
cisco Live!
Let's go

The bridge to possible

Best practice of Virtual Port Channel in VXLAN

Sonu Khandelwal
Principal Technical Marketing Engineer

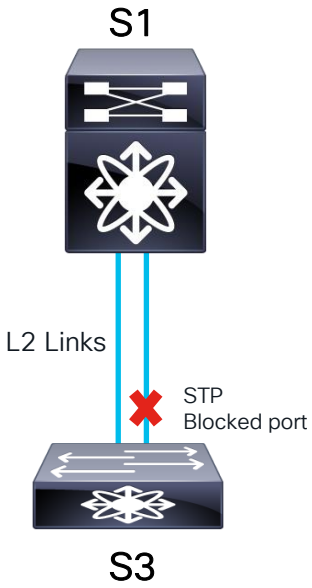
cisco Live!

BRKDCN-2912

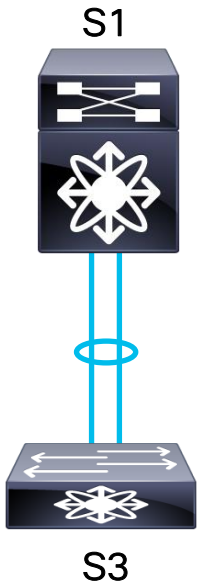
Agenda

- vPC Basics
- VXLAN overview
  - Control plane, data plane and packet walk
- vPC in VXLAN
- vPC configuration best practices
- vPC Boarder Gateway
- Automate vPC in VXLAN using Nexus Dashboard Fabric Controller (NDFC)
- Key Takeaways

Port-Channel aka EtherChannel

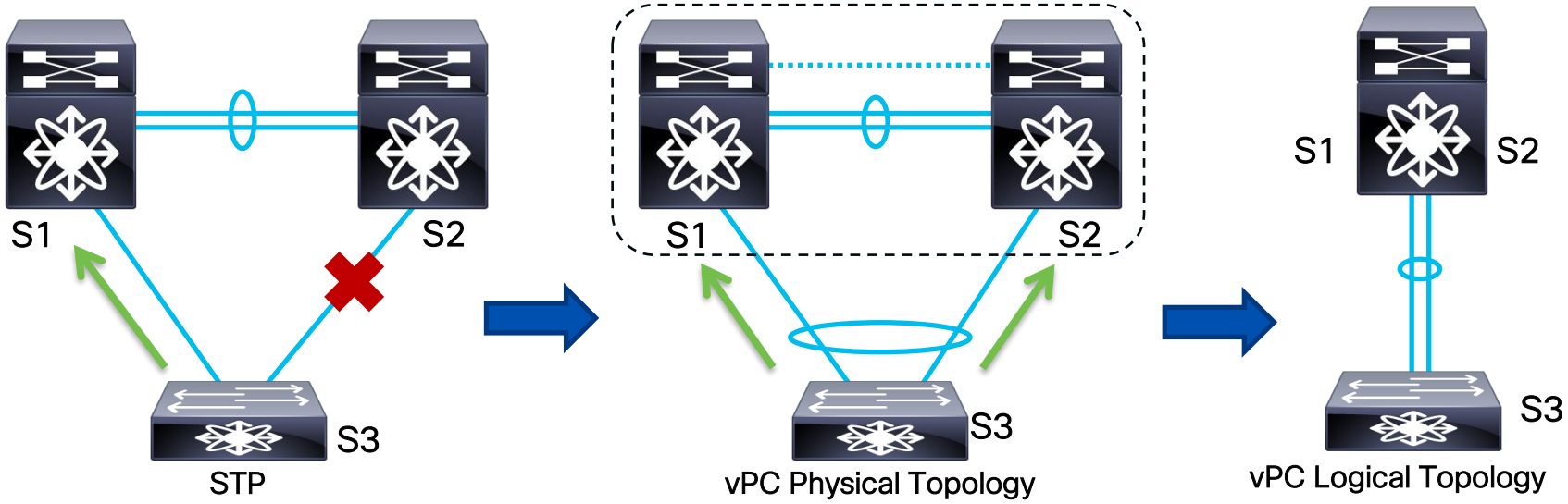


- Wasted Port
- Additional cost



- No Blocked Ports
- More Usable Bandwidth

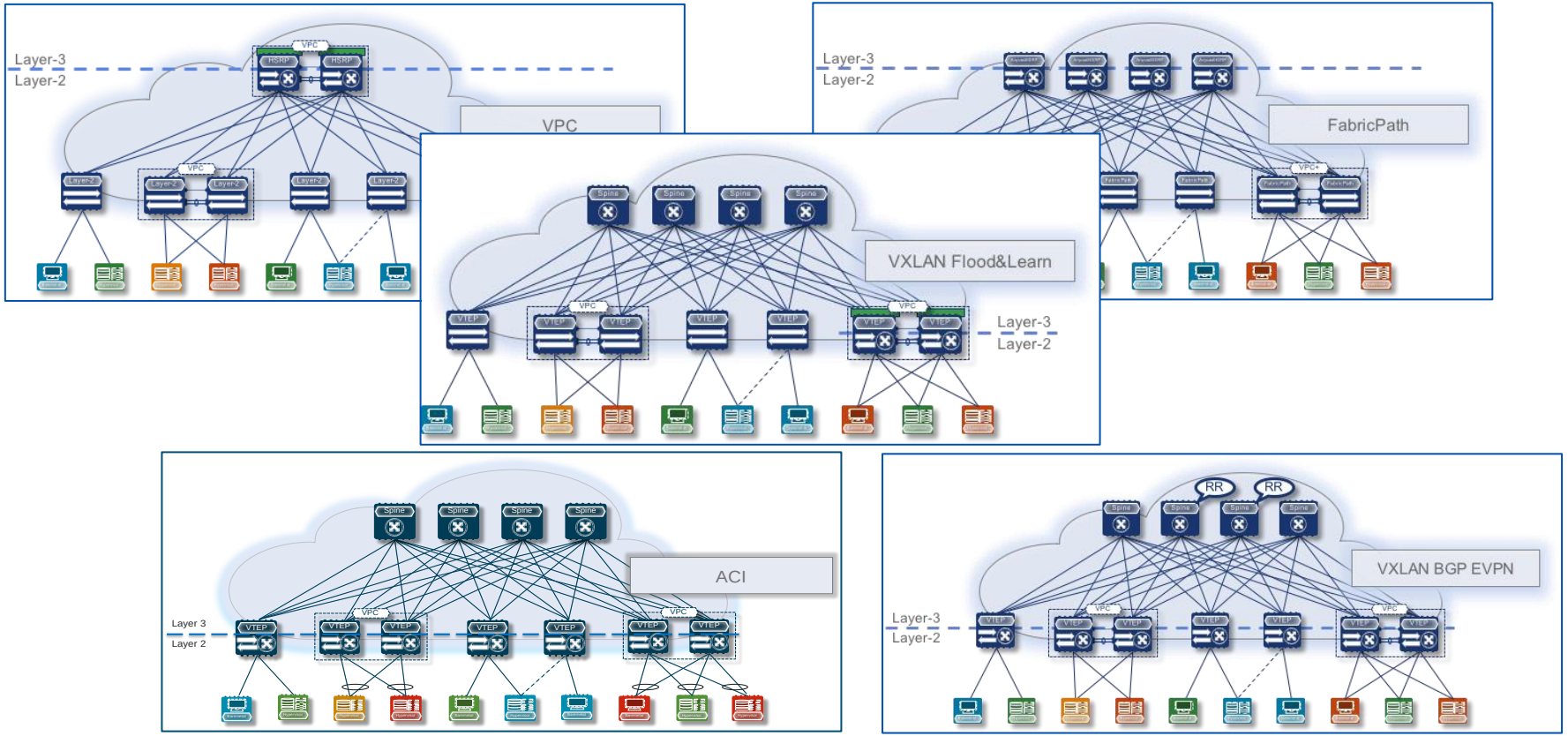
Why vPC?



- Wasted Port
- Additional cost

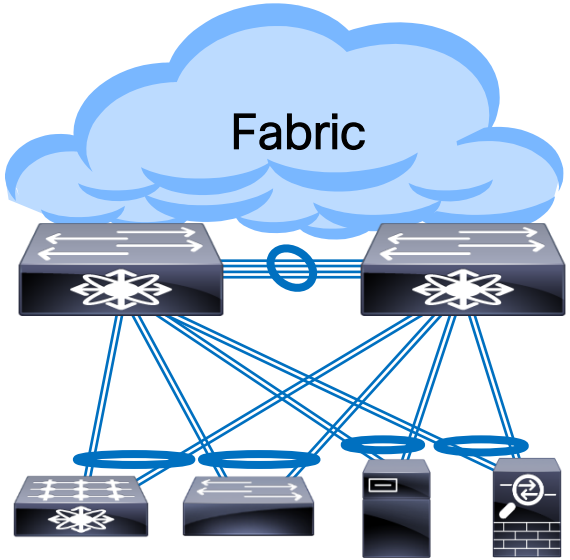
- No Blocked Ports
- More Usable Bandwidth

vPC everywhere!



Virtual Port Channel – vPC

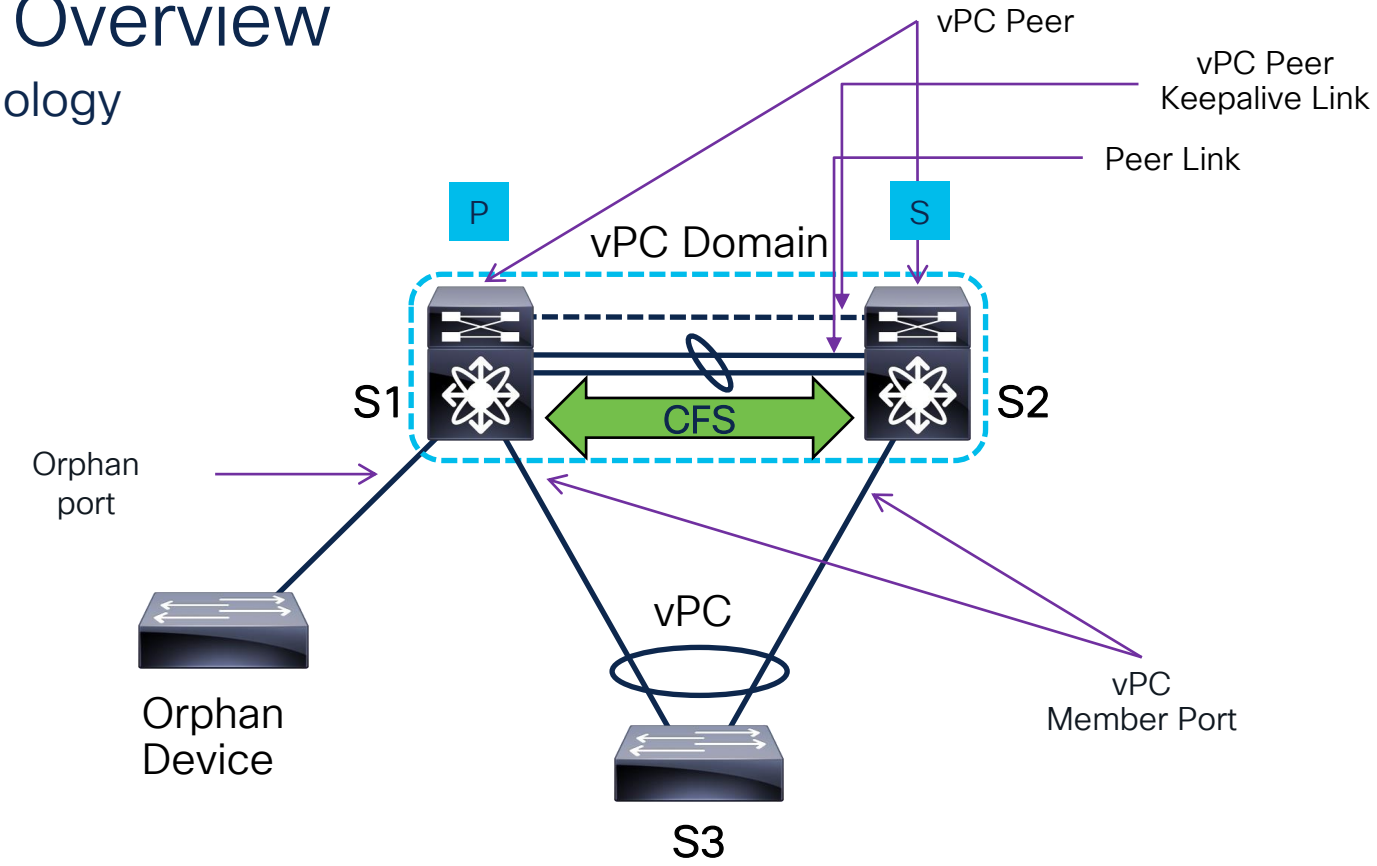
- Eliminates Spanning Tree blocked ports by providing loop-free topology
- Full link bandwidth utilization
- Provides device level redundancy
- Faster convergence over STP
- MC-LAG on Cisco Nexus Devices



vPC Basics

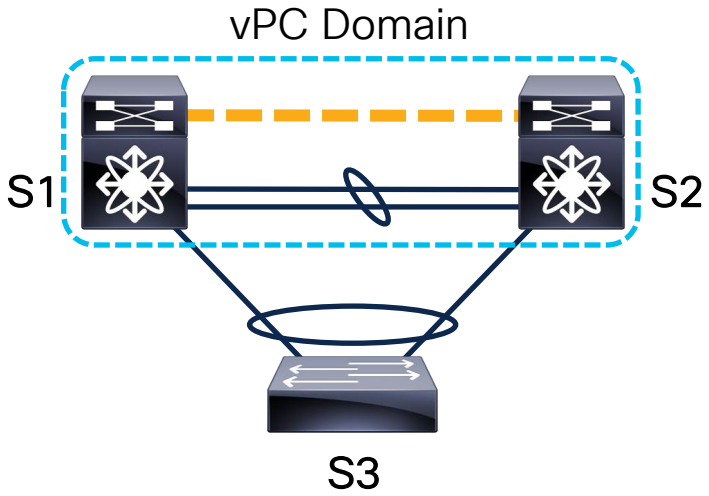
vPC Overview

Terminology



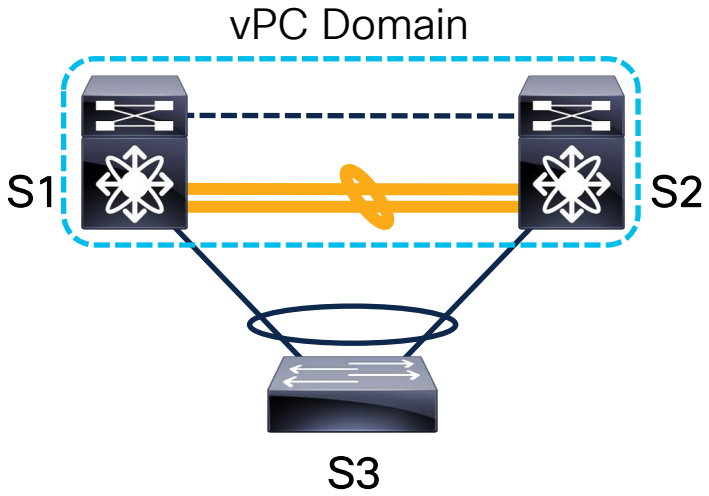
vPC Peer-keepalive link

- Carries periodic hart beet between vPC peers, to make sure both peers are up
- Uses UDP port 3200
- Sends keep alive heart beets every second



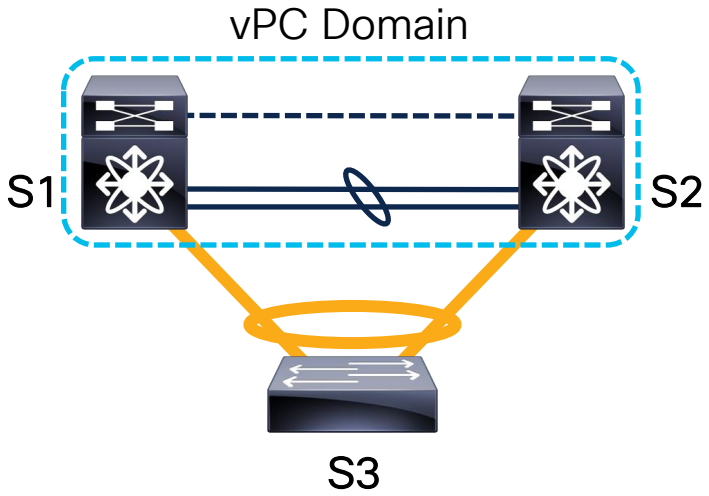
vPC Peer link

- vPC peer link is a port channel that carries:
  - vPC VLANs
  - CFS messages
  - Flooded traffic from the other peer device
  - STP BPDUs, HSRP hello messages and IGMP updates
  - Multicast traffic
- vPC imposes the rule that peer link should never be blocking



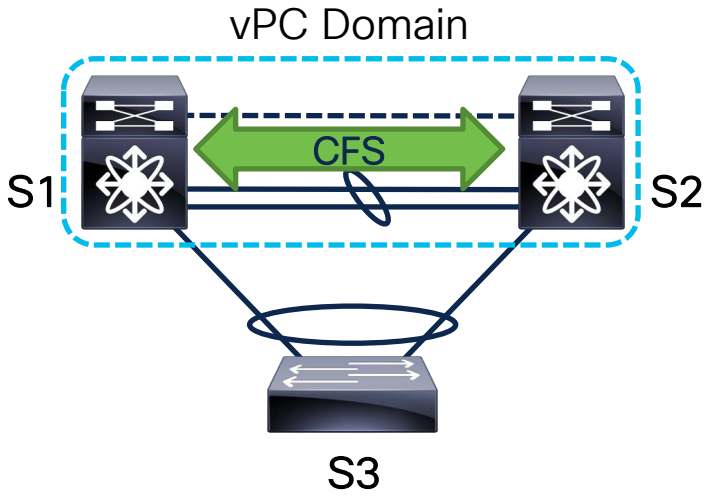
Virtual Port Channel – vPC

- Consists of port channel members of vPC
- L2 port channel
- Ports in vPC can be in access or trunk mode
- VLANs allowed on vPC need to be allowed on peer link
- LACP and Static port channel configuration



Cisco Fabric Services Protocol

- Synchronization and consistency checking mechanism
- Runs on vPC peer link
- CFS protocols mechanism:
  - Validation and comparison for consistency check
  - Synchronization of MAC addresses for member ports
  - Status of member ports advertisement
  - STP management
  - Synchronization of HSRP and IGMP snooping
- Enabled by default



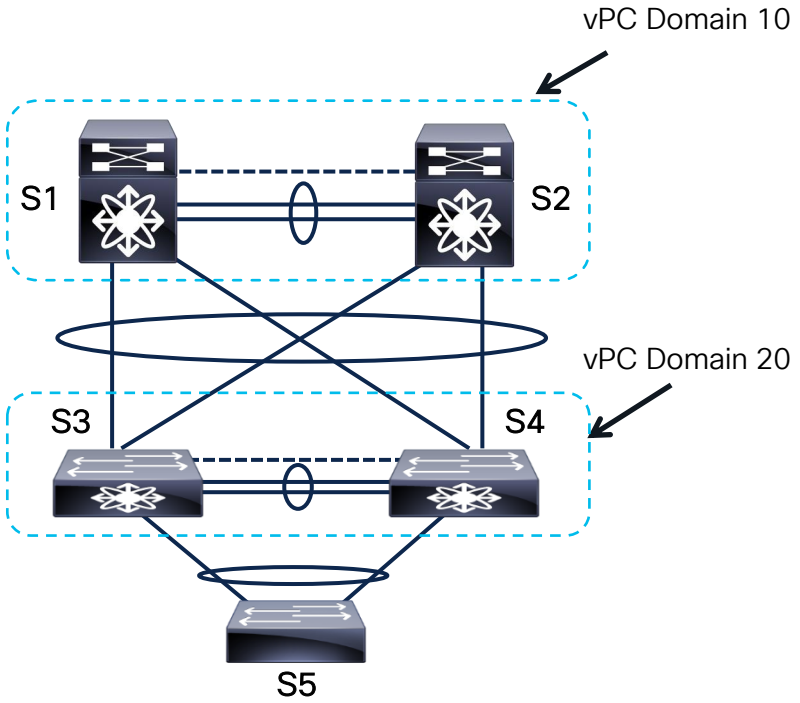
vPC Consistency check

Consistency Check	Type-1 Configuration parameters that **Must** be identical	Type-2 Configuration parameters that **Should** be identical
Configuration Parameters	• STP Mode • STP VLAN state • STP Global settings • LACP Mode • MTU	• VLAN Interface (SVI) • ACL • QoS • IGMP • HSRP
Mismatch Action	**Global Parameters** : vPC member ports on secondary are brought down **Per-interface Parameters** – vPC member ports on secondary peer set to down state	• Forwards traffic in case of inconsistency • Undesirable forwarding behavior

vPC Configuration Best Practices

vPC Domain-ID

- The vPC peer devices use the vPC domain ID to automatically assign a unique vPC system MAC address
- System MAC is used in STP BPDU, LACP BPDU, and IGMP advertisements
- You MUST use unique Domain id's for all vPC pairs defined in a contiguous layer 2 domain

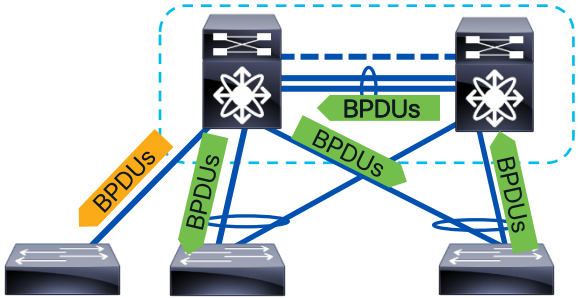



```
! Configure the vPC Domain ID – It should be unique within
the layer 2 domain
NX-1(config)# vpc domain 20

! Check the vPC system MAC address
NX-1# show vpc role
<snip>
vPC system-mac                    : 00:23:04:ee:be:14
```

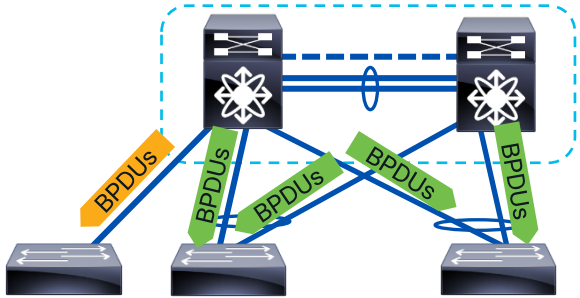
vPC Peer-Switch

- Without Peer-switch:
  - STP for vPCs controlled by vPC primary
  - vPC primary send BPDU's on STP designated ports
  - vPC secondary device proxies BPDU's to primary



- With Peer-switch:
  - Peer-Switch makes the vPC peer devices to appear as single STP root
  - BPDUs processed by the logical STP root formed by the 2 vPC peer devices

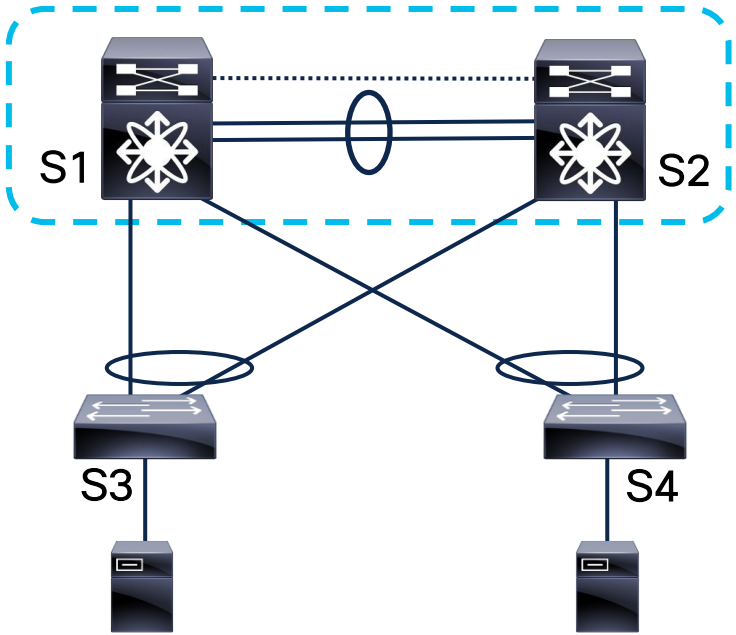
Nexus(config-vpc-domain)# peer-switch



vPC Peer-Gateway

- Allows a vPC switch to act as the active gateway for packets addressed to the peer router MAC
- Keeps forwarding of traffic local to the vPC node and avoids use of the peer link
- Allows Interoperability with features of some NAS or load-balancer devices

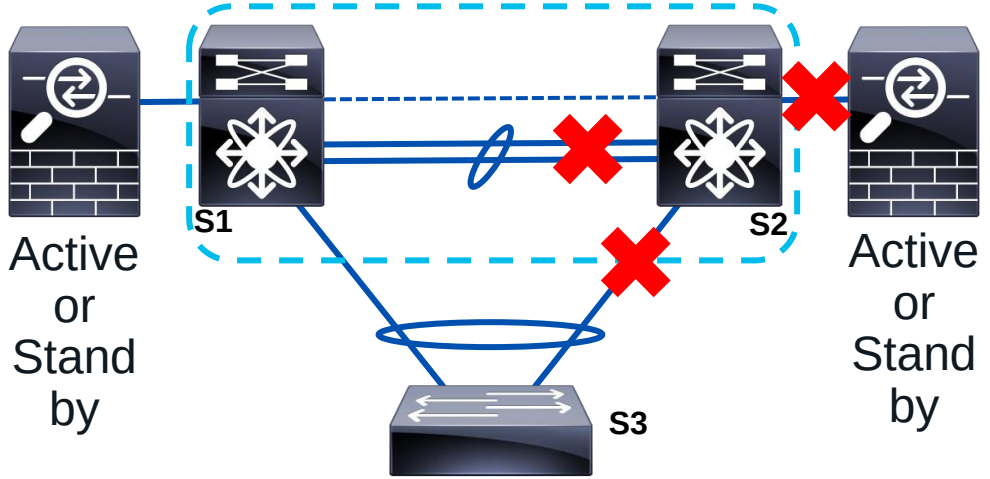
Nexus(config-vpc-domain)# peer-gateway



vPC Orphan Ports Suspend

- Single attached devices to vPC domain, will black-hole traffic if peer link fails
- With Orphan Port Suspend feature, will suspend orphan ports on vPC secondary peer
- When peer link is restored, vPC secondary restores orphan ports

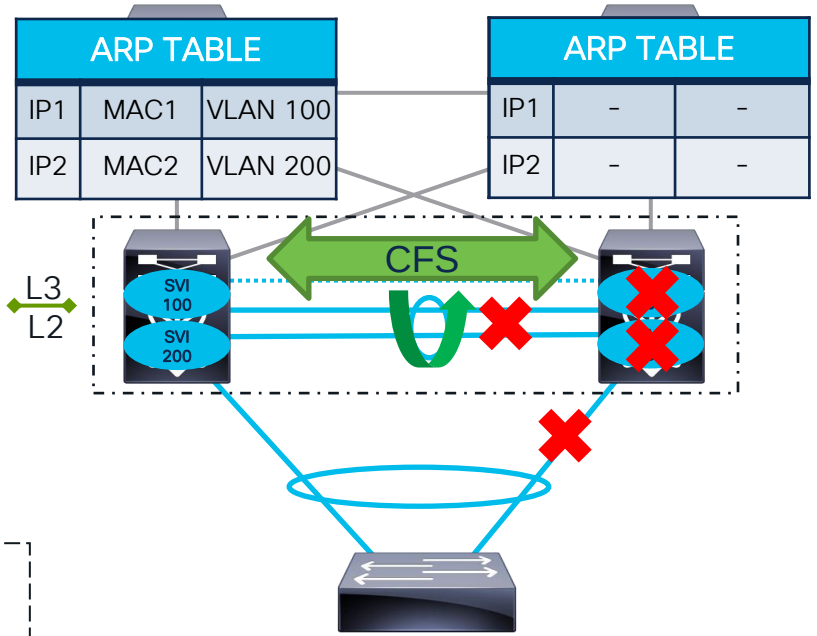
Nexus(config-if)# vpc orphan-ports suspend



vPC Configuration Best Practices

vPC ARP/ND sync

- When peer device goes down or peer link goes down, SVIs are suspended
- After restore of the peer device, or peer link, ARP table is empty - traffic black-holed
- Before bringing up SVI, peer devices synchronize ARP table over CFS
- Reduces convergence time



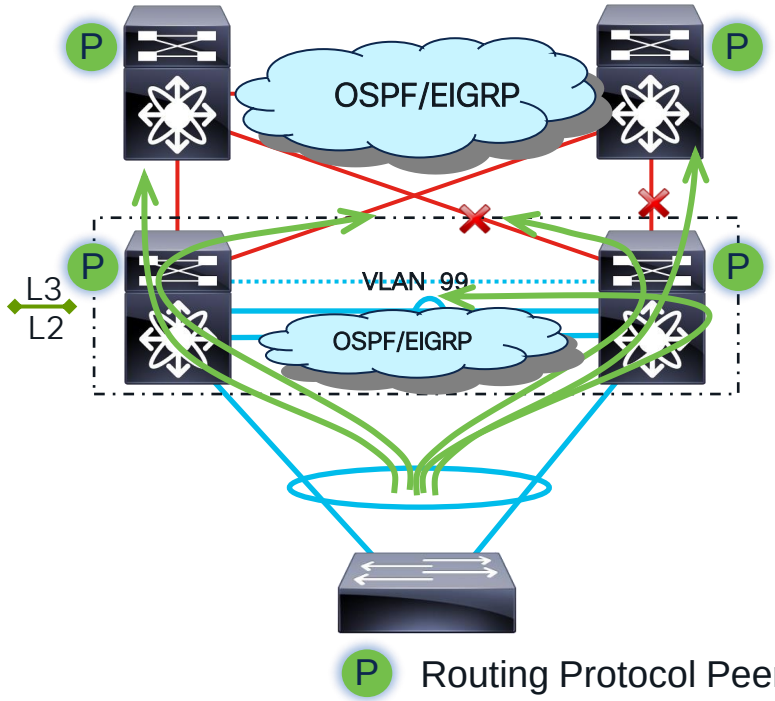

```
Nexus(config-vpc-domain)# ip arp synchronize
Nexus(config-vpc-domain)# ipv6 nd synchronize
```

