Guide

VXLAN Network with MP-BGP EVPN Control Plane

Design Guide

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Contents

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Introduction

Virtual Extensible LAN (VXLAN) is an overlay technology for network virtualization. It provides Layer-2 extension over a shared Layer-3 underlay infrastructure network by using MAC address in IP User Datagram Protocol (MAC in IP/UDP) tunneling encapsulation. The purpose of obtaining Layer-2 extension in the overlay network is to overcome the limitations of physical server racks and geographical location boundaries and achieve flexibility for workload placement within a data center or between different data centers.

The initial IETF VXLAN standards (RFC 7348) defined a multicast-based flood-and-learn VXLAN without a control plane. It relies on data-driven flood-and-learn behavior for remote VXLAN tunnel endpoint (VTEP) peer discovery and remote end-host learning. The overlay broadcast, unknown unicast, and multicast traffic is encapsulated into multicast VXLAN packets and transported to remote VTEP switches through the underlay multicast forwarding. Flooding in such a deployment can present a challenge for the scalability of the solution. The requirement to enable multicast capabilities in the underlay network also presents a challenge because some organizations do not want to enable multicast in their data centers or WAN networks.

To overcome the limitations of the flood-and-learn VXLAN as defined in RFC 7348, organizations can use Multiprotocol Border Gateway Protocol Ethernet Virtual Private Network (MP-BGP EVPN) as the control plane for VXLAN. MP-BGP EVPN has been defined by IETF as the standards-based control plane for VXLAN overlays. The MP-BGP EVPN control plane provides protocol-based VTEP peer discovery and end-host reachability information distribution that allows more scalable VXLAN overlay network designs suitable for private and public clouds. The MP-BGP EVPN control plane introduces a set of features that reduces or eliminates traffic flooding in the overlay network and enables optimal forwarding for both west-east and south-north traffic.

This document discusses the functions and configuration of MP-BGP EVPN and describes typical VXLAN overlay network designs using MP-BGP EVPN.

This document does not discuss the fundamentals of VXLAN, VXLAN in multicast-based flood-and-learn mode, or related network design options. For more information about VXLAN and VXLAN with multicast-based flood-and-learn, please refer to the following documents:

- VXLAN Overview: Cisco Nexus® 9000 Series Switches: http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html.
- VXLAN Design with Cisco Nexus 9300 Platform Switches: http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-732453.html.

This document assumes prior knowledge about BGP, MP-BGP, and BGP and Multiprotocol Label Switching (BGP/MPLS) IP VPN. For more information, refer to the following IETF RFC documents:

- RFC 4271 - Border Gateway Protocol 4 (BGP-4): https://tools.ietf.org/html/rfc4271
- RFC 4760 - Multiprotocol Extensions for BGP-4: https://tools.ietf.org/html/rfc4760
- RFC 4364 - BGP/MPLS IP VPNs: https://tools.ietf.org/html/rfc4364#page-15

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

MP-BGP EVPN Control Plane: Overview

MP-BGP EVPN is a control protocol for VXLAN based on industry standards. Prior to EVPN, VXLAN overlay networks operated in the flood-and-learn mode. In this mode, end-host information learning and VTEP discovery are both data plane driven, with no control protocol to distribute end-host reachability information among VTEPs. MP-BGP EVPN changes this model. It introduces control-plane learning for end hosts behind remote VTEPs. It provides control-plane and data-plane separation and a unified control plane for both Layer-2 and Layer-3 forwarding in a VXLAN overlay network.

The MP-BGP EVPN control plane offers the following main benefits:

- The MP-BGP EVPN protocol is based on industry standards, allowing multivendor interoperability.
- It enables control-plane learning of end-host Layer-2 and Layer-3 reachability information, enabling organizations to build more robust and scalable VXLAN overlay networks.
- It uses the decade-old MP-BGP VPN technology to support scalable multitenant VXLAN overlay networks.
- The EVPN address family carries both Layer-2 and Layer-3 reachability information, thus providing integrated bridging and routing in VXLAN overlay networks.
- It minimizes network flooding through protocol-based host MAC/IP route distribution and Address Resolution Protocol (ARP) suppression on the local VTEPs.
- It provides optimal forwarding for east-west and north-south traffic and supports workload mobility with the distributed anycast function.
- It provides VTEP peer discovery and authentication, mitigating the risk of rogue VTEPs in the VXLAN overlay network.
- It provides mechanisms for building active-active multihoming at Layer-2.

Software and Hardware Support for the MP-BGP EVPN Control Plane

Depending on the role a device plays in a MP-BGP EVPN VXLAN network, it may need to support only the control-plane functions or both the control-plane and data-plane functions of the VXLAN network with the MP-BGP EVPN control plane.

IP Transport Devices Running MP-BGP EVPN

IP transport devices provide IP routing in the underlay network. By running the MP-BGP EVPN protocol, they become part of the VXLAN control plane and distribute the MP-BGP EVPN routes among their MP-BGP EVPN peers. Devices might be MP-iBGP EVPN peers or route reflectors, or MP External BGP (MP-eBGP) EVPN peers. Their OS software needs to support MP-BGP EVPN so that it can understand the MP-BGP EVPN updates and distribute them to other MP-BGP EVPN peers using the standards-defined constructs. For data forwarding, IP transport devices perform IP routing based only on the outer IP address of a VXLAN encapsulated packet. They don’t need to support the VXLAN data encapsulation and decapsulation functions.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

VTEPs Running MP-BGP EVPN

VTEPs running MP-BGP EVPN need to support both the control-plane and data-plane functions. In the control plane, they initiate MP-BGP EVPN routes to advertise their local hosts. They receive MP-BGP EVPN updates from their peers and install the EVPN routes in their forwarding tables. For data forwarding, they encapsulate user traffic in VXLAN and send it over the IP underlay network. In the reverse direction, they receive VXLAN encapsulated traffic from other VTEPs, decapsulate it, and forward the traffic with native Ethernet encapsulation toward the host.

The correct switch platforms need to be selected for the different network roles. For IP transport devices, the software needs to support the MP-EVPN control plane, but the hardware doesn't need to support VXLAN data-plane functions. For VTEP, the switch needs to support both the control-plane and data-plane functions.

Inter-VXLAN Routing

The MP-BGP EVPN control plane provides integrated routing and bridging by distributing both the Layer-2 and Layer-3 reachability information for end hosts on VXLAN overlay networks. Communication between hosts in different subnets requires inter-VXLAN routing. BGP EVPN enables this communication by distributing Layer-3 reachability information in the form of either a host IP address route or an IP address prefix. In the data plane, the VTEP needs to support IP address route lookup and perform VXLAN encapsulation based on the lookup result. This capability is referred to as the VXLAN routing function. Not all switch hardware platforms support VXLAN routing, hence affecting the choice of hardware platform.

MP-BGP EVPN VXLAN Support on Cisco Nexus 9000 Series Switches

The MP-BGP EVPN control plane for VXLAN was introduced into Cisco® NX-OS Software Release 7.0(3)I1(1) for Cisco Nexus 9000 Series Switches. The software functions will be implemented in the Cisco NX-OS software trains for other Cisco Nexus switch platforms, such as the Cisco Nexus 7000 Series Switches, as well.

In Cisco NX-OS 7.0(3)I1(1), the Cisco Nexus 9300 platform switches support both the MP-BGP EVPN control-plane functions and the VTEP data-plane functions. The Cisco Nexus 9500 platform switches support the MP-BGP EVPN control-plane functions. The VTEP data-plane functions will be added to the Cisco Nexus 9500 platform switches in a maintenance release of Cisco NX-OS 7.0(3)I1(1). The Cisco Nexus 9300 and 9500 platforms both support inter-VXLAN routing in hardware.

Although many of the MP-BGP EVPN functions and design discussions in this document are platform independent, because the Cisco Nexus 9000 Series is the first switch platform that supports this protocol, the examples are based on the Cisco Nexus 9000 Series.

Multitenancy in MP-BGP EVPN

As an extension to the existing MP-BGP, MP-BGP EVPN inherits the support for multitenancy with VPN using the virtual routing and forwarding (VRF) construct. In MP-BGP EVPN, multiple tenants can co-exist and share a common IP transport network while having their own separate VPNs in the VXLAN overlay network.

In the EVPN VXLAN overlay network, VXLAN network identifiers (VNIs) define the Layer-2 domains and enforce Layer-2 segmentation by not allowing Layer-2 traffic to traverse VNI boundaries. Similarly, Layer-3 segmentation among VXLAN tenants is achieved by applying Layer-3 VRF technology and enforcing routing isolation between tenants by using a separate Layer-3 VNI mapped to each VRF instance. Each tenant has its own VRF routing instance. IP subnets of the VNIs for a given tenant are in the same Layer-3 VRF instance that separates the Layer-3 routing domain from the other tenants.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

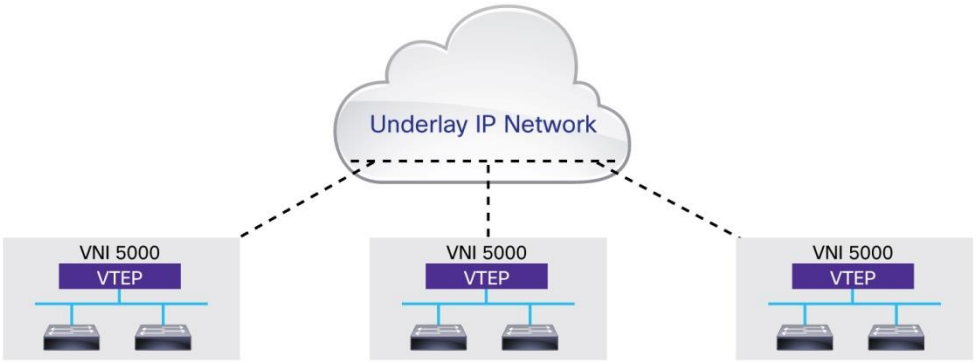
Built-in multitenancy support is an advantage of MP-BGP EVPN VXLAN compared to multicast-based flood-and-learn VXLAN and other Layer-2 extension technologies without multitenancy capabilities. It makes VXLAN technology more suitable for cloud networks, which are deployed using the multitenant model.

MP-BGP EVPN NLRI and L2VPN EVPN Address Family

Like other network routing control protocols, MP-BGP EVPN is designed to distribute network layer reachability information (NLRI) for the network. A unique feature of EVPN NLRI is that it includes both the Layer-2 and Layer-3 reachability information for end hosts that reside in the EVPN VXLAN overlay network. In other words, it advertises both MAC and IP addresses of EVPN VXLAN end hosts. This capability forms the basis for VXLAN integrated routing and bridging support.

Layer-2 MAC addresses need to be distributed because VXLAN is a Layer-2 extension technology. Unlike a traditional VLAN, which is confined in a specific location in a network and remains within the Layer-2 and Layer-3 boundary, a VNI is a virtual Layer-2 segment in the overlay network. However, from the underlay network point of view, it can span multiple noncontiguous sites, reaching beyond the Layer-2 and Layer-3 boundary of the underlay infrastructure (Figure 1). Traffic between end hosts in the same VNI needs to be bridged in the overlay network, which means that VTEP devices in a given VNI need to know about other MAC addresses of end hosts in this VNI. Distribution of MAC addresses through BGP EVPN allows unknown unicast flooding in the VXLAN to be reduced or eliminated.

Figure 1. VNI across an Underlay IP Network



Layer-3 host IP addresses are advertised through MP-BGP EVPN so that inter-VXLAN traffic can be routed to the destination end host through an optimal path. For inter-VXLAN traffic that needs to be routed to the destination end host, host-based IP routing can provide the optimal forwarding path to the exact location of the destination host.

MP-BGP EVPN can also advertise the IP subnet prefix routes of VNIs. The prefix routes can be used to route traffic to the destination hosts when the host IP routes are missing: for instance, when the host IP routes have not yet been learned by the VTEPs through MP-BGP. VTEP can also advertise the prefix routes to outside the VXLAN network if the subnets need to be routable and made known outside the VXLAN network.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

EVPN NLRI is carried in BGP using the BGP multiprotocol extension with a new address family called Layer-2 VPN (L2VPN) EVPN. Similar to the VPNv4 address-family in the BGP MPLS-based IP VPN (RFC 4364), the L2VPN EVPN address-family for EVPN uses route distinguishers (RDs) to maintain uniqueness among identical routes in different VRF instances, and uses route targets (RTs) to define the policies that determine how routes are advertised and shared by different VRF instances.

A route distinguisher is an 8-bit octet number used to distinguish one set of routes (one VRF instance) from another. It is a unique number prepended to each route so that if the same route is used in several different VRF instances, BGP can treat them as distinct routes. The route distinguisher is transmitted along with the route through MP-BGP when EVPN routes are exchanged with MP-BGP peers.

Route targets can be applied to a VRF instance to control the import and export of routes between this instance and other VRF instances. The route-target attributes for a route are distributed in the form of a BGP extended community attribute, so the BGP configuration on the devices that run MP-BGP EVPN must be enabled to generate or process extended community attributes.

In the Cisco NX-OS implementation, the BGP route distinguisher and route target can be generated automatically for ease of configuration. The BGP route distinguisher can be derived automatically from the VNI and BGP router ID of the VTEP switch, and the BGP route target can be generated automatically as the BGP AS: VNI. Alternatively, you also can manually configure the BGP route distinguisher and route target. If all the MP-BGP EVPN VTEPs in a network are Cisco Nexus switch platforms, the recommended approach is to use autogenerated route-distinguisher and route-target values. If multiple vendors' VTEP devices are interoperating, the recommended approach is to manually configure the values to avoid problems caused by the differences in vendors' implementations. For eBGP deployment scenarios in which VTEPs are in different BGP domains, the BGP route targets must be manually assigned.

