White paper
Cisco public

ACI Fabric Endpoint Learning

Contents

Introduction

This section provides an overview of the goals and prerequisites for this document.

Goals of this document

The Cisco Application Centric Infrastructure (Cisco ACI) solution can hold information about the location of MAC addresses and IPv4 (/32) and IPv6 (/128) addresses of endpoints in the Cisco ACI fabric. In addition to its use for traffic routing and bridging, endpoint information can be useful for traffic optimization, endpoint location tracking, and troubleshooting.

This document describes Cisco ACI endpoint learning behavior and deployment and presents a variety of optimization options. It focuses on specific use cases for endpoint IP address learning behavior.

Prerequisites

To best understand the design presented in this document, the reader must have a basic working knowledge of Cisco ACI technology.

For more information, see the Cisco ACI white papers available at Cisco.com: https://www.cisco.com/c/en/us/solutions/data-center-virtualization/application-centric-infrastructure/white-paper-listing.html.

Executive summary

This document covers features up to Cisco ACI Release 5.2(1g). It discusses deployment options using the data-plane learning options listed in Table 1. Detailed use cases and explanations are presented later in this document.

Table 1. Endpoint learning optimization options

Option name	Configuration location	Cisco ACI release when first introduced	Behavior	Considerations
			Benefit	
L4-L7 Virtual IPs	Tenant > Application Profiles > Application EPGs or uSeg EPGs > EPG	Release 1.2(1m)	Disables IP data-plane learning for specific endpoint IP addresses such as a Direct Server Return (DSR) virtual IP address	This option applies only to DSR.
			A workaround for DSR	
Unicast Routing	Tenant > Networking > Bridge Domains > BD	Release 1.0(1e)	Enables L3 routing and endpoint IP learning on a bridge domain	–
			Prevents IP learning by disabling it	
GARP-based EP Move Detection	Tenant > Networking > Bridge Domains > BD	Release 1.1(1j)	Uses Gratuitous Address Resolution Protocol (GARP) information to trigger an IP move when the move occurs on the same interface	ARP flooding must be enabled.
			A workaround for this behavior through which a particular IP to MAC binding changes on the same interface	

Option name	Configuration location	Cisco ACI release when first introduced	Behavior Benefit	Considerations
Limit IP Learning To Subnet	Tenant > Networking > Bridge Domains > BD	Release 1.1(1j)	Psrevents the local IP endpoint from being learned outside the subnets configured on the bridge domain Prevents mis-learning of IP addresses that may not belong to the fabric	This option does not prevent remote IP endpoint learning.
Endpoint Dataplane Learning	Tenant > Networking > Bridge Domains > BD	Release 2.0(1m)	Disables IP endpoint data-plane learning on a bridge domain A requirement for use of a service graph with Policy-Based Redirect (PBR)	This option is used only for service graphs with PBR.
IP Data-plane Learning	Tenant > Networking > Bridge Domains > BD > Subnets Tenant > Application Profiles > Application EPGs or uSeg EPGs > EPG > Subnets	Release 5.2(1g)	Disables IP endpoint data-plane learning per bridge domain subnet or IP	
IP Data-plane Learning	Tenant > Networking > VRFs > VRF	Release 4.0(1h)	Disables IP endpoint data-plane learning on a Virtual Routing and Forwarding (VRF) instance	-
Disable Remote EP Learn (on border leaf)	System > System Settings > Fabric-Wide Setting Prior to APIC Release 3.0(1k): Fabric > Access Policies > Global Policies > Fabric Wide Setting Policy > Disable Remote EP Learn	Release 2.2(2e)	Disables remote IP endpoint learning on border leaf switches for VRF instances; border leaf switches use the spine proxy exclusively	Prior to Cisco ACI Release 3.0(2h), this option requires ingress policy enforcement in the VRF instance. For second-generation leaf switches, remote IP learning is allowed only for Layer 3 multicast in order to properly forward (S, G) packets. First-generation leaf switches do not support Layer 3 multicast.
Enforce Subnet Check	System > System Settings > Fabric-Wide Setting Prior to APIC Release 3.0(2h): Fabric > Access Policies > Global Policies > Fabric Wide Setting Policy	Release 2.2(2q)*	Prevents mis-learning of IP addresses that may not belong to the fabric Limits both local and remote endpoint learning to instances only when the source IP address belongs to a bridge-domain subnet in the VRF instance	This option applies only to second-generation Cisco ACI leaf switches.

Option name	Configuration location	Cisco ACI release when first introduced	Behavior Benefit	Considerations
	> Enforce Subnet Check		Prevents mis-learning of IP addresses that may not belong to the fabric	
IP Aging	System > System Settings > Endpoint Controls > IP Aging Prior to APIC Release 3.0(1k): Fabric > Access Policies > Global Policies > Fabric Wide Setting Policy > IP Aging	Release 2.1(1h)	Tracks and ages unused IP addresses on an endpoint Prevents IP addresses from remaining stuck to an endpoint, even when the IP address is no longer used	This option is a default setting for Cisco ACI Release 2.1(1h) and later.
Rogue EP Control	System > System Settings > Endpoint Controls > Rogue EP Control	Release 3.2(1l)	Detects endpoints that move frequently and disables endpoint learning for these endpoints only Identifies rogue endpoints and minimizes impacts caused by them	Starting from Cisco ACI Release 5.2(3) it is possible to configure an "exception" list of MAC addresses to which the Rogue EP Control policy is not applied.
COOP Endpoint Dampening	This is available via API only.	Release 4.2(3)	Mitigate the impact of unreasonable amounts of endpoint updates of an endpoint on spine nodes	This is enabled by default after Cisco ACI Release 4.2(3).

* Although the Enforce Subnet Check option is first introduced in 2.2(2q) and is available on later releases on the Cisco ACI Release 2.2 release train, it is not available in Cisco ACI Release 2.3 or 3.0(1x) – for example, it is not available in releases 2.3(1f) or 3.0(1k). It is supported from 3.0(2h) onward.

Table 2 lists endpoint learning behaviors in different combinations of endpoint learning options.

Table 2. Endpoint learning behaviors comparison

Configurations			Learning behavior				
IP Data-plane Learning (onVRF, BD subnet, or EPG subnet	Endpoint Data-plane Learning (on BD for PBR)	Remote EP Learning (on border leaf)[1]	Local MAC	Local IP	Remote MAC	Remote IP (unicast)	Remote IP (multicast)
Disabled	Disabled/Enabled	Disabled/Enabled	Learned[*5]	**Not learned**[2]	**Learned**[*3 *5]	Not learned	Not Learned[*6]
Enabled	Disabled	Disabled/Enabled	Learned	**Not learned**[2]	Not learned	Not learned	Not learned
Enabled	Enabled	Disabled	Learned	Learned	Learned	**Not learned**[4]	Learned
Enabled	Enabled	Enabled	Learned	Learned	Learned	Learned	Learned

[1] This option is called "Disabled Remote EP Learning" in APIC GUI under System Settings.

[2] Not learned via the data plane but learned via the control plane, such as ARP.

[3] In first-generation Cisco ACI leaf switches, remote MAC is not learned.

[4] Not learned only on border leaf switches.

[5] In the case of the option on a BD subnet or EPG subnet, it is not learned via an endpoint-to-endpoint ARP request that does not reach a CPU. (An ARP request to a bridge domain SVI gateway is still learned.) In the case of the option on a VRF, local and remote MACs are learned via an endpoint-to-endpoint ARP request.

[6] Prior to Cisco ACI Release 5.2(1g), it is learned.

Table 3 lists the Cisco ACI leaf switches by generation.

Table 3. Cisco ACI leaf switch generations

Generation	Switch	Part number
First-generation Cisco ACI leaf switches	Cisco Nexus 9332PQ Switch	N9K-C9332PQ
	Cisco Nexus 9372PX Switch	N9K-C9372PX
	Cisco Nexus 9372PX-E Switch	N9K-C9372PX-E
	Cisco Nexus 9372TX Switch	N9K-C9372TX
	Cisco Nexus 9372TX-E Switch	N9K-C9372TX-E
	Cisco Nexus 9396PX Switch	N9K-C9396PX
	Cisco Nexus 9396TX Switch	N9K-C9396TX
	Cisco Nexus 93120TX Switch	N9K-C93120TX
	Cisco Nexus 93128TX Switch	N9K-C93128TX
Second-generation Cisco ACI leaf switches and later	Cisco Nexus 93180YC-EX	N9K-C93180YC-EX
	Cisco Nexus 93180YC-FX	N9K-C93180YC-FX
	Cisco Nexus 93108TC-EX	N9K-C93108TC-EX
	Cisco Nexus 93108TC-FX	N9K-C93108TC-FX
	Cisco Nexus 93180LC-EX	N9K-C93180LC-EX
	Cisco Nexus 9348GC-FXP	N9K-C9348GC-FXP
	Cisco Nexus 9336C-FX2	N9K-C9336C-FX2
	Cisco Nexus 93240YC-FX2	N9K-C93240YC-FX2
	Cisco Nexus 93216TC-FX2	N9K-C93216TC-FX2
	Cisco Nexus 93360YC-FX2	N9K-C93360YC-FX2
	Cisco Nexus 93180YC-FX3	N9K-C93180YC-FX3
	Cisco Nexus 93108TC-FX3P	N9K-C93108TC-FX3P
	Cisco Nexus 9316D-GX	N9K-C9316D-GX
	Cisco Nexus 93600CD-GX	N9K-C93600CD-GX
	Cisco Nexus 9364C-FX	N9K-C9364C-GX

Cisco ACI endpoint learning behavior

This section provides an overview of Cisco ACI endpoint learning behavior.

Cisco ACI and endpoints

Cisco ACI uses endpoints to forward traffic. An endpoint consists of one MAC address and zero or more IP addresses. Each endpoint represents a single networking device (Figure 1).

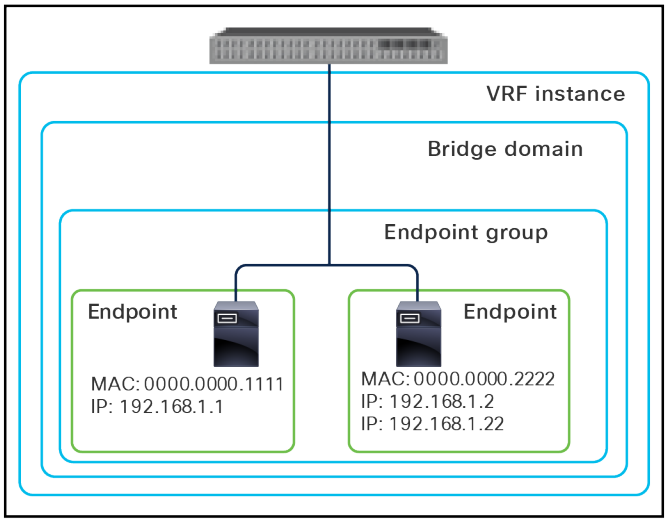


Figure 1.
Cisco ACI and endpoints

In a traditional network, three tables are used to maintain the network addresses of external devices: a MAC address table for Layer 2 forwarding, a Routing Information Base (RIB) for Layer 3 forwarding, and an ARP table for the combination of IP addresses and MAC addresses. Cisco ACI, however, maintains this information in a different way, as shown in Table 4.

Table 4. Cisco ACI and traditional networks

Traditional network		**Cisco ACI**	
Table	**Table role**	**Table**	**Table role**
RIB	• IPv4 addresses (/32 and non-/32) • IPv6 addresses (/128 and non-/128)	**RIB**	• IPv4 (non-/32*) • IPv6 (non-/128*)
MAC address table	MAC addresses	**Endpoint**	MAC and IP addresses (/32 or /128 only)
ARP table	Relationship of IP to MAC	**ARP**	Relationship of IP to MAC (only for Layer 3 outside [L3Out] connections)

* Cisco ACI bridge domain Switch Virtual Interfaces (SVI), routed port and sub-interface IP addresses, and advertised and static routes are in the RIB regardless of whether it is /32 (IPv4) or /128 (IPv6).

As Table 4 shows, Cisco ACI replaced the MAC address table and ARP table with a single table called the endpoint table. This change implies that Cisco ACI learns that information in a different way than in a traditional network. Cisco ACI learns MAC and IP addresses in hardware by looking at the packet source MAC address and source IP address in the data plane instead of relying on ARP to obtain a next-hop MAC address for IP addresses. This approach reduces the amount of resources needed to process and generate ARP traffic. It also allows detection of IP address and MAC address movement without the need to wait for GARP as long as some traffic is sent from the new host.

L3Out and regular endpoints

Although Cisco ACI mainly uses the endpoint table instead of the MAC address and ARP tables, it still uses the RIB and the ARP table. This capability is especially for L3Out communication, because the maximum number of IP addresses on a single endpoint (one MAC address) is limited, and there can be a huge number of IP addresses behind a single next-hop MAC address (external router) on a L3Out connection. For information about the number of addresses allowed, see the scalability guide for the release you are using. (For Cisco ACI Release 5.2(1g), the maximum number of entries is 4096.)

Regardless of this limitation, it is not efficient to maintain all outside IP addresses as separate /32 or /128 endpoints. Cisco ACI must know how to reach these IP addresses as prefixes through routing protocols such as Open Shortest Path First (OSPF), which is the same behavior as for traditional routers. However, Cisco ACI needs to know only the next hop (external router) for those prefixes. Because of this consideration, Cisco ACI uses a behavior similar to that in traditional networks for L3Out connectivity. The Cisco ACI L3Out domain learns the MAC address only from the data plane. IP addresses are not learned from the data plane in an L3Out domain; instead, Cisco ACI uses ARP to resolve next-hop IP and MAC relationships to reach the prefixes behind external routers.

Local endpoints and remote endpoints

A leaf switch has two types of endpoints: local endpoints and remote endpoints. Local endpoints for LEAF1 reside directly on LEAF1 (For example, directly attached), whereas remote endpoints for LEAF1 reside on other leaf endpoints (Figure 2).

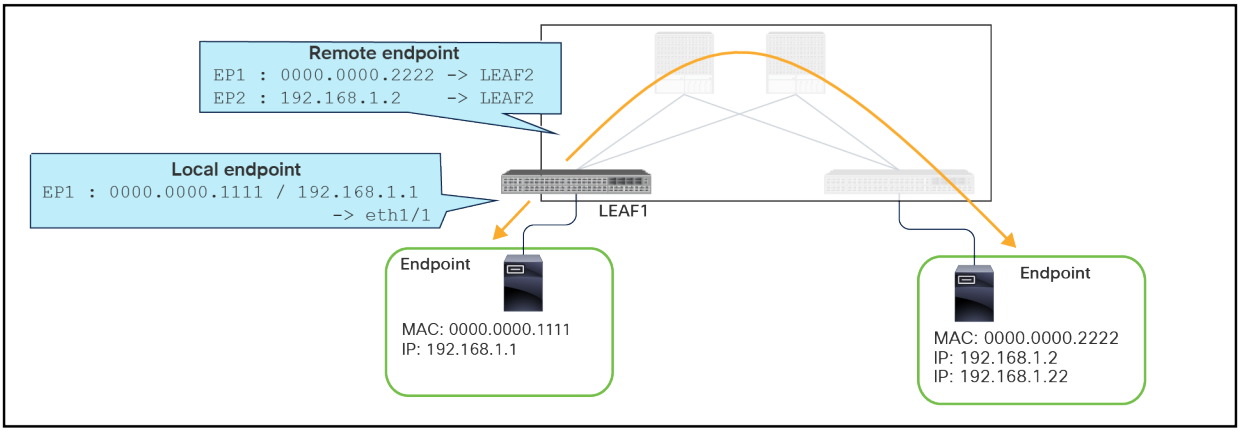


Figure 2.
Local and remote endpoints

Although both local and remote endpoints are learned from the data plane, remote endpoints are merely a cache, local to each leaf. Local endpoints are the main source of endpoint information for the entire Cisco ACI fabric. Each leaf is responsible for reporting its local endpoints to the Council Of Oracle Protocol (COOP) database, located on each spine switch, which implies that all endpoint information in the Cisco ACI fabric is stored in the spine COOP database. Because this database is accessible, each leaf does not need to know about all the remote endpoints to forward packets to the remote leaf endpoints. Instead, a leaf can forward packets to spine switches, even if the leaf does not know about a particular remote endpoint. This forwarding behavior is called spine proxy.

The meaning of remote endpoints

Because of spine proxy, Cisco ACI packet forwarding will work without remote endpoint learning. Spine proxy enables leaf switches to forward traffic directly to the COOP database located on the spine switches. Remote endpoint learning helps Cisco ACI forward packets more efficiently by allowing leaf switches to send packets directly to a destination leaf switch without using the resources on the spine switch that would be used to look up endpoints in the COOP database, which contains all the fabric endpoint information.

Remote endpoints are learned from data-plane traffic, as are local endpoints. Therefore, only leaf switches with actual communication traffic create a cache entry for remote endpoints (conversational learning) to forward the packets directly toward the destination leaf. Remote endpoints have either one MAC address or one IP address per endpoint, instead of a MAC address and IP address combination as is the case for local endpoints (as depicted in Figure 2). One reason for this difference is that the IP to MAC next-hop resolution can be performed on the destination leaf, and the next-hop MAC address is not required just to reach the destination leaf. This behavior also helps each leaf save its resources for these caches. Also, the age timer for a remote endpoint is shorter than for a local endpoint because a remote endpoint is just a cache and should not be present after the conversation has ceased and the original local endpoint on another leaf has disappeared.

The on-peer endpoint is a variant of the remote endpoint. These endpoints are remote endpoints that point to a port that is not part of a virtual Port Channel (vPC), also called an orphan port, on a vPC peer leaf. They are special endpoints because they are remote, but they are learned through vPC synchronization on the control plane instead of through data-plane learning from the actual traffic. As a result of vPC synchronization, on-peer endpoints have MAC address, IP address, and EPG information, unlike other remote endpoints, which have either bridge domain (MAC address) or VRF (IP address) information.

Table 5 summarizes the differences between local and remote endpoints.

Table 5. Differences between local and remote endpoints

Feature	Local endpoint	Remote endpoint
1 endpoint	1 MAC address and **n** IP addresses	1 MAC address or 1 IP address
Scope	Reported to spine COOP database	Only on each leaf as a cache entry
Endpoint retention timer*	900 seconds (by default)	300 seconds (by default)

* You can configure the endpoint retention timer at Tenant > Policies > Protocol > End Point Retention.

IP addresses of the local endpoint can be aged out separately depending on the IP aging policy. Refer to the IP Aging Policy section of this document for details.

Figure 3 shows an example of Command-Line Interface (CLI) output for a local and a remote endpoint on a leaf switch.

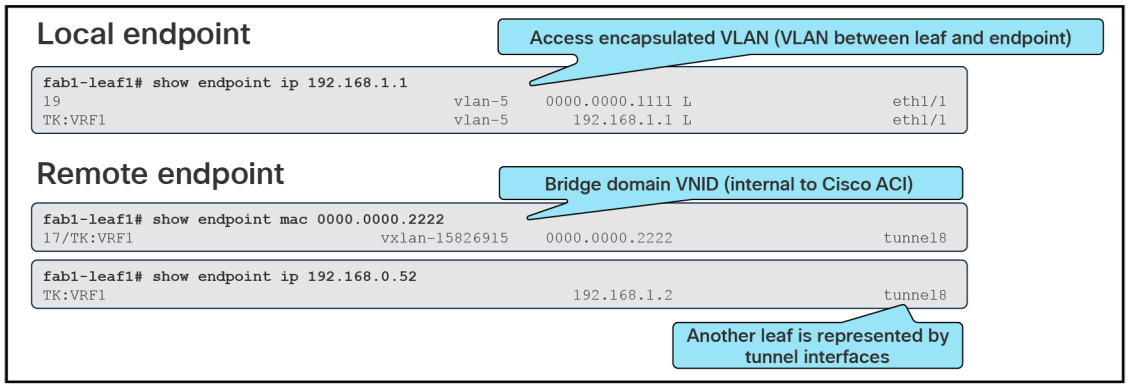


Figure 3.
Local and remote endpoint CLI output on each leaf switch

Endpoint learning

	The endpoint learning behavior described here is based on the assumption that unicast routing is enabled on the bridge domain. If unicast routing is not enabled, a Cisco ACI leaf cannot perform routing and cannot learn any IP addresses. It learns only MAC addresses and performs switching. For more information, see the Unicast Routing option section of this document.

Local endpoint learning

Cisco ACI learns the MAC (and IP) address as a local endpoint when a packet comes into a Cisco ACI leaf switch from its front-panel ports.

	Front-panel ports are southbound ports from the perspective of Cisco ACI and do not face spine switches.

A Cisco ACI leaf switch follows these steps to learn a local endpoint MAC address and IP address:

1. The Cisco ACI leaf receives a packet with a source MAC Address (MAC A) and source IP Address (IP A).
2. The Cisco ACI leaf learns MAC A as a local endpoint.

3a. If the packet is an ARP request, the Cisco ACI leaf learns IP A tied to MAC A based on the ARP header.

3b. If the packet is an IP packet and routing is performed by the Cisco ACI leaf, the Cisco ACI leaf learns IP A tied to MAC A based on the IP header.

Thus, if the packet is switched and not an ARP packet, the Cisco ACI leaf never learns the IP address but only the MAC address. This behavior is the same as traditional MAC address learning behavior on a traditional switch.

	First-generation leaf switches cannot reflect IP address movement between two MAC addresses on the same interface with the same VLAN to the endpoint database. This sort of IP address movement may occur in a high-availability failover scenario in which GARP typically is used to update IP to MAC relation on upstream network devices. This behavior is resolved by enabling the GARP-based EP Move Detection option discussed later in this document.

Remote endpoint learning

Cisco ACI learns a MAC or IP address as a remote endpoint when a packet comes into a Cisco ACI leaf switch from another leaf switch through a spine switch. When a packet is sent from one leaf to another leaf, Cisco ACI encapsulates the original packet with an outer header representing the source and destination leaf Tunnel Endpoint (TEP) and the Virtual Extensible LAN (VXLAN) header, which contains the bridge domain or VRF information of the original packet.

	Packets that are switched contain bridge domain information. Packets that are routed contain VRF information.

A Cisco ACI leaf switch follows these steps to learn a remote endpoint MAC or IP address:

1. The Cisco ACI leaf receives a packet with source MAC A and source IP A from a spine switch.
2. The Cisco ACI leaf learns MAC A as a remote endpoint if VXLAN contains bridge domain information.
3. The Cisco ACI leaf learns IP A as a remote endpoint if VXLAN contains VRF information.

Figures 4 and 5 show examples of local and remote endpoint learning.

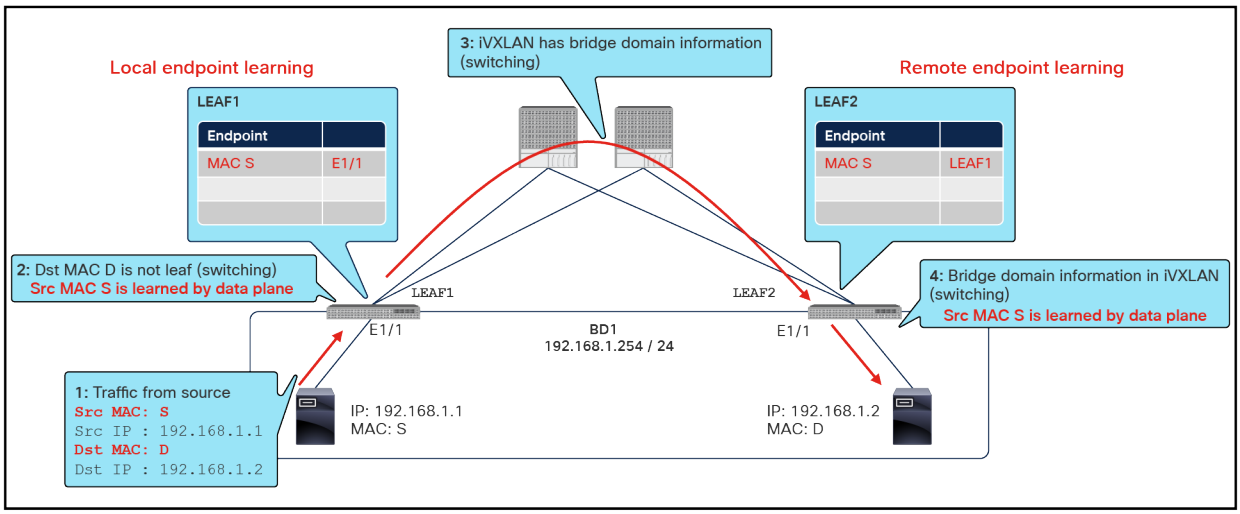


Figure 4.
Example of local and remote endpoint (MAC address) learning

In Figure 4, the packet is Layer 2 traffic without any routing on Cisco ACI. Therefore, only the MAC address (Src MAC S in the figure) is learned as a local endpoint on LEAF1 and a remote endpoint on LEAF2.

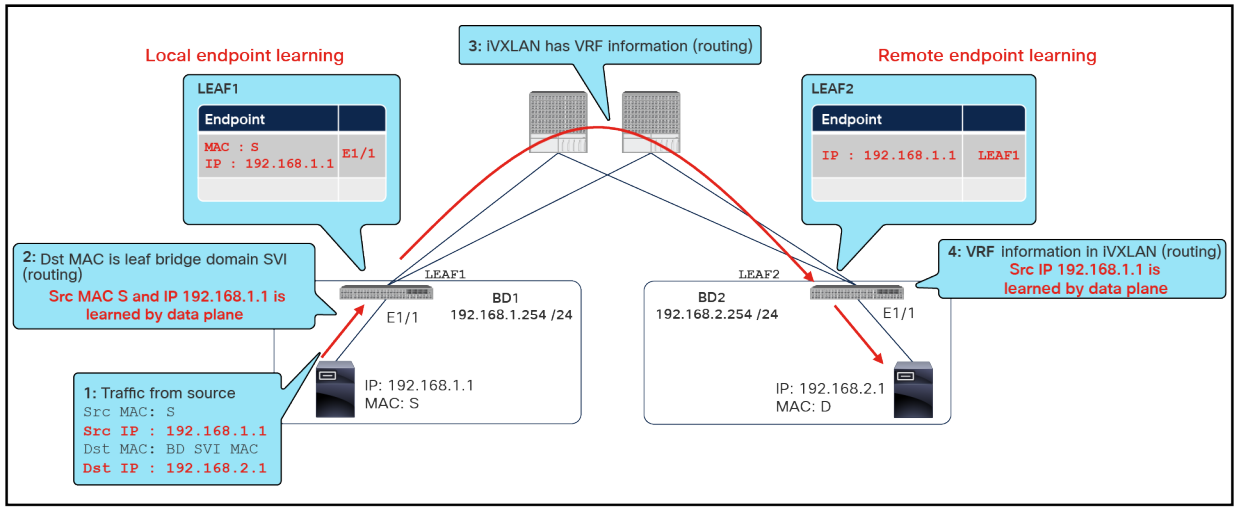


Figure 5.
Example of local and remote endpoint (IP address) learning

In Figure 5, the packet is Layer 3 traffic with the Cisco ACI bridge domain Switch Virtual Interface (SVI) as its default gateway. Therefore, both the MAC address and IP address (Src MAC S and Src IP 192.168.1.1 in the figure) are learned as a single local endpoint on LEAF1, and only IP address 192.168.1.1 is learned as a remote endpoint on LEAF2.

Aging of endpoints

The aging timer (or retention timer) of endpoints in ACI is configured through **Endpoint Retention Policy**. Local and remote endpoints have different aging intervals, and the one for remote endpoints should always be shorter than the one for local endpoints.

For local endpoints that have IP addresses associated with them, leaf switches perform **Host Tracking** at 75 percent of the aging interval if no packets from the same IP address was received during the interval. **Host Tracking** is always enabled and will send three ARP requests to the IP address in order to make sure that the IP address is still responsive. If a packet from the IP address was received at least once during the aging interval (precisely speaking, during 75 percent of the interval), **Host Tracking** is not performed, and the aging interval is reset at the timing of **Host Tracking** (75 percent of the interval). Note that the interval is not reset when the packet was received. This means if a packet was received very early in the aging interval and not thereafter, the age timer continues to run until the **Host Tracking** interval expires, which may result in the endpoint being learned for almost two times the configured aging interval in the worst case. IP addresses of the local endpoint can be aged out separately, depending on the IP aging policy. Refer to the IP Aging Policy section of this document for details.

For local endpoints without any IP addresses and remote endpoints, **Host Tracking** is not performed. If a packet from the same address was not received at all during the aging interval, the endpoint ages out and gets deleted. If a packet from the same address was received at least once during the interval, the aging is reset at the end of the interval and the endpoint remains learned for another aging interval. This means, just as the local IP endpoints do, the endpoint may remain learned for almost two times of the configured interval depending on when the last packet was received.

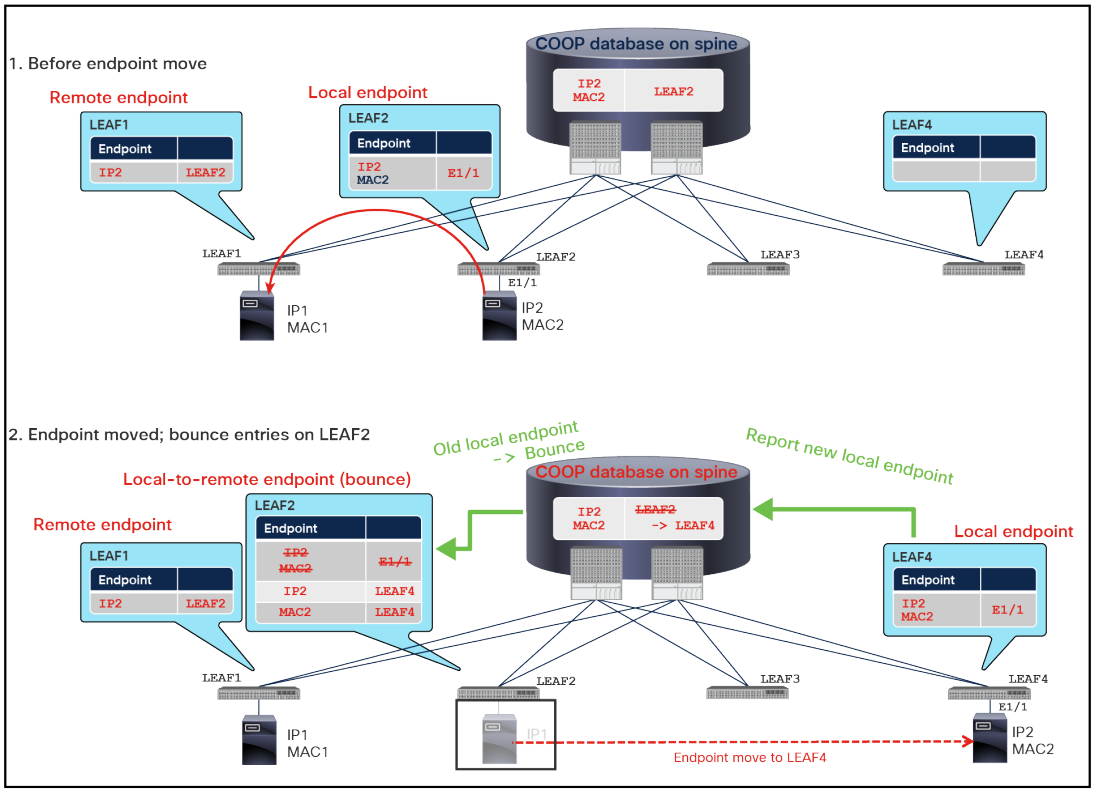
Endpoint movement and bounce entries

There are several scenarios in which an endpoint moves between two Cisco ACI leaf switches, such as a failover event or a virtual machine migration in a hypervisor environment. Cisco ACI data-plane endpoint learning detects these events quickly and updates the Cisco ACI endpoint database on a new leaf. In addition to data-plane learning, Cisco ACI uses bounce entries to manage the old endpoint information on the original leaf.

When a new local endpoint is detected on a leaf, the leaf updates the COOP database on spine switches with its new local endpoint. If the COOP database has already learned the same endpoint from another leaf, COOP will recognize this event as an endpoint move and report this move to the original leaf that contained the old endpoint information. The old leaf that receives this notification will delete its old endpoint entry and create a bounce entry, which will point to the new leaf. A bounce entry is basically a remote endpoint created by COOP communication instead of data-plane learning.

	The difference between a bounce entry and a remote endpoint is in whether or not the leaf rewrites the outer source IP address of the packet. When a packet uses a normal remote endpoint, the Cisco ACI leaf uses its own TEP address as the outer source IP address, so the remote leaf learns this packet with its own TEP. When a packet uses a bounce entry, the Cisco ACI leaf doesn't rewrite the outer source IP address, so the remote data-plane learning will behave as if the packet came from the originating leaf rather than the intermediate "bounce" leaf.
	The endpoint retention timer value (aging interval) for a bounce entry is 630 seconds by default. You can tune this value by going to Tenant > Policies > Protocol > End Point Retention, where you can also find the other endpoint retention timers.

Figure 6 shows an example of endpoint movement and a bounce entry.