Design Best Practices

Backup Routing Path

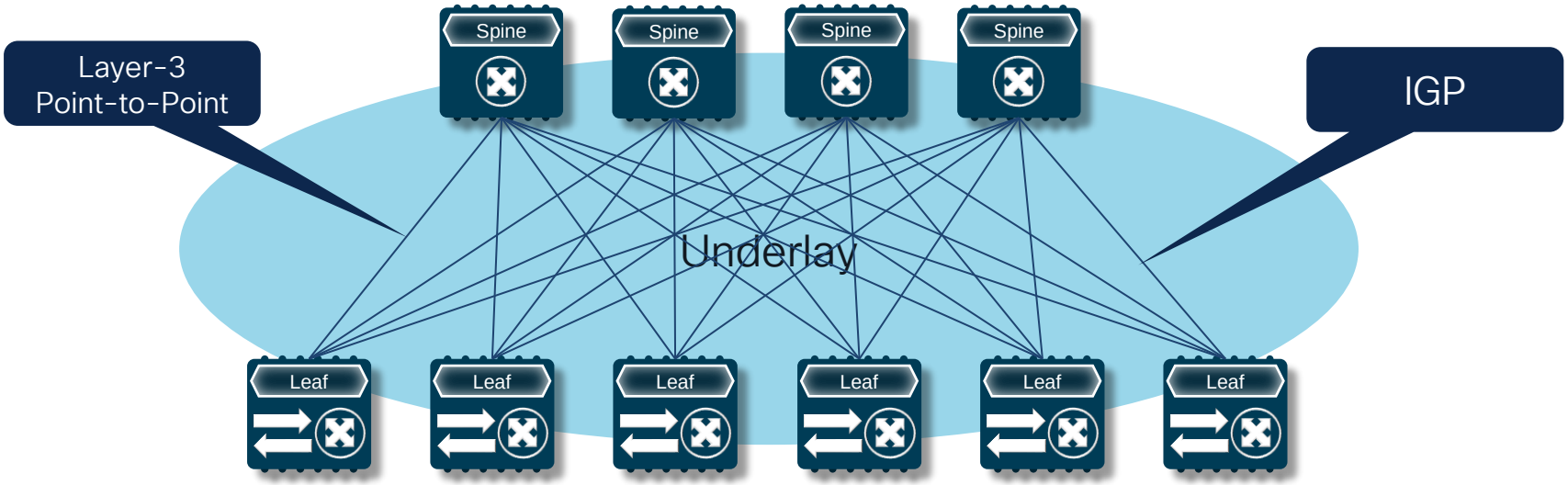
- Point-to-point dynamic routing protocol adjacency between the vPC peers to establish a L3 backup path to the core through peer link in case of uplinks failure
- Define SVIs associated with FHRP as routing passive-interfaces in order to avoid routing adjacencies over vPC peer link
- A single point-to-point VLAN/SVI (aka transit VLAN) will suffice to establish a L3 neighbor
- Alternatively, use an L3 point-to-point link between the vPC peers to establish a L3 backup path

Use one transit VLAN to establish L3 routing backup path over the vPC peer link in case L3 uplinks were to fail, all other SVIs can use passive-interfaces



VXLAN overview – control plane, data plane and packet walk

Vxlan Taxonomy : Underlay

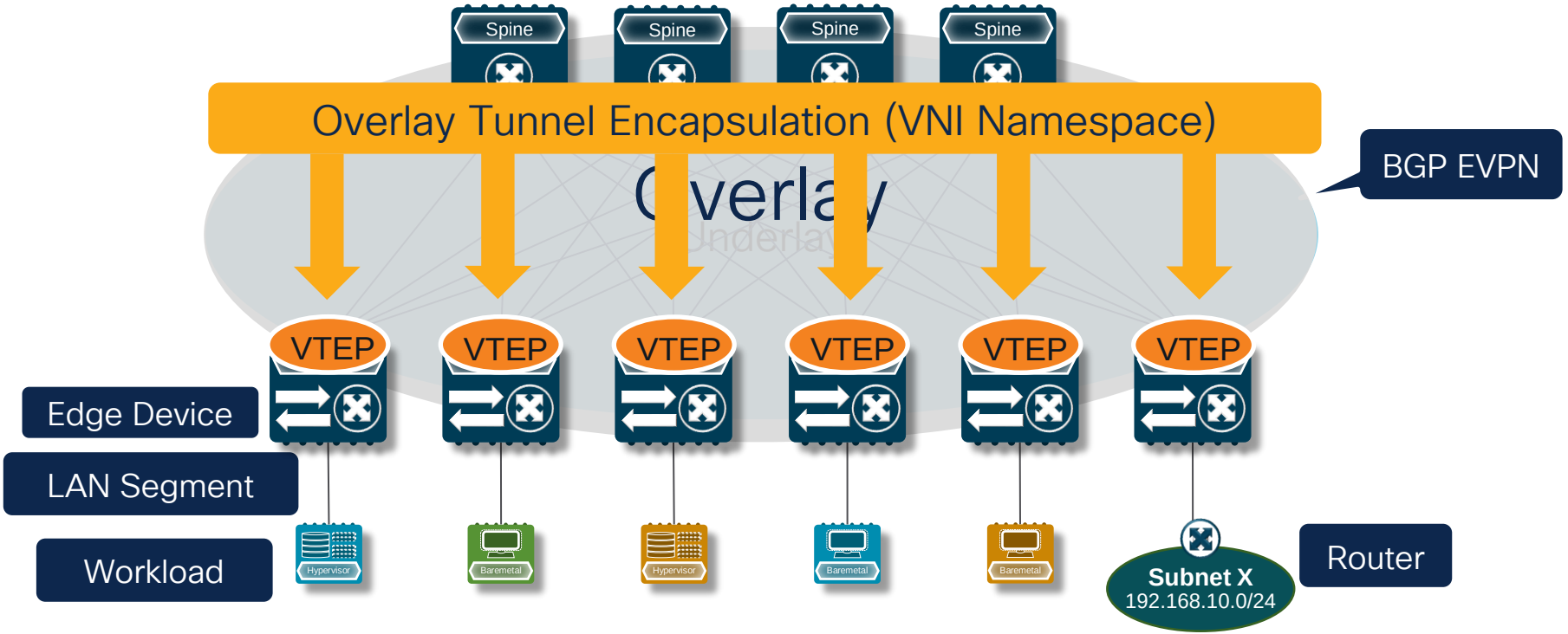


Vxlan Taxonomy : Overlay

EVPN: Ethernet VPN

VTEP: VXLAN Tunnel End-Point

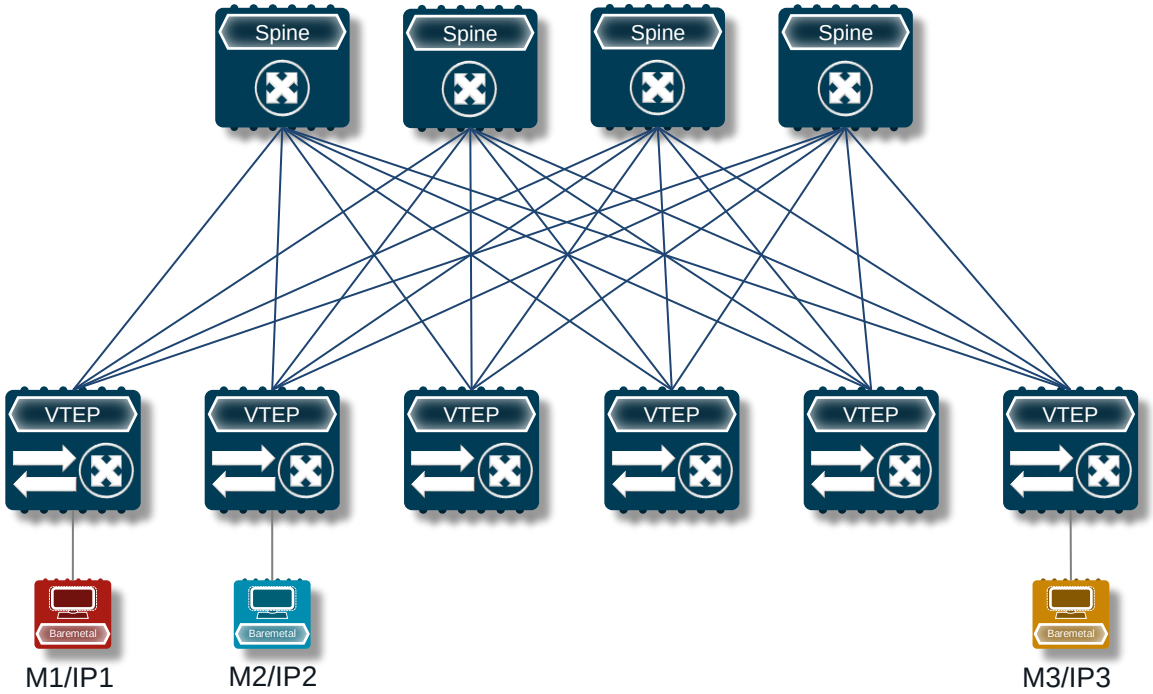
VNI/VNID: VXLAN Network Identifier



Vxlan : Control-plane

Control-Plane

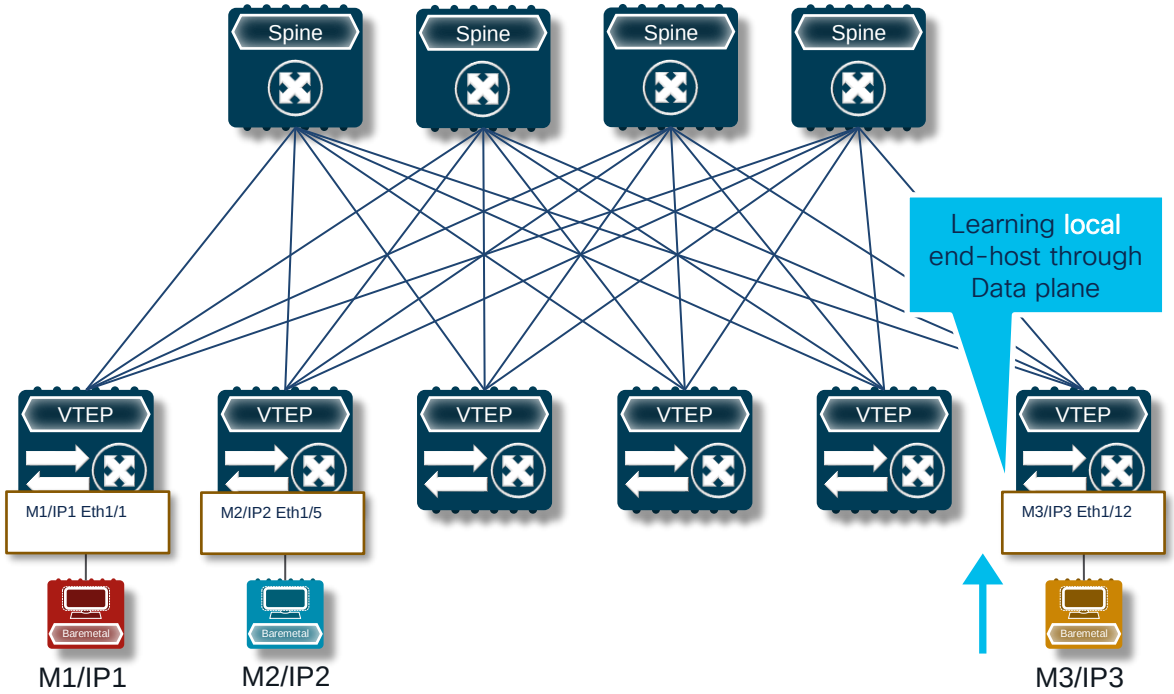
- Route Learning
  - Local Learning
  - Remote Learning
- Route Distribution
- Peer Discovery



Vxlan : Control-plane

Control-Plane

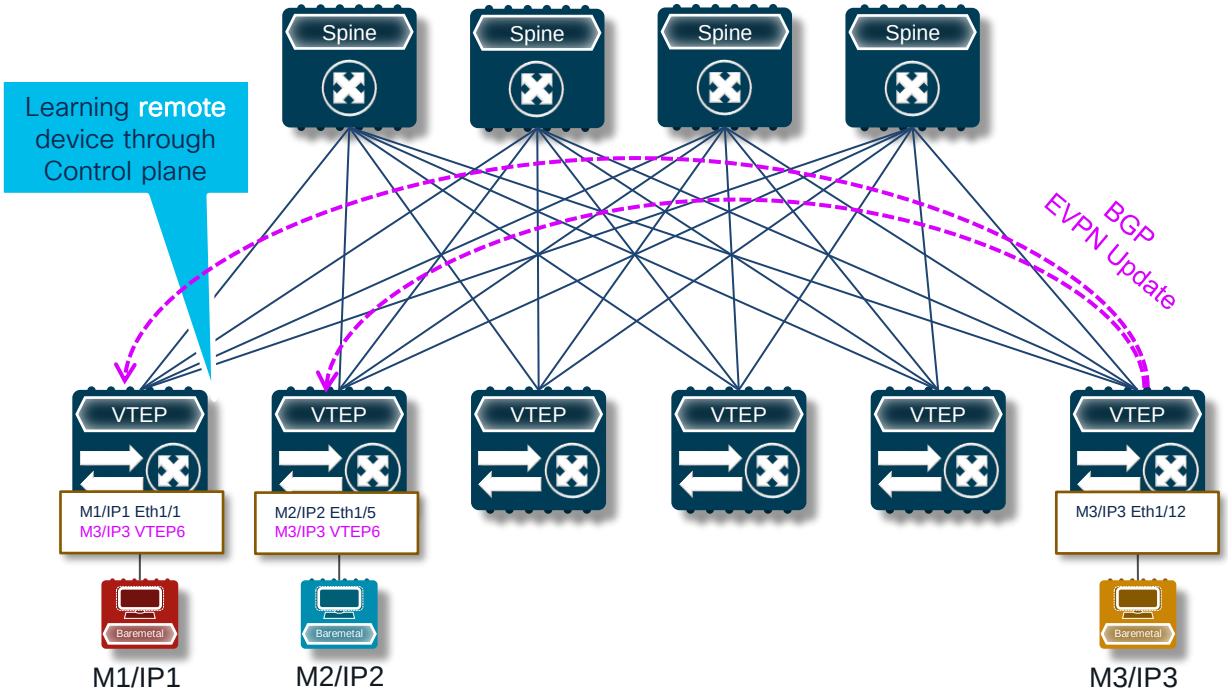
- Route Learning
  - Local Learning
  - Remote Learning
- Route Distribution
- Peer Discovery



Vxlan : Control-plane

Control-Plane

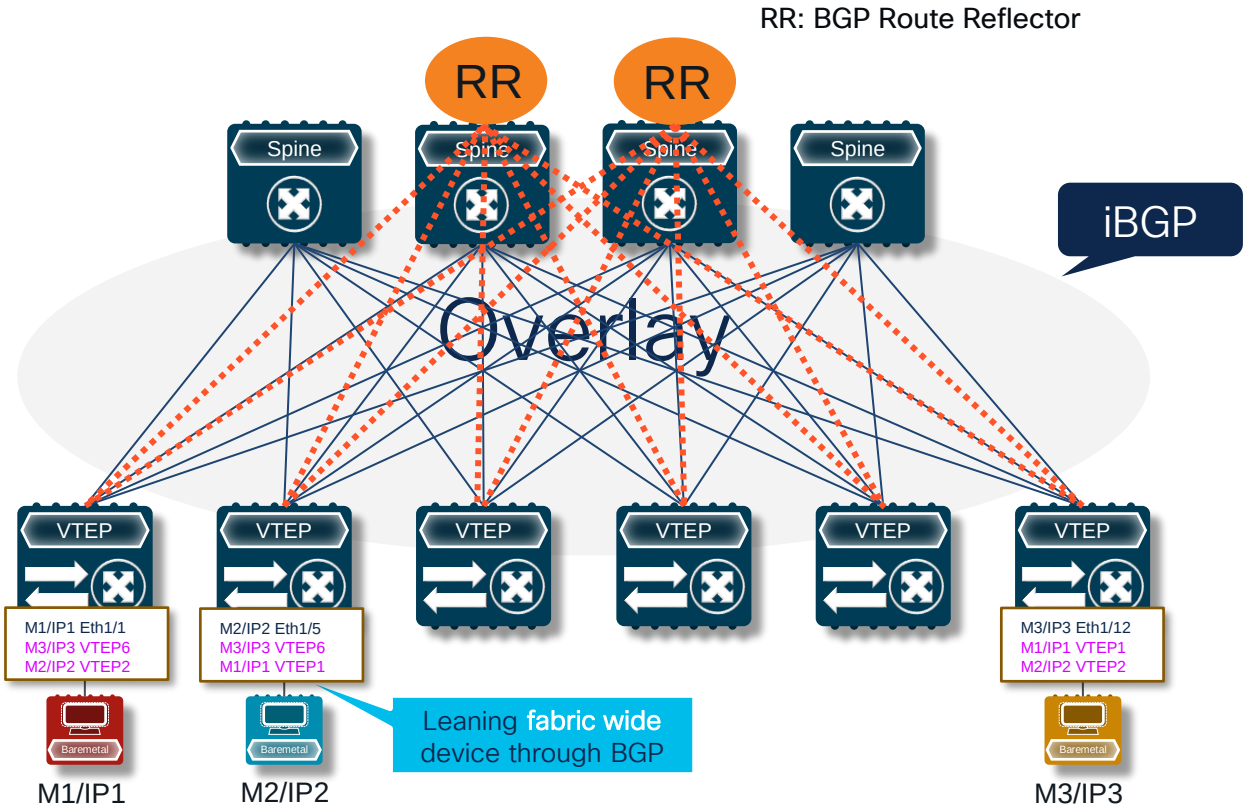
- Route Learning
  - Local Learning
  - Remote Learning
- Route Distribution
- Peer Discovery



Vxlan : Control-plane

Control-Plane

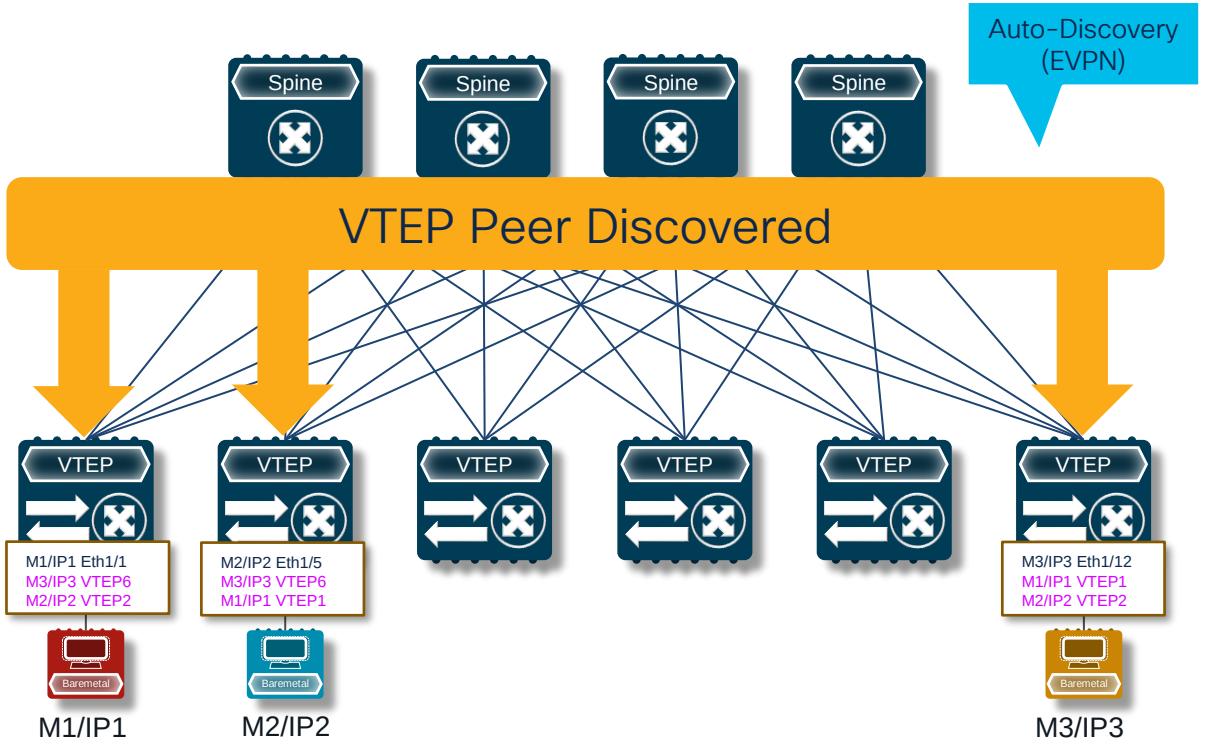
- Route Learning
  - Local Learning
  - Remote Learning
- Route Distribution
- Peer Discovery



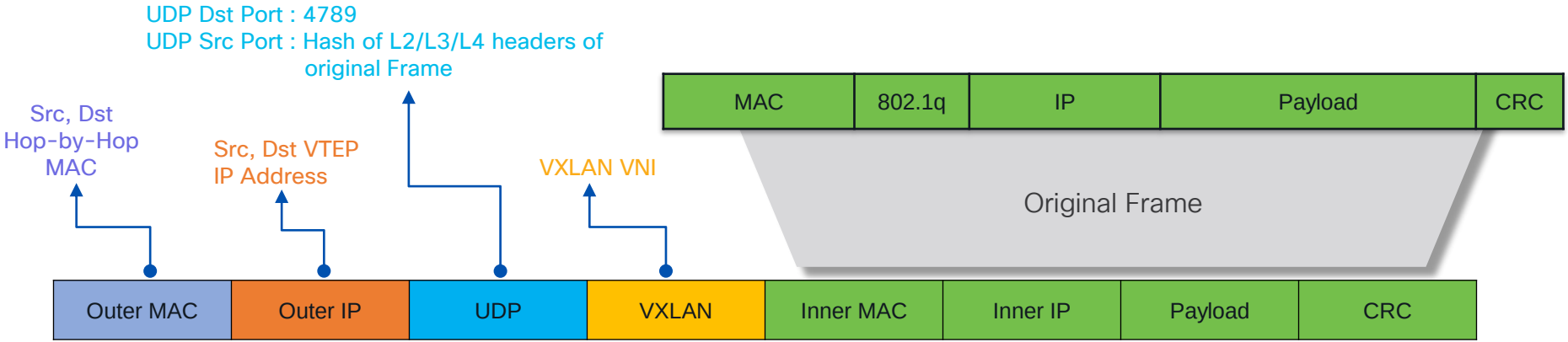
Vxlan : Control-plane

Control-Plane

- Route Learning
  - Local Learning
  - Remote Learning
- Route Distribution
- Peer Discovery



VXLAN – Data Plane



Data-Plane (VXLAN)

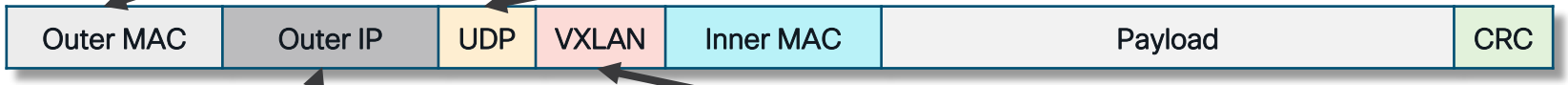
14-byte (outer MAC) + 20-byte (Outer IP) + 8-byte (UDP) + 8-byte (VXLAN) = 50 Bytes of total overhead*

*optional 4-byte if IEEE 802.1q exists

VXLAN Frame Format – MAC in IP Encapsulation

Field	Value	Bites	Total
Dest. MAC Address	Next-Hop MAC Address	48	14 Bytes (4 Bytes Optional)
Src. MAC Address	Next-Hop MAC Address	48	
VLAN Type	0x8100	16	
VLAN ID	Tag	16	
Ether Type	0x0800	16	

Field	Value	Bites	Total
Source Port	L2/L3/L4 Hash	16	8 Bytes
Destination Port	4789 (UDP)	16	
UDP Length		16	
Checksum	0x0000	16	



Field	Value	Bites	Total
IP Header	Misc. Data	72	20 Bytes
Protocol	0x11 (UDP)	8	
Header Checksum	Various	16	
Source IP	Src, VTEP IP	32	
Destination IP	Dest. VTEP IP	32	

Field	Value	Bites	Total
VXLAN Flags	RRRRIRRR	8	8 Bytes
Reserved		24	
VNI	16M Possible Segments	24	
Reserved		8	

FYI

Route Distinguisher (RD)

Auto-derived Route Distinguisher (**`rd auto`**) is based on the Type 1 encoding format as described in IETF RFC 4364 section 4.2.

The Type 1 encoding allows –

- 4-byte administrative field
- 2-byte numbering field

In NXOS with auto-rd –

- Administrative field is derived from the BGP router-id.
- Numbering field
  - For IP-VRF – VRF ID
  - For MAC-VRF – 32767 + VLAN ID

Example auto-derived Route Distinguisher (RD) –

IP-VRF with BGP Router ID 192.0.2.1 and VRF ID 6 –
RD 192.0.2.1:6

MAC-VRF with BGP Router ID 192.0.2.1 and VLAN 20 –
RD 192.0.2.1:32787

FYI

Route Target (RT)

The auto-derived Route-Target (**`route-target import/export/both auto`**) is based on the Type 0 encoding format as described in IETF RFC 4364 section 4.2

The Type 0 encoding allows –

- 2-byte administrative field
- 4-byte numbering field

With 2 byte ASN in NXOS, the auto derived Route-Target is constructed with –

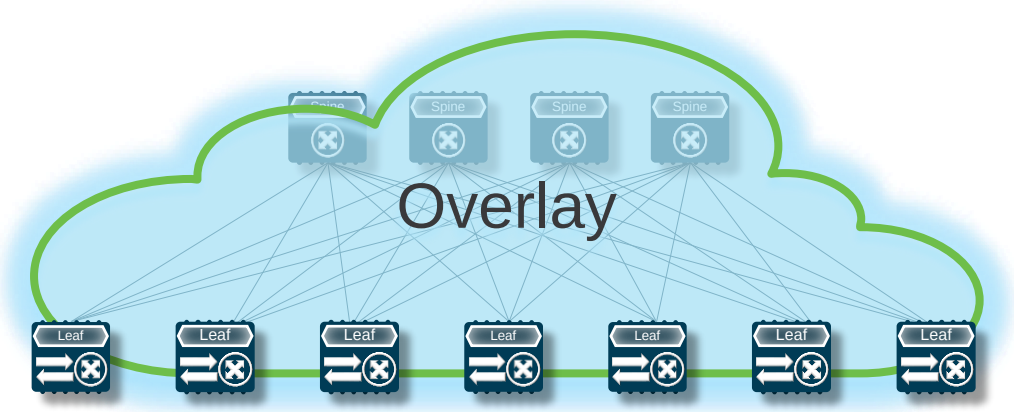
- Autonomous System Number (ASN) as the 2-byte administrative field
- Service Identifier (VNI) for the 4-byte numbering field

Example auto-derived Route Target (RT) –

IP-VRF within ASN 65001 and L3VNI 50001 –
Route Target **65001:50001**

MAC-VRF within ASN 65001 and L2VNI 30001 –
Route Target **65001:30001**

EVPN – Host and Subnet Routes



BGP EVPN NLRI*

Route Type 2 : Host Route

- Type-2 MAC only
  - MAC, Single VNI, Single Route Target
- Type-2 MAC + IP
  - MAC + IP, Two VNI, Two Route Target, Router MAC

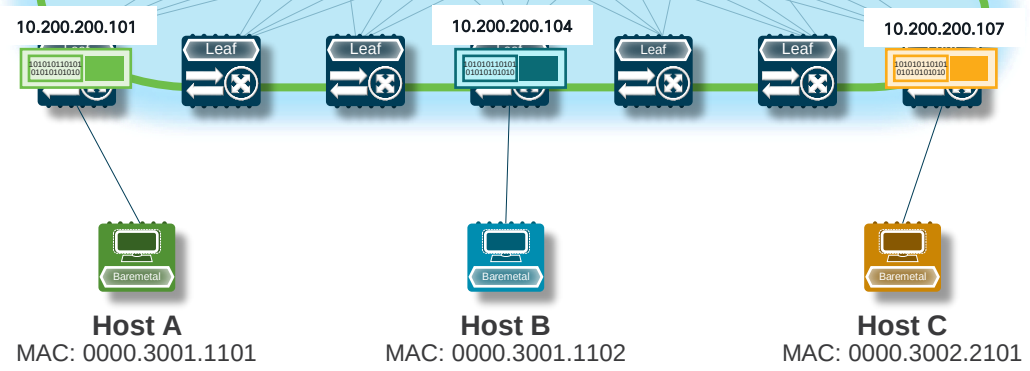
Route Type 5 : IP Prefix Route

- Internal and External Subnet Prefixes
  - IP Subnet Prefix, Single VNI, Single Route Target, Router MAC

*NLRI: Network Layer Reachability Information (BGP Update Format)

Host Advertisements (MAC Only)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / R
2	0000.3001.1101 / 48	3001, 65500:3001		
2	0000.3001.1102 / 48	3001, 65500:3001		
2	0000.3002.2101 / 48	3002, 65500:3002		



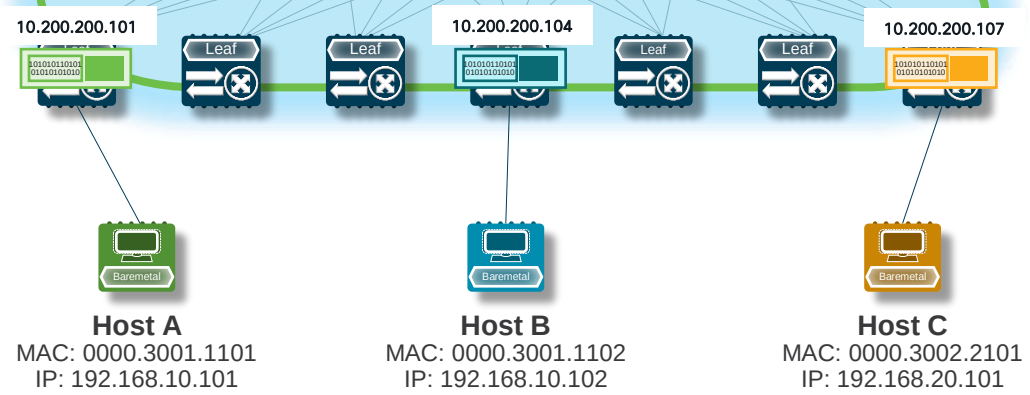
Host MAC (Route Type 2)

- MAC Address
- L2VNI
- Route Target for MAC-VRF

MAC attributes are Mandatory

Host Advertisements (MAC + IP)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / F
2	0000.3001.1101 / 48	3001, 65500:3001	192.168.10.101 /32	5000, 65
2	0000.3001.1102 / 48	3001, 65500:3001	192.168.10.102 /32	5000, 65
2	0000.3002.2101 / 48	3002, 65500:3002	192.168.20.101 /32	5000, 65



Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

Host MAC+IP (Route Type-2)

- MAC
- IP Address
- L2VNI
- Route Target for MAC-VRF
- L3VNI
- Route Target for IP-VRF
- Router MAC

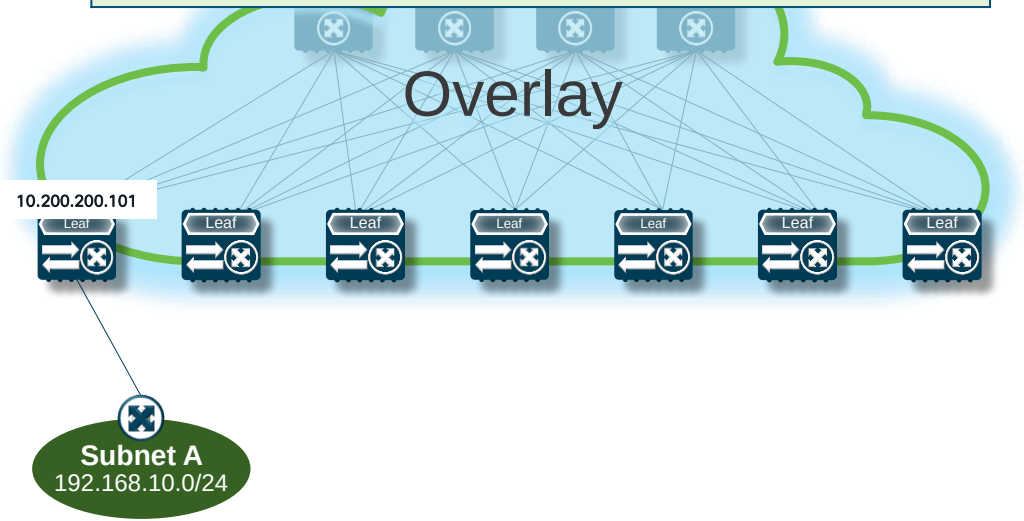
IP Attribute is Optional

Populated through ARP/ND

*L3VNI: VNI for Routing operation (”VRF-VNI”)

