Configuring VXLAN EVPN Multihoming

This chapter contains the following sections:

VXLAN EVPN Multihoming Overview

Introduction to Multihoming

Cisco Nexus platforms support vPC-based multihoming, where a pair of switches act as a single device for redundancy and both switches function in an active mode. With Cisco Nexus 9000 Series switches in VXLAN BGP EVPN environment, there are two solutions to support Layer 2 multihoming; the solutions are based on the Traditional vPC (emulated or virtual IP address) and the BGP EVPN techniques.

Traditional vPC utilizes a consistency check that is a mechanism used by the two switches that are configured as a vPC pair to exchange and verify their configuration compatibility. The BGP EVPN technique does not have the consistency check mechanism, but it uses LACP to detect the misconfigurations. It also eliminates the Peer Link that is traditionally used by vPC and it offers more flexibility as each VTEP can be a part of one or more redundancy groups. It can potentially support many VTEPs in a given group.

BGP EVPN Multihoming

When using BGP EVPN control plane, each switch can use its own local IP address as the VTEP IP address and it still provides an active/active redundancy. BGP EVPN based multihoming further provides fast convergence during certain failure scenarios, that otherwise cannot be achieved without a control protocol (data plane flood and learn).

BGP EVPN Multihoming Terminology

See this section for the terminology used in BGP EVPN multihoming:

- EVI: EVPN instance represented by the VNI.
- MAC-VRF: A container to house virtual forwarding table for MAC addresses. A unique route distinguisher and import/export target can be configured per MAC-VRF.
- ES: Ethernet Segment that can constitute a set of bundled links.
- ESI: Ethernet Segment Identifier to represent each ES uniquely across the network.

EVPN Multihoming Implementation

The EVPN overlay draft specifies adaptations to the BGP MPLS based EVPN solution to enable it to be applied as a network virtualization overlay with VXLAN encapsulation. The Provider Edge (PE) node role in BGP MPLS EVPN is equivalent to VTEP/Network Virtualization Edge device (NVE), where VTEPs use control plane learning and distribution via BGP for remote addresses instead of data plane learning.

There are 5 different route types currently defined:

- Ethernet Auto-Discovery (EAD) Route
- MAC advertisement Route
- Inclusive Multicast Route
- Ethernet Segment Route
- IP Prefix Route

BGP EVPN running on Cisco NX-OS uses route type-2 to advertise MAC and IP (host) information, route type-3 to carry VTEP information (specifically for ingress replication), and the EVPN route type-5 allows advertisements of IPv4 or IPv6 prefixes in an Network Layer Reachability Information (NLRI) with no MAC addresses in the route key.

With the introduction of EVPN multihoming, Cisco NX-OS software utilizes Ethernet Auto-discovery (EAD) route, where Ethernet Segment Identifier and the Ethernet Tag ID are considered to be part of the prefix in the NLRI. Since the end points reachability is learned via the BGP control plane, the network convergence time is a function of the number of MAC/IP routes that must be withdrawn by the VTEP in case of a failure scenario. To deal with such condition, each VTEP advertises a set of one or more Ethernet Auto-Discovery per ES routes for each locally attached Ethernet Segment and upon a failure condition to the attached segment, the VTEP withdraws the corresponding set of Ethernet Auto-Discovery per ES routes.

Ethernet Segment Route is the other route type that is being used by Cisco NX-OS software with EVPN multihoming, mainly for Designated Forwarder (DF) election for the BUM traffic. If the Ethernet Segment is multihomed, the presence of multiple DFs could result in forwarding the loops in addition to the potential packet duplication. Therefore, the Ethernet Segment Route (Type 4) is used to elect the Designated Forwarder and to apply Split Horizon Filtering. All VTEPs/PEs that are configured with an Ethernet Segment originate this route.

To summarize the new implementation concepts for the EVPN multihoming:

- EAD/ES: Ethernet Auto Discovery Route per ES that is also referred to as type-1 route. This route is used to converge the traffic faster during access failure scenarios. This route has Ethernet Tag of 0xFFFFFFFF.

- EAD/EVI: Ethernet Auto Discovery Route per EVI that is also referred to as type-1 route. This route is used for aliasing and load balancing when the traffic only hashes to one of the switches. This route cannot have Ethernet Tag value of 0xFFFFFF to differentiate it from the EAD/ES route.
- ES: Ethernet Segment route that is also referred to as type-4 route. This route is used for DF election for BUM traffic.
- Aliasing: It is used for load balancing the traffic to all the connected switches for a given Ethernet Segment using the type-1 EAD/EVI route. This is done irrespective of the switch where the hosts are actually learned.
- Mass Withdrawal: It is used for fast convergence during the access failure scenarios using the type-1 EAD/ES route.
- DF Election: It is used to prevent forwarding of the loops and the duplicates as only a single switch is allowed to decap and forward the traffic for a given Ethernet Segment.
- Split Horizon: It is used to prevent forwarding of the loops and the duplicates for the BUM traffic. Only the BUM traffic that originates from a remote site is allowed to be forwarded to a local site.

EVPN Multihoming Redundancy Group

Consider the dually homed topology, where switches L1 and L2 are distributed anycast VXLAN gateways that perform Integrated Routing and Bridging (IRB). Host H2 is connected to an access switch that is dually homed to both L1 and L2.

The access switch is connected to L1 and L2 via a bundled pair of physical links. The switch is not aware that the bundle is configured on two different devices on the other side. However, both L1 and L2 must be aware that they are a part of the same bundle.

Note that there is no Peer Link between L1 and L2 switches and each switch can have similar multiple bundle links that are shared with the same set of neighbors.

To make the switches L1 and L2 aware that they are a part of the same bundle link, the NX-OS software utilizes the Ethernet Segment Identifier (ESI) and the system MAC address (system-mac) that is configured under the interface (PO).

Ethernet Segment Identifier

EVPN introduces the concept of Ethernet Segment Identifier (ESI). Each switch is configured with a 10 byte ESI value under the bundled link that they share with the multihomed neighbor. The ESI value can be manually configured or auto-derived.

LACP Bundling

LACP can be turned ON for detecting ESI misconfigurations on the multihomed port channel bundle as LACP sends the ESI configured MAC address value to the access switch. LACP is not mandated along with ESI. A given ESI interface (PO) shares the same ESI ID across the VTEPs in the group.

The access switch receives the same configured MAC value from both switches (L1 and L2). Therefore, it puts the bundled link in the UP state. Since the ES MAC can be shared across all the Ethernet-segments on the switch, LACP PDUs use ES MAC as system MAC address and the admin_key carries the ES ID.

Cisco recommends running LACP between the switches and the access devices since LACP PDUs have a mechanism to detect and act on the misconfigured ES IDs. In case there is mismatch on the configured ES ID under the same PO, LACP brings down one of the links (first link that comes online stays up). By default, on most Cisco Nexus platforms, LACP sets a port to the suspended state if it does not receive an LACP PDU from the peer. This is based on the **lacp suspend-individual** command that is enabled by default. This command helps in preventing loops that are created due to the ESI configuration mismatch. Therefore, it is recommended to enable this command on the port-channels on the access switches and the servers.

In some scenarios (for example, POAP or NetBoot), it can cause the servers to fail to boot up because they require LACP to logically bring up the port. In case you are using static port channel and you have mismatched ES IDs, the MAC address gets learned from both L1 and L2 switches. Therefore, both the switches advertise the same MAC address belonging to different ES IDs that triggers the MAC address move scenario. Eventually, no traffic is forwarded to that node for the MAC addresses that are learned on both L1 and L2 switches.

Guidelines and Limitations for VXLAN EVPN Multihoming

See the following limitations for configuring VXLAN EVPN Multihoming:

- VXLAN EVPN Multihoming works with the iBGP or eBGP control plane. iBGP is preferred.
- If iBGP is used with VXLAN EVPN Multihoming, the administrative distance for local learned endpoints value must be lower than the value of iBGP.

Note The default value for local learned endpoints is 190, the default value for eBGP is 20, and the default value for iBGP is 200.

- If eBGP is used with VXLAN EVPN Multihoming, the administrative distance for local learned endpoints must be lower than the value of eBGP. The administrative distance can be changed by entering the **fabric forwarding admin-distance** *distance* command.

Note The default value for local learned endpoints is 190, the default value for eBGP is 20, and the default value for iBGP is 200.

