Redbooks
ibm.com/redbooks

Implementing VersaStack with Cisco ACI Multi-Pod and IBM HyperSwap for High Availability

Jaswinder Singh Saini

Jimmy John

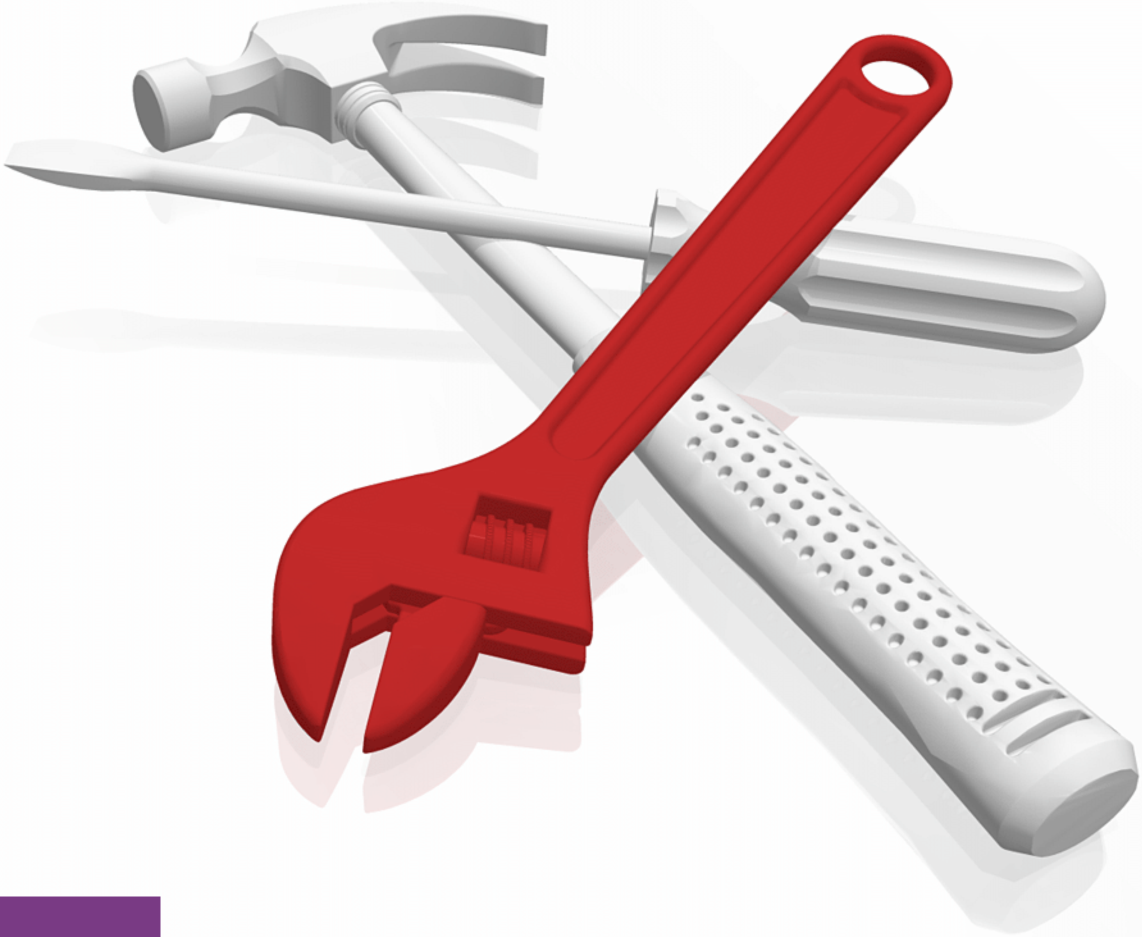
Jordan Fincher

Lee J Cockrell

Nitin D Thorve

Sreeni Edula

Vasfi Gucer



Storage

IBM

Redbooks

International Technical Support Organization

Implementing VersaStack with Cisco ACI Multi-Pod and IBM HyperSwap for High Availability

August 2018

SG24-8423-00

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

First Edition (August 2018)

This edition applies to IBM Spectrum Virtualize Version 7.8.1.4 and the associated hardware and software detailed within.

© Copyright International Business Machines Corporation 2018. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

© Copyright IBM Corp. 2018. All rights reserved.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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, MD-NC119, Armonk, NY 10504-1785, US

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 jurisdictions 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 provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises 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. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

© Copyright IBM Corp. 2018. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®
DS8000®
FlashCopy®
Global Technology Services®
HyperSwap®
IBM FlashSystem®
IBM Spectrum™
IBM Spectrum Storage™
IBM Spectrum Virtualize™
IBM®
Interconnect®
POWER®
Redbooks®
Redbooks (logo) ®
Storwize®
System Storage®
XIV®

The following terms are trademarks of other companies:

ITIL is a Registered Trade Mark of AXELOS Limited.

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

UNIX is a registered trademark of The Open Group in the United States and other countries.

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

Preface

VersaStack, an IBM® and Cisco integrated infrastructure solution, combines computing, networking, and storage into a single integrated system. It combines the Cisco Unified Computing System (Cisco UCS) Integrated Infrastructure with IBM Spectrum™ Virtualize, which includes IBM FlashSystem® 9000 and IBM FlashSystem 5030 storage offerings, for quick deployment and rapid time to value for the implementation of modern infrastructures.

Note on the Storwize rebranding: On 02/11/2020 IBM rebranded IBM Storwize storage systems as IBM FlashSystem, so for example IBM Storwize V5030 is now called IBM FlashSystem 5030. This book been updated to use the new terminology., but you might still see the “Storwize” name in some of the screenshots.

The IBM HyperSwap® high availability (HA) function allows business continuity in a hardware failure, power failure, connectivity failure, or disasters, such as fire or flooding. It is available on the IBM SAN Volume Controller and IBM FlashSystem 7200 products. This IBM Redbooks® publication covers the preferred practices for implementing Cisco VersaStack with IBM HyperSwap. The following are some of the topics covered in this book:

- Cisco Application Centric Infrastructure to showcase Cisco’s ACI with Nexus 9Ks
- Cisco Fabric Interconnects and Unified Computing System (UCS) management capabilities
- Cisco Multilayer Director Switch (MDS) to showcase fabric channel connectivity
- Overall IBM HyperSwap solution architecture
- Differences between HyperSwap and Metro Mirroring, Volume Mirroring, and Stretch Cluster
- Multisite IBM SAN Volume Controller (SVC) deployment to showcase HyperSwap configuration and capabilities

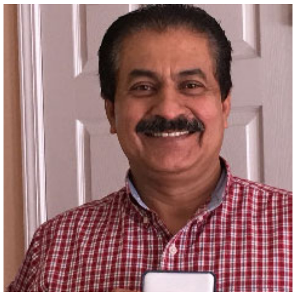
This book is intended for pre-sales and post-sales technical support professionals and storage administrators who are tasked with deploying a VersaStack solution with IBM HyperSwap.

Authors

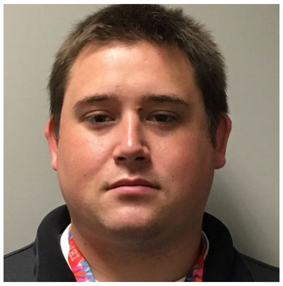
This book was produced by a team of specialists from around the world working at the Cisco Raleigh Center.

Jaswinder Singh Saini is a Storage Solution Architect working with Systems Lab services team ISA. He has been with IBM since 2012 and has been working on Accelerate, Scale, and IBM FlashSystem. He has been handling design, implementation, migrations, consolidations, POC, IT assessments, and optimization projects for leading banking and telecom clients in India.

© Copyright IBM Corp. 2018. All rights reserved.



Jimmy John is an IBM System Storage® Technical Advisor and SME for Spectrum Virtualize, SVC, and FlashSystem products. He worked as Advisory Software Engineer/PFE for SVC and Product Focal for FlashSystem platform for over 6 years. His past assignment includes being an Operational Team Lead and PFE for IBM Modular Storage Products. He has been with IBM for 18 years at various positions. His current interests are Cloud Implementation and Analytics.



Jordan Fincher is a Product Field Engineer working in Storage Support at IBM. He received his Bachelor of Science in Information Security from Western Governor University. Jordan first started his IBM career in 2012 as a Systems Engineer for the IBM Business Partner e-TechServices doing pre-sales consulting and implementation work for many IBM accounts in Florida. In 2015, Jordan started working in his current role as a Product Field Engineer for IBM Spectrum Virtualize™ storage products.

Lee J Cockrell is an IBM Technical Sales Specialist, covering several United States Federal Civilian agencies and Native American Tribal governments for IBM Federal in the Washington D.C. area. He received his B.S. in computer science from the University of Virginia, Charlottesville, VA. His current interests are in storage, cloud, and computer security. Lee joined IBM in 2010 selling storage, primarily IBM Spectrum Virtualize, SAN Volume Controller, IBM FlashSystem, IBM XIV®, and IBM DS8000®. He has worked in the storage industry since 2001, installing and configuring countless storage arrays and co-authoring expert level performance certification tests. Previously, he was a UNIX and firewall administrator in both the public and private sectors.

Nitin D Throve is an IBM Information Technology Infrastructure Architect, working with the Solution Design and Architecture Services team at IBM India in the Pune area. He received his Bachelors degree in computer engineering from MIT Academy of Engineering, Pune. His current interests are in storage, cloud, analytics, AI, and cognitive computing technologies. Nitin joined IBM in 2016 as a storage SME with skills in the IBM Spectrum Storage™ family, IBM Flash Storage family, IBM FlashSystem family, IBM Disk Systems DS8000 series, IBM replication technologies, and IBM POWER® hardware, including AIX®, VIOS, and Power virtualization, EMC Storage products, Hitachi Storage products, Netapp NAS products, Cisco, and Brocade SAN Hardware in IBM Pune India. In 2017, Nitin moved to the IBM India Solution Design and Architecture Services team as a Technical Solutions Manager, helping the IBM Global Technology Services® Solutioning team develop new logo solutions for the EMEA region. Nitin is currently assigned with a new role to support GTS Global account as an Infrastructure Architect to work on World Wide Storage Refresh and Transformation Strategy.

Sreeni Edula is a Technical Marketing Engineer in the UCS Data Center Solutions Engineering team focusing on converged and hyper-converged infrastructure solutions, prior to that he worked as a Solutions Architect at EMC Corporation. He has experience in Information Systems with expertise across Cisco Data Center technology portfolio, including DC architecture design, virtualization, compute, network, storage and cloud computing.

Vasfi Gucer is an IBM Technical Content Services Project Leader with the Digital Services Group. He has more than 20 years of experience in the areas of systems management, networking hardware, and software. He writes extensively and teaches IBM classes worldwide about IBM products. His focus has been primarily on cloud computing, including cloud storage technologies for the last 6 years. Vasfi is also an IBM Certified Senior IT Specialist, Project Management Professional (PMP), IT Infrastructure Library (ITIL) V2 Manager, and ITIL V3 Expert.

Thanks to the following people for their contributions to this project:

Ernest Keenan, Jon Tate, Erica Wazewski
Digital Services Group, Technical Content Services

Nolan Rogers, Warren Hawkins,
IBM UK

Chris O'Brien, Jawwad Memon
Cisco, Raleigh Center

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 and HyperSwap solution architecture

This chapter provides an introduction to the HyperSwap solution described in this book. The following topics are covered in this chapter:

- The business case for HyperSwap
- IBM Spectrum Virtualize and HyperSwap architecture
- Differences in HyperSwap and Enhanced Stretch Cluster
- Differences in Spectrum Virtualize copy services
- Introduction to VersaStack

© Copyright IBM Corp. 2018. All rights reserved.

1.1 The business case for HyperSwap

The IBM Spectrum Virtualize HyperSwap function is a high availability feature that provides active-active access to a volume at two sites up to 300 km apart. HyperSwap functions are available on Spectrum Virtualize systems that can support more than one I/O group.

High availability is a subset of business continuity practices. Enterprises plan for business continuity to anticipate, mitigate, and prevent disruption of their applications. Enterprises that require specific applications to function a high percentage of the time implement high availability on those most critical applications. High availability designs usually involve redundant hardware, and software capable of transferring application processes from host to host, array to array, or even site to site.

HyperSwap volumes maintain a fully independent copy of the volume data at each site. When data is written by hosts at either site, both copies are synchronously updated before the write operation is completed. If one site is no longer available, the other site continues to provide access to the volume.

Figure 1-1 shows an example FlashSystem HyperSwap configuration with underlying Volume Mirrors for additional redundancy.

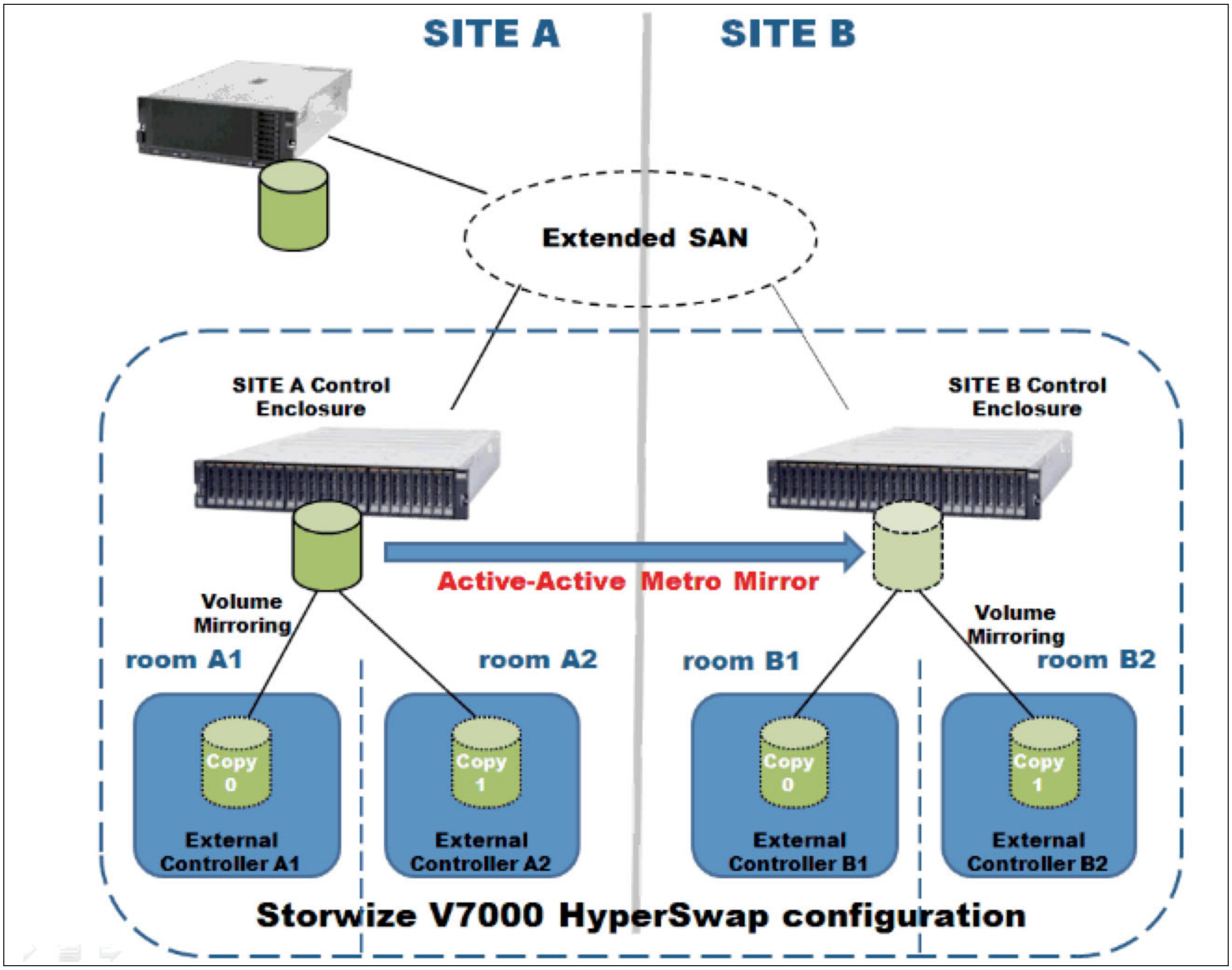


Figure 1-1 FlashSystem HyperSwap configuration

1.2 IBM Spectrum Virtualize and HyperSwap architecture

IBM Spectrum Virtualize is a software product that virtualizes storage arrays. It provides advanced features such as remote copy replication, volume copies and migration, snapshots, thin provisioning, and data duplication.

For more information, see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933.

The HyperSwap function builds on two existing Spectrum Virtualize functions: Non-Disruptive Volume Move (NDVM) and the Remote Copy features including Metro Mirror, Global Mirror, and Global Mirror with Change Volumes.

In a HyperSwap configuration, each site is an independent failure domain. If one site experiences a failure, then the other site can continue to operate without disruption. You must also configure a third site to host a quorum device or IP quorum application that provides an automatic tie-break during a link failure between the two main sites. The main site can be in the same room or across rooms in the data center, buildings on the same campus, or buildings in different cities. Different kinds of sites protect against different types of failures.

Figure 1-2 shows an example of HyperSwap used with application cluster software to ensure high availability during a total site outage.

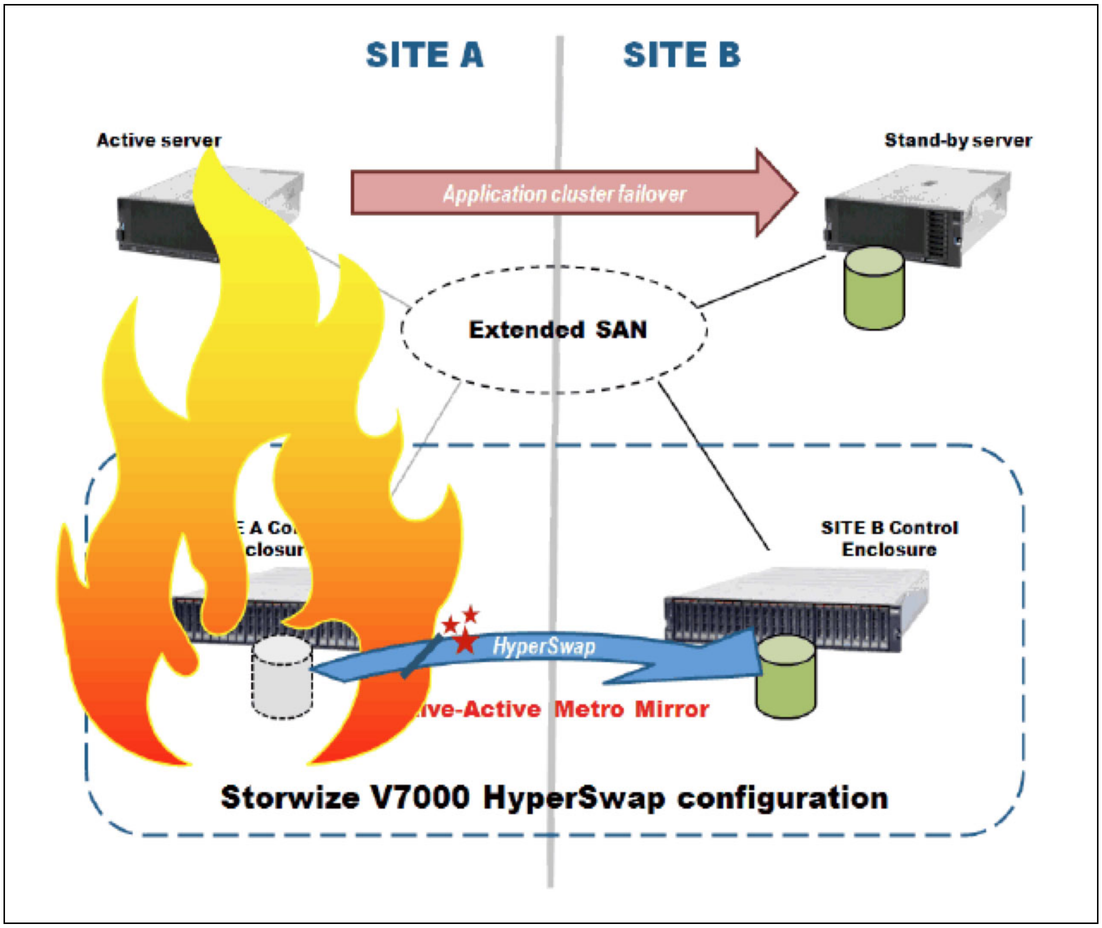


Figure 1-2 HyperSwap used with an application cluster software to ensure high availability

When the system topology is set to hyperswap, each node, controller, and host in the system configuration must have a site attribute set to 1 or 2. Both nodes of an I/O group must be at the same site. This site must be the same site as the controllers that provide the managed disks to that I/O group. When managed disks are added to storage pools, their site attributes must match. This requirement ensures that each copy in a HyperSwap volume is fully independent and is at a distinct site.

HyperSwap volumes create an active-active relationship between a copy of the volume at each site. Each volume is *active* in that it can receive and respond to I/O from the host. These relationships automatically run and switch replication direction according to which copy or copies are online and up-to-date. The relationships provide access to whichever copy is up-to-date through a single volume, which has a unique ID. Relationships can be grouped into consistency groups just like Metro Mirror and Global Mirror relationships. The consistency groups fail over consistently as a group based on the state of all copies in the group. A consistent copy of data that can be used for disaster recovery is maintained at each site.

The volume must be accessible to one or more hosts through either I/O group. The synchronizing process starts after change volumes are added to the active-active relationship.

The Small Computer System Interface (SCSI) protocol allows storage devices to indicate the preferred ports for hosts to use when they submit I/O requests. Using the Asymmetric Logical Unit Access (ALUA) state for a volume, a storage controller can inform the host of which paths are active and which ones are preferred. In a HyperSwap system topology, the nodes advertise the preferred path from a node at the same site as the host. A local node is a node that is configured at the same site as the host.

The HyperSwap function in the SAN Volume Controller (SVC) software works with the standard multipathing drivers that are available on a wide variety of host types, with no additional host support required to access the highly available volume. Where multipathing drivers support ALUA, the storage system tells the multipathing driver which nodes are closest to it and should be used to minimize I/O latency. The HyperSwap function automatically optimizes itself to minimize data transmitted between sites and to minimize host read and write latency.

1.3 Differences in HyperSwap and Enhanced Stretch Cluster

Although the capabilities of HyperSwap are similar to those of Enhanced Stretch Cluster, they use different underlying technologies to obtain those similar results. Both functions spread the nodes of the system across two sites, with storage at a third site acting as a tie breaking quorum device. Not all the environment configuration variations available with Enhanced Stretch Cluster are necessarily available for HyperSwap. Enhanced Stretch Cluster configurations consist of each I/O group being split between two sites, while both HyperSwap I/O group nodes are at the same site. Because FlashSystem node canisters cannot be split from the same enclosure, Enhanced Stretch Cluster is only available on SAN Volume Controller and v9000.

1.4 Differences in Spectrum Virtualize copy services

High availability and disaster recovery are strategies to reduce the likelihood of application outages and data loss. High availability focuses on keeping applications available in real time, whereas disaster recovery provides a way to recover from array, site, or even regional outages. Disaster recovery usually includes a secondary data copy available at a large distance from the primary.

HyperSwap is a high availability solution that uses Metro Mirror. Volume Mirror is a simple solution that keeps two copies of a volume in a system. Global Mirror is a disaster recovery solution that keeps a consistent point-in-time copy of data replicated at a remote site to decrease the possibility of a disaster affecting both sites.

This section covers the different copy services offered by Spectrum Virtualize.

1.4.1 Remote Copy

Metro Mirror and Global Mirror are two types of remote-copy operations that enable you to set up a relationship between two volumes, where updates made to one volume are mirrored on the other volume. Both support intracluster copying of a volume, in which both volumes belong to the same system and I/O group within the system, or intercluster copying of a volume, in which one volume belongs to one cluster and the other volume belongs to a different cluster. Metro Mirror replicates synchronously, whereas Global Mirror replicates asynchronously.

Metro Mirror

Metro Mirror creates a synchronous data copy from a primary volume to a secondary volume. The secondary volume can either be on the same system or on another system. With synchronous copies, hosts write to the primary volume, but the write operation is not completed until the data is confirmed written to the secondary volume. This process ensures that both the volumes have identical data when the copy operation completes. After the initial copy operation completes, the Metro Mirror function maintains a fully synchronized copy of the source data at the target site.

The write I/O flow using Metro Mirror is illustrated in Figure 1-3.

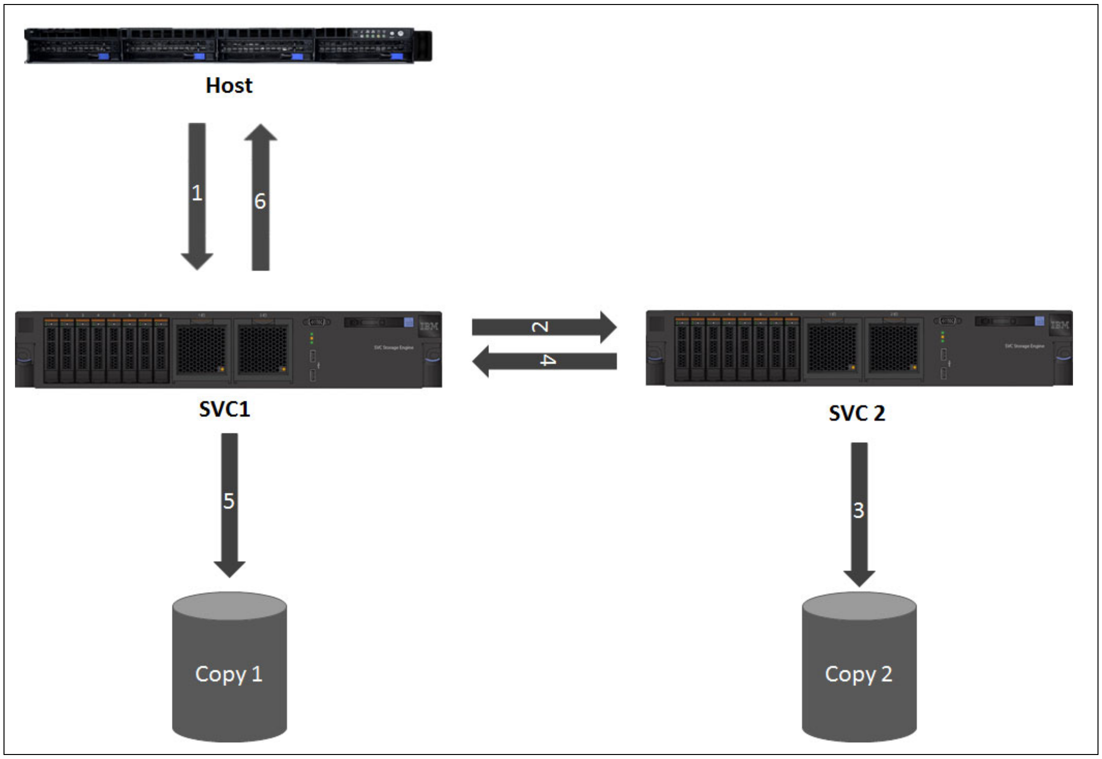


Figure 1-3 Write data flow of Metro Mirror

Global Mirror

Global Mirror creates an asynchronous data copy from a primary volume to a secondary volume. That is, the secondary volume might not have the most recent data that the primary does. The secondary usually does not catch up to the primary unless host write I/O is quiesced.

The advantage of Global Mirror is that it allows for a copy of the data at much larger distances (greater than 300 km) and latencies than Metro Mirror, with minimal performance degradation of the volume presented to the host. Global Mirror can tolerate round-trip latencies of up to 250 ms over Fibre Channel, 80 ms at 1 Gbps IP, andS 10 ms at 10 Gbps IP. At an approximately 100 km distance per ms, that would be 25,000 km, 8,000 km, or 1,000 km. The disadvantage of Global Mirror is its asynchronous design. The secondary copy is not guaranteed to be up-to-date with every change to the primary copy.

When Global Mirror operates without cycling, write operations are applied to the secondary volume as soon as possible after they are applied to the primary volume. The secondary volume is generally less than one second behind the primary volume, which minimizes the amount of data that must be recovered after a failover. However, this technique requires that a high-bandwidth link be provisioned between the two sites.

Functionally speaking, Global Mirror sends the same write I/O simultaneously to both the primary and secondary copies. However, it does not wait for the secondary to confirm the I/O, and does not guarantee a consistent secondary copy. This I/O flow is illustrated in Figure 1-4.

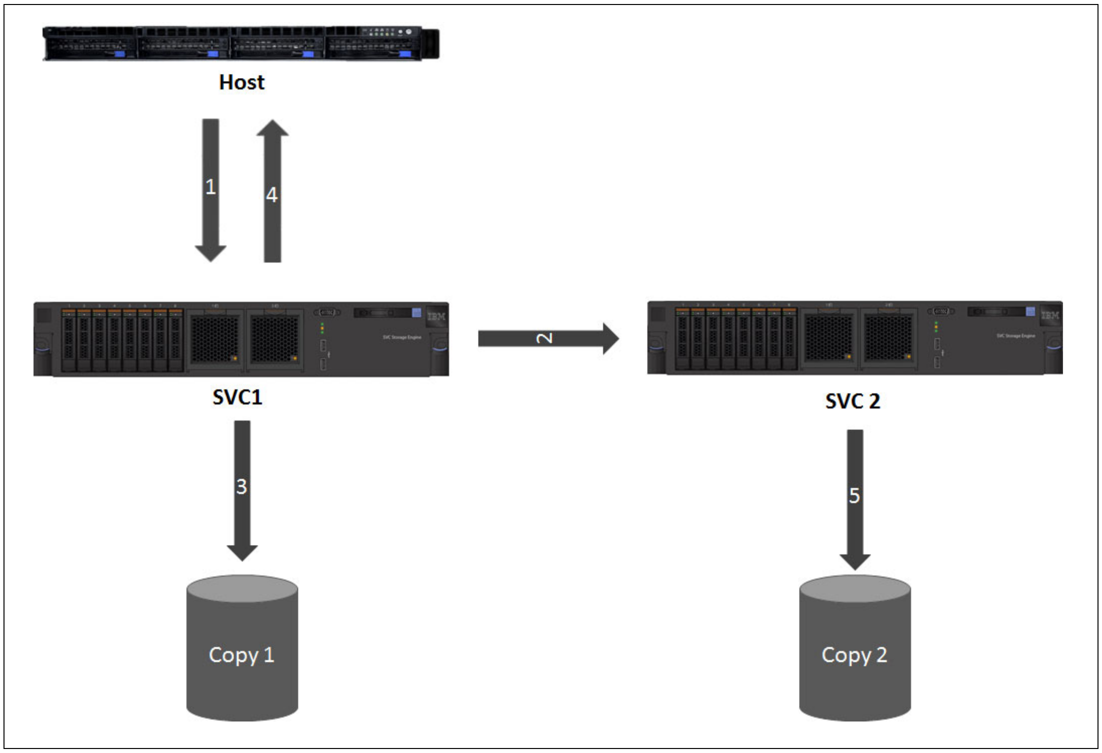


Figure 1-4 Data flow of Global Mirror

Global Mirror with Change Volumes

Global Mirror with Change Volumes allows for a consistent point-in-time volume copy to be synchronized to the secondary copy. Instead of replicating directly from the master to the auxiliary volume, a flashcopy snapshot of the master volume is taken periodically (also known as “cycling”) and this snapshot serves as the source of the replication to the secondary cluster. The tradeoff to this is that the data on the secondary site is typically seconds to minutes behind the primary copy.

1.4.2 HyperSwap

Spectrum Virtualize includes HyperSwap technology to provide a host transparent, high availability solution between two locations up to 300 km apart. HyperSwap allows a host to access a volume on multiple I/O groups at two sites, with failover between storage systems being automatic and transparent. Before the introduction of HyperSwap, the Spectrum Virtualize solutions for disaster recovery and high availability were Metro Mirror, Global Mirror, and Volume Mirror.

1.4.3 Volume Mirroring

Volume Mirroring provides data availability in a failure of internal or external storage. It creates two copies of the data within the cluster to present to the host as a single volume. However, the Volume Mirroring feature does not provide any protection in a loss of the storage system's control enclosures. To protect data against the complete loss of storage systems, host-based data availability solutions can be implemented. These solutions rely on a combination of storage system and application or operating system capabilities. They usually delegate the management of the storage loss events to the host.

The advantage of Volume Mirroring is that it provides lower latency than HyperSwap. In HyperSwap, the latency of the intersite link adds to the latency of the I/O because a write I/O is not reported to the host as complete until it is written to both sites. The main advantage of HyperSwap over Volume Mirroring is that you can maintain a synchronous copy over a longer distance. Some customers go between sites in the same city, some within the same data center for lower latency.

The data flow of volume mirroring is illustrated in Figure 1-5.

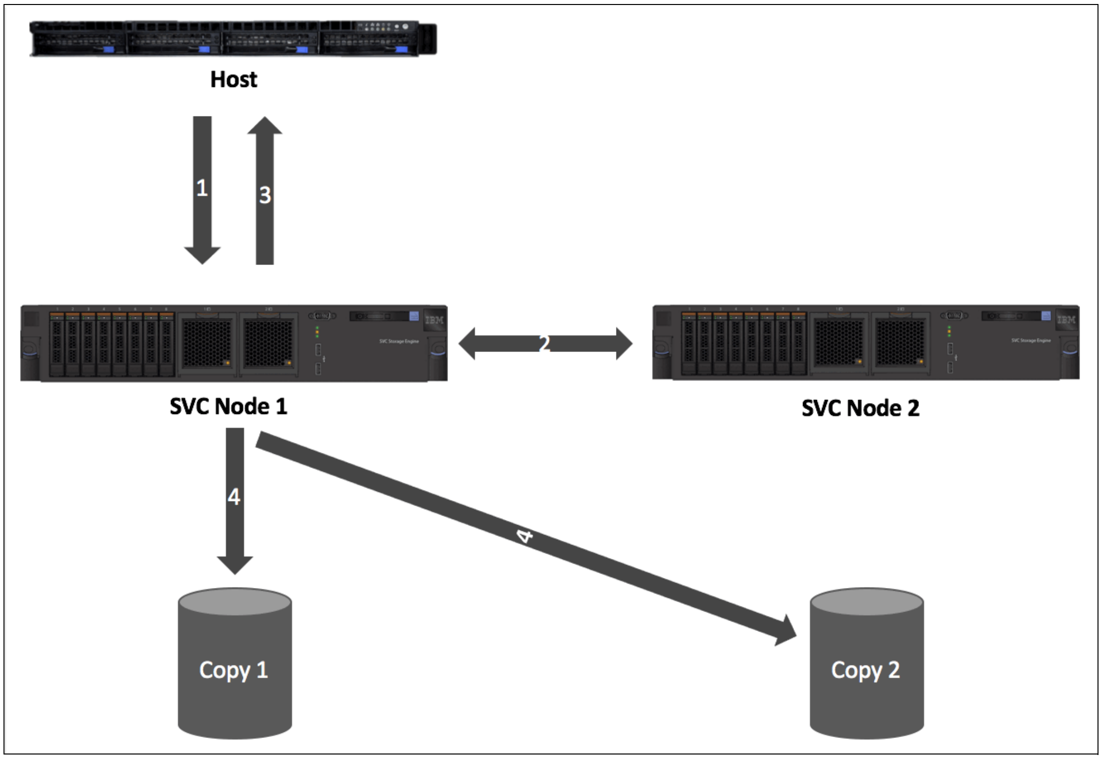


Figure 1-5 Data flow of Volume Mirroring

1.5 Introduction to VersaStack

The VersaStack solution is a pre-designed, integrated, and validated architecture for the data center. It combines Cisco UCS servers, Cisco Nexus family of switches, and Cisco MDS fabric switches; and IBM SVC, FlashSystem, and FlashSystem storage arrays into a single, flexible architecture. VersaStack is designed for high availability, with no single points of failure, while maintaining cost-effectiveness and flexibility in design to support a wide variety of workloads.

VersaStack design can support different hypervisor options and bare metal servers, and can also be sized and optimized based on customer workload requirements. The VersaStack design discussed in this document has been validated for resiliency (under fair load) and fault tolerance during system upgrades, component failures, and partial and complete loss of power scenarios.

1.5.1 Architecture

VersaStack with Cisco UCS M5 and IBM SVC architecture aligns with the converged infrastructure configurations and best practices as identified in the previous VersaStack releases. The system includes hardware and software compatibility support between all components and aligns to the configuration best practices for each of these components. All the core hardware components and software releases are listed and supported on both the Cisco compatibility list and the IBM Interoperability Matrix.

The system supports high availability at the network, compute, and storage layers such that no single point of failure exists in the design. The system uses 10 and 40 Gbps Ethernet jumbo-frame based connectivity combined with port aggregation technologies such as virtual PortChannel (vPC) for non-blocking LAN traffic forwarding. A dual SAN 16 Gbps environment provides redundant storage access from compute devices to the storage controllers.

1.5.2 Introducing Cisco ACI

Cisco ACI is a data center architecture designed to address the requirements of today's traditional networks, and to meet emerging demands that new computing trends and business factors are placing on the network. Software-defined networking (SDN) has garnered much attention in the networking industry over the past few years due to its promise of a more agile and programmable network infrastructure. Cisco ACI not only addresses the challenges of agility and network programmability that software-based overlay networks are trying to address, but it also presents solutions to the new challenges that SDN technologies are currently unable to address.

Cisco ACI uses a network fabric that employs industry-proven protocols combined with innovative technologies to create a flexible, scalable, and highly available architecture of low-latency, high-bandwidth links. This fabric delivers application instantiations using profiles that house the required characteristics to enable end-to-end connectivity. The ACI fabric is designed to support the management automation, programmatic policies, and dynamic workload provisioning. The ACI fabric accomplishes this objective with a combination of hardware, policy-based control systems, and closely coupled software to provide advantages not possible in other vendor solutions.

The Cisco ACI fabric consists of three major components:

- The Application Policy Infrastructure Controller (APIC)
- Spine switches
- Leaf switches

The ACI switching architecture is presented in a leaf-and-spine topology where every leaf connects to every spine by using 40G Ethernet interfaces. The ACI Fabric Architecture is outlined in Figure 1-6.

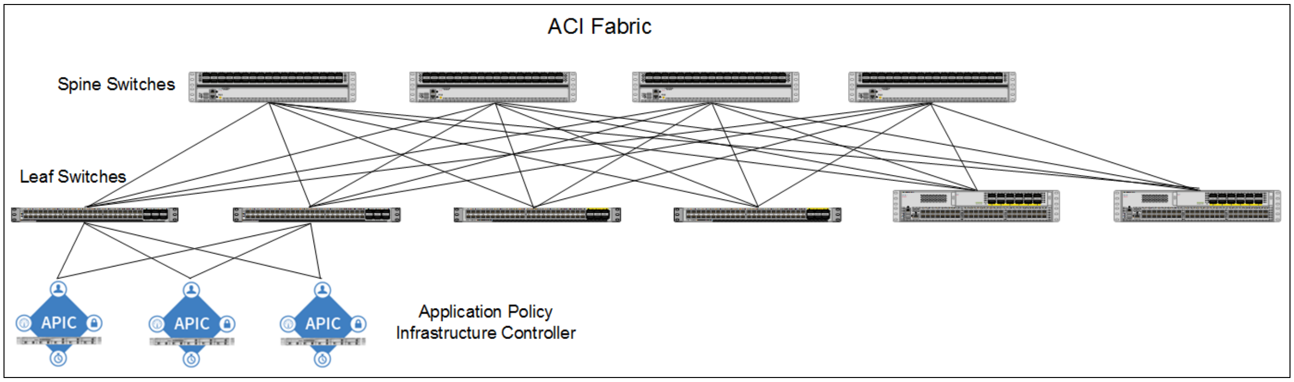


Figure 1-6   Cisco ACI switching architecture

The software controller, APIC, is delivered as an appliance, and three or more such appliances form a cluster for high availability and enhanced performance. APIC is responsible for all tasks enabling traffic transport, which include the following:

- Fabric activation
- Switch firmware management
- Network policy configuration and installation

Although the APIC acts as the centralized point of configuration for policy and network connectivity, it is never in line with the data path or the forwarding topology. The fabric can still forward traffic even when communication with the APIC is lost.

