About Cisco Floating L3Outs

About Cisco Floating L3Outs

Beginning with the Cisco Application Centric Infrastructure (ACI) release 4.2(1), you no longer need to specify multiple Layer 3 outside network connection (L3Out) logical interface paths to connect external network devices.

The floating L3Out feature enables you to configure an L3Out without specifying any L3Out interface on the local leaf. The feature saves you from having to configure multiple L3Out logical interfaces to maintain routing when virtual machines (performing a specific virtual network function) move from one host to another. Floating L3Out has been supported since Cisco ACI release 4.2(1) for VMM domains with VMware vSphere Distributed Switch (VDS).

Beginning with the Cisco ACI release 5.0(1), physical domains are also supported. This means that the same simplified configuration can be used for physical routers deployments as well or for virtual routers that are not part of a VMM domain.

Configuring L3Outs for Virtual Environments

When you connect to an external virtual router, you must configure the L3Out logical interface path from the border leaf switch to the uplink of the hypervisor where a virtual device resides. However, when hypervisor resources are aggregated into clusters, there is no guarantee that the virtual machine for the virtual function always runs on the same host.

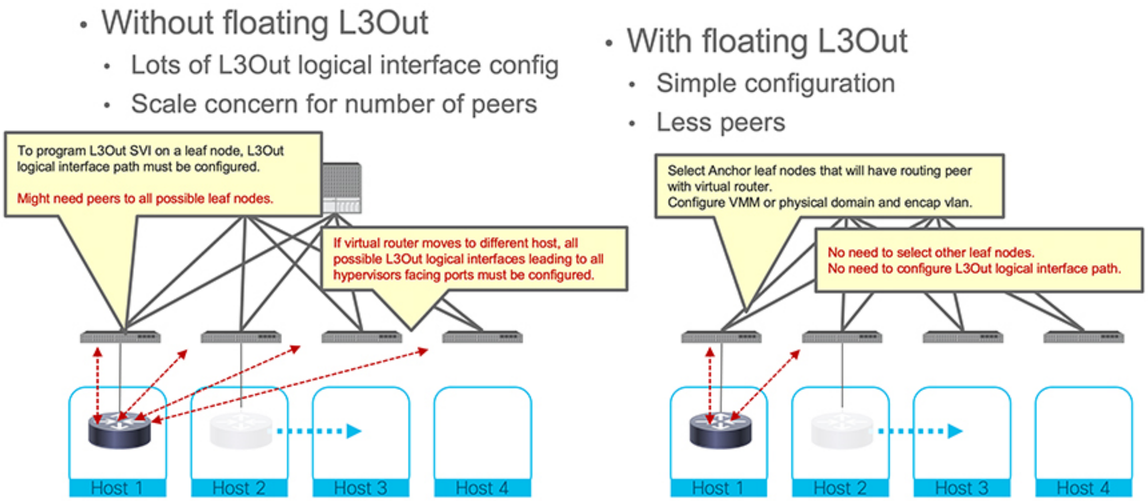
Before Cisco Application Policy Infrastructure Controller (APIC) release 4.2(1), to maintain the routing function when virtual machines moved, you had to configure all possible L3Out logical interfaces from the border leaf switches to all the hypervisors that could host the virtual machines. This extra configuration was required because L3Out switched virtual interface (SVI) and VLAN programming were not done automatically.

For example, if you had a hypervisor cluster stretched across 12 leaf switches, virtual machines could potentially move to every one of those 12 leaf switches. That meant that you had to create a policy to deploy an L3Out from every leaf node interface to every corresponding server.

However, when you configure floating L3Out, you simplify the entire process. After you configure the floating L3Out, you no longer need to configure each L3Out logical interface on all the leaf switches where the hypervisor cluster is connected.

Another benefit of floating L3Out is that only specific leaf nodes (called anchor leaf nodes in this document) will establish routing adjacencies with the external routers. This approach is beneficial for the peering scale of both Cisco ACI leaf switches and the external network devices.

