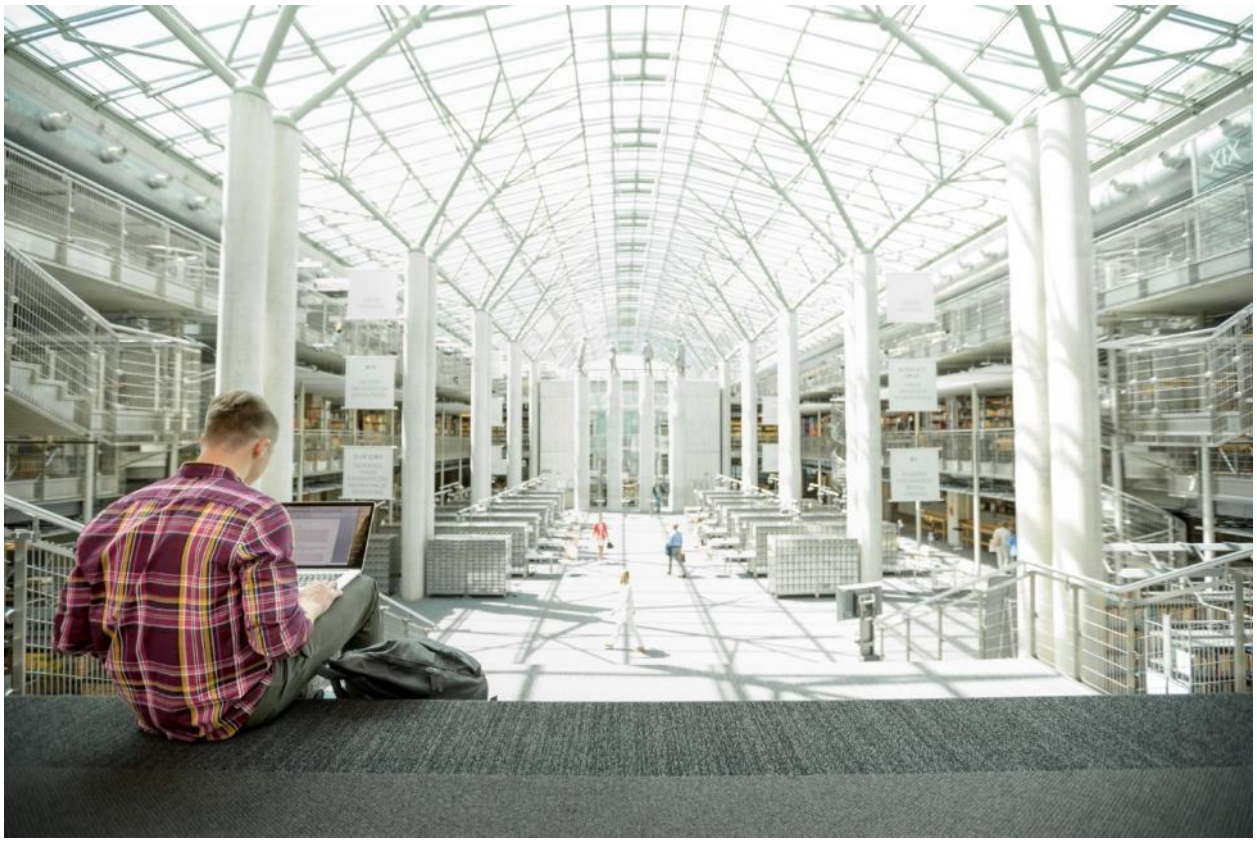


VersaStack with Cisco ACI and IBM FS9100 NVMe-accelerated Storage Design Guide

Published: January 2020

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

Cisco Validated Designs (CVDs) deliver systems and solutions that are designed, tested, and documented to facilitate and improve customer deployments. These designs incorporate a wide range of technologies and products into a portfolio of solutions that have been developed to address the business needs of the customers and to guide them from design to deployment.

The VersaStack solution described in this CVD, delivers a Converged Infrastructure platform (CI) specifically designed for software defined networking (SDN) enabled data centers, which is a validated solution jointly developed by Cisco and IBM.

In this solution deployment, Cisco Application Centric Infrastructure (Cisco ACI) delivers an intent-based networking framework to enable agility in the data center. Cisco ACI radically simplifies, optimizes, and accelerates infrastructure deployment and governance and expedites the application deployment lifecycle. IBM® FlashSystem 9100 combines the performance of flash and Non-Volatile Memory Express (NVMe) with the reliability and innovation of IBM FlashCore technology and the robust features of IBM Spectrum Virtualize.

Solution Overview

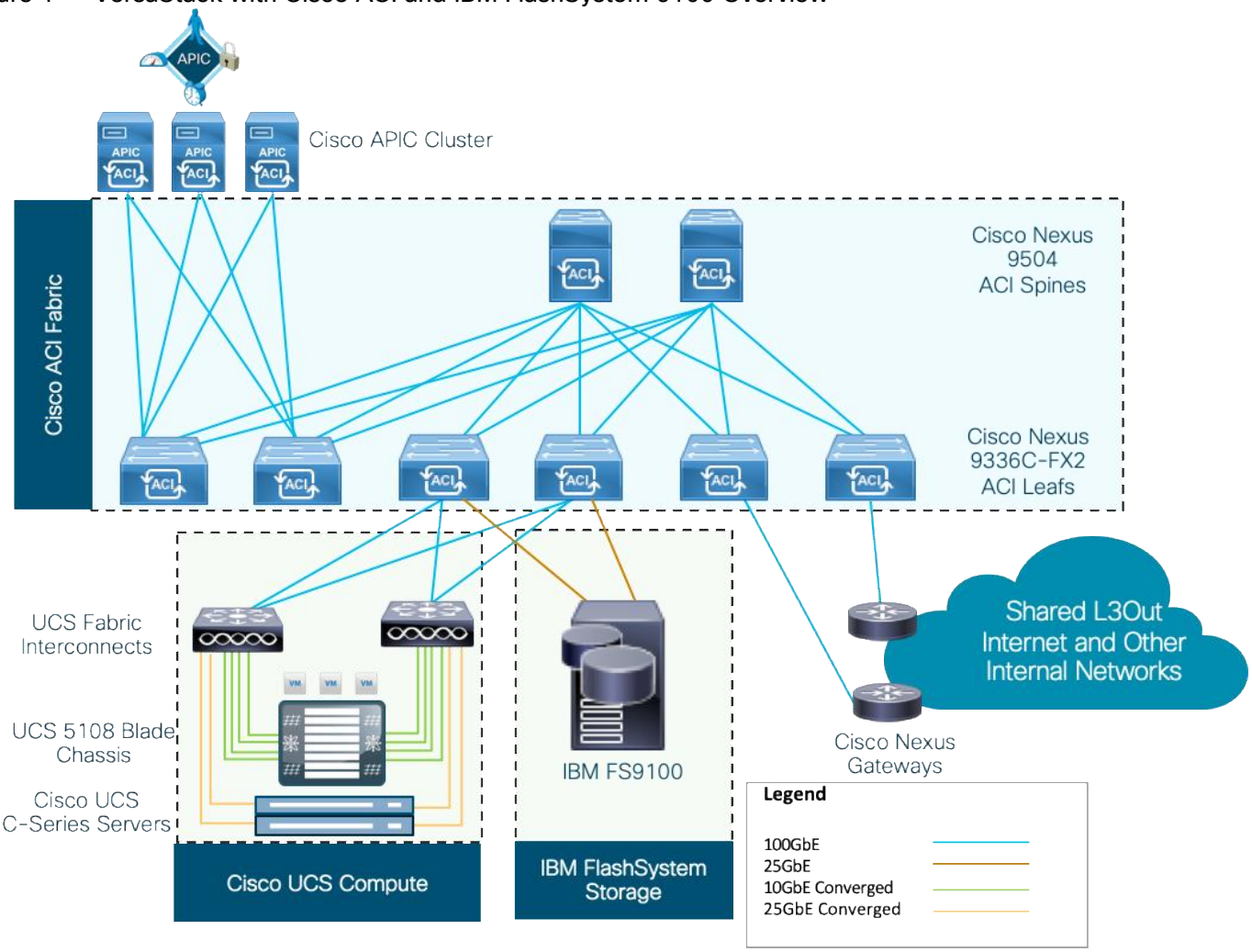
Introduction

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

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

This document describes the design of the high-performance VersaStack with Cisco ACI and IBM FlashSystem 9100 NVMe based solution. The solution is a pre-designed, best-practice data center architecture with VMware vSphere built on Cisco Unified Computing System (Cisco UCS). The solution architecture presents a robust infrastructure viable for a wide range of application workloads implemented as a Virtual Server Infrastructure (VSI). Figure 1 illustrates a high-level overview of the solution.

Figure 1 VersaStack with Cisco ACI and IBM FlashSystem 9100 Overview



Audience

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

What's New in this Release?

The VersaStack with VMware vSphere 6.7 U3 CVD introduces new hardware and software into the portfolio. The following design elements distinguish this version of VersaStack from previous models:

- Support for the Cisco UCS release 4.0(4e)
- Support for Cisco ACI 4.2
- Validation of 25GbE IP-based iSCSI storage design with Cisco Nexus ACI Fabric
- Validation of VMware vSphere 6.7 U3

For more information on the complete portfolio of VersaStack solutions, refer to the VersaStack documentation:

http://www.cisco.com/c/en/us/solutions/enterprise/data-center-designs-cloud-computing/versastack-designs.html

VersaStack Program Benefits

Cisco and IBM have carefully validated and verified the VersaStack solution architecture and its many use cases while creating a portfolio of detailed documentation, information, and references to assist customers in transforming their data centers to this shared infrastructure model.

Business Value

VersaStack combines the best-in-breed highly scalable storage controllers from IBM with the Cisco UCS B-Series and C-Series compute servers, and Cisco Nexus and MDS networking components. Quick deployment and rapid time to value allow enterprise clients to move away from disparate layers of compute, network, and storage to integrated stacks.

This CVD for the VersaStack reference architecture with pre-validated configurations reduces risk and expedites the deployment of infrastructure and applications. The system architects and administrators receive configuration guidelines to save implementation time while reducing operational risk.

The complexity of managing systems and deploying resources is reduced dramatically, and problem resolution is provided through a single point of support. VersaStack streamlines the support process so that customers can realize the time benefits and cost benefits that are associated with simplified single-call support.

Cisco Validated Designs incorporate a broad set of technologies, features, and applications to address any business needs.

This portfolio includes, but is not limited to best practice architectural design, Implementation and deployment instructions, Cisco Validated Designs and IBM Redbooks focused on a variety of use cases.

Design Benefits

VersaStack with IBM FlashSystem 9100 overcomes the historical complexity of IT infrastructure and its management.

Incorporating Cisco ACI and Cisco UCS Servers with IBM FlashSystem 9100 storage, this high-performance solution provides easy deployment and support for existing or new applications and business models. VersaStack accelerates IT and delivers business outcomes in a cost-effective and extremely timely manner.

One of the key benefits of VersaStack is the ability to maintain consistency in both scale-up and scale-down models. VersaStack can scale-up for greater performance and capacity. You can add compute, network, or storage resources as needed; or it can also scale-out when you need multiple consistent deployments such as rolling out additional VersaStack modules. Each of the component families shown in Figure 2 offer platform and resource options to scale the infrastructure up or down while supporting the same features and functionality.

The following factors contribute to significant total cost of ownership (TCO) advantages:

- Simpler deployment model: Fewer components to manage
- Higher performance: More work from each server due to faster I/O response times
- Better efficiency: Power, cooling, space, and performance within those constraints
- Availability: Help ensure applications and services availability at all times with no single point of failure

- Flexibility: Ability to support new services without requiring underlying infrastructure modifications
- Manageability: Ease of deployment and ongoing management to minimize operating costs
- Scalability: Ability to expand and grow with significant investment protection
- Compatibility: Minimize risk by ensuring compatibility of integrated components
- Extensibility: Extensible platform with support for various management applications and configuration tools

Technology Overview

The VersaStack architecture is comprised of the following infrastructure components for compute, network, and storage:

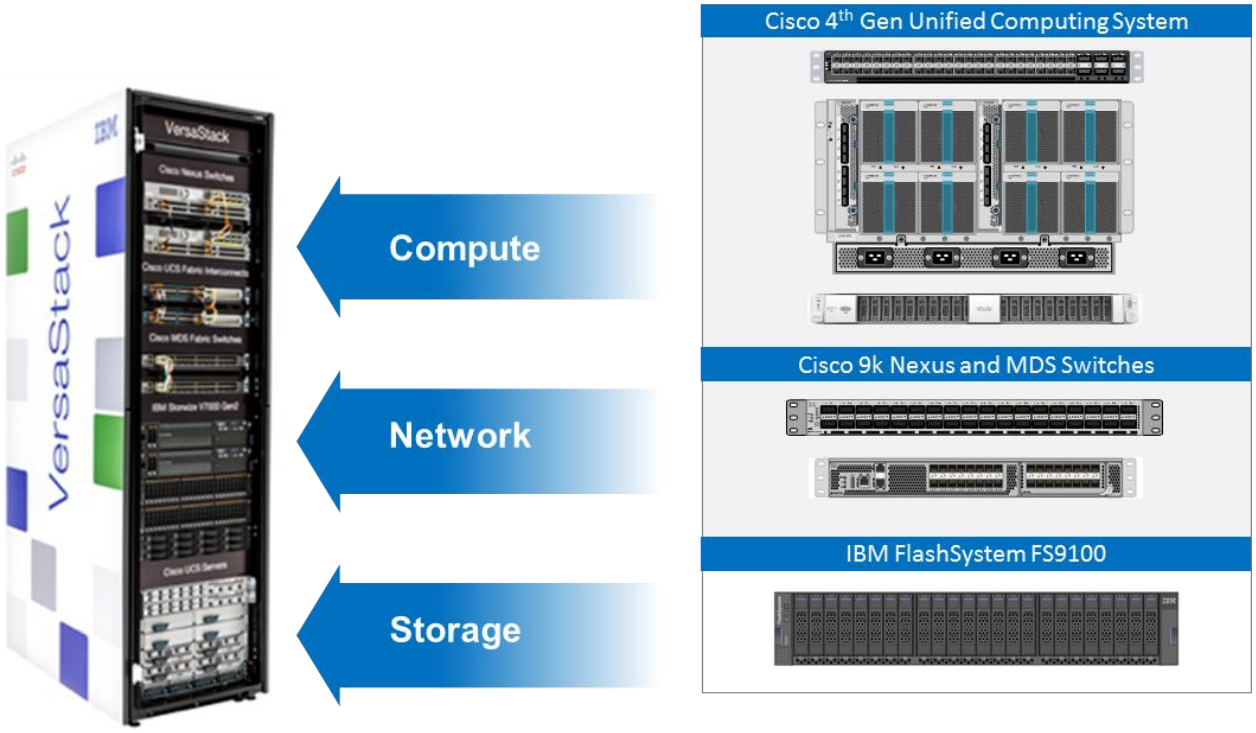
- Cisco Unified Computing System
- Cisco Nexus and Cisco MDS Switches
- IBM SAN Volume Controller, FlashSystem, and IBM Storwize family storage

These components are connected and configured according to best practices of both Cisco and IBM and provide an ideal platform for running a variety of workloads with confidence.

The VersaStack reference architecture explained in this document leverages:

- Cisco UCS 6400 Series Fabric Interconnects (FI)
- Cisco UCS 5108 Blade Server chassis
- Cisco Unified Compute System (Cisco UCS) servers with 2nd generation Intel Xeon scalable processors
- Cisco Nexus 9336C-FX2 Switches running ACI mode
- IBM FlashSystem 9100 NVMe-accelerated Storage
- VMware vSphere 6.7 Update 3

Figure 2 VersaStack with Cisco ACI and IBM FlashSystem 9100 – Components



The following sections provide a technical overview of the compute, network, storage and management components of the VersaStack solution.

Cisco Unified Computing System

Cisco Unified Computing System (Cisco UCS) is a next-generation data center platform that integrates computing, networking, storage access, and virtualization resources into a cohesive system designed to reduce total cost of ownership (TCO) and to increase business agility. The system integrates a low-latency; lossless unified network fabric with enterprise-class, x86-architecture servers. The system is an integrated, scalable, multi-chassis platform where all resources are managed through a unified management domain.

The Cisco Unified Computing System consists of the following subsystems:

- Compute - The compute piece of the system incorporates servers based on latest Intel's x86 processors. Servers are available in blade and rack form factor, managed by Cisco UCS Manager.
- Network - The integrated network fabric in the system provides a low-latency, lossless, 10/25/40/100 Gbps Ether-net fabric. Networks for LAN, SAN and management access are consolidated within the fabric. The unified fabric uses the innovative Single Connect technology to lowers costs by reducing the number of network adapters, switches, and cables. This in turn lowers the power and cooling needs of the system.
- Storage access - Cisco UCS system provides consolidated access to both SAN storage and Network Attached Storage over the unified fabric. This provides customers with storage choices and investment protection. The use of Policies, Pools, and Profiles allows for simplified storage connectivity management.
- Management - The system uniquely integrates compute, network and storage access subsystems, enabling it to be managed as a single entity through Cisco UCS Manager software. Cisco UCS Manager increases IT staff productivity by enabling storage, network, and server administrators to collaborate on Service Profiles that define the desired server configurations.

Cisco UCS Management

Cisco UCS® Manager (UCSM) provides unified, integrated management for all software and hardware components in Cisco UCS. UCSM manages, controls, and administers multiple blades and chassis enabling administrators to manage the entire Cisco Unified Computing System as a single logical entity through an intuitive GUI, a CLI, as well as a robust API. Cisco UCS Manager is embedded into the Cisco UCS Fabric Interconnects and offers comprehensive set of XML API for third party application integration.

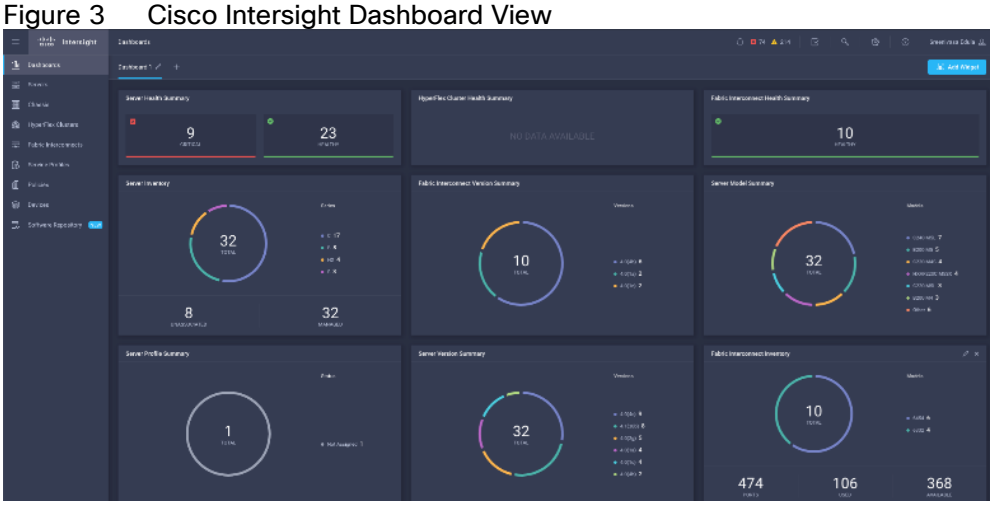
Cisco Intersight (optional)

The Cisco Intersight™ platform provides intelligent cloud-powered infrastructure management for Cisco Unified Computing System™ (Cisco UCS®) and Cisco HyperFlex™ platforms. Cisco Intersight is a subscription-based, cloud service for infrastructure management that simplifies operations by providing pro-active, actionable intelligence for operations. Cisco Intersight provides capabilities such as Cisco Technical Assistance Center (TAC) integration for support and Cisco Hardware Compatibility List (HCL) integration for compliance that Enterprises can leverage for all their Cisco HyperFlex and Cisco UCS systems in all locations. Cloud-based delivery enables Enterprises to quickly adopt the new features that are continuously being rolled out in Cisco Intersight.

Each Cisco UCS server or Cisco HyperFlex system automatically includes a Cisco Intersight Base edition at no additional cost when the customer accesses the Cisco Intersight portal and claims the device. In addition, customers can purchase the Cisco Intersight Essentials edition using the Cisco ordering tool.

A view of the unified dashboard provided by Intersight can be seen in Figure 3.

Figure 3 Cisco Intersight Dashboard View



For more information on Cisco Intersight, see:

https://www.intersight.com/help/getting_started#cisco_intersight_overview

Cisco UCS Director (optional)

Cisco UCS Director is a heterogeneous platform for private cloud Infrastructure as a Service (IaaS). It supports a variety of hypervisors along with Cisco and third-party servers, network, storage, converged and hyperconverged infrastructure across bare-metal and virtualized environments. Cisco UCS Director provides increased efficiency through automation capabilities throughout VersaStack components. The Cisco UCS Director adapter for IBM Storage and VersaStack converged infrastructure solution allows easy deployment and management of these technologies using Cisco UCS Director.

Cisco continues to invest and enhance data center automation and private cloud infrastructure as a service (IaaS) platform, Cisco UCS Director. At the same time, we are leveraging the Cisco Intersight platform to deliver additional value and operational benefits when coupled with Cisco UCS Director.

Cisco is implementing a strategy for Cisco UCS Director and Cisco Intersight to help customers transition. Cisco UCS Director can be managed by Cisco Intersight to make updates easier and improve support. The combination of Cisco UCS Director and Intersight will simplify day to day operations and extend private cloud IaaS services.

For more information, see:

https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-director/index.html#~stickynav=1

Cisco UCS Fabric Interconnects

The Cisco UCS Fabric Interconnects (FIs) provide a single point for connectivity and management for the entire Cisco Unified Computing System. Typically deployed as an active-active pair, the system's fabric interconnects integrate all components into a single, highly available management domain controlled by the Cisco UCS Manager. Cisco UCS FIs provide a single unified fabric for the system that supports LAN, SAN and management traffic using a single set of cables.

The 4th generation (6454) Fabric Interconnect (Figure 4) leveraged in this VersaStack design provides both network connectivity and management capabilities for the Cisco UCS system. The Cisco UCS 6454 offers line-rate, low-latency, lossless 10/25/40/100 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and 32 Gigabit Fibre Channel functions.

Figure 4 Cisco UCS 6454 Fabric Interconnect

Cisco UCS 5108 Blade Server Chassis

The Cisco UCS 5108 Blade Server Chassis (Figure 5) delivers a scalable and flexible blade server architecture. The Cisco UCS blade server chassis uses an innovative unified fabric with fabric-extender technology to lower total cost of ownership by reducing the number of network interface cards (NICs), host bus adapters (HBAs), switches, and cables that need to be managed. Cisco UCS 5108 is a 6-RU chassis that can house up to 8 half-width or 4 full-width Cisco UCS B-Series Blade Servers. A passive mid-plane provides up to 80Gbps of I/O bandwidth per server slot and up to 160Gbps for two slots (full-width). The rear of the chassis contains two I/O bays to house Cisco UCS Fabric Extenders for enabling uplink connectivity to the pair of FIs for both redundancy and bandwidth aggregation.

Figure 5 Cisco UCS 5108 Blade Server Chassis

Cisco UCS 2208XP Fabric Extender

The Cisco UCS Fabric extender (FEX) or I/O Module (IOM) multiplexes and forwards all traffic from servers in a blade server chassis to the pair of Cisco UCS FIs over 10Gbps unified fabric links. The Cisco UCS 2208XP Fabric Extender (Figure 6) has eight 10 Gigabit Ethernet, FCoE-capable, Enhanced Small Form-Factor Pluggable (SFP+) ports that connect the blade chassis to the FI. Each Cisco UCS 2208XP has thirty-two 10 Gigabit Ethernet ports connected through the midplane to each half-width slot in the chassis. Typically configured in pairs for redundancy, two fabric extenders provide up to 160 Gbps of I/O to the chassis.

Figure 6 Cisco UCS 2208XP Fabric Extender

Cisco UCS B-Series Blade Servers

Cisco UCS B-Series Blade Servers are based on Intel Xeon processors; they work with virtualized and non-virtualized applications to increase performance, energy efficiency, flexibility, and administrator productivity. The latest Cisco UCS M5 B-Series blade server models come in two form factors; the half-width Cisco UCS B200 Blade Server and the full-width Cisco UCS B480 Blade Server. Cisco UCS M5 server uses the latest Intel Xeon Scalable processors with up to 28 cores per processor. The Cisco UCS B200 M5 blade server supports 2 sockets, 3TB of RAM (using 24 x128GB DIMMs), 2 drives (SSD, HDD or NVMe), 2 GPUs and 80Gbps of total I/O

to each server. The Cisco UCS B480 blade is a 4-socket system offering 6TB of memory, 4 drives, 4 GPUs and 160 Gb aggregate I/O bandwidth.

