

# **Comparing Layer 3 Gateway & Virtual Machine Traffic Optimization (VMTO) For EVPN/VXLAN And EVPN/MPLS**

## **Tech Note**

Juniper Networks, Inc.  
1133 Innovation Way  
Sunnyvale, California 94089  
USA  
408-745-2000  
[www.juniper.net](http://www.juniper.net)

Copyright © 2016, Juniper Networks, Inc. All rights reserved.

## Table of Contents

1. Introduction.....	5
1.1 Overview.....	5
2. Using an IRB Interface as a Layer 3 Gateway in EVPN/MPLS Environments.....	7
2.1 IRB Interface IP Address Configuration .....	7
2.2 IRB Interface MAC Address Configuration.....	8
2.2.1 System Defined IRB Interface MAC Address Configuration.....	8
2.2.2 Statically Defined IRB Interface MAC Address Configuration .....	9
2.2.3 Caveats with Using Statically Defined IRB Interface MAC Addressing.....	10
3. Using an IRB Interface as a Layer 3 Gateway in EVPN/VXLAN Environments.....	12
3.1 MX Series Routers as the Layer 3 Default Gateway Interworking with a Layer 2 VXLAN TOR Switch ....	12
3.1.1 ARP Behavior on an EVPN/VXLAN Layer 3 Gateway .....	13
3.2 Enhancements to EVPN/VXLAN Layer 3 Gateway Operation.....	13
3.2.1 Type-2 Route (MAC+IP) Advertisements .....	13
3.2.2 Proxy (MAC+IP) Overview and Explanation.....	15
3.3 Other Enhancements of virtual-gateway-address for EVPN/VXLAN .....	17
3.4 Scale Limitation of virtual-gateway-address .....	17
4. EVPN/MPLS Enhancements .....	18
4.1 EVPN/MPLS IPv6 IRB Support.....	18
5. EVPN/MPLS and EVPN/VXLAN Caveats.....	19
5.1 EVPN/MPLS Specific Caveats or Restrictions .....	19
5.2 Common Virtual-Gateway-Address Caveats.....	19
5.2.1 Pinging the Virtual-Gateway-Address from CE Devices.....	19
5.2.2 Routing Between PE Devices and Between CEs and PE Devices .....	19
6. VMTO for EVPN/MPLS and EVPN/VXLAN .....	20
6.1 Optimizing Egress Traffic with VMTO.....	20
6.2 Optimizing Ingress Traffic with VMTO.....	21
6.3 The Current Issue with VMTO when Using EVPN/VXLAN .....	22
7. Conclusion .....	23
8. Additional Information .....	24

## Table of Figures

FIGURE 1: LAYER 3 GATEWAY FOR EVPN/VXLAN .....	5
FIGURE 2: LAYER 3 GATEWAY FOR EVPN/MPLS .....	5
FIGURE 3: MAC + IP ADVERTISEMENT EVPN/VXLAN WITH LAYER 2 / LAYER 3 PE MIX .....	16
FIGURE 4: OVERVIEW OF A NON-VMTO AND VTMO TOPOLOGY .....	20
FIGURE 5: EGRESS TRAFFIC OPTIMIZATION WITH VMTO .....	21
FIGURE 6: INGRESS TRAFFIC OPTIMIZATION WITH VMTO .....	22
FIGURE 7: INGRESS TRAFFIC PATH (NON-OPTIMIZED) WITH EVPN/VXLAN .....	23

# 1. Introduction

Ethernet VPN (EVPN) is a flexible solution that uses Layer 2 overlays to interconnect multiple edges or virtual machines (VMs) within a data center. EVPN delivers a wide range of benefits—including greater network efficiency, reliability, scalability, VM mobility, and policy control—that directly impact the bottom line of both service providers and enterprises.

There have been many questions on a particular topic related to EVPN site multi-homing. The redundant Layer 3 gateway (GW) behavior in both VXLAN and MPLS topologies, and Juniper Networks platforms' capability to support them.

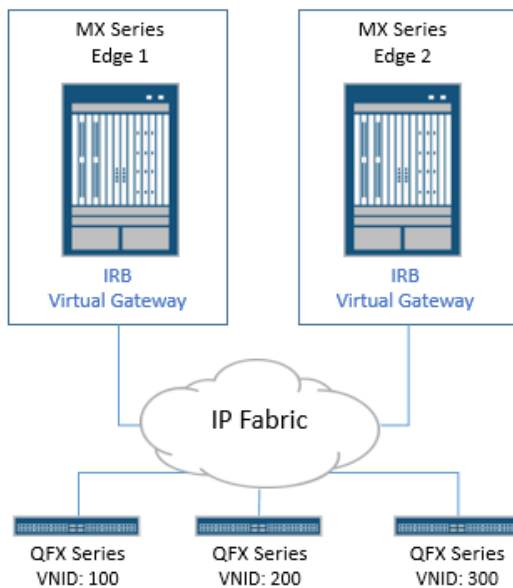
This paper will provide an overview of EVPN and describe the different ways one can configure redundant GW capabilities for EVPN with integrated routing and bridging (IRB) interface support for Virtual Extensible LAN protocol (VXLAN) and EVPN-Multiprotocol Label Switching (MPLS) deployments. The evolution of the gateway's capabilities, along with the benefits and caveats will be covered.

Finally, the paper will detail the differences in capabilities of virtual machine traffic optimization (VMTO) when using EVPN/VXLAN versus EVPN/MPLS.

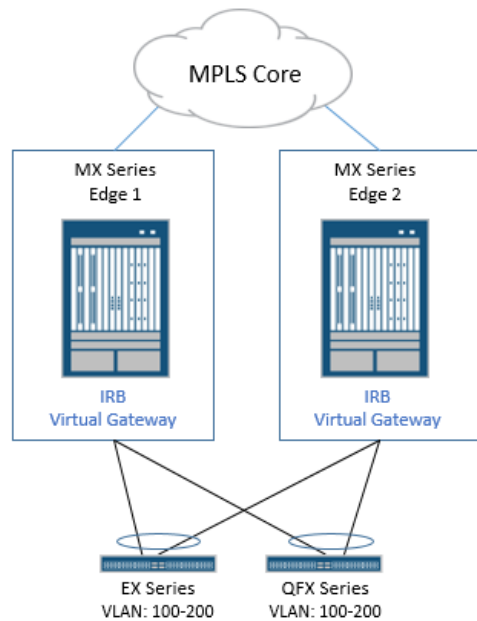
## 1.1 Overview

Currently, there are two methods you can use to configure Ethernet VPN (EVPN):

1. Using an encapsulation of MPLS
2. Using an encapsulation of Virtual Extensible LAN (VXLAN)



**Figure 1: Layer 3 Gateway For EVPN/VXLAN**



**Figure 2: Layer 3 Gateway For EVPN/MPLS**

Much consideration must be paid to the Layer 3 GW feature, which has both evolved over time and has had some differences in supported features and operations.

There are two main ways of configuring Layer 3 GW functionality on a Junos device:

1. Configure IRB interfaces directly and advertise these as GW addresses.
2. Configure an IRB interface, along with a virtual gateway address, as the default IPv4 or IPv6 address for the gateway which Layer 3 hosts to support the IRB interface redundant gateway function.

