aruba
a Hewlett Packard Enterprise company

Validated Solution Guide

Aruba Solution TME

May 28, 2025

Table of Contents

ESP Data Center Design

This guide helps IT professionals understand the following design considerations for a data center environment:

- Hardware selection
- Software selection
- Topology
- High availability
- Scalability
- Application performance
- Security

NOTE:

For the most up-to-date information on ESP Data Center solutions, refer to the Validated Solution Guide Program

Introduction

HPE Aruba Networking enables customers to build a security-first, AI-powered data center that is a modern, agile service delivery platform. Organizations of any size, distributed or centralized, can benefit from streamlined performance and improved network cost-effectiveness using HPE Aruba Networking switches and management tools.

The HPE Aruba Networking AOS-CX operating system simplifies overall operations and maintenance using a common switch operating system across the campus, branch, and data center. The system can be managed on-premises or in the cloud. AOS-CX employs robust artificial intelligence functions that continually analyze and realign network flow to ensure that the system operates seamlessly in accordance with network management best practice, without requiring manual IT staff intervention.

The use of converged Ethernet has changed the way hosts access storage within the modern data center. Dedicated storage area networks (SANs) are no longer required. Lossless Ethernet and bandwidth management protocols ensure timely reads and writes using a traditional IP LAN. The combined cost savings and operational simplicity are driving a major conversion to converged Ethernet.

At the same time, network topologies have become virtualized. Although virtualization delivers the flexibility required to meet the changing data center requirements, it can present complexity and challenges with implementation and management. HPE Aruba Networking addresses these challenges by automating installation and implementation of the AOS-CX operating system, with features such as automated device group configuration, Zero-Touch Provisioning, scheduled configuration backups, dashboard-ready network performance metrics, and built-in alerts for critical network functions.

Securing applications and hosts in a data center is critical for maintaining application availability, data integrity, and business continuity. New threats such as ransomware, data exfiltration, and denial of service continue to emerge. Policy and security enforcement requires many tools applied at different layers. The Aruba CX 10000 series switch with AMD Pensando introduces an industry-first distributed services data center switch, capable of performing inline firewall services at wire-speed in the switch itself, focusing on the high volume of east-west traffic typical in a data center environment.

Before designing a new or transformed data center, it is important to consider the organization’s current and projected strategy for hosting and accessing applications from the cloud. Determine the applications that will remain on-premises so you can establish the data center with ample storage to meet requirements.

To accommodate growth and future adaptation of the network, implementation of a spine-and-leaf underlay that supports software-defined overlay networks is highly recommended. The Aruba Networks CX 83xx, 84xx, 9300, and 10000 switching platforms provide a best-in-class suite of products featuring a variety of high-throughput port configurations, industry-leading operating system modularity, real-time analytics, and “always up” performance.

Purpose of This Guide

This guide describes the HPE Aruba Networking data center deployment, with reference for architectural options and associated hardware and software components. It explains the requirements that shaped the design and the benefits it provides. It introduces data center solutions that support options for both distributed and centralized workloads. It delivers best practice recommendations to deploy a next generation spine-and-leaf data center fabric using VXLAN and BGP EVPN.

Design Goals

The overall goal is to create a high-reliability, scalable design that is easy to maintain and adapt as business needs change. Solution components are limited to a specific set of products required for optimal operations and maintenance.

Key features of the HPE Aruba Networking data center network include:

- Zero downtime upgrades
- High throughput
- Security
- Converged storage networking
- Flexible segmentation
- Third-party integration.

This guide can be used to design new networks or to optimize and upgrade existing networks. Not intended as an exhaustive discussion of all options, the guide focuses on commonly recommended designs, features, and hardware.

Customer Use Cases

Data center networks change rapidly. The most pressing challenge is to maintain operational stability and visibility for users when moving or upgrading computing and storage resources. In addition, data center teams must continue to support the rapid pace of DevOps environments and meet growing requirements to connect directly and continue operations within the public cloud infrastructure.

In a rapidly changing landscape, it is critical that network and system engineers responsible for meeting data requirements have efficient tools to streamline and automate complex infrastructure configurations.

This guide discusses the following use cases:

- Pay-as-you-grow designs that support network and computing workload elasticity
- Ease of use and agility to deploy and manage workloads quickly by orchestrating computing, hypervisor, and network management functions
- Improved operations with data center visibility from the computing host to the overall network infrastructure
- Workload mobility, security, and multi-tenancy using standards-based overlay technologies
- Network infrastructure automation and management
- Data aggregation and pre-processing.

Data Center Connectivity Design

The HPE Aruba Networking data center provides flexible and highly reliable network designs to ensure efficient, reliable access to applications and data for all authorized users while simplifying operations and accelerating service delivery.

HPE Aruba Networking data centers are built on the following switch models:

- CX 8xxx Ethernet switches
- CX 9300 Ethernet switches
- CX 10000 Ethernet switches with Pensando
- CX 63xx Ethernet switches for out-of-band (OOB) network management.

Data Center Topologies

HPE Aruba Networking data centers support centralized and distributed workloads anywhere within an organization. Each design supports host uplink bundling to provide high throughput and resiliency for mission-critical workloads. Layer 2 domains can be deployed flexibly to meet application requirements and provide virtual host mobility.

CX switches provide a robust platform for Layer 3 services in the data center. HPE Aruba Networking data center designs are primarily implemented using CX 8xxx, CX 9300, and CX 10000 series Ethernet switches that provide low latency and high bandwidth on a fault-tolerant platform designed to carry data center traffic in a 1U form-factor.

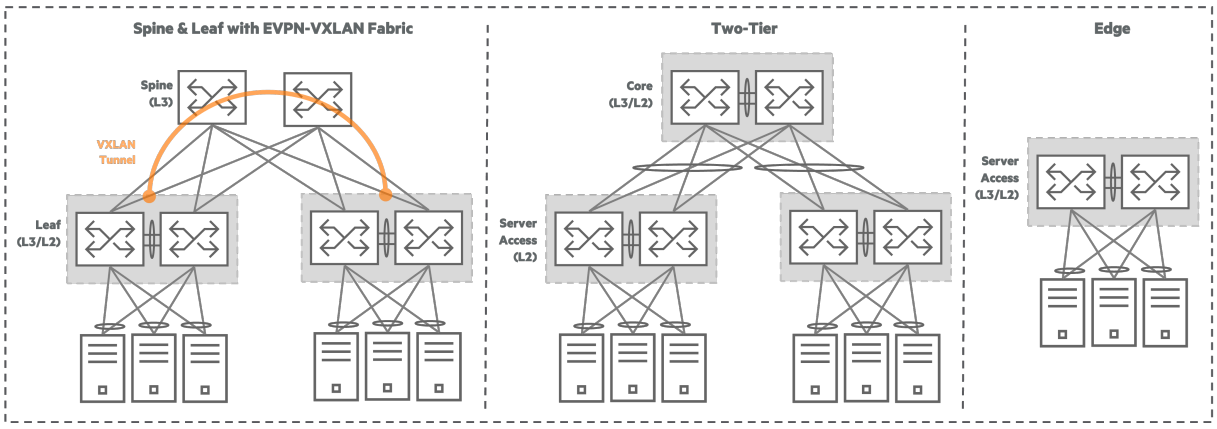


Figure 1: Aruba Data Center Designs

Spine-and-Leaf with VXLAN Fabric Overview

The most modern and flexible data center design is an EVPN-VXLAN overlay built on a spine-and-leaf routed underlay, which benefits enterprises with growing on-premises workloads and workloads spread across multiple data centers.

The spine-and-leaf underlay design ensures high reliability and horizontal scaling using redundant Layer 3 links between leaf and spine switches. This Clos-based topology provides equal-cost multipath (ECMP) routing to load balance traffic and support fast failover if a link or switch goes down. The fully meshed architecture enables capacity growth simply by adding another spine switch as needed.

An EVPN-VXLAN overlay allows ubiquitous Layer 2 adjacency across the entire data center using VXLAN tunneling. This enables customers to modernize their network while preserving legacy service requirements by connecting physically dispersed Layer 2 segments in the overlay. Physically distant data centers also can extend both Layer 2 and Layer 3 segments logically.

EVPN-VXLAN natively enables segmenting groups of resources within the data center to support multi-tenancy and separation of hosts by role such as production, development, tenant, and those requiring strict regulatory compliance.

Two-Tier Overview

Enterprises with significant, existing on-premises workloads in a single location can benefit from a two-tier data center design. The two-tier approach ensures sufficient capacity and reliability using standards-based protocols such as Link Aggregation Control Protocol (LACP), Spanning-Tree Protocol (STP), and Open Shortest Path First (OSPF). Hosts are dual-homed to two top-of-rack (ToR) switches using a Virtual Switch Extension (VSX) multi-chassis link aggregation group (MC-LAG). Each ToR switch is dual-homed to a data center core using Layer 2 VSX/MC-LAG links. Using Layer 2 between the core and server access layers supports VLAN ubiquity, and loops are prevented primarily using LACP-based MC-LAGs. The core provides Layer 3 services to data center hosts and routing to external networks.

The physical structure of a two-tier data center enables a migration path to an EVPN-VXLAN spine-and-leaf data center in the future.

Edge Data Center Overview

Enterprises that have migrated most of their workloads to the cloud and no longer require a large on-premises data center can use their existing campus network wiring closets or small server rooms to deploy small on-premises workloads.

The same AOS-CX switches that provide wired connectivity to users and Internet of Things (IoT) devices can be leveraged to provide server access.

An edge data center supports high-bandwidth and low-latency access to computing and storage resources for distributed workloads that may not be well suited to cloud deployments.

General Design Considerations

Out-of-Band Management

HPE Aruba Networking data center designs use a dedicated management LAN connecting to switch management ports and to host lights-out management (LOM) ports. Typically, a single management switch is deployed at every rack for OOB management. A dedicated management switch ensures reliable connectivity to data center infrastructure for automation, orchestration, and management without risk of disrupting management access when making changes to the data center's data plane configuration.

Top-of-Rack Design

Deploying switches in the ToR position enables shorter cable runs between hosts and switches. The result is a more modular solution with host-to-switch cabling contained inside a rack enclosure and only switch uplinks exiting the enclosure. This approach helps reduce complexity when adding racks to the data center.

In typical data centers, each rack is serviced by a redundant pair of switches. This enables connection of dual-homed hosts to two physical switches using an MC-LAG bundle for fault tolerance and increased capacity. CX switches use two different strategies to support MC-LAGs: VSX switch pairing and Virtual Switching Framework (VSF) switch stacking.

VSX enables a distributed and redundant architecture that is highly available during upgrades. It virtualizes the control plane of two switches to function as one device at Layer 2 and as independent devices at Layer 3. From a data-path perspective, each device performs an independent forwarding lookup to decide how to handle traffic. Some of the forwarding databases, such as the MAC and ARP tables, are synchronized between the two devices via the VSX control plane over a dedicated inter-switch link (ISL) trunk. Each switch builds Layer 3 forwarding databases independently.

When deploying a pair of switches in VSX mode, at least two ports must be members of the LAG assigned as the ISL, which supports control plane functions and serves as a data path between the switch pair. The ISL ports should be the same speed as the uplinks ports.

VSX requires a keepalive between members to detect a split brain condition, which occurs when communication over the ISL is no longer functional. Best practice is to configure the keepalive to use the OOBM port when using a dedicated management network. Alternatively, a loopback IP address or a dedicated low-speed physical port can be used for keepalive traffic. Loopback-based communication is supported over redundant routed paths for increased resiliency.

VSF combines a set of two to ten CX 6300 switches into a high availability switch stack using a ring topology. Data centers use VSF stacks to connect racks of 1 Gbps connected hosts to upstream leaf switches. A VSF stack operates with a single Layer 2 and Layer 3 control plane. One switch member of the stack operates in the *Conductor* role, which manages all other switch stack members. A second

member of the stack operates in the *Standby* role. The *Conductor* synchronizes state and configuration information with the *Standby* switch, so it can assume the *Conductor* role in case of failure. Monitoring for a split-brain condition is achieved using the the OOBM port of each stack member connected to a common management network.

For the most common connection speeds and backward compatibility, choose a ToR switch that supports access connectivity rates of 1, 10, and 25 Gbps on each port. These connection speeds can be increased simply by upgrading transceivers, DACs, or AOCs.

For high throughput compute racks, CX 9300 and CX 9300S switches supports 100 and 200 Gbps host connectivity with 400 Gbps switch uplinks. Breakout cabling and AOCs support connecting four QSFP56-based 100 Gbps host NICs, two QSFP56-based 200 Gbps host NICs, or two QSFP28-based 100 Gbps host NICs to one physical CX 9300-32D switch port. The 9300S can also be optimized to support 25 Gbps hosts.

Keep the following in mind when selecting ToR switches:

- **Number and type of server connections:** Typical ToR switch configurations support 48 host-facing ports, but lower-density ToR options are available in the CX 8360 series. A rack of 1 Gbps hosts can be connected using the CX 6300 series.
- **Host connectivity speed:** To simplify management, consolidate hosts connecting at the same speeds to the same racks and switches. Adapting the port speed settings of a particular interface may impact a group of adjacent interfaces. Consider interface group size when planning for a rack requiring multiple connection speeds. High speed storage and compute hosts connecting at 100 Gbps or 200 Gbps require the CX 9300-32D switch.
- **ToR-to-spine/core connectivity:** ToR switch models support a range of uplink port densities. The number and port speed of the uplinks define the oversubscription rate from the hosts to the data center fabric or data center core. For example, in a four-spine fabric deployment at 100 Gbps, a non-oversubscribed fabric can be implemented for racks of 40 servers connected at 10 Gb.
- **VSX uplink consumption:** When using VSX for redundancy, two uplink ports are consumed for ISLs providing data-path redundancy and cannot be used for spine or data center core connectivity.
- **DSS feature requirements:** The CX 10000 is required in a data center design that implements inline stateful firewall inspection using the AMD Pensando programmable DPU.
- **Cooling design:** Different ToR models are available for port-to-power and power-to-port cooling. In power-to-port configurations, an optional air duct kit can isolate hot air from servers inside the rack. Cabling can absorb heat and restrict airflow. Short cable routes and good cable management improve airflow efficiency.

Host Connectivity

A critical step in designing a data center is identifying the types of connectivity required by the computing hosts. Server hardware typically has an Ethernet RJ45 port for a lights-out management device such as HPE iLO. The lights-out port is typically connected using a Cat5e or Cat6 copper patch cable to a switch on the management LAN.

Host connections are usually 10 Gb or 25 Gb using SFP+/SFP28 fiber modules, copper direct-attach cables (DACs), or active optical cables (AOCs). DACs have limited distance support and can be harder to manage than optical cables due to the thicker wire gauge. AOCs support longer distances than DACs. AOCs are thinner and easier to manage. Both DACs and AOCs cost less than separate optical transceivers with fiber patch cables.

High-speed host connectivity is supported using QSFP-DD transceivers and AOCs with the CX 9300 switch. Both optics and AOCs can breakout a single high-speed 400 Gbps switch port into multiple lower speed 200 Gbps and 100 Gbps connections. The switch can support QSFP56 and QSFP28-based host NICs. The AOS-S and AOS-CX Transceiver Guide provides detailed information on using breakout cables and AOCs.

It is important to verify that both the host's network interface controller (NIC) and the ToR switch are compatible with the same DAC or AOC. When separate transceivers and optical cables are used, verify transceiver compatibility with the host NIC, ToR switch, and optical cable type. The supported transceiver on the host is often different from the supported transceiver on the switch. Always consult with a structured cabling professional when planning a new or upgraded data center.

When deploying a converged network for IP storage traffic, look for NIC cards that support offload of storage protocols. This minimizes latency of storage traffic by reducing the load on a host CPU.

Applications can be hosted directly on a server using a single operating system, commonly referred to as a "bare-metal" server. Multiple hosts can be virtualized on a single physical server using a hypervisor software layer. Examples include VMware ESXi or Microsoft Hyper-V.

Hypervisors contain a virtual switch that provides connectivity to each virtual machine (VM) using Layer 2 VLANs. A successful data center design should support Layer 2 and Layer 3 connectivity using untagged and VLAN-tagged ports to match the required connectivity to the server and/or virtual switch inside the server. HPE Aruba Networking Fabric Composer provides visibility and orchestration of the configuration between the server and ToR switches to ensure that connectivity is established properly.

Host mobility refers to the ability to move physical or virtual hosts in a data center network without changing the host network configuration. Especially powerful for virtualized hosts, this ensures optimized computing resources, high availability of applications, and efficient connectivity for distributed workloads. EVPN-VXLAN fabrics and two-tier data centers support flexible host mobility, allowing all data center VLANs to be present on all ToR switches. An EVPN-VXLAN design provides tunneled Layer 2 adjacency over a routed underlay, and it can logically extend Layer 2 adjacency between data center locations.

Spine-and-Leaf with EVPN-VXLAN Fabric

An EVPN-VXLAN fabric provides a virtual Layer 2 network overlay that is abstracted from the physical network underlay supporting it. This allows hosts to operate within the same VLAN network segment, even when the hosts are separated by a Layer 3 boundary, by encapsulating traffic within a tunnel. Symmetric Integrated Routing and Bridging (IRB) in EVPN-VXLAN enables Layer 3 routing between overlay network segments.

Physical Topology

The diagram below illustrates the physical connectivity for the complete set of roles in an EVPN-VXLAN solution.

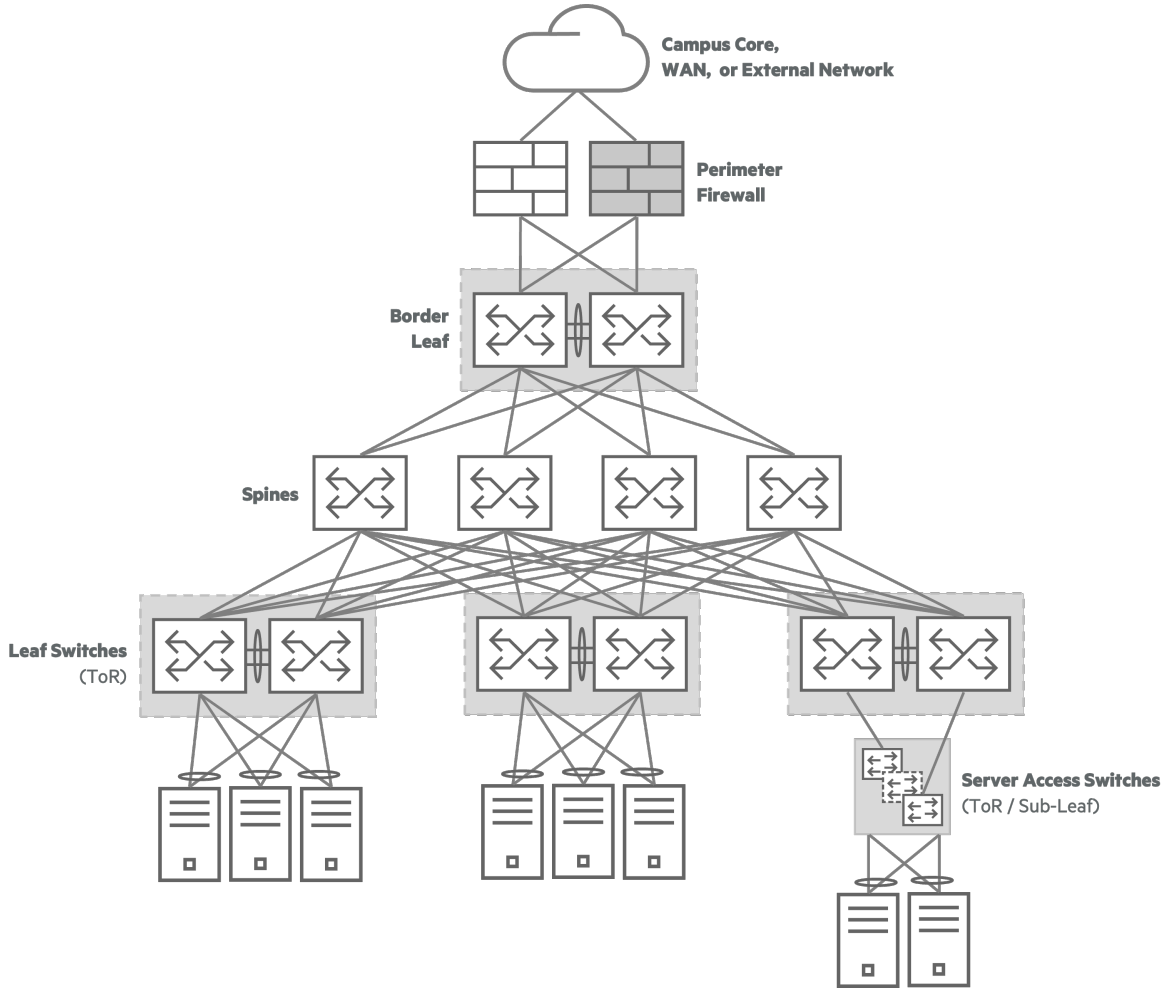


Figure 2: Physical Topology

Underlay Network Design

The underlay of an EVPN-VXLAN data center network provides IP connectivity between switches. The network underlay ensures that VXLAN tunneled traffic (the overlay network) can be forwarded between tunnel endpoints on leaf switches.

The underlay network is implemented using a 3-stage Clos-based spine-and-leaf fabric topology. It is deployed as a Layer 3 routed network, with each leaf connected to each spine using a routed-only port. The spine-and-leaf underlay topology optimizes performance, provides high availability, and reduces latency because each leaf is never more than one hop across multiple load-balanced paths to all other leaf switches.

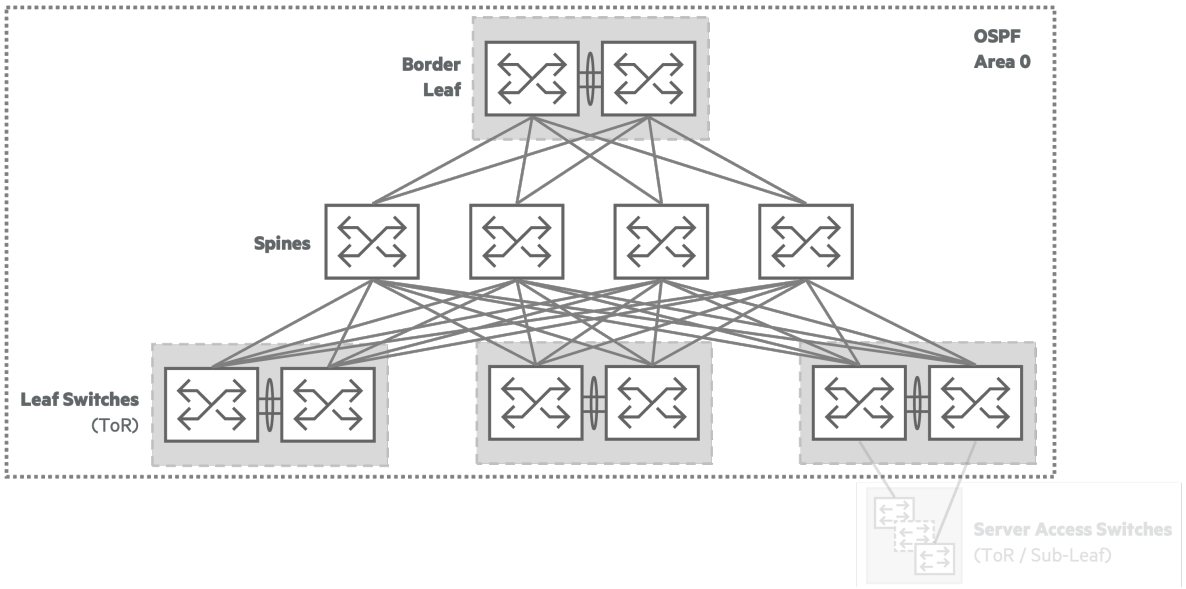


Figure 3: Underlay Network

The spine-and-leaf topology provides a flexible, scalable network design that can accommodate growth without disrupting the existing network. It is easy to begin with a small one- or two-rack fabric that can increase capacity without requiring replacement of existing hardware. Leaf switches are added to new racks to increase computing and network attached storage (NAS) capacity. Spine switches are added to increase east-west fabric capacity between leaf switches.

This topology is roughly analogous to the architecture of a chassis-based switch, where leaf switches are comparable to interface line cards and spines are comparable to the chassis fabric providing data capacity between line cards.

HPE Aruba Networking data centers typically use OSPF as the underlay routing protocol to distribute underlay IP reachability information between all fabric switches. OSPF is a widely used, well understood Interior Gateway Protocol (IGP) that provides straightforward configuration and fast convergence. When adding an EVPN-VXLAN overlay, the underlay route table is small, consisting primarily of loopback IP

addresses to establish overlay routing protocol adjacencies and VXLAN tunnel endpoint reachability. OSPF routing in the underlay also enables the selection of appropriate overlay routing protocols to support a multifabric environment. A single OSPF area and point-to-point interfaces are recommended to minimize complexity.

Setting a Layer 2 and IP maximum transmission unit (MTU) of 9198 bytes on underlay interfaces connecting spine and leaf switches avoids fragmenting the jumbo sized frames created when adding VXLAN encapsulation.

Server access switches do not participate in the routed underlay. They are connected to upstream leaf switches using Layer 2 only links.

Spine Design

The spine layer provides high-speed, routed connectivity between leaf switches. In a spine-and-leaf architecture, each leaf switch is connected to each spine switch. Each leaf-to-spine connection should use the same link speed to ensure multiple equal-cost paths within the fabric. This enables routed ECMP-based load balancing and ensures connectivity if a link goes down.

All spine switches must be the same switch model. The port capacity of the spine switch model defines the maximum number of leaf switches in a single spine-and-leaf instance. For a redundant ToR design, the maximum number of leaf racks is half the port count on the spine switch model.

A typical spine-and-leaf network begins with two spines for high availability. Spine switches are added to increase fabric capacity and fault tolerance. The impact of a spine failure is reduced for each additional spine added to the underlay. The loss of a single spine reduces overall fabric capacity in the following manner:

- 2 spines: 50% capacity reduction
- 3 spines: 33% capacity reduction
- 4 spines: 25% capacity reduction

The maximum number of spines is determined by the leaf switch model with the fewest number of uplinks. In a redundant ToR design, the maximum number of spines is the uplink port count of the leaf switch with the fewest uplinks minus two, as two of the uplinks are consumed for a VSX inter-switch link (ISL) between the ToRs. In a single ToR design, the maximum number of spines is equal to the number of uplink ports of the leaf switch with the fewest uplinks. Each ToR switch must connect to each spine for ECMP to work effectively.

A CX 9300-based spine offers high-density rack support when using breakout cabling. A CX 9300 can break out a single 400 Gbps spine port to four 100 Gbps connected leaf switch ports over single-mode fiber. It also can support two 100 Gbps connected leaf switches per spine port over multimode fiber and AOCs. This allows the CX 9300 to double or quadruple the number of racks supported over its physical port count.

Using the CX 9300 in a leaf role supports extreme horizontal CX 9300 spine scaling. When dedicating half of a 9300-32D leaf switch's available ports to host connectivity, a 5.6 Tbps fabric comprising 14 spines can be implemented to support a redundant ToR design. A 6.4 Tbps fabric comprising 16 spines can be implemented in a single ToR design. A CX 9300 spine and leaf combination can support connecting multiple links to each spine, if the required number of racks permits. This deployment model supports very high-throughput racks containing 100 Gbps connected storage and compute hosts.

Consider the following when selecting switches:

- Determine rack media and bandwidth requirements.
- Determine if single or redundant ToR switches will be installed.
- Determine how many racks are needed for current computing and storage requirements.
- Determine the spine switches required to support the planned racks.
- Design the data center network for no more than 50% capacity to leave room for growth.

When deciding where to physically place the spine switches, consider their distance from the leaf switches and the media type used to connect them. Spine-to-leaf connections are generally 40 Gb or 100 Gb fiber using quad SFP (QSFP) transceivers or AOCs, in which the cable and transceiver are integrated, similar to DACs. The CX 9300-32D can support up to 400 Gbps spine-to-leaf connections for higher speed data center applications.

Overlay Data Plane Network

An overlay network is implemented using VXLAN tunnels that provide both Layer 2 and Layer 3 virtualized network services to workloads directly attached to leaf switches. VXLAN Network Identifiers (VNIs) identify distributed Layer 2 and Layer 3 segments in a VXLAN overlay topology. Symmetric IRB enables overlay networks to support contiguous Layer 2 forwarding and Layer 3 routing across all leaf nodes.

A VXLAN Tunnel End Point (VTEP) is the function within leaf switches that handles the origination and termination of point-to-point tunnels forming an overlay network. A VTEP encapsulates and decapsulates Layer 2 Ethernet frames inside a VXLAN header in a UDP datagram. The source VTEP assigns a VNI to inform the destination VTEP of the VLAN or route table associated with the encapsulated frame. A single logical VTEP is implemented when VSX redundant leaf switches are deployed in a rack. VTEP IP addresses are distributed in the underlay network using OSPF. Spine switches provide IP transport for the overlay tunnels but do not participate in the encapsulation/decapsulation of VXLAN traffic.

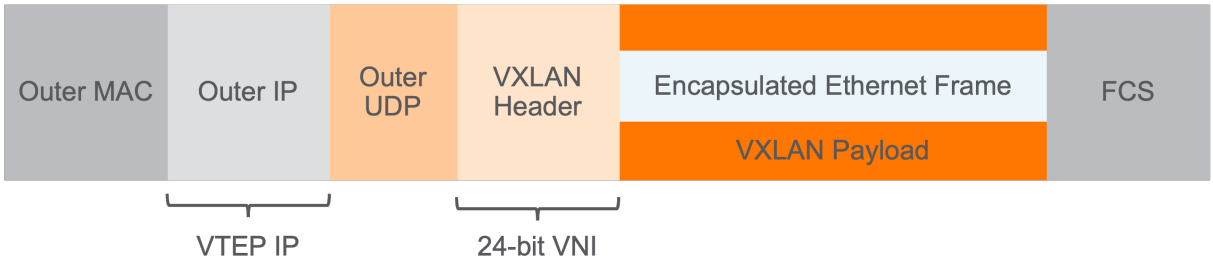


Figure 4: VXLAN Frame

The diagram below illustrates the VXLAN data plane network, a full mesh of VXLAN tunnels between leaf switch VTEPs.

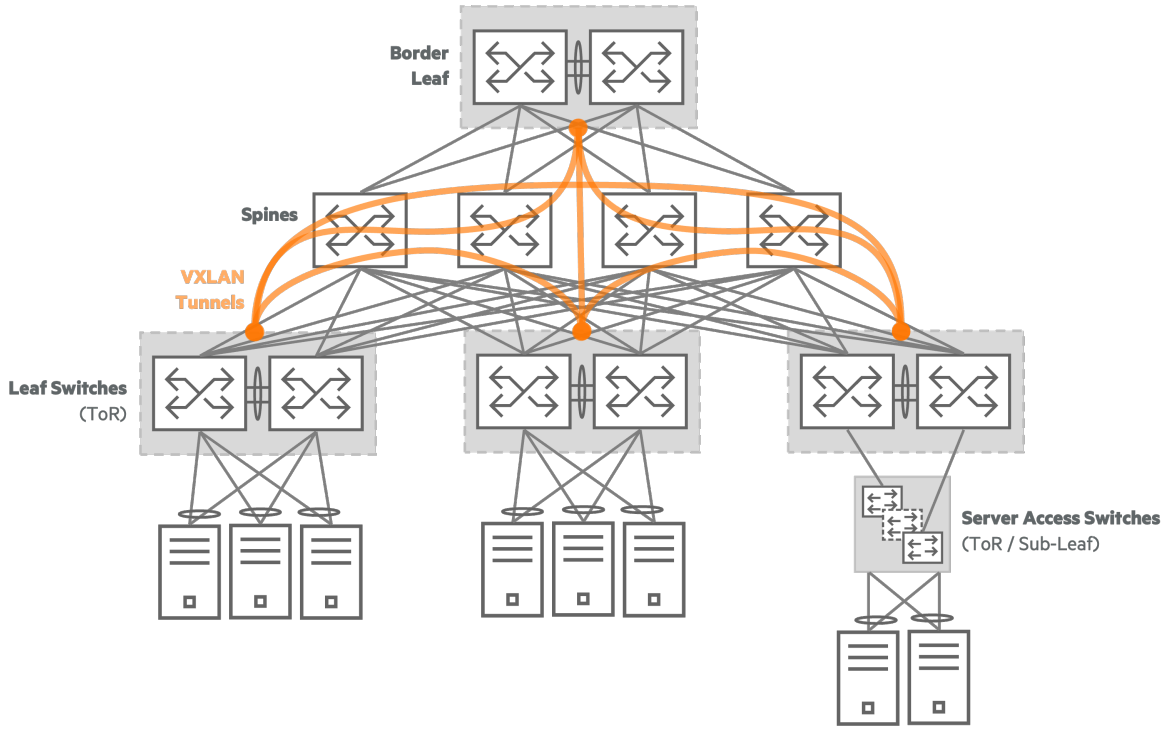


Figure 5: Overlay Network

NOTE:

Server access switches do not contain VTEPs or participate in VXLAN forwarding. They support data center host attachment to the overlay by extending VLANs from fabric leaf switches.

A Layer 2 VNI represents a single broadcast domain, similar to a traditional VLAN ID. A Layer 2 VNI is associated with a VLAN on individual switches in the fabric and collectively stitches the set of distributed VLANs into a single Layer 2 broadcast domain. When a VXLAN encapsulated frame arrives at a VTEP termination with a Layer 2 VNI, the switch unencapsulates the frame and forwards it natively based on the MAC address table of the VLAN associated with the VNI.

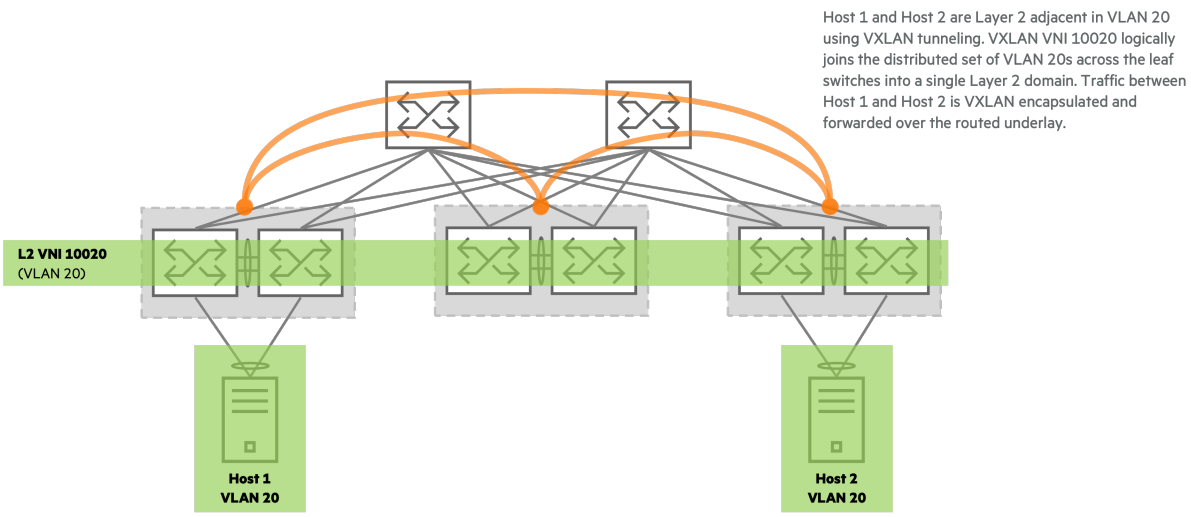


Figure 6: L2 VNI Broadcast Domain

A switch supports multiple routing domains by implementing virtual routing and forwarding instances (VRFs). Each VRF consists of a unique route table, member interfaces that forward traffic based on the route table, and routing protocols that build the route table. Different VRFs may contain overlapping IP address ranges because the individual route tables are discrete. An EVPN-VXLAN overlay must consist of at least one non-default VRF as a container for overlay VLAN SVIs and routed interfaces. Multiple VRFs can be used to provide segmentation for multi-tenancy and policy enforcement.

A Layer 3 VNI associates VXLAN encapsulated traffic with a VRF. When a VXLAN encapsulated packet arrives at a VTEP with a Layer 3 VNI, the packet is unencapsulated and forwarded using the associated VRF's route table.

Symmetric IRB uses a distributed IP gateway model. The gateways for fabric host VLANs are anycast IP addresses. The same virtual IP address is assigned to the same VLAN on each switch in the fabric using HPE Aruba Networking's Active Gateway feature. A distributed virtual MAC address is also associated with the virtual IP. This strategy supports moving active VM guests between hypervisors, which are attached to switches in different racks.

NOTE:

When using Active Gateway to create a distributed overlay IP gateway address across all leaf switches, the Active Gateway IP address also is typically assigned to the VLAN SVI on each switch to conserve IP addresses. The Active Gateway IP is not supported as a source address when using the ping command, and a unique VLAN SVI address is not available as a source IP. When testing host reachability, the ping command must specify a unique source IP address, such as a loopback IP assigned to the same VRF, to verify reachability.

Traffic from a host that must be routed hits the virtual gateway IP address assigned to its directly attached leaf switch. Both the source and destination VTEP perform routing functions, and the source VTEP assigns an L3 VNI to inform the destination VTEP of the appropriate VRF for forwarding.

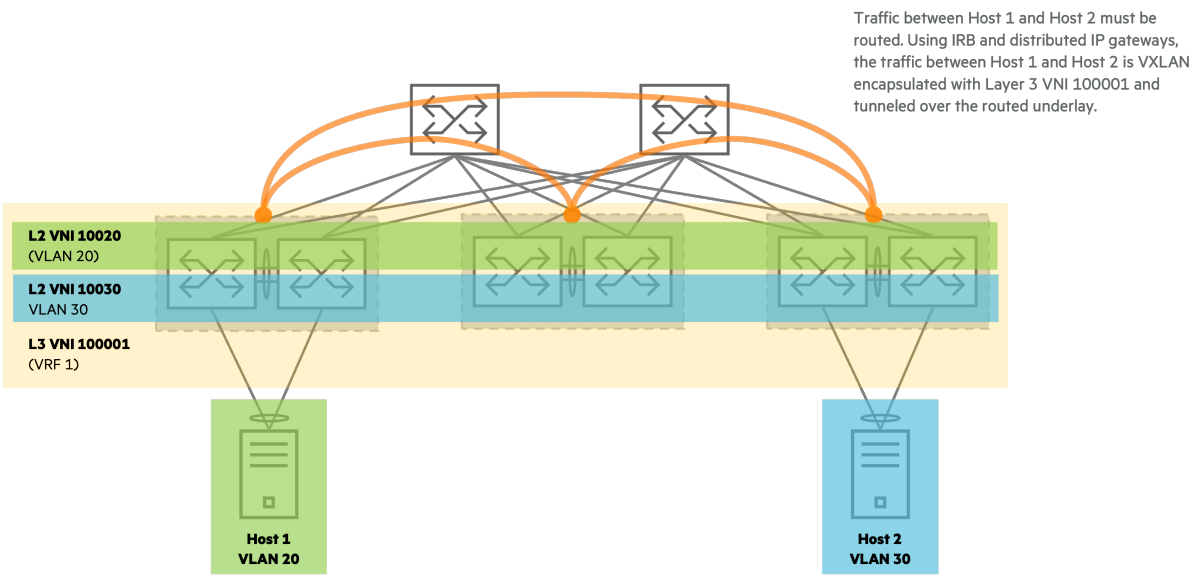


Figure 7: L3 VNI Routing Domain

Border leaf switches provide connectivity between EVPN fabric hosts and external networks such as the campus, WAN, or DMZ. A firewall is typically placed between the border leaf and external networks to enforce north/south security policy.

