Deploy a VXLAN Network with an MP-BGP EVPN Control Plane
##Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>MP-BGP EVPN Control Plane: Overview</td>
<td>4</td>
</tr>
<tr>
<td>Software and Hardware Support for the MP-BGP EVPN Control Plane</td>
<td>4</td>
</tr>
<tr>
<td>IP Transport Devices Running MP-BGP EVPN</td>
<td>4</td>
</tr>
<tr>
<td>IP Transport Devices Not Running MP-BGP EVPN</td>
<td>5</td>
</tr>
<tr>
<td>VTEPs Running MP-BGP EVPN</td>
<td>5</td>
</tr>
<tr>
<td>Inter-VXLAN Routing</td>
<td>5</td>
</tr>
<tr>
<td>MP-BGP EVPN VXLAN Support on Cisco Nexus 9000 Series Switches</td>
<td>5</td>
</tr>
<tr>
<td>MP-BGP EVPN VXLAN Support on Cisco Nexus 7000 Series and 7700 Platform Switches</td>
<td>6</td>
</tr>
<tr>
<td>Multitenancy in MP-BGP EVPN</td>
<td>6</td>
</tr>
<tr>
<td>MP-BGP EVPN NLRI and L2VPN EVPN Address Family</td>
<td>6</td>
</tr>
<tr>
<td>Integrated Routing and Bridging with the MP-BGP EVPN Control Plane</td>
<td>8</td>
</tr>
<tr>
<td>Local-Host Learning</td>
<td>8</td>
</tr>
<tr>
<td>EVPN Route Advertisement and Remote-Host Learning</td>
<td>9</td>
</tr>
<tr>
<td>Symmetric and Asymmetric Integrated Routing and Bridging</td>
<td>9</td>
</tr>
<tr>
<td>VNIs for Bridge Domains and IP VRF Instances</td>
<td>12</td>
</tr>
<tr>
<td>VTEP Peer Discovery and Authentication in MP-BGP EVPN</td>
<td>13</td>
</tr>
<tr>
<td>Distributed Anycast Gateway in MP-BGP EVPN</td>
<td>15</td>
</tr>
<tr>
<td>ARP Suppression in MP-BGP EVPN</td>
<td>16</td>
</tr>
<tr>
<td>Conclusion</td>
<td>17</td>
</tr>
<tr>
<td>For More Information</td>
<td>17</td>
</tr>
</tbody>
</table>

© 2015 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-based 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 (MP-BGP) Ethernet Virtual Private Network (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 east-west and north-south 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 mode, please refer to the following documents:


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:

MP-BGP EVPN Control Plane: Overview

MP-BGP EVPN is a control protocol for VXLAN based on IETF RFC 7342. Prior to EVPN, VXLAN overlay networks operated using the flood-and-learn model. In this model, end-host information learning and VTEP discovery are both data-plane based, 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 very well known 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 reduces 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 gateway 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 can be MP Internal BGP (MP-iBGP) EVPN peers (or route reflectors [RRs]), or MP External BGP (MP-eBGP) EVPN peers. Their OS software needs to support MP-BGP EVPN network layer reachability information (NLRI) so that the OS can understand the MP-BGP EVPN updates and distribute them to other MP-BGP EVPN peers using the standards-defined constructs. The most common example of such a function is the BGP route reflector running in the spine layer.
IP Transport Devices Not Running MP-BGP EVPN

IP transport devices play a role only in the underlay network by providing underlay Layer 3 routing and possibly underlay multicast rendezvous point (RP) functions. They do not contain VTEPs, nor do they play a role within the overlay. Because these devices don’t need to provide functions specific to the EVPN address family within MP-BGP or data-plane awareness of the VXLAN encapsulation, they can be any networking devices that meet the required Layer 3 routing and multicast requirements of the VXLAN-enabled switches and routers. The most common examples of such devices are in the spine layer of the network.

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 also 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 support, the switch needs to support both the control-plane and data-plane functions. For spines that provide only the IP underlay for the VXLAN fabric, almost any switch or router that supports the chosen routing protocol and multicast traffic can be used.

