cisco Live!

ALL IN

#CiscoLive

The bridge to possible

A Network Engineer's Blueprint for ACI Forwarding

Part 1 – Understanding ACI Forwarding

Joseph Young, ACI Technical Leader – Customer Experience

BRKDCN-3900a

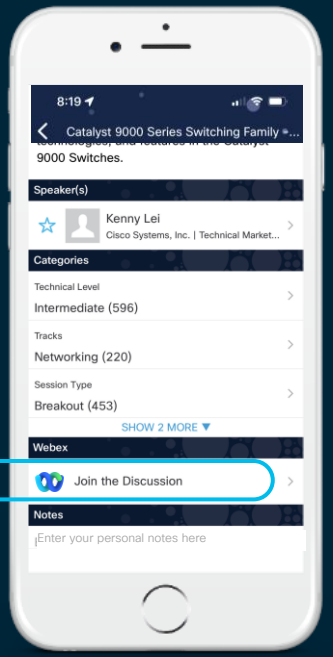
Cisco Webex App

Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space



https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-3900a

Webex spaces will be moderated by the speaker until June 17, 2022.

Agenda

- What's Different About ACI Forwarding?
  - (iVXLAN, contracts, endpoint learning)
- Proxy Forwarding
- ACI Forwarding Tables
  - Endpoint tables, routing tables, hardware lookups
- Understanding the Configuration Options
- The Anatomy of an ACI Switch

Glossary of Acronymns

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

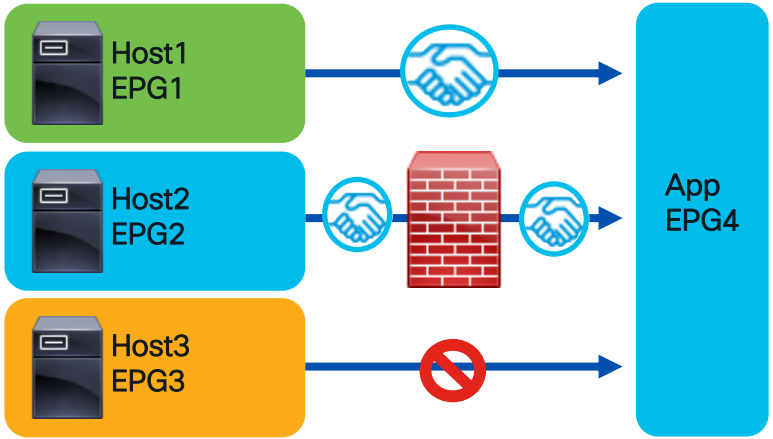
VxLAN packet acronyms

Acronyms	Definitions
d*XXX*o	Outer Destination XXX (dIPo = Outer Destination IP)
s*XXX*o	Outer Source XXX (sIPo = Outer Source IP)
d*XXX*i	Inner Destination XXX (dIPi = Inner Destination IP)
s*XXX*i	Inner Source XXX (sIPi = Inner Source IP)
GIPo	Outer Multicast Group IP
VNID	Virtual Network Identifier

What’s Different About ACI Forwarding?

What is “Application Centric”?

- Traditional networks use ACL’s to classify traffic
  - Usually based on L3 or L2 addresses
  - Makes security decisions (permit, deny, log, etc)
  - Makes forwarding decisions (policy based routing)
- ACI can classify traffic based on its EPG
- Traffic inherits the forwarding and security policy of the EPG



How is “Application Centric” Achieved?

Sources and Destinations Must be Classified into EPG’s

Endpoints

- Used by App EPG’s
- Represents the network identity of an end device
- Learned dynamically or configured statically

Policy-Prefixes

- Used by External EPG’s
- Classifies destination by longest prefix match
- Also used for shared-services
- Configured

PcTags

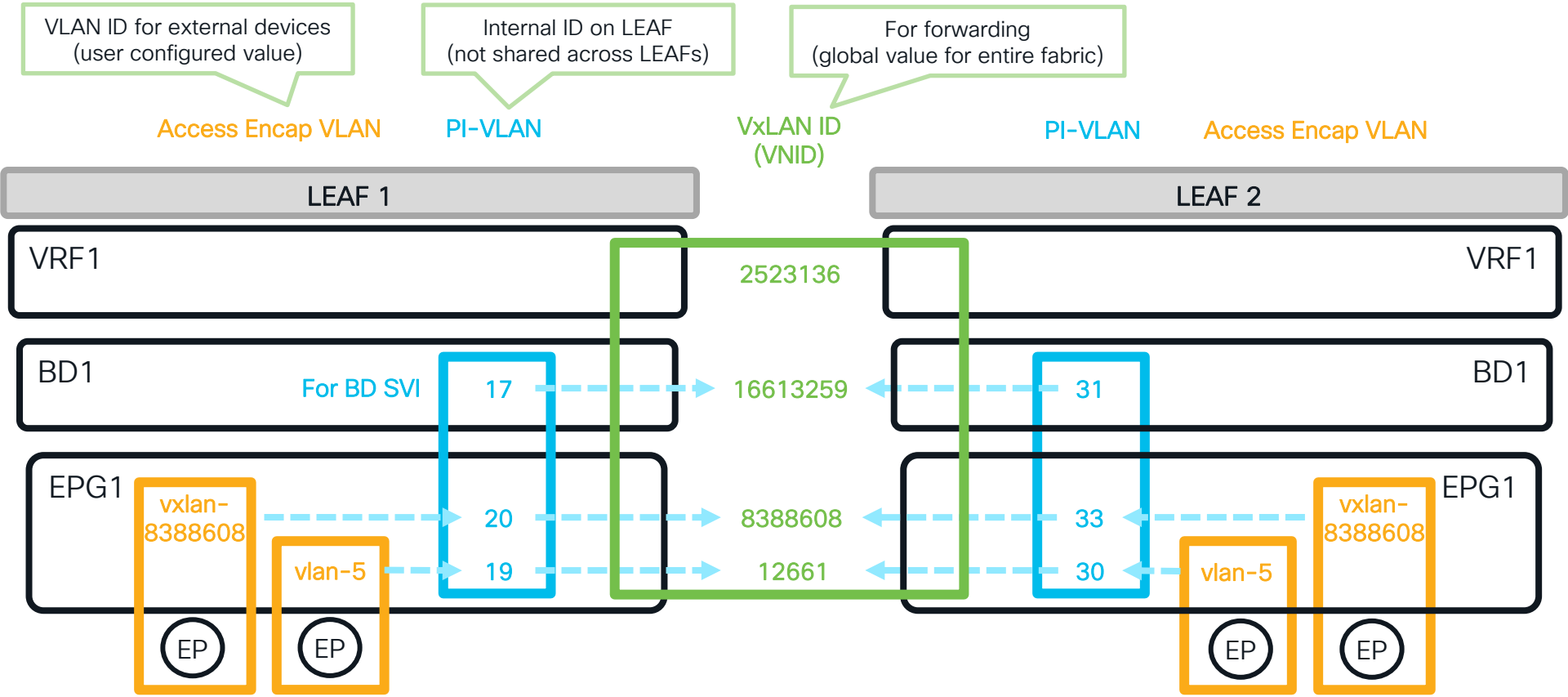
- The security ID of an EPG
- Used in contracts. Ex: Permit PcTag 1000 to PcTag 2000
- Sclass/dclass imply PcTag direction

Contracts

- Defines security and sometimes forwarding (pbr) policy between epgs
- Essentially an ACL between PcTags
- Consumer/Provider rather than src/dest

Vlan Types

※ PI-VLAN : Platform Independent VLAN



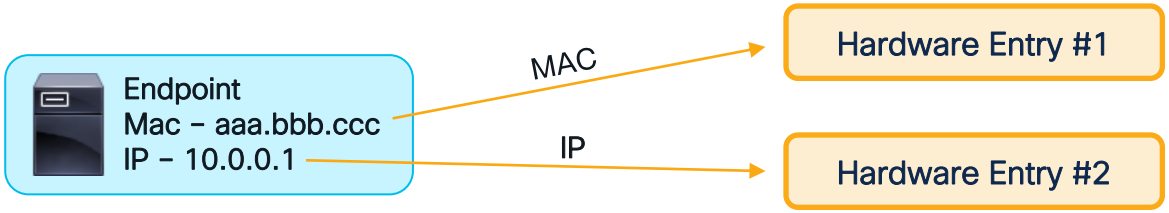
What is an Endpoint?

At the APIC level an Endpoint is a Mac address with zero or more IP/IPv6 Addresses

fvCEp
<epg-dn>/cep-00:00:00:00:0a

fvIp
<epg-dn>/cep-00:00:00:00:0a/ip-[10.0.0.10]

At the Switch level an Endpoint is a Mac address **OR** an IP/IPv6 Address



What is an Endpoint?

An Endpoint joins both forwarding and security policy

Local Learn

```
leaf103# show system internal epm end ip 192.168.200.11

MAC : 0000.1111.2222 ::: Num IPs : 1
IP# 0 : 192.168.200.11 ::: IP# 0 flags :  ::: l3-sw-hit: No
Vlan id : 2 ::: Vlan vnid : 12661 ::: VRF name : CL2022:vrf1
BD vnid : 16613259 ::: VRF vnid : 2523136
Phy If : 0x40018000 ::: Tunnel If : 0
Interface : Ethernet1/25/1
Flags : 0x80005c04 ::: sclass : 32771 ::: Ref count : 5
EP Create Timestamp : 11/01/2021 14:06:25.769904
EP Update Timestamp : 11/04/2021 18:51:54.387104
EP Flags : local|IP|MAC|host-tracked|sclass|timer|
```

Remote Learn

```
leaf103# show system internal epm endpoint ip 192.168.100.10

MAC : 0000.0000.0000 ::: Num IPs : 1
IP# 0 : 192.168.100.10 ::: IP# 0 flags :  ::: l3-sw-hit: No
Vlan id : 0 ::: Vlan vnid : 0 ::: VRF name : CL2022:vrf1
BD vnid : 0 ::: VRF vnid : 2523136
Phy If : 0 ::: Tunnel If : 0x18010001
Interface : Tunnel1
Flags : 0x80004400 ::: sclass : 49154 ::: Ref count : 3
EP Create Timestamp : 11/04/2021 16:38:13.570615
EP Update Timestamp : 11/04/2021 18:51:54.386595
EP Flags : IP|sclass|timer|
```

VNID

Interface/TEP

PcTag

What is a TEP? (Tunnel Endpoint)

- IP addresses allocated for overlay communication
- VXLAN Traffic is sent to the TEP + VNID of destination

Most Common TEP Types

TEP Type	What is it?	What is it for?
Physical TEP (PTEP)	Unique Overlay IP Address for each individual Leaf/Spine	Non-vpc dataplane, l3out communication, apic-leaf comm, etc
VPC TEP (VTEP)	Unique Overlay IP Address for each VPC Pair	Traffic destined to endpoints that are connected behind VPC
Proxy TEP	Spine Anycast IP's used for proxy traffic	Leafs send to these TEPs when doing proxy forwarding

```
a-leaf101# show ip interface loopback0
IP Interface Status for VRF "overlay-1"
lo0, Interface status: protocol-up/link-up/admin-up, iod: 4, mode: ptep
```

What are Tunnels?

- Leafs/Spines Install Tunnel Interface to each known TEP.
- Used for VXLAN Dataplane

How are Tunnels Learned?