Server Access Layer

Server access switches extend overlay VLANs from an EVPN-VXLAN leaf to a high density set of lower-speed connected hosts using an economical Layer 2 switch model. They do not participate directly in underlay routing or overlay virtualization.

Server access switches provide Layer 2 redundancy to attached hosts using VSX or Virtual Switching Framework (VSF). VSF enabled switches operate as a single logical stack across one or more racks adjacent to their uplink leaf. VSF supports redundant MC-LAGs to downstream hosts and protection against split-brain conditions using the out-of-band management port to monitor the status of stack members.

When making the transition to an EVPN-VXLAN overlay in an established data center, existing ToR switches can be Layer 2-connected as server access switches to an EVPN-VXLAN leaf as part of the transition strategy. This permits moving existing rack infrastructure into the new EVPN-VXLAN overlay on a flexible timeline without the requirement to replace existing ToR switches.

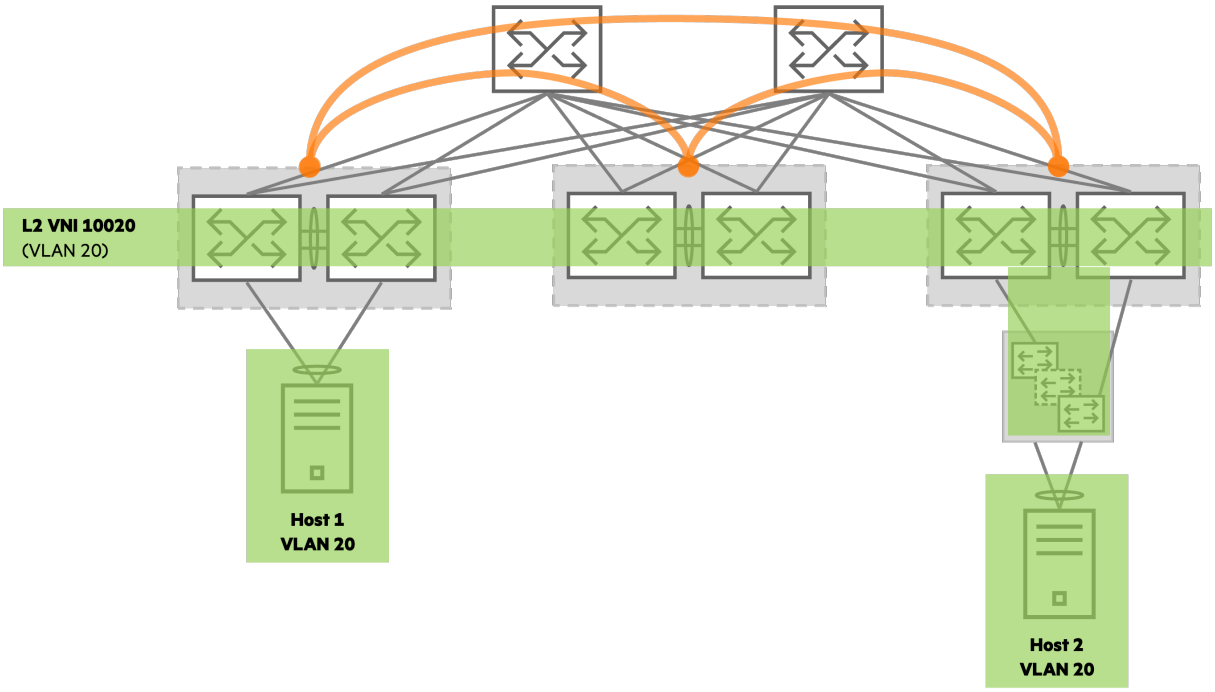


Figure 8: Server Access VLAN Extension

Overlay Control Plane

The VXLAN control plane distributes information for sharing host reachability and dynamically building VXLAN tunnels. Reachability between endpoints in a VXLAN network requires associating fabric connected endpoints with their respective VTEP and VNIs across all fabric switch members. This reachability information is used by a source VTEP to assign a VXLAN VNI in the VXLAN header and the destination VTEP IP in the IP header.

Attached hosts are learned at their uplink leaf switch using Ethernet link layer protocols. Overlay reachability information across the VXLAN fabric is distributed using Multiprotocol Border Gateway Protocol (MP-BGP) as the control plane protocol using the EVPN address family. BGP advertises both host IP and MAC prefixes. This approach minimizes flooding while enabling efficient, dynamic discovery of all hosts within the fabric.

Using a distributed control plane that dynamically populates endpoint information provides the following benefits:

- It avoids flood-and-learn techniques that can consume large amounts of bandwidth due to the replication of traffic in a large spine-and-leaf environment.
- Network configuration is simplified as fabric leaf VTEP switches automatically discover peer VTEP switches inside the fabric, building dynamic VXLAN tunnels.
- A distributed control plane provides redundancy and a consistent topology state across the data center fabric switches.
- A distributed control plane allows optimal forwarding using distributed gateways at the ToR switches. This enables default gateway addresses to remain the same across the fabric.

The use of MP-BGP with the EVPN address family provides a standards-based, highly scalable control plane for sharing endpoint reachability information with native support for multi-tenancy. For many years, service providers have used MP-BGP to offer secure Layer 2 and Layer 3 VPN services on a very large scale. Network operations are simplified by using an iBGP design with route reflectors so that peering is required only between leaf switches and two spines. iBGP is required for an individual fabric control plane, when establishing a multifabric environment. Some of the more notable BGP control plane terms include:

- **Address Family (AF):** MP-BGP supports exchanging network layer reachability information (NLRI) for multiple address types by categorizing them into address families (IPv4, IPv6, L3VPN, etc.). The Layer 2 VPN address family (AFI=25) and the EVPN subsequent address family (SAFI=70) are used to advertise IP and MAC address information between MP-BGP speakers. The EVPN address family contains reachability information for establishing VXLAN tunnels between VTEPs.
- **Route Distinguisher (RD):** A route distinguisher enables MP-BGP to carry overlapping Layer 3 and Layer 2 addresses within the same address family by prepending a unique value to the original address. The RD is only a number with no inherently meaningful properties. It does not associate an address with a route or bridge table. The RD value allows support for multi-tenancy by ensuring that a route announced for the same address range in two different VRFs can be advertised in the same MP-BGP address family.
- **Route Target (RT):** Route targets are MP-BGP extended communities used to associate an address with a route or bridge table. In an EVPN-VXLAN network, importing and exporting a common VRF route target into the MP-BGP EVPN address family establishes Layer 3 reachability for a set of VRFs defined across a number of VTEPs. Layer 2 reachability is shared across a distributed set of L2 VNIs by importing and exporting a common route target in the L2 VNI definition. Additionally, Layer 3 routes can be leaked between VRFs using the IPv4 address family by exporting route targets from one VRF that are then imported by other VRFs.
- **Route Reflector (RR):** To optimize the process of sharing reachability information between VTEPs, use route reflectors on the spines to simplify iBGP peering. This design enables all VTEPs to have the same iBGP peering configuration and eliminates the need for a full mesh of iBGP neighbors.

The MP-BGP EVPN address family consists of several route types. - **Route type 2** shares MAC address and host IP reachability information. - **Route type 5** shares IP prefixes that are reachable by a subset of fabric switches, which is most commonly used to share a default route and external prefixes from the border leaf to other leaf switches. - **Route type 3** shares VTEP IP and VNI values to establish VXLAN tunnels dynamically within a fabric.

Route type 2 MAC advertisements are associated with a VLAN based on a route-target value. The same route-target value should be imported and exported for the same VLAN ID on all switches in the fabric. This ensures complete propagation of Layer 2 reachability across the fabric. VLAN route targets can be automatically derived when using an iBGP control plane to simplify configuration and ensure consistency throughout the fabric.

The diagram below illustrates an example of sharing EVPN route-type 2 MAC address reachability using the iBGP control plane.

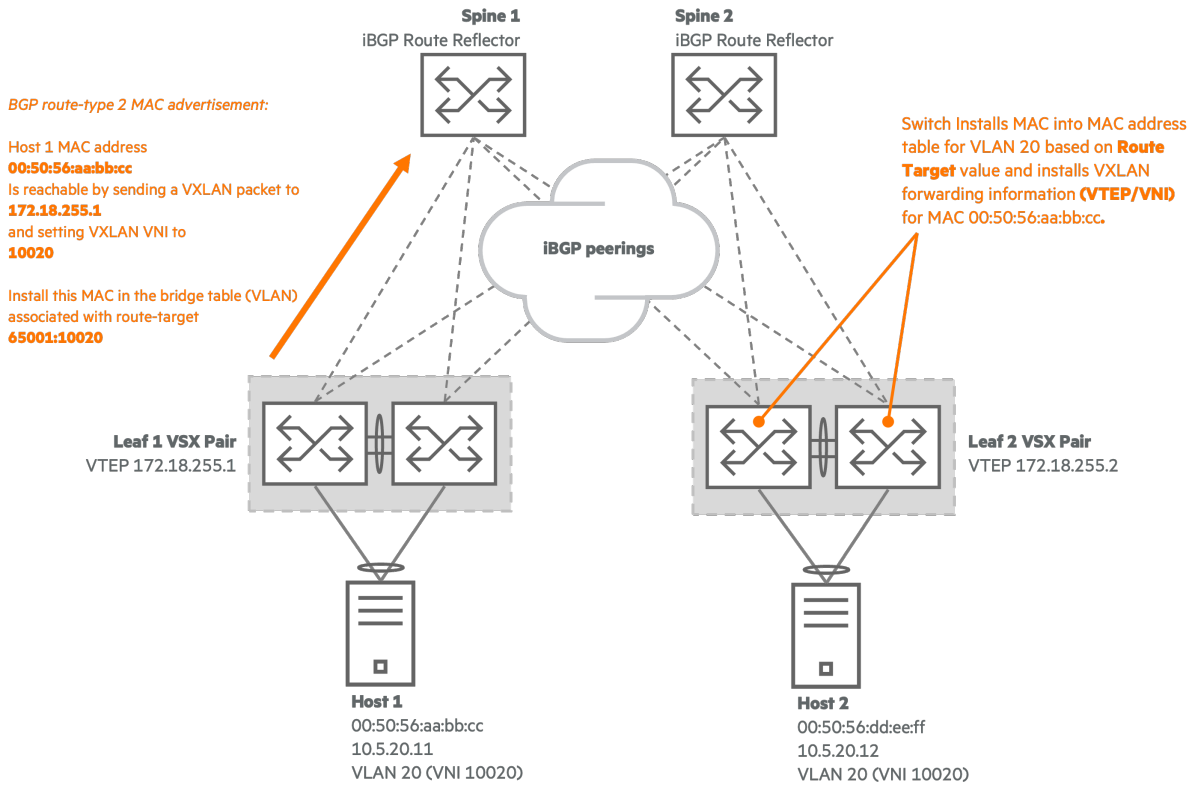


Figure 9: iBGP control plane route-type 2 advertisement

The following screenshot shows an example of an EVPN learned MAC address installed in the MAC address table with its VTEP association.

```
RSVDC-FB1-LF1-1# show mac-address-table
MAC age-time           : 300 seconds
Number of MAC addresses : 36

MAC Address          VLAN     Type                      Port
-------------------------------------------------------------
00:50:56:aa:bb:cc    20       dynamic                   lag1
00:50:56:dd:ee:ff    20       evpn                      vxlan1(172.18.255.2)
...
```

Figure 10: MAC address table with VXLAN target

Route type 5 IP prefixes are associated with a VRF based on a route-target value. A consistent route-target value should be imported and exported for the same VRF on all switches in the fabric. This ensures complete propagation of Layer 3 reachability across the fabric.

Multifabric Underlay

MP-BGP EVPN control plane peerings and VXLAN tunnel termination require establishing IP reachability to loopback interfaces between locations in a multifabric topology. External BGP (eBGP) typically shares loopback/VTEP reachability between sites.

MP-BGP uses an AS number to identify an administrative relationship between BGP speakers. BGP peers with the same AS number are members of the same administrative domain and are considered internal peers (iBGP). BGP peers with different AS numbers are considered external peers (eBGP). Internal and external BGP peers have different default behaviors and requirements. eBGP is often employed between different network segments within the same organization, because default eBGP peering behaviors are useful to the network design.

The diagram below illustrates a set of eBGP IPv4 address family peerings between border leaf switches in a two fabric topology. Layer 2 connectivity between sites is provided by a metro Ethernet circuit. Routed interfaces on each border leaf switch establish a peering relationship with each border leaf switch in the remote fabric. Loopback IP addresses are shared, to establish MP-BGP EVPN peerings and VTEP tunnel terminations.

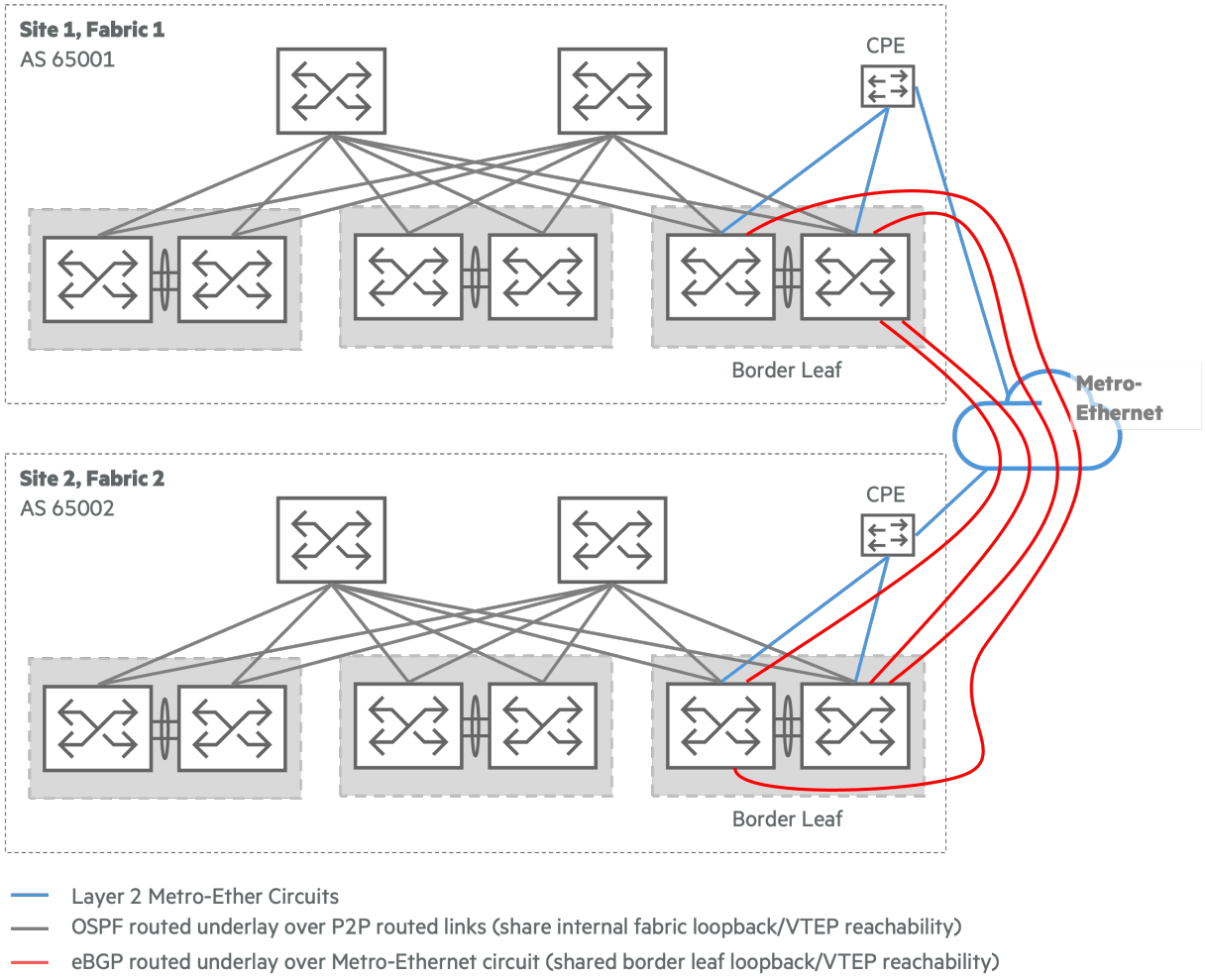


Figure 11: eBGP IPv4 simple underlay peering

Underlay eBGP peerings typically follow the physical links available between network locations. These links may not align directly with the control plane EVPN peerings. Dark fiber and metro Ethernet circuits are common connectivity options between sites.

As the number of interconnected fabrics increase, the number of high speed circuits required at the primary site may exceed the number of ports available on the border leaf switches. Available high speed ports on a spine switch can be used as part of a multifabric underlay. The WAN paths and MP-BGP IPv4 peerings vary based on each environment's variables and design preferences.

Multifabric Overlay Control Plane

MP-BGP EVPN is used as the control plane in a multifabric overlay, just as with a single fabric overlay.

iBGP is used internally within each fabric. Each leaf switch within a fabric establishes an MP-BGP EVPN address family peering with a pair of route reflectors located on two of the spine switches.

eBGP is used between fabrics to permit VXLAN traffic to be re-encapsulated and forwarded in a second tunnel and to take advantage of useful default behaviors that assist in a multifabric environment.

When more than one fabric is present in a single location, typically only one set of border leaf switches is used to establish an MP-BGP EVPN address family peering with external fabrics over the available WAN path. Any border leaf that peers between sites is called a border leader.

The diagram below illustrates MP-BGP EVPN peerings in a two fabric topology.

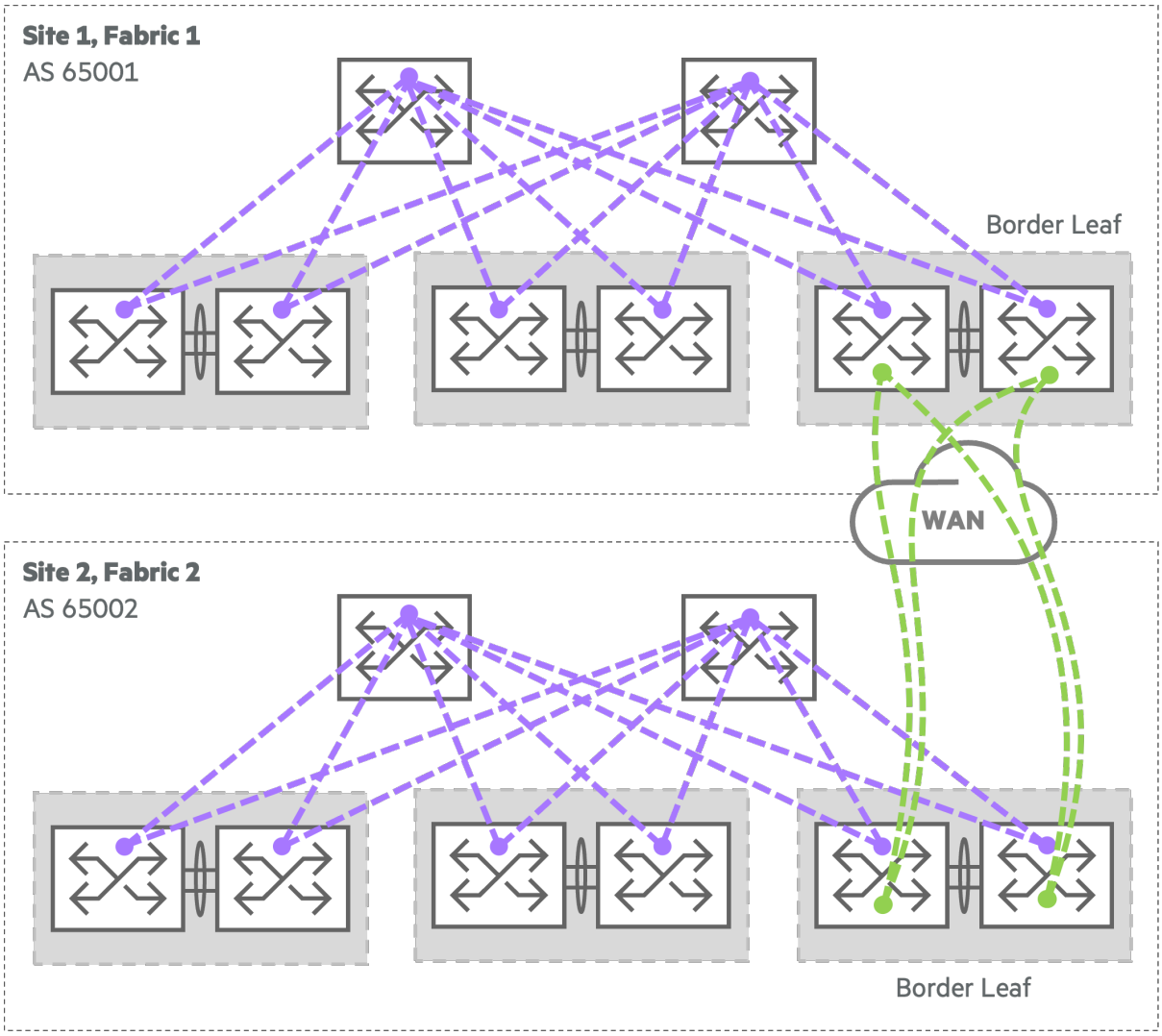


Figure 12: Multifabric BGP control plane peerings

Additional route-target values are defined to control the installation of reachability information between fabrics. Each VLAN and VRF is assigned an intra-fabric route-target during initial creation. An administrator configures an additional global route-target that is shared between fabrics for extended VLAN and VRF network segments. This strategy allows network segments that should not be extended across all fabrics to exist independently and remain a part of a local-only fabric overlay.

For example, if three fabrics had VLAN 20 in their respective overlays, EVPN route-targets (RTs) can be assigned to share VLAN 20 host reachability between two fabrics, but not the third fabric. The following example route-target assignments accomplish this goal.

- Fabric 1, VLAN 20 — **Local RT:** 65001:20, **Global RT:** 1:20
- Fabric 2, VLAN 20 — **Local RT:** 65002:20, **Global RT:** 1:20
- Fabric 3, VLAN 20 — **Local RT:** 65003:20

Global route targets also are assigned to VRFs, when extending routed IP prefixes between fabrics.

Multifabric Data Plane Network

VXLAN tunneling extends Layer 2 and Layer 3 domains across multiple EVPN-VXLAN fabrics. The fabrics can be in different pods of the same data center, in different data centers in the same campus, or in more physically distant data centers. The connection between data center fabrics is referred to as a data center interconnect (DCI).

Layer 2 and Layer 3 network segments are extended between fabrics using the same VNI values across all fabrics. For example, the same Layer 2 VNI value for VLAN 20 and the same Layer 3 VNI for VRF 1 must be the same across all fabrics.

VXLAN tunnels between fabrics are set up only between border leaf switches to maximize both local and multifabric scalability. Establishing a full mesh of tunnels only between border leaf switches eliminates the need to establish VXLAN tunnels between all VTEPs in all fabrics.

The following diagram illustrates inter-fabric and internal fabric VXLAN tunnels in a two fabric topology.

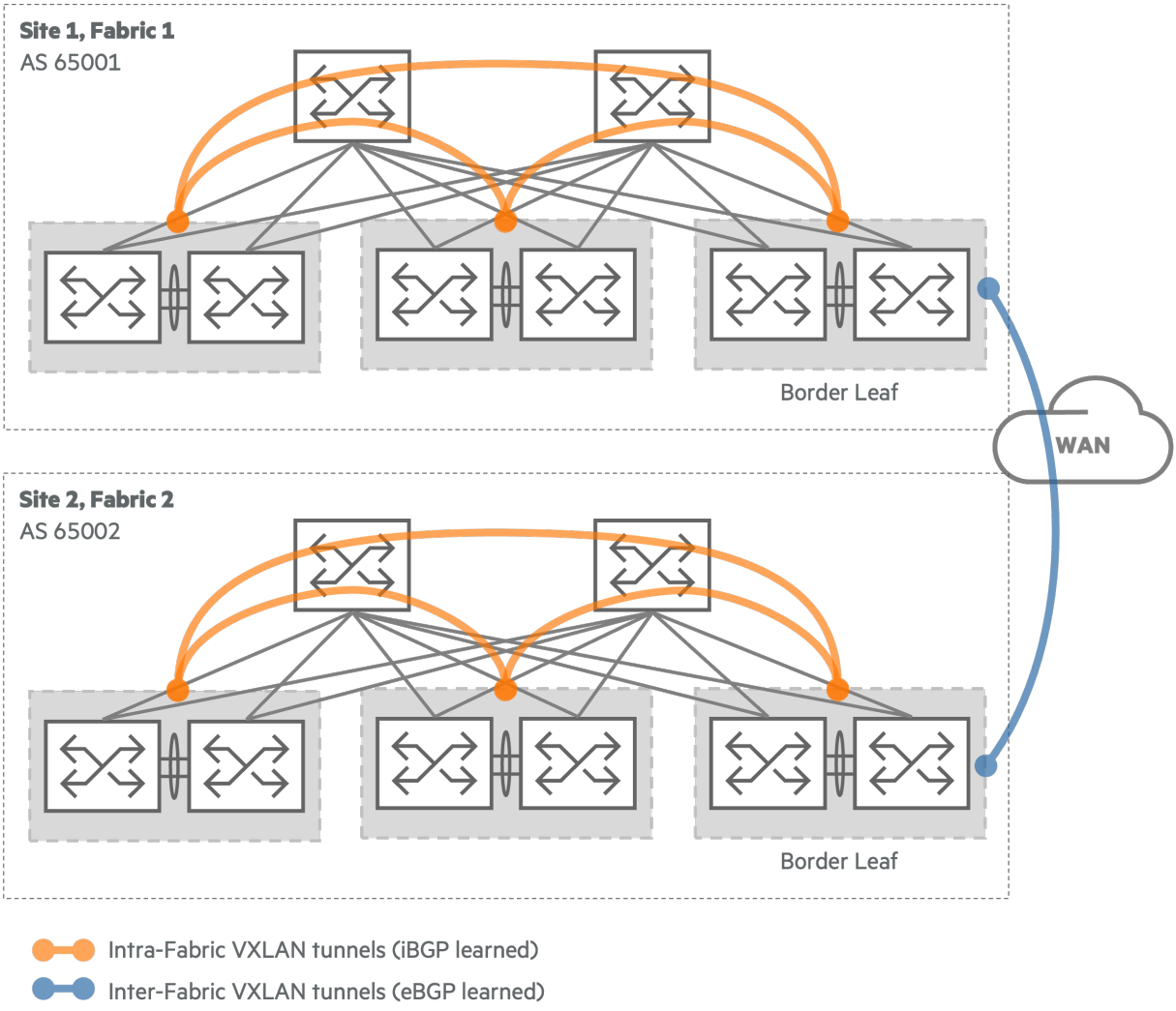


Figure 13: Inter-fabric VXLAN tunnels

VXLAN tunneled traffic between hosts within a local fabric is encapsulated at a single source VTEP and unencapsulated at a single destination VTEP. There is one logical tunnel between any two hosts in a single fabric. A full mesh of VXLAN tunnels between all VTEPs enables this forwarding model.

In a multifabric topology, traffic between hosts in different fabrics can traverse up to three VXLAN tunnels. By default, CX switches do not permit traffic received in a VXLAN tunnel to be forwarded out another VXLAN tunnel. This behavior must be disabled to allow multifabric host reachability. To protect an individual fabric from forwarding loops, VXLAN re-encapsulation can be disabled only between iBGP and eBGP dynamically learned tunnels. Within a fabric, iBGP is used to discover VXLAN VTEPs and dynamically build VXLAN tunnels. eBGP is used to discover VTEPs and establish VXLAN tunnels between fabrics.

When overlay hosts communicate between fabrics, traffic is encapsulated at the source host's directly connected leaf switch with the destination VTEP set as the same fabric's border leaf. The border leaf in the source fabric re-encapsulates the traffic with a destination VTEP of the border leaf in the destination fabric. The border leaf in the destination fabric re-encapsulates the traffic with a VTEP destination of the leaf switch directly connected to the destination host.

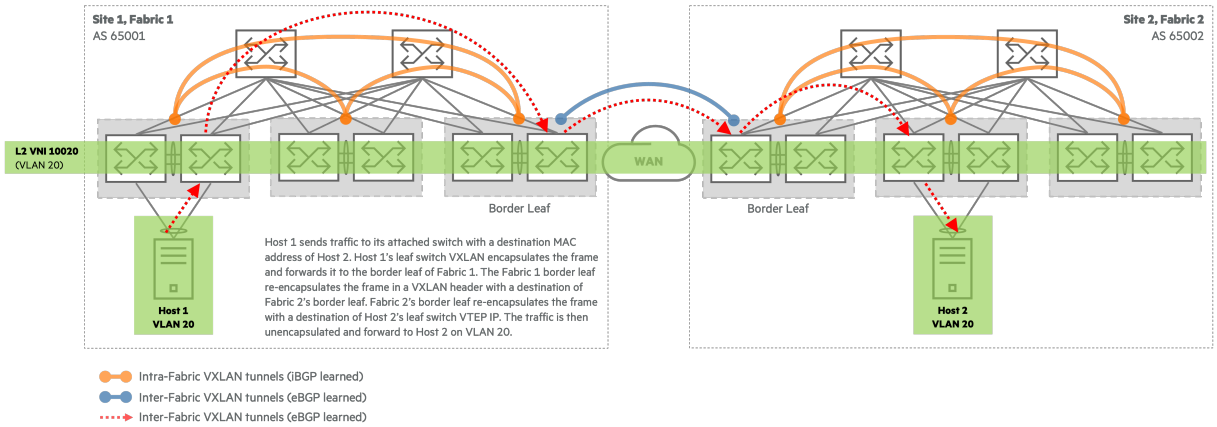


Figure 14: Inter-fabric VXLAN Host Communication

A full mesh of VXLAN tunnels between border leaf switches is established between fabrics in a multifabric topology consisting of three or more fabrics.

Two-Tier Data Center

A Two-Tier design uses traditional protocols, making it simple to deploy, operate, and troubleshoot without the need for specialized knowledge in overlay protocols or design. This architecture is appropriate for medium and small data centers, but can be implemented on a per data center pod basis in a larger environment.

NOTE:

The VSG uses Two-Tier to refer to a topology consisting of Layer 2 multi-chassis LAGs between a collapsed routed/Layer 2 core layer and a Layer 2 only set access switches compared to a spine-and-leaf network using routed links between spine and leaf layers.

Host information in a two-tier data center is populated using traditional bridge learning and ARP methods.

Topology Overview

A Two-Tier data center network implements aggregation and Layer 3 services in a data center core layer, and endpoint connectivity in a Layer 2 access layer. All access switches are Layer 2 connected to both core switches using MC-LAG for load sharing and fault-tolerance.

The physical layout of an L2 two-tier design is consistent with a two-spine spine-and-leaf architecture, which provides a future migration path to an EVPN-VXLAN overlay using a Layer 3 spine-and-leaf underlay, while protecting the investment made in Two-Tier networking equipment.

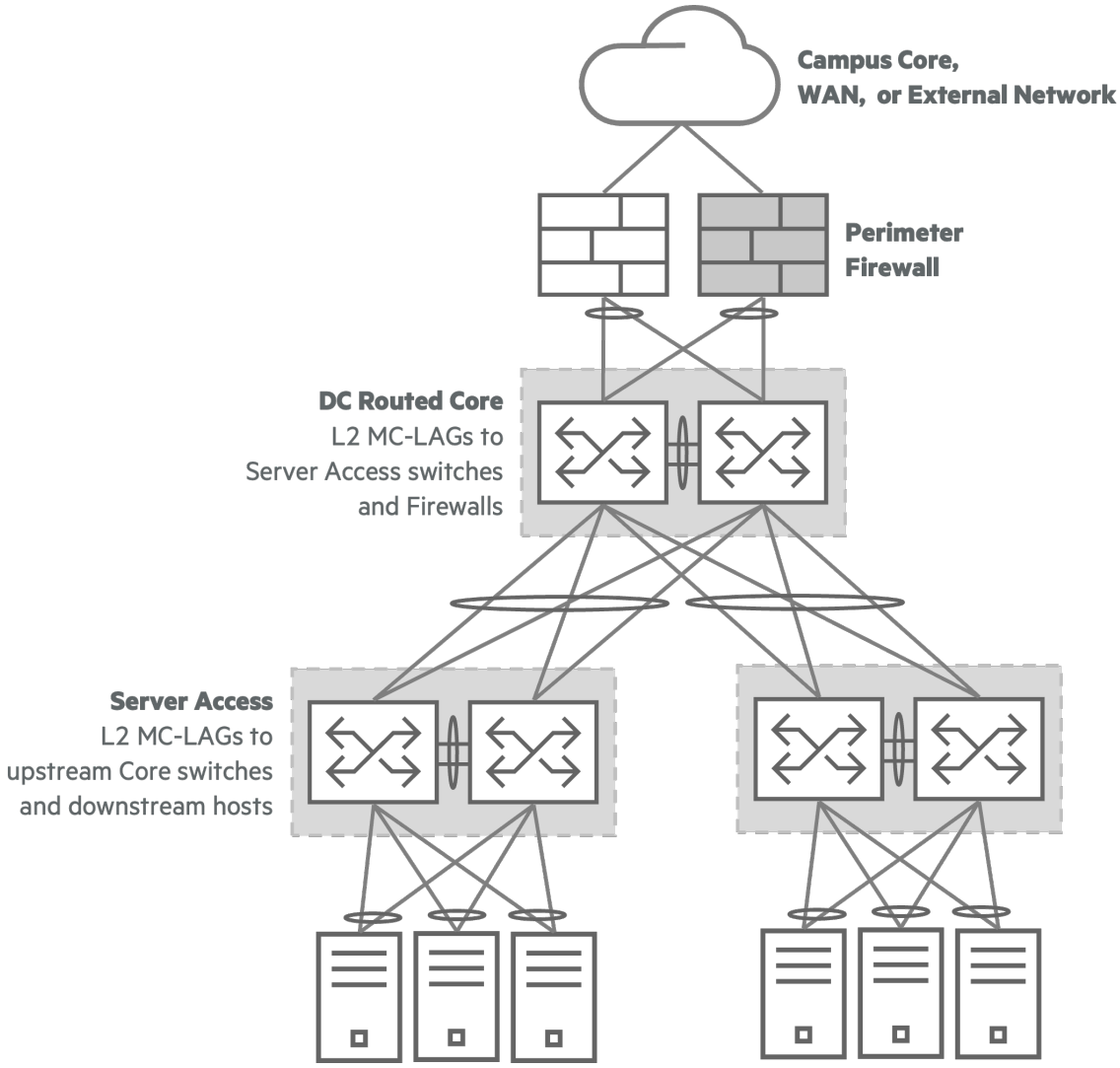


Figure 15: L2 Two-Tier Network Overview Diagram

Core Design

The core layer is deployed as a VSX pair of switches with high-density, high-bandwidth ports. This requires that both core switches are the same switch model running the same firmware version.

The port capacity of the core switches defines the maximum number of racks supported in a Two-Tier architecture. For a redundant ToR design, the maximum number of racks is half the difference of the total port count of the core switch model minus VSX and campus links (ignoring any remainder). For example, a 32-port switch using two VSX links and two campus uplinks can support 14 redundant ToR racks: (32 - 4) / 2 = 14. In a non-redundant, single-switch ToR design, the number of racks supported is equal to the port count of the core switch model minus VSX and campus links.

The core switch model also defines the maximum capacity of the data center backbone. One advantage a routed Layer 3 spine-and-leaf architecture has over a Layer 2-based Two-Tier architecture is incremental expansion of east-west throughput capacity by adding spine switches. For example, adding a spine to a two-spine fabric increases its capacity by 50%, and adding two spines will double its capacity. In an L2 two-tier design, there is a single pair of VSX switches at the core to support rack-to-rack communication. Capacity planning for an L2 two-tier data center is critical, as large capacity upgrades generally require hardware replacement.

Access-to-core connections are generally 40 Gbps or 100 Gbps fiber using quad SFP (QSFP) transceivers or AOCs. When using the CX 9300 in both core and access roles, 400 Gbps access-to-core interconnects are supported for higher speed data center applications.

Increasing initial capacity between the core and access layers can be accomplished by upgrading to higher speed transceivers or by bundling additional links in the MC-LAGs between the core and access layers. However, increasing the links in each LAG significantly reduces the number of racks supported due to increased core port consumption.

Occasionally a subset of racks will require higher capacity links to the core in order to provide high-bandwidth, centralized services. Note that inconsistent uplink capacity between the core and access switches impacts host mobility as VMs requiring increased bandwidth should be attached only to a subset of switches.

The core layer provides a Layer 2 aggregation point for access switches. Traffic between hosts on the same VLAN in different racks will traverse the core layer in VLANs configured on the MC-LAG trunks between the core and access switches. Ubiquitous Layer 2 host mobility within a Two-Tier instance can be achieved by assigning all data center VLANs to all MC-LAG links between the core and access switches using 802.1Q VLAN tagging.

Active Gateway over VSX supports using the same IP address on both core switches and eliminates the need for redundant gateway protocols such as VRRP.

The core layer provides all Layer 3 functions for data center hosts, and it provides the connection point to external networks and services.

Access Switch Design

In a Two-Tier architecture, each ToR access switch is connected to both core switches using MC-LAG to provide link load-balancing and fault tolerance.

Redundant top-of-rack pairs using VSX are recommended to add fault tolerance for downstream hosts using MC-LAG. Although the Layer 2 connectivity between the access and core switches is loop-free through the implementation of MC-LAG and LACP, spanning-tree (STP) is configured as a backup loop avoidance strategy configuration to block accidental loop creation within a rack by a data center administrator. The core VSX pair is configured with an STP priority to ensure its selection as the STP root.

MC-LAG provides better link utilization between the core and access switches over implementing Multiple Spanning-Tree (MST) instances, as all redundant links remain active for sending traffic. Traffic is forwarded on an individual LAG member link selected by a hash-based algorithm applied on a granular per-flow basis. While MST instances can also allow using multiple links between the access and core to provide fault tolerance and balance traffic, the load balancing strategy requires static configuration and limits active forwarding of traffic to a single redundant link on a per-VLAN basis.

Secure Data Center Interconnect

Overview

Organizations implement multiple data centers for a variety of reasons. These may include:

- Provide active redundancy of business-critical applications,
- Provide geographically distant backup for organizational data,
- Accommodate geo-based service delivery,
- Extend the data center to public cloud, private cloud, or colocation infrastructure.

Organizations also may have multiple data centers following mergers or acquisitions.

Data transport between remote data center locations is susceptible to eavesdropping and loss of integrity. Data centers can share data over the public Internet, commercial Layer 2 or Layer 3 transport services, cloud-provider transport services, and leased fiber- or privately-owned fiber plants.

When using the Internet, commercial, or cloud services, data in motion are visible to third parties during transmission. Privately-owned and leased fiber infrastructure are much more secure, but are susceptible to tapping at signal regeneration and termination points.

The risks of losing private data include financial loss, legal action, loss of intellectual property, and loss of reputation.

IPsec tunnels are the industry-standard method for securing data in motion over untrusted infrastructure or third-party networks, including communication between data centers.

