White Paper

Cisco VXLAN EVPN Multifabric Design

Contents

Introduction **3**

Option 1: VXLAN EVPN Multipod Fabric **3**

Option 2: VXLAN EVPN Multifabric **4**

VXLAN EVPN Multifabric Deployment Considerations **9**

VXLAN EVPN Multifabric with External Active-Active Gateways **10**

VXLAN EVPN Multifabric with Distributed Anycast Layer 3 Gateway **12**

Learning Process for Endpoint Reachability Information 14
Intrasubnet Communication across Fabrics 14
Intersubnet Communication across Fabrics 18
Host Mobility across Fabrics 23
Ingress and Egress Traffic-Path Optimization 26
Ingress Traffic-Path Optimization 28
Egress Traffic-Path Optimization 31
Keeping Interfabric Routing Through Layer 3 DCI 33

Conclusion **34**

For More Information **34**

Introduction

Recently, fabric architecture has become a common and popular design option for building new-generation data center networks. Virtual Extensible LAN (VXLAN) with Multiprotocol Border Gateway Protocol (MP-BGP) Ethernet VPN (EVPN) is essentially becoming the standard technology used for deploying network virtualization overlays in data center fabrics.

Data center networks usually require the interconnection of separate network fabrics, which may also be deployed across geographically dispersed sites. Consequently, organizations need to consider the possible deployment alternatives for extending Layer 2 and 3 connectivity between these fabrics and the differences between them.

The main goal of this document is to present one of these deployment options, using Layer 2 and 3 data center interconnect (DCI) technology between two or more independent VXLAN EVPN fabrics. This design is usually referred to as VXLAN multifabric.

To best understand the design presented in this document, the reader should be familiar with VXLAN EVPN, including the way it works in conjunction with the Layer 3 anycast gateway and its design for operation in a single site. For more information, see the following Cisco® VXLAN EVPN documents:

- VXLAN Overview: Cisco Nexus 9000 Series Switches
- VXLAN Network with MP-BGP EVPN Control-Plane Design Guide
- VXLAN Design with Cisco Nexus 9300 Platform Switches
- VXLAN Configuration Guide

To extend Layer 2 as well as Layer 3 segmentation across multiple geographically dispersed VXLAN-based fabrics, you can consider two main architectural approaches: VXLAN EVPN multipod fabric and VXLAN EVPN multifabric.

Option 1: VXLAN EVPN Multipod Fabric

You can create a single logical VXLAN EVPN fabric in which multiple pods are dispersed to different locations using a Layer 3 underlay to interconnect the pods. Also known as stretched fabric, this option is now more commonly called VXLAN EVPN multipod fabric. The main benefit of this design is that the same architecture model can be used to interconnect multiple pods within a building or campus area as well as across metropolitan-area distances.

The VXLAN EVPN multipod design is thoroughly discussed in the document at http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-737201.html.

Option 2: VXLAN EVPN Multifabric

You can also interconnect multiple VXLAN EVPN fabrics using a DCI solution that provides multitenant Layer 2 and Layer 3 connectivity services across fabrics. This document focuses on this option.

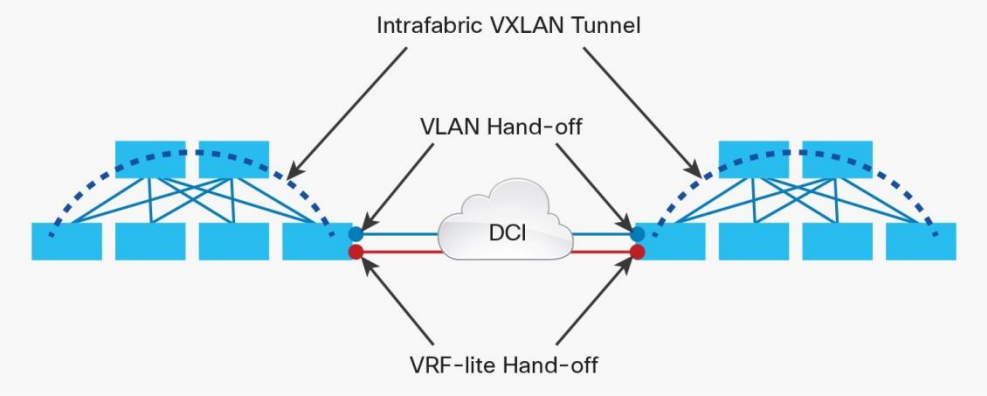
Compared to the VXLAN EVPN multipod fabric design, the VXLAN EVPN multifabric architecture offers greater independence of the data center fabrics. Reachability information for VXLAN tunnel endpoints (VTEPs) is contained in each single VXLAN EVPN fabric. As a result, VXLAN tunnels are initiated and terminated within the same fabric (in the same location). This approach helps ensure that policies, such as storm control, are applied specifically at the Layer 2 interface level between each fabric and the DCI network to limit or eliminate storm and fault propagation across fabrics.

The multifabric deployment model also offers greater scalability than the multipod approach, supporting a greater total number of leaf nodes across fabrics. The total number of leaf nodes supported equals the maximum number of leaf nodes supported in a fabric (256 at the time of the writing of this document) multiplied by the number of interconnected fabrics. A multifabric deployment can also support an overall greater number of endpoints than a multipod deployment; commonly, only a subset of Layer 2 segments is extended across separate fabrics, so most MAC address entries can be contained within each fabric. Additionally, host route advertisement can be filtered across separate fabrics for the IP subnets that are defined only locally within each fabric, increasing the overall number of supported IP addresses.

Separate VXLAN fabrics can be interconnected using Layer 2 and 3 functions, as shown in Figure 1:

- VLAN hand-off to DCI for Layer 2 extension
- Virtual Routing and Forwarding Lite (VRF-Lite) hand-off to DCI for multitenant Layer 3 extension

Figure 1. VLAN and VRF-Lite Hand-off



The maximum distance between separate VXLAN EVPN fabrics is determined mainly by the application software framework requirements (maximum tolerated latency between two active members) or by the mode of disaster recovery required by the enterprise (hot, warm, or cold migration).

After Layer 2 and 3 traffic is sent out of each fabric through the border nodes, several DCI solutions are available for extending Layer 2 and 3 connectivity across fabrics while maintaining end-to-end logical isolation.

- Layer 2 DCI: Layer 2 connectivity can be provided in several ways. It can be provided through a dedicated Layer 2 dual-sided virtual port channel (vPC) for metropolitan-area distances using fiber or dense wavelength-division multiplexing (DWDM) links between two sites. For any distance, it can be provided any valid Layer 2 overlay technology over a Layer 3 transport mechanism such as Overlay Transport Virtualization (OTV), Multiprotocol Label Switching (MPLS) EVPN, VXLAN EVPN, or Virtual Private LAN Service (VPLS).
- Layer 3 DCI: The Layer 3 DCI connection has two main purposes:
  - It advertises between sites the network prefixes for local hosts and IP subnets.
  - It propagates to the remote sites host routes and subnet prefixes for stretched IP subnets.

The DCI solutions selected for extending multitenant Layer 2 and 3 connectivity across VXLAN EVPN fabrics usually depend on the type of service available in the transport network connecting the fabrics:

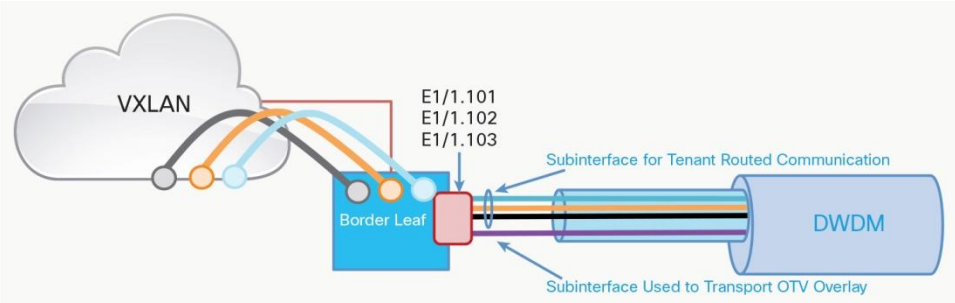
- Scenario 1: Enterprise-owned direct links (dark fibers or DWDM circuits)
  - vPC or OTV for Layer 2 extension
  - Back-to-back VRF-Lite subinterfaces for Layer 3 extension
- Scenario 2: Enterprise-owned or service provider–managed multitenant Layer 3 WAN service
  - OTV for Layer 2 extension through a dedicated VRF-Lite subinterface
  - VRF-Lite subinterfaces to each WAN Layer 3 VPN service
- Scenario 3: Enterprise-owned or service provider–managed single-tenant Layer 3 WAN service
  - OTV for Layer 2 extension across native Layer 3
  - VRF-Lite over OTV for Layer 3 segmentation

Note: In this document, OTV is the DCI solution of choice used to extend Layer 2 connectivity between VXLAN fabrics. OTV is an IP-based DCI technology designed purposely to provide Layer 2 extension capabilities over any transport infrastructure. OTV provides an overlay that enables Layer 2 connectivity between separate Layer 2 domains while keeping these domains independent and preserving the fault-isolation, resiliency, and load-balancing benefits of an IP-based interconnection. For more information about OTV as a DCI technology and associated deployment considerations, refer to the document at http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Data_Center/DCI/whitepaper/DCI3_OTV_Intro/DCI_1.html.

Scenario 1: In a metropolitan area network (MAN), you can use a direct DWDM transport for Layer 2 and 3 extension. Therefore, dedicated links for Layer 2 and 3 segmentation are established using different DWDM circuits (or direct fiber when the network is owned by the enterprise). Figure 2 shows two separate physical interfaces. One is used for Layer 3 segmentation, and the other is used for Layer 2 extension (classic Ethernet or overlay). A dedicated Layer 3 connection in conjunction with vPC is a best practice, and it is usually required for platforms that do not support dynamic routing over vPC. However, using border node platforms (such as Cisco Nexus® 7000 or 9000 Series Switches) and software releases, you also can use Layer 3 over a vPC dual-sided connection (not covered in this document).

Note: Dynamic routing over vPC is supported on Cisco Nexus 9000 Series Switches starting from Cisco NX-OS Software Release 7.0(3)I5(1).

Figure 2. Layer 2 and 3 Segmented across a DWDM Network



Nonetheless, the recommended approach is to deploy an overlay solution to provide Layer 2 DCI services such as OTV, especially if more than two sites need to be interconnected (a scenario not documented here). OTV inherently offers multipoint Layer 2 DCI services while helping ensure robust protection against the creation of end-to-end Layer 2 loops.

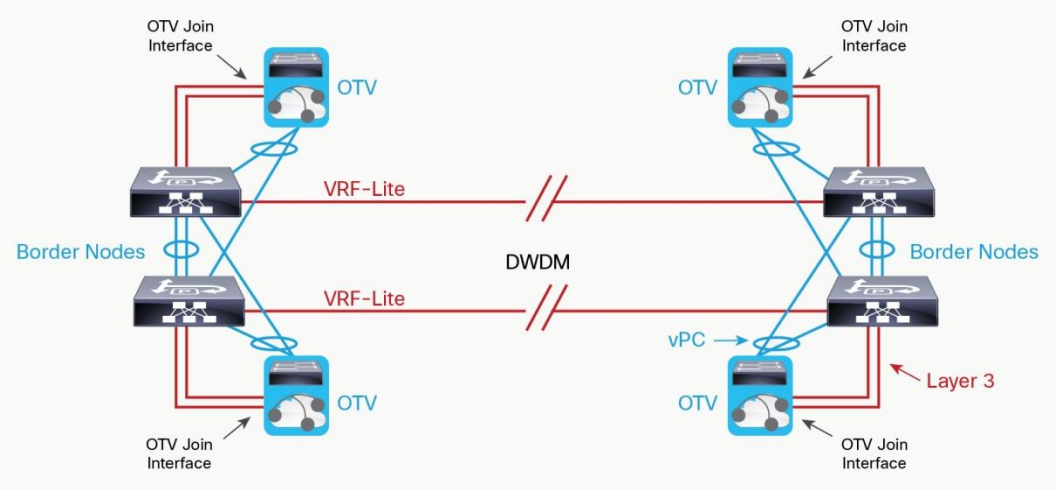
Note: At this stage, the OTV encapsulation can be built with a generic routing encapsulation (GRE) header or with a VXLAN and User Datagram Protocol (UDP) header. The latter option is preferred because it allows better load balancing of OTV traffic across the Layer 3 network that interconnects the VXLAN fabrics. However, it requires the use of F3 or M3 line cards and NX-OS Release 7.2 or later.

- OTV edge devices are connected to the border nodes with a Layer 3 and a L2 connection as shown in Figure 3, a deployment model referred to as “on a stick”. The OTV overlay is carried within a dedicated Layer 3 subinterface and VRF instance. This subinterface is carved on the same physical interface that is used to provide cross-fabric Layer 3 connectivity for all the tenants (discussed later in this section) and can belong to the default VRF instance or to a dedicated OTV VRF instance.

Note: Dual-sided vPC can be used as a Layer 2 DCI alternative for VXLAN dual-fabric deployments, though it is less preferable due to the lack of various failure isolation functions natively offered by OTV.

- Dedicated back-to-back subinterfaces carry OTV encapsulated traffic and provide Layer 3 Tenant connectivity.
  - OTV: Default VRF = E1/1.999
  - Tenant 1: VRF T1 = E1/1.101
  - Tenant 2: VRF T2 = E1/1.102
  - Tenant 3: VRF T3 = E1/1.103
  - Etc.

Figure 3. Physical View with OTV "on a Stick" to Carry Intrasubnet Communications



Scenario 2: The second scenario applies to deployments in which a Layer 3 multitenant transport service is available across separate fabrics.

This deployment model is similar to the one in scenario 1: OTV edge devices "on a stick" are used to extend Layer 2 connectivity across sites. The OTV traffic is sent out of the border nodes on dedicated subinterfaces that are mapped to a specific Layer 3 VPN service on the WAN edge router acting as a provider edge device.

At the same time, independent Layer 3 subinterfaces are deployed for each tenant for end-to-end Layer 3 communication. Each subinterface is then mapped to a dedicated Layer 3 VPN service.

Figures 4 and 5 show this scenario.