The Cisco UCS B200 M5 Blade Server (Figure 7) has been used in this VersaStack architecture.

Figure 7 Cisco UCS B200 M5 Blade Server

Each supports the Cisco VIC 1400 series adapters to provide connectivity to the unified fabric.

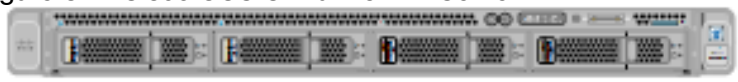
For more information about Cisco UCS B-series servers, see: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/datasheet-c78-739296.html

Cisco UCS C-Series Rack Servers

Cisco UCS C-Series Rack Servers deliver unified computing in an industry-standard form factor to reduce TCO and increase agility. Each server addresses varying workload challenges through a balance of processing, memory, I/O, and internal storage resources. The most recent M5 based C-Series rack mount models come in three main models; the Cisco UCS C220 1RU, the Cisco UCS C240 2RU, and the Cisco UCS C480 4RU chassis, with options within these models to allow for differing local drive types and GPUs.

The enterprise-class Cisco UCS C220 M5 Rack Server (Figure 8) has been leveraged in this VersaStack design.

Figure 8 Cisco UCS C220 M5 LFF Server



For more information about Cisco UCS C-series servers, see:

https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c-series-rack-servers/datasheet-listing.html

Cisco UCS Virtual Interface Card 1400

The Cisco UCS Virtual Interface Card (VIC) 1400 Series provides complete programmability of the Cisco UCS I/O infrastructure by presenting virtual NICs (vNICs) as well as virtual HBAs (vHBAs) from the same adapter according to the provisioning specifications within UCSM.

The Cisco UCS VIC 1440 is a dual-port 40-Gbps or dual 4x 10-Gbps Ethernet/FCoE capable modular LAN On Motherboard (mLOM) adapter designed exclusively for the M5 generation of Cisco UCS B-Series Blade Servers. When used in combination with an optional port expander, the Cisco UCS VIC 1440 capabilities are enabled for two ports of 40-Gbps Ethernet. In this CVD, Cisco UCS B200 M5 blade servers were equipped with Cisco VIC 1440.

The Cisco UCS VIC 1457 is a quad-port Small Form-Factor Pluggable (SFP28) mLOM card designed for the M5 generation of Cisco UCS C-Series Rack Servers. The card supports 10/25-Gbps Ethernet or FCoE. The card can present PCIe standards-compliant interfaces to the host, and these can be dynamically configured as either NICs or HBAs. In this CVD, Cisco VIC 1457 was installed in Cisco UCS C240 M5 server.

2nd Generation Intel® Xeon® Scalable Processors

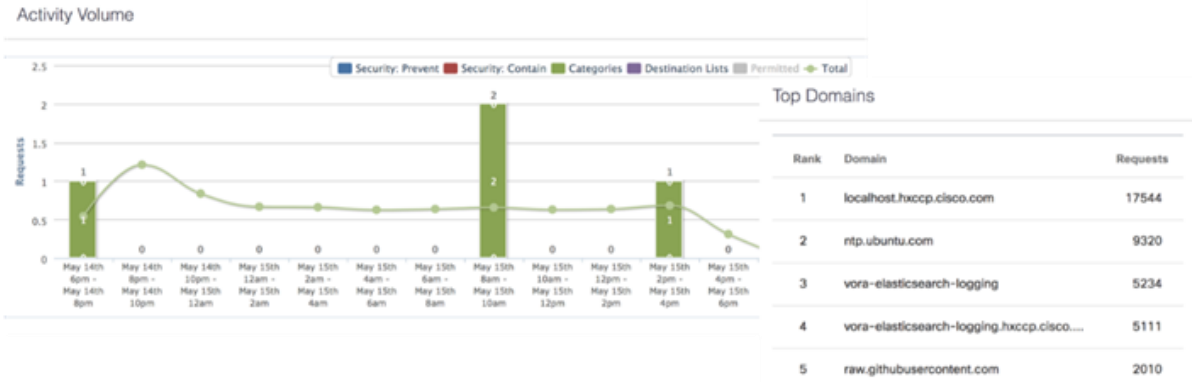
This VersaStack architecture includes the 2nd generation Intel Xeon Scalable processors in all the Cisco UCS M5 server models used in this design. These processors provide a foundation for powerful data center platforms with an evolutionary leap in agility and scalability. Disruptive by design, this innovative processor family supports new levels of platform convergence and capabilities across computing, storage, memory, network, and security resources.

Cascade Lake (CLX-SP) is the code name for the next-generation Intel Xeon Scalable processor family that is supported on the Purley platform serving as the successor to Skylake SP. These chips support up to eight-way multiprocessing, use up to 28 cores, incorporate a new AVX512 x86 extension for neural-network and deep-learning workloads, and introduce persistent memory support. Cascade Lake SP–based chips are manufactured in an enhanced 14-nanometer (14-nm++) process and use the Lewisburg chip set.

Cisco Umbrella (optional)

Cisco Umbrella is the delivery of secure DNS through Cisco's acquisition of OpenDNS. Cisco Umbrella stops malware before it can get a foothold by using predictive intelligence to identify threats that next-generation firewalls might miss. Implementation is easy as pointing to Umbrella DNS servers, and unobtrusive to the user base outside of identified threat locations they may have been steered to. In addition to threat prevention, Umbrella provides detailed traffic utilization as shown in Figure 9.

Figure 9 Traffic Breakdown of Activity Seen through Cisco Umbrella



For more information about Cisco Umbrella, see:
https://www.cisco.com/c/dam/en/us/products/collateral/security/router-security/opendns-product-overview.pdf

Cisco Workload Optimization Manager (optional)

Instantly scale resources up or down in response to changing demand assuring workload performance. Drive up utilization and workload density. Reduce costs with accurate sizing and forecasting of future capacity.

To perform intelligent workload management, Cisco Workload Optimization Manager (CWOM) models your environment as a market of buyers and sellers linked together in a supply chain. This supply chain represents the flow of resources from the datacenter, through the physical tiers of your environment, into the virtual tier and out to the cloud. By managing relationships between these buyers and sellers, CWOM provides closed-loop management of resources, from the datacenter, through to the application.

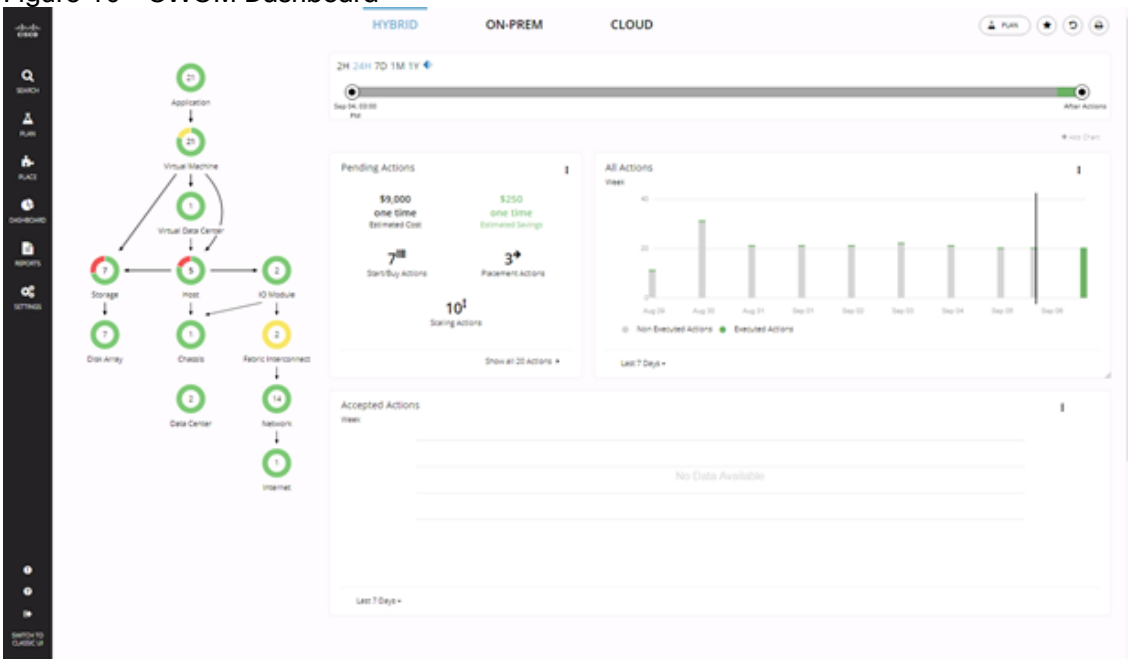
When you launch CWOM, the Home Page provides the following options:

- Planning

- Placement
- Reports
- Overall Dashboard

The CWOM dashboard provides views specific to On-Prem, the Cloud, or a Hybrid view of infrastructure, applications, and costs across both.

Figure 10 CWOM Dashboard



For more information about the full capabilities of workload optimization, planning, and reporting, see: https://www.cisco.com/c/en/us/products/servers-unified-computing/workload-optimization-manager/index.html

Cisco Application Centric Infrastructure and Nexus Switching

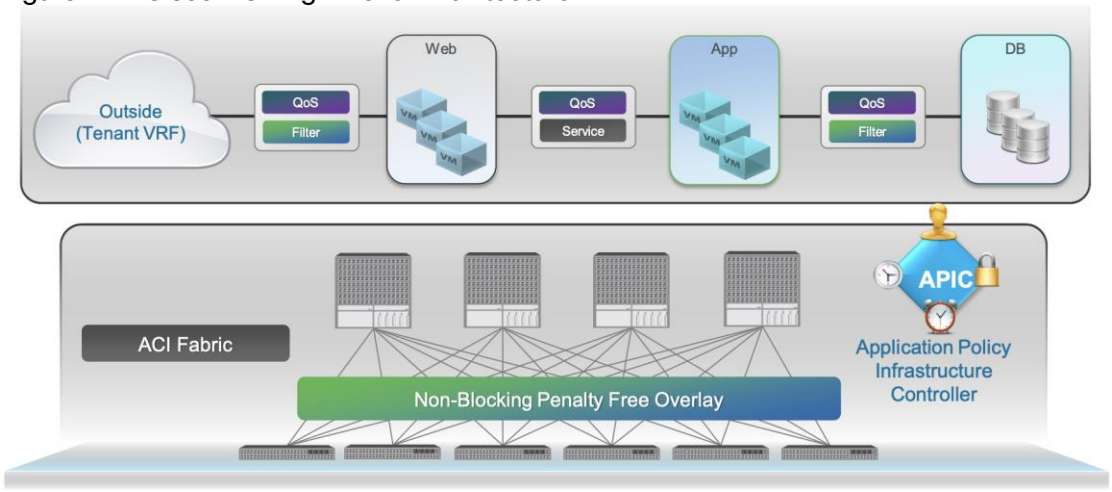
Cisco ACI is an evolutionary leap from SDN's initial vision of operational efficiency through network agility and programmability. Cisco ACI has industry leading innovations in management automation, programmatic policies, and dynamic workload provisioning. The ACI fabric accomplishes this with a combination of hardware, policy-based control systems, and closely coupled software to provide advantages not possible in other architectures.

Cisco ACI takes a policy-based, systems approach to operationalizing the data center network. The policy is centered around the needs (reachability, access to services, security policies) of the applications. Cisco ACI delivers a resilient fabric to satisfy today's dynamic applications.

Cisco ACI Architecture

The Cisco ACI fabric is a leaf-and-spine architecture where every leaf connects to every spine using high-speed 40/100-Gbps Ethernet links, with no direct connections between spine nodes or leaf nodes. The ACI fabric is a routed fabric with a VXLAN overlay network, where every leaf is VXLAN Tunnel Endpoint (VTEP). Cisco ACI provides both Layer 2 (L2) and Layer 3 (L3) forwarding across this routed fabric infrastructure.

Figure 11 Cisco ACI High-Level Architecture



Cisco Nexus 9000 Series Switches

The Cisco ACI fabric is built on a network of Cisco Nexus 9000 series switches that provide low-latency, high-bandwidth connectivity with industry proven protocols and innovative technologies to create a flexible, scalable, and highly available architecture. ACI is supported on several models of Nexus 9000 series switches and line cards. The selection of a Nexus 9000 series switch as an ACI spine or leaf switch will depend on a number of factors such as physical layer connectivity (1/10/25/40/50/100-Gbps), FEX aggregation support, analytics support in hardware (Cloud ASICs), FCoE support, link-level encryption, support for the Multi-Pod, Multi-Site design implementations and so on.

Architectural Building Blocks

The key architectural buildings blocks of the Cisco ACI fabric are:

- Application Policy Infrastructure Controller (APIC) - Cisco APIC is the unifying point of control in Cisco ACI for automating and managing the end-to-end data center fabric. The Cisco ACI fabric is built on a network of individual components that are provisioned and managed as a single entity. The APIC is a physical appliance that serves as a software controller for the overall fabric. It is based on Cisco UCS C-series rack mount servers with 2x10Gbps links for dual-homed connectivity to a pair of leaf switches and 1Gbps interfaces for out-of-band management.

- Spine Nodes – The spines provide high-speed (40/100-Gbps) connectivity between leaf nodes. The ACI fabric forwards traffic by doing a host lookup in a mapping database that contains information about the leaf node where an endpoint (IP, Mac) resides. All known endpoints are maintained in a hardware database on the spine switches. The number of endpoints or the size of the database is a key factor in the choice of a Nexus 9000 model as a spine switch. Leaf switches also maintain a database but only for those hosts that send/receive traffic through it.

The Cisco Nexus featured in this design for the ACI spine is the Nexus 9364C implemented in ACI mode (Figure 12).

Figure 12 Nexus 9364C

For more information about Cisco Nexus 9364C switch, see:
https://www.cisco.com/c/en/us/products/switches/nexus-9364c-switch/index.html

- Leaf Nodes – Leaf switches are essentially Top-of-Rack (ToR) switches that end devices connect into. They provide Ethernet connectivity to devices such as servers, firewalls, storage and other network elements. Leaf switches provide access layer functions such as traffic classification, policy enforcement, L2/L3 forwarding of edge traffic etc. The criteria for selecting a specific Nexus 9000 model as a leaf switch will be different from that of a spine switch.

The Cisco Nexus featured in this design for the ACI leaf is the Nexus 9336C-FX2 implemented in ACI mode (Figure 13).

Figure 13 Nexus 9336C-FX2

For more information about Cisco Nexus 9336C-FX2 switch, see:

https://www.cisco.com/c/en/us/products/switches/nexus-9336c-fx2-switch/index.html

For more information about Cisco ACI, see: https://www.cisco.com/c/en/us/products/cloud-systemsmanagement/application-policy-infrastructure-controller-apic/index.html

IBM Spectrum Virtualize

The IBM Spectrum Virtualize™ software stack was first introduced as a part of the IBM SAN Volume Controller (SVC) product released in 2003, offering unparalleled storage virtualization capabilities before being integrated into the IBM Storwize platform and more recently, a subset of the IBM FlashSystem storage appliances.

Since the first release of IBM SAN Volume Controller, IBM Spectrum Virtualize has evolved into the feature-rich storage hypervisor evolving over 34 major software releases, installed and deployed on over 240,000+ Storwize and 70,000 SVC engines. Managing 410,000 enclosures, virtualizing, managing and securing 9.6 Exabytes of data. Exceeding 99.999% availability.

IBM Spectrum Virtualize firmware version 8.2.1.0 provides the following features:

- Connectivity: Incorporating support for increased bandwidth requirements of modern operating systems:
  - Both 10GbE and 25GbE ports offering increased iSCSI performance for Ethernet environments.
  - NVMe-over-Fibre Channel on 16/32 Gb Fibre Channel adapters to allow end-to-end NVMe IO from supported Host Operating Systems.
- Virtualization: Supporting the external virtualization of over 450 (IBM and non-IBM branded) storage arrays over both Fibre Channel and iSCSI.
- Availability: Stretched Cluster and HyperSwap® for high availability among physically separated data centers. Or in a single site environment, Virtual Disk Mirroring for two redundant copies of LUN and higher data availability.
- Thin-provisioning: Helps improve efficiency by allocating disk storage space in a flexible manner among multiple users, based on the minimum space that is required by each user at any time.
- Data migration: Enables easy and nondisruptive moves of volumes from another storage system to the IBM FlashSystem 9100 system by using FC connectivity.

- Distributed RAID: Optimizing the process of rebuilding an array in the event of drive failures for better availability and faster rebuild times, minimizing the risk of an array outage by reducing the time taken for the rebuild to complete.
- Scalability: Clustering for performance and capacity scalability, by combining up-to 4 control enclosures together in the same cluster or connecting up-to 20 expansion enclosures.
- Simple GUI: Simplified management with the intuitive GUI enables storage to be quickly deployed and efficiently managed.
- Easy Tier technology: This feature provides a mechanism to seamlessly migrate data to the most appropriate tier within the IBM FlashSystem 9100 system.
- Automatic re-striping of data across storage pools: When growing a storage pool by adding more storage to it, IBM FlashSystem 9100 Software can restripe your data on pools of storage without having to implement any manual or scripting steps.
- FlashCopy: Provides an instant volume-level, point-in-time copy function. With FlashCopy and snapshot functions, you can create copies of data for backup, parallel processing, testing, and development, and have the copies available almost immediately.
- Encryption: The system provides optional encryption of data at rest, which protects against the potential exposure of sensitive user data and user metadata that is stored on discarded, lost, or stolen storage devices.
- Data Reduction Pools: Helps improve efficiency by compressing data by as much as 80%, enabling storage of up to 5x as much data in the same physical space.
- Remote mirroring: Provides storage-system-based data replication by using either synchronous or asynchronous data transfers over FC communication links:
  - Metro Mirror maintains a fully synchronized copy at metropolitan distances (up to 300 km).
  - Global Mirror operates asynchronously and maintains a copy at much greater distances (up to 250 milliseconds round-trip time when using FC connections).

  Both functions support VMware Site Recovery Manager to help speed DR. IBM FlashSystem 9100 remote mirroring interoperates with other IBM FlashSystem 9100, IBM FlashSystem V840, SAN Volume Controller, and IBM Storwize® V7000 storage systems.

For more information, go to the IBM Spectrum Virtualize website:
http://www03.ibm.com/systems/storage/spectrum/virtualize/index.html

IBM FlashSystems 9100

For decades, IBM has offered a range of enterprise class high-performance, ultra-low latency storage solutions. Now, IBM FlashSystem 9100 (Figure 14) combines the performance of flash and end-to-end NVMe with the reliability and innovation of IBM FlashCore technology and the rich feature set and high availability of IBM Spectrum Virtualize.

This powerful new storage platform provides:

- The option to use IBM FlashCore modules (FCMs) with performance neutral, inline-hardware compression, data protection and innovative flash management features provided by IBM FlashCore technology, or industry-standard NVMe flash drives.

- The software-defined storage functionality of IBM Spectrum Virtualize with a full range of industry-leading data services such as dynamic tiering, IBM FlashCopy management, data mobility and high-performance data encryption, among many others.
- Innovative data-reduction pool (DRP) technology that includes deduplication and hardware-accelerated compression technology, plus SCSI UNMAP support and all the thin provisioning, copy management and efficiency you'd expect from storage based on IBM Spectrum Virtualize.

Figure 14 IBM FlashSystem FS9100

The FlashSystem FS9100 series is comprised of two models; FS9110 and FS9150. Both storage arrays are dual, Active-Active controllers with 24 dual-ported NVMe drive slots. These NVMe slots cater for both traditional SSD drives, as well as the newly redesigned IBM FlashCore Modules.

The IBM FlashSystem 9100 system has two different types of enclosures: control enclosures and expansion enclosures:

- Control Enclosures
  - Each control enclosure can have multiple attached expansion enclosures, which expands the available capacity of the whole system. The IBM FlashSystem 9100 system supports up to four control enclosures and up to two chains of SAS expansion enclosures per control enclosure.
  - Host interface support includes 16 Gb or 32 Gb Fibre Channel (FC), and 10 Gb or 25Gb Ethernet adapters for iSCSI host connectivity. Advanced Encryption Standard (AES) 256 hardware-based encryption adds to the rich feature set.
  - The IBM FlashSystem 9100 control enclosure supports up to 24 NVMe capable flash drives in a 2U high form factor.
  - There are two standard models of IBM FlashSystem 9100: 9110-AF7 and 9150-AF8. There are also two utility models of the IBM FlashSystem 9100: the 9110-UF7 and 9150-UF8.
  - The FS9110 has a total of 32 cores (16 per canister) while the 9150 has 56 cores (28 per canister). The FS9100 supports six different memory configurations as shown in Table 1

Table 1 FS9100 Memory Configurations

Memory per Canister	Memory per Control Enclosure
64 GB	128 GB
128 GB	256 GB
192 GB	384 GB
384 GB	768 GB

576 GB	1152 GB
768 GB	1536 GB

- Expansion Enclosures
- New SAS-based small form factor (SFF) and large form factor (LFF) expansion enclosures support flash-only MDisks in a storage pool, which can be used for IBM Easy Tier®:
  - The new IBM FlashSystem 9100 SFF expansion enclosure Model AAF offers new tiering options with solid-state drive (SSD flash drives). Up to 480 drives of serial-attached SCSI (SAS) expansions are supported per IBM FlashSystem 9100 control enclosure. The expansion enclosure is 2U high.
  - The new IBM FlashSystem 9100 LFF expansion enclosure Model A9F offers new tiering options with solid-state drive (SSD flash drives). Up to 736 drives of serial-attached SCSI (SAS) expansions are supported per IBM FlashSystem 9100 control enclosure. The expansion enclosure is 5U high.

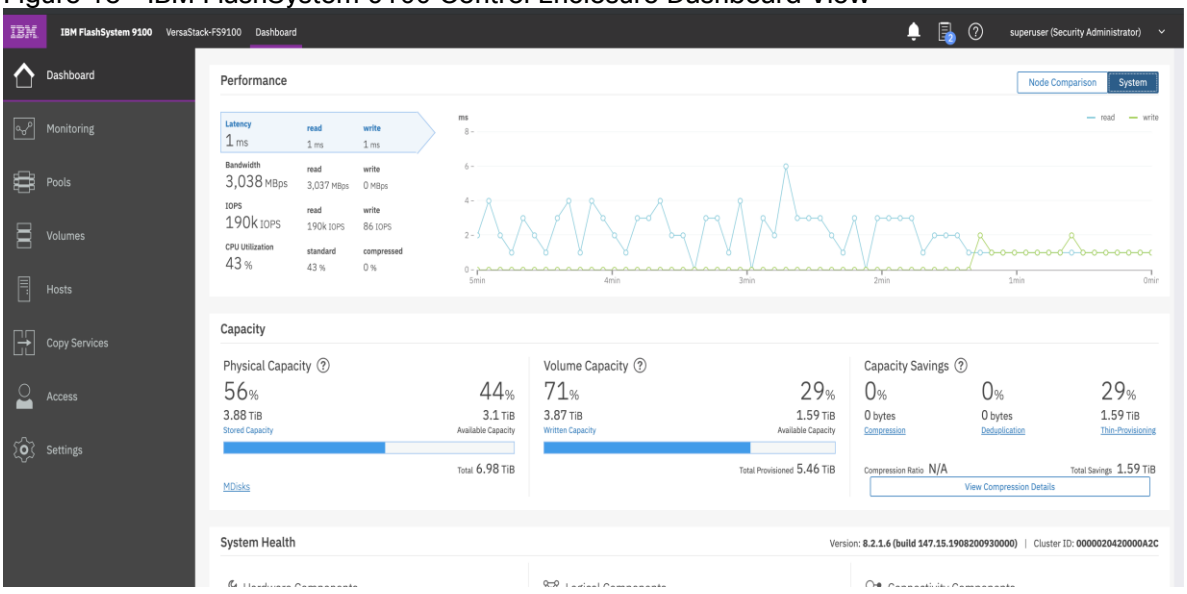
  The FS9100 supports NVMe attached flash drives, both the IBM Flash Core Modules (FCM) and commercial off the shelf (COTS) SSDs. The IBM FCMs support hardware compression and encryption with no reduction in performance. IBM offers the FCMs in three capacities: 4.8 TB, 9.6 TB and 19.2 TB, Standard NVMe SSDs are offered in four capacities, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB.

System Management and the Browser Interface

The IBM FlashSystem 9100 includes a single easy-to-use management graphical user interface (GUI) to help monitor, manage, and configure the system. The IBM FlashSystem 9100 system introduces an improved GUI with the same look and feel as other IBM FlashSystem solutions for a consistent management experience across all platforms. The GUI has an improved overview dashboard that provides all information in an easy-to-understand format and allows visualization of effective capacity.