Network Address Translation (NAT) services often are combined with IPsec services, when there is a need to accommodate overlapping IP address space. This need often arises when connecting data centers from previously independent organizations following mergers and acquisitions. Use of NAT services is also common in multi-tenant environments such as private cloud and colocation environments.

IPsec

IPsec mitigates the risks associated with transporting data between data center locations by providing:

- Data confidentiality between locations via payload encryption,
- Authentication to verify that both parties in the conversation are known and trusted,
- Data integrity checks to detect and prevent data alteration during transmission.

The simplified diagram below illustrates IPsec Encapsulating Security Payload (ESP) tunnel encryption of a standard IP packet into an IPsec ESP tunnel packet.

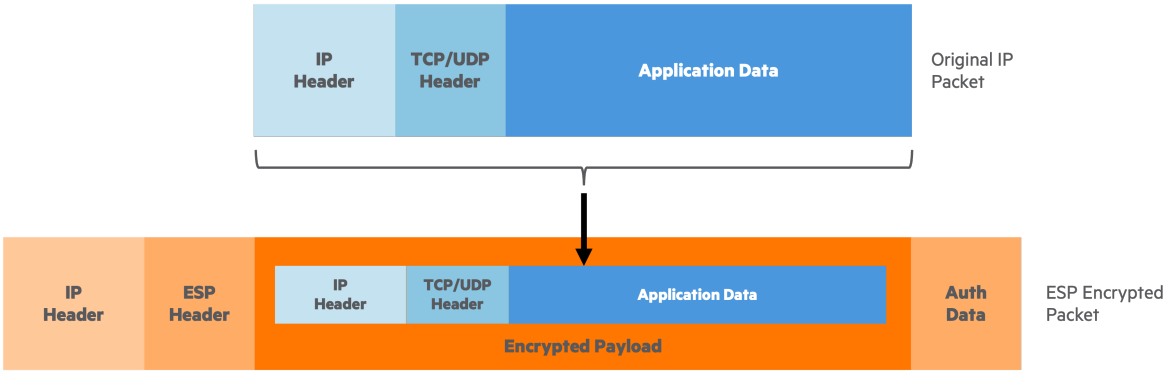


Figure 16: IPsec Site-to-Site Example

The HPE Aruba Networking CX 10000 switch implements the IPsec ESP tunnel protocol in hardware using the on-board AMD Pensando DPU. The CX 10000 also provides NAT services, often combined with IPsec services, when connecting data center locations that must accommodate overlapping IP address space.

Use Cases

Traditional Two-Tier data center topologies, EVPN-VXLAN overlay topologies, and edge data centers can use IPsec to secure communication between locations.

Site-to-Site Transport

When transmitting data using commercial transport services such as metro-Ethernet or MPLS, there is a risk that data can be observed by a third party. Using dedicated fiber substantially reduces risk, but data in motion can still be observed at signal regeneration and termination points outside an organization's direct administrative control, such as at a colocation facility.

The diagram below illustrates the concept of linking two data centers using an IPsec tunnel to protect data. IPsec high-availability and routing concepts are detailed later in this document.

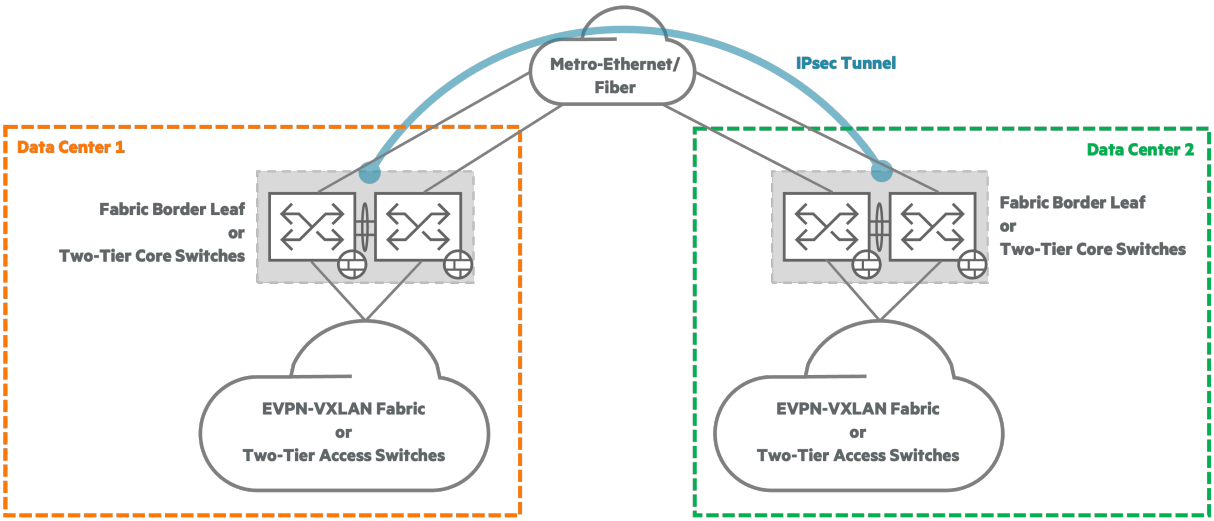


Figure 17: IPsec Site-to-Site Example

NOTE:

IPsec is a vendor-agnostic suite of protocols enabling authentication and encryption between peers in a multi-vendor environment. Deploying CX 10000 switches on both sides of an IPsec ESP tunnel is recommended for effective, efficient support and troubleshooting.

Internet Transport

Establishing connectivity among data centers using commodity Internet links provides economic benefits, but at a much higher risk of data eavesdropping and corruption, making IPsec a critical tool for securing data. IPsec tunneling is a common approach for connecting to public cloud, private cloud, and colocation data centers. SD-WAN appliances with multiple Internet uplinks typically are positioned in the data path to improve underlying network resiliency.

The diagram below illustrates the concept of linking two data centers using an IPsec tunnel over commodity Internet.

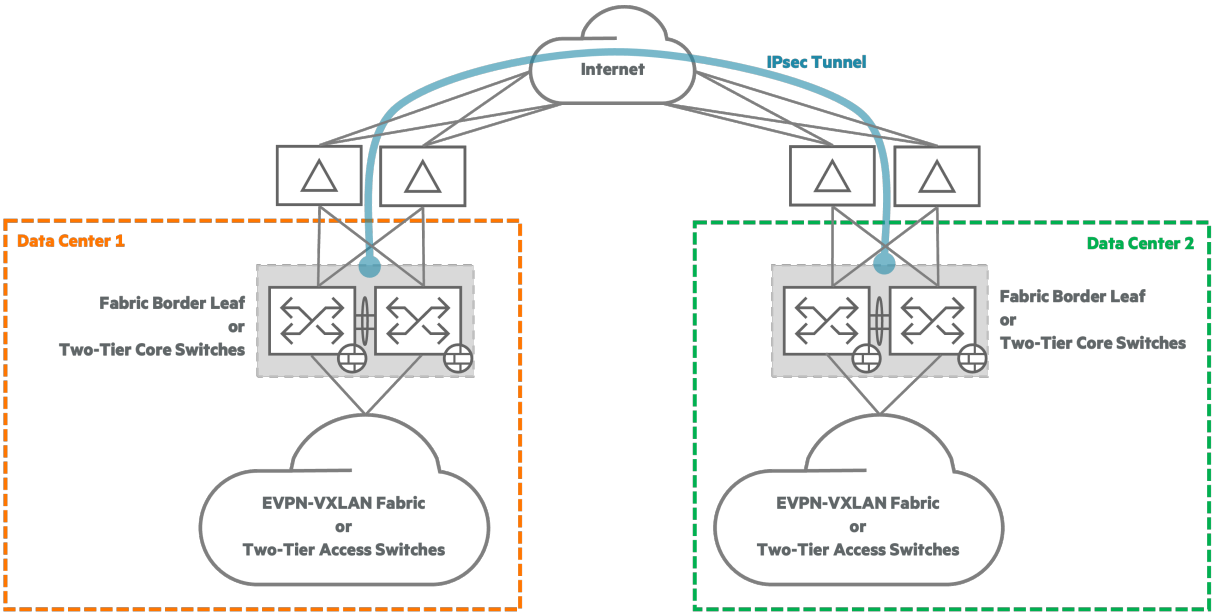


Figure 18: IPsec Over Internet

Dedicated Public Cloud Transport

Infrastructure as a Service (Iaas) public cloud providers offer direct connections into virtual cloud infrastructure. Examples include AWS Direct Connect, Microsoft's Azure ExpressRoute, and Google Cloud's Cloud Interconnect services. In each case, data transmission occurs over infrastructure outside the administrative control of an organization.

IPsec tunnels can be terminated on VPN services supplied by the public cloud provider, or on a more advanced VPN appliance provisioned in the virtual private cloud environment.

CX 10000 IPsec Overview

CX 10000 switches running AOS-CX 10.12.1000 and higher support hardware-accelerated IPsec services. AOS-CX 10.15 adds additional IPsec capabilities, including support of floating static routes to establish backup tunnels and dynamically learned BGP routes.

Each CX 10000 has two AMD Pensando DPUs. IPsec tunnel encryption and decryption are performed on the DPUs. Each DPU supports 135 Gbps full-duplex IPsec operation, for a total of 270 Gbps IPsec performance on a single CX 10000 switch.

When using an EVPN-VXLAN overlay topology, IPsec tunnels are configured on border leaf switches. When using a Two-Tier topology, IPsec tunnels are configured on the data center core switches. In most cases, the switch profile should be configured as **l3-core**.

NOTE:

A CX 10000 terminating IPsec tunnels requires assigning the **l3-core** or **spine** switch profile.

An IP address assigned to a VLAN SVI in a non-default VRF must be configured as the IPsec tunnel source. IPsec termination is not supported in the default VRF.

Traffic destined to an IPsec tunnel on a CX 10000 must be received in the same VRF context in which the tunnel is configured.

When establishing a tunnel, the CX 10000 supports passphrase and certificate-based authentication. For certificate-based authentication, the certificate authority (CA) used to generate the certificate of the remote IPsec host must be installed on the CX 10000.

A CX 10000's physical interfaces are assigned port personas that govern operational behavior for traffic processed by the DPU. When using IPsec, ports facing external networks are assigned the **uplink** persona. Ports facing data center switches and hosts are assigned the **access** persona, which includes spine links in a spine-and-leaf network.

EVPN-VXLAN

When using IPsec in an EVPN-VXLAN data center, the border leaf is dedicated to that function. Extending Layer 2 VNIs and directly connecting hosts to the CX 10000 border leaf terminating IPsec tunnels is not supported.

Inter-VRF route leaking (IVRL) is not supported on the border leaf terminating IPsec tunnels. Each overlay VRF requiring IPsec services must have its own IPsec tunnel configuration. When connecting discrete EVPN-VXLAN fabrics using IPsec, data center connectivity is typically extended on a per-VRF basis using a VRF-lite strategy.

NOTE:

An IPsec services VRF can be created to encrypt host traffic for multiple overlay VRFs by performing IVRL on a services leaf or standard leaf in the fabric that does not terminate IPsec tunnels.

Extending Layer 2 and Layer 3 overlays to establish a multifabric topology is not supported, because the default VRF that carries VXLAN encapsulated traffic does not support IPsec operation. The IPsec tunnel does not support extending Layer 2 reachability between fabrics.

Two-Tier

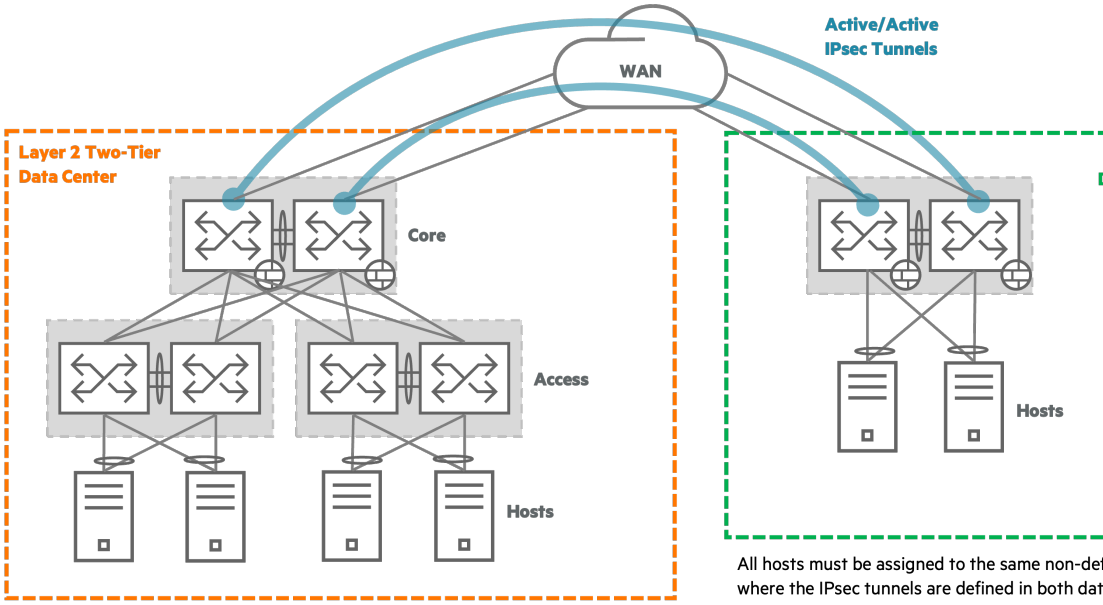
Many traditional Two-Tier data centers use the default VRF for data center host traffic. When using CX 10000 IPsec services, a non-default VRF must be configured for data center connectivity.

IVRL is not supported, when using IPsec in a Two-Tier data center. When multiple VRFs are present, each VRF can be extended to a remote data center by establishing an IPsec tunnel for each VRF using a VRF-lite strategy.

Edge Date Center

Hosts can be directly connected to a CX 10000 edge data center using IPsec to secure communication to a remote data center. All hosts must be connected to a non-default VRF. IVRL is not supported, but multiple VRFs can be extended by configuring an IPsec tunnel per VRF using a VRF-lite strategy.

The diagram below illustrates connecting an edge data center to a traditional Two-Tier data center use redundant IPsec tunnels.



Routing

Traffic routed through an IPsec tunnel is identified using two methods: static routes and BGP peerings. IPsec tunnels do not support Layer 2 bridging traffic to extend broadcast domains or VLANs between locations.

Static routes manually specify traffic by assigning a prefix with a tunnel identifier as the next-hop. As of AOS-CX 10.15, a backup IPsec tunnel is supported using a floating static route, where an assigned administrative distance to the route determines the preferred IPsec tunnel. When the tunnel associated with the preferred route fails, that route is withdrawn, and the backup route is installed into the forwarding table, instructing the switch to send traffic down the backup tunnel.

When using static routes in an EVPN-VXLAN network, the routes are redistributed into the IPv4 or IPv6 BGP address family for the same VRF context as the tunnel. Other VTEPs in the same fabric learn the redistributed route from an EVPN Type 5 advertisement originated by the border leaf.

A Two-Tier data center can use either OSPF or BGP as the routing protocol to external networks. Static routes should be redistributed into the appropriate VRF context of the routing protocol used to establish connectivity outside the data center.

When using AOS-CX 10.15 or later, BGP peerings between IPsec tunnel endpoints may be used to dynamically exchange IP prefixes that should be forwarded in the IPsec tunnel instead of static routes. The CX 10000 supports up to 1,024 learned prefixes per VRF for IPsec redirection. BGP peerings are supported when using standalone or active/active high-availability tunnels.

When using EVPN-VXLAN overlays, host routes should be filtered from the BGP exchange over IPsec. Additionally, the learned prefixes are propagated dynamically to all VTEPs in a fabric as EVPN Type 5 advertisements, without requiring additional redistribution.

When using a traditional Two-Tier topology, redistribution is not required when BGP is used to connect to external networks. When using OSPF, BGP routes must be redistributed into the OSPF instance running in the corresponding VRF context.

When using a VSX pair and route redistribution, both switch members must perform redistribution. If the VSX ISL fails, only the primary VSX member performs redistribution.

When configuring BGP adjacencies and redistribution, route maps are recommended to reduce the risk of errors and unintended network disruption.

IPsec Policy

An IPsec policy is created in Pensando Policy and Services Manager (PSM). The policy defines authentication, encryption, and high-availability parameters for a tunnel, known as a security association (SA). Both local and remote peers of the IPsec tunnel must be configured with parameters such that they can mutually agree on the Internet Key Exchange (IKE) version, an authentication method, and an encryption algorithm for the SA.

IPsec policy is combined with firewall policy, and optional NAT policy. If an explicit firewall policy is not defined, an implicit firewall policy permits all traffic. The diagram below illustrates the processing order of policies, based on the direction of traffic:

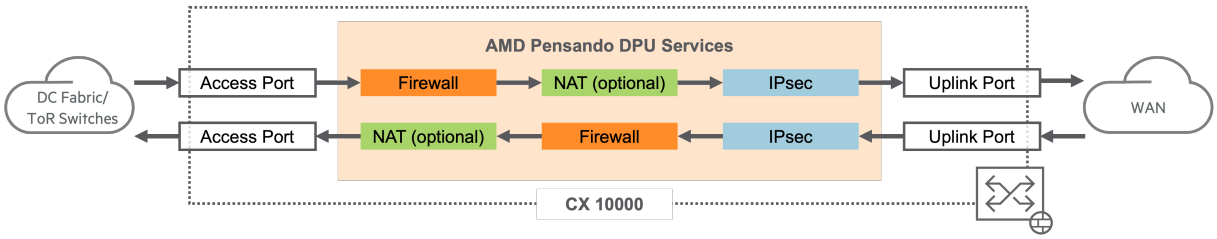


Figure 19: IPsec Policy Order

PSM Policy Distribution Targets (PDTs) provide the flexibility to apply firewall, NAT, and IPsec policy to a subset of CX 10000 switches. PDTs are defined in PSM to scope the application of IPsec related policies to the appropriate set of CX 10000 switches.

HPE Aruba Networking **Fabric Composer** is recommended for configuring PSM policy.

High Availability

In addition to standalone IPsec tunnels, the CX 10000 supports two high-availability modes when using VSX pairs:

- Active/Active
- Active/Passive

Active/Active

Active/active redundancy is the recommended high-availability strategy, where each CX 10000 in a VSX pair establishes an active tunnel to a remote site. It is recommended that the remote site uses a complementary pair of CX 10000 switches. However, the remote termination can be on a single physical or virtual device.

Active/active redundancy provides fast convergence in case of failure, since there is no need to initiate a new IPsec SA before forwarding traffic.

VSX supports flow synchronization between member pairs to share state and support asymmetric traffic forwarding. In case of switch failure, the other VSX member can take over established flows seamlessly. VSX active forwarding enables either member switch to forward decrypted IPsec traffic directly toward downstream data center hosts without consuming the ISL between the VSX pair.

The following diagram illustrates active/active IPsec tunnels between CX 10000 switches at two data center locations.

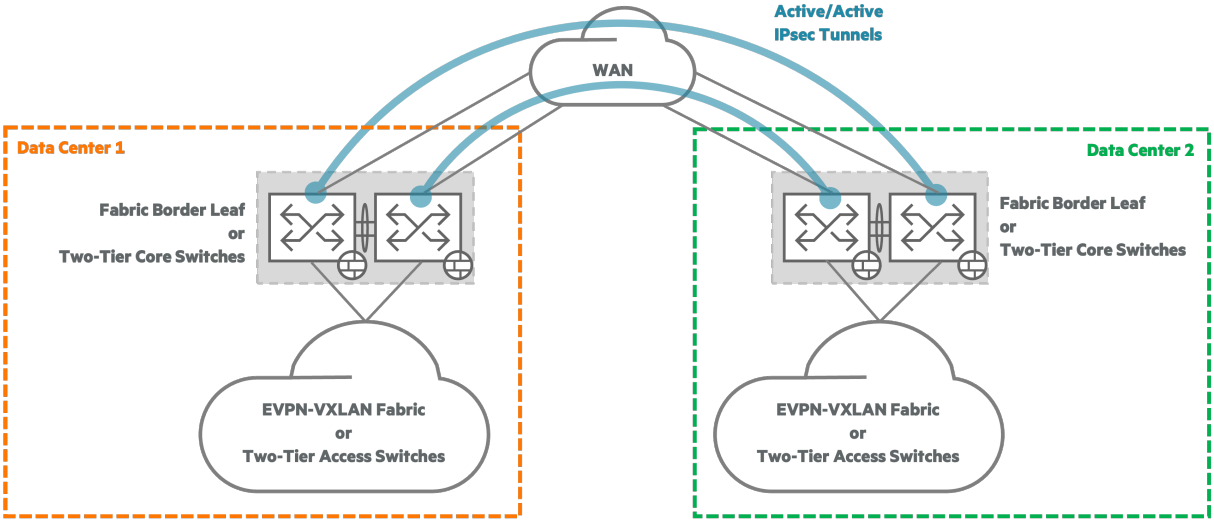


Figure 20: 10K to 10K Active/Active IPsec

The diagram below illustrates active/active IPsec tunnels between CX 10000 switches and a single remote device that supports active/active tunnels. This is common when terminating tunnels to a third-party host, such as a virtual appliance in the public cloud or a physical firewall appliance at a branch data center.

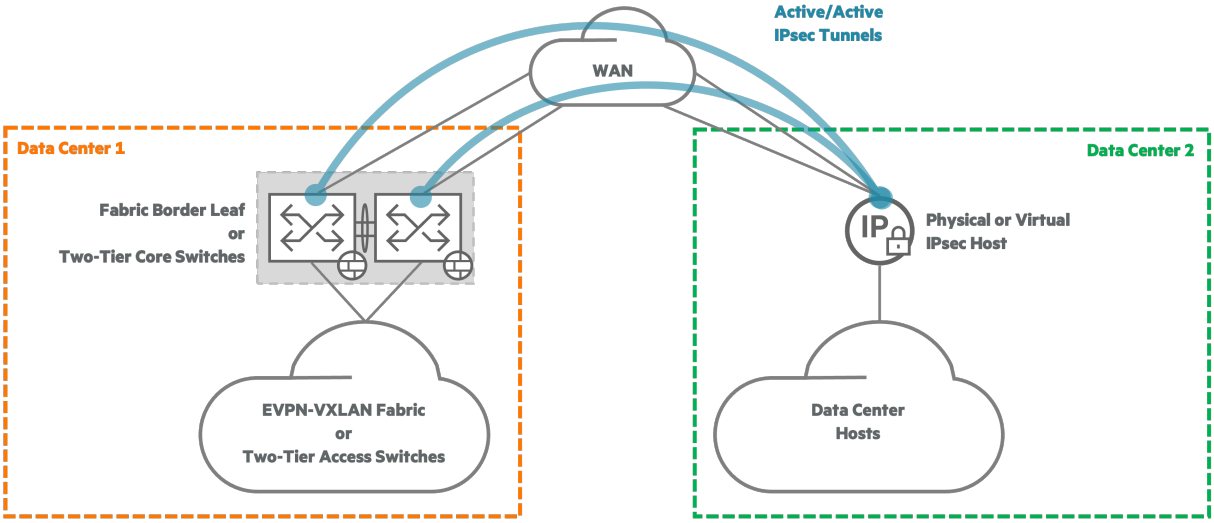


Figure 21: 10K to 3rd-party Host Active/Active IPsec

When IPsec tunnels terminate remotely on a single physical device, the IPsec appliance typically supports physical redundancy using a second appliance in passive mode that can become active if the primary appliance fails. Dual tunnels provide tunnel forwarding diversity for a single device, while physical diversity is achieved using a standby physical device.

Active/Passive

Active/passive high-availability is used when connecting to a third-party device that supports only a single active tunnel. This design has higher convergence times than an active/active design following a network failure, and it should be used only when necessary.

Active/passive operation requires Virtual Router Redundancy Protocol (VRRP) on the CX 10000 VSX pair. The active IPsec tunnel follows ownership of the VRRP Virtual IP (VIP) between the pair. The standby tunnel is associated with the VRRP standby interface. When the VRRP standby interface becomes the active VRRP VIP owner, it establishes the active IPsec tunnel. When using an MC-LAG between the VSX pair and the upstream WAN device, VRRP dual-active forwarding enables either VSX member to route received traffic destined for the VRRP VIP's MAC address.

The diagram below illustrates an active/passive IPsec tunnel topology.

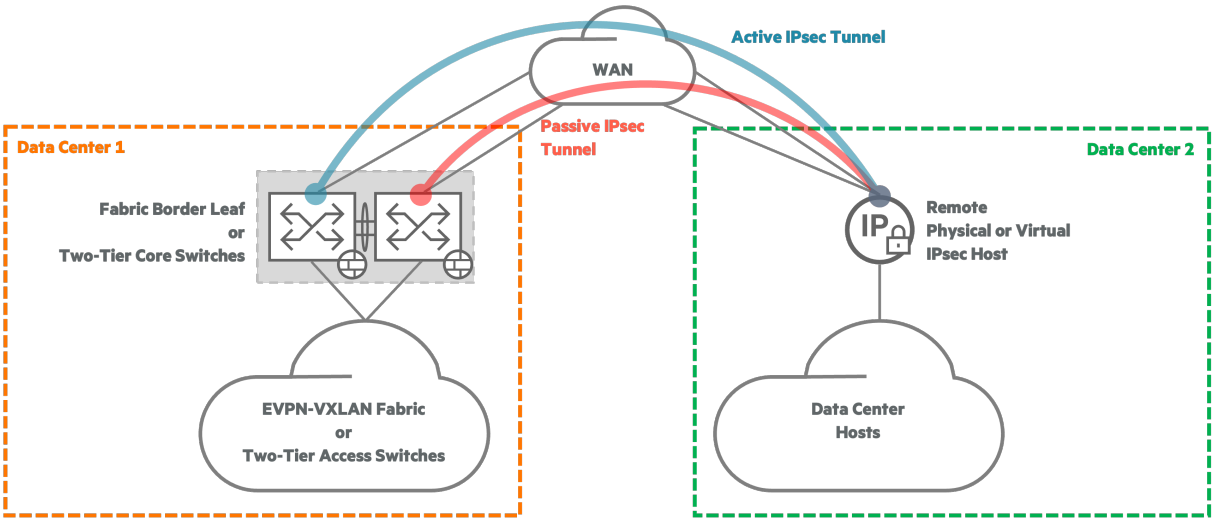


Figure 22: 10K Active/Passive IPsec

Using this strategy, it is common for outbound data center traffic to land on the CX 10000 switch that does not own the active IPsec tunnel. VSX redirects the traffic through the ISL to the CX 10000 with the active IPsec tunnel.

MTU

When IPSec tunnels connect over the Internet, the default Layer 3 IPsec interface MTU of 1500 is recommended. This value sets the TCP maximum segment size (MSS) to 1403 bytes to accommodate the 97-byte overhead of the IPsec headers, which avoids fragmentation of TCP-based traffic. Fragmentation may result in data loss.

When data center interconnections can support a higher MTU value end-to-end, such as when using leased fiber or a metro-Ethernet provider, it is recommended to set the Layer 2 and Layer 3 MTUs to the highest value supported across all devices.

Caveats

Consider the following caveats when using CX 10000 IPsec tunnels as of AOS-CX 10.15. The AOS-CX 10.15 documentation contains a complete list of caveats.

- Multicast is not supported.
- IPv6 addressing is not supported for tunnel endpoints.
- ACLs cannot be applied to IPsec tunnel interfaces.
- Routes exceeding 1024 IPsec routes per VRF are dropped. Ordering preferred routes cannot be guaranteed if more routes are learned than can be supported.
- A floating static route is not supported as a backup path for the same prefixes learned via BGP over an IPsec tunnel.

- BFD for BGP over an IPsec tunnel is not supported.
- Route-based ECMP does not support multiple IPsec tunnel next-hops.

Management

HPE Aruba Networking **Fabric Composer** provides single-pane-of-glass management of IPsec tunnels, as part of its NetOps and SecOps capabilities. Both AOS-CX and PSM components can be configured and monitored using Fabric Composer, including VRFs, interfaces, static routes, BGP peering, IPsec policy, NAT policy, and firewall policy. Fabric Composer safely enables collaborative management across multiple teams by supporting customizable permission assignments using Role-Based Access Control (RBAC).

Data Center Policy Design

The HPE Aruba Networking data center provides powerful policy management options to support multi-tenancy, achieve security goals, and protect critical data.

HPE Aruba Networking Data Center Policy Layer

Implementation of the HPE Aruba Networking data center policy layer is customized for maximum security for the chosen network architecture.

An EVPN-VXLAN design offers a rich combination of overlay technologies and traffic filtering mechanisms to isolate user and application traffic, configured primarily on leaf switches. A Layer 2 two-tier design offers many of the same filtering options, but it requires configuration at both the core and access layers.

The HPE Aruba Networking CX 10000 Distributed Services Switch (DSS) enforces east-west traffic policy using an inline stateful firewall in hardware within the switch. A DSS optimizes performance and traffic flow characteristics over a centralized firewall strategy and can replace hypervisor-based firewalls, increasing hypervisor CPU and memory resources for hosted workloads. The CX 10000 can be placed in both EVPN-VXLAN and Layer 2 two-tier architectures, but greater policy flexibility is achievable with the EVPN-VXLAN design.

NOTE:

The terms **DSS** and **CX 10000** are used interchangeably in this document.

The HPE Aruba Networking Fabric Composer integration with both VMware vCenter and AMD Pensando Policy Services Manager (PSM) provides a powerful combination for managing east-west data center policy using CX 10000 switches and VM guest policy assignment. Network and security administrators can manage all policy elements centrally, while empowering VM administrators to assign VM guests to a policy block in their own independent workflow using VM tags. Fabric Composer also supports centralized configuration of access control lists (ACLs).

This document presents components, design recommendations, and best practices for designing the policy layer for an HPE Aruba Networking data center network.

Out-of-Band Management Network

Organizations should plan to build a physically separate management LAN and role-based access control (RBAC) for network devices.

A separate management network prevents the data plane from becoming an attack surface for potential switch management compromise and ensures that data center switch reachability is not lost when modifying data plane policy.

RBAC requires login authentication against an enterprise directory, typically accomplished using either TACACS+ or RADIUS protocols with a policy server such as HPE Aruba Networking's ClearPass Policy Manager.

Use of logging facilities, log management, and log analysis also should be considered.

Segmentation and Policy Prerequisites

Data center applications are deployed in many different ways. Applications can be implemented as VMs using hypervisors or hosted in bare-metal servers. Containerized apps are highly distributed and usually require connectivity between multiple computing and service nodes. In some cases, a single data center contains applications for multiple tenants while offering a set of shared services across them. Because applications are deployed with the majority of traffic contained in the data center, it is incorrect to assume that all security threats are external.

Successful data center policy design begins by analyzing the requirements of the applications that run in the environment. It is often necessary to re-profile legacy applications when there is insufficient documentation of the requirements. From a networking perspective, application profiling should document all network connections required for an application to run successfully. These might include connections to backend databases or cloud-hosted services. It is important to understand the application profile to define proper policy that identifies which connections must be permitted and which must be denied.

Similarly, analyzing the profile of the users accessing the applications and data is also required. Never leave a data center wide open to a campus, even if it is assumed a secure environment. To restrict access properly, understand the various user profiles associated with the applications and data required. Identify on-campus, remote branch, mobile field workers, and public Internet requirements so that appropriate data center access profiles can be developed to represent their unique requirements.

Segmentation Overview

Segmentation is a logical separation of data center hosts between which a security policy can be enforced. A network segment can represent a large set of hosts such as a data center tenant, or it can be as granular as a single host.

VRF Segmentation

Establishing routing domains is a key method of segmenting internal data center communication and a requirement for an EVPN-VXLAN overlay. Network hosts that are members of one routing domain are not capable of communicating with hosts in another routing domain by default. Each routing domain contains a set of hosts allowed to communicate with one another, while traffic into or out of the routing domain can be controlled. Multiple strategies are available to share connectivity between routing domains, when necessary.

A switch supports multiple routing domains by implementing virtual routing and forwarding instances (VRFs). Each VRF consists of a unique route table, member interfaces that forward traffic based on the route table, and routing protocols that build the route table. Different VRFs may contain overlapping IP address ranges because the individual route tables are discrete.

A VRF instance can correlate with a customer, an application, a set of hosts with common security requirements (i.e., PCI), or a set of hosts with other common characteristics (i.e., production environment, development environment, etc.). This correlation enables using VRF instances as security domains by applying security policy on a network device that connects VRFs together, such as a firewall. VRF member interfaces that connect to external networks provide a natural point for implementing north-south security policy.

In an EVPN-VXLAN design, VRFs are a required construct in the overlay and are implemented on leaf switches. In a multi-fabric environment, EVPN-VXLAN allows VRF segments to be extended to additional data center locations. The default VRF is reserved for underlay connectivity.

In a Layer 2 two-tier design, VRFs are optional and are implemented at the data center core. Data center implementations large enough to require VRF segmentation are generally implemented with an EVPN-VXLAN solution, but VRFs remain a useful segmentation strategy in a Layer 2 two-tier design. By default, all hosts in a Layer 2 two-tier data center are members of the default VRF.

VRFs should be added in a thoughtful manner, because the operational complexity of a network increases as the number of VRFs grows. Minimizing complexity results in a network implementation that is easier to maintain and troubleshoot. Each organization should define its own policy that clearly states the criteria for adding a VRF to the network. For example, a service provider that supports multiple tenants has different criteria and requires more VRFs than a university data center.

Common best practice is to use the minimum number of VRFs required to achieve clearly defined organizational goals. VRFs are employed to support:

- Separate production and development application environments. This provides a development sandbox while minimizing risk to production application uptime, and it supports overlapping IP space when required.
- Applying policy to segmented traffic that requires strict regulatory compliance, such as PCI or HIPAA.
- Applying policy to traffic from hosts identified by organizational policy as requiring segmentation and possessing a common set of security requirements. These sets of hosts often share a common administrative domain.
- Isolating Layer 3 route reachability in a multi-tenancy data center, while supporting overlapping IP space.

Inter-VRF Route Forwarding (IVRF) can be used within a data center to share IP prefixes between VRFs. For example, to provide shared services in a data center, a services VRF can be created to offer a common set of resources to some or all other data center VRFs. IVRF allows Layer 3 reachability between applications in the services VRF and hosts in other VRFs.

NOTE:

IVRF can circumvent inter-VRF policy by joining previously discrete routing domains together, and it does not support overlapping IP address space.

VLAN/Subnet Segmentation

In addition to limiting broadcast domain size, a VLAN can be used to group sets of data center hosts by role, application, and administrative domain. Typically, a VLAN is associated with a single IP subnet. Traffic between VLANs must be routed, and all host traffic between VLANs is forwarded via an IP gateway interface, where security policy can be applied.

ACLs applied to Layer 3 VLAN interfaces generally are used to enforce a base policy between subnets. For more sophisticated policy requirements, a common solution is to deploy a centralized firewall and make it the default gateway. This results in a sub-optimal, inefficient traffic pattern. Routed traffic between VLANs is hairpinned at the central firewall, unnecessarily consuming extra data center bandwidth.

HPE Aruba Networking provides a more elegant data center policy option by deploying DSS (CX 10000) ToR switches which offer hardware-based, Layer 4 firewall capabilities at the host's connection layer. This model optimizes data center bandwidth capacity and eliminates the need to hairpin traffic through a central firewall.

Microsegmentation

Microsegmentation extends Layer 2 VLAN segmentation to an individual workload level using isolated and primary VLAN constructs available in the private VLAN (PVLAN) feature set. Similar to a hypervisor-based firewall, a PVLAN microsegmentation strategy provides the ability to enforce policy between VM guests on the same hypervisor. This same strategy can enforce policy between containers.

Private VLANs coupled with a DSS support microsegmentation policy enforcement local to the workload's attachment point. Data center workloads assigned to the same isolated private VLAN cannot communicate directly with each other over Layer 2. The isolated VLAN is associated to a primary VLAN on a CX 10000 which hosts a default gateway SVI for each isolated VLAN. Proxy-ARP is configured on the primary VLAN's SVI to enable communication between isolated hosts via the primary VLAN SVI.

A centralized firewall also can be used to achieve microsegmentation, but rapid growth of microsegmented workloads combined with the inefficient traffic engineering of a centralized design will exhaust overall data center bandwidth much more quickly as traffic to and from each microsegmented workload is required to be hairpinned centrally.

An EVPN-VXLAN solution using DSS leaf switches provides the most flexible policy assignment. Stateful firewall policy is assigned to the primary VLAN associated with the segmented workload's isolated VLAN. Since all traffic for a workload assigned to an isolated VLAN is forwarded to the primary VLAN on the DSS, all traffic for the individual workload is subject to policy enforcement. Both egress and ingress firewall policy can be applied to the workload.

In a Layer 2 two-tier solution using DSS access switches, firewall policy is limited to only the egress direction applied to the workload's isolated PVLAN. Policy is enforced when traffic traverses the DSS access layer toward the primary VLAN's gateway IP defined at the core.

The CX 10000 provides a unified data center microsegmentation strategy that is hypervisor agnostic (supporting VMware, Microsoft Hyper-V, KVM, etc.) and also supports bare metal servers. Using the CX 10000 in place of a hypervisor-based implementation offloads policy enforcement cycles from a VM host CPU to dedicated switch hardware.

In both the EVPN-VXLAN and Layer 2 two-tier solutions, ACL policy can be applied to the primary VLAN's SVI interface, and PVLANs can be extended across multiple switches.

Microsegmentation can be applied to a subset of hosts requiring a high level of scrutiny, or it can be applied more broadly to maximize a data center's security posture.

Policy Overview

Security policy specifies the type of traffic allowed between network segments. Network-based policy is typically enforced using a firewall or ACL. If traffic is permitted by a stateful firewall, dynamic state is created to permit return traffic for the session. An ACL is applied to traffic only in one direction with no dynamic state created.

Applying network security policy plays a significant role in reducing the attack surface exposed by data center hosts. Blocking unnecessary protocols reduces the available tactics a threat actor can use in host exploitation. Scoping allowed outbound traffic inhibits command and control structures and blocks common methods of data exfiltration. Applying intra-data center security policy constrains the options for lateral threat movement if a host has been compromised.

Data Center Perimeter Policy

Data center routes require sharing with campus and other external networks. Applying policies at the edge between the data center and external networks is the first layer of security for data center applications. They limit access to only permitted networks and hosts while monitoring those connections, and they can be implemented using perimeter firewall appliances or ACLs.

In an EVPN-VXLAN spine-and-leaf design, a pair of leaf switches is the single entry and exit point to the data center. This *border leaf* is not required to be dedicated to that function. Border and services leaf functions are commonly combined, and less frequently computing hosts are also attached. In a Layer 2 two-tier design, the core layer provides a common ingress and egress point for traffic between the data center and external networks. In both cases, this data center network edge is where a set of policies is implemented to control access into and out of the data center network.

Perimeter policy is applied at a VRF segment level. If a single VRF contains all data center hosts, a single pair of policies is configured at the data center edge: one policy for ingress traffic and a second policy for egress traffic. If multiple VRFs are configured, a unique pair of policies should be implemented on a per-VRF basis at the data center edge.

Multiple data center VRFs can be extended to the upstream edge device on a single physical interface or aggregated link using 802.1Q tagging. A VLAN is associated with each VRF, and the VLAN's corresponding SVI is used for router protocol peering purposes. The upstream edge device may have one or multiple VRFs defined. A direct VRF-to-VRF peering between a data center edge VRF and its campus VRF neighbor enables IP segmentation to be extended into the campus, which is referred to as VRF-lite.