APIC provides both a command-line interface (CLI) and graphical user interface (GUI) to configure and control the ACI fabric. APIC also exposes a northbound API through XML and JavaScript Object Notation (JSON), and an open source southbound API.

1.5.3 Overview of ACI Multi-Pod

As shown in Figure 1-7, ACI Multi-Pod represents the natural evolution of the original ACI Stretched Fabric design. It allows you to interconnect and centrally manage separate ACI fabrics.

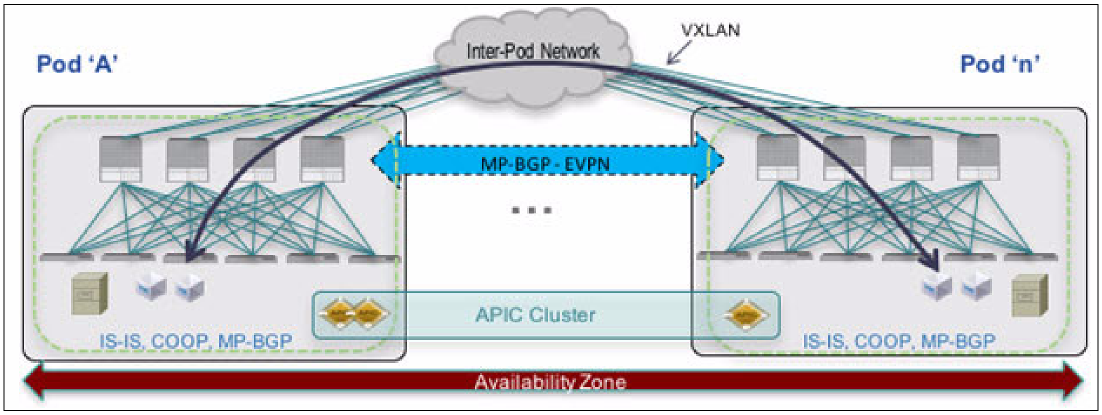


Figure 1-7   ACI Multi-Pod

ACI Multi-Pod is part of the *Single APIC Cluster/Single Domain* family of solutions. A single APIC cluster is deployed to manage all the different ACI fabrics that are interconnected. Those separate ACI fabrics are named *Pods*, and each of them looks like a regular two-tier spine-leaf fabric. The same APIC cluster can manage several Pods. To increase the resiliency of the solution, the various controller nodes that make up the cluster can be deployed across different Pods.

The deployment of a single APIC cluster simplifies the management and operational aspects of the solution because all the interconnected Pods essentially function as a single ACI fabric. Created tenents configuration (such as Virtual Route Forwardings (VRFs), Bridge Domains, and Endpoint Groups (EPGs)) and policies are made available across all the Pods, providing a high degree of freedom for connecting endpoints to the fabric. For example, different workloads that are part of the same functional group (EPG), like web servers, can be connected to (or moved across) different Pods without having to worry about provisioning configuration or policy in the new location.

At the same time, seamless Layer 2 and Layer 3 connectivity services can be provided between endpoints independently from the physical location where they are connected and without requiring any specific functionality from the network that connects the various Pods.

From a physical perspective, the different Pods are interconnected by using an Inter-Pod Network (IPN). Each Pod connects to the IPN through the spine nodes. The IPN can be as simple as a single Layer 3 device, or can be built with a larger Layer 3 network infrastructure, as will be clarified in the following sections.

2

Testing environment overview

This chapter describes the testing environment used in the making of this publication. The intention of this chapter is to serve as a reference of what we tested.

The following topics are covered in this chapter:

- Solution overview
- Cisco UCS and ACI configuration
- Cisco UCS and ACI configuration
- Virtualization design

© Copyright IBM Corp. 2018. All rights reserved.

2.1 Solution overview

For the purposes of testing HyperSwap in VersaStack, we created a full pseudo-production environment to test in. This environment consisted of the following infrastructure:

- Cisco Unified Computing System (UCS) with B-Series servers
- Cisco MDS Fibre Channel (FC) Switches
- Cisco Nexus Switches
- Cisco Application Centric Infrastructure (ACI)
- Cisco Application Policy Infrastructure Controller (APIC)
- IBM SAN Volume Controller (SVC)
- IBM FlashSystem 900 AE2
- IBM FlashSystem 5030

The overall implementation of a single site is illustrated in Figure 2-1.

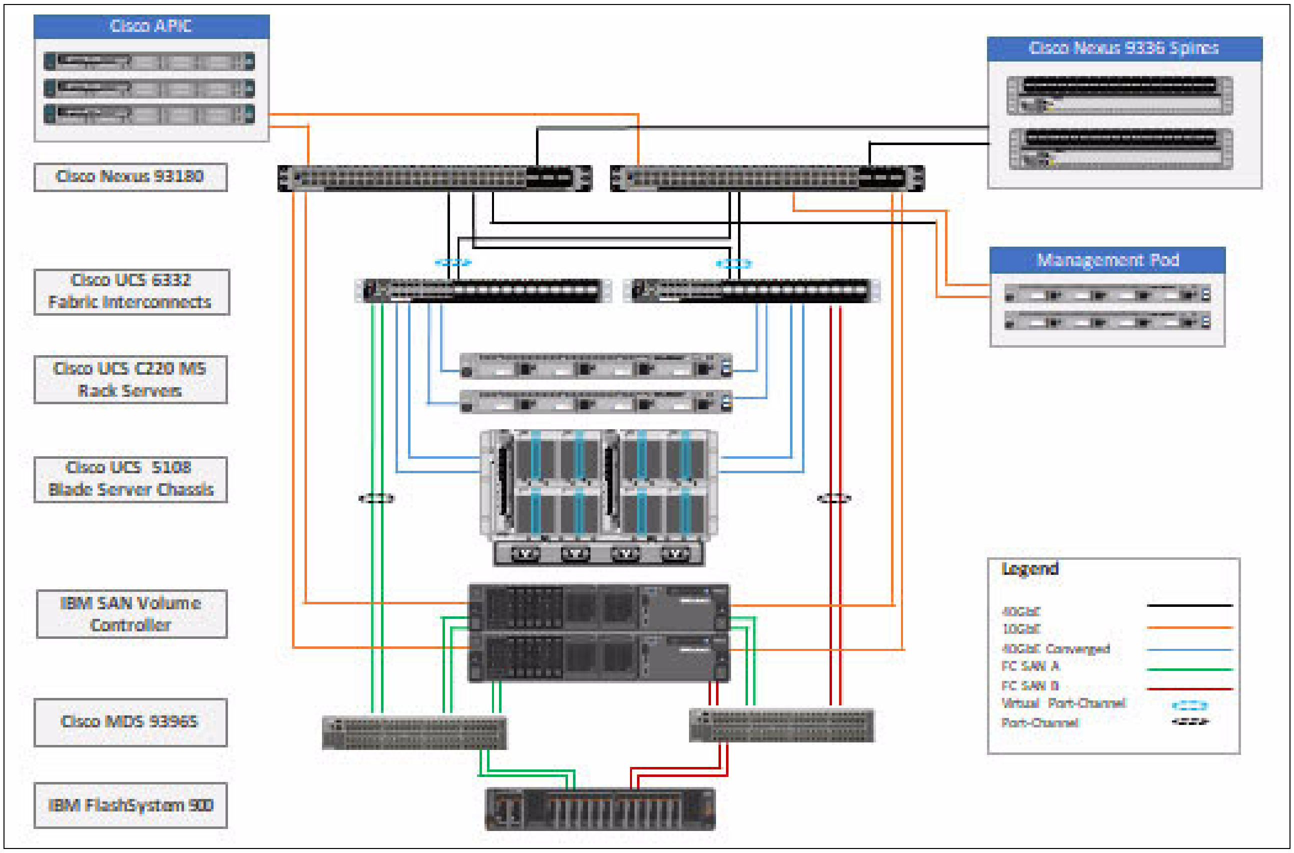


Figure 2-1 Single site topology

2.2 VersaStack with Cisco ACI Multi-Pod architecture

The Cisco ACI Multi-Pod design enables VersaStack to support multiple locations. It is further enabled by IBM HyperSwap, allowing the single site VersaStack data center architecture to be stretched across two physical locations.

The high-level VersaStack with Cisco ACI Multi-Pod and IBM HyperSwap architecture is depicted in Figure 2-2.

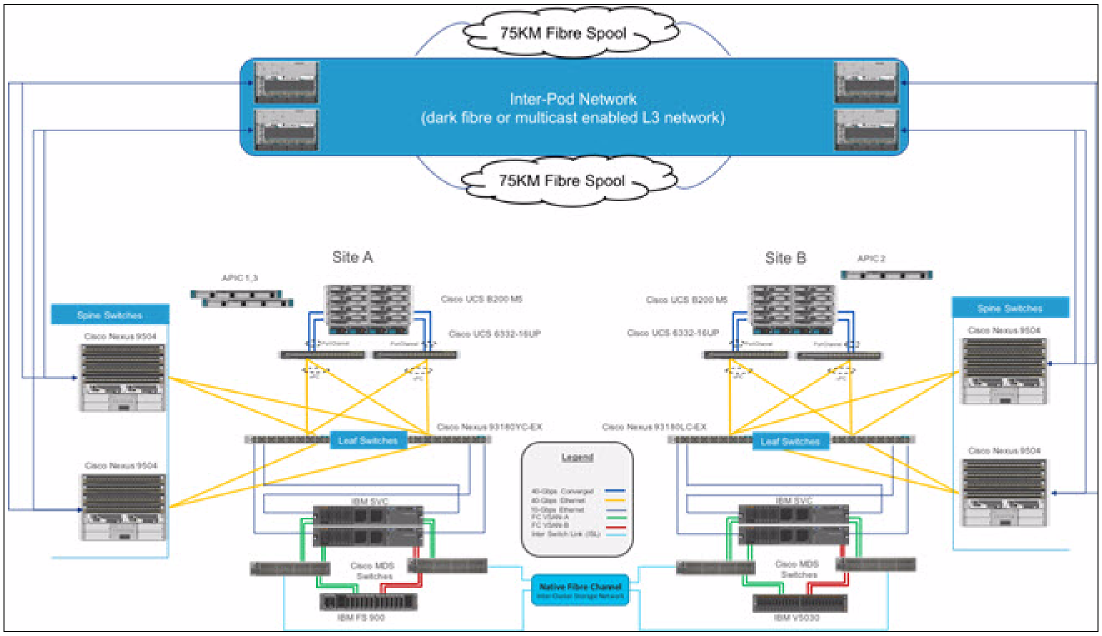


Figure 2-2 VersaStack with Cisco ACI Multi-Pod and IBM HyperSwap architecture

The following sections cover physical and logical connectivity details across the stack including various design choices at the compute, storage, virtualization, and network layers.

2.2.1 Inter-site SAN connectivity

For Inter-site SAN connectivity, IBM SVC cluster nodes across the two sites use native Fibre Channel connection through the Cisco MDS switches. In each Cisco MDS 9k switch, a storage VSAN is created and the local SVC controller connects to both the Cisco MDS 9000 switches, as shown in Figure 2-3.

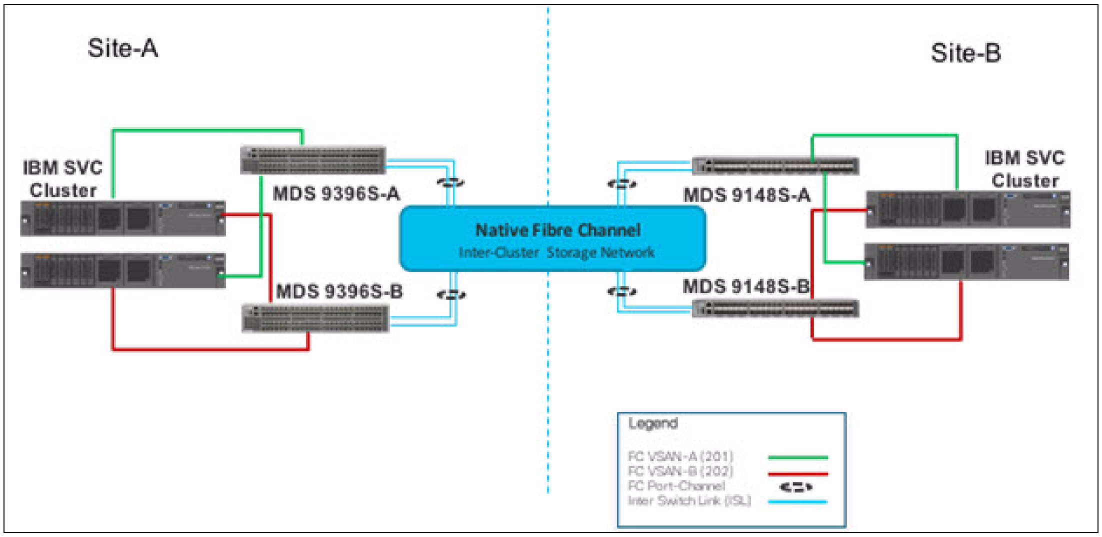


Figure 2-3 Inter-site SAN connectivity

At each of the sites, Cisco MDS 9000-A acts as the SAN-A switch and Cisco MDS 9000-B acts as the SAN-B switch. To provide FC connectivity between the sites, Fibre Channel inter-switch link (ISL) connections are configured between the Cisco MDS 9000 storage VSANs.

As shown in Figure 2-3 on page 15, two redundant paths are configured between the two sites for SAN resiliency. The ports between the Cisco MDS 9000 switches are configured as Port-Channels to aggregate bandwidth and support resiliency. By using this configuration, each IBM SVC node has access to both the IBM SVC nodes at the other site. There are multiple ways to configure ISLs to enable Fibre Channel connectivity across two distinct data centers. In this case, we have stretched the VSANs across the sites and used multiple ISLs that are aggregated into port channels functioning as one logical link.

2.3 Cisco UCS and ACI configuration

This section highlights the Cisco UCS configuration and the Cisco ACI setup.

2.3.1 Cisco UCS LAN connectivity

In the VersaStack compute design, each Cisco UCS 5108 chassis is connected to the Fabric Interconnect® (FIs) using a pair of ports from each IO Module for a combined 40G uplink, as shown in Figure 2-4.

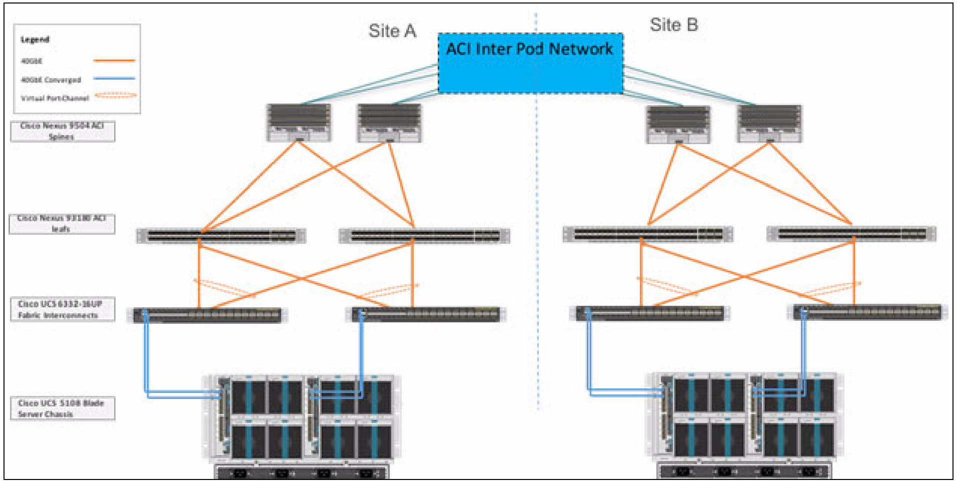

Legend
40GbE
40GbE Converged
Virtual Port-Channel
Site A
ACI Inter Pod Network
Site B
Cisco Nexus 9504 ACI Spines
Cisco Nexus 93180 ACI leafs
Cisco UCS 6332-16UP Fabric Interconnects
Cisco UCS 5108 Blade Server Chassis

Figure 2-4   UCS LAN topology

The Cisco UCS Fabric Interconnects are configured with two port-channels, one from each FI, to the Cisco Nexus 93180 leaf switches. These port-channels carry all the data and IP-based storage traffic that originated on the Cisco Unified Computing System. Virtual Port-Channels (vPC) are configured on the Cisco Nexus 93180 to provide device level redundancy.

The validated design used two uplinks from each FI to the leaf switches for an aggregate bandwidth of 160 GbE (4 x 40 GbE). The number of links can be increased based on customer data throughput requirements.

2.3.2 Cisco ACI and APIC configuration

ACI Multi-Pod solution allows interconnecting and centrally managing ACI fabrics deployed in separate, geographically dispersed data centers. In an ACI Multi-Pod solution, a single APIC cluster is deployed to manage all of the different ACI fabrics that are interconnected using an Inter-Pod Network (IPN), as shown in Figure 2-5.

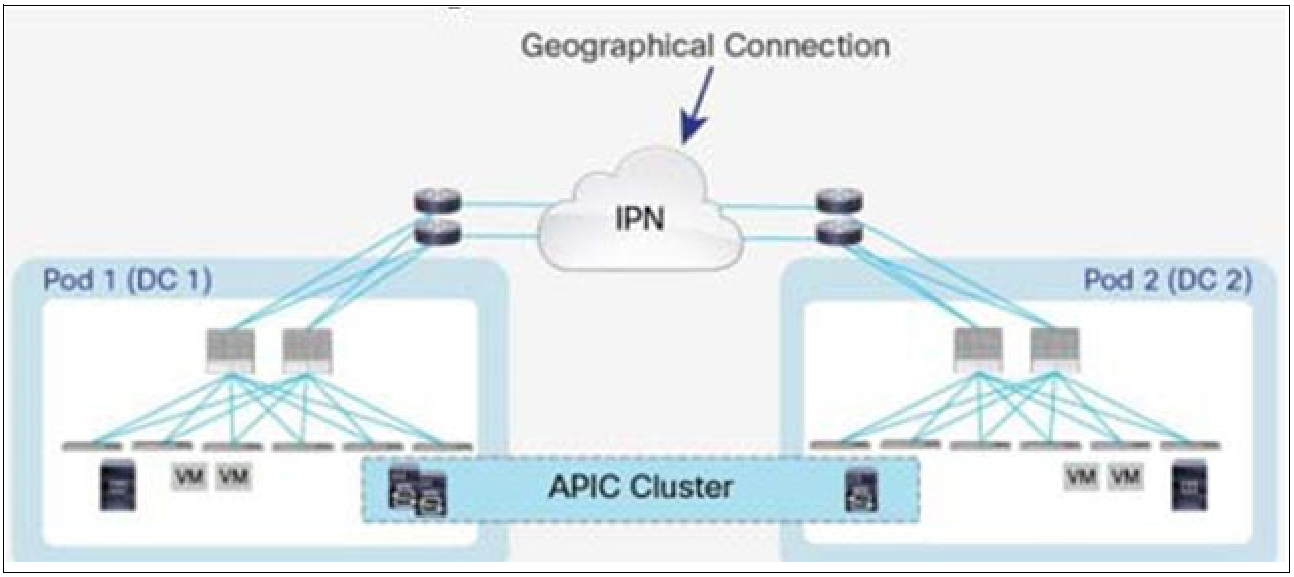


Figure 2-5 Implemented ACI design

The separate ACI fabrics are named *Pods*, and each of the Pods looks like a regular two-tier spine-leaf fabric. A single APIC cluster can manage multiple Pods and various controller nodes that make up the cluster can be deployed across these Pods for resiliency. The deployment of Multi-Pod, as shown, meets the requirement of building Active/Active Data Centers, where different application components can be deployed across Pods.

2.3.3 Inter-Pod network configuration

The Inter-Pod Network consists of two Nexus 7004 switches in each data center that are connected by using a 10 Gbps 75 km long fiber connection. The dual link design provides high availability in a link failure. Each spine is connected to each of the Nexus 7004s using a 10 Gbps connection for a fully redundant setup.

Each Nexus 7004 is configured for the following features to support the Multi-Pod network.

PIM Bidir configuration

In addition to unicast communication, Layer 2 multi-destination flows belonging to bridge domains that are extended across Pods must also be supported. This type of traffic is usually referred to as Broadcast, Unknown Unicast, and Multicast (BUM) traffic, and is exchanged by using Virtual Extensible LAN (VXLAN) data plane encapsulation between leaf nodes. Inside a Pod (or ACI fabric), BUM traffic is encapsulated into a VXLAN multicast frame, and it is always transmitted to all the local leaf nodes. To flood the BUM traffic across Pods, the same multicast that is used inside the Pod is also extended through the IPN network. PIM Bidir enables this functionality on the IPN devices.

OSPF configuration

OSPF is enabled on Spine switches and IPN devices to exchange routing information between the Spine switches and IPN devices.

DHCP relay configuration

To support auto-provisioning of configuration for all the ACI devices across multiple Pods, the IPN devices connected to the spines must be able to relay DHCP requests generated from ACI devices in remote Pods toward one or more APIC nodes active in the first Pod.

Interface VLAN encapsulation

The IPN device interfaces connecting to the ACI Spines are configured as sun-interfaces with the VLAN encapsulation value set to 4.

MTU configuration

The IPN devices are configured for a maximum supported MTU value of 9216 to handle the VXLAN overhead.

TEP pools and interfaces

In Cisco ACI Multi-Pod setup, unique Tunnel Endpoint (TEP) pools are defined on each site. The Pod connection profile uses a VXLAN TEP (VTEP) address called the External TEP (ETEP) as the anycast shared address across all spine switches in a Pod. This IP address should not be part of the TEP pool assigned to each Pod, and is therefore selected outside the two networks listed above.

2.4 IBM Storage overview

The SVC cluster in this environment is configured as a single cluster with a dual-site HyperSwap topology. In this configuration, one site is marked as primary and the other as secondary. The FlashSystem FS900 is on the primary site and the FlashSystem 5030 is on the secondary site. All connectivity including connectivity between sites flows through a standard dual fabric storage area network (SAN) consisting of Cisco MDS switches. Each SAN fabric has a single switch in each site. This configuration is illustrated in Figure 2-6.

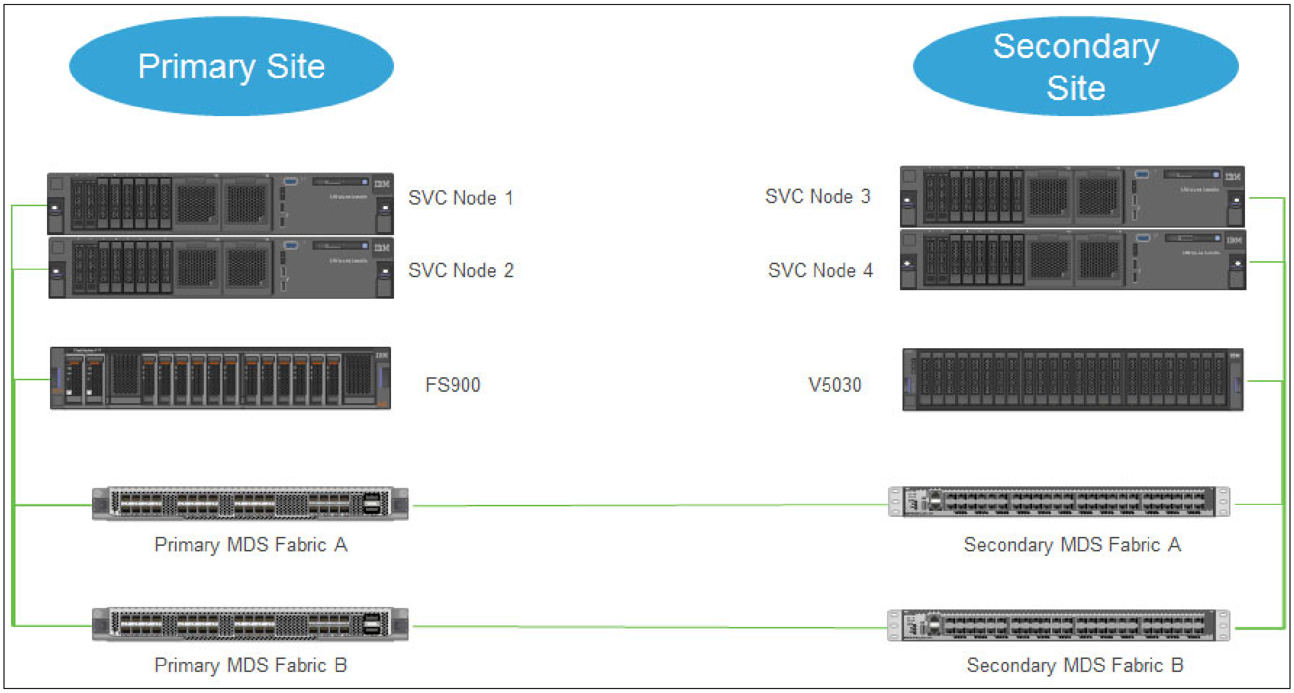


Figure 2-6 SAN topology

2.4.1 Back-end controllers

The FlashSystem 900 is carved into eight volumes of 2.5 tebibytes (TiB) each that are then presented to the SVC cluster. This FS900 controller is marked as being in the primary site and all of its volumes are SVC mdisks. These mdisks are all configured into a single storage pool with a total capacity of 16 TiB.

The FlashSystem 5030 is carved into eight volumes of 1.8 Tebibytes (TiB) each that are then presented to the SVC cluster. This V5030 controller is marked as being in the secondary site and all of its volumes are SVC mdisks. These mdisks are all configured into a single storage pool with a total capacity of 14.4 TiB.

2.4.2 Host presentation

For this solution, all host attachment to the SVC cluster is being done by using 10 Gb iSCSI with jumbo frames enabled. We allocated two of the 10 GbE ports on each SVC node to serve host communication. This configuration provides four paths per volume per site to the data. The SVC Ethernet ports are connected to the leaf switches in each respective location. The logical local area network (LAN) and wide area network (WAN) connectivity for iSCSI storage access is illustrated in Figure 2-7.

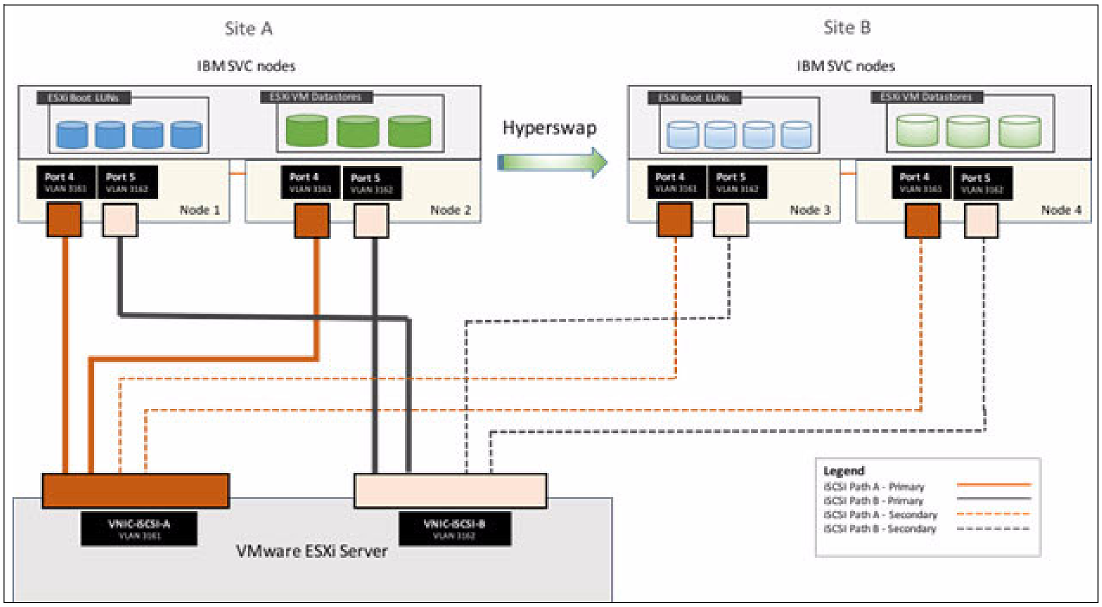


Figure 2-7 iSCSI connectivity

Hosts are assigned a site parameter based on which location they are in. In this environment, hosts are flagged as being in either the primary or secondary site depending on their location. If you use Asymmetric Logical Unit Access (ALUA), the SVC advertises preferred paths to each host to be on the same site as the site configured on the host. For example, hosts on the primary site are given preferred paths from the nodes located in the primary site. However, it is important to note that hosts should be seeing paths from all the nodes in the SVC cluster.

2.5 Virtualization design

This section highlights the general implementation of VMware ESXi and vCenter.

2.5.1 VMware vCenter

For the hypervisor environment, a common vCenter is used to manage an ESXi cluster spanning both sites. If the vCenter is to have increased availability through vSphere HA, extending through both sites, it needs to be hosted within the VersaStack and have a cross site management network that it is placed in, as was covered in 2.3.1, “Cisco UCS LAN connectivity” on page 16. If the vCenter sits outside of VersaStack environment, availability in site impact scenarios might need to be worked out by implementing vCenter HA (not covered in this document).

2.5.2 Virtual switching

The port groups for tenant applications, vMotion, iSCSI, and potentially management networks will be common to both sites, whether they are managed by a manually created vSwitch/vDS, or implemented by the Virtual Machine Manager (VMM). The VMM association is created for the first site placement of VersaStack within APIC. It creates a vDS within the vCenter that has port groups associated to VLANs from a dynamic pool that is configured within the APIC. It is then carried within the appropriate Unified Computing System (UCS) virtual Network Interface Controllers (vNICs) on both sites.

The connectivity for vMotion, iSCSI, and a potential management network will all have been created as Endpoint Groups (EPGs during the setup of the first site. Those EPGs will be joined by the UCS connectivity of the second site as static ports, automatically enabling L2 extension for that traffic.

2.5.3 VM mobility and availability

With uniform storage access configured for the hosts, vMotion, Disaster Recovery System (DRS), and HA are all options within the vSphere environment for the cluster spanning the two sites.

3

Design and planning considerations

This section helps you design a specific implementation of VersaStack and HyperSwap. The following topics are covered in this chapter:

- VMware design considerations
- VMware vCenter setup
- ESXi host installations
- Storage design considerations

© Copyright IBM Corp. 2018. All rights reserved.

3.1 VMware design considerations

This section explains the design considerations for VMware and vCenter Planning. It is not intended to be the sole source of information for implementing VMware ESXi and VMware vCenter inside of VersaStack.

3.1.1 VMware configuration checklist

The following items are required to gain the full benefit of the VMware vSphere Metro Storage Cluster (vMSC) environment. This high-level list includes the major tasks that you must complete. The details and expertise that are required to complete these tasks are beyond the intended scope of this book. For more information, see the VMware Communities and the VMware Product Interoperability Matrices.

Complete the following steps:

1. Create naming conventions for these objects:
   - Data center-wide naming VMware ESXi.
   - VMware vSphere Storage Distributed Resources Scheduler (SDRS) data stores and pools data center affinity.
   - VMware vSphere Distributed Resource Scheduler (DRS) vMotion pools data center affinity.
2. Set up *all* hardware and create a detailed inventory list. For more information, see the VMware Compatibility Guide.
   - Create an inventory list with details that cover the entire installation.
   - Mark the IBM SAN Volume Controller node names carefully and create associations in vSphere so that you know which SAN Volume Controller nodes are hosted in each data center.
3. Build two ESXi hosts in each data center for maximum resiliency:
   - Patch and update to the latest VMware patch level.
   - Follow VMware vSphere High Availability (HA) deployment.
4. Create one VM to host vCenter:
   - Update and patch vCenter.
   - Build a stretched ESXi cluster between two data centers.
   - (Optional) Implement I/O control on storage.
   - (Optional) Implement VMware vSphere Distributed Switches (VDSs).
5. Build an SDRS pool:
   - Create at least two pools to match data center affinity.
   - Differentiate between Mirrored and Non-Mirrored logical unit numbers (LUNs) if both Mirrored and Non-Mirrored LUNs are used.
   - Set the SDRS pool to manual in the beginning and monitor it before you automate it.
6. Enable DRS:
   - Create affinity rules to the ESXi host in each data center.
   - Create affinity rules to VMs, if necessary.

  - Create VM to ESXi affinity rules.
  - Set DRS to partial or automatic if the rules are trusted 100%.

3.2 VMware vCenter setup

You must implement vCenter configuration options as part of the IBM SVC HyperSwap configuration.

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

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

3.2.1 Metro vMotion vMSC

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

In the example IBM SVC HyperSwap solution, the maximum distance is 100 km (62 miles), which is an RTT of maximum 5 ms. This is not a Metro vMotion in VMotion terms, but this distance is supported.

The VMware guide to a vMSC in vSphere 6 is available in the VMware vSphere Metro Storage Cluster Recommended Practices documentation.

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

3.3 ESXi host installations

This section focuses on the design and implementation that relate to a specific ESXi configuration with an IBM SVC HyperSwap solution.

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

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

3.3.1 ESXi adapter requirements

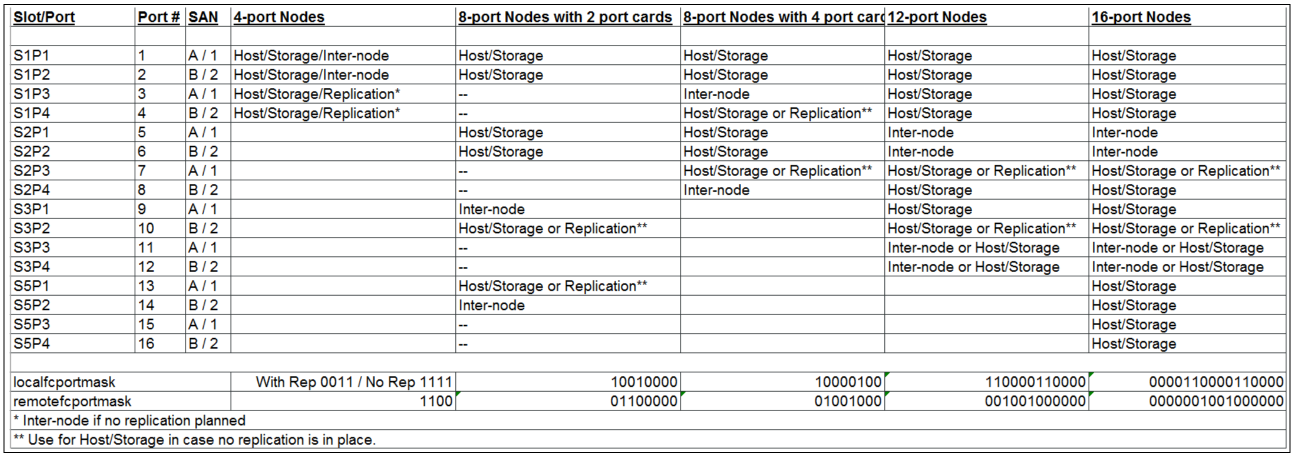
In VersaStack, the host bus adapters (HBAs) and network interface card (NIC) for each individual server are governed by the service profile assigned to the server in the UCS Manager. For explicit instructions and guidance about creating and managing adapters inside the UCS and how to assign the storage in VMware, see the Cisco Validated Design (CVD) Guide.

3.4 Storage design considerations

This section highlights the design and planning considerations for the SVC Cluster with HyperSwap solution.

3.4.1 Design considerations for IBM SVC Fibre Channel connectivity

The general guideline for Fibre Channel connectivity is to isolate different types of traffic to help mitigate against a single port or card from being overloaded. Figure 3-1 illustrates the preferred port designations for SVC nodes depending on how many ports are available per node.



Slot/Port	Port #	SAN	4-port Nodes	8-port Nodes with 2 port cards	8-port Nodes with 4 port car	12-port Nodes	16-port Nodes
S1P1	1	A / 1	Host/Storage/Inter-node	Host/Storage	Host/Storage	Host/Storage	Host/Storage
S1P2	2	B / 2	Host/Storage/Inter-node	Host/Storage	Host/Storage	Host/Storage	Host/Storage
S1P3	3	A / 1	Host/Storage/Replication*	--	Inter-node	Host/Storage	Host/Storage
S1P4	4	B / 2	Host/Storage/Replication*	--	Host/Storage or Replication**	Host/Storage	Host/Storage
S2P1	5	A / 1		Host/Storage	Host/Storage	Inter-node	Inter-node
S2P2	6	B / 2		Host/Storage	Host/Storage	Inter-node	Inter-node
S2P3	7	A / 1		--	Host/Storage or Replication**	Host/Storage or Replication**	Host/Storage or Replication**
S2P4	8	B / 2		--	Inter-node	Host/Storage	Host/Storage
S3P1	9	A / 1		Inter-node		Host/Storage	Host/Storage
S3P2	10	B / 2		Host/Storage or Replication**		Host/Storage or Replication**	Host/Storage or Replication**
S3P3	11	A / 1		--		Inter-node or Host/Storage	Inter-node or Host/Storage
S3P4	12	B / 2		--		Inter-node or Host/Storage	Inter-node or Host/Storage
S5P1	13	A / 1		Host/Storage or Replication**			Host/Storage
S5P2	14	B / 2		Inter-node			Host/Storage
S5P3	15	A / 1		--			Host/Storage
S5P4	16	B / 2		--			Host/Storage
localfcportmask			With Rep 0011 / No Rep 1111	10010000	10000100	110000110000	0000110000110000
remotefcportmask			1100	01100000	01001000	001001000000	0000001001000000
* Inter-node if no replication planned							
** Use for Host/Storage in case no replication is in place.							

Figure 3-1 Suggested SVC Port Designations on FC SAN

Figure 3-1 also provides guidelines for the `localfcportmask` and `remotefcportmask` values.

The `localfcportmask` dictates which ports the SVC permits inter-node traffic to be used on. The value can be read as a binary string where 0 represents a port that is not allowed and 1 represents a port that is allowed. This string is read from right to left to identify port numbers. For example, a `localfcportmask` value of `10000001` would represent a cluster where inter-node connectivity is allowed to take place on port IDs 1 and 8 only. In contrast, the mask of `1000001` would represent a cluster where inter-node connectivity can only occur on ports 1 and 7.

The same rules apply to `remotefcportmask`. However, this value dictates which ports are allowed by the SVC to be used for replication.

3.4.2 Fibre Channel connectivity guidelines

Our testing environment for this publication includes eight Fibre Channel (FC) ports. These connections are going into a standard dual fabric topology, as shown in Figure 3-2. Note that this diagram shows that each node is connected to each switch by using four ports (two per HBA). This configuration is intended to ensure maximum redundancy to prevent a single SAN failure from affecting the stability of the overall solution.