Figure 1: Benefits of Floating L3Out Deployment


• Without floating L3Out
• Lots of L3Out logical interface config
• Scale concern for number of peers
To program L3Out SVI on a leaf node, L3Out logical interface path must be configured.
Might need peers to all possible leaf nodes.
If virtual router moves to different host, all possible L3Out logical interfaces leading to all hypervisors facing ports must be configured.
Host 1
Host 2
Host 3
Host 4
• With floating L3Out
• Simple configuration
• Less peers
Select Anchor leaf nodes that will have routing peer with virtual router.
Configure VMM or physical domain and encap vlan.
No need to select other leaf nodes.
No need to configure L3Out logical interface path.
Host 1
Host 2
Host 3
Host 4

Configuring L3Outs for Physical Domains

Beginning with the Cisco Application Policy Infrastructure Controller (APIC) release 5.0(1), support for the floating L3Out functionality is also extended to physical domains. This enhancement enables you to use floating L3Out feature with virtual routers without VMM domain integration or to connect physical routers without L3Out logical interface path configurations.

Scenarios That Benefit from Floating L3Out

The following list provides examples of scenarios where the floating Layer 3 outside network connection (L3Out) is useful. The configuration for floating L3Out is the same for each scenario. These use cases provide additional examples to those discussed in Configuring L3Outs for Virtual Environments, on page 1.

- Physical domain: Even though a physical router does not move between different leaf nodes, using floating L3Out enables you to simplify the configuration because an L3Out logical interface configuration is not required on all the leaf switches where the external physical routers may be connected.
- Virtual firewall or virtual router that is hosted in a hypervisor cluster: Resource scheduling and allocation can be dynamically managed—for example, with the use of VMware Distributed Resource Scheduler

(DRS). The virtual machine (VM) hosting boundary therefore becomes the cluster itself, not a single host.

- Virtual firewall or virtual router with high-availability (HA): The hypervisor HA mechanism allows to restart the failed VM on any available host within the hypervisor cluster. (For example, VMware HA is a specific example of such capability. It is worth noting that this HA capability is in addition to the firewall native redundancy deployment model, such as active/active or active/standby.)
- ECMP load balancing to multiple routers: Using floating L3Out enables you to connect multiple routers to different leaf switches without requiring an L3Out logical interface configuration on each switch.
- Maintenance mode: When a hypervisor must be upgraded, VM administrators evacuate the host. That is, they perform live migration of the firewall or router VM to another host in the hypervisor cluster.
- Disaster Avoidance: In a stretched cluster, outage is expected on some nodes. The VMs, such as virtual routers, virtual firewalls or any other virtual devices, can be migrated to different hosts that are not expected to experience the outage.

Floating L3Out Topology and Terminology

This section describes an example topology for using the floating Layer 3 outside network connection (L3Out) feature. The example uses a VMM domain and a virtual port channel (vPC) deployment, but physical domains are also supported and the use of vPC is not mandatory.

- **Virtual Routers**: Virtual routers can be a router, a virtual firewall, or any other virtual device that is used as a next-hop of a static route on the Cisco Application Centric Infrastructure (ACI) fabric or establishes a routing adjacency with the ACI fabric.
- **Anchor Leaf Nodes**: In this example, there are two leaf switches acting as anchor leaf nodes (Leaf1 and Leaf2) and establishing Layer 3 adjacencies with the external routers. As of Cisco ACI release 6.0(1), the verified scalability number of anchor leaf nodes is 6 per L3Out.

  Anchor leaf nodes make use of the primary IP addresses and floating IP addresses. They can also have secondary IP and floating secondary IP addresses if needed (the purpose of all those IP addresses will be clarified later in this section of the document).
- **Non-anchor Leaf Nodes**: In this example, there are two leaf switches acting as the non-anchor leaf nodes (Leaf3 and Leaf4). The non-anchor leaf nodes do not create any adjacency with the external routers. They acts as a “pass-through” for traffic flowing between directly connected external routers and the anchor node. As of ACI release 6.0(1), the verified scalability number of non-anchor leaf nodes is 32 per L3Out.