Subnet Route Advertisements

Type	IP / Length	L3VNI / RT	Next-Hop	Seq.
5	192.168.10.0 /24	5000, 65500:5000	10.200.200.101	

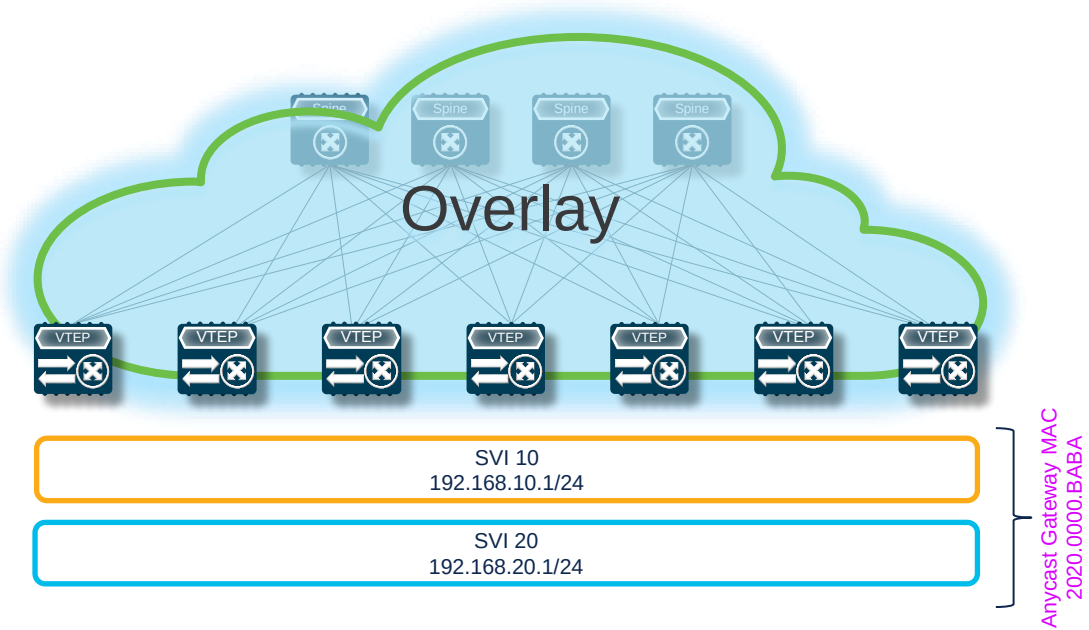


Internal and External Subnet Prefixes (Route Type-5)

- IP Prefix
- L3VNI
- Route Target for IP-VRF
- Router MAC

Populated through External Routing Protocol

Distributed Anycast Gateway (DAG)



Distributed First-Hop Routing on Edge Device

- All Edge Device share same Gateway IP and MAC address
- Pervasive Gateway approach

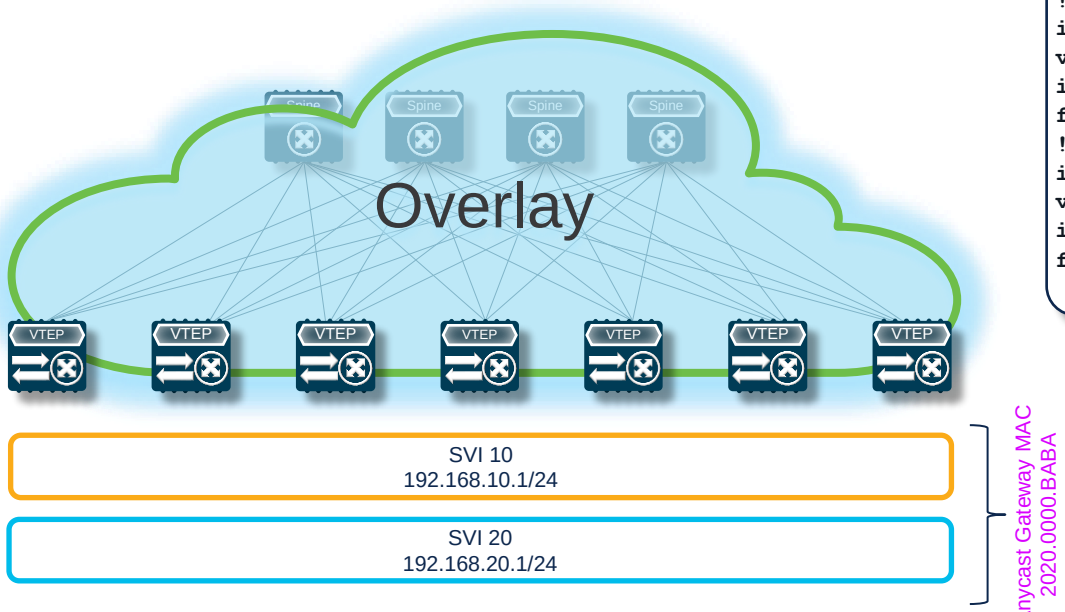
Gateway is always active

- No First-Hop redundancy protocol for hello or state exchange

Distributed and smaller state

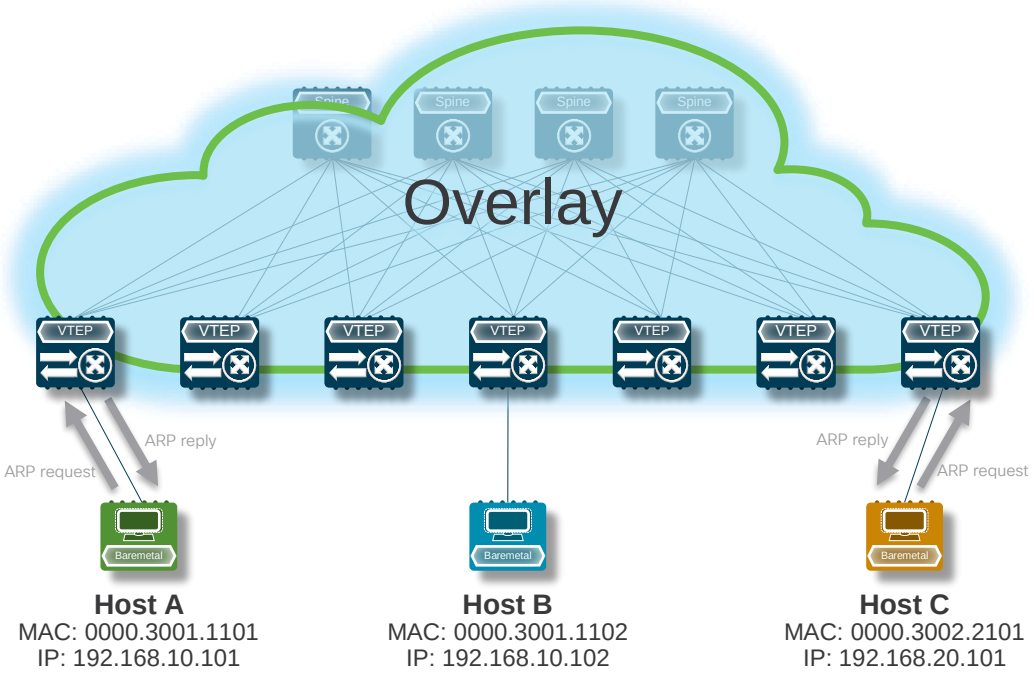
- Only local End-Points ARP entries

Distributed Anycast Gateway (DAG)




```
fabric forwarding anycast-gateway-mac 2020.0000.BABA
!
interface Vlan10
vrf member myvrf_5000
ip address 192.168.10.1/24
fabric forwarding mode anycast-gateway
!
interface Vlan20
vrf member myvrf_5000
ip address 192.168.20.1/24
fabric forwarding mode anycast-gateway
```

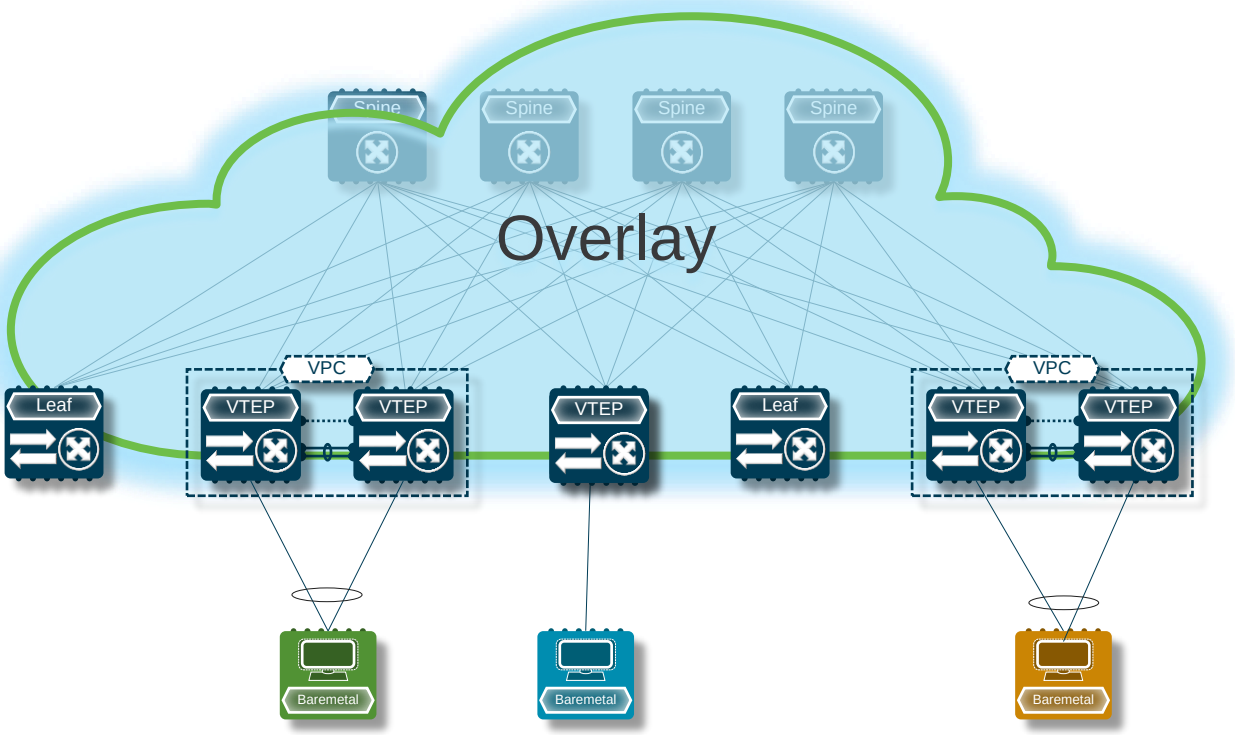
DAG- Local ARP resolution



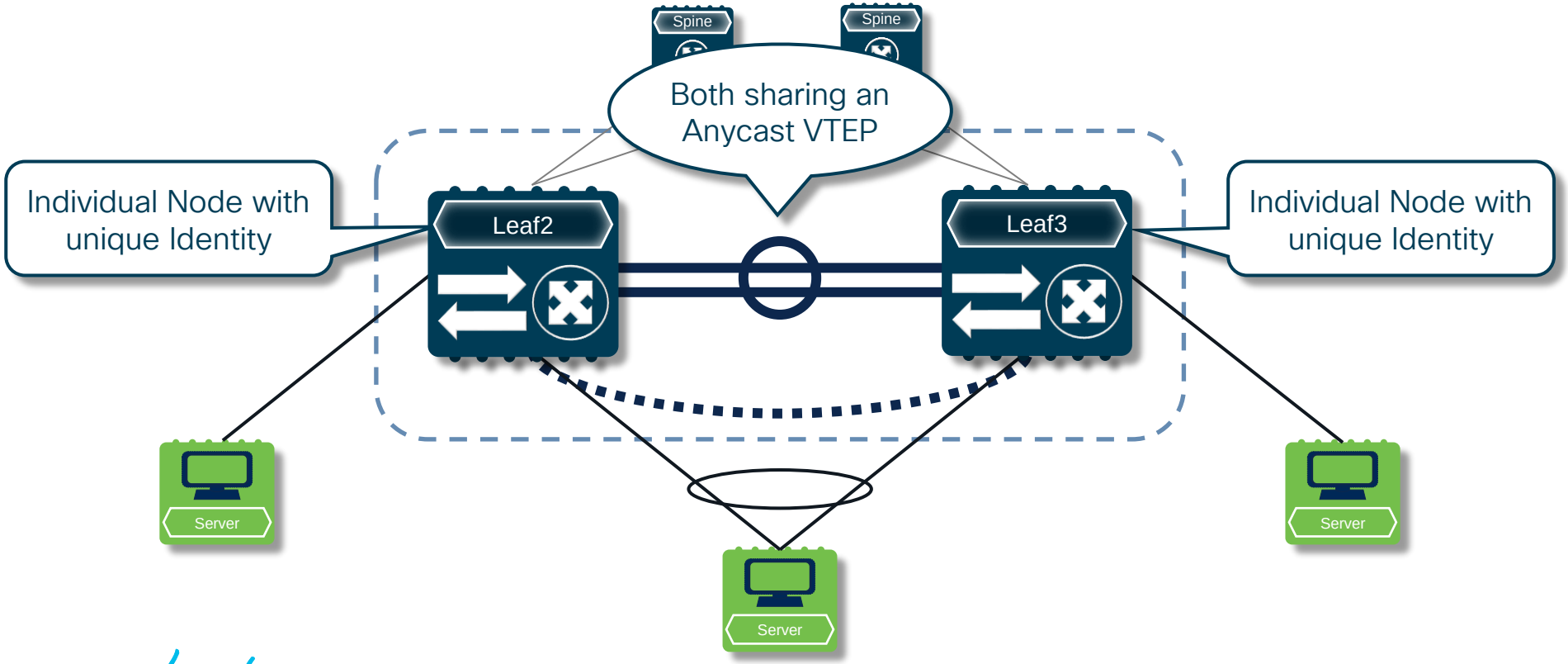
Local Ethernet Segment-based ARP Resolution for First-Hop Gateway

vPC in VXLAN

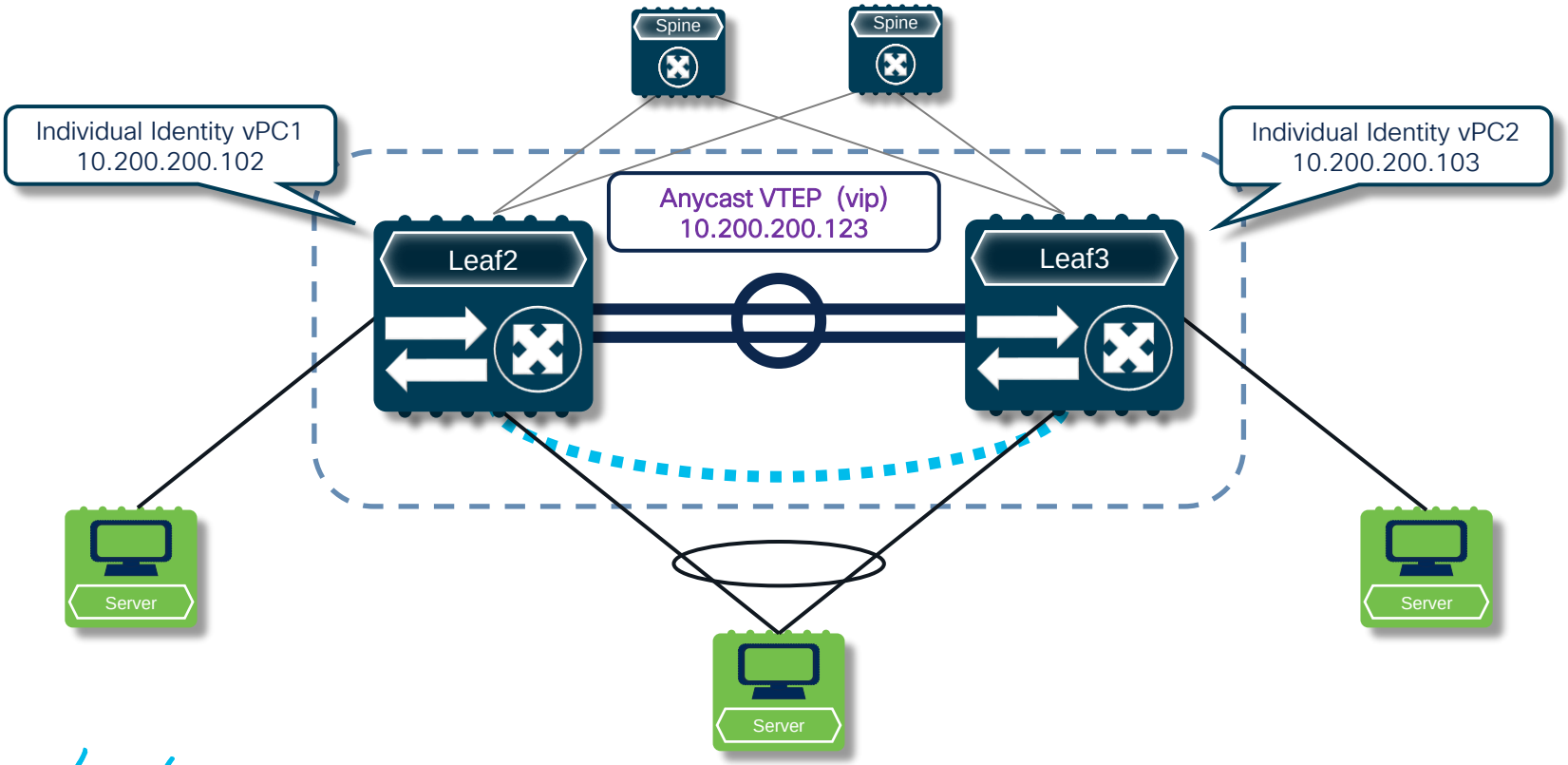
VXLAN – vPC



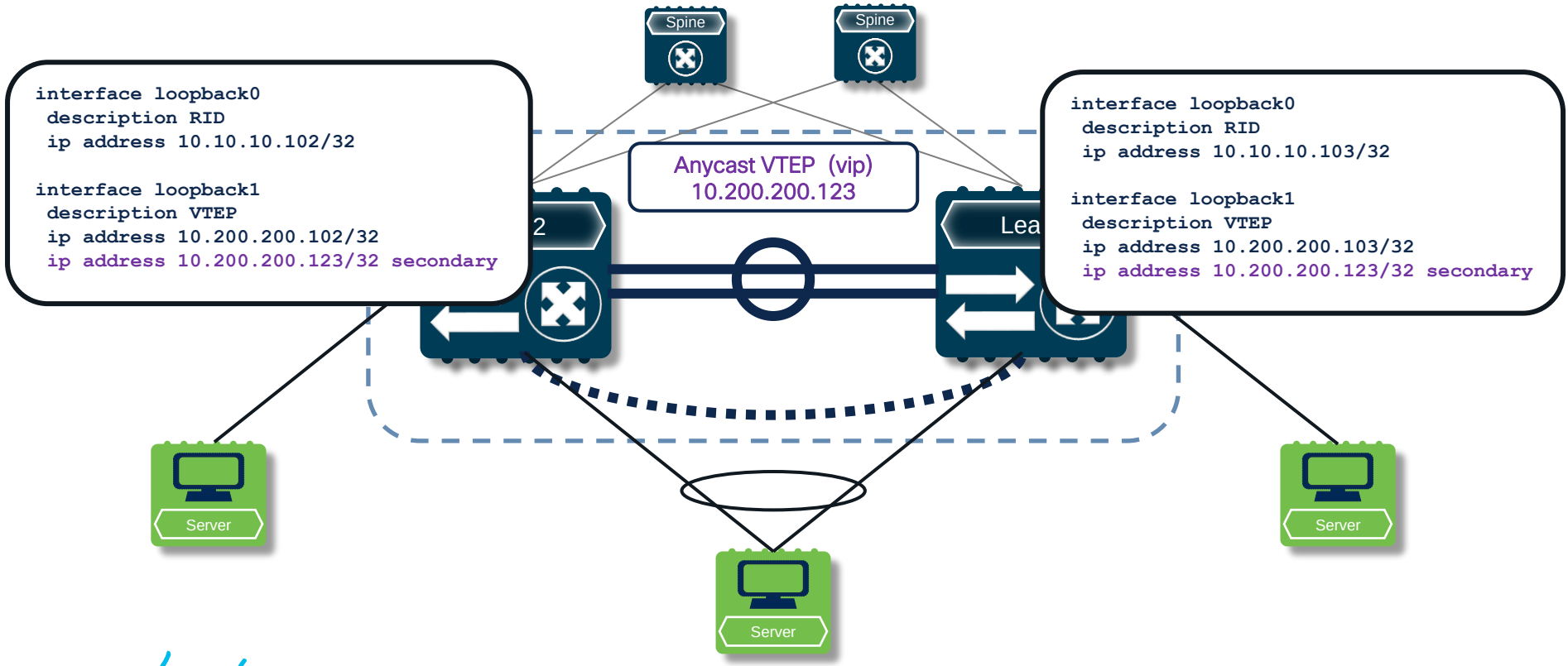
VXLAN – Anycast VTEP



Anycast VTEP and Individual VTEPs



Anycast VTEP – Virtual IP Address

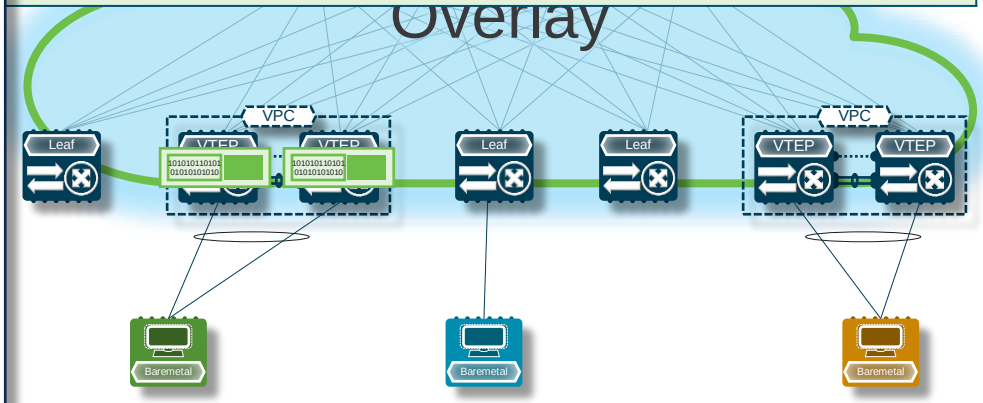


Host Advertisements with vPC – BGP EVPN

Remember the RD?

- Independent Devices in the EVPN Control-Plane
  - Individual Router and Peering
  - Unique Route Distinguisher (RD)
  - Independent Underlay Routing Devices
- Common VXLAN Device
  - Next-Hop is Anycast VTEP
  - Underlay ECMP Load Share to Anycast VTEP

NI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
1, 65500:3001	192.168.10.101 /32	5000, 65500:5000	10.200.200.123	
1, 65500:3001	192.168.10.101 /32	5000, 65500:5000	10.200.200.123	

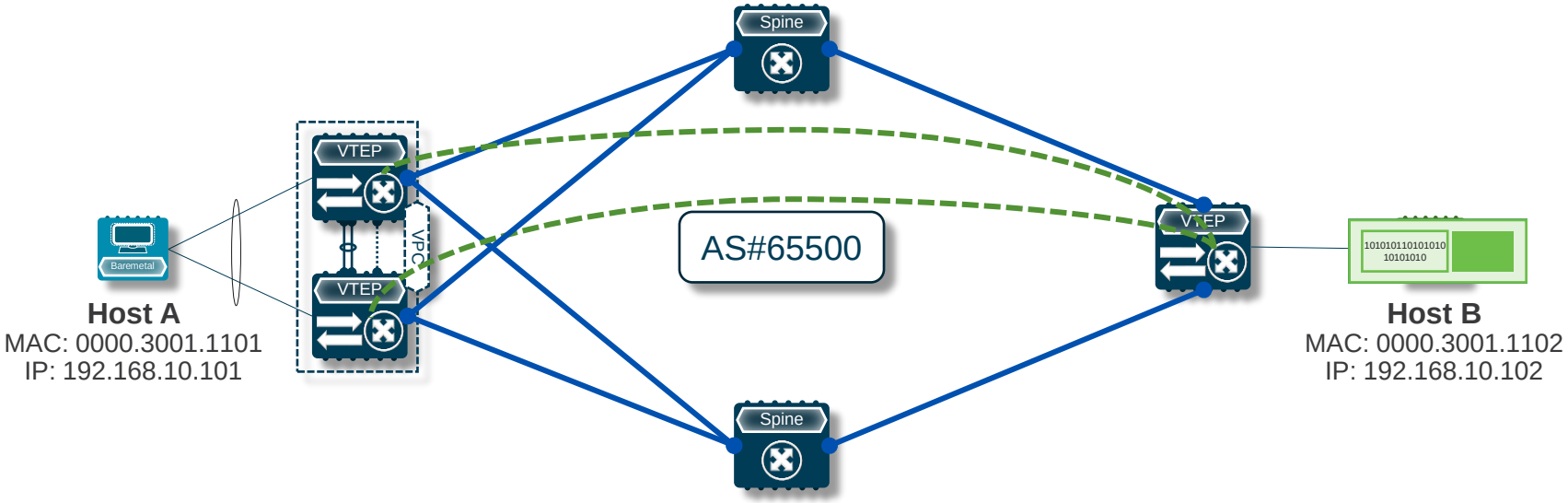


Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

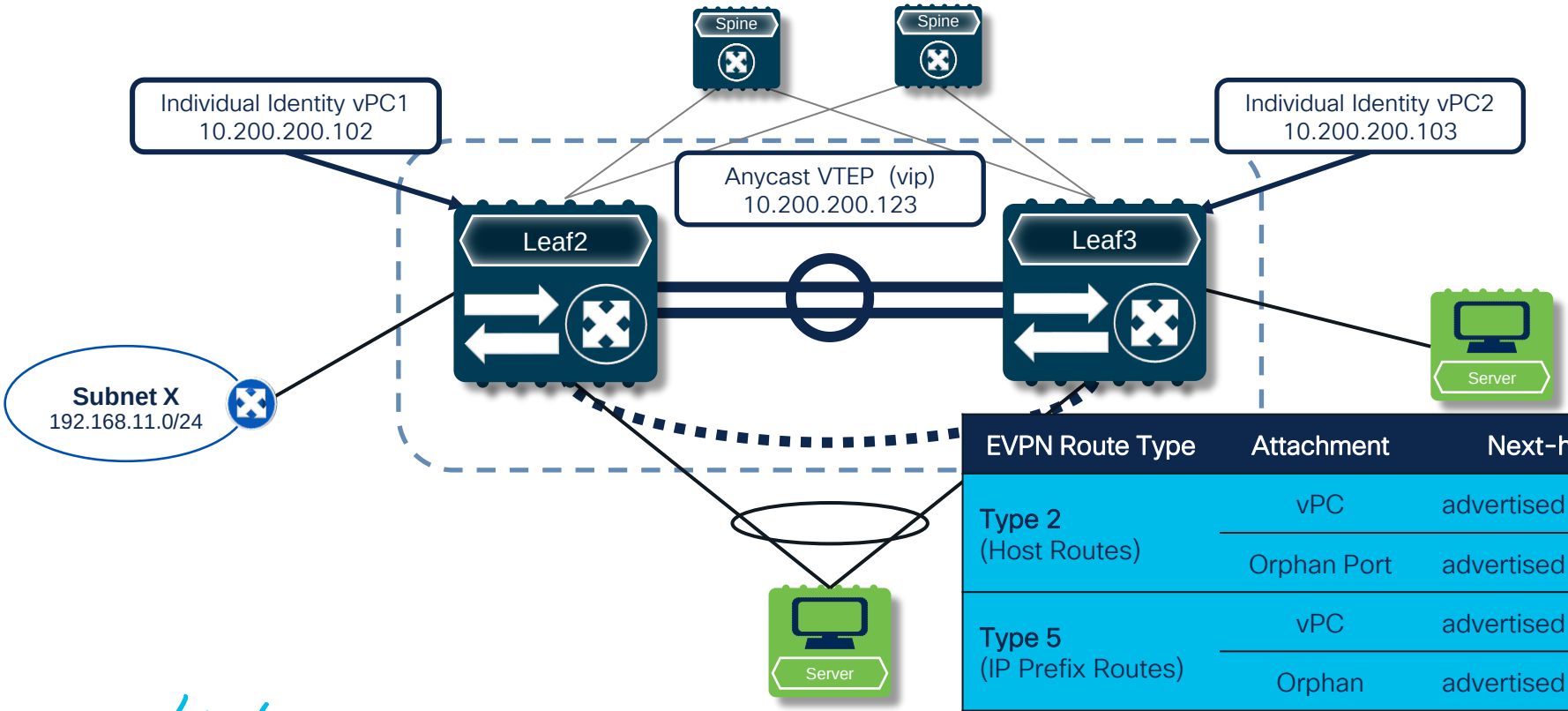
Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

ECMP to the Anycast VTEP – Underlay



Host and subnet route advertisement with Anycast VTEP

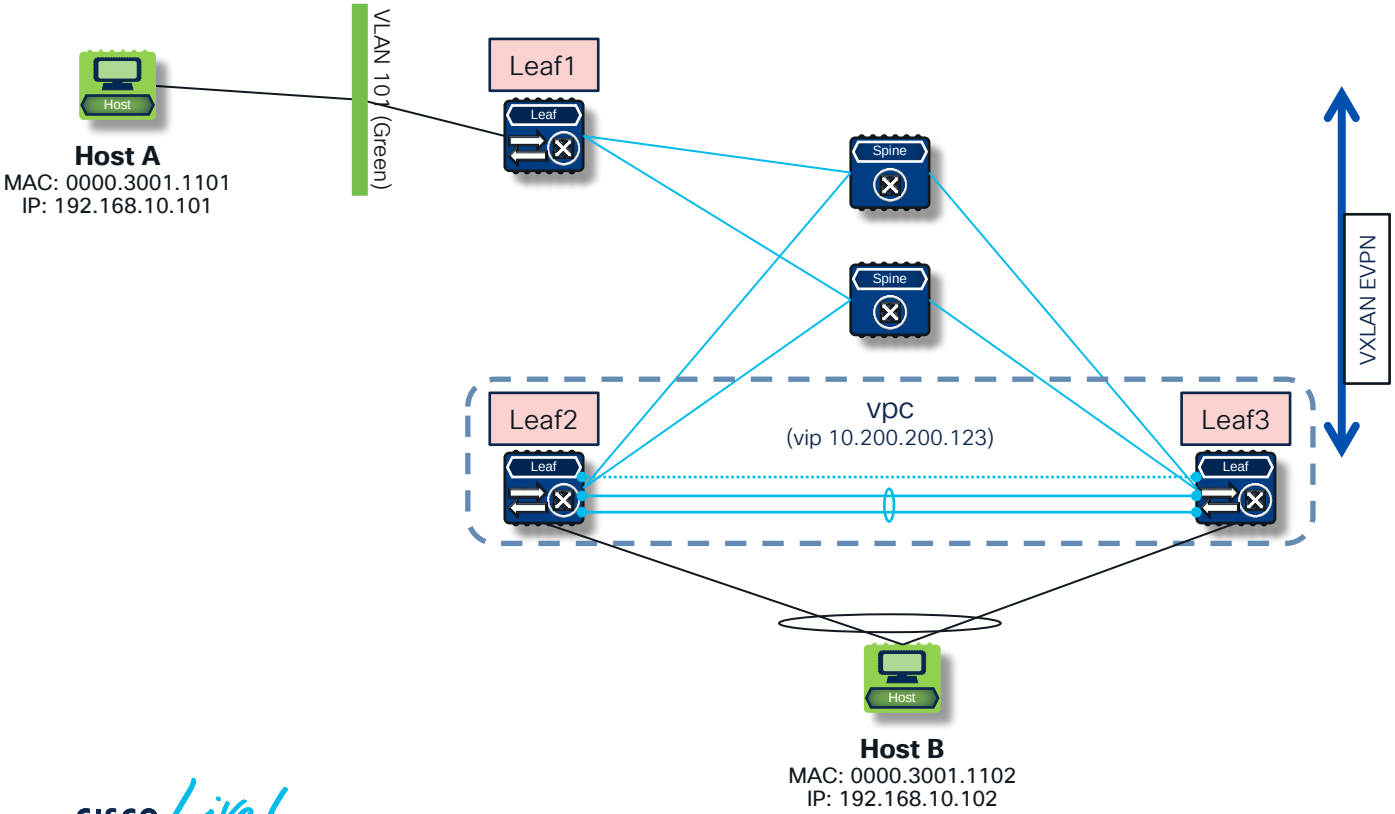


EVPN Route Type	Attachment	Next-hop
Type 2 (Host Routes)	vPC	advertised by VIP
	Orphan Port	advertised by VIP
Type 5 (IP Prefix Routes)	vPC	advertised by VIP
	Orphan	advertised by VIP

