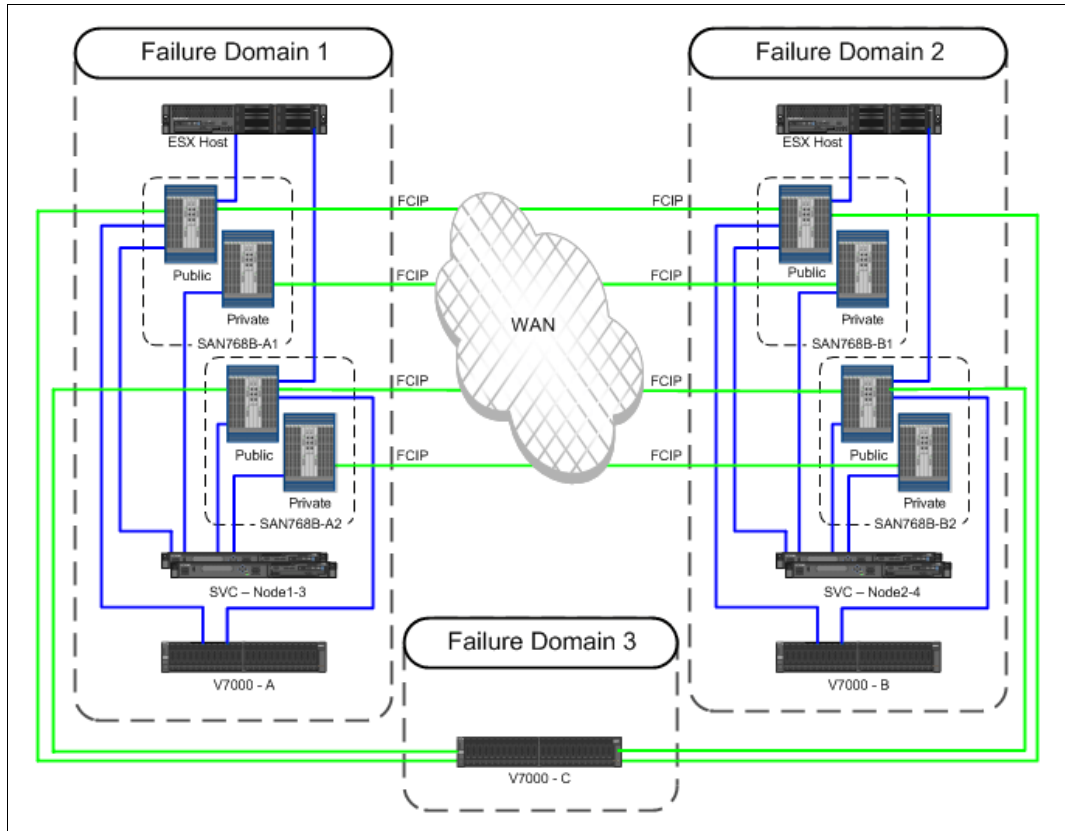


Figure 4-33 Spectrum Virtualize stretched cluster diagram with SAN connectivity

This book does not cover the physical installation nor the initial configuration. It assumes that you are familiar with the major concepts of clusters, such as nodes, I/O groups, MDisks, and quorum disks.

Also, see the IBM SAN Volume Controller Knowledge Center at:

<https://ibm.biz/BdsSSq>

4.6 Volume mirroring

The ESC I/O group uses *volume mirroring* functionality. Volume mirroring allows creation of one volume with two copies of MDisk extents. In this configuration, there are not two volumes with the same data on them. The two data copies can be in different MDisk groups. Therefore, volume mirroring minimizes the effect on volume availability if one or more MDisks fail. The resynchronization between both copies is incremental and Spectrum Virtualize starts the resynchronization process automatically.

A mirrored volume has the same functions and behavior as a standard volume. In the Spectrum Virtualize software stack, volume mirroring is below the copy services. Therefore, FlashCopy, Metro Mirror, and Global Mirror have no awareness that a volume is mirrored. Everything that can be done with a volume can also be done with a mirrored volume, including migration and expand or shrink. Like a standard volume, each mirrored volume is owned by one I/O group with a preferred node, so the mirrored volume goes offline if the

whole I/O group goes offline. The preferred node coordinates all I/O operations, reads, and writes. The preferred node can be set manually.

The three quorum disk candidates keep the status of the mirrored volume. The last status and the definition of primary and secondary volume copy (for read operations) are saved there. This means that an active quorum disk is required for volume mirroring. To ensure data consistency, Spectrum Virtualize disables mirrored volumes if there is no access to any quorum disk candidate. Therefore, quorum disk availability is an important point with volume mirroring and split I/O group configuration. Furthermore, you must allocate bitmap memory space before using volume mirroring. Use the **chiogrp** command:

```
chiogrp -feature mirror -size memory_size io_group_name|io_group_id
```

The volume mirroring grain size is fixed at 256 KB, so one bit of the synchronization bitmap represents 256 KB of virtual capacity. Therefore, a bitmap memory space of 1 MB is required for each 2 TB of mirrored volume capacity.

4.7 Read operations

Volume mirroring implements a read algorithm with one copy that is designated as the primary for all read operations. Spectrum Virtualize reads the data from the primary copy and does not automatically distribute the read requests across both copies. The first copy that is created becomes the primary by default. You can change this setting by using the **chvdisk** command:

```
chvdisk -primary copyid vdiskname
```

Starting with software version 7.2 and the introduction of the site awareness concept in the ESC configurations, read operations can be performed from either copy of the volume. Which is used depends on the site affinity between nodes running the IO operation and storage controllers. Therefore, all read operations run to the local site copy if both sites are in sync.

4.8 Write operations

Write operations are run on all copies. The storage system with the lowest performance determines the response time between Spectrum Virtualize and the storage system back-end. The cache can hide this process from the server, up to a certain level.

If a back-end write fails or a copy goes offline, a bitmap is used to track out-of-sync grains, as with other Spectrum Virtualize copy services. As soon as the missing copy is back, the controller evaluates the changed bitmap file to run an automatic resynchronization of both copies.

The resynchronization process has a similar performance impact on the system as a FlashCopy background copy or a volume migration does. The resynchronization bandwidth can be controlled with the **chvolume -syncrate** command. Host access to the volume continues during that time. This behavior can cause difficulties if there is a site failure. Since software version 6.2, it provides the **-mirrorwritepriority** volume attribute to establish priorities between strict data redundancy (**-mirrorwritepriority redundancy**) and the best performance for the volume (**-mirrorwritepriority latency**). The suggested setting when you have an ESC with the same kind of storage controller is **-mirrorwritepriority latency**.

Starting with software version 7.2 and the introduction of the site awareness concept in the ESC configurations, the write destage operation for each copy is performed by the nodes according to the site attributes of the storage controller.

4.9 Quorum disk

The quorum disk fulfills two functions for cluster reliability:

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

The quorum algorithm distinguishes between the active quorum disk and quorum disk candidates. There are three quorum disk candidates. At any time, only one of these candidates is acting as the active quorum disk. The other two 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 tiebreaker for “split brain” scenarios. Starting with version 7.6, the active quorum disk can be replaced with the IP quorum device, as described in 4.10, “IP Quorum” on page 100. In this section, standard disk quorum considerations are discussed.

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.

If the *Stretched* topology is not used, the quorum selection algorithm operates as it did with software version 7.1 and previous versions.

When the Stretched topology is enabled and automatic quorum disk selection is also enabled, three quorum disks are created, one in each site (Sites 1, 2, and 3).

If a site has no suitable MDisk, fewer than three quorum disks are automatically created. For example, if it can create only two quorum disks, only two are used.

If you are controlling the quorum by using the **chquorum** command, the choice of quorum disks must also follow the one-disk-per-site rule.

If you use the **chquorum** command to manually assign quorum disks and configure the topology as stretched, it ignores any quorum disk that is not assigned to a site. Spectrum Virtualize chooses only quorum disks that are configured to Site 3 as the active quorum disk. It chooses quorum disks that are configured to Site 1 or 2 as stand-by quorum disks.

If you do not have a quorum disk configured at each site, that might restrict when, or if, T3 recovery procedure is possible and how resilient the cluster is after site failures. Without access to a quorum disk, it cannot continue I/O when one copy of a mirrored volume goes offline.

Note: For clusters implemented with the Stretched topology, manually configure quorum devices to track which MDisk is chosen and to select the MDisk that you want to be your quorum disks.

4.9.1 Quorum disk requirements and placement

Because of the quorum disk's role in the voting process, the quorum function is not supported for internal drives on Spectrum Virtualize nodes. Inside a Spectrum Virtualize node, the quorum disk cannot act as a tiebreaker. Therefore, only managed disks (MDisks) from an external storage system are selected as quorum disk candidates. Distribution of quorum disk candidates across storage systems in different failure domains eliminates the risk of losing all three quorum disk candidates because of an outage of a single storage system or site.

Up to software version 6.1 it selects the first three Managed Disks (MDisks) from external storage systems as quorum disk candidates. It reserves some space on each of these disks per default. Spectrum Virtualize does not verify whether the MDisks are from the same disk controller or from different disk controllers. To ensure that the quorum disk candidates and the active quorum disk are in the correct sites, change the quorum disk candidates by using the `chquorum` command.

Starting with software version 6.2, and still true with Spectrum Virtualize 7.8, the quorum disk selection algorithm has changed. Spectrum Virtualize reserves space on each MDisk, and dynamically selects the quorum disk candidates and the active quorum disk. Thus the location of the quorum disk candidates and the active quorum disk might change unexpectedly. Therefore, ensure that you disable the dynamic quorum selection in a split I/O group cluster by using the `-override` flag for all three quorum disk candidates:

```
chquorum -override yes -mdisk mdisk_id|mdisk_name
```

The storage system that provides the quorum disk in a split I/O group configuration at the third site must be supported as an extended quorum disk. Storage systems that provide extended quorum support are listed at:

<https://ibm.biz/BdsvdV>

4.9.2 Automatic quorum disk selection

The CLI output in Example 4-30 shows that the cluster initially automatically assigns the quorum disks.

Example 4-30 Quorum disks assigned.

```
IBM_2145:ITS0_SVC_ESC:superuser>lsquorum
quorum_index status id name controller_id controller_name active
object_type override
0 online 1 ITS0_V7K_SITEA_SAS0 1 ITS0_V7K_SITEA_N2 yes
mdisk no
1 online 2 ITS0_V7K_SITEA_SAS1 1 ITS0_V7K_SITEA_N2 no
mdisk no
2 online 3 ITS0_V7K_SITEA_SAS2 1 ITS0_V7K_SITEA_N2 no
mdisk no
```

To change from automatic selection to manual selection, run the commands shown in Example 4-31.

Example 4-31 Changing from automatic to manual selection

```
IBM_2145:ITS0_SVC_ESC:superuser>chquorum -override yes -mdisk 1 0
IBM_2145:ITS0_SVC_ESC:superuser>chquorum -override yes -mdisk 11 1
IBM_2145:ITS0_SVC_ESC:superuser>chquorum -override yes -mdisk 6 2
```

After that process is complete, when you run the `lssquorum` command, you get output as shown in Example 4-32.

Example 4-32 Quorum changed

```
IBM_2145:ITS0_SVC_ESC:superuser>lssquorum
quorum_index status id name controller_id controller_name
active object_type override
0 online 1 ITS0_V7K_SITEA_SAS0 1 ITS0_V7K_SITEA_N2 no
mdisk yes
1 online 11 ITS0_V7K_SITEC_QUORUM 2 ITS0_V7K_SITEC_Q_N1
yes mdisk yes
2 online 6 ITS0_V7K_SITEB_SAS0 0 ITS0_V7K_SITEB_N2 no
mdisk yes
```

The output shows that the controller named `ITS0_V7K_SITEC_QUORUM`, which is in Power Domain 3, Site 3, is now the active quorum disk.

You can assign the quorum disks manually from the GUI as well. From the GUI, click **Pools** → **Pools by MDisk** as shown in Figure 4-34.



Figure 4-34 Opening the MDisk by using the Pools view

You might need to expand the pools to view all of the MDisks. Select the MDisks that you want to use for quorum disks by holding down Ctrl and selecting the three candidates. When the candidates are selected, right-click them and select **Quorum** → **Modify Quorum Disks**, as shown in Figure 4-35.

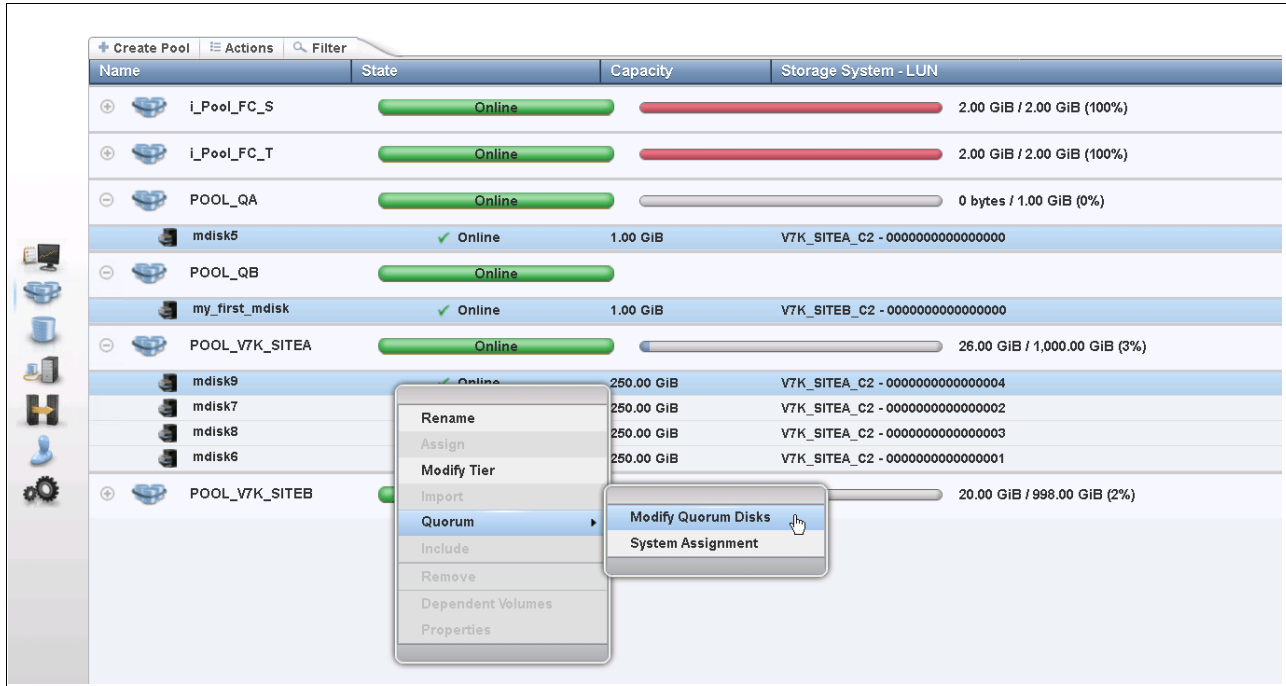


Figure 4-35 Blue color shows the MDisks that are selected for quorum

The Modify Quorum panel opens, as shown in Figure 4-36. Optionally select if you want to use the selected quorum disks even if degraded.

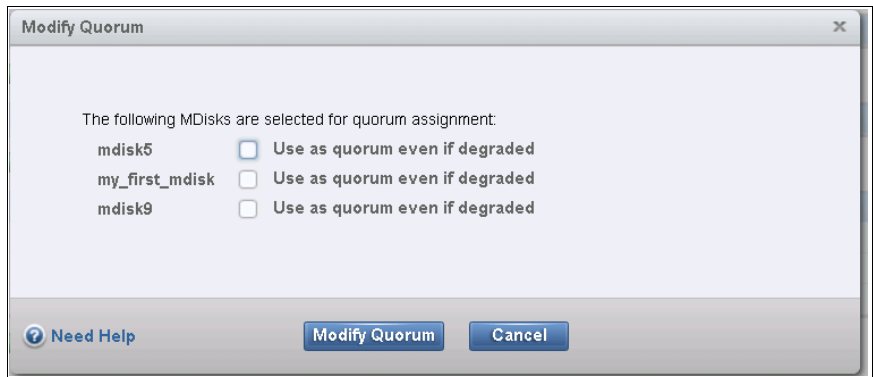


Figure 4-36 Modify Quorum confirmation panel

Finally, click **Modify Quorum** to manually assign the quorum disks in the GUI.

4.10 IP Quorum

Spectrum Virtualize version 7.6 introduced the IP Quorum feature that eliminates the requirement for Fibre Channel networking and disk storage at a 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 standard server and needs only standard network connectivity to be exploited. See Chapter 3, “Enhanced Stretched Cluster architecture” on page 27 for further details on IP quorum requirements.

To implement the IP Quorum function, the following actions 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 4-33.

Example 4-33 The mkquorumapp command

```
IBM_2145:SVC_ESC:superuser>mkquorumapp
IBM_2145:SVC_ESC: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 4-34:

Example 4-34 The lsdumps command

```
IBM_2145:SVC_ESC:superuser>lsdumps
id filename
0  reinst..trc
1  sel.000000.trc
2  ec_makevpd.000000.trc
.
multiple lines omitted
.
62 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 Example 4-35:

Example 4-35 pscp command to download the quorum app

```
pscp -unsafe -load SVC_ESC admin@SVC_ip:/dumps/ip_quorum.jar local_directory
```

If you prefer to use the GUI, click **Settings** → **Support** page, as shown in Figure 4-37. If the page is not displaying a list of individual log files, click **Show full log listing**.

(Figure 4-37 on page 101)

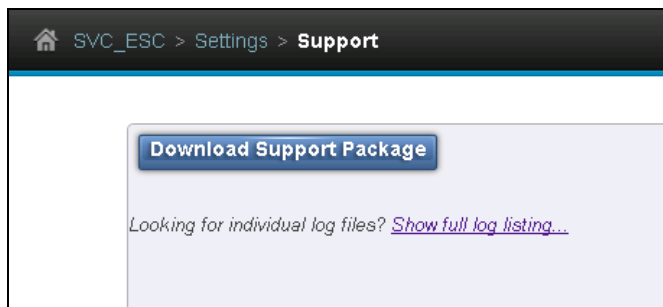


Figure 4-37 Download Support Package window

Next, right-click the row for the `ip_quorum.jar` file and choose **Download**, as shown in Figure 4-38.