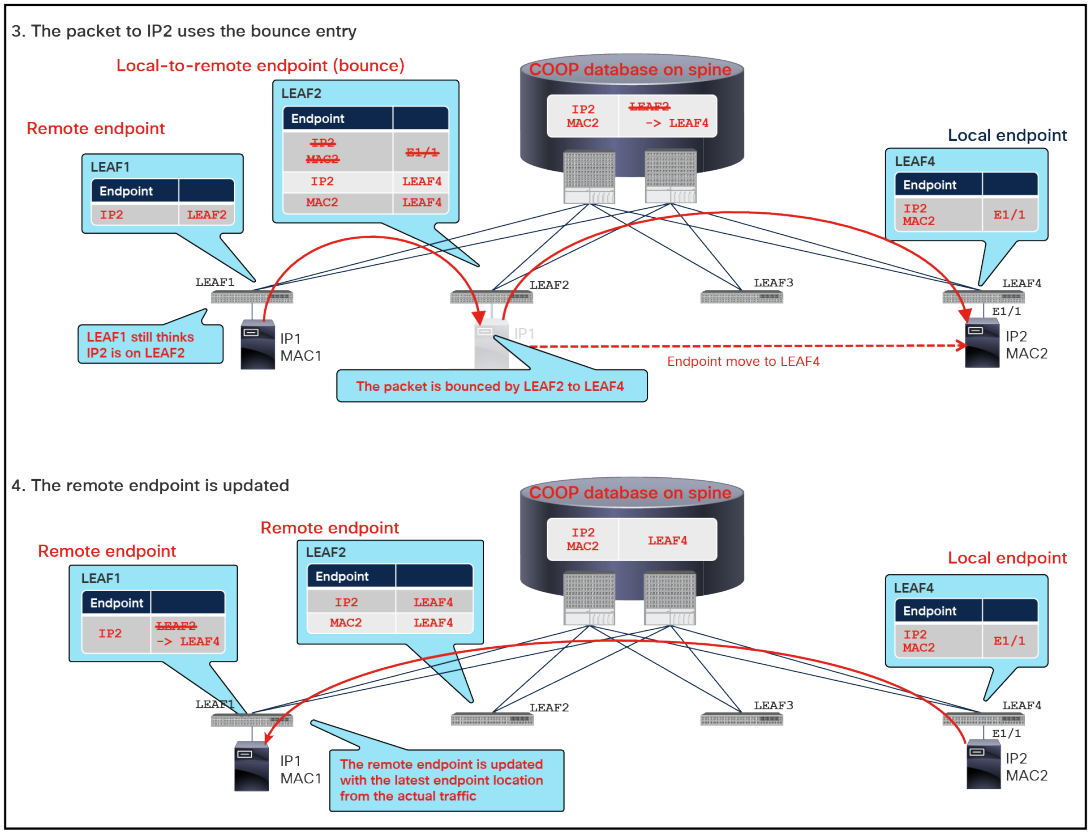


Figure 6.
Example of Cisco ACI bounce entry

At the first step in Figure 6, LEAF1 learns the remote endpoint location for IP2 pointing to LEAF2 from the data plane.

In the second step, the endpoint with MAC2 and IP2 on LEAF2 moves to LEAF4, and the new local endpoint is created on LEAF4. This new local endpoint is reported to the COOP database on the spine switches, which in turn will notify LEAF2 about this move, and LEAF2 will install bounce entries for MAC2 and IP2. Bounce entries are basically the same as remote endpoints. Hence two bounce entries, for the MAC and IP addresses, are created at step 2 in Figure 6.

Note: If only MAC2 moved from LEAF2 to LEAF4 without the associated IP, the MAC is learned as a new endpoint on LEAF4 and LEAF2 will install the bounce entry only for MAC2. At this time, the ACI Fabric is not sure about whether IP2 moved along with MAC2. Hence, LEAF2 will instead install the bounce-to-proxy entry for IP2, which points to spine proxy instead of LEAF4.

At this point, LEAF1 still has the old remote endpoint for IP2, which still points to the old location: LEAF2. If a packet is sent from LEAF1 to IP2 at this time, LEAF1 forwards it to LEAF2, instead of LEAF4, based on its remote endpoint cache. Because of the bounce entries, LEAF2 is already prepared for this sort of forwarding from leaf switches with old remote endpoints. LEAF2 will then bounce the packet to the new LEAF4 based on its bounce entries. This bounce entry is a backup mechanism for this type of scenario. Therefore, the bounce entry will not be used if the new traffic from IP2 on LEAF4 reaches LEAF1 before LEAF1 sends packets to IP2, because the old remote endpoint on LEAF1 will be updated directly by the data-plane traffic from the new leaf.

The advantage of this implementation is scale. No matter how many leaf switches have learned endpoint information, only three components will need to be updated after an endpoint moves. The three components are the COOP database, the new leaf switch to which the endpoint has moved, and the old leaf switch from which the endpoint has moved. Eventually, all other leaf switches in the fabric will update their information about the location of the endpoint through data-plane traffic. However, there are a few corner cases where the other leaf switches keep the outdated remote endpoint information even after the bounce entry ages out. This could cause traffic to be black-holed by sending the traffic to a leaf that doesn't have the destination endpoint or the bounce entry. An example of the corner case is mentioned in Disable Remote EP Learn (on border leaf) section with CSCva56754. To address this concern, the following endpoint announce messages were introduced.

Endpoint Announce Enhancement

Endpoint announce messages were enhanced to cover corner cases where stale remote endpoints need to be deleted on all leaf switches based on an endpoint learning event that happened on one specific leaf. Typically, this is not required with our bounce entry mechanism because endpoint information on leaf switches that do not own the endpoint (that is, a remote endpoint) is updated through data-plane learning through conversations after an endpoint has moved. Until that happens, the bounce entry takes care of the traffic. However, there are some corner cases in which a subsequent conversation did not take place or data-plane learning did not update the remote endpoints as expected. **Endpoint announce** messages address those corner cases. **Endpoint announce** messages are always enabled and cannot be disabled.

Trigger	Description	Integrated ACI switch release
Aging out of an IP bounce entry	When an IP bounce entry ages out, the corresponding remote endpoints on other leaf switches will be flushed if the remote endpoints are pointing to an incorrect leaf. This is not applicable when the bounce entry is of MAC address.	13.2(2l) - CSCvj17665
A pcTag change of an EPG	When the pcTag of an EPG is changed, the corresponding remote endpoints on other leaf switches will be flushed.	14.0(1h) - CSCvk22720
Unicast routing is disabled, or a BD subnet is deleted while unicast routing is enabled	When the trigger occurs, leaf switches that have the corresponding BD locally deployed will flush the corresponding IP endpoints, both local and remote regardless of endpoint announce messages. Endpoint announce messages enhance this to flush the corresponding remote IP endpoints on all other leaf switches.	14.0(1h)
A vPC port on a remote leaf pair becomes operational	When a vPC port is operational only on one of the leaf switches in the vPC pair, the remote endpoints for endpoints on the vPC port will point to the Physical TEP (PTEP) of the operational leaf. When the vPC port on the other leaf also becomes operational, those remote endpoints should now point to the Virtual TEP (VTEP) representing both of the leaf switches. This may blackhole traffic for remote leaf deployment. Endpoint announce messages flush remote endpoints that are still pointing to the PTEP of a remote leaf.	14.2(1i) - CSCvp97665

Silent hosts considerations

In the case of silent hosts, where an ACI leaf hasn't learned a local endpoint, ACI has some mechanisms to detect those silent hosts. Some of them are controlled by BD configurations. Following are explanations of each scenario with related BD configurations.

For (L2) switched traffic to an unknown MAC, the L2 Unknown Unicast option under the BD may need to be set to "Flood". This is because the ACI fabric with the L2 Unknown Unicast "Hardware-Proxy" configuration drops the L2 unicast packets on the spine in cases where the destination MAC has not been learned as an endpoint anywhere on the BD in ACI, and the COOP database doesn't have the information.

For (L3) routed traffic to an unknown IP, the ACI leaf will generate an ARP request from its BD SVI (pervasive gateway) IP toward the unknown IP in order to detect and learn the unknown IP. Only the leaf with BD SVI IP for the unknown IP subnet will generate an ARP request. This behavior is originally triggered by the spine when the spine couldn't find the unknown IP, even in the COOP database. This behavior is called silent host detection, or ARP gleaning. This behavior for (L3) routed traffic happens regardless of configuration, such as L2 Unknown Unicast or ARP flooding (mentioned below), as long as the traffic is routed to an unknown IP.

For ARP requests with a broadcast destination MAC, the ARP flooding option under the BD controls the flooding behavior. The ACI fabric will flood the ARP request within the BD when the ARP flooding option is enabled. Because the frame is flooded to the entire BD, any silent host would receive the packet and respond appropriately. In cases where the ARP flooding option is disabled, the ACI fabric will perform unicast routing against the target IP located in the ARP header instead of flooding. If the target IP is a silent host and unknown to the leaf and spines (COOP), the ACI leaf switches will follow the same procedure as the (L3) routed traffic scenario discussed earlier (ARP gleaning) and generate an ARP request from its BD SVI (pervasive gateway), even if the target IP and the source IP are in the same subnet. This implies that both enabling and disabling the ARP flooding option can detect most silent hosts.

The difference of enabling and disabling the ARP flooding option appears once ACI learns a silent host IP. One of the benefits of enabling ARP flooding is to be able to detect a silent IP that moved from one location to another without notifying an ACI leaf. Because the ARP request is flooded within the BD, even if the ACI leaf still thinks the IP is at the old location, the host with the silent IP would respond appropriately so that the ACI leaf can update its entry accordingly. If ARP flooding is disabled, the ACI leaf would keep forwarding the ARP request only to the old location until the IP endpoint ages out. On the other hand, the benefit of disabling ARP flooding is to be able to optimize traffic flow by sending the ARP request directly to the location of the target IP, assuming no endpoint moves without notifying its movement via GARP and such.

With that said, as long as an endpoint sends an ARP request to silent hosts, silent hosts can be detected by the ACI fabric regardless of the L2 Unknown Unicast option mentioned earlier, even in the case of (L2) intra subnet communication. In this case, ACI floods the ARP request or performs ARP gleaning after seeing the ARP request. Note that only the ARP request, and no other data-plane traffic, will trigger ARP gleaning in the case of (L2) intra subnet communication.

For more information about ACI BD options, refer to the following documents:

- Cisco Application Centric Infrastructure Design Guide White Paper – Switching in the overlay
- Cisco Application Centric Infrastructure Design Guide White Paper – Bridge Domain Design Considerations

vPC considerations

In the case of vPC, each encap VLAN within a bridge domain should be deployed symmetrically across both leaf switches of a vPC pair to prevent endpoint synchronization inconsistencies. This is automatically achieved when a VLAN is deployed on a vPC port. If a VLAN is deployed only on an orphan port, which is a port that is not a vPC port but still on one of the vPC pair leaf switches, ensure that your configuration deploys the same encap VLAN on both leaf switches of a vPC pair. An example configuration is where you deploy the same encap VLAN on orphan ports on both leaf switches of a vPC pair.

L3Out endpoint learning considerations

L3Out traffic behaves differently than normal endpoint traffic, as mentioned earlier in the section L3Out and regular endpoints.

Table 6 lists the main considerations for endpoint learning. A detailed example for each scenario follows.

Table 6. Endpoint learning with L3Out connections

Scenario	L3Out-specific behavior	Considerations
Scenario 1	Local endpoint learning with an incoming packet from L3Out to Cisco ACI: Only the source MAC address is learned as a local endpoint. The source IP address is not learned as a local endpoint.	–
Scenario 2	Remote endpoint learning with an incoming packet from L3Out to Cisco ACI: No source MAC or IP address is learned as a new remote endpoint by a packet.*	The endpoint retention timer for an existing remote endpoint is refreshed by this packet from L3Out, even though other information, such as the originating leaf switch, is not updated. This behavior may cause a stale remote endpoint to not age out correctly after an endpoint is migrated to L3Out from within Cisco ACI. You can use the Enforce Subnet Check feature to mitigate this situation.
Scenario 3	Remote endpoint learning with an outgoing packet to L3Out from Cisco ACI: No source MAC or IP address is learned as a new remote endpoint if the VRF mode is ingress policy enforcement. This behavior is observed only when a packet to L3Out is sourced from a first-generation leaf switch.	The endpoint retention timer for the existing remote endpoint is refreshed by this packet to L3Out, even though other information, such as the originating leaf switch, is not updated. This behavior may cause a stale remote endpoint to not age out correctly after an endpoint is moved to a different leaf. You can use the Disable Remote EP Learn feature on the border leaf to prevent this situation.
Scenario 4	Source IP address that falls under L3Out routes is not learned as an endpoint: (Note that 0.0.0.0/0 doesn't have this effect.) This behavior is observed only with second-generation leaf switches.	This behavior mitigates the unexpected endpoint learning issue caused by a spoofing packet or misconfigured endpoint.

* An exception exists for remote MAC address learning when a packet is incoming from L3Out to Cisco ACI. If ARP traffic is coming from an L3Out SVI rather than a routed-port sub-interface, ARP traffic is flooded to other leaf switches with the same L3Out SVI. This behavior could cause remote MAC address learning on another border leaf switch.

Scenario 1: Local endpoint learning with an incoming packet from L3Out

There are no special considerations for this scenario.

Scenario 2: Stale remote endpoint example with L3Out incoming traffic

Figure 7 shows an example of scenario 2 in Table 6. Although it's not a common scenario, if you perform such an operation, pay attention to this behavior.

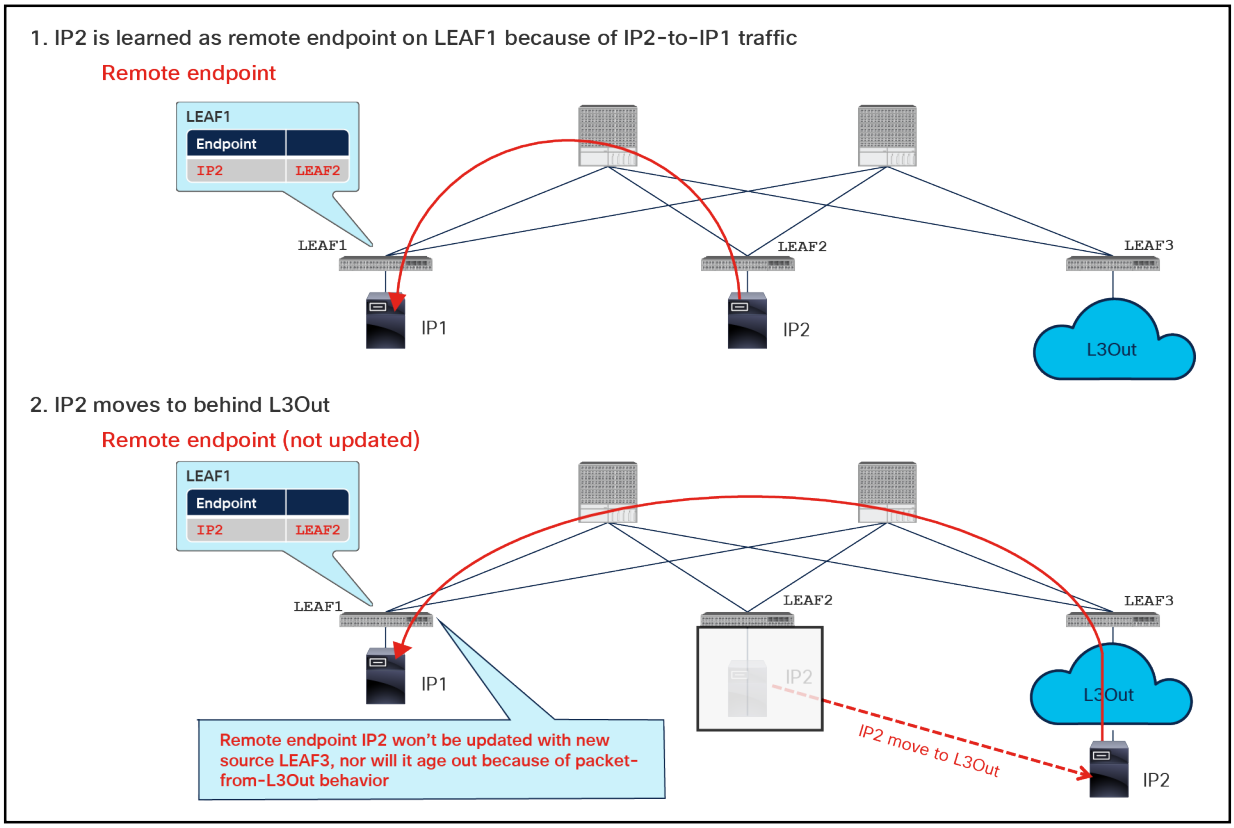


Figure 7.
Stale endpoint after endpoint migration to L3Out from Cisco ACI (scenario 2)

At first, IP2 on LEAF2 is learned on LEAF1 as a remote endpoint because of the traffic from IP2 to IP1. After that, a device with IP2 moved to a network behind the L3Out connection and resumes its communication to IP1 before the remote endpoint for IP2 on LEAF1 ages out. At this point, the remote endpoint still points to the old LEAF2 entry, instead of the new LEAF 3 entry, but this old remote endpoint will never be updated to point to LEAF3, nor will it age out because of the particular behavior on the L3Out connection, as described for scenario 2 (stale remote endpoint example with L3Out incoming traffic) in Table 6.

	In scenario 2, a bounce entry for IP2 on LEAF2 is not created because a bounce entry is created only when a Cisco ACI leaf detects the same MAC and/or IP address as a local endpoint on another leaf. Cisco ACI cannot detect this movement if the endpoint moves to an L3Out connection.

Because of this stale remote endpoint, any traffic from LEAF1 toward IP2 will fail, because LEAF1 sends packets to the wrong leaf.

This stale remote endpoint on LEAF1 needs to be manually cleared to resume communication. The command syntax to manually clear a particular remote IP endpoint is shown here:

```
LEAF1# clear system internal epm endpoint key vrf <vrf-name> ip <ip-address>
```

The command syntax to manually clear all remote endpoints (both MAC and IP) in one VRF instance is shown here:

```
LEAF1# clear system internal epm endpoint vrf <vrf-name> remote
Ex. )
 LEAF1# clear system internal epm endpoint key vrf TK:VRF1 ip 192.168.2.2
 LEAF1# clear system internal epm endpoint key vrf TK:VRF1 remote
```

Note that when a device is migrated from Cisco ACI to outside Cisco ACI, you need to consider some additional actions, such as stopping the traffic long enough for remote endpoints to age out before the migration occurs or being ready to manually clear remote endpoints.

With the Enforce Subnet Check feature, this scenario can be prevented. However, for this feature to prevent this scenario, the bridge domain subnet configuration for IP2 needs to be removed, because this feature prevents a Cisco ACI leaf from learning endpoints only when the IP address does not belong to any of the bridge domain subnets in the same VRF instance. For more information, refer to the section Enforce Subnet Check option, later in this document.

Scenario 3: Stale remote endpoint example with L3Out outgoing traffic

Figure 8 shows an example of scenario 3 in Table 6. Note that the VRF instance in this example uses ingress policy enforcement mode. This particular behavior is observed only when first-generation leaf switches are the source of traffic.

Ingress policy enforcement mode for VRF changes the location at which the contract is applied to a packet that is sourced from a normal endpoint toward an L3Out connection (traffic from a nonborder leaf to a border leaf). Prior to this feature, egress policy enforcement mode was used. In that case, the contract for this packet flow always was applied on the border leaf (egress), where TCAM capacity for contracts could be a bottleneck. With ingress policy enforcement mode, the contract for this flow is applied on a nonborder leaf (ingress). Refer to the section "Policy Control Enforcement Direction" in the ACI Fabric L3Out Guide for details about ingress policy enforcement mode in VRF instances. Cisco Application Centric Infrastructure Fundamentals also discusses this mode, in the section "Layer 3 Out for Routed Connectivity to External Networks."

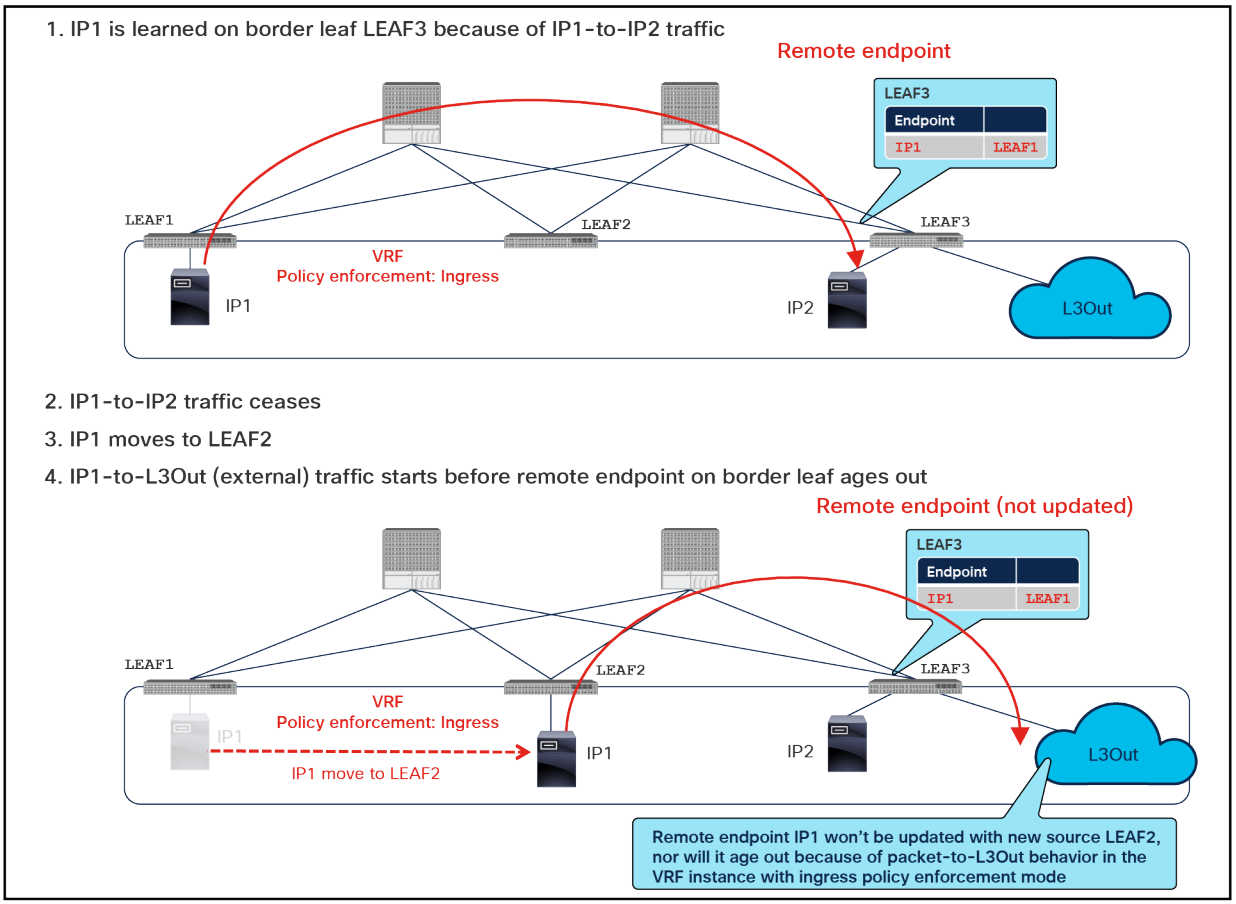


Figure 8.
Stale endpoint after endpoint move with VRF ingress enforcement mode (scenario 3)

At first, IP1 on LEAF1 is learned on border LEAF3 as a remote endpoint because of the traffic from IP1 to IP2. IP2 is a normal endpoint on border leaf LEAF3. If IP1 were sending traffic only to the external devices transiting the L3Out connection, instead of to IP2, this behavior would not create a remote endpoint for IP1 on LEAF3, because no source MAC or IP address is learned as a new remote endpoint by a packet to the L3Out connection (when the VRF mode is set to ingress policy enforcement).

After the remote endpoint is learned on LEAF3, a device with IP1 stops sending traffic to IP2 and moves to LEAF2. Next, if IP1 sends traffic toward the external devices transiting the L3Out connection, or if it starts sending traffic toward the L3Out connection before the old remote endpoint for IP1 on LEAF3 is aged out, the old remote endpoint will not to be updated with the new source information (LEAF2), and the entry will not age out because of the particular behavior described for L3Out in scenario 3 in Table 6.

Because of this stale remote endpoint, any traffic from LEAF3 toward IP1 may fail because LEAF3 sends a packet to the wrong leaf. That traffic may not fail right after the endpoint is moved because a bounce entry on LEAF1 can redirect traffic toward IP1 to the correct LEAF2. However, the traffic will start to fail as soon as the bounce entry ages out on LEAF1.

This stale remote endpoint on LEAF3 needs to be manually cleared to resume proper communication. Refer to scenario 2 for the command syntax for clearing remote endpoints.

This scenario can be prevented if LEAF3 is a dedicated border leaf, without any computing resources. This scenario can also be prevented if the Disable Remote EP Learn feature is enabled the on the border leaf. Refer to the section "Disable Remote EP Learn option (on border leaf)" for details.

	Due to this specific behavior on 1st generation leaf switches, it is generally considered a best practice to have a dedicated border leaf when there are 1st generation leaf switches in the fabric.

Scenario 4: Endpoint learning with second-generation leaf switch and L3Out

Figure 9 shows the benefits of second-generation leaf endpoint learning, as mentioned in scenario 4 in Table 6.

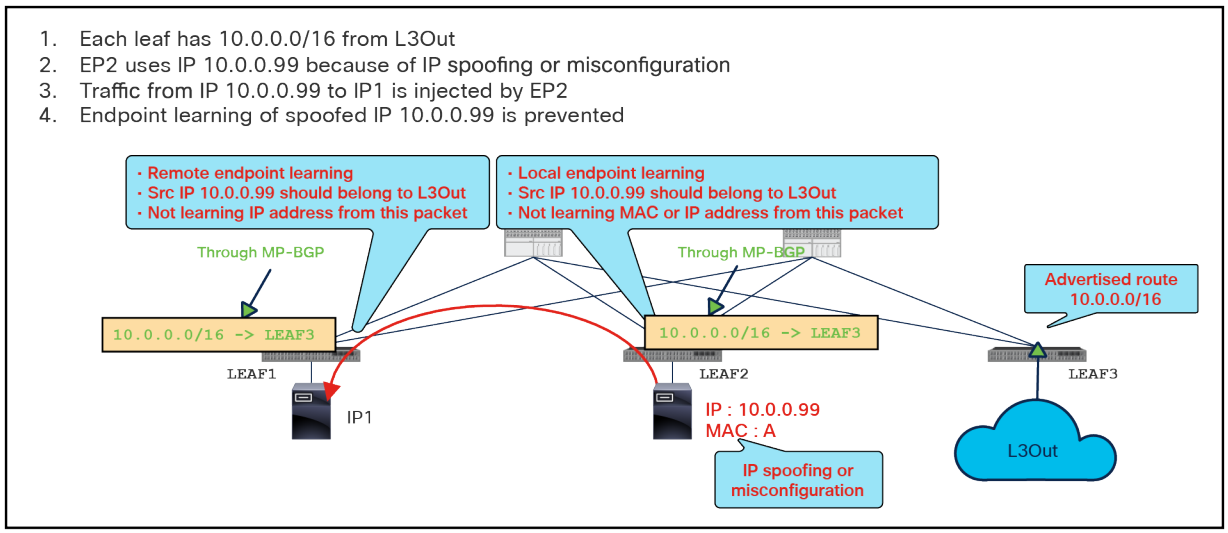


Figure 9.
Second-generation leaf switch benefits from limiting unnecessary endpoint learning (scenario 4)

In this example, the Cisco ACI fabric is receiving 10.0.0.0/16 routes from an external router through the L3Out connection on LEAF3. This route is redistributed to LEAF1 and LEAF2 through Multiprotocol Border Gateway Protocol (MP-BGP) in the Cisco ACI infrastructure network. However, because of a misconfiguration or an event such as IP spoofing, an endpoint on LEAF2 is sending packets with the source IP address 10.0.0.99, which should not exist in Cisco ACI, but should exist only behind the L3Out connection. Because of this spoofed traffic, LEAF2 will try to learn source MAC A and source IP address 10.0.0.99 as a local endpoint. Additionally, LEAF1 will try to learn the source IP address 10.0.0.99 as a remote endpoint because of Cisco ACI endpoint data-plane learning. If LEAF1 and LEAF2 are second-generation leaf switches, this learning (MAC/IP address local endpoint learning and IP address remote endpoint learning) would be prevented in this scenario, because the source IP address 10.0.0.99 is classified into routes learned from L3Out, which means that this IP address should not be local to Cisco ACI.

However, as also mentioned in Table 6, if the Cisco ACI fabric is receiving 0.0.0.0/0 route instead of 10.0.0.0/16 from an external router, this prevention mechanism will not be activated. And also, this prevention mechanism will not be activated if either the bridge domain subnets or the routes received from the L3Out connection doesn't cover 10.0.0.99.

Thus, although second-generation leaf switches provide a good built-in protection mechanism, you still should configure the Enforce Subnet Check feature. Refer to the "Enforce Subnet Check option" section later in this document for details.

Neither the built-in prevention mechanism for second-generation leaf switches nor the Enforce Subnet Check feature is available on first-generation leaf switches. Instead, you can configure the Limit IP Learning To Subnet option and the Disable Remote EP Learn option on the border leaf. Refer to the section discussing each feature to learn the differences between the features.

Advantages of Cisco ACI endpoint learning

The Cisco ACI endpoint learning capability provides efficient and scalable forwarding within the fabric. For example, with bounce entries and data-plane learning, no matter how many leaf switches the fabric contains, only three components need to be updated for endpoint move information. (For more information, refer to the section Endpoint movement and bounce entries earlier in this document.) Also, leaf switches don't have to consume their hardware resources to store information about all the endpoints on other leaf switches. Using data-plane learning, leaf switches consume resources to store only the necessary information for remote endpoints with which the leaf is actively communicating. The hardware resource savings are a huge advantage for scalable fabric.

This data-plane learning of remote endpoints, instead of relying on spine proxy (using the COOP database on spine switches) for all traffic, helps optimize traffic flow. For example, the remote endpoint learning reduces the traffic traversing the fabric by enabling the ingress leaf to enforce a contract policy and drop the packet if necessary before sending packets across the fabric. In the example in Figure 10, if a consumer leaf (LEAF1) does not know the destination endpoint (192.168.2.1) information, traffic goes to a provider leaf (LEAF2) based on spine proxy, and LEAF2 learns the source endpoint (192.168.1.1) information through data-plane learning. Then a contract policy is enforced on LEAF2, where the source and destination EPG information can be resolved. If the return traffic comes to LEAF2, a contract policy is enforced on LEAF2 that is an ingress leaf as opposed to an egress leaf (LEAF1), because LEAF2 already knows 192.168.1.1. Thus, unnecessary traffic is prevented from traversing the fabric if the traffic is denied by the contract on this ingress leaf (LEAF2).

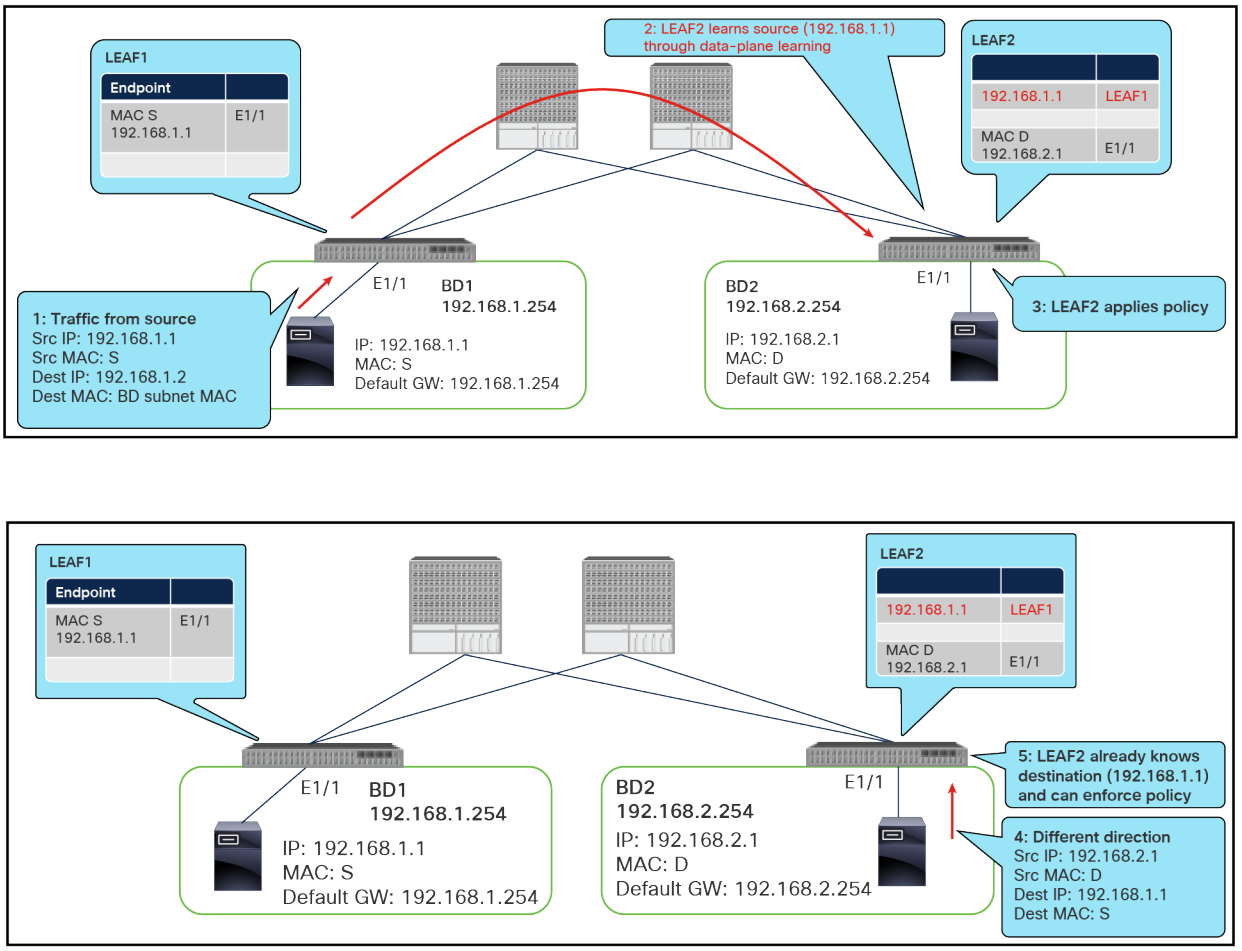


Figure 10.
Traffic flow example

Another advantage of endpoint learning through data-plane traffic is that it may help in scenarios in which the switch may have missed ARP control-plane packets previously originated by the endpoint or mistakenly cached out-of-date endpoint location information.

