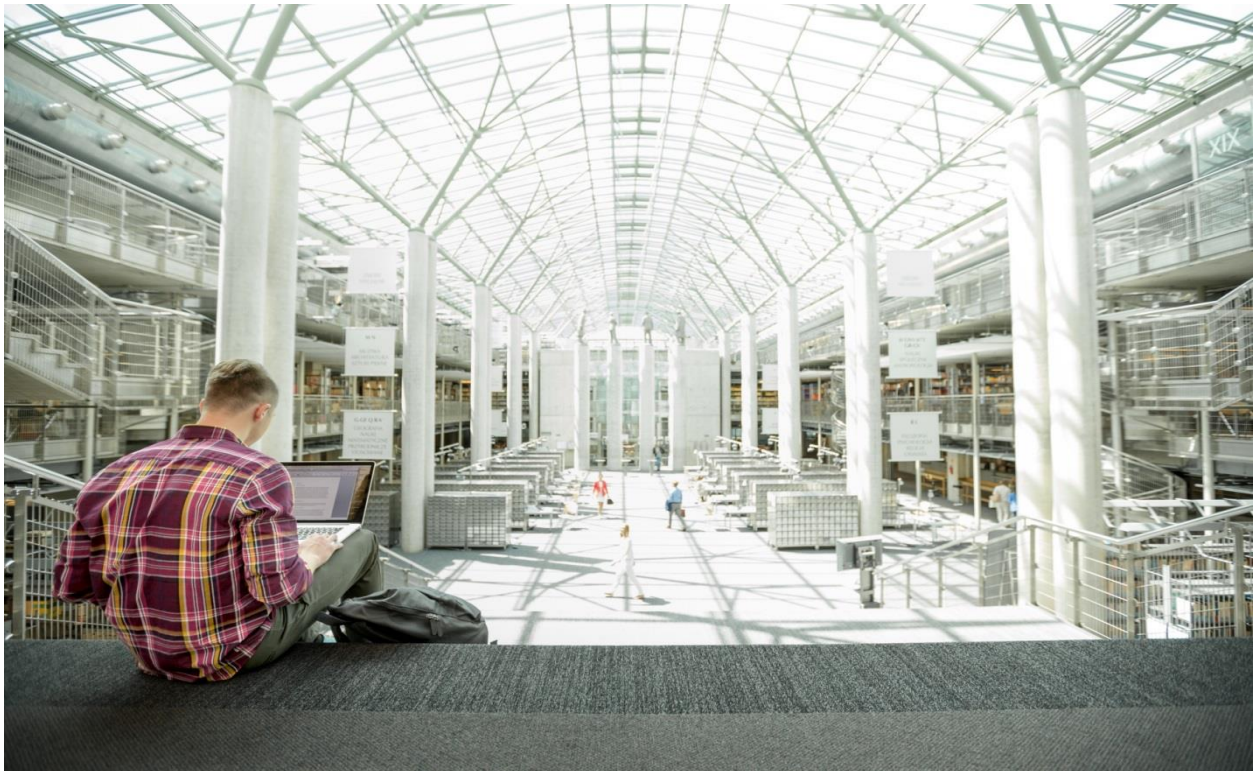


Cisco HyperFlex 3.5 Stretched Cluster with Cisco ACI 4.0 Multi-Pod Fabric Design Guide

Last Updated: September 3, 2019

CISCO VALIDATED DESIGN

About the Cisco Validated Design Program

The Cisco Validated Design (CVD) program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information, go to:

http://www.cisco.com/go/designzone.

Table of Contents

Executive Summary

Digital transformation in organizations is driving a need for application and data availability as an essential component of business success. Businesses are increasingly seeing a need for higher availability, including continuous 24x7 access to their data and applications. As a result, the data center infrastructure that hosts the data and applications are also seeing their uptime and availability requirements go up. The highly virtualized nature of today's data centers is also making the availability of the data center infrastructure more critical as the collective impact of a failure is much more catastrophic with the consolidation of multiple applications onto a single shared infrastructure.

Data center architectures typically address infrastructure availability by focusing on solutions and technologies that improve the reliability and robustness of individual components or systems within a single data center. Though high availability is crucial and necessary within a data center, it does not address data center-wide outages that are far more significant and crippling to a business. It is, therefore, essential to have a solution that spans data centers so that when a disaster takes down one data center, the data is still available from the other data center.

The **Cisco HyperFlex Stretched Cluster with Cisco ACI Multi-Pod Fabric** solution is a validated reference architecture for providing disaster avoidance and business continuity for the Virtualized Server Infrastructure (VSI) in an Enterprise data center. The solution uses an active-active data center design to maintain availability to at least one data center. Cisco HyperFlex stretched cluster provides the virtual server infrastructure in the solution. HyperFlex stretched cluster is a single cluster with nodes in the cluster distributed across both data centers. A Cisco Application Centric Infrastructure (ACI) Multi-Pod fabric provides the network fabric to interconnect the data centers. ACI Multi-Pod provides both Layer 2 extension and Layer 3 forwarding necessary for enabling application deployment in either data center with seamless connectivity and mobility. The active-active data centers can be in the same site such as different buildings in a campus environment or different sites across a metropolitan area. An ACI Multi-Pod fabric can interconnect sites separated by a distance of ~4000km for a maximum round-trip time (RTT) of 50ms while a Cisco HyperFlex stretched cluster can support a maximum RTT of 5ms or ~100km between sites. The stringent latency requirements for HyperFlex is due to application requirements, specifically read and write storage latencies required by Enterprises applications running on the cluster.

The HyperFlex stretched cluster serves as an Applications cluster in this design. Existing infrastructure outside the ACI fabric or a Management cluster within the fabric provides the management and services that HyperFlex clusters and applications hosted on the clusters require. All HyperFlex and UCS infrastructure in the data centers are centrally managed using Cisco Intersight. Cisco Intersight is a subscription-based, cloud service for infrastructure management that simplifies operations by providing pro-active, actionable intelligence for operations. It also provides capabilities such as Cisco Technical Assistance Center (TAC) integration for support and Cisco Hardware Compatibility List (HCL) integration for compliance checks. Cloud-based delivery also enables Enterprises to quickly adopt the new features that are continuously being rolled out in Cisco Intersight.

The **Cisco HyperFlex Stretched Cluster with Cisco ACI Multi-Pod Fabric** solution presented in this document is based on Cisco HyperFlex 3.5, Cisco UCS Manager 4.0, VMware vSphere 6.5, and Cisco ACI 4.0. This document is the *design guide* for the solution. The deployment guide for this solution is available at:
https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hx_35_vsi_aci_multipod.html

To understand the active-active data center design presented in this document, it is very important to first review and understand the single-site data center design based on HyperFlex VSI and Cisco ACI, which is available at:
https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hx_30_vsi_aci_32.html

Solution Overview

Introduction

The **Cisco HyperFlex Stretched Cluster with Cisco ACI Multi-Pod Fabric** solution presented in this document is a hyperconverged Virtual Server Infrastructure (VSI) solution for providing disaster avoidance and business continuity in VMware vSphere deployments. The solution uses Cisco HyperFlex stretched cluster attached to a Cisco Unified Computing System (Cisco UCS) and Cisco ACI Multi-Pod fabric to enable an active-active data center solution. The HyperFlex stretched cluster is extended across the active-active data centers to provide the VSI in each location, with ACI Multi-Pod fabric providing the seamless Layer 2 and Layer 3 connectivity between the locations.

The Cisco HyperFlex Stretched Cluster with Cisco ACI Multi-Pod Fabric solution is part of the HyperFlex VSI portfolio of solutions. For the complete portfolio of HyperFlex based solutions, see: https://www.cisco.com/c/en/us/solutions/design-zone/data-center-design-guides/data-center-hyperconverged-infrastructure.html

Audience

The audience for this document includes, but is not limited to; sales engineers, field consultants, professional services, IT managers, partner engineers, and customers that are interested in leveraging industry trends towards hyperconvergence and software-defined networking to build agile infrastructures that can be deployed in minutes and keep up with business demands.

Purpose of this Document

This document delivers an end-to-end VSI design for disaster avoidance in VMware vSphere deployments using Cisco Hyperflex Stretched Cluster for the hyperconverged infrastructure and Cisco ACI Multi-Pod fabric for the data center fabric. This document serves as the **design guide** for this solution.

What's New in this Release?

This solution introduces the Cisco HyperFlex Stretched Cluster and Cisco ACI Multi-Pod fabric to deliver a disaster avoidance solution for the data center infrastructure. The solution uses an active-active data center design to ensure availability to at least one data center in the event of a failure. The stretched cluster extends to both data centers and serves as an Application cluster in this design. Application cluster hosts workloads in either data center with access to the same networks and services from each location. Design also includes a HyperFlex standard cluster to host management services directly on the ACI fabric. Cisco Intersight centrally manages the Hyperflex and UCS infrastructure and deploys the standard cluster in this design.

The solution also uses updated components and versions to validate the design as outlined below:

- Cisco HyperFlex 3.5(2e), Cisco UCS Manager 4.0(2d), Cisco Intersight
- Cisco ACI 4.0(1h), Cisco AVE 2.0(1a) and VMware vDS 6.5.0
- VMware vSphere 6.5U2

Solution Summary

Cisco Validated Designs (CVDs) deliver systems and solutions to facilitate and accelerate customer deployments. CVDs incorporate a wide range of technologies, products, and best-practices into a portfolio of solutions that have been developed to address the business needs of our customers. For each CVD, the end-to-end design is built and validated in

the Cisco labs to ensure functionality and interoperability. The design and implementation details are then documented to provide a working template that customers can use to guide them in their data center rollouts.

The **Cisco HyperFlex Stretched Cluster with Cisco ACI Multi-Pod Fabric** solution presented in this document is a validated reference architecture for disaster avoidance and business continuity in an Enterprise data center. The solution uses an active-active data center design to ensure the availability of the Virtual Server Infrastructure (VSI) in the event of a disaster or a data center-wide failure. The solution uses the following family of infrastructure components for the compute, storage, networking and virtualization layers of the VSI in each data center.

- Cisco HyperFlex HX-series servers
- Cisco Unified Computing System (Cisco UCS)
- Cisco Application Centric Infrastructure (Cisco ACI) fabric
- Nexus 9000 family of switches (for ACI fabric and Inter-Pod Network)
- VMware vSphere

The solution uses a Cisco HyperFlex **stretched** cluster to provide the hyperconverged infrastructure for the active-active data centers. A HyperFlex stretched cluster is a single cluster that is extended across two data centers by evenly distributing the nodes in the cluster across both locations. The stretched cluster can be extended across data centers in a single location (for example, different buildings in a campus location) or different geographical locations. When there is a failure in one location, stretched clusters provide quick recovery by maintaining availability to the infrastructure in the second data center location. Stretched clusters ensure zero data loss by maintaining copies of the stored data in both locations. To meet the latency requirements of Enterprise applications hosted on the cluster, the maximum RTT and bandwidth of a stretched cluster must be <5ms (~100km) and require at least 10Gbps between sites.

The solution uses a Cisco ACI Multi-Pod fabric to provide the network fabric in each data center and to interconnect the two active-active datacenters. The ACI Multi-Pod fabric provides the Layer 2 extension and Layer 3 forwarding necessary for enabling the active-active data centers. The ACI Multi-Pod fabric in this design consists of two distinct ACI fabrics, one in each data center location and an **Inter-Pod Network (IPN)** that connects them. The HyperFlex stretched cluster nodes in a given site connect to the ACI fabric in the same site. The fabric in each site is referred to as a **Pod** in the ACI Multi-Pod architecture. Each Pod is deployed as a standard Spine-Leaf architecture (same as a single site fabric) and uses a highly-resilient design to access networks and services within the Pod as well as outside the Pod.

The solution uses 40GbE links for connectivity within the ACI fabric, and 10GbE for connectivity to APICs, IPN and networks outside the fabric. The connectivity to UCS domains and HyperFlex clusters use either 10GbE or 40GbE depending on the type of Fabric Interconnects used.

The design uses two HyperFlex clusters – a HyperFlex stretched cluster for **Application** workloads and an optional HyperFlex standard cluster for **Management**. Cisco APIC manages the virtual networking on both clusters by integrating with the Virtual Machine Manager (VMM) or VMware vCenter that manages the clusters. Management and Application clusters use VMware vDS and Cisco AVE respectively - both are APIC-controlled and managed.

Cisco Intersight manages all Cisco UCS and HyperFlex infrastructure in the solution. Cisco Intersight offers cloud-based, centralized management of Cisco UCS servers and HyperFlex nodes across all Enterprise locations and delivers unique capabilities such as:

- integration with Cisco TAC for support and case management
- proactive, actionable intelligence for issues and support based on telemetry data
- compliance check through integration with Cisco Hardware Compatibility List (HCL)
- centralized service profiles for policy-based configuration

The solution incorporates technology, design and product best practices and uses a highly resilient design across all layers of the solution. The solution was then built and verified in the Cisco labs using specific models of the different component

families (HyperFlex, Cisco UCS, ACI, VMware). The data centers were interconnected by a 75km fiber spool for validation. Table 1 lists the components used in each site. See Solution Validation section for more information.

Table 1 Solution Components per Pod

Infrastructure Domain	Component		Comments
	Pod 1	**Pod 2**	
Network (ACI MultiPod Fabric)	Cisco APIC M2 Server x 2	Cisco APIC M2 Server x 1	APIC Cluster (3-node)
	Cisco Nexus 9364C x 2	Cisco Nexus 9364C x 2	Spine Switches
	Cisco Nexus 93180YC-EX x 2	Cisco Nexus 93180YC-EX x 2	Leaf Switches – To Cisco UCS Domains
	Cisco Nexus 9372PX x 2	Cisco Nexus 9372PX x 2	Leaf Switches – Shared L3Out
	Cisco Nexus 93180YC-EX	Cisco Nexus 93180YC-EX x 2	IPN Routers
	Pod 1	**Pod 2**	
Hyperconverged Infrastructure (Cisco HyperFlex Standard & Stretched Clusters)	Cisco HX220C-M4S x 4	–	Management Cluster (4-node Standard Cluster)
	Cisco UCS 6248 FI x 2	–	
	Cisco HX220C-M5SX x 4	HX220C-M5SX x 4	Application Cluster (4-4 Stretch Cluster)
	Cisco UCS 6332 UP FI x 2	Cisco UCS 6332 UP FI x 2	
	Pod 1	**Pod 2**	
Virtualization Layer	VMware vSphere 6.5 U2 EP13	VMware vSphere 6.5 U2 EP13	Hypervisor
	vCenter Server Appliance 6.5 U2e	–	VCSA for Application Cluster; Management Cluster is managed by a VMware vCenter Server outside ACI Fabric
	VMware vDS, Cisco AVE	Cisco AVE	Virtual Switches – VMware vDS used in Management Cluster; Cisco AVE used in Application Cluster
Management & Monitoring	Cisco Intersight, Cisco UCS Manager, Cisco HyperFlex Connector, vCenter Plugins for HyperFlex and Cisco ACI		
Security	Cisco Umbrella (Cloud-based) using On-premise Virtual Appliances		https://umbrella.cisco.com

Solution Design

Cisco HyperFlex with Cisco ACI solutions bring infrastructure agility to Enterprise data centers by using an end-to-end software-defined architecture to deliver virtualized data center infrastructure. The **Cisco HyperFlex Stretched Cluster with Cisco ACI Multi-Pod Fabric** solution extends this architecture to deliver an active-active data center solution for disaster avoidance and business continuity. The solution uses a Cisco HyperFlex stretched cluster to extend the hyperconverged infrastructure across the two active-active data centers, and a Cisco ACI Multi-Pod fabric to extend the network fabric between the data center locations.

This active-active data center solution was designed to address the following key design goals.

- Disaster avoidance and business continuity in the event of a data center failure
- Design must provide access to networks and services directly from each data center location
- Ability to position workloads in either data center location while providing workload mobility
- Active management and distribution of workloads across both locations
- Quick recovery with zero data-loss
- Simplify the administration of a multi-data center environment

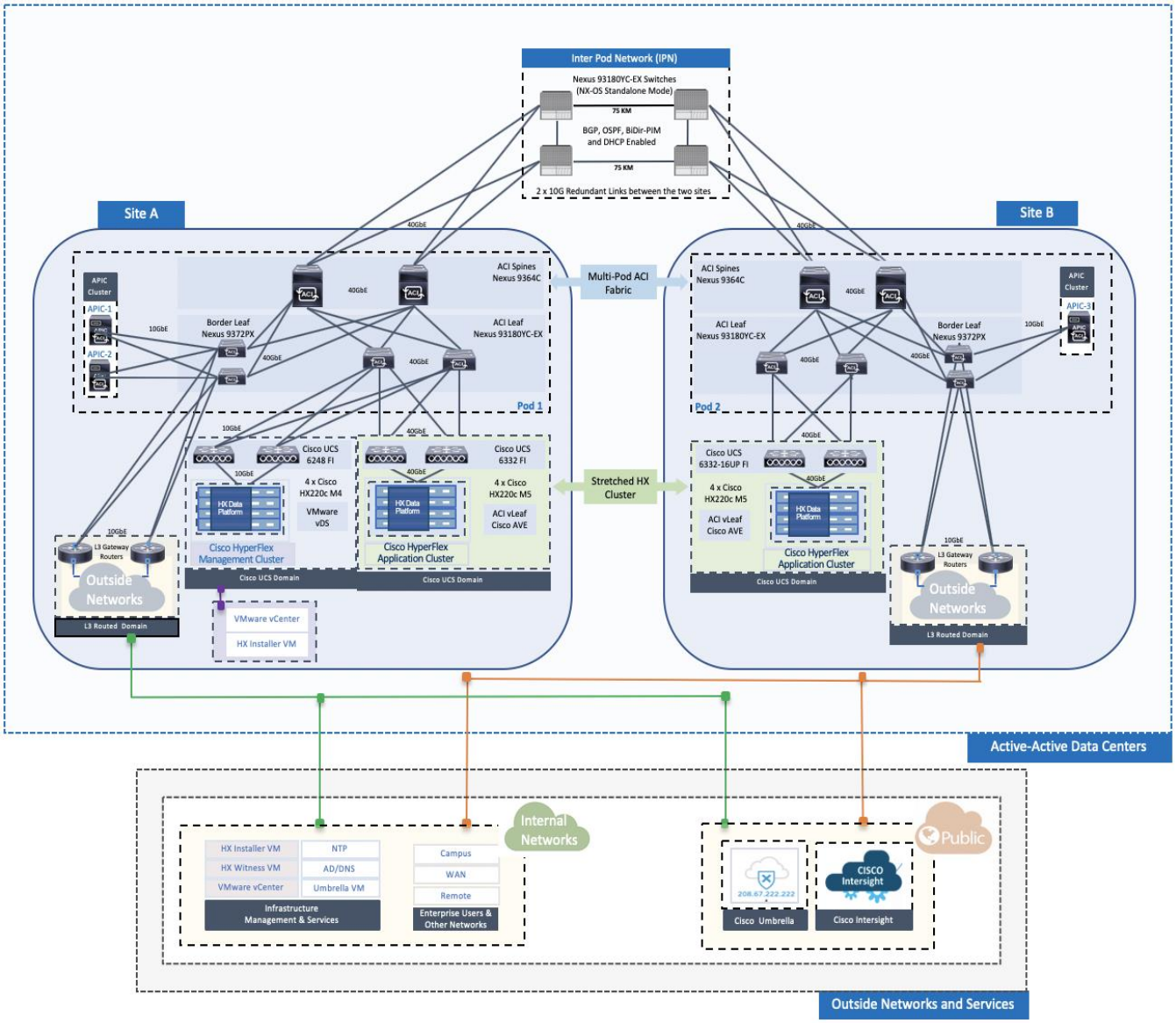
The virtual server infrastructure in the individual data centers was also designed to meet the same goals as single data center solution.

- 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 components, resources or sub-systems can be modified or upgraded to meet business needs
- Flexible design with design options for the different sub-systems in the solution, including the individual components used, storage configuration and connectivity options.
- Ability to automate and simplify by enabling integration with external automation and orchestration tools
- Incorporates technology and product-specific best practices for all components used in the solution

Topology

The end-to-end design for the active-active data center solution is shown in Figure 1.

Figure 1 High-level Design



Design Overview

A high-level overview of the design, including the capabilities and functionality provided in the solution are summarized below:

- Cisco HyperFlex brings hyperconvergence and software-defined infrastructure to the solution. Cisco HyperFlex provides the compute, storage and access networking for the virtual server infrastructure in the two active-active data centers.
- Two types of Hyperflex clusters are used in this design – a HyperFlex **stretched** cluster for hosting **Applications,** and a HyperFlex **standard** cluster for **Management**. The Management cluster is an optional part of this design.
- The HyperFlex **stretched** cluster provides the hyperconverged virtual server infrastructure for the active-active data centers. The stretched cluster is extended across the two data center locations to provide high-availability for the virtual server infrastructure in each data center. The data centers can be in a single site such as a campus

environment or in geographically separate sites such as a metropolitan area. To validate this design, the data centers are assumed to be in different geographical locations, separated by a distance of 75km.

- Cisco ACI brings software-defined networking (SDN) to the solution, with innovations and capabilities that go far beyond what traditional SDN solutions provide. Cisco ACI provides a policy-based, application-centric approach to networking that greatly simplifies the administration and rollout of applications and services. APICs provide centralized administration and management of the entire fabric which ensures configuration and policy consistency across all nodes in the fabric; there is no individual configuration of nodes in ACI.
- A Cisco ACI Multi-Pod Fabric provides the data center fabric for this multi-data center solution. A Multi-Pod fabric consists of distinct ACI fabrics or Pods interconnected by an Inter-Pod Network (IPN). Each fabric or Pod is essentially an independent, standalone ACI fabric, similar to a single-site ACI fabric. Two Pods are used in this design, one for each data center location. The HyperFlex nodes in a given datacenter location connect to the ACI fabric in that location. A pair of IPN routers in each location connects the two data centers in this solution. The ACI Multi-Pod fabric provides the Layer 2 extension and Layer 3 connectivity necessary to extend the HyperFlex stretch cluster across the active-active data centers.
- The ACI Multi-Pod fabric used in this design is managed by a single APIC cluster that greatly simplifies the administration of a multi-data center solution such as this. The APIC cluster consists of three nodes, two in first the data center and a third node in the second data center. The distribution of APIC nodes across the two active-active data centers ensures APIC availability in the event of a site failure. Additional APIC nodes can be added to the cluster for higher scale and availability.
- The optional **Management** cluster is used to host infrastructure management and application services as needed. The management cluster connects to the ACI fabric and serves as a starting point for deploying and managing other infrastructure connected to the same ACI Multi-Pod fabric. For example, the Management cluster hosts the HyperFlex installer that was used to install and deploy the HyperFlex stretched cluster in this design.
- Cisco Intersight is used to centrally manage the Cisco HyperFlex clusters and UCS domains in the different sites from the cloud. Cisco Intersight was also used to remotely install and deploy the Management cluster in this design. At this time, Cisco Intersight does not support the installation of HyperFlex stretched clusters.
- In this design, the infrastructure, network and application layer services necessary to deploy and manage active-active data center infrastructure and the applications hosted on it, are located in the Enterprise as well as in the Cloud. Within the Enterprise, the services can be either in the ACI fabric (for example, on the Management cluster) or in existing non-ACI infrastructure reachable through a Shared Layer 3 Outside connection (discussed below). In this design, services such as Microsoft Active Directory and DNS are located in the Enterprise, but outside the ACI fabric. Solution also uses cloud-based services such as Cisco Intersight and Cisco Umbrella, which are also accessed using the same Shared Layer 3 Outside connection.
- The multi-tenancy provided by the ACI fabric is leveraged in this design to create a separate tenant (**`HXV-Foundation` Tenant)** for providing HyperFlex infrastructure connectivity, separate from the tenants used for hosting applications on the HyperFlex infrastructure. ACI also uses additional tenants to provide other foundational connectivity and services within fabric. For example, system-defined tenants such as **infra, common,** and **mgmt** are used to implement, enable and manage the fabric as well as for other fabric-related functions. The ACI-defined **common** Tenant is defined by ACI to provide access to shared resources that multiple tenants require. The **`HXV-Foundation`** Tenant defined in this design can be used for any HyperFlex and UCS infrastructure in the ACI Multi-Pod fabric. An additional **Application** Tenant is also defined in this design to isolate the Applications from hyperconverged infrastructure-related functions. Customers can adapt the tenancy structure in this design as needed to meet their organizational needs. Once defined, application endpoints can be deployed in these tenants, from anywhere in the ACI Multi-Pod fabric without the needing to define the tenant or other ACI constructs (Bridge Domain, Application Profile, etc.) on a per Pod basis.
- The ACI system-defined **common** Tenant provides access to common services (application, infrastructure or network services) that other tenants in the fabric need access to. In this design, shared services provided by the ACI-defined **common** Tenant include connectivity to networks outside the ACI fabric such as an existing non-ACI

infrastructure where Active Directory, DNS etc. reside and to networks outside the Enterprise for accessing cloud-based services. Other shared services can also be easily made available through the **common** Tenant by hosting the shared service virtual machines on Management HyperFlex cluster in the ACI fabric. Shared services in the **common** Tenant, once defined, are also available and accessible from both datacenter fabrics without the need for any special Pod-specific configuration.

