White Paper
Public

The bridge to possible

FlexPod MetroCluster IP with VXLAN Multi-Site Frontend Fabric

Last Updated: July 1, 2022

Contents

Executive Summary 3
Solution Overview 4
Solution Architecture 5
Solution Design 23
Solution Validation 70
Conclusion 86
Appendix – References 87

Executive Summary

Cisco and NetApp have partnered to deliver FlexPod to support a broad range of enterprise workloads and use cases. FlexPod solutions deliver systems and solutions that are designed, tested, and documented to accelerate customer deployments while also reducing IT pain points and minimizing risks. FlexPod datacenter solutions provide a foundational architecture for building private and hybrid cloud data centers using Cisco and NetApp products and technologies. They incorporate design, technology, and product best practices to ensure a successful deployment. The designs are robust, efficient, and scalable and incorporate a range of technologies and products from Cisco and NetApp. FlexPod solutions also provide design and deployment guidance through Cisco Validated Designs and white papers that Enterprises can leverage during the different stages (planning, design, implementation, and production) of a roll-out.

This document describes the end-to-end design for a FlexPod data center solution, specifically the FlexPod MetroCluster IP solution with a Cisco VXLAN Multi-Site fabric. The solution is a disaster-recovery and business-continuity solution for the Virtualized Server Infrastructure (VSI) in an enterprise data center. The solution uses an active-active data center design to ensure availability to at least one data center at all times. The two active-active data centers can be in the same campus location or geographically dispersed across different sites in a metropolitan area. The data center infrastructure is based on NetApp's MetroCluster IP solution using All-Flash FAS (AFF) arrays for storage, Cisco Unified Computing System™ (Cisco UCS®) for the compute, and VMware vSphere for the virtualization layer using one VMware vCenter to manage the virtualized resources in both data centers. The network infrastructure in the solution is a Cisco® VXLAN BGP Ethernet VPN (EVPN) Multi-Site fabric, managed by Cisco Data Center Network Manager (DCNM) and built using Cisco Nexus 9000® series cloud-scale switches. The Cisco VXLAN fabric provides Layer 2 extension and Layer 3 forwarding between the active-active data centers, enabling applications to be deployed in either data center with seamless connectivity and mobility. The NetApp MetroCluster IP solution uses synchronous replication between data centers to ensure continuous availability to the storage data with zero data loss, zero recovery point objective (RPO), and near-zero recovery time objective (RTO). A dedicated Layer 2 network is used in this design for the storage replication traffic between sites to ensure the integrity of the replication data between data centers.

To simplify and accelerate the deployment of the end-to-end design, the solution uses a combination of GUI-driven automation, RedHat Ansible automation and traditional methods for Day 0 through Day 2 activities. For example, Cisco DCNM's Fabric Builder that provides a GUI-based automation for deploying a VXLAN fabric, is used in this solution to deploy the VXLAN fabric in the two active-active data centers. Similarly, the Day 2 network setup activities is automated using RedHat Ansible. RedHat Ansible is also used for deploying the Cisco UCS compute and VMware virtualization in both data centers. For operational simplicity, the Cisco UCS compute in the two data center locations are centrally managed from the cloud using Cisco Intersight™. Cisco Intersight is a cloud-based Software-as-a-Service (SaaS) orchestration and operations platform that uses a continuous integration/continuous development (CI/CD) model to deliver new capabilities for both private and hybrid cloud deployments. Cisco Intersight now offers Intersight Managed Mode (IMM), a new unified management architecture that uses standardized policies for all Cisco UCS infrastructure regardless of their location. Customers can also manage this solution using Cisco IMM.

The VSI infrastructure in each data center is built using VMware vSphere 7.0U1, NetApp AFF A700 storage with ONTAP 9.8, and Cisco UCS blade servers with Cisco UCS Manager (UCSM) release 4.1(3b). The VXLAN Multi-Site fabric consists of Cisco Nexus® 9000 Series switches running NX-OS 9.3(6), managed by Cisco DCNM 11.5(1). Also included in the solution are Cisco Intersight and NetApp Active IQ SaaS platforms, NetApp ONTAP System Manager; NetApp Virtual Storage Console; NFS VAAI Plug-in, and SnapCenter Plug-in for seamless VMware integration; and NetApp Active IQ Unified Manager.

Solution Overview

Introduction

The FlexPod solution is a predesigned, integrated, and validated architecture for the data center that combines Cisco UCS servers, Cisco Nexus switches, and NetApp Storage Arrays into a single, flexible architecture. FlexPod is designed for high availability, with no single points of failure, while maintaining cost-effectiveness and flexibility in the design to support a wide variety of workloads.

In this FlexPod MetroCluster IP solution, the Cisco VXLAN Multi-Site solution allows you to interconnect and centrally manage VXLAN fabrics deployed in separate, geographically dispersed data centers. NetApp MetroCluster IP provides a synchronous replication solution between two NetApp storage clusters providing storage high availability and disaster recovery in a campus or metropolitan area. The solution enables you to design a VMware vSphere-based private cloud on a distributed integrated infrastructure and deliver a unified solution that enables multiple sites to behave in much the same way as a single site while protecting data services from a variety of single-point-of-failure scenarios, including a complete site failure.

This document provides the end-to-end design of the FlexPod MetroCluster IP solution with Cisco VXLAN Multi-Site fabric. The solution was validated by a walk-through of the buildout of the virtualized server infrastructure in the two data centers directly connected to each other through the VXLAN fabric, with additional supporting infrastructure for the solution. The FlexPod solution discussed in this document has been validated for resiliency and fault tolerance during various failure scenarios as well as a simulated site failure scenario.

Audience

The audience for this document includes, but is not limited to, sales engineers, field consultants, professional services, IT managers, partner engineers, and customers who want to take advantage of an infrastructure built to deliver IT efficiency and enable IT innovation.

What's New?

The following design elements distinguish this version of FlexPod from previous FlexPod models:

- Integration of Cisco VXLAN Multi-Site fabric in FlexPod Datacenter solutions to seamlessly support multiple sites
- Integration of NetApp MetroCluster IP for synchronous data replication across the two data centers
- Simplified operations and ease of automation by using Cisco DCNM for centralized management of a VXLAN Multi-Site fabric
- Operational agility with Day 0 through Day 2 automation by using Cisco DCNM Fabric Builder for a GUI-driven, automated Day 0 buildout of the VXLAN fabric, and RedHat Ansible for Day 1+ network setup and Day 0-2 setup of the Cisco UCS and VMware virtual server infrastructure
- Support for a VMware vSphere 7.0U1 stretched cluster across the FlexPod MetroCluster IP solution sites
- Design, validation and operational aspects of this new FlexPod MetroCluster IP Datacenter design

Solution Architecture

Introduction

At a high level, the FlexPod MetroCluster IP with VXLAN Multi-Site fabric solution consists of two FlexPods, located at two sites separated by some distance, but connected and paired together to provide a highly available, highly flexible, and highly reliable data center solution that can provide business continuity despite a site failure. The FlexPod VSI at each site is connected to a larger data center fabric and the sites are interconnected through an interconnect network. In addition, the two NetApp ONTAP clusters in a MetroCluster IP configuration synchronously replicate data across sites through inter-site links (ISL) in a MetroCluster IP network, as illustrated in Figure 1. Please note that the MetroCluster IP network can utilize a dedicated network or it can be integrated into the larger data center fabric using the compliant switches configuration, when the switches and network configuration meet the requirements. In the FlexPod MetroCluster IP with VXLAN Multi-Site solution, a dedicated MetroCluster IP network is used.

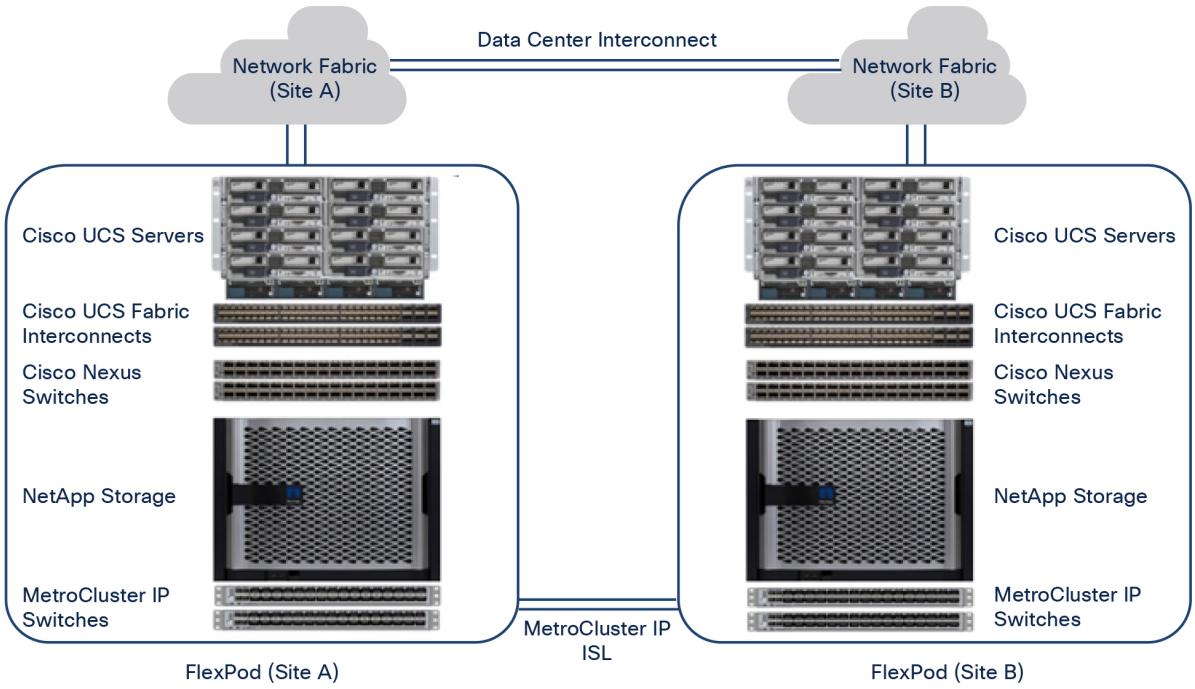


Figure 1.
FlexPod MetroCluster IP Solution – High Level

The NetApp storage requirements for the MetroCluster IP network are:

- The supported maximum distance between sites is 700 km.
- The round-trip latency must be less than or equal to 7 ms.
- Packet loss and drops due to congestion or over-subscription must be less than or equal to 0.01%.
- The supported jitter value is 3 ms for a round trip (or 1.5 ms for one way).

For additional details and requirements for the NetApp MetroCluster IP solution deployment, please refer to NetApp documentation on Install a MetroCluster IP configuration: ONTAP MetroCluster.

As stated earlier, the NetApp ONTAP storage clusters at both sites are paired together in a MetroCluster IP configuration for synchronous data replication between sites using MetroCluster IP inter-switch links (ISL).

The MetroCluster IP solution helps ensure storage availability for continued business data services if a site disaster occurs. The MetroCluster IP solution uses a high-performance Ethernet storage fabric between sites. A dedicated four-node MetroCluster IP network is used in this solution and offers a simple solution architecture that supports for up to 700 km distance between sites (refer to Figure 2).

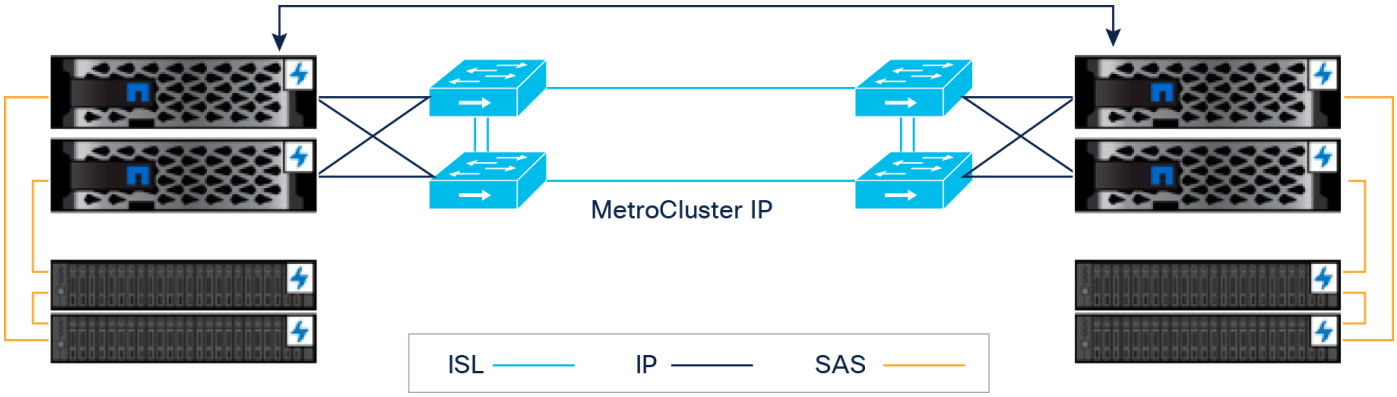


Figure 2.
Four-Node MetroCluster IP Solution

A stretched VMware cluster implementation that combines the compute resources from both sites in a high-availability configuration allows you to restart virtual machines in the surviving site if a site disaster occurs. For normal operations, VM-Host affinity rules are followed so virtual machines are running locally at a site using local site storage. If a disaster occurs, the virtual machines that are restarted at the surviving site can continue to access their storage provided by the MetroCluster IP solution.

The FlexPod configuration at each site is connected to a data center network fabric at each site that provides connectivity for the virtualized server infrastructure and for the applications hosted on that infrastructure. A VXLAN Multi-Site data center fabric is used in this solution. However, different data center network architecture options are available for connectivity within a site and for interconnecting the sites. The solution requirements and architectures are discussed in upcoming sections.

Solution Requirements

The FlexPod MetroCluster IP with VXLAN Muti-Sie fabric solution is designed to address the following key requirements:

- Business continuity and disaster recovery in the event of a complete data center (site) failure
- Flexible, distributed workload placement with workload mobility across data centers
- Direct and independent access to external networks and services from each data center location
- Site Affinity where virtual machine data is accessible locally, from the same data center site
- Quick recovery with zero data loss if a failure occurs
- Simplified administration and operation of the solution

From a design perspective, the high level design goals for the solution are:

- Resilient design across all layers of the infrastructure with no single point of failure
- Scalable design with the ability to independently scale compute, storage, and network bandwidth as needed

- Modular design where you can change, upgrade, or replace components, resources, or sub-systems as needed
- Flexible design across all layers of the solution that includes sub-system design, individual components used, and storage configuration and connectivity options
- Ability to automate and simplify by enabling integration with external automation and orchestration tools
- Incorporation of technology and product-specific best practices for all components used in the solution

The next section addresses different network architectures and design factors to consider when designing a data center infrastructure solution for business continuity and disaster recovery.

Data Center Network

Traditional data center networks are built and managed in a de-centralized manner, on a device-by-device basis. Rolling out a new application or service typically required network operators to access and configure each device individually to achieve the desired connectivity. Enterprises that automated their network operations still had to overcome significant challenges, not only in terms of time and resources but also from automation and programmability options that were both limited and complex. Enterprises not only had to manage a wide range of network protocols, technologies and design options, but they also had the burden of managing automation capabilities that varied from vendor to vendor, product to product, and sometimes even between products from the same vendor. As a result, Enterprises had significant upfront work before any automation could be developed and used. Rolling out new applications or other changes in the network could therefore take days, if not weeks, to complete and often impeded a business's ability to adapt and meet evolving business needs. To address these and other challenges, a number of architectural shifts have occurred in the past few years as outlined below.

Modern data center networks typically use a Software-Defined Networking (SDN) architecture with programmable infrastructure and centralized management entity or controller that manages the data center network as a whole, as one fabric. In this architecture, the management plane is centralized and plays a critical role in defining the behavior of the network's control and data planes. When an application needs to be deployed, the requirements are centrally defined on the controller and the controller then pushes the corresponding network configuration and policies programmatically to the relevant nodes in the fabric to achieve the desired results. This centralized approach minimizes user errors and helps ensure consistent deployment across multiple nodes. It also minimizes configuration drift because it maintains the original network intent on the controller which now serves as a single source of truth for the network fabric. In many SDN solutions, the configuration on the individual nodes are continuously verified against the original intent, and if a drift happens, you are alerted so you can resolve and synchronize the configuration. Some architectures also prevent drifts altogether by allow configuration changes only through the controller. SDN architectures, therefore, can bring agility and efficiency to enterprise networks by enabling you to deploy network changes quicker, uniformly and in a more consistent manner, with fewer errors regardless of the scale of the fabric.

SDN architectures also make it easier for enterprises to automate their network infrastructure because operators no longer have to interface with each network device individually. They can programmatically convey the network intent to the controller and the controller will then take care of implementing that intent on the individual nodes in the fabric. Automating the network fabric and the data center infrastructure is a primary goal for enterprises considering a DevOps operational model which requires access to all infrastructure as code (IaC). It is also critical for enterprises looking to achieve cloud-like agility in their private cloud and hybrid cloud deployments or to simply speed up IT operations.

Another architectural shift in modern data centers is in the network topology itself. Enterprises are migrating from traditional two- and three-tier hierarchical designs with flat two-tier Clos-based spine-and-leaf designs that are better suited for the traffic patterns seen in data center networks, where the levels of East-West traffic are significantly higher than in other parts of the enterprise. Clos-based fabrics simplifies the network topology and provides consistent and predictable routing with 3-stage forwarding with predictable performance and horizontal scaling. To realize these benefits, Clos topologies impose specific role-based characteristics on the nodes in the fabric. For example, endpoints connect only to switches in a leaf role and never to spine switches, spine switches connect to all leaf switches in the fabric but never connect directly to each other, leaf switches connect to all spine switches but not to each other for fabric connectivity, all of which brings uniformity, predictability, and most importantly, consistent performance across the network fabric (refer to Figure 3).

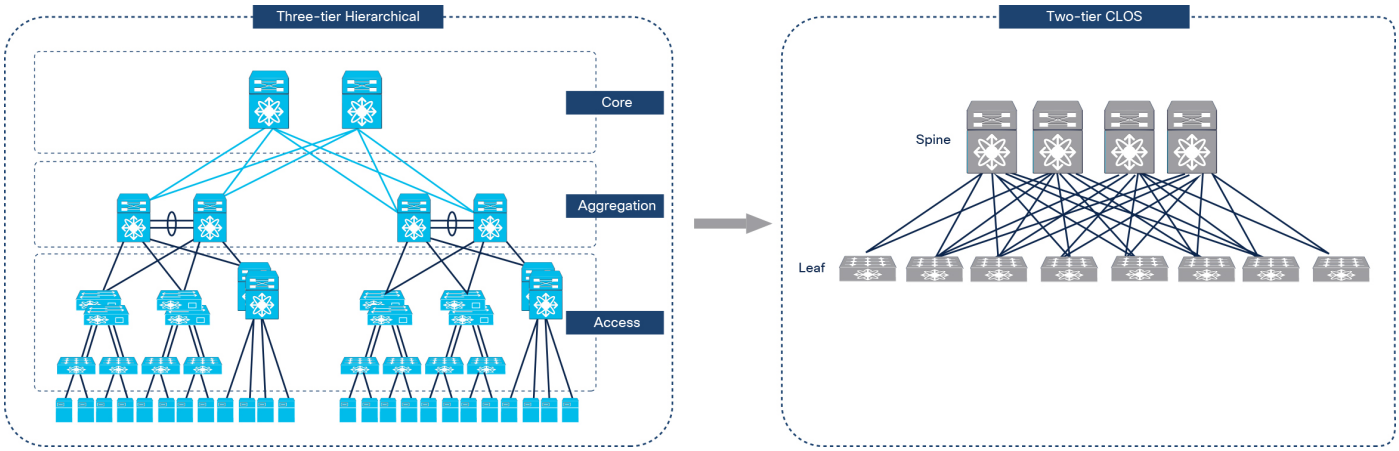


Figure 3.
Data Center Network - Traditional Three-tier Hierarchical vs. Two-tier Clos Design

The Clos-based design is also the foundation for another architectural shift in modern data center fabrics, namely Virtual Extensible LAN (VXLAN) overlay networks. VXLAN overlay networks enable Layer 2 extension and Layer 3 forwarding across a routed IP network (underlay) by establishing tunnels between leaf switches, referred to as VXLAN Tunnel Endpoints (VTEPs). VXLAN networks also provide other benefits such as :

- Scalability: VLAN networks have a 4096-segment limit. VXLAN networks can scale up to 16 million segments.
- Flexibility: Workload placement is not limited by the physical location of compute resources in the network. VXLAN networks allow you to place workloads anywhere in the network fabric, regardless of where the compute resources physically connect to the network fabric. In a data center IP network, this requires layer 2 extension and VXLAN provides it by establishing VXLAN tunnels between top-of-rack switches that the compute resources connect to. The workloads are uniquely identified in a shared data center IP network using a VXLAN segment ID or Virtual Network Identifier (VNID). Traffic using the same VNID are in the same network segment regardless of where it physically originated in the network.
- Mobility: Virtual machines and workloads in a given network segment will use the same VNID regardless of where they connect to in the network fabric. This provides the necessary layer 2 extension, enabling workloads to be moved anywhere that the data center fabric connects.

For FlexPod customers, all of the above architectural shifts in data center networks means more options and better solutions. FlexPod supports both the traditional 2- and 3-tier hierarchical designs as well as the modern SDN-based designs. The two SDN solutions supported in FlexPod are:

- Cisco Application Centric Infrastructure (Cisco ACI®) fabric - with support for Multi-Pod or Multi-Site connectivity
- Cisco DCNM-managed VXLAN MP-BGP EVPN fabric - with support for VXLAN Multi-Site connectivity

Both SDN solutions use a 2-tier Clos-based spine-and-leaf topology. FlexPod can also be deployed using traditional 2- or 3-tier hierarchical design. When a SDN fabric is used, the Cisco UCS compute and NetApp storage in the solution will connect to leaf switches in the fabric, (Cisco ACI or Cisco VXLAN MP-BGP EVPN), or to top-of-rack switches in a traditional 2- or 3-tier design. The data center network fabric for the FlexPod MetroCluster IP solution in this document is a VXLAN MP-BGP EVPN Multi-Site fabric, managed using Cisco DCNM, which serves as the centralized controller for the fabric. The solution also includes a separate Layer 2 network for synchronous replication traffic between NetApp MetroCluster IP storage systems in the active-active data centers.

All three network architectures are discussed in greater detail in upcoming sections of this document. All three architectures enable enterprises to build active-active data centers that provide business continuity and disaster recovery(BC/DR). The active-active data centers can be in a single location such as a campus environment or distributed across different geographical sites. The primary focus of each architecture is to interconnect data centers and provide seamless connectivity and mobility across data centers. The architecture must also deliver an end-to-end network design that includes a design for the network within a data center as well as a design for interconnecting them.

Before we look at the network architectures, it is important to first have a good understanding of the connectivity and flow of traffic in an enterprise data center, as outlined in the next section.

Connectivity

A data center network fabric provides the Layer-2 and Layer-3 connectivity necessary for bringing the compute and storage infrastructure online. When the infrastructure is operational, the network fabric provides Layer-2 and Layer-3 connectivity for deploying and hosting the workloads (applications, services, etc.) on that infrastructure. In a BC/DR infrastructure solution using active-active data centers, the workloads can be active in either data center, and you can move workloads between data centers as needed. The application workloads distributed across data centers should also be able to communicate across data centers, across application tiers, and between components in the same tier. This communication of data centers and workload mobility is a critical requirement to ensure business continuity in if a disaster occurs.

Traffic Flow

The traffic traversing a data center network fabric can be broadly categorized as being either infrastructure or application traffic. The infrastructure traffic is any traffic required to build, operate, and maintain the data center infrastructure. Application traffic is any traffic generated by workloads hosted on that infrastructure, when it is operational. However, note that traffic to/from infrastructure management elements hosted on the infrastructure are also categorized as infrastructure traffic since it used to operate the infrastructure, even if it is hosted on the infrastructure as application workloads are. The important distinction is that the infrastructure traffic and connectivity are considered to be foundational and a pre-requisite for hosting workloads, and also for providing capabilities such as flexible workload placement, mobility and high availability.

All data center traffic can also be characterized as:

- East-West, for traffic between endpoints connected to the same data center fabric
- North-South, for traffic to/from enterprise users in campus, WAN, or remote locations accessing applications and services hosted in the data centers or for accessing cloud services on the Internet; it could also be for connecting to management elements that manage the infrastructure within the data center

For the FlexPod BC/DR solution in this document,

- East-West traffic also includes any traffic between endpoints attached to different data centers in the active-active design; for example, the storage replication traffic between NetApp storage within and across data centers.
- North-South traffic is the same as before but it includes traffic to/from applications and services hosted in either of the data centers in the active-active design. The traffic patterns for these flows can also vary, depending on the network design. The design could use a common access point for outside/external connectivity; for example, each site could use the inter-site network to also connect to outside/external networks. Alternatively, each location could use a dedicated and independent connection for connecting to outside/external networks. For this solution, regardless of the external connectivity design, it is important to ensure that the design meets the BC/DR requirements for availability.

Network Architecture Options

The following sections provide different network architectures available for a BC/DR solution using an active-active design that provides Layer-2 and Layer-3 forwarding of traffic within a data center (East-West) and traffic from outside networks to the data center and conversely (North-South). For the FlexPod MetroCluster IP solution in this document, this paper discusses additional design options for forwarding synchronous storage replication traffic within and across data centers.

Option 1: VXLAN BGP EVPN Multi-Site Fabric

The Cisco VXLAN BGP EVPN Multi-Site fabric provides a highly flexible, scalable, and resilient network architecture for enabling business continuity and disaster recovery in modern data centers. This architecture brings Cisco's industry-leading, standards-based, innovative suite of products and technologies to deliver a programmable infrastructure with orchestration and management capabilities for meeting the needs of modern data centers.

The VXLAN BGP EVPN Multi-Site architecture uses VXLAN overlays and a 2-tier Clos-based IP underlay network, built using Nexus® 9000 Series-based spine-and-leaf switches. VXLAN overlays enable you to build Layer 2 and Layer 3 networks and extended them across data center locations, enabling multiple data centers to be interconnected with seamless forwarding of East-West traffic between them. Though the flexibility and simplicity of building VXLAN overlays across any IP underlay is extremely useful, the resulting overlay network is a flat network that spans data centers with no fault isolation neither in the control plane nor in the data plane. With Layer 2 extension, the overlays can result in large multi-data center bridge domains even when the underlay network is a well-designed, Clos-based IP fabric with hierarchical addressing, fault isolation, etc. To address this problem, Cisco, along with other industry partners have developed the VXLAN BGP EVPN Multi-Site architecture to interconnect data centers. In the VXLAN multi-site architecture, the end-to-end VXLAN overlay network is split into failure domains where the data centers and the interconnect between data centers are all in different failure domains. The overlay

segmentation provides not only fault isolation and segmentation, but also routing hierarchy without sacrificing seamless connectivity or mobility between data centers. The transition between overlay segments also provides a point of control for enforcing forwarding and security policies between data centers. This approach is also more scalable, enabling enterprises to interconnect multiple data centers as needed. Figure 4 illustrates this architecture.

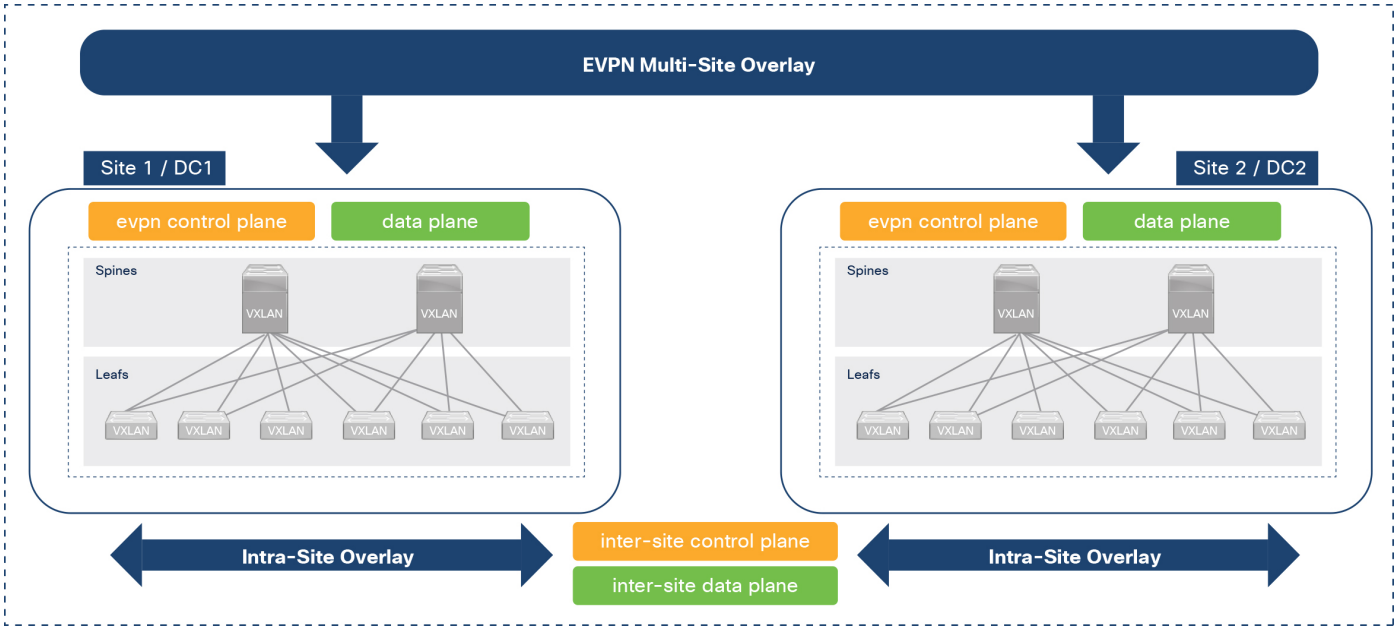


Figure 4.
VXLAN EVPN Multi-Site Architecture

FlexPod customers can take advantage of the VXLAN BGP EVPN architecture to build individual data center fabrics and extend them to multiple locations using the Multi-Site architecture. The FlexPod MetroCluster IP solution in this document uses this approach to build active-active data centers as shown in Figure 5.

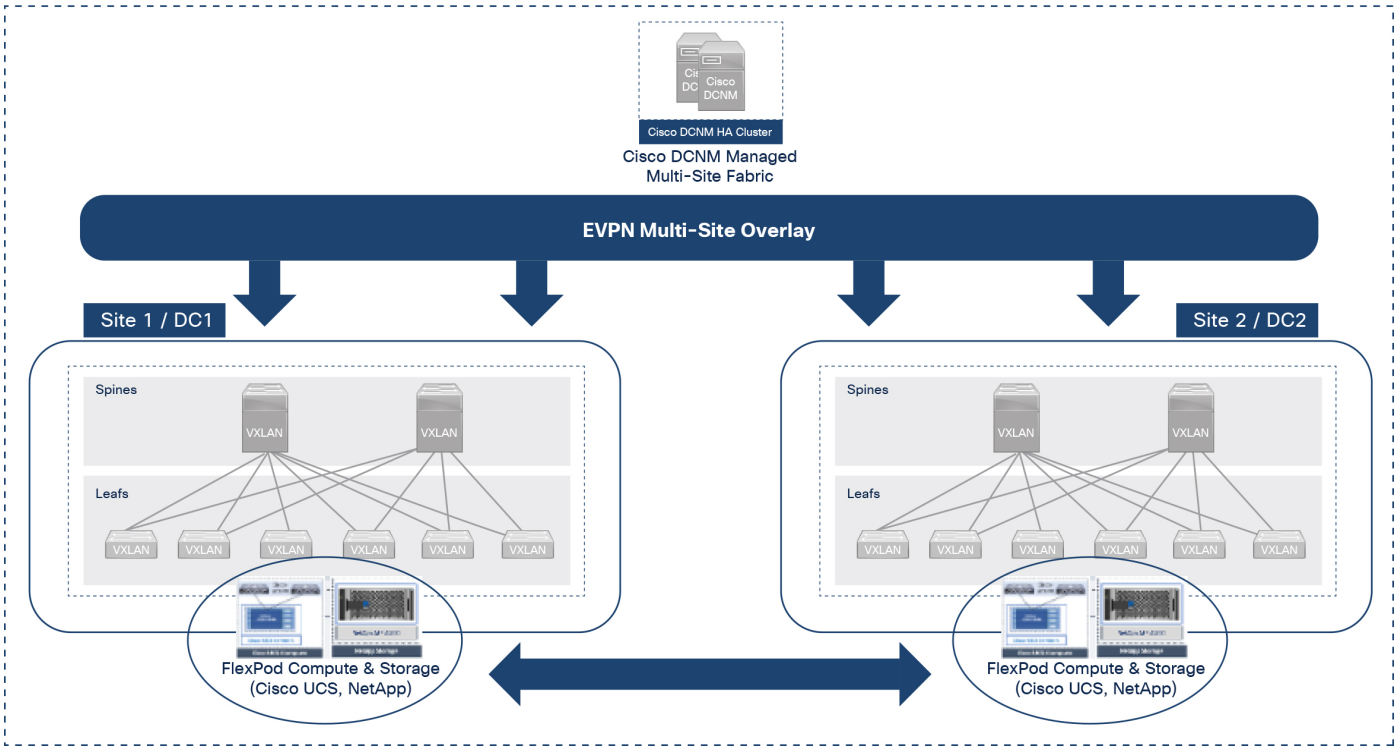
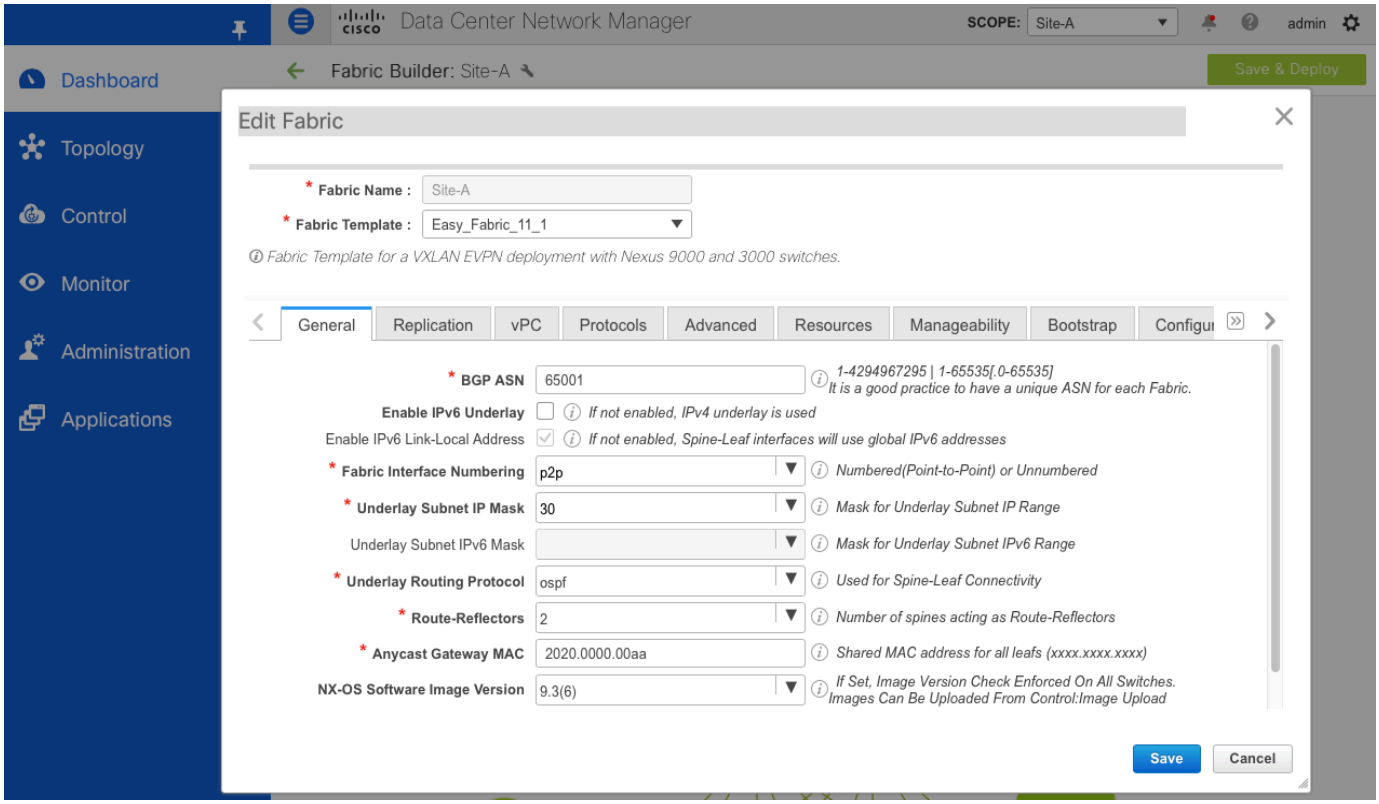


Figure 5.
Active-Active Data Centers Using VXLAN EVPN Multi-Site Architecture

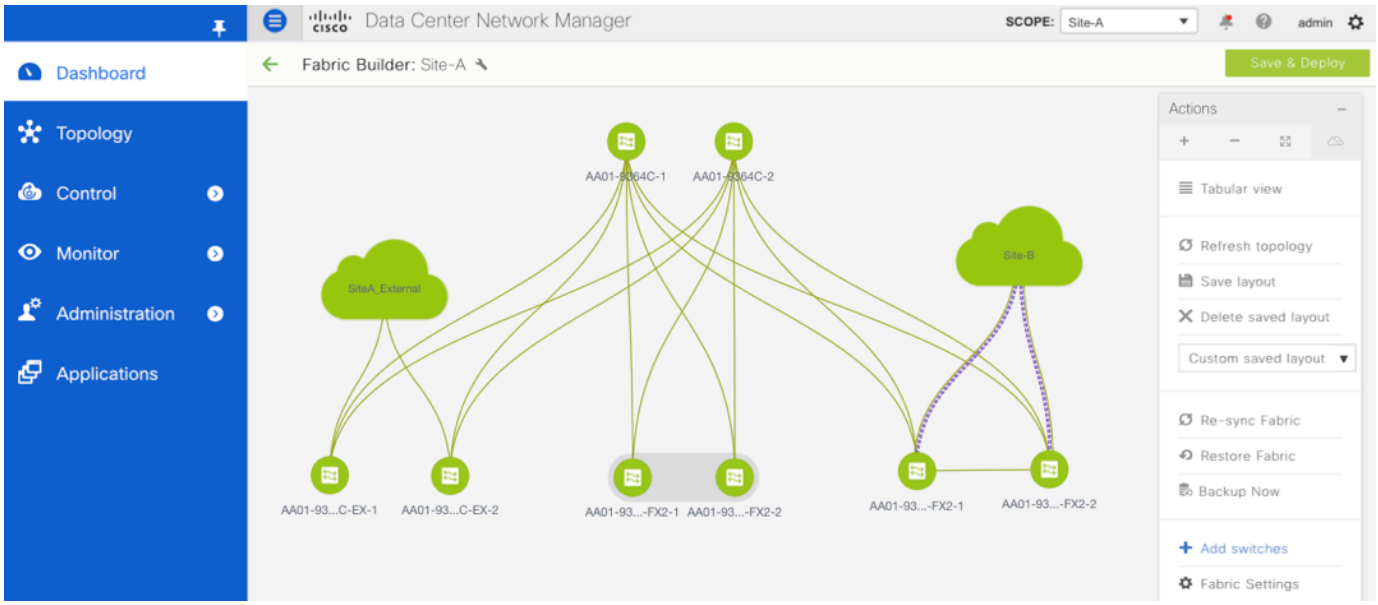
In this solution, the VXLAN EVPN Multi-Site architecture is used to provide BC/DR by interconnecting data center fabrics in two sites. Other network architectures options for enabling active/active data centers are discussed in the next two sections.

The Cisco VXLAN BGP EVPN Multi-Site fabric is managed by Cisco DCNM deployed outside the fabric. Cisco DCNM is deployed as a cluster of multiple nodes for high availability. The network fabric consisting of spines and leaf switches is built using Nexus 9000 series switches capable of supporting VXLAN overlays. The fabric provides seamless Layer 2 extension and Layer 3 forwarding for both infrastructure and application and services VMs hosted on the infrastructure.

For VXLAN fabrics, Cisco DCNM serves as an SDN controller to provision and operate the fabric. Cisco DCNM will need to first discover and add switches to a fabric. Cisco DCNM will assume a default role (spine, leaf) for the switch based on the model, but you can easily change the role (for example, from spine to leaf or leaf to border or border gateway, etc.). Cisco DCNM can also bootstrap or Pre-boot Execution Environment (PXE) boot individual switches. Cisco DCNM Fabric Builder then uses the discovered switches to deploy a complete VXLAN fabric. Fabric Builder provides a GUI-based mechanism for automating the deployment of a VXLAN fabric. Additional connectivity such as connectivity to outside/external networks (including managing the configuration on external gateways), and inter-site connectivity between data centers can also be deployed from Cisco DCNM. The following screenshot shows the Fabric Builder GUI, where the tabs represent different configuration categories.



The administrator can save the provided inputs and initiate an automated deployment of the VXLAN fabric with one click of a button. All the necessary configuration to bring up a fabric are deployed across all nodes in the fabric. The screenshot below shows the VXLAN fabric deployed in Site-A for this solution, using Fabric Builder. To bring up a switch in a VXLAN fabric typically requires 200 to 300 lines of configuration, if not more. Using DCNM Fabric Builder, therefore, significantly accelerates the deployment and ensures that it is deployed correctly, consistently and without user errors.



Cisco DCNM as an SDN controller for the fabric also makes it easier to automate the network fabric for northbound integration into a DevOps tool chain or other enterprise tools because you can develop automation for Cisco DCNM to manage the whole fabric rather than automating on a per-switch basis.

Cisco DCNM also continuously verifies the original intent against the configuration on the nodes to prevent configuration drifts and other problems. Cisco DCNM can also monitor and provide lifecycle management of the fabric after deployment. Cisco DCNM simplifies Day 2 operations by providing easy-to-use workflows follows for upgrades, Return Material Authorization (RMAs) in addition to network-wide visibility, fault monitoring, and automated consistency checks at regular intervals to prevent configuration drifts. Cisco uses pre-defined template policies for the automated deployment and configuration of all devices that Cisco DCNM manages and deploys, but administrators can also use free-form templates to customize as needed. Cisco DCNM also integrates with other Cisco orchestration and operational tools such as Network Insights and Multi-Site Orchestrator to provide centralized policy-based management of larger fabrics and for more in-depth monitoring of all fabrics.

