Deploying EVPN Multihoming in Data Center Networks

Table of contents

Introduction

The focus in this design guide is to support Multihoming in data center networks using EVPN All-Active Multihoming Services over a VXLAN (Virtual eXtensible LAN) overlay on Arista platforms. This design offers the following enhancements when compare to VPLS:

- Control plane learning of MAC and IP information. Contrary to existing Layer 2 VPN technologies, such as VPLS which learn only through the MAC address in the data-plane. Instead EVPN advertises MAC addresses using EVPN route messages.
- The control plane learning allows active-active forwarding into dual-homed environments due to split horizon and designated forwarder capability definitions. Again, in contrast to VPLS, which is exclusively active-standby and it lacks a capability to detect loops
- Using BGP route policies to control MAC (MAC+IP) advertisements, much like IP VPNs, unlike existing Layer 2 VPN technologies which rely on data-plane MAC filtering locally on each device.
- Support both IPv4 and IPv6 Overlay hosts

The configuration and guidance within the document, unless specifically noted, are based on the platforms and EOS releases noted in the table below.

Platform	Minimum Software Release
7280SE/7280R/7280R2/7500R/7500R2/7500E	EOS release 4.23.2F
7300X3, 7050X3	EOS release 4.23.2F
CCS-720XP	EOS release 4.23.2F

RFC and IETF Drafts Supported

Platform	Minimum Software Release
RFC 7432	BGP MPLS-Based Ethernet VPN
RFC 8365	A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)
draft-ietf-bess-evpn-inter-subnet-forwarding-05	Integrated Routing and Bridging in EVPN
draft-rbickhart-evpn-ip-mac-proxy-adv	Proxy IP->MAC Advertisement in EVPNs

Comparing Different MultiHoming Mechanisms

Arista EOS offers two different multihoming mechanisms:

1. Arista's Multi-Chassis Link Aggregation Group (MLAG)
2. BGP EVPN Based All-Active Multihoming (MH) mechanism

Arista's MLAG technology provides the ability to build a loop free active-active layer 2 topology. The default mode in EOS is EVPN single homing, when one "Vxlan Tunnel End Point" (VTEP) connects to one CE device using a single Ethernet link or a local Portchannel composed of more than one physical link. For redundancy, the CE device can be connected to an Arista standard MLAG using LAG.

The technology operates by allowing two physical Arista switches to appear as a single logical switch (MLAG domain). Third-party switches, servers or neighbouring Arista switches connect to the logical switch via a standard port-channel (static, passive or active) where physical links of the port-channel split across the two physical switches of the MLAG domain. With this configuration all links and switches within the topology are active and forwarding traffic, with no loops within the topology the configuration of spanning-tree becomes optional.

Two MLAG peers from a single MLAG domain share a single anycast VTEP IP address. In fact, the remote VTEP sees the two MLAG peers as a single VTEP. All the MAC-IP route advertisement uses zero ESI and the MLAG anycast VTEP IP for all the downstream hosts connecting to the MLAG .

BGP Based EVPN All-Active Multihoming mechanism is specified in RFC7432 which defines four types of route exchanged through EVPN. This mechanism allows network operators to connect a third-party switch or bare-metal servers to two or more PEs via Ethernet ports or standard port-channel.

The EVPN multihoming is a multi-vendor standard-based scale-out solution that does not require a dedicated peer link that allows more topology flexibility. It also supports a mass withdrawal mechanism to minimize traffic loss when the link connecting to a multihoming site or device fails.

	Arista Standard MLAG	EVPN A-A Multihoming
VTEP IP Address	MLAG peers appear as a single "virtual switch" and share the same VTEP IP.	Each A-A MH peer operates independently. Each A-A MH peer has its own VTEP IP.
Remote VTEP	The peers in the same MLAG domain share the same VTEP IP. The MLAG peers advertise all the hosts (connect MLAG link) with zero ESI (Ethernet Segment Identifier). The remote VTEP treats the MLAG peers as a single-homed PE/VTEP.	The MH peers advertise the hosts locally connecting to ES LAG using non-zero ESI. To achieve overlay ECMP, the remote VTEP must support Type 1 Auto Discovery Route. See [Arista Layer 2 VTEP EVPN VxLAN Route Type-1 Support]
Peer Sync	Both peers use MLAG protocol to synchronize the internal states and MAC table.	MH peers use the EVPN control plane to sync MAC and ARP/ND binding.
Peer-Link Requirement	A peer link is required to connect two MLAG peers. It is a common practise to place MLAG peers in the same rack.	No need for a "peer-link". Hence, A-A MH peers can be placed in a different location.
Downstream STP Support	Support STP for the downstream MH L2 switch.	The feature "Spanning Tree Network Super Root" enables multiple MH VTEP to act as a "Network STP Super Root".
Number of MH Peers	Up to 2 only	Up to 16 peers for EVPN/MPLS Up to 16 peers for EVPN/VxLAN (4.24.1F or later release)
On-boot/reload Delay	MLAG Reload Delay Mechanism	Support using Link Tracking Group

EVPN Multihoming Concepts

BGP based EVPN Multihoming allows a CE device to connect to multiple PE routers. The CE device can be a bare-metal server (BMS) or a switch. The multi homing mechanism prevents disruptions due to network connectivity issues from a PE router or a VTEP to the CE device. There are two different modes of operation for EVPN multihoming, namely (1) active-standby and (2) all-active.

Active-standby vs All-Active Multihoming

In active-standby and all-active multihoming modes, the CE device is connected to more than one VTEPs. An Ethernet Segment (ES) refers to the set of Ethernet links connecting to a multihomed CE device.

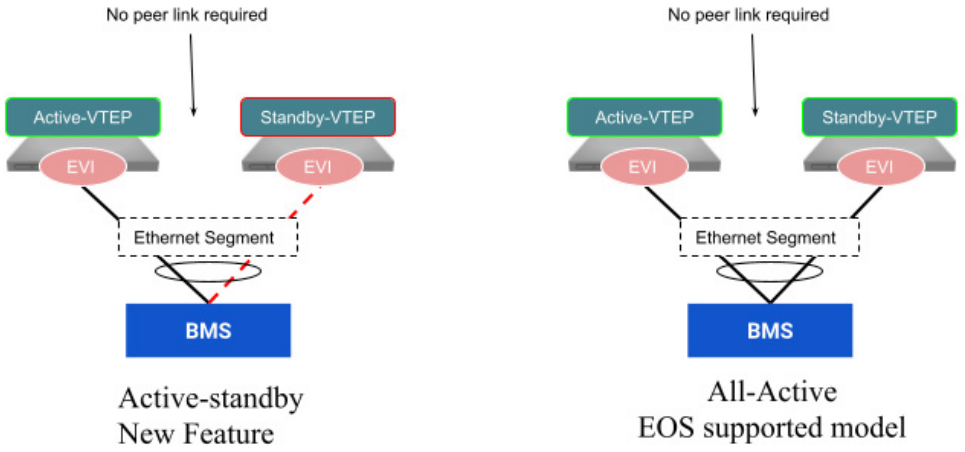

Figure 1: EVPN MH Modes

In EVPN, the Ethernet segment is operating in active-standby mode when only a single VTEP, among a group of VTEPs attached to an Ethernet segment, is allowed to forward traffic onto the Ethernet segment. On the other hand, the Ethernet segment is operating in all-active when all the multihoming PE routers attached to the Ethernet segment are allowed to forward traffic onto the Ethernet segment.

EOS currently supports single homing and all-active EVPN multihoming for connecting CE devices to PE. For all-active multihoming with EVPN VXLAN, EOS currently supports connecting up to 16 VTEPs to a CE device on a given Ethernet segment.

EVPN A-A Multihoming LAG

In all-active mode, the Ethernet segment on all the PE routers and CE devices should be configured as LAG either static or dynamic . With static LAG, there is no LACP negotiation between switches. Hence, static LAG is prone to mis-configuration.

In dynamic mode, peer switches build link aggregation groups by negotiating them sending LACP PDUs. The system identifier carried into the LACP header is used by peer devices when forming an aggregation to verify that all links are from the same switch. Each switch is assigned a globally unique LACP system ID by concatenating the LACP system priority to the system MAC address of one of its physical ports.

With Arista's standard MLAG, the MLAG uses the "MLAG domain system ID" instead. Because the MLAG domain system ID always has the same value on both MLAG switches, the connected hosts see them as a single device.

When using dynamic LAGs with EVPN Multihoming, a common "LACP system-id" value must be configured on the LAG interfaces to facilitate LACP communication with the downstream CE devices. For a given ES , the LACP system ID must be identical on all the LAG interfaces (on all multi-homing PE VTEPs) so that the downstream CE device treats the different PEs connecting with the same Ethernet Segment as a single peer device. In particular, one can configure and use a consistent LACP System ID across multiple ES LAGs which belong to the same set of MH PEs.

EVPN Multihoming Route Types

For single-homing mode, VTEPs advertise **Type 3** "Inclusive Multicast Ethernet Tag"(IMET) routes to build floodset and **Type 2** MAC-IP routes to exchange MAC routes, IP2MAC and IPv6MAC bindings using BGP control planes. In addition to these route types, EVPN Multihoming VTEPs also exchange **Type 1** "Auto-Discovery" (AD) routes and **Type 4** "Ethernet Segment" (ES) routes for each Ethernet segment. Each of them are labelled using an "Ethernet Segment Identifier" (ESI) 10 bytes value. More details will follow.

- Multihoming VTEPs always exchange type 4 ES Route for Designated Forwarder (DF) election. For a given local ES, the DF is selected on a "Per-Ethernet-VPN Instance" (per-EVI) basis. Hence, the DF for a certain "Vxlan Network Identifier" (VNI) on a given ESI may be different from the DF for another VNI on the same ESI. Only one MH VTEP can be elected as DF and only the DF is allowed to decapsulate and forward the BUM traffic destined to the connecting multihomed device. This is to avoid BUM traffic duplication.