So how did this come to be?

Initially when EVPN and Layer 3 gateway functionality were conceived, some basic assumptions were made, and RFC requirements were to be followed. These were:

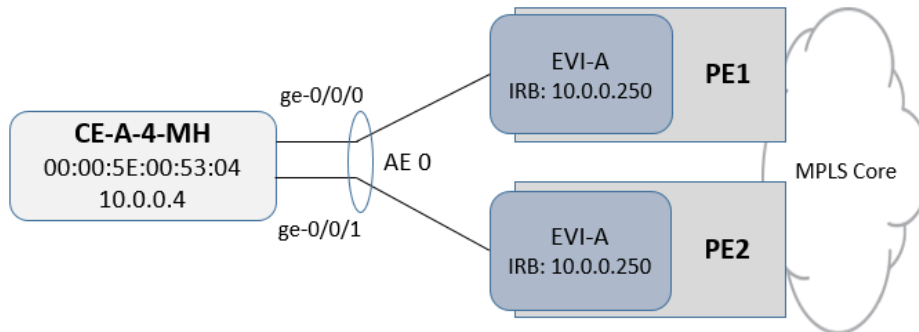
1. All PE's for an EVPN instance must have an IRB configured.
2. All PE's should have the same IP address for the GW. From the RFC, if there is a discrepancy between the GW IP addresses, an error is logged. Though it must be noted that different addresses can still be configured as both MAC/IP for advertisement to remote provider edge (PE) devices and are installed on all participating PE devices.

Where EVPN/VXLAN differed in its topology, thus causing a "diversion" from the EVPN RFC, was that not all devices within the fabric have a configured Layer 3 interface. For example, QFX5100 switches, which can only provide Layer 2 VXLAN gateway functionality, have no IRB configured in an EVPN/VXLAN topology.

As a result, the virtual gateway address (VGA) concept was developed. and must be configured in EVPN and VXLAN Data Center deployments where there is a mix of Layer 2-only and Layer 3-capable PE devices. With this VGA configuration, the communication issues that can occur when only certain EVPN PEs in a spine and leaf overlay topology have an IRB configured are eliminated.

## 2. Using an IRB Interface as a Layer 3 Gateway in EVPN/MPLS Environments

First, let's look at the EVPN MPLS use case, where it is assumed either all PE devices have an IRB configured, or none do.



### 2.1 IRB Interface IP Address Configuration

For each PE device, the IRB interface serves as the default IP gateway for each CE device connected to the PE device's access-side interfaces.

Consider the following configuration for device CE-A-4-MH in the topology pictured above:

```
chassis {
  aggregated-devices {
    ethernet {
      device-count 1;
    }
  }
}
interfaces {
  ge-0/0/0 {
    gigether-options {
      802.3ad ae0;
    }
  }
  ge-0/0/1 {
    gigether-options {
      802.3ad ae0;
    }
  }
  ae0 {
    vlan-tagging;
    mac 00:00:5E:00:53:04;
    unit 0 {
```

```

        vlan-id 100;
        family inet {
            address 10.0.0.4/24;
        }
    }
}
routing-options {
    static {
        route 0.0.0.0/0 next-hop 10.0.0.250;
    }
}

```

CE-A-4-MH is connected to the multi-homing PE devices using a static LAG interface, ae0, consisting of the underlying links to PE 1 via ge-0/0/0 and PE 2 via ge-0/0/1. The CE device is configured with the IP address 10.0.0.4 on the 10.0.0.0/24 subnet, and the IP address 10.0.0.250 is the default IP gateway for any traffic destined to subnets other than 10.0.0.0/24.

Traffic flows from the CE device to multi-homing PE devices are load balanced across the two child links of the ae0 LAG interface based on a hash of various header fields of the packets making up the flows. When starting to send inter-subnet IP traffic, the first step is for CE-A-4-MH to send an ARP request for the default gateway IP address.

Due to the nature of the LAG load balancing, the ARP request could be sent on the link to either PE 1 or PE 2. Both PE devices need to be able to reply as the default gateway and should be configured with the same IP address (10.0.0.250) on their IRB interfaces.



Multi-homing PE devices attached to the same multi-homed site should have the same IP address configured on their respective IRB interfaces.

## 2.2 IRB Interface MAC Address Configuration

EVPN IRB interfaces support two styles of configuration: system defined, and statically defined.

### 2.2.1 System Defined IRB Interface MAC Address Configuration

In the first style of configuration, each IRB interface on each PE device obtains a system-defined MAC address from the chassis MAC address pool. Using this method, these MAC addresses will be unique across the network within an EVPN instance. Therefore, every PE device will have a different address for its IRB interface inside a given EVPN instance. IRB interface MAC addresses from each PE device are synchronized in the EVPN control plane using the default gateway extended community.

Normally, an EVPN PE device will route incoming packets at Layer 3 instead of switching at Layer 2 when the destination MAC address of the incoming packet's Ethernet header matches the MAC address of the IRB interface of the receiving PE device. When an EVPN PE device receives a MAC advertisement from a remote PE device that includes the default gateway extended community, the receiving PE device will treat the received MAC address as equivalent to its own IRB interface MAC address for the purposes of IRB processing.

As described earlier, for CE-A-4-MH to send inter-subnet IP traffic it must first send an ARP request for the IP default gateway address (10.0.0.250). The ARP request will only be sent over one of the LAGs child links based on the LAG hashing parameters. The IRB interface MAC address of the responding PE will be associated with the gateway IP address in CE-A-4-MH's ARP cache.



CE-A-4-MH is able to load balance inter-subnet IP traffic flows across the links to both PE 1 and PE 2 simultaneously. However, regardless of which link the traffic is sent on, the destination MAC address in the Ethernet header of the packets being sent from CE-A-4-MH will always be the MAC address of the PE device that initially responded to the ARP request. If the destination MAC address of the traffic is set to PE 1's IRB interface MAC address, PE 2 will accept that traffic and route it as if it had been sent with PE 2's IRB interface MAC address. This is handled by the default gateway synchronization through the EVPN control plane between the PE devices.

Configurations that rely on the default gateway MAC synchronization require the least manual configuration, but this style has a weakness when a node fails. Given that the IRB MAC equivalency is learned via the EVPN routes exchanged between the PE devices, when a PE device goes down the remaining PE devices will remove all routes from the failed PE device. This includes the IRB interface MAC addresses that had been advertised with the default gateway extended community, breaking any IRB MAC equivalency.

If CE-A-4-MH had initially resolved the IP default gateway address (10.0.0.250) to PE 1's MAC address and then PE 1 fails, PE 2 will delete the routes previously advertised by PE 1. This will cause PE 2 to stop honoring packets sent with PE 1's IRB interface MAC address for Layer 3 routing. The inter-subnet packets will then be switched in the Layer 2 domain by default, breaking the inter-subnet flow for CE-A-4-MH. This causes the end hosts to re-arp for the default gateway MAC address, and depending on the end host OS this may take some time to refresh.



