Cisco live!

January 28 – February 1, 2019 · Barcelona

cisco

IMAGINE

INTUITIVE

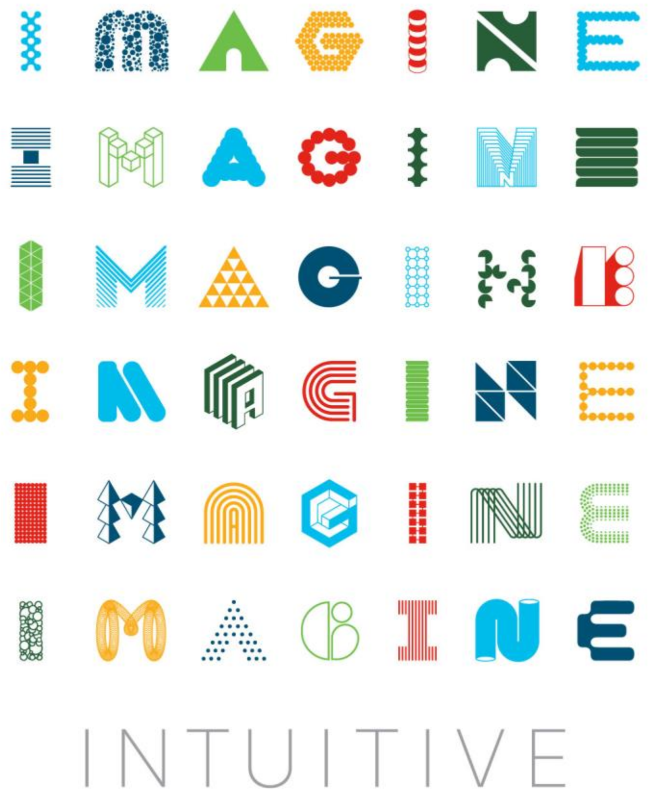
BRKACI-3101

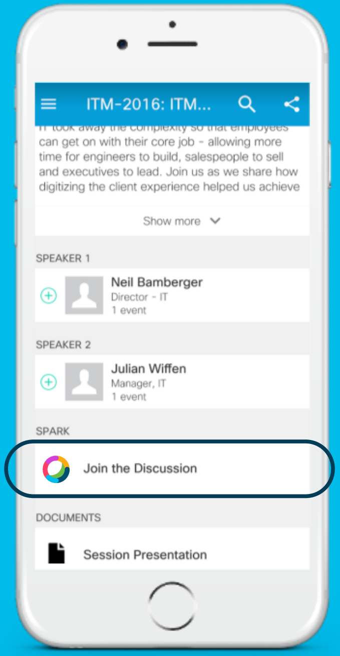
ACI Under the Hood

How Your Configuration is Deployed

Joseph Ristaino, Technical Leader – DCBU ACI Escalation

Carlo Schmidt, Technical Leader – ACI Solution Support





cs.co/ciscolivebot#BRKACI-3101

Cisco Webex Teams

Questions?

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

How

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

Agenda

- Introduction
- Building the Overlay
  - Access Policies
  - Configuration Deployment and Validation
  - Loop Prevention
- Traversing the Overlay
  - Learning, Forwarding, and Policy Enforcement
  - Shared Services and Route Leaking
  - L3outs and Routing Protocols

Acronyms/Definitions

Acronyms	Definitions	Acronyms	Definitions
ACI	Application Centric Infrastructure	LPM	Longest Prefix Match
ACL	Access Control List	MDT	Multicast Distribution Tree
APIC/IFC	Application Policy Infrastructure Controller/ Insieme Fabric Controller	MST	Multiple Spanning Tree
BD	Bridge Domain	pcTag	Policy Control Tag
COOP	Council of Oracle Protocol	PL	Physical Local
ECMP	Equal Cost Multipath	SVI	Switch Virtual Interface
EP	Endpoint	TC	Topology Change
EPG	Endpoint Group	VL	Virtual Local
FTEP/VTEP	Fabric/Virtual or VXLAN Tunnel Endpoint	VNID	Virtual Network Identifier
GIPo	Outer Group IP Address	VXLAN/iVXLAN	Virtual Extensible LAN / Insieme VXLAN
ISIS	Intermediate System to Intermediate System	XR	VXLAN Remote

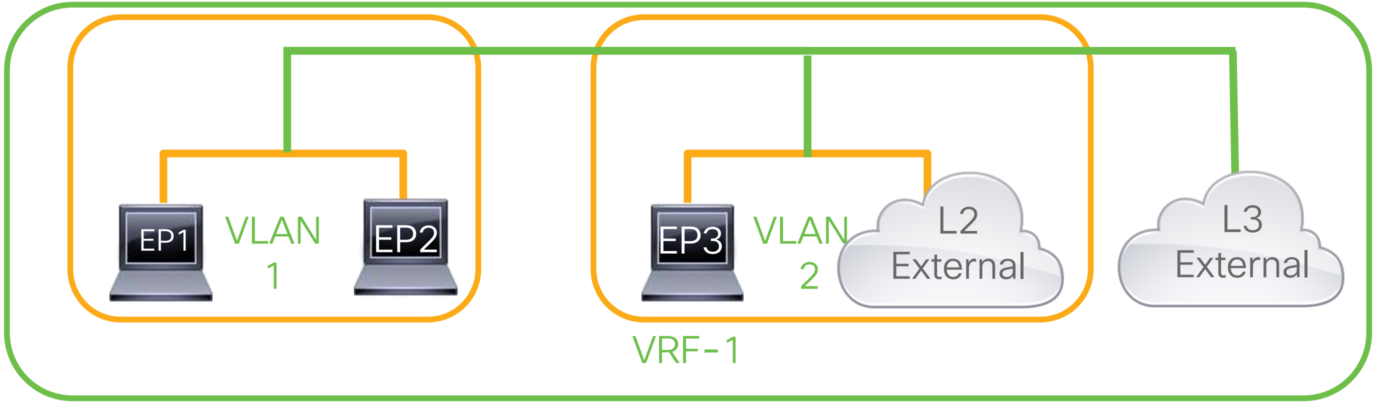
➔ Reference Slide

Introduction

Introduction

What are our basic network requirements?

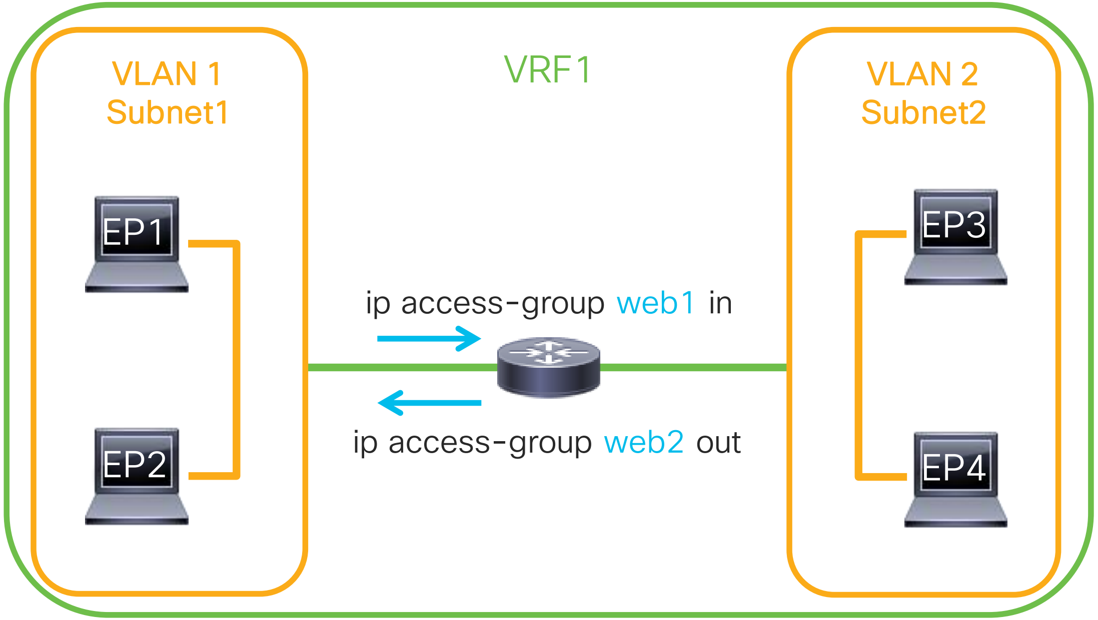
1) Provide paths for endpoints to communicate at Layer2(MAC) and Layer3(IP)
2) Provide separation of endpoint into Layer2 forwarding domains (vlan or BD)
3) Routing between IP/IPv6 subnets and allow separation of these into multiple VRFs
4) Communication to external L2 networks (DCI)
5) Communication to external L3 networks (WAN)



Introduction

What are our basic network requirements?

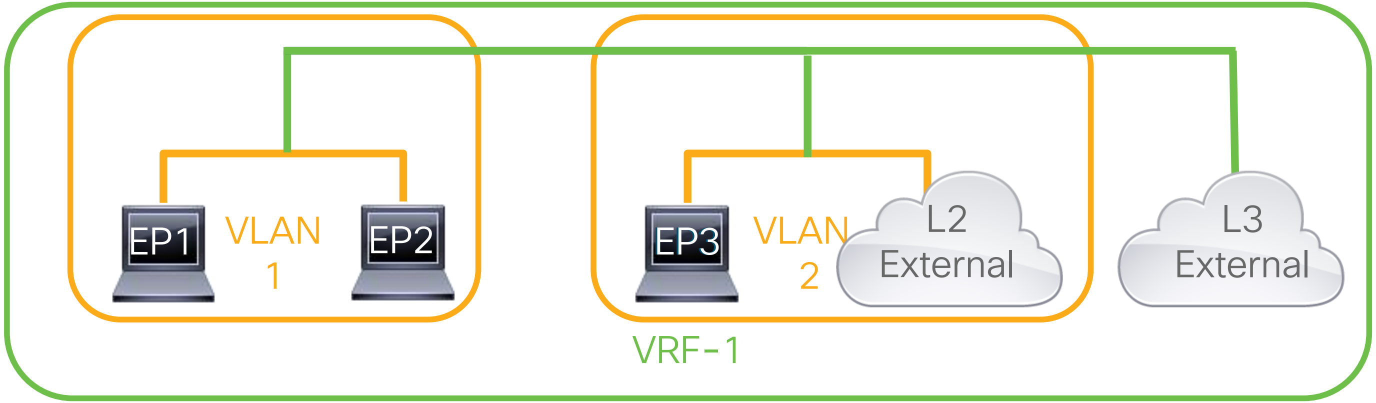
6) Allow security policies in order to limit communication to between endpoints to allowed protocols




```
ip access-list web-in
  permit tcp Subnet1 Subnet2 eq 80
ip access-list web-out
  permit tcp Subnet2 eq 80 Subnet1
```

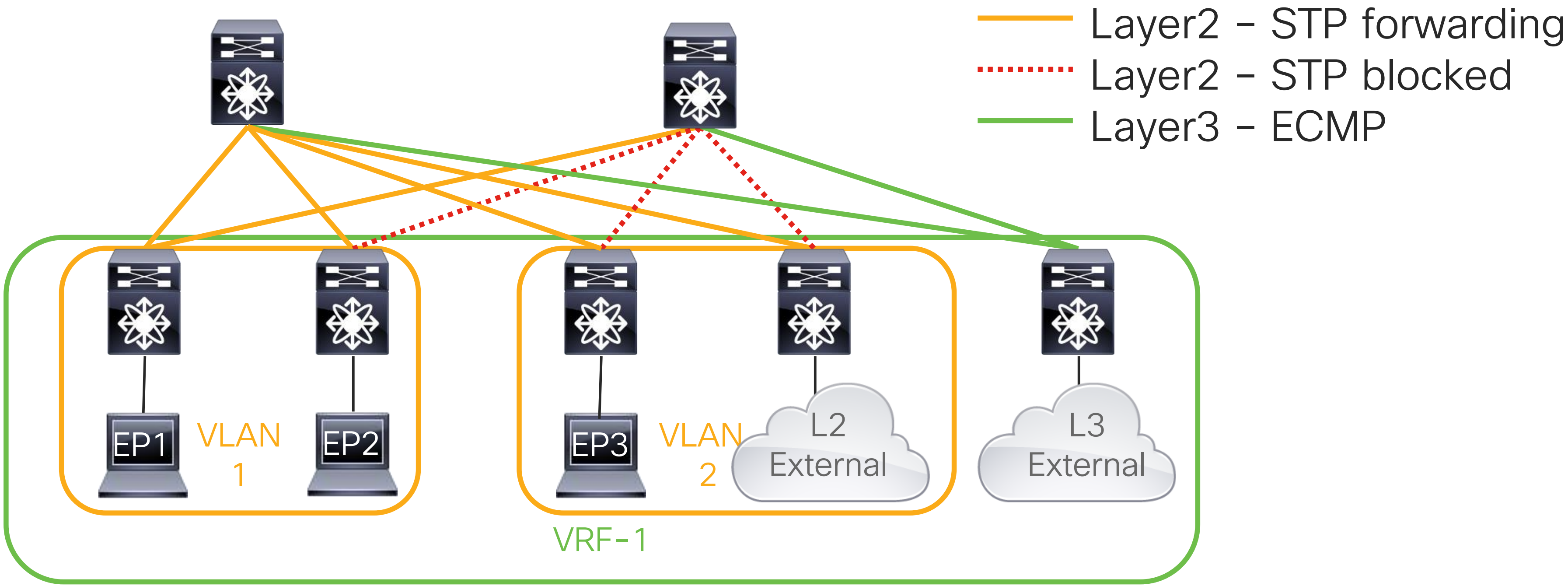
What physical topology is required?

Physical topology must support our endpoint communication (layer-2 / layer-3), and the location of endpoints within the physical network will affect the supporting design/configuration.



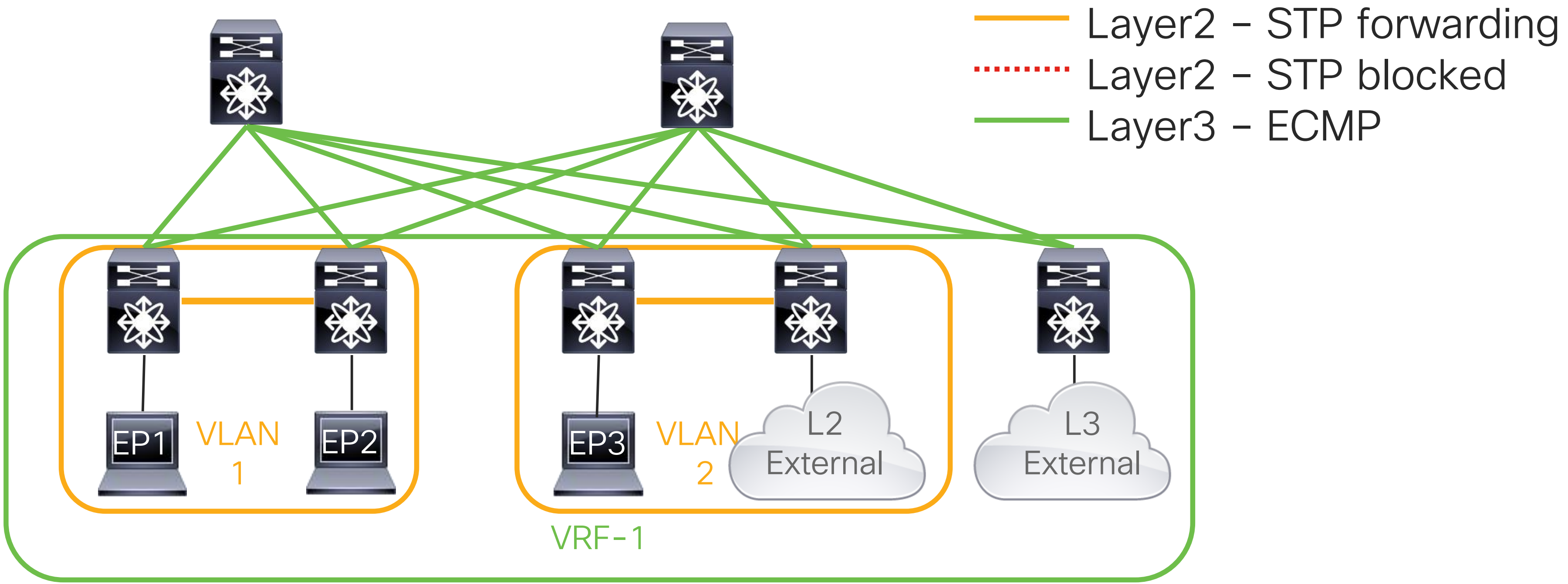
Traditional Topology – Routing at Core/Spine

STP results in unused links / limits scale / slower convergence



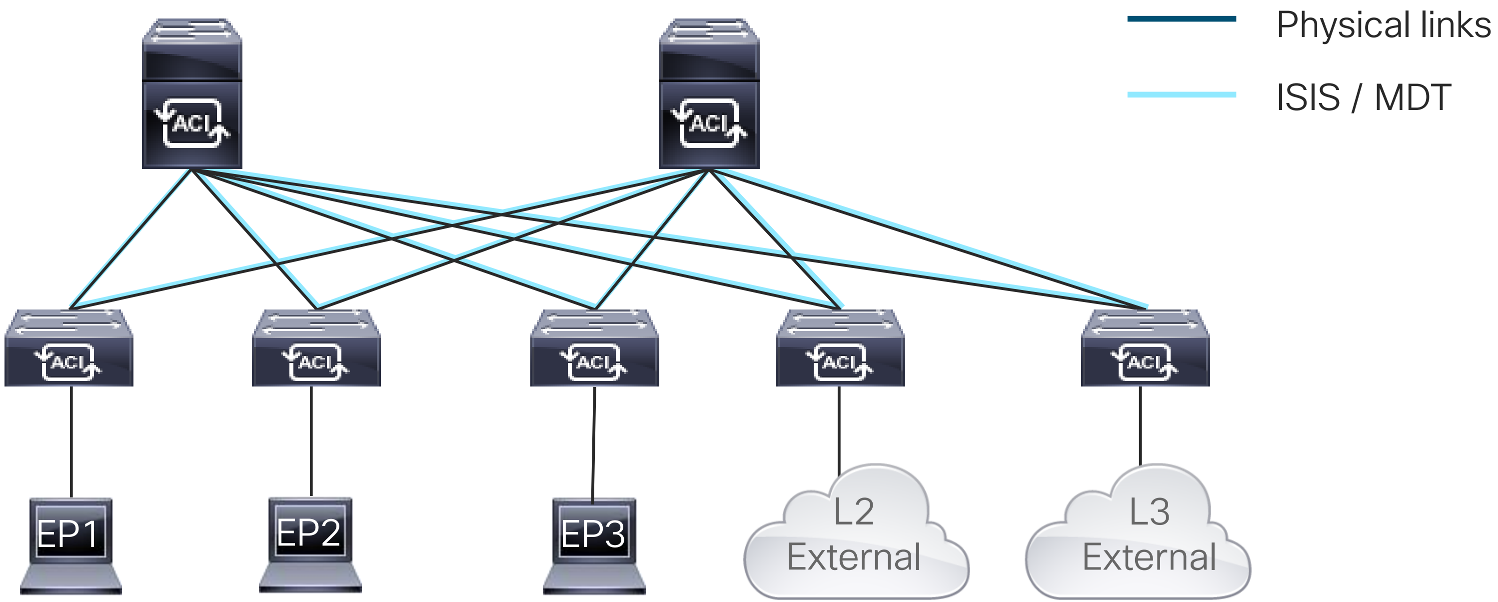
Traditional Topology – Routing at Access