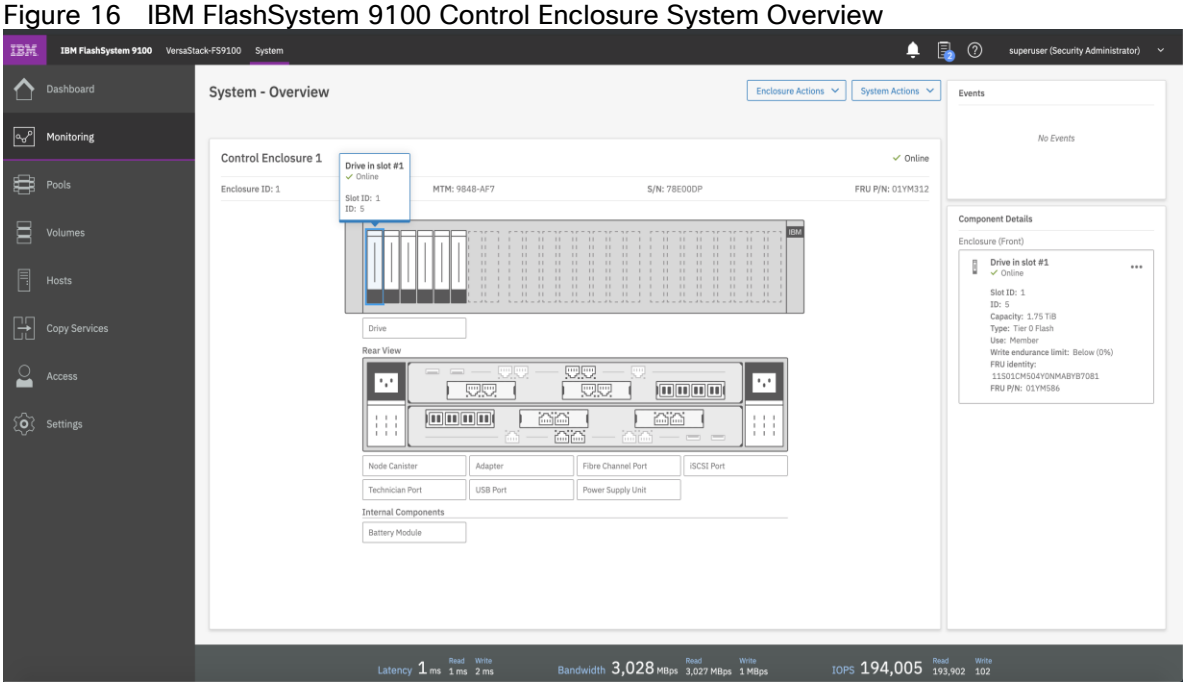
Figure 15 shows the IBM FlashSystem 9100 dashboard view. This is the default view that is displayed after the user logs on to the IBM FlashSystem 9100 system.

Figure 15 IBM FlashSystem 9100 Control Enclosure Dashboard View



In Figure 16, the GUI shows one IBM FlashSystem 9100 Control Enclosure. This is the System Overview window.

Figure 16 IBM FlashSystem 9100 Control Enclosure System Overview



The IBM FlashSystem 9100 system includes a CLI, which is useful for scripting, and an intuitive GUI for simple and familiar management of the product. RESTful API support was recently introduced to allow workflow automation or integration into DevOps environments.

The IBM FlashSystem 9100 system supports Simple Network Management Protocol (SNMP), email forwarding that uses Simple Mail Transfer Protocol (SMTP), and syslog redirection for complete enterprise management access.

VMware vSphere 6.7 Update 3

VMware vSphere is a virtualization platform for holistically managing large collections of infrastructures (resources-CPUs, storage and networking) as a seamless, versatile, and dynamic operating environment. Unlike traditional operating systems that manage an individual machine, VMware vSphere aggregates the infrastructure of an entire data center to create a single powerhouse with resources that can be allocated quickly and dynamically to any application in need.

vSphere 6.7 Update 3 (U3) provides several improvements including, but not limited to:

- ixgben driver enhancements
- VMXNET3 enhancements
- bnxtnet driver enhancements
- QuickBoot support enhancements
- Configurable shutdown time for the sfcbd service
- NVIDIA virtual GPU (vGPU) enhancements
- New SandyBridge microcode

Solution Design

The VersaStack design discussed in this document aligns with the converged infrastructure configurations and best practices identified in previous VersaStack releases. This solution focuses on integration of IBM Flash System 9100 in to VersaStack architecture with Cisco ACI and support for VMware vSphere 6.7 U3.

Requirements

The VersaStack data center is intended to provide a Virtual Server Infrastructure (VSI) that becomes the foundation for hosting virtual machines and applications. This design assumes existence of management, network and routing infrastructure to provide necessary connectivity, along with the availability of common services such as DNS and NTP, and so on.

This VersaStack solution meets the following general design requirements:

- Resilient design across all layers of the infrastructure with no single point of failure.
- Scalable design with the flexibility to add compute capacity, storage, or network bandwidth as needed.
- Modular design that can be replicated to expand and grow as the needs of the business grow.
- Flexible design that can support components beyond what is validated and documented in this guide.
- Simplified design with ability to automate and integrate with external automation and orchestration tools.
- Extensible design with support for extensions to existing infrastructure services and management applications.

The system includes hardware and software compatibility support between all components and aligns to the configuration best practices for each of these components. All the core hardware components and software releases are listed and supported on both the Cisco compatibility list:

http://www.cisco.com/en/US/products/ps10477/prod_technical_reference_list.html

and IBM Interoperability Matrix:

http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss

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

Physical Topology

The VersaStack infrastructure satisfies the high-availability design requirements and is physically redundant across the network, compute and storage stacks. Figure 17 provides a high-level topology of the system connectivity.

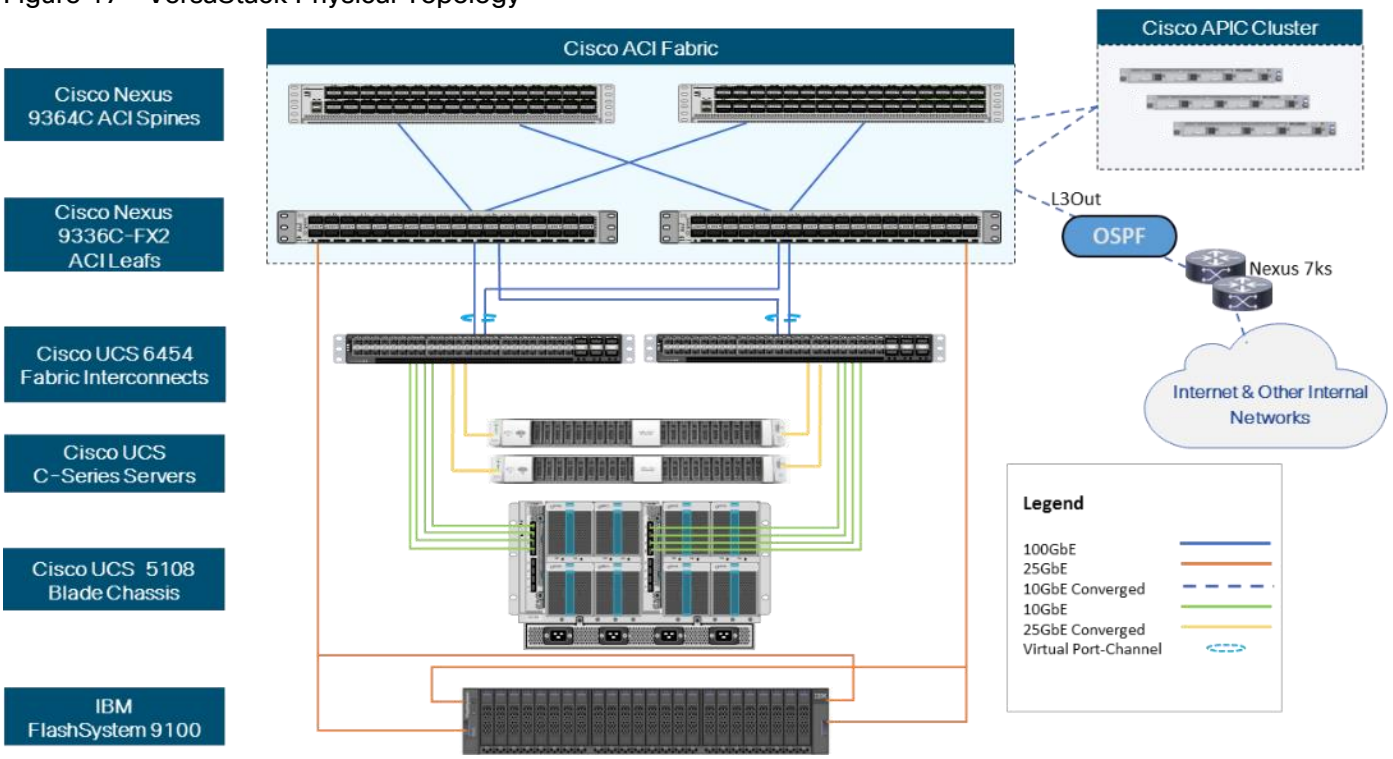
This VersaStack design utilizes Cisco UCS platform with Cisco UCS B200 M5 half-width blades and Cisco UCS C220 M5 servers connected and managed through Cisco UCS 6454 Fabric Interconnects and the integrated Cisco UCS Manager (UCSM). These high-performance servers are configured as stateless compute nodes where ESXi 6.7 U3 hypervisor is loaded using SAN (iSCSI) boot. The boot disks to store ESXi hypervisor image and configuration along with the block based datastores to host application Virtual Machines (VMs) are provisioned on the IBM Flash System 9100 storage array.

As in the non-ACI designs of VersaStack, link aggregation technologies play an important role in VersaStack with ACI solution providing improved aggregate bandwidth and link resiliency across the solution stack. Cisco UCS, and Cisco Nexus 9000 platforms support active port channeling using 802.3ad standard Link Aggregation Control Protocol (LACP). In addition, the Cisco Nexus 9000 series features virtual Port Channel (vPC) capability which allows links that are physically connected to two different Cisco Nexus devices to appear as a single " logical" port channel.

This design has the following physical connectivity between the components of VersaStack:

- 4 X 10 Gb Ethernet connections port-channeled between the Cisco UCS 5108 Blade Chassis and the Cisco UCS Fabric Interconnects
- 25 Gb Ethernet connections between the Cisco UCS C-Series rackmounts and the Cisco UCS Fabric Interconnects
- 100 Gb Ethernet connections port-channeled between the Cisco UCS Fabric Interconnect and Cisco Nexus 9000 ACI leaf's
- 100 Gb Ethernet connections between the Cisco Nexus 9000 ACI Spine's and Nexus 9000 ACI Leaf's
- 25 Gb Ethernet connections between the Cisco Nexus 9000 ACI Leaf's and IBM Flash System 9100 storage array for iSCSI block storage access

Figure 17 VersaStack Physical Topology

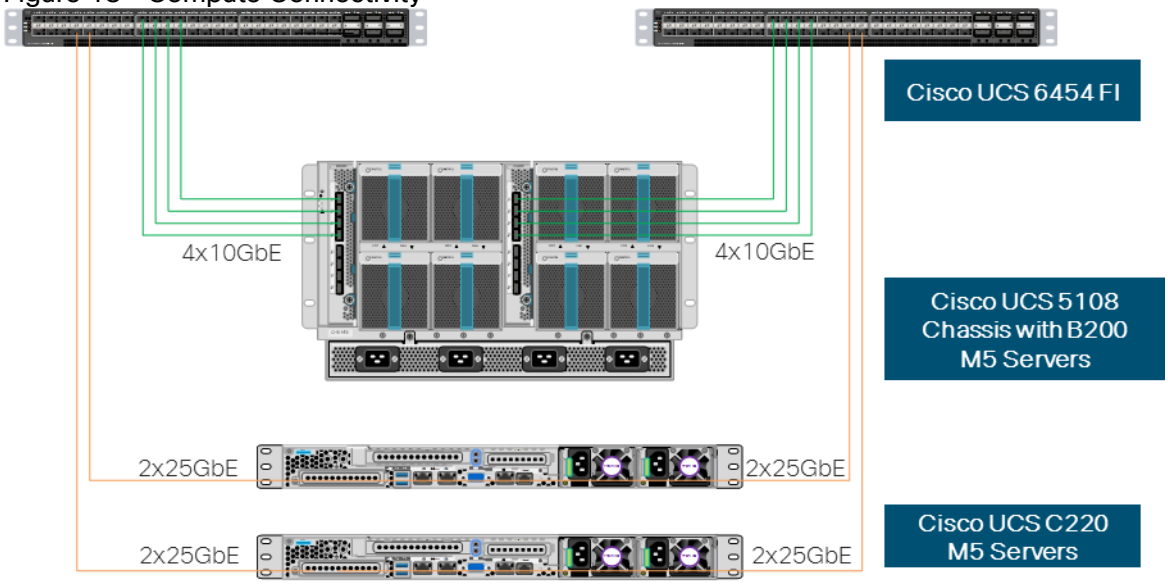


Compute Connectivity

The VersaStack compute design supports both Cisco UCS B-Series and C-Series deployments. Cisco UCS supports the virtual server environment by providing robust, highly available, and integrated compute resources centrally managed from Cisco UCS Manager in the Enterprise or from Cisco Intersight Software as a Service

(SaaS) in the cloud. In this validation effort, multiple Cisco UCS B-Series and C-Series ESXi servers are booted from SAN using iSCSI storage presented from the IBM Flash System 9100.

Figure 18 Compute Connectivity



The 5108 chassis in the design is populated with Cisco UCS B200 M5 blade servers and each of these blade servers contain one physical network adapter (Cisco VIC 1440) that passes converged fibre channel and ethernet traffic through the chassis mid-plane to the 2208XP FEXs. The FEXs are redundantly connected to the fabric interconnects using 4X10Gbps ports per FEX to deliver an aggregate bandwidth of 80Gbps to the chassis. Full population of each 2208XP FEX can support 8x10Gbps ports, providing an aggregate bandwidth of 160Gbps to the chassis.

The connections from the Cisco UCS Fabric Interconnects to the FEXs are automatically configured as port channels by specifying a Chassis/FEX Discovery Policy within UCSM.

Each Cisco UCS C-Series rack server in the design is redundantly connected to the fabric interconnects with at least one port connected to each FI to support converged traffic as with the Cisco UCS B-Series servers. Internally the Cisco UCS C-Series servers are equipped with a Cisco VIC 1457 network interface card (NIC) with quad 10/25 Gigabit Ethernet (GbE) ports. In this design, the Cisco VIC is installed in a modular LAN on motherboard (MLOM) slot, but it can also be installed on a PCIe slot using VIC 1455. The standard practice for redundant connectivity is to connect port 1 of each server's VIC card to a numbered port on FI A, and port 3 of each server's VIC card to the same numbered port on FI B. The use of ports 1 and 3 are because ports 1 and 2 form an internal port-channel, as does ports 3 and 4. This allows an optional 4 cable connection method providing an effective 50GbE bandwidth to each fabric interconnect.

Cisco UCS Server Configuration for VMware vSphere

The Cisco UCS servers are stateless and are deployed using Cisco UCS Service Profiles (SP) that consists of server identity information pulled from pools (WWPN, MAC, UUID, and so on) as well as policies covering connectivity, firmware and power control options, and so on. The service profiles are provisioned from the Cisco UCS Service Profile Templates that allow rapid creation, as well as guaranteed consistency of the hosts at the Cisco UCS hardware layer.

The ESXi nodes consist of Cisco UCS B200 M5 blades or Cisco UCS C220 M5 rack servers with Cisco UCS 1400 series VIC. These nodes are allocated to a VMware High Availability cluster to support infrastructure services and applications. At the server level, the Cisco 1400 VIC presents multiple virtual PCIe devices to the ESXi node

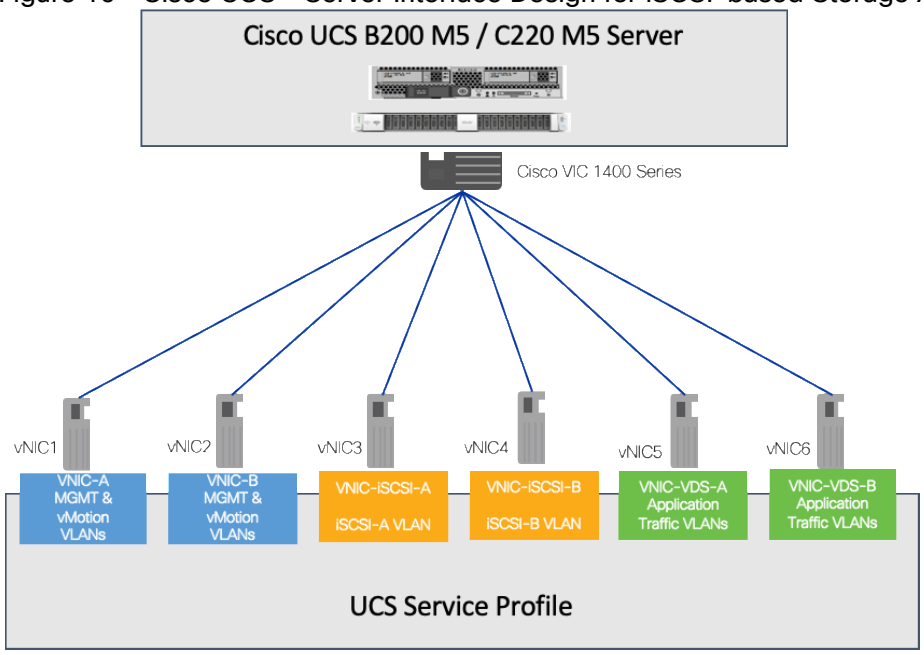
and the vSphere environment identifies these interfaces as vmnics or vmhbas. The ESXi operating system is unaware of the fact that the NICs or HBAs are virtual adapters.

In the VersaStack design with iSCSI storage, six vNICs are created and utilized as follows (Figure 19):

- One vNIC (iSCSI-A) for iSCSI SAN traffic
- One vNIC (iSCSI-B) for iSCSI SAN traffic
- Two vNICs for in-band management and vMotion traffic
- Two vNICs for application virtual machines hosted on the infrastructure. These vNICs are assigned to a distributed switch (vDS) managed by Cisco ACI

These vNICs are pinned to different Fabric Interconnect uplink interfaces and are assigned to separate vSwitches and vSphere distributed switches (VDS) based on type of traffic. The vNIC to vSwitch and vDS assignment is explained later in the document.

Figure 19 Cisco UCS – Server Interface Design for iSCSI-based Storage Access



IBM Storage Systems

IBM FlashSystem 9100 explained in this VersaStack design, is deployed as high availability storage solution. IBM storage systems support fully redundant connections for communication between control enclosures, external storage, and host systems.

Each storage system provides redundant controllers and redundant iSCSI and FC paths to each controller to avoid failures at path as well as hardware level. For high availability, the storage systems are attached to two separate fabrics, SAN-A and SAN-B. If a SAN fabric fault disrupts communication or I/O operations, the system recovers and retries the operation through the alternative communication path. Host (ESXi) systems are configured to use ALUA multi-pathing, and in case of SAN fabric fault or node canister failure, the host seamlessly switches over to alternate I/O path.

IBM FlashSystem 9100 Storage

A basic configuration of an IBM FlashSystem 9100 storage platform consists of one IBM FlashSystem 9100 Control Enclosure. For a balanced increase of performance and scale, up to four IBM FlashSystem 9100 Control Enclosures can be clustered into a single storage system, multiplying performance and capacity with each addition.

The IBM FlashSystem 9100 Control Enclosure node canisters are configured for active-active redundancy. The node canisters run a highly customized Linux-based OS that coordinates and monitors all significant functions in the system. Each Control Enclosure is defined as an I/O group and can be visualized as an isolated appliance resource for servicing I/O requests.

In this design guide, one pair of FS9100 node canisters (I/O Group 0) were deployed with in a single FS9100 Control Enclosure. The storage configuration includes defining logical units with capacities, access policies, and other parameters.

Based on the specific storage requirements and scale, the number of I/O Groups in customer deployments will vary.

IBM FlashSystem 9100 – iSCSI Connectivity

To support iSCSI-based IP storage connectivity with redundancy, each IBM FS9100 node canister is connected to each of the Cisco Nexus 9336C-FX2 leaf switches for iSCSI boot and VMware datastore access. The physical connectivity is shown in Figure 20. Two 25GbE ports from each IBM FS9100 are connected to each of the two Cisco Nexus 9336C-FX2 switches providing an aggregate bandwidth of 100Gbps for storage access. The 25Gbps Ethernet ports between the FS9100 I/O Group and the Nexus fabric are utilized by redundant iSCSI-A and iSCSI-B paths, providing redundancy for link and device failures. Additional links can be added between the storage and network components for additional bandwidth if needed.

The Nexus 9336C-FX2 switches used in the design support 10/25/40/100 Gbps on all the ports. The switch supports breakout interfaces, where each 100Gbps port on the switch can be split in to 4 X 25Gbps interfaces. In this design, a breakout cable is used to connect the 25Gbps iSCSI ethernet ports on the FS9100 storage array to the 100Gbps QSFP port on the switch end. With this connectivity, IBM SFP transceivers on the FS9100 are not required.

Connectivity between the Nexus switches and IBM FS9100 for iSCSI access depends on the Nexus 9000 switch model used within the architecture. If other supported models of Nexus switches with 25Gbps capable SFP ports are used, breakout cable is not required and ports from the switch to IBM FS9100 can be connected directly using the SFP transceivers on both sides.

Figure 20 IBM FS9100 – iSCSI Connectivity with Nexus 9336C-FX2 ACI Leaf Switch

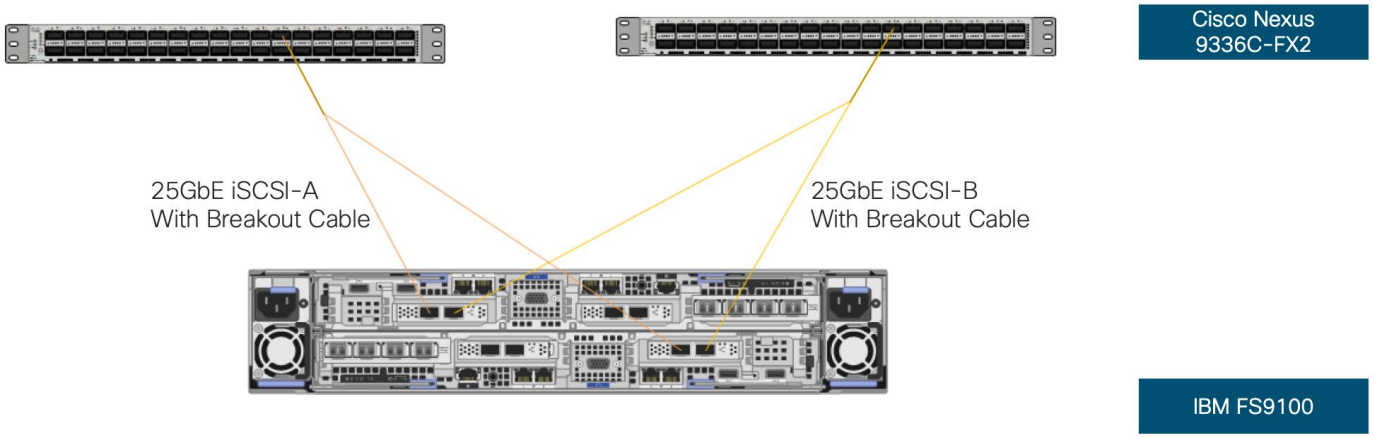
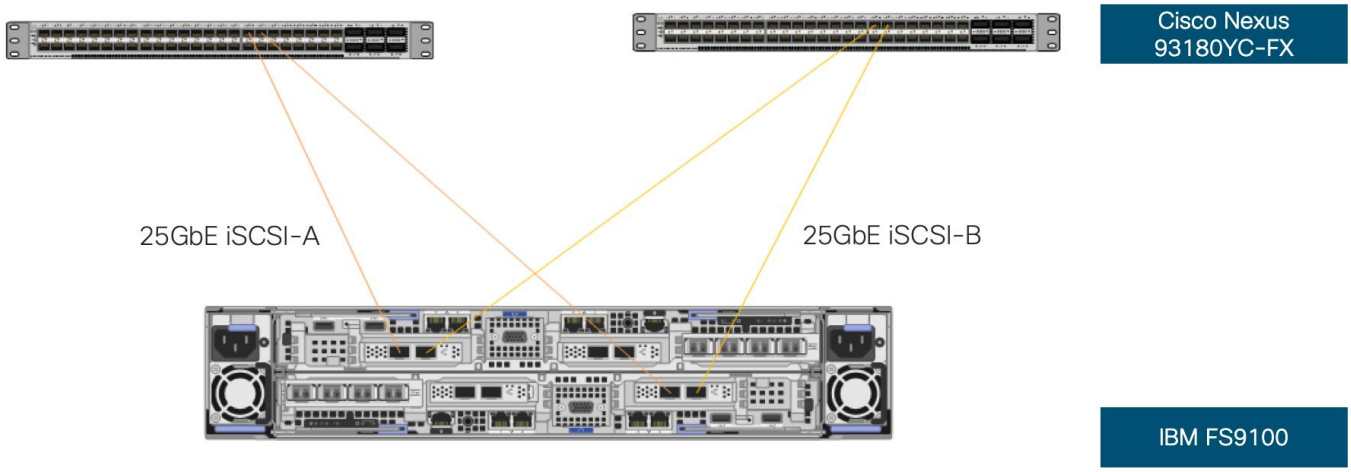


Figure 21 illustrates direct connectivity using SFP transceivers using a supported Nexus switch model that supports 25Gbps SFP ports – for example, between 93180YC-FX leaf switches and IBM FS9100. Other supported models of Nexus 9000 series switches with SFP ports can also be used for direct connectivity with the FS9100 storage array.