**Warning: IRB MAC equivalency for remote IRB interface MAC addresses is lost when the MAC routes which advertised them are withdrawn or deleted by the remote PE device.**

### 2.2.2 *Statically Defined IRB Interface MAC Address Configuration*

To overcome the node failure limitation described above, the second style of IRB interface configuration uses a static, user-defined MAC address configured on each IRB interface on every PE device in the EVPN instance.


Consider the following IRB interface configuration from PE 2 (which is replicated on PE 1):

```
user@PE2# show interfaces irb
unit 0 {
    family inet {
        address 10.0.0.250/24;
    }
    mac 00:00:5E:00:53:55;
}
```

The static, user-defined MAC address 00:00:5E:00:53:55 could be considered an "anycast MAC address" which can be configured at multiple points in the network. CE devices will always use the local PE device configured with the common IRB interface MAC address as the default gateway for Layer 3 traffic. Using this method, since each PE device within the EVPN instance has the same MAC address configured on its IRB interface, there is no longer a need to dynamically synchronize the IRB interface MAC addresses through the control plane by advertising them with the default gateway extended community.

This style of configuration requires more initial configuration time because the common, user-defined MAC must be statically configured on each PE device participating in the EVPN instance. However, it enables better fault tolerance because the CE device will use the same IRB interface MAC address when sending inter-subnet traffic regardless of where the CE device is located, or which multi-homed PE device is receiving the traffic.

Using the case discussed above, if PE 1 and PE 2 is now configured with the same static IRB interface MAC address and PE 1 fails, then PE 2 continues to forward inter-subnet traffic even without the default gateway MAC synchronization in the control plane because the IRB interface MAC address that the CE device resolves initially via ARP is present on both PE 1 and PE 2.

 Note: Statically configuring the same MAC address on the IRB interface of all PE devices in an EVPN instance is recommended to protect against node failure.

Junos OS Release 14.1R4 and later includes a configuration option to disable advertisement of IRB interface MAC addresses in the EVPN control plane in cases where when the user has configured a common MAC address on the IRB interfaces of all of PE devices in an EVPN instance:

```
user@PE2# show routing-instances
ALPHA {
  instance-type virtual-switch;
  route-distinguisher 10.255.0.2:100;
  vrf-target target:100:100;
  protocols {
    evpn {
      default-gateway do-not-advertise;
      extended-vlan-list 100;
    }
  }
  bridge-domains {
    ONE {
      domain-type bridge;
      vlan-id 100;
      interface ae0.0;
      interface ge-0/0/2.0;
      routing-interface irb.0;
    }
  }
}
```

Configuring the common, static MAC address provides both benefits of scale and synchronizing the default gateway MAC/IP addresses in large deployments.

### 2.2.3 Caveats with Using Statically Defined IRB Interface MAC Addressing

So all is good? Not quite. As was the situation in 14.1R4 for EVPN/MPLS, some issues remain.

The first issue is that configuring an IRB interface on every PE device is mandatory; second, the implementation does not support IPv6; third, the solution does not allow for any type of dynamic services such as a routing protocol or multicast to be enabled on the IRB interface. The IRB interface can *only* be used as a gateway for end hosts multi-homed to the PE devices.

As it stood, this was a very similar situation to the MC-LAG A/A usage of MAC sync between platforms, which helped greatly with scale, removing VRRP's limitation of 1000 sessions. But it also removed the option of running OSFP, PIM, or

IGMP between MC-LAG peers to upstream or downstream peers. In the case of MC-LAG, behavior was enhanced to do two things:

1. Allow routing protocols to run on the primary IP address of the IRB interface, and both of the MC-LAG peers to run routing protocols independently. The routing protocols use the primary IP address of the IRB interface and the IRB MAC address to communicate with the MC-LAG peers. The IRB MAC address of each MC-LAG peer is replicated on the other MC-LAG peer and is installed as a MAC address that has been learned on the inter-connect link between the MC-LAG peers.
2. At any given time, one of the VRRP devices is the master (active), and the other is a backup (standby). The default behavior of a VRRP backup node is to not forward incoming packets. However, when VRRP over IRB is configured in an MC-LAG active/active environment, both the VRRP master and the VRRP backup forward Layer 3 traffic arriving on the MC-AE interface. If the master fails, all the traffic shifts to the MC-AE link on the backup.


For EVPN, a similar solution to MC-LAG is required to enable dynamic protocols to use the EVPN A/A multi-homed topology. This also allows for L2-only PE devices to interwork with Layer 3-enabled PE devices (Layer 3-enabled PE devices are those with IRB interfaces configured.)

The routing protocols will be discussed further later in the document. For now, let's move on to EVPN/VXLAN. This is where the virtual gateway address was developed because of the Layer 2-only nature of the top of rack switches and end nodes hosting Layer 2-only capable hypervisor virtual switches.

### 3. Using an IRB Interface as a Layer 3 Gateway in EVPN/VXLAN Environments

The primary pain point with the current generation of VXLAN overlay designs is the capability of the devices at the top-of-rack (TOR) switch layer, the dominant issue being that they can only provide Layer 2 VXLAN gateway capabilities. As such, they do not advertise the IP address along with EVPN MAC route information for the bare metal server (BMS) that is attached to it. Therefore, no BMS host route is installed in the L3-VRF gateway device (such as an MX Series router) as a result of this EVPN MAC route advertisement from the Layer 2-only PE platform. However, ARP routes for the BMSs are installed in the kernel on L3-VRF gateway if the L3-VRF gateway receives ARP responses from BMSs.

This effectively means that MX and QFX Series devices rely on ARP/NDP to discover and install the MAC/IP bindings, which furthermore means each IRB *must* have its own interface MAC address to identify itself (for ARP) and virtual MAC address for gateway functions.

 The Contrail vRouter supports proxy ARP and advertises the IP address with the EVPN MAC route for its VM. Both Contrail vRouters and MX Series routers that host the L3-VRF for a virtual network have all the VM host routes for VMs residing in the same VN as well as in all other VNs. Traffic between the VMs, either intra-VN or inter-VN, is forwarded directly at Layer 3 between the Contrail vRouters. That is, Contrail vRouters are effectively Layer 3-capable PE devices and advertise the MAC/IP addresses.

#### 3.1 *MX Series Routers as the Layer 3 Default Gateway Interworking with a Layer 2 VXLAN TOR Switch*

MX Series routers act as the default gateway for inter-VN traffic whenever a BMS is installed through its IRB interface. To provide default gateway functionality, each IRB interface is assigned a pair of IP addresses: a unique address for the IRB interface, and an anycast IP address (referred to as the virtual gateway address (VGA).) Thus, for each IP subnet in the virtual network (VN) there is a corresponding pair of MAC addresses (a VGA MAC and an IRB MAC.) The VGA MAC address is locally unique within a VN and is the same on all Layer 3 PE devices.

The IP anycast (VGA) address along with its VGA MAC address is used to provide the default gateway function. Each BMS is configured to use the IRB interface's anycast IP address related to VN that the BMS belongs to as its default gateway address. The reason for using anycast IP and VGA MAC addresses is to achieve redundant default gateways (to be discussed further in the next section).