Figure 4. OTV Overlay "on a Stick" and VRF-Lite Sessions Run Independently

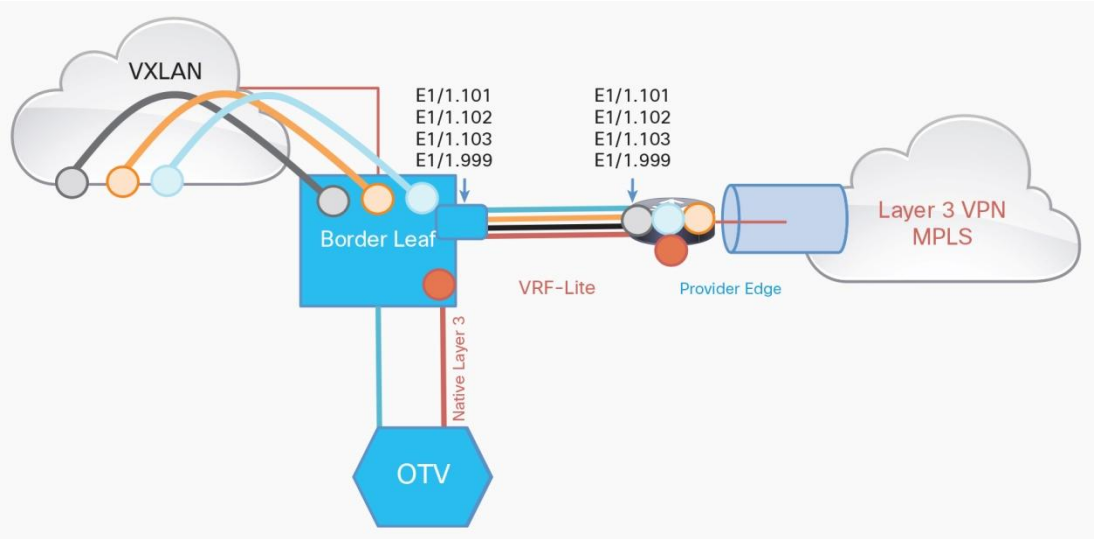
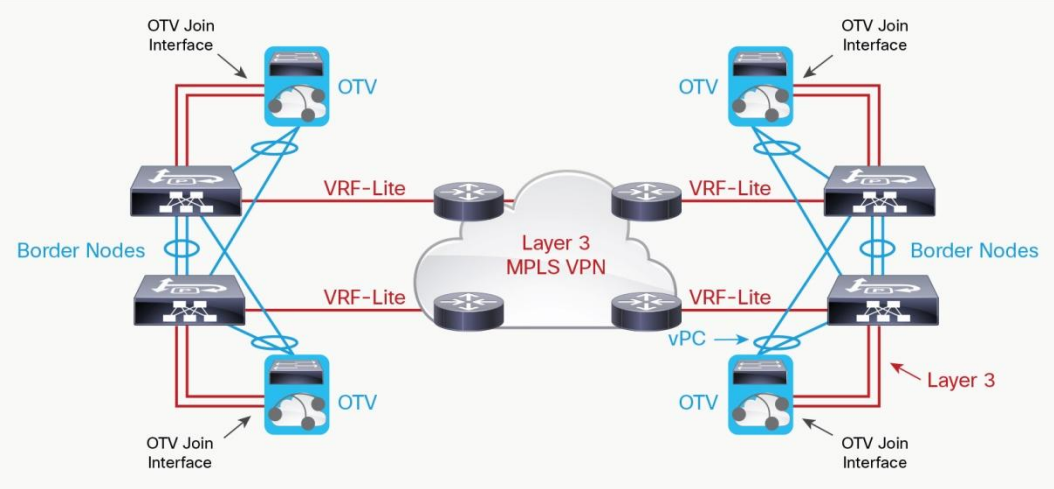


Figure 5. Physical View with OTV "on a Stick" to Carry Intrasubnet Communications: Layer 3 VPN and MPLS Core

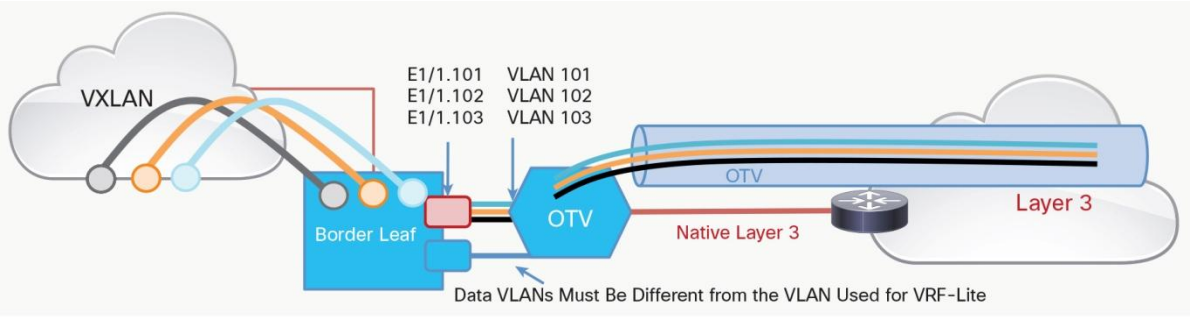


Scenario 3: In this scenario, the WAN and MAN transport network offers only single-tenant Layer 3 service (it does not support MPLS or VRF-Lite).

In this case, the Layer 2 DCI services offered by an overlay technology such as OTV can be used to establish cross-fabric Layer 2 and 3 multitenant connectivity. This connection is achieved in two steps:

- First, a Layer 2 DCI overlay service is established over the native Layer 3 WAN and MAN core.
- Next, per-tenant Layer 3 peerings are established across the fabrics over this Layer 2 overlay transport. The dedicated tenant Layer 3 interfaces to connect to the remote sites can be deployed as Layer 3 subinterfaces or as switch virtual interfaces (SVIs). Your choice has an impact on the physical connectivity between the border nodes and the OTV edge devices:
  - Use of subinterfaces for tenant-routed communication: In this case, you must deploy two separate internal interfaces to connect the border nodes to the OTV devices. As shown in Figure 6, the first interface is configured with subinterfaces to carry tenant traffic (a unique VLAN tag is associated with each tenant), and the second interface is a Layer 2 trunk used to extend Layer 2 domains between sites.

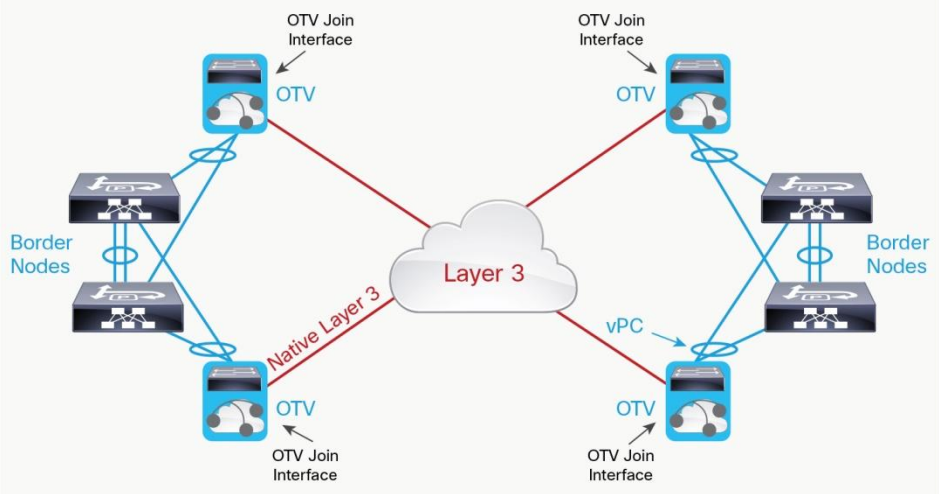
Figure 6. Layer 3 VPN Peering Over the Layer 2 Overlay Transport



Because Layer 3 routing peerings are established for each tenant to the remote border nodes across the OTV overlay service, you must enable bidirectional forwarding detection (BFD) on each Layer 3 subinterface for indirect failure detection, helping ensure faster convergence in those specific cases.

- Use of SVIs for tenant-routed communication: In this case, shown in Figure 7, the same Layer 2 trunk interface can be used to carry the VLANs associated with each tenant Layer 3 interface and the VLANs extending the Layer 2 domains across sites. Note that these two sets of VLANs must be unique, and that SVIs are not defined for the VLANs that extend the Layer 2 domains across sites.
- In this specific scenario, the join interfaces of the OTV edge devices can be connected directly to the Layer 3 core, as shown in Figure 7.

Figure 7. OTV Inline to Carry Data VLANs and VLANs for Tenants' Layer 3 Communication



VXLAN EVPN Multifabric Deployment Considerations

This section presents two use cases in which independent VXLAN EVPN fabrics are interconnected:

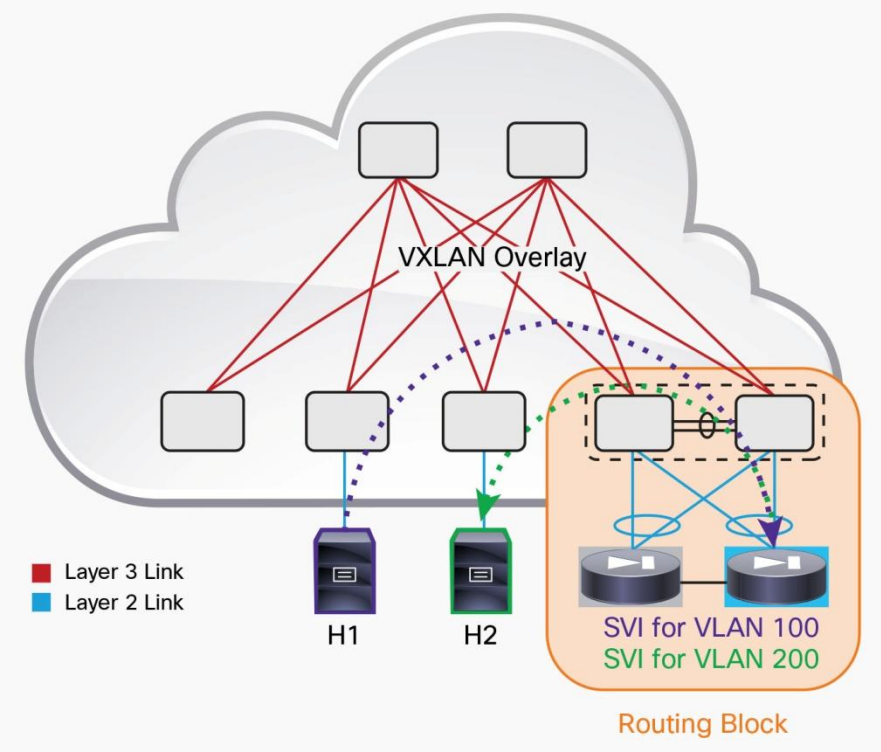
- The first use case is a more traditional deployment in which each VXLAN EVPN fabric runs in Layer 2 mode only and uses a centralized external routing block for intersubnet communication. This scenario can be a requirement when the default gateway function is deployed on external devices such as firewalls. This use case is discussed in the section "VXLAN EVPN Multifabric with External Active-Active Gateways."
- The second use case deploys the distributed functions of anycast Layer 3 gateways across both sites, reducing hairpinning workflow across remote network fabrics. This enhanced scenario can address a variety of requirements:
  - For business-continuance purposes, it is a common practice to allow transparent "hot" mobility of virtual machines from site to site without any service interruption. To reduce application latency, the same default gateways are active at both sites, reducing hairpinning across long distances for local routing purposes as well as for egress path optimization. Nonetheless, to reduce hairpinning for east-west workflows and to offer north-south traffic optimization, the same MAC and IP addresses for the default gateways must be replicated on all active routers at both sites. With the anycast Layer 3 gateway, the default gateway function is performed by all computing leaf nodes. This scenario is the main focus of this document and is discussed in greater detail in the section "VXLAN EVPN Multifabric with Distributed Anycast Layer 3 Gateway."

- For disaster-recovery purposes and for operational cost containment, enterprises may want to move endpoints using a “cold” migration process only. Consequently, only the IP address of the gateway will be replicated to simplify operation management after machines have been moved (for example, to maintain the same IP address schema after the relocation of the servers). The MAC address of the default gateway can be different on each fabric, and for a “cold” migration process, the endpoint will apply Address Resolution Protocol (ARP) on its default gateway when restarting its processes and will get the new MAC address accordingly. This use case is not covered in this document.

VXLAN EVPN Multifabric with External Active-Active Gateways

The first use case is simple. Each VXLAN fabric behaves like a traditional Layer 2 network with a centralized routing block. External devices (such as routers and firewalls) provide default gateway functions, as shown in Figure 8.

Figure 8. External Routing Block IP Gateway for VXLAN and EVPN Extended VLAN



In the Layer 2–based VXLAN EVPN fabric deployment, the external routing block is used to perform routing functions between Layer 2 segments. The same routing block can be connected to the WAN advertising the public networks from each data center to the outside and to propagate external routes to each fabric.

The routing block consists of a “router-on-a-stick” design (from the fabric’s point of view) built with a pair of traditional routers, Layer 3 switches, or firewalls that serve as the IP gateway. These IP gateways are attached to a pair of vPC border nodes that initiate and terminate the VXLAN EVPN tunnels.