Despite the advantages mentioned here, in some specific scenarios you may need to disable the endpoint learning function. The rest of this document describes these use cases in greater detail.

Endpoint learning optimization options

A variety of configuration knobs are available to set options for Cisco ACI endpoint learning. This section describes endpoint learning related knobs for EPG, bridge domain, and fabric-wide configurations. Use cases for these knobs are also presented.

Table 1, at the beginning of this document, provides a summary of all the features discussed in this section.

EPG-level configuration options

This section discusses options that apply to EPGs.

L4-L7 Virtual IPs

The L4-L7 Virtual IPs option was introduced in Cisco Application Policy Infrastructure Controller (APIC) Release 1.2(1m). This option is located at Tenant > Application Profiles > Application EPGs (Figure 11). This option is used to disable data-plane IP learning for a particular IP address for direct server return, or DSR, use cases. By default, this feature is not enabled. The L4-L7 Virtual IPs option under an uSeg EPG is not supported. DSR Virtual IP address must be part of a base EPG that is not uSeg EPG.

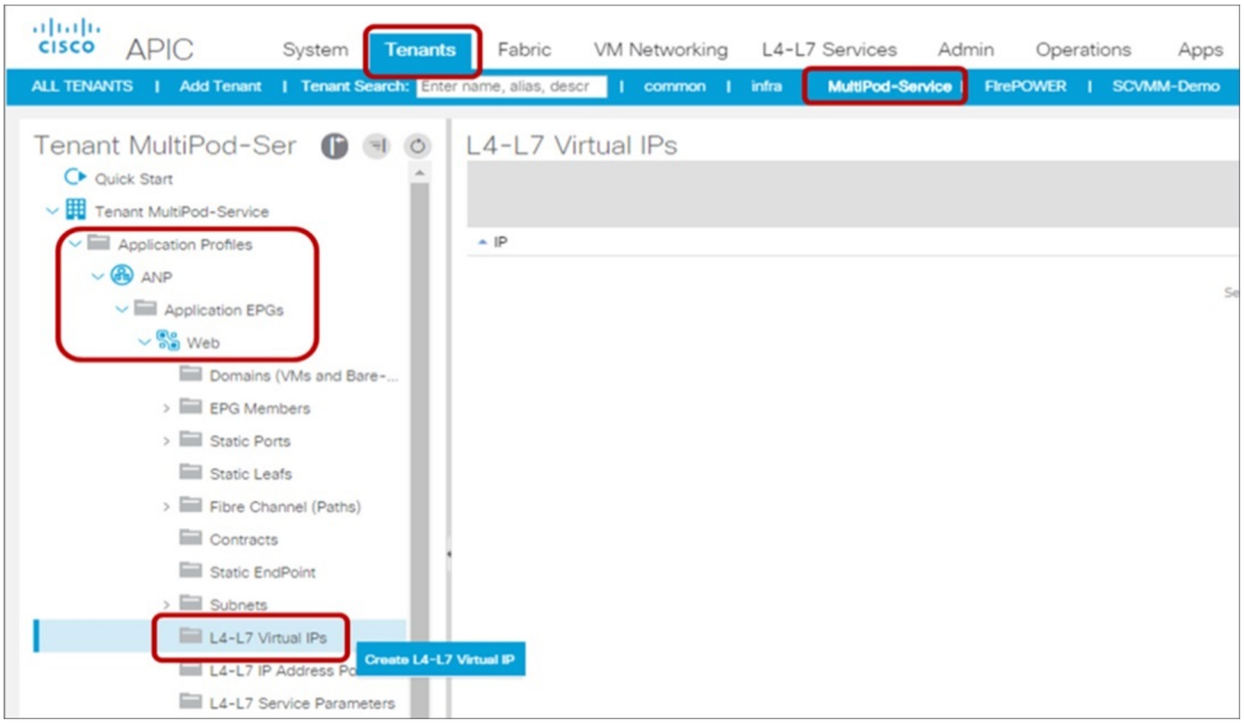


Figure 11.
L4-L7 Virtual IPs under Application EPGs

L4-L7 Virtual IPs use case

The only tested and supported use case for the L4-L7 Virtual IPs option is with Layer 2 DSR. The DSR option is deployed mainly when a large amount of return traffic is coming from a server. Typically, a load balancer is in the path between the client and the server: for both client-to-server incoming traffic and server-to-client returning traffic. If the amount of return traffic is large, the traffic will consume load-balancer resources, which will create a bottleneck. To help prevent this situation, in a DSR deployment return traffic directly goes back to the client without going through the load balancer.

In a DSR deployment, an ARP response must be suppressed on real servers. Only the load balancer is supposed to reply to ARP requests aiming to determine the MAC address associated to the virtual IP address, but real servers use the virtual IP address for server-to-client traffic. In a traditional network, this return traffic does not update IP information, but with Cisco ACI, the fabric learns the virtual IP address through data-plane IP learning, resulting in a problem.

By default, DSR does not work in Cisco ACI because of data-plane IP learning. This option disables data-plane IP learning for the specific DSR virtual IP address. Failure to disable IP learning for the DSR virtual IP address will result in IP endpoint flapping between different locations in the Cisco ACI fabric.

For example, as shown in Figure 12, 172.16.1.1 tries to connect to 192.168.1.100 (DSR virtual IP address), and the traffic goes to the load balancer because the load balancer has replied to an ARP request for 192.168.1.100. Next, the traffic is load-balanced to one of the real servers by rewriting the destination MAC address. Finally, server-to-client traffic is generated on the server. This return traffic uses 192.168.1.100 as the source IP address. The Cisco ACI fabric will learn 192.168.1.100 from different locations: from the load balancer and from real servers. Therefore, you need to prevent data-plane IP learning for the DSR virtual IP address.

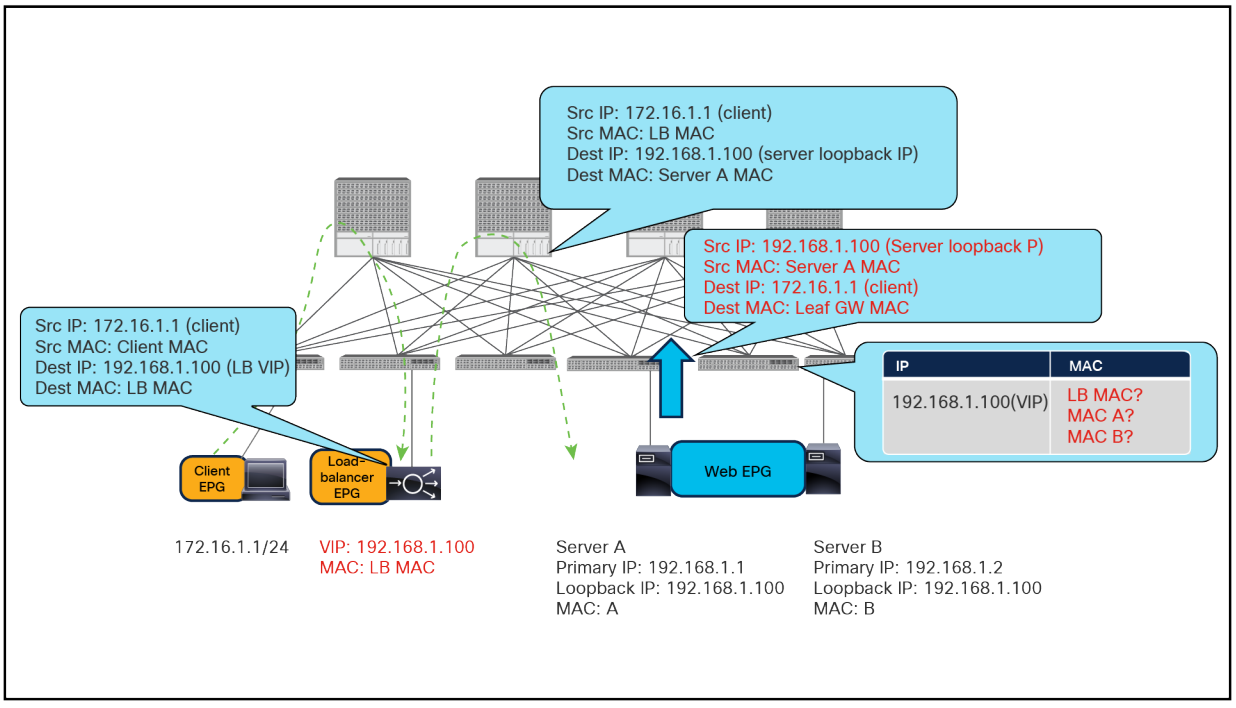


Figure 12.
Why you need to disable data-plane IP learning on the virtual IP address

With this feature enabled for the DSR virtual IP address (192.168.1.100 in this example), the Cisco ACI leaf will learn the IP address only from the control plane (ARP, GARP, or neighbor discovery) from the EPG with the DSR virtual IP address configured. Cisco ACI will also disable data-plane learning for the same IP address on related leaf switches.

The following paragraphs explain the scope of this DSR virtual IP address configuration, such as on which leaf is data-plane IP learning disabled.

The DSR configuration is downloaded to all the leaf switches on which the EPG with an L4-L7 virtual IP address is deployed, or on which an EPG with a contract with the EPG with the L4-L7 virtual IP address is deployed, regardless of the contract direction. For example, assume that you have a Client EPG, an LB EPG and a Web EPG and an L4-L7 virtual IP address configured under an LB EPG. The DSR virtual IP address configuration will be downloaded to LEAF1, LEAF2, and LEAF3, because LEAF2 has the LB EPG with the L4-L7 virtual IP address configured, and LEAF1 and LEAF3 have Web or Client EPGs that have contracts with the LB EPG (Figure 13).

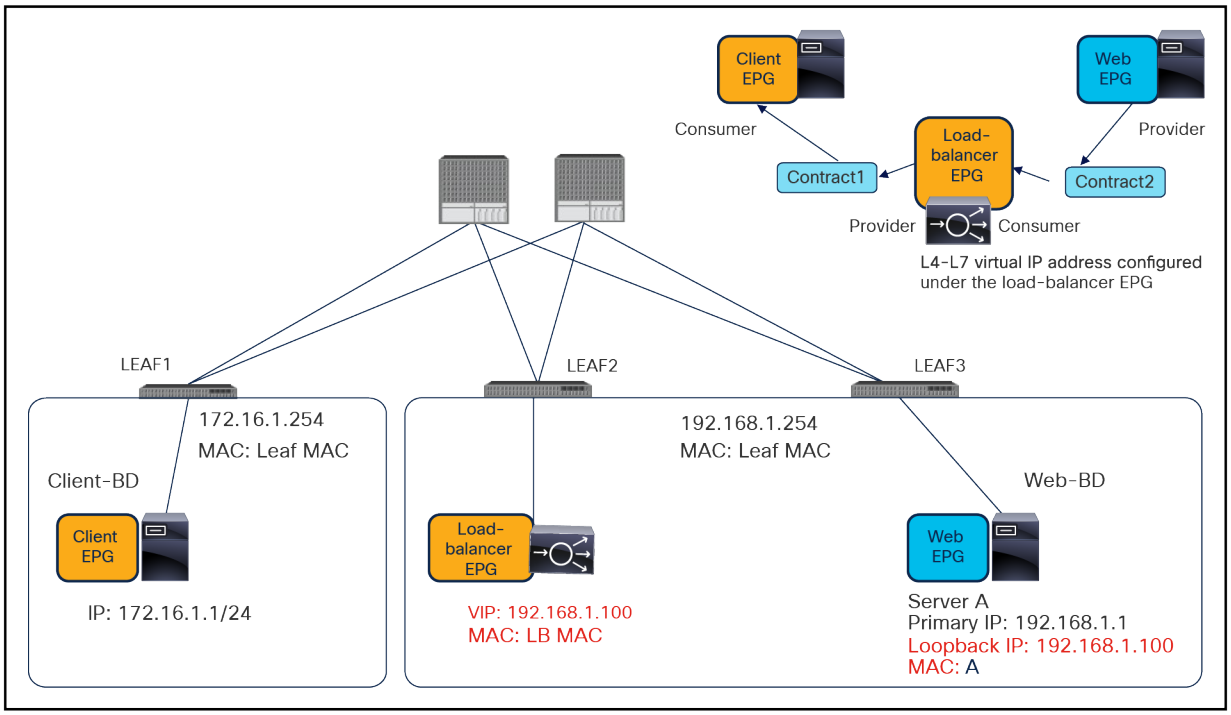


Figure 13.
Example of EPG relationships and configuration

All top-of-rack switches downloaded DSR configuration will not learn the L4-L7 virtual IP address from the data-path traffic, and they will not learn it from other EPGs, even though it's ARP, GARP, or neighbor discovery. For example, 192.168.1.100 is learned from the LB EPG through the control plane only. This behavior prevents situations in which an L4-L7 virtual IP address is learned mistakenly from a Web EPG.

For example, suppose that someone connected a web server classified to the Web EPG and forgot to suppress ARP. Even though ARP traffic is received, LEAF2 doesn't learn 192.168.1.100 from the Web EPG (Figure 14).

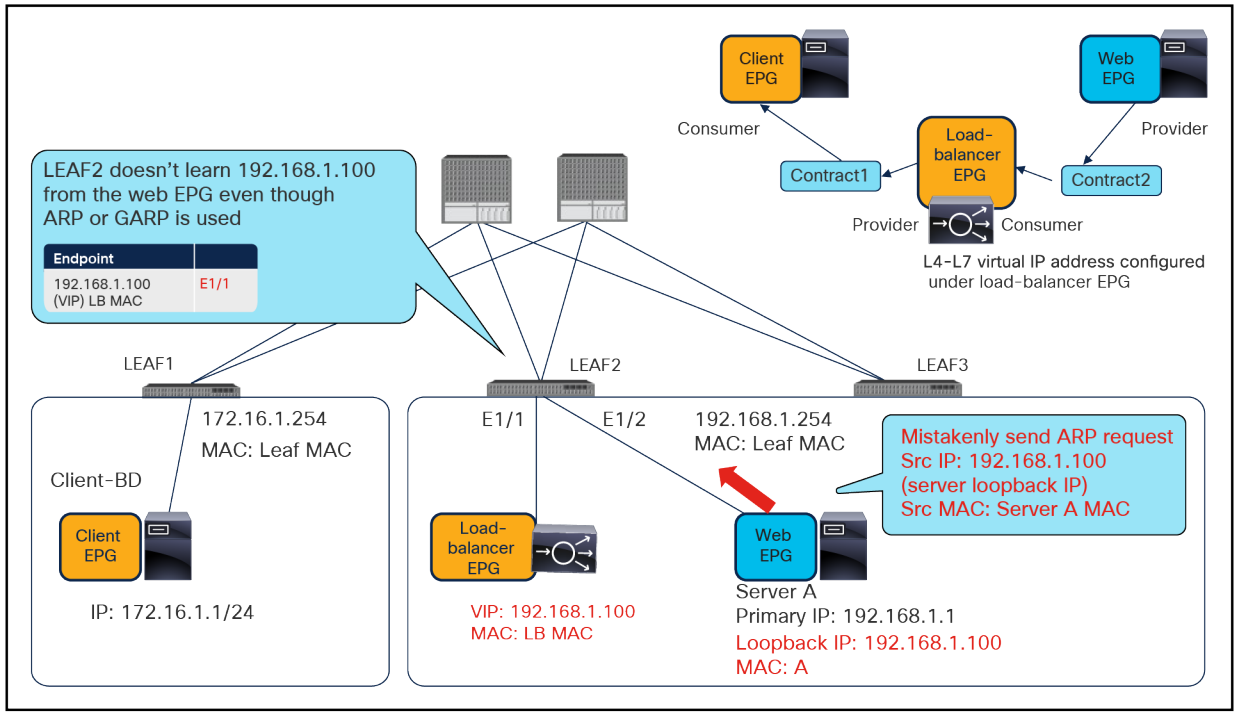


Figure 14.
Example of EPG relationships and configuration

Although DSR is described in the L4-L7 Service Deployment Guide, implementing the DSR configuration option doesn't require Cisco ACI service graph integration.

IP Data-plane Learning per host

There are three different scopes and configuration levels in IP data-plane learning as shown below:

- EPG Subnet – per host (/32 for IPv4 or /128 for IPv6)
- BD Subnet – per subnet
- VRF – per VRF

This section discusses the IP Data-plane Learning option that applies to an EPG subnet. The EPG level configuration is functionally equivalent to the per BD Subnet option, with the key difference that the EPG level configuration is meant for adding specific hosts for which IP Data-plane learning must be disabled, whereas the per BD Subnet option is for entering Subnets (for which IP- Data-plane learning must be disabled). For the option at a bridge domain subnet, please refer to IP Data-plane Learning subsection in Bridge domain-level configuration options. For the option at VRF, please refer to IP Data-plane Learning subsection in VRF-level configuration options.

The IP Data-plane Learning option at an EPG subnet was introduced in Cisco APIC Release 5.2(1g), which is to disable IP Data-plane Learning per host. This option is located at Tenant > Application Profiles > Application EPGs > Subnets (Figure 15). IP Data-plane learning is enabled by default. By using this option you can disable (or re-enable) endpoint data-plane IP learning for the host address (or addresses) that you have added under the EPG “subnet” configuration. The option can be set to “Disabled” under the following conditions:

- IP address subnet mask is /32 for IPv4 or /128 for IPv6
- “Type Behind Subnet” is “None” or “Anycast MAC’
- “No Default SVI Gateway” is checked

The BD to which this EPG configuration belongs must be set for unknown unicast flooding. This is because the ARP resolution for the specific hosts that you have configured would not work correctly otherwise. See the section L2 Unknown Unicast consideration, for details.

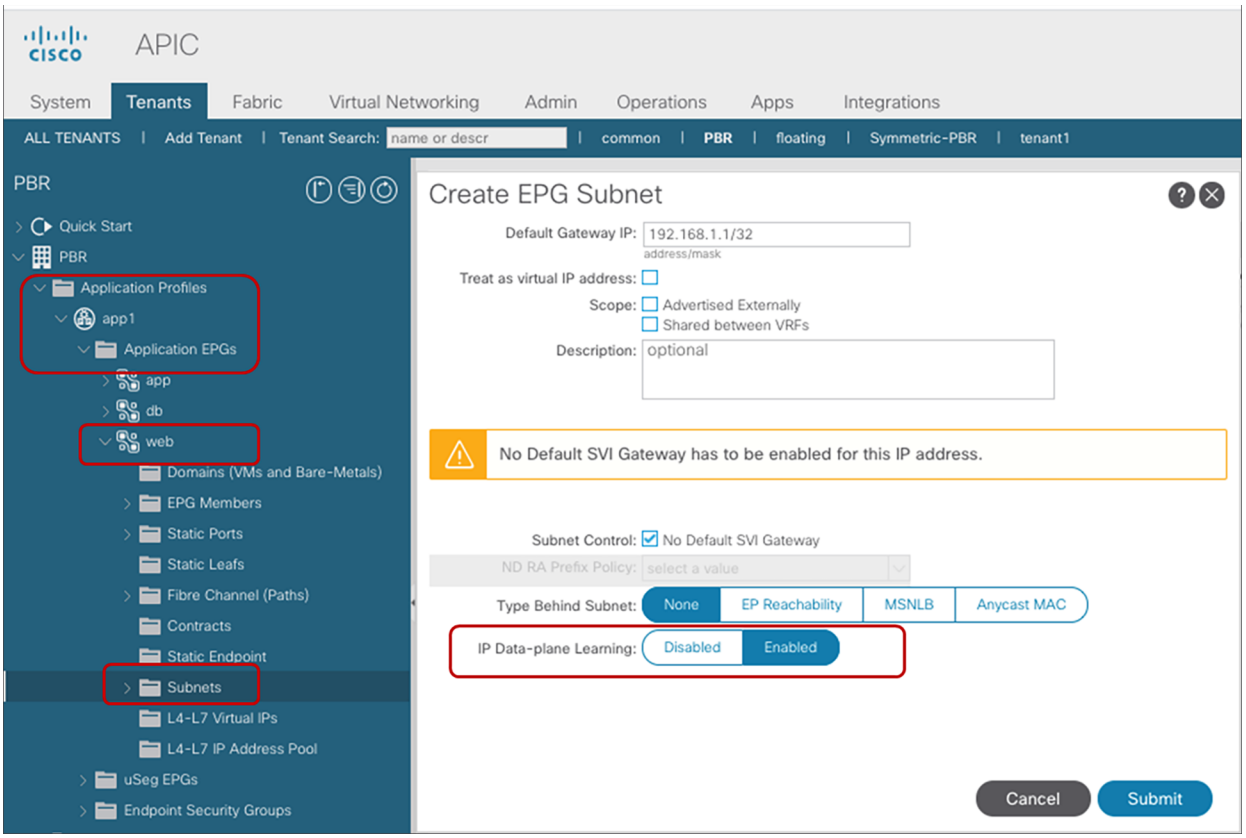


Figure 15.
Enable and disable Endpoint Data-plane Learning under the EPG subnet

For IP Data-plane Learning behavior, use cases, and considerations, please refer to Disabling IP Data-plane Learning: forwarding behavior and design considerations.

	GARP packets do not trigger endpoint learning (both MAC and IP) when this option is disabled. See the **GARP-based EP Move Detection** section for details.

	When the EPG is participating in the Multi-Site via Nexus Dashboard Orchestrator (NDO) such that the EPG is stretched or a shadow is created on the other sites, it is not supported to disable this option.

Bridge domain–level configuration options

This section discusses options that apply to bridge domains.

Unicast Routing

The **Unicast Routing** option has been implemented since the first release of Cisco ACI. It's located at Tenants > Networking > Bridge Domains in the APIC GUI (Figure 16).

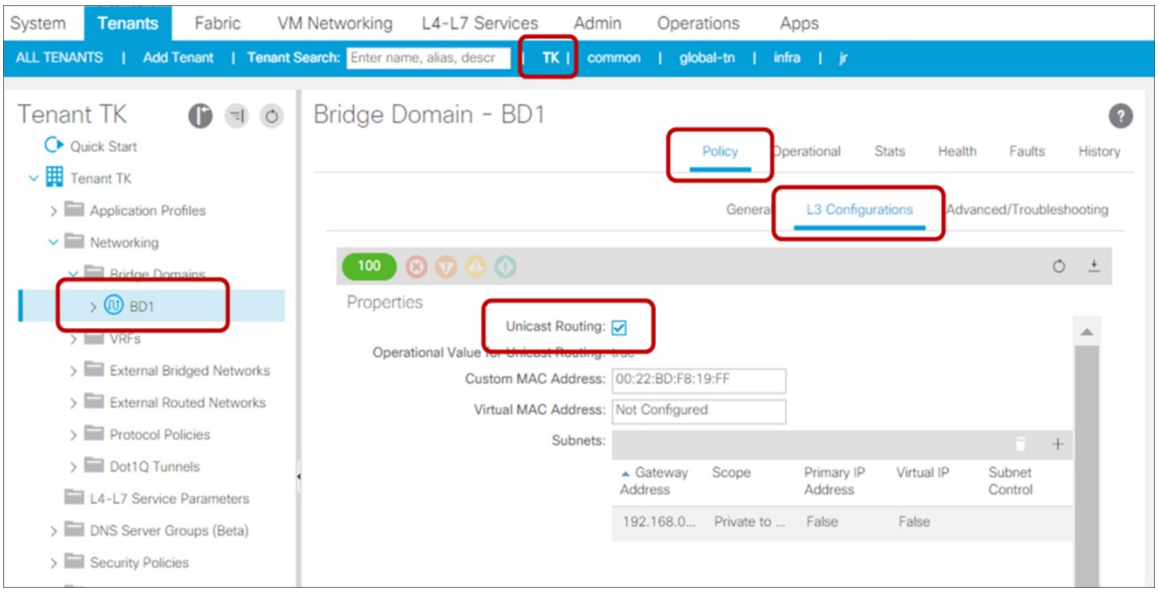


Figure 16.
Unicast Routing under the bridge domain

This feature enables IP unicast routing on the bridge domain. If this feature is not enabled, the subnets configured under the bridge domain are not pushed down to leaf switches, and routing will not occur. In addition, a bridge domain with **Unicast Routing** disabled will not learn any IP address as an endpoint. Thus, that bridge domain will be used only for Layer 2 communications, and endpoints in that bridge domain should have their default gateways outside Cisco ACI.

Unicast Routing enabled without subnet is not a recommended approach. If **Unicast Routing** is enabled without any bridge domain subnets configured, IP information in the bridge domain can still be learned through ARP in the data plane, but no routing will occur because there will be no SVI to perform routing on the bridge domain. Please refer to the following "Unicast Routing use case (disable for Layer 2 bridge domain)" section to understand the reasoning behind this recommendation. Also, be sure to enable **Enforce Subnet Check** to optimize the local and remote IP endpoint learnings based on the subnets configured under each bridge domain.

When **Unicast Routing** gets disabled, both MAC and IP endpoint information are flushed for the BD. This behavior has been enhanced to flush only IP information based on the subnets configured under the BD from the Cisco APIC Release 3.1(1i). This change was introduced through this enhancement:

CSCvd92811: L2 endpoints getting flushed when switching BD from routing to switching

When a BD subnet is deleted while **Unicast Routing** is still enabled, the flush of endpoints is also performed for the corresponding IP information.

The flushing is performed for both local and remote IP endpoints on leaf nodes where the corresponding BD SVI is deployed. Starting from the Cisco ACI switch 14.0(1) release, remote IP endpoints on leaf nodes that do not have the corresponding BD SVI are also flushed through endpoint announce messages.

Unicast Routing use case (disable for Layer 2 bridge domain)

This use case demonstrates **why unicast routing should be disabled** when a bridge domain is supposed to perform only Layer 2 switching (For example, when an endpoint's default gateway is outside Cisco ACI). A bridge domain with this configuration is referred to as a Layer 2 Bridge Domain (L2BD). Figures 17, 18, and 19 show what happens when the Unicast Routing option is not disabled on an L2BD. In this example, BD1, BD2, and the L3Out connection are in the same VRF instance.

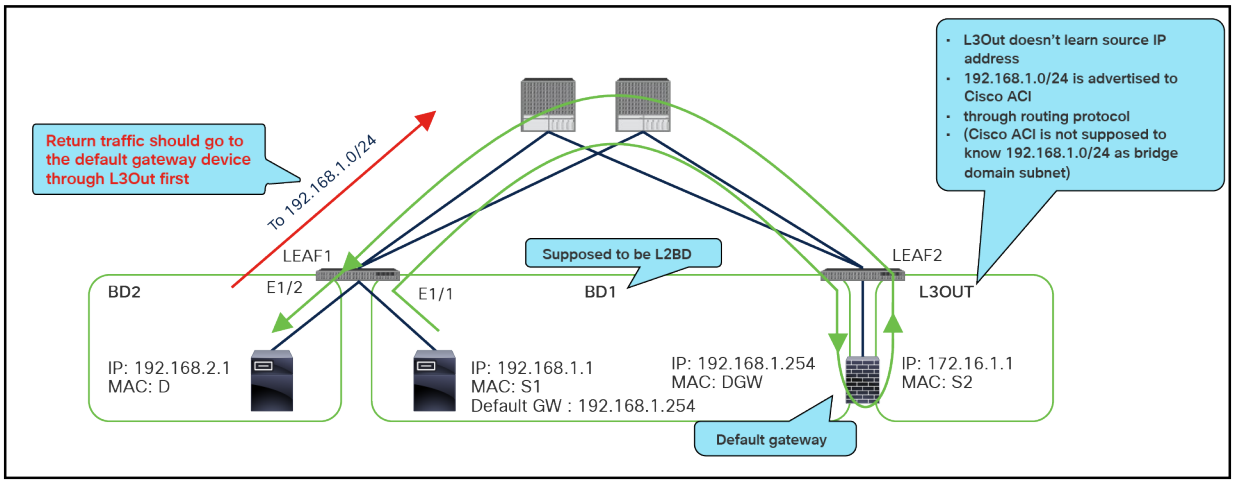


Figure 17.
Why you need to disable Unicast Routing for L2BD (part 1: expected flow)

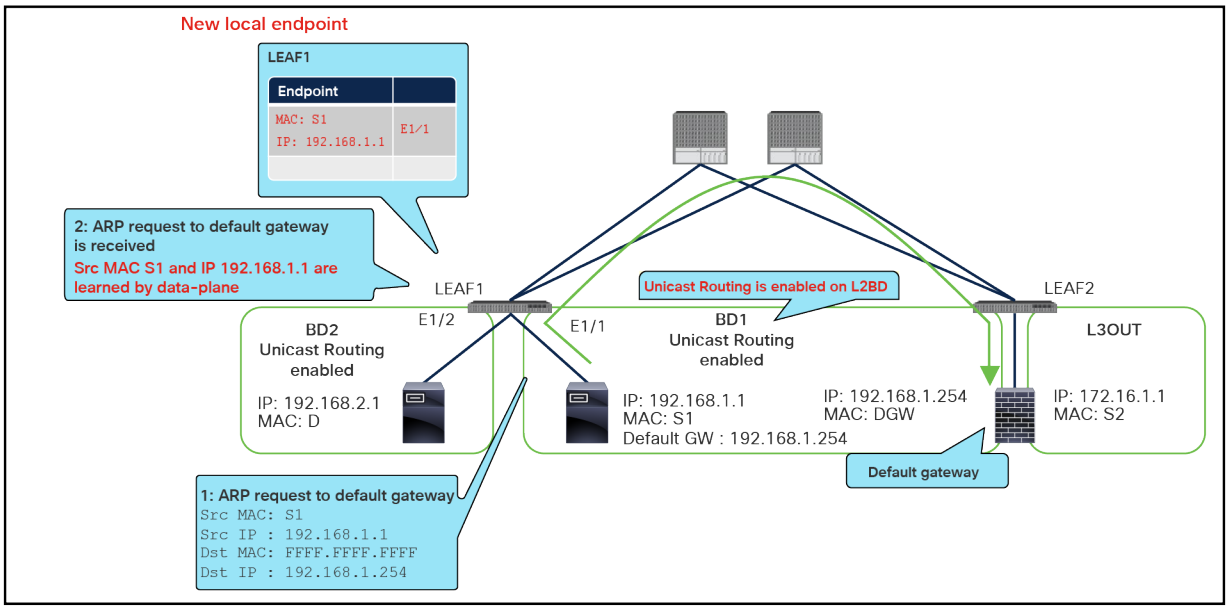


Figure 18.
Why you need to disable Unicast Routing for L2BD (part 2: IP learning on L2BD)

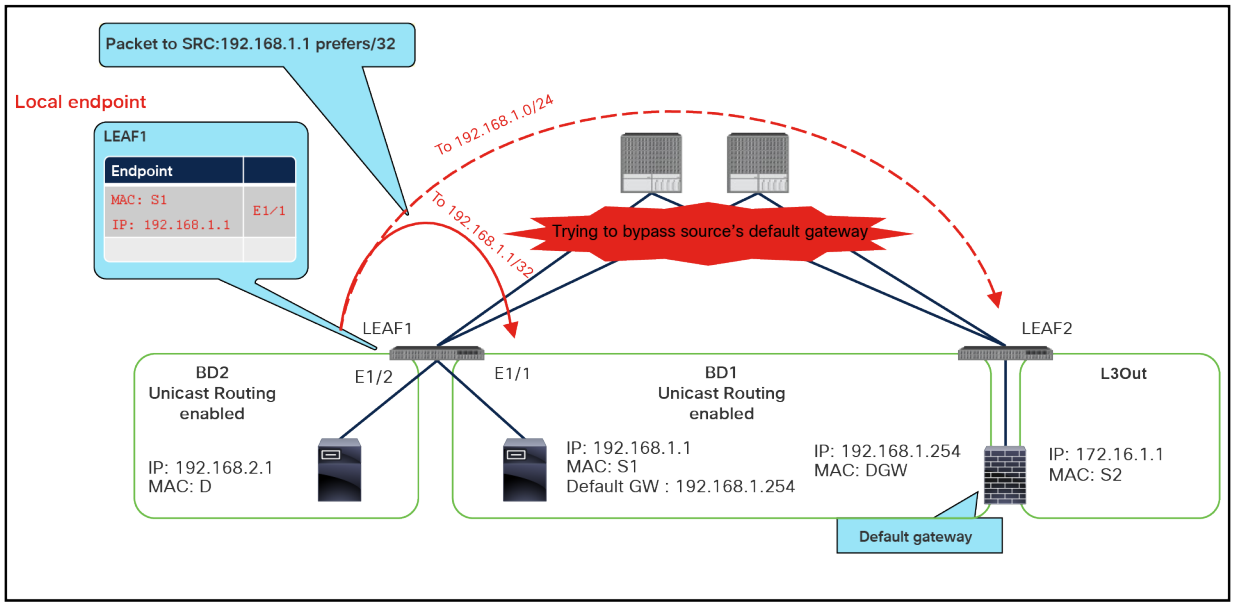


Figure 19.
Why you need to disable Unicast Routing for L2BD (part 3: problems with IP learning on L2BD)

In this example, BD1 is supposed to be L2BD. Figure 16 shows the expected traffic flow, which is through the default gateway device whenever the endpoint in BD1 (IP 192.168.1.1) talks with a device outside its subnet. However, if unicast routing is not disabled on BD1, as shown in the figures, LEAF1 learns IP 192.168.1.1 from an ARP request (Figure 17). As a result of learning the IP of this new endpoint (192.168.1.1), the traffic to 192.168.1.1 from the destination device (192.168.2.1) is trying to go directly to the actual source device bypassing the source's default gateway: for example, a firewall (Figure 18). In this scenario LEAF1 should never learn IP 192.168.1.1 from the actual host device. The traffic to 192.168.1.1 should go to the gateway device first, and the gateway device should forward the return traffic to MAC S1 (the source).