Each IRB interface is also assigned its own unique IP address and MAC address. These addresses are used for initiating ping and ARP requests. It is essential that each IRB interface has its own IP and MAC addresses so that asymmetric data paths for the ARP request and response can be avoided when the IRB interface ARPs for a host's destination MAC address.

Following on from above, the MX device advertises two EVPN MAC routes with its corresponding IP address associated with the IRB interface. However, there is no default gateway extended community attribute associated with the IRB interface's MAC route advertisement. This is because a Layer 2-only PE device does not support IRB and Layer 3 functionality today, so the remote IRB interface's MAC address has to be installed in the MAC forwarding table on the Layer 2-only PE device.

In summary, in this scenario there is no host MAC/IP route for any BMS connected to a Layer 2-only EVPN PE device because the PE device does not advertise MAC addresses with IP addresses.

This forwarding behavior is different than the normal forwarding behavior for Junos OS software. In particular, when traffic is forwarded from the MX device to the Layer 2 PE device the GW PE device does not use the composite next-hop with a

Layer 2 rewrite string. Instead it relies on the ARP route in the kernel to get the destination MAC address of attached hosts, then uses the next-hop associated with the ARP entry. The packet is then sent out with VXLAN encapsulation through EVPN. For the ARP route and its next-hop, please refer to the next section.

As a final note, the Layer 3 GW PE devices are configured with the same VGA IP address and have the same VGA MAC address, so at a very high level the redundant gateway is achieved through the EVPN all-active multi-homing feature by associating all IRB interfaces for a given VN with the same Ethernet Segment ID (ESI). To the other EVPN PE devices (Contrail and Layer 2-only PE devices) the VGA MAC address for the IRB interfaces is multi-homed to the Layer 3 GW PE devices through the same Ethernet Segment (i.e. in the same VLAN).

Per normal EVPN all-active functionality, a remote PE device builds an ECMP next-hop path to reach the IRB interface's VGA MAC address or VGA IP address based on the IRB interface's MAC route advertised by each Layer 3 GW. Traffic destined to the IRB is load balanced between the Layer 3 PE devices. If one Layer 3 EVPN GW suffers a failure, all remote PE devices are notified through the withdrawal or purge of the IRB MAC route. Then all remote PE devices update their next-hop address to the IRB interface's VGA MAC address or VGA IP address to exclude the path to the failed GW. Since the IRB interface's VGA MAC and IP addresses are still reachable through the updated next-hop, there will be no change to the ARP entries on the hosts connected to the Layer 2-only PE devices.

### 3.1.1 **ARP Behavior on an EVPN/VXLAN Layer 3 Gateway**

Given ARP is fundamental to the operation of an EVPN/VXLAN Layer 3 GW, let's take a moment to understand how it works, and more importantly in the following sections, how this influences enhancements in the GW's behavior.

A dynamic ARP route/entry is always created by the kernel if the Layer 3 GW receives an ARP packet from a remote PE device. The ARP entry points to a VXLAN ARP encapsulation next-hop and the kernel is responsible for creating the VXLAN ARP encapsulation next-hop with the target next-hop pointing to the remote VTEP IP address. When the remote VTEP IP address becomes unreachable, its corresponding remote VTEP logical interface is deleted. As a result, the kernel also deletes the corresponding dynamic ARP entries.

When a host is connected to the Layer 2-only PE device and it wants to ARP for its default gateway's MAC address, the host initiates an ARP for the IRB interface's VGA IP address. The ARP packet arrives at one of the redundant Layer 3 GW PE devices. The GW that receives the ARP request learns the host's IP/MAC binding and creates an ARP entry. The GW then sends an ARP reply with the VGA's MAC address set as the inner MAC and the outer MAC as the IRB interface's own MAC address (discussed later in this section). The ARP request is unicast back to the Layer 2 PE device with the VXLAN encapsulation. The ARP packet is de-encapsulated at the Layer 2 PE device and forwarded to the host.

For the Layer 3 routed inter-VN traffic destined to the host, when traffic is from another VN to the VN where the destination host belongs, the GW looks for the host's ARP entry. If no match is found, the MX device initiates an ARP request with the IRB interface's own IP/MAC address. The host learns the IRB interface's IP/MAC binding and adds or refreshes the ARP entry based on the ARP request packet from the Layer 3 GW. The ARP response from the host can arrive on any of the connected multi-homed Layer 2 PE devices, and it has to have a destination MAC (DMAC) of the IRB interface's own MAC address. So the ARP request is unicast back to the Layer 3 GW MX device that initiated the ARP request.

## 3.2 **Enhancements to EVPN/VXLAN Layer 3 Gateway Operation**

You might now consider the solution for EVPN VXLAN complete. However, there have been some recent additions to the capabilities that bring the complete solution very close.

### 3.2.1 **Type-2 Route (MAC+IP) Advertisements**

The first enhancement was to add the capability for EVPN/VXLAN to advertise the MAC+IP Type-2 routes between participating PE devices, making it functionally the same as the EVPN/MPLS implementation.

This may sound simple, but because not all of the PE devices are Layer 3-capable and the MAC+IP learning is implemented through ARP/NDP mechanisms, there are significant challenges in implementing this successfully.

In the case of EVPN/VXLAN, it also differs from EVPN/MPLS Layer 3 GW because it's configured using the **virtual-gateway-address** statement, whereas with EVPN/MPLS before Junos OS Release 14.2R5, you were required to configure the same IRB/MAC on all PE devices.

Recall that in the solution for EVPN/MPLS, the same MAC+IP address is configured on the IRB interface of each PE device, hence there is no need to proxy for the other Layer 3 PE device's gateway addresses. So the advertisement of the default gateway can be disabled with **default-gateway do-not-advertise** statement under the **protocols evpn** stanza. In the case of EVPN/VXLAN, this is not so; although, the VGA IP address must be configured the same on all Layer 3 PE devices, it does not exist on the Layer 2-only PE devices. Therefore, in the EVPN/VXLAN topology where not all PE devices are L3-capable, the advertisement of the default gateway should *not* be disabled. This allows for the Layer 2-only PE devices to learn the GW.

There is another option that should be configured on the Layer 3 GW PE devices for EVPN/VXLAN with Layer 2 PE devices. The **default-gateway no-gateway-community** statement under the **protocols evpn** stanza ensures that the "router" MAC address of the Layer 3 PE devices, that is the MAC address of the VGA and the IRB interface MAC address, are advertised to the Layer 2 PE device, but without the extended community option of default-gateway. Now the Layer 2 PE device can install the MAC addresses, but will not try to proxy-answer for them. It will instead just forward traffic onward to the MAC address owner, the Layer 3 PE device. This configuration is required on the current generation of Layer 2-only capable PE devices, for example QFX5100 devices, but can be used in EVPN instances where the leaf PE devices are Layer 2-capable only.

Aside from these differences, there is one more extremely important enhancement required for an EVPN/VXLAN topology with a combination of Layer 2 and Layer 3 PE devices: proxy MAC+IP advertisement capabilities on the GW PE devices. Support for this is introduced for EVPN-VXLAN to support platforms that can only support Layer 2 for VXLAN and cannot natively support Layer 3 for VXLAN.