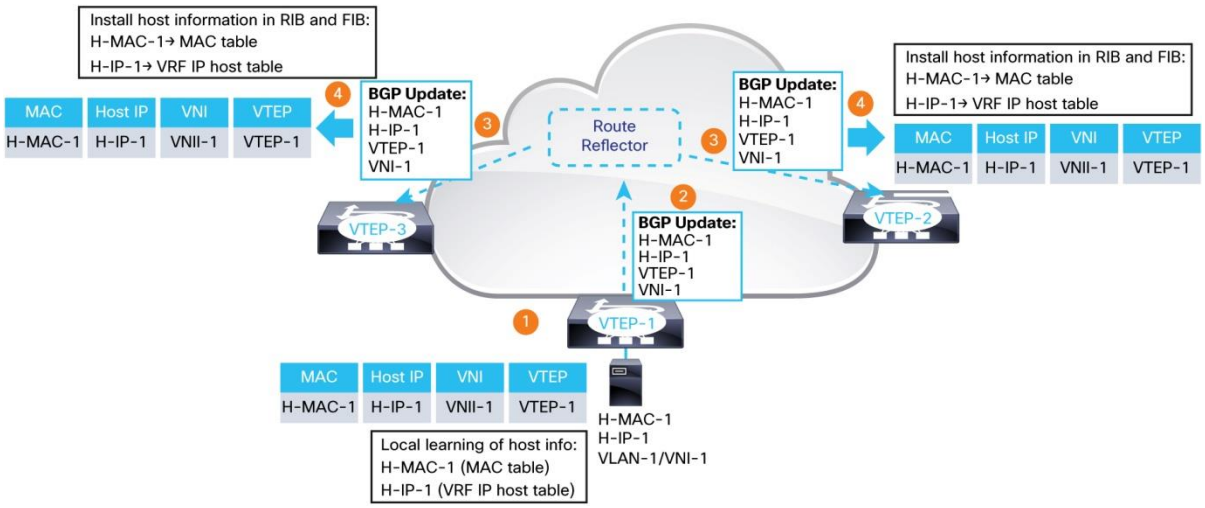
Integrated Routing and Bridging with the MP-BGP EVPN Control Plane

The MP-BGP EVPN control plane provides integrated routing and bridging by distributing both Layer-2 and Layer-3 reachability information for the end host residing in the VXLAN overlay networks. Each VTEP performs local learning to obtain MAC and IP address information from its locally attached hosts and then distributes this information through the MP-BGP EVPN control plane. Hosts attached to remote VTEPs are learned remotely through the MP-BGP control plane. This approach reduces network flooding for end-host learning and provides better control over end-host reachability information distribution.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 2 shows an example of end-host NLRI learning and distribution in an MP-iBGP EVPN using route reflectors.

Figure 2. MP-BGP EVPN Host NLRI Learning and Distribution



Local-Host Learning

A VTEP in MP-BGP EVPN learns the MAC addresses and IP addresses of locally attached end hosts through local learning. This learning can be local-data-plane based using the standard Ethernet and IP learning procedures, such as source MAC address learning from the incoming Ethernet frames and IP address learning when the hosts send Gratuitous ARP (GARP) and Reverse ARP (RARP) packets or ARP requests for the gateway IP address on the VTEP. Alternatively, the learning can be achieved by using a control plane or through management-plane integration between the VTEP and the local hosts.

EVPN Route Advertisement and Remote-Host Learning

After learning the local-host MAC and IP addresses, a VTEP advertises the host information in the MP-BGP EVPN control plane so that this information can be distributed to other VTEPs. This approach enables EVPN VTEPs to learn the remote end hosts in the MP-BGP EVPN control plane.

The EVPN routes are advertised through the L2VPN EVPN address-family. The BGP L2VPN EVPN routes include the following information:

- RD: Route distinguisher
- MAC address length: 6 bytes
- MAC address: Host MAC address
- IP address length: 32 or 128
- IP address: Host IP address (IPv4 or IPv6)
- L2 VNI: VNI of the bridge domain to which the end host belongs
- L3 VNI: VNI associated with the tenant VRF routing instance

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

MP-BGP EVPN uses the BGP extended community attribute to transmit the exported route-targets in an EVPN route. When an EVPN VTEP receives an EVPN route, it compares the route-target attributes in the received route to its locally configured route-target import policy to decide whether to import or ignore the route. This approach uses the decade-old MP-BGP VPN technology (RFC 4364) and provides scalable multitenancy in which a node that does not have a VRF locally does not import the corresponding routes. VPN scaling can be further enhanced by the use of BGP constructs such as route-target-constrained route distribution (RFC 4684).

When a VTEP switch originates MP-BGP EVPN routes for its locally learned end hosts, it uses its own VTEP address as the BGP next-hop. This BGP next-hop must remain unchanged through the route distribution across the network because the remote VTEP must learn the originating VTEP address as the next-hop for VXLAN encapsulation when forwarding packets for the overlay network.

The underlay network provides IP reachability for all the VTEP addresses that are used to route the encapsulated VXLAN packets toward the egress VTEP through the underlay network. The network devices in the underlay network need to maintain routing information only for the VTEP addresses. They don't need to learn the EVPN routes. This approach simplifies the underlay network operation and increases its stability and scalability.

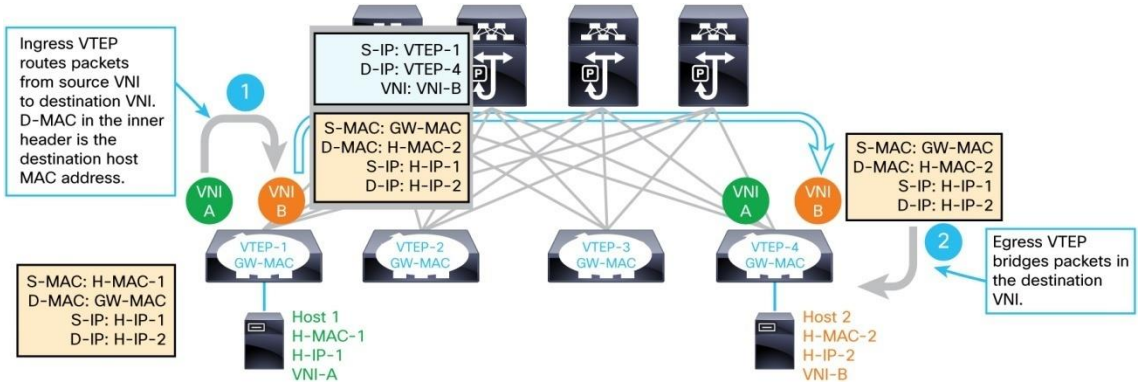
Symmetric and Asymmetric Integrated Routing and Bridging

The IETF EVPN drafts define two integrated routing and bridging (IRB) semantics: asymmetric IRB and symmetric IRB. Cisco NX-OS for Cisco Nexus switch platforms implements symmetric IRB for its scalability advantages and simplified Layer-2 and Layer-3 multitenancy support.

Asymmetric IRB

With asymmetric IRB, the ingress VTEP performs both Layer-2 bridging and Layer-3 routing lookup, whereas the egress VTEP performs only Layer-2 bridging lookup. As shown in Figure 3, with asymmetric IRB, when a packet travels between two VNIs, the ingress VTEP routes the packet from the source VNI to the destination VNI. The egress VTEP bridges the packet to the destination point within the destination VNI.

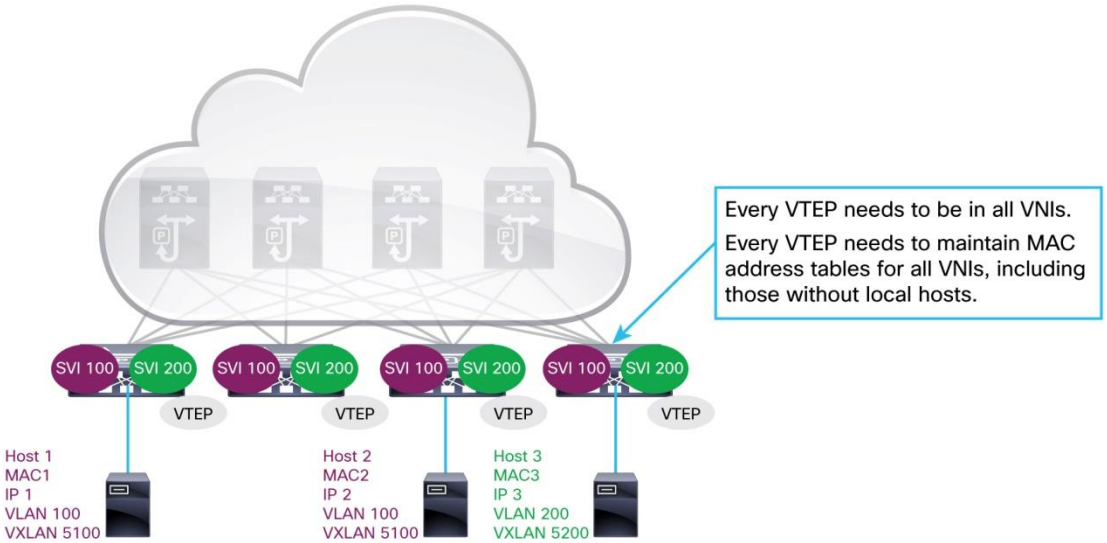
Figure 3. VXLAN Routing with Asymmetric IRB



Asymmetric IRB requires the ingress VTEP to be configured with both the source and destination VNIs for both Layer-2 and Layer-3 forwarding. Essentially, this requires each VTEP to be configured with all VNIs in the VXLAN network and to learn ARP entries and MAC addresses for all the end hosts attached to those VNIs (Figure 4). This behavior can cause scalability problems as the density of end hosts and/or the number of VXLAN VNIs in the overlay network increase.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 4. VTEP VNI Membership in Asymmetric IRB



Symmetric IRB

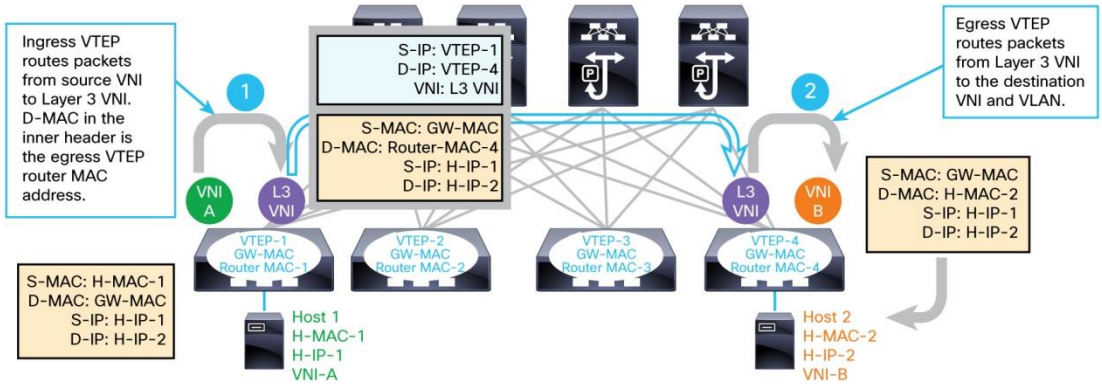
With symmetric IRB, both the ingress and egress VTEPs perform Layer-2 and Layer-3 lookups. Symmetric IRB introduces some new logical constructs:

- **Layer-3 VNI:** Each tenant VRF instance is mapped to a unique Layer-3 VNI in the network. This mapping needs to be consistent on all the VTEPs in network. All inter-VXLAN routed traffic is encapsulated with the Layer-3 VNI in the VXLAN header and provides the VRF context for the receiving VTEP. The receiving VTEP uses this VNI to determine the VRF context in which the inner IP packet needs to be forwarded. This VNI also provides the basis for enforcing Layer-3 segmentation in the data plane.
- **VTEP router MAC address:** Each VTEP has a unique system MAC address that other VTEPs can use for inter-VNI routing. This MAC address is referred to here as the router MAC address. The router MAC address is used as the inner destination MAC address for the routed VXLAN packet.

As shown in Figure 5, when a packet is sent from VNI A to VNI B, the ingress VTEP routes the packet to the Layer-3 VNI. It rewrites the inner destination MAC address to the egress VTEP's router MAC address and encodes the Layer-3 VNI in the VXLAN header. After the egress VTEP receives the encapsulated VXLAN packet, it first decapsulates the packet by removing the VXLAN header. Then it looks at the inner packet header. Because the destination MAC address in the inner packet header is its own MAC address, it performs a Layer-3 routing lookup. The Layer-3 VNI in the VXLAN header provides the VRF context in which this routing lookup is performed.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

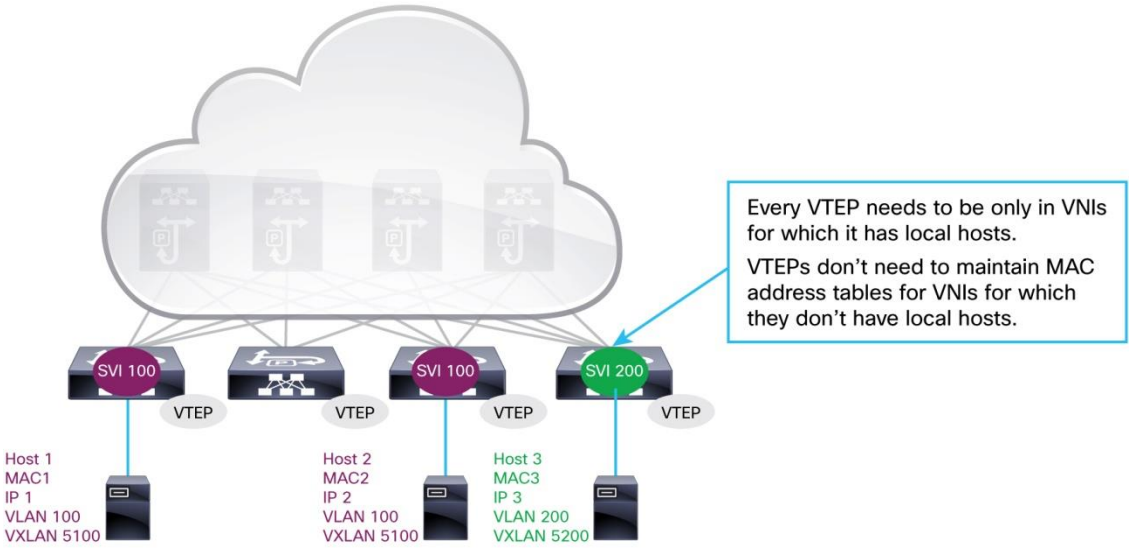
Figure 5. VXLAN Routing with Symmetric IRB



Advantages of Symmetric IRB

With symmetric IRB, the ingress VTEP doesn't need to know the destination VNI for inter-VNI routing. Therefore, VTEPs don't need to learn and maintain MAC address information for the remote hosts attached to egress VNIs for which it doesn't have local hosts (Figure 6). This approach results in better utilization of the MAC address table and ARP adjacencies on a VTEP. For example, in Figure 6 all host MAC address and ARP adjacencies in VNI-B do not need to be present on VTEP-1. As a result, the routing and bridging is more scalable than with asymmetric IRB. Cisco NX-OS implements symmetric IRB to achieve optimal learning and scaling.