- Design also provides Layer 3 connectivity outside the ACI Multi-Pod fabric to existing, non-ACI datacenter infrastructure within the Enterprise and to Internet for cloud-based services. The Layer 3 access is enabled in ACI using a Shared Layer 3 Outside (**Shared L3Out**) connection. In this design, a Shared L3Out connection is defined for each data center location or Pod to ensure independent access to external services directly from each location. In this design, the services reachable through the Shared L3Out connection include NTP, DNS, Active Directory, Cisco Umbrella Virtual Appliances (on-prem), VMware vCenter (for Management cluster), HyperFlex Installer (for Management cluster) and HyperFlex Witness node. Cloud services accessible from each data center include Cisco Intersight and Cisco Umbrella in this design.
- In the virtualization layer, ACI integrates with the Virtual Machine Manager (VMware vCenter) to dynamically orchestrate and manage the virtual networking on a VMware vDS or Cisco ACI Virtualization Edge (AVE). Cisco AVE is a virtual Leaf (vLeaf) that extends ACI policies, security and other advanced capabilities to the virtualization layer. In this design, VMware vDS is used for the Management cluster and Cisco AVE for the Applications Cluster – both are controlled and managed by the APIC cluster.

The design also incorporates and aligns with the best practices for the technologies and products used in the solution, as well as general design best practices.

System Design

The detailed design for the different sub-systems that make up the overall solution is described in this section.

ACI Multi-Pod Fabric Design

The ACI Multi-Pod fabric provides the data center fabric for the active-active data centers in this design. The fabric must be in place before any hyperconverged infrastructure can be deployed in these data centers. The Multi-Pod fabric enables the HyperFlex stretched cluster to be extended between data centers to provide the virtual server infrastructure in each data center. The Multi-Pod fabric also enables application components to be deployed in either data center, with seamless mobility across data centers.

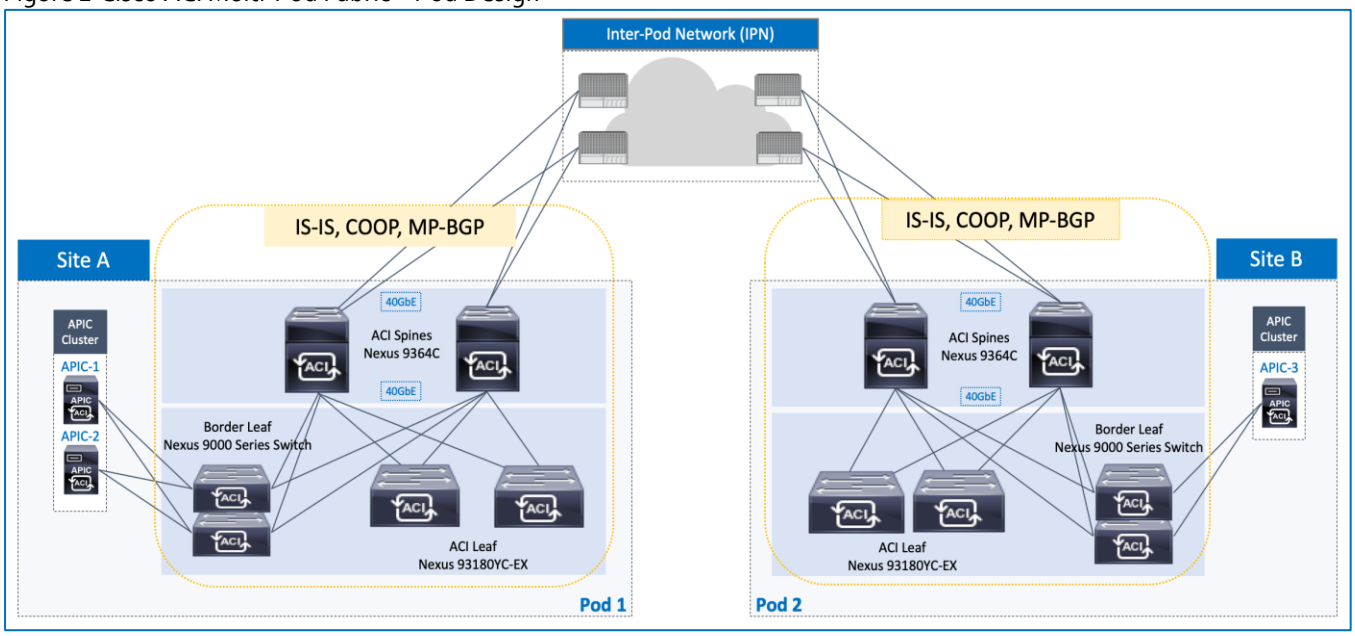
Pod Design

The Cisco ACI Multi-Pod fabric is designed to connect data centers. An ACI Multi-Pod fabric consists of distinct ACI fabrics or Pods interconnected by an Inter-Pod (IPN) network. The IPN in the ACI Multi-Pod fabric is not part of the ACI fabric, but it connects to each Pod through one or more Spine switches. The IPN design is covered in detail in the next section.

Each Pod uses a Spine-Leaf architecture, similar to a single-site ACI fabric. Pods also have independent control planes and run separate instances of the fabric protocols (IS-IS, COOP, MP-BGP). As a result, a control plane failure in one Pod does not impact or de-stabilize the control planes in the other Pods. Therefore, each Pod is a separate fault-domain from a fabric perspective. As of this writing, the latest ACI release supports up to 12 Pods in an ACI Multi-Pod fabric.

In this design, the ACI Multi-Pod fabric consists of two ACI fabrics or Pods, one in each data center, interconnected by an IPN. Each Pod is built using a spine-leaf architecture consisting of Cisco 9364C spine switches and Cisco 93180YC-EX leaf switches as shown in Figure 2. Redundant 40GbE links are used for connectivity between Spine and Leaf switches in each Pod.

Figure 2 Cisco ACI Multi-Pod Fabric – Pod Design



APIC Cluster Design

The ACI Multi-Pod fabric interconnects multiple ACI fabrics but it operates as single fabric from a management and operational perspective. A single APIC cluster manages the entire fabric and serves as a central point for management and policy definition. Once the fabric is setup, endpoints can be deployed anywhere in the fabric, on any Pod, without the need for any Pod-specific configuration. The ACI configuration for an endpoint group (EPG) is done once and it will apply to all Pods in the fabric. The seamless layer 2 extension and layer 3 reachability provided by an ACI Multi-Pod fabric make it possible for endpoints to be part of any endpoint group regardless of their location. For example, a Web Server farm hosting a company's website can have the individual web servers distributed across multiple Pods but still be part of the same EPG, with the same forwarding policies, connectivity etc. As a result, ACI Multi-Pod fabric greatly simplifies the management and operational aspects of an active-active data center solution.

To ensure APIC availability, the individual nodes in the APIC cluster are distributed across the ACI fabrics or Pods in the Multi-Pod fabric. The active-active data centers in this design uses a 3-node APIC cluster, with two APICs in Pod-1 or Site-A and one APIC in Site B or Pod-2 as shown in Figure 2. This allows each Pod to operate independently in the event of a connectivity issue between data centers or if a data center fails.

For resiliency, APIC clusters also use data sharding for the fabric configuration data it maintains. Data sharding splits the data into shards or units of data. The shards are then copied three times, with each copy assigned to a different node in the cluster. For the three node cluster used in this design, every node has a copy of each shard. If a node fails, the other two nodes will remain in read-write mode with the ability to make configuration changes as before. However, if two nodes fail, the third node will switch to read-only mode and no configuration changes will be allowed. If the two data centers becoming isolated from each other, Pod-1 with two APICs will be able to make configuration changes but Pod-2 will be in read-only mode. Once the split-brain scenario resolves, any configuration changes made in Pod-1 during the failure will be applied to the Pod-2. To address long outages, customers can also deploy a second APIC in Pod-2. This fourth APIC can be a fully active node in the cluster or it can be a backup node that is brought into service only when needed.

The size of the APIC cluster also impacts the scalability of the cluster. The 3-node APIC cluster used in this design can support up to 80 Leaf nodes across all Pods in the Multi-Pod fabric. As the fabric grows and the number of leaf and spine switches increase, additional APICs can be added to the cluster. At the time of writing, an APIC cluster can support up to 7 nodes, and up to 400 Leaf switches (max of 200 per Pod) across 12 Pods. For additional scalability information, see the Verified Scalability Guide in the References section of this document.

For the most up-to-date scalability numbers, review both the Verified Scalability Guide and the release notes for a given APIC release.

Inter-Pod Network

In a Cisco ACI Multi-Pod architecture, data centers or Pods in different locations are interconnected using an Inter-Pod Network. The Inter-Pod network is not part of the ACI fabric nor is it managed by the APIC but it is critical for enabling seamless connectivity between data centers.

In this solution, the round-trip time between data centers interconnected by the Inter-Pod network must be <5ms. ACI supports a round-trip latency of up to 50msec between data centers but the HyperFlex stretched cluster providing the infrastructure requires a latency < 5ms between data centers. The Inter-Pod network must therefore meet the <5ms latency requirement between data centers.

The Inter-Pod network provides Layer 3 connectivity for establishing Pod to Pod VXLAN tunnels between datacenters. The VXLAN tunnels enable seamless Layer 2 extension and Layer 3 forwarding between data centers.

The protocols used in the Inter-Pod network to provide Layer 2 extension and Layer 3 forwarding between data centers are:

- Open Shortest Path First (**OSPF**) for exchanging reachability information, primarily VXLAN Tunnel End Point **(TEP)** addresses between Pods. Each Pod uses a unique TEP pool that must be advertised to the other Pod in order to establish VXLAN Tunnels between Pods. Spine switches that connect to the IPN also use **proxy TEP** addresses that needs to be advertised as well. The proxy TEP addressing enables spine switches in a Pod to advertise equal cost routes for the subnets advertised from that Pod. The receiving Pod and IPN will see the routes as being reachable through multiple equal cost paths and employ Equal Cost Multipathing (**ECMP**) to distribute traffic across all spine switches in that Pod.
- Dynamic Host Configuration Protocol (**DHCP**) Relay for enabling auto-discovery and auto-provisioning of spine and leaf switches across the IPN.
- Bi-Directional Platform Independent Multicast (**BIDIR-PIM**) for forwarding Broadcast, Unknown unicast, and Multicast (**BUM**) traffic between Pods using IP multicast. BUM traffic is encapsulated in a VXLAN multicast frame and sent to leaf switches within the Pod, and to remote Pods across the Inter-Pod network. BIDIR-PIM is used in the Inter-Pod network to establish multicast flows between Pods.
- Link Layer Discovery Protocol (**LLDP**) for ease of troubleshooting. LLDP is optional but recommended across all interfaces in the Inter-Pod network

The design guidelines, considerations and best-practices for designing an Inter-Pod network are provided below:

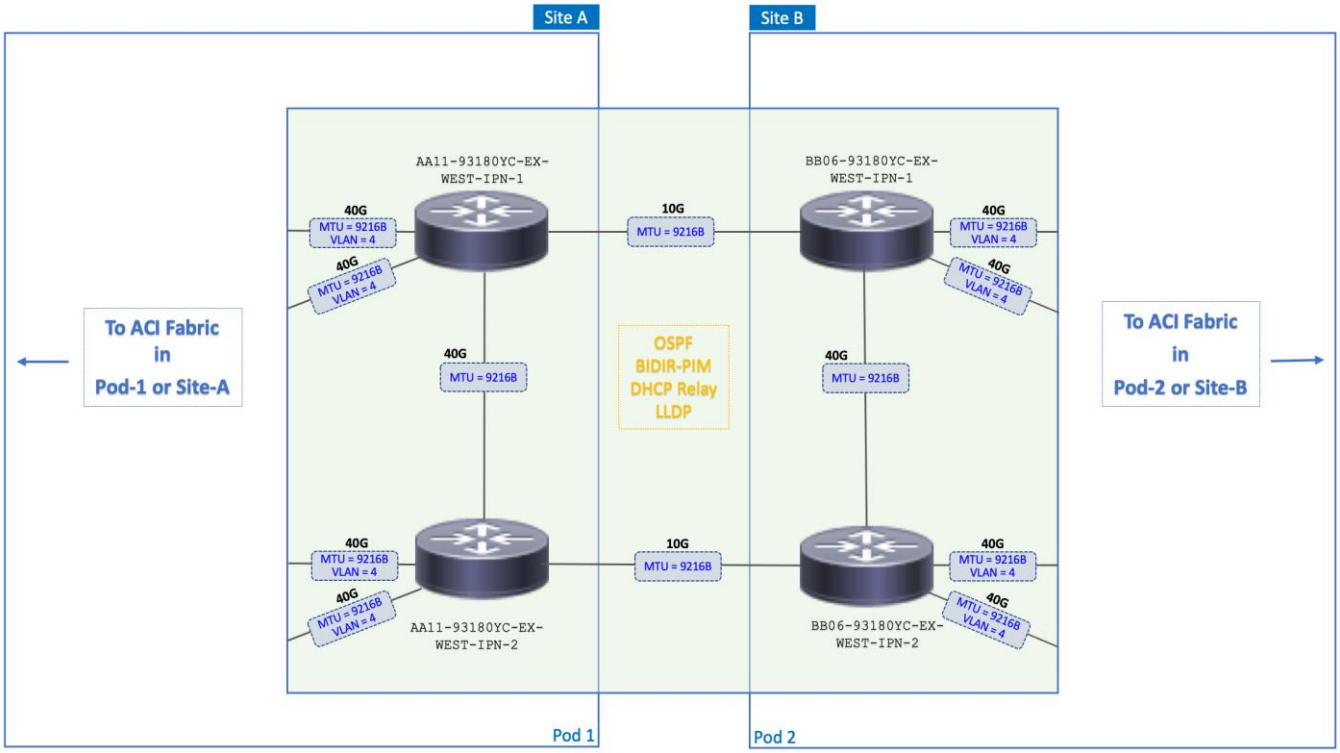
- The IPN can be a single switch or an extensive IP Network. If the Inter-Pod network is a large IP network with multiple devices, the protocols necessary to interconnect the ACI fabrics do not have to be enabled across the entire network. It can be enabled just on the devices providing IPN functionality that have direct connectivity to the Spine switches in each Pod.
- Virtual Routing and Forwarding (VRF) should be used in the IPN to isolate the traffic between the Pods. The IPN is an extension of the IP underlay in ACI that is being extended across Pods. It is best to not expose the underlay network, particularly one that interconnects multiple data centers. Therefore, VRFs should be used.
- As of this writing, OSPFv2 is the only routing protocol supported on spine switches connecting to the Inter-Pod network. ISIS is still used for routing within the Pod. In each Pod, IPN devices establish OSPF neighbor relationship with local spine switches and IPN devices in remote Pod. Reachability information, primarily VXLAN Tunnel End Point (VTEP) addresses are exchanged between Pods to establish leaf to leaf and spine to spine VXLAN tunnels between data centers.

- To support auto-discovery and auto-provisioning of spine and leaf switches in remote Pods across the IPN, IPN devices in each Pod must be able to relay DHCP requests from these switches to the APIC cluster. DHCP relay is enabled on interfaces connecting to spine switches that need to be discovered across the Inter-Pod network.
- IPN device must support a BIDIR-PIM range of at least /15. Among Nexus 9000 series switches, second generation switches support this but bot first generation switches – first generation switches can only support a max BIDIR-PIM range of /24. For other Cisco platforms, verify support before they are used as IPN devices.
- In ACI, each bridge domain is assigned a unique IP multicast group address when it is first defined. The address is allocated from a pool of multicast addresses, known as Infrastructure Global IP Outside (**Infra GIPo**) addresses. When a bridge domain is first defined in an ACI Multi-Pod fabric, the multicast group address assigned to that bridge domain will require a separate multicast group for traffic that spans both Pods. The multicast group address for forwarding BUM traffic across the IPN can be allocated from the infrastructure GIPo pool assigned for use within a Pod or a completely new pool (**System GIPo**) can be allocated for this purpose.
- BIDIR-PIM requires a Rendezvous Point (RP) for forwarding BUM traffic using IP multicast. For RP resiliency, a phantom RP can be used as a backup RP. For more details on Phantom RP – see Cisco ACI Multi-pod Configuration White Paper in the References section.
- Traffic forwarded through the IPN should not come back into the Pod before it is forwarded to the remote Pod. The traffic between Pods should be forwarded directly between IPN switches and it should not come back to the spine switches before it is forwarded. This can happen if the bandwidth on the IPN-to-IPN links are lower than that of Spine to IPN links. For example, if the IPN-IPN connectivity uses 10Gbps links while the Spine to IPN uses 40 Gbps links, then OSPF could see a lower cost route to the other Pod as being through one of the Spine switches within a Pod.
- Traffic between Pods are forwarded using VXLAN tunnels established between the spine switches in each Pod. To support VXLAN traffic , the MTU on the IPN interfaces must account for the 50 Bytes of VXLAN encapsulation overhead. If the endpoints sending traffic between data centers require Jumbo frame support, the MTU should be jumbo frame size plus 50B for VXLAN.

The IPN device(s) in an ACI Multi-Pod fabric can be any device that can support the different protocols (OSPF, BIDIR-PIM, DHCP Relay) as well as an MTU (jumbo MTU + 50 bytes) that support VXLAN across the pods.

In this design, a pair of Nexus 93180YC-EX (2nd generation) switches are deployed as IPN devices in each Pod as shown in Figure 3. The Pods are connected using two 10GbE fiber links. A 75km single-mode fiber spool is used for each link to emulate geographically dispersed data centers in Cisco Labs. The Inter-Pod network in this design provides multiple redundant paths between data centers with no single point of failure as shown in Figure 3.

Figure 3 Inter-Pod Network Between Active-Active Data Centers

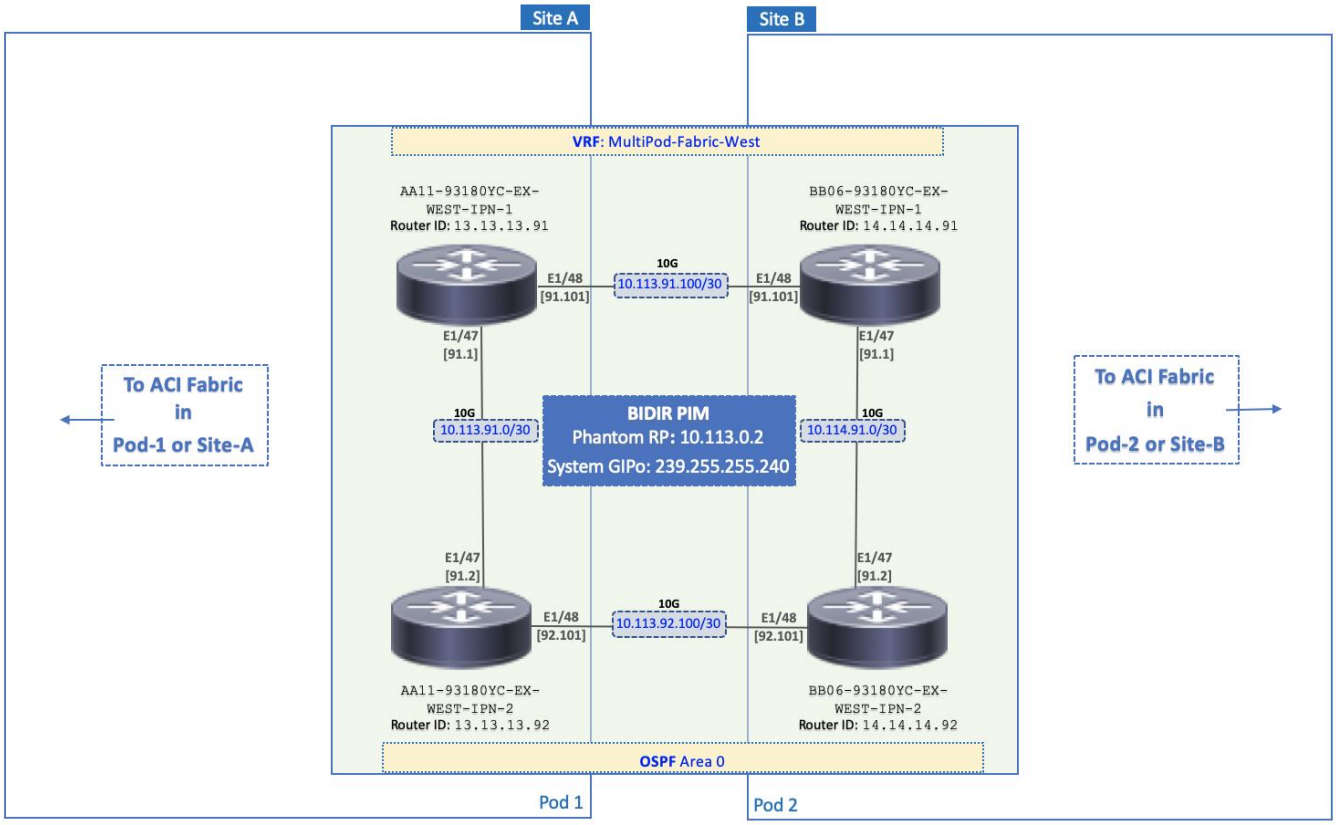


All interfaces in the Inter-Pod network are configured for an MTU of 9216 Bytes as shown in Figure 3. In this design, the endpoints that send traffic between data centers include HyperFlex nodes in the stretched cluster and virtual machines hosted on the stretched cluster. HyperFlex clusters use jumbo frames (9000 Bytes) for Storage and vMotion traffic. The specific MTU value (9216B) was chosen to maintain consistency with the default MTU used on other Cisco platforms.

A dedicated VRF is used to provide routing isolation in the Inter-Pod network as shown in Figure 4. All interfaces are put in a VRF to isolate the routing and the IP underlay between data centers. The Inter-Pod network is also configured for the following protocols:

- OSPF for exchanging VXLAN TEP addresses between data centers. All IPN routers and the spine switches that connect to it are in OSPF Area 0.
- DHCP-Relay is enabled on the IPN interface that connect to spine switches in Pod-2 so that they can be discovered by APICs in Pod-1. The APIC cluster in Pod-2 is not available at this time and cannot be brought online until the Pod-2 spine and leaf switches in Pod-2 are discovered and provisioned.
- BIDIR-PIM is also enabled on all IPN routers so that BUM traffic can be sent as multicast between Pods. All traffic is forwarded using a Rendezvous Point (RP). To provide redundancy, a backup RP model involving a Phantom RP is used in this design. A separate System GIPo, separate from the Infra GIPo is used for the multicast-group addresses in the Inter-Pod network.

Figure 4 Inter-Pod Network - Detailed Design



Pod To IPN Connectivity

To enable seamless Layer 2 extension and Layer 3 forwarding between data centers, VXLAN tunnels are established across the Inter-Pod network. The Layer 3 connectivity and the protocols necessary to achieve this are enabled in the Inter-Pod network.

The design guidelines, considerations and best-practices for connecting the ACI fabric in a Pod to the Inter-Pod network are provided below:

- Each Pod connects to the Inter-Pod network through one or more spine switches. Not all spine switches in a Pod are physically connected to the IPN. For redundancy and load distribution, at least two Spine switches should be used to connect to the IPN in each Pod. IPN Traffic between Pods are distributed across all spine switches that connect to the IPN.
- Spine switches cannot be connected back-to-back between Pods – they must connect through at least one IPN router/switch in the IPN network. The physical links that connect spine switches to the Inter-Pod network can be 10GbE/40GbE/100GbE (at the time of writing). On Nexus 9364C spine switches, 10GbE is possible on the two 1/10GbE ports or using special adapters.
- Spine switches must have an active Leaf facing link, otherwise it cannot be used by the fabric. If you are deploying new Pods and connecting them to the ACI Multi-Pod fabric – at least one leaf should be connected to the spine switches before it can be auto-discovered and auto-provisioned by APIC across the Inter-Pod network. ACI uses LLDP to determine the presence of an active Leaf.
- Each Pod requires an External VXLAN TEP (ETEP) pool, in addition to the internal TEP pool. The internal TEP pool is used for VXLAN overlay within a Pod. The external TEP pool is used for VXLAN overlay across the Inter-Pod network. Each Pod must allocate separate ETEP and internal TEP pools – they should not overlap.

The Pod-1 to Inter-Pod network and Pod-2 to Inter-Pod network design is shown in Figure 5 and Figure 6 respectively.

