External Connectivity—MPLS L3VPN

External Connectivity

Introduction to External Connectivity

Before you begin—Ensure you know about Programmable Fabric. Conceptual information is covered in the *Introduction to Cisco Programmable Fabric* and *Introducing Cisco Programmable Fabric (VXLAN/EVPN)* chapters. *Forwarding Configurations* chapter contains Virtual Extensible LAN (VXLAN) BGP EVPN fabric configurations, and the *IP Fabric Underlay* chapter contains unicast and multicast protocol configuration information for the fabric underlay.

VXLAN BGP EVPN fabric supports external connectivity, in that VXLAN BGP EVPN fabric data centers in different sites can be connected using the Data Center Interconnect functionality. Also, VXLAN and non VXLAN pods within a single site can be connected.

You can enable a VXLAN BGP EVPN fabric border leaf switch and an Autonomous System Boundary Router (ASBR) at the edge of the WAN to exchange reachability routes. This way, reachability information is transported between sites. A route exchange scenario at the border leaf or border spine switch is referred as a *handoff* scenario.

Note In this chapter, the focus is on the border leaf switch that acts as a borderPE switch (a border leaf switch with integrated MPLS PE functionality) instead of a border spine switch.

The Cisco Nexus 7000 Series switch is the primary border leaf platform for connecting a VXLAN BGP EVPN fabric to external entities since this switch, with F3 and M3 line cards, acts as the borderPE switch, and provides the combined functionality of a border leaf switch and datacenter edge device. Also, the Cisco Nexus 7000 Series switch provides higher route and VRF scaling capabilities that are required for a border leaf switch.

VM mobility across datacenters

VM mobility across VXLAN BGP EVPN datacenter fabrics works the same way as it does within the datacenter fabric. When VM mobility takes place, the VM generates RARP and GARP messages. You should enable a

Layer-2 DCI such as OTV, Classical Ethernet or VPLS to transport broadcast RARP and GARP packets generated due to the VM movement.

Note Additional configuration is not required to support VM movement across fabrics.

VM mobility across fabrics cannot take place in these scenarios:

1. VM movement takes place when MAC chaining (multiple IP addresses mapped to the same MAC address) is in effect.
2. When an end host sends a non broadcast packet such as ARP on VM move.

Points to consider for connecting two data centers:

- If you want to connect two data center fabrics, restrict the overlay within a data center. Connect two data center instances with an inter data center instance. This way, any instability in one data center will not be spread to the other. Failures can be contained since we are separating the administrative domains.
  - For example, two VXLAN fabrics in different sites are two separate domains, and the MPLS L3VPN inter data center instance forms a different, connecting domain. Here, traffic from the source data center terminates at the border leaf or borderPE switch, and a new Layer-3 inter data center instance sends the traffic to the border leaf switch or borderPE switch of the target datacenter.
- You can create a Layer-2 (OTV, etc) or Layer-3 (LISP, Multiprotocol Label Switching [MPLS] L3VPN, etc) inter datacenter instance.

Layer-3 handoff scenario using MPLS L3VPN is explained in this chapter.

Important Border leaf switch VDCs that have M3 modules do not support LISP handoff scenarios.

Layer 3 hand off scenario – MPLS L3VPN

The VXLAN BGP EVPN data center fabric can be connected across Layer-3 boundaries (to external sites and back) using MPLS L3VPN, VRF IP Routing (VRF Lite), or LISP as the mechanism of transport outside the VXLAN fabric.

Cisco Nexus 7000 Series is considered as the border leaf switch in the MPLS L3VPN scenario explained below. For the MPLS handoff scenario, the border leaf switch is referred to as a borderPE switch. In other words, a border leaf switch with integrated MPLS PE functionality is used.

Note *borderPE switch*—A Cisco Nexus 7000 Series switch with F3/M3 line cards; acts as the collapsed border leaf switch and MPLS Datacenter Provider Edge router. A BorderPE switch includes MPLS PE function. This is also referred to as a one box solution since a single device provides the combined functionality of a border leaf switch and datacenter Edge device.