Figure 21 Example: IBM FS9100 – iSCSI Connectivity with Nexus 93180YC-FX ACI Leaf Switch



Host Clusters

When managing how volumes are mapped to Hosts, IBM Spectrum Virtualize incorporates the concept of Hosts and Host Clusters. In VersaStack configuration, each VMware ESXi (or physical server) instance should be defined as an independent Host object within FS9100. If each VMware ESXi host has multiple associated FC WWPN ports (when using FibreChannel) or IQN ports when using iSCSI, it is recommended that all ports associated with each physical host be contained within a single host object.

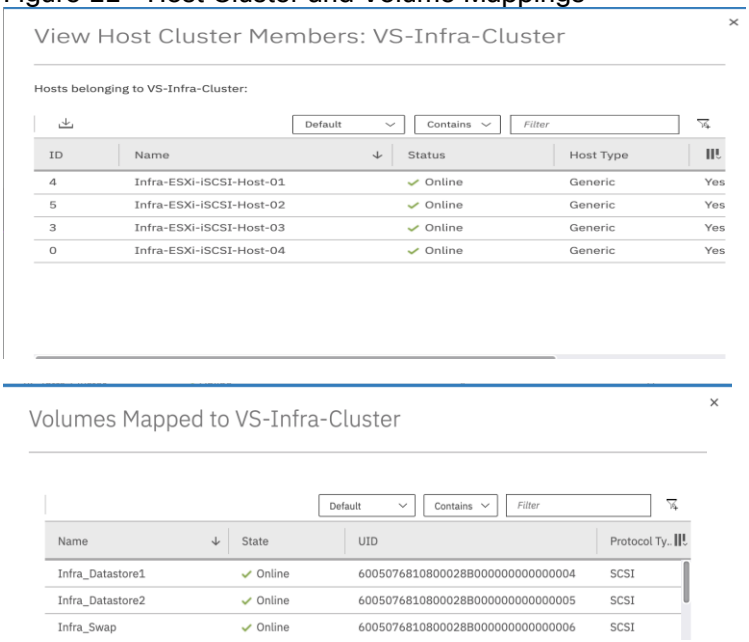
When using vSphere clustering where storage resources (data stores) are expected to be shared between multiple VMware ESXi hosts, it is recommended that a Host Cluster be defined for each vSphere cluster. When mapping volumes from the FS9100 designed for VMFS Datastores, shared Host Cluster mappings should be used. The benefits are as follows:

- All members of the vSphere cluster will inherit the same storage mappings

- SCSI LUN IDs are consistent across all members of the vSphere cluster
- Simplified administration of storage when adding/removing vSphere cluster members
- Better visibility of the Host/Host Cluster state if particular ports/SAN become disconnected

However, when using SAN boot volumes, ensure that these are mapped to the specific host via private mappings. This will ensure that they remain accessible to only the corresponding VMware ESXi host.

Figure 22 Host Cluster and Volume Mappings



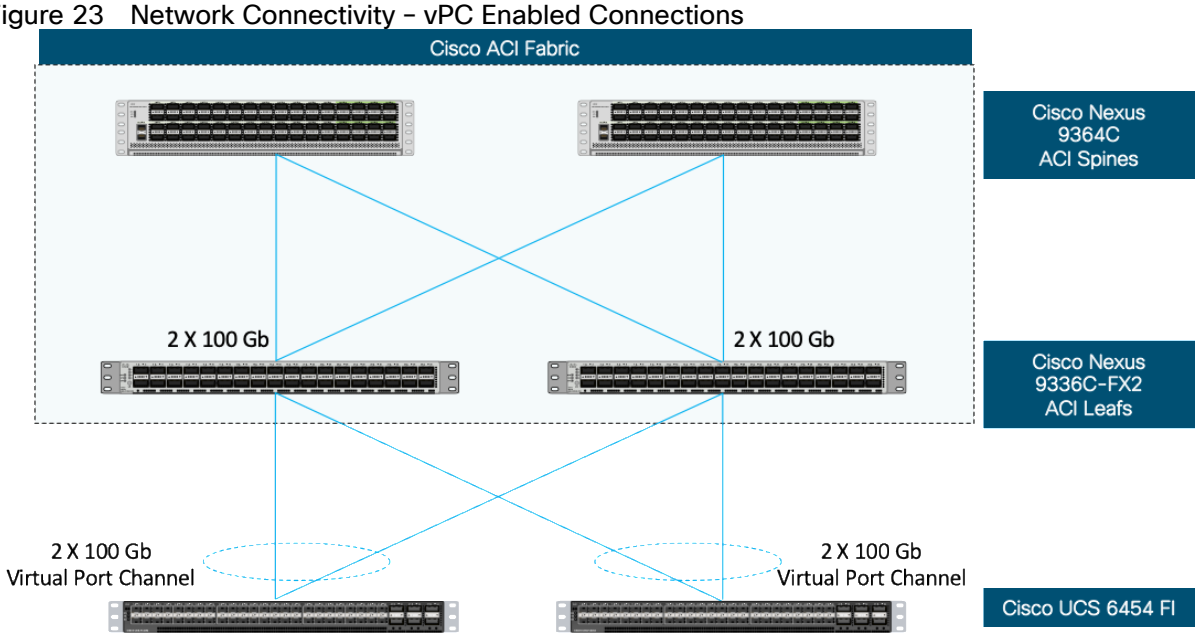
VersaStack Network Connectivity and Design

In this VersaStack design, a pair of redundant Cisco Nexus 9336C-FX2 leaf switches provide ACI based Ethernet switching fabric for iSCSI storage access for the compute and application communication. A second pair of Nexus 9000 leaf switch provides connectivity to existing enterprise (non-ACI) networks. Like previous versions of VersaStack, the core network constructs such as virtual port channels (vPC) and VLANs plays an important role in providing the necessary Ethernet based IP connectivity.

Virtual Port-Channel Design

In the current VersaStack with Cisco ACI and IBM FS9100 design, Cisco UCS FIs are connected to the ACI fabric using a vPC. Network reliability is achieved through the configuration of virtual Port Channels within the design as shown in Figure 23.

Figure 23 Network Connectivity – vPC Enabled Connections



Virtual Port Channel allows Ethernet links that are physically connected to two different Cisco Nexus 9336C-FX2 ACI Leaf switches to appear as a single Port Channel. vPC provides a loop-free topology and enables fast convergence if either one of the physical links or a device fails. In this design, two 100G ports from the 40/100G capable ports on the 6454 (1/49-54) were used for the virtual port channels.

Application Centric Infrastructure Design

The Cisco ACI design consists of Cisco Nexus 9500 and 9300 based spine/leaf switching architecture controlled using a cluster of three Application Policy Infrastructure Controllers (APICs). With the Nexus switches in place, the platform delivers an intelligently designed, high port density, low latency network, supporting up to 400G connectivity.

The Cisco Application Centric Infrastructure (ACI) fabric consists of discrete components that operate as routers and switches but are provisioned and monitored as a single entity. These components and the integrated management allow Cisco ACI to provide advanced traffic optimization, security, and telemetry functions for both virtual and physical workloads. This CVD utilizes Cisco ACI fabric-based networking as discussed in the upcoming sections.

Cisco ACI Fabric Components

The following are the ACI Fabric components:

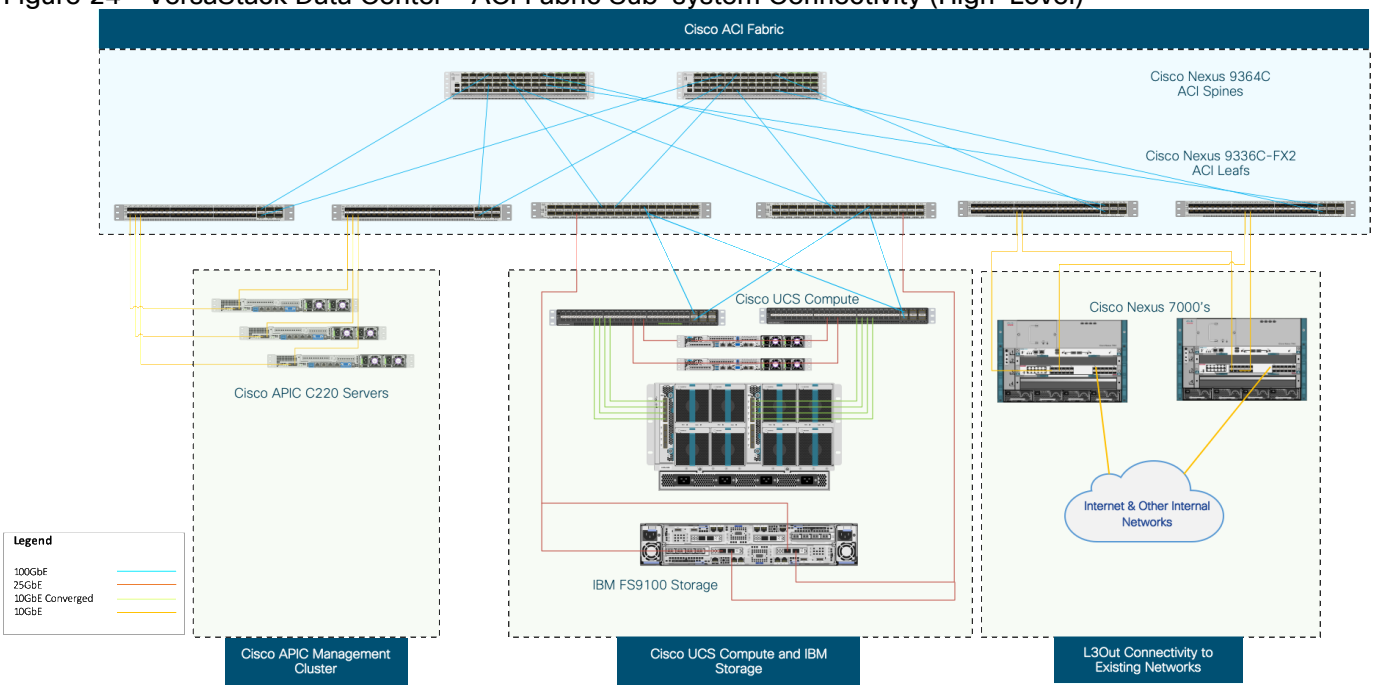
- Cisco APIC: The Cisco Application Policy Infrastructure Controller (APIC) is the unifying point of automation and management for the Cisco ACI fabric. The Cisco APIC provides centralized access to all fabric information, optimizes the application lifecycle for scale and performance, and supports flexible application provisioning across physical and virtual resources. The Cisco APIC exposes northbound APIs through XML and JSON and provides both a command-line interface (CLI) and GUI which utilize the APIs to manage the fabric.
- Leaf Switches: The ACI leaf provides physical connectivity for servers, storage devices and other access layer components as well as enforces ACI policies. A leaf typically is a fixed form factor switch such as the Cisco Nexus 9336C-FX2 switch used in the current design. Leaf switches also provide connectivity to

existing enterprise or service provider infrastructure. The leaf switches provide options starting at 1G up through 100G Ethernet ports for connectivity.

- In the VersaStack with ACI design, Cisco UCS FI, IBM FS9100 and Cisco Nexus 7000 based WAN/Enterprise routers are connected to leaf switches, each of these devices are redundantly connected to a pair of leaf switches for high availability.
- Spine Switches: In ACI, spine switches provide the mapping database function and connectivity between leaf switches. A spine switch can be the modular Cisco Nexus 9500 series equipped with ACI ready line cards or fixed form-factor switch such as the Cisco Nexus 9364C (used in this design). Spine switches provide high-density 40/100 Gigabit Ethernet connectivity between the leaf switches.

Figure 24 shows the VersaStack ACI fabric with connectivity to Cisco UCS Compute, IBM FS9100 storage, APIC Cluster for management and existing enterprise networks via Cisco Nexus 7000's:

Figure 24 VersaStack Data Center – ACI Fabric Sub-system Connectivity (High-Level)



This design assumes that the customer already has an ACI fabric in place with spine switches and APICs deployed and connected through a pair of leaf switches. In this design, an existing ACI Fabric core consisting a pair of Nexus 9364C series spine switches, a 3-node APIC cluster and a pair of Nexus 9000 series leaf switches that the Cisco APICs connect into was leveraged.

The ACI fabric can support many models of Nexus 9000 series switches as spine and leaf switches. Customers can use the models that match the interface types, speeds and other capabilities that the deployment requires - the design of the existing ACI core is outside the scope of this document. The design guidance in this document therefore focusses on attaching a Cisco UCS domain and IBM FS9100 to the existing ACI fabric and the connectivity and services required for enabling an end-to-end converged datacenter infrastructure.

The access layer connections on the ACI leaf switches to the different sub-systems in this design are summarized below:

- Cisco APICs that manage the ACI Fabric (3-node cluster)

  - Cisco APIC that manages the ACI fabric is redundantly connected to a pair of ACI leaf switches using 2x10GbE links. For high availability, an APIC cluster with 3 APICs are used in this design. The APIC cluster connects to an existing pair of ACI leaf switches in the ACI fabric.

- Cisco UCS Compute Domain (Pair of Cisco UCS Fabric Interconnects) with UCS Servers
  - A Cisco UCS Compute domain consisting of a pair of Cisco UCS 6400 Fabric Interconnects, connect into a pair of Nexus 9336C-FX2 leaf switches using port-channels, one from each FI. Each FI connects to the leaf switch-pair using member links from one port-channel.
- IBM FS9100 Storage Array
  - An IBM FS9100 control enclosure with two node canisters connect into a pair of Nexus 9336C-FX2 leaf switches using access ports, one from each node canister.
- Nexus 7000 series switches (Gateways for L3Out)
  - Nexus 7k switches provide reachability to other parts of the customer's network (Outside Network) including connectivity to existing Infrastructure where NTP, DNS, etc reside. From the ACI fabric's perspective, this is a L3 routed connection.

Cisco ACI Fabric Management Design

The APIC management model divides the Cisco ACI fabric configuration into these two categories:

- Fabric infrastructure configurations: This is the configuration of the physical fabric in terms of vPCs, VLANs, loop prevention features, Interface/Switch access policies, and so on. The fabric configuration enables connectivity between ACI fabric and access or outside (ACI) components in the enterprise network.
- Tenant configurations: These configurations are the definition of the logical constructs such as application profiles, bridge domains, EPGs, and so on. The tenant configuration enables forwarding across the ACI fabric and the policies that determine the forwarding.

ACI Fabric Infrastructure Design for VersaStack

This section describes the fabric infrastructure configurations for ACI physical connectivity as part of the VersaStack design.

Fabric Access Policies

Fabric Access Policies are an important aspect of the Cisco ACI architecture. Fabric Access Policies are defined by the Fabric Administrator and includes all the configuration and policies required to connect access layer devices to the ACI fabric. This must be in place before Tenant Administrators can deploy Application EPGs. These policies are designed to be reused as new leaf switches and access layer devices are connected to the fabric. The Fabric Access Policies at a high-level fall into the following categories (Figure 25).

Figure 25 Access Policies Categories

Fabric Access refers to access layers connections at the fabric edge to outside (ACI) devices such as:

- Physical Servers (Cisco UCS Rackmount servers, IBM Storage Controllers)
- Layer 2 Bridged Devices (Switches, Cisco UCS FI)
- Layer 3 Gateway Devices (Routers)
- Hypervisors (ESXi) and Virtual Machine Managers (VMware vCenter)

Access Policies include configuration and policies that are applied to leaf switch interfaces that connect to edge devices. Ideally, policies should be created once and reused when connecting new devices to the fabric. Maximizing reusability of policy and objects makes day-to-day operations exponentially faster and easier to make large-scale changes. They include:

- Configuring Interface Policies: Interface policies dictate interface behavior and are later tied to interface policy groups. For example, there should be a policy that dictates if CDP or LLDP is disabled and a policy that dictates if CDP or LLDP is enabled; these can be reused as new devices are connected to the leaf switches.
- Interface policy groups: Interface Policy Groups are templates to dictate port behavior and are associated to an AEP. Interface policy groups use the policies described in the previous paragraph to specify how links should behave. These are also reusable objects as many devices are likely to be connected to ports that will require the same port configuration. There are three types of interface policy groups depending on link type: Access Port, Port Channel, and vPC.
- Interface Profiles: Interface profiles include all policies for an interface or set of interfaces on the leaf and help tie the pieces together. Interface profiles contain blocks of ports and are also tied to the interface policy groups described in the previous paragraph. The profile must be associated to a specific switch profile to configure the ports.
- Switch Profiles: Switch profiles allow the selection of one or more leaf switches and associate interface profiles to configure the ports on that specific node. This association pushes the configuration to the interface and creates a Port Channel or vPC if one has been configured in the interface policy.
- Global Policies
  - Domain Profiles: Domains in ACI are used to define how different entities (for example, servers, network devices, storage) connect into the fabric and specify the scope of a defined VLAN pool. ACI defines four domain types based on the type of devices that connect to the leaf switch (physical, external bridged, external routed, and VMM domains).
  - AAEP: The Attachable Access Entity Profile provides a template for attachment point between the switch and interface profiles and the fabric resources such as the VLAN pool. The AEP can be considered the ‘glue’ between the defined physical, virtual or Layer 2 / Layer 3 domains and the fabric interfaces (logical or physical), essentially allowing to specify what VLAN tags are allowed on those interfaces. For VMM Domains, the associated AEP provides the interface profiles and policies (CDP, LACP) for the virtual environment.
  - VLAN pools: Define the range of VLANs that are allowed for use on the interfaces. VLAN pools contain the VLANs used by the EPGs the domain will be tied to. A domain is associated to a single VLAN pool. VXLAN and multicast address pools are also configurable. VLANs are instantiated on leaf switches based on AEP configuration. Allow/deny forwarding decisions are still based on contracts and the policy model, not subnets and VLANs.

In summary, ACI provides attachment points for connecting access layer devices to the ACI fabric. Interface Selector Profiles represents the configuration of those attachment points. Interface Selector Profiles are the

consolidation of a group of interface policies (for example, LACP, LLDP, and CDP) and the interfaces or ports the policies apply to. As with policies, Interface Profiles and AEPs are designed for re-use and can be applied to multiple leaf switches if the policies and ports are the same.

A high-level overview of the Fabric Access Policies in Cisco ACI architecture is shown in Figure 26.

Figure 26 Access Policy Model Overview

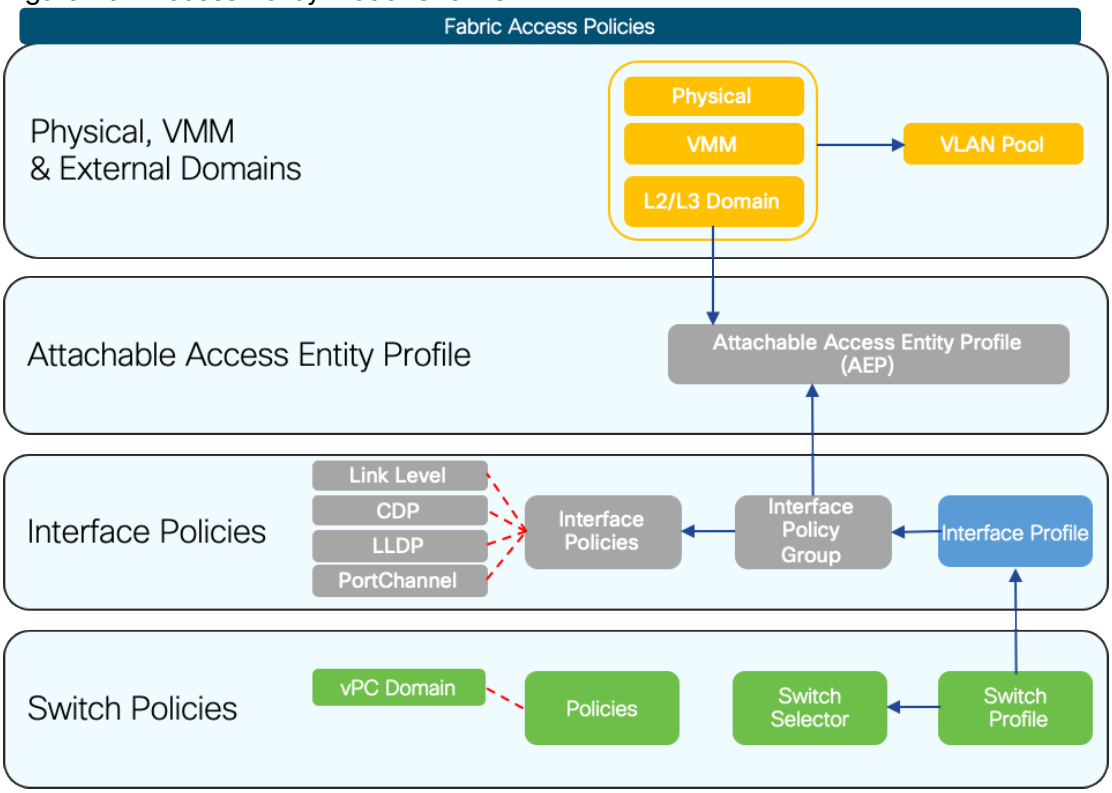


Table 2 lists some of the fabric access policy elements such as AEP's, Domain Name's and the VLAN Pool's used in this validated design:

Table 2 Validated Design – Fabric Access Policies

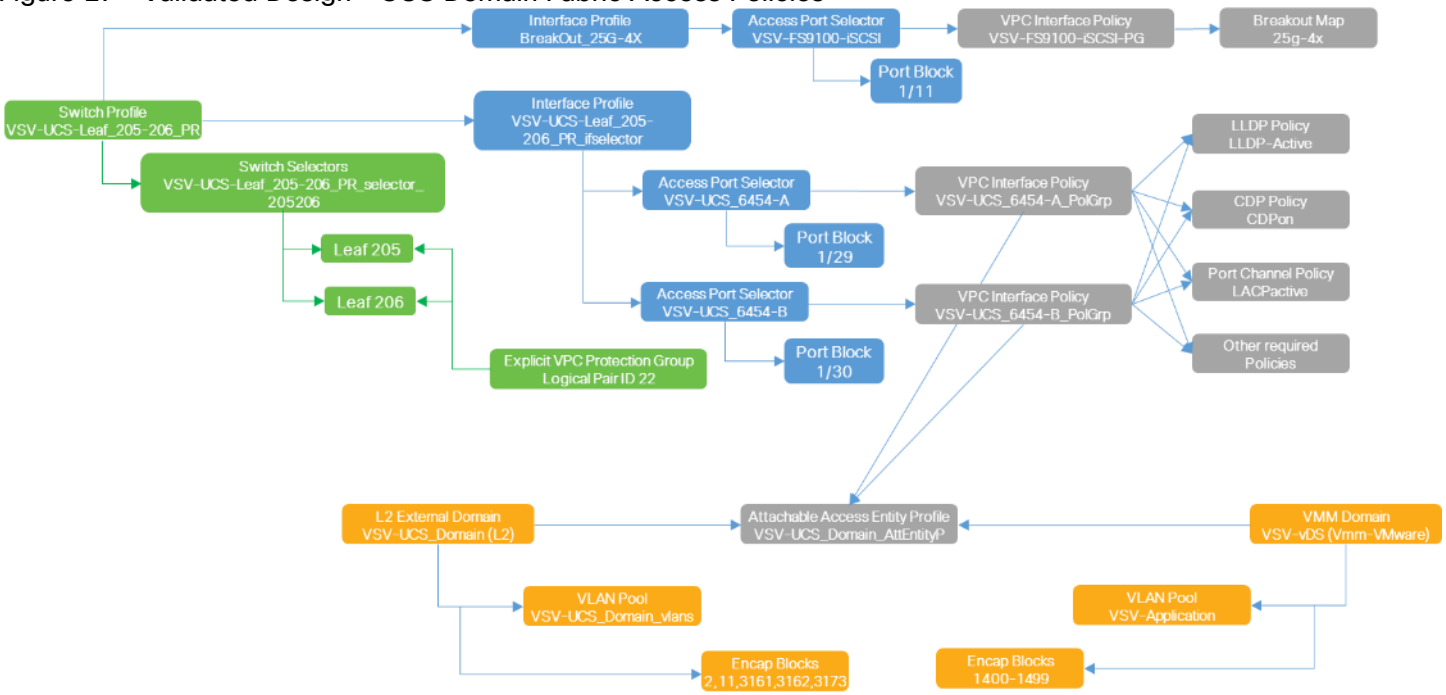
Validated Design - Fabric Access Policies				
Access Connection	AEP	Domain Name	Domain Type	VLAN Pool Name
vPC to Cisco UCS Fabric Interconnects (FI-A, FI-B)	VSV-UCS_Domain_AttEntityP	VSV-UCS_Domain	External Bridged	VSV-UCS_VLANs
vPC to Cisco UCS Fabric Interconnects (FI-A, FI-B)	VSV-UCS_Domain_AttEntityP	VSV-vDS	VMM Domain	VSV-VMM_VLANs
Redundant Connections to a pair of L3 Gateways	AA07N7k-SharedL3Out-AttEntityP	SharedOut-West-Pod2_Domain	External Routed	SharedL3Out-West-Pod2_VLANs
Redundant Connections to a pair of IBM FS9100 Storage Nodes	VSV-FS9100-A_AttEntityP VSV-FS9100-B_AttEntityP	VSV-FS9100-A VSV-FS9100-B	Physical Domains	VSV-FS9100-A_vlans VSV-FS9100-B_vlans

List of VLANs used in the validated design that are part of the defined VLAN pools are discussed in detail in the following sections of this document.