Figure 6. VTEP VNI Membership with Symmetric IRB

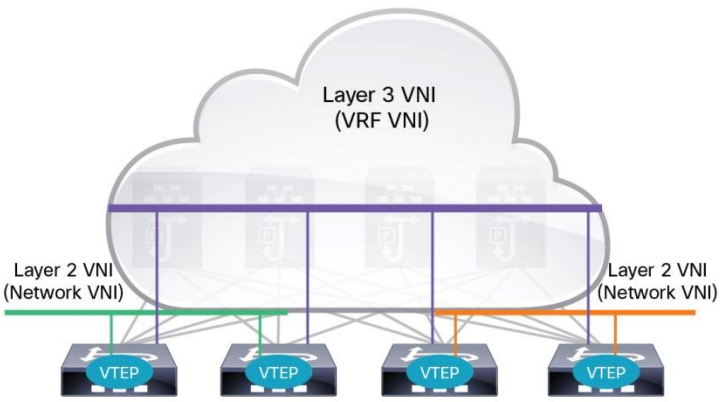


© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

VNIs for Bridge Domains and IP VRF Instances

An EVPN VXLAN tenant can have multiple Layer-2 networks, each with a corresponding VNI. These Layer-2 networks are bridge domains in the overlay network. The VNIs which are associated with them are often referred to as Layer-2 (L2) VNIs. Each tenant also needs a Layer-3 (L3) VNI for symmetric IRB if inter-VXLAN routing is needed. Although a VTEP can have all or a subset of the Layer-2 VNIs in an VXLAN EVPN, it must have the Layer-3 VNI for inter-VXLAN routing. All VTEPs in an EVPN must have the same Layer-3 VNI (Figure 7).

Figure 7. VNIs for Bridge Domain and IP VRF Instances



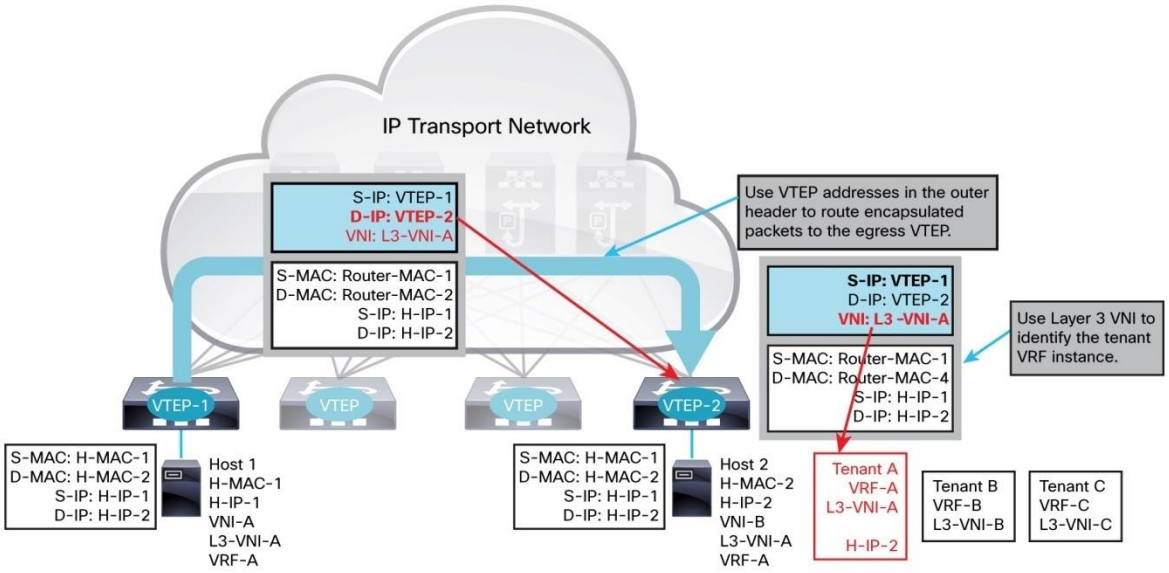

```
vlan 200
  vn-segment 20000
vlan 201
  vn-segment 20100

vlan 3900
  name l3-vni-vlan-for-tenant-1
  vn-segment 39000

vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn

interface Vlan200
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.0.0.1/24
  fabric forwarding mode anycast-gateway

interface Vlan201
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.1.0.1/24
  fabric forwarding mode anycast-gateway

interface Vlan3900
  description l3-vni-for-tenant-1-routing
  no shutdown
  vrf member evpn-tenant-1
  ip forward
```

When an EVPN VTEP performs forwarding lookup and VXLAN encapsulation for the packets it receives from its local end hosts, it uses either a Layer-2 VNI or the Layer-3 VNI in the VXLAN header, depending on whether the packets need to be bridged or routed. If the destination MAC address in the original packet header does not belong to the local VTEP, the local VTEP performs a Layer-2 lookup and bridges the packet to the destination end host that is located in the same Layer-2 VNI as the source host. The local VTEP embeds this Layer-2 VNI in the VXLAN header. In this case, both the source and destination hosts are in the same Layer-2 broadcast domain. If the destination MAC address belongs to the local VTEP switch - that is, if the local VTEP is the IP gateway for the source host, and the source and destination hosts are in different IP subnets - the packet will be routed by the local VTEP. In this case, it performs Layer-3 routing lookup. It then encapsulates the packets with the Layer-3 VNI in the VXLAN header and rewrites the inner destination MAC address to the remote VTEP's router MAC address. Upon receipt of the encapsulated VXLAN packet, the remote VTEP performs another routing lookup based on the inner IP header because the inner destination MAC address in the received packet belongs to the remote VTEP itself.

The destination VTEP address in the outer IP header of a VXLAN packet identifies the location of the destination host in the underlay network. VXLAN packets are routed toward the egress VTEP through the underlay network based on the outer destination IP address. After the packet arrives at the egress VTEP, the VNI in the VXLAN header is examined to determine the VLAN in which the packet should be bridged or the tenant VRF instance to which it should be routed. In the latter case, the VXLAN header is encoded with a Layer-3 VNI. A Layer-3 VNI is associated with a tenant VRF routing instance, so the egress VTEP can directly map the routed VXLAN packets to the appropriate tenant routing instance. Figure 8 shows this forwarding concept in symmetric IRB. This approach makes multitenancy easier to support for both Layer-2 and Layer-3 segmentation.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 8. VXLAN Packet Forwarding with Symmetric IRB Routing



VTEP Peer Discovery and Authentication in MP-BGP EVPN

Prior to MP-BGP EVPN, VXLAN didn't have a control-protocol-based VTEP peer-discovery mechanism or a method for authenticating VTEP peers. These limitations present major security risks in real-world VXLAN deployments because they allow easy insertion of a rogue VTEP into a VNI segment to send or receive VXLAN traffic.

With the MP-BGP EVPN control plane, a VTEP device first needs to establish BGP neighbor adjacency with other VTEPs or with Internal BGP (iBGP) route reflectors. In addition to the BGP updates for end-host NLRI, VTEPs exchange the following information about themselves through BGP:

- Layer-3 VNI
- VTEP address
- Router MAC address

As soon as a VTEP receives BGP EVPN route updates from a remote VTEP BGP neighbor, it adds the VTEP address from that route advertisement to the VTEP peer list. This VTEP peer list then is used as an allowed list of valid VTEP peers. VTEPs that are not on this allowed list are considered invalid or un-authorized sources. VXLAN encapsulated traffic from these invalid VTEPs will be discarded by other VTEPs.

In the data-plane forwarding, a BGP EVPN VTEP accepts VXLAN encapsulated packets only from VTEP peers that are on the allowed list. Thus, MP-BGP EVPN introduces protocol-based VTEP discovery and the capability to restrict VXLAN overlay traffic distribution to only BGP-learned VTEPs.

Along with the VTEP address that promotes VTEP peer learning, BGP EVPN routes carry VTEP router MAC addresses. Each VTEP has a router MAC address. Once a VTEP's router MAC address is distributed via MP-BGP and learned by other VTEPs, the other VTEPs use it as an attribute of the VTEP peer to encapsulate inter-VXLAN routed packets to that VTEP peer. The router MAC address is programmed as the inner destination MAC address for routed VXLAN.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

For additional security, the existing BGP Message Digest 5 (MD5) authentication can be conveniently applied to the BGP neighbor sessions so that switches can't become BGP neighbors to exchange MP-BGP EVPN routes until they successfully authenticate each other with a preconfigured MD5 Triple Data Encryption Standard (3DES) key. BGP neighbor authentication in MP-BGP EVPN is configured in the same way as previously supported in BGP. An example is shown here:

On VTEP-1

```
router bgp 100
  router-id 10.1.1.101
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.102 remote-as 100
    password 3 a667d47acc18ea6b
    update-source loopback0
    address-family ipv4 unicast
      send-community both
    address-family l2vpn evpn
      send-community both
```

- Can be configured using the command-line interface (CLI) with a clear password: **password cisco123**. The system will automatically change this password to a 3DES-encrypted password in the running configuration display.
- Both neighbors need to have the exact same password.

On VTEP-2

```
router bgp 100
  router-id 10.1.1.102
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  neighbor 10.1.1.101 remote-as 100
    password 3 a667d47acc18ea6b
    update-source loopback0
    address-family ipv4 unicast
      send-community both
    address-family l2vpn evpn
      send-community both
```

The following is a sample display of VNI peer status and information in Cisco NX-OS:

```
VTEP1-1# sh nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      10.1.1.102        Up    CP        1w3d     6412.2574.6ae7
nve1      10.1.1.134        Up    CP        1w3d      7c69.f6df.e71f

VTEP-1#
```

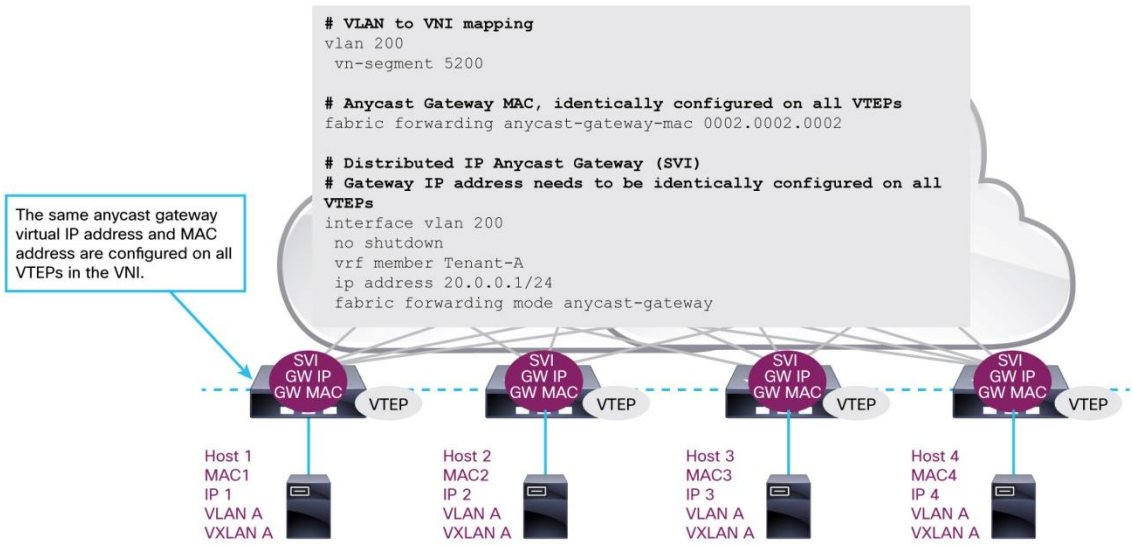
```
VTEP-1# sh nve peers peer-ip 10.1.1.102 det
Details of nve Peers:
----------------------------------------
Peer-Ip: 10.1.1.102
    NVE Interface       : nve1
    Peer State          : Up
    Peer Uptime         : 1w3d
    Router-Mac          : 6412.2574.6ae7
    Peer First VNI      : 20100
    Configured VNIs     : 20000,20100,21000,21100,39000,39010
    Provision State     : add-complete
    Route-Update        : Yes
    Peer Flags          : DisableLearn
    Learnt CP VNIs      : 20000,20100
----------------------------------------
VTEP-1#
```

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Distributed Anycast Gateway in MP-BGP EVPN

In MP-BGP EVPN, any VTEP in a VNI can be the distributed anycast gateway for end hosts in its IP subnet by supporting the same virtual gateway IP address and the virtual gateway MAC address (Figure 9). With the anycast gateway function in EVPN, end hosts in a VNI always can use their local VTEPs for this VNI as their default gateway to send traffic to outside of their IP subnet. This capability enables optimal forwarding for northbound traffic from end hosts in the VXLAN overlay network. A distributed anycast gateway also offers the benefit of seemless host mobility in the VXLAN overlay network. Because the gateway IP and virtual MAC address are identically provisioned on all VTEPs within a VNI, when an end host moves from one VTEP to another VTEP, it doesn't need to send another ARP request to re-learn the gateway MAC address.

Figure 9. Distributed Anycast Gateway in MP-BGP EVPN



ARP Suppression in MP-BGP EVPN

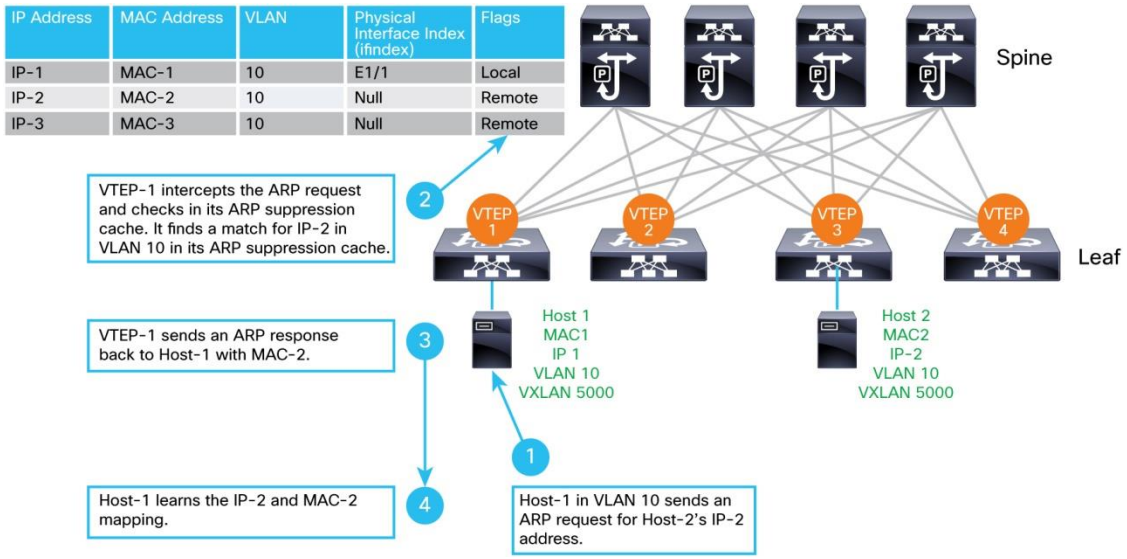
ARP suppression is an enhancement provided by the MP-BGP EVPN control plane to reduce network flooding caused by broadcast traffic from ARP requests.