Dataplane Learns

```
leaf# moquery -c tunnelIf -f 'tunnel.If.id=="tunnel1"'

id                  : tunnel1
dest                : 10.0.72.67
idRequestorDn       : sys/*/db-dtep/dtep-[10.0.72.67]
```

Through BGP (l3out routes)

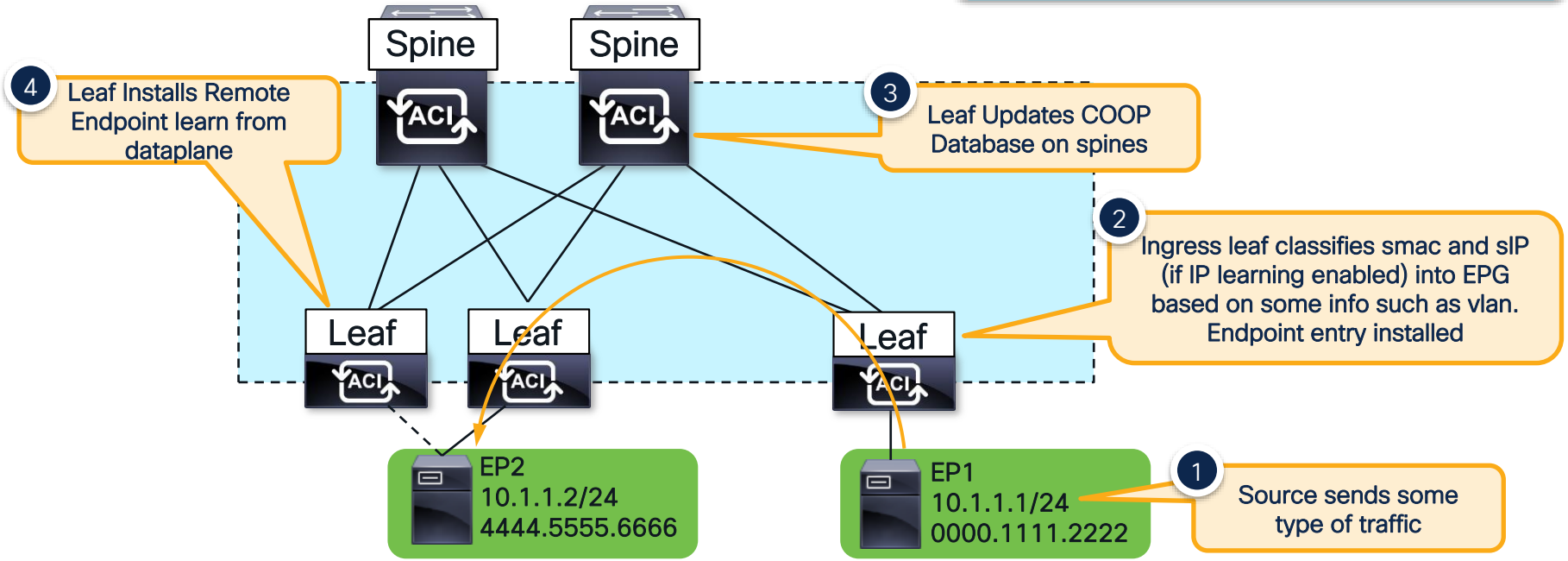
```
leaf# moquery -c tunnelIf -f 'tunnel.If.id=="tunnel1"'

id                  : tunnel1
dest                : 10.0.72.64
idRequestorDn       : sys/bgp/*/db-dtep/dtep-[10.0.72.64]
```

Local POD ISIS Database

```
leaf# moquery -c tunnelIf -f 'tunnel.If.id=="tunnel1"'

# tunnel.If
id                  : tunnel1
dest                : 10.0.152.64
idRequestorDn       : sys/isis/*/lvl-l1/db-dtep/dtep-[10.0.152.64]
```

How is an Endpoint Learned?

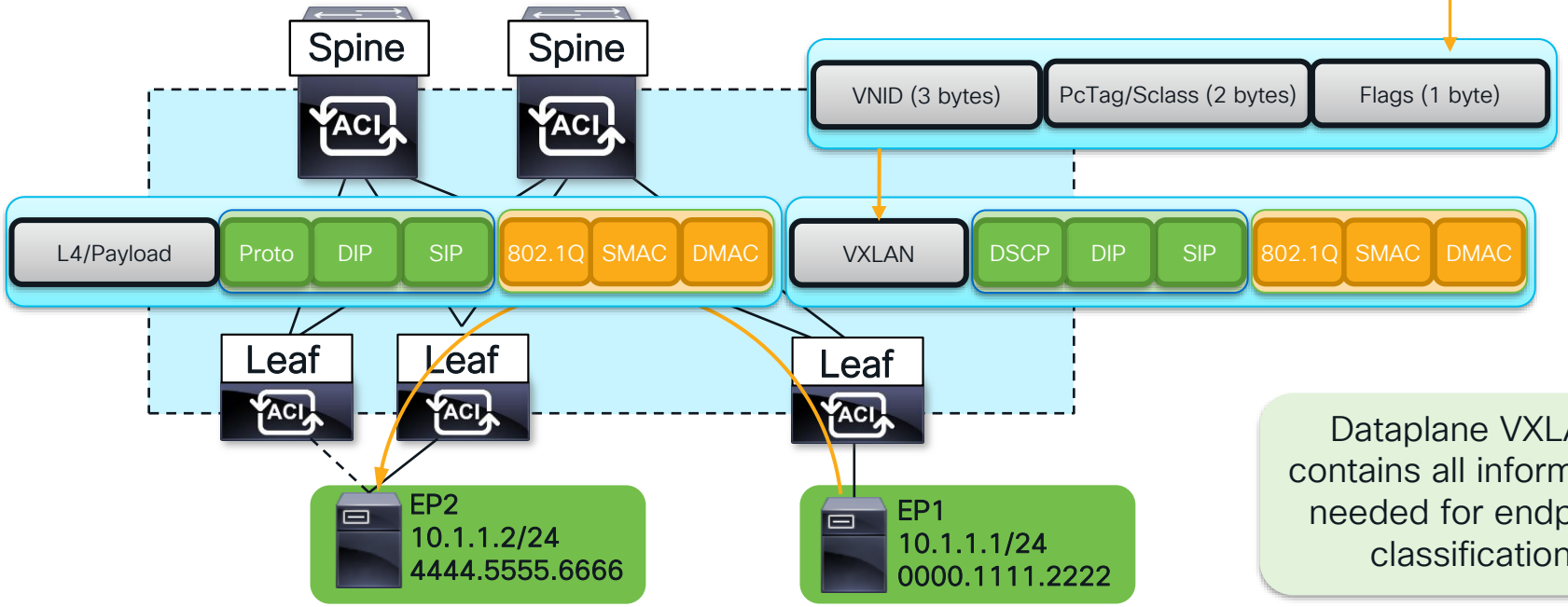
How does the Egress leaf classify traffic into the correct EPG?



Overlay iVXLAN

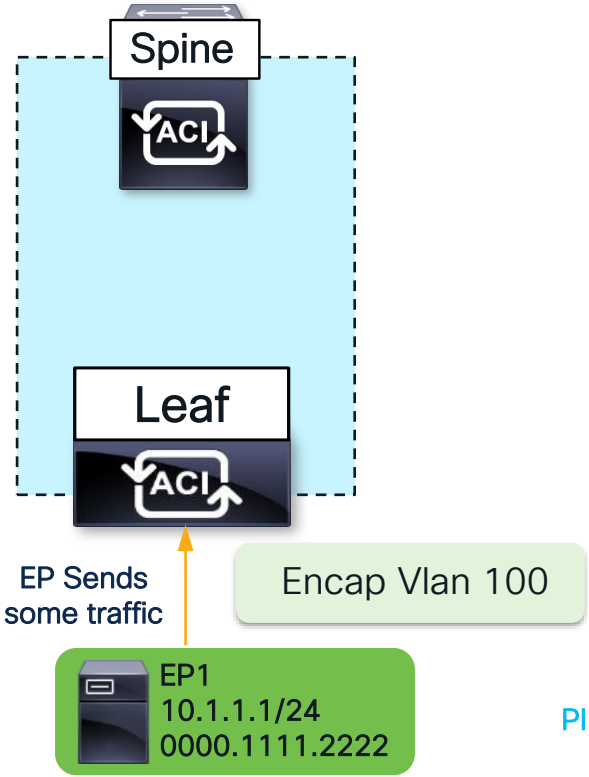
ACI uses VXLAN with some additional bits

Bit pos 4 – Source Policy Applied
Bit pos 5 – Destination Policy Applied
Bit pos 7 – Don't learn



Dataplane VXLAN contains all information needed for endpoint classification

How is an Endpoint Learned?




```
leaf103# show system internal epm vlan 2 detail

VLAN 2
VLAN type : FD vlan
hw id : 34 ::: sclass : 32771
access enc : (802.1Q, 100)
fabric enc : (VXLAN, 12661)
Object store EP db version : 4
BD vlan id : 1 ::: BD vnid : 16613259 ::: VRF vnid : 2523136
Valid : Yes ::: Incomplete : No  ::: Learn Enable : Yes
```

PI-VLAN

```
leaf103# show vlan encap-id 100

VLAN Name                             Status    Ports
---- -------------------------------- --------- -------------------------------
2    CL2022:ap1:epg2                  active    Eth1/25/3
```

Checking Endpoints

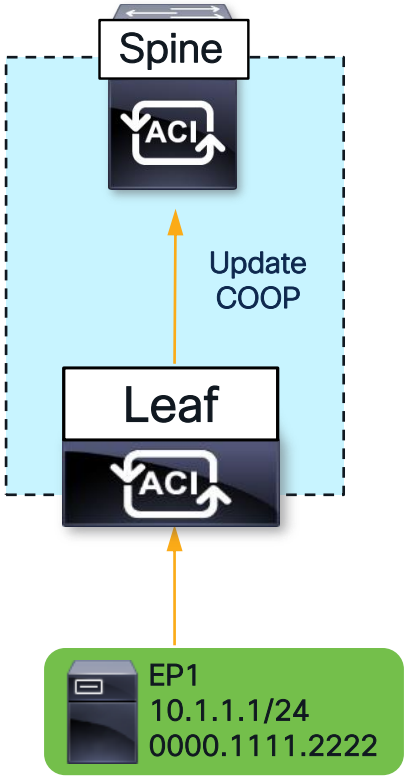
Reference commands can be run from leafs or apics

```
#Check object model for Mac Address Endpoint
moquery -c epmMacEp -f 'epm.MacEp.addr=="00:00:AA:AA:BB:BB"'

#Check object model for IP Address Endpoint
moquery -c epmIpEp -f 'epm.IpEp.addr=="192.168.200.11"'
```

Reference commands can be run from leafs only

```
#Check endpoint manager process directly
show system internal epm endpoint mac 0000.aaaa.bbbb
show system internal epm endpoint ip 192.168.200.11

#Check hardware level endpoint process directly
vsh_lc -c "show system internal epmc endpoint mac 0000.aaaa.bbbb"
vsh_lc -c "show system internal epmc endpoint ip 192.168.200.11"
```

How is an Endpoint Learned?



The Leaf Updates COOP on Spines

```
spine1005# show coop internal info ip-db | grep -B 1 -A 15
192.168.200.11
--------------------------------
IP address : 192.168.200.11
Vrf : 2523136
Flags : 0
EP bd vnid : 16613259
EP mac :  00:00:AA:AA:BB:BB
Publisher Id : 10.0.64.70
Record timestamp : 11 05 2021 17:02:56 217794556
Publish timestamp : 11 05 2021 17:02:56 220584642
Seq No: 0
Remote publish timestamp: 01 01 1970 00:00:00 0
URIB Tunnel Info
Num tunnels : 1
        Tunnel address : 10.0.64.70
        Tunnel ref count : 1
```

VNID info should match the info on leaf

Leaf TEP that owns this EP:
#From APIC
moquery -c ipv4Addr -f 'ipv4.Addr.addr=="10.0.64.70"'

Checking COOP

Reference commands can be run from spines or apics

Query COOP for l2 entry:
```
moquery -c coopEpRec -f 'coop.EpRec.mac=="00:00:AA:AA:BB:BB"'
```

Query COOP for l3 entry and get parent l2 entry:
```
moquery -c coopEpRec -x rsp-subtree=children 'rsp-subtree-filter=eq(coopIpv4Rec.addr,"1.1.1.1")' rsp-subtree-include=required
```

Query COOP for l3 only entry (such as an SVI IP):
```
moquery -c coopIpOnlyRec -f 'coop.IpOnlyRec.addr=="192.168.100.10"'
```

Query COOP for l3 ep:
```
moquery -c coopIpv4Rec -f 'coop.Ipv4Rec.addr=="192.168.100.10"'
```

How is Traffic Classified with no EP Learn?