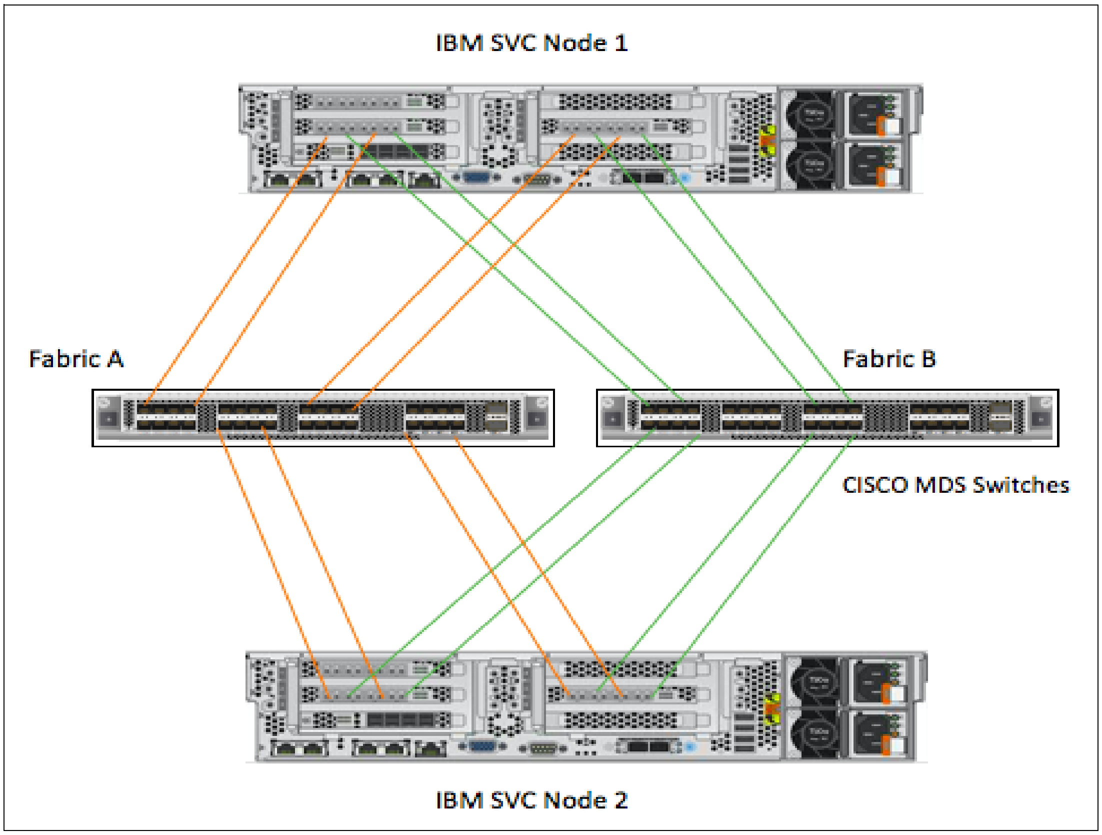


Figure 3-2 Two Node SVC Cluster Fibre Channel connectivity to Cisco MDS SAN switches

Similarly, Figure 3-3 shows the FS900 AE2 connecting to the same switches using eight ports (one per card per fabric) for the same reason that the SVC is connected this way.

Note: The FS900 AE2 allows for up to eight 16-Gbps connections (two ports per card). The FS900 AE3 allows for up to 16 16-Gbps connections (four ports per card).

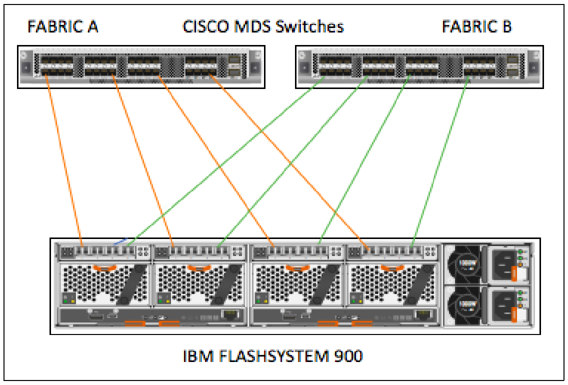


Figure 3-3 IBM FlashSystem 900 Fibre Channel connectivity with Cisco MDS switches

3.4.3 Design considerations for iSCSI connectivity

When implementing iSCSI for host attachment on Spectrum Virtualize Platforms, consider the following guidelines:

- Use a TCP/IP offload engine (TOE) when possible.
- Minimize the number of switches and routers between servers and storage.
- Enable priority flow control (PFC) when possible.
- Use iSCSI authentication (CHAP).
- Allow the switch to automatically negotiate the highest possible speed.
- Separate iSCSI traffic by using VLAN tagging (IEEE 802.1Q).
- Use the correct cabling for data transmission. For copper cabling, use CAT6 rated cables for gigabit networks and CAT 6a or CAT-7 cabling for 10-Gb implementations. For fiber, use OM3 or OM4 multimode fiber cabling.
- Enable Jumbo frames (set MTU to 9216).
- Separate management traffic from I/O traffic.
- Set iSCSI host and system storage to the same link speed.
- Use a maximum of four iSCSI sessions per node and one IQN per host.
- Configure traffic storm control on the Ethernet switches.

As with every implementation, see the IBM System Storage Interoperation Center to validate that the planned solution is supported.

3.4.4 Ethernet connectivity guidelines

All of the host attachment in our implementation is done by using 10 Gb iSCSI. Additionally, iSCSI is the preferred protocol for VersaStack host attachment. Our environment has SVC nodes with a single 4-port 10-GbE HBA. To avoid overloading our hosts with too many paths, we only used two of these ports per node. This configuration provides a total of eight paths per host (four from Site A and four from Site B). The Ethernet connectivity is depicted in Figure 3-4.

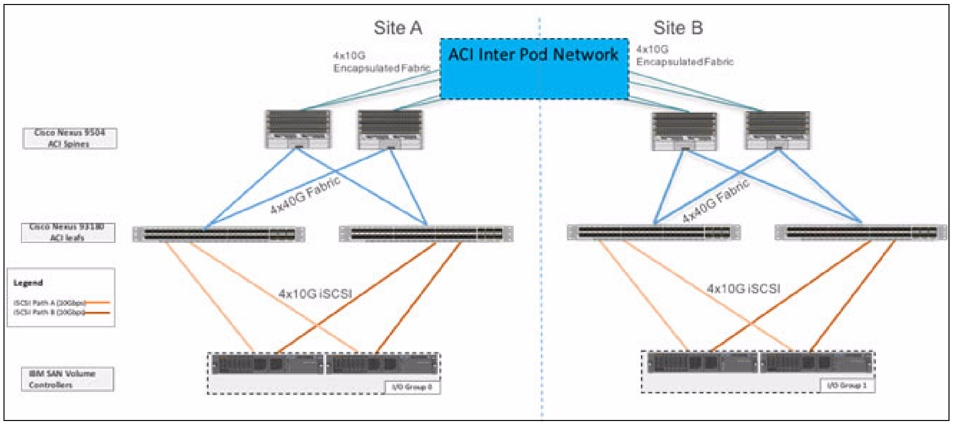
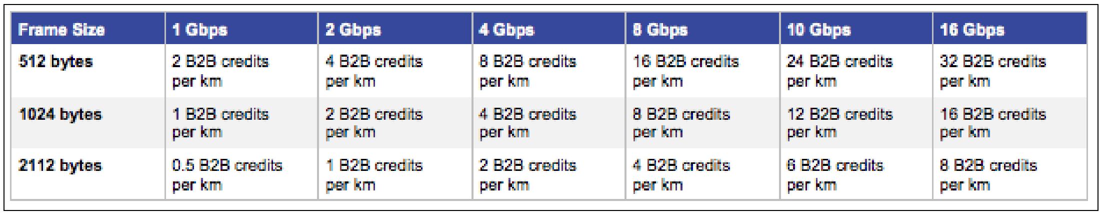


Figure 3-4 iVersaStack testing environment Ethernet connectivity

3.4.5 Design considerations for Inter-site SAN connectivity

The IBM SVC storage array communication across the sites is enabled by native Fibre Channel extension. Proper planning is necessary to avoid SAN congestion with SAN extensions depending on the customer environment. For this discussion, for example the MDS 9396S can be used for replication and backup services over long-distance native Fibre Channel links. By default, all ports on the MDS 9396S have 500 buffer-to-buffer credits (BBCs). This quantity is enough for a 62-km Fibre Channel link, assuming full-size frames of 2112 bytes. The number of BBCs can be increased up to 4095 per port using enterprise licenses. This quantity is enough for a 510-km Fibre Channel link, assuming full-size frames of 2112 bytes.

Figure 3-5 lists the number of BBCs required per kilometer of ISL at different speeds and frame sizes. Note that one BBC is needed per frame irrespective of the frame size. Also, scope should be left for a longer end-to-end path in the optical WAN/MAN infrastructure and with small frame sizes.



Frame Size	1 Gbps	2 Gbps	4 Gbps	8 Gbps	10 Gbps	16 Gbps
512 bytes	2 B2B credits per km	4 B2B credits per km	8 B2B credits per km	16 B2B credits per km	24 B2B credits per km	32 B2B credits per km
1024 bytes	1 B2B credits per km	2 B2B credits per km	4 B2B credits per km	8 B2B credits per km	12 B2B credits per km	16 B2B credits per km
2112 bytes	0.5 B2B credits per km	1 B2B credits per km	2 B2B credits per km	4 B2B credits per km	6 B2B credits per km	8 B2B credits per km

Figure 3-5 Per-kilometer BBC requirements at different speeds and frame sizes

This use case assumes the existence of an optical infrastructure between two data centers. The Cisco MDS 9000 Series switches support a wide range of optical transceivers. More details are available in the Cisco MDS 9000 Family Pluggable Transceivers data sheet.

3.4.6 Use Port-Channels

A Port-Channel is a logical interface with multiple physical ports as members. ISLs between switches should be aggregated into Port-Channels. In case of failure of a single physical port, data traffic automatically fails over to other ports in the same Port-Channel. A single physical link failure does not trigger Fibre Channel control plane (Fabric Shortest Path First (FSPF)) recalculations. Port-Channels bring resiliency and stability to your SAN.

Cisco MDS 9000 Series switches do not require any special license to use Port-Channels. Port-Channel members can reside anywhere on a switch. A single Port-Channel interface on the Cisco MDS switches can have up to 16 physical ports. Use the following guidelines for the position of Port-Channel member ports to increase scale and resilience.

Port-Channel Members on the MDS 9148S

Port-Channel members should be uniformly distributed among ports 1 - 16, 17 - 32, and 33 - 48 on the MDS 9148S. Within these three sets of 16 ports, Port-Channel members should further be uniformly distributed among port-groups. A port-group on the MDS 9148S consists of four consecutive ports. For example, ports 1 - 4 are in port-group 1, ports 5 - 8 are in port-group 2, and so on.

Using this two-level recommendation (Figure 3-6), a Port-Channel with three members should use ports 1, 17, and 33 (or any other ports as long as there is one from each set of ports mentioned above). A Port-Channel with four members should use ports 1, 17, 33, and 5 (from port-group 2 within the set of ports 1 - 16). A Port-Channel with five members should use ports 1, 17, 33, 5 (from port-group 2 within the set of ports 1 - 16), and 21 (from port-group 6 within the set of ports 17 - 32). Port-Channels with more members should be distributed by extrapolating this approach.

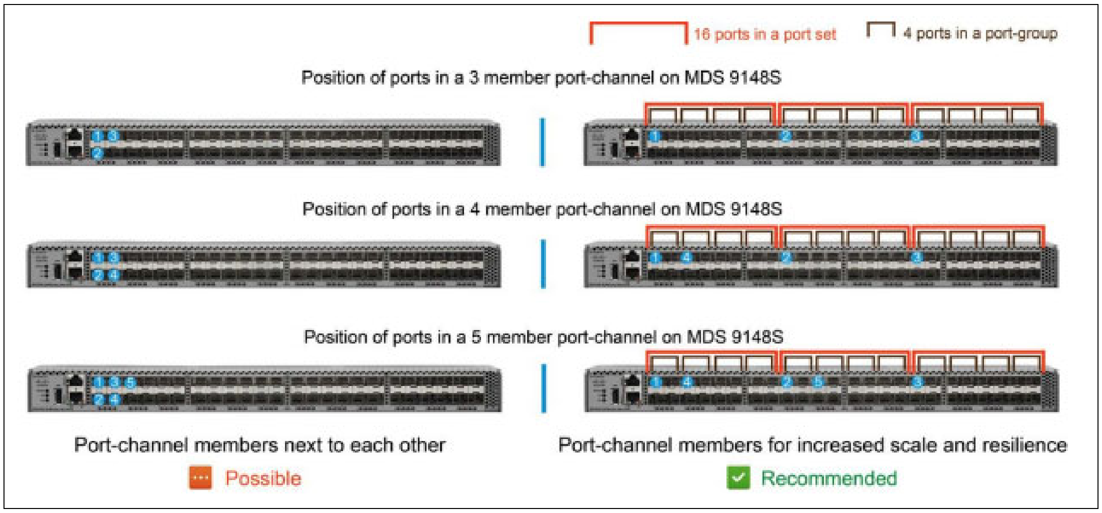

16 ports in a port set
4 ports in a port-group
Position of ports in a 3 member port-channel on MDS 9148S
Position of ports in a 4 member port-channel on MDS 9148S
Position of ports in a 5 member port-channel on MDS 9148S
Port-channel members next to each other
Possible
Port-channel members for increased scale and resilience
Recommended

Figure 3-6 Guidelines for positions of members of a Port-Channel on the MDS 9148S

Port-Channel Members on the MDS 9396S

On the MDS 9396S, Port-Channel members should be uniformly distributed among ports 1 - 8, 9 - 16, 17 - 24, and so on. Within these 12 sets of eight ports, Port-Channel members should further be uniformly distributed among port-groups. A port-group on the MDS 9396S consists of four consecutive ports. For example, ports 1 - 4 are in port-group 1, ports 5 - 8 are in port-group 2, and so on.

Using this two-level recommendation (Figure 3-7), a Port-Channel with three members should use ports 1, 9, and 17. A Port-Channel with 12 members should use ports 1, 9, 17, 25, 33, 41, 49, 57, 65, 73, 81, and 89. A Port-Channel with 13 members should use ports 1, 9, 17, 25, 33, 41, 49, 57, 65, 73, 81, 89, and 5 (from port-group 2 within the set of ports 1 - 8). Port-Channels with more members should be distributed by extrapolating this approach.

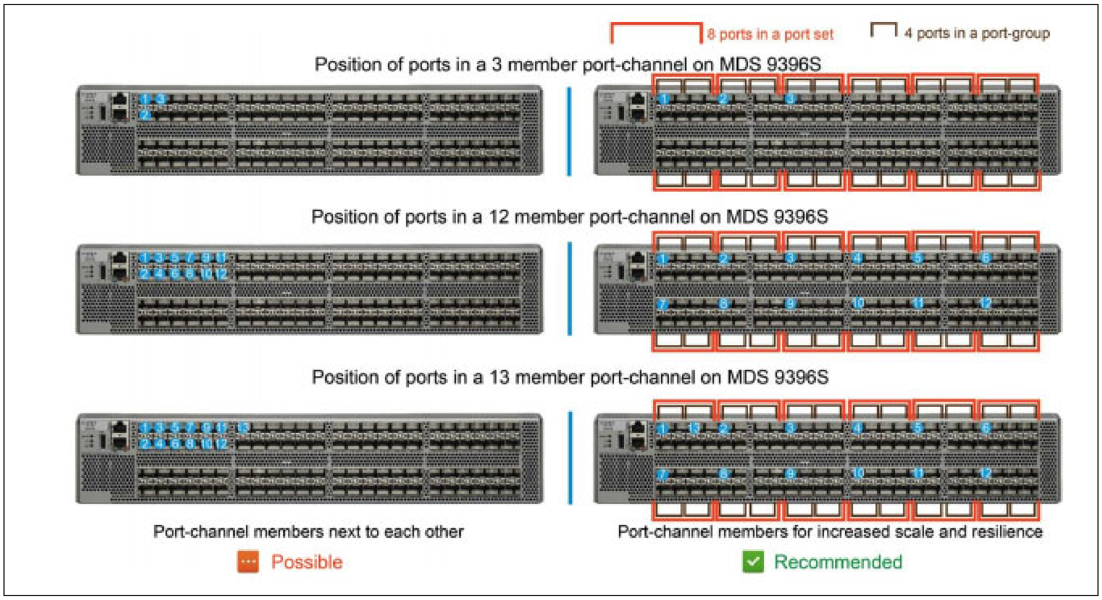

8 ports in a port set
4 ports in a port-group
Position of ports in a 3 member port-channel on MDS 9396S
Position of ports in a 12 member port-channel on MDS 9396S
Position of ports in a 13 member port-channel on MDS 9396S
Port-channel members next to each other
Possible
Port-channel members for increased scale and resilience
Recommended

Figure 3-7 Guidelines for positions of members of a Port-Channel on the MDS 9396S

The licensing model for the Cisco MDS 9000 Series switches provides flexibility to choose the port number as long as the total number of activated ports is within the licensed limit. With the 12-port base license on the MDS 9148S, you can activate any 12 ports. The activated ports need not be consecutive or the first 12 ports. As per these guidelines, you should activate ports 1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 41, and 45 with this base license.

Similarly, with the 48-port base license on the MDS 9396S, you can activate any 48 ports. As per the guidelines above, you should activate ports 1, 3, 5, 7, and so on with this base license. Ports can be activated or deactivated by using the Cisco NX-OS command port-license acquire. More information is available in the Cisco MDS 9000 Family NX-OS Licensing Guide. Consider your cable plant and future expansion plans when deciding which ports to activate.

3.4.7 Zoning

Consider the following guidelines when designing zoning:

1. Avoid using a default zone (permit all) for your production traffic.
2. Generally, use device aliases for all port worldwide names (pWWNs) in your fabric. Device aliases provide user-friendly and human-readable names to pWWNs to simplify zoning operations.
3. Enable enhanced zoning on all switches. When you begin a zoning change, the switch creates a session to lock the entire fabric to implement the change. The lock is released after the zoning change is committed. This feature helps maintain zoning database consistency between switches in the same fabric.

4. For smart zoning, generally configure a single initiator to a single target zone. However, this approach requires SAN administrators to spend a great deal of time performing configuration and management. Using smart zoning, you can create zones in which all initiators and targets are in the same zone. Cisco MDS 9000 Series switches internally create single initiator to single target zones based on your configuration in smart zones. Using smart zoning, you get operational simplicity with optimized resources.
5. Remove configuration of unused zones from your active zone set to free up resources.
6. Names of zones, zone sets, and device aliases should be descriptive and convey the meaning in a crisp format. Use any standard naming convention to maintain consistency and shorter length.

4

Implementing the HyperSwap solution

This chapter describes the implementation of the HyperSwap solution on VersaStack. The following topics are covered in this chapter:

- Deploying the VersaStack converged infrastructure
- Cisco UCS deployment
- Cisco Application Centric Infrastructure Inter-Pod deployment
- Multi-Pod setup configuration
- Site 2 Spine configuration
- Site 2 Leaf Discovery
- Fibre Channel SAN using Cisco MDS Switches
- Zoning configuration
- Cisco Multilayer Director Switch zone configuration
- IBM SVC HyperSwap planning
- Active-active Metro Mirror considerations
- IBM SVC HyperSwap configuration

© Copyright IBM Corp. 2018. All rights reserved.

4.1 Deploying the VersaStack converged infrastructure

Deploy the VersaStack converged infrastructure for the primary site, following the instructions found in the VersaStack with Cisco Application Centric Infrastructure and IBM SAN Volume Controller Cisco Validated Design Guide (CVD).

The example networks that were used during lab validation between the two sites are shown in Table 4-1.

Table 4-1 Example networks used during lab validation

Site A			Site B		
Name	**VLAN**	**Subnet**	**Name**	**VLAN**	**Subnet**
IB-Mgmt-Site-A	111	10.1.160.0/24	IB-Mgmt-Site-B	211	10.2.160.0/24
Native-VLAN	2	N/A	Native-VLAN	2	N/A
CrossSite-Mgmt (1)	311	10.3.160.0/24	CrossSite-Mgmt (2)	311	10.3.160.0/24
vMotion	3173	10.29.173.0/24	vMotion	110	10.29.173.0/24
iSCSI-A-VLAN	3161	10.29.161.0/24	iSCSI-A-VLAN	3162	10.29.161.0/24
iSCSI-B-VLAN	3162	10.29.162.0/24	iSCSI-B-VLAN	3162	10.29.162.0/24
VM-App-[1101-1150]	1101-1150	As allocated by the customer	VM-App-[1101-1150]	1101-1150	As allocated by the customer

The example virtual storage area networks (VSANs) that were used during the lab validation are shown in Table 4-2.

Table 4-2 Example VSANs used during the lab validation

Name	Node ID	Fabric
VSAN-A	101	Fabric A Public VSAN
VSAN-B	102	Fabric B Public VSAN
VSAN-A	201	Fabric A Private replication VSAN
VSAN-B	202	Fabric B Private replication VSAN

Note: Site-specific management networks rather than a dedicated cross site management network should be used as appropriate to the deployment.

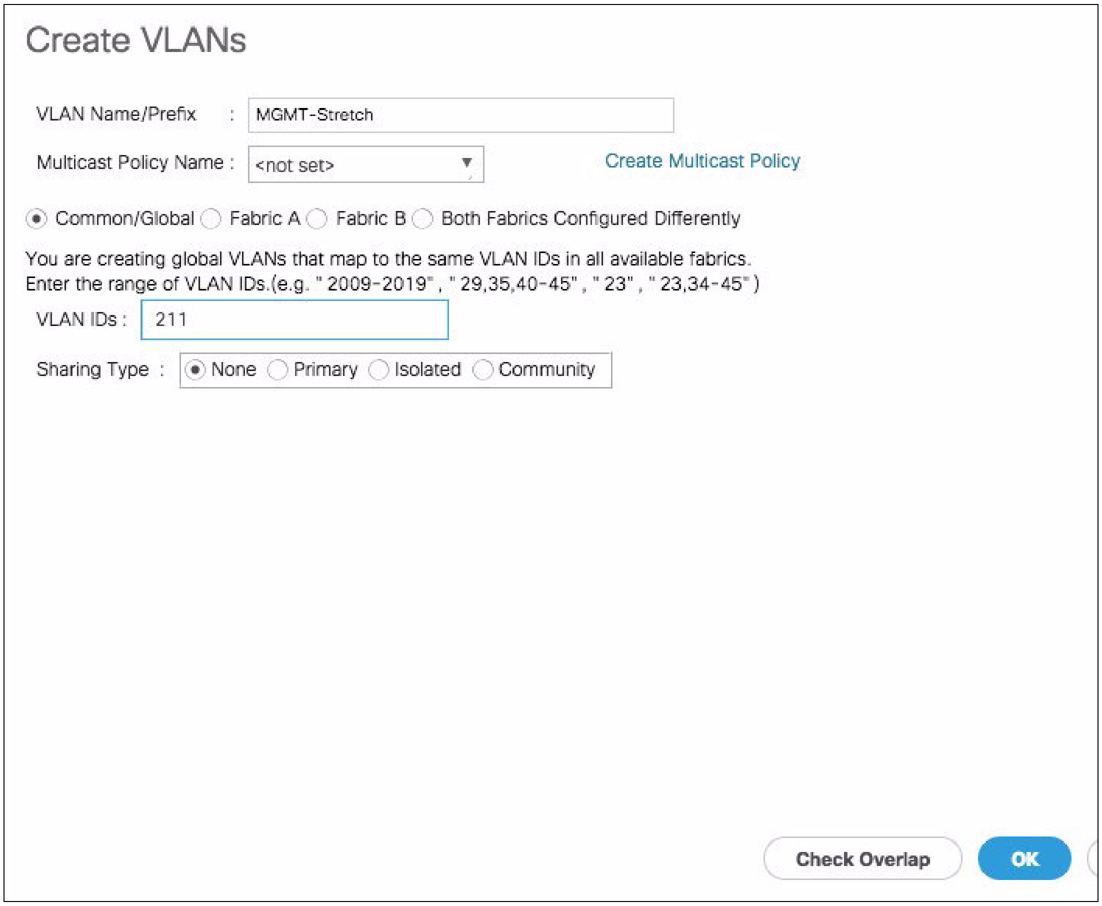
The VersaStack data center for the secondary site can be deployed by using the same CVD instructions that are used for the primary site. The virtualization instructions should be followed for the secondary site, but you ignore the setup of another vCenter.

4.2 Cisco UCS deployment

The following additional steps must be performed within the primary site during the configuration. There are no UCS deployment steps specific to Multi-Pod deployment.

The inband management site-specific VLANs and cross-site management VLAN must be configured within the Cisco Unified Computing System (UCS) Manager. In this solution, vCenter was deployed in the cross-site management VLAN to allow vCenter HA implementation as an option:

1. In Cisco UCS Manager, click the LAN tab in the navigation pane.
2. Click **LAN** → **LAN Cloud**.
3. Right-click **VLANs**.
4. Select **Create VLANs.**
5. Enter `<MGMT-Stretch>` as the name of the VLAN to be used as the stretched VLAN.
6. Keep the **Common/Global** option selected for the scope of the VLAN.
7. Enter the native VLAN ID.
8. Keep the **Sharing Type** as **None**. See Figure 4-1.



Create VLANs

VLAN Name/Prefix : MGMT-Stretch

Multicast Policy Name : <not set>

● Common/Global ○ Fabric A ○ Fabric B ○ Both Fabrics Configured Differently

You are creating global VLANs that map to the same VLAN IDs in all available fabrics.
Enter the range of VLAN IDs.(e.g. " 2009-2019" , " 29,35,40-45" , " 23" , " 23,34-45")

VLAN IDs : 211

Sharing Type : ● None ○ Primary ○ Isolated ○ Community

Figure 4-1 Create VLANs

9. Click **OK**, and then click **OK** again.
10. With this VLAN created, it must be added within Cisco UCS Manager as follows:

    LAN → **Policies** → **root (or appropriate org)** → **vNIC Templates** → **vNIC Template vNIC_Mgmt_A**

4.3 Cisco Application Centric Infrastructure Inter-Pod deployment

To deploy a Cisco Application Centric Infrastructure (ACI) Multi-Pod network, single site ACI configuration should already be in place. In our test environment, Pod1 was already configured as a single-site VersaStack environment. Complete these steps to set up the network:

1. Configure IPN device in both Site 1 and Site 2.
2. Configure Spines on Site 1.
3. Configure Multi-Pod.
4. Discover Spines and Leaves on Site 2 (new site).
5. Configure Spines on Site 2 to complete the Multi-Pod setup.

See 4.3.1, “IPN device configurations” on page 36 for details of various device links and associated IP addresses and subnets used in the example configurations.

Figure 4-2 shows IPN and Spine connectivity.

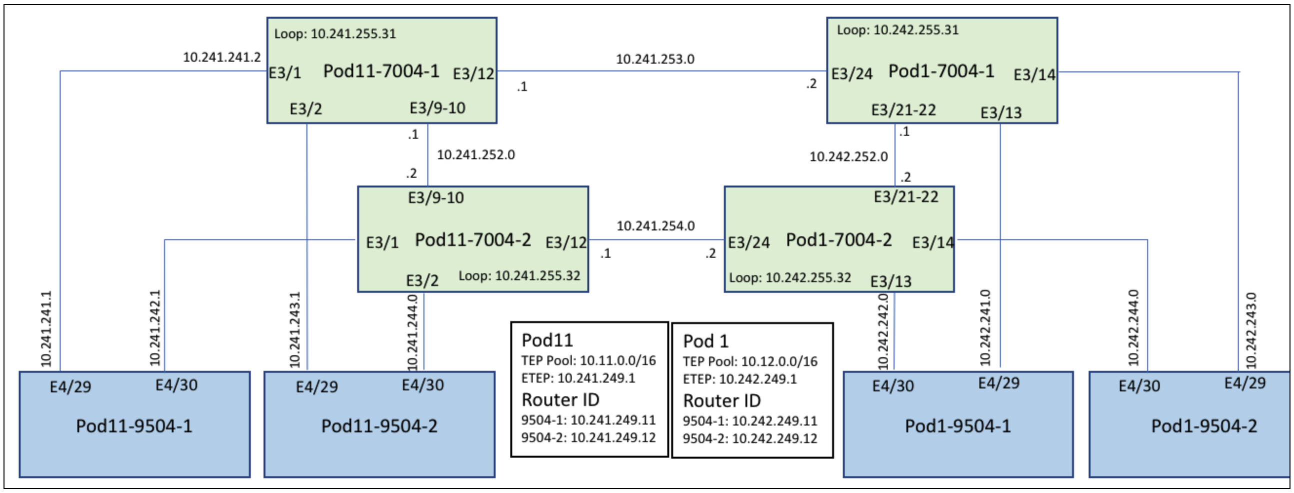


Figure 4-2 IPN and Spine connectivity

4.3.1 IPN device configurations

The following captures show the relevant configurations of the four IPN devices. The configuration enables multicast, Open Shortest Path First (OSPF), jumbo maximum transmission unit (MTU), and DHCP relay on all the Inter-Pod Network (IPN) devices. The configuration uses a single IPN device as rendezvous point (RP) for Protocol Independent Multicast (PIM) Bidir traffic. In a production network, deploying a phantom RP for high availability is preferred. For more information, see the *Bidirectional PIM Deployment Guide*.

Example 4-1 shows the configuration for Pod 1 7004-1.

Example 4-1 Configuration for Pod 1 7004-1

```
feature ospf
feature pim
feature lacp
feature dhcp
feature lldp
!
! Define the multicast groups and associated RP addresses
!
ip pim rp-address 10.241.255.1 group-list 225.0.0.0/8 bidir
ip pim rp-address 10.241.255.1 group-list 239.0.0.0/8 bidir
ip pim ssm range 232.0.0.0/8
!
service dhcp
ip dhcp relay
!
interface port-channel1
  description To Pod 1 7004-2
  mtu 9216
  ip address 10.242.252.1/30
  ip ospf network point-to-point
  ip ospf mtu-ignore
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
!
interface Ethernet3/21
  mtu 9216
  channel-group 1 mode active
  no shutdown
interface Ethernet3/22
  mtu 9216
  channel-group 1 mode active
  no shutdown
!
interface Ethernet3/13
  description to Pod1 Spine-1 interface E4/29
  mtu 9216
  no shutdown
!
interface Ethernet3/13.4
  mtu 9216
  encapsulation dot1q 4
  ip address 10.242.241.2/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  !
  ! DHCP relay for forwarding DHCP queries to APIC’s in-band IP address
  !
  ip dhcp relay address 10.12.0.1
  ip dhcp relay address [APIC 2’s IP]
  ip dhcp relay address [APIC 3’s IP]
  no shutdown
```

```
!
interface Ethernet3/14
  description Pod 11 Spine-2 interface E4/29
  mtu 9216
  no shutdown
!
interface Ethernet3/14.4
  mtu 9216
  encapsulation dot1q 4
  ip address 10.242.243.2/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.12.0.1
  ip dhcp relay address [APIC 2's IP]
  ip dhcp relay address [APIC 3's IP]
  no shutdown
!
interface Ethernet3/24
  description Link to Pod 11 7004-1 E3/12
  mtu 9216
  ip address 10.241.253.2/30
  ip ospf network point-to-point
  ip ospf mtu-ignore
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
!
interface loopback0
  description Loopback to be used as Router-ID
  ip address 10.242.255.31/32
  ip router ospf 10 area 0.0.0.0
!
interface loopback1
  description PIM RP Address
  ip address 10.241.255.1/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
!
router ospf 10
  router-id 10.242.255.31
  log-adjacency-changes
!
```

Example 4-2 shows the configuration for Pod 1 7004- 2.

Example 4-2 Configuration for Pod 1 7004 -2

```
feature ospf
feature pim
feature lacp
feature dhcp
feature lldp
!
! Define the multicast groups and associated RP addresses
```

```
!
ip pim rp-address 10.241.255.1 group-list 225.0.0.0/8 bidir
ip pim rp-address 10.241.255.1 group-list 239.0.0.0/8 bidir
ip pim ssm range 232.0.0.0/8
!
service dhcp
ip dhcp relay
!
interface port-channel1
  description To Pod 1 7004-1
  mtu 9216
  ip address 10.242.252.2/30
  ip ospf network point-to-point
  ip ospf mtu-ignore
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
!
interface Ethernet3/21
  mtu 9216
  channel-group 1 mode active
  no shutdown
interface Ethernet3/22
  mtu 9216
  channel-group 1 mode active
  no shutdown
!
interface Ethernet3/13
  description to Pod1 Spine-1 interface E4/30
  mtu 9216
  no shutdown
!
interface Ethernet3/13.4
  mtu 9216
  encapsulation dot1q 4
  ip address 10.242.242.2/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  !
  ! DHCP relay to forward DHCP queries to APIC’s in-band IP address
  !
  ip dhcp relay address 10.12.0.1
  ip dhcp relay address [APIC 2’s IP]
  ip dhcp relay address [APIC 3’s IP]
  no shutdown
!
interface Ethernet3/14
  description to Pod1 Spine-2 interface E4/30
  mtu 9216
  no shutdown
!
interface Ethernet3/14.4
  mtu 9216
  encapsulation dot1q 4
  ip address 10.242.244.2/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.12.0.1
  ip dhcp relay address [APIC 2’s IP]
  ip dhcp relay address [APIC 3’s IP]
  no shutdown
!
interface Ethernet3/24
  description Link to Pod 1 7004-2 E3/12
  mtu 9216
  ip address 10.241.254.2/30
  ip ospf network point-to-point
  ip ospf mtu-ignore
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
!
interface loopback0
  description Loopback to be used as Router-ID
  ip address 10.242.255.32/32
  ip router ospf 10 area 0.0.0.0
!
router ospf 10
  router-id 10.241.254.2
  log-adjacency-changes
!
```

Example 4-3 shows the configuration for Pod 11 7004-1.

Example 4-3 Configuration for Pod 11 7004-1

```
feature ospf
feature pim
feature lacp
feature dhcp
feature lldp
!
! Define the multicast groups and associated RP addresses
!
ip pim rp-address 10.241.255.1 group-list 225.0.0.0/8 bidir
ip pim rp-address 10.241.255.1 group-list 239.0.0.0/8 bidir
ip pim ssm range 232.0.0.0/8
!
service dhcp
ip dhcp relay
!
interface port-channel1
  description To Pod 11 7004-2
  mtu 9216
  ip address 10.241.252.1/30
  ip ospf network point-to-point
  ip ospf mtu-ignore
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
```

```
!
interface Ethernet3/9
  mtu 9216
  channel-group 1 mode active
  no shutdown
interface Ethernet3/10
  mtu 9216
  channel-group 1 mode active
  no shutdown
!
interface Ethernet3/1
  description to Pod11 Spine-1 interface E4/29
  mtu 9216
  no shutdown
!
interface Ethernet3/1.4
  mtu 9216
  encapsulation dot1q 4
  ip address 10.241.241.2/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  !
  ! DHCP relay for forwarding DHCP queries to APIC’s in-band IP address
  !
  ip dhcp relay address 10.12.0.1
  ip dhcp relay address [APIC 2’s IP]
  ip dhcp relay address [APIC 3’s IP]
  no shutdown
!
interface Ethernet3/2
  description Pod 11 Spine-2 interface E4/29
  mtu 9216
  no shutdown
!
interface Ethernet3/2.4
  mtu 9216
  encapsulation dot1q 4
  ip address 10.241.243.2/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.12.0.1
  ip dhcp relay address [APIC 2’s IP]
  ip dhcp relay address [APIC 3’s IP]
  no shutdown
!
interface Ethernet3/12
  description Link to Pod 1 7004-1
  mtu 9216
  ip address 10.241.253.1/30
  ip ospf network point-to-point
  ip ospf mtu-ignore
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
!
interface loopback0
  description Loopback to be used as Router-ID
  ip address 10.241.255.31/32
  ip router ospf 10 area 0.0.0.0
router ospf 10
  router-id 10.241.255.31
  log-adjacency-changes
!
```

Example 4-4 shows the configuration for Pod 11 7004-2.

Example 4-4 Configuration for Pod 11 7004-2

```
feature ospf
feature pim
feature lacp
feature dhcp
feature lldp
!
! Define the multicast groups and associated RP addresses
!
ip pim rp-address 10.241.255.1 group-list 225.0.0.0/8 bidir
ip pim rp-address 10.241.255.1 group-list 239.0.0.0/8 bidir
ip pim ssm range 232.0.0.0/8
!
service dhcp
ip dhcp relay
!
interface port-channel1
  description To POD 11 7004-1
  mtu 9216
  ip address 10.241.252.2/30
  ip ospf network point-to-point
  ip ospf mtu-ignore
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
!
interface Ethernet3/9
  mtu 9216
  channel-group 1 mode active
  no shutdown
!
interface Ethernet3/10
  mtu 9216
  channel-group 1 mode active
  no shutdown
!
interface Ethernet3/1
  description Pod 11 Spine-1 E4/30
  mtu 9216
  no shutdown
interface Ethernet3/1.4
  mtu 9216
  encapsulation dot1q 4
  ip address 10.241.242.2/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.12.0.1
  ip dhcp relay address [APIC 2’s IP]
  ip dhcp relay address [APIC 3’s IP]
  no shutdown
!
interface Ethernet3/2
  description To Pod 11 Spine-2 E4/30
  mtu 9216
  no shutdown
!
interface Ethernet3/2.4
  mtu 9216
  encapsulation dot1q 4
  ip address 10.241.244.2/30
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.12.0.1
  ip dhcp relay address [APIC 2’s IP]
  ip dhcp relay address [APIC 3’s IP]
  no shutdown
!
interface Ethernet3/12
  description Link to Pod 1 7004-2
  mtu 9216
  ip address 10.241.254.1/30
  ip ospf network point-to-point
  ip ospf mtu-ignore
  ip router ospf 10 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
!
interface loopback0
  description Loopback to be used as Router-ID
  ip address 10.241.255.32/32
  ip router ospf 10 area 0.0.0.0
!
router ospf 10
  router-id 10.241.255.32
  log-adjacency-changes
!
```

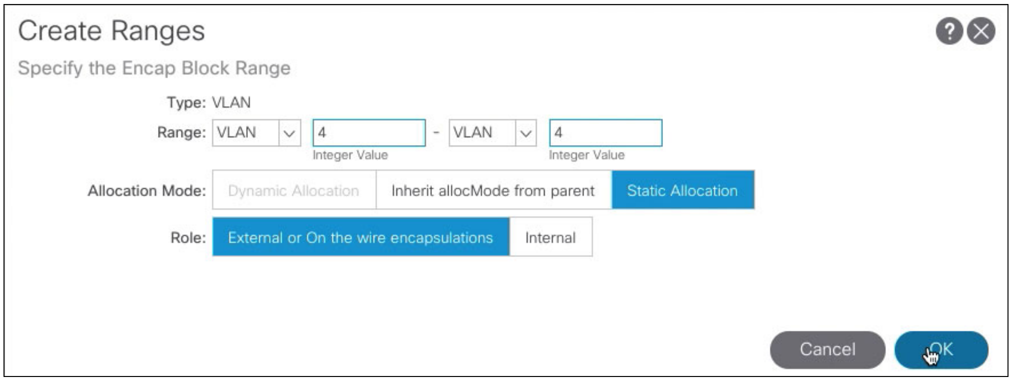
4.3.2 Spine configurations for Pod 1

The following captures show the relevant configurations of the Spine switches to enable Multi-Pod configuration. As previously stated, both sites are configured with different Tunnel End Point (TEP) pools: 10.11.0.0/16 (Pod11) and 10.12.0.0/16 (Pod1).