In addition to filtering traffic between the data center and external networks, multiple data center VRFs can peer with a single external routing instance to create a combined enforcement point for external and inter-VRF policy, when overlapping IP address space is not implemented. Policy between data center VRFs is enforced by hairpinning traffic to the upstream device.

Perimeter Firewalls

Dedicated security systems at the perimeter can offer advanced monitoring, application-aware policy enforcement, and threat detection.

Perimeter firewalls are deployed in transparent or routed mode. In transparent mode, the firewalls behave like a bump in the wire, meaning they do not participate in Layer 3 network routing. From the perspective of directly attached switches, they are no different than a transparent bridge, but the firewall forwards only explicitly permitted traffic. In routed mode, a firewall participates in the routing control plane and generally has more flexibility with deep packet inspection and policy enforcement options. It is important to note that stateful firewalls require symmetric forwarding to apply policy correctly to subsequent traffic in a session.

When multiple data center VRFs contain overlapping IP address space or VRF segmentation must be extended beyond the perimeter firewall, the firewall must support a virtualization mechanism to allow route table isolation. This can be virtualization of the firewall itself into distinct logical instances or support for VRFs.

Perimeter ACLs

When IP subnets inside the data center are designed to map to security groups or business functions, Access Control Lists (ACLs) at the border leaf can provide policy enforcement from user locations into data center applications. If subnets cannot be mapped to security groups, ACLs can become difficult to manage and scale in larger environments. The primary benefit of perimeter ACLs is that they can be implemented directly on the switching infrastructure to enforce a policy foundation from which to establish data center access. Policies implemented using switch ACLs specifically target Layer 3 and Layer 4 constructs. Switch ACLs are not stateful or application-aware.

East-West Security Policy

The majority of traffic in a modern data center is east-west traffic between the data center workloads themselves. Policy enforcement can be implemented between VRF, VLAN, and microsegmentation segments using firewalls or ACLs.

Firewalls offer more comprehensive filtering capabilities, when compared to ACLs. Firewall policy inside an HPE Aruba Networking data center can be implemented using two methods at the network layer: inline using distributed services switches (DSSs) or centrally using a firewall appliance in a services leaf.

Distributed Services Switch Policy Enforcement

The AMD Pensando programmable data processing unit (DPU) extends CX 10000 switches to include stateful firewall capabilities. Using this built-in hardware feature, firewall enforcement is provided inline as part of the switch data plane.

There are several advantages to this approach. Data paths are optimized by applying policy at the workload attachment point, without the requirement to hairpin data through a centralized firewall. Firewall policy can be granular to the host with support for microsegmentation. The Pensando DPU provides wire-rate performance that can alleviate resource consumption on hypervisor-based firewall services processing large data flows by moving firewall services to dedicated switch hardware.

CX 10000 switches are deployed as leaf switches in an EVPN-VXLAN solution and as access switches in a Layer 2 two-tier solution.

Central Firewall Policy Enforcement

Another commonly deployed policy enforcement approach is placing a firewall appliance in a central location that provides IP gateway services to data center hosts.

In an EVPN-VXLAN solution, central firewalls are connected to a services leaf and are Layer 2 adjacent to fabric hosts using the overlay network. The default gateway for hosts requiring policy enforcement is moved from the ToR to the centralized firewall. Similar to a border leaf, the services leaf is not required to be dedicated to this function. One advantage of this approach is the ease with which a Layer 2 overlay network can be used to transport host traffic to the firewall.

In a Layer 2 two-tier network, central firewalls are connected to the core switches, and the default gateway for hosts is moved from the core switches to the central firewalls.

There are several disadvantages of using a centralized firewall for both EVPN-VXLAN and Layer 2 two-tier topologies. Policy enforcement can only be applied between hosts in different subnets. A centralized firewall requires multiple switch hops from the data center host to enforce policy, which mitigates the efficiency of data path delivery in an EVPN-VXLAN model. Policy enforcement between two hosts attached to the same switch must be forwarded to the central firewall. Hairpinning a large volume of east-west traffic through a central point can create a bottleneck and reduces east-west data center bandwidth capacity. Microsegmentation can be supported centrally, but is not recommended as it requires all policy-driven data center traffic to traverse a single point in the network, significantly increasing the risk of a bottleneck.

The diagram below illustrates the inefficient traffic hairpin, when using a services leaf firewall.

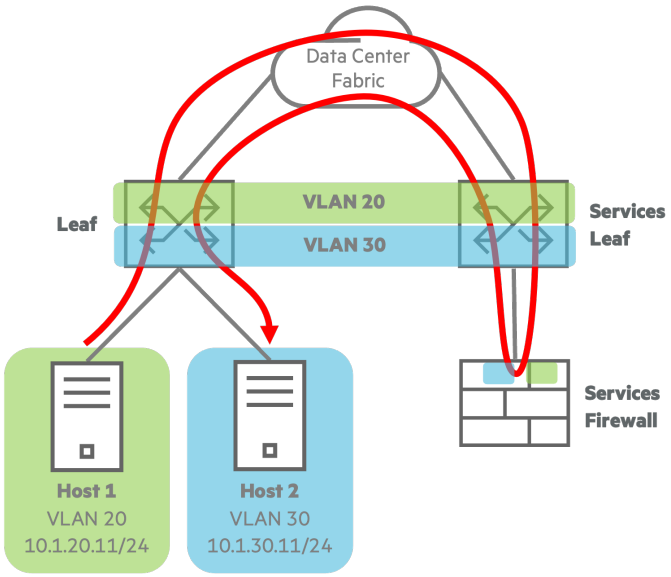


Figure 23: PSM Policy Direction Diagram

Hypervisor-Based Firewall Enforcement

Some vendors offer virtualized firewall services within a hypervisor environment. This approach can provide granular, service-level policy enforcement while allowing for the use of active gateways. VMware NSX is an example of a product that can integrate in this way.

Virtualized firewalls can consume a large volume of CPU resources, reducing CPU resources available for compute processing in VM infrastructure. The CX 10000 alleviates this pressure by moving firewall inspection to dedicated hardware on switch infrastructure that can support microsegmentation between VMs.

Applying DSS Policy

The HPE Aruba Networking CX 10000 switch with AMD Pensando delivers a powerful policy enforcement engine, referred to as a Distributed Services Switch (DSS). This section provides background information and details on how to implement DSS firewall policy.

The Pensando Policy and Services Manager (PSM) application defines policy and associated elements that are pushed to the AMD Pensando DPU. HPE Aruba Networking Fabric Composer can be used to orchestrate policy via PSM's API.

PSM Policy Foundations

PSM policy can be assigned to two different object types: *Network* and *VRF*. PSM associates a *Network* object with a VLAN configured on a DSS switch. Defining a PSM *Network* informs the switch to redirect traffic for the associated VLAN to the on-board, AMD Pensando DPU-based firewall for Layer 4 policy enforcement. Policy assigned to a *Network* is applied only to traffic in the VLAN associated with the *Network* object. PSM associates a *VRF* object with a VRF configured on a DSS switch.

Policy assigned to a *VRF* object applies to all VLANs in the associated VRF that also are associated with a *Network* object. *VRF* policy enforcement does not require that a policy is assigned to a *Network* object, but it does require that a *Network* object exists for each VLAN in the associated VRF to redirect traffic to the Pensando DPU. VLANs in a VRF with assigned PSM policy that do not have a corresponding *Network* object defined do not forward traffic to the Pensando DPU.

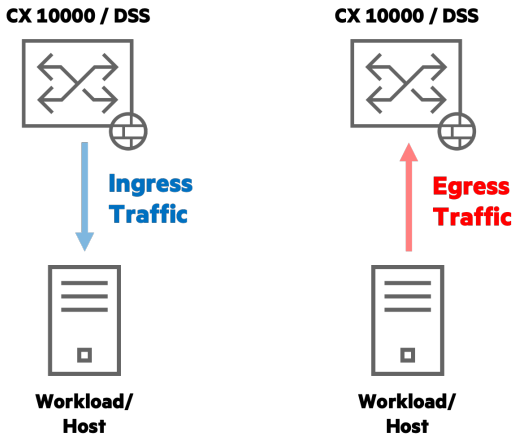
CAUTION:

Define a *Network* object for each VLAN within a VRF, if any VLAN or the VRF requires policy enforcement. Network communication failures may result, if only a subset of VLANs in a VRF have corresponding *Network* objects defined.

When traffic is redirected to the Pensando DPU for firewall enforcement, both *VRF* and *Network* policies are enforced. PSM evaluates both policy types using a logical AND function. The policies are not concatenated or evaluated sequentially. If both policies permit the traffic, the traffic is forwarded. If either policy denies the traffic, it is dropped. If a policy is not assigned at one level, policy is enforced by the other level's policy. If no policy is assigned at either level, traffic is permitted. The table below summarizes when traffic is permitted or denied.

Network Policy	VRF Policy	Result
Permit	Permit	Permit
Deny	Permit	Deny
Permit	Deny	Deny
Deny	Deny	Deny
Permit	No Policy	Permit
Deny	No Policy	Deny
No Policy	Permit	Permit
No Policy	Deny	Deny
No Policy	No Policy	Permit

PSM firewall policy is a set of rules that specify source and destination addresses, and the type of traffic allowed between them using IP protocol and port numbers. PSM policy is assigned to a *Network* or *VRF* in either an ingress or egress direction, from the perspective of the connected host. Traffic destined to a directly attached host is considered ingress, and traffic sourced from a directly attached host is considered egress. This ingress/egress relationship is the reverse of applying switch ACLs, from the perspective of the switch's network interface.



PSM Ingress Policy

Ingress policy applies to traffic destined to a host in a VLAN with an associated PSM Network, with the exception of traffic between two hosts in the same VLAN attached to the same CX 10000. Ingress policy is generally applied to filter inbound traffic destined to an application server.

Ingress policy applies to traffic routed between hosts attached to the same CX 10000, when running AOS-CX 10.10.1000 and above. Previous versions of AOS-CX are constrained to applying egress policy between hosts attached to the same switch for both routed and bridged traffic.

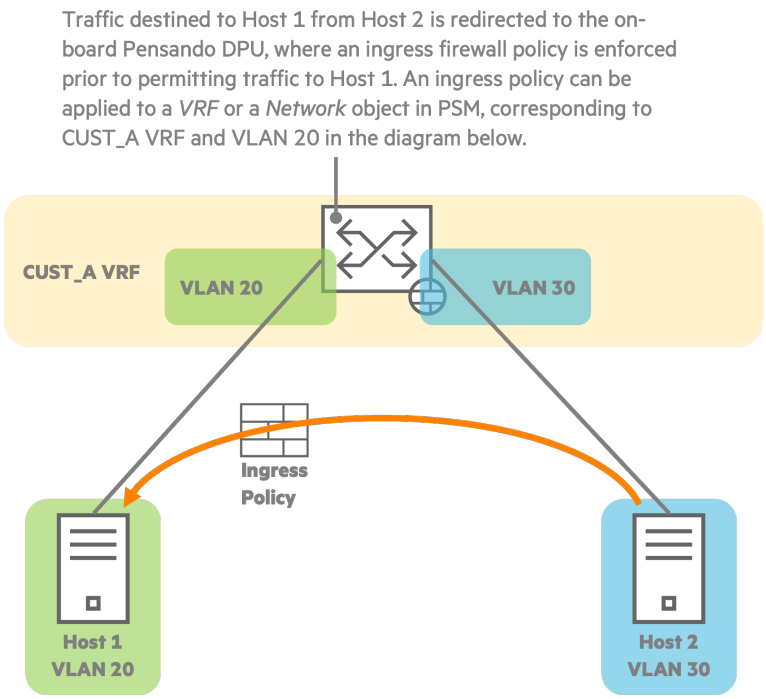


Figure 24: Ingress Intra-Switch Routed Policy Enforcement Diagram

Ingress policy is not applied to traditional Layer 2 bridged traffic between hosts on the same switch over a standard VLAN, and thus cannot be applied in a Layer 2 two-tier topology using CX 10000 access switches. When using EVPN-VXLAN, a microsegmentation strategy using private VLANs and proxy ARP can be used to apply ingress policy between hosts in the same isolated VLAN on the same switch, described below under *Microsegmentation*.

When using an EVPN-VXLAN fabric overlay, ingress policy applies to all VXLAN-forwarded traffic, both routed traffic and Layer 2 traffic between hosts in the same VLAN connected to different VTEPs. After hitting the VTEP interface on the destination switch, the traffic is forwarded to the Pensando DPU for evaluation.

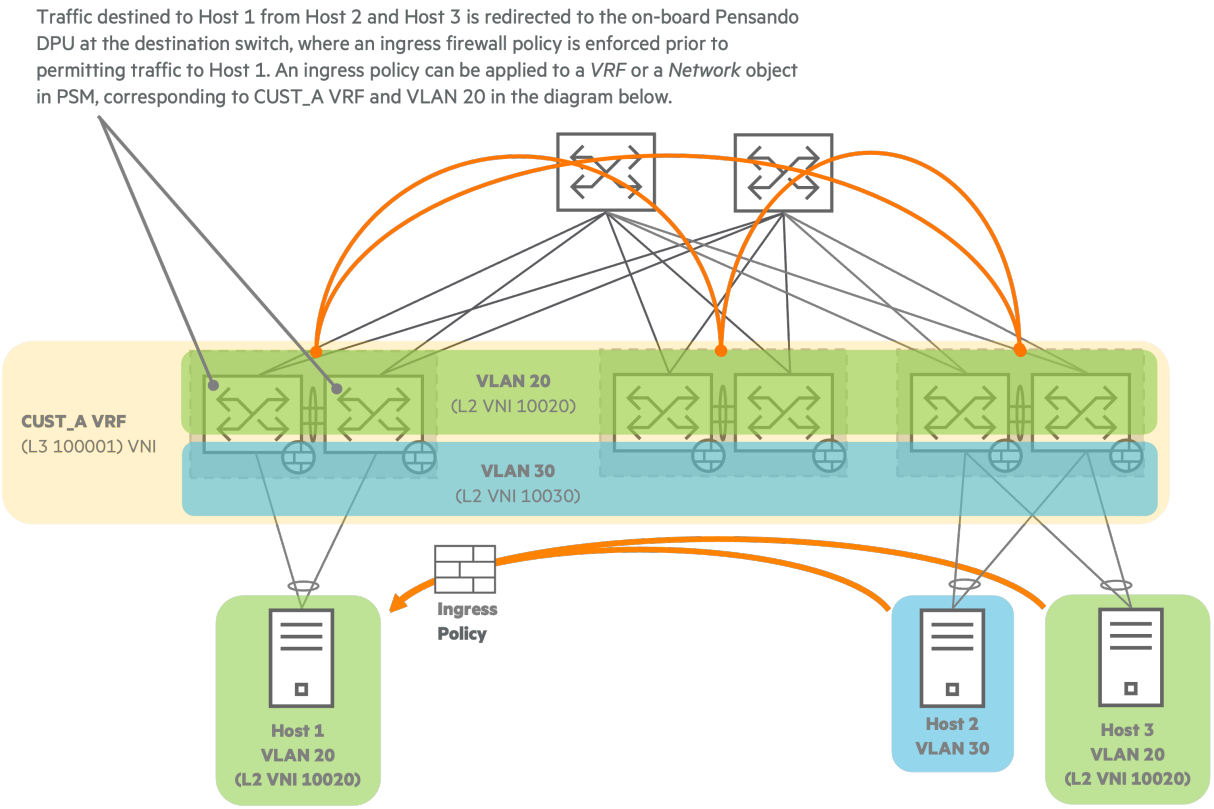


Figure 25: Ingress Layer 2 Fabric Policy Enforcement Diagram

PSM Egress Policy

Egress policy defines allowed traffic initiated by hosts directly attached to a DSS switch. Defining egress policy can protect the data center from lateral movement by a compromised host and prevent data exfiltration. For example, backend database servers may only require initiating communication with peer database servers for synchronization, DNS, NTP, authentication services, and with a local update server. This communication can be scoped by using Layer 4 filters to allow only the desired traffic type between hosts.

Egress policy is applied to all traffic sourced by hosts in an inspected VLAN, irrespective of where the destination hosts resides. Unlike ingress policy, egress policy filters Layer 2 bridged traffic between hosts in the same VLAN on the same switch. Egress policy is applied on both EVPN-VXLAN leaf switches and Layer 2 two-tier access switches to enforce inter-VLAN and microsegmentation policy.

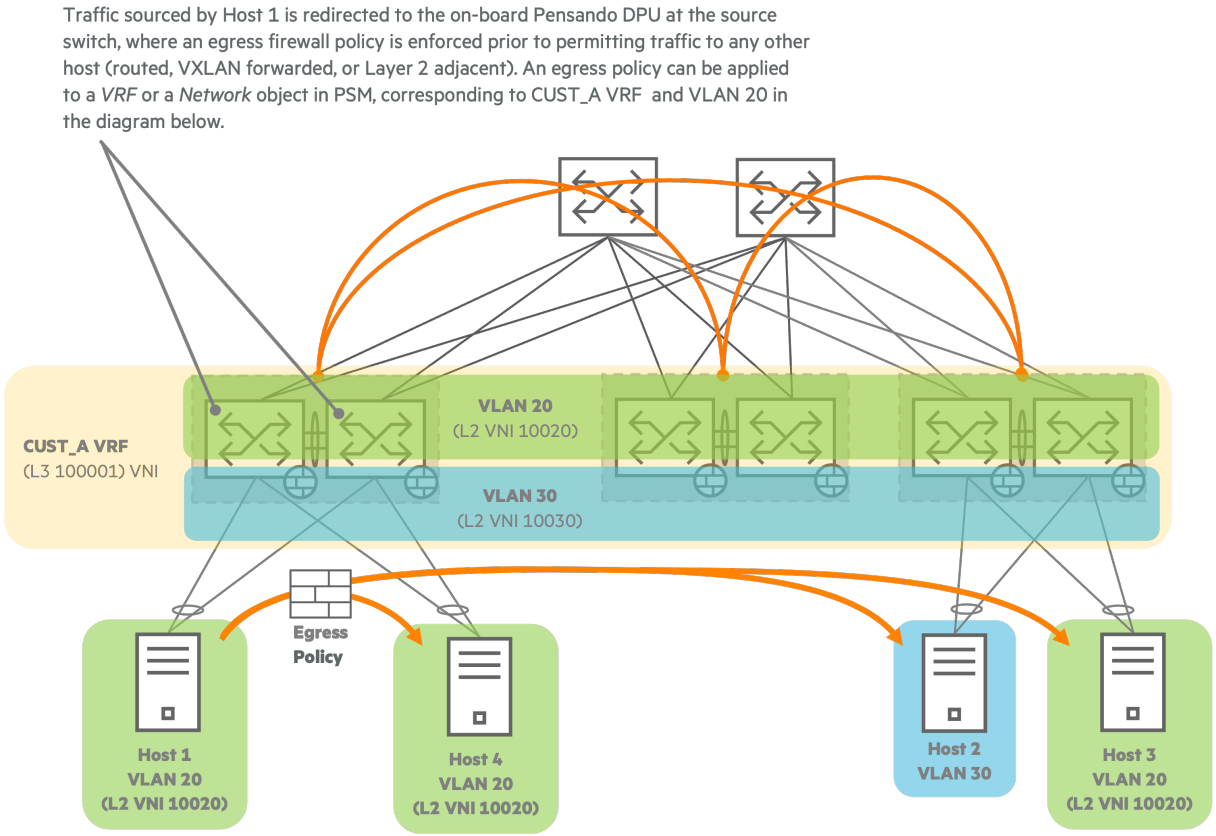


Figure 26: Egress Policy Enforcement Diagram

Communication between VMs or containers residing on the same physical host will not be forwarded to the DSS switch for policy enforcement by default. Policy between these workloads can be enforced using a PVLAN microsegmentation strategy, described below under *Microsegmentation*

CX 10000 Mixed Environment Considerations

Applying a consistent PSM-based policy across a data center fabric is achieved more easily when all leaf or access switches are CX 10000s. This supports uniform policy enforcement without the need to manage exceptions. A mixed environment of DSS- and non-DSS switches is supported, but requires additional planning.

Any VLAN redirected on one CX 10000 for policy inspection should be populated on all on DSS switches in the same fabric. Do not configure VLANs redirected for policy on DSS switches on non-DSS switches in a mixed environment to ensure proper policy enforcement and traffic flow. VLANs not redirected for policy can be populated on both DSS and non-DSS switches.

Ubiquitous host mobility within a fabric requires that all leaf and access switches support the same capabilities, which is not possible in a mixed environment. DSS stateful firewall security policies are not available on non-DSS switches, so VM mobility must be constrained in a network with a mix of DSS and non-DSS switches. For example, when using dynamic tools such as VMware's Distributed Resource

Scheduler (DRS), ensure that virtual switch and port group resources are defined to prevent automated movement of a VM guest requiring firewall services to a VM hypervisor host that is not connected to a DSS switch. Maintaining separate distinct distributed virtual switches on hypervisors is an effective method to constrain this movement.

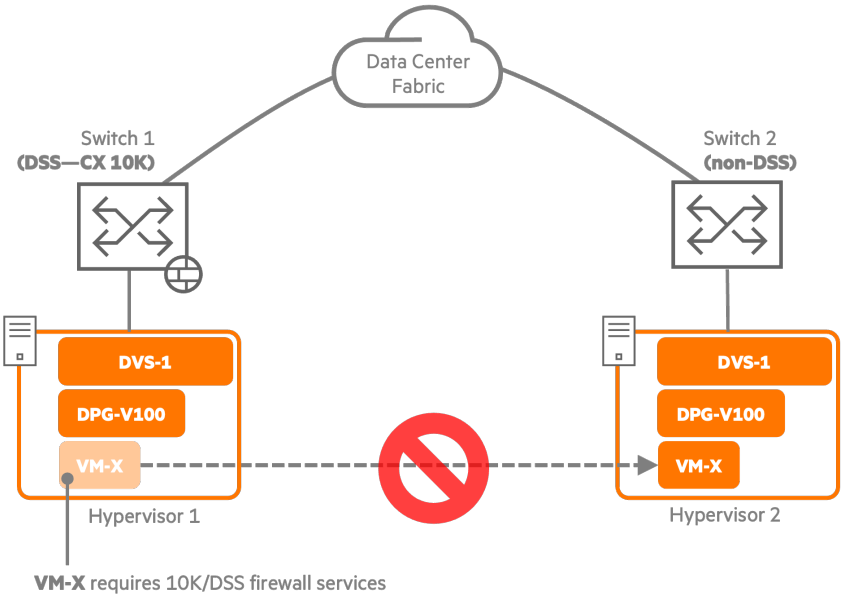


Figure 27: DSS Host Mobility Constraint Diagram

Workload Microsegmentation

Microsegmentation enables the enforcement of PSM firewall policy between VMs hosted by the same hypervisor in the same VLAN. Traffic between VMs installed on the same hypervisor will not forward traffic to the switch in a standard configuration. To perform microsegmentation, a private VLAN (PVLAN) strategy is used to force traffic from a VM to traverse the upstream DSS switch connected to the hypervisor. Enabling proxy ARP on the SVI of the primary PVLAN allows communication between hosts in the same private VLAN via the DSS switch. This strategy enables the application of both ingress and egress firewall policy between hosts in the same VLAN on the same hypervisor.

When using a Layer 2 two-tier topology, only egress policy can be applied to microsegmented traffic, because the VLAN SVI is configured at the core layer, while the CX 10000 policy engine is positioned at the access layer.

When using an EVPN-VXLAN topology, in addition to egress policy, ingress policy can be applied to microsegmented traffic, because the VLAN gateway IP is configured on every leaf switch using the active gateway feature.

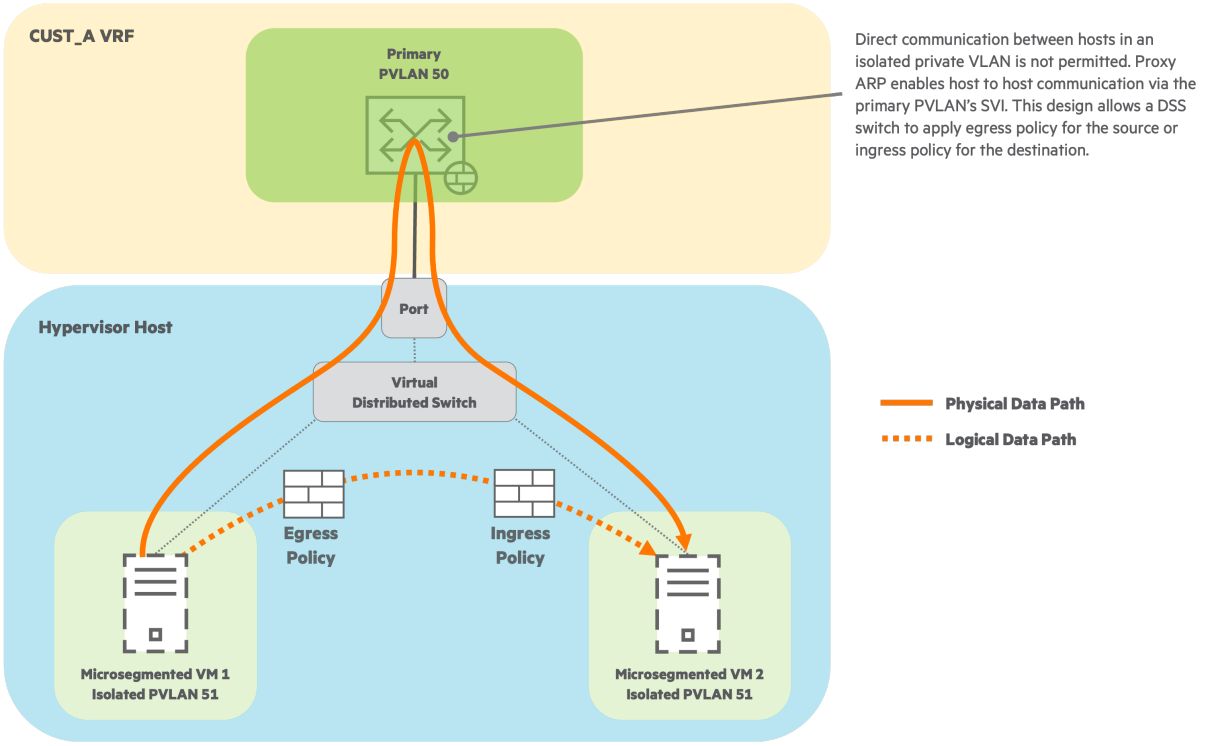


Figure 28: Microsegmentation Egress Policy Enforcement Diagram

A PVLAN microsegmentation strategy also enables policy enforcement between containers.

Fabric Composer Policy Automation

HPE Aruba Networking Fabric Composer is a GUI-based management tool for building and maintaining data center switch configuration. Fabric Composer automation tools simplify building spine-and-leaf and Layer 2 two-tier topologies, and EVPN-VXLAN overlays. Fabric Composer monitoring tools provide insight on current operational state and assist in troubleshooting processes. Fabric Composer is also a full-featured data center policy management tool. API integration with PSM and the AOS-CX operating system enable management of DSS firewall policy and switch ACLs.

Additional VMware vSphere API integration enhances Fabric Composer's policy management by enabling VMware administrators to assign DSS policy sets to individual VM guests without requiring coordination with network or security groups.

Fabric Composer assigns policy to endpoint group objects, which are sets of one or more IP addresses. A special dynamic endpoint group auto-populates VM IP address assignments from vCenter based on VM tag assignments on VM guests. When VMware administrators assign or modify VM tags, Fabric Composer automatically updates IP assignments with associated endpoint groups. These changes are propagated to PSM and switches by Fabric Composer.

PSM Policy Considerations

Rules in a policy are applied in the order they appear on the list. An implicit "deny all" rule is applied at the end of a rule set. Rules used more often should appear higher on the list.

Defining a PSM *Network* redirects all traffic for the associated VLAN to the Pensando DPU firewall as described above. Network requirements of all VLAN members must be considered when building policy rule sets.

When an ingress policy is configured, policy applies to all hosts in the destination VLAN for any routed or VXLAN-forwarded Layer 2 traffic, which includes traffic sourced from outside the data center.

When an egress policy is configured, policy applies to all traffic sourced by hosts in the VLAN, so all destinations within and outside the data center must be considered, including all hosts on the same VLAN. When defining an egress policy, rules allowing underlying services such as DNS, logging, and authentication are required.

When applying firewall policy between VM guests, a data center design must use PVLAN-based microsegmentation or assign VMs that require policy between them to different VLANs. Microsegmentation forces traffic between VMs on the same hypervisor to traverse the DSS switch, allowing policy to be applied. VMs in different VLANs also send traffic to the DSS switch for inspection.

It is recommended to apply only one PSM policy level (*Network* or *VRF*) for a particular enforcement direction. The primary advantage of defining policy at the *VRF* level is having a single policy represent the complete rule set for the entire data center in a single enforcement direction. Defining policy at the *Network* level reduces the size of an individual policy and ensures that policy applied to one *Network* does not impact another *Network*.

Mixing policy levels for a single direction can add complexity and duplication of rules in both sets of policy, although mixing policy levels is fully supported. When defining policies at both levels, one policy level requires a rule to permit any traffic at the end of the policy, which is recommended in the *Network* level policy. The policy rules defined above the “permit any” rule should use a “deny” action. This allows the *VRF* level policy to define what is permitted globally within the fabric with more granular deny restrictions applied at the *Network* level.

It is best practice to define a complete set of rules before applying a policy to a network. If the complete rule set is unknown, an “allow all” rule can be applied to collect log data on observed traffic. A complete rule set can be built by inserting rules to allow more specific traffic above the “allow all” rule. When no wanted traffic is hitting the “allow all” rule at the bottom of the rule set, remove it.

DSS Best Practices

The CX 10000 out-of-band management (OOBM) port is recommended for management plane communication with the switch. This ensures that switch configuration and policy changes do not unintentionally block switch reachability, while providing a communication path to recover from operator errors. When in-band management is required, configure the **ip source-interface psm** on the CX 10000 to permit management communication between the switch and PSM using the data plane.

When a firewall or access-control-lists are placed between DSS switch management IP interfaces and PSM, communication on a number of TCP ports must be allowed between the DSS and PSM for proper operation. The required open ports are listed in Appendix B of the **Policy and Services Manager for HPE Aruba Networking CX 10000 User Guide** available on the HPE Networking Support Portal.

Time synchronization is required to maintain DSS switch registration with PSM. Configure the same set of NTP servers for both CX 10000 switches and PSM to avoid registration issues on PSM.

PSM can consume logs during setup and testing. Production networks should export firewall logs to an external collector. Syslog and IPFix collectors must be reachable on the default VRF of DSS switches.

When using symmetric IRB, all VLANs redirected for policy enforcement should be present on all DSS switches in a fabric, and an Active Gateway should be configured on all VLAN interfaces.

When applying policy to NFS storage traffic, enable session re-use on applicable rules to permit proper NFS operation.

Data Center Storage and Lossless Ethernet

HPE Aruba Networking data centers support Data Center Bridging (DCB) protocols that create lossless Ethernet fabrics to support storage area networks, big data analytics, and artificial intelligence (AI) applications.

Storage Over Ethernet Challenges

Traditional IEEE 802.3 Ethernet relies on higher layer protocols, such as TCP, to accommodate strategies for reliable data delivery. Data transmitted over an Ethernet network can be lost between source and destination hosts, which incurs a performance penalty on applications sensitive to data loss.

Storage performance is particularly sensitive to packet loss. TCP can guarantee data delivery at the transport layer by sequencing data segments and performing retransmission when loss occurs, but the need to perform TCP retransmissions for storage significantly reduces the performance of applications depending on that storage.

Advances in storage technology, such as SSD flash memory and the Non-Volatile Memory express (NVMe) protocol, facilitate read/write storage that exceeds the performance of traditional storage networking protocols, such as FibreChannel. The performance bottleneck in a storage area network (SAN) has moved from the storage media to the network.

Remote Direct Memory Access (RDMA) was developed to provide high-performance storage communication between two networked hosts using the proprietary InfiniBand (IB) storage network. IB guarantees medium access and no packet loss, and requires a special host bus adapter (HBA) for communication. The IB HBA receives and writes data directly to host memory using dedicated hardware, bypassing both traditional protocol decapsulation and the host's primary CPU. This reduces latency, improves performance, and frees CPU cycles for other application processes.

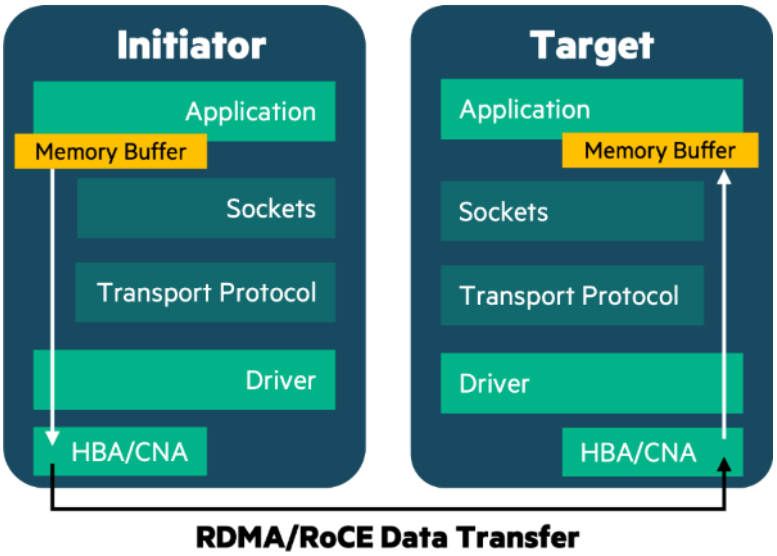


Figure 29: RoCE Data Transfer

Ethernet solutions offer high-speed networking interfaces, making them attractive options for storage communication, if the reliability issue can be solved. RMDA over converged Ethernet (RoCE) is a protocol developed by the InfiniBand Trade Association (IBTA) to extend RDMA reliability and enhanced performance over a low-cost Ethernet network. A converged network adapter (CNA) performs the task of writing received data directly to memory and enables Ethernet as the underlying communication protocol. A lossless data communication path to support RoCE is created by modifying both Ethernet host and switch behavior.

RoCE version 1 (RoCEv1) encapsulates IB Layer 3 addressing and RDMA data directly into an Ethernet frame. Ethernet replaces RDMA Layer 1 and 2 functions, and it specifies a unique EtherType value to indicate RDMA as the Ethernet payload.

Figure 30: RoCEv1 Ethernet Frame

RoCE version 2 (RoCEv2) replaces IB Layer 3 addressing with IP. It encapsulates IB Layer 4 and RDMA data into a UDP header. This strategy makes RoCEv2 routable over IPv4 and IPv6 networks. RoCEv2 is the most common implementation of RoCE.

Figure 31: RoCEv2 Packet

The lossless Ethernet optimizations implemented in CX switches improve data center performance for applications using both RoCE and non-RoCE protocols such as standard iSCSI. In addition to storage communication, RoCE enhances the performance of database operations, big data analytics, and generative AI.

Non-Volatile Memory express (NVMe) is an intra-device data transfer protocol that leverages multi-lane data paths and direct communication to the CPU provided by PCIe to move large amounts of data at a high rate with low latency. NVMe is designed specifically for solid state drives (SSDs) as a replacement for the decades-old Serial Advanced Technology Attachment (SATA) protocol. NVMe over Fabrics (NVMe-oF) extends NVMe to work between networked hosts. NVMe-oF works over multiple protocols, including RoCE.

The primary challenge running RDMA over Ethernet is overcoming the problem of link congestion, the most common cause of dropped Ethernet frames in a modern Ethernet network. Link congestion occurs when frames are received at a faster rate on a switch than can be transmitted on an outgoing port. Link congestion has two common causes. First, the receive and transmit ports on a switch are different speeds, so the higher speed port can receive data faster than transmission to the lower speed port allows. Second, a switch receives a large number of frames on multiple interfaces destined to the same outgoing interface. In both cases, the switch can queue surplus frames in a memory buffer until the outgoing port is able to transmit them. If buffer memory becomes full, additional incoming frames are dropped as long as the buffer remains full. This results in TCP retransmissions and poor application performance.

Building Reliable Ethernet

A lossless Ethernet fabric can be created by connecting a contiguous set of switches and hosts that employ a set of strategies to prevent frame drops for a particular application.

Three primary Quality of Service (QoS) strategies manage competing demands for buffer memory and switch port bandwidth: dedicated switch buffers for an application, flow-control, and guaranteed media access for an application. Combining these three strategies enables a lossless Ethernet fabric for storage and other applications.

The following table displays the key DCB protocols supported by HPE Aruba Networking CX data center switches.

Data Center Bridging Component	Description
PFC: Priority Based Flow Control	Establishes queues that do not drop packets by preventing buffer exhaustion.
ETS: Enhanced Transmission Selection	Defines bandwidth reservations for traffic classes so that lossless and lossy traffic can coexist on the same link.
DCBx: Data Center Bridging Exchange Protocol	Exchanges PFC and ETS information between devices on a link using Link Layer Discovery Protocol (LLDP) to simplify configuration.

In addition to the protocols above, CX switches support IP Explicit Congestion Notification (ECN). IP ECN is a Layer 3 flow-control method that allows any switch in the communication path to notify a traffic receiver of the presence of congestion. After receiving a congestion notification, the receiving host sends a direct, IP-based congestion notification to the traffic source to slow its data transmission rate.

Enhancements in RoCE have produced two different versions. RoCEv1 relies on the base DCB protocols in the table above and is not supported over a routed IP network. RoCEv2 enables IP routing of RoCE traffic, includes IP ECN support, and is the protocol version most often referenced by the term “RoCE.”

Priority Flow Control

Ethernet pause frames introduced link-level flow control (LLFC) to Ethernet networks in the IEEE 802.3x specification. When necessary, a traffic receiver can request a directly connected traffic source to pause transmission for a short period of time, allowing the receiver to process queued frames and avoid buffer exhaustion. The traffic source can resume transmitting frames after the requested pause period expires. The receiver also can inform the source that a pause is no longer needed, so the source can resume transmitting frames before the original pause period expires.

Priority Flow-Control (PFC) works in conjunction with quality of service (QoS) queues to enhance Ethernet pause frame function. PFC can pause traffic on a per-application basis by associating applications with a priority value. When PFC pauses traffic associated with an individual priority value, traffic assigned other priorities are unaffected and can continue to transmit.

On a link, both the CX switch and attached device must locally assign a priority to application traffic and indicate that priority to its peer on the link. Traffic priority can be signaled using either 802.1p Priority Code Point (PCP) values or Differentiated Services Code Point (DSCP) values.

PCP Priority Marking

The IEEE 802.1Qbb standard uses 802.1p PCP values in an 802.1Q header to assign application traffic priority. The three-bit PCP field allows for eight Class of Service (CoS) priority values (0-7). PCP-based PFC requires the use of trunk links with VLAN tags to add an 802.1Q header to a frame.

The diagram below illustrates the PCP bits used to specify 802.1p CoS priorities in the 802.1Q header of an Ethernet frame.