- EVPN Multihoming is supported on the Cisco Nexus 9300 platform switches only and it is not supported on the Cisco Nexus 9200, 9300-EX/-FX/-FXP/-FX2 and 9500 platform switches. The Cisco Nexus 9500 platform switches can be used as Spine switches, but they cannot be used as VTEPs.
- EVPN Multihoming requires that all switches in a given network must be EVPN Multihoming capable.. Mixing platforms with and without EVPN Multihoming is not supported.
- EVPN multihoming is not supported on FEX.
- Beginning with Cisco NX-OS Release 7.0(3)I5(2), ARP suppression is supported with EVPN multihoming.
- EVPN Multihoming is supported with multihoming to two switches only.
- To enable EVPN Multihoming, the spine switches must be running the minimum software version as Cisco NX-OS Release 7.0(3)I5(2) or later.
- Switchport trunk native VLAN is not supported on the trunk interfaces.

- Cisco recommends enabling LACP on ES PO.
- IPv6 is not currently supported.
- ISSU is not supported if ESI is configured on the Cisco Nexus 9300 Series switches.

Configuring VXLAN EVPN Multihoming

Enabling EVPN Multihoming

Cisco NX-OS allows either vPC based EVPN multihoming or ESI based EVPN multihoming. Both features should not be enabled together. ESI based multihoming is enabled using **evpn esi multihoming** CLI command. It is important to note that the command for ESI multihoming enables the Ethernet-segment configurations and the generation of Ethernet-segment routes on the switches.

The receipt of type-1 and type-2 routes with valid ESI and the path-list resolution are not tied to the **evpn esi multihoming** command. If the switch receives MAC/MAC-IP routes with valid ESI and the command is not enabled, the ES based path resolution logic still applies to these remote routes. This is required for interoperability between the vPC enabled switches and the ESI enabled switches.

Complete the following steps to configure EVPN multihoming:

Before you begin

VXLAN should be configured with BGP-EVPN before enabling EVPN ESI multihoming.

Procedure

	Command or Action	Purpose
Step 1	**evpn esi multihoming**	Enables EVPN multihoming globally.

Step 2	**address-family l2vpn evpn maximum-paths <>maximum-paths ibgp <>**
	Example:
	`address-family l2vpn evpn`
	`maximum-paths 64`
	`maximum-paths ibgp 64`
		Enables BGP maximum-path to enable ECMP for the MAC routes. Otherwise, the MAC routes have only 1 VTEP as the next-hop. This configuration is needed under BGP in Global level.

Step 3	**evpn multihoming core-tracking**	Enables EVPN multihoming core-links. It tracks the uplink interfaces towards the core. If all uplinks are down, the local ES based the POs is shut down/suspended. This is mainly used to avoid black-holing South-to-North traffic when no uplinks are available.
Step 4	**interface port-channel Ethernet-segment <>System-mac <>** **Example:**	Configures the local Ethernet Segment ID. The ES ID has to match on VTEPs where the PO is multihomed. The Ethernet Segment ID should be unique per PO.

	Command or Action	Purpose
	`ethernet-segment 11` `system-mac 0000.0000.0011`	Configures the local system-mac ID that has to match on the VTEPs where the PO is multihomed. The system-mac address can be shared across multiple POs.
Step 5	**hardware access-list tcam region vpc-convergence 256** **Example:** `hardware access-list tcam region vpc-convergence 256`	Configures the TCAM. This command is used to configure the split horizon ACLs in the hardware. This command avoids BUM traffic duplication on the shared ES POs.

VXLAN EVPN Multihoming Configuration Examples

See the sample VXLAN EVPN multihoming configuration on the switches:

```
Switch 1 (L1)

evpn esi multihoming

router bgp 1001
   address-family l2vpn evpn
   maximum-paths ibgp 2

interface Ethernet2/1
  no switchport
  evpn multihoming core-tracking
  mtu 9216
  ip address 10.1.1.1/30
  ip pim sparse-mode
  no shutdown

interface Ethernet2/2
  no switchport
  evpn multihoming core-tracking
  mtu 9216
  ip address 10.1.1.5/30
  ip pim sparse-mode
  no shutdown

interface port-channel11
  switchport mode trunk
  switchport trunk allowed vlan 901-902,1001-1050
  ethernet-segment 2011
    system-mac 0000.0000.2011
  mtu 9216


Switch 2 (L2)

evpn esi multihoming

router bgp 1001
   address-family l2vpn evpn
   maximum-paths ibgp 2
```

```
interface Ethernet2/1
  no switchport
  evpn multihoming core-tracking
  mtu 9216
  ip address 10.1.1.2/30
  ip pim sparse-mode
  no shutdown

interface Ethernet2/2
  no switchport
  evpn multihoming core-tracking
  mtu 9216
  ip address 10.1.1.6/30
  ip pim sparse-mode
  no shutdown

interface port-channel11
  switchport mode trunk
  switchport access vlan 1001
  switchport trunk allowed vlan 901-902,1001-1050
  ethernet-segment 2011
    system-mac 0000.0000.2011
  mtu 9216
```

Configuring Layer 2 Gateway STP

Layer 2 Gateway STP Overview

Beginning with Cisco NX-OS Release 7.0(3)I5(2), EVPN multihoming is supported with the Layer 2 Gateway Spanning Tree Protocol (L2G-STP). The Layer 2 Gateway Spanning Tree Protocol (L2G-STP) builds a loop-free tree topology. However, the Spanning Tree Protocol root must always be in the VXLAN fabric. A bridge ID for the Spanning Tree Protocol consists of a MAC address and the bridge priority. When the system is running in the VXLAN fabric, the system automatically assigns the VTEPs with the MAC address c84c.75fa.6000 from a pool of reserved MAC addresses. As a result, each switch uses the same MAC address for the bridge ID emulating a single logical pseudo root.

The Layer 2 Gateway Spanning Tree Protocol (L2G-STP) is disabled by default on EVPN ESI multihoming VLANs. Use the **spanning-tree domain enable** CLI command to enable L2G-STP on all VTEPs. With L2G-STP enabled, the VXLAN fabric (all VTEPs) emulates a single pseudo root switch for the customer access switches. The L2G-STP is initiated to run on all VXLAN VLANs by default on boot up and the root is fixed on the overlay. With L2G-STP, the root-guard gets enabled by default on all the access ports. Use **spanning-tree domain** <*id*> to additionally enable Spanning Tree Topology Change Notification(STP-TCN), to be tunneled across the fabric.

All the access ports from VTEPs connecting to the customer access switches are in a *desg* forwarding state by default. All ports on the customer access switches connecting to VTEPs are either in root-port forwarding or alt-port blocking state. The root-guard kicks in if better or superior STP information is received from the customer access switches and it puts the ports in the *blk l2g_inc* state to secure the root on the overlay-fabric and to prevent a loop.

Guidelines for Moving to Layer 2 Gateway STP

Complete the following steps to move to Layer 2 gateway STP:

- With Layer 2 Gateway STP, root guard is enabled by default on all the access ports.
- With Layer 2 Gateway STP enabled, the VXLAN fabric (all VTEPs) emulates a single pseudo-root switch for the customer access switches.
- All access ports from VTEPs connecting to the customer access switches are in the **Desg FWD** state by default.
- All ports on customer access switches connecting to VTEPs are either in the root-port FWD or Altn BLK state.
- Root guard is activated if superior spanning-tree information is received from the customer access switches. This process puts the ports in **BLK L2GW_Inc** state to secure the root on the VXLAN fabric and prevent a loop.
- Explicit domain ID configuration is needed to enable spanning-tree BPDU tunneling across the fabric.
- As a best practice, you should configure all VTEPs with the lowest spanning-tree priority of all switches in the spanning-tree domain to which they are attached. By setting all the VTEPs as the root bridge, the entire VXLAN fabric appears to be one virtual bridge.
- ESI interfaces should not be enabled in spanning-tree edge mode to allow Layer 2 Gateway STP to run across the VTEP and access layer.
- You can continue to use ESIs or orphans (single-homed hosts) in spanning-tree edge mode if they directly connect to hosts or servers that do not run Spanning Tree Protocol and are end hosts.
- Configure all VTEPs that are connected by a common customer access layer in the same Layer 2 Gateway STP domain. Ideally, all VTEPs on the fabric on which the hosts reside and to which the hosts can move.
- The Layer 2 Gateway STP domain scope is global, and all ESIs on a given VTEP can participate in only one domain.
- Mappings between Multiple Spanning Tree (MST) instances and VLANs must be consistent across the VTEPs in a given Layer 2 Gateway STP domain.
- Ensure that the root of an STP domain local to the VXLAN fabric is a VTEP or placed within the fabric.
- Non-Layer 2 Gateway STP enabled VTEPs cannot be directly connected to Layer 2 Gateway STP-enabled VTEPs. Performing this action results in conflicts and disputes because the non-Layer 2 Gateway STP VTEP keeps sending BPDUs and it can steer the root outside.
- Keep the current edge and the BPDU filter configurations on both the Cisco Nexus switches and the access switches after upgrading to the latest build.
- Enable Layer 2 Gateway STP on all the switches with a recommended priority and the *mst* instance mapping as needed. Use the commands **spanning-tree domain enable** and **spanning-tree mst <instance-id's> priority 8192**.
- Remove the BPDU filter configurations on the switch side first.
- Remove the BPDU filter configurations and the edge on the customer access switch.