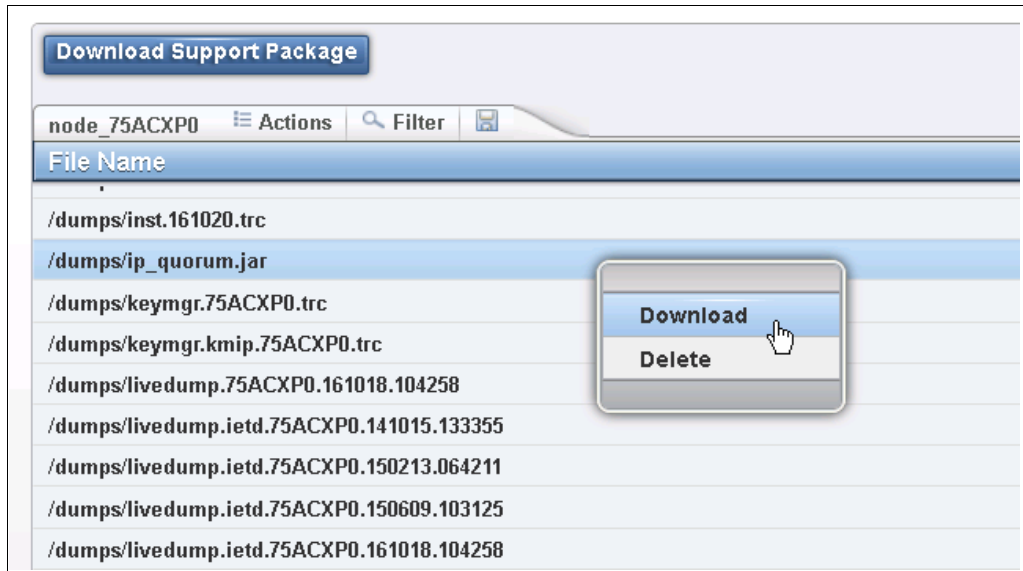


Figure 4-38 Quorum application download

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 4-39.



Figure 4-39 Select the IP Quorum

GUI support: GUI support for the IP quorum has been introduced with Spectrum Virtualize 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 4-40.

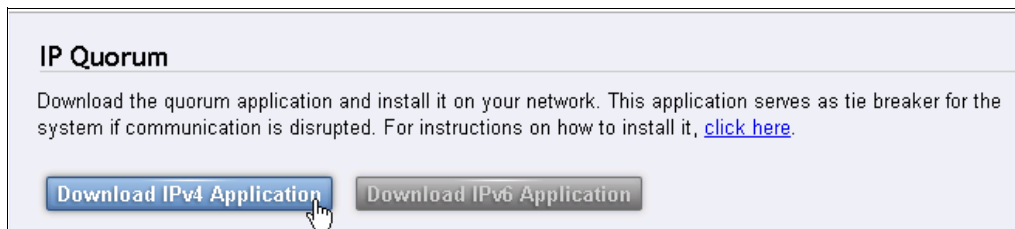


Figure 4-40 IP quorum download section

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 IP connectivity using the `ping` command from the host to the service address of each node of the Spectrum Virtualize 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 4-36.

Example 4-36 The IP quorum files

```
[root@oc2244547338 IPQ]# ls -la ip*
-rw-r--r--. 1 root root 51433 Oct 24 12:40 ip_quorum.jar
-rw-r--r--. 1 root root 2043 Oct 24 12:39 ip_quorum.log
-rw-r--r--. 1 root root 0 Oct 24 12:39 ip_quorum.log.lck
```

The `ip_quorum.log` contains the application initialization and the heartbeat 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 4-37.

Example 4-37 `ip_quorum.log` content

```
2016-10-24 12:34:25:583 Quorum CONFIG: === IP quorum ===
2016-10-24 12:34:25:591 Quorum CONFIG: Name set to null.
2016-10-24 12:34:28:846 Quorum FINE: Node 10.18.228.72:1,260 (0, 0)
2016-10-24 12:34:28:862 Quorum FINE: Node 10.18.228.73:1,260 (0, 0)
2016-10-24 12:34:28:862 Quorum CONFIG: Successfully parsed the configuration,
found 2 nodes.
2016-10-24 12:34:28:863 10.18.228.72 [9] INFO: Trying to open socket
2016-10-24 12:34:28:864 10.18.228.73 [10] INFO: Trying to open socket
2016-10-24 12:34:29:430 10.18.228.72 [9] INFO: Creating UID
2016-10-24 12:34:29:430 10.18.228.73 [10] INFO: Waiting for UID
2016-10-24 12:34:29:441 10.18.228.72 [9] FINE: <Msg [protocol=1, sequence=0,
command=CREATE_UID_REQUEST, length=0]
```

```

2016-10-24 12:34:29:708 10.18.228.72 [9] FINE: >Msg [protocol=1, sequence=0,
command=CREATE_UID_RESPONSE, length=16] Data [quorumUid=5,
clusterUid=2212453419066]
2016-10-24 12:34:29:709 10.18.228.72 [9] FINE: <Msg [protocol=2, sequence=1,
command=CONNECT_REQUEST, length=88] Data [quorumUid=5,
clusterUid=2212453419066, generationId=0, shortLeaseExtension=true,
infoText=ITS0-1.englab.brocade.com/10.18.228.170]
2016-10-24 12:34:29:710 10.18.228.73 [10] INFO: *Connecting
2016-10-24 12:34:29:711 10.18.228.73 [10] FINE: <Msg [protocol=2, sequence=0,
command=CONNECT_REQUEST, length=88] Data [quorumUid=5,
clusterUid=2212453419066, generationId=0, shortLeaseExtension=true,
infoText=ITS0-1.englab.brocade.com/10.18.228.170]
2016-10-24 12:34:29:909 10.18.228.72 [9] FINE: >Msg [protocol=1, sequence=1,
command=CONNECT_RESPONSE, length=4] Data [result=SUCCESS]
2016-10-24 12:34:29:912 10.18.228.72 [9] INFO: Connected to 10.18.228.72
2016-10-24 12:34:29:965 10.18.228.73 [10] FINE: >Msg [protocol=1, sequence=0,
command=CONNECT_RESPONSE, length=4] Data [result=SUCCESS]
2016-10-24 12:34:29:966 10.18.228.73 [10] INFO: Connected to 10.18.228.73
2016-10-24 12:34:36:369 10.18.228.73 [10] FINE: <Msg [protocol=1, sequence=1,
command=HEARTBEAT_REQUEST, length=0]
2016-10-24 12:34:36:369 10.18.228.72 [9] FINE: <Msg [protocol=1, sequence=2,
command=HEARTBEAT_REQUEST, length=0]
2016-10-24 12:34:36:372 10.18.228.73 [10] FINE: >Msg [protocol=1, sequence=1,
command=HEARTBEAT_RESPONSE, length=0]
2016-10-24 12:34:36:372 10.18.228.72 [9] FINE: >Msg [protocol=1, sequence=2,
command=HEARTBEAT_RESPONSE, length=0]
2016-10-24 12:34:51:367 10.18.228.72 [9] FINE: <Msg [protocol=1, sequence=3,
command=HEARTBEAT_REQUEST, length=0]
2016-10-24 12:34:51:368 10.18.228.73 [10] FINE: <Msg [protocol=1, sequence=2,
command=HEARTBEAT_REQUEST, length=0]
2016-10-24 12:34:51:369 10.18.228.72 [9] FINE: >Msg [protocol=1, sequence=3,
command=HEARTBEAT_RESPONSE, length=0]
2016-10-24 12:34:51:369 10.18.228.73 [10] FINE: >Msg [protocol=1, sequence=2,
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 Spectrum Virtualize cluster. To check the new quorum status through the CLI use the `lsquorum` command, as shown in Example 4-38.

Example 4-38 lsquorum output

```

IBM_2145:SVC_ESC:superuser>lsquorum
quorum_index status id name controller_id controller_name active
object_type override site_id site_name
0 online 3 mdisk3 3 V7K_SITEB_C2 no
mdisk no
1 online 4 mdisk4 3 V7K_SITEB_C2 no
mdisk no
2 online 0 my_first_mdisk 3 V7K_SITEB_C2 no
mdisk no
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 4-41 on page 105.


```
high_temp_mode off
topology standard
topology_status standard
rc_auth_method none
```

Before changing the topology to a stretched cluster, you must assign a site attribute to each node in the Spectrum Virtualize cluster as shown in Example 4-40. Optionally, you can assign a name to each site.

Example 4-40 Assigning site attribute to a node

```
IBM_2145:ITSO_SVC_ESC:superuser>chsite -name SITE_A 1
IBM_2145:ITSO_SVC_ESC:superuser>lssite
id site_name
1 SITE_A
2 SITE_B
3 SITE_C
IBM_2145:ITSO_SVC_ESC:superuser>lsnode
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
5 ITSO_SVC_NODE1_SITE_A 100006B119 500507680100B13F online 0
ITSO_SVC_ESC_0 no 2040000006481049 CF8
iqn.1986-03.com.ibm:2145.itsosvcesc.itsosvcnode1sitea 151580
151580
9 ITSO_SVC_NODE2_SITE_B 100006B074 500507680100B0C6 online 0
ITSO_SVC_ESC_0 yes 20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcesc.itsosvcnode2siteb 151523
151523
8 ITSO_SVC_NODE3_SITE_A 1000849047 50050768010027E2 online 1
ITSO_SVC_ESC_1 no 2040000204240107 8G4
iqn.1986-03.com.ibm:2145.itsosvcesc.itsosvcnode3sitea 108283
108283
11 ITSO_SVC_NODE4_SITE_B 1000871173 50050768010037E5 online 1
ITSO_SVC_ESC_1 no 20400002070411C3 8G4
iqn.1986-03.com.ibm:2145.itsosvcesc.itsosvcnode4siteb 104643
104643

IBM_2145:ITSO_SVC_ESC:superuser>lsnode ITSO_SVC_NODE1_SITE_A
id 5
name ITSO_SVC_NODE1_SITE_A
.
multiple lines omitted
.
site_id
site_name
IBM_2145:ITSO_SVC_ESC:superuser>lssite
id site_name
1 SITE_A
2 SITE_B
3 SITE_C

IBM_2145:ITSO_SVC_ESC:superuser>chnode -site 1 ITSO_SVC_NODE1_SITE_A

IBM_2145:ITSO_SVC_ESC:superuser>lsnode ITSO_SVC_NODE1_SITE_A
id 5
```

```
name ITSO_SVC_NODE1_SITE_A
```

```
.  
multiple lines omitted
```

```
.  
site_id 1  
site_name SITE_A
```

Also, assign a Storage Controller site attribute as shown in Example 4-41.

A storage controller must be available to the Spectrum Virtualize as a controller for each node or internal controller. You must configure a site attribute for each of those controllers.

Example 4-41 Controller site attribute

```
IBM_2145:ITSO_SVC_ESC:superuser>lscontroller  
id controller_name      ctrl_s/n          vendor_id         product_id_low  
product_id_high  
0 ITSO_V7K_SITEB_N2     2076              IBM               2145  
1 ITSO_V7K_SITEA_N2     2076              IBM               2145  
2 ITSO_V7K_SITEC_Q_N1  2076              IBM               2145  
3 ITSO_V7K_SITEB_N1     2076              IBM               2145  
4 ITSO_V7K_SITEA_N1     2076              IBM               2145  
5 ITSO_V7K_SITEC_Q_N2  2076              IBM               2145  
IBM_2145:ITSO_SVC_ESC:superuser>lscontroller ITSO_V7K_SITEC_Q_N1  
id 2  
controller_name ITSO_V7K_SITEC_Q_N1  
WWNN 50050768020005A8  
mdisk_link_count 1  
max_mdisk_link_count 1  
degraded no  
vendor_id IBM  
product_id_low 2145  
product_id_high  
product_revision 0000  
ctrl_s/n 2076  
allow_quorum yes  
fabric_type fc  
site_id  
site_name  
WWPN 50050768022005A8  
path_count 0  
max_path_count 0  
WWPN 50050768023005A8  
path_count 0  
max_path_count 0  
WWPN 50050768024005A8  
path_count 0  
max_path_count 0  
WWPN 50050768021005A8  
path_count 0  
max_path_count 0  
IBM_2145:ITSO_SVC_ESC:superuser>chcontroller -site 3 ITSO_V7K_SITEC_Q_N1  
IBM_2145:ITSO_SVC_ESC:superuser>lscontroller ITSO_V7K_SITEC_Q_N1  
id 2  
controller_name ITSO_V7K_SITEC_Q_N1  
WWNN 50050768020005A8
```


4.11.2 Using the GUI

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

To set up the Enhanced Stretched Cluster configuration, go the **Monitoring** → **System** window, select **Monitoring** → **System Topology**, as shown in Figure 4-42.



Figure 4-42 Modify System Topology

The Modify System Topology wizard opens, as depicted in Figure 4-43 on page 110.

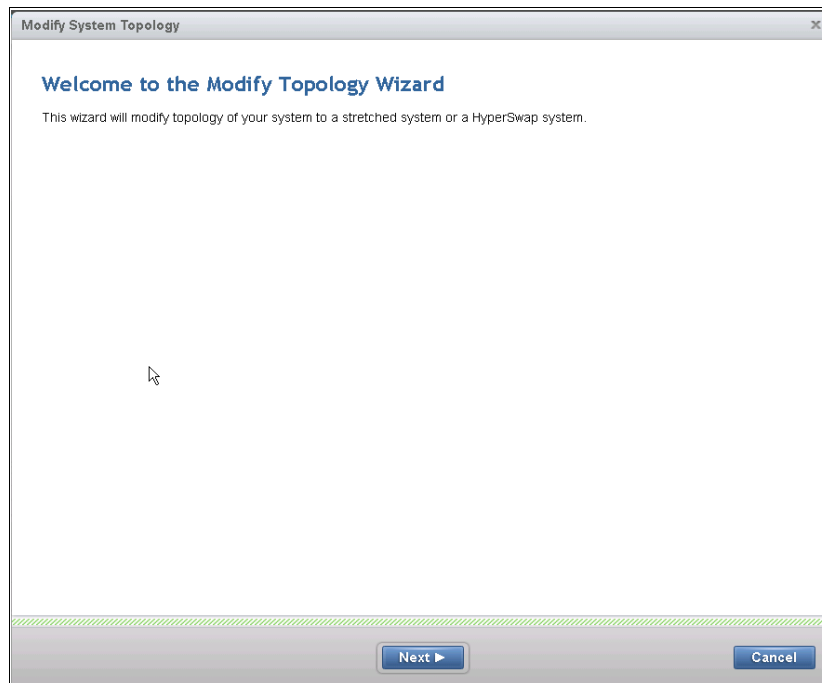


Figure 4-43 The Modify System Topology main windows

Click **Next** to start the configuration. In the sign *Assign Site Names* window you can specify the site names, as shown in Figure 4-44 on page 111.

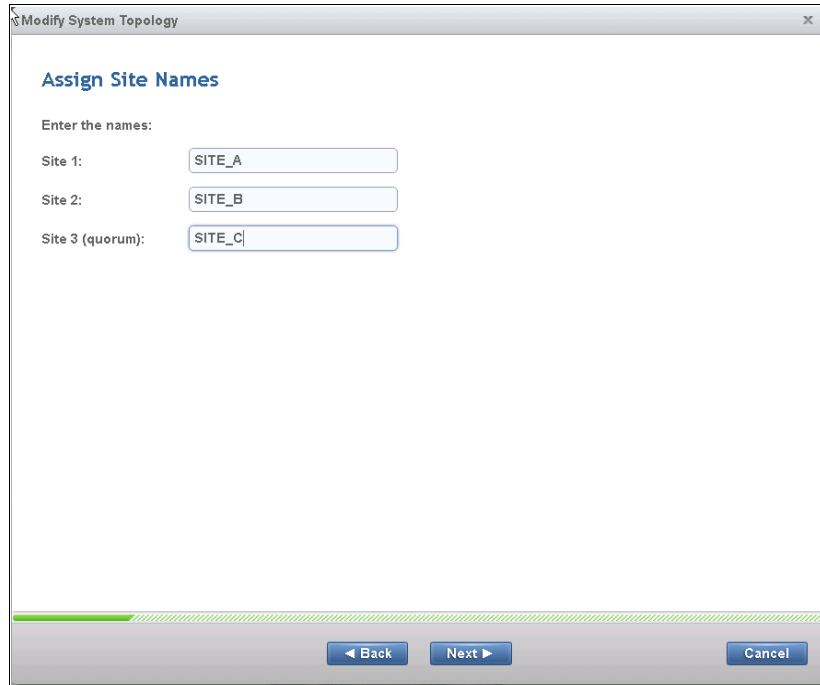


Figure 4-44 The Assign Site Names window

Click **Next** to go to the *Assign Nodes* window. Select the **Stretched System** in the **Topology** drop down menu, as shown in Figure 4-45.

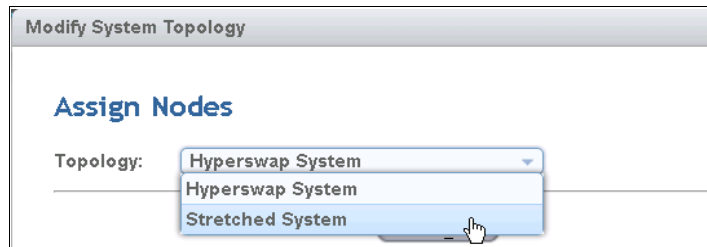


Figure 4-45 Assign the system topology

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 4-46 on page 112.

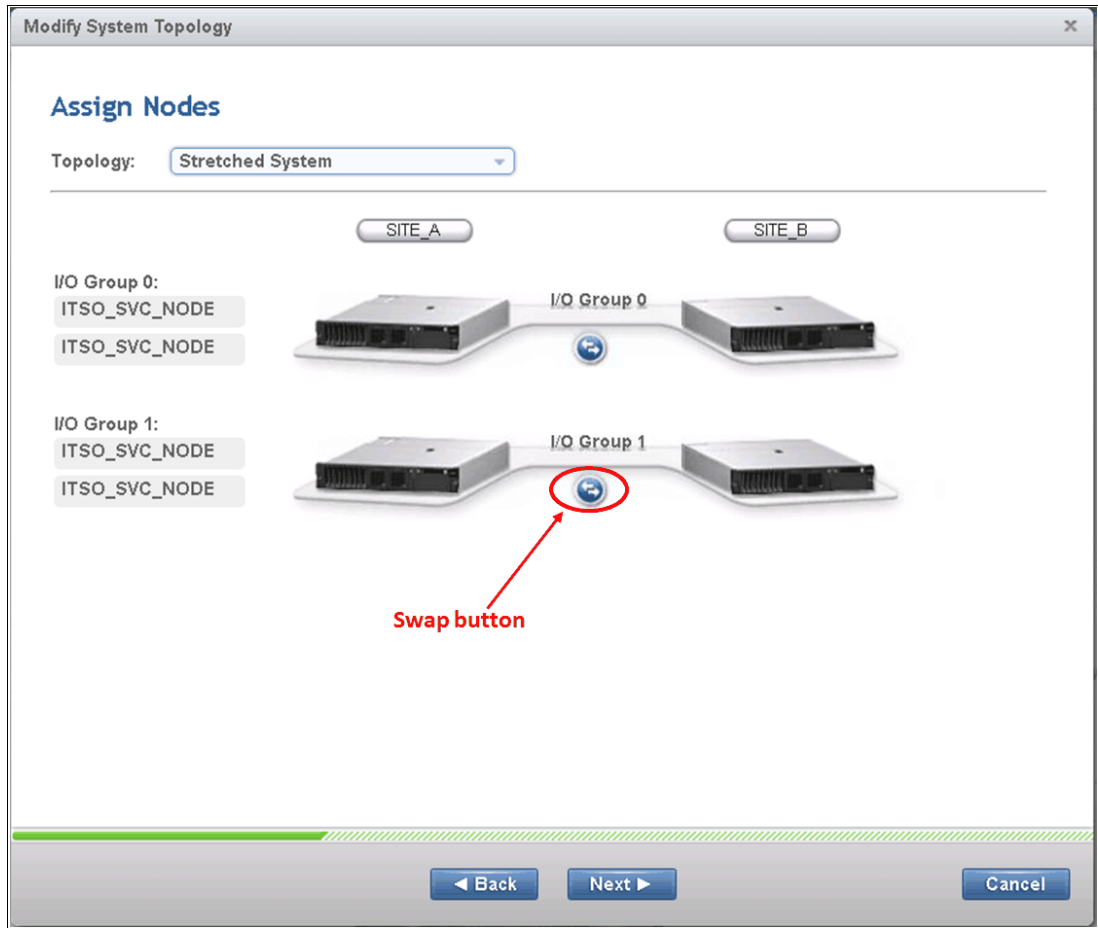


Figure 4-46 Assign Nodes window

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 4-47 on page 113.