Restricts L2 endpoint locations / requires separate links for L2 / segmented STP



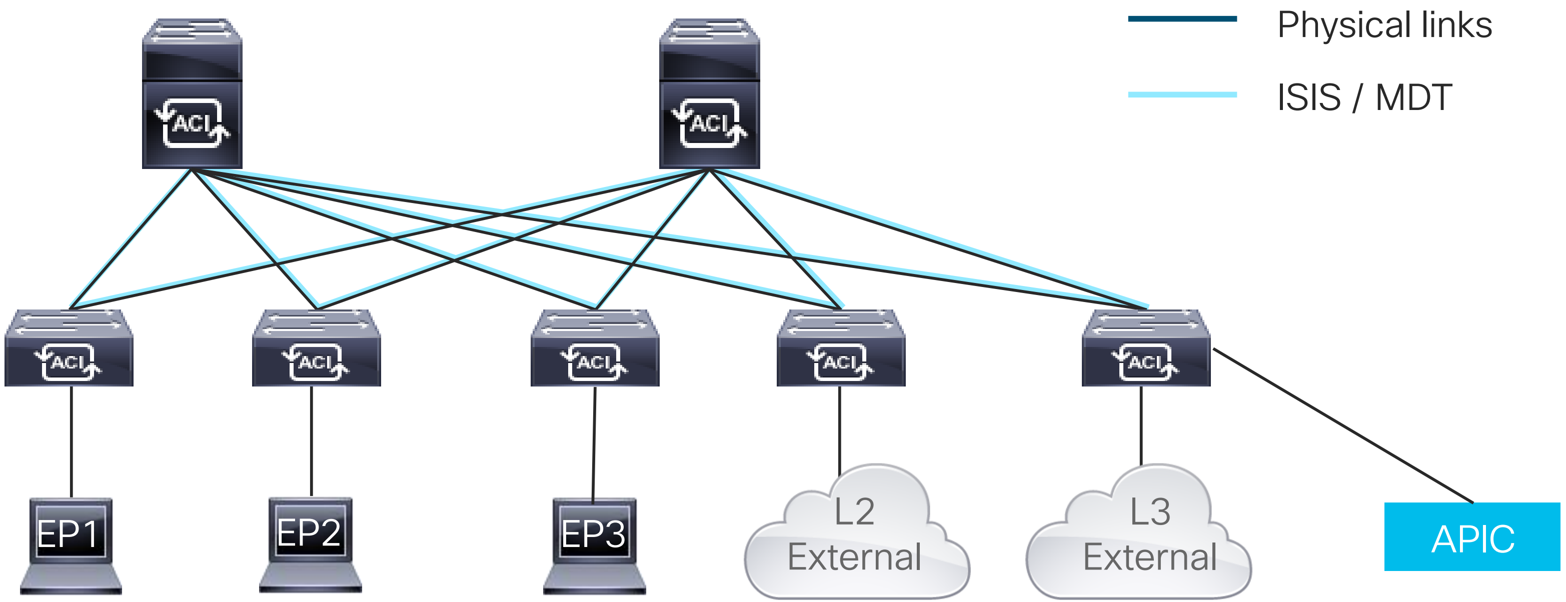
ACI Infrastructure

ISIS is run on links between spines / leaves



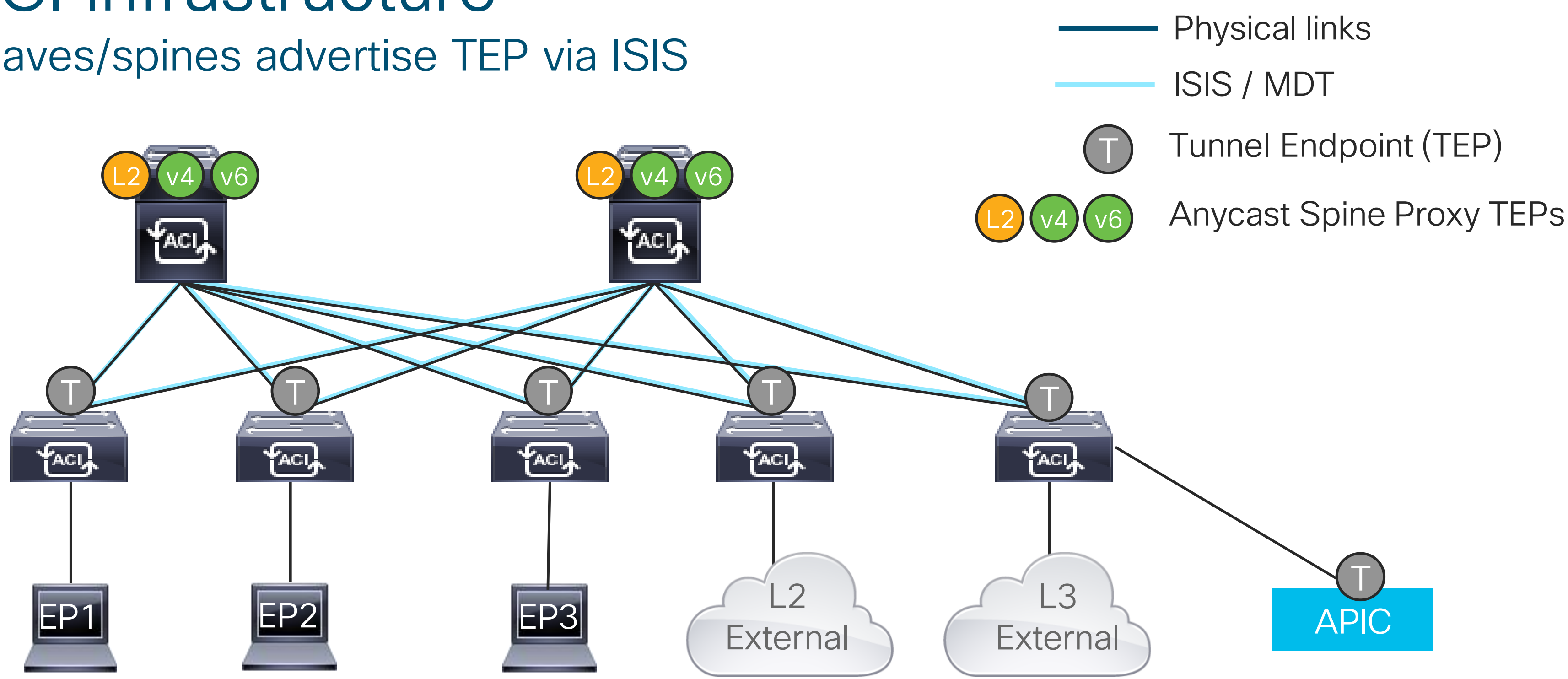
ACI Infrastructure

APICs communicate to fabric over infra vlan



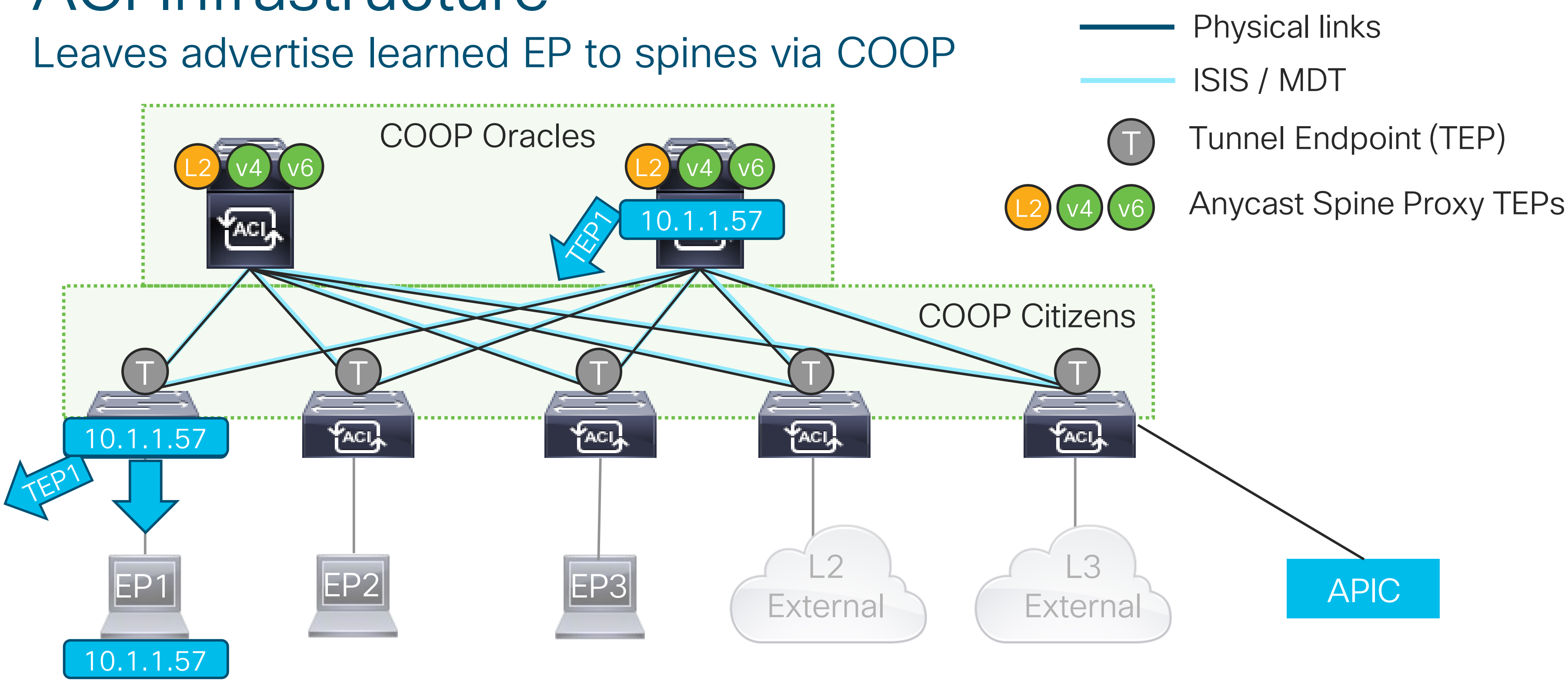
ACI Infrastructure

Leaves/spines advertise TEP via ISIS



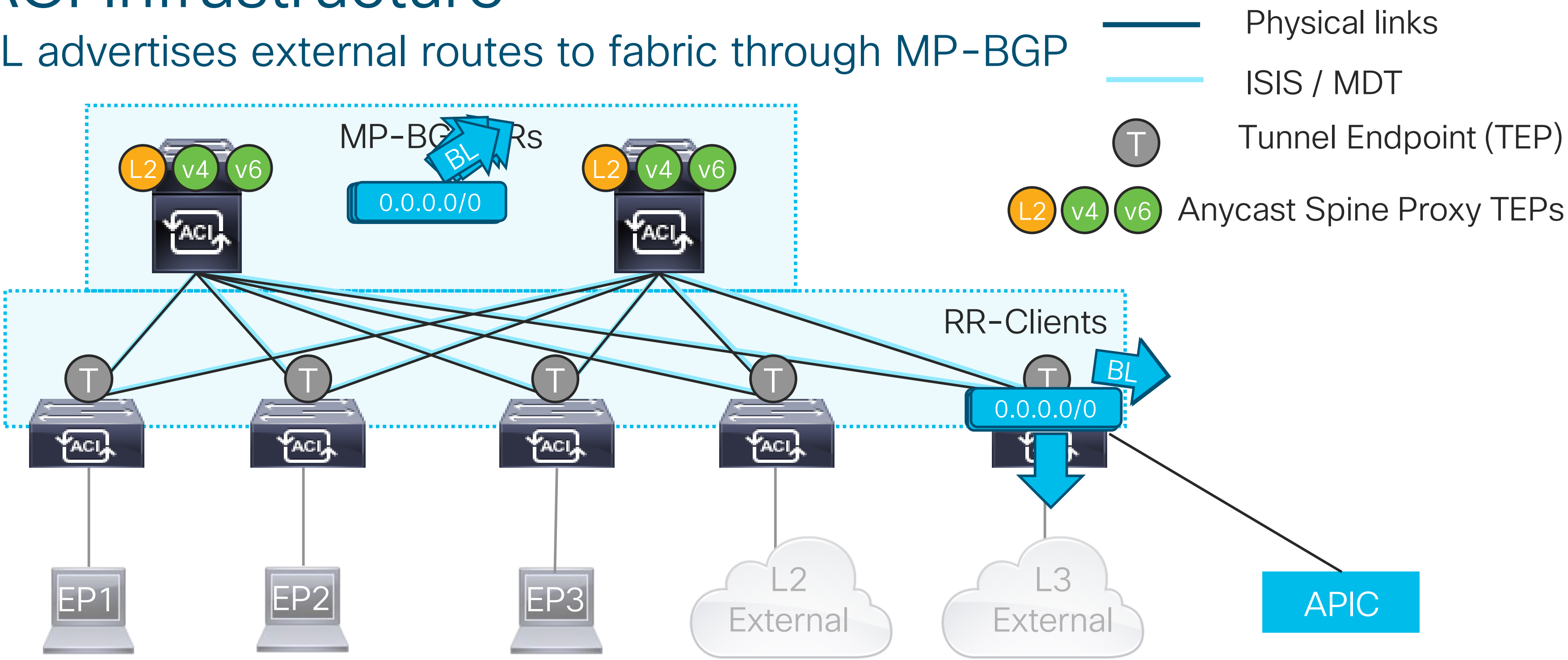
ACI Infrastructure

Leaves advertise learned EP to spines via COOP



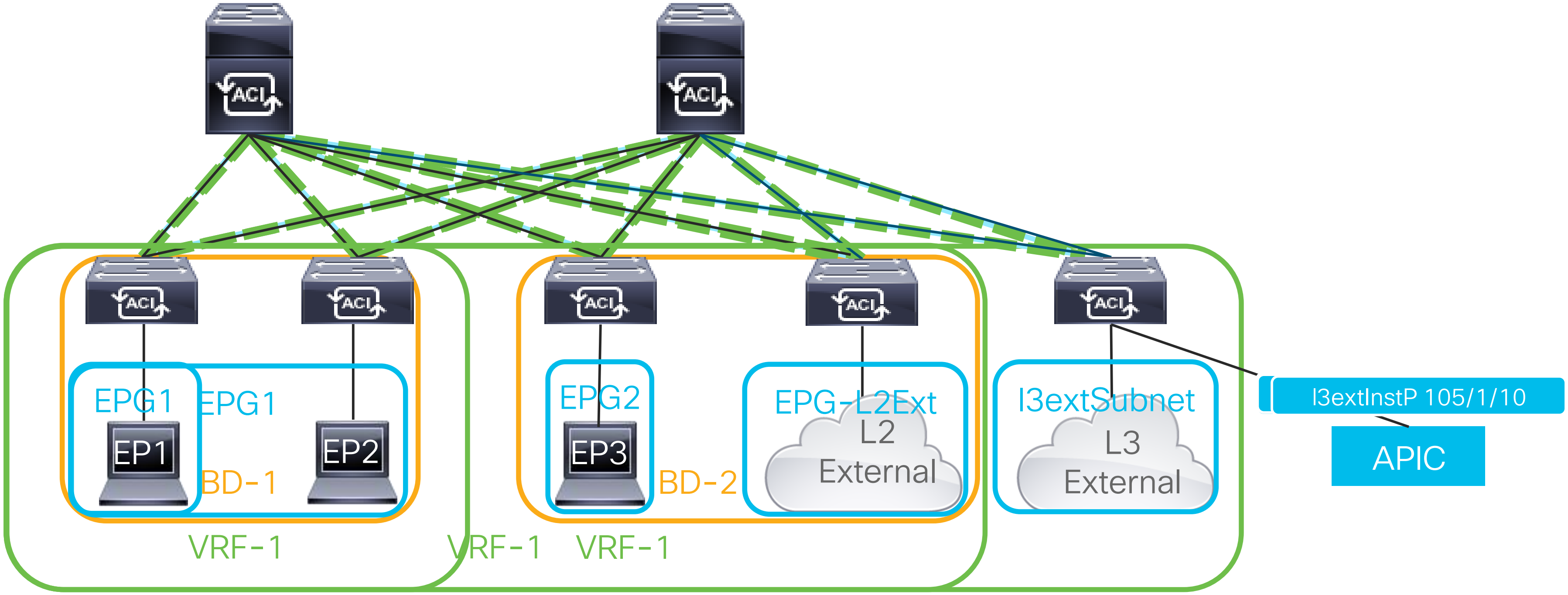
ACI Infrastructure

BL advertises external routes to fabric through MP-BGP



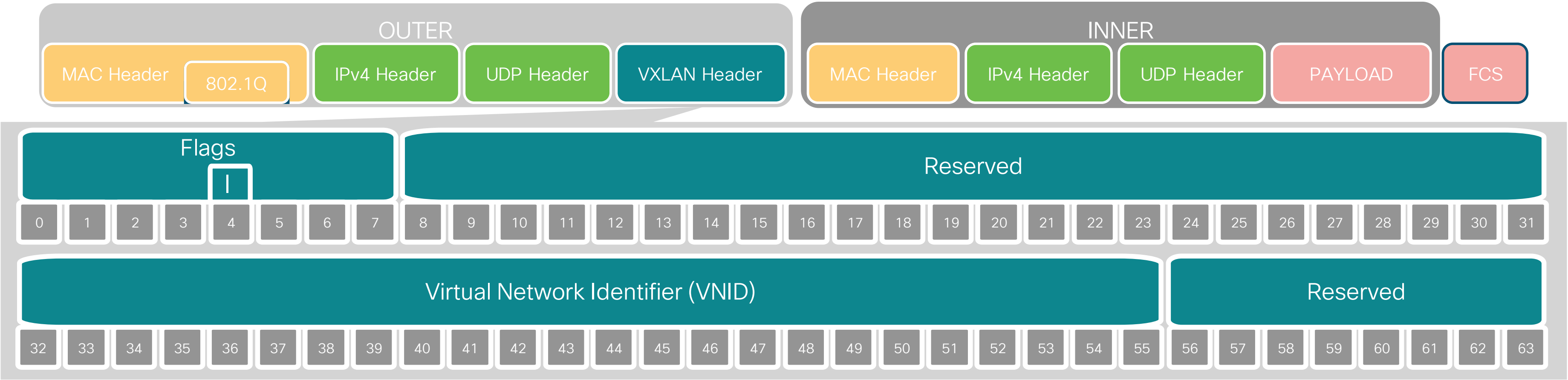
ACI Infrastructure

APIC provisions BD/VRF VXLAN overlays based on EPG attachments



VXLAN

VXLAN differentiates tunneled traffic based on VNID field.

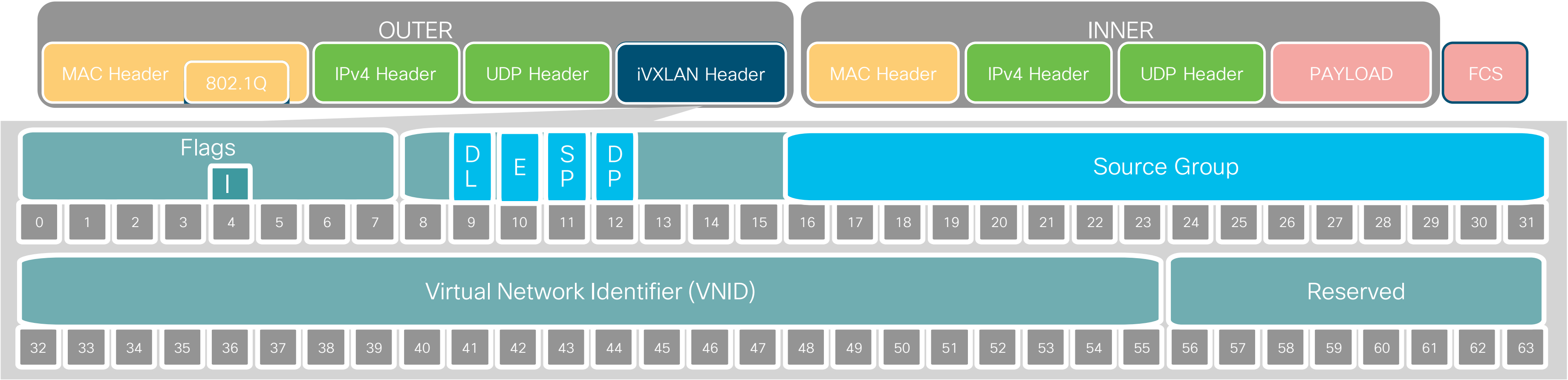


iVXLAN

In addition to differentiating traffic based on VNID, iVXLAN allows the source EPG of traffic to be identified by the Source Group (PCTAG) bits and to determine if policy was applied by source (SP) / destination (DP). Endpoint Learning can be enabled/disabled via the Don't Learn (DL) Bit.

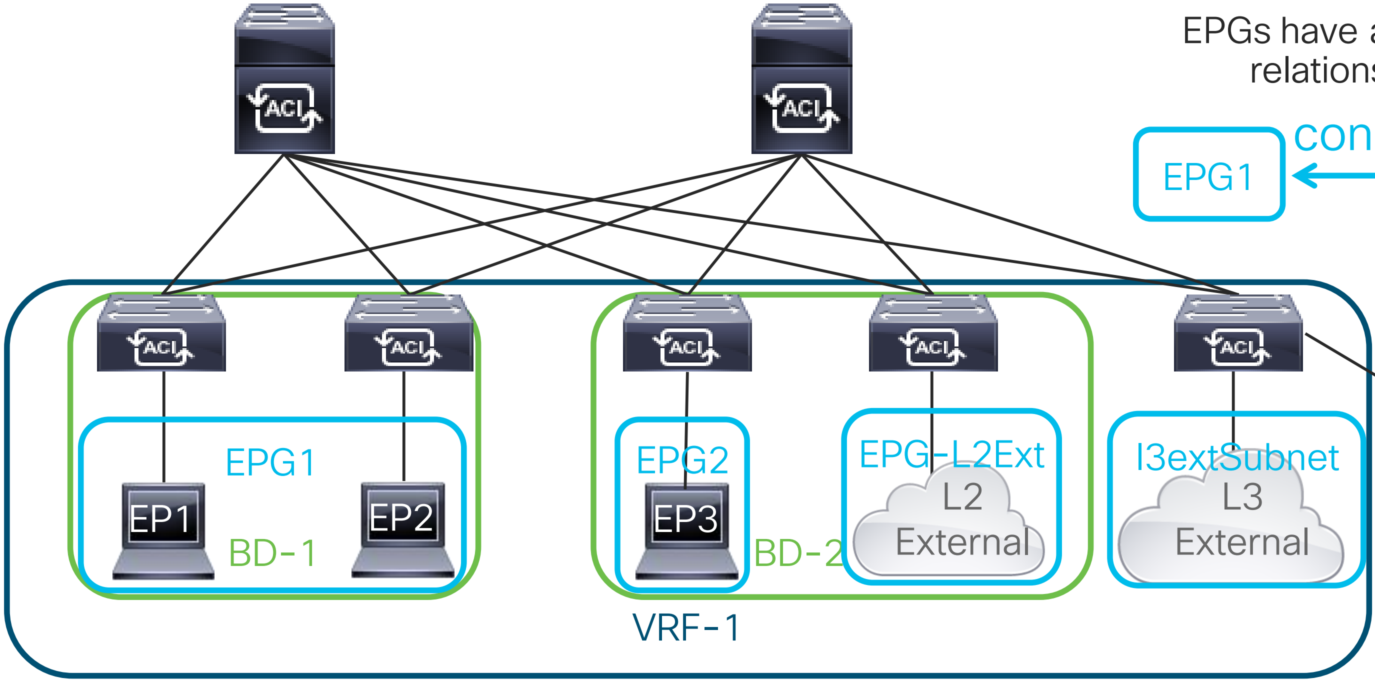
Exception (E) bit ensures packet cannot be sent back into the fabric for certain flows. Blocks Loops.

Example is Proxy Flow. Packet was proxied and should not be re-directed anywhere else.

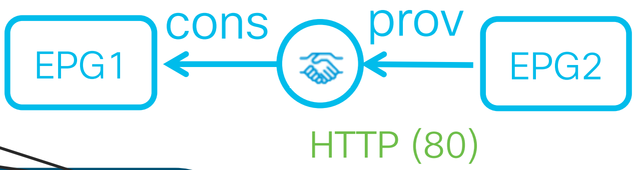


ACI Infrastructure

Policy is implemented through contracts / filters specifying allowed traffic



EPGs have a consumer / provider relationship to a contract.

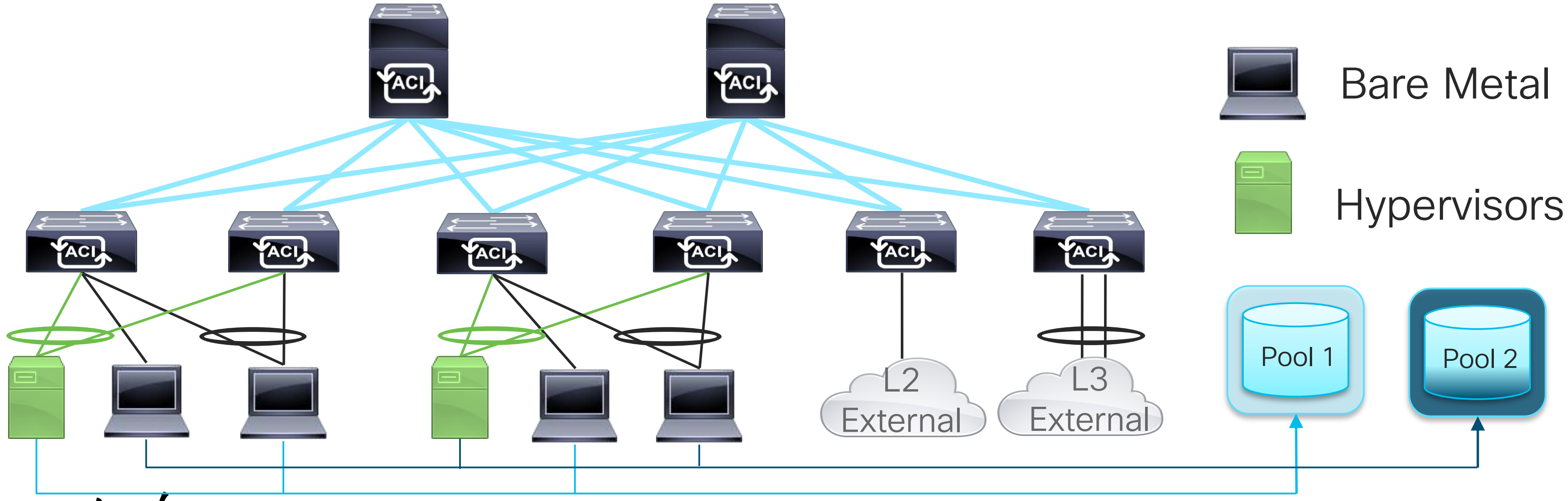


Access Policies

Access Policies

What is the goal? What are we trying to accomplish?

1) Provide consistent configurations across the whole fabric.
2) A simplified and well organized configuration, where policy is defined once and re-used.
3) Define what policies are allowed to be deployed on leafs/ports
4) Restrict Resource deployment in a multi-tenant environment.



Access Policies

Access policies refer to the configuration that is applied for physical and virtual (hypervisors/VMs) devices attached to the fabric.

Broken into a few major areas:

Switch Policy

- Policies
- Policy Groups
- Profiles

Interface Policy

- Policies
- Policy Groups
- Profiles

Global Policy

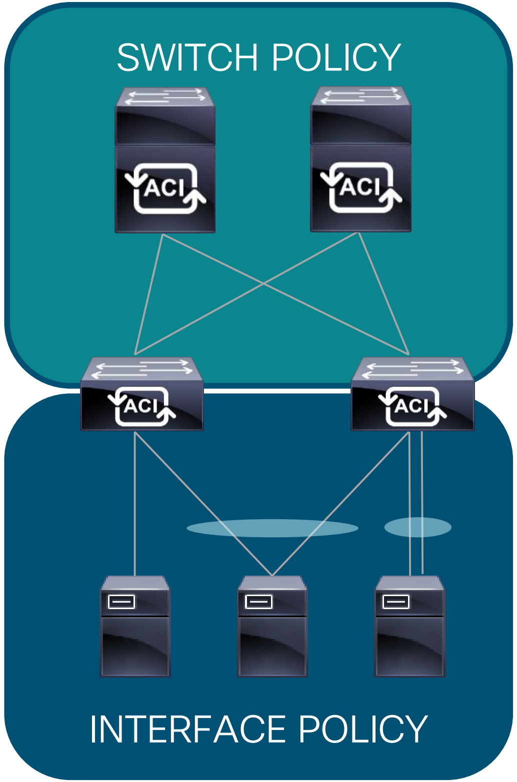
- Pools
- Domains
- Attachable Access Entity Profiles

Access Policies

Policies define protocol / feature configurations

Policy Groups select which policies should be applied

Profiles associate policy groups to switches or interfaces, through the use of selectors



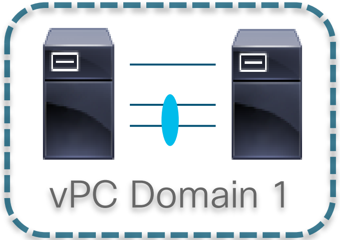
Switch Policy Types:

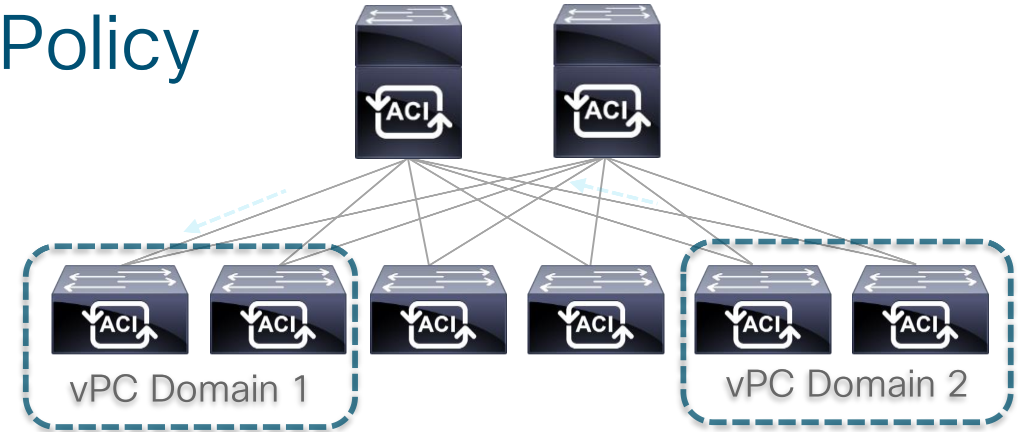
- VPC Domain
- Spanning-tree (MST)
- BFD
- Fibre-channel SAN/Node

Interface Policy Types:

- Link-level
- CDP
- LLDP
- Port-channel / LAG
- Port-channel member
- Spanning-tree
- Storm Control
- Data plane policing
- MCP
- L2 (Vlan local / global)
- Firewall

vPC Protection Group Policy





Classical vPC Domain configuration

Required configuration of domain, peer-link, and peer-keepalive link on both devices in domain

```
vpc domain 1
  peer-keepalive destination 172.168.1.2 /
    source 172.168.1.1 vrf vpc-keepalive
  peer-gateway
  ip arp synchronize

interface port-channel 20
  vpc peer-link
```

ACI vPC Domain configuration

Specify the Domain ID and the two Leaf switch IDs that form the domain pair

VPC Protection Group

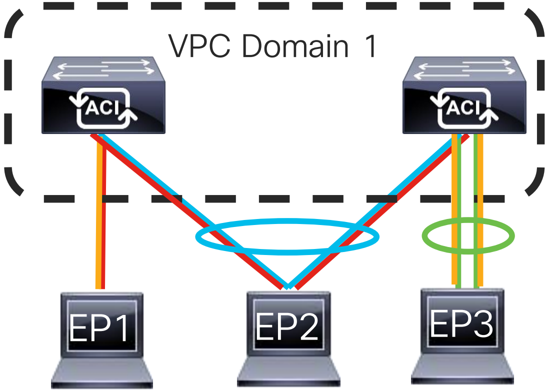
Name: vPC-Domain100
ID: 100
Switch1: 101
Switch2: 102

Interface Policies

Used to define a particular policy for a given interface level function. The intention of Interface Policies is that they are defined once and re-used among interfaces that need like policies.

Examples:

- LLDP On/Off
- CDP On/Off
- Port-Channel
    - LACP
    - Mode On
- Storm Control
- MACsec



Interface Policy Groups

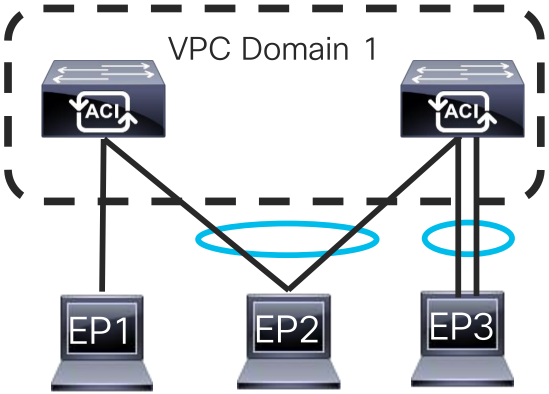
Used to specify which interface policies to be applied to a particular interface type.
It also associates an AEP (which defines which domains are allowed on the interface).

Types:

Access port (EP1)

Access Bundle Groups

- Virtual Port-channel (EP2)
- Port-channel (EP3)



Note: Separate policy groups should be created for each port-channel (standard or VPC) that you need to configure. All interfaces on leaf that are associated with a particular access bundle group reside in same channel.

Global Policy

Pools (Vlan / VXLAN)

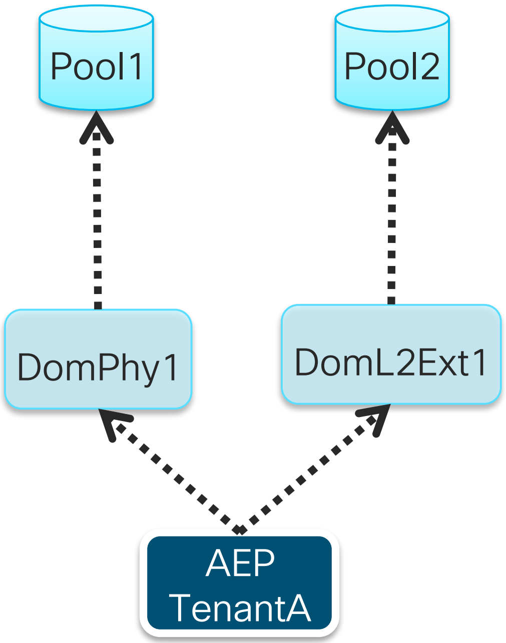
A resource pool of encapsulations that can be allocated within the fabric.

Domains (Physical / VMM / External Bridged / External Routed)

Administrative domain which selects a vlan/vxlan pool for allocation of encaps within the domain

Attachable Access Entity Profiles (AEP)

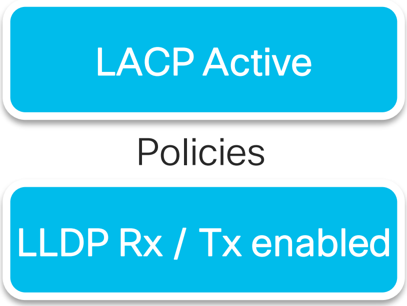
Selects one or more domains and is referenced/applied by interface policy groups.