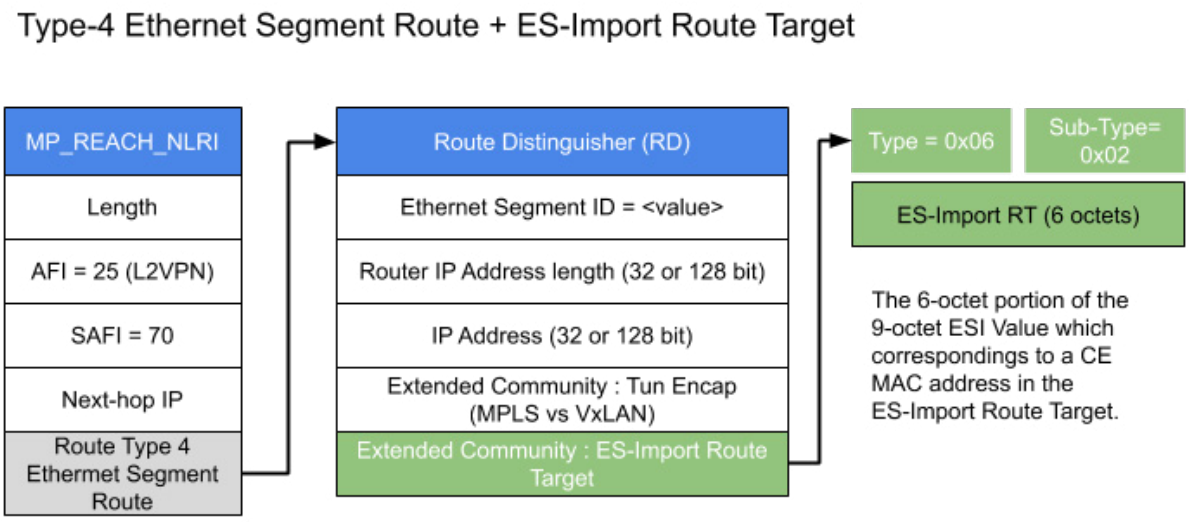

Figure 2: Type-4 Ethernet Segment Route + ES-Import Route Target

- Each multihoming PE advertises a "Ethernet Auto Discovery" Route Message per Ethernet Segment. When an ethernet segment fails (all ports in the ES go down), the multihoming PE withdraws the Auto Discovery per ES route. This withdrawal triggers all the remote PEs to reprogram the MAC address adjacencies associated with the failed ES. This mechanism provides fast convergence and it is called "**Mass Withdrawal**".

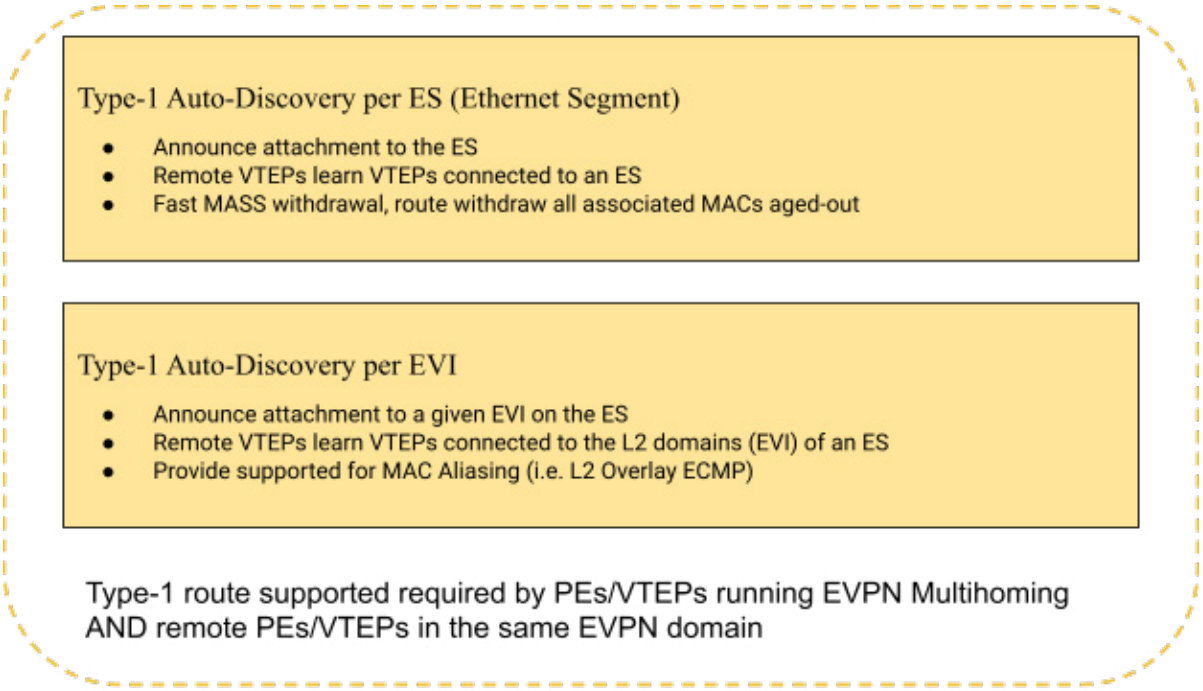

Figure 3: Type-1 Auto-Discovery Routes

- In figure 7, VTEP-1 and VTEP-2 advertise **Type 1** AD routes per ES for each local ethernet segment. The VTEPs also advertises AD routes for each EVI configured on the local ethernet segment. These routes are also called "Type 1 AutoDiscovery routes per EVI".
- The ETID field in the AD route per ES is set to MAX_ETID which has the value of 0xFFFFFFFFF.
- The ETID field in the AD route per EVI always has the value of '0' if the EVI is VLAN-based. In the case of a VLAN-aware bundle EVI, the ETID field is set to the value of the VNI.

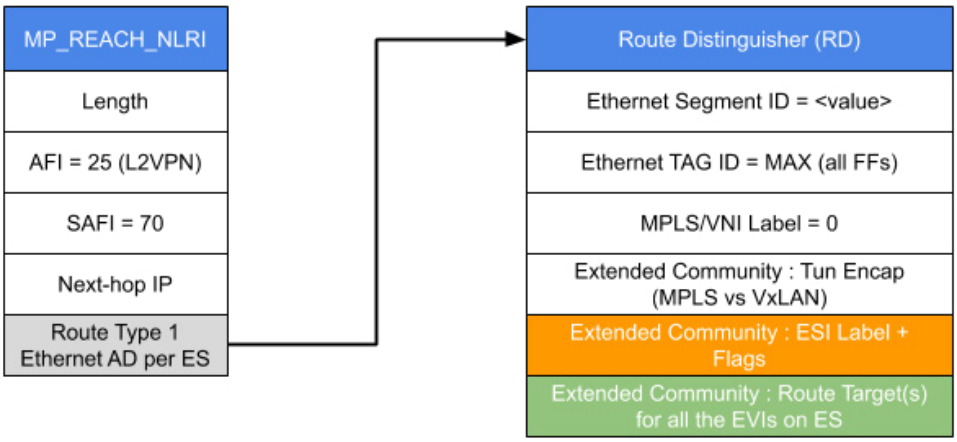

Figure 4: Type-1 Auto-Discovery Route per ES Route

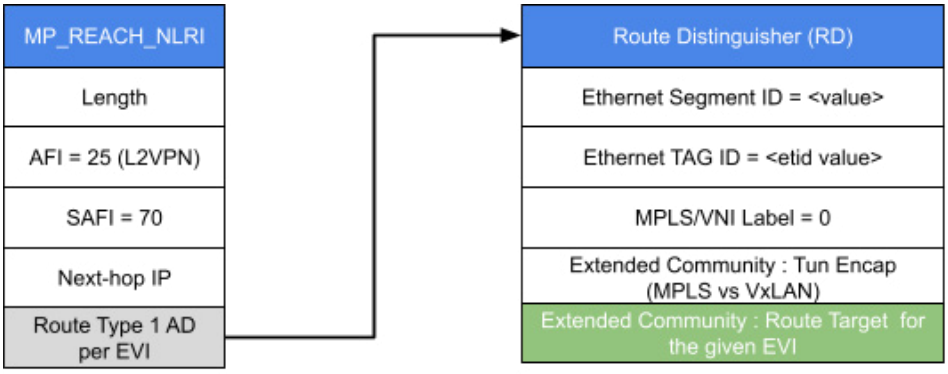

Figure 5: Type-1 Auto-Discovery Route per EVI Route

The Type 1 AD route carries the following BGP Path attributes:

- ESI, Route Distinguisher (RD) and Route Target (RT)
  - › For each EVI configured on the local ES, the sender PE generates an AD route and sets the RD, ESI and RT based on local MAC-VRF configuration. This route is known as "Auto Discovery per EVI".
  - › ESI is a 10 byte network unique number for identifying the ES in the EVPN route exchange.

ESI Type (1 octet)	ESI Value (9-octets)

0	Manual Only: 9-octets ESI value are arbitrary value
1	Manual or Auto: CE LACP system MAC (6) + port Key (2) + 00
2	Manual or Auto: Root-Bridge MAC (6) + Priority Key (2) + 00
3	Manual or Auto: System MAC (6) + Local Discriminator (3)
4	Manual or Auto: Router-ID (4) + Local Discriminator (4) + 00
5	Manual or Auto: AS number (4) + Local Discriminator (4) + 00

  - › The AD route is imported by all the multihomed peers and remote PE routers that are part of the same EVI.
- ESI label extended community
  - › This is an EVPN extended community which has a one-octet flag where the lower-order bit is called "Single-Active" bit. This bit indicates if the sending PE associated with the EtS is operating in all-active or a single-active mode. In particular, the Single-Active bit value 0 means All-Active and 1 means Single-Active.
  - › This community also carries the ESI label used for split horizon in case of EVPN MPLS multihoming. For EVPN VxLAN multihoming, the ESI label has the value zero.