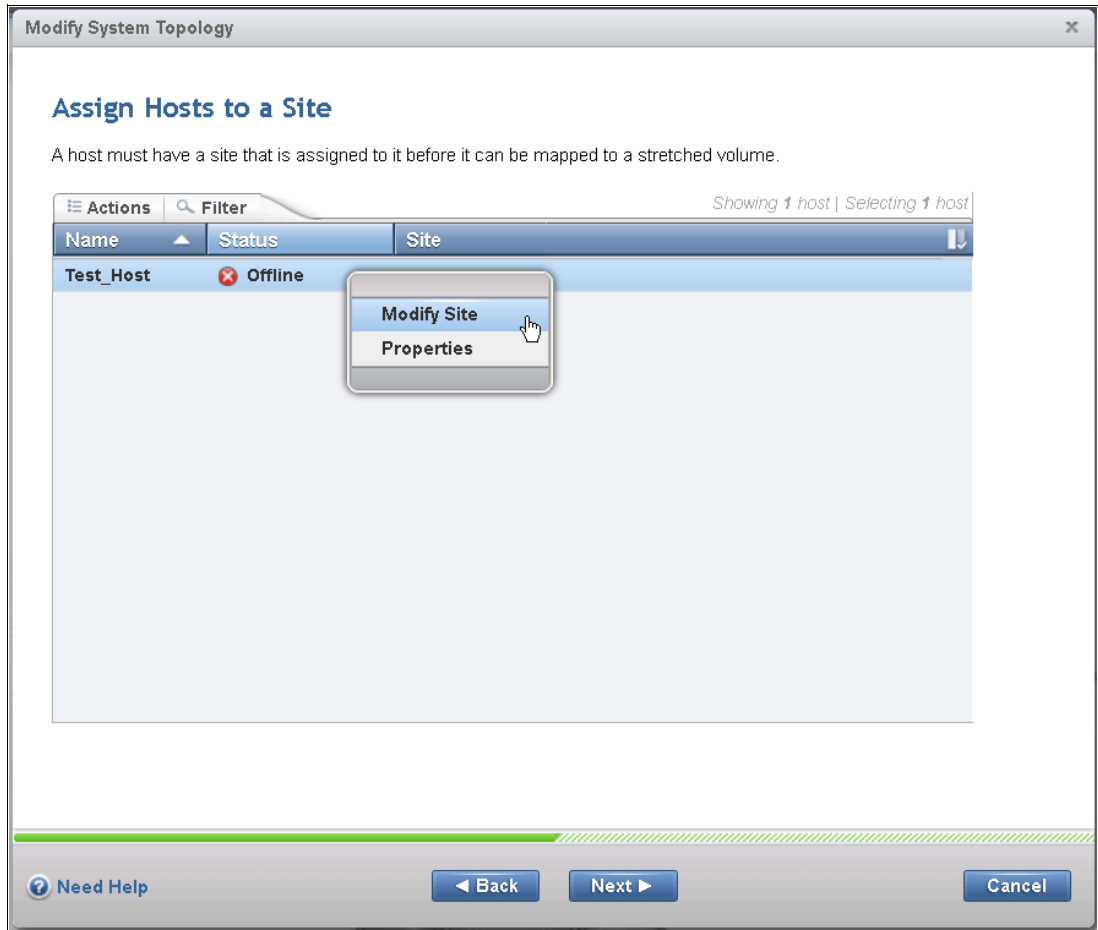


Figure 4-47 The Assign Hosts to a Site window

The site selection window opens. Select the host site from the drop down menu, as shown in Figure 4-48.

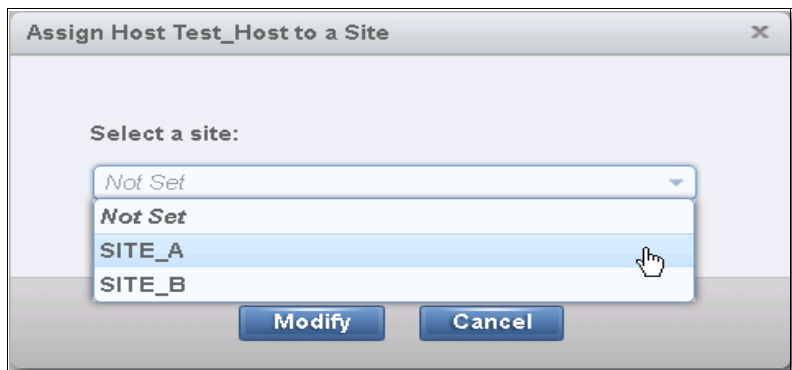


Figure 4-48 Site selection menu for hosts

When the site settings for all the hosts is completed, click **Next**. 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 4-49.

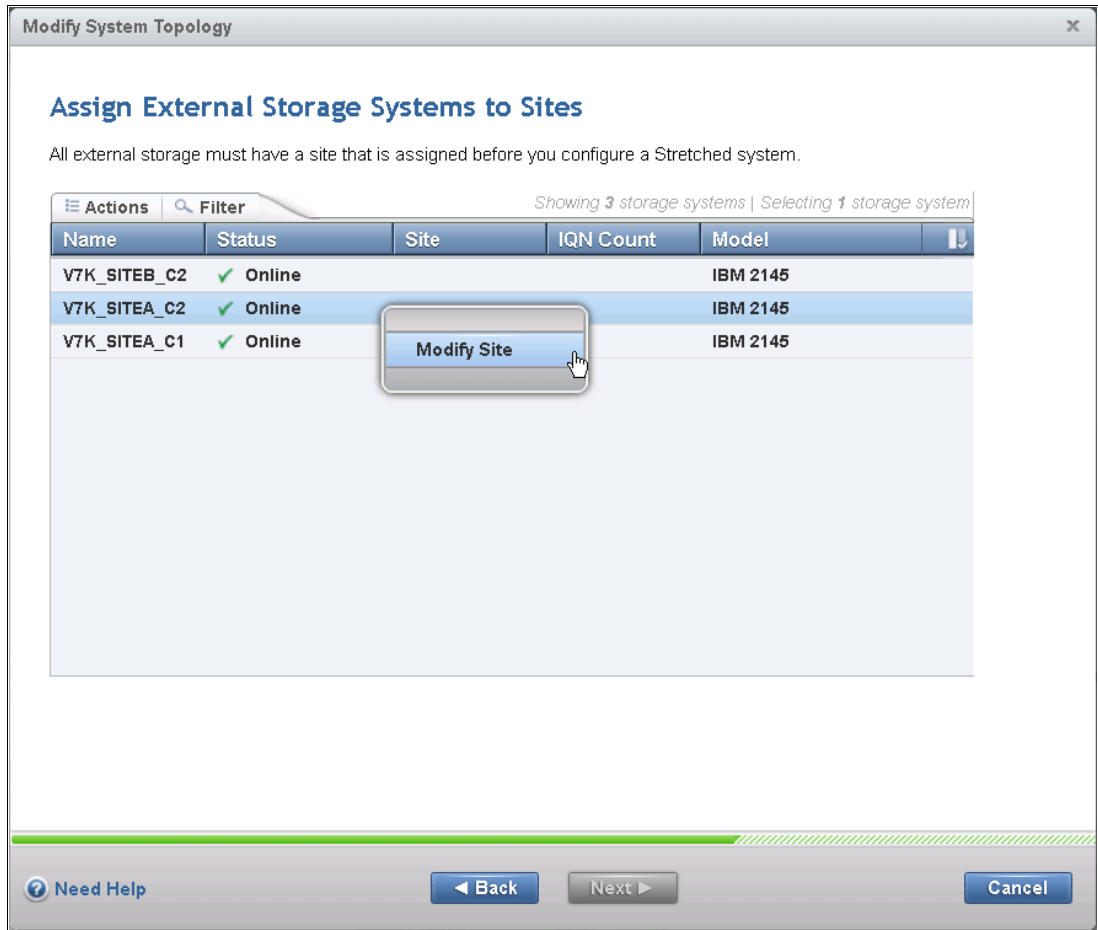


Figure 4-49 The Assign External Storage Systems to Sites window

The site selection window opens. Select the controller site from the drop down menu, as shown in Figure 4-50 on page 114.



Figure 4-50 Site selection menu for controllers

When the site settings for all the external storage is completed, click **Next**. The Summary window opens, as shown in Figure 4-51 on page 115. In this window all the configuration settings are summarized. Click **Finish** to create the configuration.

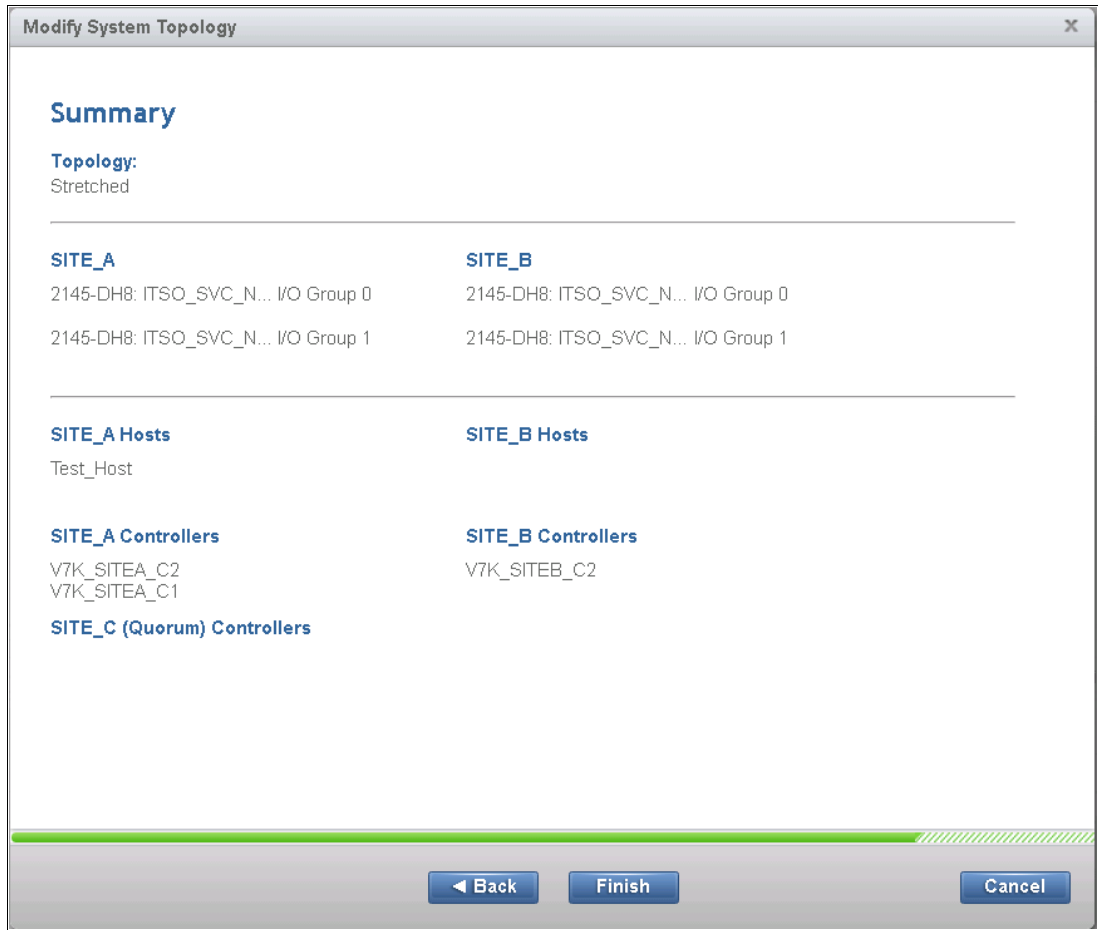


Figure 4-51 The Modify System Topology summary window.

4.12 Storage allocation to the Spectrum Virtualize

For Power Domain 1, Site 1 and Power Domain 2, Site 2, the example uses Storwize V7000 for back-end storage. Both Storwize V7000 systems are configured the same way. Power Domain 3, Site 3 has a Storwize V7000 acting as the active quorum disk.

Also, see the Storwize V7000 Knowledge Center website:

<http://pic.dhe.ibm.com/infocenter/storwize/ic/index.jsp>

Figure 4-52 shows the MDisks created in Power Domain 1.

Name	Status	Capacity	Mode	Storage System	LUN
Unassigned MDisks					
V7000SITE_RAID10	Online	0 bytes Used / 1.95 TB			
V7000SITE_RAID5	Online	510.00 GB Used / 2.22 TB			
ITSO_V7K_SITE_SA0	Online	500.00 GB	Managed	ITSO_V7K_SITE_N2	0000000000000001
ITSO_V7K_SITE_SA1	Online	500.00 GB	Managed	ITSO_V7K_SITE_N2	000000000000000B
ITSO_V7K_SITE_SA2	Online	500.00 GB	Managed	ITSO_V7K_SITE_N2	0000000000000002
ITSO_V7K_SITE_SA3	Online	500.00 GB	Managed	ITSO_V7K_SITE_N2	000000000000000D
ITSO_V7K_SITE_SSD1	Online	277.34 GB	Managed	ITSO_V7K_SITE_N2	0000000000000000
V7000SITE_RAID10	Online	0 bytes Used / 1.95 TB			
V7000SITE_RAID5	Online	510.00 GB Used / 2.22 TB			
ITSO_V7K_SITEB_SA0	Online	500.00 GB	Managed	ITSO_V7K_SITEB_N2	0000000000000014
ITSO_V7K_SITEB_SA1	Online	500.00 GB	Managed	ITSO_V7K_SITEB_N2	0000000000000015
ITSO_V7K_SITEB_SA2	Online	500.00 GB	Managed	ITSO_V7K_SITEB_N2	0000000000000016
ITSO_V7K_SITEB_SA3	Online	500.00 GB	Managed	ITSO_V7K_SITEB_N2	0000000000000017
ITSO_V7K_SITEB_SSD1	Online	277.34 GB	Managed	ITSO_V7K_SITEB_N2	0000000000000001
V7000SITEC	Online	0 bytes Used / 512.00 MB			
ITSO_V7K_SITEC_QUOR...	Online	1.00 GB	Managed	ITSO_V7K_SITEC_Q...	0000000000000000

Figure 4-52 Showing the MDisks that were created in the Storwize V7000 software

Figure 4-53 shows the volumes that are assigned to the ESC host.

Name	State	Capacity	Storage Pool	Host Mappings	UID
ESXI_VMDK_2	Online	100.00 GB	V7000SITE_RAID5	Yes	600507680183053EF800000000000092
Copy 0	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000092
Copy 1*	Online	100.00 GB	V7000SITE_RAID5	Yes	600507680183053EF800000000000092
ESXI_VMDK_3	Online	100.00 GB	V7000SITE_RAID5	Yes	600507680183053EF800000000000093
Copy 0*	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000093
Copy 1	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000093
ESXI_VMDK_4	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000094
Copy 0*	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000094
Copy 1	Online	100.00 GB	V7000SITE_RAID5	Yes	600507680183053EF800000000000094
ESXI_VMDK_5	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000095
Copy 0	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000095
Copy 1*	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000095
RR_Test	Online	10.00 GB	V7000SITE_RAID5	Yes	600507680183053EF800000000000096
Copy 0*	Online	10.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000096
Copy 1	Online	10.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF800000000000096
ESXI_VMDK_1	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF80000000000008C
Copy 0	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF80000000000008C
Copy 1*	Online	100.00 GB	V7000SITEB_RAID5	Yes	600507680183053EF80000000000008C

Figure 4-53 Volume assignment to the ESC host

4.13 Volume allocation

This section explains how to allocate volumes by using ESC. Before software version 7.5, the general recommendation was to have the volume assignments based on the local to local policy. This policy means that if a host is in Power Domain 1, Site 1, the preferred node must be in Power Domain 1, Site 1 also. With the introduction of the host site awareness in version 7.5, the host I/Os are normally performed by a node with the same site definition as the host. The preferred node definition is no longer taken in account. For this reason, no particular

recommendation must be followed for the preferred node definition, other than distributing them evenly across the cluster.

Figure 4-54 shows volumes that are assigned to the esxi_dca_p0 host.

The screenshot displays the vSphere Storage View for host `esxi_dca_p0`. The interface includes a host filter on the left with the following items:

- `esxi_dca_p0` (1 port)
- `esxi_dca_p1` (1 port)
- `esxi_dcb_p0` (1 port)
- `esxi_dcb_p1` (1 port)

The main area shows the host `esxi_dca_p0` with 1 port and Host Type: Generic. Below this is a table of volumes:

Name	State	Capacity	Storage Pool	Host Mappings	UID
ESXI_VMDK_1	Online	100.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF80000000000008C
Copy 0	Online	100.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF80000000000009C
Copy 1*	Online	100.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF80000000000008C
ESXI_VMDK_2	Online	100.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF800000000000092
Copy 0	Online	100.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF800000000000092
Copy 1*	Online	100.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF800000000000092
ESXI_VMDK_3	Online	100.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF800000000000093
Copy 0*	Online	100.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF800000000000093
Copy 1	Online	100.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF800000000000093
ESXI_VMDK_4	Online	100.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF800000000000094
Copy 0*	Online	100.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF800000000000094
Copy 1	Online	100.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF800000000000094
ESXI_VMDK_5	Online	100.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF800000000000095
Copy 0	Online	100.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF800000000000095
Copy 1*	Online	100.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF800000000000095
RR_Test	Online	10.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF800000000000096
Copy 0*	Online	10.00 GB	V7000SITEA_RAIDS	Yes	600507680183053EF800000000000096
Copy 1	Online	10.00 GB	V7000SITEB_RAIDS	Yes	600507680183053EF800000000000096

Figure 4-54 Volumes that are assigned to the esxi-dca_p0 host



VMware environment

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

- ▶ VMware configuration checklist
- ▶ VMware vCenter setup
- ▶ ESXi host installations
- ▶ VMware Distributed Resource Scheduler (DRS)
- ▶ Naming conventions
- ▶ Path Selection Policy
- ▶ VMware high availability
- ▶ VMware vStorage API for Array Integration
- ▶ VMCP VMware Component protection
- ▶ Protecting vCenter Services
- ▶ Design comments
- ▶ Script examples

5.1 VMware configuration checklist

The following items are required to gain the full benefit of the vMSC environment. This high-level list includes the major tasks that must be completed. The detail and expertise that are required to complete these tasks are beyond the intended scope of this book. Links are provided for further assistance on various topics where appropriate.