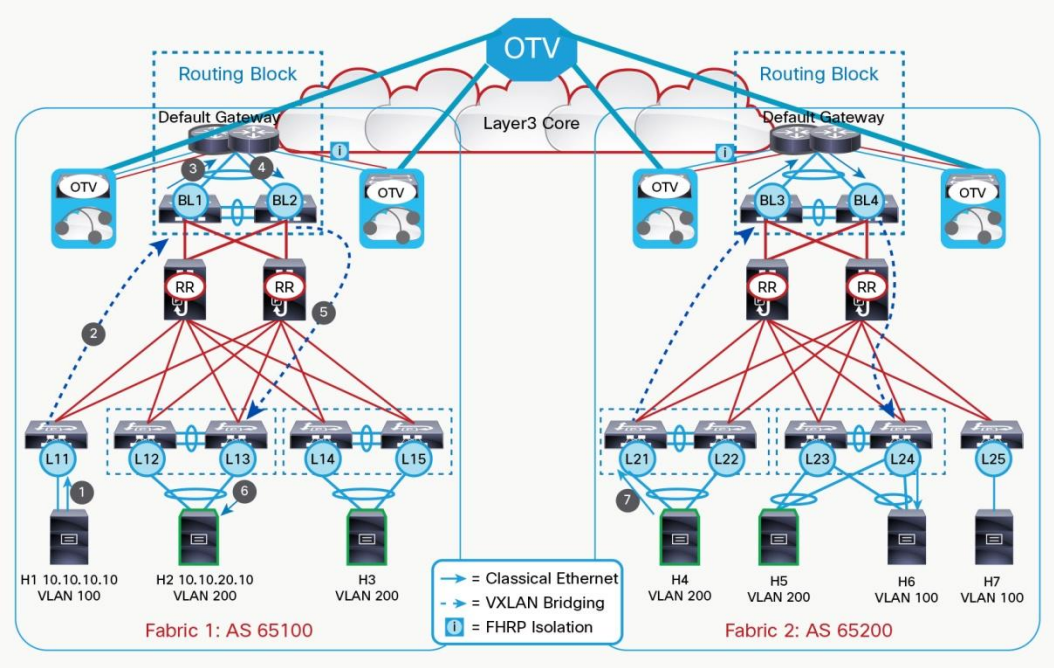
Connectivity between the IP gateways and the border nodes is achieved through a Layer 2 trunk carrying all the VLANs that require routing services.

To improve performance with active default gateways in each data center, reducing the hairpinning of east-west traffic for server-to-server communication between sites, and depending on the IP gateway platform of choice, the routing block can be duplicated with the same virtual IP and MAC addresses for all relevant SVIs on both sides. Hence, to use active-active gateways on both fabrics, you must filter communications between gateways that belong to the same First-Hop Routing Protocol (FHRP) group. With OTV as the DCI solution, the FHRP filter will be applied to the OTV control plane.

Figure 9 shows this scenario.

Note: Although VLANs are locally significant per edge device (or even per port), and the Layer 2 virtual network identifier (VNI) is locally significant in each VXLAN EVPN fabric, the following examples assume that the same Layer 2 VNIs (L2VNIs) are reused on both fabrics. The same VLAN ID was also reused on each leaf node and on the border switches. This approach was used to simplify the diagrams and packet walk. In a real production network, the network manager can use different network identifiers for the Layer 2 and 3 VNIs deployed in the individual fabrics.

Figure 9. VXLAN EVPN Layer 2 Fabric with External Routing Block



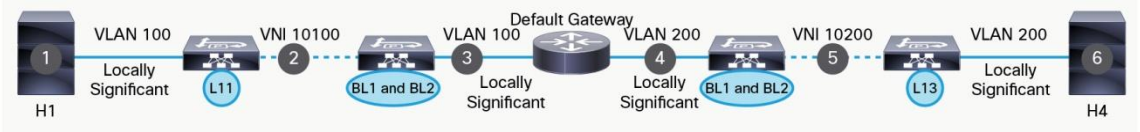


Figure 9 shows the following:

- Host H1 connected to leaf L11 in fabric 1 needs to send a data packet to host H2 connected in vPC mode to a pair of leaf nodes, L12 and L13, in the same fabric 1. Because H1 and H2 are part of different IP subnets, H1 sends the traffic to the MAC address of its default gateway, which is deployed on an external Layer 3 device connected to the same fabric 1. The communication between H1 and the default gateway uses the VXLAN fabric as a pure Layer 2 overlay service.

Note: The basic assumption is that hosts H1 and H2 have already populated their ARP tables with the default gateway information: for example, because they previously sent ARP requests to the gateway. As a consequence, the gateway also has H1 and H2 information in its local ARP table.

- Traffic from H1 that belongs to VLAN 100 is VXLAN encapsulated by leaf L11 with L2VNI 10100 (locally mapped to VLAN 100) and sent to the egress anycast VTEP address defined in border nodes 1 and 2. This address represents the next hop to reach the MAC address of the default gateway. Layer 3 equal-cost multipath (ECMP) is used in the fabric to load-balance the traffic destined for the egress anycast VTEP between the two border nodes. In the example in Figure 9, border node BL1 is selected as the destination.
- BL1 decapsulates the VXLAN frame and bridges the original frames destined for the local default gateway onto VLAN 100 (locally mapped to L2VNI 10100).
- The default gateway receives the frame destined for H2, performs a Layer 3 lookup, and subsequently forwards the packet to the Layer 2 segment on which H2 resides.
- The Layer 2 flow reaches one of the vPC-connected border nodes (BL2 in this example). BL2 uses the received IEEE 802.1q tag (VLAN 200) to identify the locally mapped L2VNI, 10200, to be used for VXLAN to encapsulates the frame. BL2 then forwards the data packet to the anycast VTEP address defined on the vPC pair of leaf nodes, L12 and L13, on which the destination H2 is connected.
- One of the receiving leaf nodes is designated to decapsulate the VXLAN frame and send the original data packet to H2.
- Routed communications between endpoints located at the remote site are kept local within fabric 2. This behavior is possible only because FHRP filtering is enabled on the OTV edge devices. In this example, H4 sends traffic destined for H6 using its local default gateway active in data center DC2. East-west routed traffic can consequently be localized within each fabric, eliminating unnecessary interfabric traffic hairpinning.

VXLAN EVPN Multifabric with Distributed Anycast Layer 3 Gateway

A distributed anycast Layer 3 gateway provides significant added value to VXLAN EVPN deployments for several reasons:

- It offers the same default gateway to all edge switches. Each endpoint can use its local VTEP as a default gateway to route traffic outside its IP subnet. The endpoints can do so not only within a fabric but across independent VXLAN EVPN fabrics (even when fabrics are geographically dispersed), removing suboptimal interfabric traffic paths. Additionally, routed flows between endpoints connected to the same leaf node can be directly routed at the local leaf layer.
- In conjunction with ARP suppression, it reduces the flooding domain to its smallest diameter (the leaf or edge device), and consequently confines the failure domain to that switch.

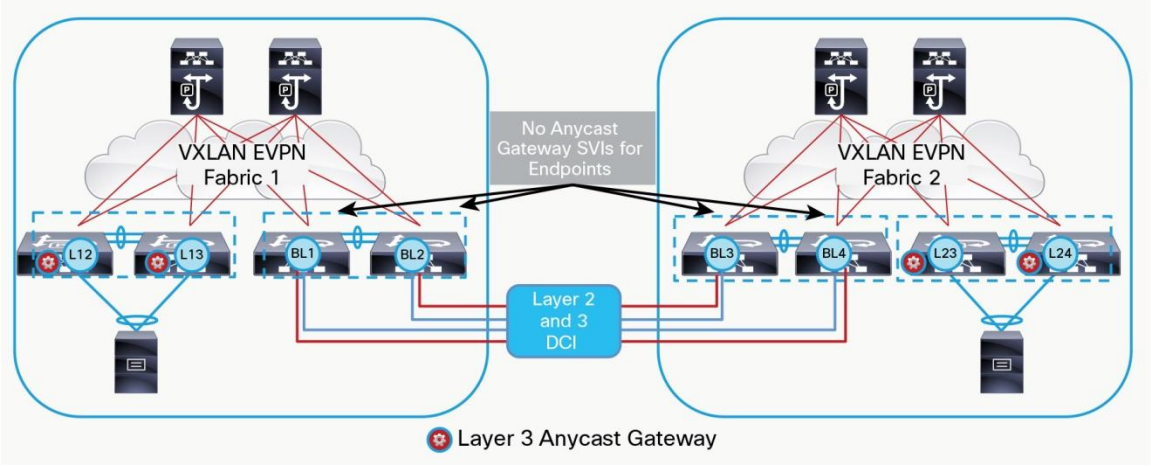
- It allows transparent host mobility, with the virtual machines continuing to use their respective default gateways (on the local VTEP), within each VXLAN EVPN fabric and across multiple VXLAN EVPN fabrics.
- It does not require you to create any interfabric FHRP filtering, because no protocol exchange is required between Layer 3 anycast gateways.
- It allows better distribution of state (ARP, etc.) across multiple devices.

In the VXLAN EVPN multifabric scenario with a distributed anycast Layer 3 gateway, the border nodes perform three main functions:

- VLAN and VRF-Lite hand-off to DCI
- MAN and WAN connection to the external Layer 3 network domain
- Connection to network services

For IP subnets that are extended between multiple fabrics, instantiation of the distributed IP anycast gateway on the border nodes is not supported. It is not supported because, with the instantiation of the distributed IP anycast gateway on the border nodes that also extend the Layer 2 network, the same MAC and IP addresses become visible on the Layer 2 extension on both fabrics, and unpredictable learning and forwarding can occur. Even if the Layer 2 network is not extended between fabrics, because it may potentially later be extended, the use of a distributed IP anycast gateway on the border node is not recommended (Figure 10).

Figure 10. No Distributed Anycast Gateway Function for Endpoints on Border Nodes



The next sections provide more detailed packet walk descriptions. But before proceeding, keep in mind that MP-BGP EVPN can transport Layer 2 information such as MAC addresses as well as Layer 3 information such as host IP addresses (host routes) and IP subnets. For this purpose, it uses two forms of routing advertisement:

- Route type 2: Used to announce host MAC and IP address information for the endpoint directly connected to the VXLAN fabric and also carrying extended community attributes (such as route-target values, router MAC addresses, and sequence numbers)
- Route type 5: Advertises IP subnet prefixes or host routes (associated, for example, with locally defined loopback interfaces) and also carrying extended community attributes (such as route-target values and router MAC addresses)

Learning Process for Endpoint Reachability Information

Endpoint learning occurs on the VXLAN EVPN switch, usually on the edge devices (leaf nodes) to which the endpoints are directly connected. The information (MAC and IP addresses) for locally connected endpoints is then programmed into the local forwarding tables.

This document assumes that a distributed anycast gateway is deployed on all the leaf nodes of the VXLAN fabric. Therefore, there are two main mechanisms for endpoint discovery:

- MAC address information is learned on the leaf node when it receives traffic from the locally connected endpoints. The usual data-plane learning function performed by every Layer 2 switch allows this information to be programmed into the local Layer 2 forwarding tables.

  You can display the Layer 2 information in the MAC address table by using the **show mac-address table** command. You can display the content of the EVPN table (Layer 2 routing information base [L2RIB]) populated by BGP updates by using the **show l2route evpn mac** command.

- IP addresses of locally connected endpoints are instead learned on the leaf node by intercepting control-plane protocols used for address resolution (ARP, Gratuitous ARP [GARP], and IPv6 neighbor discovery messages). This information is also programmed in the local L2RIB together with the host route information received through route-type-2 EVPN updates and can be viewed by using the **show l2route evpn mac-ip** command at the command-line interface (CLI).

> **Note:** Reverse ARP (RARP) frames do not contain any local host IP information in the payload, so they can be used only to learn Layer 2 endpoint information and not the IP addresses.

After learning and registering information (MAC and IP addresses) for its locally discovered endpoints, the edge device announces this information to the MP-BGP EVPN control plane using an EVPN route-type-2 advertisement sent to all other edge devices that belong to the same VXLAN EVPN fabric. As a consequence, all the devices learn endpoint information that belongs to their respective VNIs and can then import it into their local forwarding tables.

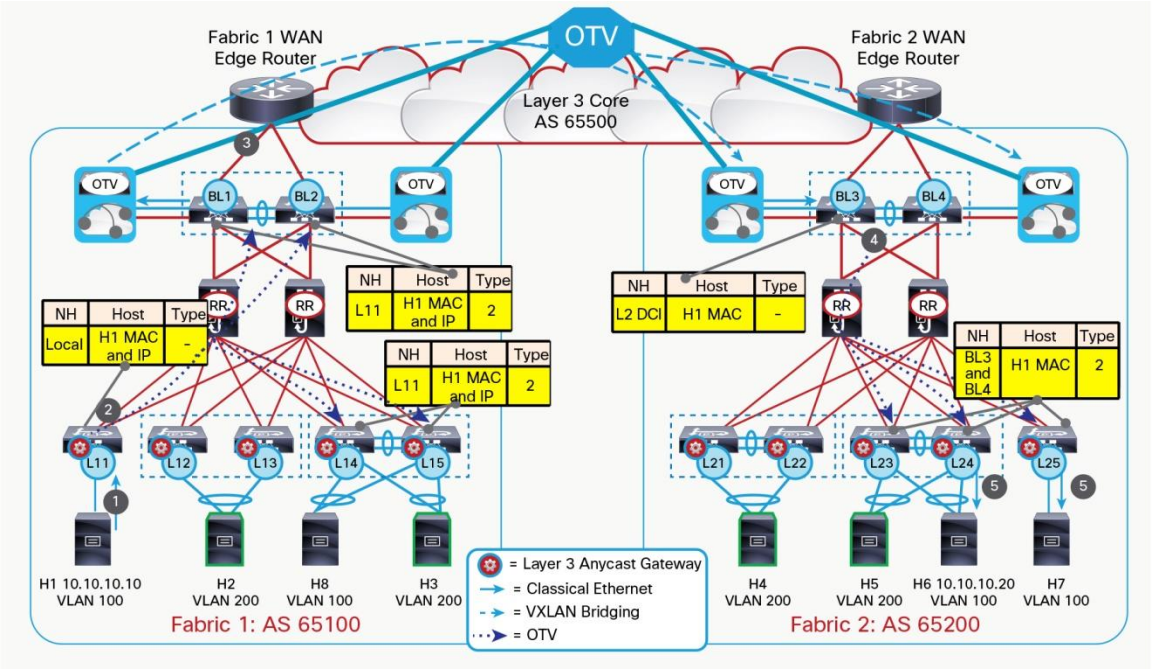
Intrasubnet Communication across Fabrics

Communication between two endpoints located in different fabrics within the same stretched Layer 2 segment (IP subnet) can be established by using a combination of VXLAN bridging (inside each fabric) and OTV Layer 2 extension services.

To better understand the way that endpoint information (MAC and IP addresses) are learned and distributed across fabrics, start by assuming that source and destination endpoints have not been discovered yet inside each fabric.