  Non-anchor leaf nodes use a floating IP address and can have floating secondary IP addresses, if needed (those IP addresses are shared by all the non-anchor leaf nodes). If it is a VMware vDS VMM domain, the floating IP address is deployed only when the virtual router is connected to the leaf node. If it is a physical domain, the floating IP address is deployed if the leaf port uses an AEP that has an L3Out domain associated to the floating L3Out. The floating IP address is the common IP address for non-anchor leaf nodes.

Configuring an L3Out creates an L3Out bridge domain on the anchor node switches. This L3Out bridge domain is usually referred to as the “L3Out’s SVIs subnet”. In the case of a floating L3Out with VMM domain, once a virtual router moves to a host connected to a non-anchor switch, Cisco Application Policy Infrastructure Controller (APIC) deploys the L3Out bridge domain on the non-anchor leaf switch as well. It also installs the floating IP address (and floating secondary IP addresses, when needed) on the non-anchor leaf switch. If an external EPG under the L3Out has a contract with another EPG, the routes to the EPG and policy enforcement rules for the contract are also installed on the non-anchor leaf switch. Although the location of the virtual router is changed, it still can maintain the routing adjacencies with the SVI interfaces deployed on the anchor leaf nodes because of the ACI capability of extending connectivity for the L3Out bridge domain across anchor and non-anchor leaf nodes.

Figure 2: Example of a Floating L3Out Topology (external routers are connected to a pair of anchor leaf nodes)

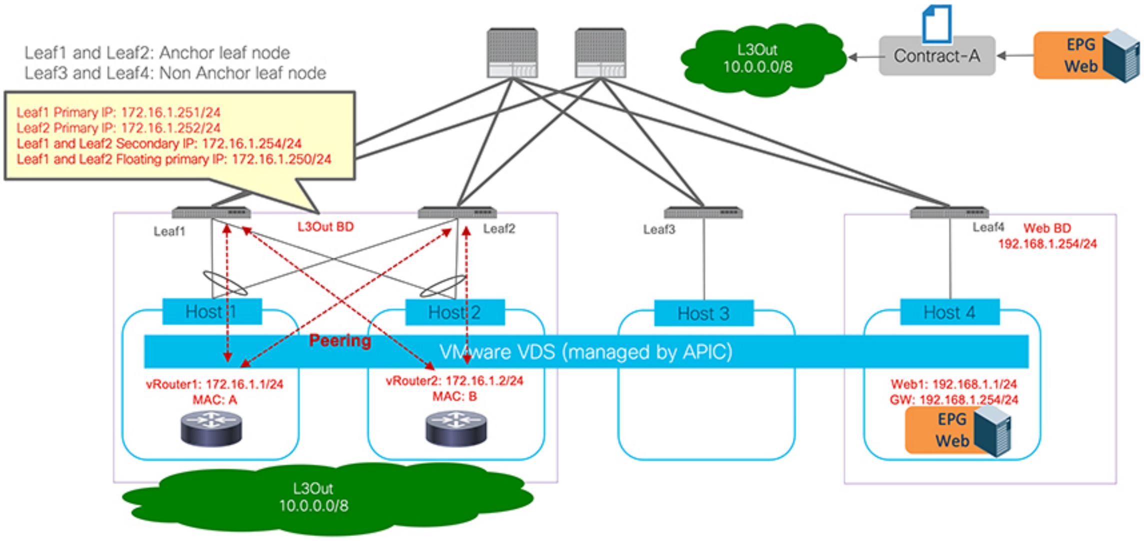
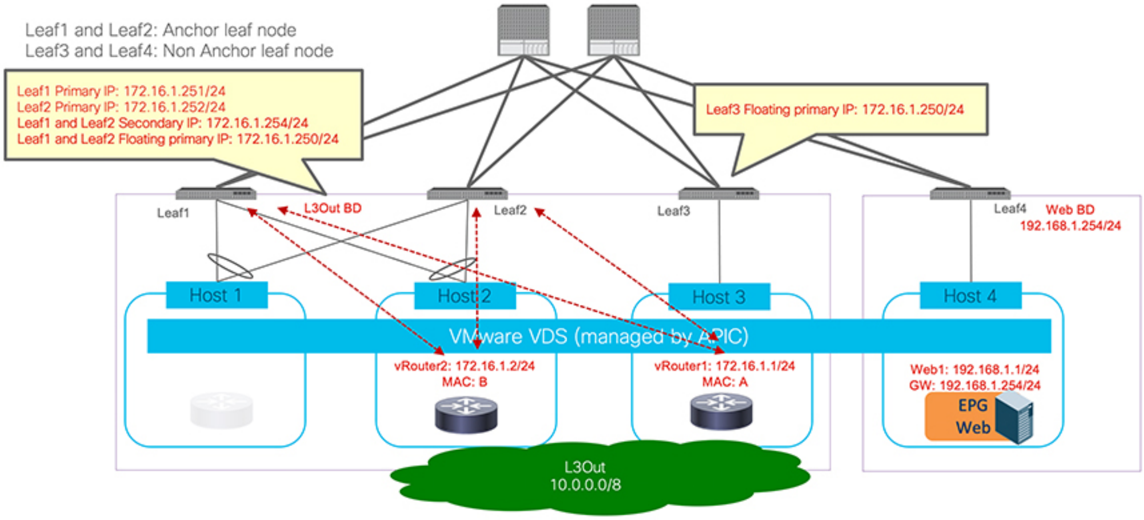