If unicast routing is disabled on BD1, which performs only Layer 2 forwarding, LEAF1 will never learn any IP address under BD1, as shown in Figure 20.

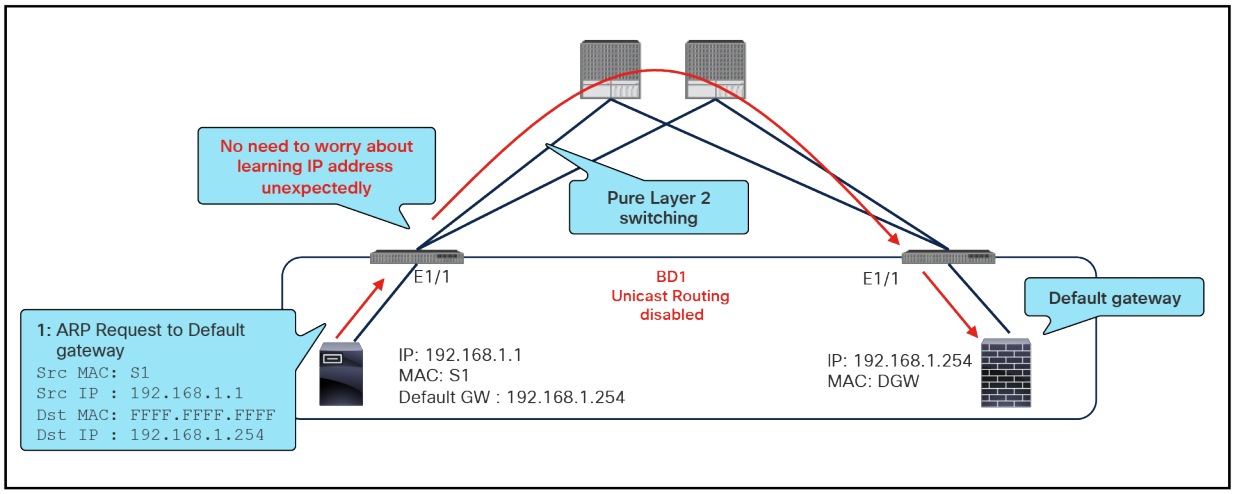


Figure 20.
Use case with Unicast Routing disabled

Figure 21 shows a conceptual image of an L2BD in which the Unicast Routing option is disabled. Because there is no routing or IP learning on this bridge domain, this L2BD is closed within its bridge domain, even though it belongs to the VRF instance. It thus could be described as being isolated from other forwarding domains within the same VRF instance.

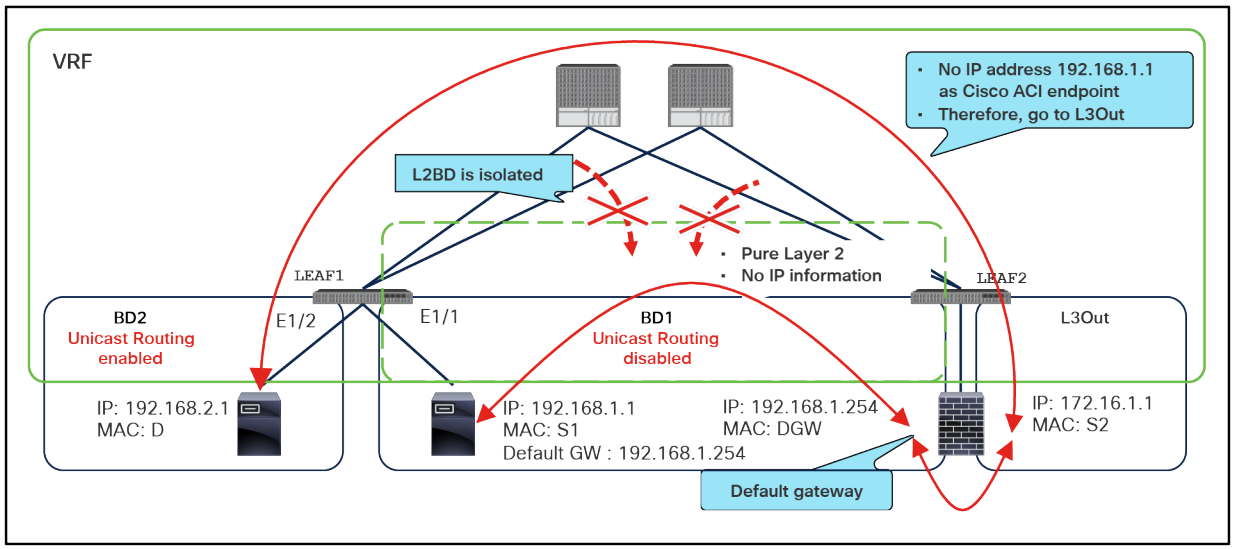


Figure 21.
L2BD concept

GARP-based EP Move Detection Mode

GARP-based EP Move Detection is a feature to address corner cases with specific conditions and hardware.

This option is located at **Tenant > Networking > Bridge Domain** (Figure 22). This option is disabled by default.

	**ARP Flooding** and **Unicast Routing** must be enabled as well for **GARP-based EP Move Detection** to take effect.

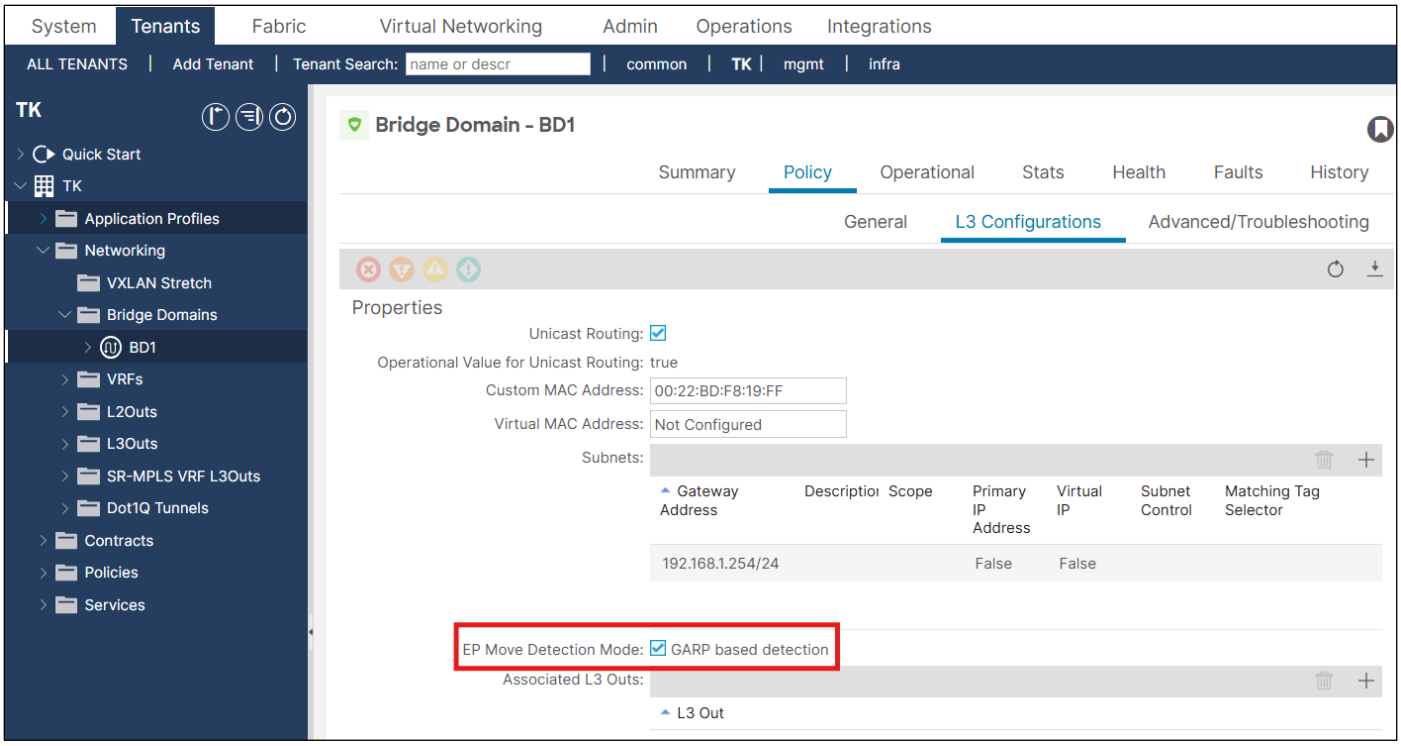


Figure 22.
EP Move Detection Mode under BD L3 configuration

The followings are the two corner cases where **GARP-based EP Move Detection** should be enabled.

On ACI gen1 switches, there is a limitation in which IP data-plane learning does not happen when two packets with the same source IP address and different MAC addresses are received on the same EPG on the same interface of an ACI switch. Such can happen when the IP is a virtual IP (VIP) and the owner of the VIP moved from one MAC address to another due to a failover. You need to enable **GARP-based EP Move Detection** to ensure that such IP address movements are recognized by gen 1 switches. This was the original issue (CSCus77627) that introduced this feature in APIC Release 1.1(1j).

On ACI gen2 or newer switches, the limitation in gen1 switches (CSCus77627) is not applicable. You do not need **GARP-based EP Move Detection** for the specific scenario mentioned above. However, on gen2 or newer switches, GARP packets do not trigger endpoint learning (both MAC and IP) when you explicitly disabled **IP Data-plane Learning per subnet**, **per host**, or when IP Data-plane Learning is implicitly disabled through Service Graph PBR (Policy Based Redirect). In such a case, you need to enable **GARP-based EP Move Detection** so that GARP packets trigger endpoint learning just as they do when **IP Data-plane Learning per subnet/per host** is not disabled.

GARP-based EP Move Detection Mode use case for gen 1 switches

As mentioned above, gen 1 switches have a limitation in which IP data-plane learning does not happen when two packets with the same source IP address and different MAC addresses are received on the same interface and the same EPG on an ACI switch. Specifically speaking, the IP learning happens with the first packet. However, the second packet with the different MAC address does not update the IP-to-MAC mapping with the new MAC address. With the example of the VIP failover mentioned above, this results in the VIP to be tied to the old MAC address on the ACI endpoint table even after the VIP ownership moved to the new MAC address on the connected device.

Consider the scenario in Figure 23. The figure shows a single VMware ESXi host attached to the Cisco ACI fabric and multiple virtual machines residing in the same EPG.

GARP-based EP Move Detection Mode use case

Consider the scenario in Figure 23. The figure shows a single VMware ESXi host attached to the Cisco ACI fabric and multiple virtual machines residing in the same EPG.

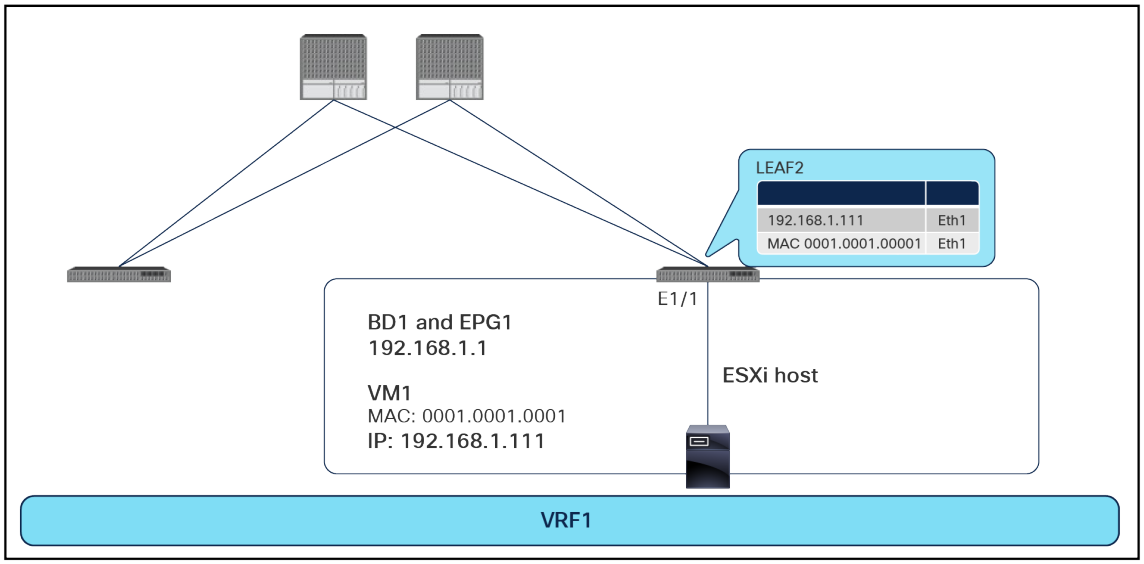


Figure 23.
Same interface and same EPG: VM1

A problem occurs when VM1 is powered off and VM2 is powered on. VM2 acquires the same IP address that previously belonged to VM1 (Figure 24). If GARP-based detection is not enabled, Cisco ACI will not update the endpoint entry for VM1 and will erroneously send traffic to the old MAC address entry that belonged to VM1. This is because we need to enable GARP-based detection in cases in which IP to MAC movement occurs on the same interface and same EPG.

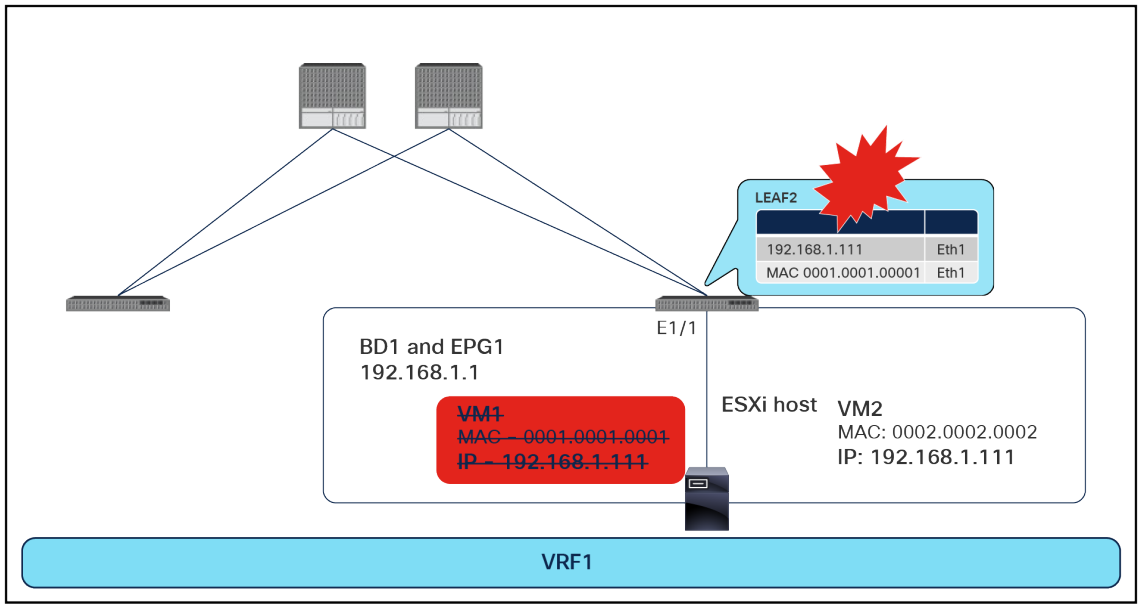


Figure 24.
Same interface and same EPG: VM2

Limit IP Learning To Subnet

The Limit IP Learning To Subnet option was originally called Enforce Subnet Check for IP Learning. It was introduced in APIC Release 1.1(1j) release with the following enhancement:

CSCuu09759: Add a configuration knob to enable/disable BD Subnet check for IP learn

It is located at Tenant > Networking > Bridge Domain (Figure 25).

Beginning with APIC Releases 2.3(1e) and 3.0(1k), this option is enabled by default with the following enhancement:

CSCvb16668: Enforce Subnet Check should be enabled by default

Prior to these releases, this option was disabled by default.

If this option is enabled, the fabric will learn only IP addresses for subnets configured on the bridge domain.

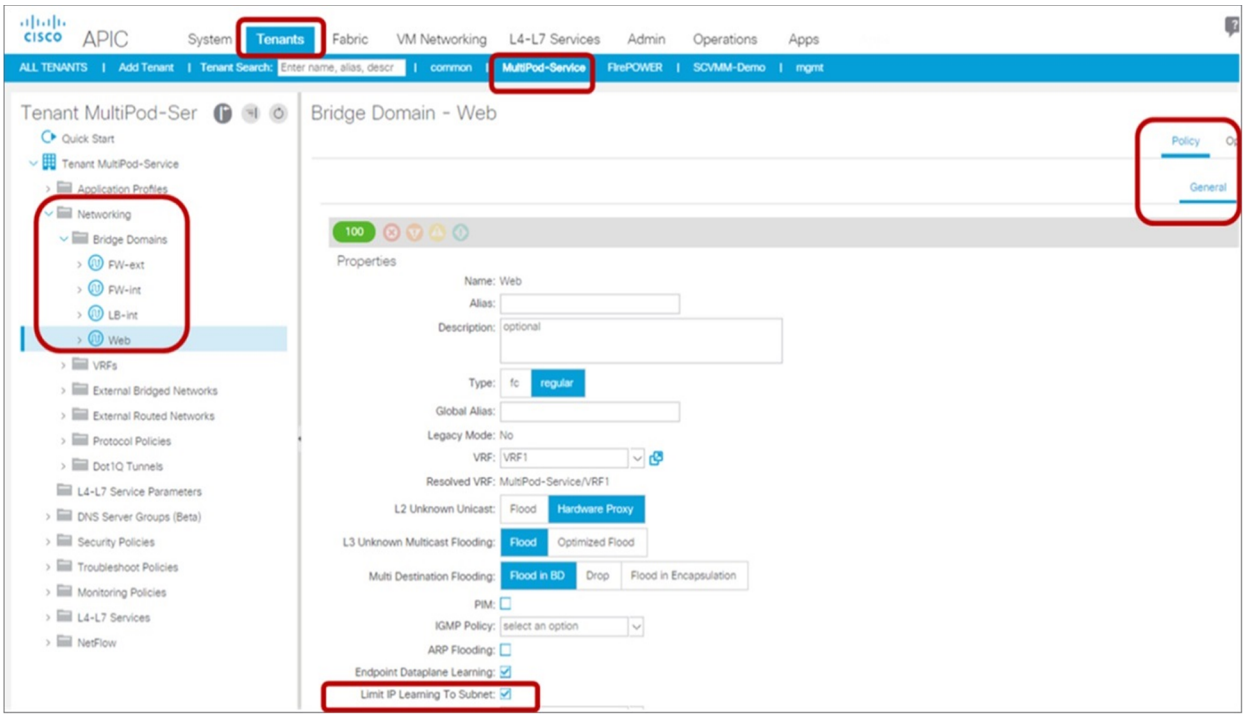


Figure 25.
Limit IP Learning To Subnet under bridge domain

Prior to Cisco ACI Release 3.0(1k), if this option is disabled or enabled on a bridge domain that was already configured, the following happens:

- Cisco ACI flushes all endpoint IP addresses learned on the bridge domain.
- Cisco ACI pauses MAC and IP address learning for 120 seconds.

This behavior has been improved from Cisco ACI Release 3.0(1k) by the enhancement, CSCve29663. From 3.0(1k), if this option is enabled on a bridge domain that had the option disabled, the following happens:

- Cisco ACI doesn't flush endpoint IP addresses that belong to the subnet. (Endpoint IP addresses that do not belong to the bridge domain subnet are flushed.)
- MAC or IP address learning are not paused for 120 seconds.

If this option is disabled on a bridge domain that had the option enabled, the following happens:

- Cisco ACI doesn't flush endpoint IP addresses learned on the bridge domain.
- MAC or IP address learning is not paused for 120 seconds.

	Prior to Cisco ACI Release 3.0(1k), if the Limit IP Learning To Subnet option was enabled when the bridge domain was configured for unicast routing, you could experience an impact of 120 seconds as the bridge domain endpoint table is flushed and endpoint learning for IP addresses would be paused (for 120 seconds).
	When migrating Layer 3 gateway (L3GW) connectivity to Cisco ACI, you can mitigate this impact by enabling the Limit IP Learning To Subnet option when the bridge domain is configured as a Layer 2–only bridge domain. After you have enabled the option, wait 120 seconds for the timer to expire. Then enable the Unicast Routing option. Because you are not learning IP endpoints on the bridge domain (because it is an L2BD), the 120-second timer will not affect the learning of new MAC-based endpoints.

From the leaf, run the command **vsh -c 'show system internal epm vlan vlan-id detail'** and look for the Learn Enable option. This option should be set to Yes (Figure 26).

```
leaf-201# vsh -c 'show system internal epm vlan 126 detail'
Warning: could not get list of reserved vlans

VLAN 126
VLAN type : FD vlan
hw id : 128 ::: sclass : 16389
access enc : (802.1Q, 1001)
fabric enc : (VXLAN, 8692)
Object store EP db version : 0
BD vlan id : 80 ::: BD vnid : 15040476 ::: VRF vnid : 2162689
Valid : Yes ::: Incomplete : No  ::: Learn Enable : Yes
pol_ctrl_flags:
Endpoint count : 0 ::: Local Endpoint count : 0
::::
```

Figure 26.
Checking that the Limit IP Learning To Subnet option is enabled

Limit IP Learning To Subnet use case

If a bridge domain is configured with a subnet address of 192.168.1.254/24, the fabric does not learn a local endpoint IP address, such as 192.168.2.1/24, that is outside this range. This behavior prevents unnecessary IP learning, as shown in Figure 27, which shows endpoints with the wrong IP address configured.

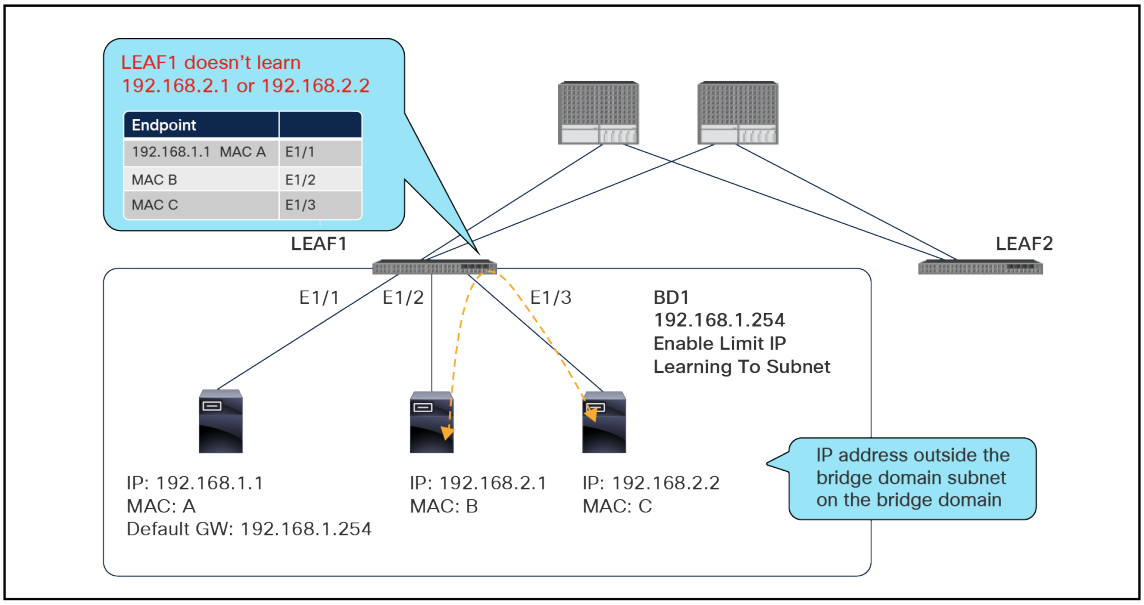


Figure 27.
Limit IP Learning To Subnet

Although this feature prevents local IP learning, the local leaf still learns the MAC address, and the remote leaf still learns the IP and MAC addresses (although the local leaf does not learn the IP address, it does not drop the packet). For example, LEAF1 doesn't learn 192.168.2.1, but it learns MAC B, and LEAF2 learns 192.168.2.1 and MAC B (Figure 28).

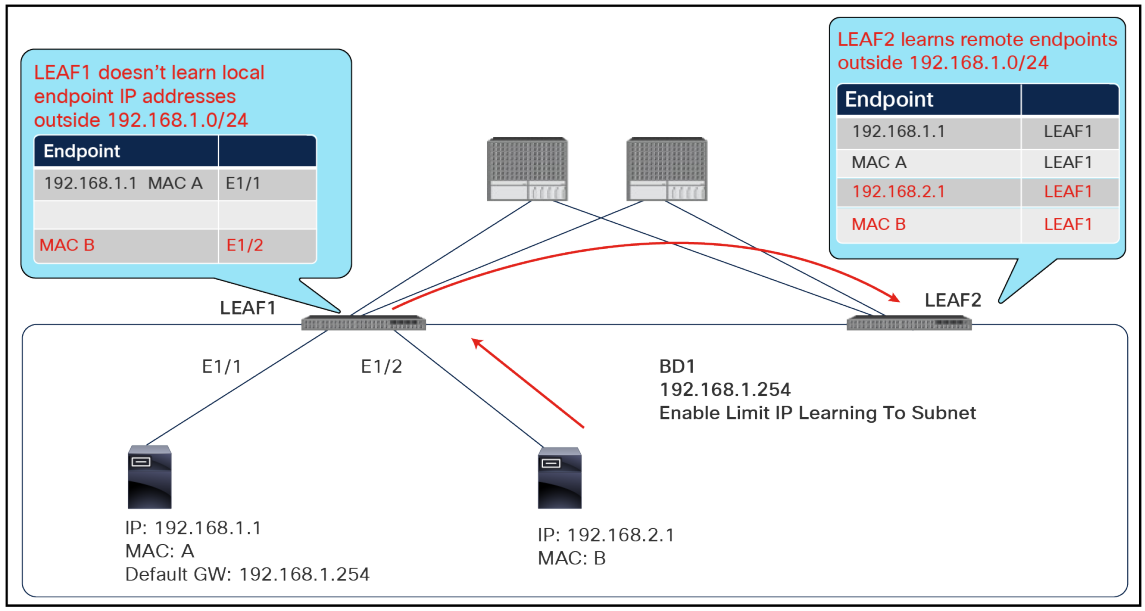


Figure 28.
Limit IP Learning To Subnet (remote IP learning)

Endpoint Dataplane Learning

The Endpoint Dataplane Learning option was introduced in Cisco APIC Release 2.0(1m). It is located at Tenant > Networking > Bridge Domain (Figure 29). Starting from APIC Release 5.0(1), this option is moved under the "Advanced/Troubleshooting" tab under the Policy tab at a bride domain.

This option is enabled by default; it enables and disables endpoint data-plane IP learning.

At the time of this writing (Cisco ACI Release 3.0(1k)), the only tested and supported use case for this option is in conjunction with service graphs with policy-based redirect, or PBR.

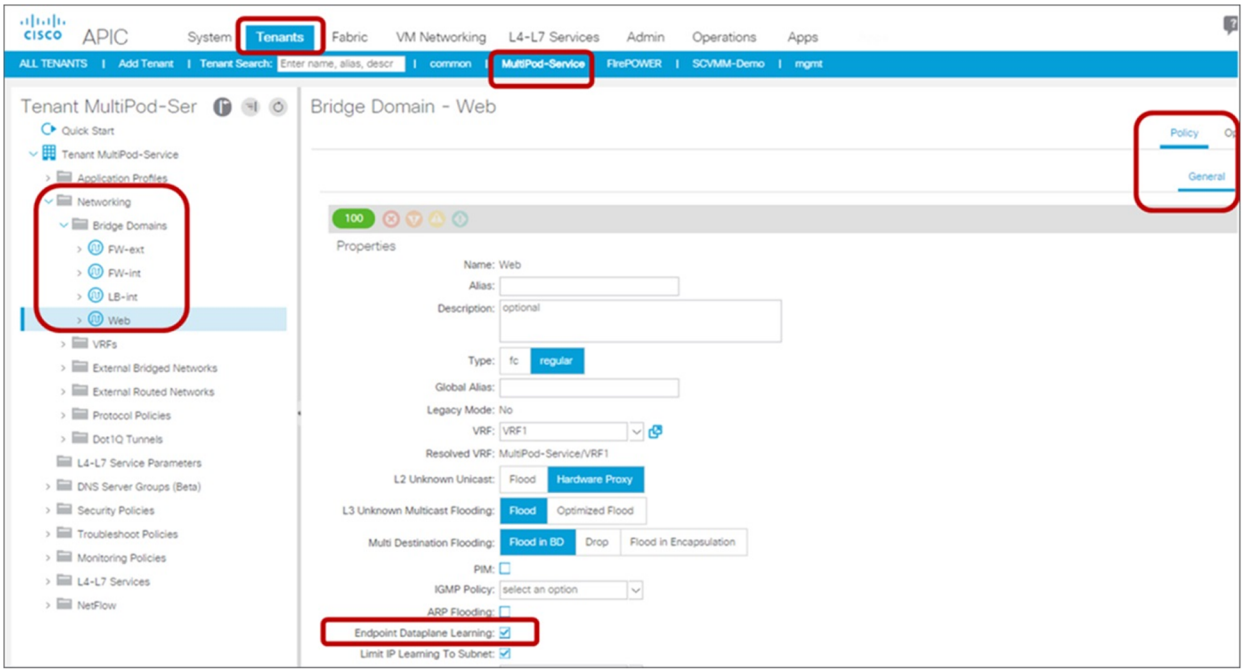


Figure 29.
Enable and disable Endpoint Dataplane Learning under the bridge domain

Note that if you disable endpoint data-plane learning, by unchecking the Endpoint Data-plane Learning option, the Limit IP Learning To Subnet option will not appear in the APIC (Figure 30). The Limit IP Learning To Subnet option is not available because IP learning on remote and local leaf switches is already disabled. Thus, as long as you disable the Endpoint Dataplane Learning option, the service leaf doesn't learn 192.168.1.1 from the Svc-internal-bridge domain in PBR example shown in the figure.

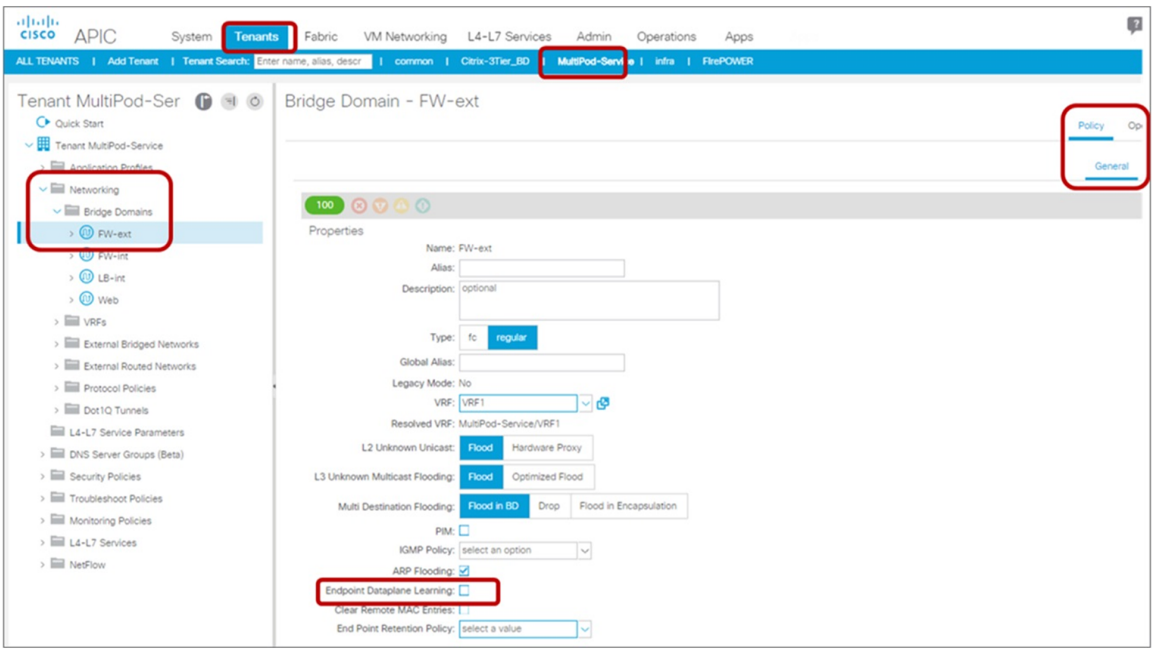


Figure 30.
Disable Endpoint Dataplane Learning and hide the Limit IP Learning To Subnet option

Disable Endpoint Dataplane Learning for PBR use case

With APIC Release 3.0 or earlier, the Endpoint Dataplane Learning option, under the bridge domain, must be disabled when that bridge domain is connected to a service graph device using the PBR feature. The service graph device with the PBR feature is typically called a PBR node. Figure 31 shows an example. This example shows bidirectional PBR with a PBR node, a firewall, inserted between the Client and Web EPGs.

Starting from APIC Release 3.1, disabling the Endpoint Dataplane Learning in the PBR node bridge domain is not mandatory if it's second-generation leaf switch. The Endpoint Dataplane Learning setting on the PBR node EPG is automatically disabled during service graph instantiation.