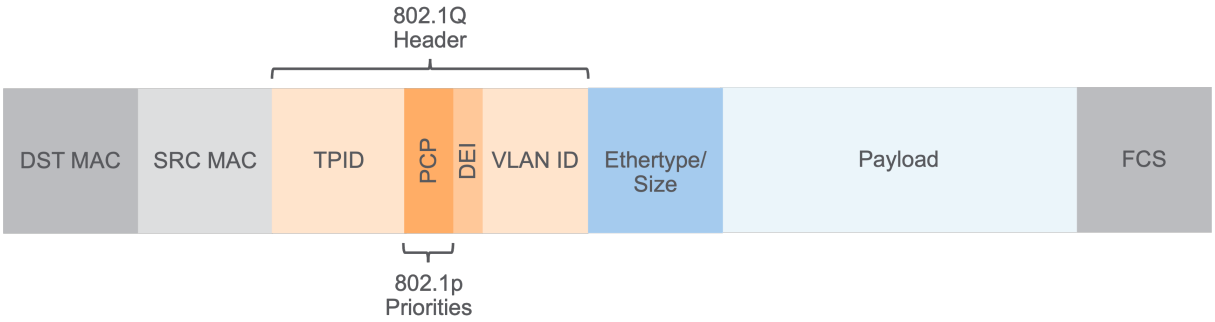


Figure 32: 802.1Q Header Diagram in Ethernet Frame

By default, there is a one-to-one mapping of CoS priorities to local priorities on the switch used for frame queueing.

DSCP Priority Marking

Lossless behavior between two data center hosts requires that both hosts and all switches in the data path have a consistent PFC configuration. If a routed-only interface is in the data path, application priority can be preserved by specifying a priority using the DSCP bits in the IP header. DSCP bits also can be used to mark application traffic priority on both 802.1Q tagged and untagged switch access ports.

The diagram below illustrates the DSCP bits located in the legacy Type-of-Service (ToS) field of the IP header.

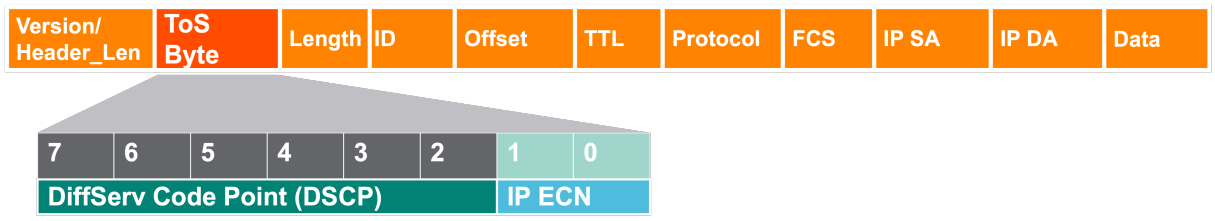


Figure 33: DSCP bits in ToS Field of IP header

The six-bit DSCP field allows for 64 DiffServ priority values. By default, DiffServ values are mapped in sequential groups of eight to each of the eight individual local-priority values.

CX switches support a mix of CoS and DSCP priority values by allowing each interface to specify which QoS marking method is trusted. When a mix of strategies is present on different switch ports, traffic must may require re-marking between Layer 2 CoS priority values and Layer 3 DSCP values.

Responding to the growth of routed spine-and-leaf data center architectures and VXLAN overlays, an increasing number of hosts and storage devices support DSCP-based priority marking. This enables consistent QoS markings across a routed domain without the need to translate between Layer 2 CoS values and Layer 3 DSCP values on network switches.

In addition to CoS and DSCP values, CX switches can apply a classifier policy to ingress traffic to assign priorities (PCP, DSCP, and local) based on header field values in the packet.

When a frame is encapsulated for VXLAN transport, the QoS DSCP priority of the encapsulated traffic is honored in the outer VXLAN packet's IP header to ensure proper queueing.

PFC Operations

CX data center switches support a special shared QoS buffer pool dedicated for lossless traffic. The CX 8325, 10000, and 9300 models support up to three lossless pools. Typically, only one lossless queue is defined for storage traffic. Each lossless pool is assigned a size, headroom capacity, and associated local priority values. The buffers assigned to a lossless pool are allocated from the total available buffer memory on the device, which are assigned to a single lossy pool by default. The CX 8100 and 8360 support a single, fixed lossless pool for smaller data centers.

Received frames are assigned a local priority value based on a mapping of PCP and DSCP values to local priority values. A frame is placed into the special lossless buffer pool when its local priority value is associated with a lossless queue. When a port's allocation of the shared lossless buffer pool nears exhaustion, packet drops are avoided by notifying the directly-connected sender to stop transmitting frames with the queue's associated priority value for a short period of time. The headroom pool stores packets that arrive at the interface after a pause in transmission was requested for the associated priority.

PFC support is included on the CX 8325, 9300, 10000, 8360, and 8100. However, traffic arriving on a CX 10000 with a QoS priority associated with a lossless queue will not be sent to the AMD Pensando Distributed Processing Unit (DPU) for policy enforcement or enhanced monitoring.

The diagram below illustrates the queuing relationship between a sender and a CX switch receiver with two queues defined using CoS priority values. All priorities are mapped to the default lossy queue or to a single lossless queue. Using two queues on the CX platform provides the best queue depth and burst absorption.

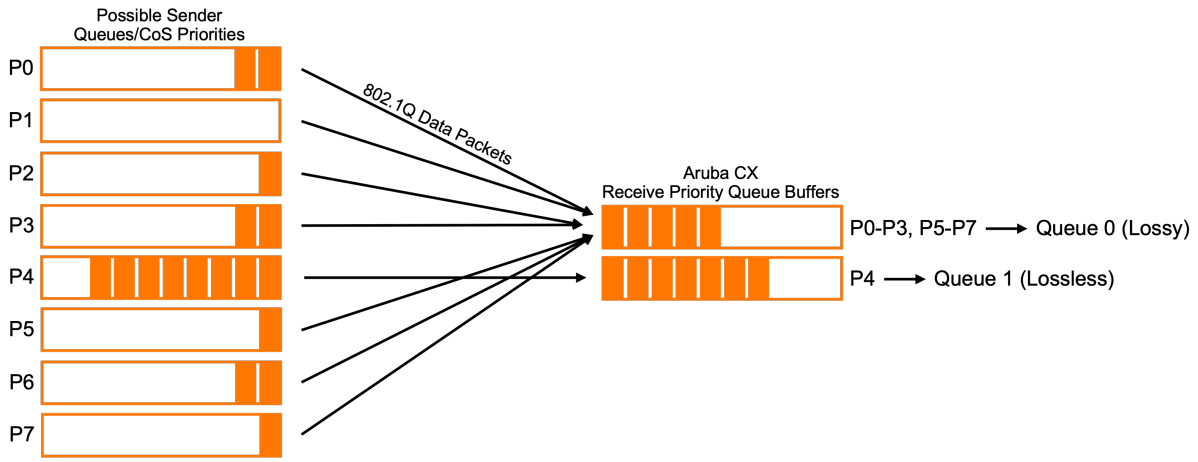


Figure 34: 802.1p Priority Based PFC

A PFC pause notification briefly stops transmissions related to a single application by its association with a priority queue number.

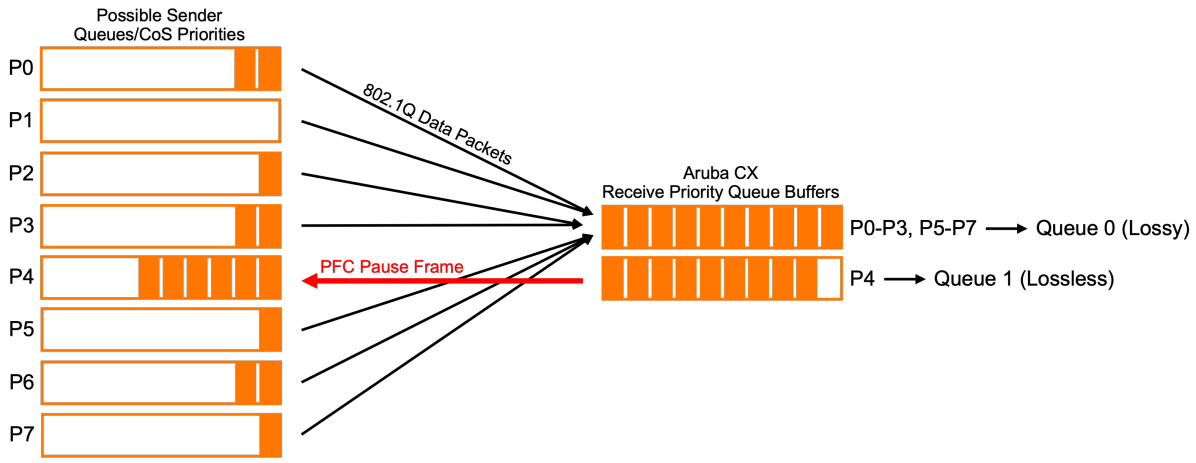


Figure 35: 802.1p Priority Based PFC

Storage is the most common application for lossless Ethernet. Applying the diagram above to a storage scenario, all storage traffic is assigned a PCP value of 4, which is mapped to local-priority 4. When storage traffic is received on the CX switch, it is placed in the lossless QoS queue dedicated for storage. Traffic assigned to the lossy queue does not impact buffer availability for the lossless storage traffic. When the lossless storage queue on the CX switch reaches a threshold nearing exhaustion, a pause frame is sent to inform the sender to pause only storage traffic. All other traffic from the sender continues to be forwarded and is placed in the shared lossy queue on the CX switch, if buffers are available.

Link-Level Flow Control (LLFC)

PFC is the preferred flow-control strategy, but it requires data center hosts to support marking traffic priority appropriately. PFC is built into specialized HBAs and is required for RoCE compliance.

LLFC can enable lossless Ethernet when implemented in combination with other QoS components for prioritization, queueing, and transmission. Many virtual and physical storage appliances do not support PFC or other DCB protocols, but LLFC is widely supported on most standard Ethernet network interface cards (NICs). Implementing LLFC extends the benefits of lossless data transport to hosts that do not support PFC and for non-RoCE protocols.

All traffic received on a switch port using LLFC is treated as lossless. It is recommended to minimize sending lossy traffic from a host connected to a link using LLFC.

When a CX switch sends an LLFC pause frame to an attached device, it pauses all traffic from that source instead of from a single targeted application. The pause in transmission gives the switch time to transmit frames in its lossless queues and prevents frame drops.

Application traffic priority is typically not signaled from a source limited to link-level flow control. In place of the source marking traffic priority, a classifier policy is implemented on the CX ingress port to identify application traffic that should be placed in a lossless queue by matching defined TCP/IP characteristics. When an interface also trusts DSCP or CoS priority values, the trusted QoS markings are honored and take precedence over a custom policy prioritization.

Enhanced Transmission Selection (ETS)

ETS allocates a portion of the available transmission time on a link to an application using its association with a priority queue number. This helps to ensure buffer availability by guaranteeing that the application traffic has sufficient bandwidth to transmit queued frames. This behavior reduces the probability of congestion and dropped frames.

Allocation of bandwidth is divided among traffic classes. ETS is implemented on CX switches using QoS scheduling profiles, where locally defined queues are treated as a traffic class. Traffic is associated with a queue by associating it with a local priority value. Traffic can be mapped to local priorities based on DSCP priorities, CoS priorities, or TCP/IP characteristics using a classifier policy.

CX 8325, 10000, and 9300 switches apply a deficit weighted round robin (DWRR) strategy to calculate a queue's bandwidth allocation by applying a weight to each queue in a scheduling profile. The following example shows the resulting percentage of bandwidth associated with a queue for the collective set of weights.

Queue Number	Weight	Guaranteed Bandwidth
Queue 0 (Lossy)	8	40%
Queue 1 (Lossless)	10	50%
Queue 2 (Lossless)	2	10%

In the example above, storage traffic can be assigned to queue 1, which guarantees storage traffic the ability to consume up to 50% of the link's bandwidth. When a class of traffic is not consuming its full allocation, other classes are permitted to use it. This enables the link to operate at full capacity, while also providing a guaranteed allocation to each traffic class. When a link is saturated, each class can consume only the bandwidth allocated to it based on the assigned weights.

Multiple scheduling profiles can be defined, but an individual port is assigned a single profile that governs its transmission schedule.

The following diagram illustrates traffic arriving on a switch, being placed in a queue, and the reserved bandwidth per queue of the outgoing port. Scheduling enforcement occurs when the outgoing port is saturated and the ingress rate for each traffic class meets or exceeds the reserved bandwidth configured on the outgoing port.

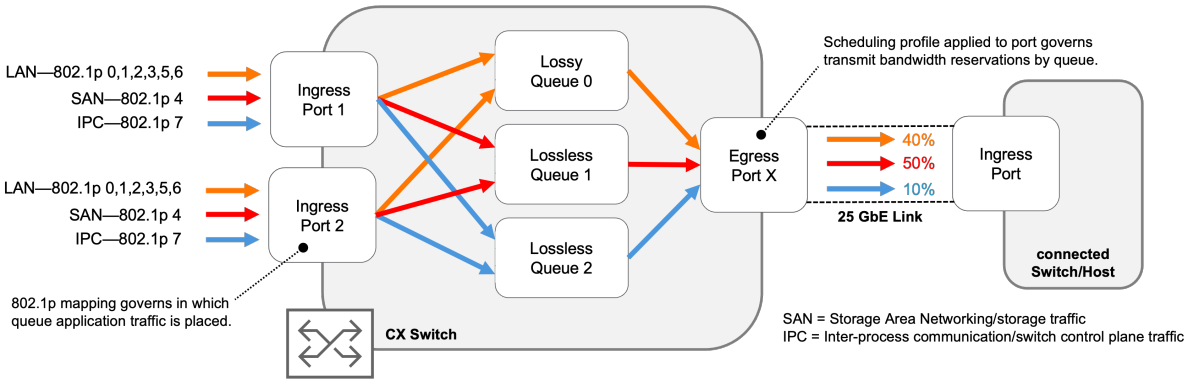


Figure 36: ETS Bandwidth Reservation

When the outgoing port is not oversubscribed, its transmission rates may be different. The unused bandwidth allocations in one class may be consumed by another class. For example, if the port is transmitting at 75% of its capacity, where 60% is from queue 0, 20% is from queue 1, and 5% is from queue 2, the switch does not need to enforce the scheduling algorithm. The lossy traffic in queue 0 is allowed to consume the unused capacity assigned to other traffic classes and transmit at a higher rate than the schedule specifies.

Data Center Bridging Exchange (DCBx)

DCBx-capable hosts dynamically set PFC and ETS values advertised by CX switches. This ensures a consistent configuration between data center hosts and attached switches. DCBx also informs compute and storage hosts of application traffic to priority mappings, which ensures that traffic requiring lossless queuing is marked appropriately. Lossless Ethernet configuration on connected hosts becomes a plug-and-play operation by removing the administrative burden of configuring PFC, ETS, and application priority mapping on individual hosts.

DCBx is a link-level communication protocol that employs Link Layer Discovery Protocol (LLDP) to share settings. PFC, ETS, and application priority settings are advertised from the switch using specific LLDP Type-Length-Value (TLV) data records. CX switches set the *willing bit* to 0 in all TLVs to indicate that it is not willing to change its configuration to match its peer's configuration. CX switches support both IEEE and Convergence Enhanced Ethernet (CEE) DCBx versions.

IP Explicit Congestion Notification (ECN)

IP ECN is a flow-control mechanism that reduces traffic transmission rates between hosts when a network switch or router in the path signals that congestion is present. IP ECN can be used between hosts separated by multiple network devices and on different routed segments. It is required for RoCEv2 compliance.

Hosts that support IP ECN set one of two reserved Type of Service (ToS) bits in the IP header to a value of 1. When a switch or router in the communication path experiences congestion, it sets the remaining zero ECN bit to 1, which informs the traffic receiver that congestion is present in the communication path.

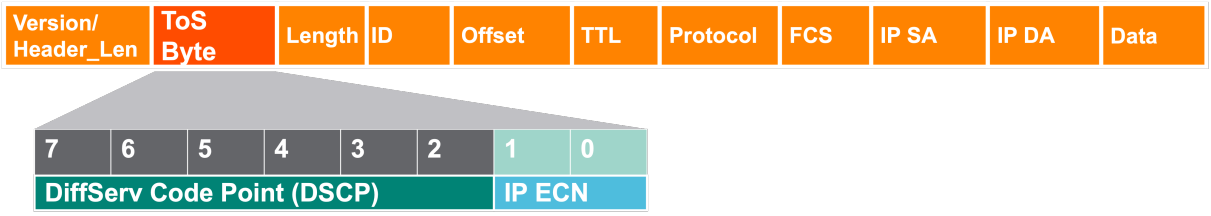


Figure 37: ECN bits in ToS Field of IP header

When the traffic receiver is notified of congestion, it signals this to the source by sending an IP unicast message. The source responds by reducing its data transmission rate for a brief period of time.

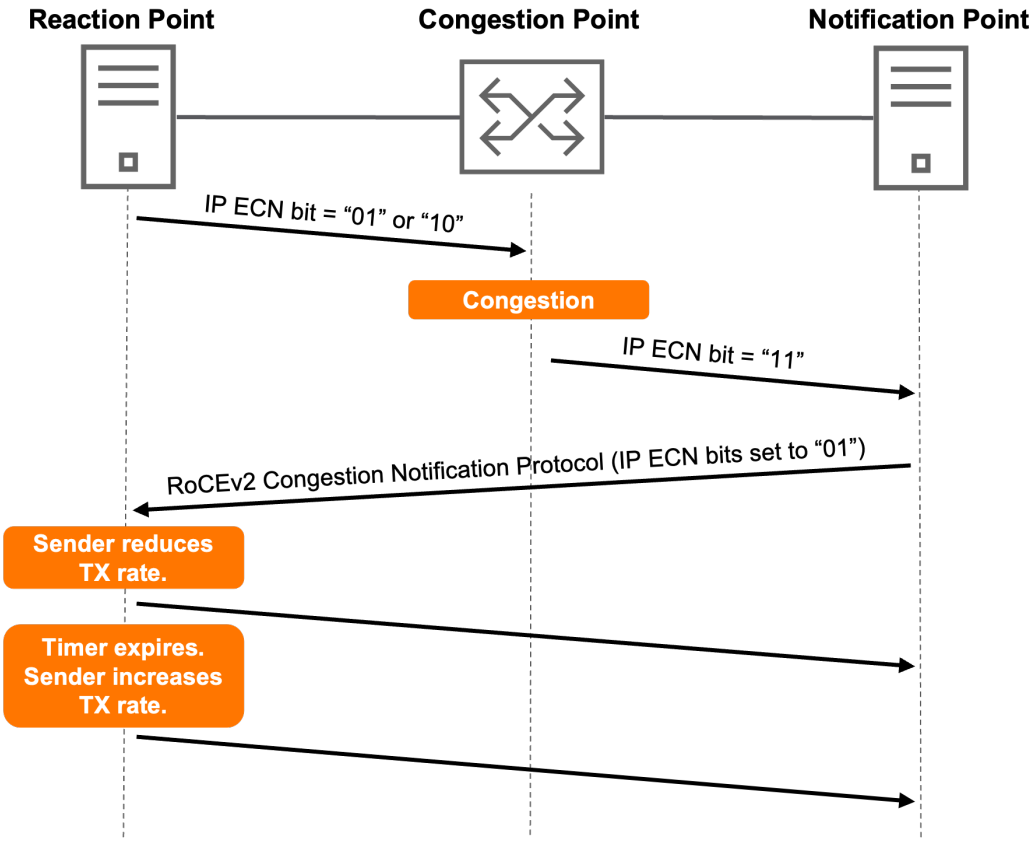


Figure 38: RoCEv2 IP ECN Process

IP ECN smooths traffic flows under most conditions, reducing the need for PFC to trigger a full pause, except as a fast acting mechanism to address microbursts.

IP ECN also can be implemented to improve the performance of non-RoCE protocols, such as iSCSI.

Data Center Networking for AI

Introduction

Artificial Intelligence (AI) has revolutionized industries, driving exponential growth in applications. However, AI workloads demand high-performance computing, low-latency networks, and scalable storage. To address these requirements, an AI-optimized data center network is designed and built on a comprehensive framework for AI-supportive networks.

Traditional data center technologies struggle to meet AI workload demands, necessitating cutting-edge solutions in compute, storage, and networking. Specialized AI data centers require tailored back-end training and front-end inference fabric designs. While Graphics Processing Units (GPUs) and InfiniBand networking have emerged as key technologies, the single-sourced, proprietary nature of InfiniBand drives up costs. In response, enterprises are embracing Ethernet as a cost-effective, open networking alternative for AI data centers, optimizing GPU performance while minimizing costs.

To accelerate AI adoption, data center networks play a vital role in optimizing GPU interconnectivity and performance. Reducing Job Completion Time (JCT) is critical for faster speed and cost savings. Furthermore, rapid market response to demand is essential for successful AI deployment. In response, the industry is shifting towards an open, competitive market fueled by GPU diversity and Ethernet, the widely deployed Layer 2 technology. This transition promises to alleviate reliance on single-vendor solutions, promoting flexibility, scalability, and cost-effectiveness.

The Data Center Network for GenAI

Advances in GenAI have made AI and machine learning (ML) a new part of corporate business tools. Data centers are the engines behind AI, and data center networks play a critical role in interconnecting and maximizing the utilization of expensive GPU servers.

GenAI training, measured by job completion time (JCT), is a massive parallel processing problem. A fast and reliable network fabric is needed to get the most out of the GPUs. The right network and design are key to optimizing ROI and maximize savings on AI applications.

A typical AI-Optimized DC Network for Generative AI workload consists of:

- **Compute Nodes:** High-performance servers with AI-optimized processors (e.g., GPUs, TPUs)
- **Storage:** High-capacity storage with low-latency access (e.g., NVMe, SSD)
- **Networking:** High-density,low-latency switches and AI-optimized network protocols
- **Software:** AI frameworks (e.g., TensorFlow, PyTorch) and network management tools

Generative AI Best Practice Architecture

The architecture for AI best practice includes frontend, backend and storage fabrics. These fabrics have a symbiotic relationship, and provide unique functions in the training and inference tasks in this architecture.

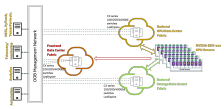


Figure 39: GPU arch

Front-End Network

The frontend network for GenAI plays a critical role in ensuring high-performance, low-latency connectivity for AI and machine learning (ML) workloads. Design considerations include utilizing high-speed Ethernet switches, such as 100GbE or 400GbE, to interconnect AI servers and storage. Additionally, implementing EVPN-VXLAN enables efficient traffic management and scalability. HPE Aruba networking CX series switches, combined with HPE Central and AFC management tools, provide a robust and automated solution. By adopting these best practices, organizations can build a high-performance frontend network optimized for AI workloads.

Storage Network

A high-performance storage fabric is crucial for AI and machine learning (ML) workloads, requiring low-latency and high-bandwidth connectivity. HPE storage fabric solution utilizes CX series switches, enabling 100GbE or 200GbE connectivity to the storage arrays. Protocols like RoCEv2, NVME are utilized on a converged ethernet fabric solution to provide lossless transport for storage traffic.

GPU Clusters

GPU clusters, also referred to as GPU Fabric provide massive parallel computing power needed to process large datasets and complex neural networks rapidly, accelerating the training time and enabling fast, efficient inference on new data. GPUs are designed with thousands of cores that can perform calculations simultaneously, ideal for the parallel processing required in training large Gen AI models. With the parallel processing capabilities of GPU's, the time needed to train complex GenAI Large Language Models (LLM) are significantly reduced. GPU clusters can be scaled up by adding more GPU nodes to handle larger and more complex datasets, as needed.

Back-end Network

Backend networks are specialized networks connecting GPU clusters for distributed Large Language Model (LLM) training, enabling high-bandwidth data transfer and efficient parallel computation. This would require a high-performance Rail-Optimzed or Rail-Only architected network, featuring low latency, robust design, no link oversubscription between workload and fabric, and lossless Ethernet DC fabric

HPE CX series switches (100/200/400GbE) address high-bandwidth and low-latency demands. AI-optimized network protocols, such as Global Load Balancing (GLB for end-to-end load balancing), efficiently manage elephant flows inherent in AI workloads. Network automation streamlines management and configuration, adapting to dynamic AI workloads. Additionally, robust security measures safeguard AI systems and data, protecting against vulnerabilities and threats.

Overview of GPU Server and Interconnects

GPU servers utilize internal high-bandwidth PCIe switches for efficient interconnectivity, facilitating communication between key components: CPU to GPU, GPU to NIC (Network Interface Card), and NIC to NIC bidirectional communication. As illustrated in the GPU Server architecture schematic, these servers employ very high-bandwidth NIC adapters (100/200/400/800G) to network and scale a large group of GPUs. This architecture is specifically designed to support the demanding requirements of training Large Language Models (LLMs).

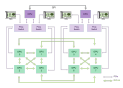


Figure 40: GPU arch

The NVIDIA NVSwitch is a high-speed switch chip (>900GB/s) connecting multiple GPUs through NVLink interfaces. It enhances intra-server communication and bandwidth while reducing latency for compute-intensive workloads.

Backend GPU Fabric for GenAI Training

The Backend Data Center fabric consist of multiple GPUs connected to form a cluster, enabling distributed training and model parallelism. Distributed training splits the training process across multiple GPUs, while model parallelism divides the model into smaller parts processed by different GPUs. Communication protocols like NCCL (NVIDIA Collective Communication Library) and MPI (Message Passing Interface) facilitate efficient communication between GPUs.

LLM training involves massive datasets and complex models, requiring thousands of GPU server nodes to process efficiently. High-performance networking (100/200/400GbE) ensures scalable and fast data transfer. Real-time communication between nodes is crucial for parallel processing. Low latency (<100 µs) enables faster iteration and convergence. LLM training is computationally intensive and time-consuming. Redundancy and failover mechanisms guarantee uninterrupted processing. Quality of Service ensures critical tasks receive sufficient bandwidth, optimising overall performance.

Packet loss and congestion significantly impair Large Language Model (LLM) training, reducing throughput, increasing latency and affecting model accuracy. This results in high Job completion time (JCT). Reliable networks should guarantee minimizing packet loss and congestion to ensure timely and accurate results.

RoCEv2 protocol was designed to provide a low cost Ethernet alternative to the InfiniBand Networks. It leverages RDMA (Remote Direct Memory Access) technology to bypass CPU overhead, reducing latency and increasing throughput. This protocol increases data delivery performance by enabling the GPUs directly access its memory as shown in figure. RoCEv2 uses PFC and ECN protocols to implement lossless behavior in the data center.

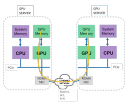


Figure 41: rail-optimized arch

The Backend GPU Fabric also enables Global Load Balancing techniques on the network links to deploy a congestion free environment to reduce latency and to support the low entropy elephant flows that are typical in a GPU message exchange communications. The low entropy elephant flow needs better load balancing than the typical ECMP that is used in enterprise DC networks. There are solutions that achieve this in the switch hardware chipset and would perform at the highest efficiency. Software driven solutions can be equally good for small/medium LLM size training.

When designing backend fabrics, the primary objective is to achieve a lossless architecture that maximizes throughput, minimizes latency, and network interference for AI traffic flows. To accomplish this, several design architectures are available, including 3-stage CLOS, NVIDIA's Rail-Optimized, and Rail-only architectures. These architectures leverage leaf switches, spine switches and super-spine switches. Notably, Rail-only simplifies design by utilizing solely leaf switches to build the backend fabric

Rail-Optimized Backend GPU Fabric

This architecture consists of leaf-spine network switches as shown in the figure. The GPU servers are connected to the leaf/spine following the rail-optimized technique. Rail-optimized archectecure is a design methodology that organizes data centers infrastructure into logical rails or paths, ensuring efficient data flow. The figure below shows a groups 8 GPU servers with 8xGPUs each. A full rail stripe populate as many GPUs as the number of ports the leaf has to create optimized paths for GPU to GPU communication

The first GPU from each (NVIDIA DXG H100) GPU server will all connect to the leaf1 switch, and all second GPUs to Leaf2 and so on. This helps optimize RDMA message flow between GPU servers by using just the Leaf switch when the communication is within the 8xGPU servers. The spine is used only when the message needs to cross over to the next logical rail of GPU servers. The figure next to following figure shows a fully populated GPU servers that shows two logical rail stripes.

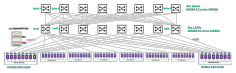


Figure 42: rail-optimized arch

Here both leaf and spine switches have 64x400GbE ports. With 1:1 subscriptions 32x400GbE ports will connect to the spines and the other 32x400GbE can connect to 32xGPU servers. The spines uses 32x400GbE ports to connect to each rail stripe.

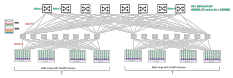


Figure 43: rail-optimized arch

The fabric implement Lossless Ethernet transport, and uses RoCEv2, PFC, ECN protocols to make sure no packets are dropped. It enables Global Load Balancing techniques on the network links to deploy a congestion free environment to reduce latency and to support the elephant flow that are typical in a GPU communications. The low entropy elephant flow needs better load balancing than the typical ECMP that is used in enterprise DC networks.

Rail-Only Architecture

The Rail-Only architecture differs significantly from Rail-optimized, eliminating spine nodes and utilizing solely leaf nodes to connect GPU servers. As illustrated, this design employs eight 64x400GbE switches, supporting up to 64 GPU servers. To scale GPU capacity, higher-port-density switches are required, limiting this architecture's scalability for large GPU clusters. However, Rail-Only suits small enterprises with smaller training LLM sizes that fit comfortably within this framework.

The fabric implement Lossless Ethernet transport using RoCEv2, PFC, ECN protocols. It enables Global Load Balancing techniques on the network links to deploy a congestion free environment to reduce latency and to support the elephant flow that are typical in a GPU communications. The low entropy elephant flow needs better load balancing than the typical ECMP that is used in enterprise DC networks.

Figure 44: rail-only arch

The Rail-Only network architecture offers significant cost savings, primarily due to the elimination of spine switches. Compared to Rail-Optimized, this design reduces capital expenditures (CapEx) by approximately 50-75%.

Validation Criteria

- Throughput: 100GbE minimum throughput per port
- Latency: <10 µs latency for real-time AI applications
- Scalability: Support for 1000+ AI nodes or applicable
- Security: Compliance with industry-standard security protocols

Conclusion

Among various designs being explored for backend training networks, Rail-Optimized and Rail-Only architectures stand out as cost-effective solutions offering balanced performance. For GenAI training backend data center fabric, hardware-based solutions provide optimal performance but come at a high cost. Software-driven approaches offer a more effective balance between features and deployment costs.

Small and Medium Enterprises (SMEs) seeking compact deployment footprints can leverage Rail-only architecture for efficient training within reasonable timelines. This approach reduces capital expenditures (CapEx) by approximately 50% compared to Rail-Optimized architecture, primarily by eliminating spine switch costs.

HPE Aruba Networking offers a comprehensive portfolio, featuring 100/200/400GbE CX series switches, powered by the AOS-CX network operating system. This combination provides a robust, lossless solution, supporting essential protocols and features for efficient data center connectivity.

Unlock the full potential of your AI initiatives with our expert-designed networking architecture, optimized for high-performance, scalability and security. Maximize throughput and minimize latency, scale effortlessly and protect your data with robust security measures. This transformative solution empowers businesses with faster insights from AI applications, enhanced collaboration and productivity, and a future-proof infrastructure, giving you the competitive edge you need to succeed.

Storage Positioning

Storage in a data center is typically deployed as a SAN, part of hyper-converged infrastructure (HCI), or as disaggregated HCI (dHCI).

SANs comprise one or more dedicated storage appliances that are connected to servers over a network. A proprietary network using storage based protocols, such as FibreChannel, can be used to connect servers to storage. However, IP-based solutions over an Ethernet network provide a high-bandwidth, low-cost option, with accelerating adoption levels. Common IP-based SAN protocols include iSCSI and RoCE.

HCI decouples the storage and compute capabilities of off-the-shelf x86 infrastructure, providing a cloud-like resource management experience. Each x86 host in the HCI environment provides both distributed storage and compute services. The local storage on an HCI cluster member can be used by any other member of the cluster. This provides a simple scaling model, where adding an additional x86 node will add both additional storage and compute to the cluster.

The HPE SimpliVity dHCI solution divides compute and storage resources into separate physical host buckets to allow flexible scaling of one resource at a time. In the traditional HCI model, both storage and compute must be increased together when adding an x86 node. This can be costly if only an increase in one resource is required. For example, if additional compute is required and storage is already adequately provisioned in an HCI solution, significant additional storage is still added to the cluster regardless of the need. dHCI supports scaling compute and storage individually, while using x86 hardware for both compute and storage services.

All the above storage models improve performance when using lossless Ethernet.

Parallel Storage Network

Traditionally, a storage network is deployed in parallel with a data network using proprietary network hardware and protocols to support the reliability needs of storage protocols such as FibreChannel and InfiniBand. TCP/IP-based storage models enabled the migration to lower-cost Ethernet-based network infrastructure, and using a parallel set of storage Ethernet switches became a common method of avoiding competition between storage and data hosts for network bandwidth.

When implementing a dedicated storage network over Ethernet, congestion resulting in dropped frames can still occur, so deploying the full suite of Layer 2 DCB protocols (DCBx, PFS, and ETS) is recommended to maximize storage networking performance.

The diagram below illustrates a dedicated Ethernet storage network deployed in parallel to a data network. Lossless Ethernet protocols are recommended even when using a dedicated storage network.

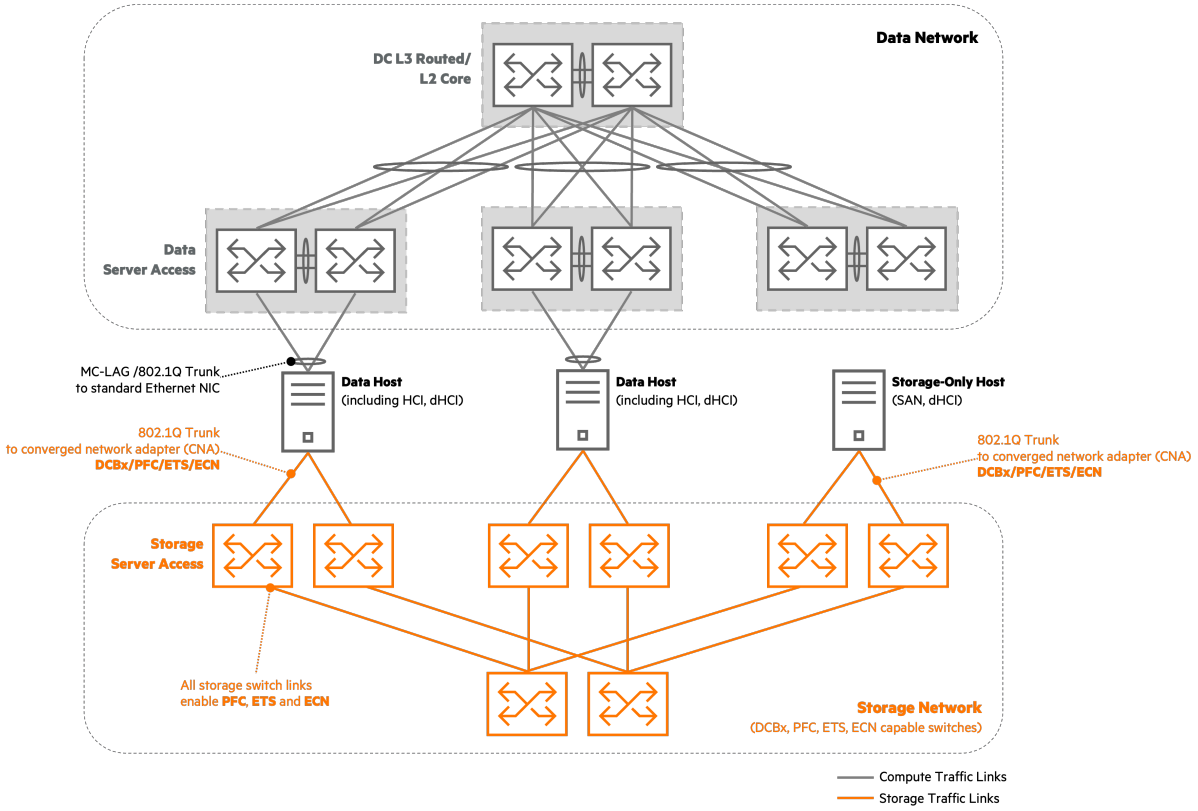


Figure 45: Parallel Storage Network

Converged Data/Storage Network

High speed top-of-rack (ToR) switches with high port density facilitate the convergence of storage and data networks onto the same physical Ethernet switch infrastructure. An organization can maximize its budgetary resources by investing in a single network capable of handling data and storage needs.

A converged storage and data network requires queueing and transmission prioritization to ensure that network resources are allocated appropriately for high-speed storage performance. IP ECN provides additional flow-control options to smooth traffic flow and improve performance. DCBx is beneficial to automate PFC and ETS host configuration.

The diagram below illustrates protocols and positioning to achieve lossless Ethernet in a two-tier data center model.

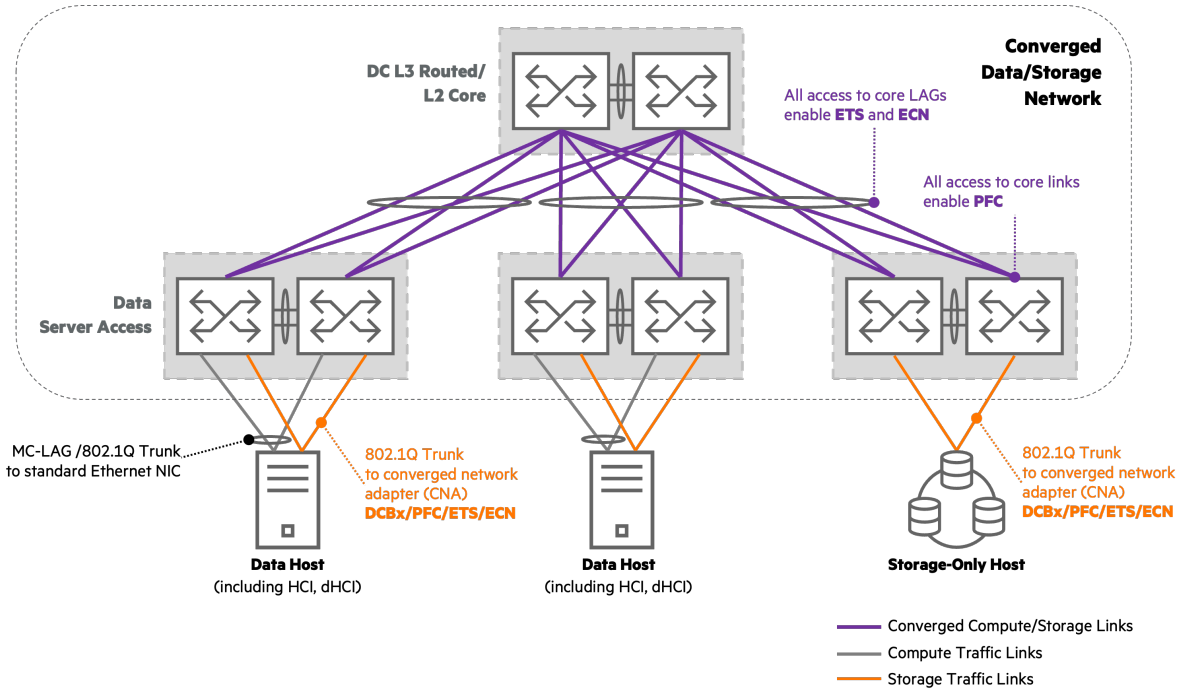


Figure 46: Parallel Storage Network

A spine and leaf network architecture allows linear scaling to reduce oversubscription and competition for network resources. This is achieved by adding spine switches to increase east-west network capacity. Spine and leaf networks use Layer 3 protocols between data center racks, which requires mapping 802.1p priorities to DSCP values to ensure consistent QoS prioritization of traffic across the network infrastructure.

iSCSI

iSCSI is one of the most prevalent general purpose SAN solutions. Standard iSCSI is TCP-based and supports routed IP connectivity, but initiators and targets are typically deployed on the same Layer 2 network. Lossless Ethernet is not a requirement for iSCSI, but it can improve overall performance. Many iSCSI storage arrays using 10 Gbps or faster network cards support PFC and ETS.