- ▶ Create naming conventions (see 5.6, “Naming conventions” on page 130):
 - Data center-wide naming ESXi
 - SDRS data stores and pools data center affinity
 - DRS vMotion pools data center affinity
- ▶ Set up all hardware, and create a detailed inventory list:
 - Follow the VMware Compatibility Guide:
<http://ibm.biz/BdxrmT>
 - Make an inventory list with details that cover the entire installation
 - Carefully mark the SAN Volume Controller node names and make associations in vSphere so that you know which SAN Volume Controller nodes are hosted in each data center.
- ▶ Build ESXhosts (for more information, see 5.4, “ESXi host installations” on page 122):
 - Two ESXhosts in each data center for maximum resiliency.
 - Patch and update to latest VMware patch level.
 - Follow VMware vSphere High Availability Deployment:
<https://ibm.biz/BdXuQD>
- ▶ Create one VM to host vCenter protected by vCenter (see 5.3, “VMware vCenter setup” on page 121):
 - Update and patch vCenter
 - Build a stretched ESXi cluster between two data centers (see 5.7.3, “HA advanced settings” on page 134)
 - Optionally, implement I/O control on storage
 - Optionally, implement VMware vSphere Distributed Switches (VDSs)
- ▶ Build an SDRS pool:
 - Make at least two pools to match data center affinity
 - Differentiate between Mirrored and Non-Mirrored LUNs if both are used
 - Set SDRS pool to manual at first and later to automatic
- ▶ Enable DRS (for more information, see 5.5, “VMware Distributed Resource Scheduler (DRS)” on page 127):
 - Make affinity rules to ESXi host in each data center
 - Make affinity rules to VMs if needed
 - Make VM to ESXi affinity rules
 - Set DRS to partial / or automatic if rules are trusted 100%

VMware Communities are a good source of information:

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

The VMware Product Interoperability Matrices are available on the VMware website:
http://partnerweb.VMware.com/comp_guide2/sim/interop_matrix.php?

5.2 VMware and Enhanced Stretched Cluster

Figure 5-1 shows an overview of the solution with VMware and Enhanced Stretched Cluster (ESC). It shows how the read/write operation is performed from an ESXi host perspective. Read/write operations are always carried out by the SAN Volume Controller node that has the same host site awareness attribute defined.

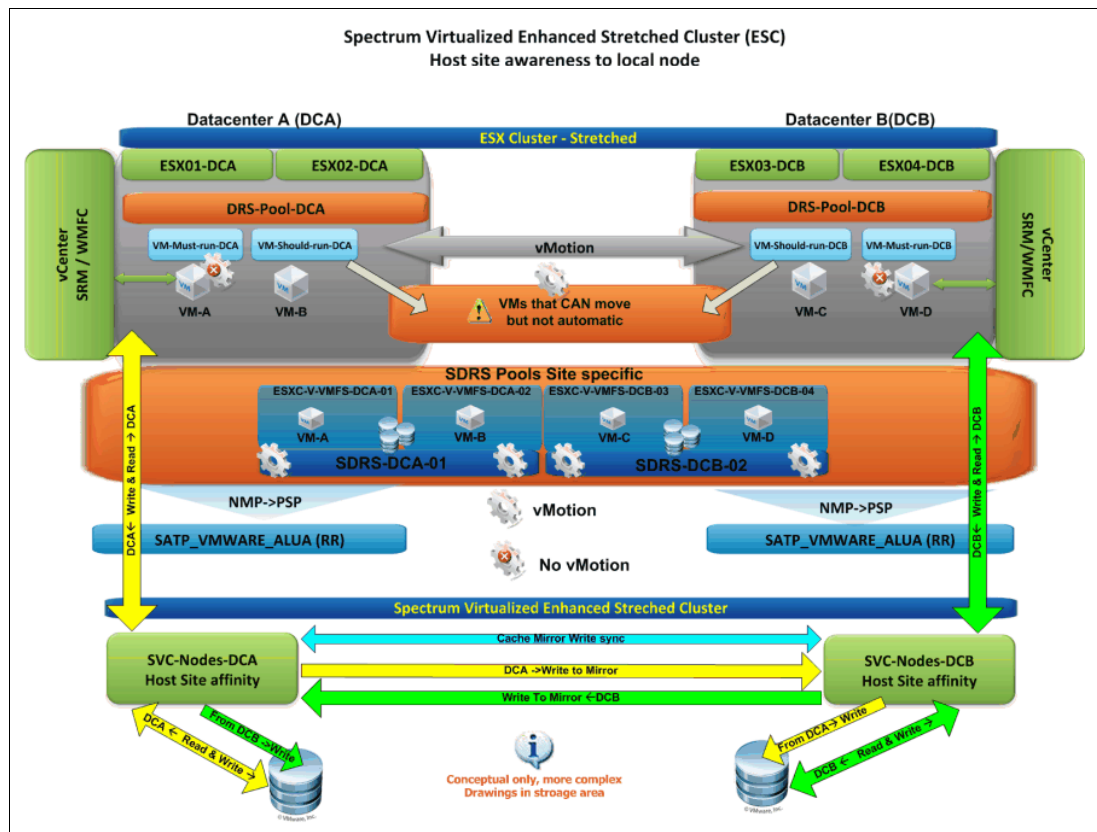


Figure 5-1 VMware stretched cluster with ESC view

5.3 VMware vCenter setup

vCenter configuration options must be implemented as part of the ESC configuration.

In an ESC, vCenter can span the two sites without problems. However, ensure that connectivity and start is set to **Automatic with host** so that, in a total failure, the vCenter automatically tries to start, along with the other vital VMs such as domain controllers, Domain Name System (DNS), and Dynamic Host Configuration Protocol (DHCP), if used.

Make affinity rules to keep these VMware components on the same primary site.

Clarification: The ESC example implementation uses vCenter 6.0.0 virtual appliance Tiny Model.

5.3.1 Metro vMotion

Use the enhanced version of vMotion, called Metro vMotion, in an ESC configuration. Metro vMotion raises the allowed latency value to 5 – 10 ms round-trip time (RTT). This increase is required when failure domains are separated by a distance of more than 300 km. The Enterprise Plus license is required for Metro vMotion.

An ESC solution has a maximum distance of 100 km, giving an RTT of a maximum of 5 ms which, in VMware terms, is not a Metro vMotion.

This is VMware's guide to a Metro Storage Cluster (vMSC) in vSphere 6:

<https://ibm.biz/BdXCzf>

vSphere Long Distance vMotion, which allows up to 100 ms RTT, is also usable in this scenario, but the example needs to keep the RTT under 5 ms due to 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 related to a specific ESXi configuration with ESC.

Attention: 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 ensure that all settings are the same between ESXi hosts, which are critical to the reliable operation of the cluster. This image can be made by using VMware Image Builder or a custom scripted installation and configuration. Standardization of the ESXi hosts safeguards against potential mismatches in configurations.

5.4.1 ESXi host HBA requirements

The HBAs for the ESXi hosts have these requirements:

- ▶ ESXi hosts require a minimum of two host bus adapters (HBAs). They must be the same type and speed.
- ▶ The HBAs must be listed in the VMware Compatibility Guide.
- ▶ 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 it does not exceed the maximum supported time of 10 ms in a Metro setup or 5 ms in a non-Metro setup. To run the latency test, use this command:

Vmkping <IP of remote ESXhost, vMotion Network>

Do a 500-ping test, pipe it to a file for later comparison, and include the date in the file name:
vmping-test01-06112013.txt

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

Example 5-1 vmping <hostname> -c 500

```
vmping ESXi-02-dcb.ibmse.local -c 500 > vmping-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 returned are consistent and repeatable.

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

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

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

Example 5-2 esxcli storage san fc list

```
# 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 (PSPs) and Native Multipath Plugins (NMPs)

VMware's overall Pluggable Storage Array (PSA) with the Native Multipath Plugin (NMP) works with third-party and VMware built-in drivers. This example uses the VMware native driver's and the NMP layer.

Figure 5-2 shows the PSA architecture.

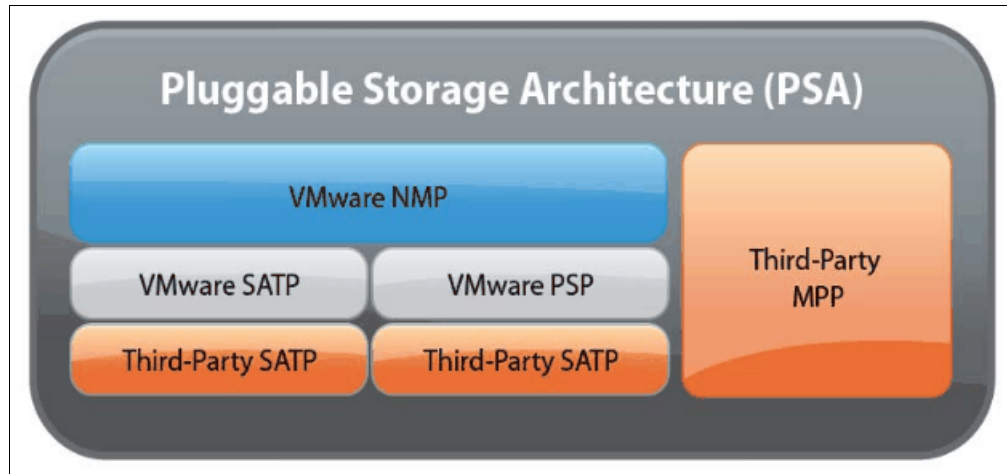


Figure 5-2 PSA plug-in architecture

For optimal performance, ESXi paths must be configured so that active paths access the SAN Volume Controller nodes that are local, meaning that they are in the same failure domain as the ESXi server.

Before Spectrum Virtualize software version 7.5, the default path selection policy was Fixed Preferred Path (path selection policy `VMW_PSP_FIXED`). Starting with Spectrum Virtualize 7.5, host site awareness allows you to use Round Robin (path selection policy `VMW_PSP_RR`) as the default. You do not need to set the preferred path any longer. However, you do need to verify that it uses the correct node (that is, the one you expect it to).

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: The example shows that the device is visible as VMW_SATP_ALUA. Ensure that this is the default by using VMW_PSP_RR.

- ▶ Spectrum Virtualize 7.5 uses VMW_SATP_ALUA with PSP_RR
- ▶ SAN Volume Controller 7.2 uses VMW_SATP_ALUA with PSP_FIXED
- ▶ SAN Volume Controller 6.2 uses VMW_SATP_SVC

Verifying the path selection policy

The current PSP for each LUN can be verified by using the command shown in Example 5-4.

Example 5-4 Verifying the path selection policy

```
esxcli storage nmp devkice 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 LUN path information from the ESXi hosts, see the VMware Knowledge Base article titled *Obtaining LUN pathing information for ESX or ESXi hosts (1003973)*:

<http://ibm.biz/BdxriP>

5.4.4 Set default PSP

Set the default PSP for the entire ESXi host, which requires that the ESXhost be restarted.

From the ESXi shell console, list the available vendors and levels, as shown in Example 5-5. Notice that 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 HyperSwapription
-----
VMW_SATP_ALUA                     VMW_PSP_MRU         Supports non-specific arrays that use the ALUA
protocol
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 lists SATP.

Example 5-6 List SATP

```
esxcli storage nmp satp list | grep SATP_ALUA
VMW_SATP_ALUA VMW_PSP_MRU Supports non-specific arrays that use the ALUA
protocol
VMW_SATP_ALUA_CX VMW_PSP_RR Placeholder (plugin not loaded)
```

For more information, see the VMware Knowledge Base article titled *Changing the default pathing policy for new/existing LUNs (1017760)*:

<https://ibm.biz/BdXCnk>

To make it easy for future disks that might be added, 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-bsp=VMW_PSP_RR --satp=VMW_SATP_ALUA
Default PSP for VMW_SATP_ALUA is now VMW_PSP_RR
```

After this command is issued, ESXhost needs to be rebooted for it to take effect.

5.4.5 Verifying Node ID path in vSphere web client

First, create the SAN Volume Controller node table. Table 5-1 shows an example.

Even with host site awareness, it is important to do this initially to verify that all components are as expected, and that you are applying the correct policy towards the nodes.

Table 5-1 Node list table example

Node ID	Data center
50:05:07:68:01:10:B1:3F	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 SAN Volume Controller nodes.

Table 5-2 ESXi-DCA 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.

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: Equally balance LUNs across HBAs on the local preferred path. With host affinity in Spectrum Virtualize software version 7.5, this is managed by the storage operation.

See 5.12.2, “PowerShell script to extract data from the entire environment and verify active and preferred paths” on page 148 for how to verify the active and preferred paths by using a script. This is still valid even with the introduction of host site awareness.

5.4.6 Path failover behavior for an invalid path

If an active path fails, the ESXi path selection policy selects an alternative path that is determined by the SAN Volume Controller node that is part of the host site awareness.

Spectrum Virtualize 7.5 decides the optimized path to the local node, and the ESXi host uses round robin (RR) to determine which of the two paths to the local node to use.

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 (preferred node) of the SAN Volume Controller is at another site on another node. It is only in the case of total site failure that the nodes that are not on the optimized path will be used. Spectrum Virtualize will automatically detect this after 25 minutes and reannounce the optimized path to become the new local node.

5.5 VMware Distributed Resource Scheduler (DRS)

Use the VMware vSphere DRS in an ESC setup because normally you do *not* want virtual machines to move to the other site. Migration should happen only in a case of site failure, or intentionally.

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

For more information about how DRS works, see the DRS Blog Guide at:

<https://ibm.biz/BdXLP>

Note: DRS rules, along with accurate and meaningful naming standards, are the most important operational considerations when you are managing an ESC.

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 case of an HA failure.

Table 5-4 DRS rules matrix

DRS-Rules	ESX-DRS-Host	Description
VM-Should-Run-In-DCA	ESX-Host-IN-DCA	VMs in Data Center A (DCA) that potentially can be migrated by vMotion to Data Center B
VM-Should-Run-IN-DCB	ESX-Host-In-DCB	VMs in Data Center B (DCB) that potentially can be migrated by 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 if you have an MSCS Cluster running with Virtual Nodes.

All virtual machines should be in a rule when they are running an ESC. Create the rules while you have the virtual machines running because you cannot create empty groups.

Since vSphere 5.5, the virtual machine to virtual machine affinity/anti-affinity rules have been changed so that whenever HA restarts a virtual machine, it respects this setting.

Figure 5-3 is an example of creating a DRS Group.

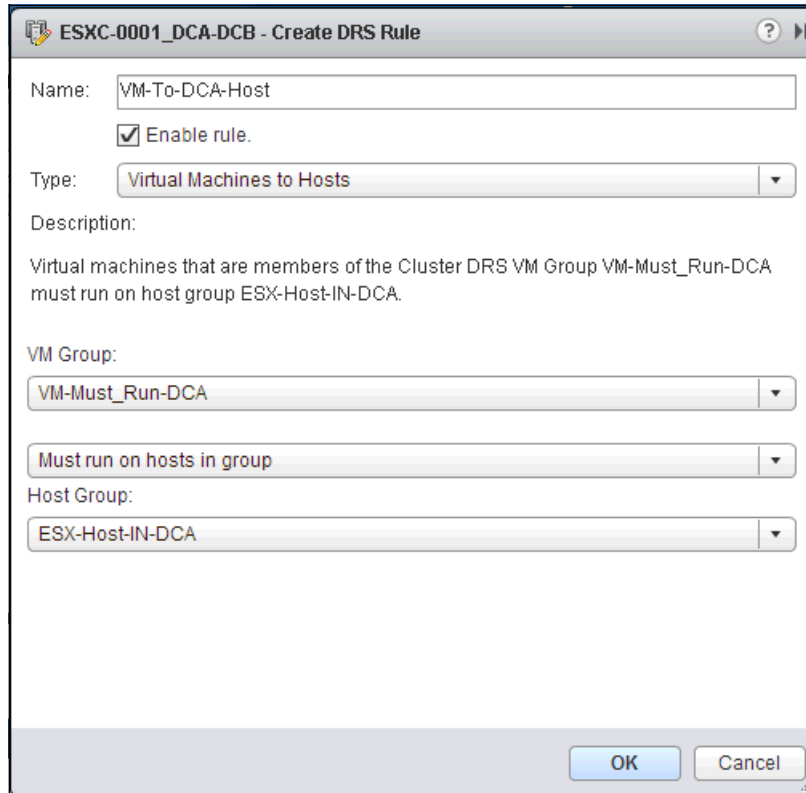


Figure 5-3 Creating DRS Group rules example

Figure 5-4 shows an example of DRS Groups/rules that are implemented in vCenter.

DRS Groups	
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 5-4 DRS-VM rules

Important: A common reason for systems encountering a Critical Event is missing and outdated guidelines in these rules.

Align your virtual machines with the corresponding groups with the data store 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.

5.6 Naming conventions

The use of a strict and well-thought-out naming convention is critical to the reliable operation of the environment, and in combination with ESC it is even more of value to know where resources are running, and be able to identify them just by their name.

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

Examples of naming standards:

ESX cluster names:

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

ESX host naming:

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

Virtual machines:

- ▶ VM-<DCA-####.<domainname.XXX>

Data stores:

- ▶ ESX_<Cluster##>_<DiskType>_<MirrorType>_<DC><#LUN_ID>_<OWNER>
- ▶ <Cluster> {just Unique} prefer a number
- ▶ <DiskType> [VMFS/NFS/RDM]
- ▶ <MirrorType>
 - M = Metro Mirrored Disk (for Synchronous disaster recovery)
 - V = Volume Mirrored Disk (for business continuity)
 - G = Global Mirror Disk (Asynchronous disaster recovery)
 - H = IBM HyperSwap (business continuity for Storwize V7000 and SAN Volume Controller two site only)
 - N = Not Mirrored
- ▶ <DC> The Preferred data center that holds the Primary Disk copy of the LUN
- ▶ <LUN_ID> OPTIONAL the Unique SCSI ID assigned by storage [0-255]
- ▶ <OWNER> Optional, if client wants Dedicated LUNs to belong to certain APPs, refer to APP-ID or name of the virtual machine that owns that LUN.

Examples of naming:

- ▶ ESXC_001_VMFS_V_DCA_01 — A Volume Mirrored LUN, Preferred from DCA
- ▶ ESXC_001_VMFS_V_DCB_02 — A Volume Mirrored LUN, Preferred from DCB
- ▶ ESXC_001_VMFS_V_DCB_03_VM-DCA_01 — With a Dedicated VM as Owner

Even datacenters and clusters should be clearly named so you clearly can identify where it is and what the cluster is used for.

Figure 5-5 shows an example of naming the data centers and clusters.