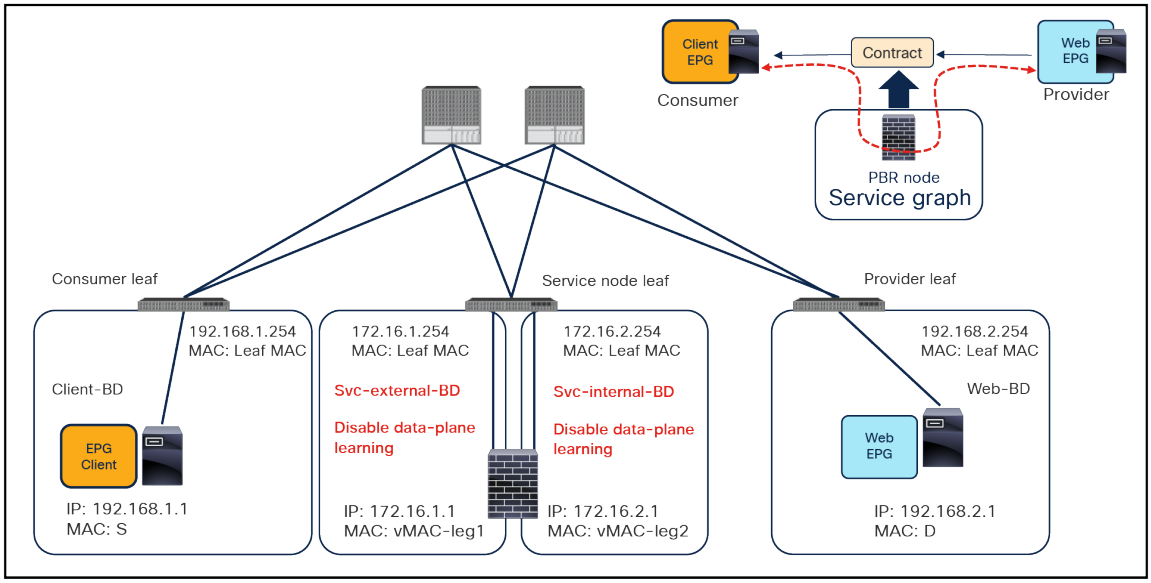


Figure 31.
Disable Endpoint Dataplane Learning (PBR use case)

You must disable the Endpoint Dataplane Learning option for a service graph with PBR because leaf switches involved in the PBR traffic flow could otherwise experience unwanted endpoint learning behavior if this option is left as enabled on the bridge domains for the PBR node.

For example, as shown in Figure 32, the source IP address of traffic coming back from the PBR node is still 192.168.1.1 even after PBR is enforced, so the provider leaf will receive packets with inner source IP address 192.168.1.1 and the outer source IP address as the service node leaf VTEP. Thus, the provider leaf will learn 192.168.1.1 through the service node leaf VTEP IP, even though 192.168.1.1 is actually under a different leaf.

If you disable data-plane learning on Svc-internal-BD, the bridge domain for the provider side of PBR node, the provider leaf doesn't learn 192.168.1.1 through the traffic from the PBR node.

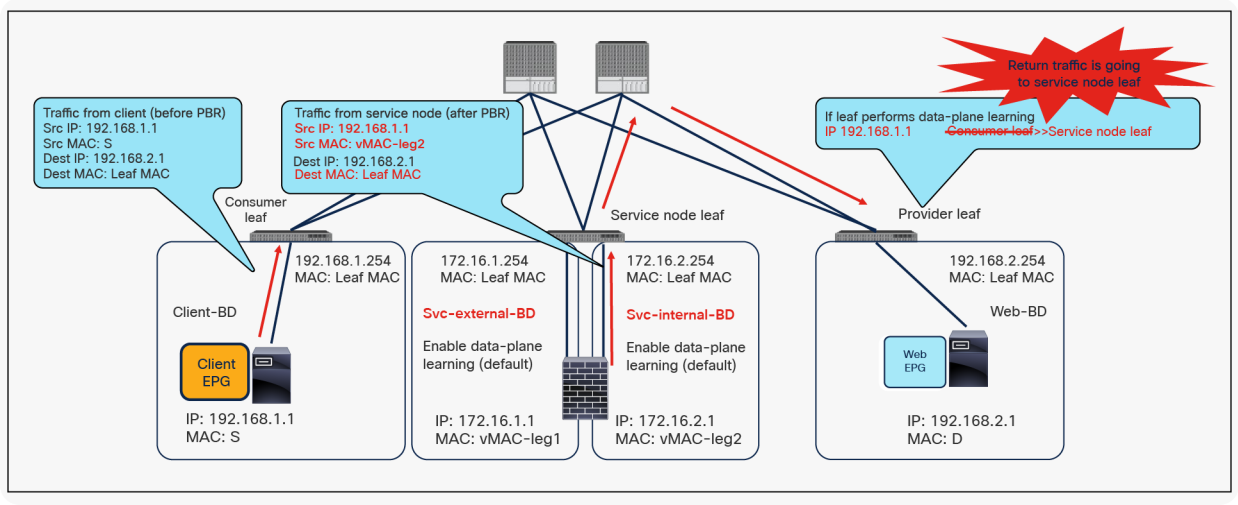
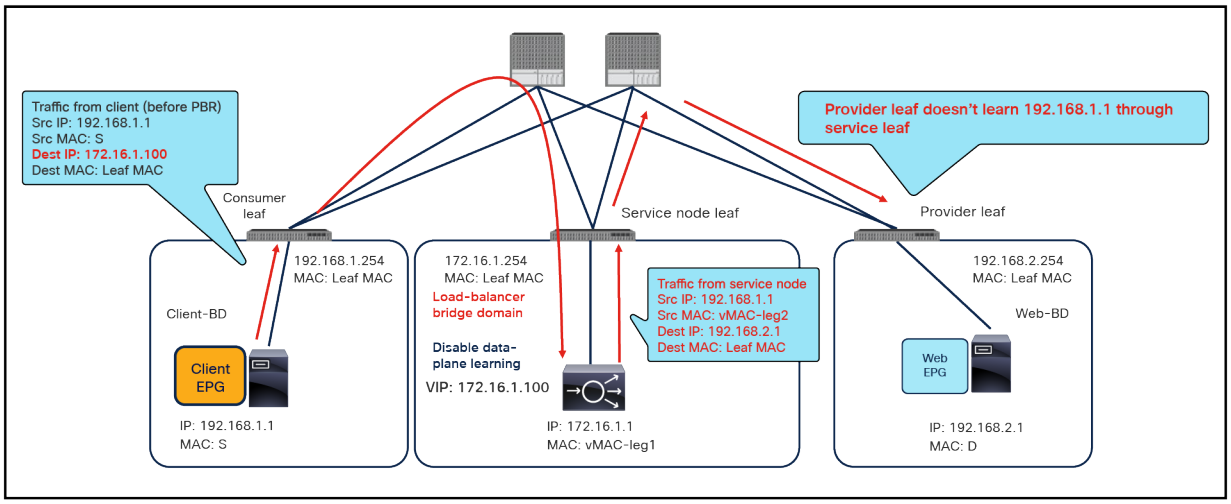


Figure 32.
Why you need to disable data-plane learning on the PBR node bridge domain

To maintain symmetric traffic, PBR for the return traffic is also required in this example. The Endpoint Dataplane Learning option needs to be set to Disabled on Svc-external-BD as well to prevent the consumer leaf switches from learning 192.168.2.1 through the service leaf after PBR is enforced.

In addition to bidirectional PBR, you can use unidirectional PBR: for instance, in the case of PBR for return traffic in load-balancer integration.

For example, as shown in Figure 33, because the destination IP address from the client is the virtual IP address on the load balancer, PBR is not required for client-to-web traffic. If the load balancer doesn't translate the source IP address, PBR for return traffic is required; otherwise, return traffic won't come back to the load balancer. You must disable data-plane learning on the Load-balancer-BD to which the load balancer and PBR node are connected, so that the provider leaf doesn't learn 192.168.1.1 through the service node leaf.



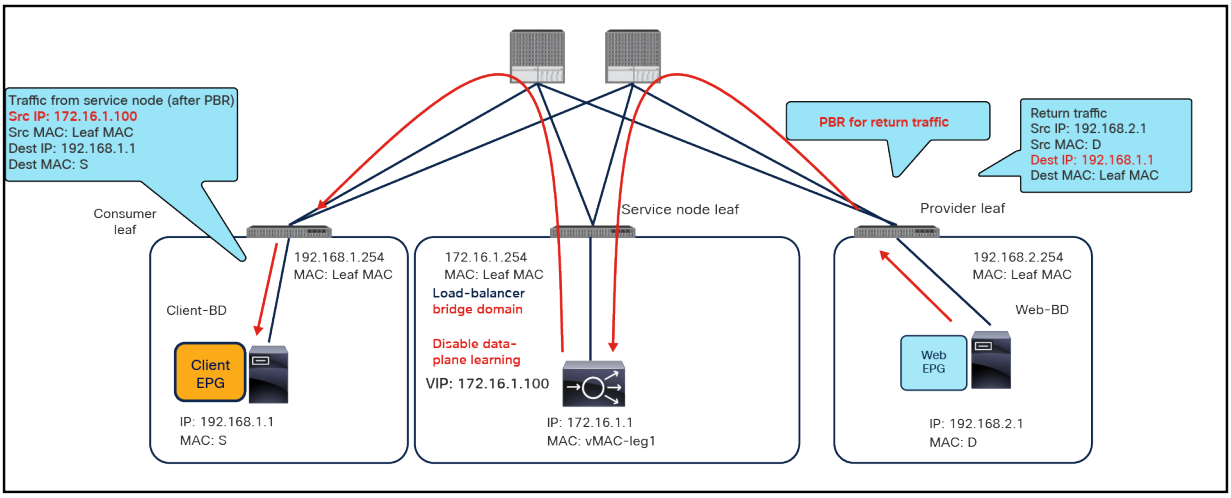


Figure 33.
Disable data-plane learning (unidirectional PBR use case)

Even if consumer and provider endpoints—for example, 192.168.1.1 and 192.168.2.1—are under the same leaf, the leaf doesn't learn local endpoints as remote endpoints through the service leaf.

IP Data-plane Learning per subnet

There are three different scopes and configuration levels in IP data-plane learning, as shown below:

- EPG Subnet – per host (/32 for IPv4 or /128 for IPv6)
- BD Subnet – per subnet
- VRF – per VRF

This section discusses the IP Data-plane Learning option that applies to a bridge domain subnet. For the option at an EPG subnet for IP, please refer to IP Data-plane Learning under EPG-level configuration options. For the option at VRF, please refer to IP Data-plane Learning under VRF-level configuration options.

The IP Data-plane Learning option per bridge domain **subnet** was introduced in Cisco APIC Release 5.2(1g). This option is located at Tenant > Networking > Bridge Domains > BD > Subnets (Figure 34). IP Data-plane Learning is enabled by default, but you can use this option to disable (or to enable again) endpoint data-plane IP learning for the bridge domain subnet.

When you disable IP Data-plane learning with this option you also need to make sure that the Subnet option "No Default SVI Gateway" is NOT checked. The "No Default SVI Gateway" option is typically selected when adding more specific subnets (e.g. when adding two /25 subnets in addition to the main /24 Subnet), but in the specific case of adding Subnets for which you want to disable IP Data-plane learning, you should not select the "No Default SVI Gateway". This is because of how the hardware is programmed when using both the global Enforce Subnet Check and (no) IP Data-plane learning.

The Bridge Domain which contains subnets where IP Data-plane learning is disabled must be configured with unknown unicast flooding instead of hardware-proxy. This is because with IP Data-plane learning disabled on a Subnet, the source MAC of endpoints is not learned from ARP traffic between the endpoints (with or without ARP proxy configured), so with the BD set for hardware-proxy the ARP resolution between new endpoints would not complete successfully. See the section L2 Unknown Unicast consideration, for details.

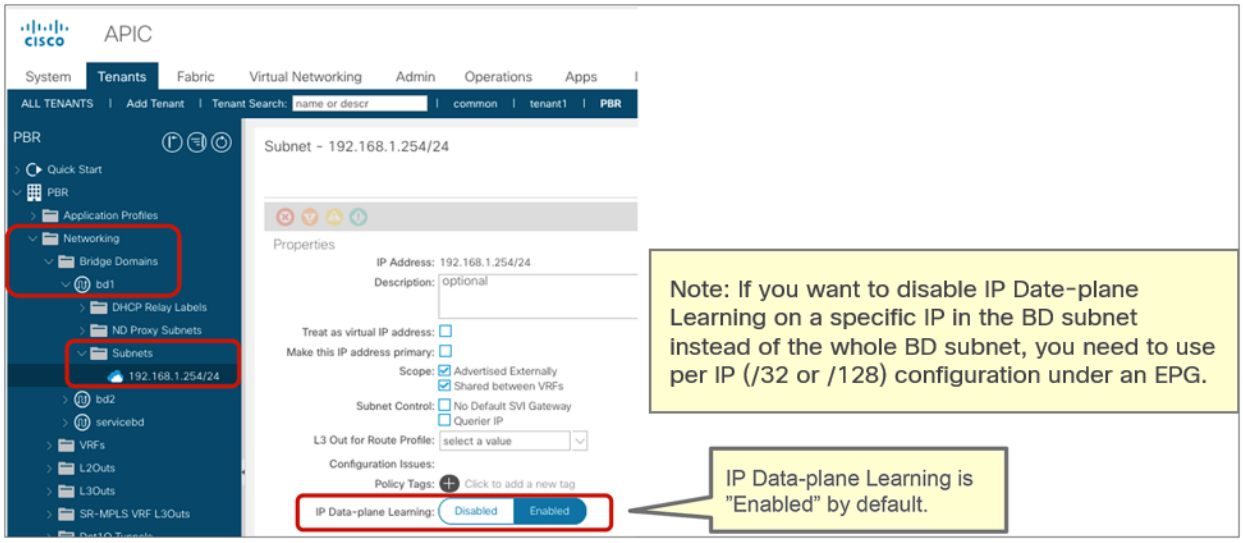


Figure 34.
Enable and disable Endpoint Data-plane Learning under the bridge domain subnet

For IP Data-plane Learning behavior, use cases, and considerations, please refer to Disabling IP Data-plane Learning: forwarding behavior and design considerations.

	GARP packets do not trigger endpoint learning (both MAC and IP) when this option is disabled. See the **GARP-based EP Move Detection** section for details.

	When the BD is participating in the Multi-Site via Nexus Dashboard Orchestrator (NDO) such that the BD is stretched or a shadow is created on the other sites, it is not supported to disable this option.

Endpoint Retention Policy

The Endpoint Retention Policy configuration is located at Tenant > Policies > Protocol > End Point Retention (Figure 35) and is referred from a Bridge Domain (BD) or a VRF (Figure 36). By default, a BD or a VRF refers to the default policy defined in the common tenant is used.

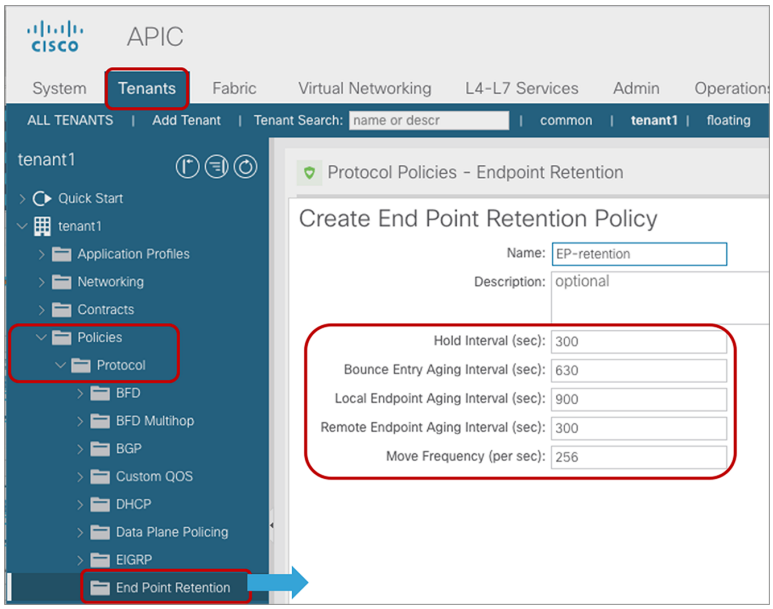


Figure 35.
Endpoint Retention Policy

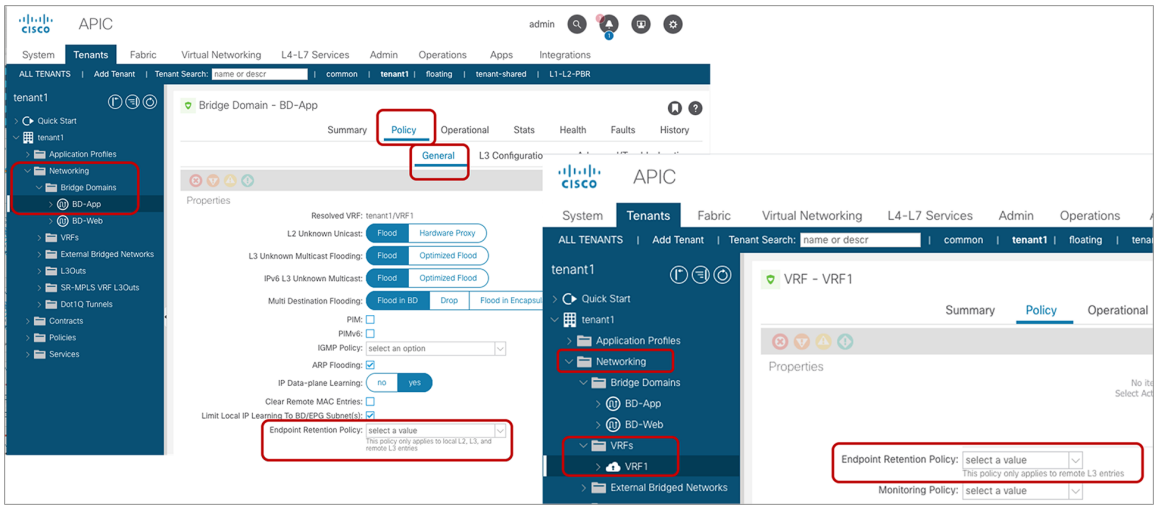


Figure 36.
Select an Endpoint Retention Policy

This option is used to specify the life cycle of endpoints using the following values:

- Hold Interval: The amount of time in seconds that endpoint learning is disabled in a bridge domain due to EP Loop Protection (BD Learn Disable) or Endpoint Move Dampening that is triggered based on the Move Frequency below. The default interval is 300 seconds.
- Bounce Entry Aging Interval: The amount of time in seconds until a bounce entry in the endpoint table on a leaf node expires. The default interval is 630 seconds.
- Local End Point Aging Interval: The amount of time in seconds that a leaf node can keep each local endpoint in its endpoint table without further updates. The default interval is 900 seconds. If 75 percent of the interval is reached, the leaf node sends three ARP requests to verify the presence of the endpoint. If no response is received, the endpoint is deleted.
- Remote End Point Aging Interval: The amount of time in seconds that a leaf node can keep each remote endpoint in its endpoint table without further updates. The default interval is 300 seconds.
- Move Frequency: The maximum number of endpoint moves that are allowed per second within a bridge domain on each leaf node. The number is counted as the total movement of any endpoint in the given BD, whether it is a single endpoint flap, a simultaneous move of multiple endpoints, or a combination of both. If the number of movements per second is exceeded the Move Frequency, the Hold Interval (described above) is triggered, and learning the new endpoint in the BD is disabled until the Hold Interval expires. The feature is called **BD Move Frequency** or **Endpoint Move Dampening**. The default is 256.

Please refer to the Cisco ACI endpoint learning behavior section of this document for details on bounce entry, local endpoints, and remote endpoints. This section mainly explains the Hold Interval and Move Frequency for Endpoint Dampening.

With the default configuration parameters above, Endpoint Move Dampening disables endpoint learning on the bridge domain for 300 seconds if the number of endpoint moves is more than 256 times per second.

Use case of Endpoint Move Dampening

Endpoint Move Dampening mitigates the impact of an unreasonable amount of endpoint moves within a short period of time (that is, 1 second to protect the ACI control plane from having to manage too many endpoint moves, which could be caused by an L2 loop. This is also to protect the ACI fabric against such issues as multiple flapping endpoints due to inappropriate configurations or designs. Figure 37 illustrates an example. An L2 loop causes a lot of endpoint moves between leaf nodes, which could potentially cost a huge ACI control plane resource. Endpoint Dampening counts endpoint moves and disables endpoint learning per bridge domain per leaf node, which allows the ACI fabric to narrow down the scope of the impact to each individual bridge domain on a specific leaf node without affecting other healthy bridge domains or leaf nodes.

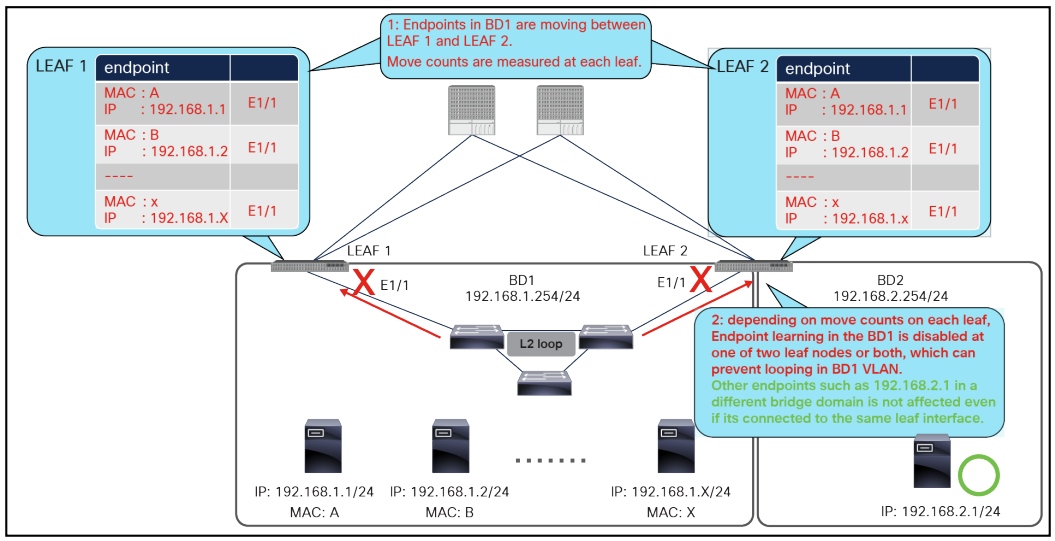


Figure 37.
Use case of Endpoint Move Dampening

Considerations for Endpoint Move Dampening are as follows:

- Endpoint Move Dampening doesn't distinguish between local or remote moves; any type of interface change is considered an endpoint move.
- Endpoint learning is disabled on bridge domains where the number of endpoint move counts per second is exceeded, and this could affect other healthy endpoints in the same bride domain. In such cases, those other endpoints still have chances to communicate through the ACI fabric because of two reasons;

  (a) Existing endpoints are still learned (not flushed).

  (b) Traffic toward new endpoints that cannot be learned for a time is simply flooded in the bridge domain if the L2 Unknown Unicast is set to flood in the bridge domain.
- However, there are two scenarios that may affect traffic for healthy endpoints in the same bridge domain.

  1. If L2 Unknown Unicast is set to hardware-proxy, traffic toward the new unlearned endpoints will be sent to a spine for spine-proxy and get dropped.
  2. If an existing endpoint moved across interfaces or VLANs or a new IP address is learned on an existing endpoint MAC address in the bridge domain with learning disabled, this results in the endpoint being flushed because ACI discovered that the existing endpoint information is no longer accurate while the learning of new endpoint information is disabled.

- If you prefer to disable endpoint learning of a specific endpoint that moves frequently instead of on the entire bride domain, please refer Rogue EP Control option.
- If Rogue EP Control is enabled, Endpoint Move Dampening will not take effect.
- If there are many IP addresses in a bridge domain that are expected to move at the same time, you might need to increase the Move Frequency to prevent endpoint learning from being disabled in the bridge domain. For example, when you failover an uplink of a server that contains hundreds of VMs, hundreds of endpoint moves will be detected on the ACI fabric in a short period of time. Or when active/standby failover takes place on an L4-L7 service device such as firewall and load balancer, the new active service device typically sends GARP for the IPs that it is going to take care of, such as an active IP and load balancer VIPs, etc., to inform the new active service.

VRF -level configuration options

IP Data-plane Learning per VRF

There are three different scopes and configuration levels in IP data-plane learning, as shown below:

- EPG Subnet – per host (/32 for IPv4 or /128 for IPv6)
- BD Subnet – per subnet
- VRF – per VRF

The IP Data-plane Learning option per VRF was introduced in Cisco APIC Release 4.0(1h). This option is located at Tenant > Networking > VRFs. It is enabled by default and enables and disables endpoint data-plane IP learning on the VRF.

For other IP Data-plane Learning configuration locations, please refer to IP Data-plane Learning subsection in Bridge domain-level configuration options and IP Data-plane Learning subsection in EPG-level configuration options.

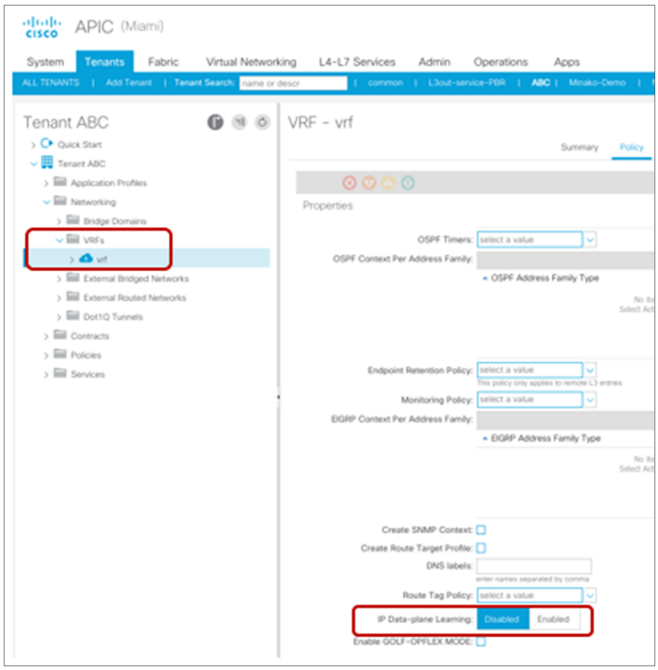


Figure 38.
Enable and disable Endpoint Data-plane Learning under the VRF

Disabling IP Data-plane Learning: forwarding behavior and design considerations

This section analyses in more details the ACI forwarding when IP Data-plane Learning is disabled. This section provides considerations and use cases that are applicable to the IP Data-plane Learning option at VRF, bridge domain subnet, and EPG subnet.

IP and MAC Learning with IP Data-plane learning disabled

When the IP Data-plane Learning option is disabled, endpoint learning behavior on an ACI leaf changes as follows:

- Local MACs and remote MACs are learned via the data plane (no change with this option).
- Local IPs are not learned via the data plane.
- Local IPs are learned from ARP/GARP/ND via the control plane.
- Remote IPs are not learned from unicast packets via the data plane.
- Remote IPs are learned from multicast packets via the data plane.

Note for the option at VRF: The above remote MAC learning behaviors apply to second-generation leaf switches. If it's a first-generation leaf switch, remote MAC is not learned, thus the hardware proxy mode on the corresponding BDs must be configured. Otherwise, L2 bridging traffics are flooded always if source and destination endpoints in the same BD are under different leaf switches. See the "First-generation leaf switch considerations" section below for detail.

Note for the option at bridge domain subnet or EPG subnet: If there are communication between endpoints in the same bridge domain, "L2 Unknown Unicast" must be set to "Flood" on the bridge domain, which means that ARP flooding must be enabled too; otherwise, ARP between endpoints in the same bridge domain will not be resolved. See the section L2 Unknown Unicast consideration, below, for details.

When the IP Data-plane Learning option is disabled, existing remote IP endpoints are flushed immediately while bounce entries are retained and age out normally. Existing local IP endpoints are not flushed either, but they will age out eventually unless control plane packets such as ARP keep them alive.

When the IP Data-plane Learning option is disabled, it is recommended to ensure IP Aging option is enabled as well. This is to ensure Host Tracking, which sends ARP/ND for a local IP endpoint at 75% of its retention timer, is always triggered to correctly track the IP status via control-plane. Without IP Aging option enabled, local IP endpoints may not age out correctly due to data-plane traffic even when the IP Data-plane Learning is disabled.

Unlike the Endpoint Data-plane Learning option under BD in the previous section, the IP Data-plane Learning option under VRF, bridge domain subnet, and EPG subnet is not restricted to PBR use cases only.

When to disable IP Data-plane Learning

The IP Data-plane Learning option must be disabled if you have a possibility that the ACI fabric receives traffics with the same-source IP address from different locations, which causes endpoint IP and MAC binding updates that occur due to data plane traffic. For example, mainframe VIPA connectivity and dynamic load balancing mode for NIC teaming on Microsoft Windows behave in this manner.

Another use case is when multiple devices share the same IP such as virtual IP (VIP) and ARP/GARP/ND is used to claim the ownership of the VIP among the devices. In that situation, those external devices may source data traffic from the same VIP at the same time, for example, when a failover is taking place. That could result in the ACI fabric learning the VIP from multiple places via the data plane. This sort of issue can be avoided by disabling IP Data-plane Learning.

See figures 39 through 40 for the example of this issue.

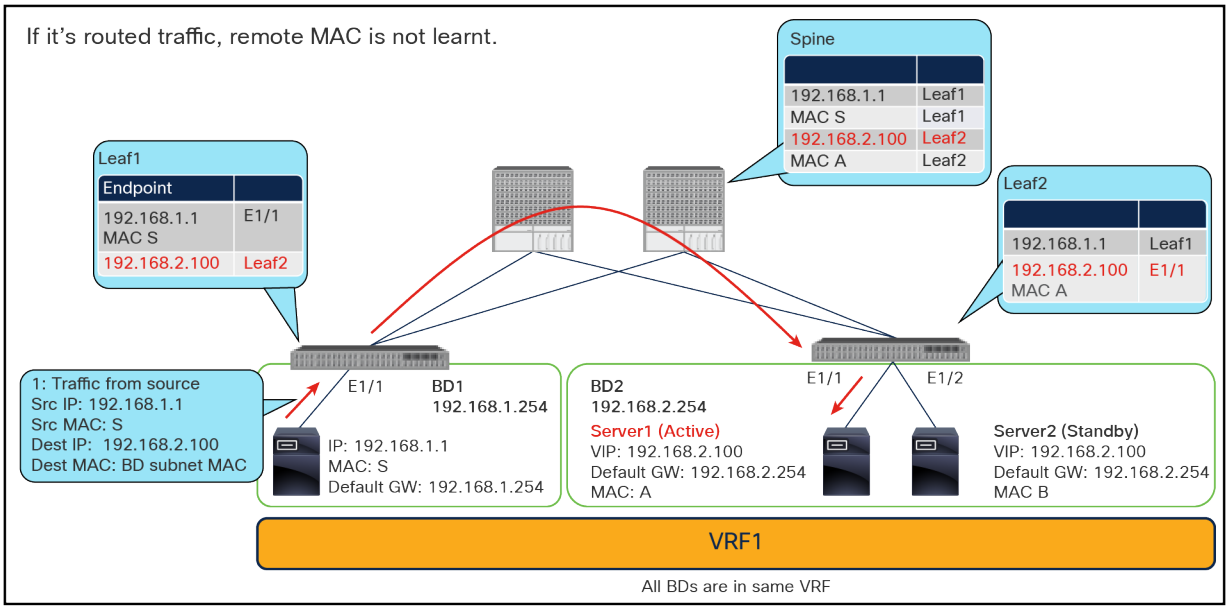


Figure 39.
Why you need to disable endpoint data-plane learning (server1 is active)

This example shows active-standby servers that share the same active IP address that is primarily owned by the active server. When server1 is active and server2 is standby, server1 takes care of 192.168.2.100 that is learned on Leaf2 E1/1.

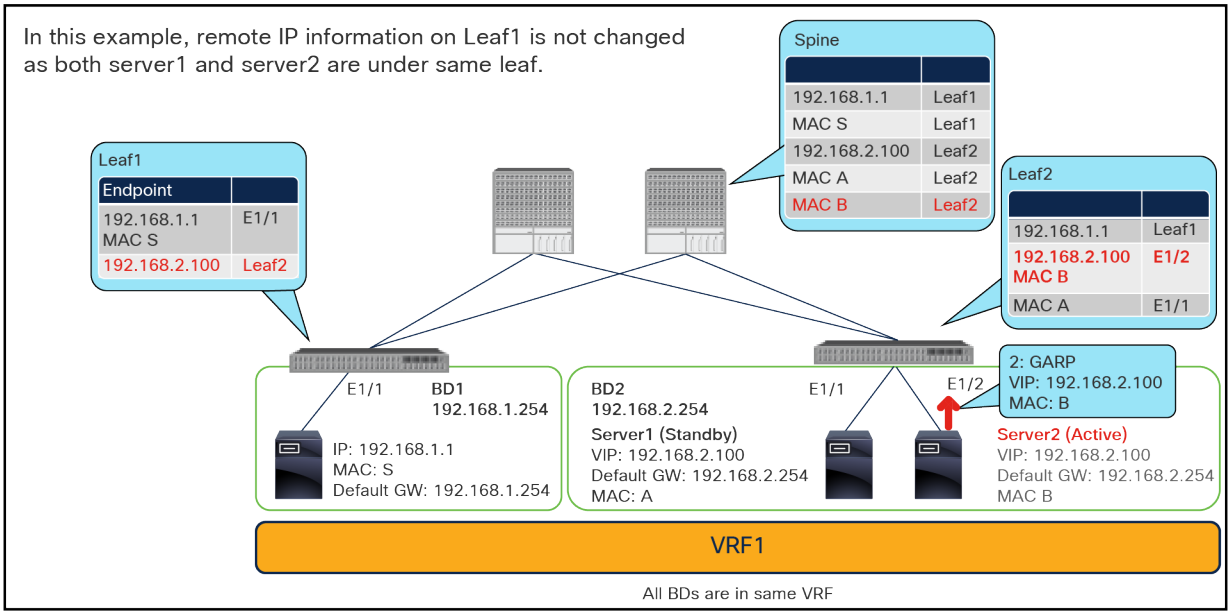


Figure 40.
Why you need to disable endpoint data-plane learning (server2 becomes active)

When server2 takes over the active role, server2 sends GARP and 192.168.2.100 is now learned on Leaf2 E1/2.