Designed Forwarder (DF)

EOS supports the DF Election procedure described in RFC 7432. Currently, EOS multihoming VTEPs always elect one of the multihoming VTEPs per EVI. This scheme allows efficient local-balancing of multi destination traffic destined for a given ES. The procedure is as follows:

1. For each local ES, the multihoming VTEP always advertises a route type 4 ES route with the associated ES-import extended community attribute.
2. After advertising the ES route, the multihoming VTEP starts a DF election timer and waits for other multihoming peers connecting to the same ES to advertise their ES routes. The default DF election timer is 3 seconds and the timer is configurable using command '***designated-forwarder election hold-time***'.
3. When the DF hold-timer expires, each multihomed VTEP builds a VTEP ordered list containing VTEP addresses connecting to the same ES.
4. Every VTEP is given an index (starting from 0) indicating its position in the ordered list. The index is used to determine which VTEP is the DF for a given EVI on the ES.
   a. For vlan-based service, the index = VNI MOD by number of MH VTEPs.
   b. For vlan-aware-bundle service, the index = (lowest VNI in the bundle) MOD by number of MH VTEPs.
5. In the figure showing vlan-based EVI below, VTEP1 has the index value of 0 while VTEP2 has the value of 1. For VNI 10010 and 10012, VTEP1 is the DF while VTEP2 is the DF for VNI 10011 and 10013.

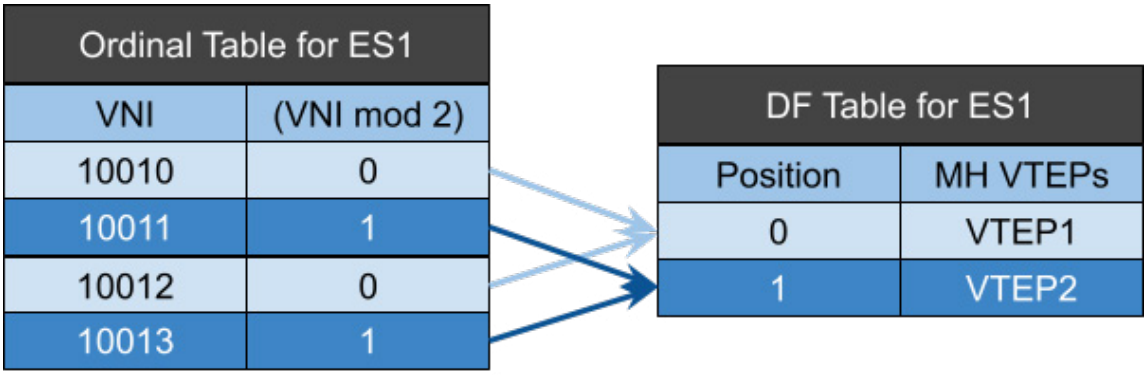

Figure 6: DF Election for Ethernet segment ES1

6. For the given ES, the elected DF MH VTEP forwards BUM traffic destined for the multihomed CE device. To avoid duplication on the ES, the non-DF VTEP blocks the BUM traffic destined for the multihomed CE device instead.
7. Support Service Carving since Designated Forwarder (DF) is elected on a per EVI basis for each ES. In particular:
   a. Allowing ECMP of the DF forwarding across all the MH A-A VTEPs connected to the ES.
   b. Only the DF router is responsible for forwarding the BUM traffic onto the ES [see later section on Unicast and BUM Traffic].

MAC Aliasing

- Because of the LAG hashing mechanism, multihoming PEs may learn only a subset of the downstream directly connected hosts on a given ES. This affects load balancing of incoming traffic from remote PEs and gives poor link utilization towards the all-active multihomed devices.
- EOS solves this problem by supporting **MAC aliasing**. For type 2 MAC-IP routes advertised by single-homed PE, the reachability of a given host is always associated with the nexthop specified in the MAC-IP route.

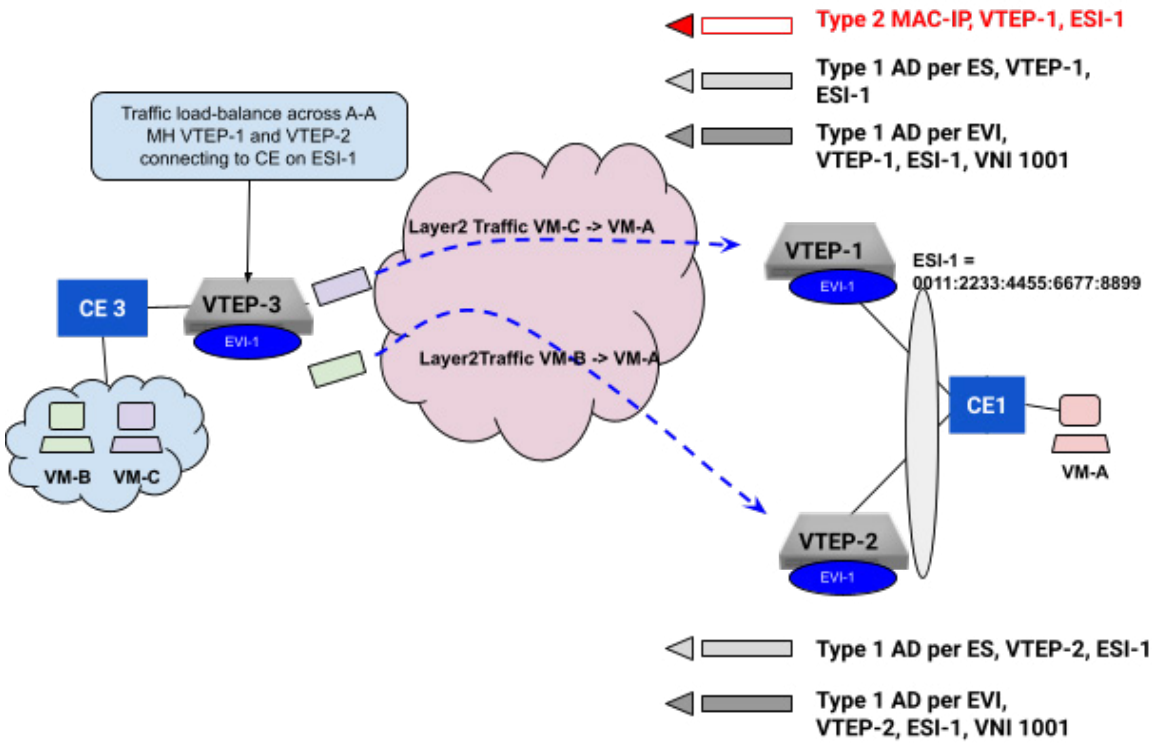

Figure 7: MAC Aliasing using Type 1 AD per AVI and Type 2 MAC-IP Route

- With MAC aliasing, the reachability of the host is determined by the set of multihomed VTEPs associated with the non-zero ESI specified in the MAC-IP route. Using the AD route per EVI, the remote PE is able to determine the set of multihomed VTEPs associated with the given ES.
- In the previous figure, VM-A behind CE-1 only sends traffic over one link of the LAG member to VTEP-1. Remote VTEP-3 learns the MAC address from the Type-2 MAC-IP route message advertised by VTEP-1.
- VTEP-2 never learns the MAC address for VM-A. Hence, VTEP-2 does not advertise a Type-2 MAC-IP route message for it's mac address.
- Because the remote PE router VTEP-3 also received the Type-1 AD per EVI route messages from both VTEP-1 and VTEP-2, it knows that VM-A must be reachable through both VTEP-1 and VTEP-2. Hence, VTEP-3 uses this info to program the MAC forwarding table to ECMP the Layer 2 traffic for VM-A traffic among VTEP-1 and VTEP-2. This mechanism is called "**MAC Aliasing**".

Mass Withdrawal

When a link inside aLAG between a multihomed CE device and a multihomed VTEP fails, the VTEP withdraws the AD route per ES. Hence, the AD route is also known as the **Mass Withdrawn** route. The other VTEPs handle the mass withdrawal in the following:

- Remote VTEPs
  - All remote VTEPs remove the nexthop from the set of multihomed VTEPs to reach the remote ESI.
  - They reprogram all the MAC routes behind the multihomed CE device with the new nexthop (new set of multihomed VTEPs).
  - In figure below, VTEP-1 withdraws the auto discovery AD route for the ES-1. VTEP-3 removes VTEP-1 from the nexthop for ESI-1. As a result, the nexthop for ESI-1 contains VTEP-2 only after the failure.