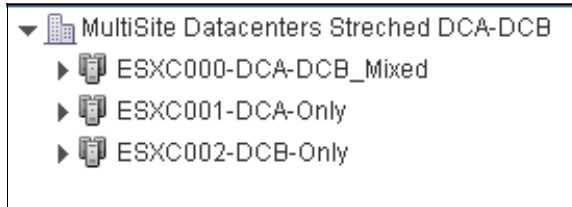


Figure 5-5 Data center and cluster naming example

Figure 5-6 shows a standard data store naming example with both ESC and HyperSwap.

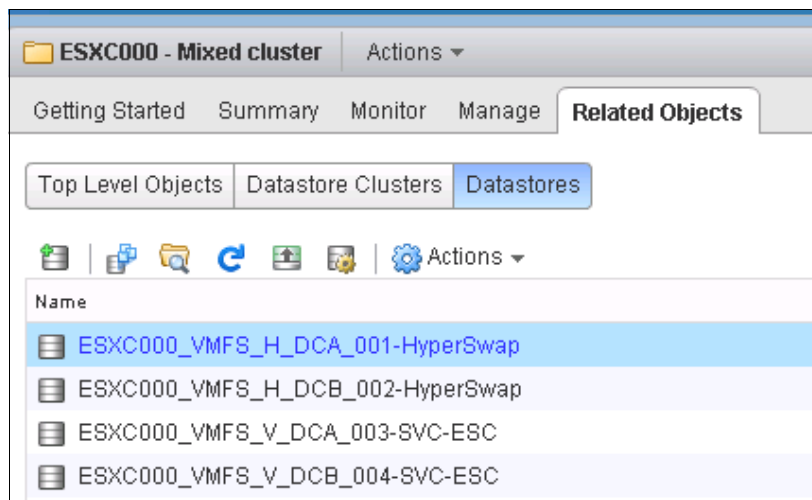


Figure 5-6 Data store naming example

And besides the actual names, use folders with a naming convention that easily identifies their contents, as shown in Figure 5-7.

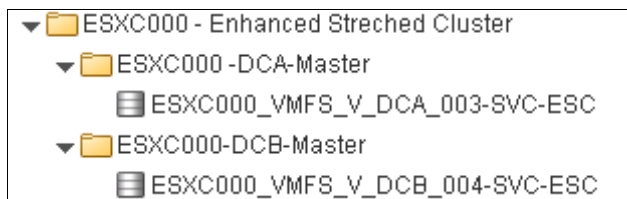


Figure 5-7 Use of folders in addition to naming standards

And to have the complete view, where you have both some ESC volumes, and HyperSwap in the same cluster as shown in Figure 5-8.

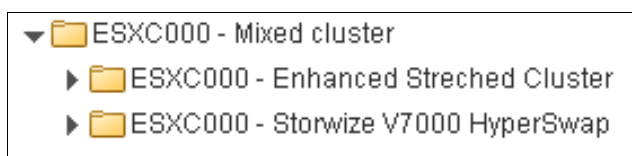


Figure 5-8 Mixed folder view

And when you look into the folders, keep the Cluster ID as part of the folder, so you keep the unique names as shown in Figure 5-9.

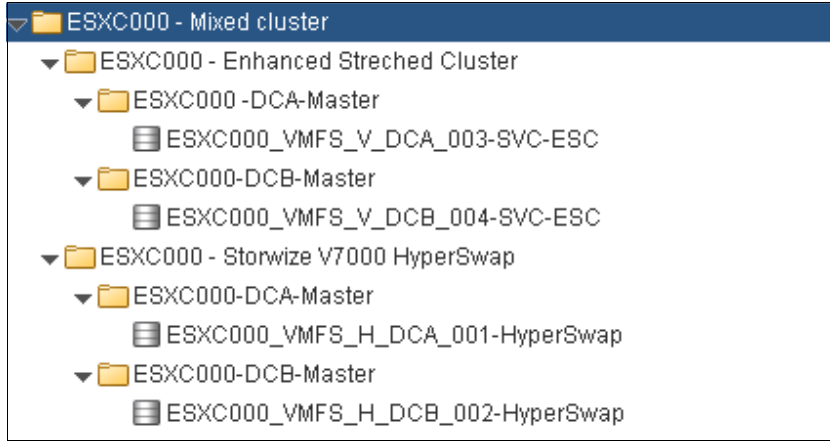


Figure 5-9 Complete folder view

As an example of how easy you can get a View with only volumes on a certain site by using Search because you have naming standards, see Figure 5-10, which shows a search using `dca` as the argument.

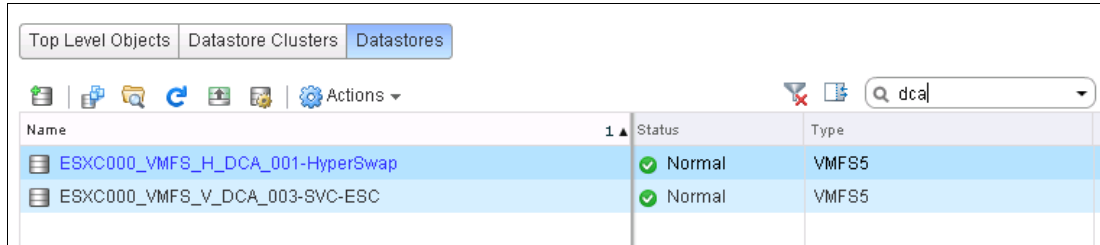


Figure 5-10 Search for data stores by reference

SDRS-Pools:

- ▶ SDRS-*<Datacenter>*<####>

For example: SDRS-DCA-001 (pool of data stores in Data Center A)

5.7 VMware high availability

Setting up a redundancy network for VMware HA is critical between the ESXi host on the cluster.

Instructions are beyond the intended scope of this book. For information about setup and configuration of VMware HA, see *VMware vSphere High Availability 5.0 Deployment Best Practices*:

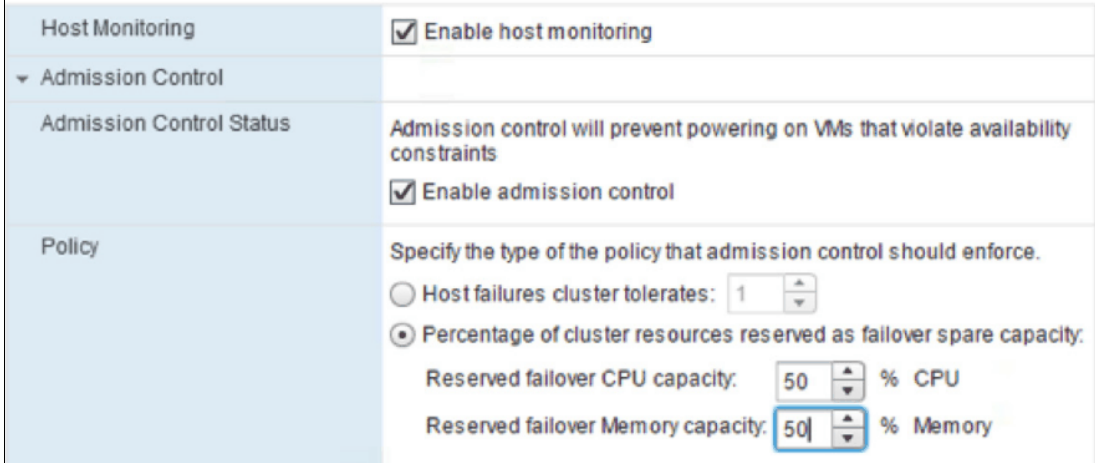
<http://ibm.biz/BdxrmN>

5.7.1 HA admission control

In a VMware ESC environment, make sure that each site can absorb the workload from the alternate site during a failure. To do so, reserve resources at each site, which are referred to as admission control.

For the vMSC environment, set the admission control policy to 50% on both CPU and memory, but obviously 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 virtual machines can move to the partner ESXi host in the other data center.

Figure 5-11 shows setting the HA Admission Control settings 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 5-11 HA admission settings to vMSC

5.7.2 HA 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 called Fault Domain Manager (FDM) is implemented in vSphere, which has completely rewritten HA code.

Basically, FDM operates at an 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, in case there is an isolation state. Generally, set the policy to **Power off, then failover** in the cluster HA setup in case a host enters the isolated state.

The VM options for the host isolation response are shown in Figure 5-12.

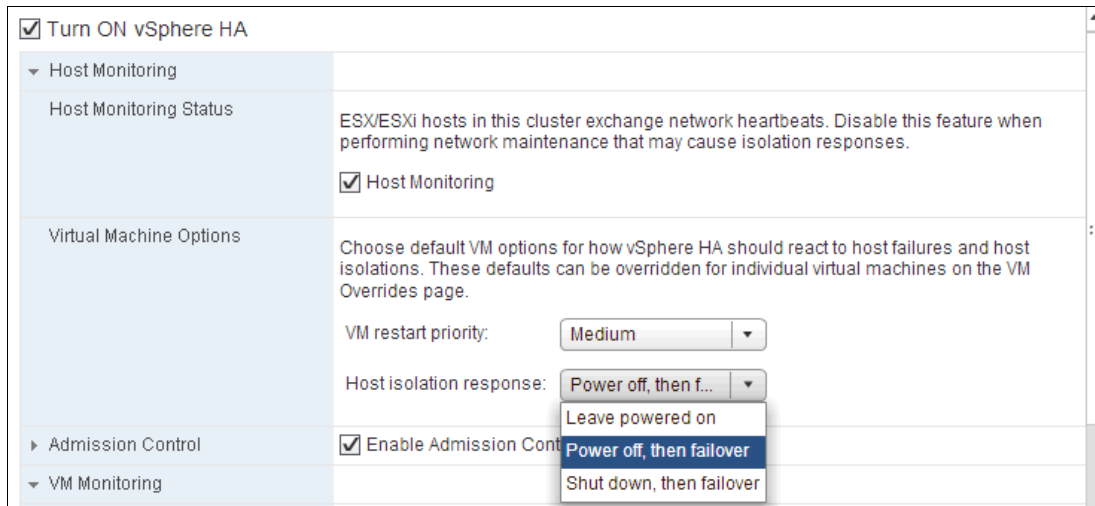


Figure 5-12 Virtual machines isolation response

Important: Be sure to review the following list because there are changes in the HA code to be aware of.

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

- ▶ Networking Heartbeat

Primary control: This checks the basic network for isolation of an ESXi host. Generally, have at least two interfaces with isolation addresses defined. For more information, see 5.7.3, “HA advanced settings” on page 134.

- ▶ HA data store Heartbeat

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

As Figure 5-13 shows, the best selection policy is **Automatically select datastores accessible from the host** for the Heartbeat.

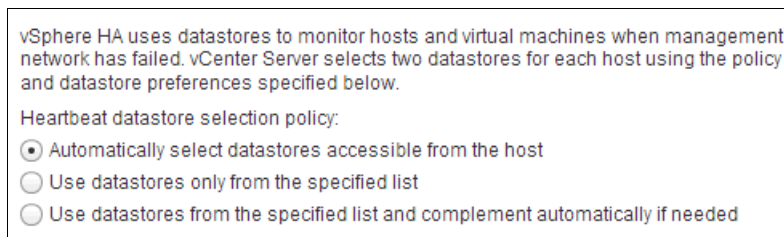


Figure 5-13 Data store selection for Heartbeat

5.7.3 HA advanced settings

Other HA settings are important for you to consider. Your environment might require additional or different ones, but these are the ones that were applicable to the example environment.

Table 5-5 provides the advanced settings that must be applied.

Remember: This is not a comprehensive list of advanced settings. The settings that are listed here are ones that are critical to this example ESC 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	This option is set to TRUE by default since Version 5.1. It allows HA to restart virtual machines that were powered off while the Permanent Device Loss (PDL) condition was in progress.

5.7.4 All Paths Down detection enhanced in vSphere 6

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

This is part of the process to secure ESXi host isolation handling of the virtual machines. Do not disable this mechanism. See VMware’s *Handling Transient APD Conditions* guide:

<https://ibm.biz/BdDwq7>

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

Figure 5-14 shows an example of the data store event in vCenter for this situation.

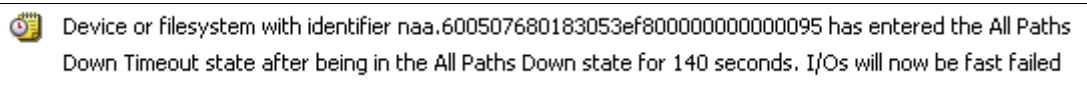


Figure 5-14 APD event

New in vSphere 6.0 is a view of which volumes have failed due to APD/PDL.

Figure 5-15 shows the APD/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				
Datastores under APD or PDL				

Figure 5-15 APD/PDL data store status page

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

<http://ibm.biz/Bdx4k7>

5.7.5 Permanent Device Loss (PDL)

With vSphere 5.5 and higher, the advanced system setting `Disk.AutoremoveOnPDL` is set to 1 by default.

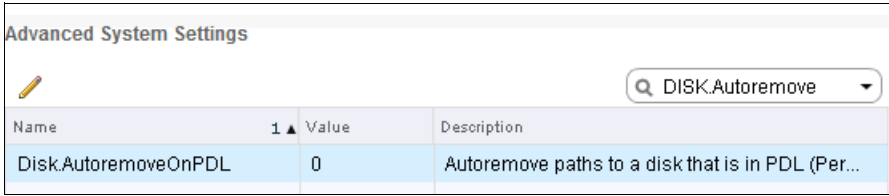
This setting is *not* recommended for use in an ESC. This is because ESC is expecting the disks to be automatically visible shortly afterward. By having the disks removed, an operator needs to manually run and rescan the disk to get them back. This process might lead to a misunderstanding of where the disks are and lead someone to think that they are gone.

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

To access the settings, open the vSphere web client and click the ESXhost, then select **Manage** → **Settings** → **Advanced system settings**, as shown in Figure 5-16.

In the filter box, click the **Edit** icon (pencil). In the **Name** field, select **DISK:Autoremove**.

Figure 5-16 shows how to find the `Disk.AutoremoveOnPDL` setting.



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

Figure 5-16 Advanced settings on PDL

Change **Advanced Option for AutoremoveOnPDL** to 0 as shown in Figure 5-17.

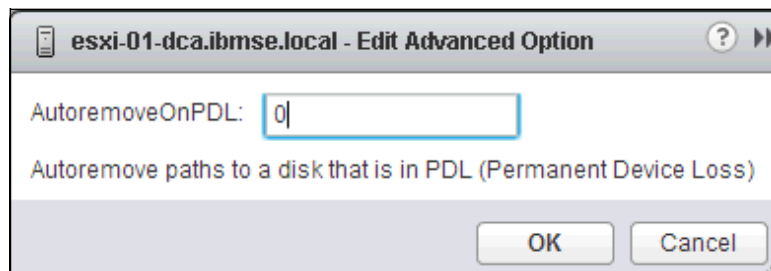


Figure 5-17 AutoremoveOnPDL setting

Ensure that the settings are the same on all ESXi hosts in your cluster. Using host profiles will make this verification easier.

Consideration: VMs not running any I/O operations might not be disabled correctly. If this problem occurs, the VM can be stopped manually from the console by using the `vmfkttools` command.

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

Also, pending I/Os or a raw device disk can hold up the PDL. 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 on 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.

5.7.6 Virtual Machine Component Protection (VMCP)

This new feature in vSphere 6 enables VMware to react to failures coming from a APD/PDL state. It is configured for the cluster, but it can be overridden by the individual virtual machine setting.

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

Important: Be sure to verify the setting on the virtual machines because it might not have been enabled on existing virtual machines if the setting in the cluster was made *before* the virtual machines were created.

There is a good blog on this subject from VMware:

<https://ibm.biz/BdXukg>

VMCP was used to create the workflow and recovery timelines shown in Figure 5-18.

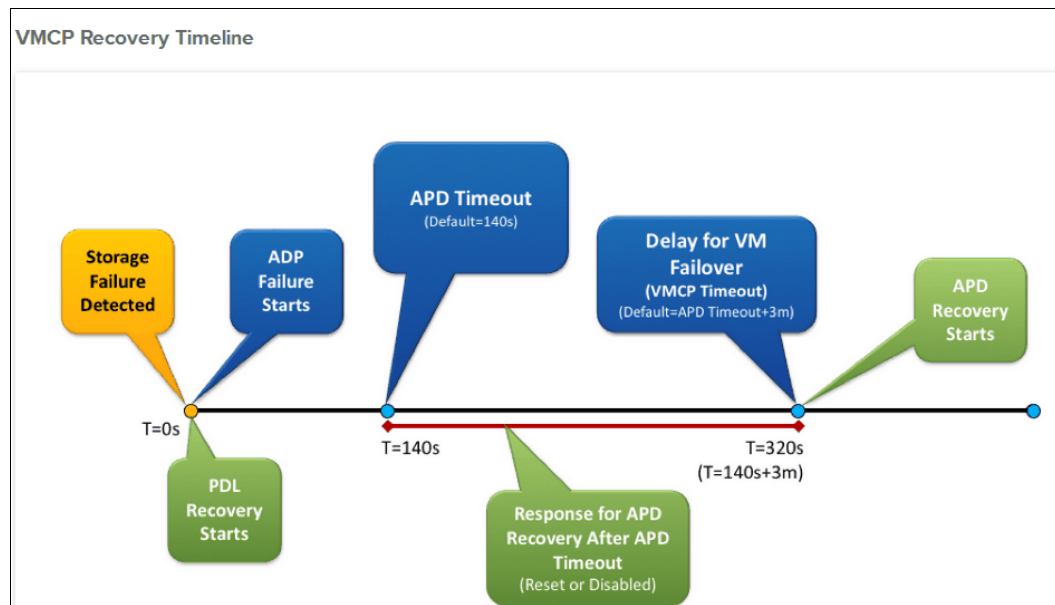


Figure 5-18 VMCP recovery timeline

The VMCP timeout can be adjusted, as in this example:

VMCP timeout =[VMCP default 180 seconds + APD TimeOut.140 seconds default] so in this case there is a 320-second default timeout before VMCP activates.

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 these can prevent a VM from being restarted. Also, if a VM has more disks spread over multiple data stores it will not be restarted if not all data stores are in an APD/PDL state.

HA can work with DRS affinity rules, and depending on your design and goals, you should take that into consideration.

Figure 5-19 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	
Profiles	
	VM anti-affinity rules Ignore rules
	VM to Host affinity rules Ignore rules

Figure 5-19 HA ignore rules

Important: VMs created before enabling VMCP do not have the APD/PDL setting enabled, so these need to be enabled manually for each VM.

Figure 5-20 shows the VM disabled for APD/PDL.

The screenshot shows a configuration window for VM settings. The following settings are visible:

- 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':

- vSphere DRS: Fully Automated
- vSphere HA: Expand for details

Buttons for 'OK' and 'Cancel' are at the bottom right.

Figure 5-20 VM disabled APD/PDL

When setting the cluster to use VMCP, be sure to pick the correct sensitivity because the preset can go from low to high, meaning from 180 seconds down to 30 seconds response time after an APD has occurred. You can also choose a custom setting, but keep in mind that having more than one setting in your solution, in case of a site failure or large amount of data stores failing, might be confusing for the operator.

