

INTUITIVE

Cisco *live!*

5-8 March 2019 • Melbourne, Australia

#CLMEL



BRKACI-3545

Mastering ACI Forwarding Behavior

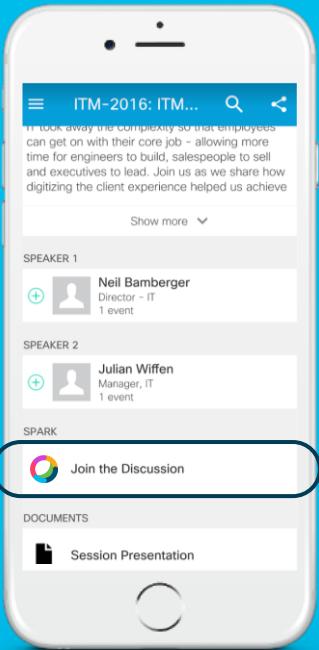
- A day in the life of a packet -

Takuya Kishida, Technical Leader, Service

Cisco *live!*



INTUITIVE



Cisco Webex Teams



Questions?

Use Cisco Webex Teams (formerly Cisco Spark) to chat with the speaker after the session

How

- 1 Open the Cisco Events Mobile App
- 2 Find your desired session in the “Session Scheduler”
- 3 Click “Join the Discussion”
- 4 Install Webex Teams or go directly to the team space
- 5 Enter messages/questions in the team space

cs.co/ciscolivebot#BRKACI-3545

Agenda

- Introduction
 - ACI Overlay VxLAN and TEP
- ACI Forwarding components
 - Endpoints, EPG, EP Learning, COOP and How it all works
 - BD, VRF forwarding scope and detailed options
 - Spine-Proxy and ARP Glean
 - Forwarding Software Architecture and ASIC Generation
- ACI Packet Walk
 - Walk through the life of a packet going through ACI

Basic Acronyms/Definitions

Reference Slide



Acronyms	Definitions
ACI	Application Centric Infrastructure
APIC	Application Policy Infrastructure Controller
EP	Endpoint
EPG	Endpoint Group
BD	Bridge Domain
VRF	Virtual Routing and Forwarding
COOP	Council of Oracle Protocol
VxLAN	Virtual eXtensible LAN

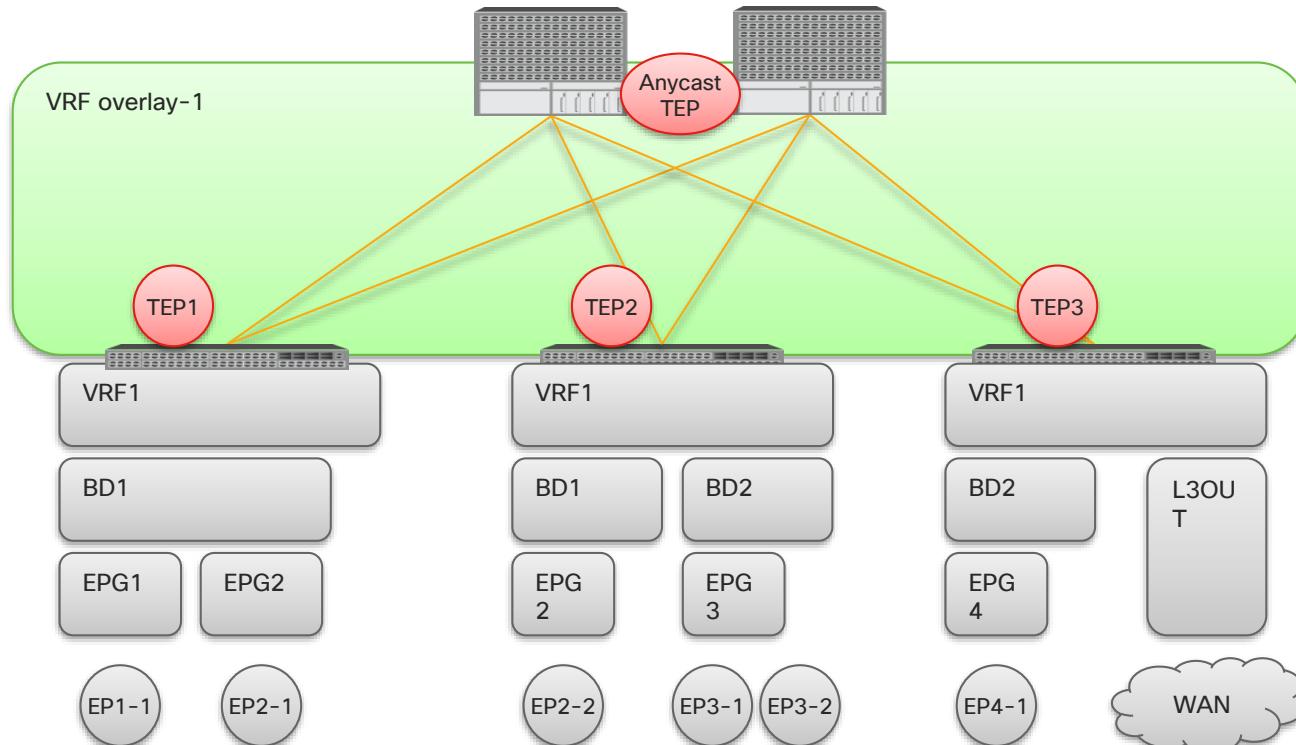
Acronyms	Definitions
dXXXo	Outer Destination XXX (dIPo = Outer Destination IP)
sXXXo	Outer Source XXX (sIPo = Outer Source IP)
dXXXi	Inner Destination XXX (dIPi = Inner Destination IP)
sXXXi	Inner Source XXX (sIPi = Inner Source IP)
GIPo	Outer Multicast Group IP
VNID	Virtual Network Identifier

Agenda

- Introduction
 - ACI Overlay VxLAN and TEP
- ACI Forwarding components
 - Endpoints, EPG, EP Learning, COOP and How it all works
 - BD, VRF forwarding scope and detailed options
 - Spine-Proxy and ARP Glean
 - Forwarding Software Architecture and ASIC Generation
- ACI Packet Walk
 - Walk through the life of a packet going through ACI

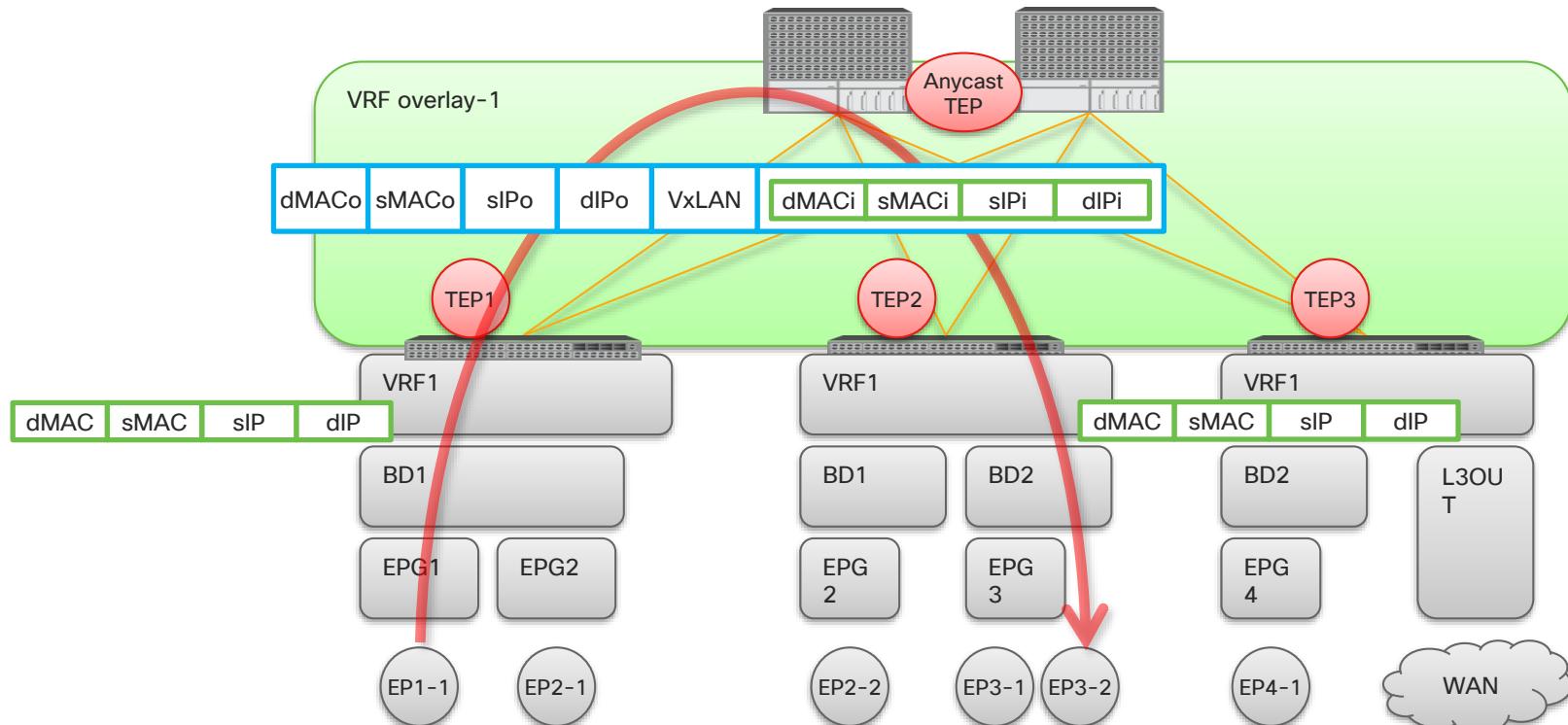
ACI Overlay VxLAN and TEP

※ TEP : Tunnel EndPoint



ACI Overlay VxLAN and TEP

※ TEP : Tunnel EndPoint



ACI Overlay VxLAN and TEP

Scenario 1 : source LEAF knows the destination (on the same LEAF)



Scenario 2 : source LEAF knows the destination (on another LEAF X)



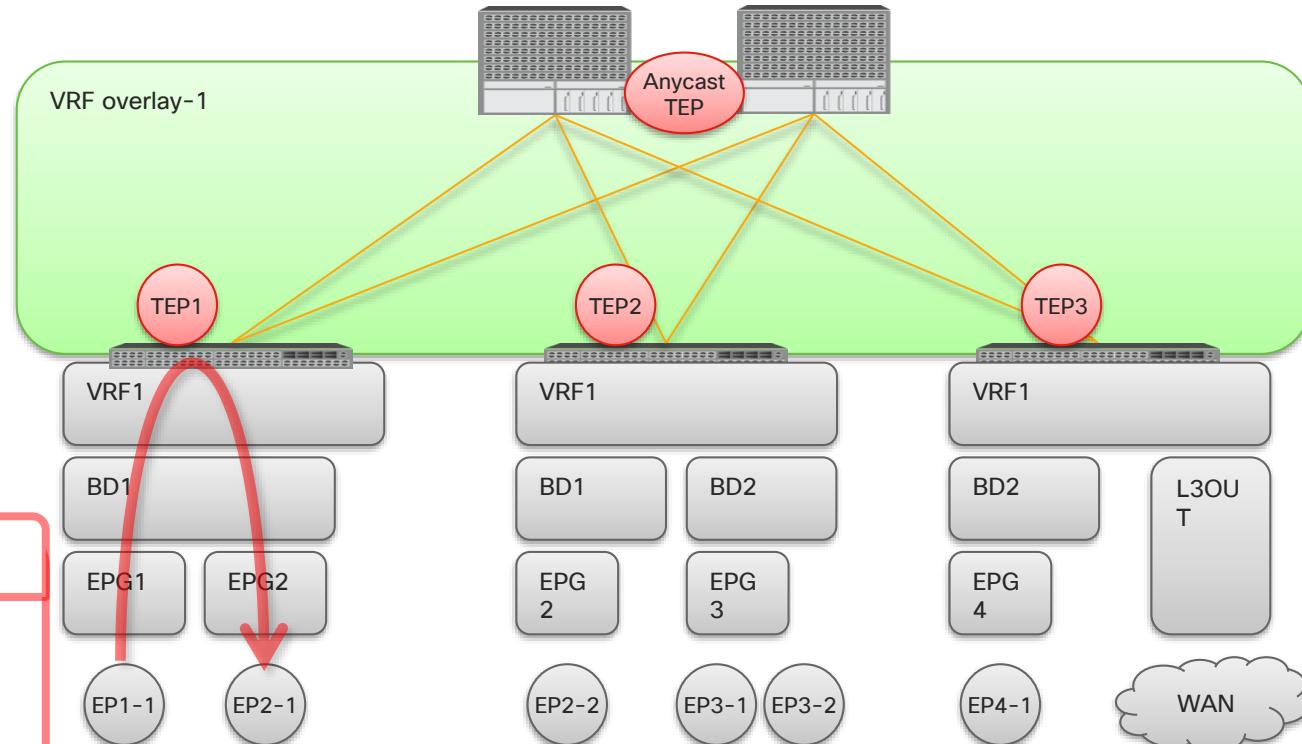
Scenario 3 : source LEAF does NOT know the destination (Spine-Proxy)



Scenario 4 : source LEAF does NOT know the destination (Flood)

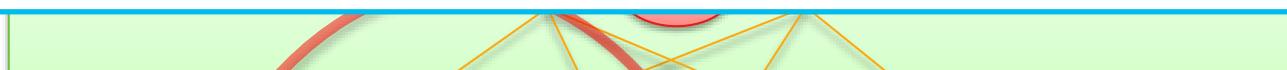


Source LEAF knows the destination (on the same LEAF)



ACI Overlay VxLAN and TEP

Scenario 1 : source LEAF knows the destination (on the same LEAF)



Scenario 2 : source LEAF knows the destination (on another LEAF X)



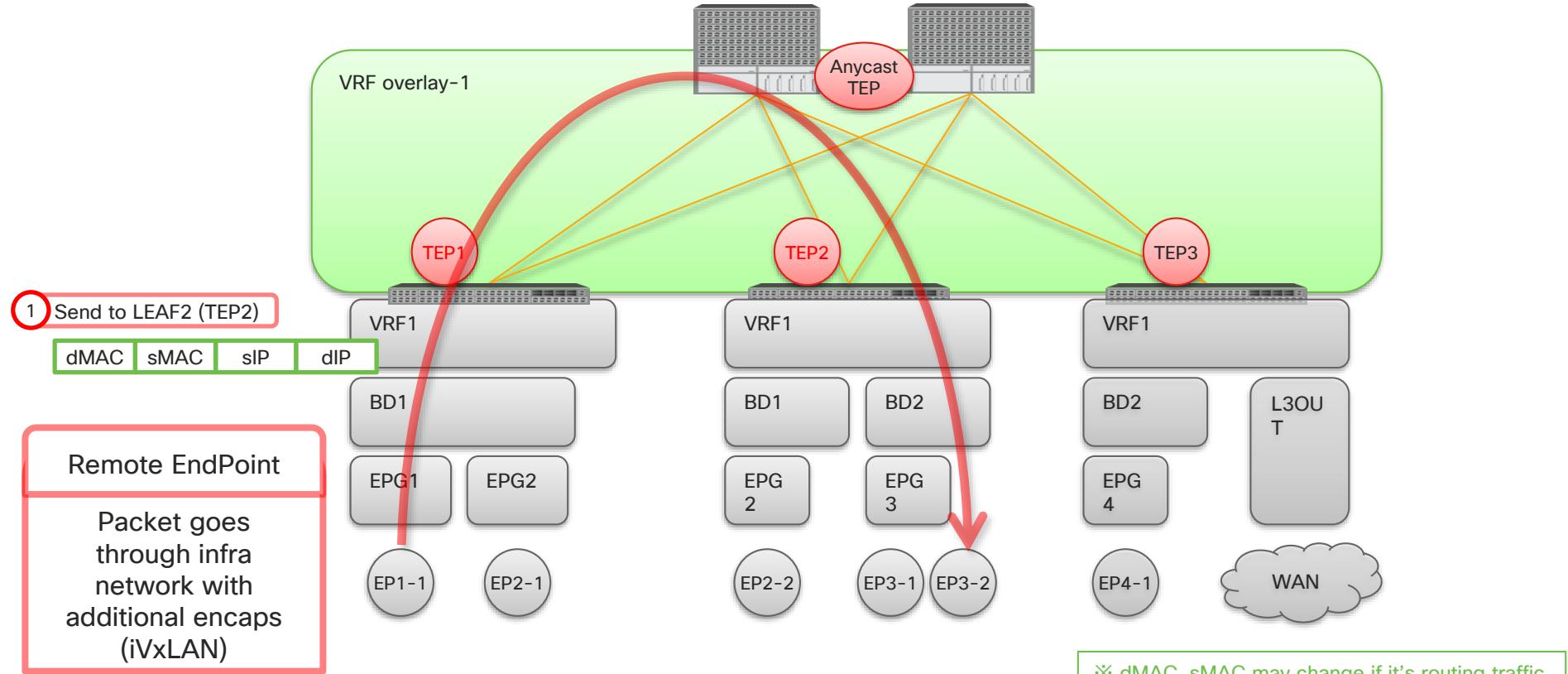
Scenario 3 : source LEAF does NOT know the destination (Spine-Proxy)



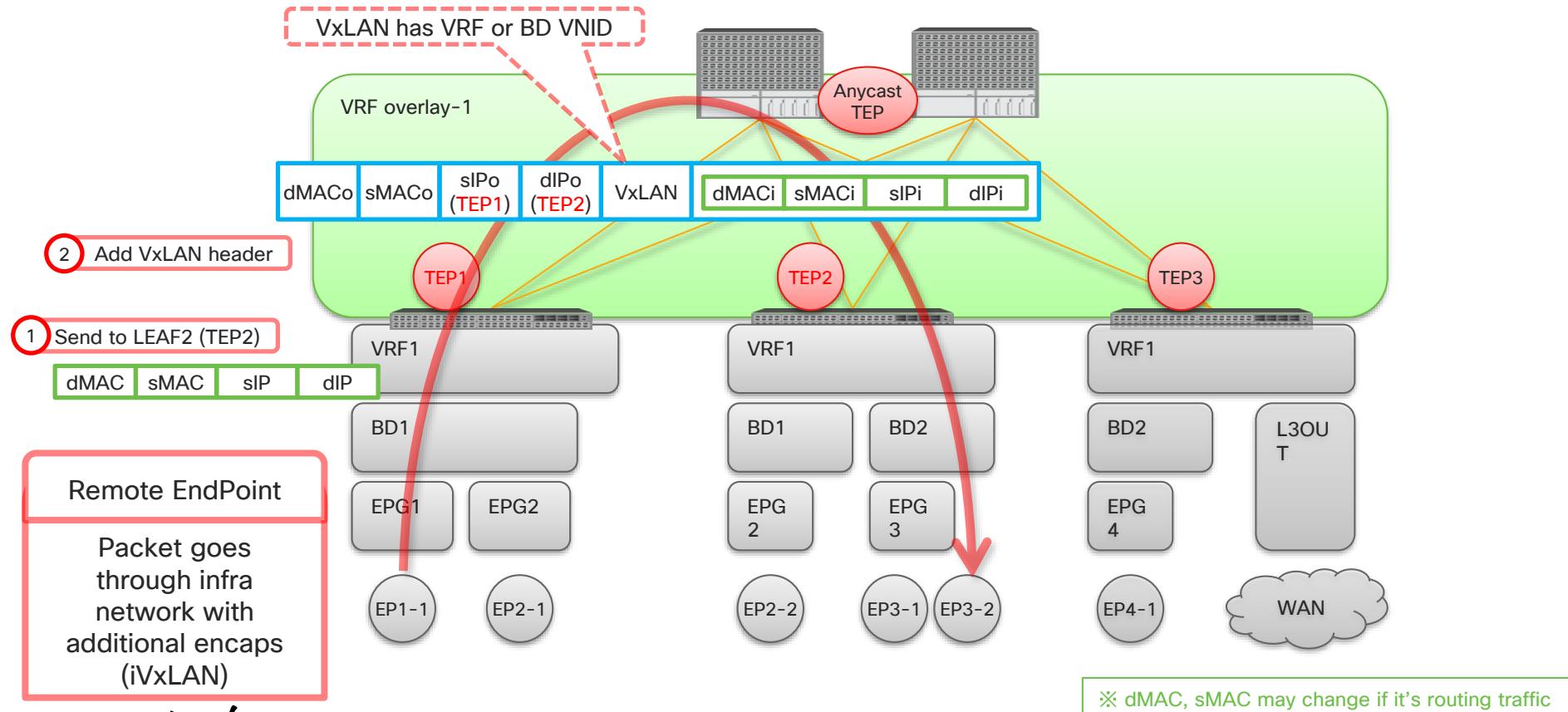
Scenario 4 : source LEAF does NOT know the destination (Flood)



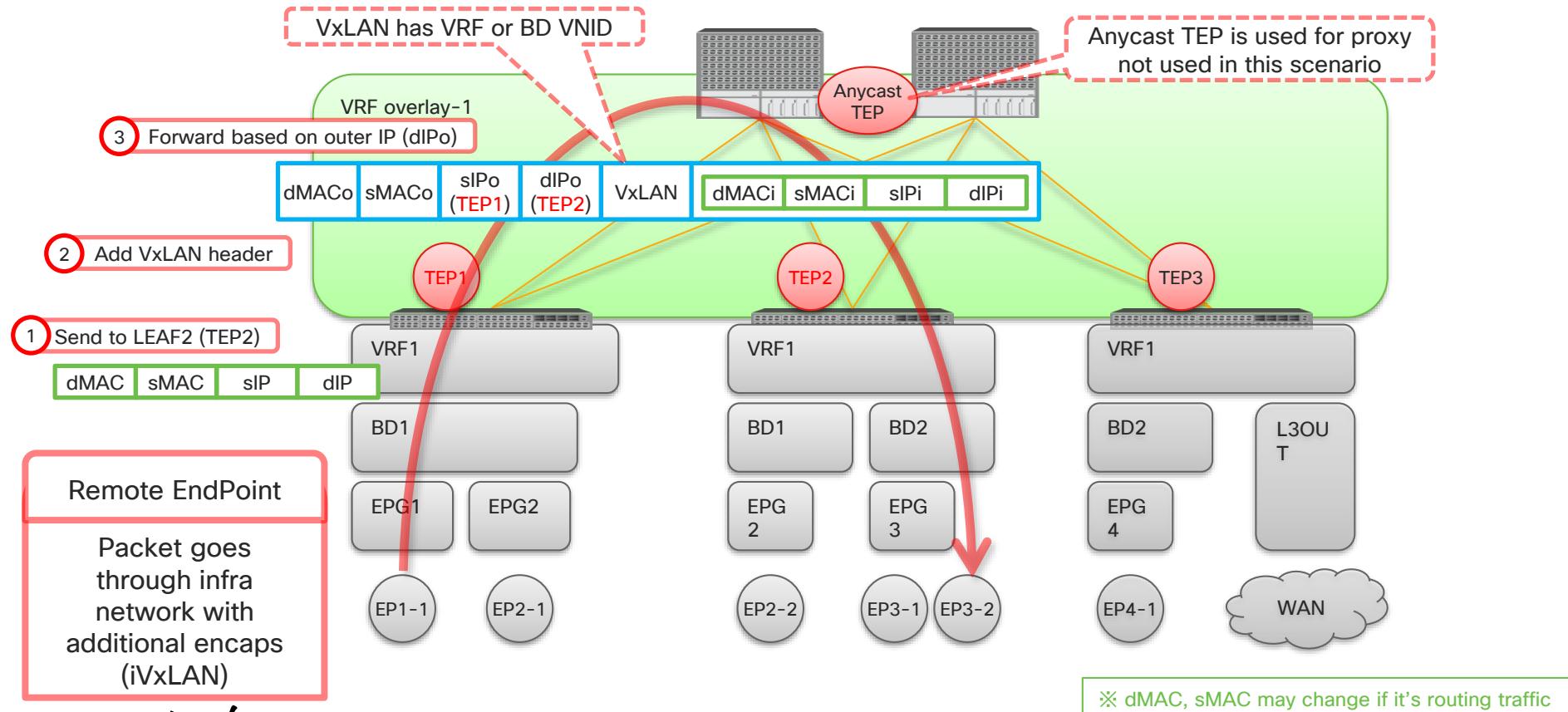
Source LEAF knows the destination (on the remote LEAF)