Now the topology converges with Layer 2 Gateway STP and any blocking of the redundant connections is pushed to the access switch layer.

Enabling Layer 2 Gateway STP on a Switch

Complete the following steps to enable Layer 2 Gateway STP on a switch.

Procedure

	Command or Action	Purpose
Step 1	**spanning-tree mode <rapid-pvst, mst>**	Enables Spanning Tree Protocol mode.
Step 2	**spanning-tree domain enable**	Enables Layer 2 Gateway STP on a switch. It disables Layer 2 Gateway STP on all EVPN ESI multihoming VLANs.
Step 3	**spanning-tree domain 1**	Explicit domain ID is needed to tunnel encoded BPDUs to the core and processes received from the core.
Step 4	**spanning-tree mst <id> priority 8192**	Configures Spanning Tree Protocol priority.
Step 5	**spanning-tree vlan <id> priority 8192**	Configures Spanning Tree Protocol priority.
Step 6	**spanning-tree domain disable**	Disables Layer 2 Gateway STP on a VTEP.

Example

All Layer 2 Gateway STP VLANs should be set to a lower spanning-tree priority than the customer-edge (CE) topology to help ensure that the VTEP is the spanning-tree root for this VLAN. If the access switches have a higher priority, you can set the Layer 2 Gateway STP priority to 0 to retain the Layer 2 Gateway STP root in the VXLAN fabric. See the following configuration example:

```
switch# show spanning-tree summary
Switch is in mst mode (IEEE Standard)
Root bridge for: MST0000
L2 Gateway STP bridge for: MST0000
L2 Gateway Domain ID: 1
Port Type Default                        is disable
Edge Port [PortFast] BPDU Guard Default  is disabled
Edge Port [PortFast] BPDU Filter Default is disabled
Bridge Assurance                         is enabled
Loopguard Default                        is disabled
Pathcost method used                     is long
PVST Simulation                          is enabled
STP-Lite                                 is disabled

Name                   Blocking Listening Learning Forwarding STP Active
---------------------- -------- --------- -------- ---------- ----------
MST0000                       0         0        0         12         12
---------------------- -------- --------- -------- ---------- ----------
1 mst                         0         0        0         12         12
```

```
switch# show spanning-tree vlan 1001

MST0000
  Spanning tree enabled protocol mstp

  Root ID    Priority    8192
             Address     c84c.75fa.6001   L2G-STP reserved mac+ domain id
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    8192   (priority 8192 sys-id-ext 0)
             Address     c84c.75fa.6001
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec
```

The output displays that the spanning-tree priority is set to 8192 (the default is 32768). Spanning-tree priority is set in multiples of 4096. The priority for individual instances is calculated as the priority and the Instance_ID. In this case, the priority is calculated as 8192 + 0 = 8192. With Layer 2 Gateway STP, access ports (VTEP ports connected to the access switches) have root guard enabled. If a superior BPDU is received on an edge port of a VTEP, the port is placed in the Layer 2 Gateway inconsistent state until the condition is cleared as displayed in the following example:

```
2016 Aug 29 19:14:19 TOR9-leaf4 %$ VDC-1 %$ %STP-2-L2GW_BACKBONE_BLOCK: L2 Gateway Backbone
 port inconsistency blocking port Ethernet1/1 on MST0000.
2016 Aug 29 19:14:19 TOR9-leaf4 %$ VDC-1 %$ %STP-2-L2GW_BACKBONE_BLOCK: L2 Gateway Backbone
 port inconsistency blocking port port-channel13 on MST0000.

switch# show spanning-tree

MST0000
  Spanning tree enabled protocol mstp
  Root ID    Priority    8192
             Address     c84c.75fa.6001
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    8192   (priority 8192 sys-id-ext 0)
             Address     c84c.75fa.6001
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Po1              Desg FWD 20000     128.4096 Edge P2p
Po2              Desg FWD 20000     128.4097 Edge P2p
Po3              Desg FWD 20000     128.4098 Edge P2p
Po12            Desg BKN*2000       128.4107 P2p *L2GW_Inc
Po13            Desg BKN*1000       128.4108 P2p *L2GW_Inc
Eth1/1        Desg BKN*2000       128.1    P2p *L2GW_Inc
```

To disable Layer 2 Gateway STP on a VTEP, enter the **spanning-tree domain disable** CLI command. This command disables Layer 2 Gateway STP on all EVPN ESI multihomed VLANs. The bridge MAC address is restored to the system MAC address, and the VTEP may not necessarily be the root. In the following case, the access switch has assumed the root role because Layer 2 Gateway STP is disabled:

```
switch(config)# spanning-tree domain disable

switch# show spanning-tree summary
Switch is in mst mode (IEEE Standard)
Root bridge for: none
L2 Gateway STP                          is disabled
Port Type Default                       is disable
Edge Port [PortFast] BPDU Guard Default  is disabled
Edge Port [PortFast] BPDU Filter Default is disabled
Bridge Assurance                        is enabled
Loopguard Default                       is disabled
Pathcost method used                    is long
PVST Simulation                         is enabled
STP-Lite                                is disabled

Name                   Blocking Listening Learning Forwarding STP Active
---------------------- -------- --------- -------- ---------- ----------
MST0000                       4         0        0          8         12
---------------------- -------- --------- -------- ---------- ----------
1 mst                         4         0        0          8         12


switch# show spanning-tree vlan 1001

MST0000
  Spanning tree enabled protocol mstp
  Root ID    Priority    4096
             Address     00c8.8ba6.5073
             Cost        0
             Port        4108 (port-channel13)
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    8192   (priority 8192 sys-id-ext 0)
             Address     5897.bd1d.db95
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec
```

With Layer 2 Gateway STP, the access ports on VTEPs cannot be in an edge port, because they behave like normal spanning-tree ports, receiving BPDUs from the access switches. In that case, the access ports on VTEPs lose the advantage of rapid transmission, instead forwarding on Ethernet segment link flap. (They have to go through a proposal and agreement handshake before assuming the FWD-Desg role).

Configuring VXLAN EVPN Multihoming Traffic Flows

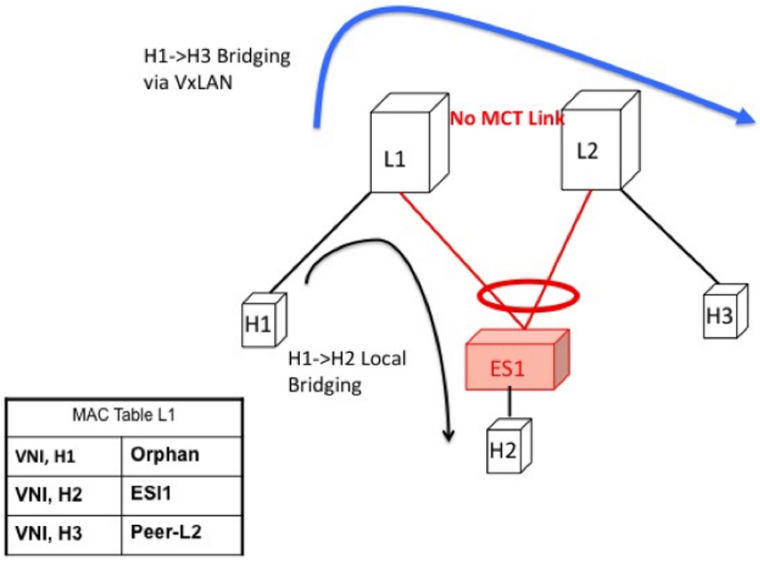
EVPN Multihoming Local Traffic Flows

All switches that are a part of the same redundancy group (as defined by the ESI) act as a single virtual switch with respect to the access switch/host. However, there is no Peer Link present to bridge and route the traffic for local access.

Locally Bridged Traffic

Host H2 is dually homed whereas hosts H1 and H3 are single-homed (also known as orphans). The traffic is bridged locally from H1 to H2 via L1. However, if the packet needs to be bridged between the orphans H1 and H3, the packet must be bridged via the VXLAN overlay.

Figure 1: Local Bridging at L1. H1->H3 bridging via VXLAN. In vPC, H1->H3 will be via Peer Link.

