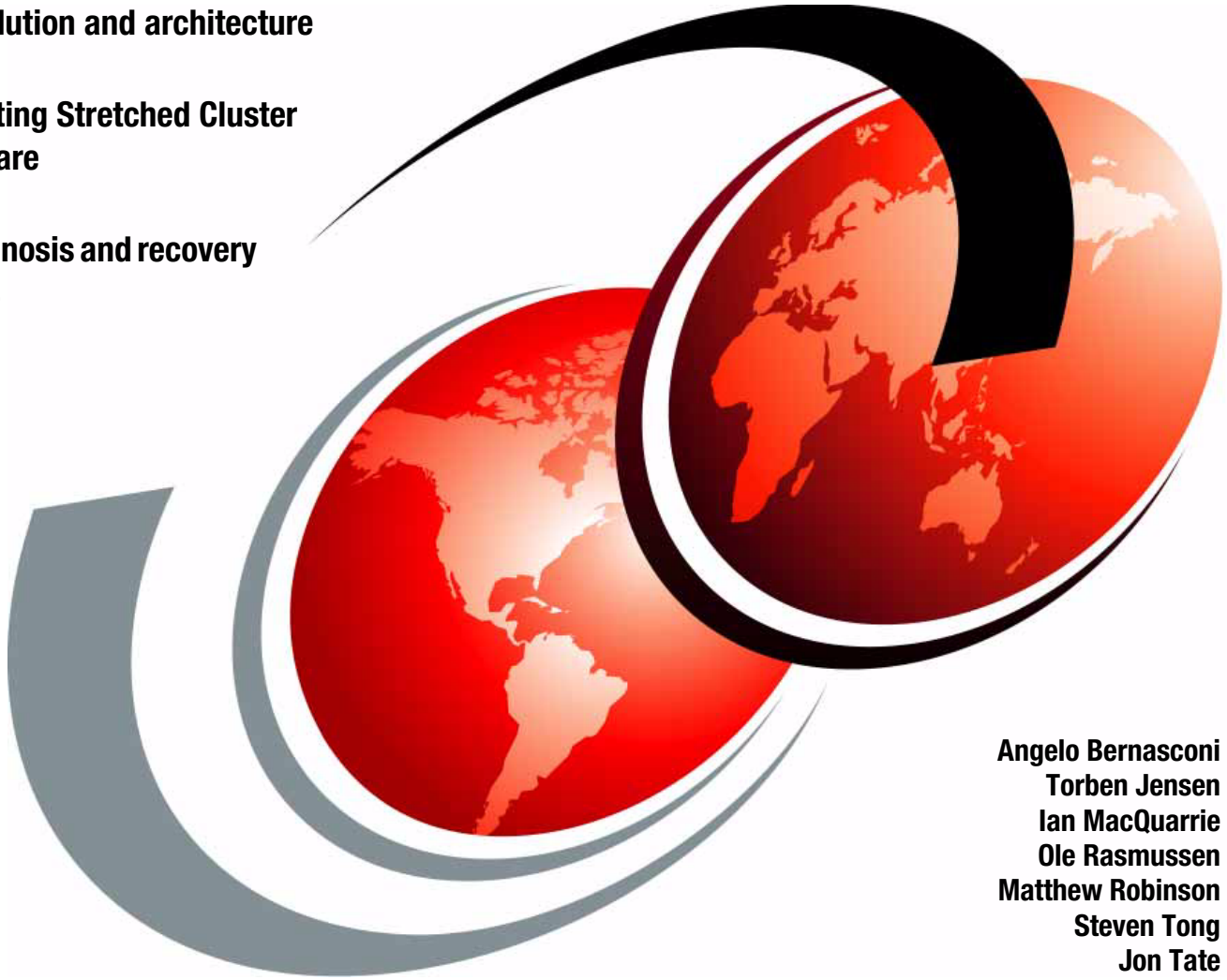
IBM

IBM SAN and SVC Stretched Cluster and VMware Solution Implementation

- Understanding the IBM Stretched Cluster solution and architecture
- Implementing Stretched Cluster with VMware
- Using diagnosis and recovery guidelines



Angelo Bernasconi
Torben Jensen
Ian MacQuarrie
Ole Rasmussen
Matthew Robinson
Steven Tong
Jon Tate

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM SAN and SVC Stretched Cluster and VMware Solution Implementation

April 2013

SG24-8072-00

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

First Edition (April 2013)

This edition applies to Version 6.4.0.2 of the IBM System Storage SAN Volume Controller. All other components are clearly identified in the book as appropriate.

Contents

Notices

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

Preface

This IBM® Redbooks® publication describes the IBM Storage Area Network and IBM SAN Volume Controller Stretched Cluster solution when combined with VMware. We describe guidelines, settings, and implementation steps necessary to achieve a satisfactory implementation.

Business continuity and continuous application availability are among the top requirements for many organizations today. Advances in virtualization, storage, and networking have made enhanced business continuity possible. Information technology solutions can now be designed to manage both planned and unplanned outages, and the flexibility and cost efficiencies available from cloud computing models.

IBM has designed a solution that offers significant functionality for maintaining business continuity in a VMware environment. This functionality provides the capability to dynamically move applications across data centers without interruption to those applications.

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

- The industry-proven VMware Metro vMotion
- IBM System Storage® SAN Volume Controller Stretched Cluster solution
- A Layer 2 IP Network and storage networking infrastructure for high performance traffic management
- DC interconnect

The team who wrote this book

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 System Storage SAN Solutions at the International Technical Support Organization, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 26 years of experience in storage software and management, services, and support, and is both an IBM Certified IT Specialist and an IBM SAN Certified Specialist. He is also the UK Chairman of the Storage Networking Industry Association.

Angelo Bernasconi is a Certified Senior Storage IT Specialist, in IBM, Italy. He has 26 years of experience in the delivery of maintenance and professional services for IBM Enterprise clients in z/OS® and Open Systems. He has a degree in Electronics and his areas of expertise include storage hardware, SAN, storage virtualization, de-duplication, and disaster recovery solutions. Angelo has written extensively about SAN and virtualization products in several IBM Redbooks and Redpaper™ publications.

Torben Jensen is an IT Specialist at IBM Global Technology Services®, Copenhagen, Denmark. He joined IBM in 1999 for an apprenticeship as an IT-System Supporter. From 2001 until 2005 he was the client representative for IBM's Internal Client platforms in Denmark. Torben joined the SAN/DISK for open systems department in March 2005. Torben provides daily and ongoing support to them as well as working with SAN designs and solutions for customers.

Ian MacQuarrie is a Senior Technical Staff Member with the IBM Systems and Technology Group located in San Jose, California. He has 26 years of experience in enterprise storage systems in a variety of test and support roles. He is currently a member of the Systems and Technology Group (STG) Field Assist Team (FAST) supporting clients through critical account engagements, availability assessments, and technical advocacy. His areas of expertise include storage area networks (SANs), open systems storage solutions, and performance analysis.

Ole Rasmussen is an IT Specialist working in IBM-SO Denmark, Copenhagen. He joined IBM during the transition of a large Danish bank's IT departments to IBM in 2004/2005, where he was working as a Technical Architect and Specialist. He has been working with the Customer Decentralized area since 1990, and his primary focus is on Windows and clustering. He has been working as an evangelist on 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.

Matthew Robinson is a storage pre-sales engineer for the IBM Systems and Technology Group Australia. Before joining IBM, Matthew completed a Bachelor of Information Technology with honours in Computer Science from Macquarie University, Australia. Matthew has 2 years of experience in implementing and designing IBM SAN Volume Controller and IBM Storwize® V7000 solutions.

Steven Tong is a corporate Solutions Architect who has been at Brocade for the last four and a half years. He is involved with developing and promoting joint partner solutions that integrate Brocade's extensive FC and IP networking portfolio. His current areas of focus include networking, storage, virtualization, and Big Data.

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

We would also like to thank the following people for their contributions:

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

Mary Connell
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:

Jim Baldyga
Silviano Gaona
Brian Steffler

Marcus Thordal
Steven Tong
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

1

Introduction

Business continuity and continuous application availability are among the top requirements for many organizations today. Advances in virtualization, storage, and networking have made enhanced business continuity possible. Information technology solutions can now be designed to manage both planned and unplanned outages, and the flexibility and cost efficiencies available from cloud computing models.

IBM has designed a solution that offers significant functionality for maintaining business continuity in a VMware environment. This functionality provides the capability to dynamically migrate applications across data centers without interruption to the applications.

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

- The industry-proven VMware Metro vMotion
- IBM System Storage SAN Volume Controller Stretched Cluster solution
- A Layer 2 IP Network and storage networking infrastructure for high performance traffic management
- DC interconnect

This chapter includes the following sections:

- SAN Volume Controller
- SAN Volume Controller Stretched Cluster solution
- SAN Volume Controller, Layer 2 IP Network, Storage Networking infrastructure, and VMware integration
- Open Data Center Interoperable Network (ODIN)

1.1 SAN Volume Controller

SAN Volume Controller is a storage virtualization system that enables a single point of control for storage resources. It helps support improved 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 are done quickly, efficiently and in real time, while also avoiding increases in administrative costs.

SAN Volume Controller supports attachment to servers through FC protocols and iSCSI protocols over IP networks at 1 Gbps and 10-Gbps speeds. These configurations can help reduce costs and simplify server configuration. SAN Volume Controller also supports FCoE protocol.

SAN Volume Controller combines hardware and software into 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 SAN Volume Controller cluster.

SAN Volume Controller configuration flexibility means that your implementation can start small and then grow with your business to manage very large storage environments.

SAN Volume Controller supports solid-state drives (SSDs) enabling scale-out high performance SSD support. The scalable architecture of this solution and the tight integration of SSDs enable businesses to take advantage of the high throughput capabilities of the SSDs. The scalable architecture is designed to deliver outstanding performance with SSDs for critical applications.

SAN Volume Controller also includes the IBM System Storage Easy Tier® function, which is designed to help 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 use of an SSD to the data that will benefit the most, helping deliver the maximum benefit even from small amounts of SSD capacity. SAN Volume Controller helps move critical data to and from SSDs as needed without application disruption.

SAN Volume Controller is designed to help increase the amount of storage capacity that is available to host applications. It does so by pooling the capacity from multiple disk systems within the SAN.

In addition, SAN Volume Controller combines various IBM technologies that include thin provisioning, automated tiering, storage virtualization, IBM Real-time Compression™, clustering, replication, multiprotocol support, and a next-generation graphical user interface (GUI). Together, these technologies are designed to enable SAN Volume Controller to deliver extraordinary levels of storage efficiency.

Because it hides the physical characteristics of storage from host systems, SAN Volume Controller help applications continue to run without disruption while you change your storage infrastructure. This advantage helps your business increase its availability to customers.

Figure 1-1 shows a high level overview of the SAN Volume Controller.

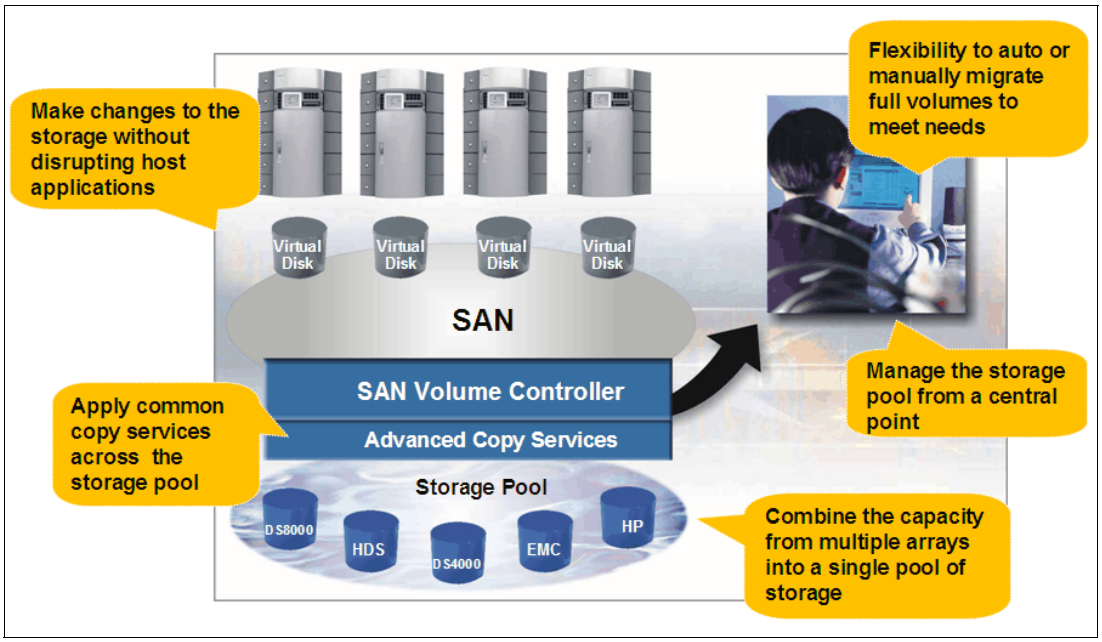


Figure 1-1 SAN Volume Controller

SAN Volume Controller includes a dynamic data migration function. This function is designed to move data from one storage system to another while still maintaining access to the data.

The SAN Volume Controller Volume Mirroring function is designed to store two copies of a volume on different storage systems. This function helps improve application availability in the event of failure or disruptive maintenance to an array or disk system. SAN Volume Controller is designed to automatically use whichever copy of the data remains available.

SAN Volume Controller is designed to enable administrators to 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 SAN Volume Controller systems at different locations. They help create copies of data for use in the event of 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 is designed to create an almost instant 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, significantly speeding application recovery.

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

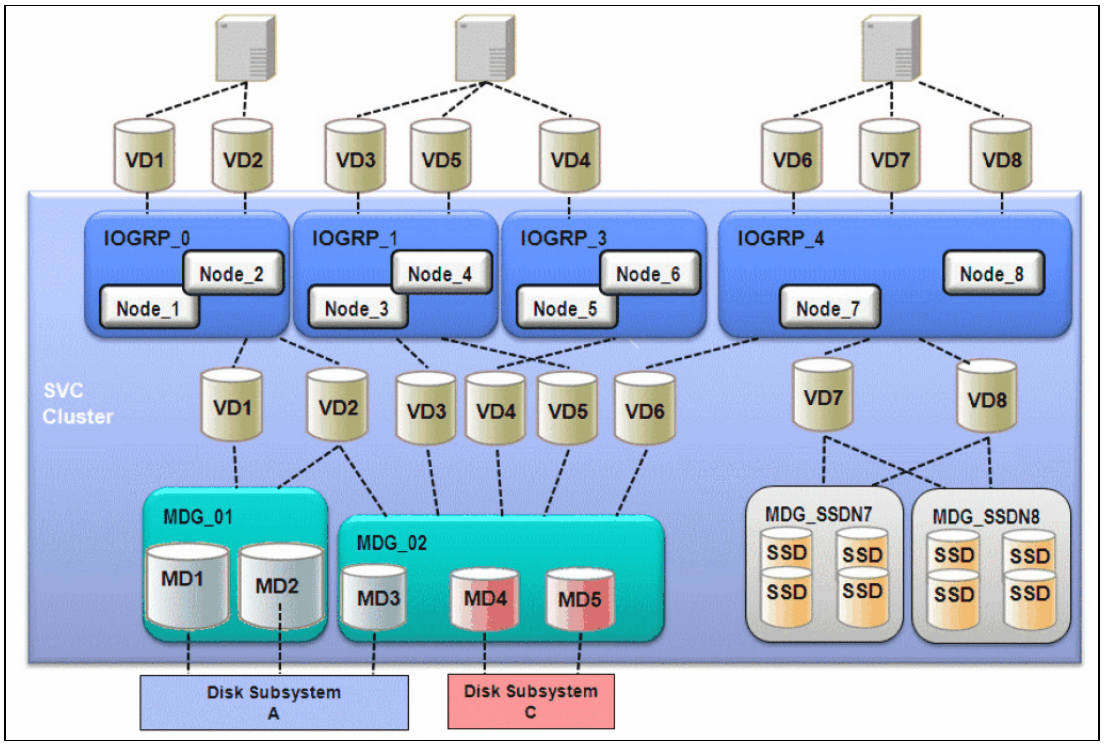


Figure 1-2 SAN Volume Controller structure

1.2 SAN Volume Controller Stretched Cluster solution

When SAN Volume Controller was first introduced, the maximum supported distance between nodes within an I/O Group was 100 m.

SAN Volume Controller 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.

With version 6.3, released in October 2011, SAN Volume Controller began supporting Stretched Cluster configurations where nodes can be separated by a distance of up to 300 km in specific configurations.

With Stretched Cluster, the two nodes in an I/O Group are separated by distance between two locations. A copy of the volume is stored at each location. This configuration means that 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 failover to a server at the alternate location and resume access to the disks. The SAN Volume Controller keeps both copies of the storage in synchronization, and the cache is mirrored between both nodes. Therefore, the loss of one location causes no disruption to the alternate location.

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. SAN Volume Controller is no exception. The SAN Volume Controller uses a tie-break mechanism that is facilitated through the implementation of a quorum disk. The SAN Volume Controller selects three quorum disks from the Managed Disks that are attached to the cluster to be used for this purpose. Usually the management of the quorum disks is transparent to the SAN Volume Controller users.

However, in a Stretched Cluster configuration, the location of the quorum disks must be assigned manually to ensure that the active quorum disk is in a third location. This configuration must be done to ensure the survival of one location in the event a failure occurs at another location.

The links between fabrics at either site have certain requirements that must be validated.

For more information about Stretched Cluster prerequisites, see:

http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102134

SAN Volume Controller is a flexible solution. You can use the storage controller of your choice at any of the three locations and with SAN Volume Controller they can be from different vendors. Also, this is all possible using the base SAN Volume Controller virtualization license with no additional charge.

SAN Volume Controller fully uses two major I/O functions that were introduced beginning with software release 4.3. These are Space-Efficient Virtual Disks (SEV), otherwise known as thin provisioning, and Virtual Disk Mirroring (VDM). The latter is a mechanism by which a single volume has two physical copies of the data on two independent Managed Disk Groups (storage pools, 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 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 enable disruptive maintenance to be run to a storage controller without any loss of access to servers and applications

Another capability that is provided by VDM is the ability to 'split' the cluster while still maintaining access to clustered servers and applications.

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 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. This process can now be done through the implementation of the SAN Volume Controller Stretched Cluster configuration.

The solution in this book focuses on the SAN Volume Controller and VMware environment. However, the SAN Volume Controller Stretched Cluster configuration can be applied to any other operating system and environment. These systems include native Microsoft Cluster, IBM AIX® Power HA, and Linux Cluster.

All the Stretched Cluster benefits and protection criteria apply, and use data protection and business continuity requirements regardless as to the operating system your application is using.

More detailed information about interoperability of the SAN Volume Controller Stretched Cluster configuration can be found at:

http://www-03.ibm.com/systems/storage/software/virtualization/svc/interop.html

1.3 SAN Volume Controller, Layer 2 IP Network, Storage Networking infrastructure, and VMware integration

Virtualization is now recognized as a key technology for improving the efficiency and cost effectiveness of a company's information technology 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 ESX servers that allows the live migration of virtual machines from one ESX server to another with no application downtime. Typically, vMotion is used to keep IT environments up and running, giving unprecedented flexibility and availability to meet the increasing demands for data. However, it is possible to migrate VMs between data centers with no downtime or user disruption.

IT can now run a secure migration of a live virtualized application and its associated storage between data centers with no downtime or user disruption. IT managers can therefore 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/environment. The following areas are key:

- Data synchronization between data centers, allowing servers, regardless of their location, to always have access to that data
- The network infrastructure that provides high performance, high reliability, and correct layer 2 extension capabilities to interconnect the data centers
- IP traffic management of client network access to the site where the application server is

The solution in this book addresses those three key areas through the combination of VMware Metro vMotion with SAN Volume Controller Stretched Cluster capabilities. This combination runs over Layer 2 IP Network and storage infrastructure.

Continuous access to data is provided by SAN Volume Controller Stretched Cluster 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 importantly, data protection and resiliency.

The combination of VMware Metro vMotion, SAN Volume Controller Stretched Cluster, 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 VMware's clustered file system, VMFS, to enable access to a virtual storage. The underlying storage that is used by the VMFS datastore is one or more volumes supplied by SAN Volume Controller that are accessible by all the vSphere hosts.

During vMotion, the active memory and precise execution 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, ensuring 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 new paradigm 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 best practices, see the "VMware Metro Storage Cluster: (vMSC)" white paper at:

http://ibm.biz/Bdx4gq

"Uniform host access configuration – When ESXi hosts from both sites are all connected to a storage node in the storage cluster across all sites. Paths presented to ESXi hosts are stretched across distance."

- Primary Benefit:
  - Fully Active/Active and workload-balanced data centers
  - Disaster/DownTime avoidance
- Secondary Benefit
  - Might be useful in a Disaster Recovery situation when combined with other processes

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

- Tier 0 applications, such as web servers in server farms
- Tier 1-3 application 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 VMware vMotion configuration.

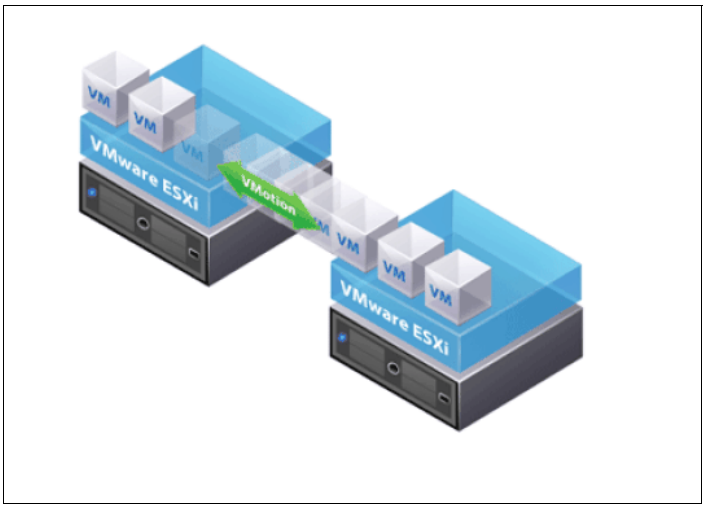


Figure 1-3 VMware vMotion

Benefits in detail

The following are the benefits of the application:

- 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 effects.

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

- User performance/load balancing between data centers

  In typical environments, a large percentage of data center capacity is set aside for “spikes” during peak demand. Backup/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 “on the fly” allows the use of external cloud resources to handle load during peak demand periods.

- Optimization to decrease power costs

  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 provides cost savings by using vSphere Distributed Power Management (DPM).

- Zero downtime maintenance

  Eliminating downtime during maintenance is a key advantage that vMotion over distance enables. Virtual machines can be relocated to a remote data center during maintenance windows, allowing users to enjoy continuous access to applications in a fully transparent manner.

- Disaster Recovery test

  A disaster recovery test can be run on live applications without affecting the business. Additionally, this process allows you to completely test the functionality of the DR site with an actual user load.

Figure 1-4 shows VMware and SAN Volume Controller Stretched Cluster.

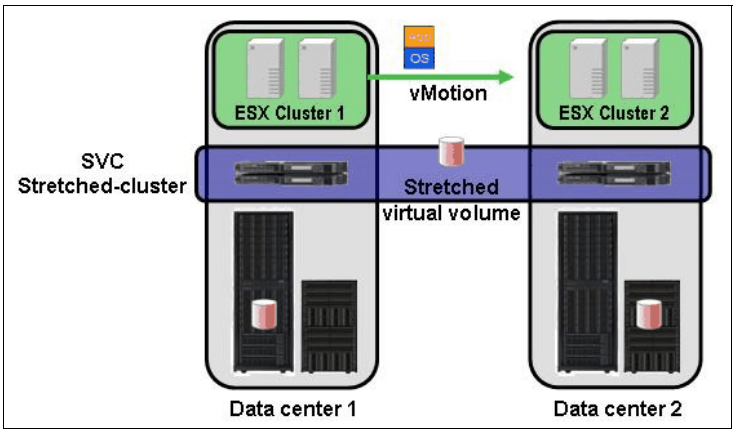


Figure 1-4 SAN Volume Controller Stretched Cluster and VMware

When to use VMware Stretched Clusters

Use VMware Stretched clusters in these situations:

- When there is a requirement for inter-site 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 RTT.
- When you are enabling multi-site 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.

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 there are environments that 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 Best practice guide at:

http://ibm.biz/Bdx4gy

1.3.2 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
- IBM SAN Volume Controller Stretched Cluster 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 Description” 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 Description” on page 13.

1.4 Open Data Center Interoperable Network (ODIN)

IBM has taken the lead to articulate the issues that are faced by agile data center networks and promote best practices. It has done so by creating a roadmap that is based on industry standards that is known as ODIN.

IBM believes that the practical, cost-effective evolution of data center networks is 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 currently address issues such as virtualization and virtual machine (VM) migration across flat Layer 2 networks, lossless Ethernet, Software-Defined Networking (SDN) and OpenFlow, and extended distance WAN connectivity (including low latency).

OLDIN provides the following benefits:

- Customer choice and future proof designs for data center networks, including vendor-neutral Requests for Quotation (RFQs)
- Lowers total cost of ownership (TCO) by enabling a multi-vendor network (for more information, see *“Gartner Group, “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

IBM has worked to develop solutions that are based on open industry standards:

- Fibre Channel and FC-IP storage technologies
- Lossless Ethernet
- Flat, Layer 2 Networks
- Distance extension options by using dark fiber WDM and MPLS/VPLS

The ODIN technical briefs can be found at:

http://www.ibm.com/systems/networking/solutions/odin.html

Figure 1-5 shows typical ODIN solutions.

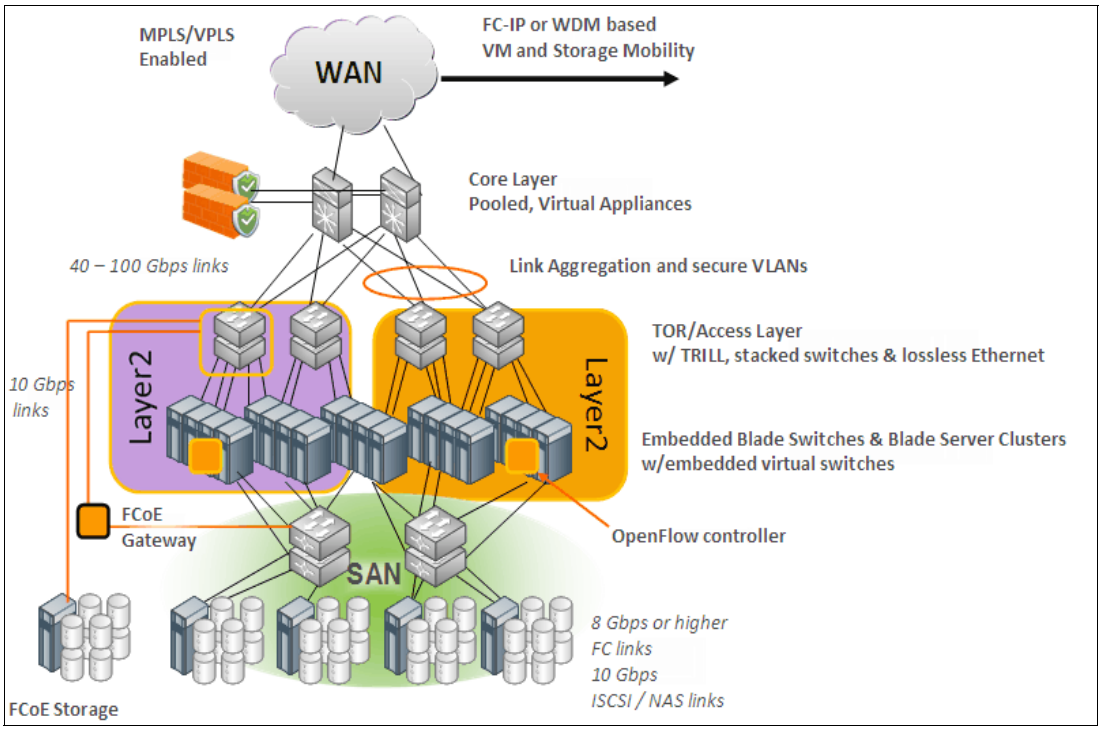


Figure 1-5 ODIN

ODIN has these major targets of ODIN:

- Support customer choice and future-proof network design choices, including vendor-neutral RFQs
- Lower total cost of ownership (TCO) by 15 – 25% as explained in *"Gartner Group, "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 that have recently emerged
- 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 SAN Volume Controller Stretched Cluster solution complies with ODIN.

2

Hardware and Software Description

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

This chapter includes the following sections:

- Hardware description
- IBM System Storage 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 is needed when you implement a SAN Volume Controller Stretched Cluster. All of the products that are mentioned can provide the functionality that is needed to implement a Stretched Cluster. It is up to you to choose the most suitable product for your environment.

Consider these hardware factors when you are implementing SAN Volume Controller Stretched Cluster:

- Distance/Latency between data centers
- Connectivity between data centers
- Bandwidth of sent data
- Customer budget
- Current customer infrastructure

All of these considerations can result in greatly varied hardware requirements. This section addresses some hardware possibilities and guidelines on what features to purchase with that hardware.

2.2 IBM System Storage SAN Volume Controller

As described in the Chapter 1, “Introduction” on page 1, this solution requires the use of the IBM System Storage SAN Volume Controller. The SAN Volume Controller provides an active-active storage interface that can allow for simple fail over and fail back capabilities in a site disruption or failure. The IBM Storwize V7000 *cannot* be configured in a Stretched Cluster configuration.

To implement a Stretched Cluster solution over 4 km, use either the 2145-CF8 or 2145-CG8 hardware models of the SAN Volume Controller controllers, as seen in Figure 2-1. Use these models because of their 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 contact or see Chapter 3, “SAN Volume Controller Stretched Cluster Architecture” on page 25 to ensure that the SAN Volume Controller node model and firmware version are supported by the solution you want to implement.

Figure 2-1 SAN Volume Controller CF8 nodes

2.3 SAN directors and switches

To implement an IBM System Storage SAN Volume Controller Stretched Cluster solution, any SAN fabrics must be extended across two data centers or failure domains. How you want to extend this fabric depends on the distance between failure domains. Your choices of architecture are outlined in Chapter 3, “SAN Volume Controller Stretched Cluster Architecture” on page 25.

This section does not address any particular WDM devices or any Ethernet infrastructure options other than FCIP devices. All of the hardware that is described is compatible with CWDM devices (by using colored SFPs), 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 are designed to increase business agility while also providing nonstop access to information and reducing infrastructure and administrative costs. The SAN768B-2 and SAN384B-2 fabric backbones, 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 robust, reliable, high-performance technology.

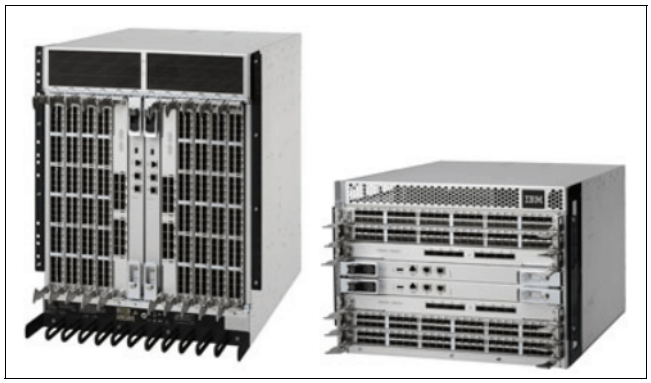


Figure 2-2 SAN768B-2 and SAN384B-2 fabric backbones

Both directors are capable of 16, 10, 8, 4, and 2-Gbps connections with the capability to have up to 512 or 256 ports. Included with the purchase of the directors is an enterprise software bundle that includes the Extended Fabrics and Trunking features. The Extended Fabrics feature is essential for implementing Stretched Cluster solutions 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 the IBM System Storage SAN384B-2 and SAN768B-2 directors, see:

http://www-03.ibm.com/systems/networking/switches/san/b-type/san768b-2/index.html

2.3.2 SAN24B-5, SAN48B-5, and SAN80B-4 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, and16 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 Stretched Cluster 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:

http://www-03.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) is configurable 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.

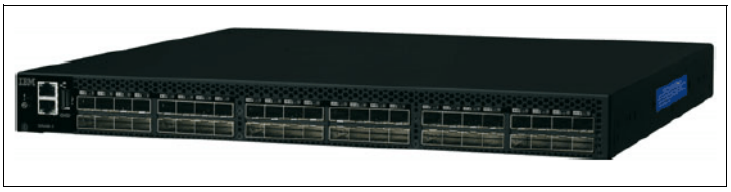


Figure 2-4 SAN48B-5 switch

When you are implementing Stretched Cluster 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:

http://www-03.ibm.com/systems/networking/switches/san/b-type/san48b-5/index.html

SAN80B-4 switch

The IBM System Storage SAN80B-4 switch, as shown in Figure 2-5, is configurable with 48, 64, or 80 active ports. It 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.

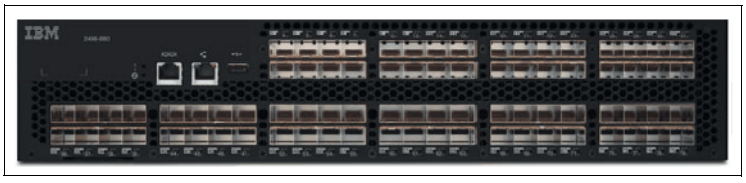


Figure 2-5 SAN80B-4 switch

The Extended Fabric feature is enabled on the SAN80B-4 by default, which makes it useful for Stretched Cluster configurations over 10 km. Because this switch is suited to larger

environments, the Trunking Activation license might have to be purchased to ensure bandwidth between failure domains.

For more information about the IBM System Storage SAN80B-4 switch, see:

http://www-03.ibm.com/systems/networking/switches/san/b-type/san80b-4/index.html

2.4 FCIP routers

When you are implementing a Stretched Cluster over long distances, it is not always possible or feasible to extend SAN fabrics by using direct Fibre Channel connectivity or Wavelength Division Multiplexing (WDM). Either the distance between the two failure domains is over 10 km, or it is too expense to lay cable or hire dark fiber.

Many dual data center environments already have existing IP connections between data centers. This configuration allows FCIP technologies to be used to enable the SAN fabric to extend across data centers while also using existing infrastructure. When you are implementing SAN Volume Controller Stretched Cluster solutions with FCIP, there are minimum bandwidth requirements that must be met so that the solutions are supported. For more information, see Chapter 3, “SAN Volume Controller Stretched Cluster Architecture” on page 25.

2.4.1 8 Gbps Extension Blade

The 8 Gbps Extension Blade is an FCIP blade, as shown in 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.

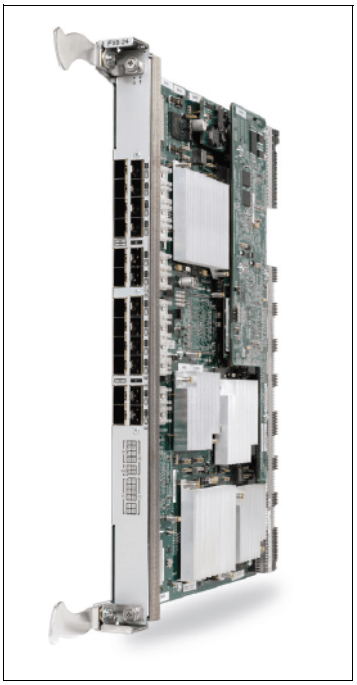


Figure 2-6 8 Gbps Extension Blade

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 10GbE Activation feature on the SAN384B-2 and SAN768B-2 directors, you can have two 10GbE ports or ten 1 GbE Ethernet ports and one 10 GbE port on the blade. Also, when you order this blade, the 8 Gbps Advanced Extension Activation on the SAN384B-2 and SAN 768B-2 directors feature must be ordered.

2.4.2 SAN06B-R extension switch

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

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, the R06 8 Gbps Advanced Extension feature, or the R06 Enterprise Package must be ordered.

For more information about the IBM System Storage SAN06B-R Extension switch, see:

http://www-03.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 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 MPLS/VPLS/VLL technology.

2.5.1 IBM System Networking switches

The following networking switches can be used in a SAN Volume Controller Stretched Cluster.

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, and 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.

The IBM RackSwitch G8124E is shown in Figure 2-8.

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 supporting IBM Virtual Fabrics. These characteristics help you reduce the number of I/O adapters to a single dual-port 10 Gb adapter, reducing cost and complexity.

Figure 2-9 shows the IBM RackSwitch G8264T.

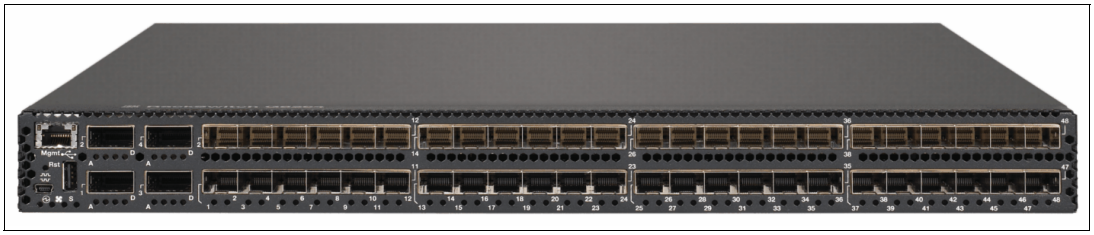


Figure 2-9 IBM RackSwitch G8264T

The RackSwitch G8264 supports up to 48 × 1/10 Gb SFP+ ports and 4 × 40 Gb QSFP+ ports. The RackSwitch G8264T supports up to 48 × 10GBase-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 a SAN Volume Controller Stretched Cluster.

Brocade MLX Series routers

The Brocade MLX Series of high-performance routers provides a rich set of high-performance IPv4, IPv6, and Multiprotocol Label Switching (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/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 is a family of compact 1U, multi-service edge/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 NetIron CES2000

The Brocade NetIron CES 2000 Series 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 Deliver Controllers

The Brocade ServerIron ADX Series, as seen 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 Brocade Application Resource Broker.

Application Resource Broker is a plug-in for vSphere that enables automated provisioning of 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 ServerIron 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- and 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 a Stretched Cluster, 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 ESX and vSphere platforms provide. This section outlines VMware ESX 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 virtual machines (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 ESX hosts and storage. These features can add to the availability of the VMs 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 datastores to allow scheduled maintenance procedures to

proceed without impacting 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 SAN Volume Controller Stretched Cluster solution, this feature is useful for moving VMs between two failure domains. You might need to move VMs to load balance across failure domains or because a failure domain needs an outage for maintenance.

For more information about vSphere vMotion, see:

http://www.vmware.com/products/vmotion/overview.html

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 copy of data from one datastore to another. With vStorage APIs for Array Integration (VAAI), this process can be offloaded to the storage subsystem, saving resources on both the vCenter host and data network.

In an SAN Volume Controller Stretched Cluster solution, this feature is useful for moving a VM's 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 migrate off a storage device that is becoming obsolete or is undergoing maintenance.

For more information about vSphere Storage vMotion, see:

http://www.vmware.com/products/storage-vmotion/overview.html

2.7.3 vSphere High Availability (HA)

vSphere HA provides cost effective, automated restart within minutes for applications in the event of hardware or operating system failures. With the addition of Fault Domain Manager, VMware HA is more reliable in operation, more easily scalable in its ability to protect virtual machines, and can provide increased uptime.

For more information about vSphere High Availability, see:

http://www.vmware.com/products/high-availability/overview.html

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 will meet availability objectives. At the time of a failure domain fail over 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).

Before you implement Site Recovery Manager, ensure that the firmware version of the SAN Volume Controller nodes is supported by the version of Site Recovery Manager that you plan to use. Table 2-1 shows Site Recovery Manager compatibility for Version 4 and 4.1.

Table 2-1 Versions of Site Recovery Manager supported by Stretched Cluster nodes

Hardware Model	SAN Volume Controller Firmware	Site Recovery Manager Version
IBM 2145-CF8	5.1.x, 6.1.x, 6.2.x, 6.3.x	4.0, 4.1
IBM 2145-CG8	6.2.x, 6.3.x	4.0, 4.1

For version 5 and later, see the IBM SAN Volume Controller compatibility matrixes at:

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

For more information about VMware vCenter Site Recovery Manager, see:

http://www.vmware.com/products/site-recovery-manager/overview.html

2.7.5 VMware Distributed Resource Scheduler (DRS)

VMware DRS dynamically balances computing capacity across a collection of hardware resources that are aggregated into logical resource pools. It continuously monitors utilization 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 migrates and allocates resources by using a set of user-defined rules and policies. These rules and policies can be used to prioritize critical or high performing VMs, ensure particular VMs never run on the same storage or host, or to save on power and cooling costs by powering off ESX servers that are not currently needed.

For more information about VMware Distributed Resource Manager, see:

http://www.vmware.com/products/drs/overview.html

2.7.6 VMware vCenter Server Heartbeat

vCenter Server Heartbeat provides unified protection for vCenter Server and its components against the broadest range of potential outages. They are protected from application, operating system, hardware, and network failures as well as external events regardless of whether vCenter Server is deployed on a physical or virtual machine.

vCenter Server Heartbeat continuously replicates information from vCenter Server to a passive standby server for rapid recovery. Data locations, registry entries, license data, and databases are all cloned automatically. When there is an availability threat or a planned maintenance window, administrators can switch over vCenter Server and all of its components from the primary server to a secondary server.

For more information about VMware vCenter Server Heartbeat, see:

http://www.vmware.com/products/vcenter-server-heartbeat/features.html

3

SAN Volume Controller Stretched Cluster Architecture

This chapter focuses on the SAN Volume Controller architecture as it applies to the Stretched Cluster configuration. It assumes a base understanding of the general SAN Volume Controller architecture.

For information about general SAN Volume Controller architecture and implementation, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

This chapter includes the following sections:

- Stretched Cluster overview
- Failure domains
- SAN Volume Controller volume mirroring
- SAN Volume Controller Stretched Cluster configurations
- Fibre Channel settings for distance
- SAN Volume Controller I/O operations on mirrored volumes

3.1 Stretched Cluster overview

In a standard SAN Volume Controller configuration, all nodes are physically located within the same rack. Beginning with version 5.1, support was provided 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, which provides protection against failures that affect a single failure domain. The initial support for Stretched Cluster that was delivered in version 5.1 contained the restriction that all communication between SAN Volume Controller node ports cannot traverse ISLs. This limited the maximum supported distance between failure domains. Starting with SAN Volume Controller 6.3, the ISL restriction was removed, which allowed the distance between failure domains to be extended to 300 km. Additionally, in SAN Volume Controller 6.3, the maximum supported distance for non-ISL configurations was extended to 40 km.

The SAN Volume Controller Stretched Cluster configuration provides a continuous availability platform whereby host access is maintained in the event of the loss of any single failure domain. This availability is accomplished through the inherent active/active architecture of 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.

3.2 Failure domains

In a Stretched Cluster configuration, the term *failure domain* is used to identify components of the SAN Volume Controller cluster that are contained within a boundary such that any failure that occurs (such as power failure, fire, and flood) is contained within that boundary. The failure therefore cannot propagate or affect components that are outside of that boundary. The components that comprise a Stretched Cluster 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 a failure of any single power source were to occur.
- 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 Stretched Cluster configuration would be 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 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 separate failure domains. Each I/O Group within the SAN Volume Controller cluster must adhere to this rule.

3.3 SAN Volume Controller volume mirroring

The Stretched Cluster configuration uses the SAN Volume Controller volume mirroring function. Volume mirroring allows the creation of one volume with two copies of MDisk extents. The two data copies, if placed in different MDisk Groups, allow volume mirroring to eliminate impact to volume availability if one or more MDisks fail. The resynchronization between both copies is incremental and is started by the SAN Volume Controller automatically. A mirrored volume has the same functions and behavior as a standard volume.

In the SAN Volume Controller software stack, volume mirroring is below the cache and copy services. Therefore, FlashCopy, Metro Mirror, Global Mirror have no awareness 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/shrink.

As with non-mirrored volumes, each mirrored volume is owned by the preferred node within the I/O Group. Thus the mirrored volume will go offline if the I/O Group goes offline. The preferred node runs all write operations to the backend disks for both copies of the volume mirror. Read operations can be run from either node in the I/O Group. However, all read operations are run only from the primary copy of the volume mirror.

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. Thus, an active quorum disk is required for volume mirroring. To ensure data consistency, SAN Volume Controller disables mirrored volumes if access to all quorum disk candidate is lost. Therefore, quorum disk availability is critical for Stretched Cluster 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. Thus, a bitmap memory space of 1 MB is required for each 2 TB of mirrored volume capacity.

3.3.2 Read operations

Volume mirroring implements a read algorithm with one copy that is designated as the primary for all read operations. SAN Volume Controller 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
```

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 SAN Volume Controller and the storage controller backend. The SAN Volume Controller cache is able to hide high backend response times from the host up to a certain level.

If a backend write fails or a copy goes offline, a bitmap is used to track out of sync grains. As soon as the missing copy is back online, SAN Volume Controller evaluates the change bitmap and runs an automatic resynchronization of both copies. The resynchronization process has a similar performance impact on the system as a FlashCopy background copy or volume migration. The resynchronization bandwidth can be controlled with the command `chvolume -syncrate`. Volume access is not impacted by the resynchronization process and is run concurrent with host I/O.

The write behavior for the mirrored copies can cause difficulties when there is a loss of a failure domain and therefore must be considered. Beginning with version 6.2, SAN Volume Controller provides the volume attribute `-mirrorwritepriority` 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
```

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

Figure 3-1 illustrates the data flow for write I/O processing. The host writes the data to the preferred node (1). The data is then mirrored to the partner node in the I/O Group (2), and then destaged to both mirror copies from the preferred node (3).