In most of these cases, the pcTag is based on a policy-prefix lookup

There will be no endpoint learn in several cases

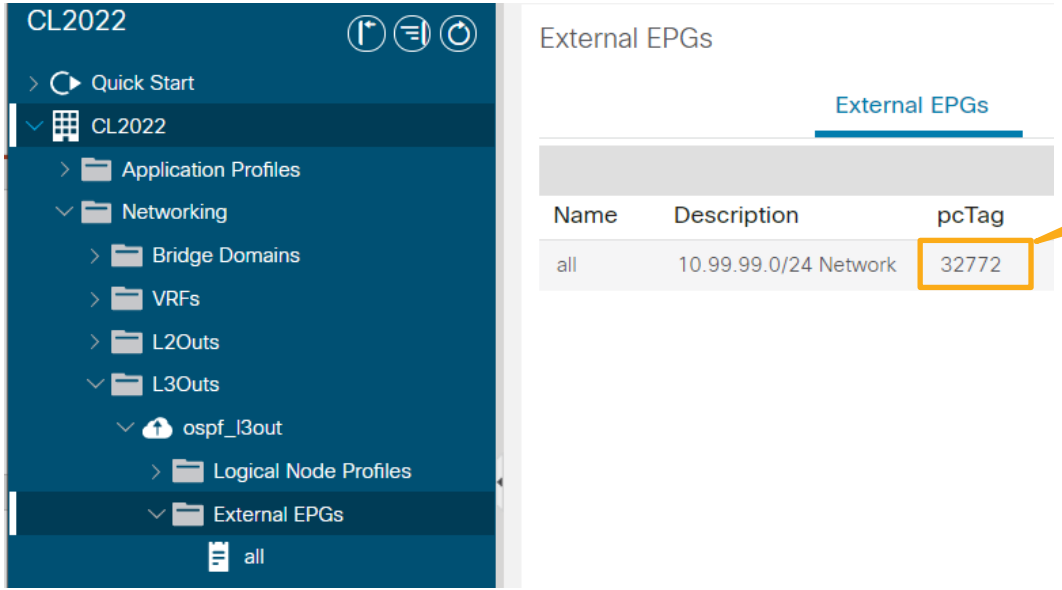
- Source/dest is behind an l3out
- Source/dest is in another vrf
- Endpoint learning is disabled by some option

If ingress leaf doesn't apply policy, egress leaf should (indicated via policy-applied bits in ivxlan header)

How is Traffic Classified with no EP Learn?

Destination Behind L3out

```
leaf101# vsh_lc -c "show forwarding route 10.99.99.100 platform vrf CL2022:vrf1"
!
Policy Prefix 10.99.99.0/24
!
vrf: 16(0x10), routed_if: 0x0 epc_class: 32772(0x8004)
```



Classification based on longest l3out policy prefix

How is Traffic Classified with no EP Learn?

Destination is unknown and is proxied

```
leaf101# show ip route 192.168.200.20 vrf CL2022:vrf1

192.168.200.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.176.66%overlay-1, [1/0], 4d05h, static, tag 4294967294
        recursive next hop: 10.0.176.66/32%overlay-1

leaf101# vsh_lc -c "show forwarding route 192.168.200.20 platform vrf CL2022:vrf1"
!
Policy Prefix 0.0.0.0/0
!
vrf: 16(0x10), routed_if: 0x0 epc_class: 1(0x1)
```

"Pervasive" indicates this is a BD or EPG subnet (fvSubnet). Send to spine proxy-addr

-pcTag of 1 indicates the fabric owns the subnet, don't apply policy
-policy applied flags not set in ivxlan header

Don't apply policy, Forward to proxy Anycast!

```
leaf101# show isis dtep vrf overlay-1 | egrep "Type|PROXY"
DTEP-Address      Role    Encapsulation   Type
10.0.176.66       SPINE   N/A             PHYSICAL,PROXY-ACAST-V4
10.0.176.65       SPINE   N/A             PHYSICAL,PROXY-ACAST-MAC
10.0.176.64       SPINE   N/A             PHYSICAL,PROXY-ACAST-V6
```

How is Traffic Classified with no EP Learn?

Shared Services Classification

Destination is in shared services **provider** EPG (different vrf)

```
leaf# show ip route 192.168.255.10 vrf CL2022:vrf1
192.168.255.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.176.66%overlay-1, [1/0], static, tag !!!, rwVnid: vxlan-2457601
        recursive next hop: 10.0.176.66/32%overlay-1

leaf# vsh_lc -c "show forwarding route 192.168.255.10 plat vrf CL2022:vrf1"
Prefix:192.168.255.0/24, Update_time:Fri Nov  5 20:57:00 2021
!
Policy Prefix 0.0.0.0/0
!
Flags: IN-HW, SHRD-SVC,
vrf: 16(0x10), routed_if: 0x0 epc_class: 36(0x24)
```

PcTag of provider epg

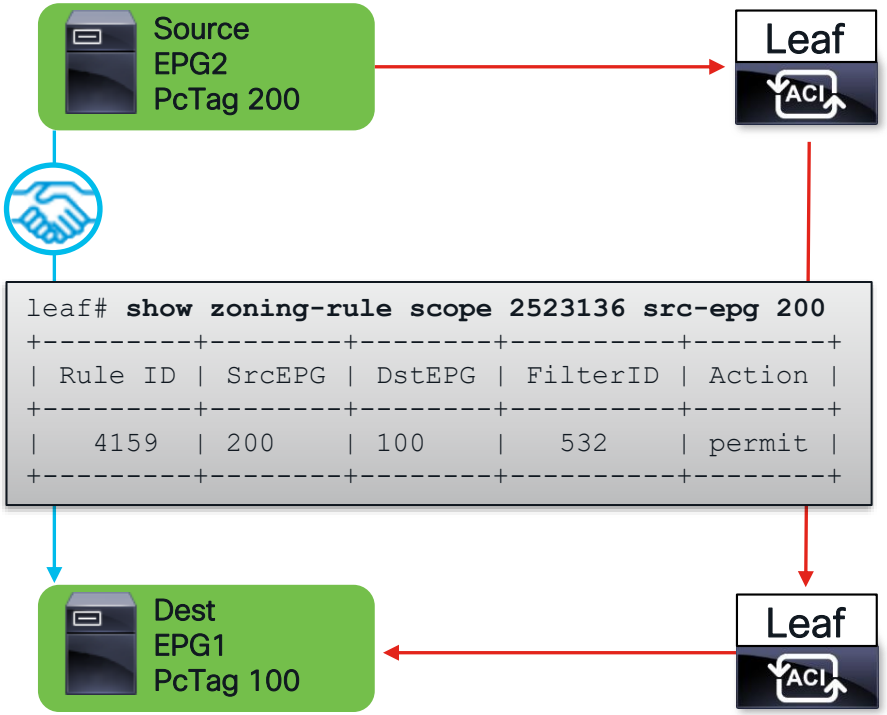
Destination is in shared services **consumer** EPG (different vrf)

```
leaf# show ip route 192.168.100.10 vrf CL2022:vrf2
192.168.100.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.176.66%overlay-1, [1/0], static, rwVnid: vxlan-2523136
        recursive next hop: 10.0.176.66/32%overlay-1

leaf# vsh_lc -c "show forwarding route 192.168.100.10 plat vrf CL2022:vrf2"
Prefix:192.168.100.0/24, Update_time:Tue Nov  9 14:34:05 2021
!
Policy Prefix 0.0.0.0/0
!
Flags: IN-HW, SHRD-SVC,
vrf: 10(0xa), routed_if: 0x0 epc_class: 14(0xe)
```

Reserved tag for shared services consumer. Policy applied in consumer vrf

Check hidden slide for impact of “Policy Control Enforcement Direction” setting

Contracts and Forwarding




```
leaf# show zoning-rule scope 2523136 src-epg 200
+---------+--------+--------+----------+--------+
| Rule ID | SrcEPG | DstEPG | FilterID | Action |
+---------+--------+--------+----------+--------+
|   4159  | 200    | 100    |   532    | permit |
+---------+--------+--------+----------+--------+
```

Ingress

Contract Found?

- Yes → Set policy-applied bits in ivxlan. Permit, deny, redir, log
- No → If LPM is BD/EPG subnet, forward and don’t set policy-applied bits in ivxlan. Otherwise, drop!

Egress

Policy-Applied Bits set?

- Yes → Don’t do contract lookup. Forward.
- No → Do contract lookup. Permit, deny, redir, log

Policy enforcement table

Where is policy enforced?

VRF Enforcement Setting

Flow Direction	INGRESS	EGRESS
EPG to unknown EPG	Applied Egress	Unchanged
EPG to known EPG	Applied Ingress	Unchanged
EPG to L3out	Applied Ingress/non-BL	Applied Egress/BL
L3out to unknown EPG	Applied Egress/non-BL	Applied Egress
L3out to known EPG	Applied Egress/non-BL	Applied Ingress/BL
L3out to L3out	Applied Ingress	Applied Egress

Policy enforcement affects only traffic to or from the L3Out. There are no behavior changes in EPG-to-EPG.

What About Flooded Traffic?

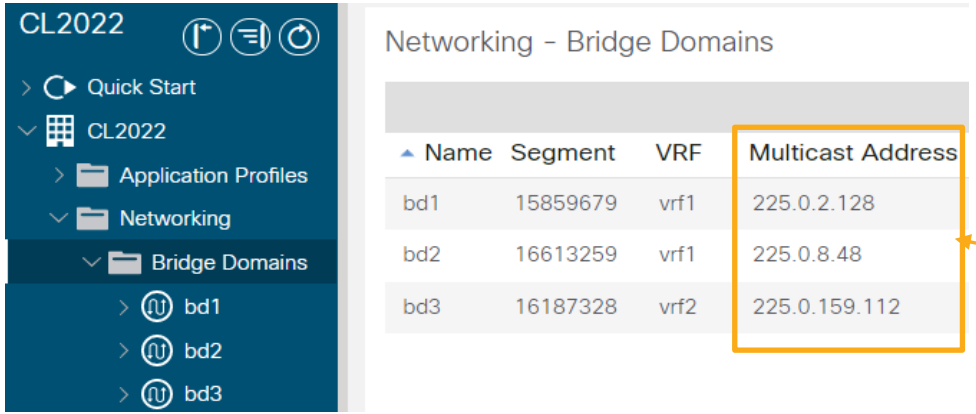
The following traffic may be flooded:

- Broadcast
- Multicast
- Unknown Unicast
- Control Plane maintenance (EP announce, fabric ARP, etc)

How does ACI flood?