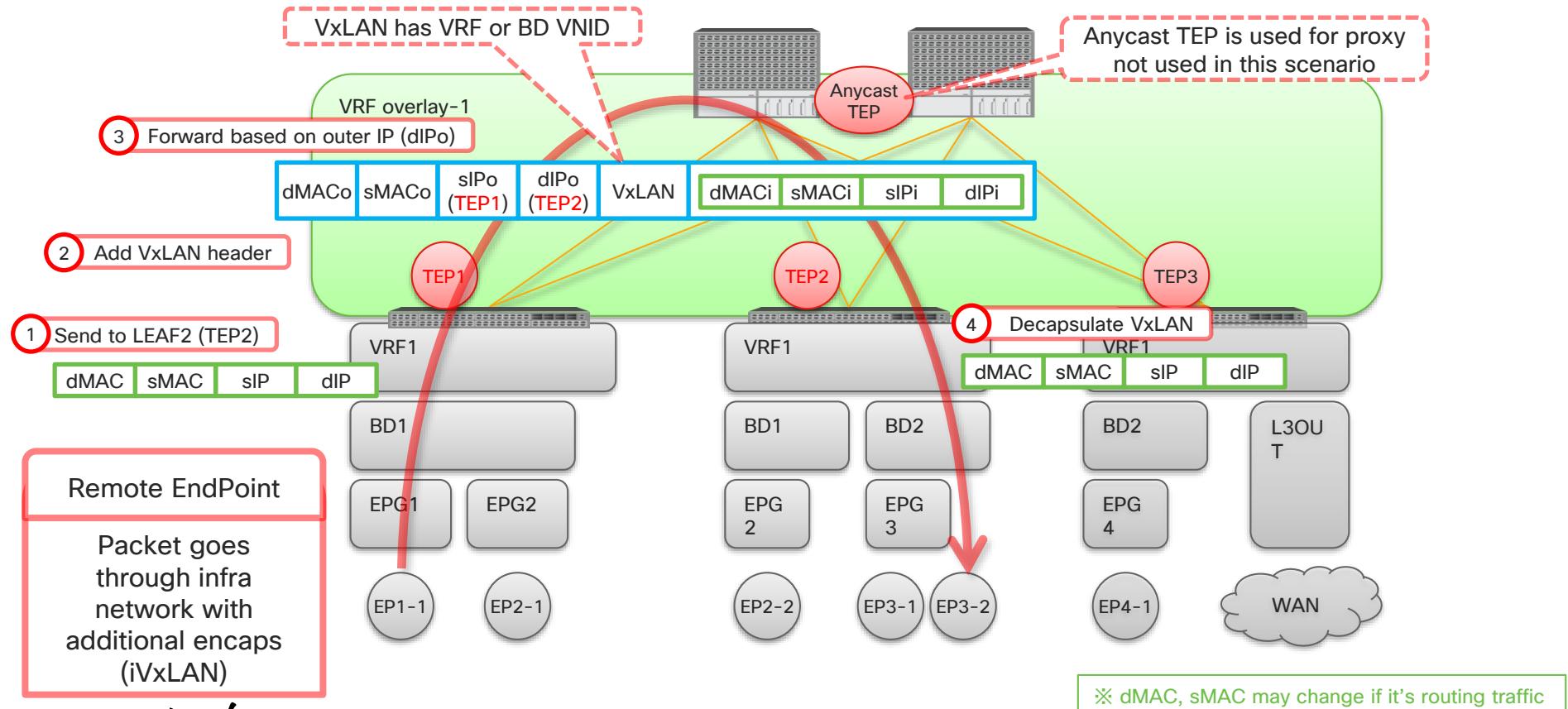
Source LEAF knows the destination (on the remote LEAF)



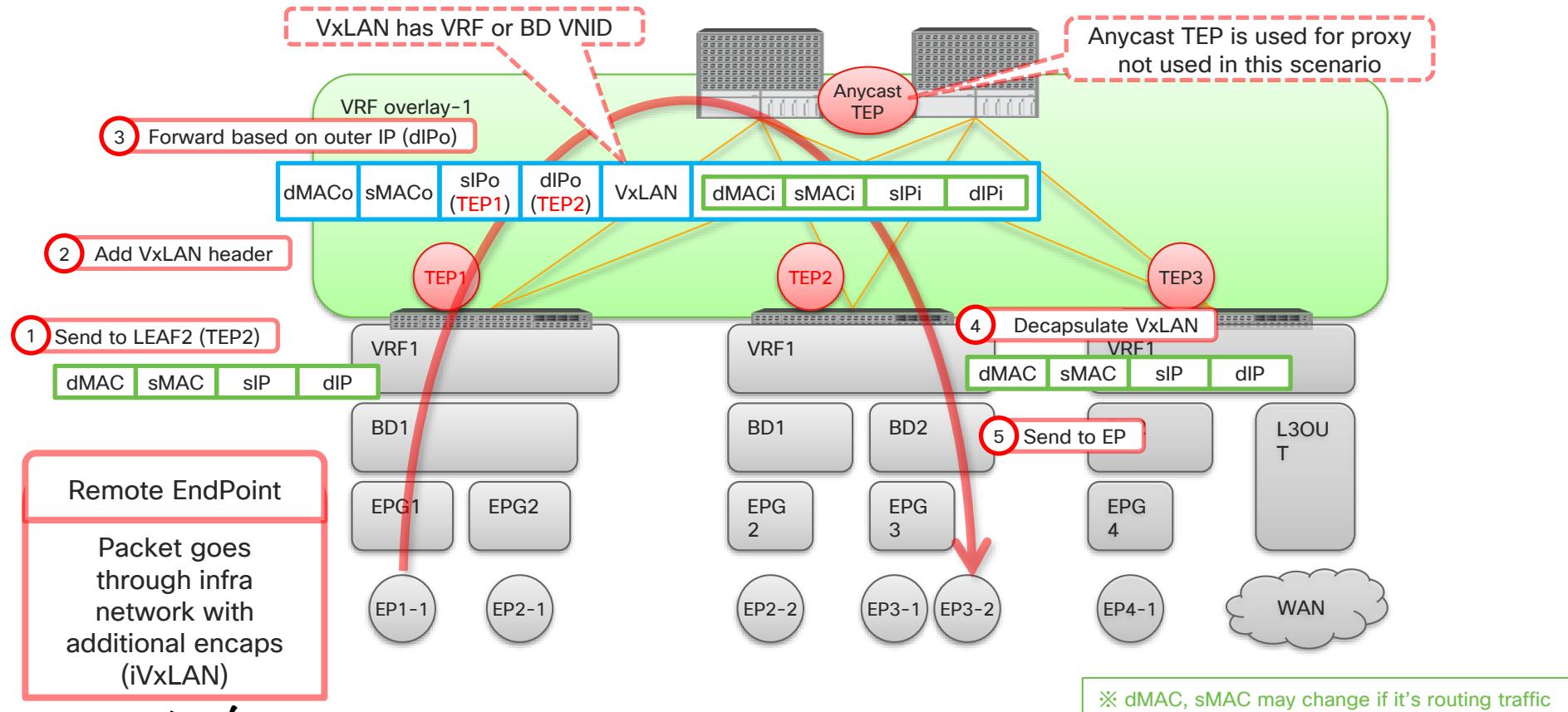
Source LEAF knows the destination (on the remote LEAF)



Source LEAF knows the destination (on the remote LEAF)

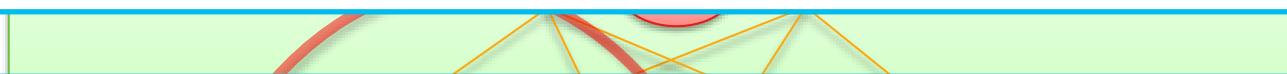


Source LEAF knows the destination (on the remote LEAF)



ACI Overlay VxLAN and TEP

Scenario 1 : source LEAF knows the destination (on the same LEAF)



Scenario 2 : source LEAF knows the destination (on another LEAF X)



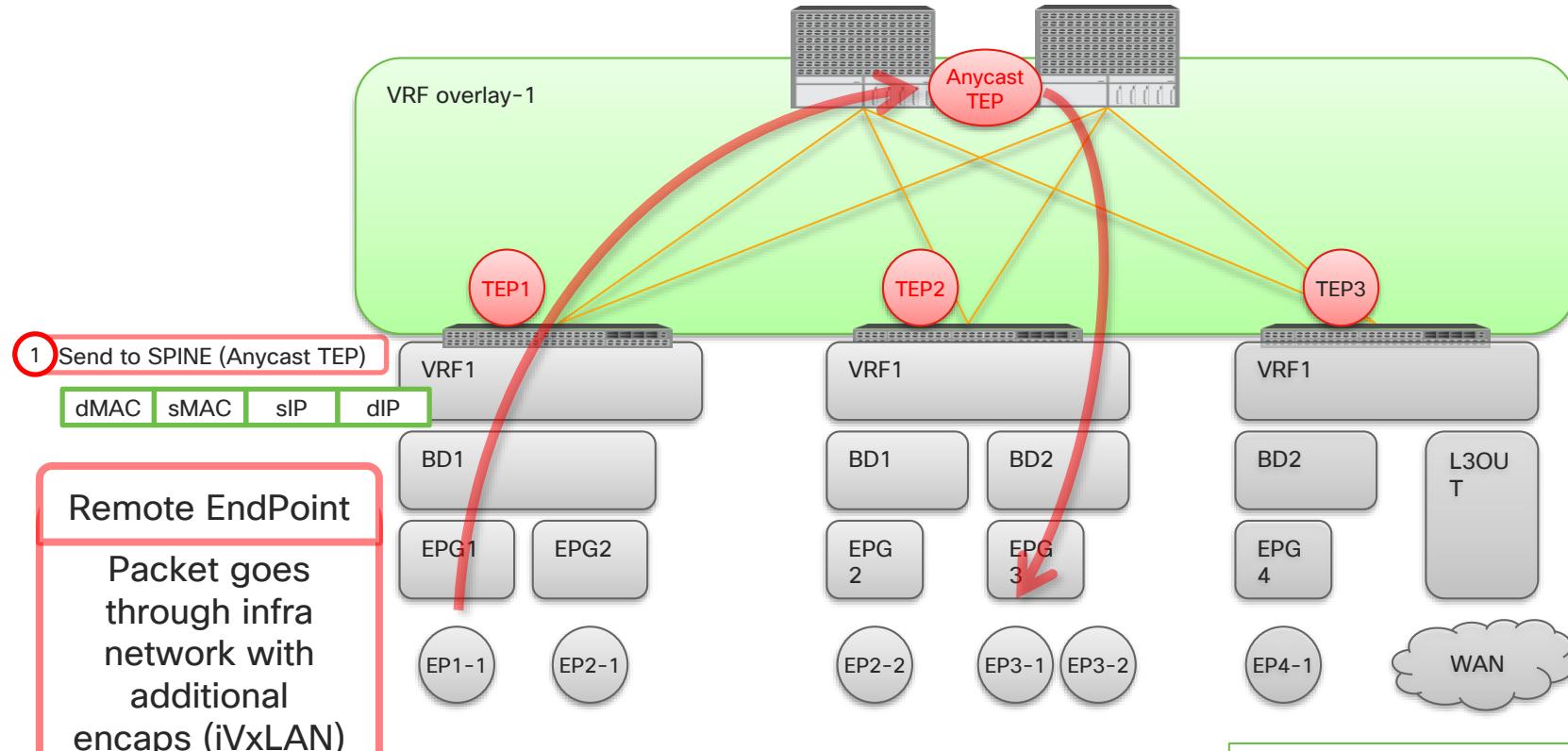
Scenario 3 : source LEAF does NOT know the destination (Spine-Proxy)



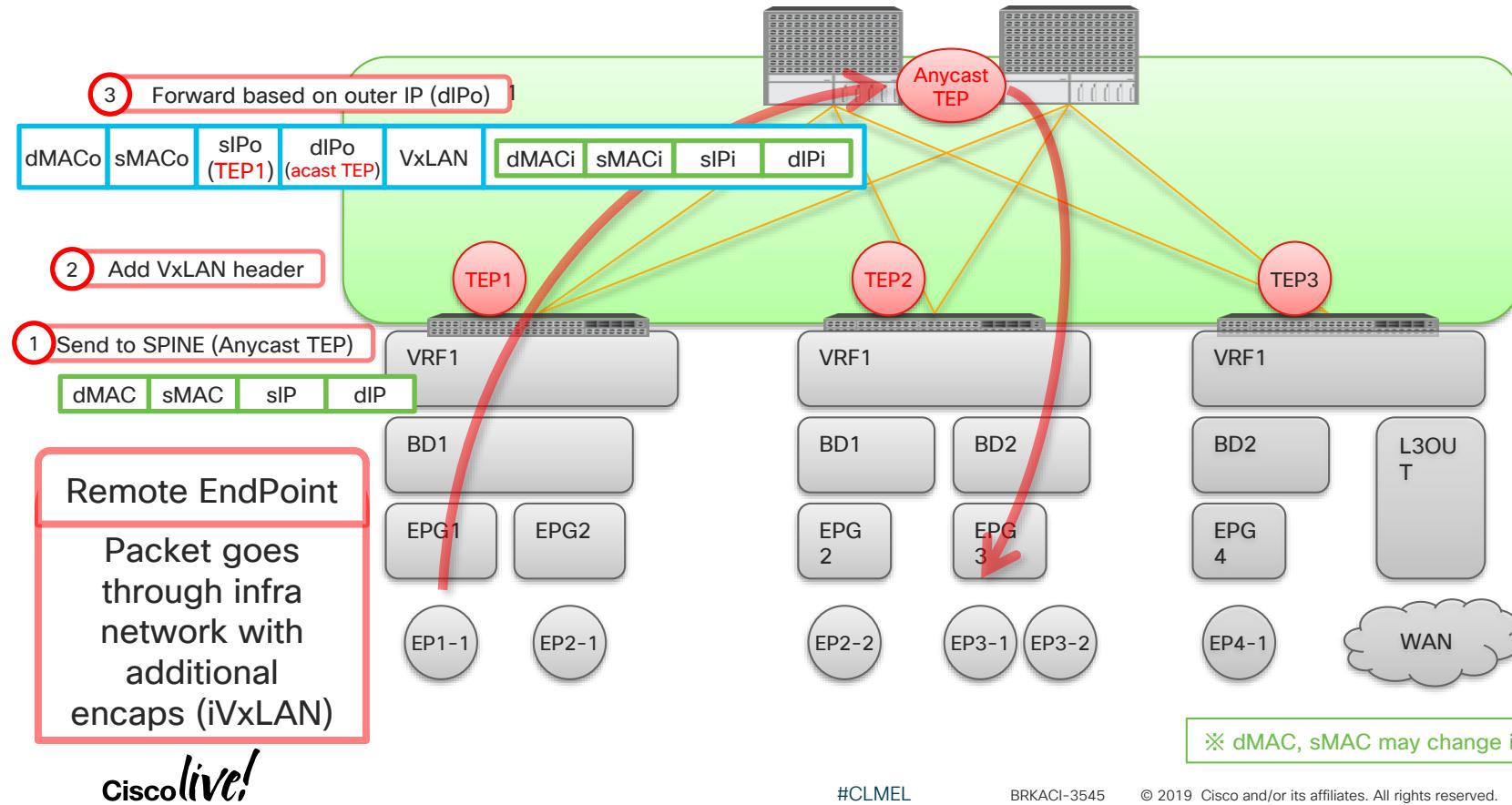
Scenario 4 : source LEAF does NOT know the destination (Flood)



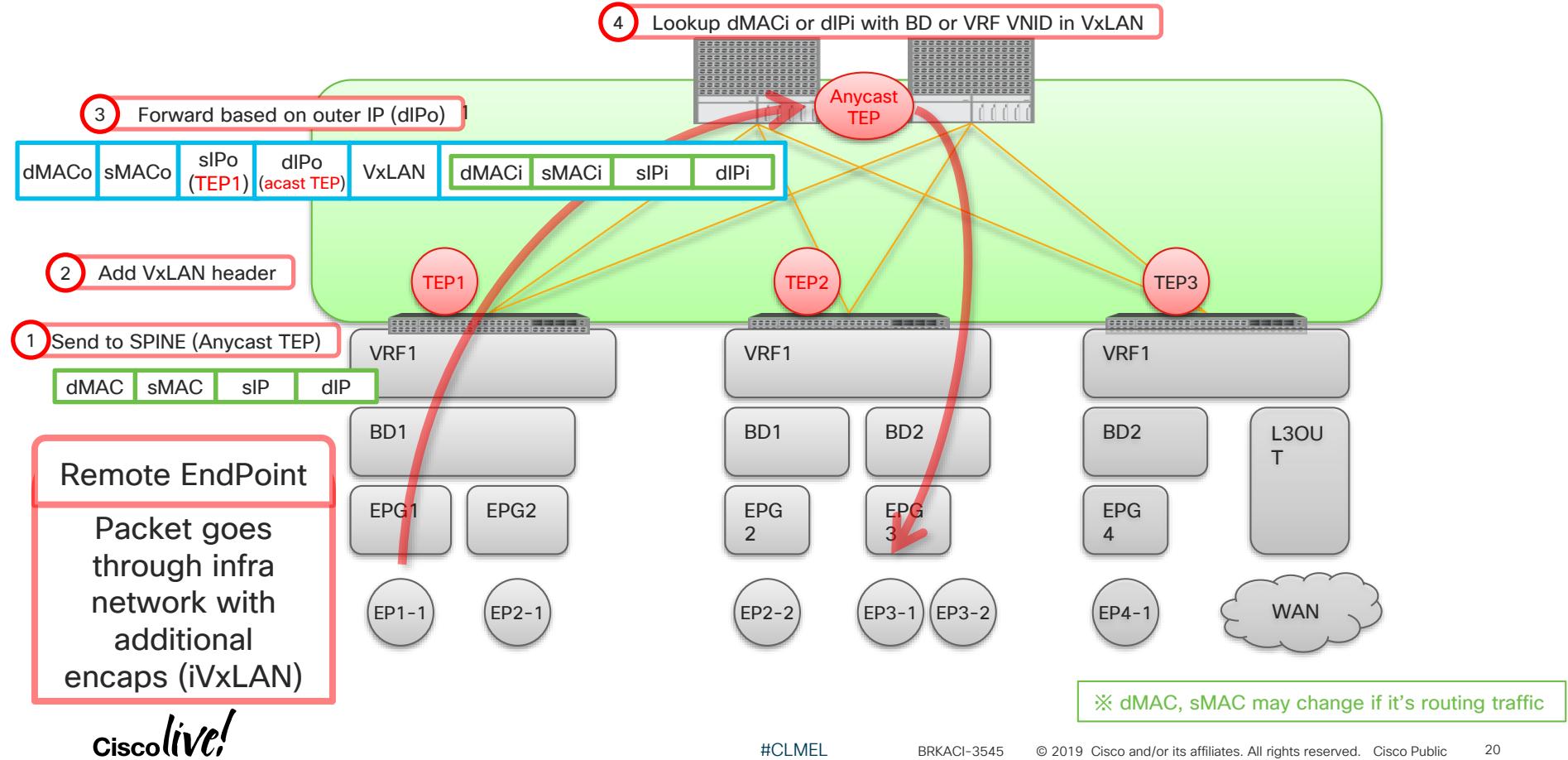
Source LEAF does NOT know the destination (Spine-Proxy)



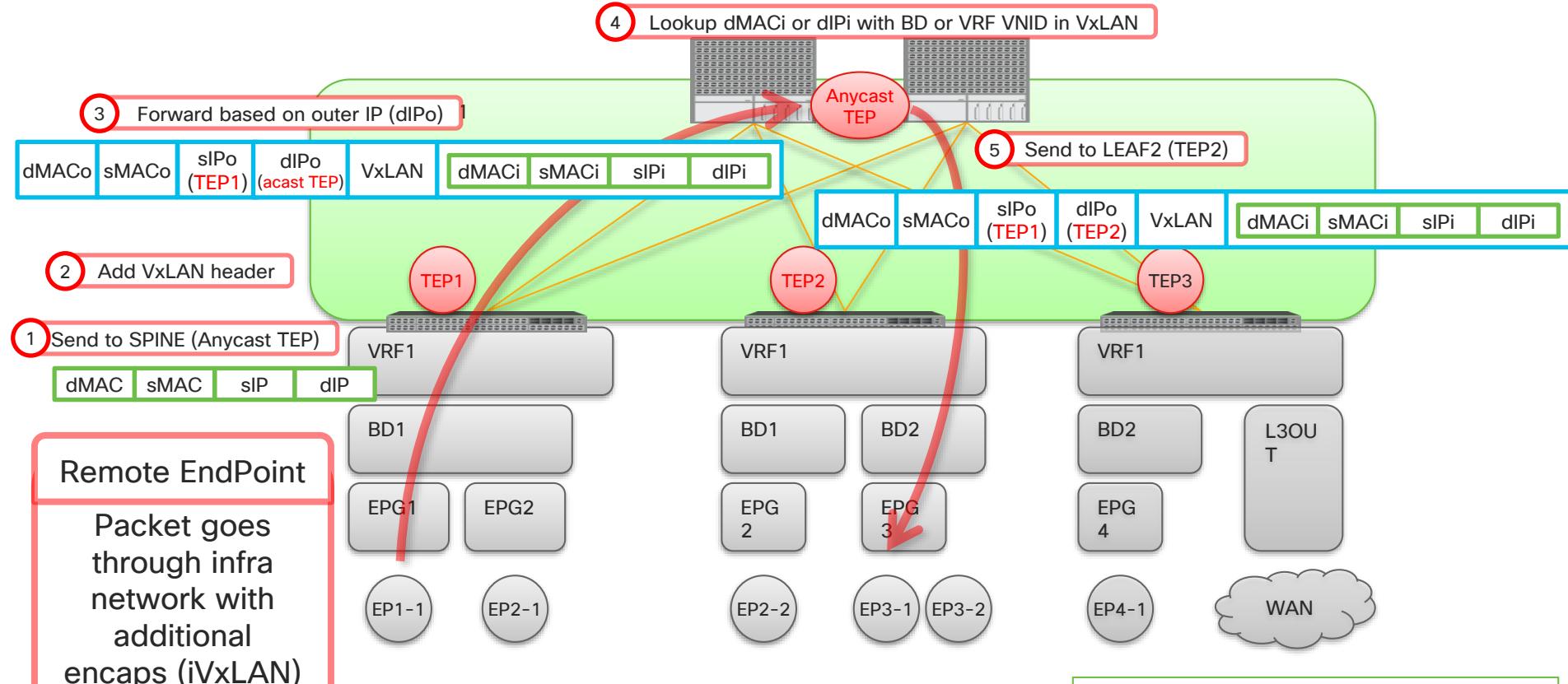
Source LEAF does NOT know the destination (Spine-Proxy)



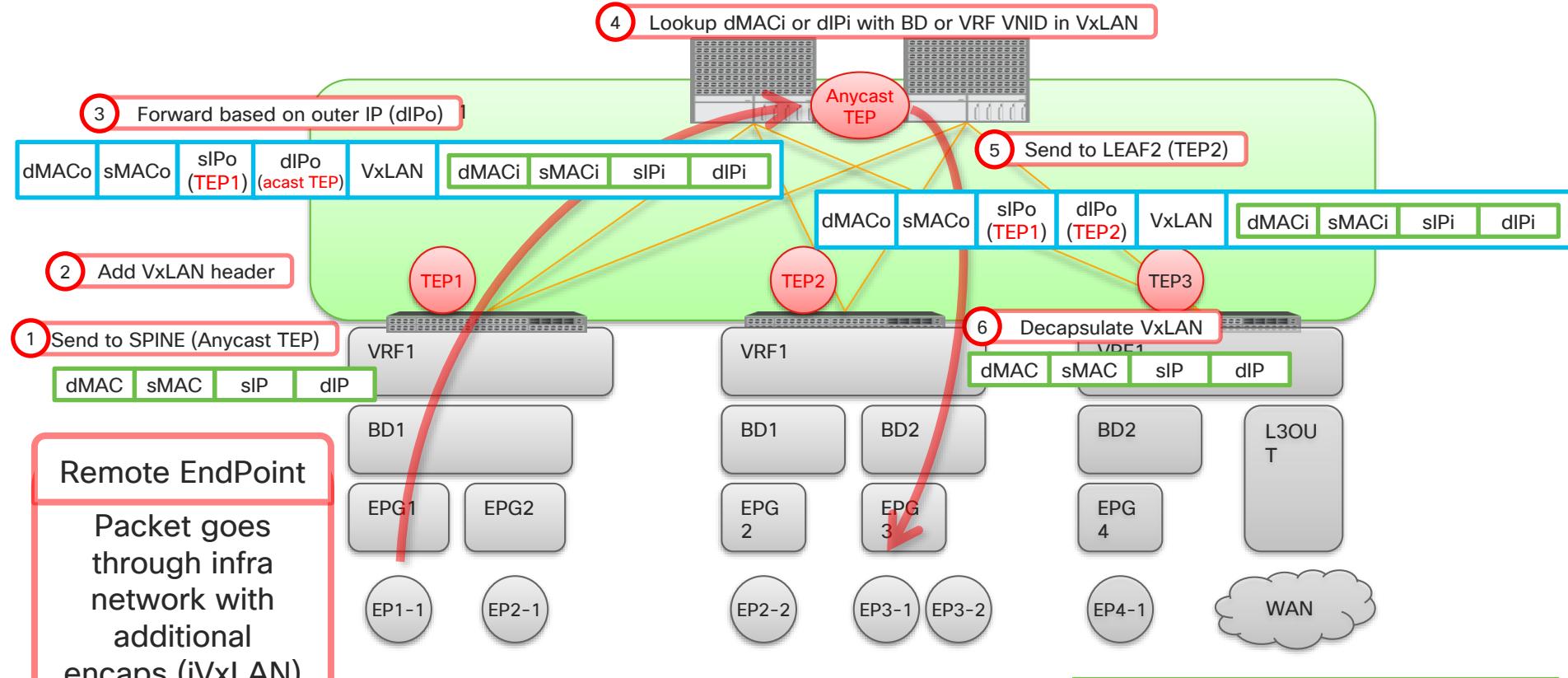
Source LEAF does NOT know the destination (Spine-Proxy)