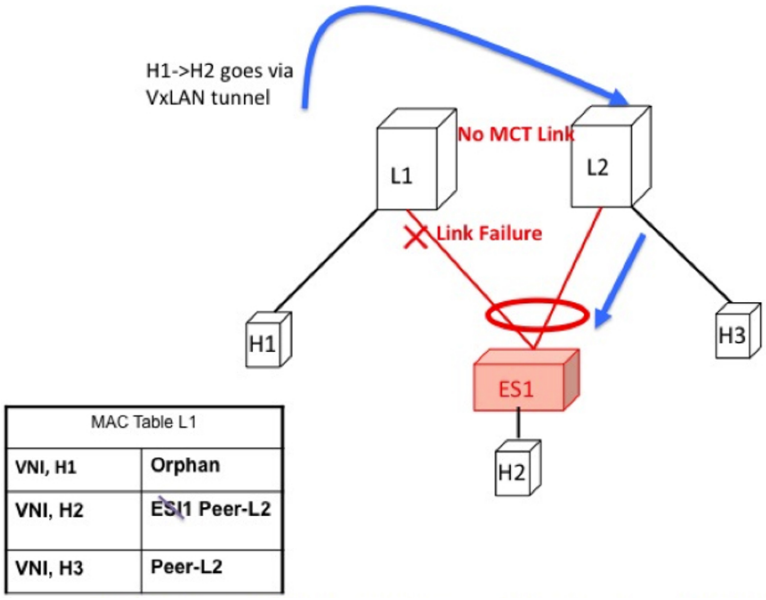


Access Failure for Locally Bridged Traffic

If the ESI link at L1 fails, there is no path for the bridged traffic to reach from H1 to H2 except via the overlay. Therefore, the local bridged traffic takes the sub-optimal path, similar to the H1 to H3 orphan flow.

Note When such condition occurs, the MAC table entry for H2 changes from a local route pointing to a port channel interface to a remote overlay route pointing to peer-ID of L2. The change gets percolated in the system from BGP.

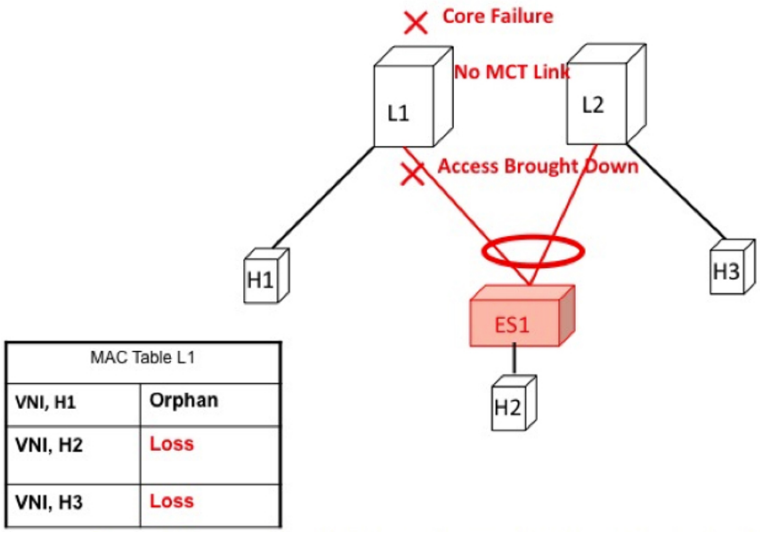
Figure 2: ES1 failure on L1. H1->H2 is now bridged over VXLAN tunnel.



Core Failure for Locally Bridged Traffic

If switch L1 gets isolated from the core, it must not continue to attract access traffic, as it will not be able to encapsulate and send it on the overlay. This means that the access links must be brought down at L1 if L1 loses core reachability. In this scenario, orphan H1 loses all connectivity to both remote and locally attached hosts since there is no dedicated Peer Link.

Figure 3: Core failure on L1. H1->H2 loses all connectivity as there is no Peer Link.



Locally Routed Traffic

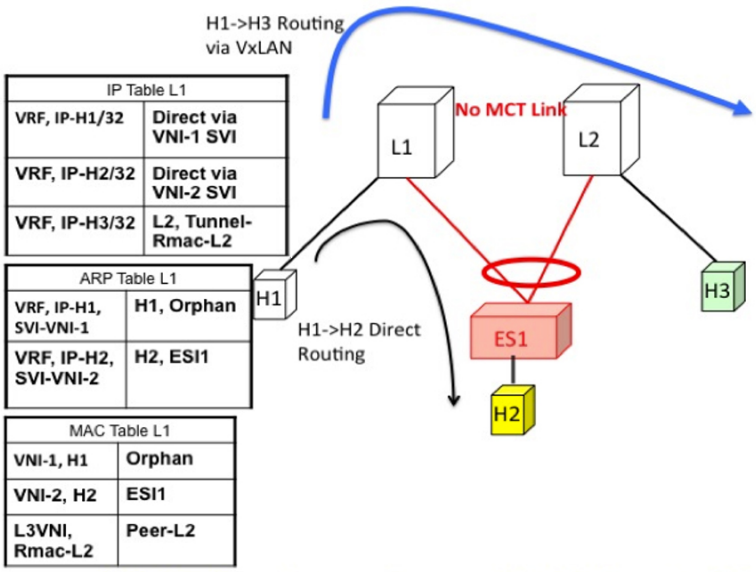
Consider H1, H2, and H3 being in different subnets and L1/L2 being distributed anycast gateways.

Any packet that is routed from H1 to H2 is directly sent from L1 via native routing.

However, host H3 is not a locally attached adjacency, unlike in vPC case where the ARP entry syncs to L1 as a locally attached adjacency. Instead, H3 shows up as a remote host in the IP table at L1, installed in the context of L3 VNI. This packet must be encapsulated in the router-MAC of L2 and routed to L2 via VXLAN overlay.

Therefore, routed traffic from H1 to H3 takes place exactly in the same fashion as routed traffic between truly remote hosts in different subnets.

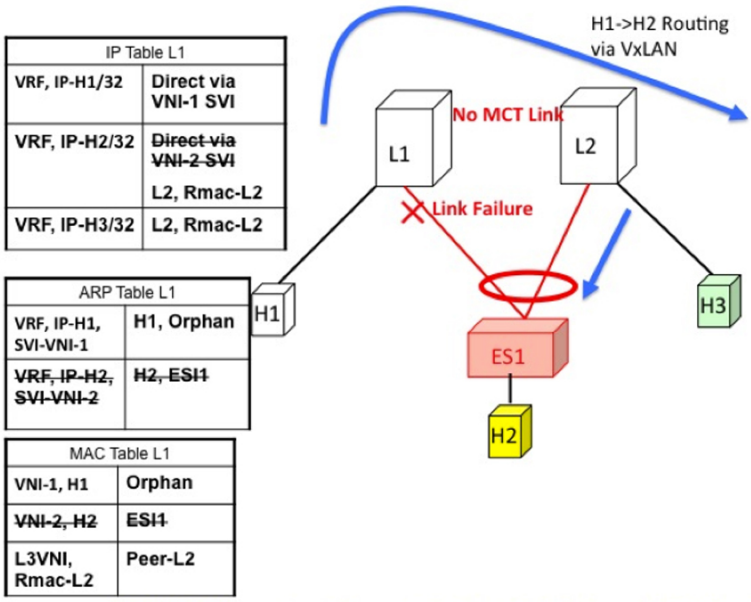
Figure 4: L1 is Distributed Anycast Gateway. H1, H2, and H3 are in different VLANs. H1->H3 routing happens via VXLAN tunnel encapsulation. In vPC, H3 ARP would have been synced via Peer Link and direct routing.



Access Failure for Locally Routed Traffic

In case the ESI link at switch L1 fails, there is no path for the routed traffic to reach from H1 to H2 except via the overlay. Therefore, the local routed traffic takes the sub-optimal path, similar to the H1 to H3 orphan flow.

Figure 5: H1, H2, and H3 are in different VLANs. ESI fails on L1. H1->H2 routing happens via VXLAN tunnel encapsulation.

