The bridge to possible

Cisco Nexus 9000 VXLAN BGP EVPN Data Center Fabrics

Fundamental Design and Implementation Guide

What will you learn?

This paper aims to provide a step-by-step guide to designing and implementing a VXLAN BGP EVPN data center. The data center architect must decide on the hardware, software, scale, protocols, architecture, and network functionality to support the application requirements. The essential decision points and criteria for designing a VXLAN BGP EVPN data center are explained in this paper. VXLAN is an overlay solution with underlay and overlay infrastructure components. Both underlay and overlay have specific technical requirements that must be designed. This paper explains the design of underlay and overlay technology in detail. The goal of this paper is for the reader to be able to design and implement a single VXLAN BGP EVPN fabric with all its essential components according to Cisco's best practices. It is assumed the reader has fundamental knowledge of VXLAN BGP EVPN and is about to embark on an exciting journey to design the data center.

Cisco Nexus 9000 NX-OS Hardware

The proper VXLAN BGP EVPN fabric hardware is not usually your first decision but later. The high-level architecture provides an idea of your network's appearance, where you will run which network function, and what network services and protocols you will use. Once the high-level architecture is finalized, the conversation about where to place which hardware becomes relevant. The hardware always runs software. The software comes with features. The choice of software will come from the features required by the hardware. The software running on networking hardware is the network operating system (NOS). The Nexus Operating System (NX-OS) is the NOS running on the Nexus 9000 switch hardware platform.

The Nexus 9000 switch can be managed via the NX-OS software's command line interface (CLI). The CLI is the traditional approach to performing configuration and system software management. Another option is to use network controller software, such as Nexus Dashboard Fabric Controller (NDFC), to manage the configuration and hardware/software life cycle management.

The VXLAN BGP EVPN fabric is a CLOS network. This network has two main variations: the 3-stage CLOS and the 5-stage CLOS. The 3-stage CLOS consists of leaf and spine nodes, as shown below.

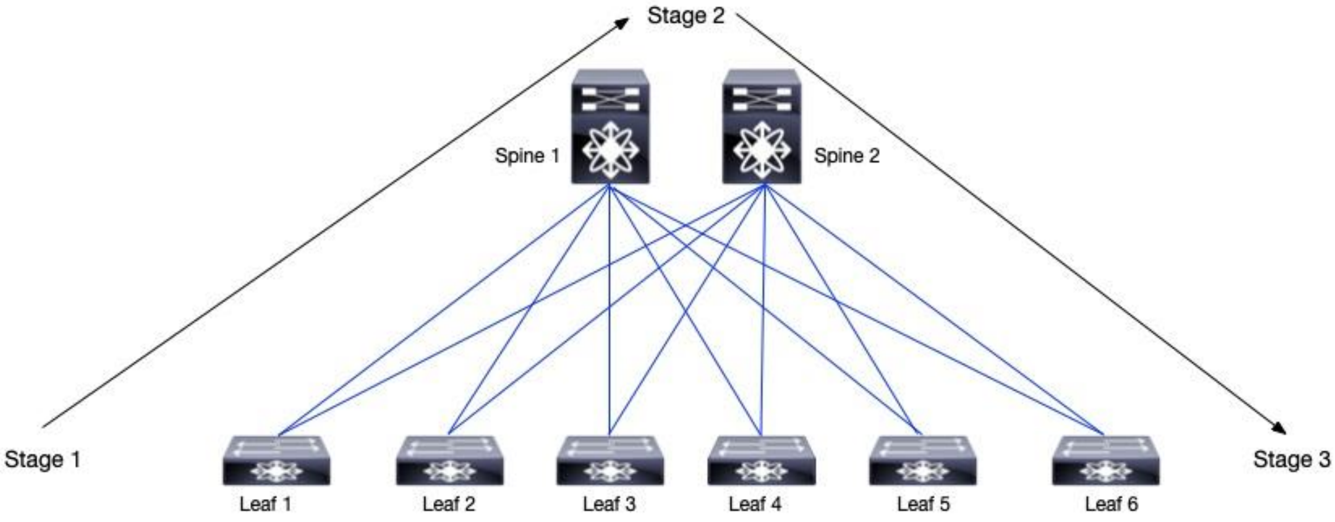


Figure 1. 3-stage CLOS Fabric

The leaf and spine nodes are the two fundamental node types in a 3-stage CLOS fabric. The users, applications, L4-L7 services, and external networks attach to the leaf device. The spine node connects to every leaf node in the fabric. All Internet Protocol (IP) traffic between the leaf nodes transit through the spines. The leaf is like a provider edge node and spine to a core node in service provider networks. Adding leaves allows you to scale out the number of servers in the network, and adding spines enables you to scale out your bandwidth capacity and increase your fabric port density to attach more servers. The more spines a leaf attaches, the lower your oversubscription ratio is.

The 5-stage CLOS fabric adds a super spine node layer that interconnects spines in different locations, allowing your site network site to scale out or compartmentalize.

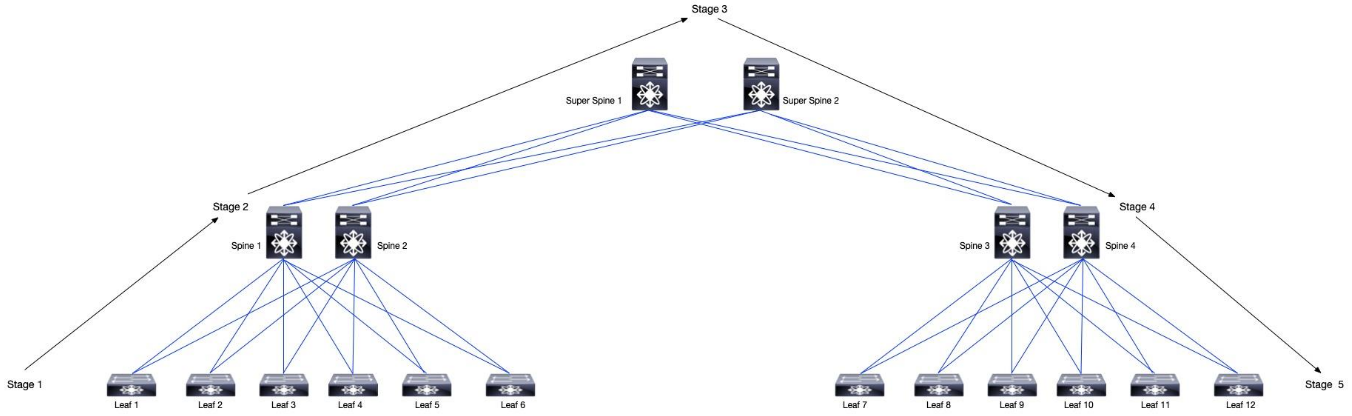


Figure 2. 5-stage CLOS Fabric

As a data center interconnect, the super spine layer can also act as IP transport to interconnect to large data center sites.

The placement of the nodes in the topology and their roles determine the suitable hardware platform, operating system, and configuration for the devices.

Leaf Switches

The leaf devices attach to endpoints or networking devices such as edge routers, firewalls, or load balancers. The type of connected device determines the role of the leaf device. The standard roles for leaf include the following:

Leaf

An edge device that functions as a VXLAN Tunnel Endpoint (VTEP). A host or network device may attach to a VTEP. If a VTEP only has servers attached, some documentation refers to such leafs as server leafs to describe specifically the purpose of the leaf. In Cisco documentation, leaf refers to a VTEP where virtual or physical servers attach to the VXLAN network.

Border Leaf

A VTEP that attaches to the edge router. The edge router can be a wide-area network (WAN) router that connects the VXLAN fabric to external networks such as the campus, Internet, or Internet service provider (ISP). The border leaf is the gateway point from VXLAN to VRF-lite handoff for north-south traffic.

Border Gateway

A VTEP in multisite VXLAN BGP EVPN fabrics is used as a data center interconnect (DCI) node to connect multiple VXLAN sites separated by a layer 3 routed inter-site network. A BGW in a VXLAN site takes packets internal to its attached site, re-originates them, and sends them out its DCI link to the remote site VTEP.

Service Leaf

The leaf switches are attached to L4 – L7 devices. Service refers to security devices operating at OSI Layer 4 – Layer 7, such as firewalls, load balancers, deep packet inspection, and intrusion detection/prevention devices.

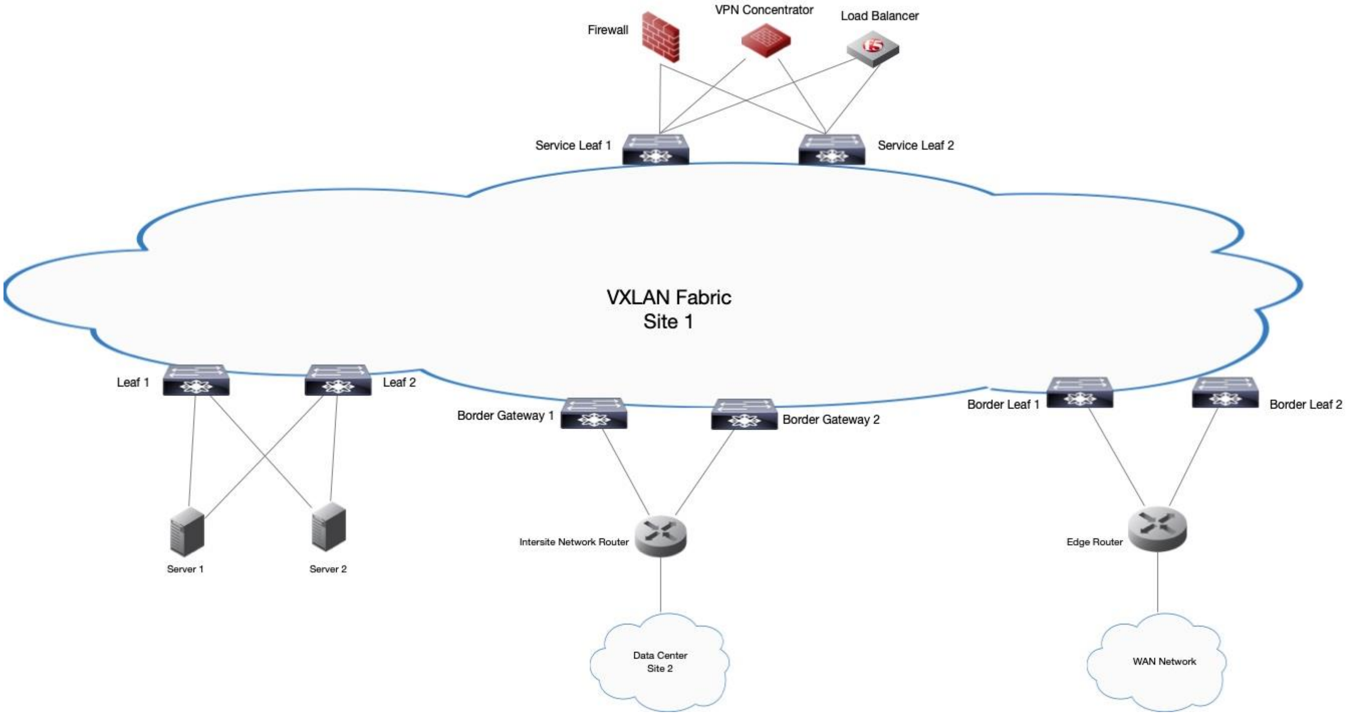


Figure 3. Leaf Roles

Servers can attach to any VTEP. If free ports are available on a Border Leaf, if the right interface types are available on the servers to consume a port, it may. What design factors must be considered when attaching servers to any border leaf? The main factors are the following:

Resource Consumption

Servers require VLANs for bridging and routing. Servers have MAC and IP addresses that must be stored in MAC and ARP tables. The VLANs map to VXLAN Network Identifiers (VNIs), creating MAC-VRFs and IP-VRF tables for MP-BGP EVPN routing and switching. Assume the Cisco Nexus 93180YC-FX3 border leaf attaches to all 48 downlink ports hypervisor-based virtual servers on the Cisco UCS-B blade server platform. Each UCS-B has eight-blade servers with 100 virtual machines per blade server. The total number of virtual machine endpoints attached to the border leaf will be 48 ports X 8 blade servers X 100 = 38,400 virtual machines. Assuming each virtual machine only has a single virtual network interface card (VNIC), 38,400 MAC and host IP entries are added to the MAC, ARP, MP-BGP EVPN, and routing table. Suppose each virtual machine has two VNICs, which doubles the number of learned routes to 76,800 MAC and IP addresses to learn and propagate by the control plane. As you add more servers, the leaf consumes more resources to maintain endpoint routes' control and data plane state. The increase in MAC-VRF and IP VRF table entries also increases the convergence time.

In a large-scale data center, hosting services for hundreds of thousands of users may require them to connect to thousands of external networks. The border leaf TCAM scale for routing must support the programming of all the external networks. If the border leaf node must scale to support connectivity to many external networks, a dedicated border leaf device is recommended for your VXLAN BGP EVPN fabric.

Traffic Profile

As you add more servers to a border leaf device, east-west traffic forwarding increases on the border leaf. The border leaf attached to the external network also handles the forwarding of north-south traffic. The switch's backplane and uplinks must scale to support the traffic bandwidth as more servers and external networks are added. It is critical to understand the application traffic profile in the data center. Suppose you attach big data platforms to your border leaf device, and replication of multiple terabytes of data starts between the servers. At the same time, your data center streams video content to users in the external network. The border leaf is susceptible to becoming a chokepoint and causing impact to east-west and north-south traffic traversing it.

Operational Simplification

Having devices with dedicated roles allows for simplifying configuration management. The configuration applied to the VTEPs is based purely on its single role. If a device only has one role, it only has one function, and the configuration related to the function is only applied. Automation and troubleshooting become easier because every device with its specific role is configured similarly. Change management is cleaner as any configuration changes to WAN-related features will not accidentally impact any servers if no servers are attached to any border leaf devices. Configuration mistakes are the most frequent cause of outages; the more functionality you add to a device, the more complex the configuration, and the higher the chances of configuration mistakes. The easier it is to manage a device, the less chance of human errors causing an outage.

The explanations above focused on the border leaf. Still, the same design factors can apply to border gateway or service leaf depending on the use case and application, security, and traffic profile requirements.

Spine Switches

The fundamental role of the spine device is to provide underlay network connectivity between the leaf nodes for unicast and multicast routing. All the leaf nodes communicate through the spines to route tenant traffic to each other. The spine becomes the transit point for all east-west traffic within the fabric. If the data center attaches to any external network using a border leaf device, the spine becomes a transit point for north-south traffic. If the data center attaches to any DCI network using a border gateway device, the spine becomes a transit point for inter-data center east-west traffic. When the DCI and WAN network connection is decoupled from the spine by attaching it to the border gateway or border leaf nodes, due to the spine's position in the topology, all traffic always traverses the spine. Coupling the spine with other functions, such as border gateway or border leaf, adds more traffic capacity requirements for the spine. The below diagram shows the various traffic patterns traversing the spine based on the role of the spine.

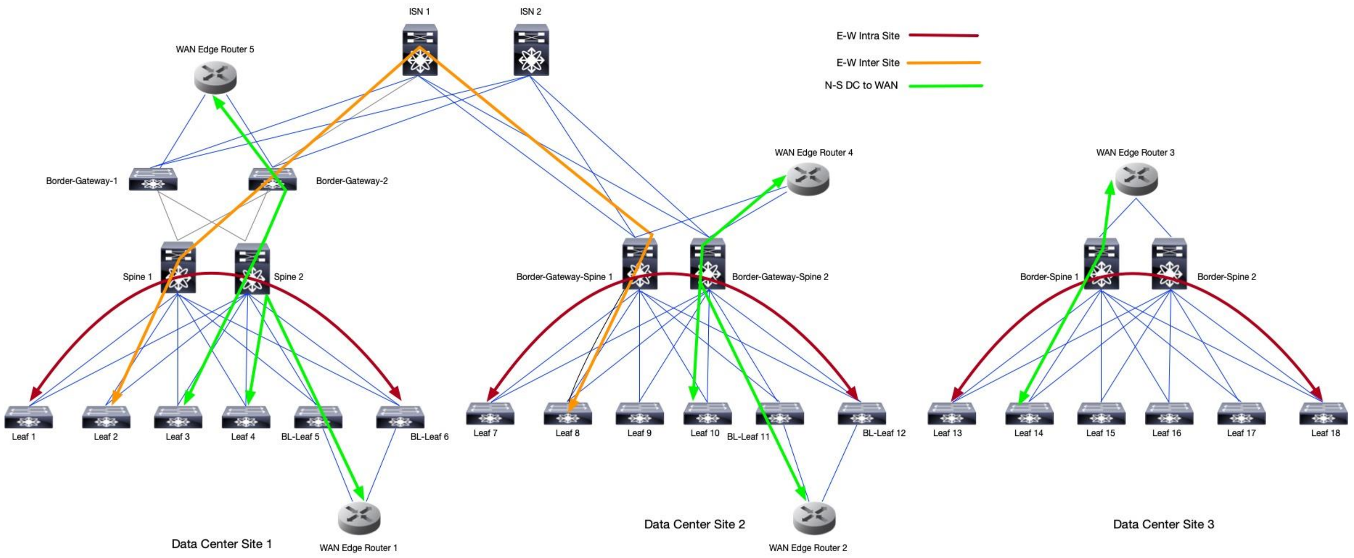


Figure 4. Spine roles traffic patterns

If the spine attaches to a WAN network, it is called a border spine. When the spine becomes the border gateway device connecting to the DCI, it is called a border gateway spine. A border gateway spine may also attach to a WAN network using the single or separate physical interfaces for DCI and WAN networks. It is still important to understand the traffic profile of your applications for all traffic patterns so the right platform is chosen for the spine nodes that can support the bandwidth and minimize impact in failure scenarios such as a link or node failures.

The decision factors to collapse border gateway or border leaf roles into a spine fall into the similar considerations mentioned earlier for leaf nodes, such as resource consumption and operational simplification.

A spine takes overlay routes learned from every leaf and pushes out to all the leaf nodes except the one from which it receives the route. The routes are learned in BGP RIB and stored in random access memory (RAM). The spine node is not a point at which tenant systems attach; therefore, a spine node doesn't have to maintain a VLAN database, ARP, MAC tables, or routing tables for each VRF or VLAN. The VTEP is the connection point to tenant systems, so all configurations and resources are required to learn and forward tenant traffic. The configuration involved in a spine is MP-BGP control plane related, as it is just a point of route distribution to all VTEP routes. The spine is part of the unicast and multicast underlay. The spine node for the unicast underlay is the most logical place to put an MP-BGP EVPN route reflector. The spine node for multicast underlay commonly acts as a rendezvous point (RP). The spine configuration is simple, providing IP transport for VXLAN traffic and distributing routing information between leaf nodes.

The resource consumed is MP-BGP EVPN BGP RIB and the underlay unicast routing table. If PIM is enabled in the underlay, then the multicast routing table is maintained. If multicast routing is enabled in the fabric, the spine must learn and distribute the MP-BGP MVPN routes. Every new overlay service added more address-family, which added more routes to maintain. Every new tenant added to the fabric also increases the routes maintained in the MP-BGP RIB for each VLAN/VRF.

A border gateway spine adds a function to the spine, an entry and exit point for the VXLAN BGP EVPN site. This introduces the responsibility of the spine to re-originate control plane and data plane traffic, making the Border Gateway Spine a point of tunnel termination and origination. A border gateway spine is now also a VTEP with the function of a DCI device. Multi-site VXLAN BGP EVPN-related and VTEP configurations must now be applied to the border gateway spine nodes, including multi-site loopbacks, multisite dci/fabric interfaces, VLANs/VRFs requiring extension across sites, VNI and EVPN-related configurations. Every new capability adds more configurations, increasing the configuration complexity. Increasing configuration complexity increases operational complexity as you have more features to understand, operate, and troubleshoot. Finally, change management risks increase as any configuration mistake made in DCI-related configurations may unintentionally cause internal fabric failures, impacting traffic within the site.

Each VLAN and VRF adds new routing and bridging databases that need to be maintained by the node in addition to existing ones. Maintenance of more control and data plane information increases the consumption of resources, which may eventually impact scale if not designed properly. The added responsibility of acting as a border gateway onto a spine will require increasing the consumption of routing and forwarding resources.

The border gateway and border spine have fully supported capabilities and valid use cases. Design is always about trade-offs. However, we can apply some design principles to decide which one is superior based on technical merit. The architecture design that will provide maximum scalability and operational simplicity is decoupling the border gateway and border leaf from the spine. The simple design principle dedicates roles to essential functions following a modular architecture. Modularity allows each component to perform a single function, simplifying the configuration, troubleshooting, and optimizing system resources. A border gateway device separate from the spine is managed and operated independently while integrated with remote sites and local spines to provide end-to-end connectivity.

Nexus Dashboard Fabric Controller

The network life cycle stages are Day 0, Day 1, Day 2, and Day N. The information technology industry defines each life cycle slightly differently depending on whether software or hardware is being implemented. In a controller-managed network infrastructure, hardware and software are being implemented, as software such as a software-defined networking (SDN) controller is being used to manage the network. The controller must be part of the network's design, implementation, and operation.