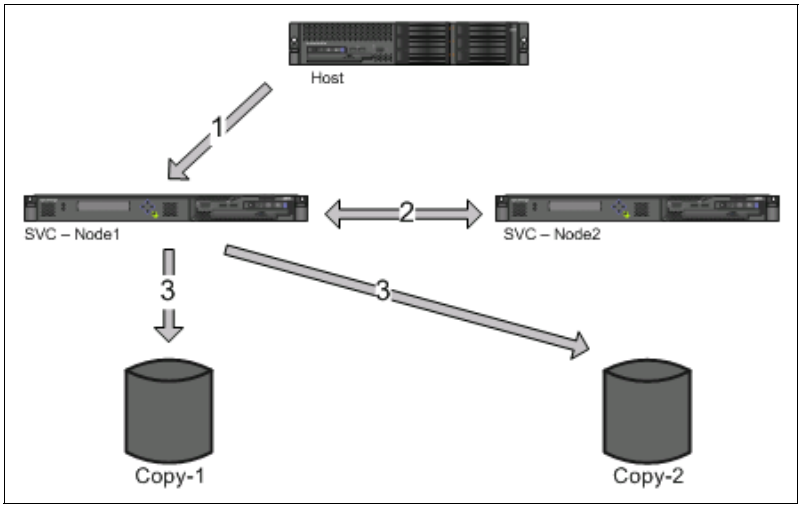


Figure 3-1 Volume Mirror: Write I/O processing

Table 3-1 summarizes the behavioral differences the **`mirrorwritepriority`** options have if slow writes were to occur to the remote volume copy (Copy-2).

Table 3-1 -mirrorwritepriority options

-mirrorwritepriority redundancy	**-mirrorwritepriority latency**
Write destage is complete after both mirror copies are updated.	Write destage is complete after Copy-1 is updated.
Cache pages are released after both copies are destaged.	Cache pages are released after Copy-1 is destaged.
Potential serialization of host I/O can occur because of cache hold for slow writes to Copy-2.	Host I/O remains asynchronous for slow writes to Copy-2.
Slow write response times on Copy-2 do not cause the mirror to go out of sync.	Slow write response times to Copy-2 (exceeding 5 seconds) cause the mirror to go out of sync. The system will stop using Copy-2 for 4-6 minutes. After this time, writes to Copy-2 resume and the synchronization process begins.
Potential performance impact for slow writes to Copy-2.	No performance impact for slow writes to Copy-2.

For write operations, consider that a failure that affects one of the failure domains might occur before or after the data is destaged from cache.

The following are the differences in behavior for these two conditions in relation to the **-mirrorwritepriority** setting:

- With **-mirrorwritepriority latency**:
  - If the failure occurs before write cache destaging, write data is still in the cache of the partner node and will be destaged as soon as possible. The volume remains online.
  - If the failure occurs after write cache destaging, data has been destaged to Copy-1. Cache pages have been released, but the write to Copy-2 has not yet occurred and is therefore not synchronized with Copy-1. The volume goes offline.
- With **-mirrorwritepriority redundancy**:
  - If the failure occurs before write cache destaging, write data is still in the cache of the partner node and will be destaged as soon as possible. The volume remains online.
  - If the site failure occurs after write cache destaging, both copies are updated and in sync. The volume remains online.

Tip: redundancy is the preferred setting for **-mirrorwritepriority** in an SAN Volume Controller Stretched Cluster configuration.

3.3.4 SAN Volume Controller quorum disk

The quorum disk fulfills two functions for cluster reliability:

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

The SAN Volume Controller 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 tie-breaker 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.

3.3.5 SAN Volume Controller 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 ½ votes for determining cluster quorum.

The SAN Volume Controller cluster manager implements a dynamic quorum. This configuration means that following a loss of nodes, if the cluster is able to continue operation, it dynamically moves the quorum disk to allow more node failure to be tolerated. This process improves the availability of the central cluster metadata, which 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 hence the nodes in the cluster, can completely change. The cluster can migrate onto

a completely separate set of nodes from the set on which it started. Within a SAN Volume Controller cluster, the quorum is defined in one of the following ways:

- More than half the nodes in the voting set are available.
- Exactly half the nodes in the voting set and the quorum disk from the voting set. When there is no quorum disk in the voting set, exactly half of the nodes in the voting set if that half includes the node that is displayed first in the voting set (configuration node).
- When there is no quorum disk in the voting set, exactly half of the nodes in the voting set if that half includes the node that is displayed first in the voting set (configuration node).

In an SAN Volume Controller Stretched Cluster configuration, the voting set is distributed across the failure domains. Figure 3-2 summarizes the behavior of the SAN Volume Controller cluster as a result of failures that affected the failure domains.

Failure Domain 1 Node 1	Failure Domain 2 Node 2	Failure Domain 3 Quorum disk	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 Quorem disk moved
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
Operational	Failed	Failed	Stopped
Failed	Operational	Failed	Stopped

Figure 3-2 SAN Volume Controller Stretched Cluster behavior

Tip: No manual fail-over or fail-back activities are required. The SAN Volume Controller is an active/active architecture that does not run fail-over/fail-back operations.

3.3.6 Quorum disk requirements and placement

Because of the quorum disk's critical role in the voting process, quorum functionality is not supported for internal drives on SAN Volume Controller nodes. Internal disks would not be able to act as a tie-breaker if access to the node was lost. Therefore, only Managed Disks (MDisks) on external storage can be selected as SAN Volume Controller 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 a failure that affects any single failure domain.

The SAN Volume Controller selects the first three MDisks from external storage controllers as quorum disk candidates. It also reserves some space on each of these disks by default. SAN Volume Controller does not verify whether the MDisks are from the same storage controller or from different storage controllers. To ensure that the quorum disk candidates and the active quorum disk are assigned to the correct failure domains, the quorum disk candidates must be manually assigned. Do so by using the **chquorum** command. Additionally, by default the SAN Volume Controller dynamically manages the location of the active quorum disk, and from time to time might change it from one candidate to another. Because of this behavior. you must disable the dynamic quorum selection for Stretched Cluster configuration by using the **-override** flag for all three quorum disk candidates:

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

The storage controller that provides the quorum disk in a Stretched Cluster configuration in the third failure domain must be supported as an *extended quorum disk.* Storage controllers that provide extended quorum support are listed at:

http://www.ibm.com/storage/support/2145

Requirement: 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 MByte/s.

Important: The following are quorum disk candidate requirements for SAN Volume Controller Stretched Cluster configuration:

- The SAN Volume Controller Stretched Cluster 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 failure domain 3.
- Dynamic quorum selection must be disabled by using **`chquorum`** command.
- Quorum disk candidates and the active quorum disk assignment must be done manually by using the **`chquorum`** command.

3.3.7 Failure scenarios in SAN Volume Controller Stretched Cluster configuration

Figure 3-3 on page 33 is used to illustrate 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 SVC Node 1 and SVC Node 2.
- Power off SVC Node 1 in Failure Domain 1: SVC Node 2 takes over the load and continues processing host I/O. SVC Node 2 changes the cache mode to write-through to avoid data loss in case SVC Node 2 also fails.
- Power off Storage System V7000-A: SAN Volume Controller 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.

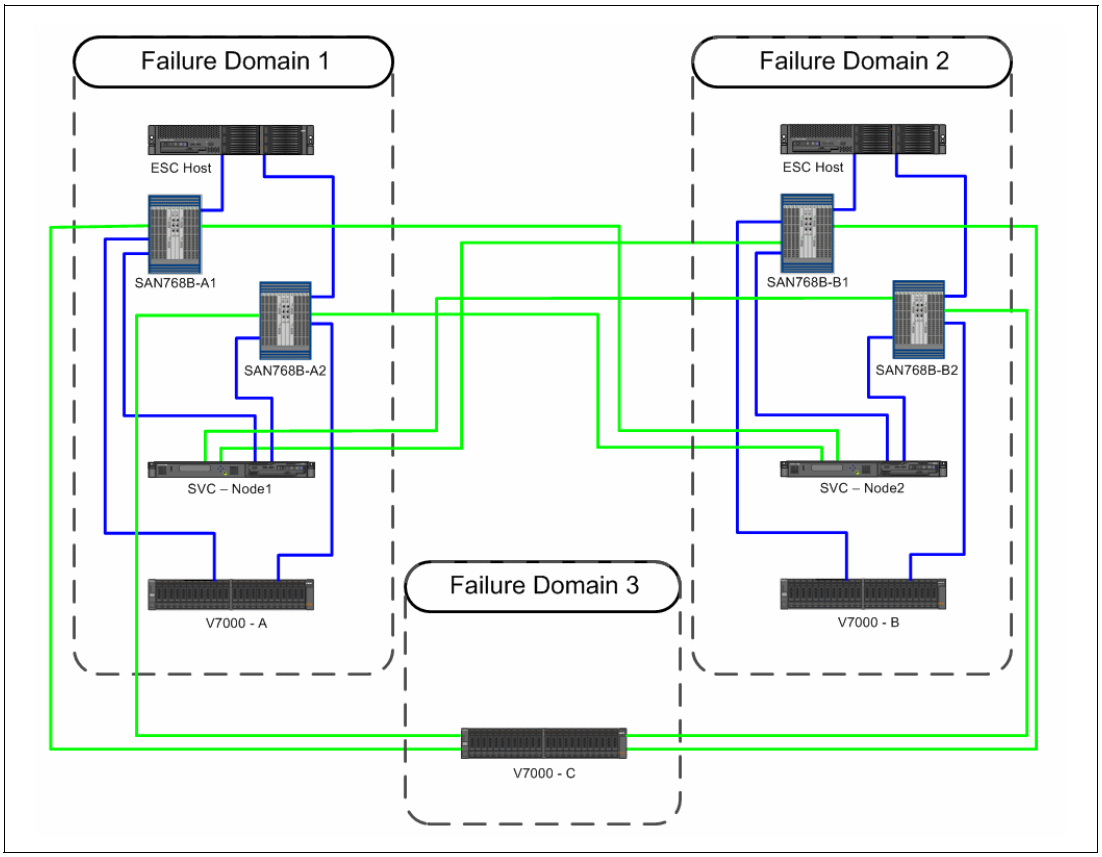


Figure 3-3 SAN Volume Controller Stretched Cluster configuration

As Table 3-2 shows, SAN Volume Controller can handle every kind of single failure automatically without impact to applications.

Table 3-2 Failure scenarios

Failure scenario	**SAN Volume Controller Streteched Cluster behavior**	**Server / 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
Single quorum storage failure	System continues to operate on the same data copy, and SAN Volume Controller selects another quorum disk candidate as the active quorum disk.	None

Failure scenario	SAN Volume Controller Streteched Cluster behavior	Server / Application impact
Failure of either Failure Domain 1 or 2. (containing SAN Volume Controller nodes)	System continues to operate on the remaining failure domain that contains SAN Volume Controller nodes.	Servers without high availability (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. SAN Volume Controller cache is disabled, which might degrade performance.
Failure of Failure Domain 3 (containing active quorum disk)	System continues to operate on both Failure Domains 1 and 2. SAN Volume Controller selects another quorum disk.	None
Access loss between Failure Domains 1 and 2 (containing SAN Volume Controller nodes)	System continues to operate the failure domain with SAN Volume Controller configuration. The node continues with operation, while the SAN Volume Controller 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. SAN Volume Controller cache is disabled, which might degrade performance.

3.4 SAN Volume Controller Stretched Cluster configurations

The SAN Volume Controller nodes of a Stretched Cluster configuration must be connected to each other by Fibre Channel or FCIP links. These links provide paths for node-to-node communication and for host access to SAN Volume Controller nodes. Stretched Cluster supports three different approaches for node-to-node intra-cluster communication between failure domains:

- Attach each SAN Volume Controller node to the Fibre Channel switches in the local and the remote failure domain directly. Thus, all node-to-node traffic can be done without traversing ISLs. This approach is referred to as Stretched Cluster *No ISL configuration.*
- Attach each SAN Volume Controller node only to local Fibre Channel switches and configure ISLs between failure domains for node-to-node traffic. This approach is referred to as Stretched Cluster *ISL configuration.*
- Attach each SAN Volume Controller 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 6.4. This approach is referred to as Stretched Cluster *FCIP configuration.*

Each of these Stretched Cluster configurations along with their associated attributes is described to assist with the selection of the appropriate configuration to meet your solution 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 consume all of the buffer-to-buffer credits. The link speed is also limited by the cable length between nodes. Table 3-3 lists the supported distances for each of the SAN Volume Controller Stretched Cluster configurations along with their associated code level and ports speed requirements.

Table 3-3 Supported distances

Configuration	SAN Volume Controller Code Level	Maximum Length	Maximum Link Speed
No ISL	5.1 or later	< 10 km	8 Gbit/s
No ISL	6.3 or later	< 20 km	4 Gbit/s
No ISL	6.3 or later	< 40 km	2 Gbit/s
ISL	6.3 or later	< 300 km	2,4,8 Gbit/s
FCIP	6.4 or later	< 300 km	2,4,8 Gbit/s

3.4.1 No ISL configuration

This configuration is similar to a standard SAN Volume Controller environment with the main difference being that nodes are distributed across two failure domains. Figure 3-4 on page 36 illustrates the *No ISL configuration.* Failure Domain 1 and Failure Domain 2 contain the SAN Volume Controller 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 similar to a standard SAN Volume Controller cluster.
- Limited hardware effort: WDM devices can be used, but are not required.

Requirements

The No ISL configuration has these requirements:

- Requires four dedicated fiber links per I/O group between failure domains.
- ISLs are not used between SAN Volume Controller nodes.
- Passive wavelength division multiplexing (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. (SAN Volume Controller 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.
- Third failure domain is required for quorum disk placement.
- Quorum disk storage system must be Fibre Channel attached.

Figure 3-4 illustrates the SAN Volume Controller Stretched Cluster *No ISL configuration*. Failure Domain 1 and Failure Domain 2 contain the SAN Volume Controller nodes along with customer data. Failure Domain 3 contains the storage subsystem that provides the active quorum disk.

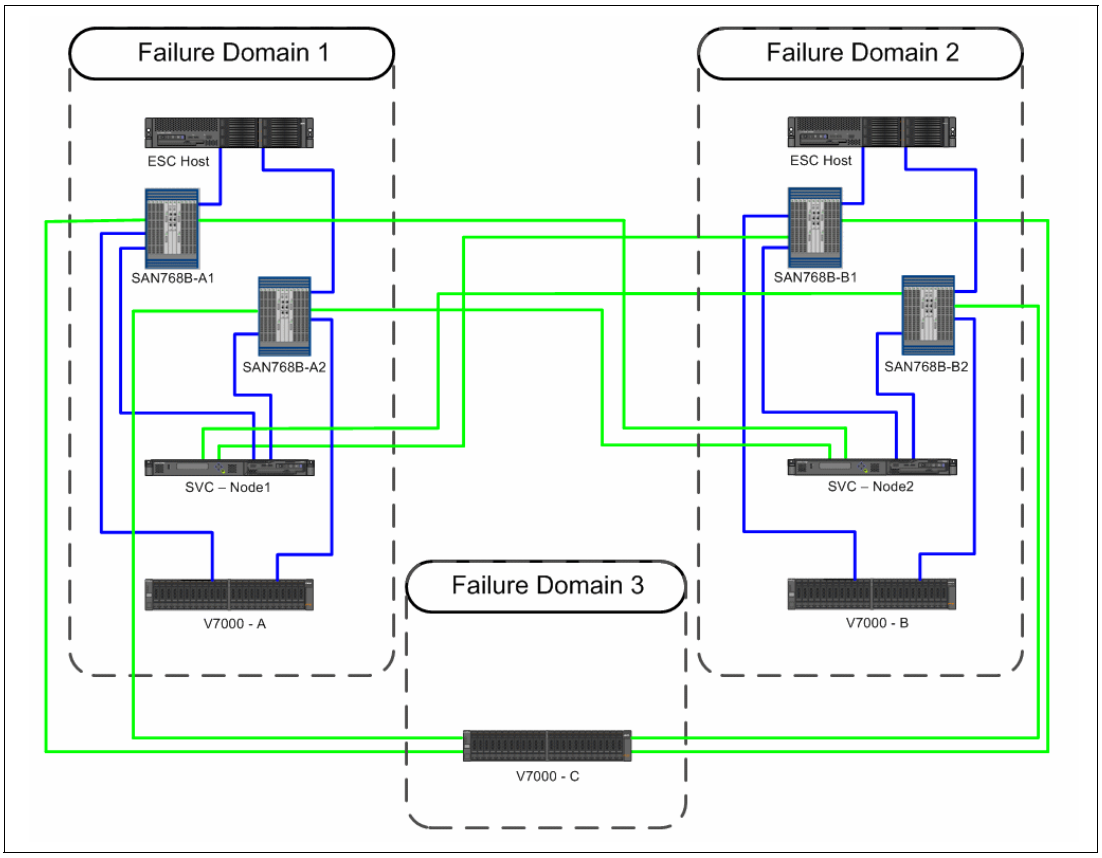


Figure 3-4 SAN Volume Controller Stretched Cluster: No ISL configuration

Zoning requirements

Zoning requirements for the SAN Volume Controller Stretched Cluster *No ISL configuration* are the same as with a standard SAN Volume Controller configuration.

- Servers access only SAN Volume Controller nodes. There is no direct access from servers to back-end storage.
- Separate zone is configured for node-to-node traffic.
- SAN Volume Controller nodes of the same I/O group do not communicate by using ISLs.
- Zones should not contain multiple back-end disk systems.

Figure 3-5 illustrates the SAN Volume Controller Stretched Cluster *No ISL configuration* with passive WDM connections between Failure Domains 1 and 2.

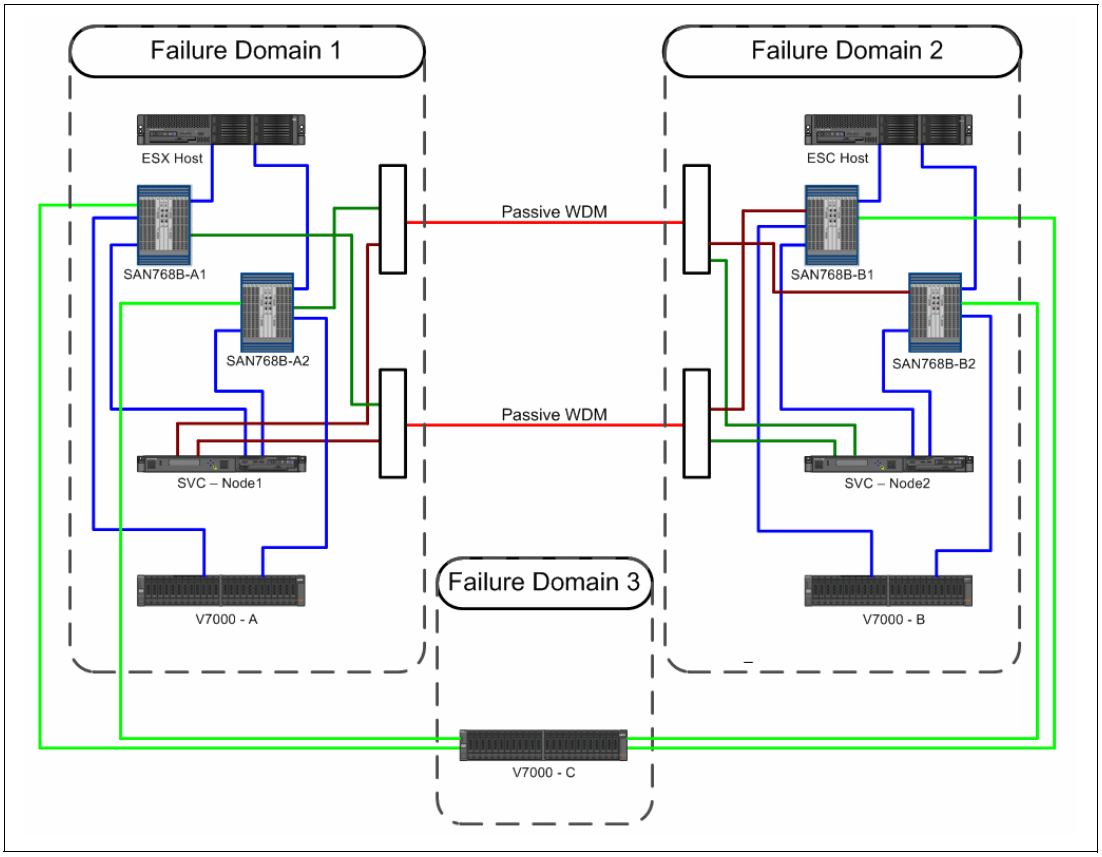


Figure 3-5 SAN Volume Controller Stretched Cluster: No ISL configuration (with WDM)

3.4.2 ISL configuration

This configuration is similar to a standard SAN Volume Controller 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. SAN Volume Controller support for ISLs was introduced in version 6.3.

The use of ISLs increases the supported distance for SAN Volume Controller Stretched Cluster configurations to 300 km. Although the maximum supported distance is 300 km, there are instances where host-dependent I/O must traverse the long distance links multiple times. Because of this, the associated performance degradation might exceed acceptable levels. To mitigate this exposure, generally 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 WDM.

Requirements

The ISL configuration has these requirements:

- Requires four dedicated fiber links per I/O group between failure domains.
- Using ISLs for node-to-node communication requires configuring two separate SANs:
  - One SAN is dedicated for SAN Volume Controller 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.
- A third failure domain is required for quorum disk placement.
- Storage controllers that contain quorum disks must be Fibre Channel attached.
- A guaranteed minimum bandwidth of 2 MBytes 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 3-6 on page 39 illustrates the *ISL configuration.* Failure Domain 1 and Failure Domain 2 contain the SAN Volume Controller nodes along with customer data. Failure Domain 3 contains the storage subsystem that provides the active quorum disk.

Zoning requirements

The SAN Volume Controller Stretched Cluster *ISL configuration* requires private and public SANs. The two SANs must be configured according to the following rules:

- Two ports of each SAN Volume Controller node are attached to the public SANs.
- Two ports of each SAN Volume Controller node are attached to the private SANs.
- 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 SAN Volume Controller Metro Mirror or Global Mirror must be attached to the public SANs.
- Failure Domain 3 (the quorum disk) must be attached to the public SAN.
- ISLs belonging to the private SANs must not be shared with other traffic, and must not be over-subscribed.

For more information, see the SAN Volume Controller Information Center or "V6.3.0 Configuration Guidelines for Extended Distance Split-System Configurations for IBM System Storage SAN Volume Controller" at:

http://www-01.ibm.com/support/docview.wss?&uid=ssg1S7003701

Figure 3-6 illustrates the SAN Volume Controller Stretched Cluster *ISL configuration*. The private and public SANs are represented as logical switches on each of the four physical switches.

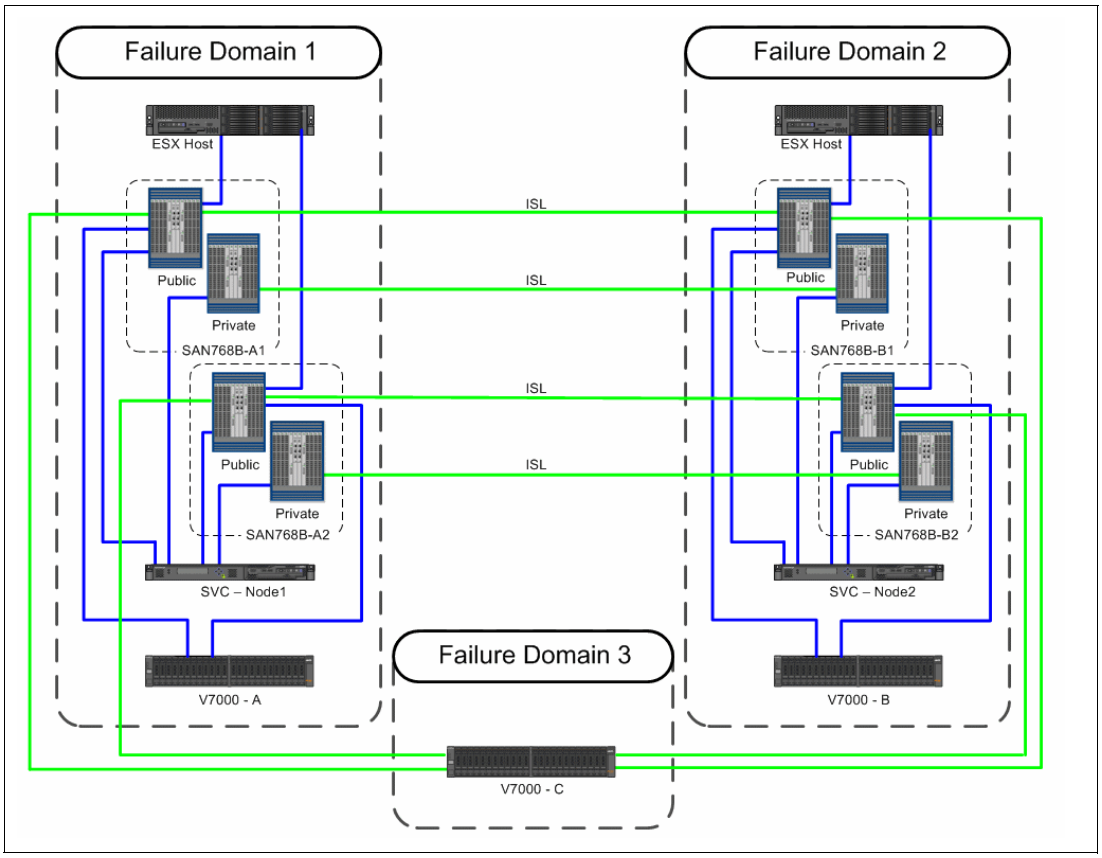


Figure 3-6 SAN Volume Controller Stretched Cluster: ISL configuration

Figure 3-7 illustrates the SAN Volume Controller 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.

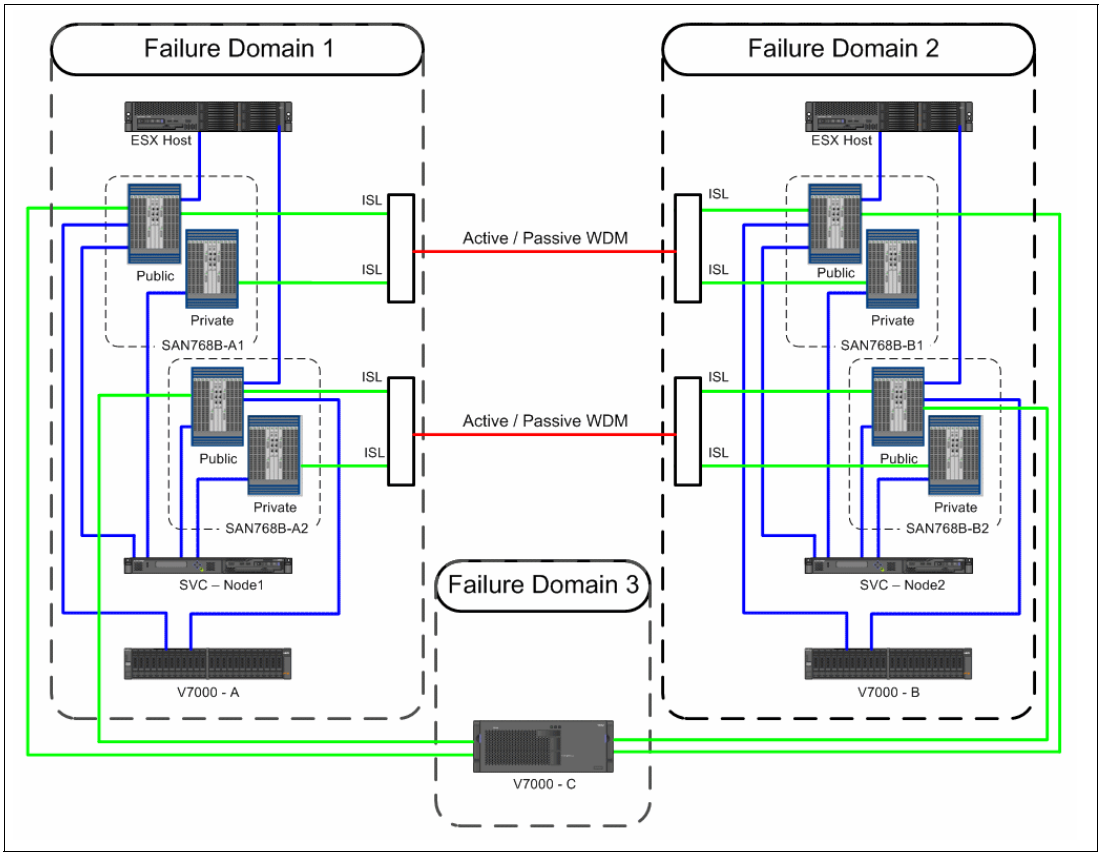


Figure 3-7 SAN Volume Controller Stretched Cluster: ISL configuration (with WDM)

3.4.3 FCIP configuration

In this configuration, FCIP links are used between failure domains. SAN Volume Controller support for FCIP was introduced in version 6.4. This configuration is variation of the ISL configuration described previously, and therefore many of the same requirements apply.

Advantages

FCIP configuration has these advantages:

- 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 SAN Volume Controller 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.

- A third failure domain is required for quorum disk placement.
- Failure domain 3 (quorum disk) must be either Fibre Channel or FCIP attached.
  - If FCIP attached, the response time to the quorum disk cannot exceed 80 ms.
- Storage controllers that contain quorum disks must be either Fibre Channel or FCIP attached.
- A guaranteed minimum bandwidth of 2 MByte/s is required for node-to-quorum traffic.
- No more than one ISL hop is supported for connectivity between failure domains.

Zoning requirements

The SAN Volume Controller Stretched Cluster *FCIP configuration* requires private and public SANs. The two SANs must be configured according to the following rules:

- Two ports of each SAN Volume Controller node are attached to the public SANs.
- Two ports of each SAN Volume Controller node are attached to the private SANs.
- 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 SAN Volume Controller Metro Mirror or Global Mirror must be attached to the public SANs.
- Failure Domain 3 (quorum disk) must be attached to the public SAN.
- ISLs belonging to the private SANs must not be shared with other traffic, and must not be over-subscribed.

Figure 3-8 illustrates the SAN Volume Controller Stretched Cluster *FCIP configuration*.

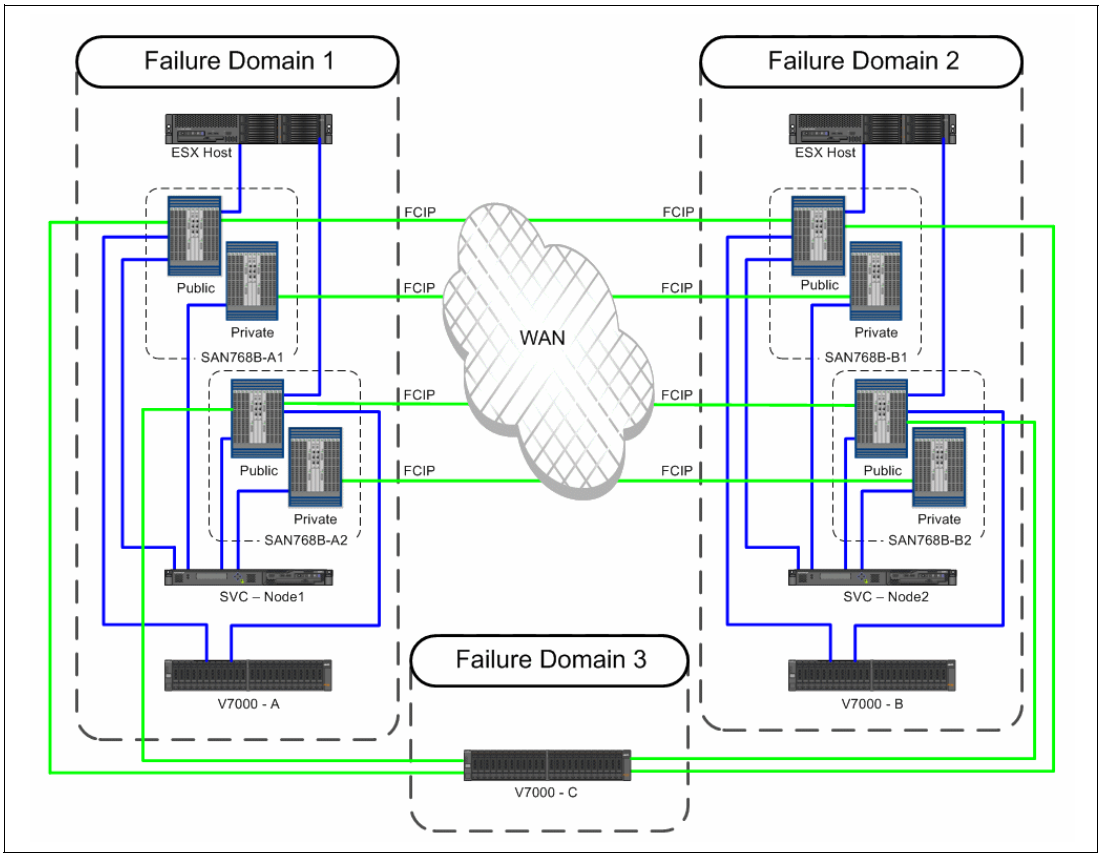


Figure 3-8 FCIP configuration

3.5 Fibre Channel settings for distance

Usage of Long Wave (LW) SFPs is an appropriate method to overcome long distances. Starting with SAN Volume Controller 6.3, active and passive DWDM/CWDM technology is 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 SAN Volume Controller Metro Mirror are also supported by SAN Volume Controller Stretched Cluster configurations.

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

Thus buffer-to-buffer credits are necessary to have multiple Fibre Channel frames in flight (Figure 3-9). 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.

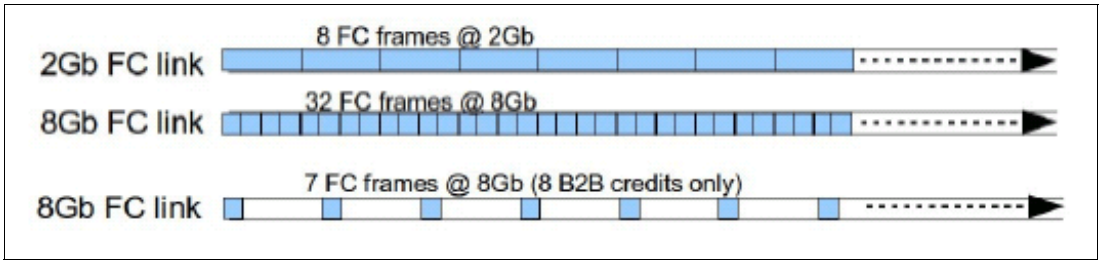


Figure 3-9 Buffer credits in flight

The calculation assumes that the other end of the link starts transmitting the acknowledgement frame R_RDY in the same moment when the last bit of the incoming frame arrives at the receiver, which is not the case. The guidelines give the minimum numbers. The performance drops dramatically if there are not enough buffer credits for the link distance and link speed. Table 3-4 illustrates the relationship between BB credits and distance.

Table 3-4 Buffer-to-buffer credits

FC link speed	B2B 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

The number of buffer-to-buffer credits that are provided by an SAN Volume Controller Fibre Channel host bus adapter (HBA) is limited. An HBA of a model 2145-CF8 node provides 41 buffer credits, which are sufficient for 10 km distance at 8 Gbit/s. The HBAs in all earlier SAN Volume Controller models provide only eight buffer credits, which are enough only for 4 km

distance with 4 Gbit/sec link speed. These numbers are determined by the HBA's hardware and cannot be changed.

Guideline: Generally, use 2145-CF8 or CG8 nodes for distances longer than 4 km to provide enough buffer-to-buffer credits.

FC switches have default settings for the B2B credits (in Brocade switches: 8 B2B credits for each port). Although the SAN Volume Controller HBAs provide 41 B2B credits, the switch stay at the default value. Thus it is necessary to adjust the switch B2B credits manually. For Brocade switches, the port buffer credits can be changed by using the **`portcfgfportbuffers`** command.

3.6 SAN Volume Controller I/O operations on mirrored volumes

There are two architectural behaviors with the SAN Volume Controller that must be considered for mirrored volumes in a Stretched Cluster configuration:

- Preferred node

  The preferred node runs all write destaging to both copies of the volume mirror. Additionally, for hosts with multipath drivers that support SAN Volume Controller preferred node, the host directs all read and write activity to the preferred node. For these hosts, the preferred node runs all read caching and write cache destaging. For hosts that do not support SAN Volume Controller preferred node, the node that is selected for read and write operations is determined by the host multipath driver.

  The preferred node is queried by supporting multipath drivers through the SCSI *Report Target Port Groups* command.

  Remember: ESX hosts do not support SAN Volume Controller preferred node, and therefore host pathing to SAN Volume Controller node alignment must be configured manually.

- Primary copy

  The primary copy of the volume mirror is 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 following initial volume creation without interrupting I/O processing.

Read operations

Volume mirroring implements a single read algorithm with one copy that is designated as the primary. SAN Volume Controller reads the data from the primary copy and does not automatically distribute the read requests across both copies. The secondary copy is read from only if the primary copy is not available. This process is independent of which node the host is reading from. Because of this behavior, for optimal performance 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 as the host. This configuration ensures that read operations are occurring locally and therefore do not experience the additional latency that is caused by traversing the long-distance links.

Figure 3-10 Illustrates the data flow for a read command. The green line represents an optimal configuration where the host in Failure Domain 1 is accessing a volume that has its primary copy set to the node that is also in Failure Domain 1. The red line is a non-optimal configuration where the host in Failure Domain 2 is accessing a volume that has its primary copy set to the node in Failure Domain 1.

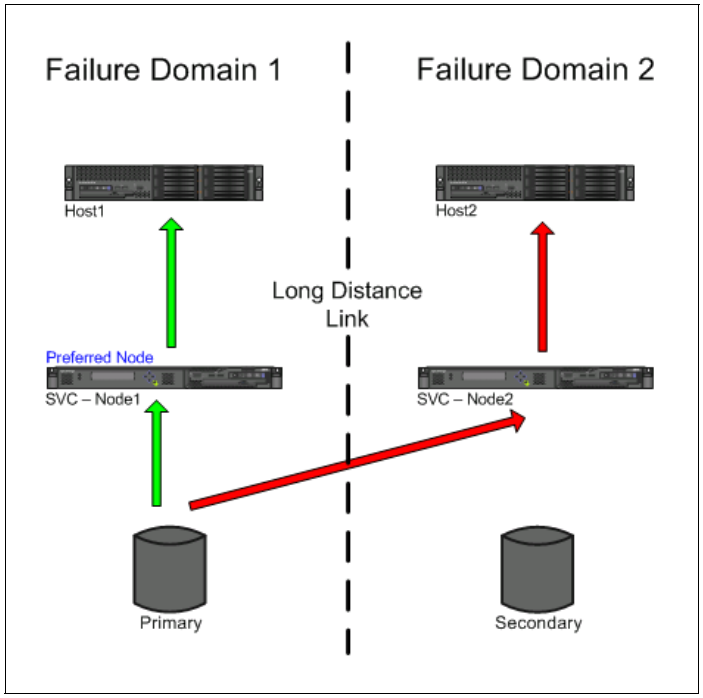


Figure 3-10 Read operation

Guidelines:

- Place the primary copy of the mirror in the same failure domain as the host that is accessing the mirrored volume. The primary copy can be changed by using the **chvdisk** command.
- Have the host access the mirrored volume through the SAN Volume Controller node that is in the same failure domain as the host.

Write operations

All writes are mirrored between nodes in the IOGroups, and therefore must traverse the long-distance links. The writes are then destaged to both copies of the volume mirror from the preferred node. 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 failure domain as the host.

Figure 3-11 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.

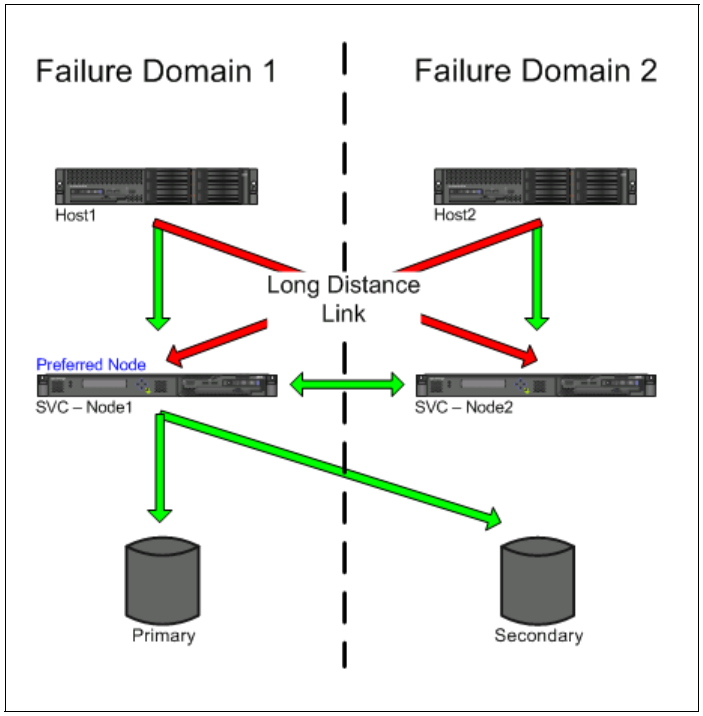


Figure 3-11 Write operation

Guideline: Have the host access the mirrored volume through the SAN Volume Controller node that is in the same failure domain as the host.

4

Implementation

This chapter provides information about how this solution is implemented.

This chapter includes the following sections:

- Test Environment
- ADX: Application Delivery Controller
- IP networking configuration
- IBM FC SAN
- IBM Storage Volume Controller using Stretched Cluster
- SAN Volume Controller volume mirroring
- Read operations
- Write operations
- SAN Volume Controller quorum disk
- Quorum disk requirements and placement
- Automatic SAN Volume Controller quorum disk selection
- Backend Storage allocation to the SAN Volume Controller Cluster
- Volume allocation
- ESXi: VMware

4.1 Test Environment

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

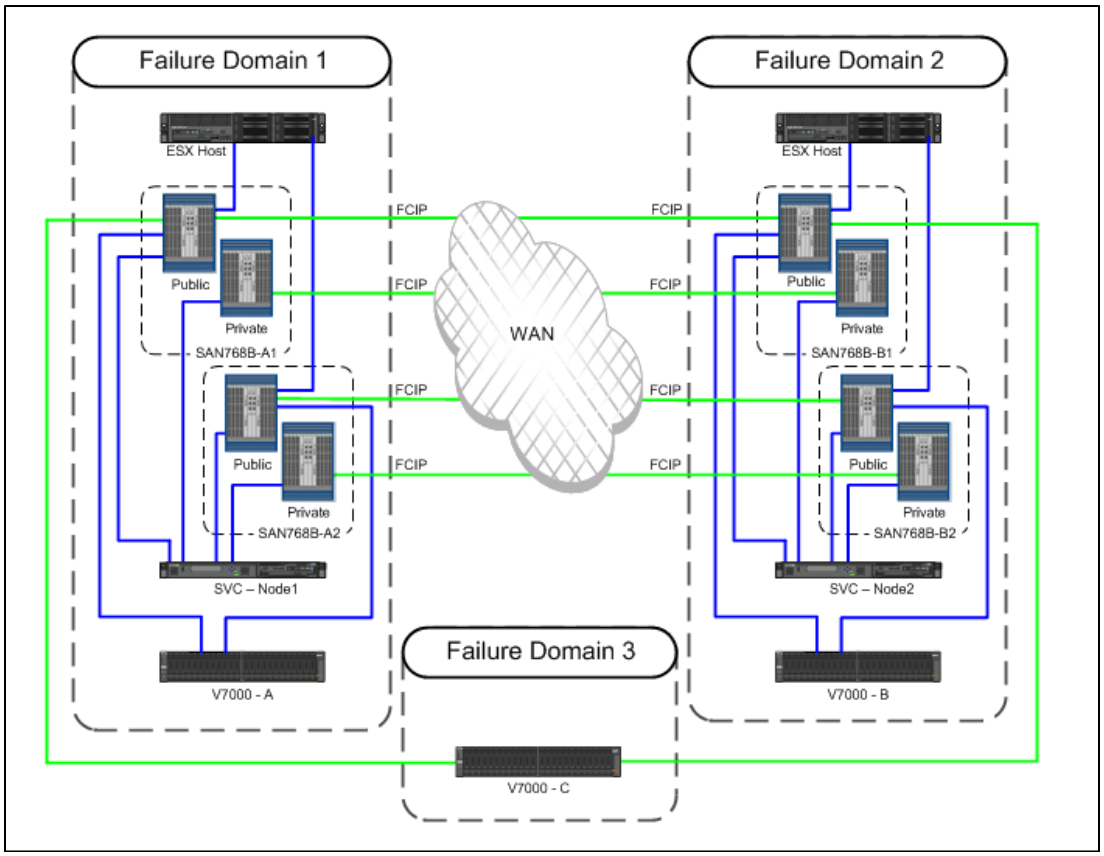


Figure 4-1 IBM SAN and SAN Volume Controller Stretched Cluster VMware solution

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 to complete the solution.

4.2 ADX: Application Delivery Controller

The Brocade ServerIron ADX provides 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 help offer seamless access for clients who connect to VMs that migrate between data centers. The example configuration uses ServerIron ADX v12.4.0.

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 ServerIron ADX in each data center has a different VIP for the same set of Real Servers. The example setup 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 from seeing 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 4-2 shows what the VIP to Real Server configuration looks like.

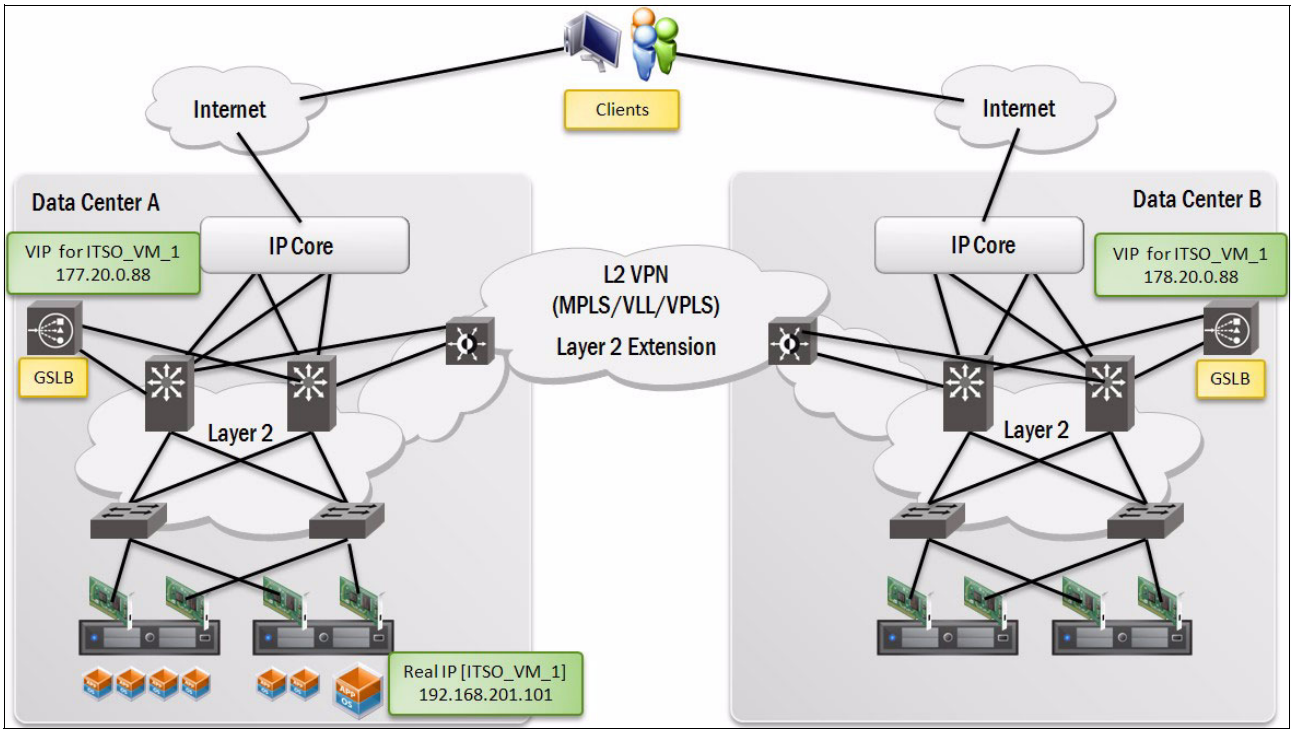


Figure 4-2 VIP and Real Server configuration for ITSO_VM_1

GSLB enables a ServerIron 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 ServerIron ADX evaluates the IP addresses in the DNS replies from the authoritative DNS server for which the ServerIron 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 that the VM is active in. 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 ServerIron ADX uses a DNS record TTL value of 10 seconds.