The control-plane and data-plane steps required to establish cross-fabric Layer 2 connectivity are highlighted in Figure 11 and described in this section.

Figure 11. VXLAN EVPN Multifabric: ARP Request Propagation Across Layer 2 DCI



1. Host H1 connected to leaf L11 in fabric 1 initiates intrasubnet communication with host H6, which belongs to remote fabric 2. As a first action, H1 generates a Layer 2 broadcast ARP request to resolve the MAC and IP address mapping for H6 that belongs to the same Layer 2 segment.
2. Leaf L11 receives the ARP packet on local VLAN 100 and maps it to the locally configured L2VNI 10100. Because the distributed anycast gateway is enabled for that Layer 2 segment, leaf L11 learns H1's MAC and IP address information and distributes it inside the fabric through an EVPN route-type-2 update. This process allows all the leaf nodes with the same L2VNI locally defined to receive and import this information in their forwarding tables (if they are properly configured to do so).

Note: The reception of the route-type-2 update triggers also a BGP update for H1's host route from border nodes BL1 and BL2 to remote border nodes BL3 and BL4 through the Layer 3 DCI connection. This process is not shown in Figure 11, because it is not relevant for intrasubnet communication.

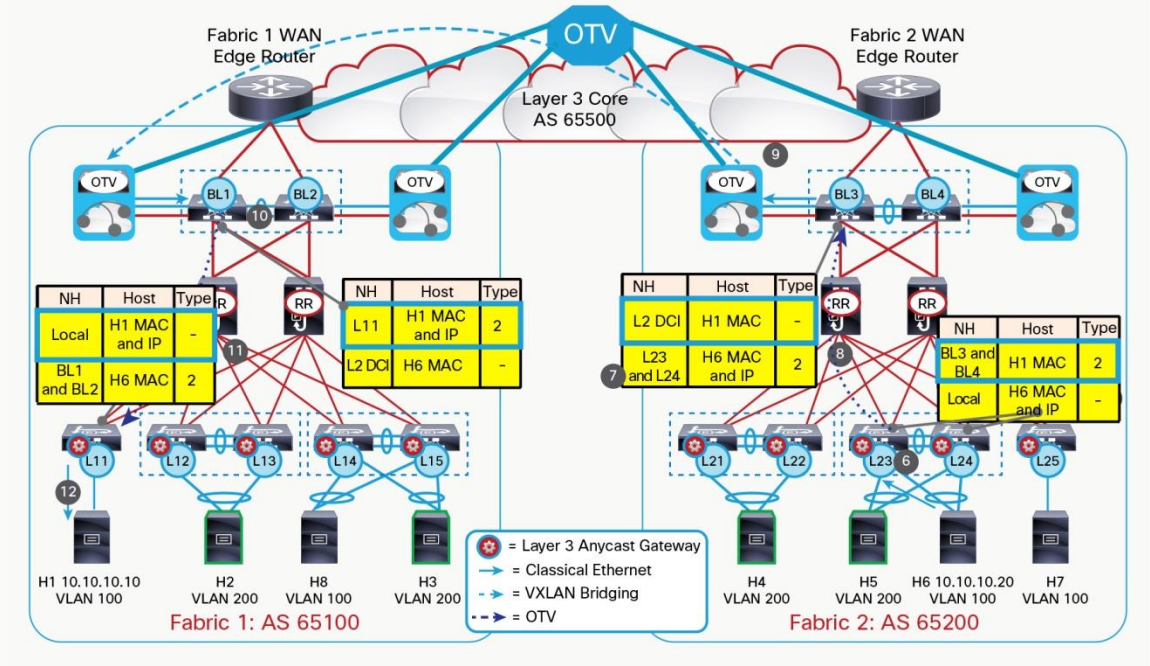
On the data plane, leaf L11 floods the ARP broadcast request across the Layer 2 domain identified by L2VNI 10100 (locally mapped to VLAN 100). The ARP request reaches all the local leaf nodes in fabric 1 that host that L2VNI, including border nodes BL1 and BL2. The replication of multidestination traffic can use either the underlay multicast functions of the fabric or the unicast ingress replication capabilities of the leaf nodes. The desired behavior can be independently tuned on a per-L2VNI basis within each fabric.

Note: The ARP flooding described in the preceding steps occurs regardless of whether ARP suppression is enabled in the L2VNI, because of the initial assumption that H6 has not been discovered yet.

3. The border nodes decapsulate the received VXLAN frame and use the L2VNI value in the VXLAN header to determine the bridge domain to which to flood the ARP request. They then send the request out the Ethernet interfaces that carry local VLAN 100 mapped to L2VNI 10100. The broadcast packet is sent to both remote OTV end devices. The OTV authoritative edge device (AED) responsible for extending that particular Layer 2 segment (VLAN 100) receives the broadcast packet, performs a lookup in its ARP cache table, and finds no entry for H6. As a result, it encapsulates the ARP request and floods it across the OTV overlay network to its remote OTV peers. The remote OTV edge device that is authoritative for that Layer 2 segment opens the OTV header and bridges the ARP request to its internal interface that carries VLAN 100.
4. Border nodes BL3 and BL4 receive the ARP request from the AED through the connected OTV inside interface and learn on that interface through the data plane the MAC address of H1. As a result, the border nodes announce H1's MAC address reachability information to the local fabric EVPN control plane using a route-type-2 update. Finally, the ARP request is flooded across L2VNI 10100, which is locally mapped to VLAN 100.
5. All the edge devices on which L2VNI 10100 is defined receive the encapsulated ARP packet, including leaf nodes L23 and L24, which are part of a vPC domain. One of the two leaf nodes is designated to decapsulate the ARP request and bridge it to its local interfaces configured in VLAN 100 and locally mapped to L2VNI 10100, so H6 receives the ARP request.

   At this point, the forwarding tables of all the devices are properly populated to allow the unicast ARP reply to be sent from H6 back to H1, as shown in Figure 12.

Figure 12. VXLAN EVPN Multifabric: ARP Reply Across Layer 2 DCI



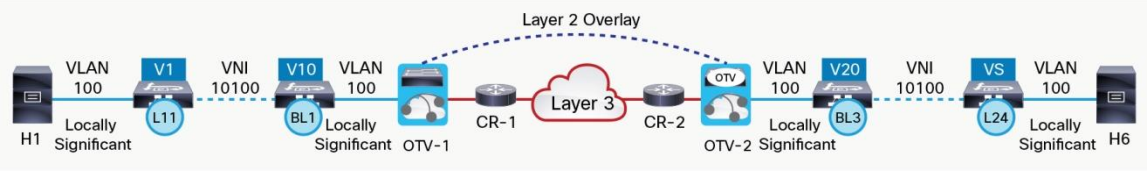
6. H6 sends an ARP unicast reply destined for H1's MAC address.
7. The reception of the packet allows leaf nodes L23 and L24 to locally learn H6's MAC and IP address information and then to generate an MP-BGP EVPN route-type-2 update to the fabric. As shown in the example in Figure 12, border nodes BL3 and BL4 also receive the update and so update their forwarding tables.

Note: The reception of the route-type-2 update also triggers a BGP update for H6's host route from border nodes BL3 and BL4 to remote border nodes BL1 and BL2 through the Layer 3 DCI connection. This process is not shown in Figure 12, because it is not relevant for intrasubnet communication.

8. Leaf L23 performs a MAC address lookup for H1 and finds as the next hop the anycast VTEP address defined on border nodes BL3 and BL4. It hence VXLAN encapsulates the response and sends it unicast to that anycast VTEP address.
9. The receiving border node (BL3 in this example) decapsulates the packet and bridges it to its Ethernet interface on which it previously learned H1's MAC address (pointing to the OTV AED). The OTV device receives the packet on the internal interface, performs a MAC address lookup for H1, and sends the packet to the remote OTV edge device that is authoritative for that VLAN. The OTV device in fabric 1 then forwards the packet to its internal interface.
10. BL1 in this specific example receives the ARP reply from the OTV device and learns the MAC address of H6. Subsequently, this information is announced to the VXLAN EVPN fabric using a route-type-2 BGP update, which is received by all relevant leaf nodes.
11. BL1 then performs a MAC address lookup for H1 and finds leaf L11 as its next hop to reach H1. Thus it VXLAN-encapsulates the ARP reply and sends it unicast to the leaf L11 VTEP address.
12. Leaf L11 decapsulates the VXLAN header and forwards the ARP reply unicast to H1.

At this stage, the leaf and border nodes in both VXLAN fabrics have fully populated their local forwarding tables with MAC reachability information for both the local and remote endpoints that need to communicate. H1 and H6 hence will always use the Layer 2 DCI connection for intrasubnet communication, as shown in Figure 13.

Figure 13. Intrasubnet Intersite Packet Walk



Note: The VLAN ID and VNID values used in this document have been kept consistent across the system for simplicity. Technically, the VLAN ID is locally significant per leaf node, and the VNID is locally significant per VXLAN fabric.

Intersubnet Communication across Fabrics

As previously mentioned, the use of a VXLAN EVPN anycast Layer 3 gateway allows you to reduce traffic hairpinning for routed communication between endpoints. Default gateways can all be active and distributed across different network fabrics on each leaf node. Therefore, you should be sure configure the default gateway used by the directly connected endpoints with the same Layer 2 virtual MAC (vMAC) address and Layer 3 virtual IP address. The distributed MAC address of the Layer 3 gateway is the same across all the defined L2VNIs in a VXLAN fabric. It is also known as the anycast gateway vMAC address.

So how does intersubnet communication occur between two VXLAN EVPN fabrics with a distributed anycast gateway?

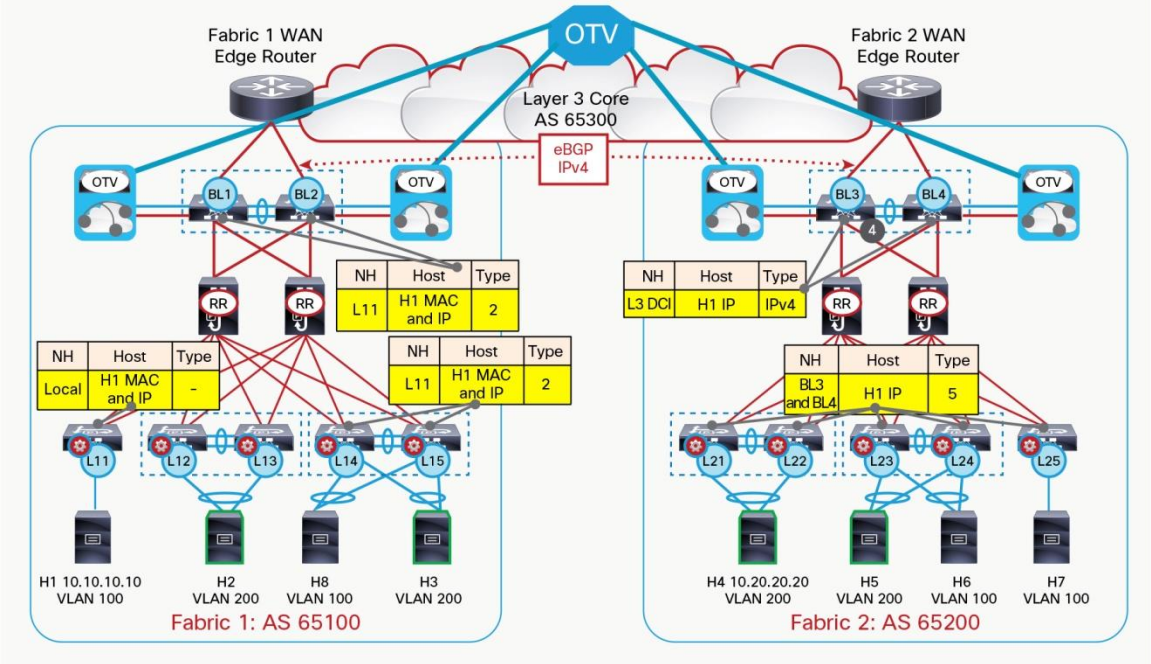
The following example describes the packet walk between two endpoints (H1 and H4) belonging to different IP subnets and communicating with each other. To better demonstrate the endpoint discovery and distribution processes, this section presents the full packet walk for the control plane (the learning and distribution of endpoint MAC and IP address information) and data plane (the data packets exchange between endpoints).

Because the VLAN ID is locally significant, this example refers to subnet_100, which is used for the Layer 2 domain on which H1 resides, and subnet_200, which is associated with the Layer 2 domain on which H4 resides.

To better show the routing behavior within the fabric, this example assumes that the first-hop router leaf L11 to which source endpoint H1 is connected has no destination IP subnet_200 locally configured (that is, no VLAN 200 is created on leaf L11).

In this scenario, for H1 to communicate with H4, it needs first to send an ARP request to its default gateway. This event then triggers a series of control-plane updates. Figure 14 shows this scenario.