Figure 5 Pod-1 to IPN Connectivity

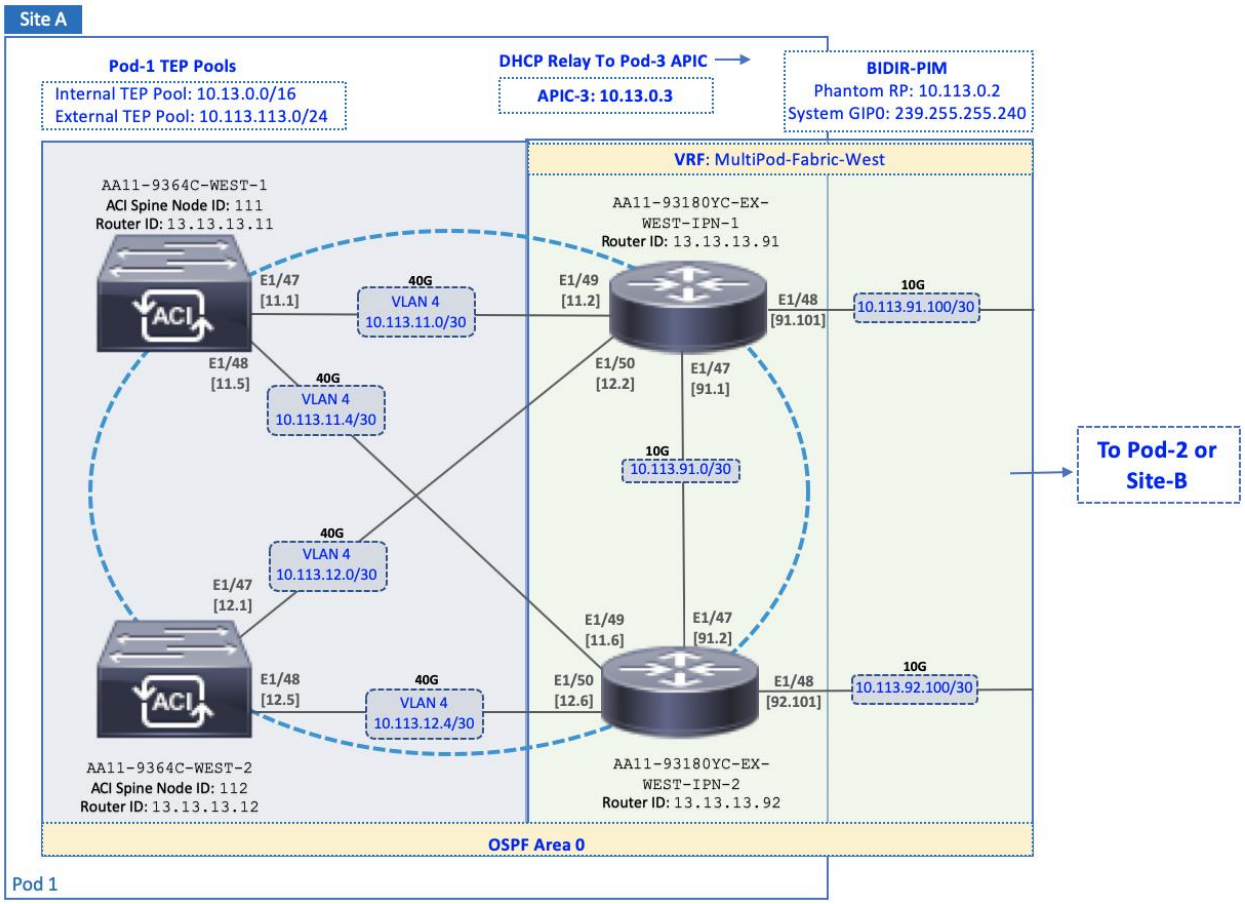
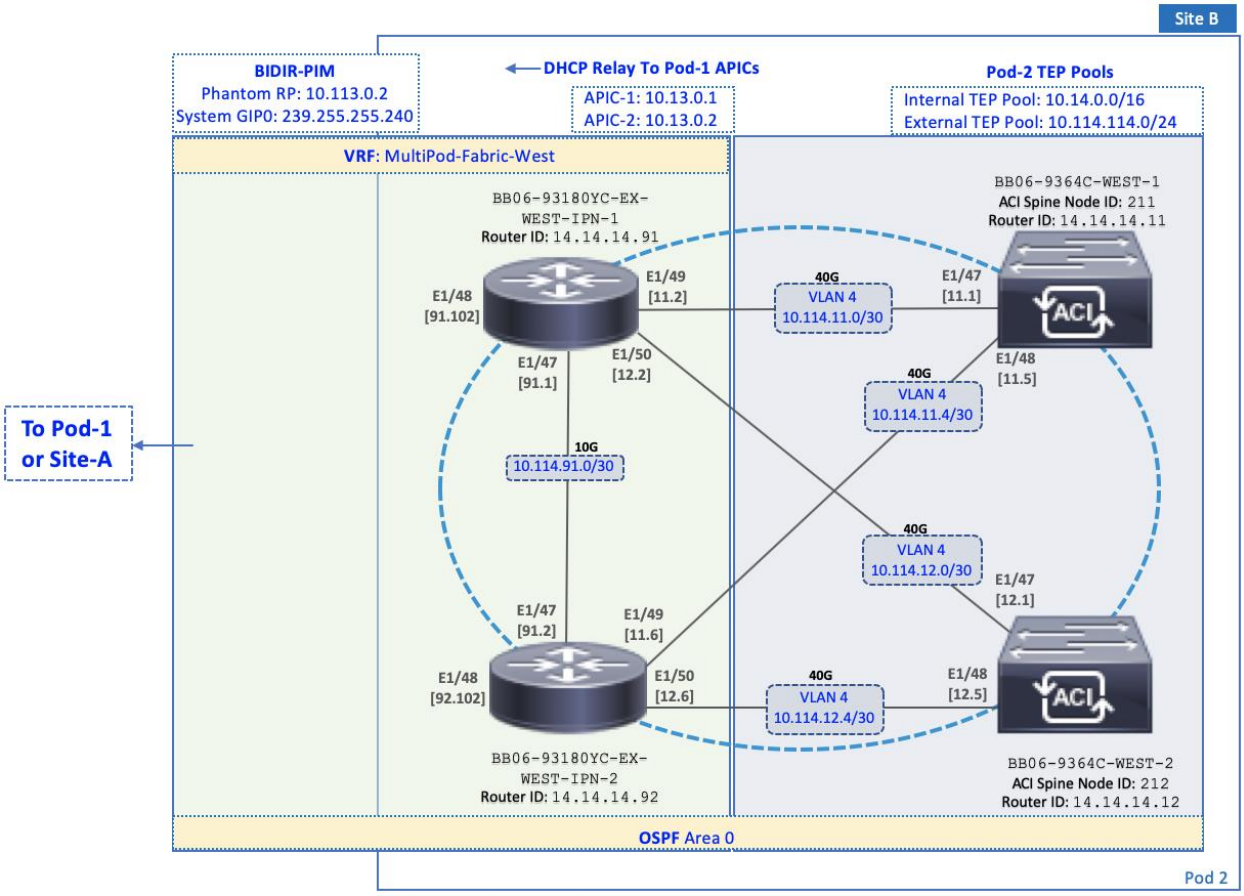


Figure 6 Pod-2 to IPN Connectivity



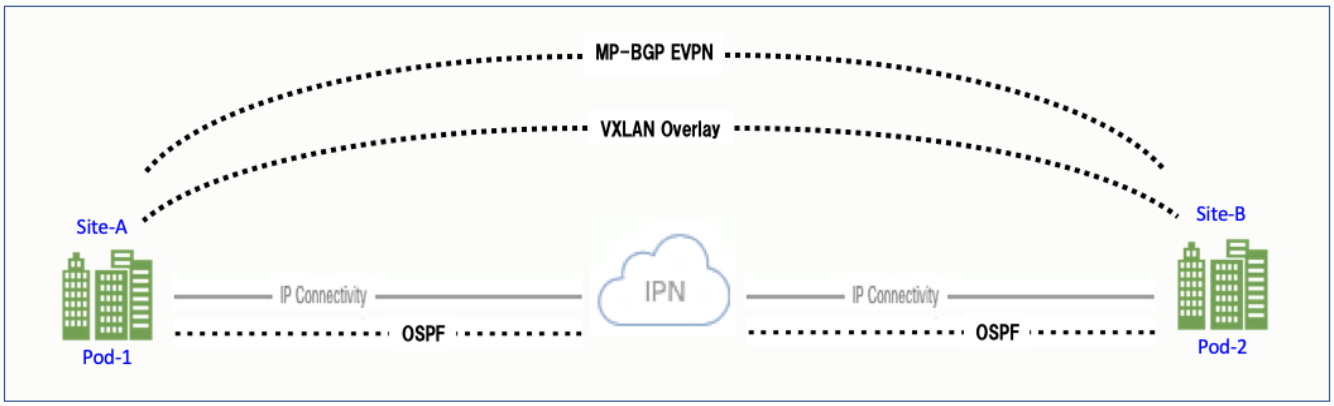
In this design, redundant 40GbE links are used for connectivity from each Pod to the Inter-Pod network as shown in Figure 5 and Figure 6. Two spine switches from each Pod connect to the IPN using multiple links, resulting in multiple paths to the IPN with no single point of failure. Customers can also use 10GbE links to connect spine switches to the IPN. Nexus 9364C spine switches is primarily a 40/100G switch but it does have two 1/10G ports that could be used for this purpose. To enable the VXLAN overlay network and establish VXLAN tunnels between the data centers, a separate external TEP pool is assigned to each Pod as shown. OSPF is enabled on the Spine switches in each Pod to connect to the IPN. To forward BUM traffic across the IPN, BIDIR-PIM is enabled. Phantom RP is used as a backup for the BIDIR-PIM RP. DHCP Relay is enabled in Pod-2 to enable zero-touch provisioning of spine, leaf and APIC in Pod-2.

Inter-Pod Design for Seamless Connectivity

The Inter-Pod network enables the IP underlay network that is necessary for establishing VXLAN tunnels between data centers. Once the physical and IP connectivity is in place, OSPF is enabled across the IPN and on the spine switches in each Pod that connect to the IPN to exchange TEP reachability information. The exchanged TEP addresses are then used to establish VXLAN tunnels between data centers. A Multi-protocol BGP (MP-BGP) EVPN session is then established to exchange endpoint reachability information.

A high-level overview of the Inter-Pod design and protocols that provide seamless Layer 2 extension and Layer 3 forwarding between Pods or data centers is shown in Figure 7.

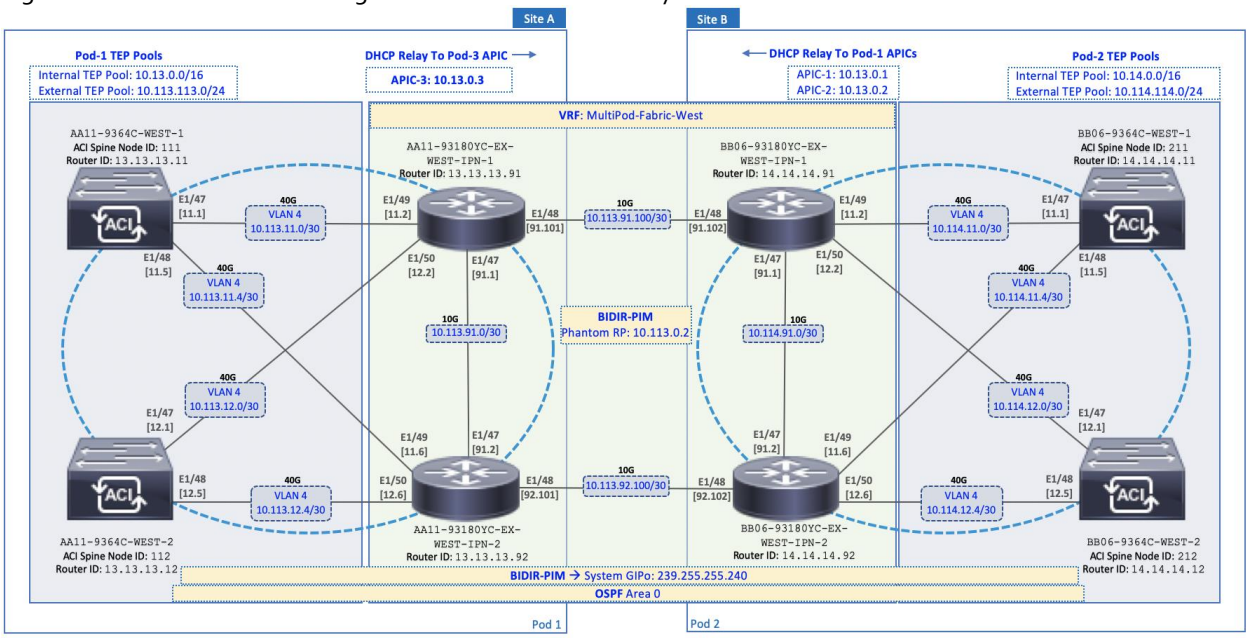
Figure 7 Inter-Pod Design for Seamless Connectivity



ACI uses similar mechanisms within each Pod or fabric to achieve the same functionality but with some notable differences. In an ACI Multi-Pod fabric, the VXLAN overlay and the IP underlay is across an external Inter-Pod network outside the ACI fabric. ACI uses ISIS for the routing protocol within each Pod or fabric but uses OSPF between Pods and across the IPN. MP-BGP EVPN session is established across the Inter-Pod network to enable endpoint (mac-address, IPv4) learning and exchange endpoint location information between data centers. Within a Pod, ACI uses the COOP protocol to exchange endpoint information. MP-BGP EVPN provides both multi-tenancy and support for multiple address families (mac-address, IPv4) that ACI requires. At least one MP-BGP EVPN peering must be in place between data centers. The peering will be between spine switches (that connect to IPN) in each Pod. However, a second MP-BGP EVPN peering is recommended between Pods for redundancy. Also, the TEP addressing for establishing VXLAN tunnels across the Inter-Pod network is allocated from an External TEP address pool but within a Pod or fabric, it is allocated from the internal TEP address pool for the Pod.

The detailed Inter-Pod design that enables seamless Layer 2 extension and Layer 3 forwarding between Pods or data centers is shown in Figure 8.

Figure 8 Detailed Inter-Pod Design for Seamless Connectivity



Accessing Outside Networks and Services

Outside networks in ACI refers to any networks outside the ACI fabric. This includes other internal (non-ACI) networks within the Enterprise as well as networks external to the Enterprise. ACI provides two main mechanisms for accessing networks and services outside the fabric as outline below:

- To connect to Layer 2 devices outside the ACI fabric, a Layer 2 bridged or Layer 2 Outside (L2Out) connection can be used.
- To connect to Layer 3 devices outside the ACI fabric, a Layer 3 routed or Layer 3 Outside (L3Out) connection can be used.

These options are independent of whether it is an ACI Multi-Pod fabric or a single-site fabric. Therefore, accessing outside networks and services in an ACI Multi-Pod fabric is the same as that of a single-site fabric. However, an ACI Multi-Pod fabric can have a higher number of outside connections if it is interconnecting multiple data centers and each data center is using one or more dedicated connections.

In an active-active data center design, each data center should have at least one connection to outside networks and services if it is important to maintain access to these services in the event of a data center failure or loss of connectivity to the data center that originally provided the access. This design assumes that both data centers in an active-active data center solution will have require equal access to the same networks and services, and therefore an outside connection is provided from both.

Also, the outside access provided in this design is a Layer 3 routed connection. The Layer 2 outside connection is typically used in migration scenarios for gradually extending an existing non-ACI bridge-group or subnet into the ACI fabric, and the gateway (if used) is currently in the non-ACI where it will remain until it can be migrated over to ACI. It is also used in certain scenarios where limited access to a subnet or service is required from within the ACI fabric, for example, access to an isolated network for management or monitoring. However, for maximum flexibility with the ability to route and access multiple subnets and services, either within the Enterprise or in the cloud, the preferred connectivity method is an L3Out connection, and therefore the method chosen for this design.

The L3Out connection in the active-active data centers provides access to the following networks and services in this design:

- Access to cloud-based services such as Cisco Intersight and Cisco Umbrella. Cisco Intersight provides centralized management of all HyperFlex and UCS clusters connected to the ACI Multi-Pod fabric, for all data center locations. Cisco Umbrella provides Enterprise users with DNS-based security when accessing the Internet or other cloud services, regardless of the location or device they use to connect.
- Access to infrastructure and application services hosted in existing Enterprise infrastructure, outside the ACI Fabric. For example, NTP, DNS, Active Directory, Cisco Umbrella Virtual Appliance are some of the services used in this design that are located in the non-ACI infrastructure.
- Connectivity to other internal networks within the Enterprise – for example, Campus, WAN, or for specific subnets such as the out-of-band management network for Cisco UCS FI and HX servers. The HyperFlex Installer VM hosted on the Management cluster requires access to management network for deploying the HyperFlex stretched cluster (Application cluster).

Shared L3Out – Design Options

In ACI, the Layer 3 outside connection can be a shared service where it is shared by multiple tenants or it can be dedicated on a per-tenant basis. In this design, the Layer 3 outside connection is envisioned as a shared or common service that all tenants can use. In ACI, the shared Layer 3 connection that ***all*** tenants can use is referred to as a **shared L3Out,** and it is typically part of the **common** Tenant. The **common** tenant is pre-defined system tenant where any objects defined in this tenant are visible to all other tenants, making it easier to position common services that many tenants will need access to.

ACI provides a number of design options for enabling a shared L3Out connection in the **common** Tenant as outlined below:

- Option 1: VRF, Bridge Domain, Subnet and L3Out in system-defined **common** Tenant
- Option 2: Bridge Domain, Subnet in user-defined Tenants but VRF and L3Out in system-defined **common** Tenant
- Option 3: VRF, Bridge Domain and Subnet in user-defined Tenants but L3Out in system-defined **common** Tenant

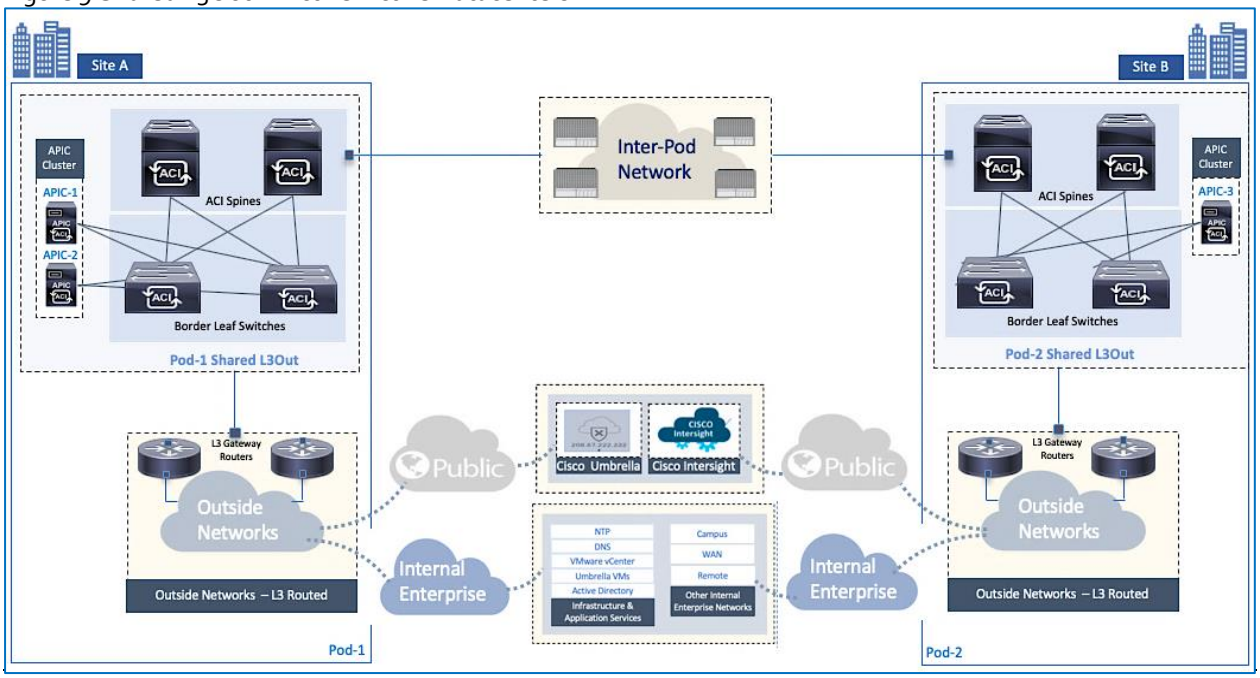
Option 3 is used in this design as it provides a more scalable L3Out solution, while enabling each Tenant using the Shared L3Out service to have separate VRF instances with overlapping address space, as long as the overlapping subnets are not leaked into the **common** Tenant. This option uses route leaking between VRFs to enable connectivity to the L3Out connection.

Shared Layer 3 connections can also be defined in other tenants. However, if the goal is for all tenants to have access to this connection (if needed), then the **common** Tenant in the ACI architecture is defined and provided for exactly this purpose. The **common** Tenant *provides* a contract for accessing the shared L3Out connection that other tenants can *consume* to gain access to outside networks.

Shared L3Out Design

The shared L3Out design used in this solution is the same for both active-active data centers. To enable a shared L3Out connection, border leaf nodes in the ACI fabric are connected to Layer 3 gateways in the outside network. The shared L3Out connections in the active-active data centers are shown in Figure 9.

Figure 9 Shared L3Out in Active-Active Datacenters



To connect each data center to outside networks using a shared L3Out, a pair of Nexus 9000 series leaf switches are deployed as ACI border leaf switches and connected to a pair of Nexus 7000 series gateway routers in the non-ACI infrastructure. The detailed shared L3Out connectivity for Pod-1 and Pod-2 are shown in Figure 10 and Figure 11, along with the ACI configuration to enable IP connectivity and routing.

Figure 10 Shared L3Out Connectivity for Pod-1

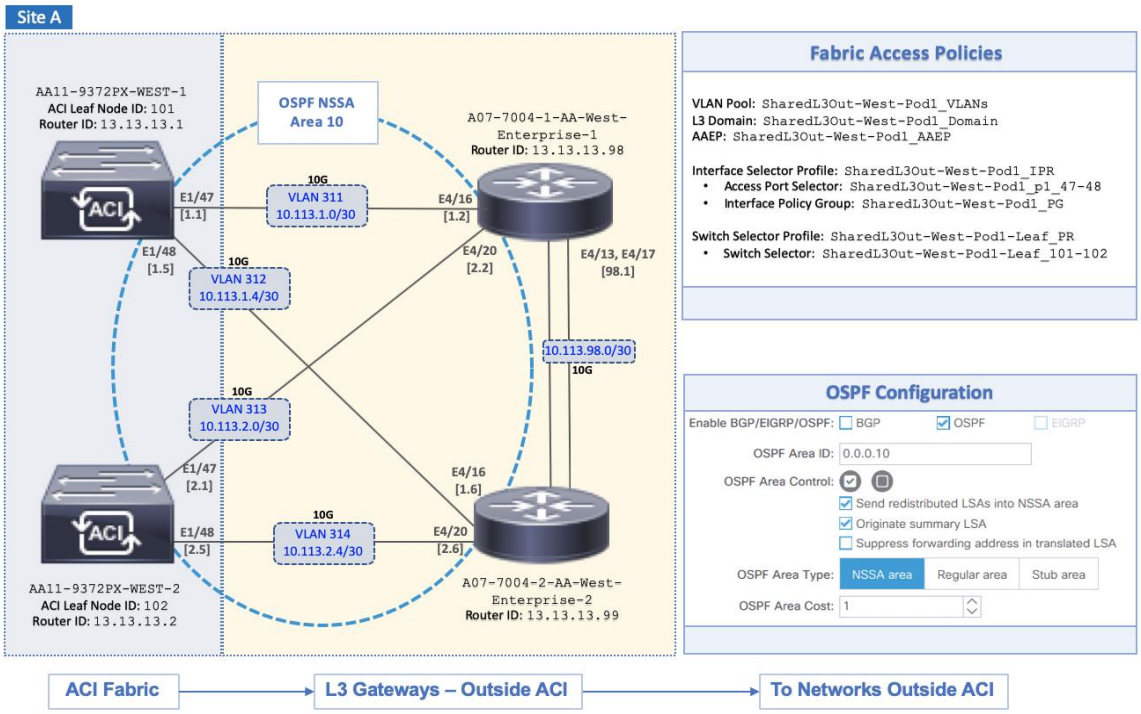
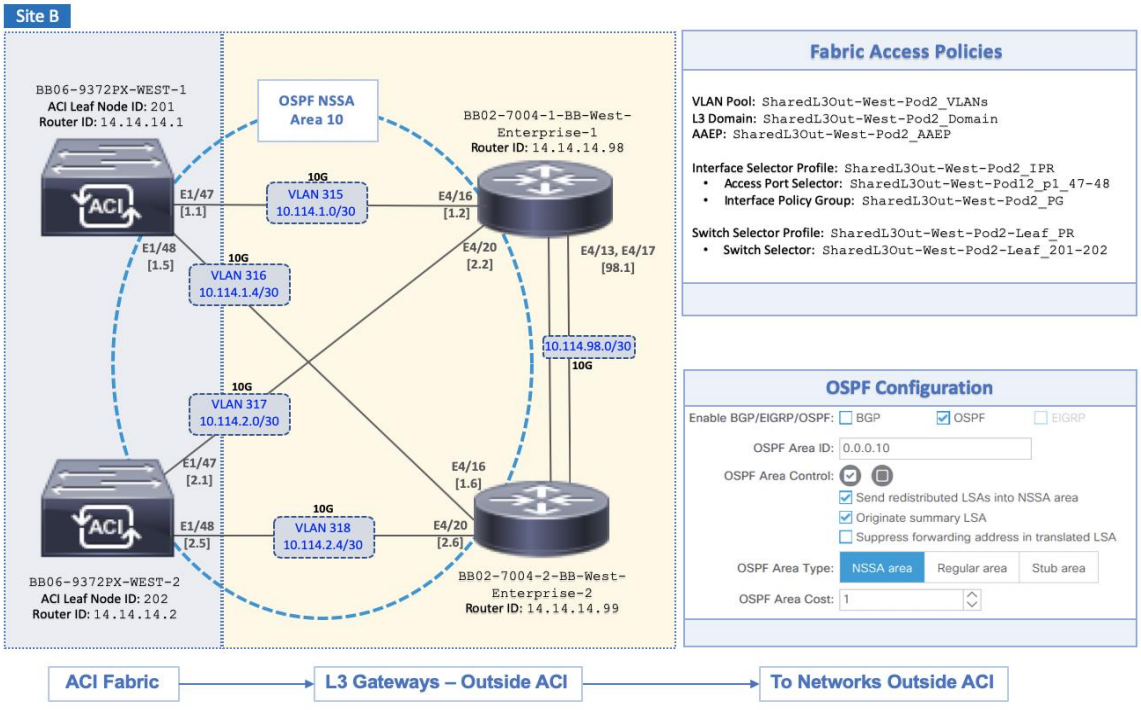


Figure 11 Shared L3Out Connectivity for Pod-2



Each border leaf switch is redundantly connected to the Nexus 7000 switches using 10GbE links. The four links between ACI leaf nodes and external routers are individual connections with a dedicated VLAN and IP subnet for each link – no link bundling is used. The border leaf switches in this design also provide connectivity to the APIC nodes in the cluster. For larger deployments, Cisco recommends using a dedicated pair of border leaf switches.