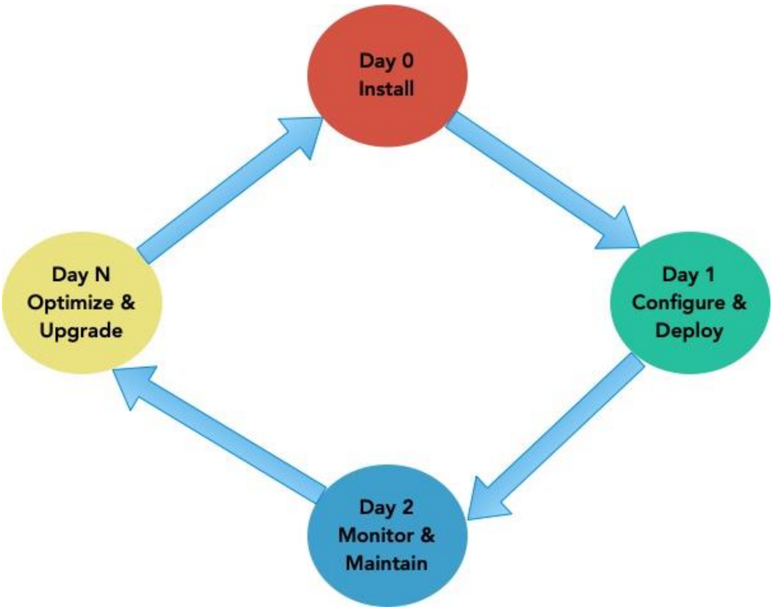


Figure 5. Network Life Cycle

Day 0 to Day N Network Life Cycle

The Day 0 to Day N network life cycle stages can be defined at a high level as follows:

Day 0

The initial network hardware, software, and connections deployment. This stage will involve racking the devices, providing power, connecting the devices with the right cables according to physical topology, provisioning console connections for management, and applying some minimal configuration to establish connectivity to the devices and between the devices. The minimal configuration can be IP addresses for managing networking devices, servers, and controllers and installing security certificates and keys to manage the infrastructure securely. If a controller is being used to automate the network provisioning and management, then a minimal configuration is required to establish connectivity between the controller and network devices. This stage sets the foundation for the network infrastructure.

Day 1

The configurations are applied for an operational network with all the technologies and features implemented per the low-level design. Once the configurations are pushed and applied to the network, the administrators must ensure the network is operational and ready for use. A network ready for use (NRFU) document is commonly used to perform various tests to verify that the network is operating as designed.

Day 2

The network's day-to-day operation, management, and monitoring. Maintenance tasks are also included, such as patching, backups, and security checks, such as bug scans. The monitoring includes Syslog and SNMP, collecting real-time streaming telemetry, and using analytics tools to generate insights to detect and rectify network incidents immediately.

Day N

Improving a network to optimize can mean upgrading existing network device configuration, software, or hardware. It may also involve making design changes to the configuration and logical or physical topology. The network may reach a state where further optimization is impossible, and the entire network needs to be replaced with a new technology or solution. The replacement will involve decommissioning the old infrastructure and replacing it with the new network, transitioning to the Day 0 phase for the new network infrastructure.

Nexus Data Center Fabric Controller (NDFC) provides complete data center life cycle management for Nexus NX-OS switches from day 0 to day N. NDFC includes configuration management, switch software and hardware life cycle management, inventory management, troubleshooting, and device monitoring capabilities.

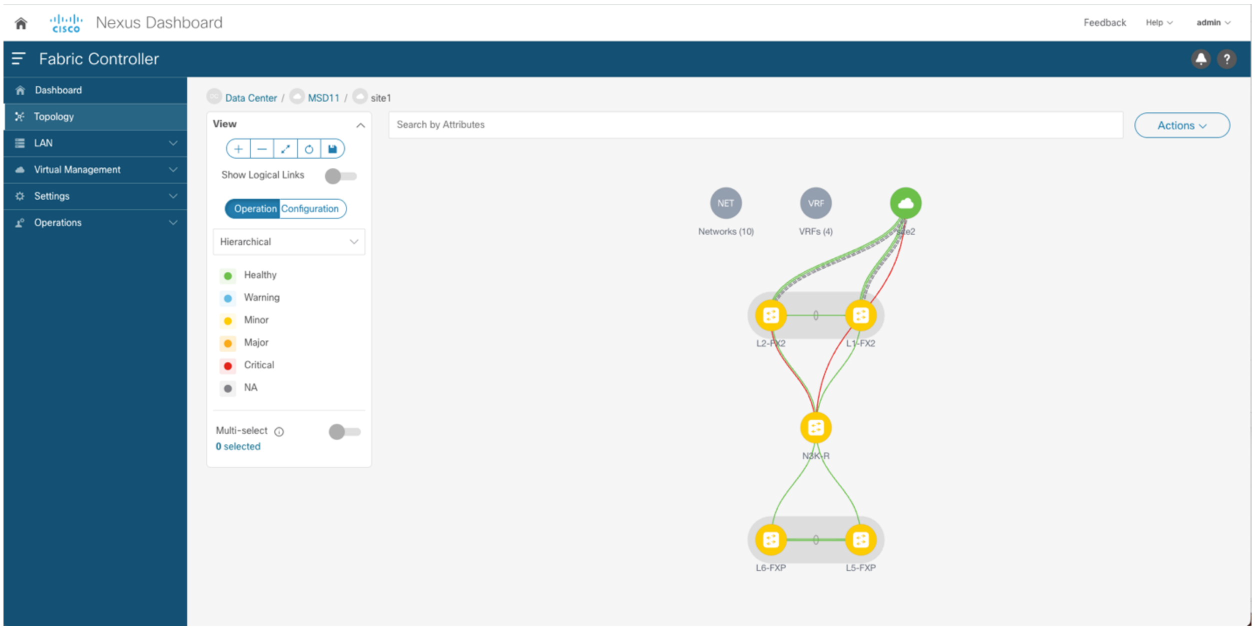


Figure 6. NDFC Topology View

NDFC as a Controller

The NDFC's strength as a controller includes the following:

- **Automation and Management of Multiple types of Data Center Architectures** – There are a variety of data center architectures existing in the data center today. Most of the controllers in the industry have focused on a single type of architecture. A controller for the SP environment will primarily focus on automating the provisioning of MPLS/SR overlay. Today, a controller in a data center may focus on the management and automation of VXLAN BGP EVPN fabrics. NDFC can automate both legacy and modern data center fabric architectures. The legacy data center architectures NDFC can automate include the NX-OS vPC-based 3-tier hierarchical (access-aggregation-core) and collapsed core (access-aggregation). The modern data center fabrics NDFC can automate include VXLAN BGP EVPN for single-site and multi-site networks.
- Multi-protocol support for Data Center Networking – NDFC can automate the provision of various logical topologies in VXLAN EVPN fabrics and legacy data center networks. For VXLAN EVPN fabrics, there are options to provision OSPF, ISIS, or eBGP as an underlay. For the VXLAN EVPN overlay, there are options to provision iBGP or eBGP. Multicast routing or ingress replication can be selected for BUM traffic handling. Various RP redundancy mechanisms, such as PIM Anycast RP or PIM Phantom RP, can be managed. The legacy data center supports protocols such as BGP, OSPF, HSRP, VRRP, and STP best practices. VXLAN over IPsec tunnels extends on-prem tenants to the public cloud, such as AWS and Azure.
- Multi-platform management – NDFC can support all the Nexus platforms, including Nexus 2k/5k/6k/7k/9k, and Cisco MDS. NDFC provides automation templates for various edge and core routing devices that run Cisco IOS-XE and Cisco IOS-XR operating systems.
- Multiple form factors and scale support – The two form factors available for NDFC are virtual ND and physical ND.

The decision to use NDFC or in-house-created automation software with open-source tools such as Ansible is a design decision impacting every phase of the data center management from day 0 to day 2. The choice must be made for the right reasons.

NDFC to Manage your Fabric

Here are some potential reasons for choosing Cisco's NDFC controller to manage your VXLAN BGP EVPN fabric:

- **Cisco Best Practices and Expertise** – Cisco is a leader in networking, specializing in developing network controller software. NDFC development and product management team have invested significant time and resources into understanding the operational requirements of VXLAN BGP EVPN fabrics from hundreds of customers and provided ready-to-use automation templates to implement green field and brownfield VXLAN BGP VPN data centers. NDFC will implement the data center according to Cisco's best practices. What if there is a unique requirement that diverts from Cisco's suggested best practices? NDFC provides free form templates for data center administrators to create custom templates to automate policy according to their needs.
- Save Time and Cost – Developing in-house software requires hiring a team with operations, automation, and programming expertise. Any software product developed requires support and maintenance. All this combined requires both operational and capital expenditure. The time to take the network to the production phase is expedited with NDFC. Cisco has already spent time and money to develop NDFC to save its customer time and cost to implement their data center fabrics. Cisco handles software defect fixes and feature enhancements, ensuring the software is secure and continuously improving with innovations.
- Cisco Customer Experience (CX) team Support – Cisco offers Technical Assistance Center (TAC) support for all their controller software, including NDFC. The Cisco TAC team provides technical support to resolve any reactive issues in production. The professional services team under CX also has data center architects experienced in NDFC and VXLAN BGP EVPN to assist Cisco customers with design and implementation.
- **Cisco Partner Support** – Cisco has thousands of partners with a presence globally, providing sales and consulting services for data center solutions, including NDFC.

The combined efforts and collaboration between Cisco engineering, CX, and Cisco partners ensure the successful adoption of NDFC in data center networks.

The NDFC controller does have open APIs that any programming or automation system may consume to automate VXLAN BGP EVPN fabrics managed by NDFC. The NDFC API details are available on Cisco Developer Portal at the following link: https://developer.cisco.com/docs/nexus-dashboard-fabric-controller/latest/#!introduction/introduction.

NDFC does have Ansible modules available. The Cisco Developer portal provides a getting started guide at the following link: https://developer.cisco.com/docs/nexus-as-code/#!ndfc-with-ansible. The open APIs and Ansible modules available in NDFC allow internal teams within the IT organization to use existing automation tools to integrate with NDFC.

VXLAN BGP EVPN Underlay Unicast Routing

The underlay network in a VXLAN BGP EVPN fabric is an IP-configured routed network that provides connectivity between the VTEPs to forward unicast traffic between the VTEPs in VXLAN tunnels. The forwarding of traffic between VTEPs requires routing protocols to exchange routing information and form a point-to-multipoint VXLAN tunnel between the VTEPs. Routing protocol implementation in any network requires the IP address to be applied first to the network infrastructure. The IP address is applied based on an IP addressing scheme for the physical and logical interfaces on the leaf and spine nodes. The logical and physical interfaces enable the routing protocols to run between the leaf and spine nodes. Once the routing protocols are up, they provide connectivity between VTEP logical interfaces, such as loopbacks used as source and destination interfaces for VXLAN tunnel interfaces.

The underlay unicast routing design requires the data center architect to decide on the following:

The network's physical topology

Some questions to consider when choosing your physical topology for the data center are:

- How many super spine, spines, and leaf nodes are required?
- How many uplinks from each leaf to the spine?
- How many links from the spine to the super spine?
- Do the endpoints need to be multi-homed to VTEPs using vPC?

The IP addressing scheme.

Some questions to consider when choosing your IP addressing scheme for the data center are:

- Will the fabric interfaces have numbered or unnumbered IP addresses?
- How many leaf nodes are in total?
- How many switches are in a vPC domain in total?
- How many spine nodes are in total?
- How many VLANs are acting as default gateways in total?
- How many VRFs are in total, and do you need to route between the VRFs?
- What network services, such as DHCP or VXLAN OAM, will the VTEPs require?
- What roles will the spine nodes support, such as route reflector, rendezvous point, border spine, or border gateway spine?

The routing protocol to run in the underlay.

Some questions to consider when choosing your underlay routing protocol for the data center are:

- Is there a preference to run a routing protocol already in the existing production environment?
- Will introducing the routing protocol require staff training?
- Does the routing protocol require multi-vendor or multi-platform support?
- Does the underlay routing protocol require dual-stack support for both IPv4 and IPv6?
- Is the requirement to support fast convergence and maximum scale?
- Is the routing protocol easy to configure, operate, and troubleshoot?

Design decisions always involve trade-offs. The data center architects should prioritize each requirement documented during the project's requirements-gathering phase and develop a proposal that addresses as many requirements as possible. The high-priority requirements are mandatory for the optimal functioning of the applications hosted on the network for the business and the operation of the network.

The CLOS fabric is a topology where every leaf node is fully meshed to all the spine nodes in a single site. Each link between the leaf and spine is called a fabric interface. The links on the VTEPs facing the attached endpoints are called host interfaces. The links attached to external devices, such as edge routers and firewalls, are called external or service interfaces. As shown below, the host, external, or service interfaces can be single-homed to one VTEP or multi-homed to two VTEPs in a vPC configuration.

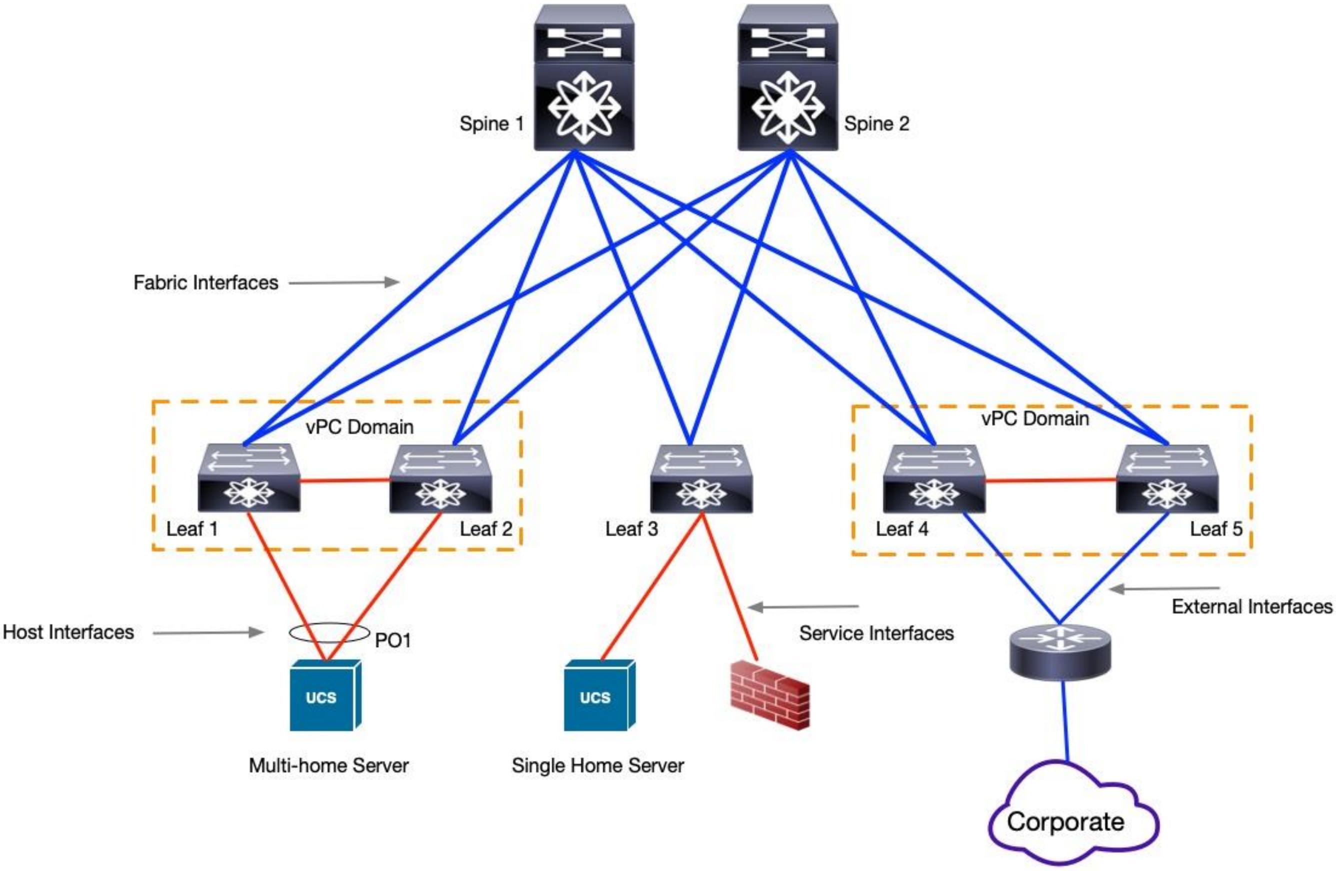


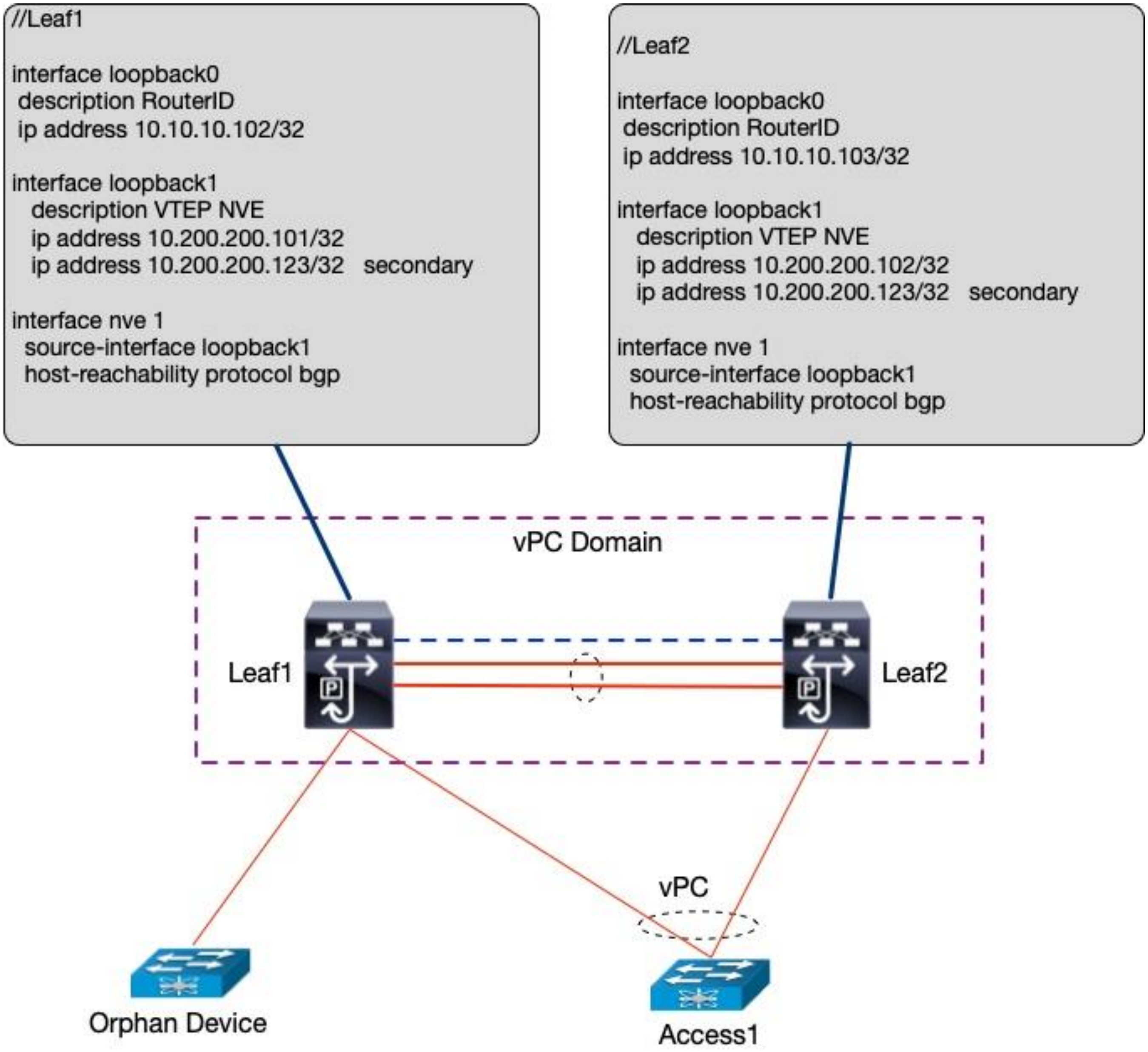
Figure 7. Fabric Link Types

The data center is where applications supporting the enterprise business are hosted. Some of the servers are running mission-critical applications. These applications need to be always available to users. The application availability is provided with network redundancy. Cisco Virtual Port-Channel (vPC) is a layer two port-channel that allows endpoints to attach to two VTEP switches (multi-chassis) with both links in an active forwarding state. Cisco vPC was first introduced in Nexus 5000/7000 platforms and continues to be popular on Nexus 9000 platforms. The Nexus 9000 platform has two variations of vPC, the standard vPC with a physical peer link and vPC Fabric Peering (vPC without a physical peer link).

Virtual Port Channel (vPC)

A VXLAN BGP fabric is a routed fabric. The underlay fabric interfaces between the leaf and spine provide layer 3 Equal Cost Multipath (ECMP) from the source to a destination IP address. The vPC member switches in the data plane behave like a single switch to the endpoints attached to the vPC interfaces. To the remote leaf switches sending traffic to an endpoint connected to a vPC member leaf switch, the vPC member leaf switches are also seen as a single next hop using a common IP address called a vPC Virtual IP (VIP). The virtual IP is an anycast IP address configured on both vPC member leaf switches.