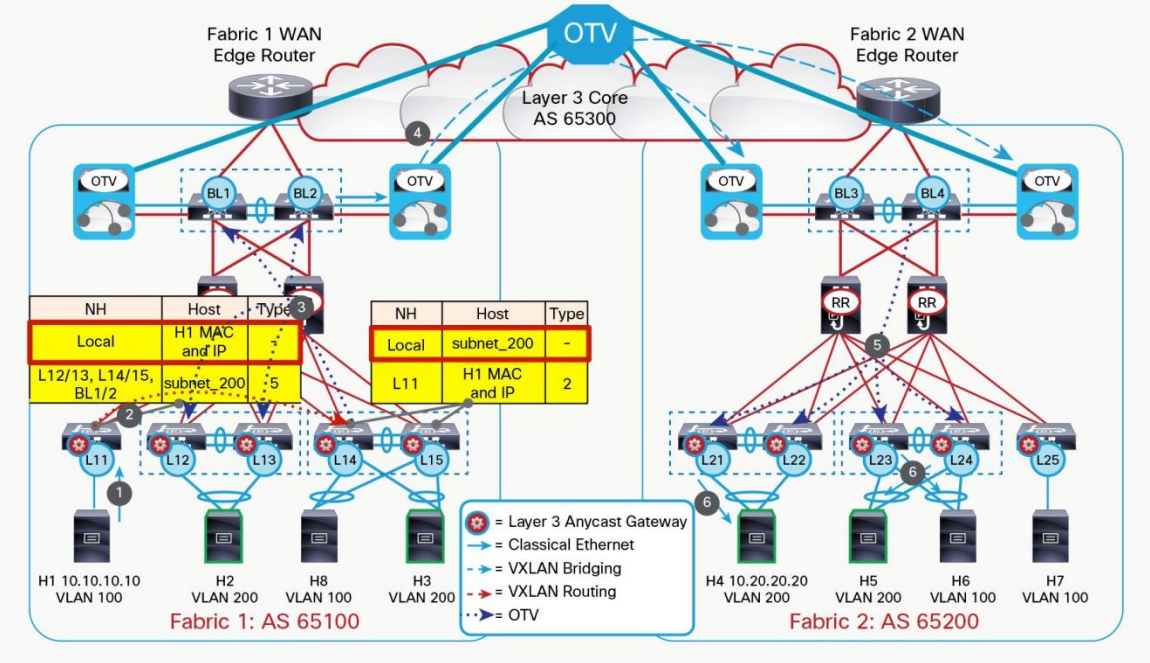
Figure 14. Forwarding Tables on Different Devices after H1 Sends ARP Request to Its Default Gateway



1. Leaf L11 consumes the received ARP request because the destination MAC address is the default gateway (locally defined). This process allows the leaf node to learn the Layer 2 and 3 reachability information for H1.
2. Leaf L11 injects this information into the fabric EVPN control plane using a route-type-2 BGP update.
3. Border leaf nodes BL1 and BL2 receive the BGP update and generate a Layer 3 update specific to H1's VRF instance using, in this specific example, an external BGP (eBGP) update across the Layer 3 DCI connection.
4. Consequently, border nodes BL3 and BL4 in fabric 2 receive the Layer 3 update from their Layer 3 DCI connection and advertise a route-type-5 update to the local MP-BGP EVPN control plane.

After the completion of the preceding steps, the forwarding tables of the devices in fabrics 1 and 2 are updated as shown in Figure 14. H1 now can start communicating with H4 as described in the following steps and shown in Figure 15.

Figure 15. VXLAN EVPN Multifabric Data Plane: Intersubnet Communication



Note: The following steps assume that destination H4 is a silent host: that is, it has not sent any ARP message or data packet, and so its MAC and IP address information is not yet known in the EVPN control plane.

1. H1 connected to leaf L11 in fabric 1 (subnet_100) generates a data packet destined for remote endpoint H4, which resides in fabric 2 in a different subnet, subnet_200. The traffic is hence sent to its default gateway (that is, the destination MAC address is the anycast gateway MAC address configured in the fabric), which is locally active on leaf L11.
2. Leaf L11 has previously received an EVPN route-type-5 update indicating that subnet_200 is reachable through multiple next hops (equal-cost paths). In the specific example shown in Figure 15, those next hops are the anycast VTEP address for leaf nodes L12 and L13 and the anycast VTEP address for leaf nodes L14 and L15, because subnet_200 exists on all these switches. You should, in fact, configure all leaf nodes to announce the locally defined IP subnet prefixes within the fabric to allow the discovery of silent hosts. For more information about this specific design point, refer to the VXLAN design guide at http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/guide-c07-734107.html.

Note: The IP prefix associated with subnet_200 is also advertised in fabric 1 (route-type-5 BGP update) by border nodes BL1 and BL2, because they receive this information through the route peering over the Layer 3 DCI connection to remote fabric 2.

In the specific example described here, leaf L11 selects the anycast VTEP address on leaf nodes L14 and L15 as the valid next hop based on the hashing result. It then routes the VXLAN encapsulated traffic to the leaf nodes through the transit L3VNI for that particular tenant (VRF instance).

3. The receiving leaf (L14 in this example) performs a Layer 3 lookup but finds no specific host route entry for H4, so it routes the traffic to locally defined subnet_200. This routing triggers an ARP request destined for H4 that is flooded across the L2VNI associated with subnet_200.
4. All the leaf nodes with the L2VNI locally defined receive the ARP request, including border nodes BL1 and BL2. One of the two border nodes decapsulates the VXLAN frame and floods the ARP request across VLAN 200 through its Layer 2 interfaces connected to the OTV devices. The OTV AED for VLAN 200 receives the ARP request and floods it across its extended overlay network to all remote OTV devices. The remote OTV AED in fabric 2 receives the ARP request and forwards it to its inside interface connecting to border nodes BL3 and BL4.

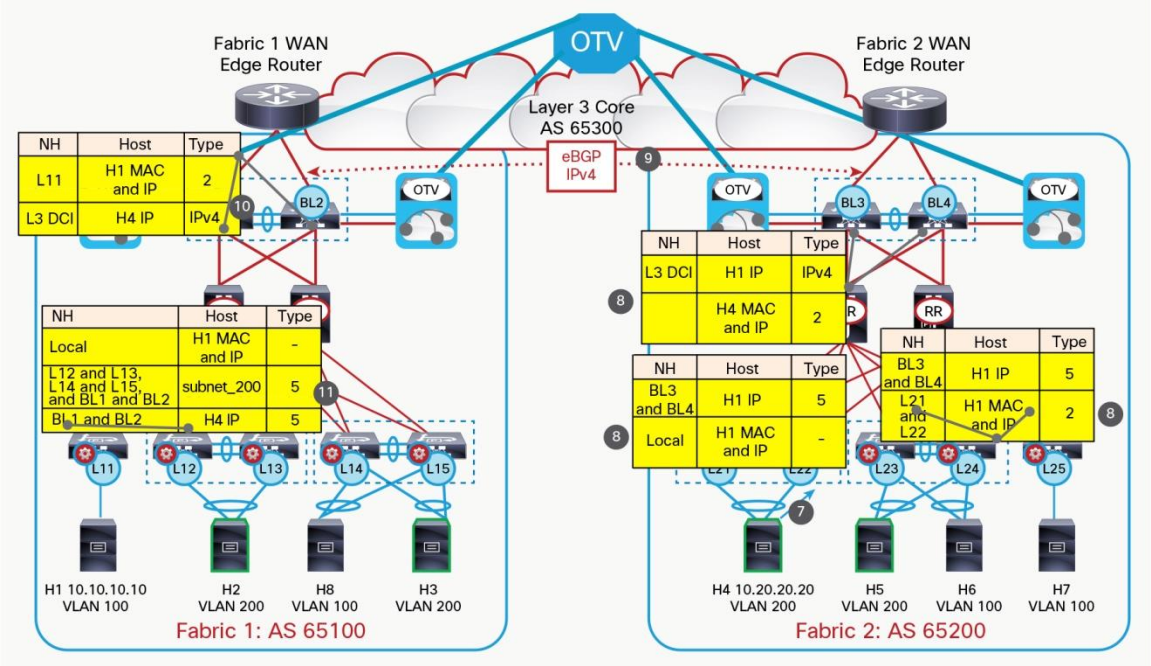
   Note that when you configure a common anycast gateway vMAC address across VXLAN fabrics, the OTV devices at each site will continuously update their Layer 2 tables, because they may receive on their Layer 2 internal interfaces ARP requests originating from endpoints connected to the local site. As a good practice, you thus should apply a route map on the OTV control plane to avoid communicating the anycast gateway MAC address information to remote OTV edge devices. You can apply a route map to the OTV Intermediate System–to–Intermediate System (IS-IS) control plane, as shown in the following configuration sample.

```
mac-list Anycast_GW_MAC_deny seq 10 deny 0001.0001.0001 ffff.ffff.ffff
mac-list Anycast_GW_MAC_deny seq 20 permit 0000.0000.0000 0000.0000.0000
route-map Anycast_GW_MAC_filter permit 10
  match mac-list Anycast_GW_MAC_deny
!
otv-isis default
  vpn Overlay0
    redistribute filter route-map Anycast_GW_MAC_filter
```

5. Border node BL4 in this specific example receives the ARP request, encapsulates the packet, and floods the packet across fabric 2 in the VXLAN Layer 2 domain associated with subnet_200.
6. The receiving leaf nodes flood the ARP request to all the local interfaces on which VLAN 200 is configured. In the specific case of leaf nodes L21 and L22, L21 is designated to forward the traffic to the vPC connection to H4, which then receives the ARP request.

   Figure 16 shows the series of control-plane events triggered by the ARP reply originating from H4.

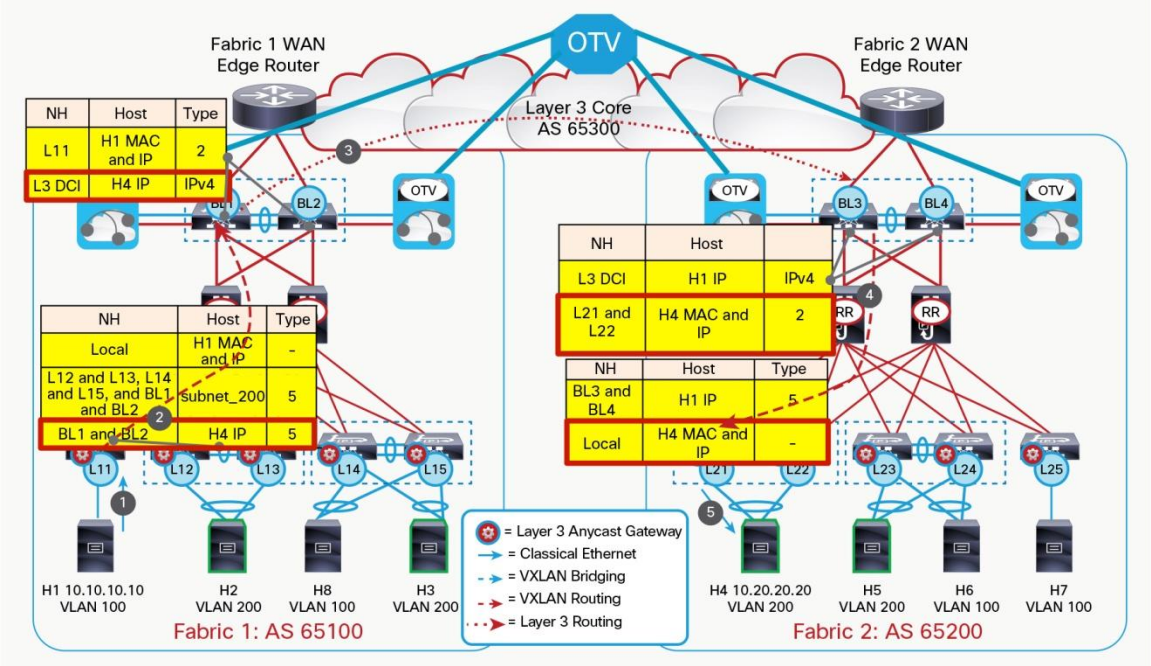
Figure 16. Control-Plane Updates Triggered by the H4 ARP Reply

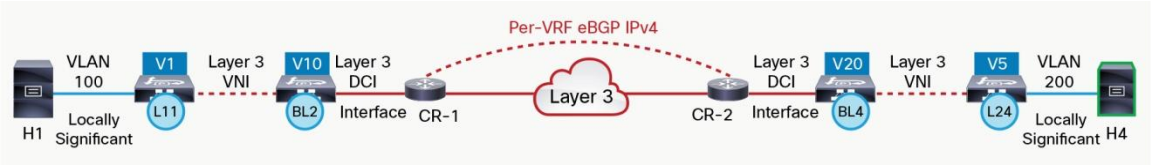


7. H4 sends an ARP reply to the MAC address that sourced the request (the anycast gateway MAC address identifying leaf L14 in fabric 1). However, because the same global anycast gateway MAC address is identically configured at both sites, local leaf L22 locally consumes the ARP reply. As a consequence, the ARP response will never get back to the original sender in fabric 1. This is not a problem, because, as clarified in the following steps, this process allows H4's discovery to be triggered in both local fabric 2 and remote fabric 1.
8. The received ARP reply triggers on leaf nodes L21 and L22 the local discovery of H4. As a consequence, H4's Layer 2 and 3 reachability information is announced to fabric 2 through an MP-BGP EVPN route-type-2 update. All the leaf nodes that have locally configured the L2VNI associated with subnet_200 (in this example, leaf nodes L23 and L24 and border nodes BL3 and BL4) receive this information.
9. Border nodes BL3 and BL4 advertise H4's host route across the Layer 3 DCI connection to border nodes BL1 and BL2 in fabric 1.
10. Border nodes BL1 and BL2 receive the BGP update from fabric 2 and generate an MP-BGP EVPN route-type-5 update in fabric 1 for H4's host-route information.
11. Now all the local leaf nodes can import H4's host route into their forwarding tables, specifying the anycast IP address of border nodes BL1 and BL2 as the next hop.

At this point, leaf L11 has learned host route information for H4, which changes the forwarding behavior of the next data packet generated by the locally connected endpoint H1 destined for H4, as described in the next steps and shown in Figure 17.

Figure 17. VXLAN EVPN Multifabric: Intersubnet Communication (Continued)





1. H1 generates another data packet destined for H4 and sends it to its local default gateway on leaf L11.
2. Leaf L11 performs a Layer 3 lookup for H4, and this time it finds the host route pointing to the anycast VTEP address on border nodes BL1 and BL2 as the next hop. It hence encapsulates and unicasts the data packet to that anycast address.
3. The receiving border node decapsulates the VXLAN packet and uses the retrieved L3VNI information to perform a Layer 3 lookup in H4's routing domain. The result is a host route pointing to a next-hop router reachable through the Layer 3 DCI connection. Hence, traffic is sent to remote border nodes BL3 and BL4 in fabric 2.
4. Border nodes BL3 and BL4 receive the traffic destined for H4 and perform a Layer 3 lookup in the appropriate VRF instance. Having previously received a route-type-2 update, the border node encapsulates the packet with a VXLAN header and routes it to the anycast VTEP address defined on leaf nodes L21 and L22 using the transit L3VNI defined for that specific tenant.
5. Leaf L21 receives the packet and routes it locally to H4.

Note that routed communication between endpoints connected to separate IP subnets in different VXLAN fabrics always use the Layer 3 DCI connection between fabrics, whereas the Layer 2 DCI is reserved exclusively for intrasubnet flows. This behavior occurs regardless of whether the IP subnets are stretched or locally defined inside each fabric.