When ARP suppression is enabled for a VNI, its VTEPs each maintain an ARP suppression cache table for known IP hosts and their associated MAC addresses in the VNI segment. As illustrated in Figure 10, when an end host in the VNI sends an ARP request for another end host IP address, its local VTEP intercepts the ARP request and checks for the ARPed IP address in its ARP suppression cache table. If it finds a match, the local VTEP sends an ARP response on behalf of the remote end host. The local host learns the MAC address of the remote host in the ARP response. If the local VTEP doesn't have the ARPed IP address in its ARP suppression table, it floods the ARP request to the other VTEPs in the VNI. This ARP flooding can occur for the initial ARP request to a silent host in the network. The VTEPs in the network don't see any traffic from the silent host until another host sends an ARP request for its IP address and it sends an ARP response back. After the local VTEP learns about the MAC and IP address of the silent host, the information is distributed through the MP-BGP EVPN control plane to all other VTEPs. Any subsequent ARP requests do not need to be flooded.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Because most end hosts send GARP or RARP requests to announce themselves to the network right after they come online, the local VTEP will immediately have the opportunity to learn their MAC and IP addresses and distribute this information to other VTEPs through the MP-BGP EVPN control plane. Therefore, most active IP hosts in VXLAN EVPN should be learned by the VTEPs either through local learning or control-plane-based remote learning. As a result, ARP suppression reduces the network flooding caused by host ARP learning behavior.

Figure 10. ARP Suppression in MP-BGP EVPN



MP-BGP EVPN VTEP Configuration

This section summarizes the steps for configuring MP-BGP EVPN VTEP.

Step 1. Perform the initial configuration of each VTEP switch.

Enable the VXLAN and MP-BGP EVPN control plane.

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

- `feature nv overlay` — Enable VXLAN.
- `feature vn-segment-vlan-based` — Enable VLAN-based VXLAN (currently the only mode).
- `feature bgp` — Enable BGP.
- `nv overlay evpn` — Enable the EVPN control plane for VXLAN.

Other features may need to be enabled.

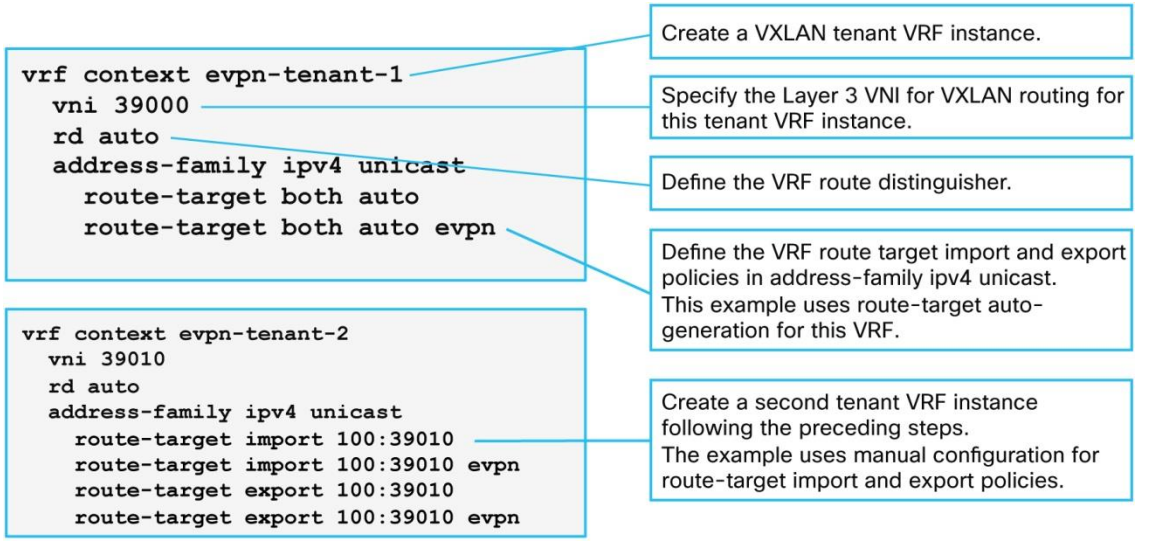
```
feature ospf
feature pim
feature interface-vlan
```

- `feature ospf` — Enable Open Shortest Path First (OSPF) if it's chosen as the underlay Interior Gateway Protocol (IGP) routing protocol.
- `feature pim` — Enable IP Protocol-Independent Multicast (PIM) routing in the underlay network.
- `feature interface-vlan` — Enable VLAN switch virtual interfaces (SVIs) if the VTEP needs to be the IP gateway and route for the VXLAN VLAN IP subnet.

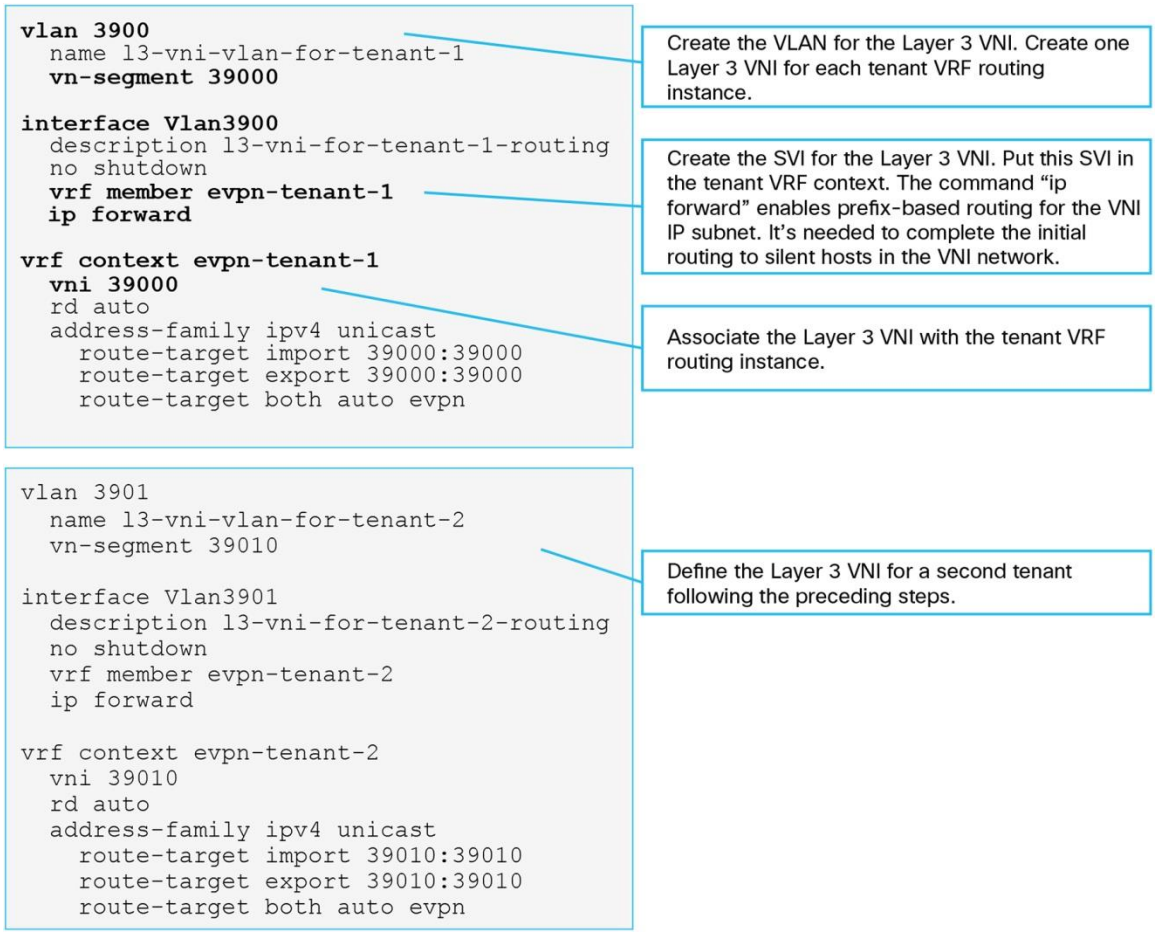
© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Step 2. Configure the EVPN tenant VRF instance.

The following example shows a configuration for two tenant VRF instances:



Step 3. Create a Layer-3 VNI for each tenant VRF instance.



© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Step 4. Configure EVPN Layer-2 VNIs for Layer-2 networks.

This step involves mapping VLANs to Layer-2 VNIs and defining their EVPN parameters.

```
vlan 200
  vn-segment 20000
vlan 210
  vn-segment 21000
```

Map the VLAN to the VXLAN VNI.

```
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 21000 l2
    rd auto
    route-target import auto
    route-target export auto
```

Under the EVPN configuration, define the route distinguisher and route target import and export policies for each Layer 2 VNI.

Step 5. Configure the SVI for Layer-2 VNIs and enable the anycast gateway under the SVI.

```
interface Vlan200
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.1.1.1/8
  fabric forwarding mode anycast-gateway


interface Vlan210
  no shutdown
  vrf member evpn-tenant-1
  ip address 21.1.1.1/8
  fabric forwarding mode anycast-gateway
```

Create the SVI for a Layer 2 VNI. Associate it with the tenant VRF instance.

All VTEPs for this VLAN and VNI should have the same SVI IP address as the distributed IP gateway.

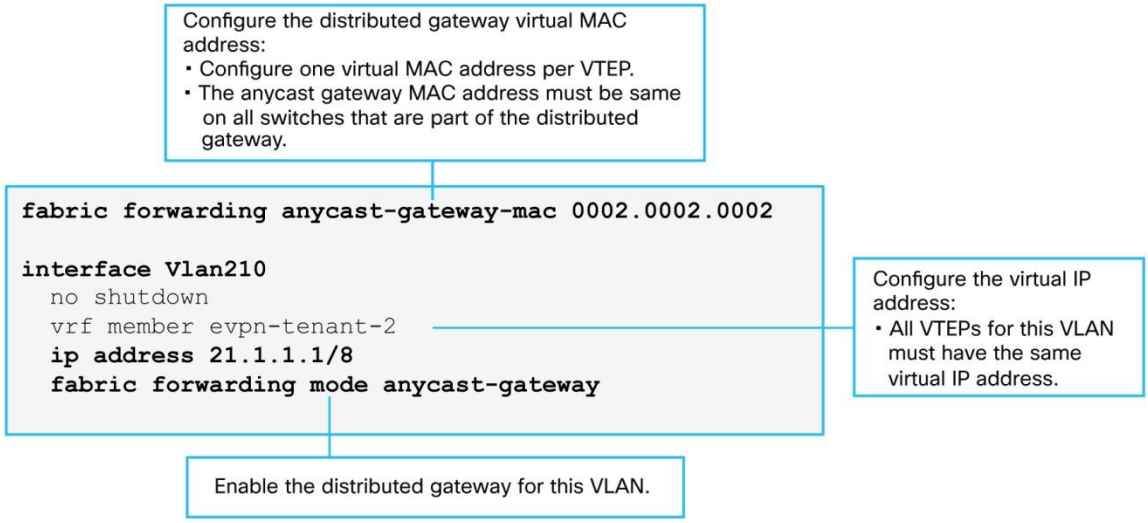
Enable the distributed anycast gateway for this VLAN and VNI.

Step 6. Configure the EVPN distributed anycast gateway.

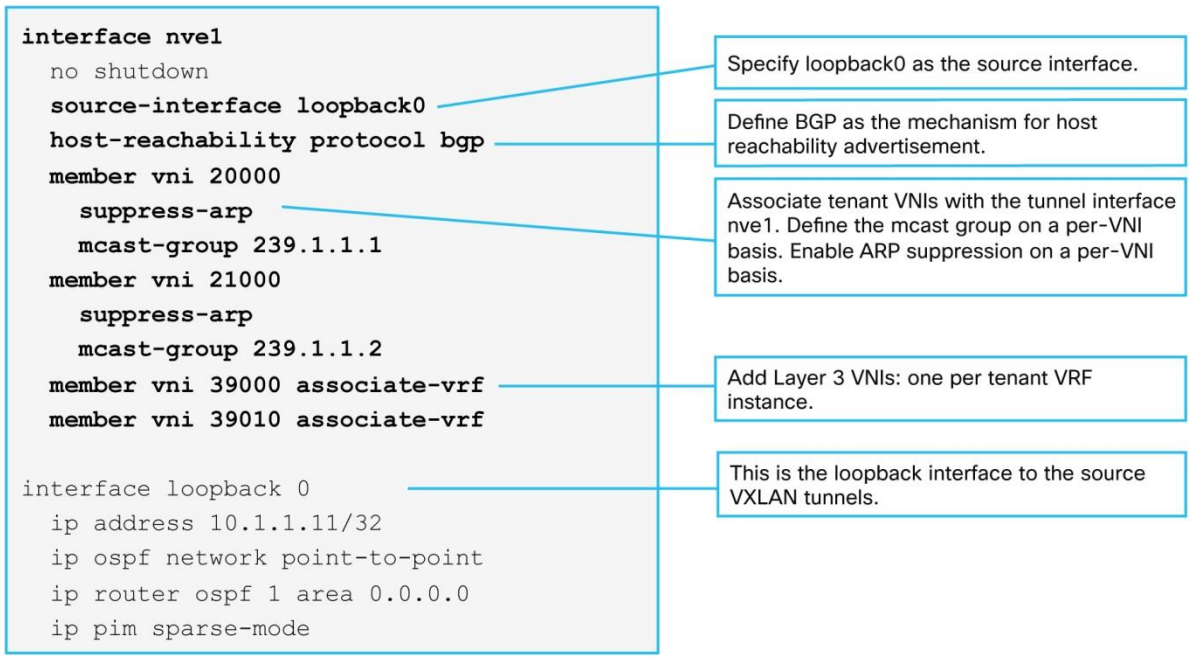
This step includes configuring the anycast gateway virtual MAC address for each VTEP and the anycast gateway IP address for each VNI.

All the VTEPs in the EVPN domain **must** have the same anycast gateway virtual MAC address and the same anycast gateway IP address for a given VNI for which they function as the default IP gateway.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.



Step 7. Configure VXLAN tunnel interface nve1 and associate Layer-2 VNIs and Layer-3 VNIs with it.



© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Step 8. Configure MP-BGP on the VTEPs.

```
router bgp 100
  router-id 10.1.1.11
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended

  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
  vrf evpn-tenant-2
    address-family ipv4 unicast
      advertise l2vpn evpn
```

- Use address-family ipv4 unicast for prefix-based routing.
- Use address-family l2vpn evpn for evpn host routes.
- Define the MP-BGP neighbors. Under each neighbor, define address-family ipv4 unicast and l2vpn evpn.
- Send extended community in address-family l2vpn evpn to distribute EVPN route attributes.
- Under address-family ipv4 unicast for each tenant VRF instance, enable advertising for EVPN routes.

Step 9. Configure the iBGP route reflector.

```
router bgp 100
  router-id 10.1.1.1
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  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 10.1.1.11
    inherit peer vtep-peer
  neighbor 10.1.1.12
    inherit peer vtep-peer
  neighbor 10.1.1.13
    inherit peer vtep-peer
  neighbor 10.1.1.14
    inherit peer vtep-peer
```

- Use address-family ipv4 unicast for prefix-based routing.
- Use address-family l2vpn evpn for EVPN VXLAN host routes. Retain all the route-target attributes.
- Use an iBGP route-reflector client peer template.
- Send both standard and extended communities in address-family ipv4 unicast.
- Send both standard and extended communities in address-family l2vpn evpn.