Defining the above fabric access policies for the UCS domain and IBM FS9100 results in the following policies and relationships (Figure 27 and Figure 28) for VersaStack. Once the policies and profiles are in place, they can be re-used to add new leaf switches and connect new endpoints to the ACI fabric. Note that the Policies, Pools and Domains defined are tied to the Policy-Groups and Profiles associated with the physical components (Switches, Modules, and Interfaces) to define the access layer design in the ACI fabric.

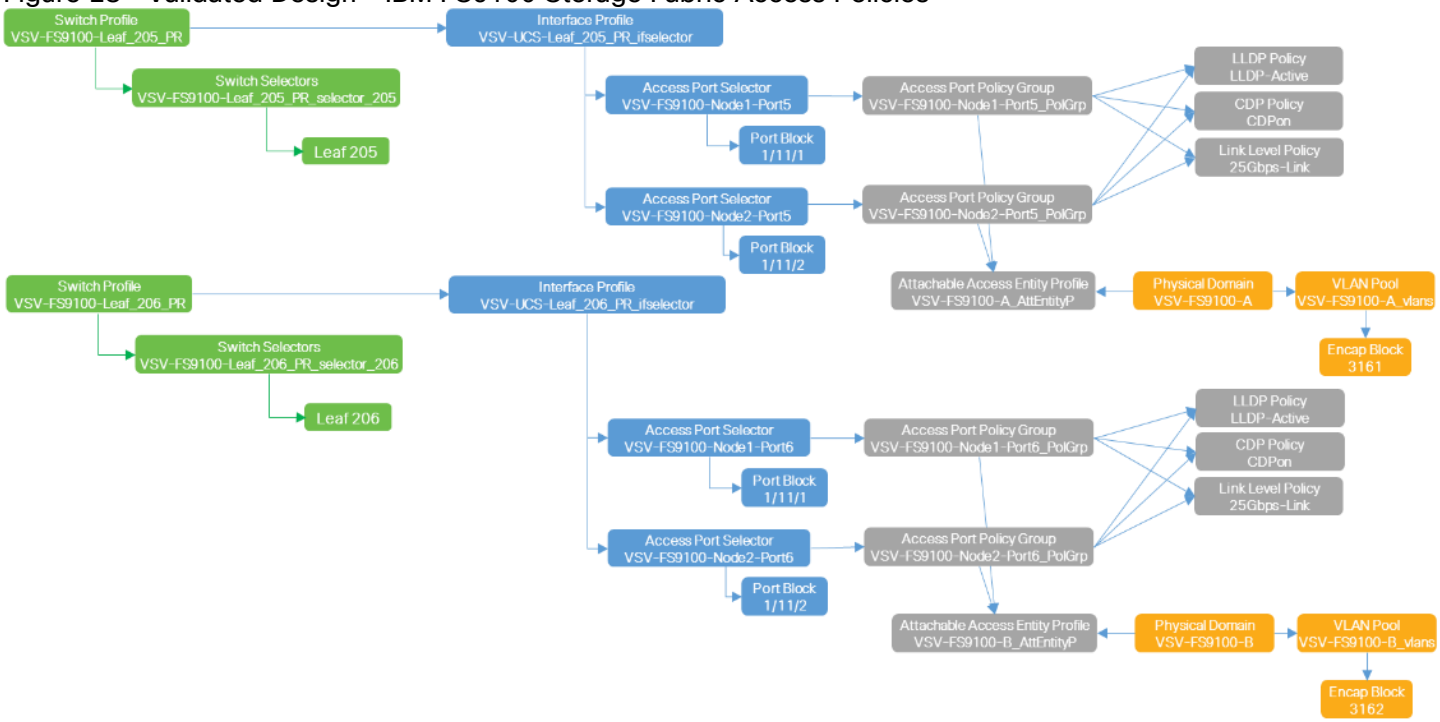
Fabric Access Policies used in the VersaStack design to connect the Cisco UCS and VMM domains are shown in Figure 27:

Figure 27 Validated Design – UCS Domain Fabric Access Policies



The Fabric Access Policies used in the VersaStack design to connect the IBM FS9100 are shown in Figure 28. The breakout policy was configured in addition to the other interface policies to convert a 100 GbE port on the Nexus 9336C-FX2 leaves to 4 X 25 GbE ports for iSCSI connectivity on the IBM FS9100.

Figure 28 Validated Design – IBM FS9100 Storage Fabric Access Policies



Cisco ACI Tenant Design

ACI delivers multi-tenancy using the following ACI constructs:

- Tenant: A tenant is a logical container which can represent an actual tenant, organization, application or a construct to easily organize information. 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, one or more VRF contexts, one or more bridge domains, and one or more EPGs can be defined according to application requirements.

  VersaStack with ACI design recommends the use of an infrastructure tenant called " VSV-Foundation" to isolate all infrastructure connectivity to a single tenant. This tenant will provide compute to storage connectivity for iSCSI-based SAN environment as well as to provide access to the management infrastructure. The design also utilizes the predefined " common" tenant to provide in-band management infrastructure connectivity for hosting core services required by all the tenants such as DNS, AD etc. In addition, each subsequent application deployment requires creation of a dedicated tenant.

- 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. Because VRFs use separate forwarding instances, IP addressing can be duplicated across VRFs for multitenancy. In the current design, each tenant typically used a single VRF.

- Application Profile: An application profile models application requirements and contains one or more End Point Groups (EPGs) as necessary to provide the application capabilities. Depending on the application and connectivity requirements, VersaStack with ACI design uses multiple application profiles to define multi-tier applications as well as to establish storage connectivity.

- Bridge Domain: A bridge domain represents a L2 forwarding construct within the fabric. One or more EPG can be associated with one bridge domain or subnet. In ACI, a bridge domain represents the broadcast domain and the bridge domain might not allow flooding and ARP broadcast depending on the configuration.

The bridge domain has a global scope, while VLANs do not. Each endpoint group (EPG) is mapped to a bridge domain. A bridge domain can have one or more subnets associated with it and one or more bridge domains together form a tenant network.

- End Point Group: An End Point Group (EPG) is a collection of physical and/or virtual end points that require common services and policies. An EPG example is a set of servers or VMs on a common VLAN segment providing a common function or service. While the scope of an EPG definition is much wider, in the simplest terms an EPG can be defined on a per VLAN basis where all the servers or VMs on a common LAN segment become part of the same EPG.

  In the VersaStack with ACI design, various application tiers, ESXi VMkernel ports for Management, iSCSI and vMotion, and interfaces on IBM storage devices are mapped to various EPGs.

- Contracts: Contracts define inbound and outbound traffic filter, QoS rules and Layer 4 to Layer 7 redirect policies. Contracts define the way an EPG can communicate with another EPG(s) depending on the application requirements. Contracts are defined using provider-consumer relationships; one EPG provides a contract and another EPG(s) consumes that contract. Contracts utilize filters to limit the traffic between the applications to certain ports and protocols.

  Cisco ACI Tenant Model overview and relationship between the constructs is show in Figure 29.

Figure 29 ACI Tenant Model Overview

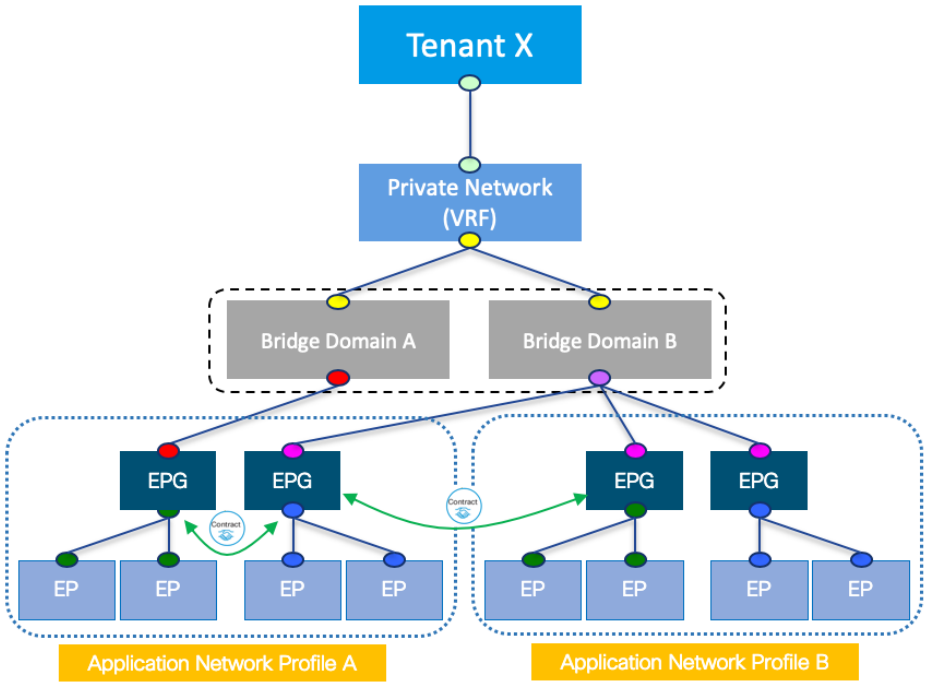
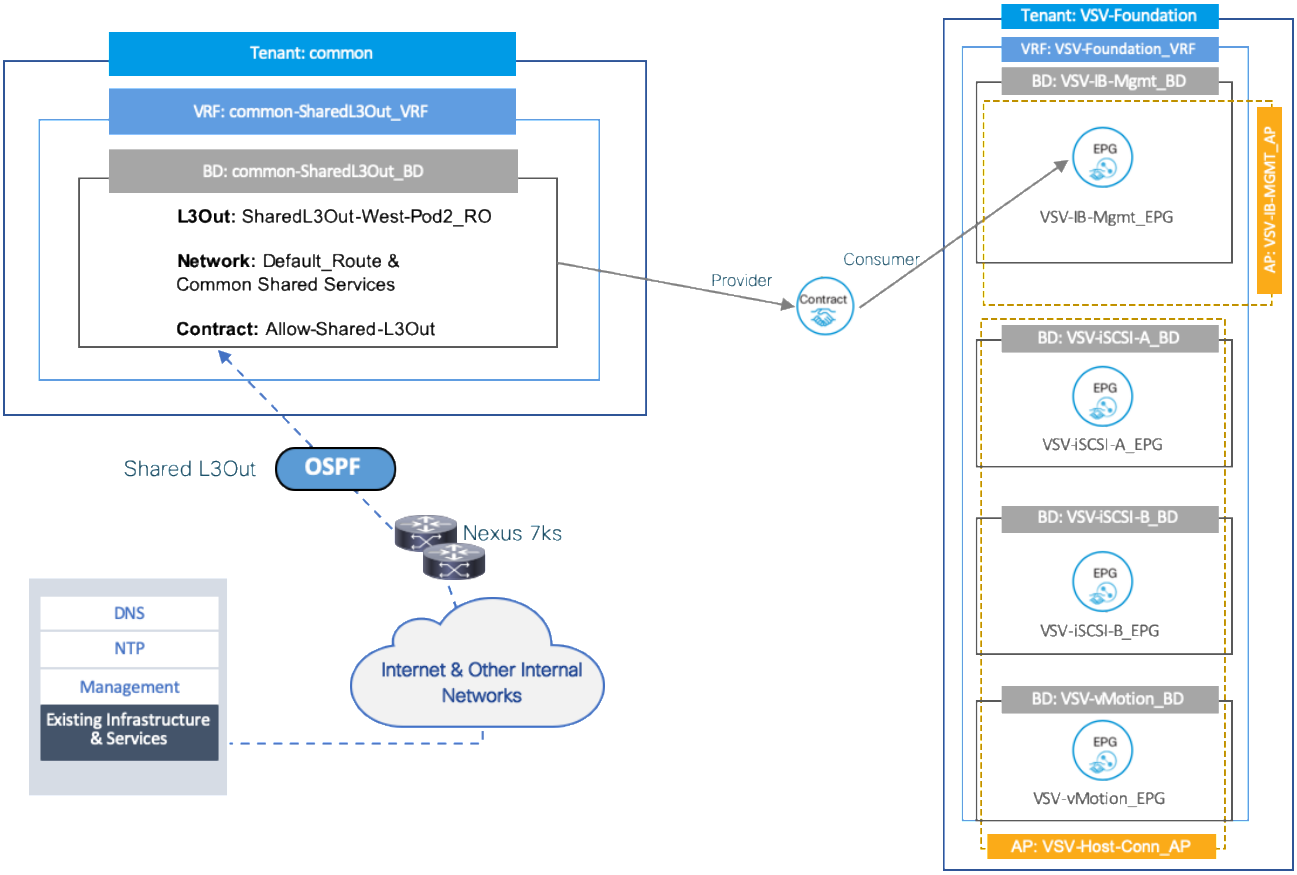


Figure 30 illustrates the high-level relationship between various ACI Tenant elements as deployed in the validated architecture by highlighting the Foundation tenant. As shown in the figure, a Tenant can contain one or more application profiles and an application profile can contain one or more EPGs. Devices in the same EPG can talk to each other without any special configuration. Devices in different EPGs can talk to each other using contracts and associated filters. A tenant can also contain one or more VRFs and bridge domains. Different application profiles and EPGs can utilize the same VRF or the bridge domain. The subnet can be defined within the EPG but is preferably defined at the bridge domain.

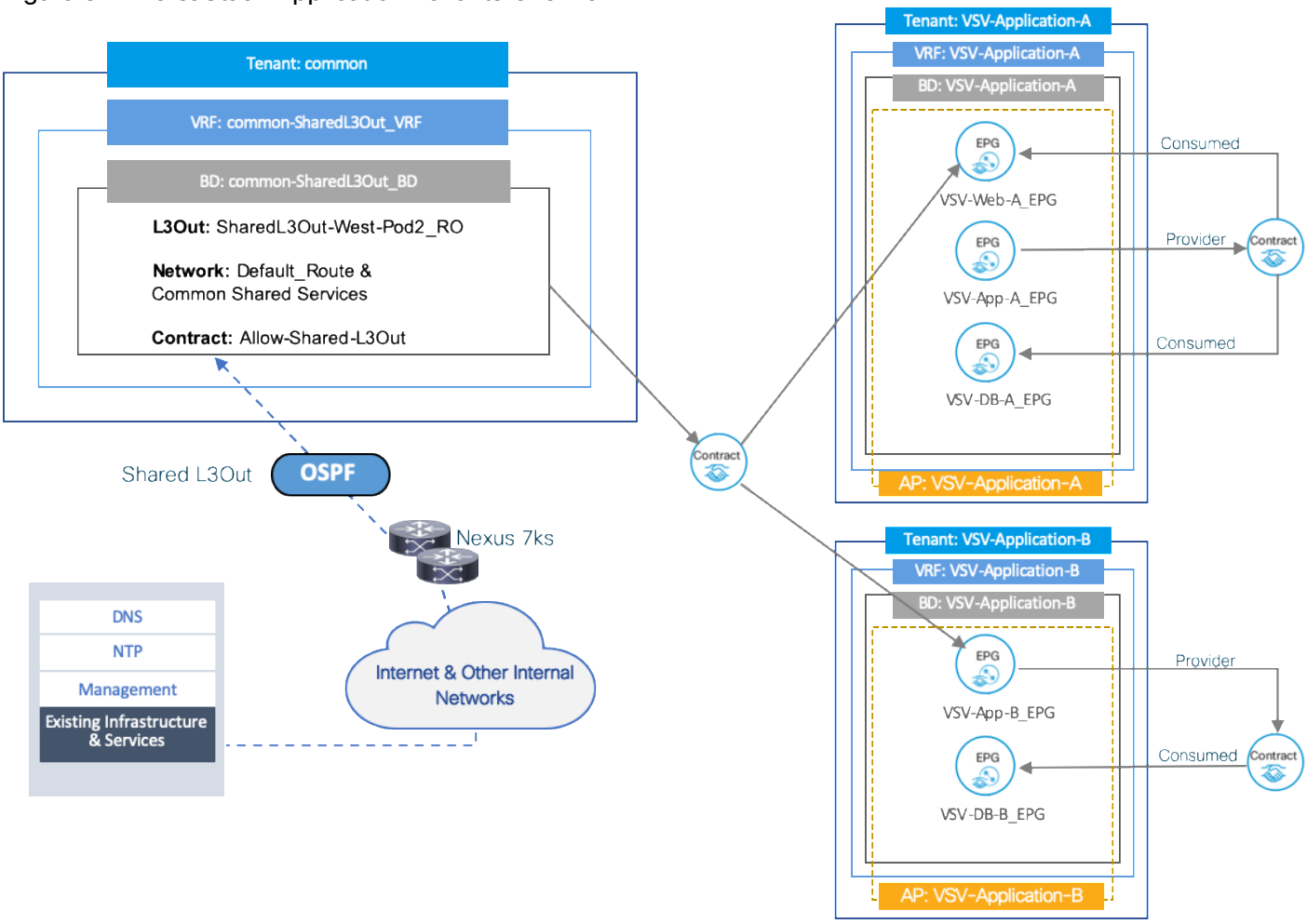
Figure 30 VersaStack Foundation Tenant Overview



Specifically, in the Foundation Tenant shown in Figure 30, are two Application Profiles which are acting as logical groupings of the EPGs within the Foundation Tenant. In the Foundation Tenant, each EPG has its own bridge domain with the subnet used by each EPG specified within the bridge domain.

There are two Application Tenants configured with in the validated architecture as shown in Figure 31. The same relationships exist between the differing tenant elements as with the Foundation Tenant, but the Application Tenant was provisioned with all EPGs in the same Application Profile and all EPGs in the same Bridge Domain. In this design, the subnet was set within the bridge domain and shared by all EPGs, but the member endpoints in different EPGs do not have connectivity amongst each other without a contract in place though they are in the same subnet.

Figure 31 VersaStack Application Tenants Overview



The connectivity for the Application-A EPGs shown within the tenant breaks down to both Web and DB having connectivity to App, but not each other, and only Web having connectivity to outside networks.

The Application-B shown within the tenant represents an application with two EPGs and App having connectivity to DB and outside networks.

End Point Group (EPG) Mapping within ACI

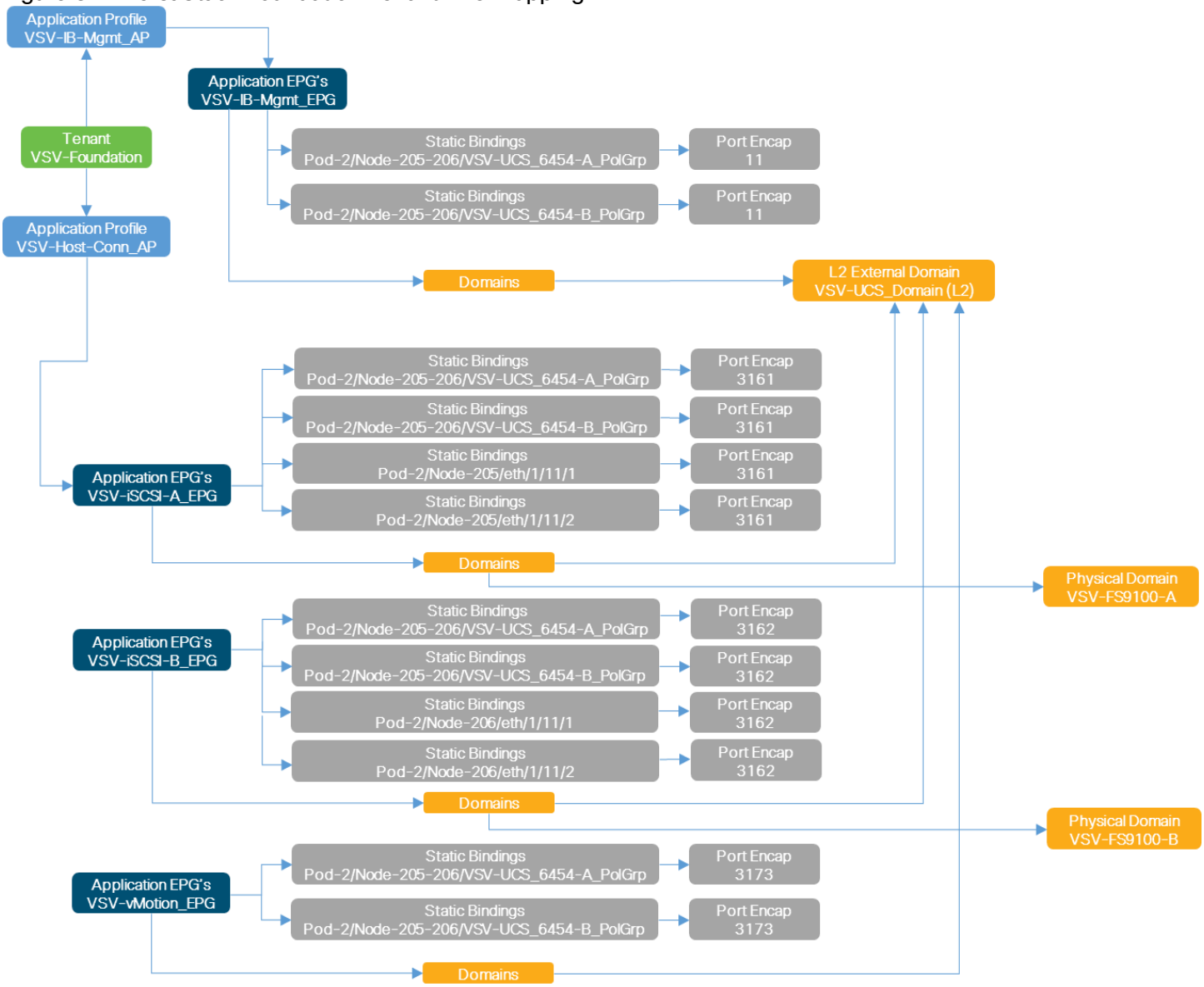
Once the Access policies and Tenant configurations are in place, the EPGs within the tenants need to be linked to the ACI networking domains (created as shown in Fabric Access Policies section above), hence, making the link between the logical object representing workload (the EPG) and the physical or virtual switches where the endpoints generating the workload reside.

In ACI, endpoint traffic is associated with an EPG in one of the following ways:

- Statically mapping a Path/VLAN to an EPG
- Associating an EPG with a Virtual Machine Manager (VMM) domain thereby allocating a VLAN dynamically from a pre-defined pool in APIC

VersaStack Foundation Tenant EPG's mapping to UCS physical domain uses static mapping. The mapping and VLAN assignment as configured within the validated architecture is depicted in Figure 32.