Create VLAN Pool

To create the VLAN Pool, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Access Policies**.
2. In the left pane, expand **Pools**, right-click **VLAN**, and select **Create VLAN Pool**.
3. Provide a Name for the VLAN Pool (`MultiPod-vlans`), select **Static Allocation**, and click **+** to add a VLAN range.
4. Enter a single VLAN 4, select **Static Allocation**, and click **OK** (Figure 4-3).



Create Ranges

Specify the Encap Block Range

Type: VLAN

Range: VLAN 4 - VLAN 4

Allocation Mode: Static Allocation

Role: External or On the wire encapsulations

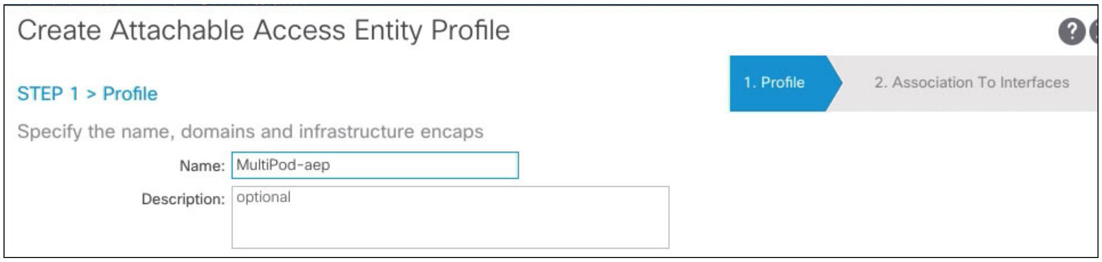
Figure 4-3 Create Ranges window

5. Click **Submit** to finish create the VLAN Pool.

Create AEP for Spine connectivity

To create an Attachable Access Entity Profile (AEP) for Spine connectivity, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Access Policies**.
2. In the left pane, expand **Global Policies**, right-click **Attachable Access Entity Profile**, and select **Create Attachable Access Entity Profile**.
3. Provide a name for the AEP (`MultiPod-aep`) and click **Next** (Figure 4-4).



Create Attachable Access Entity Profile

STEP 1 > Profile

Specify the name, domains and infrastructure encaps

Name: MultiPod-aep

Description: optional

Figure 4-4 Attachable Access Entity Profile

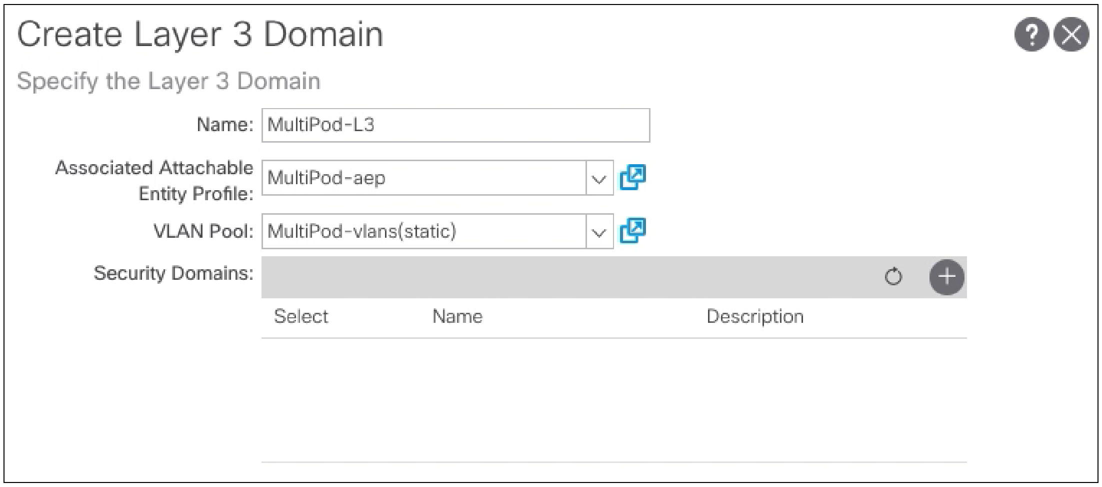
4. Click **Finish** to complete the AEP creation without adding any interfaces.

Create External Routed Domain for Spine connectivity

To create an External Routed Domain for Spine connectivity, complete the following steps:

1. Log in to the Application Policy Infrastructure Controller (APIC) GUI and click **Fabric** → **Access Policies**.
2. In the left pane, expand **Physical and External Domains**, right-click **External Routed Domains**, and select **Create Layer 3 Domain**.

3. Provide a name for the Layer 3 domain (`MultiPod-L3`).
4. From the **Associated Attachable Entity Profile** menu, select the recently created AEP (**MultiPod-aep**).
5. From the **VLAN Pool** menu, select the recently created VLAN Pool (**MultiPod-vlans**), as shown in Figure 4-5.



Create Layer 3 Domain

Specify the Layer 3 Domain

Name: MultiPod-L3

Associated Attachable Entity Profile: MultiPod-aep

VLAN Pool: MultiPod-vlans(static)

Security Domains:

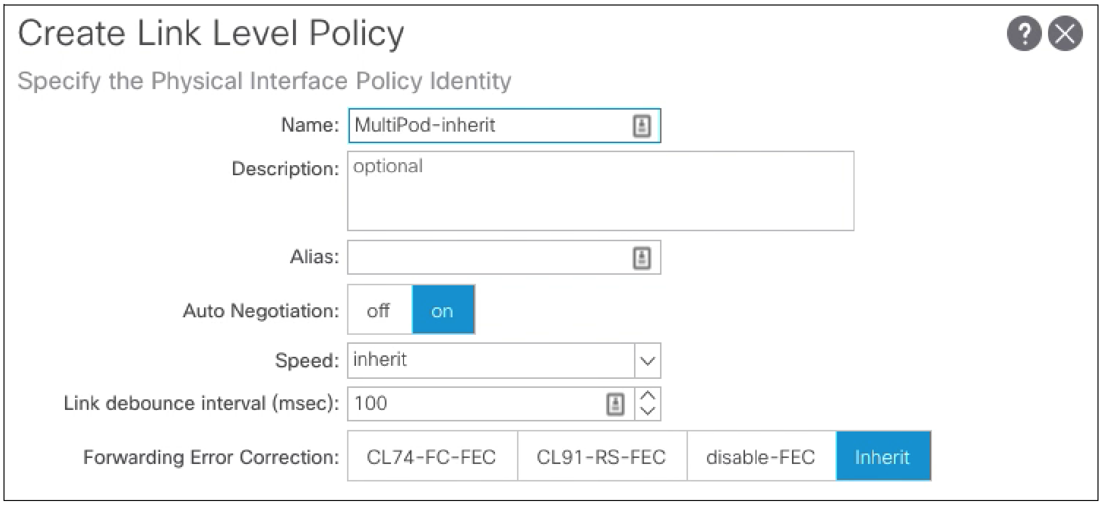
Figure 4-5 Create layer 3 domain

6. Click **Submit** to finish creating the Layer 3 domain**.**

Create Link Level Interface policy for Spine connectivity

To create a Link Level Interface policy for Spine connectivity, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Access Policies**.
2. In the left pane, expand **Interface Policies** → **Policies** → **Link Level**.
3. Right-click **Link Level** and select **Create Link Level Policy**.
4. Provide a name for the Link Level Policy (`MultiPod-inherit`) and ensure that **Auto Negotiation** is set to **on** and **Speed** is set to **inherit**, as shown in Figure 4-6.



Create Link Level Policy

Specify the Physical Interface Policy Identity

Name: MultiPod-inherit

Description: optional

Alias:

Auto Negotiation: off [✓ on]

Speed: inherit

Link debounce interval (msec): 100

Forwarding Error Correction: CL74-FC-FEC CL91-RS-FEC disable-FEC [✓ Inherit]

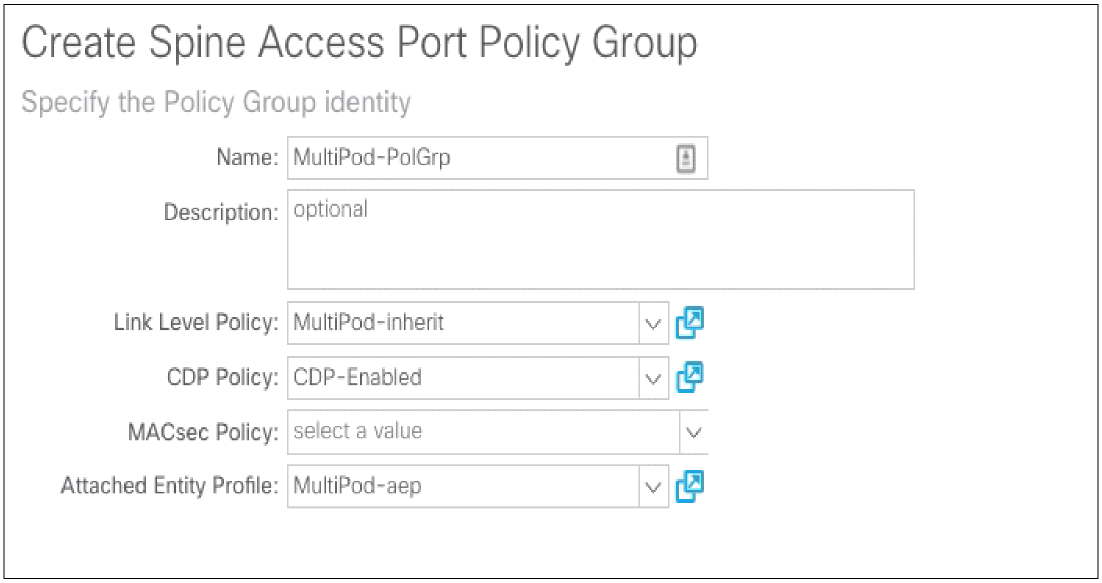
Figure 4-6 Create Link Level Policy

5. Click **Submit** to create the policy**.**

Create Spine Policy Group

To create a Spine Policy Group, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Access Policies**.
2. In the left pane, expand **Interface Policies** → **Policy Groups**.
3. Right-click **Spine Policy Group** and select **Create Spine Access Port Policy Group**.
4. Provide a name for the Spine Access Port Policy Group (`MultiPod-PolGrp`).
5. From the **Link Level Policy** menu, select the recently created policy (**MultiPod-Inherit**).
6. From the **CDP Policy** menu, select the previously created policy to enable Cisco Discovery Protocol (CDP) (**CDP-Enabled**).
7. From the **Attached Entity Profile** menu, select the recently created AEP (**MultiPod-aep**), as shown in Figure 4-7.



Create Spine Access Port Policy Group

Specify the Policy Group identity

Name: MultiPod-PolGrp

Description: optional

Link Level Policy: MultiPod-inherit

CDP Policy: CDP-Enabled

MACsec Policy: select a value

Attached Entity Profile: MultiPod-aep

Figure 4-7 Create Spine Policy Group

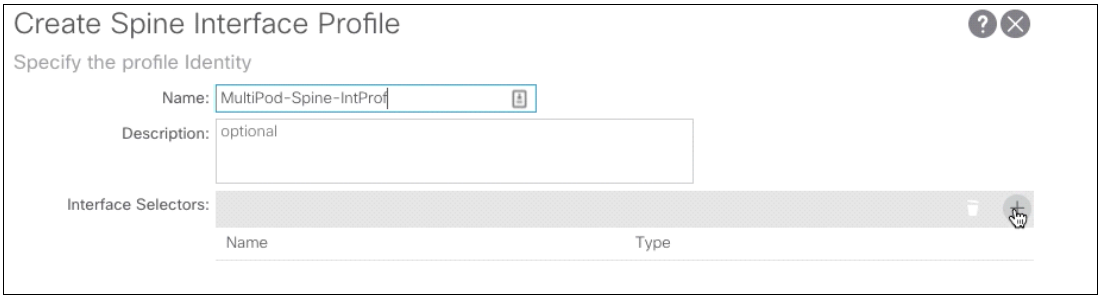
8. Click **Submit**.

Create Spine Interface Profile

To create a Spine Interface Profile, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Access Policies**.
2. In the left pane, expand **Interface Policies** → **Profiles.**
3. Right-click **Spine Profiles** and select **Create Spine Interface Profile**.
4. Provide a **Name** for the Spine Interface Profile (`MultiPod-Spine-IntProf`).

5. Click **+** to add **Interface Selectors**, as shown in Figure 4-8.



Create Spine Interface Profile

Specify the profile Identity

Name: MultiPod-Spine-IntProf

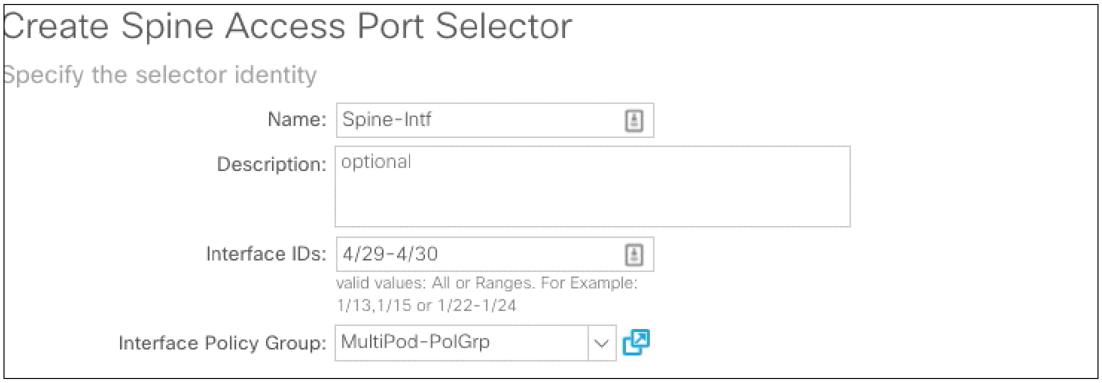
Description: optional

Interface Selectors:

Name	Type

Figure 4-8 Create Spine Interface Profile

6. Provide a name for the Spine Access Port Selector (`Spine-Intf`).
7. For Interface IDs, add interfaces that connect to the two IPN devices (4/29-4/30).
8. From the **Interface Policy Group** menu, select the recently created Policy Group (**MultiPod-PolGrp**), as shown in Figure 4-9.



Create Spine Access Port Selector

Specify the selector identity

Name: Spine-Intf

Description: optional

Interface IDs: 4/29-4/30

valid values: All or Ranges. For Example: 1/13,1/15 or 1/22-1/24

Interface Policy Group: MultiPod-PolGrp

Figure 4-9 Create Spine access Port Selector

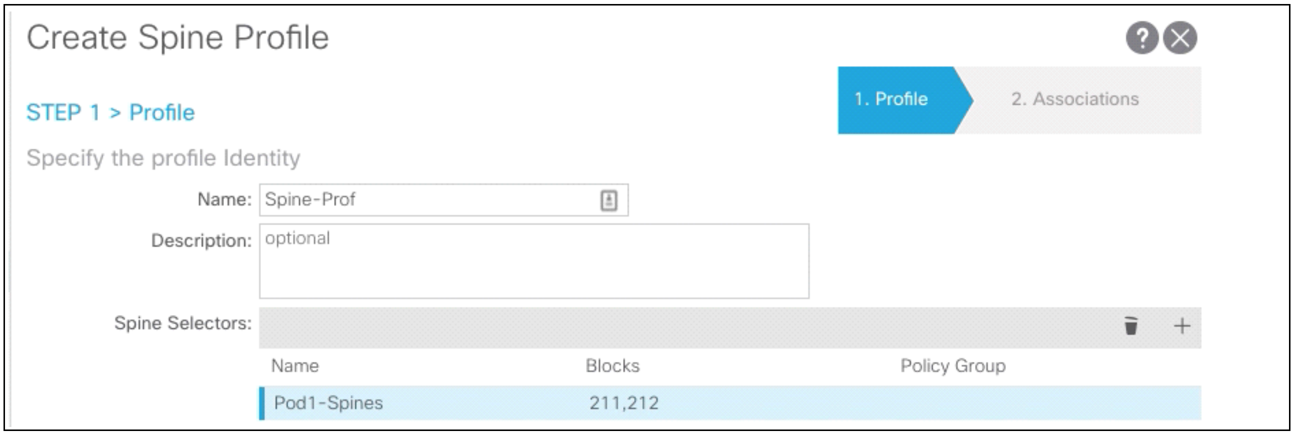
9. Click **OK** to finish creating the Access Port Selector.
10. Click **Submit** to finish creating the Spine Interface Profile.

Create Spine Profile

To create a Spine Profile, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Access Policies**.
2. In the left pane, expand **Switch Policies** → **Profiles**.
3. Right-click **Spine Profiles** and select **Create Spine Profile**.
4. Provide a **Name** for the Spine Profile (`Spine-Prof`).
5. Click **+** to add **Spine Selectors**.

6. Provide a **Name** for the Spine Selector (`Pod1-Spines`) and from the drop-down menu under Blocks, select the spine switch IDs (211-212), as shown in Figure 4-10.



Create Spine Profile

STEP 1 > Profile

Specify the profile Identity

Name: Spine-Prof

Description: optional

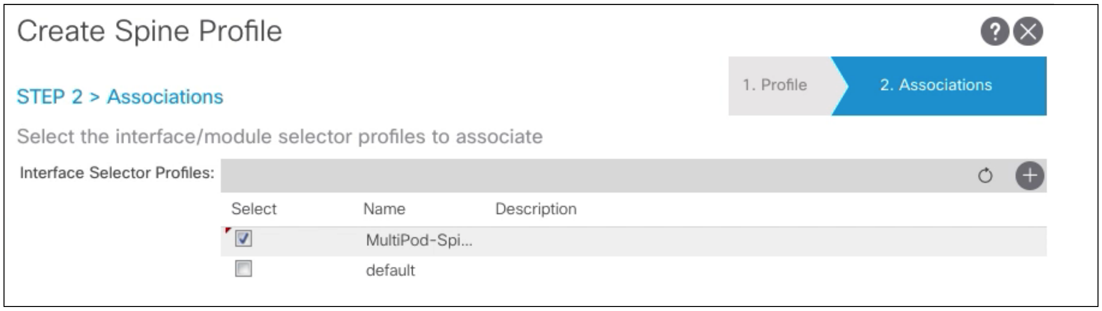
Spine Selectors:

Name	Blocks	Policy Group
Pod1-Spines	211,212	

Figure 4-10 Create Spine Profile - Step1

7. Click **Update** and then **Next**.

8. For the Interface Selector Profiles, select the recently created interface selector profile (**MultiPod-Spine-intProf**), as shown in Figure 4-11.



Create Spine Profile

STEP 2 > Associations

Select the interface/module selector profiles to associate

Interface Selector Profiles:

Select	Name	Description
☑	MultiPod-Spi...	
□	default	

Figure 4-11 Create Spine Profile - Step 2

9. Click **Finish** to complete creating the Spine Profile.

4.4 Multi-Pod setup configuration

When the original VersaStack with ACI setup was completed, a Pod (1) and the TEP Pool (10.12.0.0/16) were created as part of the setup. In this configuration step, Pod and TEP addresses are defined for the second Site and Multi-Pod configuration is completed on the APIC. The Pod ID used for second Site is `11` and the TEP Pool used is `10.11.0.0/16`.

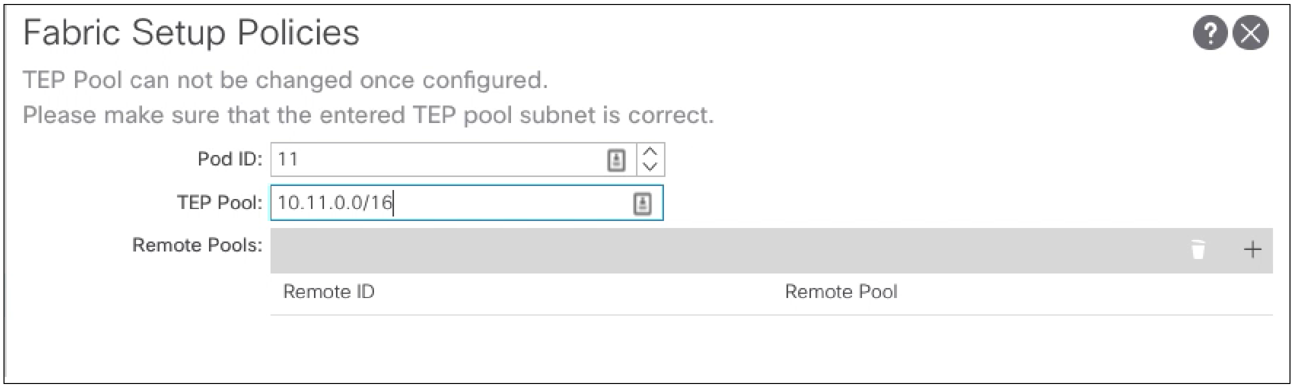
4.4.1 Set up Pod and TEP Pool for Site 2

To set up the Pod and TEP Pool for SIte 2, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Inventory**.

2. In the left pane, right-click **Pod Fabric Setup Policy** and select **Setup Pods**.

3. Enter the **Pod ID** and **TEP Pool** for Site 2, as shown in Figure 4-12.



Fabric Setup Policies

TEP Pool can not be changed once configured.
Please make sure that the entered TEP pool subnet is correct.

Pod ID: 11

TEP Pool: 10.11.0.0/16

Remote Pools:

Remote ID	Remote Pool

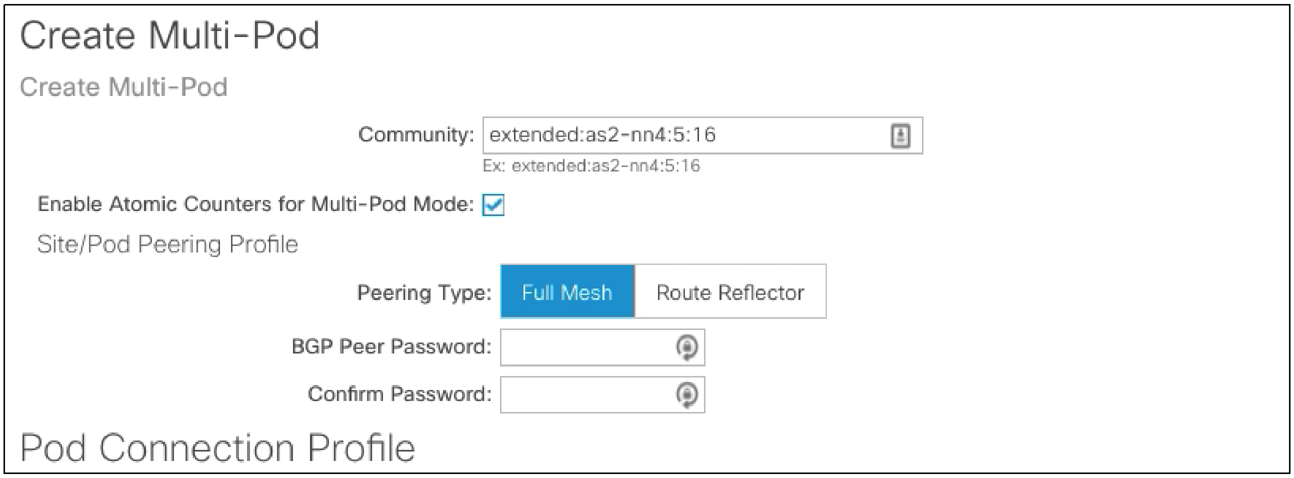
Figure 4-12 Set up Pod and TEP Pool for Site 2

4. Click **Submit**.

4.4.2 Create Multi-Pod

To create a Multi-Pod, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Inventory**.
2. In the left pane, right-click **Pod Fabric Setup Policy** and select **Create Multi-Pod**.
3. Provide a **Community** string (`extended:as2-nn4:5:16`)
4. Select **Enable Atomic Counters for Multi-pod Mode**.
5. Select **Peering Type** as **Full Mesh** (because there are only two sites), as shown in Figure 4-13.



Create Multi-Pod

Create Multi-Pod

Community: extended:as2-nn4:5:16

Ex: extended:as2-nn4:5:16

Enable Atomic Counters for Multi-Pod Mode: ☑

Site/Pod Peering Profile

Peering Type: [✓ Full Mesh] Route Reflector

BGP Peer Password:

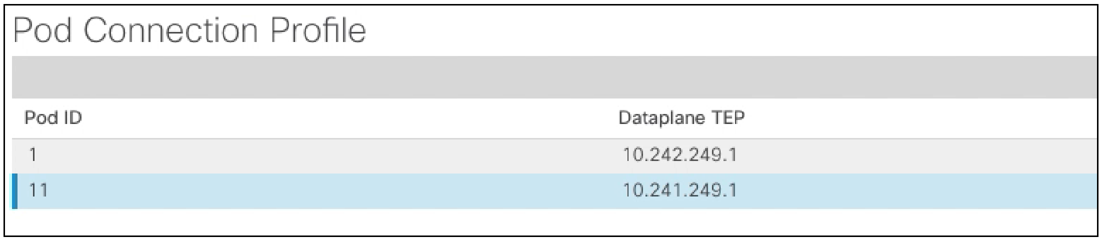
Confirm Password:

Pod Connection Profile

Figure 4-13 Create Multi-Pod

6. Click **+** to add a Pod Connection Profile.
7. Add Pod 1 and provide the Dataplane TEP or External Tunnel Endpoint (ETEP) shared by Spines at Site 1, and click **Update**.

8. Add Pod 11 and provide the name of Dataplace TEP or ETEP to be shared by Spines at Site 2 and click **Update**, as shown in Figure 4-14.

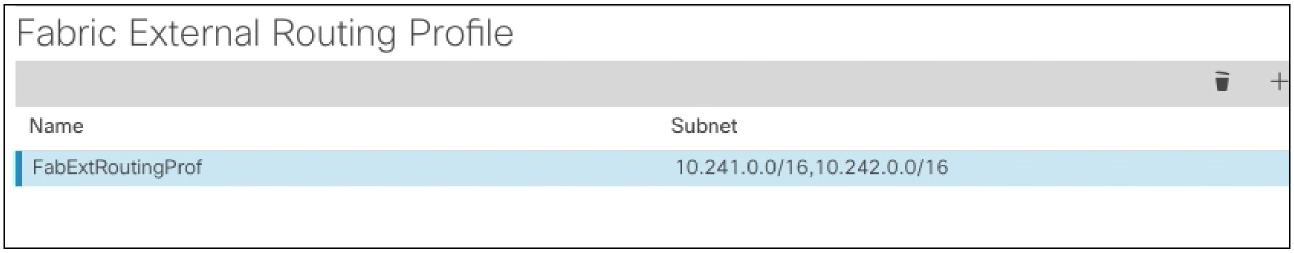


Pod Connection Profile

Pod ID	Dataplane TEP
1	10.242.249.1
11	10.241.249.1

Figure 4-14 Pod Connection Profile

9. Click **+** to add a Fabric External Routing Profile.
10. Provide a name (`FabExtRoutingProf`) and define the subnets that are used for defining point-to-point connections between Spines and IPN devices. In this guide, all the point-to-point connections are within following two subnets: `10.241.0.0/16` and `10.242.0.0/16`, as shown in Figure 4-15. Click **Update**.



Fabric External Routing Profile

Name	Subnet
FabExtRoutingProf	10.241.0.0/16,10.242.0.0/16

Figure 4-15 Fabric External Routing Profile

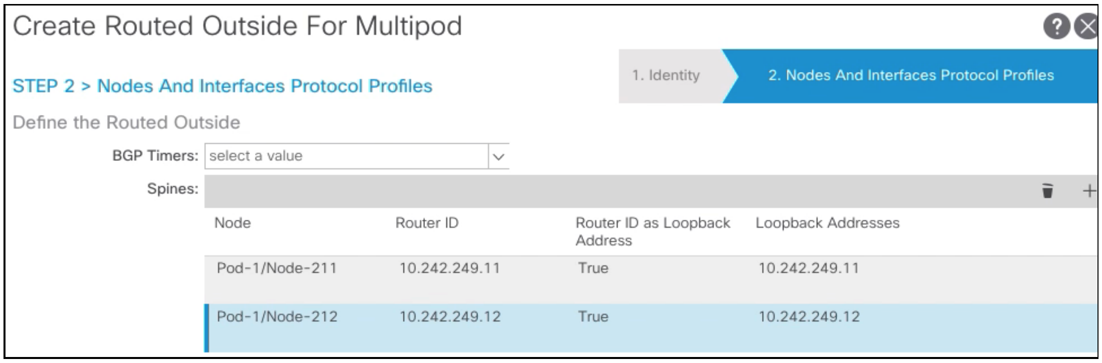
11. Click **Submit** to complete the Multi-Pod configuration.

4.4.3 Create Routed Outside for Multipod

To create the Routed Outside for Multipod, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Inventory**.
2. In the left pane, right-click **Pod Fabric Setup Policy** and select **Create Routed Outside for Multipod**.
3. Provide the OSPF Area ID as configured on the IPN devices (0.0.0.0).
4. Select the OSPF Area Type (Regular area).
5. Click **Next**.
6. Click **+** to add the first Spine.

7. Select the first Spine (**Pod-1/Node-211**) and add **Router ID (Loopback)** as shown in Figure 4-16, and click **Update**.



Create Routed Outside For Multipod

STEP 2 > Nodes And Interfaces Protocol Profiles

Define the Routed Outside

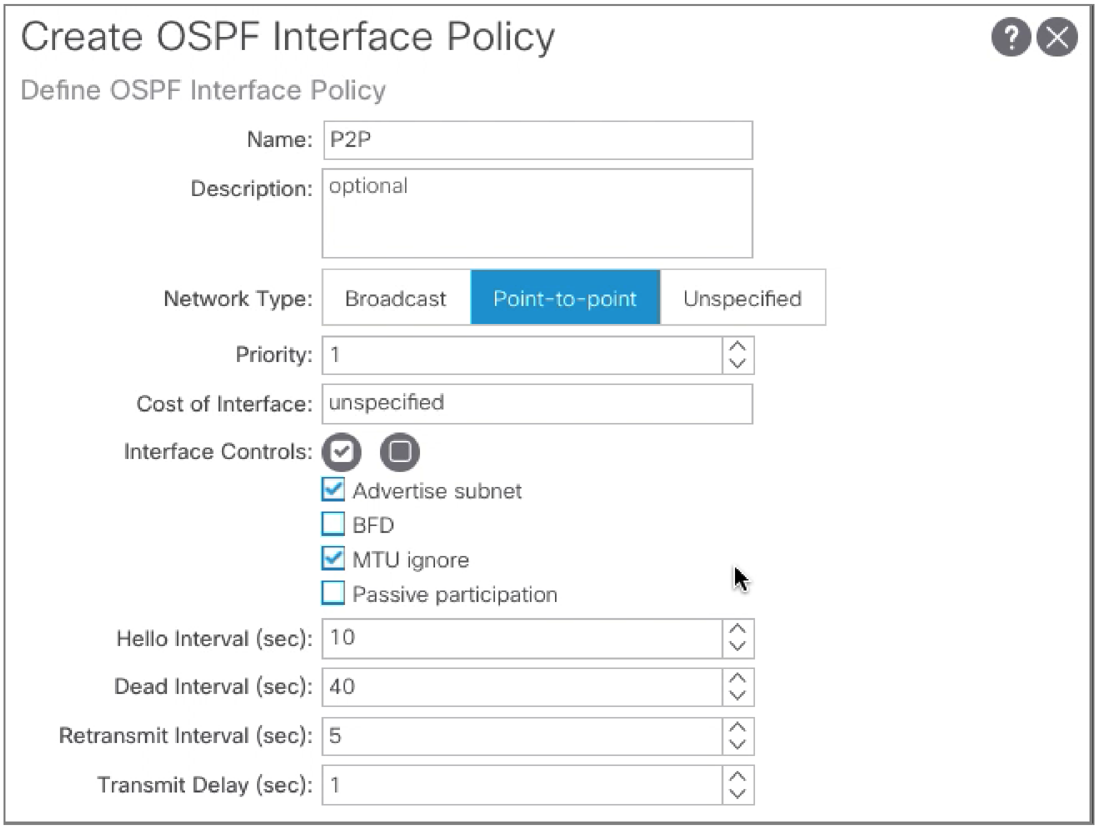
BGP Timers: select a value

Spines:

Node	Router ID	Router ID as Loopback Address	Loopback Addresses
Pod-1/Node-211	10.242.249.11	True	10.242.249.11
Pod-1/Node-212	10.242.249.12	True	10.242.249.12

Figure 4-16 Create Routed Outside for Multipod

8. Click **+** to add the second Spine.
9. Select the second Spine (**Pod-1/Node-212**) and add **Router ID (Loopback)** as shown in Figure 4-16, and click **Update**.
10. From the **OSPF Profile For Sub-Interfaces** menu, select **Create OSPF Interface Policy**.
11. Provide a **Name** for the Policy (`P2P`).
12. Select **Point-to-Point** as the **Network Type**.
13. Select **Advertise Subnet** and **MTU ignore**, as shown in Figure 4-17.



Create OSPF Interface Policy

Define OSPF Interface Policy

Name: P2P

Description: optional

Network Type: Broadcast [✓ Point-to-point] Unspecified

Priority: 1

Cost of Interface: unspecified

Interface Controls:

☑ Advertise subnet

□ BFD

☑ MTU ignore

□ Passive participation

Hello Interval (sec): 10

Dead Interval (sec): 40

Retransmit Interval (sec): 5

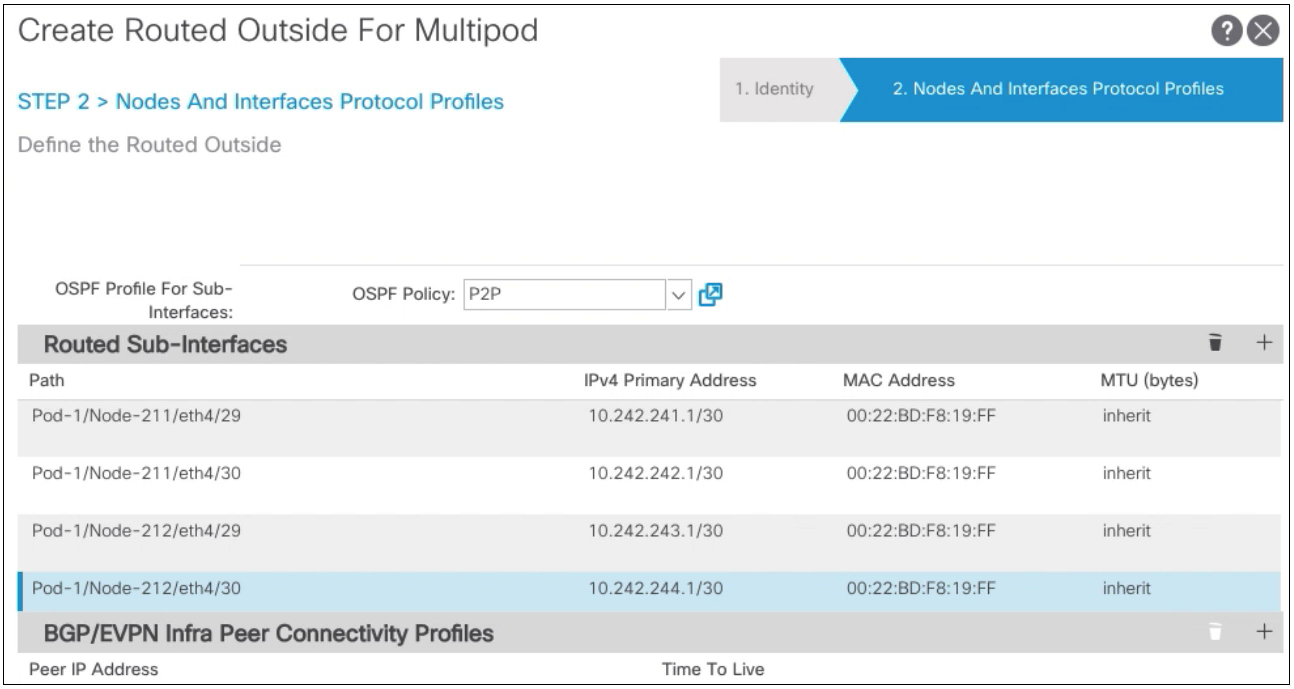
Transmit Delay (sec): 1

Figure 4-17 Create OSPF Interface Policy

14. Click **Submit**

15. Click **+** to add Routed Sub-Interfaces.

16. Add all four interfaces (Path) and their respective IP addresses to connect the Spines to IPN devices, as shown in Figure 4-18.