Virtual Port-Channel VTEP in MP-BGP EVPN VXLAN

Virtual Port-Channel (vPC) VTEP combines the two technologies, vPC and VXLAN, to provide device-level redundancy for VTEPs. A pair of vPC switches share the same VTEP address, often referred to as the anycast VTEP address, and function as a logical VTEP. The other VTEPs in the network see the two switches as a single VTEP with the anycast VTEP address. When both the vPC VTEP switches are up and running, they load share in an active-active configuration. If one vPC switch goes down, the other switch takes over the entire traffic load so that the failure event doesn't cause loss of connectivity for the devices connected to the vPC pair.

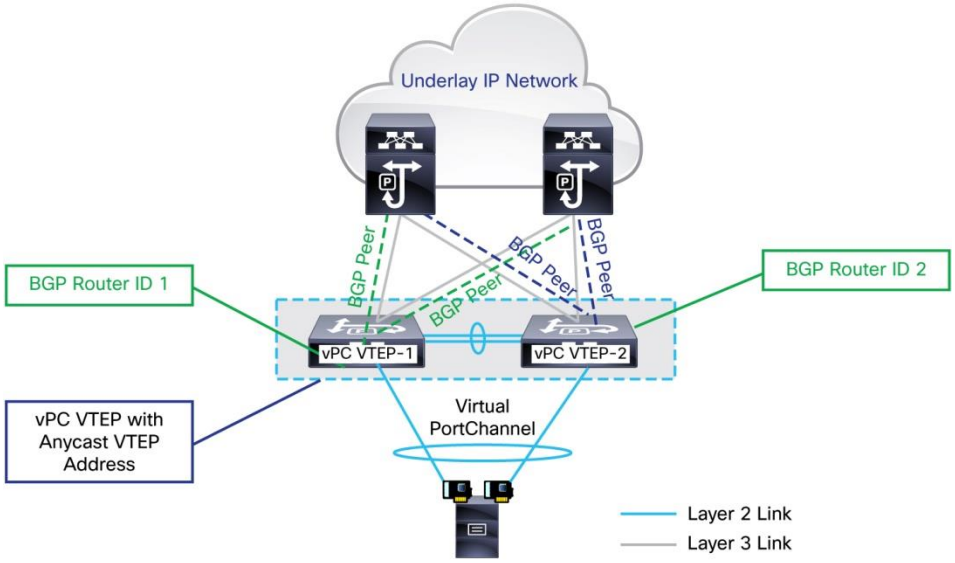
© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

The MP-BGP EVPN control plane in Cisco NX-OS is implemented to work transparently with vPC VTEP. With an MP-BGP EVPN control plane, vPC VTEPs continue to function as a single logical VTEP with the anycast VTEP address for VTEP functions, but they operate as two separate entities from the perspective of MP-BGP. They have different router IDs for BGP, form BGP neighbor adjacency with the BGP peers separately, and advertise EVPN routes independently. In the EVPN routes, they both use the anycast VTEP address as the next hop so that the remote VTEPs can use the learned EVPN routes and encapsulate packets using the anycast VTEP address as the destination in the outer IP header of encapsulated packets.

EVPN vPC VTEP Configuration

The vPC VTEP switches are configured to use a secondary IP address on the loopback interface as the VTEP address for the source of the VXLAN tunnels (interface nve1). The rest of the EVPN VXLAN configuration remains the same as for a standard single VTEP. Both switches need to have their own BGP configurations with a unique router ID. Figure 11 illustrates the concept of the MP-BGP EVPN vPC VTEP. MP-BGP uses the anycast VTEP address as the next hop when building BGP updates for EVPN routes.

Figure 11. MP-BGP EVPN vPC VTEPs



A sample vPC VTEP configuration is shown here.

vPC VTEP-1 Configuration

```
interface nve1
 no shutdown
  source-interface loopback0
  host-reachability protocol bgp
  member vni 20000
    suppress-arp
    mcast-group 239.1.1.1
  member vni 20100
    suppress-arp
    mcast-group 239.1.1.2
```

This is the VXLAN tunnel

Interface loopback0 is the source of the VXLAN tunnels.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
  member vni 39000 associate-vrf

interface loopback0
  ip address 10.1.1.13/32
  ip address 10.1.1.134/32 secondary
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

router bgp 100
  router-id 10.1.1.13
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 20100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn

n9396-vPC-VTEP-1#
```

This secondary IP address is used as the anycast VTEP address. Both vPC VTEPs need to be configured with the exact same anycast VTEP address.

The BGP instance has its own router ID: 10.1.1.13.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

vPC VTEP-2 Configuration

```
interface nve1
  no shutdown
  source-interface loopback0
  host-reachability protocol bgp
  member vni 20000
    suppress-arp
    mcast-group 239.1.1.1
  member vni 20100
    suppress-arp
    mcast-group 239.1.1.2

  member vni 39010 associate-vrf


interface loopback0
  ip address 10.1.1.14/32
  ip address 10.1.1.134/32 secondary
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

router bgp 100
  router-id 10.1.1.14
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
```

This is the VXLAN tunnel interface.

Interface loopback0 is the source of the VXLAN tunnels.

This secondary IP address is used as the anycast VTEP address. Both vPC VTEPs need to be configured with the exact same anycast VTEP address.

The BGP instance has its own router ID: 10.1.1.14.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
    route-target export auto
  vni 20100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn


n9396-vPC-VTEP-2#
```

vPC VTEP MP-BGP Status and EVPN Route Updates

To their MP-BGP neighbors, vPC VTEPs appear as two separate neighbors. The following is an example **of show bgp l2vpn evpn summary** output from a BGP neighbor of the vPC VTEPs:

```
spine-9508-1# sh bgp l2vpn evpn summary
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 10.1.1.1, local AS number 100
BGP table version is 75, L2VPN EVPN config peers 4, capable peers 4
13 network entries and 13 paths using 1716 bytes of memory
BGP attribute entries [12/1728], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.1.1.11       4   100    8247    8262       75    0    0    5d17h 6
10.1.1.12       4   100    8254    8259       75    0    0    1d08h 3
10.1.1.13       4   100    8258    8409       75    0    0    1d16h 2
10.1.1.14       4   100    8257    8455       75    0    0    1d16h 2
```

The two vPC VTEPs are shown as two separate BGP neighbors.

The two vPC VTEPs advertise EVPN routes with the same anycast VTEP address as the BGP next hop. Examples of route advertisements from the two vPC VTEPs are shown here.

On VTEP-1

```
n9396-vPC-VTEP-1# sh bgp l2vpn evpn neighbors 10.1.1.1 advertised-routes

Peer 10.1.1.1 routes for address family L2VPN EVPN:
BGP table version is 94, local router ID is 10.1.1.13
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-
injected
```

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

  Network      Next Hop      Metric   LocPrf   Weight Path
Route Distinguisher: 10.1.1.11:32967

Route Distinguisher: 10.1.1.11:32968

Route Distinguisher: 10.1.1.11:32977

Route Distinguisher: 10.1.1.12:2

Route Distinguisher: 10.1.1.12:6

Route Distinguisher: 10.1.1.13:32967    (L2VNI 20000)
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                    10.1.1.134                          100      32768 i
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                    10.1.1.134                          100      32768 i

Route Distinguisher: 10.1.1.13:32977    (L2VNI 21000)

Route Distinguisher: 10.1.1.14:32967

Route Distinguisher: 10.1.1.13:3    (L3VNI 39000)

n9396-vPC-VTEP-1#
```

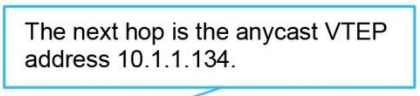
The next hop is the anycast VTEP address 10.1.1.134.

On VTEP-2.

```
n9396-vPC-VTEP-2# sh bgp l2vpn evpn neighbors 10.1.1.1 advertised-routes

Peer 10.1.1.1 routes for address family L2VPN EVPN:
BGP table version is 117, local router ID is 10.1.1.14
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-
injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 10.1.1.11:32967

Route Distinguisher: 10.1.1.11:32968

Route Distinguisher: 10.1.1.11:32977

Route Distinguisher: 10.1.1.12:2
```

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
Route Distinguisher: 10.1.1.12:6

Route Distinguisher: 10.1.1.13:32967

Route Distinguisher: 10.1.1.14:32967    (L2VNI 20000)
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                     10.1.1.134                          100      32768 i
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                     10.1.1.134                          100      32768 i

Route Distinguisher: 10.1.1.14:32977    (L2VNI 21000)

Route Distinguisher: 10.1.1.14:3    (L3VNI 39000)

n9396-vPC-VTEP-2#
```



On the other VTEPs, the EVPN routes are learned with the anycast VTEP as the next hop. The following snippet is from the **show bgp l2vpn evpn** output on a remote VTEP for the same routes as advertised in the preceding example:

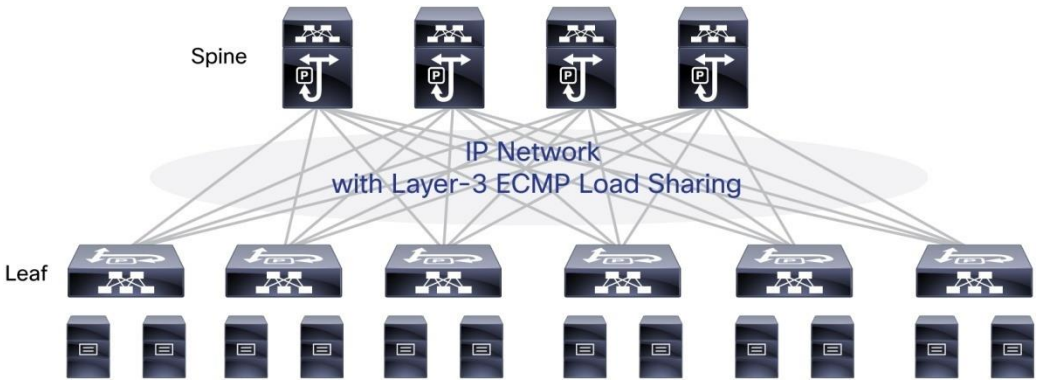
```
Route Distinguisher: 10.1.1.14:32967
* i[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                     10.1.1.134                          100          0 i
*>i                  10.1.1.134                          100          0 i
*>i[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                     10.1.1.134                          100          0 i
* i                  10.1.1.134                          100          0 i
```

MP-BGP EVPN VXLAN Fabric Design

Increasing numbers of organizations are looking at the two-tier spine-and-leaf fabric architecture when deploying new scalable data center networks (Figure 12). The two-tier fabric design provides the flexibility needed for a network to grow to accommodate applications' ever-increasing requirements for connectivity density and forwarding capacity. The fabric runs as a Layer-3 network to take advantage of the proven stability and scalability of existing Layer-3 routing protocols such as Open Shortest Path First (OSPF), BGP, and Intermediate System to Intermediate System (IS-IS).

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 12. Two-Tier Spine-Leaf Fabric Architecture



With a Layer-3 fabric, Layer-2 domains are contained under each leaf switch. For applications that assume direct Layer-2 adjacency among the computing nodes, this design could restrict workload placement. VXLAN can be deployed to extend Layer-2 domains over the Layer-3 fabric to achieve workload placement flexibility. This section discusses some typical design options for VXLAN fabric using the MP-BGP EVPN control plane for route distribution and multi-tenancy support.

MP-BGP EVPN is a new address family in BGP and uses mechanisms in BGP that are independent of the address family. It doesn't mandate the use of either iBGP or eBGP. This flexibility makes it easier for organizations to transition from their current data center BGP designs to the MP-BGP EVPN VXLAN design, This approach also provides flexibility in assignment of BGP autonomous system numbers (ASNs).This section discusses both MP-iBGP EVPN and MP-eBGP EVPN designs.

VXLAN Fabric with MP-iBGP EVPN

With MP-iBGP EVPN design, all MP-BGP speakers are in the same BGP autonomous system. To simplify the iBGP peering topology, iBGP route reflectors are often deployed in the network. An IGP routing protocol of choice can be deployed to provide IP reachability for VTEP addresses in the underlay network. Depending on the software capability and scalability, iBGP route reflectors can be placed on either the spine layer or the leaf layer, or they can be in dedicated devices for greater scalability.

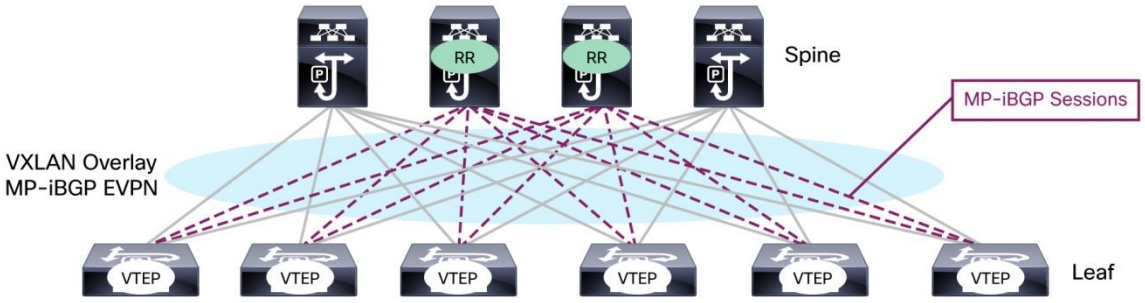
MP-iBGP Route Reflector on the Spine Layer

In this design, leaf switches are VTEP devices. They run MP-iBGP and peer with a pair of route reflectors that are running on the spine switches. This design requires the chosen spine devices to have the MP-BGP EVPN software functions, but they don't need to be VTEPs.

Figure 13 shows a sample MP-iBGP EVPN VXLAN fabric with iBGP route reflectors (RRs) on the spine layer. In this design, each VTEP leaf has two iBGP neighbors that are the two spine BGP route reflectors. Each spine BGP route reflector has all the VTEP leaf nodes as route reflector clients and reflects EVPN routes for the VTEP leaf nodes.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 13. MP-iBGP EVPN VXLAN Fabric Design with Route Reflectors on the Spine Layer



The following sample shows the MP-iBGP configuration on VTEP leaf nodes in this design:

```
n9396-vtep-1# sh run bgp

!Command: show running-config bgp
!Time: Fri Jan 23 07:38:48 2015

version 7.0(3)I1(1)
feature bgp

router bgp 100
  router-id 10.1.1.11
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended

  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 20100 l2
```

Configure the two spine BGP route reflectors as two iBGP neighbors. Under each neighbor, send extended community in address-family l2vpn evpn. EVPN routes use extended community to carry EVPN attributes.