In this mode, the redundant Layer 3 GW learns the MAC+IP binding via ARP/NDP, and subsequently advertises the MAC+IP Type-2 routes with the BGP next-hop set to the EVPN PE device that initially advertised the MAC route. Or more simply, it enables the MAC+IP Layer 3 capability required for Type-2 routes on behalf of the Layer 2-only PE devices.

Proxy MAC+IP advertisement is enabled with the following command:

```

irb {
  unit 0 {
    proxy-macip-advertisement;
    family inet {
      address 10.34.0.1/24;
    }
    mac 00:00:5E:00:53:00;
  }
}

```

There are two main challenges with the current overlay environment with a mix of Layer 2-only and Layer 3 / Layer 2-capable PE devices without the **proxy-macip-advertisement** statement enabled:

- If the GW PE devices don't advertise on behalf of the MAC+IP for the Layer 2-only PE devices, then Type-2 route advertisements would not be supported at all in this solution. The EVPN remote routers (MX3 in Figure 3 below) cannot learn the MAC+IP through ARP because it's not supported over an EVPN pseudo-wire.
- Even if one of the redundant GWs is rebooted and comes back up, traffic is directed toward the redundant GW but it has to resolve the IP to MAC binding using ARP/NDP, and this process takes time. There is a traffic loss even though the other redundant Layer 3 GWs are up and running and can pass traffic. The same problem is seen when a new node is added as a redundant Layer 3 GW to increase the capacity of the network.

The **proxy-macip-advertisement** statement also enables another important piece of functionality in the advertisements of MAC+IP routes in EVPN for VXLAN. In the EVPN/MPLS implementation, the MAC routes are advertised in one Type-2 advertisement, along with a corresponding MAC+IP Type-2 route update by default. This means that the receiving device now has these /32 host IP routes installed in its EVPN database and locally in the IRB interface's routing table (inet.0 or .vrf).

With EVPN/VXLAN, when the **proxy-macip-advertisement** statement is not enabled, only the MAC routes are sent between PE devices; when enabled, both the IP and MAC host routes are installed on the receiving PE devices.

### 3.2.2 Proxy (MAC+IP) Overview and Explanation

The figure below details the forwarding process with the **proxy-macip-advertisement** feature enabled, and we will step through the learning and forwarding in this environment.

Topology details:

- T1 and T2 – Layer 2-only EVPN-VXLAN PE devices that connect bare metal servers BMS1 and BMS2.
- MX1 and MX2 – EVPN-VXLAN PE devices providing redundant Layer 3 GW functionality for EVPN-VXLAN.
- MX3 – another Layer 3 EVPN PE device, *not* acting as the GW for T1 or T2

Notes:

- T1 and T2 advertise only MAC routes in EVPN Type-2 routes
- T1 and T2 do not advertise the MAC+IP Type-2 route
- MX1 and MX2 (redundant Layer 3 GWs) use ARP/NDP to resolve the IP to MAC bindings. Since the MAC+IP bindings are not generated by T1 and T2, each of the redundant gateway devices does ARP/NDP to resolve the bindings (BMS1 – M+IPA & BMS2 – M+IPB)

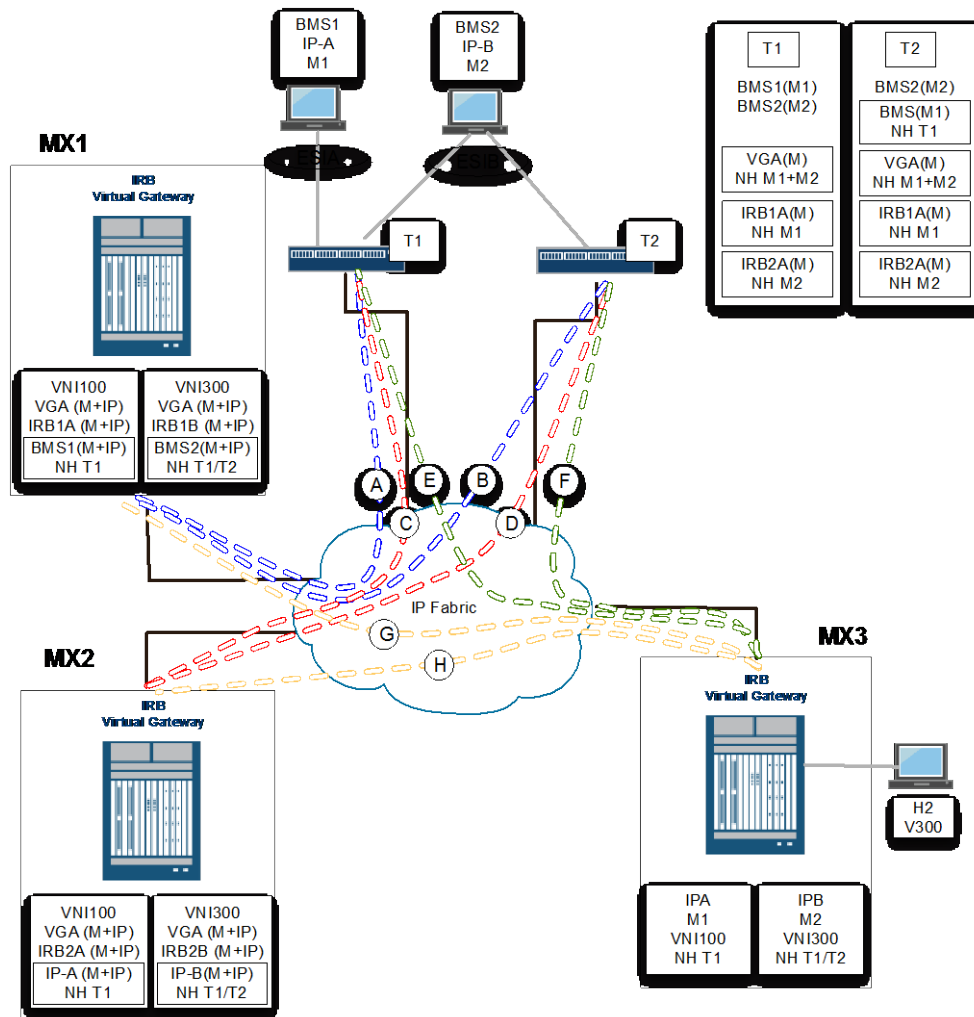


Figure 3: MAC + IP Advertisement EVPN/VXLAN With Layer 2 / Layer 3 PE Mix

Now let's step through the learning and forwarding process.

- a) A single-homed device is connected to T1 with has a MAC address, M1, and an associated IP address, IP-A.
- b) A multi-homed device is connected to T1 and T2 with has a MAC address, M2, and an associated IP address, IP-B. The Ethernet Segment Identifier (ESI) for this device is ESIB.
- c) T1 will advertise the Type-2 route for M1 (no IP address).
- d) T1 and T2 will advertise Type-2 route for M2 with ESI set to ESIB.
- e) T1 and T2 are configured as Layer 2-only EVPN devices in MX1 and MX2.
- f) The redundant Layer 3 GW PE devices will advertise the MAC+IP Type-2 route with the BGP next-hop set to T1 when the ARP resolves the IP-A binding to M1. They build this entry by snooping the ARP to map the MAC to IP binding, and use the next-hop information set in the route advertisement from T1.