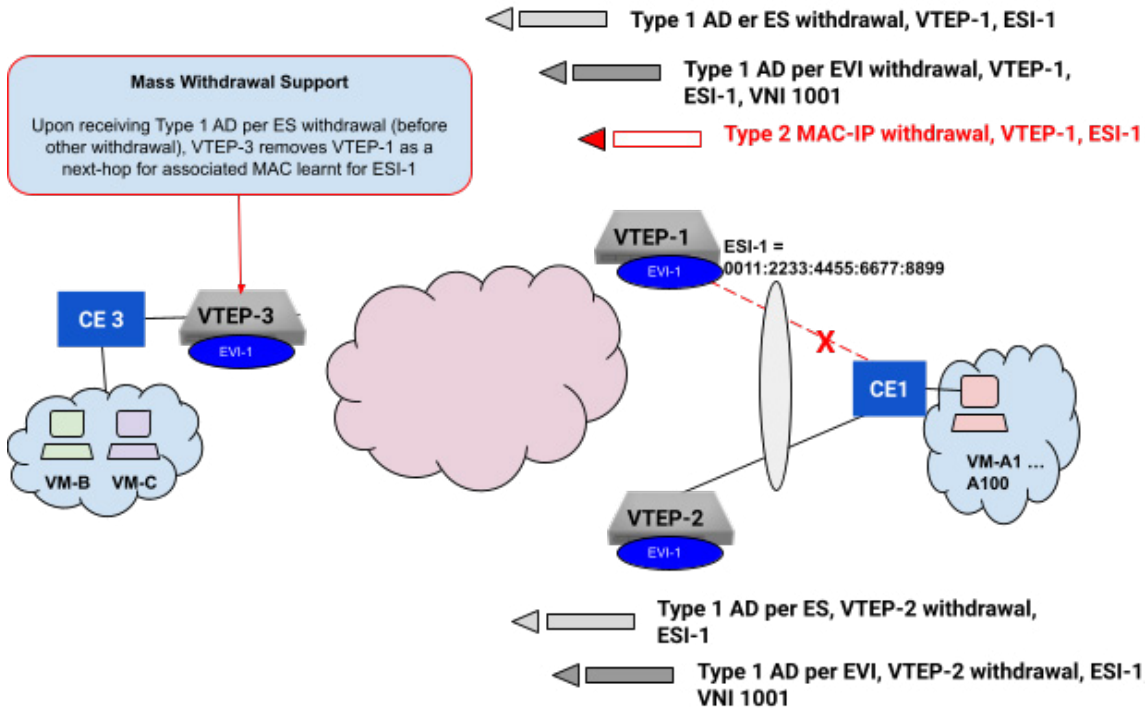

Figure 8: Mass Withdrawal using Type 1 AD per ES Route Message

- Multihomed VTEP Peer
  - When the other multihomed VTEP Peer had received and learned the same set of MAC addresses on the ES, the mass withdrawal does not affect routes routing to the local ES.
  - When the other multihomed VTEP Peer had not received and learned the same set of MAC addresses on the ES, the other multihomed VTEP peer removed the MACs from its MAC forwarding table. Traffic towards flushed MAC addresses are flooded on the local ESt until the MAC addresses are learned again.
  - In figure above after the ES link failed between VTEP-1 and CE-1, VTEP-2 will flush the MAC address for multihomed CE-1 if VTEP-2 has not learned its MAC address on its local ES. The downstream traffic toward CE-1 would be flooded until VTEP2 received and learned the MAC address for CE-1.

Unicast and BUM Traffic

With EVPN all-active multihoming, packet forwarding can be handled differently based on the type of traffic: known unicast or BUM.

- Unicast Traffic
  - For traffic sourced from downstream CE, the traffic is load-balanced to all the connected multihomed PE routes on the Ethernet segment.
  - For traffic from remote PE toward downstream CE, the remote PE uses "mac aliasing" to determine the set of multihomed PEs for the ES where the target CE resides. Traffic from the remote PE devices is load-balanced to all the multihomed PE devices connected on the same ES.

- BUM Traffic
  - For BUM traffic sourced from remote VTEP, the remote PE HERs (Head End Replicates) the traffic to all the other VTEPs. The traffic is filtered based on the DF election. When the VXLAN encapsulated BUM traffic reaches the VTEPs, only the DF accepts the packets and forward them to the downstream CE while the non-DF VTEP always discards it to avoid traffic duplication on the ES..
  - Consider BUM traffic originating from VM-C in Figure 9. VTEP-3 HERs the BUM traffic and sends the packets to the Spine. DF is elected on a per (ES, VNI) basis. For example, VTEP-1 is the elected DF for (ESI-1, VNI-A) and VTEP-2 is the DF for (ESI-2, VNI-A). VTEP-2 must decap and bridge the BUM traffic to orphan host VM-E but blocked the traffic from going onto the ESI-1 LAG to Server-1 since VTEP-2 is the non-DF for ESI-1.
  - With EVPN VXLAN, the split-horizon filtering rule for all-active ES is based on the source IP address of the VTEP since VXLAN encapsulation does not include the ESI Label.
    - Each VTEP tracks the IP address(es) of other multihomed EVPN VTEPs sharing the same ES.
    - When a VTEP receives a BUM packet from the overlay network, it examines the source IP address in the VXLAN tunnel header (source VTEP) and filters out the packet on all local interfaces or ethernet links connected to the ESs eventually shared with the source VTEP.
  - With this split-horizon filtering mechanism, the ingress VTEP should always perform replication to all directly attached ESs (regardless of DF status) for all BUM traffic arriving from downstream access ports. This is referred to as "**Local Bias**" (RFC 8365).

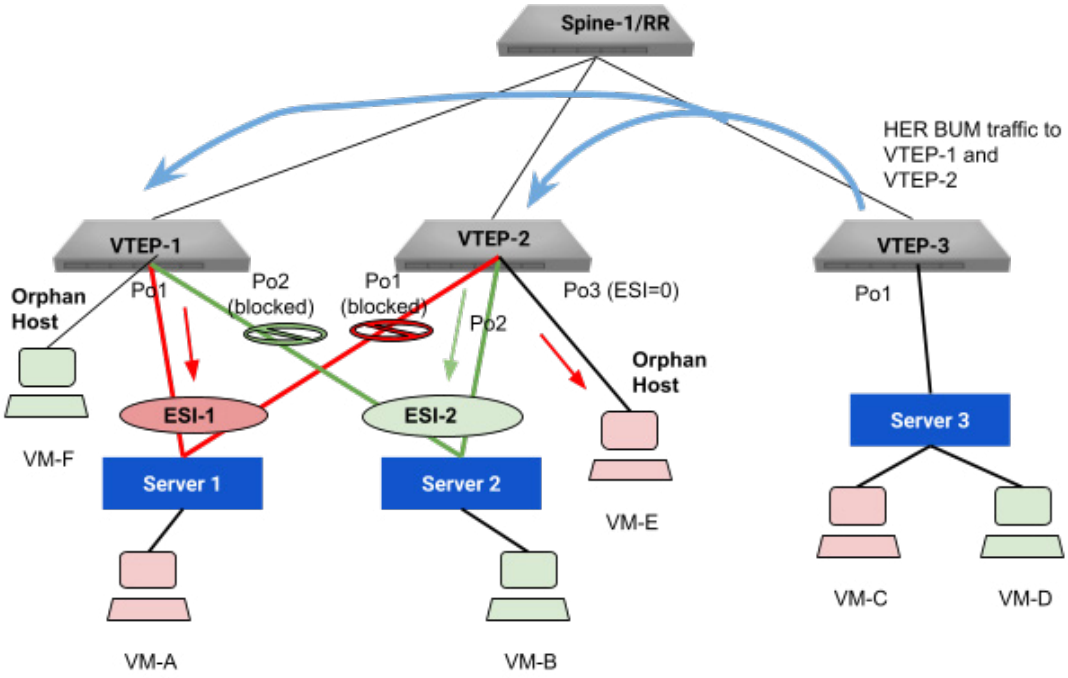

Figure 9: Split Horizon with Ethernet Segment

- For BUM traffic originated from orphan host VM-F in figure 9, VTEP-1 replicates the traffic onto directly attached ES Po2 to support Local Bias. When the multihomed VTEP-2 (peer) receives the BUM packet:
  - VTEP-2 performs the split-horizon filter check and blocks the BUM traffic from bridging onto its local ESI-2 because VTEP-1 shares the same ESI-2.
  - VTEP-2 will still decap the received VXLAN BUM frames and flood it out other local interfaces that are members of the bridge domain represented by the EVI, but not part of the ESI. In figure 6, VTEP-2 floods the BUM frames to single homed hosts VM-E connecting to zero-ES LAG Po3.

EVPN A-A Multihoming and IRB

In figure below, CE1 is attached to the Ethernet segment ES-1 which connects with a pair of multihoming TORs VTEP-1 and VTEP-2. EOS VTEPs support IRB (Integrated Routing and Bridging) described in the inter-subnet draft (draft-sajassi-l2vpn-evpn-inter-subnet-forwarding). Refer to EOS User Manual for more detailed explanation on IRB mechanism and configuration.

Consider a remote TOR VTEP-3 and a CE-3 attached to it. In this example, VM-A, VM-B and VM-C belong to 3 different subnets. When VM-B wants to communicate with VM-A, it sends traffic to its L3 default gateway VTEP-3. If VTEP-3 has the same IRB as VTEP-1 and VTEP-2, where VM-A is dual homed, VTEP-3 does the routing lookup and bridges the packet on the same bridging domain as VM-A. The VNI in the VXLAN encapsulation is the L2 domain of VM-A. This mechanism is called **Asymmetric IRB**.