Inter-VXLAN Routing

The MP-BGP EVPN control plane provides integrated routing and bridging by distributing both the Layer 2 (MAC address) and Layer 3 (IP address) reachability information for end hosts and IP prefixes 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 that lookup. This capability is referred to as the VXLAN routing function. Not all switch hardware platforms support VXLAN routing, thus 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 in Cisco® NX-OS Software Release 7.0(3)I1(1) for Cisco Nexus 9000 Series Switches. In 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 both the MP-BGP EVPN control-plane and data-plane functions starting in NX-OS Software Release 7.0(3)I1(2). In NX-OS Software Release 7.0(3)I1(1) the Cisco Nexus 9500 platform switches only support the control-plane functions.
MP-BGP EVPN VXLAN Support on Cisco Nexus 7000 Series and 7700 Platform Switches

In NX-OS 7.2(0)D1(1), the Cisco Nexus 7000 Series Switches and 7700 platform added MP-BGP VXLAN support, allowing them to participate in both the control plane and the data plane. With the Cisco Nexus F3-Series line card, full data-plane capability is possible. Without an F3-Series line card, control-plane capability is still possible. Today, the Cisco Nexus 7000 Series or 7700 platform switch can be a MP-BGP route reflector (with or without the F3-Series line card), VXLAN Layer 3 gateway, border leaf, spine, or border spine. In NX-OS 7.2, the Cisco Nexus 7000 Series or 7700 platform switch in the border leaf or border spine role can either hand off to Cisco Locator/ID Separation Protocol (LISP) or use Virtual Routing and Forwarding Lite (VRF-Lite).

NX-OS 7.2 does not support leaf functionality. Therefore, Layer 2 VXLAN Overlay features are not supported. This support will be added in a future release.

Multitenancy in MP-BGP EVPN

In an extension to the existing MP-BGP, MP-BGP EVPN inherits the support for multitenancy with VPN using the VRF construct. In MP-BGP EVPN, multiple tenants can co-exist and share a common IP transport network while having their own separate VPDNs 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.

For example, tenant A is assigned layer segment (VNIs) 10000 to 10004. All five VNIs are assigned to VRF A. Tenant B is assigned VNIs 10010 to 10014, and all VNIs are assigned to VRF B. With this assignment in place, all VNIs are routed through the same underlay network topology, but tenant A’s VNIs will not be able to communicate with tenant B’s VNIs.

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, or 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 the 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 to a specific location in a network that 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’s point of view, the VNI can span multiple noncontiguous sites, reaching beyond the Layer 2 and Layer 3 boundaries of the underlay infrastructure (Figure 1). Traffic between end hosts in the same VNI needs to be bridged in the overlay network. Therefore, 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.
**Note:** VXLAN with a MP-BGP EVPN control plane supports MAC address-only routes. The IP information is optional for route-type 2 (MAC and IP address advertisement) and can result in redundant MAC address route objects in the route information base (MAC address advertisement and MAC and IP address advertisement).

**Figure 1.** VNI across an Underlay IP Network

RFC 7342 defines a standards-based way of advertising Layer 3 host IP addresses 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 destination host. The following is an example of a MAC and IP address host advertisement in MP-BGP EVPN:

**Example Type-2 Route:**

```
Route Distinguisher: 10.1.1.13:32967   (L2VNI 30200)
*>1[2]:[0]:[0]:48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
```

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 yet to be 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.

The 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 EVPV uses route distinguishers (RDs) to maintain uniqueness among identical routes in different VRF instances, and 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 EVPV 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 EVPV must be enabled to generate or process extended community attributes.
In the 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 autonomous system (AS) VNI. Alternatively, you 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 possible 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.

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

![MP-BGP EVPN Host NLRI Learning and Distribution](image)

**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

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 MP-BGP VPN standard (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 source 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 (RFC 7432) define two integrated routing and bridging (IRB) models: asymmetric IRB and symmetric IRB. NX-OS for the Cisco Nexus Family switches 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.
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, it requires every 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 or the number of VXLAN VNIs in the overlay network increases.

Symmetric IRB

With symmetric IRB, both the ingress and egress VTEPs perform Layer 2 and Layer 3 lookups. Symmetric IRB introduces a few 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 the 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.

**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 they don’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 VLAN 200 and VNI 5200 do not need to be present in the leftmost VTEP. As a result, routing and bridging is more scalable than with asymmetric IRB. NX-OS implements symmetric IRB to achieve optimal learning and scaling. In addition, with symmetrical IRB you can choose to scope switch configurations so that they have switch virtual interfaces (SVIs) only for segments that are located behind the respective switch, thus enabling an optimized configuration.
Figure 6. VTEP VNI Membership with Symmetric IRB

Every VTEP needs to be only in VNIs for which it has local hosts.
VTEPs don’t need to maintain MAC address tables for VNIs for which they don’t have local hosts.

VTEP VNI Membership with Symmetric IRB

Figure 7. VNIs for Bridge Domain and IP VRF Instances

Layer 2 Bridging in an EVPN VXLAN Fabric

When the destination MAC address in the original packet header does not belong to the local VTEP, then VXLAN bridging needs to occur. In other words, the originating VTEP will perform a Layer 2 lookup and bridge the packet to the destination VTEP.