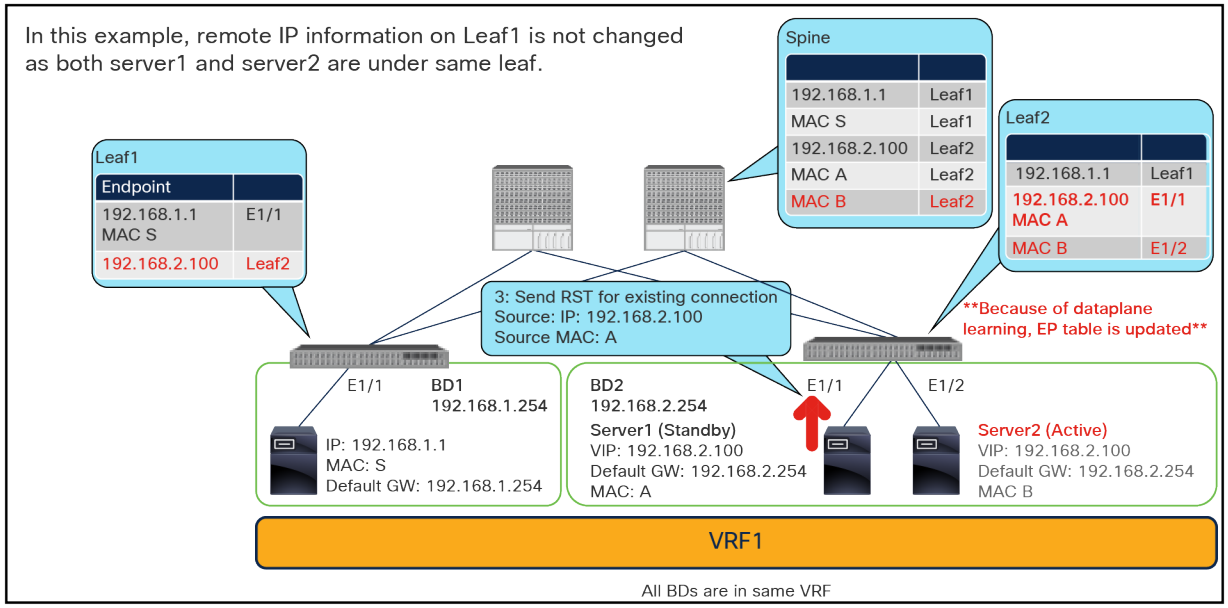


Figure 41.
Why you need to disable endpoint data-plane learning (server1 sends TCP RESET)

A problem occurs if server1 sends a TCP RESET from the VIP address 192.168.2.100 to terminate existing TCP sessions that server1 had been handling. This TCP RESET results in data-plane IP learning for the VIP address 192.168.2.100 on Leaf2 E1/1.

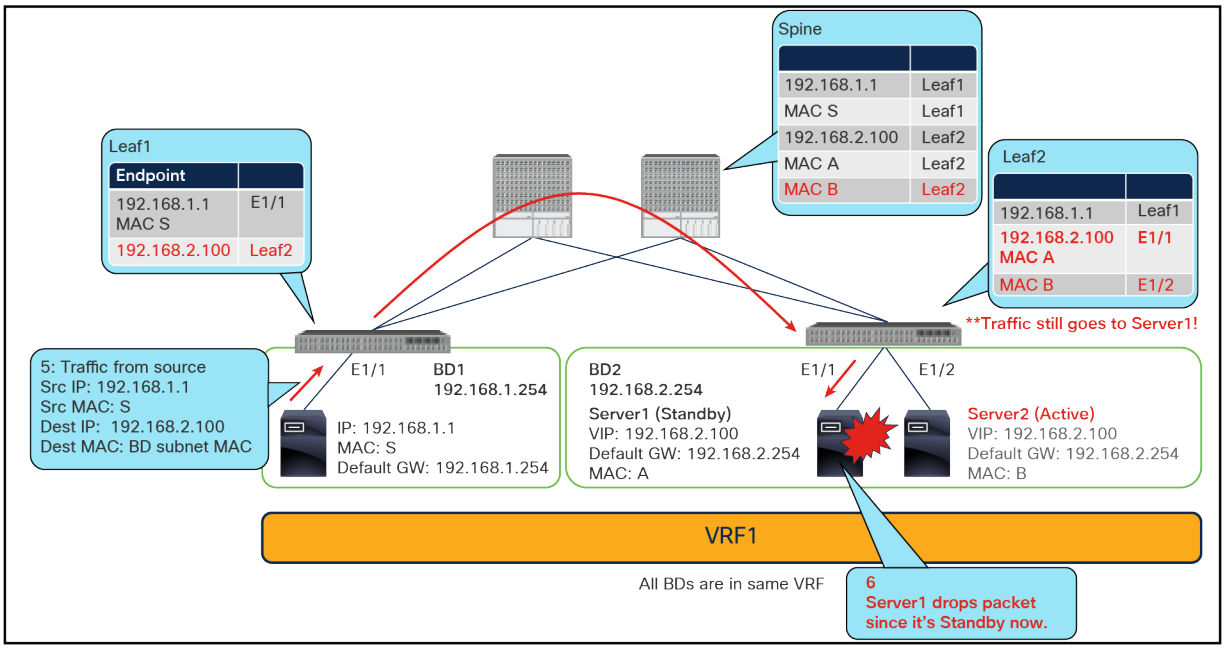
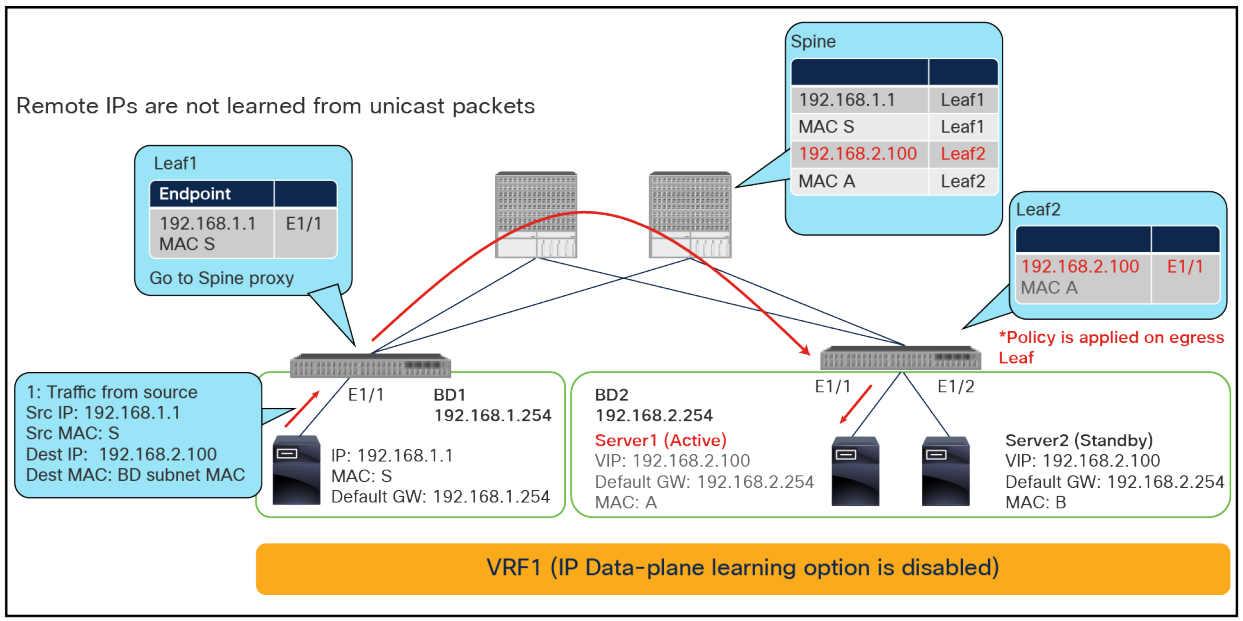
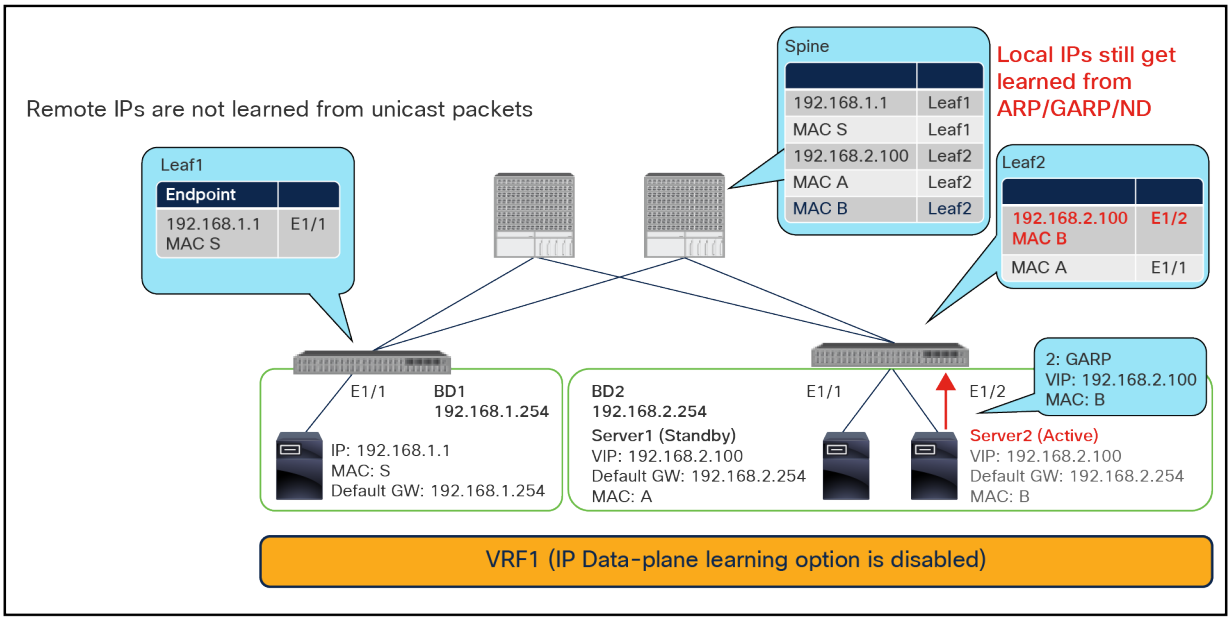


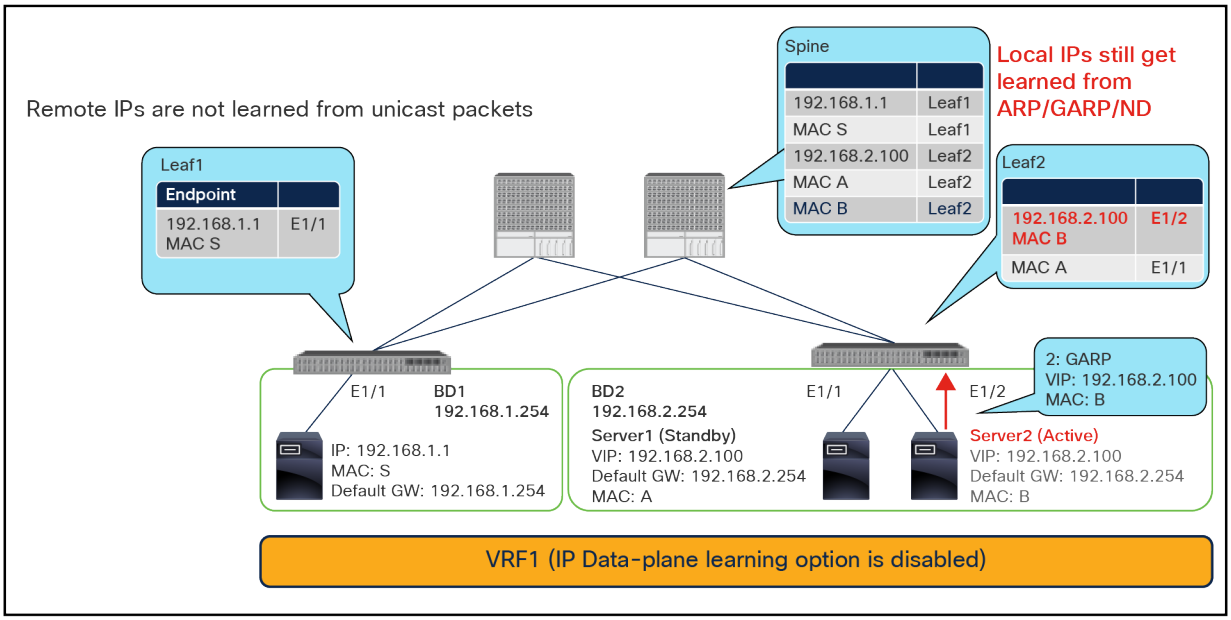
Figure 42.
Why you need to disable endpoint data-plane learning (traffic goes to server1)

Unless server2 sends GARP for the VIP again, traffic toward the VIP 192.168.2.100 will be forwarded to server1 that is currently standby, and that drops packet.

Figure 43 explains how this issue can be avoided by disabling IP Data-plane Learning. If IP Data-plane Learning is disabled, endpoint learning information is not updated through RST from server1.







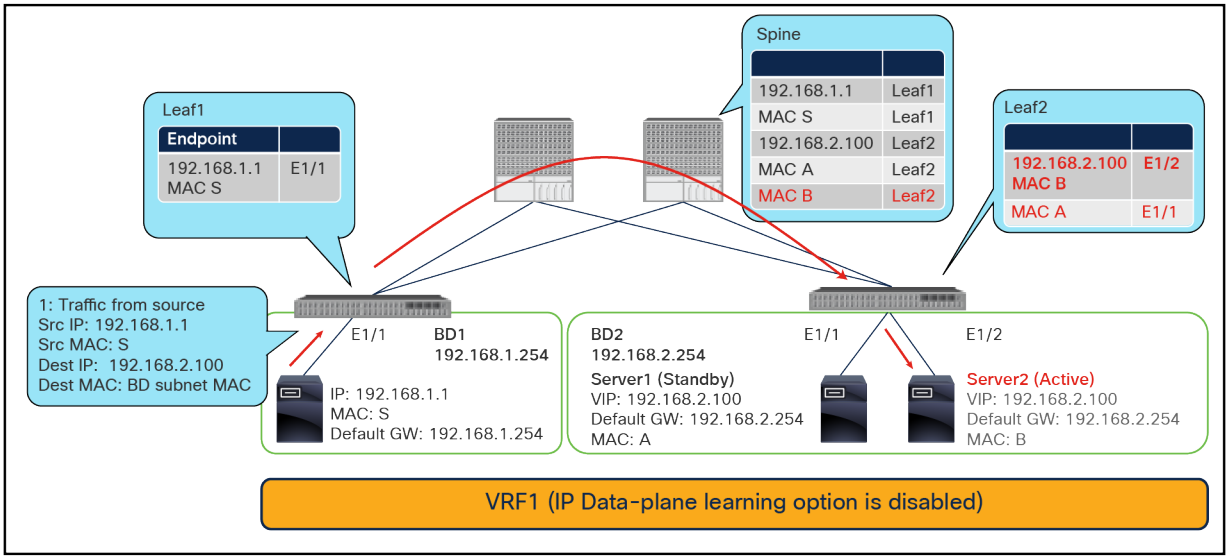


Figure 43.
Use case with disable endpoint data-plane learning

Note: In this EPG-to-EPG routed-traffic example, the contract policy will be applied on the egress leaf as the remote endpoint IP is not learned. If it is EPG-to-EPG bridged traffic, the contract policy can be applied on the ingress leaf since remote MAC is still learned.

For a DSR use case, use of the L4-L7 Virtual IPs option is still recommended as the L4-L7 VIP option can prevent learning VIP from other EPGs via both the control plane and data plane.

First-Generation leaf switch considerations

This is applicable to the option at VRF only because the other options are available after Cisco APIC Release 5.2(1g) that doesn't support first-generation leaf switch anymore. On first-generation leaf switches, remote MACs are not learned when the IP Data-plane Learning option under VRF is disabled. Thus, the L2 Unknown Unicast option under the BDs must be set to hardware-proxy mode. This is to take care of first-generation leaf limitations where a L2 Unknown Unicast packet is flooded from one leaf to another leaf pair on which the destination MAC is locally learned on a vPC. If the vPC leaf pair is first generation, both leaf switches will send it out to their respective destination vPC interface. This results in a packet from both leaf switches, which is seen as duplicate.

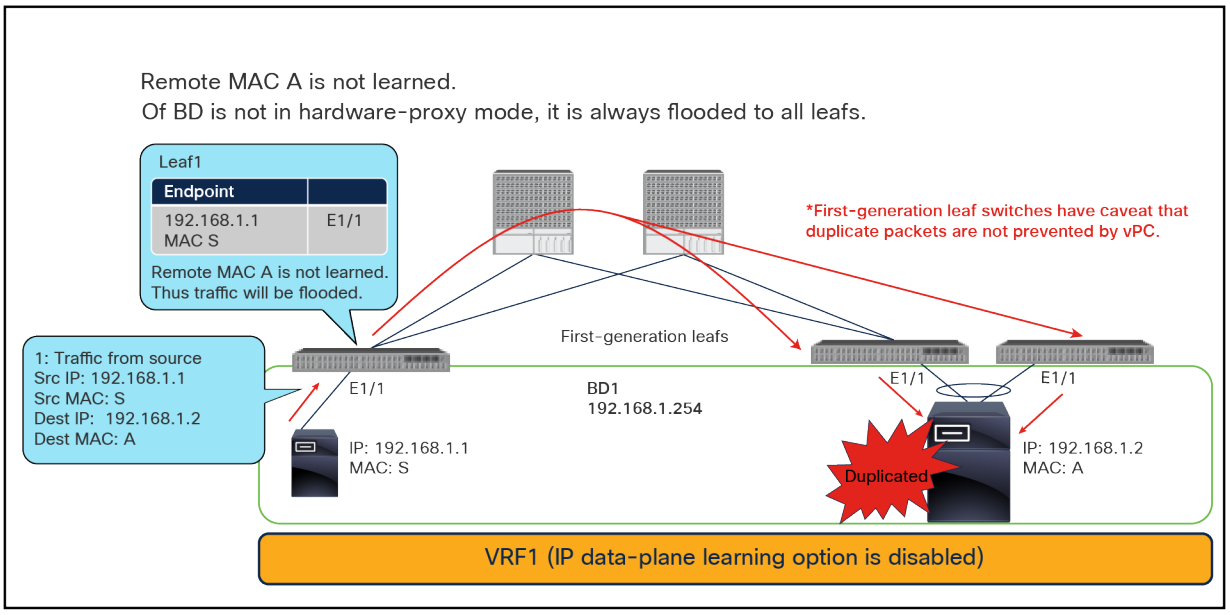
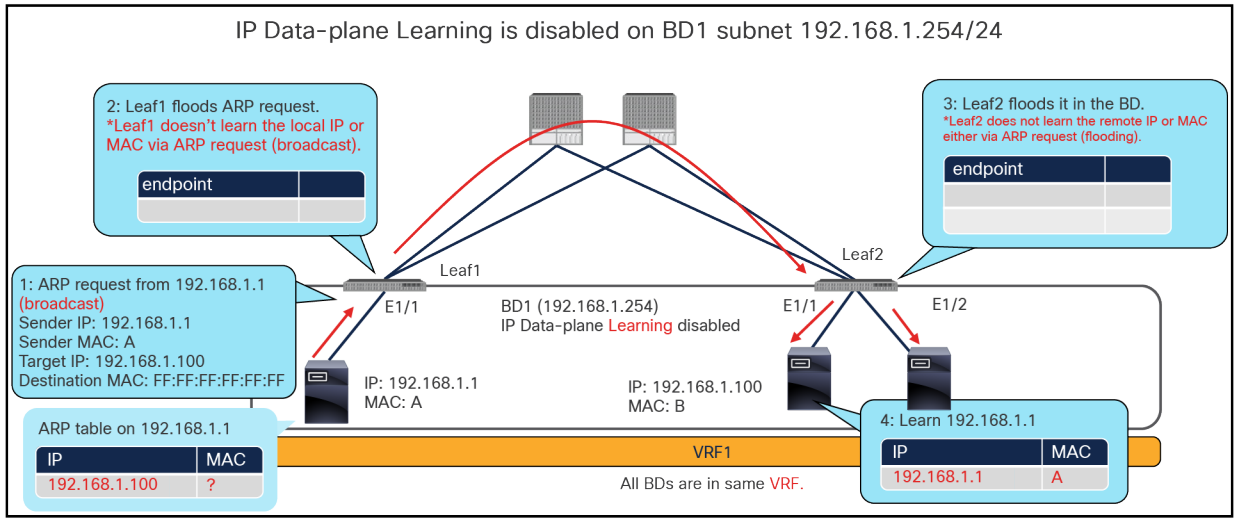


Figure 44.
Why hardware-proxy mode must be set (first-generation leaf switch)

L2 Unknown Unicast considerations

If IP Data-plane Learning is disabled by using either a per bridge domain subnet or per EPG subnet option, either local or remote MAC is not learned through an endpoint-to-endpoint ARP request, although it is still learned through an ARP request to a bridge domain SVI gateway or through any other traffic. Thus, the L2 Unknown Unicast option under the bridge domain must be set to Flooding mode to take care of ARP resolution between endpoints in the same bridge domain subnet.

Figure 45 shows an example. Because either local or remote MAC is not learned on both Leaf1 and Leaf2, the unicast ARP reply at step 6 and 7 needs to be flooded, which requires the L2 Unknown Unicast option to be “Flooding” mode.



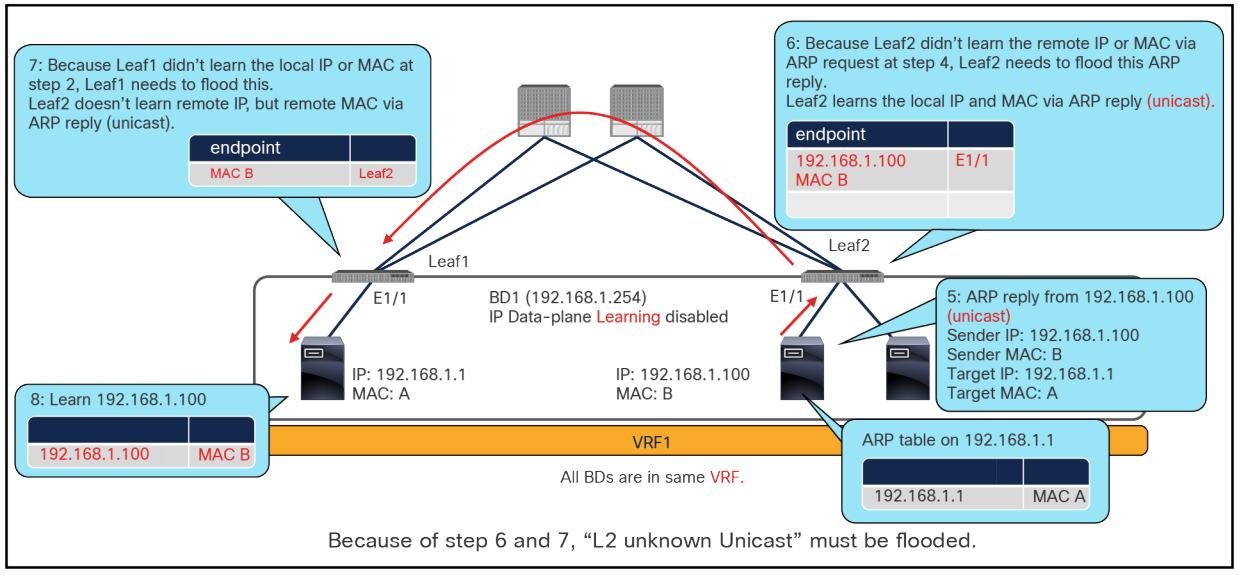
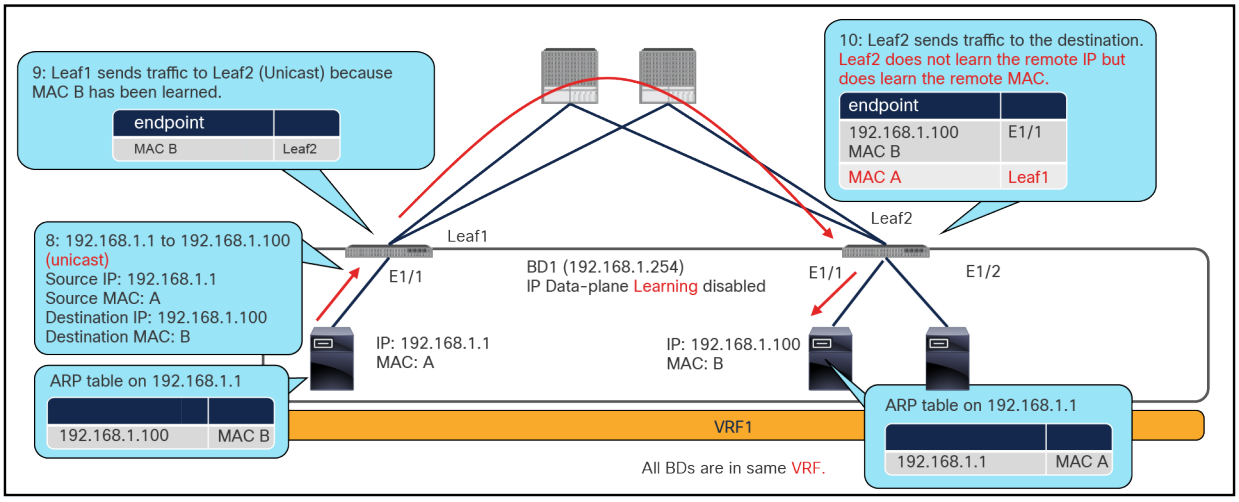


Figure 45.
Why L2 Unknown Unicast must be set to "Flood" (ARP resolution)

After the ARP resolution, L2 unicast traffic between endpoints works as illustrated in Figure 46. Because it is a unicast traffic, not an ARP request traffic, the remote MAC (MAC A) is learned on Leaf2 at Step 10.



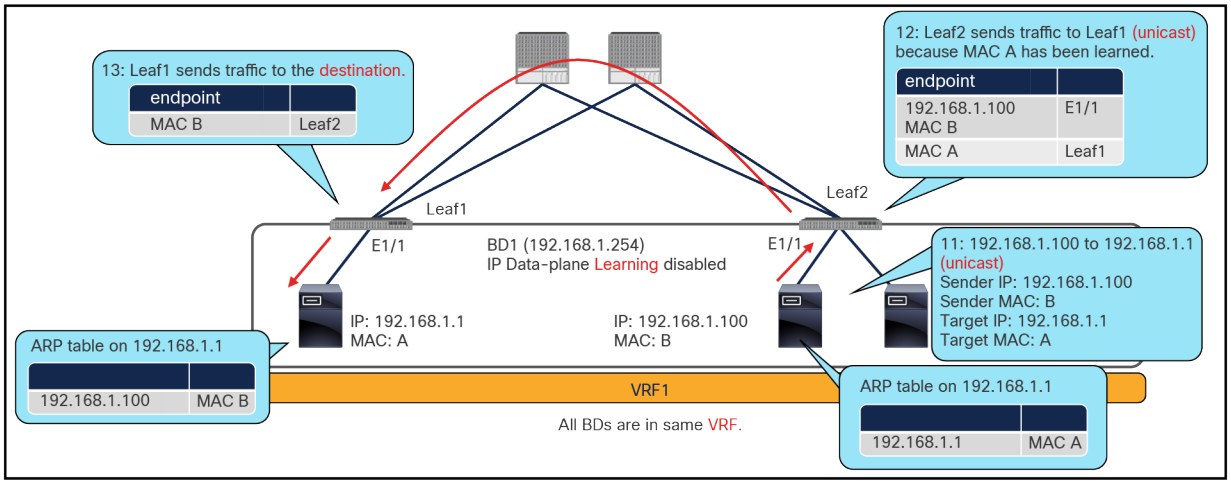


Figure 46.
Why L2 Unknown Unicast must be set to “Flood” (unicast traffic after ARP resolution)

Fabric-level configuration options

This section discusses options that apply to the entire fabric.

Disable Remote EP Learn (on border leaf)

The Disable Remote EP Learn option was first introduced in APIC Release 2.2(2e) with the following enhancement:

CSCuz19695: Stale endpoint on Border Leaf after EP move

In APIC Release 2.0, this option is located at Fabric > Access Policies > Global Policies > Fabric Wide Setting Policy (Figure 47). For APIC Release 3.0(1k) and later, it is located at System > System Settings > Fabric Wide Setting (Figure 48). This option is disabled by default. Prior to Cisco ACI Release 3.0(2h), this option requires ingress policy enforcement in the VRF instance.

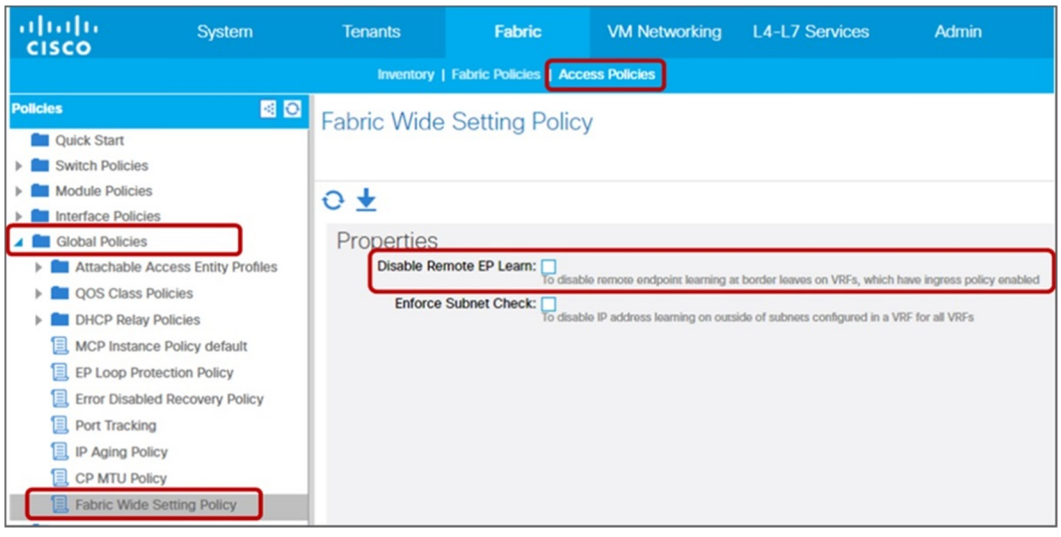


Figure 47.
Disable Remote EP Learn under Fabric-Wide Setting Policy (APIC Release 2.0)

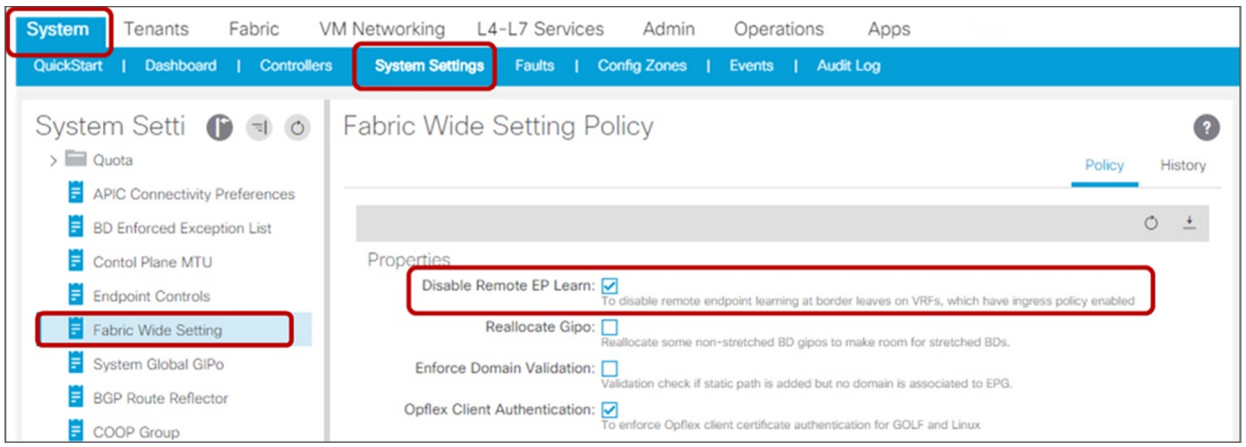


Figure 48.
Disable Remote EP Learn under Fabric Wide Setting (APIC Release 3.0)

When this feature is enabled, remote IP endpoint learning at the VRF instance is disabled on border leaf switches. However, border leaf may still learn remote IP endpoints from IP multicast routing packets, because of a limitation in the Cisco ACI IP multicast routing implementation. This exception applies only when a second-generation switch is used as the border leaf because Cisco ACI IP multicast routing is supported only starting with second-generation switches. This feature doesn't disable remote MAC endpoint learning.

Disable Remote EP Learn use case 1

The Disable Remote EP Learn feature was originally introduced to address scenario 3 in the "L3Out endpoint learning considerations" section. In this scenario, IP1 on LEAF1 is learned as a remote endpoint on border LEAF3 due to communication with normal endpoint IP2 on LEAF3. The potential problem here is that this remote endpoint could become stale. It could become stale after IP1 ceases communication with IP2 and moves to LEAF2 while IP1 is still continuing to send traffic toward the L3Out connection on LEAF3. Because of this traffic from source IP1 toward the L3Out connection on LEAF3, in a VRF instance with ingress policy enforcement mode, the remote endpoint on LEAF3 for IP1 pointing to the previous LEAF1 does not age out, nor is it updated with a new source leaf, LEAF2. (Refer to the discussion of scenario 3 in the "L3Out endpoint learning considerations" section for details and figures.)