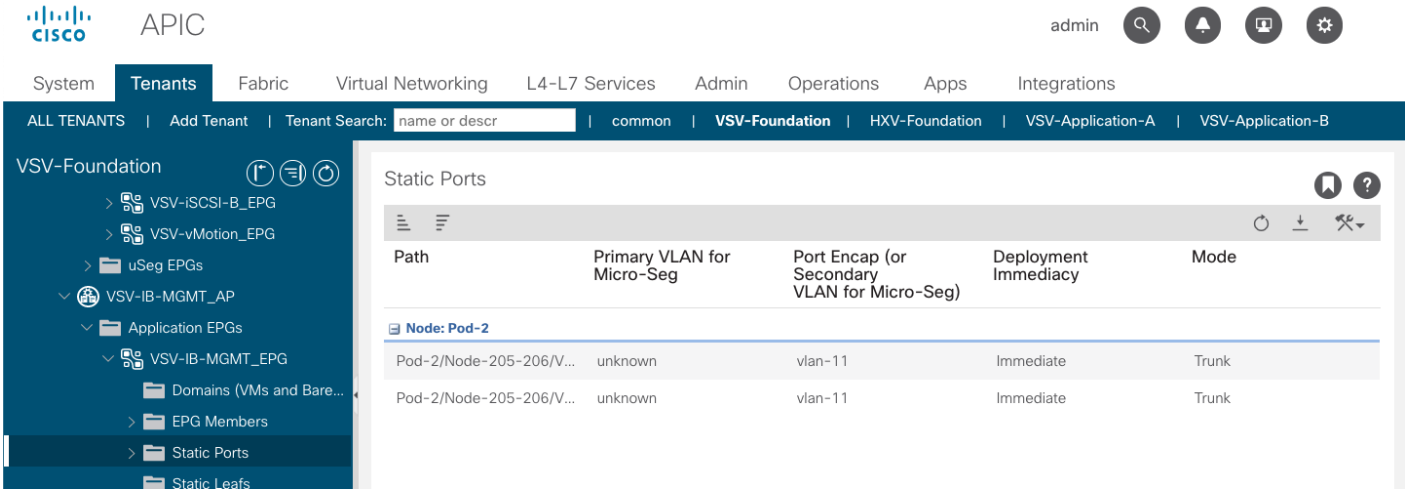
Figure 32 VersaStack Foundation Tenant EPG Mapping



Statically mapping of Path/VLAN to an EPG is useful for:

- Mapping bare metal servers to an EPG
- Mapping vMotion VLANs on the Cisco UCS/ESXi Hosts to an EPG
- Mapping iSCSI VLANs on both the Cisco UCS and the IBM storage systems to appropriate EPGs
- Mapping the management VLAN(s) from the existing infrastructure to an EPG

Figure 33 Example – EPG Static Path Binding

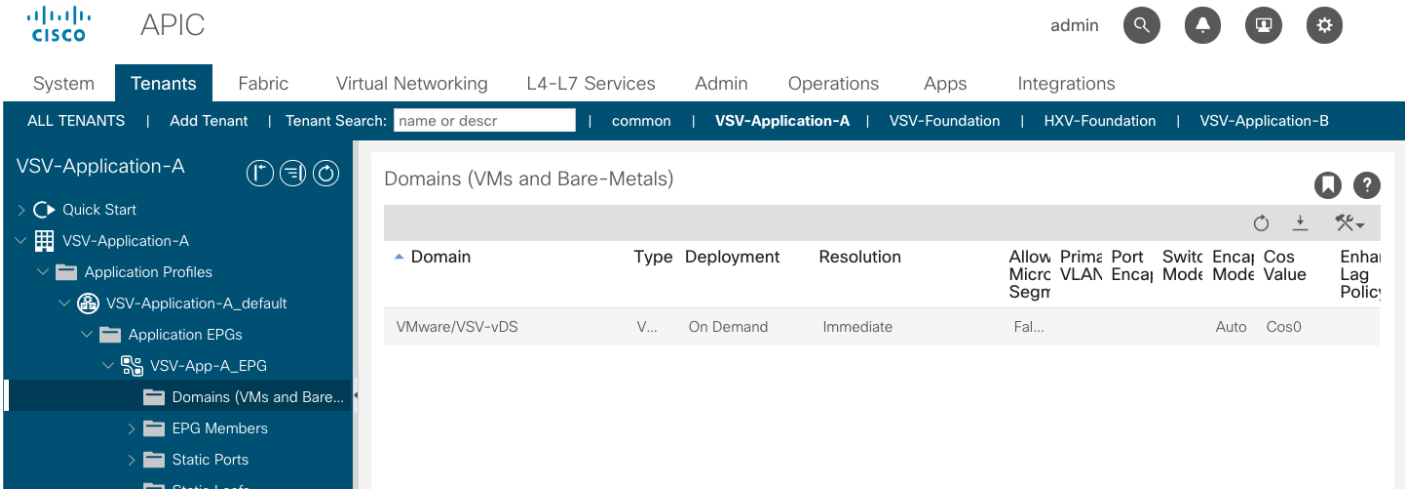


For Application tenants, the EPGs are linked to the VMM domain and dynamically leverages the physical paths to UCS domain and VLAN assignments.

Dynamically mapping a VLAN to an EPG by defining a VMM domain is useful for:

- Deploying VMs in a multi-tier Application requiring one or more EPGs
- Potentially deploying application specific IP based storage access within the application tenant environment

Figure 34 Example – EPG Assigned to Virtual Machine Manager



VLAN Design

To enable connectivity between compute and storage layers of the VersaStack and to provide in-band management access to both physical and virtual devices, several VLANs are configured and enabled on various paths. The VLANs configured in VersaStack design include:

- iSCSI VLANs to provide access to iSCSI datastores including boot LUNs
- Management and vMotion VLANs used by compute and vSphere environment

- A pool of VLANs associated with ACI Virtual Machine Manager (VMM) domain. VLANs from this pool are dynamically allocated by APIC to application end point groups

These VLAN configurations are explained in the following sections.

VLANs in Cisco ACI

VLANs in an ACI Fabric do not have the same meaning as VLANs in a regular switched infrastructure. The VLAN tag for a VLAN in ACI is used purely for classification purposes. In ACI, data traffic is mapped to a bridge domain that has a global scope therefore local VLANs on two ports might differ even if they belong to the same broadcast domain. Rather than using forwarding constructs such as addressing or VLANs to apply connectivity and policy, ACI utilizes End Point Groups (EPGs) to establish communication between application endpoints.

Table 3 lists various VLANs configured for setting up the VersaStack UCS environment.

Table 3 EPG VLANs to Cisco UCS Compute Domain

vPC to Cisco UCS Fabric Interconnects	VLAN Name and ID	VLAN ID Name Usage
Domain Name: VSV-UCS_Domain Domain Type: External Bridged (L2) Domain VLAN Scope: Port-Local Allocation Type: Static VLAN Pool Name: VSV-UCS_Domain_vlans	Native VLAN (2)	VLAN 2 used as Native VLAN instead of default VLAN (1)
	IB-MGMT-VLAN (11)	Management VLAN to access and manage the servers
	vMotion (3173)	VMware vMotion traffic
	iSCSI-A (3161)	iSCSI-A path for booting both UCS B-Series and C-Series servers and datastore access
	iSCSI-B (3162)	iSCSI-B path for booting both UCS B-Series and C-Series servers and datastore access

iSCSI VLAN Configuration

To provide redundant iSCSI paths, two VMkernel interfaces are configured to use dedicated NICs for host to storage connectivity. In this configuration, each VMkernel port provided a different path that the iSCSI storage stack and its storage-aware multi-pathing plug-ins can use.

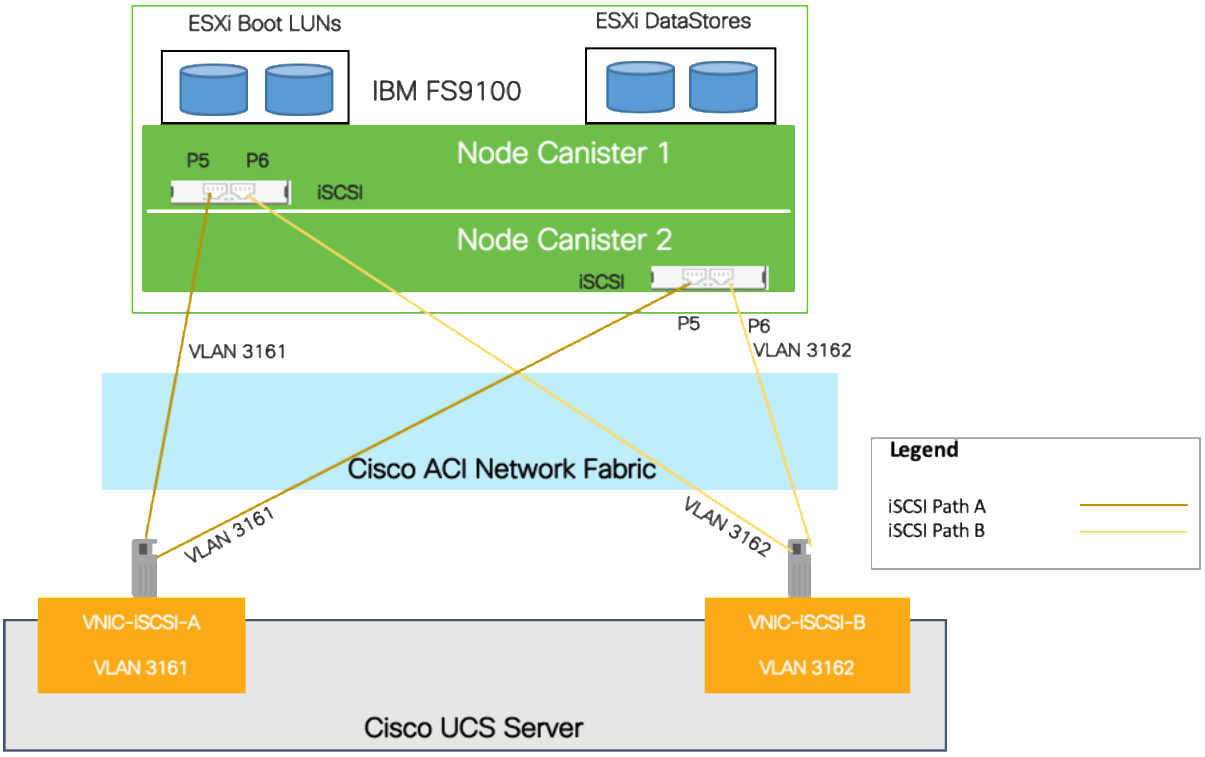
To setup iSCSI-A path between the ESXi hosts and the IBM FS9100 node canisters, VLAN 3161 is configured on the Cisco UCS, Cisco ACI and on the IBM FS9100 interfaces. To setup iSCSI-B path between the ESXi hosts and the IBM FS9100, VLAN 3162 is configured on the Cisco UCS, Cisco ACI and on the appropriate IBM FS9100 node interfaces. Within Cisco UCS service profile, these VLANs are enabled on vNIC-iSCSI-A and vNIC-iSCSI-B interfaces respectively. The iSCSI VLANs are set as native VLANs on the UCS server vNICs to enable boot from SAN functionality. Figure 35 shows the iSCSI connectivity between the UCS server and IBM storage array. The iSCSI traffic passes through the Cisco ACI fabric, and ACI policies gets applied to iSCSI traffic. Each ESXi host ends up with 2 active optimized paths, and 2 active un-optimized paths to its boot and datastore LUNs.

Table 4 EPG VLANs to IBM FS9100 Storage Nodes

vPC to Cisco UCS Fabric Interconnects	VLAN Name and ID	VLAN ID Name Usage

vPC to Cisco UCS Fabric Interconnects	VLAN Name and ID	VLAN ID Name Usage
Domain Name: VSV-FS9100-A VSV-FS9100-B Domain Type: Bare Metal (Physical) VLAN Scope: Port-Local Allocation Type: Static VLAN Pool Name: VSV-FS9100-A_vlans VSV-FS9100-B_vlans	iSCSI-A (3161)	Provides access to boot, application data and datastore LUNs on IBM FS9100 via iSCSI Path-A
	iSCSI-B (3162)	Provides access to boot, application data and datastore LUNs on IBM FS9100 via iSCSI Path-B

Figure 35 VLAN Mapping for iSCSI Storage Access



Virtual Machine Networking VLANs for VMware vDS

The access layer connectivity from ACI leaf to a VMM domain is through the vPCs to the UCS domain hosting the virtual environment. APIC integration with VMware vCenter enables EPGs to be deployed in the VMM domain. To communicate with a virtual endpoint in an EPG, VLANs are dynamically assigned based on VM events from VMware vCenter. As VMs come online, vCenter notifies the APIC and a VLAN is allocated from the pool. The EPG VLANs used in the VersaStack design for connectivity to virtual endpoints are listed in Table 5

Table 5 EPG VLANs to VMM Domain

vPC to Cisco UCS Fabric Interconnects	VLAN Name and ID	VLAN ID Name Usage

vPC to Cisco UCS Fabric Interconnects	VLAN Name and ID	VLAN ID Name Usage
Domain Name: VSV-vDS Domain Type: VMM Domain VLAN Scope: Port-Local Allocation Type: Dynamic VLAN Pool Name: VSV-Application	(1400-1499)	VLANs for Application EPGs hosted on Cisco UCS Servers. The physical connectivity to the EPG virtual endpoints is through the vPCs to Fabric Interconnects in the Cisco UCS domain. APIC to VMM integration is used to dynamically assign VLANs as new virtual endpoints come online.

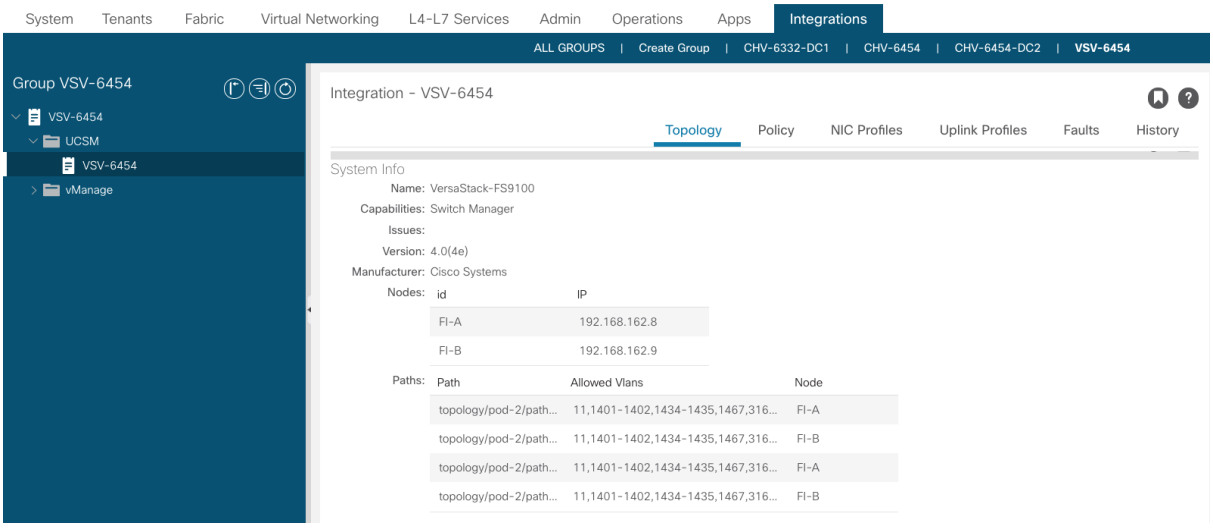
When using VMware vDS in the VersaStack ACI setup, a pool of 100 VLANs,1400-1499, were defined to be used on-demand by the VM Networking. VLANs from this pool are dynamically assigned to the EPGs mapped to the Virtual Machine Manager (VMM) domain.

Virtual Machine Manager (VMM) Domains

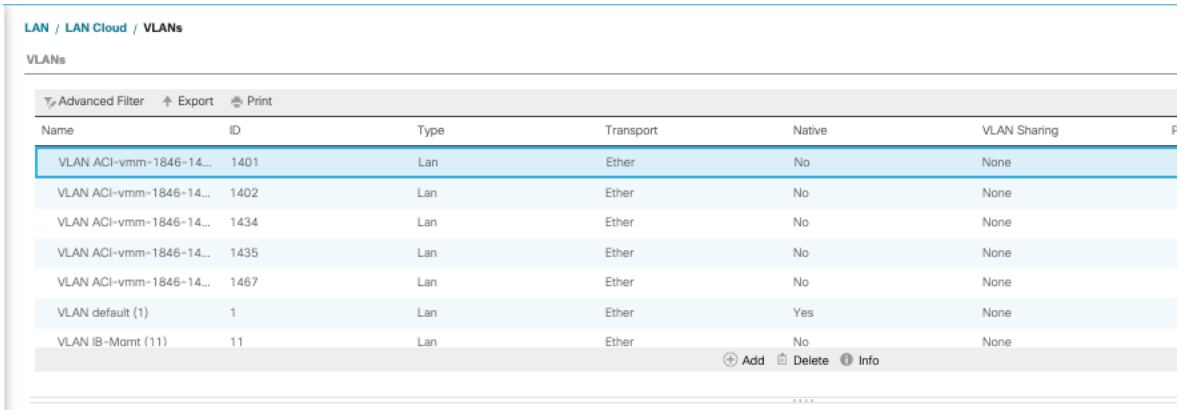
In a VMware vCenter environment, Cisco APIC controls the creation and configuration of the VMware vSphere Distributed Switch (VDS) or the Cisco Application Virtual Switch (AVS). Once the virtual distributed switches are deployed, APIC communicates with the switches to publish network policies that are applied to the virtual workloads including creation of port groups for VM association. A VMM domain contains multiple EPGs and hence multiple port groups. To position an application, the application administrator deploys the VMs using VMware vCenter and places the VMNIC into the port group defined for the appropriate application tier.

Cisco UCS Integration with ACI

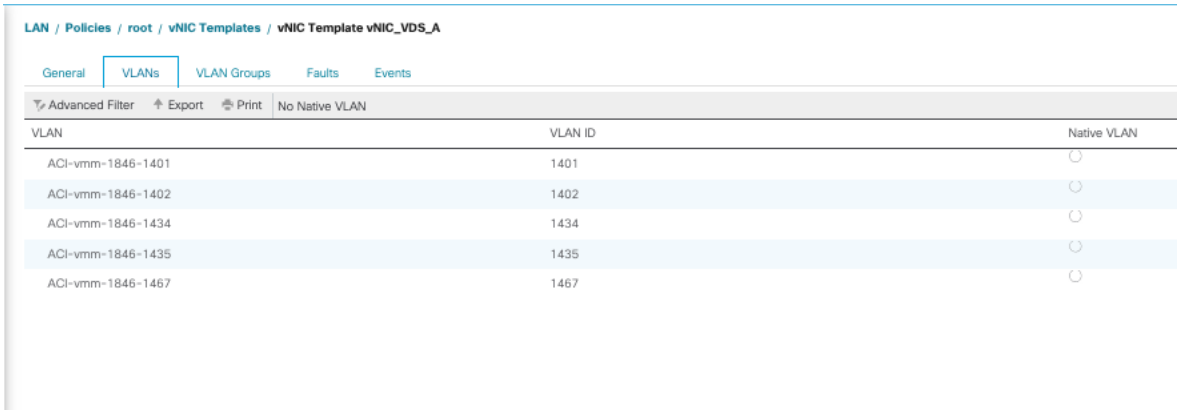
New to ACI 4.1 is an integration with Cisco UCS Manager (UCSM) to allow VMM synchronization of VLANs dynamically allocated to the EPGs to be configured within UCSM using an ACI app named ExternalSwitch.



With ExternalSwitch installed, the integration manager connection can be created allowing the allocated VLANs to be configured within Cisco UCS:



As well as the automatic insertion into the designated vNIC templates:

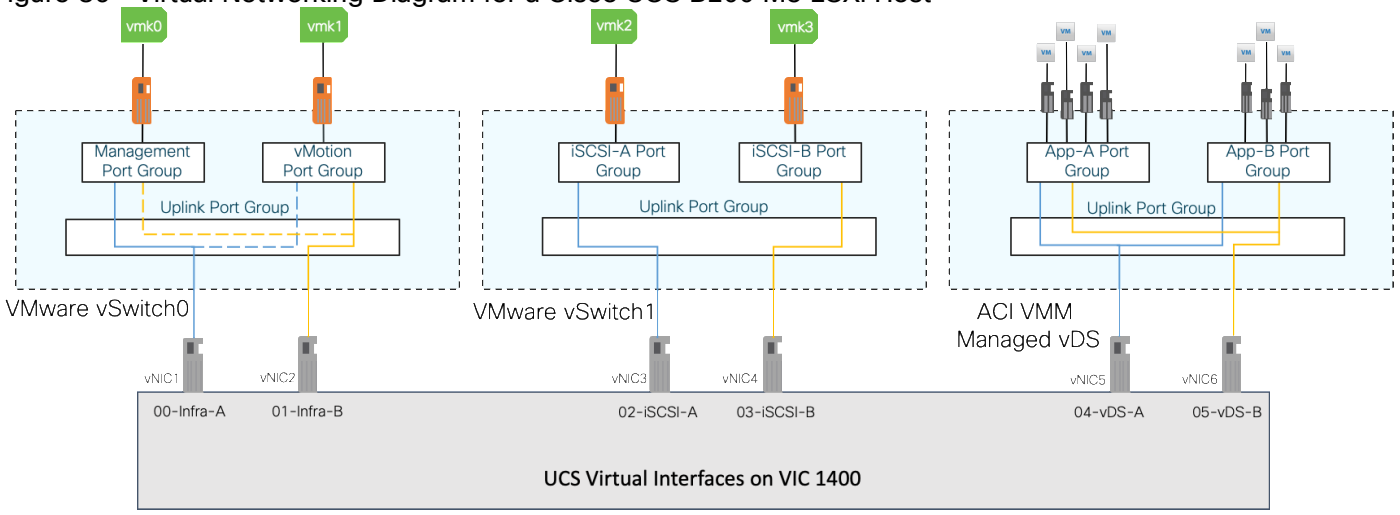


Virtual Switching Architecture

A tenant application deployment utilizes port groups on APIC controlled distributed switch (VDS). However, for some of the core connectivity such as out-of-band management access, vSphere vMotion and storage LUN access using iSCSI vSphere vSwitches are deployed. To support this multi-vSwitch requirement, multiple vNIC interfaces are setup in Cisco UCS services profile and storage, management and data VLANs are then enabled on the appropriate vNIC interfaces.

Figure 36 shows the distribution of VMkernel ports and VM port-groups on VMware ESXi server. For an ESXi server, supporting iSCSI-based storage access, In-band management and vMotion traffic is handled by a Foundation Services vSwitch and iSCSI-A and iSCSI-B traffic is handled by a dedicated iSCSI vSwitch. The VMkernel configuration of both management and vMotion are pinned in an active/standby configuration setting on opposing links, to keep these types of traffic contained within a particular Cisco UCS fabric interconnect when switching this traffic between ESXi hosts, thus preventing the need to send it up into the Nexus switch to pass between fabrics. The resulting ESXi host configuration therefore has a combination of 2 vSwitches and a single APIC-Controlled distributed switch which handles application (tenant) specific traffic.

Figure 36 Virtual Networking Diagram for a Cisco UCS B200 M5 ESXi Host



Onboarding Infrastructure Services

In an ACI fabric, all the applications, services and connectivity between various elements are defined within the confines of tenants, application profiles, bridge domains and EPGs as discussed earlier. The tenant configured to provide the infrastructure services is named VSV-Foundation. The VSV-Foundation tenant enables compute to storage connectivity for accessing iSCSI datastores, enables VMware vMotion traffic and provides ESXi hosts and VMs access to existing management infrastructure. The Foundation tenant comprises of three bridge domain's, one per EPG. Since there are no overlapping IP address space requirements, VSV-Foundation tenant consists of a single VRF called VSV-Foundation_VRF.