In a VXLAN fabric, each leaf is also called a VXLAN Tunnel Endpoint (VTEP). Each VTEP's tunnel interface is called a Network Virtualization Edge (NVE) interface. The NVE interface is bound to a dedicated loopback interface separate from the loopback used for the routing protocol router ID. The NVE interface IP is the next hop to reach all the tenant endpoints and networks attached to a VTEP. The NVE interface loopback is assigned two IP addresses, the primary and secondary IP address. The primary IP (PIP) address is unique to each vPC member switch. The secondary IP is the vPC VIP, an anycast IP address common to vPC member switches. The vPC VIP address encapsulates all VXLAN traffic, including unicast and multicast. Cisco documentation commonly uses Loopback1 as the loopback interface bounded to the NVE interface, but it is possible to use other loopback number interfaces.



```
//Leaf1

interface loopback0
 description RouterID
 ip address 10.10.10.102/32

interface loopback1
   description VTEP NVE
   ip address 10.200.200.101/32
   ip address 10.200.200.123/32   secondary

interface nve 1
  source-interface loopback1
  host-reachability protocol bgp
```

```
//Leaf2

interface loopback0
 description RouterID
 ip address 10.10.10.103/32

interface loopback1
   description VTEP NVE
   ip address 10.200.200.102/32
   ip address 10.200.200.123/32   secondary

interface nve 1
  source-interface loopback1
  host-reachability protocol bgp
```

Figure 8. vPC IP Addresses

The following table summarizes when the vPC VIP is used as the next hop for host and prefix routes.

EVPN Route Type	Attachment	Next-hop
Type 2 (Host Routes)	vPC	advertised by VIP
	Orphan Port	advertised by VIP
Type 5 (IP Prefix Routes)	vPC	advertised by VIP
	Orphan	advertised by VIP

Table 1. vPC VIP Next Hop Route Advertisements

The loopback 0 interface is commonly used as the router ID IP for underlay routing protocols such as OSPF and ISIS. OSPF and ISIS automatically select a loopback interface as a router ID IP address. However, it is best practice to manually set the router ID for underlay IGP to the loopback interface assigned for layer three routing protocol router-id purposes. The 32-bit router-id is usually set manually under the BGP process.

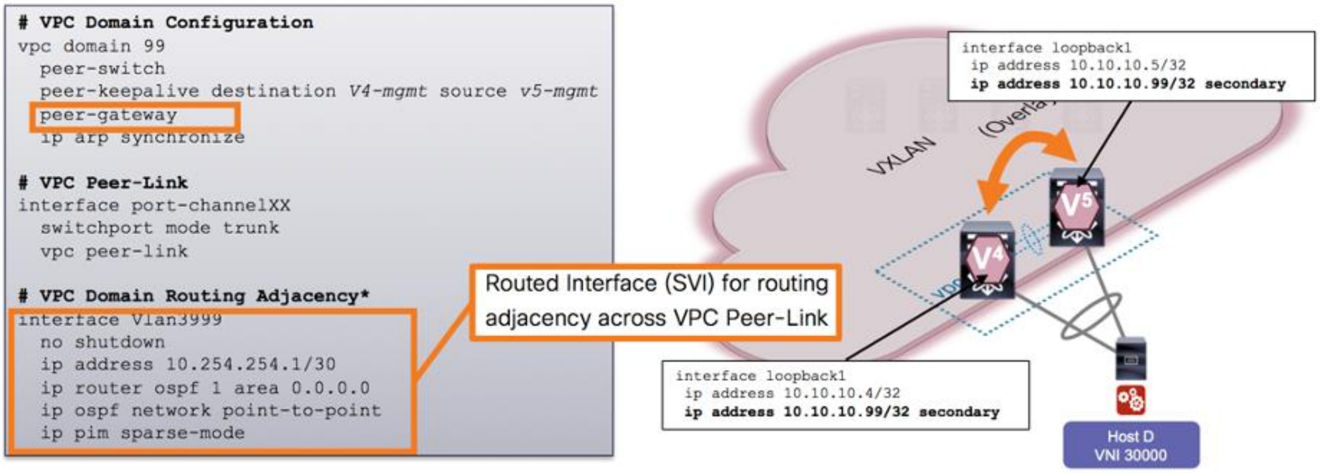
Separating loopback interfaces, one for the underlay routing protocol and the other for the VTEP NVE interface, provides stability in the network. The stability is provided during failure scenarios. The failure in the vPC domain includes configuration mistakes. The vPC consistency checker validates the configuration mismatch to ensure both vPC member switches have the same vPC-related configuration. The devices connected to the vPC member switch receive the same LACP system ID when negotiating the vPC port channel. The vPC member switches are represented as a single switch in the data plane; therefore, their forwarding-related configurations must match. The vPC consistency checker shut down the NVE interface by shutting down the NVE tunnel source interface. If the underlay routing protocol and NVE interface share the same loopback, shutting down the shared loopback will bring the NVE interface and underlay routing protocol down. A data plane configuration issue in the vPC system impacts the control plane in the underlay. It is recommended that the loopback be separated according to function to avoid this shared fate.

A key design decision for vPC is the IP address assignment for the NVE and router-id loopback interfaces. A total of 3 IP addresses must be allocated on each, which are the following:

- Unique IP address for underlay routing protocol router-id.
- Unique primary IP address for NVE interface.
- Anycast secondary IP address for NVE interface vPC VIP.

The servers attached to vPC switches are multihomed to two-leaf switches. The leaf switch connects to the spines, and the fabric interface between the leaf and spines can fail. If the traffic from the server is hashed to a leaf switch with the failed uplink fabric interface to one of the spines, the traffic should have a path using the vPC peer-link to the peer vPC switch to use its uplink interface to forward traffic to one of the spines. The vPC peer-link is a layer 2 port-channel interface. A switch virtual interface (SVI) is created on vPC member switches inside a VLAN that acts as a backup SVI for traffic redirection. The backup SVI extends the underlay network between the vPC member switches. The underlay network unicast routing protocols such as ISIS, OSPF, or eBGP must establish peering on the backup SVI. For BUM traffic and multicast routing PIM Sparse mode is enabled in the underlay. The underlay network in a VXLAN BGP EVPN fabric is in the default VRF table.

The following image shows a configuration example of a backup SVI in VLAN 3999. The underlay routing protocols in the example are OSPF and PIM. The vPC peer-link must allow VLAN 3999 in the port-channel trunk interface.



```
# VPC Domain Configuration
vpc domain 99
  peer-switch
  peer-keepalive destination V4-mgmt source v5-mgmt
  peer-gateway
  ip arp synchronize

# VPC Peer-Link
interface port-channelXX
  switchport mode trunk
  vpc peer-link

# VPC Domain Routing Adjacency*
interface Vlan3999
  no shutdown
  ip address 10.254.254.1/30
  ip router ospf 1 area 0.0.0.0
  ip ospf network point-to-point
  ip pim sparse-mode
```

Routed Interface (SVI) for routing adjacency across VPC Peer-Link

```
interface loopback1
 ip address 10.10.10.5/32
 ip address 10.10.10.99/32 secondary
```

```
interface loopback1
 ip address 10.10.10.4/32
 ip address 10.10.10.99/32 secondary
```

Figure 9. vPC Back Up SVI

Note: The vPC peer-gateway feature must be enabled on both peers to facilitate NVE RMAC/VMAC programming on both peers. For peer-gateway functionality, at least one backup routing SVI must be enabled across peer-link and configured with PIM. This provides a backup routing path when VTEP loses complete connectivity to the spine. Remote peer reachability is rerouted over peer-link in his case.

The backup SVI VLAN are categorized as infrastructure VLANs as they are not mapped to VXLAN L2 VNIs. The infrastructure VLAN is defined with the following commands:

```
nexus(config)# system nve infra-vlans <1-3967>
```

A separate layer 3 routed interface between vPC member switches as a backup path is also supported. The underlay network is extended on a separate layer 3 routed interface between the vPC member switches.

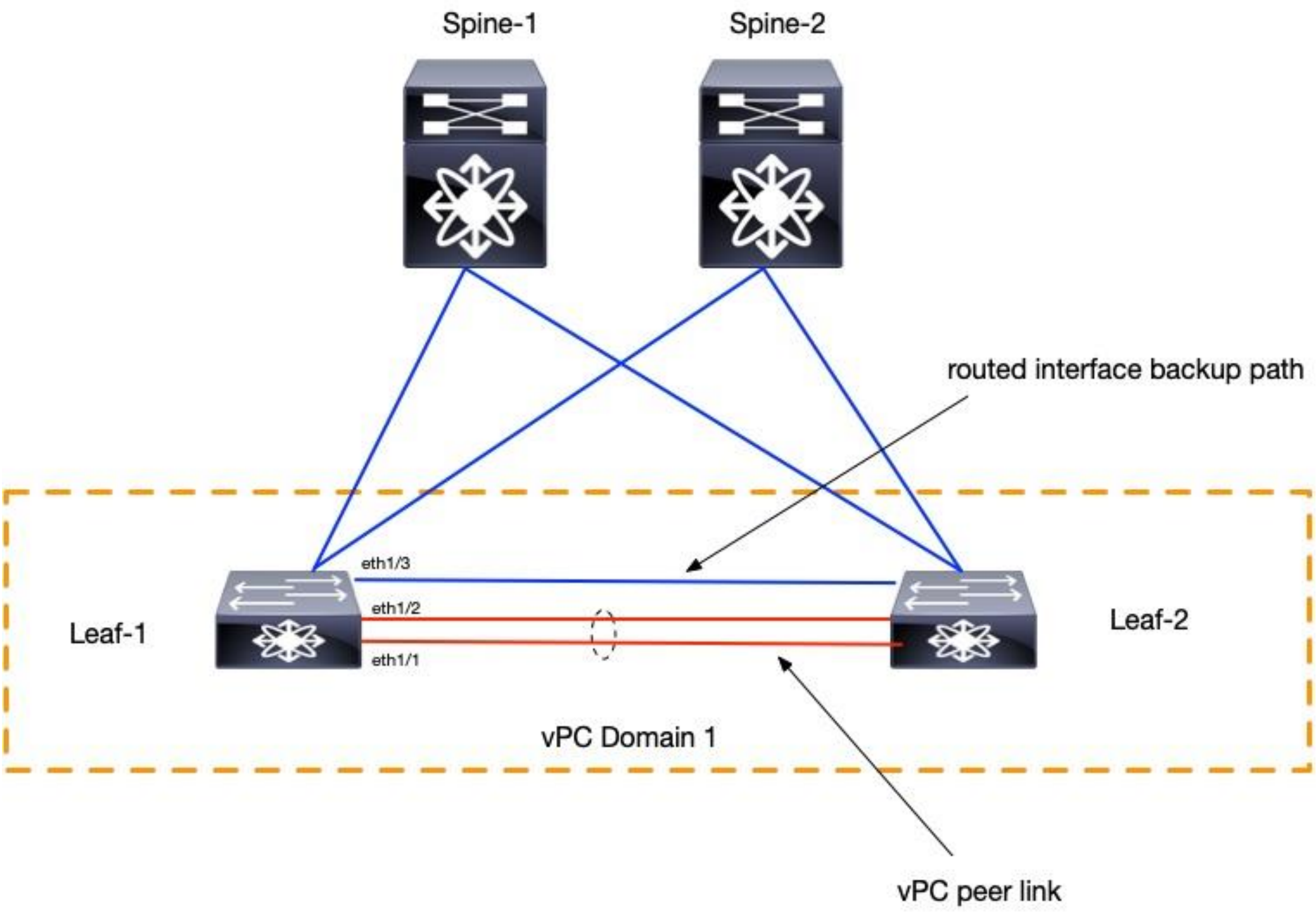


Figure 10. Layer 3 Routed Interface Backup Path

The first advantage of using a separate layer 3 interface for the backup path is redundancy and fault isolation. A failure of the vPC peer link will not impact the backup path traffic. Another advantage of using a separate layer 3 interface for the backup path is since the backup path is not shared with the vPC peer link, there's no risk of congestion due to backup traffic competing with regular VXLAN or vPC synchronization traffic. The disadvantage of using a separate layer 3 interface for the backup path is that it consumes one extra physical interface.

The vPC peer link backup SVI and separate layer 3 routed interface for the backup path are both supported architectures. Some of the key considerations for both approaches include the following:

Traffic Engineering

If your backup path needs to handle traffic only during a failure, you might configure routing policies or metrics to ensure that it only becomes active when the primary path fails. This can be done by adjusting OSPF/ISIS/EBGP.

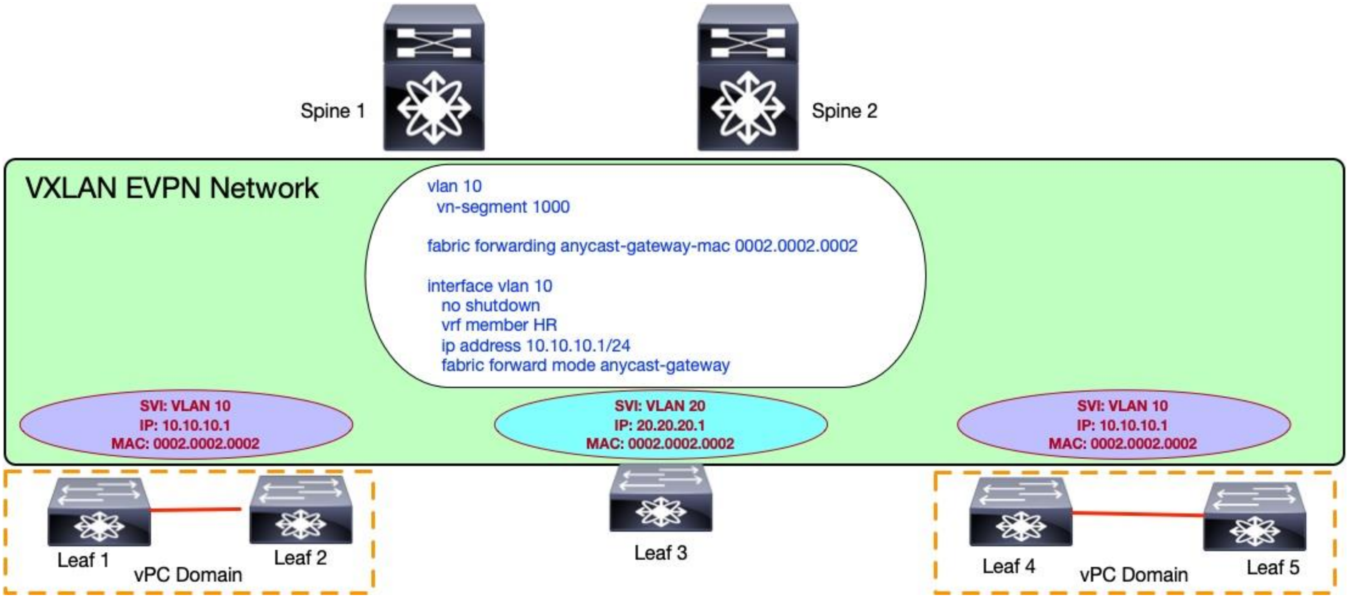
Redundancy and Failure Scenarios

Consider how failure is detected and how traffic should be rerouted in both designs. Bidirectional Forwarding Detection (BFD) or other failure detection mechanisms can help ensure quick failover.

Operational Complexity

A backup SVI over the vPC Peer Link is simpler but might offer different fault tolerance and flexibility than a separate Layer 3 routed backup path. The latter might be more complex but is often preferred in large-scale, high-availability environments where traffic resilience and fault isolation are critical.

In a VXLAN EVPN fabric, every VTEP is assigned the same MAC address for every L2 VNI SVI interface, acting as a default gateway. A common gateway MAC assigned across all VTEPs for all networks is a Distributed Anycast Gateway (DAG).



```
vlan 10
  vn-segment 1000

fabric forwarding anycast-gateway-mac 0002.0002.0002

interface vlan 10
  no shutdown
  vrf member HR
  ip address 10.10.10.1/24
  fabric forward mode anycast-gateway
```

Figure 11. Distributed Anycast Gateway

The DAG is a pervasive gateway approach. A network shares the same gateway IP and the globally assigned DAG MAC address. In the above figure, the VLAN 10 gateway IP and MAC are the same on leaf 1, leaf 2, leaf 4 and leaf 5. The DAG aligns with the peer-gateway capability in vPC, providing active/active forwarding without a need to implement First Hop Redundancy Protocols (FHRP) like HSRP or VRRP. The peer-gateway capability allows both vPC member switches to act as gateways for the connected endpoints destined to the router MAC of the vPC switch without requiring the packet to be sent out the peer-link to the primary vPC switch.

The VLANs that leverage DAG are usually extended across VTEPS in the data center, allowing virtual machine mobility. As the endpoint attachments to the network are moved from one VTEP to the other, the ARP entry for the gateway MAC does not need to be resolved again on the endpoint's ARP cache, as the gateway MAC and IP remain consistent across VTEPs provisioned with the same VLAN.

In the infrastructure IP planning, the VLANs that require layer 2 extensions in the data center need to be allocated a single shared gateway IP per VLAN and a common anycast gateway MAC across all VLANs. Note down every VTEP with the VLAN attached and configure the required gateway configurations. VLANs do not need to be connected to every VTEP in the data center. The VLANs are attached based on security, scale, traffic pattern, and application requirements. Place the VLANs on switches based on application requirements.

The following guidelines must be followed when implementing vPC VTEPs in VXLAN fabrics:

- Bind NVE to a loopback address separate from other loopback addresses required by Layer 3 protocols. A best practice is to use a dedicated loopback address for VXLAN and a separate loopback for layer 3 underlay protocols.
- The loopback address used by NVE needs to be configured to have a primary IP address and a secondary IP address. The secondary IP address is used for all VXLAN encapsulated traffic, including multicast and unicast.
- The primary IP address is unique and is used by Layer 3 protocols. The secondary IP address on the loopback is necessary because the interface NVE uses it for the VTEP IP address. The secondary IP address must be the same on both vPC peers.
- vPC peers must have the following identical configurations:
    - Consistent VLAN to vn-segment mapping.
    - Consistent NVE1 binding to the same loopback interface
    - Using the same secondary IP address.
    - Using different primary IP addresses.
    - Consistent VNI to group mapping.

If the configurations above are inconsistent between the vPC member switches, the NVE loopback interface will be shut down administratively on the vPC secondary VTEP.

vPC Fabric Peering

The vPC multihoming technology requires a physical peer link between the vPC member switches. The VXLAN BGP EVPN CLOS fabric architecture provides each leaf device with a fully meshed connection between the spine and leaf layers. If a leaf in a vPC has 2 spines then it has 2 physical paths to send any traffic to its vPC peer leaf. The fabric architecture provides high bandwidth redundant routed path between every leaf through the spine for data and control plane traffic. The physical peer link requirement between vPC can be removed now. In vPC Fabric peering, the vPC peer link is converted to a logical connection between the vPC member switches called the vPC virtual peer-link. The vPC fabric peering creates two tunnels: the data plane tunnel, which carries the VXLAN traffic, and the control plane tunnel, which syncs data plane states such as ARP, MAC, IGMP, etc, between the vPC leaf switches using Cisco Fabric Services over IP (CFSoIP). The traditional vPC with the physical peer link uses CFS over Ethernet (CFSoE).

Like any tunnel, the vPC virtual peer link requires a source and terminating interface to form the tunnel. A loopback interface is also required. The loopback interface used for the vPC virtual peer link cannot be the same as the peer keep alive or the loopback used for the VTEP IP address. The loopback used for the underlay unicast routing protocol, such as router-id OSPF or BGP can be used as a loopback for the vPC peer link. The best practice recommendation is to have a dedicated loopback for the vPC virtual peer link. The separate loopback utilized for separate functions provides fault isolation. The keep-alive can be out-of-band or inband. The out-of-band peer keep-alive can use the mgmt0 or dedicated interface. The inband can use a dedicated loopback over the fabric.