Option 2: Cisco ACI Multi-Pod Fabric

The Cisco ACI fabric is an application-centric, policy based, secure SDN architecture for the data center. To enable active-active data centers for BC/AR, you can extend the network fabric to multiple data centers using the Cisco ACI Multi-Pod architecture. The Cisco ACI architecture also uses a 2-tier Clos topology, built using Nexus 9000 series-based spine-and-leaf switches and VXLAN overlays for forwarding East-West traffic within and across data centers. However, unlike the other two architectures presented in this document, Cisco ACI builds the network fabric with minimal input from the network administrator or operator. The fabric is pre-designed based on Cisco's years of networking expertise, and deployed using design, product, and technology best practices. Cisco ACI hides the complexities of designing and deploying the IP underlay and the VXLAN overlay, and instead focusses on enabling a data center fabric ready for deploying applications. The architecture assumes that the primary goal of a data center fabric is for supporting applications and therefore, Cisco ACI focusses on accelerating, simplifying, and optimizing application roll outs and on the networking and policies associated with that roll out, rather than on how the network is built.

Cisco ACI also uses an intent-based framework to deliver agility. It captures high-level intent (application, business) in the form of a policy and translates the policy into network constructs to provision the networking, security, and other services on the individual switches that make up the fabric. The Cisco Application Policy Infrastructure Controller (Cisco APIC) serves as a centralized controller for the fabric and manages it as one system, with tight integration between hardware and software, physical and virtual elements, and innovative Cisco Application-Specific Integrated Circuits (ASICs) that bring exceptional capabilities to modern data centers. Cisco APIC is a critical architectural component of the Cisco ACI solution. It is typically deployed as a high-availability cluster (3 nodes minimum) with the ability to horizontally scale as needed by adding more nodes. It is the unified point of automation and management for the Cisco ACI fabric, policy enforcement, and health monitoring.

Cisco ACI architecture is very extensible, and it provides several architectural options to meet different enterprise requirements. They include Cisco ACI Multi-Pod for interconnecting data center fabrics managed by the same Cisco APIC cluster, Cisco ACI Remote Leaf for edge/satellite deployments, Cisco ACI Mini Fabric, Cisco ACI Multi-Site with Multi-Site Orchestrator for interconnecting multiple data center sites, and Cisco Cloud ACI with Cloud APIC and Cisco Cloud Services Router 1000V (Cisco CSR 1000V) for public cloud deployments (Amazon AWS, Microsoft Azure). The Cisco ACI architecture can orchestrate seamless connectivity across these environments using a common policy model. The common policy and operating model significantly reduces the cost and complexity of managing multi-data center and multi-cloud data center deployments.

FlexPod customers can leverage the Cisco ACI architecture for the data center fabric, and then expand and extend the fabric using one of the above architectural options. For an active-active data center design such as the one in this solution, you can use both Cisco ACI Multi-Pod and Cisco ACI Multi-Site design to build and interconnect data center fabrics. However, there are some important factors that you should consider when deciding which one to use.

The Cisco ACI Multi-Site architecture is designed to interconnect ACI fabrics in different geographical sites or regions and to enable a large scale-out architecture for the network fabric. Each fabric is referred to as a "Site" in the Cisco ACI Multi-Site architecture, with each site managed by a separate APIC cluster. Each fabric is also part of a larger Cisco ACI Multi-Site fabric that helps to ensure connectivity between sites using a common policy model that extends end to end, across all data center sites. A critical element of this architecture is the Cisco Multi-Site Orchestrator (Cisco MSO). Cisco MSO provides a single point of administration for all inter-site policies which is then extended to the different APIC clusters managing the different ACI fabrics. Cisco MSO also automates and manages the interconnect between ACI fabrics to enable multi-tenant Layer 2 and Layer 3 connectivity between sites. Multiprotocol BGP (MP-BGP) EVPN sessions are established to exchange endpoint reachability information, and VXLAN tunnels are established for forwarding Layer 2 and Layer 3 traffic across sites. Cisco MSO is also a common point of automation that can be very helpful in automating large ACI deployments.

The Cisco ACI Multi-Pod is designed to interconnect ACI networks in different physical data center locations but within the same campus or metropolitan area. The ACI networks are managed as one fabric, using a single APIC cluster. The ACI network in each data center is referred to as a "Pod" in the Cisco ACI Multi-Pod architecture. Each Pod can be considered a separate availability zone where each Pod is a separate network fault domain because each Pod runs separate instances of control plane protocols, though all are part of the same ACI fabric and are managed by the same APIC cluster. Cisco ACI Multi-Pod is administratively and operationally simpler by managing multiple data center networks as a single, logical fabric. Policy changes are seamlessly deployed across all Pods because they are part of the same ACI fabric and they are managed by the same APIC. Pods are, therefore, part of the same tenant change domain, where any changes for a given tenant are immediately applied to all Pods.

Cisco ACI Multi-Pod is commonly used in active-active data center deployments that also have some or all of the following requirements :

- Both data centers are part of the same Virtual Machine Manager (VMM) domain.
- VMware Distributed Resource Scheduler (DRS) initiates workload mobility between data centers.
- vSphere High Availability or Fault Tolerance is configured across data centers.
- The data centers have application clustering that requires Layer 2 extension with broadcast, unicast, and multicast between data centers.
- Clustered network services are (firewalls, load balancers) across data centers.

Based on the requirements for this solution, Cisco ACI Multi-Pod architecture is a better fit than Cisco ACI Multi-Site for building active-active data centers for business continuity and disaster recovery. For more details, refer to the Design Guide and Deployment Guide of the validated FlexPod MetroCluster IP solution with Cisco ACI Multi-Pod.

Option 3: Hierarchical Layer 2/Layer 3

You also can build data center networks using hierarchical Layer 2/Layer 3 (IP) designs with a network of Layer 2 or Layer 3 switches at each tier. The network design typically involves an access layer (same as other network architectures), but also an aggregation layer and core layers. Traditionally, these aggregation layers are Layer 3, with the core being Layer 2 or Layer 3. Figure 6 shows a hierarchical 3-tier architecture. You also can use enterprise data center networks with this type of architecture or a variation of this model (for example, collapsed core/aggregation) in a FlexPod MetroCluster IP solution for BC/DR.

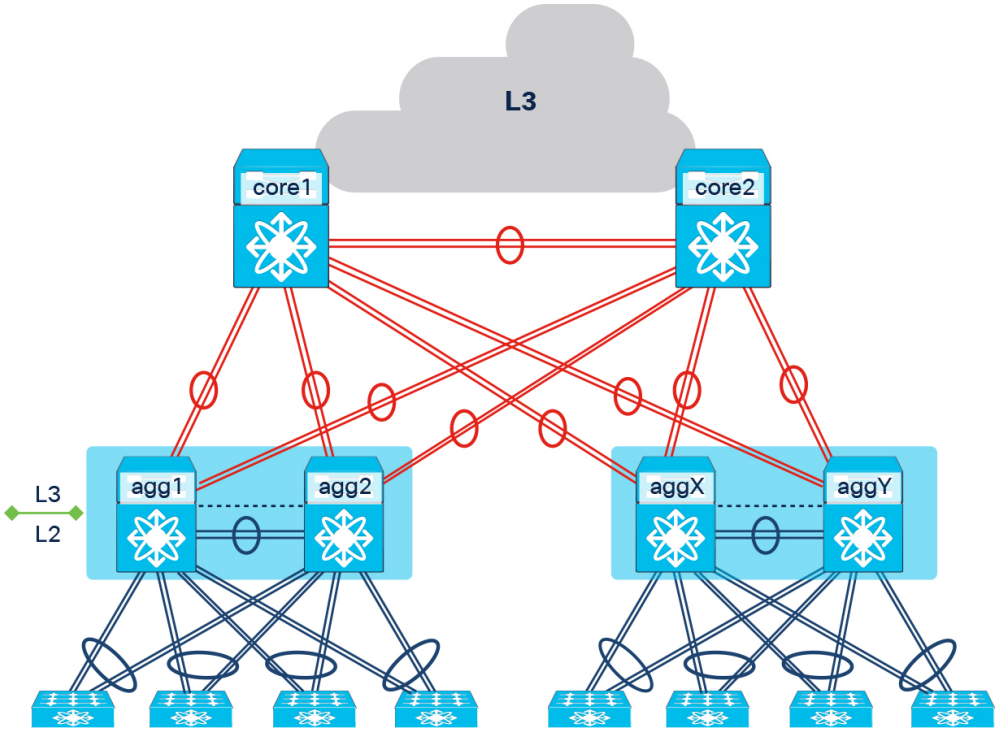


Figure 6.
Hierarchical 3-Tier Architecture

The connectivity between the data centers would require a Data Center Interconnect (DCI) technology such as Overlay Transport Visualization (OTV) over dark fiber, Dense Wavelength Division Multiplexing (DWDM) or IP transport. Alternatively, you also could use Multiprotocol Label Switching (MPLS)-based EVPNs or other Layer 2 and Layer 3 interconnect technologies.

Extended Layer 2 Design

For a small solution deployment that might be cost-sensitive, simply extending Layer 2 across sites will require the least amount of hardware to implement a FlexPod MetroCluster IP solution. For site-to-site connectivity that uses extended Layer 2 connectivity natively, you can achieve additional cost optimizations by sharing the Inter-Switch Link (ISL) between sites for MetroCluster IP and non-MetroCluster IP usage if you have sufficient bandwidth. In addition, you can use the Cisco Nexus switches in a FlexPod as compliant MetroCluster IP switches if they meet the requirements of the compliant switches, thus eliminating the need for additional dedicated MetroCluster IP switches.

Following are the general requirements for a MetroCluster IP solution with compliant switches:

- Only platforms that support MetroCluster IP and provide dedicated ports for switchless cluster interconnects are supported, including NetApp AFF A300, AFF A320, AFF A400, AFF A700, AFF A800, FAS8200, FAS8300, FAS8700, FAS9000, ASA AFF A400, ASA AFF A700, and ASA AFF A800.
- You can connect the MetroCluster IP interface to any switch port that you can configure to meet the requirements.
- The speed of the switch ports required depends on the platform. For example, it is 25 Gbps for AFF A300, 40 Gbps for AFF A700, and 40/100 Gbps for AFF A800.
- The ISLs must be 10 Gbps or higher and must be sized appropriately for the load on the MetroCluster configuration.
- The MetroCluster IP configuration must be connected to two networks, and each MetroCluster IP node must be connected to two network switches.
- The network must meet additional requirements for ISL sharing, cabling, and required settings on intermediate switches.
- The maximum transmission unit (MTU) of 9216 must be configured on all switches that carry MetroCluster IP traffic.
- The compliant switches must support Quality of Service (QoS)/traffic classification, explicit congestion notification (ECN), Layer 4 port-VLAN load-balancing policies to preserve order along the path, and Layer 2 Flow Control (L2FC).
- The cables connecting the nodes to the switches must be purchased from NetApp and supported by the switch vendor.

Figure 7 shows two identical FlexPod configurations, one at each site, that are connected by the ISLs. You can separate the NetApp ONTAP MetroCluster IP sites by up to 700 km. There are two types of connections between the Nexus switches and the NetApp storage controllers; they are used for client data traffic and MetroCluster IP data replication between the two ONTAP clusters. The ONTAP Mediator monitoring the solution from a third site enables the solution to perform an automated, unplanned switchover when one of the sites experiences a failure.

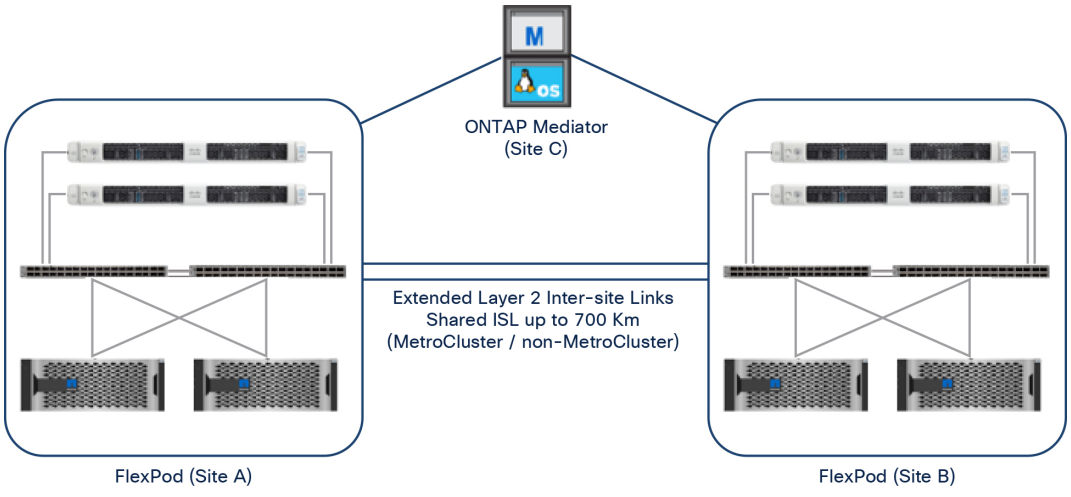


Figure 7.
FlexPod MetroCluster IP Solution Architecture with ISL Sharing and Compliant FlexPod Switches

For details about the requirements of using shared Layer 2 ISL connectivity for both MetroCluster IP and non-MetroCluster IP traffic between sites, deploying a MetroCluster IP solution with compliant switches, and using the ONTAP storage controllers that support MetroCluster IP deployment with compliant switches, please refer to NetApp documentation on Install a MetroCluster IP configuration: ONTAP MetroCluster and information on NetApp Hardware Universe.

Compute

The Cisco UCS compute infrastructure in FlexPod solutions is a revolutionary architecture for deploying and managing servers in the data center. The Cisco UCS infrastructure can now be deployed and managed from Cisco Intersight in IMM mode or be managed using Cisco UCS Manager. Following are some unique differentiators of Cisco UCS servers managed using Cisco UCS Manager running on Cisco UCS Fabric Interconnects. The benefits of using Cisco Intersight are discussed in the next section.

Cisco UCS Manager

- Embedded management: When the Cisco UCS servers are managed using the embedded firmware in the Fabric Interconnects, it eliminates the need for any external physical or virtual devices to manage the servers.
- Unified fabric: Cisco UCS servers in a blade-server chassis or rack servers connect to the Cisco UCS 6000 Series Fabric Interconnects, a single Ethernet cable is used for LAN, storage-area network (SAN), and management traffic. This converged I/O results in reduced cables, Small Form-Factor Pluggables (SFPs), and adapters, thereby reducing capital and operational expenses of the overall solution.
- Auto-discovery: By simply inserting the blade server in to a chassis or by connecting the rack server to a fabric interconnect, the compute resources are discovered and added to the inventory automatically without any management intervention. The combination of unified fabric and auto-discovery enables the wire-once architecture that Cisco UCS provides. Therefore, Cisco UCS architecture enables the compute capacity to be easily extended by using the existing external connectivity to LAN, SAN, and management networks.
- Policy-Based Resource Classification: When Cisco UCS Manager discovers a compute resource, , it can automatically classify it to a given resource pool based on defined policies. The policy-based resource classification capability is particularly useful in multi-tenant cloud deployments.
- Combined rack and blade server management: Cisco UCS Manager can manage Cisco UCS B-Series blade servers and Cisco UCS C-Series rack servers under the same Cisco UCS domain. This feature, along with stateless computing, makes compute resources truly hardware form-factor–agnostic.
- Model-based management architecture: The Cisco UCS Manager architecture and management database is model-based and data-driven. An open Extensible Markup Language (XML) application programming interface (API) is provided to operate on the management model. This API enables easy and scalable integration of Cisco UCS Manager with other management systems.
- Policies, pools, and templates: The management approach in Cisco UCS Manager is based on defining policies, pools, and templates, instead of cluttered configuration, enabling a simple, loosely coupled, data-driven approach in managing compute, network, and storage resources.
- Loose referential integrity: In Cisco UCS Manager, a service profile, port profile, or policies can refer to other policies or logical resources with loose referential integrity. A referred policy cannot exist at the time of authoring the referring policy or a referred policy can be deleted even though other

policies are referring to it. With this feature different subject matter experts can work independently from each-other, providing great flexibility where different experts from different domains, such as network, storage, security, server, and virtualization, work together to accomplish a complex task.

- Policy resolution: In Cisco UCS Manager, you can create a tree structure of organizational unit hierarchy that mimics the real-life tenants and/or organization relationships. You can define various policies, pools, and templates at different levels of organization hierarchy. A policy referring to another policy by name is resolved in the organizational hierarchy with the closest policy match. If no policy with a specific name is found in the hierarchy of the root organization, then the special policy named “default” is searched. This policy-resolution practice enables automation-friendly management APIs and provides great flexibility to owners of different organizations.
- Service profiles and stateless computing: A service profile is a logical representation of a server, carrying its various identities and policies. You can assign this logical server to any physical compute resource as far as it meets the resource requirements. Stateless computing enables procurement of a server within minutes; this task used to take days in older server management systems.
- Built-in multi-tenancy support: The combination of policies, pools, and templates; loose referential integrity; policy resolution in the organizational hierarchy; and a service profiles-based approach to compute resources makes Cisco UCS Manager inherently friendly to multi-tenant environments typically observed in private and public clouds.
- Extended memory: The enterprise-class Cisco UCS B200 M5 Blade Server extends the capabilities of the Cisco UCS portfolio in a half-width blade form factor. The Cisco UCS B200 M5 harnesses the power of the latest Intel Xeon Scalable series processor family CPUs with up to 3 TB of RAM (using 128-GB dual inline memory modules [DIMMs]): These modules allow the required ratio of huge virtual machines to physical servers in many deployments or allow large memory operations required by certain architectures such as big data.
- Simplified QoS: Even though Fibre Channel and Ethernet are converged in the Cisco UCS fabric, built-in support for QoS and lossless Ethernet makes it seamless. Network QoS is simplified in Cisco UCS Manager by representing all system classes in one GUI panel.

Cisco Intersight Platform

The Cisco Intersight platform is a SaaS based orchestration and operations platform that provides global management of Cisco UCS infrastructure located anywhere. The Cisco Intersight platform provides a holistic approach to managing distributed computing environments from the core to the edge. Its virtual appliance (available in the Essentials edition) offers you deployment options while still offering all the benefits of SaaS. This deployment flexibility can enable you to achieve a higher level of automation, simplicity, and operational efficiency.

Cisco UCS systems are fully programmable infrastructures. The Cisco Intersight platform provides a RESTful API to provide full programmability and deep integrations with third-party tools and systems. The platform and the connected systems are DevOps-enabled to facilitate continuous delivery. Customers have come to appreciate the many benefits of SaaS infrastructure-management solutions. Cisco Intersight monitors the health and relationships of all the physical and virtual infrastructure components. Telemetry and configuration information is collected and stored in accordance with Cisco’s information security requirements. The data is isolated and displayed through an intuitive user interface. The virtual appliance feature enables you to specify what data is sent back to Cisco with a single point of egress from the customer network.

The cloud-powered intelligence of Cisco Intersight can assist organizations of all sizes. Because the Cisco Intersight software gathers data from the connected systems, it learns from hundreds of thousands of devices in diverse customer environments. It combines this data with Cisco best practices to enable Cisco Intersight to evolve and become smarter. As the Cisco Intersight knowledge base increases, it reveals trends, and provides information and insights through the recommendation engine.

In addition to Cisco UCS server status and inventory, Cisco Intersight provides the Cisco UCS server Hardware Compatibility List (HCL) check for Cisco UCS server drivers. In this FlexPod validation, you can use the HCL check to verify that the correct Cisco UCS Virtual Interface Card (VIC) nfnic and nenic drivers are installed.

Cisco UCS B200 M5 Blade Servers

The Cisco UCS B200 M5 Blade Server (Figure 8) used in this FlexPod solution is a half-width blade.

Figure 8.
Cisco UCS B200 M5 Blade Server

It features:

- 2nd Gen Intel Xeon Scalable and Intel Xeon Scalable processors with up to 28 cores per socket
- Up to 24 Dial-on-Demand Routing 4 (DDR4) DIMMs for improved performance with up to 12 DIMM slots ready for Intel Optane DC Persistent Memory
- Up to two GPUs
- Two Small-Form-Factor (SFF) drive slots
  - Up to 2 Secure Digital (SD) cards or M.2 Serial Advanced Technology Attachment (SATA) drives
  - Up to 80 Gbps of I/O throughput with the Cisco UCS 6454 Fabric Interconnects

For more information about the Cisco UCS B200 M5 Blade Servers, please visit: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/datasheet-c78-739296.html.

Cisco UCS C220 M5 Rack Servers

The Cisco UCS C220 M5 Rack Server shown in Figure 9 is a high-density 2-socket rack server that can also be used to provide compute resources in this FlexPod solution.

Figure 9.
Cisco UCS C220 M5 Rack Server

It features:

- 2nd Gen Intel Xeon Scalable and Intel Xeon Scalable processors, 2-sockets
- Up to 24 DDR4 DIMMs for improved performance with up to 12 DIMM slots ready for Intel Optane DC Persistent Memory
- Up to 10 SFF 2.5-inch drives or 4 Large-Form-Factor (LFF) 3.5-inch drives (77 TB storage capacity with all NVMe PCIe Solid-State Disks [SSDs])
- Support for 12-Gbps serial-attached SCSI (SAS) modular redundant array of independent disks (RAID) controller in a dedicated slot, leaving the remaining PCIe Generation 3.0 slots available for other expansion cards
- Modular LAN-On-Motherboard (mLOM) slot that you can use to install a Cisco UCS VIC without consuming a PCIe slot
- Dual embedded Intel x550 10GBASE-T LAN-On-Motherboard (LOM) ports
- Up to 100 Gbps of I/O throughput with Cisco UCS 6454 Fabric Interconnects

For more information about the Cisco UCS B200 M5 Rack Servers, please visit: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/datasheet-c78-739281.html.

Cisco UCS 6400 Series Fabric Interconnects

The Cisco UCS Fabric Interconnects provide a single point for connectivity and management for all Cisco UCS servers (blade or rack) that connect to it. Typically deployed as an active-active pair, the system fabric interconnects integrate all components into a single, highly available management domain controlled by Cisco UCS Manager. The fabric interconnects manage all I/O efficiently and securely at a single point, resulting in deterministic I/O latency regardless of the topological location of a server or virtual machine in the system.

The Cisco UCS Fabric Interconnect provides both network connectivity and management capabilities for Cisco Unified Computing System. IOM modules in the blade server chassis support port channeling and, thus, better use of bandwidth. The IOMs support virtualization-aware networking in conjunction with the Fabric Interconnects and Cisco Virtual Interface Cards (VIC).

The Cisco UCS 6400 Series Fabric Interconnect is a core part of Cisco Unified Computing System, providing both network connectivity and management capabilities for the system. The Cisco UCS 6400

Series offers line-rate, low-latency, lossless 10/25/40/100 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and 32 Gigabit Fibre Channel functions.

The Cisco UCS 6454 54-Port Fabric Interconnect is a one-rack-unit (1RU) 10/25/40/100 Gigabit Ethernet, FCoE, and Fibre Channel switch that offers up to 3.82-Tbps throughput and up to 54 ports. The switch has twenty-eight 10-/25-Gbps Ethernet ports, four 1-/10-/25-Gbps Ethernet ports, six 40-/100-Gbps Ethernet uplink ports, and 16 unified ports that can support 10-/25-Gbps Ethernet ports or 8-/16-/32-Gbps Fibre Channel ports. All Ethernet ports can support FCoE.

The Cisco UCS 64108 Fabric Interconnect is a 2RU top-of-rack switch that mounts in a standard 19-inch rack such as the Cisco R-Series rack. The 64108 is a 10/25/40/100 Gigabit Ethernet, FCoE, and Fiber Channel switch that offers up to 7.42-Tbps throughput and up to 108 ports. The switch has 16 unified ports (port numbers 1–16) that can support 10-/25-Gbps SFP28 Ethernet ports or 8-/16-/32-Gbps Fibre Channel ports, seventy-two 10-/25-Gbps Ethernet SFP28 ports (port numbers 17–88), eight 1-/10-/25-Gbps Ethernet SFP28 ports (port numbers 89–96), and twelve 40-/100-Gbps Ethernet QSFP28 uplink ports (port numbers 97–108). All Ethernet ports can support FCoE. Although the Cisco UCS 64108 FI is supported in the FlexPod solution, it was not validated in this solution.

For more information about the Cisco UCS 6400 Series Fabric Interconnects, refer to the Cisco UCS 6400 Series Fabric Interconnects Data Sheet.

Cisco UCS 2408 Fabric Extender

The Cisco UCS 2408 connects the I/O fabric between the Cisco UCS 6454 Fabric Interconnect and the Cisco UCS 5100 Series Blade Server chassis, enabling a lossless and deterministic converged fabric to connect all blades and chassis together. Because the fabric extender is similar to a distributed line card, it does not perform any switching and is managed as an extension of the fabric interconnects. This approach removes switching from the chassis, reducing overall infrastructure complexity and enabling Cisco UCS to scale to many chassis without multiplying the number of switches needed, reducing total cost of ownership (TCO), and allowing all chassis to be managed as a single, highly available management domain.

The Cisco UCS 2408 Fabric Extender has eight 25-Gigabit Ethernet, FCoE-capable, SFP28 ports that connect the blade chassis to the fabric interconnect. Each Cisco UCS 2408 provides 10-Gigabit Ethernet ports connected through the midplane to each half-width slot in the chassis, giving it a total of 32 10-GbE interfaces to Cisco UCS blades. Typically configured in pairs for redundancy, two fabric extenders provide up to 400 Gbps of I/O from Cisco UCS 6400 to 5108 fabric interconnect chassis.

Cisco UCS 1400 Series VICs

Cisco VICs support Cisco SingleConnect technology, which provides an easy, intelligent, and efficient way to connect and manage computing in your data center. Cisco SingleConnect unifies LAN, SAN, and systems management into one simplified link for rack servers and blade servers. This technology reduces the number of network adapters, cables, and switches needed and radically simplifies the network, reducing complexity. Cisco VICs can support 256 Express (PCIe) virtual devices, either virtual Network Interface Cards (vNICs) or virtual Host Bus Adapters (vHBAs), with a high rate of I/O Operations Per Second (IOPS), support for lossless Ethernet, and 10-/25-/40-/100-Gbps connection to servers. The PCIe Generation 3 x16 interface helps ensure optimal bandwidth to the host for network-intensive applications, with a redundant path to the fabric interconnect. Cisco VICs support NIC teaming with fabric failover for increased reliability and availability. In addition, they provide a policy-based, stateless, agile server infrastructure for your data center.

The Cisco VIC 1400 Series is designed exclusively for the M5 generation of Cisco UCS B-Series Blade Servers and Cisco UCS C-Series Rack Servers. The adapters can support 10-/25-/40-/100-GE and FCoE. They incorporate Cisco's next-generation Converged Network Adapter (CNA) technology and offer a comprehensive feature set, providing investment protection for future feature software releases.

Solution Design

Solution Summary

The FlexPod MetroCluster IP with Cisco VXLAN EVPN Multi-Site solution is a data center solution that uses an active-active data center design to provide BC/DR in the event of a disaster or a data center–wide failure. The solution helps ensure the availability of the virtual server infrastructure (VSI) in at least one data center site and provides seamless workload mobility by enabling Layer 2 extension and Layer 3 forwarding across sites. The two active-active data centers can be in different geographically locations as long as the bandwidth, round-trip latency (<7 ms), round-trip jitter (<3 ms) and drop (<0.01%) requirements for the NetApp MetroCluster IP storage are met.

As with other FlexPod solutions, this solution incorporates technology, design and product best practices and uses a highly resilient design across all layers of the solution. FlexPod data center solutions are typically built using different models from the following component families for compute, storage, and network:

- Cisco Unified Computing System (Cisco UCS)
- Cisco Nexus switches
- Cisco MDS switches (not included in this design)
- NetApp Fabric Attached Storage (FAS)/All-Flash FAS (AFF)/All SAN Array (ASA) systems

The critical components of the virtualized server infrastructure in the active-active data center design solution are as follows:

- Cisco Unified Computing System (Cisco UCS)
- NetApp MetroCluster IP storage using NetApp AFF A700 arrays
- Cisco VXLAN BGP EVPN Multi-Site network fabric, managed using Cisco DCNM
- Nexus 9000 Family of switches (VXLAN fabric and Inter-Site Network)
- VMware vSphere

This solution also uses a separate, dedicated Layer 2 network for the MetroCluster IP synchronous data-replication traffic between data centers. Several MetroCluster IP switches are supported for a typical FlexPod configuration, including the Cisco Nexus 3132Q-V, 3232C, and 9336C-FX2. The Nexus 3132Q-V switches are used for this solution. These switches serve as both ONTAP cluster switches and MetroCluster IP switches.

The solution was built and verified in Cisco labs using the following models of products from the different component families (Cisco UCS Software, NetApp storage, Cisco Nexus switches, and VMware). The components used in each data center site are shown in Table 1.

Table 1. Solution Components per Data Center Site

FlexPod Data Center	Component		Notes
Network (Cisco DCNM Managed VXLAN BGP EVPN Multi-Site Fabric)	**Site A**	**Site B**	
	Cisco Nexus 9364C x 2	Cisco Nexus 9504 x 2	Spine Switches
	Cisco Nexus 9336C-FX2 x 2	Cisco Nexus 9336C-FX2 x 2	Leaf Switches – To Cisco UCS Domains and NetApp
	Cisco Nexus 93180LC-EX x 2	Cisco Nexus 93180LC-EX x 2	Border Leaf Switches – For External or Outside connectivity
	Cisco Nexus 93240YC-FX2 x 2	Cisco Nexus 93240YC-FX2 x 2	Border Gateways with CloudSec encryption support for inter-site connectivity
Storage Infrastructure (NetApp MetroCluster IP)	**Site A**	**Site B**	
	NetApp AFF A700	NetApp AFF A700	Storage Controller
	DS-224C x 2	DS-224C x 2	Disk Shelf
Compute Infrastructure (Cisco UCS Servers)	**Site A**	**Site B**	
	Cisco UCS 6454 FI x 2	Cisco UCS 6454 FI x 2	Fabric Interconnects
	Cisco UCS 5108 chassis with 2 x IOM 2408, 3 x UCS B200 M5 servers with VIC 1440 MLOM adapter	Cisco UCS 5108 chassis with 2 x IOM 2408, 3 x UCS B200 M5 servers with VIC 1440 MLOM adapter	Blade Server Chassis with Servers
Virtualization Layer (VMware)	**Site A**	**Site B**	
	VMware vSphere 7.0U1	VMware vSphere 7.0	Hypervisor
	VMware vSwitch, vDS	VMware vSwitch, vDS	Virtual Switching
Management & Monitoring	vCenter Server Appliance 7.0U1 Cisco DCNM (HA cluster) 11.5(1), Cisco Network Insights (Optional) Cisco Intersight, Cisco UCS Manager NetApp ActiveIQ Unified Manager, NetApp Virtual Storage Console, NetApp ONTAP Mediator		

Table 2 lists the components used in the replication network for the NetApp MetroCluster IP solution:

Table 2. Solution Components for Data Replication Network in Each Site

FlexPod Data Center	Component		Description
NetApp MetroCluster IP	**Site A**	**Site B**	
	Cisco Nexus 3132Q-V x 2	Cisco Nexus 3132Q-V x 2	Synchronous Replication Network Switches

Solution Topology

Figure 10 shows the end-to-end design for the active-active data center solution.

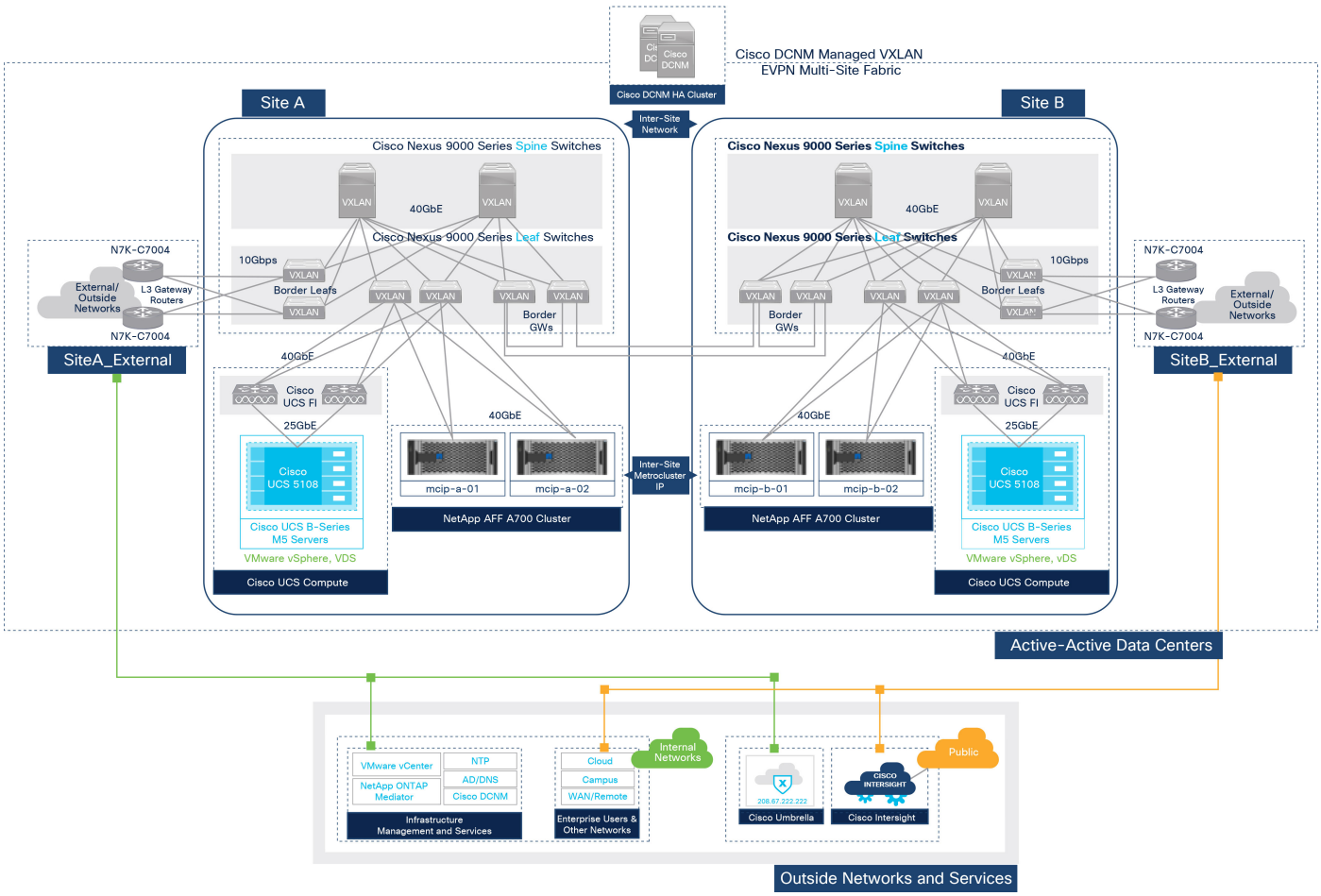


Figure 10.
Solution Design – High-Level

Figure 11 shows the synchronous replication network for NetApp MetroCluster IP that is used in this solution.

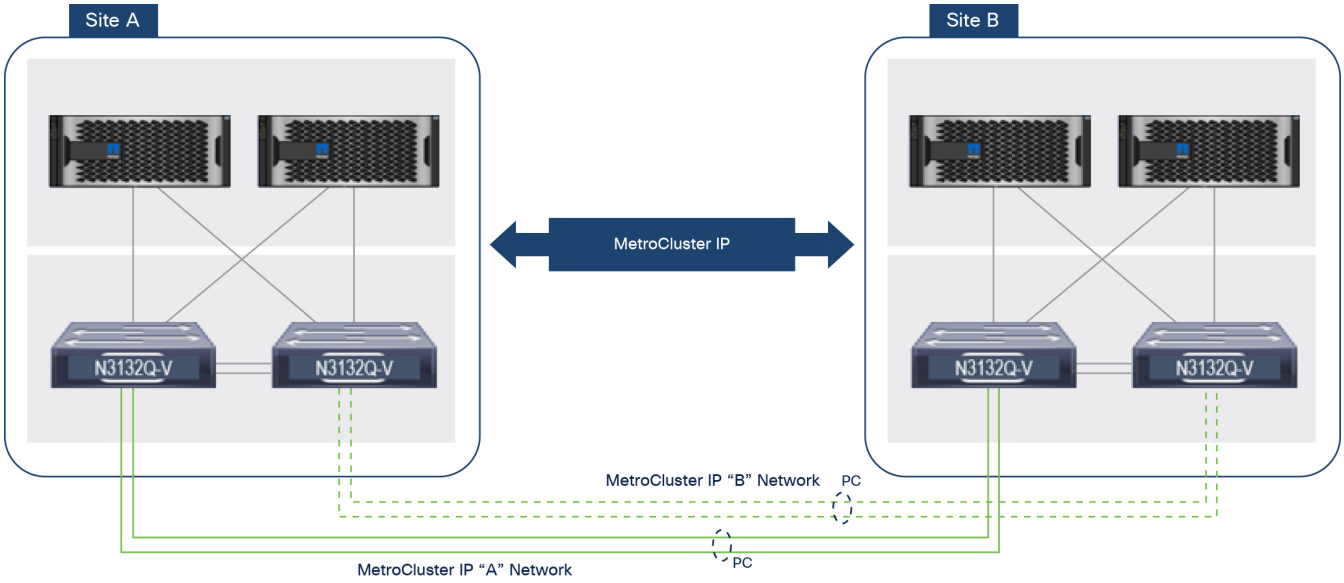


Figure 11.
NetApp MetroCluster IP Network – High-Level Design

Network Design

Cisco VXLAN EVPN Multi-Site Fabric

The Cisco VXLAN Multi-Site fabric provides the network fabric for the two active-active data centers and the connectivity between the data center sites in this FlexPod MetroCluster IP solution. The end-to-end network consists of distinct VXLAN fabrics, one in each data center site, interconnected by an inter-site network (ISN). Cisco DCNM, serving as SDN controller, centrally manages the end-to-end fabric. Cisco DCNM is deployed as a cluster of 3 nodes for high availability. The end-to-end fabric is designed to be highly resilient, with no single point of failure. The solution design incorporates technology and product best practices and was built using Nexus 9000 Series switches. The solution is horizontally scalable without compromising on latency or performance.

Figure 12 shows a high-level view of the end-to-end VXLAN EVPN Multi-Site fabric used in this FlexPod solution.

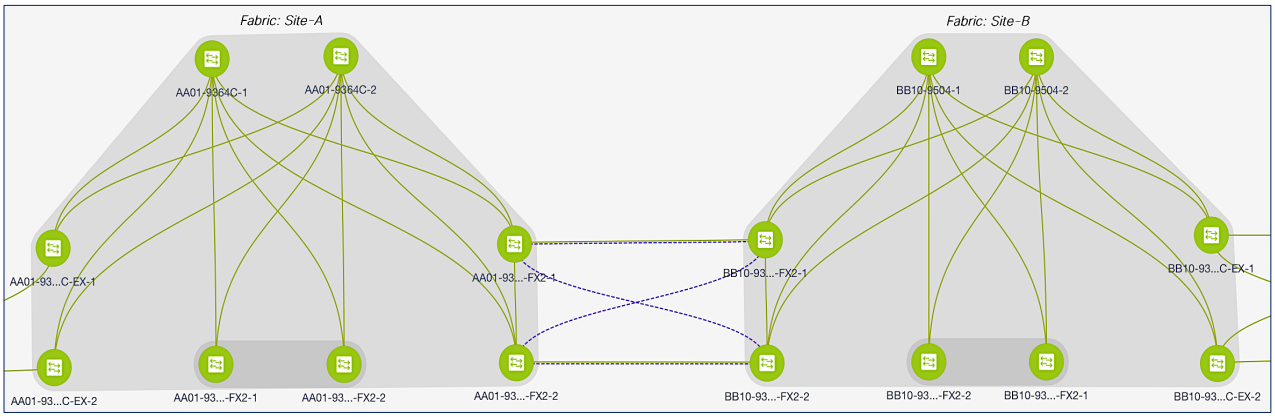


Figure 12.
Cisco DCNM – VXLAN EVPN Multi-Site Fabric Topology

In upcoming sections, the following aspects of the this VXLAN Multi-Site design is discussed in detail:

- Intra-site design for connectivity within a site, including connectivity from each site to outside/external networks
- Inter-site design for connectivity between sites
- Tenancy design
- Connectivity for FlexPod Infrastructure
- Connectivity for applications hosted on FlexPod infrastructure
- High availability

Fabric Design - Intra-Site

As stated before, each data center in the VXLAN EVPN Multi-Site architecture is a separate VXLAN fabric, built using Nexus 9000 Series switches in a 2-tier spine-leaf Clos topology. Figures 13 and 14 show the high-level topology intra-site design for the two data centers (Site-A, Site-B) in the solution. The design at each site is highly resilient, with no single points of failure. Cisco DCNM manages each site as well as the end-to-end multi-site fabric. Cisco DCNM is located outside the VXLAN multi-site fabric. The fabric switches in both sites have out-of-band IP reachability to Cisco DCNM for management. Both sites are designed and deployed identically in this solution.