When PFC is not supported, LLFC can be used to achieve a lossless Ethernet fabric. Separate switching infrastructure can be implemented to avoid competition between storage and compute traffic, but lossless Ethernet enables the deployment of a single converged network to reduce both capital and operating expenditures.

The following diagram illustrates the components of a converged data and iSCSI storage network.

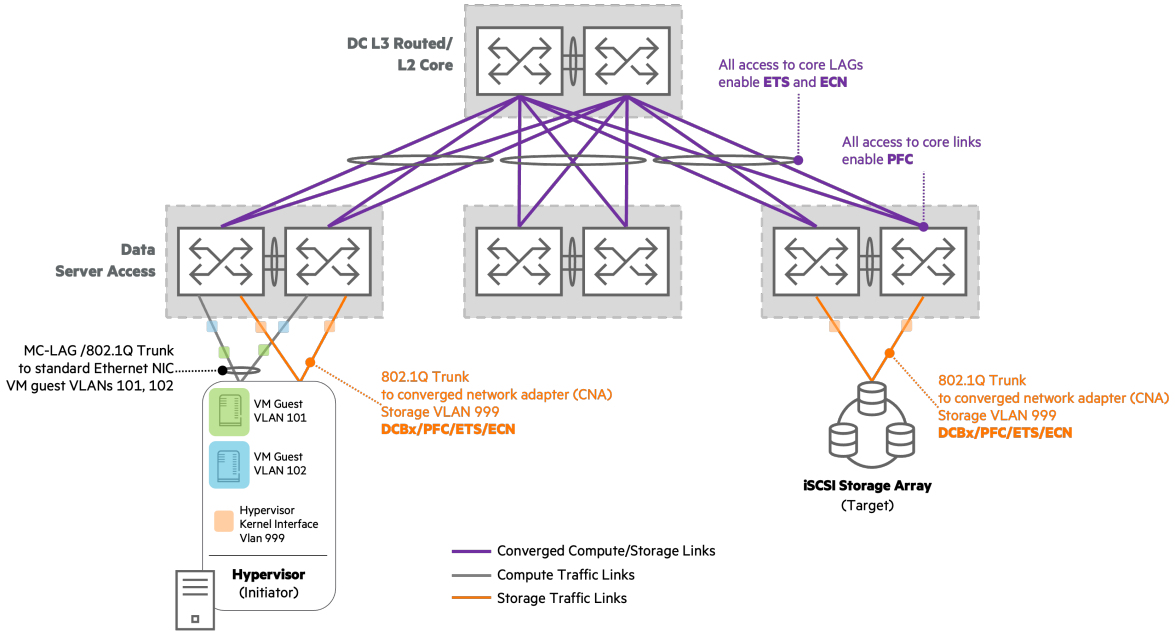


Figure 47: iSCSI L2 Lossless Topology

High Availability

Applications using lossless Ethernet are typically critical to an organization's operations. To maintain application availability and business continuity, redundant links from Top-of-Rack (ToR) switches provide attached hosts continued connectivity in case of a link failure. Use a data center network design that provides redundant communication paths and sufficient bandwidth for the application. The Data Center Connectivity Design guide details network design options.

CX Switch Lossless Ethernet Support

The following illustration summarizes HPE Aruba Networking CX data center switch support for lossless Ethernet and storage protocols, and the feature requirements for common storage protocols.

Feature		Aruba CX 10.12+ supported features					Ethernet-based Storage Network protocols required features							
							iSCSI		RoCE		NVMe-over			
		CX 8100	CX 8360	CX 8325	CX 10000	CX 9300	iSCSI	Lossless iSCSI	RoCE v1	RoCE v2	RDMA RoCE v1	RDMA RoCE v2	RDMA iWARP	TCP
DCB Protocols	Priority-based Flow Control (PFC) IEEE 802.1Qbb & 802.3bd	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		
	Enhanced Transmission Selection (ETS) IEEE 802.1Qaz	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		
	Data Center Bridging Exchange (DCBx) IEEE 802.1AB	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		
L3 Notification Protocol - IP Explicit Congestion Notification (ECN) RFC3168		✓	✓	✓	✓	✓		(*)		✓		(*)		
Lossless Pools per port		1	1	3	3	3								
Lossless PFC priorities per port		2	2	7	7	7								
Asymmetric PFC				✓	✓	✓								

(*) Not a mandatory requirement; just a recommendation.

Figure 48: CX Switch Protocol Support Matrix

HPE Storage Validation for CX Switches

Single Point Of Connectivity Knowledge (SPOCK) is a database that compiles validated compatibility for HPE Storage components, including CX switches. HPE Aruba Networking CX 8325 and CX 9300 series switches have been SPOCK validated and approved by the HPE Storage Networking Team.

Data Center Multicast

Multicast is an efficient method of sending the same traffic to multiple hosts. Traffic from a source can be delivered to a distributed group of interested listeners, allowing a single traffic flow to reach multiple recipients. HPE Aruba Networking supports IPv4 and IPv6 multicast; however, this document focuses on IPv4.

Overview

HPE Aruba Networking data centers support multicast traffic between hosts within a data center and between data center and external hosts. Protocol Independent Mode–Sparse Mode (PIM-SM), PIM–Bidirectional (PIM-BIDIR), Source-Specific Multicast (SSM), and the Internet Group Management Protocol (IGMP) support multicast routing in EVPN-VXLAN and traditional Two-Tier data center architectures.

Multicast traffic is identified as having a destination IP address in the 224.0.0.0/4 IPv4 address block, and it is forwarded between routers in a network based on a dedicated multicast routing table. Each unique address in the reserved block is referred to as a multicast group. Both unicast and multicast traffic use the same IP headers, but routing table maintenance and traffic forwarding decisions have key differences.

A unicast routing table associates unicast IP prefixes with interfaces on the router. When a unicast packet is received by a router, it performs a lookup in the local forwarding table to find the longest prefix match for the destination address in the IP header. An exact prefix match for the destination address is not required, and is likely not present, except in the case of EVPN-VXLAN overlays that share host routes. If the destination IP address does not fall within the range of a prefix installed in the forwarding table, the packet is dropped. A default route that matches all addresses can be assigned to provide a default forwarding path for traffic that does not match a more specific prefix. Only the interface associated with the longest prefix match is selected to forward the received packet. When more than one interface is associated with the same longest prefix, the router selects only one of the interfaces to forward the packet based on a load balancing algorithm.

A multicast routing table associates IP multicast groups with one or more outgoing router interfaces. When a multicast packet is received by a router, it performs a lookup for an exact match of the destination IP address, which is the multicast group. If an exact match is not found, the packet is dropped. If a match is found, the packet is forwarded on all interfaces in a list associated with the multicast group.

Multicast enables sending a single data flow to multiple recipients along multiple network paths. The sender does not need to maintain a traffic flow for each interested listener, but can simply send a single traffic stream with a multicast group destination IP, and all interested hosts can receive the same data. This reduces the CPU burden on the host sourcing data and minimizes traffic utilization in the network for applications that can take advantage of this delivery method. Multicast traffic is commonly used for service discovery, host imaging, telephony, and video applications.

Multicast Components

Multicast routers require a method to efficiently establish forwarding state between sources and interested receivers. Typically, only a small number of hosts in a network are interested in receiving traffic for a multicast group, and those receivers are not aware of source IP addresses.

Similar to building unicast routing tables with protocols like OSPF and BGP, multicast routing combines several control plane protocols and strategies to ensure that multicast flows are delivered to all interested receivers, while only consuming capacity on appropriate links and minimizing router resource consumption.

PIM-SM

PIM-SM is the recommended multicast routing protocol in an HPE Aruba Networking data center, and it is the primary protocol used to build multicast route tables.

In general, PIM-SM routing is scoped to a PIM-SM domain, which is roughly defined as a contiguous set of PIM-SM speaking routers that agree which member of the domain performs the router rendezvous point (RP) function. A PIM domain typically includes the full set of campus and data center switches performing routing functions. PIM-SM operates in both traditional and overlay network architectures.

PIM-SM maintains neighbor adjacencies with directly connected PIM peers. PIM speakers primarily add and remove multicast route table entries after receiving PIM join and prune messages from their neighbors, which communicate interest in receiving traffic for a specific multicast group. Multicast route entries contain two main components for a given multicast group: an incoming interface (IIF) and an outgoing interface list (OIL). When a multicast packet is received on a PIM router, the packet is replicated and forwarded on all interfaces included in the OIL for the group.

PIM routers directly connected to Layer 2 LAN and VLAN segments that contain sources and receivers are referred to as designated routers (DRs).

Rendezvous Point (RP)

In a PIM-SM network, hosts interested in receiving traffic for a multicast group are generally not aware of source IP addresses, so this information is maintained on behalf of the receivers using network infrastructure. One or more PIM-SM routers are assigned the RP role and manage the mapping of multicast groups to their source IP addresses. It is best practice to assign a router loopback IP for the RP function to ensure that the interface is always up.

PIM-SM only permits the assignment of a single IP address to function as the RP for a given range of multicast groups. Typically, the full range of multicast IP addresses (224.0.0.0/4) is assigned to one RP address. In order to provide RP redundancy, an anycast strategy is used, where the same RP IP address is configured on two different physical switches.

Each source for a multicast group is registered with the RP by the source's DR, typically the source's IP gateway, using a unicast PIM register message. The RP sends a Register-Stop message after the information has been added the to RP's source address table. Periodic null register messages from the source's DR maintain state for the multicast group's source address at the RP.

When using an anycast RP for redundancy, it is important to note that unicast PIM register messages are delivered to only one of the anycast RPs. Both switches require a complete set of source addresses to ensure a fully functional environment. The Multicast Source Discovery Protocol (MSDP) solves this problem by sharing multicast group source address entries between the two anycast RPs, which ensures that both switches have a complete database of multicast group to source IPs.

Bootstrap Router (BSR)

All the PIM-SM routers in a PIM domain must be configured with the RP's IP address manually or using an automation method.

The Bootstrap Router (BSR) mechanism, built into the PIM-SM protocol, provides a dynamic method of selecting the RP and distributing the RP's address throughout the PIM domain. The BSR process reduces the administrative burden of configuring RP information manually on all CX switches that perform multicast routing.

The BSR process occurs in two phases. First, an election selects one of the PIM-SM routers configured as a candidate BSR to serve as the BSR for the PIM domain. The identity of the selected BSR is distributed throughout the PIM domain. Second, PIM-SM routers configured as candidate RPs advertise a specific interface IP to the BSR as a possible RP, typically a loopback IP. The BSR selects one of the candidate RPs as the active RP for the domain, and then informs all other members in the PIM domain of the selected RP's address. Candidate RPs should be assigned priority values to influence the RP selection process.

When using an anycast RP, two candidate RPs advertise the same IP loopback address with the same priority value. It is best practice to configure the same two PIM-SM routers as both candidate BSRs and candidate RPs, where both advertise the same candidate RP address and priority value. This minimizes configuration, provides RP redundancy, and simplifies troubleshooting.

Internet Group Management Protocol (IGMP)

IGMP identifies the hosts on a Local LAN or VLAN segment interested in receiving multicast traffic using two methods.

When a host is interested in receiving traffic from a multicast group, it informs IGMP speaking PIM-SM routers of its interest by sending an unsolicited IGMP membership report to a well-known multicast address. The PIM DR for the local network starts building multicast state that allows source traffic to reach the interested listener as described in the following Routed Multicast Flow Setup section.

One or more PIM-SM routers are configured as IGMP queriers for a local LAN or VLAN segment. One of the configured IGMP routers is selected to operate as the active IGMP querier for the local network. The IGMP querier periodically sends solicitations for interest in multicast groups with current state on the router and a general query for all groups. Hosts interested in receiving multicast traffic respond with an IGMP membership report.

In an HPE Aruba Networking data center, the VLAN SVIs of the VSX switches providing IP gateway services are configured as IGMP queriers.

IGMPv3 is the recommended protocol version and is used by default on AOS-CX switches.

IGMP Snooping

IGMP Snooping is enabled on Layer 2 switches to monitor IGMP communications between hosts and multicast routers. By listening to IGMP messages, switches discover local ports with downstream receivers interested in specific multicast groups. Each IP multicast group has a corresponding MAC address. Based on IGMP snooping data, the switch installs multicast group MAC addresses into the Layer 2 forwarding table. This ensures that multicast traffic is forwarded only to hosts interested in receiving traffic for a multicast group on only the necessary Layer 2 links required to reach the receiver, rather than flooding the traffic to all ports in same VLAN.

NOTE:

There is not a one-to-one mapping of MAC addresses to IP multicast groups. The last 23 bits of a multicast group's IP address, which correspond to the range 00:00:00–7F:FF:FF, are appended to the 01:00:5E Organization Unit Identifier (OUI) to create a MAC address that represents the group and can be installed in the Layer 2 forwarding table. This strategy results in MAC address oversubscription, where each MAC address potentially represents 32 IP multicast groups. This should be considered when selecting multicast group addresses for applications to ensure Layer 2 forwarding optimization.

IGMP snooping configuration is recommended on all AOS-CX switches performing Layer 2 operations, when using multicast. This enhances network performance by not consuming capacity on links without downstream receivers. Downstream hosts not interested in the traffic do not need to process the packets.

Routed Multicast Flow Setup

The network path between multicast receivers and sources is referred to as a distribution tree, where the receivers are the leaves of the tree, PIM router links are branches, and the root of the tree is the source.

Multicast route tables specify two types of entries: (*,G) and (S,G). For both types of entries, the "G" represents the multicast group. In a (*,G) route entry, the "*" represents all possible sources for a specific multicast group. In an (S,G) entry, the "S" represents an individual source address.

NOTE:

(*,G) is pronounced “star comma gee” and (S,G) is pronounced “ess comma gee.”

Both multicast route types help build the multicast forwarding table and specify two components per entry: an outgoing interface list (OIL) and an incoming interface (IIF). When routing state exists for a group, multicast traffic received from an upstream source is forwarded downstream toward all interested receivers on every interface included in the combined (*,G) and (S,G) OILs.

The receiver’s DR is responsible for starting the process of building distribution trees and routing state.

Rendezvous Point Tree

When the first interested listener for a multicast group sends an IGMP join for a multicast group, the receiver’s DR has no knowledge of the group’s sources. Since the RP is aware of all multicast sources in a PIM domain, the receiver’s DR begins building a distribution tree toward the RP. This path is referred to as the rendezvous point tree (RPT).

The receiver’s DR adds the interface for the receiver’s network segment to the outgoing interface list (OIL) in a (*,G) route entry. The DR then sends a (*,G) PIM join toward the RP based on reverse path forwarding (RPF), which determines the interface with the closest unicast routing distance to the RP’s IP address. The same interface on which the (*,G) join was sent is added as the incoming interface of the (*,G) route entry.

The upstream PIM-SM neighbor receiving the (*,G) PIM join creates a (*,G) route entry and adds the interface on which the join was received to the OIL for the multicast group. It then sends a (*,G) join toward the RP using RPF to select the nearest interface, again adding the interface as the IIF for the (*,G) route entry. This process is repeated until the RP receives the (*,G) PIM join.

Shortest-Path Tree (SPT)

The RP contains a list of all known sources in the PIM domain, so it is used to facilitate the initial traffic flow between sources and interested receivers over a routed network.

When an RP receives a (*,G) PIM join from a PIM neighbor, it creates a (*,G) route entry for the group and adds the interface for the PIM join to the OIL. It then consults its list of known sources for the requested multicast group. For each known source, the RP builds a shortest-path tree (SPT).

The following process is repeated for each source address known by the RP for the group. The RP adds an (S,G) route entry. The RPF process consults the unicast routing table to determine the interface with the shortest routed distance towards a source. This interface is added as the IIF for the (S,G) route entry, and a PIM (S,G) join is sent to the PIM neighbor on that interface toward the source. The upstream

PIM-SM neighbor receiving the (S,G) PIM join creates an (S,G) route entry and adds the interface it was received on to the OIL for the multicast group, and it then sends an (S,G) join toward the source based on RPF, adding this interface as the IIF for the (S,G) route entry. This process is repeated until the DR for the multicast source receives the (S,G) PIM join.

The following diagram provides a simple example of establishing the initial RPT, and the SPT from the RP to the source, with only a single router in the path for each. After distribution trees are built, traffic flows in the reverse direction of building the PIM multicast routing state.

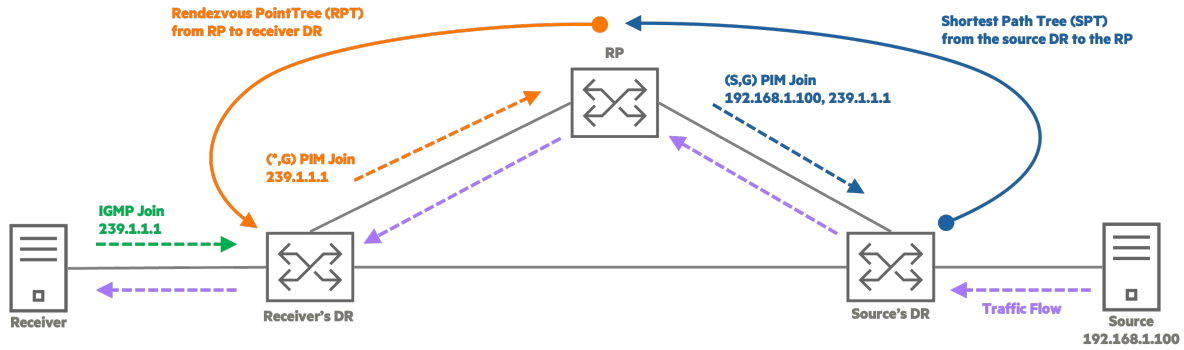


Figure 49: Initial RPT/SPT Build

Routed Multicast Path Optimization

After the SPT is built from the RP to the source, multicast traffic can flow from the source to the interested receiver. Multicast traffic is forwarded from the source to the RP using the SPT. The RP then forwards multicast group traffic toward interested listeners using the RPT.

However, the combined distribution trees through the RP may not be the shortest path between the source and receiver. After multicast traffic from a source arrives at the receiver's DR, a source IP is known for building a more optimized path. The receiver's DR adds an (S,G) route entry for each source observed for a multicast group. It then initiates building an SPT toward an individual source using the same process described above. After the receiver's DR observes traffic on the SPT, the source address is pruned from the RPT by sending a special PIM prune message to the RP that removes only the specified source from the RPT. This permits additional sources to come online and deliver traffic to multicast group receivers using the RPT. After the RP receives the PIM prune message on the RPT, it may prune the SPT for the same source.

Building the SPT from the interested receiver's DR to the source's DR ensures optimal use of network capacity and reduces resource consumption on the RP.

The following illustration depicts optimized traffic flow after cutover to the SPT from the receiver's DR to the source, and the source has been pruned from both the RPT and the SPT path from the RP to the source.

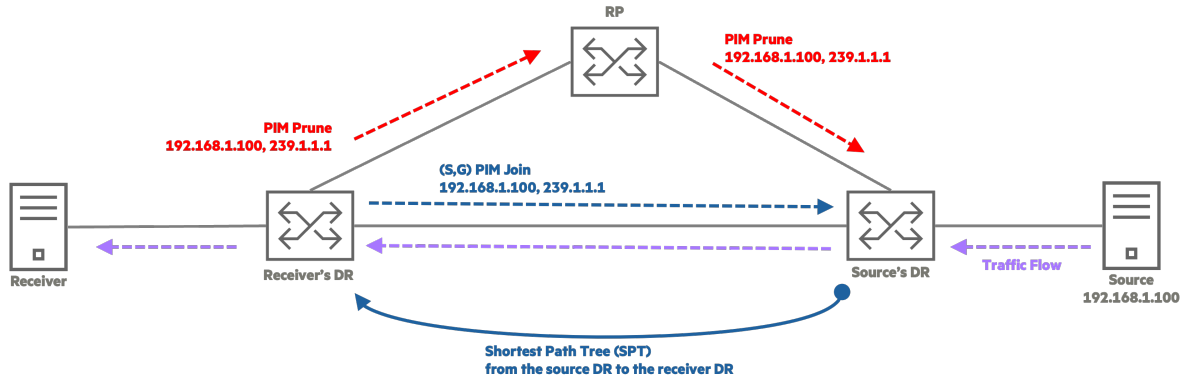


Figure 50: Receiver DR SPT Cutover

VSX Considerations

When implementing PIM-SM and IGMP functions on a VSX pair, each switch operates as different routers that share the same LAN segments.

By default, only one member of the VSX pair functions as the PIM DR for downstream hosts. The DR is responsible for building the RPTs and SPTs for multicast traffic flows. When timely multicast recovery is required following a VSX member failure, a VSX pair can enable PIM active-active mode. In this mode, one member of the pair is designated the DR and the second member is designated the proxy-DR, which is in a backup role. Both the DR and proxy-DR initiate building RPTs and SPTs, but only the DR forwards traffic to interested receivers. This allows for fast recovery in case of DR failure, since multicast traffic is already streaming to the proxy DR. As soon as a DR failure is detected, the proxy DR can begin forwarding multicast traffic without the delay of building new multicast routing state. When implementing a Two-Tier topology, the data center core switches should be configured with active-active PIM.

When using an EVPN-VXLAN overlay with Integrated Routing and Bridging (IRB), host-facing SVI interfaces are configured to operate with the behavior of a DR on both VSX members, irrespective of DR role assignment, by assigning the **ip pim-sparse vsx-virtual-neighbor** command on the interface. IGMP and PIM joins are processed in the same manner by both VSX members, and both members actively forward received multicast traffic to downstream hosts. In an EVPN-VXLAN environment, both VSX members are assigned the same logical VTEP address. Unlike the active-active configuration above, only one member of the VSX pair receives any individual packet as part of a multicast stream from a remote VTEP in the fabric, based on the load balancing algorithm of the underlay network. Therefore, both VSX members must actively forward multicast traffic to ensure delivery to the receiver, and duplicate delivery of of multicast packets to downstream hosts does not occur in this design.

Two-Tier Multicast Operation

In a Two-Tier data center, the core layer switches operate as PIM-SM routers. They support multicast routing between data center hosts and between the data center and external networks. Typically, the data center core switches learn the campus RPs using the BSR mechanism. Active-active PIM provides fast failover of multicast traffic forwarding in case of a core switch failure. The core switch selected as the PIM DR reports directly-connected multicast sources to the campus RP. The core switches also learn about multicast receivers using IGMP.

IGMP snooping optimizes Layer 2 multicast forwarding on both core and access layer switches, so that core switches only forward traffic to downstream access switches with interested receivers, and access switches only forward traffic to ports with directly attached receivers.

The diagram below identifies the multicast features enabled for different roles related to the Two-Tier data center.

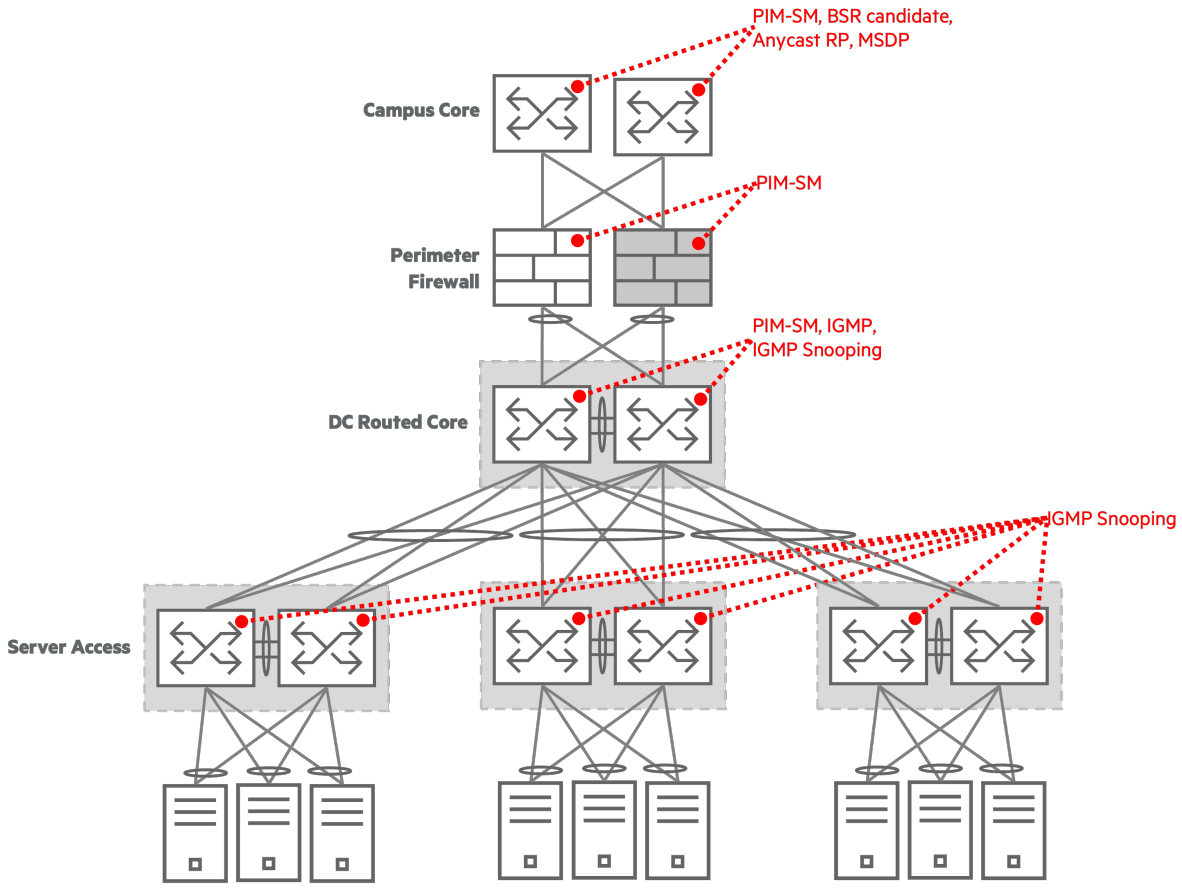


Figure 51: L2 Two-Tier Multicast Roles

RP Placement

When a data center is attached to a campus network, the campus RP can be used by data center switches and learned via the BSR mechanism.

When an external RP is unavailable, an anycast RP can be configured on the Two-Tier routed core to support routed multicast.

EVPN-VXLAN Multicast Operations

HPE Aruba Networking's EVPN-VXLAN implementation enables efficient routing of multicast traffic for single or multiple data center fabrics using PIM-SM. IPv4 multicast forwarding for both Layer 2 and Layer 3 is supported, and additional IGMP and PIM optimizations are implemented to accommodate the overlay network topology.

The diagram below identifies the multicast features enabled for different roles in the EVPN-VXLAN data center. PIM is enabled per overlay VRF and configured on overlay interfaces. IGMP is configured on host-facing overlay interfaces. A separate PIM adjacency is formed between the border leaf switches and firewalls for each overlay VRF.

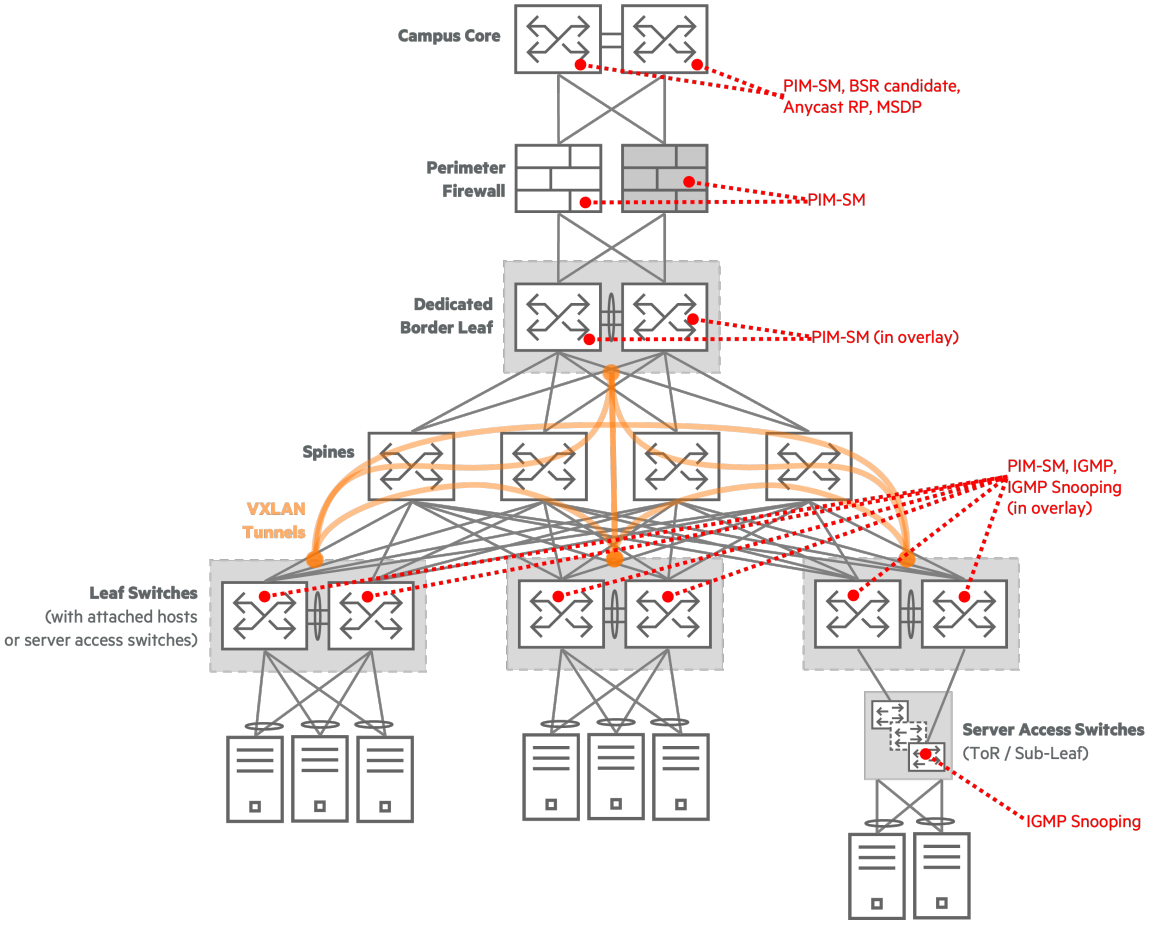


Figure 52: EVPN-VXLAN Multicast Roles

NOTE:

When hosts are positioned at the border leaf, IGMP and IGMP snooping are also required on the border leaf switches.

Native support for Layer 2 and Layer 3 multicast protocols over EVPN-VXLAN tunnels simplifies multicast overlay configuration and troubleshooting. IGMP and PIM-SM optimize IP multicast forwarding, constraining multicast group traffic only to VTEPs with interested listeners. This reduces network traffic and possible congestion.

VXLAN Bridged Multicast

In a VXLAN overlay, Layer 2 multicast traffic is bridged logically between sources and receivers in the same Layer 2 VNI (VLAN) across VTEPs.

By default, when a VTEP receives multicast traffic from an attached source, it replicates and forwards the traffic to all other VTEPs configured with the same Layer 2 VNI. Remote VTEPs receiving the multicast traffic flood it out all ports configured for the VLAN associated with the Layer 2 VNI. All fabric hosts in the same VLAN receive the multicast traffic, irrespective of their interest in the multicast group.

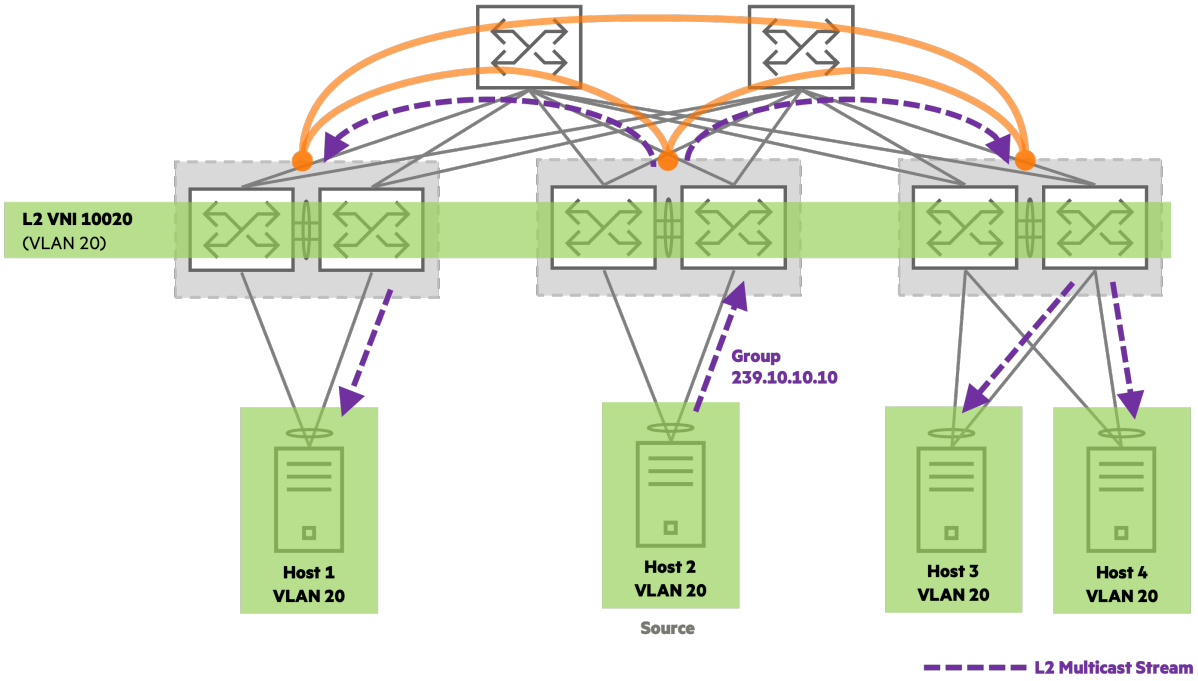


Figure 53: "** VXLAN L2 multicast without IGMP snooping **"

IGMP snooping optimizes Layer 2 multicast forwarding in a VXLAN overlay. When enabled, VTEPs forward IGMP joins and leave messages to peer VTEPs configured with the same L2 VNI on which the IGMP message was received. Each VTEP updates its local Layer 2 multicast forwarding table based on the shared IGMP messages. This enables the source-connected VTEP to forward multicast traffic only to VTEPs with receivers interested in a specific multicast group.

The diagram below illustrates multicast forwarding with IGMP snooping optimizations. One host sends an IGMP join to express interest for a multicast group. The IGMP join is forwarded to all other VTEPs. The source attached VTEP forwards multicast traffic only to VTEPs with an interested receiver. The receiver's VTEP forwards traffic only to the individual host that made the IGMP join.

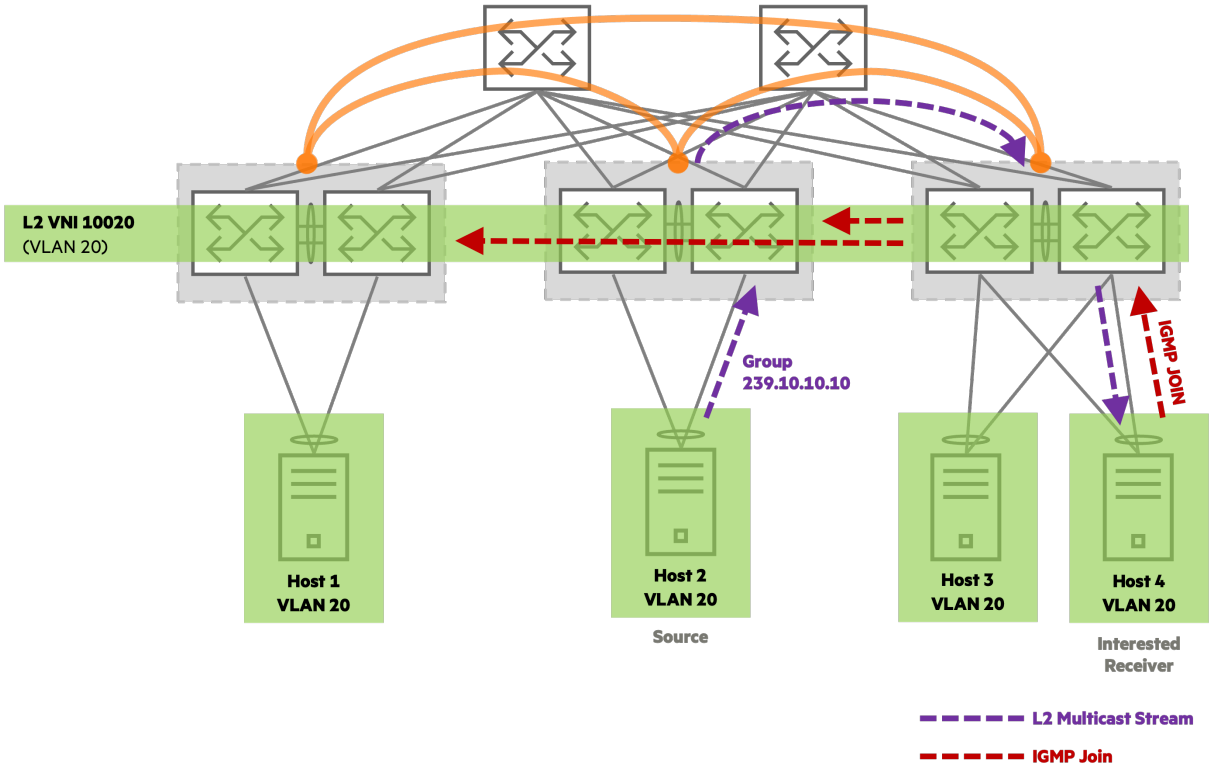


Figure 54: "** VXLAN L2 Multicast with IGMP snooping **"

IGMP Querier Positioning

In an HPE Aruba Networking VXLAN overlay, IGMP joins, leaves, queries, and responses are flooded to all VTEPs in a fabric. When using symmetric IRB for unicast traffic, IGMP queriers are configured on all host-facing VLAN SVIs.

When using a centralized gateway, the IGMP querier should be configured on the centralized gateway and the centralized gateway provides routed multicast functions.

VXLAN Routed Multicast

An IRB deployment with IP gateways local to each VTEP provides better scalability than a centralized gateway. In this model, each leaf switch establishes a PIM adjacency with each remote VTEP in the fabric using a virtual interface associated with the Layer 3 VNI. The result is a full mesh of PIM adjacencies in the data center fabric for each overlay VRF.

When multicast senders and receivers are in different subnets in a VXLAN overlay, Layer 3 multicast traffic is routed between subnets in the same VRF. When the receiver's Layer 2 VNI is present on the source VTEP, AOS-CX multicast uses asymmetric IRB by routing the traffic locally at the source VTEP and sending the traffic to the receiver's VTEP using the appropriate Layer 2 VNI.