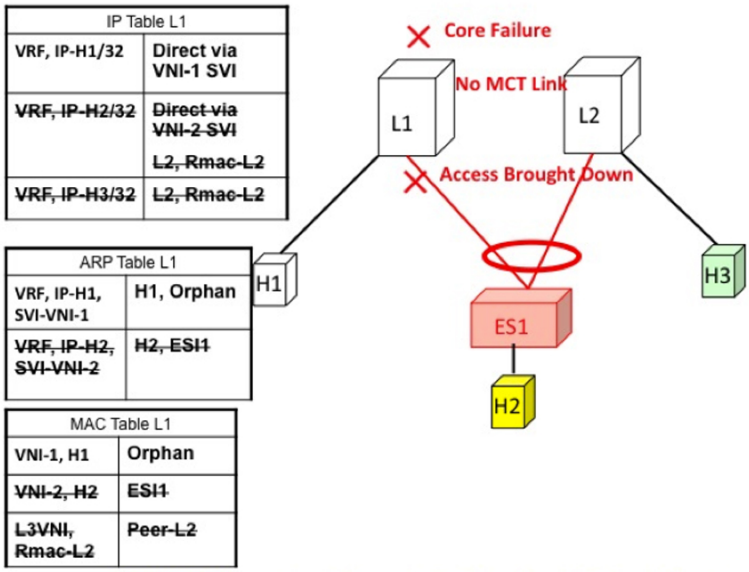


Core Failure for Locally Routed Traffic

If switch L1 gets isolated from the core, it must not continue to attract access traffic, as it will not be able to encapsulate and send it on the overlay. It means that the access links must be brought down at L1 if L1 loses core reachability.

In this scenario, orphan H1 loses all connectivity to both remote and locally attached hosts as there is no dedicated Peer Link.

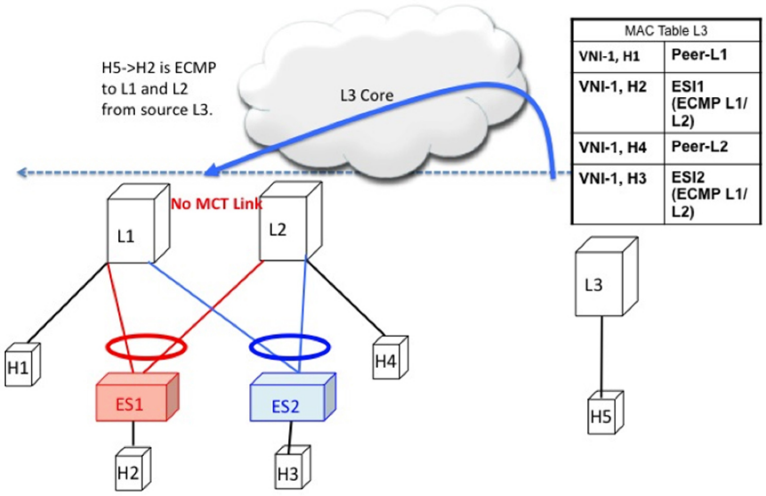
Figure 6: H1, H2, and H3 are in different VLANs. Core fails on L1. Access is brought down. H1 loses all connectivity.



EVPN Multihoming Remote Traffic Flows

Consider a remote switch L3 that sends bridged and routed traffic to the multihomed complex comprising of switches L1 and L2. As there is no virtual or emulated IP representing this MH complex, L3 must do ECMP at the source for both bridged and routed traffic. This section describes how the ECMP is achieved at switch L3 for both bridged and routed cases and how the system interacts with core and access failures.

Figure 7: Layer 2 VXLAN Gateway. L3 performs MAC ECMP to L1/L2.



Remote Bridged Traffic

Consider a remote host H5 that wants to bridge traffic to host H2 that is positioned behind the EVPN MH Complex (L1, L2). Host H2 builds an ECMP list in accordance to the rules defined in RFC 7432. The MAC table at switch L3displays that the MAC entry for H2 points to an ECMP PathList comprising of IP-L1 and IP-L2. Any bridged traffic going from H5 to H2 is VXLAN encapsulated and load balanced to switches L1 and L2. When making the ECMP list, the following constructs need to be kept in mind:

- Mass Withdrawal: Failures causing PathList correction should be independent of the scale of MACs.
- Aliasing: PathList Insertions may be independent of the scale of MACs (based on support of optional routes).

Below are the main constructs needed to create this MAC ECMP PathList:

Ethernet Auto Discovery Route (Type 1) per ES

EVPN defines a mechanism to efficiently and quickly signal the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet Segment. Having each PE advertise a set of one or more Ethernet A-D per ES route for each locally attached Ethernet Segment does this.

Ethernet Auto Discovery Route (Route Type 1) per ES		
NLRI	Route Type	Ethernet Segment (Type 1)
	Route Distinguisher	Router-ID: Segment-ID (VNID << 8)
	ESI	<Type: 1B><MAC: 6B><LD: 3B>
	Ethernet Tag	MAX-ET
	MPLS Label	0
ATTRS	ESI Label Extended Community ESI Label = 0	Single Active = False
	Next-Hop	NVE Loopback IP
	Route Target	Subset of List of RTs of MAC-VRFs associated to all the EVIs active on the ES

MAC-IP Route (Type 2)

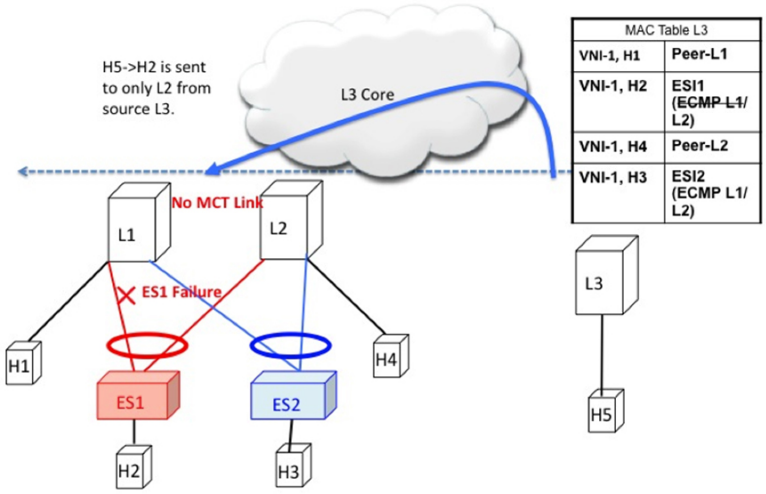
MAC-IP Route remains the same as used in the current vPC multihoming and NX-OS single-homing solutions. However, now it has a non-zero ESI field that indicates that this is a multihomed host and it is a candidate for ECMP Path Resolution.

MAC IP Route (Route Type 2)		
NLRI	Route Type	MAC IP Route (Type 2)
	Route Distinguisher	RD of MAC-VRF associated to the Host
	ESI	<Type : 1B><MAC : 6B><LD : 3B>
	Ethernet Tag	MAX-ET
	MAC Addr	MAC Address of the Host
	IP Addr	IP Address of the Host
	Labels	L2VNI associated to the MAC-VRF L3VNI associated to the L3-VRF
ATTRS	Next-Hop	Loopback of NVE
	RT Export	RT configured under MAC-VRF (AND/OR) L3-VRF associated to the host

Access Failure for Remote Bridged Traffic

In the condition of a failure of ESI links, it results in mass withdrawal. The EAD/ES route is withdrawn leading the remote device to remote the switch from the ECMP list for the given ES.

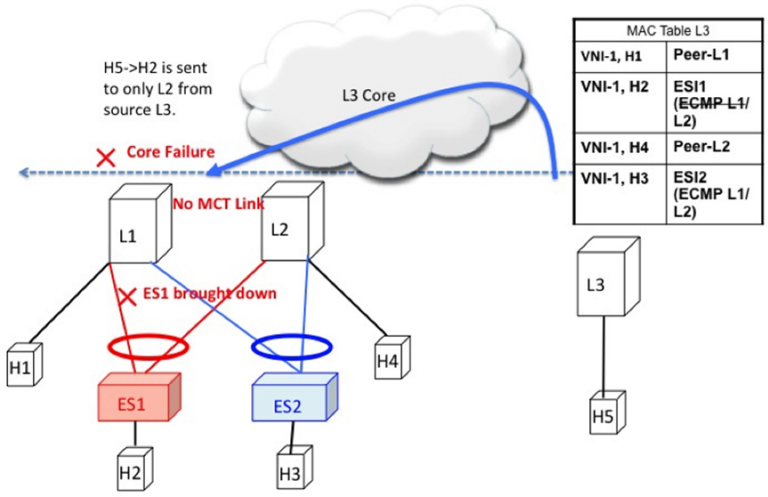
Figure 8: Layer 2 VXLAN Gateway. ESI failure on L1. L3 withdraws L1 from MAC ECMP list. This will happen due to EAD/ES mass withdrawal from L1.



Core Failure for Remote Bridged Traffic

If switch L1 gets isolated from the core, it must not continue to attract access traffic, as it is not able to encapsulate and send it on the overlay. It means that the access links must be brought down at L1 if L1 loses core reachability.