Layer 2 Bridging in an EVPN VXLAN Fabric

When the destination MAC address in the original packet header does not belong to the local VTEP, then VXLAN bridging needs to occur. In other words, the originating VTEP will perform a Layer 2 lookup and bridge the packet to the destination VTEP.
Layer 3 Routing in an EVPN VXLAN Fabric

If the destination MAC address matches the anycast gateway MAC address, then VXLAN routing needs to occur. In this case, the originating VTEP will perform a Layer 3 lookup and then encapsulate the packet with the Layer 3 VNI. The destination VTEP will then receive the traffic and will perform another routing lookup based on the inner IP header. The destination VTEP knows to perform this step because the destination MAC address in the packet belongs to the destination VTEP itself.

The destination VTEP address, which resides 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 this outer destination IP address. After the packet arrives at the egress VTEP, the VNI in the VXLAN header is examined to determine whether the packet needs to be bridged or 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.

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 can easily allow insertion of a rogue VTEP into a VNI segment. After a rogue VTEP has been inserted into the segment, it can 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 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 unauthorized sources. VXLAN-encapsulated traffic from these invalid VTEPs will be discarded by other VTEPs.

For 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. After a VTEP's router MAC address is distributed through 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.

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 65500
  router-id 10.1.1.101
  log-neighbor-changes
  address-family ipv4 unicast
  neighbor 10.1.1.102 remote-as 65500
      password 3 a667d47acc18ea6b
      update-source loopback0
  address-family ipv4 unicast
  address-family l2vpn evpn
  send-community both
```

- This entry 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 65500
  router-id 10.1.1.102
  log-neighbor-changes
  address-family ipv4 unicast
  neighbor 10.1.1.101 remote-as 65500
      password 3 a667d47acc18ea6b
      update-source loopback0
  address-family ipv4 unicast
  address-family l2vpn evpn
  send-community both
```
The following is a sample display of VNI peer status and information in Cisco NX-OS:

<table>
<thead>
<tr>
<th>VTEP1-1# sh nve peers</th>
<th>State</th>
<th>LearnType</th>
<th>Uptime</th>
<th>Router-Mac</th>
</tr>
</thead>
<tbody>
<tr>
<td>nve1 10.1.1.102 Up</td>
<td>CP</td>
<td>1w3d</td>
<td>6412.2574.6ae7</td>
<td></td>
</tr>
<tr>
<td>nve2 10.1.1.134 Up</td>
<td>CP</td>
<td>1w3d</td>
<td>7c69.f6df.e71f</td>
<td></td>
</tr>
</tbody>
</table>

VTEP-1# sh nve peers peer-ip 10.1.1.102 det
Details of nve Peers:

- Peer-IP: 10.1.1.102
  - VNE Interface: nve1
  - Peer State: Up
  - Peer Uptime: 1w3d
  - Router-Mac: 6412.2574.6ae7
  - Peer First VNI: 30210
  - Configured VNIs: 30200,30210,50000,50001
  - Provision State: add-complete
  - Route-Update: Yes
  - Peer Flags: DisableLearn
  - Learnt CF VNIs: 30200,30210

Distributed Anycast Gateway in MP-BGP EVPN

In MP-BGP EVPN, any VTEP participating in a VNI can use the distributed anycast gateway feature 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, an end host in a VNI can always use its local VTEP for this VNI as its default gateway to send traffic outside its 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 transparent host mobility in the VXLAN overlay network. Because the gateway IP address and virtual MAC address are identically provisioned on all VTEPs within a VNI, when an end host moves from one VTEP to another, it doesn’t need to send another ARP request to relearn 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 ARP-resolved 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 then learns the MAC address of the remote host in the ARP response. If the local VTEP doesn’t have the ARP-resolved 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 an ARP response is sent back. After the local VTEP learns about the MAC and IP addresses 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.

Because most end hosts send GARP or RARP requests to announce themselves to the network immediately after they come online, the local VTEP immediately has 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.

In addition to ARP suppression, VXLAN and EVPN provide built-in loop mitigation through a BGP dampening mechanism. When movement of a host is detected more than a set number of times within a defined time period, BGP will stop updating the associated route tables. By default, NX-OS stops updating the route table when a host moves more than three times within 10 minutes.

Figure 10. ARP Suppression in MP-BGP EVPN
Conclusion

With the addition of the MP-BGP EVPN control plane, VXLAN fabrics are now more scalable and intelligent. By introducing features such as anycast distributed gateways and symmetrical IRB into the fabric, Cisco Nexus switches provides best-in-class forwarding features while providing robust and high-performance fabric networks. Because the MP-BGP EVPN control plane is an open standard, VXLAN also helps achieve many enterprises’ goal of an open and interoperable data center.

For More Information