Create Routed Outside For Multipod

STEP 2 > Nodes And Interfaces Protocol Profiles

Define the Routed Outside

OSPF Profile For Sub-Interfaces: OSPF Policy: P2P

Routed Sub-Interfaces

Path	IPv4 Primary Address	MAC Address	MTU (bytes)
Pod-1/Node-211/eth4/29	10.242.241.1/30	00:22:BD:F8:19:FF	inherit
Pod-1/Node-211/eth4/30	10.242.242.1/30	00:22:BD:F8:19:FF	inherit
Pod-1/Node-212/eth4/29	10.242.243.1/30	00:22:BD:F8:19:FF	inherit
Pod-1/Node-212/eth4/30	10.242.244.1/30	00:22:BD:F8:19:FF	inherit

BGP/EVPN Infra Peer Connectivity Profiles

Peer IP Address	Time To Live

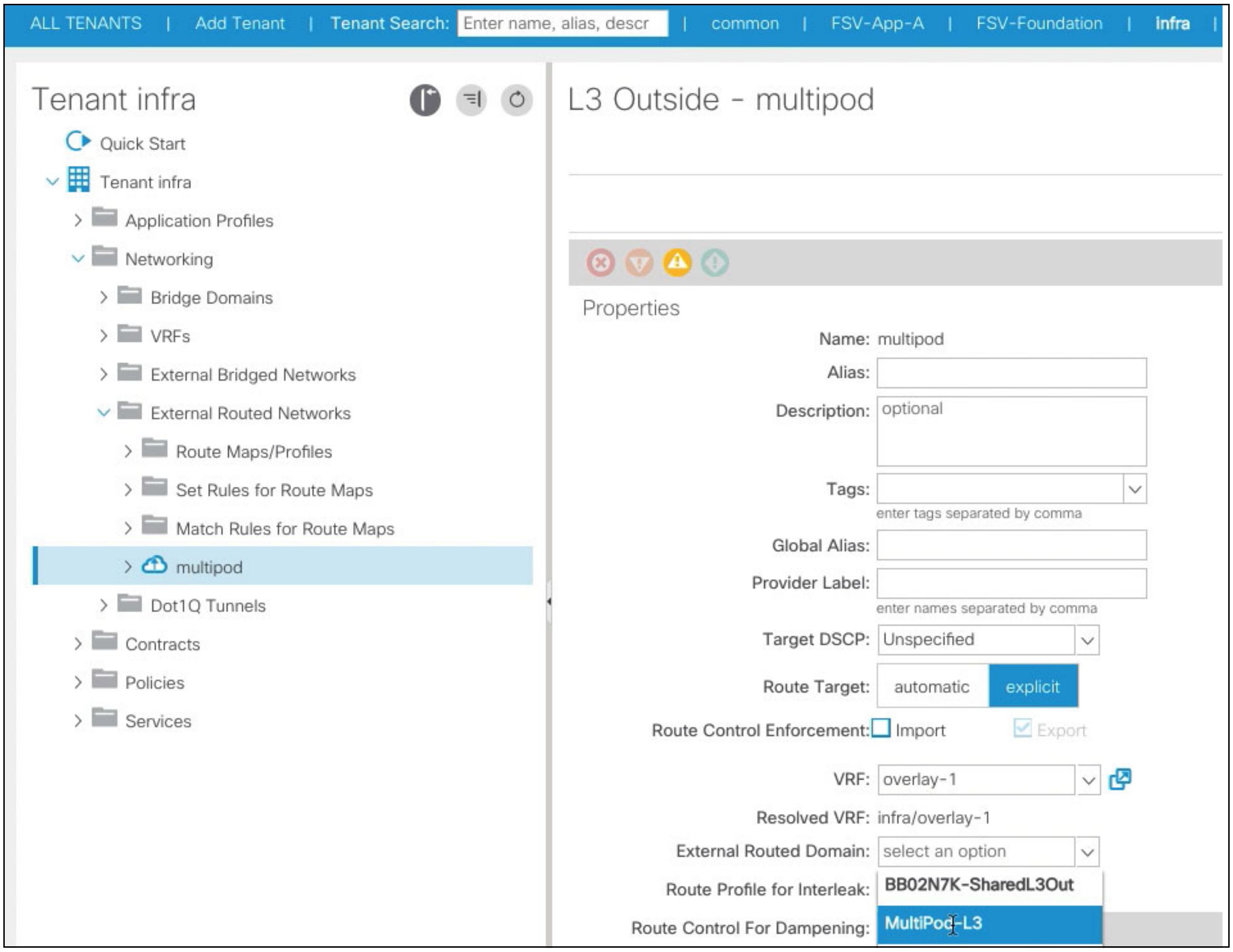
Figure 4-18 Create Routed Outside for Multipod

17. Click **Finish**.

18. Click **Tenants** → **Infra**.

19. In the left pane, expand **Networking** → **External Routed Networks** and click **multipod**.

20. From the **External Routed Domain** menu on the main page, select **MultiPod-L3**, as shown in Figure 4-19.



Name: multipod

Alias:

Description: optional

Tags: enter tags separated by comma

Global Alias:

Provider Label: enter names separated by comma

Target DSCP: Unspecified

Route Target: automatic explicit

Route Control Enforcement: □ Import ☑ Export

VRF: overlay-1

Resolved VRF: infra/overlay-1

External Routed Domain: select an option

Route Profile for Interleak: BB02N7K-SharedL3Out

Route Control For Dampening: MultiPod-L3

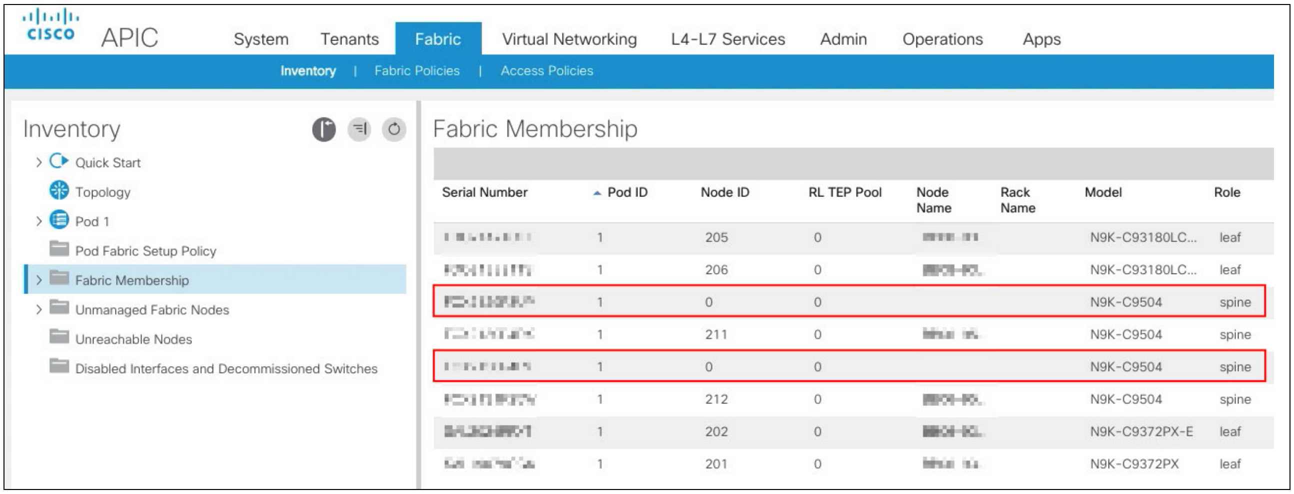
Figure 4-19 L3 Outside: multipod

21. Click **Submit**

The Site-1 spine configuration is now complete. Log in to the IPN devices to verify the OSPF routing and neighbor relationship.

4.4.4 Site 2 Spine Discovery

With the Multi-Pod network configured on Site 1, the Spines on the Site 2 should now be visible under the **Fabric** → **Inventory** → **Fabric Membership**, as shown in Figure 4-20.



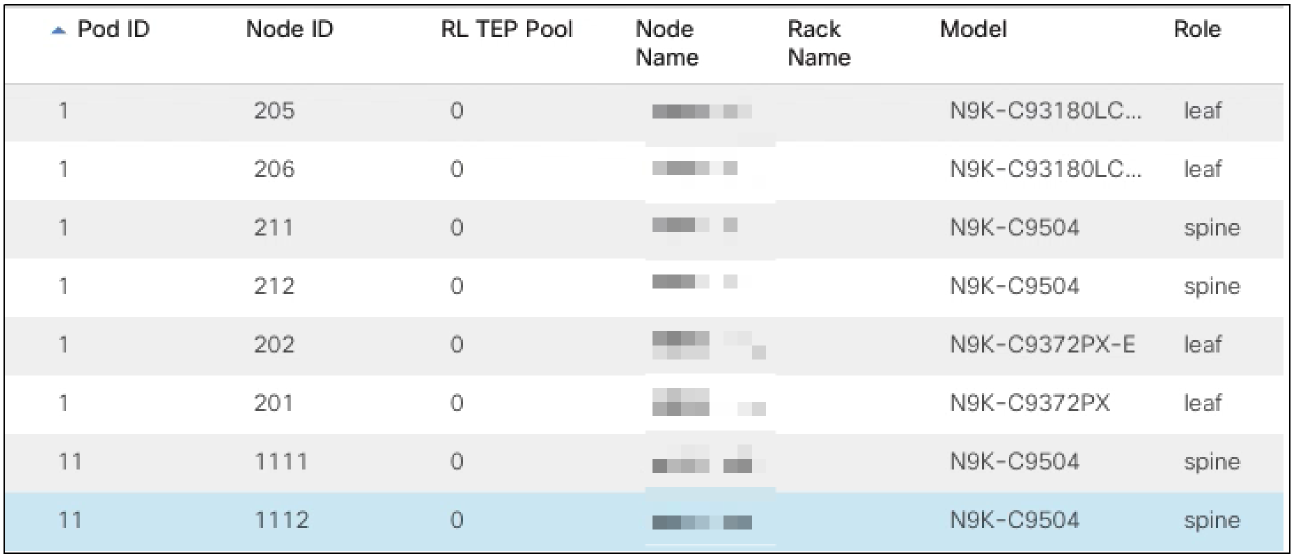
Serial Number	Pod ID	Node ID	RL TEP Pool	Node Name	Rack Name	Model	Role
[redacted]	1	205	0	[redacted]		N9K-C93180LC...	leaf
[redacted]	1	206	0	[redacted]		N9K-C93180LC...	leaf
[redacted]	1	0	0			N9K-C9504	spine
[redacted]	1	211	0	[redacted]		N9K-C9504	spine
[redacted]	1	0	0			N9K-C9504	spine
[redacted]	1	212	0	[redacted]		N9K-C9504	spine
[redacted]	1	202	0	[redacted]		N9K-C9372PX-E	leaf
[redacted]	1	201	0	[redacted]		N9K-C9372PX	leaf

Figure 4-20 Fabric Membership

4.4.5 Set up Pod and Node IDs for the Site 2 Spines

To set up Pod and Node IDs for the Site 2 Spines, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Inventory** → **Fabric Membership**, as shown in Figure 4-21.
2. In the main window, double-click and update Pod ID (11), Node ID (1111 and 1112), and Node Names for both the new Spines.



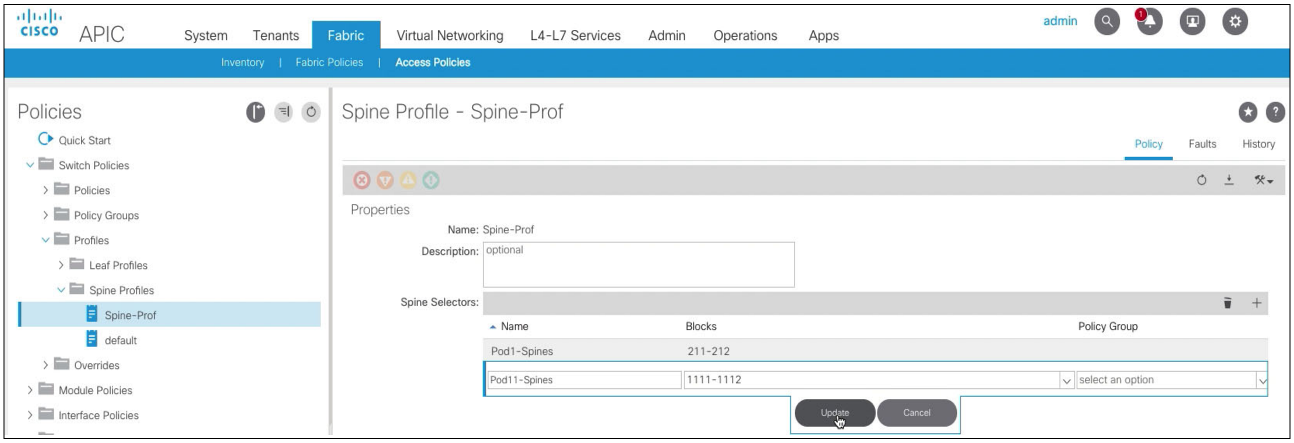
Pod ID	Node ID	RL TEP Pool	Node Name	Rack Name	Model	Role
1	205	0	[redacted]		N9K-C93180LC...	leaf
1	206	0	[redacted]		N9K-C93180LC...	leaf
1	211	0	[redacted]		N9K-C9504	spine
1	212	0	[redacted]		N9K-C9504	spine
1	202	0	[redacted]		N9K-C9372PX-E	leaf
1	201	0	[redacted]		N9K-C9372PX	leaf
11	1111	0	[redacted]		N9K-C9504	spine
11	1112	0	[redacted]		N9K-C9504	spine

Figure 4-21 Fabric membership

4.5 Site 2 Spine configuration

After adding the Spines to the Fabric, configure the Spines on Site 2 to correctly communicate with the IPN devices on Site 2. The leaf switches from Site 2 will not be visible unless this process is complete.

1. Log in to the APIC GUI and click **Fabric** → **Access Policies**.
2. From the left pane, expand **Switch Policies** → **Profiles** → **Spine Profiles**.
3. Select the previously created Spine Profile (**Spine-Prof**).
4. In the main window, click **+** to add additional (Site 2) Spines.
5. Provide a **Name** and select the Node IDs for the two Spines (1111-1112), as shown in Figure 4-22.



Spine Profile - Spine-Prof

Properties

Name: Spine-Prof

Description: optional

Spine Selectors:

Name	Blocks	Policy Group
Pod1-Spines	211-212	
Pod11-Spines	1111-1112	select an option

Figure 4-22 Spine Profiles

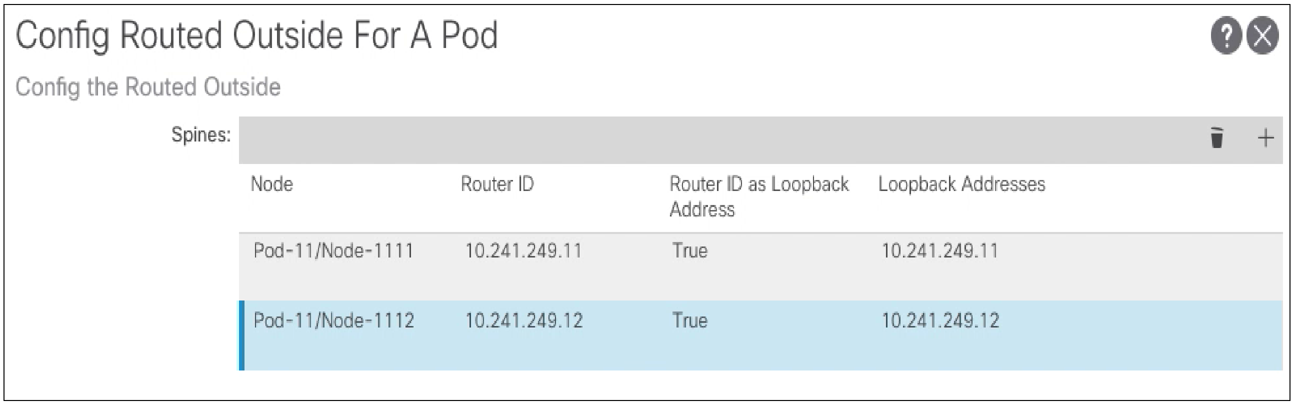
6. Click **Update.**

4.5.1 Create Routed Outside for Site 2 Spines

To create a Routed Outside for the Site 2 Spines, complete the following steps:

1. Log in to the APIC GUI and click **Fabric** → **Inventory**.
2. In the left pane, right-click **Pod Fabric Setup Policy** and select **Create Routed Outside for A Pod**.
1. Click **+** to add the first Spine.
2. Select the first Spine (**Pod-11/Node-1111**) and add **Router ID (Loopback)**, and click **Update**.
3. Click **+** to add second Spine.

4. Select the second Spine (**Pod-11/Node-1112**) and add **Router ID (Loopback)**, as shown in Figure 4-23, and click **Update**.



Config Routed Outside For A Pod

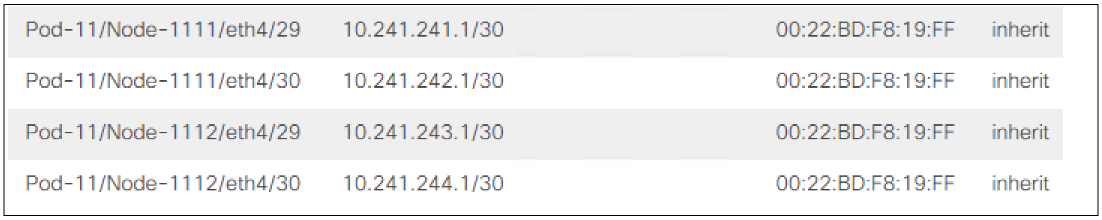
Config the Routed Outside

Spines:

Node	Router ID	Router ID as Loopback Address	Loopback Addresses
Pod-11/Node-1111	10.241.249.11	True	10.241.249.11
Pod-11/Node-1112	10.241.249.12	True	10.241.249.12

Figure 4-23   Config Routed Outside For A Pod

5. Click **+** to add Routed Sub-Interfaces.
6. Add all four interfaces (Path) and their respective IP addresses connecting the Spines to IPN devices, as shown in Figure 4-24.



Pod-11/Node-1111/eth4/29	10.241.241.1/30	00:22:BD:F8:19:FF	inherit
Pod-11/Node-1111/eth4/30	10.241.242.1/30	00:22:BD:F8:19:FF	inherit
Pod-11/Node-1112/eth4/29	10.241.243.1/30	00:22:BD:F8:19:FF	inherit
Pod-11/Node-1112/eth4/30	10.241.244.1/30	00:22:BD:F8:19:FF	inherit

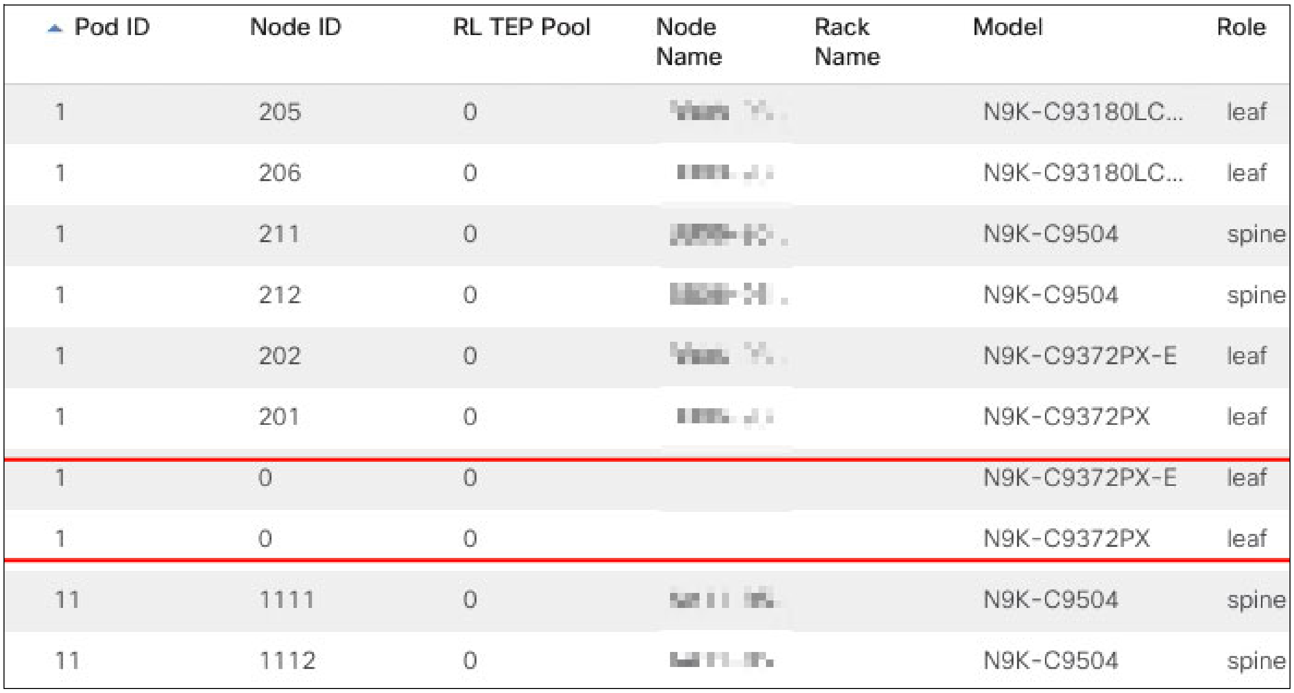
Figure 4-24   Add all interfaces

7. Click **Submit**.

The Site-2 spine configuration is now complete. Log in to the IPN devices to verify the OSPF routing and neighbor relationship.

4.6 Site 2 Leaf Discovery

With the Multi-Pod network configured on Site 2, all the Leaf switches on Site 2 should now be visible under **Fabric** → **Inventory** → **Fabric Membership.** In the main window, double-click and update Pod ID (11), Node ID, and Node Names for the leaf devices, as shown in Figure 4-25.



Pod ID	Node ID	RL TEP Pool	Node Name	Rack Name	Model	Role
1	205	0	[redacted]		N9K-C93180LC...	leaf
1	206	0	[redacted]		N9K-C93180LC...	leaf
1	211	0	[redacted]		N9K-C9504	spine
1	212	0	[redacted]		N9K-C9504	spine
1	202	0	[redacted]		N9K-C9372PX-E	leaf
1	201	0	[redacted]		N9K-C9372PX	leaf
1	0	0			N9K-C9372PX-E	leaf
1	0	0			N9K-C9372PX	leaf
11	1111	0	[redacted]		N9K-C9504	spine
11	1112	0	[redacted]		N9K-C9504	spine

Figure 4-25 Fabric Membership

With the discovery and addition of the Site 2 devices, the Multi-Pod portion of the configuration is now complete.

4.6.1 Configuring the Secondary Site to the VMM

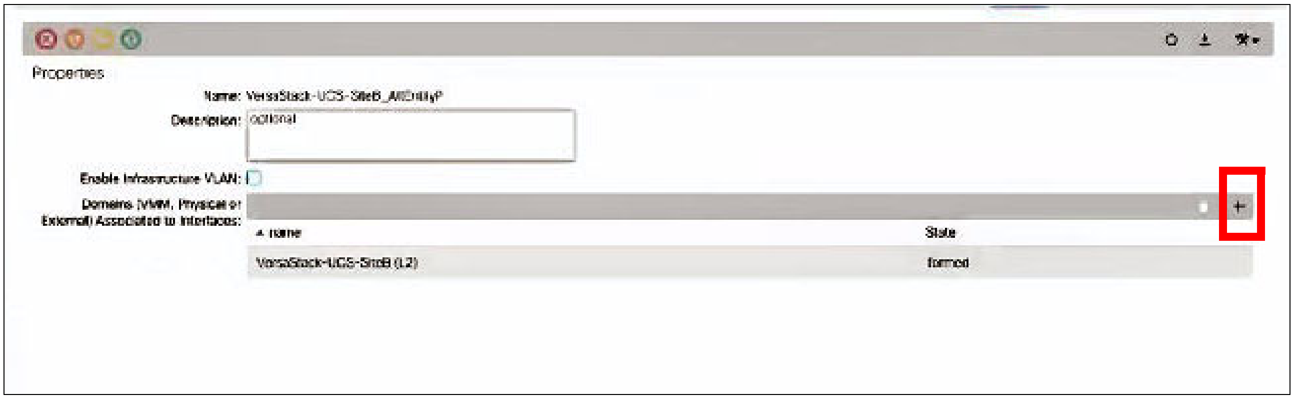
A new Virtual Machine Manager (VMM) domain for the vCenter will not be created for the second site. Instead, the existing VMM domain that associates the ACI fabric to the APIC generated VMware Distributed Switch (vDS) will be associated to the UCS AEP that was created for the secondary site UCS connection.

Note: The AEP for the second site UCS is automatically created during the APIC virtual port channel (vPC) creation wizard for connecting UCS Fabric Interconnects to the Nexus leafs that they are associated with.

The VMM can be associated from within the APIC GUI by completing the following steps:

1. Click the Fabric tab, and select **External Access Policies**.
2. Under External Access Policies in the left column, select **Policies** → **Global** → **Attachable Access Entity Profiles** and select the AEP for the secondary site UCS.

3. Click the **+** mark to the far right of the **Domains** option within Properties, as shown in Figure 4-26.



Properties

Name: VersaStack-UCS-SiteB_AttEntityP

Description: optional

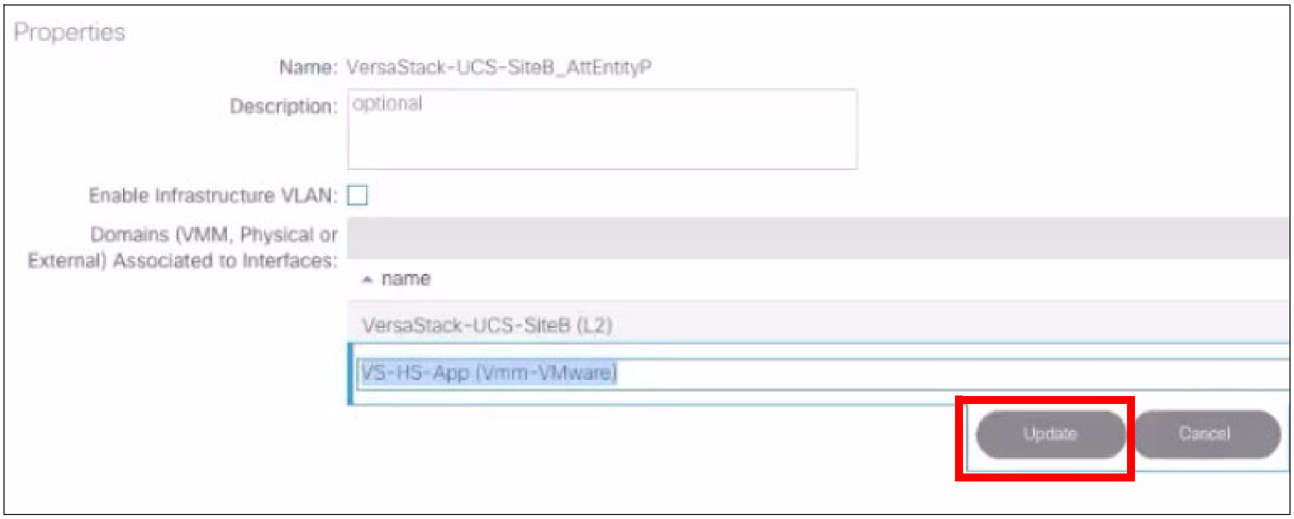
Enable Infrastructure VLAN: □

Domains (VMM, Physical or External) Associated to Interfaces:

▲ name	State
VersaStack-UCS-SiteB (L2)	formed

Figure 4-26 Domains option: Properties

4. Click **Continue** past any Policy Usage Warning windows that are displayed.
5. Select the **VMM f**rom the menu that appears within Domains.
6. Right-click **Networking** and select **Create Bridge Domain**.
7. Click **Update** to add the VMM, as shown in Figure 4-27.



Properties

Name: VersaStack-UCS-SiteB_AttEntityP

Description: optional

Enable Infrastructure VLAN: □

Domains (VMM, Physical or External) Associated to Interfaces:

▲ name
VersaStack-UCS-SiteB (L2)
VS-HS-App (Vmm-VMware)

Figure 4-27 Click Update

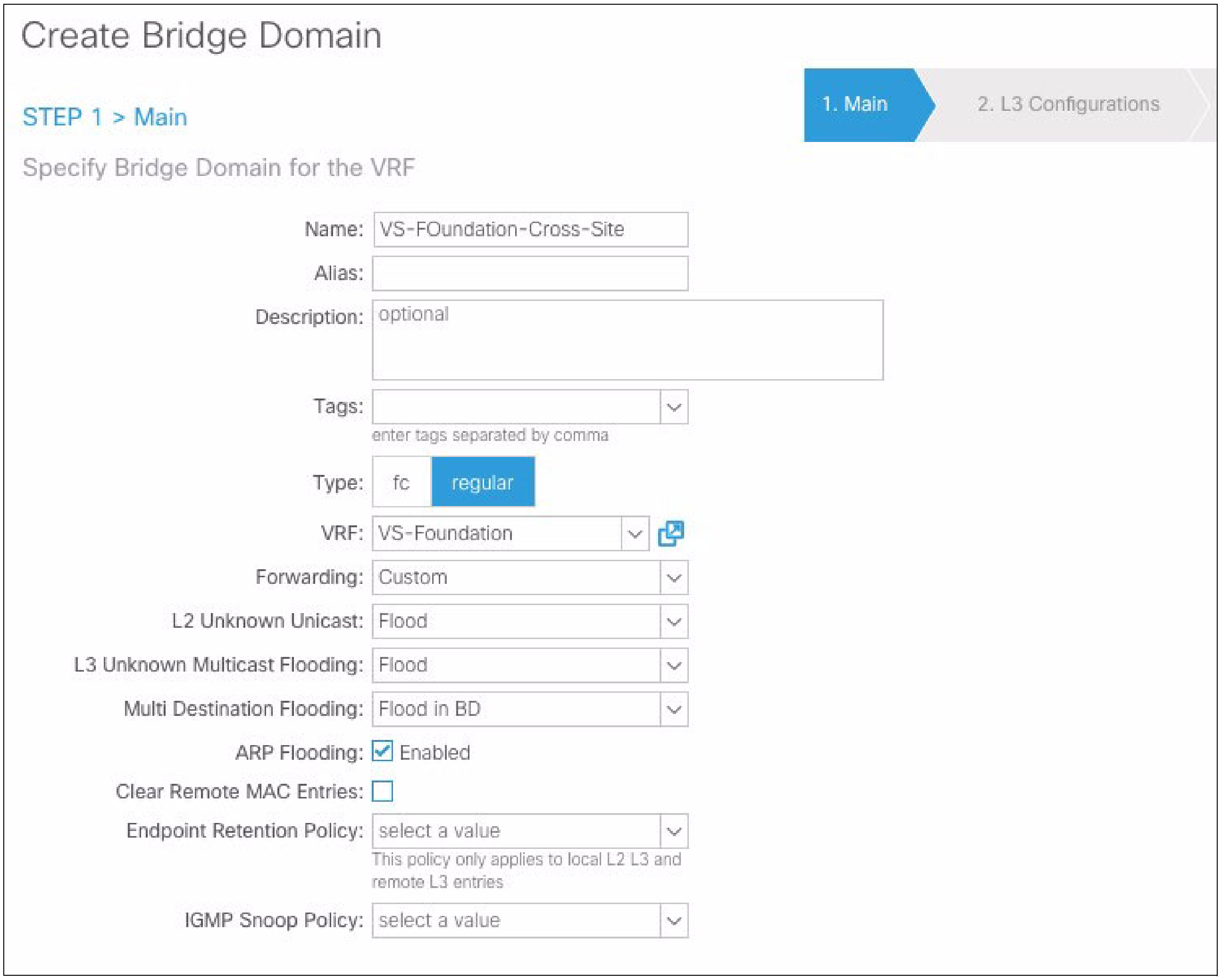
4.6.2 Create a Cross-Site Bridge Domain

The cross site VersaStack infrastructure Endpoint Groups (EPGs) will share a common Bridge Domain (BD). Optionally, there can be a dedicated BD per EPG, but a common BD was used in the validated environment.

To create the cross site BD, complete the following steps from within the APIC GUI:

1. Select the Tenants **t**ab, and select the **Tenant VS-Foundation** or another named infrastructure tenant of the VersaStack.
2. Right-click **Networking** and select **Create Bridge Domain**.
3. Provide a name for the cross site BD (`Foundation-Cross-Site`).

4. Select the **VRF** to be **VS-Foundation**.
5. Change **Forwarding** to **Custom**.
6. Change **L2 Unknown Unicast** to **Flood** and click **Next**, as shown in Figure 4-28.



Create Bridge Domain

STEP 1 > Main

Specify Bridge Domain for the VRF

Name: VS-FOundation-Cross-Site
Alias:
Description: optional
Tags:
enter tags separated by comma
Type: fc regular
VRF: VS-Foundation
Forwarding: Custom
L2 Unknown Unicast: Flood
L3 Unknown Multicast Flooding: Flood
Multi Destination Flooding: Flood in BD
ARP Flooding: ☑ Enabled
Clear Remote MAC Entries: □
Endpoint Retention Policy: select a value
This policy only applies to local L2 L3 and remote L3 entries
IGMP Snoop Policy: select a value

Figure 4-28 Create Bridge Domain

7. Leave Step 2 options as the defaults, and click **Next**.
8. Leave Step 3 options as the defaults, and click **Finish**.

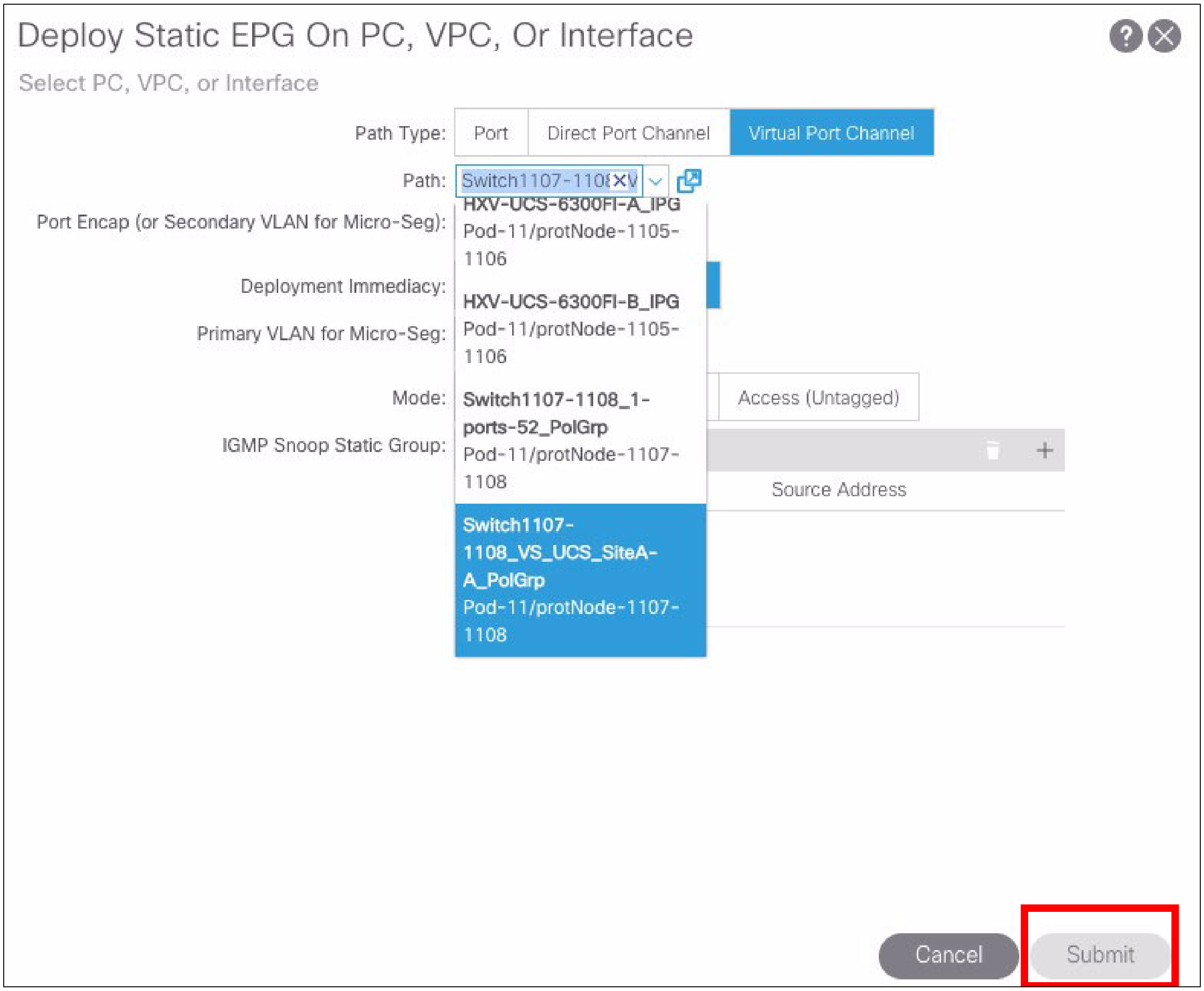
4.6.3 Configuring Cross-Site EPGs

As with the VMM configuration, the Application Profiles and EPGs that will be used across both sites have been created, except for an Application Profile. The following EPGs will be created, or have previously been created during the single site setup, within the VS-Foundation tenant:

- vMotion
- iSCSI-A
- iSCSI-B
- IB-Mgmt (site specific)
- MGMT-Stretch (cross site)

The EPGs can be configured from within the APIC GUI by completing the following steps (the example involves a vMotion EPG):

1. Select the Tenants tab, and select the **Tenant VS-Foundation** or another named infrastructure tenant of the VersaStack.
2. Within **Application Profiles**, select **Host-Connectivity** and select **vMotion EPG**. Host-Connectivity is the example Application Profile from the previous deployment guide.
3. Right-click **vMotion EPG** and select **Deploy Static EPG on PC, VPC, or Interface**.
4. Select the following options:
   - **Path Type**: **Virtual Port Channel**.
   - **Path**: ***Leaf Switch connection for site being added for the UCS-A Policy Group***.
   - **Port Encap**: ***<leave as VLAN>***, set Integer Value to vMotion VLAN.
   - **Deployment Immediacy**: **Immediate**.
5. Click **Submit**, as shown in Figure 4-29.



Deploy Static EPG On PC, VPC, Or Interface

Select PC, VPC, or Interface

Path Type: Port | Direct Port Channel | Virtual Port Channel

Path: Switch1107-1108

Port Encap (or Secondary VLAN for Micro-Seg):

Deployment Immediacy:

Primary VLAN for Micro-Seg:

Mode: Access (Untagged)

IGMP Snoop Static Group:

Source Address

HXV-UCS-6300FI-A_IPG
Pod-11/protNode-1105-1106

HXV-UCS-6300FI-B_IPG
Pod-11/protNode-1105-1106

Switch1107-1108_1-ports-52_PolGrp
Pod-11/protNode-1107-1108

Switch1107-1108_VS_UCS_SiteA-A_PolGrp
Pod-11/protNode-1107-1108

Figure 4-29   Deploy Static EPG on PC, VPC, or Interface

6. Repeat steps 3-5 to add the UCS-B Policy Group to the vMotion EPG.

7. Select **Policy** within the **vMotion EPG**, and select **General** within the **Policy** options.
8. Set the **Bridge Domain the VS-Foundation-Cross-Site BD** that was previously created.
9. Click **Submit**.

Repeat steps 1-9 for both iSCSI EPGs, and the appropriate IB-Mgmt EPG that can be extended across sites.

4.7 Fibre Channel SAN using Cisco MDS Switches

The following section assumes familiarity with general Fibre Channel (FC) SAN design and technologies. Typically, the SAN design has servers and storage that connect into dual, redundant fabrics. The tested configuration has a redundant fabric design that uses two Cisco MDS 9336S SAN switches at production site and Cisco MDS 9148S SAN switches at the DR Site. Each Cisco MDS SAN switch is equipped with FC 16 Gbps ports.

The IBM SVC HyperSwap system as tested was implemented in an ISL configuration that requires two types of SAN:

- Public SAN: In this type, server hosts, storage, and SVC nodes are connected. Data storage traffic traverses the Public SAN.
- Private SAN: Only the SVC node canisters connect into the Private SAN, which is used for cluster and replication traffic. The Virtual Fabric feature of the Cisco MDS Switches was used to implement segregated Private and Public SANs on the same chassis.

The following sections describe how the SAN was implemented.

4.8 Zoning configuration

This section describes the zoning configurations for Private SANs. The Public VSAN configuration remains the same as single site VersaStack architetcure. See the *VersaStack Design Guide* for information about Public VSAN configuration.

4.8.1 Private SAN

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

4.8.2 Public SAN

In an SVC Cluster HyperSwap configuration, the Public SAN is used for hosts, external storage, and quorum communication.

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

- Quorum zone: This zone contains all the SVC ports that connect to the fabric (on both logical switches) and all the ports of the external storage controller that connect to the fabric.
- Host zones: These zones contain at least one port for each IBM SVC node canister and only one host port (single initiator zone).
- Storage Zone: These zones contain all the SVC ports that connect to the fabric and all the FlashSystem 900 and IBM V5030 ports.

4.9 Cisco Multilayer Director Switch zone configuration

The following steps configure zoning for the worldwide port names (WWPNs) for setting up communication between SVC nodes across the sites using Cisco Multilayer Director Switch (MDS). WWPN information for various nodes can be easily collected by using the command **`show flogi database`**.

> **Note:** SAN switch configuration for communication between SAN Volume Controller (SVC) nodes and the FS900 and V5030 storage systems at each site is not covered in the following procedure. For more information, see the VersaStack with IBM SVC and Application Centric Infrastructure (ACI) design guides.

See Table 4-1 on page 34 to identify the ports where IBM nodes are connected to the MDS switches. In this configuration step, various zones are created to enable communication between all the IBM nodes. Take care when connecting two Cisco MDS switches to merge the fabrics. For information and guidelines when configuring multi-site SAN fabrics, see Zone Merge Behavior When Two MDS Switches Have Different Active Zoneset Names Are Connected.