Figure 9: Layer 2 VXLAN Gateway. Core failure at L1. L3 withdraws L1 from MAC ECMP list. This will happen due to route reachability to L1 going away at L3.

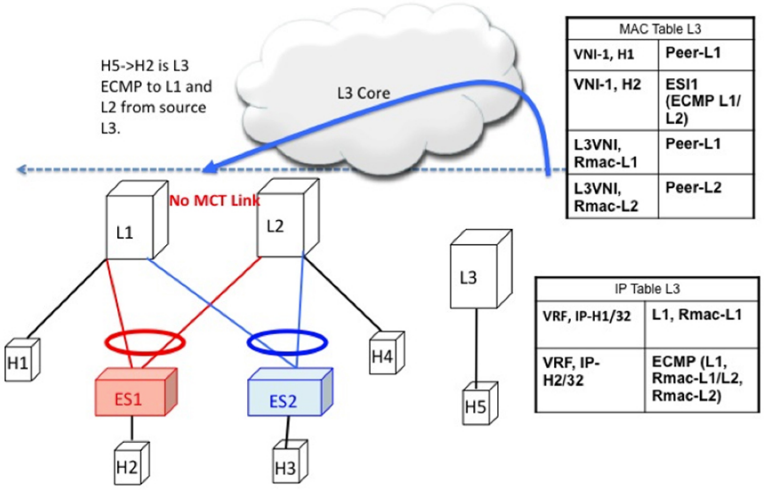


Remote Routed Traffic

Consider L3 being a Layer 3 VXLAN Gateway and H5 and H2 belonging to different subnets. In that case, any inter-subnet traffic going from L3 to L1/L2 is routed at L3, that is a distributed anycast gateway. Both

L1 and L2 advertise the MAC-IP route for Host H2. Due to the receipt of these routes, L3 builds an L3 ECMP list comprising of L1 and L2.

Figure 10: Layer 3 VXLAN Gateway. L3 does IP ECMP to L1/L2 for inter subnet traffic.



Access Failure for Remote Routed Traffic

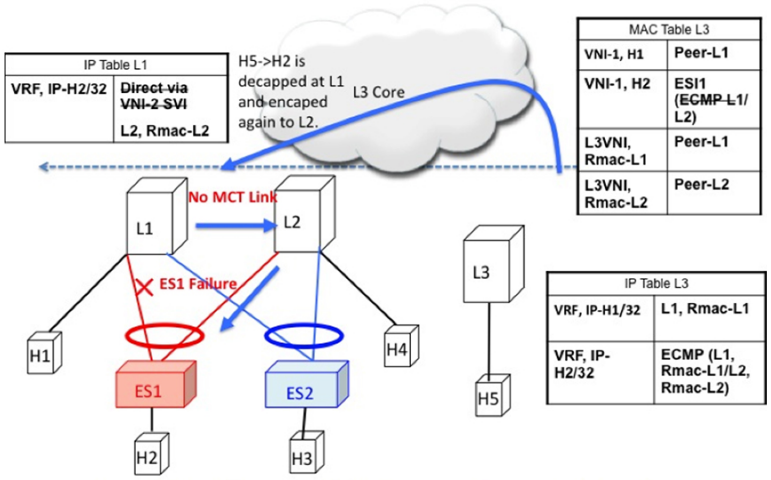
If the access link pointing to ES1 goes down on L1, the mass withdrawal route is sent in the form of EAD/ES and that causes L3 to remove L1 from the MAC ECMP PathList, leading the intra-subnet (L2) traffic to converge quickly. L1 now treats H2 as a remote route reachable via VxLAN Overlay as it is no longer directly connected through the ESI link. This causes the traffic destined to H2 to take the suboptimal path L3->L1->L2.

Inter-Subnet traffic H5->H2 will follow the following path:

- Packet are sent by H5 to gateway at L3.
- L3 performs symmetric IRB and routes the packet to L1 via VXLAN overlay.
- L1 decaps the packet and performs inner IP lookup for H2.
- H2 is a remote route. Therefore, L1 routes the packet to L2 via VXLAN overlay.
- L2 decaps the packet and performs an IP lookup and routes it to directly attached SVI.

Hence the routing happens 3 times, once each at L3, L1, and L2. This sub-optimal behavior continues until Type-2 route is withdrawn by L1 by BGP.

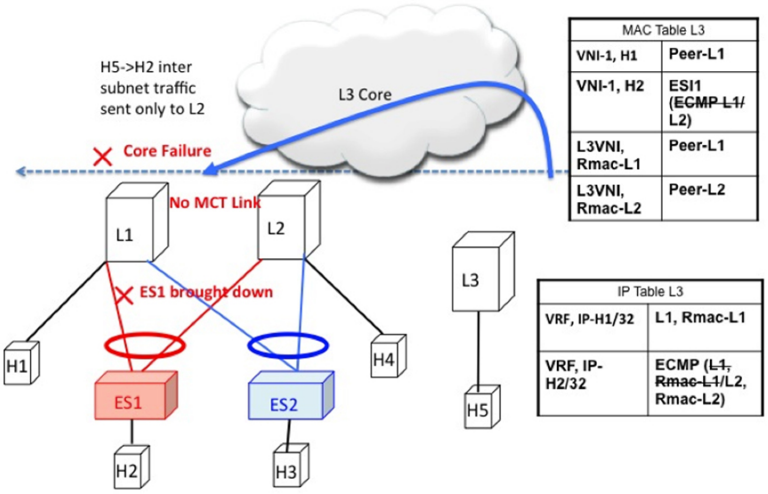
Figure 11: Layer 3 VXLAN Gateway. ESI failure causes ES mass withdrawal that only impacts L2 ECMP. L3 ECMP continues until Type2 is withdrawn. L3 traffic reaches H2 via suboptimal path L3->L1->L2 until then.



Core Failure for Remote Routed Traffic

Core Failure for Remote Routed Traffic behaves the same as core failure for remote bridged traffic. As the underlay routing protocol withdraws L1's loopback reachability from all remote switches, L1 is removed from both MAC ECMP and IP ECMP lists everywhere.

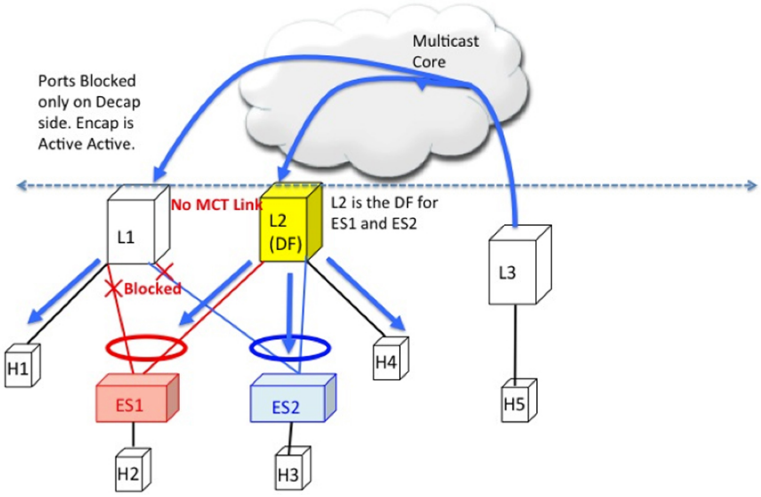
Figure 12: Layer 3 VXLAN Gateway. Core failure. All L3 ECMP paths to L1 are withdrawn at L3 due to route reachability going away.



EVPN Multihoming BUM Flows

NX-OS supports multicast core in the underlay with ESI. Consider BUM traffic originating from H5. The BUM packets are encapsulated in the multicast group mapped to the VNI. Because both L1 and L2 have joined the shared tree (*, G) for the underlay group based on the L2VNI mapping, both receive a copy of the BUM traffic.

Figure 13: BUM traffic originating at L3. L2 is the DF for ES1 and ES2. L2 decapsulates and forwards to ES1, ES2 and orphan. L1 decapsulates and only forwards to orphan.