Tip: Keep settings uniform across all virtual machines in the same cluster.

VMCP Sensitivity is key to get things back up and running. In the case of a Split Brain scenario, you need an aggressive setting to let HA decide and force only one instance of the virtual machine. Figure 5-21 shows that the VMCP Sensitivity settings in HA are high.

The screenshot shows the 'VM monitoring sensitivity' configuration window. The 'Preset' option is selected, and the slider is positioned at 'High'. The 'Custom' option is also visible with the following settings:

- Failure interval: 30 seconds
- Minimum uptime: 120 seconds
- Maximum per-VM resets: 3
- Maximum resets time window: Within 1 hrs

Figure 5-21 VMCP Sensitivity setting

5.7.7 Storage failure detection flow

The following blog illustrates the entire process in a storage failure scenario:

<https://ibm.biz/BdXukg>

Figure 5-22 shows storage failure detection flow.

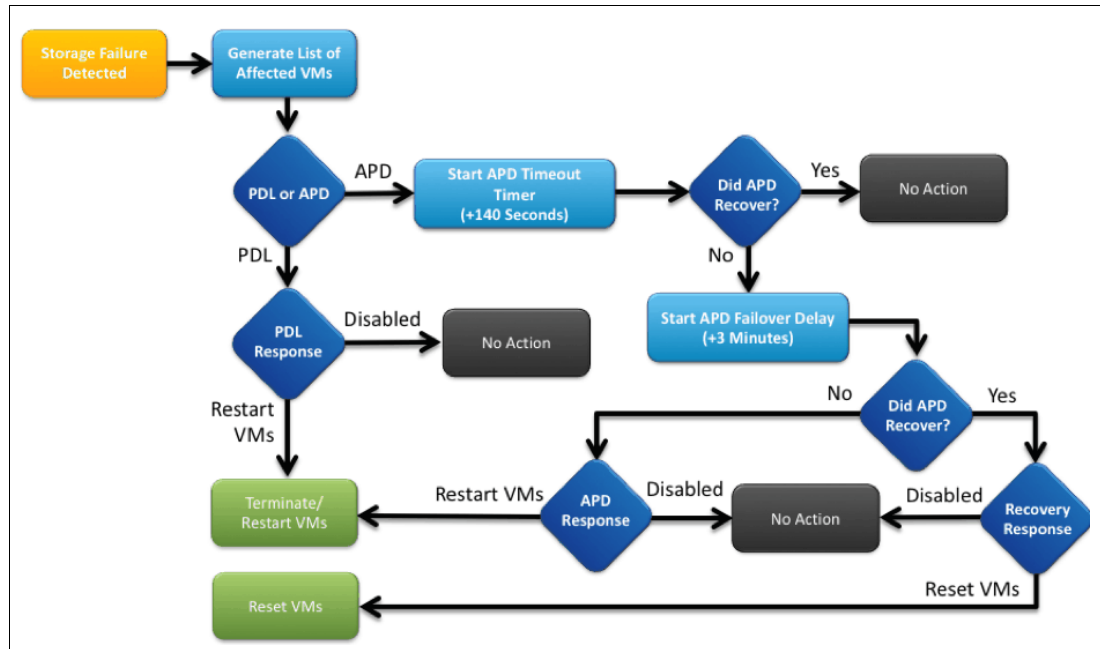


Figure 5-22 Storage failure detection flow

5.8 VMware vStorage API for Array Integration

VMware vStorage API for Array Integration (VAAI) is supported if it is listed on the hardware compatibility list. Spectrum Virtualize 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.

Several commands that can be used to check VAAI status. Example 5-8 shows using 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 following commands and check whether the default interval value is set to 1, which means it is enabled (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:
Description: 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:
Description: 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:
  Description: Enable hardware accelerated VMFS data initialization (requires compliant hardware)
```

5.9 Protecting vCenter Services

Secure the VMs from hardware failure by running vCenter as a virtual machine on the primary data center.

VMware has a list of possible solutions to secure this important component. This website 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 5-23 shows the vCenter supported 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 ¹	No	No	No	Yes ⁵	No
5.0	Yes ¹	No	No	No	Yes ⁵	No
5.1	Yes ¹	No	No	No	Yes ⁵	No
5.5	Yes ¹	No	Yes ⁴	Yes ⁷	Yes ⁵	No
6.0	Yes ¹	Yes ²	Yes	Yes ⁶	No	Yes ³

Figure 5-23 vCenter failover solution support matrix

Note: Using vCenter 6.0 it is not supported to use Heartbeat.

Medium/large solutions should implement WSFC 5.9.1, “vCenter Availability Based Windows Server Failover Clustering (WSFC)”.

Tiny/small solution can use vCenter appliance with FT enabled as an alternative.

Figure 5-24 shows vCenter scaling for solutions based on size.

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 5-24 vCenter scaling

5.9.1 vCenter Availability Based Windows Server Failover Clustering (WSFC)

The following links to a VMware Guide that shows how to build a cluster using Microsoft Cluster technologies to provide the best failover solution, and to protect the vCenter components.

<https://ibm.biz/BdXLtJ>

Figure 5-25 shows the VMware best practice guide to building a cluster using Microsoft Cluster.

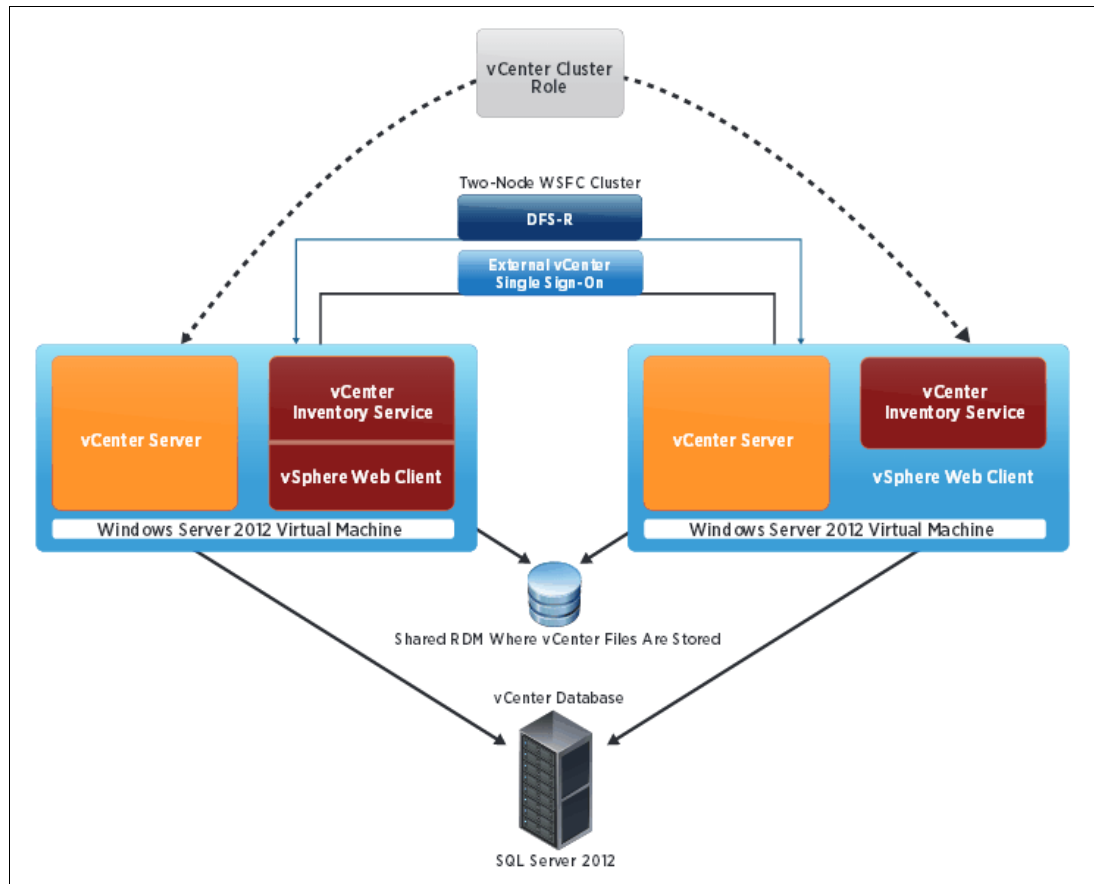


Figure 5-25 WSFC Microsoft cluster running VMware vCenter

5.10 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 a product such as RVTools to extract all vital data from the vCenter database to CSV files. You can use these files when you are planning for recovery to detect connections and missing relationships in the virtual infrastructure.

Recovery is more than just getting things started. It is getting them started in a way that the infrastructure can start to function. Therefore, categorization of virtual machines is important. At a minimum, complete these steps:

- ▶ 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 virtual machines in visible folders, and ensure that restart priorities are clearly understood and documented.
- ▶ Create and enforce naming standards. For more information, see 5.6, “Naming conventions” on page 130.

All of this is in addition to the normal, day-to-day planning, such as backup operations. Floor space management is also important so that you know the physical location of servers.

Print the basic 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 some closed servers on the floor.

5.10.1 VMware alternatives to minimize the impact of a complete site failure (“split brain” scenario)

To help bring back vital virtual machines after a site failure, consider including these products in your plan.

vCenter Failover Method

This method is described in 5.9, “Protecting vCenter Services” on page 141.

vCenter Site Recovery Manager

Site Recovery Manager (SRM) is a software suite that takes care of an entire site failure scenario.

All of the above can be used in different scenarios and become part of any recovery plan. SRM interacts with vReplicator if used.

A Split Brain situation is controlled by VMware HA, where the master ESX host is the controlling part in each ESX Cluster.

With VMCP in aggressive mode, the Split Brain is solved because VMware Master Node ensures that only one instance of the virtual machine is running. For more information, see 5.7.6, “Virtual Machine Component Protection (VMCP)” on page 137.

5.10.2 Investigating 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 has many knowledge base articles about investigating HA failures. Because it always is a critical situation, create a VMware support case as your first step.

There is a good troubleshooting guide here:

<https://ibm.biz/BdXCgD>

A SYSLOG tool that can collect and analyze the data, like VMware Log Insight™ which is license based, can help your analysis of the situation a lot.

Prioritized plan:

- ▶ Create a VMware support case.
- ▶ Ensure that the ESXi management cluster is running first, if applicable.
- ▶ Ensure that vCenter is operational. If you have vRecovery running, use that to activate the copy on the remaining site.
- ▶ Start VMs, such as data centers, DNS/WINS.

Make sure that storage is running at the site as expected.

Investigate whether any of the data stores are offline, inactive, or not functional for any reason before you start any VMs.

Depending on the situation, look at the vCenter logfiles and or the ESX logfiles.

The location of log files for VMware products is available here:

<https://ibm.biz/BdXCgy>

The Windows version of VMware vCenter Server 6.0 logs are located in the %ALLUSERSPROFILE%\VMware\vCenterServer\logs folder.

The VMware vCenter Server Appliance 6.0 logs are located in the /var/log/vmware/ folder.

Figure 5-26 shows the vCenter logfiles locations.

vCenter Server	vCenter Server Appliance	Description
vmware-vpx\vpzd.log	vpzd/vpzd.log	The main vCenter Serverlog
vmware-vpx\vpzd-profiler.log	vpzd/vpzd-profiler.log	Profile metrics for operations performed in vCenter Server
vmware-vpx\vpzd-alert.log	vpzd/vpzd-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
vmkdir	vmkdir	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 5-26 vCenter logfiles

For more information about the location of VMware vCenter Server 6.0 log files, see:

<https://ibm.biz/BdXCgs>

Monitor the ESXi host log files, either from SSH directly or through vCenter, if it is still running. You can find it 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 some of them. If the vmkernel.log file has been 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, there is no way to recover automatically. The situation needs to be resolved at the storage array fabric layer to restore connectivity to the host.

All ESXi hosts that were affected by the APD might require a reboot. Table 5-6 shows the sense code for PDL.

APD/PDL in vSphere 6 is optimized to react on these failures, as already described in 5.7.4, “All Paths Down detection enhanced in vSphere 6” on page 135.

Table 5-6 Sense code for PDL

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

5.11 Design comments

When you create an ESC 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 have access to fail over or vMotion instantly, without rezoning your SAN disk or switching to the mirrored copy of the disk.

A VMware stretched cluster with an ESC is best managed by implementing rules that, under normal operation, bind the virtual machines to each data center that is part of an ESC. Use affinity rules to secure it, and use vMotion to prevent disasters and to balance loads.

Use vMSC in enterprise solutions where distance is the key factor and the RTT calculation is measured in ms. From the VMware perspective, the solution is accepted up to 10 ms RTT. However, in this example, ESC is based on a 3 ms solution, which is capable of up to 300 km.

The only limit is the response times to the disk that are at the remote site, which 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 has expanded vMotion capabilities using Long Distance vMotion up to 100 ms, this solution is not capable of supporting more than 5 ms delay due to the 300 KM distance of the fiber optic cable.

5.12 Script examples

The following script examples can help with testing VM mobility and checking the preferred path policy.

The example is planned to operate on a large scale, so you need to be able to quickly detect and organize the disks. The fastest way to do this is to use vSphere PowerCLI or Microsoft PowerShell scripts.

Both scripts must be modified to suit your environment. Also, have someone who is familiar with PowerShell implement the scripts.

Reminder: You use these scripts at your own risk. However, because they do *not* change anything, they are unlikely to cause any harm.

5.12.1 PowerShell test script to move VMs 40 times between two ESXi hosts

It is important to make sure that the infrastructure is stable by testing the capability to move VMs between the ESXhost in the clusters. To do so, use this sample script to automate the process.

The script requires a PowerShell working environment with VMware automation tools installed. The script can then be run from the Power CLI 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 }
#
```



```

### 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 dont 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
datstores, 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 ","}
```



Enhanced Stretched Cluster diagnostic and recovery guidelines

This chapter addresses Enhanced Stretched Cluster (ESC) diagnostic and recovery guidelines. These features help you understand what is happening in your ESC environment after a critical event. This knowledge is crucial when you are making decisions to alleviate the situation. You might decide to wait until the failure in one of the two sites is fixed or 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. Some of them are specific to the example lab environment and some are common with every ESC installation.

Before starting a recovery action after a disaster is declared, it is important that you are familiar with all of the recovery steps that you are going to follow. Test and document 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 run the recovery action with IBM Support engaged.

This chapter includes the following sections:

- ▶ Solution recovery planning
- ▶ Recovery planning
- ▶ Enhanced Stretched Cluster diagnosis and recovery guidelines

6.1 Solution recovery planning

In the context of the ESC 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 solution is designed, incorporate a solution recovery plan into the process.

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 described in this chapter.

6.2 Recovery planning

To achieve the most benefit from the ESC configuration, postinstallation 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 effect on the client applications. Correct planning and configuration backup also helps minimize downtime.

You can categorize the recovery in the following ways:

- ▶ Recover a fully redundant configuration in the surviving site without ESC.
- ▶ Recover a fully redundant configuration in the surviving site with ESC 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 configuration, complete these steps:

1. Collect a detailed configuration. To do so, run a daily Spectrum Virtualize configuration backup with the CLI commands shown in Example 6-1.

Example 6-1 Saving the Spectrum Virtualize configuration

```
IBM_2145:ITS0_SVC_ESC:superuser>svconfig backup
.....
CMMVC6155I SVCCONFIG processing completed successfully
IBM_2145:ITS0_SVC_SPLIT:superuser>lsdumps
id filename
0 151580.trc.old
.
.
24 SVC.config.backup.xml_151580
```

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 SVC ESC
admin@10.17.89.251:/tmp/SVC.config.backup.xml_151580 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 Spectrum Virtualize commands to issue

```
lsystem
lssite
lsnode
lsnode node <nodes name>
lsnodevpd <nodes name>
lsiogrp
lsiogrp <iogrps name>
lscontroller
lscontroller <controllers name>
lsdiskgrp
lsdiskgrp <mdiskgrps name>
lsmdisk
lsquorum
lsquorum <quorum id>
lsdisk
lshost
lshost <host name>
lshostvdiskmap
```

From the output of these commands and the .xml file, you have a complete picture of the ESC infrastructure. Remember that the SAN Volume Controller FC ports worldwide node names (WWNNs), so you can reuse them during the recovery operation that is described in 6.3.3, “Recovery guidelines” on page 167.

Example 6-4, which is contained in the .xml file, shows what you need to re-create an ESC environment after a critical event.

Example 6-4 XML configuration file

```
<object type="node" >
  <property name="id" value="1" />
  <property name="name" value="node_151580" />
  <property name="UPS_serial_number" value="100014P293" />
  <property name="WWNN" value="500507680110B13F" />
  <property name="status" value="online" />
  <property name="IO_Group_id" value="0" />
  <property name="IO_Group_name" value="io_grp0" />
  <property name="partner_node_id" value="2" />
  <property name="partner_node_name" value="node_151523" />
  <property name="config_node" value="yes" />
  <property name="UPS_unique_id" value="2040000044802243" />
  <property name="port_id" value="500507680140B13F" />
  <property name="port_status" value="active" />
  <property name="port_speed" value="8Gb" />
  <property name="port_id" value="500507680130B13F" />
  <property name="port_status" value="active" />
  <property name="port_speed" value="8Gb" />
  <property name="port_id" value="500507680110B13F" />
  <property name="port_status" value="active" />
  <property name="port_speed" value="8Gb" />
  <property name="port_id" value="500507680120B13F" />
  <property name="port_status" value="active" />
  .
lines omitted for brevity
  .
  <property name="service_IP_address" value="10.17.89.251" />
  <property name="service_gateway" value="10.17.80.1" />
```

```
<property name="service_subnet_mask" value="255.255.240.0" />
<property name="service_IP_address_6" value="" />
<property name="service_gateway_6" value="" />
<property name="service_prefix_6" value="" />
```

You can also get this information from the `.txt` command output that is shown in Example 6-5.

Example 6-5 lsnode example output command

```
IBM_2145:ITSO_SVC_ESC:superuser>lsnode 1
id 1
name ITSO_SVC_NODE1_SITE_A
UPS_serial_number 100006B119
WWNN 500507680100B13F
status online
IO_group_id 0
IO_group_name io_grp0
partner_node_id 2
partner_node_name ITSO_SVC_NODE1_SITE_B
config_node yes
UPS_unique_id 2040000006481049
port_id 500507680140B13F
port_status active
port_speed 8Gb
port_id 500507680130B13F
port_status active
port_speed 8Gb
port_id 500507680110B13F
port_status active
port_speed 8Gb
port_id 500507680120B13F
port_status active
port_speed 8Gb
hardware CF8
iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea
iscsi_alias
failover_active no
failover_name ITSO_SVC_NODE1_SITE_B
failover_iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1siteb
failover_iscsi_alias
panel_name 151580
enclosure_id
canister_id
enclosure_serial_number
service_IP_address 10.17.89.253
service_gateway 10.17.80.1
service_subnet_mask 255.255.240.0
service_IP_address_6
service_gateway_6
service_prefix_6
service_IP_mode static
service_IP_mode_6
```

For more information about backing up your configuration, see the IBM Spectrum Virtualize Information Center:

<https://ibm.biz/Bdsvxb>

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 cabling, and ensure that it is fully documented.
6. Back up your SAN zoning. The zoning backup can be done by using your Fibre Channel (FC) switch command-line interface 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 command-line interface (CLI) commands. You can also use the appropriate utility to back up the entire configuration.