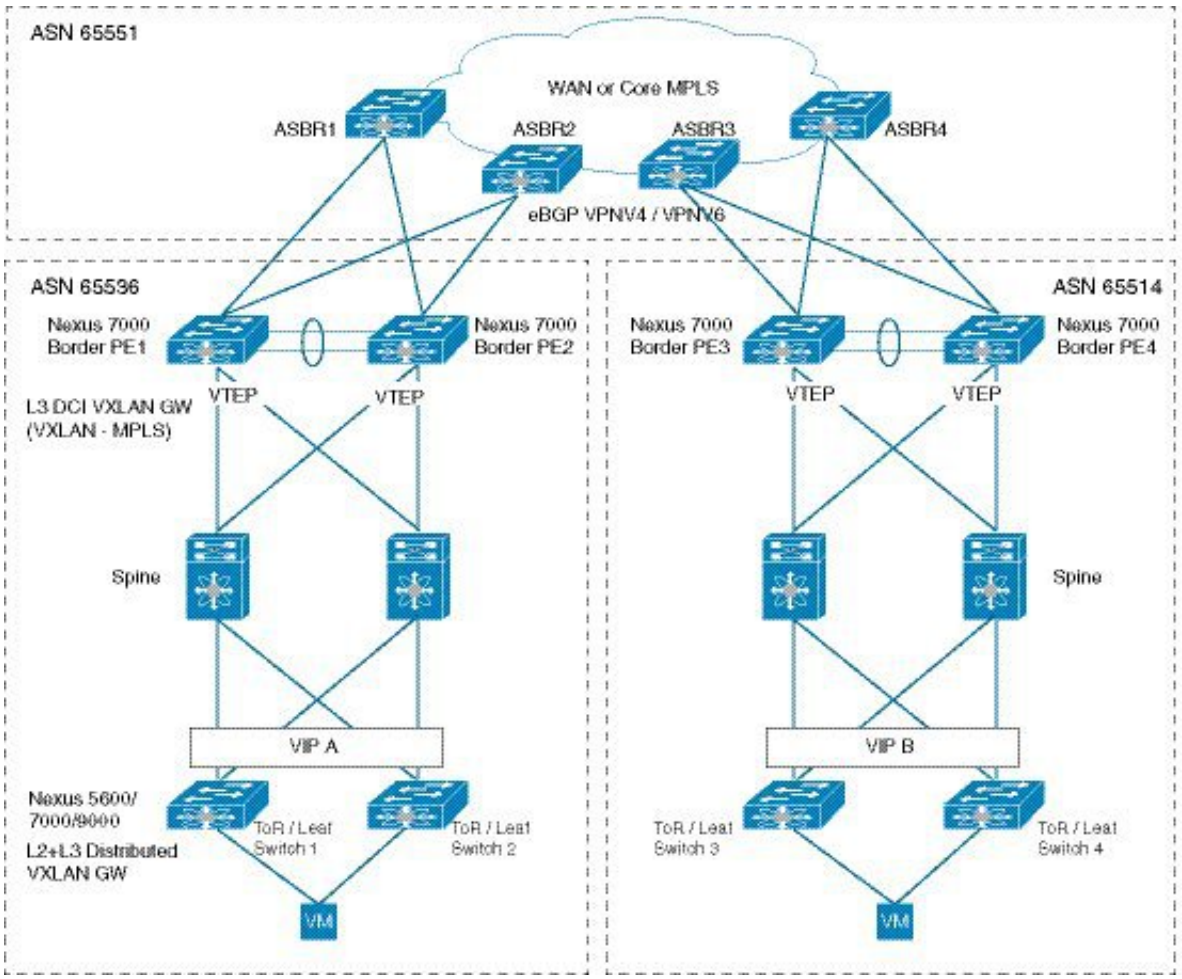
VXLAN BGP EVPN - MPLS L3VPN DCI scenario – In brief

- Two VXLAN BGP EVPN fabric data centers are depicted at the left and the right of the image (below). Routes within a fabric pod are shared between all VTEPs, including with the Cisco Nexus 7000 Series borderPE switches (that are in a vPC setup).
- The borderPE switches and the connected Autonomous System Boundary Routers (ASBRs) of the WAN are configured to pass on routes between each other. For example, routes within the VXLAN BGP EVPN fabric (left) are sent to the WAN ASBR, and (necessary) reachability routes within the WAN are sent to the borderPE switch on the VXLAN fabric (left).
- The WAN ASBR and the VXLAN fabric (right) are also connected in the same way.
- The WAN ASBRs act as MPLS PE nodes that belong to the WAN autonomous system number (ASN), and this ASN is different than that of the 2 VXLAN BGP EVPN fabrics.

As a result, the data center pods depicted in the left and right sides of the image are connected through the WAN using MPLS L3VPN.

Note In addition to connecting data center fabrics, these Layer-3 solutions (MPLS L3VPN, VRF Lite, and LISP) are also used to connect campus networks.