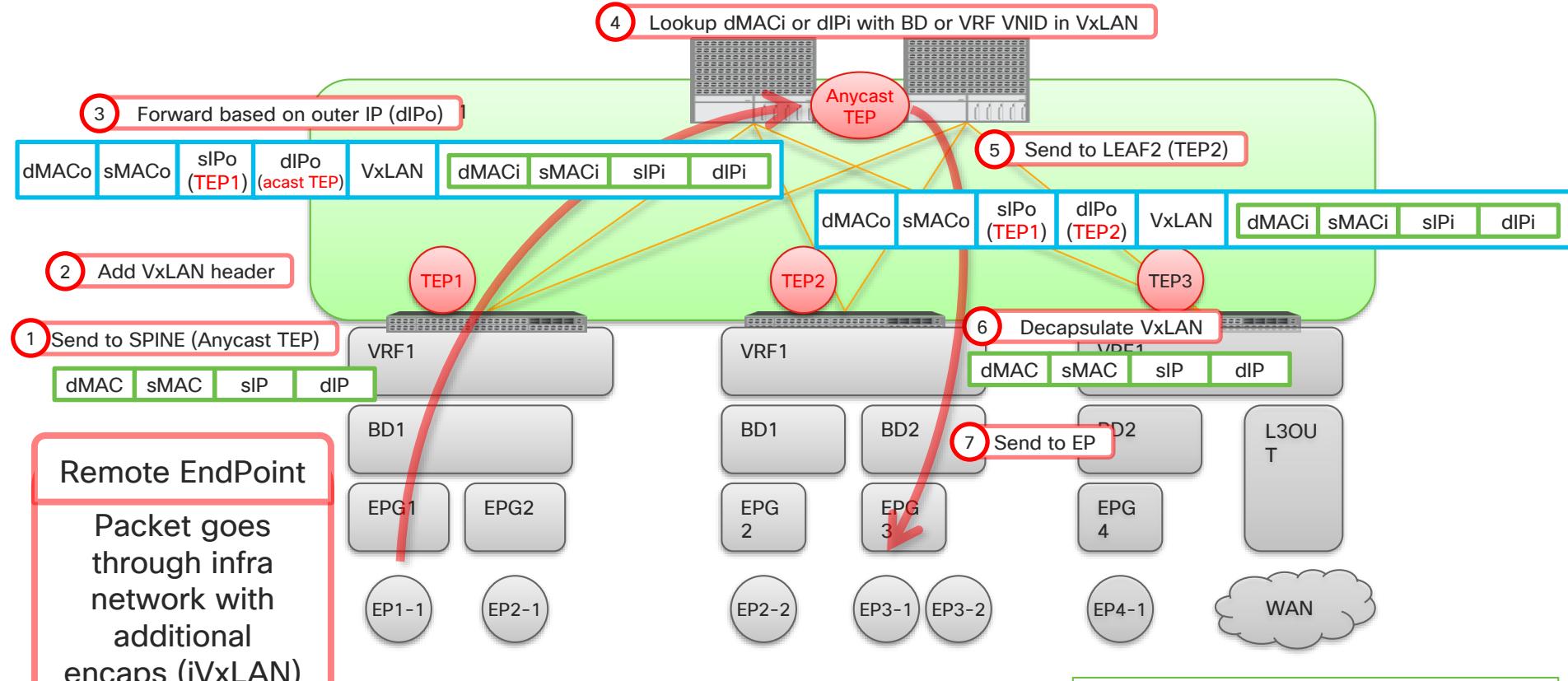
Source LEAF does NOT know the destination (Spine-Proxy)



Source LEAF does NOT know the destination (Spine-Proxy)



Source LEAF does NOT know the destination (Spine-Proxy)



ACI Overlay VxLAN and TEP

Scenario 1 : source LEAF knows the destination (on the same LEAF)



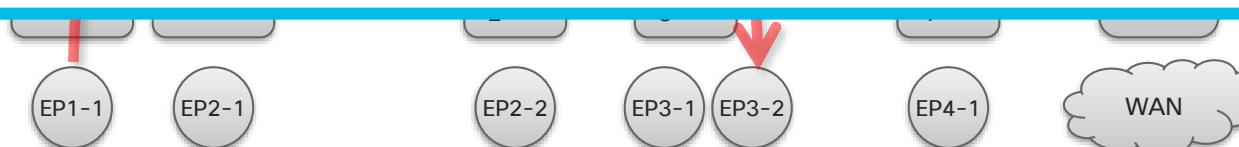
Scenario 2 : source LEAF knows the destination (on another LEAF X)



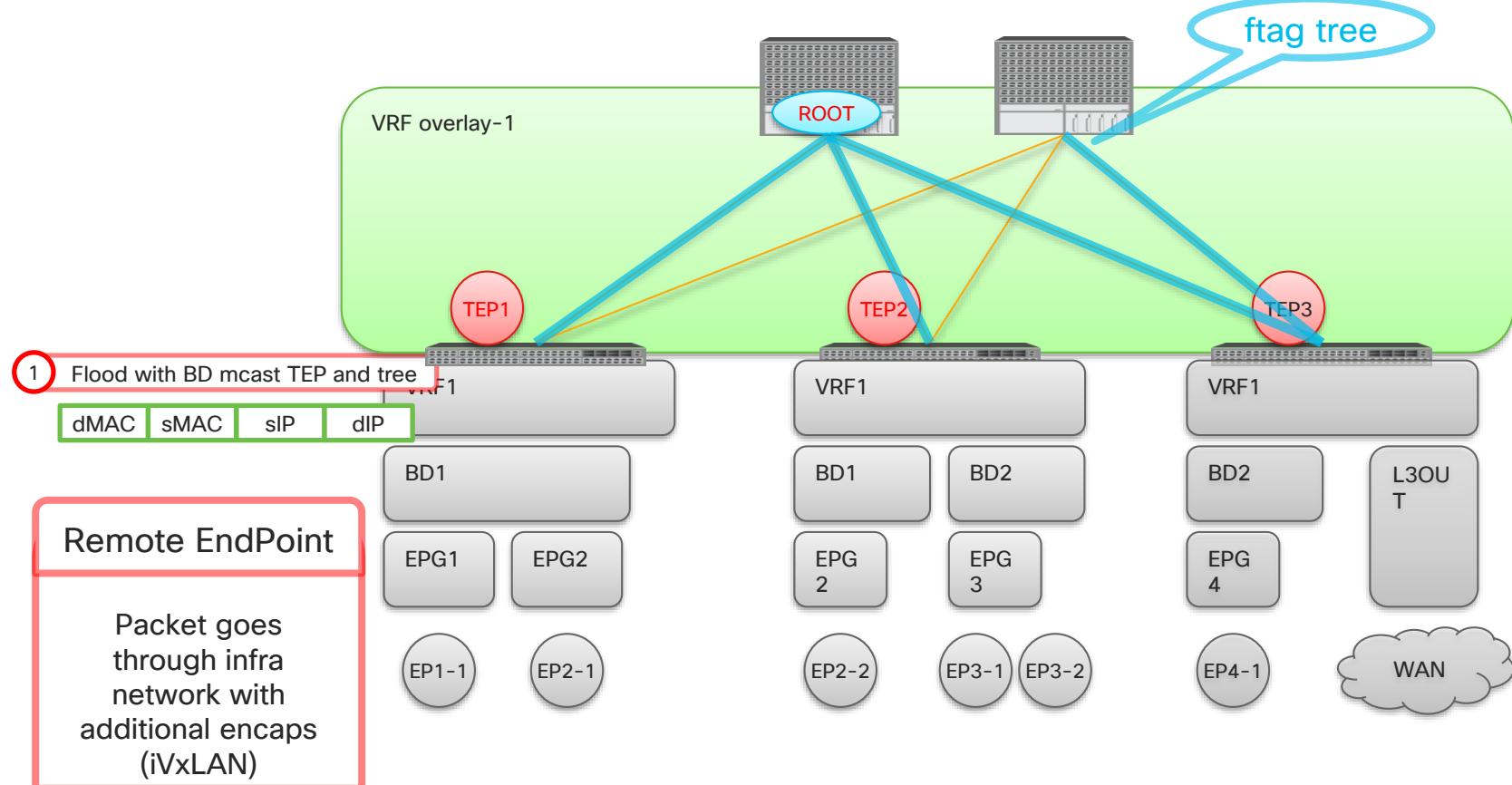
Scenario 3 : source LEAF does NOT know the destination (Spine-Proxy)



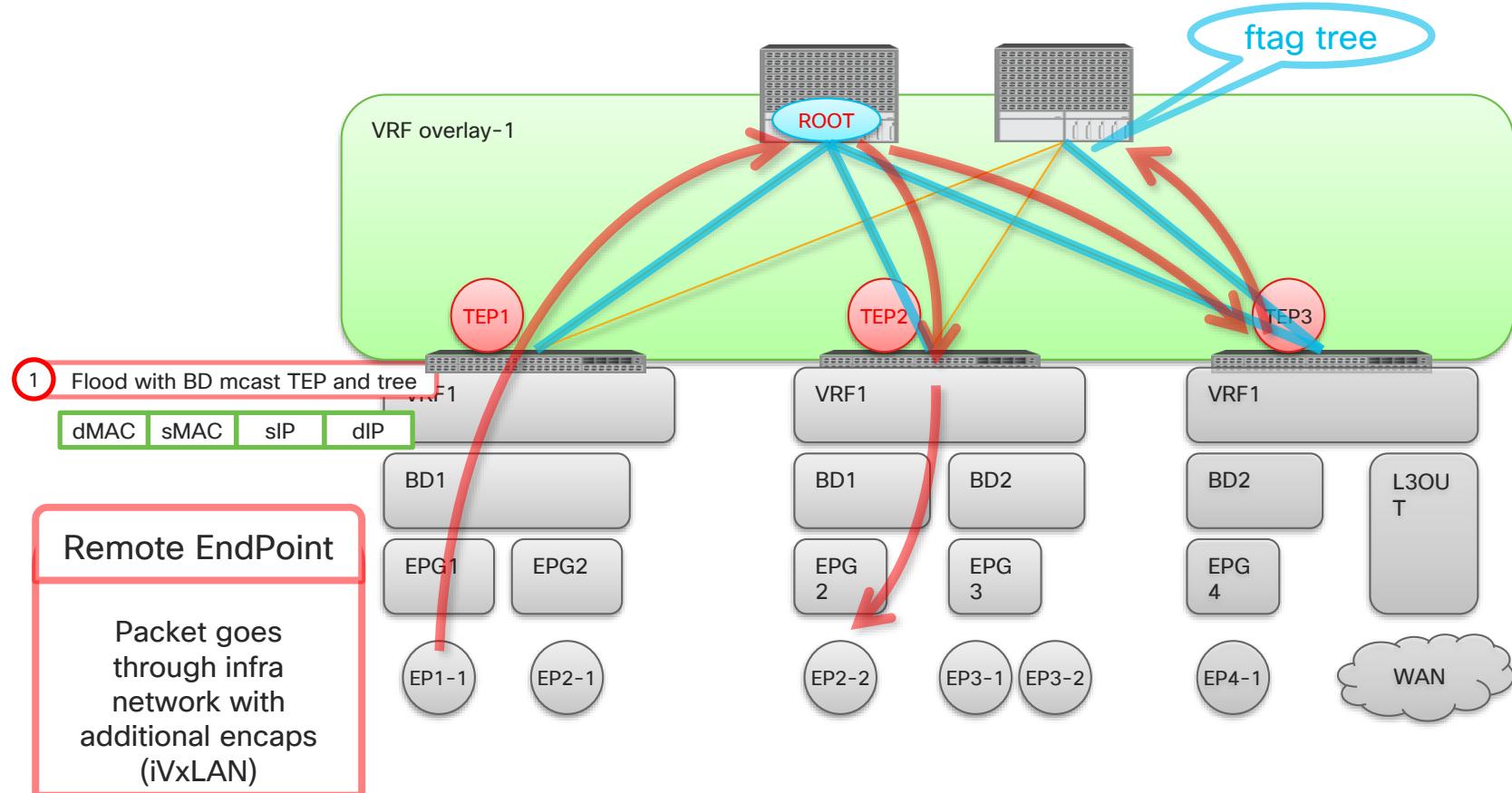
Scenario 4 : source LEAF does NOT know the destination (Flood)



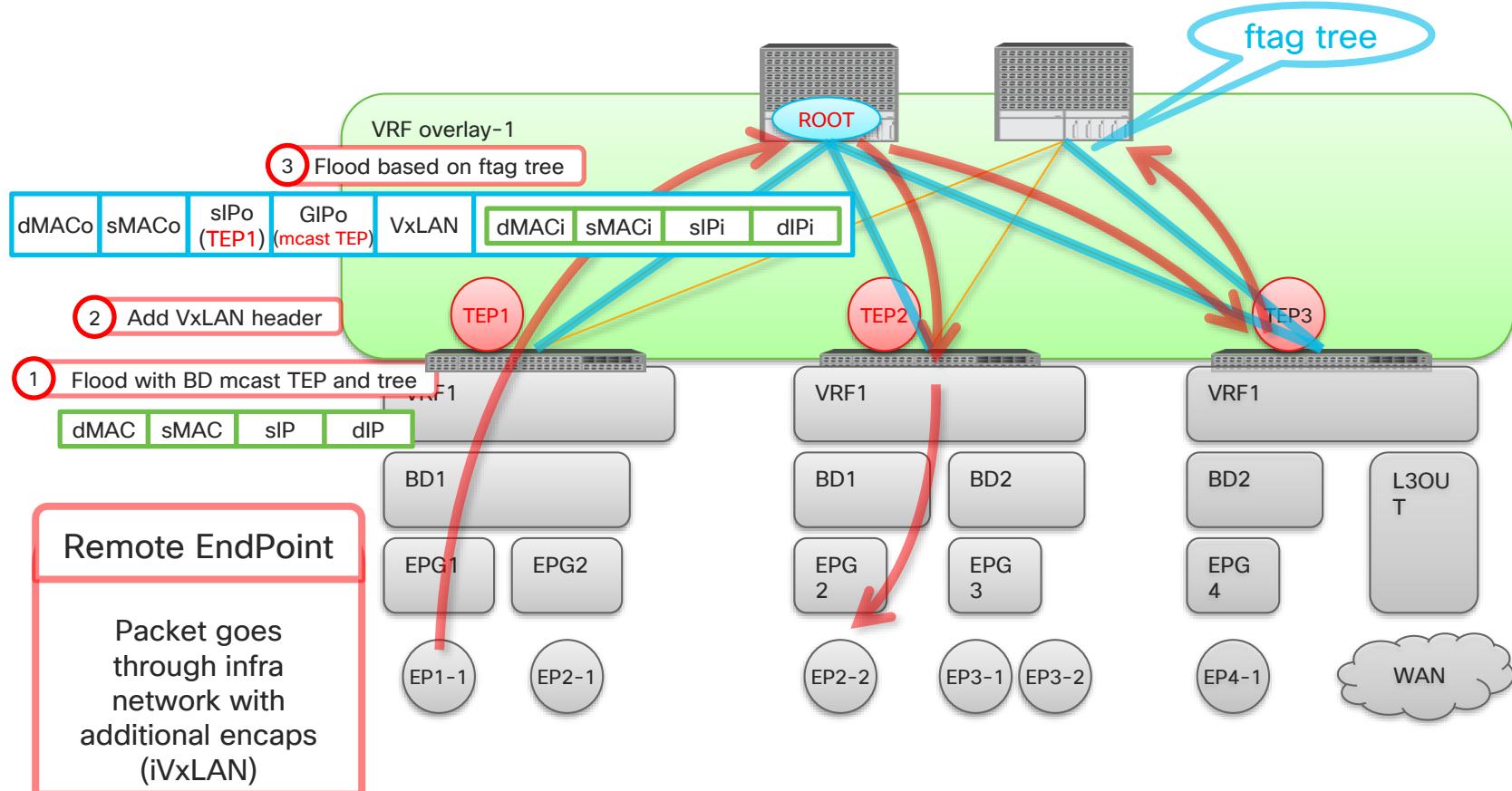
Source LEAF does NOT know the destination (Flood)



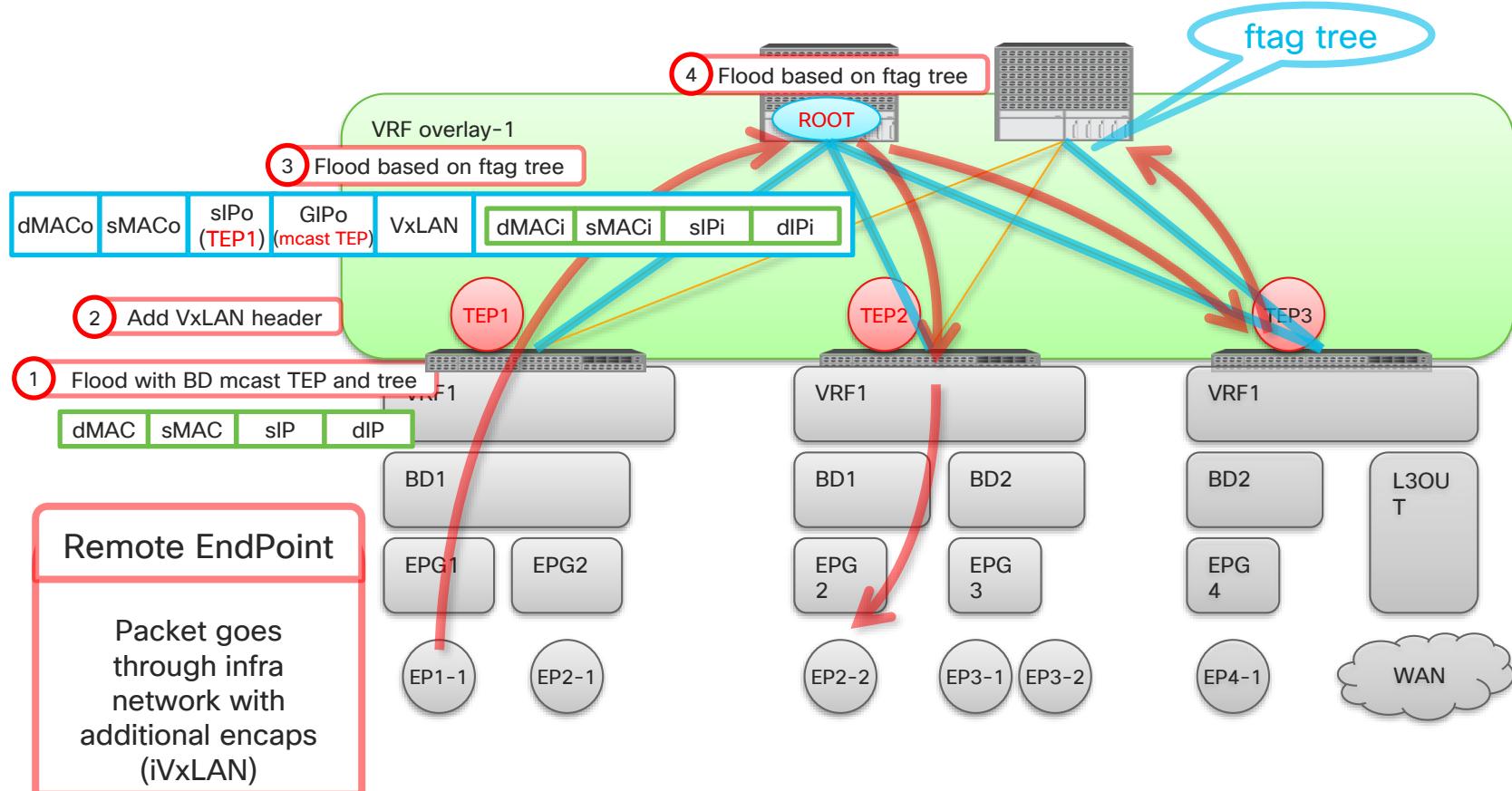
Source LEAF does NOT know the destination (Flood)



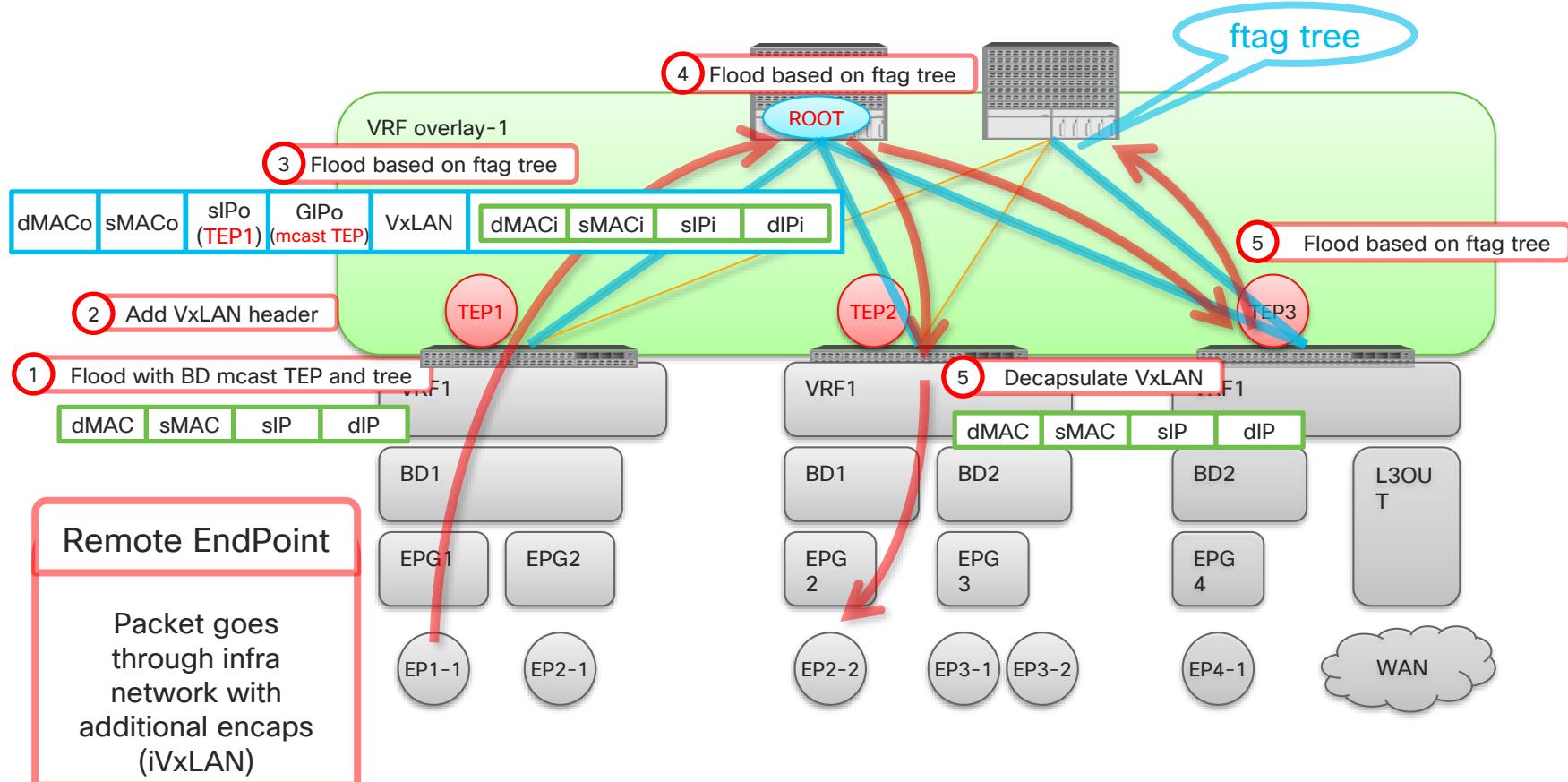
Source LEAF does NOT know the destination (Flood)



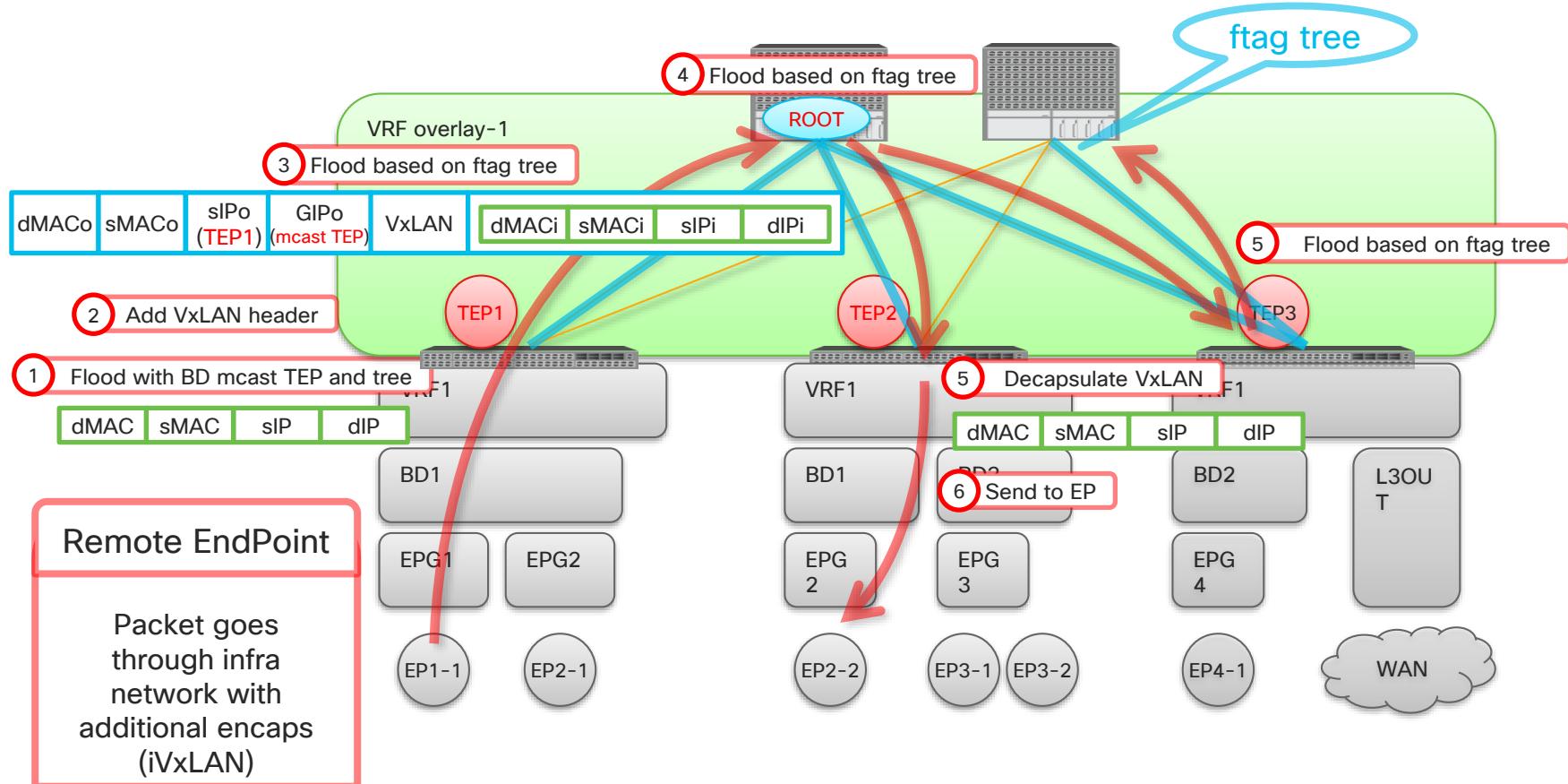
Source LEAF does NOT know the destination (Flood)