Example 6-6 shows how to save the information in a .txt file by using CLI commands.

Example 6-6 Zoning information example

```
Public_A1:FID111:admin> switchshow
switchName:      Public_A1
switchType:      121.3
switchState:     Online
switchMode:      Native
switchRole:      Principal
switchDomain:    11
switchId:        fffc0b
switchWwn:       10:00:00:05:33:b5:3e:01
zoning:          ON (ITSO_Public1)
switchBeacon:    OFF
FC Router:       OFF
Allow XISL Use: OFF
LS Attributes:   [FID: 111, Base Switch: No, Default Switch: No, Address Mode 0]
```

Index	Slot	Port	Address	Media	Speed	State	Proto		
12	1	12	0bfc0	-- --	Online	VE VE-Port	10:00:00:05:33:97:a5:01		
"Public_B1" (downstream)									
192	8	0	0bcfc0	id	N16	No_Light	FC		
193	8	1	0bcf80	id	N16	No_Light	FC		
194	8	2	0bcf40	id	N8	Online	FC	F-Port	50:05:07:68:01:40:b1:3f
196	8	4	0bcec0	id	N8	Online	FC	F-Port	50:05:07:68:02:10:00:ef
197	8	5	0bce80	id	N8	Online	FC	F-Port	50:05:07:68:02:20:00:ef
198	8	6	0bce40	id	N8	Online	FC	F-Port	50:05:07:68:02:10:00:f0
199	8	7	0bce00	id	N8	Online	FC	F-Port	50:05:07:68:02:20:00:f0

```
Public_A1:FID111:admin> fabricshow
Switch ID      Worldwide Name          Enet IP Addr    FC IP Addr      Name
-----
11: fffc0b 10:00:00:05:33:b5:3e:01 10.17.85.251    0.0.0.0         >"Public_A1"
21: fffc15 10:00:00:05:33:97:a5:01 10.17.85.195    0.0.0.0         "Public_B1"
```

The Fabric has 2 switches

```
Public_A1:FID111:admin> cfgshow
Defined configuration:
cfg:  ITSO_Public1
      V7K_SITEB; SVCN2P1_DS5100Cont1; SVCN2P1_DS5100Cont2;
      V7K_SITEA; SVCN1P1_DS5100Cont2; SVCN2P1_V7KSITEB;
      SVCN1P1_DS5100Cont1; SVCN2P1_V7KSITEA; SVCN1P1_V7KSITEB;
      SVCN1P1_V7KSITEA
zone: SVCN1P1_DS5100Cont1
      ITSO_SVC_N1_P1; ITSO_DS5100_Cont1_P3; ITSO_DS5100_Cont1_P1
zone: SVCN1P1_DS5100Cont2
      ITSO_SVC_N1_P1; ITSO_DS5100_Cont2_P1; ITSO_DS5100_Cont2_P3
.
```

```

lines omitted for brevity
.
zone: V7K_SITEA
    ITS0_V7K_SITEA_N1_P2; ITS0_V7K_SITEA_N1_P1;
    ITS0_V7K_SITEA_N2_P1; ITS0_V7K_SITEA_N2_P2
zone: V7K_SITEB
    ITS0_V7K_SITEB_N2_P1; ITS0_V7K_SITEB_N1_P2;
    ITS0_V7K_SITEB_N1_P1; ITS0_V7K_SITEB_N1_P4
alias: ITS0_DS5100_Cont1_P1
    20:16:00:A0:B8:47:39:B0
alias: ITS0_DS5100_Cont1_P3
    20:36:00:A0:B8:47:39:B0
alias: ITS0_DS5100_Cont2_P1
    20:17:00:A0:B8:47:39:B0

```

```

lines omitted for brevity
.
alias: ITS0_V7K_SITEA_N2_P2
    50:05:07:68:02:20:00:F0
alias: ITS0_V7K_SITEB_N1_P1
    50:05:07:68:02:10:54:CA
alias: ITS0_V7K_SITEB_N1_P2
    50:05:07:68:02:20:54:CA
alias: ITS0_V7K_SITEB_N1_P4
    50:05:07:68:02:40:54:CA
alias: ITS0_V7K_SITEB_N2_P1
    50:05:07:68:02:10:54:CB

```

```

Effective configuration:
cfg: ITS0_Public1
zone: SVCN1P1_DS5100Cont1
    50:05:07:68:01:40:b1:3f
    20:36:00:a0:b8:47:39:b0
    20:16:00:a0:b8:47:39:b0
zone: SVCN1P1_DS5100Cont2
    50:05:07:68:01:40:b1:3f
    20:17:00:a0:b8:47:39:b0
    20:37:00:a0:b8:47:39:b0
zone: SVCN1P1_V7KSITEA
    50:05:07:68:02:20:00:ef
    50:05:07:68:01:40:b1:3f
    50:05:07:68:02:10:00:ef
    50:05:07:68:02:10:00:f0
    50:05:07:68:02:20:00:f0

```

```

lines omitted for brevity
.
zone: V7K_SITEA
    50:05:07:68:02:20:00:ef
    50:05:07:68:02:10:00:ef
    50:05:07:68:02:10:00:f0
    50:05:07:68:02:20:00:f0
zone: V7K_SITEB
    50:05:07:68:02:10:54:cb
    50:05:07:68:02:20:54:ca
    50:05:07:68:02:10:54:ca
    50:05:07:68:02:40:54:ca

```

During the implementation, use WWNN zoning. During the recovery phase after a critical event, reuse the same domain ID and same port number that was 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, “Recovery guidelines” on page 167.

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 ESC implementation, you can use different vendors’ storage subsystems. Configure those subsystems according to the Spectrum Virtualize guidelines to be used for volume mirroring.

Back up your storage subsystem configuration so that you can re-create the same environment during a critical event when you re-establish your stretched cluster infrastructure in a different site with new storage subsystems.

For more information, see 6.3.3, “Recovery guidelines” on page 167.

- a. As an example, for IBM DS3, DS4, or DS5 storage subsystems, save a copy of an up-to-date subsystem profile, as shown in Figure 6-1, in a safe place.

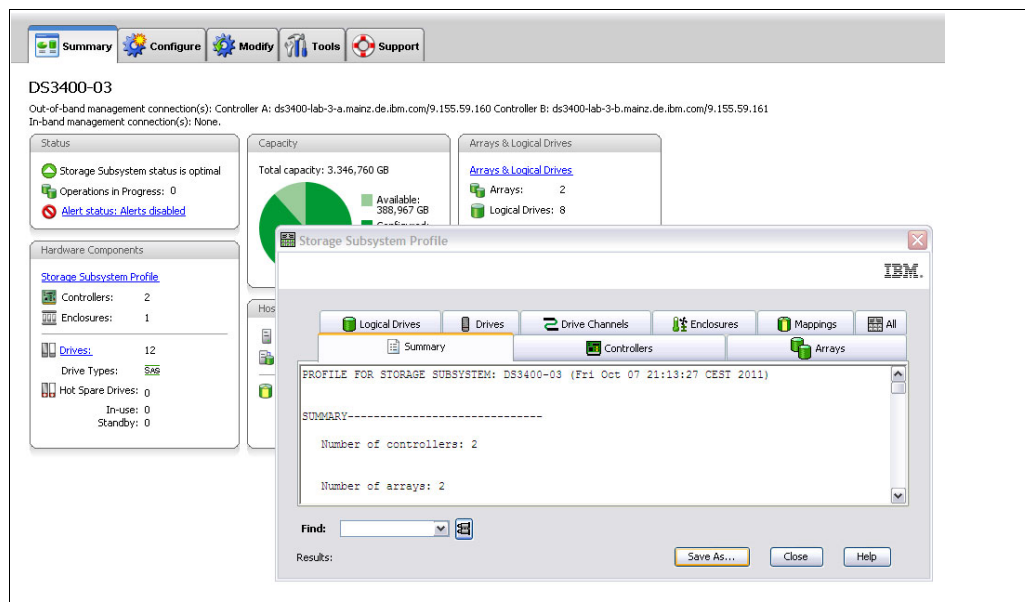


Figure 6-1 DS storage subsystem back-up configuration example

- b. For the IBM DS8000® storage subsystem, save the output of the CLI commands in .txt format, as shown in Example 6-7.

Example 6-7 DS8000 commands

```
lsarraysite -l
lsarray -l
lsrank -l
lsextpool -l
lsfbvol -l
lshostconnect -l
lsvolgrp -l
showvolgrp -lunmap <SVC vg_name>
```

- c. For the IBM XIV® Storage System, save the output of the XCLI commands in .txt format, as shown in Example 6-8.

Example 6-8 XIV subsystem commands

```
host_list
host_list_ports
mapping_list
vol_mapping_list
pool_list
vol_list
```

- d. For IBM Storwize V7000, collect configuration files and the output report as described in 6.2, “Recovery planning” on page 154.
- e. For any other supported storage vendor’s products, see their documentation for MDisk configuration and mapping.

6.3 Enhanced Stretched Cluster diagnosis and recovery guidelines

This section provides guidelines for diagnosing a critical event in one of the two sites where the ESC 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

An ESC environment can have many different critical event scenarios. Some of them can be handled by using standard (*business as usual*) recovery procedures. This section addresses all of the operations that are required to recover from a *complete site failure*.

Some parts of the recovery depend on the environment design. This section shows what was done to diagnose the situation and, later, to recover the ESC. Most of these steps are basic and can be used in every environment and configuration.

Because of the importance of the success of the recovery action, do not improvise this action. Instead, perform all of the steps under the direction of IBM Support.

The following list includes some of the scenarios that you might face and their required recovery actions:

- ▶ Back-end storage box failure in one failure domain: Business as usual because of volume mirroring.
- ▶ Partial SAN failure in one failure domain: Business as usual because of SAN resilience.
- ▶ Total SAN failure in one failure domain: Business as usual because of SAN resilience, but pay attention to the effect on performance. You need to take appropriate action to minimize the effect on applications.
- ▶ SAN Volume Controller node failure in one failure domain: Business as usual because of IBM Spectrum Virtualize and SAN Volume Controller high availability.
- ▶ 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 Site 3, where the active quorum disk is located. This scenario is covered in 6.3.2, “Diagnosis guidelines” on page 161.

topology_status dual_site

rc_auth_method none

IBM_2145:ITSO_SVC_ESC:superuser>lsiogrp

id	name	node_count	vdisk_count	host_count
0	ITSO_SVC_ESC_0	2	2	4
1	ITSO_SVC_ESC_1	2	2	4
2	io_grp2	0	0	4
3	io_grp3	0	0	4
4	recovery_io_grp	0	0	0

IBM_2145:ITSO_SVC_ESC:superuser>lsnode

id	name	UPS_serial_number	WWNN	status	IO_group_id
1	ITSO_SVC_NODE1_SITE_A	100006B119	500507680100B13F	online	0
ITSO_SVC_ESC_0 yes 2040000006481049 CF8					
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsosvcnode1sitea 151580					
2	ITSO_SVC_NODE2_SITE_B	100006B074	500507680100B0C6	online	0
ITSO_SVC_ESC_0 no 20400000064801C4 CF8					
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsosvcnode2siteb 151523					
3	ITSO_SVC_NODE3_SITE_A	1000849047	50050768010027E2	online	1
ITSO_SVC_ESC_1 no 2040000204240107 8G4					
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsosvcnode3sitea 108283					
4	ITSO_SVC_NODE4_SITE_B	1000871173	50050768010037E5	online	1
ITSO_SVC_ESC_1 no 20400002070411C3 8G4					
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsosvcnode4siteb 104643					

IBM_2145:ITSO_SVC_ESC:superuser>lscontroller

id	controller_name	ctrl_s/n	vendor_id	product_id_low
0	ITSO_V7K_SITEB_N2	2076	IBM	2145
1	ITSO_V7K_SITEA_N2	2076	IBM	2145
2	ITSO_V7K_SITEC_Q_N1	2076	IBM	2145
3	ITSO_V7K_SITEB_N1	2076	IBM	2145
4	ITSO_V7K_SITEA_N1	2076	IBM	2145
5	ITSO_V7K_SITEC_Q_N2	2076	IBM	2145

IBM_2145:ITSO_SVC_ESC:superuser>lsmdiskgrp

id	name	status	mdisk_count	vdisk_count	capacity	extent_size	free_capacity
0	V7000SITEA_RAID5	online	5	4	2.22TB	256	1.83TB
400.00GB 400.00GB 400.00GB 17 80 auto inactive							
no 0.00MB 0.00MB 0.00MB 0.00MB 0.00MB							
1	V7000SITEB_RAID5	online	5	4	2.22TB	256	1.83TB
400.00GB 400.00GB 400.00GB 17 80 auto inactive							
no 0.00MB 0.00MB 0.00MB 0.00MB 0.00MB							
2	V7000SITEC	online	1	0	512.00MB	256	512.00MB
0.00MB 0.00MB 0.00MB 0 80 auto inactive							
no 0.00MB 0.00MB 0.00MB 0.00MB 0.00MB							
3	V7000SITEA_RAID10	online	4	0	1.95TB	256	1.95TB
0.00MB 0.00MB 0.00MB 0 80 auto inactive							
no 0.00MB 0.00MB 0.00MB 0.00MB 0.00MB							
4	V7000SITEB_RAID10	online	4	0	1.95TB	256	1.95TB
0.00MB 0.00MB 0.00MB 0 80 auto inactive							
no 0.00MB 0.00MB 0.00MB 0.00MB 0.00MB							


```
18 ITSO_V7K_SITEB_SAS7 online managed 4 V7000SITEB_RAID10 500.0GB
0000000000000004 ITSO_V7K_SITEB_N2
600507680282018b300000000000080000000000000000000000000000000000000000 generic_hdd
```

```
IBM_2145:ITSO_SVC_ESC: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 ESXi_VMDK_2 0 ITSO_SVC_ESC_0 online many many 100.00GB many
600507680183053EF80000000000092 0 2 not_empty 0 no
0
1 ESXi_VMDK_3 1 ITSO_SVC_ESC_1 online many many 100.00GB many
600507680183053EF80000000000093 0 2 not_empty 0 no
0
3 ESXi_VMDK_4 0 ITSO_SVC_ESC_0 online many many 100.00GB many
600507680183053EF80000000000095 0 2 not_empty 0 no
0
46 ESXi_VMDK_1 1 ITSO_SVC_ESC_1 online many many 100.00GB many
600507680183053EF8000000000008C 0 2 not_empty 0 no
0
```

```
IBM_2145:ITSO_SVC_ESC:superuser>lsquorum
quorum_index status id name controller_id controller_name active
object_type override
0 online 1 ITSO_V7K_SITEA_SAS0 1 ITSO_V7K_SITEA_N2 no mdisk
yes
1 online 6 ITSO_V7K_SITEB_SAS0 0 ITSO_V7K_SITEB_N2 no mdisk
yes
2 online 11 ITSO_V7K_SITEC_QUORUM 2 ITSO_V7K_SITEC_Q_N1 yes mdisk
yes
```

From the CLI command output that is shown in Example 6-9 on page 161 you can see these aspects of the configuration:

- ▶ The clustered system is accessible through the CLI and it is in its stretched topology.
- ▶ The SAN Volume Controller nodes are online, and one of them is the configuration node.
- ▶ The I/O groups are in the correct state.
- ▶ The subsystem storage controllers are connected.
- ▶ The managed disk (MDisk) groups are online.
- ▶ The MDisks are online.
- ▶ The volumes are online.
- ▶ The three quorum disks are in the correct state.

Now, check the Volume Mirroring status by running a CLI command against each volume as shown in Example 6-10.

Example 6-10 Volume mirroring status check

```
IBM_2145:ITSO_SVC_ESC:superuser>lsvdisk ESXi_VMDK_1
id 46
name ESXi_VMDK_1
IO_group_id 1
IO_group_name ITSO_SVC_ESC_1
status online
mdisk_grp_id many
.
multiple lines omitted
.
copy_id 0
```

```

status online
sync yes
primary no
mdisk_grp_id 0
mdisk_grp_name V7000SITEA_RAID5
.
multiple lines omitted
.
copy_id 1
status online
sync yes
primary yes
mdisk_grp_id 1
mdisk_grp_name V7000SITEB_RAID5

multiple lines omitted

```

From the CLI command output in Example 6-10, you can see these aspects of the configuration:

- ▶ The volume is online.
- ▶ The storage pool name and the MDisk name are *many*, which means that Volume Mirroring is in place.
- ▶ Copy_id 0 is *online*, in *sync*.
- ▶ Copy_id 1 is *online*, in *sync*.

If you have several volumes to check, you can create a customized script directly from the Spectrum Virtualize shell. You can find useful scripts on the Storwize Scripting wiki on IBM developerWorks:

<http://ibm.co/1hdCYkA>

Critical event scenario analysis

In this scenario, the environment experienced a critical event that caused the complete loss of Site 1.

Follow these steps to get a complete view of any damage and to gather enough information about key elements to determine what your next recovery actions need to be:

1. Is IBM Spectrum Virtualize and SAN Volume Controller system management available through 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 System Storage Spectrum Virtualize Troubleshooting Guide*, GC27-2284.

2. Is system login possible?

YES: System is online, continue with Step 3.

NO: System is offline or has connection problems.

- a. Check your connections, cabling, and node front panel event messages.
- b. Verify the system status by using the Service Assistant menu or the node front panel. For more information, see *IBM System Storage Spectrum Virtualize Troubleshooting Guide*, GC27-2284.

3. Bring a part of the system online for further diagnostic tests:
 - a. Using a browser, connect to one of the SAN Volume Controller node's service IP addresses:


```
https://<service_ip_add>/service/
```
 - b. Log in with your cluster GUI password.
 - c. After login, you are redirected to the Service Assistant menu. From that menu, you can try to bring at least a part of the clustered system online for further diagnostic tests.
4. If the Spectrum Virtualize system management is available, run these checks:
 - a. Check the status by running the CLI commands that are shown in Example 6-11.

Example 6-11 lssystem example

```
IBM_2145:ITSO_SVC_ESC:superuser>lssystem
```

- b. Check the status of the nodes as shown in Example 6-12.

Example 6-12 Node status example

```
IBM_2145:ITSO_SVC_ESC:superuser>lspath
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
1 ITSO_SVC_NODE1_SITE_A 100006B119 500507680100B13F online 0
ITSO_SVC_ESC_0 yes 2040000006481049 CF8
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsovcnode1sitea 151580
151580
2 ITSO_SVC_NODE2_SITE_B 100006B074 500507680100B0C6 offline 0
ITSO_SVC_ESC_0 no 20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsovcnode2siteb 151523
151523
3 ITSO_SVC_NODE3_SITE_A 1000849047 50050768010027E2 online 1
ITSO_SVC_ESC_1 no 2040000204240107 8G4
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsovcnode3sitea 108283
108283
4 ITSO_SVC_NODE4_SITE_B 1000871173 50050768010037E5 offline 1
ITSO_SVC_ESC_1 no 20400002070411C3 8G4
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsovcnode4siteb 104643
104643
```

Observe the following statuses in Example 6-12:

- The *config node* role is still on Node 1 but might change in some cases where the lost node was the config node.
- Node 1 and 3 are online.
- Node 2 and 4 are offline.