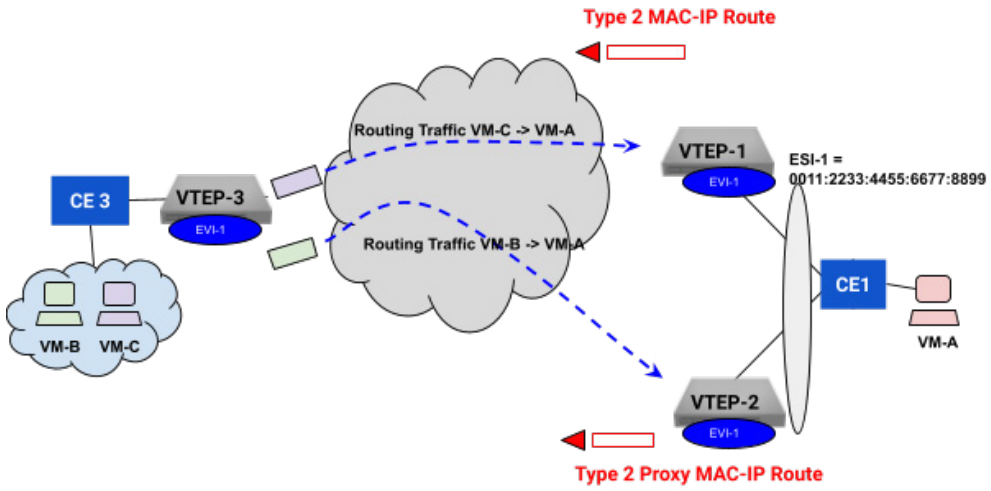

Figure 10: L3 ECMP Traffic Flow with Symmetric IRB

What if VTEP-3 does not have the IRB as VM-A? With Symmetric IRB, VTEP-3 performs routing lookup in the L3VNI tenant VRF where both VM-B and VM-A belongs to. In this example, the routing lookup should result in an L3 ECMP list (VTEP-1/VTEP-2) which is built using the following mechanism:

- If VM-A is locally learnt only by one of the multihoming VTEPsdue to port-channel hashing, only one of them advertises the MAC-IP route for VM-A.
- After VTEP-2 processes the VM-A's MAC-IP route advertisement from VTEP-1, it realizes that the target host VM-A in the MAC-IP route is part of its local ES ESI-1.

- VTEP-2 generates a proxy MAC-IP route advertisements with VTEP-2 as the next-hop. This proxy mechanism is described in [draft-rbickhart-evpn-ip-mac-proxy-adv-01]. The proxy MAC-IP route carries the ND extended community which has the "**Proxy Flag**" set.
- With the **Proxy MAC-IP mechanism**, VTEP-3 receives two MAC-IP routes for VM-A. This allows VTEP-3 to construct an L3 ECMP list (VTEP-1/VTEP-2) for host VM-A connecting to multihoming CE1.
- For the same bridge domain, the corresponding EVI configuration should always use a unique RD on each VTEP. The best practice is to use Type 1 RD where the source VTEP IP address is the first field of the RD. With a unique RD for the same MAC-VRF on different PE, the BGP Route Reflectors (RRs) should advertise two MAC-IP routes, one from each MH VTEPs. If different MH VTEPs configure the same RD for the same MAC-VRF, the RRs don't advertise two MAC-IP routes. Instead, the RRs always advertise only the bestpath. Hence VTEP-3 cannot perform L3 overlay ECMP with the same RD.

Multihoming Configuration

Connecting a L2 switch to Multihoming PE

Starting from 4.25.1F EOS supports a new feature "Spanning tree network super root. This feature enables the switch to act as STP network super root. When this feature is enabled, it sends STP BPDUs with bridge MAC address and source MAC address as 0000.0000.0001 and priority as 0. This enables topologies like EVPN All-Active multihoming to use STP and detect Layer-2 loops. The user can choose multiple switches to act as STP network super root.

This feature supports EVPN All-Active multihoming configurations. To enable this feature, configure "spanning-tree root support".

Pre 4.25.1F EVPN All-Active Multihoming mechanism does not support Spanning Tree Protocol (STP) on the ES LAG connecting with downstream CE devices. In case a L2 switch is used as CE, it is possible to have L2 loops if it is also connected to other downstream L2 switches. Hence, the network operators should make sure that all the downstream L2 connections should be "loop free". There are two ways to disable STP:

- Configure "***spanning-tree BPDU filter enable***" on the port-channels used for Ethernet segment.
- As a second option - if the VLANs configured on the ES LAG are used exclusively, configure "***no spanning-tree vlan <vlan id>***" to disable STP for those VLANs.

MLAG supports STP for Layer-2 loop detection. In fact, most customers enable STP in their MLAG(s) to ensure no downstream Layer-2 loops due to mis-cabling or mis-configuration.

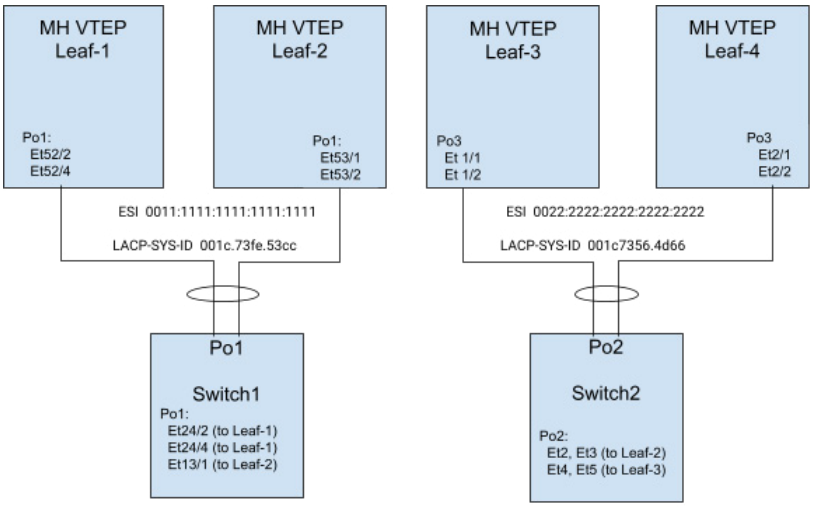

Figure 11: Multihoming LAG Interface Configuration

Ethernet Segment ID

- ESI is a 10 bytes network unique number for identifying the ES in the EVPN route exchange.
- Currently, EOS supports Type 0 ESI only which takes 9 octets on user input. The ES LAG connecting to the same CE device should always have the same ES ID. One scheme is to assign the first 6 octets of the ESI using the system MAC address of the connected CE device and set the rest of 3 octets to zero.
- The PE that is connected to a multihomed CE device exchanges Type 4 ES route messages which consists of an ESI and ES-import extended community to learn other PE routers connected to the same ES.

ES-Import Route Target

- PE routers construct ES-import based route-target filter to allow Type-4 ES routes with matching ES-import route-target extended community to be imported.
- All the LAG interfaces (on different PE routers) for a given ES should have identical ES-import route-target value.
- ES-Import route-target takes 6 octets on user input. One scheme is to assign it using the system MAC address of the CE device connecting to the multi-homing PEs.

ES LAG	CE System MAC	ES-Import Route Target	Ethernet Segment ID
Leaf-1 PO1	1111.1111.1111	11:11:11:11:11:11	0011:1111:1111:1100:0000
Leaf-2 PO2	1111.1111.1111	11:11:11:11:11:11	0011:1111:1111:1100:0000
Leaf-3 PO3	2222.2222.2222	22:22:22:22:22:22	0022:2222:2222:2200:0000
Leaf-4 PO4	2222.2222.2222	22:22:22:22:22:22	0022:2222:2222:2200:0000

ES LAG Interface

- Configure all the port-channel interfaces using identical interface number for the same ES.

ESI	Leaf-1	Leaf-2
0011:1111:1111:1100:0000	Port-channel-1	Port-channel-1
0022:2222:2222:2200:0000	Port-channel-2	Port-channel-2

LACP System-ID for ES LAG

- Because Po1 on Leaf-1 and Po2 Leaf-2 are part of the same ES, they are both configured with the same "LACP System-id". Likewise the LACP system-id on the ES LAG interfaces of Leaf-3 and Leaf-4 should be the same.
- One way to select the LACP system-id is to choose the lower System MAC addresses between the two peers. Hence, the LACP system-id based on this scheme:

ES LAG	Ethernet Segment	LACP System-ID	System MAC
Leaf-1 PO1	0011:1111:1111:1100:0000	001c.73fe.53cc	001c.76aa.3daa
Leaf-2 PO2	0011:1111:1111:1100:0000	001c.73fe.53cc	001c.73fe.53cc
Leaf-3 PO3	0022:2222:2222:2200:0000	001c.7356.4d66	001c.7356.4d66
Leaf-4 PO4	0022:2222:2222:2200:0000	001c.7356.4d66	001c.81ee.7d1a

LACP Port Range

- A global configuration command must be used to configure LACP port-id range for all PEs connected to multihomed CE(s). When the port-id range command is configured, it limits the range of allocated port-ids to the one specified.
- If the CE is receiving the packet from two different ports in the same LAG (one each to a MH PE) with the same port-id, the CE device interprets (wrongly) that the port has moved from one LAG to a different LAG. Hence, configure the range for the port-id on the PE router to ensure there are no overlapping ranges for the PEs in the same ES.
- One scheme to assign the port range is simply based on the total number of physical ports on a PE router. For instance, one can set the range to be 100 as long as the total number of ports on a PE router is less than or equal to 100 ports.

Leaf-1 (MH VTEP)	Leaf-2 (MH VTEP)
(config)#lacp port-id range 1 100	(config)#lacp port-id range 101 200

Ethernet Segment Configuration

Leaf-1 (MH VTEP-1)1)	Leaf-2 (MH VTEP-2)
! ! Disable SPT for those VLANs used for Ethernet Segment ! spanning-tree mode rapid-pvst no spanning-tree vlan-id 2-4094 ! interface Port-Channel1 switchport trunk allowed vlan 100-102 switchport mode trunk switchport ! evpn ethernet-segment identifier 0011:1111:1111:1100:0000 route-target import 11:11:11:11:11:11 ! lacp system-id 001c.73fe.53cc !	! ! Disable SPT for those VLANs used for Ethernet Segment ! spanning-tree mode rapid-pvst no spanning-tree vlan-id 2-4094 ! interface Port-Channel1 switchport trunk allowed vlan 100-102 switchport mode trunk switchport ! evpn ethernet-segment identifier 0011:1111:1111:1100:0000 route-target import 11:11:11:11:11:11 ! lacp system-id 001c.73fe.53cc !