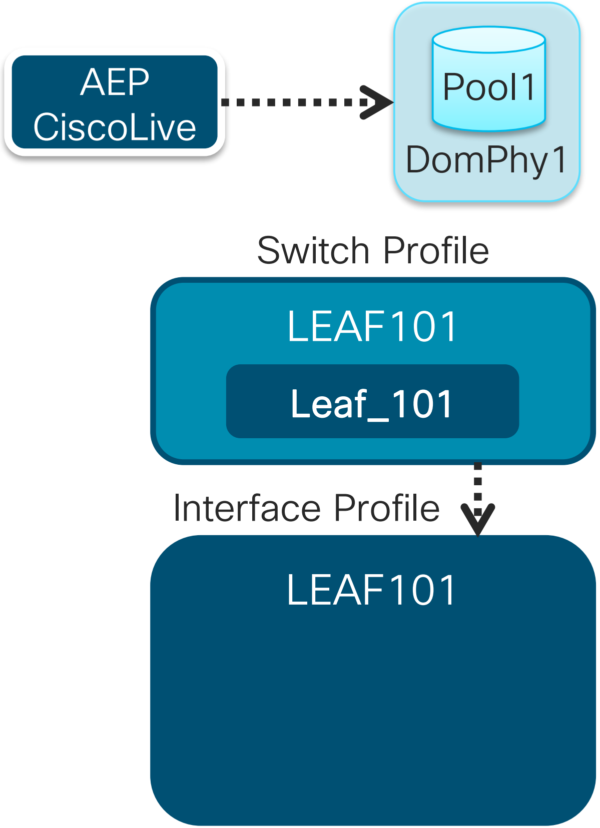


Access Policy Example

General Configuration (reused for many interfaces):

1) Configure a physical domain and vlan pool
2) Create an AEP and associate physical domain
3) Create switch/interfaces profiles for leaf (LEAF101)
   - very easy to apply configurations if you create a switch/interface profile for each leaf and one for each VPC domain pair
4) Configure Interface policies (LACP / LLDP)

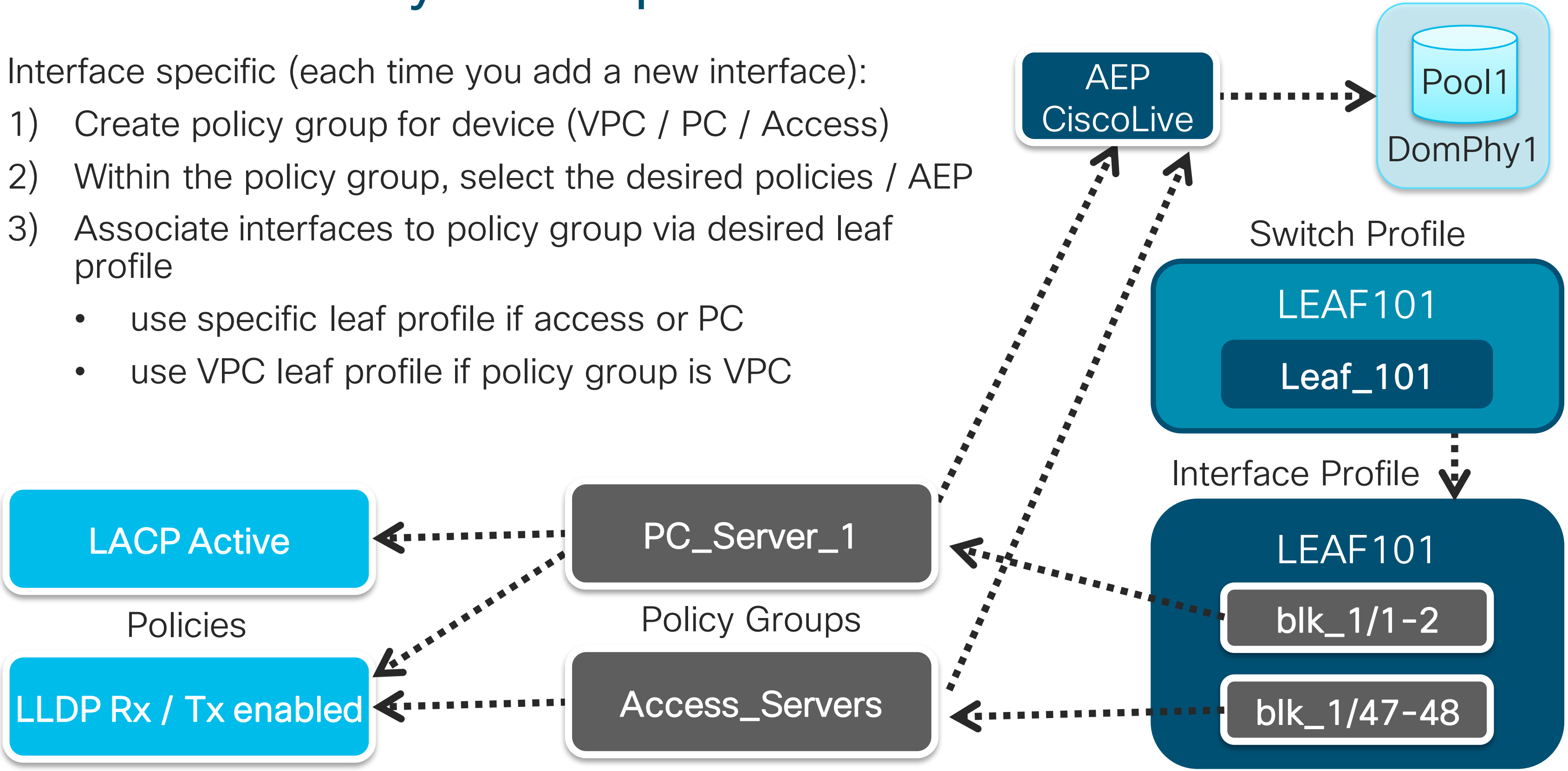




Access Policy Example

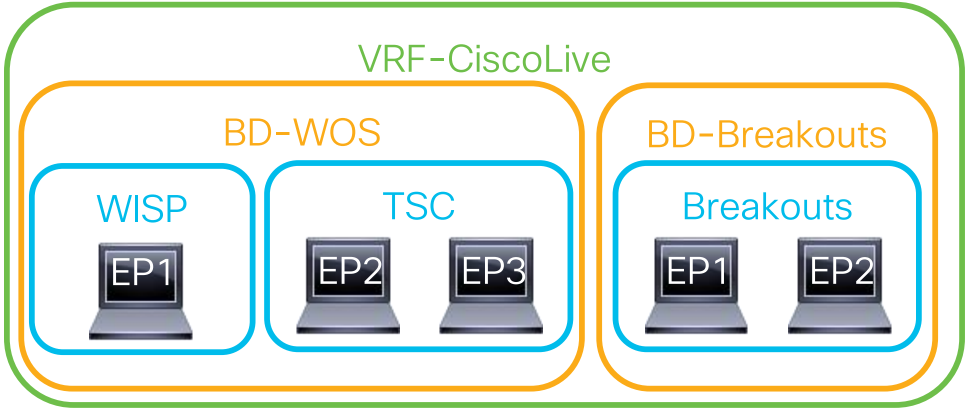
Interface specific (each time you add a new interface):

1) Create policy group for device (VPC / PC / Access)
2) Within the policy group, select the desired policies / AEP
3) Associate interfaces to policy group via desired leaf profile
   - use specific leaf profile if access or PC
   - use VPC leaf profile if policy group is VPC



Configuration Deployment and Validation

VRF/BD/EPG Logical Configuration



Classical configuration steps

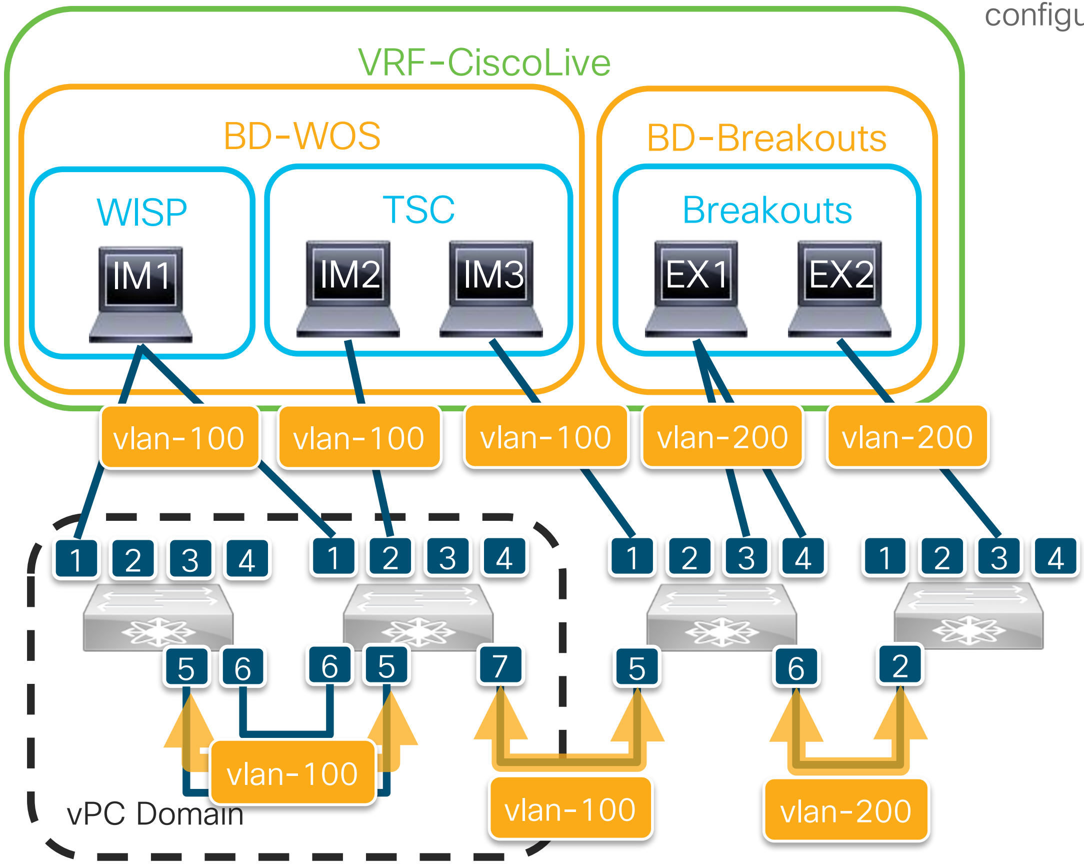
- Create VRF
- Create Vlans
- Create Vlan interfaces
  - Associate to VRF
  - Assign Subnets / configure gateway redundancy
- Assign encapsulation to interfaces

ACI Logical configuration

- Create Tenant
  - Create VRF
  - Create BDs
    - Associate to VRF
    - Define a Subnet (optional)
  - Create App Profile
    - Create EPGs
    - Associate to Domain
    - Define a Subnet (optional)

Classical VRF/BD config

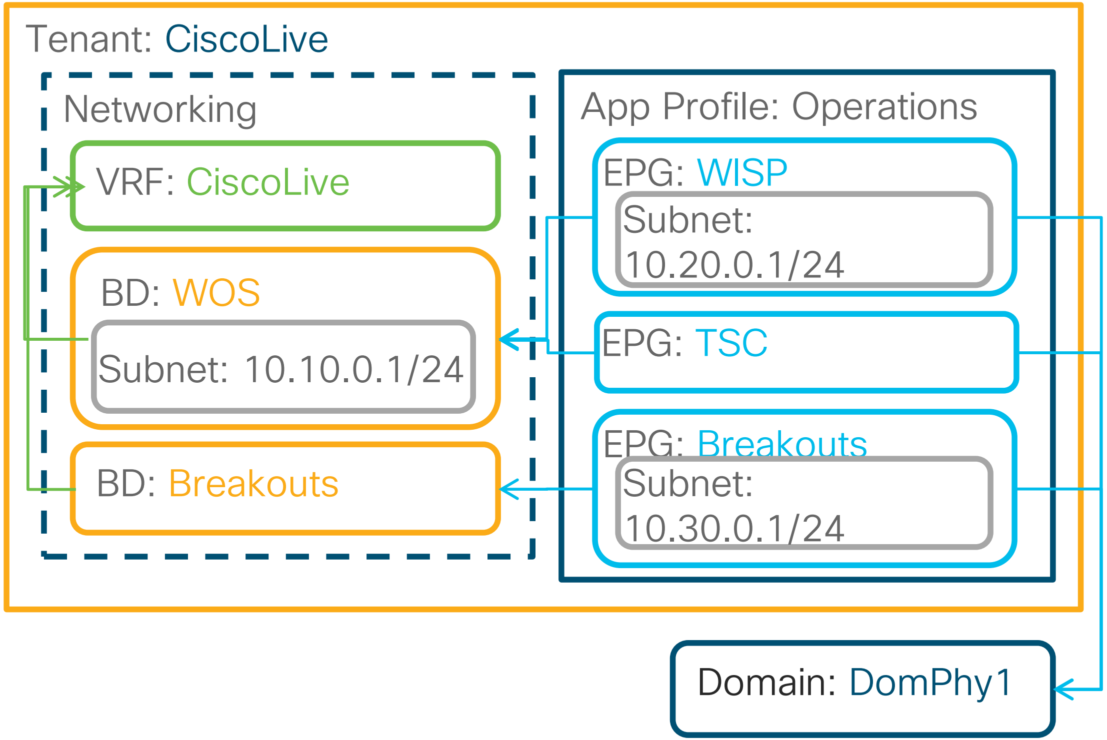
Each node must be individually configured with the VRF, associated vlans/BDs, and an SVI with unique IP. For gateway redundancy, HSRP must also be configured.




```
vrf context CiscoLive
vlan 100
  name WOS
vlan 200
  name Breakouts
feature interface-vlan
feature hsrp
interface Vlan100
  vrf member CiscoLive
  ip address 10.10.0.2/24
  ip address 10.20.0.2/24 secondary
  hsrp 100
    ip 10.10.0.1
interface Vlan200
  vrf member CiscoLive
  ip address 10.30.0.2/24
  hsrp 200
    ip 10.30.0.1

interface Ethernet1/1
  switchport trunk vlan allowed 100
interface Port-channel1
  switchport access vlan 200
```

ACI Logical Configuration



- Create Tenant
  - Create VRF
  - Create BDs
    - Associate to VRF
    - Define a Subnet (optional)
  - Create an App Profile
    - Create EPGs
    - Associate to Domain
    - Define a Subnet (optional)

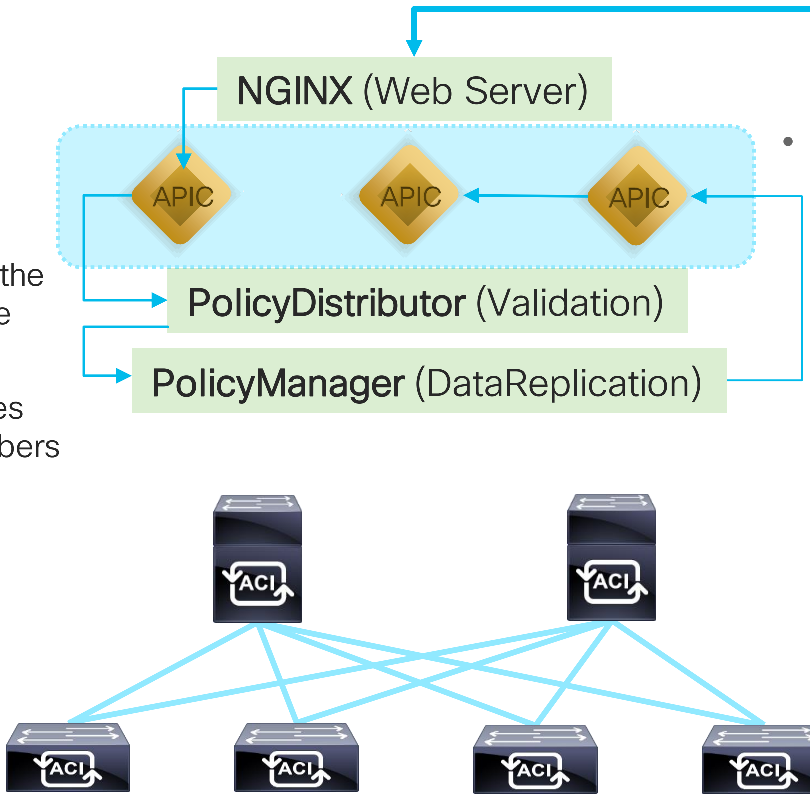
What have we accomplished?
Specified the logical configuration that should be deployed on each leaf where EPG is deployed. We also restricted which interfaces can deploy the EPG through Domain associations.

ACI Logical Configuration Deployment

NGINX Receives REST API Call and Parses Request

PolicyDistributor Validates the Configuration is Deployable

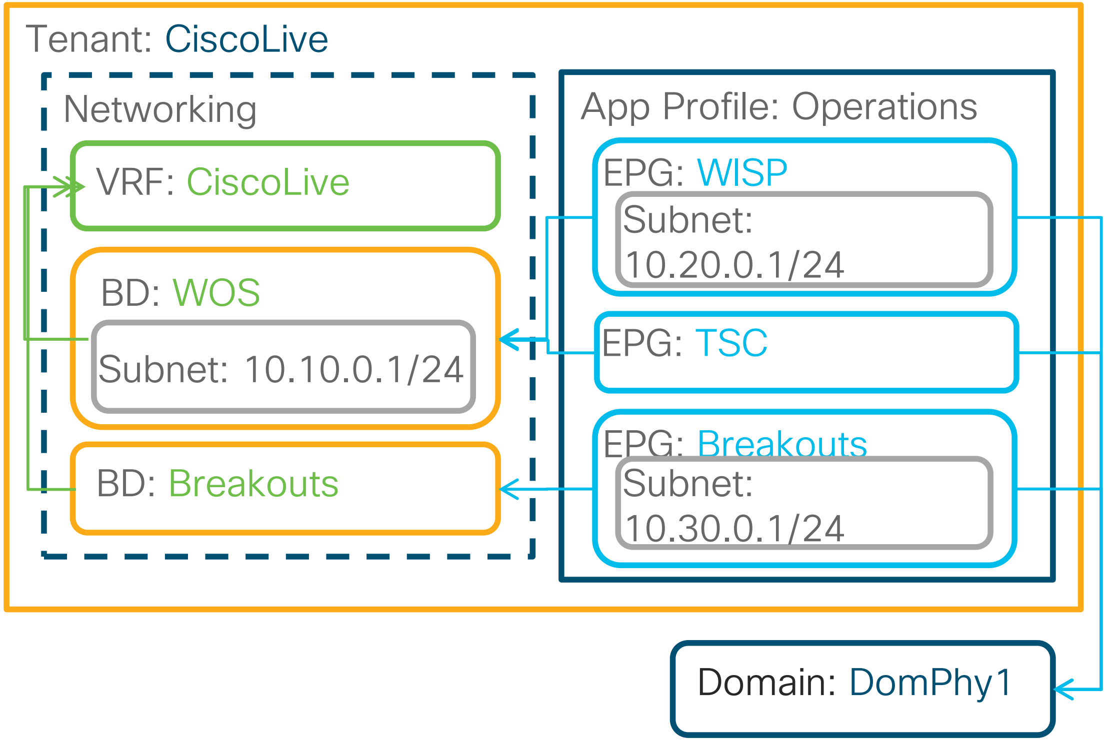
PolicyManager Writes the Config to DB and Distributes Data to other Cluster Members



- Create Tenant (fvTenant)
  - Create VRF (fvCtx)
  - Create BDs (fvBD)
    - Associate to VRF
    - Define a Subnet (optional)
  - Create an App Profile
    - Create EPGs (fvAEPg)
    - Associate to Domain
    - Define a Subnet (optional)

NOTE: No Policy is Pushed to Switches Yet…

Overlay Fabric Allocations



VRF-VNID – allocated per VRF

- (unique within fabric)

BD-VNID – allocated per BD

- (unique within fabric)

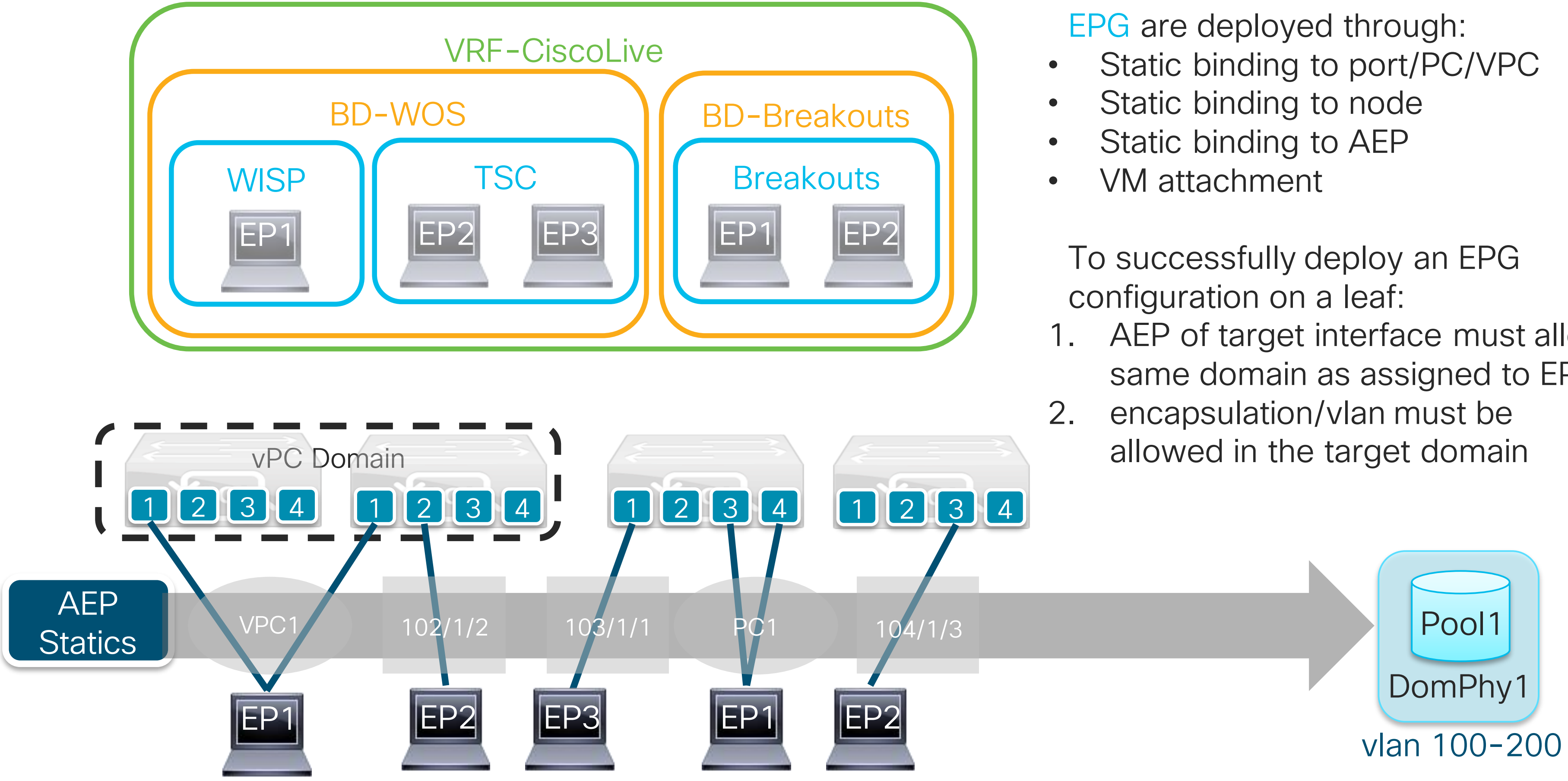
EPG-VNID – allocated from **vlan pool** (domain specific) and is unique within fabric

- Used for STP BPDU flooding and flood in encap for unknown unicast traffic

PCTAG – allocated per EPG

- FABRIC-global if shared service provider
- VRF-local otherwise

EPG Deployment to Leaf



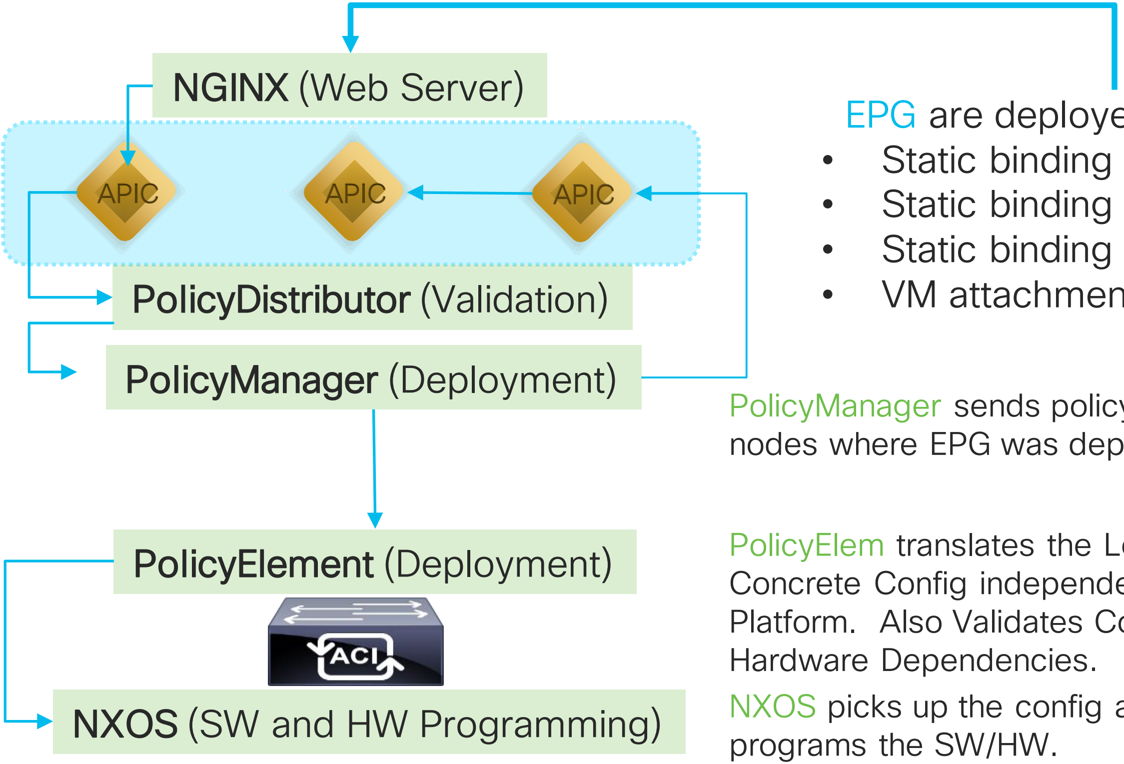
EPG are deployed through:

- Static binding to port/PC/VPC
- Static binding to node
- Static binding to AEP
- VM attachment

To successfully deploy an EPG configuration on a leaf:

1. AEP of target interface must allow same domain as assigned to EPG
2. encapsulation/vlan must be allowed in the target domain

ACI EPG Configuration Deployment



EPG are deployed through:

- Static binding to port/PC/VPC
- Static binding to node
- Static binding to AEP
- VM attachment

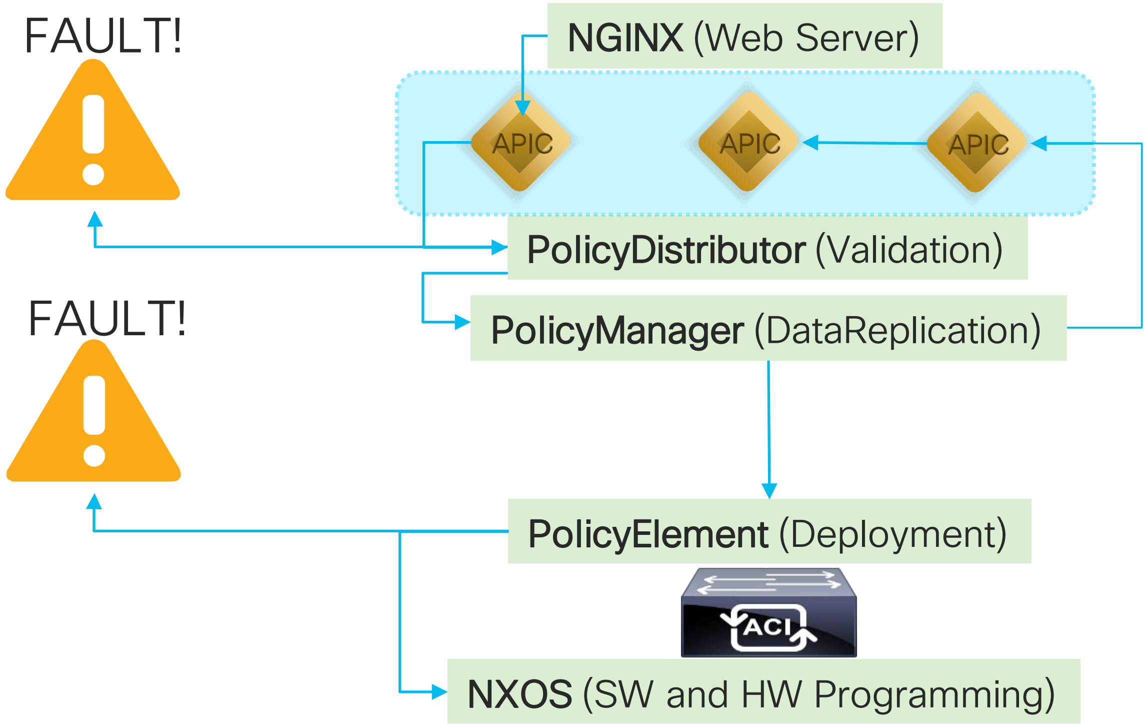
PolicyManager sends policy to appropriate nodes where EPG was deployed.