Designated Forwarder

It is important that only one of the switches in the redundancy group decaps and forwards BUM traffic over the ESI links. For this purpose, a unique Designated Forwarder (DF) is elected on a per Ethernet Segment basis. The role of the DF is to decap and forward BUM traffic originating from the remote segments to the destination local segment for which the device is the DF. The main aspects of DF election are:

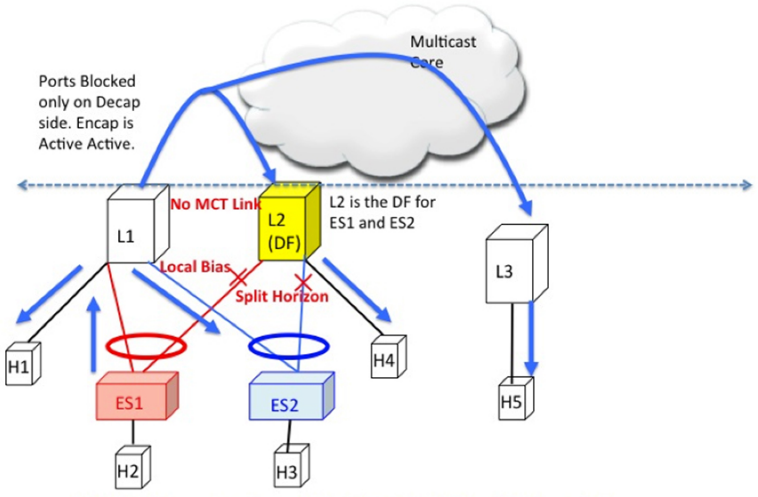
- DF Election is per (ES, VLAN) basis. There can be a different DF for ES1 and ES2 for a given VLAN.
- DF election result only applies to BUM traffic on the RX side for decap.
- Every switch must decap BUM traffic to forward it to singly homed or orphan links.
- Duplication of DF role leads to duplicate packets or loops in a DHN. Therefore, there must be a unique DF on per (ES, VLAN) basis.

Split Horizon and Local Bias

Consider BUM traffic originating from H2. Consider that this traffic is hashed at L1. L1 encapsulates this traffic in Overlay Multicast Group and sends the packet out to the core. All switches that have joined this multicast group with same L2VNI receive this packet. Additionally, L1 also locally replicates the BUM packet on all directly connected orphan and ESI ports. For example, if the BUM packet originated from ES1, L1 locally replicates it to ES2 and the orphan ports. This technique to replicate to all the locally attached links is termed as local-bias.

Remote switches decap and forward it to their ESI and orphan links based on the DF state. However, this packet is also received at L2 that belongs to the same redundancy group as the originating switch L1. L2 must decap the packet to send it to orphan ports. However, even through L2 is the DF for ES1, L2 must not forward this packet to ES1 link. This packet was received from a peer that shares ES1 with L1 as L1 would have done local-bias and duplicate copies should not be received on ES2. Therefore L2 (DF) applies a split-horizon filter for L1-IP on ES1 and ES2 that it shares with L1. This filter is applied in the context of a VLAN.

Figure 14: BUM traffic originating at L1. L2 is the DF for ES1 and ES2. However, L2 must perform split horizon check here as it shares ES1 and ES2 with L1. L2 however



Ethernet Segment Route (Type 4)

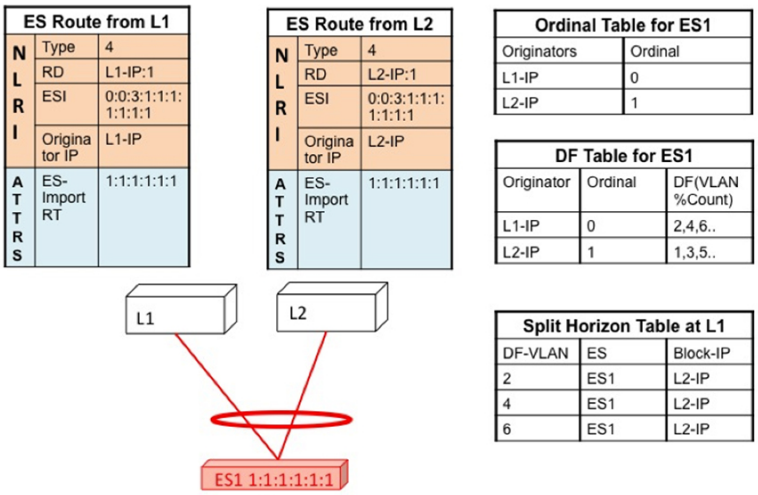
The Ethernet Segment Route is used to elect the Designated Forwarder and to apply Split Horizon Filtering. All the switches that are configured with an Ethernet Segment originate from this route. Ethernet Segment Route is exported and imported when ESI is locally configured under the PC.

Ethernet Segment Route (Route Type 4)		
NLRI	Route Type	Ethernet Segment (Type 4)
	RD	Router-ID: Base + Port Channel Number
	ESI	<Type : 1B><MAC : 6B><LD : 3B>
	Originator IP	NVE loopback IP
ATTRS	ES-Import RT	6 Byte MAC derived from ESI

DF Election and VLAN Carving

Upon configuration of the ESI, both L1 and L2 advertises the ES route. The ESI MAC is common between L1 and L2 and unique in the network. Therefore, only L1 and L2 import each other's ES routes.

Figure 15: If VLAN % count equals to ordinal, take up DF role.



Core and Site Failures for BUM Traffic

If the access link pertaining to ES1 fails at L1, L1 withdraws the ES route for ES1. This leads to a change triggering re-compute the DF. Since L2 is the only TOR left in the Ordinal Table, it takes over DF role for all VLANs.

BGP EVPN multihoming on Cisco Nexus 9000 Series switches provides minimum operational and cabling expenditure, provisioning simplicity, flow based load balancing, multi pathing, and fail-safe redundancy.

Configuring VLAN Consistency Checking

Overview of VLAN Consistency Checking

In a typical multihoming deployment scenario, host 1 belonging to VLAN X sends traffic to the access switch and then the access switch sends the traffic to both the uplinks towards VTEP1 and VTEP2. The access switch does not have the information about VLAN X configuration on VTEP1 and VTEP2. VLAN X configuration mismatch on VTEP1 or VTEP2 results in a partial traffic loss for host 1. VLAN consistency checking helps to detect such configuration mismatch.

For VLAN consistency checking, CFSoIP is used. Cisco Fabric Services (CFS) provides a common infrastructure to exchange the data across the switches in the same network. CFS has the ability to discover CFS capable switches in the network and to discover the feature capabilities in all the CFS capable switches. You can use CFS over IP (CFSoIP) to distribute and synchronize a configuration on one Cisco device or with all other Cisco devices in your network.

CFSoIP uses multicast to discover all the peers in the management IP network. For EVPN multihoming VLAN consistency checking, it is recommended to override the default CFS multicast address with the **cfs ipv4 mcast-address** <*mcast address*> CLI command. To enable CFSoIP, the **cfs ipv4 distribute** CLI command should be used.

When a trigger (for example, device booting up, VLAN configuration change, VLANs administrative state change on the ethernet-segment port-channel) is issued on one of the multihoming peers, a broadcast request

with a snapshot of configured and administratively up VLANs for the ethernet-segment (ES) is sent to all the CFS peers.

When a broadcast request is received, all CFS peers sharing the same ES as the requestor respond with their VLAN list (configured and administratively up VLAN list per ES). The VLAN consistency checking is run upon receiving a broadcast request or a response.

A 15 seconds timer is kicked off before sending a broadcast request. On receiving the broadcast request or response, the local VLAN list is compared with that of the ES peer. The VLANs that do not match are suspended. Newly matched VLANs are no longer suspended.

VLAN consistency checking runs for the following events:

- Global VLAN configuration: Add, delete, shut, or no shut events.

  Port channel VLAN configuration: Trunk allowed VLANs added or removed or access VLAN changed.
- CFS events: CFS peer added or deleted or CFSoIP configuration is removed.
- ES Peer Events: ES peer added or deleted.

The broadcast request is retransmitted if a response is not received. VLAN consistency checking fails to run if a response is not received after 3 retransmissions.

VLAN Consistency Checking Guidelines and Limitations

See the following guidelines and limitations for VLAN consistency checking:

- The VLAN consistency checking uses CFSoIP. Out-of-band access through a management interface is mandatory on all multihoming switches in the network.
- It is recommended to override the default CFS multicast address with the CLI **cfs ipv4 mcast-address** <*mcast address*> command.
- The VLAN consistency check cannot detect a mismatch in **switchport trunk native vlan** configuration.
- CFSoIP and CFSoE should not be used in the same device.
- CFSoIP should not be used in devices that are not used for VLAN consistency checking.
- If CFSoIP is required in devices that do not participate in VLAN consistency checking, a different multicast group should be configured for devices that participate in VLAN consistency with the CLI **cfs ipv4 mcast-address** <*mcast address*> command.

Configuring VLAN Consistency Checking

Use the **cfs ipv4 mcast-address** <*mcast address*> CLI command to override the default CFS multicast address. Use the **cfs ipv4 distribute** CLI command to enable CFSoIP.