Advertise EVPN routes to address-family ipv4 unicast. This step is optional. It's needed in case this VTEP routes to an external device such as a WAN edge router and needs to distribute EVPN routes to the outside.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
    rd auto
    route-target import auto
    route-target export auto
  vni 21000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 21100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn
vrf context evpn-tenant-2
  rd auto
  address-family ipv4 unicast
    route-target import 39010:39010
    route-target export 39010:39010
    route-target both auto evpn


n9396-vtep-1#
```

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

The following sample shows an MP-iBGP configuration on the spine BGP route reflector:

```
feature bgp
nv overlay evpn
router bgp 100
  router-id 10.1.1.1
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all

 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 10.1.1.11
    inherit peer vtep-peer
  neighbor 10.1.1.12
    inherit peer vtep-peer
  neighbor 10.1.1.13
    inherit peer vtep-peer
  neighbor 10.1.1.14
    inherit peer vtep-peer
```

Enable MP-BGP l2vpn evpn.

Use address-family l2vpn evpn for VXLAN EVPN routes. The original route-target attributes must be retained while advertising EVPN routes from one iBGP route-reflector client to the others. This requirement is important to allow the routes to be received by the other route-reflector clients.

Use the iBGP RR client peer template.

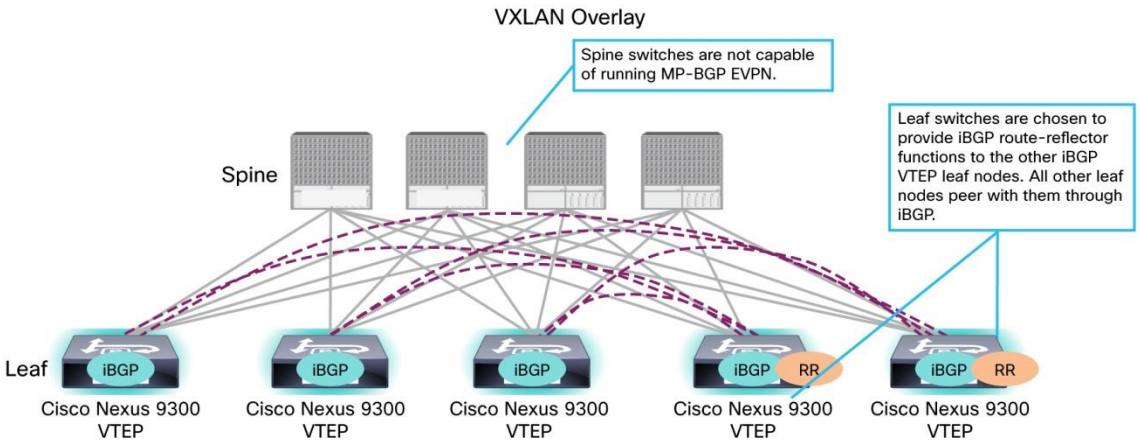
Send both standard and extended community in address-family l2vpn evpn.

VTEP leaf nodes are iBGP route-reflector clients.

MP-iBGP Route Reflector on the Leaf Layer

Placement of BGP route reflectors on the spine layer is an intuitive design for MP-iBGP EVPN. It requires the chosen spine devices to support the software functions of the MP-iBGP EVPN protocol so that they can process and distribute MP-iBGP updates for EVPN routes. If the spine devices are not capable of running MP-BGP EVPN, then the BGP route-reflector functions need to be moved to the leaf layer, where leaf switches support MP-BGP EVPN and VTEP functions (Figure 14).

Figure 14. MP-iBGP EVPN Fabric Design with BGP Router Reflector Functions on the Leaf Layer



In this design, the spine switches don't participate in the MP-BGP EVPN control plane at all. They run the underlay network routing protocol to establish IP reachability for the VTEP addresses and for the iBGP peering addresses if they are not the same as the VTEP addresses: for instance, on vPC VTEPs.

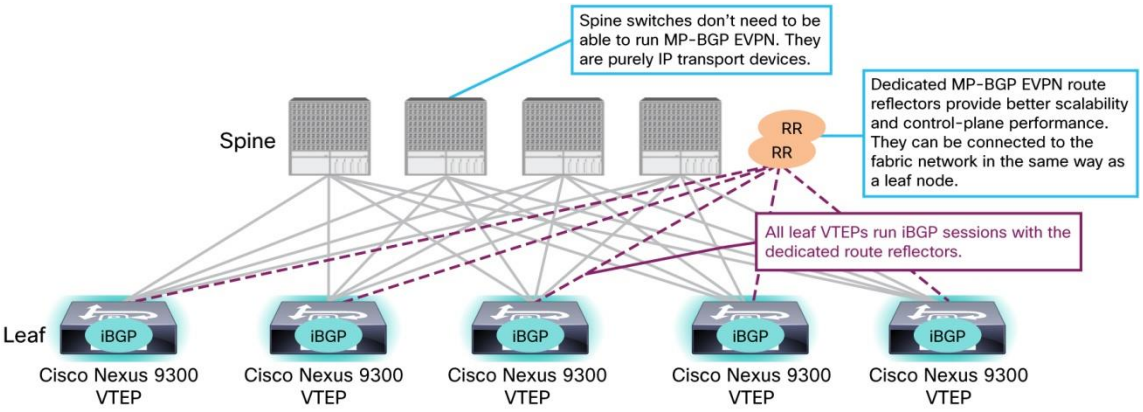
© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

MP-iBGP with Dedicated Route Reflectors

The role of MP-iBGP route reflectors in EVPN is the same as for the standard iBGP route reflectors, which is to reflect BGP updates between iBGP peers so that they don't need to form a fully meshed iBGP peering topology. This approach significantly simplifies iBGP topology and makes the protocol more scalable. Because the route reflector functions are purely a control-plane functions, BGP route reflectors don't need to be in the data-plane forwarding path. This feature allows great flexibility in route-reflector placement and platform selection.

An option for a scalable design is to use dedicated devices as route reflectors, out of the data path (Figure 15). The chosen devices need to support MP-BGP EVPN and must have the appropriate BGP control-plane scalability and computing power needed for fast convergence. The use of dedicated route reflectors eliminates the MP-BGP EVPN function requirements in the spine layer. It also removes the burden from the VTEP leaf nodes of having to run the BGP route-reflector functions in addition to performing data forwarding. Although logically the VTEP leaf nodes have direct iBGP neighbor adjacency with the route reflectors, the route reflectors can be physically connected to the VXLAN fabric network in the same way as leaf nodes and have the iBGP sessions between VTEP leafs and route reflectors to go through multiple hops (usually 2) in the fabric underlay network. Routing considerations need to be applied so that the underlay data paths between VTEP addresses don't go through the route reflectors. This requirement helps ensure that the route reflectors are out of the data forwarding path.

Figure 15. MP-iBGP EVPN Design with Dedicated Route Reflectors



VXLAN Fabric with MP-eBGP EVPN

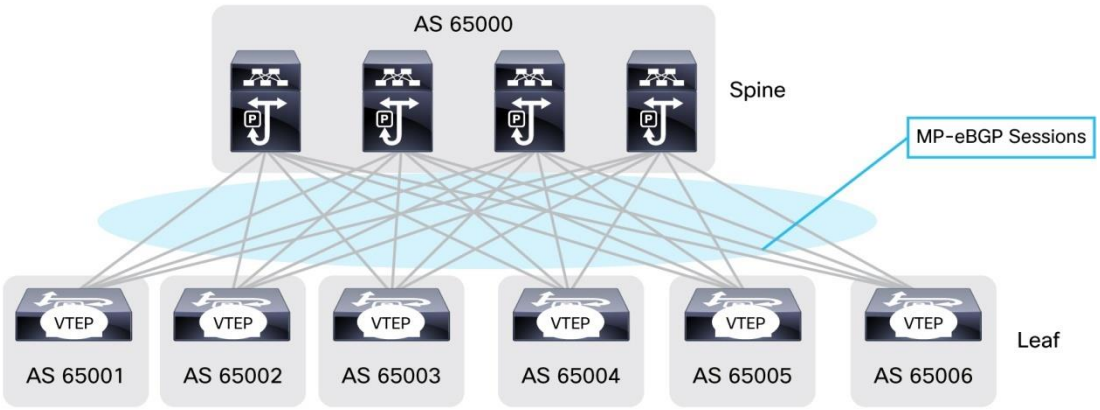
Although a MP-iBGP EVPN design is common practice, some organizations choose to run eBGP between their leaf and spine layers. MP-BGP EVPN has the flexibility to work with both iBGP and eBGP. EVPN with MP-eBGP peering is a viable design option. An eBGP design offers several options for BGP autonomous system(AS) allocation. Figure 16 shows a design with each VTEP leaf in its own unique BGP AS, and Figure 17 shows another design in which all VTEP leaf nodes are in the same AS, but they all peer through eBGP with the spine switches.

Because MP-BGP EVPN is an extension of BGP, it inherits the standard BGP behaviors. In an MP-BGP EVPN network, some of the default behaviors are not desired. For example, when a BGP router advertises BGP routes to an eBGP peer, by default it changes the BGP next hop to its own IP address. In MP-BGP EVPN, when a VTEP initiates a BGP update to advertise its EVPN routes, it uses its own VTEP address as the BGP next hop. This next hop needs to be preserved throughout the hop-by-hop BGP route distribution so that the other VTEPs can receive the EVPN routes with the original VTEP address as the next hop and can use this route to initiate VXLAN tunneling in the data plane.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Therefore, the eBGP on the spine switches needs to be configured so that it does not change the BGP next hop. A BGP router also may modify BGP community attributes when sending eBGP routes. In MP-EVPN, this change could cause route-target attributes in the EVPN routes to be modified or removed. Therefore, additional configuration needs to be applied on the intermediate eBGP peers to help ensure that they retain all route-target attributes.

Figure 16. MP-eBGP EVPN VXLAN Fabric with VTEP Leaf Nodes in Unique Autonomous Systems



Because every VTEP has a unique BGP AS in this design, route-target auto-generation in NX-OS will result in different route-targets on VTEPs for the same VNI. It is recommended to manually configure import and export route targets to ensure VTEPs have the same route target configuration for the same Layer-3 VRF instance and for the same EVPN Layer-3 VNI.

The following sample shows the MP-BGP configuration for a spine switch and a VTEP leaf as shown in Figure 16. The MP-BGP configuration on a spine switch includes the application of outbound policy on the spine switches so that it doesn't change the eBGP route next hop. The example also shows the manual route-target configuration on a VTEP leaf for both Layer-3 VRF instances and EVPN Layer-2 VNIs.

```
[BGP configuration on a spine switch as in Figure 16 design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65000
  router-id 10.1.1.1
  address-family ipv4 unicast
    redistribute direct route-map permitall
  address-family l2vpn evpn
    nexthop route-map permit-all
    retain route-target all
  neighbor 192.167.11.2 remote-as 65001
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
      route-map permit-all out
  neighbor 192.168.12.2 remote-as 65002
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
      route-map permit-all out
```

Set next-hop policy to not change the next-hop attributes.

Retain routes with all route targets when advertising the EVPN BGP routes to eBGP peers.

Set outbound policy to advertise all routes to this eBGP neighbor.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
[Manual Configuration for import & export route-targets on a
VTEP leaf in Figure 16 design]

vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 65001:39000
    route-target import 65001:39000 evpn
    route-target export 65001:39000
    route-target export 65001:39000 evpn


vrf context evpn-tenant-2
 vni 39010
  rd auto
  address-family ipv4 unicast
    route-target import 65001:39010
    route-target import 65001:39010 evpn
    route-target export 65001:39010
    route-target export 65001:39010 evpn

evpn
  vni 20000 l2
    rd auto
    route-target import 65001:20000
    route-target export 65001:20000
   vni 21000 l2
    rd auto
    route-target import 65001:21000
    route-target export 65001:21000
```

Manually configure import and export route-targets for the Layer-3 VRF instance evpn-tenant-1.

Manually configure import and export route-targets for the Layer-3 VRF instance evpn-tenant-2.

Manually configure import and export route-targets for the Layer-2 VNIs under EVPN configuration.

```
[BGP configuration on a leaf switch as in Figure 16
design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65001
  address-family ipv4 unicast
  neighbor 192.167.11.1 remote-as 65000
    address-family ipv4 unicast
      allowas-in
      send-community extended
    address-family l2vpn evpn
      send-community extended
  neighbor 192.168.11.1 remote-as 65000
    address-family ipv4 unicast
      send-community extended
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
```

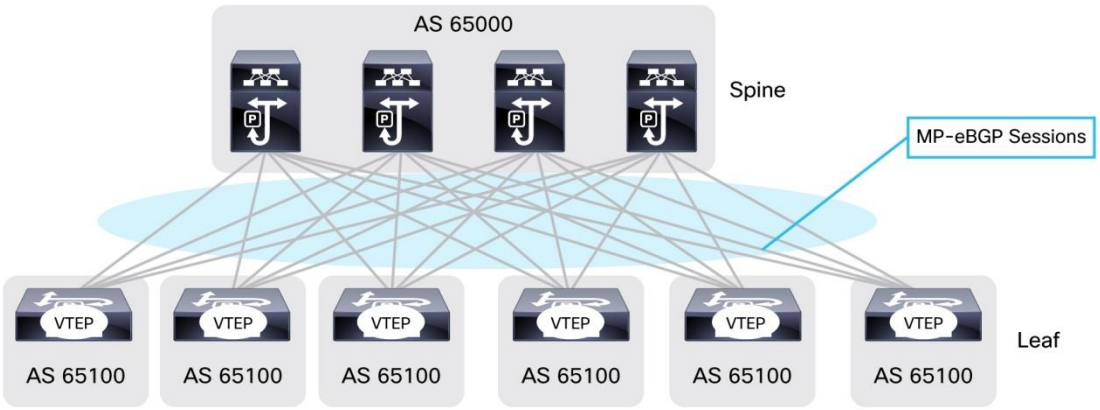
Spine switch 1 is an eBGP neighbor.

Spine switch 2 is an eBGP neighbor.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 17 depicts a MP-eBGP design with all leaf nodes in the same autonomous system, but they each peer with the spine nodes through MP-eBGP.

Figure 17. MP-eBGP Design with VTEP Leaf Nodes in the Same BGP Autonomous System



The following sample shows a configuration for a VTEP leaf and spine switch design, as shown in Figure 17. In addition to the configuration in the Figure 16 design, the spine switches in Figure 17 need to have **peer-as-check** disabled because they need to pass MP-BGP EVPN routes between two eBGP neighbors that are in the same BGP autonomous system. The VTEP leaf nodes in Figure 17 need to have **allowas-in** enabled so that they accept BGP routes from the other VTEPs that are in the same BGP autonomous system as they are. Because all the VTEP leafs are in the same BGP autonomous system in this design, it is suitable to use system auto-generated import and export route targets for the Layer-3 VRF instances and the EVPN Layer-2 VNIs.

```
[BGP configuration on a spine switch as in Figure 17 design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65000
  router-id 10.1.1.1
  address-family ipv4 unicast
    redistribute direct route-map permit-all
  address-family l2vpn evpn
    nexthop route-map permit-all
    retain route-target all
  neighbor 192.167.11.2 remote-as 65100
    address-family ipv4 unicast
    address-family l2vpn evpn
      disable-peer-as-check
      send-community extended
      route-map permit-all out
  neighbor 192.168.12.2 remote-as 65100
    address-family ipv4 unicast
    address-family l2vpn evpn
      disable-peer-as-check
      send-community extended
      route-map permit-all out
```

Set next-hop policy to not change the next-hop attributes.