Intra-Site Design - Site-A

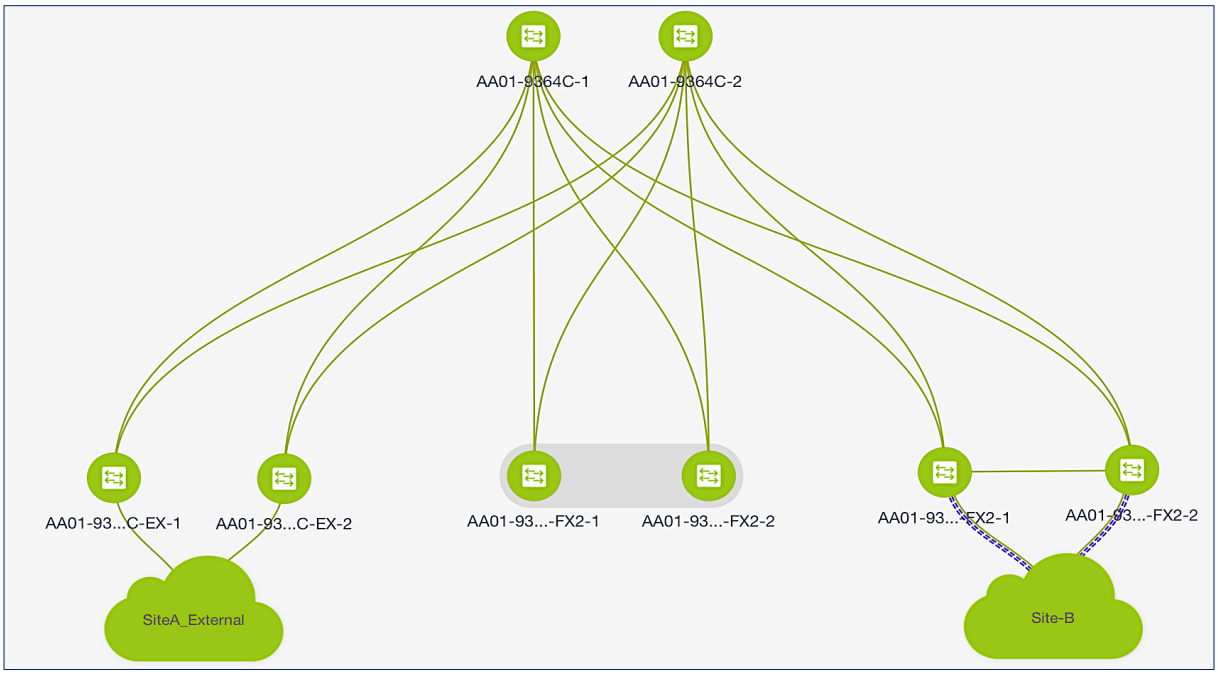


Figure 13.
Cisco DCNM - Site-A Data Center Fabric

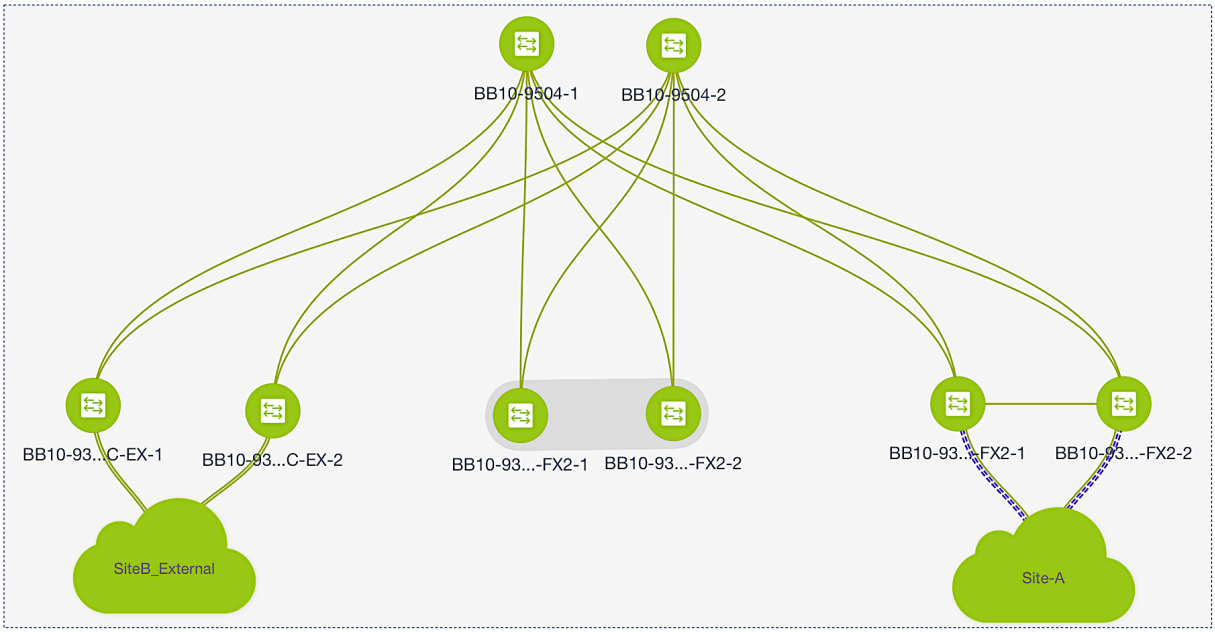


Figure 14.
Cisco DCNM – Site-B Data Center Fabric (Intra-Site Design)

A pair of spine switches and three pairs of leaf switches are deployed in each data center site. One pair of leaf switches are used for connecting to FlexPod compute and storage infrastructure that provides the VSI in each data center. A second pair of leaf switches are deployed as border leaf switches for connecting to an outside/external network from each site. The third pair of leaf switches serve as border gateway switches for inter-site connectivity between the active-active data centers. In addition to providing high-speed core connectivity, the spine switches provide redundant route-reflector (RR) functions for Internal Border Gateway protocol (iBGP) and Rendezvous-Point (RP) functions for IP Multicast deployed in each fabric. You can collapse the functions that the 3-leaf switch pairs provide into a minimal fabric consisting of 2 spine switches and 2 leaf switches. However, for scalability, Cisco recommends deploying these functions on separate leaf switch pairs.

The leaf switches in each site connect to both spine switches using 40-GE links. Note that spine switch pairs have no direct links, nor do leaf switch pairs in a Clos topology. The border gateway switches do have a cross-link connecting them, but this link is not part of the Clos-based intra-site fabric; it is an external link for inter-site connectivity.

With Cisco DCNM Fabric Builder you can provision and deploy the fabric in each site using the **Easy_Fabric_11_1** template.

The fabric settings specified for each site include:

- IP addressing (IPv4 or IPv6)
- Interior Gateway Protocol (IGP) routing protocol (Intermediate System-to-Intermediate System [IS-IS], Open Shortest Path First (OSPF), or Exterior Border Gateway Protocol [eBGP])
- Overlay routing and address learning (iBGP, EPVN Address Family)
- Layer 2 multi-destination traffic (BUM) traffic handling (Ingress Replication, IP Multicast)
- IP Multicast (Protocol Independent Multicast-Sparce Mode [PIM-ASM] or Bidirectional PIM (BiDir-PIM)

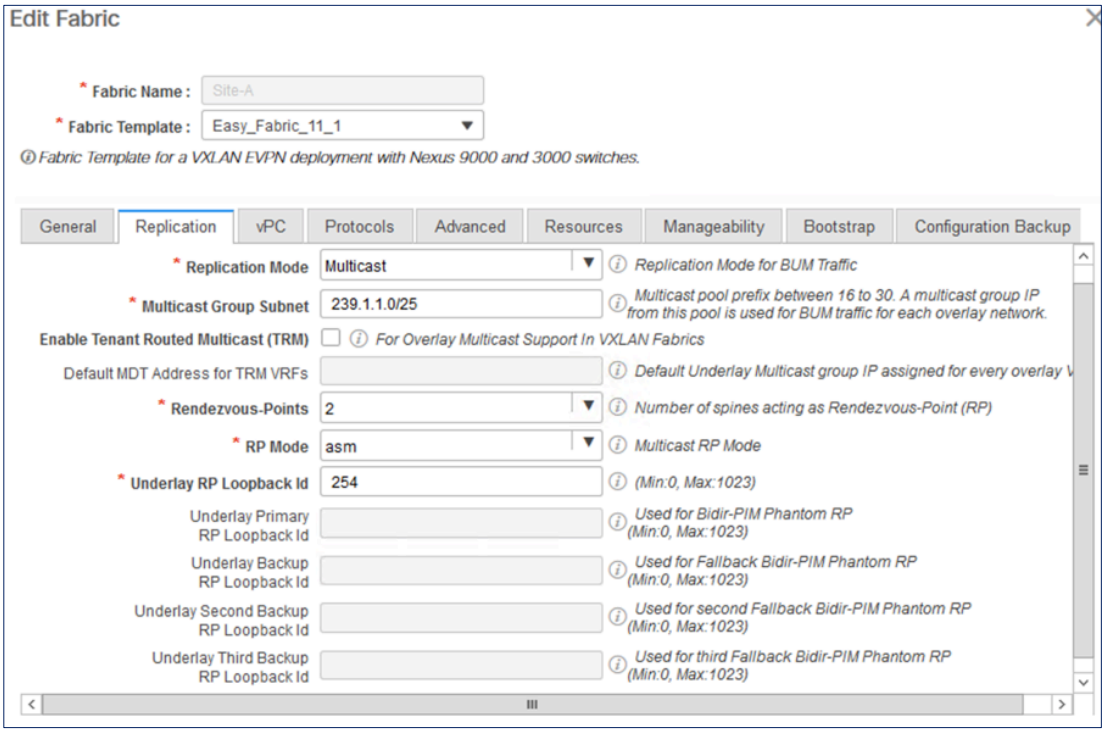
In this design, all links in the underlay are deployed as point-point links using an IPv4/30 subnet mask with OSPF for routing in the underlay for VXLAN Tunnel Endpoint (VTEP)-to-VTEP reachability. IP Multicast using PIM-ASM is used for forwarding BUM traffic through the fabric. All links in the fabric are also configured to use jumbo MTU. The "**Intra-Site Design Considerations**" section below provides additional information and reasoning for the choices made in this solution.

Intra-Site Design Considerations

This section discusses the different design options and factors to consider when deploying a VXLAN EVPN fabric.

- Forwarding of Broadcast, Unknown Unicast, and Multicast (BUM) traffic: Switched Ethernet networks use a flood-and-learn mechanism to forward traffic to unknown (yet-to-be learned) destinations or to a small subset of destinations or to all destinations in the same Layer 2 broadcast domain. This data-plane method helps ensure reachability and enables address learning. When the Layer 2 Ethernet segment is extended across a VXLAN fabric, a similar mechanism is needed, but the BUM traffic must now be forwarded across an IP transport or underlay so that Ethernet endpoints can continue to operate as they normally do, unaware that they are connecting through a VXLAN fabric. To provide this capability, VXLAN fabrics also use a similar data-plane flood-and-learn mechanism to flood traffic to all VTEPs in the fabric handling traffic for that Layer 2 network. VXLAN provides two options for flooding BUM traffic across an IP underlay: IP Multicast or Ingress Replication. With Ingress (headend) Replication, the local VTEP or leaf receiving BUM traffic replicates and sends an individual copy to each remote VTEP. If IP Multicast is used, the VTEP or leaf forwards that traffic to the IP Multicast group associated with that Layer 2 network. VXLAN fabric uses the Internet Group Management Protocol (IGMP)/PIM and VTEPs to join the multicast group in order to forward and receive BUM traffic for a given Layer 2 network. Cisco recommends using IP Multicast because it is a more efficient method for forwarding traffic to multiple remote destinations or VTEPs across an IP network. It also limits the scope of the flooding to only those VTEPs handling traffic for that Layer 2 network with Ethernet endpoints connected to it. Cisco DCNM Fabric Builder deploys the VXLAN fabric in this solution, and it follows Cisco best-practice recommendations and deploys IP Multicast by default. This FlexPod design, therefore, uses IP Multicast for forwarding BUM traffic across the VXLAN fabric in each data center site. For BUM forwarding across data centers, see "**Inter-site Design – Interconnecting Data Centers"** section of this document.
- Underlay IP Multicast: A multicast routing protocol is necessary when running IP Multicast for forwarding BUM traffic across a VXLAN fabric. This type of multicast is different from any IP multicast that you may run in the overlay for applications or services deployed in the edge networks. Cisco VXLAN fabrics support both PIM-ASM and PIM-Bidir for forwarding BUM traffic within the data center. Both protocols use Rendezvous-Point (RP), and the spine switches in each data center are centralized and therefore an ideal location for RP functions. Rendezvous-Points are deployed on two spine switches in each data center for redundancy. In this FlexPod solution, PIM-ASM is used in both data center sites, and Cisco DCNM Fabric Builder automatically configures the necessary configuration is when the fabric is deployed.
- IP Multicast Addressing: The IP Multicast deployed for flooding BUM traffic within a data center site requires multicast group addresses to be allocated/associated for each network being deployed for forwarding BUM traffic for that network across the VXLAN fabric. However, as the number of VXLAN segments grows, the number of multicast groups and the forwarding states that need to be maintained on the fabric switches also grow. For this reason, the IP Multicast group address assigned for each network defaults to the same address unless explicitly specified otherwise. Using

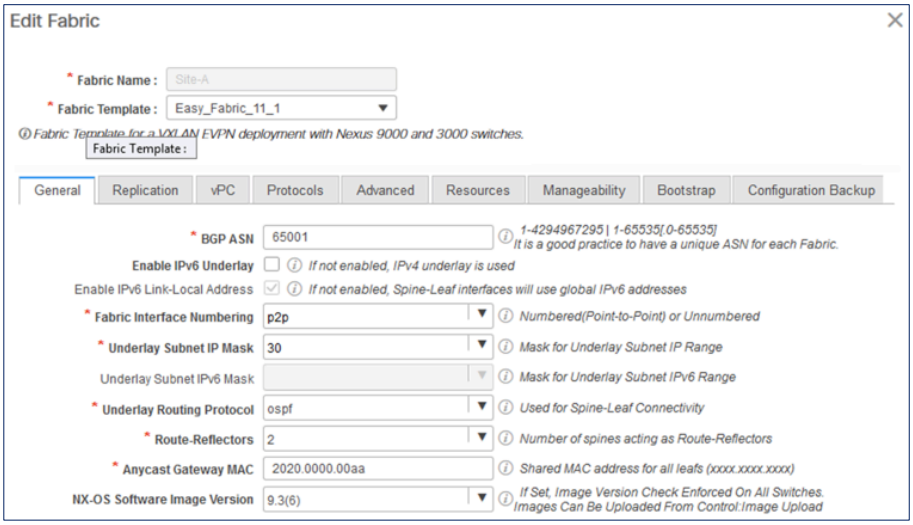
the same multicast group reduces the control plane resources used, but it also means that a VTEP could see multi-destination traffic for networks that are not deployed or handled by that VTEP. However, the VTEP will ensure that the BUM traffic forwarded to Ethernet endpoints in the edge network is only traffic destined for that network. The VTEP will forward only those packets whose VNID in the VXLAN header matches the VNID of a local segment. You can change this setup to suit the needs of your deployment. For this FlexPod solution, it is recommended that infrastructure networks, particularly the storage data network, use a dedicated IP Multicast group to isolate the traffic, an action that also makes it easier for you to monitor and troubleshoot if needed. The following screenshot shows the multicast configuration used in this solution for Site-A. This configuration is identical in Site-B.



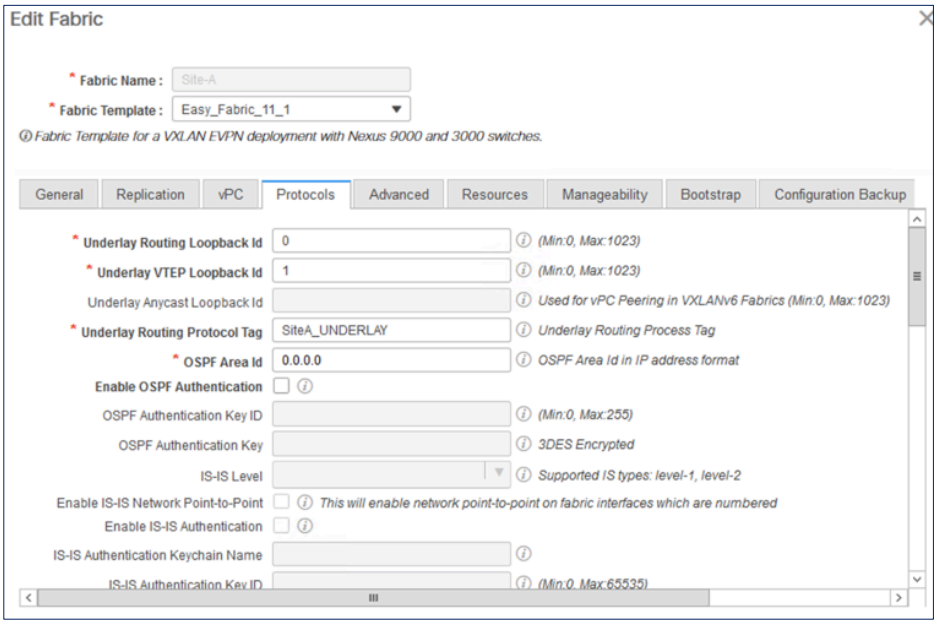
- Address learning – MP-BGP EVPN: VXLAN fabrics use flood-and-learn, using either IP Multicast or Ingress Replication for forwarding BUM traffic across an IP underlay. In a VXLAN fabric, the flood-and-learn method provides endpoint reachability, address learning, as well as discovery of remote VTEPs for the learned addresses. This solution uses IP Multicast to efficiently forward the BUM traffic to the VTEPs handling traffic for that network. However, large amounts of multicast traffic can still limit the scalability of a data center fabric. With VXLANs, Layer 2 segments can now span large boundaries, even across data centers. To address this situation, the Internet Engineering Task Force (IETF) standardized on the Multiprotocol Border Gateway Protocol (MP-BGP), an Internet-scale routing technology and Ethernet VPNs (EVPNs) to do address learning in the control plane. The EVPN address family in BGP carries both the Media Access Control (MAC) and IP address information of the endpoints along with other information such as the network and tenant (or Virtual Route Forwarding [VRF]) instances to which they belong. This method reduces the flooding of BUM traffic due to unknown unicasts and provides a more optimal forwarding of the Layer 2 and Layer 3 traffic within a VXLAN fabric. By default, Cisco DCNM Fabric Builder deploys the VXLAN fabric using BGP so that when you deploy networks you can use the EVPN address family to advertise and learn addresses.

- Address Resolution Protocol (ARP) flooding across a VXLAN fabric: When endpoints in an edge network originate an ARP request using a broadcast MAC address as the destination, the receiving VXLAN leaf or VTEP switch floods that traffic across the fabric to all remote VTEPs using the multicast-group address associated with that network. However, if ARP suppression is enabled on the Cisco VXLAN fabric, Cisco leaf switches will first inspect the ARP request and do a local lookup for the destination IP that it is trying to resolve. If local VTEP has learned the location of the destination IP (with MP-BGP VPN), it will respond to that ARP request locally without flooding that ARP request across the VXLAN fabric. ARP suppression is supported only when the fabric is doing Layer 3 forwarding. This feature is enabled in this solution.
- Integrated Routing and Bridging (IRB): VXLAN fabrics support IRB where each ingress leaf switch or VTEP can do either Layer 2 or Layer 3 forwarding for the local networks attached to it. The edge networks are mapped to a Layer 2 or Layer 3 VXLAN segment (VNI) and then forwarded across the fabric. The Layer 3 VNI provides Layer 3 segmentation (VRF) to support multi-tenancy. Similarly, Layer 2 VNI provides segmentation for Layer 2 traffic through the fabric. The IRB in VXLAN fabrics can be of two types: symmetric or asymmetric. Cisco VXLAN fabrics and Nexus switches support symmetric IRB because it is more scalable and less complex from a configuration perspective. Therefore, symmetric IRB is used in this FlexPod solution for the data center fabric in each site. As you deploy networks, Cisco DCNM deploys the corresponding configuration on the appropriate VTEPs as needed.
- Underlay network: An underlay routing protocol is necessary for reachability between leaf (VTEP) switches in a VXLAN fabric. An IGP such as OSPF or IS-IS is recommended because it can use the Clos-based spine-and-leaf topology to provide multiple equal-cost forwarding paths between leaf switches, with rapid convergence during network failures. You also can use BGP for the underlay routing, but IGP is preferred. OSPF is used in this FlexPod solution.

  The interface addressing that Cisco DCNM Fabric Builder deploys for connectivity between leaf- and-spine switches is either IP unnumbered or point-to-point with a /30 or /31 subnet mask. This setup minimizes the number of IP addresses required for the underlay network in a VXLAN fabric. Alternatively, the underlay network can also use IPv6. The following screenshot shows the interface addressing specified for the VXLAN fabric in Site-A.



Cisco DCNM Fabric Builder deploys multiple loopback interfaces in the VXLAN fabric it deploys. Loopbacks are used as router IDs for the underlay routing protocol, as the VTEP IP address in the VXLAN encapsulated packets, for multicast Rendezvous-Point, etc. The following screenshot shows the underlay loopbacks used for the VXLAN fabric in Site-A.



- Overlay routing: As discussed earlier, BGP, specifically internal BGP (iBGP) is used for overlay routing in VXLAN fabrics. iBGP requires either a full mesh connectivity between all the switches or they all need to peer with a route reflector . Any routes (EVPN) that a route reflector receives from one leaf switch will be forwarded to the other leaf switches that it is peered with. Route reflectors are typically deployed in a central location, and the spine switches are a good location for route reflectors in a spine-and-leaf leaf topology. Cisco DCNM Fabric Builder deploys the BGP and route-reflector configuration on relevant switches when the fabric is deployed. Two route reflectors set up are deployed by default on spine switches in each site.
- Distributed anycast gateway (GW): To facilitate flexible workload placement, endpoint mobility, and optimal traffic forwarding across a data center fabric, VXLAN fabrics use distributed anycast gateways, where each Leaf switch serves as a first-hop default gateway for the local endpoints in that network. To enable this function, the same gateway IP address and virtual anycast gateway MAC address (2020.0000.00aa) is used across all leaf switches where the Layer 3 network is deployed. This setup helps ensure that the ARP entries for the Gateway IP are still valid. Even when an endpoint in the network moves to another leaf switch in the same data center, the ARP entry for the Gateway IP remains the same. VXLAN fabrics, therefore, do not require protocols (Hot Standby Router Protocol [HSRP], First Hop Redundancy Protocol [FHRP]) that ensure the availability of a default gateway because each leaf switch is a default gateway that can provide routing for local edge networks. As Layer 3 networks are deployed, Cisco DCNM will deploy the corresponding anycast GW function on relevant leaf switches as needed.

- Multitenancy: You can use MP-BGP to enable multitenancy in VXLAN fabrics. The MP-BGP with EVPN address family uses the same concepts as an MPLS-based Layer 3 VPN (L3VPN) to provide tenant separation in the control plane. Similar to MPLS L3VPNs, a Route Distinguisher (RD) ensures the global uniqueness of addresses belonging to different VPNs (or VRFs) when advertising them to other BGP peers, and route targets (RT) will associate the addresses to a VRF instance for flexible exporting and importing of routes between peers in the same VRF and across VRFs. In the data plane, VXLAN uses VNIDs to provide segmentation in the overlay network by mapping edge networks to a VNID and by enforcing VNID and VRF boundaries. Two tenants are used in this design—one for infrastructure connectivity and one for applications. You can deploy additional tenants as needed. The infrastructure or "Foundation" tenant isolates all connectivity required to build and maintain the virtualized server infrastructure; it includes compute, storage, and virtualization layer-related connectivity and management. The "Application" tenant is used for applications hosted on the infrastructure. As tenants are provisioned, Cisco DCNM deploys the corresponding configuration as needed. The RD and RT that is deployed for the "Foundation" VRF on a leaf switch in this FlexPod design is shown below.

```
vrf context fpv-foundation_vrf
  description FPV_Foundation_VRF
  vni 30000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
```

- MTU: VXLAN uses a MAC-in-IP/User Datagram Protocol (UDP) encapsulation, resulting in a 50B overhead on VXLAN-tagged frames. However, per IETF, intermediate switches in a VXALN fabric can fragment packets but a VTEP cannot. To avoid all fragmentation in the fabric, the MTU within the fabric should be at least 50B higher than the MTU of the edge traffic being transported by the fabric. The Cisco DCNM Fabric Builder implements best-practice recommendations and uses a default MTU of 9216B.
- VNID allocation and naming conventions: The VXLAN fabric deployed in this design uses VNIDs in the 20000s range for Layer 2 networks and 30000s range for Layer 3 networks. Similarly, all naming includes a "_<type>" tag to indicate the type of object or construct the name is referring to. Both of these are not required but can be helpful for troubleshooting purposes.

Intra-Site Design - Core Connectivity

Core connectivity is the connectivity between leaf-and-spine switches in each data center fabric. Figure 15 shows the detailed physical topology of the VXLAN MP-BGP EVPN core in each site, built using Cisco Nexus 9000 Series switches (spine-and-leaf switches).

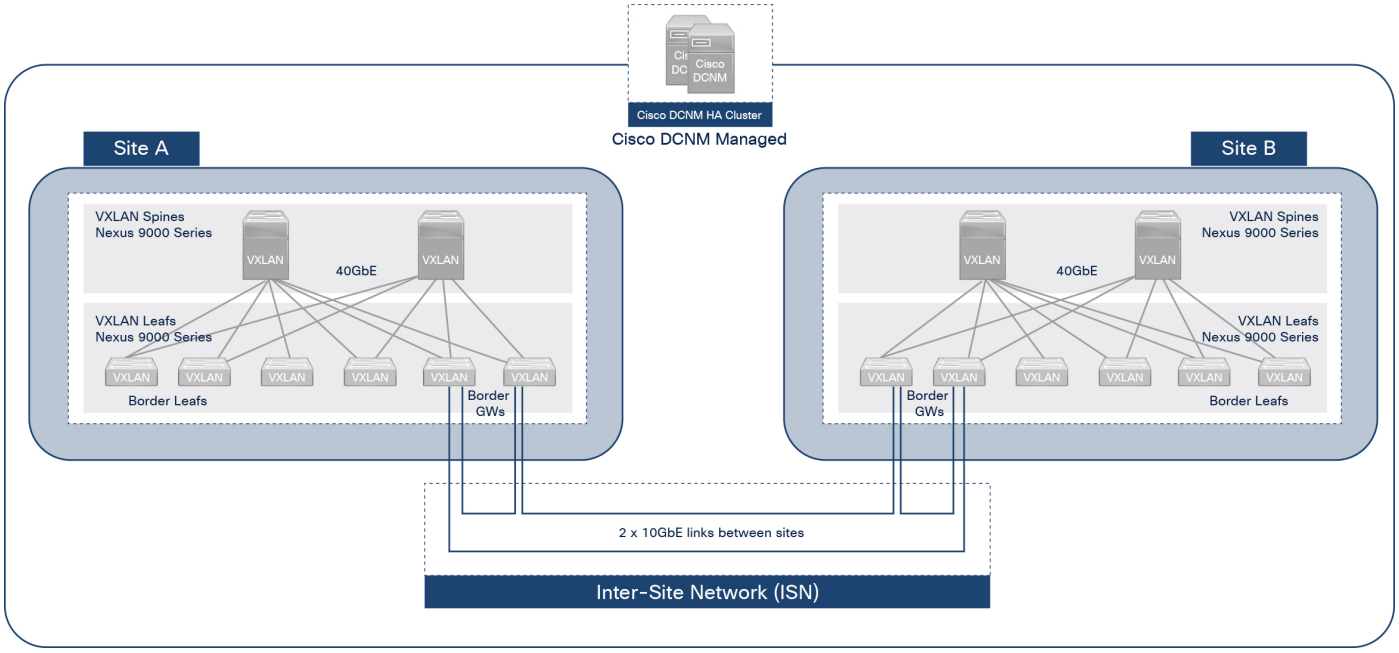


Figure 15.
Intra-Site VXLAN Fabric – Core Connectivity

The three pairs of leaf switches provide connectivity to outside/external networks, FlexPod compute, and storage infrastructure using Virtual PortChannels (vPCs) and to the second data center in the active-active design. All links in the core are 40GbE links.

Intra-Site Design - Edge Connectivity

Edge connectivity is the connectivity from the VXLAN fabric leaf switches to endpoints in the edge network. For Layer 2 connectivity, Cisco switches support link aggregation using Link Aggregation Control Protocol (LACP) to connect to endpoints in the edge network. The connectivity can be a port channel with multiple links from a single leaf switch or a vPC, where the multiple links span a leaf switch pair acting as a single logical entity. Link aggregation provides both higher aggregate bandwidth and resiliency, but vPCs provide a higher level of resiliency by providing both node and link-level resiliency and, therefore, preferred when possible.

In this FlexPod design, connectivity to both Cisco UCS compute and NetApp storage infrastructure in each site uses vPCs, and connect to the same leaf switch pair as shown in Figures 16 and 17.

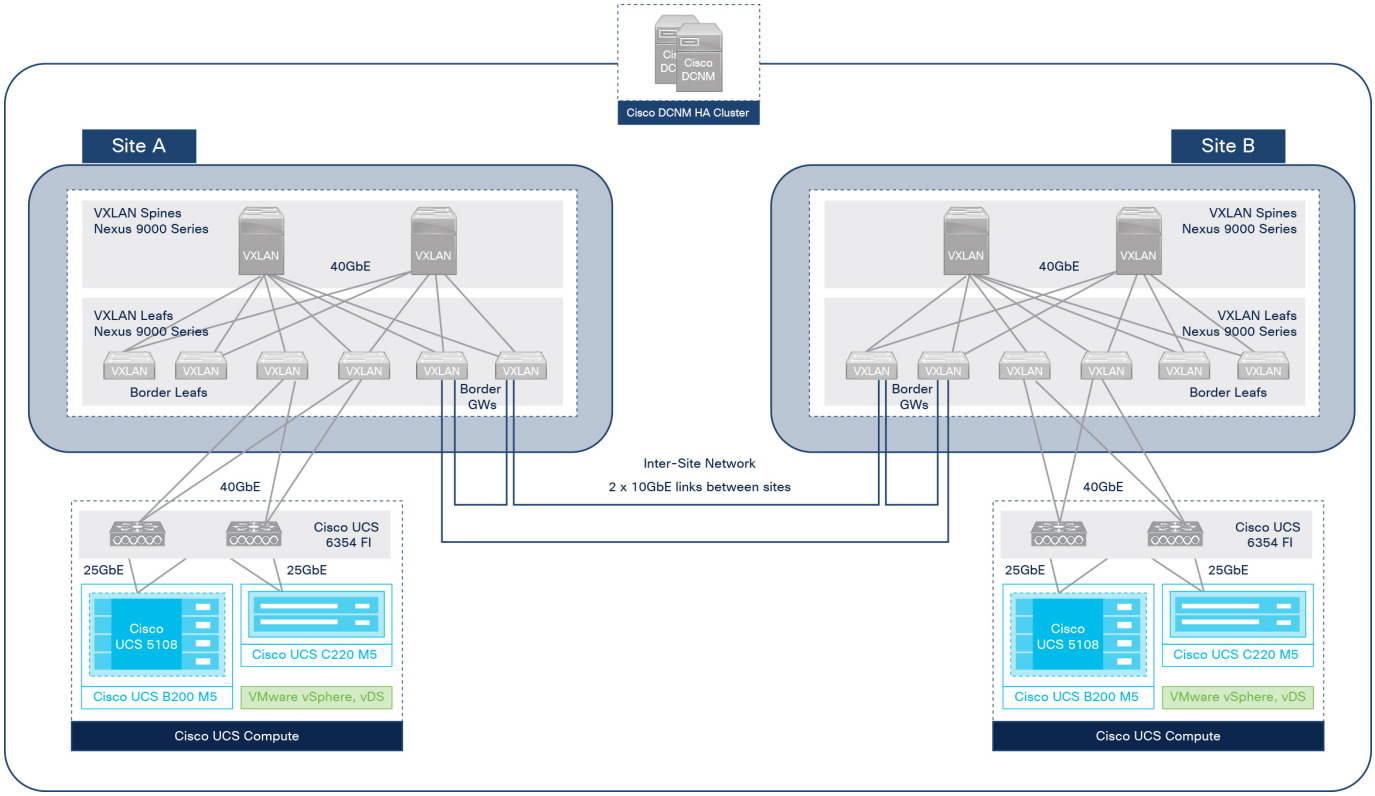


Figure 16.
Intra-Site VXLAN Fabric – Connectivity to Cisco UCS Compute

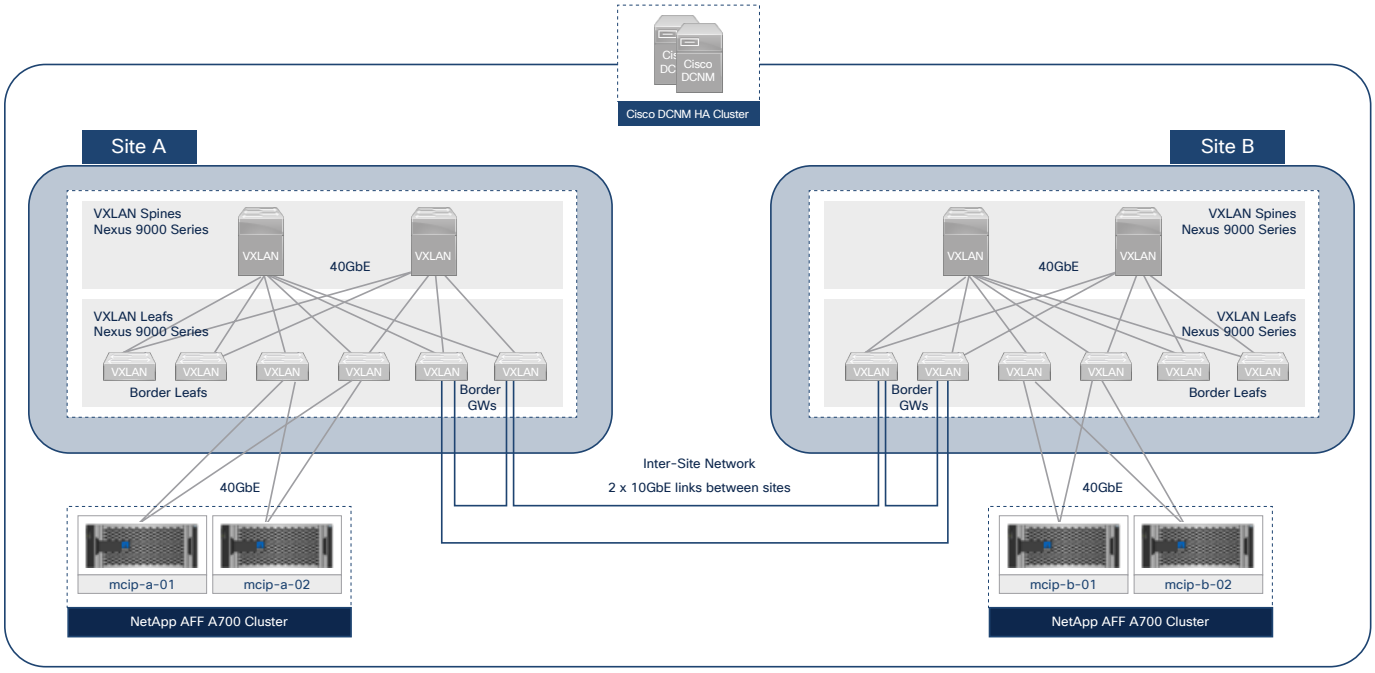


Figure 17.
Intra-Site VXLAN Fabric – Connectivity to NetApp Storage

The vPC configuration for each site is also identical in both sites for vPCs going to Cisco UCS domain and NetApp arrays.

Intra-Site Design - Connectivity to Outside/External Networks (VMware vCenter, NetApp ONTAP Mediator)
External connectivity refers to the connectivity from the VXLAN data center fabric to networks outside the VXLAN fabric, either internal or external to the enterprise. In this FlexPod design, external connectivity is necessary to connect to networks and services that are necessary for deployment, managment, and operation of the FlexPod infrastructure. Applications deployed on the FlexPod infrastructure will likely require access to outside/external networks and services as well. In the FlexPod MetroCluster IP solution, VMware vCenter and NetApp ONTAP Mediator are two critical components that are located outside the fabric. Both data center sites have separate connections to outside networks for seamless failover and business continuity if a failure occurs.

The outside/external connectivity can be Layer 2 for scenarios such as migrating workloads to/from infrastructures in other parts of the enterprise. However, for most scenarios such as accessing services on the Internet or infrastructure located elsewhere in the internal enterprise network, the connectivity is typically Layer 3. Cisco VXLAN fabrics offer multiple options for enabling Layer 3 connectivity to outside networks, including VRF-to-VRF hand-off to an MPLS-VPN network or an IP network using VRF or VRF-Lite, or the hand-off could be to a non-VRF IP network. VRF-to-VRF hand-off enables multitenancy in the VXLAN fabric to be extended to external routed domains.

The external connectivity in this FlexPod MetroCluster IP solution uses VRF-to-VRF hand-off to an IP network. The border leaf switches in each site connects to external gateways with VRF-Lite deployed on the gateway side. You can use Cisco DCNM Fabric Builder to deploy the external connectivity. Cisco DCNM uses a separate fabric for the external gateways that each site uses for outside connectivity. This external fabric is deployed using the **External_Fabric_11_1** template. Figure 18 shows a high-level view of the connectivity from each site (Site-A, Site-B) to the external gateways in the external fabrics (**SiteA_External, SiteB_External**).

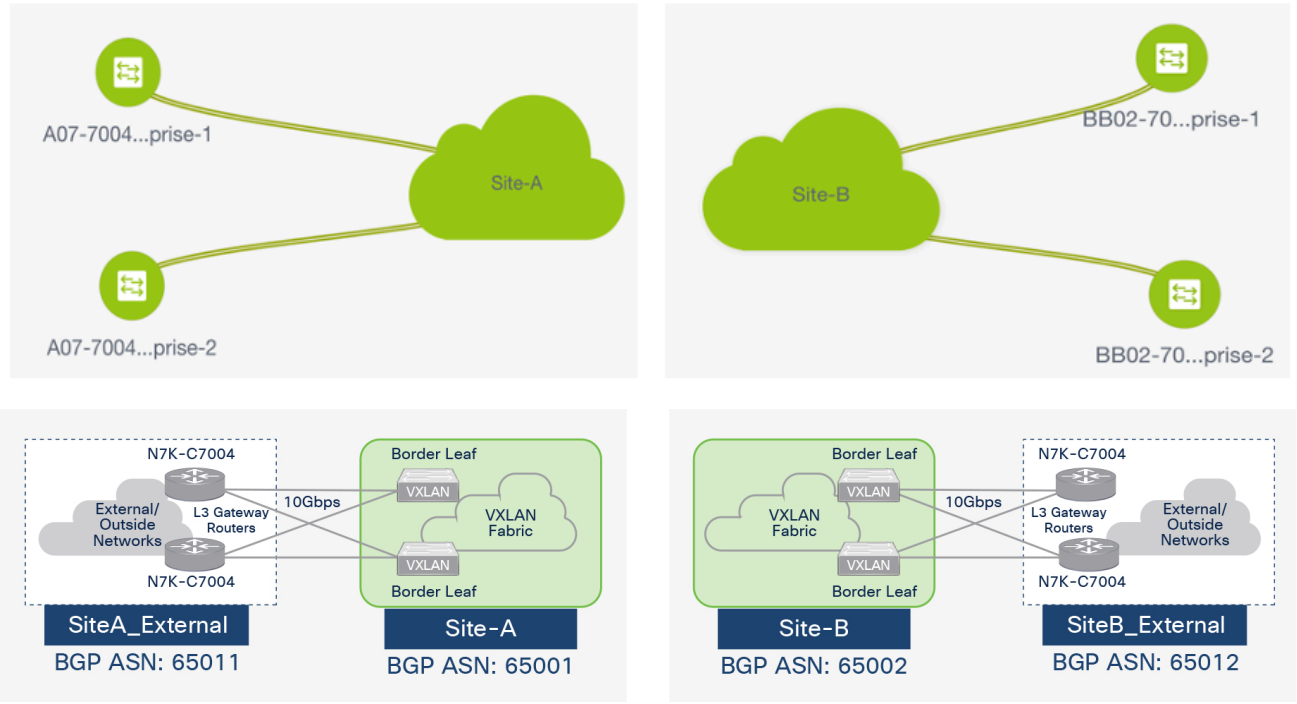


Figure 18.
External Connectivity in Site-A and Site-B - High-Level View

In this design, the external gateways in the external fabric are a pair Cisco Nexus 7000 Series switches connected using redundant 10-GE links to the border leaf switches in Site-A and Site-B fabrics. The external gateways are in different BGP Autonomous Systems. Figure 19 shows the physical connectivity between the fabrics in each site (**Site-A** <-> **SiteA_External**, **Site-B** <-> **SiteB_External**).