PolicyElem translates the Logical Config to Concrete Config independent of Hardware Platform. Also Validates Config against Hardware Dependencies.

NXOS picks up the config and programs the SW/HW.

ACI EPG Configuration Deployment

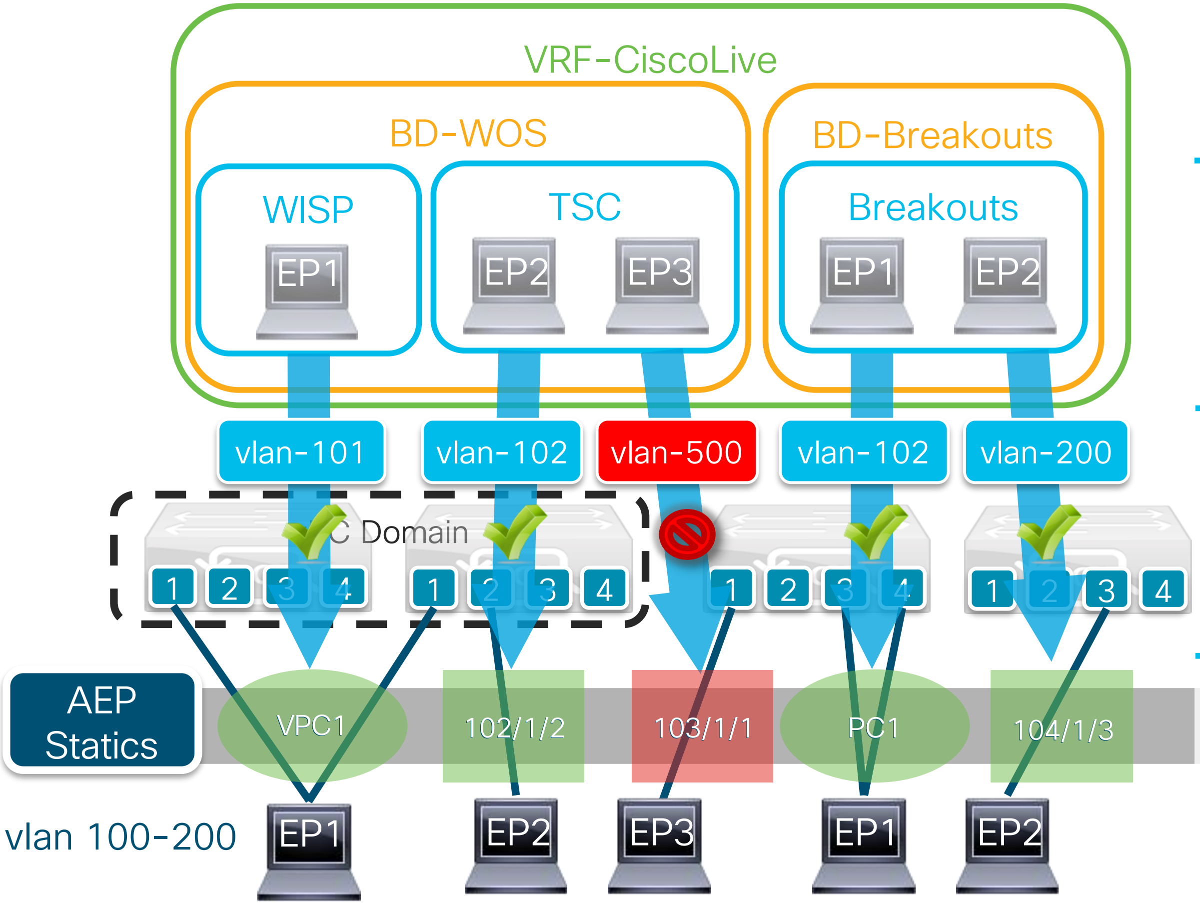
- Why is this Useful?



Logical Configuration Errors are Detected by APIC. Faults are raised.

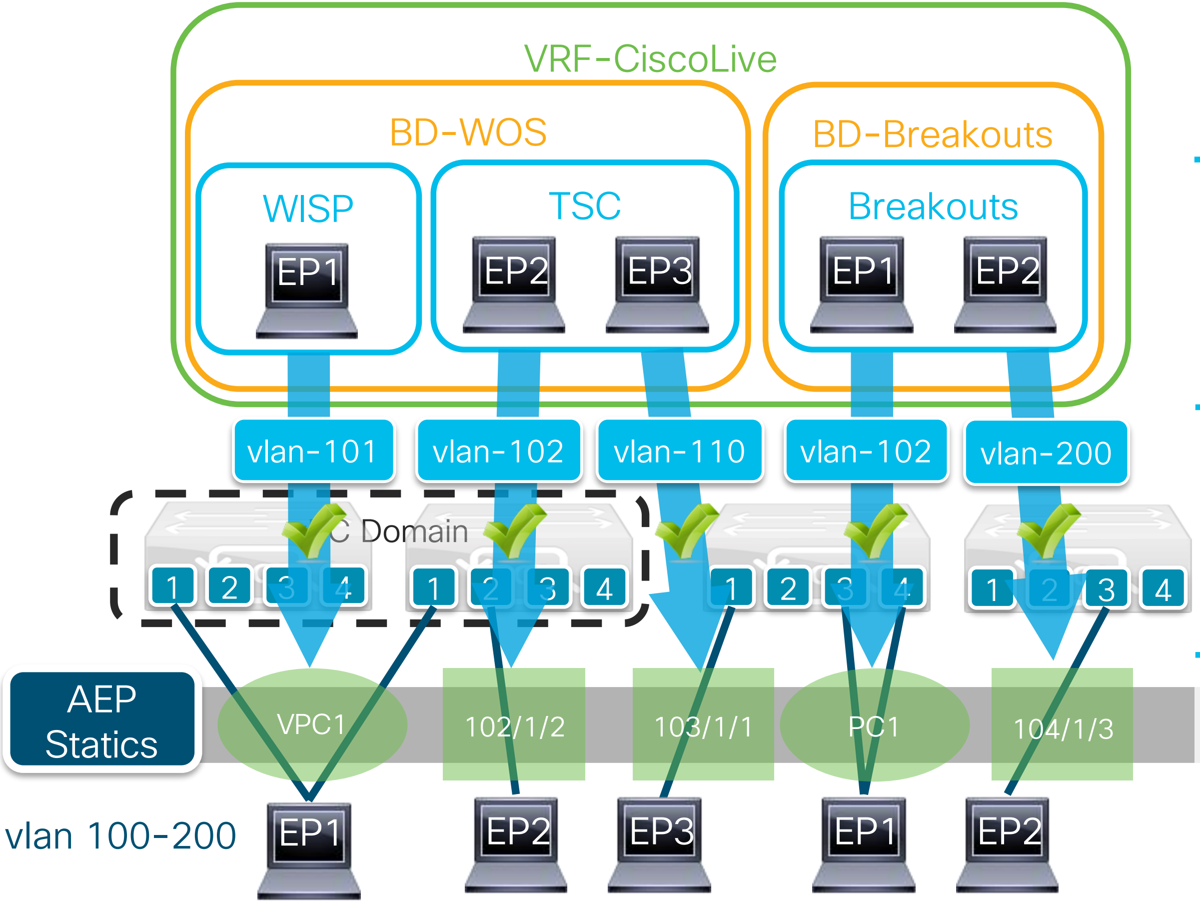
Platform Configuration Errors are Detected by Switch. Faults are raised.

EPG Static Path Deployment



Leaf101	BD-WOS	vlan-101
	VRF-CiscoLive	10.10.0.1/24
		10.20.0.1/24
Leaf102	BD-WOS	vlan-101
		vlan-102
	VRF-CiscoLive	10.10.0.1/24
		10.20.0.1/24
Leaf103	BD-Breakouts	vlan-102
	VRF-CiscoLive	10.10.0.1/24
Leaf104	BD-Breakouts	vlan-200
	VRF-CiscoLive	10.30.0.1/24

EPG Static Path Deployment



Leaf101	BD-WOS	vlan-101
	VRF-CiscoLive	10.10.0.1/24 10.20.0.1/24
Leaf102	BD-WOS	vlan-101 vlan-102
	VRF-CiscoLive	10.10.0.1/24 10.20.0.1/24
Leaf103	BD-Breakouts	vlan-102
	VRF-CiscoLive	10.30.0.1/24 10.10.0.1/24
	BD-WOS	vlan-110
Leaf104	BD-Breakouts	vlan-200
	VRF-CiscoLive	10.30.0.1/24

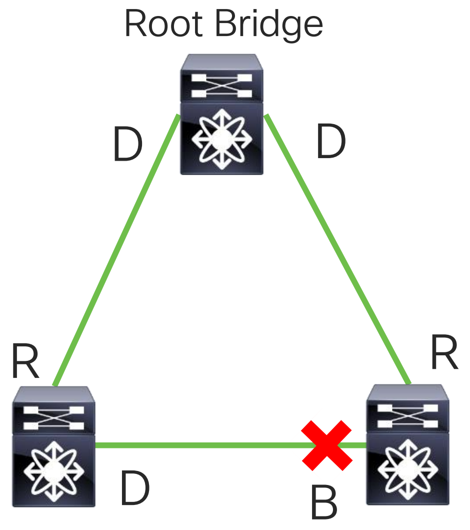
L2Outs and Loop Prevention

Spanning Tree

Classical behavior

- STP BPDUs (PVST or MST) are generated by each switch in the topology.
- STP root is elected and interface forwarding is calculated to prevent loops by blocking some interfaces.
  - All interfaces with best-path (highest bandwidth) towards root bridge will be forwarding.
  - Backup paths will be put in a blocking state by the switch with worst path towards root on the affected path (usually based on either the bridge identifier or port priority)
- Topology changes (TC) trigger MAC addresses to be flushed in received vlan, allowing traffic reconvergence based on new topology

Role	Description
R	Root port
D	Designated port
B	(Blk) Blocking port

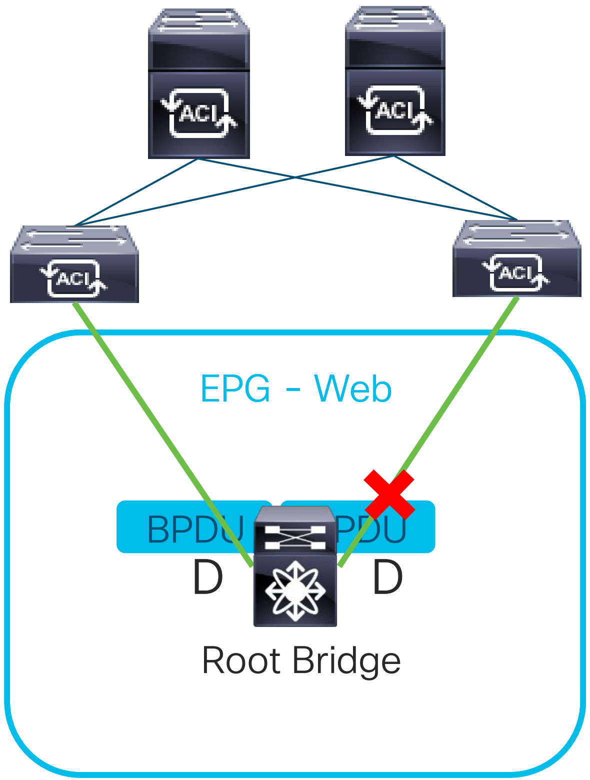


Spanning Tree

ACI floods BPDUs in the fabric encap

- ACI leaves don’t participate in spanning tree (generate BPDUs or block any ports)
- STP BPDUs (PVST or MST) are flooded within the fabric/EPG encap (allocated per vlan encap in a domain)
- Leaves flush endpoints in the EPG if a TC BPDU is received.
  - Spanning Tree Domain policy determines which EPGs to flush for MST domain TCs

NOTE: MST BPDUs are untagged and require an untagged/native EPG to be deployed on all interfaces connected to MST domain (this includes L3outs using SVIs)



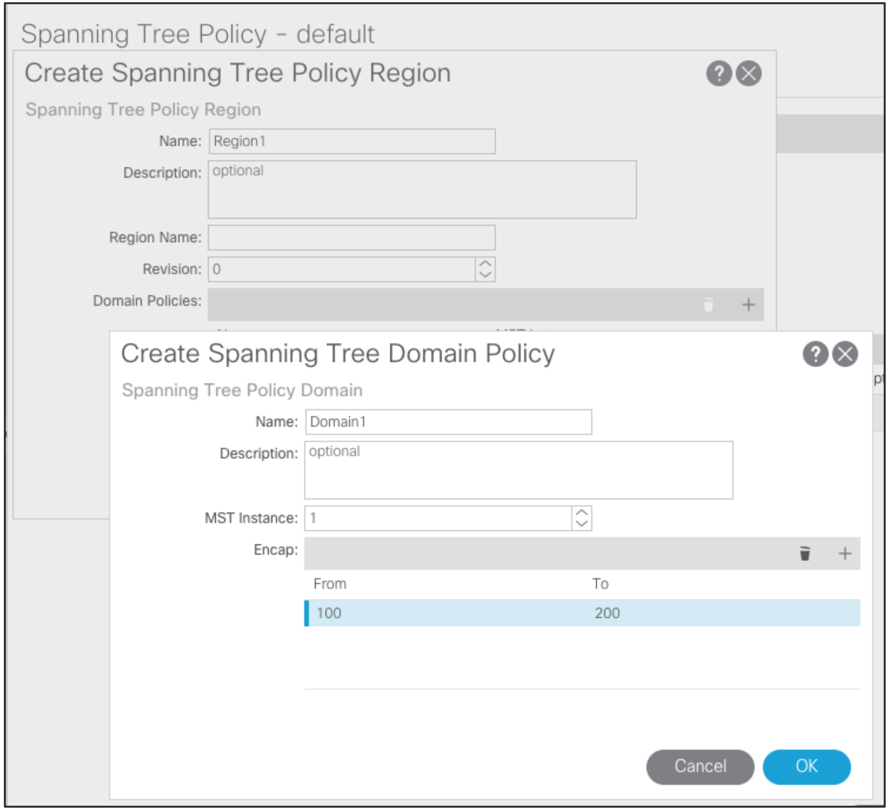
Spanning Tree Domain Policy

ACI MST Configuration

Configuration is fabric-wide and supports multiple regions for use within different tenants/domains.

Any ports connecting to MST switches within the same region MUST have untagged static-path.

Each MST region should have it's own EPG for BPDU flooding.



Fabric -> Access Policies -> Policies -> Switch -> Spanning Tree -> default

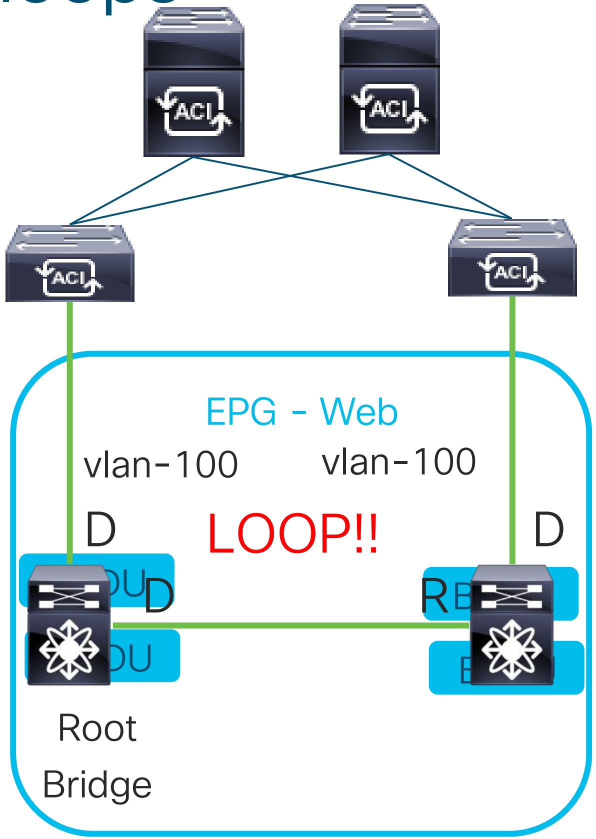
- Add a Region Policy
- Add a Domain Policy for each MST instance within the region (instance 0 is implicit)
    - Add vlan blocks

Common mistakes that cause loops

Missing untagged/native EPG in MST region

MST BPDUs are sent untagged by switches and will only be accepted by leaf if an EPG is deployed with an untagged/native EPG path binding.

All interfaces connected to a common MST region should have the same EPG deployed (this is to ensure BPDU is flooded to all of the MST switches connected to fabric).

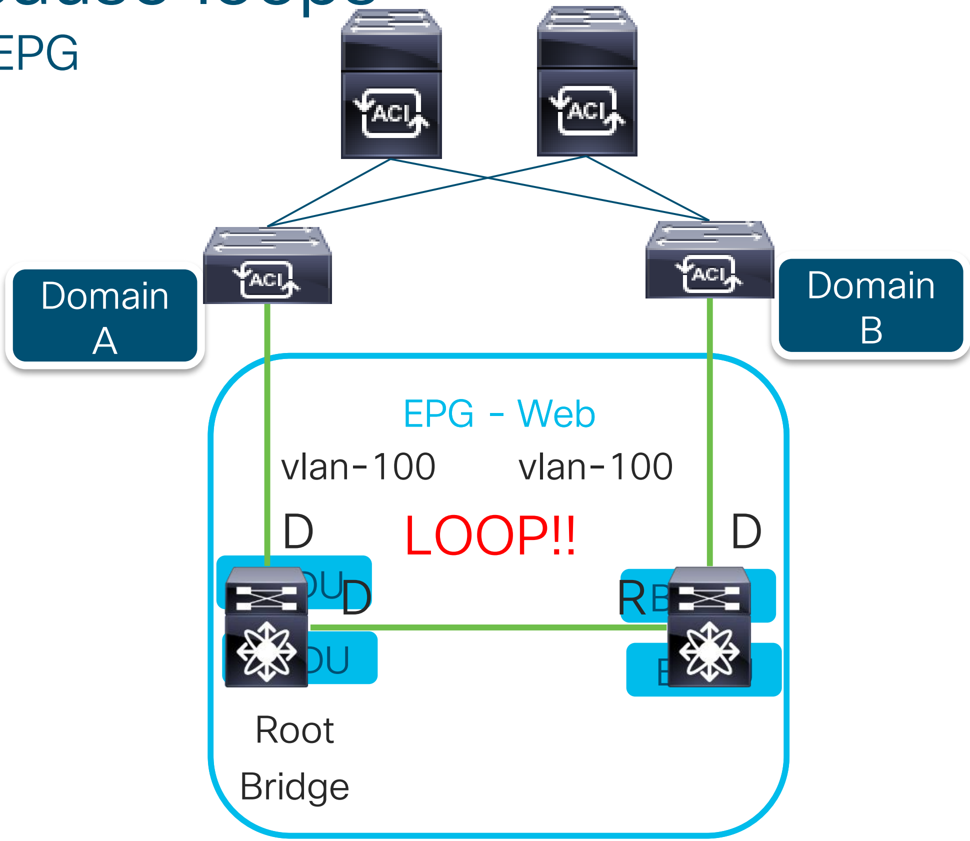


Common mistakes that cause loops

Multiple fabric encaps used for same EPG

BPDUs are flooded within the fabric encap of an EPG (allocated based on domain/vlan pool).

In order for BPDUs to be flooded properly, all interfaces within the EPG that are connected to external bridges MUST reside in the same physical or L2 external domain and vlan encapsulation.



Common STP Misconfiguration

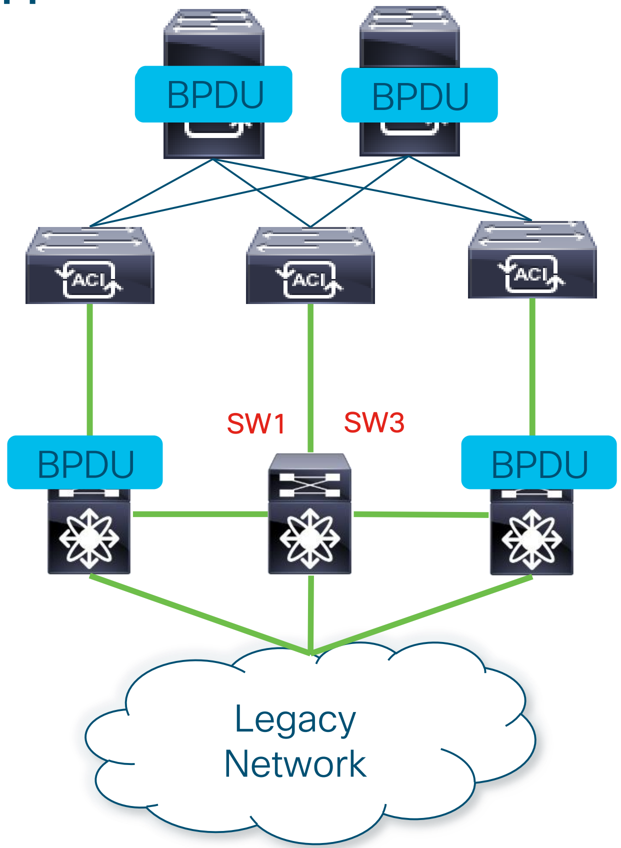
STP Link Type Must Be Shared

Since BPDU's are flooded, ACI acts as a HUB from an STP Perspective.

Full Duplex Links default to Spanning-Tree Link-Type PTP.

If multiple switches connect to ACI on separate links, Link-Type must be set to Shared to allow processing of multiple BPDU's on the same interface.

```
Root(config-if)#spanning-tree link-type shared
```



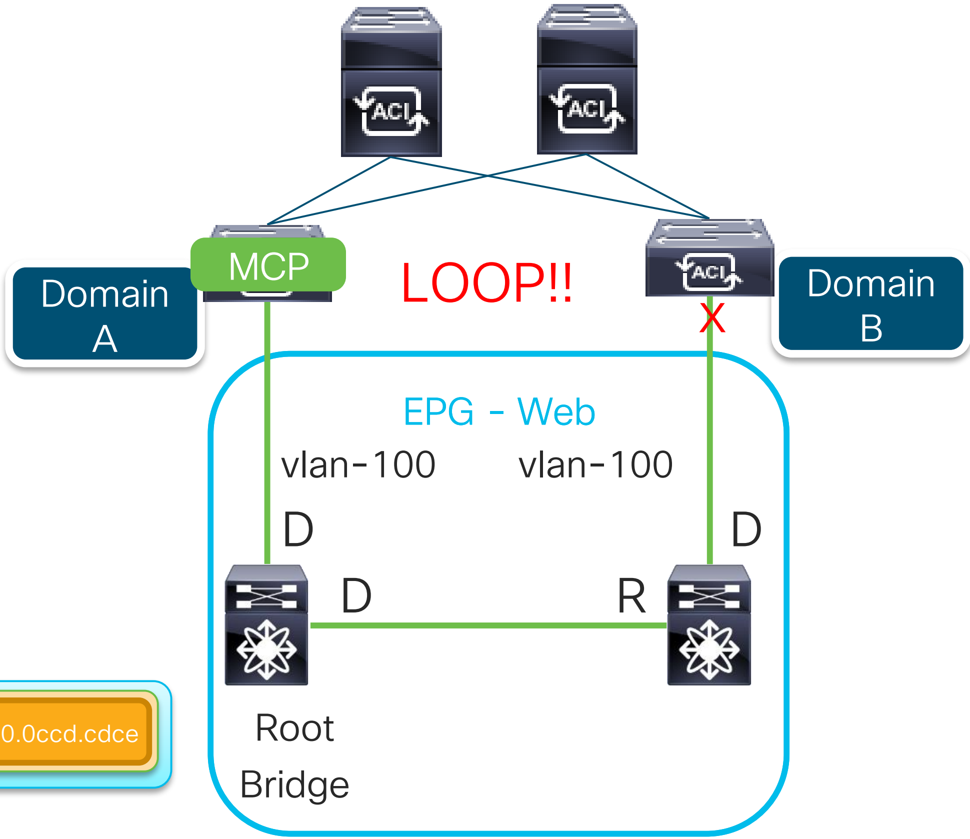
Loop Prevention - MCP
Mis-Cabling Protocol

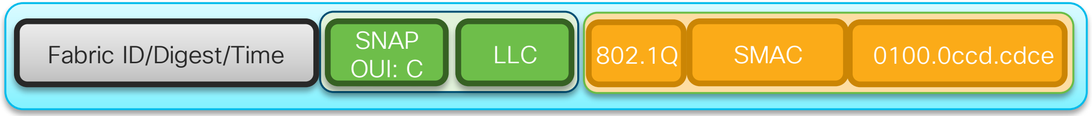
Mis-Cabling Protocol can be used to detect loops. With MCP, a special frame is sent out with a multicast destination MAC so that the downstream devices will flood it.

MCP Can be sent on a per VLAN basis.