A routing protocol is then enabled across the layer 3 connection to exchange routes between the ACI and non-ACI domains. OSPF is used in this design. In this design, OSPF learns routes to outside networks, and advertises ACI routes to outside networks. Routes learned by ACI in the common Tenant are then shared with other ACI Tenants by providing and consuming contracts between these Tenants. In this design, a default route is learned from the Layer 3 gateways and advertises tenant subnets to the outside infrastructure. Note that this requires ACI tenant routes to be leaked to the common Tenant and then advertised outside the fabric. The leaked routes for each Tenant must be unique – overlapping subnets should not be leaked. OSPF metrics on Cisco Nexus 7000 switches can be optionally used to influence path preferences.

The ACI constructs and design for enabling and accessing a shared L3Out service is shown in Figure 12. These include:

- A single *External Routed Network* under tenant **common** to connect ACI fabric to Cisco Nexus 7000s using OSPF.
- A unique private VRF (`common-SharedL3Out_VRF`) network and a dedicated external facing bridge domain (`common-SharedL3Out_BD`) is defined under the **common** tenant. This private network is setup with OSPF to provide connectivity to external infrastructure.
- The access layer configuration for the two shared L3Out connections in the active-active datacenters must be separately configured. The ACI constructs for the two shared L3Outs are shown in Figure 12.
- The shared L3Out created in the **common** Tenant *provides* an external connectivity contract (`Allow-Shared-L3Out`) that can be *consumed* from any tenant. Contracts created in **common** Tenant are visible to all tenants. Therefore, the contract to the shared L3Out is also accessible by all tenants.
- When other tenants *consume* the contract, the Tenant subnets shared by the tenants will get advertised to the outside infrastructure. These tenant will also learn the routes to outside networks, to access the external infrastructure networks and endpoints. The outside routes in this design is a single *default* route.

By defining a shared L3Out in **common** tenant, the contract is provisioned as part of the L3Out configuration and it would automatically be available in all other tenants to consume, without doing any additional configuration since the objects (contracts in this case) from the **common** tenant are available in all other tenants. If the shared L3Out was deployed in any other tenant, the contract would have to be explicitly exported from that tenant to each tenant where this contract needs to be consumed.

Figure 12 ACI Constructs and Design for Shared L3Out

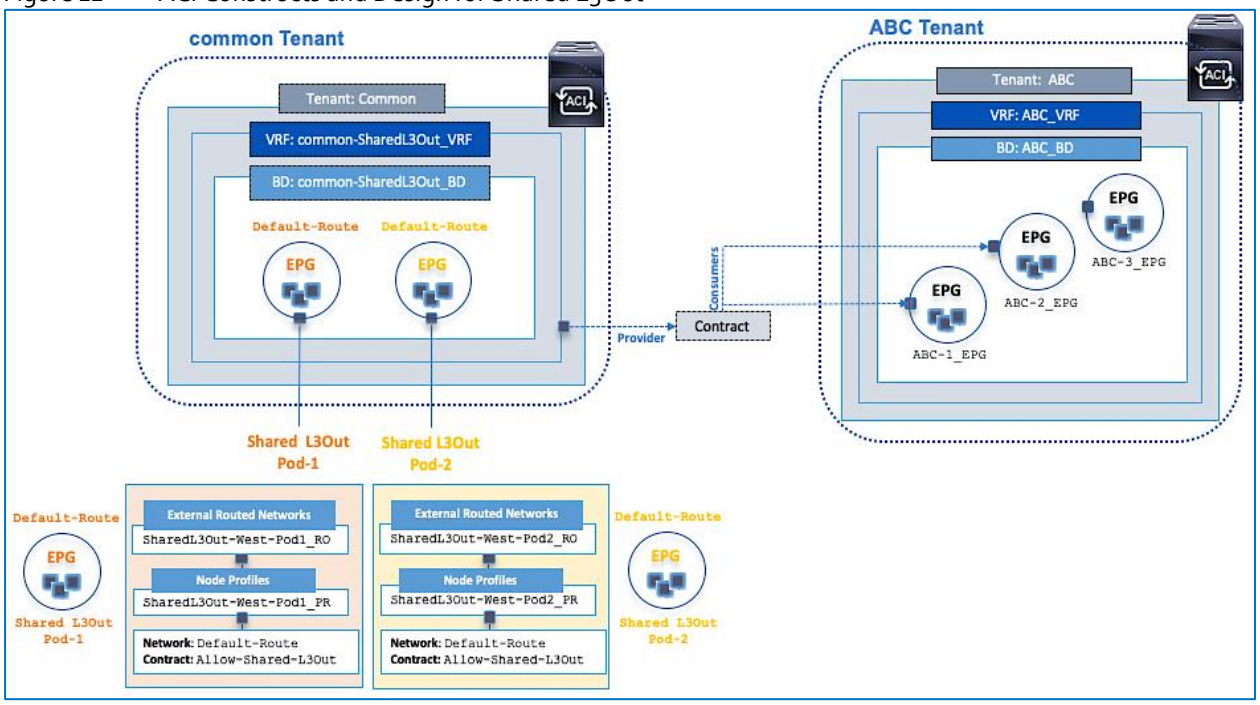
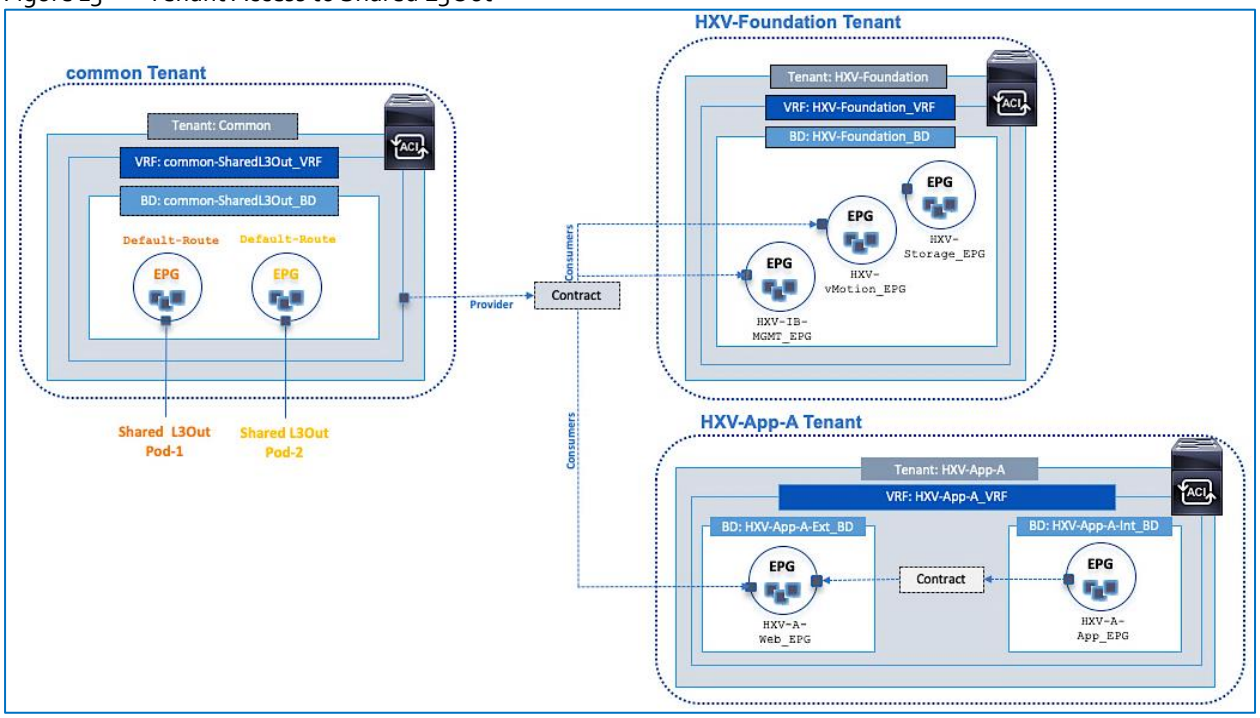


Figure 13 shows two user tenants (`HXV-Foundation, HXV-App-A`) in this design *consuming* the shared L3Out contract provided by the **common** Tenant to enable access to networks and services outside the ACI fabric.

Figure 13 Tenant Access to Shared L3Out



High Availability

The active-active data centers enable access to critical applications and services hosted in either data center location. To provide business continuity in the event of a site failure or a data center failure, a highly-resilient design is used throughout the ACI Multi-Pod fabric. High-availability is implemented within a Pod as well across Pods as discussed in previous sections. Some of the high-availability provided in this design are summarized below:

- Pod Connectivity: The connectivity within a Pod is the same for both active-active data centers. Redundant links are used between Spine and Leaf switches and from Leaf switches to access layer devices such as Cisco UCS Fabric Interconnects in the HyperFlex UCS domains, and non-ACI routers that provide connectivity to outside networks. Virtual Port-channels (vPCs) are used between leaf switches and HyperFlex UCS domains to provide both node and link-level redundancy. APIC nodes are also dual-homed to different leaf switches to provide redundant connectivity to the fabric. Connectivity from each Pod to the IPN is through two Spines switches and use multiple links to provide both node and link redundancy.
- Inter-Pod Connectivity between Pods: The Inter-Pod network in this design uses two IPN routers and two Spine switches for Pod to IPN connectivity in each data center location. Each IPN router is dual-homed to the Spine switches in that location. IPN routers also connect to remote IPN routers to provide two redundant paths between the sites, with no single point of failure.
- APIC Clustering: To provide resiliency and scalability, an APIC cluster consisting of multiple nodes are used to manage an ACI Multi-Pod fabric. APIC cluster uses data sharding to maintain three copies of the fabric configuration data, one on each node in the cluster. The nodes are distributed across both Pods in this design so that each site has a local APIC available in the event of a failure in the other site.
- ACI Multi-Pod architecture: By enabling distinct fabrics to be interconnected, the architecture enables a second fabric in a second location for use as a second data center, thereby providing redundant fabrics for an active-active data center design.
- Fault Isolation: ACI Multi-Pod fabric is designed to interconnect data centers and operate as a single fabric but each Pod is also a separate failure domain. To provide fault-isolation, ACI runs separate instances of the control plane protocols (IS-IS, COOP, MP-BGP) in each Pod so that an issue in one Pod does not de-stabilize the other.
- Connectivity from each Pod to Outside networks and services: To enable each site or Pod to operate as an independent datacenter, a shared L3Out is established from each data center location so that access to critical networks and services are available directly from that site.

Multi-tenancy

The ACI architecture is designed for multi-tenancy. Multi-tenancy enables the administrator to partition the fabric along organizational or functional lines to form different tenants. Tenants enable domain-specific management by defining a Tenant Administrator.

Tenant is a logical container for grouping applications and their networking and security policies. A tenant represents a unit of isolation from a policy perspective. This container can represent an actual tenant, an organization, an application or a group based on some other criteria. All application configurations in ACI are all done within the context of a tenant. Tenants will typically have multiple **Application Profiles**, **EPGs** and associated tenant networking that includes **VRF**s, **Bridge Domains** and **External Bridged or Routed Networks**.

Tenants can be system-defined or user-defined. The three system-defined tenants on the ACI fabric are **mgmt**, **infra**, **common** tenants. The **common** Tenant in ACI is designed for providing shared services to Tenants. For example, shared services such Microsoft Active Directory (AD), Domain Name System (DNS), etc. can be deployed in the **common** Tenant and any tenant can access it by *consuming* the contract provided by the **common** Tenant for that service. The contract will be available within the tenant for it to consume, without any additional configuration. Contract provided by a tenant will typically not show up in another tenant – this is unique to the **common** Tenant.

The user tenants are defined as needed by the administrator, to meet the needs of the organization. The two user tenants defined in this design are:

- `HXV-Foundation` Tenant
- `HXV-App-A` Tenant

Customers can deploy additional as needed to meet the needs of their business.

Tenant Design

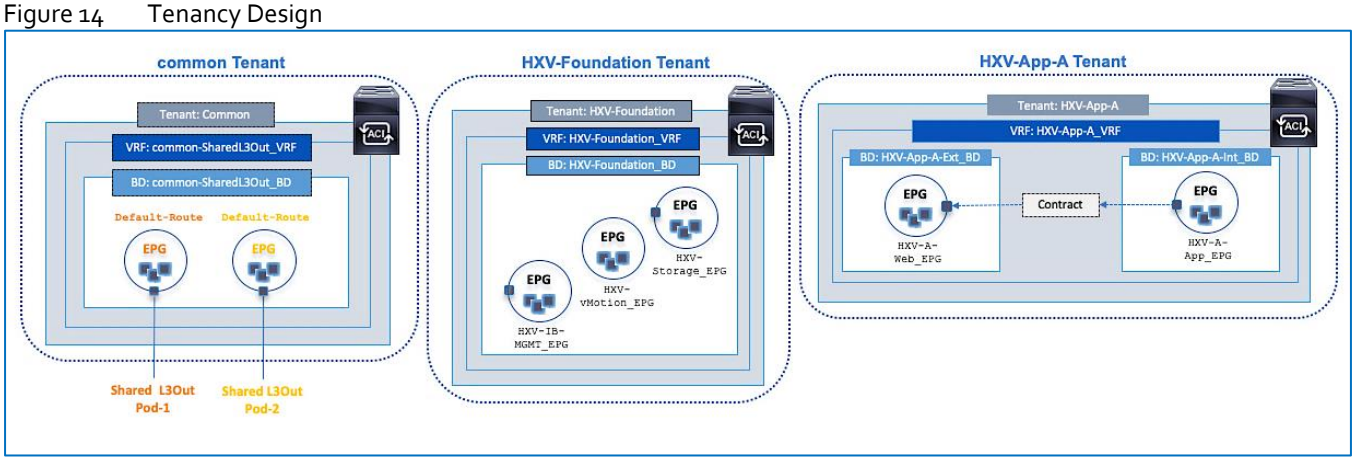
The tenancy design in an ACI fabric in terms of the number of tenants and what it is intended for, can be based on a number of factors. In this design, the tenancy design is based on the connectivity requirements of the endpoints, primarily HyperFlex clusters and the virtual machines hosted on the cluster, once the cluster is up and running. As such, two tenants are defined to meet the requirements outlined below:

- `HXV-Foundation` Tenant for HyperFlex infrastructure and services. This tenant provides the infrastructure connectivity and services necessary to deploy and manage a HyperFlex cluster, and to access services enabled by the cluster, primarily storage. It can be used for any HyperFlex cluster deployed in the ACI Multi-Pod fabric. In this design, `HXV-Foundation` Tenant provides infrastructure connectivity for both HyperFlex clusters, the standard cluster for Management and the stretched cluster for Applications.
- `HXV-App-A` Tenant for application virtual machines. The virtual machines will be hosted on the HyperFlex clusters. The clusters will need to be up and running, with compute, storage and virtualization in place before any virtual machines can be deployed on this cluster.

ACI also uses a number of system-defined tenants (`infra, mgmt`) to operate and manage the fabric. It also provides a **common** Tenant (discussed above) for shared or common services that multiple Tenants need access.

The three tenants and the associated ACI constructs are shown in Figure 14.

Figure 14 Tenancy Design



ACI Constructs

Cisco ACI architecture uses a number of design constructs that enable connectivity through the ACI fabric. The key design elements in ACI are summarized below.

- Tenant – A tenant is a logical container which can represent an actual tenant, organization, application or a construct for grouping. From a policy perspective, a tenant represents a unit of isolation. All application configurations in Cisco ACI are part of a tenant. Within a tenant, multiple VRF contexts, bridge domains, and EPGs can be defined according to application requirements.

- VRF – Tenants can be further divided into Virtual Routing and Forwarding (VRF) instances (separate IP spaces) to further separate the organizational and forwarding requirements for a given tenant. A Tenant can have multiple VRFs. IP addressing can be duplicated across VRFs for multitenancy.
- Bridge Domain – A bridge domain (BD) is a L2 forwarding construct that represents a broadcast domain within the fabric. A bridge domain is associated with a single tenant VRF but a VRF can have multiple bridge domains and endpoints. The endpoints in a BD can be anywhere in the ACI fabric, distributed across multiple leaf switches. To minimize flooding across the fabric, ACI provides several features such as learning of endpoint addresses (Mac/IP/Both), forwarding of ARP Requests directly to the destination leaf node, maintaining a mapping database of active remote conversations, local forwarding, and probing of endpoints before they expire. Subnet(s) can be associated with a BD to provide an L3 gateway to the endpoints in the BD.
- End Point Group – An End Point Group (EPG) is a collection of physical and/or virtual end points that require common services and policies, independent of their location. Endpoints could be physical servers, VMs, storage arrays, etc. For example, a Management EPG could be a collection of endpoints that connect to a common segment for management. Each EPG is associated with a single bridge domain but a bridge domain can have multiple EPGs mapped to it.
- Application Profile – An application profile (AP) models application requirements and contains one or more EPGs as necessary to provide the application capabilities. A Tenant can contain one or more application profiles and an application profile can contain one or more EPGs.
- Contracts – Contracts are rules and policies that define the interaction between EPGs. Contracts determine how applications use the network. Contracts are defined using provider-consumer relationships; one EPG provides a contract and another EPG(s) consumes that contract. Contracts utilize inbound/outbound filters to limit the traffic between EPGs or applications based EtherType, IP protocols, TCP/UDP port numbers and can specify QoS and L4-L7 redirect policies.

ACI Constructs in an ACI Multi-Pod Fabric

The connectivity that endpoints require from the ACI fabric are defined and enabled through different ACI constructs. For a given endpoint group, the ACI constructs that enable the connectivity is configured once. New endpoints can now be added to the group and leverage the already defined ACI constructs to receive the same connectivity. In an ACI Multi-Pod fabric, endpoints can be added to this endpoint group from anywhere in the fabric, including the two active-active data centers without the need for any additional configuration.

As previously stated, an ACI Multi-Pod fabric is a managed as a single ACI fabric using a single APIC cluster. As such, the ACI constructs (Tenant, VRF, Bridge Domain, Application Profile, EPG, Contracts) and policies for enabling endpoint forwarding through the fabric are not Pod specific. Once an endpoint group is created and associated configuration is complete, it is available fabric-wide, enabling endpoints to be added to that EPG from anywhere in the fabric, including any data center or Pod. As discussed in earlier sections, ACI will provide the Layer 2 extension and Layer 3 forwarding necessary for enabling seamless connectivity across the fabric, including the different data centers in an ACI Multi-Pod fabric. Application workloads and other virtual machines can therefore be positioned quickly and easily from any Pod or data center by adding it the endpoint group. The workloads can also be moved between data centers without any additional configuration. Therefore, for an endpoint, the configuration for enabling forwarding is done *once,* at the endpoint group level. This is true for any ACI fabric as long as it is being managed by a single APIC cluster.

Though the forwarding configuration is done only once, the access layer configuration for attaching endpoints to the fabric will need to be done for every attachment point. This is to be expected since the connectivity outside the fabric can vary depending on the type of endpoint and connectivity method used. Also, once the access layer policies are defined, it can be re-used across the fabric for different access layer connections. The access layer configuration for endpoints will be discussed in greater detail in an upcoming section. Therefore, the ACI configuration necessary to connect an endpoint and enable forwarding in an ACI Multi-Pod fabric is *identical* to that of a single-site ACI fabric.

Tenant Design for HyperFlex Infrastructure

As stated earlier, the HyperFlex infrastructure in this design includes a standard HyperFlex cluster for Management and a stretched HyperFlex cluster for Applications. Both clusters connect to the ACI Multi-Pod fabric. Management cluster connects to Pod-1 while the Application cluster spans both Pods.

Endpoints in this design are either part of the `HXV-Foundation` Tenant or `HXV-App-A` Tenant and may use services provided by the **common** Tenant (for example, shared L3Out service provided by **common** Tenant). The infrastructure connectivity and services that HyperFlex nodes (or endpoint in ACI) in a cluster require are provided by the `HXV-Foundation` Tenant. The connectivity required by the application virtual machines hosted on the stretched HyperFlex cluster are provided by the `HXV-App-A` Tenant in this design. Note that application virtual machines are not deployed in the standard HyperFlex cluster in this design, but they can be, in which case it would use the `HXV-App-A` Tenant per this design. Conversely, management virtual machines can be deployed in the Application cluster, and if they do, it would use the `HXV-Foundation` Tenant.

The specific HyperFlex infrastructure connectivity and services provided by the `HXV-Foundation` Tenant in this design are:

- Connectivity to in-band management network: The HyperFlex ESXi hosts and storage controller virtual machine (SCVM) (on every HyperFlex node) communicate over the same in-band management network in the HyperFlex architecture. The connectivity for these end points are provided by the in-band EPG. In this design, *both* clusters share the same in-band management network.
- Connectivity to storage data network: The HyperFlex ESXi hosts and storage controller virtual machine (SCVM) (on every HyperFlex node) also communicate over the storage-data network to enable and access storage services in the HyperFlex architecture. The communication between nodes on network is critical for the health of the cluster, to provide storage services and for the basic functioning of the distributed storage platform. Since the health of the storage cluster and the integrity of the data relies on this network, a *separate* storage-data network is used for each cluster to isolate this network as much as possible.
- Connectivity to VMware vMotion network: To support VMware vMotion for virtual machines hosted on the HyperFlex ESXi hosts, the hosts needs connectivity to the VMware vMotion network. In this design, *both* clusters share the same VMware vMotion network.
- Connectivity for infrastructure management and services network (optional): Multiple networks can be defined as needed to serve a variety of functions. The networks can be used to provide or access services. In this design, virtual machines that require connectivity to other networks for Infrastructure management are deployed in this network. The infrastructure management virtual machines are also hosted on the Management cluster. VMware vCenter and HyperFlex installer are two of the virtual machines that use this network. In this design, the Installer VM requires connectivity to multiple networks and devices to install and deploy the HyperFlex stretched cluster. ACI will route the traffic as needed to enable this connectivity.

In an ACI, all connectivity and services are defined using ACI constructs (Tenants, Application profiles, Bridge domains and EPGs). To provide the connectivity outlined above that HyperFlex endpoints (HyperFlex ESXi nodes, Storage Controller VMs) require, the HyperFlex networks that these endpoints communicate on in the UCS and HyperFlex domains, are mapped to end point groups in ACI. EPGs and other ACI construct enable ACI to provide these endpoints with connectivity through the fabric, both between endpoints in the same network as well as outside networks such as the shared L3Out in the **common** Tenant. The HyperFlex ESXi nodes and storage controller VMs are part of the `HXV-Foundation` Tenant in this design.

The ACI constructs that enable connectivity for HyperFlex infrastructure are summarized below .