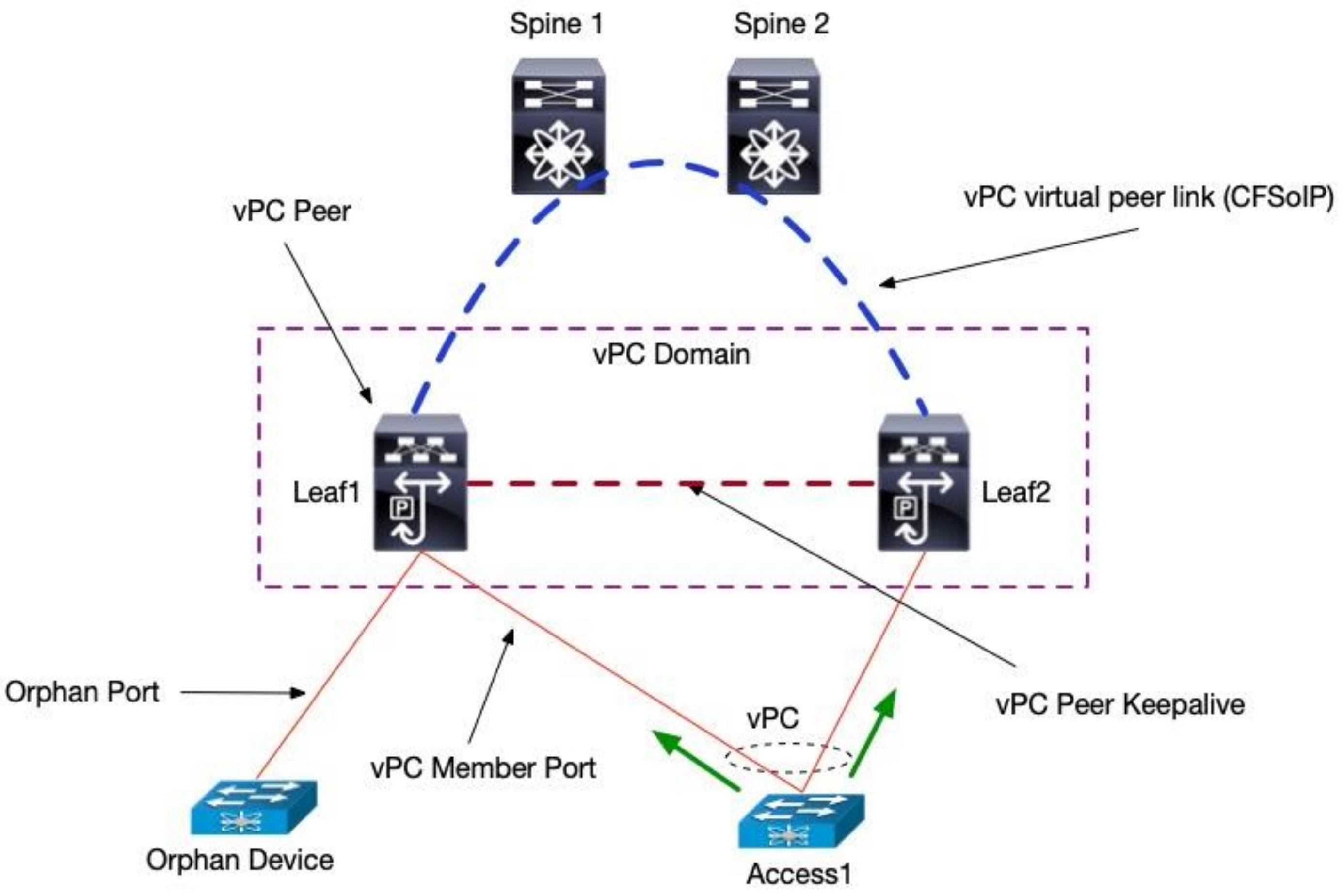


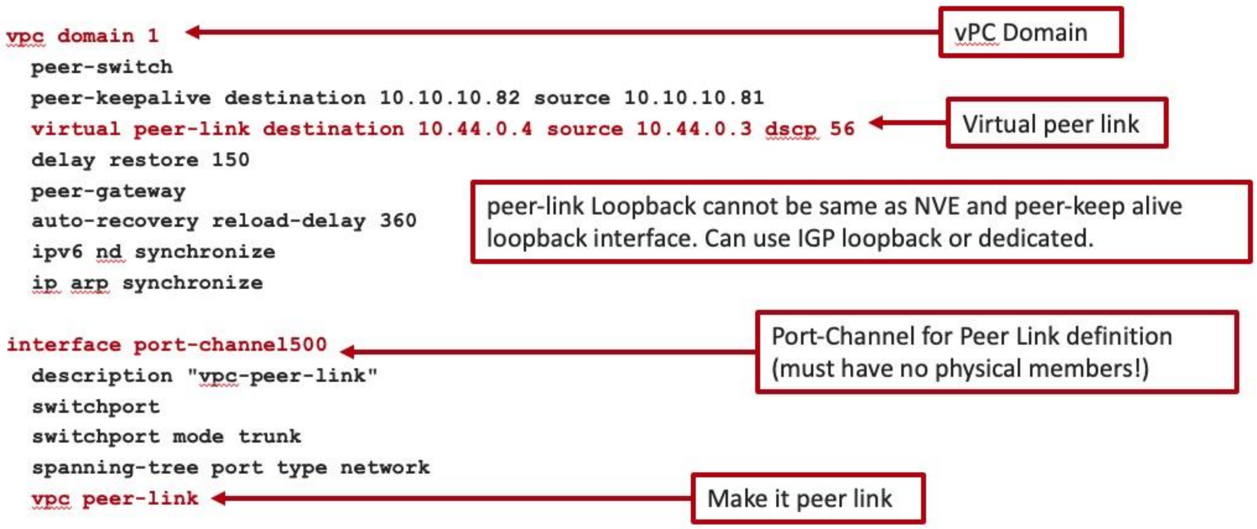
Figure 12. vPC Fabric Peering Architecture

The control plane requirements and behavior change slightly compared to traditional vPC. Some key concepts to remember are the following:

- The advertise PIP and RMAC configurations are mandatory in vPC fabric peering if EVPN Type 5 is used with this feature. vPC and Orphan Type 5 routes will be advertised with (PIP,RMAC).
- An orphan Type-2 host is advertised using PIP and RMAC. A vPC Type-2 host is advertised using vPC VIP and VMAC.
- Some platforms may require TCAM carving of the region **ing-flow-redirect** before configuring vPC Fabric Peering. TCAM carving requires saving the configuration and reloading the switch before using the feature. Please confirm the platform steps from the Nexus 9000 NXOS VXLAN configuration guide.

The steps to configure vPC fabric peering on the vPC leafs are the following:

Step 1. Define the vPC Domain.



```
vpc domain 1
  peer-switch
  peer-keepalive destination 10.10.10.82 source 10.10.10.81
  virtual peer-link destination 10.44.0.4 source 10.44.0.3 dscp 56
  delay restore 150
  peer-gateway
  auto-recovery reload-delay 360
  ipv6 nd synchronize
  ip arp synchronize

interface port-channel500
  description "vpc-peer-link"
  switchport
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link
```

vPC Domain

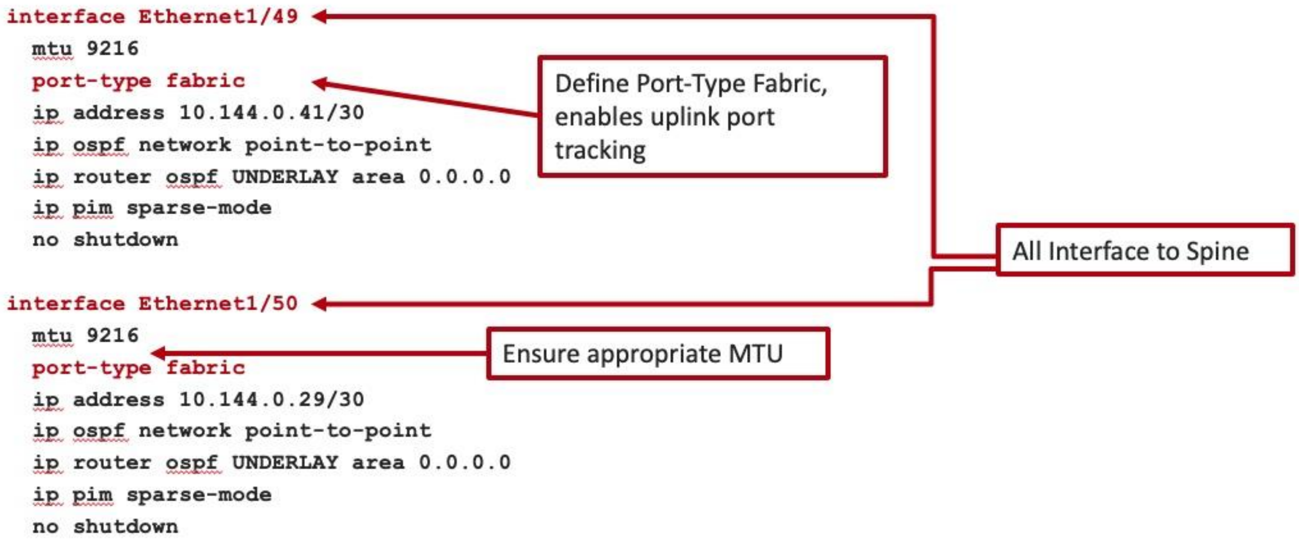
Virtual peer link

peer-link Loopback cannot be same as NVE and peer-keep alive loopback interface. Can use IGP loopback or dedicated.

Port-Channel for Peer Link definition (must have no physical members!)

Make it peer link

Step 2. Define the uplinks to the Spine.



```
interface Ethernet1/49
  mtu 9216
  port-type fabric
  ip address 10.144.0.41/30
  ip ospf network point-to-point
  ip router ospf UNDERLAY area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet1/50
  mtu 9216
  port-type fabric
  ip address 10.144.0.29/30
  ip ospf network point-to-point
  ip router ospf UNDERLAY area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```

Define Port-Type Fabric, enables uplink port tracking

All Interface to Spine

Ensure appropriate MTU

In the above configuration snippet, when defining the virtual peer link, dscp 56 is also applied for virtual peer link traffic. The virtual peer link synchronizes the data plane state; it is a control plane tunnel. The traffic in the virtual peer link must be treated as a high priority. The spine nodes must match this peer-link traffic and allocate them to the strict priority queue during congestion. The below shows a configuration snippet from the spine node.

```
//Classify traffic by matching the DSCP value (DSCP 56 is the default value).
class-map type qos match-all CFS
  match dscp 56

//Set traffic to the qos-group that corresponds with the strict priority queue for the
appropriate spine switch.
policy-map type qos CFS
  class CFS
    Set qos-group 7
```

```
//Assign a classification service policy to all interfaces toward the VTEP (the leaf layer
of the network).
interface Ethernet 1/1
  service-policy type qos input CFS
```

IP Address Design

The IP address design for any network is foundational to a good network design. Poor IP address design can cause challenges with the control plane scale due to a lack of summarization, risk of IP address overlap and exhaustion due to lack of scheme, and poor allocation practice. A good IP address scheme helps understand the network's logical topology, which results in quick troubleshooting. Designing an IP address scheme always takes time in any greenfield network but getting it right from the start is vital.

In a VXLAN BGP EVPN fabric, the IP address is required in the underlay and overlay network protocols. In the network, both unicast and multicast routing protocols require an IP address. The roles of the nodes and their place in the network also determine what and how many IP addresses a node requires. The spine device does not have endpoints attached and usually does not require a gateway IP address assigned. A border leaf device peering to an external router using the VRF-lite extension will require sub-interfaces configured per VRF and the IP address assigned to those interfaces with a routing protocol or static routing configured to the edge router. VTEPs used only for endpoint attachment do not require the extra IP address for WAN edge router interconnection. Another important factor influencing how many IP addresses a network requires is whether network infrastructure services have been enabled. For example, if DHCP Relay or VXLAN OAM is configured, extra interfaces are needed to support the network communication and enable those services. Before you start defining your IP address scheme, remember the design principle to understand all the roles, places in the network, and network protocols and services required on the network device. Calculate the maximum IP address required per device role, leaving room for fabric endpoint and network node growth.

The VXLAN BGP EVPN fabric leaf node with endpoints attached for unicast routing requires the following IP addresses:

- **Fabric Interface IP**: The fabric interface is the layer 3 routed interface between the leaf and spine. The IP address assignment can be allocated using point-to-point (P2P) IP. The p2p IP has a /30 or /31 mask. The p2p IP is assigned to the physical interface. The other option for fabric interface IP address assignment is IP unnumbered. The IP unnumbered interface borrows an IP address from another interface, such as a loopback. In VXLAN BGP EVPN fabrics, the fabric interfaces borrow the IP address from the loopback allocated for the underlay of the IGP unicast routing protocol, commonly configured as loopback 0. The best practice is to use unnumbered interfaces because it reduces the number of IP addresses required in the transit interfaces between leaf and spine, simplifying automation and IP address allocation management.
- **Loopback Interface IP**: A leaf node will at least require a loopback interface for layer 3 underlay routing and NVE interface. The vPC and vPC fabric Peering also have different IP address requirements. The vPC fabric peering may require a unique loopback for virtual peer link or peer keepalive.
- **L2 VNI Gateway IP**: Every L2 VNI network IP address is required for the SVI for the VLAN mapped to L2 VNI. The SVI IP is the gateway IP address for the host attached to the network.
- **Network Infrastructure Services**: DHCP Relay and VXLAN OAM are commonly used features on VTEPs. Each VTEP requires a unique loopback interface in a VRF to uniquely identify the leaf for

DCHP and OAM packet forwarding and receiving. It is impossible to use SVI gateway IPs as they are shared across all VTEPs due to the distributed anycast gateway configuration to support endpoint mobility.

Note: Border Leaf or Border Gateway devices will require other IP addresses to support their role and function. For those use cases, please reference the Nexus 9000 NXOS VXLAN configuration guide on cisco.com, particularly the chapters on Multisite VXLAN BGP EVPN and External VRF Connectivity.

The VXLAN BGP EVPN fabric spine node for unicast routing requires the following IP addresses:

- **Fabric Interface IP**: The fabric interface is the layer 3 routed interface between the leaf and spine. The IP address assignment can be allocated using point-to-point (P2P) IP. The p2p IP has a /30 or /31 mask. The p2p IP is assigned to the physical interface. The other option for fabric interface IP address assignment is IP unnumbered. The IP unnumbered interface borrows an IP address from another interface, such as a loopback. In VXLAN BGP EVPN fabrics, the fabric interfaces borrow the IP address from the loopback allocated for the underlay of the IGP unicast routing protocol, commonly configured as loopback 0. The best practice is to use unnumbered interfaces because it reduces the number of IP addresses required in the transit interfaces between leaf and spine, simplifying automation and IP address allocation management.
- **Loopback Interface IP**: A spine node will at least require a loopback interface for layer 3 underlay routing.

Note: The above assumes a Spine node with no other functions overloaded, such as a Border Gateway or Border Spine. Cisco NXOS supports a variety of architectures. The more roles added to the Spine, the more IP addresses will be required to support those functions and roles. Please reference the Nexus 9000 NXOS VXLAN Configuration Guide on cisco.com, particularly the chapters on Multisite VXLAN BGP EVPN and External VRF Connectivity.

IGP vs eBGP Underlay

An important design decision is which underlay protocol is best for a VXLAN BGP EVPN fabric. VXLAN is the data plane protocol that works with the MP-BGP EVPN control plane protocol. History teaches us a lot to help us answer questions for today. The MP-BGP control plane has been used with Multiprotocol Label Switching (MPLS) for two decades. The latest evolution of MPLS with Segment Routing (SR) continues to use MP-BGP control plane protocol. In the MPLS networks, MP-BGP's role is to signal the exchange of VPN or overlay routing information. The largest MPLS networks in the world all use IGPs such as OSPF or ISIS in the underlay for provider edge node connectivity through the core network. IGP had a simple function to provide fast ingress PE to egress PE connectivity. The provider core routers had no overlay routes in their routing table unless the core device had a route reflector or route server role assigned to it. If IGPs have succeeded in the largest MPLS overlay networks for two decades, does it not prove that IGPs should be the choice of underlay even in a VXLAN BGP EVPN network?

A famous **informational** RFC 7938 titled "*Use of BGP for Routing in Large-Scale Data Centers*" proposed using BGP as a routing protocol to build massively scalable data centers. The RFC proposed using eBGP in the CLOS fabric. EBGP routing concepts and features were explained to optimize BGP to behave like an IGP. Optimization techniques to support ECMP using maximum-paths and converge faster by lowering the Minimum Route Advertisement Interval (MRAI) timers. The functionality that an IGP provides by default use BGP knobs to make BGP behave like an IGP. Is it possible to create a very large-scale data center fabric using eBGP routing protocol? Yes, it is. Is it worth the effort to tweak BGP to make it behave like an IGP? For some, the answer is yes because those networks are probably non-overlay networks to simplify the network system as much as possible. Some industries simplify network systems by removing as many features as possible. The ideal system for some sectors is a network system with a lightweight operating system, energy efficiency, minimal resource usage, minimal price, and easily replaceable. One major factor for running network nodes with minimal protocols and features is to avoid defects and reduce security attack vectors. These are all valid reasons, but they are specific to the industry and network operational model and goals. They do not apply to all sectors.

A vital concluding point on this informational RFC 7938 is that it does **not** discuss or propose using eBGP as an underlay routing protocol in a VXLAN BGP EVPN fabric or any overlay running in the data center. Therefore, referencing this informational RFC 7938 is not a strong argument for using eBGP as an underlay in VXLAN BGP EVPN fabrics.

A network is a complex system. A fundamental design principle applied when building any complex system is to break up that complex system into smaller problems or functions. The individual sub-systems are simple and provide a simple function to other systems. Modifying the code into multiple libraries or smaller functions is standard practice in software engineering. The goal is not to overload one system with too many functions. Overloading functions on one system adds technical debt. Overloading a single protocol like BGP to do underlay and overlay routing is overloading the BGP protocol with multiple functions. The separation of IGP for underlay and BGP for overlay allocates specialized functions to each of these protocols to deliver what each protocol is designed to do best.

Cisco Nexus 9000 NX-OS fully supports eBGP VXLAN BGP EVPN fabrics with IGP and eBGP underlays.

Cisco's best practice is to use IGP, such as OSPF or ISIS, in the underlay with an iBGP EVPN overlay. Some industries prefer to use eBGP to support such requirements; Cisco Nexus NX-OS fully supports eBGP VXLAN BGP EVPN fabrics.

The following table compares IGP vs eBGP as an underlay protocol:

Feature	IGP	eBGP
Extensibility	ISIS and OSPFv3 are multi-protocol. OSPFv2 only routes IPv4 prefix.	BGP supports multiple address families including L2VPN, L3VPN, Flowspec etc
ECMP	IGPs support it natively	BGP requires maximum path knob and need to meet certain BGP best path decision criteria.
Unnumbered Interfaces	Supported	Need IPv6 underlay with LLA peering. IPv4 underlay requires point to point routed interfaces with IP address.
Training/Knowledge	IGPs is well known by data center engineers for decades. Commonly deployed between distribution and core layers or even routed access networks down to access layer.	MP-BGP is a relatively new technology in the data center. Data centers have predominately been 3-tier Access/Distribution/Core networks. BGP was between Distribution and FW or Core and WAN router. Fabric path, OTV and Trill were also Layer 2 overlays with ISIS underlays in the data center.
Troubleshooting	Multiple protocols to monitor and troubleshoot IGP + MP-BGP.	A single protocol to configure, monitor and troubleshoot.
Fault Tolerance	Underlay and overlay are decoupled. Protocol map to a function.	eBGP failure can bring down overlay and underlay.

Feature	IGP	eBGP
Open Standards	ISIS, OSPFv2 and OSPFv3 all are open standards.	BGP is open standard.

Table 2. VXLAN IGP vs eBGP Underlay Protocol

OSPF Routing

Open Shortest Path First (OSPF) is a link-state routing protocol classified as an Interior Routing Protocol (IGP). RFC 2328 describes OSPFv2, and RFC 2740 describes OSPFv3. OSPF calculates the best path using the shortest path first (SPF) algorithm and uses cost as its metric. The primary functions of any underlay routing protocol are:

1. The advertisement of the control plane peering interface IP address. The MP-BGP EVPN control plane is used in VXLAN BGP EVPN for unicast routing and bridging. BGP EVPN peering uses the loopback IP address to peer. OSPF provides reachability for BGP EVPN address family peering. A separate loopback is used for BGP router-id.
2. The advertisement for the VTEP NVE interface peering address. The VTEP NVE interface IP (VIP or PIP) is the next hop address for host and prefix routes. A local VTEP should be able to reach the remote VTEP to reach the attached host or network. A separate loopback is used for the VTEP NVE interface.

The underlay routing protocol should provide fast convergence even when a node or link failure occurs to support the above two functions effectively.

OSPF for multiple paths with equal metrics will install the paths in the routing table and forward packets using equal-cost multipath (ECMP). The VXLAN fabric is a CLOS architecture; to scale out endpoints, the leaf nodes are added, and the spine nodes are added to scale out the leaf. The more spines you have, the more paths from one ingress VTEP to egress VTEP. OSPF will automatically learn of those alternative paths with equal metrics and install them into the routing table.

OSPF is a well-known open standard protocol. It is simple to configure, and skilled engineers in OSPF routing protocol are easy to find. A valuable feature of OSPF is the ability to isolate a node from the network path using **max-metric router-lsa** command. The router advertises itself as a stub by advertising router LSAs with maximum metric value. This feature is handy when installing a new spine, a VXLAN BGP EVPN fabric. A spine can be introduced in the fabric to increase the spine-to-leaf bandwidth and allow more leaf switches to be added to the fabric without overloading the existing spines. Introducing an additional spine provides path redundancy. If one spine goes down or becomes unreachable, traffic can still be routed through the remaining spines, ensuring high availability and fault tolerance. With additional spines, the fabric can distribute traffic more evenly. VXLAN uses ECMP (Equal-Cost Multi-Path) routing, meaning traffic can be spread across multiple paths between the leaf and spine layers. This helps in avoiding bottlenecks and ensures more efficient use of available links.

Once the spine is configured and tested for production, it is introduced into the network path by removing the **max-metric router-lsa** command The CLI for the command with options is shown below:

```
max-metric router-lsa [external-lsa [<max-metric-value>]] [include-stub] [on-startup
{<seconds> | wait-for bgp <tag>}] [summary-lsa [<max-metric-value>]]
```

Stub route advertisements can be configured with the following optional parameters:

on startup: Sends stub route advertisements for the specified announcement time.

wait for BGP: Sends stub router advertisements until BGP converges.

The configuration snippet below shows how to enable the stub router advertisements on startup for the default 600 seconds:

```
configure terminal
router ospf underlay
    max-metric router-lsa on-startup
```

The following OSPF protocol deployment considerations for an underlay IP network in a VXLAN fabric are given below:

Maximum Transmission Unit (MTU)

VXLAN packets cannot be fragmented. It is recommended to use jumbo frames. Most server network interface cards (NICs) support up to 9000 bytes. The MTU of 9216 bytes should be used in the underlay. An MTU of 9216 bytes on each interface on the path between the VTEPs accommodates maximum server MTU + 50 bytes VXLAN overhead. The host facing interfaces on the VTEPs should also be configured with the same MTU as the fabric interfaces.

Underlay Address Family

The underlay comprises the links between the spine and leaf switches, loopback interfaces for router ID, and the VTEP NVE interface. OSPFv3 is enabled for IPv6 underlay, and OSPFv2 is enabled for IPv4 only underlay. The OSPF IGP must be enabled on the interfaces between the leaf and spine, plus the loopback interfaces.

OSPF Area

The general recommendation is to use a single area, 0, in the underlay. Multiple areas in the underlay are possible to configure but unnecessary because the underlay's role is to provide VTEP and BGP peering reachability. Even in a single site with 1000 leaf VXLAN BGP EVPN fabric with unnumbered fabric interfaces and two loopbacks (one for router-id and VTEP NVE interface), the total number of underlay networks learned is just 2000. The modern data center switches can easily support 2000 LSAs in an OSPF topology database with insignificant impact on TCAM resources and memory. Depending on the routing hardware template and routing protocol implemented, the Nexus 9000 NXOS platform can support 100s of thousands to even up to 1 million routes. Older platforms with limited CPU/memory resources in campus and wide area networks (WAN) infrastructure OSPF areas have their benefits. In a very large-scale VXLAN BGP EVPN fabric, such as a multi-pod fabric in different buildings or geographic locations with hundreds of VTEPs in each location, then the Cisco recommendation would be to move to VXLAN BGP EVPN Multisite architecture for large-scale VXLAN BGP EVPN fabrics. Each VXLAN site becomes an isolated fault domain, providing a more stable and scalable network. Please reference the VXLAN EVPN Multi-Site Design and Deployment White Paper.

OSPF Network Type

By default, any ethernet interface is understood as an OSPF broadcast network, which requires a designated router or backup designated router election. DR/BDR election creates unnecessary election and database overhead (from the LSA type 2). The fabric interface between the leaf and spine are point-to-point links, not a multiaccess broadcast network with multiple OSPF routers attached. In a point-to-point link, only two network nodes are connected. Therefore, a DR/BDR election is not beneficial as the database is only synchronized between two devices. The recommendation is to define the fabric interfaces between the leaf and spine as OSPF interface type point-to-point.

OSPF Fabric Interface IP Address

OSPF supports point-to-point IP addressing with a/30 and /31 subnet mask. The IP unnumbered does not require an IP address on the fabric interface between the leaf and the spine. The IP unnumbered approach consumes fewer IP addresses.

This section concludes with a basic configuration example with OSPFv2 as an underlay routing protocol.

Note: The topology below is used for all OSPF underlay routing configuration examples.

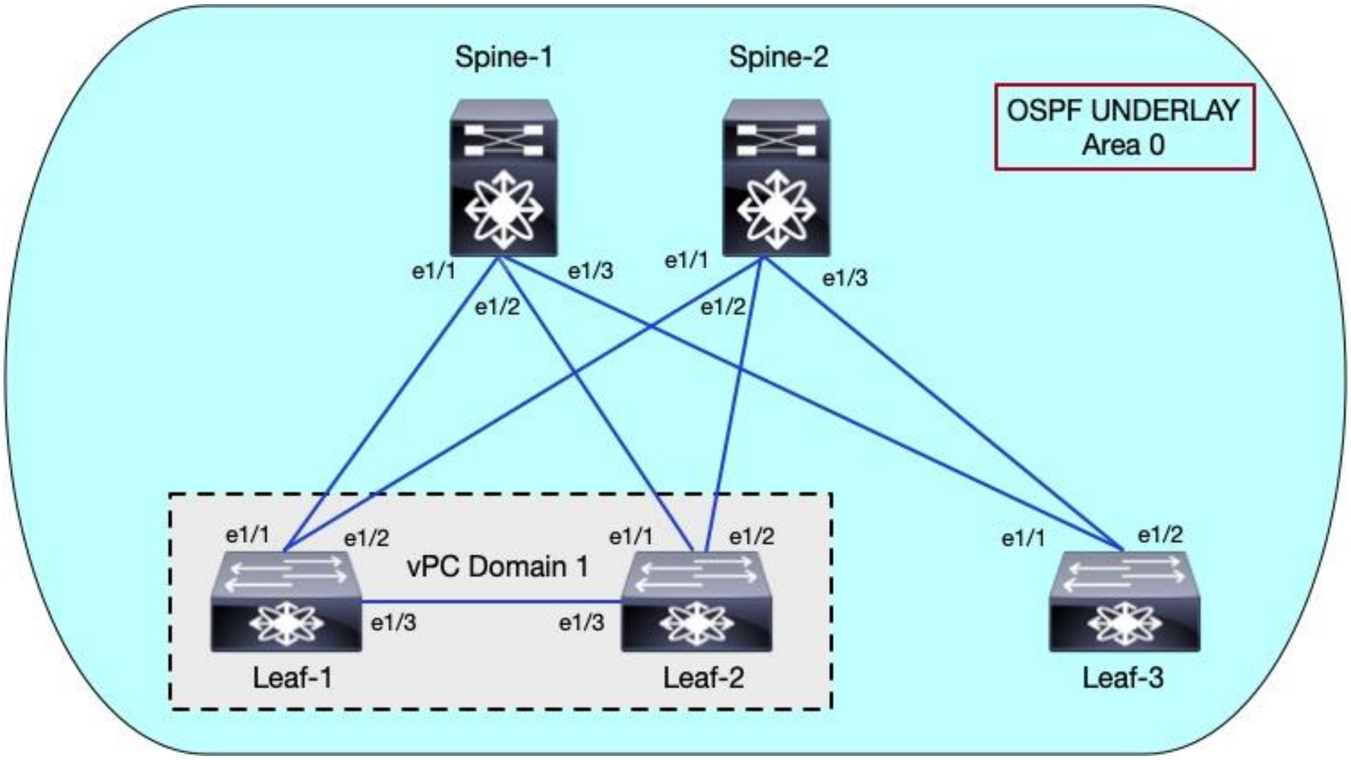


Figure 13. OSPFv2 Underlay Routing Topology

Leaf Switch

The configuration of OSPFv2 as an underlay routing protocol in a VXLAN BGP EVPN Fabric consists of the following steps on the leaf switches:

Step 1. Enable OSPF, Interface VLAN, vPC, and LACP features. The Interface VLAN feature is required to configure SVIs. The LACP feature is needed to negotiate port channels dynamically. The VPC feature enables virtual port channels.

```
feature ospf
feature interface-vlan
feature lacp
feature vpc
```

Step 2. Configure OSPF process UNDERLAY

```
router ospf UNDERLAY
router-id 192.168.3.3/32
```

Step 3. Configure loopback 0 for OSPF router-id and loopback1 for NVE interface on the leaf switches.

```
interface loopback0
    no shutdown
    description Router-ID Interface
    ip address 192.168.3.3/32
    ip router ospf UNDERLAY area 0.0.0.0

interface loopback 1
    no shutdown
    description VTEP NVE interface
    ip address 192.168.3.4/32
    ip address 192.168.34.34/32 secondary
    ip router ospf UNDERLAY area 0.0.0.0
```

Step 4. Configure fabric interfaces connecting the leaf to the spine on the leaf switch. Below is an example of Leaf-1 using unnumbered interfaces.

```
interface ethernet 1/1
    no shutdown
    description link to Spine-1 eth 1/1
    no switchport
    mtu 9216
    medium p2p
    ip unnumbered loopback 0
    ip router ospf UNDERLAY area 0.0.0.0
    ip ospf network point-to-point
    no ip redirect
    no ipv6 redirect
interface ethernet 1/2
    no shutdown
    description link to Spine-2 eth1/1
    no switchport
    mtu 9216
    medium p2p
    ip unnumbered loopback 0
    ip router ospf UNDERLAY area 0.0.0.0
    ip ospf network point-to-point
    no ip redirect
    no ipv6 redirect
```

Step 5. Configure vPC on Leaf-1 and Leaf-2. The ethernet 1/3 is the physical VPC peer link. The VLAN3600 is the backup SVI for traffic redirection through the peer link for fabric interface failure events. The example below shows the configuration for Leaf-1. A similar configuration applies to Leaf-2.

```
feature vpc
    interface ethernet 1/3
    no shutdown
    description member of port-channel 500 (vpc-peer-link)
    switchport
    switchport mode trunk
    channel-group 500 mode active

vpc domain 1
    peer-switch
    peer-gateway
    peer-keepalive destination 10.0.4.15 source 10.0.4.14
    delay restore 150
    auto-recovery reload-delay 360
    ipv6 nd synchronize
    ip arp synchronize

int port-channel 500
    no shutdown
    description "vpc-peer-link"
    switchport
    switchport mode trunk
    spanning-tree port type network
    vpc peer-link

vlan 3600
    name svi_vpc
    interface vlan 3600
    no shutdown
    description SVI for the vPC Peer link
    mtu 9216
    ip address 172.16.34.3 255.255.255.0
    no ip redirects
    no ipv6 redirects
    ip router ospf UNDERLAY area 0.0.0.0
    no shutdown
```

Spine Switch

The configuration of OSPFv2 as an underlay routing protocol in a VXLAN BGP EVPN Fabric consists of the following steps on the Spine switches:

Step 1. Enable the OSPF feature and process UNDERLAY.

```
feature ospf
router ospf UNDERLAY
router-id 192.168.1.1
```

Step 2. Configure Loopback0 for OSPF router-id.

```
interface loopback0
no shutdown
Description Routing-ID Interface
    ip address 192.168.1.1/32
    ip router ospf UNDERLAY area 0.0.0.0
```

Step 3. Configure the fabric interfaces connecting to the leaf switches. Below example uses unnumbered interfaces.

```
interface ethernet 1/1
    no shutdown
    description link to Leaf-1 eth 1/1
    no switchport
    mtu 9216
    medium p2p
    ip unnumbered loopback 0
    ip router ospf UNDERLAY area 0.0.0.0
    ip ospf network point-to-point
    no ip redirect
    no ipv6 redirect
interface ethernet 1/2
    no shutdown
    description link to Leaf-2 eth 1/1
    no switchport
    mtu 9216
    medium p2p
    ip unnumbered loopback 0
    ip router ospf UNDERLAY area 0.0.0.0
    ip ospf network point-to-point
    no ip redirect
    no ipv6 redirect
```

As a verification step, check if your vPC peering is up and that your OSPF neighbors are adjacent on the VTEPs.

```
leaf-1# show vpc
    Legend:
                    (*) - local vPC is down, forwarding via vPC peer-link
    vPC domain id                     : 1
    Peer status                       : peer adjacency formed ok
    vPC keep-alive status             : peer is alive
    Configuration consistency status  : success
    Per-vlan consistency status       : success
    Type-2 consistency status         : success
    vPC role                           : primary
    Number of vPCs configured          : 0
    Peer Gateway                       : Enabled
    Dual-active excluded VLANs         : -
    Graceful Consistency Check         : Enabled
    Auto-recovery status               : Enabled, timer is off.(timeout = 360s)
    Delay-restore status               : Timer is off.(timeout = 150s)
    Delay-restore SVI status           : Timer is off.(timeout = 10s)
    Operational Layer3 Peer-router     : Disabled
    Virtual-peerlink mode              : Disabled

    vPC Peer-link status
    ---------------------------------------------------------------------
    id    Port   Status Active vlans
    --    ----   ------ -------------------------------------------------
    1     Po500  up     1,3600

On Leaf-1 the peer link Po500 is up, adjacency is formed, and peer is detected as alive.
leaf-1# show ip ospf neighbors
      OSPF Process ID UNDERLAY VRF default
      Total number of neighbors: 3
      Neighbor ID     Pri State            Up Time  Address         Interface
      192.168.1.1       1 FULL/ -          02:33:06 192.168.1.1     Eth1/1
      192.168.2.2       1 FULL/ -          02:33:13 192.168.2.2     Eth1/2
      192.168.4.4       1 FULL/BDR         01:23:56 172.16.34.4     Vlan3600
    pod-4-leaf-1#
    vPC Peer-link status
    ---------------------------------------------------------------------
    id    Port   Status Active vlans
    --    ----   ------ -------------------------------------------------
    1     Po500  up     1,3600
```

Leaf-1 is OSPF neighbors with Spine-1 and Spine-2. The last OSPF peer is Leaf-2, which peers through the backup SVI VLAN 3600.

ISIS Routing

The Intermediate-to-Intermediate System (IS-IS) protocol is a link-state routing protocol based on the OSI standard. In IS-IS, the intermediate system refers to a router, and the end system refers to a host connected to the IS-IS network. The protocol was designed initially to route the Connectionless Network Protocol (CLNP). Later, it added support for routing IPv4 and IPv6, commonly referred to as integrated ISIS. The IS-IS protocol carries any routing information in the Type Length Value (TLV) fields in the IS-IS link state packets.

ISIS is a hierarchical protocol but more flexible than OSPF because it does not require all inter-area routes to traverse the backbone area. A two-level hierarchy defines the ISIS network hierarchy. Level 2 is the backbone level, and level 1 is a stub level, like a totally stubby area in OSPF. Stub areas in link state protocols do not allow inter-area or external routing information into the area. All inter-area routes in ISIS must use an exit node called a level-1-2 router, like an area border router in OSPF. The level-1-2 router peers with routers inside a level-1 stub area and backbone level-2 routers. The level-1-2 router will have an independent topology database for level-2 and level-1 peers. The level-1 routers inside the stub area will only have link state information of routers within their area. The level-1 routers use a default route to reach the level-1-2 router to exit its area.

The Cisco Nexus 9000 NX-OS platform has validated the use of IS-IS Level 1 and IS-IS Level 2 only as an underlay routing protocol on all VXLAN BGP EVPN fabric nodes. This implies that in a VXLAN BGP EVPN site, all the leaves and spines must be configured as IS-IS Level 1 or Level 2 nodes.

The advantages of running ISIS in a VXLAN BGP EVPN fabric as an underlay protocol are:

Fast convergence

ISIS uses incremental SPF. IP Prefix information is leaves (encoded using TLVs) in the branch of the SPF tree. Any change in the prefix state does not trigger a complete SPF calculation; it only triggers a partial SPF for the area of the topology where the prefix has changed state, allowing for fewer computational resources to calculate a new path.

Extensible with multiprotocol support

Like BGP, ISIS is an extensible protocol. As mentioned, IPv6/IPv4 prefix information is encoded as TLVs in ISIS LSPs. ISIS routes over CLNS protocol, so it is not dependent on IP to build an SPF tree. ISIS has been used as an underlay in Cisco Fabric Path layer 2 overlay to route MAC addresses. ISIS is also an underlay with Cisco ACI fabric to carry VTEP reachability information. ISIS is also used in OTV to carry MAC addresses. ISIS supports IPv4 and IPv6 routing in a single routing instance. OSPFv2 only supports IPv4 routing, while OSPFv3 supports IPv6 and IPv4 routing, but IPv6 is mandatory. Thus, you cannot use OSPFv3 to implement only an IPv4 underlay. If there is the possibility of migrating from IPv4 to IPv6 routing in the underlay, then ISIS would be a good choice because, from the same routing protocol instance, both address families are supported without having to replace or run multiple routing protocols.

The main disadvantage of ISIS is that it uses networking terminology that is very different from OSPF, a much more common IP routing protocol in the enterprise. It also uses a different routed protocol with a different address structure. A network address in the CLNS protocol is called a Network Service Access Point (NSAP). An NSAP address assigned as the router ID of an ISIS router is called a Network Entity Title address. The NET address is an NSAP address with an NSAP value equal to zero. A single NET address is assigned to a router. The format of an NET address is shown below.

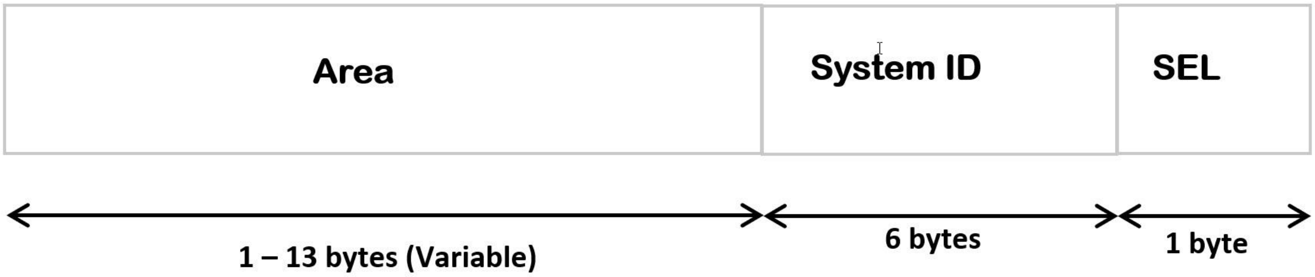


Figure 14. ISIS NET Address Format

The NET address comprises the IS-IS system ID, uniquely identifying this IS-IS instance in the area and the area ID. For example, if the NET ID is 49.0001.0010.0100.1074.00, the system ID is 0010.0100.1074, and the area ID is 49.0001.

This section concludes with a basic configuration example with ISIS as an underlay routing protocol.

Note: The topology below is used for all ISIS underlay routing configuration examples.

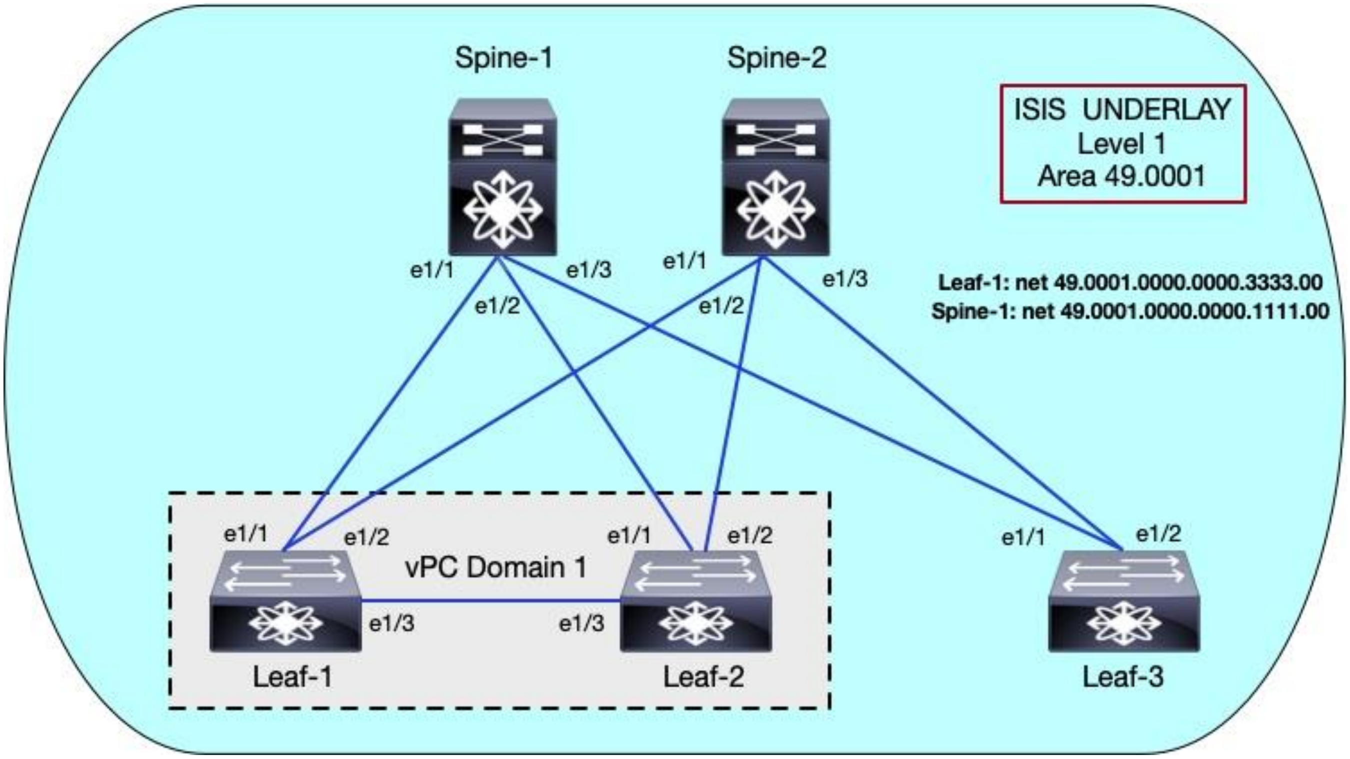


Figure 15. ISIS underlay routing topology

Leaf Switch

The configuration of ISIS as an underlay routing protocol in a VXLAN BGP EVPN Fabric consists of the following steps on the leaf switches:

Step 1. Enable ISIS, Interface VLAN, VPC, and LACP features. The Interface VLAN feature is required to configure SVIs. The LACP feature is needed to negotiate port channels dynamically. The VPC feature enables virtual port channels.

```
feature isis
feature interface-vlan
feature lacp
feature vpc
```

Step 2. Configure ISIS process UNDERLAY. The NET address is required.