If that frame is received back on a leaf in the fabric, it will err-disable the interface if ONE of the following conditions are met:

1. MD5 Digest is the same
2. Send time is within ~2s of receive time



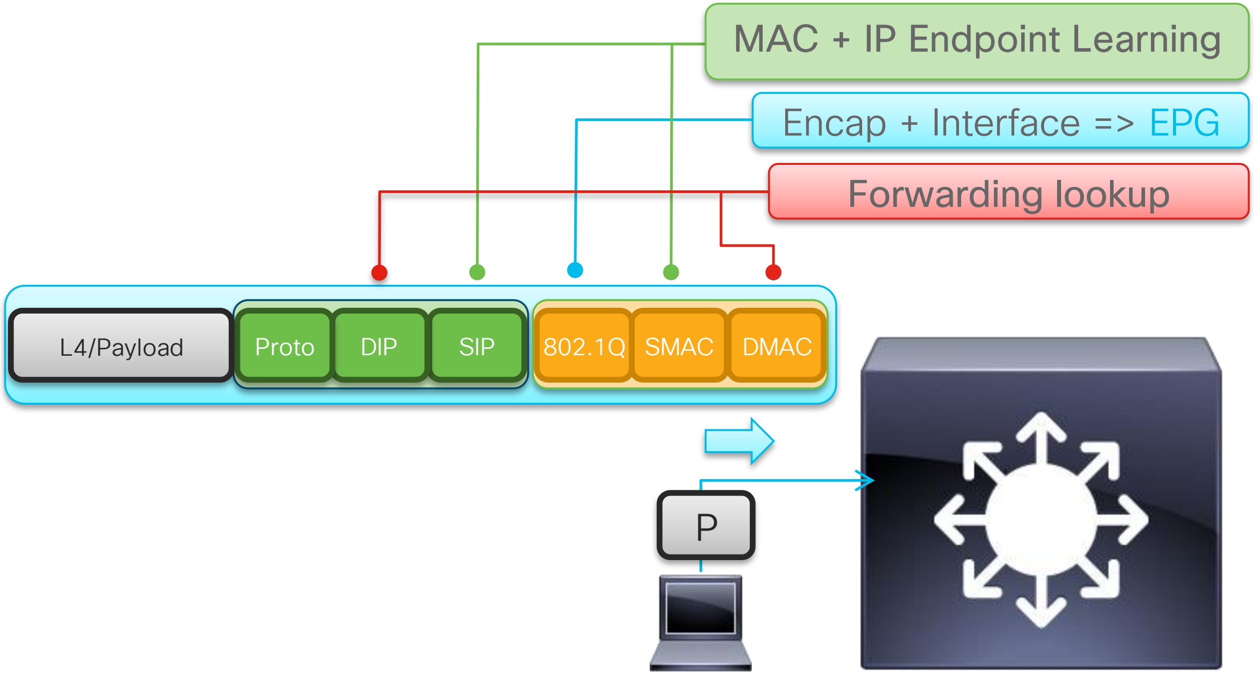


Agenda

- Introduction
- Building the Overlay
  - Access Policies
  - VRFs, Bridge Domains, and Endpoint Groups
  - L2Outs and Loop Prevention
- Traversing the Overlay
  - Learning, Forwarding, and Policy Enforcement
  - Shared Services and Route Leaking
  - L3outs and Routing Protocols
  - MultiPod

Learning, Forwarding, and Policy Enforcement

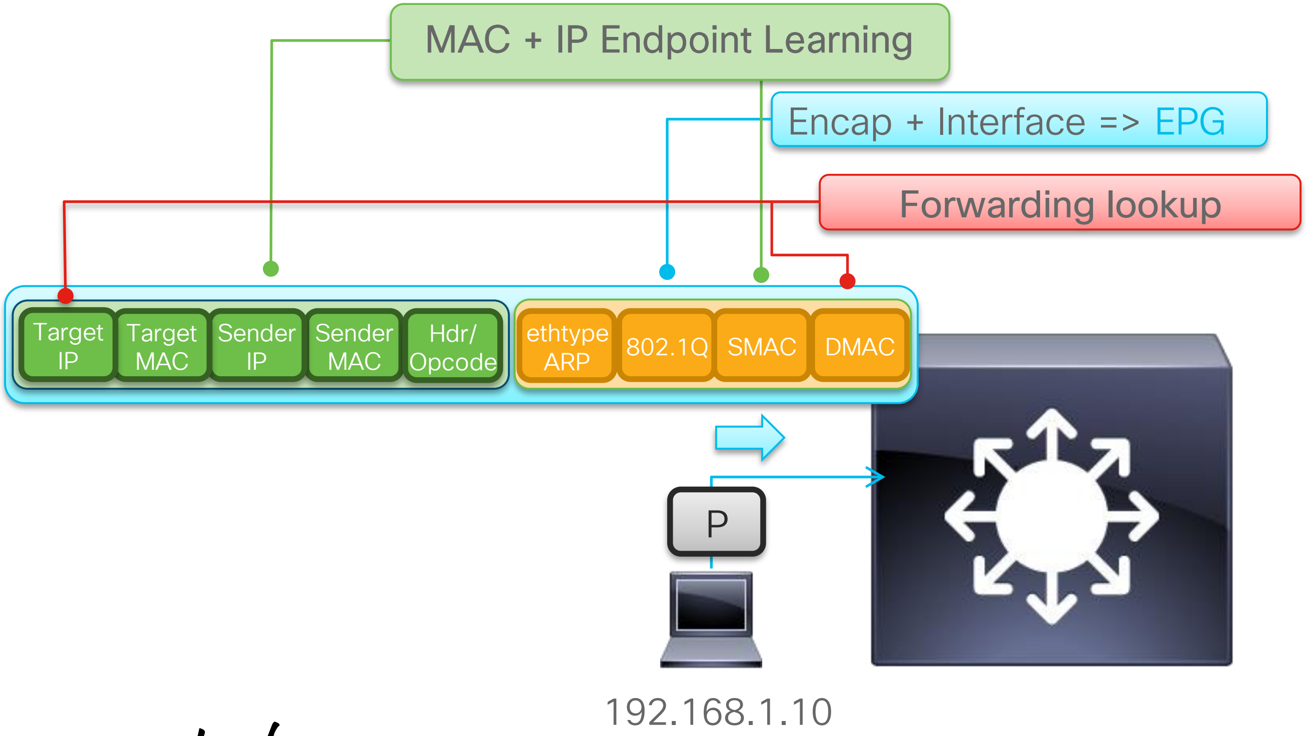
ACI Learning and Forwarding (MAC and IP)



192.168.1.10

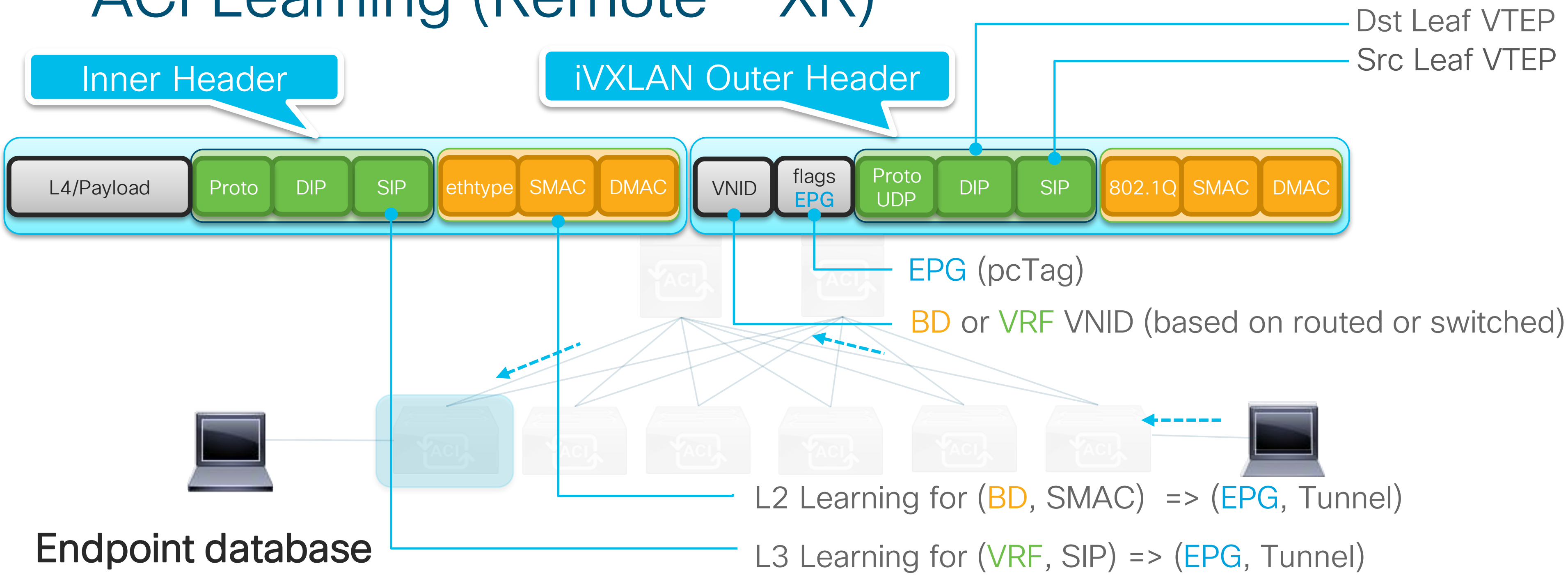
Packet flow	MAC	IP
Switched	Learned	X
Routed	**Learned**	**Learned**
L3Out	Learned	X

ACI Learning and Forwarding (ARP)



Packet flow	MAC	IP
Switched	learning	X
Routed	Learned	Learned
ARP	**Learned**	**Learned**
L3Out	Learned	X

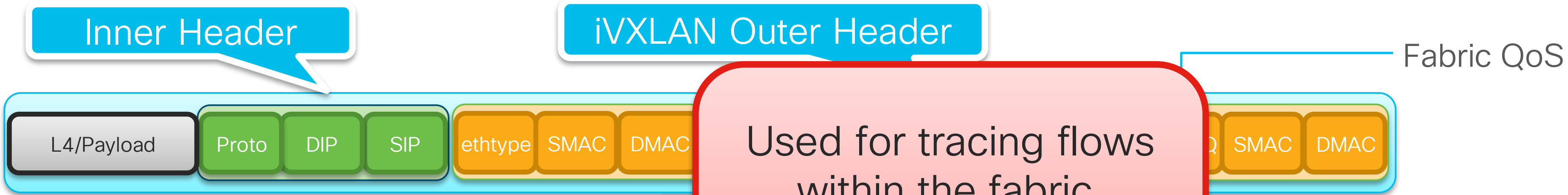
ACI Learning (Remote - XR)



Endpoint database

VLAN/ Domain	Encap	MAC/IP Address	Info	Interface	EPG
BD Name	BD VNID	SMAC		Tunnel oSIP	VXLAN Flags
VRF Name	VRF VNID	SIP		Tunnel oSIP	VXLAN Flags

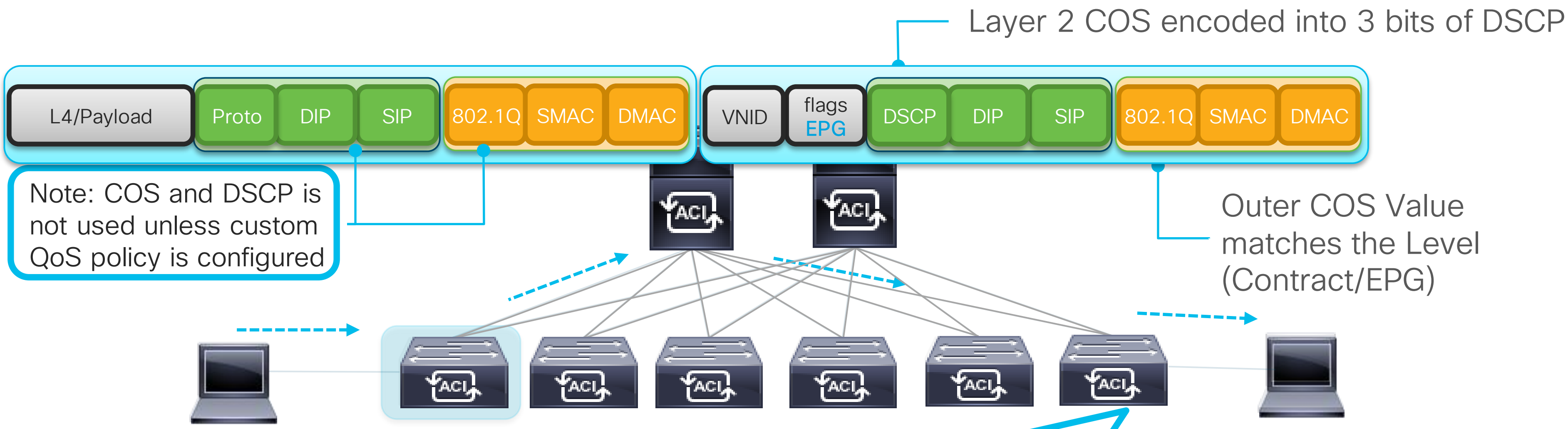
ACI Forwarding and QoS



Used for tracing flows within the fabric. Reserved for CPU generated traffic

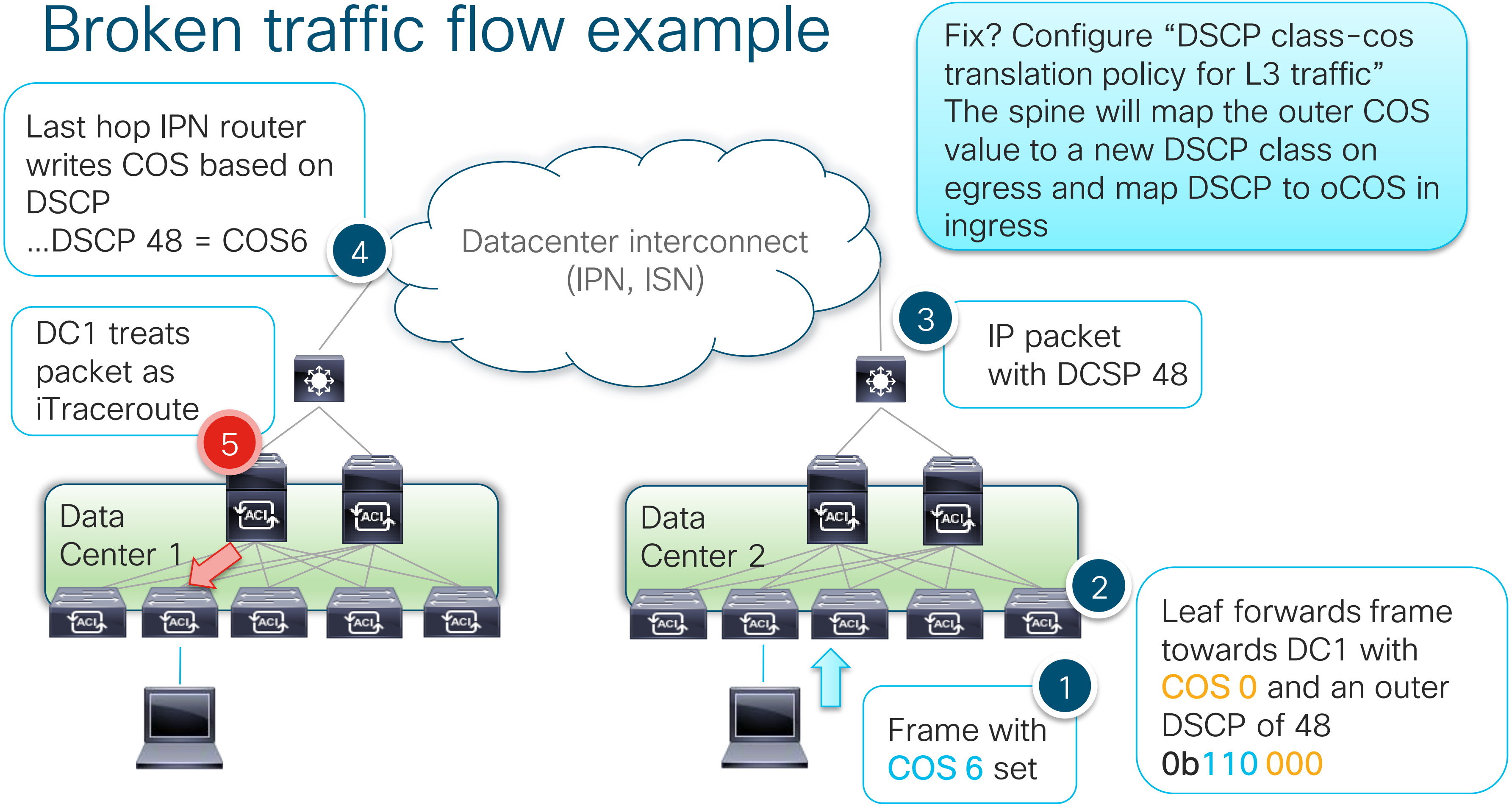
COS	Function	
3, 4, 5	APIC, SPAN, Con	
6	iTraceroute	Punted on Leaf
0 1 2	Level 1 Level 2 Level 3 (Default)	User Traffic 1 Priority
2 + DEI 3 + DEI 5 + DEI	Level 4 Level 5 Level 6	New in 4.0! User Traffic 5 Priority

ACI Forwarding and QoS – Preserve COS

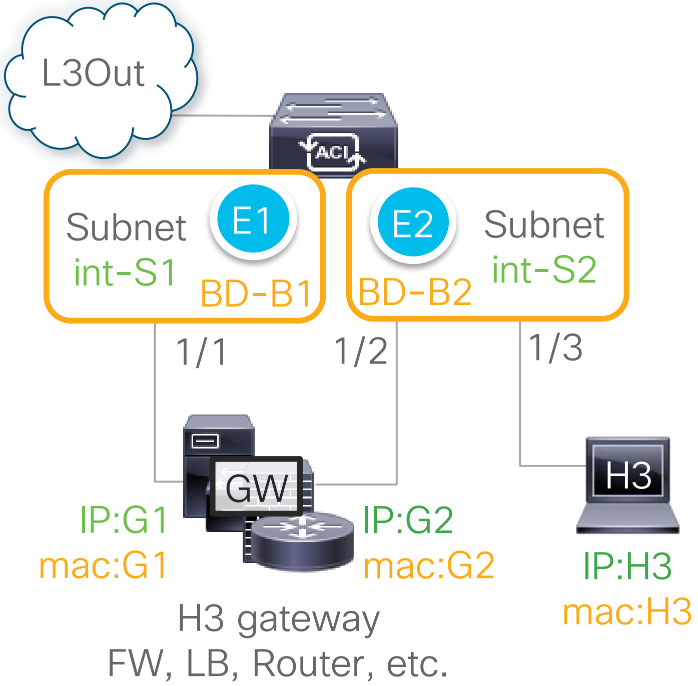


Configure **Dot1p Preserve**! The egress leaf will look at the 3 MSB bits of the DSCP value to know which COS value to use for packet rewrite

Broken traffic flow example



Broken Traffic Flow Example



- A Layer3 gateway device (**GW**) is connected to the fabric via a normal BD/EPG. Host **H3** is using GW as its gateway for a subset of traffic.
- The initial EP database show the IP's and MACs learned in the correct locations.

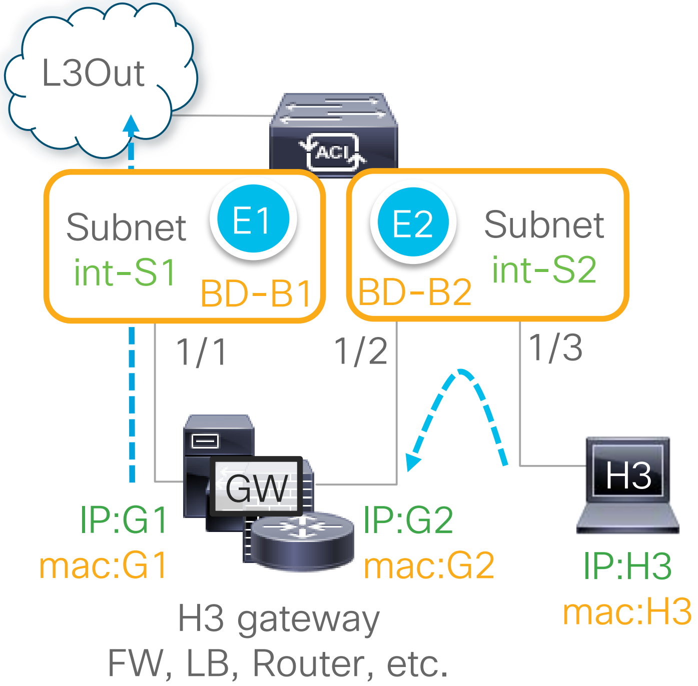
MAC EP Database

BD	MAC	EPG	Port
BD-B1	mac:G1	E1	1/1
BD-B2	mac:G2	E2	1/2
BD-B2	mac:H3	E2	1/3

IP EP Database

Vrf	IP	MAC	EPG	Port
v1	IP:G1	mac:G1	E1	1/1
v1	IP:G2	mac:G2	E2	1/2
v1	IP:H3	mac:H3	E2	1/3

Broken Traffic Flow Example



- H3 sends a frame to GW on BD-B2 (L2 switched through the fabric). GW routes the frame and sends it toward the fabric to be routed out.
- Fabric performs IP learning on routed traffic, IP:H3 moves to mac:G1 on EGP E1, port 1/1

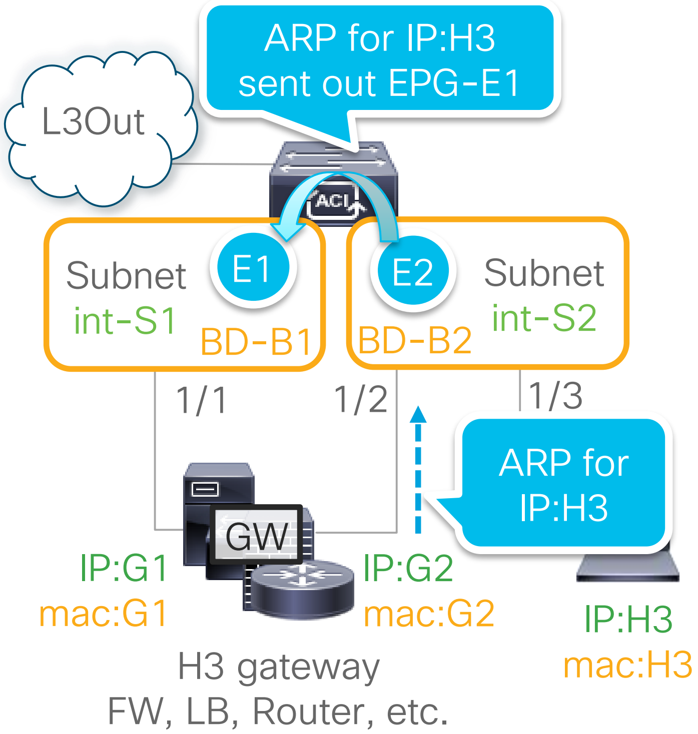
MAC EP Database

BD	MAC	EPG	Port
BD-B1	mac:G1	E1	1/1
BD-B2	mac:G2	E2	1/2
BD-B2	mac:H3	E2	1/3

IP EP Database

Vrf	IP	MAC	EPG	Port
v1	IP:G1	mac:G1	E1	1/1
v1	IP:G2	mac:G2	E2	1/2
v1	IP:H3	mac:G1	E1	1/1

Broken Traffic Flow Example



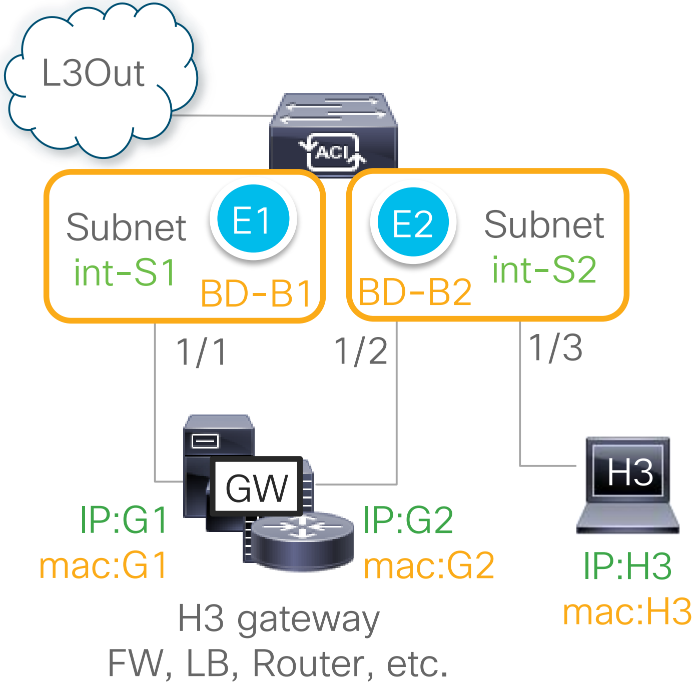
What's Broken?

- ARP to IP:H3 may fail since the IP is pointing to the wrong port
- Routed traffic to IP:H3 may be policy dropped since it's classified in EPG-E1 instead of EPG-E2
- IP:H3 may rapidly move within the fabric.

IP EP Database

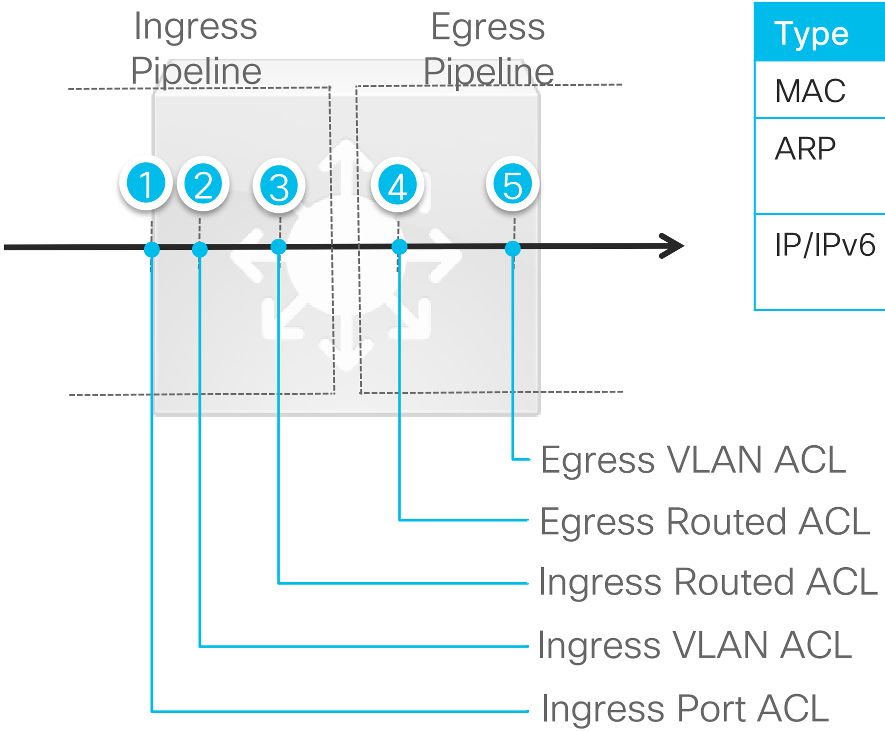
Vrf	IP	MAC	EPG	Port
v1	IP:G1	mac:G1	E1	1/1
v1	IP:G2	mac:G2	E2	1/2
v1	IP:H3	mac:G1	E1	1/1

Broken Traffic Flow Example



1. Connect devices that perform routing functionality to L3Outs.
2. **Disable unicast routing** on BD-B2 and enable ARP flooding so only MAC is examined when forwarding ARP instead of performing (VRF,IP) lookup on ARP target-IP
3. Enable NAT on routed device connected to internal BD. In this way, source IP address will be translated preventing fabric from learning IP address in wrong location
4. Disable **IP data-plane learning for VRF**
5. Enable **IP subnet prefix** check on BD-B1 or enable global subnet check. This will prevent learning of IP's outside of the subnets configured under the BD.