Figure 1: DCI using MPLS L3VPN



VXLAN BGP EVPN - MPLS L3VPN DCI scenario - In more detail

Route distribution and data flow between the VXLAN pod and the WAN MPLS is explained below.

Route distribution within the VXLAN pod, and subsequent export of VXLAN pod routes to the WAN MPLS

The routes within the VXLAN BGP EVPN pod should be exported to the WAN MPLS, thereby extending Layer-3 reachability from the WAN MPLS to the ToRs within the VXLAN BGP EVPN fabric.

- The BGP EVPN control plane in the VXLAN BGP EVPN fabric ensures distribution of routes between ToR/leaf switch VTEPs (including the borderPE switch) within the fabric. ToRs will forward the attached end host IP and MAC addresses, /32 (or /128, for IPv6 addresses) ‘Host IP + MAC’ routes, and Layer-2 and Layer-3 VXLAN VNIs using the EVPN Route Type 2/5 option. Based on the import route target (RT) configured on the border leaf switch, the switch will import the /32 (or /128) routes into appropriate VRF tables.
- The borderPE switch advertises the VRF default route to the ToR/leaf switches. If the ToR/leaf switch nodes receive the same route from multiple borderPE switches, it results in ECMP at the ingress ToR/leaf switch.

- The borderPE switch should be configured to *reoriginate* the /32 (or /128) routes on an eBGP VPNv4 session towards the WAN MPLS. The borderPE switch does MPLS switching towards the WAN MPLS and VXLAN BGP EVPN routes are reoriginated into the L3VPN address family.

VXLAN fabric to WAN Data flow—Let us say a host in the VXLAN fabric (left) sends traffic to a host in the other VXLAN fabric (right). A high level data plane flow is depicted below:

- The VXLAN packet reaches the borderPE switch.
- A VXLAN VNI lookup happens. This lookup maps to the appropriate bridge domain.
- The subsequent MAC address lookup points to a bridge domain interface on the borderPE switch, and the packet is bridged to the BDI interface. The BDI interface then points to the corresponding VRF table.

Note The BDI is used for VRF routing and has no IP.

- The destination IP address is checked in the VRF IP table, wherein the corresponding L3VPN MPLS encapsulation towards the WAN MPLS is pointed at.
- The packet is sent to the WAN ASBR.

Importing of WAN or external routes into the VXLAN BGP EVPN borderPE switch

After WAN or external routes are imported into the borderPE switch, they are re-advertised to the ToR/leaf switches in the VXLAN BGP EVPN fabric with the borderPE VTEP acting as the next hop for the ToR/leaf switches, thereby extending Layer-3 reachability from the ToR/leaf switches to the WAN.

- The borderPE (Cisco Nexus 7000 Series) switch will function as an EVPN based Layer-3 VXLAN Gateway to provide an overlay routing function for Layer-3 IP traffic between the VXLAN overlay fabric and the WAN. This includes Layer-3 flows between the VPNv4/6 and/or IP end points outside the VXLAN pod (like hosts in the other data center, etc).
- The borderPE switch receives routes from L3VPN ASBRs to enable connectivity to the WAN. This is achieved through an external BGP VPNv4 or VPNv6 session with the WAN ASBR device. The L3VPN routes from the WAN will be imported into local VRF tables in the borderPE switch.
- The borderPE switch should be configured to re-originate the L3VPN routes into the EVPN address family so as to send the routes via BGP EVPN control plane from the borderPE switch towards the ToR/leaf switches. The ToR/leaf switches will import these routes into the appropriate VRF.
- Necessary configuration knobs will need to be added in BGP under **VRF**, and under **neighbor evpn** address family to originate a default route towards EVPN neighbors and drop all other routes.

WAN to VXLAN BGP EVPN fabric Data flow—Traffic from the WAN ASBR arrives at the borderPE switch. A high level flow is depicted below:

- The packet arrives with a local, per VRF VPN label (advertised earlier by the borderPE switch). The MPLS label lookup points to the correct VRF table.
- The (VRF, IP) lookup results in a /32 (or /128) route that points to the VXLAN tunnel end point adjacency for the ToR/leaf switch VTEP on the fabric facing per VRF BDI interface.
- The packet is VXLAN encapsulated with the router MAC (RMAC) address of the remote ToR/leaf switch VTEP and sent on the VRF BDI interface.

- The FIB lookup drives the VXLAN encapsulation towards the designated remote ToR/leaf switch VTEP.