Retain all the route-target attributes when advertising the EVPN BGP routes to eBGP peers.

The VTEP leaf is an eBGP peer. All VTEPs are in the same BGP autonomous system: AS 65100.

Disable peer-as-check for this neighbor.

Set outbound policy to advertise all routes to this eBGP neighbor.

VTEP leaf is an eBGP peer. All VTEPs are in the same BGP autonomous system: AS 65100.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
[BGP configuration on a leaf switch in Figure 17 design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65001
  address-family ipv4 unicast
  neighbor 192.167.11.1 remote-as 65000
    address-family ipv4 unicast
      allowas-in
      send-community extended
    address-family l2vpn evpn
      allowas-in
      send-community extended
  neighbor 192.168.11.1 remote-as 65000
    address-family ipv4 unicast
      allowas-in
      send-community extended
    address-family l2vpn evpn
      allowas-in
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
```

Spine switch 1 is an eBGP neighbor.

Allow BGP routes with the local autonomous system in the autonomous system path from this neighbor.

Spine switch 2 is an eBGP neighbor.

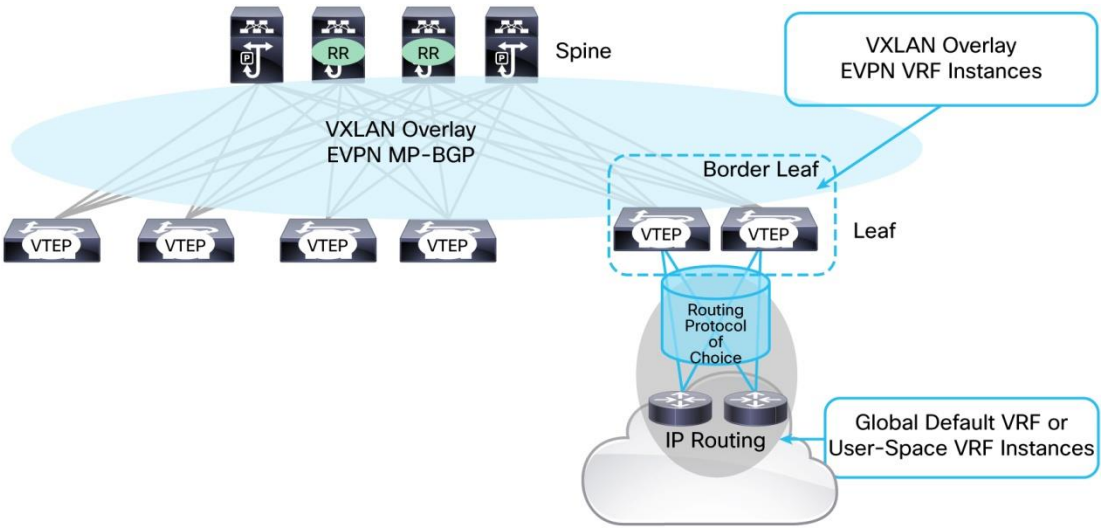
Allow BGP routes with the local autonomous system in the autonomous system path from this neighbor.

External Routing for MP-BGP EVPN VXLAN

In most organizations, the data center is not isolated from the rest of the network, including the campus network, WAN, and Internet. When EVPN VXLAN fabric is deployed in the data center, it needs to maintain connectivity with these networks that are external to the VXLAN fabric.

With the standard spine-and-leaf fabric architecture, external connectivity can be achieved by using border leaf nodes to connect to the outside routing devices. Figure 18 illustrates such a design with a pair of border leaf switches.

Figure 18. Border Leaf Switches for External Routing of MP-BGP EVPN VXLAN Fabric



The border leaf switch runs MP-BGP EVPN on the inside with the other VTEPs in the VXLAN fabric and exchanges EVPN routes with them. At the same time, it runs the normal IPv4 or IPv6 unicast routing in the tenant VRF instances with the external routing device on the outside. The routing protocol can be regular eBGP or any IGP of choice. By design, MP-BGP EVPN automatically imports the BGP routes learned in the IPv4 or IPv6 unicast address family into the L2VPN EVPN address family.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

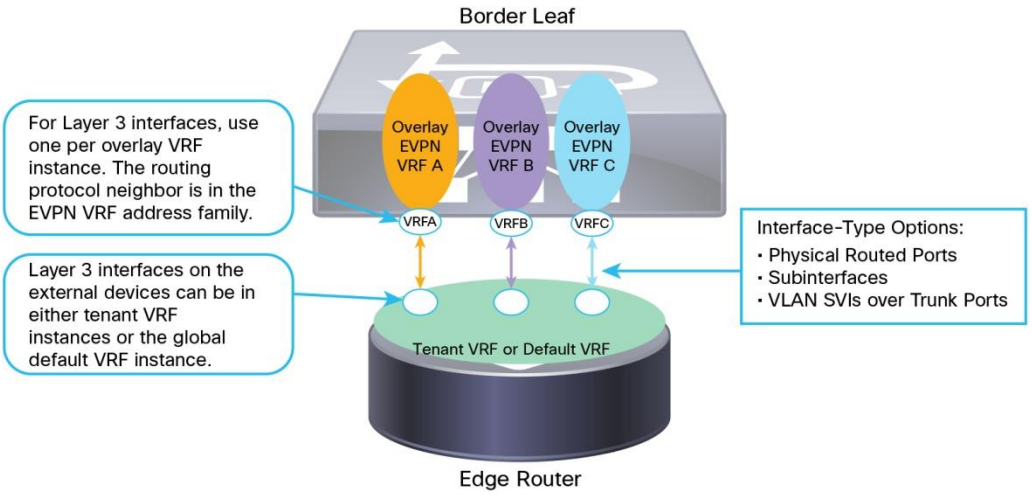
Therefore, after a border leaf switch learns the external routes, it can advertise them to the EVPN domain as EVPN routes so that other VTEP leaf nodes can also learn about the external routes for sending outbound traffic.
The border leaf switch can also be configured to send EVPN routes learned in the L2VPN EVPN address family to the IPv4 or IPv6 unicast address family and advertise them to the external routing device. Therefore, if any public subnets exist in the VXLAN fabric, they can be advertised to the outside so that the inbound traffic from the outside to these public subnets can be routed to the VXLAN fabric.

Because MP-BGP EVPN has built-in multitenancy, Layer-3 subnets in the VXLAN overlay network are in a tenant VRF routing instance. Different tenants can maintain their separate Layer-3 routing instances by default. Therefore, external routing for different tenants needs to be provided separately. The border leaf needs to have a Layer-3 interface to the outside for each tenant VRF instance for which it runs external routing (Figure 19).

To extend such Layer-3 routing segmentation among different tenants to the external network, the external router can also place its Layer-3 interfaces for the border leaf in tenant VRF instances. The routing sessions between the border leaf and the external router will run in VRF-lite on both sides.

In designs that terminate the Layer-3 segmentation on the VXLAN border leaf, the external router can run all the routing sessions in the default routing table. In this case, the routes from different tenant routing instances in the VXLAN fabric will be merged into the same default routing table on the outside. Because the tenants essentially share the external routing in this type of design, the IP addresses of the VXLAN tenants cannot overlap.

Figure 19. MP-BGP EVPN VXLAN Fabric External Routing with Multitenancy



Sample Configuration for eBGP Between the VXLAN EVPN Border Leaf and the External Router

The following is a sample configuration with eBGP routing between the VXLAN border leaf and the external router. The eBGP session is in the tenant VRF instance on the border leaf, but in the default routing table for the external router for shared external routing.

On the border leaf, BGP is configured to advertise the VXLAN IP subnet prefixes. By default, BGP advertises the MP-BGP EVPN IP host routes. Route filtering is applied in the sample configuration to block the/32 IP host routes so that only prefix routes are advertised to the external router. Because the outside doesn't need the specific host routes for inbound traffic, this approach allows better router scalability for external routing.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

On the VXLAN Border Leaf:

```
router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      network 20.0.0.0/24
    neighbor 30.10.1.2 remote-as 200
      address-family ipv4 unicast
       prefix-list outbound-no-hosts out
ip prefix-list outbound-no-hosts seq 5 deny 0.0.0.0/0 eq 32
ip prefix-list outbound-no-hosts seq 10 permit 0.0.0.0/0 le 32
```

The eBGP neighbor is on the outside. It's in address-family ipv4 unicast of the tenant VRF routing instance.

For better scalability, apply prefix-list to filter out /32 IP host routes. Advertise prefix routes only to the external eBGP neighbor.

BCP Configuration on the External Router:

```
router bgp 200
  router-id 10.1.1.18
  log-neighbor-changes
  address-family ipv4 unicast
    network 100.0.0.0/24
    network 100.0.1.0/24
  neighbor 30.10.1.1 remote-as 100
    address-family ipv4 unicast
```

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

In the preceding example, the VNI subnet route 20.0.0.0/24 is advertised to the external router through VRF-lite eBGP as shown in the global routing table, as follows:

```
N9372TX-2-ext# sh ip bgp 20.0.0.0/24
BGP routing table information for VRF default, address family IPv4 Unicast
BGP routing table entry for 20.0.0.0/24, version 36
Paths: (1 available, best #1)
Flags: (0x00001a) on xmit-list, is in urib, is best urib route

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop
  AS-Path: 100 , path sourced external to AS
    30.10.1.1 (metric 0) from 30.10.1.1 (20.0.0.1)
      Origin IGP, MED not set, localpref 100, weight 0

  Path-id 1 not advertised to any peer
N9372TX-2-ext#
N9372TX-2-ext# sh ip route 20.0.0.0/24
IP Route Table for VRF "default"

20.0.0.0/24, ubest/mbest: 1/0
    *via 30.10.1.1, [20/0], 1w2d, bgp-200, external, tag 100
N9372TX-2-ext#
```

The routes learned from the external router are distributed to the VXLAN fabric by the border leaf through the MP-BGP EVPN protocol. The following sample shows the capture of an external route on an internal VTEP. The VTEP learns the external route from the border leaf through the route reflector. The route is distributed through MP-BGP EVPN.

```
n9396-vtep-1# sh vrf evpn-tenant-1 detail
VRF-Name: evpn-tenant-1, VRF-ID: 3, State: Up
    VPNID: unknown
    RD: 10.1.1.11:3
    VNI: 39000
    Max Routes: 0  Mid-Threshold: 0
    Table-ID: 0x80000003, AF: IPv6, Fwd-ID: 0x80000003, State: Up
    Table-ID: 0x00000003, AF: IPv4, Fwd-ID: 0x00000003, State: Up

n9396-vtep-1#

n9396-vtep-1# sh bgp l2vpn evpn rd 10.1.1.11:3 100.0.0.0
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.1.1.11:3    (L3VNI 39000)
BGP routing table entry for [5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/224, version 324
Paths: (1 available, best #1)
Flags: (0x00001a) on xmit-list, is in l2rib/evpn

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from 10.1.1.16:3:[5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/120
  AS-Path: NONE, path sourced internal to AS
    10.1.1.16 (metric 3) from 10.1.1.1 (10.1.1.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 39000
      Extcommunity:  RT:100:39000 ENCAP:8 Router MAC:6412.2574.6ae7
      Originator: 10.1.1.16 Cluster list: 10.1.1.1

  Path-id 1 not advertised to any peer

n9396-vtep-1#
```

The external route is distributed through EVPN and imported into the tenant VRF instance.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
n9396-vtep-1# sh ip bgp vrf evpn-tenant-1 100.0.0.0
BGP routing table information for VRF evpn-tenant-1, address family IPv4 Unicast
BGP routing table entry for 100.0.0.0/24, version 70
Paths: (1 available, best #1)
Flags: (0x08041a) on xmit-list, is in urib, is best urib route
  vpn: version 75, (0x100002) on xmit-list

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from unknown dest
  AS-Path: NONE, path sourced internal to AS
    10.1.1.16 (metric 3) from 10.1.1.1 (10.1.1.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 39000
      Extcommunity:  RT:100:39000 ENCAP:8 Router MAC:6412.2574.6ae7
      Originator: 10.1.1.16 Cluster list: 10.1.1.1

  VRF advertise information:
  Path-id 1 not advertised to any peer

  VPN AF advertise information:
  Path-id 1 not advertised to any peer
n9396-vtep-1#
n9396-vtep-1# sh ip route vrf evpn-tenant-1 100.0.0.0/24
IP Route Table for VRF "evpn-tenant-1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

100.0.0.0/24, ubest/mbest: 1/0
    *via 10.1.1.16%default, [200/0], 01:01:14, bgp-100, internal, tag 100 (evpn)segid: 0x9858 tunnelid:
0xa010110 encap: 1

n9396-vtep-1#
```

This is the external route.

The next hop is the VTEP address of the border leaf.

The tenant is VRF L3 VNI.

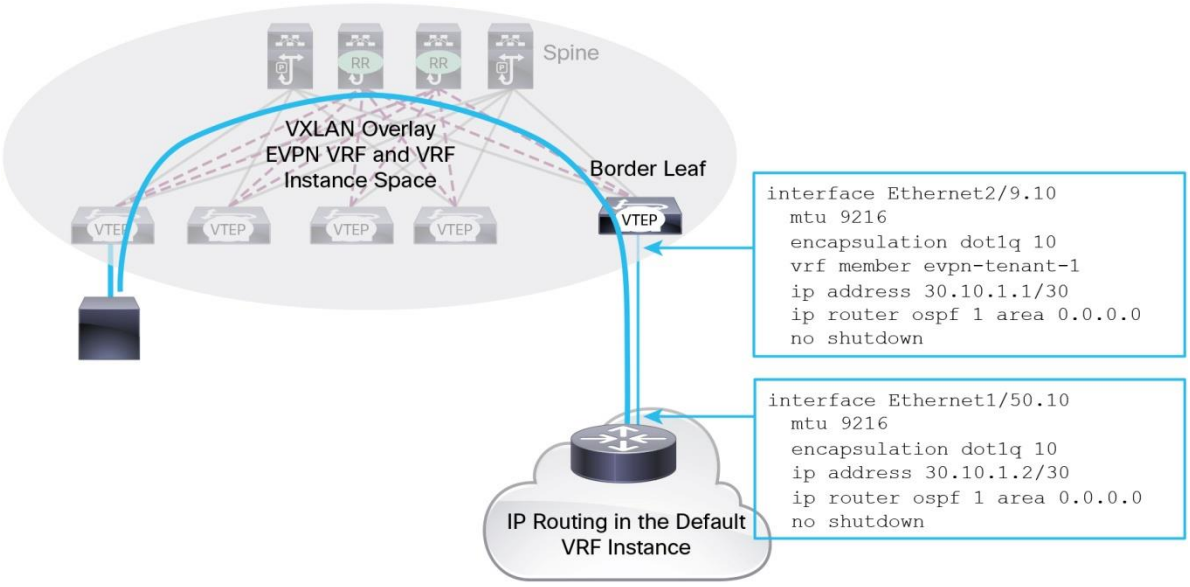
10.1.1.16 is the BGP router ID of the border leaf. 10.1.1.1 is the spine route reflector.