Classical Policy Enforcement

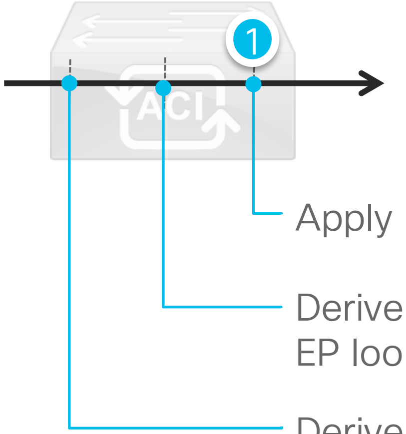


Type	Access Control Entry (ACE) Format
MAC	action src/mask dst/mask ethertype [PD filters]
ARP	action opcode srcIp/mask dstIp/mask srcMac/mask dstMac/mask [PD filters]
IP/IPv6	action protocol srcIp/mask srcPort/mask dstIp/mask dstPort/mask [PD filters]

- Multiple logical locations where ACLs can be applied depending on what type of traffic and what type of filters are needed (**very flexible**)
- ACE primarily based on src and dst values within frame (may be hard to maintain)
- ACLs often need to be configured and maintained on multiple devices in the network

ACI Policy Enforcement

Scope	Access Control Entry (ACE) Format
VRF	action src-EPG dst-EPG [filters]
VRF	permit any any (unenforced mode)



Apply Policy

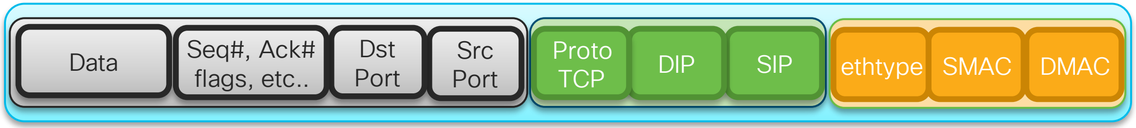
Derive destination EPG pcTag
EP lookup, IP Prefix

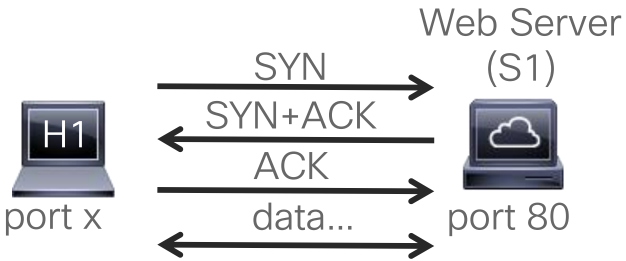
Derive source EPG pcTag
local EP, IP Prefix, or Encap

- Policy is created based on contract between EPGs with support for L2/L3/L4 filters similar to traditional ACLs.
- Leaf derives source EPG pcTag based on:
  - match in **EP database**
    src MAC for L2 traffic or src IP for L3 traffic
  - **longest-prefix match** against src IP
    (IP-based EPG or L3Out external EPG)
  - ingress **port + encap**
- Leaf derives destination EPG pcTag based on:
  - match in **EP database**
    dst MAC for L2 traffic or dst IP for L3 traffic
  - **longest-prefix match** against dst IP
    (L3Out external EPG or shared-services)
- Rules are programmed with scope of VRF. Policy lookup is always (VRF, src-EPG, dst-EPG, filter).
- Allow traffic between all EPGs without a contract by setting the VRF to **unenforced** mode

ACI Policy Enforcement

Reference TCP Packet



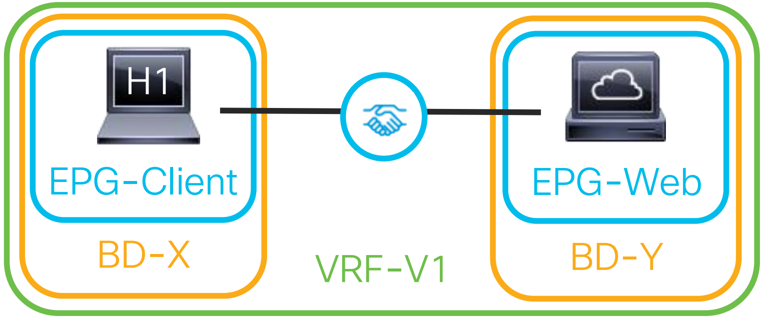


Classical Switch ACL

Generally applied at one or more L3 boundaries assuming H1 and S1 are in different subnets

```
ip access-list web
  permit tcp host H1 host S1 eq 80
  permit tcp host S1 eq 80 host H1
```

ACI Contract



EPG-Web is Providing a service on port 80

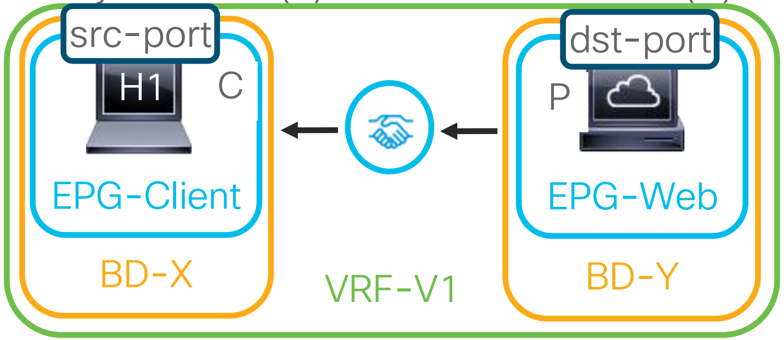
ACI Desired Behavior

Scope	Access Control Entry
VRF-V1	permit tcp EPG-Client EPG-Web eq 80
VRF-V1	permit tcp EPG-Web eq 80 EPG-Client

How do we get here?

ACI Policy Enforcement

- ❑ Identify Provider (P) EPG and Consumer (C) EPG



  - With a bidirectional contract, the 'provider' will be the dst-port filters and the 'consumer' will be the src-port filters (opposite of contract arrows)

- ❑ Create Filters

Name	EthType	Proto	Src Port	Dst Port
flt-1	IP	TCP	Any	80
flt-2	IP	TCP	80	Any

- ❑ Create a contract, subject, and filter(s). Apply to EPGs EGP-Web as provider and EPG-Client as consumer

Option 1 – Unidirectional filters
Apply both flt-1 and flt-2 to subject

flt-1 (C to P) and flt-2 (P to C)
permit tcp Consumer Provider eq 80
permit tcp Provider eq 80 Consumer

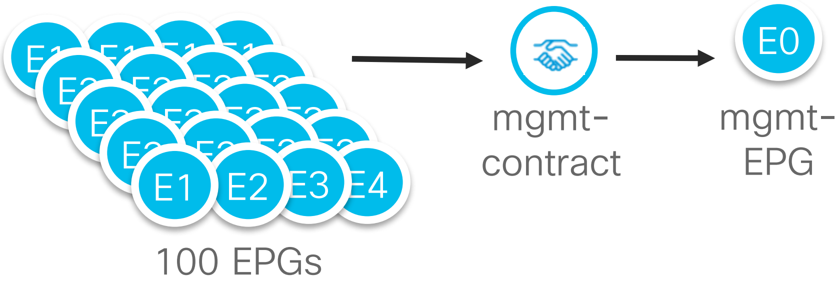
Option 2 – Bidirectional filters with reverse ports

flt-1 (C to P implied)
permit tcp Consumer Provider eq 80

flt-1 + apply both directions
permit tcp Consumer Provider eq 80
permit tcp Provider Consumer eq 80

flt-1 + apply both directions + reverse ports
permit tcp Consumer Provider eq 80
permit tcp Provider eq 80 Consumer

Only flt-1 needed!



Name	EthType	Proto	Src Port	Dst Port
flt-ssh	IP	TCP	1-65535	22
flt-snmp	IP	UDP	1-65535	161

- 100 EPGs all providing a basic management contract to a single consumer EPG.
- TCAM Utilization Calculation (Approximate)
  ~= (entries in contract)(# of Cons)(# of Providers)(2)
  ~= 2 * 1 * 100 * 2
  ~= 400 entries in hardware

- Policy CAM utilization increases by over 6400
  Why?

High Policy CAM Utilization Example

Name	EthType	Proto	Src Port	Dst Port
flt-ssh	IP	TCP	1-65535	22
flt-snmp	IP	UDP	1-65535	161

Expanded

permit tcp E1 eq 1 E0 eq 22
permit tcp E1 2-3 E0 eq 22
permit tcp E1 4-7 E0 eq 22
permit tcp E1 8-15 E0 eq 22
permit tcp E1 16-31 E0 eq 22
permit tcp E1 32-63 E0 eq 22
permit tcp E1 64-127 E0 eq 22
permit tcp E1 128-255 E0 eq 22
permit tcp E1 256-511 E0 eq 22
permit tcp E1 512-1023 E0 eq 22
permit tcp E1 1024-2047 E0 eq 22
permit tcp E1 2048-4095 E0 eq 22
permit tcp E1 4096-8191 E0 eq 22
permit tcp E1 8192-16383 E0 eq 22
permit tcp E1 16384-32767 E0 eq 22
permit tcp E1 32768-65535 E0 eq 22

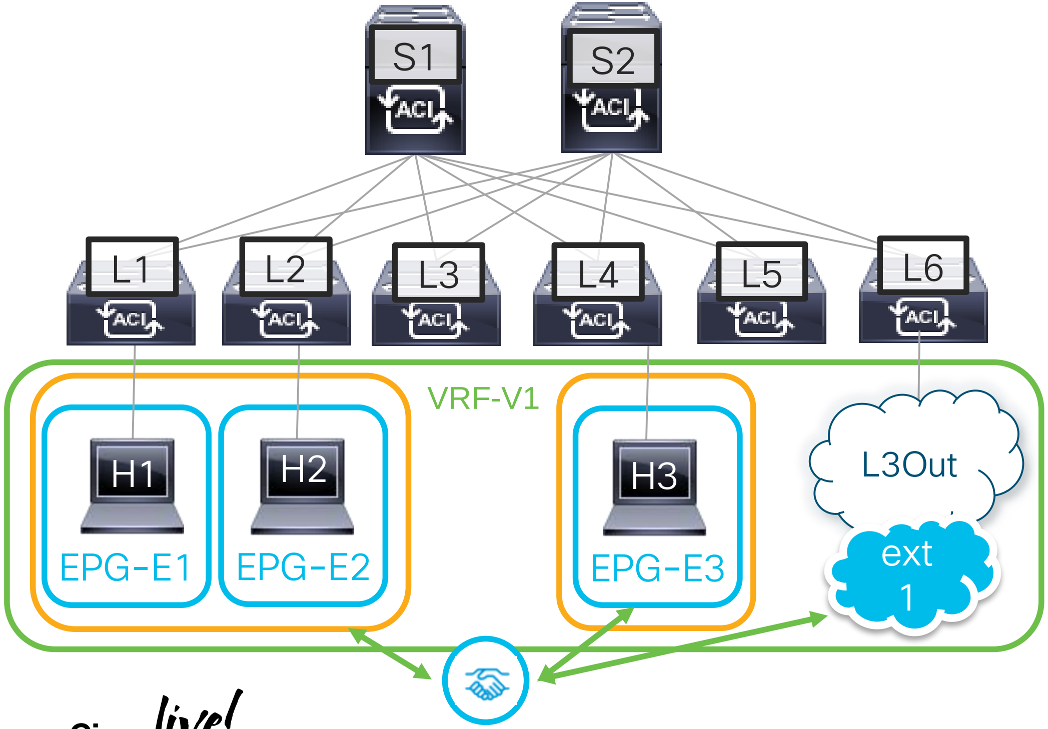
- Port Ranges
  Policy CAM, as with any TCAM, uses a value and mask to perform matching.
- Matching a single port utilizes only one entry in TCAM.
- Using a range of ports may need to be expanded to multiple entries in hardware depending on the start and end values.

How to fix this issue?

- Use port 0-65535 or 'unspecified' source port
  => utilization down from 6400 to 400 entries
- Consider using VzAny if all EPGs in the VRF need it
  => utilization down from 400 to 4 entries

ACI Preferred Group

Allow any any for a subset of EPGs

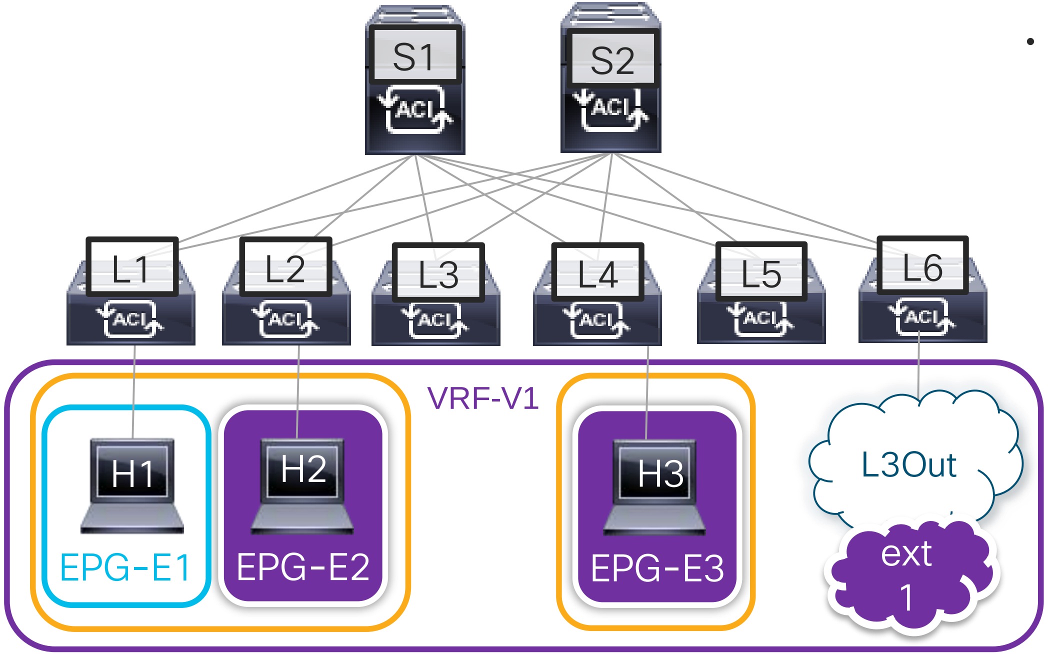


- EPGs that are part of the preferred group do not require contracts to communicate with each other
- EPGs and External EPGs can be configured to included or excluded from the preferred group
  - EPGs which are excluded, have hardware rules programmed to prevent communication to EPGs which are included

Contract	VRF	Action	Src	Dst	Filter
C1	V1	permit	E2	E3	all
	V1	permit	E3	E2	all
	V1	permit	E2	ext1	all
	V1	permit	E3	ext1	all
	V1	permit	ext1	E2	all
implicit	V1	deny	any	any	all

ACI Preferred Groups

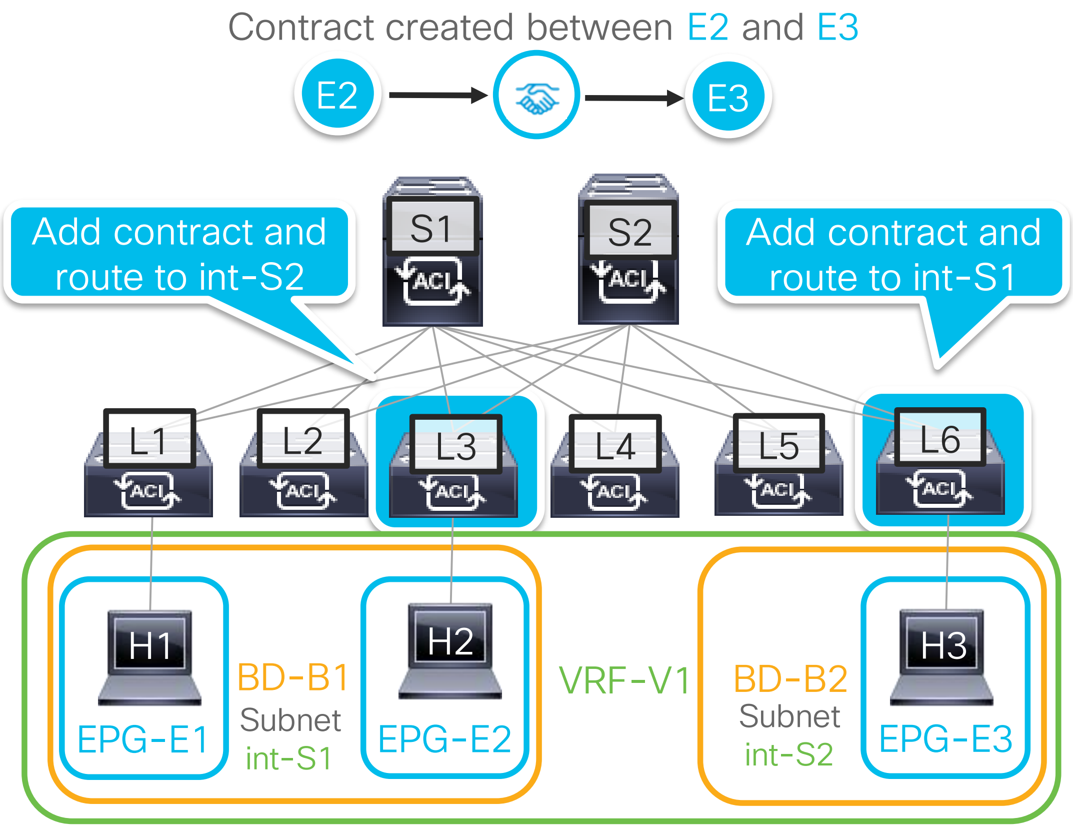
Allow any any for a subset of EPGs



- **Only recommended if the majority of EPGs require unenforced policy**
- Deny rules are installed for EPGs outside of the preferred groups
- Contracts can still be used to enable communication between excluded and included EPGs

Contract	VRF	Action	Src	Dst	Filter
implicit	V1	deny	any	E1	all
	V1	deny	E1	any	all
implicit	V1	permit	any	any	all

ACI Contracts and Resource Utilization



- BD-B1 and BD-B2 each have a subnet defined. Subnet int-S1 on BD-B1 exists on L1 and L3, while subnet int-S2 for BD-B2 exists on L6

When creating the contract between E2 and E3:

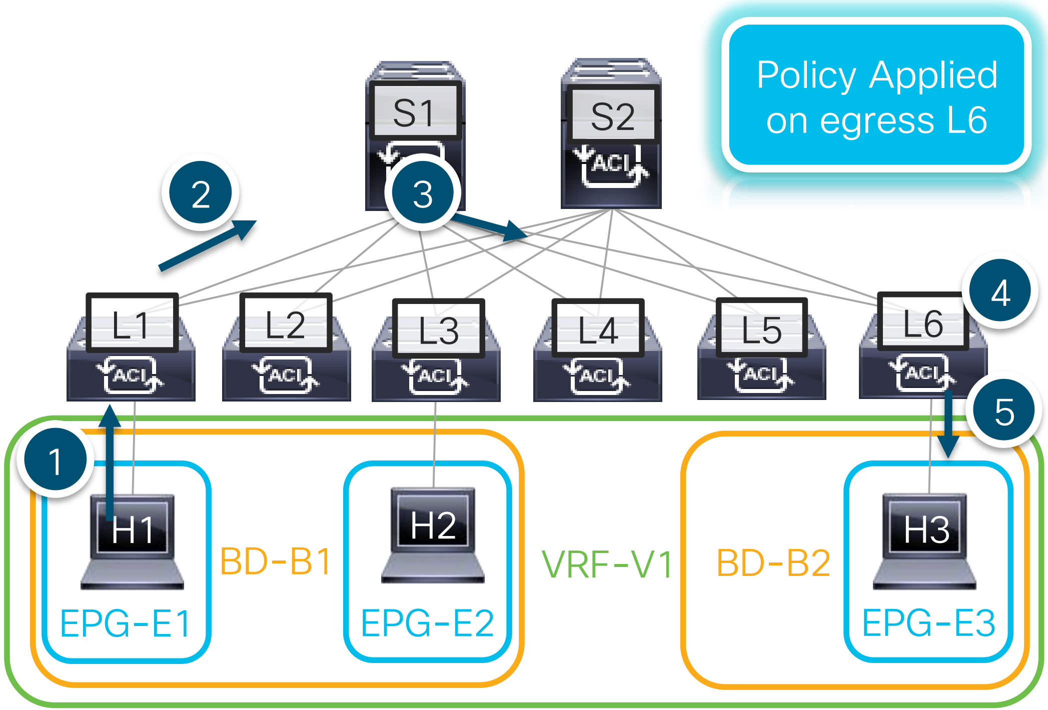
- Program contract rule between E2 and E3 in TCAM. Add **Static route** for int-S1 created on **L6** pointing to spine proxy.
- Program contract rule between E2 and E3 in TCAM. Add **Static route** for int-S2 created on **L3** pointing to spine proxy.

- Contracts are only programmed on leafs that have provider/consumer EPGs. BD routes are only programmed on leafs that need them!

Contracts contribute to both **policy** AND **routing** entries on leafs!

ACI Policy Enforcement

Unknown Layer3 Unicast



ARP has resolved on hosts for ACI GW
L1 has not learned H3 from L6

1. H1 sends layer3 unicast frame to H3 (destination MAC of BD-B1).

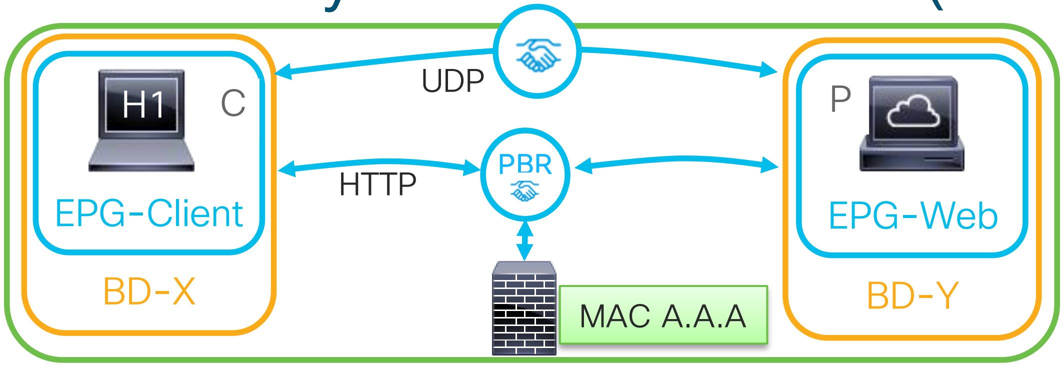
2. L1 performs layer3 lookup on H3 destination IP and pervasive route pointing to the Spine Proxy.
   L1 does not set policy applied bits - frame is sent to Proxy TEP with **EPG-E1** ***(PCTag)*** and **VRF-V1** set in VXLAN header.

3. Spine receives frame and preforms proxy lookup. Frame is sent to L6.

4. L6 does layer3 lookup on H3 destination IP in VRF-V1. Hit in local EP database and derives destination **EPG-E3** ***(PCTag)***. Policy check is enforced

5. L6 forwards traffic to H3 with appropriate encap if permitted by contract

ACI Policy Based Redirect (PBR)



- Contract can now redirect traffic to service device (FW, LB etc) for inspection prior to allowing

Name	EthType	Proto	Src Port	Dst Port	Action
flt-1	IP	TCP	Any	80	Redirect (Grp 1)
flt-2	IP	UDP	Any	Any	Permit

Name	Dest MAC	Dest BD	Tunnel Int
Redir-Grp1	A.A.A	ServiceBD	Mac Proxy

1) Create L4-L7 Device
   Define Interface, VLAN, etc.
2) Create redirect policy
   Contains the MAC & IP of service Device
3) Create Graph Template & check Redirect
4) Apply Graph template between two EPGs
   Creates redirect contract
   Can be reused with different EPGs

Shared Services and Route Leaking

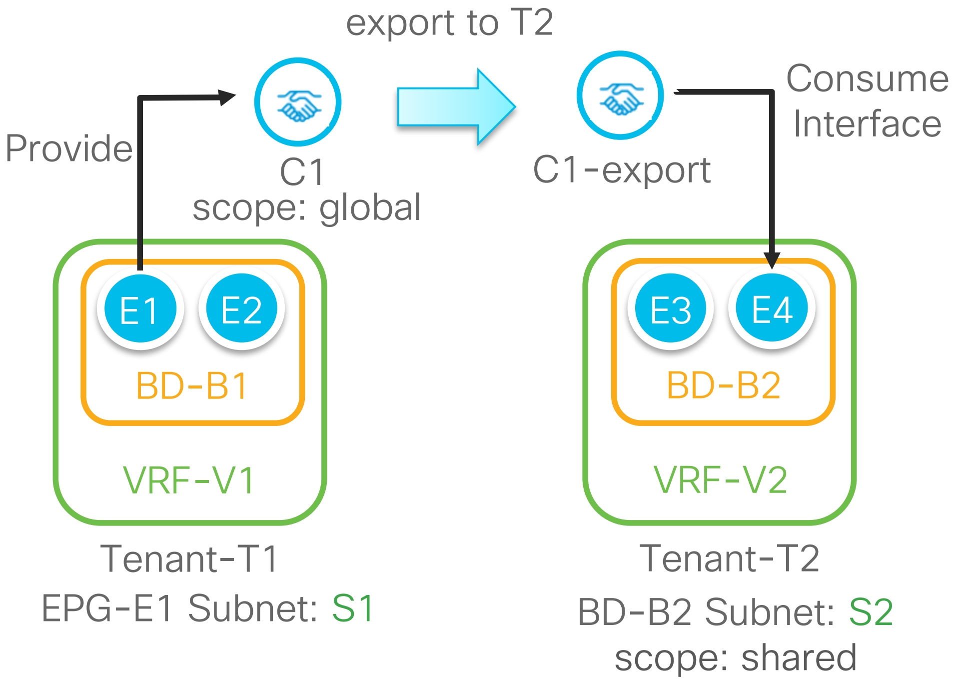
ACI Shared Services

- What is a shared service?
- Shared Service (**Route Leaking**) enables traffic between endpoints in different VRFs.
- A shared service EPG provider is an EPG that provides a contract consumed by an EPG in a **different** VRF

Restrictions

- **Provider Subnet** must be defined under the **provider EPG**
- Both provider and consumer subnets must have scope set to **shared**
- contract needs correct scope
- VzAny not supported as provider

Scope:
- ☐ Private to VRF
- ☐ Advertise Externally
- ✔ Share Between VRFs



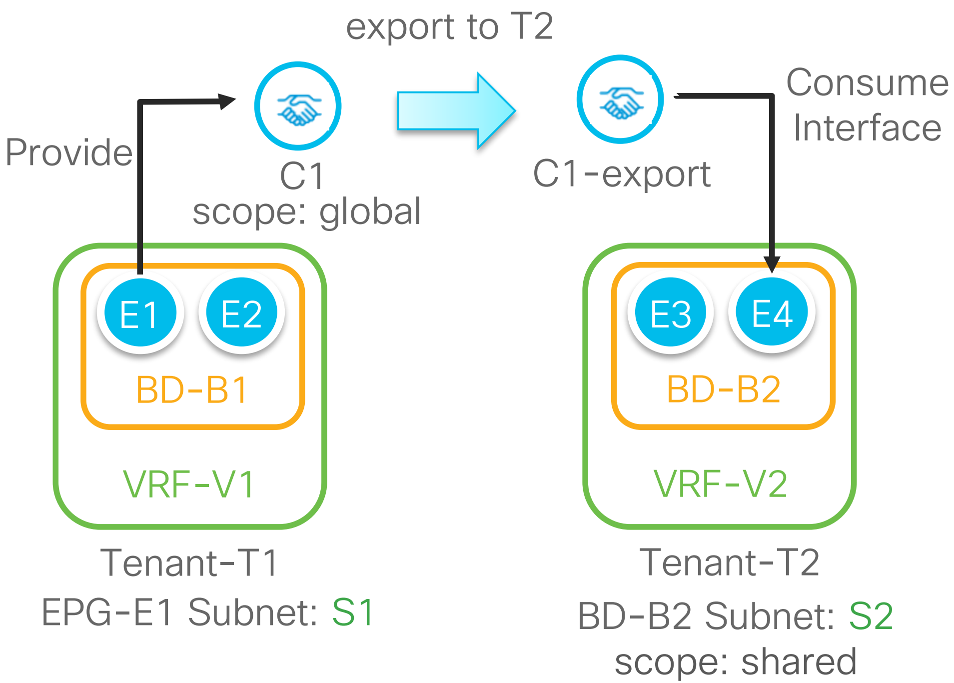
VRF	Route	pcTag	Flags
V1	S1	1	proxy
V2	S2	1	proxy

VRF	EPG	pcTag
V1	E1	49155
V1	E2	49156
V2	E3	16387
V2	E4	49155

ACI Shared Services

- What happens in the fabric?

- EPG-E1 is now a shared service provider. It is reallocated a fabric unique pcTag (<16384)
- All subnets on **consumer BD** programmed in **provider VRF**
- **Provider subnet** programmed in **consumer VRF** with pcTag of provider EPG

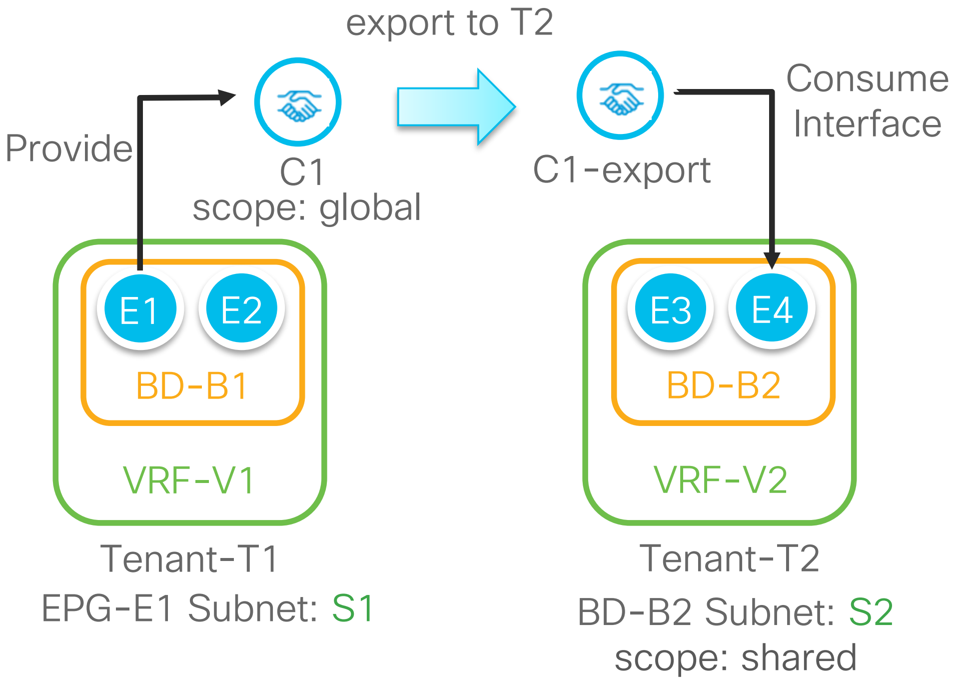


VRF	Route	pcTag	Flags
V1	S1	1	proxy
V1	S2	**1**	proxy, **rewrite VNID(V2)**
V2	S2	1	proxy
V2	S1	17	proxy, **rewrite VNID(V1)**

VRF	EPG	pcTag
V1	E1	17
V1	E2	49156
V2	E3	16387
V2	E4	49155

ACI Shared Services

- What happens in the fabric?