You must configure the following items:

1. VIP address assignment that outside clients will use, and the corresponding Real Server configuration behind the VIP
2. GSLB configuration
3. ARB server installation
4. ADX registration in ARB plug-in
5. VM mobility enablement in ARB plug-in in vCenter

4.2.1 VIP and Real Server configuration

For the ServerIron 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 (Example 4-1). In this example, select port 50 to bind for ITSO_VM_1.

Example 4-1 Data Center A - ServerIron 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 ServerIron 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 **gslb-disable** 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 - ServerIron 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 GSLB Configuration

The ServerIron ADX can either act as proxy for either local or remote DNS servers, or be populated with host information to respond with. Configure the ServerIron ADX in Data Center A to respond with DNS request 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 ServerIron 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 ServerIron ADX in Data Center A for GSLB (Example 4-4).

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

Example 4-5 ServerIron 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, corresponding to Data Center A, is Active. It is active because the Real Server definition for ITSO_VM_1 behind VIP 178.20.0.88 was initially disabled.

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

```
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 ARB Server Installation

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

- Processor - Minimum 2 cores and 2 GHz
- Memory - Minimum 2-GB RAM
- Windows Server 2003, Windows Server 2008 R2, or 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 ServerIron 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 that you want to install the ARB plug-in on. Have the FQDN/IP address and login credentials of this vCenter server before you install ARB.

Figure 4-3 on page 53 shows the ARB prerequisites.

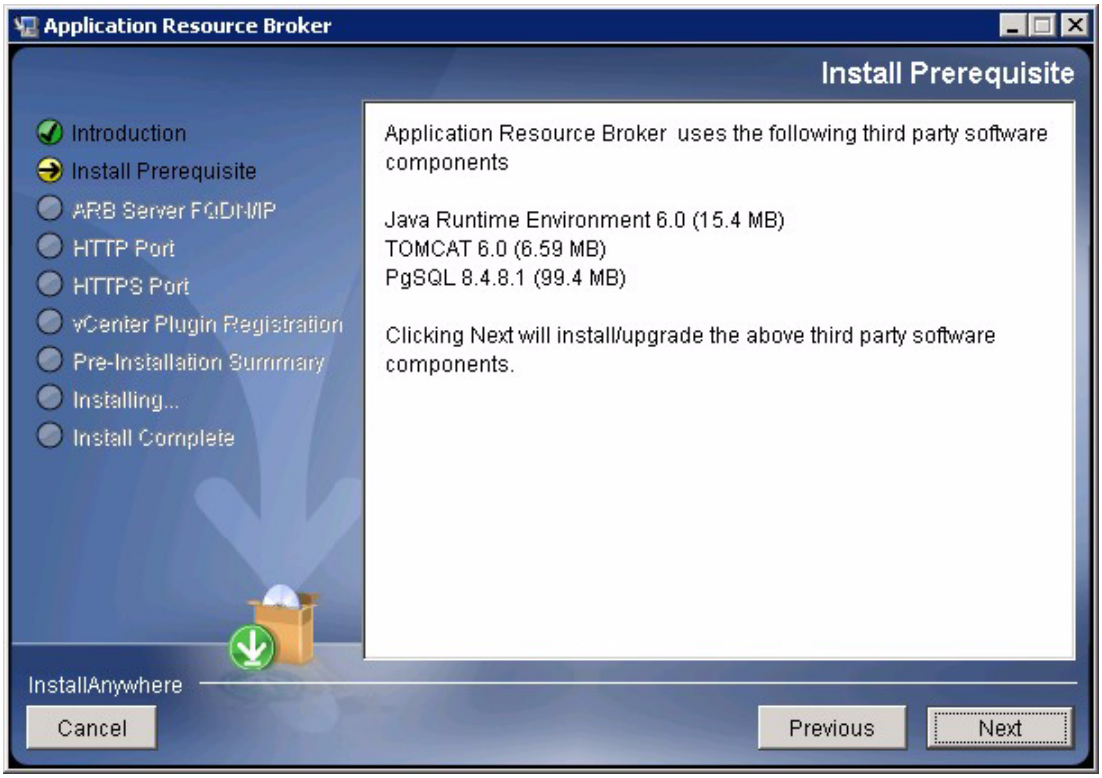


Figure 4-3 ARB Installation Prerequisites

Figure 4-4 shows ARB Server IP address settings.

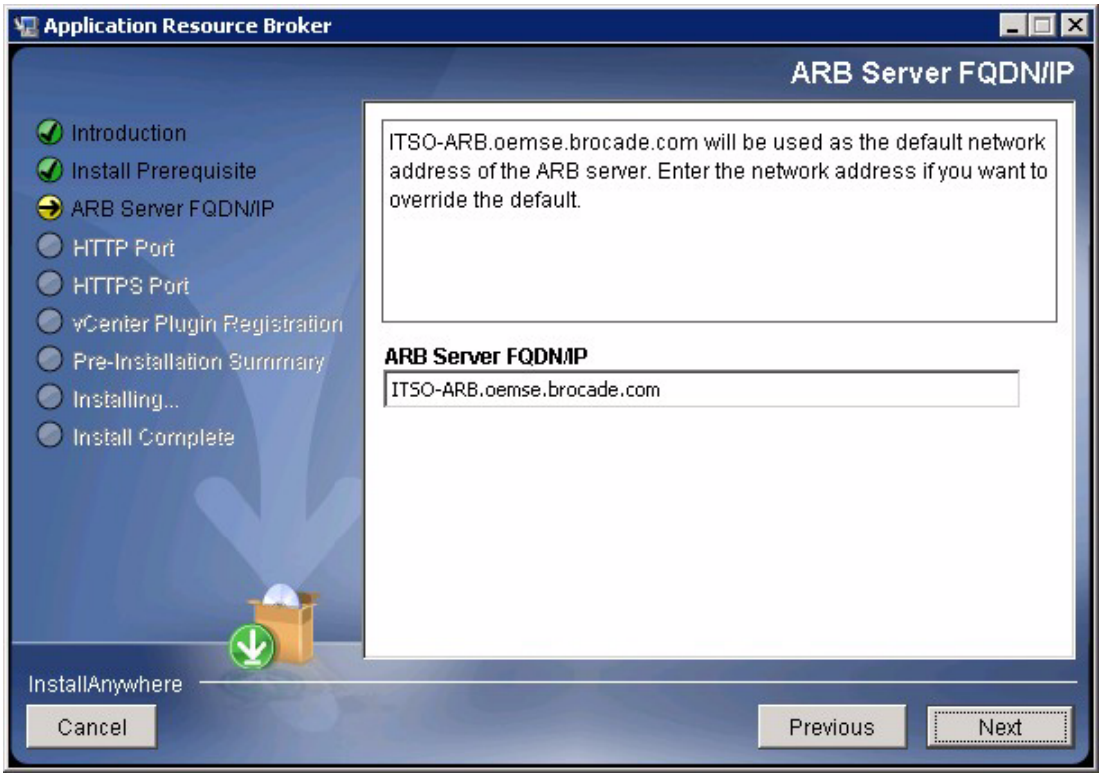


Figure 4-4 ARB Server FQDN/IP address settings

Figure 4-5 shows the ARB vCenter server settings.

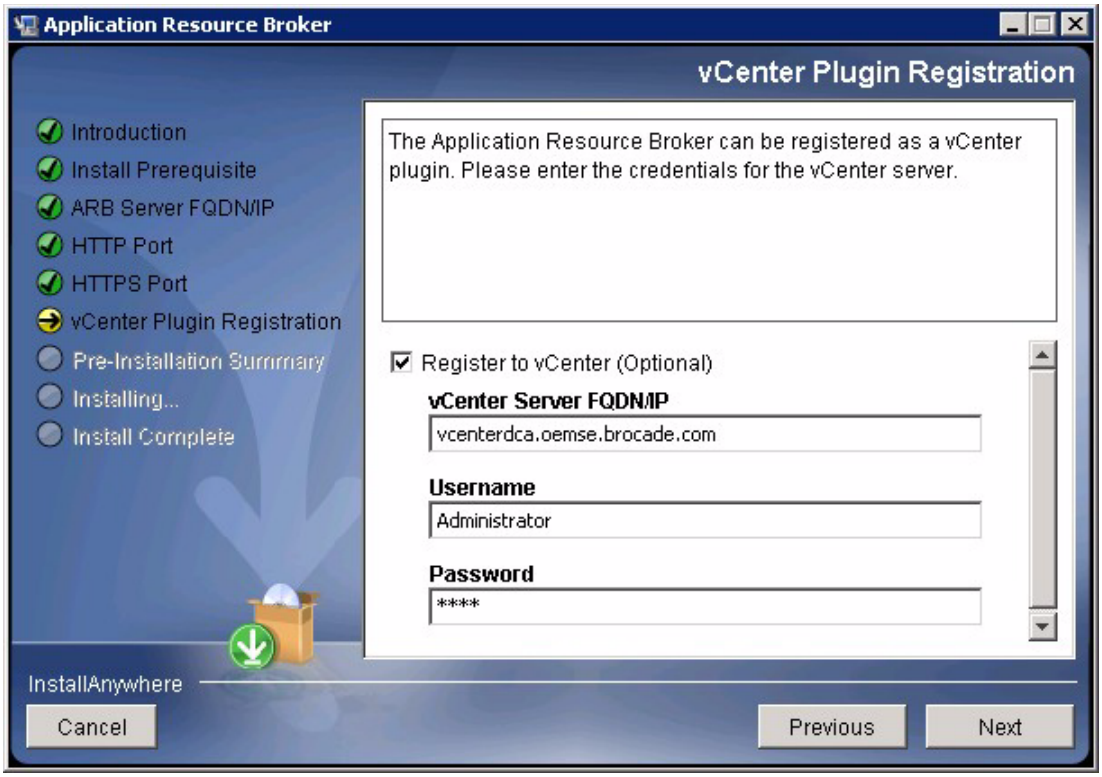


Figure 4-5 ARB vCenter server settings

In Figure 4-6, the ARB installation is complete.

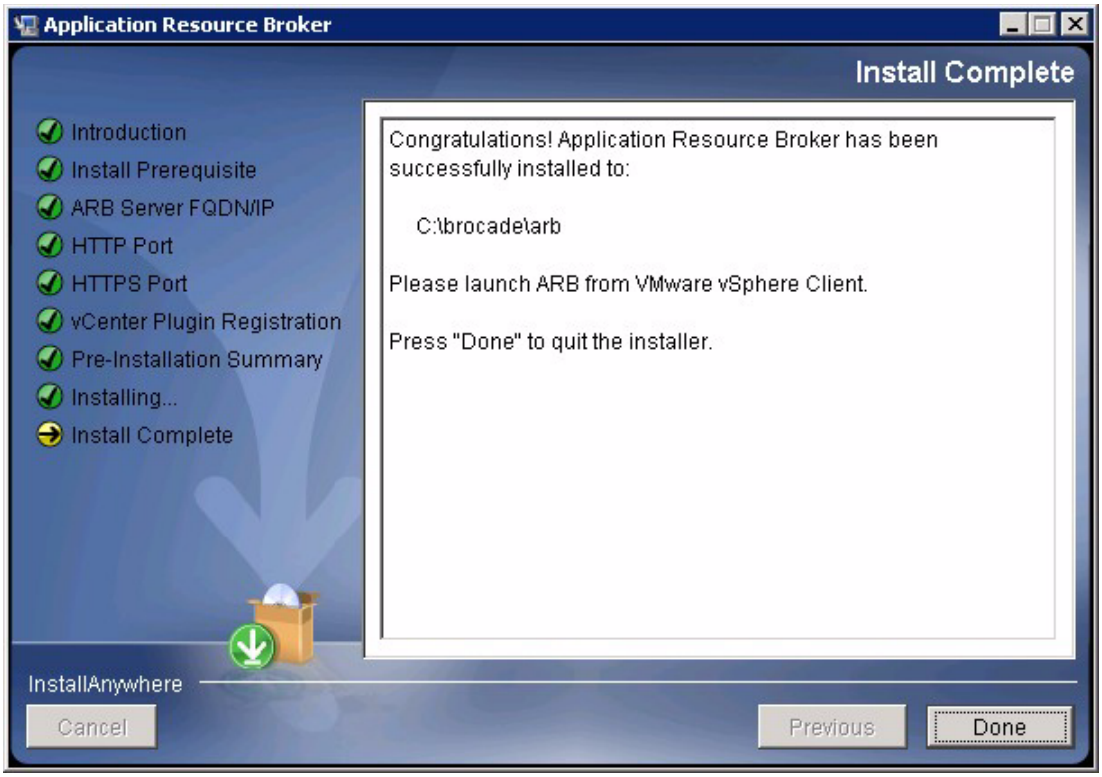


Figure 4-6 ARB installation complete

After you complete the ARB installation, connect to your vCenter server using vCenter client and check that it was installed correctly by clicking **Plug-ins** → → **Manage Plug-ins** and seeing whether it is installed and Enabled as shown in Figure 4-7.

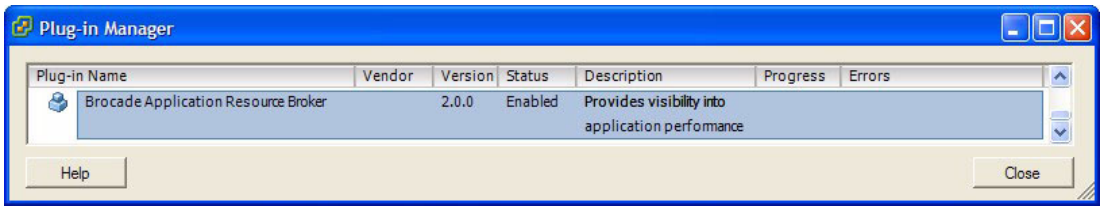


Figure 4-7 Viewing the Brocade ARB plug-in in vCenter

4.2.4 ADX registration in ARB plug-in

You can access the ARB plug-in interface in vCenter client by clicking the Cluster level and then the Application Resource Broker tab. Click the ADX Devices tab within the ARB window to register your ADX devices as seen in Figure 4-8.

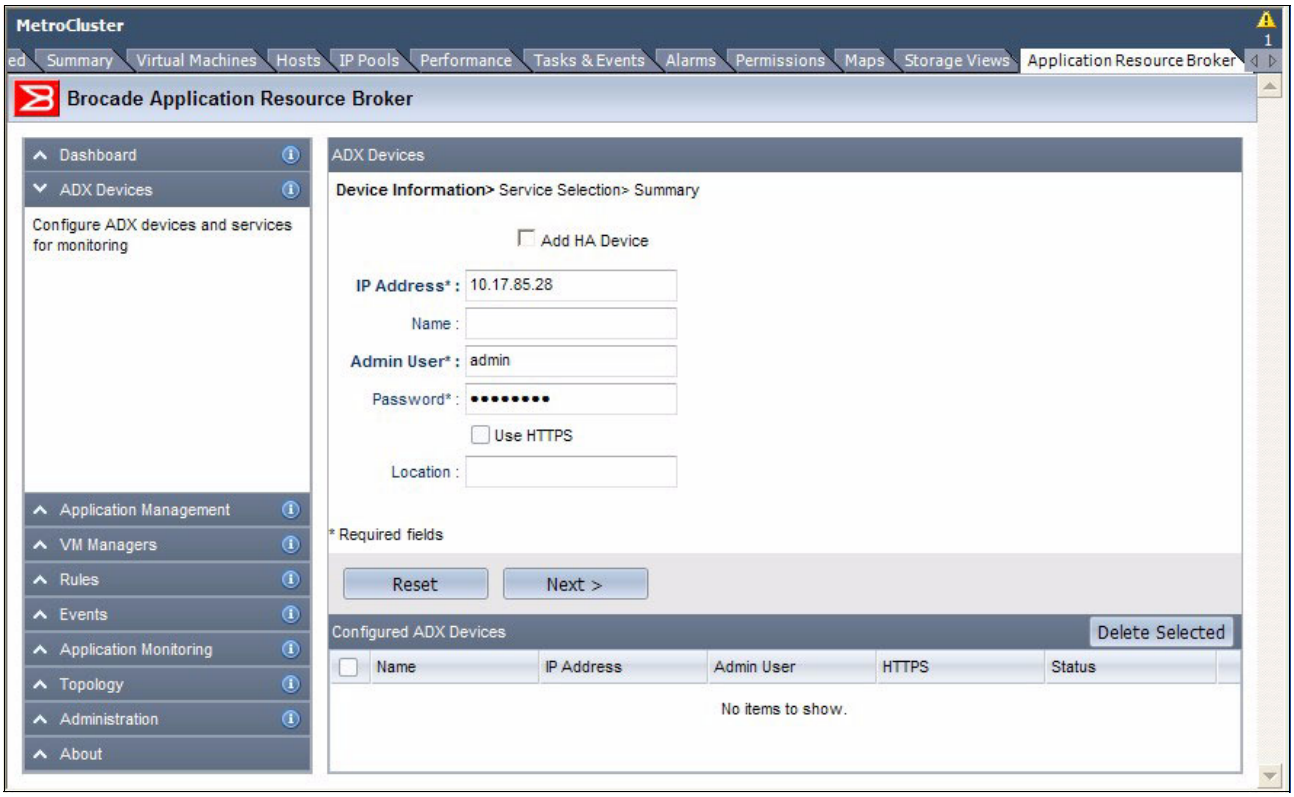


Figure 4-8 Brocade ARB window in vCenter client: ADX Devices

Enter the IP management address of your ADX device along with the HTTP user name and password (default admin/password). HTTPS can be used as well if configured. After that is done, click **Next**.

Select the configured Virtual Servers that you want ARB to monitor as seen in Figure 4-9.

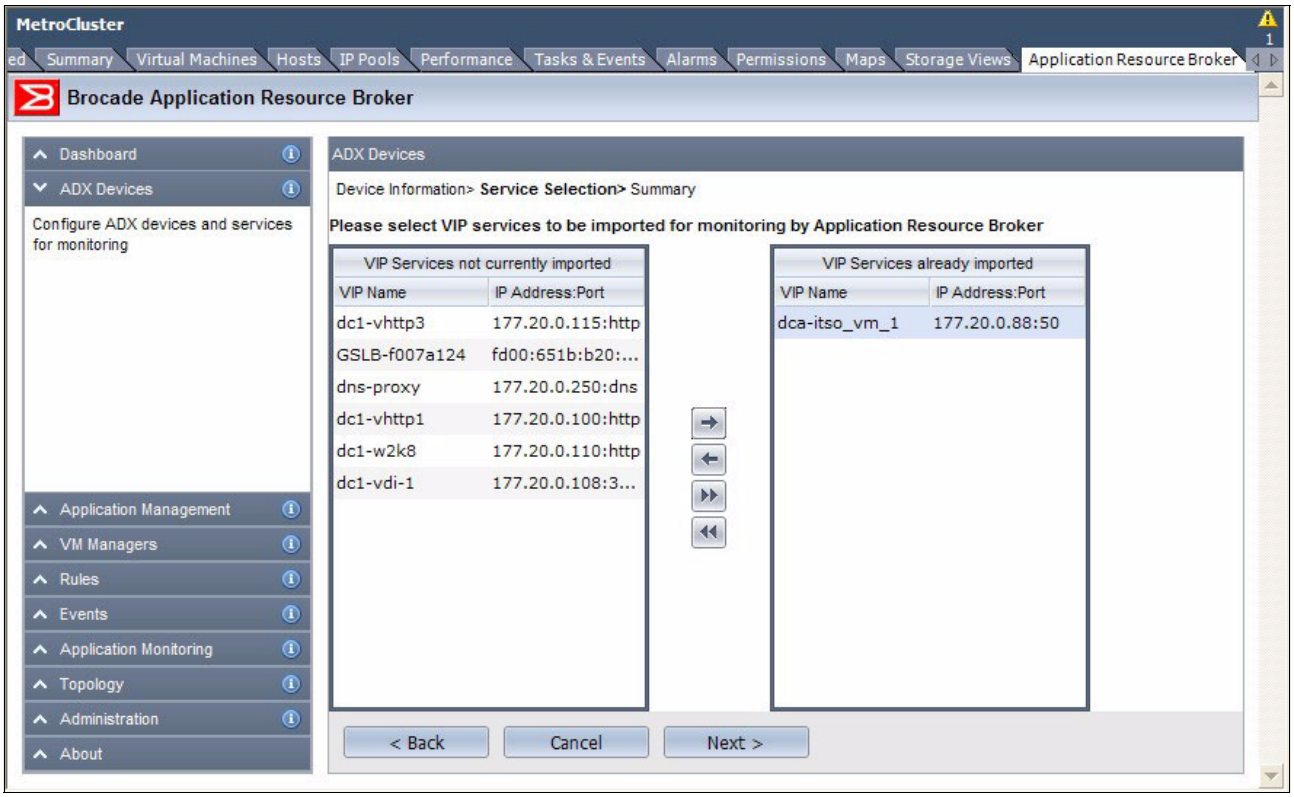


Figure 4-9 Brocade ARB configuration: Selecting the Virtual Servers for monitoring

On this ADX, select **dca-itso_vm_1** as the VIP that you want monitored. The next window is a summary window for confirmation. Do the same ServerIron ADX registration steps for the ADXs in Data Center A and Data Center B. When that is done, the ADX Devices page should look like Figure 4-10.

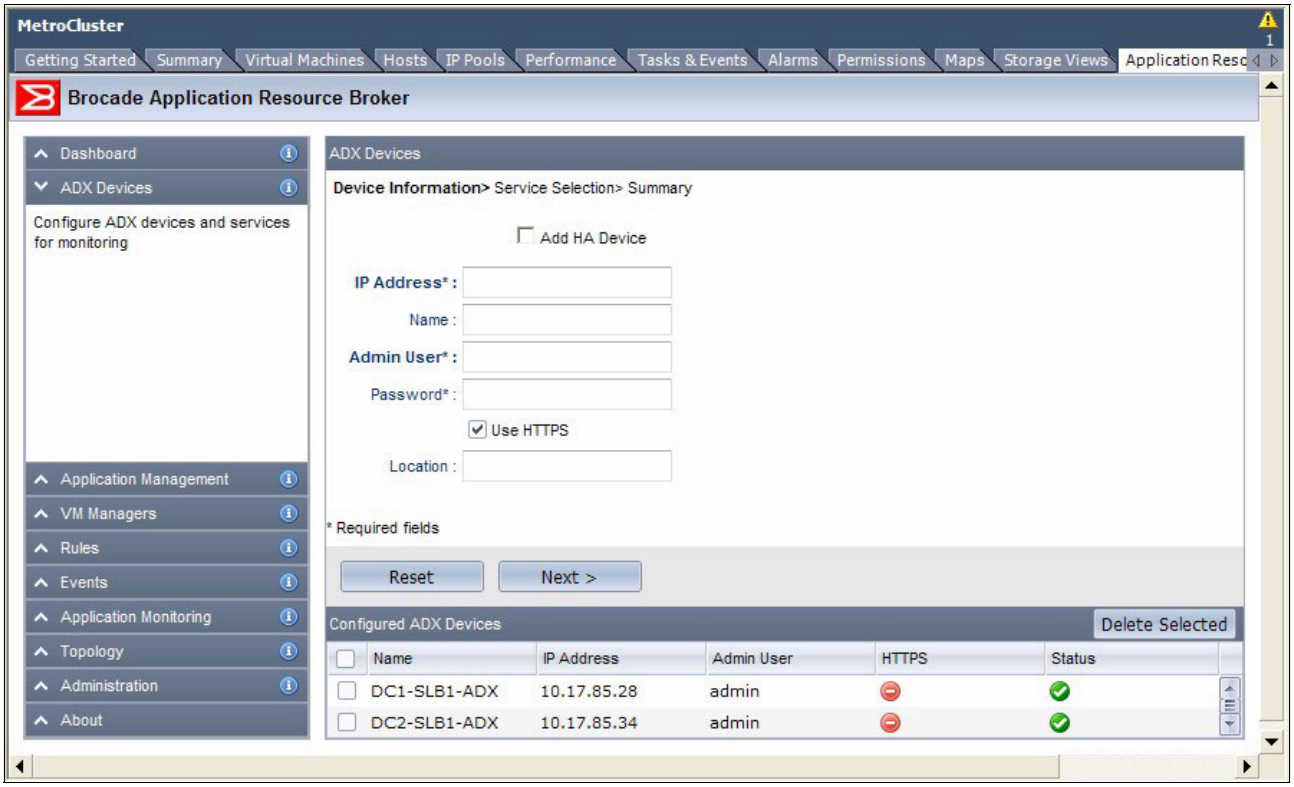


Figure 4-10 Brocade ARB configuration: Both ADXs are now registered

4.2.5 VM mobility enable in 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 seen in Figure 4-11.

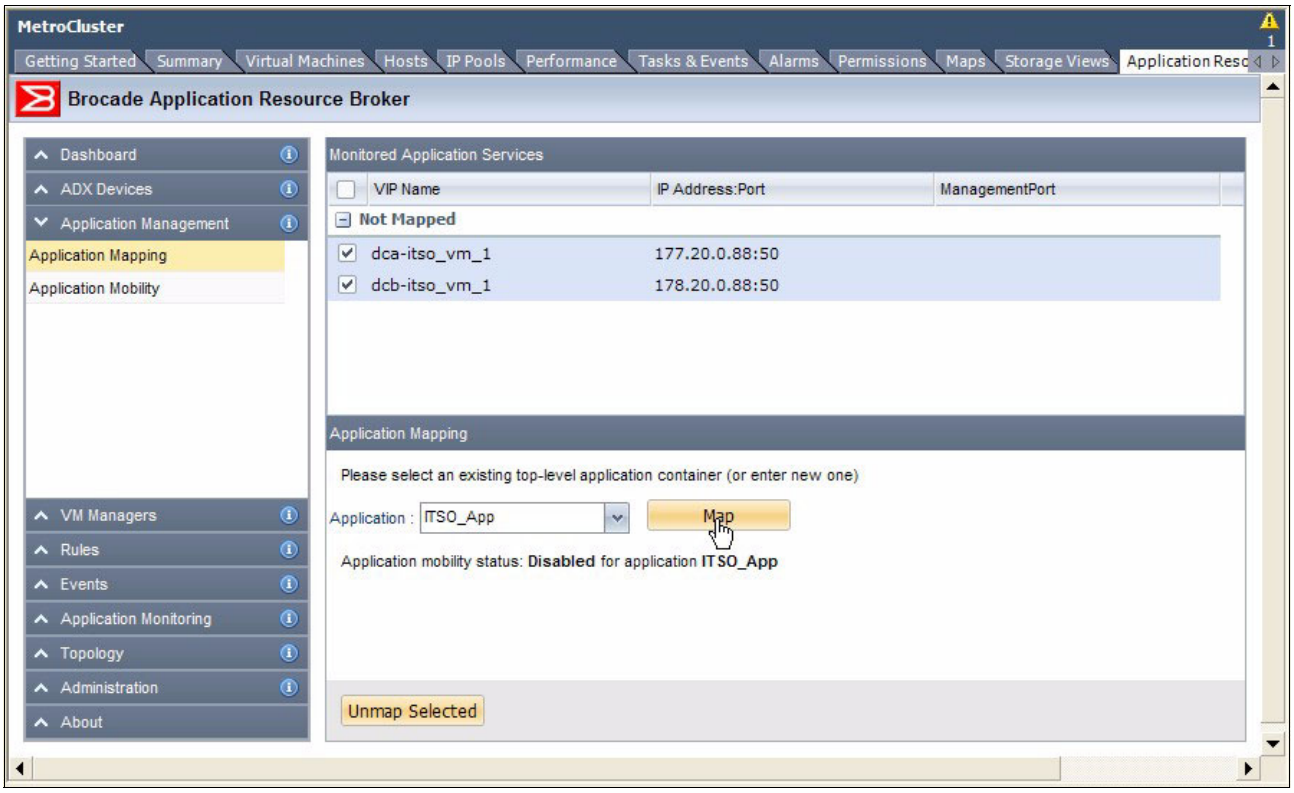


Figure 4-11 Brocade ARB configuration: Registering VIPs to a monitored application

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

After this step is completed, both virtual servers are mapped under ITSO_App, but vMotion Mobility is Disabled.Right-click the application and select **Enable mobility** as seen in Figure 4-12.

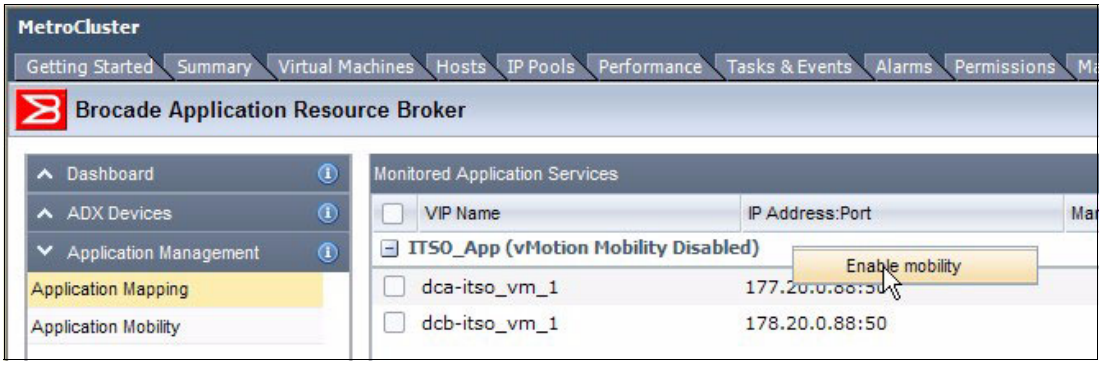


Figure 4-12 Brocade ARB configuration: Enabling VM mobility for a mapped application

Finally, check your configuration and which VIP is active, and which VIP is backup, under the Application Mobility tab as seen in Figure 4-13.

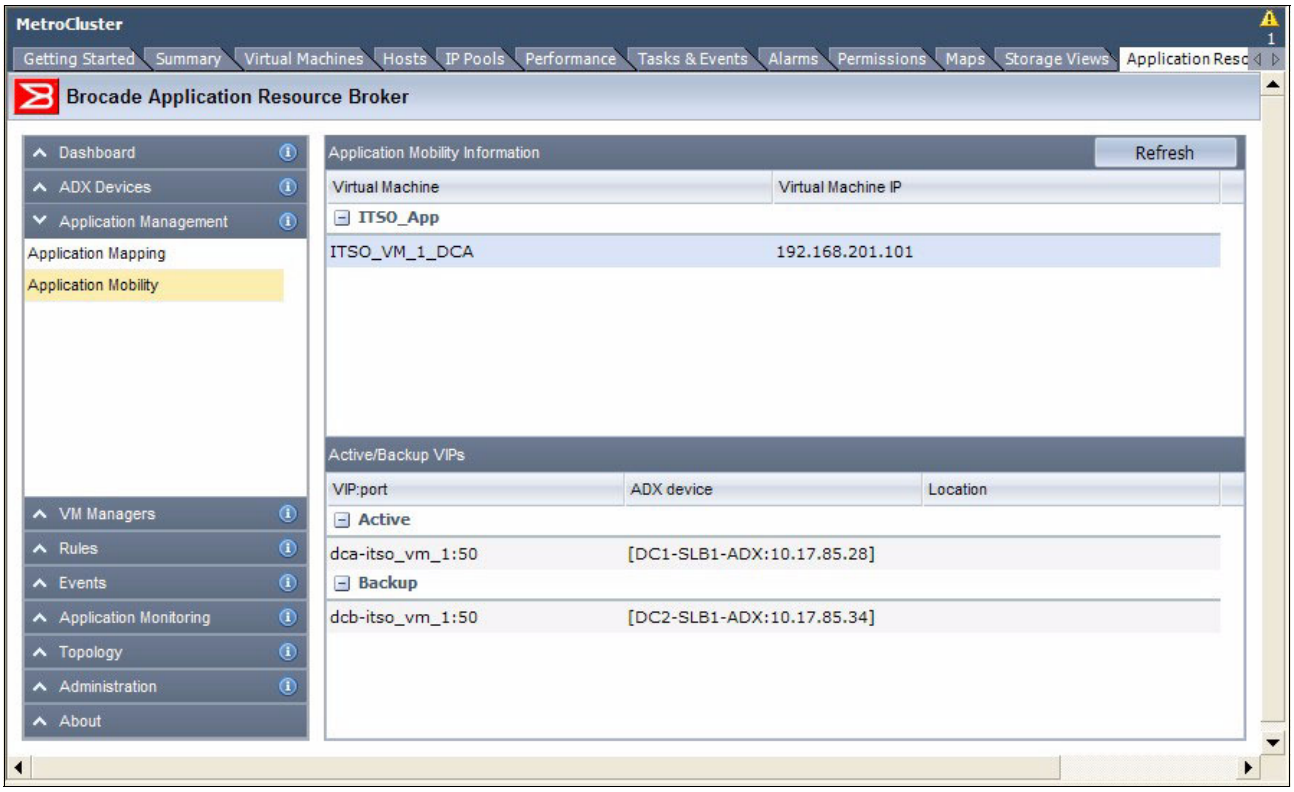


Figure 4-13 Brocade ARB: Checking the Active/Backup VIP for a particular VM

4.2.6 Additional references

More references can be found in the following manuals:

- *ServerIron ADX Administration Guide supporting v12.4.00*
- *ServerIron ADX Server Load Balancing Guide supporting v12.4.00*
- *ServerIron ADX Global Server Load Balancing Guide supporting v12.4.00*
- *ServerIron ADX Application Resource Broker Administrator's Guide v2.0.0*

4.3 IP networking configuration

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

Figure 4-14 shows the IP networking areas.

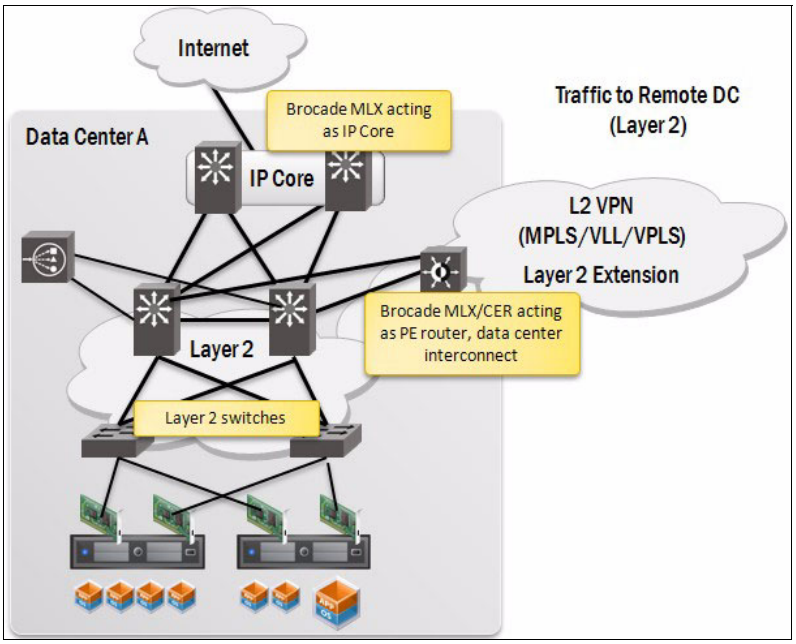


Figure 4-14 High-level IP network architecture

There are three IP networking components that 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 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 flavor of Spanning Tree Protocol. With higher density Top of Rack switches, scalable logical switch capabilities such as stacking or Virtual Switch Clustering, and passive patch paneling in each rack to a higher density End of Row switch, the overall data center network might be collapsed into a flat Layer 2 “Edge” connecting directly into the IP Core over either Layer 2 or Layer 3.
- IP Core: The Layer 2 network connects into the IP core, which then connects out to the WAN Edge/Internet. The example configuration connects from the Layer 2 switches to the IP core over Layer 2. The IP core made up of Brocade MLXe devices that provide aggregation and routing capabilities between the various subnetworks.
- Brocade MLXe/CER acting as the data center interconnect or Provider Edge (PE) router: For Layer 2 extension, the example uses standards-compliant L2 VPN technology that uses MPLs/VPLS/VLL.

The actual lab configuration is shown in Figure 4-15.

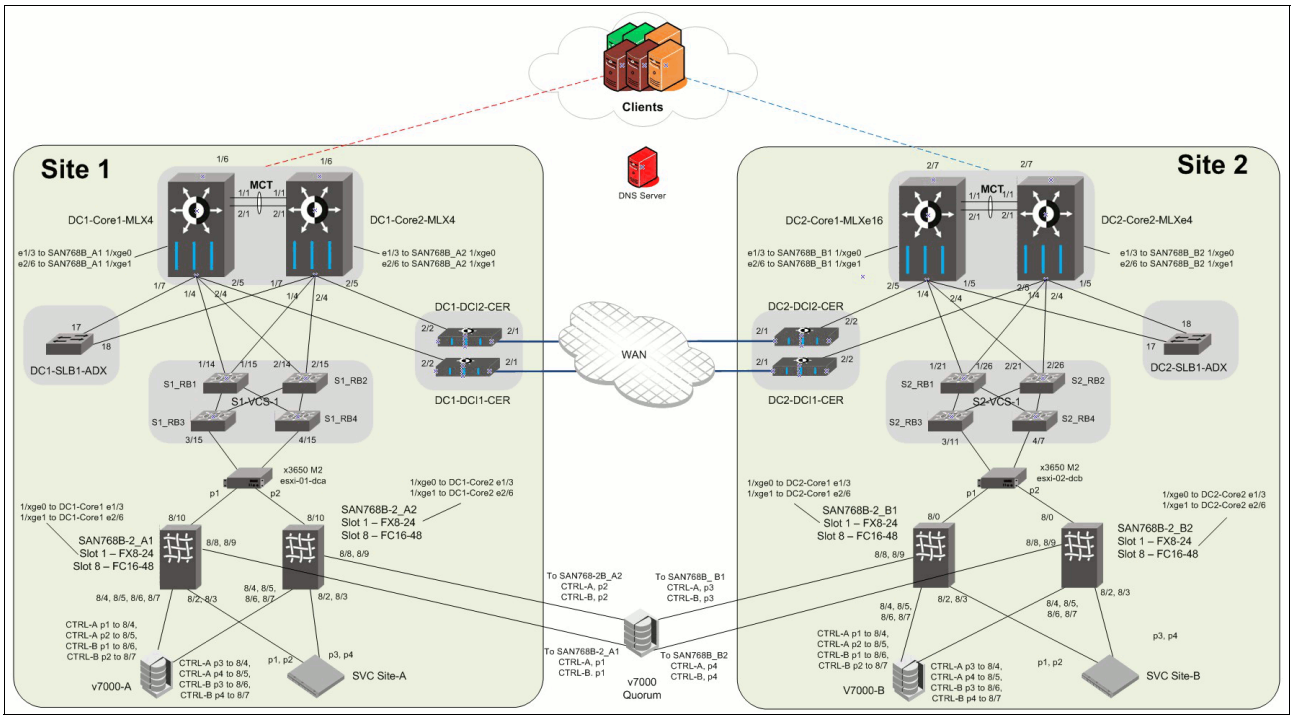


Figure 4-15 IBM SAN Volume Controller Stretched Cluster lab architecture

A closer view of just the Data Center A (Site 1) connection is shown in Figure 4-16.

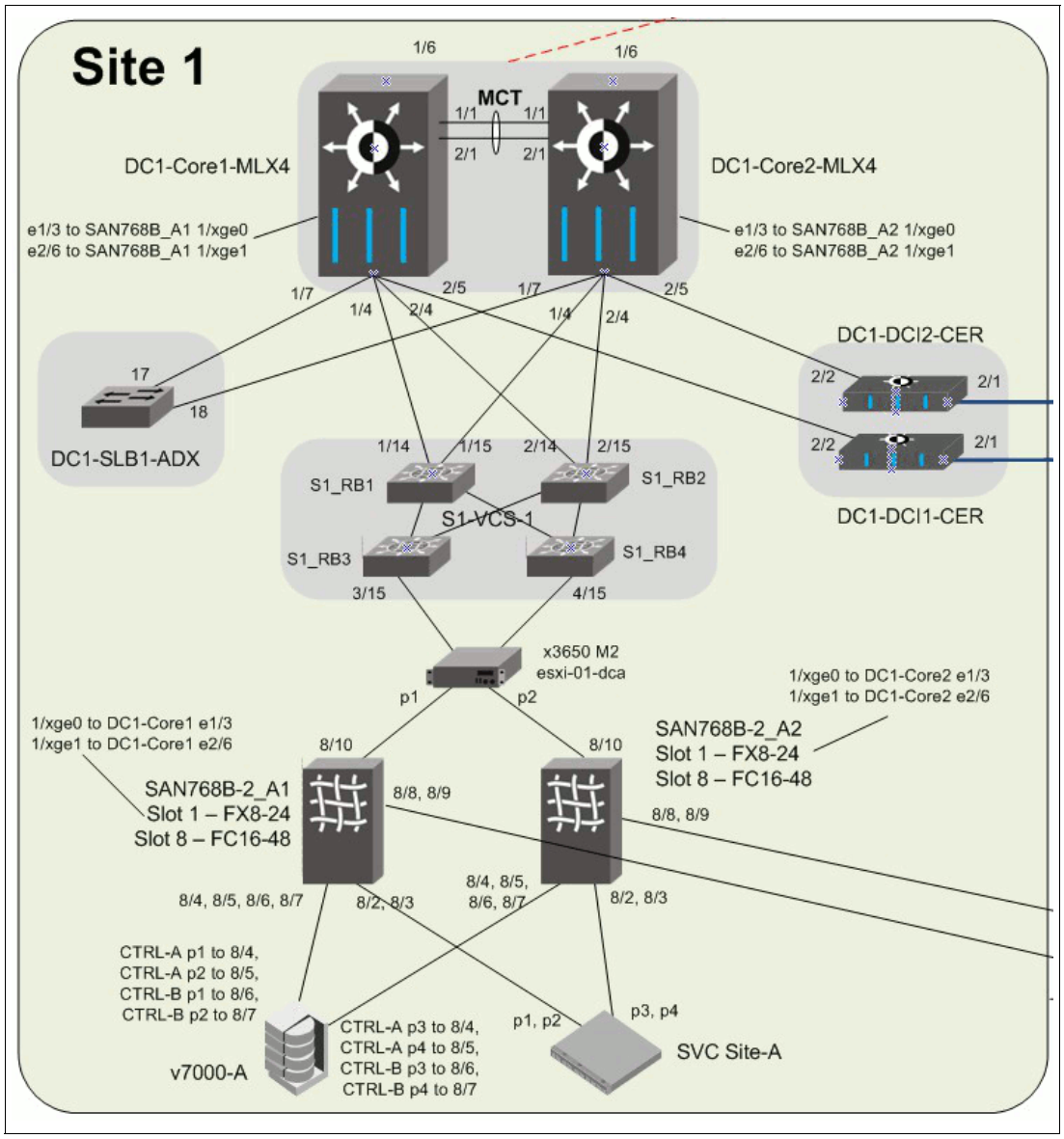


Figure 4-16 IBM SAN Volume Controller Stretched Cluster: Data Center A topology

The example lab configuration consisted of these components:

- An IBM x3650 server that is ESXi host ‘esxi-01-dca1.’ This host has two 10 GbE ports and two 16 Gbps FC ports.
- The 16-Gbps FC ports 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 will cut 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 all on 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(conf-if-te-4/0/15)# switchport mode trunk
S1_RB4(conf-if-te-4/0/15)# switchport trunk allowed vlan add 700-703
```

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

Now that the port channels are created on both switches, perform 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 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, including the ESXi host configuration, is the same. Otherwise, fragmentation might occur.

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

- Hardening the switch by using RBAC (for example, RADIUS/TACACS+), user accounts, disabling Telnet, and restricting login to certain VLANs or IP ranges
- Configuring SNMP, Syslog, or sFlow monitoring servers
- Configuring an NTP server
- Enabling LLDP and CDP, or extra network visibility
- Configuring quality of service for specific traffic classes, or trusting the 802.1p settings that are sent by ESXi
- Configuring security and Access Control Lists

Additional references

More references for the Brocade VDX can be found in the following manuals:

- *Network OS Administrator's Guide supporting v2.1.1*
- *Network OS Command Reference Manual support v2.1.1*

4.3.2 IP Core (MLXe) configuration

Keep in mind the following considerations when configuring IP Core (MLXe).

VLAN configuration

Start by defining the VLANs that the MLXe will be carrying traffic over. These are the same four VLANS, 700-703, as defined in the Layer 2 switch network. However, the MLXe will also be carrying FCIP traffic as follows:

- 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 essentially three different types of network traffic access patterns for traffic that passes through the MLXe. The following is an example of the different network traffic types and interfaces applicable on DC1-Core1-MLXe4:

- Traffic that only needs access to 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.

Additionally, whereas the Management traffic might need to be routed to a different subnetwork, hence the creation of 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 and also the external client traffic (eth 1/6) and to the ServerIron 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 MLXes (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 VRRPe 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 will pass 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. Do 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 that is the primary port in the LAG.