Link Tracking Group (on boot)

To force the ES LAG interface to remain in the errdisabled state after switch reload, configure a "Link Tracking Group" feature which specifies (1) the upstream core interfaces and down stream ES Lag interfaces (2) how long the LAG remains error disabled after core interfaces come up. Depending on the scale, it's good to stagger the recovery-delay for the ES LAG interfaces. This mechanism is similar to the "MLAG reload delay" used in Arista standard MLAG switches. It allows time to program the learned remote BGP EVPN routes before attracting traffic.

The default "links minimum" value for a Link State Group is 1. Use CLI "**show link tracking group**" to check the Link State Group status. Note there is the other failure recovery behavior by default: when all the upstream links are down, the downstream ES links will be shut down automatically to prevent blackholing traffic from ES links.

Leaf-1 (MH VTEP-1)	Leaf-2 (MH VTEP-2)

link tracking group ES1
  recovery delay 500
!
link tracking group ES2
  recovery delay 520
!
interface Port-Channel1
  description EVPN-MH-Link1
  link tracking group ES1 downstream
interface Port-Channel2
  description EVPN-MH-Link2
  link tracking group ES2 downstream
 	
! All SPINE/Core facing links are tracked as upstream
interface Ethernet49/1-54/1
  link tracking group ES1 upstream
  link tracking group ES2 upstream
	link tracking group ES1
	  recovery delay 500
	!
	link tracking group ES2
	  recovery delay 520
	!
	interface Port-Channel101
	  description EVPN-MH-Link1
	  link tracking group ES1 downstream
	interface Port-Channel102
	  description EVPN-MH-Link2
	  link tracking group ES2 downstream
	
	! All SPINE/Core facing links are tracked as upstream
	interface Ethernet49/1-54/1
	  link tracking group ES1 upstream
	  link tracking group ES2 upstream

EVPN IRB

For a detailed configuration guide on EVPN IRB, refer to following document:

https://eos.arista.com/multi-tenant-evpn-vxlan-irb-configuration-verification-guide-ebgp-overlay-underlay/. Here are some best practises for configuring IRB with A-A MH PE routers:

- Always configure a unique RD on EVI on different PE routers sharing the same ES.
- Each PE router should have a unique VTEP-IP. For eBGP peering, each A-A MH PE router should have its own ASN as well. If the A-A MH Peers sharing the same ES are having the same ASN in the overlay (i.e. iBGP overlay), the EVPN peering should configure "allowas-in" so that local hosts can sync the MAC address between MH peers.
- On the remote PE, configure "maximum-paths <# of overlay paths>" under all the IP VRFs to enable L3 ECMP.

Verification

LACP Status

Multihomed Switch

```
#show lacp internal
LACP System-identifier: 8000,00-1c-73-38-22-d3
State: A = Active, P = Passive; S=ShortTimeout, L=LongTimeout;
       G = Aggregable, I = Individual; s+=InSync, s-=OutOfSync;
       C = Collecting, X = state machine expired,
       D = Distributing, d = default neighbor state
                          |  Partner                                                Actor
   Port           Status   |  Sys-id                          Port#       State          OperKey
PortPriority
------------ --------------|------------------------------ ----------- ------------- ----------
--- ------------
Port Channel Port-Channel1:
   Et13/1         Bundled  |  8000,00-01-02-03-04-05            53       ALGs+CD         0x0001
32768
   Et24/2         Bundled  |  8000,00-01-02-03-04-05            98       ALGs+CD         0x0001
32768
   Et24/4         Bundled  |  8000,00-01-02-03-04-05           100       ALGs+CD         0x0001
32768
```

EVPN Peering

First, verify that the EVPN sessions are Established on the Spines and Leafs:

Spine1

```
Spine1.00:45:12#sh bgp evpn sum
BGP summary information for VRF default
Router identifier 8.8.8.8, local AS number 65000
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  1.1.1.1          4  65001          28397     28431    0    0 02:20:06 Estab   131    131
  2.2.2.2          4  65002          30189     15295    0    0 10:08:43 Estab   68     68
  7.7.7.7          4  65007           2123     37679    0    0 10:11:42 Estab   22     22
```

Spine2

```
Spine2....22:28:33#sh bgp evpn sum
BGP summary information for VRF default
Router identifier 9.9.9.9, local AS number 65000
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  1.1.1.1          4  65001           1409      1341    0    0 18:08:37 Estab   20     20
  2.2.2.2          4  65002           1397      1364    0    0 18:10:43 Estab   22     22
  7.7.7.7          4  65004           1379      1399    0    0 18:08:06 Estab   51     51
```

A-A MH Leaf-1

```
Leaf-1.01:51:00#sh bgp evpn sum
BGP summary information for VRF default
Router identifier 1.1.1.1, local AS number 65001
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  8.8.8.8          4  65000          24005     25993    0    0 02:10:44 Estab   181    181
  9.9.9.9          4  65000          23971     27624    0    0 02:10:44 Estab   181    181
```

A-A MH Leaf-2

```
Leaf-2.01:51:00#sh bgp evpn sum
BGP summary information for VRF default
Router identifier 2.2.2.2, local AS number 65002
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  8.8.8.8          4  65000          14945     30058    0    0 09:59:41 Estab   244    244
  9.9.9.9          4  65000          14821     32857    0    0 09:57:26 Estab   244    244
```

Remote Single Homed VTEP (Leaf-7)

```
Leaf-7.11:39:56#sh bgp evpn sum
BGP summary information for VRF default
Router identifier 7.7.7.7, local AS number 65001
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  8.8.8.8          4  65000           1900      1845    0    0    1d00h Estab   75     75
  9.9.9.9          4  65000           1899      1857    0    0    1d00h Estab   75     75
```

EVPN IMET Routes Exchange

For L2 ECMP, the multi-homed VTEPs need to advertise Type-1 AD per ES and Type-1 AD per EVI routes. These Type-1 routes combined with the Type-2 MAC-IP routes form the L2 ECMP aka Aliasing route.

Before verifying L2 ECMP, verify that the Type-3 IMET routes are received.

For instance, Leaf-7 builds floodset to include Leaf-1 (1.1.1.1) and Leaf-2 (2.2.2.2) using EVPN Type 3 IMET routes with originator-IP 1.1.1.1 and 2.2.2.2:

Remote Single Homed VTEP (Leaf 7)

```
Leaf-7.11:45:24#sh bgp evpn route-type imet 1.1.1.1 vni 100
BGP routing table information for VRF default
Router identifier 7.7.7.7, local AS number 65001
Route status codes: s - suppressed, * - valid, > - active, # - not installed, E - ECMP head, e
- ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
 * >Ec   RD: 1.1.1.1:100 imet 1.1.1.1
                                 1.1.1.1               -       100     0       65000 65002 i
 *  ec   RD: 1.1.1.1:100 imet 1.1.1.1
                                 1.1.1.1               -       100     0       65000 65002 i
Leaf-7.11:45:31#sh bgp evpn route-type imet 2.2.2.2 vni 200
BGP routing table information for VRF default
Router identifier 7.7.7.7, local AS number 65001
Route status codes: s - suppressed, * - valid, > - active, # - not installed, E - ECMP head, e
- ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
 * >Ec   RD: 2.2.2.2:200 imet 200 2.2.2.2
                                 2.2.2.2               -       100     0       65000 65003 i
 *  ec   RD: 2.2.2.2:200 imet 200 2.2.2.2
                                 2.2.2.2               -       100     0       65000 65003 i
```

VXLAN Floodset

Once verified IMET routes are received, check VXLAN floodset is populated as a result of these Type-3 IMET routes.

Remote Single Homed VTEP (Leaf 7)

```
Leaf-7.11:45:45#sh int vxlan1
Vxlan1 is up, line protocol is up (connected)
  Hardware is Vxlan
  Source interface is Loopback1 and is active with 7.7.7.7
  Replication/Flood Mode is headend with Flood List Source: EVPN
  Remote MAC learning via EVPN
  VNI mapping to VLANs
  Static VLAN to VNI mapping is
    [100, 100]        [101, 101]        [102, 102]        [200, 200]
    [201, 201]        [202, 202]
  Dynamic VLAN to VNI mapping for 'evpn' is
    [3961, 5001]      [3962, 5000]
  Note: All Dynamic VLANs used by VCS are internal VLANs.
        Use 'show vxlan vni' for details.
  Static VRF to VNI mapping is
   [vrf_100-102, 5000]
   [vrf_200-202, 5001]
  Headend replication flood vtep list is:
   100 1.1.1.1         2.2.2.2
   101 1.1.1.1         2.2.2.2
   102 1.1.1.1         2.2.2.2
  MLAG Shared Router MAC is 0000.0000.0000
```

Type 1 AutoDiscovery Route

Now let's look at the Type-1 AD per ES and Type-1 AD per EVI route that will help us form an L2 ECMP.