The fabric is stitched to the VPN service

- Once the tenant flows within the 2 data centers are stitched to the IP VPN service, routes from data center (left) are connected to data center (right) and vice versa.

BorderPE switches in a vPC setup

The two borderPE switches are configured as a vPC. In a VXLAN vPC deployment, a common, virtual VTEP IP address (secondary loopback IP address) is used for communication. The common, virtual VTEP uses a system specific router MAC address. The Layer-3 prefixes or default route from the borderPE switch will be advertised with this common virtual VTEP as the next hop, and the ToR/leaf switch VTEPs will install the default route and/or prefix route with a single BGP path to the borderPE common, virtual VTEP IP address.

Advertising primary IP address (PIP)

On a vPC enabled leaf or border leaf switch, by default all Layer-3 routes are advertised with the secondary IP address (VIP) of the leaf switch VTEP as the BGP next-hop IP address. Prefix routes and leaf switch generated routes are not synced between vPC leaf switches. Using the VIP as the BGP next-hop for these types of routes can cause traffic to be forwarded to the wrong vPC leaf or border leaf switch and black-holed. The provision to use the primary IP address (PIP) as the next-hop when advertising prefix routes or loopback interface routes in BGP on vPC enabled leaf or border leaf switches allows users to select the PIP as BGP next-hop when advertising these types of routes, so that traffic will always be forwarded to the right vPC enabled leaf or border leaf switch.

The configuration command for advertising the PIP is **advertise-pip**.

A sample configuration is given below:

(config) #

```
router bgp 65536
  address-family l2vpn evpn
      advertise-pip
      advertise-system-mac
```

The **advertise-pip** command lets BGP use the PIP as next-hop when advertising prefix routes or leaf generated routes if vPC is enabled. The **advertise-system-mac** command lets BGP advertise Route-type-2 routes that includes VIP and router-mac information. This is needed for solving an issue in EVPN decapsulation on remote leaf switches when PIP is used as next-hop in the BGP advertisement.

BorderPE switch to WAN ASBR link failure scenario

If there is link failure between one of the two borderPE switches and the connected L3VPN ASBR, the switch will withdraw the BGP routes that are being advertised towards the fabric and traffic re-convergence happens through the redundant border leaf switch.

For border leaf switches (in a two box solution), the default route will be withdrawn when both links to the DC Edge router fail. For a borderPE switch, advertising default route is not recommended.

Configuration for the VXLAN BGP EVPN - MPLS L3VPN DCI scenario

The following configurations enabled on the borderPE switch establish a Layer-3 link along with the MPLS/LDP configuration between the borderPE switch and the WAN ASBR. After configurations on the borderPE switch and the WAN ASBR device, routes are exchanged between the VXLAN fabric borderPE switch and the MPLS WAN ASBR.

Important Ensure that you follow these implementation pointers:

- This document only contains Cisco Nexus 7000 Series borderPE switch related configurations. To complete Layer-3 DCI configurations, you should also enable corresponding configurations on the WAN ASBR.
- To forward traffic across the borderPE switch (from the VXLAN BGP EVPN fabric towards the WAN ASBR and the other way round), the **fabric forwarding switch-role border** command should be mandatorily configured on the borderPE switch. Since the change of switch role requires a switch reload (through **write erase** and **reload** commands), ensure that this command is included in the startup configuration. For the borderPE Layer-3 extension auto configuration feature, use the **fabric forwarding switch-role border dci-node** command.
- The physical interface connecting the VXLAN BGP EVPN fabric should be different from the IP/MPLS WAN facing interface. The same physical Layer-3 interface or sub interface should not be used to connect the VXLAN BGP EVPN Fabric and the WAN Core.
- For an F3 only VDC, the Layer-3 backup link that is used to protect the WAN facing interfaces from failure should not be the peer link or SVI of the VLANs extended over the peer link. *A separate Layer-2 interface with a dedicated VLAN/SVI, or a separate Layer-3 interface or sub interface should be used*.
- For an M3-F3 VDC, traffic received from the VXLAN BGP EVPN fabric should not be forwarded to a Layer-3 sub interface. VXLAN terminated traffic can only be forwarded over a Layer-3 physical interface.
- Host Mobility Manager (HMM) CLIs are removed with the **no feature fab forwarding** command even though the **nv overlay evpn** command is present.

  The **feature fabric forwarding** command is not needed if **nv overlay evpn** is already configured.