Host Mobility across Fabrics

This section discusses support for host mobility when a distributed anycast gateway is configured across multiple VXLAN EVPN fabrics.

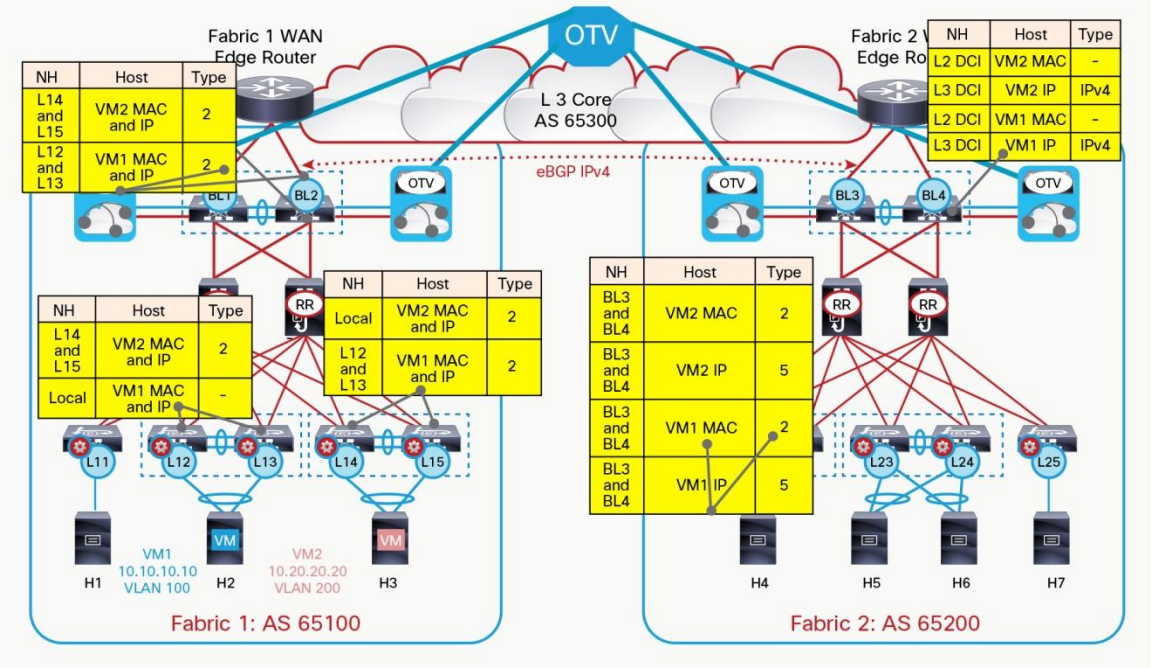
> **Note:** If you are deploying Cisco Nexus 9000 Series Switches as leaf nodes, the scenario described in this section requires NX-OS Release 7.0(3)I2(2e) or later.

In this scenario, VM1 belonging to VLAN 100 (subnet_100) is hosted by H2 in fabric 1, and VM2 on VLAN 200 (subnet_200) initially is hosted by H3 in the same fabric 1. Destination IP subnet_100 and subnet_200 are locally configured on leaf nodes L12 and L13 as well as on L14 and L15.

This example assumes that the virtual machines (endpoints) have been previously discovered, and that Layer 2 and 3 reachability information has been announced across both sites as discussed in the previous sections.

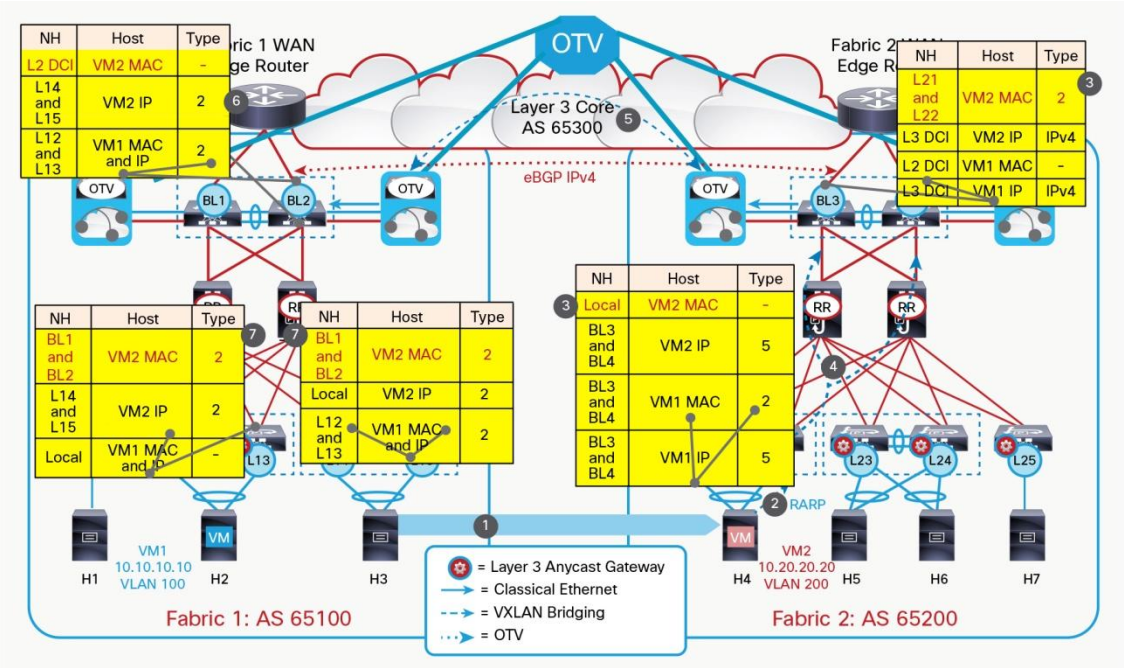
Figure 18 highlights the content of the forwarding tables on different leaf nodes in both fabrics before virtual machine VM2 is migrated to fabric 2.

Figure 18. Content of Forwarding Tables before Host Migration



The following steps show the process for maintaining communication between the virtual machines in a host mobility scenario, as depicted in Figure 19.

Figure 19. VXLAN EVPN Multifabric and Host Mobility



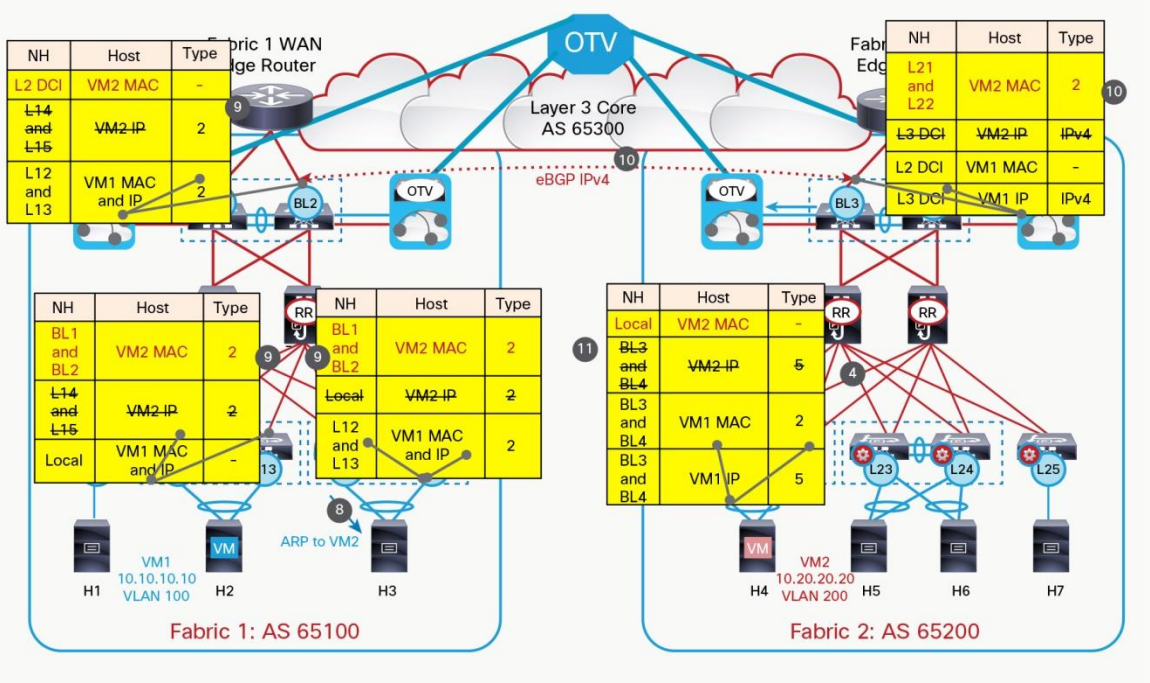
1. For operational purposes, virtual machine VM2 moves to host H4 located in fabric 2 and connected to leaf nodes L21 and L22.
2. After the migration process is completed, assuming that VMware ESXi is the hypervisor used, the virtual switch generates a RARP frame with VM2's MAC address information.

> **Note:** With other hypervisors, such as Microsoft Hyper-V or Citrix Xen, a GARP request is sent instead, which includes the source IP address of the sender in the payload. As a consequence, the procedure will be slightly different than the one described here.

3. Leaf L22 in this example receives the RARP frame and learns the MAC address of VM2 as locally connected. Because the RARP message can't be used to learn VM2's IP address, the forwarding table of the devices in fabric 2 still points to border nodes BL3 and BL4 (that is, VM2's IP address is still known as connected to fabric 1). Leaf L22 also sends an MP-BGP EVPN route-type-2 update in fabric 2 with VM2's MAC address information. When doing so, it increases the sequence number associated with this specific entry and specifies as the next hop the anycast VTEP address of leaf nodes L21 and L22. The receiving devices update their forwarding tables with this new information.
4. On the data plane, the RARP broadcast frame is also flooded in fabric 2 and reaches border nodes BL3 and BL4, which forward it to the local OTV devices.
5. The OTV AED in fabric 2 forwards the RARP frame across the Layer 2 DCI overlay network to reach the remote OTV devices in fabric 1. The OTV AED device in fabric 1 forwards the frame to the local border nodes.
6. Border nodes BL1 and BL2 learn the MAC address of VM2 from the reception of the RARP frame as locally attached to their Layer 2 interfaces connecting to the OTV AED device. As a consequence, one of the border nodes advertises VM2's MAC address information in fabric 1 with a route-type-2 BGP update using a new sequence number (higher than the previous number).
7. The forwarding tables for all relevant local leaf nodes in fabric 1 are updated with the information that VM2's MAC address is now reachable through the anycast VTEP address of border nodes BL1 and BL2.

At this point, all the devices in fabrics 1 and 2 have properly updated their forwarding tables with the new VM2's MAC address reachability information. This process implies that intrasubnet communication to VM2 is now fully reestablished. However, VM2's IP address still is known in both fabrics as connected to the old location (that is, to leaf nodes L14 and L15 in fabric 1), so communications still cannot be routed to VM2. Figure 20 shows the additional steps required to update the forwarding tables of the devices in fabrics 1 and 2 with the new reachability information for VM2'S IP address.

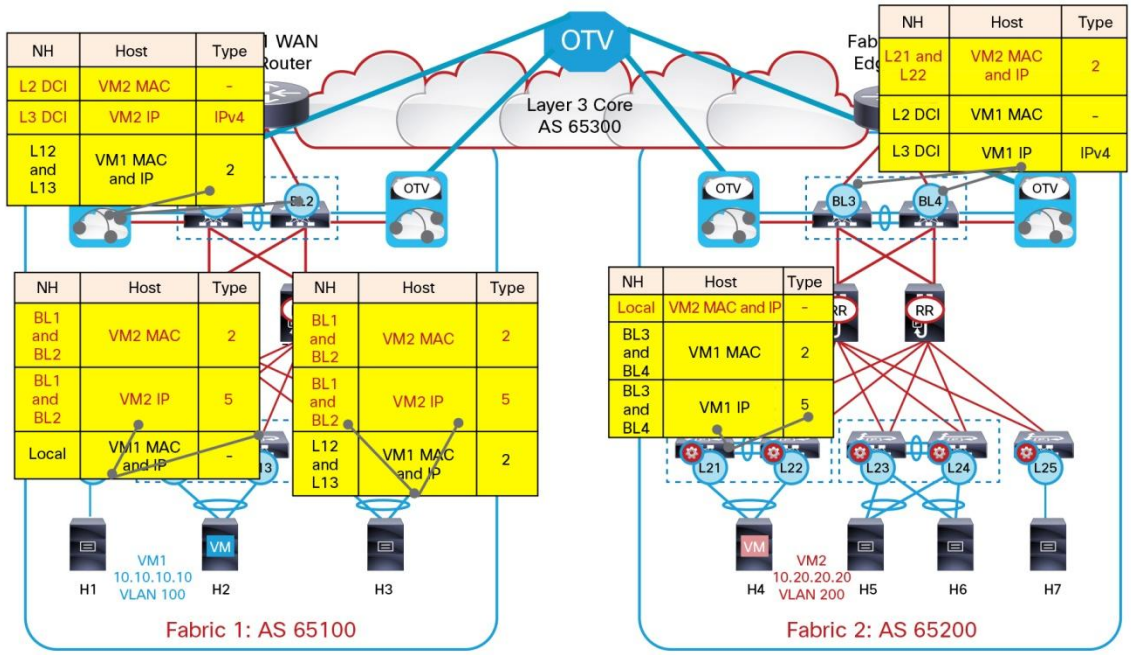
Figure 20. Propagation of VM2's Reachability Information Toward Fabric 1



8. The reception of the route-type-2 MAC address advertisement on leaf nodes L14 and L15 triggers a verification process to help ensure that VM2 is not locally connected anymore. Note that ARP requests to VM2 are locally sent out the local interface to which VM2 was originally connected as well as to fabric 1 and subsequently to fabric 2 through the Layer 2 DCI connection. The ARP request reaches VM2, which responds, allowing leaf nodes L21 and L22 to update the local ARP table and trigger the consequent control-plane updates discussed previously and shown in Figure 16.
9. After verification that VM2 has indeed moved away from leaf nodes L14 and L15, one of the leaf nodes withdraws VM2's IP reachability information from local fabric 1, sending an MP-BGP EVPN update. This procedure helps ensure that this information can be cleared from the forwarding tables of all the devices in fabric 1.
10. Because border nodes BL1 and BL2 also receive the withdrawal of VM2's IP address, they update the border nodes in the remote fabric to indicate that this information is not reachable anymore through the Layer 3 DCI connection.
11. As a consequence, border nodes BL3 and BL4 also withdraw this information from remote VXLAN EVPN fabric 2, allowing all the local devices to clear this information from their tables.