Source LEAF does NOT know the destination (Flood)

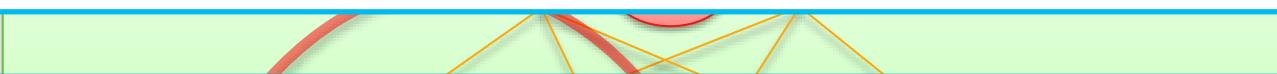


Source LEAF does NOT know the destination (Flood)



ACI Overlay VxLAN and TEP

Scenario 1 : source LEAF knows the destination (on the same LEAF)



Scenario 2 : source LEAF knows the destination (on another LEAF X)



Scenario 3 : source LEAF does NOT know the destination (Spine-Proxy)



Scenario 4 : source LEAF does NOT know the destination (Flood)



**How does LEAF pick one of these scenario?
➤ based on EP information**

Agenda

- Introduction
 - ACI Overlay VxLAN and TEP
- ACI Forwarding components
 - Endpoints, EPG, EP Learning, COOP and How it all works
 - BD, VRF forwarding scope and detailed options
 - Spine-Proxy and ARP Glean
 - Forwarding Software Architecture and ASIC Generation
- ACI Packet Walk
 - Walk through the life of a packet going through ACI

ACI Forwarding Component 1

- Endpoint
- EPG (EndPoint Group)
- VLAN Type in ACI
- Endpoint Type
- Endpoint Learning
- COOP (Council of Oracle Protocol)

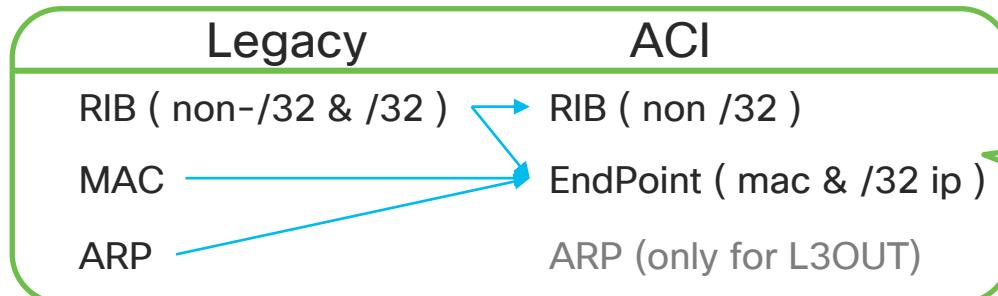
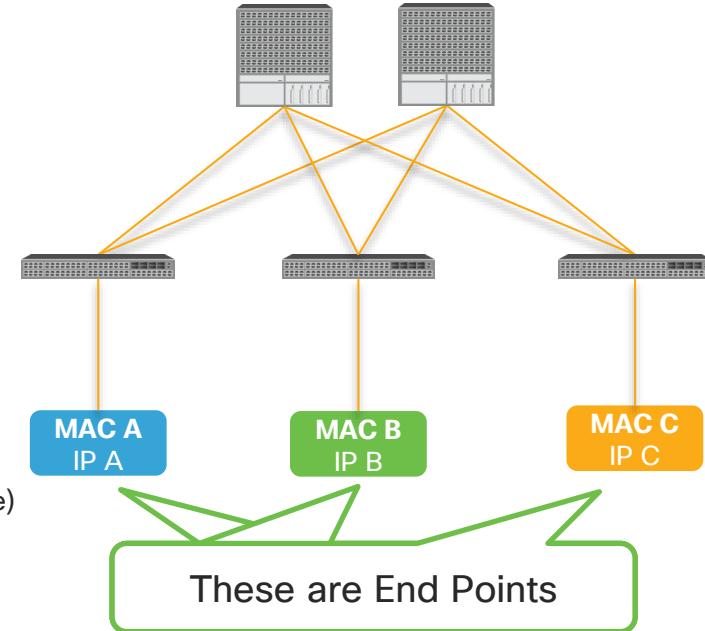
End Point (EP)

What is an EP?

- It stands for hosts, in other words MAC address with IP(s)
 - sometimes MAC only
 - IP in EP is always /32

What Forwarding Table is used?

- End Point Table
 - host information (MAC and /32 IP address)
- LPM(Longest Prefix Match) Table
 - non /32 IP route information (exception: /32 for SVI or L3OUT route)



Cisco live!

RIB : Routing Information Base

#CLMEL

BRKACI-3545

© 2019 Cisco and/or its affiliates. All rights reserved. Cisco Public

34

Forwarding table lookup order

1. EndPoint Table (show endpoint)
2. RIB (show ip route)

End Point Group (EPG)

What is an EPG?

- Logical grouping of hosts (EPs)
- Each EPG belongs to a Bridge Domain (BD).

What is the EPG used for?

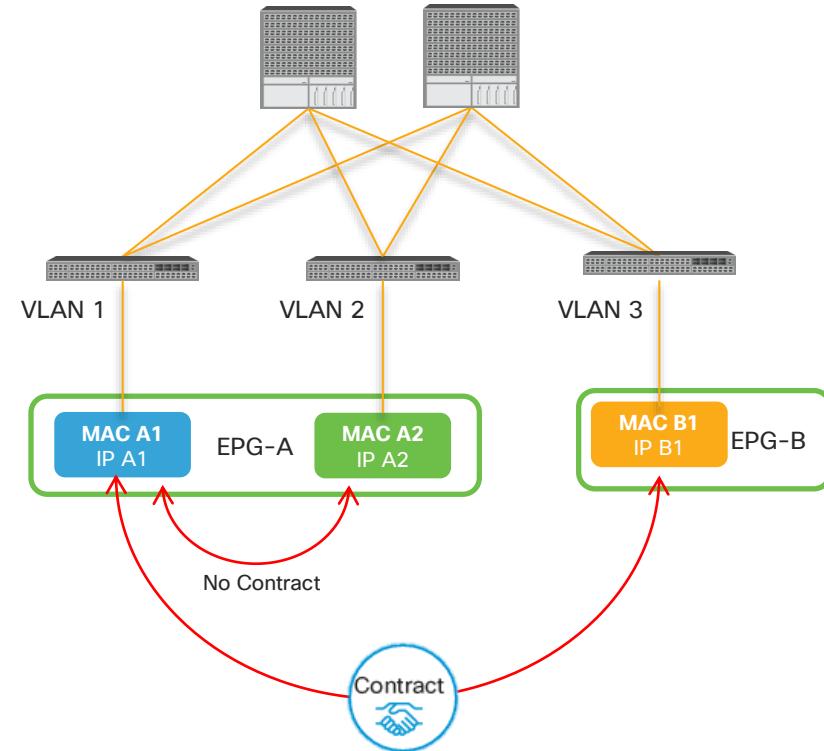
- **To implement traffic filtering**
 - Traffic within the same EPG doesn't get blocked
 - Traffic across EPGs always blocked without a contract
- A contract is applied between EPGs to allow traffic

What is a VLAN in ACI?

- VLAN is just an identifier to classify EPs to each EPG
- BD is the L2 domain instead of VLAN

What happens with forwarding?

- It will be done by BD or VRF to which EPG belongs



How to check End Points

From APIC GUI (Fabric perspective)

End Point	MAC	IP	Learning Source	Hosting Server	Report Contrc Name	Interface	Multicast Address	Encap
EP-00-00:00:00:51:51	00:00:00:00:51:51	192.168.1.1	learned	---	---	Pod-1/Node-101/eth1/1 (learned)	---	vlan-51
EP-00-00:11:11:51:51	00:00:11:11:51:51	192.168.1.11	learned	---	---	Pod-1/Node-102/eth1/27 (learned)	---	vlan-51

Fabric Wide Visibility
shows where EPs are learned

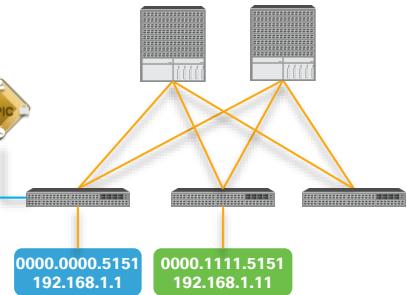
From LEAF CLI (LEAF perspective)

```
leaf1# show endpoint vrf TK:VRF1
```

Legend:

s - arp	O - peer-attached	a - local-aged	S - static
V - vpc-attached	p - peer-aged	M - span	L - local
B - bounce	H - vtep		

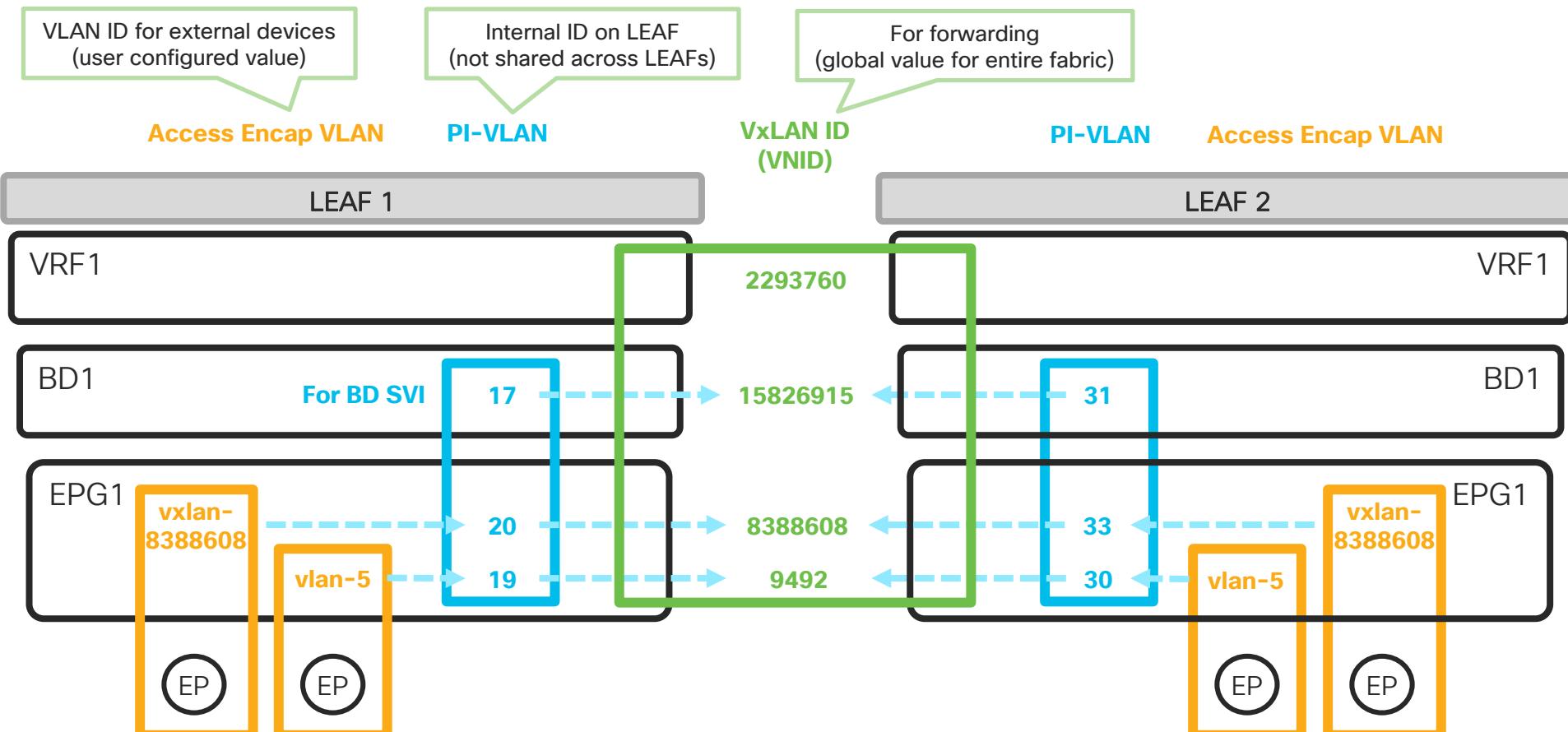
VLAN/ Domain	Encap VLAN	MAC Address IP Address	MAC Info/ IP Info	Interface
TK:VRF1		192.168.1.11		tunnel18
17/TK:VRF1	vxlan-15826915	0000.1111.5151		tunnel18
19	vlan-5	0000.0000.5151 L		eth1/1
TK:VRF1		192.168.0.51 L		eth1/1



Good for forwarding verification
shows how EPs look from each LEAF

VLAN types in ACI

* PI-VLAN : Platform Independent VLAN



PI-VLAN for EPG and BD CLI



- Endpoint Table

```
leaf1# show endpoint ip 192.168.0.51
19          vlan-5      0000.5555.1111 L      eth1/1
TK:VRF1     vlan-5      192.168.0.51 L      eth1/1

PI-VLAN      Access Encap VLAN
```

- VLAN Table

```
leaf1# show vlan id 17,19 extended
VLAN Name Status Ports
---- --
17 TK:BD1 active Eth1/1, Eth1/2, Po6
19 TK:AP1:EPG1 active Eth1/1

VLAN Type Vlan-mode Encap
---- --
17 enet CE vxlan-15826915
19 enet CE vlan-5
```

PI-VLAN Access Encap VLAN

Cisco live!



PI-VLAN for EPG and BD CLI

```
leaf1# show vlan id 17,19 extended
```

VLAN	Name	Status	Ports
17	TK:BD1	active	Eth1/1, Eth1/2, Po6
19	TK:AP1:EPG1	active	Eth1/1

```
VLAN Type Vlan-mode Encap
```

17	enet	CE	vxlan-15826915
19	enet	CE	vlan-5

EPG
PI-VLAN

```
leaf1# show system internal epm vlan 19
```

VLAN ID	Type	Access Encap (Type Value)	Fabric Encap	H/W id	BD VLAN	Endpoint Count
19	FD vlan	802.1Q	5 8294	14	17	2

BD
PI-VLAN

End Point Types

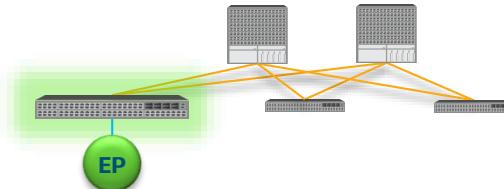


Physical Local Endpoint (PL)

- An endpoint attached to this LEAF

```
fab1-leaf1# show endpoint ip 192.168.0.51
```

19	vlan-5	0000.5555.1111	L	eth1/1
TK:VRF1	vlan-5	192.168.0.51	L	eth1/1



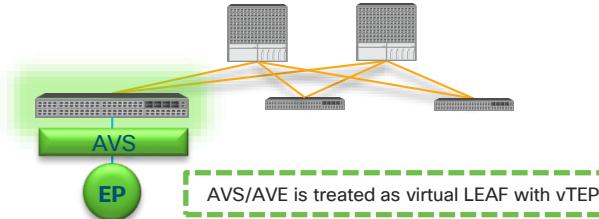
Virtual Local Endpoint (VL)

- An endpoint on AVS/AVE attached to this LEAF

```
fab1-leaf1# show endpoint ip 192.168.66.2
```

14	vxlan-8388608	0050.5680.34eb	L	tunnel10
TK:VRF1	vxlan-8388608	192.168.66.2	L	tunnel10

Access Encap VLAN (VxLAN)



Remote Endpoint (Xr)

- An endpoint on a remote LEAF

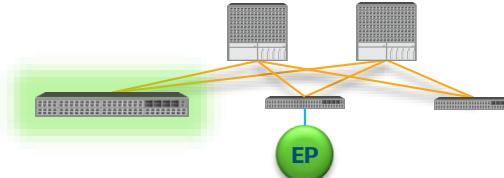
```
fab1-leaf1# show endpoint mac 0000.5555.2222
```

17/TK:VRF1	vxlan-15826915	0000.5555.2222	tunnel8
------------	----------------	----------------	---------

BD VNID (not Access Encap VLAN)

```
fab1-leaf1# show endpoint ip 192.168.0.52
```

TK:VRF1	192.168.0.52	tunnel8
---------	--------------	---------

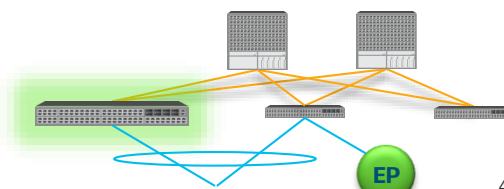


On-Peer Endpoint

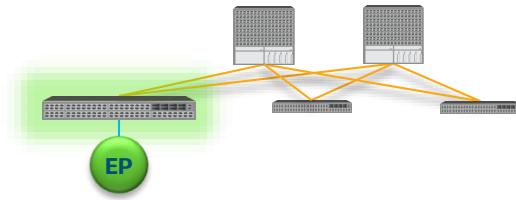
- An endpoint connected to an orphan port on vPC peer

```
fab1-leaf1# show endpoint ip 192.168.0.52
```

19	vlan-5	0000.5555.2222	O	tunnel8
TK:VRF1	vlan-5	192.168.0.52	O	tunnel8



End Point Learning (Local EP)



PI-VLAN ID(19) of EPG
to which MAC belongs

```
leaf1# show endpoint ip 192.168.0.51
```

```
19  
TK:VRF1
```

vlan	mac	ip	port
vlan-5	0000.5555.1111	L	eth1/1
vlan-5	192.168.0.51	L	eth1/1

VRF to which
MAC & IP belong

Access Encap VLAN
of EP(MAC+IP)

Local Endpoint (MAC)

A leaf learns **MAC A** as **local** if a packet with **src MAC A** comes in from its **front panel port**.

Local Endpoint (/32 host IP)

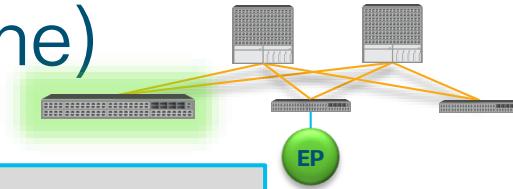
A leaf learns **IP A /32** as **local**

- if a packet with **src IP A** comes in from its **front panel port AND IP lookup** is done on ACI.
(which means IP addr is learned **only when** a leaf handles **L3 traffic**)
or
- if **ARP request** with **sender IP A** comes in from its **front panel port**. (regardless of ARP Flooding setup)

EPG/BD/VRF is based on Access Encap VLAN ID

What APIC GUI shows are these local Endpoints

End Point Learning (Remote EP = cache)



PI-VLAN(17) of BD
and VRF to which
MAC belongs

```
fab1-leaf1# show endpoint mac 0000.5555.2222  
17/TK:VRF1 vxlan-15826915 0000.5555.2222 tunnel18
```

BD VNID

VRF to which
MAC & IP belong

```
fab1-leaf1# show endpoint ip 192.168.0.52  
TK:VRF1 192.168.0.52 tunnel18
```

tunnel represents
destination leaf TEP

Remote Endpoint (MAC)

A leaf learns **MAC A** as **remote** when **L2 traffic** with **src MAC A** comes in from **SPINE**.

Remote Endpoint (/32 host IP)

A leaf learns **IP A** as **remote** when **L3 traffic** with **src IP A** comes in from **SPINE**.

- Remote MAC and remote IP is learned separately
- BD(for MAC) / VRF(for IP) is based on VNID in VxLAN header

VNID is
BD when L2 traffic
VRF when L3 traffic (not both)





How to check Tunnel Interface (TEP)

```
leaf1# show int tunnel 8 | grep Tun
```

Tunnel8 is up

Tunnel protocol/transport is ivxlan

Tunnel source 11.0.200.92/32 (lo0)

Tunnel destination 11.0.48.95

TEP IP address

```
leaf1# acidiag fnvread
```

ID	Pod ID	Name	Serial Number	IP Address	Role	State	LastUpdMsgId
101	1	leaf1	FDO20160AAA	11.0.200.92/32	leaf	active	0
102	1	leaf2	FDO20160BBB	11.0.48.95/32	leaf	active	0
103	1	leaf3	FDO20240CCC	11.0.200.91/32	leaf	active	0
201	1	spine1	FGE12345678	11.0.200.94/32	spine	active	0
202	1	spine2	FGE87654321	11.0.200.93/32	spine	active	0

```
admin@apic1:~> moquery -c vpcDom | egrep 'virtualIp|dn|#'
```

```
# vpc.Dom
```

```
dn : topology/pod-1/node-101/sys/vpc/inst/dom-1
```

virtualIp : 11.0.64.65/32

```
# vpc.Dom
```

```
dn : topology/pod-1/node-102/sys/vpc/inst/..
```

virtualIp : 11.0.64.65/32

```
# vpc.Dom
```

```
dn : topology/pod-1/node-103/sys/vpc/inst/dom-2
```

virtualIp : 11.0.192.64/32

```
# vpc.Dom
```

```
dn : topology/pod-1/node-104/sys/vpc/inst/dom-2
```

virtualIp : 11.0.192.64/32

Tunnel may point to vPC vTEP
This is vTEP for vPC LEAF 101-102

COOP (End Point Learning on Spine)

SPINEs do NOT learn EP from data plane like LEAF

SPINEs receive all EP data from Leaf

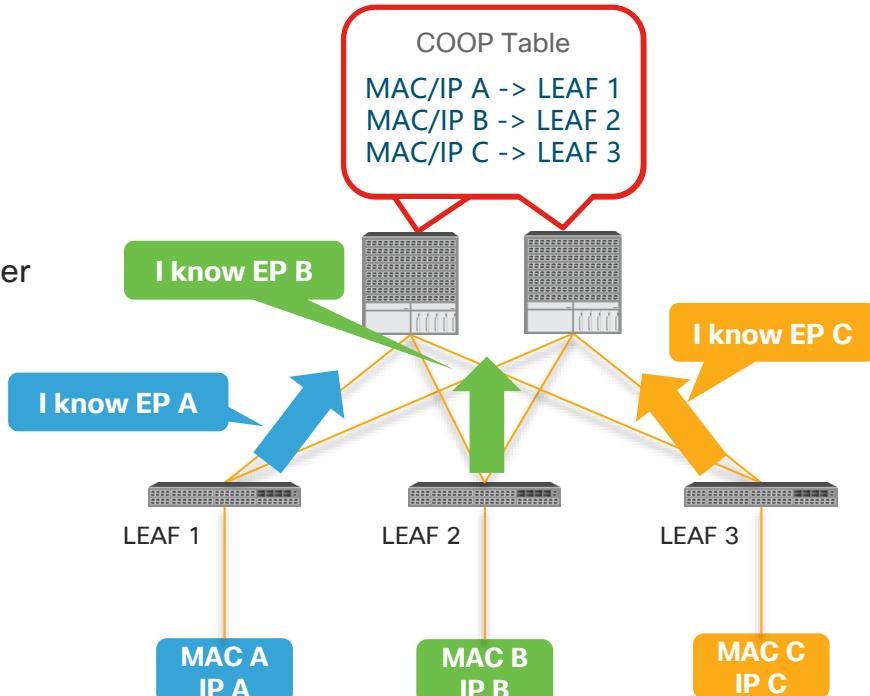
1. LEAF learns EP (either MAC or/and IP) as local
2. LEAF reports local EP to Spine via COOP process
3. SPINE stores these in COOP DB and synchronize with other SPINEs

What is the purpose of COOP?

When Leaf doesn't know dst EP, LEAF can forward packet to Spine in order to let Spine decide where to send.
This behavior is called **Spine-Proxy**.

Note :

- Normally SPINE doesn't push COOP DB entries to each LEAF. It just receives and stores. The exception is for bounce entries.
- Remote Endpoints are stored on each Leaf nodes as cache. This is not reported to Spine COOP.