On BorderPE1 switch, enable VXLAN BGP EVPN features

(config) #

```
install feature-set fabric
feature-set fabric
feature fabric forwarding
feature interface-vlan
feature ospf  (OR feature isis)
feature nv overlay
feature bgp
feature vni
nv overlay evpn
install feature-set mpls
feature-set mpls
feature mpls l3vpn
feature mpls ldp
```

The VXLAN feature related configurations shown above are already enabled on all switches in the VXLAN BGP EVPN fabric. This has been included here included for completeness only.

Note The **install feature-set fabric** command should only be used in the admin VDC. When using a VDC, ensure the VDC is of type F3 or M3, for EVPN. A sample configuration is given below:

(config) #

```
vdc BorderPE1
   limit-resource module-type f3
```

Configure the anycast gateway MAC address

(config) #

```
fabric forwarding anycast-gateway-mac 0202.0002.0002
```

On BorderPE1, configure a bridge domain and associate a Layer-3 network VNI

(config) #

```
fabric forwarding switch-role border
system bridge-domain 2500-3500
system fabric bridge-domain 2500-2999
vni 31000
bridge-domain 2500
   member vni 31000
```

On BorderPE1, create a VRF and associate the previously configured VNI to it. Then, enable importing and exporting of routes between the VXLAN BGP EVPN fabric and the MPLS WAN side, and create the default routes to be injected into the VXLAN fabric

(config) #

```
vrf context vni-31000
  vni 31000
  rd auto
  address-family ipv4 unicast
      route-target import 65551:1
      route-target export 65551:1
      route-target both auto
      route-target both auto evpn
```

Type **exit** and configure IPv6 route import/export.

```
  address-family ipv6 unicast
     route-target import 65551:1
     route-target export 65551:1
     route-target both auto
     route-target both auto evpn
```

65551:1 refers to the import/export of WAN routes.

Within the VXLAN BGP EVPN fabric, NX-OS automatically assigns the correct route target. It is recommended that the commands **route-target both auto** and **route-target both auto evpn** are used on the ToR/Leaf switches too.

Note The route target *65551:1* should be the same as the route target configured on the connected WAN ASBR since the importing/exporting of MPLS information is based on this route target

By using route targets as the glue, the BGP EVPN control plane (in the VXLAN fabric) and the BGP L3VPN control plane (from the fabric to the WAN) are connected. Similarly, for a tenant VRF, the same route target should be enabled on ToR/leaf switches and the border leaf switch(es).

On BorderPE1, configure a bridge domain and BDI for the VRF

(config) #

```
interface BDI 2500
   no shutdown
   mtu 9192
   vrf member vni-31000
   no ip redirects
   ip forward
   ipv6 address use-link-local-only
   ipv6 forward
   no ipv6 redirects
```

After the above configuration, we ensure that a bridge domain interface is designated for Layer-3 traffic transportation.

Important The **interface BDI** configuration is not required to be configured manually, when the profile vrf-tenant-profile is configured. As soon the VRF context is configured, NX-OS automatically calls the profile vrf-tenant-profile and applies appropriate configurations. vrf-tenant-profile is always added when you deploy the BorderPE switch with POAP from DCNM. The resulting interface BDI configuration can be verified with the **show run inter bdi 2500 expand-port-profile** command.

The vrf-tenant-profile configuration is given below:

```
configure profile vrf-tenant-profile
vni $vrfSegmentId
bridge-domain $bridgeDomainId
member vni $vrfSegmentId
interface bdi $bridgeDomainId
vrf member $vrfName
ip forward
no ip redirects
ipv6 forward
ipv6 address use-link-local-only
no ipv6 redirects
mtu 9192
no shutdown
```

On BorderPE1, add the Layer-3 VRF VNI to the overlay

(config) #

```
interface nve 1
   no shutdown
   source-interface loopback 1
   host-reachability protocol bgp
   member vni 31000 associate-vrf
```

In the above configuration, we ensure that the Layer-3 VNI is associated with the VXLAN VTEP.