Figure 15 Endpoint Groups for HyperFlex Infrastructure

Endpoint Groups

EPG	Notes
HXV-IB-MGMT_EPG	In-Band Management Network – Endpoints in this group include ESXi Management VMkernel interface and HyperFlex SCVM* Management interfaces on HX nodes
HXV-CL0-StorData_EPG	Storage Data Network for HyperFlex standard cluster – Endpoints in this group include ESXi Storage Data VMkernel interface and HyperFlex SCVM Storage Data interfaces on HX nodes
HXV-CL1-StorData_EPG	Storage Data Network for HyperFlex stretched cluster – Endpoints in this group include ESXi Storage Data VMkernel interface and HyperFlex SCVM Storage Data interfaces on HX nodes
HXV-vMotion_EPG	VMware vMotion Network – Endpoints in this group include ESXI vMotion VMkernel interfaces on HX nodes

* SCVM – Storage Controller VM – one on each HX node in a cluster

Figure 16 Tenant Networking for HyperFlex Infrastructure

Tenant Networking

Tenant	VRF	Bridge Domains	Associated EPG	Notes
HXV-Foundation	HXV-Foundation_VRF	HXV-IB-MGMT_BD	HXV-IB-MGMT_EPG	
		HXV-Storage_BD	HXV-CL0-StorData_EPG	For HyperFlex standard cluster
		HXV-CL1-Storage_BD	HXV-CL1-StorData_EPG	For HyperFlex stretched cluster
		HXV-vMotion_BD	HXV-vMotion_EPG	

Figure 17 Application Profiles for HyperFlex Infrastructure

Application Profiles

Application Profiles	EPG	Notes
HXV-IB-MGMT_AP	HXV-IB-MGMT_EPG	
HXV-vMotion_AP	HXV-vMotion_EPG	
HXV-Storage_AP	HXV-CL0-StorData_EPG HXV-CL1-StorData_EPG	Same application profile is used for for HX nodes in both standard and stretched clusters

The relationship between the various ACI constructs used for enabling connectivity for the HyperFlex infrastructure are shown in the following figures:

Figure 18 ACI Constructs for In-Band Management and vMotion

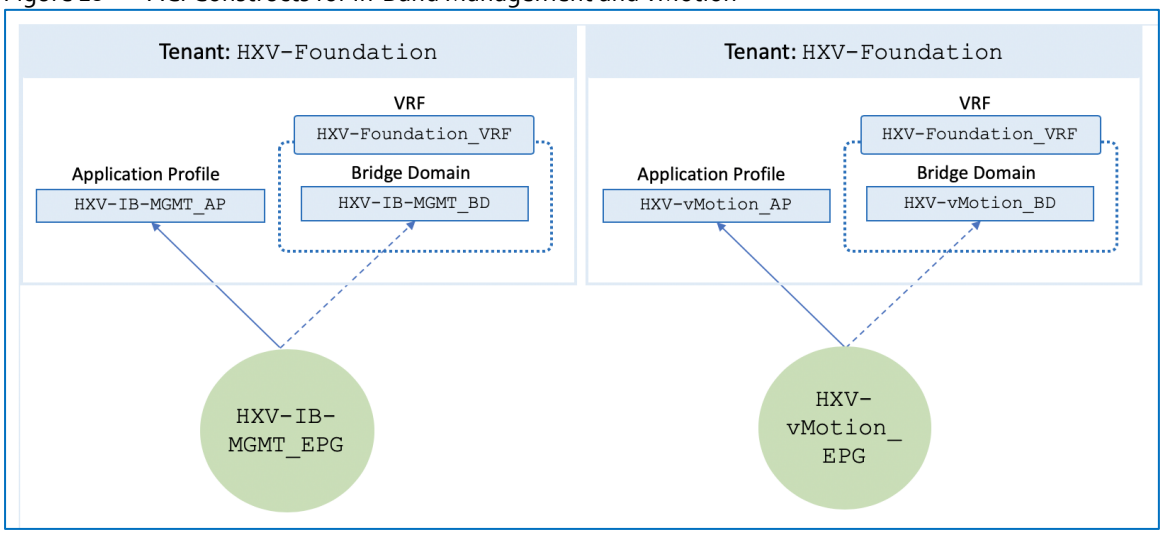
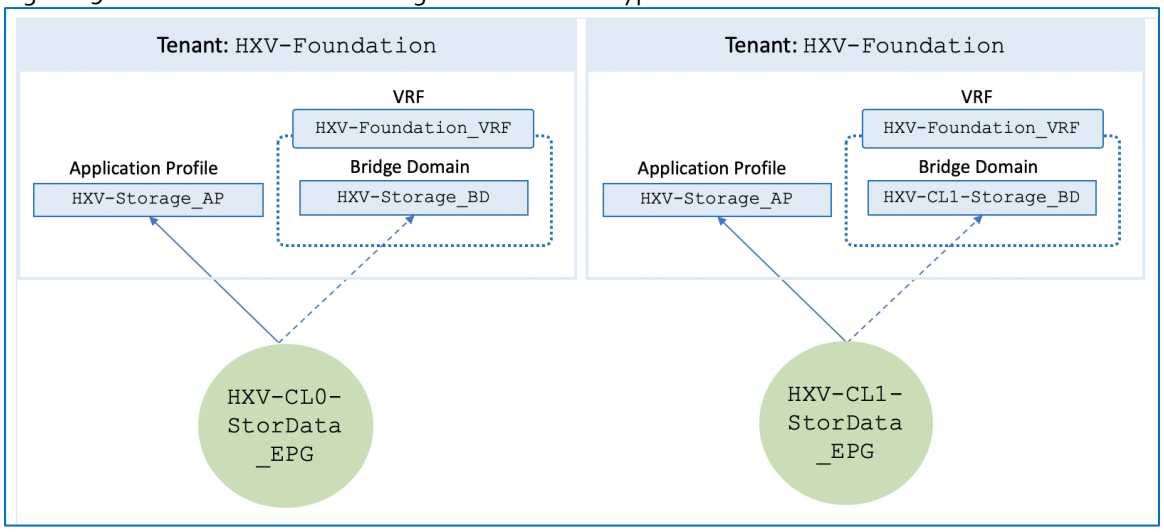


Figure 19 ACI Constructs for Storage Data EPGs for HyperFlex Standard and Stretched Clusters

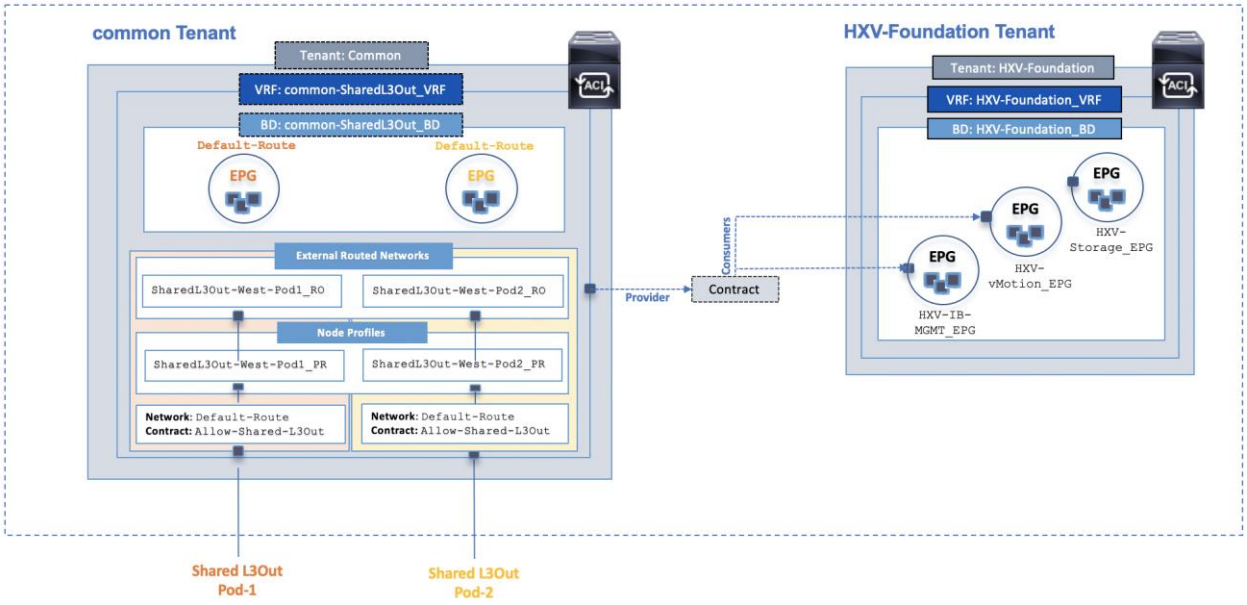


Enable Connectivity to Outside Networks and Services

To enable connectivity to outside networks and services for HyperFlex infrastructure in the `HXV-Foundation` Tenant, the contract *provided* by the **common** Tenant for the shared L3Out needs to be *consumed* by the endpoints (EPGs) that require that connectivity. In this design, HyperFlex endpoints (ESXi nodes, SCVM) in the management network and vMotion networks are allowed access to the outside networks. The HyperFlex storage data network, specifically the endpoints in that network are not allowed (and cannot since it is not enabled for L3 forwarding) to access. This segment should be isolated as much as possible and should not need access to the shared L3Out and the networks and services reachable through it.

Figure 20 shows the ACI fabric connectivity for accessing networks and services using the shared L3Out. This connectivity is available to endpoints in `HXV-Foundation` Tenant that have *consumed* the contract provider for L3Out. The *consumed* contract enables access from both active-active data centers. HyperFlex endpoints each Pod will use the shared L3out connection from that Pod, as routing will direct the traffic via the shortest path.

Figure 20 Connectivity to Outside Networks - HyperFlex Infrastructure



Both HyperFlex clusters will leverage the connectivity above, independent of the location. The key difference in this active-active design is that the ACI Multi-Pod fabric has two independent paths to access the same services from within the fabric.

Enabling Access Layer Connectivity to HyperFlex Clusters and UCS Domains

Before any virtual server infrastructure can be deployed in the active-active datacenters, the ACI Multi-Pod fabric must provide access layer connectivity to the UCS domains and HyperFlex servers that provide the compute, storage and server networking infrastructure for each data center. The access layer connectivity includes the following:

- Physical connectivity to the UCS domains that HyperFlex clusters connect to. A Cisco UCS domain consists of a pair of UCS Fabric Interconnects. In a HyperFlex deployment, HyperFlex servers are dual-homed to the Fabric Interconnects in the UCS domain.
- Access Layer to configuration and setup to enable connectivity from the UCS domain and HyperFlex clusters to the fabric.

The access layer connectivity provided by the ACI Multi-Pod fabric for HyperFlex clusters and UCS domains are the same as that of a single site ACI fabric. The only difference here is that the connections could be in any Pod.

Connectivity to UCS Domains for HyperFlex Clusters

The physical connectivity to the UCS domains for the HyperFlex stretched cluster is shown in Figure 21 and Figure 22.

Figure 21 Connectivity to Application Cluster UCS Domain in Site-A/Pod-1

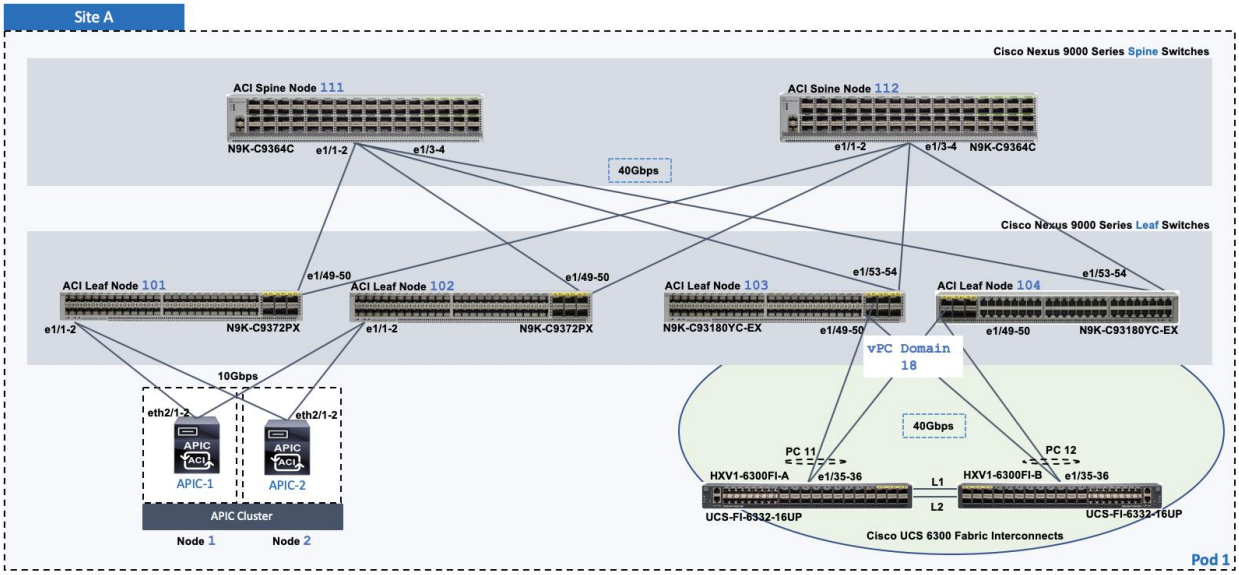
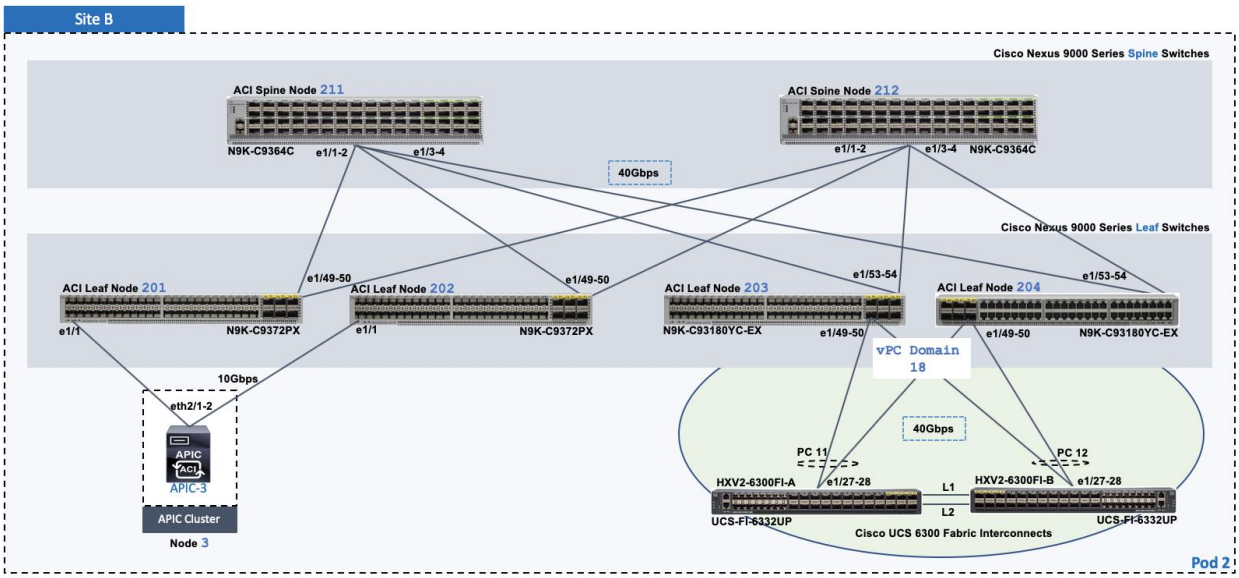


Figure 22 Connectivity to Application Cluster UCS Domain in Site-B/Pod-1



For the HyperFlex stretched cluster (Application cluster), a pair of Cisco UCS 6300 Series Fabric Interconnects are connected using 40Gbps links to a pair of Leaf switches in each Pod. In each Pod, two virtual Port Channels (vPCs) will be established from the Leaf switches to each Cisco UCS Fabric Interconnect (FI-A, FI-B). The vPC will enable link bundling to enable higher aggregate bandwidth and availability between the ACI fabric and UCS domains.

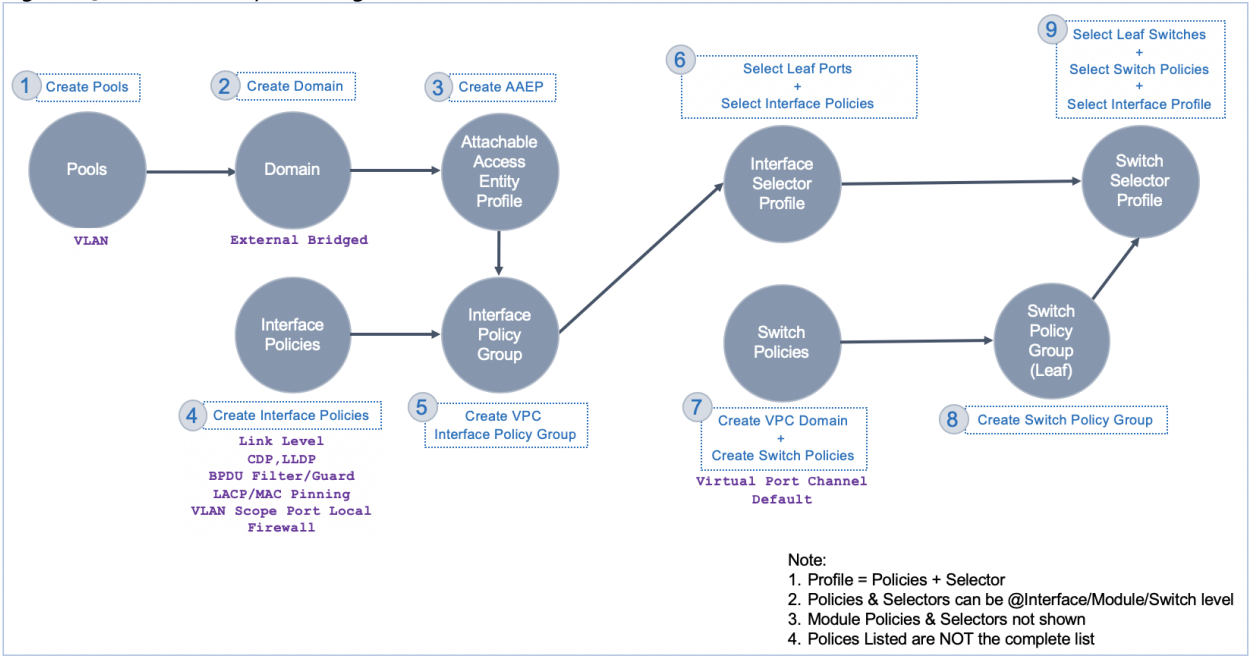
For the HyperFlex standard cluster (Management cluster), connectivity similar to the one used for the stretched cluster in Pod-1 is used. The HyperFlex standard cluster connects to a pair of Cisco 6200 Fabric Interconnects in Pod-1 and use 10GbE for the vPC links between the ACI leaf switches and Cisco UCS Fabric Interconnects.

Access Layer Design - To UCS Domain

In ACI, fabric access policies capture the access layer design for connecting to access layer devices. The workflow for connecting access layer devices in ACI involves defining policies and then applying the policies to leaf switch interfaces that

connect to access layer device. Figure 23 is a high-level workflow of the access layer setup in ACI that would create vPCs and configure access ports that connect to Cisco UCS Fabric Interconnects in each data center.

Figure 23 Access Layer Configuration Workflow



The fabric access polices that enable access layer connectivity to two UCS domains for the HyperFlex stretched cluster are provided below.

Figure 24 Access Layer Design – VLAN Pools, Domain and AAEP

vPC to UCS 6300 FIs

VLAN Pool Name	Allocation Mode	VLAN	VLAN Name	Description
HXV-UCS_VLANs	Static	118	hxv-inband-mgmt	Management (InBand) Network for ESXi Hypervisor and Storage Controller VM (SCVM) on HX nodes
		3018	hxv-vmotion	HX vMotion Network
		3218	hxv1-storage-data	HX Storage Data Network – a unique VLAN should be used for each HX cluster deployed

vPC to UCS 6300 FIs

Domain Name	Domain Type	VLAN Pool Name	Connects To
HXV-UCS_Domain	External Bridged Domain	HXV-UCS_VLANs	Cisco UCS Domain

vPC to UCS 6300 FIs

AAEP Name	Domain Name	VLAN Pool Name	Connects To
HXV-UCS_AAEP	HXV-UCS_Domain	HXV-UCS_VLANs	Cisco UCS Domain

Figure 25 Access Layer Design – Leaf Interface Profiles and Policies

vPC to UCS 6300 FIs

Interface Policy Name	Associated AAEP	Description
40Gbps-Link	HXV-UCS_AAEP	Configures link for 40Gbps
CDP-Enabled		Enables CDP
LLDP-Enabled		Enables LLDP
BPDU-FG-Enabled		Enables BPDU Guard
VLAN-Scope-Local		Configures VLAN Scope to be Local
LACP-Active		Enables LACP

vPC to UCS 6300 FIs

Interface Policy Group Name	Interface Policy Name	Associated AAEP
HXV-UCS-6300FI-A_IPG HXV-UCS-6300FI-B_IPG	40Gbps-Link CDP-Enabled LLDP-Enabled BPDU-FG-Enabled VLAN-Scope-Local LACP-Active	HXV-UCS_AAEP

vPC to 6300 FIs

Leaf Interface Profile Name	Access Port Selector	Interface Policy Group
HXV-UCS-6300FI_IPR	HXV-UCS_p1_49	HXV-UCS-6300FI-A_IPG
	HXV-UCS_p1_50	HXV-UCS-6300FI-B_IPG

Figure 26 Access Layer Design – Leaf Switch Profiles and Policies

Pod 1

vPC to UCS 6300 FIs

Switch Policy Name	VPC Explicit Protection Group	vPC Domain ID	Node ID
Virtual Port Channel default	HXV-UCS-Leaf_103-104_VPC_ExPG	18	103, 104

Pod 1

vPC to UCS 6300 FIs

Leaf Profile Name	Leaf Selectors	Leaf Interface Profile
HXV-UCS-Leaf_103-104_IPR	HXV-UCS-Leaf_103-104	HXV-UCS-6300FI_IPR

Pod 2

vPC to UCS 6300 FIs

Switch Policy Name	VPC Explicit Protection Group	vPC Domain ID	Node ID
Virtual Port Channel default	HXV-UCS-Leaf_203-204_VPC_ExPG	18	203, 204

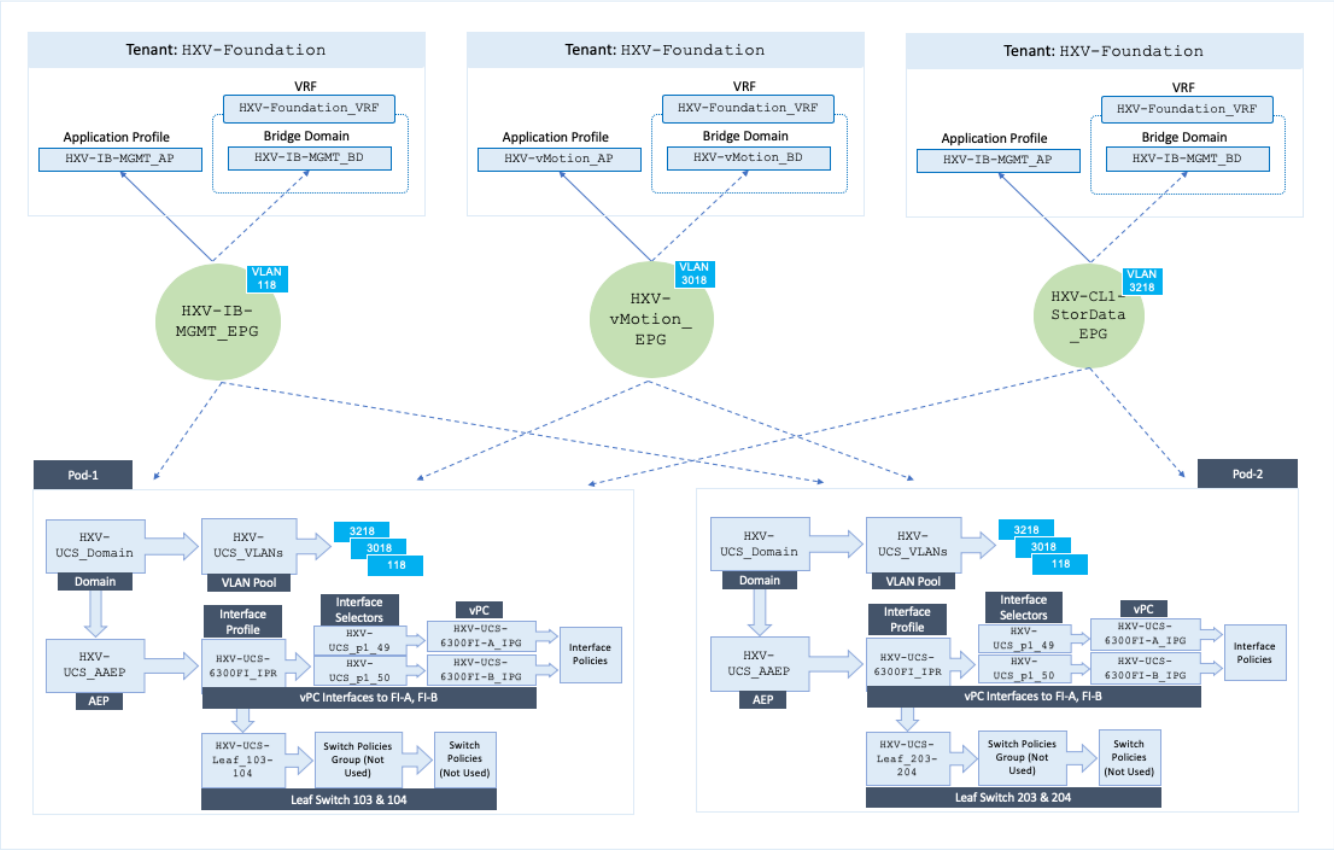
Pod 2

vPC to UCS 6300 FIs

Leaf Profile Name	Leaf Selectors	Leaf Interface Profile
HXV-UCS-Leaf_203-204_IPR	HXV-UCS-Leaf_203-204	HXV-UCS-6300FI_IPR

Defining the fabric access policies for the UCS domain results in the following (see Figure 27) relationships between the endpoint group and access layer connection when the EPG is deployed. It will also result in access layer connectivity using vPCs to both UCS domains in the HyperFlex stretched cluster.