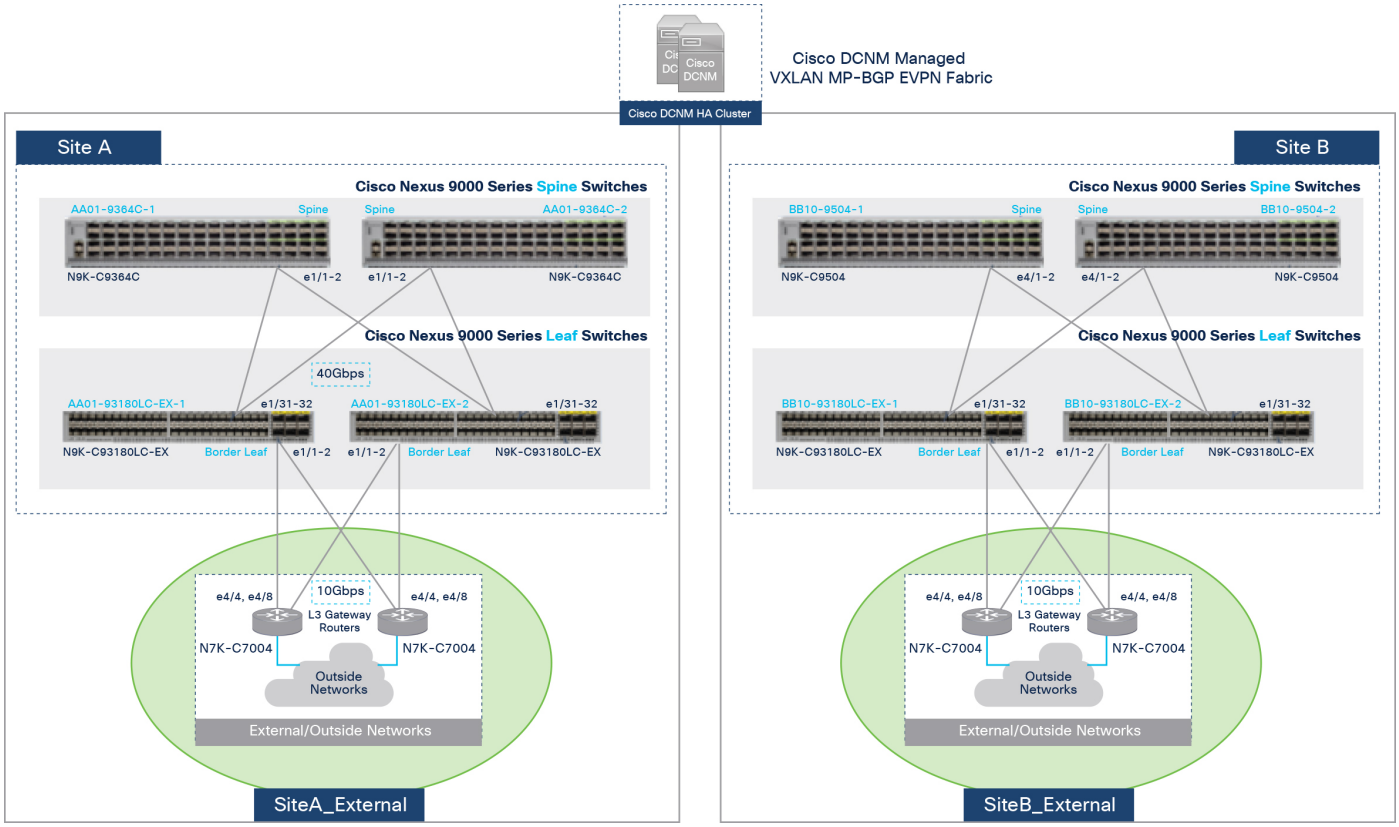
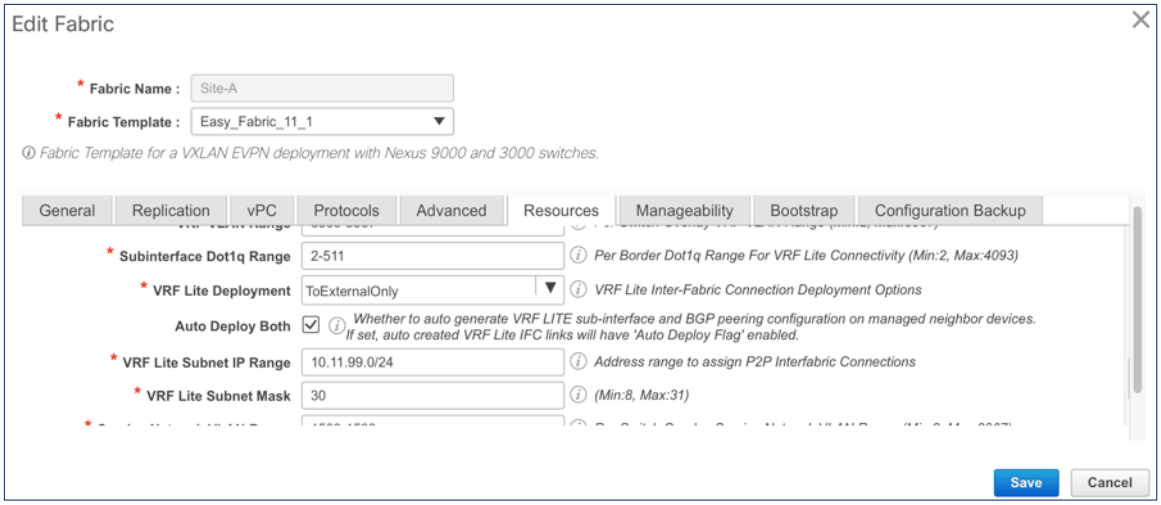
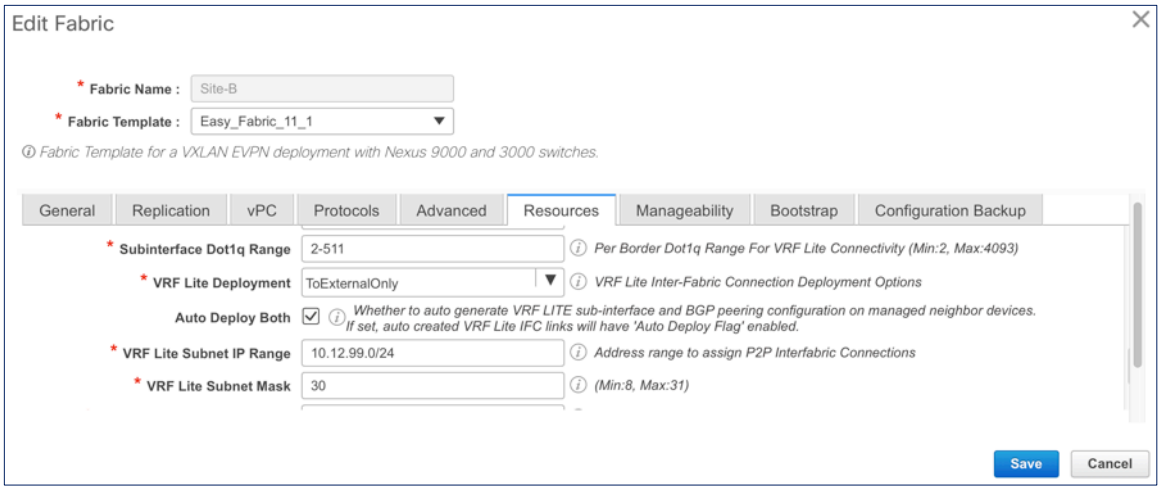


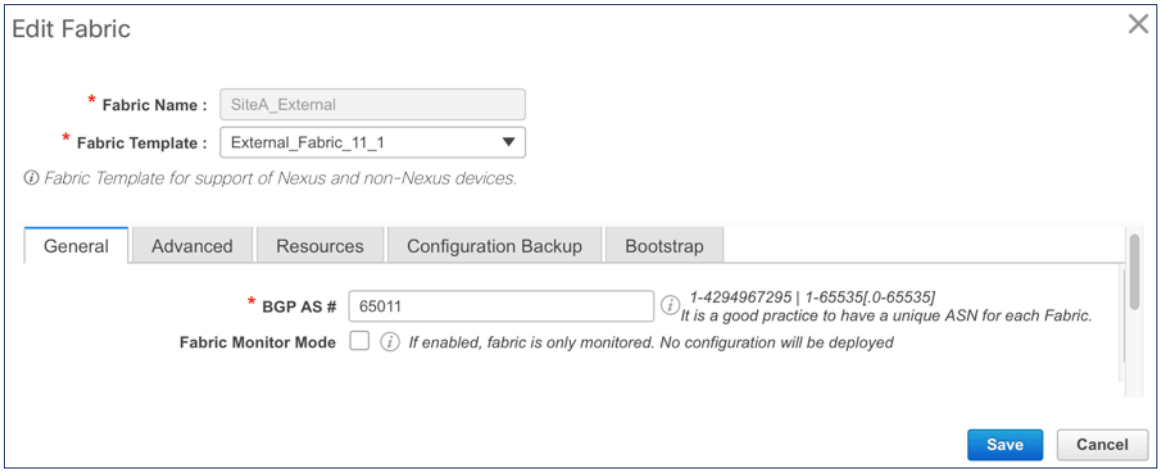
Figure 19.
External Connectivity in Site-A and Site-B – Detailed View

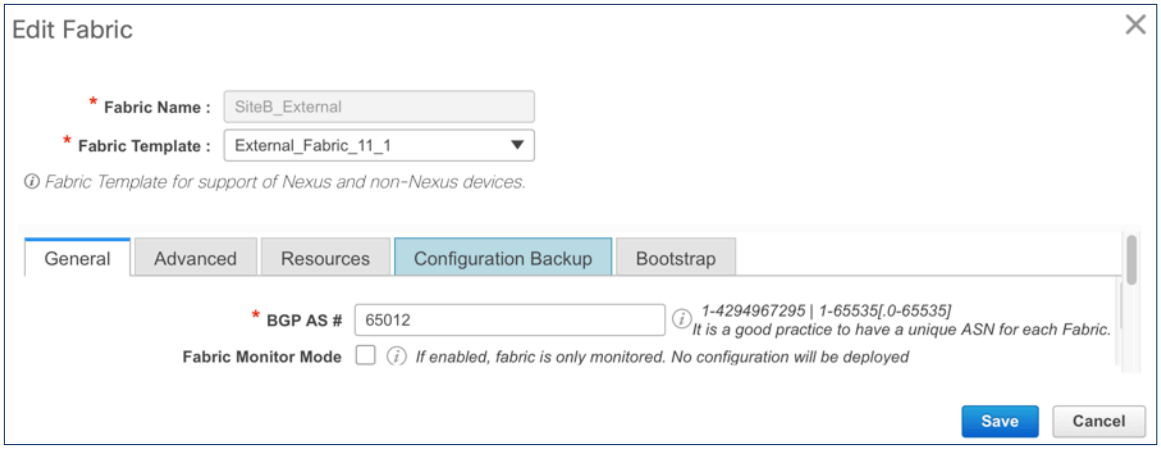
You can deploy the external gateways in a **managed** or **monitored** mode. This solution uses managed mode for the external gateways, which allows Cisco DCNM to deploy the VRF-Lite setup on the external gateways that connect to the border leaf switches in each site. The following screenshots show the Cisco DCNM Fabric Builder configuration parameters for external connectivity from Site-A and Site-B, respectively.





Following are screenshots of the Cisco DCNM Fabric Builder configuration parameters for the external fabrics (SiteA_External, SiteB_External) that Site-A and Site-B, connect to, respectively.

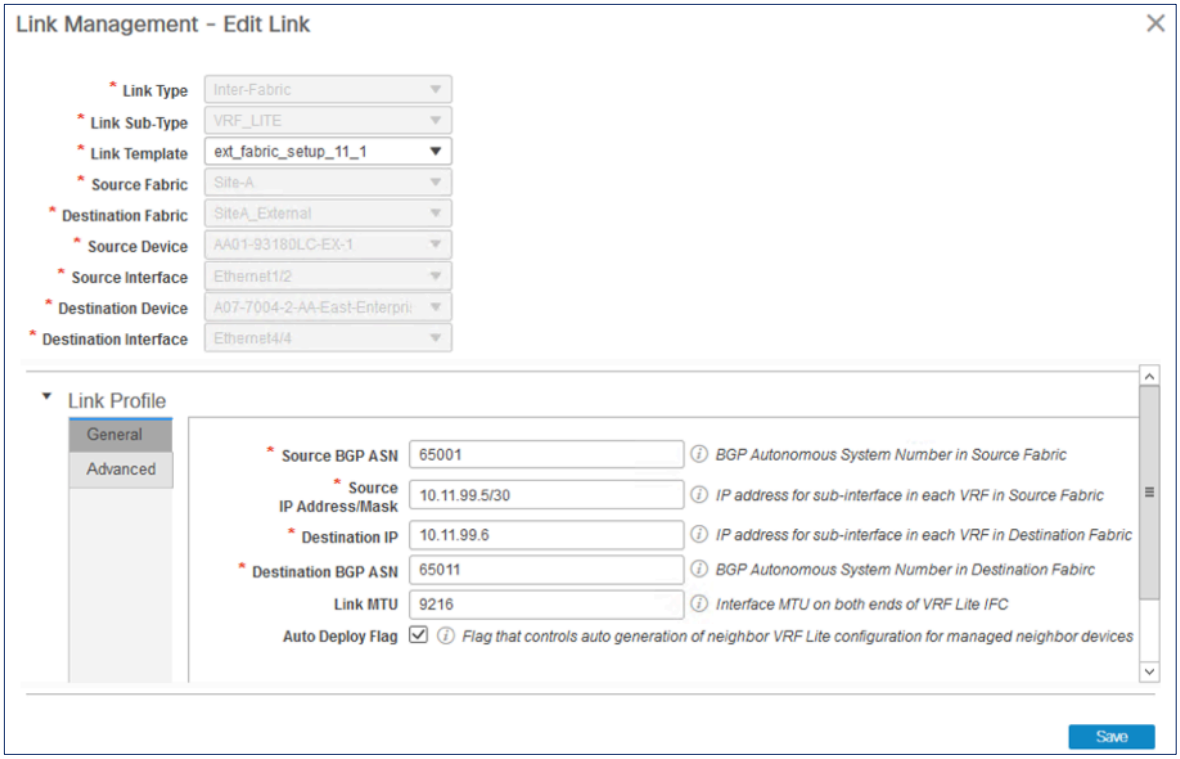




The access-layer connectivity from each site to the external gateways is enabled through Inter-Fabric links configured for IEEE 802.1Q trunking. For redundancy, each border leaf switch connects to both external gateways. The connectivity is enabled from a routed, VLAN tagged, VRF interface on the border leaf switch to a routed, VLAN tagged VRF-Lite interface on the external gateway switch. The connectivity is enabled on a per VRF basis, one for each tenant that requires connectivity to external/outside networks. In this design, external/outside connectivity is enabled in both sites for the **FPV-Foundation_VRF**.

The following screenshot shows the Inter-Fabric connectivity between the switches at Site-A. The Fabric Name shows that the connectivity is between fabrics and the policy is an **ext_fabric_setup_11_1** policy. The setup in Site-B is similar to that of Site-A shown in this screenshot:

Following is a screenshot of the detailed link-level configuration for one link between the VXLAN fabric and external fabric in Site-A. The remaining links in Site-A and Site-B links are set up similarly.



When the external-facing links and the initial setup is complete as outlined previously, the tenants and VRF interfaces for FlexPod Infrastructure connectivity or for applications hosted on the FlexPod infrastructure should be deployed on the border leaf switches to enable external connectivity for those tenants.

Inter-Site Design - Interconnecting Data Centers

The VXLAN EVPN Multi-Site architecture provides seamless Layer 2 and Layer 3 extension between individual VXLAN EVPN fabrics (typically in different sites) that are being interconnected. You can extend VXLAN EVPN fabrics by using other Data Center Interconnect (DCI) technologies; however this design provides a more integrated and scalable design that is also transport-independent. For more details about the Multi-Site approach used in this design, refer to the IETF draft (draft-sharma-multi-site-evpn).

The connectivity between data center sites in a VXLAN EVPN Multi-Site architecture is enabled through the inter-site network. Border Gateways (BGWs) are used to interconnect each site to the inter-site network for east-west traffic flow between data center sites. You can deploy BGWs as standalone leaf switches or combine the function with spine switches already in each site. This design uses standalone leaf switches as BGWs for a more scalable design which is particularly important in large enterprise data centers. You also can use leaf switches for connectivity to networks outside the data centers for north-south traffic, but this design is not the one used in this solution. Outside connectivity is done through separate border leaf switches (refer to the "**Intra-Site Design - Connectivity to Outside/External Networks"** section for more details about this design. You also can deploy BGWs as vPC gateways for connecting to endpoints, typically for network services such as firewalls and load balancers. Alternatively, you can deploy them as anycast BGWs when no endpoints are BGWs; this deployment model is used in this solution. For data-plane scalability, you can deploy up to four BGWs in each site at the time of writing this document. This solution uses two BGWs, but you can add additional BGWs as needed. For BUM traffic forwarding between sites, the BGWs always use ingress replication. However, within a site, they can use either PIM ASM or ingress replication, and you do not need use the same technology for each site being interconnected.

The EVPN Multi-Site architecture uses VXLAN encapsulation for the data plane, requiring 50 or 54 bytes of overhead, so a minimal MTU of 1550 or 1554 is necessary in the inter-site network as well. In this design, a jumbo MTU of 9216 is used across the end-to-end VXLAN fabric. The BGWs provide separation between the internal VXLAN fabric and the external or inter-site VXLAN network by implementing internal and external VTEP functions for connecting to the internal and external networks respectively. To the internal fabric, the BGWs in a site are anycast BGWs (A-BGWs); they provide a common anycast virtual IP (VIP) address that is used for all data-plane communication between sites. A dedicated loopback IP address is allocated for this VIP. The distributed BGWs with anycast VIP enable you to use Equal Cost Multi-Pathing (ECMP) to provide active data forwarding across all BGWs for load distribution and redundancy. For BUM traffic, a BGW is elected as the designated-forwarder for each Layer 2 VNI. The election process distributes the designated-forwarder functionality for the different networks across the different A-BGWs. The A-BGWs will forward BUM traffic for one or more networks typically. Failure detection and the failover of VIP and designated-forwarder function to other BGWs is an important advantage of the VXLAN EVPN Multi-Site architecture. Internal and external interfaces on the BGWs are specially configured to understand their role in the network and tracked to detect failure quickly. Seamless Layer 2 and Layer 3 extension between sites will be available as long as one BGW with one internal- and external-facing interface is available in each site.

The control plane for inter-site connectivity uses Multiprotocol External BGP (MP-eBGP), unlike intra-site connectivity, which can use either eBGP or iBGP. For control-plane scalability, you can deploy route servers in the inter-site network and provide the functions similar to route reflectors in iBGP. Route servers are recommended when three or more sites are being connected. Route servers are not used in this FlexPod solution because it is an active-active, two-data-center solution. However, not using a centralized entity for route peering means that the BGWs in one site will need full-mesh eBGP connectivity to BGWs in the remote site.

Figure 20 shows the inter-site design with back-to-back gateways that is used in this FlexPod solution.

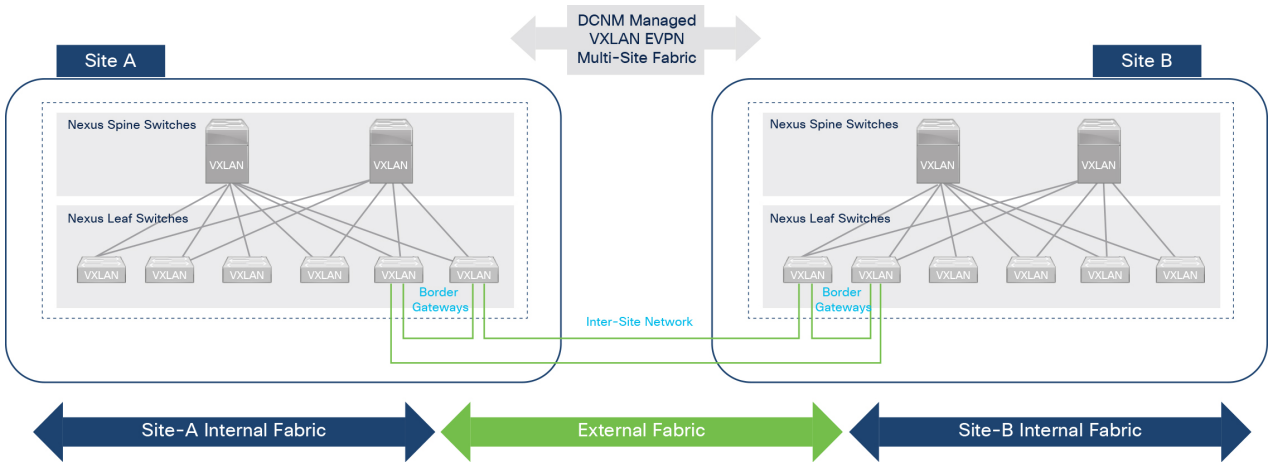


Figure 20.
Inter-Site Design with Back-to-Back Border Gateways

Note that the anycast BGWs in each site have direct IP connectivity between them, resulting in a square topology in the inter-site network; it is required for proper BUM-traffic handling during normal operations and failure scenarios.

In this FlexPod solution, Cisco DCNM deploys and manages the inter-site connectivity. Cisco DCNM Fabric Builder deploys a multi-site domain (MSD) fabric using the **MSD_Fabric_11_1** template, and the existing

fabrics (**Site-A, Site-B, SiteA_External**, **SiteB_External**) are then added and integrated into this new MSD fabric with Cisco DCNM managing the end-to-end network.

Figure 21 shows the physical connectivity between sites in the end-to-end MSD fabric. The border gateways used in this solution are a pair of Cisco Nexus 93240YC-FX2 Switches. The connectivity between the BGWs in each site and across sites are 10-GE, enabling BGWs to establish full-mesh eBGP sessions across all BGWs in the inter-site network. Within a site, BGWs connect to spine switches using 40-GbE links, the same as other leaf switches in each fabric.

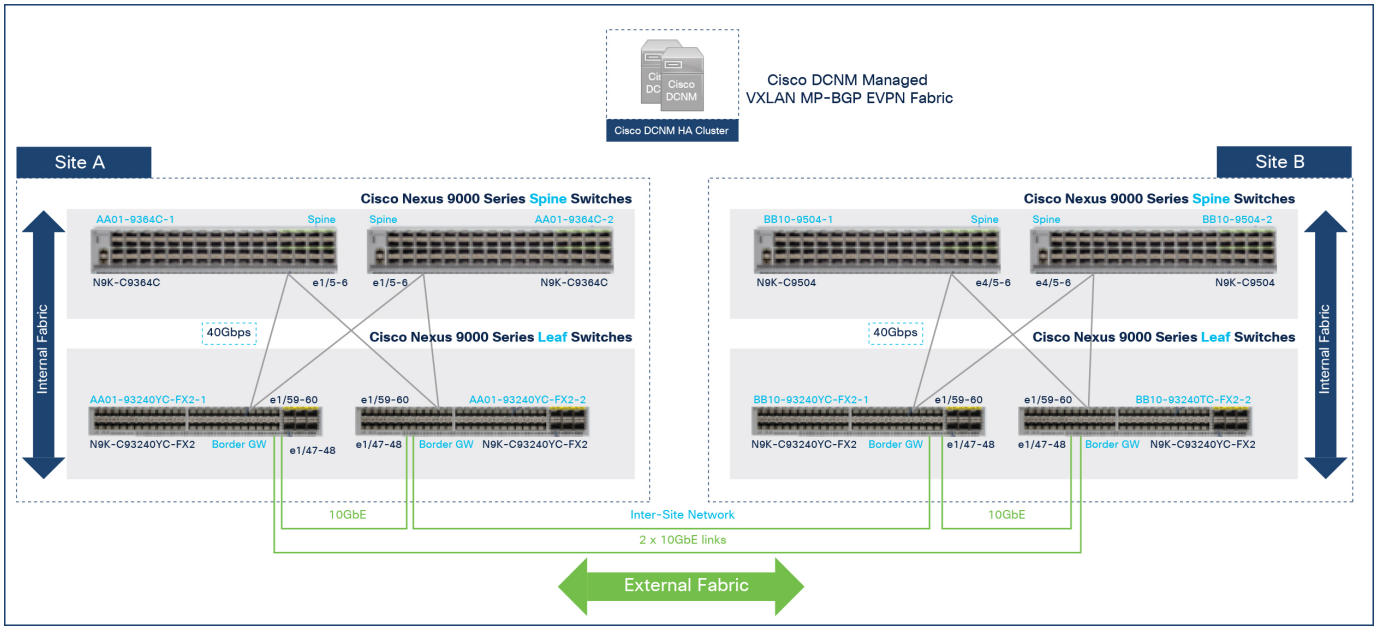
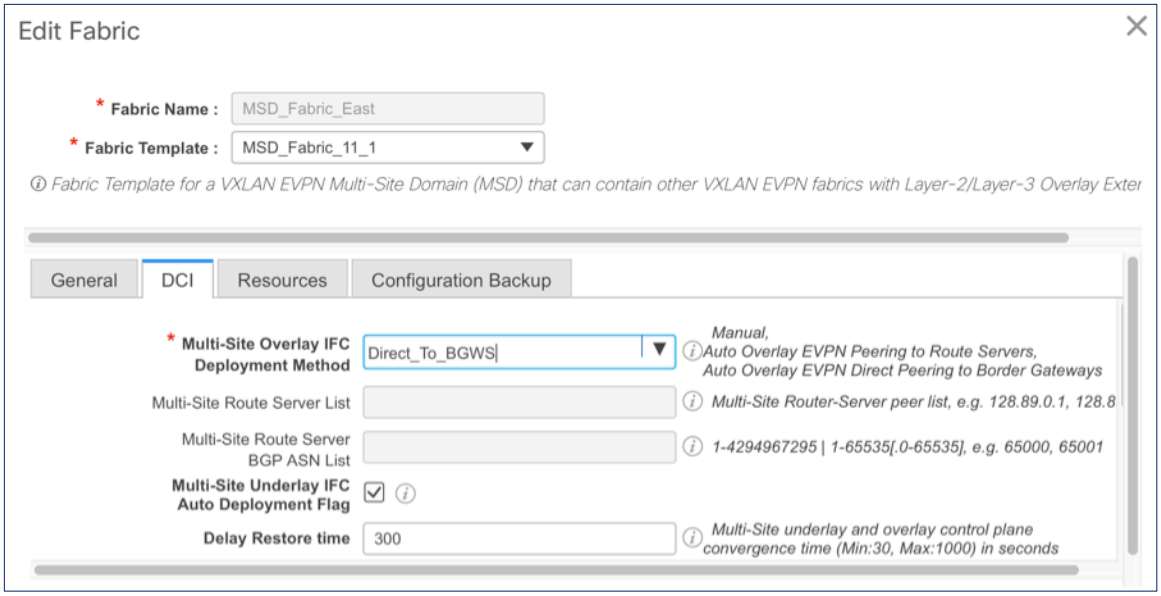
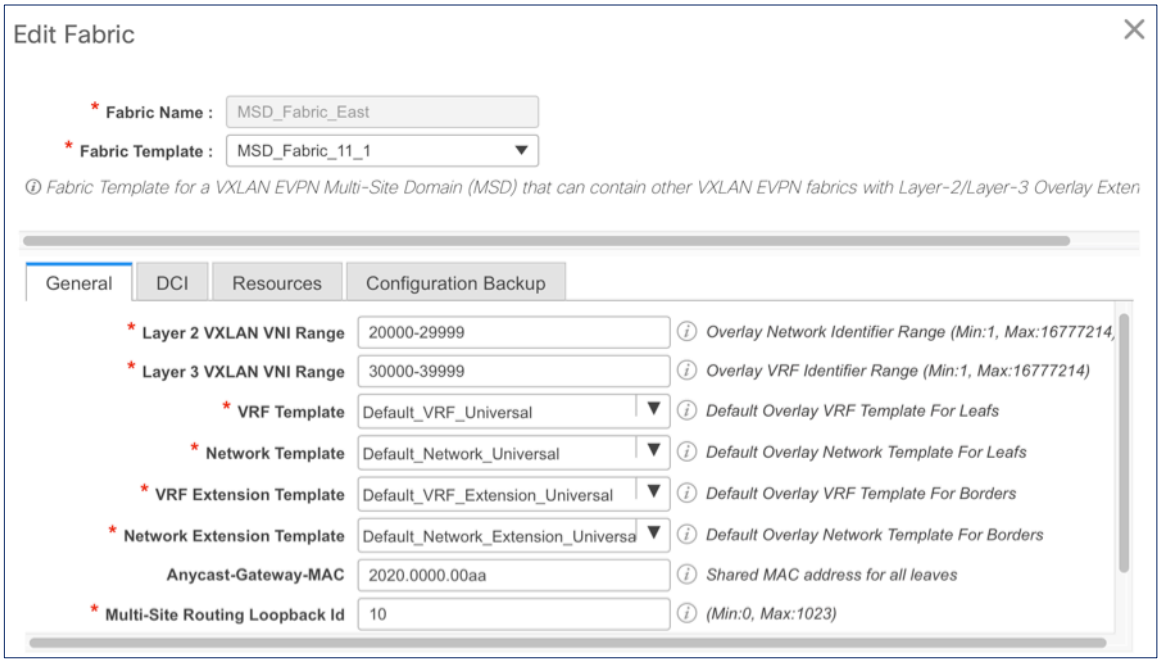


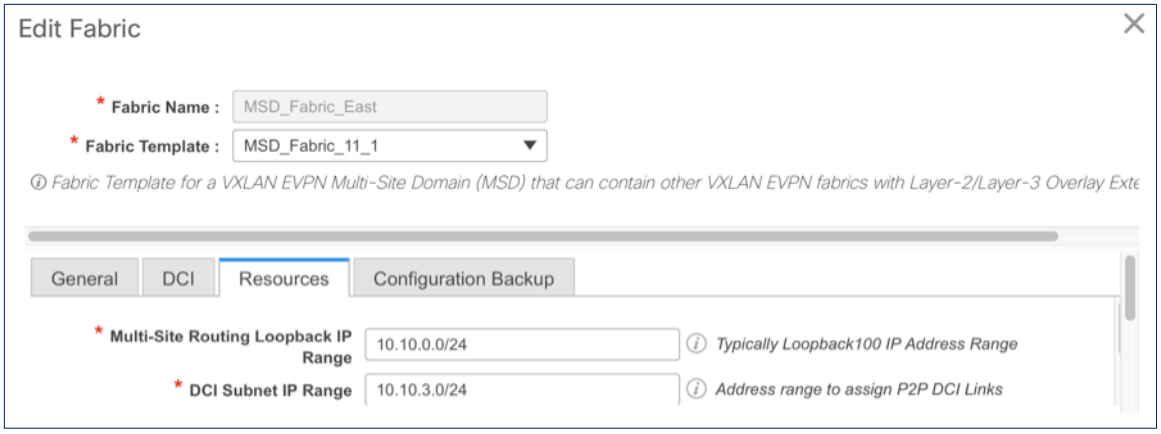
Figure 21.
Inter-Site Design – Physical Connectivity

The BGWs are a transition point from the internal fabric to external fabric, making it a good location for enforcing policies (QoS, security) between data centers. In this design, this transition point is also a bandwidth transition point, so it is important to ensure that critical traffic does not experience congestion as the load increases. Cisco recommends monitoring the inter-site links and traffic flows to have a baseline understanding of the bandwidth and latency when it is first deployed and post-deployment so you can track changes to the baseline performance and take actions before any performance problems occur. This monitoring is particularly important for the high-bandwidth, latency-sensitive storage access flows in this solution that could traverse the inter-site links under certain failure conditions. The actions you could take include adding more links, increasing the available bandwidth, and avoiding congestion altogether, or using QoS to prioritize the more critical traffic if congestion occurs.

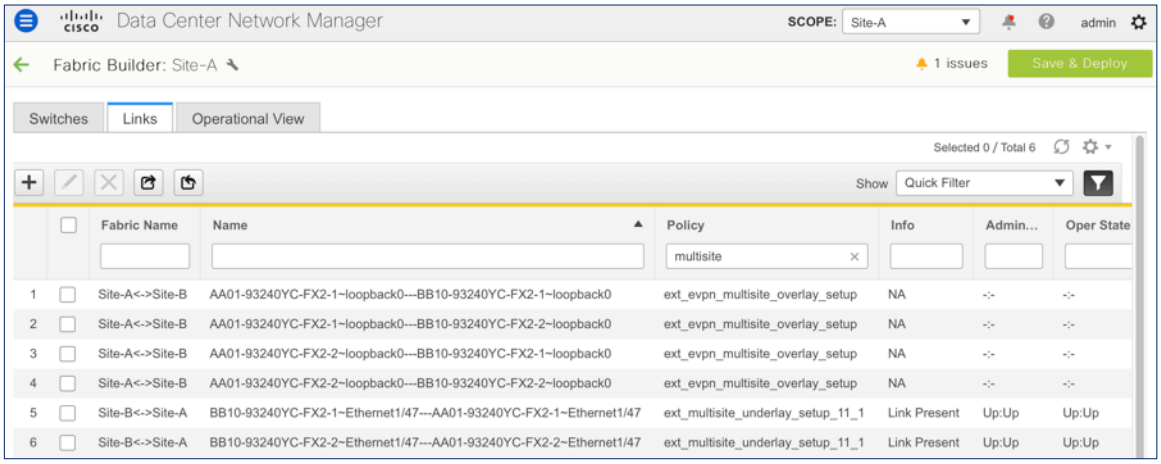
Following are screenshots of three MSD fabrics that Cisco DCNM Fabric Builder creates and deploys for inter-site connectivity. Note that the back-to-back BGW design is specified by selecting **Direct_To_BGWS** from the drop-down list.

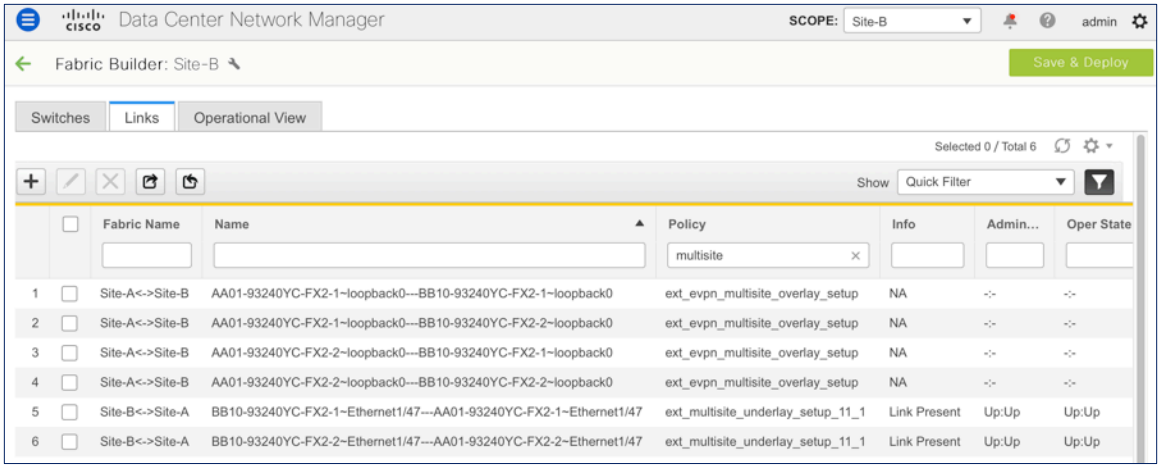




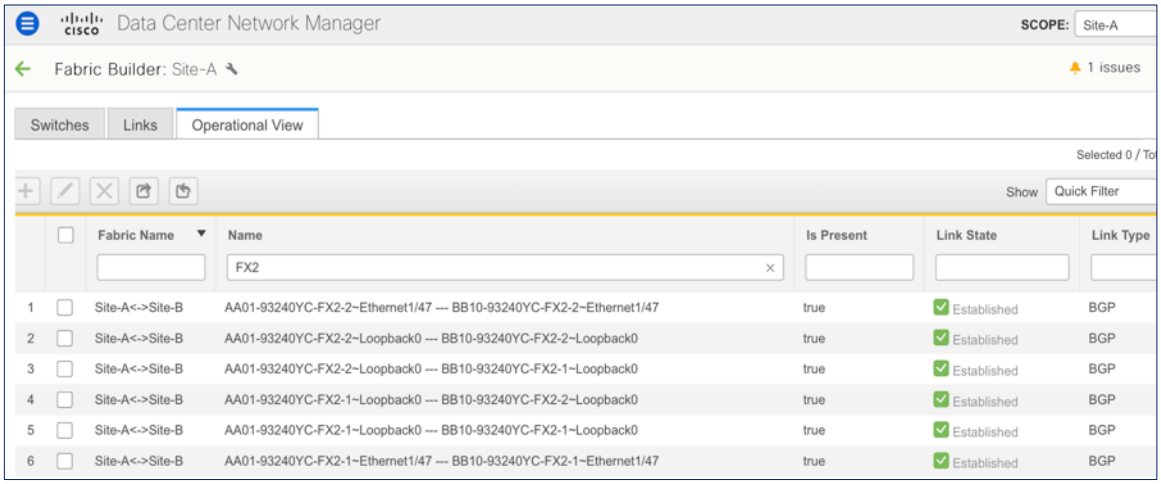


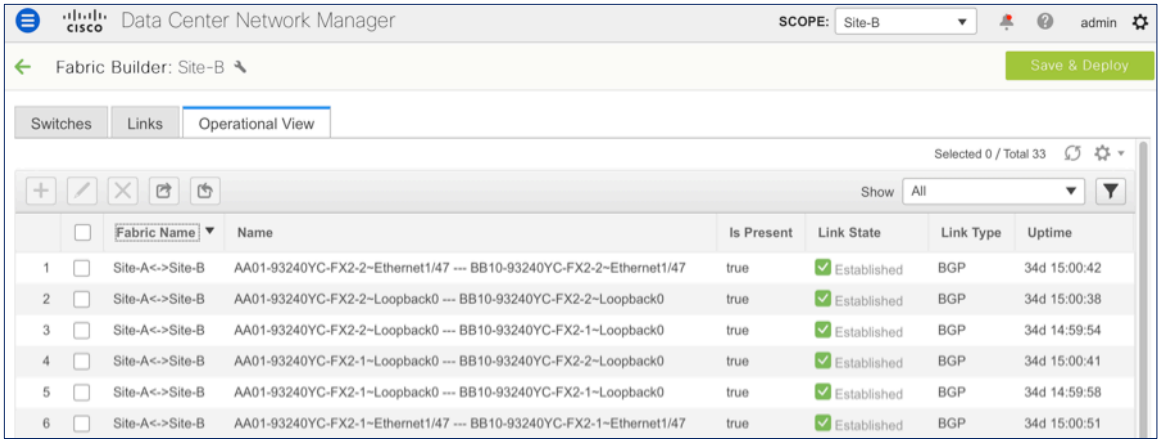
The access-layer connectivity and setup between sites are done by deploying the **ext_multisite_underlay_setup_11_1** and **ext_evpn_multisite_overlay_setup** policies on the physical and loopback interfaces used for inter-site connectivity. Two 10-GE links (from e1/47 of each BGW) provide connectivity between sites and form the underlay IP transport for the inner-site network. The loopbacks are used for the VXLAN overlay setup to establish overlay connectivity between sites. The following screenshots show the connectivity and policies used between sites in this solution.





The following screenshots show the eBGP sessions established between BGWs in the inter-site network.





Tenancy Design

The VXLAN MP-BGP EVPN is designed for multitenancy, which enables enterprises to partition the fabric along organizational or functional lines. You can also base the tenancy design on other factors. The tenancy design in this solution is based on connectivity requirements. Two tenants are used in this FlexPod design: **FPV-Foundation_VRF** and **FPV-Application_VRF**. The foundation tenant is used for all FlexPod compute, storage, and virtual infrastructure connectivity in the FlexPod solution. It includes all connectivity required to stand up the virtual server infrastructure in a given data center (site) and for connectivity between VSI components in each data center. It also includes connectivity for any management or operational tools that use to manage the infrastructure. The application tenant, on the other hand, is for any application workloads hosted on the FlexPod virtual server infrastructure. You can deploy additional tenants as needed to meet the needs of your deployment.

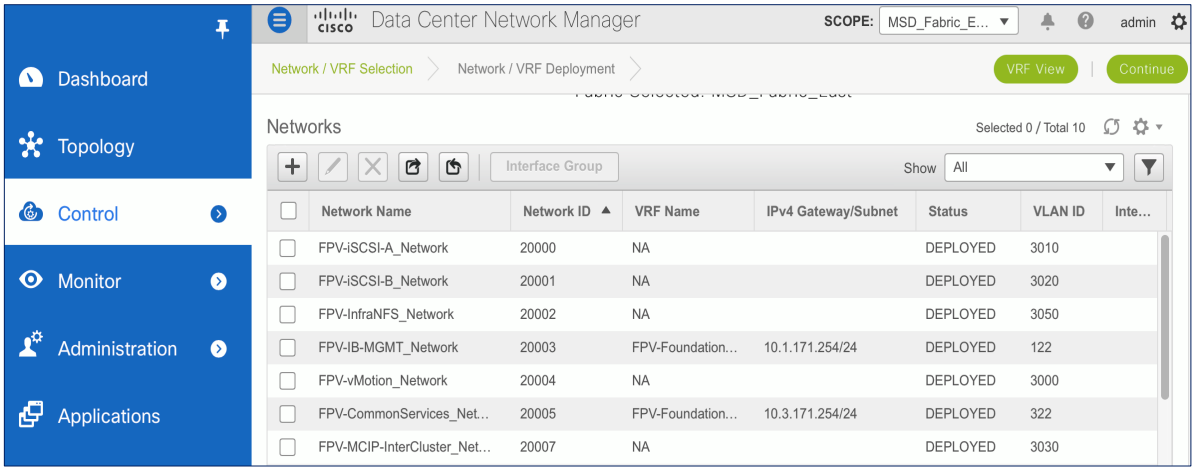
FlexPod Infrastructure Connectivity

The FlexPod infrastructure networks provides connectivity to/from Cisco UCS Compute, NetApp AFF A700 storage, and VMware vCenter and vSphere hosts. As described in the "**Intra-Site Design – Edge Connectivity**" section, vPCs are used for connecting the FlexPod compute and storage infrastructure in the edge or access-layer network to the leaf switches in the VXLAN fabric. To enable connectivity beyond the leaf switches, the VXLAN fabric still needs to enable connectivity for these FlexPod infrastructure networks; these networks are part of the **FPV-Foundation_VRF** tenant.