Next, configure the virtual routing interface with an IP address that the two MLXes will 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 details 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 capable of handling the entire Internet routing table with advanced MPLS/VPLS/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

More references for the Brocade MLXe can be found in the *Brocade MLX Series and NetIron Family Configuration Guide supporting r05.3.00*.

4.3.3 Data Center Interconnect (CER) 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/VLL technology. This is typically a service that is 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/VLL capabilities.

It is outside of the intended 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 as shown for reference in Example 4-22 and Example 4-23 on page 72.

Example 4-22 DC1-DCI1-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 DCIvl700 700
  vll-peer 90.90.90.10
  vlan 700
   tagged e 2/2

 vll DCIvl701 701
  vll-peer 90.90.90.10
  vlan 701
   tagged e 2/2

 vll DCIvl702 702
  vll-peer 90.90.90.10
  vlan 702
   tagged e 2/2

 vll DCIvl703 703
  vll-peer 90.90.90.10
  vlan 703
   tagged e 2/2

 vll DCIvl705 705
  vll-peer 90.90.90.10
  vlan 705
   tagged e 2/2

 vll DCIvl706 706
```

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

 vll DCIvl707 707
  vll-peer 90.90.90.10
  vlan 707
   tagged e 2/2

 vll DCIvl708 708
  vll-peer 90.90.90.10
  vlan 708
   tagged e 2/2
```

Example 4-23 shows the configuration of SC2-CDCI1-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 DCIvl700 700
  vll-peer 80.80.80.10
  vlan 700
   tagged e 2/2

 vll DCIvl701 701
  vll-peer 80.80.80.10
  vlan 701
   tagged e 2/2

 vll DCIvl702 702
  vll-peer 80.80.80.10
  vlan 702
   tagged e 2/2
```

```
vll DCIvl703 703
 vll-peer 80.80.80.10
 vlan 703
  tagged e 2/2

vll DCIvl705 705
 vll-peer 80.80.80.10
 vlan 705
  tagged e 2/2

vll DCIvl706 706
 vll-peer 80.80.80.10
 vlan 706
  tagged e 2/2

vll DCIvl707 707
 vll-peer 80.80.80.10
 vlan 707
  tagged e 2/2

vll DCIvl708 708
 vll-peer 80.80.80.10
 vlan 708
  tagged e 2/2
```

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

4.4 IBM FC SAN

The following section assumes that you are familiar with general FC SAN design and technologies. Typically, SAN design has servers and storage that are connected into dual, redundant fabrics. The lab configuration had a redundant fabric design that used two IBM SAN768B-2 chassis at each data center site. Each IBM SAN768B-2 was equipped with these components:

- IBM FC 8-Gbps FCIP Extension blade in Slot 1
- IBM FC 16 Gbps 48-port blade in Slot 8

IBM SAN Volume Controller Stretched Cluster requires two types of SAN:

- Public SAN: Where server hosts, storage, and SAN Volume Controller nodes connect. Data storage traffic traverses the Public SAN.
- Private SAN: Only the SAN Volume Controller 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 4-17 shows the various Public and Private SAN connections in different colors.

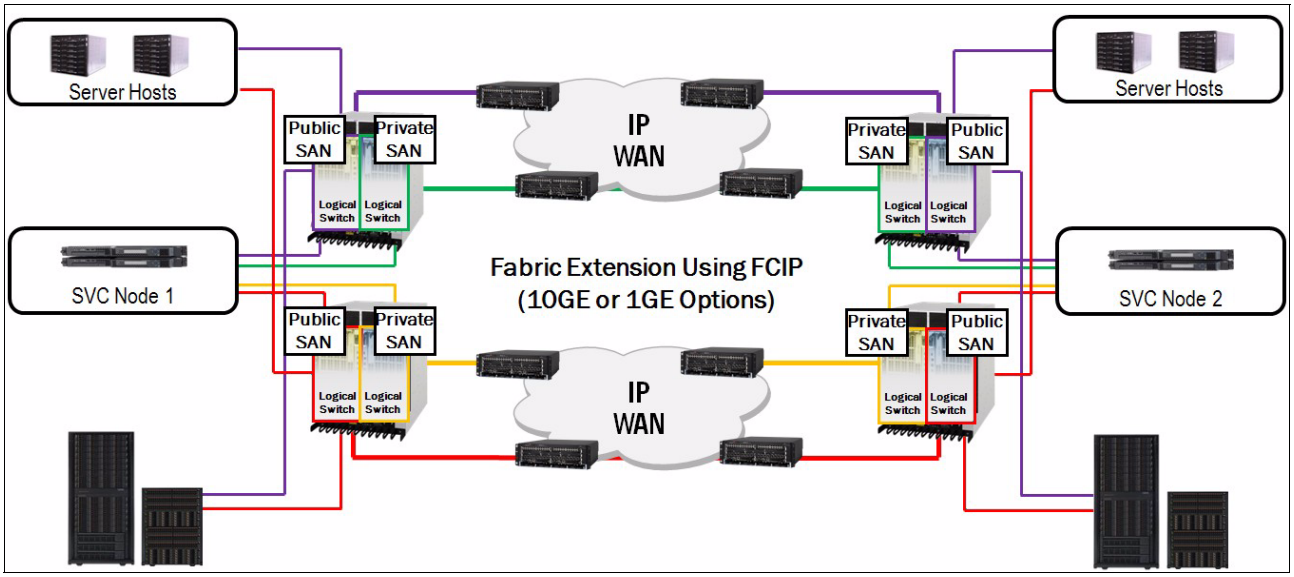


Figure 4-17 Public and Private SAN connections at each data center site

The actual lab environment is shown in Figure 4-18.

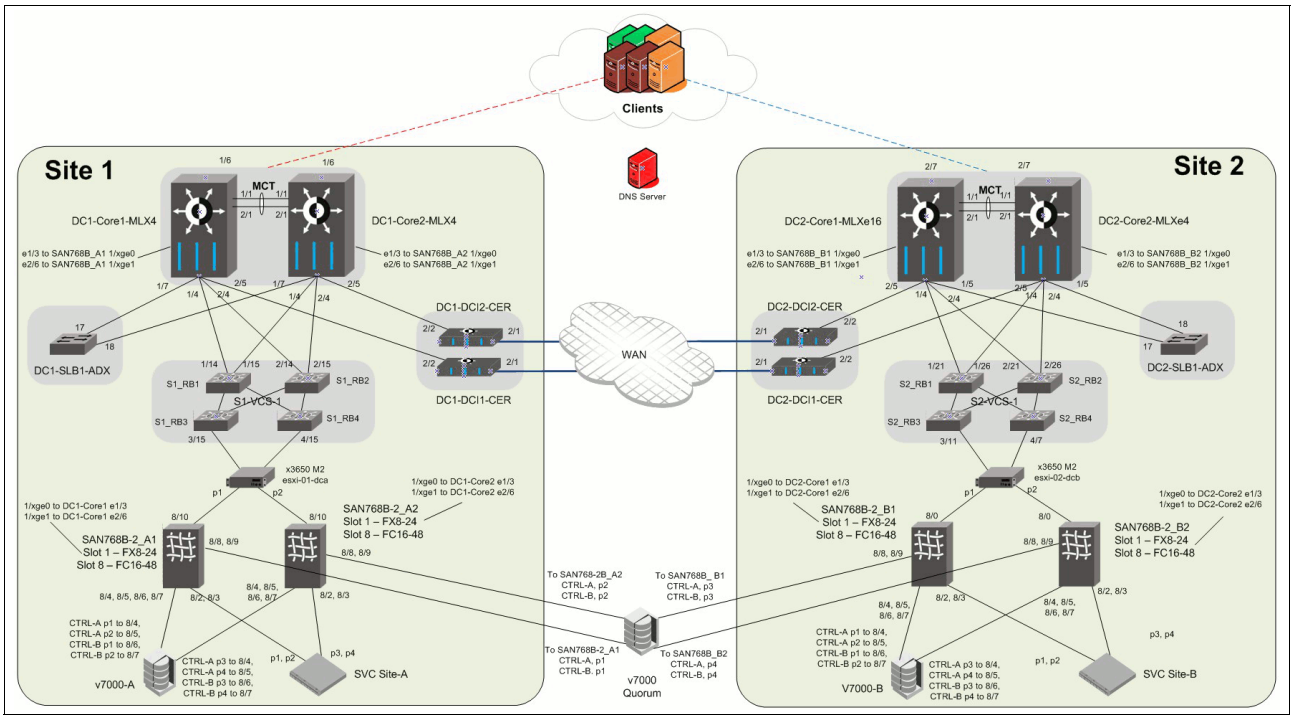


Figure 4-18 IBM SAN Volume Controller Stretched Cluster lab topology

4.4.1 Creating the logical switches

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

Figure 4-19 shows a closer picture of the SAN port topology in Data Center A.

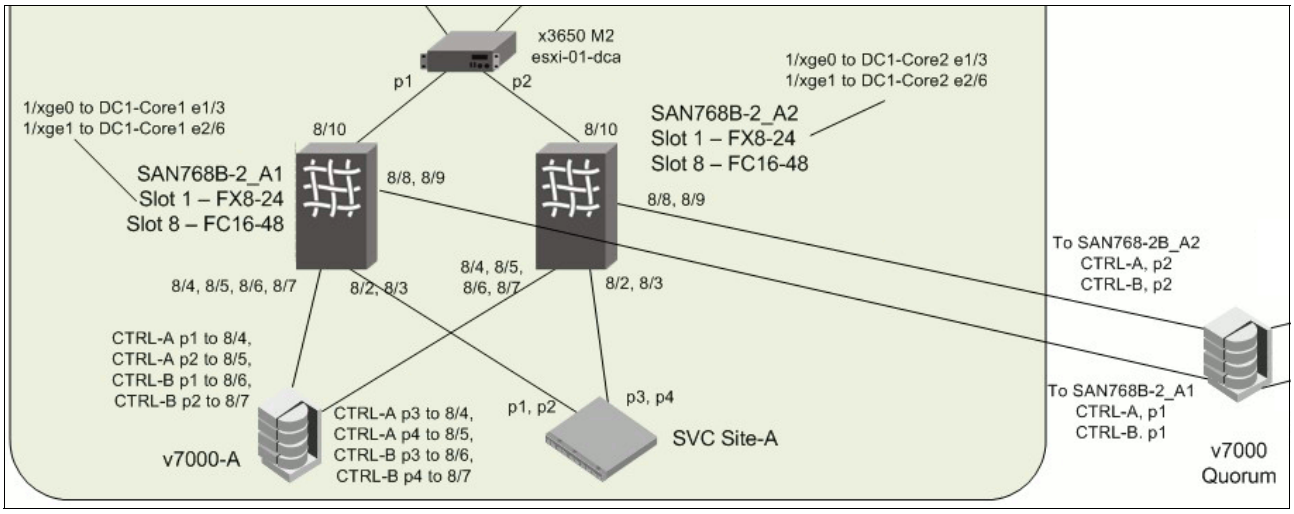


Figure 4-19 Data Center A: SAN port topology

Figure 4-20 shows a closer picture of the SAN port topology in Data Center B.

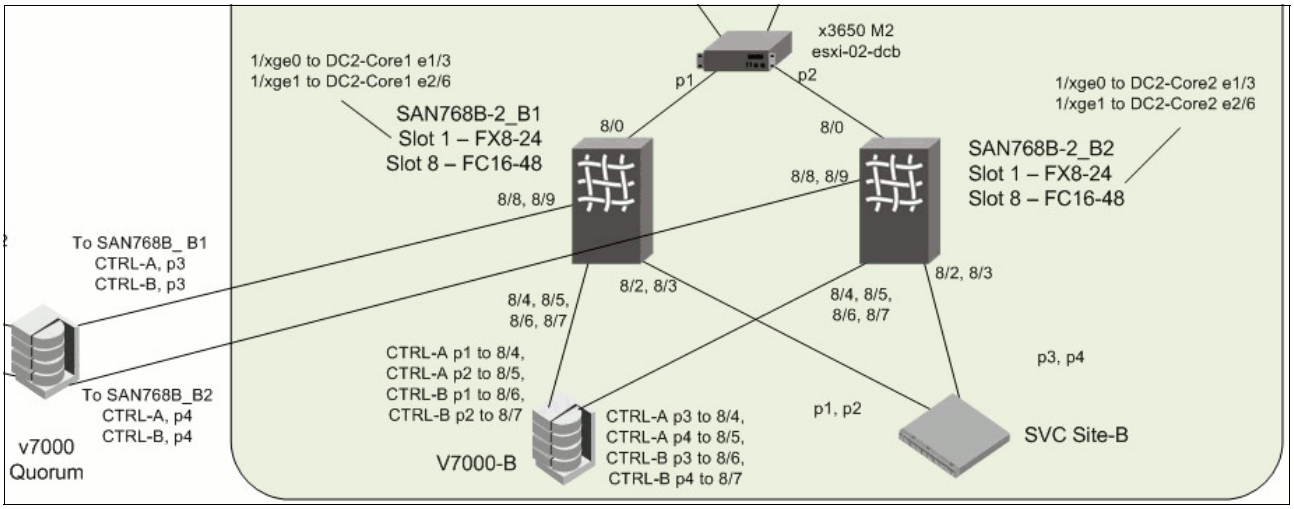


Figure 4-20 Data Center B: SAN port topology

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 Name	Physical Switch	Logical Switch	LS #	Ports
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.

The following are two examples of creating a logical switch through the CLI and through IBM Network Advisor. 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 logical switch Public_A1 on SAN768B-2_A1.

Example 4-24 Creating the logical switch Public_A1

```
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 logical switch Public_B1 on SAN768B-2_B1 by using IBM Network Advisor. 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 4-21.

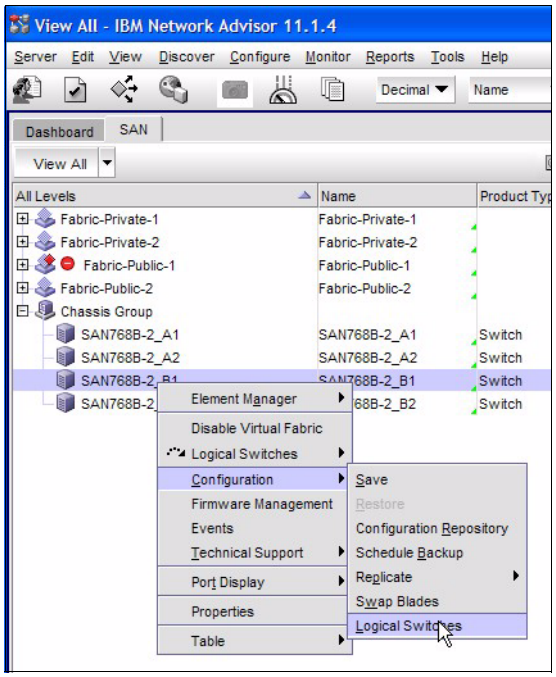


Figure 4-21 Entering the Logical Switches configuration menu

Clarification: In Figure 4-21, 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, which is a single-switch fabric that is named Fabric-Public-1A for now.

Now, create Public_B1 in Data Center B and then the corresponding FCIP tunnels for this fabric to merge.

In the Logical Switches window, check to make sure that the Chassis you selected is SAN768B-2_B1. Then, select **Undiscovered Logical Switches** in the right window and click **New Switch** as shown in Figure 4-22.

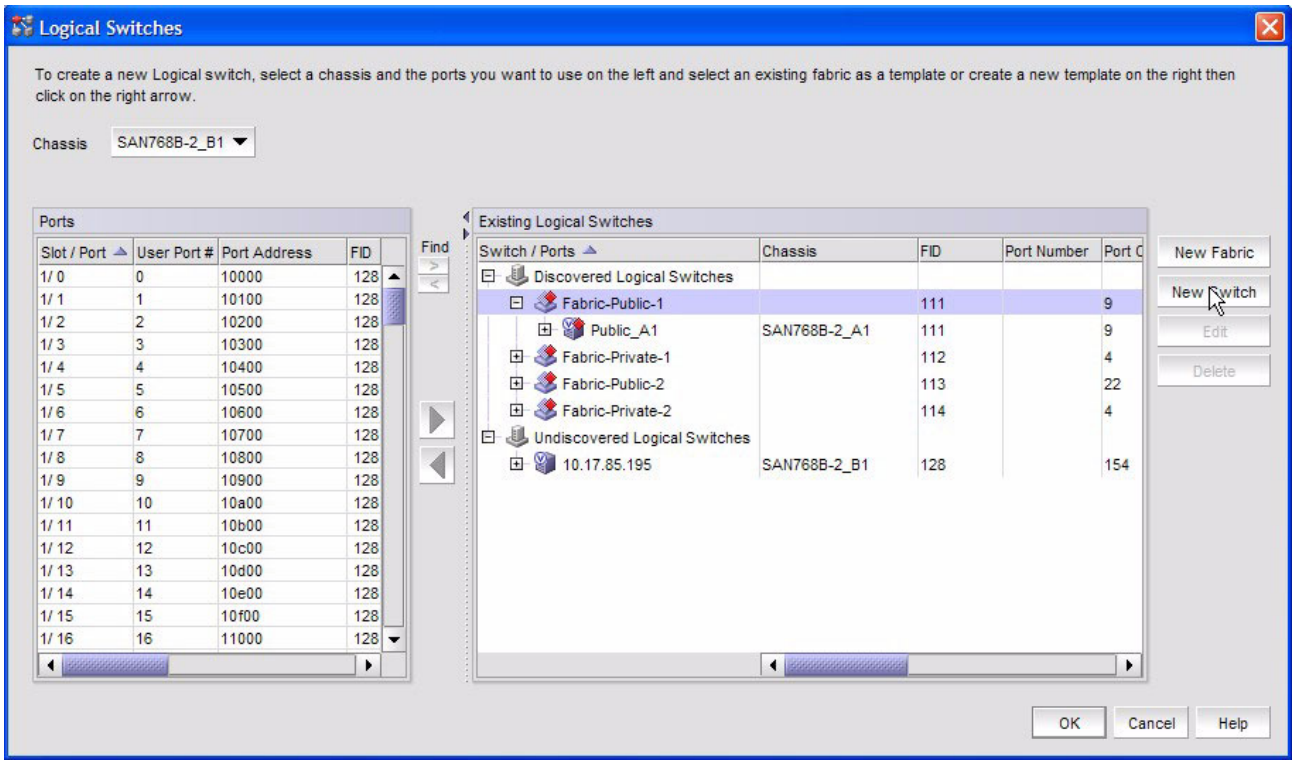


Figure 4-22 Creating a new logical switch on SAN768B-2_B1

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 will eventually merge with. However, set it the same to remain consistent. This process is shown in Figure 4-23.

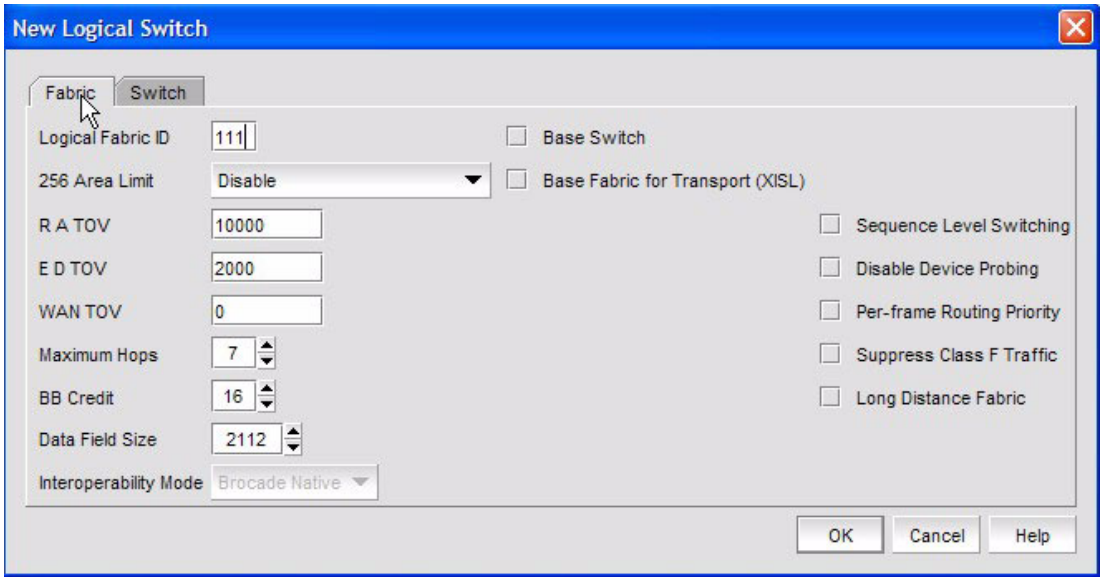


Figure 4-23 Setting the Logical Fabric ID

Next, click the Switch tab and provide a switch name of **Public_B1** as shown in Figure 4-24.

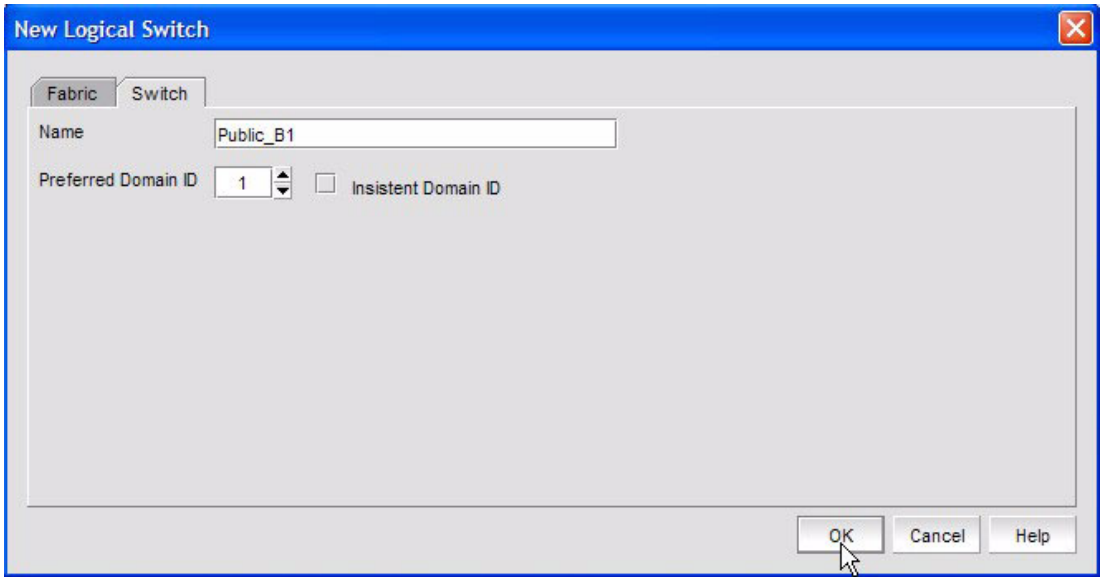


Figure 4-24 Setting the switch name for Public_B1

Now that the new logical switch construct is created, move the ports from SAN768B-2_B1 to Public_B1 as shown in Figure 4-25.

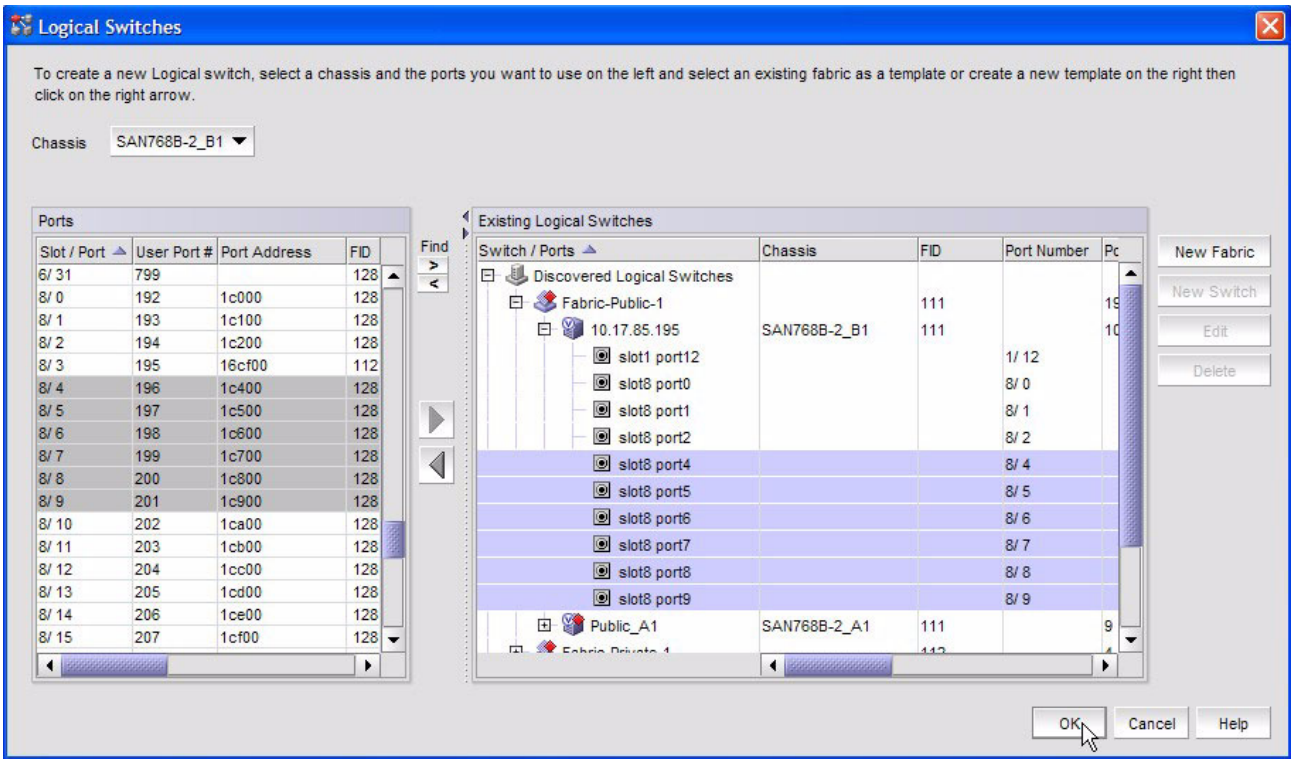


Figure 4-25 Selecting the ports from SAN768B-2_B1 to move to logical switch Public_B1

Review the information in the confirmation window, and then click **Start** begin the process. Figure 4-26 shows the message that confirms the creation.

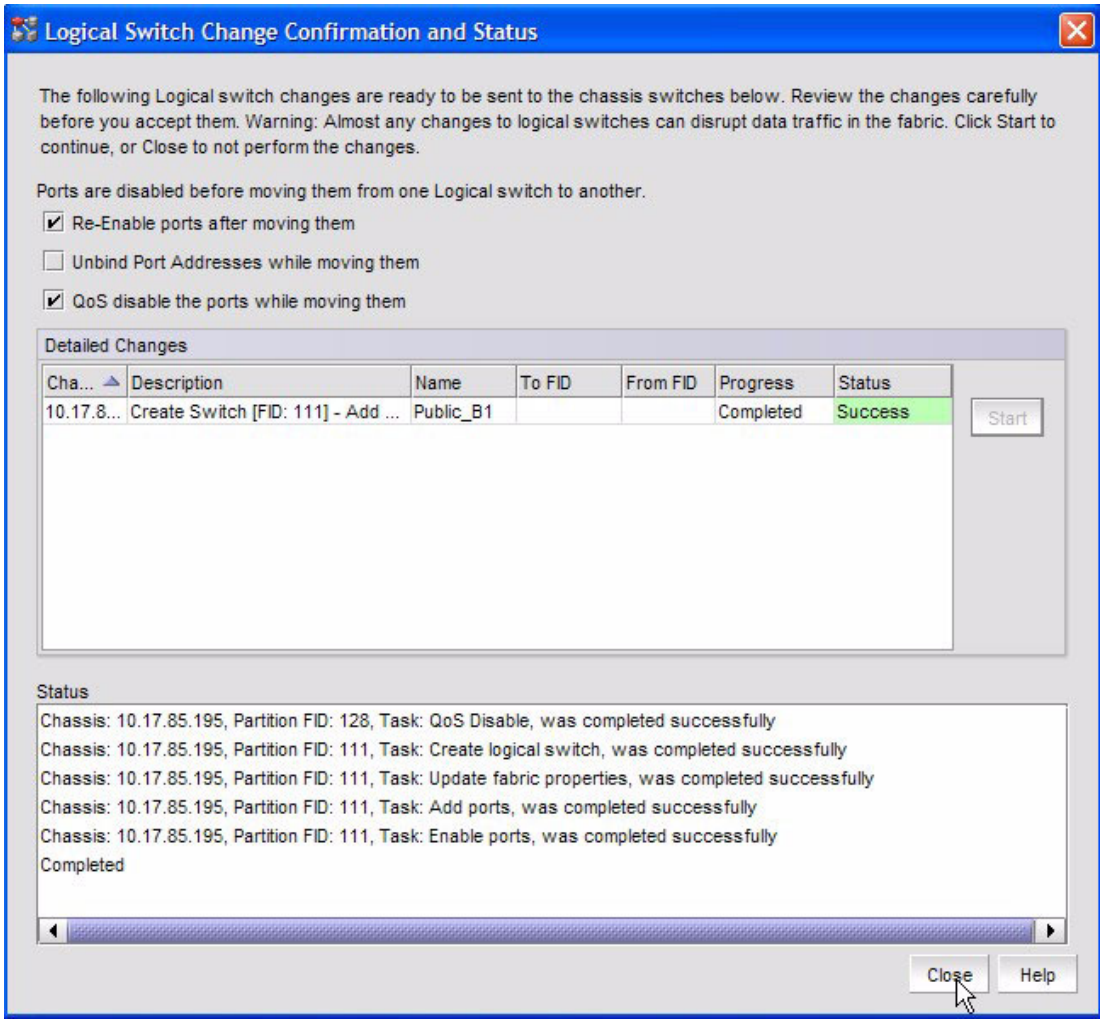


Figure 4-26 Successfully creating Public_B1

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 4-27.

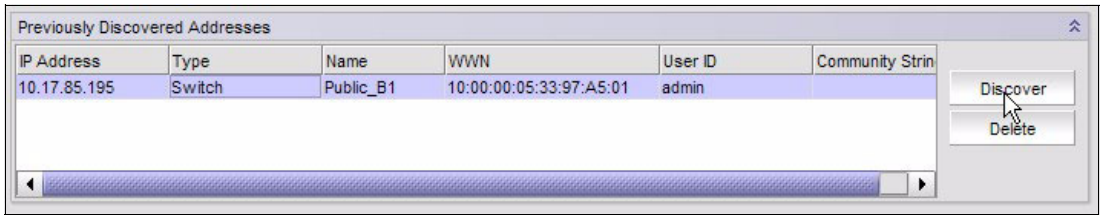


Figure 4-27 Rediscovering SAN768B-2_B1 to add Public_B1

In the Fabric Discovery window, name the new fabric **Fabric-Public-1B** for now as shown in Figure 4-28.

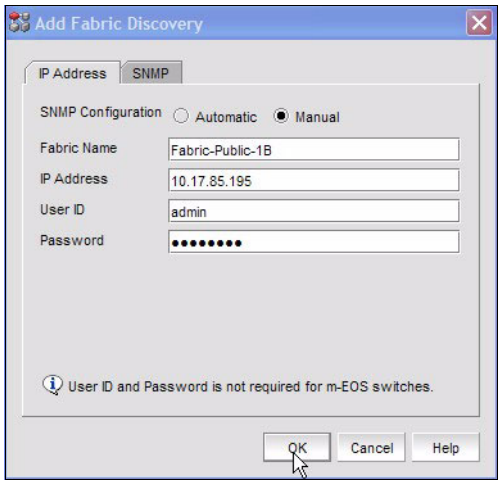


Figure 4-28 Discovering Fabric-Public-1B

Choose to only discover and monitor Public_B1 and not the base SAN768B-2_B1 switch as shown in Figure 4-29.

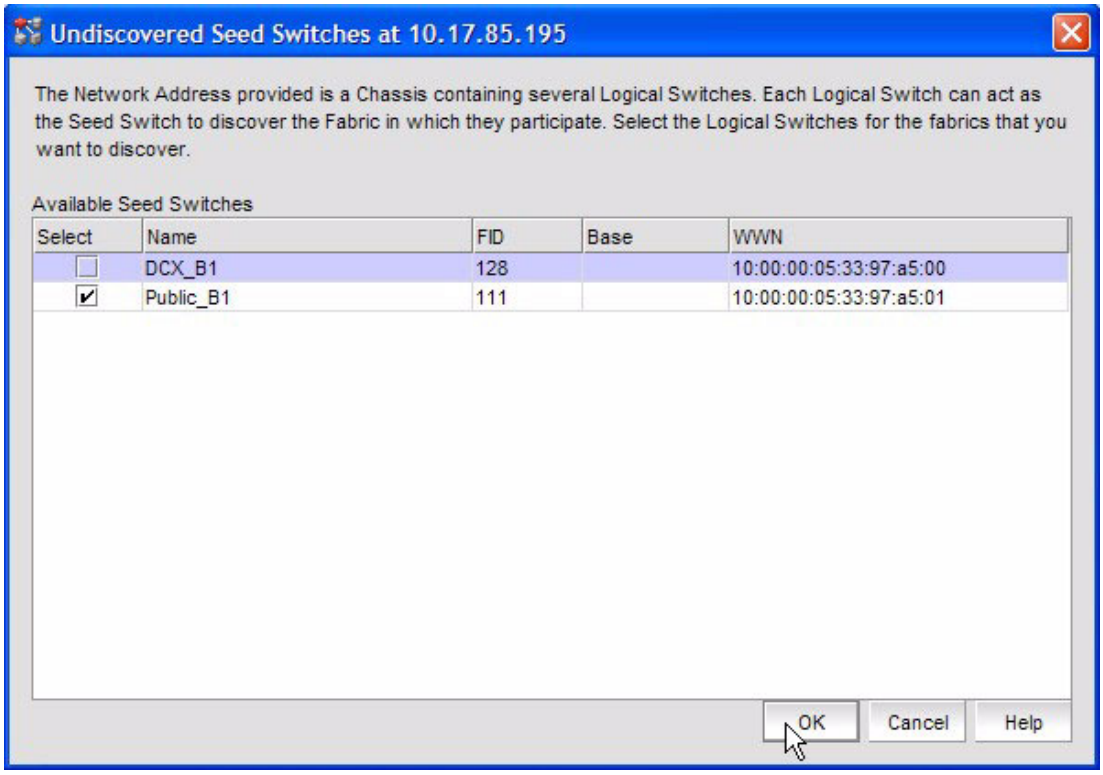


Figure 4-29 Selecting to monitor only Public_B1

Fabric-Public-1B with Public_B1 should now be visible in the SAN view as seen in Figure 4-30.

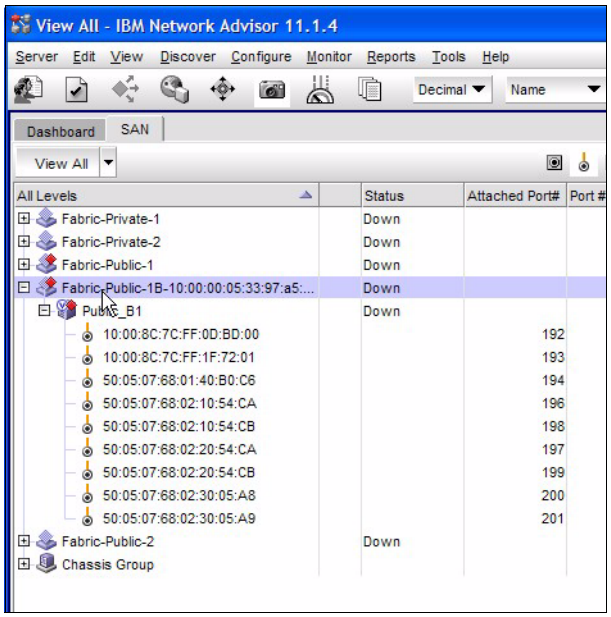


Figure 4-30 Fabric-Public-1B successfully discovered

4.4.2 Creating FCIP tunnels

This section addresses how to create the FCIP tunnel connectivity through the 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 off 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 --traceroute 1/xge1 -s 192.168.76.10 -d
192.168.76.20
Traceroute 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
Traceroute 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 SAN Volume Controller 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 validate that the tunnel was established as seen in Example 4-29.

Example 4-29 Creating an FCIP tunnel on Public_B1 and validating establishment

```
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 fciptunnel 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 IBM Storage Volume Controller using Stretched Cluster

The SAN Volume Controller code that is used is based on code level 6.4.0.2 (build 65.0.1207120000). The back-end storage that is used in the example is a Storwize V7000 running a code level of 6.4.0.2 (build 65.0.1207120000). The third site, the quorum storage, was on a Storwize V7000 running a code level of 6.4.0.2 (build 65.0.1207120000).

For a full list of supported extended Quorum devices, see the IBM Support site at:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S1003907

Figure 4-31 shows the components that are used for the SAN Volume Controller Stretched Cluster and the SAN connectivity.

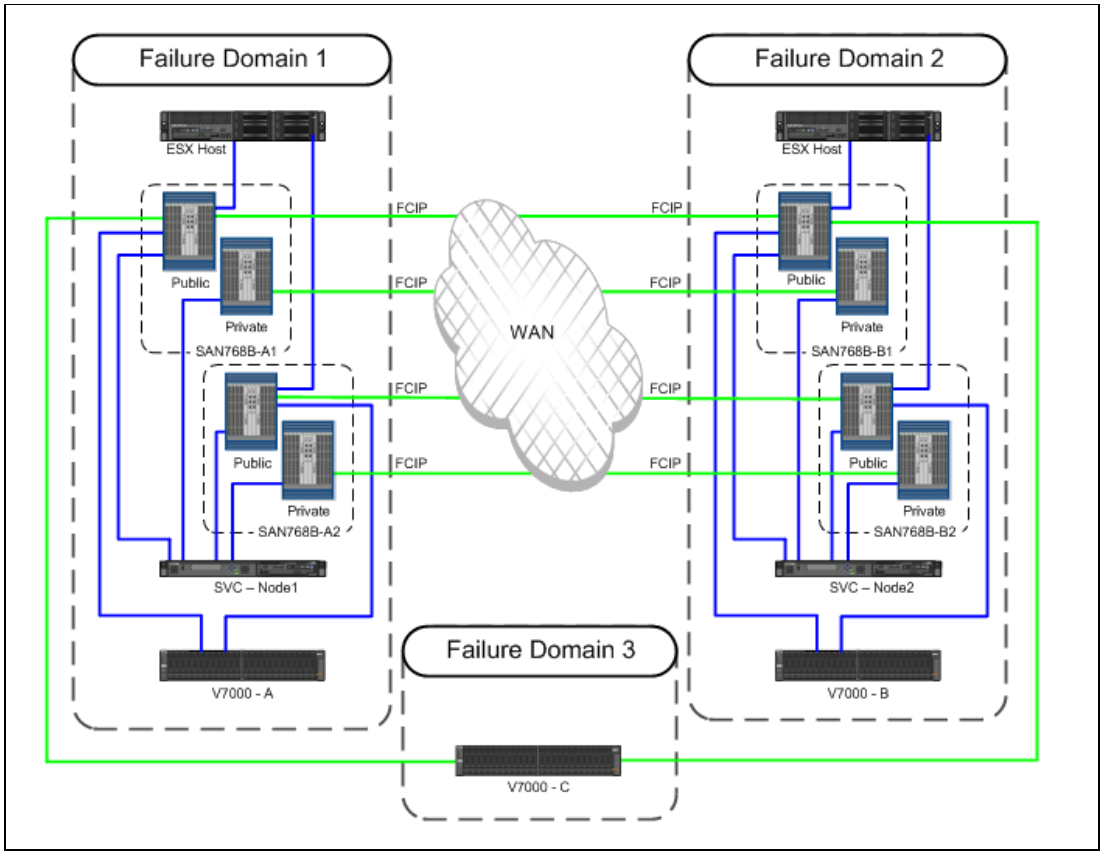


Figure 4-31 SAN Volume Controller Stretched Cluster diagram with SAN connectivity

This book does not cover the physical installation nor the initial configuration. It is intended to supplement the *V6.3.0 Configuration Guidelines for Extended Distance Split-System Configurations for IBM System Storage SAN Volume Controller* available at:

http://www-01.ibm.com/support/docview.wss?&uid=ssg1S7003701

This book assumes that you are familiar with the major concepts of SAN Volume Controller clusters such as nodes, I/O groups, MDisks, and quorum disks. If you are not, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

For more information, see the IBM SAN Volume Controller Information Center at:

http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp

4.6 SAN Volume Controller volume mirroring

The Stretched Cluster I/O group uses SAN Volume Controller 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. Thus, volume mirroring can minimize the impact to volume availability if one or more MDisks fails. The resynchronization between both copies is incremental. SAN Volume Controller starts the resynchronization process automatically.

A mirrored volume has the same functions and behavior as a standard volume. In the SAN Volume Controller software stack, volume mirroring is below the cache and 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 be done with a mirrored volume as well, including migration and expand/shrink. Like a standard volume, each mirrored volume is owned by one I/O group with a preferred node. Thus the mirrored volume goes offline if the whole I/O group goes offline. The preferred node runs all IO 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. Thus, an active quorum disk is required for volume mirroring. To ensure data consistency, SAN Volume Controller 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 prior usage of 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, and so one bit of the synchronization bitmap represents 256 KB of virtual capacity. Therefore, 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 designated as the primary for all read operations. SAN Volume Controller 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, which can be changed by the command **chvdisk -primary**.

4.8 Write operations

Write operations are run on all copies. The storage system with the lowest performance determines the response time between SAN Volume Controller and the storage system

back-end. The SAN Volume Controller cache is able to 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 SAN Volume Controller copy services. As soon as the missing copy is back, SAN Volume Controller evaluates the change bitmap 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. The resynchronization bandwidth can be controlled with the command **chvolume -syncrate**. Host access to the volume continues during that time. This behavior can cause difficulties in a site failure. Since version 6.2, SAN Volume Controller provides the volume attribute **-mirrorwritepriority** to prioritize between strict data redundancy (**-mirrorwritepriority redundancy**) and best performance (**-mirrorwritepriority latency**) for the volume. The default setting is **-mirrorwritepriority latency** to maintain compatibility with earlier versions.

4.9 SAN Volume Controller quorum disk

The quorum disk fulfills two functions for cluster reliability:

- Act as tiebreaker in split brain scenarios
- Save critical configuration metadata

The SAN Volume Controller quorum algorithm distinguishes between the active quorum disk and quorum disk candidates. There are three quorum disk candidates. Only one of these candidates acts as the active quorum disk. The other two are reserved, and 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 and is used in T3 recovery.

Attention: Place each quorum disk in one site (failure domain). Set the quorum disk in the third site (quorum site) as the active quorum disk.

4.10 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 SAN Volume Controller nodes. Inside an SAN Volume Controller node, the quorum disk cannot act as a tie-breaker. Therefore, only Managed Disks (MDisks) from external storage system are selected as SAN Volume Controller 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 version 6.1, SAN Volume Controller 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. SAN Volume Controller 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 are in the correct sites, change the quorum disk candidates by using the command **chquorum**.

Starting with SAN Volume Controller 6.2, the quorum disk selection algorithm changed. SAN Volume Controller 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 extended quorum disk. Storage systems that provide extended quorum support are listed at:

http://www.ibm.com/storage/support/2145

4.11 Automatic SAN Volume Controller quorum disk selection

The CLI output in Example 4-30 shows that the SAN Volume Controller cluster initially has automatically assigned the quorum disks.

Example 4-30 Quorum disks assigned.

```
IBM_2145:ITSO_SVC_SPLIT: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 yes
mdisk       no
1            online 2  ITSO_V7K_SITEA_SAS1 1             ITSO_V7K_SITEA_N2 no
mdisk       no
2            online 3  ITSO_V7K_SITEA_SAS2 1             ITSO_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:ITSO_SVC_SPLIT:superuser>chquorum -override yes -mdisk 110
IBM_2145:ITSO_SVC_SPLIT:superuser>chquorum -override yes -mdisk 1 1
IBM_2145:ITSO_SVC_SPLIT:superuser>chquorum -override yes -mdisk 6 2
```

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

Example 4-32 Quorum changed

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name                controller_id controller_name
active object_type override
0            online 10 ITSO_V7K_SITEC_Q    5             ITSO_V7K_SITEC_Q_N2 yes
mdisk       yes
1            online 6  ITSO_V7K_SITEB_SAS0 0             ITSO_V7K_SITEB_N2   no
mdisk       yes
2            online 4  ITSO_V7K_SITEA_SAS3 1             ITSO_V7K_SITEA_N2   no
mdisk       yes
```

The output shows that the controller named ITSO_V7K_SITEC_Q, 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 Figure 4-32.