- Flooded traffic is sent to the BD GiPo (l2 flood) or VRF GiPo (l3 flood)
- The GiPo is an overlay multicast address allocated to a BD or VRF
- Flooding is done on a loop-free tree called an FTAG

Security policy NOT applied

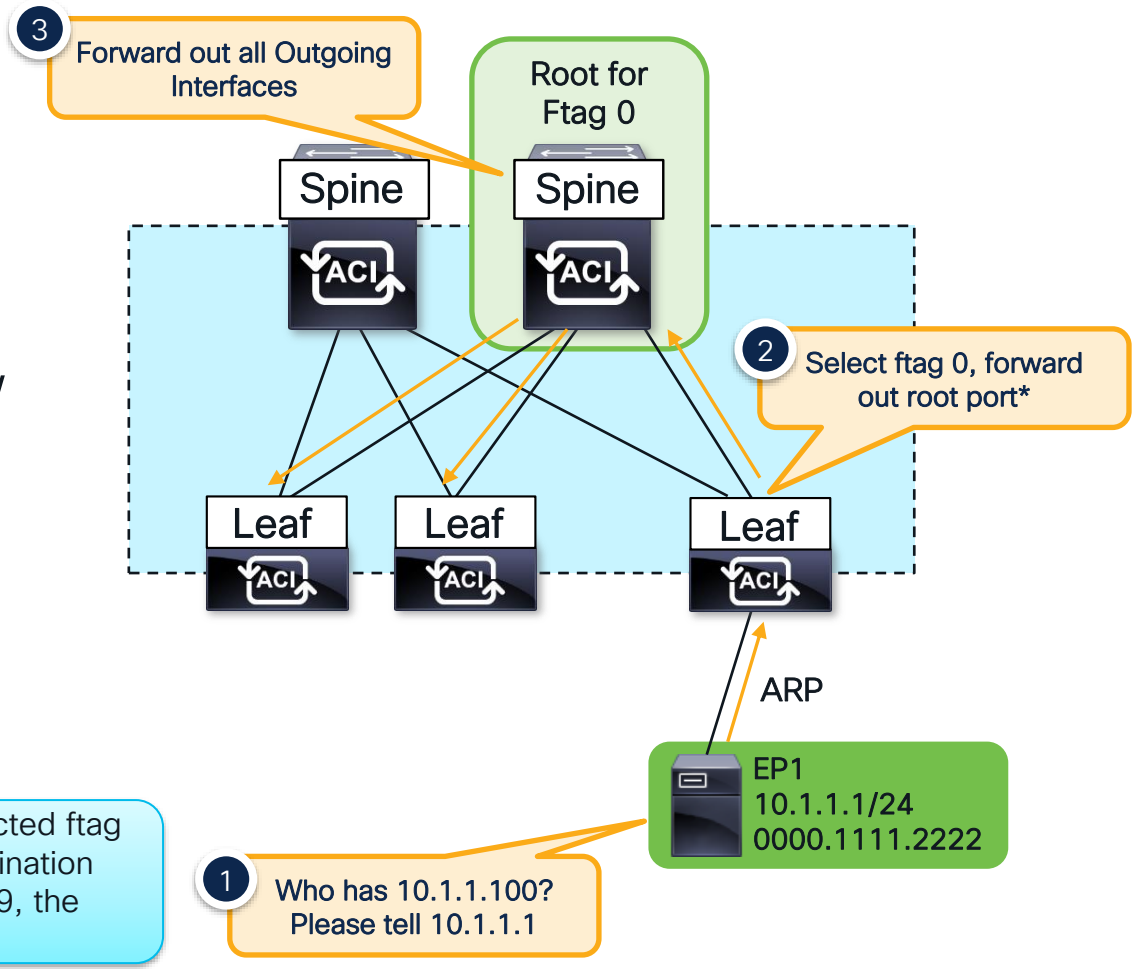


Name	Segment	VRF	Multicast Address
bd1	15859679	vrf1	225.0.2.128
bd2	16613259	vrf1	225.0.8.48
bd3	16187328	vrf2	225.0.159.112

GiPo

What are FTAGs?

- FTAGs are loop-free trees within the overlay used by flooded traffic
- FTAGs are picked per flow from values 0 – 0xc
- One spine is root for each tree
- Outgoing interfaces calculated by ISIS



*Note, the ingress leaf communicates the selected ftag to the rest of the fabric by adding it to the destination gipo. If the gipo is 225.0.0.0 and the ftag is 0x9, the destination address would be 225.0.0.9

Checking FTAGs

Find the outgoing interfaces for a tree

Check FTAG tree on ingress leaf

```
leaf101# show isis internal mcast routes ftag

FTAG Routes
====================================

 FTAG ID:   0  [Enabled] Cost:(   1/   7/   0)
 ---------------------------------
    Root port: Ethernet1/54.6
    OIF List:
      Ethernet1/53.5
!
!ommitted rest of ftags
```

Leaf forwards to root port and any additional OIFs

Check FTAG tree on root spine

```
spine1005# show isis internal mcast routes ftag

FTAG Routes
================================================

 FTAG ID:   0 [Root] [Enabled] Cost:(   0/   0/   0)
 ---------------------------------
    Root port: -
    OIF List:
      Ethernet1/1.20
      Ethernet1/2.21
      Ethernet1/3.19
!ommitted rest of ftags
```

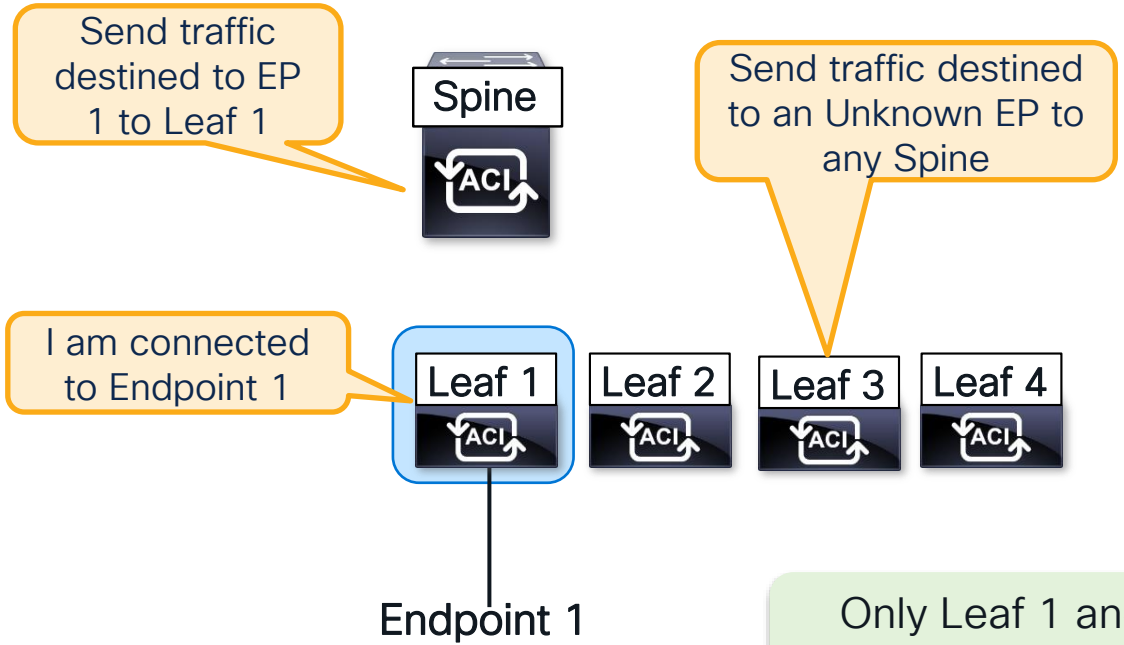
This spine is the root for ftag 0

Forward out all of these interfaces

Proxy Forwarding

What is Proxy Forwarding?

Why? Scaling out Endpoint Learning



Spines own separate anycast TEP's for mac, ipv4, and ipv6 proxy lookups

Only Leaf 1 and Spines have to program Endpoint 1 in hardware

How to check the Spine-Proxy TEP

```
leaf1# show ip route vrf CL2022:vrf1

192.168.0.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.16.64%overlay-1, [1/0], 00:21:39, static
```

BD Subnet (Pervasive Route)

next-hop should be SPINE-PROXY

```
leaf1# show isis dteps vrf overlay-1 | grep PROXY
10.0.16.65          SPINE    N/A              PHYSICAL,PROXY-ACAST-MAC
10.0.16.64          SPINE    N/A              PHYSICAL,PROXY-ACAST-V4
10.0.16.67          SPINE    N/A              PHYSICAL,PROXY-ACAST-V6
```

next-hop of Pervasive Route is IPv4 Spine Proxy TEP

Three types of Spine Proxy TEP

- Proxy-Acast-MAC
  - ✓ Spine-Proxy for L2 traffic (L2 Unknown Unicast mode “Hardware Proxy”)
- Proxy-Acast-V4
  - ✓ Spine-Proxy for IPv4 traffic (includes ARP Request with ARP Flooding mode “OFF”)
- Proxy-Acast-V6
  - ✓ Spine-Proxy for IPv6 traffic

What is COOP?

COOP is the proxy-database of ACI

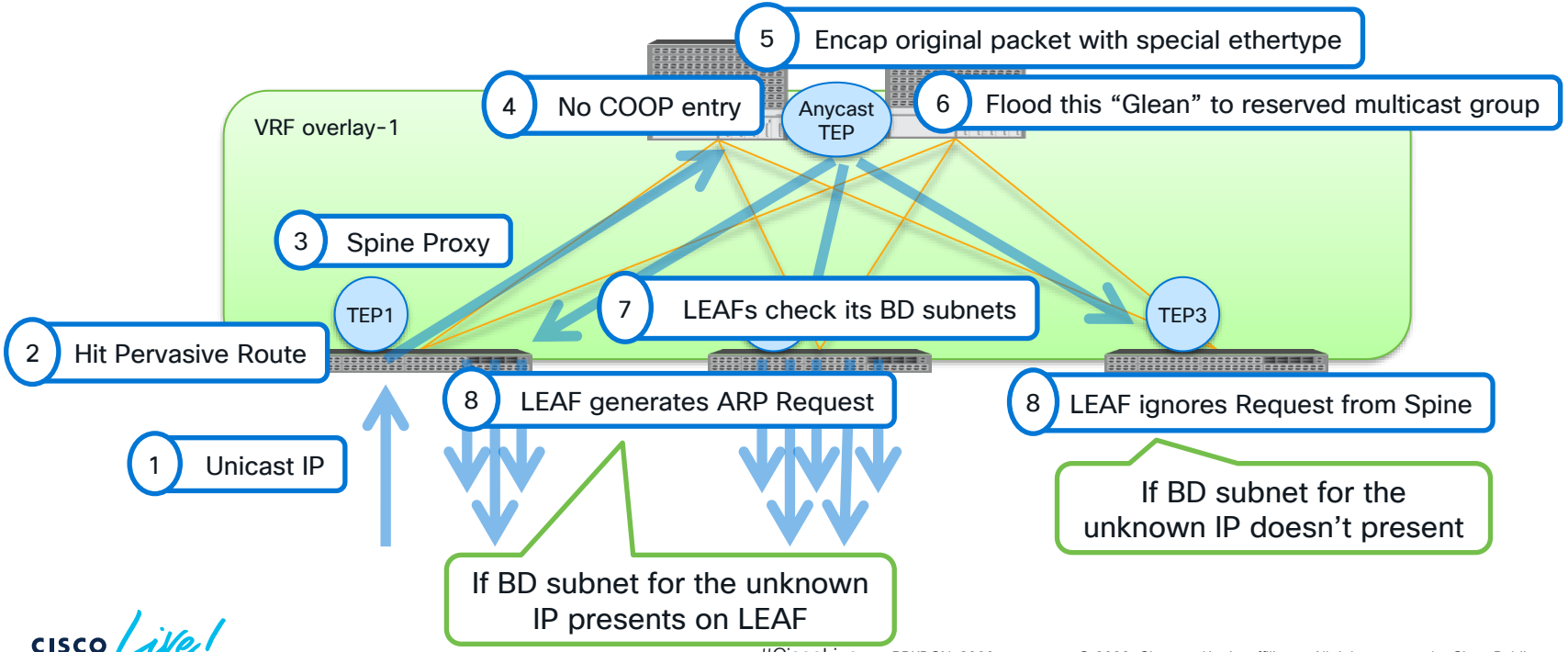
- Council of Oracles Protocol – A TCP protocol for citizens (Leafs) to publish records to oracles (Spines).
- Used for announcing endpoints, fabric owned IP's, multicast information, and more
- Synced across Pods/Sites with BGP EVPN
- Each Endpoint Record contains all information to forward (VNID, leaf TEP, mac, etc)
- COOP records pushed into hardware on spines
- For modular spines, scale is achieved by pushing each EP onto only two Fabric Modules

What if the Endpoint isn't in COOP? (ARP Glean)