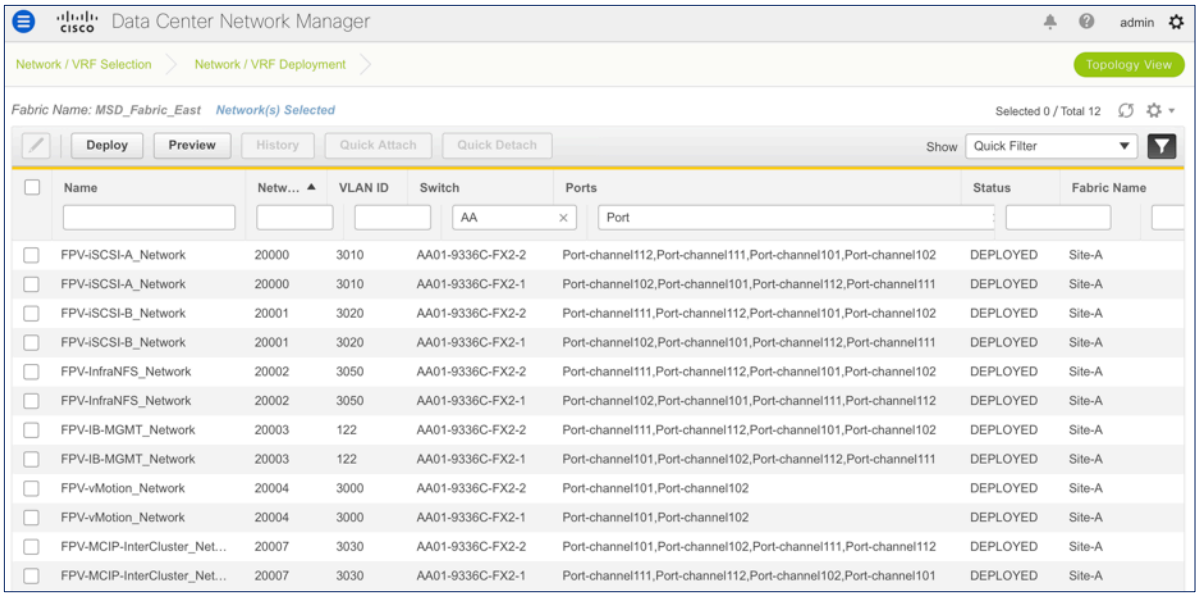
The FlexPod infrastructure connectivity that provides the **FPV-Foundation_VRF** in the VXLAN fabric follow:

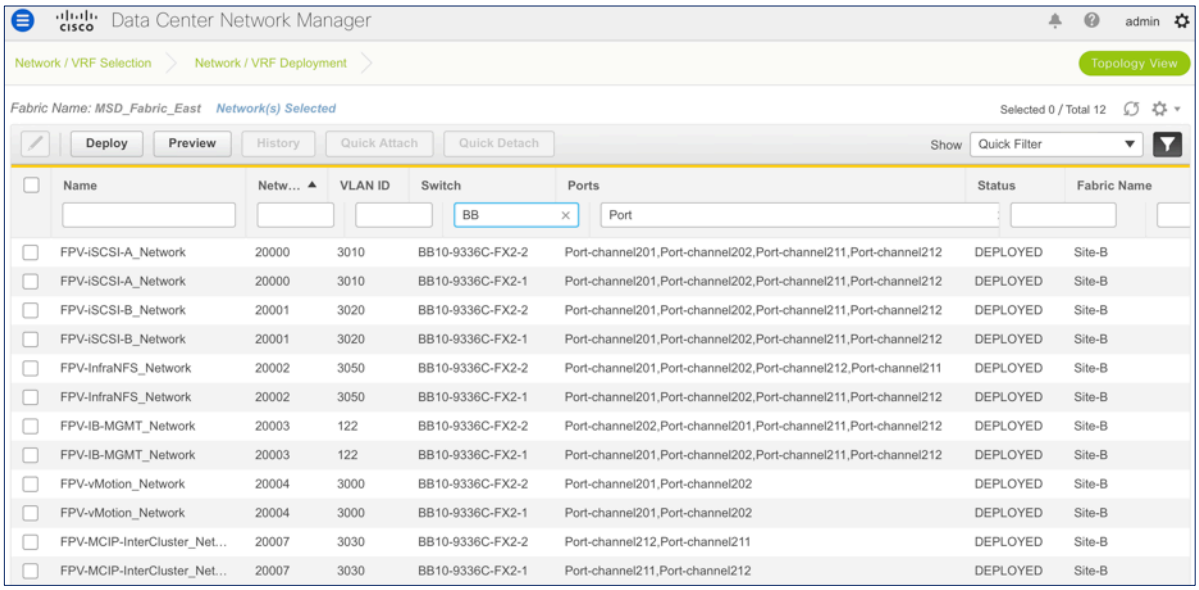
- Connectivity for Internet Small Computer Systems Interface over IP (iSCSI) Boot: This connectivity enables Cisco UCS servers to boot using iSCSI with boot datastores hosted on the NetApp storage cluster. The two iSCSI networks provide redundant iSCSI paths to the NetApp array. The FPV-iSCSI-A_Network and FPV-iSCSI-B_Network in the VXLAN fabric enable iSCSI boot connectivity between the endpoints. These networks are deployed as Layer 2 networks in the VXLAN fabric.
- Connectivity for MetroCluster IP Inter-Cluster network: This connectivity is for Inter-Cluster logical interface (LIF) communication. Every Inter-Cluster LIF on the local cluster must be able to communicate with every Inter-Cluster LIF on the remote cluster. ONTAP cluster peering and the MetroCluster IP configuration replication go through the Inter-Cluster network.
- In-band management: This connectivity is for in-band management communication, which is used primarily by VMware ESXi hosts and vCenter. The in-band management network and the connectivity between these endpoints are enabled by the FPV-InBand-SiteA_Network in the VXLAN fabric. This network is referred to as Site1-IB in the Cisco UCS and VMware portion of the configuration. This network is deployed as a Layer 3 network with the default gateway in the VXLAN fabric.
- Connectivity to Network File System (NFS) datastores: This connectivity is used primarily for accessing NFS datastores hosted on the NetApp storage cluster. The **FPV-InfraNFS_Network** in the VXLAN fabric enables the NFS datastore access. This network is deployed as a Layer 2 network in the VXLAN fabric.
- Connectivity to VMware vMotion network: To support VMware vMotion for the virtual machines hosted on the FlexPod infrastructure, the hosts need connectivity to a VMware vMotion network. The vMotion network and the connectivity between ESXi hosts in the cluster are enabled by the **FPV-vMotion_Network**. It is deployed as a Layer 2 network in the VXLAN fabric.
- Connectivity for infrastructure management and services network (optional): The **FPV-CommonServices_Network** in the VXLAN fabric enables connectivity for infrastructure management, services, and other operational tools used in this FlexPod design. This network is deployed as a Layer 3 network with the default gateway in the VXLAN fabric.

Following is a screenshot of the FlexPod Infrastructure networks used in this solution:



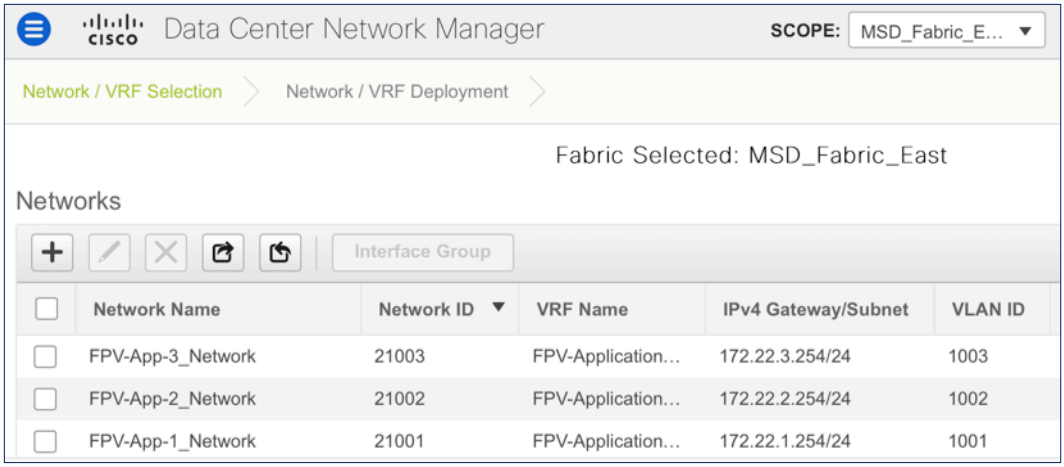
Following are screenshots that show the access-layer connectivity for the previous networks to the FlexPod infrastructure in the access/edge network for Sites A and B, respectively. Note that port-channels [1-2] are part of the vPC to the Cisco UCS domain and port-channels [3-4] are part of the vPC going to NetApp storage.





Connectivity for Applications and Services hosted on FlexPod Infrastructure

The applications hosted on the FlexPod VSI require connectivity through the VXLAN fabric; it is provided by the networks in the **FPV-Application_VRF** tenant. The following screenshot shows the application networks enabled in the VXLAN fabric for validating this design.



These networks are enabled on the vPC to provide access-layer connectivity to Cisco UCS software where the application virtual machines that use these networks are running.

Cisco Data Center Network Manager Design

The Cisco DCNM serves as a controller; it manages the end-to-end VXLAN Multi-Site network in the solution. Cisco DCNM provides a dashboard for a single point of management with automated fabric deployment, automatic consistency-checking, automatic remediation, and device lifecycle management. For monitoring and visibility, Cisco DCNM provides a real-time health summary for fabric, devices, and topology, with correlated visibility for fabric, and triggered alarms. Cisco DCNM also offers numerous workflows for ease of deployment and operations (return materials authorization [RMA], install, upgrade), in addition to customizable Python++ templates for the VXLAN Fabric Builder used to deploy the end-to-end fabric. Cisco DCNM also provides, on a per-switch basis, the configuration deployment history of the nodes, including underlay, overlay, and interface configurations. For automation, Cisco DCNM provides northbound Representational State Transfer application programming interfaces (REST APIs) and serves as a single point of integration for the entire fabric. The DCNM HTML5 GUI uses the same REST APIs for all GUI functions.

In this solution, Cisco DCNM is deployed in a third site, separate from the two data center sites in the solution. The VXLAN fabric in each data center site has out-of-band management connectivity to Cisco DCNM in the third site, helping ensure independent access to the fabric from DCNM at all times. Cisco DCNM is not necessary for data plane forwarding; it is strictly for fabric provisioning, configuration changes, day 2 operations, and management. Figure 22 shows the Cisco DCNM connectivity to the VXLAN fabrics in each data center

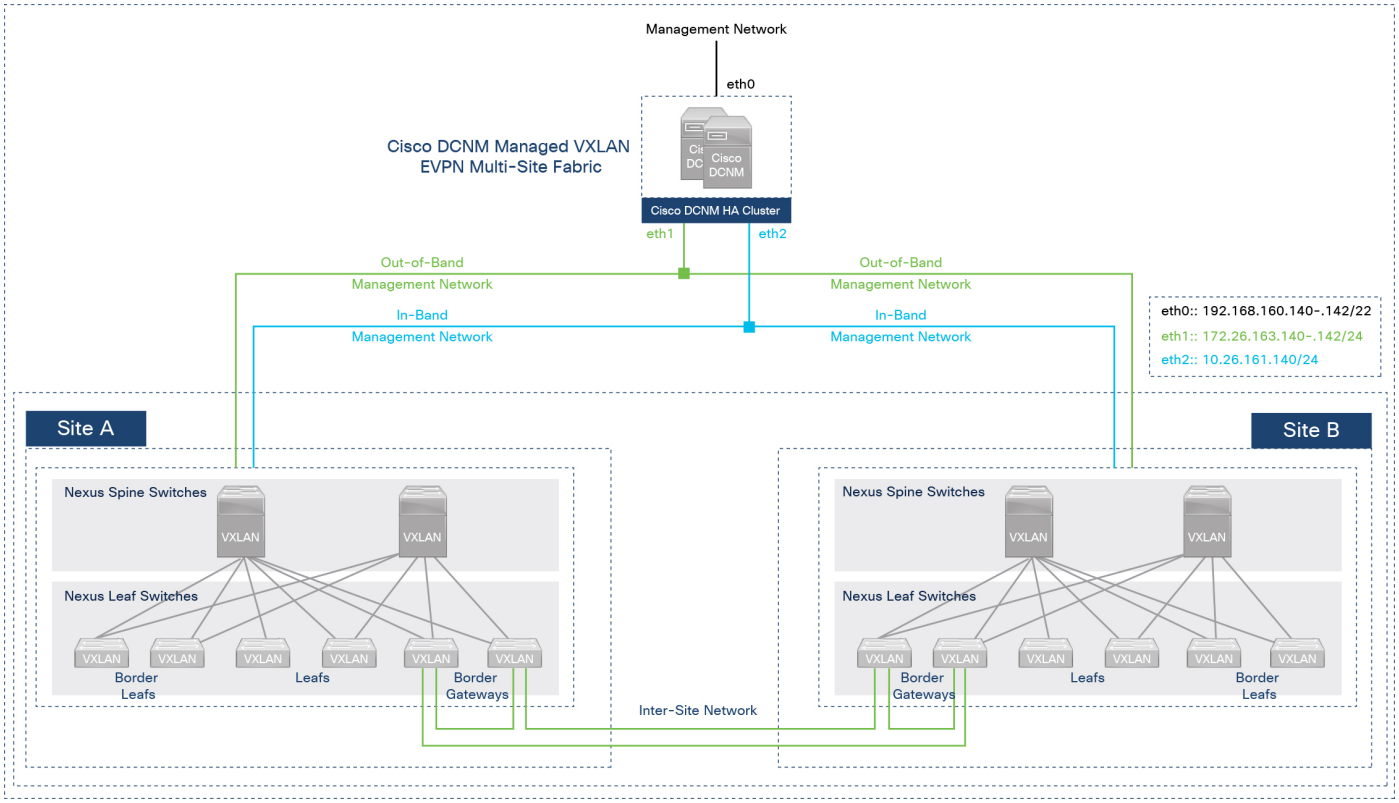


Figure 22.
Connectivity to Cisco DCNM

Cisco DCNM is deployed as a cluster of two virtual machines in native high-availability (HA) mode. With native HA, the two Cisco DCNMs run as active and standby applications. The embedded databases of both active and standby appliances are synchronized in real time. You can add additional compute nodes or worker nodes for scalability; three additional worker virtual machines are used in this solution. The DCNM and compute servers are deployed on three physical servers. The eth0, eth1, and eth2 interfaces of the Cisco DCNM and compute nodes in a clustered mode must be in the same network or Layer 2 adjacent.

The Cisco DCNM GUI is accessed through the management network. Cisco DCNM connects to the VXLAN fabric in each site through the out-of-band (OOB) management network, which is used for management and provisioning, and the in-band (IB) management network for the Endpoint Locator and telemetry features that are typically bandwidth-intensive.

High Availability

High availability is a critical consideration for any data center infrastructure design, and more so for a disaster-recovery solution such as this one. The active-active VSI in each data center delivers continuous access to mission-critical workloads, with each site providing backup and seamless failover for instances of failure. You can deploy applications and services in either data center location using local resources (compute, storage) or remote resources, depending on the type of failure. To achieve availability at the data center level, the sub-systems that make up data center infrastructure (compute, storage, network, virtualization) must provide complementary capabilities in each active-active data center, with the ability to fail over to the second data center if a failure occurs in the first one. High availability is also important within a data center to handle smaller failures with minimal impact.

For the network sub-system, the VXLAN EVPN Multi-Site architecture used in this solution provides the network fabrics in each data center location and the connectivity between them, as well as the ability to fail over by extending connectivity and services across data centers. The solution also provides high availability for the network fabric in each data center and in the inter-site network between them, with no single points of failure. The end-to-end network is resilient at the physical link and node level as well as across higher layers of the infrastructure stack. The high availability features the network provides include:

- VXLAN Multi-Site architecture: The architecture fundamentally provides high availability by enabling interconnection of independent fabrics. This feature allows deployment and interconnection of a second fabric and data center to the first data center, thereby enabling the active-active design used in this solution. The architecture also provides fault containment and isolation between sites because each site is a separate failure domain, helping ensure that a failure in one active site does not affect the other.
- Intra-site connectivity: The connectivity within a site is the same for both active-active data centers. Endpoints connect to top-of-rack leaf switches, and each leaf switch connects to all the spine switches in that data center site. This setup provides redundancy while also enabling multiple IP Equal-Cost Multipath (ECMP) routes between leaf switches for VTEP-to-VTEP connectivity. The VXLAN fabric is deployed using two spine switches that serve as redundant BGP route reflectors for the fabric. The routing protocols deployed in the fabric uses the physical-layer connectivity to provide multiple ECMP paths between VTEPs for redundancy and load distribution.
- Inter-Site connectivity: Two border gateways in each data center site connect to BGWs in the remote data center, providing two redundant paths between sites. The BGWs establish eBGP sessions for inter-site connectivity.
- Access-layer connectivity: Two leaf switches are used in this design to connect to FlexPod infrastructure. vPCs are used to connect to Cisco UCS Fabric Interconnects to NetApp storage in each site, providing node and link-level redundancy in the access layer.
- Connectivity to outside networks and services: To enable each site to operate as an independent data center, the design uses separate connections from each site for reachability to outside networks, helping ensure access to critical services directly from each data center.
- Cisco DCNM clustering: To provide resiliency and scalability, a Cisco DCNM cluster consisting of multiple nodes is used to manage the end-to-end VXLAN EVPN Multi-Site fabric. The cluster is located outside the fabric, with reachability to both sites. However, both sites have independent connectivity such that a failure in one site will not affect the ability of Cisco DCNM to communicate and manage the other.

Data-Replication Network – NetApp MetroCluster IP Storage

Figure 23 shows the switch connectivity for the MetroCluster IP storage fabric and intra-cluster fabric within each site and between the two sites. For the A700 storage controllers, each A700 node is connected to both Cisco Nexus 3132Q-V 40G switches at its local site. The connections are used for intra-cluster and MetroCluster IP traffic from the intra-cluster and MetroCluster IP node ports. The intra-cluster ISL between switches at each site is required for the ONTAP cluster communications. The MetroCluster IP ISLs between sites carry traffic for MetroCluster IP storage data and nonvolatile RAM (NVRAM) replication data between sites. The redundant switches and redundant connections configuration provide a highly available MetroCluster IP solution to ensure business continuity because it avoids single-point-of-failure scenarios and a single site failure. The details of the connectivity of the MetroCluster IP switches at each site are also shown in Figure 23 and Tables 3 and 4.

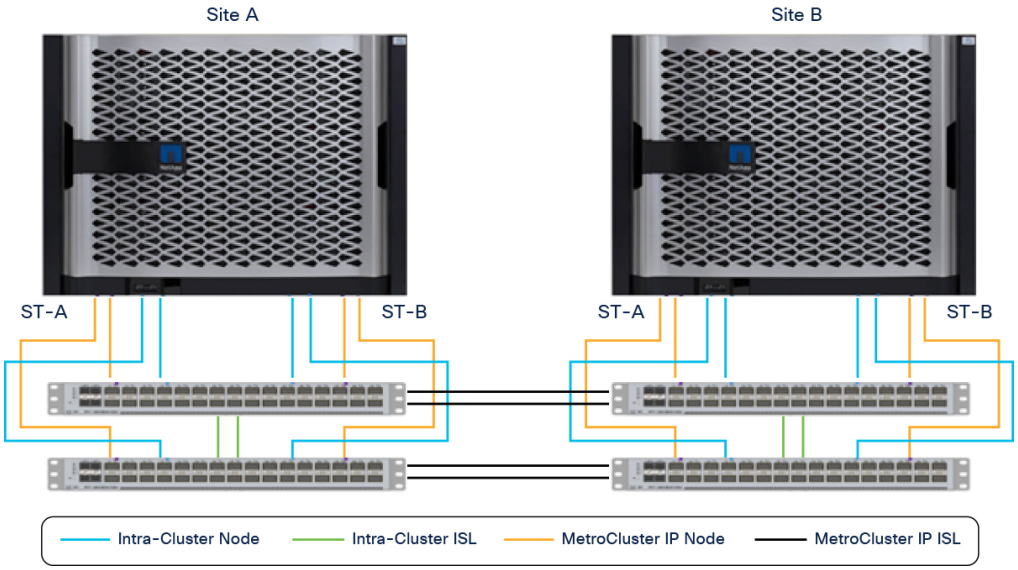


Figure 23.
NetApp MetroCluster IP Storage Network

Table 3. Switch Port Usage for Intra-Cluster and MetroCluster IP Connectivity for Site A

Switch Device	Switch Port	Description	Connected Device / Port
AA01-C3132QV-1	Eth1/1	Intra-Cluster Node Port	mcip-a-01 / e4a
AA01-C3132QV-1	Eth1/2	Intra-Cluster Node Port	mcip-a-02 / e4a
AA01-C3132QV-1	Eth1/7	Intra-Cluster ISL Port	AA01-C3132Q-2 / Eth1/7
AA01-C3132QV-1	Eth1/8	Intra-Cluster ISL Port	AA01-C3132Q-2 / Eth1/8
AA01-C3132QV-1	Eth1/9	MetroCluster IP Node Port	mcip-a-01 / e5a
AA01-C3132QV-1	Eth1/10	MetroCluster IP Node Port	mcip-a-02 / e5a
AA01-C3132QV-1	Eth1/15	MetroCluster IP ISL Port	AA13-C3132Q-1 / Eth1/15
AA01-C3132QV-1	Eth1/16	MetroCluster IP ISL Port	AA13-C3132Q-1 / Eth1/16
AA01-C3132QV-2	Eth1/1	Intra-Cluster Node Port	mcip-a-01 / e4e
AA01-C3132QV-2	Eth1/2	Intra-Cluster Node Port	mcip-a-02 / e4e
AA01-C3132QV-2	Eth1/7	Intra-Cluster ISL Port	AA01-C3132Q-1 / Eth1/7
AA01-C3132QV-2	Eth1/8	Intra-Cluster ISL Port	AA01-C3132Q-1 / Eth1/8
AA01-C3132QV-2	Eth1/9	MetroCluster IP Node Port	mcip-a-01 / e5b
AA01-C3132QV-2	Eth1/10	MetroCluster IP Node Port	mcip-a-02 / e5b
AA01-C3132QV-2	Eth1/15	MetroCluster IP ISL Port	AA13-C3132Q-2 / Eth1/15
AA01-C3132QV-2	Eth1/16	MetroCluster IP ISL Port	AA13-C3132Q-2 / Eth1/16

Table 4. Switch Port Usage for Intra-Cluster and MetroCluster IP Connectivity for Site B

Switch Device	Switch Port	Description	Connected Device / Port
AA13-C3132QV-1	Eth1/1	Intra-Cluster Node Port	mcip-b-01 / e4a
AA13-C3132QV-1	Eth1/2	Intra-Cluster Node Port	mcip-b-02 / e4a
AA13-C3132QV-1	Eth1/7	Intra-Cluster ISL Port	AA13-C3132Q-2 / Eth1/7
AA13-C3132QV-1	Eth1/8	Intra-Cluster ISL Port	AA13-C3132Q-2 / Eth1/8
AA13-C3132QV-1	Eth1/9	MetroCluster IP Node Port	mcip-b-01 / e5a
AA13-C3132QV-1	Eth1/10	MetroCluster IP Node Port	mcip-b-02 / e5a
AA13-C3132QV-1	Eth1/15	MetroCluster IP ISL Port	AA01-C3132Q-1 / Eth1/15
AA13-C3132QV-1	Eth1/16	MetroCluster IP ISL Port	AA01-C3132Q-1 / Eth1/16
AA13-C3132QV-2	Eth1/1	Intra-Cluster Node Port	mcip-b-01 / e4e
AA13-C3132QV-2	Eth1/2	Intra-Cluster Node Port	mcip-b-02 / e4e
AA13-C3132QV-2	Eth1/7	Intra-Cluster ISL Port	AA13-C3132Q-1 / Eth1/7
AA13-C3132QV-2	Eth1/8	Intra-Cluster ISL Port	AA13-C3132Q-1 / Eth1/8
AA13-C3132QV-2	Eth1/9	MetroCluster IP Node Port	mcip-b-01 / e5b
AA13-C3132QV-2	Eth1/10	MetroCluster IP Node Port	mcip-b-02 / e5b
AA13-C3132QV-2	Eth1/15	MetroCluster IP ISL Port	AA01-C3132Q-2 / Eth1/15
AA13-C3132QV-2	Eth1/16	MetroCluster IP ISL Port	AA01-C3132Q-2 / Eth1/16

For deployment simplification, NetApp provides reference configuration files (RCF) and an RCF file generator for the supported switches along with the procedures to deploy the RCF files on each of the four MetroCluster IP switches. Refer to MetroCluster IP Reference Configuration Files for more information.

For additional supported MetroCluster IP switches, details of the ports used for the various connectivity to other storage controllers, and best-practice configuration implementations, please refer to NetApp Hardware Universe, MetroCluster IP Solution Architecture and Design (TR-4689), and Install a MetroCluster configuration: ONTAP MetroCluster.

Storage

FlexPod converged infrastructures provide the data management and protection features and capabilities from NetApp AFF/FAS/ASA storage arrays with Cisco UCS and Cisco Nexus network infrastructures to deliver highly available, highly scalable, and highly flexible solutions that you can easily deploy. With verified architectures that use components that are supported on the interoperability matrices and compatibility listings of NetApp, Cisco, and VMware, the thoroughly tested FlexPod solutions minimize deployment risks and accelerate time to value for solution deployments.

Data fabric powered by NetApp storage systems facilitates the mobility of data across the hybrid cloud ecosystem and enables businesses to seamlessly move data from where it was generated to where it is needed. You can choose the optimal location and platforms for your data and take advantage of the NetApp ONTAP data management features and capabilities both on premises and in the cloud. You can deploy storage solutions with a high-performance all-flash storage, or a hybrid storage system with both SSDs and hard disks to meet your performance, capacity, and cost objectives using the NetApp AFF, FAS, and ASA storage systems. Refer to NetApp Hardware Universe for information about the ONTAP AFF/FAS/ASA storage systems that support MetroCluster IP configuration and the details about their MetroCluster IP interfaces and supported cables.

ONTAP 9.8

ONTAP is the industry-leading flagship data management software from NetApp that enables you to seamlessly manage and protect your data wherever it lives, whether on-premises, at the edge, or in the cloud. Following are some highlights of the new features and enhancements released in ONTAP 9.8 that are relevant to FlexPod:

- Data fabric enhancements: S3 protocol integration in ONTAP allows you to use the FAS system as a FabricPool target and SnapMirror backup to the S3 cloud target.
- SAN enhancements: A new SnapMirror Business Continuity (SMBC) solution is available in public preview, as well as persistent ports enabled by default on all All SAN Arrays, Nonvolatile Memory Express (NVMe) protocol co-existence, and increased size limitations.
- File access protocol enhancements: This release offers ONTAP auditing schema reference, support for SHA-2 Lightweight Directory Access Protocol (LDAP) password hashes, encryption support between ONTAP and domain controllers, NFSv4.2 support, and improved NFSv4.1 performance with nconnect multiple TCP connections (up to 16) support for a single NFS mount.
- FlexGroup support: ONTAP 9.8 offers FlexGroup support for VMware datastore, Virtual Storage Console, and SnapCenter backup.
- Data protection enhancements: NetApp Volume Encryption (NVE) support for root volume encryption to protect root volume, new unspecified retention period for the volume, SnapMirror Cloud using qualified vendor applications, increased concurrent SnapMirror transfer limits, and volume move support for SnapLock.
- NetApp MetroCluster updates: Updates include AFF A250 support for MetroCluster IP, Cisco 9336C switch support, simplified interface for MetroCluster management, MetroCluster IP switchover and switchback support with System Manager, and unmirrored aggregate support.
- Hardware support updates: Updates include expanded platform support for NS224 drive shelves; support for AFF A250, FAS 500f, and ASA AFF A800 platforms; and support for the Nexus 3232C 100GbE Switch (X190100 and X190100R) as a storage switch to connect NS224 NVMe drive shelves to additional platforms.

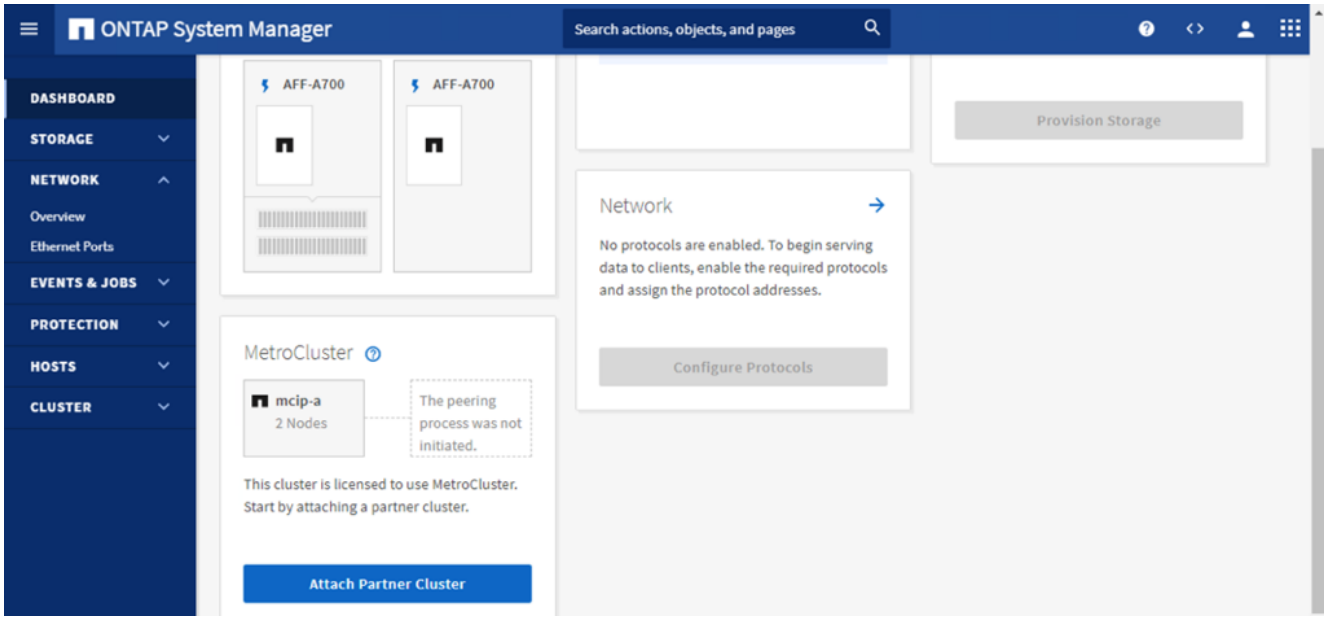
For more information about enhancements, new features, and supported platforms, refer to the ONTAP 9 Release Notes and NetApp Hardware Universe for details.

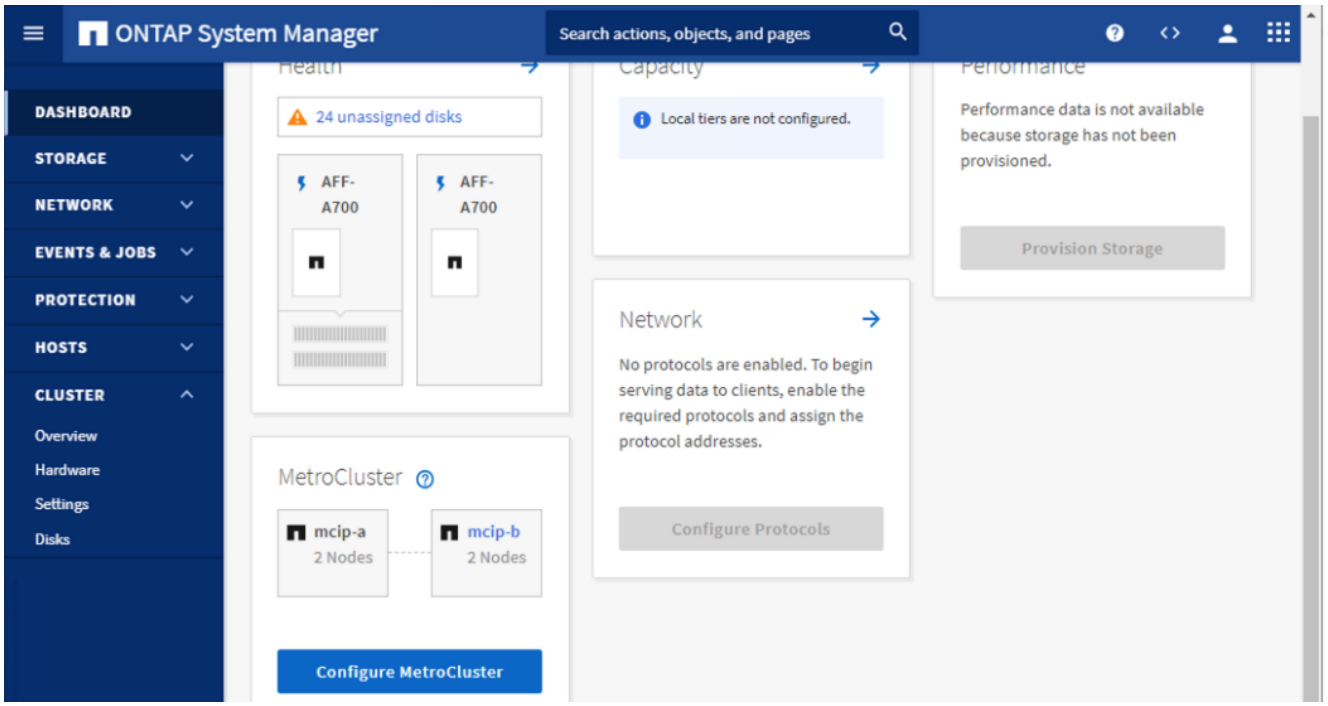
MetroCluster IP Configuration

When the two physically and geographically separated ONTAP clusters are configured to work together as an ONTAP MetroCluster IP solution, the data written to either of the ONTAP clusters is automatically synchronously mirrored to the other site. In addition to synchronously replicating the client write data, the configuration details of the ONTAP MetroCluster IP solution itself are also replicated to the other site so that the two sites stay in sync from the perspectives of both stored data and cluster configuration. The data and configuration replication happens “under the hood” of the ONTAP software and therefore is transparent to the clients performing the I/O after MetroCluster is configured. This set-it-once-and-forget-it configuration approach greatly simplifies the operations of an ONTAP MetroCluster IP solution.

To configure the MetroCluster IP solution for two ONTAP clusters, the underlying network connectivity between the two sites must be in place before peering the ONTAP clusters and configuring MetroCluster IP. The configuration of MetroCluster IP solution involves several steps using the cluster shell interface. Starting with ONTAP 9.8, you can also use System Manager as a simplified interface for the configuration

and management of a MetroCluster solution. Following are screenshots of the System Manager dashboards where workflows to attach partner cluster and MetroCluster configuration are initiated. The first one shows attachment of a partner cluster from the System Manager Dashboard MetroCluster pane, and the second one shows configuration of a MetroCluster in the System Manager Dashboard MetroCluster pane.





The System Manager guided configuration steps greatly simplify the MetroCluster IP deployment. For details about how to set up a MetroCluster site, set up MetroCluster peering, configure a MetroCluster site, and perform MetroCluster switchover and switchback with System Manager, please visit ONTAP System Manager documentation on Manage MetroCluster Sites. Additional information related to the installation of the ONTAP MetroCluster IP solution is available in the Install a MetroCluster IP Configuration: ONTAP MetroCluster document.

Mirrored and Unmirrored Aggregate

To accommodate and mirror the data coming from the other site, the storage in the MetroCluster IP solution is partitioned to store data both locally and remotely with a mirrored aggregate configuration consisting of two matching disk pools, pool 0 for local storage and pool 1 for remote storage. Because client data is synchronized at both sites, data can be served from the remote site when a situation requires it, such as in a simulated testing scenario or when a real disaster occurs.

With ONTAP 9.8, you can optionally create unmirrored data aggregates for data that does not require the redundant mirroring that the MetroCluster IP configuration provides. The unmirrored aggregates must be local to the cluster owning them. For volumes and Logical Unit Numbers (LUNs) created on the unmirrored aggregate, only local clients that do not have the requirement to survive a site failure should use them.

In-Band Management, Inter-Cluster, and Data Storage Protocol Connectivity

The in-band management, inter-cluster, and IP-based data storage protocol connectivity are provided by interconnecting the storage controllers at each site to the VXLAN leaf switches in a redundant configuration. Each storage controller is connected to two leaf switches, as shown in Figure 24. The two connections are configured as part of an interface group for increased resiliency. On the Nexus switch side, those ports are configured as vPC.

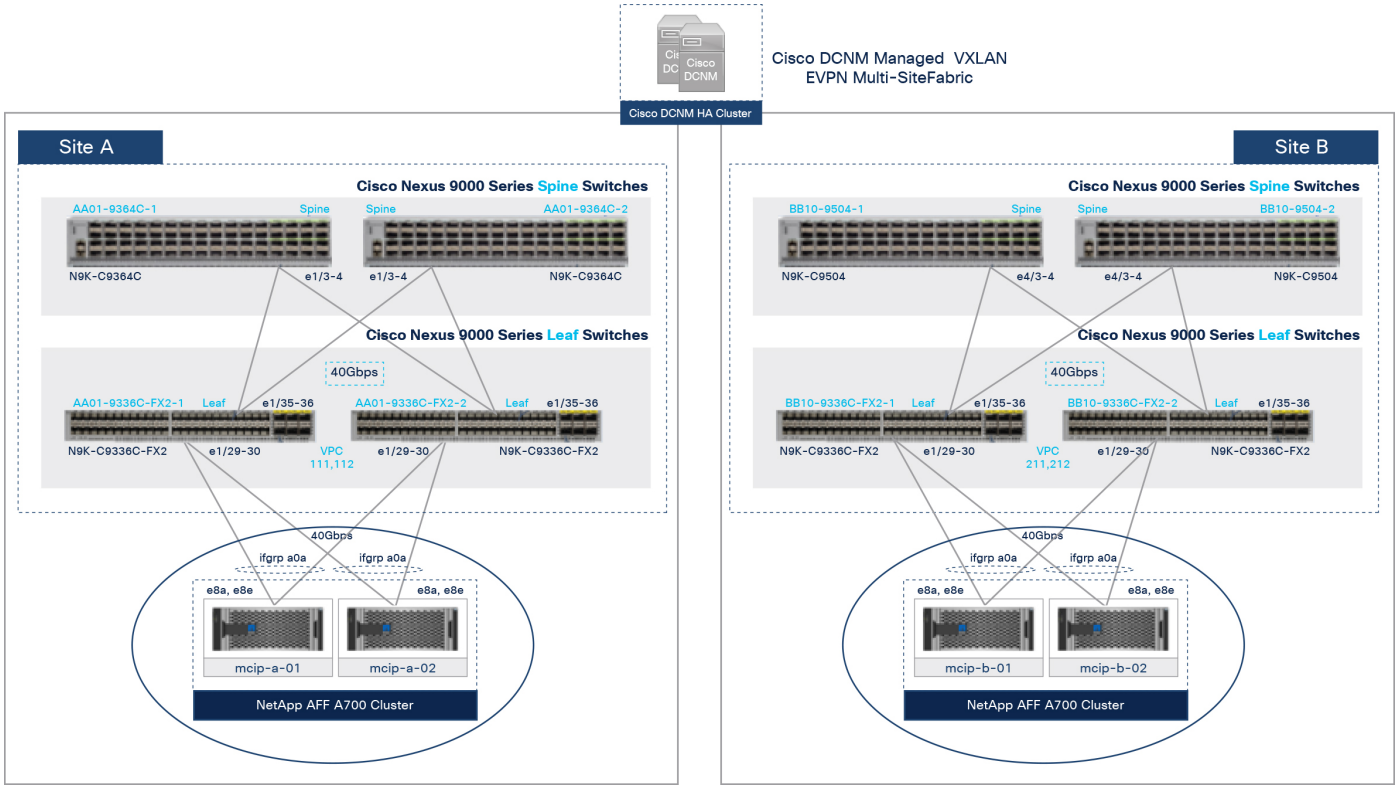


Figure 24.
NetApp Storage In-Band Management, Inter-Cluster, and Data Protocol Connectivity

The in-band management, inter-cluster, and NFS/iSCSI data storage protocols use VLANs. VLAN ports are created on the interface group to segregate the different types of traffic. Logical interfaces for the respective functions are created to use the corresponding VLAN ports. Figure 25 shows the relationship between the physical connections, interface groups, VLAN ports, and logical interfaces.

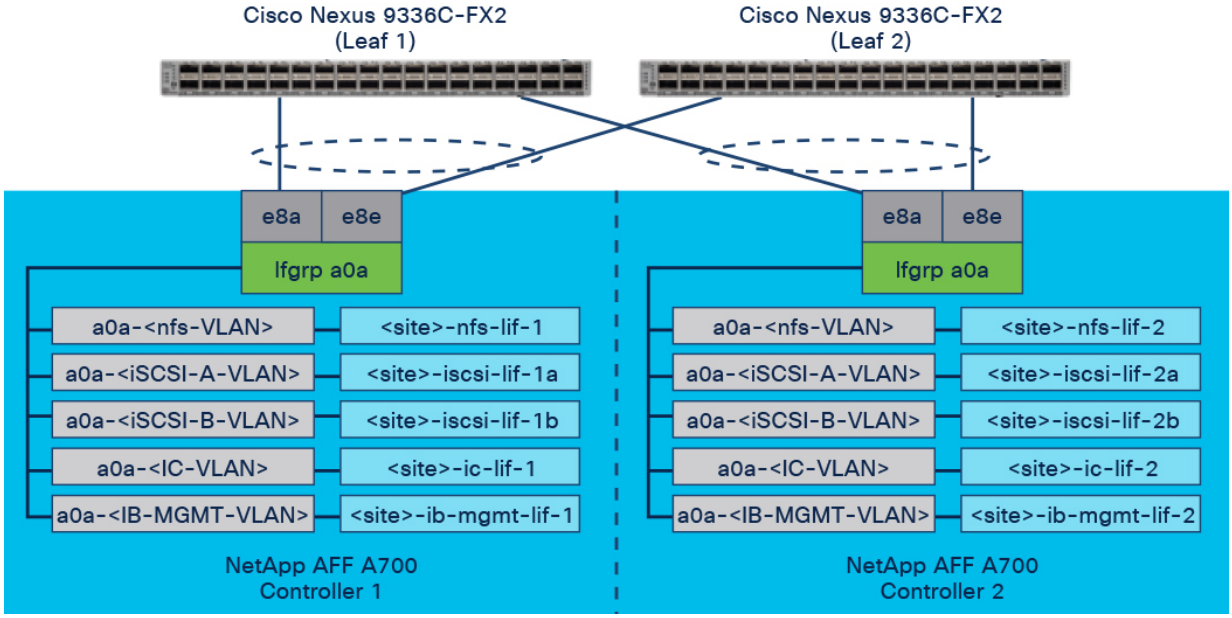


Figure 25.
Relationship Between Physical Connections, Interface Groups, VLAN Ports, and Logical Interfaces

Compute Design

Figure 26 shows the Cisco UCS compute infrastructure design for the FlexPod MetroCluster IP solution. The design is identical in both active-active data centers, down to the server models and interface cards used on the individual servers. The compute infrastructure does not have to be the same across sites, but it is highly recommended in a FlexPod MetroCluster IP solution to ensure seamless workload placement and mobility across data centers, especially during a failure event.