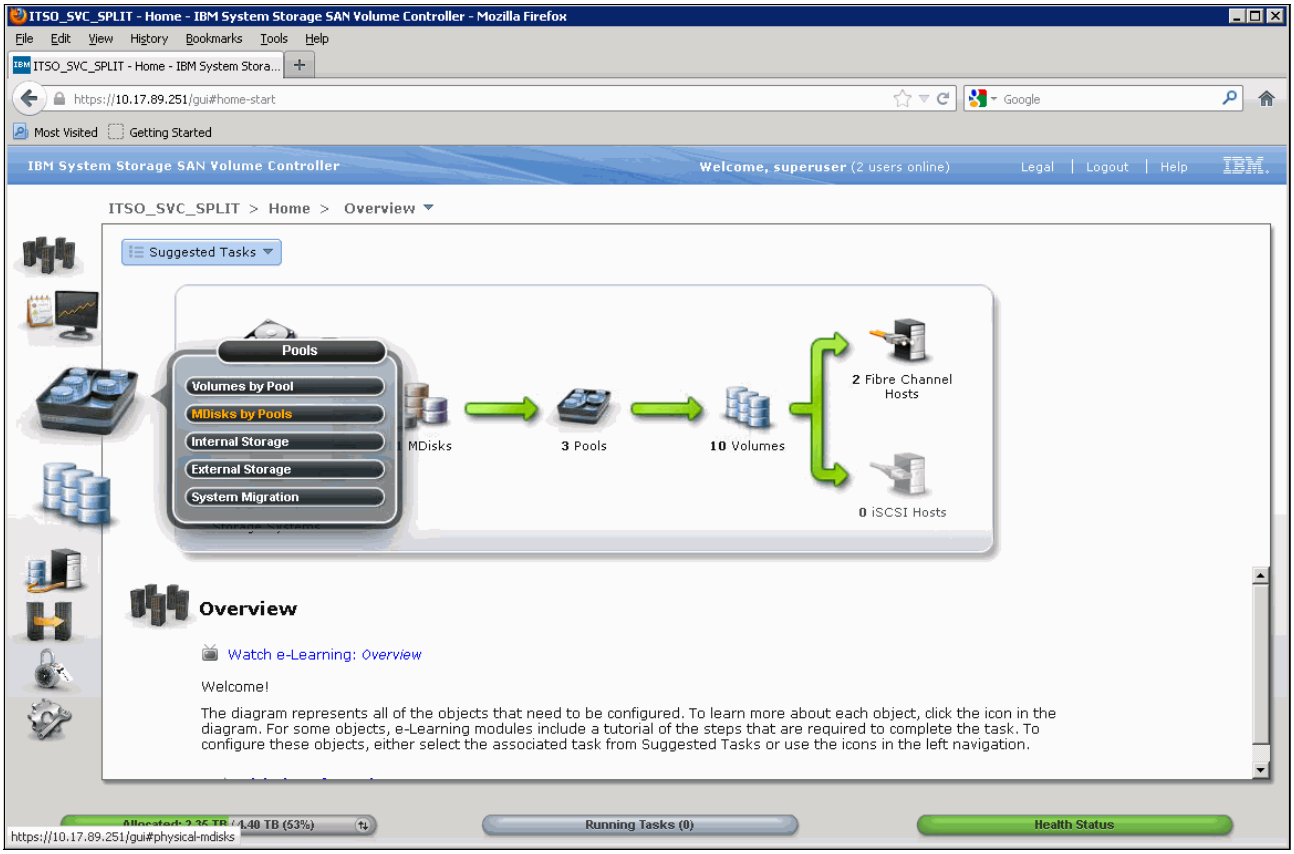


Figure 4-32 Opening the MDisk by using the Pools view

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 Ctrl down and selecting the three candidates. When the candidates are selected, right-click them and select **Quorum** → **Edit Quorum**, as shown in Figure 4-33.

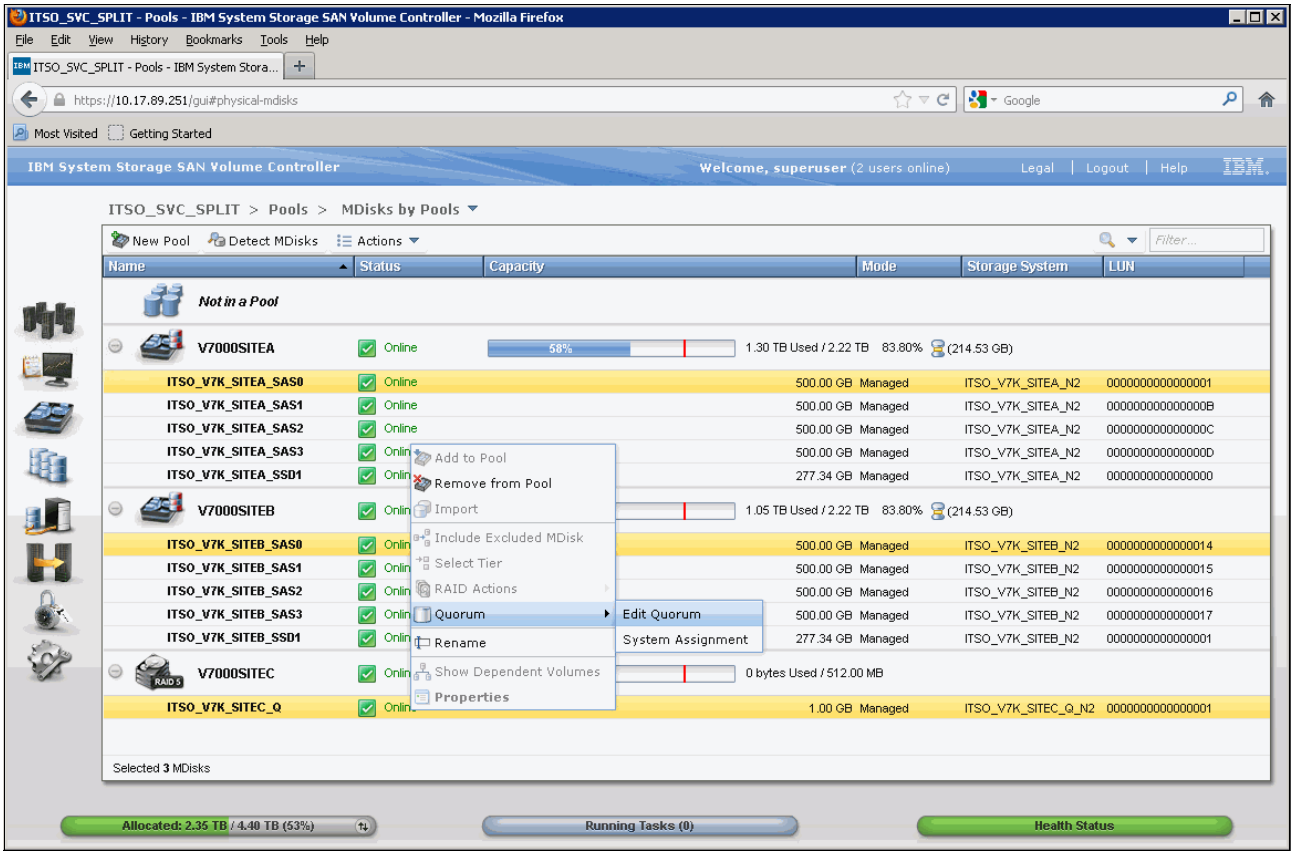


Figure 4-33 Yellow color shows the MDisks that are selected for quorum

Choose the MDisks for quorum assignment. In this example, select a split-site configuration and an MDisk that is in site 3 as shown in Figure 4-34.

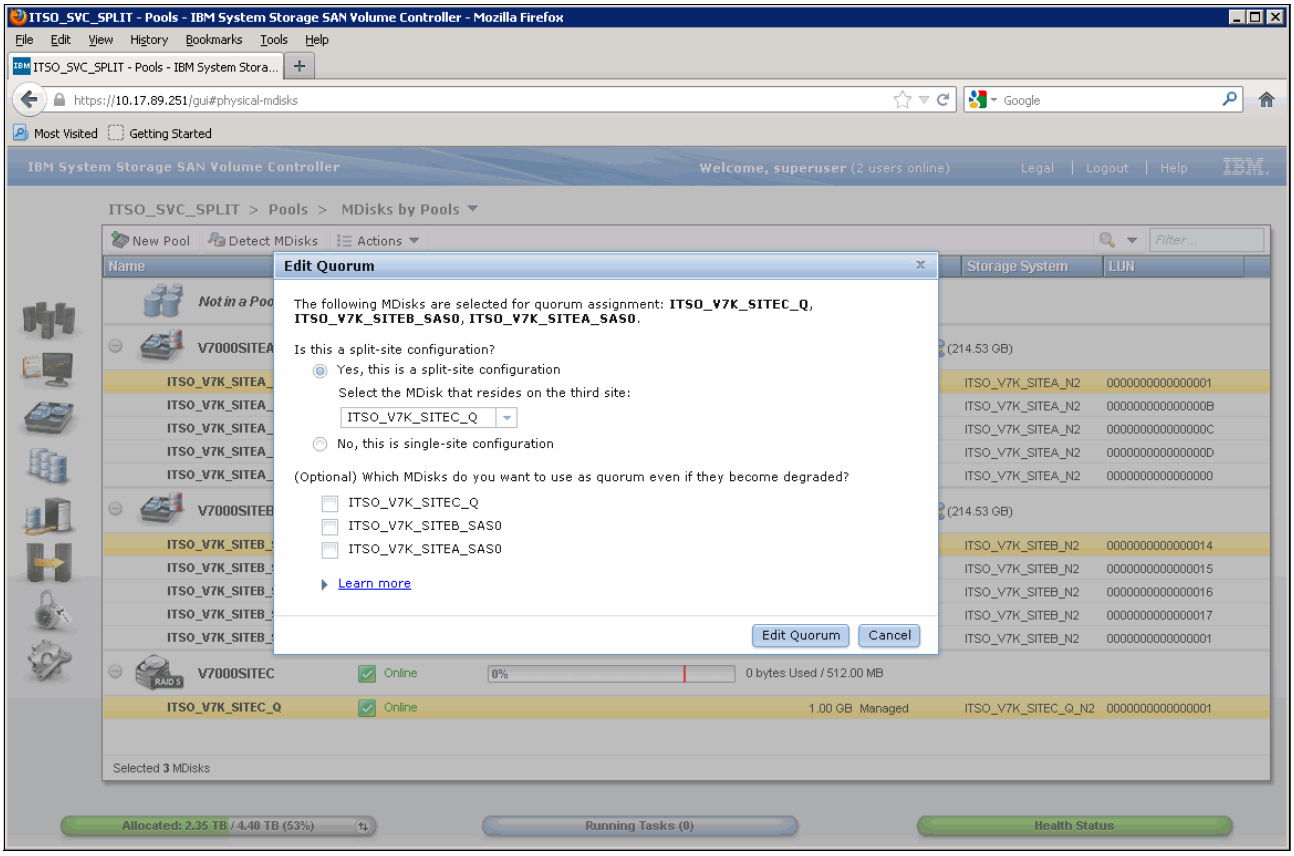


Figure 4-34 MDisk that is selected for site 3 (Failure Domain 3)

Finally, click **Edit** to manually assign the quorum disks in the GUI, as shown in Figure 4-35.

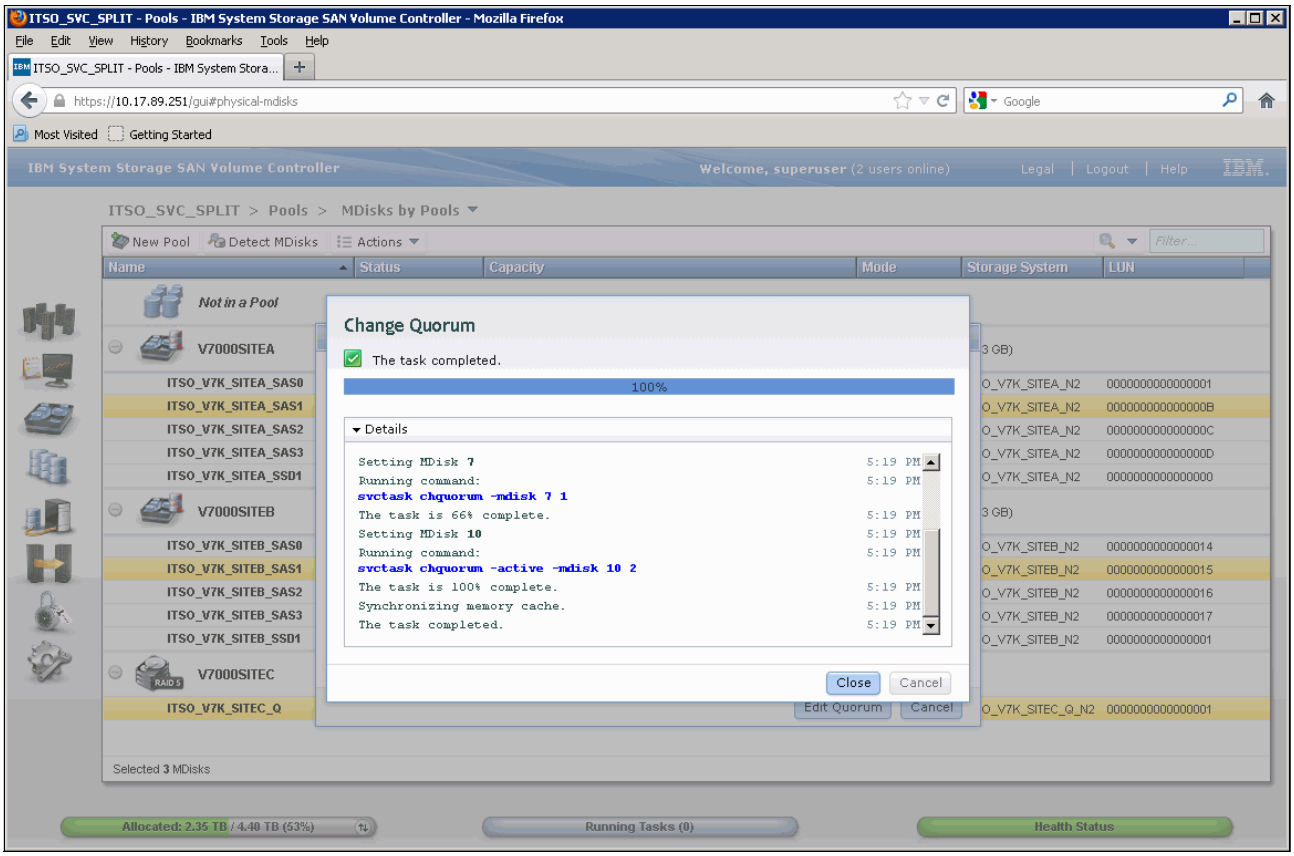


Figure 4-35 Quorum disk is assigned

4.12 Backend Storage allocation to the SAN Volume Controller Cluster

For Power Domain 1/ Site 1 and Power Domain 2/ Site 2, the example uses Storwize V7000 for backend storage. Both V7000 are configured the same way. Power Domain 3 / Site 3 has a Storwize V7000 acting as the active Quorum disk.

For more information about how to implement Storwize V7000, see *Implementing the IBM Storwize V7000 V6.3*, SG24-7938. Also, see the Storwize V7000 information center website at:

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

Figure 4-36 shows the MDisks created in Power Domain 1.

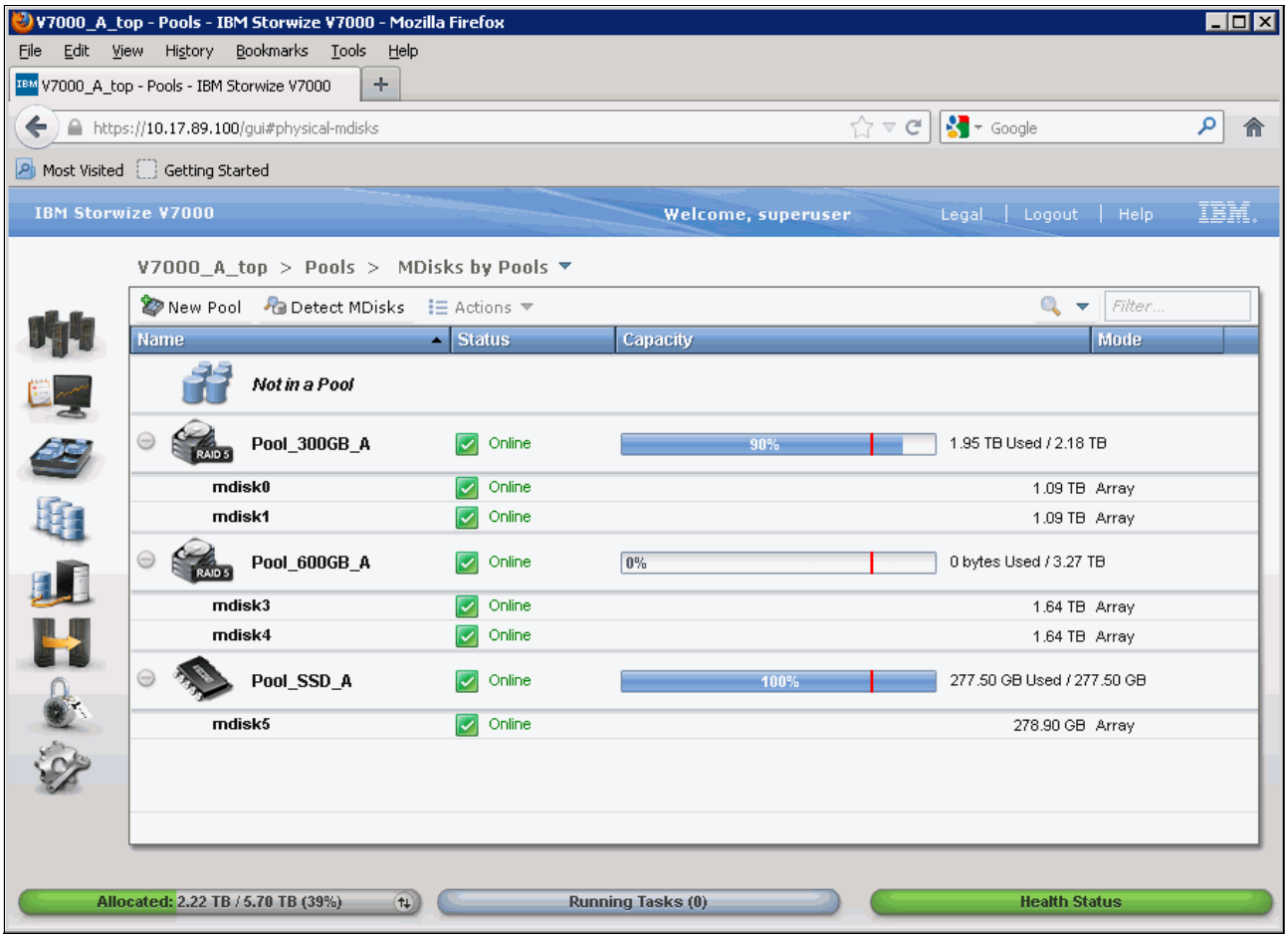


Figure 4-36 Showing the MDisks that have been created in the V7000

Figure 4-37 shows the volumes that are assigned to the SAN Volume Controller Stretched Cluster host.

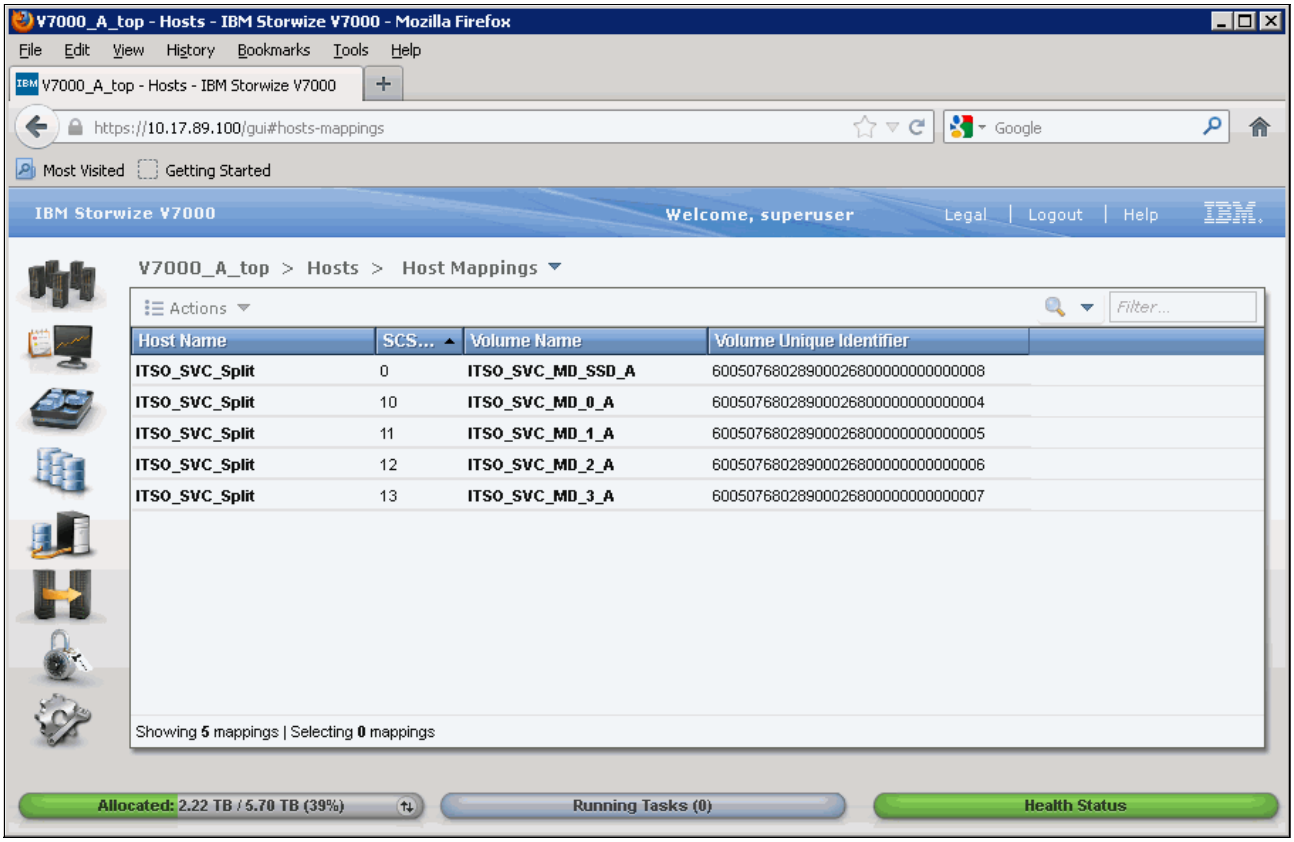


Figure 4-37 Volume assignment to the SAN Volume Controller Stretched cluster host

4.13 Volume allocation

The volume allocation using the SAN Volume Controller Stretched Cluster solution is explained here. All volume assignments are 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.

Furthermore, copy 0 of the volume mirror (also referred to as the primary copy), must also be in Power Domain 1 / Site 1. This configuration ensures that during normal operations, there are no unnecessary roundtrips for the I/O operations.

Figure 4-38 shows volumes that are assigned to Host ESXI-01-DCA.

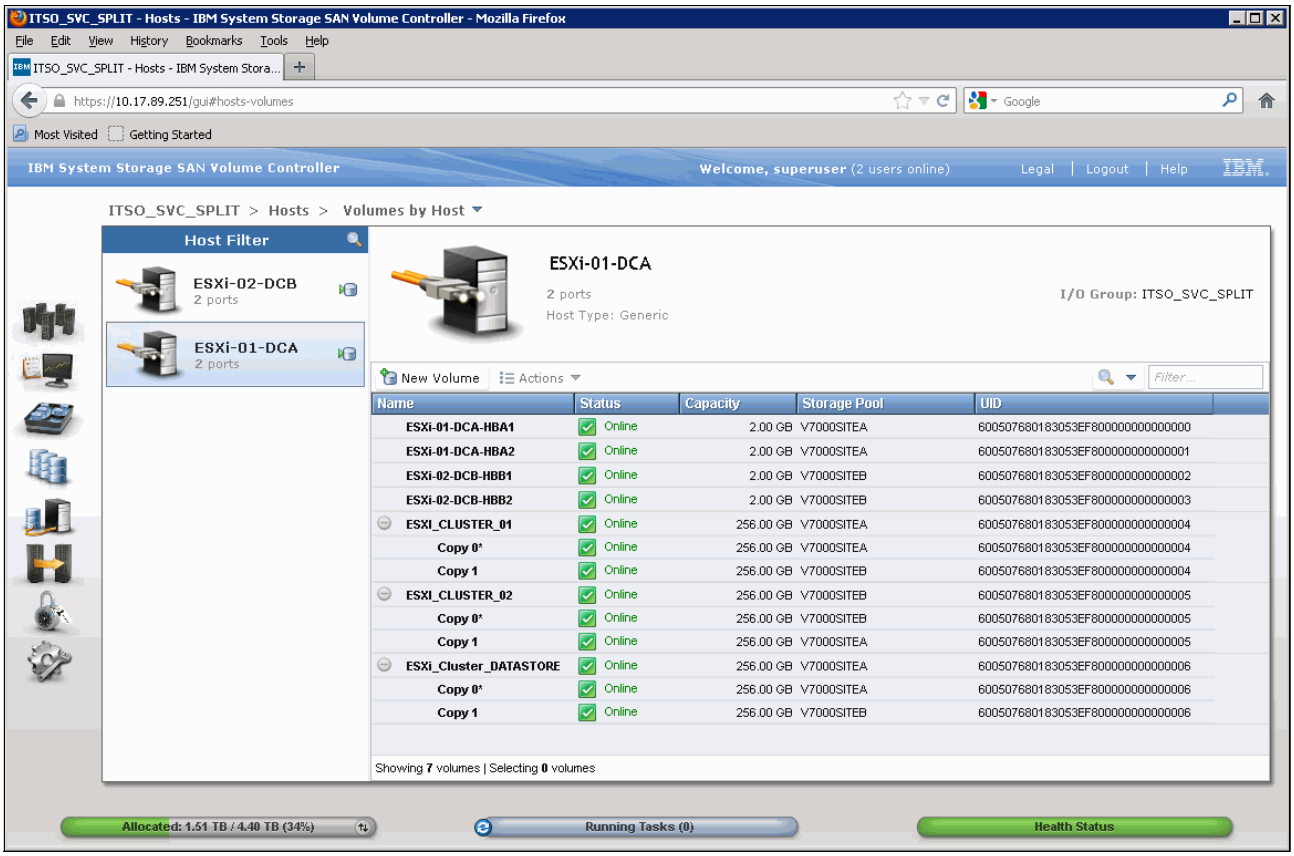


Figure 4-38 Volumes that are assigned to Host ESXI-01-DCA

Figure 4-39 shows the I/O from a local perspective.

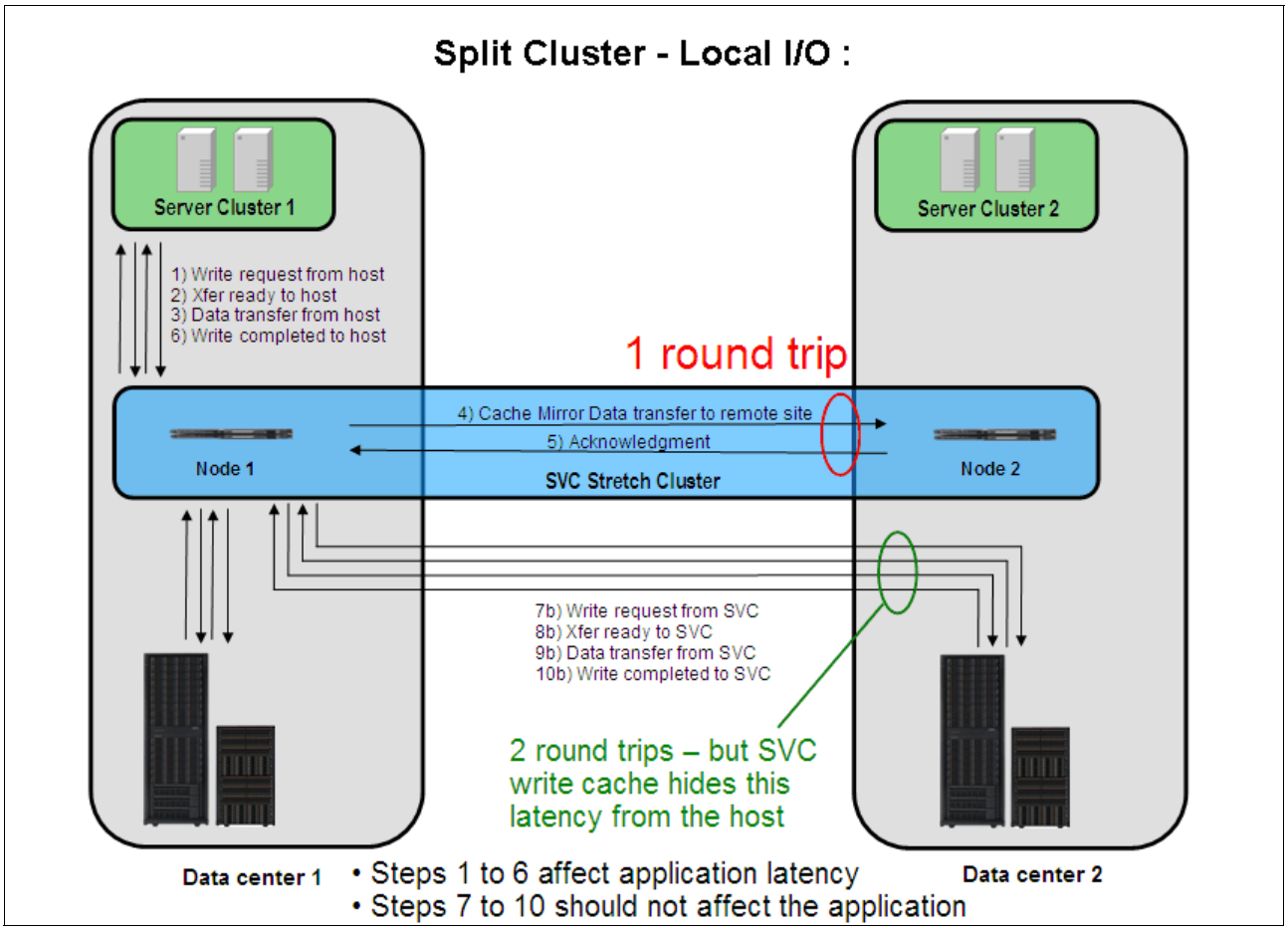


Figure 4-39 Local I/O diagram

Figure 4-40 shows the I/O from a remote perspective.

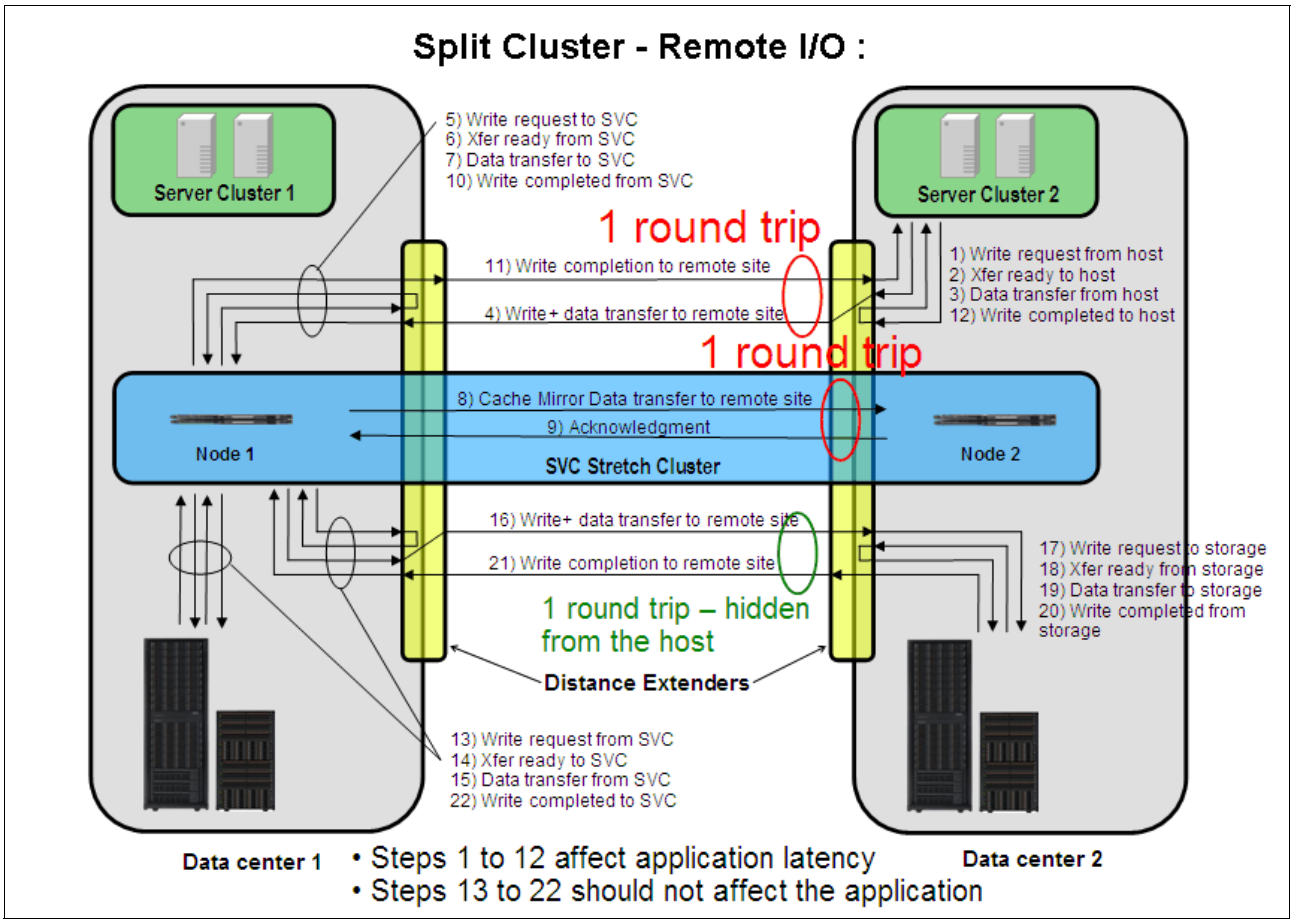


Figure 4-40 Remote I/O diagram

4.14 ESXi: VMware

To create a VMware virtual Distributed Switch (vDS), right-click the cluster and select **New vSphere Distributed Switch** as shown in Figure 4-41.

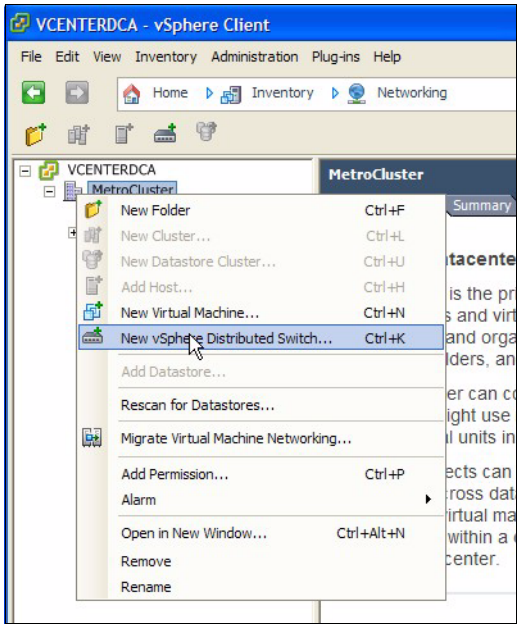


Figure 4-41 Creating a vSphere Distributed Switch

In the next window, select **vSphere Distributed Switch Version: 5.0.0**. Then, select the maximum number of physical ports per host for this vDS. The example keeps the default at four even though the lab configuration can have a maximum of two 10 GbE ports for uplinks. Also, provide a name for the vDS of **dvSwitch-DCA_DCB** as shown in Figure 4-42.

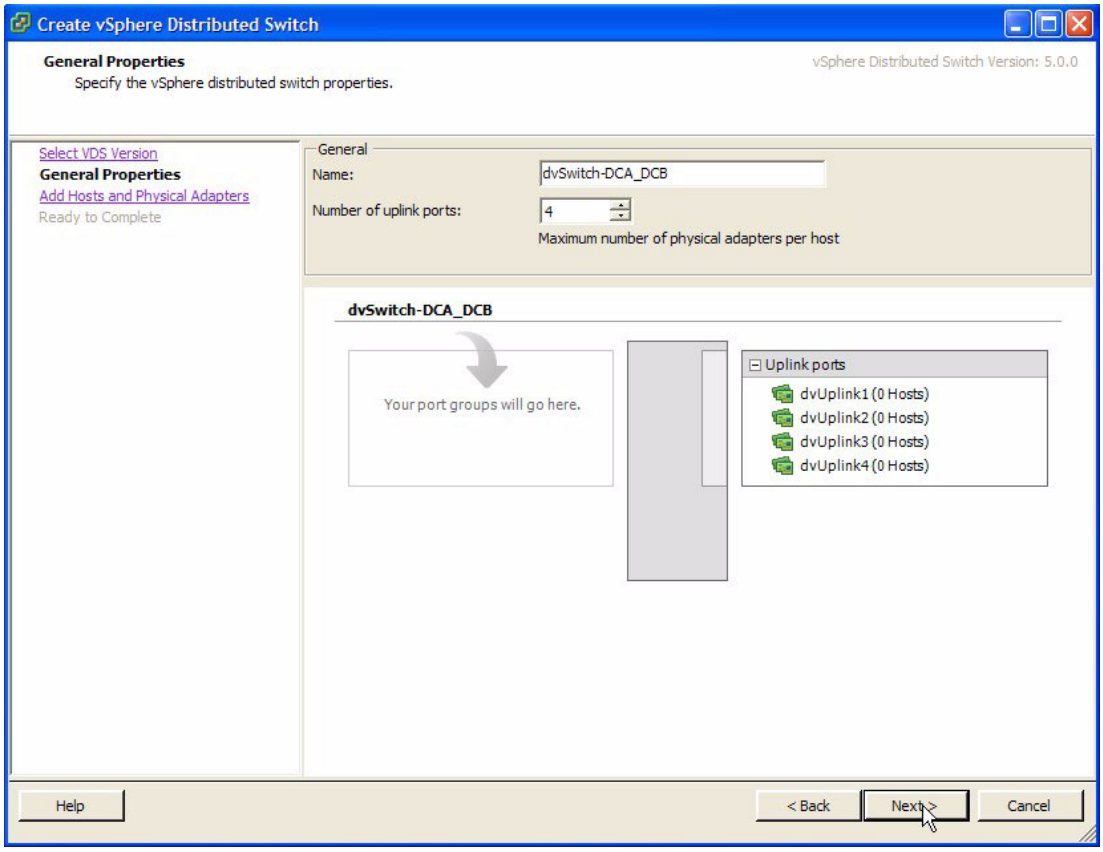


Figure 4-42 Naming the vDS and choosing the number of uplink ports per host

Add ports from each host. In the example, the vDS is already created so the two 10 GbE ports (vmnic2 and vmnic3) on each host are not available as shown in Figure 4-43.

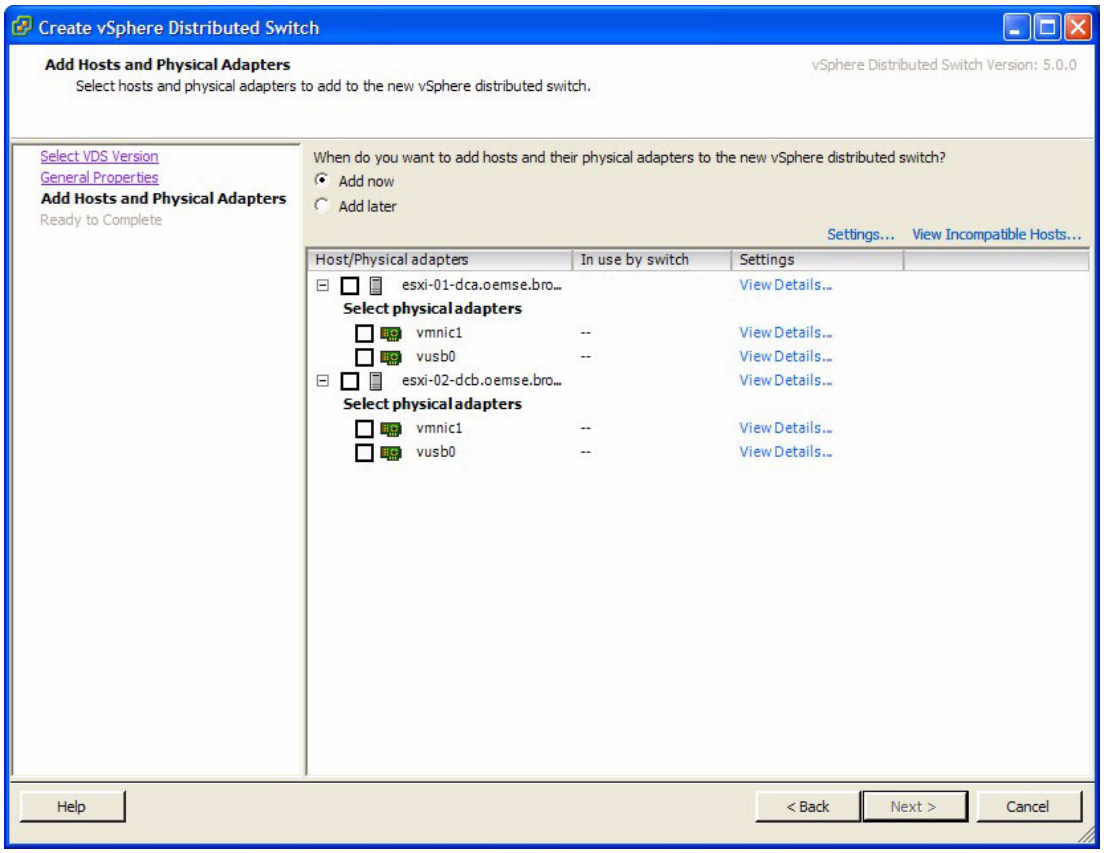


Figure 4-43 Selecting the physical ports on each host to add to the vDS

After the vDS is created and assigned physical uplinks on each host, create Port Groups. In the example, a few Port Groups have already been created. Create a fourth Port Group for Fault Tolerant Traffic by clicking **New Port Group** as seen in Figure 4-44.

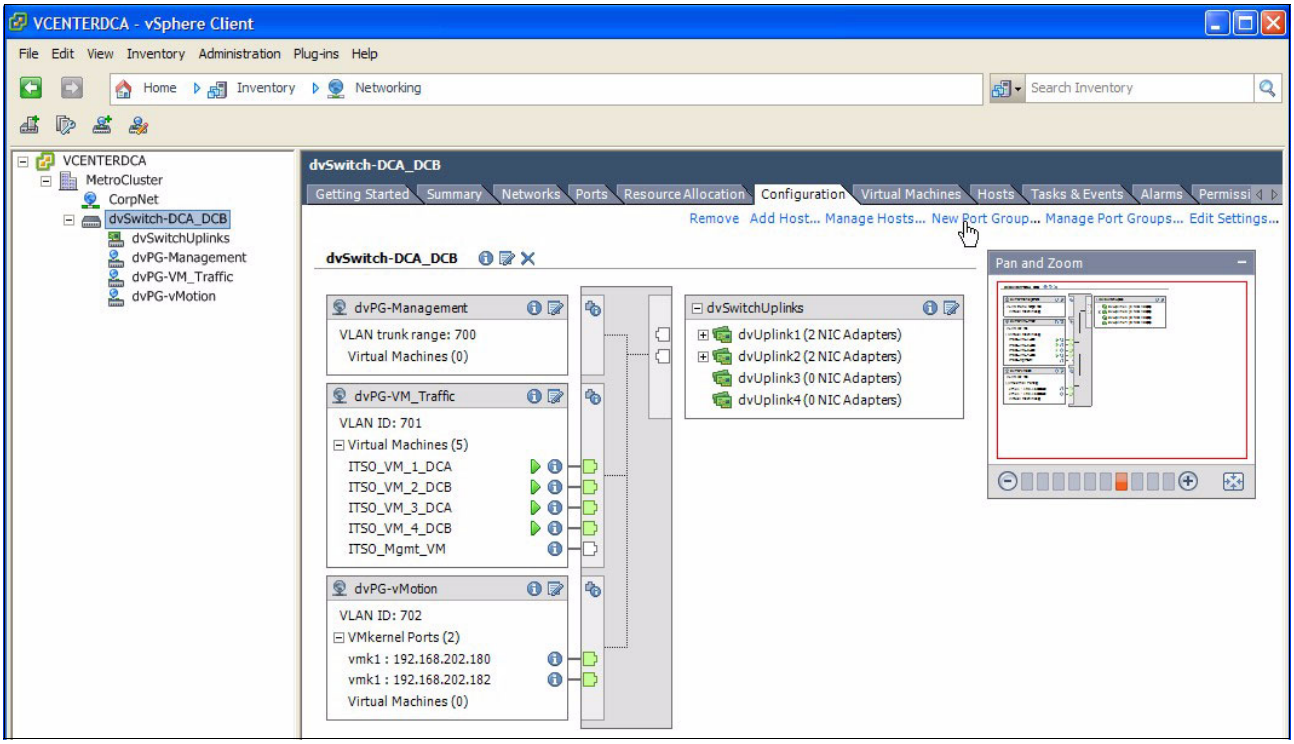


Figure 4-44 Creating a Port Group

Enter the Port Group name, **dvPG-Fault_Tolerant**, and assign it a VLAN ID of **703** as shown in Figure 4-45.

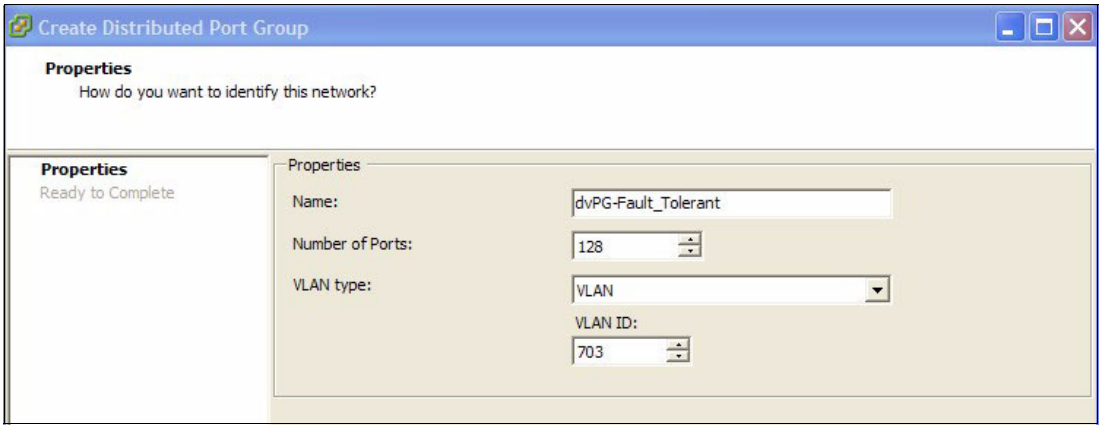


Figure 4-45 Creating a Port Group: Naming and VLAN assignment

After the Port Group is created, you can modify its attributes further by clicking **Properties** in the Port Group as seen in Figure 4-46.

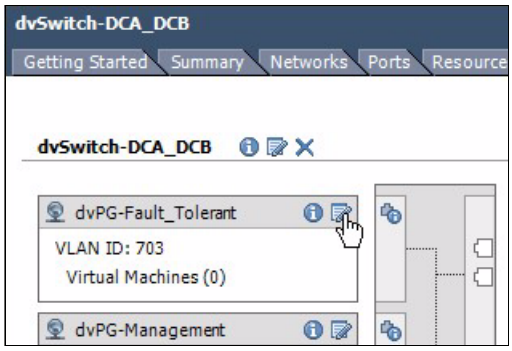


Figure 4-46 Properties for a Port Group

You can adjust the **Teaming and Failover** policies in the Port Group Settings window as seen in Figure 4-47.