Figure 27 Fabric Access Policies - To Cisco UCS Domain



A similar set of fabric access policies are used for the access layer connectivity to the UCS domain for the HyperFlex standard cluster (Management).

Integration with Virtual Machine Manager

The Virtual Machine Manager (VMM) integration enables Cisco APIC to control and manage the creation and operation of distributed virtual switches running on ESXi hosts. In this design, the APIC integrates with VMware vCenter to manage the creation and operation of virtual distributed switches running on ESXi hosts. To deploy new application virtual machines, endpoint groups can be created in the ACI fabric and APIC will dynamically allocate VLAN for the new EPG and create a corresponding port-group for that EPG in the VMM domain, with network policies applied. Virtual machines can now be added to that EPG. Multiple EPGs and port-groups can be created this way.

The APIC can integrate with vCenter to manage a VMware vSphere Distributed Switch (vDS) or a Cisco ACI Virtual Edge (AVE). This design uses an APIC-controlled Cisco ACI Virtual Edge for guest (application) virtual machines hosted on Applications cluster and VMware vDS in the Management cluster. In a HyperFlex deployment, infrastructure endpoint groups and VLANs are deployed on VMware vSwitch and these VLANs will remain on the vSwitch. However, the VM network VLANs will be migrated to vDS or Cisco AVE in this design.

For VMM integration to work correctly with HyperFlex ESXi hosts, the following need to be setup correctly:

- Virtual NICs on hosts must be configured to use either CDP or LLDP; only one can be enabled at a time
- VLAN pool for the VMM domain must be enabled on the HyperFlex host's virtual NIC for VM networks and on the Fabric Interconnect uplinks connecting to the leaf switches.

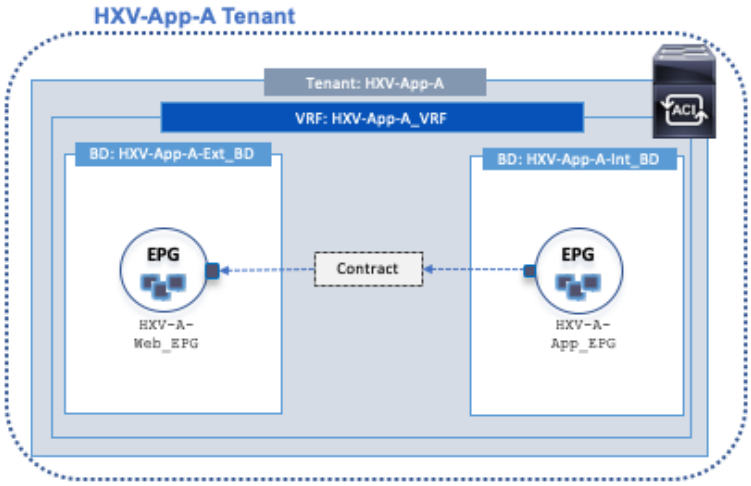
HyperFlex Installer can take care of the VLAN pool configuration if it is known during the install process. It can also be created using the HyperFlex post-install script.

Onboarding Applications

Once the infrastructure and virtualization setup necessary for hosting Applications is complete, application can be deployed on the Application HyperFlex cluster in either datacenter. In this design, a multi-tier application is deployed where ACI must explicitly allow access between the different tiers of the application.

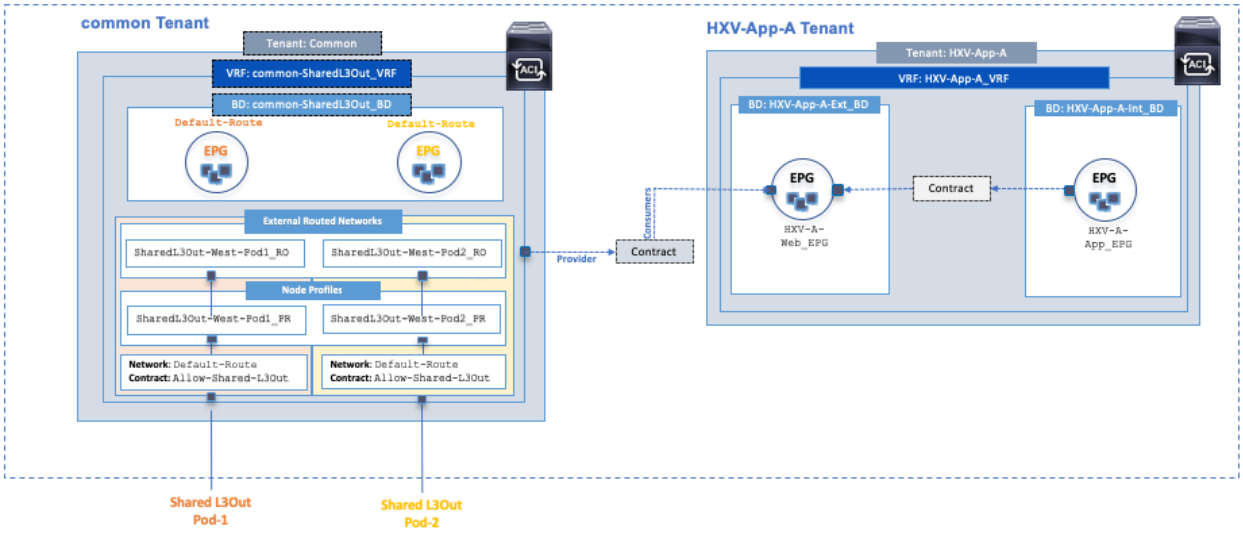
Figure 28 shows the ACI constructs used for deploying a sample two-tier application in this design.

Figure 28 Application Tenant



To enable access to the shared L3Out service provided by **common** Tenant, the contract *provided* for the L3Out is *consumed* by Application Tenant as shown in Figure 29.

Figure 29 Connectivity to Outside Networks - Applications



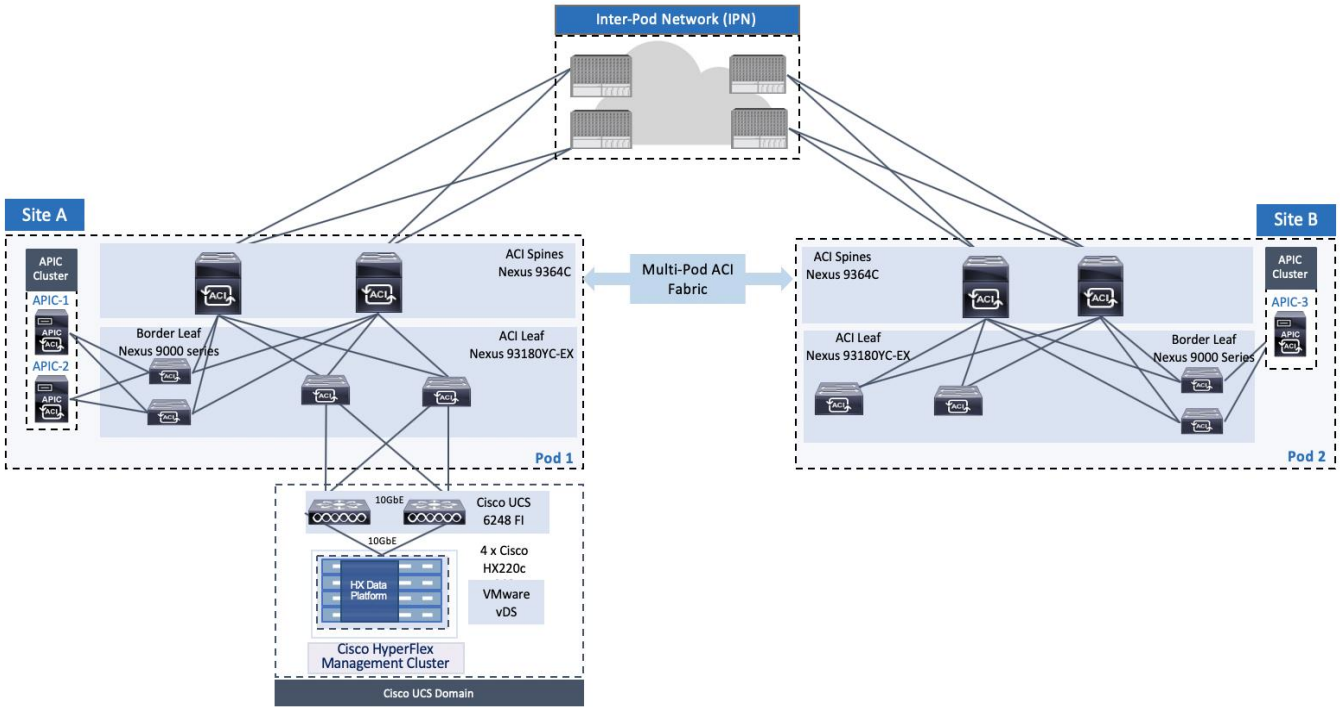
HyperFlex Virtual Infrastructure Design

Two HyperFlex clusters provide the virtual server infrastructure in this design – one to host Management virtual machines (optionlao and another for Application workloads. The HyperFlex clusters and UCS domains in the ACI Multi-Pod fabric are all centrally managed from the cloud using Cisco Intersight.

Management HyperFlex Cluster

The Management cluster is optional in this design as the services hosted on this cluster can be provided from a customer's existing infrastructure or through other methods. In this design, the Management cluster serves as a dedicated cluster for hosting services necessary for managing and operating other HyperFlex and UCS infrastructure directly from within the ACI fabric. Infrastructure and management services hosted outside the ACI fabric are also used in this design. ACI provides direct access to these services from each Pod using the shared L3Out connection in each data center location. A HyperFlex standard cluster is used for the Management cluster in Pod-1 as shown in Figure 30.

Figure 30 Management HyperFlex Cluster

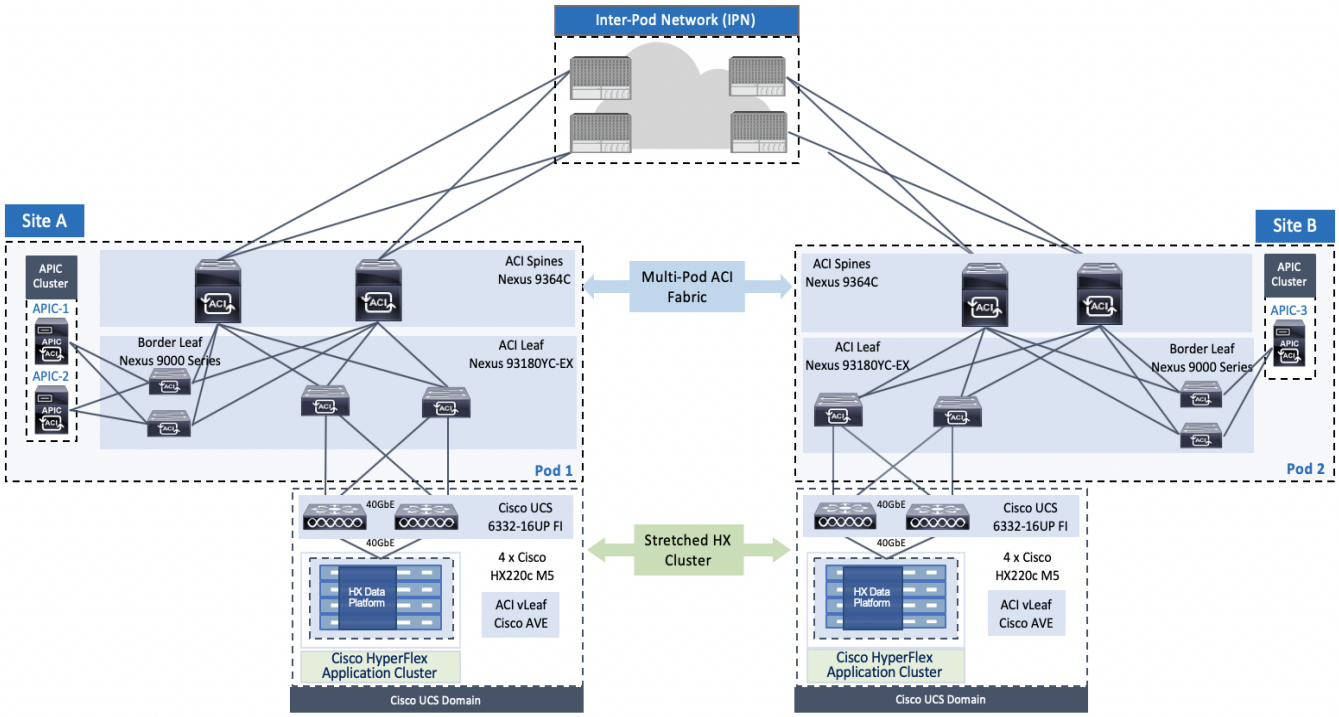


The Management cluster is deployed from the cloud using Cisco Intersight. The HyperFlex nodes in the management cluster and the virtual machines hosted on the cluster can access networks and services outside the ACI fabric using the shared L3Out connection in Pod-1.

Application HyperFlex Cluster

The Application cluster in this design spans both data centers, enabling application virtual machines to be deployed in either data center with seamless connectivity and mobility if needed. A HyperFlex stretched cluster is used for the Application cluster. The nodes in the stretched cluster are evenly distributed between Pod-1 and Pod-2 and provides the virtual server infrastructure for the active-active data centers as shown in Figure 31 .

Figure 31 Application HyperFlex Cluster



The stretched cluster is deployed using an Installer VM running on the Management HyperFlex cluster. The Pod-1 HyperFlex nodes in the application cluster and the virtual machines hosted on it can access networks and services outside the ACI fabric using the shared L3Out connection in Pod-1. Similarly, Pod-2 HyperFlex nodes in the application cluster and the virtual machines hosted on it can access networks and services outside the ACI fabric using the shared L3Out connection in Pod-2.

Cisco UCS Networking Design

The HyperFlex clusters in this design use dedicated UCS domains to connect to the ACI fabric. The HyperFlex stretched cluster uses a dedicated UCS domain per Pod to connect the nodes in that Pod. A UCS domain consists of a pair of Cisco UCS 6x00 series Fabric Interconnects and the servers that connect to it. A single Cisco UCS domain can support multiple HyperFlex clusters, the exact number depends on the size of the cluster and the port-density on the Fabric Interconnect model chosen. Cisco UCS manager that manages the Cisco HyperFlex and UCS servers in the UCS domain, runs on the Fabric Interconnects. In this design, the UCS domains and the associated HyperFlex clusters are managed from the cloud using Cisco Intersight. Cisco Intersight offers centralized management of all Cisco UCS and HyperFlex infrastructure in an Enterprise.

The HyperFlex clusters in each Pod connect to the ACI fabric in that location through a pair of Cisco UCS 6x00 series Fabric Interconnects.

Unified Fabric – Cisco UCS Fabric Interconnects

Cisco UCS Fabric Interconnects (FI) are an integral part of the HyperFlex system. The fabric interconnects providing a unified fabric for integrated LAN, SAN and management connectivity for all HyperFlex servers that connect to the Fabric Interconnects. Fabric Interconnects provide a lossless and deterministic switching fabric, capable of handling I/O traffic from hundreds of servers.

Cisco UCS Fabric Interconnects are typically deployed in pairs to form a single management cluster but with two separate network fabrics, referred to as Fabric A or FI-A and Fabric B or FI-B. Cisco UCS Manager that manages the UCS domain, runs on the Fabric Interconnects. In a UCS domain, one FI is the primary, and the other is the secondary. Each FI has its own

IP address and a third roaming IP that serves as the cluster IP address for management. This primary/secondary relationship is only for the management cluster and has no effect on data transmission. The network fabric on both Fabric Interconnects are active at all times, forwarding data on both network fabrics while providing redundancy in the event of a failure. A HyperFlex cluster connects to the ACI fabric through a UCS domain, with every node in the cluster connecting to both Fabric Interconnects in the Cisco UCS domain.

The Fabric Interconnect model used in a UCS domain will determine the link speeds that can be used for connecting upstream to the ACI fabric and downstream to the servers. Two Fabric Interconnect models are used in this design though other models and uplinks are also supported.

- Cisco UCS 6200 series fabric interconnects provide a 10GbE unified fabric with 10GbE uplinks for northbound connectivity to the ACI fabric and 10GbE downlinks for southbound connectivity to HyperFlex servers.
- Cisco UCS 6300 series fabric interconnects provide a 40GbE unified fabric with 40GbE uplinks for northbound connectivity to the ACI fabric and 40GbE downlinks for southbound connectivity to HyperFlex servers.

For higher bandwidth and resiliency, the Fabric Interconnects connect to the ACI fabric using multiple 40GbE or 10GbE links in a virtual Port-channel (vPC) configuration.

Uplink Connectivity to Data Center Network Fabric

The Cisco UCS Fabric Interconnects in this design connect to Nexus 9000 series leaf switches in the Cisco ACI fabric and provide uplink or northbound connectivity to other parts of the Enterprise. Multiple UCS domains are used in this design to support the Management and Applications cluster. The Application cluster uses a HyperFlex stretched cluster with 2 pairs of UCS Fabric Interconnects, one in each site, to provide connectivity to the HyperFlex nodes in that site.

For redundancy and bandwidth, multiple links from each FI are used for uplink connectivity to data center fabric. Cisco UCS FI supports 802.3ad standards for aggregating links into a port-channel (PC) using Link Aggregation Protocol (LACP). Multiple links on each FI are bundled together in a port-channel and connected to upstream switches in the data center network. The port-channel provides link-level redundancy and higher aggregate bandwidth for LAN, SAN and Management traffic to/from the UCS domain. The switches in the data center fabric that connect to single FI are bundled into a virtual Port Channel (vPC). vPC enables links that are physically connected to two different switches to be bundled such that it appears as a "single logical" port channel to a third device (in this case, FI). This PC/vPC based design has many benefits such as:

- Higher resiliency - both link and node-level redundancy
- Higher uplink bandwidth by bundling links
- Flexibility to increase the uplink bandwidth as needed by adding more links to the bundle.

All uplinks on the Cisco UCS FIs operate as trunks, carrying multiple 802.1Q VLAN IDs across the uplinks. And all VLAN IDs defined on Cisco UCS should be trunked across all available uplinks. This is important as traffic may need to be forwarded between servers in the UCS domain but use different fabric (FI-A, FI-B) as its primary data path. There are also failure scenarios where a VIC or an internal fabric level port or link failure results in traffic that normally does not leave the Cisco UCS domain, to now be forced over the Cisco UCS uplinks for intra-domain connectivity. Reachability through the second fabric may also be needed for maintenance events such as FI firmware upgrade that may require a fabric to be rebooted.

UCS Networking Design for Application Cluster

The Application cluster consists of 4+4 node HyperFlex stretch cluster that connects to the ACI fabric in each Pod through a pair of Cisco UCS 6332 Fabric Interconnects. The stretched cluster connectivity to the ACI fabric in Pod-1 and Pod-2 are shown in Figure 32 and Figure 33 respectively.

Figure 32 UCS Networking in Pod-1 for Application Cluster

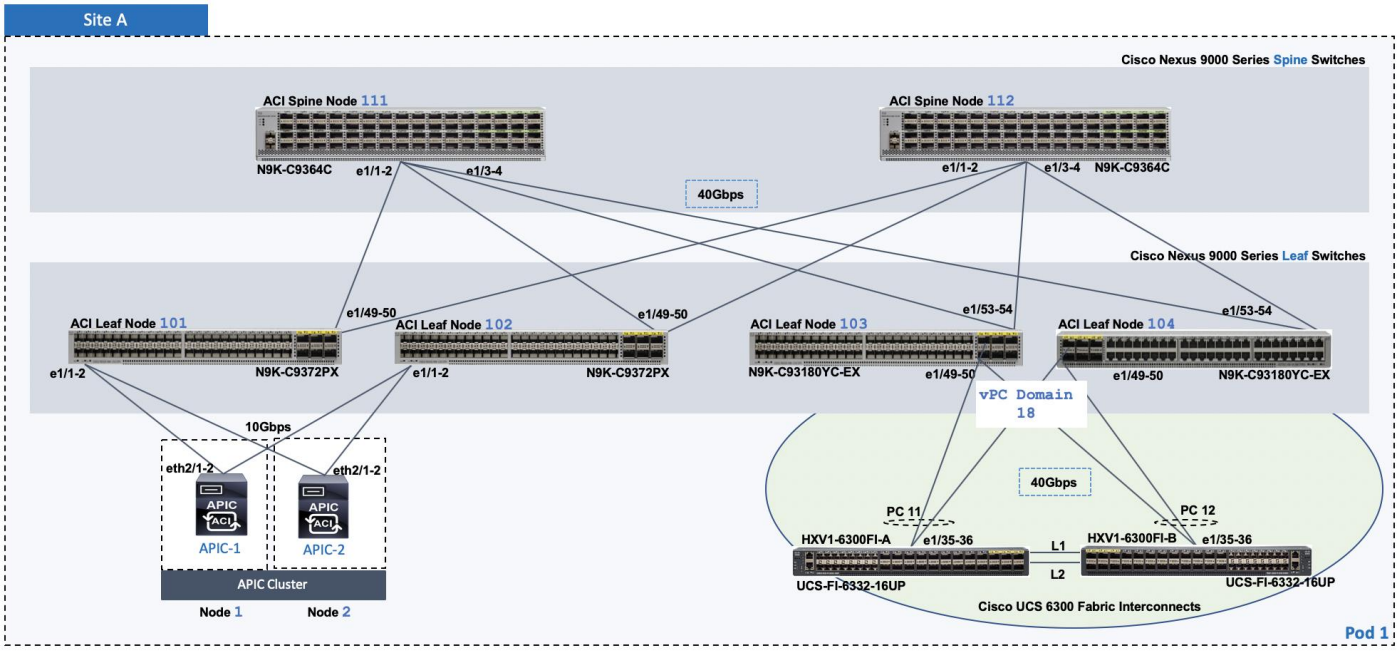
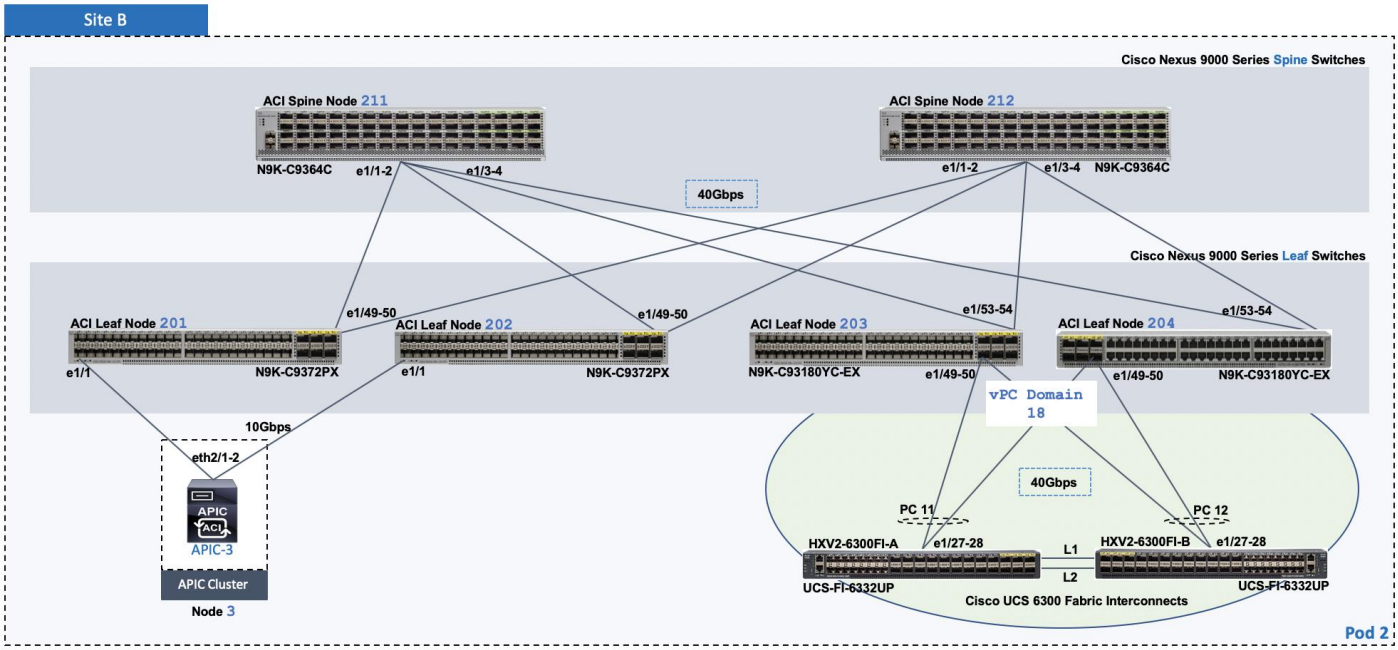


Figure 33 UCS Networking in Pod-2 for Application Cluster



To connect to the upstream ACI fabric, the Cisco UCS 6300 Fabric Interconnects are connected to a pair of upstream Nexus 9000 ACI leaf switches as follows:

- 2 x 40GbE links from FI-A to Nexus leaf switches, one to each Leaf switch
- 2 x 40GbE links from FI-B to Nexus leaf switches, one to each Leaf switch

The FI side ports are configured to be a port-channel, with vPC configuration on the Nexus leaf switches. The two links from separate Nexus 9000 leaf switches in the ACI fabric that connect to a specific FI is configured to be part of the same vPC.