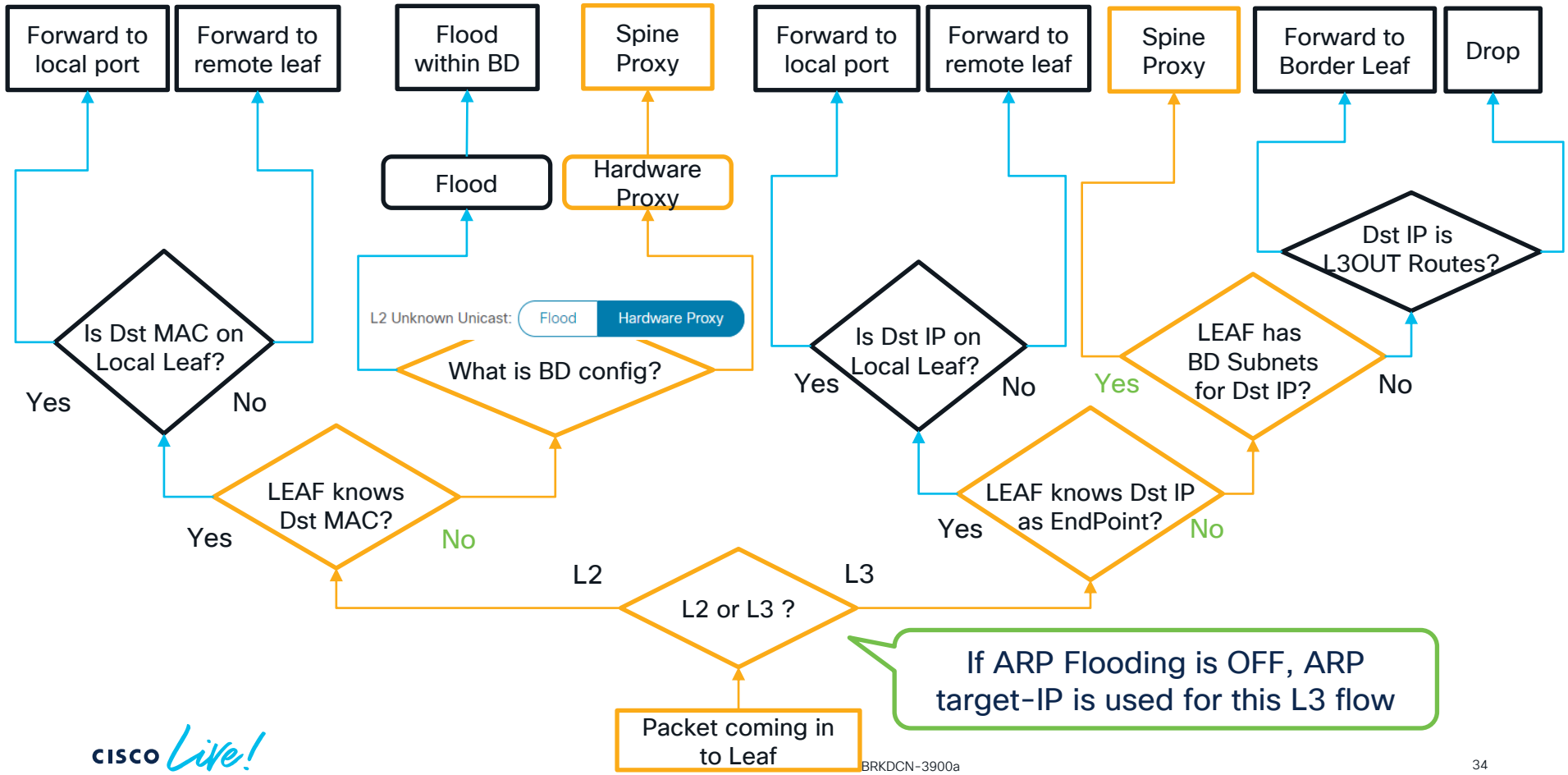
What if Spine's COOP DB doesn't know the destination when proxy'ed?

✗ L2 Traffic : Drop

✓ L3 Traffic : ARP Glean



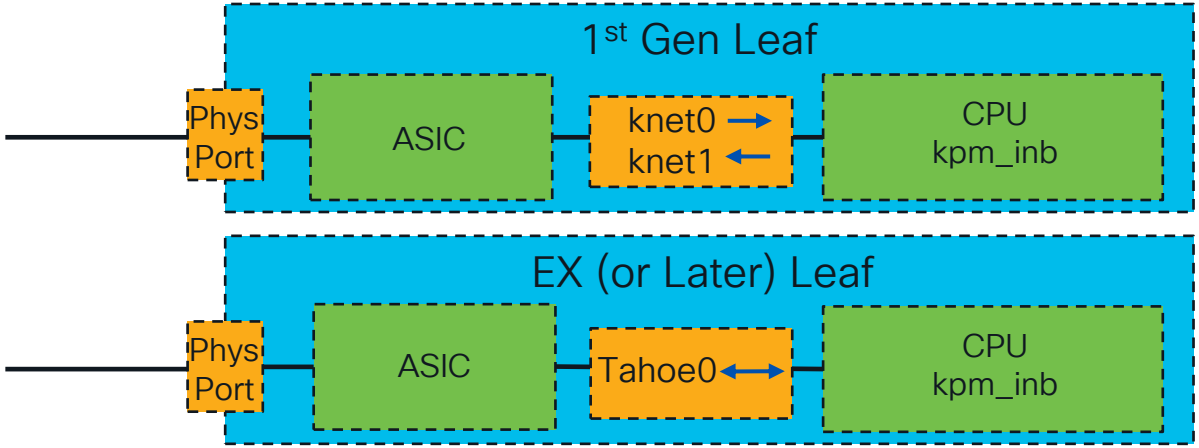
Spine Proxy Summary



Capturing a Glean with Tcpdump

ACI Leafs and Spines contain pseudo interfaces for traffic to and from the CPU

- Traffic on the on the knet or tahoe pseudo interface will have a special ieth header. It must be decoded.
- Starting in 3.2 the *knet_parser.py* script is available on the switch cli to decode



- For traffic going to the cpu check knet0 and kpm_inb
- For traffic coming from the cpu check knet1 and kpm_inb
- For traffic to and from the cpu check Tahoe0 and kpm_inb

*Note, not all traffic will show up on the kpm_inb interface. However, all traffic shows on the pseudo interface
*Gen1 and 2 Modular spines use psdev0, psdev1, and psdev2 interfaces. Gen 2 fixed spines use tahoe0. Gen 1 fixed spines use knet0-3

Capturing a Glean with Tcpdump

Gen2 or Later Leaf

Egress Leaf Verification

```
tcpdump -xxxvei tahoe0 -w /bootflash/tahoe0.pcap
knet_parser.py --file /bootflash/tahoe0.pcap --pcap --decoder tahoe

Frame 111
  Time: 2019-05-16T16:56:33.059831+00
  Header: ieth_extn CPU Receive
        sup_qnum:0x14, sup_code:0x21, istack:ISTACK_SUP_CODE_SPINE_GLEAN(0x21)
  Header: ieth
        sup_tx:0, ttl_bypass:0, opcode:0x6, bd:0x120e, outer_bd:0x27, dl:0, span:0, traceroute:0, tclass:0
        src_idx:0x3a, src_chip:0x0, src_port:0x19, src_is_tunnel:1, src_is_peer:1
        dst_idx:0x0, dst_chip:0x0, dst_port:0x0, dst_is_tunnel:0
  Len: 148
  Eth: 000d.0d0d.0d0d > 0100.5e7f.fff1, len/ethertype:0x8100(802.1q)
  802.1q: vlan:2, cos:5, len/ethertype:0x800(ipv4)
  ipv4: 10.0.116.64 > 239.255.255.241, len:130, ttl:249, id:0x0, df:0, mf:0, offset:0x0, dscp:32, prot:17(udp)
  udp: (ivxlan) 0 > 48879, len:110
  ivxlan: n:1, l:1, i:1,
        vnid: 0x2b0000
        lb:0, dl:1, exception:0, src_policy:0, dst_policy:0, src_class:0  c0
        mcast(routed:0, ingress_encap:0/802.1q), ac_bank:0, src_port:0x0
  Eth: 000c.0c0c.0c0c > ffff.ffff.ffff, len/ethertype:0xfff2(aci-glean)
  ipv4: 172.16.1.1 > 172.16.2.2, len:84, ttl:63, id:0x71f9, df:1, mf:0, offset:0x0, dscp:0, prot:1(icmp)
  icmp: echo request id:0x9092, seq:0x1980
```

Decode type should be tahoe for tahoe interface

RX sup traffic rather than TX

Switch recognizes this as a Glean

Traffic that triggered Glean

Capturing a Glean with Tcpdump

Egress Leaf Verification

Gen1 Leaf Example

knet0 would show Rx traffic (similar output as Tahoe0)

```
tcpdump -xxxvei knet0 -w /bootflash/knet0.pcap
knet_parser.py --file /bootflash/knet0.pcap --pcap --decoder knet
```

knet1 would show Tx traffic

```
tcpdump -xxxvei knet1 -w /bootflash/knet1.pcap
knet_parser.py --file /bootflash/knet1.pcap --pcap --decoder knet
```

No decode necessary for kpm_inb (cpu) interface…Gleans aren't easily readable

```
tcpdump -xxxvei kpm_inb ether proto 0xfff2
a-leaf102# tcpdump -xxxvei kpm_inb ether proto 0xfff2
tcpdump: listening on kpm_inb, link-type EN10MB (Ethernet), capture size 65535 bytes
15:27:37.663580 00:0c:0c:0c:0c:0c (oui Unknown) > Broadcast, ethertype Unknown (0xfff2), length 94:
        0x0000:  ffff ffff ffff 000c 0c0c 0c0c fff2 4500
        0x0010:  0054 aa4b 4000 3f01 825d 0404 0464 0303
        0x0020:  0396 0800 0dc6 2384 38db 5275 dd5c 0000
        0x0030:  0000 9e35 0100 0000 0000 1011 1213 1415
        0x0040:  1617 1819 1a1b 1c1d 1e1f 2021 2223 2425
        0x0050:  2627 2829 2a2b 2c2d 2e2f 3031 3233
```

Layer 3 Unicast – Glean Scenario

Verify ARP on Remote Leaf

Egress Leaf Verification

```
a-leaf205#show ip arp internal event-history event | grep -F -B 1 172.16.2.2
73) Event:E_DEBUG_DSF, length:127, at 316928 usecs after Wed May  1 08:31:53 2019
Updating epm ifidx: 1a01e000 vlan: 105 ip: 172.16.2.2, ifMode: 128 mac: 0000.1111.2222
75) Event:E_DEBUG_DSF, length:152, at 316420 usecs after Wed May  1 08:31:53 2019
log_collect_arp_pkt; sip = 172.16.2.2; dip = 172.16.2.254; interface = Vlan104;info = Garp Check adj:(nil)
77) Event:E_DEBUG_DSF, length:142, at 131918 usecs after Wed May  1 08:28:36 2019
log_collect_arp_pkt; dip = 172.16.2.2; interface = Vlan104;iod = 138; Info = Internal Request Done
78) Event:E_DEBUG_DSF, length:136, at 131757 usecs after Wed May  1 08:28:36 2019
log_collect_arp_glean;dip = 172.16.2.2;interface = Vlan104;info = Received pkt Fabric-Glean: 1
79) Event:E_DEBUG_DSF, length:174, at 131748 usecs after Wed May  1 08:28:36 2019
log_collect_arp_glean; dip = 172.16.2.2; interface = Vlan104; vrf = CiscoLive2020:vrf1; info = Address in PSVI subnet or special VIP
```

Endpoint Learn Installed

Response Received

ARP Request is generated by leaf

Glean Received, Dst IP is in BD Subnet

How ACI Builds Forwarding Tables

Building Adjacency Tables

ACI combines ARP and MAC Tables into the Endpoint Table

Legacy Behavior

- ARP/ND tables map Layer 3 to Layer 2
- ARP/ND tables are updated by control-plane messages
- MAC Address Table used for switching decisions
- Mac Address Table updated by dataplane

ACI Behavior

- Endpoint table contains endpoints, which are Layer 2 addresses OR Layer 3 addresses OR a combination of Layer 2 and Layer 3 addresses
- By default, both Layer 2 and Layer 3 information is updated by dataplane
- Used for security and forwarding policy

Building Endpoint Tables

Endpoints can be programmed via software process or by hardware dataplane learns (HAL)

Resource	Table Info	Commands to Verify
Supervisor	EPM – Endpoint Manager Sup process for managing endpoints.	show system internal epm endpoint mac <addr> show system internal epm endpoint ip <addr>
Line Card	EPMC – Endpoint Manager Client Line card process that sits between hardware layer (HAL) and EPM	vsh_lc –c “show system internal epmc endpoint mac <addr>” vsh_lc –c “show system internal epmc endpoint ip <addr>”
Asic	HAL – Hardware Abstraction Layer View of what is programmed into the ASIC.	vsh_lc –c "show plat internal hal ep l2 mac <addr>" vsh_lc –c "show plat internal hal ep l3 ip <ip/pfx len>“ ! !L3 Endpoints are put into HW Routing Table vsh_lc –c "show plat internal hal l3 routes | grep EP"

What about ARP?