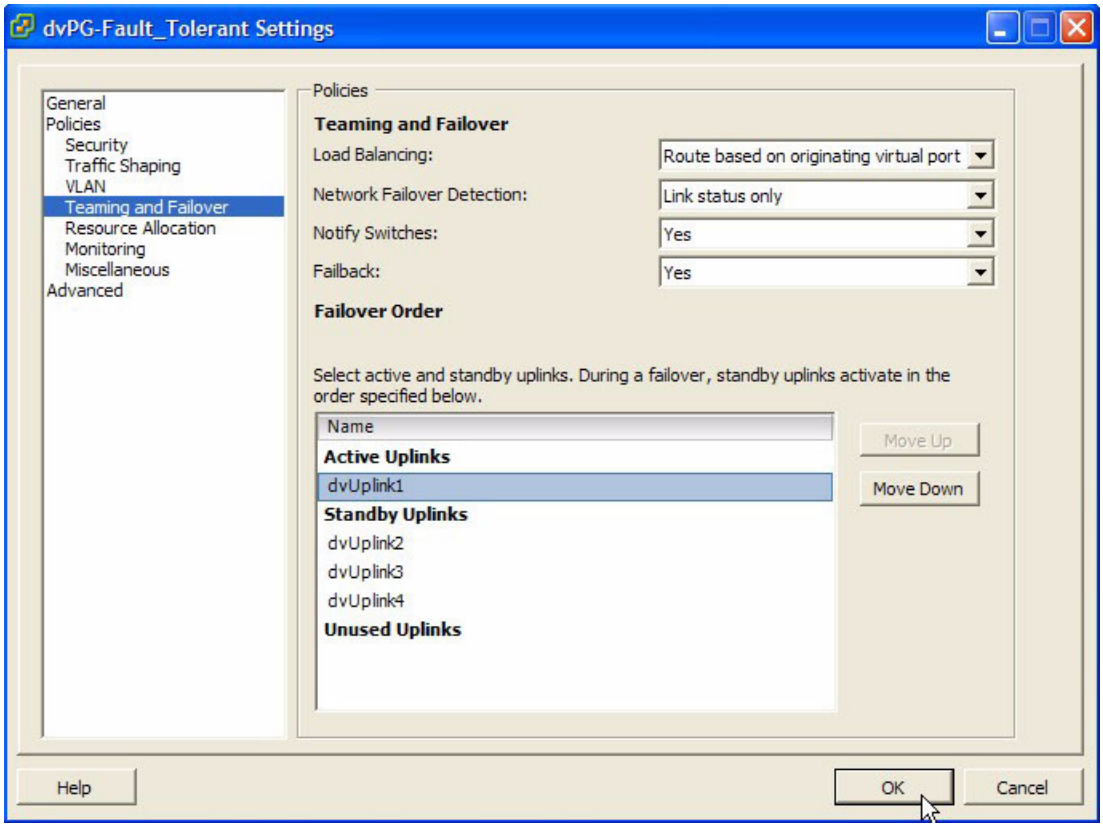


Figure 4-47 Adjusting the Teaming and Failover policies for a port group

To enable vMotion, a VMkernel interface must be created with vMotion enabled. Click **Host Level** → **Configuration** → **vSphere Distributed Switch**. To create a VMkernel interface, click **Manage Virtual Adapters** as seen in Figure 4-48.

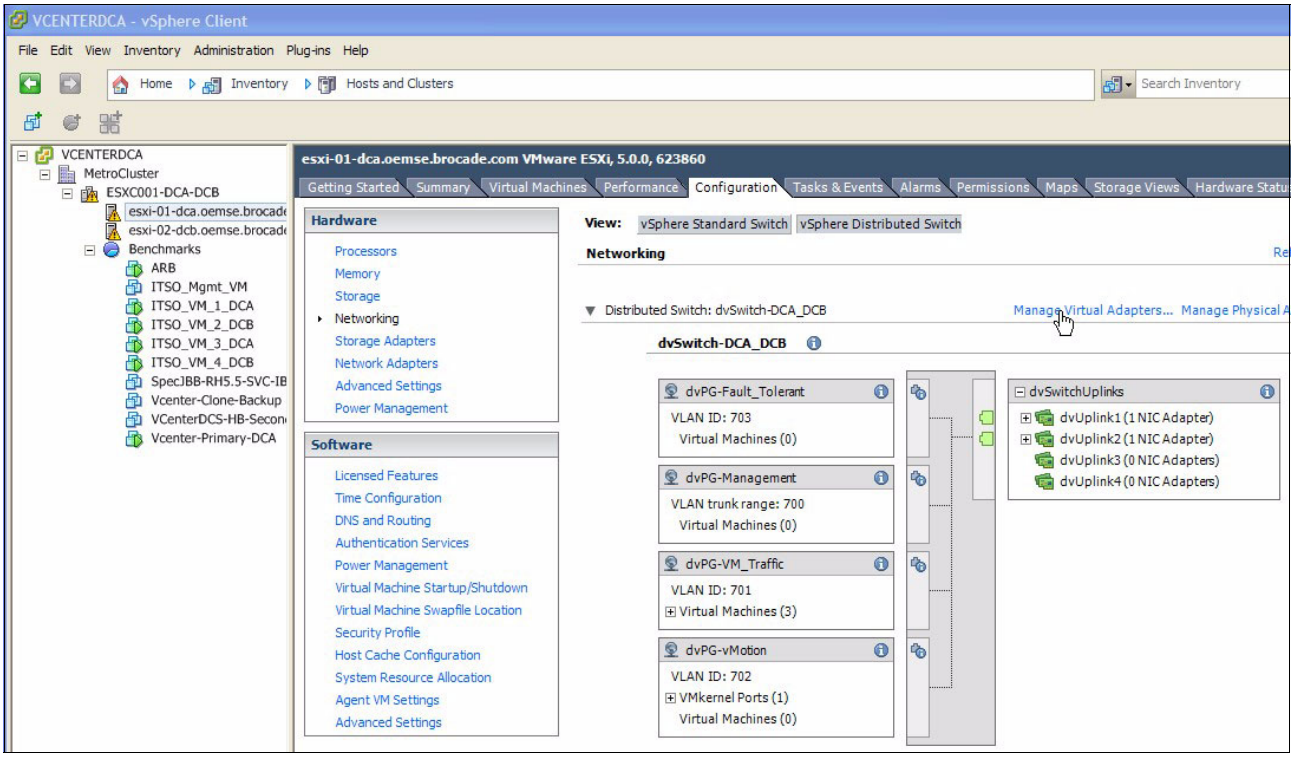


Figure 4-48 Creating a VMkernel interface attached to a vDS

Click **Add** in the **Manage Virtual Adapters** window to create a new virtual adapter. The virtual adapter type should be VMkernel. At the Connection Settings prompt, select a Port Group for this VMkernel interface and then make sure to select **Use this virtual adapter for vMotion** as seen in Figure 4-49.

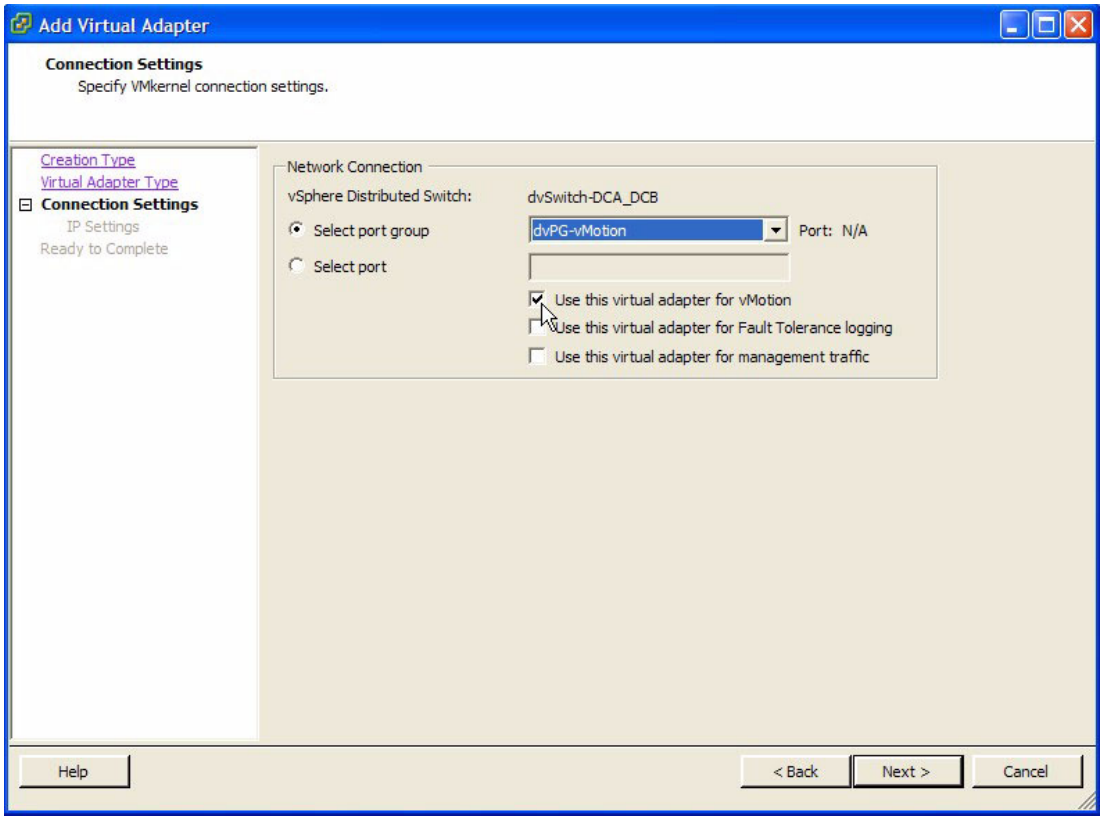


Figure 4-49 Configuring a VMkernel interface for vMotion

Finally, enter an IP address in the same subnet as your vMotion network to finish creating the interface as seen in Figure 4-50.

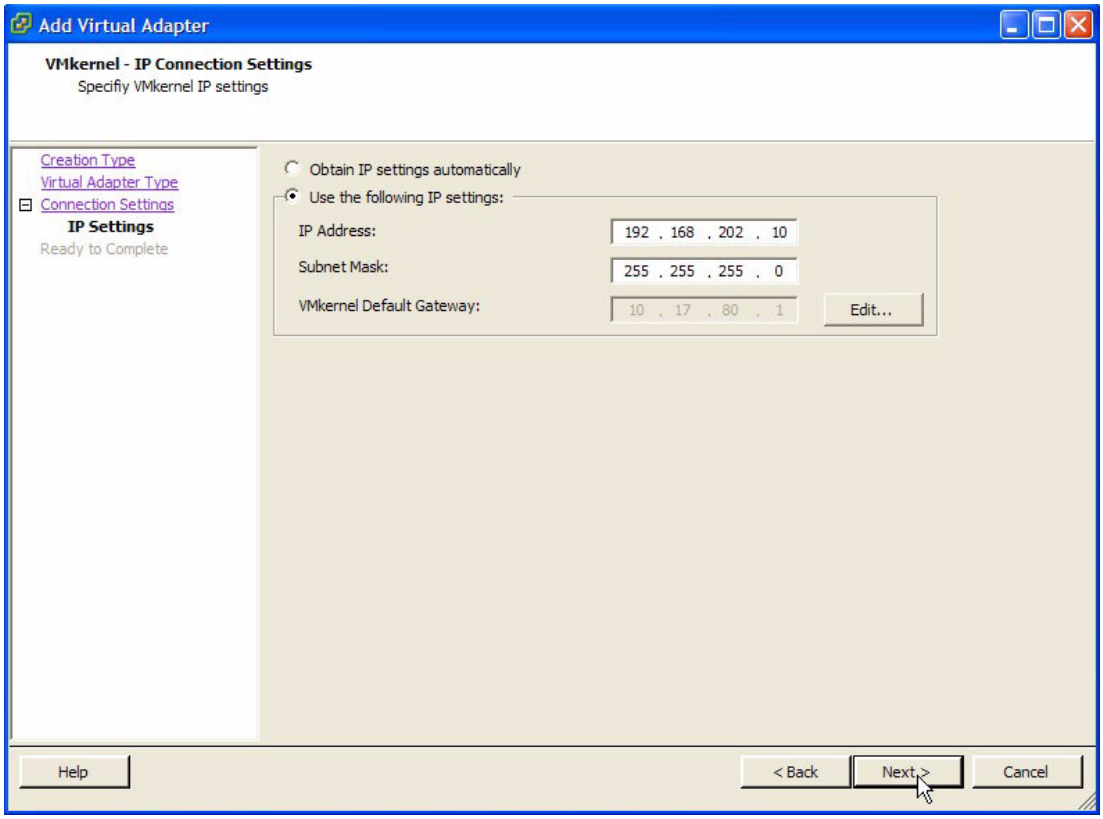


Figure 4-50 Assigning an IP address to the vMotion VMkernel interface

5

VMware environment

This chapter addresses the steps that are needed to create a VMware environment.

This chapter includes the following sections:

- VMware vMSC certification program
- VMware configuration checklist
- VMware vCenter setup
- ESX host installations
- VMware Distributed Resource Scheduler (DRS)
- Naming conventions
- VMware High Availability (HA)
- VMware vStorage API for Array Integration (VAAI)
- vCenter Heartbeat setup
- Overall design comments
- Scripting examples

5.1 VMware vMSC certification program

VMware Metro Storage Cluster (vMSC) is a VMware configuration that is defined within the VMware Hardware Compatibility List. Vendors providing support for vMSC configurations must be certified by using the VMware vMSC certification process.

IBM has successfully completed the vMSC for the SAN Volume Controller Stretched Cluster configuration. The VMware KB article that describes support for the SAN Volume Controller Stretched Cluster configuration can be found at:

http://ibm.biz/Bdxr2t

5.2 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 outside the scope of this book. Links are provided to assist on various topics.

Tip: VMware Communities are a good source of information:

http://communities.vmware.com/community/vmtn/server/vcenter

- Create naming conventions (for more information, see 5.6, “Naming conventions” on page 122):
  - Data center (DC) wise naming ESX
  - SDRS Datastores and Pools DC Affinity
  - DRS vMotion Pools DC Affinity
  - VM Naming Towards DC Affinity
- Set up ALl Hardware and create a detailed Inventory List:
  - Follow HCL List from VMware: http://ibm.biz/BdxrmT
  - Make an Inventory List
- Build ESXhosts (for more information, see 5.4, “ESX host installations” on page 112):
  - Two ESXhosts in each DC for maximum resiliency.
  - Patch and update to latest VMware Patch level.
  - Follow VMware’s Best Practice Guide: http://www.vmware.com/pdf/Perf_Best_Practices_vSphere5.0.pdf
  - Follow vSphere High Availability Deployment Best Practices: http://ibm.biz/BdxrmN
  - Set PSP and NMP (for more information, see 5.4.3, “Path Selection Policies (PSP) and Native MultiPath Drivers (NMP)” on page 113)
- Create one VM to host vCenter protected by vCenter. For more information, see 5.3, “VMware vCenter setup” on page 111):
  - Update and patch vCenter
  - Optionally, set up a Heartbeat (for more information, see 5.9, “vCenter Heartbeat setup” on page 128)

- Build a Stretched ESX Cluster between two data centers (for more information, see 5.7.3, “HA advanced settings” on page 125):
  - Optionally, implement IO control on Storage
  - Optionally, implement Virtual Distributed Switches (VDSs)
- Build an SDRS Pool:
  - Make at least two pools to match DC affinity
  - Differentiate between Mirrored and Non-Mirrored LUNs if both are used.
  - Set SDRS pool to manual
- Enable DRS (for more information, see 5.5, “VMware Distributed Resource Scheduler (DRS)” on page 118):
  - Make affinity rules to ESX Host
  - Make affinity rules to VMs
  - Make VM to ESX affinity rules
  - Set DRS to partial / or automatic if rules are trusted 100%

5.3 VMware vCenter setup

Several vCenter configuration options must be implemented as part of the SAN Volume Controller Stretched Cluster configuration.

Clarification: The example implementation uses vCenter 5.0 and MS SQL Database.

5.3.1 vCenter Heartbeat

Use the vCenter Heartbeat function to make the vCenter disaster resilient. This function requires two vCenters, with each in a separate failure domain. Perform the installation in accordance with VMware best practices.

5.3.2 Metro vMotion

Use the enhanced version of vMotion, called Metro vMotion, in a SAN Volume Controller Stretched Cluster configuration. Metro vMotion raises the allowed latency value from 5 ms to 10 ms RTT (Round Trip Time). This increase is required when failure domains are separated by a distance of more than 40 km. The Enterprise Plus License is required to obtain Metro vMotion.

Figure 5-1 lists the required platforms from the vSphere compatibility vCenter 5.0 matrix.

Platform	VMware vCenter Server 5.0 U1
VMware vCenter Site Recovery Manager 5.0.1	✓
VMware vCenter Server Heartbeat 6.4 U1	✓
VMware ESXi 5.0 U1	✓

Figure 5-1 vSphere compatibility

The vCenter application view in Figure 5-2 shows the icons for the vCenter Heartbeat and IBM SAN Volume Controller Storage plug-ins.

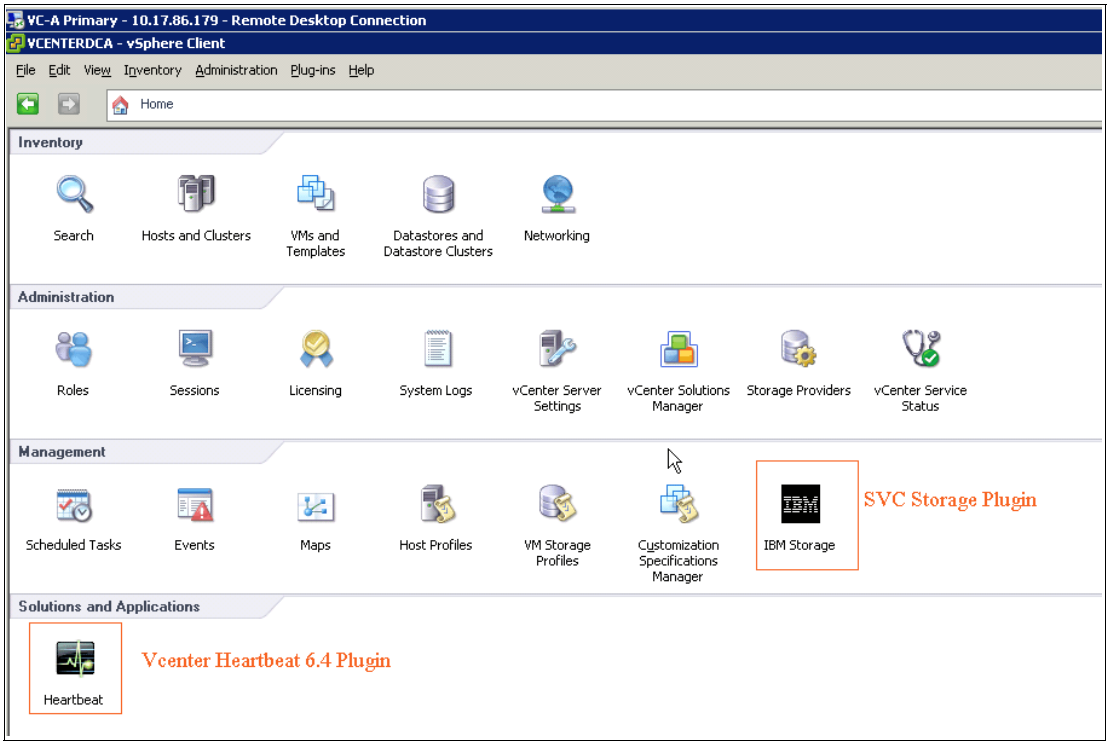


Figure 5-2 VC Application view.

5.4 ESX host installations

This chapter does not go into detail about the installation and setup of an ESX host. Instead, it focuses on design and implementation as it relates to a specific ESX Stretched Cluster configuration.

Attention: Adhere to all VMware best practice configuration guidelines for the installation of ESX hosts.

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

5.4.1 ESX host HBA requirements

The HBAs for the ESX hosts have these requirements:

- ESX hosts require a minimum of two HBAs of the same type and speed.
- The HBAs must be listed in the VMware hardware compatibility list (HCL).
- HBA firmware levels must be current and supported according to the VMware HCL and IBM SAN Volume Controller interoperability guide.

5.4.2 Initial verification

Check the latency RTT between ESX hosts to ensure that it does not exceed the maximum supported time of 10 ms. To run the latency test, use this command:

```
Vmkping <Ip of remote ESXhost, vMotion Network).
```

Verify that the ping times returned are consistent and repeatable.

Guideline: Keep a record of the ping times for future reference. This record will assist with troubleshooting if required at some point 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 impacted by other traffic on the network.

Example 5-1 shows the command to verify that the adapters are online and are functional.

Example 5-1 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 (PSP) and Native MultiPath Drivers (NMP)

For optimal performance, the ESX pathing must be configured such that active paths are accessing the SAN Volume Controller nodes that are local (in the same failure domain) to the ESX server.

To do so, obtain the node UUIDs from both the local and remote nodes. After the node UUIDs are identified, the preferred path must be configured for each of the ESX hosts.

To ensure that ESX paths are configured to the local nodes, the pathing policy must be set to VMW_PSP_FIXED. In addition, the preferred path must be set manually to enable Failback to Preferred. Even when the preferred path is set by default, it will *not* go back to the preferred path unless Failback to Preferred is enabled.

For the SAN Volume Controller Stretched Cluster configuration, the path selection policy must be set to VMW_PSP_FIXED and the preferred path be set to the local node. This configuration avoids extra I/O latency that is introduced from traversing the long-distance links.

The default path selection policy is VMW_PSP_FIXED. However, check to make sure that it has not been changed.

Guideline: Create and maintain a table that contains node UUIDs that identify local and remote nodes. Reference this table as needed to ensure that paths remain in the optimal configuration.

The script provided in **5.11.2, "PowerShell script to extract data from entire environment and verify active and preferred paths" on page 131** can be used to obtain this information from the SAN Volume Controller cluster.

Verifying path selection policy

The current path selection policy for each LUN can be verified by using the command that is shown in Example 5-2.

Example 5-2 Verifying the path selection policy

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

For more information about how to obtain LUN pathing information from the ESX hosts, see:

http://ibm.biz/BdxriP

Setting the default path selection policy

Set the default path selection policy for the entire ESX host, which requires that the ESXhost be restarted.

From the ESX Shell console, first list the available vendors and levels as shown in Example 5-3.

Example 5-3 Listing vendors and levels

```
esxcli storage nmp satp list
Name                 Default PSP       Description
-------------------  ----------------  ----------------------------------------
VMW_SATP_SVC         VMW_PSP_FIXED     Supports IBM SVC
VMW_SATP_MSA         VMW_PSP_MRU       Placeholder (plugin not loaded)
VMW_SATP_ALUA        VMW_PSP_MRU       Placeholder (plugin not loaded)
VMW_SATP_DEFAULT_AP  VMW_PSP_MRU       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_FIXED_AP  Placeholder (plugin not loaded)
VMW_SATP_SYMM        VMW_PSP_FIXED     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
```

Refer to the output in Example 5-3 on page 114 to set the default path selection policy to fixed for all new paths as shown in Example 5-4.

Example 5-4 Setting the default to fixed

```
esxcli storage nmp satp set --default-psp VMW_PSP_FIXED --satp VMW_SATP_SVC
Default PSP for VMW_SATP_SVC is now VMW_PSP_FIXED
```

The fixed path with Array Preference (VMW_PSP_FIXED_AP) policy was introduced in ESX/ESXi 4.1. It works for both Active/Active and Active/Passive storage arrays that support ALUA. This policy queries the storage array for the preferred path based on the array's preference. If no preferred path is specified by the user, the storage array selects the preferred path based on a specific criteria.

The VMW_PSP_FIXED_AP policy has been removed from ESXi 5.0. For ALUA arrays in ESXi 5.0, the PSP MRU is normally selected. However, some storage arrays must use Fixed. To check which PSP is recommended for your storage array, see the Storage/SAN section in the VMware Compatibility Guide or contact your storage vendor.

For more information, see the KB article from VMware at:

http://ibm.biz/BdxrKn

Assigning the preferred path

Set PSP to FIXED by using VMware_SATP_SVC. The path that is highlighted in blue in Figure 5-3 on page 116 shows it is the preferred path. Its target WWPN ending with b1:3f is on a node that is local to the ESX host.

Click **ESXHost** → **Storage** → **Datastore** as shown in Figure 5-3 to change the preferred path from the VC-GUI.

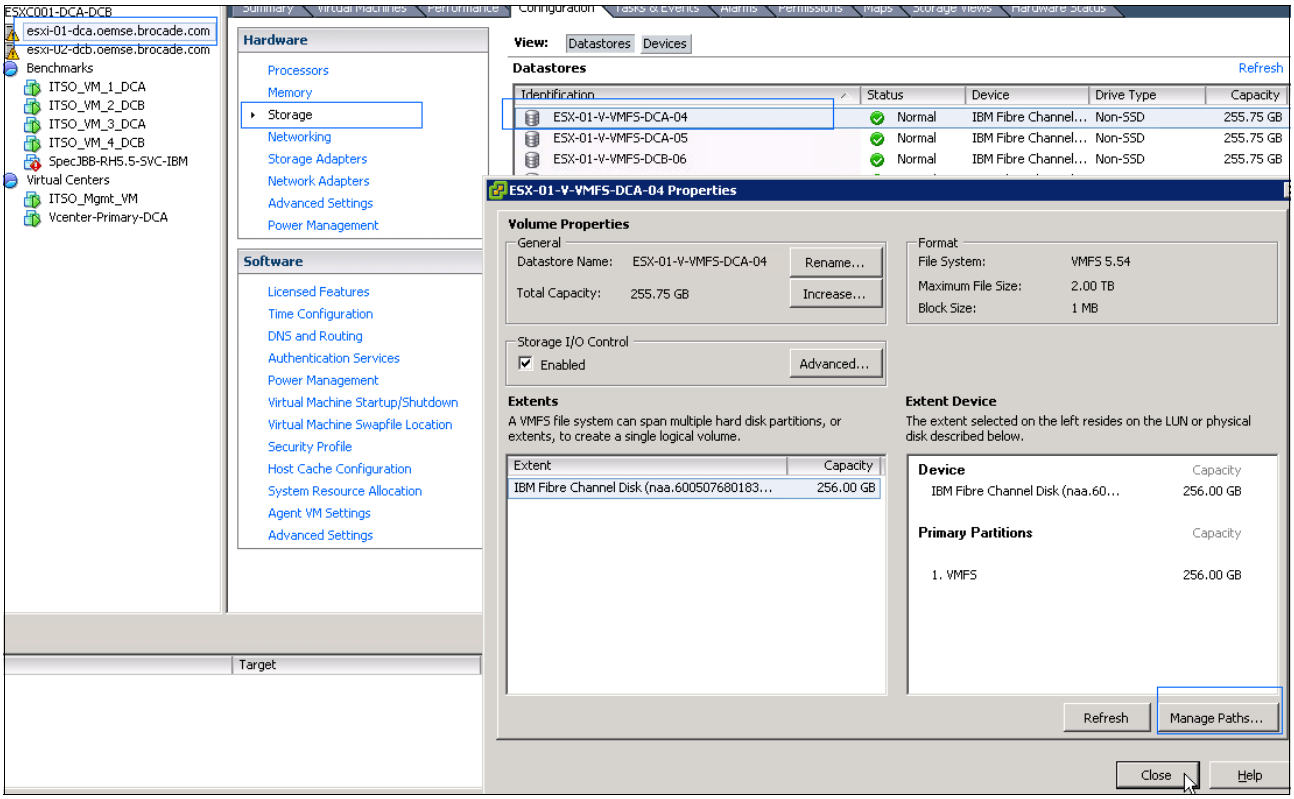


Figure 5-3 Verifying the Preferred Path from VC-GUI

Click **Manage Paths** to open the Manage Path view as shown in Figure 5-4.

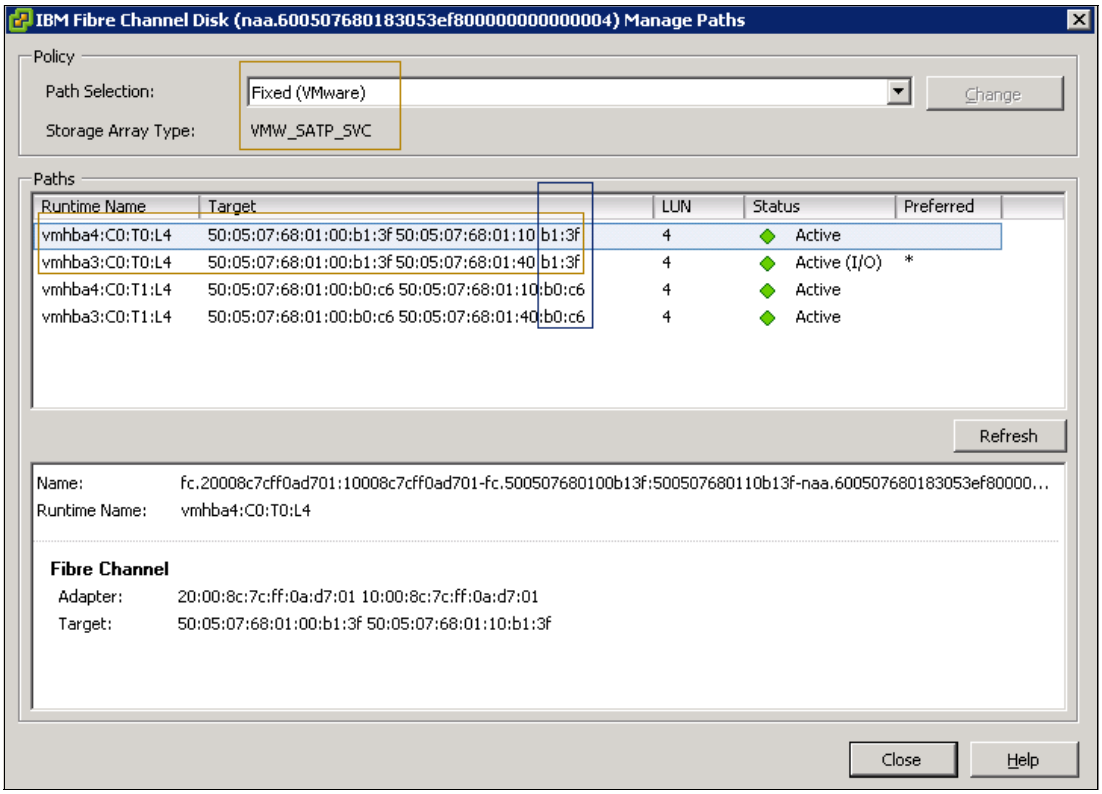


Figure 5-4 View of the path for a specific datastore

Ensure that the PSP is set to Fixed (VmWare)

Verify that the Target information matches the expected ID of the Preferred Local SAN Volume Controller Node according to the inventory plan. In this case b1:3f is in Data Center A (DCA) and b0:c6 is in DCB.

Guideline: Equally balance LUNs across HBAs on the local preferred path. For example, assign odd LUN IDs to HBA3 and even LUN IDs to HBA4 as a standardized rule.

Path failover behavior for a dead path

If an active path fails, the ESX path selection policy randomly selects an alternative path. Because the path selection is random, the path selected might be one configured to a remote node, which can degrade performance. If this occurs, you can manually switch the path back to an available path configured to a local node. If another preferred path is not manually selected, the original preferred path automatically becomes the active path after the path failure is repaired and the path returns to an operational state.

In large environments, managing the ESX path configuration and ensuring it is optimally aligned with the SAN Volume Controller nodes can require significant effort. You can make this process easier by using scripts to extract the current configuration data from the environment and present it in a format where it can be quickly reviewed and acted upon if necessary.

ESX path optimization in an SAN Volume Controller Stretched Cluster configuration

The SAN Volume Controller architecture requires that all read operations are run from the primary copy of the volume mirror. This configuration makes it necessary to change the primary copy to the alternate mirror after the VMs are moved to the remote site. This process must be done to keep the primary mirror copy local to the ESX server that is accessing the mirrored volume. This configuration avoids the extra latency that is incurred by traversing the long-distance links.

Under normal operations, it does not make sense to change the primary copy after running a vMotion on one or a subset of VMs. The change affects all VMs that share a volume, including ones that are not being moved. Exceptions are cases where a VM has a dedicated volume assigned to it, and when a moved VM has a critical application running on it and performance must be optimized for that application. In general, the primary copy is changed only when most VMs are moved to the remote site, or when all VMs are moved to the remote site such as in the case of a disaster or facility maintenance.

5.5 VMware Distributed Resource Scheduler (DRS)

Before you use DRS, you must create an ESX cluster, and enabled DRS in the menus.

Figure 5-5 shows the Main DRS/HA view where DRS and HA are enabled.

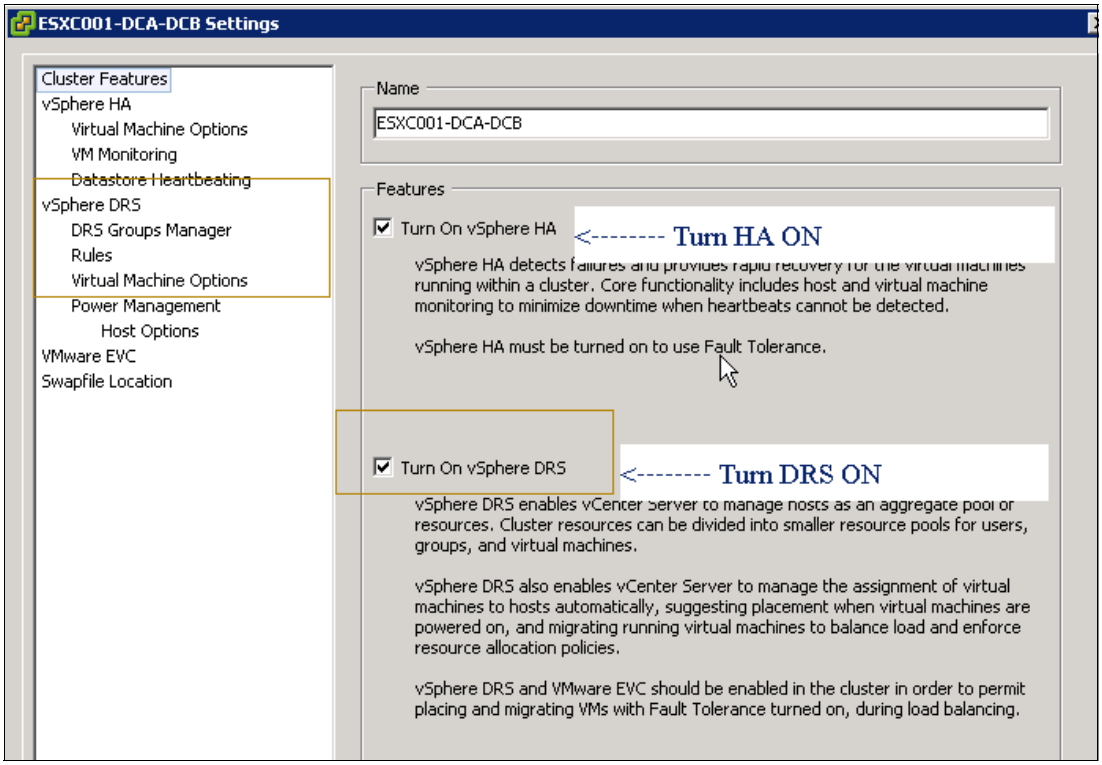


Figure 5-5 DRS enabled

DRS rules, along with accurate and meaningful naming standards, are the most important operational considerations when you are managing a VMware Metro Storage Cluster.

DRS mode can be set to automatic under normal conditions if the appropriate rules (Table 5-1) are always in place.

Table 5-1 DRS rules matrix

DRS-Rules	ESX-DRS-Host	Description
VM-Should-Run-In-DCA	ESX-Host-IN-DCA	VMs in Datacenter A (DCA), that potentially can be vMotion to DCB
VM-Should-Run-IN-DCB	ESX-Host-In-DCB	VMs in Datacenter B (DCB), that potentially can be vMotion to DCA
VM-Must_Run-DCA	ESX-Host-In-DCA	No vMotion, stick to DCA
VM-Must-Run-DCB	ESX-Host-In-DCB	No vMotion, stick to DCB

Figure 5-6 shows the DRS Rules View after you implement the rules in Table 5-1.

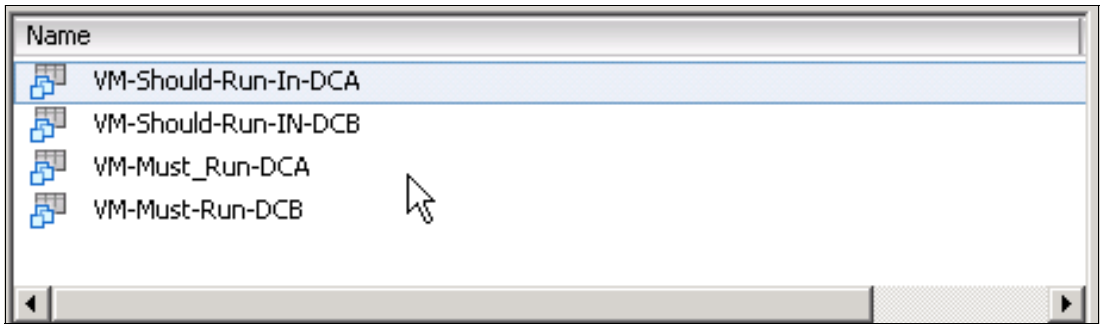


Figure 5-6 DRS Rules view

The Should-Run rules apply to VMs that can be moved to the alternate site to manage pre-disaster situations. Figure 5-7 shows the Should-Run rules on DCA.

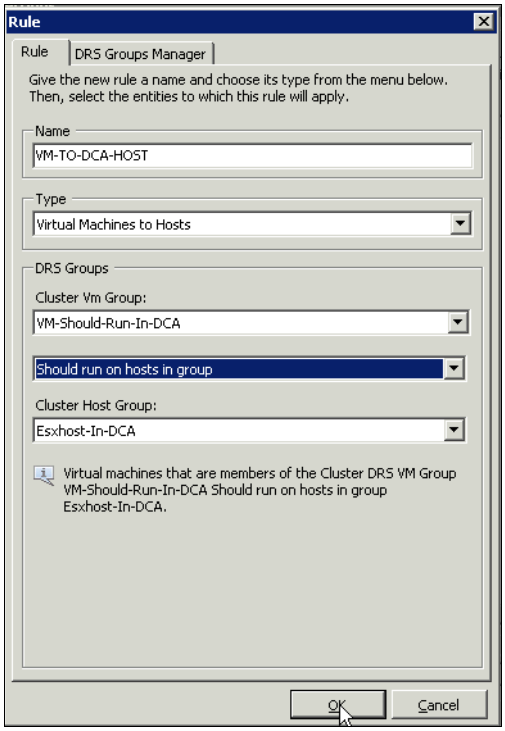


Figure 5-7 Should-Run rules on DCA

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 Heartbeat Primary/Secondary.

Figure 5-8 shows the Should-Run-Rules-DCB.

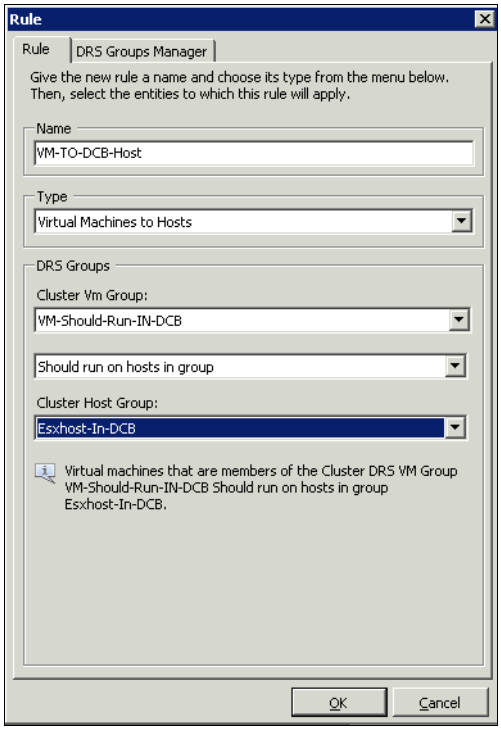


Figure 5-8 Should-Run-Rules-DCB

Figure 5-9 shows an example of DRS-VM rules that are implemented in VC.

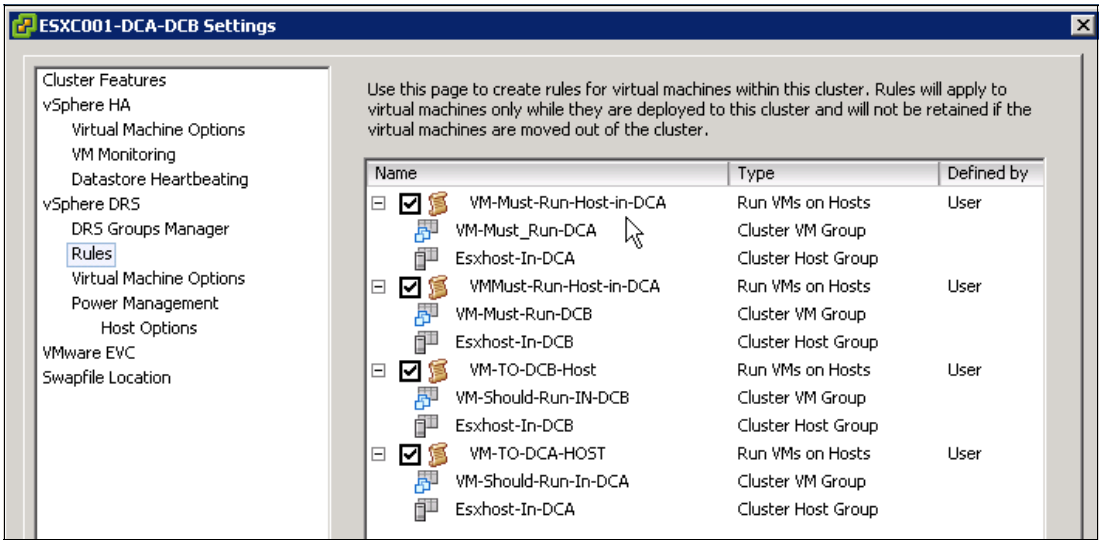


Figure 5-9 DRS-VM rules

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

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

For VMs where the running site is not important, specific rules do not need to be defined. In these cases, DRS-vMotion moves them automatically if it determines the performance can be improved by doing so.

5.6 Naming conventions

The use of a strict and well thought out naming convention is critical to the reliable operation of the environment.

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 Datastores are running at the correct site.

Example 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>

Datastores:

- ESX_<Cluster##>_<DiskType>_<MirorType>_<DC><#LUN_ID>_<OWNER>
- <Cluster> {just Unique} prefer a number
- <DiskType> [VMFS/NFS/RDM]
- <MirrorType>
  - M=Metro Mirrored Disk
  - V= Volume Mirrored Disk (what is used in vMSC)
  - G= Global Mirror Disk (Asynchronous) between two Storage Clusters
  - N= Not Mirrored
- <DC> The Preferred data center that holds the Primary Disk copy of the LUN
- <LUN_ID> 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_01_VMFS_V_DCA_01 — A Volume Mirrored LUN, Preferred from DCA
- ESXC_01_VMFS_V_DCB_02 — A Volume Mirrored LUN, Preferred from DCB
- ESXC_01_VMFS_V_DCB_03_VM-DCA_01 — With a Dedicated VM as Owner

Figure 5-10 shows a datastore naming standard example.

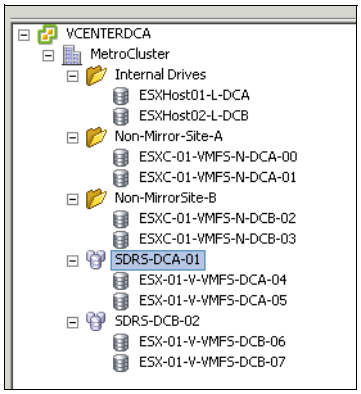


Figure 5-10 Datastore naming example

SDRS-Pools:

- SDRS-<Datacenter><####>
- IE: SDRS-DCA-001 (Pool of datastores in Data Center A)

5.7 VMware High Availability (HA)

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

For information about setup and configuration of VMware HA, see *vSphere High Availability Deployment Best Practices* at:

http://ibm.biz/BdxrmN

5.7.1 HA Admission Control

In a VMware Stretched Cluster environment, make sure that each site can absorb the workload from the alternate site in the event of 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%. 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 ESX host in the other data center.

5.7.2 HA Heartbeating

Heartbeating is a method for detecting possible downtime of an ESX host to enable recovery actions that are based on the defined policies. A new feature called Fault Domain Manager (FDM) has been implemented in vSphere 5, which is 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 decided on an isolation state itself, but now FDM and vCenter decide that.

When an ESX host is isolated from the other ESX host, which means that under normal conditions it follows the rules of an Isolation state.

One of the rules that are needed is what to do with the VMs on the host in case of an isolation state. Generally, setting the policy for the virtual machines to “Leave Powered On” in the cluster HA setup in case a host enters the isolated state.

Important: Review the following list of items because of the changes that have been made to HA.

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

- Networking Heartbeating

  Primary Control: Checks the basic network for isolation of an ESX host. Generally, have at least two interfaces with isolation addresses defined. For more information, see 5.7.3, “HA advanced settings” on page 125.

- HA Datastore Heartbeating

  Secondary Control: Generally for VMware, allow vCenter to find the best possible datastores for control. However, you can manually set the datastore that you think is the best, and where the ESX host has the most connections to.

Figure 5-11 shows the datastore that is used for heartbeat isolation detection when your preferred datastores are selected.