4.9.1 Cisco MDS - A Switch

Log in to the MDS switch and complete the following steps:

1. Configure lower switch priority for replication VSAN on the primary site Cisco MDS switch, ensuring that the switch on the primary site is the principal switch:

   ```
   fcdomain priority 31 vsan 201
   ```

2. Configure inter-switch link (ISL) ports on switches at both sites. Depending on the number of links between the sites, use one of these procedures:

   a. To configure Port-Channel with multiple ISLs, issue these commands:

   ```
   switch(config)# interface fc1/x
   switch(config-if)# switchport trunk allowed vsan 201
   switch(config-if)# port-license acquire
   switch(config-if)# no shutdown
   ```

b. To configure Port-Channel with multiple ISLs, issue these commands:

```
switch(config)# int port-channel <9>
switch(config-if)# switchport trunk mode on
switch(config-if)# int fc 1/x
switch(config-if)# channel-group <9>
switch(config-if)# switchport trunk allowed vsan 201
```

3. Configure all the relevant ports (Table 4-1 on page 34) on Cisco MDS as follows:

```
interface fc1/x
  port-license acquire
  no shutdown
!
```

4. Create the VSAN and add all the ports from Table 4-1 on page 34:

```
vsan database
  vsan 201 interface fc1/x
  vsan 201 interface fc1/x
  <..>
```

5. The WWPNs obtained from **show flogi database** are used in this step. Replace the variables with actual WWPN values.

```
device-alias database
  device-alias name SVC-Clus-Node1-FC1 pwwn <Actual PWWN for Node1 FC1>
  device-alias name SVC-Clus-Node2-FC1 pwwn <Actual PWWN for Node2 FC1>
  device-alias name SVC-Clus-Node3-FC1 pwwn <Actual PWWN for Node3 FC1>
  device-alias name SVC-Clus-Node4-FC1 pwwn <Actual PWWN for Node4 FC1>
device-alias commit
```

6. Create the zones and add device-alias members for the SVC inter-node and SVC nodes to storage system configurations:

```
zone name Inter-Node-HS vsan 201
member device-alias SVC-Clus-Node1-FC1
member device-alias SVC-Clus-Node2-FC1
member device-alias SVC-Clus-Node3-FC1
member device-alias SVC-Clus-Node4-FC1
```

7. Add zones to the zoneset:

```
zoneset name crosssite vsan 201
    member Inter-Node-HS
```

8. Activate the zoneset:

```
zoneset activate name crosssite vsan 201
```

Note: Validate that all the host bus adapters (HBAs) are logged in to the MDS switch. The SVC nodes and storage systems should be powered on.

9. Validate that all the HBAs are logged in to the switch using the **show zoneset active** command:

```
VersaStack-SVC-FabA# show zoneset active

  zone name Inter-Node-HS vsan 201
  * fcid 0x400240 [pwwn 50:05:07:68:0c:11:93:c8] [SVC-Clus-Node1-FC1]
  * fcid 0x660800 [pwwn 50:05:07:68:0c:11:74:c9] [SVC-Clus-Node3-FC1]
  * fcid 0x400280 [pwwn 50:05:07:68:0c:11:93:c2] [SVC-Clus-Node2-FC1]
  * fcid 0x660b00 [pwwn 50:05:07:68:0c:21:74:c1] [SVC-Clus-Node4-FC1]
```

10. Save the configuration:

```
copy run start
```

4.9.2 Cisco MDS - B Switch

Log in to the MDS switch and complete the following steps:

1. Configure lower switch priority for replication VSAN on the primary site Cisco MDS switch, ensuring that the switch on the primary site is the principal switch:

   ```
   fcdomain priority 31 vsan 202
   ```

2. Configure ISL ports on switches at both sites. Depending on the number of links between the sites, use one of these procedures:

   a. To configure Port-Channel with multiple ISLs, issue these commands:

      ```
      switch(config)# interface fc1/x
      switch(config-if)# switchport trunk allowed vsan 202
      switch(config-if)# port-license acquire
      switch(config-if)# no shutdown
      ```

   b. To configure Port-Channel with multiple ISLs, issue these commands:

      ```
      switch(config)# int port-channel <9>
      switch(config-if)# switchport trunk mode on
      switch(config-if)# int fc 1/x
      switch(config-if)# channel-group <9>
      switch(config-if)# switchport trunk allowed vsan 202
      ```

3. Configure all the relevant ports (Table 4-1 on page 34) on Cisco MDS as follows:

   ```
   interface fc1/x
     port-license acquire
     no shutdown
   !
   ```

4. Create the VSAN and add all the ports from Table 4-1 on page 34:

   ```
   vsan database
     vsan 202 interface fc1/x
     vsan 202 interface fc1/x
     <..>
   ```

5. The WWPNs obtained from **show flogi database** are used in this step. Replace the variables with actual WWPN values.

   ```
   device-alias database
     device-alias name SVC-Clus-Node1-FC8 pwwn <Actual PWWN for Node1 FC8>
     device-alias name SVC-Clus-Node2-FC8 pwwn <Actual PWWN for Node2 FC8>
     device-alias name SVC-Clus-Node3-FC8 pwwn <Actual PWWN for Node3 FC8>
     device-alias name SVC-Clus-Node4-FC8 pwwn <Actual PWWN for Node4 FC8>
   device-alias commit
   ```

6. Create the zones and add device-alias members for the SVC inter-node and SVC nodes to storage system configurations:

   ```
   zone name Inter-Node-HS vsan 208
   member device-alias SVC-Clus-Node1-FC8
   member device-alias SVC-Clus-Node2-FC8
   member device-alias SVC-Clus-Node3-FC8
   member device-alias SVC-Clus-Node4-FC8
   ```

7. Add zones to the zoneset:

```
zoneset name crosssite vsan 202
    member Inter-Node-HS
```

8. Activate the zoneset:

```
zoneset activate name crosssite vsan 202
```

Note: Validate that all the HBAs are logged in to the MDS switch. The SVC nodes and storage systems should be powered on.

9. Validate that all the HBAs are logged in to the switch by using the **`show zoneset active`** command:

```
VersaStack-SVC-FabB# show zoneset active
zone name Inter-Node-HS vsan 202
  * fcid 0x770340 [pwwn 50:05:07:68:0c:24:93:c2] [SVC-Clus-Node1-FC8]
  * fcid 0x940500 [pwwn 50:05:07:68:0c:24:74:c9] [SVC-Clus-Node3-FC8]
  * fcid 0x770320 [pwwn 50:05:07:68:0c:24:93:c8] [SVC-Clus-Node2-FC8]
  * fcid 0x940600 [pwwn 50:05:07:68:0c:34:74:c1] [SVC-Clus-Node4-FC8]
```

10. Save the configuration:

```
copy run start
```

4.10 IBM SVC HyperSwap planning

Before you implement any IBM SVC cluster solution, perform a capacity planning exercise in terms of both physical resources and connectivity requirements. A good starting point is to collect as much workload information as possible. The workload analysis helps you to determine the number of I/O Groups (control enclosures) and the type and amount of storage to be virtualized.

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

For new SVC Cluster implementations, ensure that the attached devices, host, and levels of the I/O stack components (HBA firmware, device drivers, and multipathing) are supported. To do so, consult the System Storage Interoperation Center (SSIC).

Also, check the V7.8.x Supported Hardware List, Device Driver, Firmware, and Recommended Software Levels for IBM SVC Cluster.

Finally, check the full list of supported product list and documentation for SAN Volume Controller (2145, 2147).

The IBM SVC Cluster in the lab environment ran version 7.8.1.4. The Virtualized storage on FlashSystem 5030 also ran version 7.8.1.4.

4.11 Active-active Metro Mirror considerations

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

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

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

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

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

Example 4-5 Output of the lsrcrelationship command

```
IBM_2145:VersaStack-SVC:superuser>lsrcrelationship 0
id 0
name rcrel1
master_cluster_id 0000020320A044E2
master_cluster_name VersaStack-SVC
master_vdisk_id 6
master_vdisk_name vdisk4
aux_cluster_id 0000020320A044E2
aux_cluster_name VersaStack-SVC
aux_vdisk_id 0
aux_vdisk_name ESXi_iSCSI_5_Boot_Valid
primary master
consistency_group_id 0
consistency_group_name Boot_LUN
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type activeactive
cycling_mode
cycle_period_seconds 300
master_change_vdisk_id 27
```

```
master_change_vdisk_name vdisk6
aux_change_vdisk_id 8
aux_change_vdisk_name vdisk5
```

HyperSwap enforces the following rules when you define the active-active relationships:

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

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

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

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

With the SVC HyperSwap function, you can group multiple active-active Metro Mirror relationships together for high availability (HA) by creating Consistency Groups. The use of Consistency Groups is important where an application spans many volumes and requires that the data is consistent across more volumes. All of the relationships that belong to a Consistency Group must be consistent in terms of Master and Auxiliary site definition. In fact, when a relationship is added to anon-empty Consistency Group, it sets the Master and Auxiliary roles according to the replication direction of the Consistency Group.

HyperSwap Volumes in Consistency Groups all switch direction together. Therefore, the direction in which a set of active-active relationships in a Consistency Group replicates depends on which of the two sites has the most host I/O across all HyperSwap Volumes.

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

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

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

4.11.1 Quorum disk considerations

The quorum disk fulfills two functions for cluster reliability:

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

Starting with Spectrum Virtualize version 7.6, the quorum disk acting as a tiebreaker can be replaced with an IP quorum device, as described in 4.11.2, “IP Quorum” on page 69. An IBM SVC HyperSwap solution that does not use the IP Quorum feature requires an external controller to provide the active quorum disk.

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

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

However, enabling HyperSwap configuration with fewer than one quorum disk for each site is not recommended because it can seriously affect the operation during split-brain and rolling disaster scenarios.

To modify the quorum disk assignment, use either the CLI or GUI. The CLI output in Example 4-6 shows that the SVC Cluster initially automatically assigns the quorum disks.

Example 4-6 Modifying the quorum disk assignment

```
IBM_2145:VersaStack-SVC:superuser>lsquorum
quorum_index status id name   controller_id controller_name active object_type override site_id
site_name
0            online 3  mdisk1 4             controller1     no     mdisk       no       2
Secondary
1            online 10 mdisk8 0             FlashSystem 900 no     mdisk       no       1
Primary
3            online                                         yes    device      no
192.168.163.78/192.168.163.78
```

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

Example 4-7 Changing from automatic selection to manual selection

```
IBM_2145:VersaStack-SVC:superuser>chquorum -override yes -mdisk 2 2
```

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

Example 4-8 Setting the active quorum disk

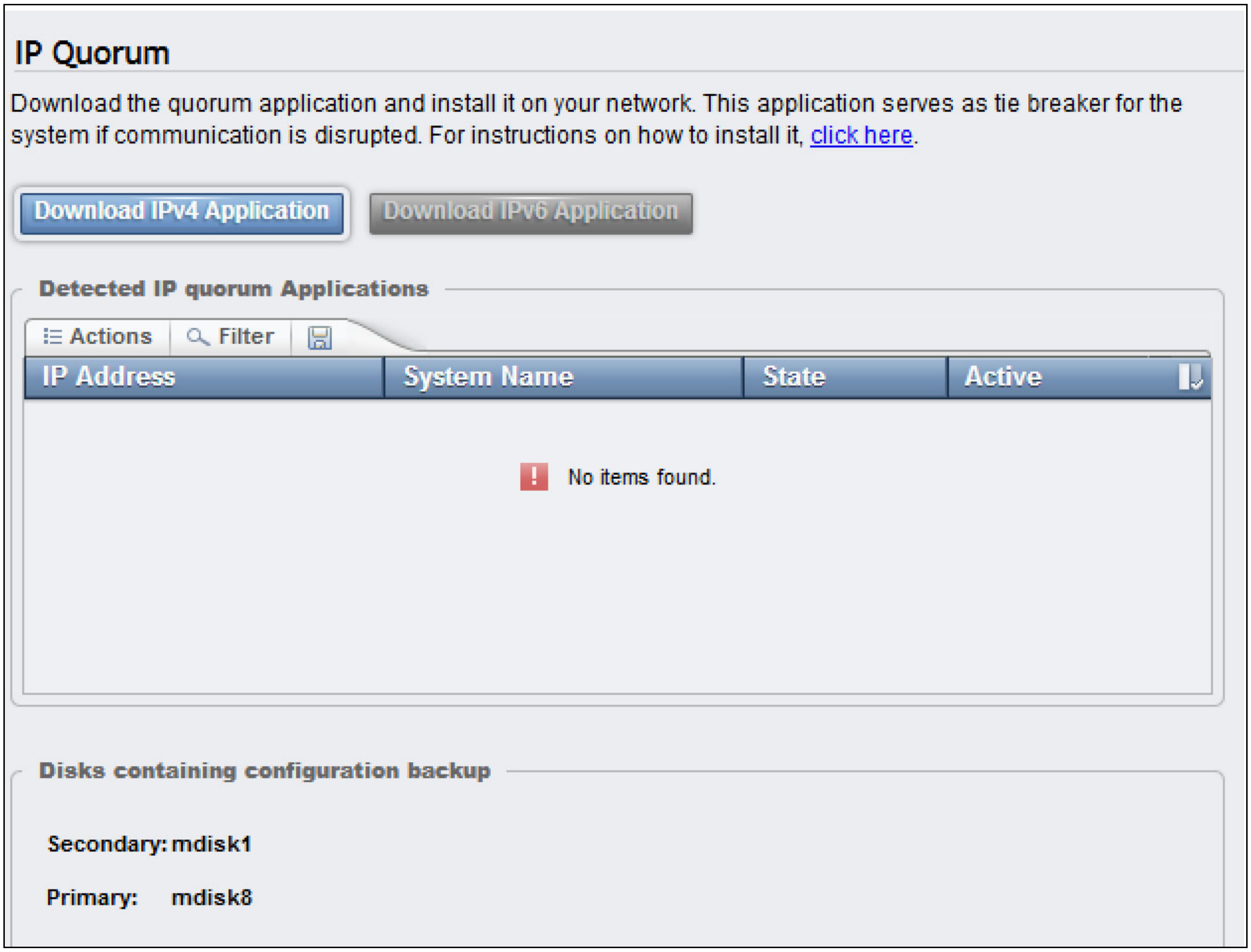
```
IBM_2145:VersaStack-SVC:superuser>chquorum -active 2
```

4.11.2 IP Quorum

IBM SVC firmware version 7.6 introduced the IP Quorum feature that eliminates the requirement for Fibre Channel networking and disk storage at third site. This feature deploys a Java application to a third site that acts as a tie-breaker in split-brain scenarios.

The Java application runs on a standard server and needs only standard network connectivity to be used.

Figure 4-30 shows the main page to download the IP Quorum file.



IP Quorum

Download the quorum application and install it on your network. This application serves as tie breaker for the system if communication is disrupted. For instructions on how to install it, click here.

Download IPv4 Application
Download IPv6 Application
Detected IP quorum Applications
Actions
Filter

IP Address	System Name	State	Active

No items found.

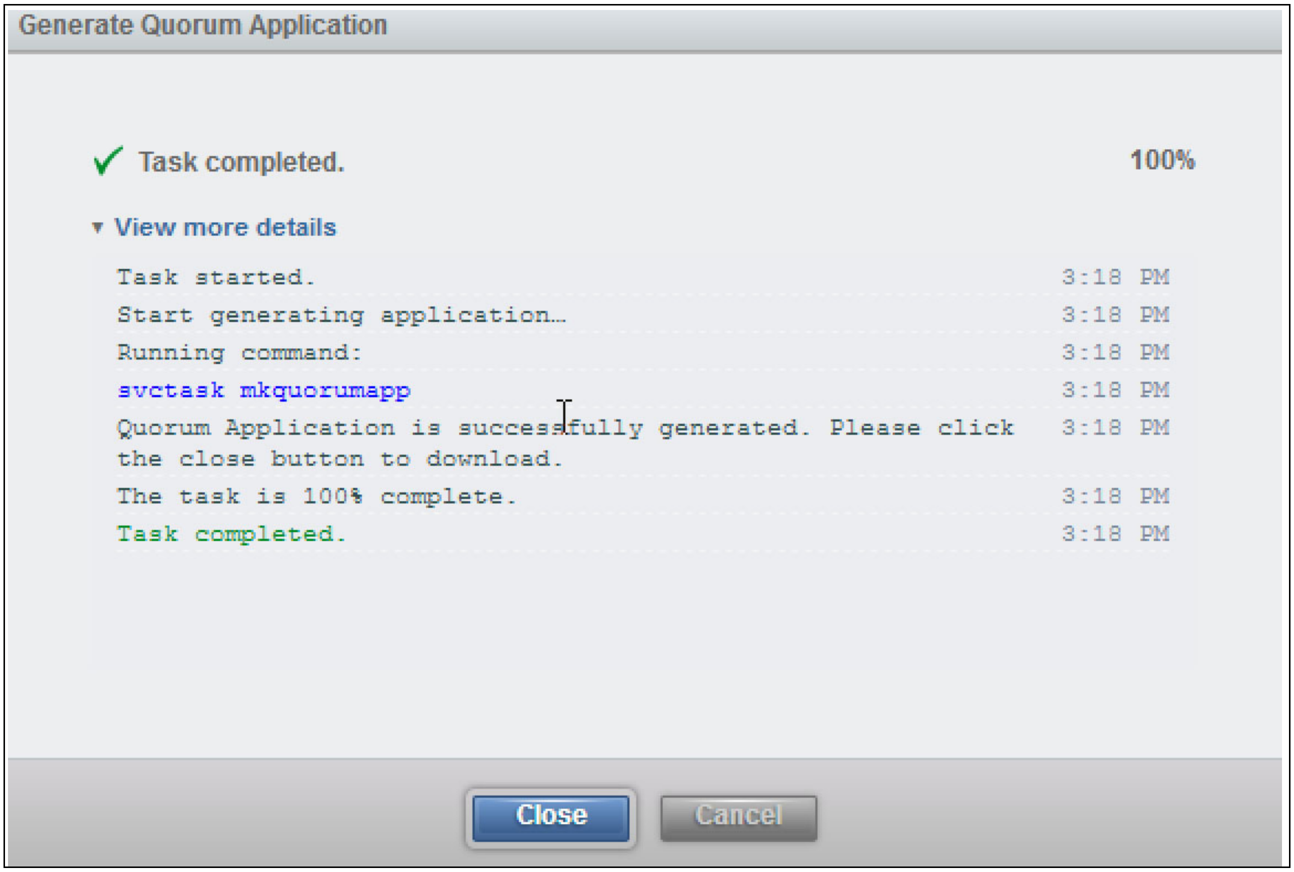
Disks containing configuration backup

Secondary: mdisk1

Primary: mdisk8

Figure 4-30 Main page to download IP Quorum file

After you select the option of IP Quorum, a wizard to run IP Quorum opens, as shown in Figure 4-31.



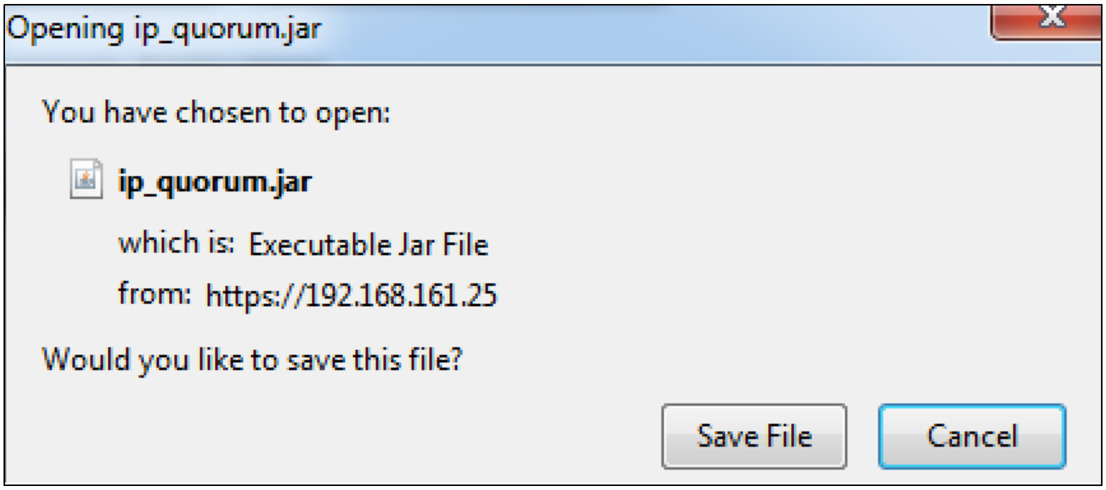
✓ Task completed. 100%

▾ View more details

```
Task started.                                                 3:18 PM
Start generating application…                                 3:18 PM
Running command:                                              3:18 PM
svctask mkquorumapp                                           3:18 PM
Quorum Application is successfully generated. Please click    3:18 PM
the close button to download.
The task is 100% complete.                                    3:18 PM
Task completed.                                               3:18 PM
```

Figure 4-31 Quorum application

Figure 4-31 shows the **mkquorumapp** command that creates the quorum file. This file, as shown in Figure 4-32, can be downloaded and installed at the server side to configure IP Quorum.



You have chosen to open:

ip_quorum.jar

which is: Executable Jar File

from: https://192.168.161.25

Would you like to save this file?

Figure 4-32 IP Quorum JAR file

This file can be installed on the Server that needs to be configured as the IP Quorum, as shown in Figure 4-33.

```
[root@localhost ~]# /opt/ibm/java-x86_64-80/bin/java -jar ip_quorum.jar &
[1] 2334
[root@localhost ~]# === IP quorum ===
Name set to null.
Successfully parsed the configuration, found 4 nodes.
Trying to open socket
Trying to open socket
Trying to open socket
Trying to open socket
Creating UID
Waiting for UID
Waiting for UID
Waiting for UID
*Connecting
*Connecting
*Connecting
Connected to 192.168.161.31
Connected to 192.168.161.32
Connected to 192.168.161.29
Connected to 192.168.161.30
```

Figure 4-33 IP Quorum installation

4.12 IBM SVC HyperSwap configuration

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

4.12.1 Configuring the SVC HyperSwap system topology using GUI

Spectrum Virtualize version 7.6 introduced GUI support for nonstandard topology systems such as Enhanced Stretched Cluster and HyperSwap topologies. To configure an SVC HyperSwap system topology using the GUI, complete the following steps:

1. To set up the HyperSwap configuration, click **Monitoring** → **System**, and then select **Actions** → **Modify System Topology**, as shown in Figure 4-34.

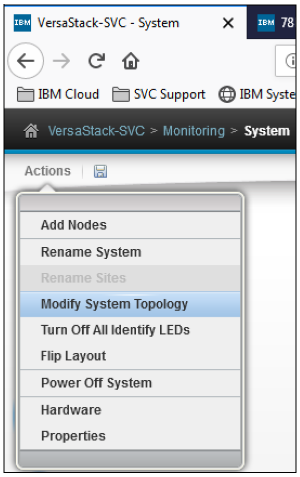
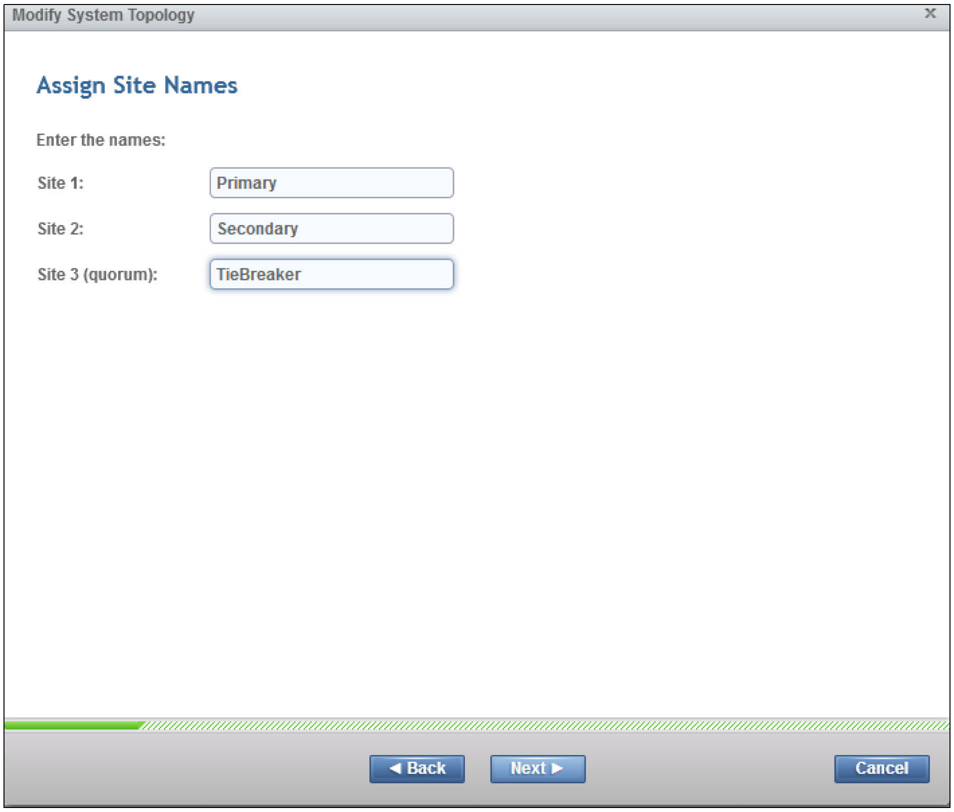


Figure 4-34 Modifying the system topology to configure HyperSwap

2. Configure site names such as primary, secondary, and Quorum Site (Tie Breaker), as shown in Figure 4-35. In this case, we are using IP Quorum for the active quorum.



Enter the names:

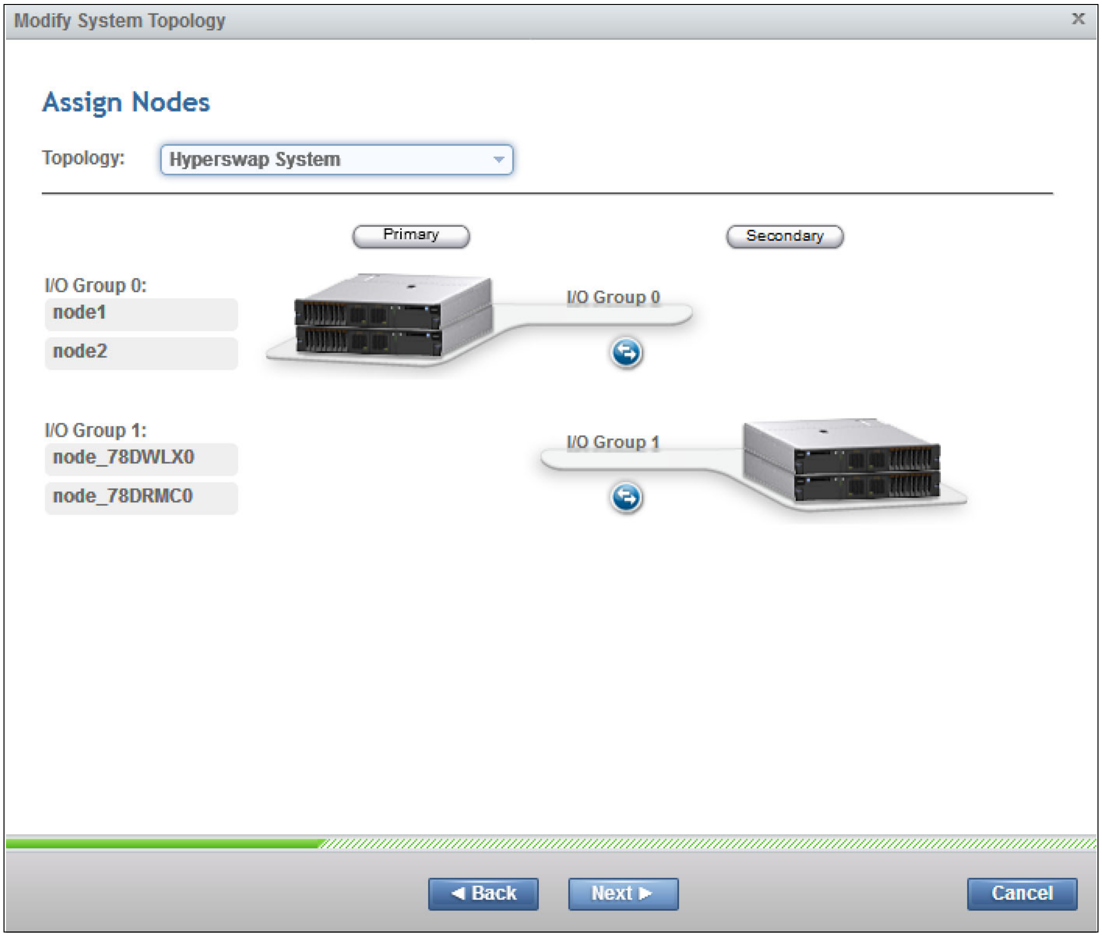
Site 1: Primary

Site 2: Secondary

Site 3 (quorum): TieBreaker

Figure 4-35 Configure site names

3. After site names are defined, assign the topology type and assign nodes to each site. Figure 4-36 shows selecting the HyperSwap topology and assigning nodes to each site.



Modify System Topology

Assign Nodes

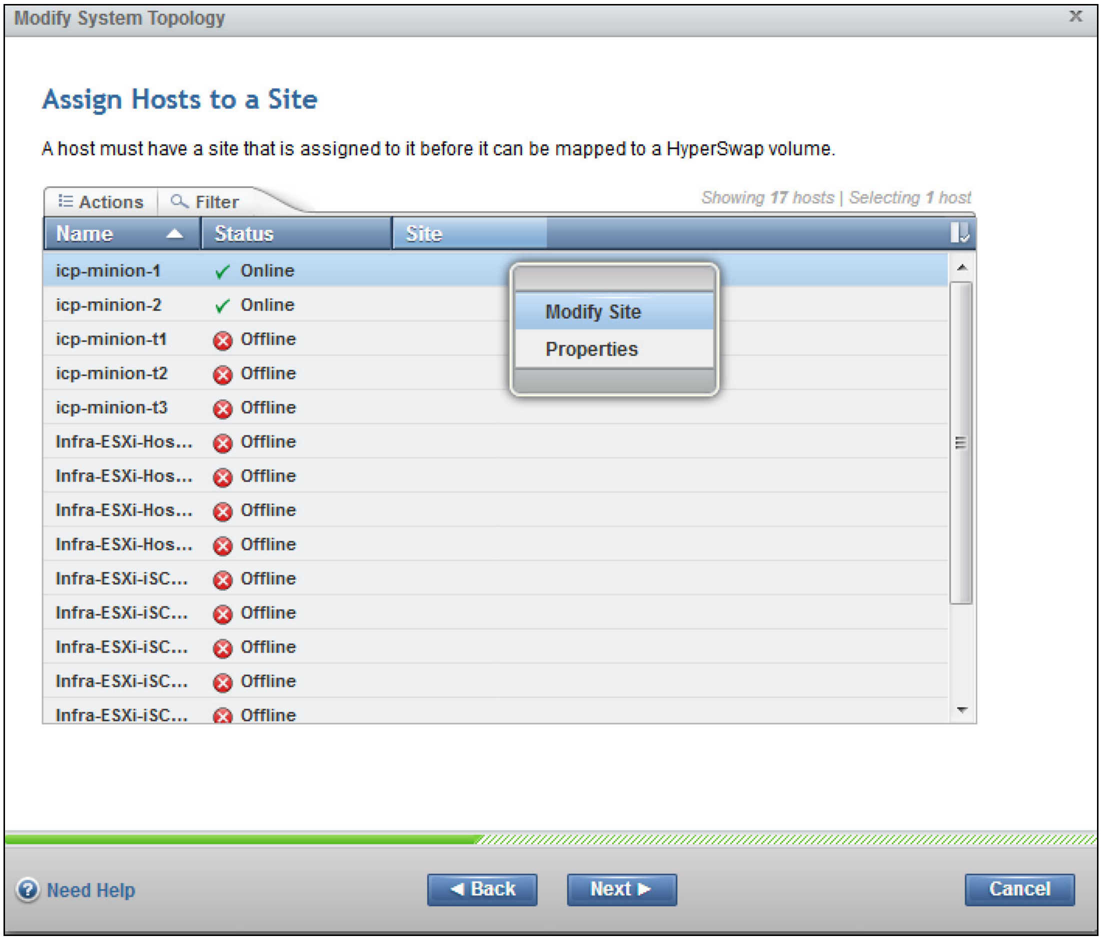
Topology: Hyperswap System

I/O Group 0:
node1
node2

I/O Group 1:
node_78DWLX0
node_78DRMC0

Figure 4-36 Assign nodes

4. Add a Host to the site in a HyperSwap configuration, as shown in Figure 4-37. Hosts need to be mapped to a site before HyperSwap volumes can be mapped to them.



Modify System Topology

Assign Hosts to a Site

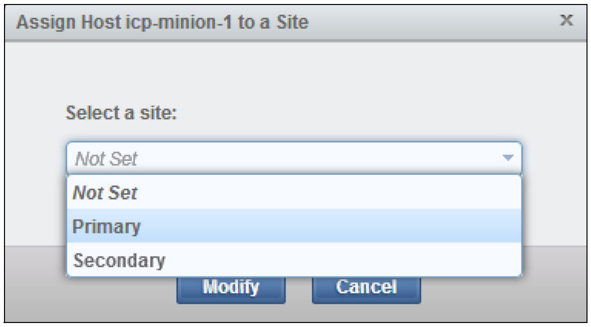
A host must have a site that is assigned to it before it can be mapped to a HyperSwap volume.

Showing 17 hosts | Selecting 1 host

Name	Status	Site
icp-minion-1	✓ Online	
icp-minion-2	✓ Online	
icp-minion-t1	☒ Offline	
icp-minion-t2	☒ Offline	
icp-minion-t3	☒ Offline	
Infra-ESXi-Hos...	☒ Offline	
Infra-ESXi-Hos...	☒ Offline	
Infra-ESXi-Hos...	☒ Offline	
Infra-ESXi-Hos...	☒ Offline	
Infra-ESXi-iSC...	☒ Offline	
Infra-ESXi-iSC...	☒ Offline	
Infra-ESXi-iSC...	☒ Offline	
Infra-ESXi-iSC...	☒ Offline	
Infra-ESXi-iSC...	☒ Offline	

Figure 4-37 Adding a Host to the site

5. Figure 4-38 shows the site options provided to assign to an individual host. Each host should be assigned to the site corresponding to their physical location.



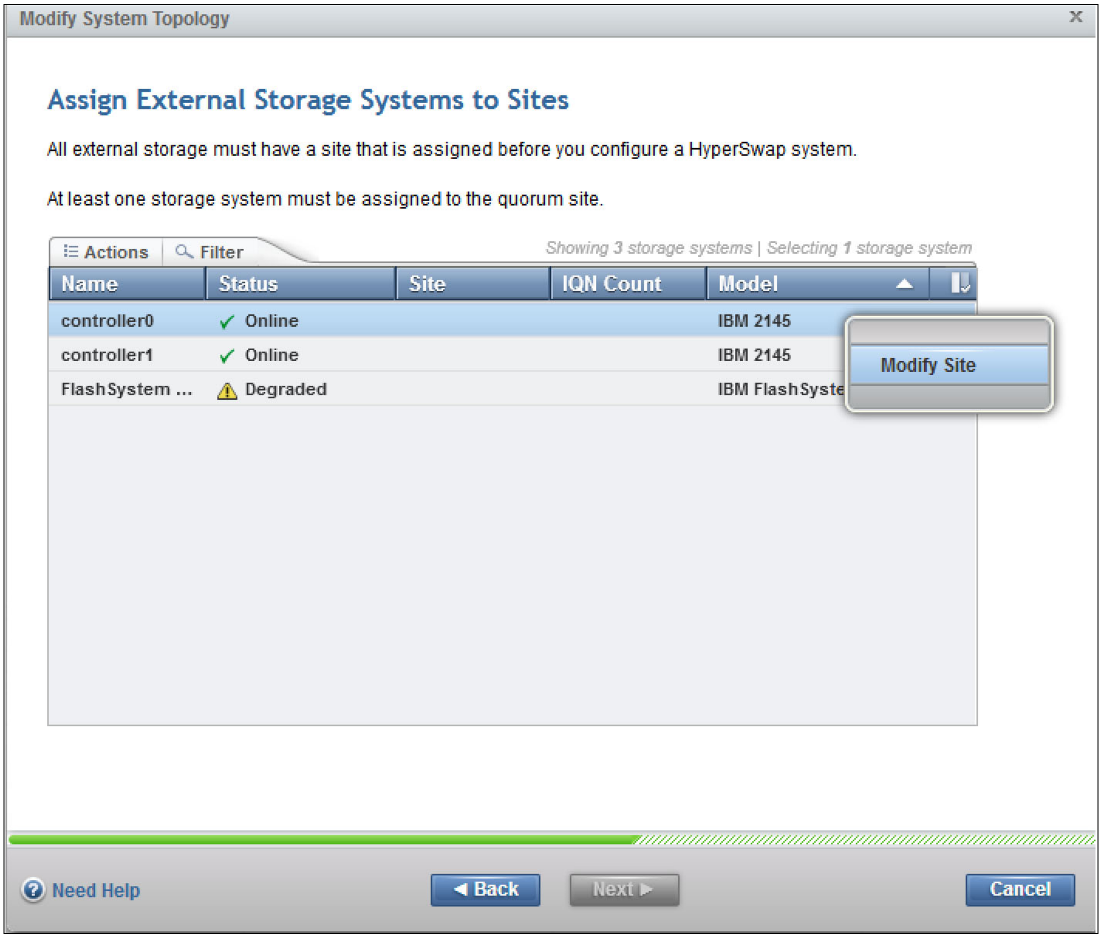
Assign Host icp-minion-1 to a Site

Select a site:

Not Set

Figure 4-38 Options for assigning host to a site

6. Figure 4-39 shows adding the controllers. As part of implementing HyperSwap, just like with hosts you need to configure sites for the controllers to provide site awareness.



Modify System Topology

Assign External Storage Systems to Sites

All external storage must have a site that is assigned before you configure a HyperSwap system.

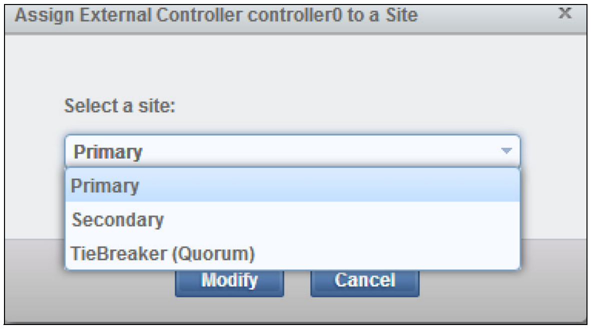
At least one storage system must be assigned to the quorum site.

Showing 3 storage systems | Selecting 1 storage system

Name	Status	Site	IQN Count	Model
controller0	✓ Online			IBM 2145
controller1	✓ Online			IBM 2145
FlashSystem ...	⚠ Degraded			IBM FlashSyste

Figure 4-39 Modify system topology

7. Figure 4-40 shows the site options that you can assign to an individual controller. Assign each controller to the site corresponding to their physical location.