Agenda

- Introduction
 - ACI Overlay VxLAN and TEP
- ACI Forwarding components
 - Endpoints, EPG, EP Learning, COOP and How it all works
 - BD, VRF forwarding scope and detailed options
 - Spine-Proxy and ARP Glean
 - Forwarding Software Architecture and ASIC Generation
- ACI Packet Walk
 - Walk through the life of a packet going through ACI

ACI Forwarding Component 2

- Pervasive Gateway (BD SVI)
- Forwarding Scope (VRF or BD)
- Forwarding mode in BD

Pervasive Gateway(BD SVI)

The screenshot shows the Cisco ACI Tenant TK configuration interface. On the left, under 'Networking > Bridge Domains > BD1 > Subnets', two subnets are listed: 'SN 192.168.0.254/24' and 'SN 192.168.1.254/24'. A red box highlights this list. On the right, the 'Properties' tab for 'Subnet - 192.168.0.254/24' is shown. It includes fields for IP Address (192.168.0.254/24), Description (optional), and various configuration options like 'Scope' (Private to VRF selected), 'Subnet Control' (radio button selected), and 'L3 Out for Route Profile'.

```
leaf1# show ip route vrf TK:VRF1
```

192.168.0.0/24, ubest/mbest: 1/0, attached, direct, **pervasive**
*via 10.0.184.64%overlay-1, [1/0], 04:32:16, static

192.168.0.254/32, ubest/mbest: 1/0, attached
*via 192.168.0.254, vlan10, [1/0], 04:32:16, local, local

BD SVI with PI-VLAN

Pervasive route

Pervasive SVI

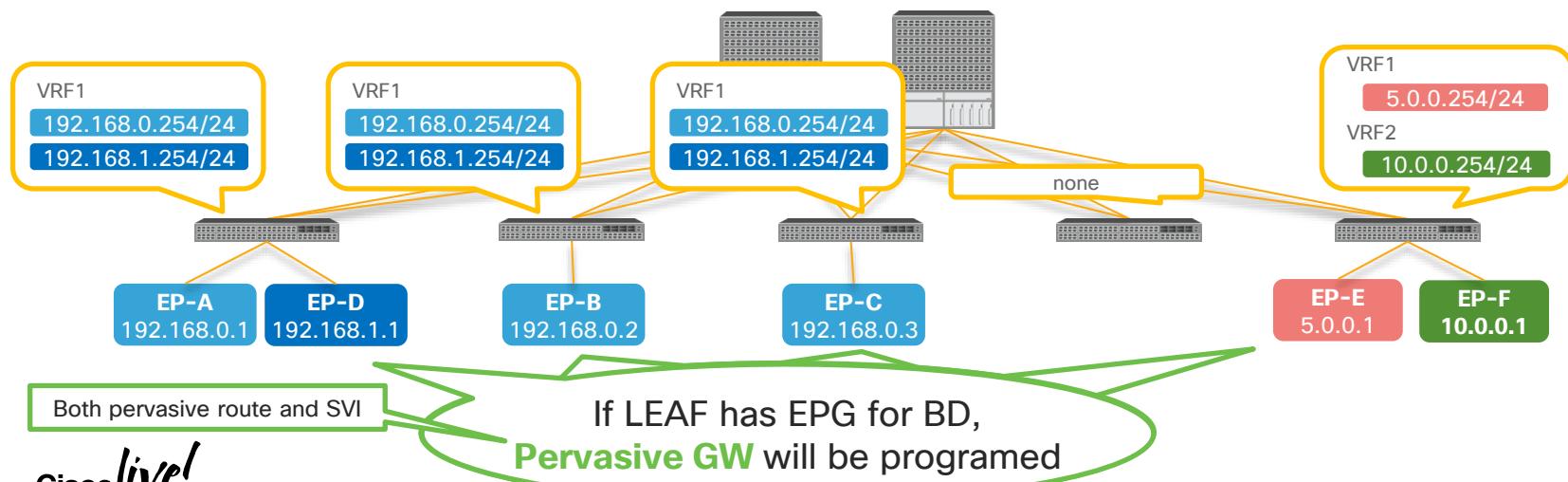
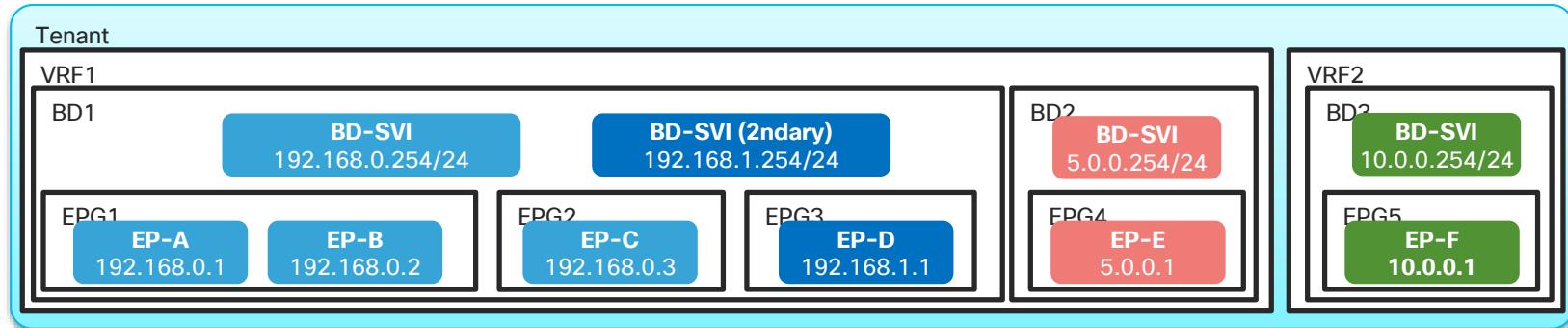
What is pervasive GW for?

- To be a default GW for EPs in the Fabric
 - All EPs can have consistent gateway IP address one hop away
- To represent subnets(IP ranges) for a BD
 - ACI knows which BD may have potential hidden/silent EPs

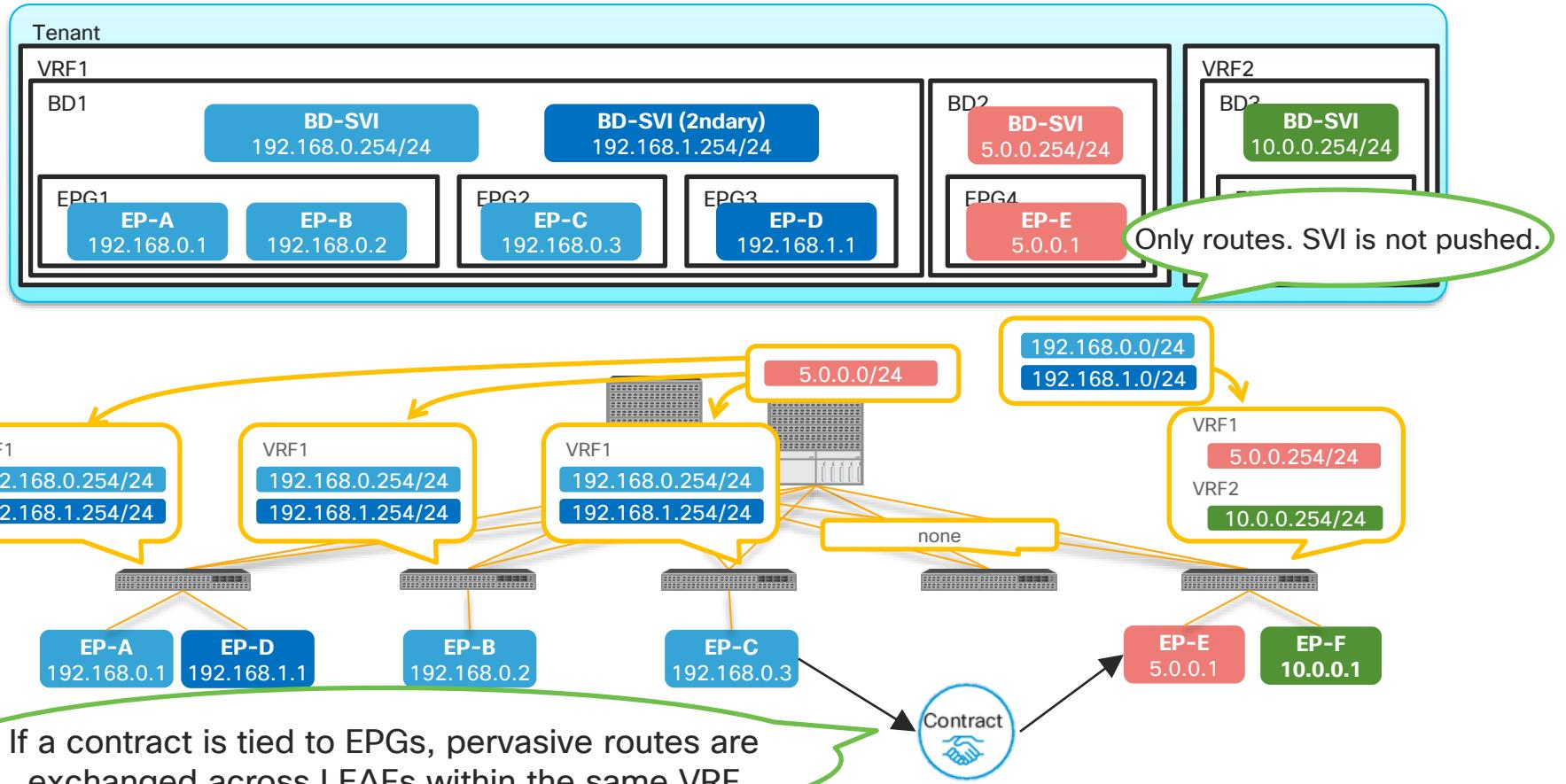
How is pervasive GW deployed?

- Installed as an SVI on LEAFs
 - PI-VLAN for BD is used to represent a pervasive GW SVI
 - A pervasive SVI has secondary IP when multiple pervasive GWs are configured on the same BD
 - User can choose a primary address

Pervasive Gateway(BD SVI) example



Pervasive Gateway(BD SVI) example



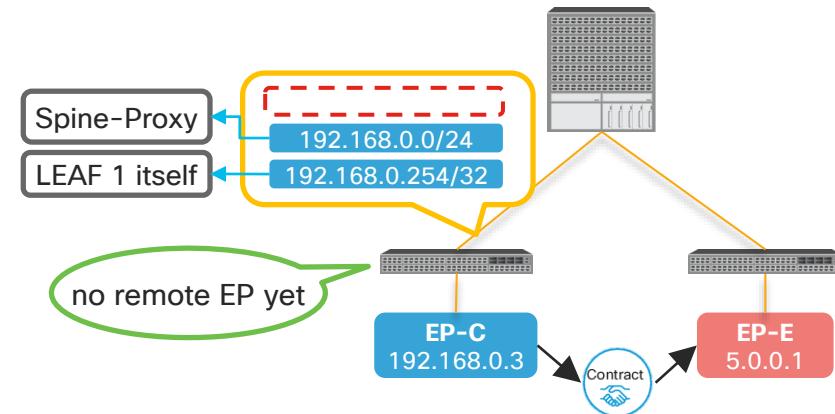
Pervasive Gateway(BD SVI) cont.

Why does ACI push pervasive routes to other LEAFs after a contract?

➤ Pervasive routes are required for Spine-Proxy



what if no pervasive route, no remote EP?

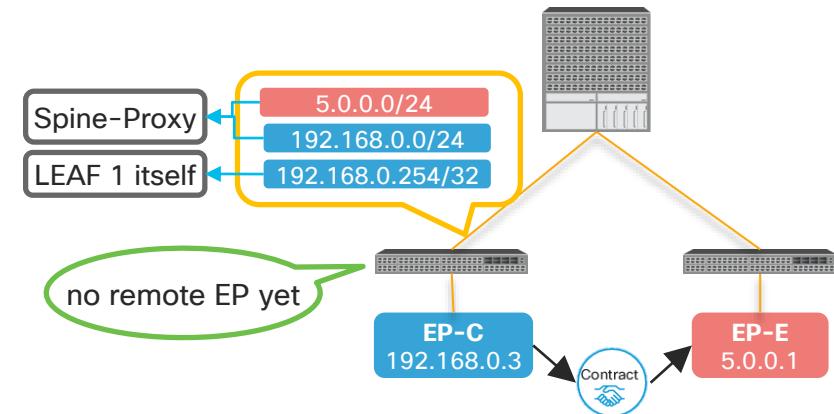


No Spine-Proxy for 5.0.0.1

It may be either dropped or forwarded to L3OUT if a default route exists



with pervasive route and no remote EP?



Spine-Proxy for 5.0.0.1

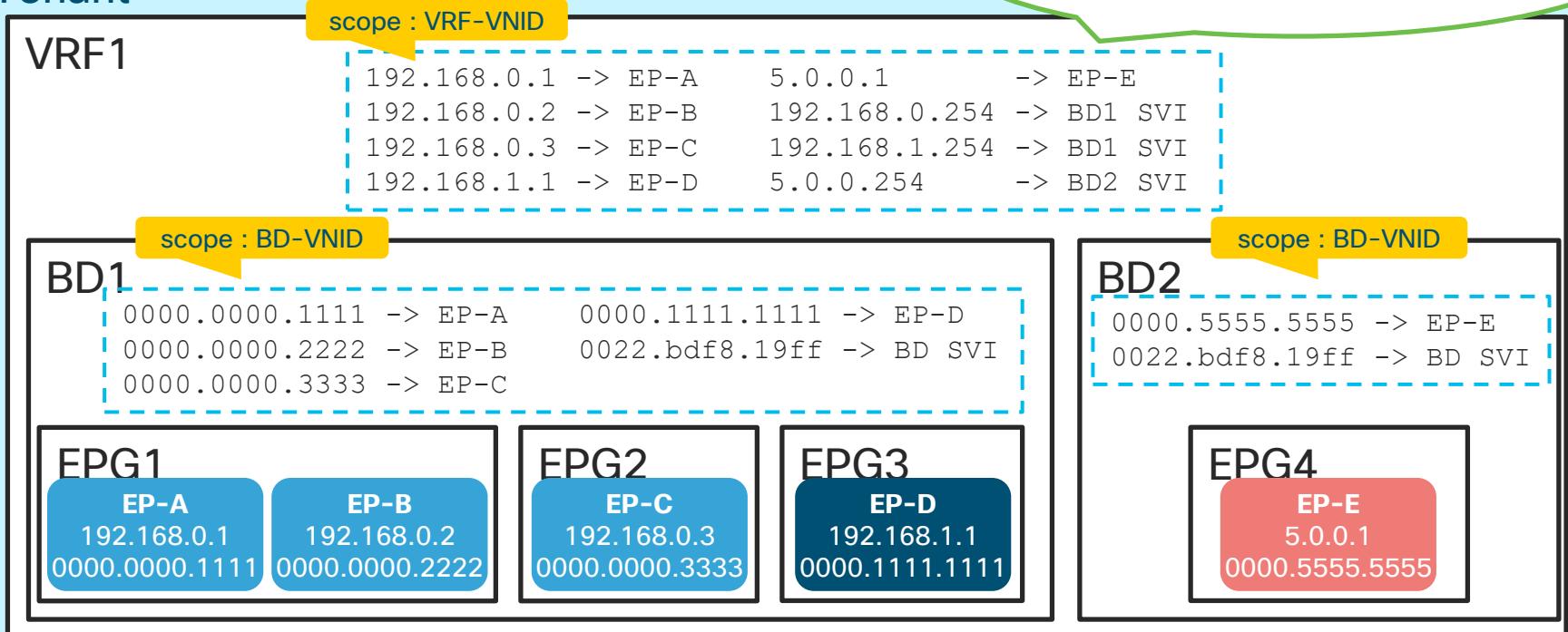
With the contract, ACI knows the LEAF needs to reach out to 5.0.0.0/24

ACI Forwarding Component 2

- Pervasive Gateway (BD SVI)
- **Forwarding Scope (VRF or BD)**
- Forwarding mode in BD

Forwarding Scope

Tenant



Forwarding Scope

L2 traffic(=same subnet) use only MAC
hence BD lookup only

Tenant

VRF1

scope : VRF-VNID

192.168.0.1	->	EP-A	5.0.0.1	->	EP-E
192.168.0.2	->	EP-B	192.168.0.254	->	BD1 SVI
192.168.0.3	->	EP-C	192.168.1.254	->	BD1 SVI
192.168.1.1	->	EP-D	5.0.0.254	->	BD2 SVI

BD1

scope : BD-VNID

0000.0000.1111	->	EP-A	0000.1111.1111	->	EP-D
0000.0000.2222	->	EP-B	0022.bdf8.19ff	->	BD SVI
0000.0000.3333	->	EP-C			

EPG1

EP-A

192.168.0.1

0000.0000.1111

EP-B

192.168.0.2

0000.0000.2222

EPG2

EP-C

192.168.0.3

0000.0000.3333

EPG3

EP-D

192.168.1.1

0000.1111.1111

EPG4

EP-E

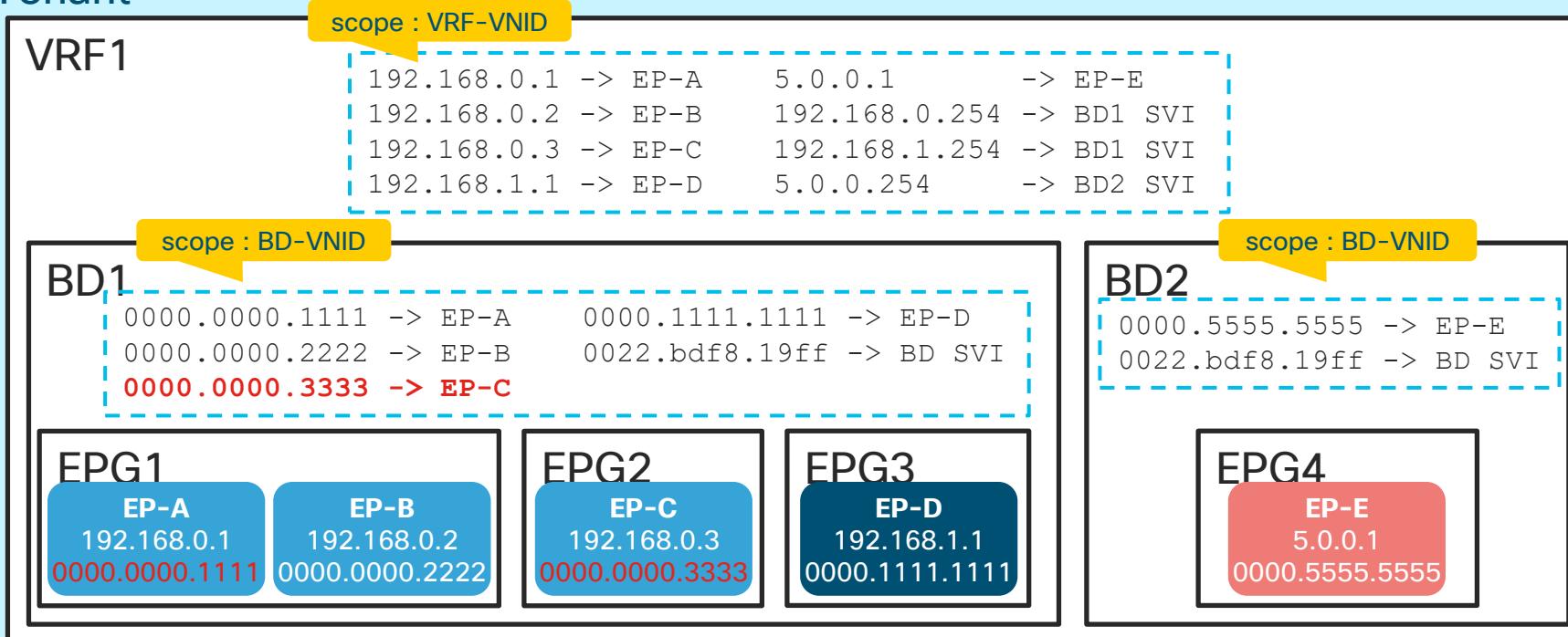
5.0.0.1

0000.5555.5555

Forwarding Scope

It's same even if EPG is different

Tenant



Forwarding Scope

Tenant

VRF1

scope : VRF-VNID

192.168.0.1	->	EP-A	5.0.0.1	->	EP-E
192.168.0.2	->	EP-B	192.168.0.254	->	BD1 SVI
192.168.0.3	->	EP-C	192.168.1.254	->	BD1 SVI
192.168.1.1	->	EP-D	5.0.0.254	->	BD2 SVI

L3 traffic(=different subnet) use IP Lookup

1. Dst MAC hits default gw svi mac

2. IP Lookup in VRF

even though EPs are in the same BD

BD1

scope : BD-VNID

0000.0000.1111	->	EP-A	0000.1111.1111	->	EP-D
0000.0000.2222	->	EP-B	0022.bdf8.19ff	->	BD SVI
0000.0000.3333	->	EP-C			

EPG1

EP-A

192.168.0.1

0000.0000.1111

EP-B

192.168.0.2

0000.0000.2222

EPG2

EP-C

192.168.0.3

0000.0000.3333

EPG3

EP-D

192.168.1.1

0000.1111.1111

EPG4

EP-E

5.0.0.1

0000.5555.5555

Forwarding Scope

It's same even if BD is different

Tenant

VRF1

scope : VRF-VNID

192.168.0.1	->	EP-A	5.0.0.1	->	EP-E
192.168.0.2	->	EP-B	192.168.0.254	->	BD1 SVI
192.168.0.3	->	EP-C	192.168.1.254	->	BD1 SVI
192.168.1.1	->	EP-D	5.0.0.254	->	BD2 SVI

BD1

scope : BD-VNID

0000.0000.1111	->	EP-A	0000.1111.1111	->	EP-D
0000.0000.2222	->	EP-B	0022.bdf8.19ff	->	BD SVI
0000.0000.3333	->	EP-C			

EPG1

EP-A

192.168.0.1

EP-B

192.168.0.2

EPG2

EP-C

192.168.0.3

EPG3

EP-D

192.168.1.1

EPG4

EP-E

5.0.0.1

BD2

scope : BD-VNID

0000.5555.5555	->	EP-E
0022.bdf8.19ff	->	BD SVI

ACI Forwarding Component 2

- Pervasive Gateway (BD SVI)
- Forwarding Scope (VRF or BD)
- **Forwarding mode in BD**

ACI BD Forwarding Option

The screenshot shows the Cisco ACI Bridge Domain configuration interface for Tenant TK. The left sidebar lists various tenant and network configurations. The main window displays the properties for Bridge Domain BD1, specifically focusing on the L3 Configurations tab.

L3 Configurations Tab: This tab is highlighted with a red box. It contains several configuration options:

- L2 Unknown Unicast:** Options include "Flood" and "Hardware Proxy".
- L3 Unknown Multicast Flooding:** Options include "Flood" and "Optimized Flood".
- Multi Destination Flooding:** Options include "Flood in BD", "Drop", and "Flood in Encapsulation".
- PIM:** A checkbox for PIM is present.
- IGMP Policy:** A dropdown menu with the placeholder "select an option".
- ARP Flooding:** A checked checkbox.
- Endpoint Dataplane Learning:** A checked checkbox.
- Limit IP Learning To Subnet:** A checked checkbox.
- End Point Retention Policy:** A dropdown menu with the placeholder "select a value".
- IGMP Snoop Policy:** A dropdown menu with the placeholder "select a value".

Properties Panel: This panel is highlighted with a blue box and contains the following information:

- Unicast Routing:** A checked checkbox.
- Operational Value for Unicast Routing:** true
- Custom MAC Address:** 00:22:BD:F8:19:FF
- Virtual MAC Address:** Not Configured

Callout Boxes:

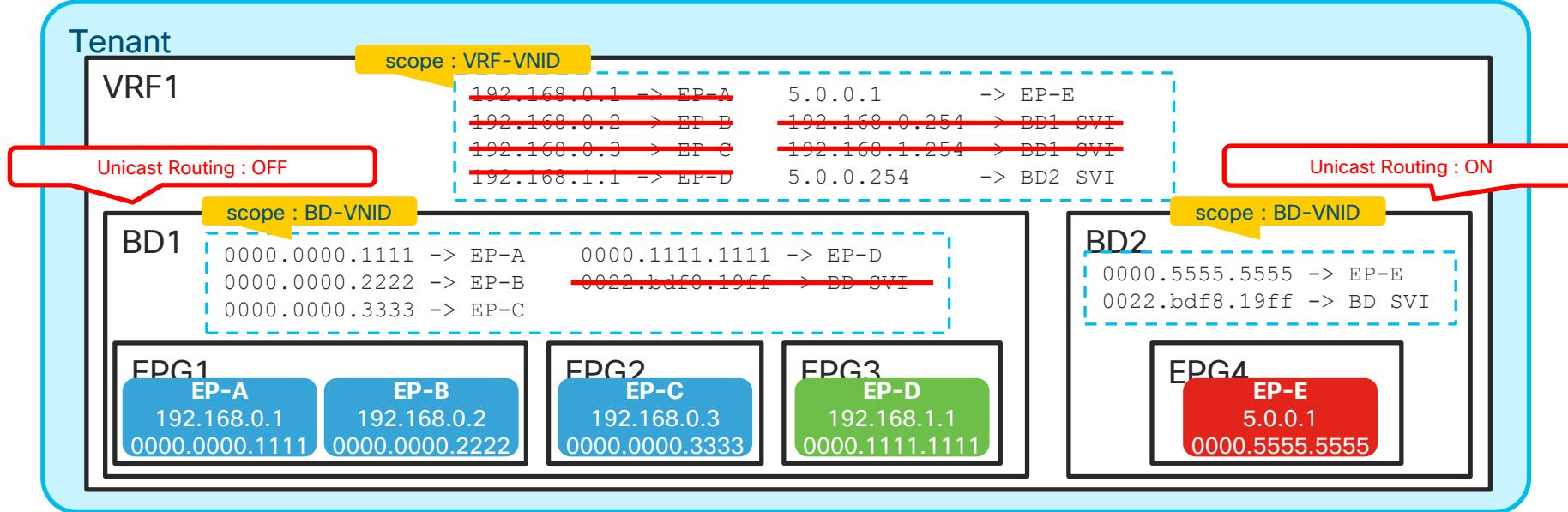
- A green callout box points to the "Unicast Routing" section in the Properties panel, containing the following list:
 - Unicast Routing
 - L2 Unknown Unicast
 - L3 Unknown Multicast Flooding
 - Multi Destination Flooding
 - ARP Flooding
- A red callout box points to the "L2 Unknown Unicast" and "L3 Unknown Multicast Flooding" sections in the L3 Configurations tab.