Remote VTEP Leaf-7

```
Leaf-7.06:25:28#sh bgp evpn route-type auto-discovery esi 0011:1111:1111:1100:0000
BGP routing table information for VRF default
Router identifier 7.7.7.7, local AS number 65001
Route status codes: s - suppressed, * - valid, > - active, # - not installed, E - ECMP head, e
- ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
 * >Ec    RD: 1.1.1.1:100 auto-discovery 0 0011:1111:1111:1100:0000
                                 1.1.1.1               -       100     0       65000 65002 i
 * ec    RD: 1.1.1.1:100 auto-discovery 0 0011:1111:1111:1100:0000
                                 1.1.1.1               -       100     0       65000 65002 i
 * >Ec    RD: 2.2.2.2:100 auto-discovery 0 0011:1111:1111:1100:0000
                                 2.2.2.2               -       100     0       65000 65003 i
 *  ec    RD: 2.2.2.2:100 auto-discovery 0 0011:1111:1111:1100:0000
                                 2.2.2.2               -       100     0       65000 65003 i
 * >Ec    RD: 2.2.2.2:200 auto-discovery 200 0011:1111:1111:1100:0000
                                 2.2.2.2               -       100     0       65000 65003 i
 *  ec    RD: 2.2.2.2:200 auto-discovery 200 0011:1111:1111:1100:0000
                                 2.2.2.2               -       100     0       65000 65003 i
```

```
 * >Ec    RD: 2.2.2.2:200 auto-discovery 201 0011:1111:1111:1100:0000
                                2.2.2.2                -       100     0       65000 65003 i
 *  ec    RD: 2.2.2.2:200 auto-discovery 201 0011:1111:1111:1100:0000
                                2.2.2.2                -       100     0       65000 65003 i
 * >Ec    RD: 2.2.2.2:200 auto-discovery 202 0011:1111:1111:1100:0000
                                2.2.2.2                -       100     0       65000 65003 i
 *  ec    RD: 2.2.2.2:200 auto-discovery 202 0011:1111:1111:1100:0000
                                2.2.2.2                -       100     0       65000 65003 i
 * >Ec    RD: 1.1.1.1:1 auto-discovery 0011:1111:1111:1100:0000
                                1.1.1.1                -       100     0       65000 65002 i
 *  ec    RD: 1.1.1.1:1 auto-discovery 0011:1111:1111:1100:0000
                                1.1.1.1                -       100     0       65000 65002 i
 * >Ec    RD: 2.2.2.2:1 auto-discovery 0011:1111:1111:1100:0000
                                2.2.2.2                -       100     0       65000 65003 i
 *  ec    RD: 2.2.2.2:1 auto-discovery 0011:1111:1111:1100:0000
                                2.2.2.2                -       100     0       65000 65003 i
```

At this point, the Type-1 AD per ES, Type-1 AD per EVI, and Type-3 IMET routes have been received. - Multiple copies means we can do ECMP right.

Local MultiHomed Host

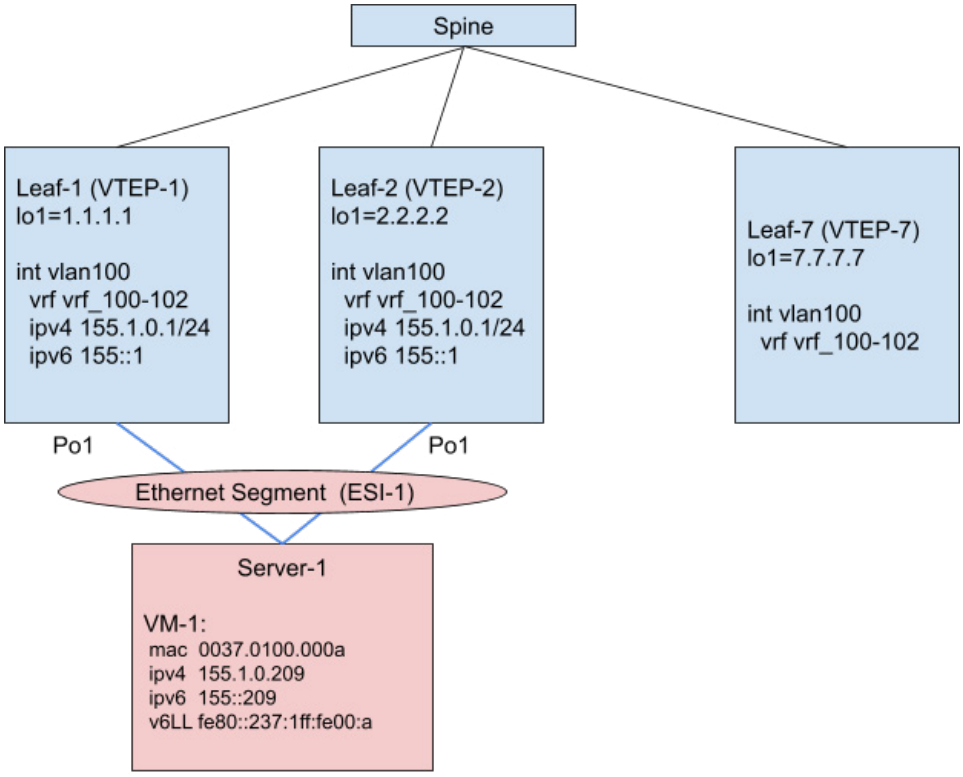

Figure 12: MAC-IP route advertisement for A-A Multihoming host

Step 1a: A-A MH Leaf-1 learns ARP binding of VM-1 (155.1.0.209), in VLAN 100 because of hashing load balancing algorithm (your mileage may vary)

A-A MH Leaf-1 IPV4 ARP Binding

```
Leaf-1.23:28:18#sh ip arp vrf vrf_100-102 155.1.0.209
Address         Age (sec)  Hardware Addr   Interface
155.1.0.209           N/A  0037.0100.000a  Vlan100, Port-Channel1
```

Step 1b: Likewise, A-A MH Leaf-2 learns the IPv6 ND binding of VM-1 (155::209), in VLAN 100

A-A MH Leaf-2 IPv6 ND Binding

```
Leaf-2.01:13:19#sh ipv6 nei vrf vrf_100-102 155::209
IPv6 Address                                  Age Hardware Addr    State Interface
155::209                                      N/A 0037.0100.000a   REACH Vl100, Port-Channel1
```

Step 2a: Leaf-1 advertises EVPN Type 2 MAC-IP route to its EVPN peers

A-A MH Leaf-1 originates Type 2 MAC-IP for IP2MAC binding 155.1.0.209

```
Leaf-1.23:53:22#sh bgp evpn route-type mac-ip 155.1.0.209 next-hop 0.0.0.0 detail
BGP routing table information for VRF default
Router identifier 1.1.1.1, local AS number 65001
BGP routing table entry for mac-ip 0037.0100.000a 155.1.0.209, Route Distinguisher:
1.1.1.1:100
 Paths: 1 available
  Local
    - from - (0.0.0.0)
      Origin IGP, metric -, localpref -, weight 0, valid, local, best
      Extended Community: Route-Target-AS:100:100 Route-Target-AS:5000:5000
TunnelEncap:tunnelTypeVxlan EvpnRouterMac:28:99:3a:25:fd:d3
      VNI: 100 L3 VNI: 5000 ESI:
```

- A-A MH Leaf-1 originates an EVPN Type 2 MAC-IP route:
  - BGP nexthop = 0.0.0.0
  - RD value 1.1.1.1:100
  - ESI: 0011:1111:1111:1100:0000
- This EVPN Type 2 MAC-IP route has two route-targets:
  - RTtAS:100:100 is MAC-VRF RT
  - RTAS:5000:5000 is IP-VRF RT
- This EVPN Type 2 MAC-IP route has two VNIs:
  - VNI: 100 is L2 VNI
  - VNI: 5000 is L3 IP-VRF VNI

Step 2b: in the same way, Leaf-2 advertises EVPN Type 2 MAC-IP (IPV6 host) route to its EVPN peers

A-A MH Leaf-2 originates Type 2 MAC-IP for IPv6 ND binding 155::209

```
Leaf-2.01:13:32#sh bgp evpn route-type mac-ip 155::209 next-hop 0.0.0.0 detail
BGP routing table information for VRF default
Router identifier 2.2.2.2, local AS number 65002
BGP routing table entry for mac-ip 0037.0100.000a 155::209, Route Distinguisher: 2.2.2.2:100
 Paths: 1 available
  Local
    - from - (0.0.0.0)
      Origin IGP, metric -, localpref -, weight 0, valid, local, best
      Extended Community: Route-Target-AS:100:100 Route-Target-AS:5000:5000
TunnelEncap:tunnelTypeVxlan EvpnRouterMac:28:99:3a:d0:ca:10
      VNI: 100 L3 VNI: 5000 ESI: 0011:1111:1111:1100:0000
```

Step 3a: A-A MH Leaf-2 receives this EVPN Type 2 MAC-IP update and re-originates a proxy MAC-IP route as indicated by the proxy flag (pflag) in the EVPN ND/ARP Extended community

A-A MH Leaf-2 (MAC-IP Route)

```
Leaf-2.23:27:50# sh bgp evpn route-type mac-ip 155.1.0.209 next-hop 1.1.1.1
BGP routing table information for VRF default
Router identifier 2.2.2.2, local AS number 65002
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
 * >Ec    RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155.1.0.209
                                 1.1.1.1               -       100     0       65000 65001 i
 *  ec    RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155.1.0.209
                                 1.1.1.1               -       100     0       65000 65001 i

Leaf-2.23:27:50# sh ip arp remote vlan 100
ARP remote bindings
VLAN IP Address  MAC Address
---- ----------- --------------
100  155.1.0.209 0037.0100.000a

Leaf-2.23:27:50# sh bgp evpn route-type mac-ip 155.1.0.209 next-hop 0.0.0.0 detail
BGP routing table information for VRF default
Router identifier 2.2.2.2, local AS number 65002
BGP routing table entry for mac-ip 0037.0100.000a 155.1.0.209, Route Distinguisher:
2.2.2.2:100
 Paths: 1 available
  Local
    - from - (0.0.0.0)
      Origin IGP, metric -, localpref -, weight 0, valid, local, best
      Extended Community: Route-Target-AS:100:100 Route-Target-AS:5000:5000
TunnelEncap:tunnelTypeVxlan EvpnRouterMac:28:99:3a:d0:ca:10 EvpnNdFlags:pflag
      VNI: 100 L3 VNI: 5000 ESI: 0011:1111:1111:1100:0000
```

Step 3b: likewise, A-A MH Leaf-1 receives this EVPN Type 2 MAC-IP (IPV6 host) route and re-originate a proxy MAC-IP route