The above connectivity provides the UCS domain with redundant paths and 160Gbps (40Gbps per link x 2 uplinks per FI x 2 FI) of aggregate uplink bandwidth to/from the ACI fabric. The uplink bandwidth can be increased as needed by adding additional connections to the port-channel.

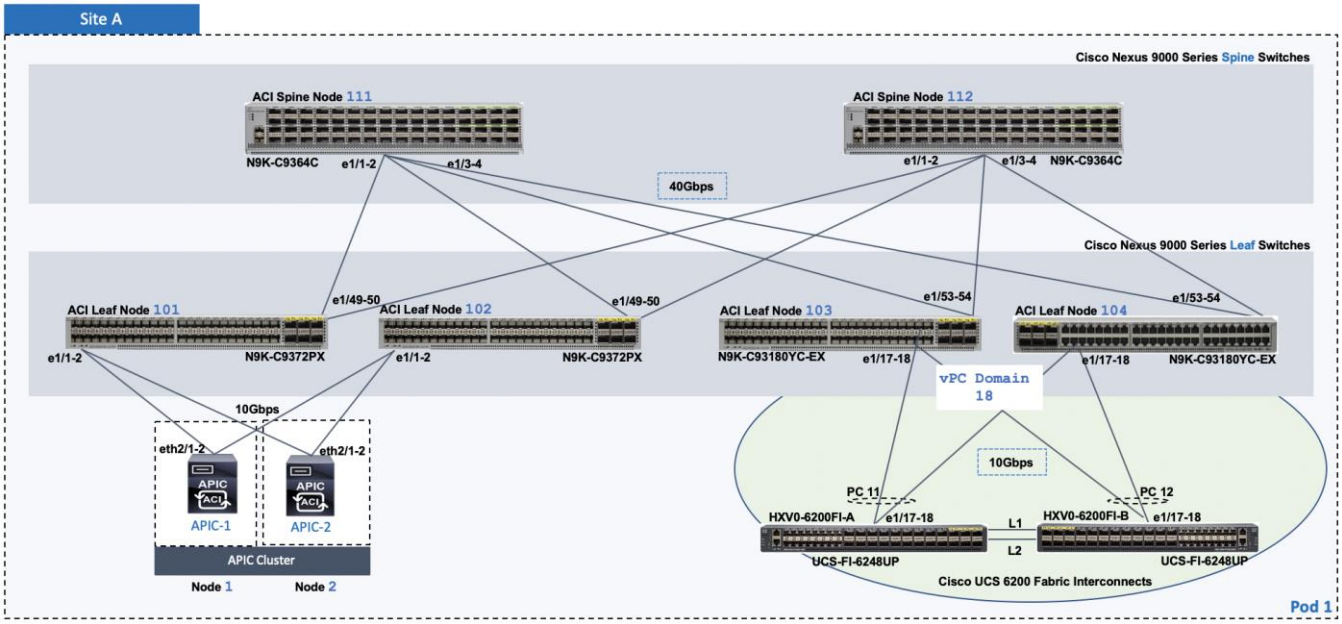
The VLANs for in-band management, vMotion, storage data and VM network vlans are enabled on the uplinks of Fabric Interconnects in both data centers. The VLANs are also enabled on the individual vNIC templates going to each server in the HX cluster.

Each server in the cluster uses a VIC 1387 adapter with two 40Gbps uplink ports to connect to each FI, forming a path through each fabric (FI-A, FI-B). The two uplink ports are bundled in a port-channel to provide 2x40Gbps of uplink bandwidth from each server and redundancy in the event of a failure.

UCS Networking Design for Management Cluster

The Management cluster consists of 4-node HyperFlex standard cluster that connects to the ACI fabric in Pod-1 through a pair of Cisco UCS 6200 series Fabric Interconnects as shown in Figure 34.

Figure 34 UCS Networking in Pod-1 for Management Cluster



The FI side ports are configured to be a port-channel, with vPC configuration on the Nexus leaf switches. The two links from separate Nexus 9000 leaf switches in the ACI fabric that connect to a specific FI is configured to be part of the same vPC.

The above connectivity provides the UCS domain with redundant paths and 40Gbps (10Gbps per link x 2 uplinks per FI x 2 FI) of aggregate uplink bandwidth to/from the ACI fabric. The uplink bandwidth can be increased as needed by adding additional connections to the port-channel.

Each server in the cluster uses two 10Gbps uplink ports to connect to each FI, forming a path through each fabric (FI-A, FI-B). The two uplink ports are bundled in a port-channel to provide 2x10Gbps of uplink bandwidth from each server and redundancy in the event of a failure.

The VLANs for in-band management, vMotion, storage data and VM network vlans are enabled on the uplinks of Fabric Interconnects in both data centers. The VLANs are also enabled on the individual vNIC templates going to each server in the HX cluster.

Connectivity for HyperFlex Installation

To deploy a HyperFlex system, the HyperFlex installer (HyperFlex Installer VM or Cisco Intersight) requires the following connectivity to the UCS domain where the servers are.

- IP connectivity to the management interfaces of both Fabric Interconnects – this is typically an out-of-band network dedicated for management.
- IP connectivity to the external management IP address of each server in the HX cluster. This IP address comes from an IP Pool (**ext-mgmt**) defined as a part of the service profile template for configuring the servers in the cluster. The IP address is assigned to the CIMC interface on each server which is reachable through the out-of-band management network of the Fabric Interconnects.

The out-of-band network that provides the above connectivity is not part of the ACI fabric in this design. However, connectivity between the ACI fabric in each Pod and the out-of-band network is through the Shared L3Out connection in that Pod.

Other Considerations – Jumbo Frames

HyperFlex uses jumbo frames for storage and vMotion. Jumbo frames should therefore be enabled end-to-end, across the ACI Multi-Pod fabric to prevent service interruptions or during planned maintenances such as Cisco UCS firmware upgrades, or when a cable or port failure would cause storage traffic to traverse northbound to the ACI fabric.

Cisco HyperFlex Networking Design

The Cisco HyperFlex system requires multiple VLANs and IP subnets to enable connectivity between the different sub-components within the HyperFlex system. In a HyperFlex stretched cluster, the installer will deploy identical VLANs and networking in both data center locations since it is essentially a single cluster.

The networking required in a HyperFlex system are covered in this section.

HyperFlex VLANs and Subnets

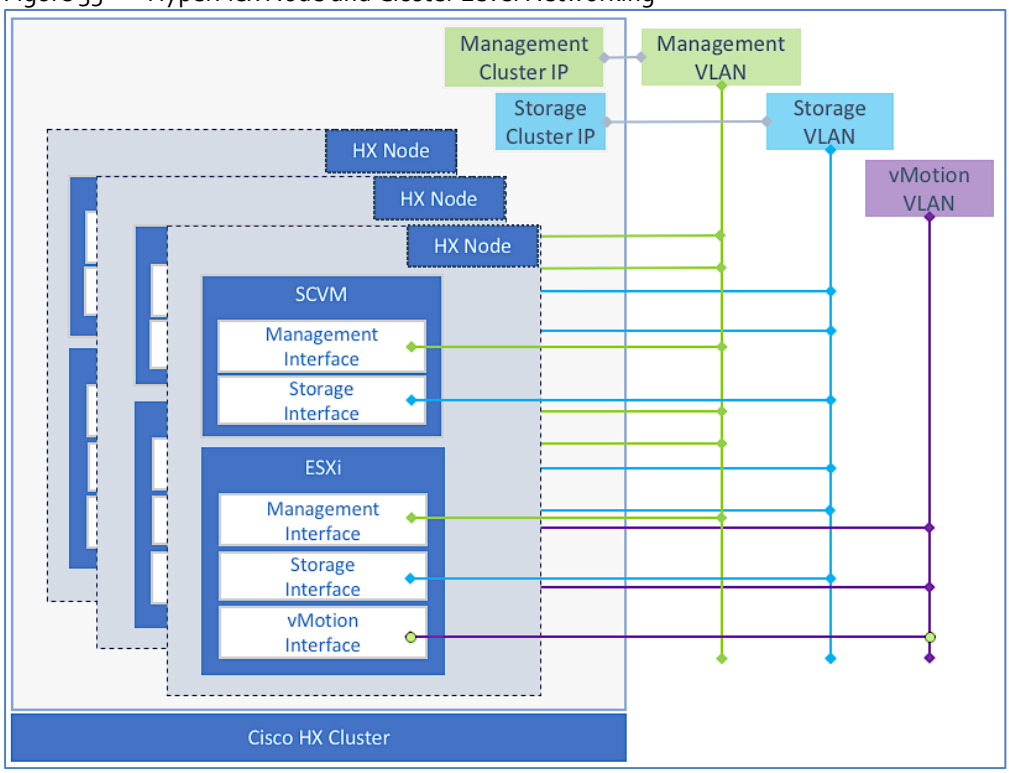
The networking required within a HyperFlex system can be categorized into the following groups:

- **Management:** The HyperFlex system requires network connections to manage the ESXi hosts and the storage platform controller virtual machines (SCVM) in the cluster. The SCVM management interfaces required for a HX cluster includes following interfaces:
  - SCVM Management Interfaces
  - SCVM Replication Interfaces
  - Roaming HX Cluster Management IP – one per HX cluster
- **Storage:** The HyperFlex system requires network connections to manage the storage sub-system in a given HX cluster. These connections are used by the Cisco Data Platform software (HXDP) to manage the HX Distributed File System. The storage interfaces required for a HX cluster are:
  - SCVM Storage Interface
  - Roaming HX Cluster Storage IP – one per HX cluster
  - VMkernel interface for storage access on each ESXi host in the HX cluster
- **vMotion:** To enable vMotion for guest VMs on ESXi hosts in the HX cluster, a VMkernel interface for each ESXi host in the HX cluster is required.

- **Virtual Machines:** The HX system will need to provide connectivity to the guest virtual machines deployed on hosts in the HX cluster so that they can be accessed by other VMs and users in the organization.

The HyperFlex networking for Management, Storage and vMotion at the HX node and cluster level are shown in Figure 35.

Figure 35 HyperFlex Node and Cluster Level Networking



HyperFlex Networking Guidelines and Best Practices

The design guidelines and best practices for HyperFlex networking and their use in this design are as follows.

- The HyperFlex storage data must be on a dedicated network and therefore must be in a separate VLAN, should not be used for other traffic.
- The HyperFlex installer, during installation, requires reachability to the components in the UCS domain, specifically the Fabric Interconnect's cluster management IP address and the external management (**ext-mgmt)** IP addresses of the UCS servers (HX nodes). Cluster management IP is a roaming IP that is assigned to the primary Fabric Interconnect and both the cluster IP and the external management IP of the HX nodes are reachable via the dedicated management ports on each FI.
- All storage and vMotion traffic in a HyperFlex system is configured to use jumbo frames by default. Jumbo frames enable IP traffic to use a Maximum Transmission Unit (MTU) size of 9000 bytes. Larger MTU value enables each IP packet to carry a larger payload, therefore transmitting more data per packet, and consequently sending and receiving data faster. The HyperFlex installer will configure the uplinks (vNICs) on all servers in the HX cluster to use a jumbo frame MTU for storage and vMotion. Links end-to-end must also be configured for jumbo frames.
- Replication Networking is setup after the initial install by the installer or Cisco Intersight. Replication was not validated in this design. For a detailed discussion on HyperFlex Replication – see the **Cisco HyperFlex 3.0 for Virtual Server Infrastructure with VMware ESXi** design guide listed in the References section.

Validated Design – HyperFlex VLANs and Subnets

Figure 36 shows the VLANs that the HyperFlex Installer (HyperFlex Installer VM or Cisco Intersight) uses to provision the HyperFlex stretched cluster, HyperFlex standard cluster, UCS domains for each cluster and the vSphere environment. For each network type listed, HyperFlex will create a corresponding VMware virtual switch (vSwitch) on each ESXi host in the HyperFlex cluster. The installer will also provision the virtual switches with port-groups for each of the VLANs listed. These VLANs are also configured in the UCS Fabric Interconnects on its two uplinks to the ACI fabric and on the vNICs to the HyperFlex nodes.

If replication is enabled (to a second cluster), a VLAN will need to be allocated for this. The replication VLAN will map to a port-group on the inband management vSwitch. Replication networking is not part of the initial automated install of the HX cluster. Replication was not validated in this design.

Figure 36 HyperFlex VLANs

Network Type	VLAN Name	VLAN	HyperFlex Networks
In-Band Management Network	hxv-inband-mgmt	118	ESXi Hypervisor & Storage Controller VM (SCVM) Management
vMotion Network	hxv-vmotion	3018	vMotion
Storage Data Network	hxv0-storage-data	3118	Storage Data Network – for Management HX cluster
Storage Data Network	hxv1-storage-data	3218	Storage Data Network – for Application HX cluster
VM Network	hxv-vm-network-1118 hxv-vm-network-1218	1118 - 1218	HX VM Network – on Management HX cluster
VM Network	hxv-vm-network-2118 hxv-vm-network-2218	2118 - 2218	HX VM Network – on Application HX cluster

The HyperFlex Installer uses the VM Network VLAN pool to create VLANs in the UCS Fabric and port-groups in vSphere. Additional VLANs can be added for VM network as needed. The VM network VLANs are initially mapped to port-groups on a VMware virtual switch but they can be migrated to an APIC-controlled VMware virtual distributed switch (vDS), Cisco AVS, or Cisco ACI Virtual Edge after the initial install.

The infrastructure (4093) required for VxLAN tunneling to the APIC-controlled Cisco ACI Virtual Edge used in this design, is provisioned directly from Cisco UCS Manager, after the initial HyperFlex install.

Virtual Networking Design

The HyperFlex installer deploys the Cisco HyperFlex system with a pre-defined virtual networking design on the ESXi hosts in the cluster. The virtual networking for a HyperFlex stretched cluster is identical across all hosts in the cluster regardless of their location. The design segregates the different types of traffic through the HyperFlex system using different VMware virtual switches (vSwitch). Four virtual switches are created by the HyperFlex installer as summarized below. Each vSwitch is assigned two uplinks – the uplink adapters seen by the host at the hypervisor level are virtual NICs (vNICs) created on the VIC converged network adapter installed on the HX server. The vNICs for each server are created in Cisco UCS Manager using service profiles. Installer creates the vNICs as well.

The virtual Switches created by the installer are:

vswitch-hx-inband-mgmt: This is the default ESXi vSwitcho which is renamed by the ESXi kickstart file as part of the automated installation. The switch has two uplinks, active on fabric A and standby on fabric B – jumbo frames are not enabled on these uplinks. The following port groups are created on this switch:

- Port group for the standard ESXi Management Network. The default ESXi VMkernel port: **vmk0**, is configured as a part of this group on each ESXi HX node.
- Port Group for the HyperFlex Storage Platform Controller Management Network. The SCVM management interfaces is configured as a part of this group on each HX node.
- If replication is enabled across two HX clusters, a third port group should be deployed for VM snapshot replication traffic.

The VLANs associated with the above port-groups are all tagged VLANs (not native VLANs) in Cisco UCS vNIC templates. Therefore, the VLANs are also explicitly configured in ESXi/vSphere.

vswitch-hx-storage-data: This vSwitch is created as part of the automated installation. The switch has two uplinks, active on fabric B and standby on fabric A – jumbo frames are enabled on these uplinks (recommended):

- Port group for the ESXi Storage Data Network. The ESXi VMkernel port:**vmk1** is configured as a part of this group on each HX node.
- Port group for the Storage Platform Controller VMs. The SCVM storage interfaces is configured as a part of this group on each HX node.

The VLANs associated with the above port-groups are all tagged VLANs (not native VLANs) in Cisco UCS vNIC templates. Therefore, the VLANs are also explicitly configured in ESXi/vSphere.

vswitch-hx-vm-network: This vSwitch is created as part of the automated installation. The switch has two uplinks, active on both fabrics A and B – jumbo frames are not enabled on these uplinks. The VLANs associated with the above port-groups are all tagged VLANs (not native VLANs) in Cisco UCS vNIC templates. Therefore, the VLANs are also explicitly configured in ESXi/vSphere.

vmotion: This vSwitch is created as part of the automated installation. The switch has two uplinks, active on fabric A and standby on fabric B – jumbo frames are enabled on these uplinks (recommended). The IP addresses of the VMkernel ports (vmk2) are configured by using post_install script. The VLANs associated with the above port-groups are all tagged VLANs (not native VLANs) in Cisco UCS vNIC templates. Therefore, the VLANs are also explicitly configured in ESXi/vSphere.

Once the cluster is operational, additional virtual machine networks may need to be added to roll out new applications and services. This requires the VLANs to be provisioned in the ACI fabric, on the Fabric Interconnects and on the ESXi hosts in the cluster through APIC, UCSM and vCenter respectively. To minimize some of the provisioning, ACI provides VMM integration with VMware vCenter. As new applications and services are deployed on the ACI side, APIC will leverage this integration to dynamically allocate a VLAN for use by new application or service outside the ACI fabric and provision the corresponding virtual-networking through vCenter. To enable this dynamic provisioning, APIC will need to manage the virtual switching – using either a VMware virtual distributed switch (vDS) or Cisco ACI Virtual Edge (AVE). APIC will also require a pre-defined pool to allocate VLANs from.

In this design, multiple virtual networking options are used as outlined below:

- VMware vSphere vSwitch is used for critical infrastructure and management services. This includes in-band management (`vswitch-hx-inband-mgmt`), storage data (`vswitch-hx-storage-data`) and VM migration (`vmotion`) networks.
- VMM integration using APIC-controlled VMware vDS is used for non-HyperFlex infrastructure and management services. This includes VM networks for services such as VMware vCenter and HyperFlex Installer. In this design, these services are hosted on the Management HyperFlex cluster and are used to deploy and manage other HX

clusters in the ACI fabric. The Installer and vCenter in the Management cluster will be used to deploy and manage the stretched HyperFlex cluster. The VM networks and the vSwitch (`hx-vm-network`) provisioned by the installer and the post-install script will be deleted in order to migrate the uplinks to the APIC-controlled vDS.

- VMM integration using APIC-controlled Cisco AVE for application VM networks. In this design, the stretched HyperFlex cluster is used to host Applications that users access. The VM networks and the vSwitch deployed by the Installer and the post-install script will be deleted in order to migrate the uplinks to the APIC-controlled Cisco AVE.

Design Options

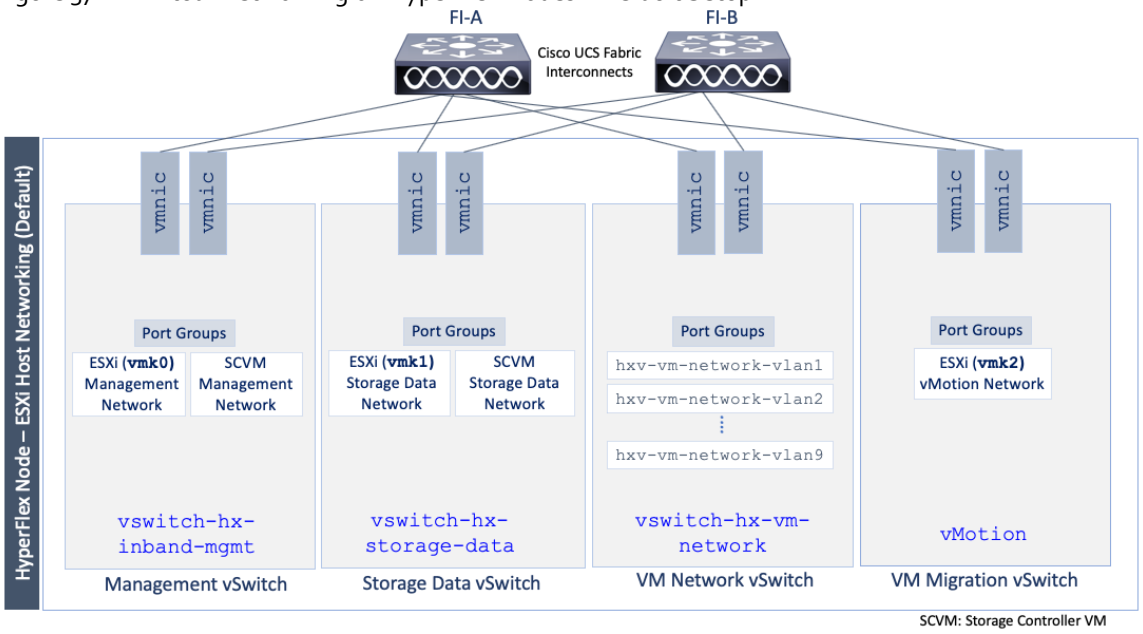
This section discusses the virtual networking options available in this design.

Virtual Networking Using VMware vSphere vSwitch

Cisco HyperFlex Installer deploys the following default virtual network design on each ESXi host in the HyperFlex cluster, this is the **default** setup. To support this multi-vSwitch environment, the HX Installer will user service profiles on Cisco UCS Manager to configure each HyperFlex host with multiple virtual NIC (vNIC) interfaces which are then used as uplinks for each vSphere vSwitch. Cisco UCS Manager then enables the VLANs for management, storage, vMotion, and application traffic on the appropriate vNIC interfaces.

Figure 37 shows the default virtual networking deployed by the HX Installer on ESXi hosts.

Figure 37 Virtual Networking on HyperFlex Nodes – Default Setup



Virtual Networking Using VMware vSphere vDS or Cisco AVE

Cisco ACI can also manage the virtual networking in a Cisco HyperFlex environment using a APIC controlled VMware vSphere Distributed Switch (vDS) or Cisco ACI Virtual Edge (AVE). The default virtual networking deployed on Cisco HyperFlex systems can be converted to use either a VMware vDS or Cisco AVE instead of the default VMware virtual switches deployed by the HyperFlex installer.