ARP Tables are still used in ACI for…

- L3outs
- Overlay adjacencies
  - VXLAN Endpoints (AVE, K8s, Openstack, etc)
  - APIC / Fabric node adjacencies

Resource	Table Info	Commands to Verify
Supervisor	Adjacency Manager. Programmed by ARP process.	show ip arp vrf <name>
Line Card	UFIB	vsh_lc –c “show forwarding adjacency <ip>”
Asic	HAL – Hardware Abstraction Layer View of what is programmed into the ASIC.	vsh_lc –c "show plat internal hal l3 routes”

Building Routing Tables

Resource	Table Info	Commands to Verify
Supervisor	URIB / MRIB – the unicast and multicast routing tables. Programmed by route protocol	show ip route x.x.x.x/y vrf <name> show ip mroute x.x.x.x/y vrf <name>
Line Card	UFIB / MFIB – the unicast and multicast forwarding tables on the Line Card	vsh_lc -c "show forwarding route <ip/pfx len> vrf <name>“ vsh_lc -c "show forwarding multicast route vrf <name>"
Asic	HAL – Hardware Abstraction Layer View of what is programmed into the ASIC.	vsh_lc -c "show platform internal hal l3 routes vrf <name>” vsh_lc -c "show platform internal hal l3 mcast routes vrf <name>” vsh_lc -c "show plat internal hal l3 routes vrf <name>" | grep MC

Troubleshooting TIP

Check Endpoint Table before Routing Table

When Troubleshooting Layer 3 Flows Always…

1) Check if there is an Endpoint Learn

```
show endpoint ip <addr>
show system internal epm endpoint ip <addr>
```

If not then…

2) Check if there is a BD (pervasive) static route

If not then…

3) Check if there is an External Route

```
show ip route x.x.x.x/y vrf <name>
```

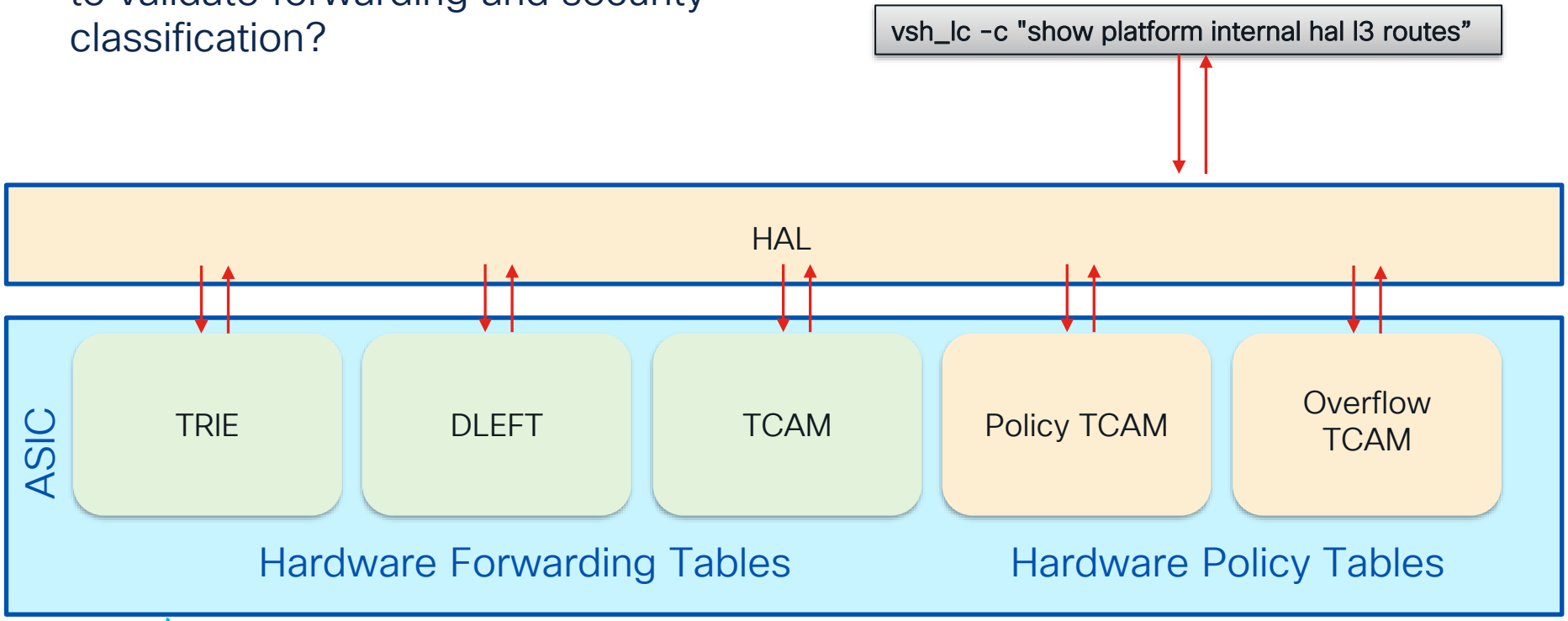
Programming Contracts

Resource	Table Info	Commands to Verify
Supervisor	Policy Manager. Programmed by leaf policy-element process	show zoning-rules
Line Card	ACLQOS	vsh_lc -c “show system internal aclqos zoning-rules”
Asic	HAL – Hardware Abstraction Layer View of what is programmed into the ASIC.	vsh_lc -c “show plat internal hal objects policy zoningrule”

HAL – Hardware Abstraction Layer

Applicable to EX and Later Hardware

Wouldn't it be great if there was a single point to validate forwarding and security classification?



HAL – Hardware Abstraction Layer

Applicable to EX and Later Hardware

L3 Lookup of Hardware Tables

```
module-1# show plat internal hal l3 routes vrf CL2022:vrf1
-----------------------------------------!!-------------------------
|     |                       |    | LID |!!                        |
| VRF | Prefix/Len            | RT| Type|!!!|CLSS| Flags             |
|-----|-----------------------|---|-----|!!!|----|------------------|
|-----|-----------------------|---|-----|!!!|----|------------------|
| 4626|    192.168.100.10/ 32| EP  TRIE|!! c002| le,bne,sne, dl    |
| 4626|        10.99.99.0/ 24| UC  TCAM|!! 8004| sc,spi,dpi        |
| 4626|     192.168.255.0/ 24| UC  TCAM|!!   24| sc,spi,dpi, dr    |
| 4626|    192.168.200.11/ 32| EP  TRIE|!! 8003| sc, le,sne        |
 -----------------------------------------!!-------------------------|
```

Much more info available in full output!

Consolidated view of routes for Endpoints, Shared Services, and External routes

PcTag from destination EPG...used for contract lookup

HAL – Hardware Abstraction Layer

L2 Lookup of Hardware Tables

Applicable to EX and Later Hardware

```
module-1# show platform internal hal ep l2 all
==============================================================

     BD            EP                  L2        L2          S
BdId Name       T  Mac                 IfId      Ifname      Class
==============================================================
b    BD-11      Pl 00:00:11:11:22:22 1a010000 Eth1/17     c003
1a   BD-26      Xr 00:00:22:22:33:33 18010004 Tunnel4     400f
21   BD-33      Pl 00:00:22:22:33:33 16000002 Po3         4002
```

Much more info available in full output!

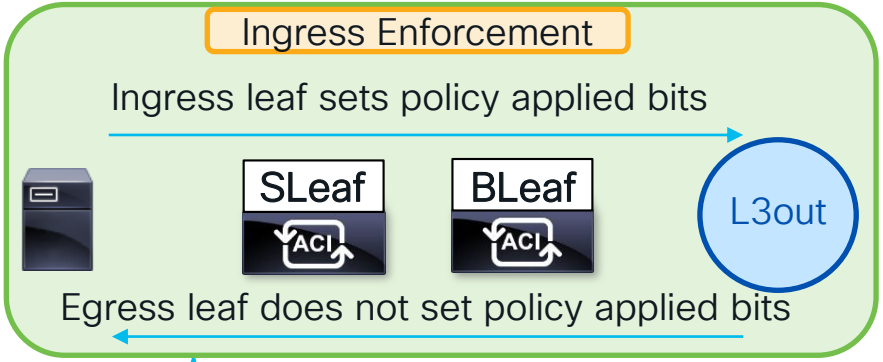
Consolidated view of all learned Mac Addresses

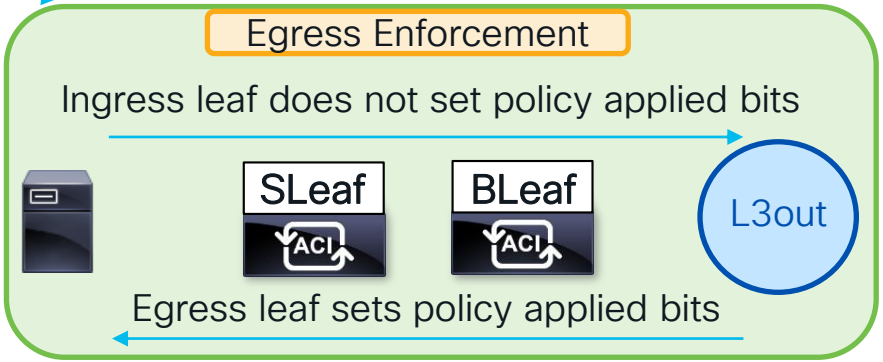
PcTag from destination EPG...used for contract lookup

Understanding the Configuration Options

VRF Level Forwarding Options

Feature	What Does it Do?
Policy Control Enforcement Preference	If disabled, policy is never applied between EPGs. If enabled, contracts are enforced.
IP Dataplane Learning	If Disabled, ACI uses legacy behavior for learning endpoints. Layer 3 endpoints are learned by ARP/GARP/ND and Layer 2 endpoints are learned by dataplane.
Policy Control Enforcement Direction	If set to Ingress, contract enforcement for l3out flows is done on service leaf. Egress enables enforcement on Border Leaf (requires remote learning to be enabled)





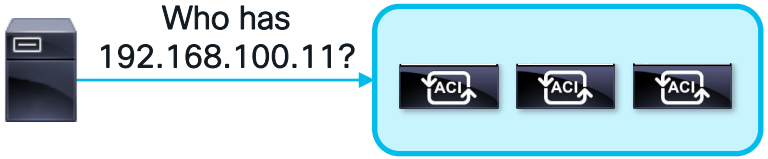
Bridge-Domain Level Forwarding Options

Feature	What Does it Do?
L3 Unknown Multicast Flooding	For non-link-local L3 multicast traffic in a PIM-disabled BD, should a leaf with no snooping entries flood in BD (flood) or wait for joins (OMF)?
Multidestination Flooding	For L2 mcast and broadcast, flood, drop, or flood within epg encap? If flooding with EPG encap, proxy-arp is required for cross-epg L2 communication
L2 Unknown Unicast	If destination mac is unicast and unknown, flood or proxy to spines?

Proxied, L2 Unknown Unicast is dropped if the Destination MAC isn't known in COOP

Bridge-Domain Level Forwarding Options

Feature	What Does it Do?
Limit IP Learning to Subnet	Only learn IP's if they are within the configured BD subnet for local learns.
Unicast Routing	Enable IP learning as well as routing (if a BD subnet is configured)
Disable IP Dataplane Learning	Only for PBR! Only local MAC's are learned via DP. IP's and remote macs learned via ARP.
ARP Flooding	When disabled, ARP is unicast routed based on the Target IP (if known)




```
leaf# show endpoint ip 192.168.100.11
leaf# show ip route 192.168.100.11 vrf CL2022:vrf1

192.168.100.0/24, ubest/mbest: 1/0, direct, pervasive
    *via 10.0.176.66%overlay-1, [1/0], 01w00d, static
        recursive next hop: 10.0.176.66/32%overlay-1
```

Proxy!