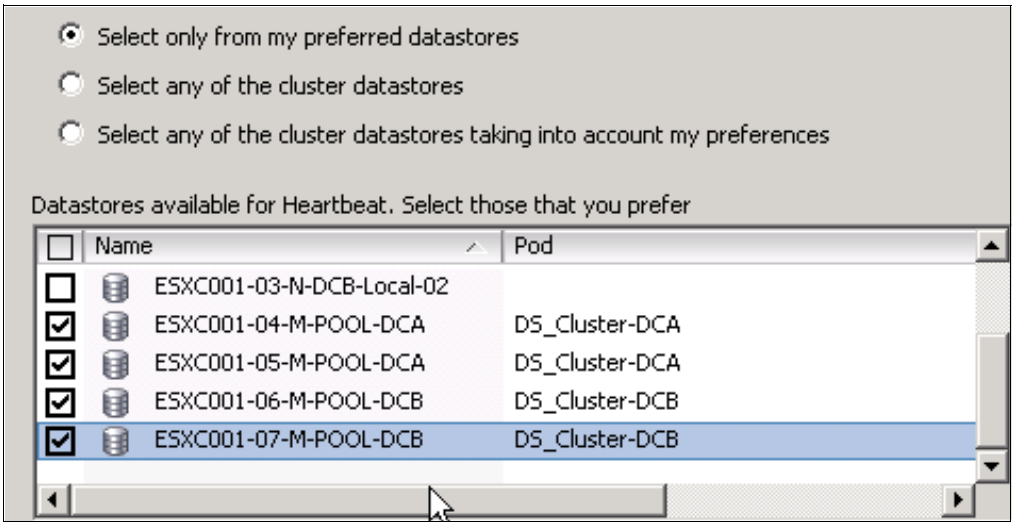


Figure 5-11 Datastore selection for Heartbeat

If you want to allow VMware to decide, it will look as shown in Figure 5-12.

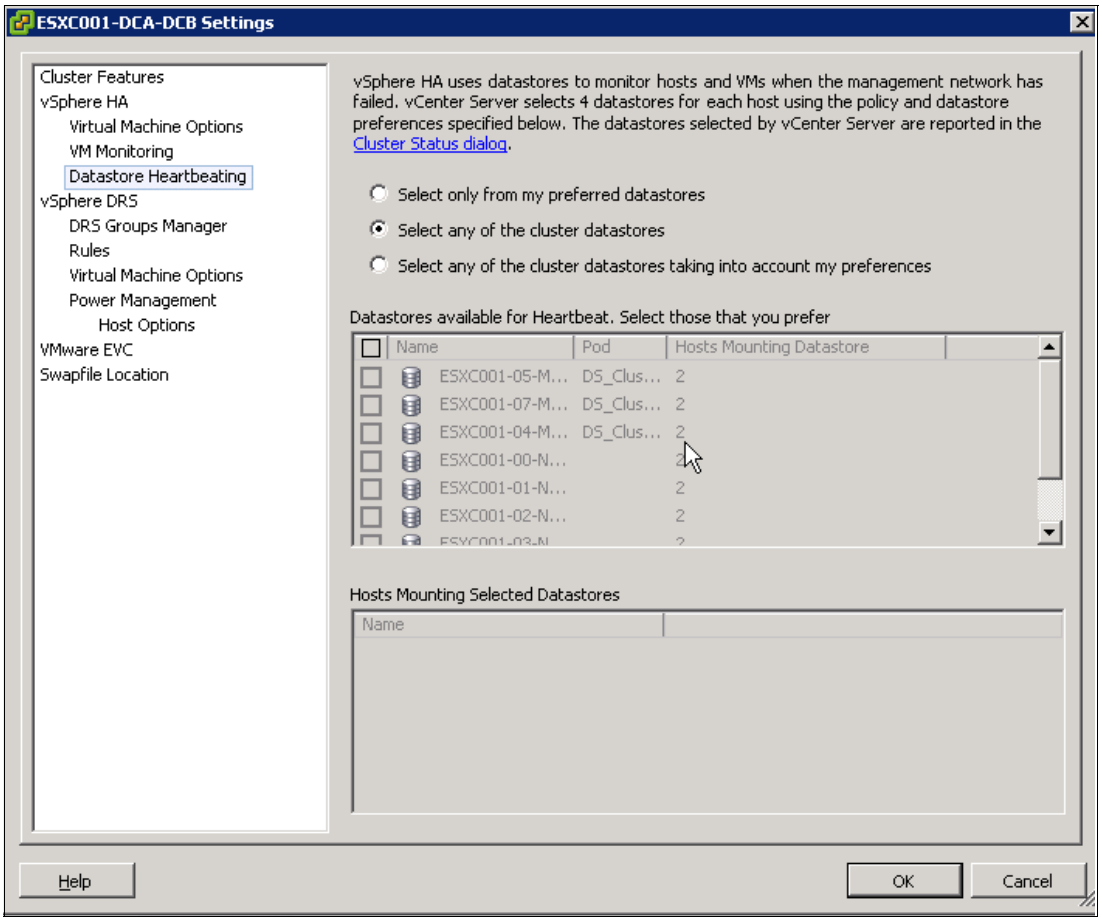


Figure 5-12 Datastore Heartbeating

5.7.3 HA advanced settings

HA has other settings that are important for you to consider. Your environment might require others that you must consider, but these are the ones that were applicable to the example environment.

Table 5-2 provides the list of 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 implementation.

Table 5-2 HA advanced settings

HA string	HA value	Short explanation
das.heartbeatDsPerHost	4	Number of Heartbeat datastores that are used in the cluster.
das.maskCleanShutdownEnabled	true	PDL Enabled Kill VM state.
das.Isolationadress0	10.0.0.1	Default GW on mng. Kernel 1.

HA string	HA value	Short explanation
das.isolationadress2	10.0.1.1	Def-.Gateway for 2'cond Mng interface used for Heartbeat. Can be vMotion, but then you need to add the next one also.
das.allowvMotion	true	Allows vMotion to be a part of the isolation check.

5.7.4 All Paths Down (APD) detection

vSphere 5.0 Update 1 introduced a new mechanism that uses SCSI Sense Codes to determine whether a VM is on a datastore that is in an APD state. This state triggers a VM HA failover because the system cannot restart the VM on the same ESX host.

This is part of the process to secure an ESX host isolation handling of the virtual machines. Generally, do not disable this mechanism.

5.7.5 Permanent Device Loss (PDL)

PDL is a state where the VM is on an ESX host memory, but the datastore is lost. When this state occurs, any new I/O sent by the VM is killed (this is a UNIX based kill) so another ESX host can restart it. Generally, do not disable this mechanism.

Consideration: VMs not running any I/O might not be killed correctly. If this problem occurs, the VM can be killed manually from the console by using vmfktools.

In vSphere 5.0, you must set PDL manually. In the folder `/etc/vmware #`, add the following line to the file: `settings`. Use the VI Editor to add this line, and note that this command is case-sensitive. Do not copy and paste onto the command line:

```
disk.terminateVMOnPDLDefault True
```

Restart the ESX host after the file has been created.

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

For more information about PDL/ADL states, see the VMware document at:

http://ibm.biz/Bdx4k7

5.8 VMware vStorage API for Array Integration (VAAI)

VMware VAAI is supported if listed on the HCL list for the SAN Volume Controller. SAN Volume Controller version 6.4.0.2 is supported, but version 6.4.0 is not.

With VMware vSphere 5.x, you do not need to install a plug-in to support VAAI if the underlying storage controller supports VAAI.

There are several commands that can be used to check VAAI status. Example 5-5 shows using the command `esxcli storage core device vaai status get`.

Example 5-5 Checking the VAAI status

```
esxcli storage core device vaai status get
VAAI Plugin Name:
ATS Status: supported
Clone Status: supported
Zero Status: supported
Delete Status: unsupported
```

To determine whether VAAI is enabled, issue the three commands and check if the default interval value is set to 1, which means it is enabled as shown in Example 5-6.

Example 5-6 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 vCenter Heartbeat setup

VMware vCenter Server Heartbeat allows administrators to monitor and protect multiple instances of VMware vCenter Server from one location. Heartbeat requires no hardware configuration dependencies and automatically detects standard VMware vCenter Server configuration upon installation, providing instant monitoring and protection. These features allow you to protect vCenter in case of failures on the primary data center.

For more information about in installation, see the vCenter Heartbeat Installation Guide from VMware at:

http://www.vmware.com/products/vcenter-server-heartbeat/overview.html

To install and set up the vCenter Heartbeat, generally use the Virtual to Virtual (V2V) method.

5.9.1 Heartbeat Virtual to Virtual (V2V)

V2V is the supported architecture if vCenter Server is already installed on the production (Primary) server that runs on a virtual machine. The Secondary virtual machine must meet the minimum requirements. It is also the IBM preferred solution.

The specifications of the Secondary virtual machine must match that of the Primary virtual machine as follows:

- Similar processor (including resource management settings)
- Memory configuration (including resource management settings)
- Appropriate resource pool priorities
- Each virtual machine that is used in the V2V pair must be on a separate ESX host to guard against failure at the host level.

The vCenter application view in Figure 5-13 shows the icon for the vCenter Heartbeat plug-in.

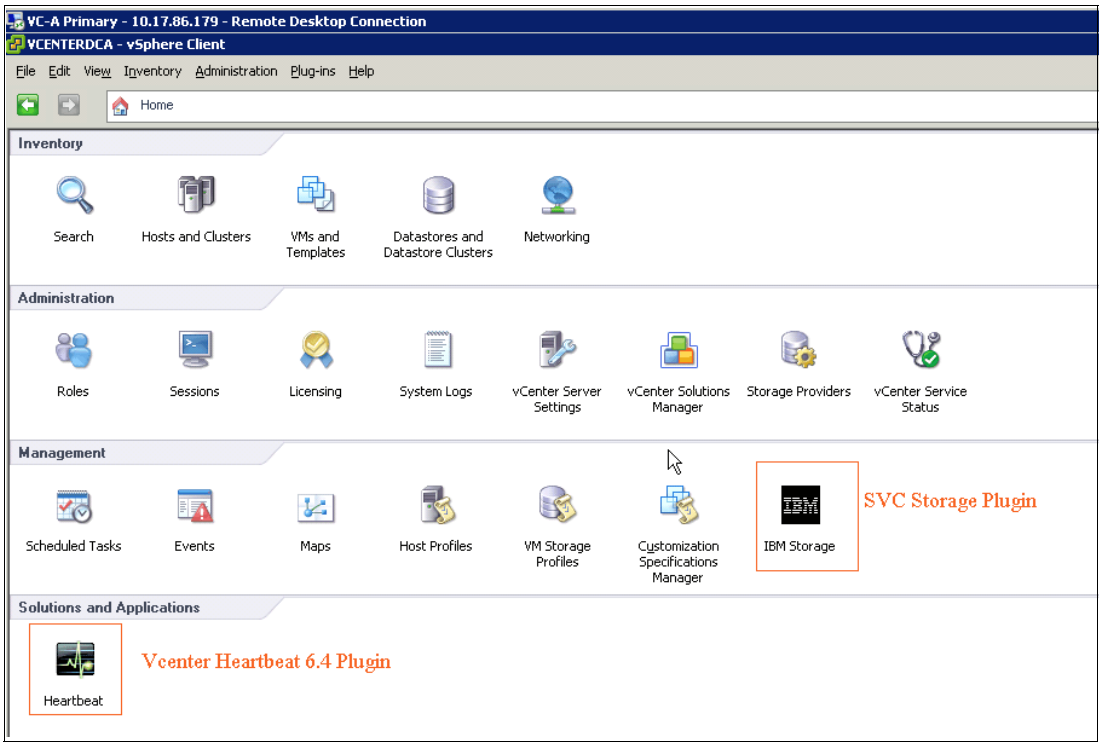


Figure 5-13 VC application view

5.9.2 Why vCenter as Virtual

Generally, secure the VMs from hardware failure by running vCenter as a virtual machine on the primary data center. Have the secondary VM, which is a part of the heartbeat setup, also running as virtual but on the secondary data center. Have the secondary VM be a clone of the primary VM.

5.10 Overall design comments

When you create a SAN Volume Controller Stretched Cluster together with a VMware Stretched Cluster, you combine options to prevent a disaster and allow access to the datastores across the clusters. With this configuration, you have access to failover or vMotion instantly, without having to rezone your SAN disk, or having to switch to the mirrored copy of the disk.

A VMware Stretched Cluster is managed best by implementing rules that, under normal operation, bind the virtual machines to each data center that is part of an SAN Volume Controller Stretched Cluster. Use affinity rules to secure it, and vMotion to prevent disasters.

Generally, use vMSC in enterprise solutions, where distance is the key factor, and the actual calculation is measured in ms. From the VMware perspective, the solution is accepted up to 10 ms RTT. However, in this solution Stretched SVC Cluster 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 then 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.

5.11 Scripting examples

The following scripting examples can help you with testing VM mobility, and to check for the preferred path policy.

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

Remember: You use these scripts at your own risk. If you are in any doubt, do not run them.

5.11.1 PowerShell script to move VMs between two ESX hosts

It is important to make sure that the infrastructure is stable by actually 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 Command 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 5-7 shows the vMotion test script.

Example 5-7 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, since we start from 0)

#[Change] the name here to your vCenter:
Connect-VIServer -server "vCenterDCA"

## Perform 40 vMotions bewteen 2 ESXhosts in each datacenter
do {
#
```

```
# Get VM information and its current location
#
   $vmname = get-vmhost -VM $vm -ErrorAction SilentlyContinue

   if ( $vmname.Name -eq $dest1 ) { $mdest = $dest2 }
   else { $mdest = $dest1 }
   #
   # Migrate VM to Destination
   #
   write-output "Performing Migration # +$cnt+ To: $mdest"
   migrateVM $vm $mdest
   $cnt++
# If fast vMotion, you can lower this A bit , 300 = 5 min.
   start-sleep -s 300
   }
while ($cnt -lt $NumberOfvMotions)
```

5.11.2 PowerShell script to extract data from entire environment and verify active and preferred paths

This script extracts the preferred path and active state on datastores into comma-separated values (CSV) files to verify the actual preferred path settings on all HBAs and datastores. The script does not modify or set anything. It only writes the settings into a CSV file.

Note: Look for: **#[Change]**, which indicates where you must change the script to match the names in your environment.

Example 5-8 shows the script.

Example 5-8 Script to verify active and preferred paths

```
# Copy & Paste the entire Text into a Tesxt file
#[Change] the name here: "vCenterDCA" to the name of your Virtual Center.
Connect-VIServer -server "vCenterDCA"
foreach($ds in (Get-Datastore | where {$_.Type -eq "VMFS"} )){
    $ds.ExtensionData.Info.Vmfs.Extent | %{
        $datastores.Add($_.DiskName,$ds.Name)
    }
}
$report = @()

# Find the Corresponding NAA names, and State

for each($esx in Get-VMHost){
   $PrefArray = @()
        ### AllOne line ->
        foreach($lun in (Get-ScsiLun -VMHost $esx)){
      $PrefArray =  Get-ScsiLunPath -ScsiLun $lun  |`
      Select @{N="Host";E={$esx.Name}},Preferred,SanID,State,LunPath,`
                @{N="DS";E={$dsTab[$_.ScsiCanonicalName]}} | Where-Object
{$_.preferred -eq "true"}
        ## One liner END.
```

```
   $report += $PrefArray
     }
}
########################## rename Target CSV file ###################
## Change the name of the File: C:\Cluster-pref.csv and the Delimiter if not an
“;”
$report| Export-Csv  -Path c:\Cluster-pref.csv -Delimiter ","
```

6

SAN Volume Controller Stretched Cluster diagnostics and recovery guidelines

This chapter addresses cluster diagnostics and recovery guidelines. These features help you understand what is happening in your Stretched Cluster 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.

This chapter includes the following sections:

- Solution recovery planning
- SAN Volume Controller recovery planning
- VMware recovery planning
- SAN Volume Controller diagnosis and recovery guidelines

6.1 Solution recovery planning

In the context of the Stretched Cluster environment, solution recovery planning is more application-oriented. Therefore, any plan must be made together with the client application owner. In every customer environment, when a business continuity or disaster recovery solution is designed, incorporate a solution recovery plan into the process.

It is imperative to identify high-priority applications that are critical to the nature of the business. You should then create a plan to recover those applications, in tandem with the other elements described in this chapter.

6.2 SAN Volume Controller recovery planning

To achieve the most benefit from the SAN Volume Controller Stretched Cluster configuration, post installation planning must include several important steps. These steps ensure that your infrastructure can be recovered with either the same or a different configuration in one of the surviving sites with minimal impact for the client applications. Correct planning and configuration backup also helps minimize possible downtime.

You can categorize the recovery in the following way:

- Recover a fully redundant SAN Volume Controller configuration in the surviving site without Stretched Cluster.
- Recover a fully redundant SAN Volume Controller configuration in the surviving site with Stretched Cluster implemented in the same site or on a remote site.
- Recovery of 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 SAN Volume Controller, complete these steps:

1. Collect a detailed SAN Volume Controller configuration. To do so, run a daily based SAN Volume Controller configuration backup with the CLI commands shown in Example 6-1.

Example 6-1 Saving the SAN Volume Controller configuration

```
IBM_2145:ITSO_SVC_SPLIT:superuser>svcconfig backup
.....................................................................
CMMVC6155I SVCCONFIG processing completed successfully
IBM_2145:ITSO_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 split iogrp
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 SAN Volume Controller commands to issue

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

From the output of these commands and the .xml file, you have a complete picture of the Stretched Cluster infrastructure. Remember the SAN Volume Controller FC ports WWNNs so you can reuse them during the recovery operation that is described in 6.4.3, “SAN Volume Controller Recovery guidelines” on page 157.

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

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 ommitted 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 as shown in Example 6-5.

Example 6-5 lsnode example output command

```
IBM_2145:ITSO_SVC_SPLIT: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 how the backup of your configuration, see:

http://pic.dhe.ibm.com/infocenter/svc/ic/index.jsp?topic=%2Fcom.ibm.storage.svc.console.doc%2Fsvc_svconfigbackup_1esjw7.html

4. Create an up-to-date, high-level copy of your configuration where all elements and connections are described.
5. Create a standard labeling schema and name convention for your FC or ETH cabling, and ensure that it is fully documented.

6. Back up your SAN zoning. The zoning backup can be done by using your FC switch/director command-line interface or GUI.

   The essential zoning configuration data, domain id, zoning, alias, configuration, and zone set can be saved in a .txt file by using the output from the CLI commands. You can also use the appropriate utility to back up the entire configuration.

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

Example 6-6 Zoning 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 0bfcc0 -- -- 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
                ITSO_V7K_SITEA_N1_P2; ITSO_V7K_SITEA_N1_P1;
                ITSO_V7K_SITEA_N2_P1; ITSO_V7K_SITEA_N2_P2
 zone:  V7K_SITEB
                ITSO_V7K_SITEB_N2_P1; ITSO_V7K_SITEB_N1_P2;
                ITSO_V7K_SITEB_N1_P1; ITSO_V7K_SITEB_N1_P4
 alias: ITSO_DS5100_Cont1_P1
                20:16:00:A0:B8:47:39:B0
 alias: ITSO_DS5100_Cont1_P3
                20:36:00:A0:B8:47:39:B0
 alias: ITSO_DS5100_Cont2_P1
                20:17:00:A0:B8:47:39:B0
.
lines omitted for brevity
.
alias: ITSO_V7K_SITEA_N2_P2
                50:05:07:68:02:20:00:F0
 alias: ITSO_V7K_SITEB_N1_P1
                50:05:07:68:02:10:54:CA
 alias: ITSO_V7K_SITEB_N1_P2
                50:05:07:68:02:20:54:CA
 alias: ITSO_V7K_SITEB_N1_P4
                50:05:07:68:02:40:54:CA
 alias: ITSO_V7K_SITEB_N2_P1
                50:05:07:68:02:10:54:CB

Effective configuration:
 cfg:   ITSO_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/director because of SAN extension with

ISL. For more information, see 6.4.3, “SAN Volume Controller Recovery guidelines” on page 157.

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 backend storage subsystems configuration.

   In your Stretched Cluster implementation, you can use different vendors storage subsystems. Configure those storage subsystems according to the SAN Volume Controller guidelines to be used for Volume Mirroring.

   Back up your storage subsystem configuration so you can re-create the same environment in a critical event when you re-establish your Stretched Cluster infrastructure in a different site with new storage subsystems.

   For more information, see 6.4.3, “SAN Volume Controller Recovery guidelines” on page 157.

   a. As an example, for DS3XXX, DS4XXX, or DS5XXX storage subsystems, save in a safe place a copy of an up-to-date subsystem profile as shown in Figure 6-1.

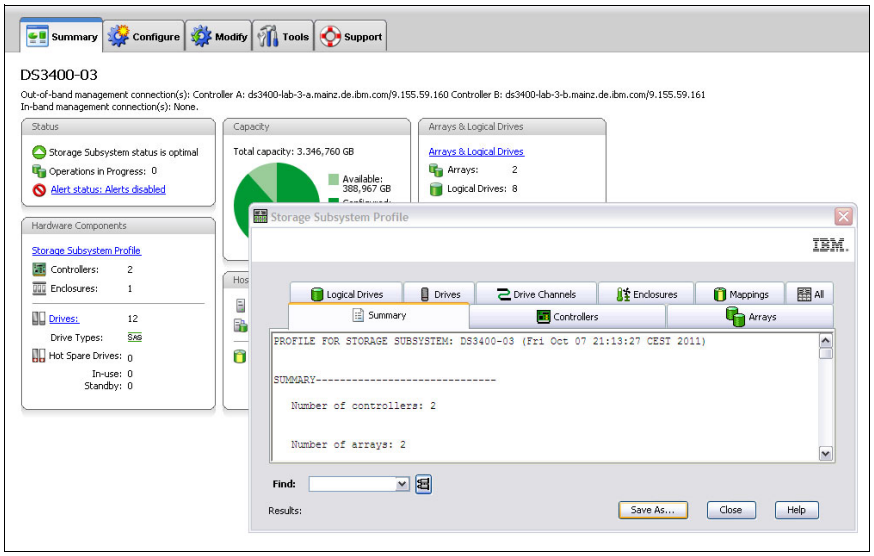


Figure 6-1 DSXYYY back-up configuration example

   b. For the IBM DS8000® storage subsystem, save the output of the SAN Volume Controller 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 subsystem, save the output of the XCLI commands in `.txt` format as shown in Example 6-8.

Example 6-8 XIV commands

```
host_list
host_list_ports
mapping_list
vol_mapping_list
pool_list
vol_list
```

d. For Storwize V7000, collect configuration files and output report as done for SAN Volume Controller in 6.2, “SAN Volume Controller recovery planning” on page 134.

e. For any other supported storage vendor, see their documentation to save details where it will be easy to find the SAN Volume Controller MDisk configuration and mapping.

6.3 VMware recovery planning

If the implementation guidelines have been followed and the inventory (CADB) is up to date, recovery ultimately depends on the situation.

The VMware environment can be documented in several ways. You can use PowerShell, or even a freeware product such as RVTools, to extract all vital data from the vCenter database to CSV files. These files can be used when planning for recovery, and to detect connections and even missing relations in the virtual infrastructure.

Most recovery planning is about having good naming conventions kept up to date, and procedures that can understand and follow this data.

Recovery is not only a question of just getting things started, but getting them started in a way that the infrastructure can start to function. Therefore, virtual machines categorization is important. At a minimum, complete these steps:

- Extract data from vCenter (by using RVtools or another tool) and save 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 122.

All of this is in addition to the normal, day to day planning such as backup. Floor space management is also important so that the physical location of servers to start, and in which order, is known.

Keep the basic IP plan printed or saved to a device that does not require a network to be available. In a worst case scenario, you might not have anything to work apart from besides paper and some closed servers on the floor.

To mitigate against total failure, install and use Site Recovery Manager. This solution moves all virtual machine failures to a separate domain, and does need a separate standby domain. This solution is not active/active, but Site Recovery Manager can be combined with SAN Volume Controller Stretched Cluster to use its remote mirroring (Metro Mirror or Global Mirror) capabilities.

6.4 SAN Volume Controller diagnosis and recovery guidelines

This section provides some guidelines about diagnosing a critical event in one of the two sites where the Split I/O Group has been implemented.

With these guidelines, you can determine the extent of any damage, what is still running, what can be recovered with which impact on the performance.

6.4.1 Critical event scenarios and complete domain failure

A Stretched Cluster environment can face many different critical event scenarios. Some of them can be handled by using standard (Business as Usual) recovery procedures. This section addresses in detail all the operations that are required to recover from a *complete site failure*.

The following is the list of scenarios you can face and their required recovery actions:

- Backend storage box failure in one failure domain: Business As Usual because of SAN Volume Controller 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 performance impact. You will eventually need to take appropriate action to minimize application impacts.
- SAN Volume Controller node failure in one failure domain: Business As Usual because of SAN Volume Controller High Availability.
- Complete site failure in one failure domain: This is the scope of chapter. All the required operations are detailed starting with 6.4.2, “SAN Volume Controller diagnosis guidelines” on page 141.

6.4.2 SAN Volume Controller diagnosis guidelines

This section provides some guidelines about how to diagnose a critical event in one of the two sites where the Stretched Cluster is implemented.

The Stretched Cluster configuration that is used in the examples is shown in Figure 6-2.

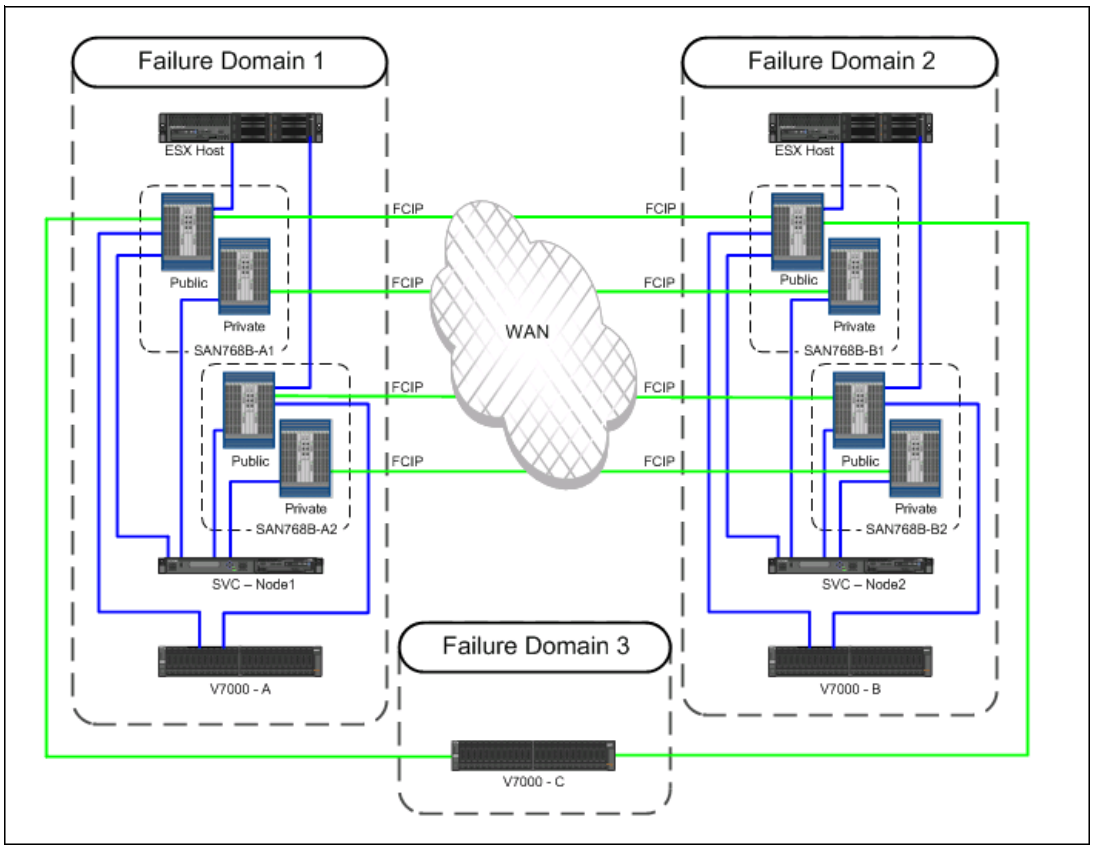


Figure 6-2 Environment diagram

The configuration that is implemented must be consistent with one of the supported configurations in 3.4, “SAN Volume Controller Stretched Cluster configurations” on page 34.

Up and running scenario analysis

In this scenario, all components are up and running and all the guidelines were applied when implementing the solution as shown in the CLI command output in Example 6-9.

Example 6-9 Running example

```
IBM_2145:ITSO_SVC_SPLIT:superuser>svcinfo lscluster
id               name           location partnership bandwidth id_alias
0000020060C14FBE ITSO_SVC_SPLIT local                          0000020060C14FBE

IBM_2145:ITSO_SVC_SPLIT:superuser>lsiogrp
id name            node_count vdisk_count host_count
0  ITSO_SVC_SPLIT  2          0           0
1  io_grp1         0          0           0
2  io_grp2         0          0           0
3  io_grp3         0          0           0
4  recovery_io_grp 0          0           0

IBM_2145:ITSO_SVC_SPLIT: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  ITSO_SVC_NODE1_SITE_A 100006B119        500507680100B13F online 0
ITSO_SVC_SPLIT yes          20400000064810 49 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea             151580
2  ITSO_SVC_NODE1_SITE_B 100006B074        500507680100B0C6 online 0
ITSO_SVC_SPLIT no           20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1siteb             151523

IBM_2145:ITSO_SVC_SPLIT: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_SITEA_N1   2076                 IBM                  2145
3  ITSO_V7K_SITEB_N1   2076                 IBM                  2145
5  ITSO_V7K_SITEC_Q_N2 2076                 IBM                  2145
6  ITSO_V7K_SITEC_Q_N1 2076                 IBM                  2145

IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdiskgrp
id name       status mdisk_count vdisk_count capacity extent_size free_capacity
virtual_capacity used_capacity real_capacity overallocation warning easy_tier
easy_tier_status compression_active compression_virtual_capacity
compression_compressed_capacity compression_uncompressed_capacity
0  V7000SITEA online 5          0          2.22TB   256        2.22TB       0.00MB
0.00MB       0.00MB       0            80      auto     active          no
0.00MB                        0.00MB                           0.00MB
1  V7000SITEB online 5          0          2.22TB   256        2.22TB       0.00MB
0.00MB       0.00MB       0            80      auto     active          no
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


IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdisk
id name                status mode    mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name   UID                                                              tier
0  ITSO_V7K_SITEA_SSD0 online managed 0            V7000SITEA     277.3GB  0000000000000000
ITSO_V7K_SITEA_N2 600507680289000268000000000000080000000000000000000000000000000000
generic_ssd
1  ITSO_V7K_SITEA_SAS0 online managed 0            V7000SITEA     500.0GB  000000000000000A
ITSO_V7K_SITEA_N2 600507680289000268000000000000040000000000000000000000000000000000
generic_hdd
2  ITSO_V7K_SITEA_SAS1 online managed 0            V7000SITEA     500.0GB  000000000000000B
ITSO_V7K_SITEA_N2 600507680289000268000000000000050000000000000000000000000000000000
generic_hdd
3  ITSO_V7K_SITEA_SAS2 online managed 0            V7000SITEA     500.0GB  000000000000000C
ITSO_V7K_SITEA_N2 600507680289000268000000000000060000000000000000000000000000000000
generic_hdd
4  ITSO_V7K_SITEA_SAS3 online managed 0            V7000SITEA     500.0GB  000000000000000D
ITSO_V7K_SITEA_N2 600507680289000268000000000000070000000000000000000000000000000000
generic_hdd
5  ITSO_V7K_SITEB_SSD1 online managed 1            V7000SITEB     277.3GB  0000000000000001
ITSO_V7K_SITEB_N2 600507680282018b300000000000000400000000000000000000000000000000000
generic_ssd
6  ITSO_V7K_SITEB_SAS0 online managed 1            V7000SITEB     500.0GB  0000000000000014
ITSO_V7K_SITEB_N2 600507680282018b300000000000000000000000000000000000000000000000000
generic_hdd
```

```
7  ITSO_V7K_SITEB_SAS1 online managed 1            V7000SITEB     500.0GB  0000000000000015
ITSO_V7K_SITEB_N2 600507680282018b3000000000000001000000000000000000000000000000000
generic_hdd
8  ITSO_V7K_SITEB_SAS2 online managed 1            V7000SITEB     500.0GB  0000000000000016
ITSO_V7K_SITEB_N2 600507680282018b3000000000000002000000000000000000000000000000000
generic_hdd
9  ITSO_V7K_SITEB_SAS3 online managed 1            V7000SITEB     500.0GB  0000000000000017
ITSO_V7K_SITEB_N2 600507680282018b3000000000000003000000000000000000000000000000000
generic_hdd
10 ITSO_V7K_SITEC_Q    online         managed 2            V7000SITEC     1.0GB
0000000000000001 ITSO_V7K_SITEC_Q_N2
600507680283801ac8000000000000005000000000000000000000000000000000 generic_hdd

IBM_2145:ITSO_SVC_SPLIT: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-01-DCA-HBA1 0           ITSO_SVC_SPLIT online 0            V7000SITEA     1.00GB
striped                         600507680183053EF800000000000000 0           1
empty            0             no        0
1  ESXi-01-DCA-HBA2 0           ITSO_SVC_SPLIT online 0            V7000SITEA     1.00GB
striped                         600507680183053EF800000000000001 0           1
empty            0             no        0
2  ESXi-02-DCB-HBB1 0           ITSO_SVC_SPLIT online 1            V7000SITEB     1.00GB
striped                         600507680183053EF800000000000002 0           1
empty            0             no        0
3  ESXi-02-DCB-HBB2 0           ITSO_SVC_SPLIT online 1            V7000SITEB     1.00GB
striped                         600507680183053EF800000000000003 0           1
empty            0             no        0
4  ESXI_CLUSTER_01  0           ITSO_SVC_SPLIT online many         many           256.00GB
many                            600507680183053EF800000000000004 0           2
empty            0             no        0
5  ESXI_CLUSTER_02  0           ITSO_SVC_SPLIT online many         many           256.00GB
many                            600507680183053EF800000000000005 0           2
empty            0             no        0
6  ESXi_Cluster_DATASTORE_A 0           ITSO_SVC_SPLIT online many         many
256.00GB many                             600507680183053EF800000000000006 0
2          empty            0             no        0
7  ESXI_Cluster_DATASTORE_B 0           ITSO_SVC_SPLIT online many         many
256.00GB many                             600507680183053EF800000000000007 0
2          empty            0             no        0



IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name               controller_id controller_name     active
object_type override
0           online 10 ITSO_V7K_SITEC_Q    5            ITSO_V7K_SITEC_Q_N2 yes    mdisk
yes
1           online 6  ITSO_V7K_SITEB_SAS0 0            ITSO_V7K_SITEB_N2   no     mdisk
yes
2           online 4  ITSO_V7K_SITEA_SAS3 1            ITSO_V7K_SITEA_N2   no     mdisk
yes
```

From the SAN Volume Controller CLI command output that is shown in Example 6-9 on page 142 you can see these aspects of the configuration:

- The SAN Volume Controller clustered system is accessible through the CLI.
- 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 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

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk ESXI_CLUSTER_01
id 4
name ESXI_CLUSTER_01
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
status online
mdisk_grp_id many
mdisk_grp_name many
capacity 256.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 600507680183053EF800000000000004
throttling 0
preferred_node_id 1
fast_write_state empty
cache readwrite
udid
fc_map_count 0
sync_rate 95
copy_count 2
se_copy_count 0
filesystem
mirror_write_priority latency
RC_change no
compressed_copy_count 0
access_IO_group_count 1

copy_id 0
status online
sync yes
primary yes
mdisk_grp_id 0
mdisk_grp_name V7000SITEA
type striped
mdisk_id
mdisk_name
.
lines omitted for brevity
.
easy_tier_status active
tier generic_ssd
tier_capacity 0.00MB
tier generic_hdd
```

```
tier_capacity 256.00GB
compressed_copy no
uncompressed_used_capacity 256.00GB

copy_id 1
status online
sync yes
primary no
mdisk_grp_id 1
mdisk_grp_name V7000SITEB
type striped
mdisk_id
mdisk_name
uncompressed_used_capacity 256.00GB
.
lines omitted for brevity
.
```

From the SAN Volume Controller CLI command output in the Example 6-10 on page 145 you can see these aspects of the configuration:

- The volume is online.
- The storage pool name and the MDisk name are *many*, which means Volume Mirroring is in place.
- Copy id 0 is *online*, in *sync* and it is the *primary.*
- Copy id 1 is *online*, in *sync* and it is the *secondary.*

If you have several volumes to check, you can create a customized script directly from the SAN Volume Controller shell. Some useful scripts can be found at:

https://www.ibm.com/developerworks/mydeveloperworks/wikis/home/wiki/W1d985101fbfa_4ae7_a090_dc535355ae7e/page/Show%20Volume%20Copies?lang=en

Critical event scenario analysis

In this scenario, the SAN Volume Controller environment has experienced a critical event that caused the complete loss of one of the sites, Site 1.

Complete the following steps to gain a complete view of any damage and to gather enough decision elements to determine what your next recovery actions will be:

1. Is SAN Volume Controller system management available through GUI or CLI?
   a. Is SAN Volume Controller system login possible?

      YES: SAN Volume Controller system is online, continue with step 2

      NO: SAN Volume Controller system is offline or suffering connection problems.

      i. Check your connections, cabling, and node front panel event messages.

      ii. Verify the SAN Volume Controller system status by using Service Assistant menu or node front panel. For more information, see *IBM System Storage SAN Volume Controller Troubleshooting Guide,* GC27-2284.

      iii. Bring a part of the SAN Volume Controller system online for further diagnostic tests.

      iv. Using a browser connect to one of the SAN Volume Controller node's service IP addresses:

      `https://<service_ip_add>/service/`

Log in with your SAN Volume Controller Cluster GUI password as shown in Figure 6-3.

Figure 6-3 SAN Volume Controller Service assistant tool login

After the login, you are redirected to the Service Assistant menu as shown in Figure 6-4

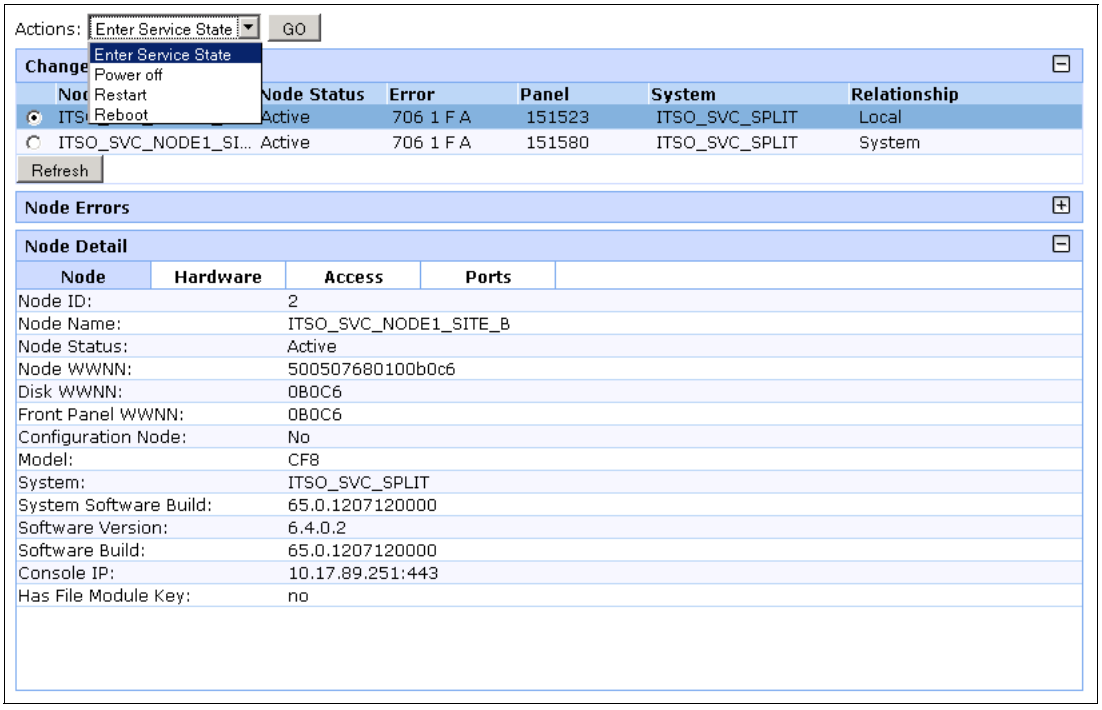


Figure 6-4 Service Assistant menu

From the Service Assistant menu, you can try to bring at least a part of the SAN Volume Controller Clustered system online for further diagnostic tests. For further and detailed information about Service Assistant menu, see *Implementing the IBM System Storage SAN Volume Controller V6.3* SG24-7933.

2. If the SAN Volume Controller system management is available:
   a. Check the status by running the SAN Volume Controller CLI commands that are shown in Example 6-11.

Example 6-11 lscluster example

```
IBM_2145:ITSO_SVC_SPLIT:superuser>svcinfo lscluster
id               name           location partnership bandwidth id_alias
0000020060C14FBE ITSO_SVC_SPLIT local                          0000020060C14FBE

IBM_2145:ITSO_SVC_SPLIT:superuser>svcinfo lscluster ITSO_SVC_SPLIT
id 0000020060C14FBE
name ITSO_SVC_SPLIT
location local
partnership
bandwidth
.
lines ommitted for brevity
.
total_mdisk_capacity 4.4TB
space_in_mdisk_grps 4.4TB
space_allocated_to_vdisks 0.00MB
total_free_space 4.4TB
statistics_status on
statistics_frequency 15
required_memory 0
gm_max_host_delay 5
tier generic_ssd
tier_capacity 554.50GB
tier_free_capacity 554.50GB
tier generic_hdd
tier_capacity 3.90TB
tier_free_capacity 3.90TB
email_contact2
email_contact2_primary
email_contact2_alternate
total_allocated_extent_capacity 1.50GB
has_nas_key no
auth_service_type tip
layer replication
rc_buffer_size 48
IBM_2145:ITSO_SVC_SPLIT:superuser>
```

The example shows that the SAN Volume Controller clustered system looks accessible. The same view in the GUI is shown in Figure 6-5.

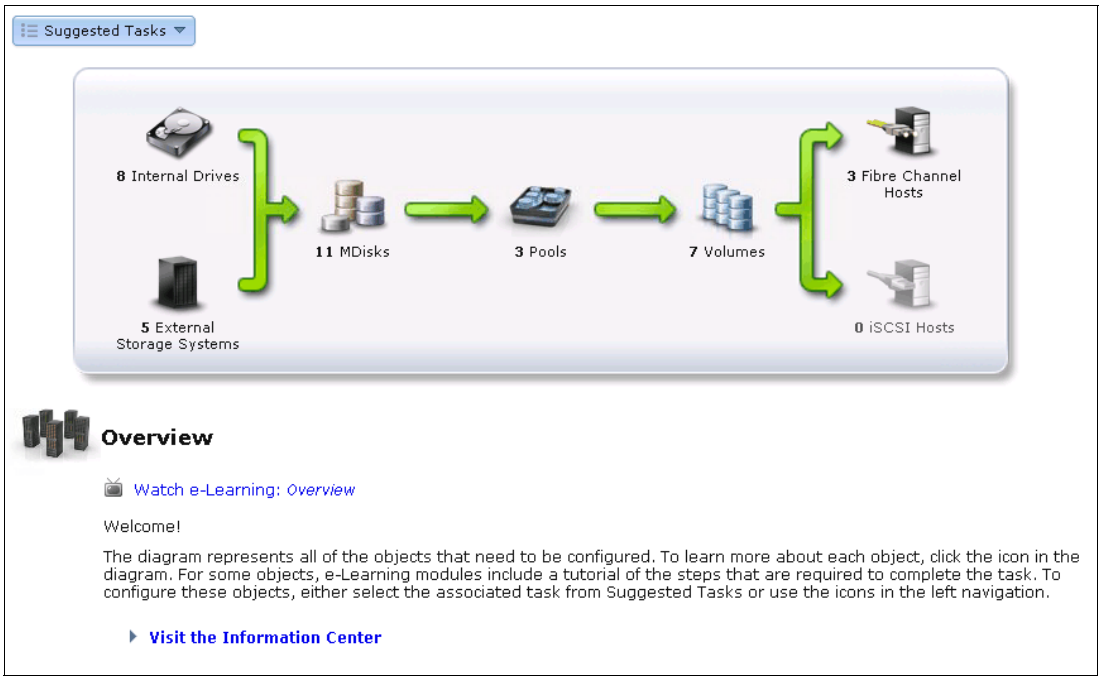


Figure 6-5 GUI example

b. Check the status of the nodes as shown in Example 6-12.

Example 6-12 Node status example

```
IBM_2145:ITSO_SVC_SPLIT: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  ITSO_SVC_NODE1_SITE_A 100006B119        500507680100B13F offline 0
ITSO_SVC_SPLIT no          2040000006481049 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea             151580
2  ITSO_SVC_NODE2_SITE_B 100006B074        500507680100B0C6 online  0
ITSO_SVC_SPLIT yes         2040000006480 1C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb             151523

IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode ITSO_SVC_NODE1_SITE_A
id 1
name ITSO_SVC_NODE1_SITE_A
UPS_serial_number 100006B119
WWNN 500507680100B13F
status offline
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
partner_node_id 2
partner_node_name ITSO_SVC_NODE2_SITE_B
config_node no
UPS_unique_id 2040000006481049
port_id 500507680140B13F
port_status inactive
port_speed 8Gb
port_id 500507680130B13F
port_status inactive
```

```
port_speed 8Gb
port_id 500507680110B13F
port_status inactive
port_speed 8Gb
port_id 500507680120B13F
port_status inactive
port_speed 8Gb
hardware CF8
iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea
iscsi_alias
failover_active no
failover_name ITSO_SVC_NODE2_SITE_B
failover_iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb
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

IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode ITSO_SVC_NODE2_SITE_B
id 2
name ITSO_SVC_NODE2_SITE_B
UPS_serial_number 100006B074
WWNN 500507680100B0C6
status online
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
partner_node_id 1
partner_node_name ITSO_SVC_NODE1_SITE_A
config_node yes
UPS_unique_id 20400000064801C4
port_id 500507680140B0C6
port_status active
port_speed 8Gb
port_id 500507680130B0C6
port_status active
port_speed 8Gb
port_id 500507680110B0C6
port_status active
port_speed 8Gb
port_id 500507680120B0C6
port_status active
port_speed 8Gb
hardware CF8
iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb
iscsi_alias
failover_active no
failover_name ITSO_SVC_NODE1_SITE_A
failover_iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea
failover_iscsi_alias
panel_name 151523
enclosure_id
```

```
canister_id
enclosure_serial_number
service_IP_address 10.17.89.254
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
```

Observe the following statuses in Example 6-12 on page 149:

- The *config node* role has moved from node 1 to node 2.
- Node 1 is offline.
- FC ports in node 1 are inactive.
- Node 2 is online.
- FC ports in node 2 are still online.