On BorderPE1, establish a multihop external BGP session to the WAN ASBR and enable forwarding of L2VPN routes towards the WAN ASBR

(config) #

```
router bgp 65536
   neighbor 209.165.200.225 remote-as 65551
      update-source loopback 100 -> Optional
      ebgp-multihop 10 -> Optional
      address-family vpnv4 unicast
        send-community both
        import l2vpn evpn reoriginate
```

If the BGP session to the WAN ASBR is on a directly connected interface, and the peering is done on the interface address, then the **ebgp-multihop** and **update-source** commands are not required. Also, configure the VPNv6 address family as shown below:

```
     address-family vpnv6 unicast
        send-community both
        import l2vpn evpn reoriginate
```

Attention In the above configurations, the L2VPN EVPN information is being imported into the VPNv4/VPNv6 address families so that the routes in the VXLAN BGP EVPN fabric can be sent over VPNv4/VPNv6 to the connected WAN ASBR. When the WAN ASBR sends routes to the border leaf switch, the received VPNv4/VPNv6 routes need to be sent into the VXLAN BGP EVPN control plane. To achieve that, the VPNv4/VPNv6 information (L3VPN routes) is imported into the L2VPN EVPN address family, as shown below.

Configure the BGP EVPN neighbor within the fabric.

(config) #

```
router bgp 65536
   neighbor 10.2.2.1 remote-as 65536
      update-source loopback 0
      address-family l2vpn evpn
         send-community both
         import vpn unicast reoriginate
```

On BorderPE1, create the VRF under the BGP configuration to advertise the L2VPN EVPN address family (routes) within the VRF

(config) #

```
router bgp 65536
  vrf vni-31000
     address-family ipv4 unicast
       advertise l2vpn evpn
       maximum-paths 2
       label-allocation-mode per-vrf
       exit
     address-family ipv6 unicast
       advertise l2vpn evpn
       maximum-paths 2
       label-allocation-mode per-vrf
```

Note Alternatively, you can enable a 0/0 default route origination in each tenant VRF (on the border leaf switch). The ToRs/leaf switches will import the default route, resulting in a VRF default route towards the border leaf switch in the ToR switch tenant VRFs.

Creation of default route and route maps on the borderPE1 switch

Enable default route origination in each VRF, for IP4 and IPv6 address families

(config) #

```
vrf context vni-31000
  ip route 0.0.0.0/0 null 0 254
  ipv6 route 0::/0 null 0 254
```

The ToR switches will import this default route, resulting in a VRF default route for tenant VRF *vni-31000* on the ToR switches.

Create a route-map to ensure that the default route from the WAN ASBR is preferred over a default route from other sources

(config) #

```
route-map PREFER-EXTERNAL-DEFAULT permit 100
   set local-preference 50
```

Apply the route map to tenant VRF vni-31000 address families (IP4 and IPv6 unicast)

(config) #

```
vrf vni-31000
     address-family ipv4 unicast
       network 0.0.0.0/0 evpn route-map PREFER-EXTERNAL-DEFAULT
       exit
     address-family ipv6 unicast
       network 0.0.0.0/0 evpn route-map PREFER-EXTERNAL-DEFAULT
```

Restrict default routes generated or learned on the BorderPE switch from being distributed to the external WAN ASBR

(config) #

```
ip prefix-list default-route seq 5 permit 0.0.0.0/0 le 1
ipv6 prefix-list default-route-v6 seq 5 permit 0::0/0
route-map DENY-DEFAULT-ROUTE deny 10
   match ip address prefix-list default-route
   exit
route-map DENY-DEFAULT-ROUTE permit 1000
exit
route-map DENY-DEFAULT-ROUTE-v6 deny 100
  match ipv6 address prefix-list default-route-v6
  exit
route-map DENY-DEFAULT-ROUTE-v6 permit 1000
```

BGP specific default route configurations

Restrict default routes generated in the fabric from being distributed to external neighbors through BGP

(config) #

```
router bgp 65536
   neighbor 209.165.200.225 remote-as 65551
      address-family vpnv4 unicast
         route-map DENY-DEFAULT-ROUTE out
         exit
      address-family vpnv6 unicast
         route-map DENY-DEFAULT-ROUTE-v6 out
```

After the above configuration, we ensure that default routes will not be included in the VPNv4 and VPNv6 routes sent to the WAN ASBR.

Glossary

RD—Route Distinguisher

RT—Route Target

RR—BGP route reflector. A route reflector reflects incoming routes to all other leaf switch nodes. Typically, spine switches are configured as route reflectors.

BorderPE switch—A Cisco Nexus 7000 Series switch with an F3 or M3 line card; acts as the collapsed border leaf switch and Cisco Nexus 7000 Series data center edge switch. A BorderPE switch includes the MPLS PE function. This is also referred to as a one box solution.

A two box solution comprises of two switches (A Cisco Nexus 5600 Series or 7000 Series border leaf switch + a Cisco Nexus 7000 Series data center edge switch) to route IP frames from an external network into the VXLAN BGP EVPN fabric