- g) The redundant Layer 3 GW PE devices (MX1 and MX2) will advertise two MAC+IP Type-2 routes with the BGP next-hop set to T1 and T2 when the ARP resolves the IP-B binding to M2. They build this entry by snooping the ARP to map the MAC to IP binding, and use the next-hop information set in the route advertisement from T1 and T2.
- h) When a Layer 2-only EVPN TOR switch learns multiple router MAC addresses (via the default gateway MAC route), it should map all of them to the nearest Layer 3 EVPN PE devices. If it does not, since the device itself does not route and only looks at the router's MAC address as Layer 2 MAC, it will send it to the original router-MAC hosting PE device (which means that distributed routing will not happen.)
- i) The other Layer 3 PE device (MX3) will receive this MAC+IP from MX1 and MX2, and will install the MAC address to point to the next-hop IP address associated with this MAC+IP route. So in data path, a packet with a destination of that MAC address will be forwarded to the TOR switch instead of the MX device. There are two case scenarios to cover:
  - o For inter-VNI traffic between hosts BMS1 and BMS2 in VNI100, behind T1 and T2, and host H2 in VNI300, behind MX3, traffic will first traverse VTEP tunnels **a, b, c or d** (shown in the figure above) from T1 or T2 to the GW, MX1 or MX2 (the tunnel is selected based on hashing). The GW will then route the inter-VNI traffic to VNI300 and forward it on through tunnels **g and h** to MX3. MX3 will forward it on to host H2. Return traffic will go directly over tunnels **e or f** to the appropriate TOR switch, T1 or T2.
  - o For Layer 2-only traffic, T1 or T2 will forward directly over tunnels **e or f**.
- j) Lastly, when the MAC route is withdrawn by T1 and T2, the respective MAC+IP bindings are withdrawn by MX1 and MX2.

### **3.3 Other Enhancements of virtual-gateway-address for EVPN/VXLAN**

The final Layer 2 enhancement to mention for VXLAN is the configuration of the MAC address for the virtual gateway address, where the **virtual-gateway-v4-mac** statement is a configuration option under the IRB logical interface. This static virtual MAC address will be the MAC address used for forwarding inter-VNI traffic. This will now allow Layer 2 switches between the Layer 2-only PE devices and a host to learn the virtual MAC address and prevent flooding. Without this setting configured, the switch just learns the VRRP MAC to virtual GW IP address binding, but it never learns the VRRP MAC through the data packet.

### **3.4 Scale Limitation of virtual-gateway-address**

There is no limitation on the number of virtual gateway addresses that can be configured on an IRB interface. Every IRB interface address can have a corresponding virtual gateway address.

However, currently the maximum number of PE devices that can have the same virtual gateway address is 64.

MX240 and larger devices support 16K IRB interfaces today. This is expected to increase to 64K in the future.

## 4. EVPN/MPLS Enhancements

### 4.1 *EVPN/MPLS IPv6 IRB Support*

Starting with Junos OS Release 14.2R5, EVPN/MPLS support for IPv6 is available on GW PE devices' IRB interfaces. Along with support for IPv6, the **virtual-gateway-address** configuration options have been adopted as well. The main reason for this update was because in IPv6 the neighbor advertisement (NA) is only processed if the local router has a local address configured that matches the destination address in the received packet. For this to be assured to happen, both PE devices must have the same link-local address configured on the interfaces to towards the CE devices. This requirement necessitates that there be an IRB link-local address and a VGA address. Therefore, the concept of the virtual gateway address—originally implemented for EVPN/VXLAN Layer 2-only PE support—was adopted for EVPN/MPLS.

The following must happen for IPv6 NA to work:

- The packet will be processed only if the destination IP address in the packet is hosted locally by the PE device (this is possible only if we have configured same global and link-local IP address on PE1 and PE2)
- The IP address is configured on an IRB interface that is part of same EVPN instance.
- The “solicited” flag is set in the NA packet.

Only after all these checks, we form the IPv6 neighbor on PE2. The route is then propagated to PE1 using MP-BGP.

The configuration of two IPv6 GW PE devices is detailed below:

#### PE1

```

irb {
  unit 1 {
    family inet6 {
      address 2001:db8::192:168:15:1/120 {
        virtual-gateway-address 2001:db8::192:168:15:10;
      }
    }
    address fe80::10/120;
  }
}

```

#### PE2

```

irb {
  unit 1 {
    family inet6 {
      address 2001:db8::192:168:15:2/120 {
        virtual-gateway-address 2001:db8::192:168:15:10;
      }
    }
  }
}

```

```

        address fe80::10/120;
    }
}

```

## 5. EVPN/MPLS and EVPN/VXLAN Caveats

### 5.1 *EVPN/MPLS Specific Caveats or Restrictions*

- Currently when using EVPN/MPLS, one must configure an IRB interface on all participating PE devices with the optional **virtual-gateway-address** statement. There is no support for a mix of Layer 2-only and L2/L3 PE devices in an EVPN/MPLS environment.
- EVPN/MPLS is not currently supported on QFX10000 Series devices.
- The virtual gateway address feature can be configured with EVPN/MPLS for both IPv6 and IPv4, however one will not see the virtual gateway IP/MAC address in the EVPN database (instance), as the local MAC/IP address is not advertised to other PE devices. For IRB interfaces on EVPN/MPLS, it is required that the same virtual gateway IP address be configured on all the PE devices so there is no functional impact with the routing daemon blocking publishing of the local VIP/VMAC address to other devices.