VXLAN vPC packet walk

Bridging Packet Walk –vPC

vPC Host

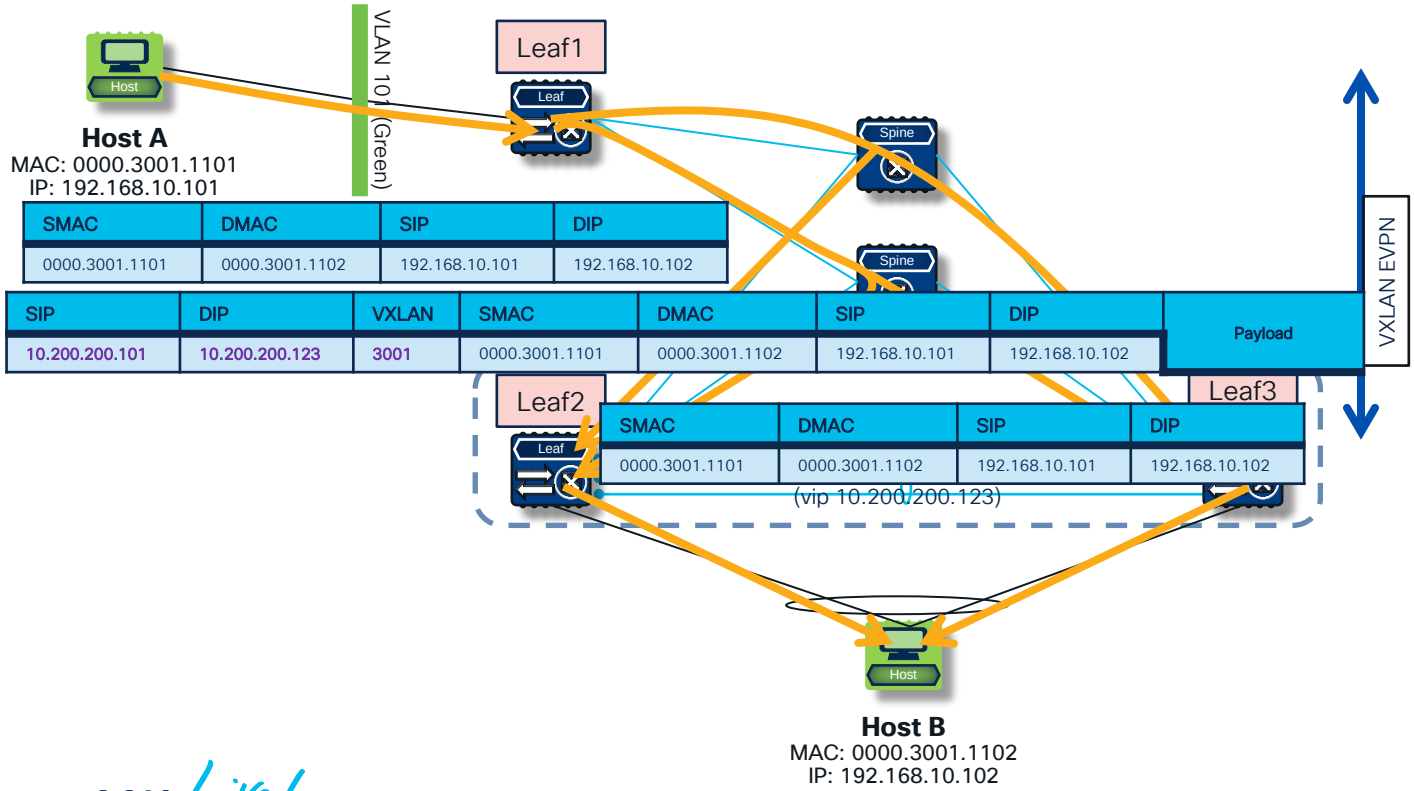


EVPN Control-Plane	
RD	10.10.10.102:32777
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123

EVPN Control-Plane	
RD	10.10.10.103:32777
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123

Bridging Packet Walk –vPC

vPC Host

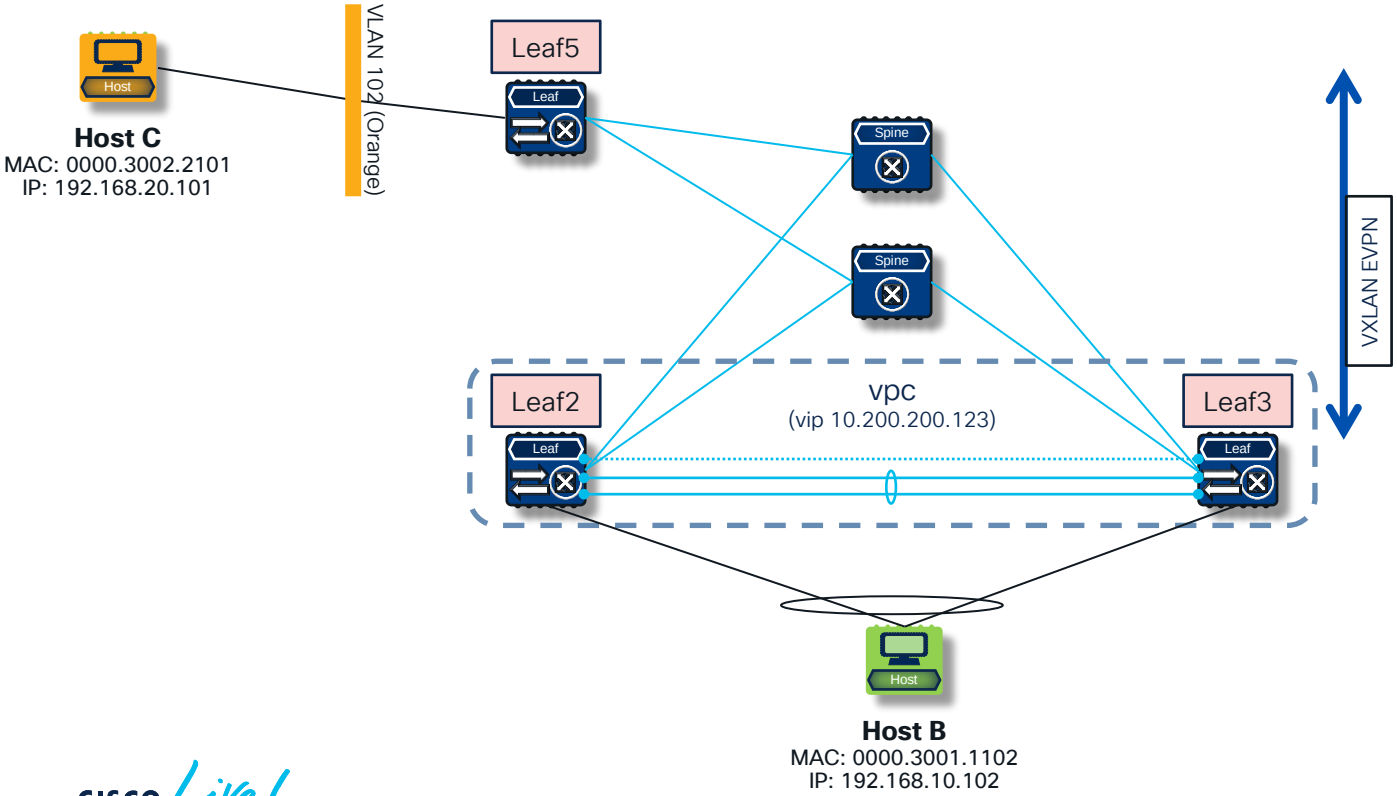


EVPN Control-Plane	
RD	10.10.10.102:32777
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123

EVPN Control-Plane	
RD	10.10.10.103:32777
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123

Routing Packet Walk –vPC

vPC Host

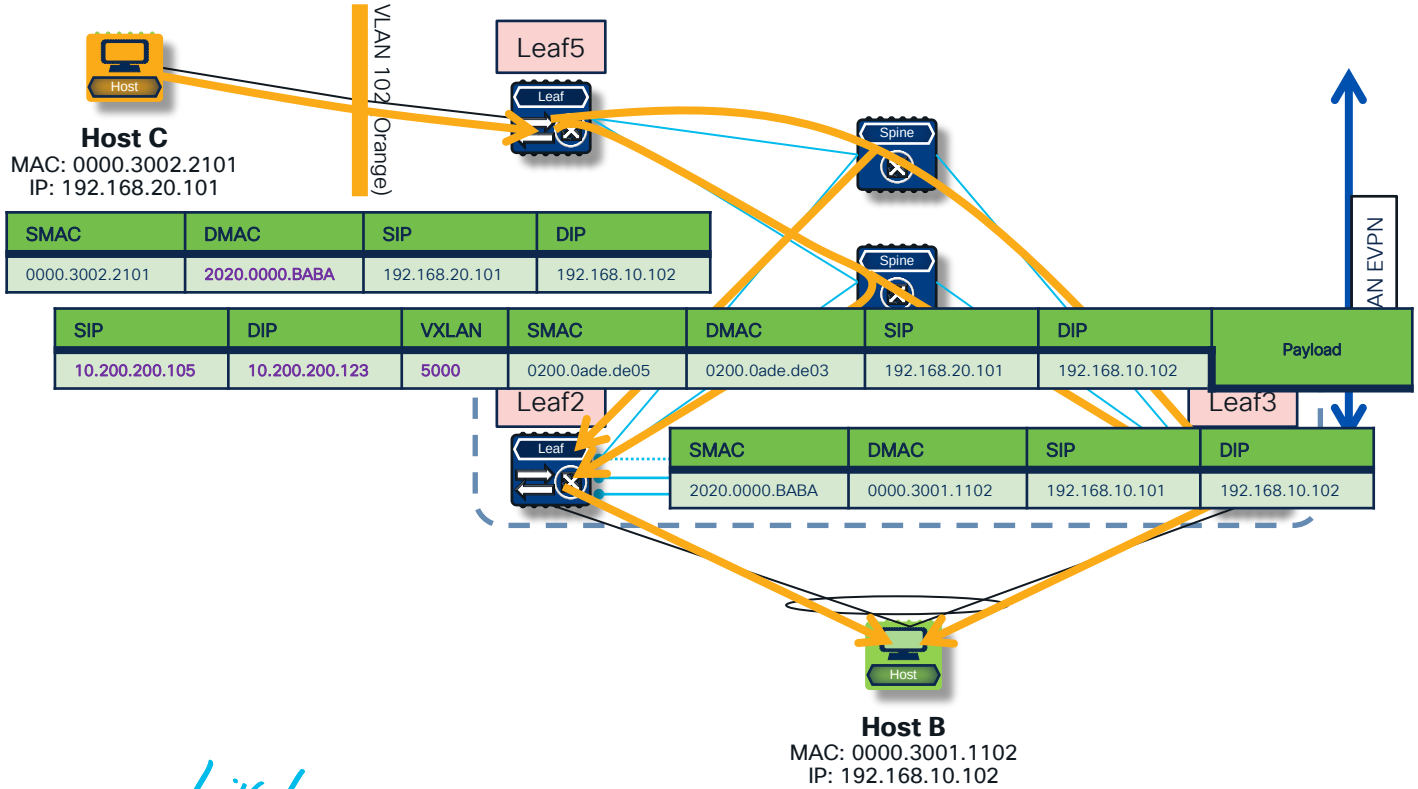


EVPN Control-Plane	
RD	10.10.10.102:4
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de02

EVPN Control-Plane	
RD	10.10.10.103:4
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de03

Routing Packet Walk –vPC

vPC Host



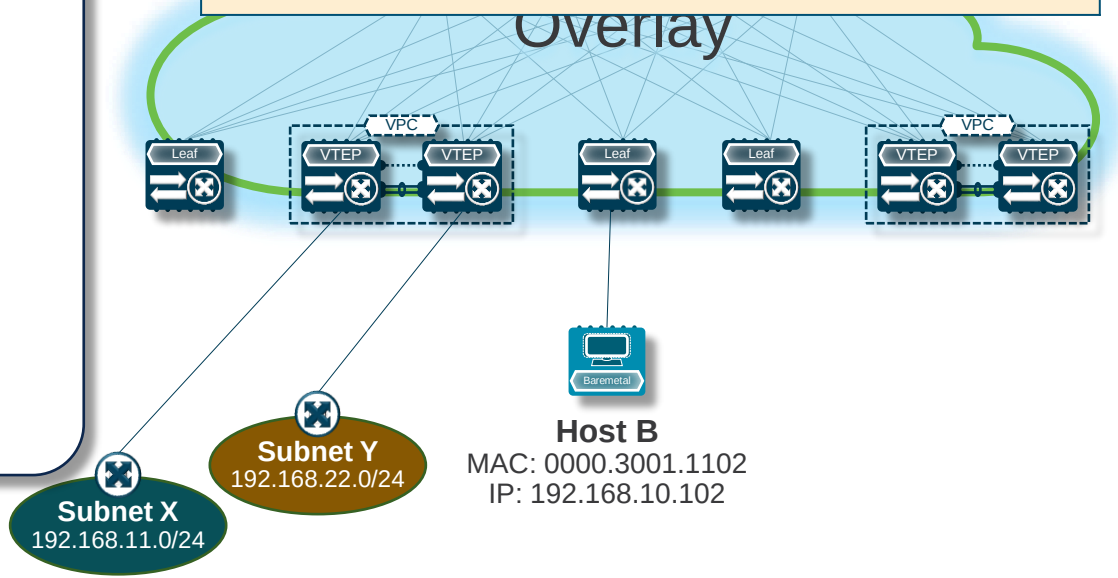
EVPN Control-Plane	
RD	10.10.10.102:4
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de02

EVPN Control-Plane	
RD	10.10.10.103:4
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de03

Subnet Route Advertisement with vPC

- Subnet Route Advertisement
  - Route Type 5
  - Next-Hop is Anycast VTEP
- Ensure Sync of Subnet
  - Dual-Connect Networks (Point-2-Point not Layer-3 over vPC)
  - Synchronize Routing Table
  - Can cause inefficient forwarding

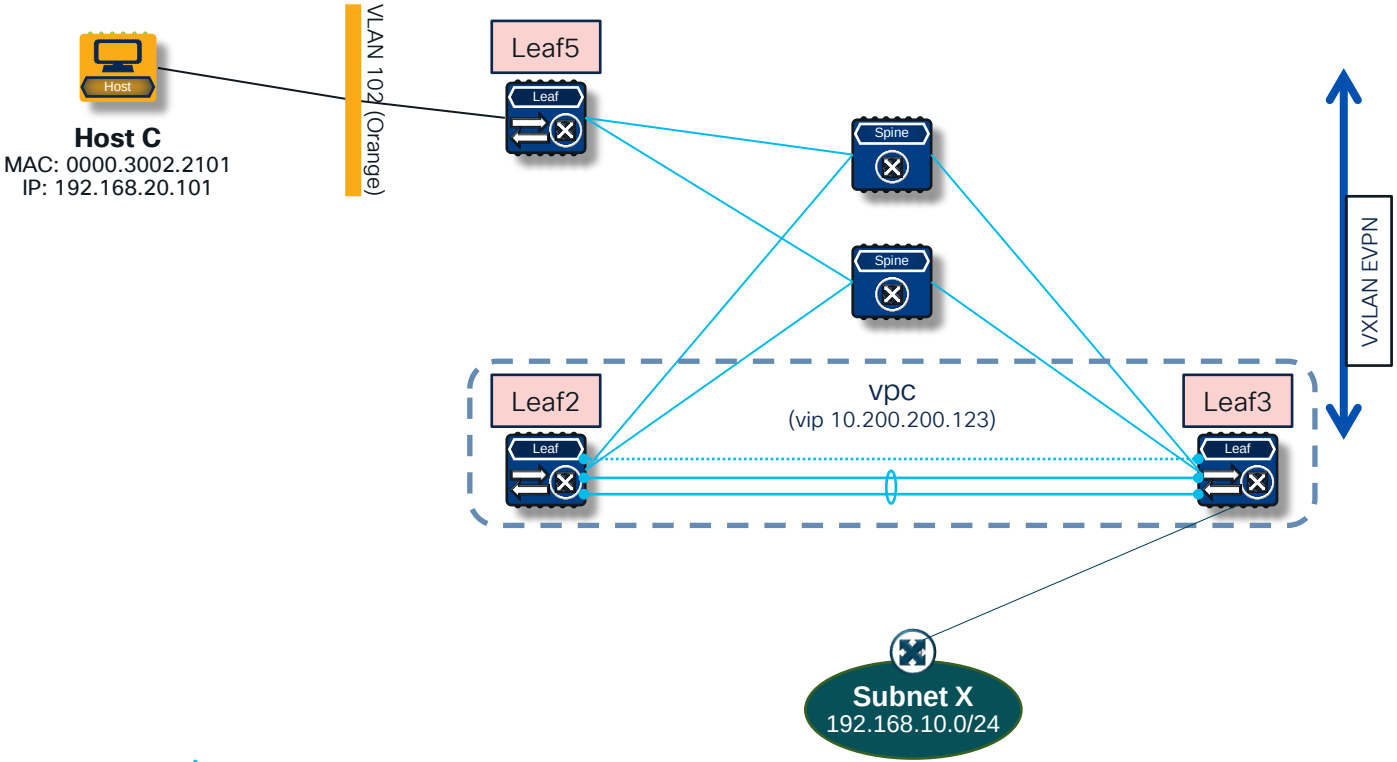
Type	IP / Length	L3VNI / RT	Next-Hop	Seq.
5	192.168.11.0 /24	5000, 65500:5000	10.200.200.123	
5	192.168.22.0 /24	5000, 65500:5000	10.200.200.123	



Routing Packet Walk –vPC

Subnet route

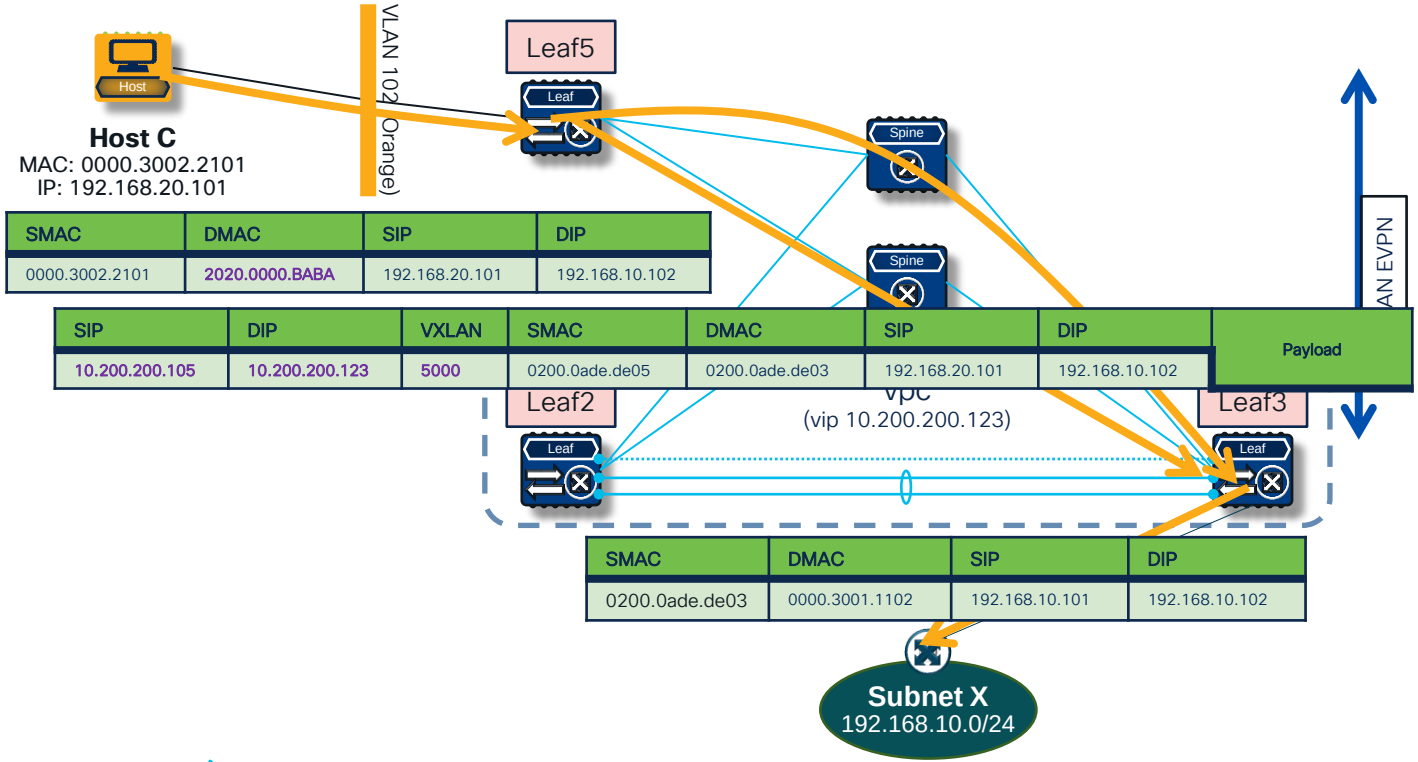
EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de03



Routing Packet Walk –vPC

Subnet route

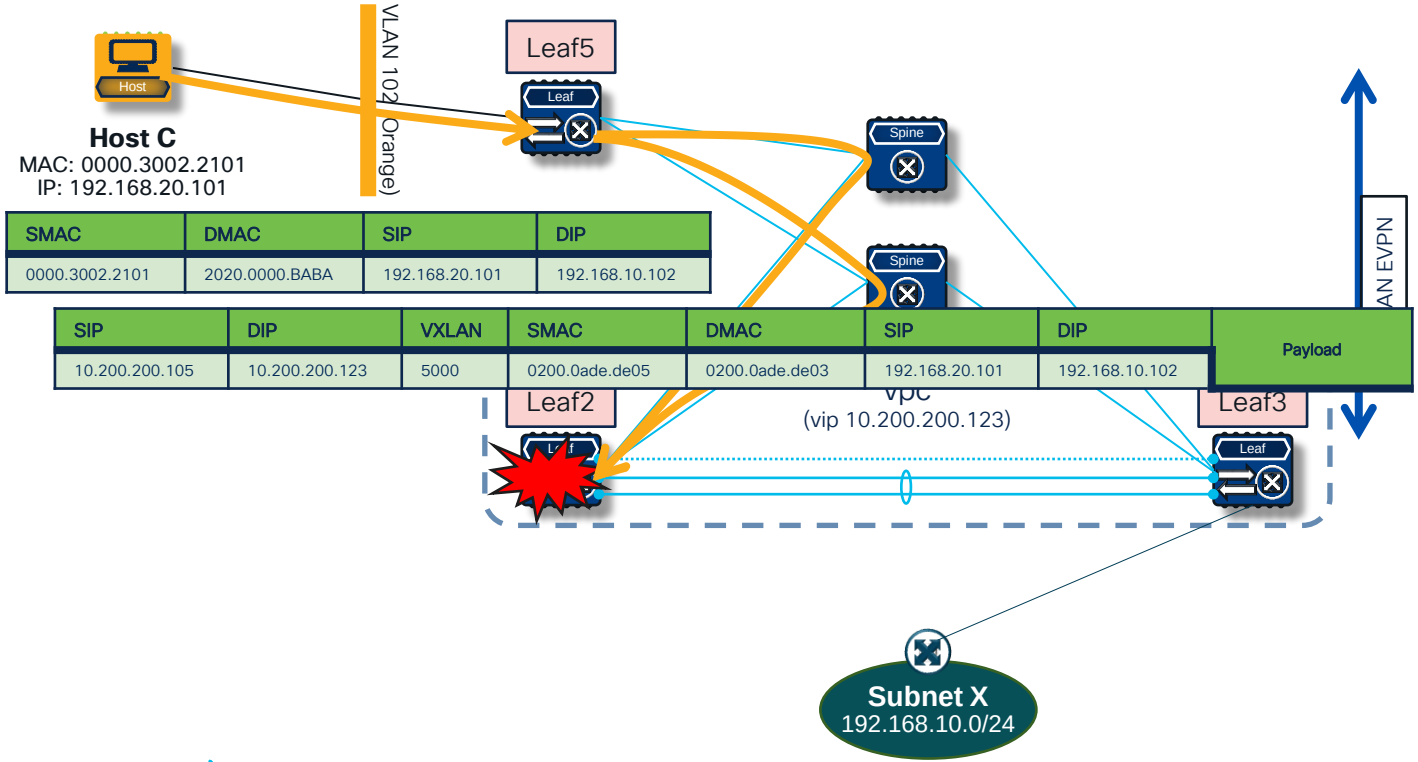
EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de03



Routing Packet Walk –vPC

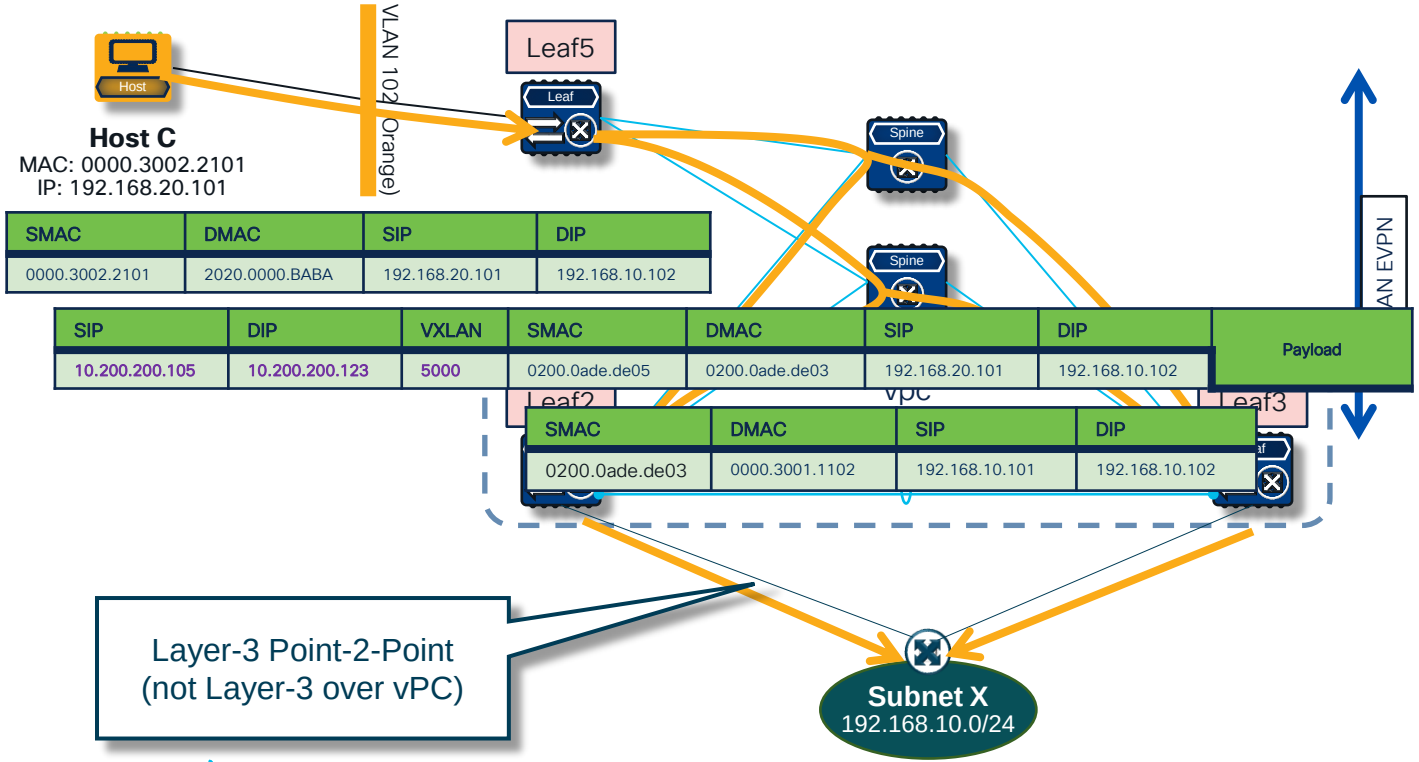
Subnet route

EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de03



Routing Packet Walk –vPC

Subnet route



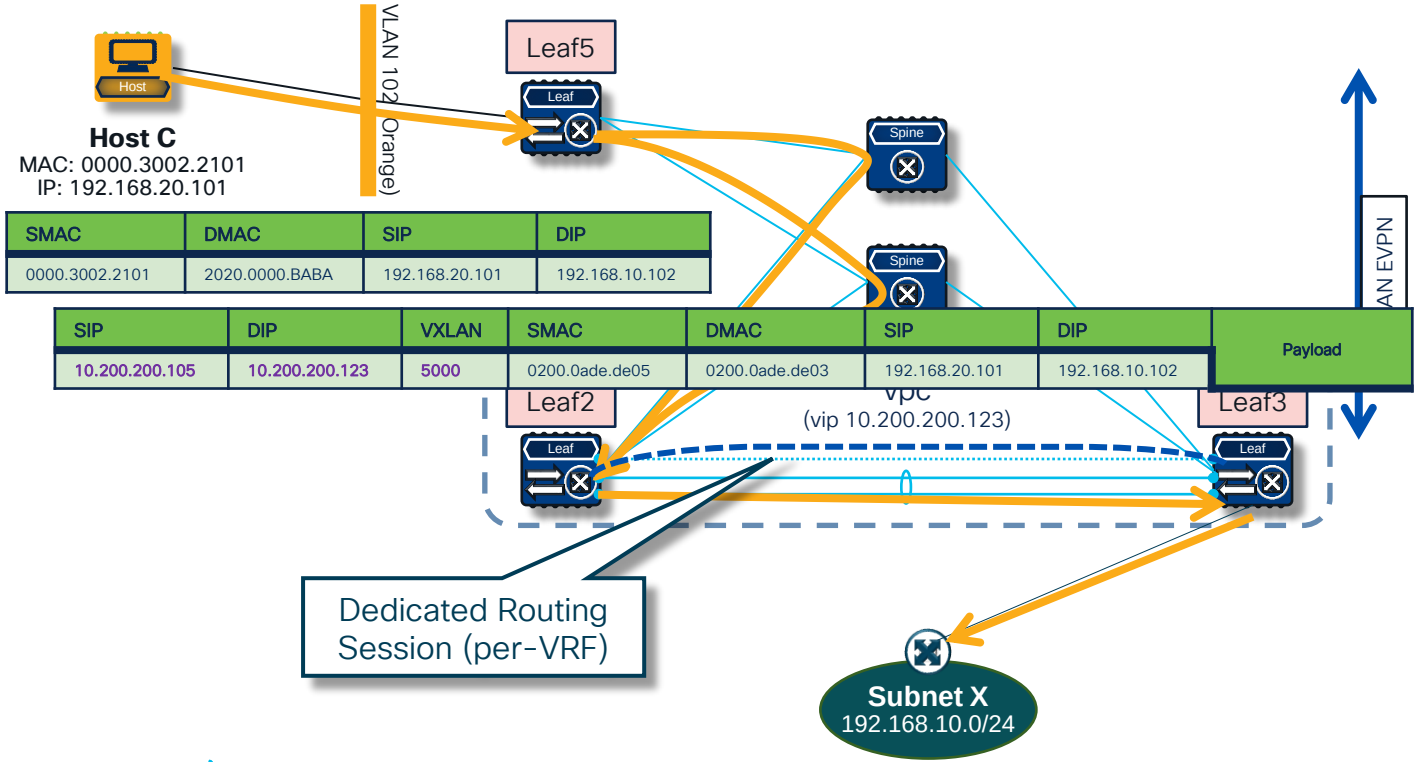
EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de03

EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de02

Routing Packet Walk –vPC

Subnet route

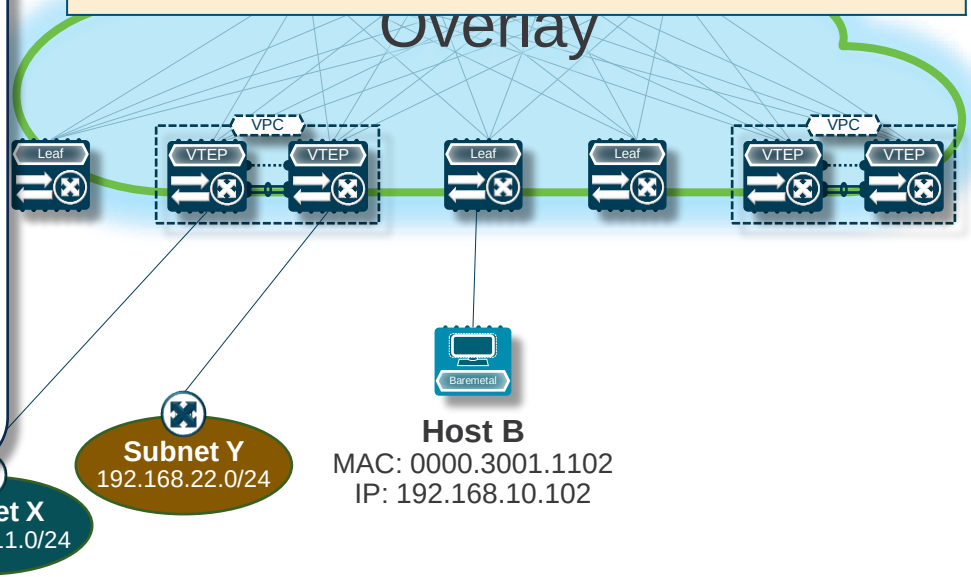
EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ade.de03



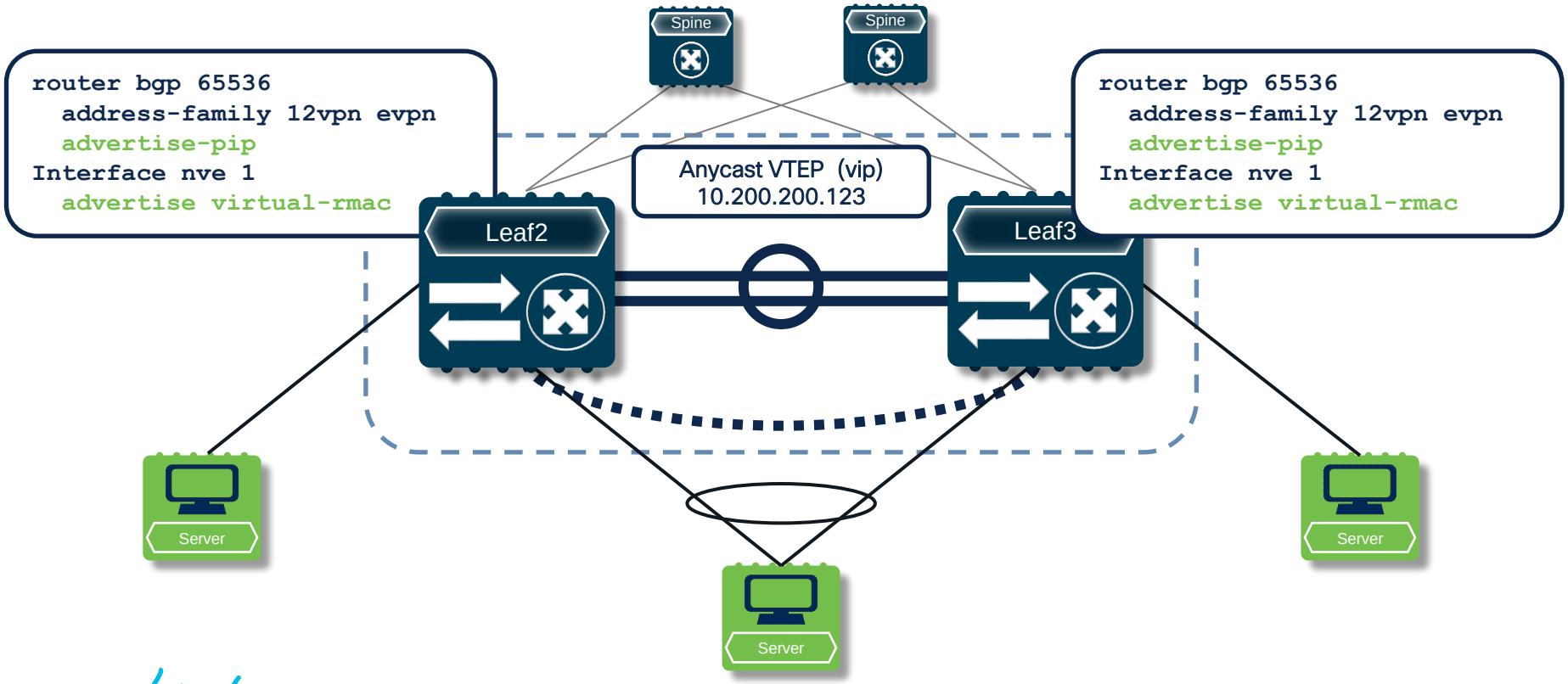
Advertise Primary IP Address

- Subnet Route Advertisement
  - Route Type 5
  - Next-Hop is individual VTEP
- Advertise Route Type 5 with individual VTEP IP (PIP)
- Prevent usage of peer link for traffic hashed to remote peer
- Routes will be advertised with RMAC address

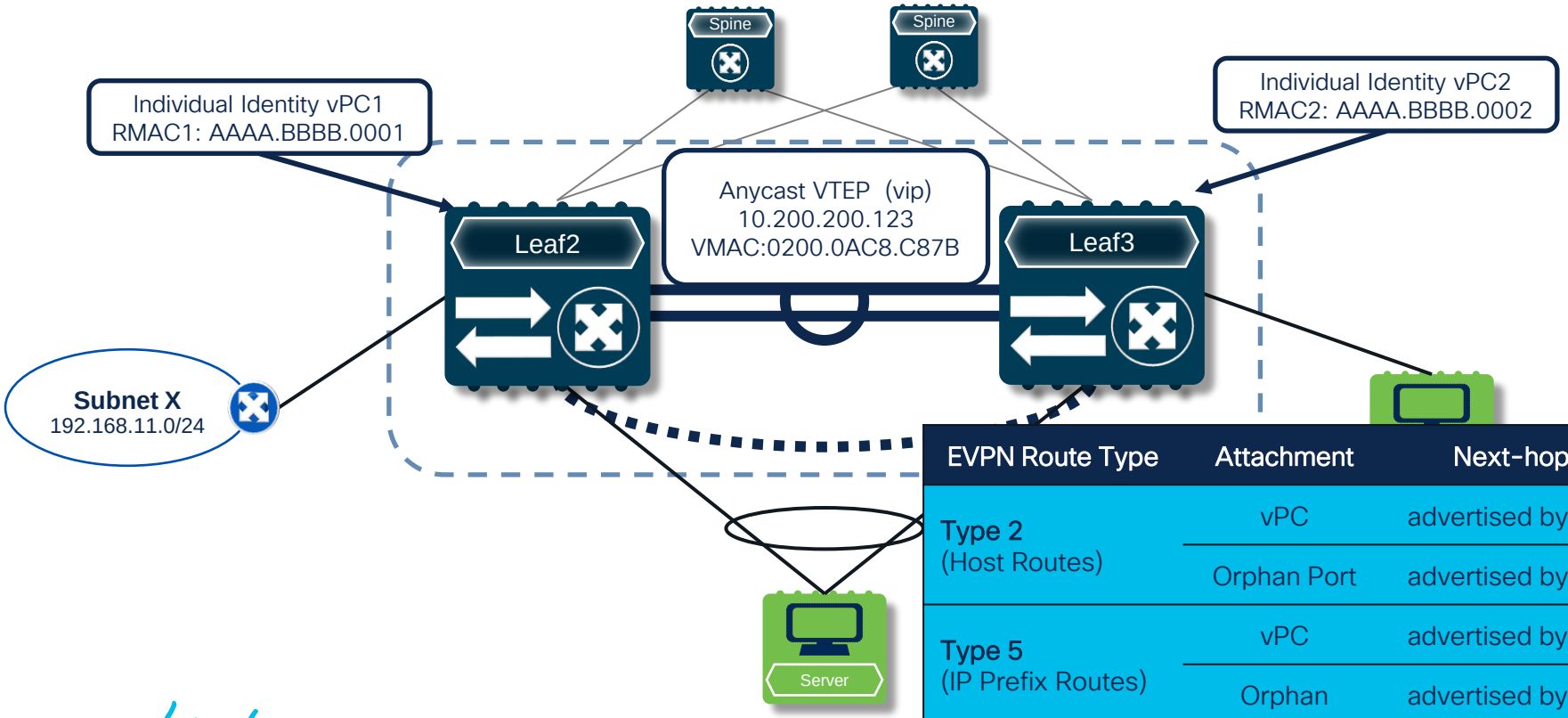
Type	IP / Length	L3VNI / RT	Next-Hop	Seq.
5	192.168.11.0 /24	5000, 65500:5000	10.200.200.102	
5	192.168.22.0 /24	5000, 65500:5000	10.200.200.103	



Advertise Primary IP Address



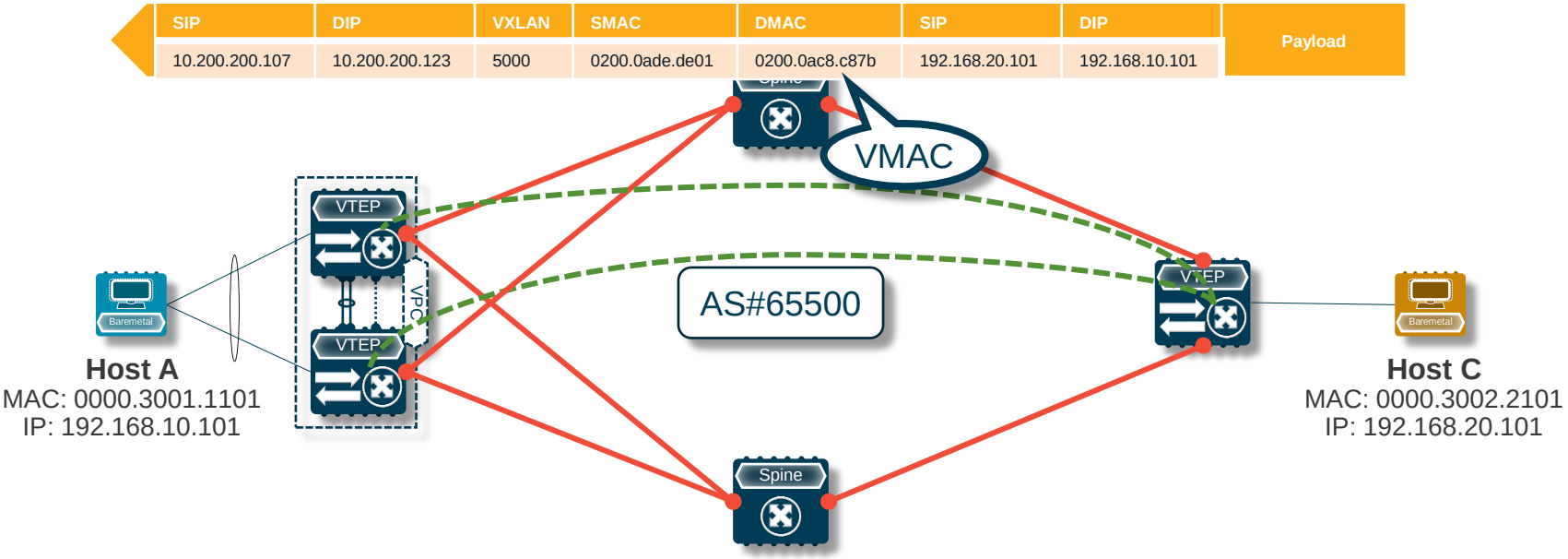
Host and subnet route advertisement with “advertise-PIP”



EVPN Route Type	Attachment	Next-hop
Type 2 (Host Routes)	vPC	advertised by VIP
	Orphan Port	advertised by VIP
Type 5 (IP Prefix Routes)	vPC	advertised by PIP
	Orphan	advertised by PIP

Routing to a vPC Domain – VXLAN

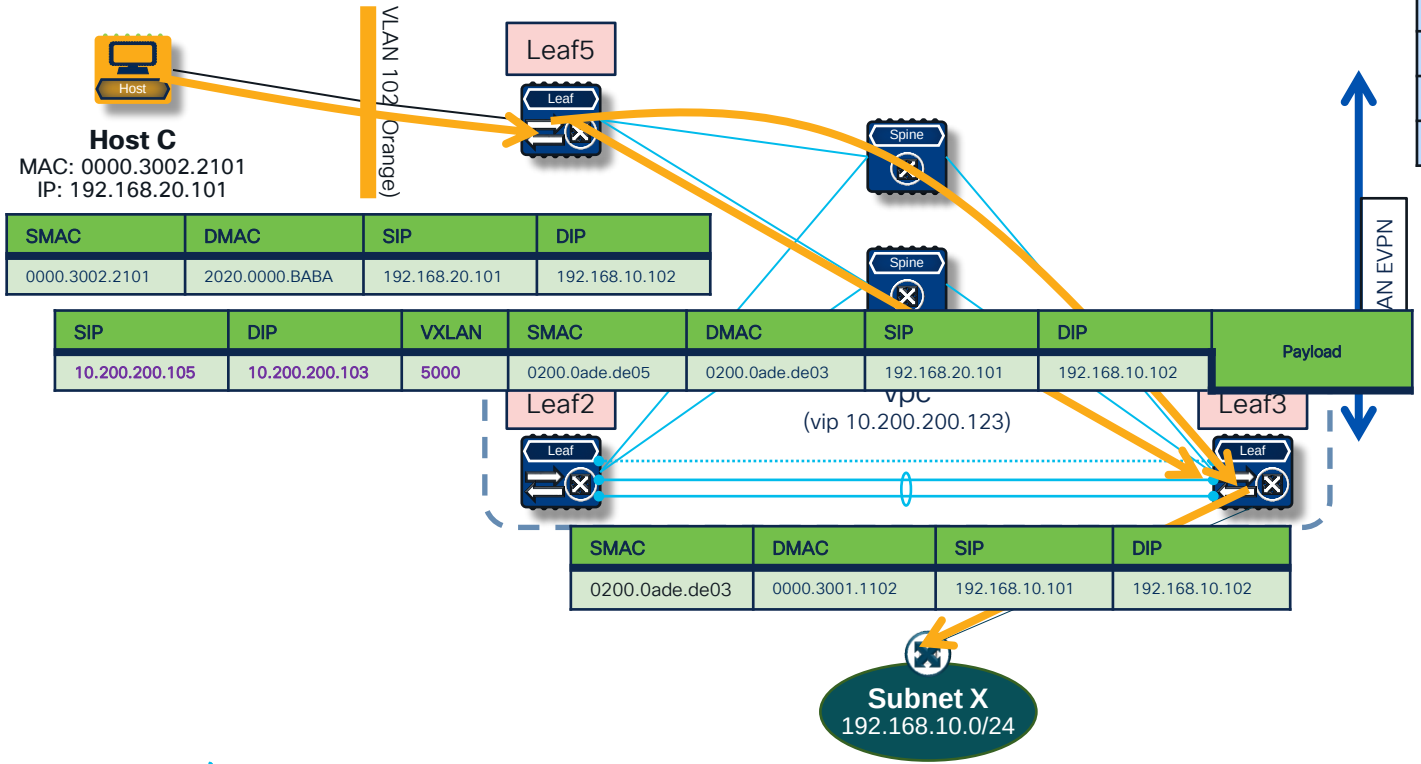
Advertise-PIP



Routing Packet Walk–subnet attached to single peer

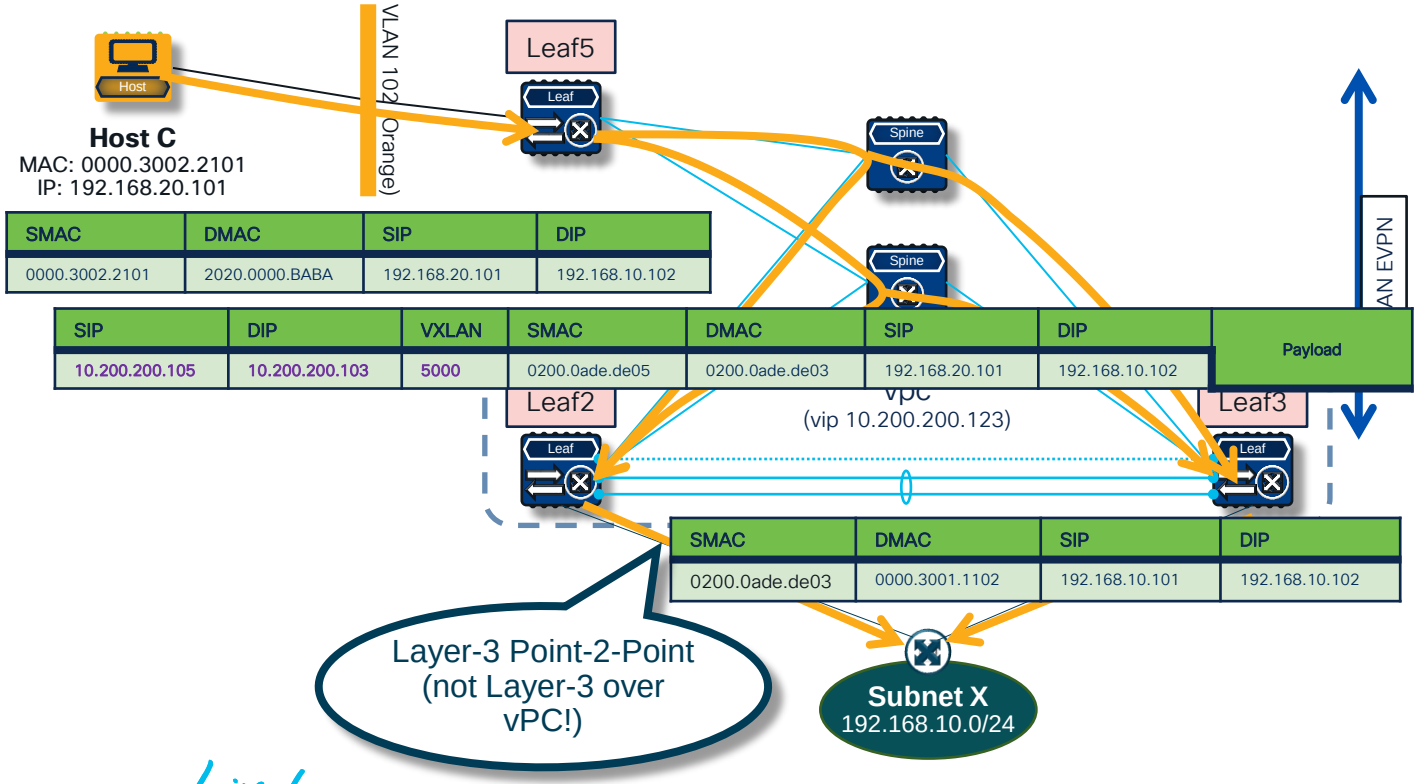
Advertise PIP

EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.103
Ext. Community	0200.0ade.de03



Routing Packet Walk–subnet attached to both peers

Advertise PIP



EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.103
Ext. Community	0200.0ade.de03

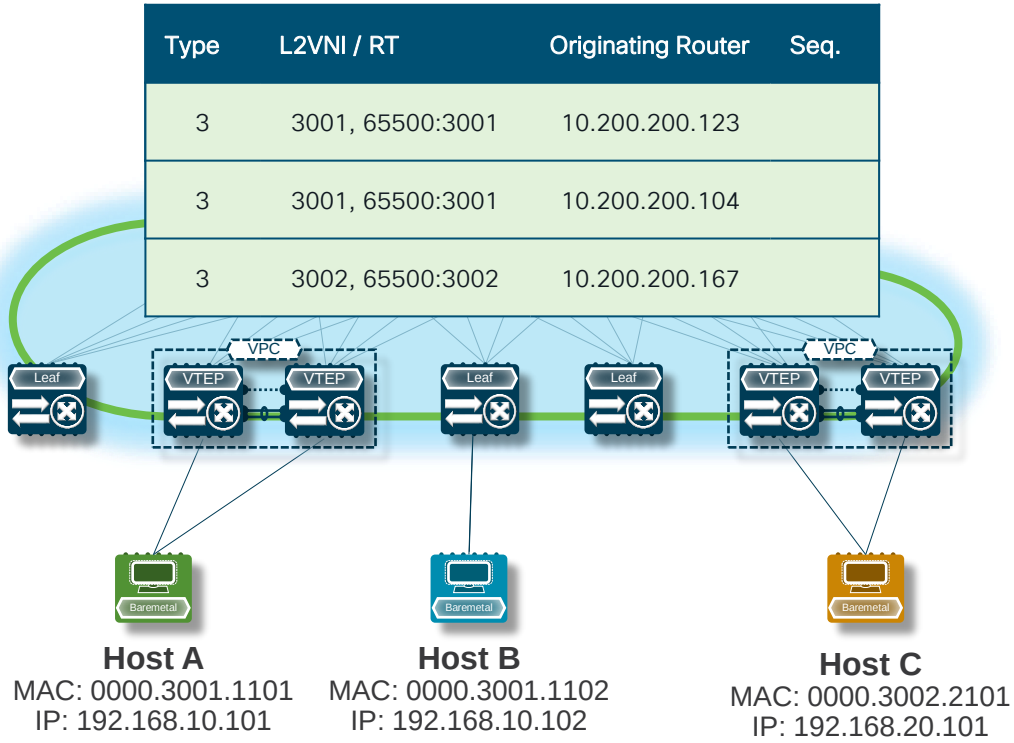
EVPN Control-Plane	
Type	5
IP / Length	192.168.10.0/ 24
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.102
Ext. Community	0200.0ade.de02

BUM Traffic Forwarding with vPC in VXLAN

BUM traffic in vPC

Ingress Replication

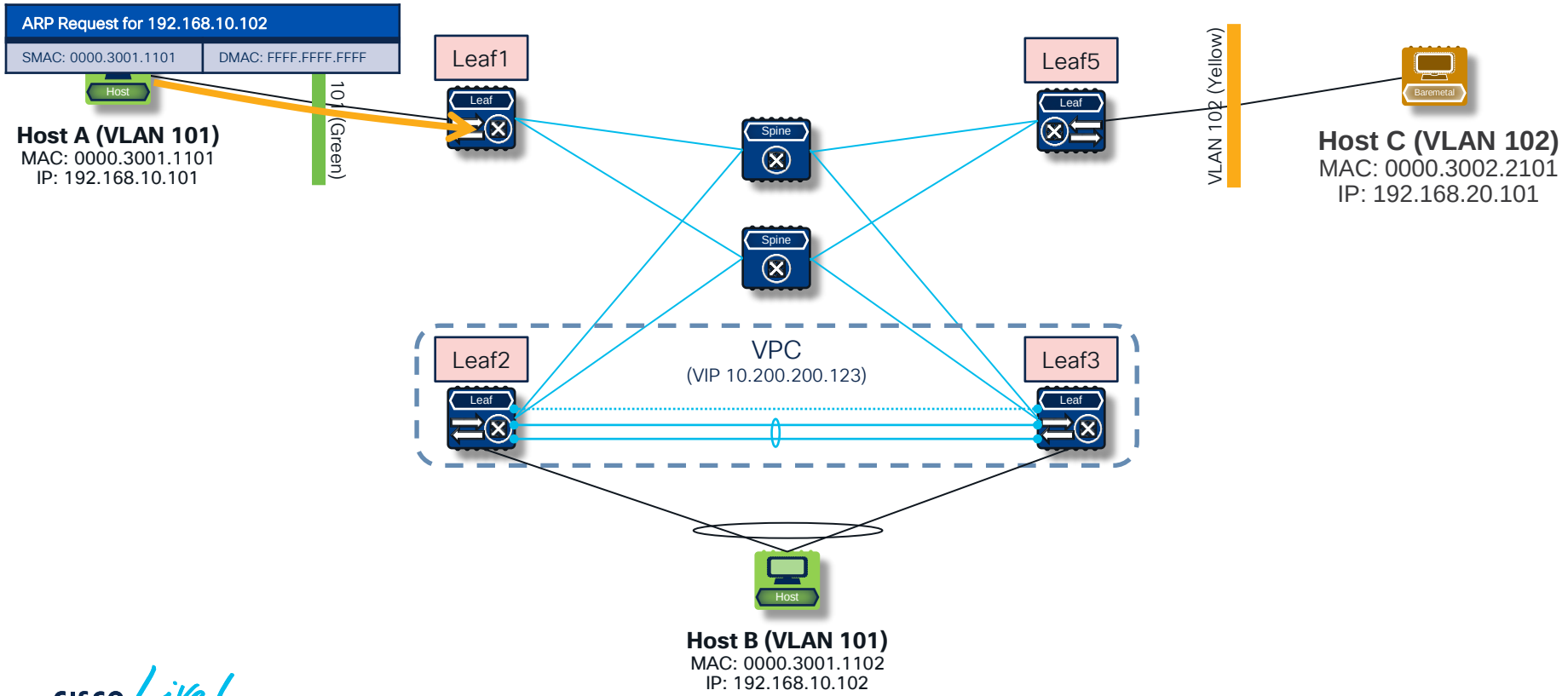
- BUM traffic is replicated at ingress VTEP to each VTEP with same L2VNI
- Dynamic List of VNI to VTEP paring
  - Uses EVPN Type 3 routes
- Anycast VTEP is used as next-hop for vPC attached host
- Can be inefficient for large scale deployments



BUM traffic in vPC

Ingress Replication

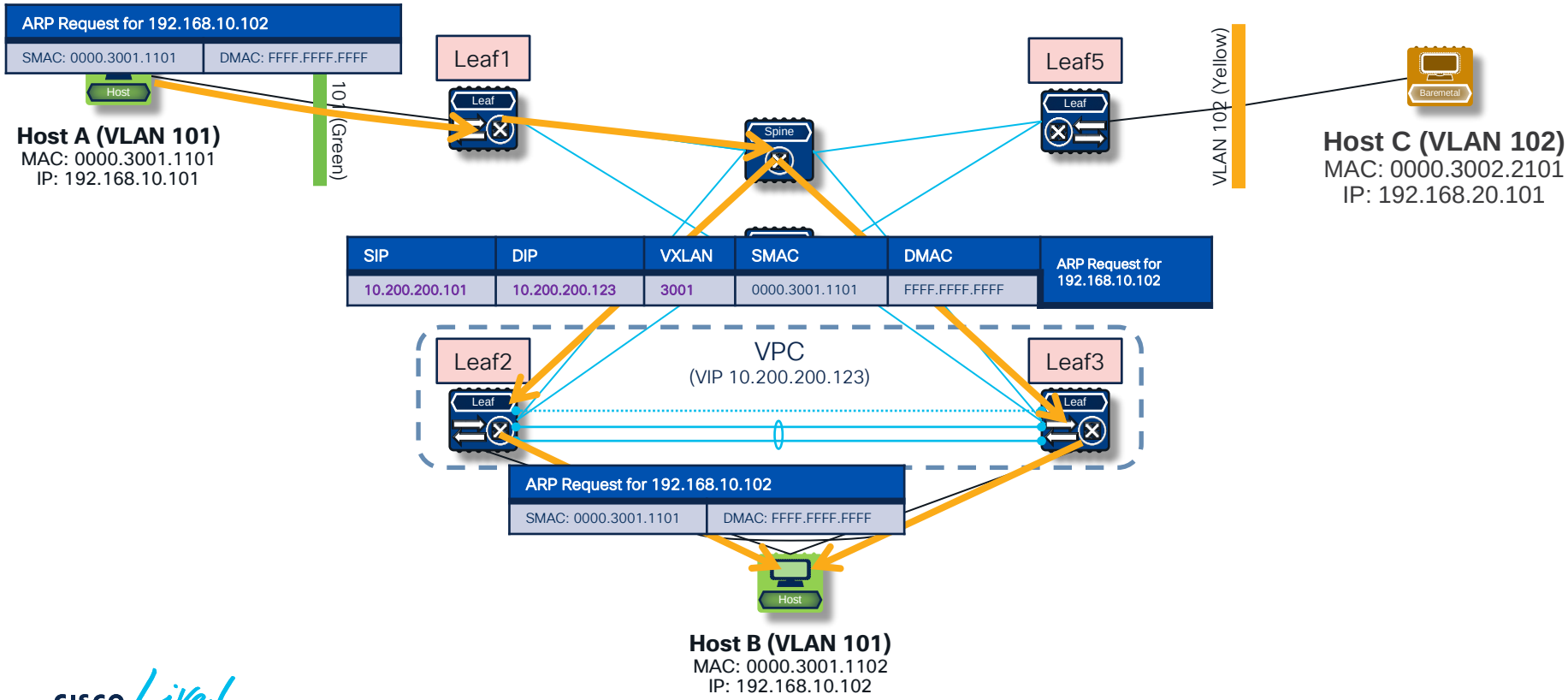
EVPN Control-Plane	
Type	3
L2VNI / RT	3001/ 65500:3001
Origin. Router	10.200.200.123



BUM traffic in vPC

Ingress Replication

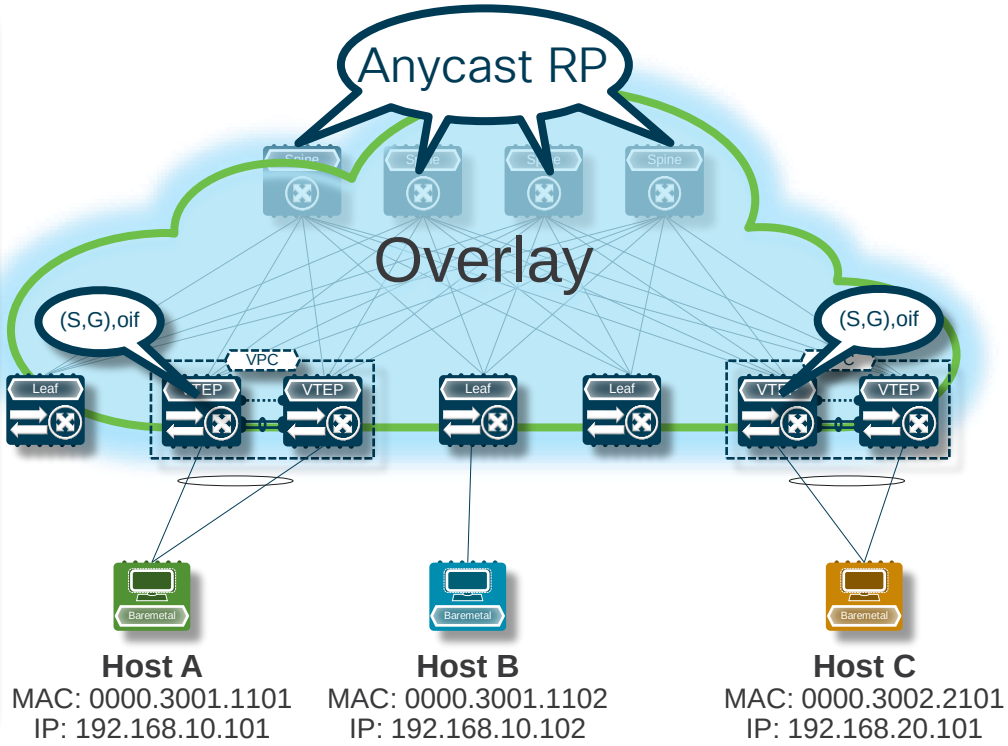
EVPN Control-Plane	
Type	3
L2VNI / RT	3001/ 65500:3001
Origin. Router	10.200.200.123