A-A MH Leaf-1 (MAC-IP Route)

```
Leaf-1.01:18:56#sh bgp evpn route-type mac-ip 155::209 next-hop 2.2.2.2
BGP routing table information for VRF default
Router identifier 1.1.1.1, local AS number 65001
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
 * >Ec    RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155::209
                                 2.2.2.2               -       100     0       65000 65002 i
 *  ec    RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155::209
                                 2.2.2.2               -       100     0       65000 65002 i

Leaf-1.01:18:56# sh ipv6 nei remote vlan 100
ARP remote bindings
VLAN IP Address MAC Address
---- ---------- --------------
100  155::209   0037.0100.000a

Leaf-1.01:18:56# sh bgp evpn route-type mac-ip 155::209 next-hop 0.0.0.0 detail
BGP routing table information for VRF default
Router identifier 1.1.1.1, local AS number 65001
BGP routing table entry for mac-ip 0037.0100.000a 155::209, Route Distinguisher: 1.1.1.1:100
 Paths: 1 available
  Local
    - from - (0.0.0.0)
      Origin IGP, metric -, localpref -, weight 0, valid, local, best
      Extended Community: Route-Target-AS:100:100 Route-Target-AS:5000:5000
TunnelEncap:tunnelTypeVxlan EvpnRouterMac:28:99:3a:25:fd:d3 EvpnNdFlags:pflag
      VNI: 100 L3 VNI: 5000 ESI: 0011:1111:1111:1100:0000
```

Step 4a: A-A MH Leaf-2 install received MAC-IP in local ARP table:

For IP2MAC binding learnt from ARP traffic, the binding is aged based on the ARP aging timeout configuration. On the other hand, if the IP2MAC binding is learnt via the EVPN control plane, the subsequent ARP entry isn't aged out. Therefore, the EVPN bindings alway show as '-' under the "Age" column.

A-A MH Leaf-2 IPV4 ARP Binding

```
Leaf-2.23:55:54#sh ip arp vrf vrf_100-102 155.1.0.209
Address         Age (sec)  Hardware Addr   Interface
155.1.0.209             -  0037.0100.000a  Vlan100, Port-Channel1
```

Step 4b: similarly, A-A MH Leaf-1 install received MAC-IP (IPV6) in local ARP table:

A-A MH Leaf-1 IPV6 ARP Binding

```
Leaf-1.01:18:59#show ipv6 neighbors vrf vrf_100-102 155::209
IPv6 Address                              Age Hardware Addr    State Interface
155::209                                    - 0037.0100.000a   REACH Vl100, Port-Channel1
```

L2 ECMP

The Type-2 MAC-IP routes received from either Leaf-1 or Leaf-2 need to be installed to form an L2 ECMP.

Remote EVPN Peer Leaf-7 (Single Homed)

```
Leaf-7.00:09:29#sh bgp evpn route-type mac-ip 0037.0100.000a
BGP routing table information for VRF default
Router identifier 7.7.7.7, local AS number 65007
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
 * >Ec    RD: 1.1.1.1:100 mac-ip 0037.0100.000a
                                 1.1.1.1               -       100     0       65000 65001 i
 *  ec    RD: 1.1.1.1:100 mac-ip 0037.0100.000a
                                 1.1.1.1               -       100     0       65000 65001 i
 * >Ec    RD: 2.2.2.2:100 mac-ip 0037.0100.000a
                                 2.2.2.2               -       100     0       65000 65002 i
 *  ec    RD: 2.2.2.2:100 mac-ip 0037.0100.000a
                                 2.2.2.2               -       100     0       65000 65002 i
 * >Ec    RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155.1.0.209
                                 1.1.1.1               -       100     0       65000 65001 i
 *  ec    RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155.1.0.209
                                 1.1.1.1               -       100     0       65000 65001 i
 * >Ec    RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155.1.0.209
                                 2.2.2.2               -       100     0       65000 65002 i
 *  ec    RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155.1.0.209
                                 2.2.2.2               -       100     0       65000 65002 i
 * >Ec    RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155::209
                                 1.1.1.1               -       100     0       65000 65001 i
 *  ec    RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155::209
                                 1.1.1.1               -       100     0       65000 65001 i
 * >Ec    RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155::209
                                 2.2.2.2               -       100     0       65000 65002 i
 *  ec    RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155::209
                                 2.2.2.2               -       100     0       65000 65002 i
```

There are 12 mac-ip routes in total, reflected by the two Spines (8.8.8.8 and 9.9.9.9) acting as RRs, where :

- 6 of them originated from A-A MH Leaf-1 (1.1.1.1) where 2 routes carry VM-1's MAC address only, 2 routes for IPv4 ARP binding and 2 routes for IPv6 ND binding.
- 6 of them originated from A-A MH Leaf-2 (2.2.2.2) where 2 routes carry VM-1's MAC address only, 2 routes for IPv4 ARP binding and 2 routes for IPv6 ND binding.

Remote Peer Leaf-7 supports "MAC aliasing" which makes use of the received Type 1 AD route and Type 2 MAC-IP route to perform L2 ECMP for VM-1. Check MAC programming for VM-1 in remote Peer Leaf-7:

Remote Peer Leaf-7 supports "MAC aliasing" which makes use of the received Type 1 AD route and Type 2 MAC-IP route to perform L2 ECMP for VM-1. Check MAC programming for VM-1 in remote Peer Leaf-7:

Remote EVPN Peer Leaf-7 (Single Homed)

```
Leaf-7.06:37:50#sh vxlan address-table address 0037.0100.0001
          Vxlan Mac Address Table
----------------------------------------------------------------------

VLAN  Mac Address     Type      Prt  VTEP             Moves   Last Move
----  -----------     ----      ---  ----             -----   ---------
 100  0037.0100.0001  EVPN      Vx1  1.1.1.1          6       6:45:40 ago
                                     2.2.2.2
Total Remote Mac Addresses for this criterion: 1
```

L3 ECMP (Symmetric IRB with Proxy MAC-IP)

Remote MH Peer Leaf-7 received MAC-IP route for VM-1 from both A-A MH leaf-1 and leaf-2:

Remote Single homed VTEP leaf-7

```
Leaf-7.00:07:21#sh bgp evpn route-type mac-ip 155.1.0.209
BGP routing table information for VRF default
Router identifier 7.7.7.7, local AS number 65007
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local Nexthop

          Network          Next Hop            Metric  LocPref Weight  Path
 * >Ec   RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155.1.0.209
                    1.1.1.1          -       100     0       65000 65001 i
 *  ec   RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155.1.0.209
                    1.1.1.1          -       100     0       65000 65001 i
 * >Ec   RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155.1.0.209
                    2.2.2.2          -       100     0       65000 65002 i
 *  ec   RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155.1.0.209
                    2.2.2.2          -       100     0       65000 65002 i

Leaf-7.00:11:44#sh bgp evpn route-type mac-ip 155::209
BGP routing table information for VRF default
Router identifier 7.7.7.7, local AS number 65007
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local Nexthop

          Network          Next Hop            Metric  LocPref Weight  Path
 * >Ec   RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155::209
                    1.1.1.1          -       100     0       65000 65001 i
*  ec   RD: 1.1.1.1:100 mac-ip 0037.0100.000a 155::209
                    1.1.1.1          -       100     0       65000 65001 i
 * >Ec   RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155::209
                    2.2.2.2          -       100     0       65000 65002 i
 *  ec   RD: 2.2.2.2:100 mac-ip 0037.0100.000a 155::209
                    2.2.2.2          -       100     0       65000 65002 i
```

Remote MH Peer Leaf-7 installs both IPv4 and IPv6 hosts (from received MAC-IP) in its local routing table

Remote A-A MH Leaf-7

```
Leaf-7.00:06:13#show ip route vrf vrf_100-102 155.1.0.209

VRF: vrf_100-102
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B - BGP, B I - iBGP, B E - eBGP,
       R - RIP, I L1 - IS-IS level 1, I L2 - IS-IS level 2,
       O3 - OSPFv3, A B - BGP Aggregate, A O - OSPF Summary,
       NG - Nexthop Group Static Route, V - VXLAN Control Service,
       DH - DHCP client installed default route, M - Martian,
       DP - Dynamic Policy Route, L - VRF Leaked,
       RC - Route Cache Route

 B E      155.1.0.209/32 [200/0] via VTEP 2.2.2.2 VNI 5000 router-mac 28:99:3a:d0:ca:10
                                 via VTEP 1.1.1.1 VNI 5000 router-mac 28:99:3a:25:fd:d3

Leaf-7.00:12:58#show ipv6 route vrf vrf_100-102 155::209

VRF: vrf_100-102
Routing entry for 155::209
Codes: C - connected, S - static, K - kernel, O3 - OSPFv3, B - BGP, R - RIP, A B - BGP
Aggregate, I L1 - IS-IS level 1, I L2 - IS-IS level 2, DH - DHCP, NG - Nexthop Group Static
Route, M - Martian, DP - Dynamic Policy Route, L - VRF Leaked, RC - Route Cache Route

 B        155::209/128 [200/0]
           via VTEP 2.2.2.2 VNI 5000 router-mac 28:99:3a:d0:ca:10
           via VTEP 1.1.1.1 VNI 5000 router-mac 28:99:3a:25:fd:d3
```

Summary

EVPN All-Active Multihoming mechanism is an IETF standard supported by many -vendors. This technology adds high availability with path redundancy in Data center network designs. It offers the following benefits:

- Overlay ECMP for both IP and IPv6 hosts
- All-Active PE routers redundancy to protect against both Ethernet Segment Link and PE node failure.

References

- EVPN Design Guide
- Multi-Tenant EVPN VXLAN IRB Configuration & Verification Guide

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office 1390
Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor

Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard

#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062