```
router isis UNDERLAY
    net 49.0001.0000.0000.3333.00
    is-type level-1
    set-overload-bit on-startup 60
```

Note: The **set-overload-bit** command on the Nexus NX-OS switch signals other devices not to use the switch as a transit hop when calculating the shortest path. You can optionally configure the overload bit temporarily on startup. In the above example, the **set-overload-bit** command sets the overload bit on startup to 60 seconds, which gives the underlay and overlay protocols 60 seconds to converge during startup.

Step 3. Configure loopback 0 for BGP peering and loopback1 for NVE interface on the leaf switches.

```
interface loopback0
    no shutdown
    Description Routing-ID loopback Interface
    ip address 192.168.3.3/32
    ip router isis UNDERLAY
interface loopback 1
    no shutdown
    description VTEP NVE loopback interface
    ip address 192.168.3.4/32
    ip address 192.168.34.34/32 secondary
    ip router isis UNDERLAY
```

Step 4. Configure fabric interfaces connecting the leaf to the spine on the leaf switch. Below is an example of Leaf-1 using unnumbered interfaces.

```
interface ethernet 1/1
    no shutdown
    description link to Spine-1 eth 1/1
    no switchport
    mtu 9216
    medium p2p
    ip unnumbered loopback 0
    ip router isis UNDERLAY
    no ip redirect
    no ipv6 redirect
interface ethernet 1/2
    no shutdown
    description link to Spine-2 eth1/1
    no switchport
    mtu 9216
    medium p2p
    ip unnumbered loopback 0
    ip router isis UNDERLAY
    no ip redirect
    no ipv6 redirect
```

Step 5. Configure vPC on Leaf-1 and Leaf-2. The ethernet 1/3 is the physical VPC peer link. The VLAN3600 is the backup SVI for traffic redirection through the peer link for fabric interface failure events. The example below shows the configuration for Leaf-1. A similar configuration applies to Leaf-2.

```
interface ethernet 1/3
    no shutdown
    description member of port-channel 500 (vpc-peer-link)
    switchport mode trunk
    channel-group 500 mode active

vpc domain 1
    peer-switch
    peer-gateway
    peer-keepalive destination 10.0.4.15 source 10.0.4.14
    delay restore 150
    auto-recovery reload-delay 360
    ipv6 nd synchronize
    ip arp synchronize

int port-channel 500
    no shutdown
    description "vpc-peer-link"
    switchport mode trunk
    spanning-tree port type network
    vpc peer-link

vlan 3600
    name svi_vpc
    interface vlan 3600
    description SVI for the vPC Peer link
    mtu 9216
    ip address 172.16.34.3 255.255.255.0
    no ip redirects
    no ipv6 redirects
    ip router isis UNDERLAY
    no shutdown
```

Spine Switch

The configuration of ISIS as an underlay routing protocol in a VXLAN BGP EVPN Fabric consists of the following steps on the Spine switches:

Step 1. Enable the ISIS feature and process UNDERLAY.

```
feature isis
router isis UNDERLAY
    net 49.0001.0000.0000.1111.00
```

```
        is-type level-1
        set-overload-bit on-startup 60
```

Step 2. Configure Loopback0 for BGP peering.

```
    interface loopback0
        no shutdown
        Description Routing-ID loopback Interface
        ip address 192.168.1.1/32
        ip router isis UNDERLAY
```

Step 3. Configure the fabric interfaces connecting to the leaf switches. Below example uses unnumbered interfaces.

```
    interface ethernet 1/1
        no shutdown
        description link to Leaf-1 eth 1/1
        no switchport
        mtu 9216
        medium p2p
        ip unnumbered loopback 0
        ip router isis UNDERLAY
        no ip redirect
        no ipv6 redirect

    interface ethernet 1/2
        no shutdown
        description link to Leaf-2 eth1/1
        no switchport
        mtu 9216
        medium p2p
        ip unnumbered loopback 0
        ip router isis UNDERLAY
        no ip redirect
        no ipv6 redirect
```

As a verification step, check if your vPC peering is up and that the ISIS neighbors are adjacent on the VTEPs.

```
leaf-1# show vpc
    Legend:
                    (*) - local vPC is down, forwarding via vPC peer-link
    vPC domain id                       : 1
    Peer status                         : peer adjacency formed ok
    vPC keep-alive status               : peer is alive
    Configuration consistency status    : success
    Per-vlan consistency status         : success
    Type-2 consistency status         : success
    vPC role                          : primary
    Number of vPCs configured         : 0
    Peer Gateway                      : Enabled
    Dual-active excluded VLANs        : -
    Graceful Consistency Check        : Enabled
    Auto-recovery status              : Enabled, timer is off.(timeout = 360s)
    Delay-restore status              : Timer is off.(timeout = 150s)
    Delay-restore SVI status          : Timer is off.(timeout = 10s)
    Operational Layer3 Peer-router    : Disabled
    Virtual-peerlink mode             : Disabled

    vPC Peer-link status
    ---------------------------------------------------------------------
    id    Port   Status Active vlans
    --    ----   ------ -------------------------------------------------
    1     Po500  up     1,3600
```

On Leaf-1 the peer link Po500 is up, adjacency is formed, and peer is detected as alive.

```
spine-1# show isis adjacency
    IS-IS process: UNDERLAY VRF: default
    IS-IS adjacency database:
    Legend: '!': No AF level connectivity in given topology
    System ID       SNPA            Level  State  Hold Time  Interface
    Leaf-1          N/A             1      UP     00:00:25   Ethernet1/1
    Leaf-2          N/A             1      UP     00:00:24   Ethernet1/2
    0000.0000.5555  N/A             1      UP     00:00:25   Ethernet1/3
```

The Spine-1 has ISIS adjacency with all the VTEPs. The Leaf-1 also has peering with both spines and its peer vPC switch Leaf-2.

```
leaf-1# show isis adjacency
    IS-IS process: UNDERLAY VRF: default
    IS-IS adjacency database:
    Legend: '!': No AF level connectivity in given topology
    System ID       SNPA            Level  State  Hold Time  Interface
    Leaf-2          5000.0003.0007  1      UP     00:00:08   L2_VLAN3600
    Spine-1         N/A             1      UP     00:00:22   Ethernet1/1
    Spine-2         N/A             1      UP     00:00:29   Ethernet1/2
```

VXLAN BGP EVPN Underlay Multicast Routing

Broadcast, unknown unicast, and multicast (BUM) packets are multi-destination traffic. The unicast routing table does not contain BUM addresses. The VTEP, by default, uses flooding to forward BUM traffic. The BUM traffic will arrive on every VTEP with the L2 VNI configured. The L2 VNI is a broadcast domain in a

VXLAN fabric. The Nexus 9000 NX-OS operating system maps a single VLAN to a single broadcast domain. BUM traffic flooding is handled using ingress replication (IR) or multicast routing in the underlay.

The IR is the unicast mode of handling BUM traffic. A host sends a single BUM packet to the VTEP. The VTEP device encapsulates the BUM traffic into a VXLAN packet, multiplies it, and creates copies equal to the number of NVE peers. The below picture displays IR.

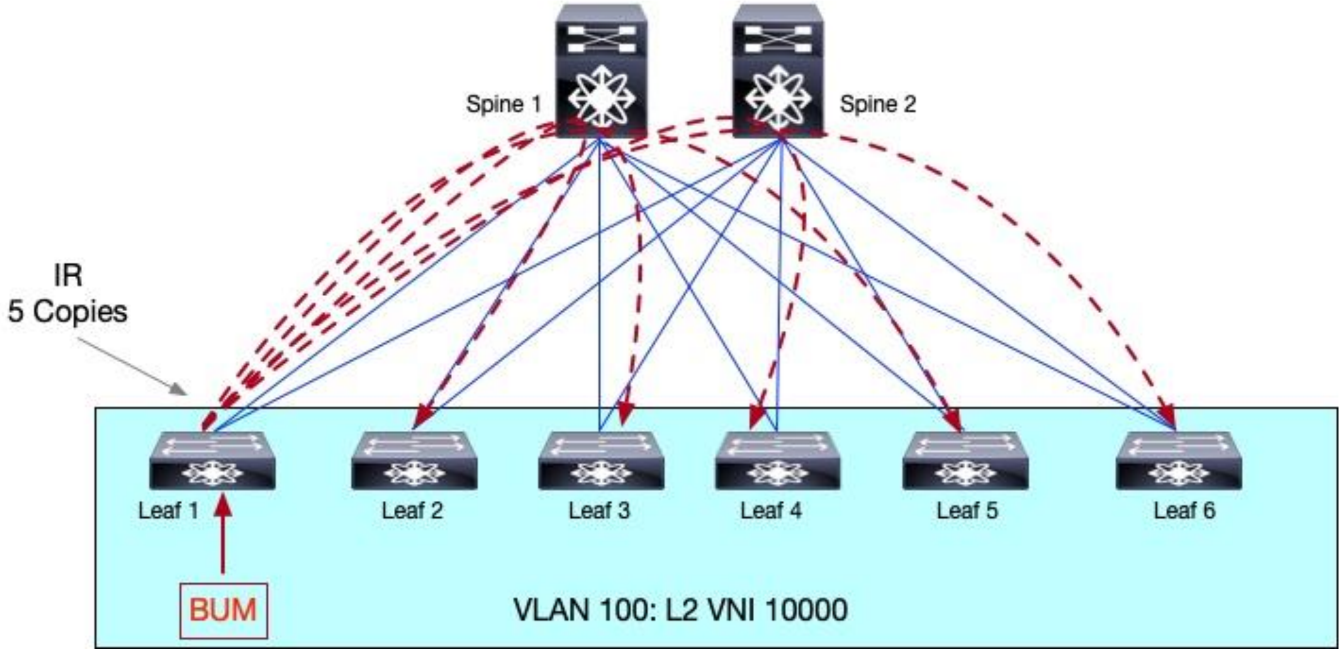


Figure 16. Ingress Replication

The NVE peer list on a VTEP can be statically or dynamically maintained. To configure a static NVE peer, apply the following commands:

```
interface nve1
  source-interface loopback 1
  member vni 30001
  ingress-replication protocol static
  peer-ip <remote vtep IP>
```

As networks are removed and added according to the movement of workloads in the data center network, the NVE neighbor table will have to be updated to ensure the BUM traffic for each VNI is replicated to remote VTEPs that are members of the VNI. Manually updating NVE neighbor tables across all the VTEPs is operationally challenging. It will create situations when VTEPs receive BUM traffic when not required or VTEPs don’t receive BUM traffic when they should. Dynamic IR uses the BGP EVPN Type 3 – Inclusive Multicast Ethernet Tag Route to create a list of remote VTEPs for ingress replication. The BGP EVPN Type 3 routes are sent to all remote VTEPs as soon as a VNI is provisioned on a VTEP to signal interest in receiving BUM traffic. The example below provides the basic commands for dynamic IR to define a VTEP NVE neighbor for VNI 30001.

```
interface nve1
  source-interface loopback 1
  member vni 30001
  ingress-replication protocol bgp
```

To verify the NVE peers on a VTEP, use the **show nve peers** command. Notice that the output below in the LearnType column indicates CP. CP stands for control plane learning, meaning the NVE peer was learned dynamically using BGP EVPN.

```
switch# show nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      1.1.1.53         Up    CP        05:21:58 f8c2.8846.e07f
nve1      1.1.1.112        Up    CP        07:08:37 2cd0.2d51.9f1b
nve1      1.1.1.114        Up    CP        07:04:49 00a6.cab6.bbbb
```

The multicast routing instance in the underlay allows optimal forwarding of BUM traffic between the VTEPs received on the tenant interfaces. The multicast protocols supported in the underlay on Nexus 9000 NX-OS switches are PIM ASM Sparse Mode and PIM BiDir. PIM Sparse mode creates one source tree per VTEP per multicast group. PIM BiDir creates one shared tree per multicast group. The RP redundancy mechanism varies between the two flavors of PIM ASM.

The PIM Sparse Mode uses PIM Anycast RP (RFC 4610) for RP redundancy and sharing between multiple RPs. A group of routers called the Anycast-RP set share a common IP address, allowing multiple routers to act as RPs for a single multicast group. In a VXLAN fabric, the spine serves as RPs for the underlay. The FHR leaf registers their attached sources to one of the RPs. Each RP with registered sources sync S,G state information using a unique IP address assigned to each RP providing active/active redundancy. Thus, two loopbacks are required for PIM Anycast-RP on the RPs, one loopback that uniquely identifies each RP and one loopback with the common anycast IP address.

The topology and sample configuration elaborate on how PIM is configured on the leaf and spines with PIM Anycast-RP.

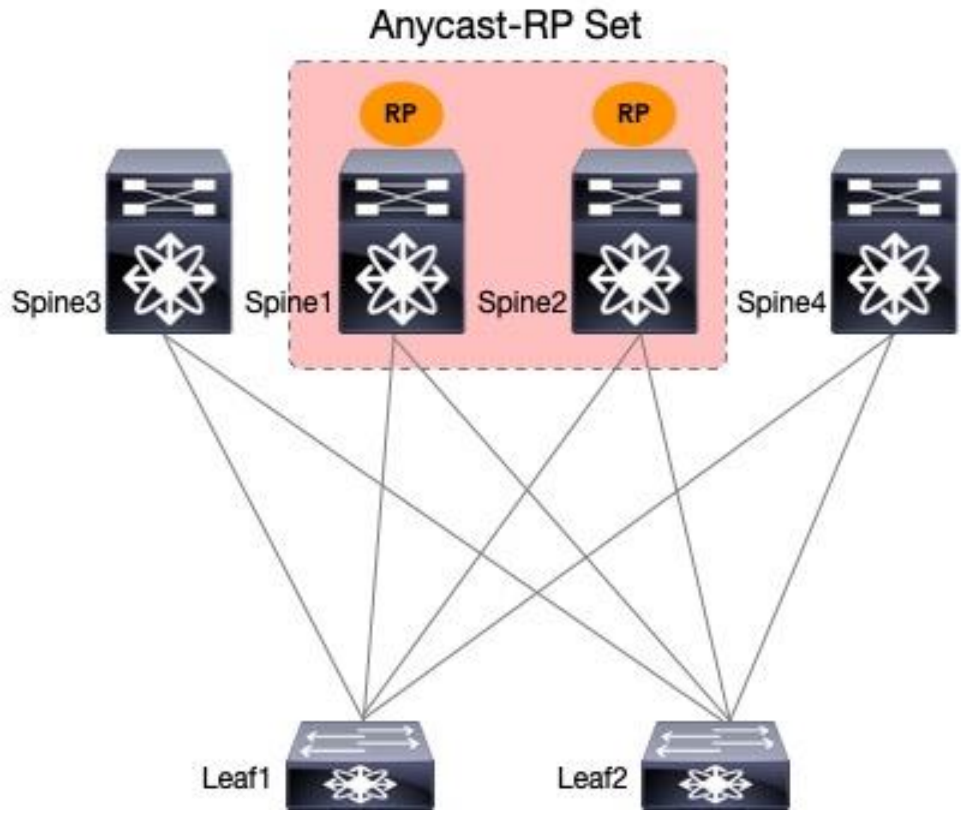


Figure 17. PIM Anycast RP

Spine-1 PIM Anycast-RP Configuration

```
interface loopback0
 description RID
 ip address 192.168.1.1/32
 ip pim sparse-mode

interface loopback254
 description RP
 ip address 192.168.100.100/32

ip pim rp-address 192.168.100.100 group-list 239.0.0.0/24
ip pim anycast-rp 192.168.100.100 192.168.1.1
ip pim anycast-rp 192.168.100.100 192.168.2.2
```

Note: Spine-2 will have a similar configuration to Spine-1 as shown above.

Leaf-1 PIM Anycast-RP Configuration

```
ip pim rp-address 192.168.100.100 group-list 239.0.0.0/24
```

PIM ASM Bidir uses phantom RP as its RP redundancy mechanism. The phantom RP provides active/standby redundancy. The spines have a dedicated loopback interface for RP addresses with the same IP address but a different subnet mask length. The primary RP will have a longer subnet mask length than the secondary RP for the same multicast group. The RP address specified in all the nodes will be a different IP address in the same subnet as the RP loopbacks of the spines. The RP address is not an IP address applied to any interfaces.

The topology and sample configuration elaborate on how PIM is configured on the leaf and spines with **PIM Phantom RP**.

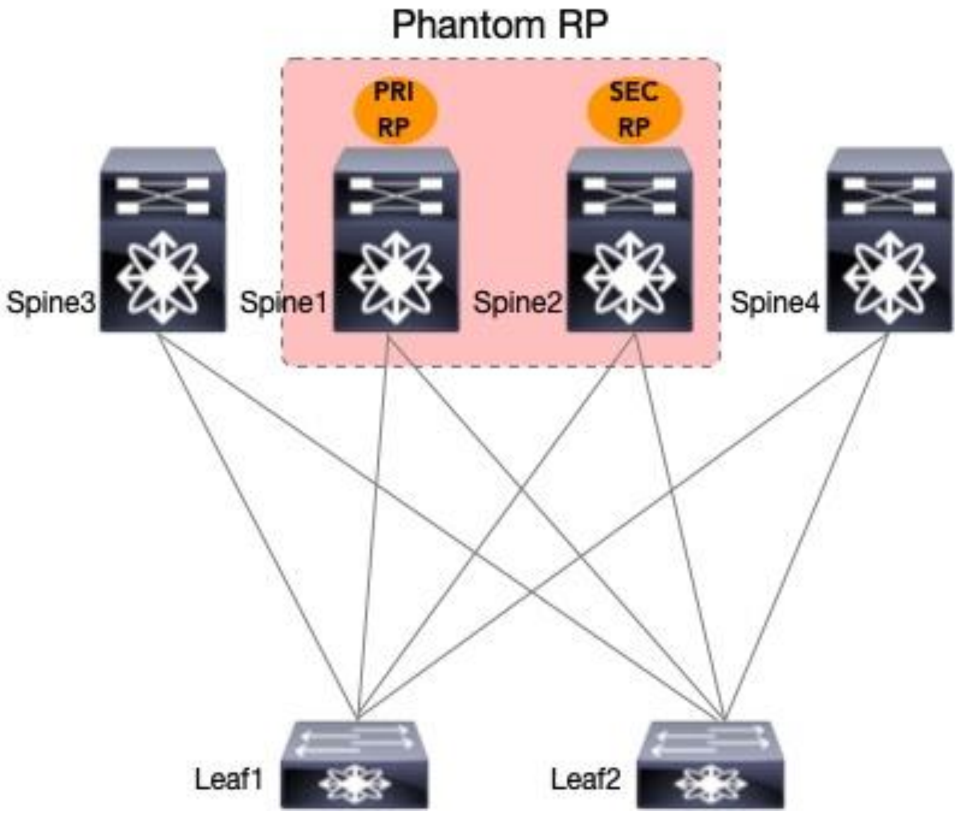


Figure 18. PIM Phantom RP

Spine-1 PIM Phantom RP Configuration

```
interface loopback254
  description RP
  ip address 192.168.100.2/30
  ip pim sparse-mode

interface loopback0
  description RID
  ip address 192.168.1.1/32
  ip pim sparse-mode

ip pim rp-address 192.168.100.1 group-list 239.0.0.0/24 bidir
```

Spine-2 PIM Phantom RP Configuration

```
interface loopback254
```

```
  description RP
  ip address 192.168.100.2/29
  ip pim sparse-mode

interface loopback0
  description RID
  ip address 192.168.2.2/32
  ip pim sparse-mode

ip pim rp-address 192.168.100.1 group-list 239.0.0.0/24 bidir
```

Leaf-1 PIM Phantom RP Configuration