BUM traffic in vPC

Multicast Mode

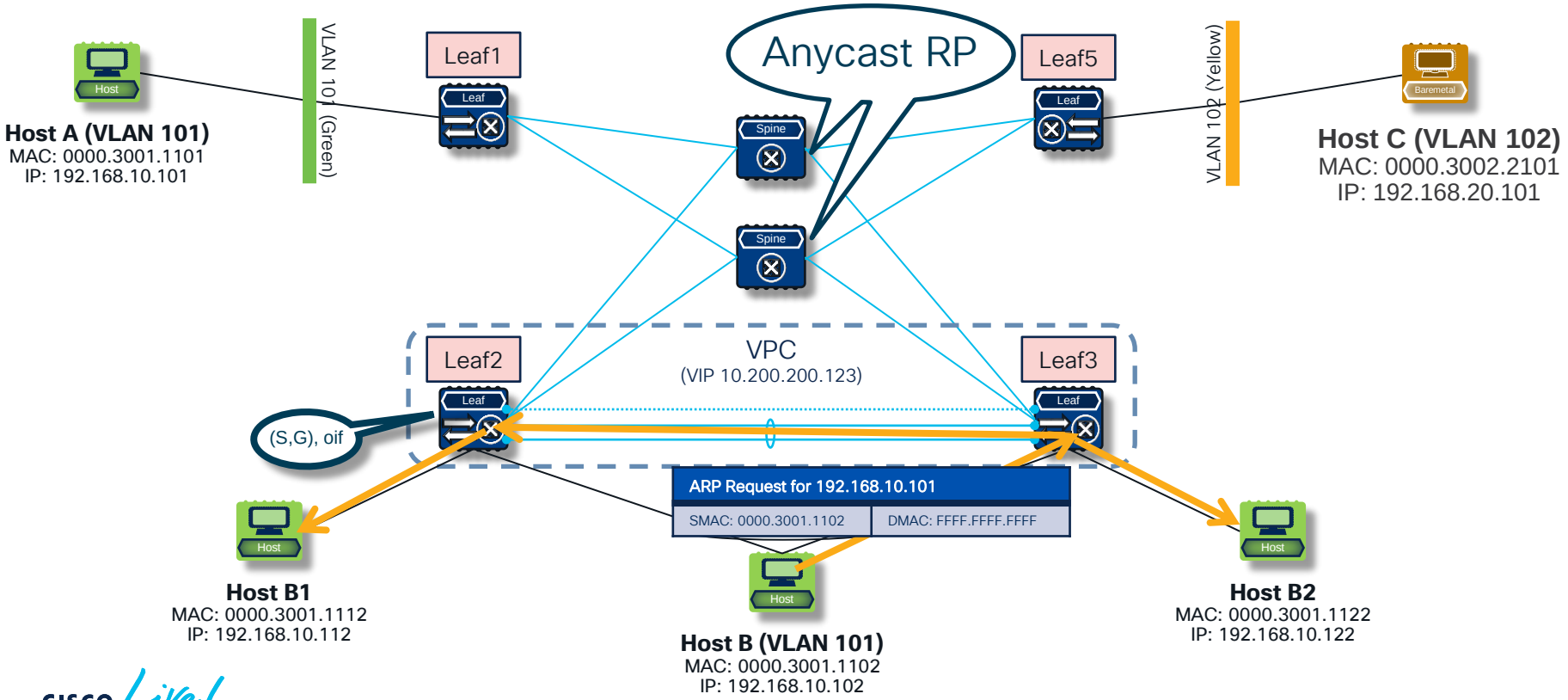
- BUM traffic is flooded to the multicast group
- Map a L2 VNI to a multicast group
- PIM Network in Underlay
- Any vPC peer can encapsulate traffic, and send to anycast RP
- One of the vPC peer is elected as decapsulation node
  - Peer with lower cost to RP will be elected as decapsulation node
  - Same cost to RP, vPC Primary will be elected



BUM traffic – Sender in VPC

Multicast Mode

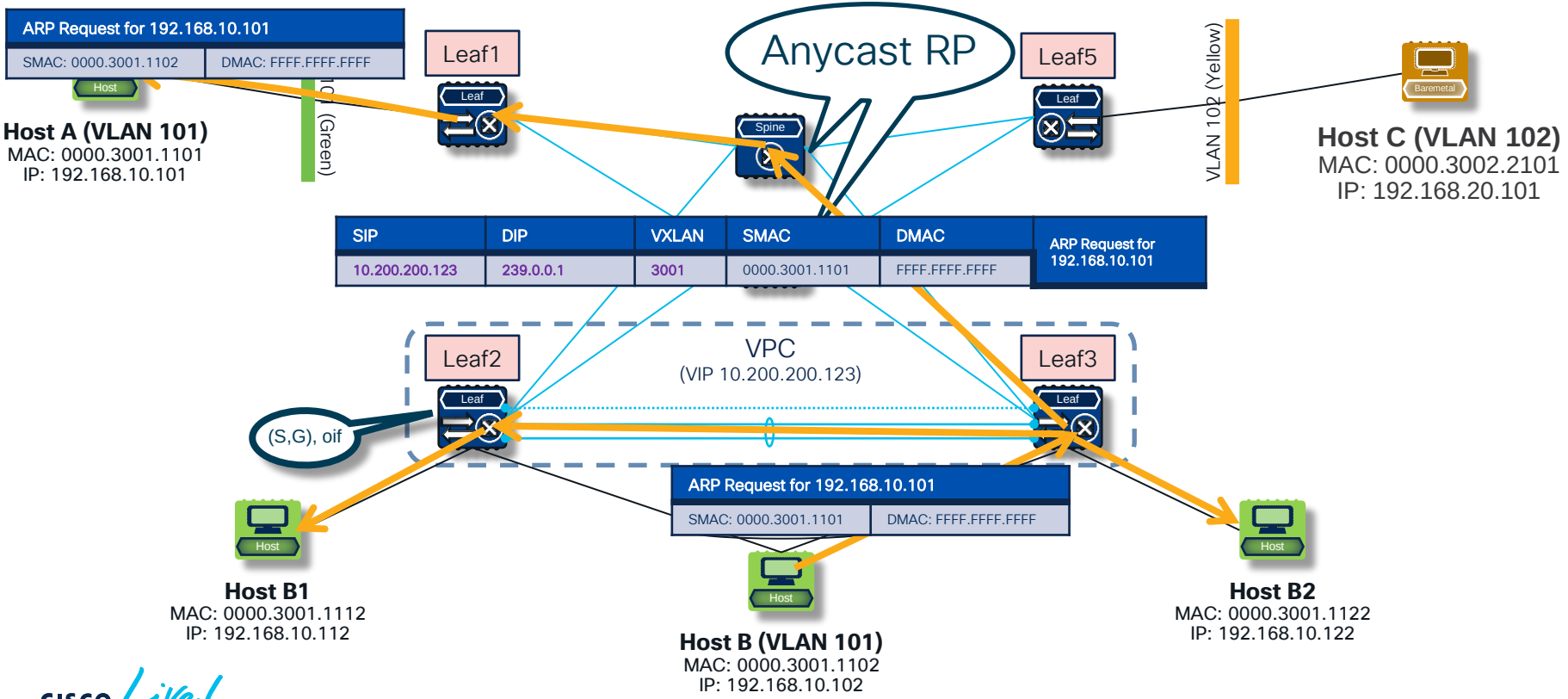
L2 VNI mapped to multicast group 239.0.0.1



BUM traffic – Sender in VPC

Multicast Mode

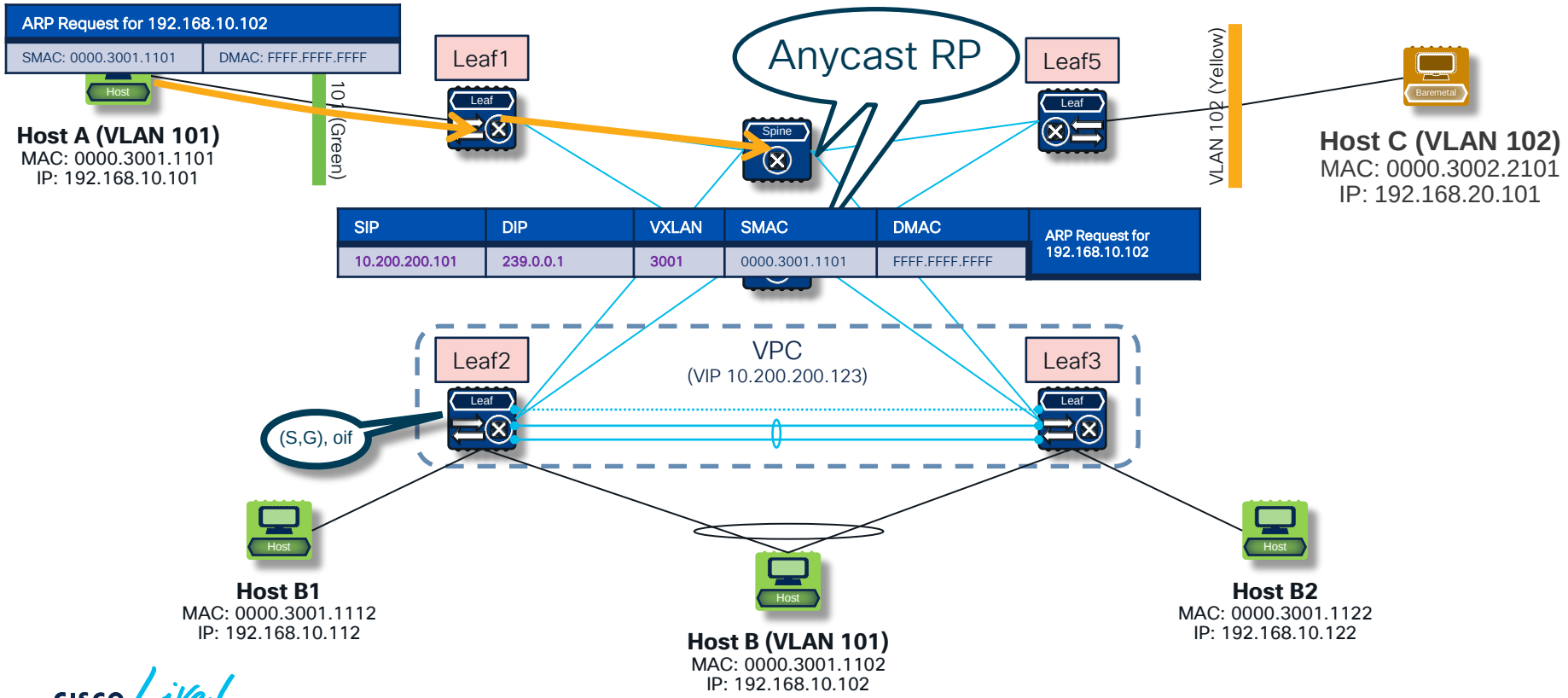
L2 VNI mapped to multicast group 239.0.0.1



BUM traffic – Receiver in VPC

Multicast Mode

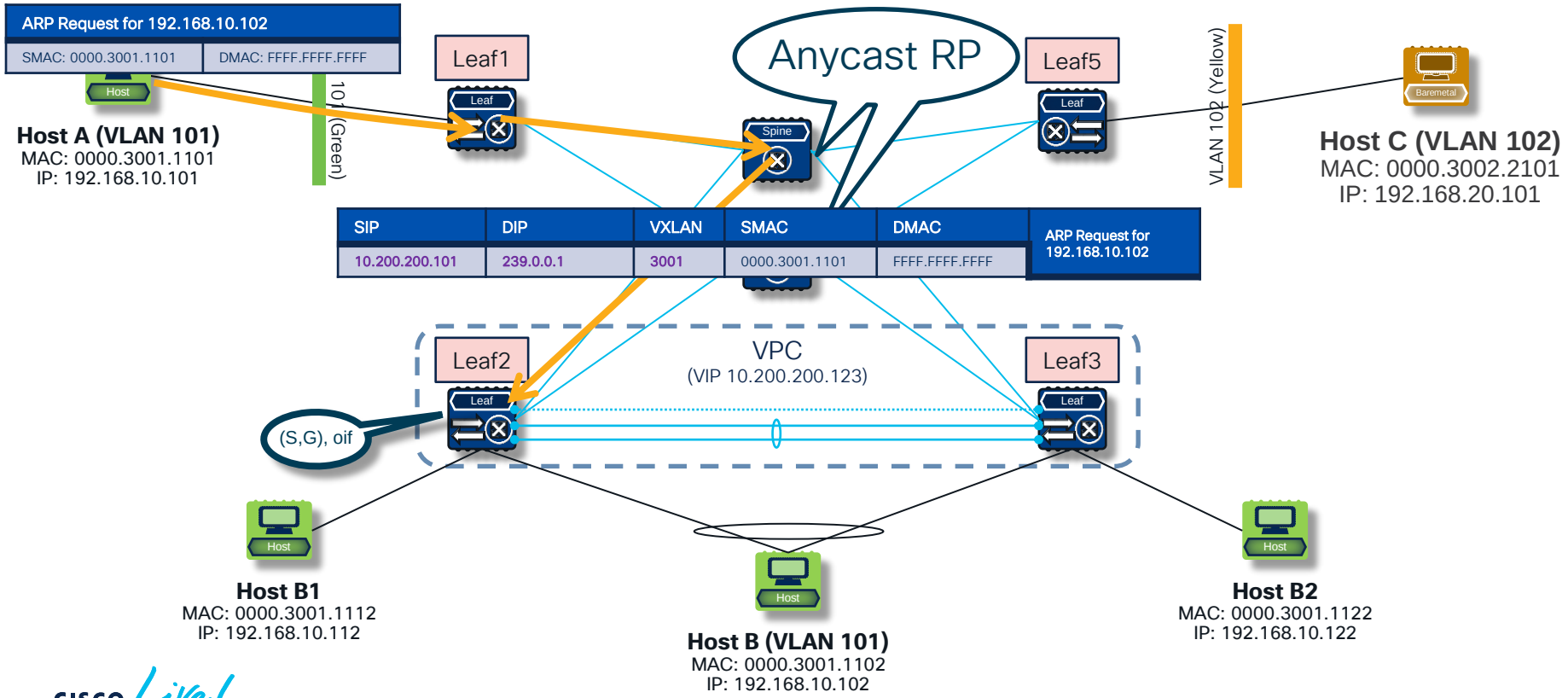
L2 VNI mapped to multicast group 239.0.0.1



BUM traffic – Receiver in VPC

Multicast Mode

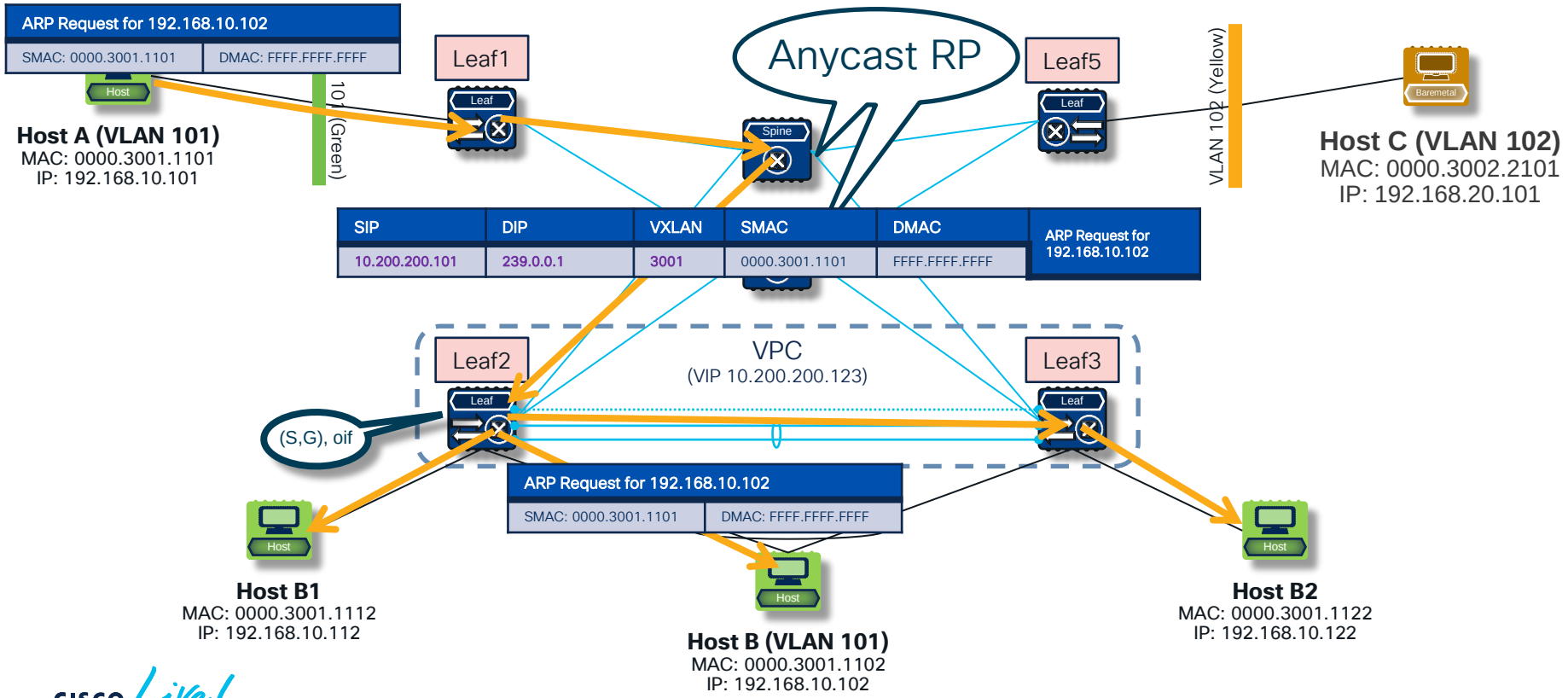
L2 VNI mapped to multicast group 239.0.0.1



BUM traffic – Receiver in VPC

Multicast Mode

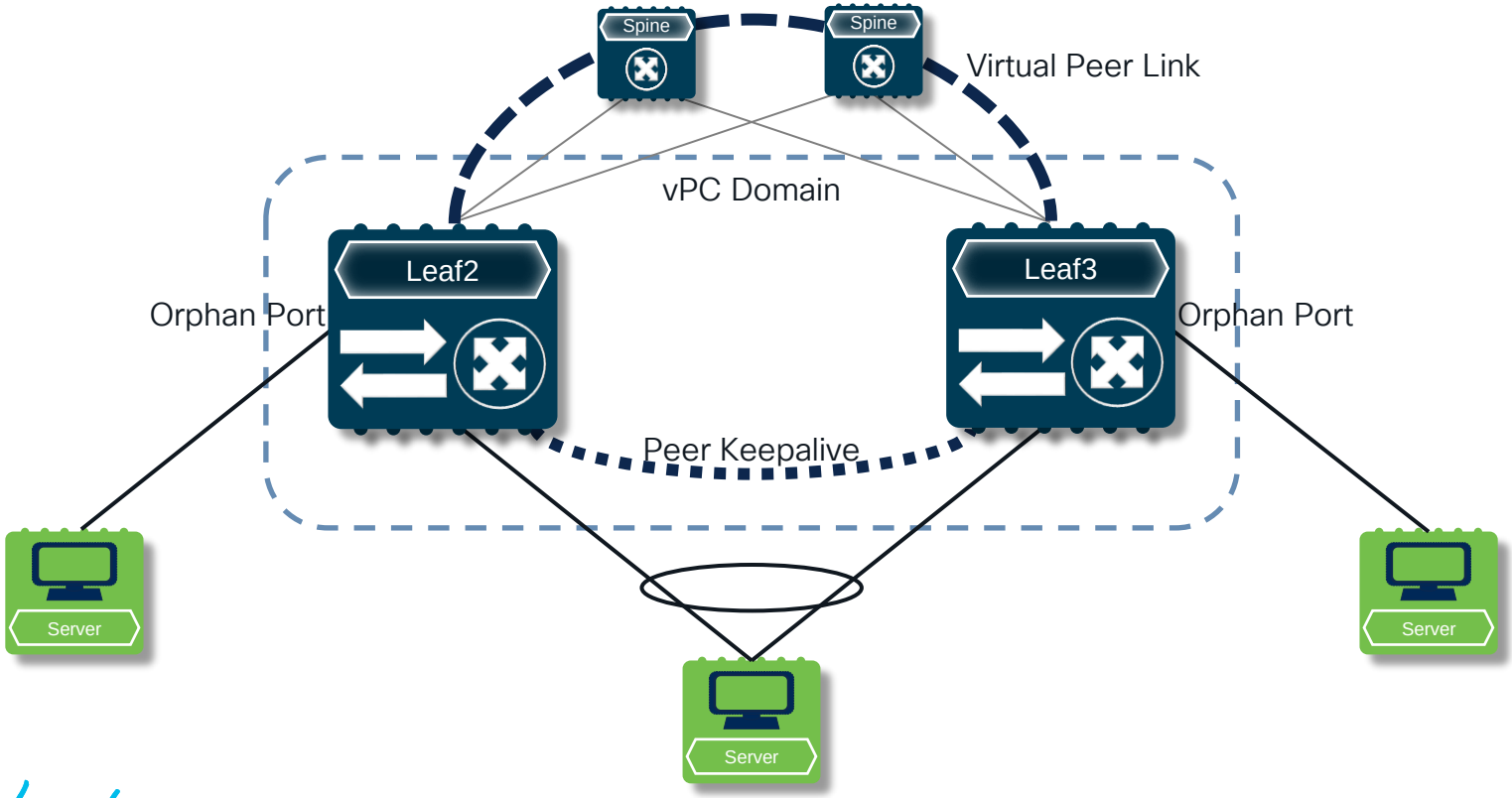
L2 VNI mapped to multicast group 239.0.0.1



vPC Fabric Peering

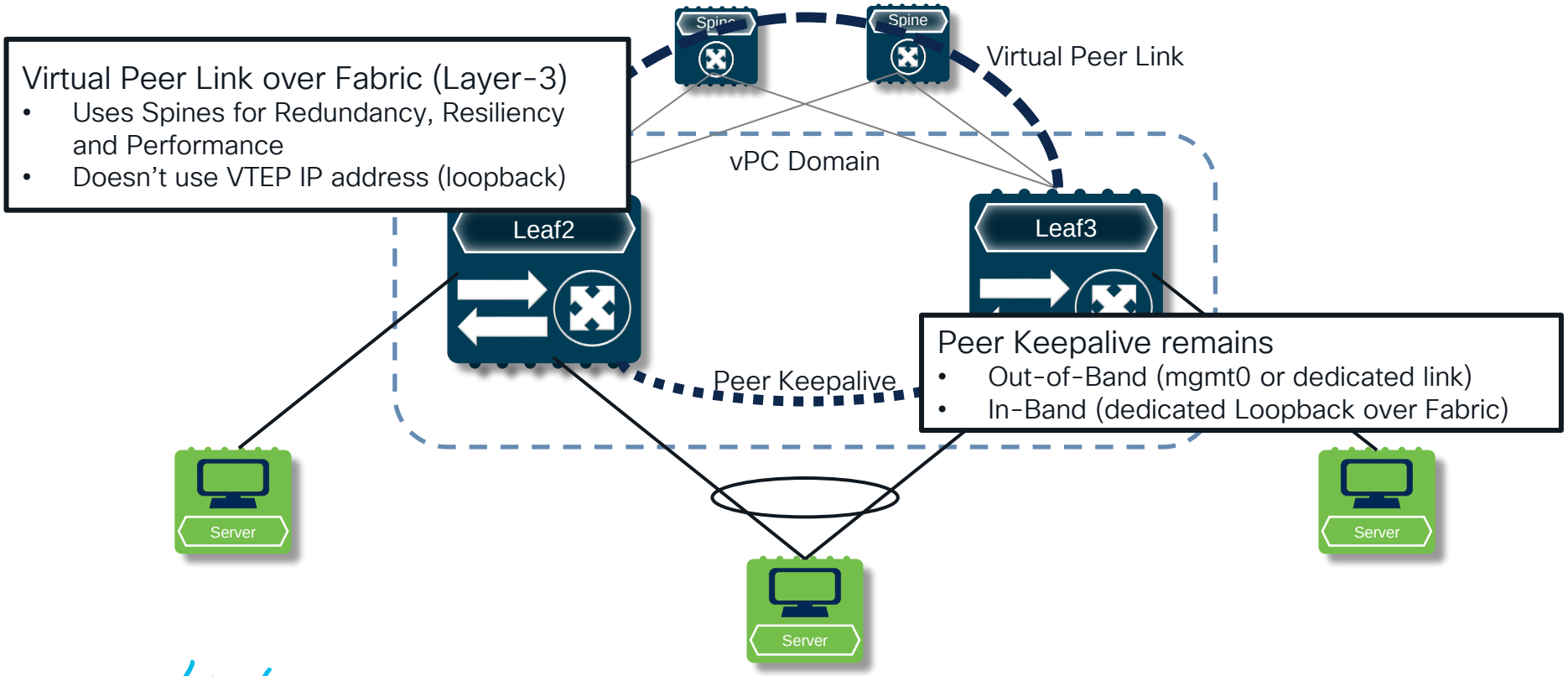
vPC with Fabric Peering

for VXLAN BGP EVPN



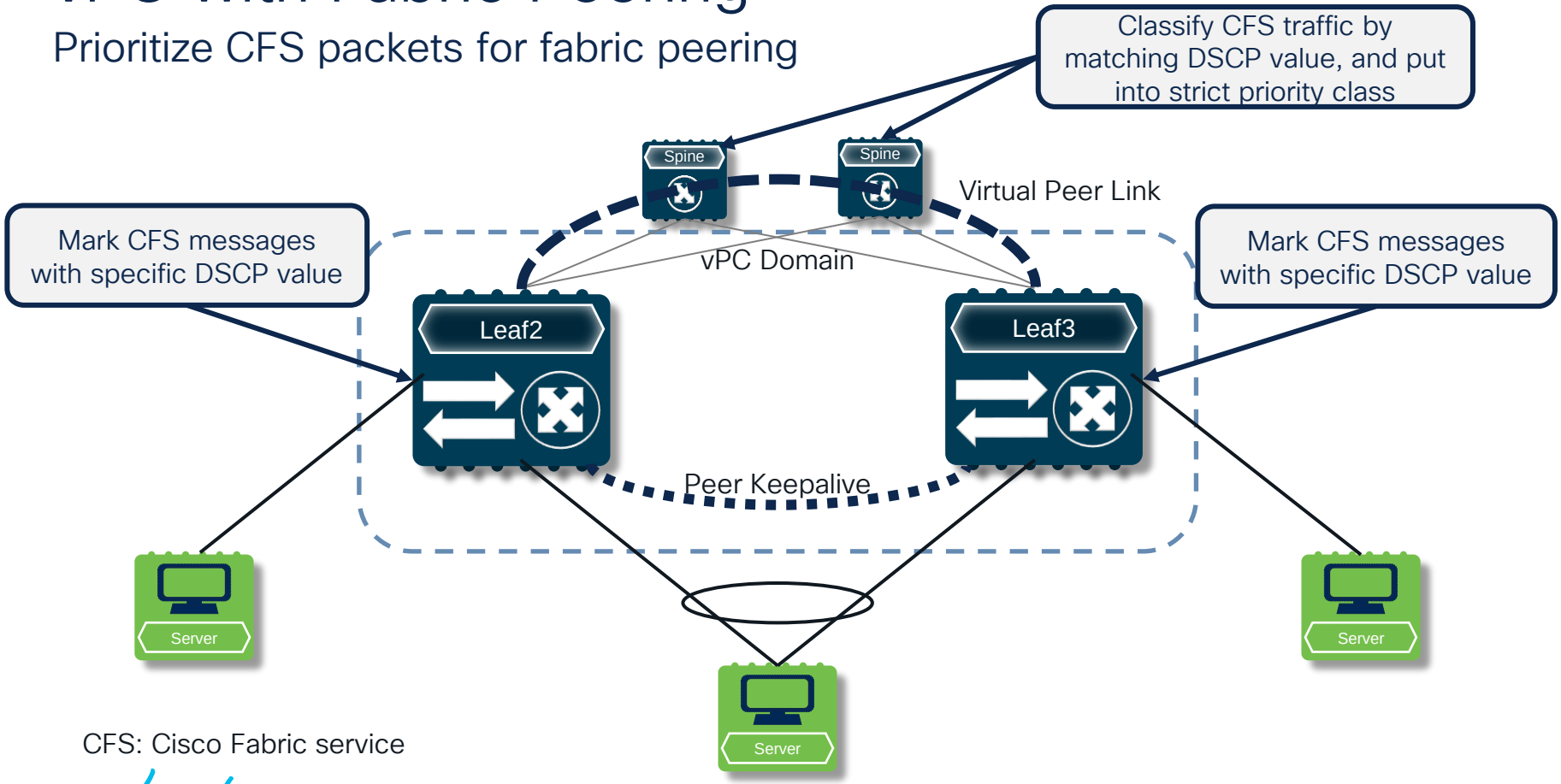
vPC with Fabric Peering

for VXLAN BGP EVPN



vPC with Fabric Peering

Prioritize CFS packets for fabric peering



CFS: Cisco Fabric service

Configuration – Define vPC Domain

```
vpc domain 1
  peer-switch
  peer-keepalive destination 10.10.10.82 source 10.10.10.81
  virtual peer-link destination 10.44.0.4 source 10.44.0.3 dscp 56
  delay restore 150
  peer-gateway
  auto-recovery reload-delay 360
  ipv6 nd synchronize
  ip arp synchronize

interface port-channel500
  description "vpc-peer-link"
  switchport
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link
```

vPC Domain

Virtual peer link

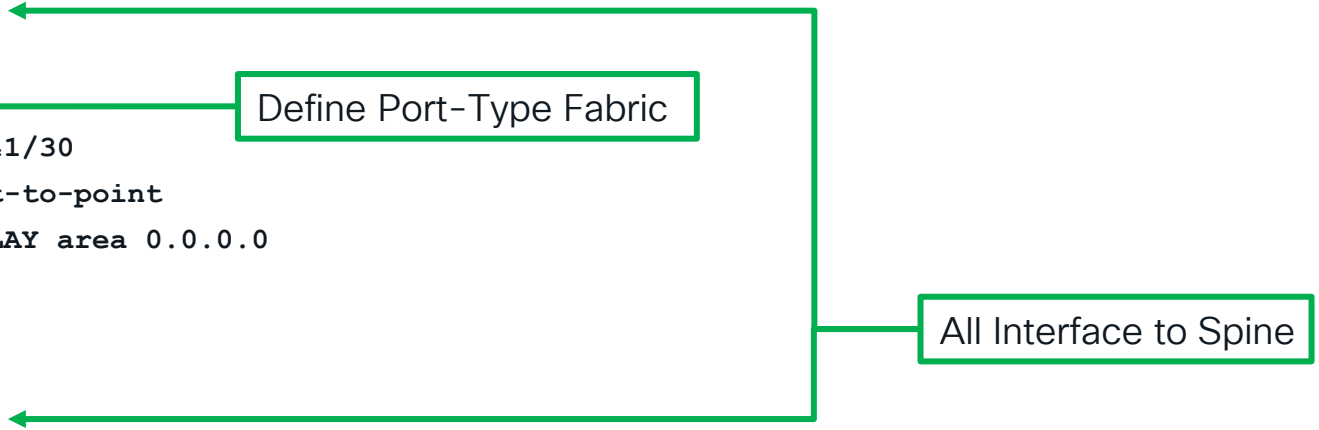
Port-Channel for Peer Link definition (must have no physical members!)