You can prevent this stale-remote-endpoint scenario by using the Disable Remote EP Learn option on the border leaf. If you enable this option, border leaf LEAF3 prevents the remote IP endpoint from being learned (Figure 49). Because there is no remote endpoint, there will be no stale endpoint.

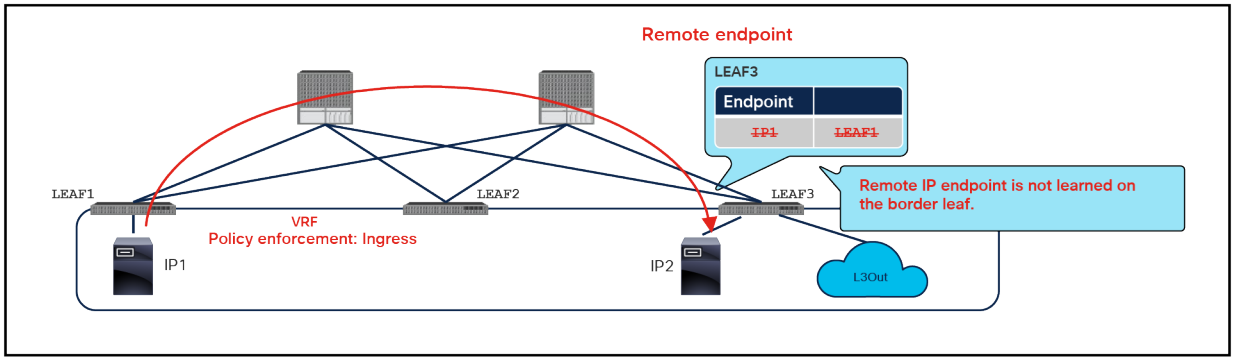


Figure 49.
Disable Remote EP Learning example on border leaf

This particular example applies only to first-generation leaf switches sourcing traffic toward the border leaf, as mentioned in the scenario 3 discussion earlier in this document. However, this feature can be used on second-generation leaf switches to prevent unexpected remote endpoint learning on a border leaf as mentioned in use case 2.

Disable Remote EP Learn use case 2

The Disable Remote EP Learn option prevents another issue when the same encapsulation VLAN SVI in one L3Out is deployed on multiple leaf switches. Figures 50, 51, and 52 show one possible scenario. However, users no longer need to use this option for this use case when your ACI version supports the endpoint announce feature (CSCvj17665) that is mentioned in the Endpoint movement and bounce entries section.

This use case applies to mainly second-generation leaf switches due to the limitation CSCva56754 mentioned below. But it could apply to first-generation leaf switches as well when the endpoint retention timer is not configured properly.

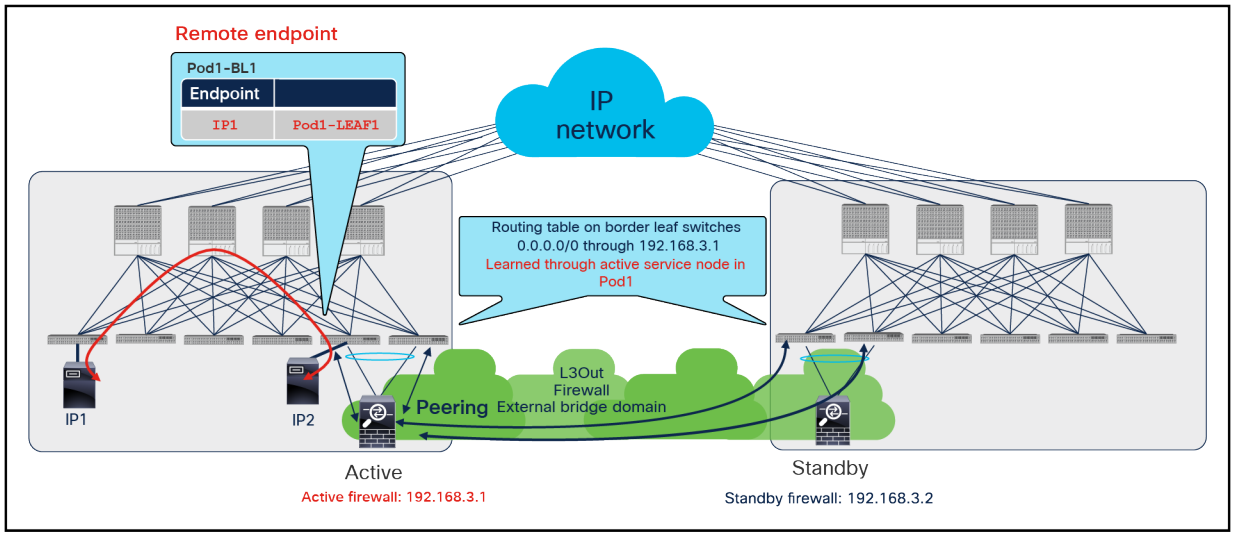


Figure 50.
Disable Remote EP Learn use case 2 (part 1)

In Figure 51, Cisco ACI is configured as Multi-Pod and each pod has a firewall (in active-standby mode) connected through the same L3Out connection with the same encap VLAN SVI on each leaf pair. When multiple border leaf switches are configured with the same encap VLAN SVI in one L3Out connection, all border leaf switches belong to the same Layer 2 domain (that is, the L3Out bridge domain). Therefore, all border leaf switches in each pod directly peer with the active firewall on Pod1 at the same time, and all border leaf switches have routes pointing toward the active firewall in Pod1.

At this time, in Figure 51, an endpoint with IP1 is sending traffic to another endpoint with IP2 on the border leaf in the same Pod1. This traffic causes one of the border leaf switches in Pod1 to learn the remote IP endpoint for IP1 pointing to LEAF1.

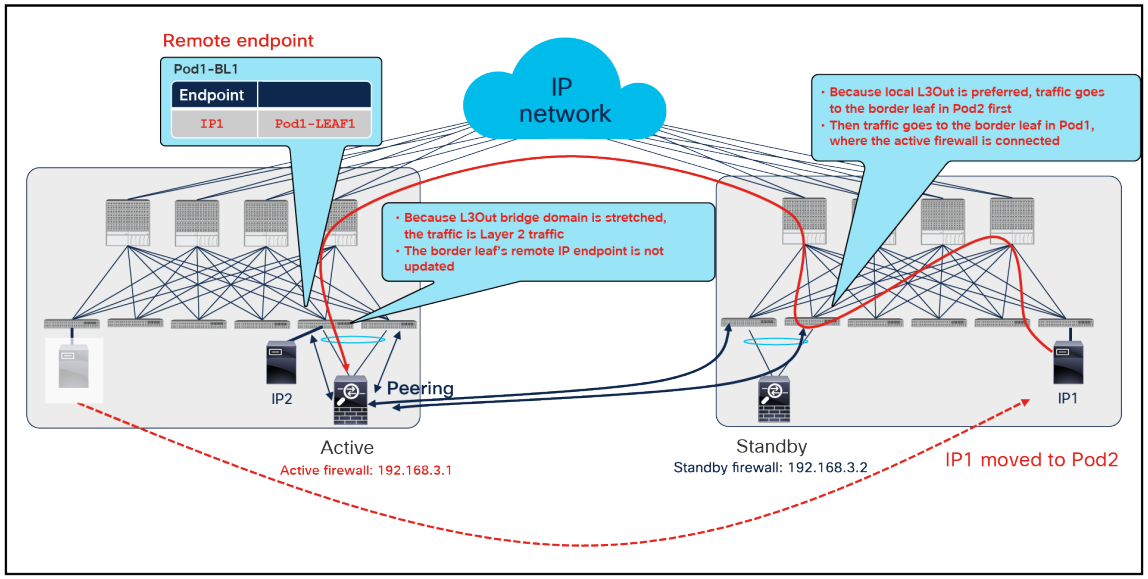


Figure 51.
Disable Remote EP Learn use case 2 (part 2)

Figure 52 shows what happens when IP1 in Figure 51 is migrated to Pod2. If IP1, now in Pod2, is trying to reach the firewall, the traffic is forwarded to the border leaf switches in Pod2 because they know the routes directly from the active firewall, and the local border leaf switches are preferred over the border leaf switches in another pod.

Next, the traffic is looked up on the border leaf in Pod2, and the next-hop MAC address for the active firewall is resolved on it through an ARP entry on the same border leaf. However, the active firewall is not physically connected to the same border leaf. Hence, the traffic is bridged to the border leaf switches in Pod1 through the Inter-Pod Network (IPN) connection. This traffic does not update the previous remote endpoint for IP1 on the border leaf in Pod1 because this traffic is switched and not routed. Therefore, only the remote MAC address is learned, not the IP address.

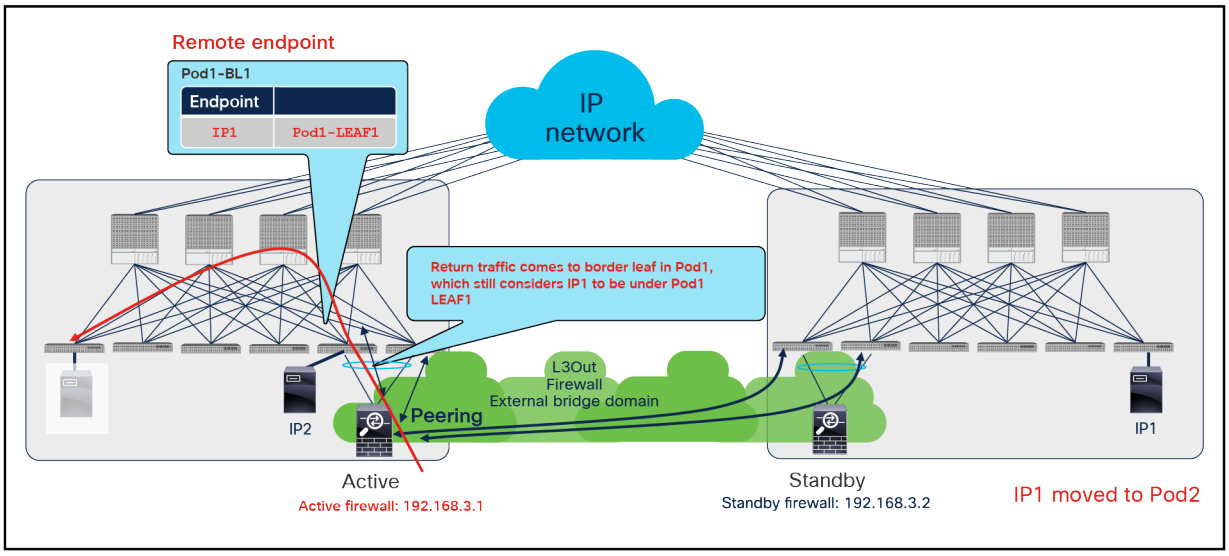


Figure 52.
Disable Remote EP Learn use case 2 (part 3)

Because of what happens in Figure 53, return traffic from the active firewall, or any other traffic to IP1 from the border leaf, hits the previous stale remote endpoint for IP1 pointing to the previous leaf, LEAF1.

As long as the previous leaf has a bounce entry, traffic will be forwarded to the new location in pod2. If the bounce entry is aged out and the border leaf still has a remote endpoint for IP1 pointing to the previous leaf, it could cause a loss of traffic toward IP1 from this border leaf. This loss of traffic should usually not happen because a remote endpoint ages out before a bounce entry ages out by default. However, the aging timer for a remote IP endpoint is improperly updated by L2 bridged traffic only on second generation leaf switches due to the following limitation.

CSCva56754 ACI: remote IP endpoint is not aging out due to L2 (bridged) traffic

With that said, this use case always applies to second generation leaf switches. For first-generation leaf switches, it depends on age timer (retention timer) configuration for bounce and remote endpoint.

If the Disable Remote EP Learn option is enabled, the border leaf switches in each pod will not learn the remote endpoint on that VRF instance in the first place, which can prevent this concern.

	This topology with vPC for two border leaf pairs is supported only on second-generation leaf switches starting from Cisco ACI Release 2.3(1) regardless of whether a multiple-pod or single-pod design is used.
	The topology with a normal port channel or access port (For example, one border leaf switch for each firewall) for two border leaf switches—one for each—is supported regardless of the generation of the leaf switch, starting from Cisco ACI Release 2.2(2), regardless of whether a multiple-pod or single-pod design is used.

A note for vPC and Disable Remote EP Learn

Although it is considered a best practice to have dedicated border leaf switches as mentioned in scenario 3 for “L3Out endpoint learning considerations” section, there may be some cases where L3OUT is deployed on a single leaf and at the same time the same leaf is a part of vPC for non-L3OUT resources. In this scenario, the border leaf itself does not learn a remote IP endpoint directly from data plane when **Disable Remote EP Learn** is activated. However, the vPC peer that is not a border leaf may learn a remote IP endpoint and the information will be synced to its vPC peer that is the border leaf. This behavior is expected and recorded with the following ID.

CSCvi50954: Disabling remote EP learning doesn't disable learning on VPC peer of Border Leaf.

Figure 53 depicts this behavior.

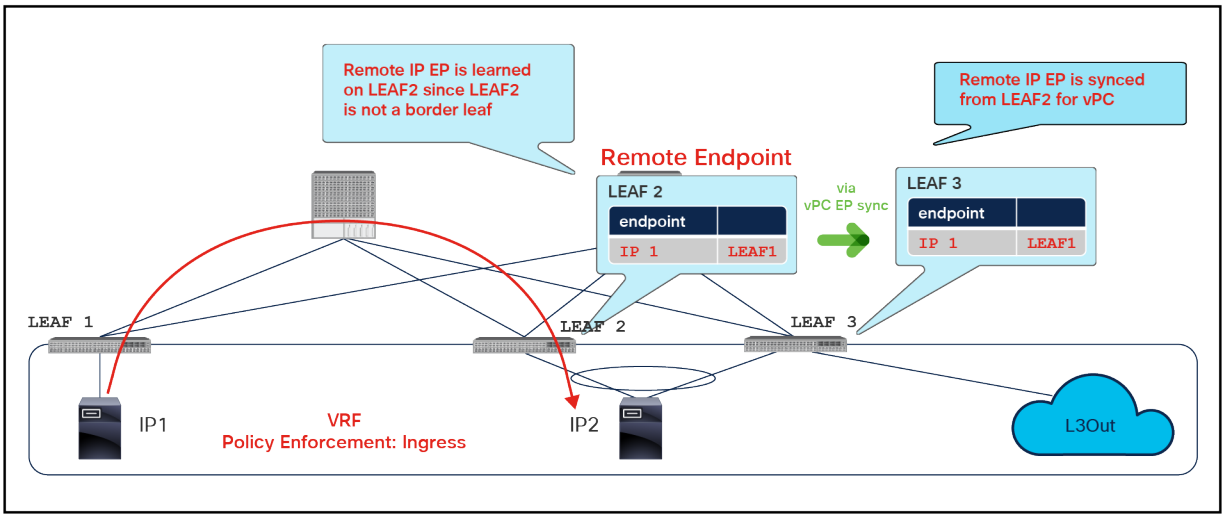


Figure 53.
A note for vPC and Disable Remote EP Learn

Enforce Subnet Check

The Enforce Subnet Check option was first introduced in APIC Releases 2.2(2q) and 3.0(2h) with the following enhancement:

CSCvf43074: ACI knob to limit IP EP learning to available BD subnets under the same VRF

In APIC Release 2.2(2q), the option is located at Fabric > Access Policies > Global Policies > Fabric Wide Setting Policy (Figure 54).

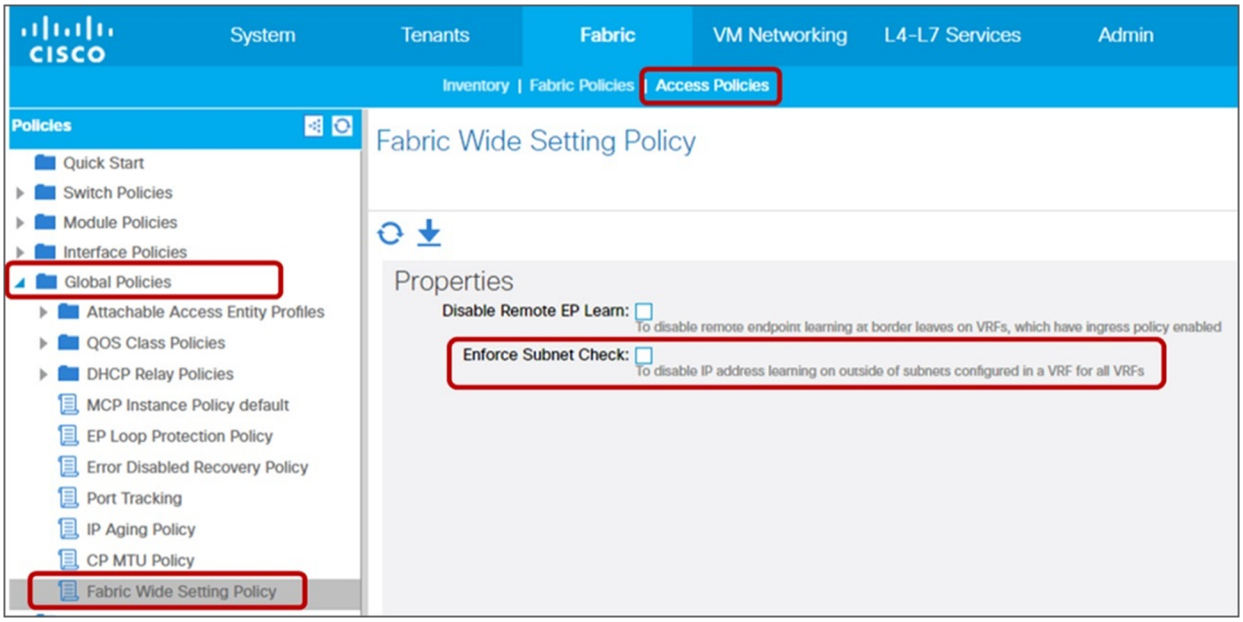


Figure 54.
Enforce Subnet Check under Fabric Wide Setting Policy (APIC Release 2.2(2q))

In APIC Release 3.0(2h) and later, it is located at System > System Settings > Fabric Wide Setting (Figure 55).

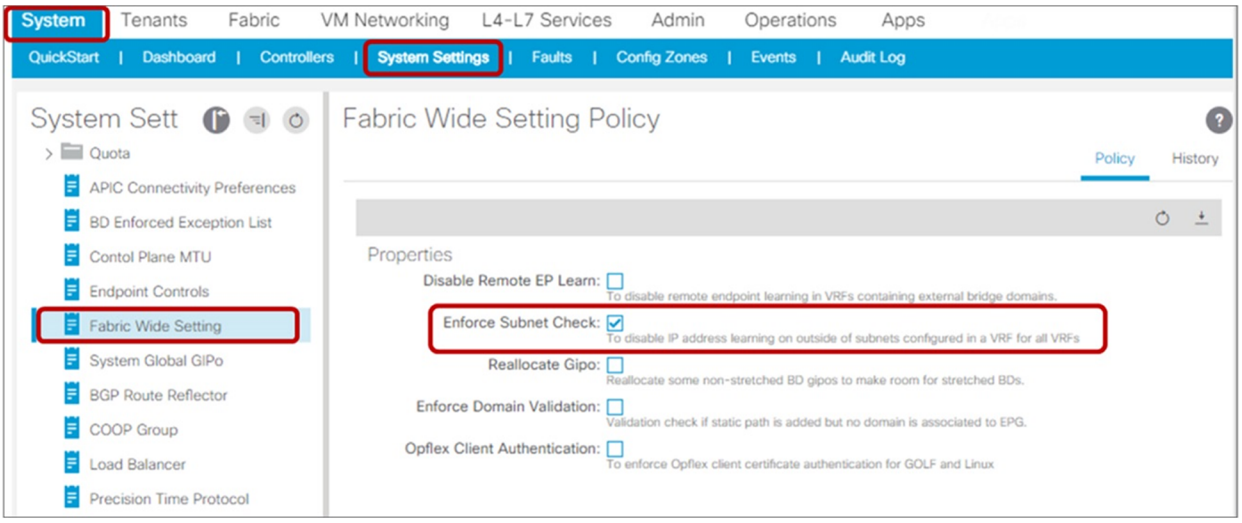


Figure 55.
Enforce Subnet Check under Fabric Wide Setting Policy (APIC Release 3.0(2h))

This feature is available only on second-generation leaf switches.

This feature enforces subnet checks at the VRF level, when Cisco ACI learns the IP address as an endpoint from the data plane. Although the subnet check scope is the VRF instance, this feature can be enabled and disabled only globally under Fabric Wide Setting Policy. You cannot enable this option only in one VRF instance.

This feature is disabled by default.

This feature optimizes the learning of both local and remote endpoints across the fabric. Hence, this can be considered a superior version of the **Limit IP Learning To Subnet** option under each bridge domain. **Enforce Subnet Check** uses the bridge domain subnets in each VRF as the indicator for the validity of the endpoint learning.

- On the ingress leaf (local endpoint learning):

  The option enforces bridge domain–level subnet checks for local endpoint learning. When this feature is enabled, the Cisco ACI leaf learns an IP address and MAC address as a new local endpoint only when the source IP address of the incoming packet belongs to one of the ingress bridge domain subnets.

  This behavior is almost the same as Limit IP Learning To Subnet option under the bridge domain. The difference is that Limit IP Learning To Subnet limits only IP learning if the source IP address of a packet doesn't belong to an ingress bridge domain subnet, whereas this feature limits learning of both the MAC address and IP address when IP learning is triggered but yet prevented because the source IP address doesn't belong to an ingress bridge domain subnet. Please note that, regardless of the source IP range, the Cisco ACI leaf still learns MAC address if the packet is a bridging traffic because the leaf does not check the IP header or whether or not it has the IP header for bridging traffic.

  Thus, Enforce Subnet Check enables slightly stronger checks than Limit IP Learning To Subnet. This check will be enabled on all bridge domains, and you cannot turn the checks on and off per bridge domain. Therefore, Limit IP Learning To Subnet is not required when this feature is enabled.

- On the egress leaf (remote endpoint learning):

  This option enforces VRF-level subnet checks for remote endpoint learning. When this feature is enabled, the egress Cisco ACI leaf will learn an IP address as a remote endpoint only when the source IP address of the incoming packet belongs to a bridge domain subnet (of that VRF) that is present on the egress ACI leaf.

  This behavior prevents IP spoofing scenarios, in which an endpoint sends a packet with an unexpected source IP address that does not belong to any of the bridge domains on the VRF instance, such as an IP address that exists behind the L3Out connection.

When this feature is enabled, Cisco ACI flushes all local IP endpoints outside bridge domain subnets and all remote IP endpoints.

Enforce Subnet Check use case

Figures 56 and 57 show use case examples that provide details about the behavior of the Enforce Subnet Check option.

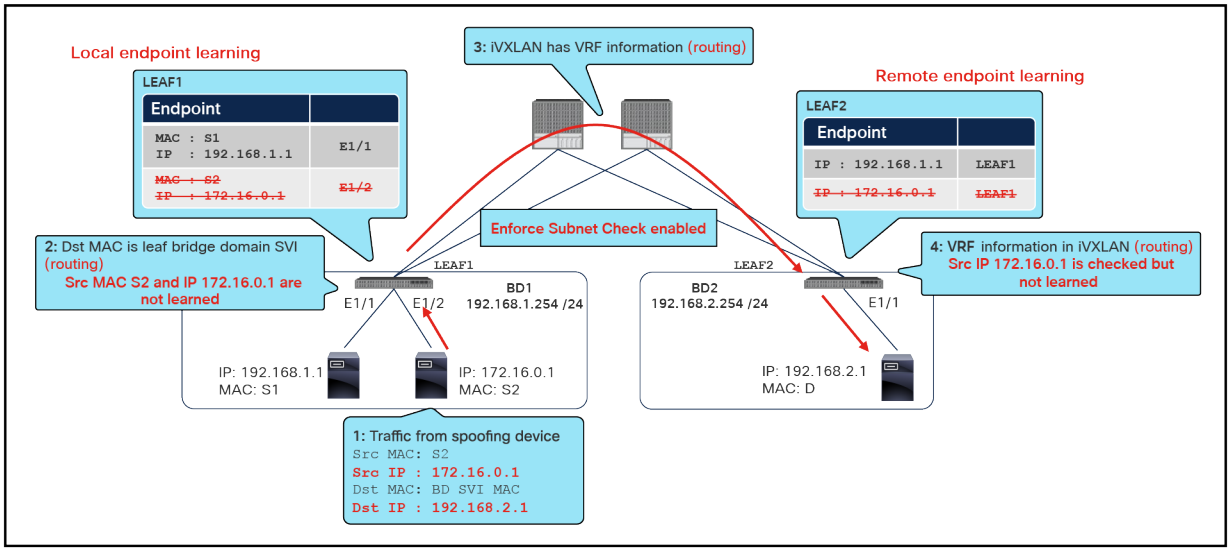


Figure 56.
Enforce Subnet Check example 1

When Enforce Subnet Check is enabled, as in Figure 57, LEAF1 does not learn either MAC S2 or IP 172.16.0.1 as a local endpoint, because 172.16.0.1 doesn't belong to ingress BD1. LEAF2 doesn't learn IP 172.16.0.1 as a remote endpoint, because 172.16.0.1 doesn't belong to any of the bridge domain subnets on LEAF2 in the same VRF instance. If 172.16.0.1 is learned as a local endpoint on LEAF1 and the remote endpoint on LEAF2 before this feature is enabled, those two endpoints are cleared after this feature is enabled.

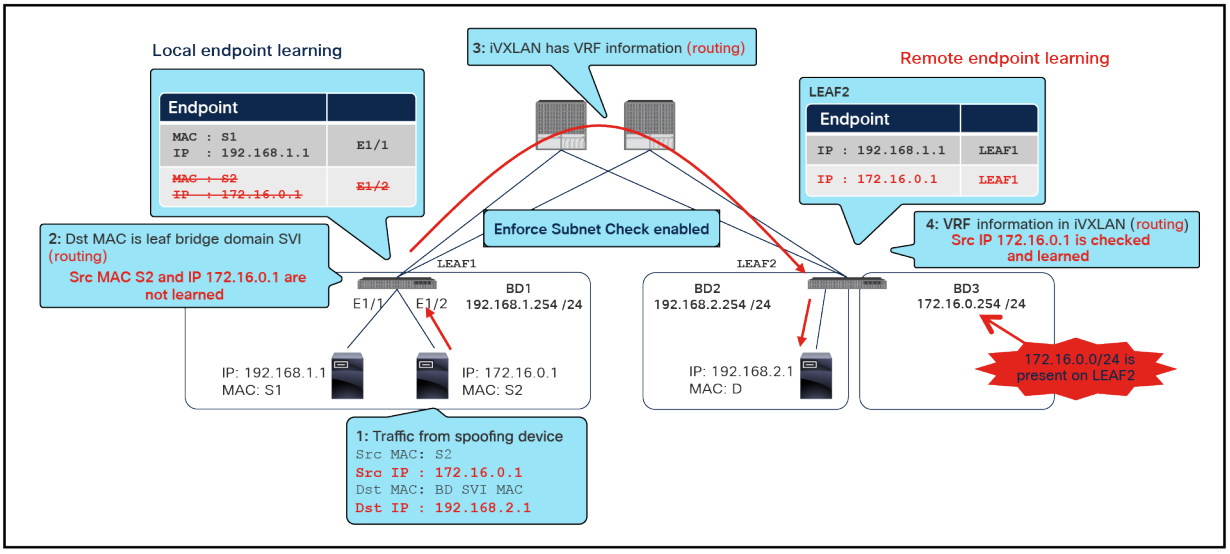


Figure 57.
Enforce Subnet Check example 2

Figure 58 shows an example in which a remote IP endpoint is still learned even though neither the ingress nor the egress bridge domain contains the spoofed IP subnet (172.16.0.1). As mentioned previously, the remote IP endpoint learning check is performed with all bridge domain subnets on the remote leaf under the same VRF instance. In Figure 53, LEAF2 includes BD3 with 172.16.0.0/24 configured, which is why unexpected remote IP endpoint learning is not prevented in this scenario. A common reason that LEAF2 includes this bridge domain subnet is that static binding or the Virtual Machine Manager (VMM) domain for BD3 may be configured on LEAF2 ports. Another reason is that an EPG in BD2 on LEAF2 may have a contract with another EPG in BD3 on another leaf; because of the contract, LEAF2 installs a route for BD3 subnets, called a pervasive route, so that EPG in BD2 on LEAF2 can be routed to BD3 on another leaf.

A note for enabling Enforce Subnet Check

When Enforce Subnet Check feature is enabled, be aware of the following defect.

CSCvh17285: Endpoint learning from ARP stops working on L2 BDs with Enforce Subnet Check Enabled

Due to the defect, when Enforce Subnet Check is enabled, ACI will not be able to learn MAC address via ARP/GARP in L2 BD where Unicast Routing is disabled. This will not impact BDs with Unicast Routing enabled. The impact from this defect is ACI may age out the MAC endpoint and will not be able to re-learn it in the L2 BD unless there is non-ARP traffic coming in from the MAC address. Another typical impact is ACI will not be able to detect MAC move even when a host sends a GARP to let ACI know about its movement.

These impact from the defect can be mitigated by setting L2 Unknown Unicast Flood in the L2 BD. Although the MAC learning in the L2 BD from ARP/GARP will still not occur due to the defect even with L2 Unknown Unicast Flood, ACI will be able to forward the traffic to the MAC by flooding it even after the MAC ages out due to this defect. If L2 Unknown Unicast is set to Hardware-Proxy and the MAC ages out, ACI will not be able to find the MAC endpoint in SPINE COOP database and have to drop the packet. Also L2 Unknown Unicast Flood option in the L2 BD is generally recommended as a best practice.

IP Aging Policy

The IP Aging Policy was first introduced in APIC Release 2.1(1h) with the following enhancement:

CSCut23815 ACI: unused local IP endpoint should be aged out separately from its MAC endpoint

This configuration is disabled by default to keep the same behavior with the older release.

For APIC Release 2.0, this option is located at Fabric > Access Policies > Global Policies > IP Aging Policy (Figure 58). For APIC Release 3.0(1k) and later, it is located at System > System Settings > Endpoint Controls > IP Aging (Figure 59).

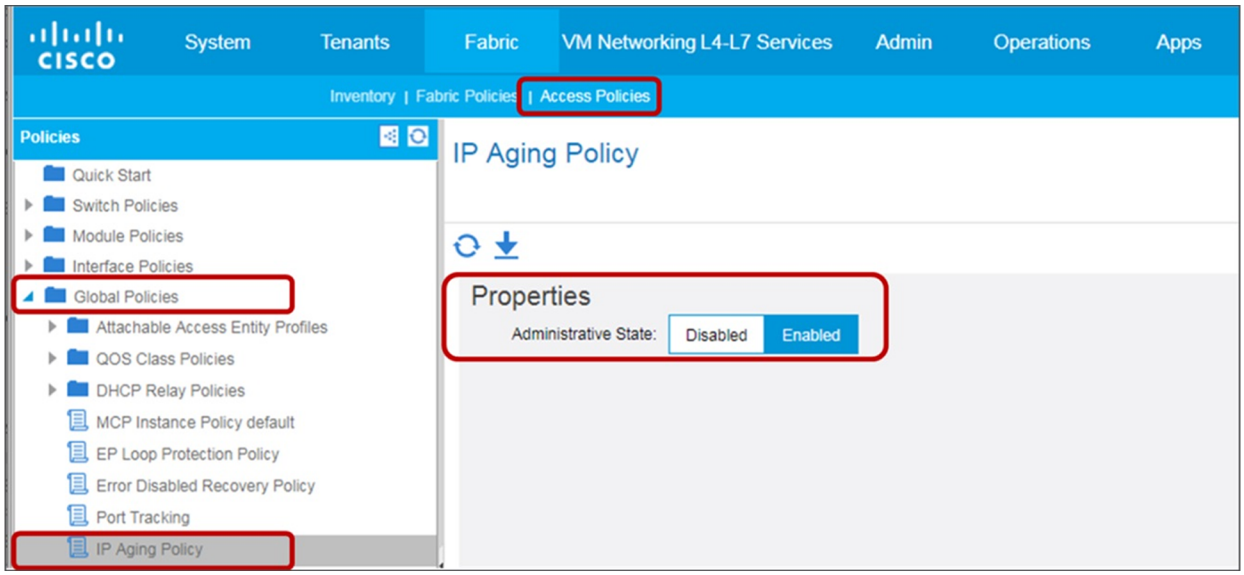


Figure 58.
IP Aging (APIC Release 2.0)

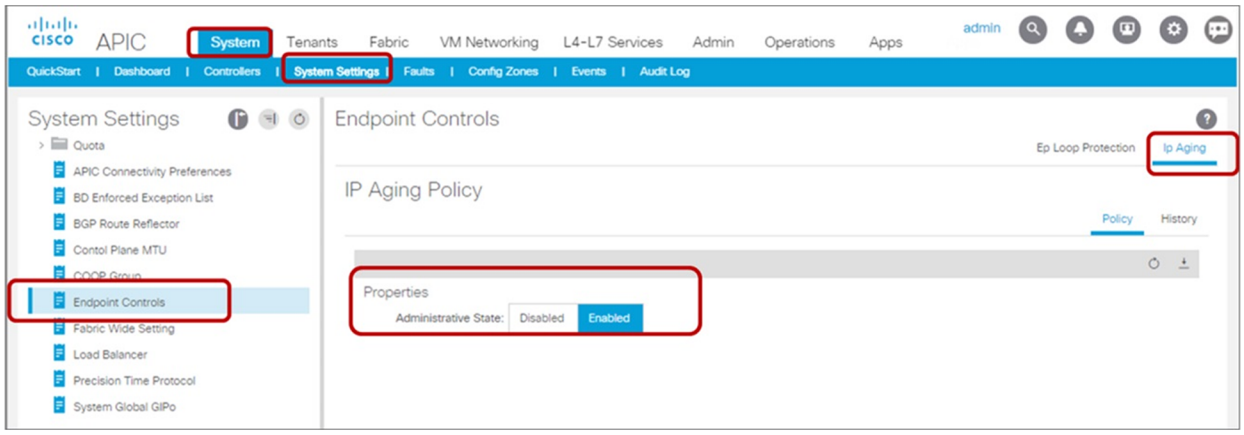


Figure 59.
IP Aging (APIC Release 3.0)

The IP aging policy tracks and ages unused IP addresses on an endpoint. Tracking is performed by using the endpoint retention policy, which is configured for the bridge domain to send ARP requests (for IPv4) and neighbor solicitations (for IPv6) at 75 percent of the local endpoint aging interval. When no response is received from an IP address that IP address is aged out.