EPG Level Forwarding Options

Feature	What Does it Do?
Flood in Encapsulation	Feature is enabled for just the EPG (rather than all epg's in the BD). Requires proxy arp for L2 traffic between encaps.
L4-L7 Virtual IP's	Designed for Direct Server Return flows. This disables dataplane learning per IP. IP is learned by ARP/ND.
Disable DP Learning Per-IP/Prefix	Disables dataplane learning for non DSR scenarios. More specific than VRF-level option

New in 5.2

Global Forwarding Options

Feature	What Does it Do?
Enforce Subnet Check	Don’t learn an IP (both local and remote) if it is not within a configured BD subnet in the VRF.
Disable Remote EP Learning on BL’s	Remote IP learning is disabled for Unicast flows on a leaf in a specific VRF if an l3out exists in the same VRF

Multicast sources are still learned

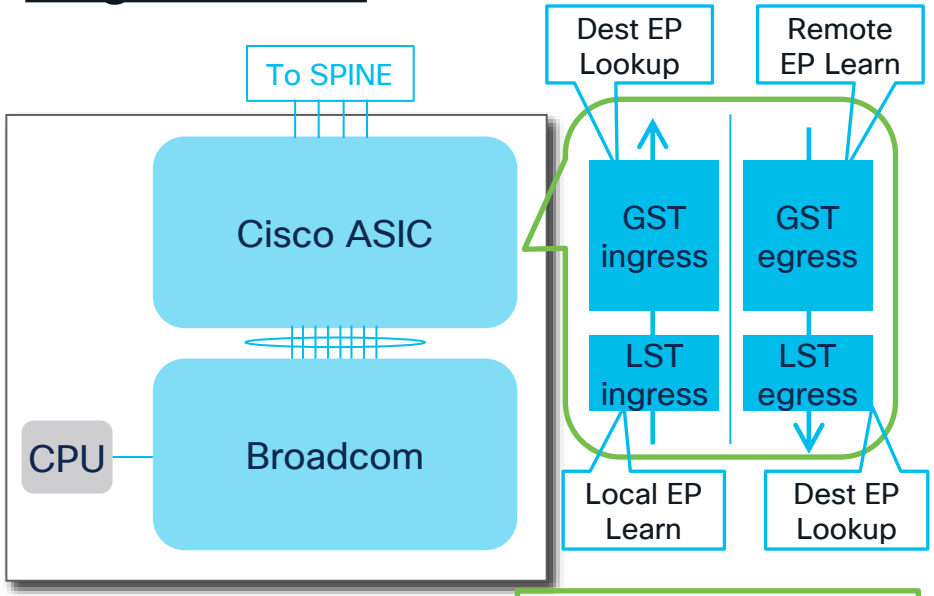
Also implicitly disabled when intersite l3out is configured

The Anatomy of an ACI Switch

LEAF ASIC Generations

※ LST: Local Station Table, GST: Global Station Table
※ FP Tile: Forwarding and Policy Tile

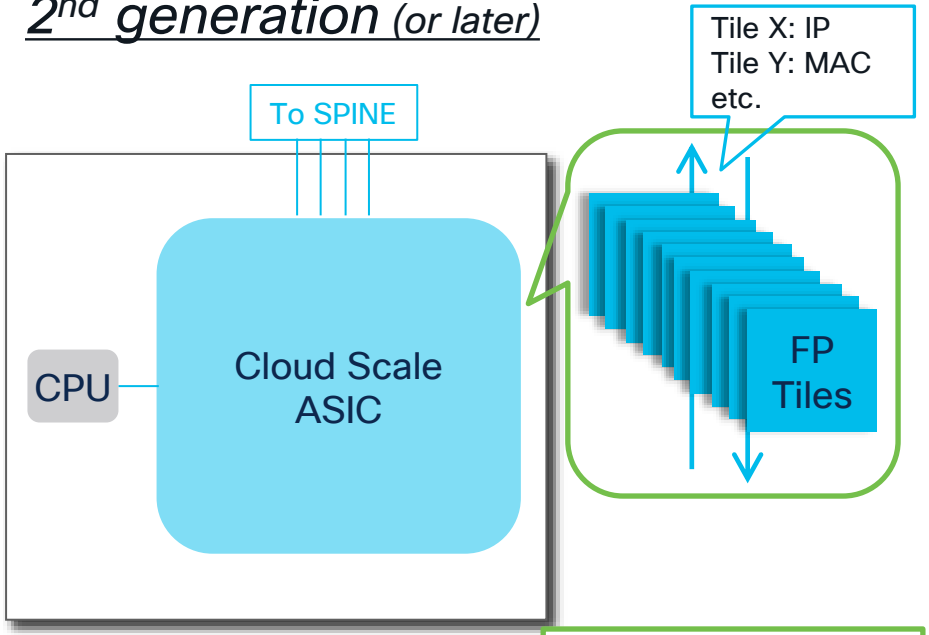
1st generation



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

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

2nd generation (or later)



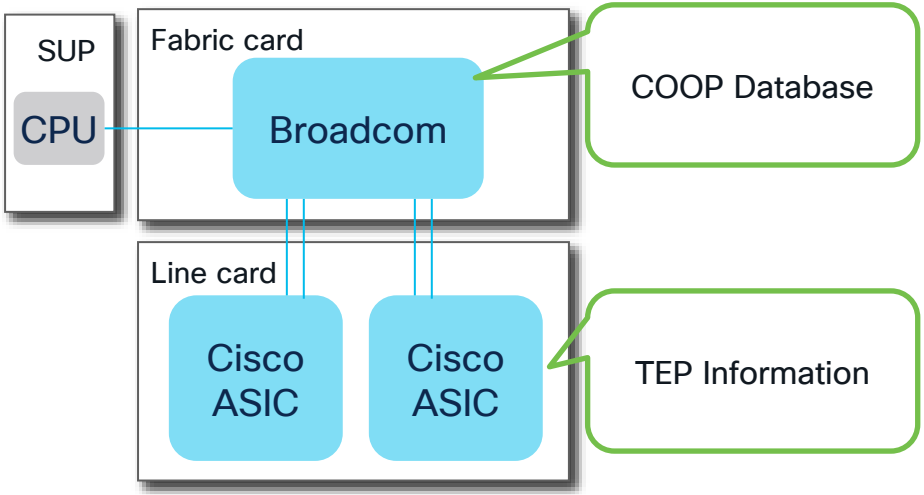
N9K-C*-EX
N9K-C*-FX
N9K-C*-FX2
N9K-C*-FX3
N9K-C*-FXP
N9K-C*-GX
N9K-C*-GX2

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

SPINE ASIC Generations

※ number of ASIC per card depends on model

1st generation

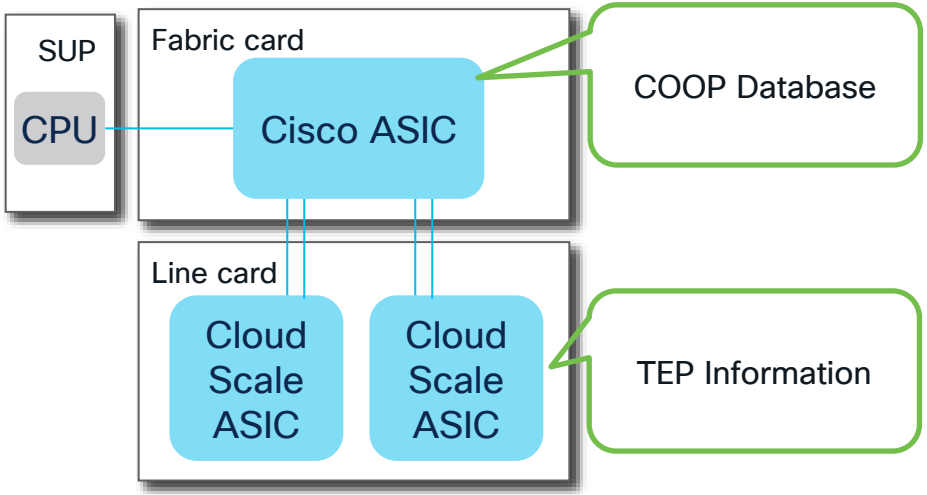


Line card
N9K-X9736PQ

Box spine
N9K-C9336PQ

Fabric card
N9K-C9504-FM
N9K-C9508-FM
N9K-C9516-FM

2nd generation (or later)