Make it peer link

Configuration – Define Uplink to Spine

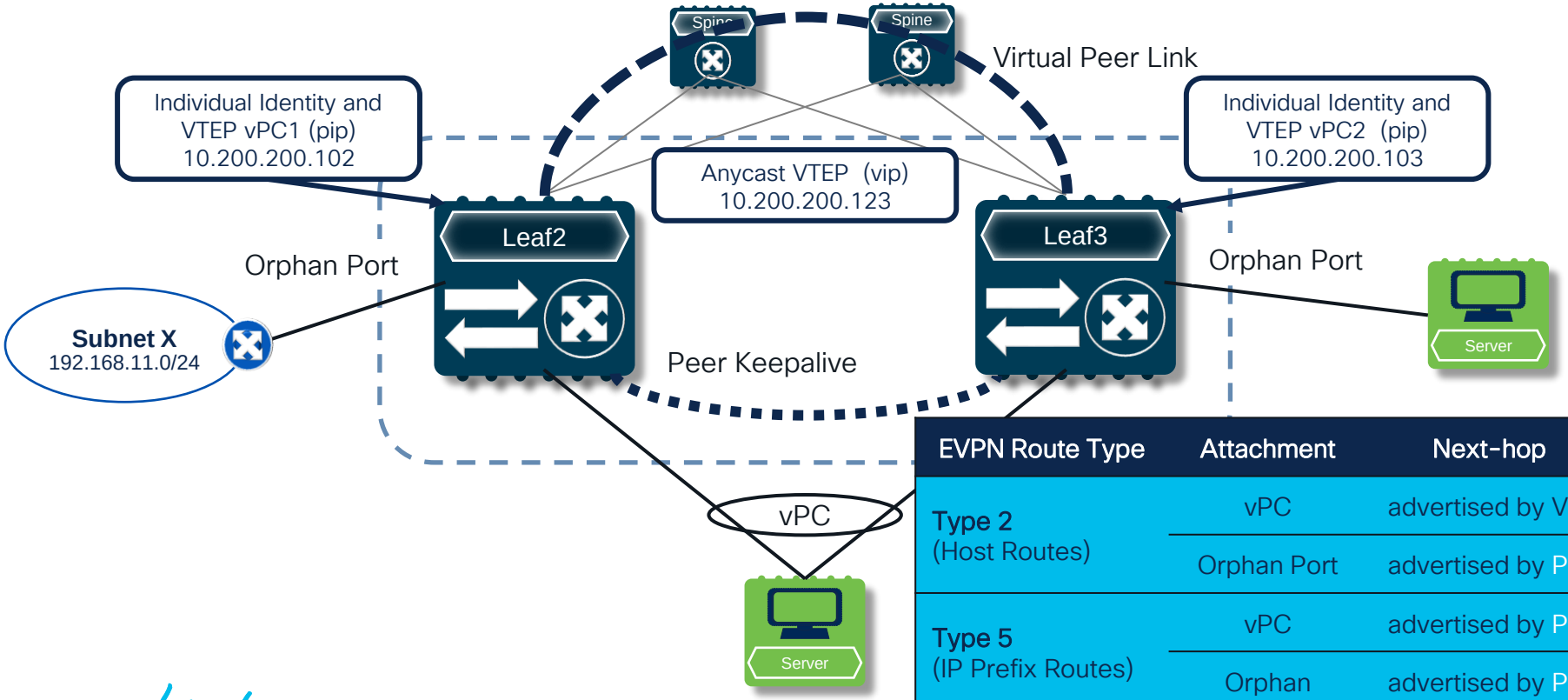
```
interface Ethernet1/49
  mtu 9216
  port-type fabric
  ip address 10.144.0.41/30
  ip ospf network point-to-point
  ip router ospf UNDERLAY area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet1/50
  mtu 9216
  port-type fabric
  ip address 10.144.0.29/30
  ip ospf network point-to-point
  ip router ospf UNDERLAY area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```



Host Attachment

vPC with Fabric Peering

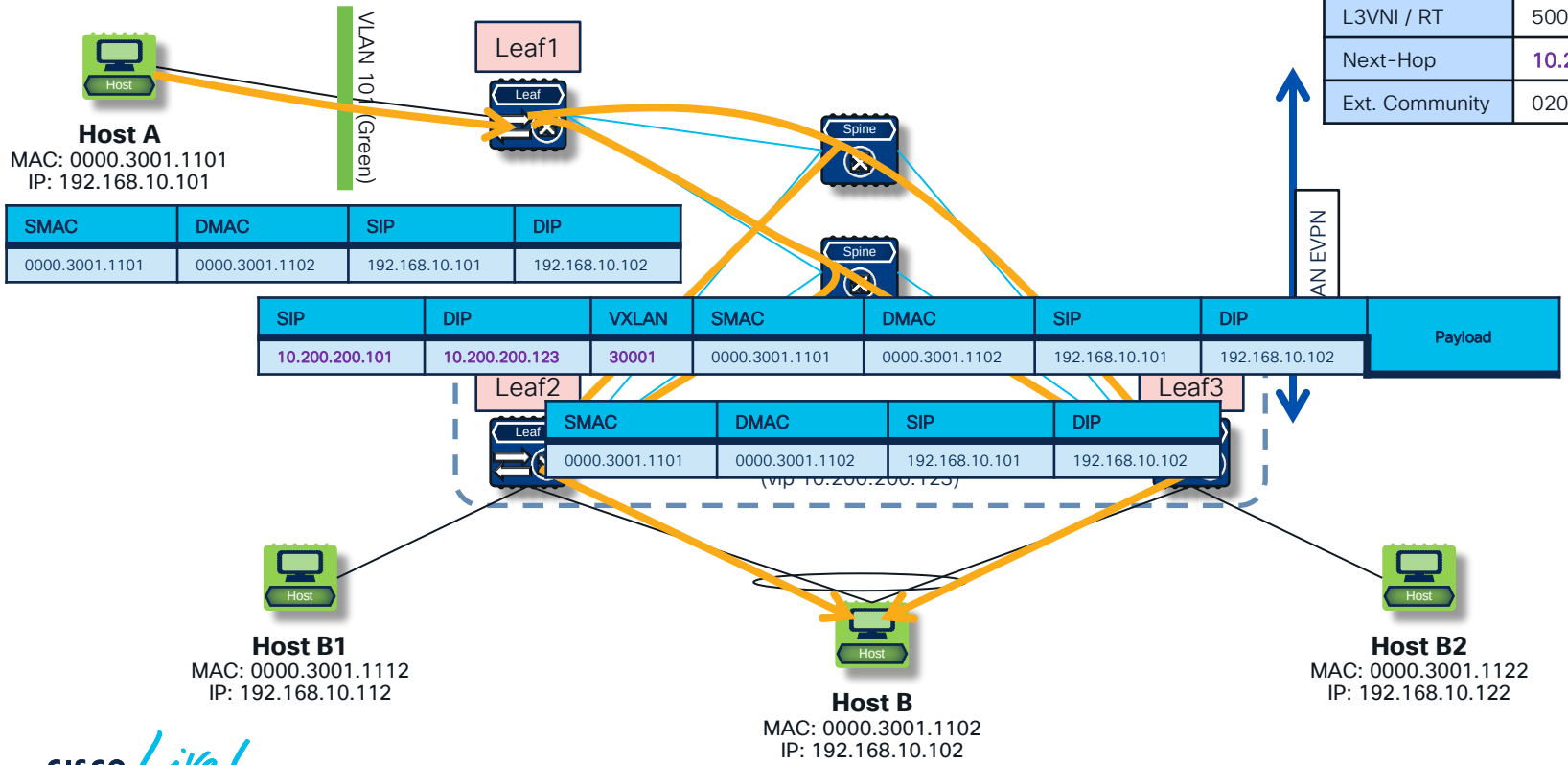


EVPN Route Type	Attachment	Next-hop
Type 2 (Host Routes)	vPC	advertised by VIP
	Orphan Port	advertised by PIP
Type 5 (IP Prefix Routes)	vPC	advertised by PIP
	Orphan	advertised by PIP

Packet Walk – vPC with Fabric Peering

vPC Host

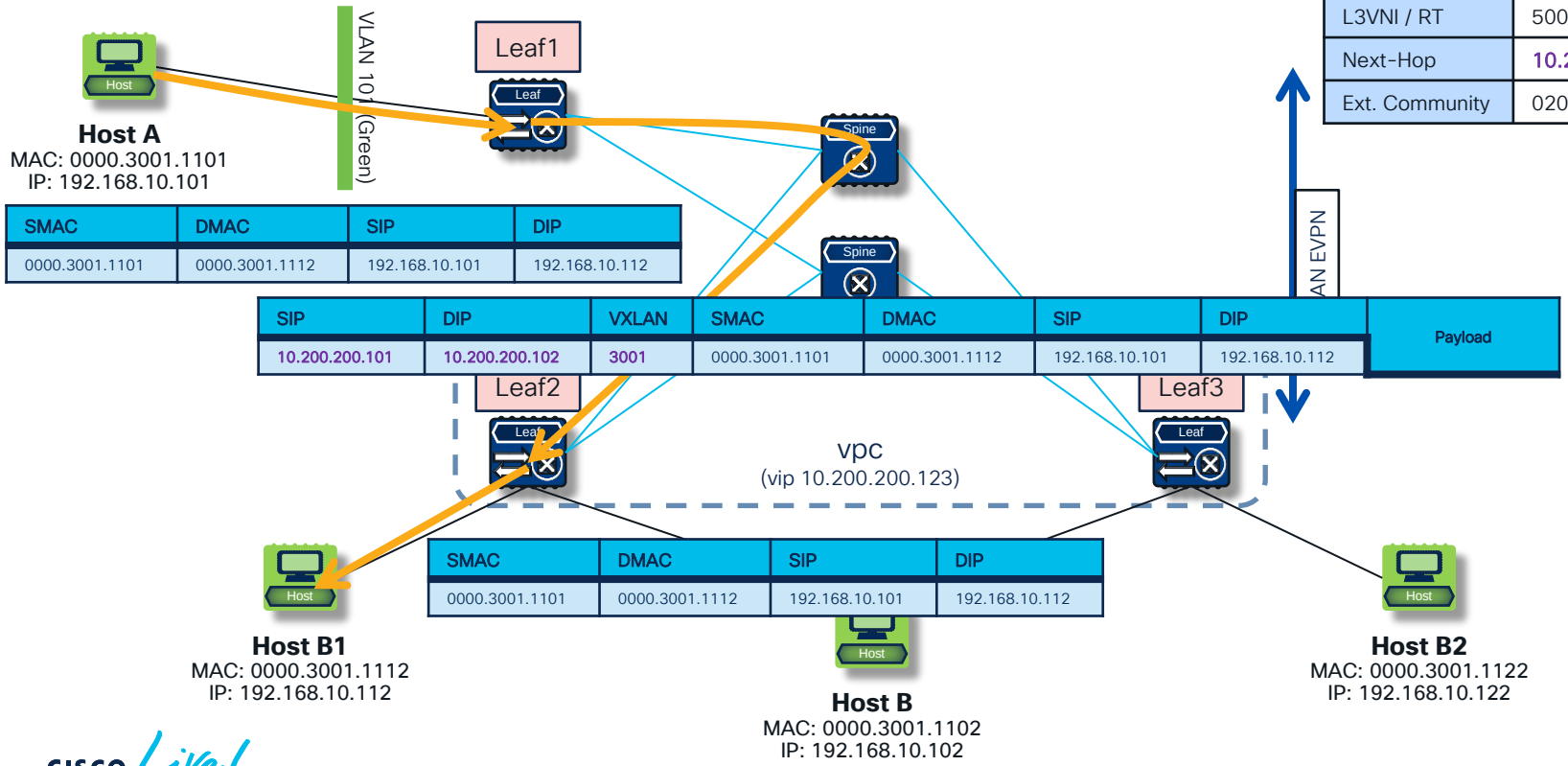
EVPN Control-Plane	
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ac8.c87b



Packet Walk – vPC with Fabric Peering

Orphan Host

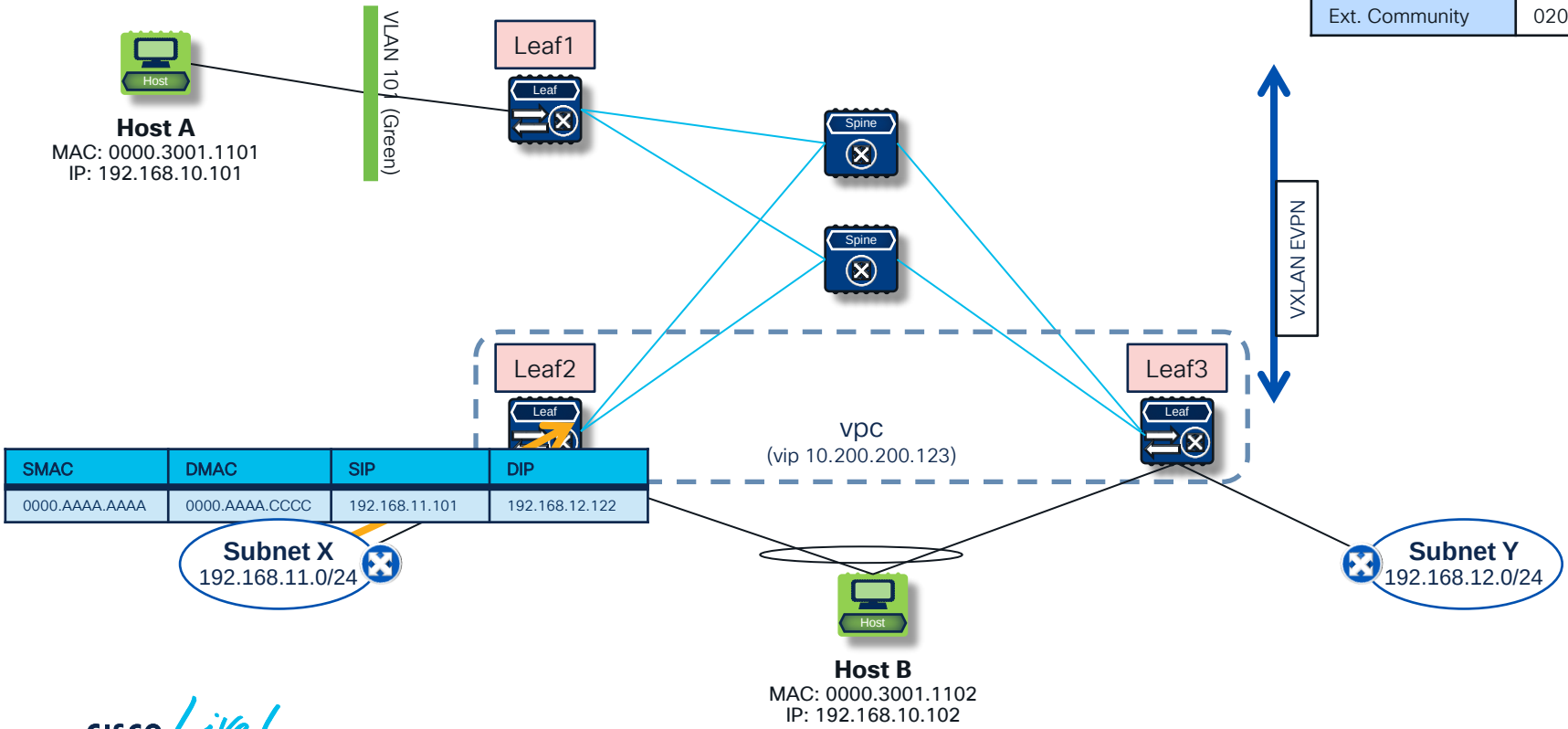
EVPN Control-Plane	
Type	2
MAC / Length	0000.3001.1112 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.112 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.102
Ext. Community	0200.0ade.de02



Packet Walk – vPC Fabric Peering

Orphan Networks

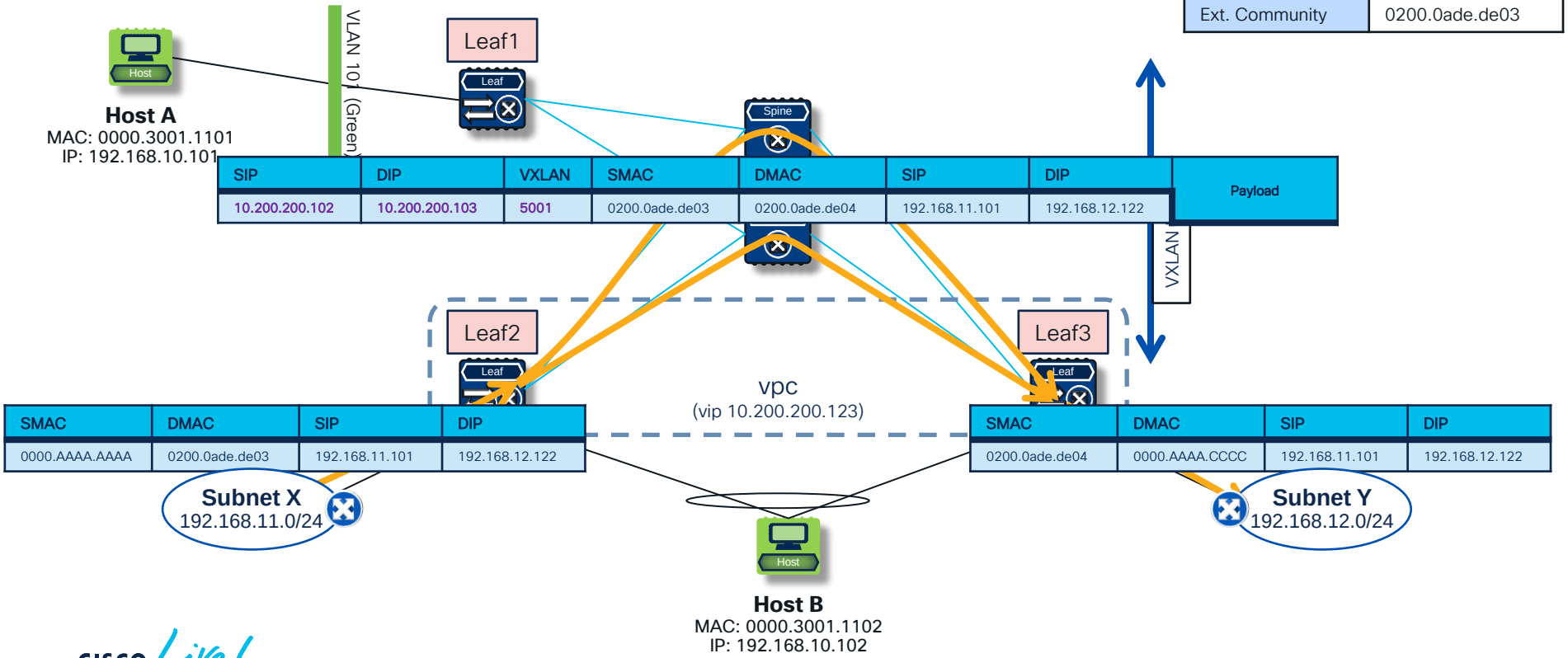
EVPN Control-Plane	
Type	5
IP / Length	192.168.12.0 / 24
L3VNI / RT	5001 / 65500:5000
Next-Hop	10.200.200.103
Ext. Community	0200.0ade.de03



Packet Walk – vPC Fabric Peering

Orphan Networks

EVPN Control-Plane	
Type	5
IP / Length	192.168.12.0 / 24
L3VNI / RT	5001 / 65500:5000
Next-Hop	10.200.200.103
Ext. Community	0200.0ade.de03

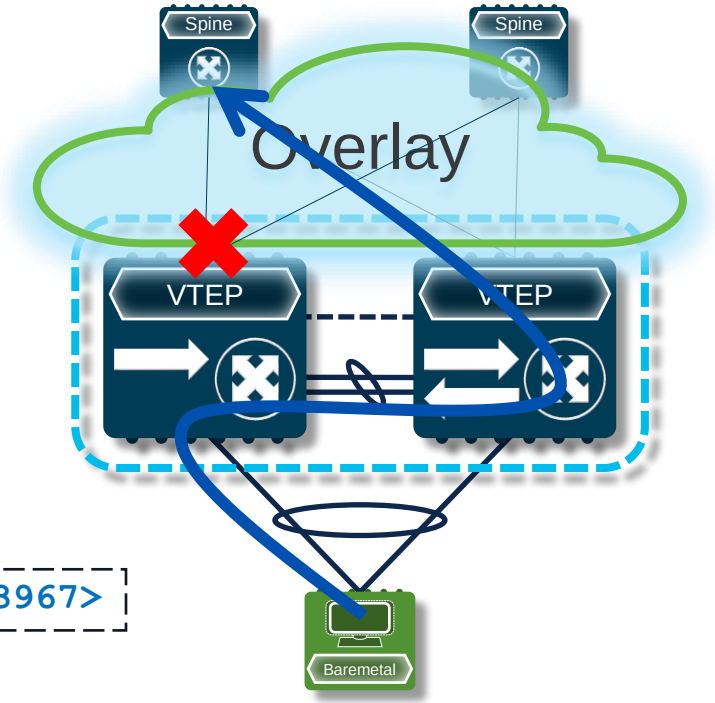


vPC configuration best practices

vPC Infrastructure VLANs

- Infrastructure VLANs are used for Backup Routing Path, in default VRF
- Infrastructure VLAN is present on Peer Link
- Used in case of failure of uplinks on a vPC peer

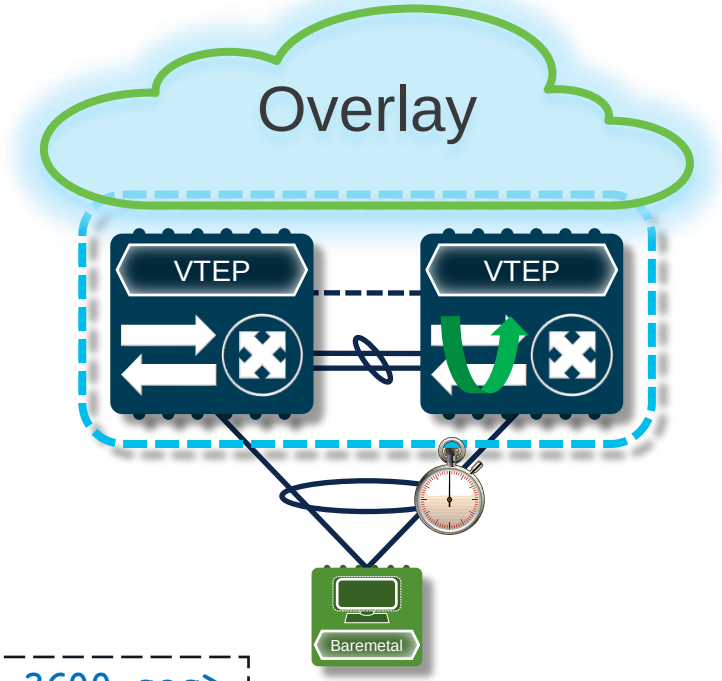
```
Nexus (config)# system nve infra-vlans <1-3967>
```



vPC Configuration Best Practices

vPC Delay Restore

- After vPC peer reload, traffic might be black-holed, before L3 connectivity is reestablished
- vPC link bring up can be delayed to allow Underlay and Overlay Convergence
- Allows encapsulation path to converge
- Default time 150 seconds

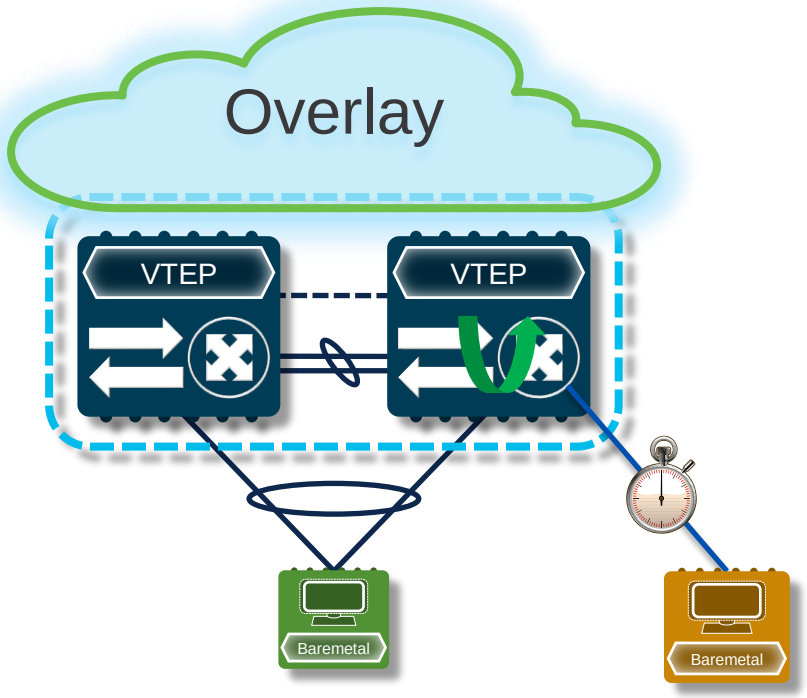



```
Nexus(config-vpc-domain)# delay restore <1-3600 sec>
```

vPC Configuration Best Practices

Orphan Port Delay Restore

- After vPC peer reload, traffic might be black-holed, before L3 connectivity is reestablished
- Orphan port bring up can be delayed to allow Underlay and Overlay Convergence
- Allows encapsulation path to converge
- Default time is equal as vPC delay restore time

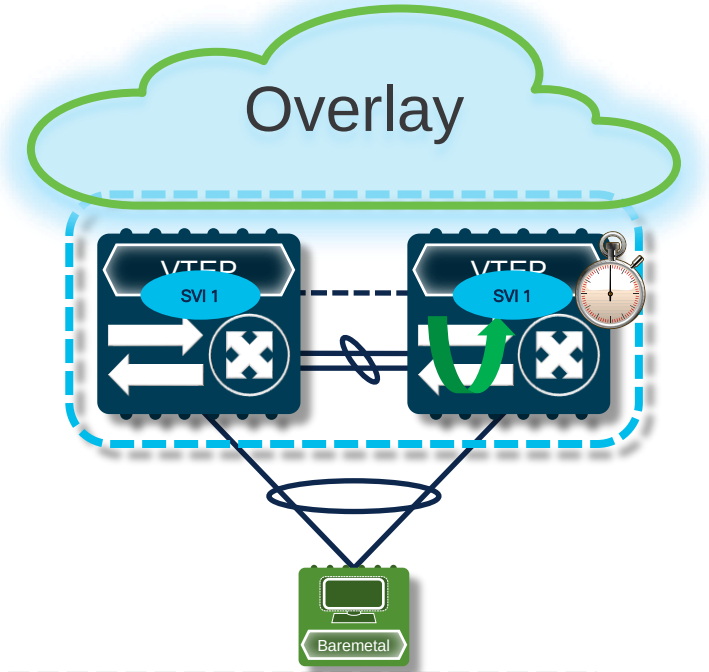



```
Nexus(config-vpc-domain)# delay restore orphan-port <0-300 sec>
```

vPC Configuration Best Practices

SVI Delay Restore

- After vPC peer reload, traffic might be black-holed, before L3 connectivity is reestablished
- SVI bring up can be delayed to allow Underlay and Overlay Convergence
- Allows encapsulation path to converge
- Default time 10 seconds

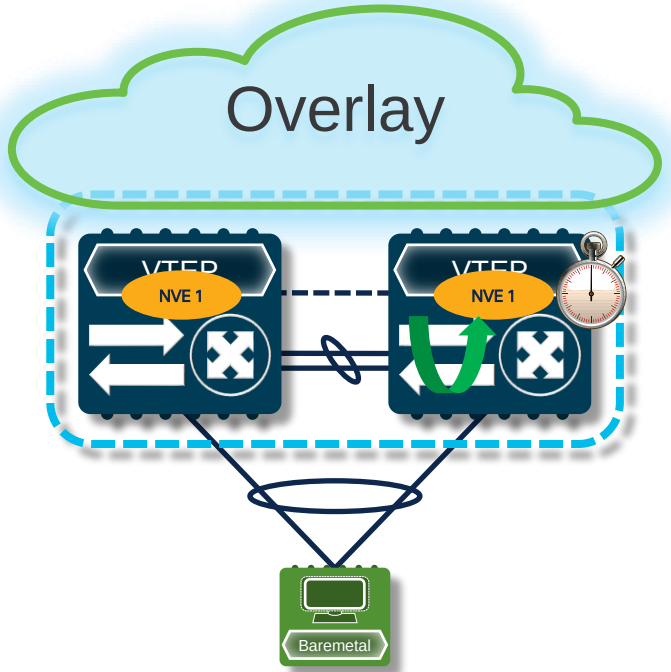



```
Nexus (config-vpc-domain)# delay restore interface-vlan <1-3600 sec>
```

vPC Configuration Best Practices

NVE Hold-Down timer

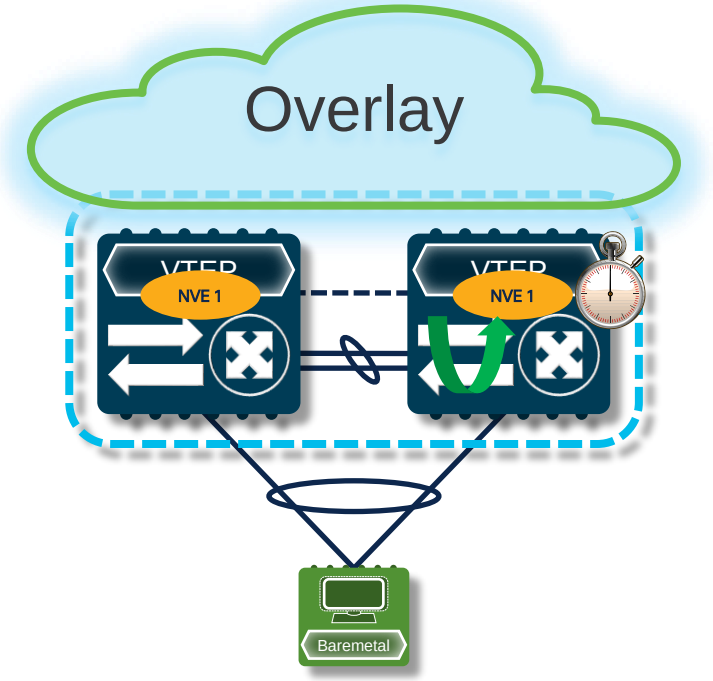
- After vPC peer reload, traffic going to Anycast VTEP hashed to the peer will be black-holed
- Advertisements of NVE loopback interface can be suppressed until overlay has converged
- NVE loopback interface bring up can be delayed using hold-down timer
- For proper overlay convergence, hold-down time needs to be longer than delay restore time
- Default time 180 seconds