```
ip pim rp-address 192.168.100.1 group-list 239.0.0.0/24 bidir
```

Underlay Multicast Design Consideration

The decision to use IR or multicast in the underlay to handle BUM depends on the answer to the following:

- Will there be requirements to route multicast traffic in the overlay?

  Tenant Routed Multicast (TRM) provides multicast routing capabilities in a VXLAN BGP EVPN fabric using MP-BGP NGMVPN control plane signaling.

  Note: To read more about VXLAN BGP EVPN TRM, please reference the white paper titled “Tenant Routed Multicast in Cisco Nexus 9000 VXLAN BGP EVPN Data Center Fabrics - Fundamental Concepts and Architectures” on cisco.com.

  TRM requires a PIM ASM underlay multicast routing. TRM is not supported with PIM Bidir in the underlay. If multicast routing is introduced in the future, it is better to implement a multicast underlay to support the potential requirement proactively. Migrating from IR to PIM ASM multicast underlay will be a complex and impactful implementation, causing traffic outages.

- What is the data center operations team’s comfort level with multicast routing?

  The network team needs more expertise in multicast routing and is uncomfortable managing any infrastructure with multicast technologies. Therefore, the team decided not to implement PIM in the underlay.

- The VTEP performing the IR has abundant network packet processing power, and BUM traffic bandwidth consumption is negligible.

  The BUM traffic profile impacts the VTEP's work by creating copies and utilizing the fabric interface bandwidth capacity. If the BUM traffic is a multicast packet that contains some text data for a market feed for a stock market trading floor financial application. Then, the amount of replicated data is small at the head end VTEP. But suppose the BUM traffic is a 4K video multicast stream. In that case, the VTEP will have to create copies of a much larger data set, delaying packet processing, and the duplicate copies of the video streaming application will consume more bandwidth in the fabric. If the VTEP is a platform that can perform IR in hardware, has good backplane capacity, and the BUM traffic is not high bandwidth-consuming, IR is a viable option. If the decision is to proceed with IR in the underlay, always verify the VTEP peer scale number for the VTEP. If the VTEP peer scale for the VTEP is 254, only 254 can receive the IR traffic a VTEP.

Underlay Multicast Design Considerations

To conclude this section, underlay multicast design considerations are listed below.

- The spines should act as the Rendezvous-Point for the underlay multicast domain.
- Use RP redundancy models such as PIM Anycast RP or Phantom RP across multiple spines.
- The scale of the underlay multicast group is 512. Reserve a range of multicast groups for L2VNI and L3VNI (TRM).
- It is possible to map different VNIs to different multicast groups mapped to different RP for load balancing.
- The same replication protocol and mode must be configured on all the VTEPs in a single site.
- The same L2VNI to multicast group mapping must be configured across all VTEPs.
- Configuring IR or multicast on an L2 VNI basis is possible for BUM traffic handling.

VXLAN BGP EVPN Overlay Unicast Routing

The MP-BGP EVPN control plane function is unicast bridging and routing. The MP-BGP EVPN AF can advertise internal endpoints' MAC addresses using the EVPN Type 2 MAC route and advertise internal endpoints' IP addresses using the EVPN Type 2 MAC-IP route. The internal subnet prefixes and the external IP network advertisements are signaled using the EVPN Type 5 IP Prefix route. One of the benefits of VXLAN BGP EVPN is the ability to do integrated routing and bridging (IRB) using a single control plane and address family.

The underlay routing protocol design was explained in the previous section. This section focuses on overlay routing design and implementation. The overlay will only be discussed for unicast routing. The overlay multicast routing is discussed in detail in the following white paper on cisco.com: "Tenant Routed Multicast in Cisco Nexus 9000 VXLAN BGP EVPN Data Center Fabrics - Fundamental Concepts and Architectures".

The overlay tunnel encapsulates the tenant/VRF traffic and routes the packet from the ingress VTEP to the egress VTEP. Each VTEP is the location of the attached tenants/VRFs. The VTEPs advertise the tenant routes to the spine using MP-BGP EVPN. Hence, understanding the device's role in the fabric determines what type of routes the node must advertise. The type of routes the node must advertise determines the kind of networking protocols and configurations that need to be enabled on the node. For example, the spine in an iBGP Fabric with no other function, such as border gateway or border leaf, must only take EVPN routes learned from the source VTEPs and reflect them to the destination VTEPs. In a spine node, VXLAN tunnel or VRF-related configurations are not required.

The underlay network has three unicast routing protocol options: OSPF, ISIS, and eBGP. The overlay routing is only MP-BGP EVPN, but the overlay EVPN peering can be iBGP or eBGP. The table below summarizes the underlay and overlay protocol combination supported on Nexus 9000 NX-OS IPv4 fabric.

Underlay Protocol	Overlay EVPN Peering
OSPFv2	iBGP or eBGP
ISIS	iBGP or eBGP
eBGP	eBGP

Table 3. Underlay and overlay unicast routing combinations

Note: Cisco recommends using an IGP in the underlay with an iBGP EVPN overlay peering.

The following section provides an example of implementing BGP EVPN overlays for iBGP fabrics.

iBGP Fabrics

In an iBGP fabric, the leaf and spine are part of one autonomous system. The spine's role in the EVPN overlay is to take the routes learned from the leaf and propagate them to the other leaves in the fabric. The spine is a BGP route reflector in an iBGP fabric. An important design decision is how many route reflectors should be implemented. At least two route reflectors should be implemented in a fabric with two or more spines. The two route reflectors ensure high availability while not consuming unnecessary memory resources on the leaf. Every route reflector introduced into the fabric creates one additional copy of the same route a leaf must store in its BGP table. Every leaf should be a client of every route reflector. The peering sessions should use the loopback interfaces as standard iBGP best practice.

The below topology is for the configuration example covering the steps to implement iBGP fabric.

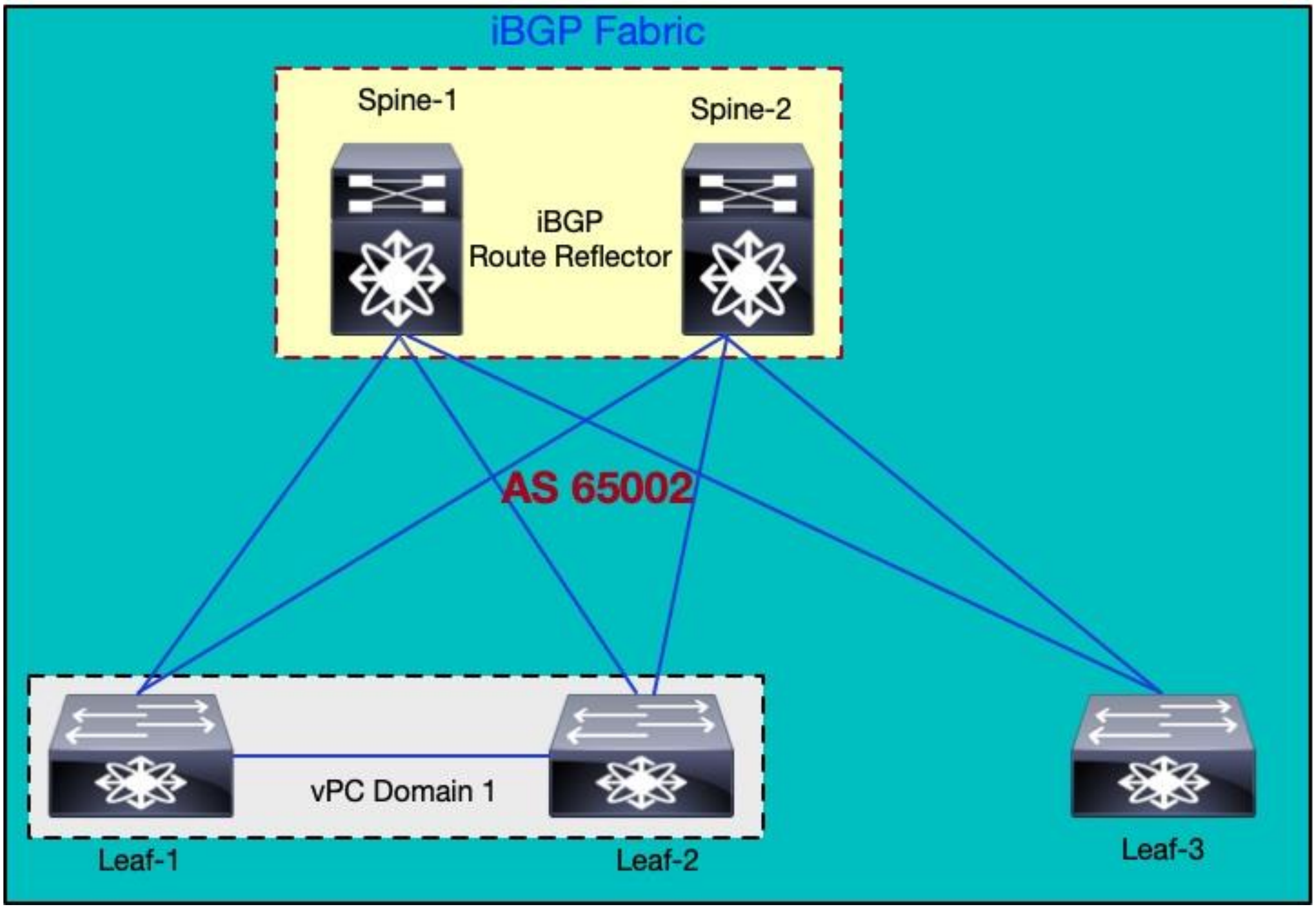


Figure 19. iBGP Fabric Topology

Note: The below configuration steps only outline the BGP EVPN overlay-related commands. The underlay commands explained in the previous section must be implemented before these steps.

The sequence of steps to provision the overlay is divided into three parts which are:

1. The tenant VRFs (IP-VRF) and VLANs (MAC-VRF) on the leaf.
2. The BGP EVPN neighbor configuration on leaf and spine.
3. The VXLAN NVE interface on the leaf.

Leaf Switch

The steps to configure the overlay network in an iBGP EVPN VXLAN fabric on the leaf devices are the following:

Step 1. Enable the VXLAN and MP-BGP EVPN feature.

```
feature nv overlay
feature vn-segment-vlan-based
feature bgp
nv overlay evpn
```

Step 2. Configure tenant VLANs for L2 and L3 VNI.

```
vlan 201
    name L2-VNI-201-Tenant1
    vn-segment 50201

vlan 202
    name L2-VNI-202-Tenant1
    vn-segment 50202

vlan 999
    name L3-VNI-999-Tenant1
    vn-segment 50999
```

The above commands map VLANs to VNIs. VLAN 201 to VLAN 202 are for bridging (L2VNIs) while VLAN 999 is for routing (L3VNIs). Beginning with Cisco NX-OS Release 10.2(3)F, the new L3VNI mode is supported on Cisco Nexus 9000 switches. The new CLI for L3 VNI does not require mapping a VLAN to L3VNI, which also removes the requirement to provision an SVI interface, saving on VLANs and increasing the scale of VNIs supported on a given leaf node. The L3VNI is created under the VRF as shown below.

```
vrf context vxlan-999
    vni 50999 L3
```

See *Cisco Nexus 9000 Series NX-OS VXLAN Configuration Guide* for more guidelines and recommendations on new L3VNI CLI commands.

Step 3. Create IP-VRF and the SVI interface for L3 VNI.

```
vrf context Tenant-1
    vni 50999
    rd auto
    address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
```

```
interface vlan999
    no shutdown
    vrf member Tenant-1
    ip forward
```

Note: The new L3VNI cli will not require the creation of an SVI interface for VRF. The new L3 VNI cli mode will also increase the L2 VNI scale to 4000 and L3 VNI scale to 2000. Please reference the Nexus 9000 NXOS Verified Scalability Guide.

Step 4. Configure anycast gateway MAC address.

```
fabric forwarding anycast-gateway-mac 0000.2222.3333
```

Step 5. Create MAC-VRF and the SVI interfaces for L2VNI.

```
interface Vlan201
    no shutdown
    vrf member Tenant-1
    no ip redirects
    ip address 172.16.201.1/24
    fabric forwarding mode anycast-gateway

interface Vlan202
    no shutdown
    vrf member Tenant-1
    no ip redirects
    ip address 172.16.202.1/24
    fabric forwarding mode anycast-gateway

evpn

vni 50201 l2
    rd auto
    route-target import auto
    route-target export auto

vni 50202 l2
    rd auto
    route-target import auto
    route-target export auto
```

Step 6. Configure the VXLAN NVE interface.

```
interface loopback1
    description VTEP NVE Interface
    ip address 192.168.3.4/32
    ip address 192.168.34.34/32 secondary
    ip router isis UNDERLAY

interface nve1
    no shutdown
    source-interface loopback1
    host-reachability protocol bgp
  member vni 50201
    mcast-group 239.0.0.201
  member vni 50202
    mcast-group 239.0.0.202
  member vni 50999 associate-vrf
```

Step 7. Configure the BGP EVPN peering to the spines.

```
route-map permit all
router bgp 65002
    address-family l2vpn evpn
    retain route-target all
  neighbor 192.168.1.1
    address-family l2vpn evpn
    send-community both
  neighbor 192.168.2.2
    address-family l2vpn evpn
    send-community both
vrf Tenant-1
    address-family ipv4 unicast
    redistribute direct route-map permitall
```

Spine Switch

The steps to configure the overlay network in an iBGP EVPN VXLAN fabric on the spine devices are the following:

Step 1. Enable MP-BGP EVPN feature.

```
feature bgp
nv overlay evpn
```

Note: If the spine’s only role is a spine and not a Border Spine or Border Gateway Spine, it is unnecessary to enable VXLAN features.

Step 2. Configure BGP EVPN peering to leaves.

```
router bgp 65002
    router-id 192.168.1.1
    log-neighbor-changes
    address-family ipv4 unicast
    address-family l2vpn evpn
```

```
template peer vtep-peer
  remote-as 100
  update-source loopback0
  address-family ipv4 unicast
    send-community both
  route-reflector-client
  address-family l2vpn evpn
    send-community both
  route-reflector-client
neighbor 192.168.3.3
  inherit peer vtep-peer
neighbor 192.168.4.4
  inherit peer vtep-peer
neighbor 192.168.5.5
  inherit peer vtep-peer
```

VXLAN BGP EVPN Edge Routing

The data center network hosts application data and services. Its servers provide application services for enterprise network domains such as campus, branches, other data centers, and remote sites. The data center network connects to external networks through the border leaf device.

The border leaf device connects to an external router called an edge router. The border leaf has fabric interfaces to the VXLAN BGP EVPN fabric that connects to the spines and external interfaces that connect to the edge router. The external interface to extend the tenant VRF is configured with the tenant VRF. The routing between the border leaf and edge router can be static or dynamic. The border leaf supports OSPF and eBGP routing for external network connectivity to the edge router.

The eBGP routing is the recommended protocol for external network connectivity between the border leaf and edge router. The BGP protocol advertises routes from EVPN to AF IPv4/IPV6 without redistribution, simplifying the configuration. The eBGP also has many routing manipulation and control policy capabilities, making it suitable for inter-domain network routing.

The VXLAN BGP EVPN tenant VRFs can terminate at the border leaf or extend to the edge router. If the tenant VRF extends from the border leaf to the edge router, network segmentation extends beyond the data center fabric. The extension of the tenant VRF to the edge router is called the VXLAN BGP EVPN to VRF-lite handoff, which is like MPLS Inter-AS Option A. The forwarding between the border leaf and edge router in the VRF-lite handoff is native IPv4 or IPv6 forwarding without encapsulation.

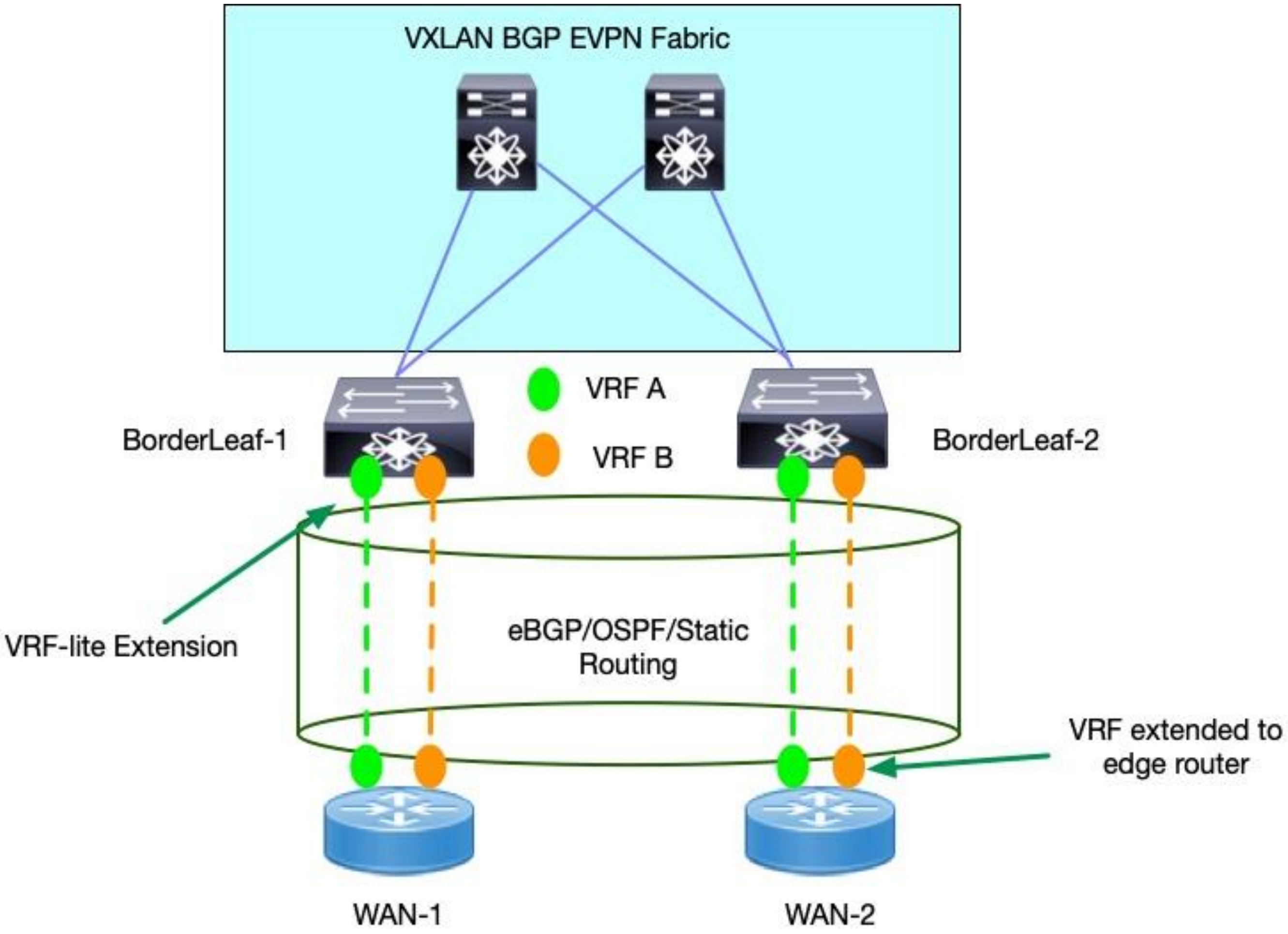


Figure 20. VRF-Lite Handoff

The edge router can also be a point of tenant VRF convergence. The border leaf can extend the tenant VRF to the edge router, but the edge router consumes all the tenant VRF routes into a single default VRF routing table. The edge router fuses the tenant VRFs, allowing inter-VRF routing through the edge router. The edge router in this implementation is called a fusion router.

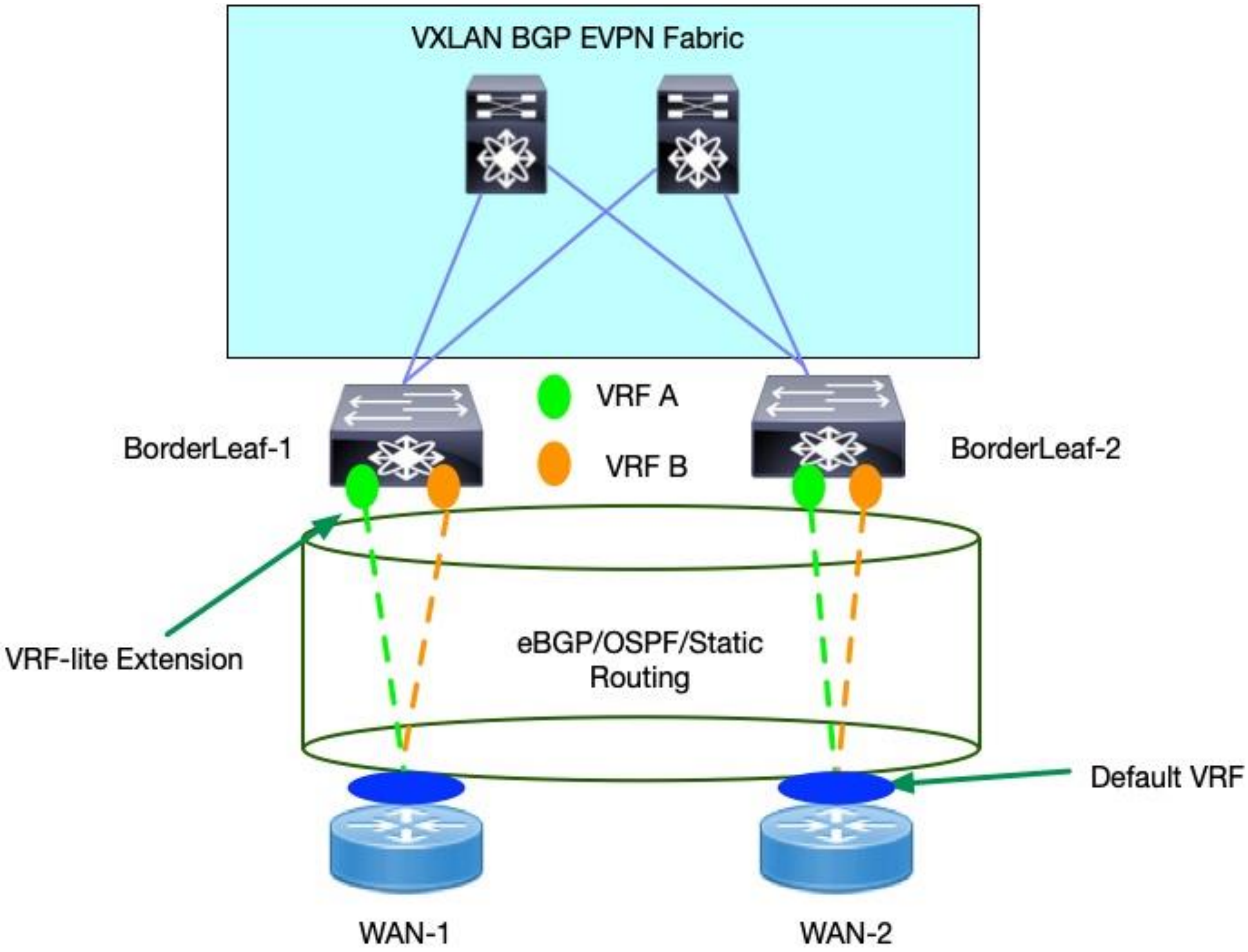


Figure 21. VRF-lite Handoff to Fusion Router

Note: A routed firewall can also be put in place of the edge router. Such a perimeter-routed firewall will act as an inter-tenant firewall.

The border leaf is the routing update control point for filtering the routes advertised from the data center out to the external network and from the external network into the data center. Using eBGP between the border leaf makes it suitable at such points in the network to allow network administrators to use BGP's powerful policy tools to define routing policy at the border leaf. To advertise externally learned into the data center as EVPN Type 5 prefix routes, the command advertise l2vpn evpn is applied under the BGP VRF.