### 5.2 *Common Virtual-Gateway-Address Caveats*

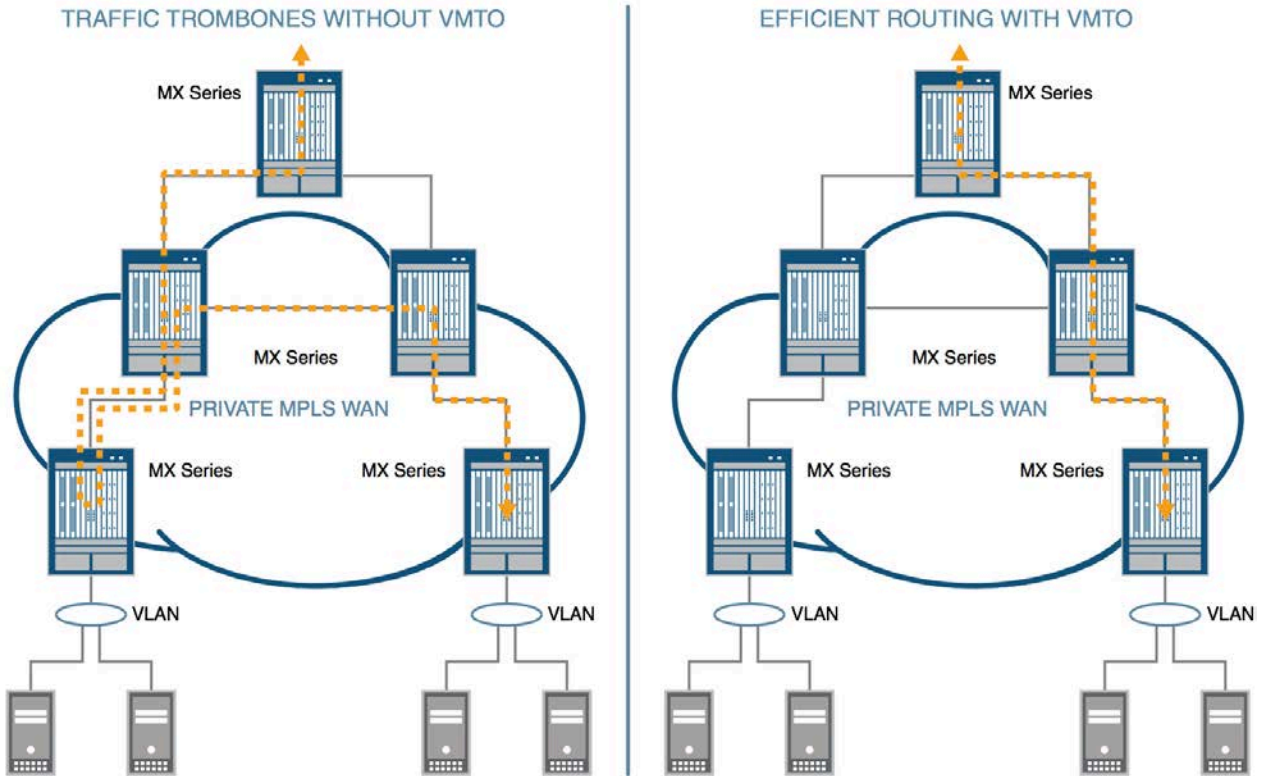
#### 5.2.1 *Pinging the Virtual-Gateway-Address from CE Devices*

Although an ARP will reach the GW devices, currently Junos software does not allow for the virtual gateway address to respond directly. This capability is the same as the accept-data feature for VRRP.

#### 5.2.2 *Routing Between PE Devices and Between CEs and PE Devices*

Currently, running a routing protocol (BGP or an IGP) is not supported in an EVPN/MPLS or EVPN/VXLAN environment. Essentially, this means one cannot enable routing protocols on the IRB interfaces on the PE devices. This limitation also includes PIM.

## 6. VMTO for EVPN/MPLS and EVPN/VXLAN



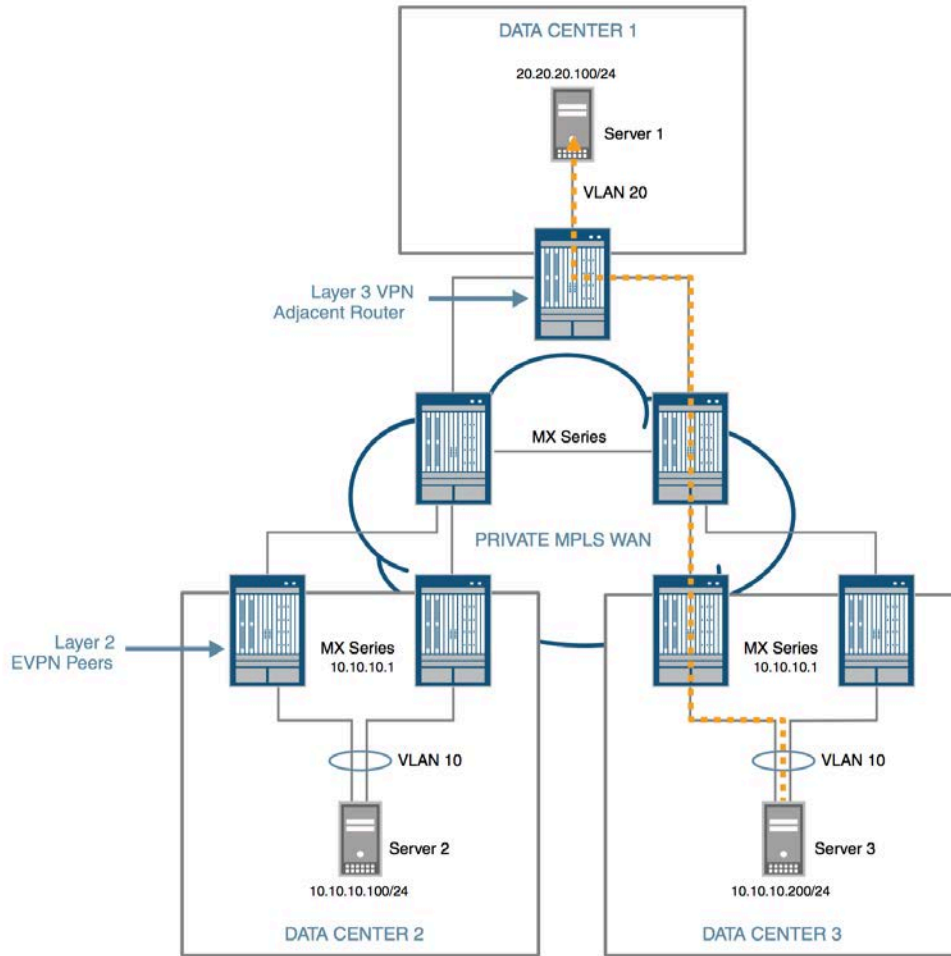
**Figure 4: Overview of a non-VMTO and VMTO Topology**

Virtual machine traffic optimization (VMTO) is the ability for Layer 2-connected data centers to route traffic optimally outbound, and for a remote Layer 3 router to route return traffic inbound directly back to a host that has been relocated.

VMTO enables a customer to migrate one server (or VM) within the same, or between different data centers, without traffic “tromboning.” Because a VM is unaware that it has moved, it does not flush its ARP table, and it continues to send inter-VLAN packets to its configured default gateway. However, after the move this default gateway may no longer be the optimal default gateway.

Figure 4 shows the difference between traffic flows without, and then with, VMTO. The diagram on the left shows how “traffic trombones” can develop between VLANs if VMTO is not implemented. For example, when a VM from one MX Series router is moved to another MX Series router, but keeps its original gateway address, this results in inefficient traffic flow and causes the traffic to loop back and forth between MX Series routers.

### 6.1 Optimizing Egress Traffic with VMTO



**Figure 5: Egress Traffic Optimization with VMTO**

To optimize egress traffic and avoid the trombone effect, packets from Server 1 to Server 3 as shown in Figure 5 are not tromboned for outbound traffic. This is achieved by configuring all GW devices to use the same IP/MAC address on either their IRB interface MAC/IP addresses or virtual gateway MAC/IP addresses. In Figure 5, Server 3 can send egress packets to either GW using VLAN 10.