During this event, the system has lost 50% of the SAN Volume Controller clustered system resources, but it is still up and running.

This information can also be seen in the GUI as shown in Figure 6-6.

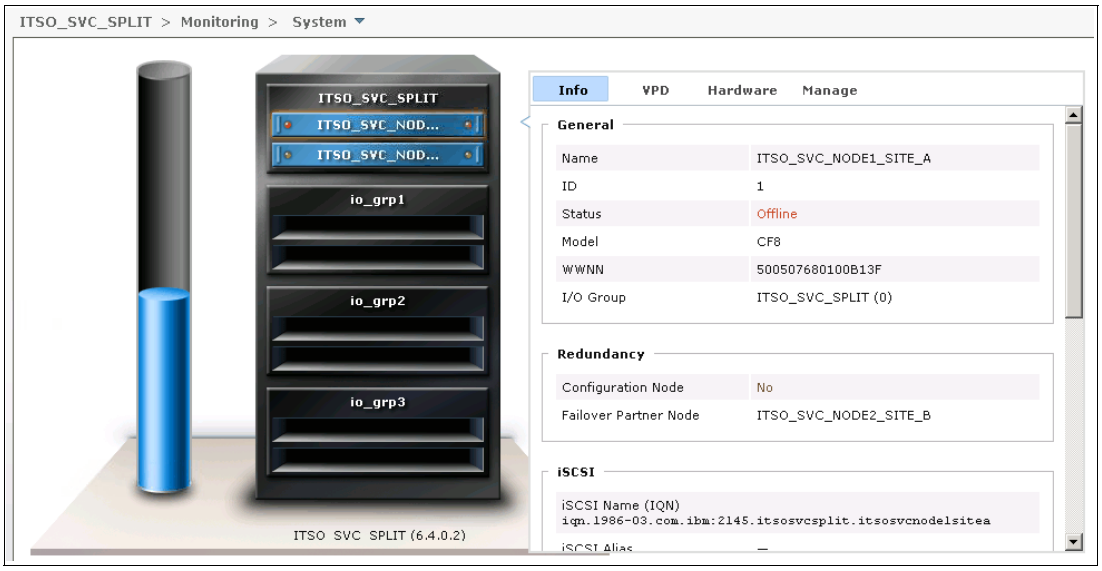


Figure 6-6 Node status with GUI

c. Check the IO Group status as shown in Example 6-13.

Example 6-13 I/O Group status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsiogrp
id name            node_count vdisk_count host_count
0  ITSO_SVC_SPLIT  2          8           2
1  io_grp1         0          0           0
2  io_grp2         0          0           0
3  io_grp3         0          0           0
4  recovery_io_grp 0          0           0
```

As you can see, the I/O Group still reports two nodes per I/O Group.

d. Check the quorum status as shown in Example 6-14.

Example 6-14 Quorum status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name                controller_id controller_name     active
object_type override
0            online 10 ITSO_V7K_SITEC_Q    5             ITSO_V7K_SITEC_Q_N2 yes
mdisk       yes
1            online 6  ITSO_V7K_SITEB_SAS0 0             ITSO_V7K_SITEB_N2   no
mdisk       yes
2            online 5  ITSO_V7K_SITEB_SSD1 0             ITSO_V7K_SITEB_N2   no
mdisk       ignored
```

The active quorum disk is still active because it was not affected by the critical event. However, the quorum index 1 that is in the site that suffered the power failure has flagged it with override ignored. It is flagged because if the original resource in DS3400_09 goes offline and another resource is used instead, the override field in **`lsquorum`** shows as **`ignored`**.

e. Check the controller's status as shown in Example 6-15. Some controllers might be offline as expected, or might have *path_count = 0* because of the site failure.

Example 6-15 Controller status

```
IBM_2145:ITSO_SVC_SPLIT: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_SITEA_N1   2076                 IBM                  2145
3  ITSO_V7K_SITEB_N1   2076                 IBM                  2145
5  ITSO_V7K_SITEC_Q_N2 2076                 IBM                  2145
6  ITSO_V7K_SITEC_Q_N1 2076                 IBM                  2145

IBM_2145:ITSO_SVC_SPLIT:superuser>lscontroller ITSO_V7K_SITEA_N2
id 1
controller_name ITSO_V7K_SITEA_N2
WWNN 50050768020000F0
mdisk_link_count 5
max_mdisk_link_count 5
degraded yes
vendor_id IBM
product_id_low 2145
product_id_high
product_revision 0000
ctrl_s/n 2076
allow_quorum yes
WWPN 50050768023000F0
path_count 0
max_path_count 6
WWPN 50050768024000F0
path_count 0
max_path_count 6
WWPN 50050768021000F0
path_count 0
max_path_count 6
WWPN 50050768022000F0
path_count 0
max_path_count 6
```

As you can see in the output, some controllers are still accessible from the SAN Volume Controller system. Others are no longer accessible because the power loss in site 1 has affected the SAN Volume Controller node, the storage subsystem, and the FC SAN switches.

The same information can be gotten from the GUI as shown in Figure 6-7.

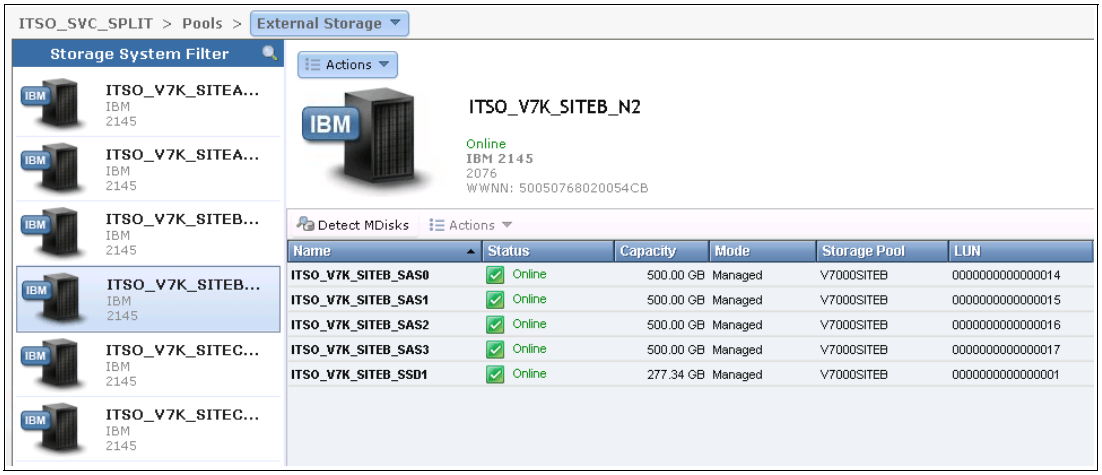


Figure 6-7 Online controllers

The offline controller is shown in Figure 6-8.

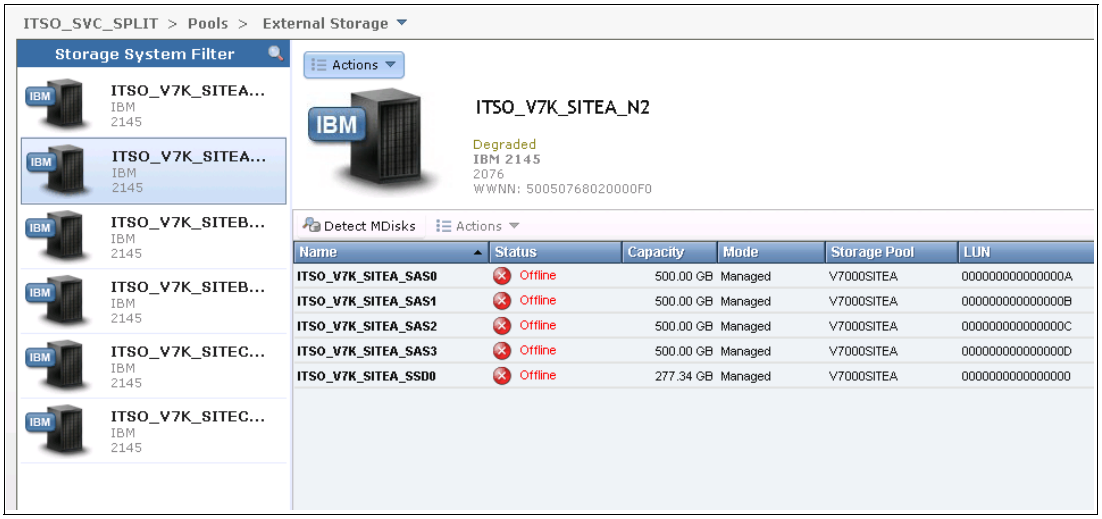


Figure 6-8 Offline controller

f. Check the MDisk group status as shown in Example 6-16.

Example 6-16 MDisk group status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdiskgrp
id name       status  mdisk_count vdisk_count capacity extent_size free_capacity
virtual_capacity used_capacity real_capacity overallocation warning easy_tier
easy_tier_status compression_active compression_virtual_capacity
compression_compressed_capacity compression_uncompressed_capacity
0  V7000SITEA offline 5           6           2.22TB   256         1.22TB       1.00TB
1.00TB       1.00TB       45            80      auto      active          no
0.00MB                      0.00MB                          0.00MB
```

```
1  V7000SITEB online  5          6          2.22TB   256        1.22TB       1.00TB
1.00TB       1.00TB       45           80      auto     active          no
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
IBM_2145:ITSO_SVC_SPLIT:superuser>
```

Because of the critical event, some are offline and others are still online. The ones offline are those that had space allocated on the storage subsystem that suffered the critical event.

g. Check the MDisk status as shown in Example 6-17.

Example 6-17 MDisk status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdisk
id name               status  mode    mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name     UID                                                              tier
0  ITSO_V7K_SITEA_SSD0 offline managed 0            V7000SITEA     277.3GB
0000000000000000 ITSO_V7K_SITEA_N2
600507680289000268000000000000080000000000000000000000000000000000 generic_ssd
1  ITSO_V7K_SITEA_SAS0 offline managed 0            V7000SITEA     500.0GB
000000000000000A ITSO_V7K_SITEA_N2
600507680289000268000000000000040000000000000000000000000000000000 generic_hdd
2  ITSO_V7K_SITEA_SAS1 offline managed 0            V7000SITEA     500.0GB
000000000000000B ITSO_V7K_SITEA_N2
600507680289000268000000000000050000000000000000000000000000000000 generic_hdd
3  ITSO_V7K_SITEA_SAS2 offline managed 0            V7000SITEA     500.0GB
000000000000000C ITSO_V7K_SITEA_N2
600507680289000268000000000000060000000000000000000000000000000000 generic_hdd
4  ITSO_V7K_SITEA_SAS3 offline managed 0            V7000SITEA     500.0GB
000000000000000D ITSO_V7K_SITEA_N2
600507680289000268000000000000070000000000000000000000000000000000 generic_hdd
5  ITSO_V7K_SITEB_SSD1 online  managed 1            V7000SITEB     277.3GB
0000000000000001 ITSO_V7K_SITEB_N2
600507680282018b3000000000000004000000000000000000000000000000000 generic_ssd
6  ITSO_V7K_SITEB_SAS0 online  managed 1            V7000SITEB     500.0GB
0000000000000014 ITSO_V7K_SITEB_N2
600507680282018b3000000000000000000000000000000000000000000000000 generic_hdd
7  ITSO_V7K_SITEB_SAS1 online  managed 1            V7000SITEB     500.0GB
0000000000000015 ITSO_V7K_SITEB_N2
600507680282018b3000000000000001000000000000000000000000000000000 generic_hdd
8  ITSO_V7K_SITEB_SAS2 online  managed 1            V7000SITEB     500.0GB
0000000000000016 ITSO_V7K_SITEB_N2
600507680282018b3000000000000002000000000000000000000000000000000 generic_hdd
9  ITSO_V7K_SITEB_SAS3 online  managed 1            V7000SITEB     500.0GB
0000000000000017 ITSO_V7K_SITEB_N2
600507680282018b3000000000000003000000000000000000000000000000000 generic_hdd
10 ITSO_V7K_SITEC_Q    online  managed 2            V7000SITEC     1.0GB
0000000000000001 ITSO_V7K_SITEC_Q_N2
600507680283801ac8000000000000050000000000000000000000000000000000 generic_hdd
```

Again, because of the critical event, some are offline and others are still online. The ones offline are those that had space allocated on the storage subsystem that suffered the critical event.

You can get the same information from the GUI as shown in Figure 6-9.

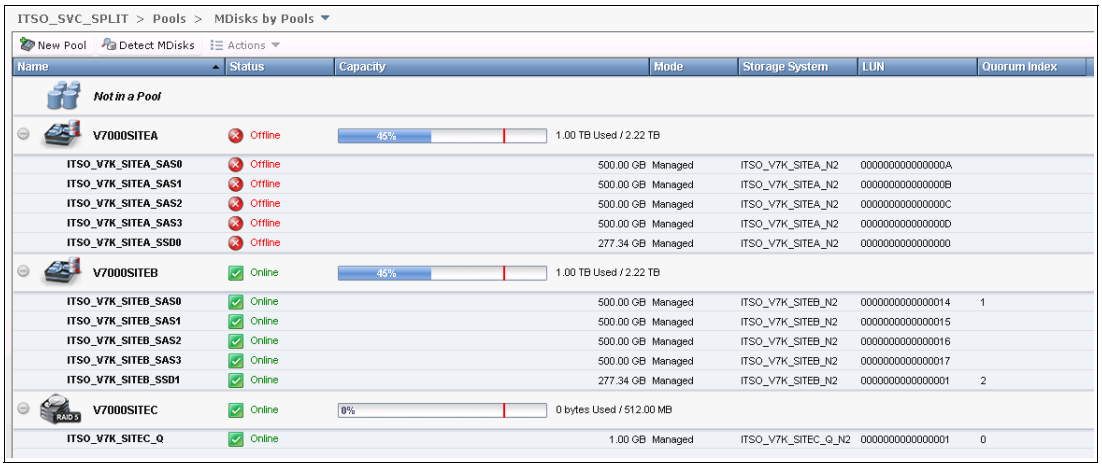


Figure 6-9 MDisk status with GUI

h. Check the volume status as shown in Example 6-18.

Example 6-18 Volume status

```
IBM_2145:ITSO_SVC_SPLIT: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-01-DCA-HBA1         0           ITSO_SVC_SPLIT offline  0            V7000SITEA
2.00GB   striped                              600507680183053EF800000000000000 0
1          not_empty        0             no        0
1  ESXi-01-DCA-HBA2         0           ITSO_SVC_SPLIT offline  0            V7000SITEA
2.00GB   striped                              600507680183053EF800000000000001 0
1          not_empty        0             no        0
2  ESXi-02-DCB-HBB1         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000002 0
1          empty            0             no        0
3  ESXi-02-DCB-HBB2         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000003 0
1          empty            0             no        0
4  ESXI_CLUSTER_01          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                              600507680183053EF800000000000004 0
2          empty            0             no        0
5  ESXI_CLUSTER_02          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                              600507680183053EF800000000000005 0
2          empty            0             no        0
6  ESXi_Cluster_DATASTORE_A 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                              600507680183053EF800000000000006 0
2          empty            0             no        0
7  ESXI_Cluster_DATASTORE_B 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                              600507680183053EF800000000000007 0
2          empty            0             no        0
```

As you can see from the output in Example 6-18, you have lost 50% of the resources because of a loss of power in site 1. The volumes are not offline, but in a degraded state. Volume Mirroring acted to ensure business continuity. The volumes are still accessible from the hosts that are still running in site 2 where power is still present.

Some other volumes are offline because they are not protected by the Volume Mirror feature, but this was known during the planning of this environment.

It might be helpful to use the **filtervalue** option in the CLI command to reduce the number of lines that are produced and volumes to check as shown in Example 6-19.

Example 6-19 Volume status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk -filtervalue status=degraded
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
2  ESXi-02-DCB-HBB1         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000002 0
1          empty            0             no        0
3  ESXi-02-DCB-HBB2         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000003 0
1          empty            0             no        0
4  ESXI_CLUSTER_01          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                               600507680183053EF800000000000004 0
2          empty            0             no        0
5  ESXI_CLUSTER_02          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                               600507680183053EF800000000000005 0
2          empty            0             no        0
6  ESXi_Cluster_DATASTORE_A 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                               600507680183053EF800000000000006 0
2          empty            0             no        0
7  ESXI_Cluster_DATASTORE_B 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                               600507680183053EF800000000000007 0
2          empty            0             no        0
```

As you can see from the output in Example 6-19, one copy of each volume is offline and you can also see which storage pool it is related to.

You can get the same information from the GUI as shown in Figure 6-10.

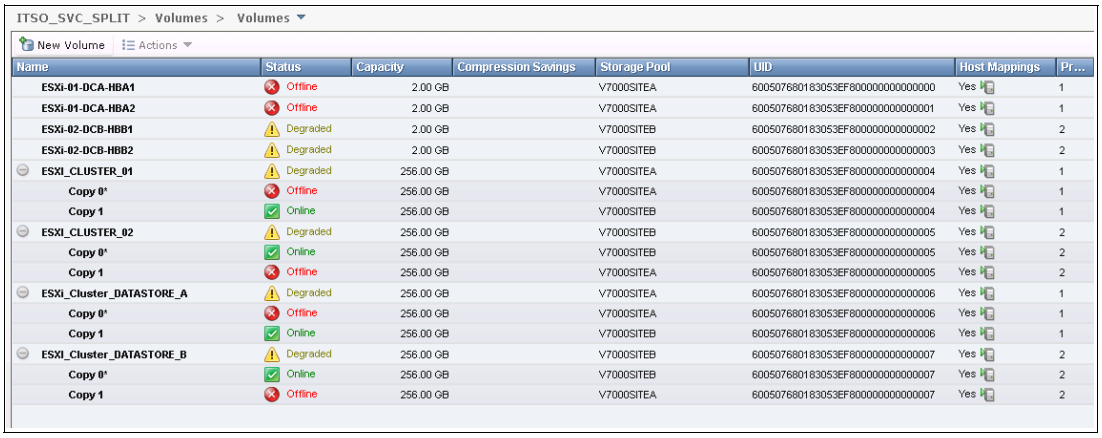


Figure 6-10 Volume status with GUI

As you can see from Figure 6-10, you can easily see which resources are online, which are not, and why you have a degraded status related to each volume.

3. Check path status.

   Check the status of the storage paths from your hosts point of view by using your multipathing software commands. For SAN Volume Controller, the best multipath software to use is Subsystem Device Driver (SDD). For more information about SDD commands, see:

   http://www-01.ibm.com/support/docview.wss?rs=540&context=ST52G7&uid=ssg1S7000303

You can also see the *Multipath Subsystem Device Driver User's Guide*, GC52-1309-03.

You can also verify the SDD vpath device configuration by entering the **`lsvpcfg`** or **`datapath query device`** command.

All of these steps are also valid for a limited failure where we are facing a failure with limited impact in one of the sites.

In a limited failure, it can be helpful to use the following steps to verify the status of your Split I/O Group infrastructure.

4. Check your SAN by using the FC switch or director CLI, or web interface to verify any failure.
5. Check the FC and FC/IP connection between the two sites by using the FC switch or director CLI, or web interface to verify any eventual partial failure.
6. Check the storage subsystem status by using its own management interface to verify any failure.

These steps help you identify the root cause and the impact of the event on your infrastructure. After you have this information, select one of the following strategic decisions:

- Wait until the failure in one of the two sites is fixed
- Declare a disaster and start with the recovery actions that are described in 6.4.3, “SAN Volume Controller Recovery guidelines” on page 157.

If you decide to wait until the failure in one of the two sites is fixed, when the resources become available again, the SAN Volume Controller Split I/O Group will be fully operational. In addition, the following events occur:

- Automatic Volume Mirroring resynchronization takes place
- Missing nodes rejoin the SAN Volume Controller clustered system

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 6.4.3, “SAN Volume Controller Recovery guidelines” on page 157.

6.4.3 SAN Volume Controller Recovery guidelines

This section explores some recovery scenarios. Regardless of the scenario, the common starting point is the complete loss of site 1 or site 2 caused by a severe 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 so that the surviving site configuration is rebuilt so that it provides the same performance and availability characteristics as it did before

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 supply to recover your Stretched Cluster configuration

If you cannot recover the site in a reasonable time, you must take some 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 will 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 brand 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 (test environment) at the moment?
- 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 some 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, and monetary steps to take. These steps must be part of a recovery plan to create a minimal impact to applications and therefore service levels.

The recovery guidelines assume that you have answered the questions and decided to recover a fully redundant configuration in the same surviving site. The solution involves new SAN Volume Controller nodes, new storage subsystems, and new FC SAN devices. Reusing SAN Volume Controller nodes, storage, and SAN devices that are already available is also covered, as well as guidelines on how to plan a failback scenario.

If you must recover your Split I/O Group infrastructure, involve IBM Support as early as possible.

Recovery guidelines for the example configuration

The following recovery guidelines assume that you have decided to recover a fully redundant configuration in the same surviving site by supplying new SAN Volume Controller nodes, storage subsystems, and FC SAN devices. This recovery action is based on a decision to recover the Stretched Cluster infrastructure at the same performance characteristics as before the critical event. However, the solution has limited business continuity because the Stretched Cluster is recovered at one site only.

The description assumes that you have already received and installed a new SAN Volume Controller node, FC switches, and backend storage subsystems.

Figure 6-11 shows the new recovery configuration.

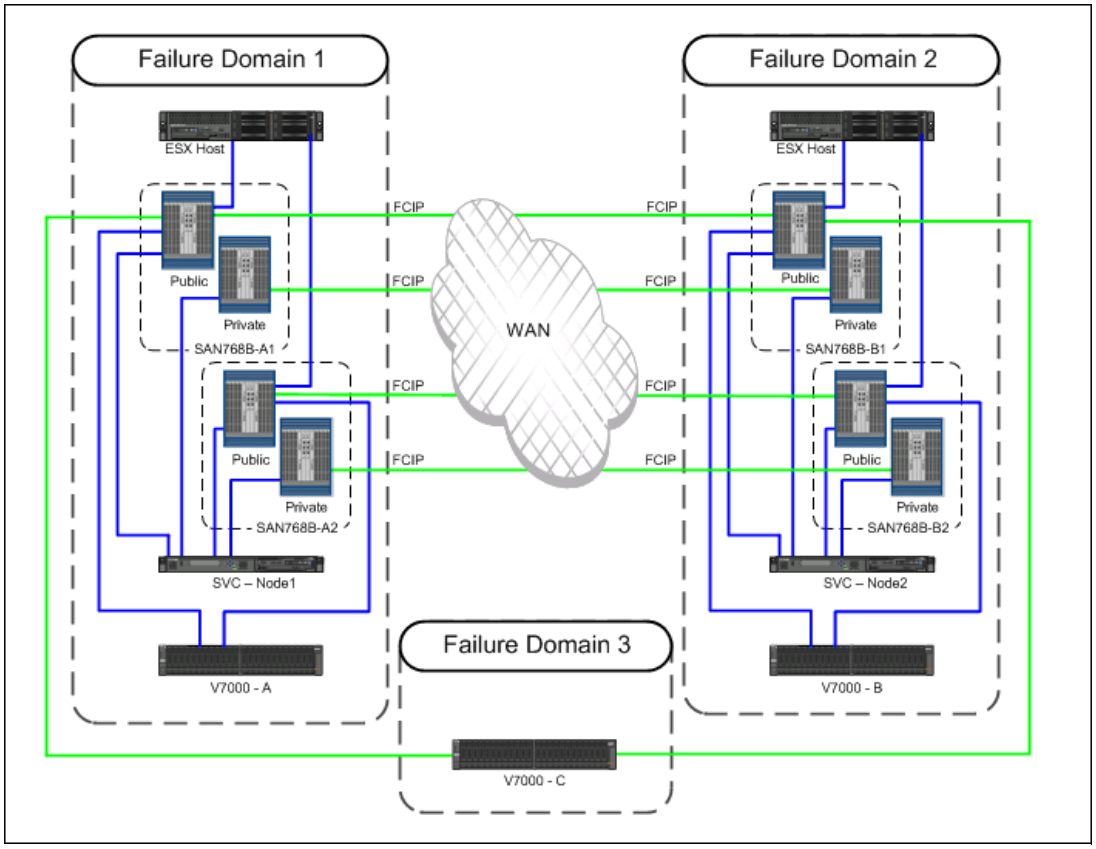


Figure 6-11 New recovery configuration in surviving site

The configuration will be recovered exactly as it was, even if has been recovered in the same site. You can make it easier in the future to implement this configuration over distance when a new site is provided by completing the following major steps:

1. Disconnect the FCIP links between Failure Domains
2. Uninstall/reinstall all the new devices in the new site
3. Reconnect the FCIP links between Failure Domains

The following steps must be run to recover your Split I/O Group configuration as it was before the critical event in the same site. They are completed after you install the new devices.

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 just configuring the new switches with the same domain id as before, and connecting them to the surviving switches through the passive WDM. The WWPN zoning is then automatically propagated to the new switches.
3. Connect, if possible, the new storage subsystems to the same FC switch ports as before the critical event. SAN Volume Controller to storage zoning must be reconfigured to be able to see the new storage subsystem's WWNN. Old WWNNs can be removed, but take care to remove the right ones because you have only one volume copy active.

4. Do not connect the SAN Volume Controller node FC yet. Wait until directed to do so by the SAN Volume Controller node's WWNN change procedure.
5. Remove the offline node from the SAN Volume Controller system configuration with the CLI commands shown in Example 6-20.

Example 6-20 Remove node command

```
IBM_2145:ITSO_SVC_SPLIT: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  ITSO_SVC_NODE1_SITE_A 100006B119        500507680100B13F offline 0
ITSO_SVC_SPLIT no          2040000006481049 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea             151580
2  ITSO_SVC_NODE2_SITE_B 100006B074        500507680100B0C6 online  0
ITSO_SVC_SPLIT yes         20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb             151523

IBM_2145:ITSO_SVC_SPLIT:superuser>rmnode ITSO_SVC_NODE1_SITE_A

IBM_2145:ITSO_SVC_SPLIT: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
2  ITSO_SVC_NODE2_SITE_B 100006B074        500507680100B0C6 online 0
ITSO_SVC_SPLIT yes         20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb             151523
```

You can also use the GUI as shown in Figure 6-12.

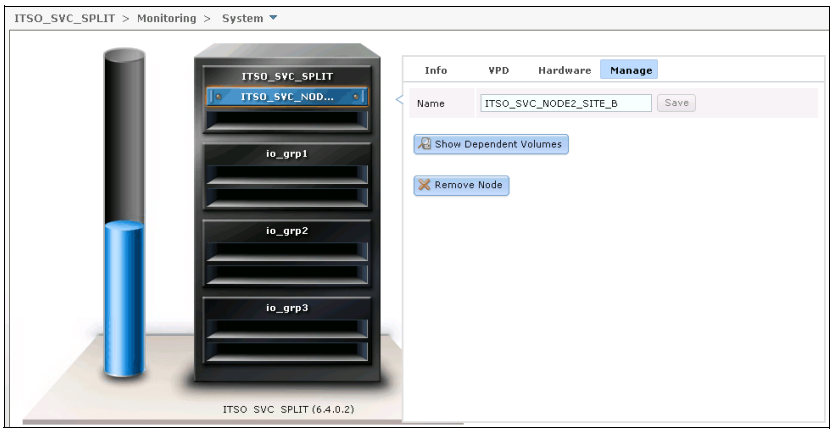


Figure 6-12 Removing a node by using the GUI

6. Remove the Volume Mirroring definitions. First, identify which copy id is offline for each volume by using the SAN Volume Controller CLI command shown in Example 6-21.

Example 6-21 Identifying Volume copy ID

```
IBM_2145:ITSO_SVC_SPLIT: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-01-DCA-HBA1        0          ITSO_SVC_SPLIT offline  0           V7000SITEA
2.00GB   striped                          600507680183053EF800000000000000 0
1          not_empty        0             no        0
```

```
1  ESXi-01-DCA-HBA2        0          ITSO_SVC_SPLIT offline  0           V7000SITEA
2.00GB   striped                          600507680183053EF800000000000001 0
1          not_empty        0             no        0
2  ESXi-02-DCB-HBB1        0          ITSO_SVC_SPLIT degraded 1           V7000SITEB
2.00GB   striped                          600507680183053EF800000000000002 0
1          empty            0             no        0
3  ESXi-02-DCB-HBB2        0          ITSO_SVC_SPLIT degraded 1           V7000SITEB
2.00GB   striped                          600507680183053EF800000000000003 0
1          empty            0             no        0
4  ESXI_CLUSTER_01         0          ITSO_SVC_SPLIT degraded many        many
256.00GB many                          600507680183053EF800000000000004 0
2          empty            0             no        0
5  ESXI_CLUSTER_02         0          ITSO_SVC_SPLIT degraded many        many
256.00GB many                          600507680183053EF800000000000005 0
2          empty            0             no        0
6  ESXi_Cluster_DATASTORE_A 0          ITSO_SVC_SPLIT degraded many        many
256.00GB many                          600507680183053EF800000000000006 0
2          empty            0             no        0
7  ESXI_Cluster_DATASTORE_B 0          ITSO_SVC_SPLIT degraded many        many
256.00GB many                          600507680183053EF800000000000007 0
2          empty            0             no        0

IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk ESXi_Cluster_DATASTORE_A
id 6
name ESXi_Cluster_DATASTORE_A
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
status degraded
mdisk_grp_id many
mdisk_grp_name many
capacity 256.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 600507680183053EF800000000000006
throttling 0
preferred_node_id 0
fast_write_state empty
cache readwrite
udid
fc_map_count 0
sync_rate 95
copy_count 2
se_copy_count 0
filesystem
mirror_write_priority redundancy
RC_change no
compressed_copy_count 0
access_IO_group_count 1

copy_id 0
status offline
sync no
primary yes
mdisk_grp_id 0
```

```
mdisk_grp_name V7000SITEA
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 256.00GB
real_capacity 256.00GB
free_capacity 0.00MB
overallocation 100
autoexpand
warning
grainsize
se_copy no
easy_tier on
easy_tier_status active
tier generic_ssd
tier_capacity 15.75GB
tier generic_hdd
tier_capacity 240.25GB
compressed_copy no
uncompressed_used_capacity 256.00GB

copy_id 1
status online
sync yes
primary no
mdisk_grp_id 1
mdisk_grp_name V7000SITEB
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 256.00GB
real_capacity 256.00GB
free_capacity 0.00MB
overallocation 100
autoexpand
warning
grainsize
se_copy no
easy_tier on
easy_tier_status active
tier generic_ssd
tier_capacity 14.00GB
tier generic_hdd
tier_capacity 242.00GB
compressed_copy no
uncompressed_used_capacity 256.00GB
```

You can also use the GUI as shown in Figure 6-13.

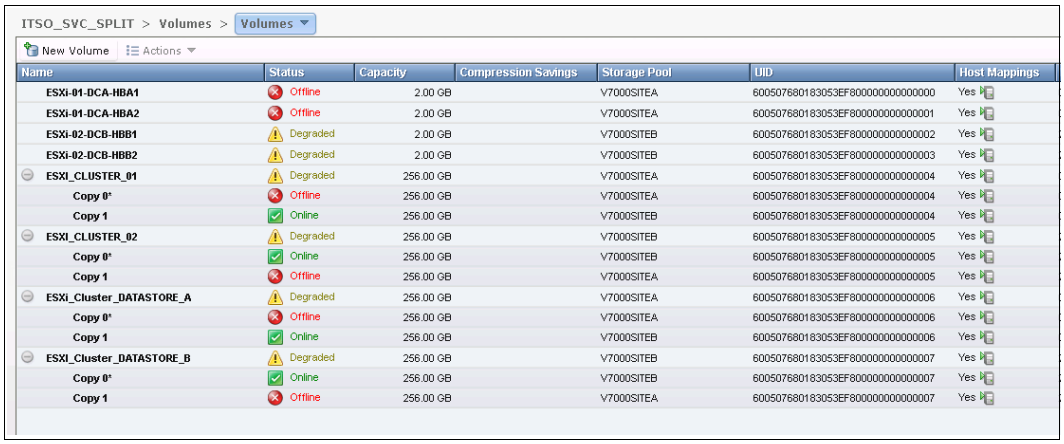


Figure 6-13 Identifying the offline copy ID with the GUI

7. Remove each identified offline Volume Mirroring copy with the SAN Volume Controller CLI command shown in Example 6-22.

Example 6-22 rmvdiskcopy output

```
IBM_2145:ITSO_SVC_SPLIT:superuser>rmvdiskcopy -copy 0 ESXi_Cluster_DATASTORE_A

IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk ESXi_Cluster_DATASTORE_A
id 6
name ESXi_Cluster_DATASTORE_A
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
status degraded
mdisk_grp_id 1
mdisk_grp_name V7000SITEB
capacity 256.00GB
type striped
formatted no
mdisk_id
mdisk_name
FC_id
FC_name
RC_id
RC_name
vdisk_UID 600507680183053EF800000000000006
throttling 0
preferred_node_id 0
fast_write_state empty
cache readwrite
udid
fc_map_count 0
sync_rate 95
copy_count 1
se_copy_count 0
filesystem
mirror_write_priority redundancy
RC_change no
compressed_copy_count 0
access_IO_group_count 1

copy_id 1
```

```
status online
sync yes
primary yes
mdisk_grp_id 1
mdisk_grp_name V7000SITEB
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 256.00GB
real_capacity 256.00GB
free_capacity 0.00MB
overallocation 100
autoexpand
warning
grainsize
se_copy no
easy_tier on
easy_tier_status active
tier generic_ssd
tier_capacity 14.00GB
tier generic_hdd
tier_capacity 242.00GB
compressed_copy no
```

You can also use the GUI as shown in Figure 6-14.

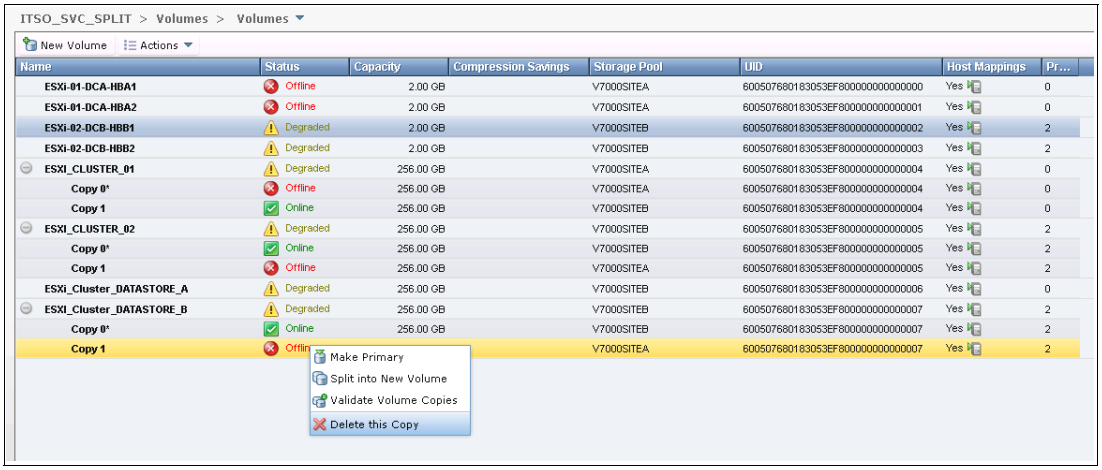


Figure 6-14 Deleting the copy with the GUI

8. Power on the new node, but leave the FC cable disconnected
9. Change the new node WWNN by using the following procedure:
   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 it is part of an SAN Volume Controller clustered system. If this is a new node from IBM, this error should not occur.

Change the WWNN of the replacement node to match the WWNN that you recorded earlier by following these steps:

b. From the front panel of the new node, navigate to the Node panel, and then navigate to the Node WWNN panel.

   Press and hold the down button, press and release the select button, and then release the down button. Line one should be Edit WWNN and line two is the last five numbers of this new node's WWNN.

c. 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.

d. Press the up or down button to increment or decrement the character that is displayed.

   The characters wrap F to 0 or 0 to F.

e. Press the left navigation button to move to the next field or the right navigation button to return to the previous field, and repeat step d for each field. At the end of this step, the characters that are displayed must be the same as the WWNN you recorded in step a.

f. Press the select button to retain the characters that you have updated and return to the WWNN panel.

g. 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 f and g instruct you to do. After step f it might seem that the WWNN is changed, but it is step g that applies the change.

h. Ensure that the WWNN has changed by following step a again.

10. 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 impact 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 will cause port ids to change, which can affect the hosts' access to volumes or cause problems with 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

11. Issue the SAN Volume Controller CLI command as shown in Example 6-23 to verify that the last five characters of the WWNN are correct.

Example 6-23 Verifying candidate node with the correct WWNN

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsnodecandidate
id               panel_name UPS_serial_number UPS_unique_id    hardware
500507680100B13F 151580     100006B119        2040000006481049 CF8
```

Important: If the WWNN does not match the original node's WWNN exactly as recorded, repeat steps 8b to 8g.

12. 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 SAN Volume Controller CLI commands shown in Example 6-24.

Example 6-24 Adding a node

```
IBM_2145:ITSO_SVC_SPLIT:superuser>addnode -wwnodename 500507680100B13F -iogrp 0
Node, id [3], successfully added

IBM_2145:ITSO_SVC_SPLIT: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
3  ITSO_SVC_NODE1_SITE_A 100006B119        500507680100B13F online 0
ITSO_SVC_SPLIT no          2040000006481049 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea             151580
2  ITSO_SVC_NODE2_SITE_B 100006B074        500507680100B0C6 online 0
ITSO_SVC_SPLIT yes         20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb             151523
IBM_2145:ITSO_SVC_SPLIT:superuser>
```

You can also use the GUI as shown in Figure 6-15.

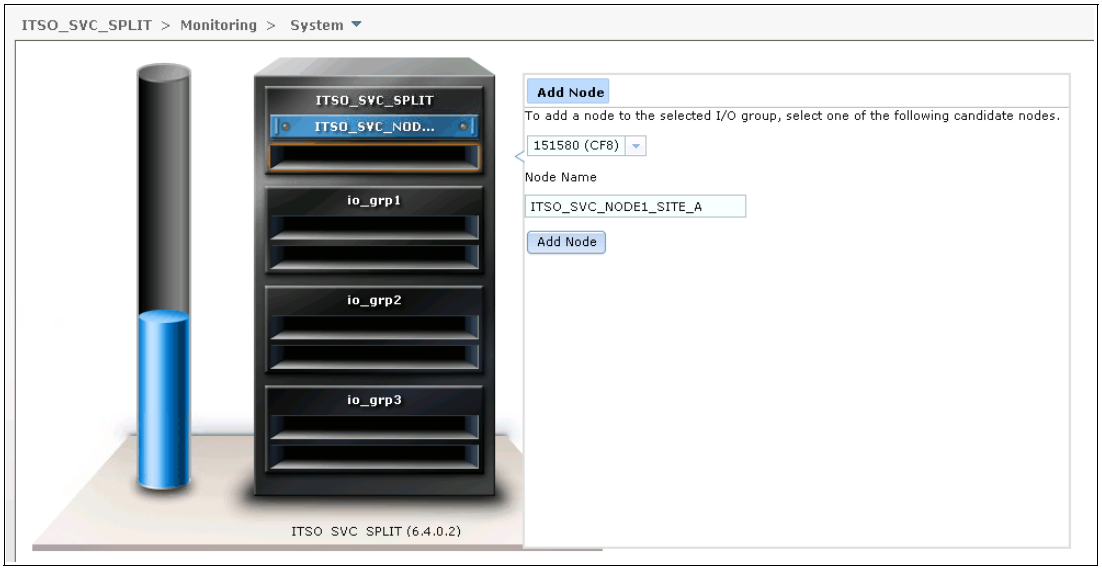


Figure 6-15 Adding a node by using the GUI

13. Verify that all volumes for this I/O group are back online and are no longer degraded. If the node replacement process is being done disruptively (no I/O is occurring to the I/O group),

wait a while (generally 30 minutes) to ensure that the new node is back online and available to take over before you do the next node in the I/O group.

Use the SAN Volume Controller CLI command that is shown in Example 6-25 to verify that all volumes for this I/O group are back online and are no longer degraded.

Example 6-25 No longer degraded volume

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk -filtervalue status=degraded
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk -filtervalue status=offline
```

14. Use the CLI to discover the new MDisk supplied by the new back-end storage subsystem. The MDisk is displayed as status online with a mode of *unmanaged* as shown in Example 6-26.

Example 6-26 New MDisk discovered

```
IBM_2145:ITSO_SVC_SPLIT:superuser>detectmdisk

IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdisk
id name                status         mode    mdisk_grp_id mdisk_grp_name capacity
ctrl_LUN_#      controller_name    UID
tier
0  ITSO_V7K_SITEA_SSD0 online unmanaged 0            V7000SITEA     277.3GB
0000000000000000 600507680289000268000000000000080000000000000000000000000000000000
generic_ssd
1  ITSO_V7K_SITEA_SAS0 online         unmanaged 0            V7000SITEA     500.0GB
000000000000000A 600507680289000268000000000000040000000000000000000000000000000000
generic_hdd
2  ITSO_V7K_SITEA_SAS1 online         unmanaged 0            V7000SITEA     500.0GB
000000000000000B 600507680289000268000000000000050000000000000000000000000000000000
generic_hdd
3  ITSO_V7K_SITEA_SAS2 online         unmanaged 0            V7000SITEA     500.0GB
000000000000000C 600507680289000268000000000000060000000000000000000000000000000000
generic_hdd
4  ITSO_V7K_SITEA_SAS3 online         unmanaged 0            V7000SITEA     500.0GB
000000000000000D 600507680289000268000000000000070000000000000000000000000000000000
generic_hdd
```

15. Add the MDisks to the storage pool by using the SAN Volume Controller CLI commands as shown in Example 6-27. Then, re-create the MDisk to storage pool relationship that existed before the critical event.

Important: Remove the previous MDisks that are still defined in each storage pool but no longer physically exist before you add the newly discovered MDisks. They might be displayed in an offline or degraded state to the SAN Volume Controller.

Example 6-27 Adding new MDisk to Storage pool

```
IBM_2145:ITSO_SVC_SPLIT:superuser>addmdisk -mdisk ITSO_V7K_SITEA_SAS0 V7000SITEA
```

Remember: After you have readded your newly discovered MDisks to the storage pool, the three quorum tandem will be automatically fixed. Check this with the SAN Volume Controller CLI command as shown in Example 6-28 on page 168.

16. Check the Quorum Disk status as shown in Example 6-28.

Example 6-28 Quorum status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name                controller_id controller_name     active
object_type override
0           online 10 ITSO_V7K_SITEC_Q    5            ITSO_V7K_SITEC_Q_N2 yes    mdisk
yes
1           online 6  ITSO_V7K_SITEB_SAS0 0            ITSO_V7K_SITEB_N2   no     mdisk
yes
2           online 4  ITSO_V7K_SITEA_SAS3 1            ITSO_V7K_SITEA_N2   no     mdisk
yes
```

17. Reactivate Volume Mirroring for each volume in accordance with your Volume Mirroring requirements to re-create the same business continuity infrastructure as before the critical event. Use the SAN Volume Controller CLI command that is shown in Example 6-29.

Example 6-29 addvdiskcopy example

```
IBM_2145:ITSO_SVC_SPLIT:superuser>addvdiskcopy -mdiskgrp V7000SITEA
ESXi_Cluster_DATASTORE_A
Vdisk [6] copy [0] successfully created
```

18. Check the status of your Volume Mirroring synchronization progress with the SAN Volume Controller CLI command shown in Example 6-30.

Example 6-30 vdisksyncprogress example

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisksyncprogress
```

You can speed up the synchronization progress with the **chvdisk** command. However, the more speed you give to the synchronization process, the more impact on the overall performance you might have.

19. Consider rebalancing your Split I/O Group configuration to have the Volume Mirroring Primary copy related with the storage pool and preferred node as it was before the critical event. This configuration is useful even if they are now in the same site. Doing that will help you with an eventual future stretch of your configuration when a new remote site becomes available. You can do that using SAN Volume Controller CLI command shown in Example 6-31.

Example 6-31 Change Volume primary copy ID

```
IBM_2145:ITSO_SVC_SPLIT:superuser>chvdisk -primary 0 ESXi_Cluster_DATASTORE_A
```

All your volumes are now accessible from your hosts point of view and the recovery action has not affected your applications.

Some operations have been run by using the CLI, CLI and GUI, or just the GUI to show the different possibilities you have.

Related publications

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

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933
- *Implementing the IBM Storwize V7000 V6.3*, SG24-7938
- *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116
- *Real-time Compression in SAN Volume Controller and Storwize V7000*, REDP-4859
- *IBM System Storage SAN Volume Controller Best Practices and Performance Guidelines*, SG24-7521

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

ibm.com/redbooks

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.storage.ibm.com
- SAN Volume Controller supported platform:

  http://www-1.ibm.com/servers/storage/support/software/sanvc/index.html
- Download site for Windows SSH freeware:

  http://www.chiark.greenend.org.uk/~sgtatham/putty
- IBM site to download SSH for AIX:

  http://oss.software.ibm.com/developerworks/projects/openssh
- 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-1.ibm.com/servers/storage/support/software/sdd/index.html
- IBM TotalStorage Virtualization home page:

  http://www-1.ibm.com/servers/storage/software/virtualization/index.html

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

(0.2”spine)
0.17”<->0.473”
90<->249 pages

IBM SAN and SVC Stretched Cluster and VMware Solution Implementation

Understanding the IBM Stretched Cluster solution and architecture

Implementing Stretched Cluster with VMware

Using diagnosis and recovery guidelines

This IBM Redbooks publication describes the IBM Storage Area Network and IBM SAN Volume Controller Stretched Cluster solution when combined with VMware. We describe guidelines, settings, and implementation steps necessary to achieve a satisfactory implementation.

Business continuity and continuous application availability are among the top requirements for many organizations today. Advances in virtualization, storage, and networking have made enhanced business continuity possible. Information technology solutions can now be designed to manage both planned and unplanned outages, and the flexibility and cost efficiencies available from cloud computing models.

IBM has designed a solution that offers significant functionality for maintaining business continuity in a VMware environment. This functionality provides the capability to dynamically move applications across data centers without interruption to those applications.

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

- The industry-proven VMware Metro vMotion
- IBM System Storage SAN Volume Controller Stretched Cluster solution
- A Layer 2 IP Network and storage networking infrastructure for high performance traffic management
- DC interconnect

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-8072-00 ISBN 0738438138