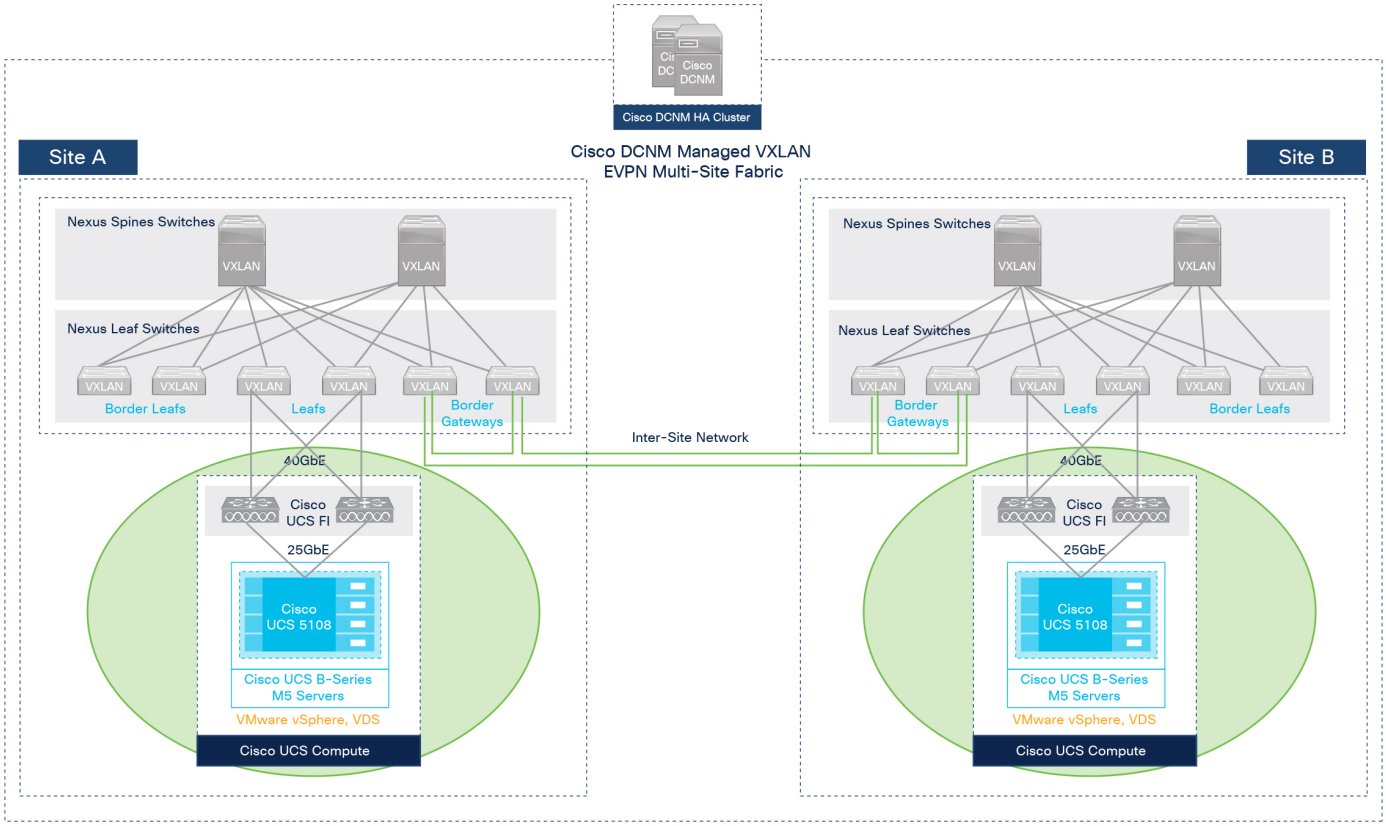


Figure 26.
Cisco UCS Compute Design – High-Level View

The Cisco UCS compute infrastructure consists of Cisco UCS servers connected to a pair of Cisco UCS Fabric Interconnects. These fabric interconnects are an integral part of Cisco Unified Computing System, providing both network connectivity and management capabilities for the system. The servers can be Cisco UCS B-Series blade servers in a UCS 5108 B-Series server chassis, UCS Managed C-Series rack servers, or UCS S-Series storage servers. The fabric interconnects can be the Cisco UCS 6400 Series (6454, 64108) for 1-/10-/25-/40-/100-Gbps Ethernet or 8-/16-/32-Gbps Fibre Channel connectivity or the Cisco UCS 6300 Series (6332, 6332-16UP, 6324 for UCS Mini) for 10-/40-Gbps Ethernet and 4-/8-/16-Gbps Fibre Channel connectivity. The Cisco UCS 5108 B-Series blade server chassis uses fabric extenders (FEXs) or IOMs that plug into the back of the chassis to connect to fabric interconnects. You can use the Cisco UCS 2408, 2208XP, or 2204XP models of FEXs when connecting to Cisco UCS 6400 Series fabric interconnects and Cisco UCS 2304, 2208XP, or 2204XP fabric extenders when connecting to Cisco UCS 6300 Series fabric interconnects. The FEXs are extensions of the fabric interconnects and act as remote line cards to form a distributed modular system. Fabric extenders and fabric interconnects are typically deployed in pairs for redundancy. A pair of Cisco UCS fabric interconnects, together with the servers connected to it, form a single, highly available, management and server networking domain referred to as a Cisco UCS domain.

Two Cisco UCS domains are deployed in this FlexPod solution to provide the compute infrastructure in the active-active data centers. The Cisco UCS domain in each site connects to a VXLAN fabric in the same site for:

- Connectivity to NetApp storage, either in the local site or in a remote site: This connectivity is necessary for enabling iSCSI SAN boot of the servers and for accessing NFS datastores.
- Connectivity to outside networks and services: The Cisco UCS servers in each site are part of the same VMware vSphere cluster, and this connectivity is necessary to deploy, operate, and manage the cluster using VMware vCenter located outside the VXLAN EVPN Multi-Site fabric. When the VMware vSphere cluster is operational, the virtual machines deployed on the cluster may also require connectivity to outside networks and services.
- Connectivity between Cisco UCS servers (ESXi hosts), either in the local site or in a remote site: This connectivity is necessary for resiliency features such as vSphere high availability, vMotion, and DRS.
- Connectivity for applications and services hosted on the Cisco UCS compute: When the VMware vSphere cluster is operational, the application (or services) components or tiers in the same site or remote site will require connectivity to each other. Enterprise users in other parts of the enterprise network (campus, WAN, remote) also will need to access applications and services hosted in either data center.

Figure 27 shows the detailed compute design for the Cisco UCS domain in each data center.

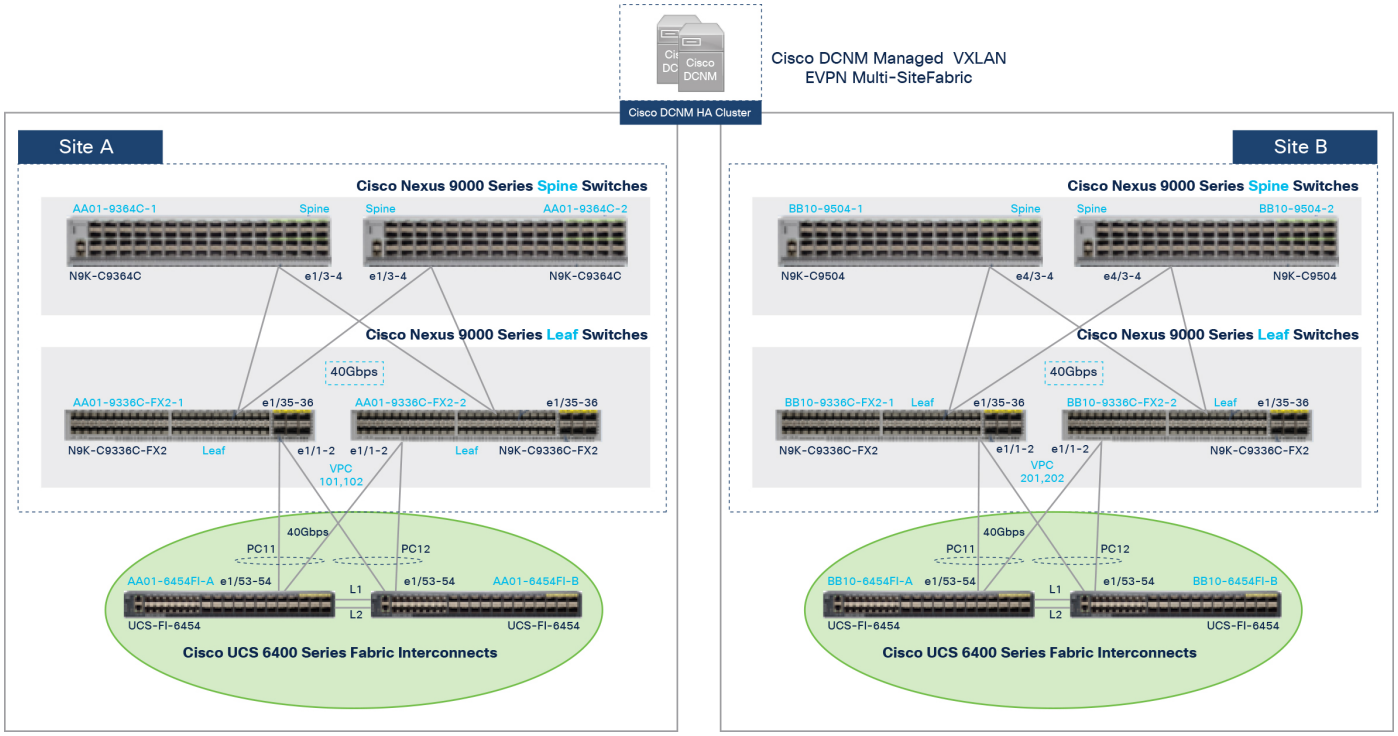


Figure 27.
Cisco UCS Compute Design – Detailed View

The Cisco UCS domain in each data center consists of a pair of Cisco UCS 6454 fabric interconnects connected to a Cisco UCS 5108 server chassis equipped with two Cisco UCS 2408 fabric extenders and three Cisco UCS B200M5 blade servers. Each FEX connects to one of the fabric interconnects using two 25-GE links to provide an aggregate uplink bandwidth of two 50-Gbps per FEX or 100 Gbps from each server chassis. Each fabric interconnect connects to the VXLAN fabric using two 40-GE links, to provide an aggregate uplink bandwidth of two 80 Gbps per fabric interconnect or 160 Gbps from the UCS domain. You can add additional links between the FEX and fabric interconnect and from the fabric interconnect to the VXLAN fabric to increase the uplink bandwidth as needed.

Cisco UCS Blade Server Chassis-to-Fabric Interconnect Connectivity

A pair of Cisco UCS 2408 fabric extenders are used in this solution for connecting the Cisco UCS 5108 blade server chassis to the Cisco UCS 6454 fabric interconnects in each site. The Cisco UCS 2408 FEX provides up to eight 25-GE links on each fabric to connect to 10-/25-GE ports on the Cisco UCS 6400 Series fabric interconnect as shown in Figure 28. The FEX links are deployed in port-channel mode where the links are bundled to provide higher uplink bandwidth from each chassis (and servers). Two 25-GE links in a port channel are used in this design. A Cisco UCS 2408 fabric extender can provide up to 200 Gbps of bandwidth from the servers in the chassis to each 6400 Series fabric interconnect.

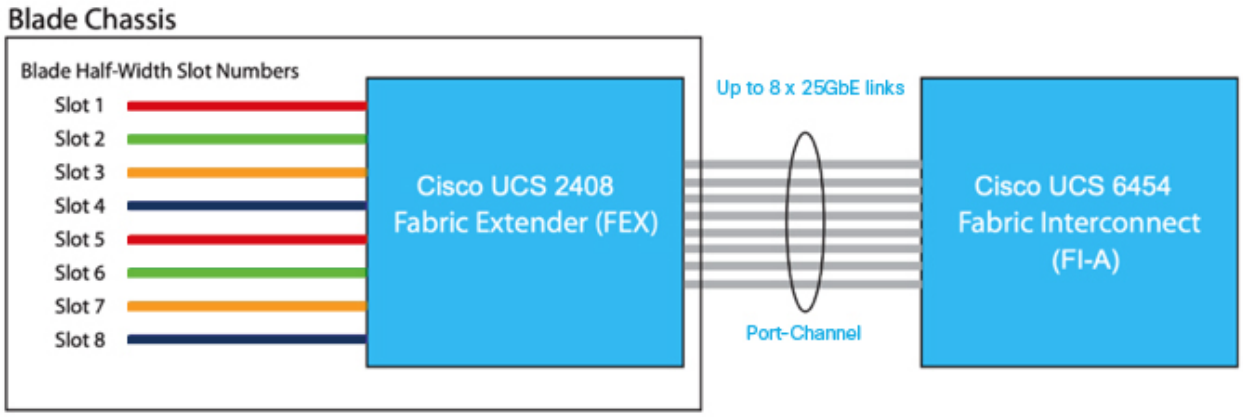


Figure 28.
Fabric Extender-to-Fabric Interconnect Connectivity

Cisco UCS Blade Server(s)-to-Fabric Extender Connectivity

The Cisco UCS B200M5 servers in the solution are deployed with a Cisco UCS VIC 1440 adapter in the mLOM slot for 40-GE uplink connectivity from each server. The different connectivity options available from the Cisco UCS VIC 1440 in a Cisco UCS B200 M5 server to the Cisco UCS 2408 FEX in a Cisco UCS blade server chassis are outlined as follows. The different connectivity options enable higher uplink bandwidth from each server. You can deploy the option that best meets your needs. Option 1 is used in this solution.

Option 1:
In this design, a single Cisco UCS VIC 1440 is deployed in the mLOM slot of each Cisco UCS B200M5 blade server. Figure 29 shows the connectivity within the Cisco UCS chassis, between the Cisco UCS VIC 1440 and Cisco UCS 2408 FEX.

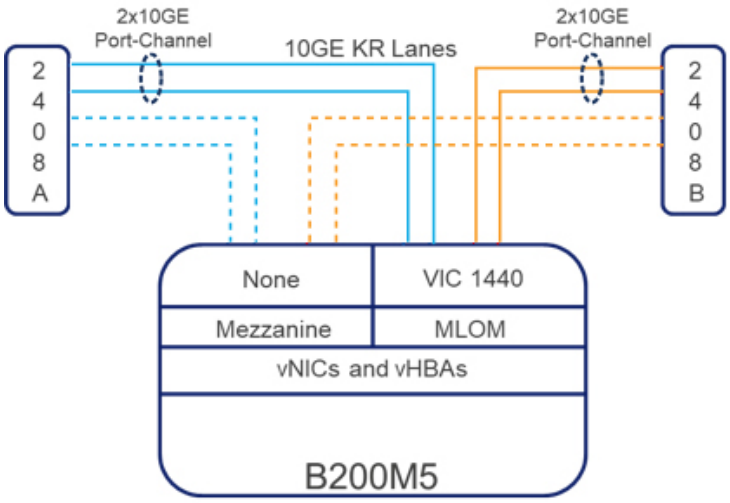


Figure 29.
Cisco UCS Blade Server Connectivity to UCS 2408 FEX Using UCS VIC 1440

Two 10-GE KR lanes connect the Cisco UCS VIC 1440 in the mLOM slot of each server to each Cisco UCS 2408 FEX in the chassis. This combination of components provides 20-GE vNICs/vHBAs to each FEX, but individual network flows or TCP sessions on these vNICs have a maximum speed of 10 Gbps. The aggregate bandwidth (with multiple flows) is 40 Gbps to each server in this design.

Option 2:
In this design, a Cisco UCS VIC 1440 and a port expander are deployed on each Cisco UCS B200M5 server for higher aggregate bandwidth per server. Cisco UCS 2408 FEX and Cisco UCS VIC 1440 support using a port expander in the mezzanine slot. Figure 30 shows the connectivity within the Cisco UCS chassis, from the Cisco UCS VIC 1440 and port expander to the Cisco UCS 2408 FEX.

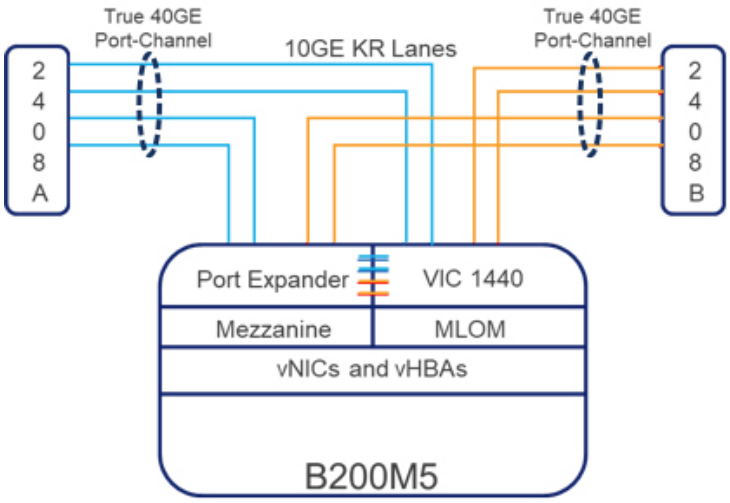


Figure 30.
Cisco UCS Blade Server Connectivity to UCS 2408 FEX Using UCS VIC 1440 and Port Expander

Two 10-GE KR lanes connect each Cisco UCS 2408 FEX to the Cisco UCS VIC 1440 in the mLOM slot. Two more 10-GE KR lanes connect each 2408 FEX to the port expander in the mezzanine slot. The port expander also connects its KR lanes to the Cisco UCS VIC 1440. This combination of components along with timing of the KR lanes provides true 40-GE vNICs/vHBAs per FEX, but individual network flows or TCP sessions on these vNICs have a maximum speed of 25 Gbps since 25-Gbps links are used in the port channel between the Cisco UCS 2408 IOM and the fabric interconnect. The aggregate bandwidth (with multiple flows) is 80 Gbps to each server in this design.

Option 3:
In this design, a Cisco VIC 1480 is deployed in the mezzanine slot instead of the port expander along with the Cisco UCS VIC 1440 for higher aggregate bandwidth per server. Figure 31 shows the connectivity within the Cisco UCS chassis, from the Cisco UCS VIC 1440 and VIC 1480 to the Cisco UCS 2408 FEX.

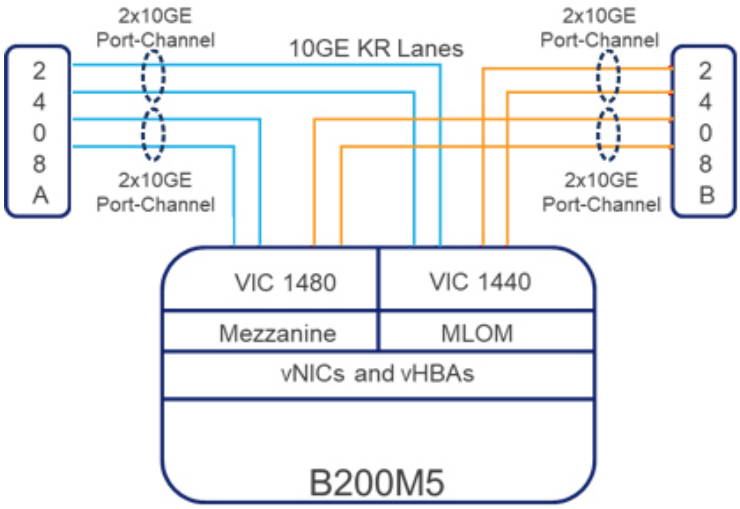


Figure 31.
Cisco UCS Blade Server Connectivity to Cisco UCS 2408 FEX Using UCS VIC 1440 and 1480

Rack-Server Connectivity to Cisco UCS Fabric Interconnects

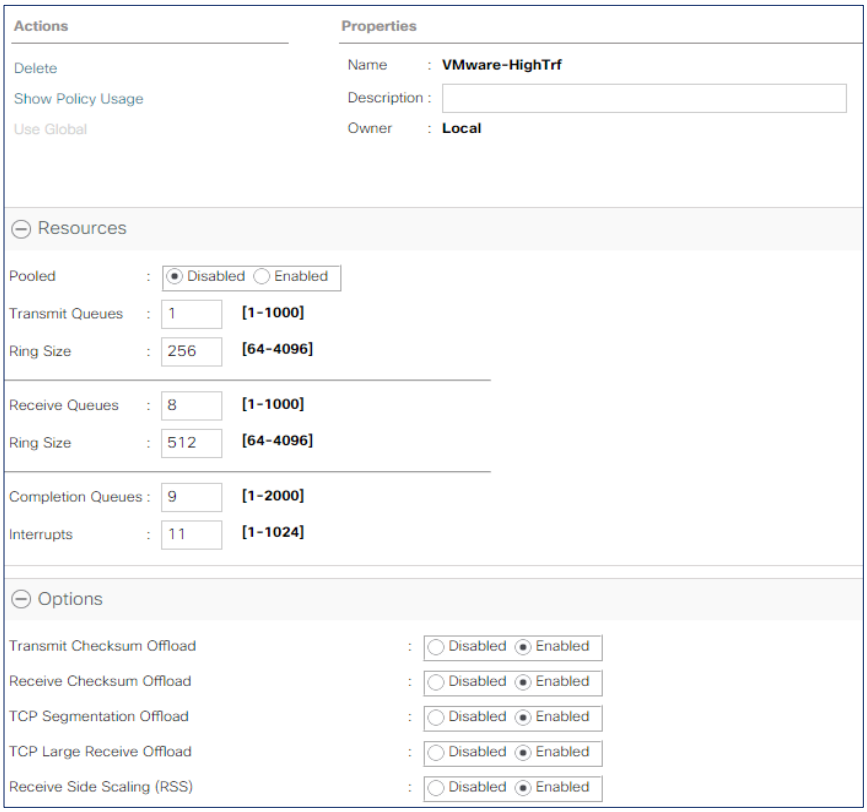
You can also use Cisco UCS C-Series servers in this solution. The Cisco UCS C-Series servers support both Cisco UCS VIC 1455 and 1457. These adapters provide uplink connectivity, and you can connect them directly to Cisco UCS 6400 fabric interconnects. The connections from the fabric interconnects to the VIC are either single-or dual-link port channels. If 25-GE interfaces on all four links are used, and vNICs/vHBAs are 50 Gbps, with an aggregate of 100 Gbps to each server. Individual network flows or TCP sessions on these vNICs have a maximum speed of 25 Gbps.

Note: When using Cisco UCS VIC 1455 and 1457 with Cisco UCS 6300 fabric interconnects, only single-link port channels, or one 10-GE connection to each fabric interconnect, is supported.

Cisco vNIC Ethernet Adapter Policy

The Ethernet adapter policy governs the host-side behavior of the vNIC, including how the adapter handles traffic. For example, you can use these policies to change default settings for queues, interrupt handling, enhance performance, and receive side scaling (RSS) hash. Cisco UCS provides a default set of Ethernet adapter policies, including the recommended settings for each supported server operating system. Operating systems are sensitive to the settings in these policies.

A custom VMware-HighTrf Ethernet adapter policy was configured in this solution according to the section "Configuring an Ethernet Adapter Policy to Enable eNIC Support for RSS on VMware ESXi" in the Cisco UCS Manager Network Management Guide, Release 4.1. This policy is designed to provide higher performance on vNICs with a large number of TCP sessions by providing multiple receive queues serviced by multiple CPUs. The following screenshot shows the **VMware-HighTrf** Ethernet Adapter Policy.



Connectivity from Cisco UCS Fabric Interconnects to VXLAN Fabric

Figure 32 shows the connectivity to Cisco UCS domains in the active-active data centers. Each fabric interconnect uses two 40-GE links in a port- channel to connect to top-of-rack leaf switches in the VXLAN fabric, for an aggregate bandwidth of two 80-Gbps per fabric interconnect or 160 Gbps for each Cisco UCS domain in the solution. You can add links as needed to increase this bandwidth.

Figure 32.
Cisco UCS Domain – Uplink to VXLAN Fabric

The top-of-rack leaf switches in the VXLAN fabric that the fabric interconnects connect to in each data center are configured for vPCs. The leaf switches in each data center are deployed as a vPC leaf switch pair, and virtual peer links are used instead of physical links in the vPC setup. The vPC configuration is also deployed by Cisco DCNM in this solution.

FlexPod Infrastructure and Application Networks – Cisco UCS Domain

Table 5 lists the FlexPod infrastructure and application networks defined in the Cisco UCS domain. These VLANs are trunked on the uplinks to the VXLAN fabric, and the fabric provides the necessary connectivity for those networks.

Table 5. FlexPod Infrastructure and Application Networks Defined in Cisco UCS Domain

Network Type	VLAN Name	VLAN	Description
In-Band Management	IB-MGMT_VLAN	122	ESXi Management
vMotion	vMotion_VLAN	3000	VMware vMotion
Storage Data	iSCSI-A_VLAN	3010	Storage Access – iSCSI boot of Cisco UCS (ESXi) servers
Storage Data	iSCSI-B_VLAN	3020	Storage Access – iSCSI boot of Cisco UCS (ESXi) servers
Storage Data	NFS_VLAN	3050	Storage Data Network for accessing NFS datastores
VM Network	VM-Network-1_VLAN VM-Network-2_VLAN VM-Network-3_VLAN	1001-1003	Application & Service VM Hosted on FlexPod Infrastructure

Virtualization Design

In the FlexPod MetroCluster IP solution with VXLAN Multi-Site fabric, a single VMware vCenter manages the VSI in each data center. The hosts in both data centers are part of a single ESXi cluster that spans both data centers. A VMware vCenter server virtual machine is deployed on existing infrastructure that could be thought of as a third site and not part of the VXLAN multi-site fabric. VMware vCenter has independent connectivity to both sites for reachability to hosts in each data center that is part of the same ESXi cluster.

For application and services virtual machines deployed on the ESXi cluster, NFS and iSCSI datastores are created on both NetApp AFF A700 systems at both sites to host the local virtual machines. Under normal conditions, virtual machines in a given site will access storage from the NetApp arrays in the same site. However, you can deploy virtual machines in either location because the design uses active-active data centers that provide flexible workload placement and mobility. As part of implementation design, the distribution of the virtual machines across the two sites must be determined, and some virtual machines are hosted primarily in site A while the others are hosted in site B. You can determine this virtual machine and application distribution across the two sites according to your site preferences and requirements. For optimal virtual-machine performance, the virtual-machine disks should be hosted on the local NetApp AFF A700 systems to avoid additional latency and traffic across the WAN links under normal operation. VMware DRS is configured with site affinity rules to make sure the virtual machines adhere to site preference requirements.

For a site failure scenario, virtual machines running at the failed site must be restarted at the surviving site. To accommodate virtual machines that normally run at both sites, all the iSCSI and NFS shared datastores must be mounted on all the ESXi hosts to ensure a smooth vMotion operation of virtual machines between sites. For additional best practices on configuring VMware for a Metro storage cluster, refer to the VMware vSphere Metro Storage Cluster (vMSC) and VMWare KB for deploying NetApp MetroCluster documentation.

Virtual Networking

Each host in the cluster is deployed using identical virtual networking regardless of its location. The design separates the different traffic types using VMware virtual switches (vSwitch) and VMware Virtual Distributed Switches (vDS). The VMware vSwitch is used primarily for the FlexPod infrastructure networks and vDS for application networks, but it is not required. The virtual switches (vSwitch, vDS) are deployed with two uplinks per virtual switch; the uplinks at the ESXi hypervisor level are referred to as vmnics and virtual NICs (vNICs) on Cisco UCS Software. The vNICs are created on the Cisco UCS VIC adapter in each server using Cisco UCS service profiles. Eight vNICs are defined, two for vSwitch0, two for vDS1, two for vSwitch1, and two for the iSCSI uplinks.

vSwitch0 is defined during VMware ESXi host configuration, and it contains the FlexPod infrastructure management VLAN and the FlexPod infrastructure NFS VLAN. The ESXi host VMkernel (VMK) ports, for management, and infrastructure NFS are placed on vSwitch0. An infrastructure management virtual machine port group is also placed on vSwitch0 for any critical infrastructure management virtual machines that are needed. It is important to place such management infrastructure virtual machines on vSwitch0 instead of the vDS because if the FlexPod infrastructure is shut down or power cycled and you attempt to activate that management virtual machine on a host other than the host on which it was originally running, it will boot up fine on the network on vSwitch0. It is particularly important if VMware vCenter is the management virtual machine. If vCenter were on the vDS and moved to another host and then booted, it would not be connected to the network after booting up.

It is important to note that the vDS deployed will span all hosts in the cluster, so the vDS will span both active-active data centers in this design. The vDS deployed for this solution can contain port groups for application tenant iSCSI, NFS VMKs, virtual machines (for in-guest iSCSI and NFS), tenant management, and other virtual-machine networks. The vDS uplinks have the VMware-HighTrf UCS Ethernet adapter policy, providing more queuing and throughput on the adapters.

vMotion VMK is placed on a vSwitch (vSwitch1), but you can also place it on the vDS to allow for application of QoS to vMotion if necessary. If you us the Teaming and Failover policy of the port group, you can pin vMotion to the switching fabric B uplink with the fabric A uplink as standby so that under normal conditions vMotion uses a specific fabric. A final note with vMotion is that if a server vNICs are 40 Gbps or greater, you can provision three vMotion VMKs in the same subnet on the server to allow vMotion to establish multiple sessions and take advantage of the higher bandwidth. You also can optionally place Infrastructure NFS on the vDS for higher performance, but it is left on vSwitch0 for administrative simplicity. You can also assign the VMware-HighTrf with the Cisco UCS Ethernet adapter policy to the vSwitch vNICs for higher performance. You should pin tenant iSCSI port groups to the appropriate switching fabric uplink with the other fabric uplink set as unused.

For an application and services virtual machine deployed on the FlexPod virtual infrastructure, you can add new networks and port groups as needed. You must also add the corresponding VLANs to Cisco UCS Manager and to the vDS vNIC templates. On both vSwitches and the vDS, all port groups initially use the Route based on originating virtual port hashing method. If multiple ports in the same port group are configured on a virtual machine, or for better VMK distribution, consider using the Route based on source MAC hash method. Do not use the Route based on IP hash method because that method requires port channeling configuration on the connected switch ports.

You also can deploy additional vDSs (maybe on a per-tenant basis) with dedicated vNIC uplinks, allowing for role-based access control (RBAC) for visibility and/or alignment of vDS to the respective tenant manager in vCenter. However, you do not need to have separate vDSs for each tenant. You can add any new networks to the vDS used in this FlexPod design.

Two iSCSI boot vSwitches are used in this design. Cisco UCS iSCSI boot requires separate vNICs for iSCSI boot. These vNICs use iSCSI VLAN of the appropriate fabric as the native VLAN and are attached to the appropriate iSCSI boot vSwitch. Optionally, you could also deploy iSCSI networks on vDS by deploying a new vDS or using an existing one.

Table 6 lists the FlexPod infrastructure and application networks, the corresponding VLANs on Cisco UCS, the vNICs created on Cisco UCS for use as uplinks on ESXi hosts, and the corresponding VMware configuration for them. Figure 33 shows a layout of the ESXi virtual networking design.

Table 6. Server Networking – Cisco UCS and VMware vSphere

VLAN ID	HyperFlex & UCS VLAN Names	UCS Server Uplink Names	UCS Server vNIC(s)	VMware Virtual Switch	VMware vmnic(s)	Description
122	IB-MGMT_VLAN	vSwitch0-A	vNIC 01	vSwitch0	vmnic0	ESXi Management Network
3050	NFS_VLAN	vSwitch0-B	vNIC 02		vmnic1	ESXi Storage Data Access - NFS
3000	vMotion_VLAN	vSwitch1-A	vNIC 05	vSwitch1	vmnic2	VMware vMotion Network
		vSwitch1-B	vNIC 06		vmnic3	
3010	iSCSI-A_VLAN	iSCSI-A	vNIC 07	iScsiBootvSwitch	vmnic4	ESXi Storage Data Access - iSCSI
3020	iSCSI-B_VLAN	iSCSI-B	vNIC 08	iScsiBootvSwitch-B	vmnic5	ESXi Storage Data Access - iSCSI
1001-1103	VM-Network-1_VLAN VM-Network-2_VLAN VM-Network-3_VLAN	vDS1-A	vNIC 09	vDS1	vmnic6	Application & Service VM Hosted on FlexPod Infrastructure
		vDS1-B	vNIC 10		vmnic7	

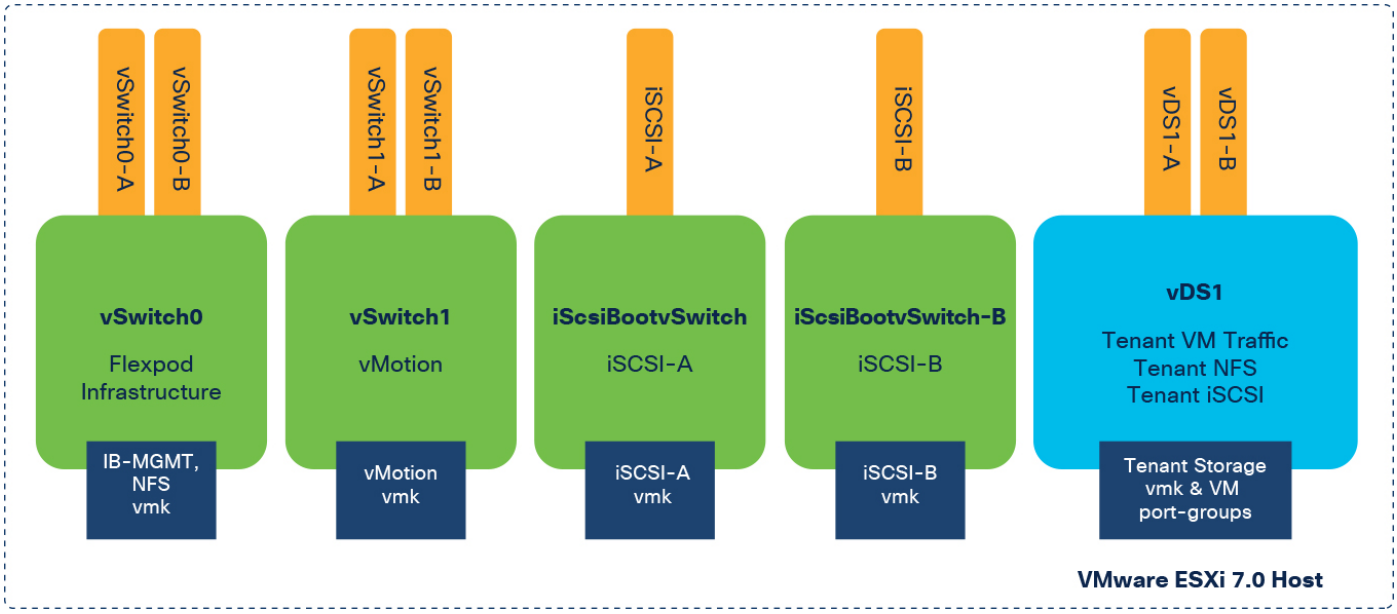


Figure 33.
Virtual Networking – vNIC Design on ESXi Hosts

End-to-End Design

iSCSI SAN Boot

NetApp recommends implementing SAN boot for the Cisco UCS servers in the FlexPod solution. Implementing SAN boot enables you to safely secure the operating system with the NetApp storage system, providing better performance and flexibility. In this design, iSCSI SAN boot is validated.

In iSCSI SAN boot, each Cisco UCS server is assigned two iSCSI vNICs (one for each SAN fabric) that provide redundant connectivity all the way to the storage. The 40-Gb Ethernet storage ports that are connected to the Nexus switches, in this example e8a and e8e, are grouped together to form one logical interface group (ifgrp) (in this example, a0a). The iSCSI virtual LANs (VLANs) are created on the ifgrp and the iSCSI LIFs are created on the iSCSI VLAN ports. Each iSCSI boot LUN is mapped to the server that boots from it through the iSCSI LIFs by using igroups. This feature enables only the authorized server to have access to the boot LUN (refer to Figure 34).

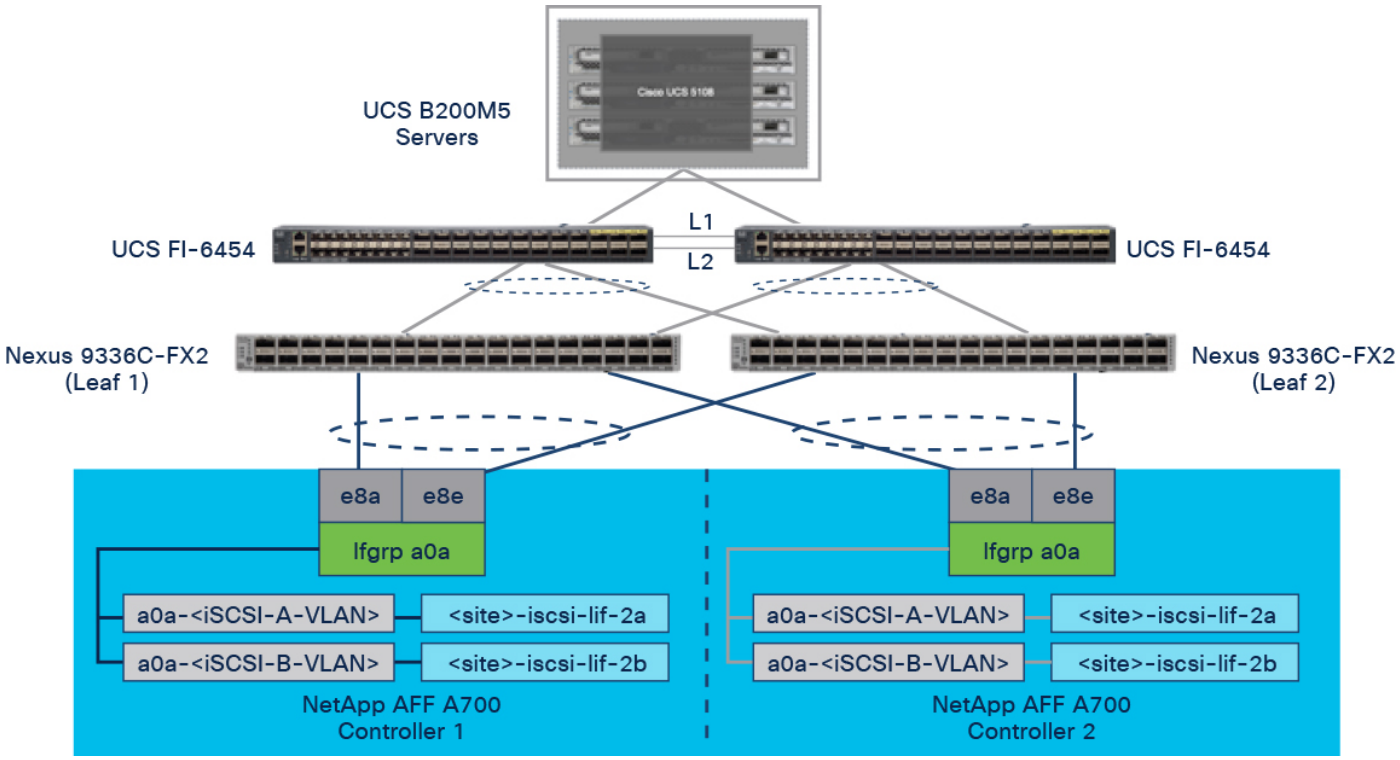


Figure 34.
Connectivity for iSCSI SAN Boot of Cisco UCS Servers from NetApp Storage

End-to-End IP Network Connectivity

Figure 35 shows the end-to-end IP connectivity between ESXi hosts on Cisco UCS B200M5 servers and NetApp MetroCluster IP storage for iSCSI/NFS storage access within a site. Each site also has connectivity to an identically deployed second site. The Cisco VXLAN Multi-Site fabric provides connectivity between port-channeled 40-/100-Gbps connected Cisco UCS compute and NetApp storage in each data center, as well as connectivity between data centers. vPCs extend from VXLAN leaf switches to both the AFF A700 controllers and the Cisco UCS 6454 fabric interconnects in each site.

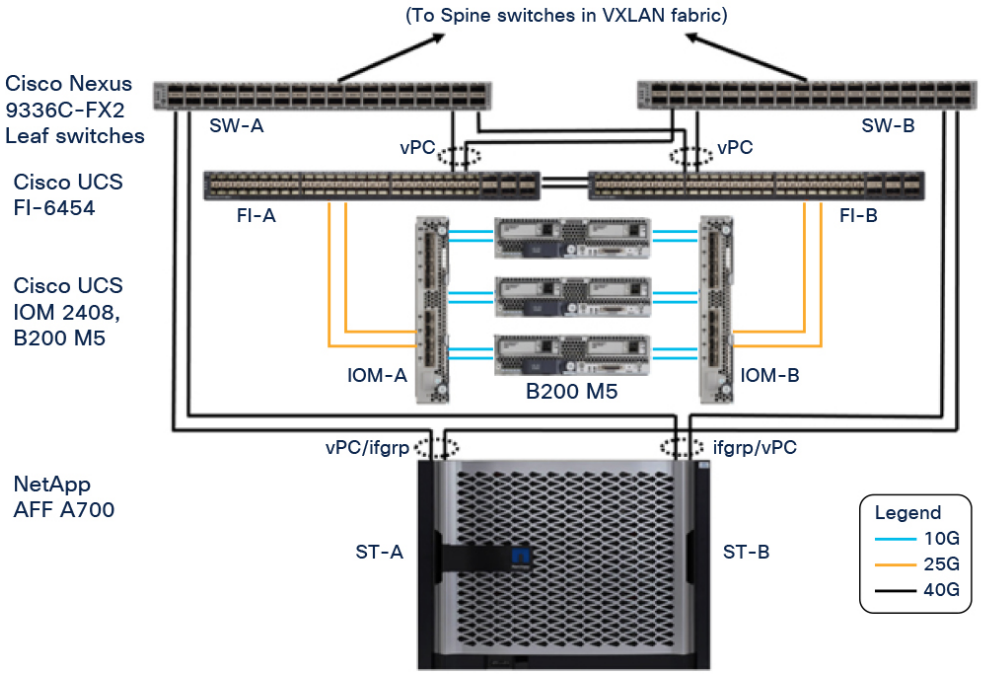


Figure 35.
Component-Level Connectivity in Each Data Center

The flow of storage access traffic (IP traffic) from ESXi host/Cisco UCS server to NetApp storage is summarized as follows; from left to right,

- The traffic flow starts from the Cisco UCS B200 M5 server/ESXi host, equipped with a Cisco UCS VIC 1440 adapter and port expander that provides 40-GE connectivity through each fabric (FI-A, FI-B) from the server.
- The traffic flow traverses timed 10-Gb KR lanes of the Cisco UCS 5108 chassis backplane into the Cisco UCS 2408 IOM (FEX).
- The traffic then flows through the IOM to the fabric interconnects. Each IOM to the fabric interconnect provides up to eight 25-GE links automatically configured as port channels during chassis association.
- Traffic then continues from the Cisco UCS 6454 fabric interconnects into the Cisco Nexus 9336C-FX2 VXLAN fabric leaf switches through bundled (port-channel) 40-/100-GE ports that go to both Nexus leaf switches. The leaf-switch pair is configured for vPC and therefore presents itself as a single switch, but with node-level redundancy.
- Traffic destined for the remote data center then traverses the VXLAN Multi-Site fabric to the NetApp arrays in the remote data center.
- Traffic destined to the NetApp arrays in the local site will be switches by the leaf switch pair to AFF A700 controllers in the same site. The NetApp A700s connect to both Nexus leaf switches using 40-GE links in a port-channel bundle. The leaf switch pair is configured for vPC and therefore presents itself as a single switch, but with node-level redundancy.