The end result is the proper update of VM2's IP address information in the forwarding tables of all the nodes in both fabrics, as shown in Figure 21.

Figure 21. End State of the Forwarding Tables for Nodes in Fabrics 1 and 2



At this point, Layer 2 and 3 communication with VM2 can be fully reestablished.

Ingress and Egress Traffic-Path Optimization

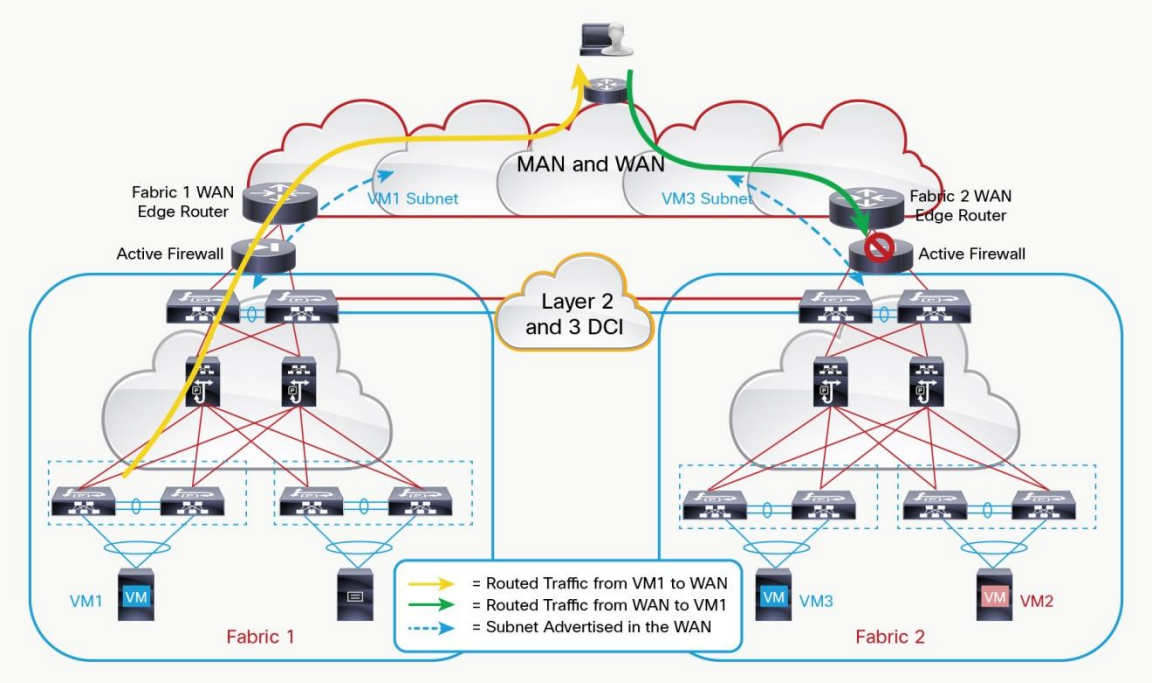
In the VXLAN multifabric design discussed in this document, each data center normally represents a separate BGP autonomous system (AS) and is assigned a unique BGP autonomous system number (ASN).

Three types of BGP peering are usually established as part of the VXLAN multifabric solution:

- MP internal BGP (MP-iBGP) EVPN peering sessions are established in each VXLAN EVPN fabric between all the deployed leaf nodes. As previously discussed, EVPN is the intrafabric control plane used to exchange reachability information for all the endpoints connected to the fabric and for external destinations.
- Layer 3 peering sessions are established between the border nodes of separate fabrics to exchange IP reachability information (host routes) for the endpoints connected to the different VXLAN fabrics and the IP subnets that are not stretched (east-west communication). Often, a dedicated Layer 3 DCI network connection is used for this purpose. In a multitenant VXLAN fabric deployment, a separate Layer 3 logical connection is required for each VRF instance defined in the fabric (VRF-Lite model). Although either eBGP or IGP routing protocols can be used to establish interfabric Layer 3 connectivity, the eBGP scenario is the most common and is the one discussed in this document.
- Per-VRF eBGP peering sessions are frequently used for WAN connectivity to exchange IP reachability information with the external Layer 3 network domain (north-south communication). A common best practice and the recommended approach is to deploy eBGP for this purpose.

When extending IP subnets across separate VXLAN fabrics, you need to consider the paths used for ingress and egress communication. You particularly should avoid establishing an asymmetric path such as one shown in Figure 22 (virtual machines VM1 and VM3 are part of the same extended IP subnet that is advertised in the MAN and WAN), which would cause communication failure when independent stateful network services are deployed across sites.

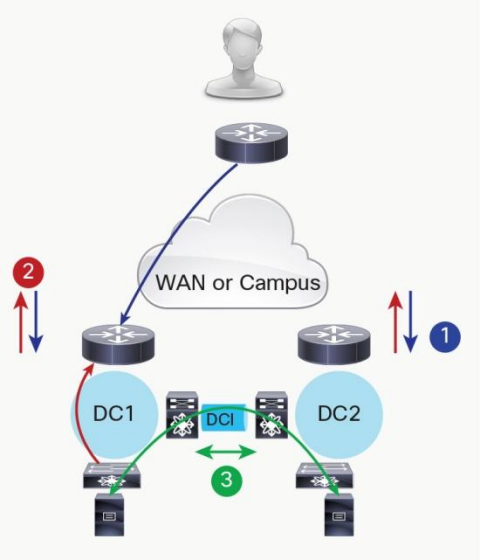
Figure 22. Establishment of Undesirable Asymmetric Flow



The solution presented in this document avoids establishing asymmetric traffic paths by following three main design principles, illustrated in Figure 23:

- Traffic originating from the external Layer 3 domain and destined for endpoints connected to a specific VXLAN fabric should always come inbound through the site's local border nodes (ingress traffic-path optimization).
- Traffic originating from data center endpoints and destined for the external Layer 3 domain should always prefer the outbound path through the local border nodes (outbound traffic-path optimization).
- All east-west routed communication between endpoints that are part of different data center sites should always prefer the dedicated Layer 3 DCI connection if it is available.

Figure 23. Ingress and Egress Traffic-Path Optimization and East-West Communication



Ingress Traffic-Path Optimization

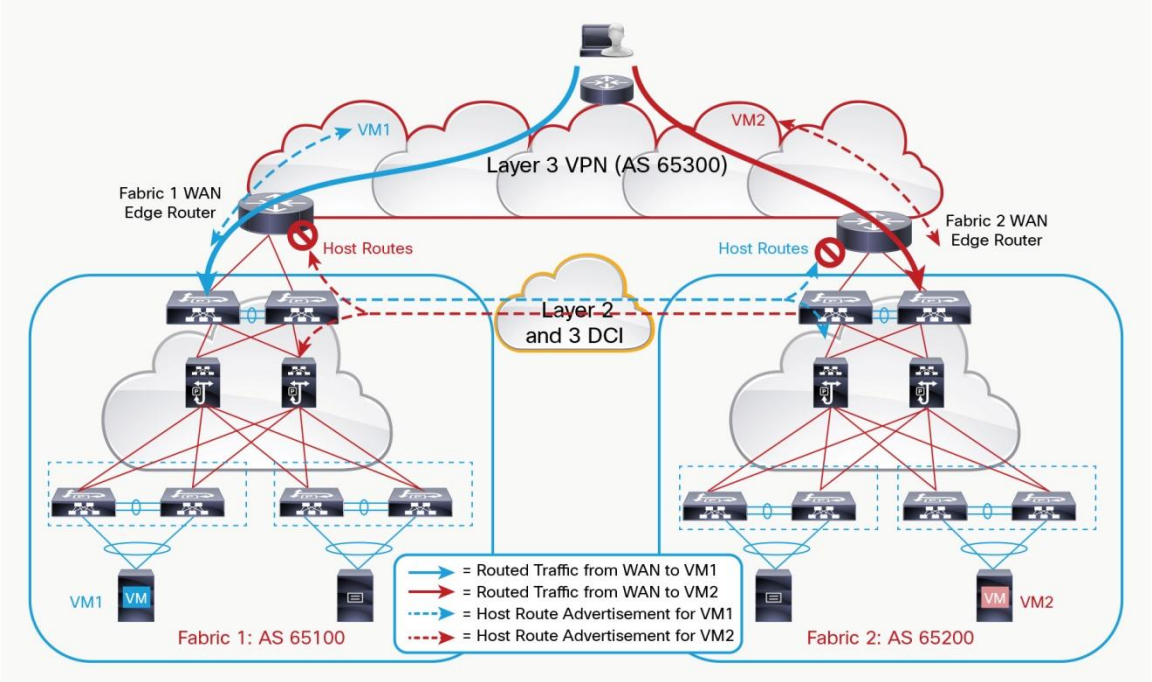
This document proposes the use of host-route advertisement from the border nodes to the local WAN edge routers to influence and optimize ingress traffic flows. After the WAN edge routers receive the host routes, two approaches are possible:

- You can inject specific host-route information from each VXLAN fabric into the MAN or WAN so that incoming traffic can be optimally steered to the destination. This method is usually applicable when a Layer 3 VPN hand-off is deployed to the WAN, so that host routes can be announced in each specific Layer 3 VPN service. Before adopting this approach, be sure to assess the scalability implications of the solution for consumption of local resources on the data center WAN edge and remote routers and within the WAN provider network.
- In designs in which advertisement of host routes in the MAN or WAN is not desirable because of scalability concerns or is not possible (as, for example, in many cases in which the MAN or WAN is managed by a service provider), you can deploy Cisco Location Identifier Separation Protocol (LISP) as an IP-based hand-off technology. LISP deployment is beyond the scope of this document. For more information about LISP and LISP mobility, see http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Data_Center/DCI/5-0/LISPmobility/DCI_LISP_Host_Mobility.html.

The rest of this section focuses on the deployment model using a Layer 3 VPN hand-off to the MAN or WAN. The goal is to help ensure that the border nodes in each VXLAN fabric always advertise to the local WAN edge routers only host routes for endpoints connected to the local fabric.

As shown in Figure 24, the border nodes at a given site receive host route information for endpoints connected to remote fabrics through the route peering established over the dedicated Layer 3 DCI connection. You therefore must help ensure that these host routes advertised from the Layer 3 DCI are never sent to the local WAN edge routers. Host routing from the WAN always should steer the traffic to the fabric to which those destinations are connected, and a less specific route should be used only in specific WAN isolation scenarios discussed later in this section.

Figure 24. Border Nodes Advertising Only Host Routes for Locally Connected Endpoints



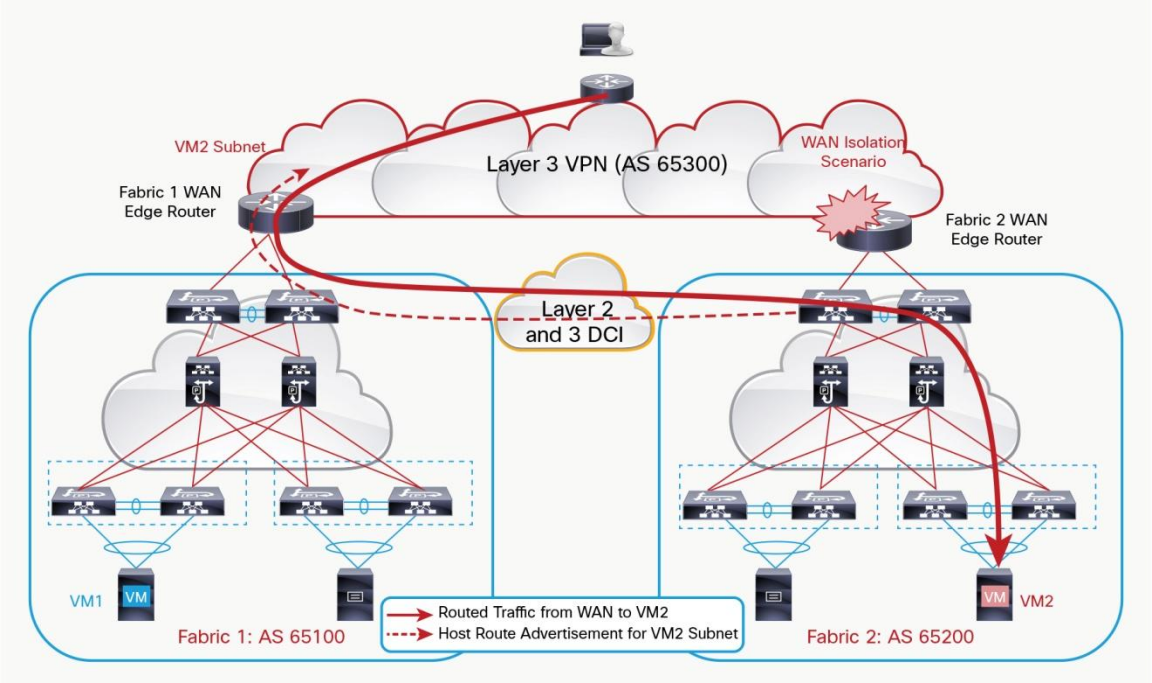
Many different approaches can be used to achieve the behavior shown in Figure 24. The example shown here simply proposes configuring the border nodes in a given fabric so that they do not announce to the local WAN edge routers host route information received through the Layer 3 DCI connection.

```
ip as-path access-list 1 permit "_65200_”
!
ip prefix-list MATCH-HOST-ROUTES seq 5 permit 0.0.0.0/0 eq 32
!
ip access-list ANY
  10 permit ip any any
!
route-map DENY-HOST-ROUTES-FROM-REMOTE-DCs deny 10
  match as-path 1
  match ip address prefix-list MATCH-HOST-ROUTES
!
route-map DENY-HOST-ROUTES-FROM-REMOTE-DCs permit 20
  match ip address ANY
!
router bgp 65100
  vrf Tenant-1
    neighbor 10.0.1.1
      description Local WAN Edge Device
      remote-as 65300
      address-family ipv4 unicast
        route-map DENY-HOST-ROUTES-FROM-REMOTE-DCs out
```

Note: The same configuration must be applied to all the border nodes facing the WAN edge and deployed across VXLAN fabrics.