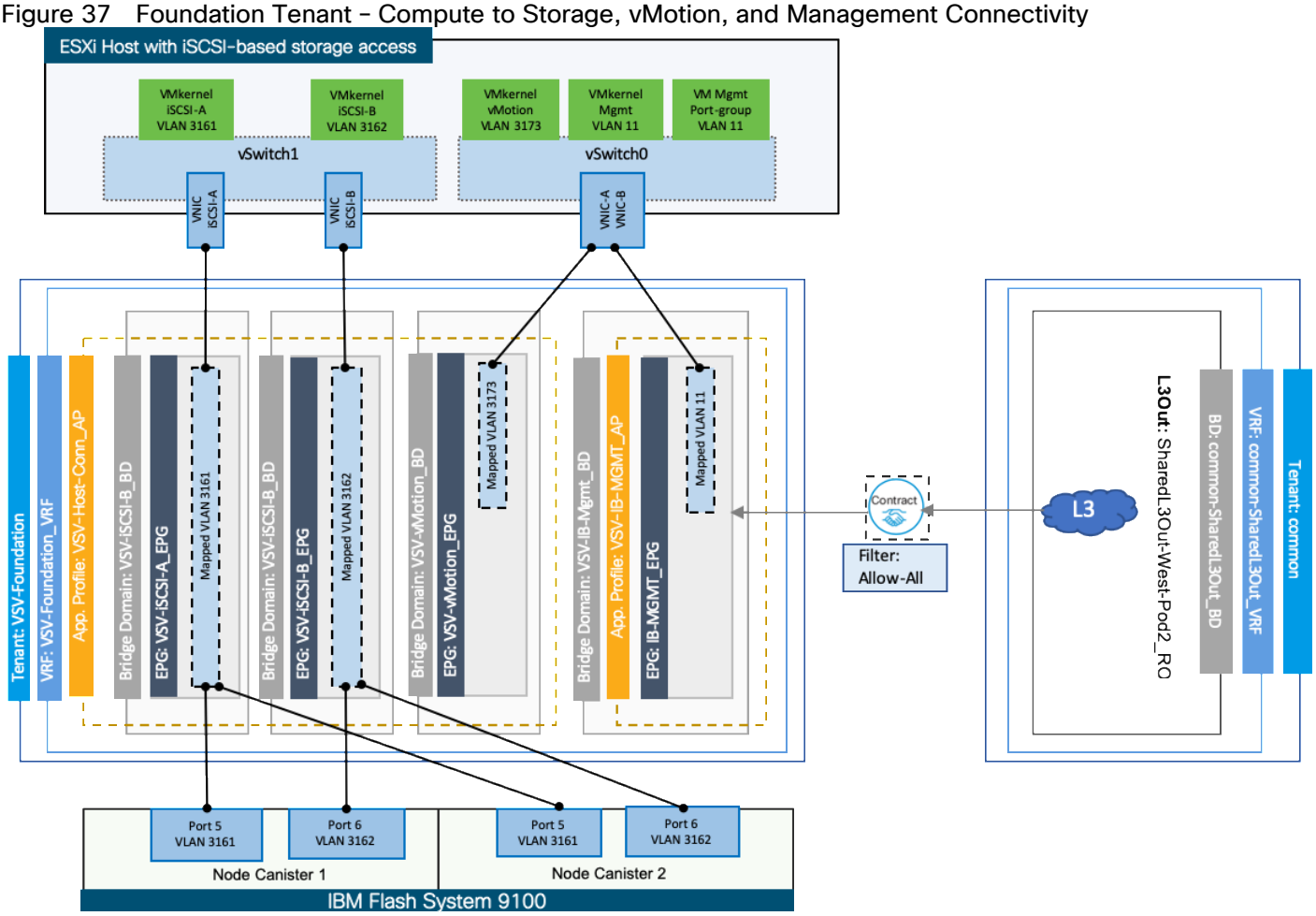
VSV-Foundation tenant is configured with two different Application Profiles:

- VSV-Host-Conn_AP: This application profile contains EPGs to support compute to storage connectivity as well as VMware vMotion traffic. The three EPGs defined under this application profile are: VSV-iSCSI-A_EPG, VSV-iSCSI-B_EPG and VSV-vMotion_EPG.
- VSV-IB-Mgmt_AP: This application profile provides ESXi host and VMs connectivity to existing management segment through the Common tenant via L3Out network.

Foundation Tenant EPG Design and Connectivity for iSCSI-based Storage

Figure 37 shows an overview of ACI design covering ESXi host and IBM storage connectivity details and the relationship between various ACI elements for the iSCSI-based storage access.

Figure 37 Foundation Tenant – Compute to Storage, vMotion, and Management Connectivity



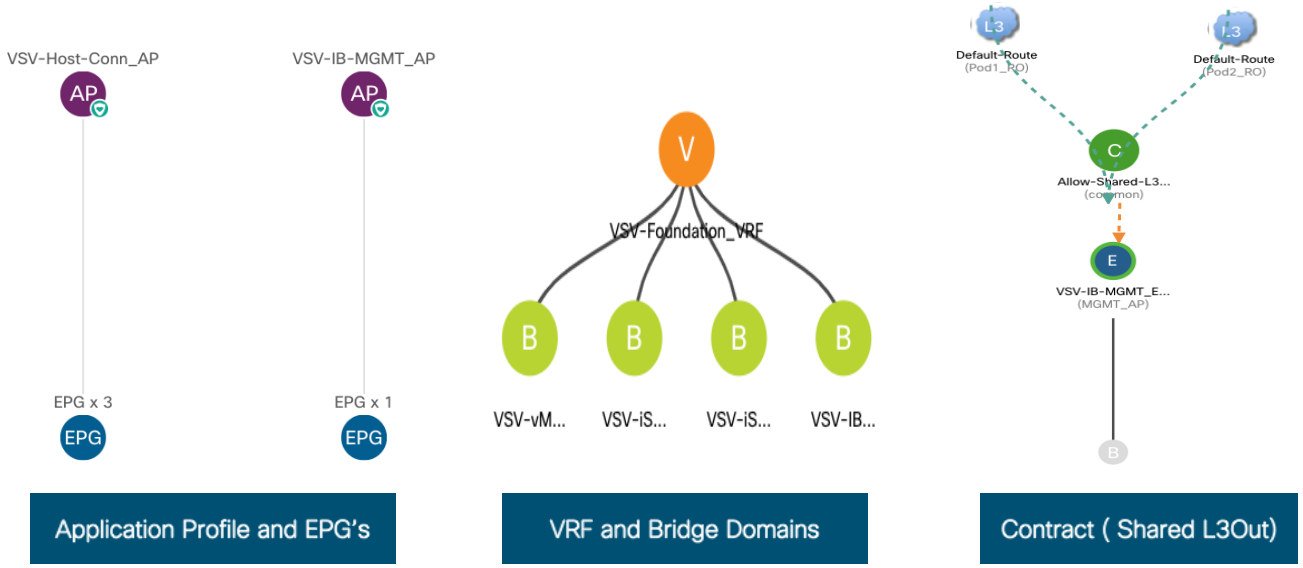
The following ACI constructs are defined in the *VSV-Foundation* Tenant configuration for the iSCSI-based storage access:

- Tenant: VSV-Foundation
- VRF: VSV-Foundation_VRF
- Application Profile VSV-Host-Conn-AP consist of three EPGs:
  - VSV-iSCSI-A_EPG statically maps the VLANs associated with iSCSI-A interfaces on the IBM storage controllers and Cisco UCS Fabric Interconnects (VLAN 3161)
    - Bridge Domain: VSV-iSCSI-A_BD
  - VSV-iSCSI-B_EPG statically maps the VLANs associated with iSCSI-B interfaces on the IBM storage controllers and Cisco UCS Fabric Interconnects (VLAN 3162)
    - Bridge Domain: VSV-iSCSI-B_BD
  - VSV-vMotion_EPG statically maps vMotion VLAN (3173) on the Cisco UCS Fabric Interconnects
    - Bridge Domain: VSV-vMotion_BD
- Application Profile VSV-IB-MGMT_AP consist of one EPG:

- VSV-IB-MGMT_EPG statically maps the management VLAN (11) on the Cisco UCS Fabric Interconnects. This EPG is configured to provide VMs and ESXi hosts access to the existing management network via Shared L3Out connectivity. As shown in Figure 37, this EPG utilizes the bridge domain VSV-IB-Mgmt_BD.

ACI constructs created after the deployment of VersaStack Foundation Tenant are in Figure 38:

Figure 38 VersaStack Design – ACI Constructs created for Foundation Tenant



Onboarding Multi-Tier Application

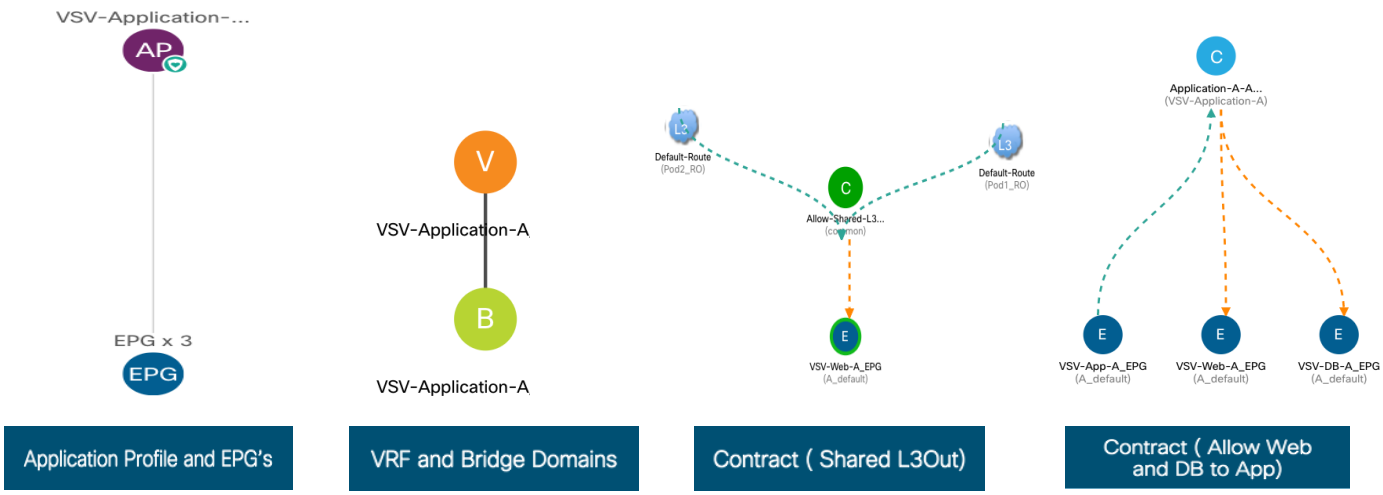
As previously mentioned, the ACI constructs for onboarding a multi-tier application include defining a new tenant, VRF(s), bridge domain(s), application profile(s), end point group(s), and the contract(s) to allow communication between various tiers of the application.

To deploy a sample three-tier application, following elements were configured:

- A new Tenant called VSV-Application-A is defined to host the application
- A VRF called VSV-Application-A_VRF is defined under the tenant to provide the tenant IP address space
- A bridge domain Application-A_BD is associated with the tenant
- An application profile, VSV-Application_AP is utilized to deploy the application.
- Three EPGs, VSV-App-A-Web, VSV-App-A-App and VSV-App-A_DB are associated with the VMM domain to host Web, APP and DB tiers of the application
- A contract to allow communication between the application tiers is defined. This contract is "provided" by the EPG VSV-App-A-App and "consumed" by the EPG's VSV-App-A and VSV-Web-A.

ACI constructs created after the deployment of a VersaStack 3-Tier Application Tenant are shown in Figure 39:

Figure 39 VersaStack Design – ACI Constructs Created for 3-Tier Application



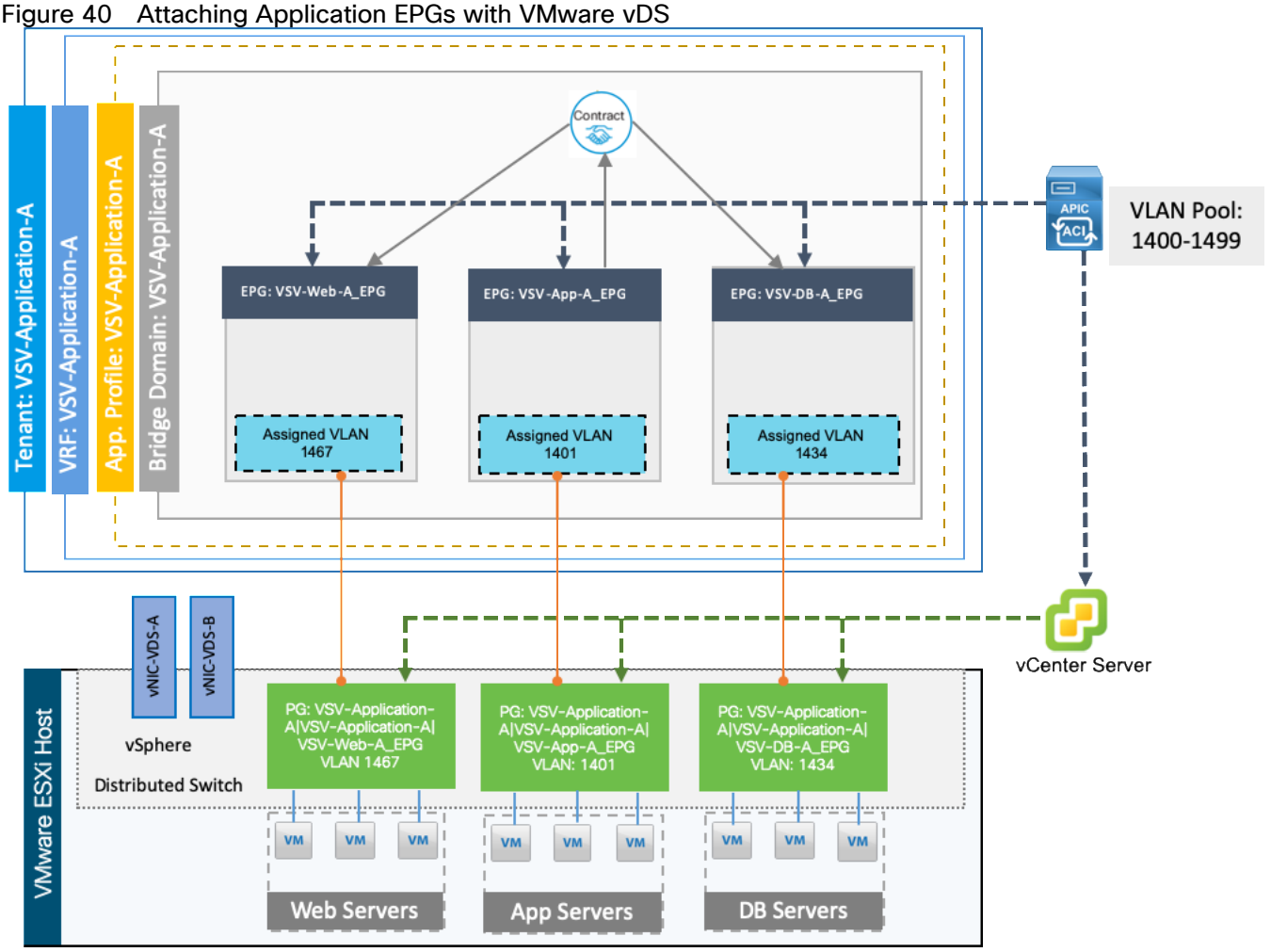
The following sections explain the deployment details for VMware vDS.

Port Group creation for VMware vDS

Cisco APIC integrates with the VMware vCenter instances to transparently extend the Cisco ACI policy framework to VMware vSphere workloads. When application EPGs are attached to a VMware vDS based VMM domain, Cisco APIC assigns VLANs from a pre-defined pool and uses its connection to the VMware vCenter to create new port groups on the VMware vDS. These port groups are used to deploy application VMs in the appropriate application tier. The port group name is determined using following format: "Tenant_Name | Application Profile_Name | EPG_Name".

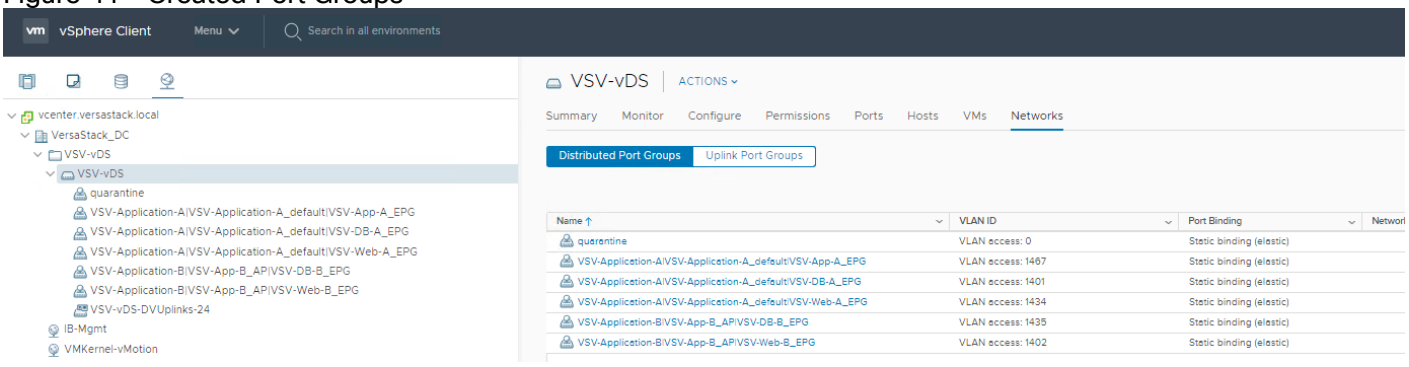
For example, as shown in Figure 40, when a new EPG VSV-App-A-Web is defined under application profile VSV-Application-A (that belongs to tenant VSV-Application-A), a VLAN from the dynamic VLAN pool (1467 in this example) gets assigned to this EPG and a new port group named VSV-Application-A|VSV-Application-A|VSV-Web-A_EPG is automatically created on the VMware vDS. When a virtualization administrator assigns a VM NIC to this port group, all the network policies including security (contracts), L4-L7 and QoS automatically get applied to the VM communication.

Figure 40 Attaching Application EPGs with VMware vDS



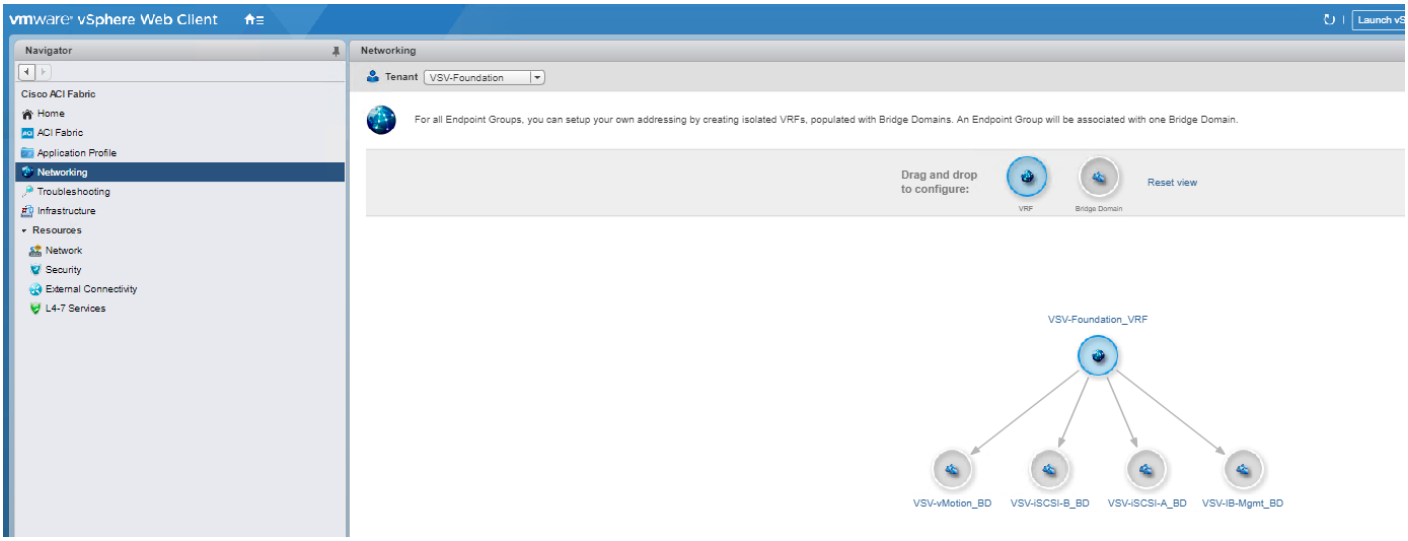
The port groups created on the VersaStack vDS switch is shown in Figure 41:

Figure 41 Created Port Groups



ACI Plug-in for vCenter

The ACI Plugin for vCenter allows a subset of the commands available through the APIC GUI to be invoked directly from the vCenter vSphere Web Client.



This subset of commands from the APIC GUI includes tenant creation and components, including:

- Application Profiles
- VRFs
- Bridge Domains
- EPGs
- Contracts

These commands provide the vSphere administrator basic abilities for managing and creating the tenant construct as it interfaces with the upstream ACI network.

ACI VMware plugin is only supported with vSphere Flash based Web Client.

External Network Connectivity – Shared Layer 3 Out

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 a pre-defined system tenant where any objects defined in this tenant are visible to all other tenants, making it easier to position common services in which many tenants will need access.

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.

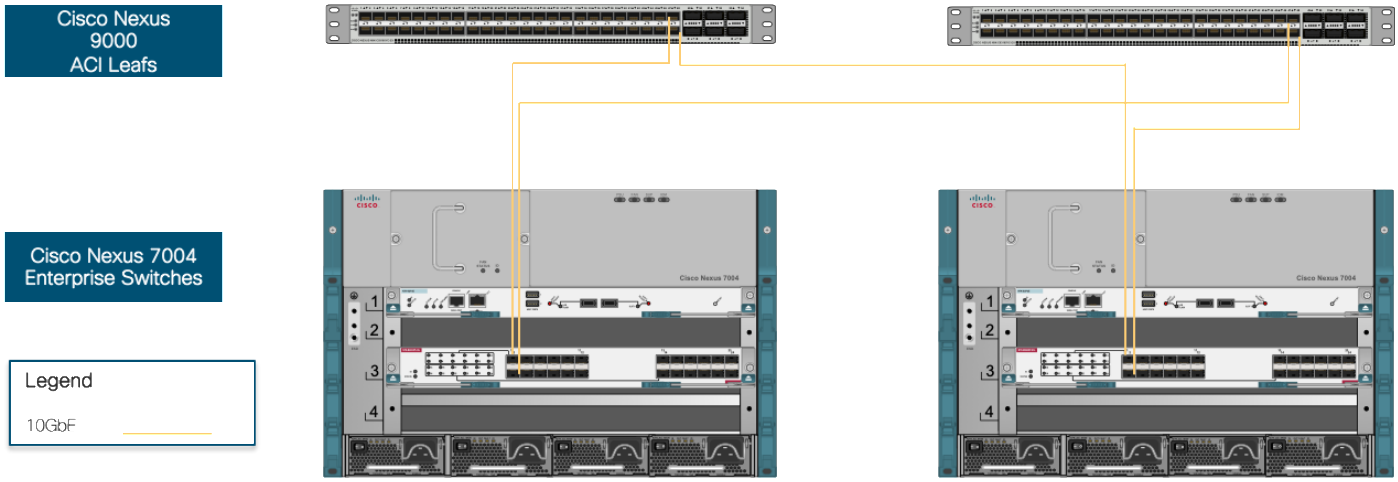
Shared L3Out Design

To enable a shared L3Out connection, border leaf nodes in the ACI fabric are connected to Layer 3 gateways in the outside network. To connect the 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 are shown in Figure 43, along with the ACI configuration to enable IP connectivity and routing.