Figures 36 and 37 show s graphical view of the previously described traffic flow for storage access from Cisco UCS B-Series servers.

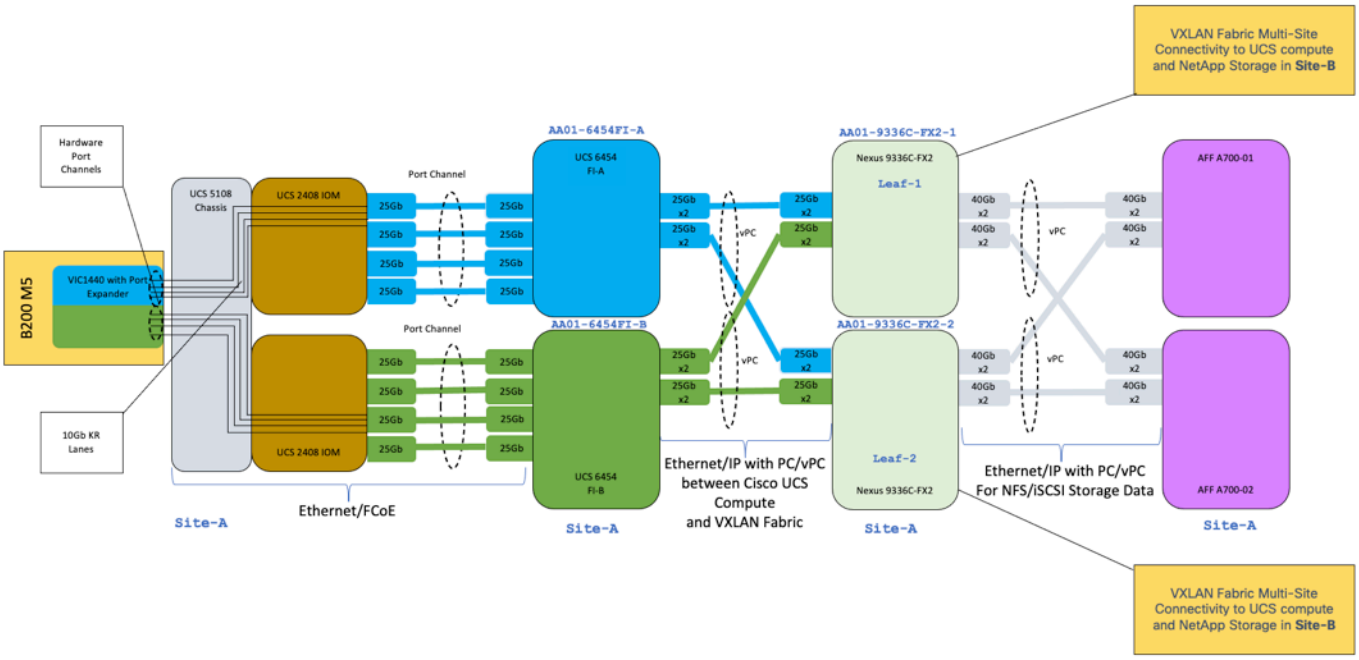


Figure 36.
End-to-End IP (iSCSI/NFS) Connectivity Between Cisco UCS B-Series Server and NetApp Storage in Site-A

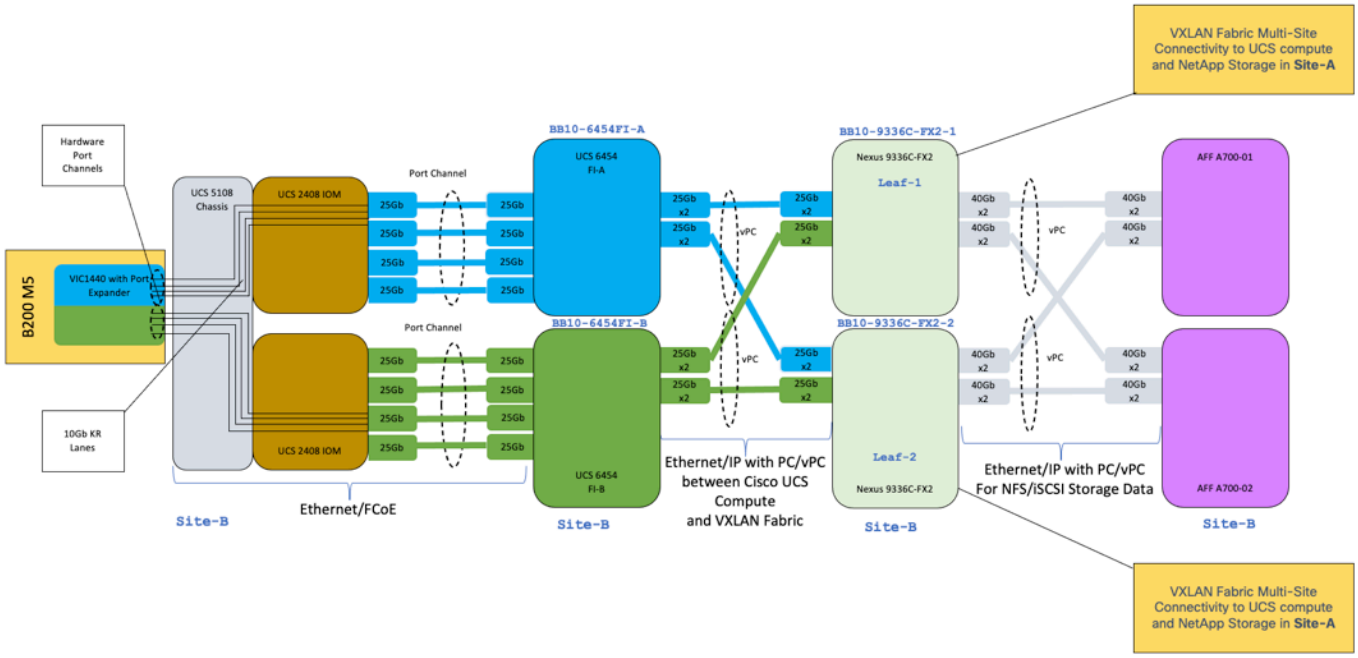


Figure 37.
End-to-End IP (iSCSI/NFS) Connectivity Between Cisco UCS B-Series Server and NetApp Storage in Site-B

On Cisco UCS C-Series servers, the traffic flow is similar to that of Cisco UCS B-series. Although a given flow is limited to 25 Gbps because of the IOM uplinks and port connections of the 1455/1457, multiple flows can exhaust the greater available bandwidth going across the port channels of the network fabric. Figures 38 and 39 show a graphical view of this traffic flow for storage access from Cisco UCS C-Series servers.

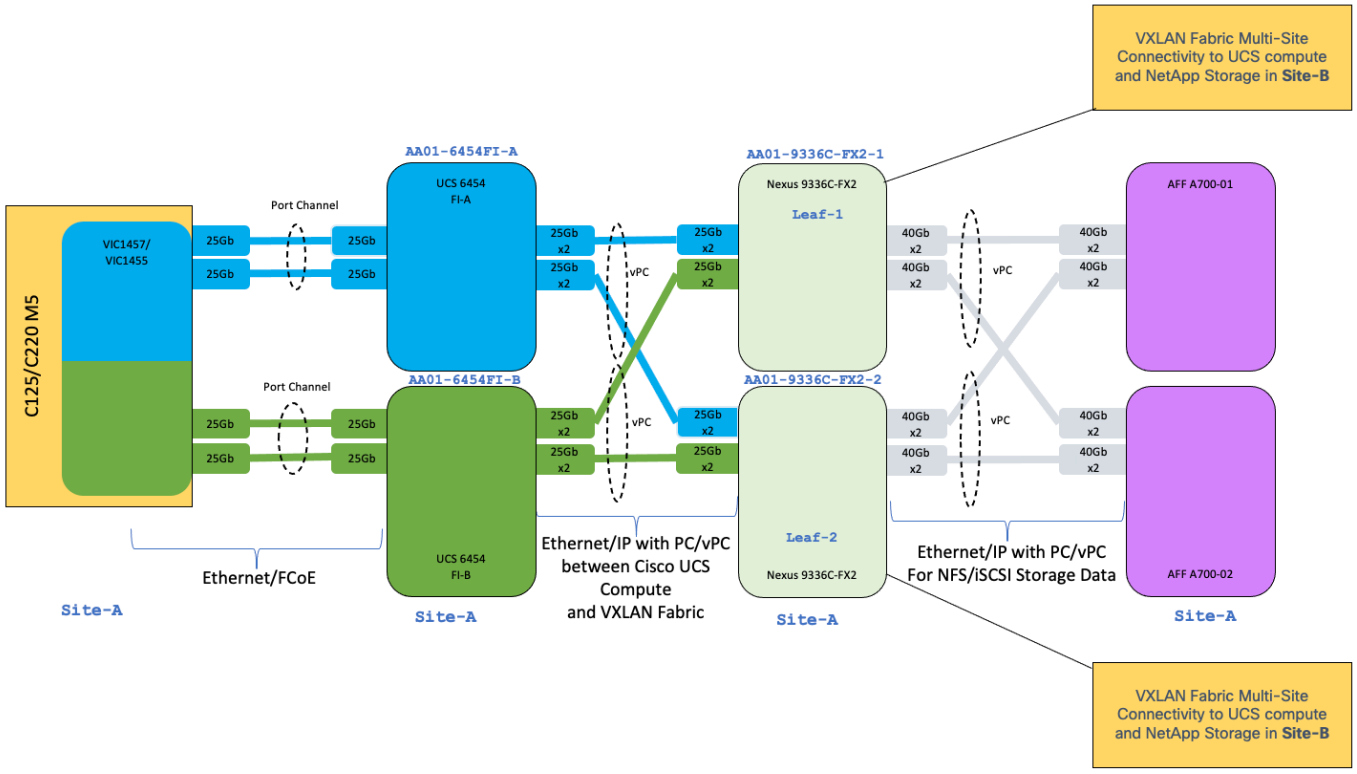


Figure 38.
End-to-End IP (iSCSI/NFS) Connectivity Between Cisco UCS C-Series Server and NetApp Storage in Site-A

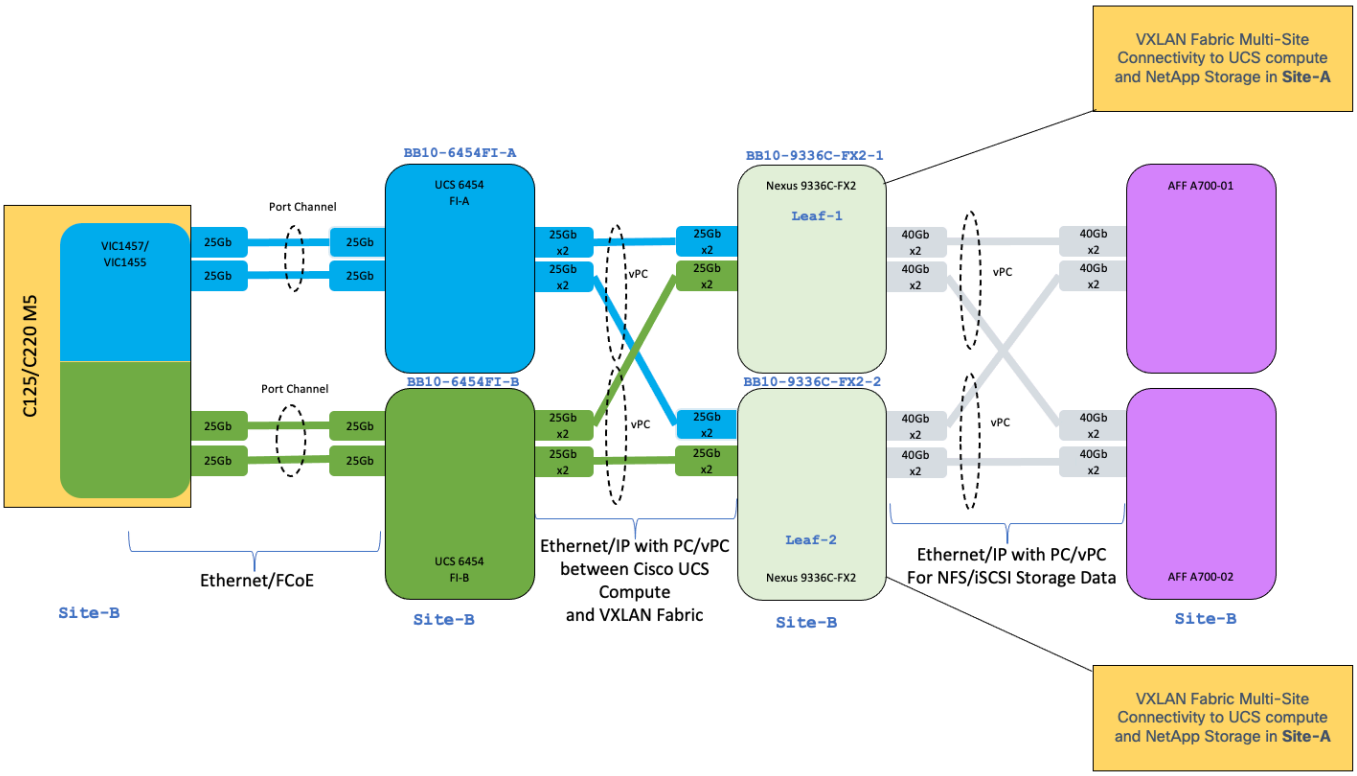


Figure 39.
End-to-End IP (iSCSI/NFS) Connectivity Between Cisco UCS C-Series Server and NetApp Storage in Site-B

Note: The previous figures show additional connections to what was used during validation, but they constitute an example of what you could deploy, depending on port and adapter availability. Please refer to physical topology diagrams in the "Solution Design" section for specific information about what was used for validation.

Solution Validation

This section provides details of the validation done for this solution.

Hardware and Software

Table 7 lists the hardware and software versions used for the solution validation testing.

Table 7 Hardware and Software Used for Solution Validation Testing

Component		Software	Count
Network	Cisco DCNM	11.5(1)	3 virtual-machine clusters
	Nexus 9336C-FX2	9.3(6)	4 (2 per site)
	Other Nexus spine and leaf switches in the fabric	9.3(6)	8 (4 per site)
Compute	Cisco Intersight Platform	–	–
	Cisco UCS Fabric Interconnect 6454	4.1(3a)	4 (2 per site)
	Cisco UCS IOM 2408 on UCS 5108 blade server chassis	4.1(3a)	4 (2 per site)
	Cisco UCS B200 M5 servers	4.1(3a)	4 (2 per site)
	VIC firmware - Cisco VIC 1440 on Cisco UCS B200 M5 (PID: UCSB-MLOM-40G-04)	5.1(3a)	4 (2 per site)
	Cisco VIC Driver	1.0.33.0	6 (3 per site)
Virtualization	VMware ESXi	7.0U1	6 (3 per site)
	VMware vCenter	7.0U1	1
Storage	NetApp AFF A700	ONTAP 9.8	4 (2 per site)
	Cisco Nexus 3132Q-V	9.3(5)	4 (2 per site)
	NetApp Virtual Storage Console	9.7.1P1	1
	NetApp NFS Plug-in for VMware VAAI	2.0	6 (3 per site)
	NetApp Active IQ Unified Manager	9.8P1	1
	NetApp SnapCenter Plug-in for VMware vSphere	4.4	1
	NetApp ONTAP Mediator	1.2	1

Validated Scenarios

The deployed FlexPod MetroCluster IP Datacenter with the Multi-Site VXLAN fabric solution is validated for its desired solution functions and various failure scenarios for which the solution is designed to protect.

Solution Functions Validation

The FlexPod Datacenter with Cisco VXLAN, NetApp MetroCluster IP, and VMware vSphere 7.0U1 solution is validated for successful infrastructure configuration, high availability, and business continuity across two sites. A variety of test cases are used to verify solution functions and simulate partial and complete site failure scenarios. To minimize duplication with the tests already performed in the existing FlexPod VXLAN solution Cisco Validated Design Program, the focus of this paper is on the MetroCluster IP aspects of the solution. Some general FlexPod validations are included for practitioners to go through for their implementation validations.

For the solution validation, one Windows 10 virtual machine per ESXi host was created on all ESXi hosts at both sites. The IOMeter tool was installed and used to generate I/O to two virtual data disks that are mapped from local NFS and iSCSI datastores. The IOMeter workload parameters configured were 16-KB I/O, 75% read, and 50% random, with 8 outstanding I/O commands for each data disk. For most of the test scenarios performed, the continuation of IOMeter I/O serves as an indication that the scenario did not cause a data service outage.

ESXi Host iSCSI SAN Boot Test

The ESXi hosts in the solution are configured to boot from iSCSI SAN. Using iSCSI SAN boot simplifies the blade-server management when replacing a blade because the service profile of the server can be associated with a new blade for it to boot up without making any additional configuration changes.

In addition to booting an ESXi host located at a site from a local iSCSI LUN, testing was also performed to boot an ESXi host from the remote site mirrored copy. This step helps ensure that the ESXi hosts can boot up during a maintenance or disaster scenario, when the local storage is not available, to verify disaster-recovery capability.

For the AFF A700 used for the design verification, an iSCSI LUN hosted by a storage controller has multiple redundant paths that an initiator can access. A host can get to the LUN through the two iSCSI VLANs/fabrics to the LUN hosting controller as well as through the high-availability partner of the controller (refer to Figure 40).

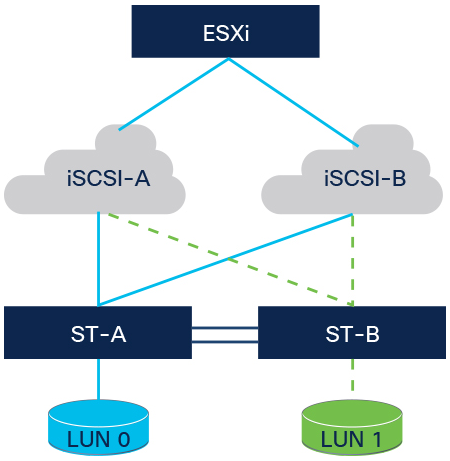
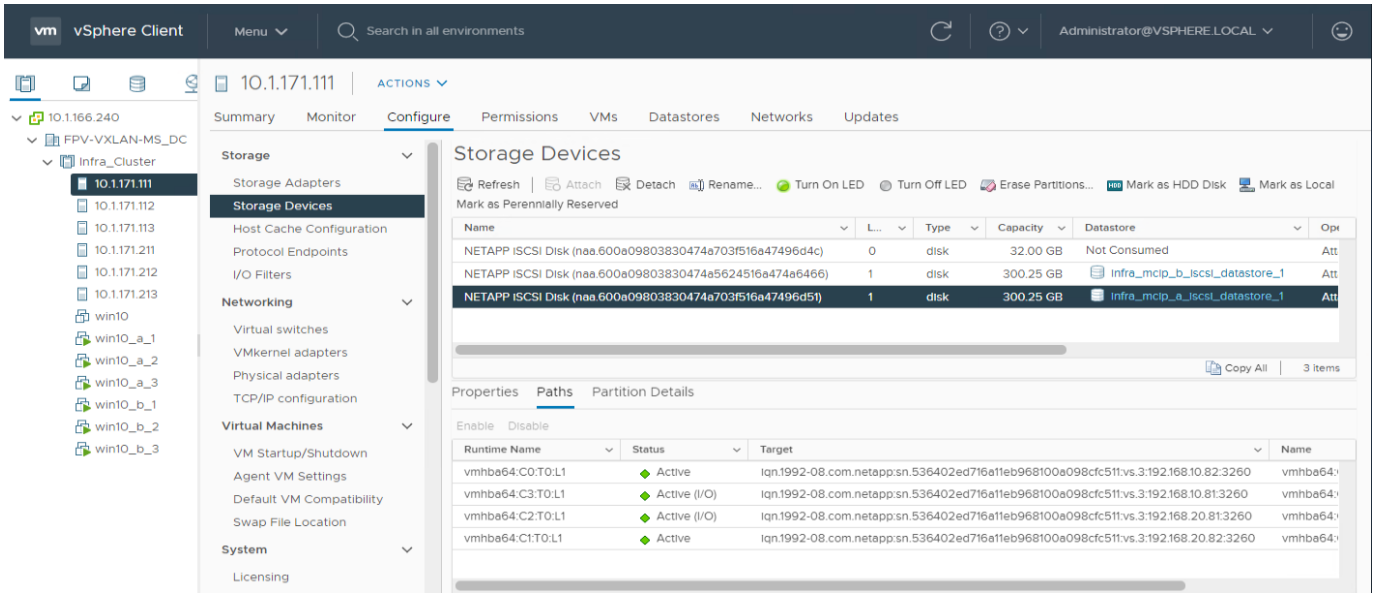


Figure 40.
LUNs Presented to an ESXi Host Are Accessible Through Multiple Paths

For the LUN ID 0 in figure, the two paths through storage controller A (ST-A) are active/optimized paths, whereas the other two paths through storage controller B (ST-B) are active/non-optimized. For LUN ID 1, it is just the opposite. The following storage device path information screenshot shows how the ESXi host sees the two type of device paths. The two active/optimized paths are shown as having “active (I/O)” path status, whereas the two active/non-optimized paths are shown only as “active”.

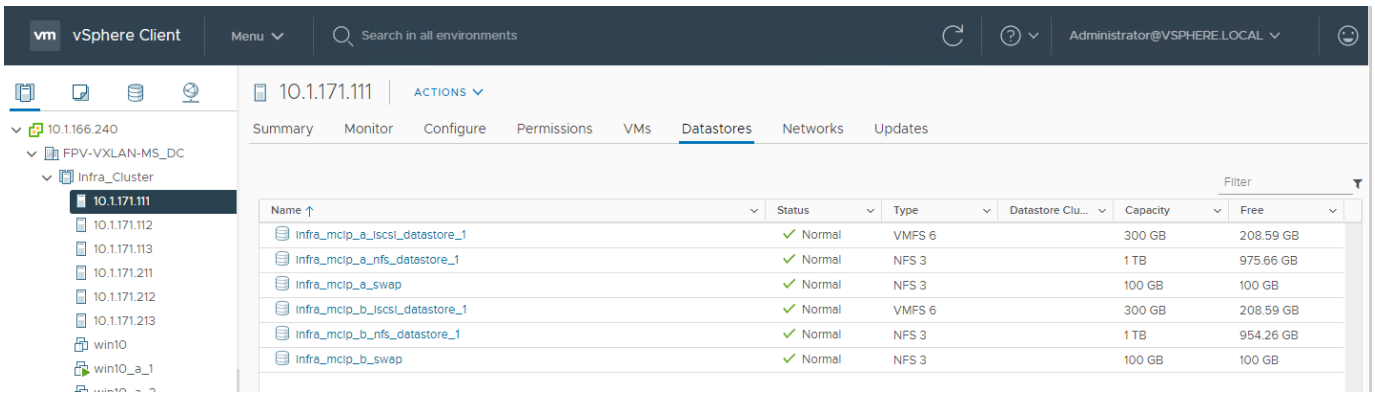


When one of the storage controllers goes down for maintenance or upgrade, the two paths that go through the down controller will no longer be available and will show up with a path status of dead instead.

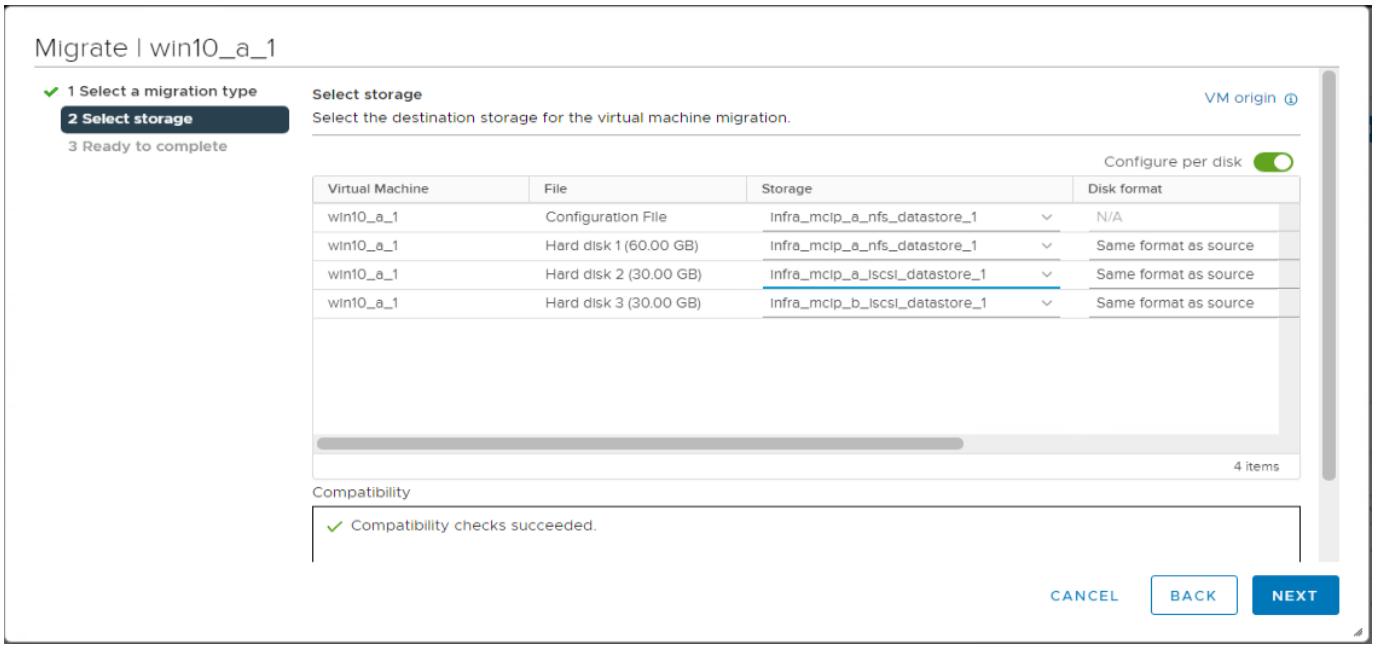
If site disaster or a scheduled MetroCluster switchover occurs, the surviving storage cluster takes over the data services of the storage cluster that went down. Because the LUN identities are preserved and the data has been replicated synchronously, an ESXi host can still boot up from its iSCSI SAN boot LUN from the remote MetroCluster IP storage.

ESXi Host iSCSI and NFS Shared Datastore and VMware vMotion Test

The FlexPod MetroCluster IP solution supports multiple protocols, including the iSCSI and NFS protocols used for this validation. To allow virtual machines to use storage services with different protocols and from either MetroCluster sites, the iSCSI and NFS datastores from both sites must be mounted by all the hosts in the cluster. This mounting enables a virtual machine to migrate its virtual disks between the datastores using different protocols and served from both sites. The following screenshot shows hosts configured to mount iSCSI and NFS datastores from both sites.



You can migrate virtual-machine disks between available NFS and iSCSI datastores from both sites, as shown in the following screenshot:



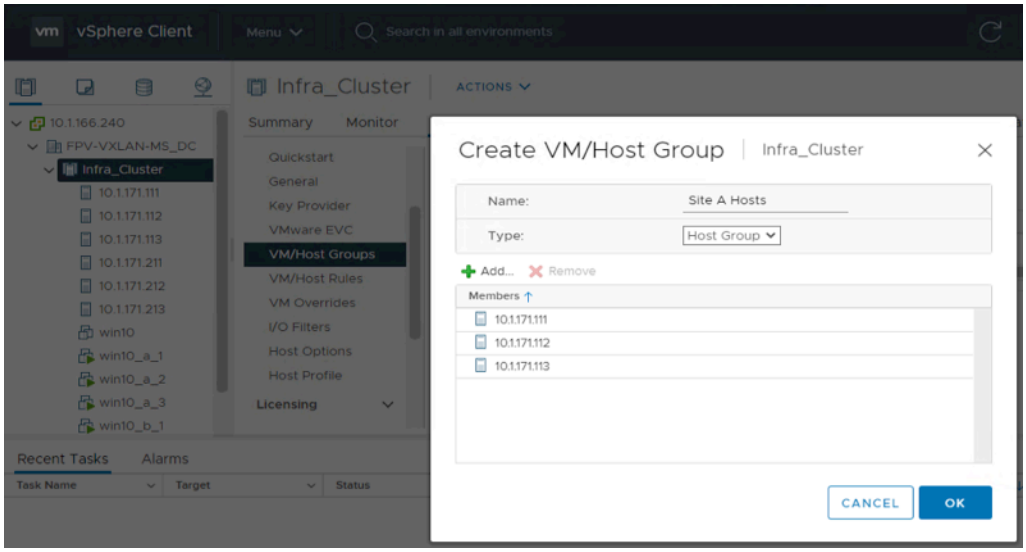
For performance considerations, it is optimal to have virtual machines using storage from their local storage cluster to reduce disk I/O latencies. To protect data and achieve zero recovery point objective with no data loss, the MetroCluster solution performs synchronous replication of data written to storage across sites. As a result, for a write operation, the latency will include the time needed for the data to be transferred between the MetroCluster sites so it can be mirrored remotely. On the other hand, for a read operation, the data can be returned directly by the local storage cluster without the distance latency when virtual machines are configured to use only local storage.

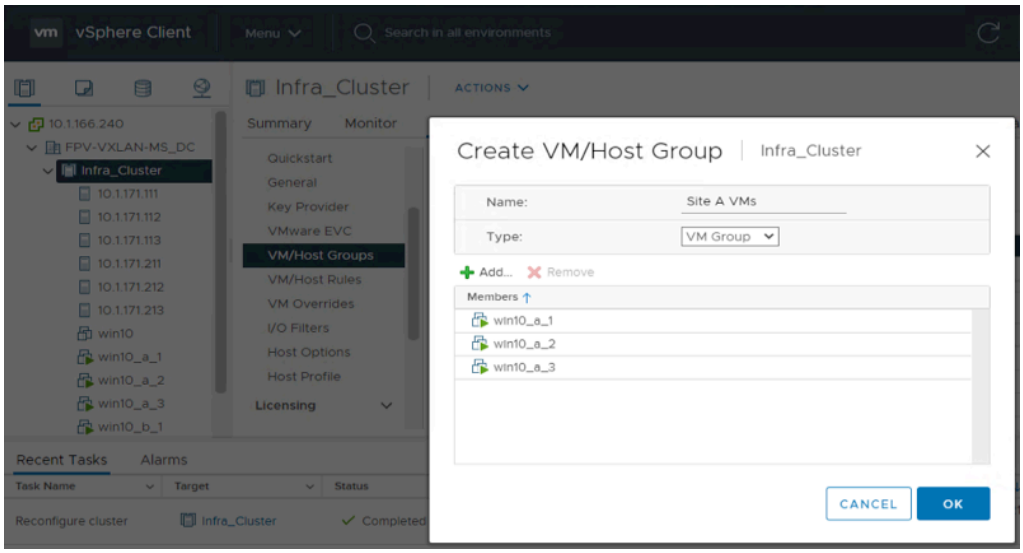
VMware vMotion and VM-Host Affinity Test

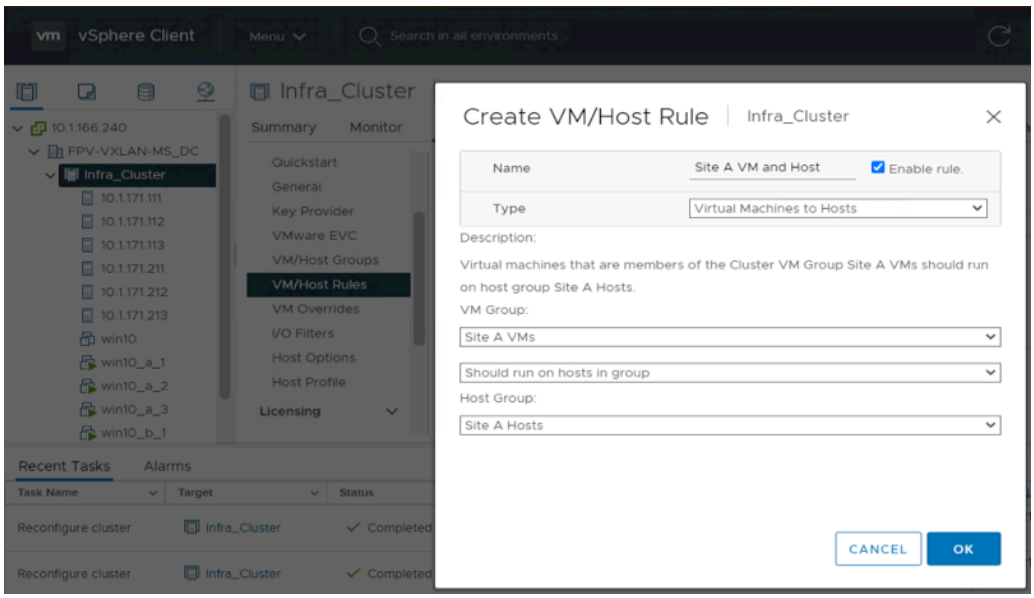
To enable virtual machines to run on any ESXi host at both of the MetroCluster sites, all ESXi hosts must mount the iSCSI and NFS datastores from both sites. If the datastores from both sites are properly mounted by all ESXi hosts, you can migrate a virtual machine between hosts with vMotion and still maintain access to all its virtual disks created from those datastores.

For a virtual machine that uses local datastores, its access to virtual disks becomes remote if it is migrated to a host at the remote site and thus increasing read operation latency due to distance. Therefore, it is a best practice to keep virtual machines on the hosts that have local access to the storage it uses. By using VM-Host affinity groups, you can create a virtual-machine group and a host group for virtual machines and hosts located at a particular site. If you use a VM-Host affinity rule, you can specify the virtual machines at one site to run on the host group at that site. To allow virtual-machine migration across sites during a site maintenance or disaster scenario, use the “Should run on hosts in group” specification for the rule to have that flexibility.

The following screenshots show the Site-A host group, the Site-A virtual-machine group, and the Site-A virtual-machine to host affinity rule, respectively.







Tests of vMotion of virtual machines to hosts at the same site as well as across sites were performed and were successful. After manually migrating a virtual machine across sites, the VM-Host affinity rule activates and migrates the virtual machine back to the group where it should run.

Additional features and functions of a FlexPod configuration available in a single-site VXLAN deployment were not repeated as part of this effort. For information and documentation, please refer to the FlexPod Datacenter with VXLAN Single-Site Cisco Validated Design Program design guide and deployment guide. Some of those value propositions include:

- Simplified host hardware replacement with service profile migration
- Simplified VMware integration with NetApp Virtual Storage Console and NetApp NFS Plug-in for VMware VAAI
- Virtual-machine consistent backup and recovery with NetApp SnapCenter Plug-in for VMware

High Availability and Disaster Recovery

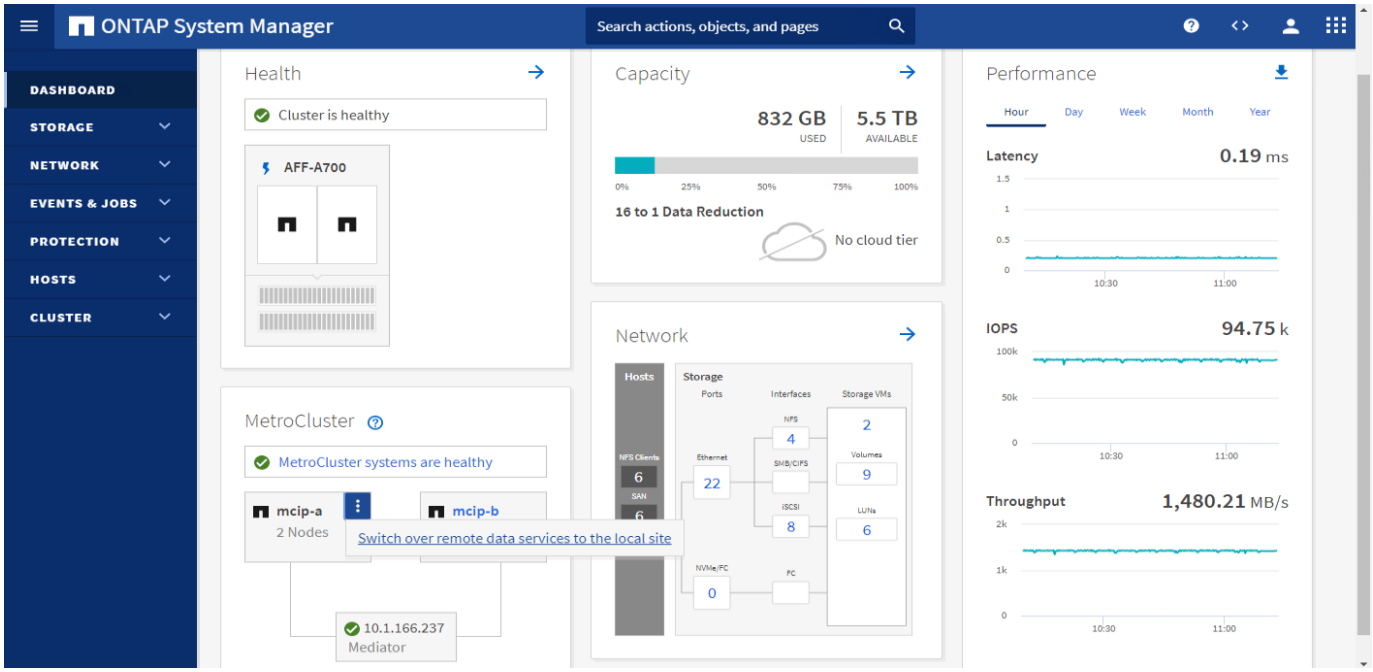
The FlexPod MetroCluster IP solution protects data services for a variety of single-point-of-failure scenarios as well as for a site disaster. The redundant design implemented at each site provides high availability, and the MetroCluster IP implementation with synchronous data replication across sites protects data services from a sitewide disaster of one site. To demonstrate the high availability and disaster recovery capabilities, the following scenarios are tested for this white paper:

- Planned negotiated MetroCluster switchover and switchback for disaster recovery testing and site maintenance
- Unplanned MetroCluster switchover and negotiated switchback to handle simulated sitewide disaster and recovery
- MetroCluster connectivity failures

Planned Negotiated MetroCluster Switchover and Switchback for Disaster Recovery Testing and Site Maintenance

Planned negotiated switchover and switchback operations should be performed on the solution after initial deployment to determine whether the solution was properly deployed. The testing can help identify connectivity and/or configuration problems that could lead to I/O disruptions. Testing and resolving any connectivity or configuration problems regularly helps ensure uninterrupted data services when a real site disaster occurs.

To initiate a switchover operation, you can issue a command from the ONTAP cluster shell or request to switch over remote data services to the local site from the ONTAP System Manager Dashboard MetroCluster pane, as shown in the following screenshot:



After a switchover operation is requested, the storage I/O at the remote site pauses briefly and then the remote data services resume at the local site. Afterwards, you can perform the required maintenance at the remote site. After testing or maintenance is completed for the remote site, you can return the local data services that belong to the remote site with the switchback operation from either the ONTAP System Manager GUI or from an ONTAP cluster shell.

You can perform a variety of storage maintenance without needing a planned switchover operation if the maintenance at one site is performed one controller at a time. An example is the storage controller firmware upgrade. The nondisruptive upgrade process performs firmware upgrade for one controller at a time. While a controller is being upgraded, its data services are failed over to its local partner. Other example maintenance scenarios include shutting down a controller to add an adapter for additional network connectivity or replacing a failed adapter. You can perform these types of maintenance tasks on one controller at a time with local storage failover without affecting data services. They do not require a switchover operation to take place.

Unplanned MetroCluster Switchover and Negotiated Switchback to Handle Simulated Sitewide Disaster and Recovery

An unplanned MetroCluster switchover can occur during a disaster simulation or when a real disaster happens; for example, one site experiences power outage due to a hurricane moving through the area. For this scenario, no planned invocation of the switchover operation occurs. Instead, the ONTAP Mediator, which monitors the MetroCluster IP solution from a third site, will detect a site failure condition and enable the MetroCluster IP solution to perform an automated unplanned switchover (refer to Figure 41).

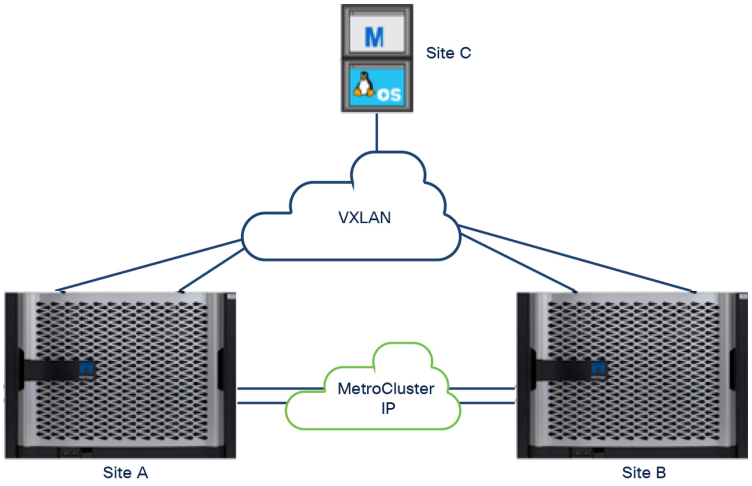


Figure 41.
ONTAP Mediator Monitors the MetroCluster IP Solution from a Third Site

Thanks to the highly available design for a FlexPod solution, redundancies are built into the solution to avoid single-point-of-failure scenarios. As a result, the un-planned switchover happens only when both ONTAP storage controllers at a site experience failure. If only one of the controllers fails, storage failover happens, and the surviving controller provides the local data services for the site.

To simulate a site disaster scenario, both storage controllers at one site were powered off by using the service processor commands, resulting in the unexpected power loss to the storage cluster and a sudden stop of data services at one site. Under such circumstances, the ONTAP Mediator that is running at a third site detects the failure and an unplanned switchover happens for the surviving site to start providing data services on behalf of the failed site. From an application perspective, the data services paused briefly and then the data services resumed normally.