IP Aging Policy use case

Before this option was available, an endpoint (such as an interface on a virtual machine) might have unused IP addresses stuck on the same MAC address. For example, when booting, a Microsoft Windows virtual machine that does not receive a Dynamic Host Configuration Protocol (DHCP) address (and does not have a static IP address) will automatically obtain an address from the 169.254.0.0/16 address range, as shown in Figure 60.

Figure 60.
IP aging before address is obtained

At some point, the virtual machine will obtain a routable address, and the endpoint will then consist of one MAC address and two IP addresses, as shown in Figure 61.

Figure 61.
IP aging after address is obtained

The potential problem in these examples (prior to the IP Aging Policy) is that Cisco ACI could allow stale IP components of the endpoint to be retained indefinitely (or until someone manually clears the entry on the leaf).

Cisco ACI would see the endpoint as all three components (the MAC, IP1, and IP2 addresses). If traffic is received from any one of these components, the entries for all three would be kept active.

Now that IP Aging is available, Cisco ACI will send a unicast ARP packet at 75 percent of the configured endpoint retention timer for all IP addresses that belong to the endpoint. If no response is received from that particular IP address, it will be aged out of the endpoint table. (Note that the MAC address and responding IP address for the endpoint will be retained.)

Rogue EP Control

Rogue EP Control was first introduced in APIC Release 3.2(1l). This configuration is disabled by default to keep the same behavior with the older release.

This option is located at System > System Settings > Endpoint Controls > Rogue EP Control (Figure 62).

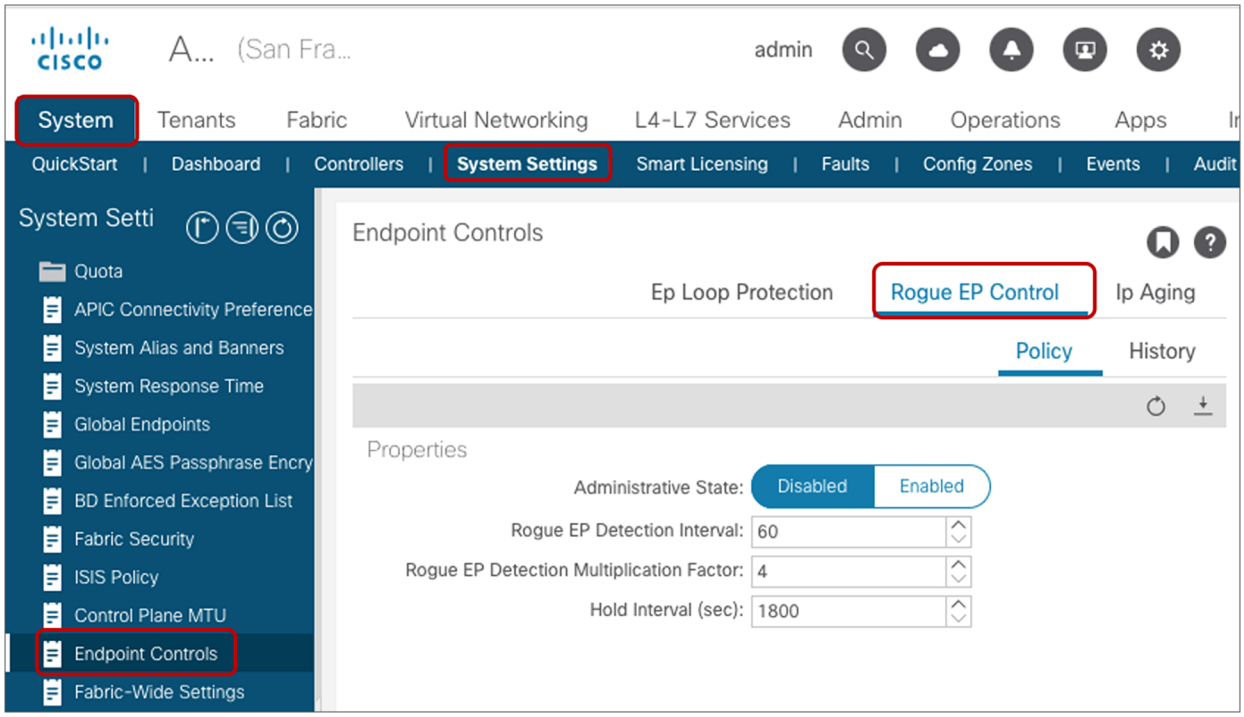


Figure 62.
Rogue EP Control

Rogue EP Control detects an endpoint that moves frequently and pauses the endpoint learning of the specific endpoint. While the endpoint is marked as rogue, the last information before the learning was paused is kept as a static endpoint to prevent abnormal endpoint updates from impacting packet forwarding and resources such as CPU on the leaf switch. The learning of the endpoint will resume after the hold interval.

Although **Rogue EP Control** is enabled globally, each leaf individually tracks the endpoint movements and pauses the endpoint learning. This means that when an endpoint is marked as rogue on leaf1 and the endpoint learning is paused on leaf1, learning of the same endpoint can still happen on other leaf switches unless the other leaf switches also marked the endpoint as rogue from their point of view.

Because **Rogue EP Control** pauses the learning of the specific endpoint as opposed to disabling endpoint learning of the entire bridge domain or shutting down the interface from which the endpoint was learned, the feature affects the traffic only for the endpoint that was marked as rogue without affecting other healthy endpoints in the same bridge domain or on the same interface.

Rogue EP Control raises a fault when an endpoint is marked as rogue to help an administrator to identify the rogue endpoint.

The detection criteria can be configured by using the following values:

- Rogue EP Detection interval: to specify the time in seconds to detect rogue endpoints. The default is 60 seconds. The supported range is 30 to 3600 seconds.
- Rogue EP Detection Multiplication Factor: The endpoint is declared rogue if the endpoint moves more than this number within the Rogue EP Detection interval. The default is 4. Starting from ACI 5.2(3), the new default is 6. When upgrading to 6.0(3) or newer, and the value is set to the old default value 4, it's automatically updated to 6 after the upgrade. The supported range is 2 to 10. When a value larger than 10 is configured, fault F3146 is raised, and switches use the value 10.
- Hold Interval: the amount of time the endpoint is being handled as rogue and kept as the static endpoint. After this interval, the endpoint is deleted. The default is 1800 seconds (30 minutes). The supported range depends on the release. With ACI releases prior to ACI 5.2(3) the configurable range is 1800 to 3600. Starting with ACI 5.2(3) you can configure a minimum hold interval of 300 seconds (5 minutes).

For example, if the Rogue EP Control is enabled with the default configuration parameters above, the ACI fabric declares an endpoint rogue if the endpoint moves more than four times (six after ACI 5.2(3)) in 60 seconds and disables learning for the endpoint for 1800 seconds. The rogue endpoint will be static on the leaf node, interface, and VLAN where it was detected right before the declaration of rogue.

After the hold-interval, rogue endpoints will be deleted.

During the hold time, it is possible that the endpoints that were quarantined on the bridge domain may experience traffic disruption. If you want to reduce the potential downtime to last less than the hold time you can delete the Rogue endpoints manually even before the hold interval expires by using the option Clear Rogue Endpoints at Fabric > Inventory > Pod_number > Leaf_name (Figure 63).

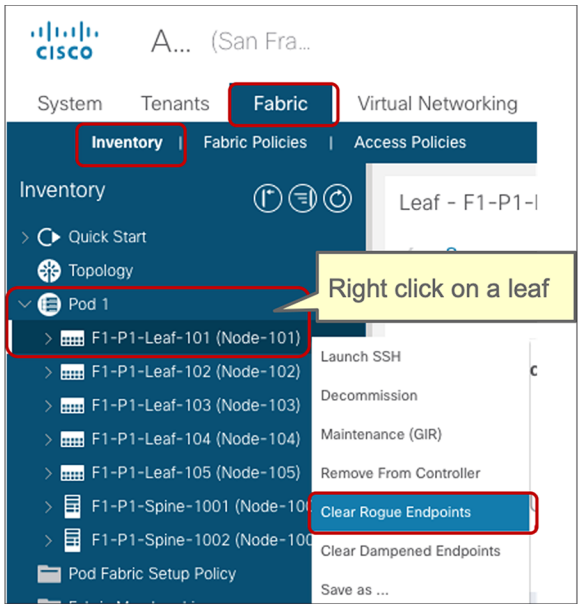


Figure 63.
How to clear rogue endpoints on a leaf

Rogue EP Control protects the ACI fabric and limits the impact of temporary loops by quarantining endpoints only in the bridge domain where they occur, but it can also cause unnecessary glitches when for instance a firewall fails over and for a short time two firewalls may send traffic with the same MAC address before converging. In order to address this requirement starting from Cisco ACI Release 5.2(3) it is possible to configure an “exception” list of MAC addresses to which the Rogue EP Control policy is not applied (or in fact it is applied but it is not as strict).

This option can be used only If the BD is a L2 BD (i.e. if the BD is not configured for IP routing) and it is found under Tenant > Networking > Bridge Domains > BD > Policy > Advanced/Troubleshooting Tab, where it is called Rogue/COOP Exception List.

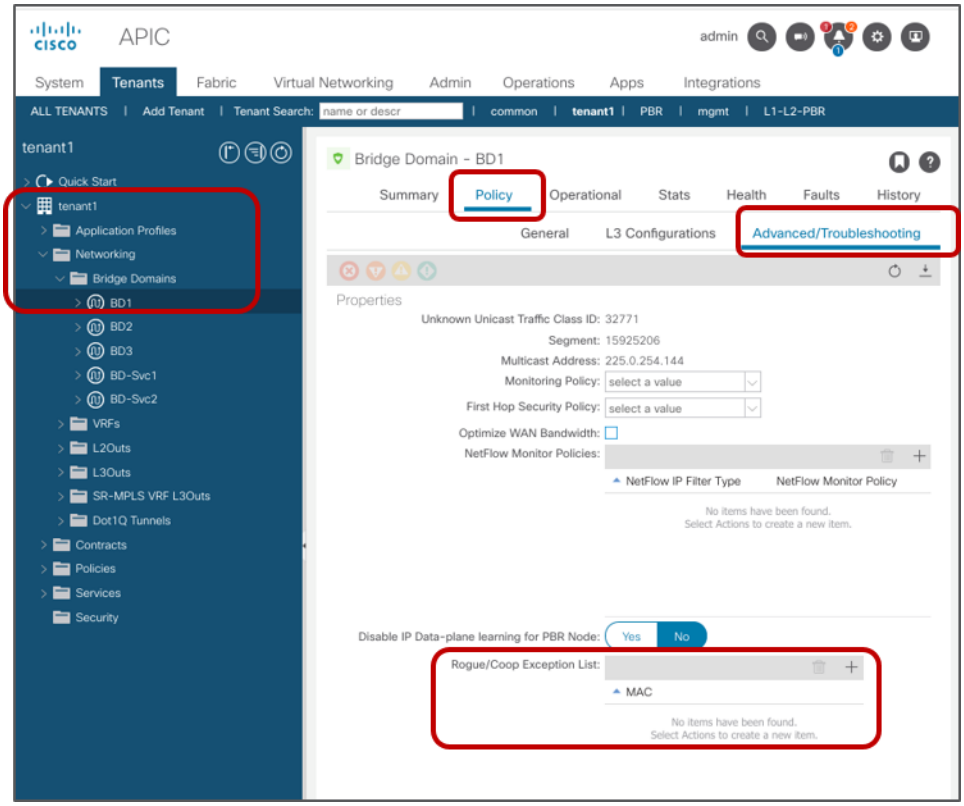


Figure 64.
Rogue/COOP Exception List

The MAC addresses that are entered in the BD(s) in the Rogue/Coop Exception List benefit from a less strict Rogue EP control policy, but if they move more than 3000 times in 10 minutes they are also going to be quarantined.

The reason why this option is only applicable to L2 BDs is because if there is routing enabled and ACI leaf nodes detect IP moves they may quarantine the endpoint IP even if the MAC is in the exception list. If you need to “disable” Rogue EP on IP addresses that are moving too often because of a valid design reason (and not because of a misconfiguration), you can disable IP Data-plane learning either per-VRF, per-BD-subnet or per-EPG. This configuration will prevent the traffic from the same IP that keeps flapping between ACI leaf nodes ports from continuously updating the endpoint IP information, hence Rogue EP Control won't take effect. If instead the reason for the IP address flapping is not due to data plane traffic, but to continuous ARP responses from different hosts/MAC addresses, Rogue EP Control will still take effect.

As of ACI 5.2(3) you can only enter a total of 100 MAC addresses to be excluded from Rogue EP Control. This is a global limit, not a per-BD limit, which means that the sum of all MAC addresses entered in the Rogue/COOP Exception List across all BDs in the ACI fabric must be less or equal to 100.

Rogue EP Control use case

Rogue EP Control is meant to protect the ACI fabric against issues such as a specific flapping endpoint due to inappropriate configurations or designs. This is to protect also the ACI control plane from having to manage too many endpoint moves, which could be caused by L2 loops. The figure below illustrates an example. The server is mistakenly configured as active/active NIC teaming, which sends packets with the same source IP from multiple NICs with different MAC addresses. This causes the endpoint flaps between leaf nodes. Without the Rogue EP Control, although Endpoint Move Dampening via Endpoint Retention Policy might be able to detect such frequent endpoint moves, it will disable learning on the entire bridge domain, which affects other endpoints in the same bridge domain.

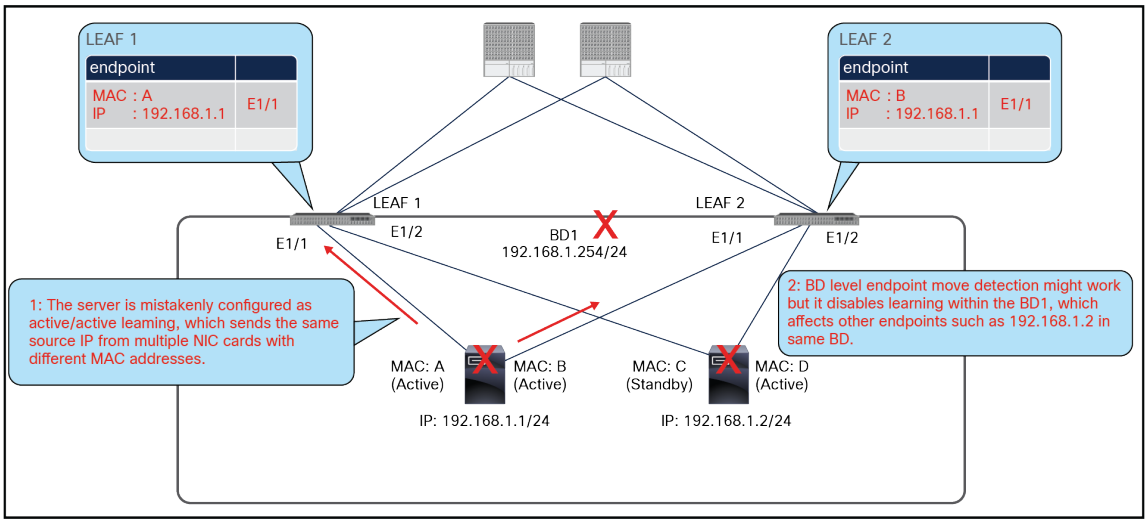


Figure 65.
Why Rogue EPG Control is required

With the Rogue EP Control enabled, once the endpoint is marked as rogue, a fault is raised and learning is disabled for the endpoint only, which allows other endpoints in the same bridge domain to function as usual as shown in the figure below. After the administrator identifies the issue based on the fault, the design that caused this issue can be corrected, for example by changing the NIC teaming configuration or disabling IP Data-plane Learning on the VRF.

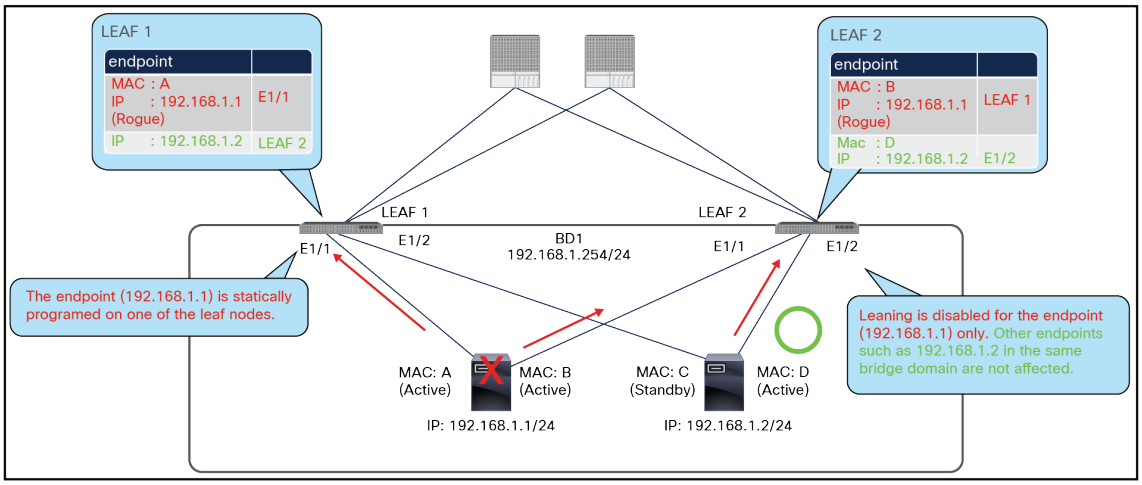


Figure 66.
With Rogue EPC Control enabled

Considerations for enabling Rogue EP Control are as follows:

- If the Rogue EP Control is enabled, Ep Loop Protection and Endpoint Move Dampening via Endpoint Retention Policy will not take effect.
- It doesn't distinguish between local or remote moves; any type of interface change is considered an endpoint move.
- The endpoint moves for Rogue EP Control are counted for MAC and IP addresses separately, even though an endpoint in ACI contains both MAC and IP addresses. This is because IP addresses may move around, and be learned on multiple MAC addresses while the MAC addresses themselves didn't move. When the move count of an IP address exceeds the threshold, only the IP address is marked as rogue. When the move count of a MAC address exceeds the threshold, the MAC and any IP addresses associated to the MAC at that time are marked as rogue.
- Changing the Hold Interval will not affect existing rogue endpoints' hold timer.
- This feature works within a site. This feature is not designed for detecting endpoint moves between sites or between the main location and the remote leaf location.
- Disabling the Rogue EP Control will clear all existing rogue endpoints.
- Because of caveats in APIC Release 4.1, you must disable Rogue EP Control before upgrading to or from APIC Release 4.1.

Ep Loop Protection

Ep Loop Protection is disabled by default.

Prior to APIC Release 3.0, this option is located at System > System Settings > Endpoint Controls > Ep Loop Protection. For APIC Release 3.0 and later, this option is located at System > System Settings > Endpoint Controls > Ep Loop Protection (Figure 67). The hold interval is the one defined in the BD level End Point Retention Policy.

Now that Rogue EP Control is available after APIC Release 3.2, Rogue EP Control instead of Ep Loop Protection is recommended.

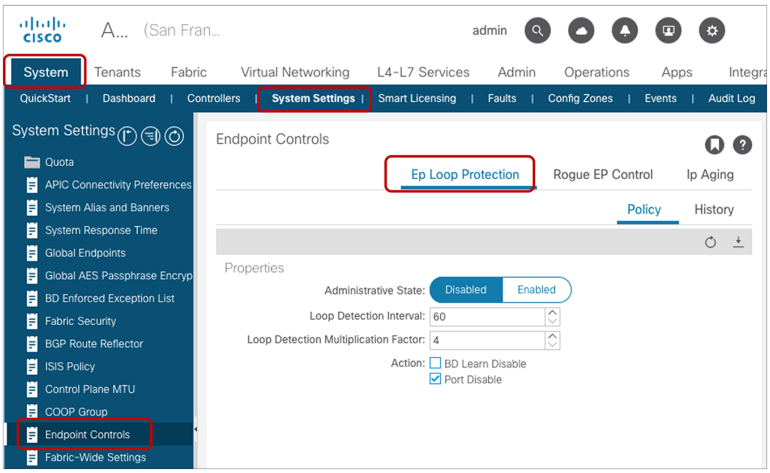


Figure 67.
Ep Loop Protection

While Ep Loop Protection has similar intentions (that is, finding a loop and disabling learning) to Rogue EP Control, it disables learning on the bridge domain and/or disables the port (error-disable), which could affect healthy endpoints connected to the same interface or the same bridge domain. Thus, use of Rogue EP Control instead of Ep Loop Protection is recommended after APIC Release 3.2. Please be aware that the detection mechanism of Ep Loop Protection is strictly for a loop, because it triggers the move only when the move is between the same two interfaces. Ep Loop Protection cannot detect movements for an endpoint that moves around randomly.

COOP Endpoint Dampening

The COOP Endpoint Dampening was introduced in APIC Release 4.2(3). This configuration is enabled by default.

This option is not available on GUI as of APIC Release 5.0(1). To disable COOP Endpoint Dampening, disableEpDampening needs to be set to “true” via API.

Post URL: https://APIC IP/api/policymgr/mo/.xml

Body:

```
<polUni>
    <infraInfra>
        <infraSetPol disableEpDampening="true"></infraSetPol>
    </infraInfra>
</polUni>
```

COOP Endpoint Dampening is used to mitigate the impact of unreasonable amounts of endpoint updates on spine nodes by ignoring the endpoint updates of the particular endpoint only. When a spine node identifies a dampened endpoint, a fault is raised, and the spine notifies all leaf nodes to ignore the update from the endpoint.

Detection criteria are based on the penalty value calculations that are based on types of endpoint-related events, as shown in Table 7. The penalty value is decreased by 50 percent every five minutes.

Table 7. Penalty calculation

Event	Penalty value	Note
Learn new IP address	0	New IP address is learned
Learn additional IP address	2	Additional IP address is learned with an existing endpoint MAC.
Delete IP address	50	Delete the remote endpoint IP address after the IP address is learned
Learn the deleted IP address	50	Learn the remote endpoint IP address after the IP address is deleted
Delete IP address	400	Delete the local endpoint IP address after the IP address is learned
Learn the deleted IP address	400	Learn the local endpoint IP address after the IP address is deleted
Endpoint move	200	An endpoint moves to different interface.
IP move	200	An IP moves to different MAC The penalty is high for this event because it causes two route updates to BGP
URIB programming	50	Spine tunnel interface status change (up/down) for an endpoint

The penalty value is calculated per IP address. For example, if the penalty value of the endpoint is 4000 and the number of IP addresses in the endpoint is 2, the penalty value per IP is 4000/2 = 2000. When the penalty value per IP exceeds the critical threshold (4000), the endpoint status is changed to “Critical” from “Normal.” If an endpoint stays in a “Critical” state for more than five minutes or the penalty value per IP address exceeds the freeze threshold (10000), the endpoint status will go into a “Freeze” (dampening) state, and the update of the endpoint will be ignored. When the penalty value per IP address becomes below the reuse threshold (2500), the endpoint status becomes “Normal” (nondampening). It requires ten minutes for the penalty value to be decreased by 75 percent (10000 * 0.5 * 0.5 =2500). The thresholds are not user-configurable values.

The dampened endpoints can be manually recovered via GUI at Fabric > Inventory > Pod_number > Leaf_name or Spine_Name > Clear Rogue Endpoints (Figure 63). This operation has to be executed on all spine nodes and on the source leaf node of the endpoint. If the dampened endpoint is still in the endpoint table on the leaf, the endpoint is published to the spine COOP database. If not, the dampened endpoint is deleted on the spine COOP database after two minutes.

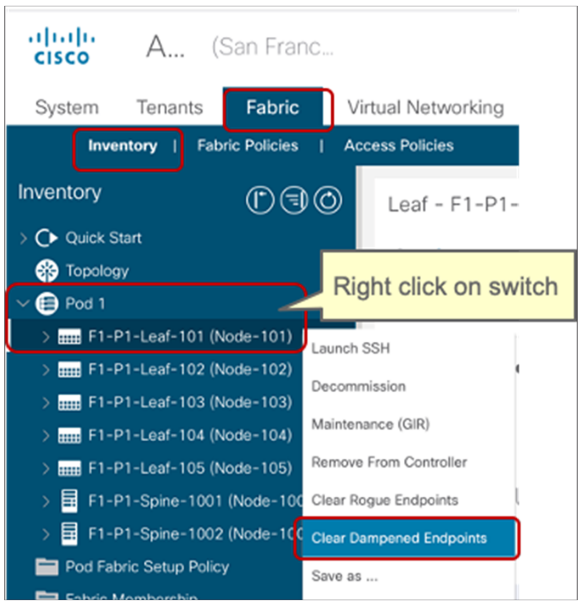


Figure 68.
How to clear COOP Dampened Endpoints

COOP Endpoint Dampening use case

COOP Endpoint Dampening is meant to protect the spine nodes control plane from having to manage too many endpoint updates, which could be caused by multiple IPs on an endpoint. The figure below illustrates an example. The malicious endpoint has a lot of IP addresses using the same MAC address, which sends packets with the same source MAC using different source IPs. This potentially can cause many endpoint updates, and Rogue Endpoint Control might not be able to detect such frequent endpoint updates because these are not endpoint moves.

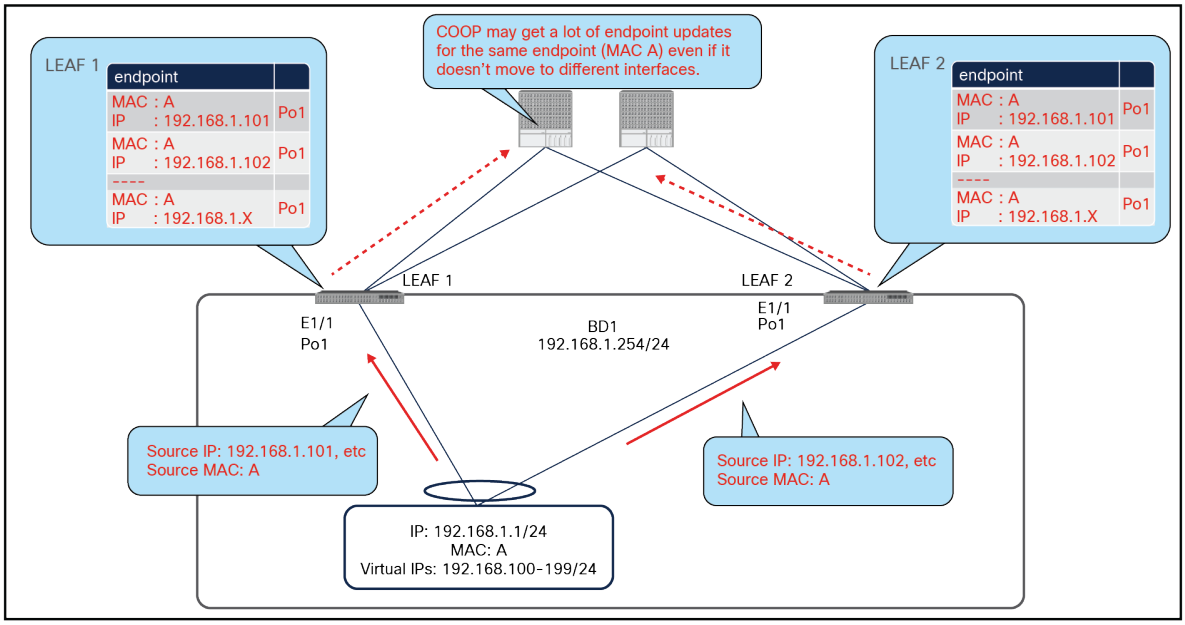


Figure 69.
Why COOP Endpoint Dampening is required

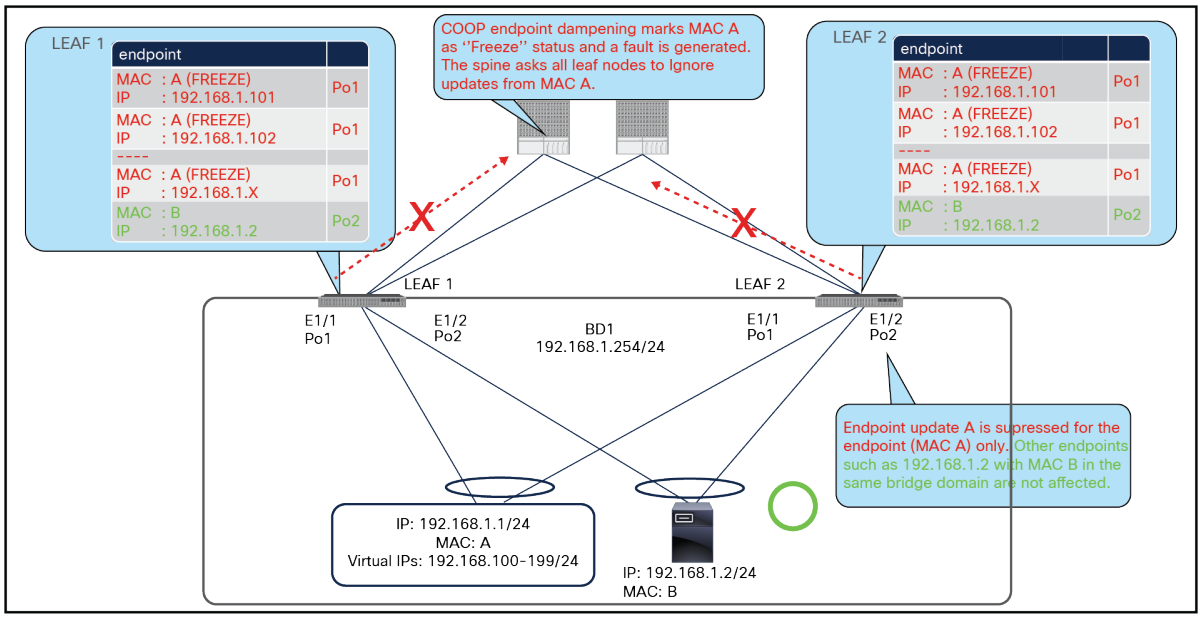


Figure 70.
With COOP Endpoint Dampening enabled

Considerations for COOP Endpoint Dampening are as follows:

- This feature works within a site that includes an ACI Remote Leaf.
- Disabling COOP Endpoint Dampening will recover all existing dampened endpoints that are in “Freeze” state. Endpoints in “Critical” state will be “Normal” when the penalty value decreases to the Reuse threshold.

Best practices for configuring endpoint learning on Cisco ACI

Cisco ACI fundamentally handles endpoint learning in a different manner than traditional network devices. This difference gives Cisco ACI the unique advantage of being able to limit flooding of ARP, unknown unicast, and other traffic types. As Cisco ACI has evolved, the best way to configure Cisco ACI has evolved as well. This section presents a list of recommended configurations for endpoint learning that you should use, depending on the hardware that you have installed.

For optimal fabric operations, you should use settings that cause Cisco ACI to learn only IP addresses that are configured on a bridge domain subnet. The options you use to enable the desired behavior depends on the generation of Cisco ACI leaf switches in your fabric.

First-generation leaf switches

For first-generation leaf switches, the following configurations are recommended for optimal endpoint update and forwarding behavior:

- Bridge domain–level configurations
  - Limit IP Learning to Subnet
- Fabric-level configurations
  - IP Aging Policy
  - Disable Remote EP Learn (on border leaf)
    - Prior to Cisco ACI Release 3.0(2h), prerequisite is to set Tenant > Networking > VRFs > Policy Control Enforcement to Ingress on your VRF instances

Second-generation leaf switches

For second-generation leaf switches, the following configurations are recommended for optimal endpoint update and forwarding behavior:

- Fabric-level configurations
  - IP Aging Policy
  - Disable Remote EP Learn (on border leaf)
    - Prior to Cisco ACI Release 3.0(2h), prerequisite is to set Tenant > Networking > VRFs > Policy Control Enforcement to Ingress on your VRF instances
    - Only on APIC release prior to the enhancement for endpoint announce (CSCvj17665)
  - Enforce Subnet Check

Second-generation leaf switches don't need Limit IP Learning To Subnet because the Enforce Subnet Check option, which is available only starting from the second-generation switches, is superior to the Limit IP Learning To Subnet feature. Please refer to the Enforce Subnet Check section for details.

Fabrics with both first- and second-generation leaf switches

For fabrics with a mix of first- and second-generation leaf switches, the following configurations are recommended for optimal endpoint update and forwarding behavior:

- Bridge domain–level configurations
  - Limit IP Learning To Subnet
- Fabric-level configurations
  - IP Aging
  - Disable Remote EP Learn (on border leaf)
    - Prior to Cisco ACI Release 3.0(2h), prerequisite is to set Tenant > Networking > VRFs > Policy Control Enforcement to Ingress on your VRF instances
  - Enforce Subnet Check

First-generation leaf switches in a mixed environment will ignore the Enforce Subnet Configuration. However, in a mixed environment, the subnet bridge domain check (which is triggered by the Enforce Subnet Check option) will be enforced on all leaf switches.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C11-739989-24 10/25