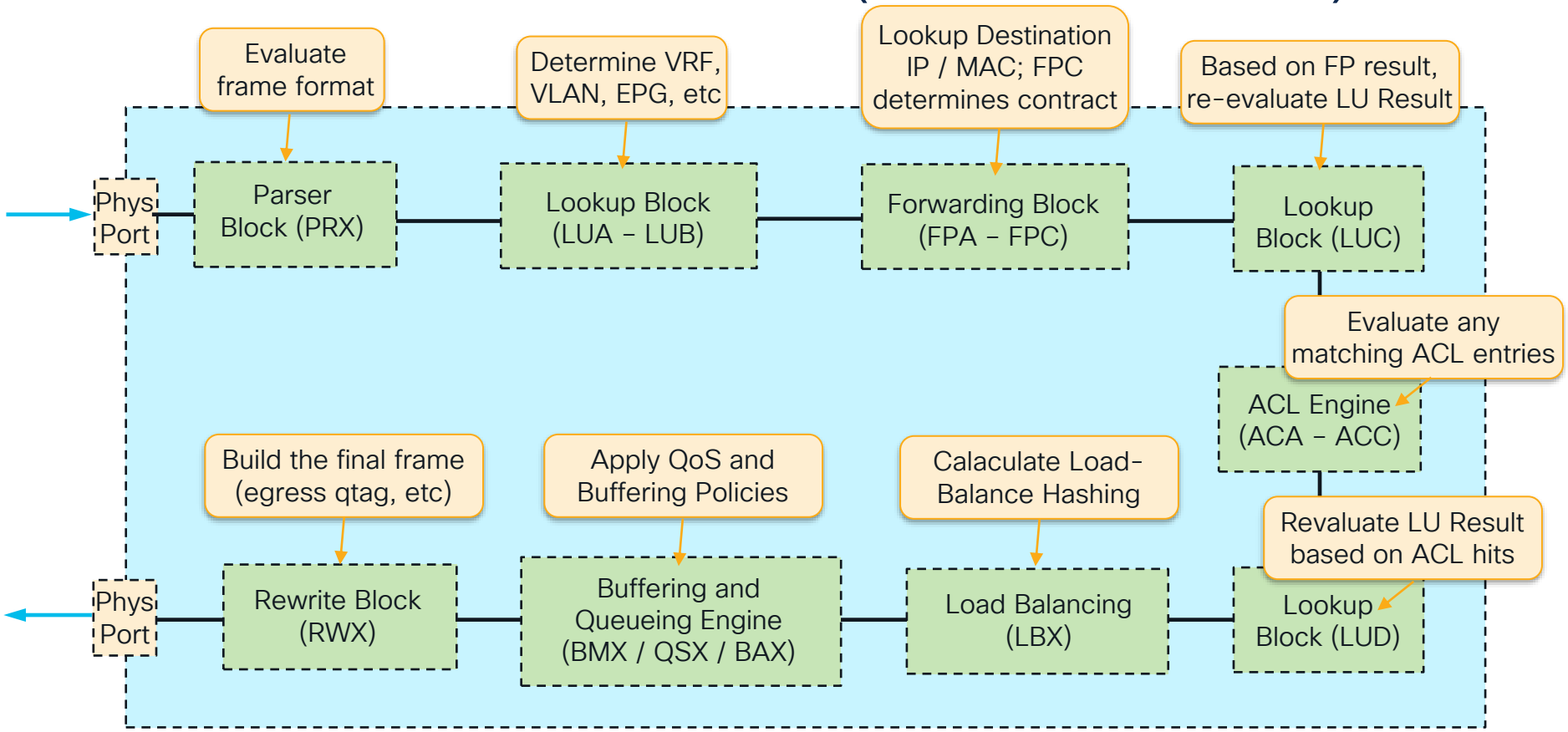


Line card
N9K-*X

Box spine
N9K-*C
N9K-*X

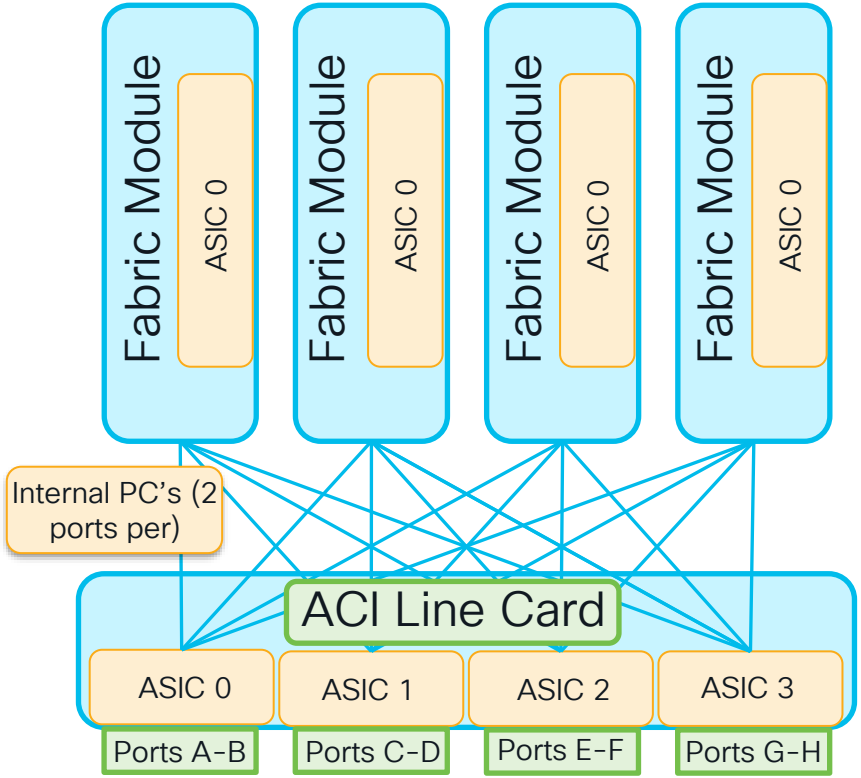
Fabric card
N9K-C*FM-E
N9K-C*FM-E2
N9K-C*FM-G

Inside an ACI Switch ASIC (Gen 2 and Later)



※ number of ASIC per card depends on model

Inside an ACI Modular Spine



What are the strange IP's on the Fabric Modules?

```
sp# vsh -c "slot 26 show plat internal hal l3 routes"
40.0.99.139/ 32
3.124.199.13/ 32
0.156.151.177/ 32
```

Where are the linecard forwarding tables?

```
sp# vsh -c "slot 2 show plat internal hal l3 routes"
<no output>
```

Inside an ACI Modular Spine

How is traffic forwarded?

For Proxied Traffic

- Depending on if the dest IP is the L2 or L3 Proxy TEP the VRF VNID + Dest IP OR BD VNID + Dest MAC is used to hash a synthetic Dest IP and VRF ID
- Synthetic information is used on LC to hash the uplink port to FM
- FM routing lookup is based on Synthetic IP
- Each Synthetic IP is owned by two FM's
- FM uses vnTag to tell egress LC which front panel port to use

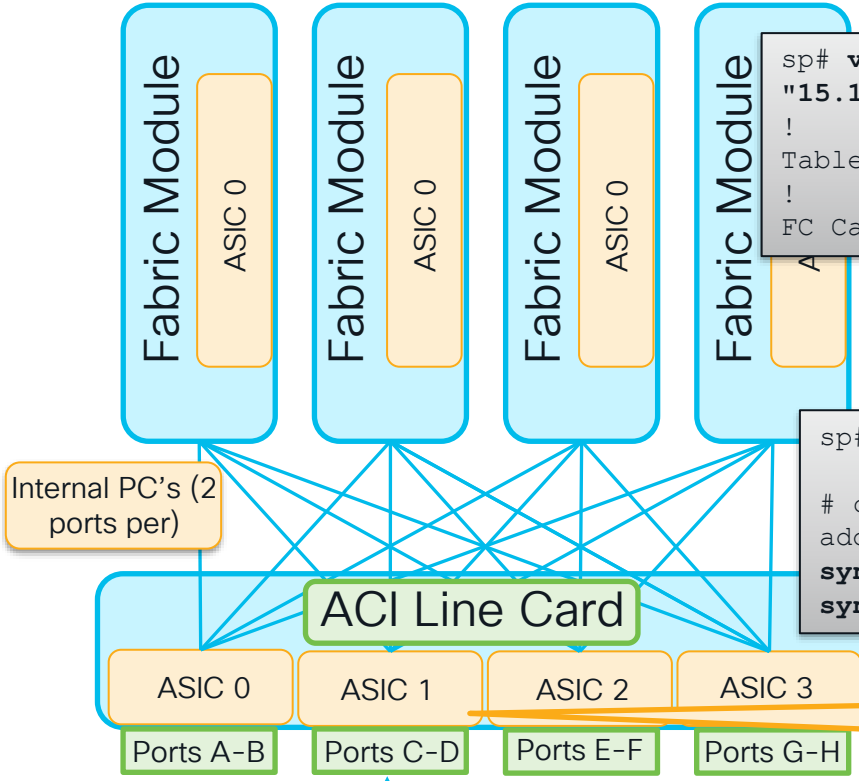
Inside an ACI Modular Spine

How is traffic forwarded?

For Transit Traffic

- Line card hashes across ALL FM uplinks
- ALL FM’s have overlay TEP routes
- FM uses vnTag to tell egress LC which front panel port to use

Inside an ACI Modular Spine




```
sp# vsh -c "slot 26 show forwarding route platform" | grep -A 10
"15.180.164.253"
!
Table: (IN-HW) Type=100 Vrf=750 Synth=0
!
FC Cards/ASICs : FC22/ASIC-0     FC26/ASIC-0
```

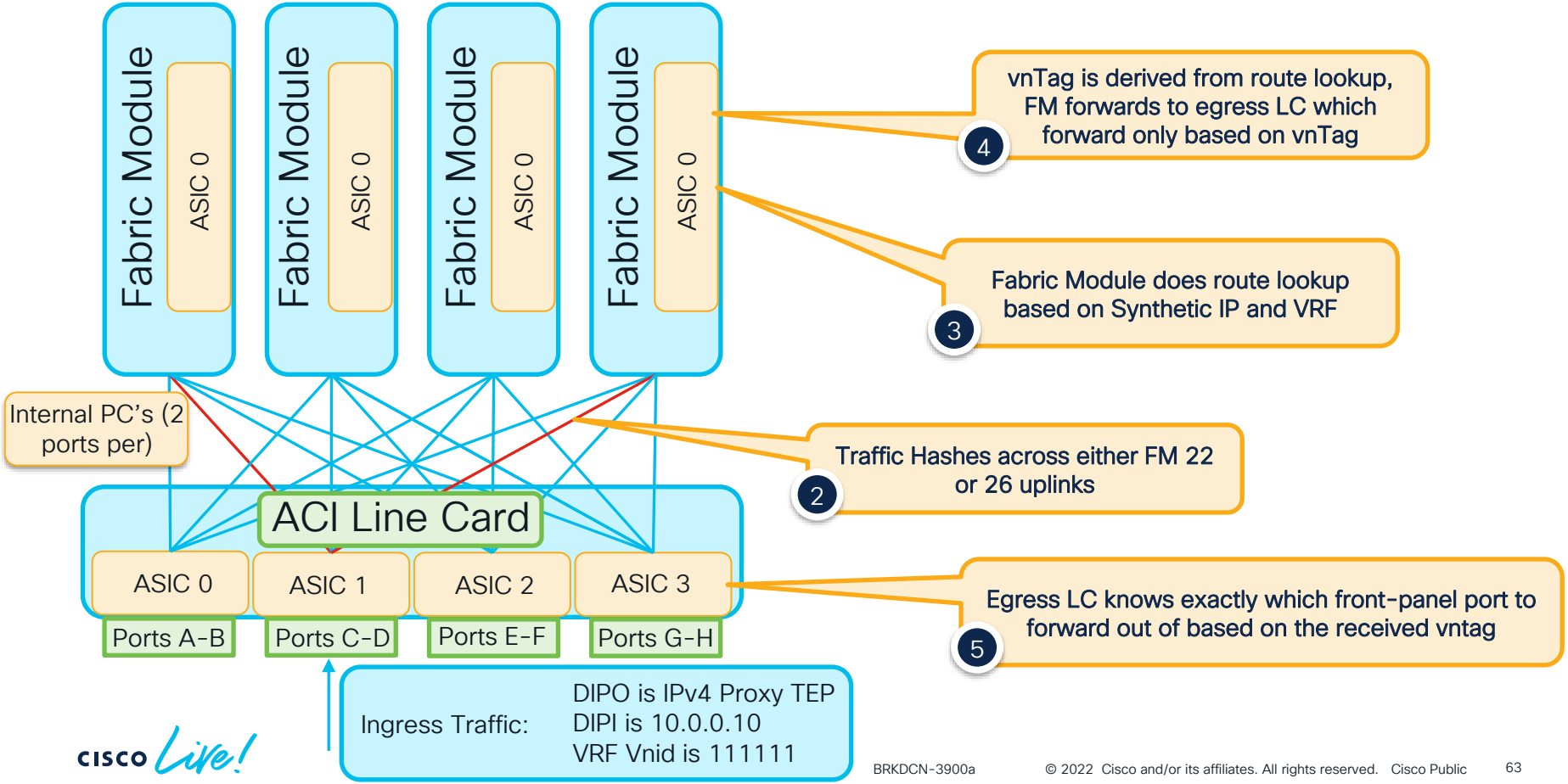
Synthetic info programmed on FM's 22 and 26

```
sp# moquery -c coopIpv4Rec -f 'coop.Ipv4Rec.addr=="10.0.0.10"'

# coop.Ipv4Rec
addr          : 10.0.0.10
synthIp       : 15.180.164.253
synthVrf      : 250
```

1 Ingress LC hashes Synth IP and VRF based on Real Tenant IP + VRF

Ingress Traffic:
DIPO is IPv4 Proxy TEP
DIPI is 10.0.0.10
VRF Vnid is 111111

Inside an ACI Modular Spine



Technical Session Surveys

- Attendees who fill out a minimum of four session surveys and the overall event survey will get Cisco Live branded socks!
- Attendees will also earn 100 points in the Cisco Live Game for every survey completed.
- These points help you get on the leaderboard and increase your chances of winning daily and grand prizes.

Cisco Learning and Certifications

From technology training and team development to Cisco certifications and learning plans, let us help you empower your business and career. www.cisco.com/go/certs

Pay for Learning with Cisco Learning Credits

(CLCs) are prepaid training vouchers redeemed directly with Cisco.

Learn

Cisco U.
IT learning hub that guides teams and learners toward their goals

Cisco Digital Learning
Subscription-based product, technology, and certification training

Cisco Modeling Labs
Network simulation platform for design, testing, and troubleshooting

Cisco Learning Network
Resource community portal for certifications and learning

Train

Cisco Training Bootcamps
Intensive team & individual automation and technology training programs

Cisco Learning Partner Program
Authorized training partners supporting Cisco technology and career certifications

Cisco Instructor-led and Virtual Instructor-led training
Accelerated curriculum of product, technology, and certification courses

Certify

Cisco Certifications and Specialist Certifications
Award-winning certification program empowers students and IT Professionals to advance their technical careers

Cisco Guided Study Groups
180-day certification prep program with learning and support

Cisco Continuing Education Program
Recertification training options for Cisco certified individuals

Here at the event? Visit us at **The Learning and Certifications lounge at the World of Solutions**

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Thank you

cisco Live!

ALL IN

#CiscoLive