Leaf1 and Leaf2: Anchor leaf node
Leaf3 and Leaf4: Non Anchor leaf node
Leaf1 Primary IP: 172.16.1.251/24
Leaf2 Primary IP: 172.16.1.252/24
Leaf1 and Leaf2 Secondary IP: 172.16.1.254/24
Leaf1 and Leaf2 Floating primary IP: 172.16.1.250/24
L3Out
10.0.0.0/8
Contract-A
EPG
Web
Leaf1
L3Out BD
Leaf2
Leaf3
Leaf4
Web BD
192.168.1.254/24
Host 1
Host 2
Host 3
Host 4
Peering
VMware VDS (managed by APIC)
vRouter1: 172.16.1.1/24
MAC: A
vRouter2: 172.16.1.2/24
MAC: B
Web1: 192.168.1.1/24
GW: 192.168.1.254/24
EPG
Web
L3Out
10.0.0.0/8

Figure 3: Example of a Floating L3Out Topology (an external router is moved to a non-anchor leaf node)


Leaf1 and Leaf2: Anchor leaf node
Leaf3 and Leaf4: Non Anchor leaf node
Leaf1 Primary IP: 172.16.1.251/24
Leaf2 Primary IP: 172.16.1.252/24
Leaf1 and Leaf2 Secondary IP: 172.16.1.254/24
Leaf1 and Leaf2 Floating primary IP: 172.16.1.250/24
Leaf3 Floating primary IP: 172.16.1.250/24
Leaf1
L3Out BD
Leaf2
Leaf3
Leaf4
Web BD
192.168.1.254/24
Host 1
Host 2
Host 3
Host 4
VMware VDS (managed by APIC)
vRouter2: 172.16.1.2/24
MAC: B
vRouter1: 172.16.1.1/24
MAC: A
Web1: 192.168.1.1/24
GW: 192.168.1.254/24
EPG
Web
L3Out
10.0.0.0/8

As mentioned above, anchor and non-anchor leaf nodes that are defined as part of the floating L3Out configuration make use of the following IP addresses.