To enable or disable the VLAN consistency checking, use the new **vlan-consistency-check** CLI command that has been added under the **evpn esi multihoming** mode.

```
switch (config)# sh running-config | in cfs
cfs ipv4 mcast-address 239.255.200.200
cfs ipv4 distribute
```

```
switch# sh run | i vlan-consistency
evpn esi multihoming
    vlan-consistency-check
```

Displaying Show command Output for VLAN Consistency Checking

See the following show commands output for VLAN consistency checking.

To list the CFS peers, use the **sh cfs peers name nve** CLI command.

```
switch# sh cfs peers name nve

Scope      : Physical-ip
--------------------------------------------------------------------------
 Switch WWN              IP Address
--------------------------------------------------------------------------
 20:00:f8:c2:88:23:19:47 172.31.202.228                          [Local]
                         Switch
 20:00:f8:c2:88:90:c6:21 172.31.201.172                          [Not Merged]
 20:00:f8:c2:88:23:22:8f 172.31.203.38                           [Not Merged]
 20:00:f8:c2:88:23:1d:e1 172.31.150.132                          [Not Merged]
 20:00:f8:c2:88:23:1b:37 172.31.202.233                          [Not Merged]
 20:00:f8:c2:88:23:05:1d 172.31.150.134                          [Not Merged]
```

The **show nve ethernet-segment** command now displays the following details:

- The list of VLANs for which consistency check is failed.
- Remaining value (in seconds) of the global VLAN CC timer.

```
switch# sh nve ethernet-segment
ESI Database
----------------------------------------
ESI: 03aa.aaaa.aaaa.aa00.0001,
   Parent interface: port-channel2,
  ES State: Up
  Port-channel state: Up
  NVE Interface: nve1
   NVE State: Up
   Host Learning Mode: control-plane
  Active Vlans: 3001-3002
   DF Vlans: 3002
   Active VNIs: 30001-30002
  CC failed VLANs: 0-3000,3003-4095
  CC timer status: 10 seconds left
  Number of ES  members: 2
  My ordinal: 0
  DF timer start time: 00:00:00
  Config State: config-applied
  DF List: 201.1.1.1 202.1.1.1
  ES route added to L2RIB: True
  EAD routes added to L2RIB: True
```

See the following Syslog output:

```
switch(config)# 2017 Jan ?7 19:44:35 Switch %ETHPORT-3-IF_ERROR_VLANS_SUSPENDED: VLANs
2999-3000 on Interface port-channel40 are being suspended.
(Reason: SUCCESS)

After Fixing configuration
2017 Jan ?7 19:50:55 Switch %ETHPORT-3-IF_ERROR_VLANS_REMOVED: VLANs 2999-3000 on Interface
 port-channel40 are removed from suspended state.
```

Configuring ESI ARP Suppression

Overview of ESI ARP Suppression

Ethernet Segment Identifier (ESI) ARP suppression is an extension of the ARP suppression solution in VXLAN EVPN. It optimizes the ESI multihoming feature by significantly decreasing ARP broadcasts in the data center.

The host normally floods the VLAN with ARP requests. You can minimize this flooding by maintaining an ARP cache locally on the leaf switch. The ARP cache is built by:

- Snooping all ARP packets and populating the ARP cache with the source IP address and MAC bindings from the request
- Learning IP host or MAC address information through BGP EVPN IP or MAC route advertisements

With ESI ARP suppression, the initial ARP requests are broadcast to all sites. However, subsequent ARP requests are suppressed at the first-hop leaf switch and answered locally if possible. In this way, ESI ARP suppression significantly reduces ARP traffic across the overlay. If the cache lookup fails and the response cannot be generated locally, the ARP request can be flooded, which helps with the detection of silent hosts.

ESI ARP suppression is a per-VNI (L2 VNI) feature and is supported only with VXLAN EVPN (distributed gateway). This feature is supported in L2 (no SVI) and L3 modes for releases prior to Cisco NX-OS Release 7.0(3)I5(2) but only in L3 mode for later releases.

Limitations for ESI ARP Suppression

See the following limitations for ESI ARP suppression:

- ESI multihoming solution is supported only on Cisco Nexus 9300 Series switches at the leafs.
- ESI ARP suppression is only supported in L3 [SVI] mode.
- ESI ARP suppression cache limit is 64K that includes both local and remote entries.

Configuring ESI ARP Suppression

For ARP suppression VACLs to work, configure the TCAM carving using the **hardware access-list tcam region arp-ether 256** CLI command.

```
Interface nve1
   no shutdown
   source-interface loopback1
   host-reachability protocol bgp
```

```
  member vni 10000
      suppress-arp
  mcast-group 224.1.1.10
```

Displaying Show Commands for ESI ARP Suppression

See the following Show commands output for ESI ARP suppression:

```
switch# show ip arp suppression-cache ?
 detail        Show details
   local          Show local entries
   remote     Show remote entries
   statistics  Show statistics
   summary Show summary
   vlan           L2vlan

switch# show ip arp suppression-cache local

Flags: + - Adjacencies synced via CFSoE
       L - Local Adjacency
       R - Remote Adjacency
       L2 - Learnt over L2 interface
      PS - Added via L2RIB, Peer Sync
      RO - Dervied from L2RIB Peer Sync Entry

Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags              Remote
 Vtep Addrs

61.1.1.20       00:07:54 0000.0610.0020  610 port-channel20     L
61.1.1.30       00:07:54 0000.0610.0030  610 port-channel2       L[PS RO]
61.1.1.10       00:07:54 0000.0610.0010  610 Ethernet1/96        L


switch# show ip arp suppression-cache remote
 Flags: + - Adjacencies synced via CFSoE
        L - Local Adjacency
        R - Remote Adjacency
        L2 - Learnt over L2 interface
          PS - Added via L2RIB, Peer Sync
          RO - Dervied from L2RIB Peer Sync Entry
          Ip Address      Age      Mac Address         Vlan    Physical-ifindex    Flags
    Remote Vtep Addrs
 61.1.1.40       00:48:37 0000.0610.0040   610     (null)                      R
 VTEP1, VTEP2.. VTEPn


switch# show ip arp suppression-cache detail
       Flags: + - Adjacencies synced via CFSoE
       L - Local Adjacency
       R - Remote Adjacency
       L2 - Learnt over L2 interface
       PS - Added via L2RIB, Peer Sync
       RO - Derived from L2RIB Peer Sync Entry

             Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags
     Remote Vtep Addrs
 61.1.1.20       00:00:07 0000.0610.0020  610 port-channel20     L
 61.1.1.30       00:00:07 0000.0610.0030  610 port-channel2       L[PS RO]
 61.1.1.10       00:00:07 0000.0610.0010  610 Ethernet1/96        L
 61.1.1.40       00:00:07 0000.0610.0040  610 (null)                    R
 VTEP1, VTEP2.. VTEPn
```

```
switch# show ip arp suppression-cache summary
 IP ARP suppression-cache Summary
 Remote          :1
 Local              :3
 Total               :4
switch# show ip arp suppression-cache statistics
 ARP packet statistics for suppression-cache
 Suppressed:
 Total 0, Requests 0, Requests on L2 0, Gratuitous 0, Gratuitous on L2 0
 Forwarded :
 Total: 364
  L3 mode :      Requests 364, Replies 0
  Request on core port 364, Reply on core port 0
                 Dropped 0
  L2 mode :      Requests 0, Replies 0
                 Request on core port 0, Reply on core port 0
                 Dropped 0
 Received:
 Total: 3016
 L3 mode:       Requests 376, Replies 2640
  Local Request 12, Local Responses 2640
                 Gratuitous 0, Dropped 0
  L2 mode :      Requests 0, Replies 0
                 Gratuitous 0, Dropped 0
```

```
switch# sh ip arp multihoming-statistics vrf all
ARP Multihoming statistics for all contexts
Route Stats
============
 Receieved ADD from L2RIB        :1756 | 1756:Processed ADD from L2RIB Receieved DEL from
L2RIB        :88 | 87:Processed DEL from L2RIB Receieved PC shut from L2RIB    :0 |
1755:Processed PC shut from L2RIB Receieved remote UPD from L2RIB :5004 | 0:Processed remote
 UPD from L2RIB
ERRORS
=======
Multihoming ADD error invalid flag         :0
Multihoming DEL error invalid flag         :0
Multihoming ADD error invalid current state:0
Multihoming DEL error invalid current state:0
Peer sync DEL error MAC mismatch           :0
Peer sync DEL error second delete          :0
Peer sync DEL error deleteing TL route     :0
True local DEL error deleteing PS RO route :0

switch#
```