Note: Please check a whitepaper "ACI Fabric EP Learning" for EP learning options
<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739989.html>

Unicast Routing

Unicast Routing:

On



If Unicast Routing is disabled, (BD1 in above)

- IP Learning is disabled on BD
 - BD SVI is disabled
- => Only L2 Forwarding is available

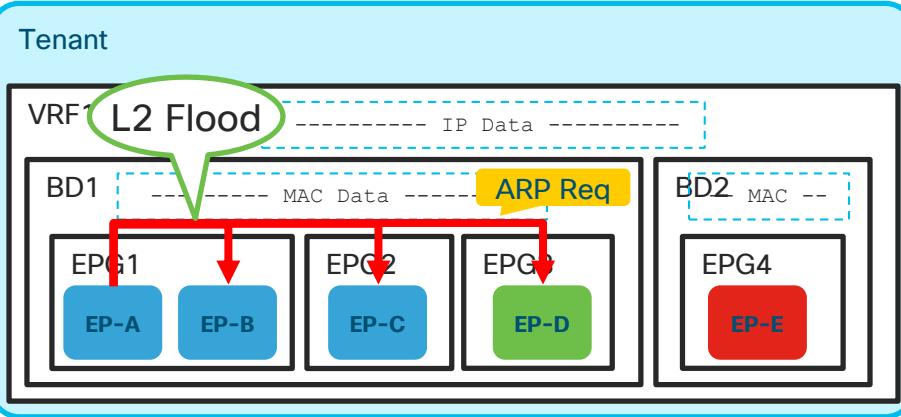
In above example :

- EP-A <-> EP-B : GOOD (L2 forwarding)
- EP-A <-> EP-C : GOOD (L2 forwarding)
- EP-A <-> EP-D : FAIL (L3 forwarding)
- EP-A <-> EP-E : FAIL (L3 forwarding)

ARP Flooding

ARP Flooding: On

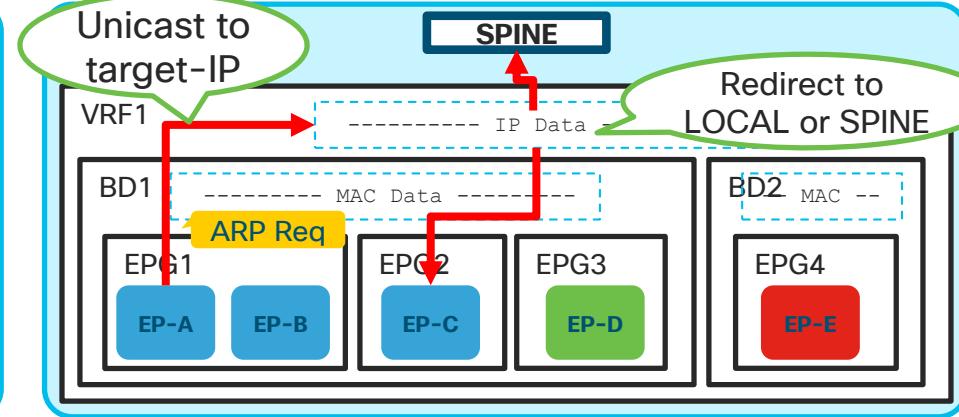
- ARP Flood On



Always **flood** ARP Request **within the same BD**

- Flood as broadcast if DST-MAC is FFFF.FFFF.FFFF
- Flood to other Leaf switches through Spine
- EP IP Data is not used for forwarding but still Sender-IP is learned if Unicast Routing is enabled.
- Good option when BD is supposed to be pure L2 without Unicast Routing like legacy VLAN

- ARP Flood Off (= Spine-Proxy)



ARP Request is handled as **L3 Unicast** with Target-IP

- If IP is **learned** on ingress Leaf,
 - Ingress Leaf forwards ARP Req **directly to dest**
- If IP is **not learned** on ingress Leaf,
 - Ingress Leaf forwards ARP Req **to Spine** Spine-Proxy
 - Spine will forward it to Leaf on which DstIP resides
- If IP is **not learned** even on Spine,
 - Drop and **ARP Glean** (only within BD)

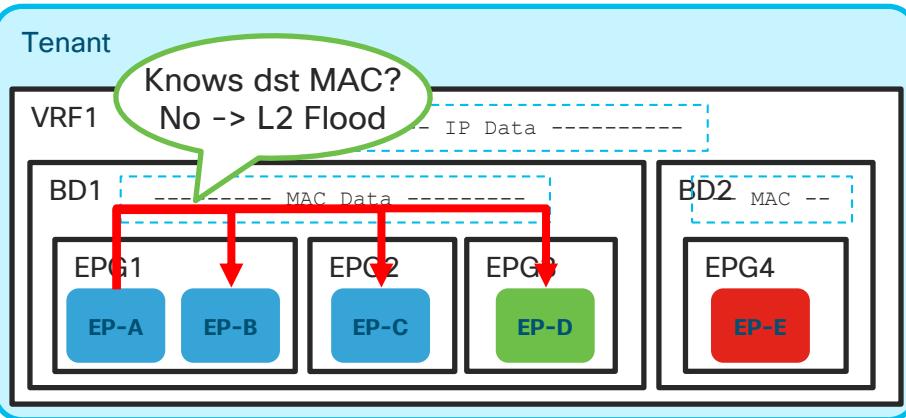
* ARP is not filtered by a contract by default

* if dst mac is not bcast, ARP request is bridged based on dst mac regardless of ARP Flooding mode

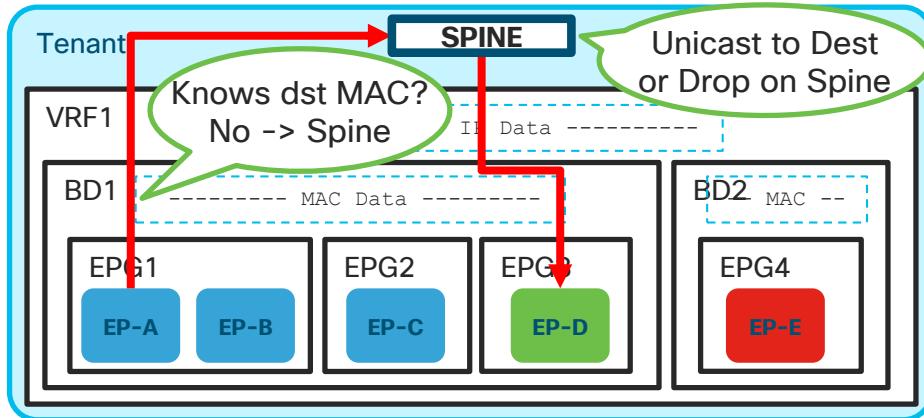
L2 Unknown Unicast

L2 Unknown Unicast: Flood Hardware Proxy

- Flood



- Hardware Proxy (= Spine-Proxy)



Always **flood** L2 Unknown Unicast **within the same BD**

- Flood as well as legacy VLAN.
- Flood happens locally and on other Leaf switches.
- Good option when BD is supposed to be pure L2 without Unicast Routing as in legacy VLAN
- Good option when there are silent L2 hosts

L2 Unknown Unicast is sent to Spine-Prox

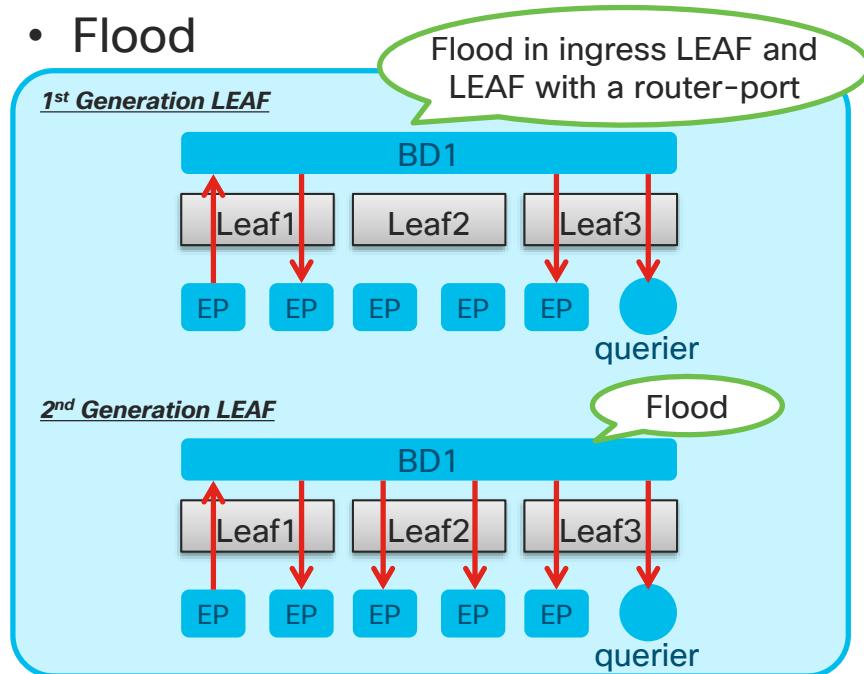
- If DST-MAC is **learned** on Spine,
 - Spine forwards it **directly to dest Leaf**
- If DST-MAC is **not learned** even on Spine
 - Drop

※EP IP Data is not used even if Leaf knows DST-IP. L2 Unknown is still L2 traffic.

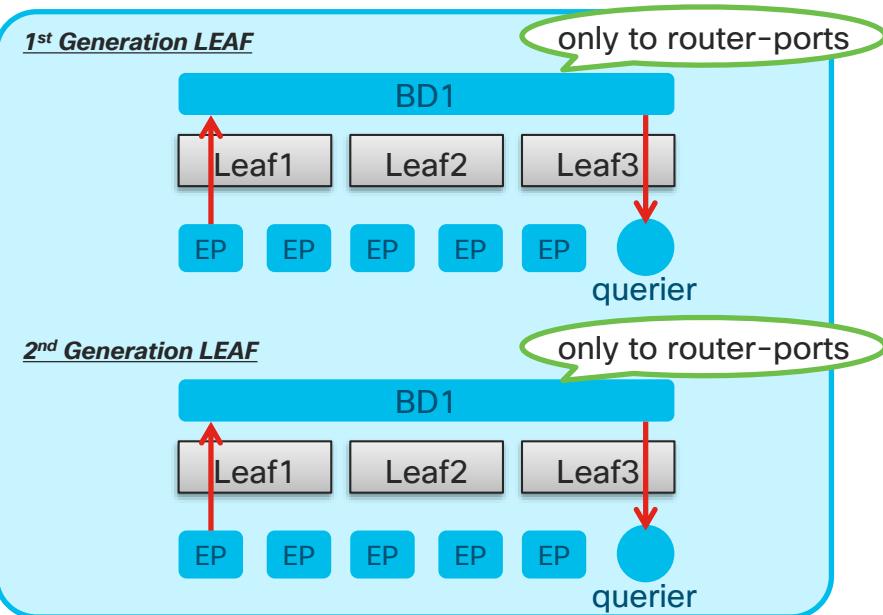
L3 Unknown Multicast Flooding

L3 Unknown Multicast Flooding: Flood Optimized Flood

- Flood



- OMF (Optimized Multicast Flood)



L3 Unknown Multicast = IP multicast group unknown to LEAF IGMP snooping
➤ Controls flooding **unknown** IGMP snooping groups

Multi Destination Flooding

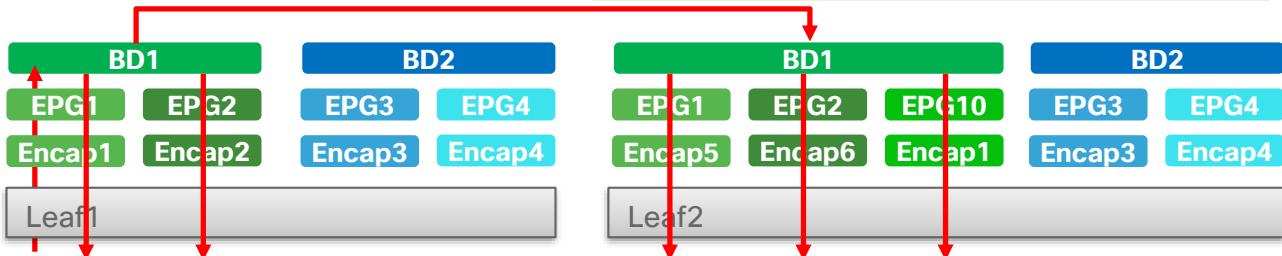
Multi Destination Flooding: Flood in BD Drop Flood in Encapsulation

Flooding mode for L2 multicast, Broadcast and link-local

This mode does not apply to OSPF/OSPFv6, BGP, EIGRP, CDP, LACP, LLDP, ISIS, IGMP, PIM, ST-BPDU, ARP/GARP, RARP, ND

- Flood in BD**

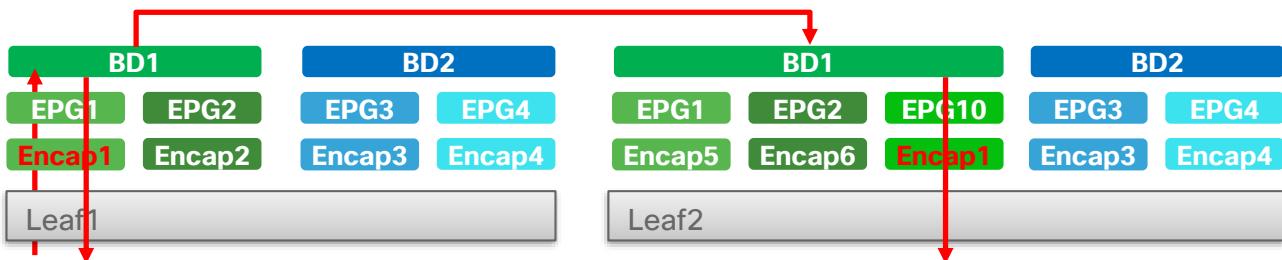
Flood within the same BD regardless of EPG or VLAN.



Behavior change from 3.1

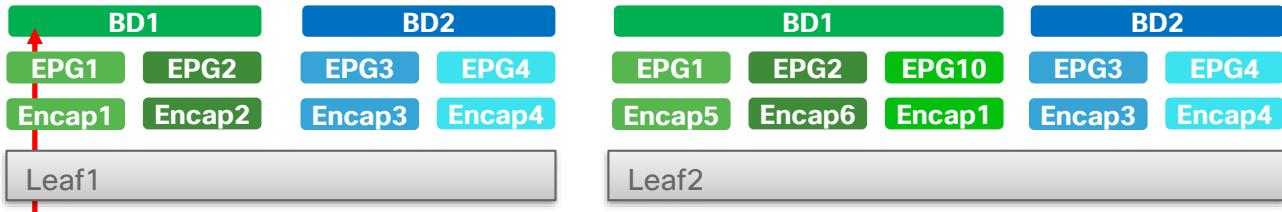
- Flood in Encapsulation**

Flood within the same access encap VLAN and BD regardless of EPG.



- Drop**

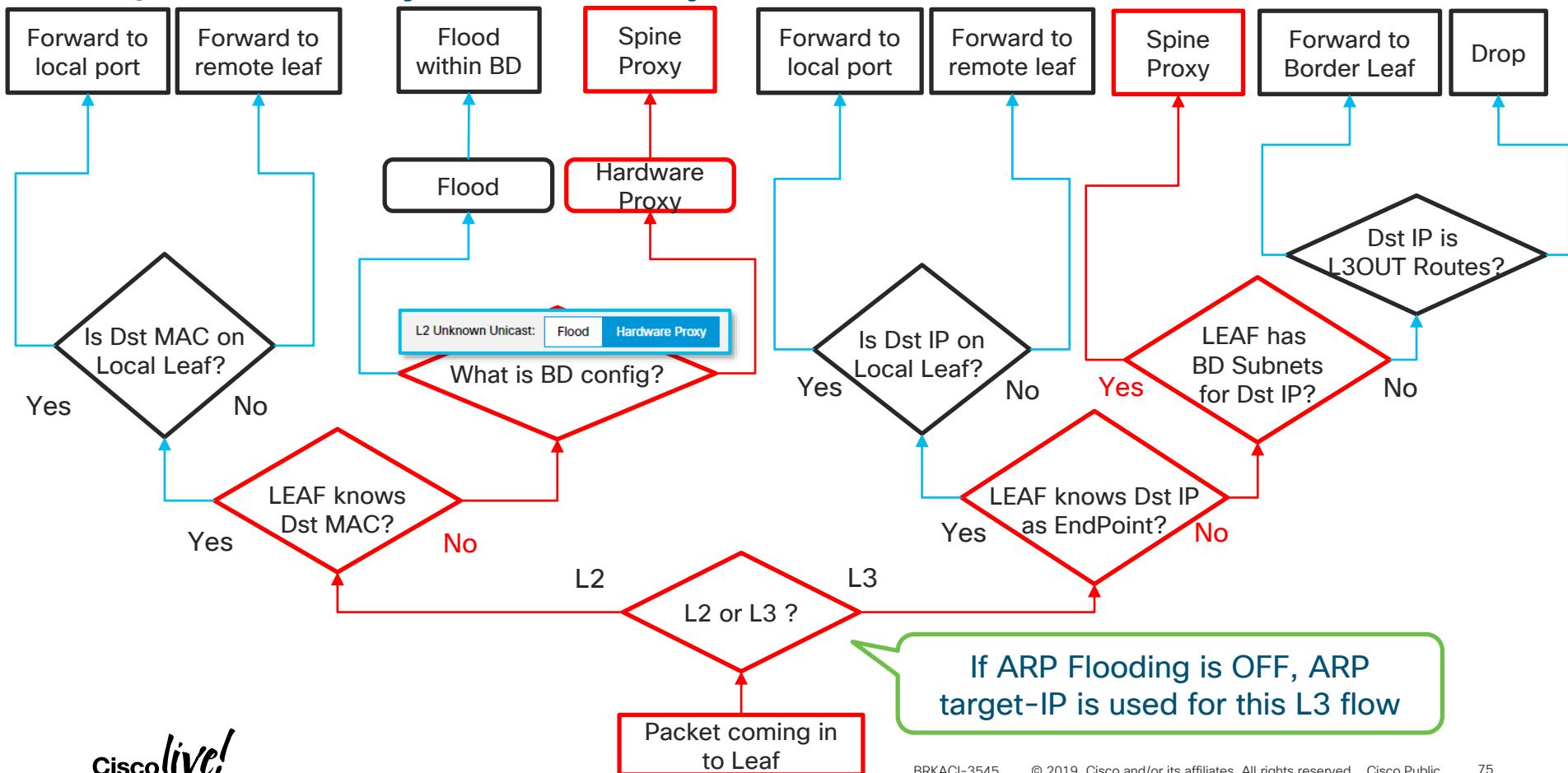
No Flood. Just drop.



Agenda

- Introduction
 - ACI Overlay VxLAN and TEP
- ACI Forwarding components
 - Endpoints, EPG, EP Learning, COOP and How it all works
 - BD, VRF forwarding scope and detailed options
 - Spine-Proxy and ARP Glean
 - Forwarding Software Architecture and ASIC Generation
- ACI Packet Walk
 - Walk through the life of a packet going through ACI

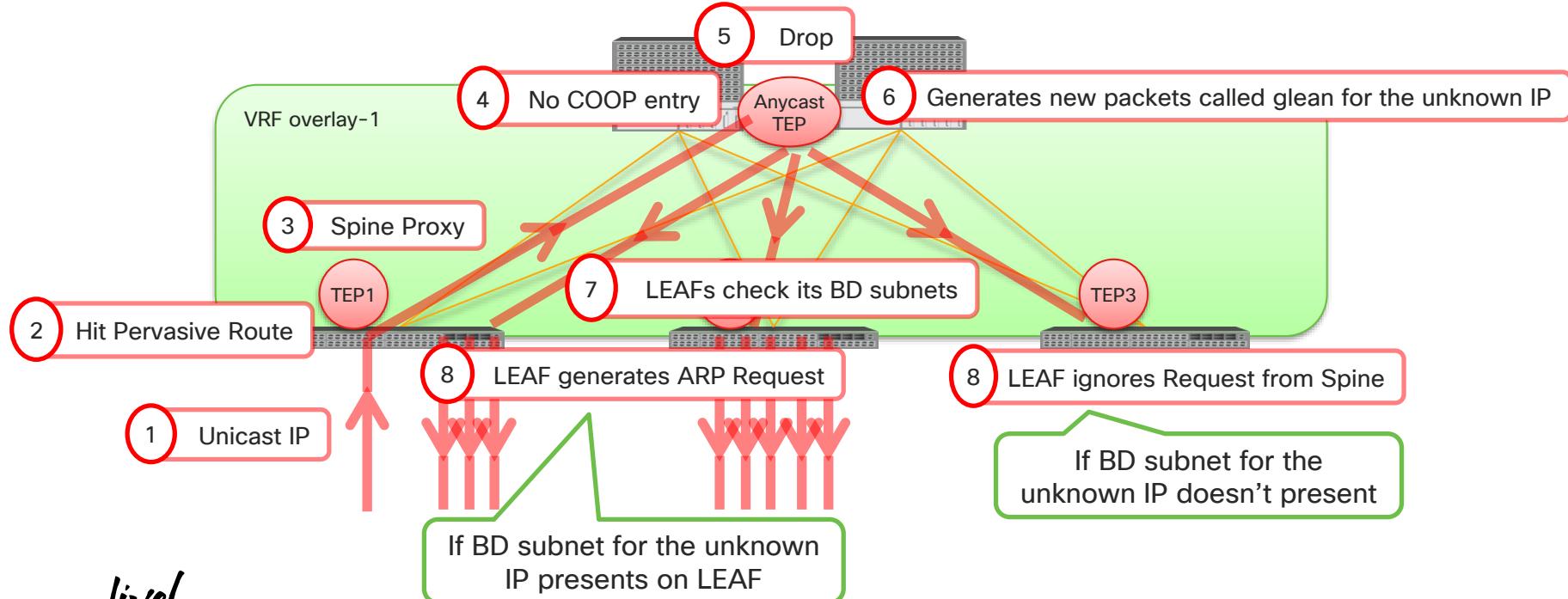
Spine Proxy Summary



ARP Glean (Silent Host Tracking)

What if even SPINE COOP doesn't know the destination when proxy'ed?

- ✓ L2 Traffic : Drop
- ✓ L3 Traffic : ARP Glean



Agenda

- Introduction
 - ACI Overlay VxLAN and TEP
- ACI Forwarding components
 - Endpoints, EPG, EP Learning, COOP and How it all works
 - BD, VRF forwarding scope and detailed options
 - Spine-Proxy and ARP Glean
 - Forwarding Software Architecture and ASIC Generation
- ACI Packet Walk
 - Walk through the life of a packet going through ACI

ACI Forwarding Table & Software Architecture

on the Supervisor Engine:

EPM (EndPoint Manager): manages host MAC & IP learning

uRIB (Unicast RIB): contains the unicast routing information

Policy Mgr: manages contracts between EPGs or L3OUT.

on the Linecards:

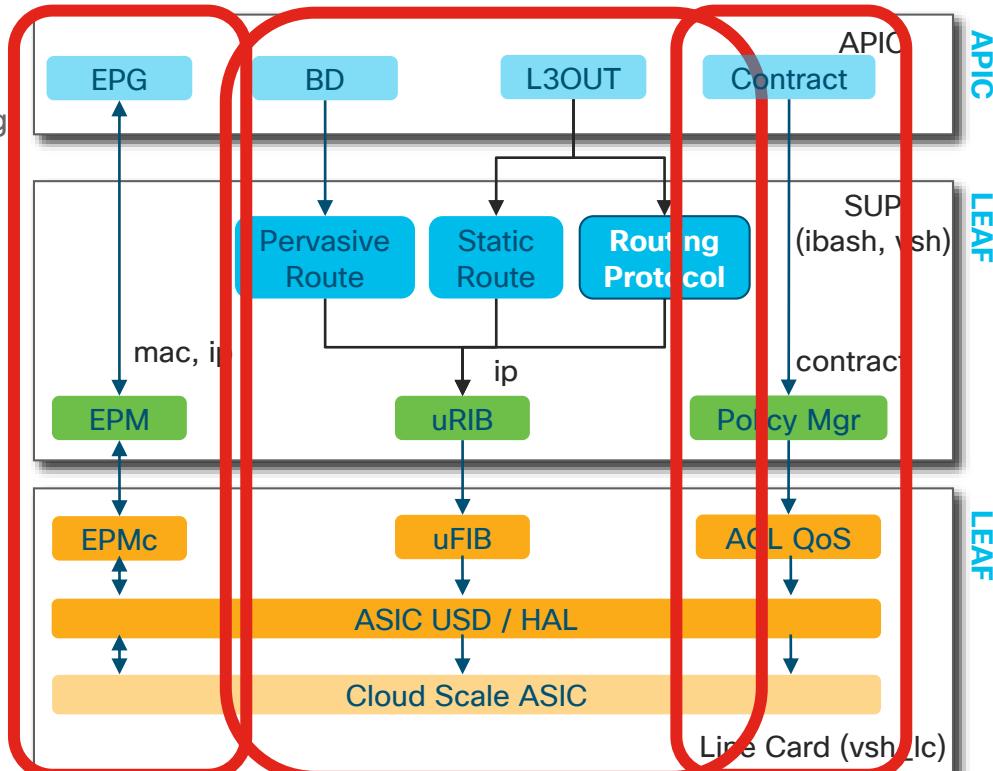
EPMc (EndPoint Manager Client): learns host MAC & IP addresses from hardware(dataplane) via HAL

uFIB (Unicast FIB): programs the hardware unicast routing table via HAL

ACL QoS: programs contracts via HAL

HAL (Hardware Abstraction Layer): pass the messages between hardware(ASIC) and software

Cisco live!