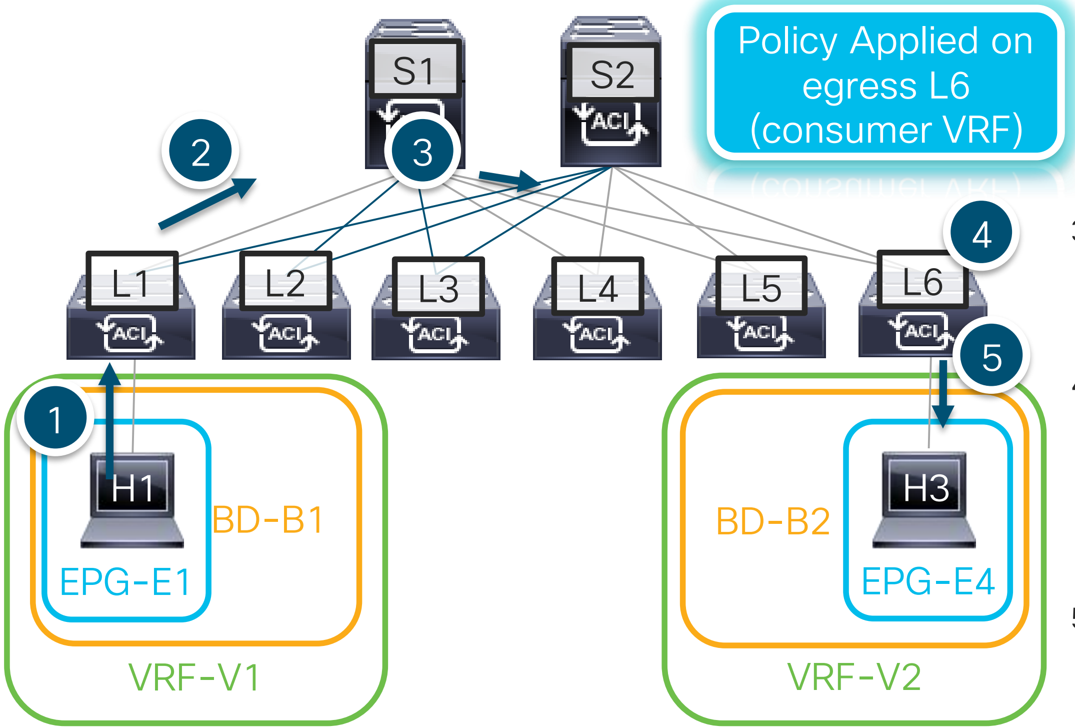
- EPG-E1 is now a shared service provider. It is reallocated a fabric unique pcTag (<16384)
- All subnets on **consumer BD** programmed in **provider VRF**
- **Provider subnet** programmed in **consumer VRF** with pcTag of provider EPG
- Policy enforcement always performed in **consumer** VRF. Therefore, **contracts** are always **programmed** in **consumer** VRF.

Contract	VRF	Action	Src	Dst	Filter
C1	V2	permit	E4	E1	flt1
	V2	permit	E1	E4	*flt1
	V1	-	-	-	-

No Rule added in provider VRF

Shared Service Forwarding

- From Provider E1 to Consumer E4



1. H1 sends packet toward gateway in EPG-E1 with destination IP of H3

2. L1 performs layer3 lookup for H3 in VRF-V1 and hits LPM entry for H3 subnet. LPM entry points to proxy with **VNID rewrite** info for **VRF-V2**. Packet is sent to Spine Anycast IPv4 Proxy VTEP with VRF-V2 VNID EPG-E1 set in VXLAN header. **No policy applied** in provider VRF

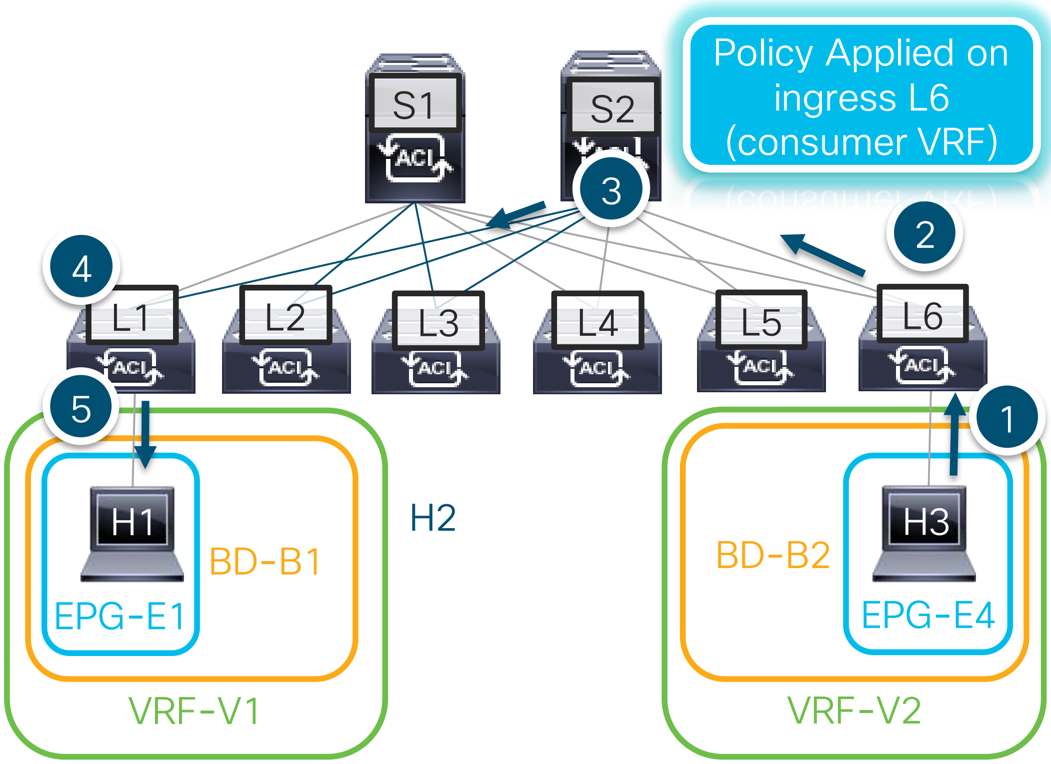
3. Spine performs proxy lookup for H3 IP in VRF-V2. Normal Proxy behavior to forward packet to VTEP of L6

4. L6 performs layer3 lookup on H3 destination IP in VRF-V2. Hit in local EP database and derives destination EPG-E4 L6 **applies policy** between EPG-E1 and EPG-E6

5. If permitted, traffic forwarded to H3 with appropriate encap

Shared Service Forwarding

- From Consumer E4 to Provider E1



1. H3 sends packet toward gateway in EPG-E4 with destination IP of H1
2. L6 performs layer3 lookup for H1 in VRF-V2 and hits LPM entry for H1 subnet. LPM entry points to proxy with **VNID rewrite** info for **VRF-V1** and **pcTag** of EPG-E1.
   L6 **applies policy** between EPG-E4 and EPG-E1 in consumer VRF-V2.
   If permitted, packet is sent to Spine Anycast IPv4 Proxy VTEP with VRF-V1 VNID and EPG-E4 set in VXLAN
3. Spine performs proxy lookup for H1 IP in VRF-V1. If unknown drops the packet. Else forward to VTEP of L1
4. L1 performs layer3 lookup on H1 destination IP in VRF-V1. Hit in local EP database and derives destination EPG-E1 Policy already applied by L6
5. Traffic is forwarded to H1 with appropriate encap

Contract Review

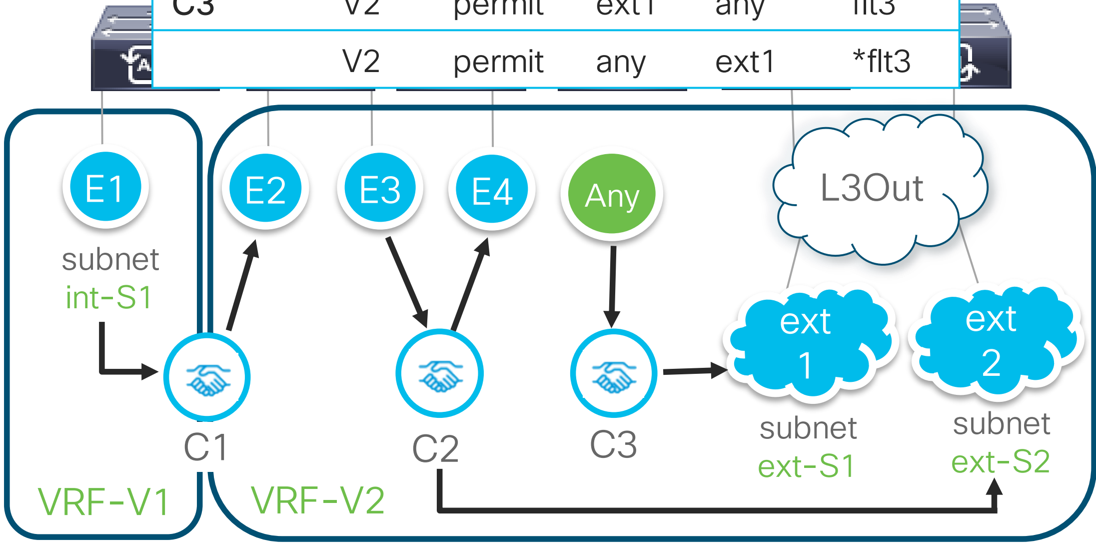
- Shared Service EPGs
  EPGs that provide contract consumed by EPG in a different VRF: E1, E2*
- Application EPGs
  E1, E2, E3, E4
- External EPGs
  configured on L3Out and classified based on IP prefix: ext1, ext2
- VzAny
  Represents all EPGs in a single VRF: Any

Contract Assumptions for this Example :

- All contract subjects have both directions and reverse filters enabled.

Policy TCAM

Contract	VRF	Action	Src	Dst	Filter
C1	V2	permit	E2	E1	flt1
	V2	permit	E1	E2	*flt1
C2	V2	permit	E4	E3	flt2
	V2	permit	E3	E4	*flt2
	V2	permit	ext2	E3	flt2
	V2	permit	E3	ext2	*flt2
C3	V2	permit	ext1	any	flt3
	V2	permit	any	ext1	*flt3



L3outs and Routing Protocols

Basic Connectivity

Layer3 Out: L3Out-1
VRF: VRF-V1
Layer-3 Domain: DomL3

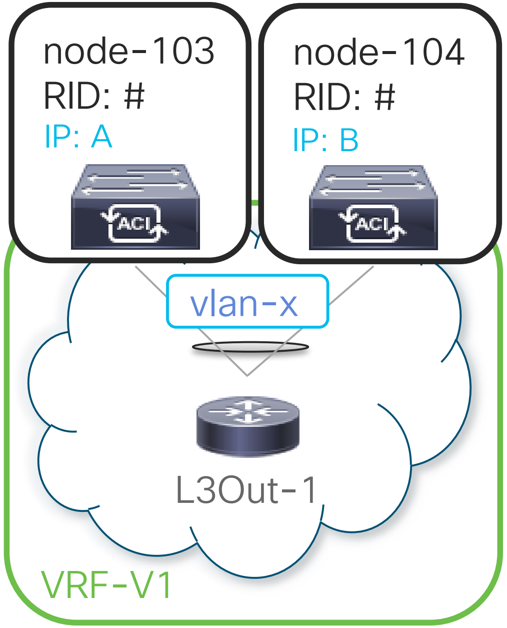
Logical Node Profile: node-103-104

node: node-103
Router-ID: #

node: node-104
Router-ID: #

Logical Interface Profile: ipv4-lif

path: topology/pod1/...vpcX
type: ext-svi, encap: vlan-x
IP-A, IP-B, MTU, MAC



Create the L3Out

- Associate VRF and L3 Domain
- Create Logical Node Profile and associate fabric nodes to the L3Out.
- Create Logical Interface Profile
- Specify Path attributes containing physical interface, encapsulation, and IPs

Configuring Routing Protocols

Layer3 Out: L3Out-1
VRF: VRF-V1
Layer-3 Domain: DomL3

Logical Node Profile: node-103-104

node: node-103
Router-ID: #

node: node-104
Router-ID: #

Logical Interface Profile: ipv4-lif

path: topology/pod1/...vpcX
type: ext-svi, encap: vlan-x
IP-A, IP-B, MTU, MAC

Enable Protocol
- BGP (fabric ASN configured in fabric pod policy)
- OSPF - area, area-type, area-configuration
- EIGRP - ASN

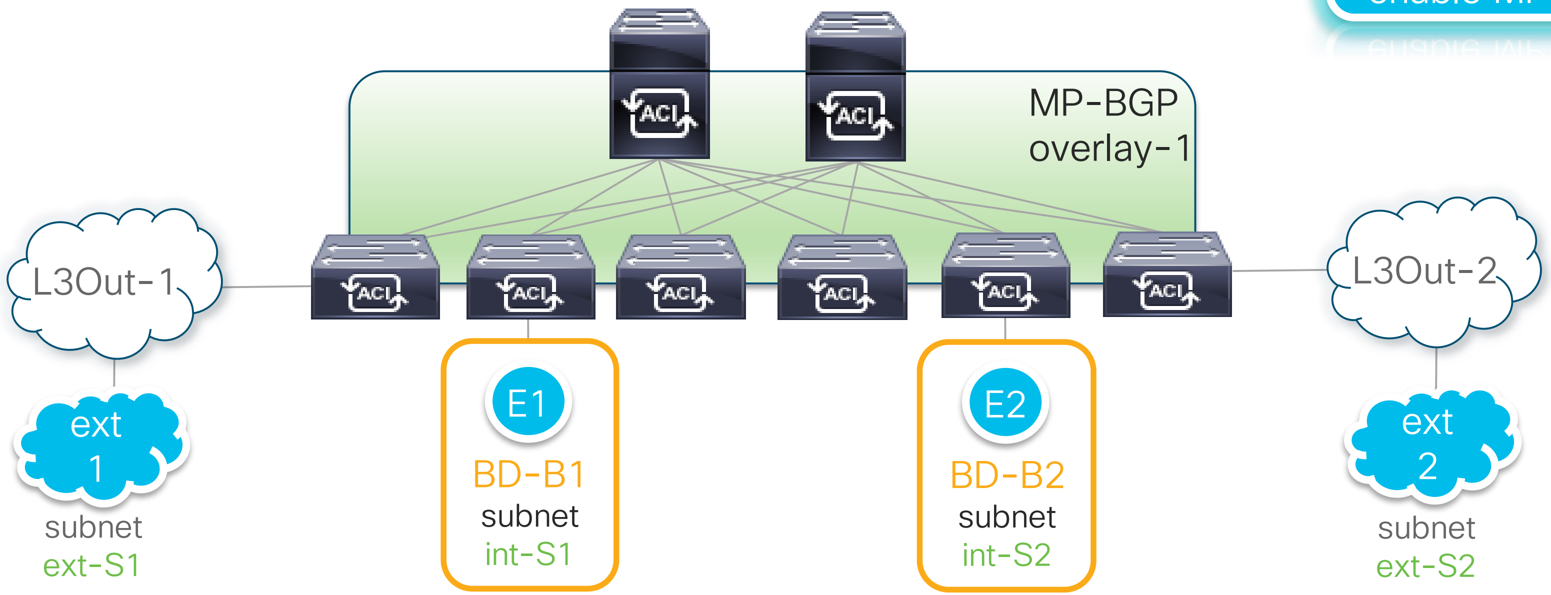
Static Routes configured per node
BGP Peer Profiles tied to Loopback

BGP Peer Profiles tied to Interface
OSPF Interface Policy and Authentication
EIGRP Interface Policy
BFD Interface Policy

Majority of configuration under the interface/peer policies...

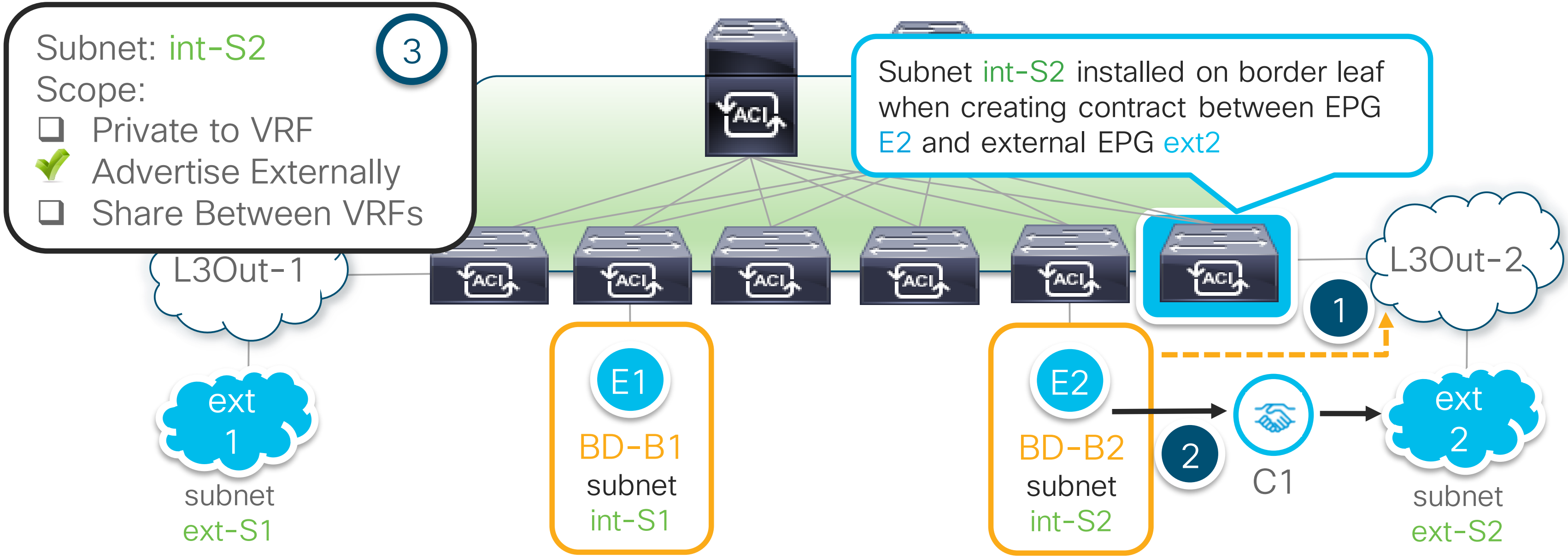
Types of Fabric Routes

Ensure BGP RR is configured to enable MP-BGP



- Internal Routes: Subnets defined under create static pervasive routes within the fabric.
- External Routes: Routes learned via a routing protocol or static routes configured under an L3Out. These routes are redistributed into MP-BGP and advertise to all leafs that contain the VRF
- Transit Routes – Routes advertised between L3Outs.

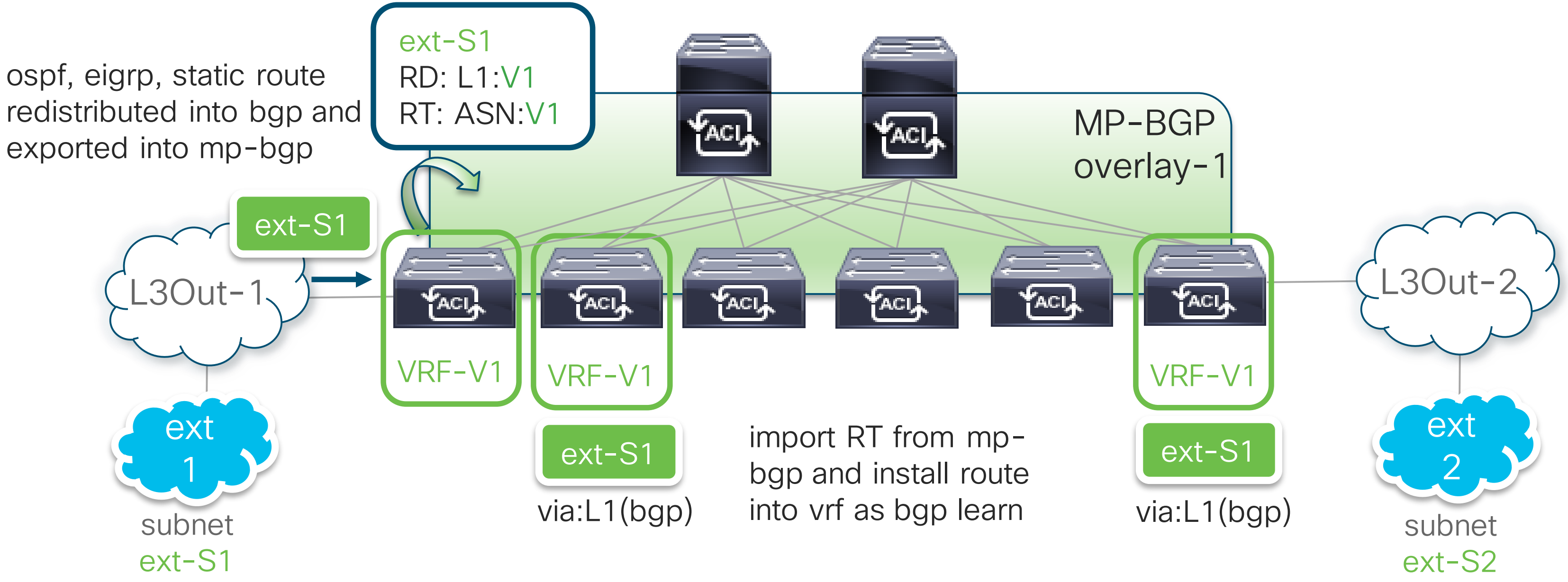
Types of Fabric Routes – Internal Routes



There are three requirements to advertise Internal Routes out an L3Out:

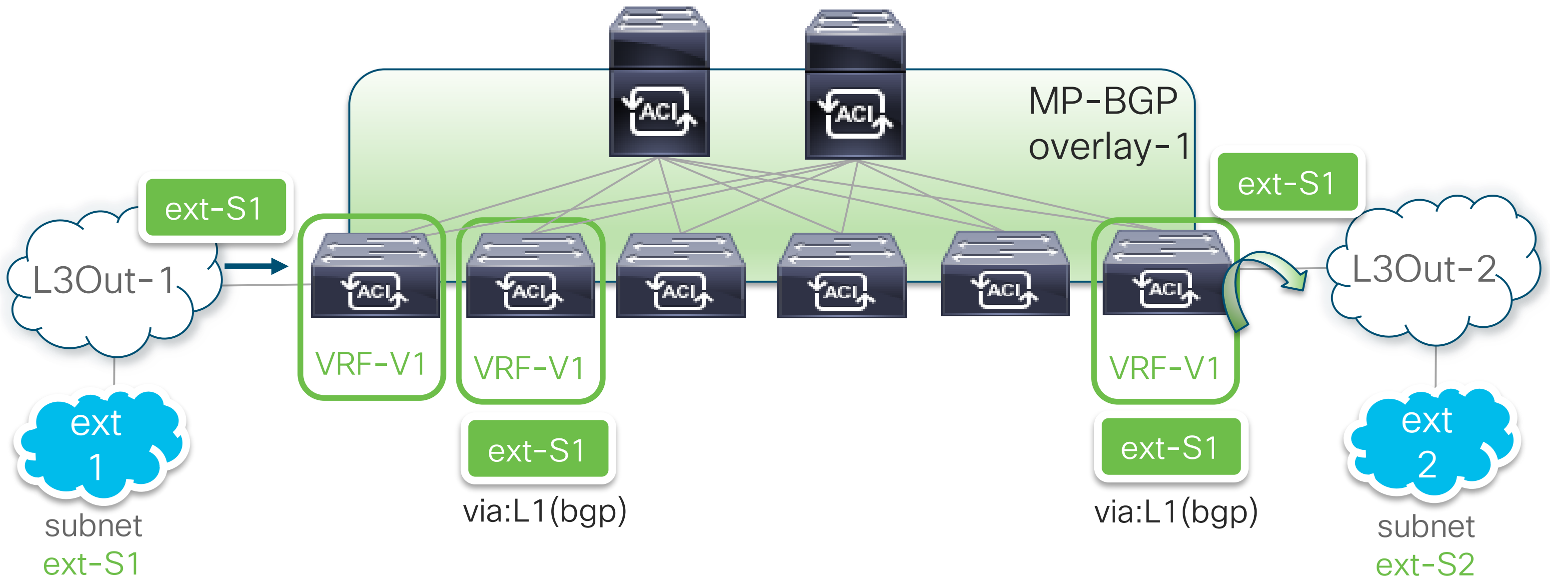
1. The BD must be associated with the L3Out*
   The association adds prefix entry to route map controlling advertised routes
2. A contract must exists between an EPG within the BD and an external EPG on the L3Out.
   The contract creates internal BD route on border leaf (cannot advertise route until it exists locally)
3. The subnet must have a public scope (Advertise Externally)

Types of Fabric Routes – External Routes



- External Routes from ospf, eigrp, or static are redistributed on the border leaf into the local bgp process.
- The bgp route is exported into MP-BGP with a route-target (RT) of the corresponding VRF. Each leaf in the fabric with the VRF present will import the RT and install the route. External routes on the non-originating border leaf will be seen as bgp learned routes.
- External Routes are controlled via Import Route Control flag

Types of Fabric Routes – Transit Routes



- In this example, external route ext-S1 is a Transit Route when advertised out L3Out-2.
- If OSPF or EIGRP on L3Out-2, ext-S1 is redistributed from BGP into the IGP and advertised.
- Transit Routes are controlled via Export Route Control flag

Configure L3Out External Network

Define an External Network, ext1 in this example

- Note: At least one external network required to bring up L3Out interfaces on border leaf

- Add Subnet to External Network

Prefix-based EPG for Contracts:

- External Subnets for the External EPG
- Shared Security Import

Route Control

- Export Route Control
- Import Route Control
- Shared Route Control
- Aggregate Export
- Aggregate Import
- Aggregate Shared Routes

Categorize options

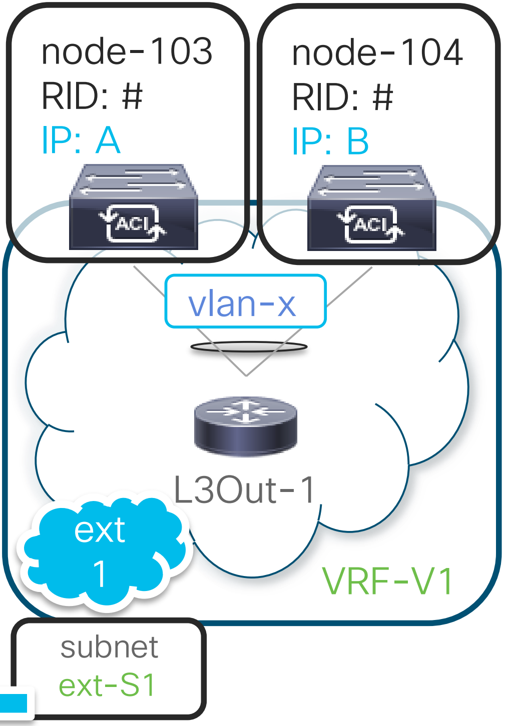
Scope:

❑ Export Route Control Subnet
❑ Import Route Control Subnet
❑ External Subnets for the External EPG
❑ Shared Route Control Subnet
❑ Shared Security Import Subnet

Aggregate:

❑ Aggregate Export
❑ Aggregate Import
❑ Aggregate Shared Routes

Subnet options



External Subnets for the External EPG

Previously: Import-Security

External Subnet for the External EPG is used to classify dataplane packets into external EPG for policy enforcement.