```
router bgp 65002
 vrf Tenant-1
   address-family ipv4 unicast
    advertise l2vpn evpn
```

Suppose there is a requirement to optimize edge router TCAM resources, and the number of internal routes to the data center is substantial. In that case, the internal data center routes can be summarized using the aggregate-address command below.

```
router bgp 65002
```

```
 vrf Tenant-1
   address-family ipv4 unicast
    aggregate-address 10.1.0.0/16 summary-only
```

The data center will host many applications for different groups of users. Some users may be part of the intranet, and others on the extranet. Maybe the requirement is that extranet users cannot access specific applications; therefore, at the border leaf, the prefixes for the restricted applications are filtered toward the edge router.

The case of a multi-tier application with web, application, and database tiers. The users only need to interface with the web front end. The web is proxy by a load balancer, usually with a VIP address. The load balancer VIP address can only be advertised out the border leaf to external network users. A common requirement is to filter all host routes and only advertise the internal subnet routes. To filter host routes, a filter can be created as shown below:

```
#Prefix List to Filter host routes for v4
ip prefix-list host-route seq 5 permit 0.0.0.0/0 eq 32
```

The prefix list above matches all host routes and is then used in a route map, as shown below.

```
route-map denyhost-filter deny 10
match ip address prefix-list host-route
route-map denyhost-filter permit 20
```

The route map to filter host routes is then applied outbound to the BGP neighbor.

```
router bgp 65002
vrf Tenant-1
 neighbor 10.10.1.2
  remote-as 65003
  address-family ipv4 unicast
  route-map denyhost-filter out
```

The border leaf can be a single exit point to the external network, removing the requirement to advertise all external networks inside the data center fabric. The border leaf can act as the default gateway device for the data center fabric. The border leaf injects a default route into the BGP EVPN table and advertises to the spine to distribute to all the VTEPs inside the VXLAN site. A default route must first exist in the routing table to inject the default route into BGP. Suppose a default route is not learned via a dynamic routing protocol from the edge router. In that case, a static default route can be configured for each tenant VRF extended to the external network as shown below.

```
vrf context Tenant-1
 ip route 0.0.0.0/0 10.10.1.2 //next hop IP on the edge router.
```

A precaution that must be taken to avoid routing loops is to filter the default route sourced from the border leaf from being advertised out to the edge router. A prefix list filter is necessary to match the default route and applied outbound to the edge router to prevent the default route intended for the data center from leaking to the external network.

```
#Prefix List to Filter default routes for v4
ip prefix-list default-route seq 5 permit 0.0.0.0/0
```

Apply the prefix list matching default route to route map as shown below.

```
route-map denydefaultroute-filter deny 10
```

```
match ip address prefix-list default-route
route-map denydefaultroute-filter permit 20
```

Under the BGP process, inside the tenant VRF advertise the default route and apply the route map, denying the default route outbound to the edge router.

```
router bgp 65002
 vrf Tenant-1
   address-family ipv4 unicast
   network 0.0.0.0/0
   neighbor 10.10.1.2
     remote-as 65003
     address-family ipv4 unicast
      route-map denydefaultroute-filter out
```

The following topology is used for the configuration example, which puts all the above concepts into one case study.

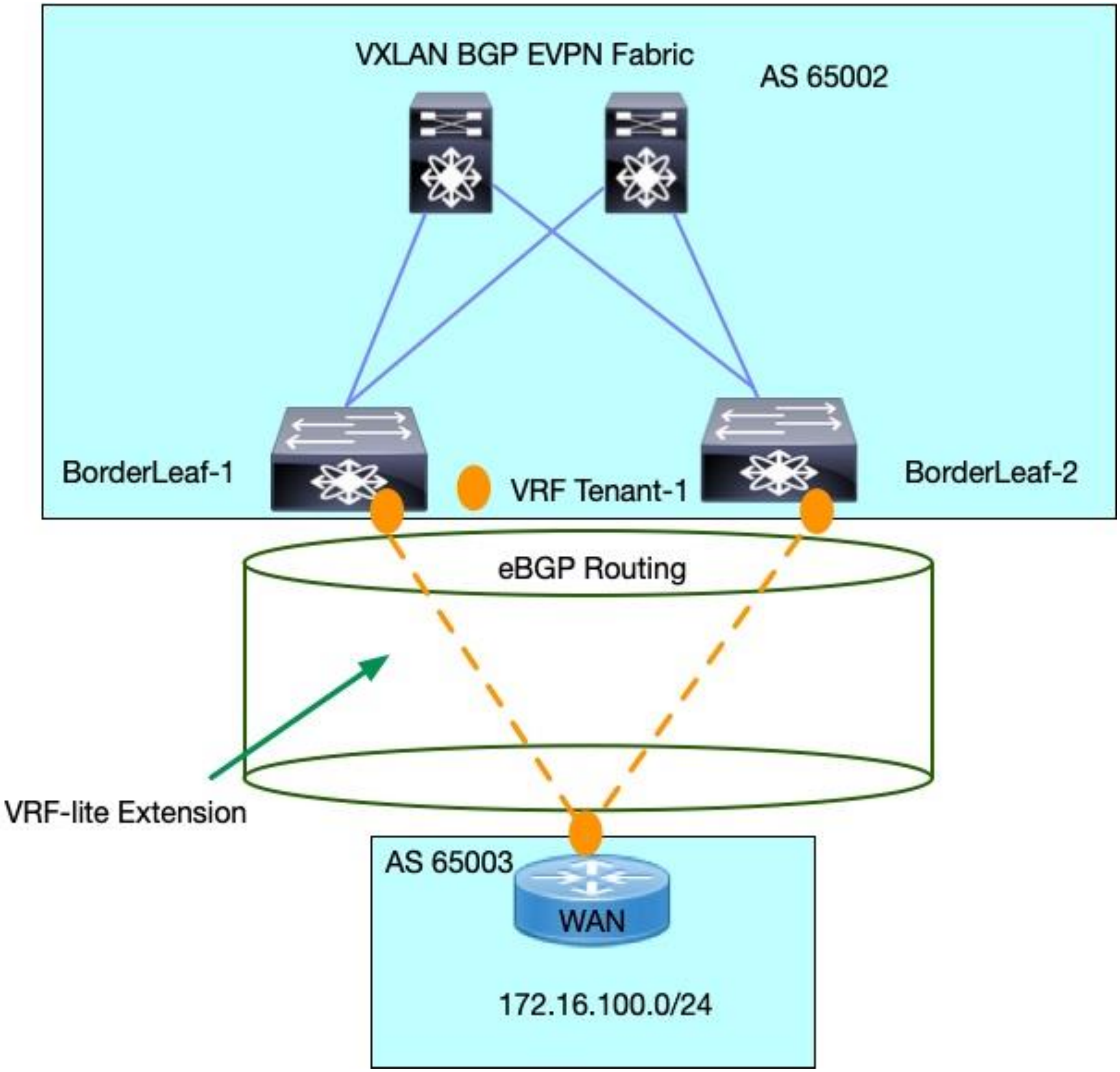


Figure 22. Border Leaf to external network connectivity

Border Leaf

Assume the requirement is to filter the host and default routes out to the external network. The steps to configure the border leaf external network connectivity are as follows:

Step 1. Configure the default route, ip prefix list and route map to filter host and default route.

```
ip prefix-list host-route seq 5 permit 0.0.0.0/0 eq 32
ip prefix-list default-route seq 5 permit 0.0.0.0/0
```

```
route-map extcon-rmap-filter deny 10
match ip address prefix-list default-route
route-map extcon-rmap-filter deny 20
match ip address prefix-list host-route
route-map extcon-rmap-filter permit 1000

vrf context Tenant-1
ip route 0.0.0.0/0 10.10.1.2
```

Step 2. Configure the routed dot1q sub-interfaces on the border leaf links connecting to the WAN router.

```
//BorderLeaf-1
interface eth1/4
mtu 9216
no switchport
no shutdown
interface eth1/4.10
encapsulation dot1q 10
vrf member Tenant-1
ip address 10.10.1.1/30

//BorderLeaf-2

interface eth1/4
mtu 9216
no switchport
no shutdown
interface eth1/4.10
encapsulation dot1q 10
vrf member Tenant-1
ip address 10.10.2.1/30
```

Step 3. Enable BGP peering on the border leaf to the WAN router.

```
//BorderLeaf-1
router bgp 65002
    address-family l2vpn evpn
    vrf Tenant-1
    address-family ipv4 unicast
    advertise l2vpn evpn
    maximum-paths 2
    network 0.0.0.0/0
  neighbor 10.10.1.2 remote-as 65003
```

```
    address-family ipv4 unicast
  route-map extcon-rmap-filter out

//BorderLeaf-2

router bgp 65002
 address-family l2vpn evpn
 vrf Tenant-1
 address-family ipv4 unicast
    advertise l2vpn evpn
    maximum-paths 2
    network 0.0.0.0/0
  neighbor 10.10.2.2 remote-as 65003
    address-family ipv4 unicast
  route-map extcon-rmap-filter out
```

Finally, on the WAN router, the routed interfaces to the border leaf are configured, as shown below.

```
interface eth1/1
  description linktoBL1
  mtu 9216
  no switchport
  no shutdown

interface eth1/1.10
  encapsulation dot1q 10
  vrf member Tenant-1
  ip address 10.10.1.2/30
  no shutdown

interface eth1/2
  description linktoBL2
  mtu 9216
  no switchport
  no shutdown

interface eth1/2.10
  encapsulation dot1q 10
  vrf member Tenant-1
  ip address 10.10.2.2/30
  no shutdown
```

Then enable the BGP peering on the WAN router to the border leaf.

```
feature bgp
```

```
router bgp 65003
  vrf member Tenant-1
  address-family ipv4 unicast
  network 172.16.100.0/24
  maximum-paths 2
  neighbor 10.10.1.1 remote-as 65002
   address-family ipv4 unicast
  neighbor 10.10.2.1 remote-as 65002
   address-family ipv4 unicast
```

vPC Border Leaf

The border leaf nodes can be multi-homed to a WAN edge router in a vPC domain. The links from the vPC border leaf devices will not be layer 2 port channels, but point-to-point routed interfaces. The edge router advertises networks to the border leaf using a dynamic routing protocol such as OSPF or BGP. The preferred protocol is eBGP, which removes the requirement to redistribute another protocol into the BGP EVPN control plane in the VXLAN fabric and avoids routing loops that may happen when using a different peering protocol.

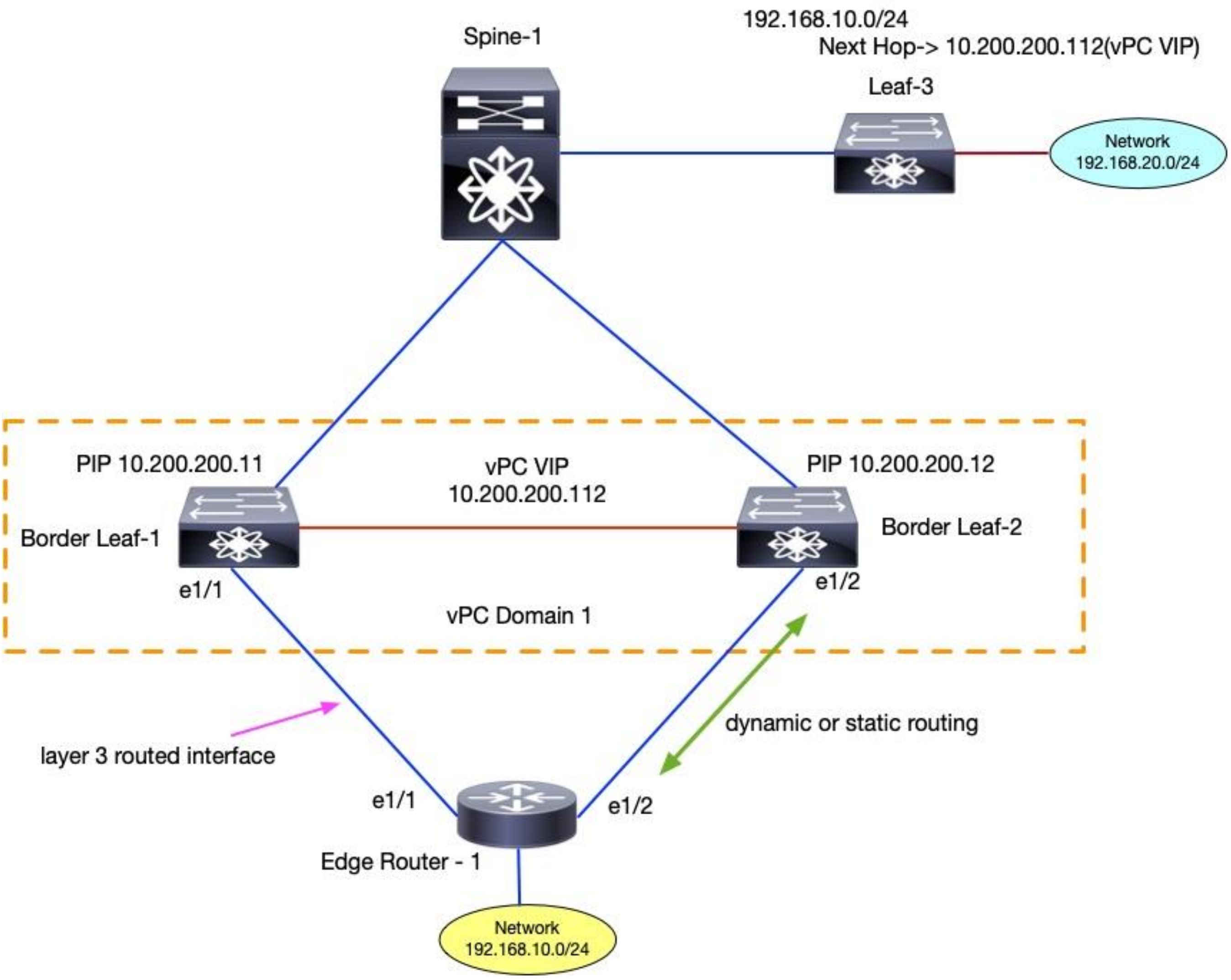


Figure 23. vPC Border Leaf

In the above diagram, the external network 192.168.10.0/24 is reachable via both border leaf nodes. On leaf 3, the next hop to reach the external network is the vPC VIP of the border leaf nodes. The packet sent from Leaf 3 to the external network will hash to either of the border leaves through the ECMP path. The border leaves can then forward the packet to the edge router to reach the external network.

The vPC member switches synchronize forwarding state information, such as ARP, MAC, etc., for host routes learned on vPC interfaces. The routes learned from the edge router are Type 5 prefix routes, which are not synchronized between vPC member switches. This implies that the vPC member switches are unaware it's vPC peer switch is attached to the same external network. The consequence is if any of the edge router interfaces were to fail on the border leaf, there would be no path to redirect the traffic to the vPC member switch, and the packet would be dropped.

To illustrate this scenario, leaf 3 sends the packet to the external network using vPC VIP as the next hop. The packet arrives at border leaf 1 as it has advertised the external network with vPC VIP as the next hop. if the interface eth1/1 between border leaf 1 and edge router goes down due to a link failure, border leaf 1 will drop the packet as it has no alternative path to route the packet. The reason of the packet drop is that leaf 3 views both border leaves with a single IP address identity, the vPC VIP. The transient spine node underlay routing protocols install border leaf 1 and border leaf 2 as the next hops to reach the vPC VIP.

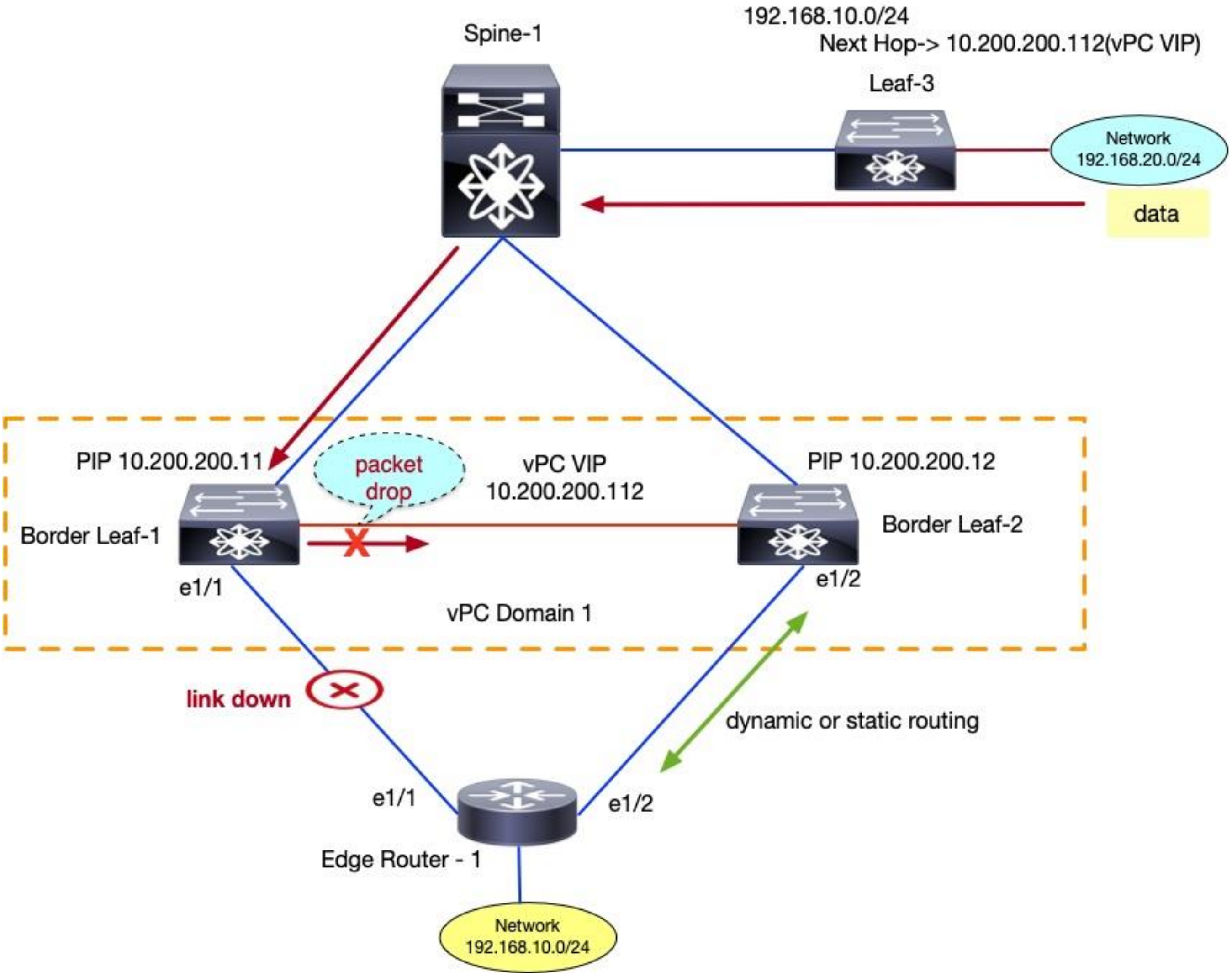


Figure 24. vPC Border Leaf Edge Router Multihoming Packet Drop

To avoid tenant traffic blackholing, the recommendation is to configure the vPC border leaf to use its primary IP address as the next hop for all routes learned from the WAN edge router. The command to use the PIP address for EVPN Type 5 prefix routes is shown below.

```
router bgp 65536
   address-family l2vpn evpn
   advertise-pip
interface nve 1
   advertise virtual-rmac
```

The next hop for EVPN Type 2 host routes vPC and orphan attached devices are advertised with vPC VIP and VMAC addresses. The VMAC is a shared MAC address for the vPC domain derived from 02.00 + 4 Bytes of VIP converted in HEX. The next hop for EVPN Type 5 prefix routes vPC and orphan attached networks are advertised with PIP and RMAC addresses. The RMAC address is unique to each VTEP device.

The border leaf device can also be configured as vPC Fabric Peering VTEPs and peer with an edge router. All prefixes learned from the edge router are injected inside the VXLAN BGP EVPN fabric as EVPN type 5 routes. The advertise PIP and RMAC configurations are mandatory in vPC fabric peering if EVPN Type 5 routes are learned from external routers. vPC and Orphan Type 5 routes will be advertised with (PIP,RMAC).

Conclusion

VXLAN BGP EVPN fabric is a networking architecture with underlay and overlay layers. Each layer must be designed with the function of each layer in mind. The underlay and overlay provide routing and switching for unicast and multicast traffic. Designing any network requires asking “why?” and understanding the tradeoff of taking one approach. VXLAN BGP EVPN fabric can be created using different architectures such as 3-stage or 5-stage CLOS. The network devices can have functions such as border leaf, border spine, border gateway, border gateway spine, service, or server leaf. The placement of these devices in the network allows flexible architectures to be designed for various data center applications and scale requirements. The Nexus 9000 NXOS VXLAN BGP EVPN fabric provides an open standard solution supporting large-scale data centers.