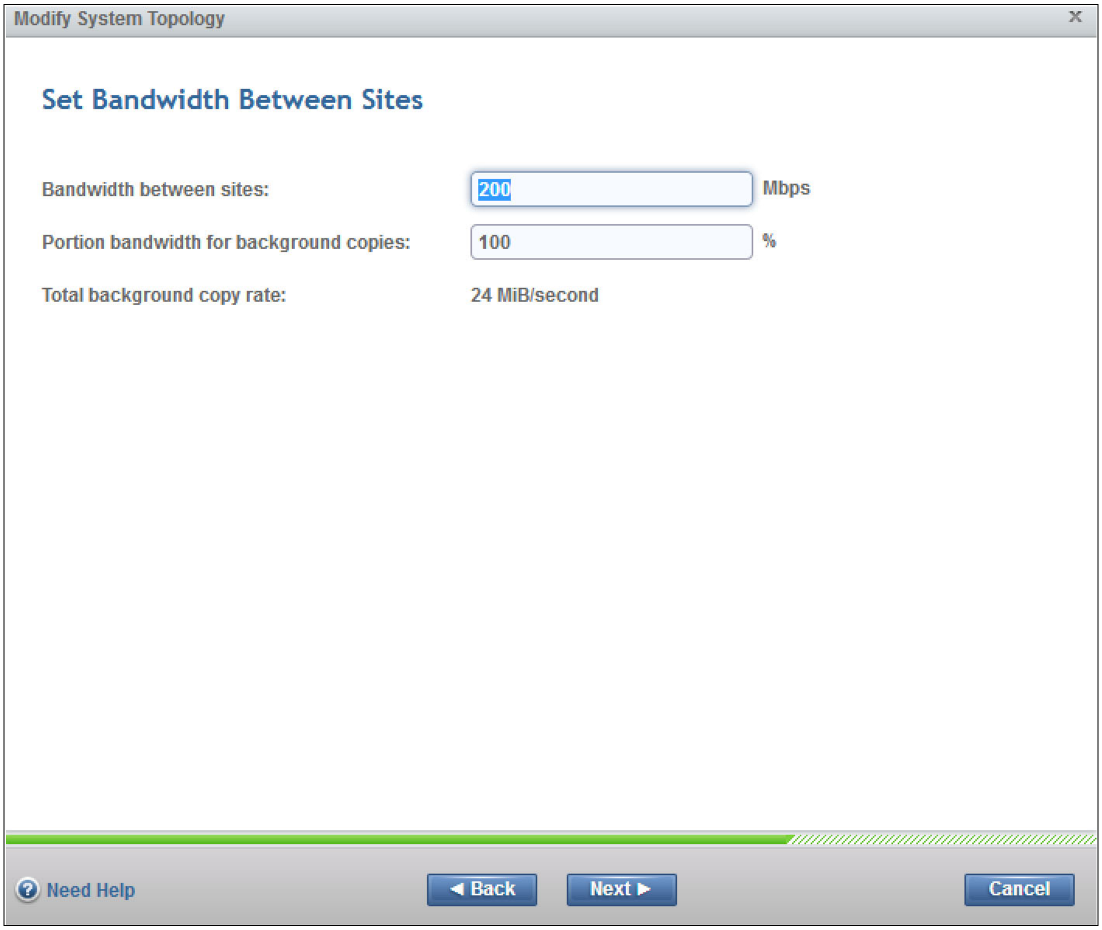


Assign External Controller controller0 to a Site

Select a site: Primary

Figure 4-40 Site options provided to assign to an individual controller

8. Figure 4-41 shows the columns for bandwidth and background copy rates for the HyperSwap Configuration. The settings for these parameters should be in line with your environment. The link bandwidth should be the same as the bandwidth available between sites on the FC SAN.



Set Bandwidth Between Sites

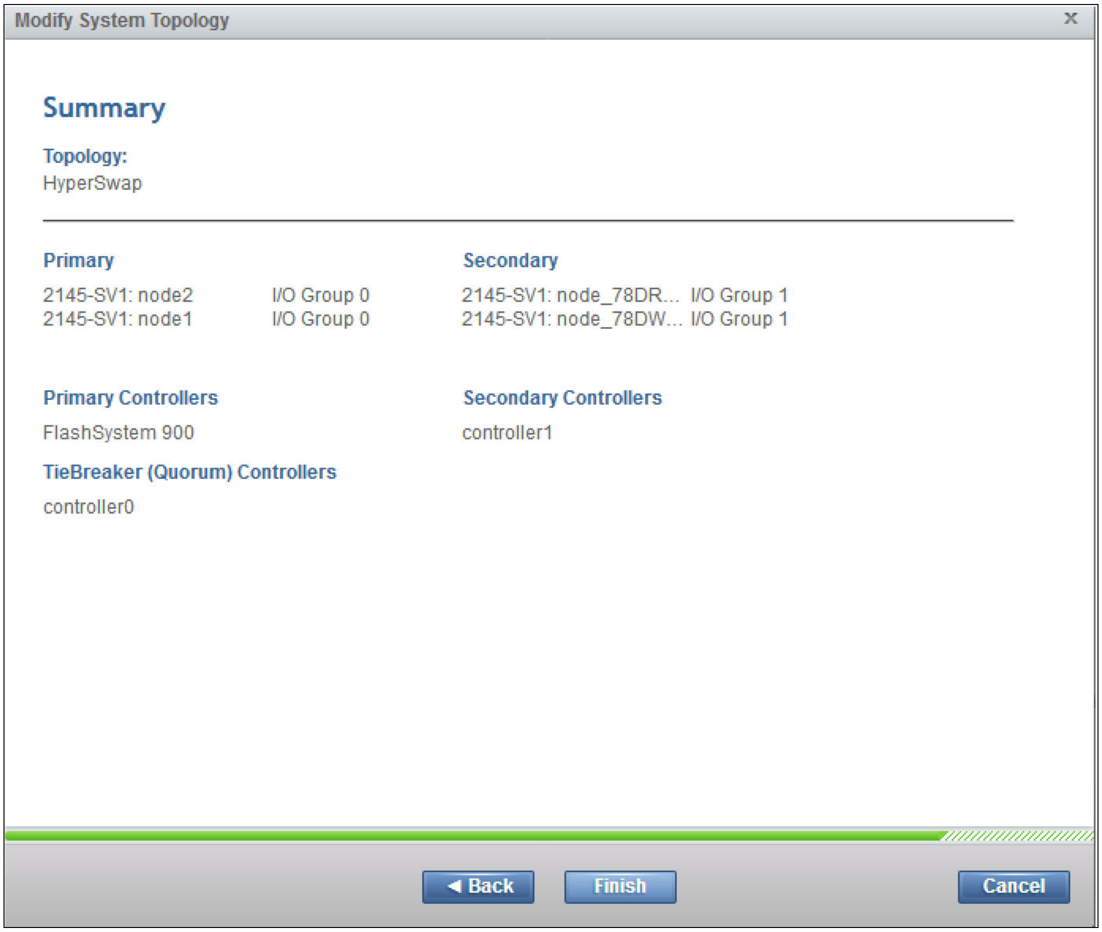
Bandwidth between sites: 200 Mbps

Portion bandwidth for background copies: 100 %

Total background copy rate: 24 MiB/second

Figure 4-41 Setting the bandwidth

9. Figure 4-42 shows the details of the HyperSwap configuration to be reviewed before committing the topology change.



Modify System Topology

Summary

Topology:
HyperSwap

Primary		Secondary	
2145-SV1: node2	I/O Group 0	2145-SV1: node_78DR...	I/O Group 1
2145-SV1: node1	I/O Group 0	2145-SV1: node_78DW...	I/O Group 1

Primary Controllers
FlashSystem 900

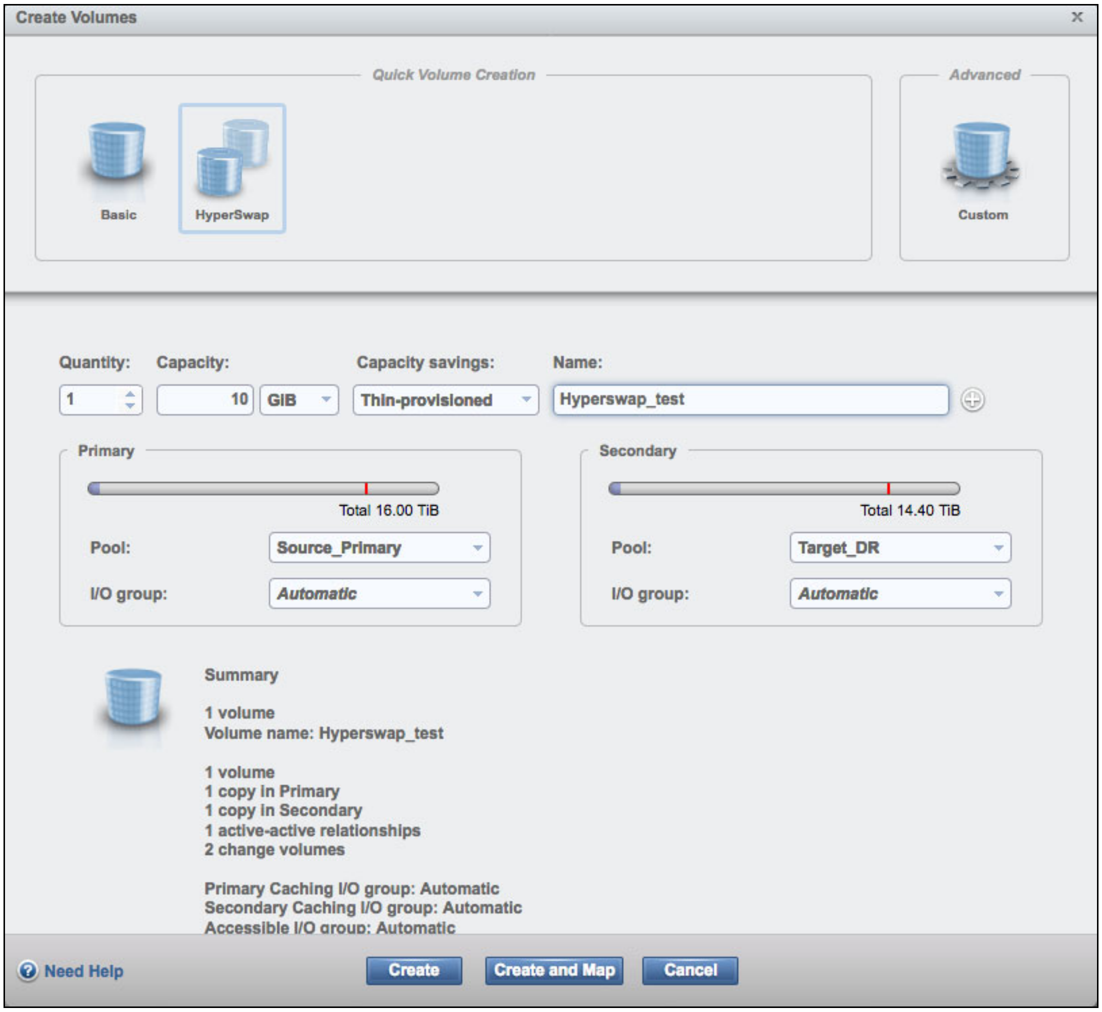
Secondary Controllers
controller1

TieBreaker (Quorum) Controllers
controller0

Figure 4-42 Summary

4.12.2 Creating HyperSwap Volumes using GUI

Figure 4-43 shows creating a HyperSwap Volume. After the HyperSwap Configuration has been completed, the Create Volume tab shows **HyperSwap Volume** as an option. After the **HyperSwap Volume** option is selected, the details of the volumes or both primary and secondary site are displayed. HyperSwap volumes with the Generic or Thin provisioning feature can be created and mapped to the hosts.



Quantity: 1
Capacity: 10 GiB
Capacity savings: Thin-provisioned
Name: Hyperswap_test

Primary
Total 16.00 TiB
Pool: Source_Primary
I/O group: Automatic

Secondary
Total 14.40 TiB
Pool: Target_DR
I/O group: Automatic

Summary

1 volume
Volume name: Hyperswap_test

1 volume
1 copy in Primary
1 copy in Secondary
1 active-active relationships
2 change volumes

Primary Caching I/O group: Automatic
Secondary Caching I/O group: Automatic
Accessible I/O group: Automatic

Figure 4-43 Creating volumes

4.12.3 Converting standard volumes into HyperSwap volumes

Figure 4-44 shows converting a standard volume (Existing Volumes) to a HyperSwap volume in a HyperSwap configuration. Right-click the volume and select **Add a volume copy** to the target site.

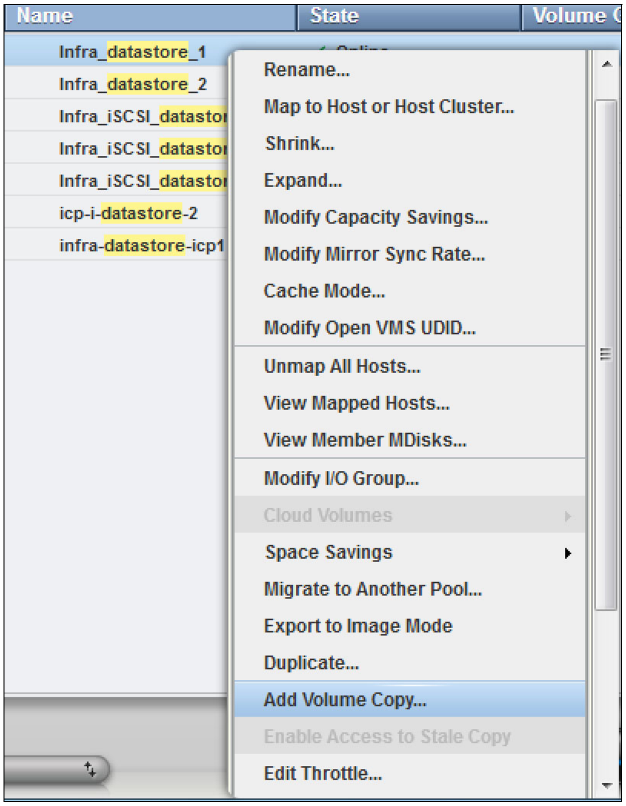
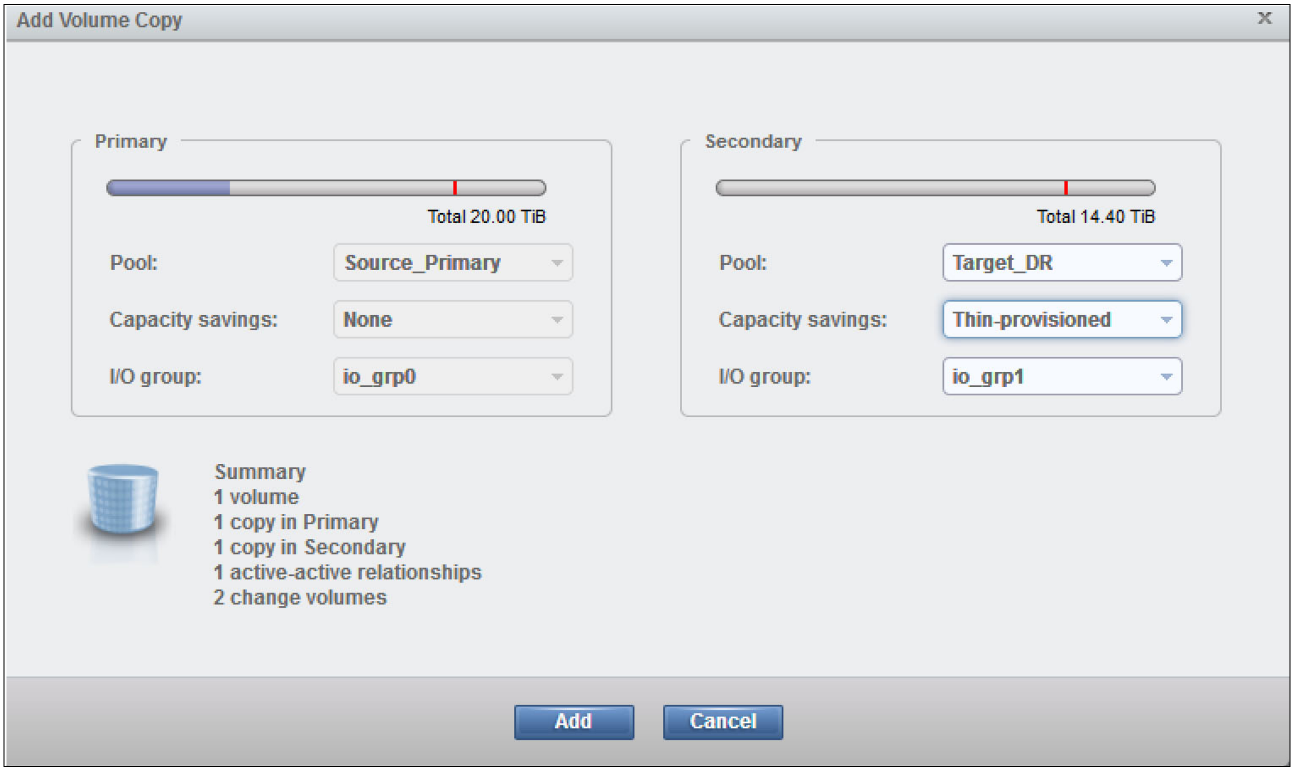


Figure 4-44 Add a volume copy

A wizard opens that prompts you to specify the pool capacity savings and iogrp that you want to add the volume copy to, as shown in Figure 4-45. After you are finished, click **Add**.



Add Volume Copy

Primary

Total 20.00 TiB

Pool: Source_Primary

Capacity savings: None

I/O group: io_grp0

Secondary

Total 14.40 TiB

Pool: Target_DR

Capacity savings: Thin-provisioned

I/O group: io_grp1

Summary
1 volume
1 copy in Primary
1 copy in Secondary
1 active-active relationships
2 change volumes

Figure 4-45 Add volume copy

5

SVC HyperSwap diagnostic and recovery guideline

This chapter addresses IBM Spectrum Virtualize HyperSwap diagnostic and recovery guidelines. These features help you understand what is happening in your Spectrum Virtualize HyperSwap environment after a critical event. This knowledge is crucial when you decide to alleviate the situation. You might decide to wait until the failure in one of the two sites is fixed, or declare a disaster and start the recovery action.

All of the operations that are described are guidelines to help you in a critical event or a rolling disaster. Several of the operations are specific to our lab environment, and several of the operations are common with every Spectrum Virtualize HyperSwap installation.

Before you start a recovery action after a disaster is declared, it is important that you be familiar with all of the recovery steps. We strongly suggest testing and documenting a recovery plan that includes all of the tasks that you must perform, according to the design and configuration of your environment. It is also best to execute the recovery action with IBM Support engaged.

This chapter includes the following sections:

- Solution recovery planning
- Failure testing
- Other failure cases

© Copyright IBM Corp. 2018. All rights reserved.

5.1 Solution recovery planning

In the context of the Spectrum Virtualize HyperSwap environment, solution recovery planning is more application-oriented. Therefore, any plan must be made with the client application's owner. In every IT environment, when a business continuity or disaster recovery (DR) solution is designed, incorporate a solution recovery plan into the process.

High availability (HA) and disaster recovery are facets of business continuity:

- High availability attempts to keep business applications and information technology available for as high a percentage of time as possible. High availability strategies include avoiding disaster scenarios in the first place and continuing application availability despite hardware or software failures, or natural disasters.
- Disaster recovery is the process of recovering from an information technology failure caused by a hardware or software failure, or natural disaster. Although not every information technology environment is considered highly available, disaster recovery occurs after high availability has failed.

Therefore, high availability is a strategy to prevent an outage from occurring, whereas disaster recovery is the process of recovering from an outage.

The HyperSwap HA function in the IBM Spectrum Virtualize software allows business continuity in a hardware failure, power failure, connectivity failure, or disasters such as fire or flooding.

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

5.1.1 Critical event scenarios and complete site or domain failure

Many critical event scenarios can occur in a Spectrum Virtualize HyperSwap environment. Certain events can be handled by using standard (*business as usual*) recovery procedures. A business as usual recovery usually entails an application or storage environment that is still up and running, but perhaps in a non-redundant state due to hardware or software failure, or a disaster taking out one site. Business processes run as usual while repairs are made to return the environment to a redundant state. This section addresses all of the required operations to recover from a *complete site failure*.

Certain parts of the recovery depend on the environment design. This section lists the actions to diagnose the situation and to recover the Spectrum Virtualize HyperSwap. However, most of the steps are basic and can be used in every environment and configuration.

> **Important:** Because of the high importance of successful recovery, do not improvise these actions. Instead, perform all steps with the assistance of IBM Support.

The following list includes several scenarios that you might face and their required recovery actions:

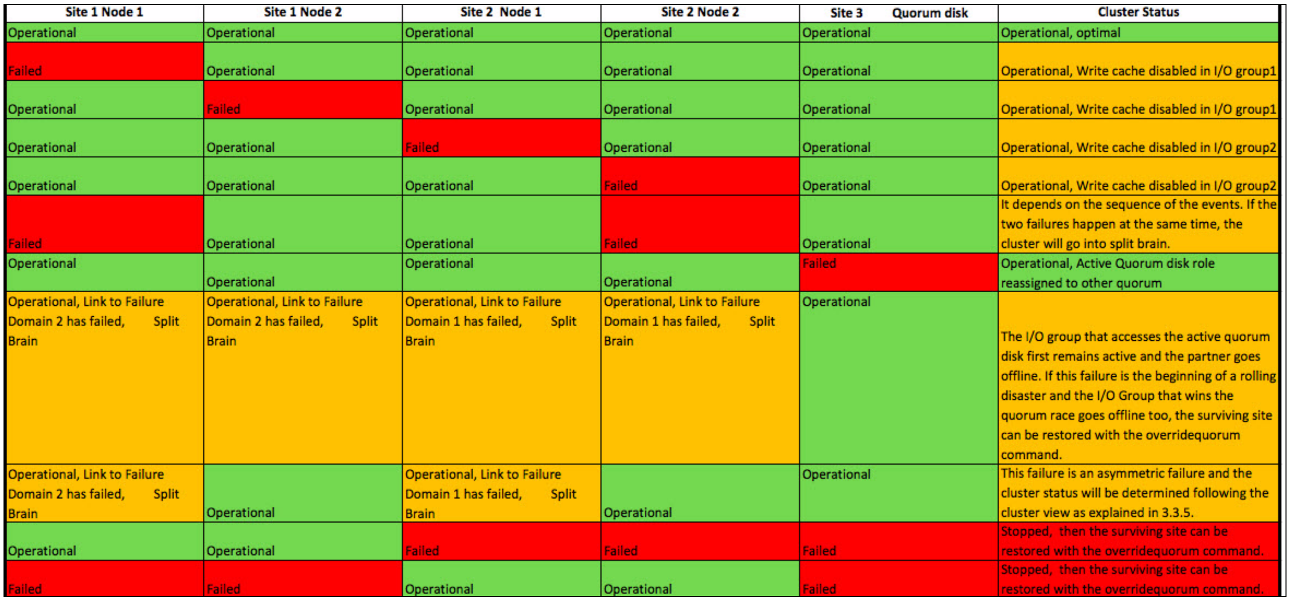
- Back-end storage box failure in one failure domain: Only possible if you are virtualizing the external storage controller. Use business as usual recovery provided by Spectrum Virtualize HyperSwap active-active Metro Mirror.
- Partial SAN failure in one failure domain: Use business as usual recovery because of SAN resilience.

- Total SAN failure in one failure domain: Use business as usual recovery because of SAN resilience, but pay attention to the performance impact. You need to act to minimize the impact on applications.
- Spectrum Virtualize HyperSwap node failure in one failure domain: Use business as usual recovery because of SVC HyperSwap high availability.

> **Note:** The recovery actions of the underlying storage environment will vary in every event. Therefore, engage IBM Support before taking any system recovery actions on any IBM Storage subsystems.

5.2 Failure testing

This section highlights the failure testing and results that we conducted in the test environment. Generally speaking, the failure scenarios in HyperSwap can be summarized as shown in Figure 5-1.

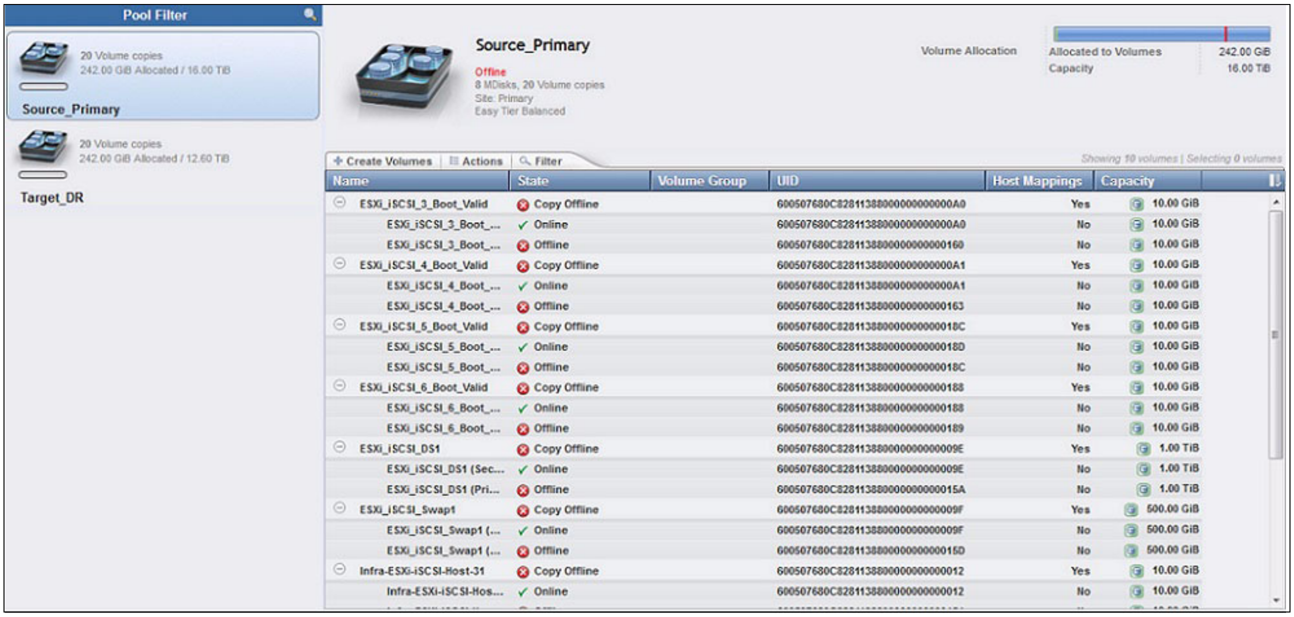


Site 1 Node 1	Site 1 Node 2	Site 2 Node 1	Site 2 Node 2	Site 3 Quorum disk	Cluster Status
Operational	Operational	Operational	Operational	Operational	Operational, optimal
Failed	Operational	Operational	Operational	Operational	Operational, Write cache disabled in I/O group1
Operational	Failed	Operational	Operational	Operational	Operational, Write cache disabled in I/O group1
Operational	Operational	Failed	Operational	Operational	Operational, Write cache disabled in I/O group2
Operational	Operational	Operational	Failed	Operational	Operational, Write cache disabled in I/O group2
Failed	Operational	Operational	Failed	Operational	It depends on the sequence of the events. If the two failures happen at the same time, the cluster will go into split brain.
Operational	Operational	Operational	Operational	Failed	Operational, Active Quorum disk role reassigned to other quorum
Operational, Link to Failure Domain 2 has failed, Split Brain	Operational, Link to Failure Domain 2 has failed, Split Brain	Operational, Link to Failure Domain 1 has failed, Split Brain	Operational, Link to Failure Domain 1 has failed, Split Brain	Operational	The I/O group that accesses the active quorum disk first remains active and the partner goes offline. If this failure is the beginning of a rolling disaster and the I/O Group that wins the quorum race goes offline too, the surviving site can be restored with the overridequorum command.
Operational, Link to Failure Domain 2 has failed, Split Brain	Operational	Operational, Link to Failure Domain 1 has failed, Split Brain	Operational	Operational	This failure is an asymmetric failure and the cluster status will be determined following the cluster view as explained in 3.3.5.
Operational	Operational	Failed	Failed	Failed	Stopped, then the surviving site can be restored with the overridequorum command.
Failed	Failed	Operational	Operational	Failed	Stopped, then the surviving site can be restored with the overridequorum command.

Figure 5-1 HyperSwap failure scenarios overview

5.2.1 Loss of access to Mdisks

The SVC primary site has an IBM FS900 as external storage controller. All of the storage in this site is backed by the FS900. To simulate a failure of the storage controller, we shut down the FS900. Figure 5-2 shows the result of this failure from the SVC perspective. A single volume copy from the FS900 pool is offline but the volume as a whole is still online and accessible through the redundant copy on the secondary site.



Pool Filter

20 Volume copies
242.00 GiB Allocated / 16.00 TiB
Source_Primary

20 Volume copies
242.00 GiB Allocated / 12.60 TiB
Target_DR

Source_Primary
Offline
8 MDisks, 20 Volume copies
Site: Primary
Easy Tier Balanced

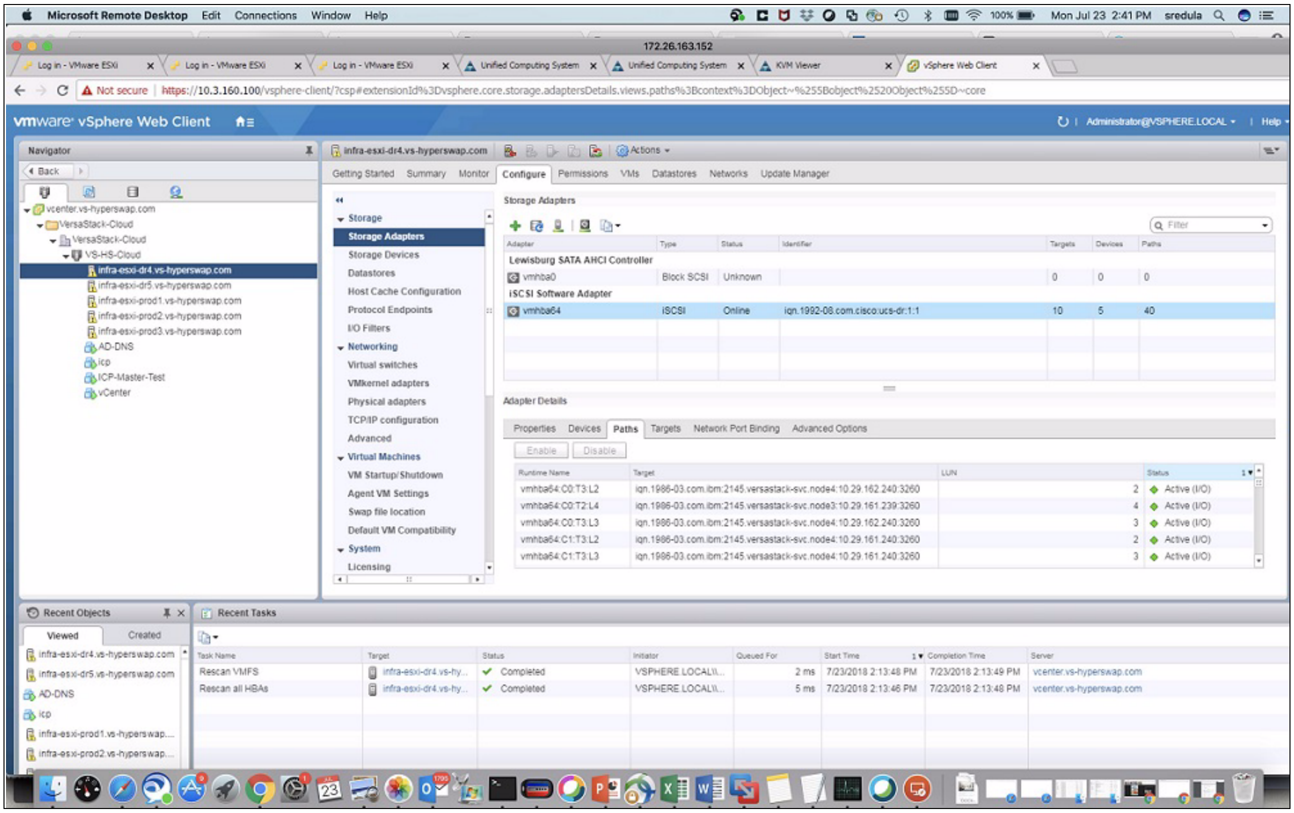
Volume Allocation
Allocated to Volumes 242.00 GiB
Capacity 16.00 TiB

Showing 10 volumes | Selecting 0 volumes

Name	State	Volume Group	UID	Host Mappings	Capacity
ESXi_iSCSI_3_Boot_Valid	Copy Offline		600507680C82811388000000000000A0	Yes	10.00 GiB
ESXi_iSCSI_3_Boot_...	✓ Online		600507680C82811388000000000000A0	No	10.00 GiB
ESXi_iSCSI_3_Boot_...	Offline		600507680C8281138800000000000160	No	10.00 GiB
ESXi_iSCSI_4_Boot_Valid	Copy Offline		600507680C82811388000000000000A1	Yes	10.00 GiB
ESXi_iSCSI_4_Boot_...	✓ Online		600507680C82811388000000000000A1	No	10.00 GiB
ESXi_iSCSI_4_Boot_...	Offline		600507680C8281138800000000000163	No	10.00 GiB
ESXi_iSCSI_5_Boot_Valid	Copy Offline		600507680C828113880000000000018C	Yes	10.00 GiB
ESXi_iSCSI_5_Boot_...	✓ Online		600507680C828113880000000000018D	No	10.00 GiB
ESXi_iSCSI_5_Boot_...	Offline		600507680C828113880000000000018C	No	10.00 GiB
ESXi_iSCSI_6_Boot_Valid	Copy Offline		600507680C8281138800000000000188	Yes	10.00 GiB
ESXi_iSCSI_6_Boot_...	✓ Online		600507680C8281138800000000000188	No	10.00 GiB
ESXi_iSCSI_6_Boot_...	Offline		600507680C8281138800000000000189	No	10.00 GiB
ESXi_iSCSI_DS1	Copy Offline		600507680C828113880000000000009E	Yes	1.00 TiB
ESXi_iSCSI_DS1 (Sec...	✓ Online		600507680C828113880000000000009E	No	1.00 TiB
ESXi_iSCSI_DS1 (Pri...	Offline		600507680C828113880000000000015A	No	1.00 TiB
ESXi_iSCSI_Swap1	Copy Offline		600507680C828113880000000000009F	Yes	500.00 GiB
ESXi_iSCSI_Swap1 (...	✓ Online		600507680C828113880000000000009F	No	500.00 GiB
ESXi_iSCSI_Swap1 (...	Offline		600507680C828113880000000000015D	No	500.00 GiB
Infra-ESXi-iSCSI-Host-31	Copy Offline		600507680C8281138800000000000012	Yes	10.00 GiB
Infra-ESXi-iSCSI-Hos...	✓ Online		600507680C8281138800000000000012	No	10.00 GiB

Figure 5-2 Volumes by Pool during FS900 failure

Host connectivity to the auxiliary volume was verified while Mdisks at site 1 went offline, as shown in Figure 5-3.



https://10.3.160.100/vsphere-client/?csp#extensionId%3Dvsphere.core.storage.adaptersDetails.views.paths%3Bcontext%3DObject~%255Bobject%2520Object%255D~core

Storage Adapters

Adapter	Type	Status	Identifier	Targets	Devices	Paths
Lewisburg SATA AHCI Controller						
vmhba0	Block SCSI	Unknown		0	0	0
iSCSI Software Adapter						
vmhba64	iSCSI	Online	iqn.1992-08.com.cisco:ucs-dr:1:1	10	5	40

Adapter Details

Properties | Devices | Paths | Targets | Network Port Binding | Advanced Options

Enable | Disable

Runtime Name	Target	LUN	Status
vmhba64:C0:T3:L2	iqn.1986-03.com.ibm:2145.versastack-svc.node4:10.29.162.240:3260	2	◆ Active (I/O)
vmhba64:C0:T2:L4	iqn.1986-03.com.ibm:2145.versastack-svc.node3:10.29.161.239:3260	4	◆ Active (I/O)
vmhba64:C0:T3:L3	iqn.1986-03.com.ibm:2145.versastack-svc.node4:10.29.162.240:3260	3	◆ Active (I/O)
vmhba64:C1:T3:L2	iqn.1986-03.com.ibm:2145.versastack-svc.node4:10.29.161.240:3260	2	◆ Active (I/O)
vmhba64:C1:T3:L3	iqn.1986-03.com.ibm:2145.versastack-svc.node4:10.29.161.240:3260	3	◆ Active (I/O)

Recent Tasks

Task Name	Target	Status	Initiator	Queued For	Start Time	Completion Time	Server
Rescan VMFS	infra-esxi-dr4.vs-hy...	✓ Completed	VSPHERE.LOCAL\\...	2 ms	7/23/2018 2:13:48 PM	7/23/2018 2:13:49 PM	vcenter.vs-hyperswap.com
Rescan all HBAs	infra-esxi-dr4.vs-hy...	✓ Completed	VSPHERE.LOCAL\\...	5 ms	7/23/2018 2:13:46 PM	7/23/2018 2:13:48 PM	vcenter.vs-hyperswap.com

Figure 5-3 VMware host shows all paths active

This result proves that there was no disruption to the host during the storage failure on the primary site in a properly configured HyperSwap cluster.

5.2.2 Secondary site SAN switch failure

The next test conducted was to shut down the Cisco MDS switches in the secondary site. This action broke communication between the nodes in the secondary site. The result was that the nodes in the secondary site lease expired. These nodes remained offline in starting state until connectivity was restored, as shown in Figure 5-4.