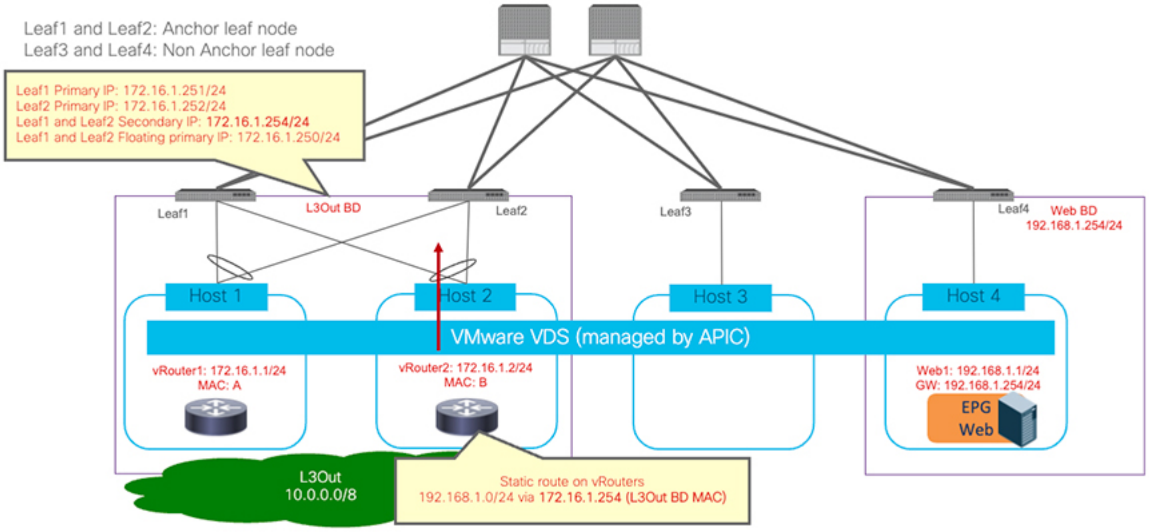
- Primary IP address: the unique IP address assigned to an SVI interface on each leaf node part of the L3Out (it is the real IP of the leaf node). In the case of floating L3Out, the provisioning of an SVI interface with a unique primary IP address is required on each anchor leaf node and it's used for establishing L3Out peering adjacencies with the external routers.

- Secondary IP address (optional): additional IP address assigned to the SVI interface for anchor leaf nodes, which can be used for the use cases below:
    - Common IP address shared by the anchor leaf nodes, acting as a virtual IP and used when external network devices are connected using static routing. The external network devices will set the next-hop gateway of the static routes to this specific IP.
    - Unique IP addresses per anchor leaf node for a secondary IP subnet provisioned in addition to the primary IP subnet.
    - Common IP address shared by the anchor leaf nodes for the secondary IP subnet (to be used for static routing, as described above for the primary IP subnet).
- Floating (primary) IP address: the floating IP is programmed on anchor and non-anchor nodes in order to program a layer-3 interface on the floating SVI. This will program the same MAC address on all anchor and non-anchor switches and allows the switch to directly forward traffic received from external routers into the fabric. It is used for ARP resolution from an anchor leaf node.
- Floating secondary IP address (optional): a common IP provisioned on anchor and non-anchor leaf nodes. Used only if multiple subnets are used in the same external bridge domain (SVI). Floating secondary IP is not supposed to be used for external communication.

The figure below illustrates an example.