During this event, the system lost 50% of the ESC system resources, but it is still up and running with the 50% of resources.

If the critical event was not a rolling disaster and stopped with the loss of Site 2, the host applications that were running on Site 1 can still run on this site. The applications that were running on Site 2 can be moved with host clustering or high availability (HA) functionality to Site 1.

Later, Site 2 can be recovered or rebuilt without any impact on the production systems in the same or another location.

In some scenarios, or in the case of a rolling disaster, connectivity to Site 3 might also be lost, and the site that apparently survived will be lost, too. This is because the critical event was triggered from this site (Site 2), and the first site that went offline was frozen by the disaster recovery feature and set as offline.

Assuming that Site 1 wins the quorum race and that the critical event was a rolling disaster that also affected Site 3 (where the active quorum is located), the cluster is stopped and needs to be recovered from the only site that is still physically available (in this example, Site 2). 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, “Recovery guidelines” on page 167.

6.3.3 Recovery guidelines

This section explores recovery scenarios. Regardless of the scenario, the common starting point is the complete loss of Site 1 or Site 2 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 that was introduced with software version 7.2, using the 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 ESC configuration

If you cannot recover the site in a reasonable time, you must take recovery actions. Consider these questions to determine the appropriate 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, extra SAN Volume Controller nodes, and extra SAN components. Consider these questions about the extra components:

- ▶ Do you plan to use new nodes that are supplied by IBM?
- ▶ Do you plan to reuse other, existing SAN Volume Controller nodes, which might be being 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 newly acquired SAN Volume Controller nodes and storage?
- ▶ Do you plan to use new back-end storage subsystems?
- ▶ Do you plan to configure free space on the surviving storage subsystems to host the space that is required for volume mirroring?

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 ESC 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 involves ordering and installing new SAN Volume Controller nodes, new storage subsystems, and new FC SAN devices before you begin the steps that follow.

This recovery action is based on a decision to recover the ESC infrastructure at the same performance characteristics as before the critical event. However, the solution has limited business continuity because the ESC is recovered at only one site.

Figure 6-3 shows the new recovery configuration.

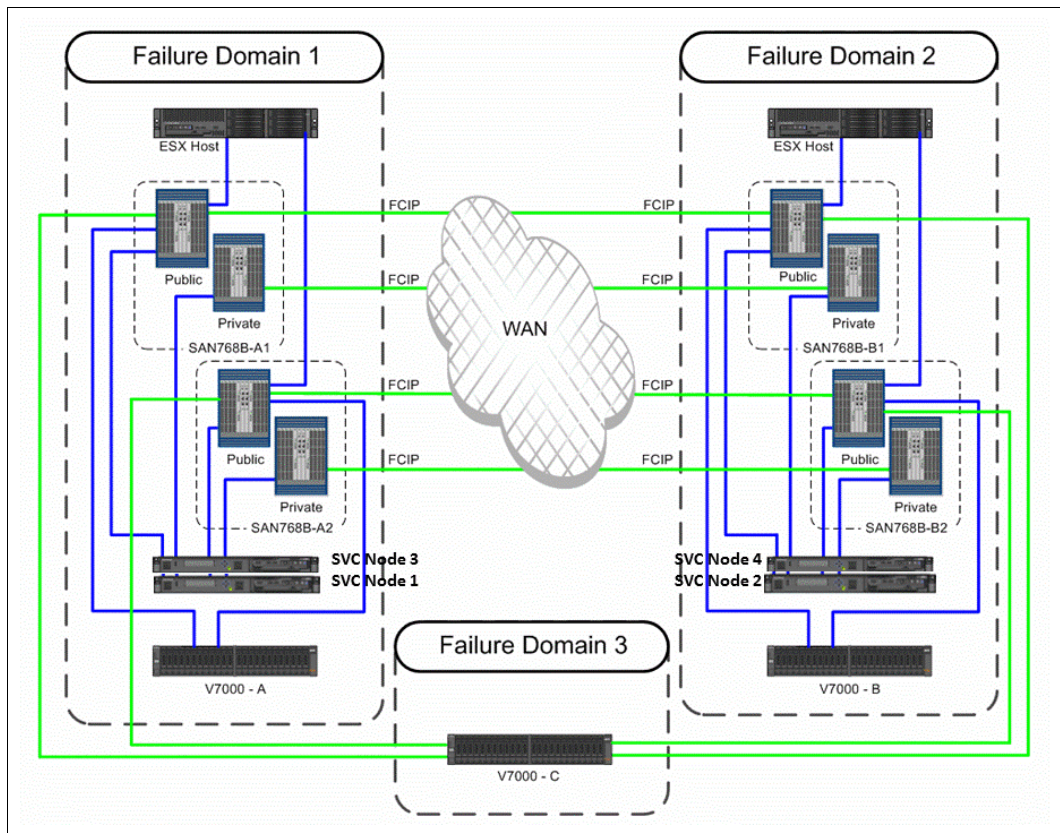


Figure 6-3 New recovery configuration in surviving site

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 these major steps:

1. Be sure that the active quorum disk is available and visible from the site that is not to be moved.
2. Disconnect the FCIP links between failure domains.
3. Uninstall and reinstall devices that you plan to reuse, and install all of the new devices in the new sites.

4. Reconnect the FCIP links between failure domains.

Steps to restore your ESC configuration in the same site

Complete the following steps to restore your ESC configuration as it was before the critical event in the same site. They are finished after you install the new devices.

Important: Before you perform any recovery action, be sure that the previous environment or site cannot come back to life with a device or nodes that still have earlier configuration data. This would 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).

1. Restore your back-end storage subsystem configuration as it was, starting from your backup. LUN masking can be done in advance because the SAN Volume Controller node's WWNN is already known.
2. Restore your SAN configuration exactly as it was before the critical event. You can do so 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.
3. If possible, connect the new storage subsystems to the same FC switch ports as before the critical event. SVC-to-storage zoning must be reconfigured to be able to see the new storage subsystem's WWNN. Previous WWNNs can be removed, but take care to remove the correct ones, because you have only one volume copy active.
4. Do not connect the SAN Volume Controller node FC ports yet. Wait until directed to do so by the SAN Volume Controller node's WWNN change procedure.

Note: All of the recovery action is run by using the CLI. At the time of writing, there is no support within the GUI to run the required recovery action.

5. Connect to one of your SAN Volume Controller nodes that is in the frozen site. You are going to recover that node with the CLI by using its Service IP address and running the **sainfo lsservicenodes** command, as shown in Example 6-13.

Example 6-13 sainfo lsservicenodes command

```
IBM_2145:ITS0_SVC_ESC:superuser>sainfo lsservicenodes
panel_name cluster_id      cluster_name      node_id node_name      relation
node_status error_data
151523      0000020061214FBE Cluster_10.17.89.251 2      ITS0_SVC_NODE2_SITE_B local
Starting 551 151580 108283
104643      0000020061214FBE Cluster_10.17.89.251 4      ITS0_SVC_NODE4_SITE_B cluster
Starting 551 151580 108283
```

As Example 6-13 on page 169 shows, the two nodes are in Starting state, with a 551 error code waiting for the 151580 and 108283 missing resources that are the names related to the missing nodes in the rolling disaster.

6. Run the **overridequorum** command as shown in Example 6-14. No return message is expected, just the prompt, and you will lose the connection in a few seconds.

Example 6-14 overridequorum command

```
IBM_2145:ITS0_SVC_ESC:superuser>satask overridequorum -force
```

Now a new ESC is created with just the available resources online and with the previous resources offline. That is because these resources are still in the configuration files and in the quorum disk that is used to re-create the cluster.

7. Run the **lssystem** and **lsnode** commands from the CLI, using your regular cluster management IP address to show the new cluster created and the online and offline resources, as shown in Example 6-15. The new cluster that is created has the same management IP address as the previous one.

Example 6-15 New cluster and the online and offline resources

```

IBM_2145:ITS0_SVC_ESC:superuser>lssystem
id 000020060E150B6
name ITS0_SVC_ESC
location local
.
multiple line omitted
.
console_IP 10.17.89.251:443
id_alias 000020060C14FBE
.
multiple lines omitted
.
topology stretched
topology_status recovered_site_2
rc_auth_method none

IBM_2145:ITS0_SVC_ESC:superuser>lsnode
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
1 ITS0_SVC_NODE1_SITE_A 100006B119          500507680100B13F offline 0
ITS0_SVC_ESC_0 no                2040000006481049 CF8
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsosvcnode1sitea 151580
151580
2 ITS0_SVC_NODE2_SITE_B 100006B074          500507680100B0C6 online 0
ITS0_SVC_ESC_0 yes                20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsosvcnode2siteb 151523
151523
3 ITS0_SVC_NODE3_SITE_A 1000849047          50050768010027E2 offline 1
ITS0_SVC_ESC_1 no                2040000204240107 8G4
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsosvcnode3sitea 108283
108283
4 ITS0_SVC_NODE4_SITE_B 1000871173          50050768010037E5 online 1
ITS0_SVC_ESC_1 no                20400002070411C3 8G4
iqn.1986-03.com.ibm:2145.cluster10.17.89.251.itsosvcnode4siteb 104643
104643

```

Example 6-15 on page 170 shows that the new cluster ID created with the same management IP address and the two SAN Volume Controller nodes that were in a Starting state with error code 551 are now online. Notice that the cluster topology status is now `recovered_site_2` and its topology is still stretched.

8. Remove the two offline SAN Volume Controller nodes with the **rmnode** command.
9. Remove the volume mirroring definitions. First, identify which copy ID is offline for each volume by using the **rmvdiskcopy** command or GUI.
10. Remove the offline storage pool. First, identify which pools are offline, and then remove them with the **rmmdiskgrp** command.

11. Power on the new supplied node, if you have not done that already. Set its Service IP address, but leave the FC cable disconnected.
12. Change the new nodes WWNN by using the following procedure to set the same WWNN as the lost nodes. It is important that this is done in this sequence. Also, do not change any zoning configuration in the surviving site:

Warning: The next procedure is for 2145-CG8 Nodes. Before starting be sure to check the relevant website in case there is more up to date instructions:

For 2145-CG8 Nodes:

http://www.ibm.com/support/knowledgecenter/en/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_replacingnodecg8modelsnondisrupttask.html

For 2145-DH8 Nodes:

http://www.ibm.com/support/knowledgecenter/en/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_replacingnodedh8modelsnondisrupttask.html

For 2145-SV1 Nodes:

http://www.ibm.com/support/knowledgecenter/en/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_replacingnode_sv1_nondisrup.html

- a. Power on the replacement node from the front panel with the Fibre Channel cables and the Ethernet cable disconnected.

You might receive error 540, "An Ethernet port has failed on the 2145," and error 558, "The 2145 cannot see the Fibre Channel fabric or the Fibre Channel card port speed might be set to a different speed than the Fibre Channel fabric." These errors are expected because the node was started with no fiber-optic cables connected and no LAN connection.

If you see Error 550, "Cannot form a cluster due to a lack of cluster resources," this node still thinks that it is part of a clustered system. If this is a new node from IBM, this error should not occur.

- b. Change the WWNN of the replacement node to match the WWNN that you recorded earlier by following these steps:
 - i. From the front panel of the new node, navigate to the **Node** panel, and then navigate to the **Node WWNN** panel.
 - ii. Press and hold the Down button, press and release the Select button, and then release the Down arrow button. Line one is Edit WWNN, and line two is the last five numbers of this new node's WWNN.
 - iii. Press and hold the Down button, press and release the Select button, and then release the Down button to enter WWNN edit mode. The first character of the WWNN is highlighted.

Tip: When you are changing the WWNN, you might receive error 540, "An Ethernet port has failed on the 2145," and error 558, "The 2145 cannot see the FC fabric or the FC card port speed might be set to a different speed than the Fibre Channel fabric." These errors are expected because the node was started with no fiber-optic cables connected and no LAN connection. However, if this error occurs while you are editing the WWNN, you are taken out of Edit mode with partial changes saved. You must then reenter Edit mode by starting again at Step b.

- iv. Press the Up or Down arrow to increase or decrease the character that is displayed. The characters wrap F to 0 or 0 to F.
- v. Press the left arrow button to move to the next field or the right navigation button to return to the previous field, and repeat Step b for each field. At the end of this step, the characters that are displayed must be the same as the WWNN that you recorded in Step a.
- vi. Press the Select button to retain the characters that you updated, and return to the WWNN panel.
- vii. Press the Select button again to apply the characters as the new WWNN for the node.

You must press the Select button twice as Steps vi and vii instruct you to do. After Step vi, it might seem that the WWNN is changed, but it is Step vii that applies the change.

- c. Ensure that the WWNN changed by repeating Step b.

13. Connect the node to the same FC switch ports as it was before the critical event.

This is the key point of the recovery procedure. Connecting the new SAN Volume Controller nodes to the same SAN ports and reusing the same WWNN avoids rebooting, rediscovering, and reconfiguring. This, in turn, avoids creating any effect from the host point of view as the lost disk resources and paths are restored.

Important: Do *not* connect the new nodes to different ports at the switch or director. Doing so causes port IDs to change, which can affect the hosts' access to volumes or cause problems when adding the new node back into the clustered system.

If you are not able to connect the SAN Volume Controller nodes 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

14. Issue the CLI command as shown in Example 6-16 to verify that the last five characters of the WWNN are correct. The new nodes are in candidate state, ready to be added to the SAN Volume Controller cluster.

Example 6-16 Verifying candidate node with the correct WWNN

```
IBM_2145:ITS0_SVC_ESC:superuser>lsnodecandidate
id                panel_name UPS_serial_number UPS_unique_id hardware
500507680100B13F 151580    100006B119      2040000006481049 CF8
50050768010027E2 108283    1000849047      2040000204240107 8G4
```

Important: If the WWNN does not match the original node's WWNN exactly as recorded, repeat steps 12b and 12c.

15. Add the node to the clustered system, and ensure that it is added back to the same I/O group as the original node with the CLI commands shown in Example 6-17.

Example 6-17 Adding a node

```
IBM_2145:ITS0_SVC_ESC:superuser>addnode -panelname 151580 -iogrp 0 -site 1
Node, id [5], successfully added
IBM_2145:ITS0_SVC_ESC:superuser>addnode -panelname 108283 -iogrp 0 -site 1
```


21. Consider running the quorum-assigning procedure to reassign the three-quorum disk according to your new back-end storage subsystem by using the GUI, as shown in Figure 6-4.

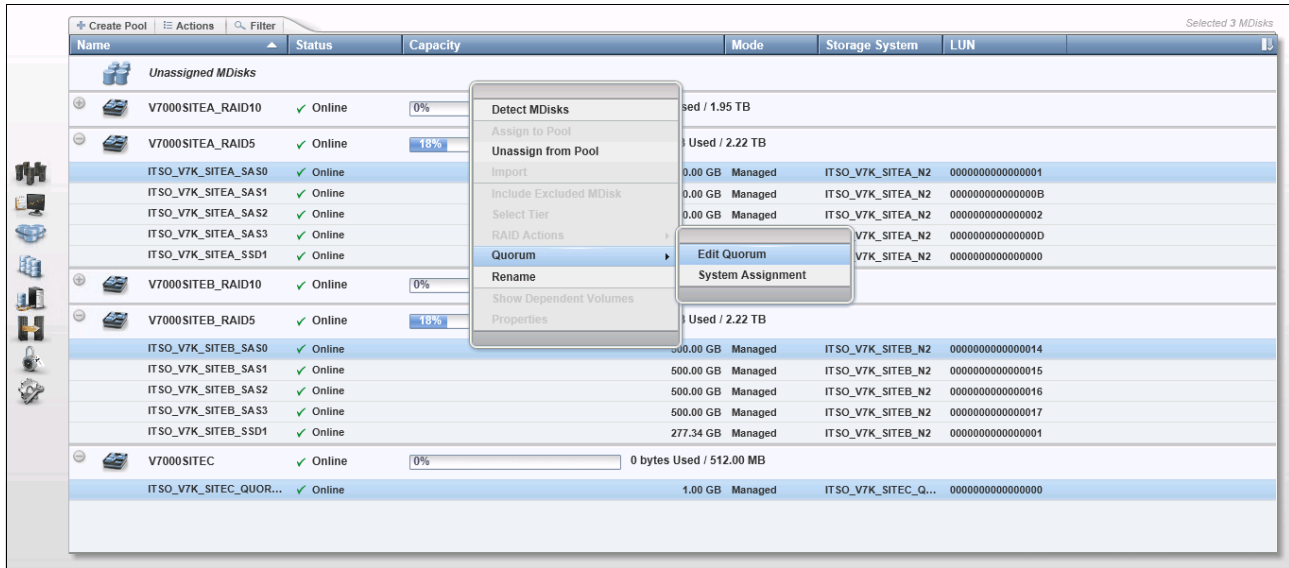


Figure 6-4 Quorum assignment by using the GUI

22. Check the quorum disk location with the command shown in Example 6-24.

Example 6-24 Isquorum command

```

IBM_2145:ITSO_SVC_ESC:superuser>lsquorum
quorum_index status id name controller_id controller_name active
object_type override
0 online 11 ITSO_V7K_SITEC_QUORUM 2 ITSO_V7K_SITEC_Q_N1 yes
mdisk yes
1 online 7 ITSO_V7K_SITEB_SAS1 0 ITSO_V7K_SITEB_N2 no
mdisk yes
2 online 3 ITSO_V7K_SITEA_SAS2 1 ITSO_V7K_SITEA_N2 no
mdisk yes

```

All of your volumes are now accessible from your host's point of view. The recovery action is finished, and the ESC is active again.

All of these operations are guidelines to help you in a critical event or a disaster. Some of them are specific to the example lab environment, and some are common to every ESC installation.

Make sure that you have a tested recovery plan in place, and always engage IBM Support at the earliest possible time if you need to initiate a recovery of any sort.

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 (be sure to check for update versions/editions) 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 V7.4*, SG24-7938
- ▶ *Implementing the IBM System Storage SAN Volume Controller V7.4*, 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>

VMware online resources

The following website provides additional VMware resources:

<http://www.vmware.com/support/pubs/>

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 TotalStorage Multipath Subsystem Device Driver User's Guide*, SC30-4096

Websites

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>
- ▶ Download site for Windows SSH freeware:
<http://www.chiark.greenend.org.uk/~sgtatham/putty>
- ▶ Open source site for SSH for Windows and Mac:
<http://www.openssh.com/windows.html>

- ▶ Cygwin Linux-like environment for Windows:
<http://www.cygwin.com>
- ▶ Sysinternals home page:
<http://www.sysinternals.com>
- ▶ Subsystem Device Driver download site:
<http://www.ibm.com/servers/storage/support/software/sdd/index.html>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



SG24-8211-01

ISBN 0738441104

Printed in U.S.A.

Get connected