The following diagram illustrates routing multicast traffic from a source in VLAN 30 to a host in VLAN 20 in a VXLAN overlay. Layer 2 IGMP snooping optimizations (not depicted) still constrain forwarding at the target VTEP only to hosts that expressed interest in the multicast group.

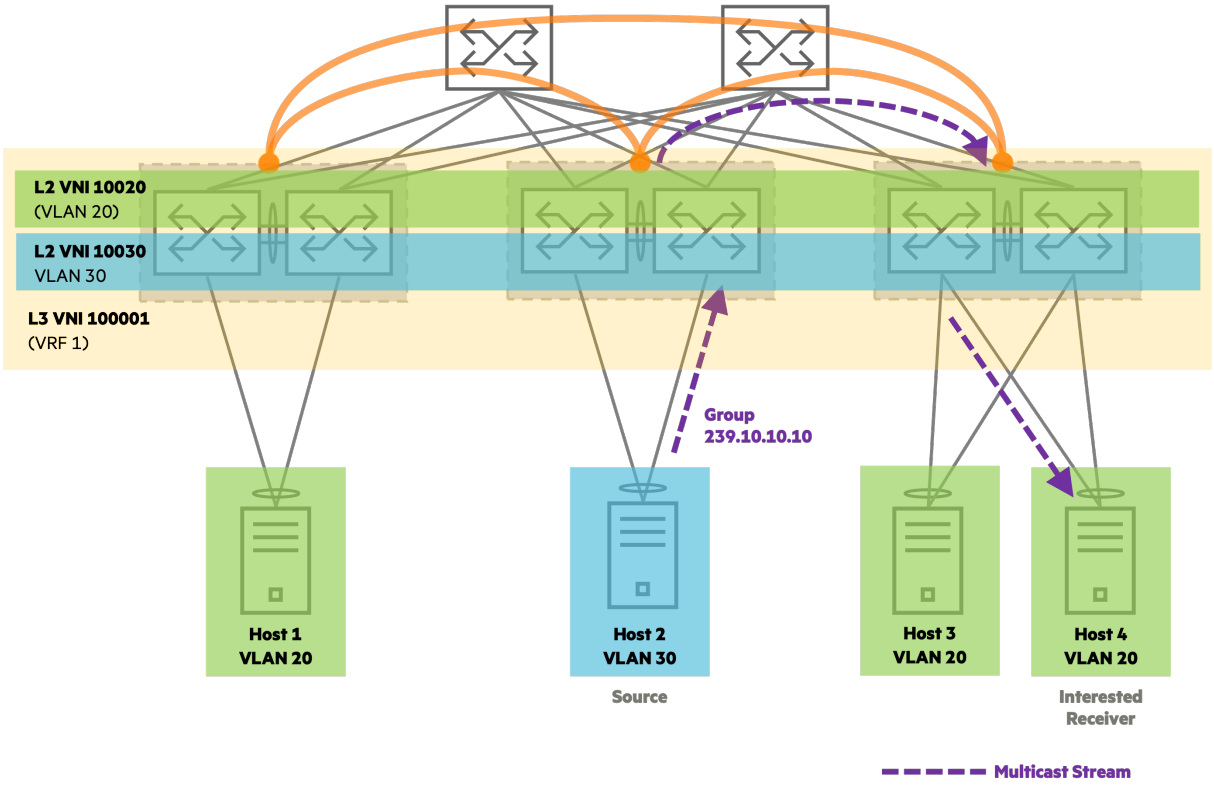


Figure 55: L3 Multicast

When a multicast source begins sending traffic for a multicast group in an EVPN-VXLAN environment, its locally attached VTEP registers the multicast source's unicast IP address for the multicast group with the PIM-SM domain's RP. When unicast routing uses symmetric IRB, the VLAN SVI IP address connected to the source is not unique, so a unique loopback IP in the VTEP's overlay VRF is configured to originate the PIM Register message. When the RP is located external to the EVPN-VXLAN fabric, a route must be advertised to the external network that allows RP communication back to these loopback addresses, so the RP can send PIM Register-Stop messages and build an SPT to the source's DR.

RP Placement

When a data center is attached to a campus network, the campus RP can used by EVPN-VXLAN leaf switches and learned via the BSR mechanism.

When an external RP is unavailable, an anycast RP can be configured on a redundant pair of leaf switches, typically the border leaf. When configuring the RP on a VSX leaf pair, it is common to configure an RP per overlay VRF.

Data Center Management

HPE Aruba Networking supports on-premises and cloud-based options for managing a data center. HPE Aruba Networking Central is a cloud-based service that provides configuration, alerting, and powerful AI-Insights into network communication. HPE Aruba Networking Fabric Composer is on-premises software that automates building EVPN-VXLAN underlays and overlays, orchestrates firewall policy with AMD Pensando Policy Services Manager and access control list (ACL) policy on switches, integrates with VMware vCenter, and provides useful alerting and visualization tools.

Data Center Services Layer

HPE Aruba Networking data center solutions include management plane choices that enable an organization to apply the approach that best suits its needs.

- HPE Aruba Networking Central provides a cloud management solution for the end-to-end networking solution.
- HPE Aruba Networking Fabric Composer is an on-premises fabric automation tool that provides a simplified, workflow-based method of fabric configuration.
- AMD Pensando Policy and Services Manager (PSM) provides management and monitoring of DPU services contained in HPE Aruba Networking CX 10000 switches.
- HPE Aruba Networking NetEdit provides the same multidevice configuration editor and topology mapper now found in Central in an on-premises offering.

Central

HPE Aruba Networking Central is designed to simplify the deployment, management, and optimization of network infrastructure. The use of integrated Artificial Intelligence (AI)-based Machine Learning (ML), and Unified Infrastructure management provides an all-encompassing platform for digital transformation in the enterprise.

Central provides advanced services to facilitate transformational data center rollouts. The NetEdit-style MultiEditor capability integrated into Central makes it possible to deploy complex, multi-device, multilayer configurations from the cloud to the data center. The Network Analytics Engine provides real-time alerts on the state of switches and allows for rapid analysis of intermittent problems. Central is cloud-hosted for elasticity and resilience, which also means that users need not be concerned with system maintenance or application updates.

Workflow-based configurations within Central enable efficient, error-free deployments of HPE Aruba Networking solutions anywhere in the world. The workflows are based on common best-practice approaches to network configuration. They enable new devices to be brought online quickly using new or existing network configurations.

AIOps

According to Gartner Glossary, "AIOps combines big data and machine learning to automate IT operations processes, including event correlation, anomaly detection and causality determination."

HPE Aruba Networking AIOps, driven by Central, eliminates manual troubleshooting tasks, reduces average resolution time, and automatically discovers network optimizations. HPE Aruba Networking's next-generation AI uniquely combines network and user-centric analytics to identify and inform personnel of anomalies. It also applies decades of networking expertise to analyze and provide prescriptive actions.

AI Assist uses event-driven automation to trigger the collection of troubleshooting information, identify issues before they impact the business, and virtually eliminate the time-consuming process of log file collection and analysis. After log information is collected automatically, IT staff are alerted with relevant logs that can be viewed and shared with HPE Aruba Networking TAC, who can assist more quickly with root cause determination and remediation.

Fabric Composer

HPE Aruba Networking Fabric Composer provides API-driven automation and orchestration capabilities for data centers. Fabric Composer discovers data center infrastructure and automates provisioning for both spine-and-leaf fabric and Layer 2 two-tier topologies. Fabric Composer ensures a consistent and accurate configuration of a spine-and-leaf data center with or without deployment of an overlay network.

Fabric Composer orchestrates a set of switches as a single entity called a *fabric* and enables the operator to orchestrate data center resources using an application-centric approach to visualizing network and host infrastructure. It supports managing multiple fabrics and also performs day-to-day operations across rack-scale computing and storage infrastructure.

Visualization of the data center network fabric includes physical and virtual network topologies as well as host infrastructure through integration with AOS-CX, HPE iLO Amplifier, HPE SimpliVity, VMware vSphere, and other leading data center products. In addition to providing a complete view across the fabric, Fabric Composer makes network provisioning accessible to others beside high-level network staff. It provides a secure platform for orchestrated deployment of host and networking resources across the fabric using a guided workflow user interface.

Fabric Composer product integration with vSphere dynamically modifies network security policy by monitoring VM attributes such as IP assignment and VM tags. This automation empowers VMware administrators to add or remove hosts from firewall and ACL policy enforcement by modifying vCenter tags associated with a VM guest.

Fabric Composer is recommended for new data center deployments based on a spine-and-leaf fabric topology. It is particularly helpful when also deploying an EVPN-VXLAN overlay. Fabric Composer configures both underlay and overlay routing automatically using basic IP information provided by the operator.

Fabric Composer facilitates stitching multiple fabrics together to support extending an overlay across multiple locations.

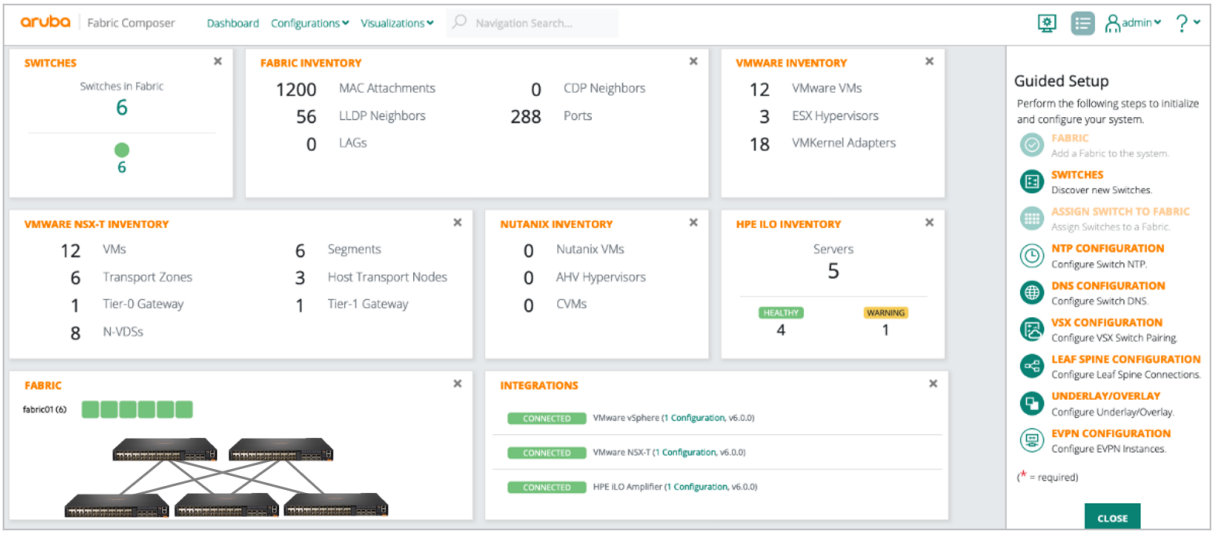


Figure 56: Aruba Fabric Composer

AMD Pensando Policy and Services Manager

The AMD Pensando Policy and Services Manager (PSM) provides an API-based platform for programming and monitoring CX 10000 DPUs. PSM is the firewall policy authority for associated switches.

Fabric Composer integration with PSM enables single-pane-of-glass configuration and orchestration of both the switch fabric and PSM firewall services. PSM also can be managed independently using its web-based GUI or REST API.

NetEdit

HPE Aruba Networking NetEdit helps IT teams automate the configuration of multiple switches and ensure that deployments are consistent, conformant, and error-free. It enables automation workflows without the overhead of programming by providing operators with a user-friendly interface similar to command line. NetEdit also provides a dynamic network topology view to ensure an up-to-date view of the network.

When deploying an HPE Aruba Networking data center using on-premises tools, NetEdit may be deployed for detailed configuration management.

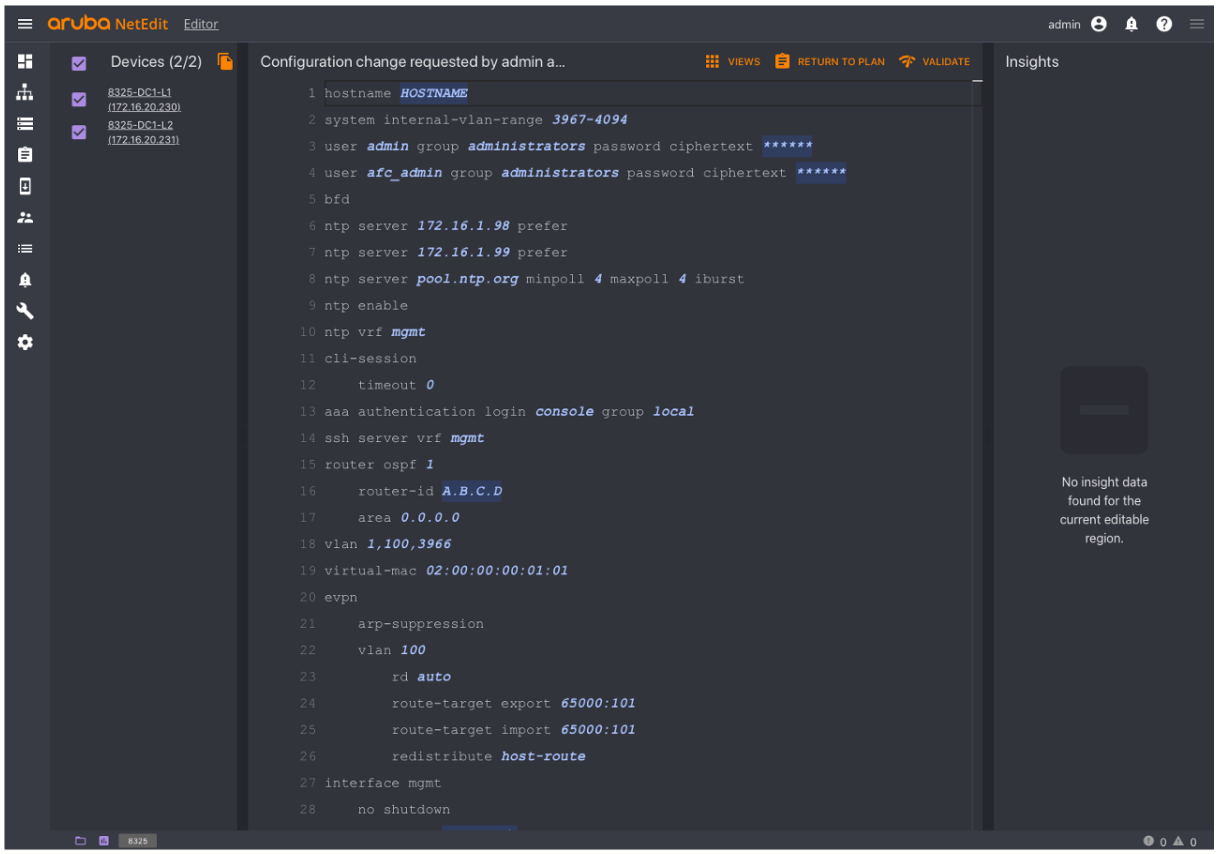


Figure 57: NetEdit

AOS-CX Ansible Collection

The HPE Aruba Networking Developer Hub provides documentation and tooling to support Ansible automation using the AOS-CX Ansible Collection.

Ansible is an open-source orchestration framework maintained by Red Hat. It automates provisioning, configuration management, and application deployment. Ansible playbooks are a powerful and flexible method of automating any CX switch-based topology.

Network Analytics Engine

HPE Aruba Networking's Network Analytics Engine (NAE) provides a built-in framework for monitoring and troubleshooting networks. It automatically analyzes network events to provide unprecedented visibility into outages and anomalies. Using these insights, IT personnel can detect problems in real time and analyze trends to predict or even avoid future security and performance issues

A built-in time-series database provides event and correlation history along with real-time access to network-wide insights to help operators deliver better user experiences. Rules-based, real-time monitoring and intelligent notifications automatically correlate to configuration changes. Integrations with NetEdit and third-party tools such as ServiceNow and Slack provide the ability to generate alerts to trigger actions within the IT service management process.

NAE runs within the AOS-CX operating system in the CX 6xxx, CX 8xxx, CX 9300, and CX 10000 switch series. NAE agents test for conditions on the switch, its neighboring devices, and traffic passing through the network, and then take actions based on test results.

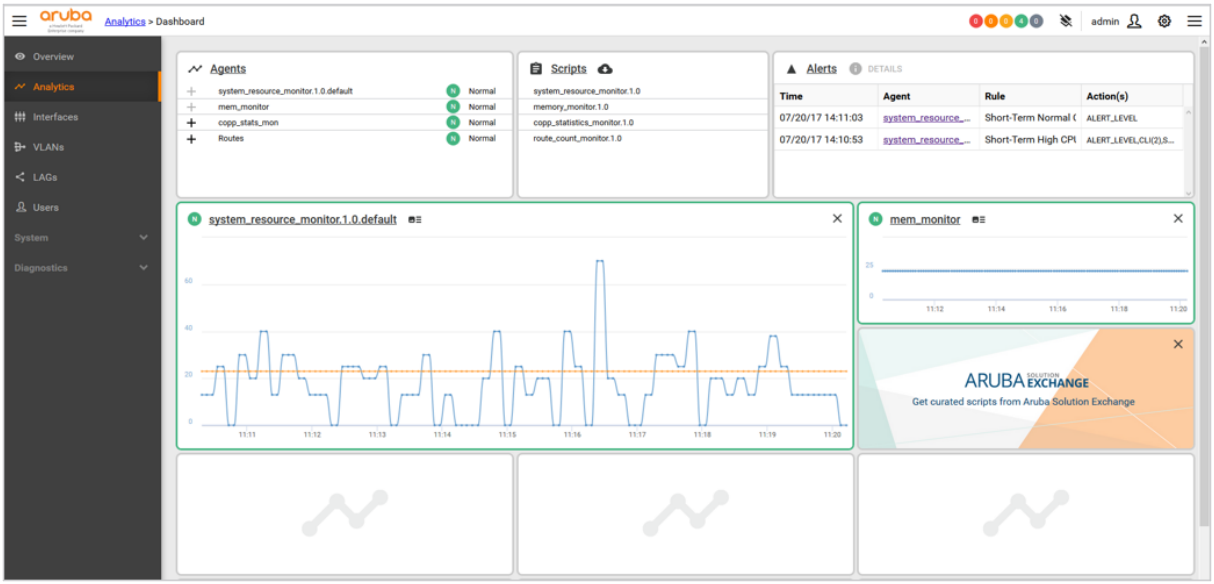


Figure 58: Network Analytics Engine

Choosing an Approach

In general, small, edge-connected data centers are best managed using Central to ensure consistent configuration anywhere in the world.

Layer 2 two-tier data centers can be managed by Central or Fabric Composer. Central provides a single, cloud-based management platform for both campus and data center networks.

Plans to build a spine-and-leaf data center topology should include Fabric Composer. When deploying an EVPN-VXLAN overlay, Fabric Composer is highly recommended to simplify the configuration of underlay and overlay services as well as Layer 3 segments. When deploying PSM with CX 10000 switches, Fabric Composer is recommended to manage firewall rule and policy creation.

Additional Data Center Services

Planning a data center network involves more than just designing the physical network infrastructure. It also is necessary to ensure that services are available to bring switches and hosts online and to ensure that devices can send log messages to a syslog server accessible to people and applications.

It may be useful to leverage the Zero Touch Provisioning (ZTP) capabilities of HPE Aruba Networking switches. To use ZTP, the network must provide a Dynamic Host Configuration Protocol (DHCP) server on a management LAN with a route to the Internet. In addition to the default gateway address, devices also require at least one domain name service (DNS) server to resolve hostnames required for connectivity to Central and the HPE Aruba Networking Activate service.

Network Time Protocol (NTP) ensures that log data from across the network and in the cloud is time-stamped correctly for later analysis. NTP also is required for public key infrastructure (PKI) to function correctly. PKI is required for a variety of access security approaches today. A log management or security information and event management (SIEM) solution also is a part of most modern data centers, which can be used to establish baselines for all switches in the network.

Data Center Reference Architectures

HPE Aruba Networking data center reference architectures support high-availability computing racks using redundant top-of-rack (ToR) switches in EVPN-VXLAN overlay and traditional topologies.

EVPN-VXLAN Spine and Leaf

The HPE Aruba Networking EVPN-VXLAN solution is built on a physical spine-and-leaf topology, which optimizes performance and provides a horizontally scalable design that accommodates data center growth. The Layer 3 links between spine and leaf switches enable adding spine capacity without disrupting existing network components. A data center can start with two spine switches, and then add spine switches in the future when additional capacity is required. The figure below shows the reference architecture with two spine switches and dual-ToR switches.

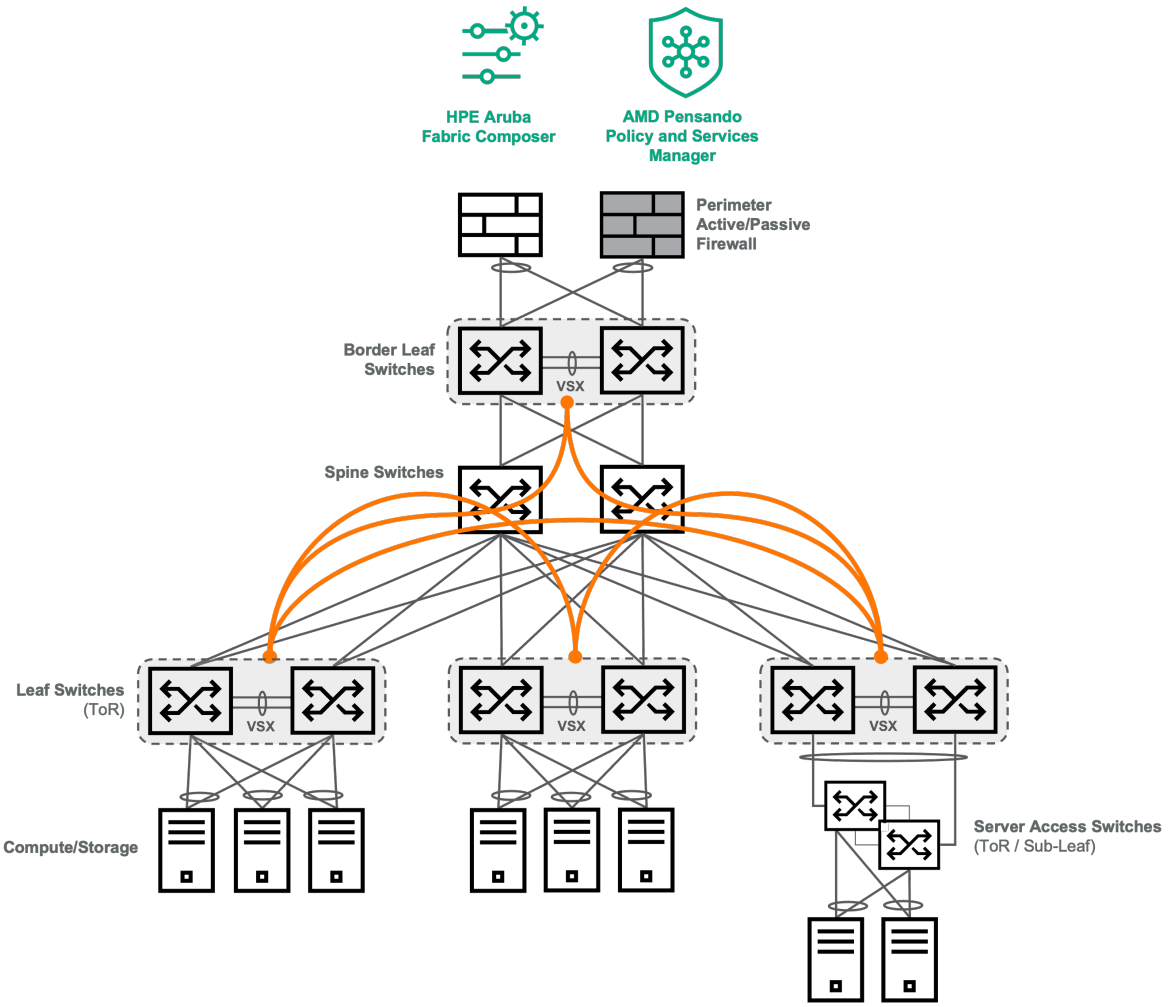


Figure 59: Spine and Leaf: Dual Top of Rack

Certain application environments do not require high availability at the individual computing host. In this case, a single ToR switch per rack provides a more cost-effective data center network. In this type of implementation, host positioning and non-switch redundancy mechanisms must be considered, because a ToR switch under maintenance affects connectivity to all computing hosts in the rack. Spine and leaf deployments can include a mix of both single and dual ToR racks.

Two-Tier

The Two-Tier topology physically resembles a spine-and-leaf design with two spines. Fault tolerance is achieved using multi-chassis Layer 2 link aggregation between the core and access layers, in contrast to the Layer 3 links used in a spine-and-leaf solution. The VSX feature enables upgrading and removing individual switches without disrupting other network components. The core size is fixed at two switches, which makes upgrading physical links and aggregation bundles the primary methods of increasing bandwidth capacity between access and core switches.

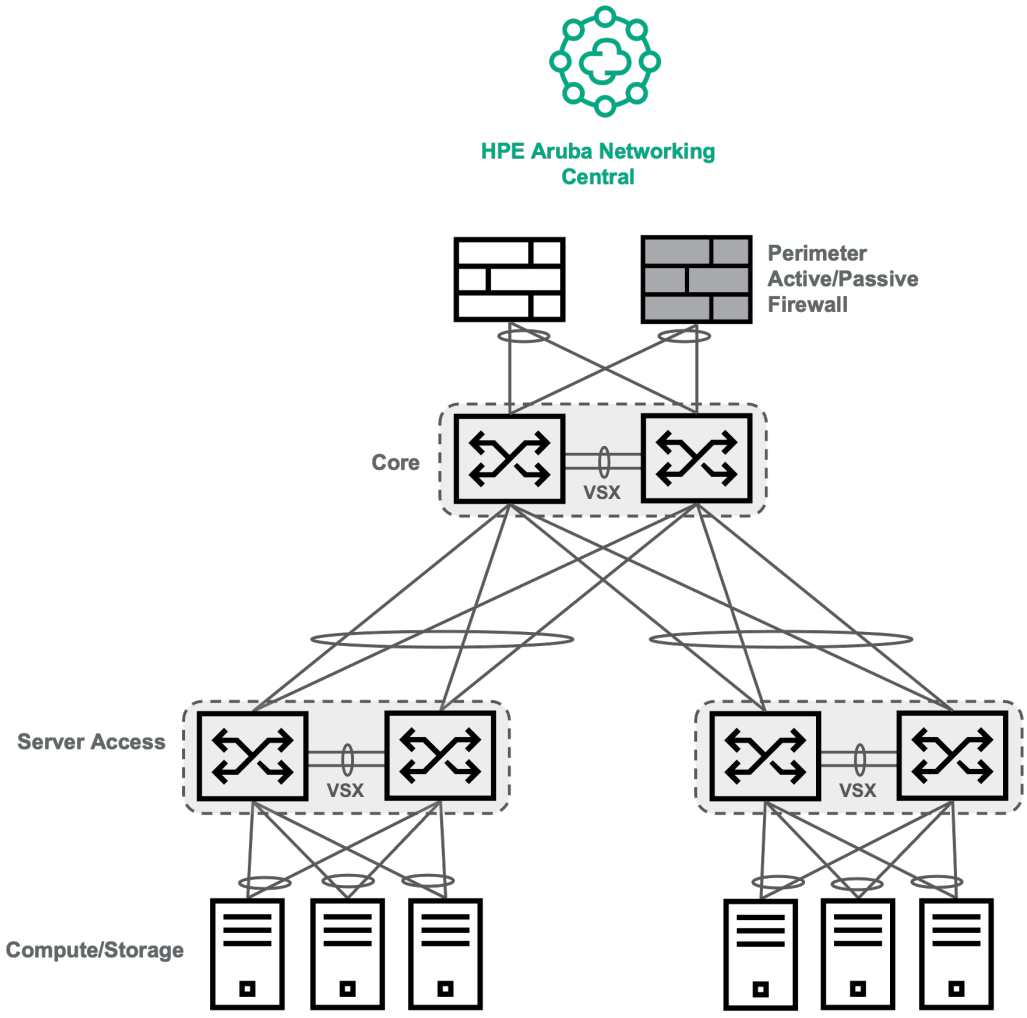


Figure 60: Two-Tier Solution

Reference Architecture Components Selection

The following section provides guidance for hardware selection based on computing host, availability, and bandwidth requirements.

HPE Aruba Networking CX Data Center Switch Overview

The HPE Aruba Networking CX portfolio offers five 1U fixed configuration data center switch models.

- The CX 8325 model offers high ToR port density for 10 and 25 Gbps connected hosts.
- The CX 10000 adds enhanced features along with the same ToR port density.
- The CX 8100 offers high ToR port density for small and medium data centers consisting of 1 and 10 Gbps connected hosts.
- The CX 9300 offers the highest throughput capacity and the most flexibility in a 1U form-factor.
- The CX 9300S offers high throughput ToR capacity for 100 and 200 Gbps connected hosts.
- The CX 8360 model offers a variety of port configurations for small and medium sized topologies.

The CX 10000 distributed services switch (DSS) supports non-switching features to consider when selecting a ToR switch. In addition to inline stateful firewall enforcement and enhanced traffic visibility, it includes IPsec encryption services, DDoS protection, and NAT.

All models offer the following data center switching capabilities:

- High-speed, fully distributed architecture with line-rate forwarding
- High availability and in-service ToR upgrades with VSX
- Cloud-native and fully programmable modern operating system built on a microservices architecture
- Error-free network configuration with software-defined orchestration tools
- Distributed analytics and guided troubleshooting to provide full visibility and rapid issue resolution
- Hot-swappable and redundant load-sharing fans and power supplies
- Front-to-back and back-to-front cooling options for different data center designs
- Jumbo frame support for 9198 byte frames
- Advanced Layer 2 and Layer 3 features to support an EVPN-VXLAN overlay
- Distributed active gateways to support host mobility.

The HPE Aruba Networking CX 6300 model offers an economical Layer 2 ToR for racks with a high number of 1 Gbps connected hosts.

EVPN-VXLAN Solution Switches

The HPE Aruba Networking reference architecture for an EVPN-VXLAN data center includes switches in two roles: spine and leaf.

Spine Switches

The EVPN-VXLAN architecture is built around spine switches with high-density, high-speed ports. The primary function of spine switches is to provide high-speed routed capacity between tunnel endpoints for VXLAN encapsulated traffic. When choosing a spine switch, primary design considerations are:

- Port density
- Ports speeds
- Maximum routes in BGP RIB.

HPE Aruba Networking 1U switches support a range of data center fabric sizes, offering 400 Gbps, 100 Gbps, and 40 Gbps connections to leaf switches.

The CX 9300-32D offers the greatest spine capacity and flexibility in the 1U switch lineup.

- When using a CX 9300S-32C8D leaf switch, a maximum of eight CX 9300-32D spines can connect up to 32 leaf racks in a single ToR switch topology or 16 leaf racks in a dual ToR switch topology using 400 Gbps links. This configuration targets high-speed compute and AI applications using 100 and 200 Gbps connected hosts.
- When using the CX 9300-32D as both spine and leaf switches, it supports up to 32 leaf racks in a single ToR switch topology or up to 16 leaf racks in a dual ToR switch topology using 400 Gbps links over single-mode or multimode fiber optic cable. This configuration supports 400/200/100-Gbps leaf connected compute and AI applications.
- Using the CX 9300-32D as both spine and leaf switches supports extreme horizontal spine scaling. A single ToR topology supports up to 16 spines, and a dual ToR topology supports up to 15 spines, delivering a respective non-oversubscribed fabric capacity of 6.4 Tbps or 6.0 Tbps to each leaf rack.
- The CX 9300-32D spine can double (64 single ToR/32 dualToR) or quadruple (128 single ToR/64 dual ToR) the number of leaf racks supported over its physical port count when using breakout cabling combined with 100 Gbps connections to CX 8xxx and CX 10000 leaf switches. Single-mode transceivers and fiber are required to support four leaf switches per spine port. Two leaf switches per spine port are supported over multimode fiber or when using AOCs.
- The CX 9300-32D spine can support a mix of 400 Gbps links to service leaf racks and 100 Gbps links to standard computing racks to alleviate centralized service congestion points. A CX 9300-32D based spine also provides an upgrade path from 100 Gbps to 400 Gbps for up to 32 leaf switches by replacing a CX 8xxx leaf with a CX 9300 or 9300S switches.

The CX 8325 and CX 8360 offer cost-effective, high-speed spine capacity using 40/100 Gbps links.

- The CX 8325 can support up to 32 leaf racks in a single ToR switch topology or up to 16 computing racks in a dual ToR switch topology.
- The CX 8360 can support up to 12 leaf racks in a single ToR switch topology or up to six computing racks in a dual ToR switch topology.

The table below summarizes the spine SKUs available and their corresponding leaf rack capacity.

SKU	Description	Maximum Leaf Rack Capacity
R9A2	9300-32D: 32-port 400 GbE QSFP-DD, front-to-back airflow	**400G to CX 9300/9300S leaf:**32 single ToR / 16 dual ToR
		100G to CX 8xxx/10000 leaf (single-mode fiber):128 single ToR / 64 dual ToR (400G eDR4 to 4 x 100G FR1)
		100G to CX 8xxx/10000 leaf (multimode fiber or AOC):64 single ToR / 32 dual ToR (400G SR8 to 2 x 100G SR4 or AOC breakout cable)
R9A3	9300-32D: 32-port 400 GbE QSFP-DD, back-to-front airflow	**400G to CX 9300/9300S leaf:**32 single ToR / 16 dual ToR
		100G to CX 8xxx/10000 leaf (single-mode fiber):128 single ToR / 64 dual ToR (400G eDR4 to 4 x 100G FR1)
		100G to CX 8xxx/10000 leaf (multimode fiber or AOC):64 single ToR / 32 dual ToR (400G SR8 to 2 x 100G SR4 or AOC breakout cable)
JL626A	8325-32C: 32-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	32 single ToR / 16 dual ToR
JL627	8325-32C: 32-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	32 single ToR / 16 dual ToR
JL708C	8360-12C v2: 12-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	12 single ToR / 6 dual ToR
JL709	8360-12C v2: 12-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	12 single ToR / 6 dual ToR

The table below lists the optics that support CX 9300 spine connectivity over structured cabling:

SKU	Description	Comments
R9B4	400G QSFP-DD MPO-16 SR8 100m MMF Transceiver	Supports 400G connections between CX 9300 switches over multimode optical fiber. Suports 2 x 100G connections in breakout mode to CX 8xxx/10000 switches using 100G QSFP28 MPO SR4 transceivers (JL309A).

SKU	Description	Comments
R9B42A	400G QSFP-DD MPO-12 eDR4 2km SMF Transceiver	Supports 400G connections between CX 9300 switches over single-mode optical fiber. Suports 4 x 100G connections in breakout mode to CX 8xxx/10000 switches using 100G QSFP28 LC FR1 transceivers (R9B63A).
JL309	100G QSFP28 MPO SR4 MMF Transceiver	When installed in CX 8xxx/10000, supports a 100G connection to CX 9300 400G SR8 (R9B41A) in breakout mode.
R9B63A	100G QSFP28 LC FR1 SMF 2km Transceiver	When installed in CX 8xxx/10000, supports a 100G connection to CX 9300 400G eDR4 (R9B42A) in breakout mode.

The table below lists the available AOC breakout cables for connecting CX 9300 spines to CX 8xxx/10000 leaf switches:

SKU	Description
R9B60A	3m 200G QSFP-DD to 2x QSFP28 100G AOC
R9B58A	7m 200G QSFP-DD to 2x QSFP28 100G AOC
R9B62A	15m 200G QSFP-DD to 2x QSFP28 100G AOC
R9B61A	30m 200G QSFP-DD to 2x QSFP28 100G AOC
R9B59A	50m 200G QSFP-DD to 2x QSFP28 100G AOC

Leaf Switches

The HPE Aruba Networking data center reference architecture primarily uses six models as 1U data center ToR switches.

- The CX 8325 series and CX 10000 switches support high-density host racks using 1 GbE / 10 GbE / 25 GbE ports.

- The CX 9300-32D in a leaf role is intended to connect 100 GbE, 200 GbE, and 400 GbE high-throughput hosts to a CX 9300-32D spine using 400 Gbps links.
- The CX 9300S supports 100 GbE and 200 GbE high-throughput hosts to a CX 9300-32D spine. It also can be optimized for 25 GbE connected hosts. Additionally, the 9300S provides secure border leaf options using high-speed MACsec interfaces.
- The CX 8100 offers high ToR port density for small and medium data centers with 1 GbE and 10 GbE host ports.

- The CX 8360 series offers a variety of models that support 1GbE / 10 GbE RJ45 ports, and flexible variations of 1 GbE, 10 GbE, 25 GbE, and 50 GbE modular transceiver ports.

The CX 10000 distributed services switch (DSS) adds inline firewall features typically provided by dedicated firewall appliances attached to a services leaf or VM hypervisors attached to leaf switches. The CX 10000 also offers IPsec encryption between data centers, NAT, DDoS, and enhanced telemetry services. The CX 10000 switch should be selected when these features are required by downstream hosts or to meet other data center goals. DSS features are not available on other CX switch models. A mix of DSS and non-DSS ToR leaf switch models can connect to a common spine.

Redundant ToR designs require at least four uplink ports for a two-spine switch topology. A minimum of two ports connect to spine switches and two additional ports are members of a high-speed VSX ISL. The CX 9300S is an exception that can connect all eight 400 Gbps uplink ports to spine switches, when using 200 Gbps ports for the VSX ISL. A non-redundant ToR design requires at least two high-speed uplink ports for a two-spine topology.

The table below summarizes the leaf SKUs available and their corresponding supported designs.