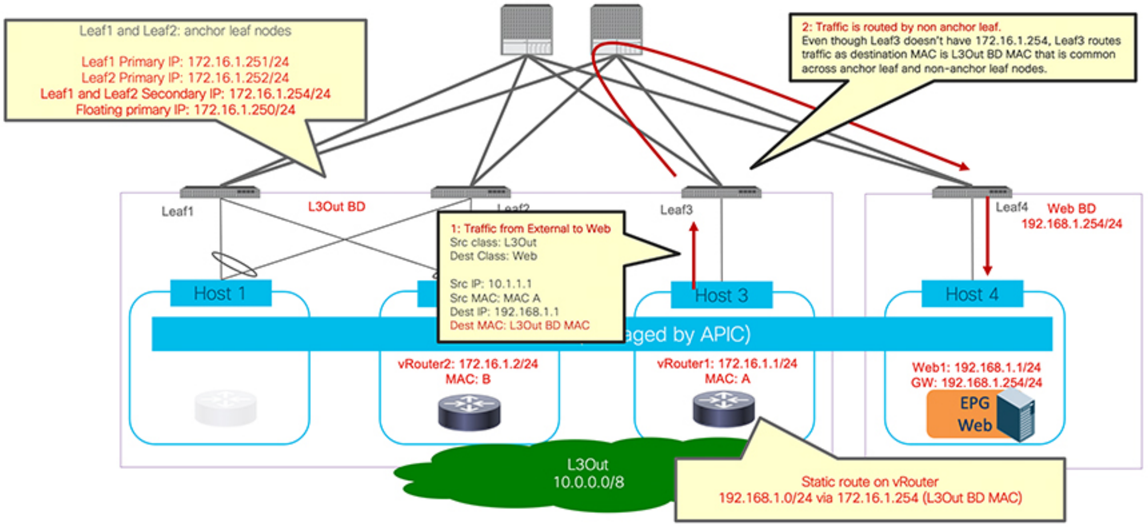
- Primary IP address: 172.16.1.251 is Leaf1 primary IP address and 172.16.1.252 is Leaf2 primary IP address (those are the anchor leaf nodes).
- Secondary IP address (optional): 172.16.1.254 is the secondary IP address of Leaf1 and Leaf2. 172.16.1.254 is used as the next-hop of the static route on the external devices to reach the IP subnet 192.168.1.0/24 deployed inside the ACI fabric.
- Floating (primary) IP address: 172.16.1.250/24 is the floating IP address that is used for ARP resolution.
- If another subnet is required using the same SVI VLAN encapsulation, additional secondary IP addresses and floating secondary IP addresses can be added under the same floating SVI. For example, 172.16.2.254 as the secondary IP address and 172.16.2.250 as the floating secondary IP address.

Figure 4: Example of IP addresses: secondary IP is used as the next-hop of the static route


Leaf1 and Leaf2: Anchor leaf node
Leaf3 and Leaf4: Non Anchor leaf node
Leaf1 Primary IP: 172.16.1.251/24
Leaf2 Primary IP: 172.16.1.252/24
Leaf1 and Leaf2 Secondary IP: 172.16.1.254/24
Leaf1 and Leaf2 Floating primary IP: 172.16.1.250/24
Leaf1
L3Out BD
Leaf2
Leaf3
Leaf4
Web BD
192.168.1.254/24
Host 1
Host 2
Host 3
Host 4
VMware VDS (managed by APIC)
vRouter1: 172.16.1.1/24
MAC: A
vRouter2: 172.16.1.2/24
MAC: B
Web1: 192.168.1.1/24
GW: 192.168.1.254/24
EPG
Web
L3Out
10.0.0.0/8
Static route on vRouters
192.168.1.0/24 via 172.16.1.254 (L3Out BD MAC)

Because non-anchor leaf nodes instantiate the floating (primary) IP with the same MAC address as the next-hop IP (172.16.1.254 in the example above) used in the static route, even if the external router is moved under a non-anchor leaf node, traffic is directly routed by the non-anchor leaf node.