This behavior was verified when the IOMeter I/O to the local datastore experienced a brief pause after the local cluster was powered off, and I/O resumed shortly afterwards. Checking on the surviving cluster, it did a switchover operation and was serving I/O on behalf of the cluster that was powered off. After power was returned to the controllers and the controllers activated, a manual switchback operation was initiated to return the data services from the remote site back to the local storage cluster.

MetroCluster IP Connectivity and Switch Failures

A variety of potential connectivity or switch failure and switch maintenance events could affect the MetroCluster IP solution operations. The following lists the scenarios validated for this white paper (refer to Figure 42):

- Connectivity of a MetroCluster IP interface to a MetroCluster IP switch failed.
- Connectivity of one of the links between a MetroCluster IP switch pair between sites failed.
- Connectivity of both links between a MetroCluster IP switch pair between sites failed.
- Connectivity of all links between both MetroCluster IP switch pairs failed.
- A single MetroCluster IP switch failure or reboot led to multiple connection failures

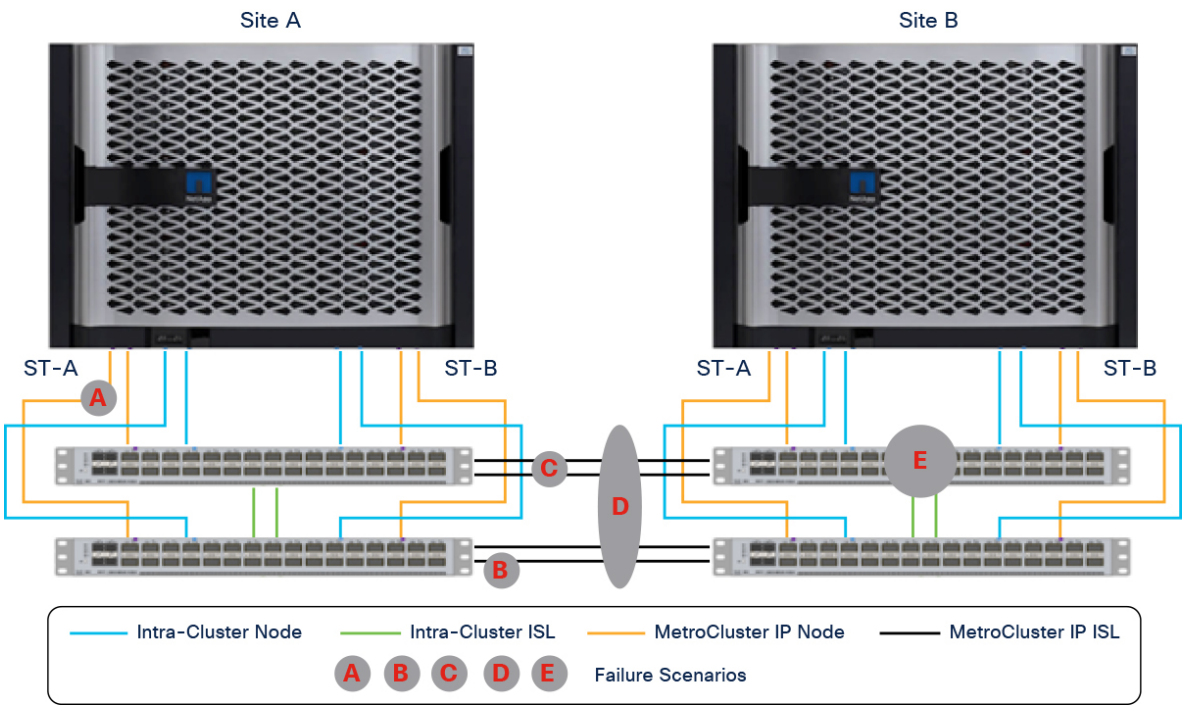
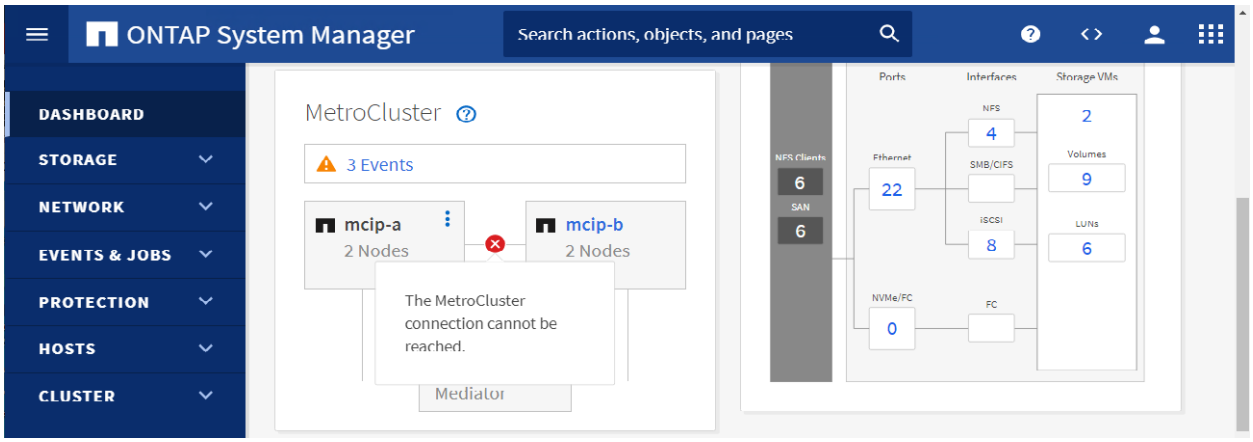


Figure 42.
MetroCluster IP Connectivity and Switch Failure Scenarios

For all the scenarios discussed previously, I/O continues in the virtual machines. However, some slight differences in their affects for the various test scenarios exist. For scenarios A, B, C, and E in the figure, despite the overall bandwidth impact as a result of the failures, the communication between site A and site B continues to be available thanks to the redundant solution design. As a result, the MetroCluster IP synchronous data replication continues despite the failure, and the data aggregates are still mirrored and in sync. When the failure affects only one link or one MetroCluster IP fabric, the MetroCluster IP functions will continue with potentially reduced bandwidth because of the connectivity or switch failure.

For the failure scenario D, the entire MetroCluster IP site-to-site communication was affected when all links between the two sites were down. The synchronous data replication between the sites was not possible as a result. Even though I/O continued at each site, the data is protected only locally. The ONTAP System Manager dashboard indicated MetroCluster connection could not be reached and the RAID status of the mirrored data aggregates was moved from normal to degraded state. The following Dashboard screenshot shows that the MetroCluster connection could not be reached when all ISL links were severed.



The following ONTAP data aggregate status output shows degraded RAID status when synchronous data replication between sites is not possible.

```
mcip-a::> metrocluster show

Configuration: IP-fabric

Cluster                       Entry Name             State
----------------------------- ---------------------- ---------------------
Local: mcip-a
                                Configuration State  configured
                                              Mode   normal
                                AUSO Failure Domain  auso-on-cluster-disaster
Remote: mcip-b
                                Configuration State  configured
                                              Mode   normal
                                AUSO Failure Domain  auso-on-cluster-disaster

mcip-a::> aggr show

Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
aggr0_mcip_a_01
            1.04TB   92.42GB   91% online       1 mcip-a-01        raid_dp,
                                                                    mirror
                                                                    degraded
aggr0_mcip_a_02
            1.06TB   116.6GB   09% online       1 mcip-a-02        raid_dp,
                                                                    mirror
                                                                    degraded
data_mcip_a_01
            2.80TB    2.75TB    2% online       5 mcip-a-01        raid_dp,
                                                                    mirror
                                                                    degraded
data_mcip_a_02
            2.80TB    2.78TB    1% online       6 mcip-a-02        raid_dp,
                                                                    mirror
                                                                    degraded
4 entries were displayed.

mcip-a::>
```

When the links between site A and site B were restored after the failure scenario was removed, data synchronization resumed, and the RAID status was restored to normal. Please note that the amount of time the RAID status stayed in the degraded state after the links were restored depended on the time duration of the severed links. It will take longer to synchronize the data when the failure duration is longer.

Network Failures

With the FlexPod MetroCluster IP solution connected to the Multi-Site VXLAN network at both sites, there are a lot of network connections and switches. To provide a highly available FlexPod solution, redundancy was built into the design with redundant components and cables for the solution. The network failure scenarios validated for this paper follow (refer to Figure 43):

- Partial network failure at one site with a single uplink failure on one fabric interconnect
- Partial network failure at one site with dual uplink failures on one fabric interconnect

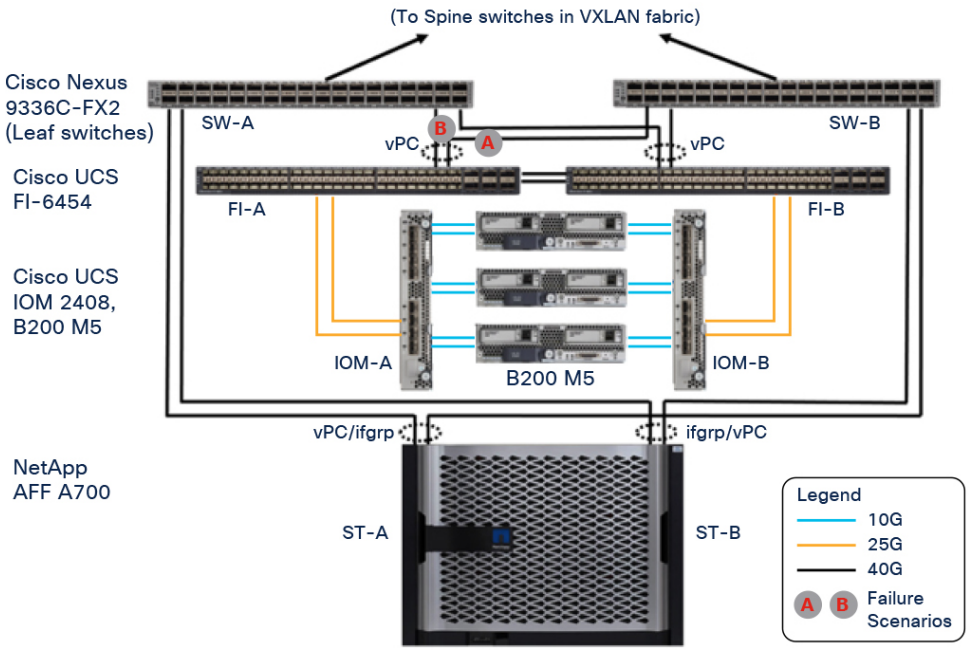


Figure 43.
Network Failure Scenarios for the Fabric Interconnect to Network Fabric Uplink Connections

Each fabric interconnect has two uplinks to the network, and they are connected to two switches for redundancy. When a single uplink failed (failure was created by disabling an uplink using Cisco UCS Manager), the affected fabric interconnect could still reach the network fabric from the remaining uplink. However, when both uplinks of a fabric interconnect were disabled, traffic from the servers had to go through the other fabric interconnect to reach the network fabric. No problems were observed for IOMeter I/O during the testing of either of the two test scenarios.

Many more potential connectivity or switch failure scenarios are possible within the Multi-Site VXLAN fabric. If the failure represents a single point of failure, the design of the Multi-Site VXLAN fabric would have no problem accommodating the failure. As a result, those failures are not specifically called out or validated.

Compute Failures

With a stretched VMware cluster implementation having three servers at each site, the solution can accommodate various planned and unplanned compute failures. Here are the scenarios validated in testing:

- Planned server reboot or maintenance by evacuating virtual machines with vMotion first
- Unplanned server failure

For a planned server maintenance or reboot, virtual machines on the server that will be serviced can be live migrated to other servers at the same site to allow local storage access. This scenario is the same as the already-validated vMotion scenario.

To simulate an unplanned server failure scenario, the Cisco UCS Manager was used to reset one of the servers using the power cycle method without a graceful shutdown. Because VMware HA is enabled for the VMware cluster, the virtual machines originally running on the affected server will be restarted on another server automatically. The configured VM-Host site affinity rule will help ensure that the affected virtual machines are restarted on servers within the same Host affinity group/site as the affected server.

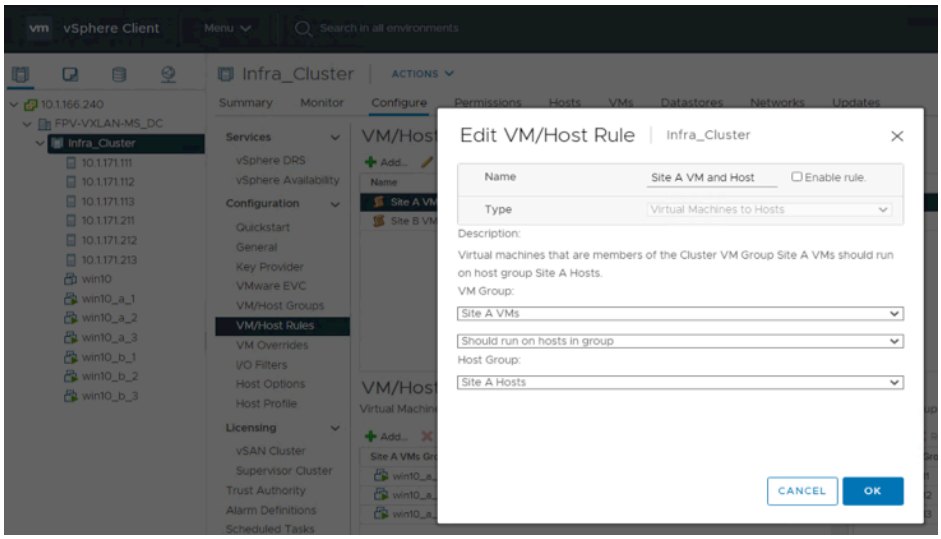
Complete Site Maintenance or Site Failure

A site may need site maintenance, experience power loss, or may possibly be affected by a natural disaster such as a hurricane. Therefore, it is crucial that you exercise planned and unplanned site failure scenarios to help ensure that your FlexPod MetroCluster IP solution is properly configured to survive such failures. The following site-related scenarios were validated during the testing.

- Planned site maintenance scenario by migrating virtual machines offsite and performing a MetroCluster switchover
- Unplanned site outage scenario by powering off servers and storage controllers for disaster simulation

To get a site ready for a planned site maintenance, a combination of migrating virtual machines off the site that would be affected with vMotion and negotiating a MetroCluster switchover are needed to migrate virtual machines and data services to the alternative site. Testing was performed in two different orders, vMotion first followed by MetroCluster switchover and MetroCluster switchover first followed by vMotion, to confirm that virtual machines continue to run, and data services are not interrupted.

Before performing the planned migration, you should disable the VMware cluster “run VMs on hosts” VM/Host site affinity rule for the site that will go down so the affected virtual machines won’t be automatically migrated back to their original site after they are manually migrated to the other site. In the testing, virtual machines were successfully migrated to the other site and the data services continued without problems. After virtual machines and storage services have been migrated, you can power off servers, storage controllers and disk shelves, and switches and perform site maintenance. The following screenshot illustrates the VM/Host site affinity rule disabled for the affected site before migrating virtual machines off for planned maintenance.



When site maintenance is completed, you should re-enable the “run VMs on hosts” VM/Host site affinity rule for the site that was disabled so virtual machines can return to their original site when the servers are booted back up. You also should initiate a MetroCluster switchback operation so the data services can be returned to their original site as well.

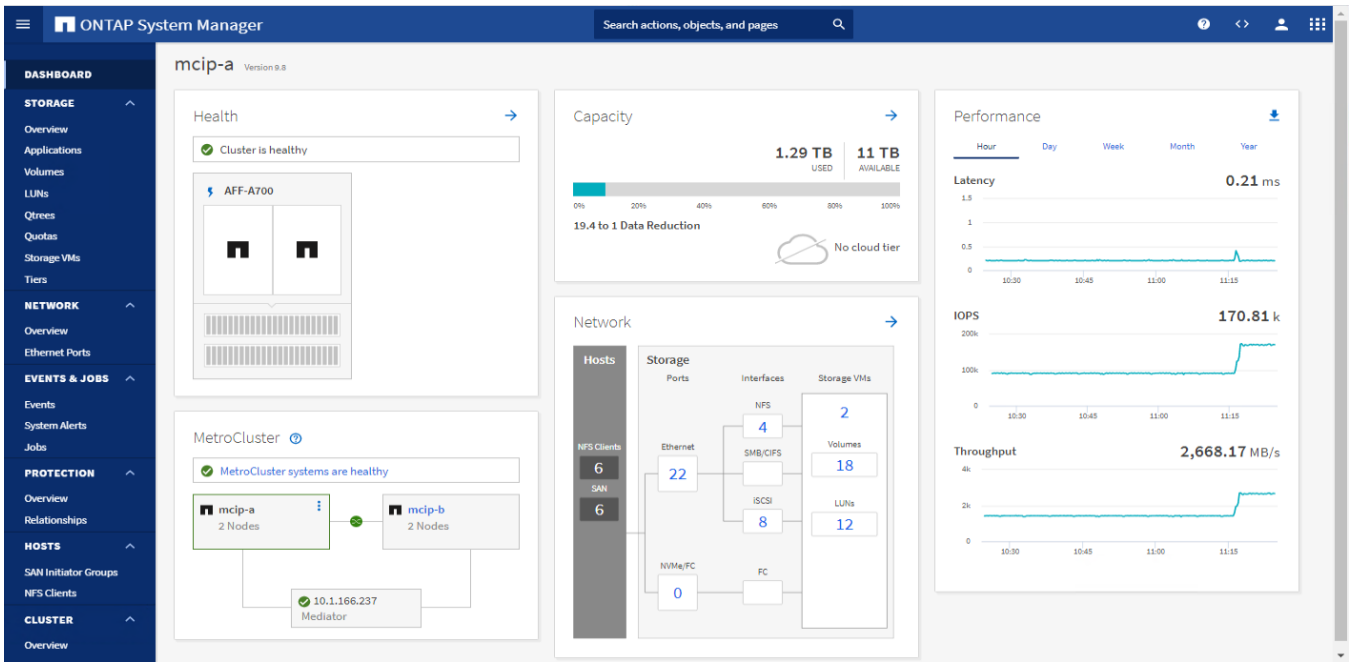
For the unplanned site disaster simulation, the servers and storage controllers were powered off to simulate a site disaster. The VMware HA feature detects the downed virtual machines and restarts those virtual machines on the surviving site. In addition, the ONTAP Mediator running at a third site detects the

MetroCluster IP site failure and the surviving MetroCluster IP site initiates a switchover and starts providing data services for the down site.

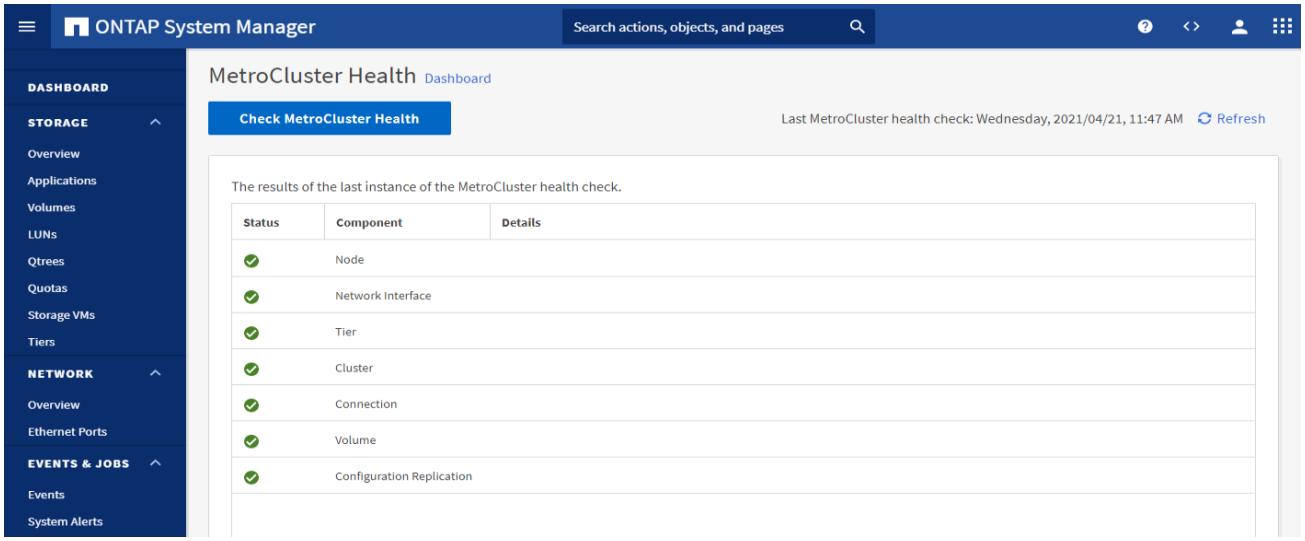
Storage Monitoring

ONTAP System Manager

ONTAP System Manager is a simple and versatile product that enables you to easily configure and manage ONTAP clusters. System Manager simplifies common storage tasks such as creating volumes, LUNs, qtrees, shares, and exports, all of which save time and help prevent errors. It also provides dashboards for you to quickly examine the health of the ONTAP cluster and critical metrics such as capacity usage and performance indicators such as latency, IOPs, and throughput. For the MetroCluster IP solution, a MetroCluster pane on the dashboard shows MetroCluster health, connectivity, and Mediator information. The following screenshot shows the ONTAP System Manager Dashboard, which provides cluster health and important metrics.



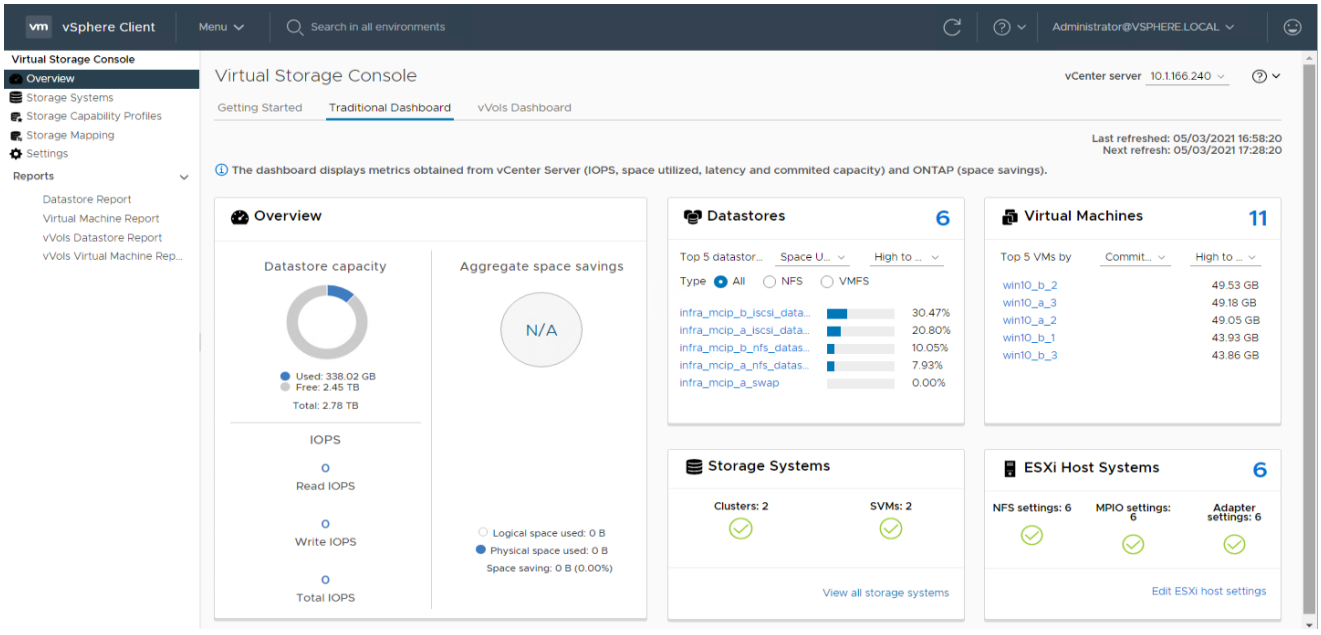
When you click the *MetroCluster Systems are healthy* link in the ONTAP System Manager MetroCluster pane, you can see additional details of the MetroCluster health status. Also, you can check its health status on demand by clicking the *Check MetroCluster Health* button on the MetroCluster Health screen. The results of the health check are summarized for the various components. The following screenshot shows the MetroCluster Health dashboard.



Virtual Storage Console

NetApp Virtual Storage Console (VSC) for VMware vSphere is a vSphere client plug-in that is integrated with VMware vCenter to provide end-to-end lifecycle management for virtual machines in VMware environments that use NetApp storage systems. VSC provides visibility into the NetApp ONTAP storage environment from within the vSphere web client. VMware administrators can easily perform tasks that improve both server and storage efficiency while still using RBAC to define the operations that administrators can perform.

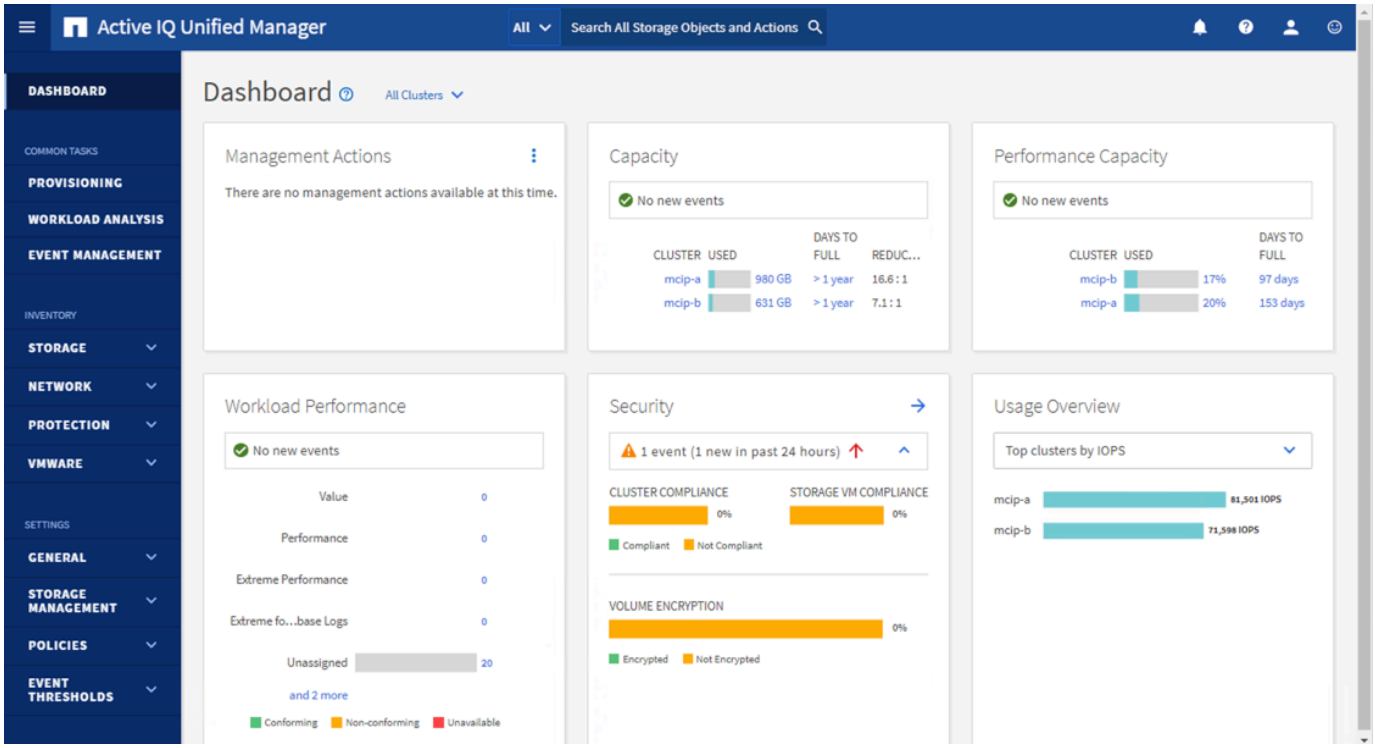
VSC uses NetApp technologies to deliver comprehensive, centralized management of ONTAP storage operations in both SAN and network attached storage (NAS)-based VMware infrastructures. These operations include discovery, health, capacity monitoring, and datastore provisioning. VSC delivers tighter integration between storage and virtual environments, greatly simplifies virtualized storage management, and helps deliver excellent performance from virtualized storage environments. After it is installed and configured, VSC provides a view of the storage environment from a VMware administrator's perspective and optimizes storage and host configurations for use with NetApp ONTAP storage systems. The following screenshot shows the VSC dashboard view, which displays important indicators.



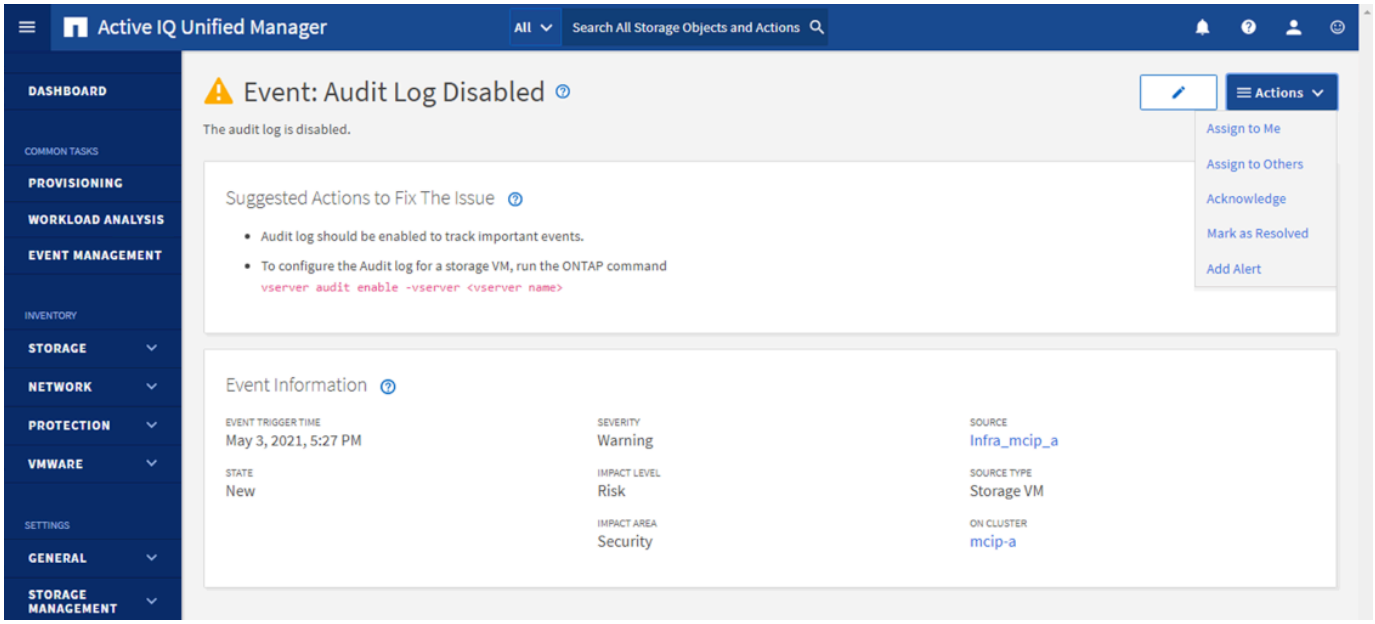
Active IQ Unified Manager

NetApp Active IQ Unified Manager is a comprehensive monitoring and proactive management tool for NetApp ONTAP systems to help manage the availability, capacity, protection, and performance risks of your storage systems and virtual infrastructure. You can deploy Unified Manager on a Linux server, on a Windows server, or as a virtual appliance on a VMware host.

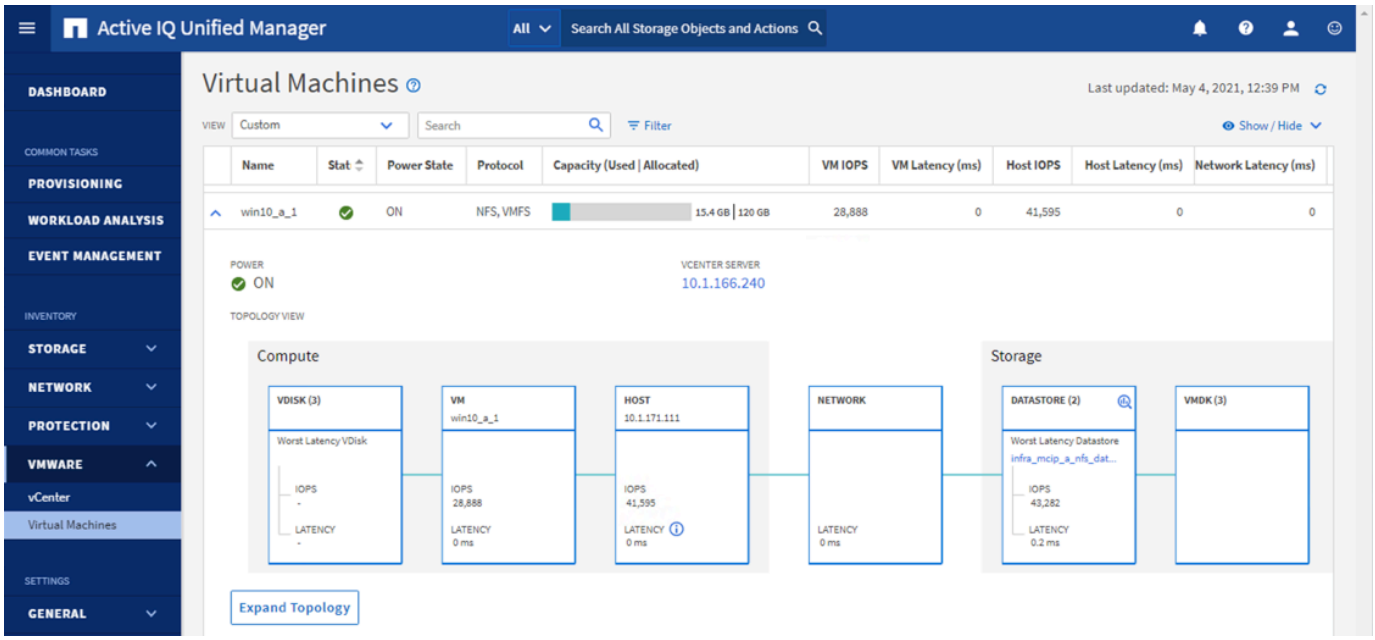
Active IQ Unified Manager enables you to monitor your ONTAP storage clusters, VMware vCenter server, and virtual machines from a single redesigned, intuitive interface that delivers intelligence from community wisdom and artificial intelligence (AI) analytics. It provides comprehensive operational, performance, and proactive insights into the storage environment and the virtual machines running on it. When a problem occurs on the storage or virtual infrastructure, Unified Manager notifies you about the details of the problem to help with identifying its root cause. It provides additional details for the events along with the suggested actions to resolve the problem. The following screenshot shows the Active IQ Unified Manager dashboard, which provides usage, capacity, performance, and security insights.



The following screenshot shows the Active IQ Unified Manager, which provides event details and suggested actions to resolve the problem:



The virtual-machine dashboard gives you a view into the performance statistics for the virtual machines that you can investigate the entire I/O path from the vSphere host down through the network and finally to the storage. Some events also provide remedial actions that you can take to rectify the problem. You can configure custom alerts for events so that when problems occur, you are notified through email and Simple Network Management Protocol (SNMP) traps. The following screenshot shows the Active IQ Unified Manger virtual-machines dashboard:



Refer to Active IQ Unified Manager Documentation Resources for more information about Active IQ Unified Manager.

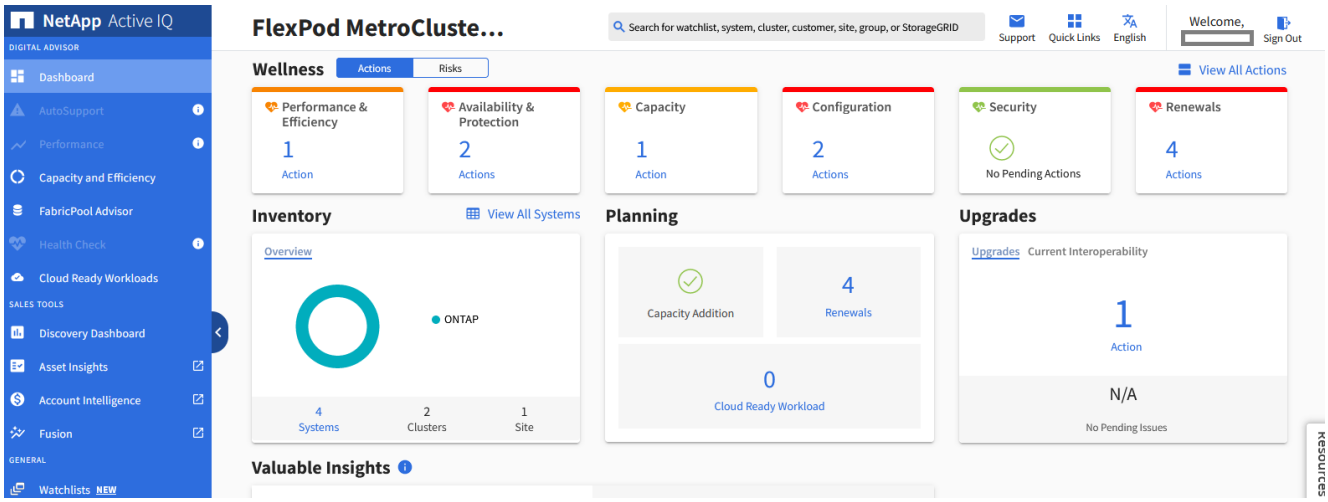
Active IQ

NetApp Active IQ is a cloud service that provides proactive care and optimization of your NetApp environment, leading to reduced risk and higher availability. Active IQ uses community wisdom and AIOps artificial intelligence to provide proactive recommendations and risk identification. The latest release of

Active IQ offers an enhanced user interface and a personalized experience with Active IQ Digital Advisor dashboards. It allows smooth and seamless navigation with its intuitiveness throughout different dashboards, widgets, and screens. It provides insights that help you detect and validate important relationships and meaningful differences based on the data that different dashboards present.

Watchlists are a way to organize a group of systems inside Active IQ Digital Advisor and create custom dashboards based on the system grouping. They provide quick access to only the group of storage systems you want, without having to sort or filter those you don't want. You can create a watchlist by using the storage system serial numbers to start a new dashboard.

The following screenshot shows the Active IQ dashboard, which provides an at-a-glance view into your systems and actions you can take to improve system wellness.



Refer to Active IQ Digital Advisor documentation resources for more information about Active IQ.

Conclusion

The FlexPod MetroCluster IP Datacenter with VXLAN Multi-Site fabric solution uses an active-active data center design to provide business continuity and disaster recovery. The solution interconnects two data centers deployed in separate, geographically dispersed locations. The NetApp MetroCluster IP solution uses synchronous replication to protect business-critical data services against site failure to achieve zero recovery point and low recovery time objectives. The data center sites can be up to 700 km apart if the network performance characteristics meet the requirements. The NetApp ONTAP Mediator and VMware vCenter deployed at a third site help you monitor the MetroCluster IP operations and manage the stretched highly available VMware cluster solution. Cisco Intersight enables the Cisco UCS compute in two data center sites to be centrally managed from the cloud. Redhat Ansible (or Hashicorp Terraform) automation can be leveraged to further simplify and accelerate the deployment of infrastructure, applications and services in Enterprise data centers. For VXLAN fabrics, Cisco DCNM Fabric Builder can also be used to automate the build out of the data center fabrics in the two sites. Additional monitoring tools are available from Cisco and NetApp so you can easily monitor the solution and gain insights on its operations. The flexibility and scalability of FlexPod enables you to start out with a right-sized infrastructure that can grow and evolve as your business requirements change. This validated design, which peers two FlexPod together with NetApp MetroCluster IP and uses a VMware stretched cluster with high-availability features, enables you to reliably deploy VMware vSphere-based private cloud on a distributed and integrated infrastructure, thereby delivering a solution that is resilient to many single-point-of-failure scenarios as well as a site failure to protect critical business services.

Appendix – References

FlexPod

1. FlexPod Datacenter with VMware vSphere 7.0, Cisco VXLAN Single-Site Fabric, and NetApp ONTAP 9.7 Design Guide
2. FlexPod Datacenter with VMware vSphere 7.0, Cisco VXLAN Single-Site Fabric, and NetApp ONTAP 9.7 Deployment Guide
3. FlexPod Datacenter with VMware vSphere 7.0 Design Guide
4. FlexPod Datacenter with VMware vSphere 7.0 and NetApp ONTAP 9.7 Deployment Guide
5. FlexPod Datacenter with Cisco ACI Multi-Pod, NetApp MetroCluster IP, and VMware vSphere 6.7 Design Guide
6. FlexPod Datacenter with Cisco ACI Multi-Pod with NetApp MetroCluster IP and VMware vSphere 6.7 Deployment Guide
7. FlexPod MetroCluster IP solutions with ONTAP 9.7 and compliant switches

NetApp Storage

1. Install a MetroCluster IP Configuration: ONTAP MetroCluster
2. MetroCluster IP Reference Configuration Files
3. NetApp Hardware Universe
4. ONTAP 9 Release Notes
5. ONTAP System Manager: Manage MetroCluster Sites
6. Active IQ Digital Advisor documentation resources
7. Active IQ Unified Manager documentation resources
8. TR-4883: FlexPod Datacenter with ONTAP 9.8, ONTAP Storage Connector for Cisco Intersight, and Cisco Intersight Managed Mode
9. TR-4689: MetroCluster IP Solution Architecture and Design

Cisco VXLAN BGP EVPN Multi-Site Fabric

1. VXLAN EVPN Multi-Site Design and Deployment White Paper
2. draft-sharma-multi-site-evpn - Multi-site EVPN based VXLAN using BGWs
3. RFC-7432 (BGP MPLS-based Ethernet VPN)
4. BRKDCN-2035 (VXLAN BGP EVPN-based multipod, multifabric, and multisite architecture)
5. BRKDCN-2125 (overlay management and visibility with VXLAN)
6. Cisco programmable fabric with VXLAN BGP EVPN configuration guide

VMware

1. VMware vSphere Metro Storage Cluster (vMSC)
2. VMware vSphere 5.x, 6.x and 7.x support with NetApp MetroCluster (2031038)

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA 222085.2 07/22