※ ASIC USD (User Space Driver) is only for 1st generation ASIC



ACI Forwarding Table & Software Architecture

on the Supervisor Engine: **ibash (default)**

EPM show endpoint
show system internal epm

uRIB show ip route vrf xxx

Policy Mgr show system internal policymgr

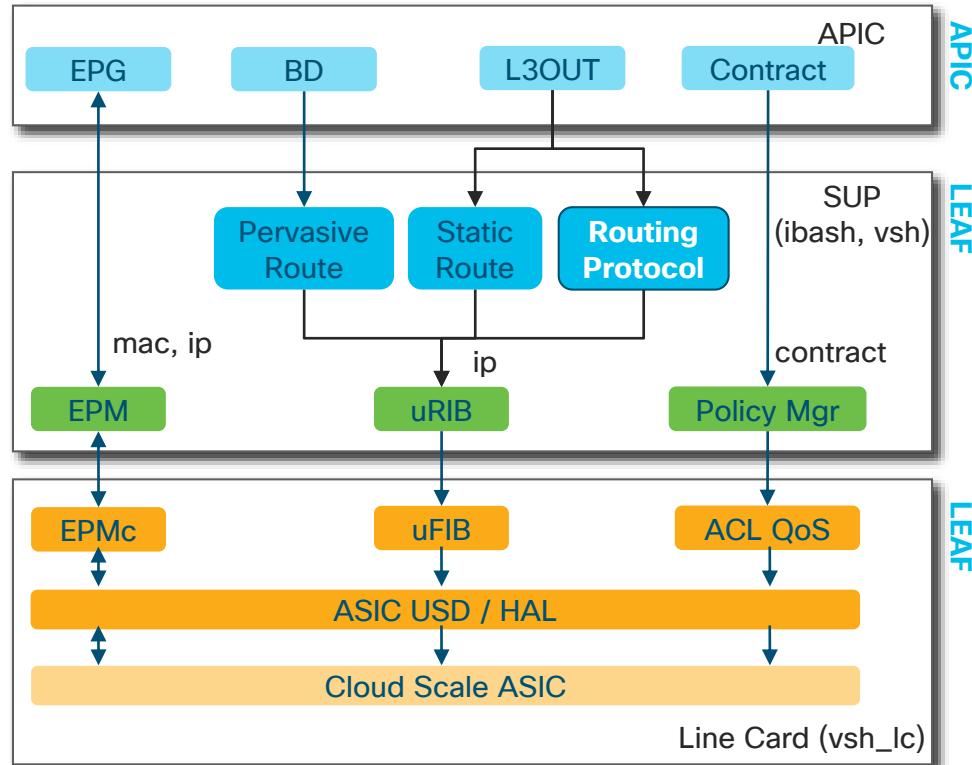
on the Linecards: **vsh_lc**

EPMC show system internal epmc ...

uFIB show forwarding ...

ACL QoS show system internal aclqos ...

HAL show platform internal hal ...

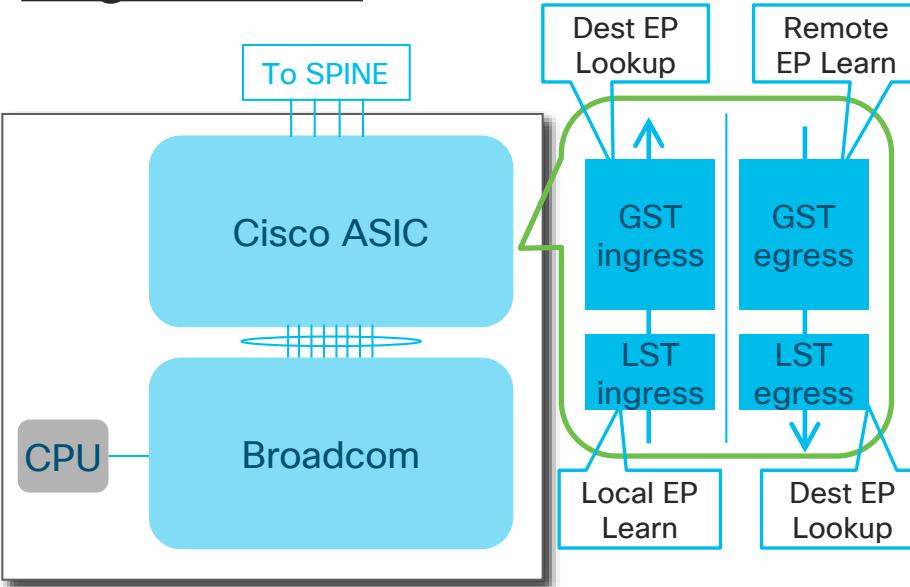


※ ASIC USD (User Space Driver) is only for 1st generation ASIC

LEAF ASIC Generations

※ LST: Local Station Table, GST: Global Station Table
※ FP Tile: Forwarding and Policy Tile
※ HAL: Hardware Abstraction Layer

1st generation

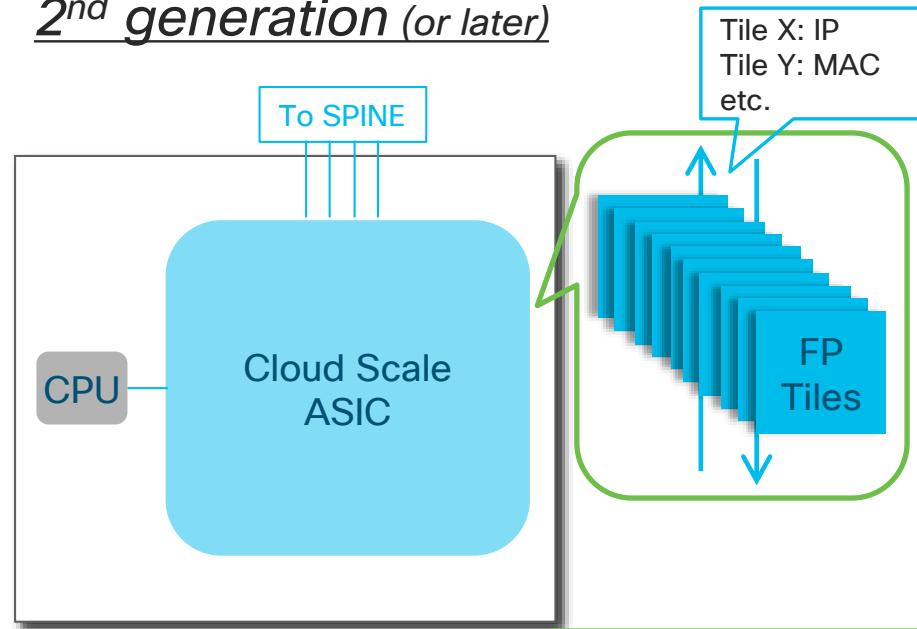


N9K-C9332PQ N9K-C9396PX
N9K-C9372PX N9K-C9396TX
N9K-C9372PX-E N9K-C93120TX
N9K-C9372TX N9K-C93128TX
N9K-C9372TX-E

Cisco live!

- Complete separation of
 - + Ingress and Egress
 - + Source Learn and Destination Lookup
- Separate GST/LST for IP and MAC

2nd generation (or later)

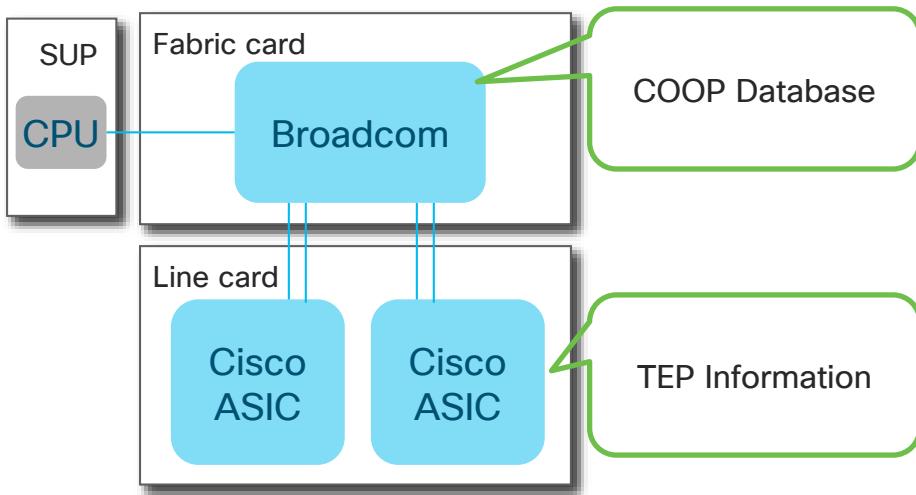


N9K-C93180YC-EX N9K-C93180YC-FX
N9K-C93108TC-EX N9K-C93108TC-FX
N9K-C93180LC-EX N9K-C9348GC-FXP
N9K-C9336C-FX2
N9K-C93240YC-FX2

- More flexible/scalable with configurable tiles
- Abstracted with HAL
- Tile X for both source learn and destination lookup

SPINE ASIC Generations

1st generation



Line card

N9K-X9736PQ

Fabric card

N9K-C9504-FM

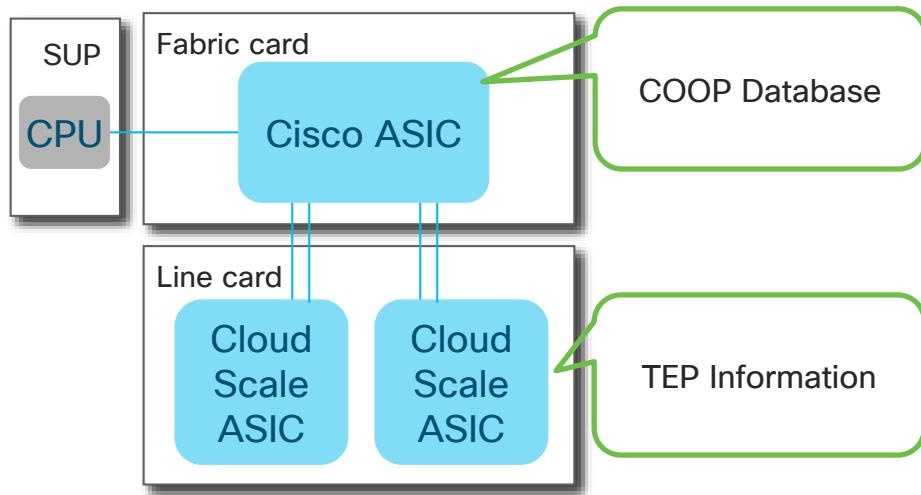
N9K-C9508-FM

N9K-C9516-FM

Box spine

N9K-C9336PQ

2nd generation (or later)



Line card

N9K-X9732C-EX

N9K-X9736C-FX

Fabric card

N9K-C9504FM-E

N9K-C9508FM-E

N9K-C9508FM-E2

N9K-C9516FM-E2

Box spine

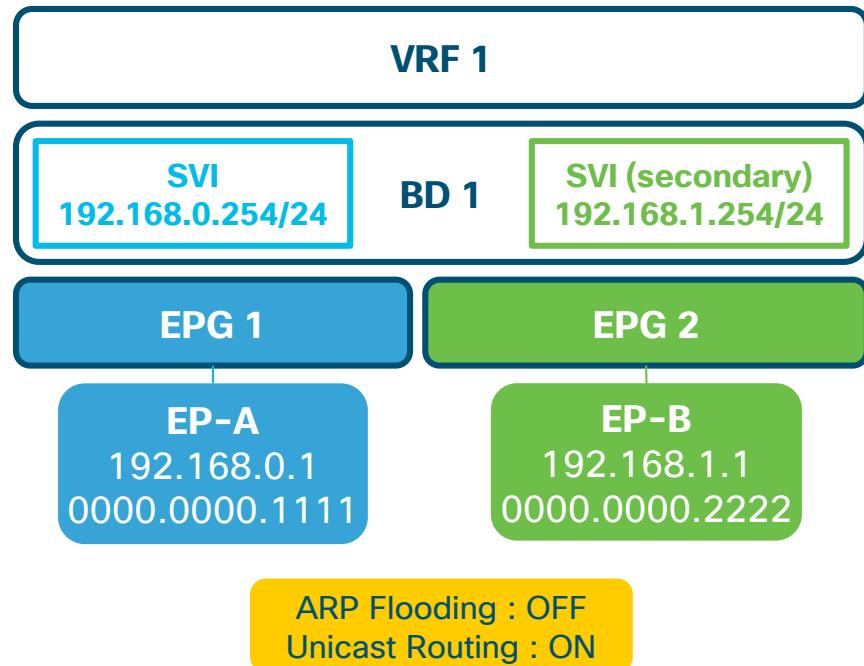
N9K-C9364C

N9K-C9332C

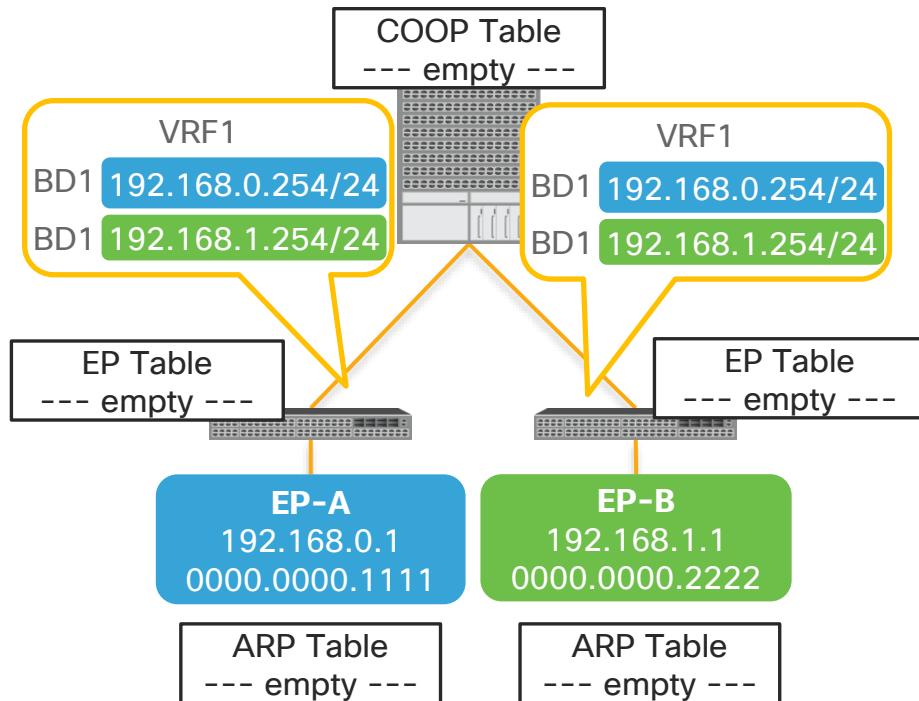
Agenda

- Introduction
 - ACI Overlay VxLAN and TEP
- ACI Forwarding components
 - Endpoints, EPG, EP Learning, COOP and How it all works
 - BD, VRF forwarding scope and detailed options
 - Spine-Proxy and ARP Glean
 - Forwarding Software Architecture and ASIC Generation
- ACI Packet Walk
 - Walk through the life of a packet going through ACI

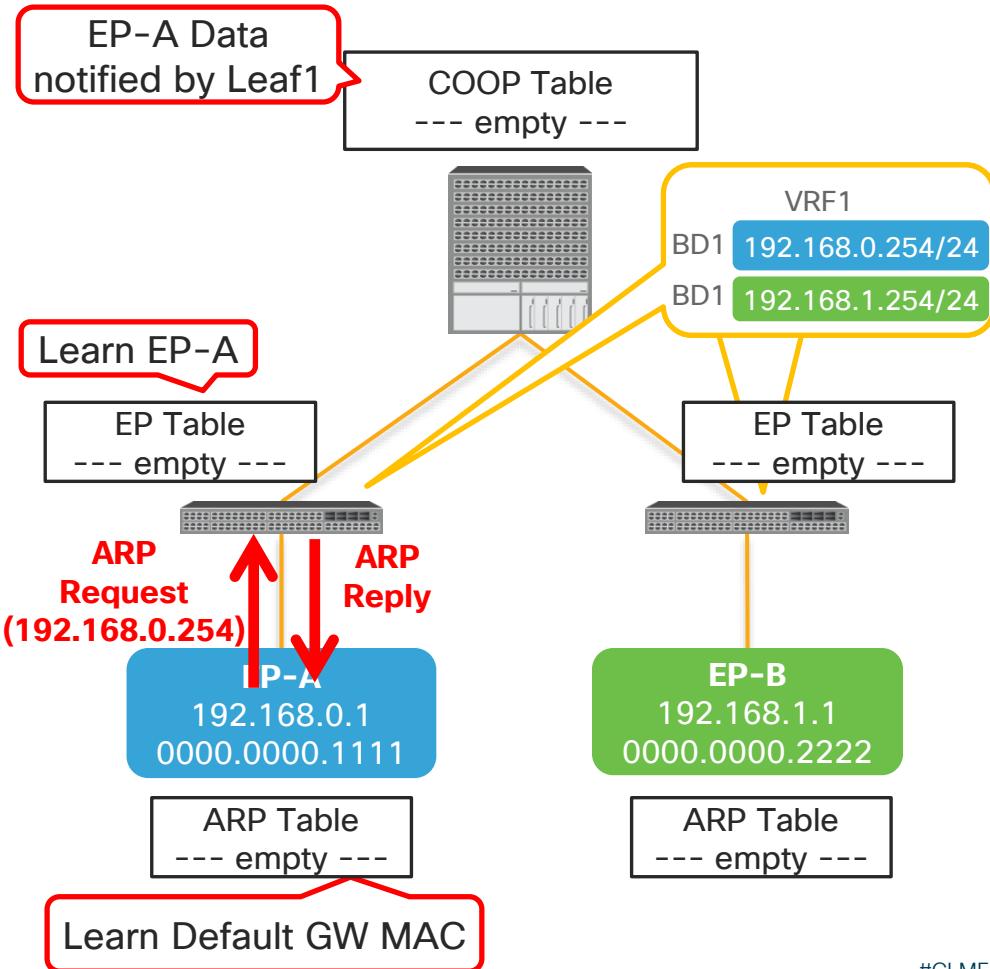
• Logical Topology



• Physical Topology

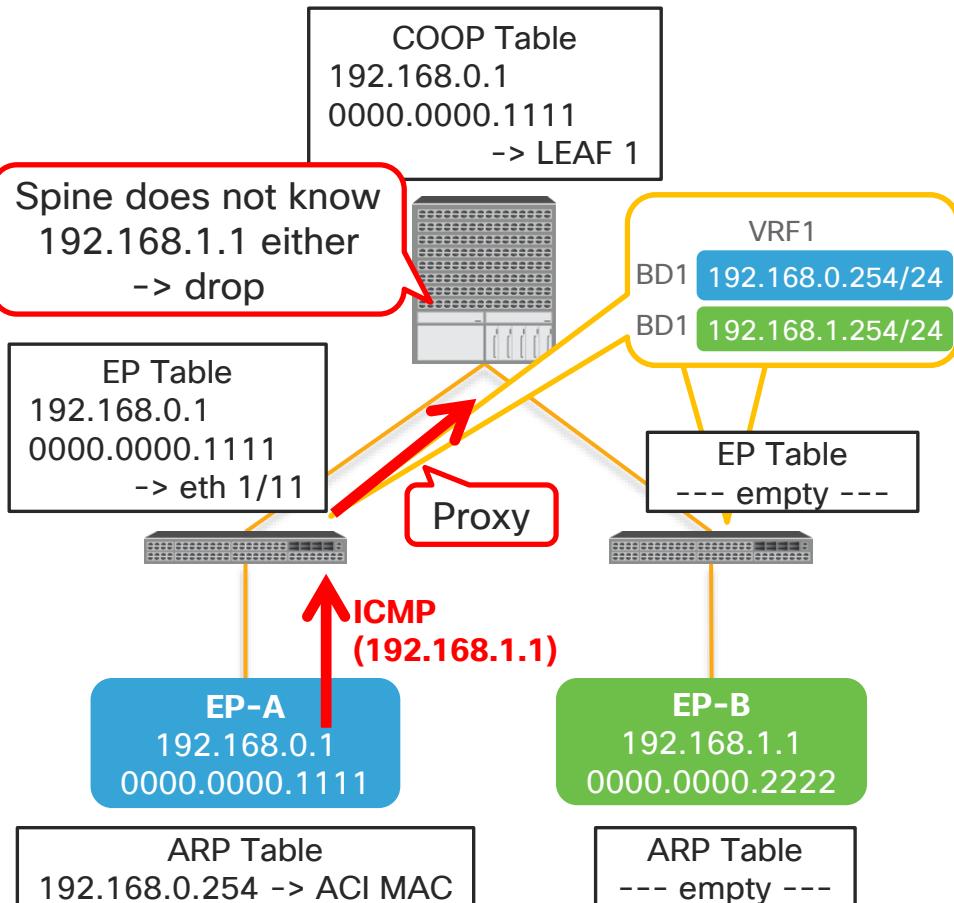


PING : EP-A (192.168.0.1) -> EP-B (192.168.1.1)



1. ARP Request to default GW

1. ARP Req is sent out to GW (192.168.0.254)
2. LEAF1 learns src IP/MAC from ARP.
➤ Leaf1 notify that to Spine COOP
3. LEAF1 sends ARP reply to EP-A.



1. ARP Request to default GW

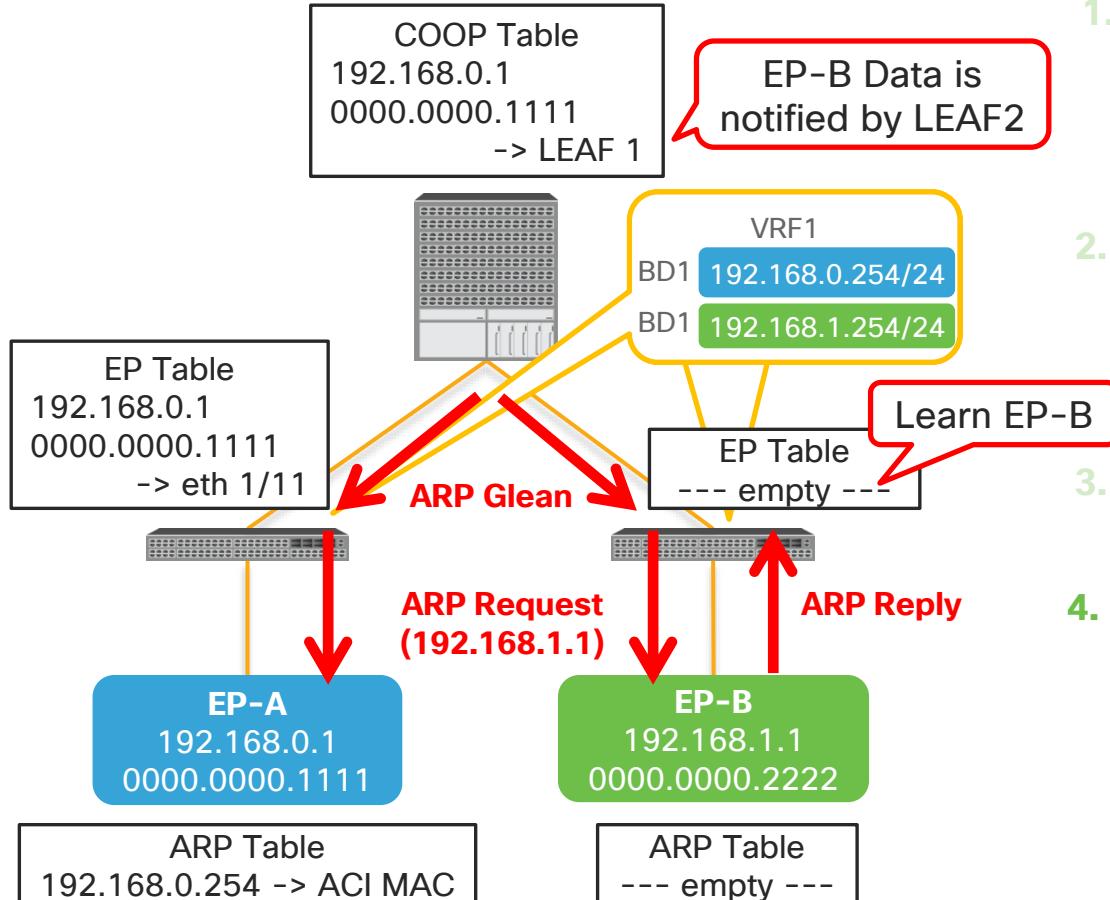
1. ARP Req is sent out to GW (192.168.0.254)
2. LEAF1 learns src IP/MAC from ARP.
 - Leaf1 notify that to Spine COOP
3. LEAF1 sends ARP reply to EP-A.

2. ICMP from EP-A to EP-B (192.168.1.1)

1. Dst MAC is ACI MAC (BD SVI router-mac)
 - L3 Lookup within VRF
2. LEAF1 doesn't know 192.168.1.1 but knows it's subnet (192.168.1.0/254)
 - Spine-Proxy

3. Spine COOP lookup

1. COOP doesn't know 192.168.1.1 either
 - drop



1. ARP Request to default GW

1. ARP Req is sent out to GW (192.168.0.254)
2. LEAF1 learns src IP/MAC from ARP.
 - Leaf1 notify that to Spine COOP
3. LEAF1 sends ARP reply to EP-A.