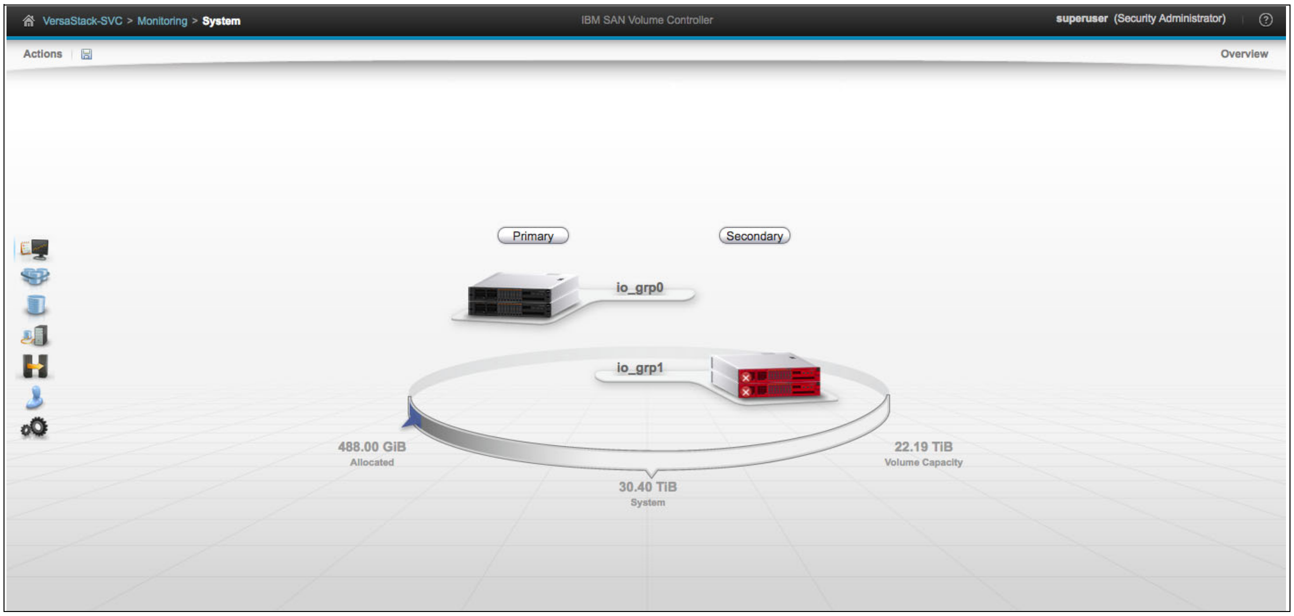
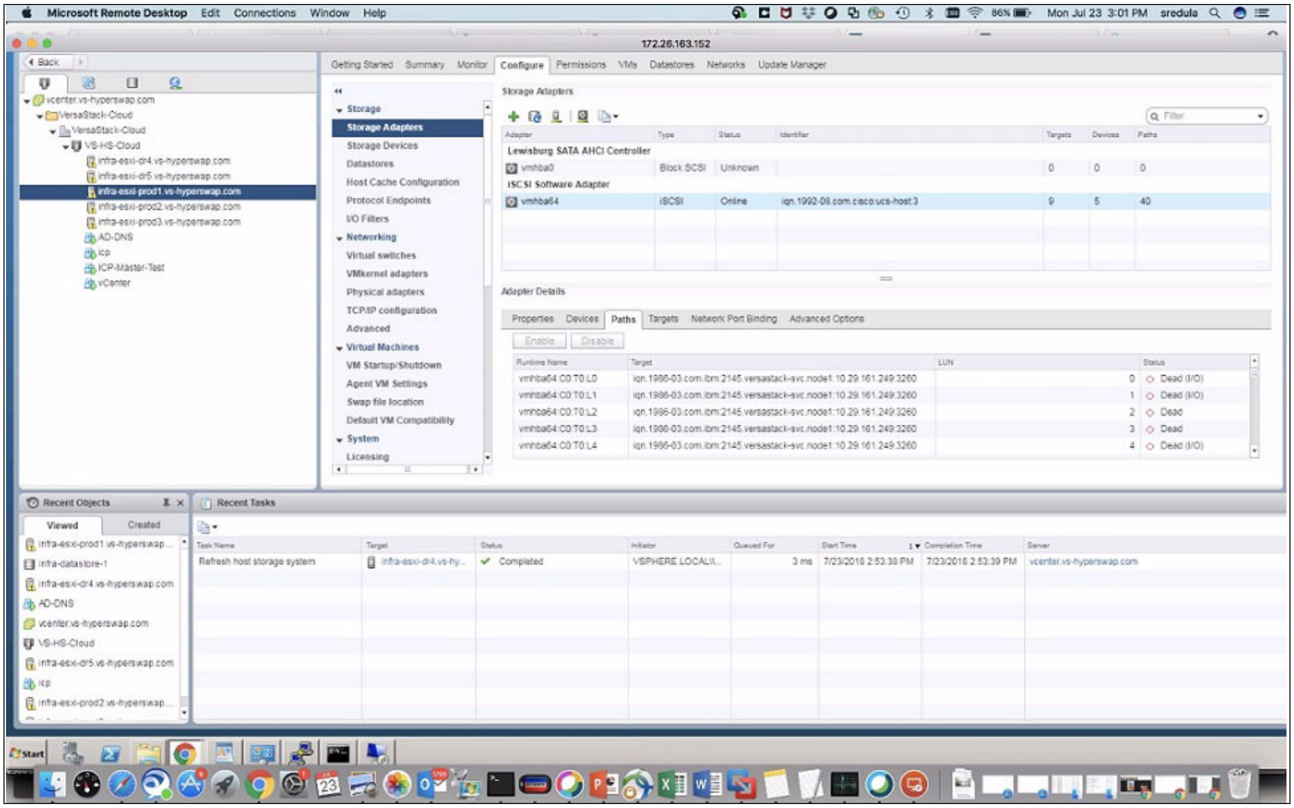


Figure 5-4 SVC Response to secondary site SAN failure

Hosts in the secondary site maintained access to the data through the primary site iogrp, as shown in Figure 5-5.



Storage Adapters

Adapter	Type	Status	Identifier	Targets	Devices	Paths
Lewisburg SATA AHCI Controller						
vmhba0	Block SCSI	Unknown		0	0	0
iSCSI Software Adapter						
vmhba64	iSCSI	Online	iqn.1992-08.com.cisco:ucs-host:3	9	5	40

Adapter Details

Properties Devices Paths Targets Network Port Binding Advanced Options

Enable Disable

Runtime Name	Target	LUN	Status
vmhba64:C0:T0:L0	iqn.1986-03.com.ibm:2145.versastack-svc.node1:10.29.161.249:3260	0	◇ Dead (I/O)
vmhba64:C0:T0:L1	iqn.1986-03.com.ibm:2145.versastack-svc.node1:10.29.161.249:3260	1	◇ Dead (I/O)
vmhba64:C0:T0:L2	iqn.1986-03.com.ibm:2145.versastack-svc.node1:10.29.161.249:3260	2	◇ Dead
vmhba64:C0:T0:L3	iqn.1986-03.com.ibm:2145.versastack-svc.node1:10.29.161.249:3260	3	◇ Dead
vmhba64:C0:T0:L4	iqn.1986-03.com.ibm:2145.versastack-svc.node1:10.29.161.249:3260	4	◇ Dead (I/O)

Recent Tasks

Task Name	Target	Status	Initiator	Queued For	Start Time	Completion Time	Server
Refresh host storage system	infra-esxi-dr4.vs-hy...	✓ Completed	VSPHERE.LOCAL\\...	3 ms	7/23/2018 2:53:38 PM	7/23/2018 2:53:39 PM	vcenter.vs-hyperswap.com

Figure 5-5 VMware hosts show half of the paths dead

5.2.3 Primary site iogrp failure

Another test that we performed was a hard shutdown of both SVC nodes in the primary site. Similar to the SAN outage, this action breaks inter-node communication between the primary and secondary sites. The secondary site nodes race to the active quorum. After the site reaches it, the nodes remain online. The SVC results in the loss of the iogrp and the volume copy in the primary site as show in Figure 5-6 and Figure 5-7.

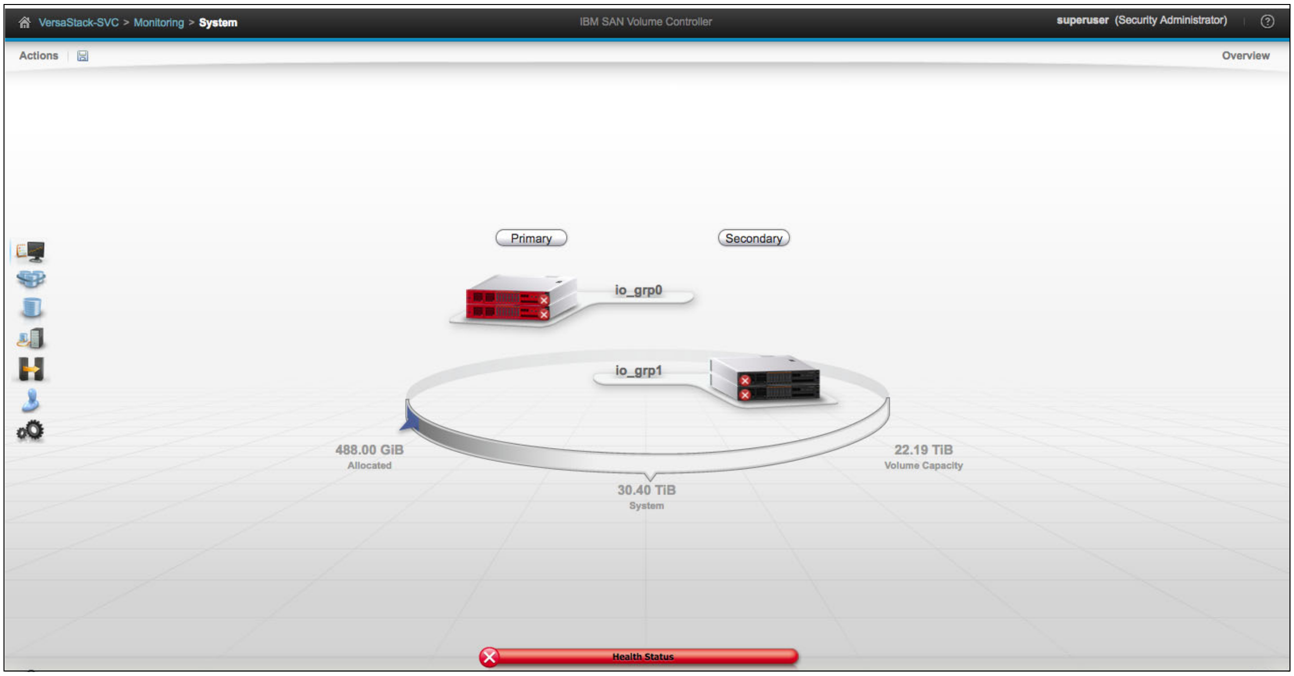
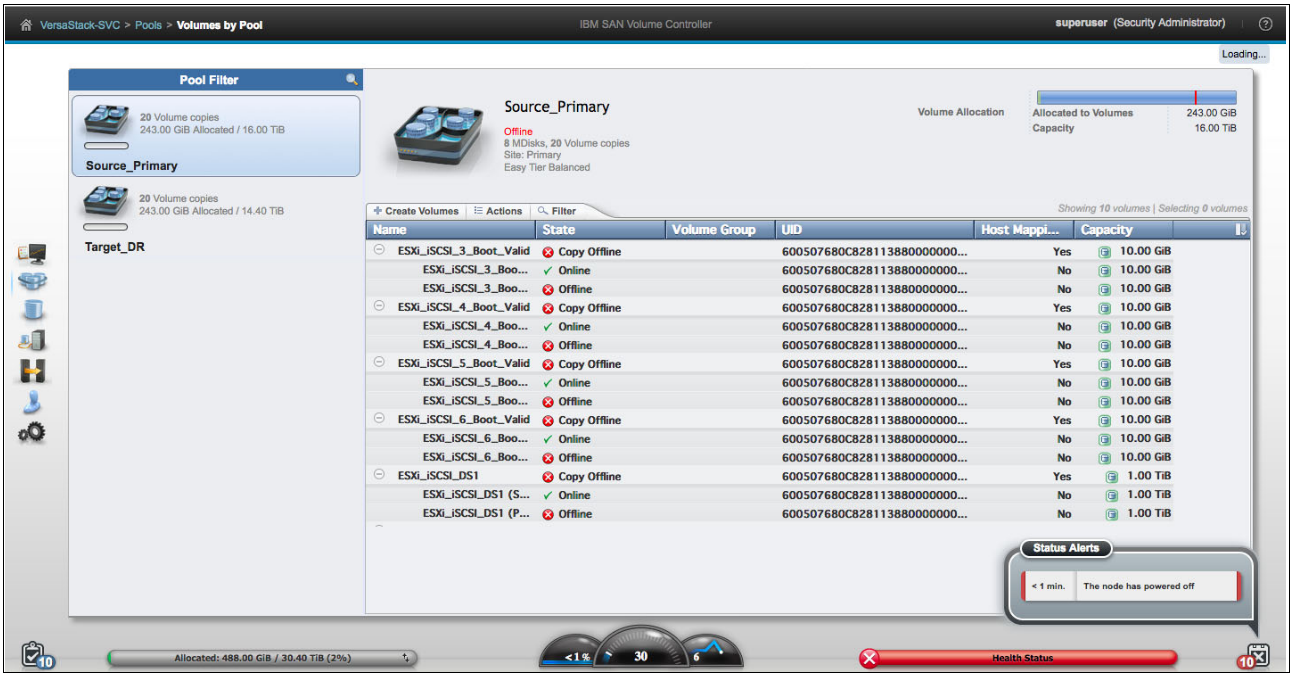


Figure 5-6 Loss of primary site nodes

Figure 5-7 shows the primary site volume copy offline.


VersaStack-SVC > Pools > Volumes by Pool
IBM SAN Volume Controller
superuser (Security Administrator)
Loading...
Pool Filter
+ Create Volumes
Actions
Filter
Status Alerts
Allocated: 488.00 GiB / 30.40 TiB (2%)
<1 %
30
6
Health Status
10
10

Source_Primary
20 Volume copies
243.00 GiB Allocated / 16.00 TiB

Target_DR
20 Volume copies
243.00 GiB Allocated / 14.40 TiB

Source_Primary
Offline
8 MDisks, 20 Volume copies
Site: Primary
Easy Tier Balanced

Volume Allocation		
	Allocated to Volumes	243.00 GiB
	Capacity	16.00 TiB

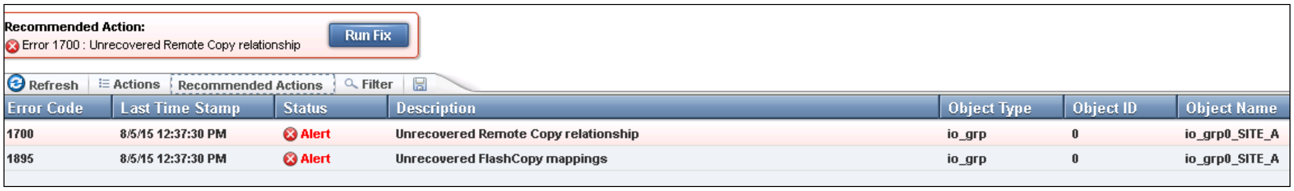
Showing 10 volumes | Selecting 0 volumes

Name	State	Volume Group	UID	Host Mappi...	Capacity
ESXi_iSCSI_3_Boot_Valid	Copy Offline		600507680C828113880000000...	Yes	10.00 GiB
ESXi_iSCSI_3_Boo...	✓ Online		600507680C828113880000000...	No	10.00 GiB
ESXi_iSCSI_3_Boo...	Offline		600507680C828113880000000...	No	10.00 GiB
ESXi_iSCSI_4_Boot_Valid	Copy Offline		600507680C828113880000000...	Yes	10.00 GiB
ESXi_iSCSI_4_Boo...	✓ Online		600507680C828113880000000...	No	10.00 GiB
ESXi_iSCSI_4_Boo...	Offline		600507680C828113880000000...	No	10.00 GiB
ESXi_iSCSI_5_Boot_Valid	Copy Offline		600507680C828113880000000...	Yes	10.00 GiB
ESXi_iSCSI_5_Boo...	✓ Online		600507680C828113880000000...	No	10.00 GiB
ESXi_iSCSI_5_Boo...	Offline		600507680C828113880000000...	No	10.00 GiB
ESXi_iSCSI_6_Boot_Valid	Copy Offline		600507680C828113880000000...	Yes	10.00 GiB
ESXi_iSCSI_6_Boo...	✓ Online		600507680C828113880000000...	No	10.00 GiB
ESXi_iSCSI_6_Boo...	Offline		600507680C828113880000000...	No	10.00 GiB
ESXi_iSCSI_DS1	Copy Offline		600507680C828113880000000...	Yes	1.00 TiB
ESXi_iSCSI_DS1 (S...	✓ Online		600507680C828113880000000...	No	1.00 TiB
ESXi_iSCSI_DS1 (P...	Offline		600507680C828113880000000...	No	1.00 TiB

< 1 min.	The node has powered off

Figure 5-7 SVC primary site volume copy offline

The hosts in this case marked the paths to the primary site nodes as dead. However, the storage remained available through the secondary site paths, as shown in Figure 5-8.



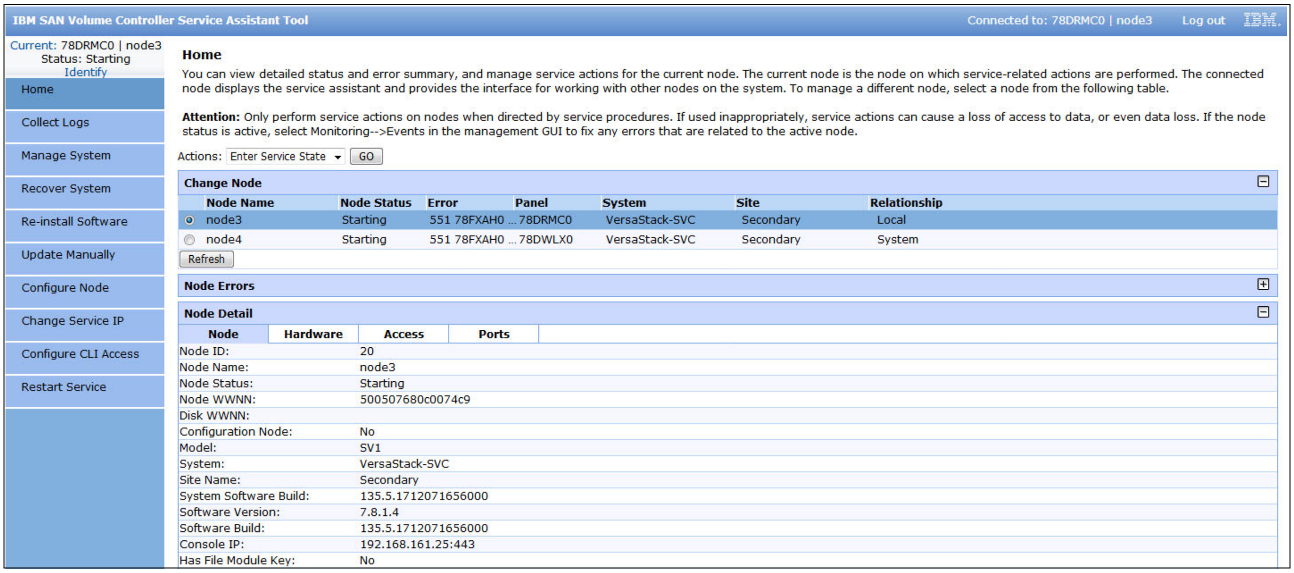
Recommended Action:
Error 1700 : Unrecovered Remote Copy relationship

Error Code	Last Time Stamp	Status	Description	Object Type	Object ID	Object Name
1700	8/5/15 12:37:30 PM	Alert	Unrecovered Remote Copy relationship	io_grp	0	io_grp0_SITE_A
1895	8/5/15 12:37:30 PM	Alert	Unrecovered FlashCopy mappings	io_grp	0	io_grp0_SITE_A

Figure 5-8 Error message

5.2.4 Loss of the active quorum and primary site nodes

For this test, we shut down the server that contained the IP Quorum application. We then shut down the SVC nodes in the primary site. The result of these actions was a storage outage, as shown in Figure 5-9.



IBM SAN Volume Controller Service Assistant Tool

Connected to: 78DRMC0 | node3

Current: 78DRMC0 | node3
Status: Starting

Home

You can view detailed status and error summary, and manage service actions for the current node. The current node is the node on which service-related actions are performed. The connected node displays the service assistant and provides the interface for working with other nodes on the system. To manage a different node, select a node from the following table.

Attention: Only perform service actions on nodes when directed by service procedures. If used inappropriately, service actions can cause a loss of access to data, or even data loss. If the node status is active, select Monitoring-->Events in the management GUI to fix any errors that are related to the active node.

Actions: Enter Service State

Change Node

	Node Name	Node Status	Error	Panel	System	Site	Relationship
●	node3	Starting	551 78FXAH0 ...	78DRMC0	VersaStack-SVC	Secondary	Local
○	node4	Starting	551 78FXAH0 ...	78DWLX0	VersaStack-SVC	Secondary	System

Node Errors

Node Detail

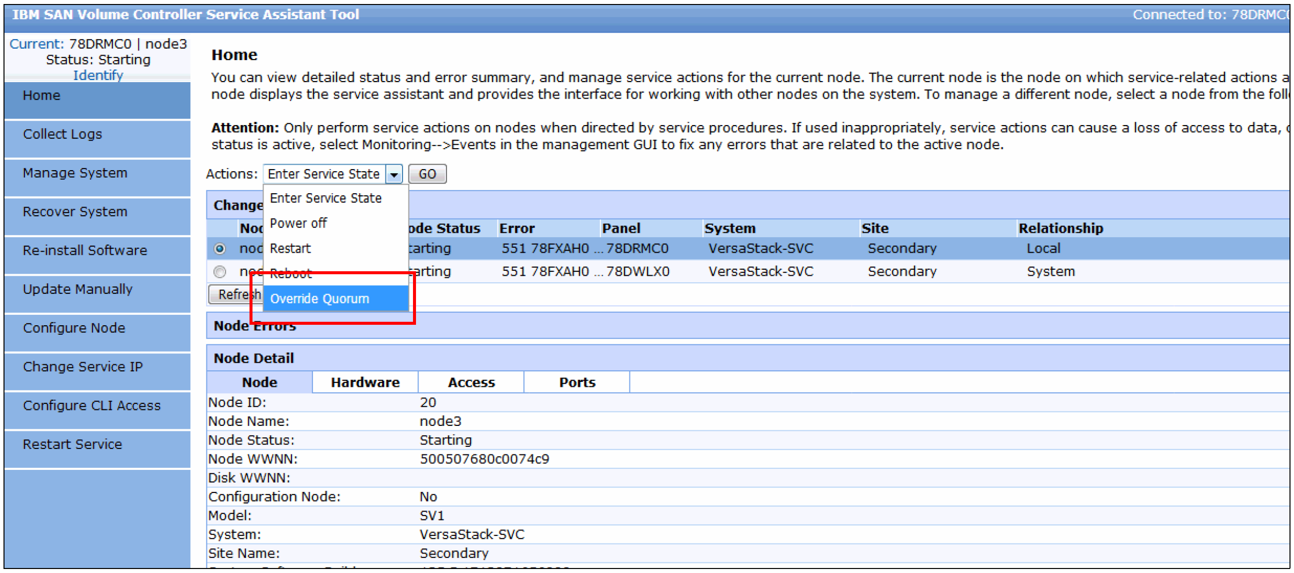
Node ID:	20
Node Name:	node3
Node Status:	Starting
Node WWNN:	500507680c0074c9
Disk WWNN:	
Configuration Node:	No
Model:	SV1
System:	VersaStack-SVC
Site Name:	Secondary
System Software Build:	135.5.1712071656000
Software Version:	7.8.1.4
Software Build:	135.5.1712071656000
Console IP:	192.168.161.25:443
Has File Module Key:	No

Figure 5-9 Service GUI after quorum and site loss

In this scenario, you can override the active quorum and bring storage access up on the secondary site. However, doing so discards the primary site data and requires the user to resynchronize all of the data back to the primary site manually.

> **Important:** Before performing the recovery action, ensure that the previous environment or site cannot come back to life with a device or node canister that still has earlier configuration data. This situation will cause serious problems in the environment or site that you are working on. Take any appropriate action to ensure that they cannot come back to life again (link disconnection, power down, and so on).

The procedure to override the active quorum is to select that option in the service GUI, as shown in Figure 5-10.



Current: 78DRMC0 | node3
Status: Starting

Home

You can view detailed status and error summary, and manage service actions for the current node. The current node is the node on which service-related actions a
node displays the service assistant and provides the interface for working with other nodes on the system. To manage a different node, select a node from the foll

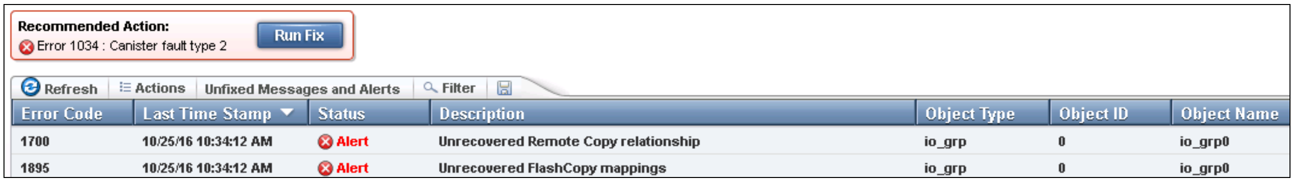
Attention: Only perform service actions on nodes when directed by service procedures. If used inappropriately, service actions can cause a loss of access to data,
status is active, select Monitoring-->Events in the management GUI to fix any errors that are related to the active node.

	Noc	ode Status	Error	Panel	System	Site	Relationship
●	noc	tarting	551 78FXAH0 ...	78DRMC0	VersaStack-SVC	Secondary	Local
○	noc	tarting	551 78FXAH0 ...	78DWLX0	VersaStack-SVC	Secondary	System

Node ID:	20
Node Name:	node3
Node Status:	Starting
Node WWNN:	500507680c0074c9
Disk WWNN:	
Configuration Node:	No
Model:	SV1
System:	VersaStack-SVC
Site Name:	Secondary

Figure 5-10 Override Quorum option

After this process is complete, the nodes on the secondary site warm start and become active. The consequence of this process is that the remote copy relationships and FlashCopy® maps associated with the HyperSwap volumes will be unrecoverable. The specific errors for this situation are shown in Figure 5-11.



Recommended Action:
Error 1034 : Canister fault type 2

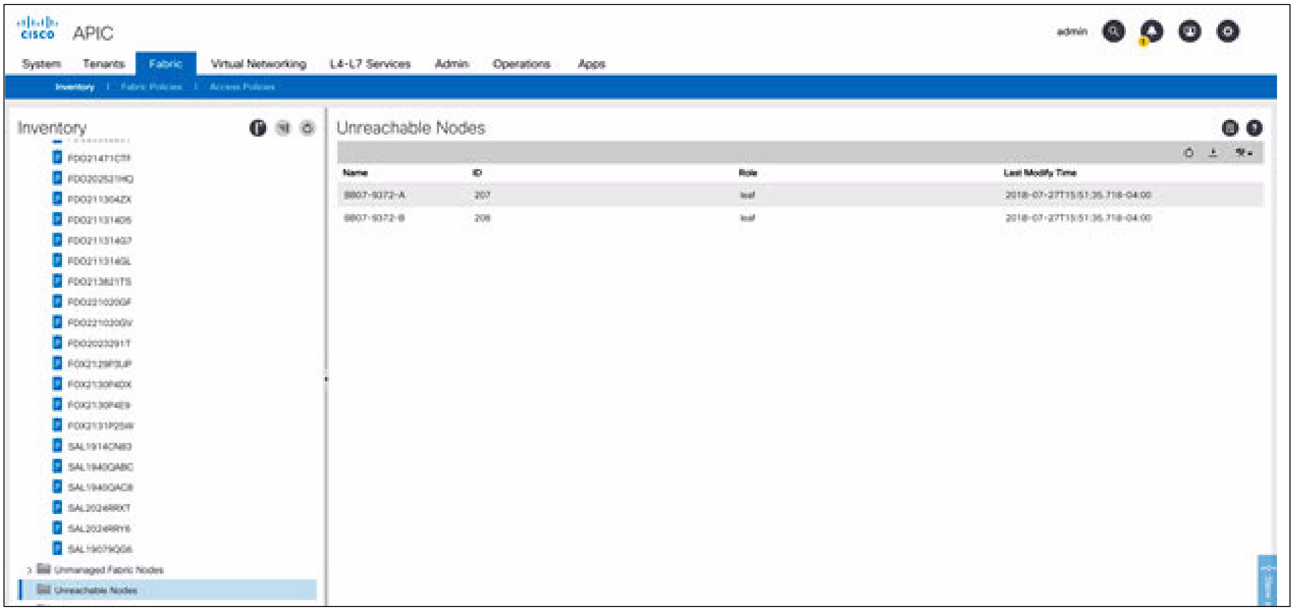
Error Code	Last Time Stamp	Status	Description	Object Type	Object ID	Object Name
1700	10/25/16 10:34:12 AM	Alert	Unrecovered Remote Copy relationship	io_grp	0	io_grp0
1895	10/25/16 10:34:12 AM	Alert	Unrecovered FlashCopy mappings	io_grp	0	io_grp0

Figure 5-11 Error messages

The recovery procedure for these errors will guide you to delete and re-create all of the remote copy and FlashCopy mappings affected. It is typically simpler to delete and re-create the volume offline volume copies associated with the HyperSwap volumes because this is a single operation that addresses both FlashCopy and remote copy relationships.

5.2.5 IP network failure

The test was conducted by shutting down the Cisco ACI Leaf switches in the primary site. This action interrupted the network communication to the ESXi servers in the primary site (Figure 5-12). All storage and network access to the ESXi servers was disrupted, with Cisco ACI leafs going offline. The result is the VMs running on the ESXi servers at the primary site vMotioned to the secondary site ESXi servers. Normal HA rules apply in an entire network at one site down situation.

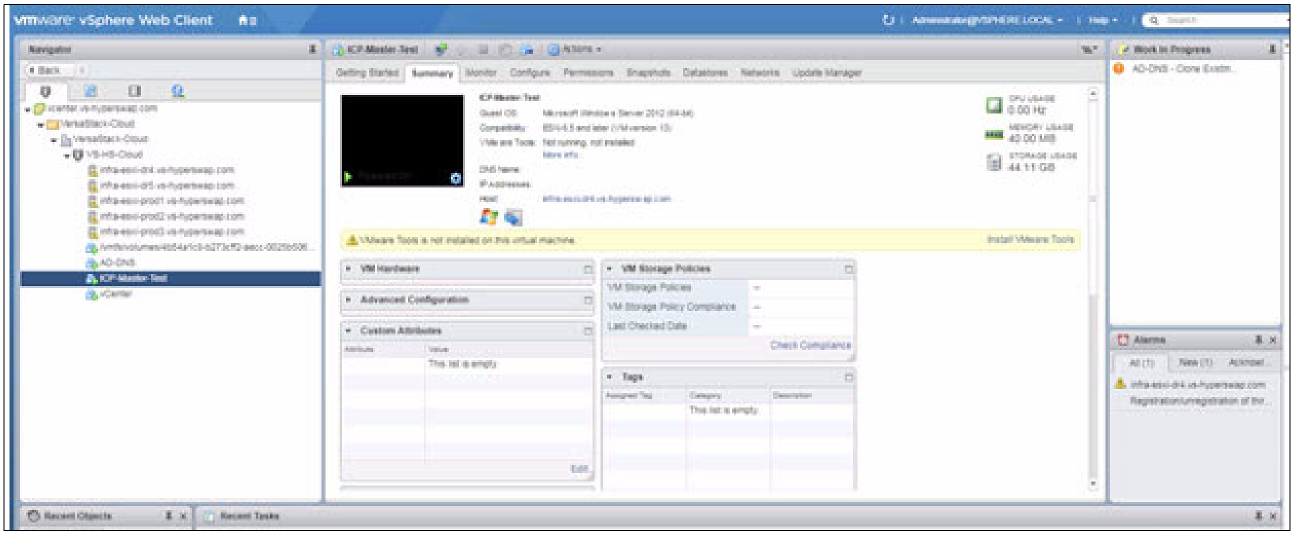


Unreachable Nodes

Name	ID	Role	Last Modify Time
8807-9372-A	207	leaf	2018-07-27T15:51:35.718-04:00
8807-9372-B	208	leaf	2018-07-27T15:51:35.718-04:00

Figure 5-12 Primary network failure

Figure 5-13 illustrates sample VM (`ICP_Master_Test`) running on one of the ESXi servers at the primary site.



ICP-Master-Test

Guest OS: Microsoft Windows Server 2012 (64-bit)
Compatibility: ESXi 6.5 and later (VM version 13)
VMware Tools: Not running, not installed
More info
DNS Name:
IP Addresses:
Host: infra-esxi-dr4.vs-hyperswap.com

VMware Tools is not installed on this virtual machine.

VM Storage Policies: --
VM Storage Policy Compliance: --
Last Checked Date: --

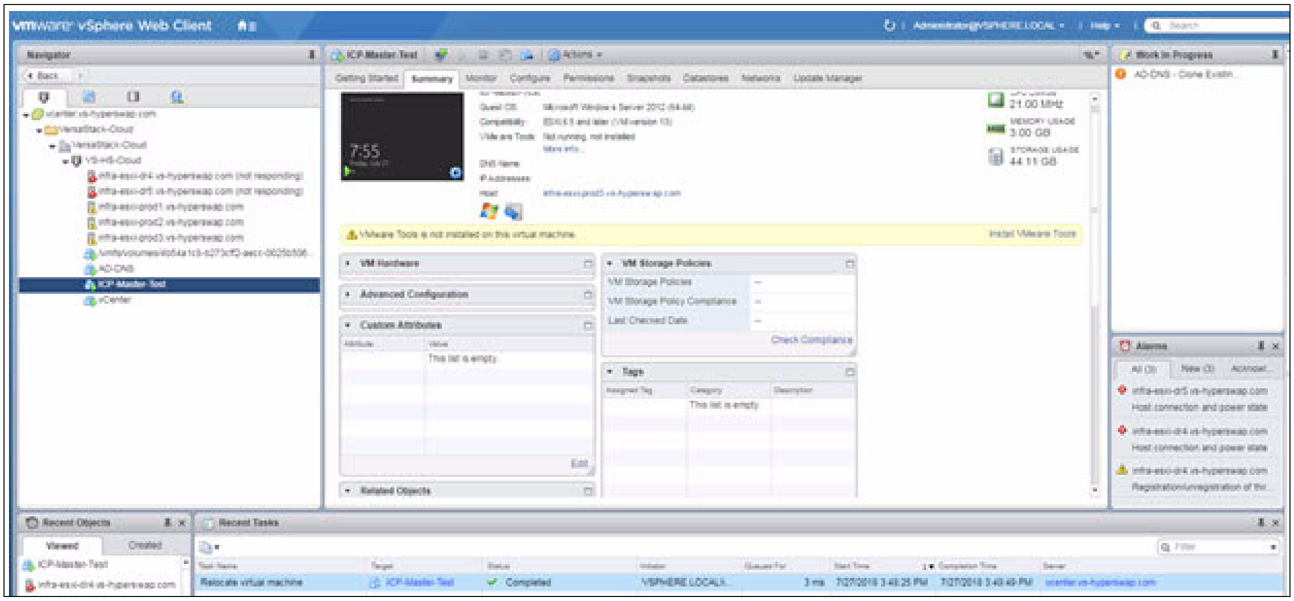
Attribute	Value
	This list is empty.

Assigned Tag	Category	Description
	This list is empty.	

infra-esxi-dr4.vs-hyperswap.com
Registration/unregistration of thir...

Figure 5-13 Sample VM running on one of the ESXi servers at primary site

The VM running at primary site moved to the ESXi server deployed at secondary site after the primary site network failure, as shown in Figure 5-14.



Guest OS: Microsoft Windows Server 2012 (64-bit)
Compatibility: ESXi 6.5 and later (VM version 13)
VMware Tools: Not running, not installed

Host: infra-esxi-prod3.vs-hyperswap.com

VMware Tools is not installed on this virtual machine. Install VMware Tools

infra-esxi-dr5.vs-hyperswap.com Host connection and power state
infra-esxi-dr4.vs-hyperswap.com Host connection and power state
infra-esxi-dr4.vs-hyperswap.com Registration/unregistration of thi...

Task Name	Target	Status	Initiator	Queued For	Start Time	Completion Time	Server
Relocate virtual machine	ICP-Master-Test	Completed	VSPHERE.LOCAL\...	3 ms	7/27/2018 3:48:25 PM	7/27/2018 3:48:49 PM	vcenter.vs-hyperswap.com

Figure 5-14 VMs moved to secondary site after the network failure

5.3 Other failure cases

This section provides information about failure cases that were not explicitly tested as part of this publication.

5.3.1 Handling volumes with bad blocks

In a production environment, an SVC HyperSwap cluster can report bad blocks on a back-end storage controller. SVC Cluster will log event ID 1840 on affected Mdisks on the back-end Storage Controller. In this scenario, the recovery action is typically to overwrite the affected data using the host or restoring any missing data from backup after the back-end controller has been repaired. Alternatively, you can refresh the volume copy by deleting the corrupted volume copy and re-creating it. Contact IBM Support before performing any action plan that involves bad blocks to evaluate the full range of recovery options and their potential impact on the solution.

5.3.2 Active quorum loss and dual site failure (T3)

A data-center-wide power outage can cause dual site nodes to go offline. When nodes do not shut down gracefully, there might not be enough time to save inflight data or Spectrum Virtualize configuration data to the quorum disks. In such scenarios, all Spectrum Virtualize nodes might come up with error codes 550/578 when they are powered on after an outage:

- Error 550: Cannot form a cluster due to a lack of cluster resources.

  The node cannot become active in a cluster because it is unable to connect to enough other cluster resources. The cluster resources are the nodes in the system and the active quorum disk, which can be a SAN-attached MDisk or a drive.

The node needs to be able to connect to a majority of the resources before that group will form an online cluster. This requirement prevents the cluster from splitting into two or more active parts, with both parts independently performing I/O.

The error data lists the missing resources. This data includes a list of nodes and optionally either a drive that is operating as the quorum disk or a LUN on an external storage system that is operating as the *quorum disk*.

- Error 578: The state data was not saved following a power loss.

  On startup, the node was unable to read its state data. When this situation happens, the node expects to be automatically added back into a clustered system. However, if it is not joined to a clustered system in 60 sec, it raises this node error. This error is a critical node error, and user action is required before the node can become a candidate to join a clustered system.

Engage IBM support with a Severity 1 ticket if error combination 550/578 is reported. If the errors are not induced by other hardware failures or as a result of maintenance actions, IBM support should be engaged run a Tier 3 (T3) recovery process.

Spectrum Virtualize cluster data is stored and updated continuously on its quorum drives. Whenever a change in configuration occurs such as creating a VDisk or deleting a VDisk, these changes are updated on quorum disks. Spectrum Virtualize is designed to run a **`cron`** job every night at 1:00 AM and back up its configuration data. This data is stored in quorum drives as well. When a T3 recovery is run, stored configuration data is used to recover the configuration and host mappings. The T3 operation only deals with recovering configuration data, not customer data. However, with every successful T3 recovery data, access is restored to respective hosts.

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.

- *IBM HyperSwap for IBM FlashSystem A9000 and A9000R*, REDP-5434
- *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317
- *Implementing a VersaStack Solution by Cisco and IBM with IBM Storwize V5030, Cisco UCS Mini, Hyper-V, and SQL Server*, SG24-8407
- *Storwize HyperSwap with IBM i*, REDP-5490

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

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- VersaStack Datacenter with Cisco Application Centric Infrastructure and IBM SAN Volume Controller CVD

  https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/versastack_aci_svc_vmw6.html
- Recommended practices for configuring multi-site SAN fabrics

  https://www.cisco.com/c/en/us/support/docs/storage-networking/mds-9000-series-multilayer-switches/46202-zoning-switches.html
- Bidirectional PIM Deployment Guide

  https://www.cisco.com/c/dam/en/us/products/collateral/ios-nx-os-software/multicast-enterprise/prod_white_paper0900aecd80310db2.pdf
- System Storage Interoperation Center (SSIC)

  http://www.ibm.com/systems/support/storage/ssic/interoperability.wss
- V7.8.x Supported Hardware List, Device Driver, Firmware, and Recommended Software Levels for IBM SVC Cluster

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

© Copyright IBM Corp. 2018. All rights reserved.

- List of supported product list and documentation for SVC

  https://www.ibm.com/support/home/product/5329743/SAN_Volume_Controller_(2145,_2147)

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Redbooks

Implementing VersaStack with Cisco ACI Multi-Pod and IBM HyperSwap for

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

SG24-8423-00

ISBN 0738457086

Printed in U.S.A.

Get connected

Redbooks®

ibm.com/redbooks