Figure 38 Virtual Networking on HyperFlex Nodes – using VMware vDS

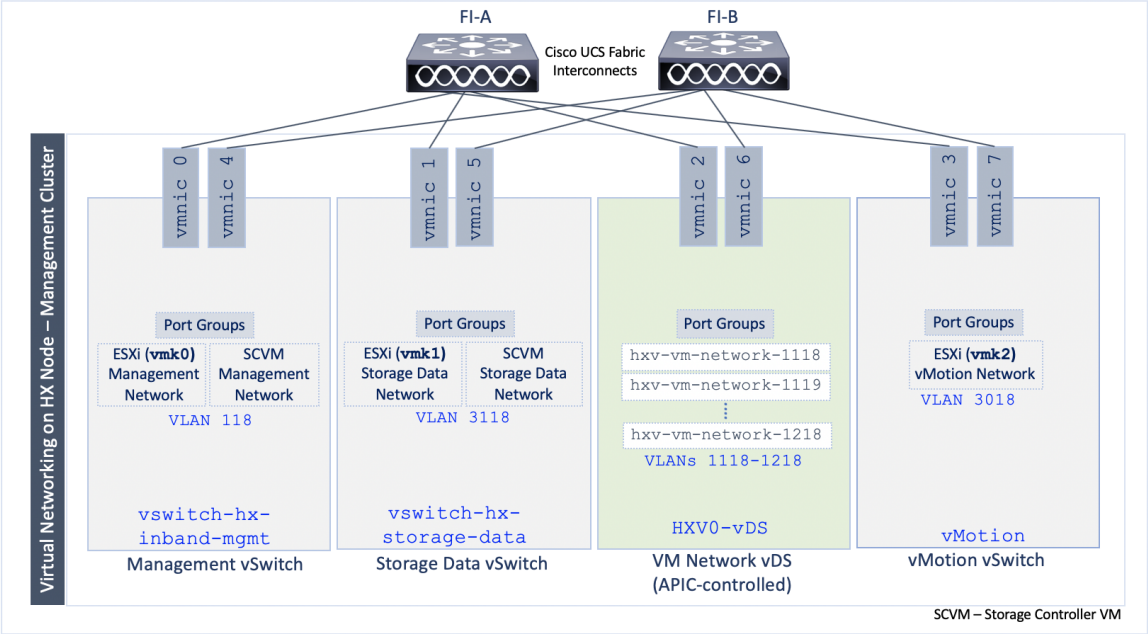
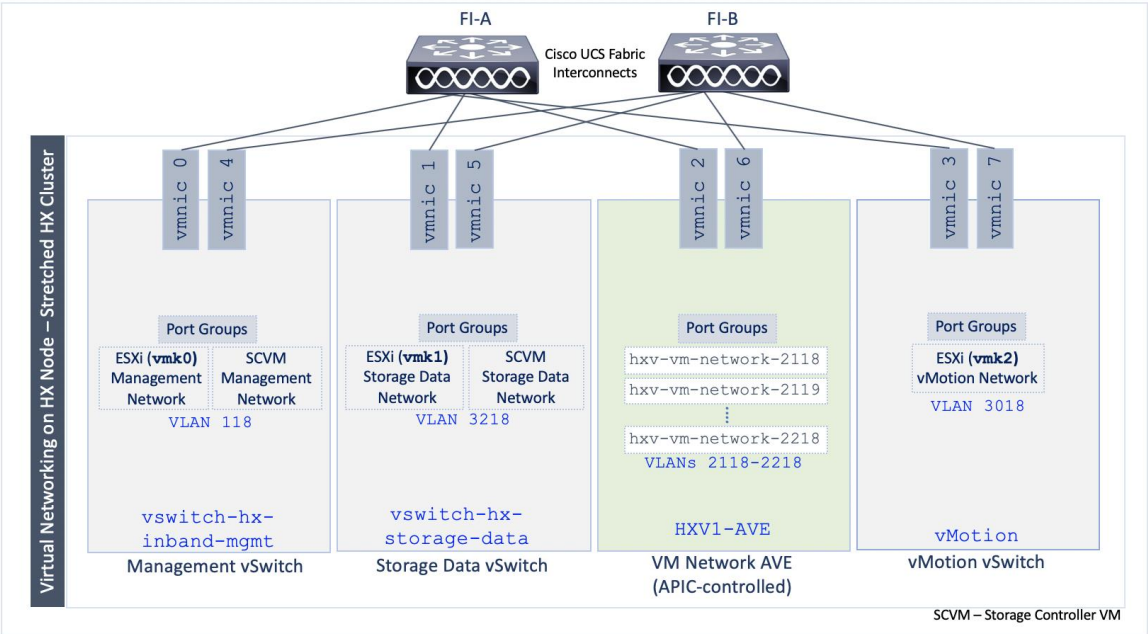


Figure 39 Virtual Networking on HyperFlex Nodes – using Cisco AVE



This design uses Cisco AVE for the guest VM (application or services) networking but all infrastructure connectivity remains on VMware vSwitch(s). The infrastructure access is maintained on VMware vSwitch(s) to ensure access to management, storage and vMotion on the host to maximize availability under failure scenarios that result in connectivity issues between host and VMware vCenter.

Cisco ACI Virtual Edge – for Application Tenant Traffic

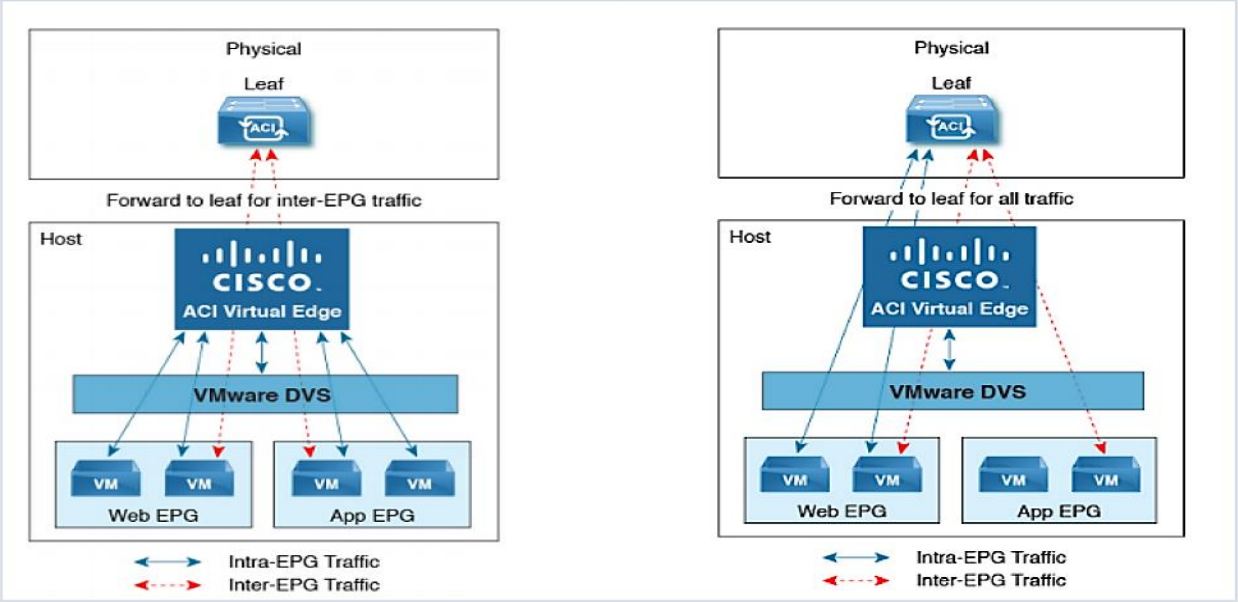
Cisco ACI Virtual Edge (AVE) is the next generation of application virtual switches for Cisco ACI environments. Unlike, AVS, Cisco AVE is hypervisor-independent, runs in the user-space, and leverages the native VMware vSphere Distributed Switch (VDS) to operate. Cisco AVS is purpose-built for Cisco ACI and operates as a virtual leaf (vLeaf) switch, managed by Cisco APIC. Cisco AVE is supported as of Cisco APIC release 3.1(1i) and VMware vCenter Server 6.0. Cisco AVE

Cisco AVE is a distributed services VM that leverages the hypervisor-resident VMware vDS and uses OpFlex protocol for control plane communication. Cisco AVE supports two modes of operation for data forwarding: **Local Switching** and **No Local Switching**.

Local Switching Mode: In this mode, Cisco AVE forwards all intra-EPG traffic locally but all inter-EPG traffic is forwarded to the leaf switch. Also, this mode supports both VLAN and VXLAN encapsulation for forwarding traffic to the leaf. The encapsulation type is specified during VMM integration, when Cisco AVE VMM domain is created. You can also specify that both encapsulations be used in a given VMM domain.

No Local Switching Mode: In this mode, Cisco AVE forwards all traffic (intra-EPG and inter-EPG traffic) to the leaf switch. Only VXLAN encapsulation type is supported in this mode.

Figure 40 Cisco AVE – Local vs. No Local Switching



If **Local switching** mode is used, either a range of VLANs or a single infra-VLAN must be specified when using VLAN and VXLAN encapsulations respectively. The specified VLAN(s) have local scope as they are only significant to the layer 2 segment between Cisco AVE and ACI Leaf switch.

As stated earlier, Cisco AVE leverages the VMware vDS to operate and the VDS is configured for private VLAN (PVLAN) mode. When a Cisco AVE based VMM domain is created on Cisco APIC, they must associate the domain with a range of VLANs to be used for PVLAN pair association of port groups on the DVS. Server administrators do not need to associate PVLANs to port groups on vCenter because Cisco APIC automatically associates PVLAN pairs with the corresponding ACI EPGs.

Also, Cisco AVE can be deployed using local or remote storage – local storage is recommended. Cisco AVE virtual machine configuration can be lost if the ESXi host or Cisco AVE VM is removed or moved from vCenter.

Validated Design – Virtual Networking on HyperFlex Nodes

The virtual networking on Cisco HyperFlex systems attached to a Cisco ACI Fabric can be converted to use either a VMware vDS or Cisco AVE instead of the default VMware virtual switches deployed by the HyperFlex installer. In this design, the virtual machines using VM networks are migrated from the VM network vSwitch to either a VMware vDS (Management cluster) or an APIC-controlled Cisco AVE switch (Application cluster). However, the default HyperFlex infrastructure networks, management, storage data and vMotion will remain on the default virtual switches created by the HyperFlex installer. The resulting ESXi networking design is therefore a combination of 3 virtual switches and a fourth virtual switch for application traffic that is migrated to either a VMware vDS or an APIC-Controlled Cisco AVE switch.

Table 2 and Table 3 lists the ESXi networking deployed by the HX Installer on the HX nodes in the Management and Application clusters respectively.

Table 2 Virtual Switching on HyperFlex Nodes in Management Cluster

VLAN ID	HyperFlex & UCS VLAN Names	HX Server Uplinks (vNICs)	Virtual Switch	ESXi Port Groups
118	hxv-inband-mgmt	hv-mgmt-a, hv-mgmt-b	vswitch-hx-inband-mgmt	Management Network (ESXi)
				Storage Controller Management Network (SCVM)
3218	hxv1-storage-data	storage-data-a, storage-data-b	vswitch-hx-storage-data	Storage Hypervisor Data Network (ESXi)
				Storage Controller Data Network (SCVM)
1118-1218	hxv-vm-network1118-hxv-vm-network1218	vm-network-a, vm-network-b	vswitch-hx-vm-network	hxv-vm-network-1118 - hxv-vm-network-1218
3018	hxv-vmotion	hv-vmotion-a, hv-vmotion-b	vMotion	vmotion-3018

Table 3 Virtual Switching on HyperFlex Nodes in Applications Cluster

VLAN ID	HyperFlex & UCS VLAN Names	HX Server Uplinks (vNICs)	Virtual Switch	ESXi Port Groups
118	hxv-inband-mgmt	hv-mgmt-a, hv-mgmt-b	vswitch-hx-inband-mgmt	Management Network (ESXi)
				Storage Controller Management Network (SCVM)
3218	hxv1-storage-data	storage-data-a, storage-data-b	vswitch-hx-storage-data	Storage Hypervisor Data Network (ESXi)
				Storage Controller Data Network (SCVM)
2118-2218	hxv-vm-network2118-hxv-vm-network2218	vm-network-a, vm-network-b	vswitch-hx-vm-network	hxv-vm-network-2118 - hxv-vm-network-2218
3018	hxv-vmotion	hv-vmotion-a, hv-vmotion-b	vMotion	vmotion-3018

Any port-groups for application virtual machines on the **vswitch-hx-vm-network** vSwitch and the uplinks will be migrated to Cisco AVE (Application cluster) or VMware vDS (Management cluster) using VMM integration. Cisco AVE is deployed in **local switching** mode with VxLAN tunneling. The infrastructure VLAN (v4093) is required for VxLAN tunneling and must be added to the ACI Leaf switches and to the Cisco UCS Fabric (uplinks and vNICs to the HX servers) before the APIC can create the Cisco AVE switch in the vSphere environment.

vSphere High Availability Recommendations

The VMware vSphere high availability setup is critical for the operation of a HyperFlex stretched cluster. HyperFlex installation configures many VMware features that a stretched cluster requires such as vSphere high availability, DRS, virtual machine and datastore host-groups, site-affinity, etc. In addition, customers should also enable the following vSphere high availability settings in VMware vCenter:

- vSphere Availability: vSphere high availability should be enabled but keep Proactive high availability disabled
- Failure Conditions and responses:
  - Enable Host Monitoring
  - For Host Failure Response, select Restart VMs
  - For Response for Host Isolation, select Power off and restart VMs

  - For Datastore with PDL, select Power off and restart VMs
  - For Datastore with PDL, select Power off and restart VMs (conservative)
  - For VM Monitoring: Customer can enable this if they prefer. It is typically disabled.
- Admission Control: select Cluster resource percentage for Define host failover capacity by
- Datastore Heartbeats: Select Use datastores only from the specified list and select HyperFlex datastores in each site
- Advanced Settings:
  - select False for `das.usedefaultisolationaddress`
  - select an IP address in Site A for `das.isolationaddress0`
  - select an IP address in Site B for `das.isolationaddress1`
- For additional details, see Operating Cisco HyperFlex Data Platform Stretched Clusters white paper in the References section of this document. This white paper is also referenced in the next section.

HyperFlex Stretched Cluster Design Guidelines and Recommendations

For a detailed discussion on the HyperFlex stretched cluster architecture and design, please review the following white paper: Operating Cisco HyperFlex Data Platform Stretched Clusters

Solution Validation

To verify the design, the solution was built in the Cisco Labs with all components integrated and validated to ensure interoperability and to confirm that the design is ready for deployment. This section provides a summary of the validation done for this CVD.

Validated Hardware and Software

Table 4 lists the hardware and software versions used during the solution validation. The versions used have been certified within interoperability matrixes supported by Cisco and VMware.

Table 4 Hardware and Software Versions

Infrastructure Domain	Component		Software	Notes
Network (ACI MultiPod Fabric)	**Pod 1**	**Pod 2**		
	Cisco APIC M2 Server x 2 (APIC-SERVER-M2)	Cisco APIC M2 Server x 1 (APIC-SERVER-M2)	4.0.1h	3-node APIC Cluster
	Cisco Nexus 9364C x 2 (N9K-C9364C)	Cisco Nexus 9364C x 2 (N9K-C9364C)	aci-n9000-dk9.14.0.1h	ACI Spine Switches
	Cisco Nexus 93180YC-EX x 2 (N9K-C93180YC-EX)	Cisco Nexus 93180YC-EX x 2 (N9K-C93180YC-EX)	aci-n9000-dk9.14.0.1h	ACI Leaf Switches for HyperFlex and UCS Domains
	Cisco Nexus 9372PX x 2 (N9K-C9372PX)	Cisco Nexus 9372PX x 2 (N9K-C9372PX)	aci-n9000-dk9.14.0.1h	ACI Border Leaf Switches for Shared L3Out
	Cisco Nexus 93180YC-EX x 2 (N9K-C93180YC-EX)	Cisco Nexus 93180YC-EX x 2 (N9K-C93180YC-EX)	NX-OS 9.2(1)	IPN router deployed in NX-OS Standalone Mode
Hyperconverged Infrastructure (Cisco HyperFlex Standard & Stretched Clusters)				
	Witness VM		1.0.4	Deployed in existing infrastructure, outside ACI; Available as an OVA
	Pod 1	**Pod 2**		
	Cisco HX220c M4S x 4 (HX220C-M4S)	–	3.5(2e)*	• 4-node Management Cluster (Standard Cluster); • Cisco HyperFlex Hybrid M4 Nodes with 10G VIC 1227 (UCSC-MLOM-CSC-02)
	Cisco UCS 6248 FI x 2 (UCS-FI-6248UP)	–	4.0(2d)	1RU 10G Fabric Interconnect with 48 ports
	Cisco HX220C-M5SX x 4 (HX220C-M5SX)	HX220C-M5SX x 4 (HX220C-M5SX)	3.5(2e)*	• 8-node Application Cluster (4-4 Stretch Cluster); • Cisco HyperFlex Hybrid M5 Nodes with 40G VIC 1387 (UCSC-MLOM-C40G-03)
	Cisco UCS 6332 FI x 2 (UCS-FI-6332-16UP)	Cisco UCS 6332 FI x 2 (UCS-FI-6332UP)	4.0(2d)	• Pod 1 FI: 1RU, 40G FI with 40 ports (24 fixed ports) • Pod 2 FI: 1RU, 40G FI with 32 fixed ports
Virtualization	**Pod 1**	**Pod 2**		
	VMware vSphere 6.5U2 EP13	VMware vSphere 6.5U2 EP13	6.5U2 EP13	Hypervisor – Custom Cisco Build: 13004031
	VMware vCenter Server Appliance 6.5 U2e	–	6.5U2e	• VCSA for Application Cluster; Management Cluster is managed by a VCSA outside the ACI Fabric • Version: 6.5.0.23100 Build Number 11347054
	VMware vDS, Cisco AVE	Cisco AVE	2.0(1a)	Virtual Switches • VMware vDS used in Management Cluster • Cisco AVE used in Application Cluster
Security				
	Cisco Umbrella			Cloud-based security for Enterprise; Virtual Appliances(Optional) deployed on-premise: https://umbrella.cisco.com
Management & Monitoring				
	Cisco UCS Manager		4.0(2d)	Management Cluster is managed by a VMware vCenter Server outside ACI Fabric
	Cisco HyperFlex Connector			Virtual Switches – VMware vDS in Management Cluster and Cisco AVE in Application Cluster
	Cisco Intersight			Cloud-based Management Tool
	Cisco HyperFlex vCenter Plugin		3.0.1.29754	6.5 Flash client – added by HX Installer
	Cisco ACI vCenter Plugin		3.2.2000.12	
Tools				
	HX Bench, VdBench			Load Generation Tools

The solution was primarily validated on HyperFlex release 3.5(2d), and 3.5(2e) for a subset of test cases.

Interoperability

To use other hardware models or software versions in this design, verify interoperability using the following matrixes. Also, review the release notes for release and product documentation.

- Cisco UCS and HyperFlex Hardware and Software Interoperability Tool
- Cisco ACI Recommended Release
- Cisco ACI Virtualization Compatibility Matrix
- Cisco APIC and ACI Virtual Edge Support Matrix
- VMware Compatibility Guide

Solution Validation

The solution was validated for basic data forwarding by deploying virtual machine running VdBench and IOMeter tools. The system was validated for resiliency by failing various aspects of the system under load. Examples of the types of tests executed include:

- Failure and recovery of various links and components between the sites and within each site.
- Failure events triggering vSphere high availability between sites.
- Failure events triggering vMotion between sites.
- All tests were performed under load, using load generation tools. Different IO profiles representative of customer deployments were used.

Summary

The **Cisco HyperFlex Stretched Cluster with Cisco ACI Multi-Pod Fabric** solution for VMware vSphere deployments delivers an active-active data center architecture that can protect against disasters and provide business continuity in the event of a site or data center-wide failure. The HyperFlex stretched cluster used in the design enables the virtual server infrastructure to be extended across different geographical sites to provide availability with quick recovery and zero data-loss in the event of a site failure. The ACI Multi-Pod fabric in the design provides the necessary Layer 2 and Layer 3 connectivity between the sites to enable the active-active data center. The ACI Multi-Pod fabric is centrally and uniformly managed using a single APIC cluster which greatly simplifies the administration of a multi-data center solution as the tenant configuration and policies are only done once and not on a per site basis. The APIC cluster is also distributed across both sites to provide availability in the event of a site failure. The solution was validated to provide customers and partners with a reliable reference design to deploy their own active-active data center solutions.

References

Cisco HyperFlex

- Cisco HyperFlex Virtual Server Infrastructure 3.0 with Cisco ACI 3.2 and VMware vSphere 6.5: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hx_30_vsi_aci_32.html
- Cisco HyperFlex 3.0 for Virtual Server Infrastructure with VMware ESXi: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hyperflex_30_vsi_esxi.html
- Operating Cisco HyperFlex HX Data Platform Stretch Clusters:https://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/hyperflex-hx-series/operating-hyperflex.pdf
- Cisco HyperFlex Systems Stretched Cluster Guide, Release 3.5: https://www.cisco.com/c/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/HyperFlex_Stretched_Cluster/3_5/b_HyperFlex_Systems_Stretched_Cluster_Guide_3_5.html
- Comprehensive Documentation for Cisco HyperFlex: https://http://hyperflex.io
- Comprehensive Documentation Roadmap for Cisco HyperFlex: https://www.cisco.com/c/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/HX_Documentation_Roadmap/HX_Series_Doc_Roadmap.html

Cisco UCS

- Cisco Unified Computing System: http://www.cisco.com/en/US/products/ps10265/index.html
- Cisco UCS 6300 Series Fabric Interconnects: http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-6300-series-fabric-interconnects/index.html
- Cisco UCS 5100 Series Blade Server Chassis: http://www.cisco.com/en/US/products/ps10279/index.html
- Cisco UCS 2300 Series Fabric Extenders: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-6300-series-fabric-interconnects/datasheet-c78-675243.html
- Cisco UCS 2200 Series Fabric Extenders: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-6300-series-fabric-interconnects/data_sheet_c78-675243.html
- Cisco UCS B-Series Blade Servers: http://www.cisco.com/en/US/partner/products/ps10280/index.html
- Cisco UCS C-Series Rack Mount Servers: http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c-series-rack-servers/index.html
- Cisco UCS VIC Adapters: http://www.cisco.com/en/US/products/ps10277/prod_module_series_home.html
- Cisco UCS Manager: http://www.cisco.com/en/US/products/ps10281/index.html
- Cisco UCS Manager Plug-in for VMware vSphere Web Client: http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/sw/vmware_tools/vCenter/vCenter_Plugin_Release_Notes/2_0/b_vCenter_RN_for_2x.html

Cisco ACI Fabric

- ACI Multi-Pod White Paper: https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737855.html

- Cisco ACI Multi-Pod Configuration Whitepaper: https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739714.html
- Verified Scalability Guide for Cisco APIC Release 4.0(1) https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/4-x/verified-scalability/Cisco-ACI-Verified-Scalability-Guide-401.html
- Cisco Nexus 9000 Series Switches: http://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/index.html
- Transceiver Compatibility Matrix for Cisco Switches: https://tmgmatrix.cisco.com/
- Cisco Application Centric Infrastructure – Data center and Virtualization: https://www.cisco.com/c/en_au/solutions/data-center-virtualization/aci.html
- Cisco Application Centric Infrastructure – Cisco Data center: https://www.cisco.com/go/aci
- Cisco ACI Fundamentals: https://www.cisco.com/c/en/us/td/docs/switches/data center/aci/apic/sw/1-x/aci-fundamentals/b_ACI-Fundamentals.html
- Cisco ACI Infrastructure Release 2.3 Design Guide: https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737909.pdf
- Cisco ACI Infrastructure Best Practices Guide: https://www.cisco.com/c/en/us/td/docs/switches/data center/aci/apic/sw/1-x/ACI_Best_Practices/b_ACI_Best_Practices.html

Virtualization Layer

- VMware vCenter Server: http://www.vmware.com/products/vcenter-server/overview.html
- VMware vSphere: https://www.vmware.com/products/vsphere

Security

- Integrating Cisco Umbrella to Cisco HyperFlex and Cisco UCS Solutions: https://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/hyperflex-hx-series/whitepaper-c11-741088.pdf

Interoperability Matrixes

- Cisco UCS and HyperFlex Hardware Compatibility Matrix: https://ucshcltool.cloudapps.cisco.com/public/
- Cisco ACI Recommended APIC and Cisco Nexus 9000 Series ACI-Mode Switches Releases: https://www.cisco.com/c/en/us/td/docs/switches/data center/aci/apic/sw/recommended-release/b_Recommended_Cisco_ACI_Releases.html
- Cisco ACI Virtualization Compatibility Matrix: https://www.cisco.com/c/dam/en/us/td/docs/Website/data center/aci/virtualization/matrix/virtmatrix.html
- Cisco APIC and ACI Virtual Edge Support Matrix: https://www.cisco.com/c/dam/en/us/td/docs/Website/data center/aveavsmatrix/index.html
- VMware Compatibility Guide: http://www.vmware.com/resources/compatibility

About the Authors

Archana Sharma, Technical Leader, Cisco UCS Data Center Solutions, Cisco Systems Inc.

Archana Sharma is Technical Marketing Engineer with over 20 years of experience at Cisco on a range of technologies that span Data Center, Desktop Virtualization, Collaboration, and other Layer2 and Layer3 technologies. Archana is focused on systems and solutions for Enterprise and Provider deployments, including delivery of Cisco Validated designs for over 10 years. Archana is currently working on designing and integrating Cisco UCS-based Converged Infrastructure solutions. Archana holds a CCIE (#3080) in Routing and Switching and a Bachelor's degree in Electrical Engineering from North Carolina State University.

Acknowledgements

- Haseeb Niazi, Technical Marketing Engineer, Cisco Systems, Inc.
- Ramesh Isaac, Technical Marketing Engineer, Cisco Systems, Inc.