- An **IP prefix** is installed into leaf TCAM to **classify** traffic to/from the external network and assign correct pcTag for policy enforcement

Subnet: ext-S2/mask
Scope:
✔ External Subnets for the External EPG

EPG to pcTag

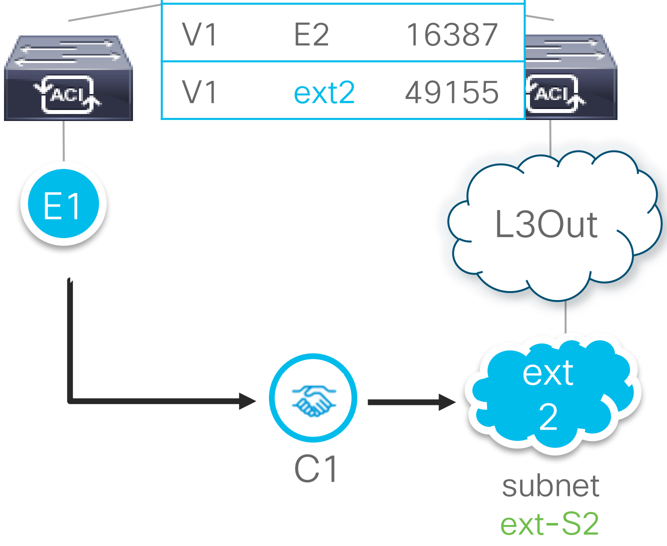
VRF	EPG	pcTag
V1	E1	49156
V1	E2	16387
V1	ext2	49155

Host Table

VRF	EP	PcTag	Dst
V1	Host1	49156	Leaf1
V1	Host2	16387	Leaf2

LPM Table

VRF	Subnet	PcTag	Dst
V1	int-S1	1	Proxy
V1	ext-S2	49155	L3Out



- Apply policy between src E1(49156) and dst ext2(49155)

Import Route Control

*Import Route Control supported only for OSPF & BGP

Import Route Control is used to filter External Routes received on an L3Out

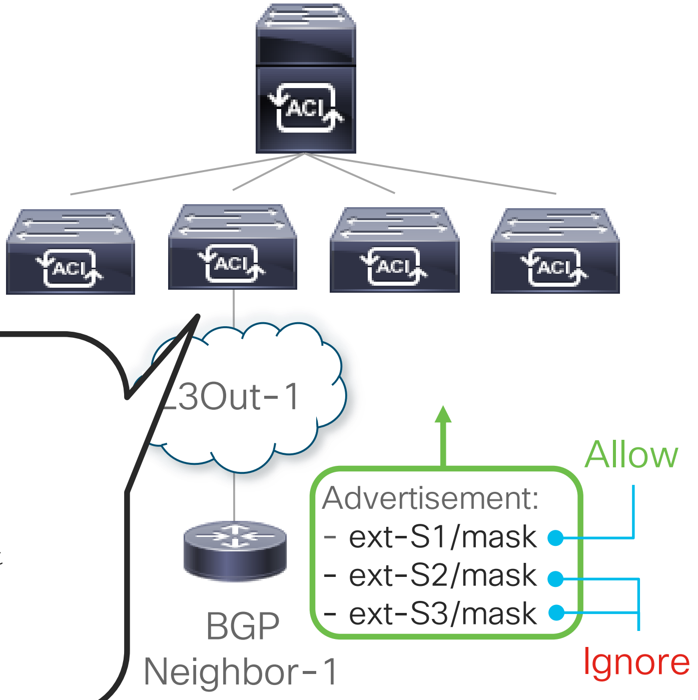
Subnet: ext-S1/mask
Scope:
✔ Import Route Control Subnet

- A route-map is created per BGP neighbor to filter incoming routes. Subnets defined with the import flag will be added to corresponding prefix list to allow in remote routes.
- The import flag must be enabled on the L3Out to set import flag per external subnet.
- By default, import is disabled on the L3Out

```
neighbor neighbor-1
  Inbound route-map imp-l3out-vrf
  Outbound route-map exp-l3out-vrf

route-map imp-l3out-peer-vrf permit
  match: prefix-list IPv4-network-vrf-exc-ext-inferred-import-dst

ip prefix-list IPv4-network-vrf-exc-ext-inferred-import-dst
  permit ext-S1/mask
```



Aggregate Import

*Aggregate Import supported only for 0.0.0.0/0 or ::/0

Import Route Control allows fabric to permit a specific prefix. Instead of creating each prefix advertised by a neighbor, multiple prefixes can be aggregated together by using the Aggregate Import flag.

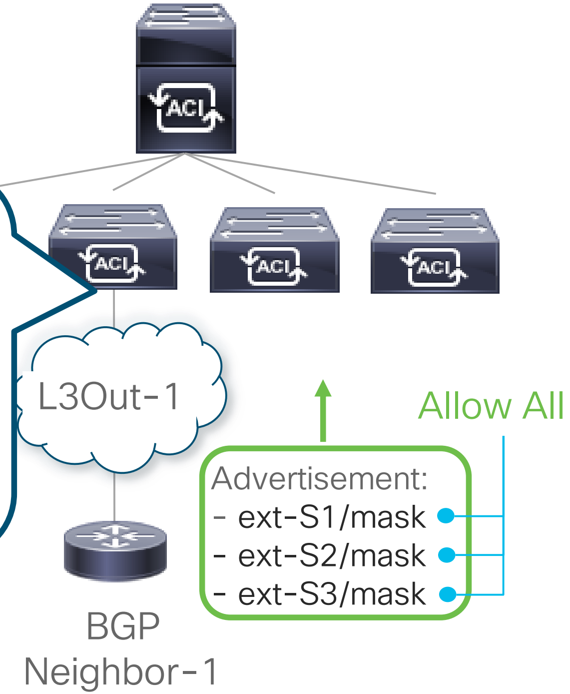
Subnet: 0.0.0.0/0
Scope:
- [x] Import Route Control Subnet

Aggregate:
- [x] Aggregate Import

```
neighbor neighbor-1
  Inbound route-map imp-l3out-vrf
  Outbound route-map exp-l3out-vrf

route-map imp-l3out-peer-vrf permit
  match prefix-list IPv4-network-vrf-exc-ext-inferred-import-dst

ip prefix-list IPv4-network-vrf-exc-ext-inferred-import-dst
  permit 0.0.0.0/0 le 32
```



Export Route Control & Aggregate Export

Subnet: ext-S1/mask
Scope:
✓ Export Route Control Subnet

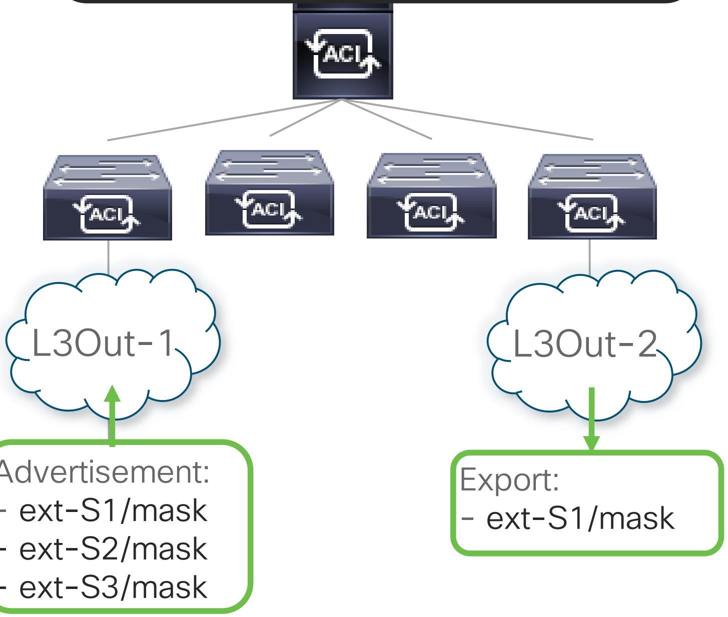
Export Route Control allows Transit Routes to be advertised out of the fabric.

- Export control does NOT affect pervasive BD SVIs, they are only advertised when the BD is associated with the L3Out.
- Similar to import route control subnet, a prefix list with corresponding exported subnets is created to allow routes to be advertised out

Aggregate Export is identical concept to aggregate import, allowing prefixes to be aggregated together in export direction.

Subnet: 0.0.0.0/0
Scope:
✓ Export Route Control Subnet
Aggregate:
✓ Aggregate Export

Export all Transit Routes within VRF



Shared L3Out

Similar to Shared Services, a **Shared L3Out** uses **contracts** to leak routes between VRFs. The leaked routes can be:

int-S1 subnet from VRF-V1 to VRF-V2

ext-S2 subnet from VRF-V2 into VRF-V1

Similar Restrictions as Shared Services

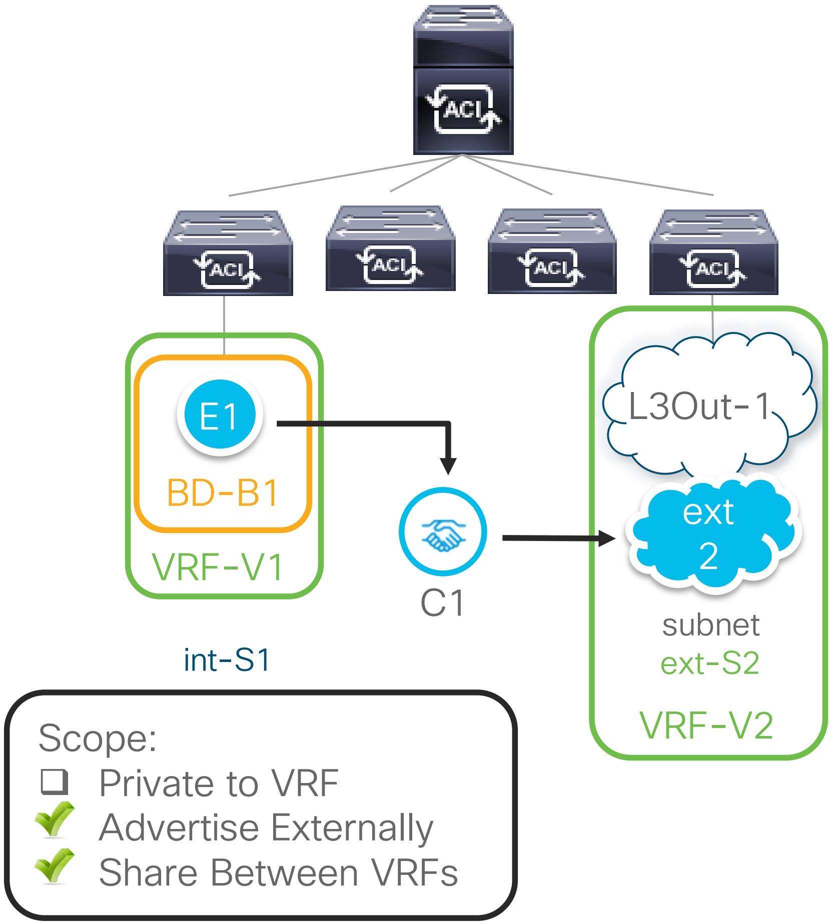
If the **application EPG** is **providing** the contract for shared L3Out, the **internal** subnet must be defined **under the EPG**.

If the **external EPG** is **providing** the contract for shared L3Out, then internal subnet can be defined either under the EPG or the BD

Internal subnet must have **shared** and Advertise Externally(**public**) scope.

Contract must be appropriately scoped.

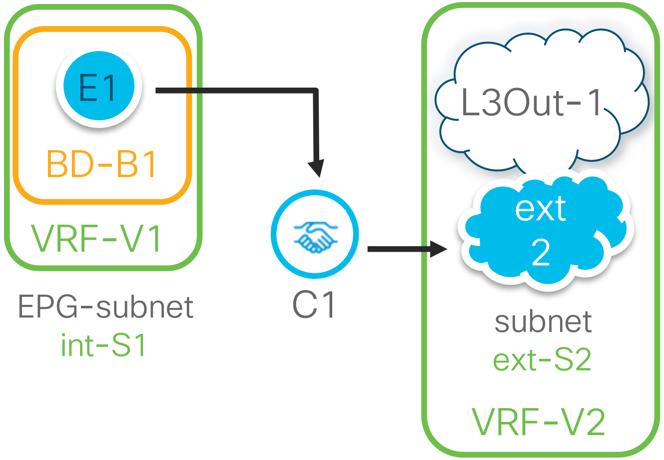
For shared L3Out, shared subnet must be globally unique within the entire ACI fabric.



Cisco live!

Shared L3Out

What happens in the fabric when contract is added?



- Internal Route int-S1 leaked into VRF-V2
  ext-S2 route not leaked into VRF-V1 yet...
- Shared-Service prefix list added to route-map permitting advertisement of int-S1. External routers can now learn int-S1 through OSPF, EIGRP, or BGP on VRF-V2.
  No need to associate BD to shared L3Out, route controlled by contract!
- Shared-Service contract programmed onto leaf to allow traffic flow.
- Problems:
  - VRF-V1 does not have return route to ext-S2
  - Even though rule is programmed, return traffic from VRF-V1 can't derive destination pcTag so no policy available to enforce

Assume: VRF-V2 has a route to ext-S2 through static or dynamic route

VRF	Route	pcTag	Flags
V1	int-S1	1	proxy
V2	ext-S2	ext2	L3Out
V2	int-S1	E1	proxy, **leak->V1**

Shared L3Out

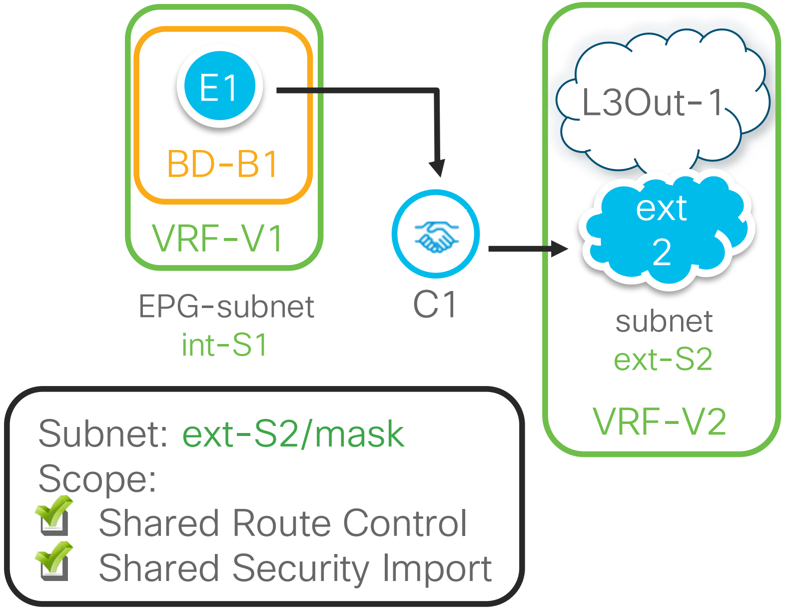
Completing the Configuration

Shared Route Control flag allows external route to be leaked into EPG context.

- In this example, adding shared route control to the external subnet allows ext-S2 to be leaked into VRF-V1, but pcTag set to reserved drop value.

Shared Security Import is used to classify dataplane packets into external EPG for policy enforcement for shared prefixes

- In this example, adding shared security import to the external subnet created a prefix-based EPG in any-VRF* for the external subnet ext-S2 with pcTag of EPG-ext2.



VRF	Route	pcTag	Flags
V1	int-S1	1	proxy
V2	ext-S2	ext2	L3Out
V2	int-S1	E1	proxy, **leak->V1**
V1	ext-S2	ext2	L3Out, **leak->V2**

Aggregate Shared

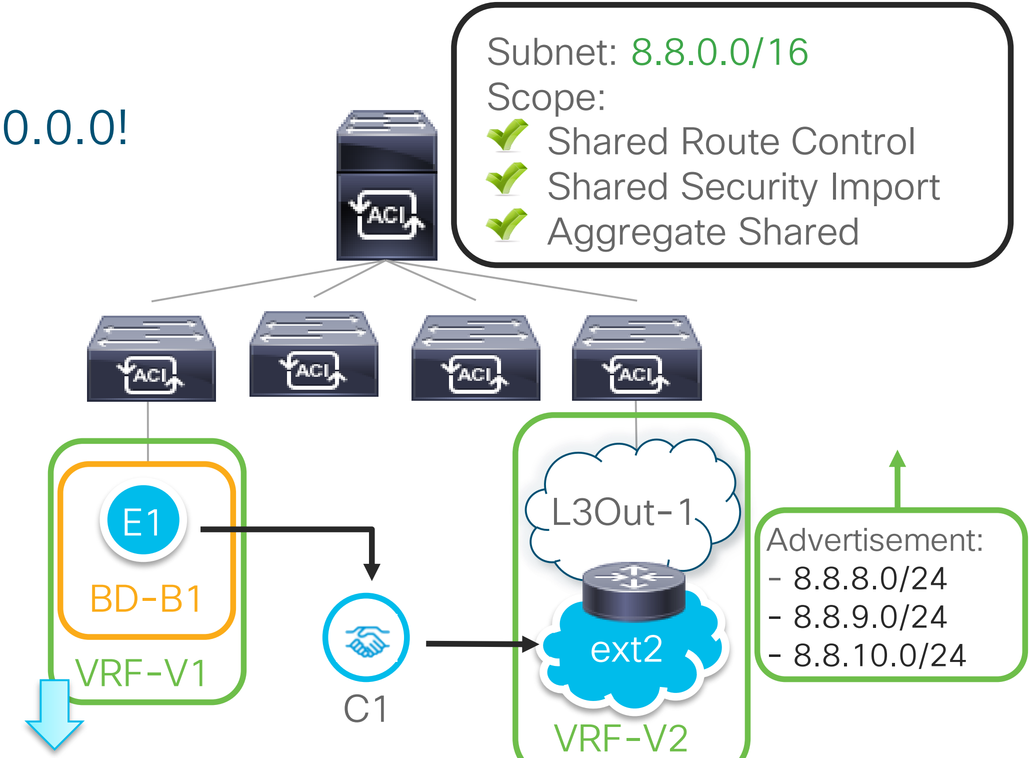
Supported for any prefix, not just 0.0.0.0!

Aggregate Shared flag allows multiple prefixes from L3Out to be shared/leaked into another VRF.
In this example, a /16 prefix is configured with aggregate shared flag set. The external router advertised multiple /24 subnets within the range. Each are leaked into VRF-V1

Restrictions
Shared Route control subnets cannot be a subset of Shared Security import. For example:
8.8.0.0/16
- shared security import + shared route control + aggregate shared

8.8.10.0/24
- shared route control (only)

Traffic on VRF-V1 toward 8.8.10.0/24 dropped



Forwarding Table

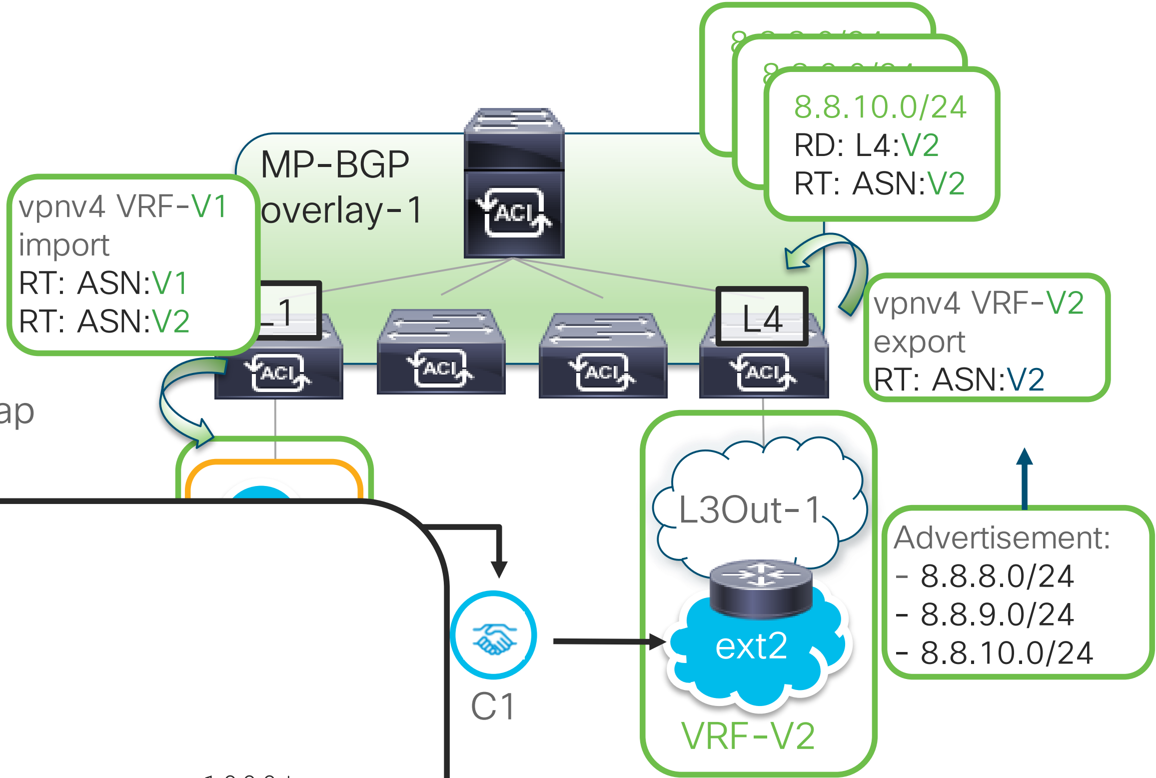
VRF	Route	pcTag	Flags
V1	8.8.8.0/24	ext2	L3Out, **leak->V2**
V1	8.8.9.0/24	ext2	L3Out, **leak->V2**
V1	8.8.10.0/24	ext2	L3Out, **leak->V2**

Cisco live!

Aggregate Shared

How does this work?

- Leaf4 exports routes into MP-BGP with route-target for VRF V2
- Leaf1 imports routes with route-targets from both VRF-V1 and VRF-V2 into V1 vrf. Routes are filtered with route-map based on subnet control flags




```
leaf101# show bgp process vrf V1
 Import route-map V1-shared-svc-leak
 Import RT list:
       ASN:V1
       ASN:V2
...
route-map V1-shared-svc-leak, permit, sequence 1000*
  Match clauses:
    ip address prefix-lists: IPv4-V2-V1-shared-svc-leak

ip prefix-list IPv4-V2-V1-shared-svc-leak
   seq 3 permit 8.8.0.0/16 le 32
```

L3 External Subnet Review

- External Subnets for the External EPG (Security Import)
  Used to classify dataplane packets into external EPG for policy enforcement

- Export Route Control
  filter Transit Routes advertised out of the fabric.

- Import Route Control
  filter External Routes received on an L3Out

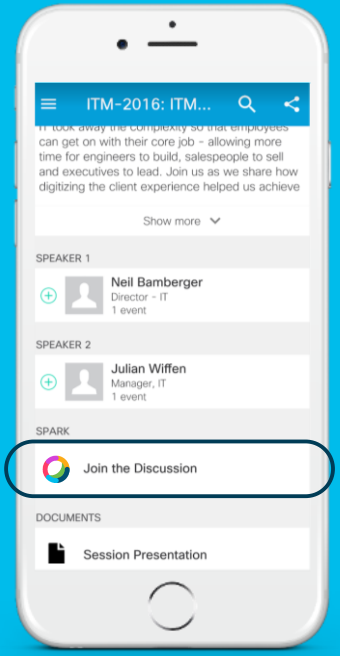
- Shared Security Import
  used to classify dataplane packets into external EPG for policy enforcement for shared/leaked prefixes

- Shared Route Control
  Allows external route to be leaked into another VRF

- Aggregate Export - allows prefixes to be aggregated together in export direction (0/0 or ::/0 only)
- Aggregate Import - allows prefixes to be aggregated together in import direction (0/0 or ::/0 only)
- Aggregate Shared Route - allows prefixes to be aggregated together for shared route control

Agenda

- Introduction
- Building the Overlay
  - Access Policies
  - VRFs, Bridge Domains, and EPGs
  - L2Outs and Loop Prevention
- Traversing the Overlay
  - Learning, Forwarding, and Policy Enforcement
  - Shared Services and Route Leaking
  - L3outs and Routing Protocols



cs.co/ciscolivebot#BRKACI-3101

Cisco Webex Teams

Questions?

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

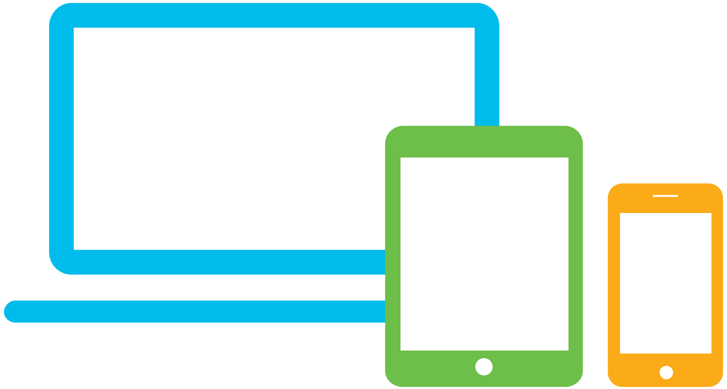
How

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

Complete your online session survey

- Please complete your Online Session Survey after each session
- Complete 4 Session Surveys & the Overall Conference Survey (available from Thursday) to receive your Cisco Live T-shirt
- All surveys can be completed via the Cisco Events Mobile App or the Communication Stations

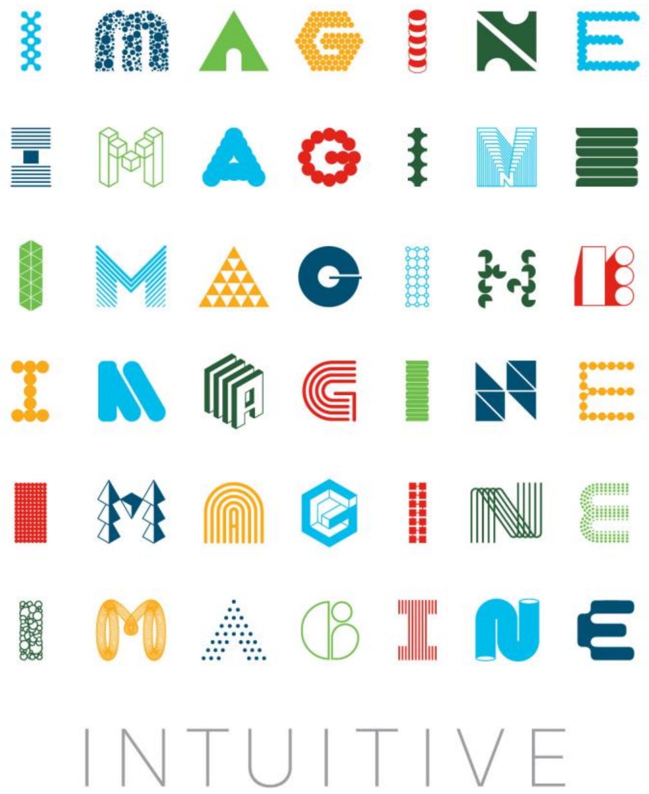
Don’t forget: Cisco Live sessions will be available for viewing on demand after the event at ciscolive.cisco.com



Continue Your Education

- Demos in the Cisco Showcase
- Walk-in self-paced labs
- Meet the engineer 1:1 meetings
- Related sessions

Thank you



IMAGINE

INTUITIVE