SKU	Description	Rack Design	Spine Design
R8P13A	10000-48Y6C: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 6-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR	2–4 switches
R8P14A	10000-48Y6C: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 6-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR	2–4 switches
JL624A	8325-48Y8C: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 8-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR	2–6 switches
JL625A	8325-48Y8C: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 8-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR	2–6 switches
S0F82A	9300S-32C8D: 32-port QSFP28 100G 8p QSFP-DD 400G, front-to-back airflow	High-density / Dual or Single ToR	**400G 9300-32D spine:**2-8 switches
S0F84A	9300S-32C8D: 32-port QSFP28 100G 8p QSFP-DD 400G, back-to-front airflow	High-density / Dual or Single ToR	**400G 9300-32D spine:**2–8 switches

SKU	Description	Rack Design	Spine Design
R9A29	9300-32D: 32-port 100/200/400 GbE QSFP-DD, 2-port 10G SFP+, front-to-back airflow	High-density / Dual ToR	**9300-32D spine:**2–15 switches
		High-density / Single ToR	**9300-32D spine:**2–16 switches
R9A30	9300-32D: 32-port 100/200/400 GbE QSFP-DD, 2-port 10G SFP+, back-to-front airflow	High-density / Dual ToR	**9300-32D spine:**2–15 switches
		High-density / Single ToR	**9300-32D spine:**2–16 switches
JL704(	8360-48Y6C v2: 48-port with up to 22 ports of 50GbE, 44-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 10/25 GbE SFP+/SFP28 with MACsec, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR	2 switches
JL705C	8360-48Y6C v2: 48-port with up to 22 ports of 50GbE, 44-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 10/25 GbE SFP+/SFP28 with MACsec, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR	2 switches
JL706(	8360-48XT4C: 48-port 100M / 1GbE / 10GbE BASE-T, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR	2 switches
JL707C	8360-48XT4C: 48-port 100M / 1GbE / 10GbE BASE-T, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR	2 switches
R9W9(	8100-48XF4C: 48-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR	2 switches
R9W91A	8100-48XF4C: 48-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR	2 switches

SKU	Description	Rack Design	Spine Design
R9W92	8100-40XT8XF4C: 40-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE BASE-T, 8-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR	2 switches
R9W93A	8100-40XT8XF4C: 40-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE BASE-T, 8-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR	2 switches
R9W86	8100-24XF4C: 24-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Medium-density / Dual ToR	2 switches
R9W87A	8100-24XF4C: 24-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Medium-density / Dual ToR	2 switches
R9W88	8100-24XT4XF4C: 24-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE 10GBASE-T, 4-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Medium-density / Dual ToR	2 switches
R9W89A	8100-24XT4XF4C: 24-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE 10GBASE-T, 4-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Medium-density / Dual ToR	2 switches
JL700	8360-32Y4C v2: 32-port with up to 12 ports of 50GbE, 28-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 10/25 GbE SFP+/SFP28 with MACsec, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Medium-density / Dual ToR	2 switches
JL701C	8360-32Y4C v2: 32-port with up to 12 ports of 50GbE, 28-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 10/25 GbE SFP+/SFP28 with MACsec, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Medium-density / Dual ToR	2 switches
JL710	8360-24XF2C v2: 24-port 1/10 GbE SFP/SFP+, 2-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Medium-density / Single ToR	2 switches
JL711C	8360-24XF2C v2: 24-port 1/10 GbE SFP/SFP+, 2-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Medium-density / Single ToR	2 switches
JL702	8360-16Y2C v2: 16-port 1/10/25 GbE SFP/SFP+/SFP28, 2-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Low-density / Single ToR	2 switches

SKU	Description	Rack Design	Spine Design
JL703C	8360-16Y2C v2: 16-port 1/10/25 GbE SFP/SFP+/SFP28, 2-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Low-density / Single ToR	2 switches

NOTE:

Three CX 9300S-32C8D bundles are Trade Agreement Act (TAA) compliant with the same capabilities listed in the table above.S0F81A (front-to-back air flow)S0F83A (back-to-front air flow)S0F87A (back-to-front air flow and DC power supplies)

Server Access Switches

CX 6300 and CX 8100 switches can be used to extend VLANs from a leaf switch to adjacent racks. This strategy provides an economical solution for connecting a rack with a high number of low-speed connected hosts. CX 6300 server access switches are typically connected to CX 8325 or 10000 leaf switches. CX 6300 models support both built-in and modular power supplies

SKU	Description	Power Supplies
JL663	6300M: 48-port 10/100/1000Base-T, 4-port 1/10/25/50 GbE SFP/SFP+/SFP28/SFP56, port/side-to-power airflow	Modular/Redundant
JL762A	6300M: 48-port 10/100/1000Base-T, 4-port 1/10/25/50 GbE SFP/SFP+/SFP28/SFP56 Bundle, back-to-front/side airflow	Modular/Redundant
JL664	6300M: 24-port 10/100/1000Base-T, 4-port 1/10/25/50 GbE SFP56, port/side-to-power airflow	Modular/Redundant
JL658A	6300M: 24-port 1/10 GbE SFP/SFP+, 4-port 1/10/25 GbE SFP/SFP+/SFP28, port/side-to-power airflow	Modular/Redundant
JL667	6300F: 48-port 10/100/1000Base-T, 4-port 1/10/25/50 GbE SFP/SFP+/SFP28/SFP56, port/side-to-power airflow	Built-in/Non-Redundant
JL668A	6300F: 24-port 10/100/1000Base-T, 4-port 1/10/25/50 GbE SFP/SFP+/SFP28/SFP56, port/side-to-power airflow	Built-in/Non-Redundant
R9W9	8100-48XF4C: 48-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Modular/Redundant
R9W91A	8100-48XF4C: 48-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Modular/Redundant
R9W9	8100-40XT8XF4C: 40-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE BASE-T, 8-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Modular/Redundant

SKU	Description	Power Supplies
R9W93A	8100-40XT8XF4C: 40-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE BASE-T, 8-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Modular/Redundant
R9W8	8100-24XF4C: 24-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Modular/Redundant
R9W87A	8100-24XF4C: 24-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Modular/Redundant
R9W8	8100-24XT4XF4C: 24-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE 10GBASE-T, 4-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Modular/Redundant
R9W89A	8100-24XT4XF4C: 24-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE 10GBASE-T, 4-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Modular/Redundant

EVPN-VXLAN Architecture Capacity Planning

The following section provides capacity planning guidance for the HPE Aruba Networking data center spine-and-leaf reference architecture.

Bandwidth Calculations

A spine-and-leaf network design provides maximum flexibility and throughput in a data center implementation. To achieve the greatest level of performance, a spine-and-leaf topology can be designed for zero oversubscription of bandwidth. This results in a data center network that will never be congested because the bandwidth available to hosts is equal to the bandwidth between leaf-and-spine switches.

A significant advantage of a spine-and-leaf design is the ability to add capacity as needed simply by adding additional spine switches and/or increasing the speed of the uplinks between leaf-and-spine switches. A rack with 40 dual-homed servers with 10 GbE NICs could theoretically generate a total load of 800G of traffic. For that server density configuration, a 1:1 (non-oversubscribed) fabric could be built with four spine switches using 4x100 GbE links on each. In practice, most spine-and-leaf topologies are built with server-to-fabric oversubscription ratios between 2:1 and 6:1.

Network and Compute Scaling

The HPE Aruba Networking data center reference architecture provides capacity for most deployments. Distributed gateways and symmetric IRB forwarding optimize fabric capacity consumption. Total fabric capacity can be increased incrementally by adding spines to accommodate growing host compute requirements over time. The CX 10000 DSS switch enables policy enforcement without changing spine-and-leaf traffic optimizations.

The border leaf is typically the node with the highest control plane load since it handles both internal and external connections. Route summarization is a good practice to reduce the redistribution of IP prefixes among domains. Both CX 10000 and 9300S switches support secure border leaf capabilities to external networks and between fabrics.

The HPE Aruba Networking data center reference architecture was tested thoroughly in an end-to-end solution environment that incorporates best-practice deployment recommendations, applications, and load profiles that represent production environments.

Refer to the product data sheets on HPE Aruba Networking Campus Core and Aggregation Switches for detailed specifications not included in this guide.

Two-Tier Solution Switches

The HPE Aruba Networking reference architecture for a Two-Tier data center includes switches in two roles: core and access.

Core Switches

The Two-Tier architecture is built around a pair of core switches with high-density, high-speed ports. The core switches provide fast Layer 2 switching between data center computing racks and all Layer 3 functions for the data center, including IP gateway services, routing between subnets, routed connectivity outside of the data center, and multicast services. The primary design considerations when choosing a spine switch are:

- Port density
- Ports speeds
- MAC address table size
- ARP table size
- IPv4/IPv6 route table size

HPE Aruba Networking 1U switch models support a full range of small to large data center core options.

The CX 9300-32D offers the most capacity and flexibility in the core role of the 1U switch lineup. - When using the CX 9300-32D in both core and access roles, it supports up to 28 computing racks in a single ToR switch topology or up to 14 computing racks in a dual ToR switch topology using 400 Gbps links over single-mode or multimode fiber optic cable. - A CX 9300-32D core can double (56 single ToR/28 dualToR) or quadruple (112 single ToR/56 dual ToR) the number of supported access racks when using breakout cabling combined with 100 Gbps connections to CX 8xxx and CX 10000 access switches. Single-mode transceivers and fiber are required to support four leaf switches per spine port. Two leaf switches per spine port are supported over multimode fiber or when using AOCs.

CX 8325 and CX 8360 offer cost-effective, high-speed core capacity using 40/100 Gbps links. - The CX 8325 can support up to 28 access racks in a single ToR switch topology or up to 14 access racks in a dual ToR switch topology. - The CX 8360 can support up to 8 access racks in a single ToR switch topology or up to four access racks in a dual ToR switch topology.

The table below summarizes the core switch SKUs available and their corresponding access rack capacity, assuming two core ports are consumed per core switch for redundant external connectivity in addition to the two VSX ISL ports.

SKU	Description	Maximum Access Rack Capacity
JL62€	8325-32C: 32-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	28 single ToR / 14 dual ToR
JL627A	8325-32C: 32-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	28 single ToR / 14 dual ToR
JL708	8360-12C v2: 12-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	8 single ToR / 4 dual ToR
JL709C	8360-12C v2: 12-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	8 single ToR / 4 dual ToR
R9A2!	9300-32D: 32-port 400 GbE QSFP-DD, front-to-back airflow	**400G to CX 9300/9300S access:** 28 single ToR / 14 dual ToR
		100G to CX 8xxx/10000 access (single-mode fiber):112 single ToR / 56 dual ToR (400G eDR4 to 4 x 100G FR1)
		100G to CX 8xxx/10000 access (multimode fiber or AOC):56 single ToR / 28 dual ToR (400G SR8 to 2 x 100G SR4 or AOC breakout cable)
R9A30A	9300-32D: 32-port 400 GbE QSFP-DD, back-to-front airflow	**400G to CX 9300/9300S access:**28 single ToR / 14 dual ToR
		100G to CX 8xxx/10000 access (single-mode fiber):112 single ToR / 56 dual ToR (400G eDR4 to 4 x 100G FR1)
		100G to CX 8xxx/10000 access (multimode fiber or AOC):56 single ToR / 28 dual ToR (400G SR8 to 2 x 100G SR4 or AOC breakout cable)

The table below lists the optics that support CX 9300 core connectivity over structured cabling:

SKU	Description	Comments
R9B41A	400G QSFP-DD MPO-16 SR8 100m MMF Transceiver	Supports 400G connections between CX 9300/9300S series switches over multimode optical fiber. Suports 2 x 100G connections in breakout mode to CX 8xxx/10000 switches using 100G QSFP28 MPO SR4 transceivers (JL309A).
R9B42A	400G QSFP-DD MPO-12 eDR4 2km SMF Transceiver	Supports 400G connections between CX 9300/9300S series switches over single-mode optical fiber. Suports 4 x 100G connections in breakout mode to CX 8xxx/10000 switches using 100G QSFP28 LC FR1 transceivers (R9B63A).
JL309A	100G QSFP28 MPO SR4 MMF Transceiver	When installed in CX 8xxx/10000, supports a 100G connection to CX 9300 400G SR8 (R9B41A) in breakout mode.
R9B63A	100G QSFP28 LC FR1 SMF 2km Transceiver	When installed in CX 8xxx/10000, supports a 100G connection to CX 9300 400G eDR4 (R9B42A) in breakout mode.

The table below lists the available AOC breakout cables for connecting a CX 9300-32D core to CX 8xxx/10000 access switches:

SKU	Description
R9B60A	3m 200G QSFP-DD to 2x QSFP28 100G AOC
R9B58A	7m 200G QSFP-DD to 2x QSFP28 100G AOC
R9B62A	15m 200G QSFP-DD to 2x QSFP28 100G AOC
R9B61A	30m 200G QSFP-DD to 2x QSFP28 100G AOC
R9B59A	50m 200G QSFP-DD to 2x QSFP28 100G AOC

Access Switches

The HPE Aruba Networking data center reference architecture includes six access switch models. All models are 1U ToR switches.

- The CX 8325 series and CX 10000 switches support high-density racks using 1 GbE / 10 GbE / 25 GbE host ports.
- The CX 8360 series offers a variety of models supporting 1GbE / 10 GbE RJ45 ports, and flexible variations of 1 GbE, 10 GbE, 25 GbE, and 50 GbE modular transceiver ports.
- The CX 8100 series offers a cost effective model for 1 GbE / 10 GbE connected hosts.

- The CX 9300-32D in an access role is intended to connect 100 GbE and 200 GbE high-throughput hosts to a CX 9300-32D core layer using 400 Gbps links.
- The CX 9300S supports 100 GbE and 200 GbE high-throughput hosts to a CX 9300-32D core, but it also can be optimized for 25 GbE connected hosts.

The CX 10000 distributed services switch (DSS) adds inline firewall features typically provided by dedicated firewall appliances attached to the core or VM hypervisors attached to access switches. The CX 10000 switch should be selected when these features are required by downstream hosts, or to meet other data center goals. DSS features are not available on other CX switch models. A mix of DSS and non-DSS switches connected to a common core is supported.

The table below summarizes the access switch SKUs available.

SKU	Description	Rack Design
R8P13A	10000-48Y6C: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 6-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR
R8P14A	10000-48Y6C: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 6-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR
JL624A	8325-48Y8C: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 8-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR
JL625A	8325-48Y8C: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 8-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR
R9A29A	9300-32D: 9300-32D 32-port 100/200/400 GbE QSFP-DD, 2-port 10G SFP+, front-to-back airflow	High-density / Dual ToR
R9A30A	9300-32D: 9300-32D 32-port 100/200/400 GbE QSFP-DD, 2-port 10G SFP+, back-to-front airflow	High-density / Dual ToR
S0F82A	9300S-32C8D: 32-port QSFP28 100G 8p QSFP-DD 400G, front-to-back airflow	High-density / Dual ToR
S0F82A	9300S-32C8D: 32-port QSFP28 100G 8p QSFP-DD 400G, front-to-back airflow	High-density / Dual ToR
R9W90A	8100-48XF4C: 48-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR
R9W91A	8100-48XF4C: 48-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR
R9W92A	8100-40XT8XF4C: 40-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE BASE-T, 8-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR
R9W93A	8100-40XT8XF4C: 40-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE BASE-T, 8-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR

SKU	Description	Rack Design
R9W86A	8100-24XF4C: 24-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR
R9W87A	8100-24XF4C: 24-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR
R9W88A	8100-24XT4XF4C: 24-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE 10GBASE-T, 4-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR
R9W89A	8100-24XT4XF4C: 24-port 100M / 1GbE / 2.5GbE / 5GbE / 10GbE 10GBASE-T, 4-port 1/10 GbE SFP/SFP+, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR
JL704C	8360-48Y6C v2: 48-port with up to 22 ports of 50GbE, 44-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 10/25 GbE SFP+/SFP28 with MACsec, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR
JL705C	8360-48Y6C v2: 48-port with up to 22 ports of 50GbE, 44-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 10/25 GbE SFP+/SFP28 with MACsec, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR
JL706C	8360-48XT4C v2: 48-port 100M / 1GbE / 10GbE 10GBASE-T, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	High-density / Dual ToR
JL707C	8360-48XT4C v2: 48-port 100M / 1GbE / 10GbE 10GBASE-T, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	High-density / Dual ToR
JL700C	8360-32Y4C v2: 32-port with up to 12 ports of 50GbE, 28-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 10/25 GbE SFP+/SFP28 with MACsec, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Medium-density / Dual ToR
JL701C	8360-32Y4C v2: 32-port with up to 12 ports of 50GbE, 28-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 10/25 GbE SFP+/SFP28 with MACsec, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Medium-density / Dual ToR
JL710C	8360-24XF2C v2: 24-port 1/10 GbE SFP/SFP+, 2-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Medium-density / Single ToR
JL711C	8360-24XF2C v2: 24-port 1/10 GbE SFP/SFP+, 2-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Medium-density / Single ToR
JL702C	8360-16Y2C v2: 16-port 1/10/25 GbE SFP/SFP+/SFP28, 2-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	Low-density / Single ToR

SKU	Description	Rack Design
JL703C	8360-16Y2C v2: 16-port 1/10/25 GbE SFP/SFP+/SFP28, 2-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	Low-density / Single ToR

NOTE:

Three CX 9300S-32C8D bundles are Trade Agreement Act (TAA) compliant with the same capabilities listed in the above table.S0F81A (front-to-back air flow)S0F83A (back-to-front air flow)S0F87A (back-to-front air flow and DC power supplies)

Out-of-Band Management Switches

The HPE Aruba Networking data center reference architecture uses a management LAN built on dedicated switching infrastructure to ensure reliable connectivity to data center infrastructure for automation, orchestration, and traditional management access. The table below lists the recommended switch models.

SKU	Description	Host ports
JL667A	CX 6300F 48-port 1 GbE and 4-port SFP56 Switch	48
JL668A	CX 6300F 24-port 1 GbE and 4-port SFP56 Switch	24
JL663A	CX 6300M 48-port 1 GbE and 4-port SFP56 Switch	48
JL664A	CX 6300M 24-port 1 GbE and 4-port SFP56 Switch	24
JL724A	6200F 24G 4SFP+ Switch	24
JL726A	6200F 48G 4SFP+ Switch	48
JL678A	6100 24G 4SFP+ Switch	24
JL676A	6100 48G 4SFP+ Switch	48

Fabric Composer

HPE Aruba Networking Fabric Composer is offered as a self-contained ISO or virtual machine OVA and can be installed in both virtual and physical host environments as a single instance or as a high-availability, three-node cluster. Fabric Composer can manage EVPN-VXLAN spine-and-leaf fabric and Two-Tier topologies. Fabric Composer is available as an annual per-switch software subscription.

SKU	Description	Supported Switches
R7G99AAE	Aruba Fabric Composer Device Management Service Tier 4 Switch 1 year Subscription E-STU	9300, 10000, 8360, 8325, 6400, 8400
R7H00AAE	Aruba Fabric Composer Device Management Service Tier 4 Switch 3 year Subscription E-STU	9300, 10000, 8360, 8325, 6400, 8400
R7H01AAE	Aruba Fabric Composer Device Management Service Tier 4 Switch 5 year Subscription E-STU	9300, 10000, 8360, 8325, 6400, 8400
R8D18AAE	Aruba Fabric Composer Device Management Service Tier 3 Switch 1 year Subscription E-STU	6300
R8D19AAE	Aruba Fabric Composer Device Management Service Tier 3 Switch 3 year Subscription E-STU	6300
R8D20AAE	Aruba Fabric Composer Device Management Service Tier 3 Switch 5 year Subscription E-STU	6300

The Fabric Composer solutions overview provides additional information.

AMD Pensando Policy and Services Manager

The AMD Pensando Policy and Services Manager (PSM) runs as a virtual machine OVA on a host. PSM requires vCenter for installation. It is deployed as a high-availability, quorum-based cluster of three VMs.

PSM supports CX 10000 series switches. Management of PSM is integrated into Fabric Composer.

PSM can be downloaded from the HPE Networking Support Portal. Entitlement to PSM is included by adding the following required SKU when purchasing a CX 10000 switch.

SKU	Description
R9H25AAE	CX 10000 Base Services License

NetEdit

HPE Aruba Networking's NetEdit software runs as a VM OVA on a host. NetEdit is available from the HPE Networking Support Portal.

Ordering information for NetEdit is provided at the end of this data sheet.

Reference Architecture Physical Layer Planning

The following section provides guidance for planning the physical layer of data center switches.

Cables and Transceivers

Refer to the following documents to ensure that supported cables and transceivers are selected when planning physical connectivity inside the data center:

HPE Server Networking Transceiver and Cable Compatibility Matrix

HPE Aruba Networking ArubaOS-Switch and ArubaOS-CX Transceiver Guide

Interface Groups

For ToR configurations that require server connectivity at multiple speeds, it is important to note that setting the speed of a port might require adjacent ports to operate at that same speed.

CX 8325 and CX 10000 host facing ports have a default speed of 25GbE. Changing the speed to 10GbE will impact groups of 12 ports on the CX 8325 and groups of four ports on the CX 10000. Some CX 8360 switches use interface groups and others support individual port speed settings without impacting adjacent ports. CX 9300-32D switches allow individual ports to operate at different speeds. The CX 9300S 400 Gbps ports support individual speed settings, while the remaining 100G and 200G ports can be assigned two speed modes in interface groups of four.

The following diagram illustrates 9300S port groups:

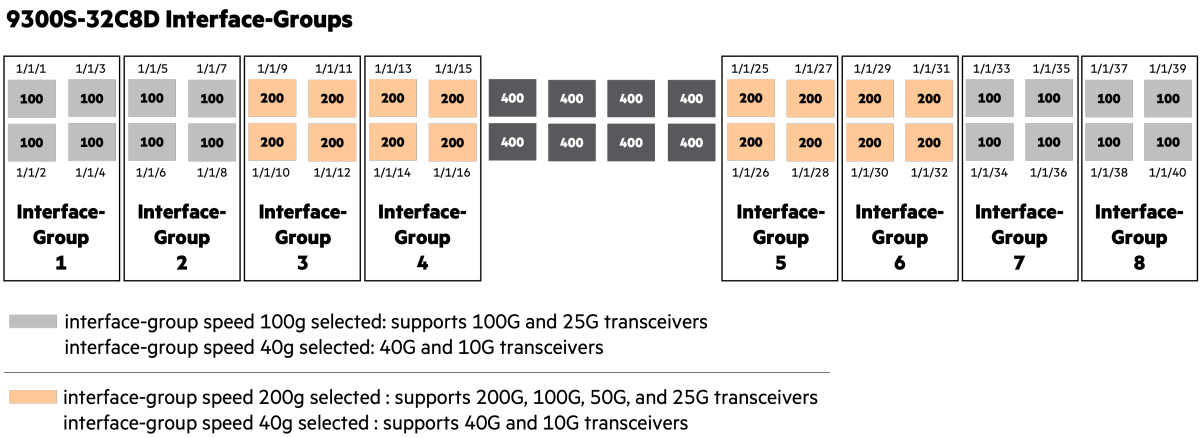


Figure 61: CX 9300S Interface Groups

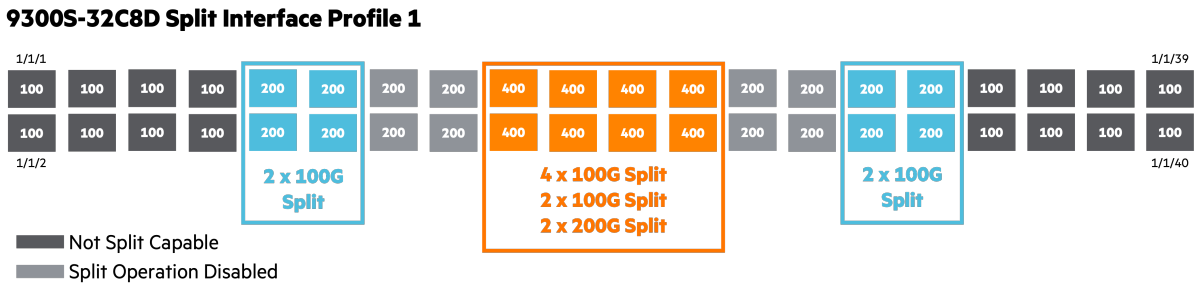
Split Ports

Split ports enable an individual high-speed interface to establish multiple lower speed links using active optical breakout cables or optical transceivers.

The CX 9300-32D can split an individual 400 Gbps port into 4 x 100 Gbps, 2 x 100 Gbps or 2 x 200 Gbps links.

The CX 9300S supports two split interface profile modes that optimize split port capabilities for 100 Gbps or 25 Gbps operational requirements. The default profile (profile 1) optimizes 100 Gbps operation. In this mode, the eight 400 Gbps ports can be split into 4 x 100 Gbps, 2 x 100 Gbps, or 2 x 200 Gbps links, and eight 200 Gbps ports can be split into 2 x 100 Gbps links.

The following diagram illustrates split port operation on the CX 9300s using split interface profile 1 with interface-groups 3 and 6 set to 200 Gbps operation:



:::info Currently shipping HPE Aruba Networking 200G to 2 x 100G AOC split cables support only Q-DD interfaces. These are supported in the CX 9300S 400G interfaces, but not in the 200G QSFP28/56 interfaces. Future cabling options will support 200G to 2 x 100G split operation on CX9300S 200G ports.When a CX 9300S 200G port group is set to 40 Gbps operation in split interface profile 1, the ports are capable of 2 x 10 Gbps split. Split interface profile 2 is recommended when optimizing the 9300S for 25 Gbps or 10 Gbps operation.

:::

The CX 9300S split interface profile 2 optimizes 25 Gbps operation, where six 200 / 100 / 40 Gbps ports can be split into four 25 Gbps links. The number of 400 Gbps ports supporting split operation is reduced to four, when using split interface profile 2.

The following diagram illustrates split port operation on the CX 9300s using split interface profile 2 with interface-groups 4 and 5 set to 200 Gbps operation:

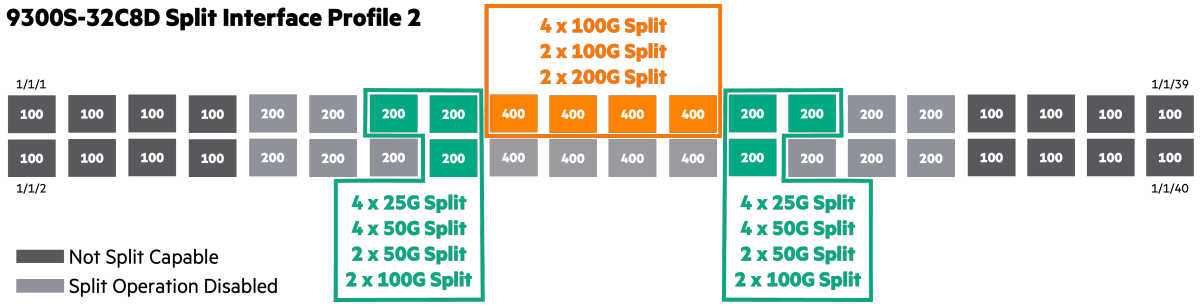


Figure 62: CX 9300S Split Interface Profile 1

NOTE:

When the CX 9300S 200G ports in interface-group 4 or 5 are set to 40 Gbps operation (depicted in green in the diagram above), ports within that group only support 4 x 10 Gbps or 2 x 10 Gbps split operation. The CX 9300S requires a reboot to switch between split interface port profiles.

The QSA28 network adapter (845970-B21) supports 25 Gbps and 10 Gbps optics in QSFP28 ports and 10 Gbps optics in QSFP+ ports. The QSA28 can be used with the CX 9300S to enable lower port speed operation on ports that do not support split operation or have split operation disabled due to the port profile selection.

Most other platforms can split a 40/100 Gbps port into four lower-speed connections (4x10 Gb/s or 4x25 Gb/s).

Refer to the HPE Aruba Networking ArubaOS-Switch and ArubaOS-CX Transceiver Guide when selecting supported breakout cables, adapters, and transceivers.

Media Access Control Security (MACsec)

MACsec is a standard defined in IEEE 802.1AE that extends standard Ethernet to provide frame-level encryption on point-to-point links. This feature is typically used in environments where additional layers of data confidentiality are required or where it is impossible to physically secure the network links between systems.

MACsec can be used to encrypt communication between switches within a data center, between two physically separate data center locations over a data center interconnect (DCI), or between switches and attached hosts.

The table below details MACsec support in the HPE Aruba Networking switch portfolio:

SKU	Description	Number of MACsec Ports
S0F82A	9300S-32C8D: 32-port QSFP28 100G 8p QSFP-DD 400G, front-to-back airflow	16 QSFP+/QSFP28*Future firmware upgrade will provide additional:8 x QSFPDD (400 GbE) ports8 x QSFP28/56 ports*

SKU	Description	Number of MACsec Ports
S0F84A	9300S-32C8D: 32-port QSFP28 100G 8p QSFP-DD 400G, back-to-front airflow	16 QSFP+/QSFP28*Future firmware upgrade will provide additional:8 x QSFPDD (400 GbE) ports8 x QSFP28/56 ports*
JL704C	8360-48Y6C v2: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 6-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	4 SFP+/SFP28,2 QSFP+/QSFP28
JL705C	836048Y6C v2: 48-port 1/10/25 GbE SFP/SFP+/SFP28, 6-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	4 SFP+/SFP28,2 QSFP+/QSFP28
JL700C	8360-32Y4C v2: 32-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 40/100 GbE QSFP+/QSFP28, front-to-back airflow	4 SFP+/SFP28
JL701C	8360-32Y4C v2: 32-port 1/10/25 GbE SFP/SFP+/SFP28, 4-port 40/100 GbE QSFP+/QSFP28, back-to-front airflow	4 SFP+/SFP28

Scale Validation

HPE Aruba Networking's test lab performs multidimensional scale validation of data center architectures. A comprehensive, solution-level test case for each architecture is implemented using recommended best practices.

The validated scale values below represent specific test cases and are not intended to indicate the maximum achievable scale for a specific architecture. The test case is intended to provide a sample reference of achievable scale across multiple switch resources, in contrast to unidimensional data sheet values that specify maximum values for a feature in isolation. Each customer environment is unique and may require optimizing resources in a different manner.

Topology architectures are connected to a high performance testing platform that generates large-scale client traffic.

Spine and Leaf with EVPN-VXLAN Overlay

The spine-and-leaf/EVPN-VXLAN data center was validated using CX 8325-32C spine switches and CX 10000-48Y6C leaf switches.

The following diagram illustrates the HPE Aruba Networking test lab's topology (simulated racks not depicted).

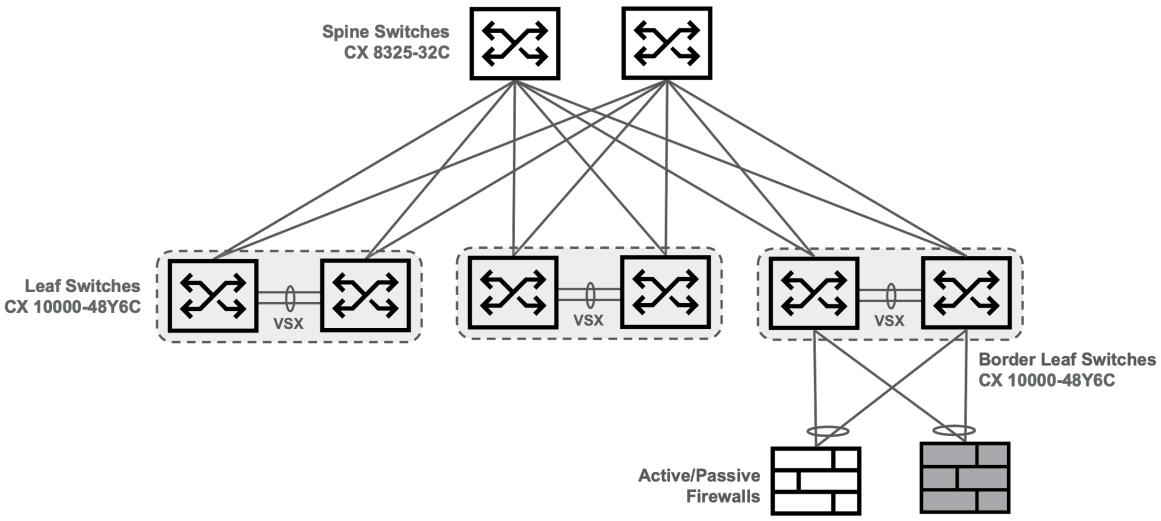


Figure 63: HPE Aruba Networking NTL S&L Test Topology

The underlay uses IPv4 routed-only ports between spine and leaf switches and a single OSPF area to share loopback and VTEP reachability. The testing environment consists of three physical racks with redundant leaf switches and 13 simulated racks to support a total of 16 overlay VTEPs. The testing platform simulates non-redundant leaf switches, resulting in a lower number of underlay OSPF adjacencies than when using a purely physical setup, which does not affect EVPN-VXLAN overlay scale testing parameters.

Layer 2 and Layer 3 overlay scalability were tested. Sixty-four VRFs were defined, each with five VLANs [three standard VLANs, an isolated private VLAN (PVLAN), and a primary PVLAN]. Dual-stacked VLAN SVIs were defined on standard VLANS and primary PVLANs. HPE Aruba Networking's Active Gateway feature provided a dual-stacked, distributed Layer 3 gateway on each leaf switch. Both ARP and ND suppression were enabled.

Two VLAN SVIs per VRF were defined on each border leaf to connect to a pair of external firewalls. Bidirectional Forwarding Detection (BFD) was enabled on external BGP peerings for fast routing failure detection.

Hardware and Firmware

The following switch models and firmware versions were tested in the designated roles:

Switch Role	Switch Model	Firmware Version	Mode	Forwarding Profile
Spine	8325-32C	10.13.1000	Standalone	Spine
Leaf	10000-48Y6C	10.13.1000	VSX	Leaf
Border Leaf	10000-48Y6C	10.13.1000	VSX	Leaf

NOTE:

The internal switch architecture of the 10000-48Y6C is based on the 8325-48Y8C. Validated values for the 10000-48Y6C also apply to the 8325-48Y8C.

Switch Scale Configuration

The following per-switch configuration values established Layer 3 and Layer 2 scale for the testing environment.

Feature	Spine	Leaf	Border Leaf
Underlay OSPF Areas	1	1	1
Underlay OSPF Interfaces	19	3	3
Underlay BGP Peers	19	2	2
Overlay VRFs	N/A	64	64
Overlay VLANs (including one transit VLAN per VRF)	N/A	387	515
Overlay Primary PVLANs	N/A	64	64
Overlay Isolated PVLANs (one per primary)	N/A	64	64
Overlay BGP Peers to External Networks	N/A	N/A	128
BGP IPv4 Route Maps (In + Out)	0	0	128
BGP IPv6 Route Maps (In + Out)	0	0	128
VXLAN EVPN L3 VNIs	N/A	64	64
VXLAN EVPN L2 VNIs	N/A	256	256
Dual-stack overlay external-facing SVIs	N/A	N/A	128
Dual-stack overlay host SVIs	N/A	256	256
SVIs with DHCPv4 Relay	N/A	255	255
SVIs with DHCPv6 Relay	N/A	255	255
Dual-stack Aruba Active Gateway SVIs	N/A	256	256
Unique Active Gateway virtual MACs	N/A	1	1
Host MC-LAG		48	48

Multidimensional Dynamic Table Values

The following table values were populated during the solution test.

Feature	Spine	Leaf	Border Leaf
Underlay OSPF Neighbors	19	3	3
MAC	N/A	38339	38651
IPv4 ARP	19	37288	37543
IPv6 ND	N/A	26374	26758
IPv4 Routes (Underlay + Overlay)	608	37066/1250*	37080/1250*
IPv6 Routes (Overlay)	N/A	26694/640*	26848/656*
Underlay BGP Peers	19	2	2
Loop Protect interfaces	N/A	6976	5568

NOTE:

*The AOS-CX “show ip route” and “show ipv6 route” command outputs include /32 and /128 EVPN host routes, which do not consume a route table entry. In the table above, the first value represents the number of displayed routes when using a show route command. The second number represents the number of actual route entries consumed in the route table during the test.

Two-Tier Architecture

The Two-Tier data center was validated using CX 8360-12C core switches and two types of server access switches: 8360-48XT4Cv2 and 8100-40XT8XF4C. A total of four server access racks were connected to the VSX-redundant core.

The following diagram illustrates the HPE Aruba Networking test lab’s topology (simulated access racks not depicted).

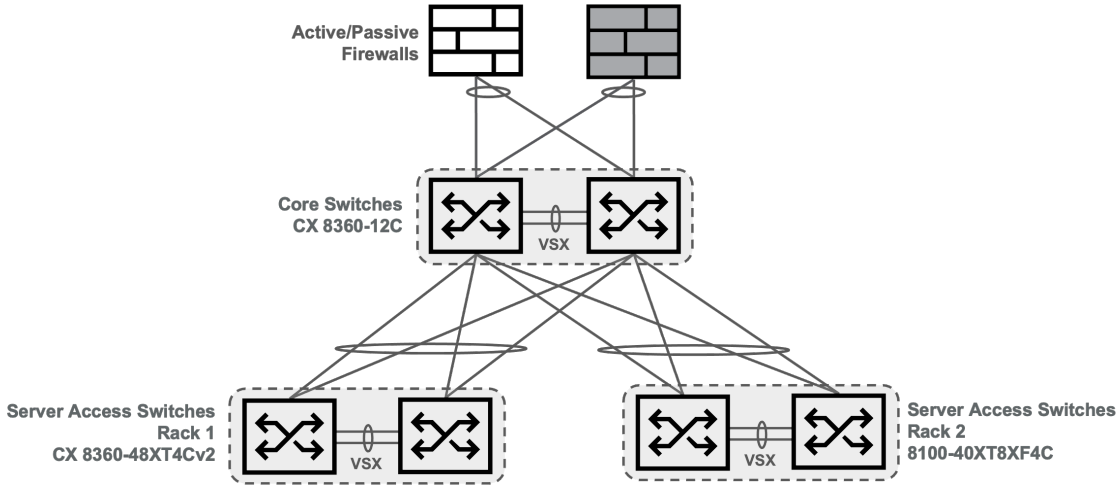


Figure 64: HPE Aruba Networking NTL S&L Test Topology

Four VRFs were defined, with 128 VLANs assigned per VRF (127 server facing VLANs and one transit VLAN). HPE Aruba Networking's Active Gateway feature provided host gateway redundancy on the core switches.

OSPFv2 and OSPFv3 were used for IPv4 and IPv6 routing on a transit VLAN between core switches and external firewalls. BFD was enabled for fast OSPF neighbor failure detection.

PIM-SM, IGMP, and MLD were enabled on core routed interfaces. IGMP and MLD snooping were enabled on server access switches.

MSTP was enabled with a single instance.

Hardware and Firmware

The following switch models and firmware versions were tested in the designated roles:

Switch Role	Switch Model	Firmware Version	Mode	Forwarding Profile
Core	8360-12C	10.13.1000	VSX	Aggregation-Leaf
Server Access	8360-48XT4Cv2	10.13.1000	VSX	Aggregation-Leaf
Server Access	8100-40XT8XF4C	10.13.1000	VSX	N/A

Configured Test Scale

The following per-switch configuration values established Layer 3 and Layer 2 scale for the testing environment.

Feature	Core	Server Access (8360)	Server Access (8100)
VRFs	4	N/A	N/A
ACL Routed VLAN IPv4 Ingress Entries	4096	N/A	N/A
ACL Routed VLAN IPv6 Ingress Entries	4096	N/A	N/A
OSPF Areas	1	N/A	N/A
OSPF Interfaces	8	N/A	N/A
Dual-stack PIM Interfaces	516	N/A	N/A
VLANs	516	512	512
VLAN SVI (dual-stack)	512	N/A	N/A
SVIs with DHCPv4 Relay	511	N/A	N/A
SVIs with DHCPv6 Relay	511	N/A	N/A
Active-Gateway virtual IP (dual-stack)	512	N/A	N/A
Active-Gateway virtual MAC	1	N/A	N/A

Feature	Core	Server Access (8360)	Server Access (8100)
Host MC-LAG	N/A	48	48

Multidimensional Dynamic Table Values

The following table values were populated during the solution test.

Feature	Core	Server Access (8360)	Server Access (8100)
MAC	25109	25600	25600
IPv4 ARP	25109	N/A	N/A
IPv6 ND	49685	N/A	N/A
IPv4 IGMPv3 Groups	1024	256	256
IPv4 Multicast Routes	2036	N/A	N/A
IPv6 MLDv2 Groups	268	67	67
IPv6 Multicast Routes	240	N/A	N/A
PIM-SM Neighbors	516	N/A	N/A
IPv4 Routes	16471	N/A	N/A
IPv6 Routes	5528	N/A	N/A
Dual-stack OSPF Neighbors	8	N/A	N/A
OSPF BFD Neighbors	8	N/A	N/A

www.arubanetworks.com

3333 Scott Blvd. Santa Clara, CA 95054
1.844.472.2782 | T: 1.408.227.4500 | FAX: 1.408.227.4550