Figure 5: Example of IP addresses: secondary IP is used for the next-hop of the static route (an external router is connected to a non-anchor leaf node)


Leaf1 and Leaf2: anchor leaf nodes
Leaf1 Primary IP: 172.16.1.251/24
Leaf2 Primary IP: 172.16.1.252/24
Leaf1 and Leaf2 Secondary IP: 172.16.1.254/24
Floating primary IP: 172.16.1.250/24
2: Traffic is routed by non anchor leaf.
Even though Leaf3 doesn't have 172.16.1.254, Leaf3 routes traffic as destination MAC is L3Out BD MAC that is common across anchor leaf and non-anchor leaf nodes.
Leaf1
L3Out BD
Leaf2
Leaf3
Leaf4
Web BD
192.168.1.254/24
1: Traffic from External to Web
Src class: L3Out
Dest Class: Web
Src IP: 10.1.1.1
Src MAC: MAC A
Dest IP: 192.168.1.1
Dest MAC: L3Out BD MAC
Host 1
Host 3
Host 4
aged by APIC)
vRouter2: 172.16.1.2/24
MAC: B
vRouter1: 172.16.1.1/24
MAC: A
Web1: 192.168.1.1/24
GW: 192.168.1.254/24
EPG
Web
L3Out
10.0.0.0/8
Static route on vRouter
192.168.1.0/24 via 172.16.1.254 (L3Out BD MAC)

- **Traffic Flow**: Regardless of the use of dynamic peering or static routing between the external routers and anchor leaf node, before any virtual routers move, external to internal (L3Out-to-Web) traffic through an anchor node goes to the spine switch and then to the web endpoint in Host 4. Return traffic (Web-to-L3Out) goes back to the virtual router through an anchor leaf node.

  If a virtual router moves to Host 3 under non-anchor Leaf3 as illustrated in Figure 5: Example of IP addresses: secondary IP is used for the next-hop of the static route (an external router is connected to a non-anchor leaf node), on page 6, external-to-internal (L3Out-to-Web) traffic comes to the fabric through Leaf3 and then to the web endpoint in Host 4 through the spine switch.

The return traffic (Web-to-L3Out) goes back to an anchor leaf node, and then goes back to the virtual router through the non-anchor leaf node as illustrated in Figure 6: Return Traffic Flow Steered Toward the Anchor Leaf Node, on page 7 and Figure 7: Traffic Flow Bouncing between Anchor and Non-Anchor Leaf Nodes, on page 7. It's because the anchor leaf nodes learn the external route via the virtual router and redistributes the route to the other leaf nodes.