This is the iBGP route. The next hop is the VTEP address of the border leaf.

Sample Configuration for OSPF Between the VXLAN EVPN Border Leaf and the External Router

The example in Figure 20 uses OSPF as the external routing protocol on the EVPN VXLAN border leaf to exchange routes with the outside. For multitenancy, the example uses subinterfaces for routing between the border leaf and the external router. With subinterfaces, multiple tenants can share the same physical links for external routing, with one subinterface for each tenant VRF routing instance on the border leaf. In this example, the routing on the external router is in the default VRF instance. You also can extend the tenant VRF instances on the external device by configuring VRF-lite subinterfaces on it.

Figure 20. EVPN VXLAN External Routing with OSPF



© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

The relevant configuration on the border leaf is shown here:

```
ip prefix-list bgp-ospf-no-hosts seq 5 permit 0.0.0.0/0 eq 32
route-map permit-bgp-ospf deny 5
  match ip address prefix-list bgp-ospf-no-hosts
route-map permit-bgp-ospf permit 10
route-map permit-ospf-bgp permit 10

router ospf 1
  router-id 10.1.1.16
  vrf evpn-tenant-1
    redistribute bgp 100 route-map permit-bgp-ospf

router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
      redistribute ospf 1 route-map permit-ospf-bgp
```

Redistribute BGP routes to OSPF. Filter out /32 IP host routes.

A BGP router will modify route targets in l2vpn evpn routes when it is an autonomous system boundary router. The original route target must be retained.

Redistribute OSPF to BGP. Advertise the redistributed routes to L2VPN EVPN.

In this design, the border leaf learns external routes through OSPF in the tenant VRF instances. It redistributes the routes to MP-BGP within the VRF instances and then advertises them through MP-BGP L2VPN EVPN to the internal VTEPs.

The following example shows external route distribution on the border leaf:

```
n9396-border-leaf# sh ip route 100.0.0.0/24 vrf evpn-tenant-1
IP Route Table for VRF "evpn-tenant-1"

100.0.0.0/24, ubest/mbest: 1/0
    *via 30.10.1.2, Eth2/9.10, [110/2], 01:43:07, ospf-1, intra

n9396-border-leaf# sh bgp l2vpn evpn 100.0.0.0 vrf evpn-tenant-1
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.1.1.16:3    (L3VNI 39000)
BGP routing table entry for [5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/224, version 325
Paths: (1 available, best #1)
Flags: (0x00000a) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    10.1.1.16 (metric 0) from 0.0.0.0 (10.1.1.16)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 39000
      Extcommunity:  RT:100:39000

  Path-id 1 advertised to peers:
    10.1.1.1          10.1.1.2

n9396-border-leaf#
```

This is an external route learned through OSPF in the tenant VRF.

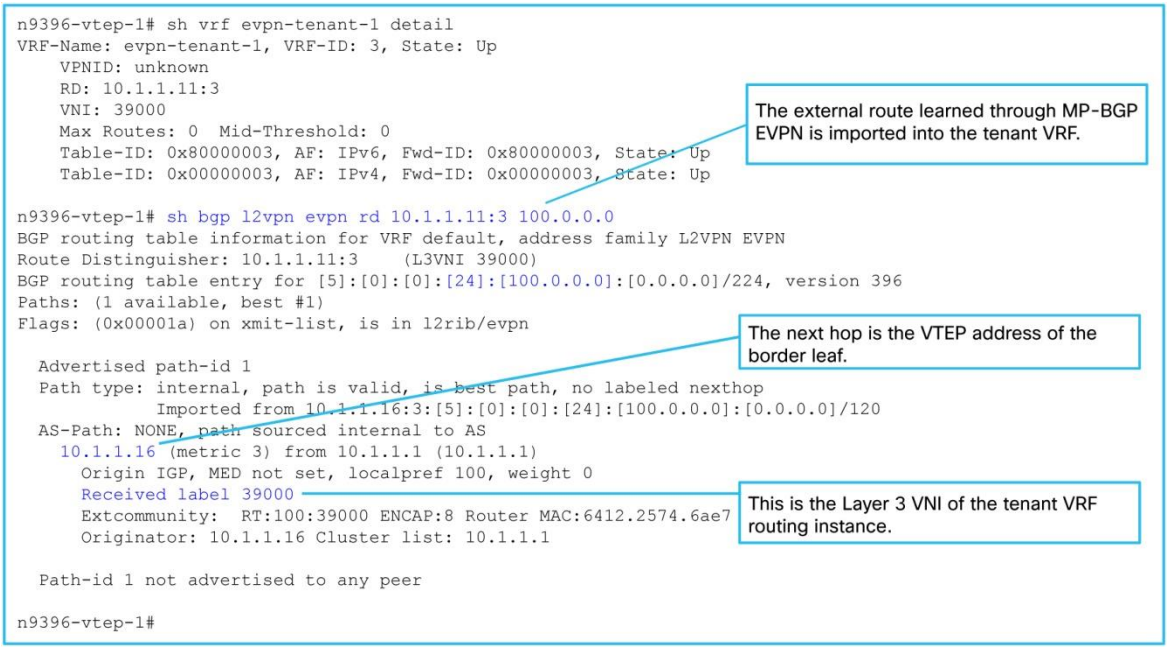
The external OSPF route is redistributed to BGP and distributed to other VTEPs through MP-BGP L2VPN EVPN.

The BGP next hop is the VTEP address of the border leaf.

The MP-BGP EVPN route is advertised to the BGP peers.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

The internal VTEPs learn the external routes through MP-BGP EVPN:



Scalability Considerations for the EVPN VXLAN Border Leaf Nodes

The VXLAN border leaf nodes are the connection points of a VXLAN fabric network to the outside. They learn external routes and redistribute them to other VTEPs through MP-BGP EVPN. At the same time, they advertise to the outside the public subnets that are on the VXLAN fabric.

Distribution of External Routes to the EVPN VXLAN Fabric

A border leaf may receive a large number of external routes from the outside. Because border leaf nodes are normally the exit gateway for the fabric's internal devices, all the external routes may not need to be distributed to the fabric. Instead, you may want to summarize the routes before advertising them to MP-BGP EVPN. In some cases, advertising a default route to the fabric on a per-tenant basis can be sufficient. Reducing the number of distributed external routes helps ensure that the internal VTEP devices do not run out of the longest-prefix-match (LPM) routing table resources. This approach also reduces the MP-BGP EVPN control plane burden on the internal VTEPs, resulting in better control-plane performance.

EVPN VXLAN Fabric Internal Network Advertisements to the Outside

Some Layer-3 subnets in an EVPN VXLAN overlay network need to be reachable from the outside. The border leaf nodes need to advertise the Layer-3 reachability information for these public subnets. MP-BGP EVPN may distribute both IP host routes and inside subnet prefix routes on the outside. In the routing protocol session between the border leaf and the external router, you can apply filters to avoid sending the internal IP host routes to the outside. In most of cases, LPM prefix routes for the public subnets are what the outside network needs to send traffic to the VXLAN fabric.

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

EVPN Tenant Scalability on the Border Leaf Nodes

The border leaf provides external connectivity for the tenants in the VXLAN overlay network. They need to participate in all the tenant VRF routing instances for which they serve as border leaf nodes. In building a large-scale multitenancy design, follow the requirements for the maximum number of EVPN Layer-3 VRF instances that a border leaf can support.

IP Host Route Scalability on the Border Leaf Nodes

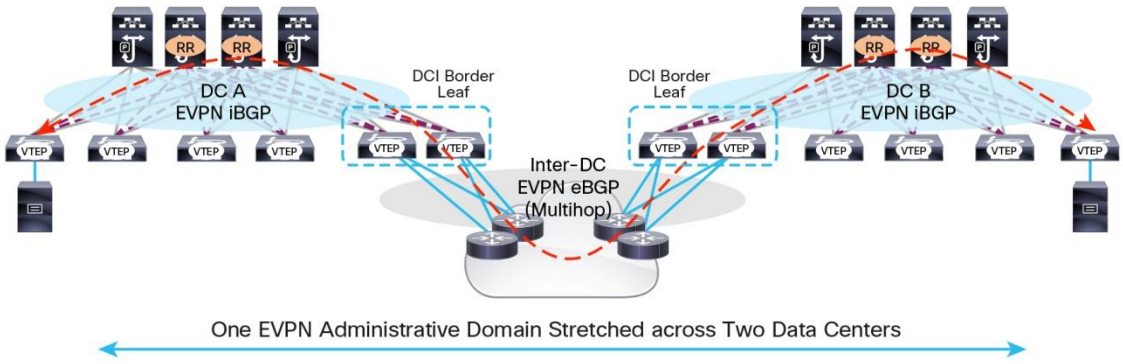
To achieve optimal forwarding for inbound traffic destined for internal end hosts, the border leaf needs to perform IP host-based routing for end hosts in the tenant public subnets. This requirement implies that the border leaf needs to learn and program the host routes in the hardware forwarding table for IP host routes. The IP host table size dictates the total number of end hosts that can be present in the tenant public subnets.

Data Center Interconnect for MP-BGP EVPN VXLAN

Although Overlay Transport Virtualization (OTV) and Virtual Private LAN Service (VPLS) remain the most proven Layer-2 data center interconnect (DCI) solutions, VXLAN with an MP-BGP EVPN control plane can offer an alternative under certain deployment conditions. When VXLAN is deployed within data centers, use of it for interconnection between data centers can simplify the overall network design and reduce operational complexity, providing a unified network overlay solution for traffic both within and between data centers.

Figure 21 illustrates a simple data center and DCI design with MP-BGP EVPN VXLAN. In this design, each data center maintains its own BGP autonomous system and deploys EVPN VXLAN fabric running MP-iBGP with route reflectors for simplicity and scalability. Between data centers, the DCI border leaf nodes run multihop MP-eBGP EVPN with each other. Consequently, the two data centers are joined together to form one unified MP-BGP EVPN routing domain. In the control plane, EVPN routes are distributed through the iBGP-eBGP-iBGP path between the data centers. In the data plane, when an end host in data center A sends traffic to another host in data center B, the data packets traverse one VXLAN tunnel and are encapsulated by the ingress VTEP in data center A and decapsulated by the egress VTEP in data center B. This approach provides highly effective DCI data forwarding in the overlay network.

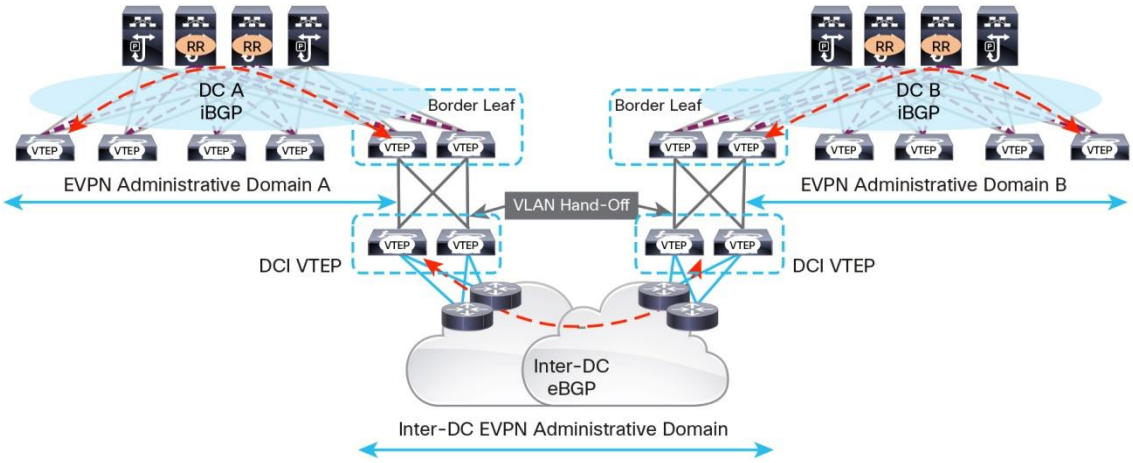
Figure 21. DCI Solution with a Unified MP-BGP EVPN Administrative Domain



© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 22 shows another DCI design with MP-BGP EVPN. It has a separate MP-iBGP EVPN domain for each data center, and it joins them together through an inter-data center MP-eBGP EVPN domain between the DCI VTEPs. The MP-eBGP session between the DCI VTEPs needs to be multihop if the VTEPs are not directly connected. This design provides the flexibility of deployment of different EVPN operational and functional models in each data center. It also allows greater scalability within a data center in terms of intra-data center VTEP peering because each data center has its own atomic EVPN domain.

Figure 22. DCI with Separate MP-BGP EVPN Administrative Domains



Conclusion

MP-BGP EVPN changes the paradigm for the VXLAN overlay network. It introduces control-plane learning to provide a consistently signaled forwarding database in any size of network instead of relying on flooding and learning. MP-BGP EVPN is based on an industry-standard draft and a collaborative effort by multiple vendors and service providers working together to develop a simple and interoperable technology. It provides integrated bridging and routing for overlay networks for optimized delivery of traffic. With MP-BGP EVPN capabilities in Cisco NX-OS Software and VXLAN routing capabilities in Cisco Nexus 9000 Series hardware, you can use Cisco Nexus 9000 Series Switches to build highly scalable, robust, and high-performance VXLAN overlay fabric networks.

For More Information

- IETF Draft - BGP MPLS-based Ethernet VPN: https://tools.ietf.org/html/draft-ietf-l2vpn-evpn-11
- IETF Draft - Network virtualization overlay solution with EVPN: https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay-00
- IETF Draft - Integrated routing and bridging in EVPN: https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-00
- IETF Draft - IP prefix advertisement in EVPN: https://tools.ietf.org/html/draft-rabadan-l2vpn-evpn-prefix-advertisement-02
- RFC 4271 - Border Gateway Protocol 4 (BGP-4): https://tools.ietf.org/html/rfc4271

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

- RFC 4760 - Multiprotocol extensions for BGP-4: https://tools.ietf.org/html/rfc4760
- RFC 4364 - BGP/MPLS IP VPNs: https://tools.ietf.org/html/rfc4364#page-15
- VXLAN overview - Cisco Nexus 9000 Series Switches: http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html
- VXLAN design with Cisco Nexus 9300 platform switches: http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-732453.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

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

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

Printed in USA C07-734107-01 03/16

© 2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.