## 6.2 Optimizing Ingress Traffic with VMTO

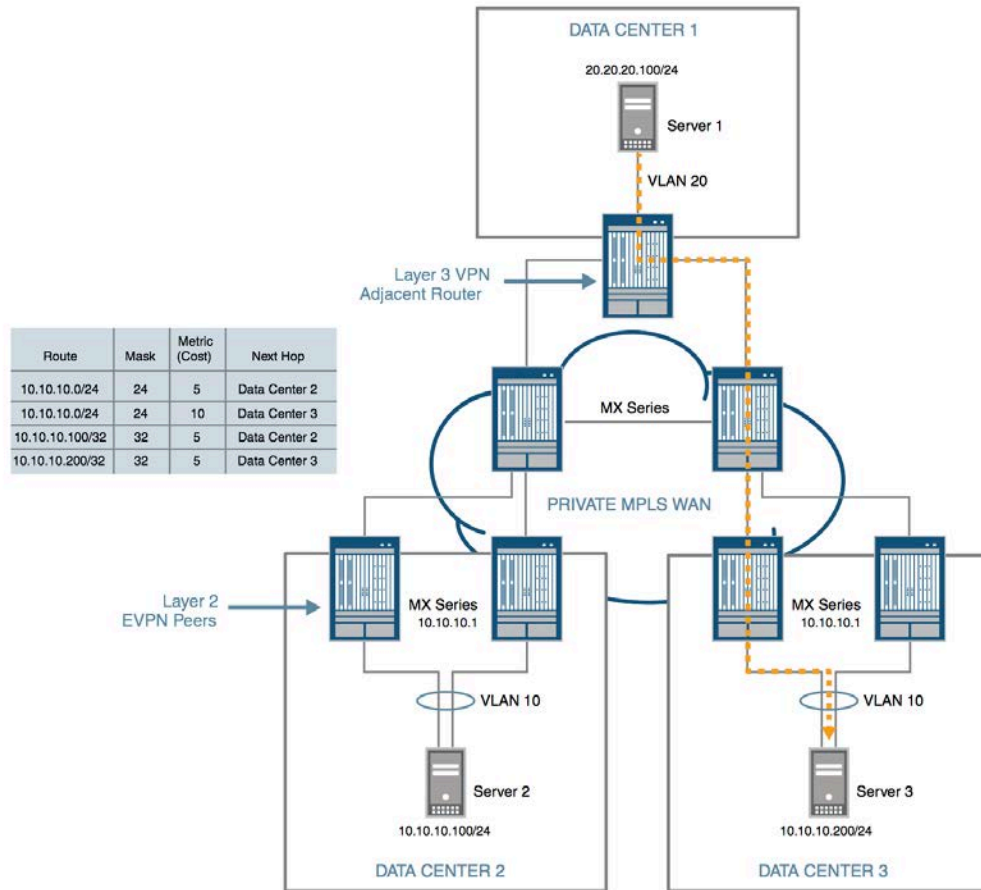


Figure 6: Ingress Traffic Optimization with VMTO

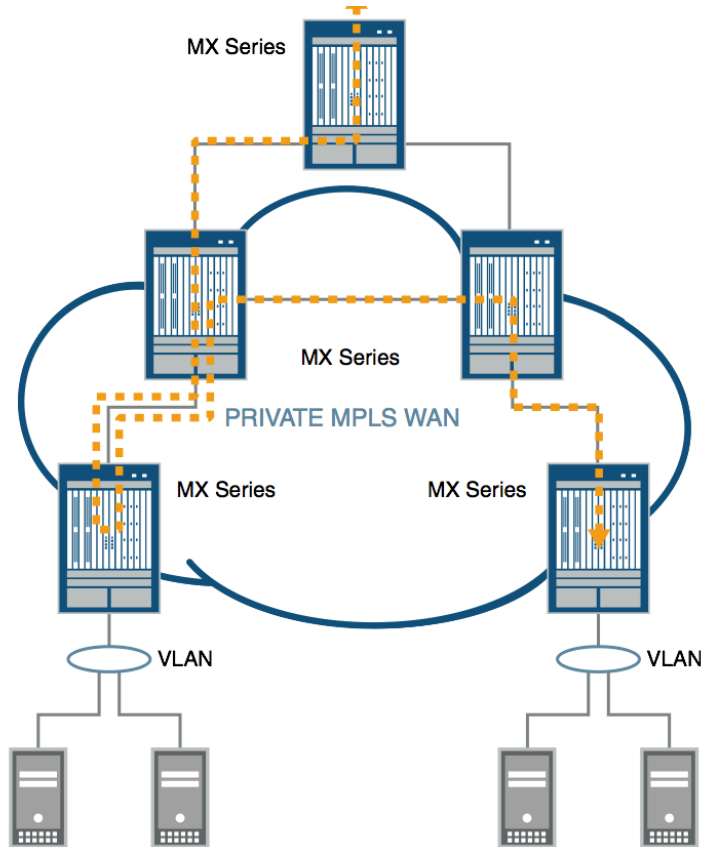
VMTO also addresses a similar problem with ingress routing. To optimize ingress traffic and eliminate the trombone effect between VLANs, use VMTO to send packets from an external Layer 3-connected site to an internal host.

In Figure 6, Server 1 seeks to send packets to Server 3 optimally. Because of the VM move, the edge router has no knowledge of the host IP address's new location. Consequently, it routes the traffic across the WAN to the original data center due to a lower-cost route for the subnet. Then the edge router at that original data center sends the packet to the destination data center (as shown earlier, in Figure 4).

VMTO addresses this inefficiency. In addition to sending a summary route of 10.10.10.0/24 (of the subnet in question) to the data center edge routers, VMTO also sends host routes that represent the location of the local servers. With VMTO, ingress traffic destined for Server 3 is sent directly across the WAN from Data Center 1 to destination Data Center 3, thus eliminating the ingress trombone effect.

### 6.3 The Current Issue with VMTO when Using EVPN/VXLAN

Now the issue currently is that this works as above for EVPN/MPLS; however, for EVPN/VXLAN there is a missing piece of functionality. With EVPN/VXLAN, Junos OS does *not* advertise the MAC+IP Type-2 host routes into internal routing tables or any VRF table.



**Figure 7: Ingress Traffic Path (non-optimized) with EVPN/VXLAN**

This being the case, the ingress routing will not be optimized when a VM moves from one DC to another because the /32 host routes will not be exported. Figure 7 above details the ingress path when using EVPN/VXLAN and routing ingress from a remote layer 3 site. In this case there will still be reachability, however traffic will flow over the suboptimal path to the DC that is primary at Layer 3 for the VLAN prefix. It will then be forward over the Layer 2 path between DCs and on to the destination host.

## 7. Conclusion

This paper described the various ways to configure redundant GW capabilities for EVPN with IRB support for EVPN/VXLAN and EVPN/MPLS deployments, including the evolution of the gateway’s capabilities, as well as benefits and caveats.

This paper also detailed the differences in capabilities VMTO when using EVPN/VXLAN versus EVPN/MPLS.

## 8. Additional Information

The following links provide further information on the topics covered in this paper.

[Configuring EVPN with IRB Solution](#)

[Day One: Using Ethernet VPNs for Data Center Interconnect](#)

[Improve Data Center Interconnect, L2 Services with Juniper's EVPN](#)