Figure 6: Return Traffic Flow Steered Toward the Anchor Leaf Node

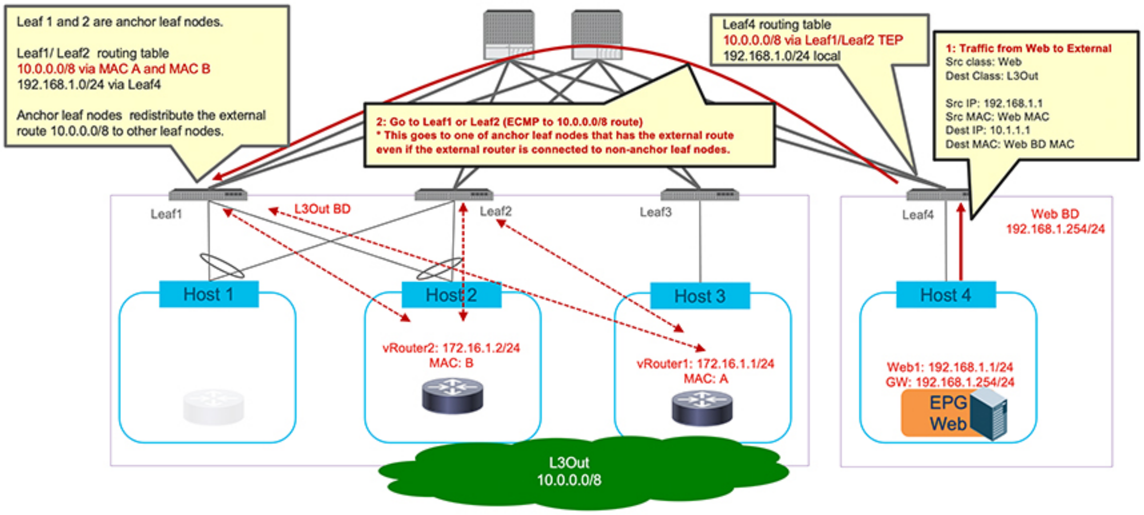
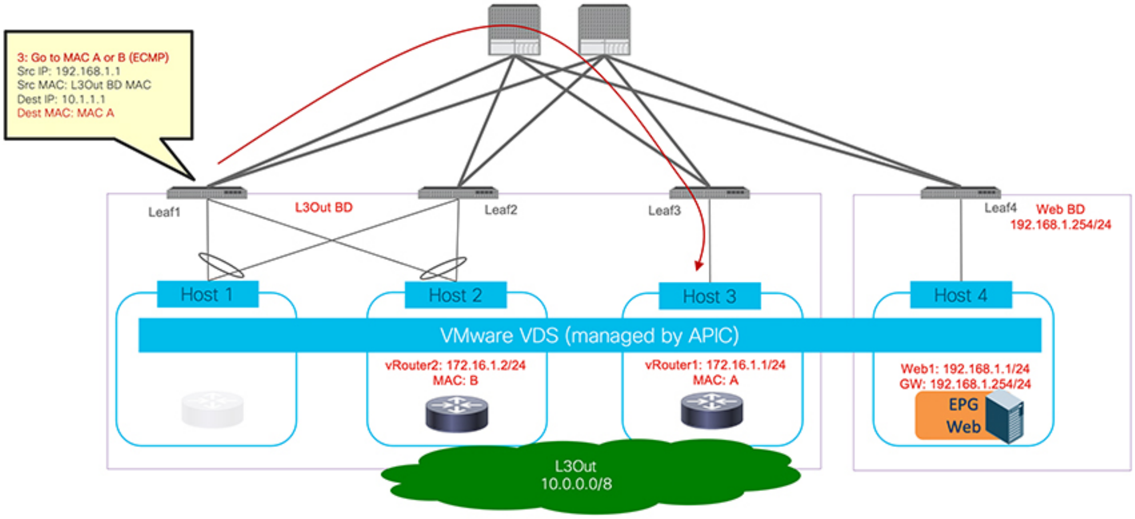

Leaf 1 and 2 are anchor leaf nodes.
Leaf1/ Leaf2 routing table
10.0.0.0/8 via MAC A and MAC B
192.168.1.0/24 via Leaf4
Anchor leaf nodes redistribute the external route 10.0.0.0/8 to other leaf nodes.
2: Go to Leaf1 or Leaf2 (ECMP to 10.0.0.0/8 route)
* This goes to one of anchor leaf nodes that has the external route even if the external router is connected to non-anchor leaf nodes.
Leaf4 routing table
10.0.0.0/8 via Leaf1/Leaf2 TEP
192.168.1.0/24 local
1: Traffic from Web to External
Src class: Web
Dest Class: L3Out
Src IP: 192.168.1.1
Src MAC: Web MAC
Dest IP: 10.1.1.1
Dest MAC: Web BD MAC
Leaf1
L3Out BD
Leaf2
Leaf3
Leaf4
Web BD
192.168.1.254/24
Host 1
Host 2
Host 3
Host 4
vRouter2: 172.16.1.2/24
MAC: B
vRouter1: 172.16.1.1/24
MAC: A
Web1: 192.168.1.1/24
GW: 192.168.1.254/24
EPG
Web
L3Out
10.0.0.0/8

Figure 7: Traffic Flow Bouncing between Anchor and Non-Anchor Leaf Nodes



Note

You can avoid this suboptimal path by using Cisco ACI Release 5.0. For more information, see Avoiding Suboptimal Traffic From a Cisco ACI Internal Endpoint to a Floating L3Out.