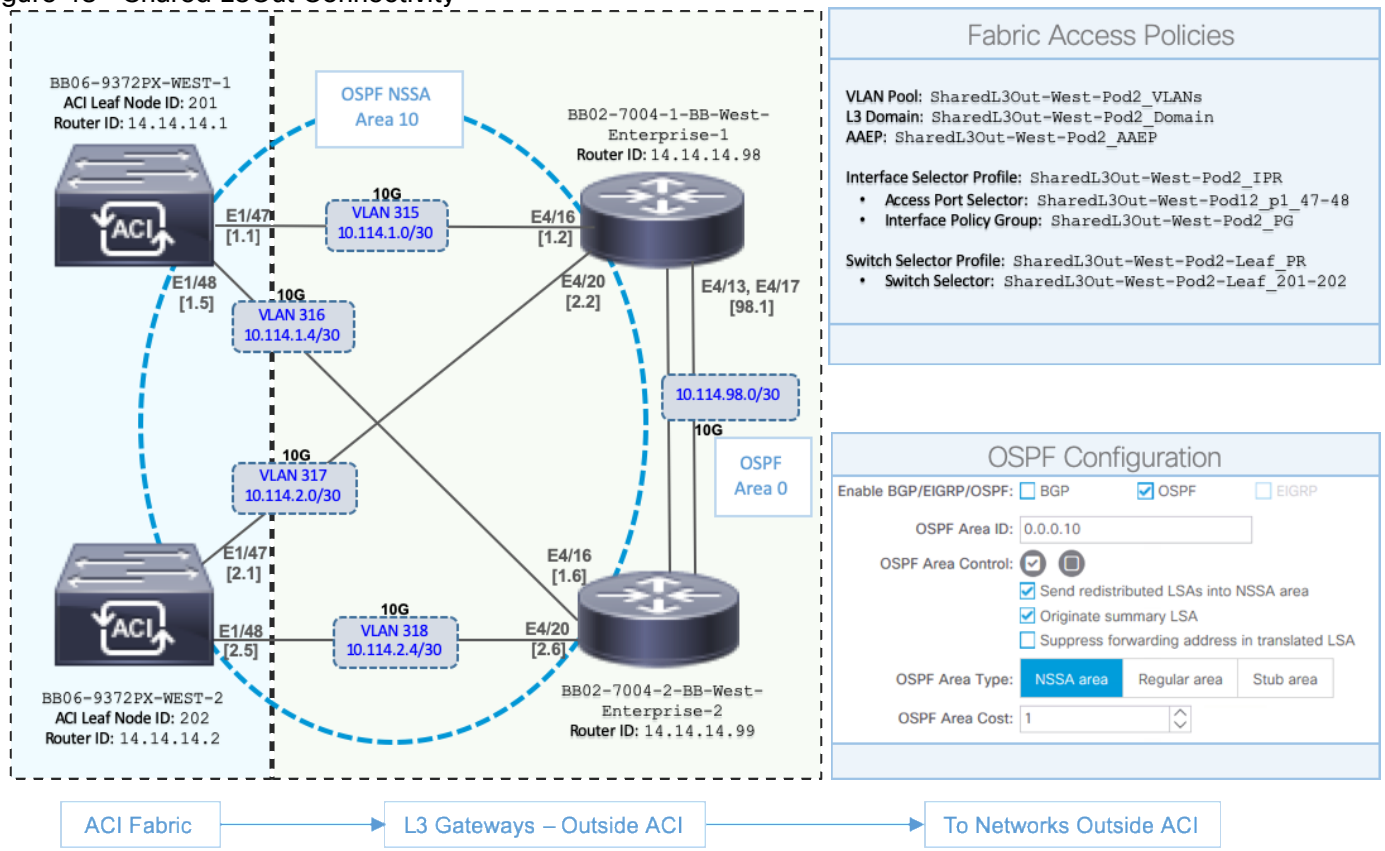
Each border leaf switch is redundantly connected to the Nexus 7000 switches using 10GbE links. The four links between ACI leaf nodes and external routers are individual connections with a dedicated VLAN and IP subnet for each link – no link bundling is used. The border leaf switches in this design also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

Figure 42 ACI – Physical Connectivity to Existing Infrastructure



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.

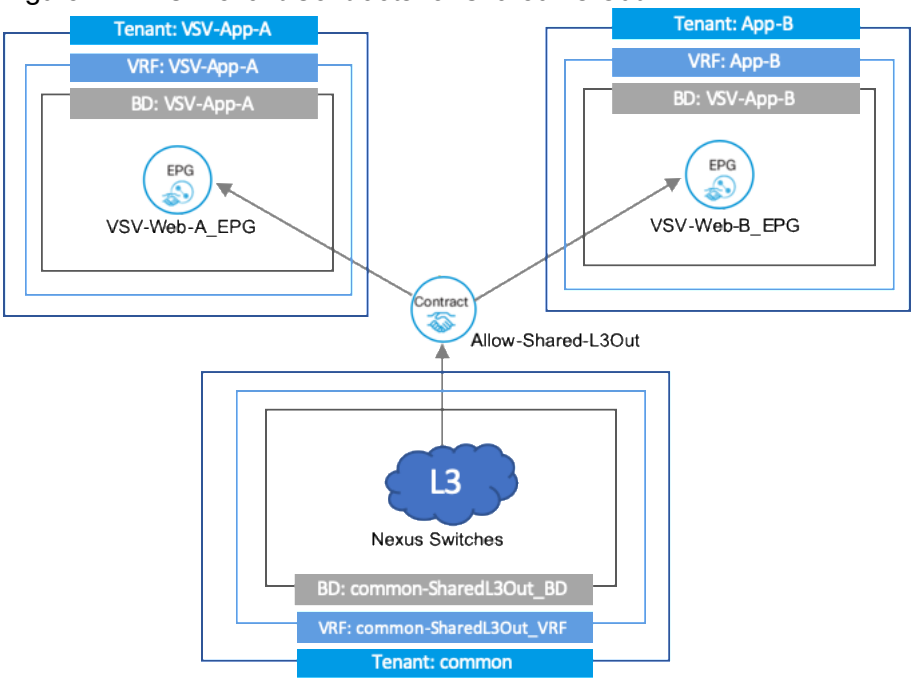
Figure 43 Shared L3Out Connectivity



The ACI constructs and design for enabling and accessing a shared L3Out service is shown in Figure 44. 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 is defined under the common tenant, which is setup with OSPF to provide connectivity to external infrastructure.
- 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 tenants 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.

Figure 44 ACI Tenant Contracts for Shared L3 Out



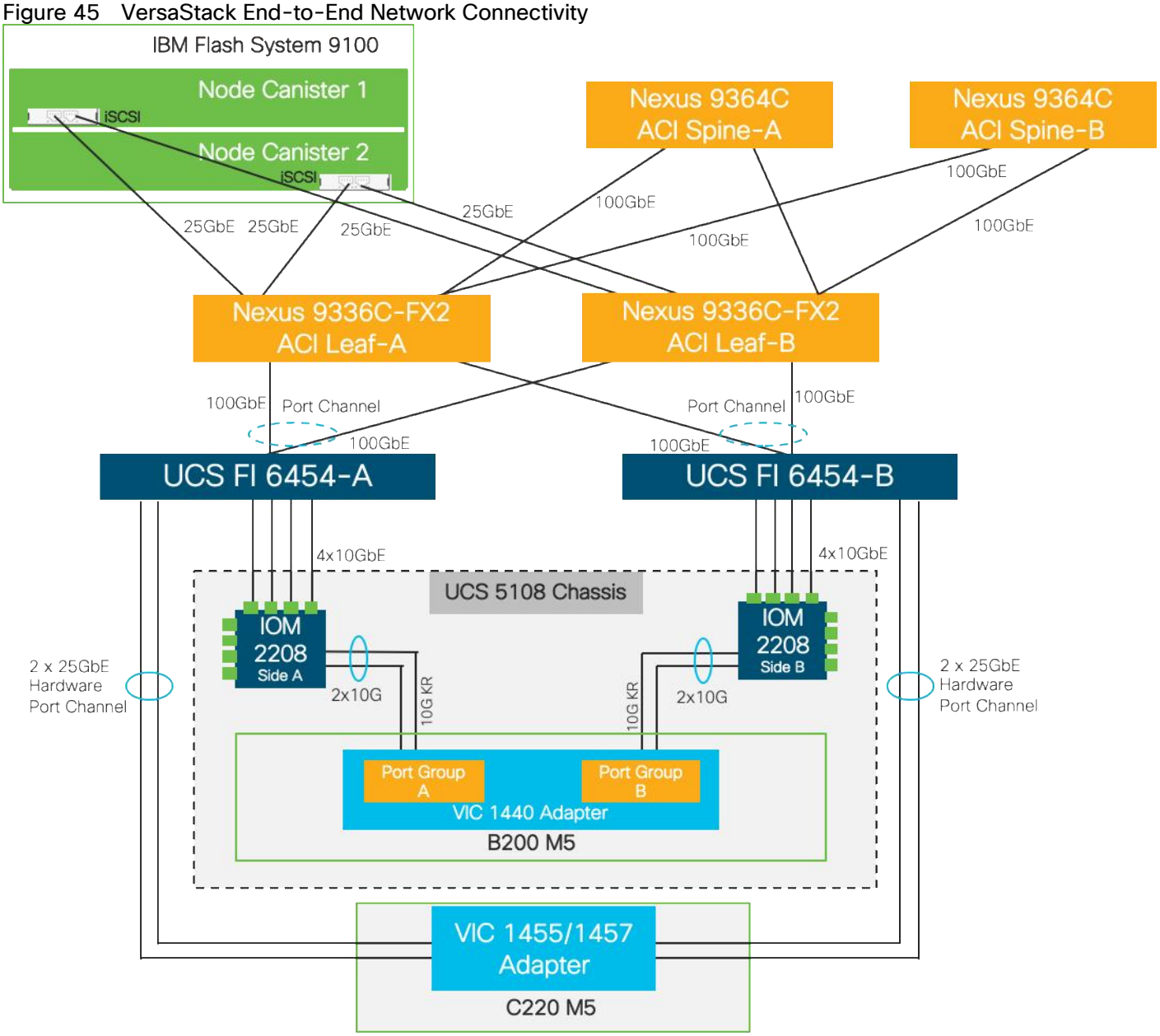
By defining a shared L3Out in common tenant, the contract is provisioned as part of the L3Out configuration and it would automatically be available in all other tenants to consume, without doing any additional configuration since the objects (contracts in this example) from the common tenant are available 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.

VersaStack End-to-End Core Network Connectivity

The Cisco Nexus 9336C-FX2 used in this design supports flexible port speeds and provides 25Gbps connectivity to the IBM FS9100 storage controllers for iSCSI traffic and 100 Gbps connectivity to the Cisco UCS FIs for iSCSI storage and application traffic.

- The Cisco UCS M5 Servers are equipped with a VIC 1400 Series adapter
  - In the Cisco B200 M5 server, a VIC 1440 provides 2x10 Gbps to IOM A and 2x10Gbps to IOM B via the Cisco UCS Chassis 5108 chassis backplane
  - In the Cisco C220 M5 server, a VIC 1457 is used with 2x25 Gbps connections port-channeled to FI-A and 2x25 Gbps connections port-channel to FI-B
- Each IOM is connected to its respective Cisco UCS 6454 Fabric Interconnect using a port-channel for 4-8 links
- Each Cisco UCS 6454 FI connects to the Nexus 9336C-FX2 leaf switches through 2x100Gbps virtual port channels
- The IBM FS9100 is connected to both Nexus 9336C-FX2 leaf switches using QSFP100G-4SFP25G cable with 4x25 Gbps connections to provide redundant paths

Figure 45 VersaStack End-to-End Network Connectivity



Design Considerations

VersaStack designs incorporate connectivity and configuration best practices at every layer of the stack to provide a highly available best performing integrated system. VersaStack is a modular architecture that allows customers to adjust the individual components of the system to meet their particular scale or performance requirements. This section describes some of the design considerations for the current design and a few additional design selection options available to the customers.

VersaStack Scalability Considerations

VersaStack can be scaled-up or scaled-out depending on the needs of your business, applications, workloads, and users.

Cisco UCS

Cisco UCS scales linearly, with no performance degradation, the limits of scaling are found only in the port count of the Fabric Interconnects. UCS environment in VersaStack with a pair of Cisco UCS 6454 Fabric Interconnects can scale up to 20 Chassis and up to 160 blades in a single Cisco UCS management domain. Cisco UCS can scale to multiple Cisco UCS Domains with Cisco UCS Central or Cisco Intersight within and across data centers globally.

Cisco Network Fabric

ACI delivers software flexibility with the scalability of hardware performance that provides a robust transport network for today's dynamic workloads. Application Centric Infrastructure (ACI) in the data center is a holistic architecture with centralized automation and policy-driven application profiles.

Policy re-use enables the existing ACI configuration to be leveraged for adding additional compute clusters with minimal configuration as the Foundation tenant and other ACI constructs defined for the first cluster can be re-used for expansion.

To find the for maximum verified scalability limits for Cisco Application Centric Infrastructure (Cisco ACI) parameters, see: https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/4-x/verified-scalability/Cisco-ACI-Verified-Scalability-Guide-412.html

IBM FS9100 Storage

IBM FlashSystem 9100 system has a scalable architecture that enables flash capacity to be added (scaled up) to support multiple applications. The virtualized system can also be expanded (scaled-out) to support higher IOPS and bandwidth, or the solution can be simultaneously scaled up and out to improve capacity, IOPS, and bandwidth while maintaining MicroLatency.

A single IBM FlashSystem 9100 storage system consists of one control enclosure with internal storage, representing a 2U building block. The FlashSystem 9100 control enclosure can support multiple attached expansion enclosures. Expansion enclosures can be dynamically added with virtually no downtime.

For balanced increase of performance and scale, up to four IBM FlashSystem 9100 control enclosures can be clustered into a single storage system, multiplying performance and capacity with each addition. Clustering FlashSystem 9100 will scale the performance with additional NVMe storage. With four-way system clustering, the size of the system can be increased to a maximum of 3,040 drives.

Jumbo Frames

Enabling jumbo frames in a VersaStack environment optimizes throughput between devices by enabling larger size frames on the wire while reducing the CPU resources to process these frames. VersaStack supports wide variety of traffic types (vMotion, NFS, iSCSI, control traffic, and so on) that can benefit from a larger frame size. In this validation effort the VersaStack was configured to support jumbo frames with an MTU size of 9000. In VMware vSphere, the jumbo frames are configured by setting MTU sizes at both vSwitches and VMkernel ports. On IBM storage systems, the interface MTUs are modified to enable the jumbo frame.

Traditional switching fabrics typically us a 1500 bytes MTU and must be configured to support Jumbo frames. However, the ACI fabric, by default uses an MTU of 9366 bytes on core facing ports of leaf and spine switches and 9000 bytes on access ports of leaf switches. Therefore, no configuration is necessary to support Jumbo frames on an ACI fabric.

When setting the Jumbo frames, it is important to make sure MTU settings are applied uniformly across the stack to prevent fragmentation and the negative performance.

Cisco Best Practices

The following Cisco design best practices and recommendations were used as references in this design.

ACI Best Practices

The best practices for deploying a basic ACI fabric are detailed in the Cisco Application Centric Infrastructure Design Guide White Paper found here: https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737909.html

Performance and Tuning for Cisco UCS M5 Servers

The following white paper was referenced for adjusting the BIOS options to optimize the UCS servers for virtualization workloads: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/whitepaper_c11-740098.pdf

Intel Xeon Scalable 2nd Generation Processor (Cascade Lake) Recommendations

The recommendations for CPU and memory configurations for differing application workloads utilizing the Intel Xeon Scalable 2nd generations processors within Cisco UCS can be found here: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/whitepaper-c11-742358.pdf

IBM FS9100 Storage Considerations

Based on the workload requirements, the FS9100 system should be sized with appropriate cache and I/O cards and ports. Configure a balanced system with performance and capacity targeted volumes and spread the resources by using multiple volumes and combining them at the host. If you're running many workloads, then a single volume might be good enough for each workload. If the balance of the system is leaning towards DRP, consider two Data Reduction Pools.

Port Bandwidth

A single Fibre Channel port can deliver over 1.5 Gbps (allowing for overheads) and an FC card in each canister with 8 ports can deliver more than 12 Gbps. An NVMe device can perform at over 1 Gbps.

A single Fibre Channel port can deliver 80,000 - 100,000 IOPS with a 4 Kb block size. An FC card in each canister with 8 ports can deliver up to 800,000 IOPS. An IBM FlashSystem 9100 can support over 1.1 million 4 Kb read miss IOPS.

So, if you have more than 12 NVMe devices, use two Fibre Channel cards per container, and a third Fibre Channel card enables you to achieve up to 33 Gbps. If you want to drive more than 600,000 IOPS, use two Fibre Channel cards per container.

Cache

256 GB per system (128 GB base plus a 128 GB upgrade) is a good starting point. If you're using data reduction pool or making heavy use of copy services, add a further 128 GB per system. As your capacity increases (especially with the 19.2 TB FCM devices) add more cache to accommodate more of the working set (most accessed workloads, excluding snapshots, backups, and so on). A truly random working set might not benefit from a right-sized cache. If you're consolidating from multiple controllers, consider at least matching the amount of cache across those controllers.

Data Reduction Pools

VersaStack with the inclusion of IBM FlashSystem 9100 supports Data Reduction Pools. Data Reduction Pools are a new type of storage pool that implement several techniques, such as thin-provisioning, compression, and

deduplication, to reduce the amount of physical capacity required to store data. Savings in storage capacity requirements translate into reduction in the cost of storing the data.

Compression and deduplication are not mutually exclusive, one or both or neither features can be enabled. If the volume is deduplicated and compressed, data is deduplicated first, and then compressed. Therefore, deduplication references are created on the compressed data stored on the physical domain.

Data Reduction Pools are useful if the underlying media does not have hardware acceleration (for example, when using NVMe SSDs instead of Flash Core Modules), or if the goal is to ensure the most data reduction possible, by enabling deduplication. When using Flash Core Modules, take care to ensure that compression is not being used as a space saving method when creating volumes to present to hosts. Little benefit will be gained from trying to compress data that has already been compressed.

Distributed RAID

IBM FlashSystem 9100 requires grouping multiple individual NVMe drives together into a Distributed RAID array. Each array (sometimes referred to as Managed Disk or mdisk) must be added to a storage pool from which Volumes can be presented to hosts. A given array can only be assigned to a single pool at any time.

As a technology, Distributed RAID benefits from having large numbers of drives per array, both in read/write performance, and the time taken to rebuild in the event of a drive failure. In Spectrum Virtualize 8.2.1 firmware, the implementation of Distributed RAID does not allow for the expansion of an array, so the only way to increase the size of a storage pool is to create additional Distributed RAID arrays and add them to the pool.

However, depending on the number of NVMe drives installed in the FlashSystem 9100, and the application requirements, consider whether either Data Reduction Pools, Standard Pools or both are required. If both pools are required, divide the number of drives between the two pools and create the arrays with as many drives as possible.

Creating one 24-drive array is more efficient and is better performing than two 12-drive arrays, unless both pool types are required in which create two 12-drive arrays.

Multiple Volumes

IBM FlashSystem FS9100 is optimized for multiple volumes, and around 30 volumes are required to unlock the maximum performance. A workload can become unnecessarily limited when backed by fewer volumes, and a single volume is limited to up to 10% of the ultimate performance. This is due to the relationship between the host interface adapter port(s), and how internal resources within Spectrum Virtualize are allocated to CPU threads/cores. Adding volumes initially scales performance linearly and enables the workload to be balanced across the ports and canisters.

The following IBM Redbook can be referred for IBM FlashSystem 9100 Best Practices and Performance Guidelines: http://www.redbooks.ibm.com/redpieces/pdfs/sg248448.pdf

Deployment Hardware and Software

Hardware and Software Revisions

Table 6 lists the hardware and software versions used for the solution validation. It is important to note that Cisco, IBM, and VMware have interoperability matrices that should be referenced to determine support for any specific implementation of VersaStack. Please see the following links for more information:

- IBM System Storage Interoperation Center
- Cisco UCS Hardware and Software Interoperability Tool
- VMware Compatibility Guide

Table 6 Validated Hardware and Software Revisions

Layer	Device	Image	Comments
Compute	Cisco UCS Fabric Interconnects 6400 Series, Cisco UCS B200 M5 and Cisco UCS C220 M5	4.0 (4e)	Includes the Cisco UCS-IOM 2208XP, Cisco UCS Manager, Cisco UCS VIC 1440 and Cisco UCS VIC 1457
	Cisco nenic Driver	1.0.29.0	Ethernet driver for Cisco VIC
	Cisco nfnic Driver	4.0.0.40	FCoE driver for Cisco VIC
Network	Cisco APIC	4.2(1j)	ACI Controller
	Cisco Nexus Switches	N9000-14.2(1j)	ACI Leaf Switches
	Cisco ExternalSwitch	1.1	UCS Integration with ACI
Storage	IBM FlashSystem 9110	8.2.1.6	Software version
Virtualization	VMware vSphere ESXi	6.7 update 3	Software version
	VMware vCenter	6.7 update 3	Software version
	Cisco ACI Plugin	4.2.1000.10	VMware ACI Integration

Validation

Test Plan

The solution was validated by deploying multiple VMs running tools such as IOMeter and Vdbench. The system was validated for resiliency by failing various aspects of the system under the load. The following sections provide examples of the tests executed for this solution.

Cisco UCS Validation

- Failure and recovery of links from Cisco UCS Chassis (IOM) and Cisco UCS C-Series servers to FI-A and FI-B
- Rebooting Cisco UCS FI, one at a time
- Removing the physical cables between FI and Cisco Nexus 9336C-FX2 switches to simulate path failure

Network Validation

- Fail/power off both Cisco 9336C-FX2 switches, one after another
- Failure and recovery of links in the vPC from UCS FI and Cisco Nexus 9336C-FX2 switches
- Failure and recovery of physical links from Cisco Nexus switches and the Cisco Nexus 7000 switches

Storage Validation

- Failure and recovery of the links between Cisco Nexus 9336C-FX2 leaf switches and IBM FS9100 storage controllers
- Failure and recovery of IBM FS9100 each storage node in the enclosure
- VMware ESXi host iSCSI SAN Boot support for multipath failover

vSphere Validation

- Failure and recovery of ESXi hosts in a cluster (rebooting of hosts, shutting down of hosts, and so on)
- In case of a host failure, verify VM auto restart within the high availability cluster
- VM vMotion across ESXi servers

Summary

VersaStack delivers a platform for enterprise and cloud datacenters using Cisco UCS Blade and Rack Servers, Cisco Fabric Interconnects, Cisco Nexus 9000 switches, Cisco MDS switches, and Fibre Channel or iSCSI attached IBM Storage Arrays.

VersaStack with Cisco ACI and IBM FS9100 NVMe-accelerated Storage solution is designed to simplify the data center evolution to a shared cloud-ready infrastructure by using an application driven policy model. The solution delivers an application centric architecture with centralized automation that combines software flexibility with the hardware performance. Cisco and IBM have created a platform that is both flexible and scalable for multiple use cases and applications. From virtual desktop infrastructure to SAP®, VersaStack can efficiently and effectively support business-critical applications running simultaneously from the same shared infrastructure. The flexibility and scalability of VersaStack also enable customers to start out with a right-sized infrastructure that can ultimately grow with and adapt to their evolving business requirements.

References

Products and Solutions

Cisco Unified Computing System:

http://www.cisco.com/en/US/products/ps10265/index.html

Cisco UCS 6400 Series Fabric Interconnects:

https://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-6400-series-fabric-interconnects/tsdproducts-support-series-home.html

Cisco UCS 5100 Series Blade Server Chassis:

http://www.isco.com/en/US/products/ps10279/index.html

Cisco UCS B-Series Blade Servers:

http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-b-series-blade-servers/index.html

Cisco UCS C-Series Rack Servers:

http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c-series-rack-servers/index.html

Cisco UCS 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 Intersight:

https://www.cisco.com/c/en/us/products/servers-unified-computing/intersight/index.html

Cisco Nexus 9000 Series Switches:

http://www.cisco.com/c/en/us/support/switches/nexus-9000-series-switches/tsd-products-support-serieshome.html

Cisco Application Centric Infrastructure:

http://www.cisco.com/c/en/us/solutions/data-center-virtualization/application-centric-infrastructure/index.html

Cisco Data Center Network Manager:

https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-data-center-networkmanager/index.html

Cisco UCS Director:

https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-director/index.html

VMware vCenter Server:

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

VMware vSphere:

https://www.vmware.com/tryvmware_tpl/vsphere-55_evalcenter.html

IBM FlashSystem 9100:

https://www.ibm.com/us-en/marketplace/flashsystem-9100

Interoperability Matrixes

Cisco UCS Hardware Compatibility Matrix:

http://www.cisco.com/c/en/us/support/servers-unified-computing/unified-computing-system/productstechnical-reference-list.html

VMware and Cisco Unified Computing System:

http://www.vmware.com/resources/compatibility

IBM System Storage Interoperation Center:

http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss

About the Authors

Sreenivasa Edula, Technical Marketing Engineer, UCS Data Center Solutions Engineering, Cisco Systems, Inc.

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

Warren Hawkins, Virtualization Test Specialist for IBM Spectrum Virtualize, IBM

Working as part of the development organization within IBM Storage, Warren Hawkins is also a speaker and published author detailing best practices for integrating IBM Storage offerings into virtualized infrastructures. Warren has a background in supporting Windows and VMware environments working in second-line and third-line support in both public and private sector organizations. Since joining IBM in 2013, Warren has played a crucial part in customer engagements and, using his field experience, has established himself as the Test Lead for the IBM Spectrum Virtualize™ product family, focusing on clustered host environments

Acknowledgements

For their support and contribution to the design, validation, and creation of this Cisco Validated Design, the authors would like to thank:

- Haseeb Niazi, Technical Marketing Engineer, Cisco Systems, Inc.
- Archana Sharma, Technical Marketing Engineer, Cisco Systems, Inc.