2. ICMP from EP-A to EP-B (192.168.1.1)

1. Dst MAC is ACI MAC (BD SVI router-mac)
 - L3 Lookup within VRF
2. LEAF1 doesn't know 192.168.1.1 but knows it's subnet (192.168.1.0/254)
 - Spine-Proxy

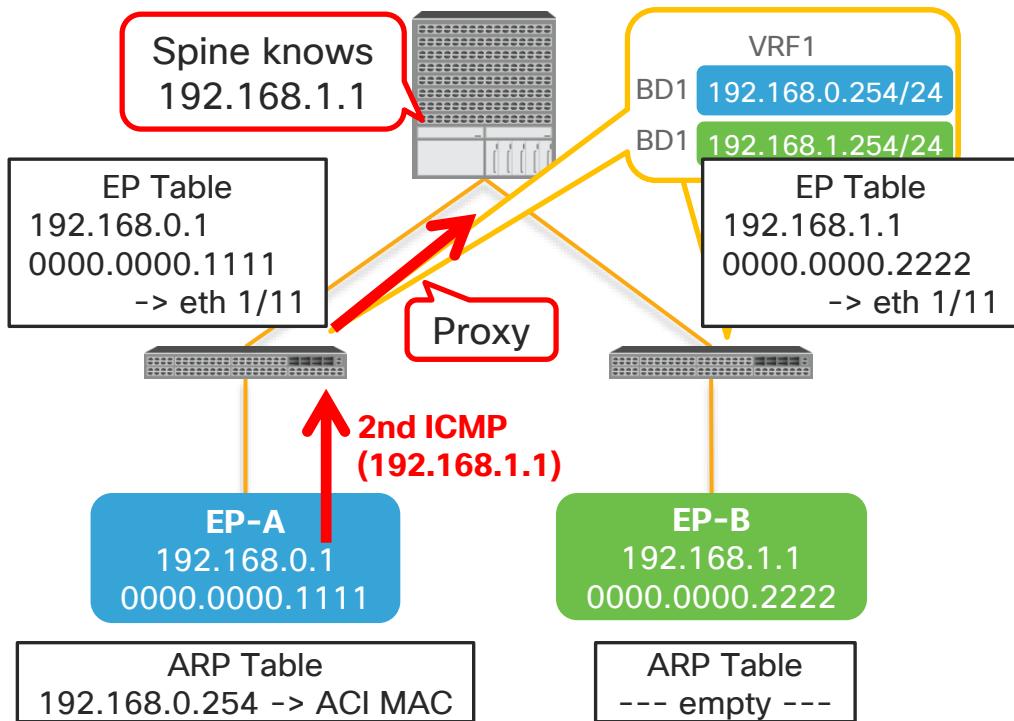
3. Spine COOP lookup

1. COOP doesn't know 192.168.1.1 either
 - drop

4. ARP Glean for 192.168.1.1 to each LEAFs

1. LEAF1 and LEAF2 has a BD with 192.168.1.0/24 subnet
 - Both LEAFs generates an ARP Request for 192.168.1.1 out of ports on the BD
2. EP-B sends ARP Reply to LEAF2
3. LEAF2 learns EP-B IP/MAC
 - LEAF2 notifies that to Spine COOP

COOP Table	
192.168.0.1	192.168.1.1
0000.0000.1111	0000.0000.2222
-> LEAF 1	-> LEAF 2



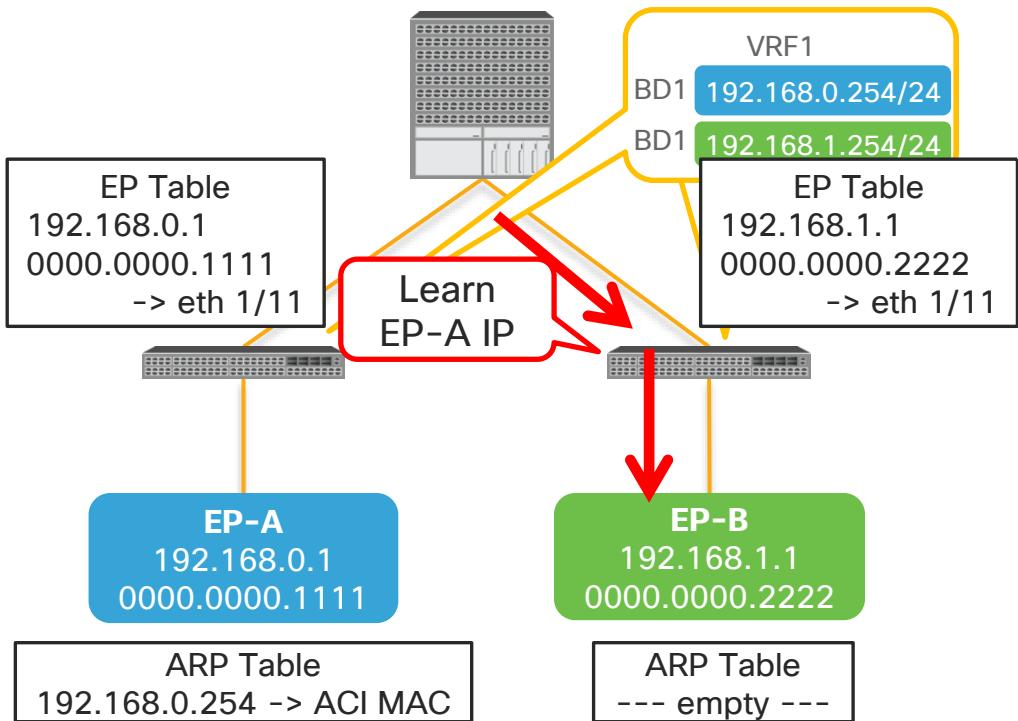
4. EP-A sends 2nd ICMP to EP-B (192.168.1.1)

1. Dst MAC is ACI MAC (BD SVI router-mac)
 - L3 Lookup within VRF
2. LEAF1 still doesn't know 192.168.1.1 but knows it's subnet (192.168.1.0/254)
 - Spine-Proxy

5. Spine COOP lookup for 2nd ICMP

1. Now COOP knows 192.168.1.1
2. Spine sends it to Leaf2

COOP Table	
192.168.0.1	192.168.1.1
0000.0000.1111	0000.0000.2222
-> LEAF 1	-> LEAF 2



4. EP-A sends 2nd ICMP to EP-B (192.168.1.1)

1. Dst MAC is ACI MAC (BD SVI router-mac)
 - L3 Lookup within VRF
2. LEAF1 still doesn't know 192.168.1.1 but knows it's subnet (192.168.1.0/254)
 - Spine-Proxy

5. Spine COOP lookup for 2nd ICMP

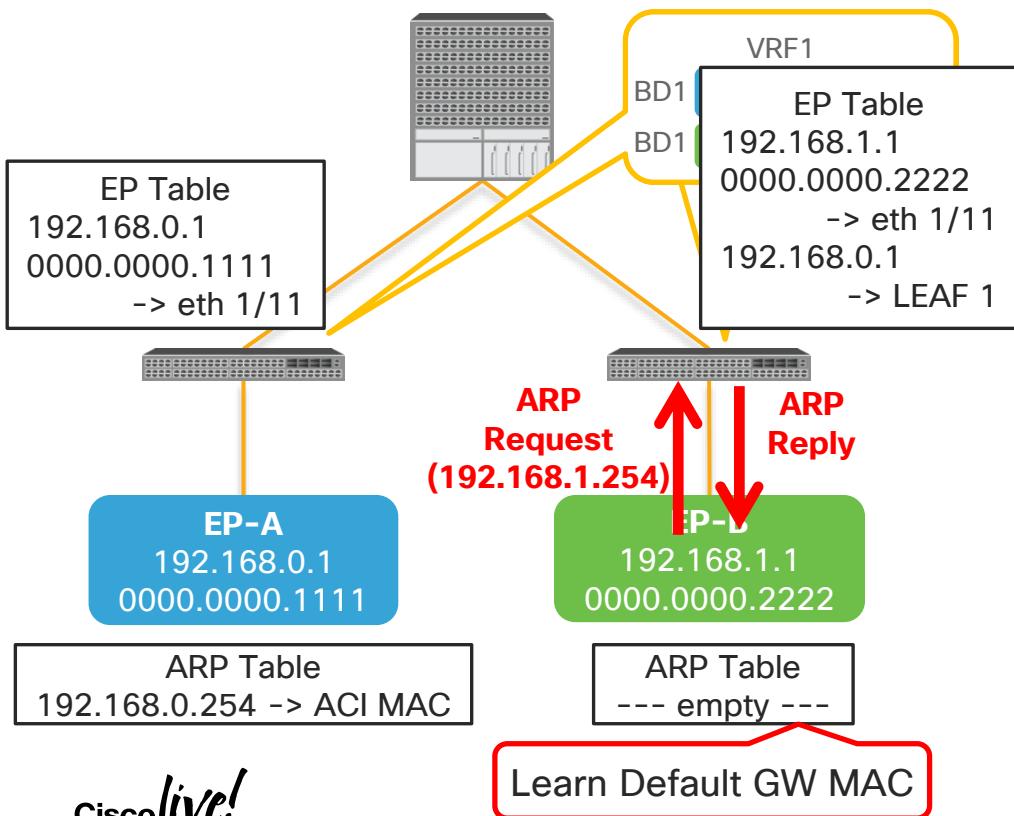
1. Now COOP knows 192.168.1.1
2. Spine sends it to Leaf2

6. LEAF2 learns EP-A as a remote EP

- The packet is routed = sent out with VRF VNID.
- Only IP is learned

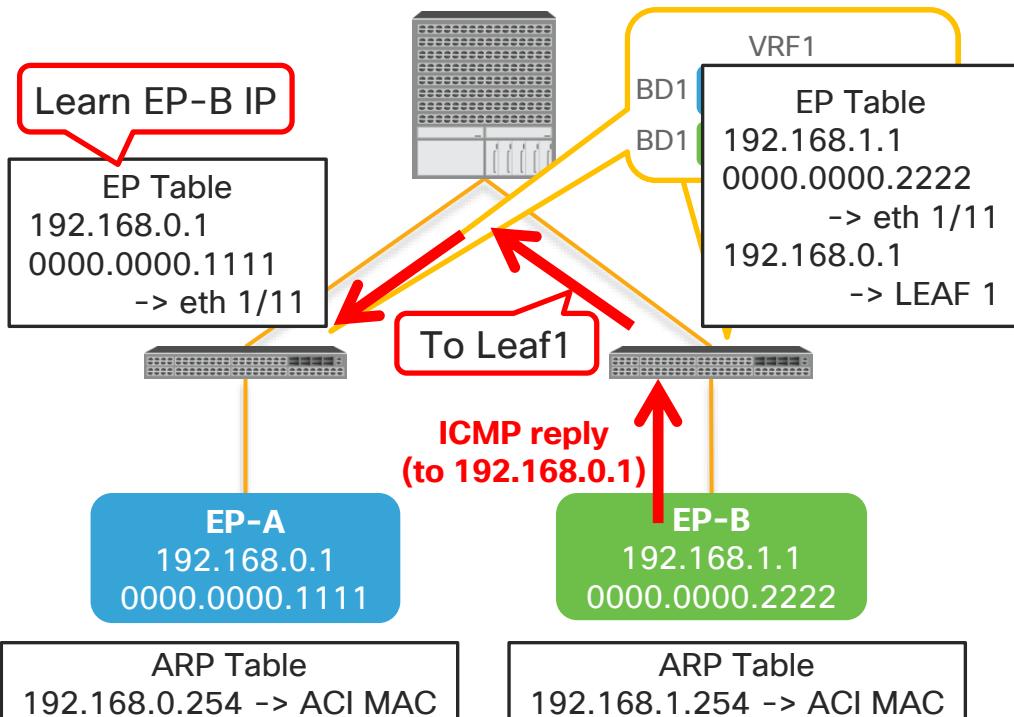
7. LEAF2 sends it out to EP-B

COOP Table	
192.168.0.1	192.168.1.1
0000.0000.1111	0000.0000.2222
-> LEAF 1	-> LEAF 2



4. EP-A sends 2nd ICMP to EP-B (192.168.1.1)
 1. Dst MAC is ACI MAC (BD SVI router-mac)
 - L3 Lookup within VRF
 2. LEAF1 still doesn't know 192.168.1.1 but knows it's subnet (192.168.1.0/254)
 - Spine-Proxy
5. Spine COOP lookup for 2nd ICMP
 1. Now COOP knows 192.168.1.1
 2. Spine sends it to Leaf2
6. LEAF2 learns EP-A as a remote EP
 - The packet is routed = sent out with VRF VNID.
 - Only IP is learned
7. LEAF2 sends it out to EP-B
8. EP-B resolves ARP for its gateway (192.168.1.254)

COOP Table	
192.168.0.1	192.168.1.1
0000.0000.1111	0000.0000.2222
-> LEAF 1	-> LEAF 2



4. EP-A sends 2nd ICMP to EP-B (192.168.1.1)

1. Dst MAC is ACI MAC (BD SVI router-mac)
 - L3 Lookup within VRF
2. LEAF1 still doesn't know 192.168.1.1 but knows it's subnet (192.168.1.0/254)
 - Spine-Proxy

5. Spine COOP lookup for 2nd ICMP

1. Now COOP knows 192.168.1.1
2. Spine sends it to Leaf2

6. LEAF2 learns EP-A as a remote EP

- The packet is routed = sent out with VRF VNID.
- Only IP is learned

7. LEAF2 sends it out to EP-B

8. EP-B resolves ARP for its gateway (192.168.1.254)

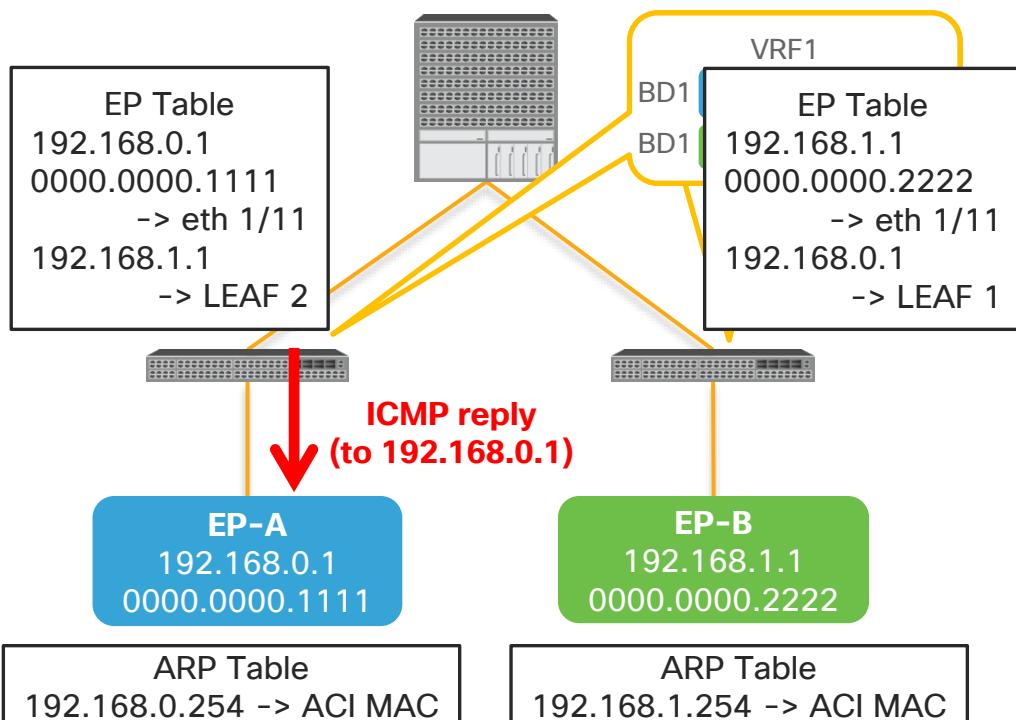
9. EP-B sends ICMP reply

1. LEAF2 already knows where EP-A IP is
 - Directly sends it to LEAF1

10. LEAF1 learns EP-B IP as a remote EP

- Only IP is learned as well

COOP Table	
192.168.0.1	192.168.1.1
0000.0000.1111	0000.0000.2222
-> LEAF 1	-> LEAF 2



4. EP-A sends 2nd ICMP to EP-B (192.168.1.1)

1. Dst MAC is ACI MAC (BD SVI router-mac)
 - L3 Lookup within VRF
2. LEAF1 still doesn't know 192.168.1.1 but knows it's subnet (192.168.1.0/254)
 - Spine-Proxy

5. Spine COOP lookup for 2nd ICMP

1. Now COOP knows 192.168.1.1
2. Spine sends it to Leaf2

6. LEAF2 learns EP-A as a remote EP

- The packet is routed = sent out with VRF VNID.
Only IP is learned

7. LEAF2 sends it out to EP-B

8. EP-B resolves ARP for its gateway (192.168.1.254)

9. EP-B sends ICMP reply

1. LEAF2 already knows where EP-A IP is
 - Directly sends it to LEAF1

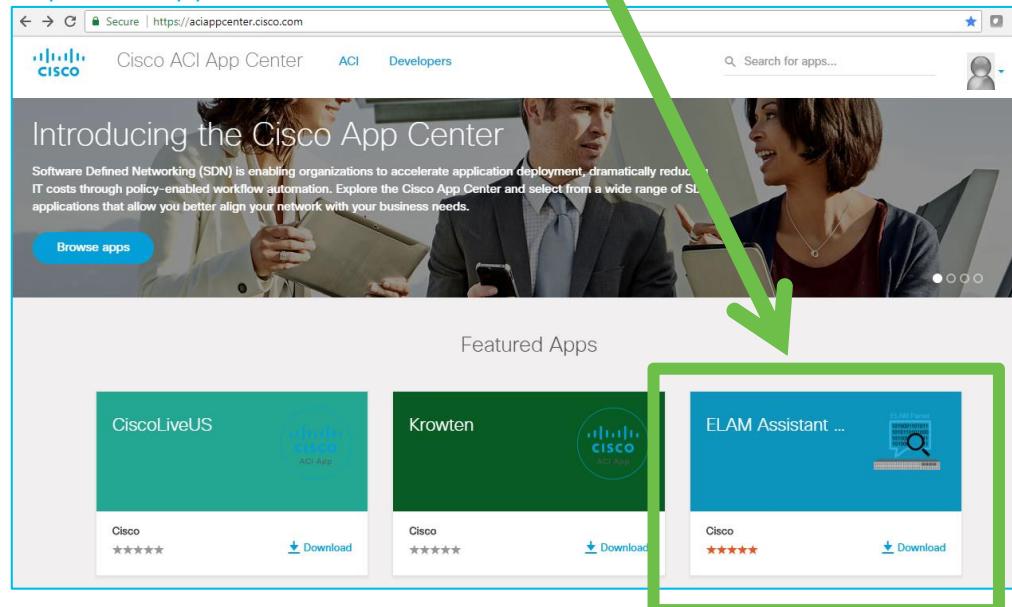
10. LEAF1 learns EP-B IP as a remote EP

- Only IP is learned as well

ELAM Assistant in ACI AppCenter

Interested in more detail packet forwarding verification ?
➤ ELAM Assistant!!

<https://aciappcenter.com>



ELAM (Embedded Logic Analyzer Module)

- Perform an ASIC level packet capture

ELAM Assistant

- You can perform ELAM like a TAC engineer!
- With a nicely formatted result report

Detail Explanations:

- <https://aciappcenter.cisco.com/elam-assistant-beta-2-1n.html>
 - How to use video, pictures
 - A download link for ELAM Assistant
- <https://learningnetwork.cisco.com/docs/DOC-34985>
 - ACI webinar for ELAM Assistant

ELAM Assistant in ACI AppCenter (example)

1. Perform ELAM

Set Parameters

Triggered!!

Capture a packet with ELAM (Embedded Logic Analyzer Module)

ELAM PARAMETERS

name your capture : (optional)

Status	Node	Direction	Source I/F	Parameters (Outer Header)	(Inner Header)
Triggered	node-101	from frontport	any	dst_ip 192.168.1.4 src_ip 192.168.1.1	+ + + +
Set	node-102	from SPINE	any	dst_ip 192.168.1.4 src_ip 192.168.1.1	+ + + +
Triggered	node-103	from SPINE	any	dst_ip 192.168.1.4 src_ip 192.168.1.1	+ + + +

▶ Set ELAM(s) ⚙ Check Trigger

admin

APIC

System Tenants Fabric Virtual Networking L4-L7 Services Admin Operations Apps

Apps | Faults

node-101 (fab5-leaf1)
node-102 (fab5-leaf2)
node-103 (fab5-leaf3)
node-104 (fab5-leaf4)
node-105 (fab5-leaf5)
node-107 (fab5-leaf7)
node-201 (fab5-spine1)
slot 1 (N9K-X9732C-EX)

ELAM Assistant in ACI AppCenter (example)

2. Read a report

Packet Forwarding Information	
Destination is Local port	
Destination Port	eth1/15
Destination Logical Port	Po1
EPG Classification (pcTag)	
Destination EPG pcTag (dclass)	0x4006 (16390)
Source EPG pcTag (sclass)	0x2ABA (10938)
Policy Applied	1 (Contract was applied)
Lookup Drop	
drop reason	no drop

Zoom

The screenshot shows the ELAM Assistant interface in the ACI AppCenter. The main window displays a detailed packet capture for node-103. On the left, a sidebar lists nodes: node-101 (fab05-leaf1), node-102 (fab05-leaf2), node-103 (fab05-leaf3), node-104 (fab05-leaf4), node-105 (fab05-leaf5), node-107 (fab05-leaf7), and node-201 (fab05-spine1). The central pane shows the "ELAM REPORT FILES" section with a file named "node-103_elam_report.txt" from June 9, 2018, at 07:49:51. Below it is the "ELAM REPORT PARSE RESULT" section, which contains extensive details about the captured packet, including trigger information, L2 headers, L3 headers, and L4 headers. A zoomed-in view of the "Packet Forwarding Information" section is shown at the bottom right.

ELAM Assistant

node-103

ELAM REPORT FILES

file name: node-103_elam_report.txt date: 2018-06-09 07:49:51+00:00

ELAM REPORT PARSE RESULT

Captured Packet Information -- [node-103_elam_report.txt]

Trigger Information

Packet Direction	SPINE → LEAF	
Incoming Interface	eth1/50	Slice Source ID(Sa) 0x38 in "show plat int hal12 port god"

Outer L2 Header

Destination MAC	000C.8C0C.8C0C
Source MAC	0000.0000.000D
CoS	0x0 (0)
Access Encap VLAN	0x2 (2)

Outer L3 Header

L3 Type	IPv4
DSCP	0x0 (0)
Don't Fragment Bit	0x0
TTL	0x1F (31)
IP Protocol Number	17 (UDP)
Destination IP	11.0.192.67 fab05-leaf3
Source IP	11.0.192.70 fab05-leaf3

Inner L3 Header

Destination MAC	0000.5454.5454
Source MAC	0000.0000.5151
L3 Type	IPv4
DSCP	0x0 (0)
Don't Fragment Bit	0x0
TTL	0xFF (255)
IP Protocol Number	1 (ICMP)
Destination IP	192.168.1.4
Source IP	192.168.1.1

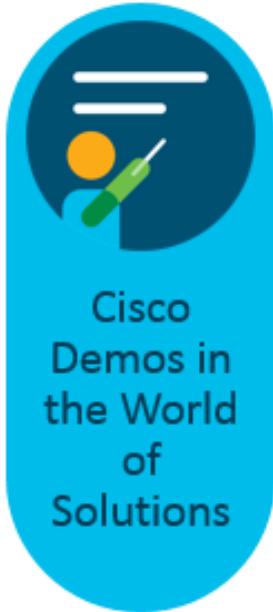
Outer L4 Header

L4 Type	IPvLAN
Don't Learn Bit	0x0
Src Policy Applied Bit	0x1
Dst Policy Applied Bit	0x1
Src EPG (sclass/src pcTag)	0x2ABA (10938) TKA_P1_EPG1-1
VRF or BD VNI	0xF87FAA (16285610) TK_BD1

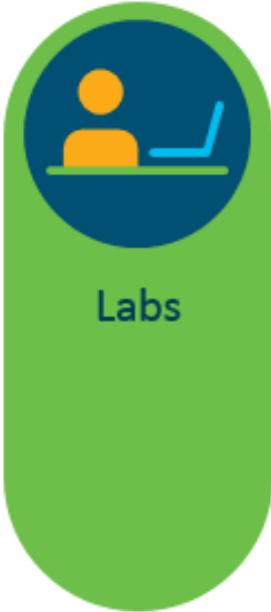
Packet Forwarding Information

Destination is Local port	eth1/15	
Destination Port	eth1/15	Ovec (0x14) in "show plat int hal12 port god"
Destination Logical Port	Po1	LID (0x4C) in "show plat int hal12 port pr"
EPG Classification (pcTag)	0x4006 (16390)	TK_AP1_EPG1-4
Destination EPG pcTag (dclass)	0x4006 (16390)	TK_AP1_EPG1-4
Source EPG pcTag (sclass)	0x2ABA (10938)	TK_AP1_EPG1-4
Policy Applied	1 (Contract was applied)	
Lookup Drop	no drop	

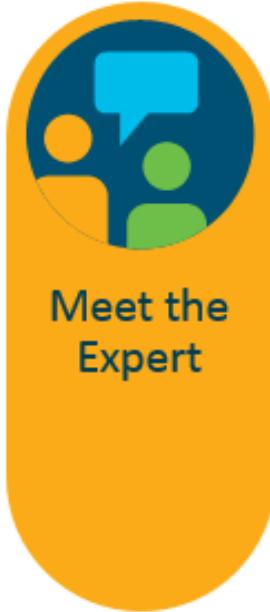
Continue your education



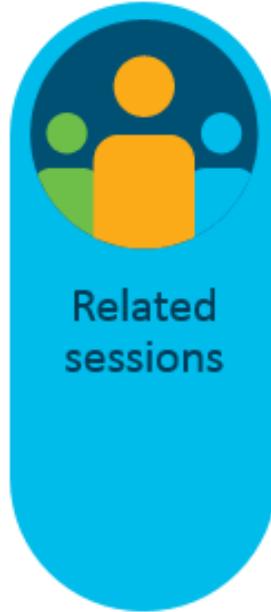
Cisco
Demos in
the World
of
Solutions



Labs



Meet the
Expert



Related
sessions

Complete Your Online Session Evaluation

- Give us your feedback and receive a complimentary **Cisco Live 2019 Power Bank** after completing the overall event evaluation and 5 session evaluations.
- All evaluations can be completed via the Cisco Live Melbourne Mobile App.
- Don't forget: Cisco Live sessions will be available for viewing on demand after the event at:

<https://cisco.com/ciscolive/on-demand-library/>





Thank you



INTUITIVE



INTUITIVE