Note that other prefixes received from remote sites are still accepted, as long as they are not host routes. This behavior is required mainly to handle the potential WAN isolation scenario in which a given fabric loses connectivity to the MAN or WAN (as a result of a dual failure of the WAN edge routers or a WAN or MAN outage). In that case, traffic originating from the external routed domain and destined for an endpoint connected to the WAN-isolated fabric should be steered to a different VXLAN fabric following a less specific route (usually the IP prefix for the subnet to which the destination endpoint is connected). This scenario is shown in Figure 25.

Figure 25. Inbound Traffic in a WAN Isolation Scenario



Note: The same considerations apply if VXLAN fabric 1 experiences a WAN isolation scenario.

When virtual machine VM1 migrates to data center DC2, the host route information is updated across the VXLAN fabrics, as previously described in the section "Host Mobility Across Fabrics." As a consequence, local border nodes BL3 and BL4 will start advertising VM1's host route to the fabric 2 WAN edge router with AS 65200 of fabric 2, and border nodes BL1 and BL2 will stop sending the same host route information to the fabric 1 WAN edge router. As a consequence, traffic destined for VM1 and originating from the Layer 3 MAN or WAN will be steered directly to fabric 2.

Egress Traffic-Path Optimization

After you have optimized the ingress traffic, you usually should do the same for the egress traffic to maintain symmetry for the communications with the external Layer 3 domain. As previously mentioned, this optimization is mandatory for deployment across sites with independent stateful network services such as firewalls.

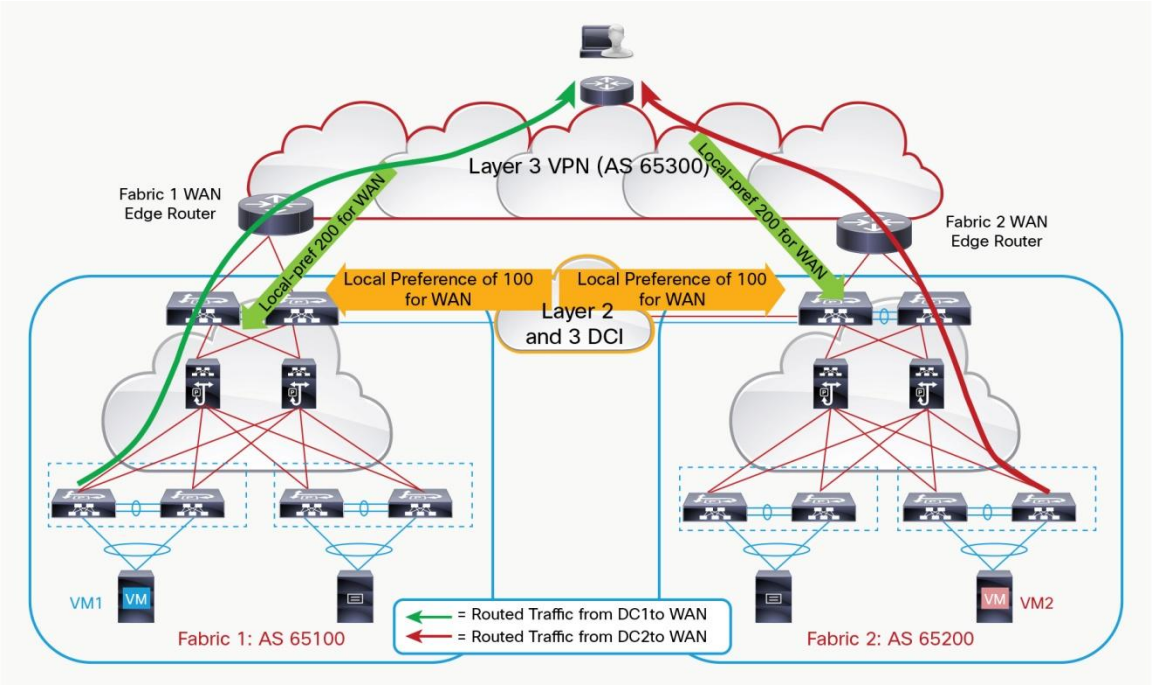
To force the egress traffic to prefer the local WAN connection, you can modify the local-preference value for the prefixes learned through peering with the local WAN edge routers. The local preference is an attribute that routers exchange in the same autonomous system and that tells the autonomous system which path to prefer to reach destinations that are external to it. A path with a higher local-preference value is preferred. The default local-preference value is 100.

In the configuration sample shown here, one of the border nodes in VXLAN EVPN fabric 1 is configured to assign a higher local preference (200) to all the prefixes received from the WAN edge routers in fabric 1 in data center DC1 using the route map **INCREASE-LOCAL-PREF-FOR-WAN-ROUTES.**

```
ip access-list ANY
  10 permit ip any any
!
route-map INCREASE-LOCAL-PREF-FOR-WAN-ROUTES permit 10
  match ip address ANY
  set local-preference 200
!
router bgp 65100
  vrf Tenant-1
    neighbor 10.0.1.1
      remote-as 65300
      description eBGP Peering with WAN Edge router in DC1
      address-family ipv4 unicast
        route-map INCREASE-LOCAL-PREF-FOR-WAN-ROUTES in
```

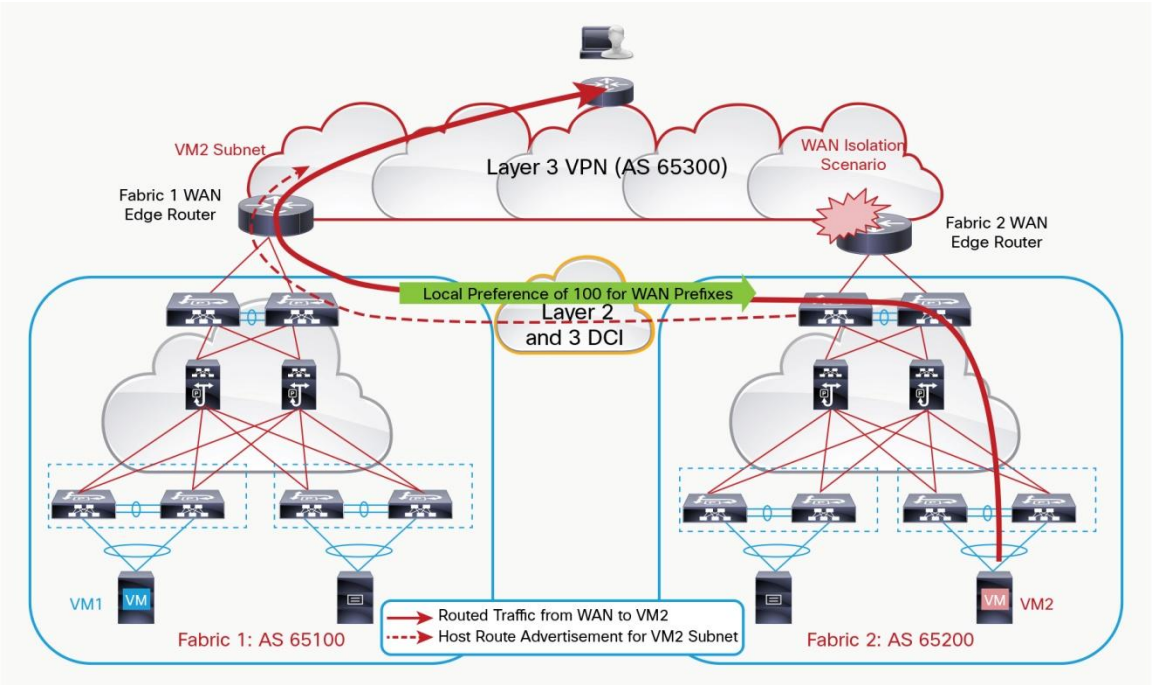
The same external prefixes received through the Layer 3 DCI connection would instead have the default local-preference value of 100, so the local path will always be preferred (steered with the local preference 200), as shown in Figure 26.

Figure 26. Local Preferences for Outbound Traffic



In the WAN isolation scenario previously considered, the border nodes in isolated VXLAN fabric 2 would start using the external prefixes received on the Layer 3 DCI connection from the border nodes in VXLAN fabric 1. This behavior still helps ensure that inbound and outbound communication between VM2 and the WAN remain symmetrical, which you can verify by comparing previous Figure 25 to Figure 27.

Figure 27. Outbound Traffic in a WAN Isolation Scenario



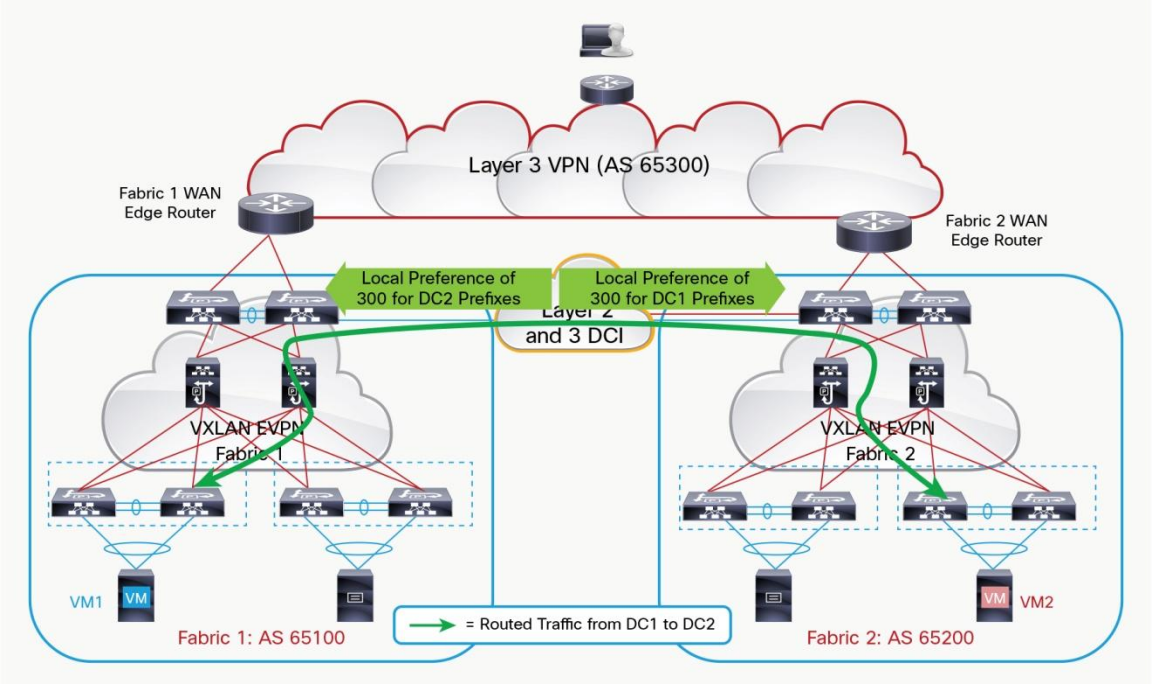
Keeping Interfabric Routing Through Layer 3 DCI

The last requirement is to help ensure that all communications between endpoints belonging to different VXLAN fabrics preferably are established using the dedicated Layer 3 DCI connection. This route is desirable because this connectivity usually is characterized by lower latency and higher bandwidth than the path through the MAN or WAN.

To meet this requirement, you must help ensure that even if a prefix (host route or IP subnet) belonging to fabric 2 is received in fabric 1 through the WAN edge router, this latter information is considered less preferable than the information received through the Layer 3 DCI connection.

Recall the configuration previously discussed to optimize the egress traffic flows. In this case, all the routes received by the border nodes from the WAN edge routers are characterized by a local-preference value of 200. You then must add a route map (**INCREASE-LOCAL-PREF-FOR-REMOTE-HOST-ROUTES**) to the routing updates received from the remote border nodes to help ensure that all the IP prefixes belonging to the remote fabric (that is, received by eBGP updates originating from the autonomous system of the remote data center) have a higher local-preference value (300 in the example in Figure 28).

Figure 28. Increase the Local-Preference Value to Keep Interfabric Routing Through Layer 3 DCI



The following sample shows the required configuration.

```
ip as-path access-list 2 permit "^65200$”
!
ip access-list ANY
  10 permit ip any any
!
route-map INCREASE-LOCAL-PREF-FOR-REMOTE-FABRIC-ROUTES permit 10
  match as-path 2
```

```
    set local-preference 300
route-map INCREASE-LOCAL-PREF-FOR-REMOTE-FABRIC-ROUTES permit 20
  match ip address ANY
!
router bgp 65100
  vrf Tenant-1
    neighbor 172.16.1.1
      remote-as 65200
      description eBGP Peering with BL Node 1 in DC2
      address-family ipv4 unicast
        route-map INCREASE-LOCAL-PREF-FOR-REMOTE-FABRIC-ROUTES in
```

As a result of this configuration, all the intersite communication stays on the Layer 3 DCI connection and will use the MAN or WAN path only if this connection completely fails.

Conclusion

VXLAN EVPN multifabric is a hierarchical network design consisting of individual fabrics interconnected together. The design described in this document focuses on the individuality of the data center domains, allowing independent scale and, more important, independent failure domains. The connectivity between the individual fabric domains is independent of the connectivity being used in the data center, and thus a natural separation is achieved.

OTV is the recommended technology for providing Layer 2 extension while maintaining failure containment. With this capability and the additional attributes that OTV offers for DCI, modern data center fabrics can be extended in an optimized way.

A variety of solutions can be used to extend multitenant Layer 3 connectivity across fabrics, mainly depending on the nature of the transport network interconnecting them.

Specific deployment considerations and configurations can be used to keep inbound and outbound traffic flows symmetrical. Symmetry is desirable to optimize access to data center resources, which may be spread across different fabrics. Symmetry is mandatory when independent sets of stateful network services (such as firewalls) are deployed in separate fabrics.

For More Information

http://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/white-paper-listing.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.

Printed in USA C11-738503-00 03/17