```
Nexus(config-if-nve)# source-interface hold-down <1-1500 sec>
```

Optimize Convergence for Orphan ports

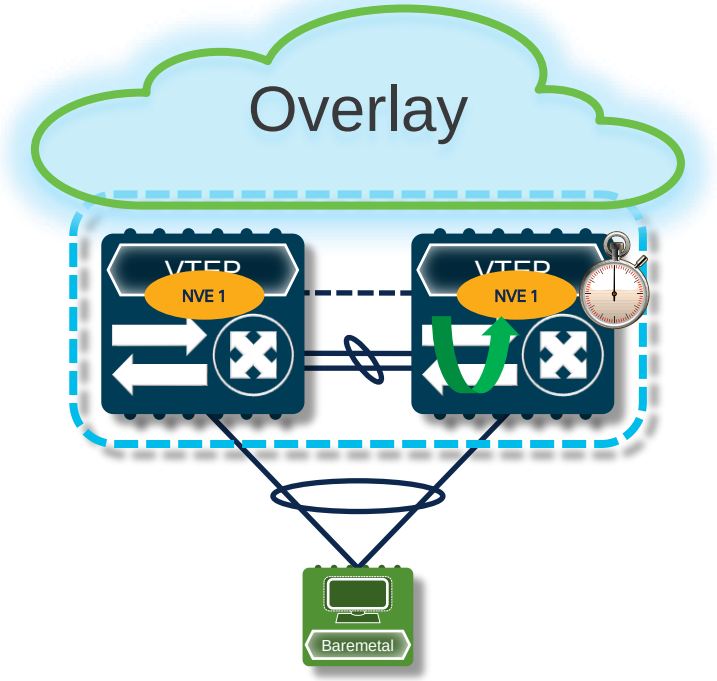
- After vPC peer reload, NVE source-loopback is kept down by using hold-down timer to avoid traffic blackholing
- During this period, no traffic (vPC and Orphan) is attracted by this peer.
- While the dual-attached hosts are being served by the other peer
- It delays convergence of Orphan Type-2 & Type-5 routes which are advertised using PIP.



Optimize Convergence for Orphan ports

Split Loopbacks

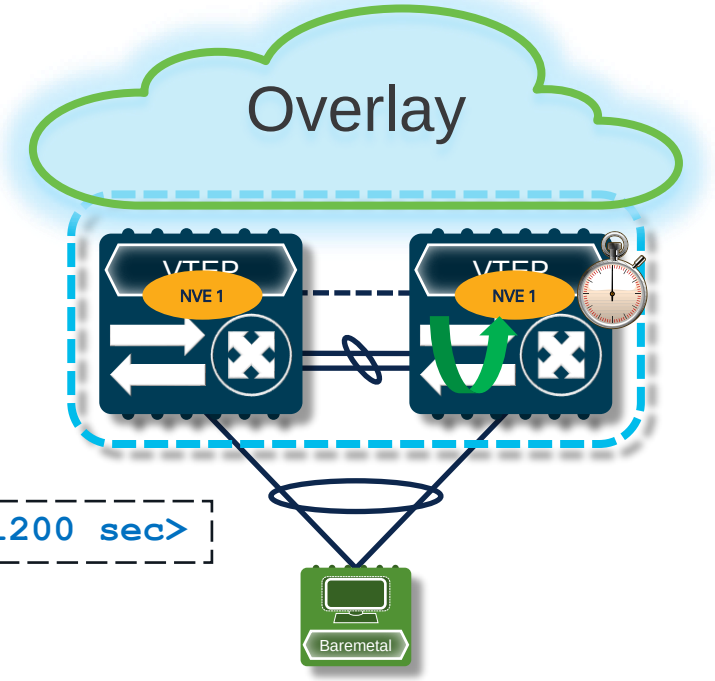
- VTEP's PIP and VIP are assigned to separate loopbacks
- It allows PIP loopback to brought up before the VIP (Anycast) loopback.
- Improves convergence for following routes advertised using PIP -
  - Type-2 Orphan
  - Type-5 subnet routes



Optimize Convergence for Orphan ports

Split Loopbacks

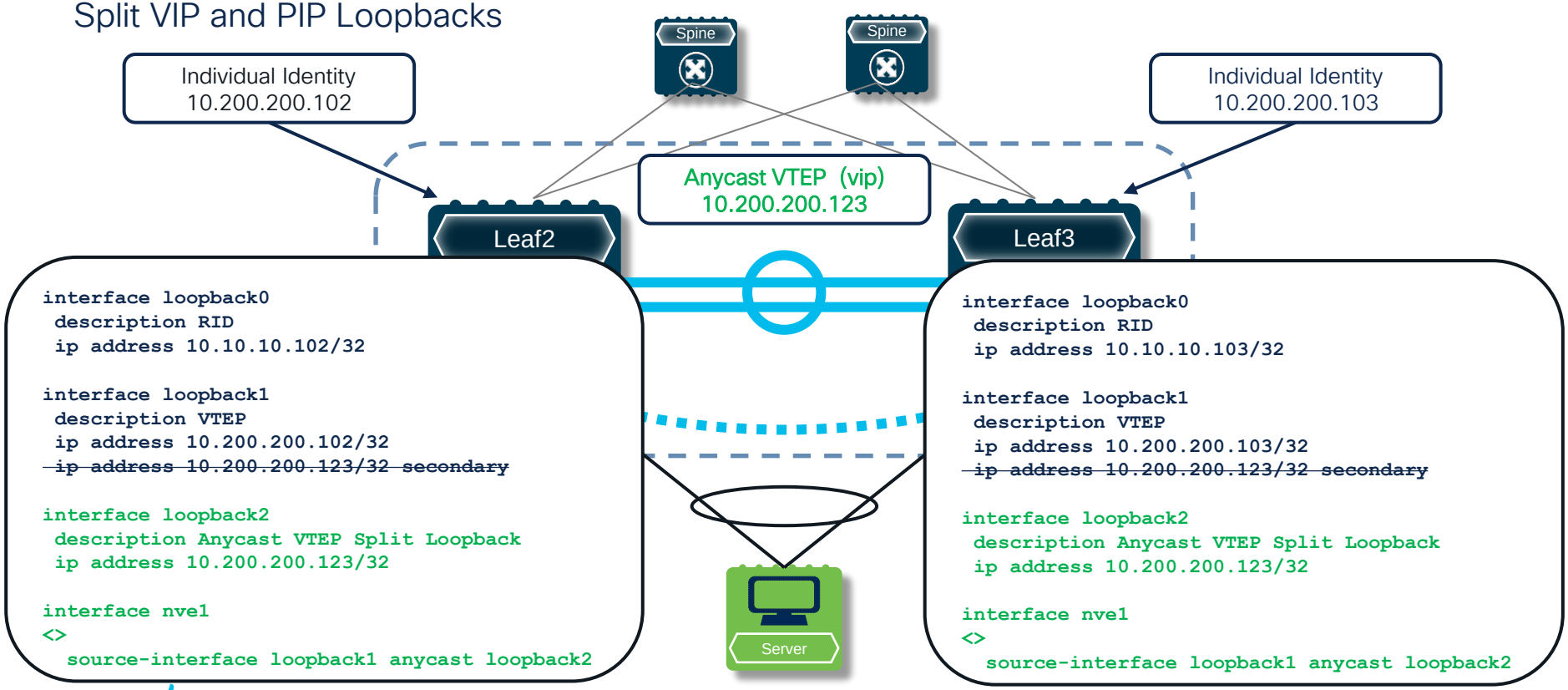
- Bringing up PIP loopback interface is controlled by Fabric Ready timer
- Default is 3/4th of source-interface hold-down-timer (i.e. 135 seconds)




```
Nexus(config-if-nve)# fabric-ready time <1-1200 sec>
```

Optimize Convergence for Orphan ports

Split VIP and PIP Loopbacks




```
interface loopback0
 description RID
 ip address 10.10.10.102/32

interface loopback1
 description VTEP
 ip address 10.200.200.102/32
 ip address 10.200.200.123/32 secondary

interface loopback2
 description Anycast VTEP Split Loopback
 ip address 10.200.200.123/32

interface nve1
<>
  source-interface loopback1 anycast loopback2
```

```
interface loopback0
 description RID
 ip address 10.10.10.103/32

interface loopback1
 description VTEP
 ip address 10.200.200.103/32
 ip address 10.200.200.123/32 secondary

interface loopback2
 description Anycast VTEP Split Loopback
 ip address 10.200.200.123/32

interface nve1
<>
  source-interface loopback1 anycast loopback2
```

VXLAN vPC Consistency Checking

vPC consistency check in VXLAN requires:

- The same VLAN-to-VNI mapping on both vPC peers
- SVI present for VLANs mapped to VNI on both vPC peers
- The same VNI needs to use the same BUM traffic transport mechanism on both VTEPs
- When a VNI uses multicast replication, both VTEPs need to use the same multicast group for this VNI

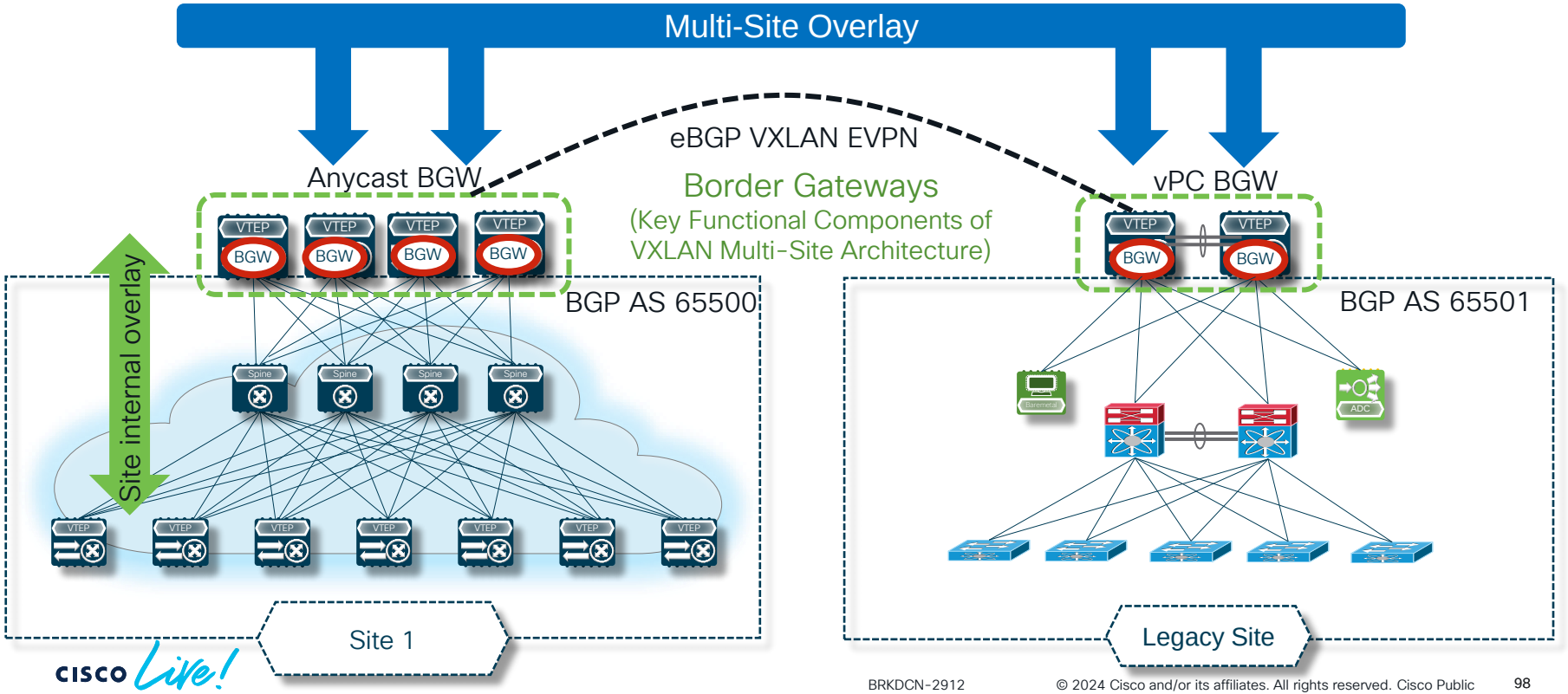
When vPC VTEP consistency check failed:

- The NVE loopback interface will be admin shutdown on the vPC secondary VTEP

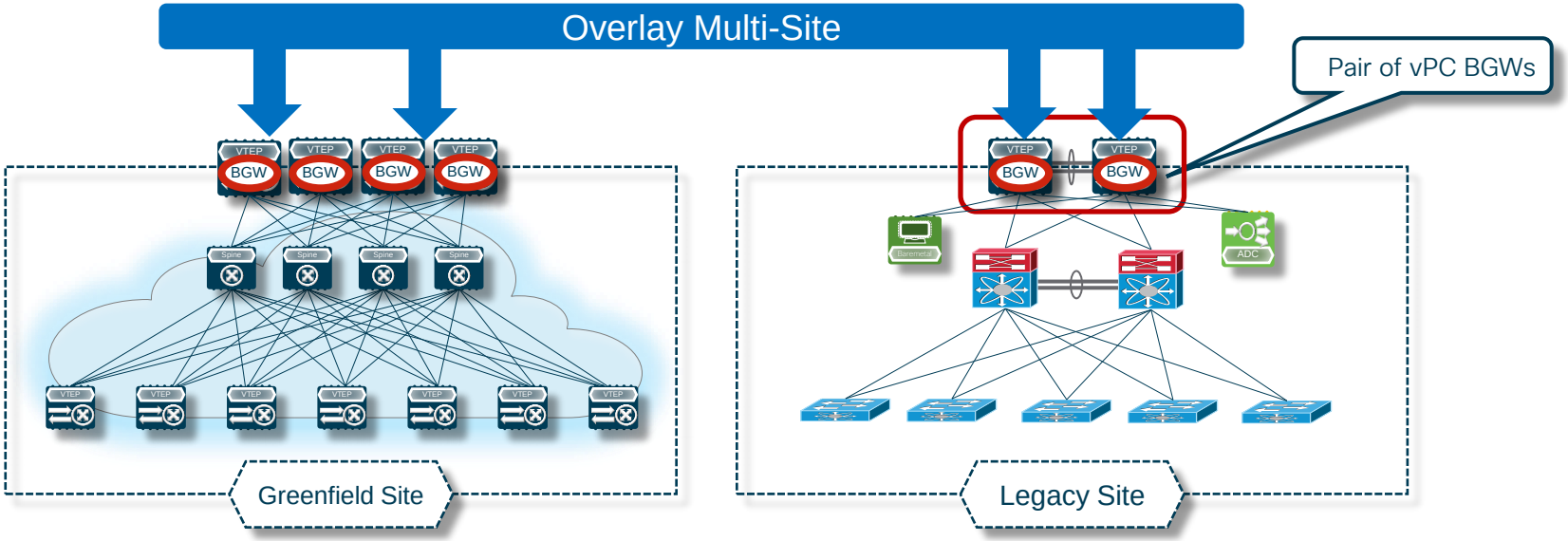
vPC Border Gateway

VXLAN Multi-Site

Scalable L2 and L3 DCI for both Greenfield and brownfield Sites



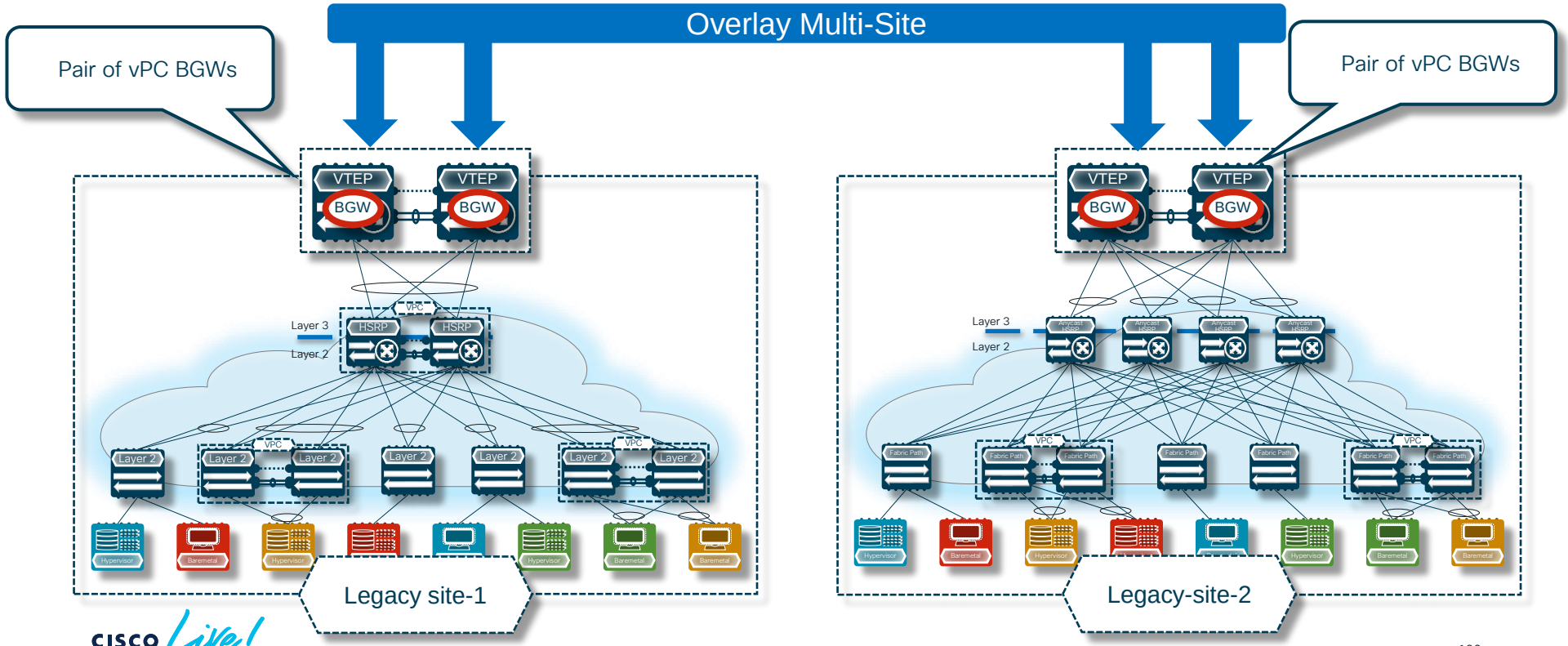
vPC Border Gateway
Connecting Legacy to Greenfield site



- Coexistence and/or migration use cases
  - Extend Layer-2 and Layer-3 multi-tenant connectivity across sites
- Deploy a pair of vPC BGW in the legacy site

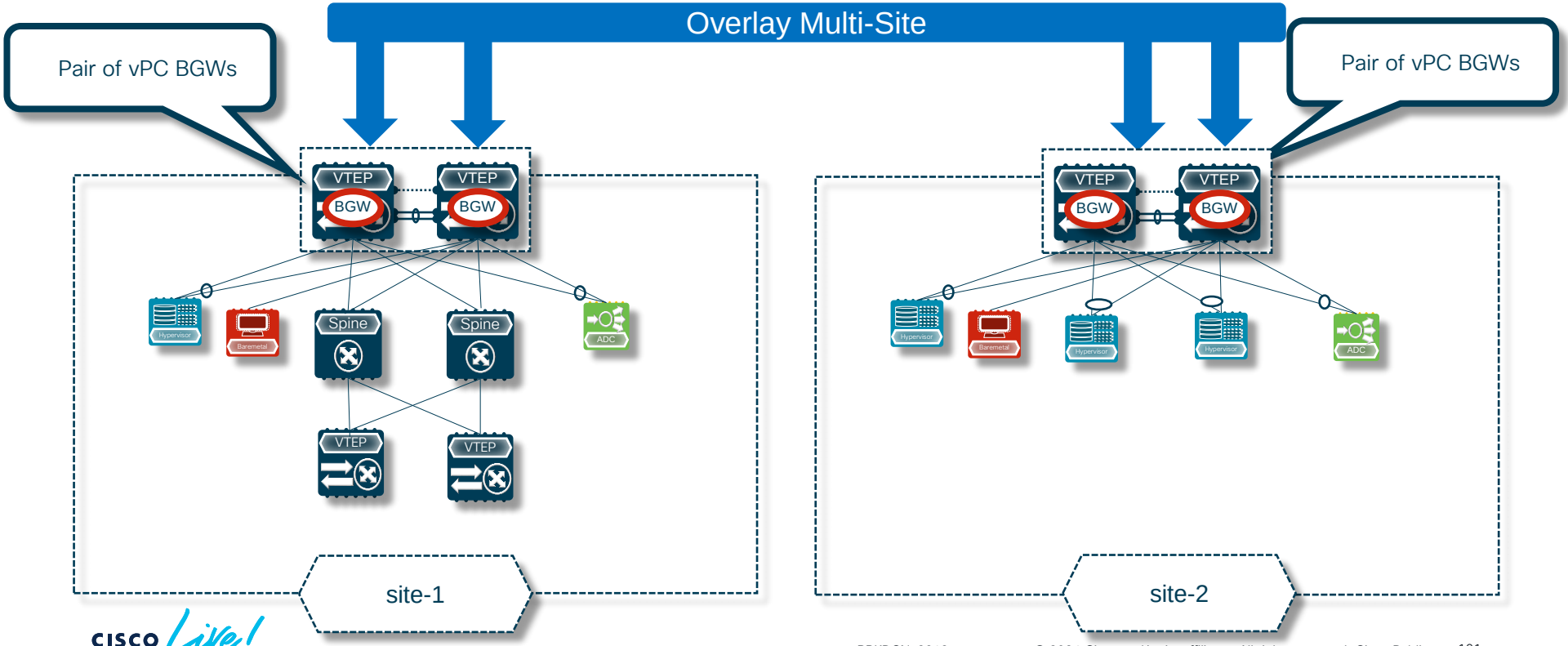
vPC Border Gateway

Connecting legacy sites

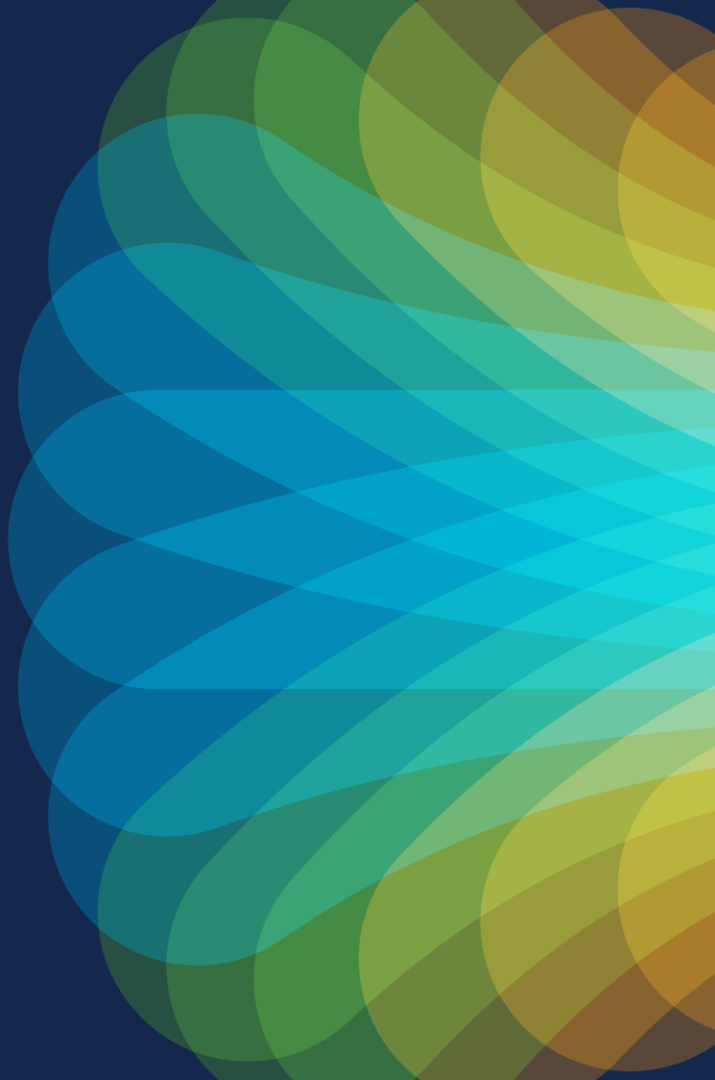


vPC Border Gateway

Connecting small sites with endpoints attached to vPC BGW



Automate VXLAN with vPC using Nexus Dashboard Fabric Controller (NDFC)



cisco Live!

Nexus Dashboard Fabric Controller (NDFC)

A comprehensive data center automation tool

NDFC helps you easily and reliably deploy, operate and maintain VXLAN-EVPN fabric

Day-0

Bootstrap, deploy

Day-1

Provision, maintain, monitor, operate

Day-2 with ND Insights

Troubleshoot, plan, grow

Scale out with ND Orchestrator

Multi-site and cloud acceleration

Why NDFC

Step into SDN via VXLAN BGP EVPN

End to End Automation

Single Pane of Glass for Day-0/Day-1 Provisioning

Config and Compliance across Cisco Products

Image Management

Change Control & Rollback

Simplify Complex Network Operations

Automate, Manage, and Interconnect Multi-Fabric topologies

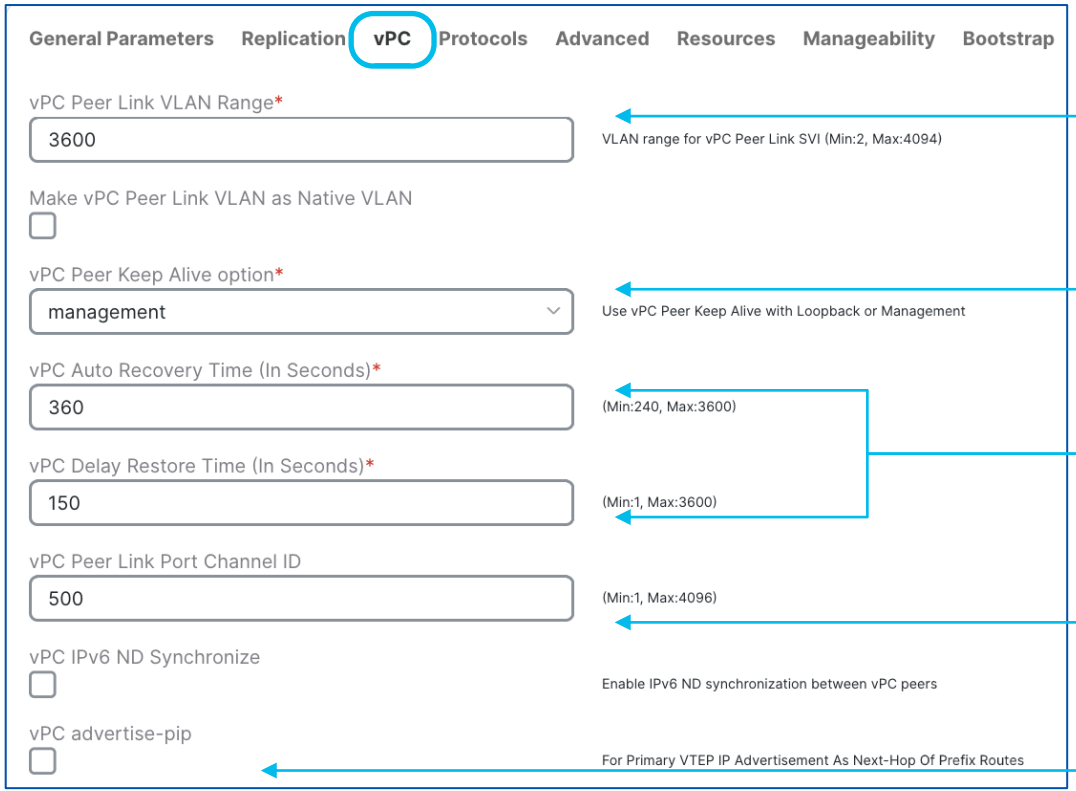
L4-L7 Service Insertion and Service Chaining

Compute Visibility

Programmability and Orchestration

RMA Workflow

NDFC VXLAN Fabric Template : vPC Parameters



Define system nve infra-vlans

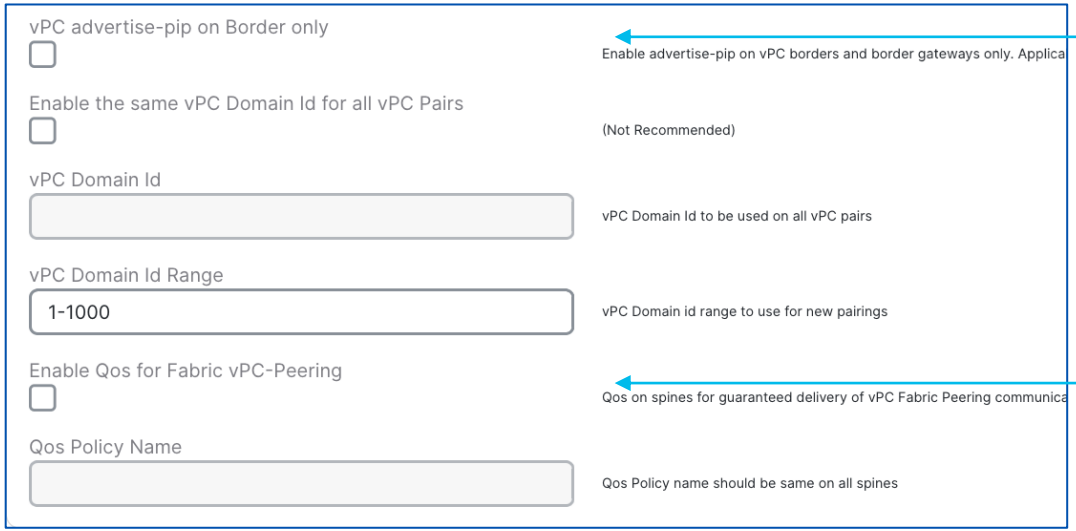
PKA (Mgmt/Loopback)

Auto Recovery and Delay Restore timers

Define Peer-link ID

Enable advertise-pip

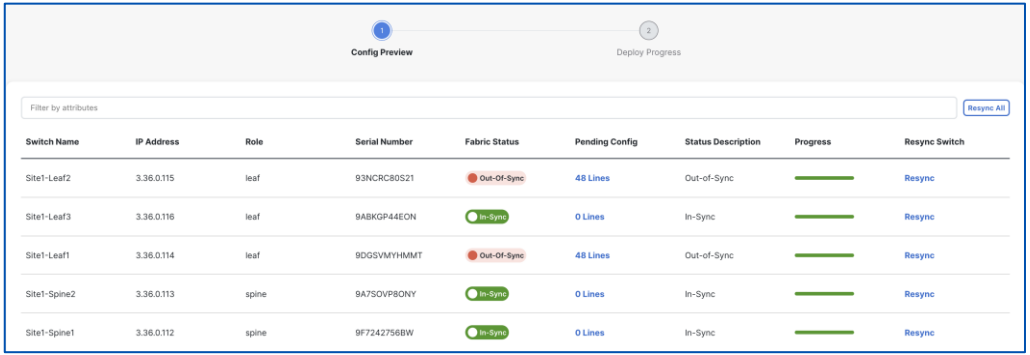
NDFC VXLAN Fabric Template : vPC Parameters

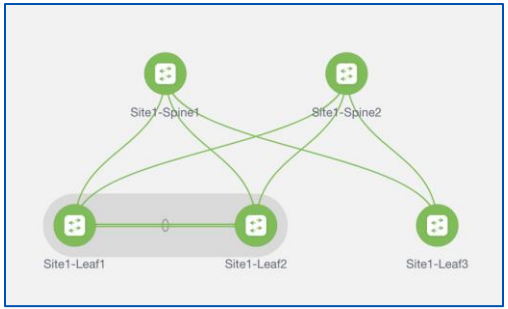


Advertise-pip only for Border

Enable QoS on Spines for VPC Fabric peering

NDFC : vPC Peering





NDFC : vPC Config Generated

Site1-Leaf-1	Site1-Leaf-2

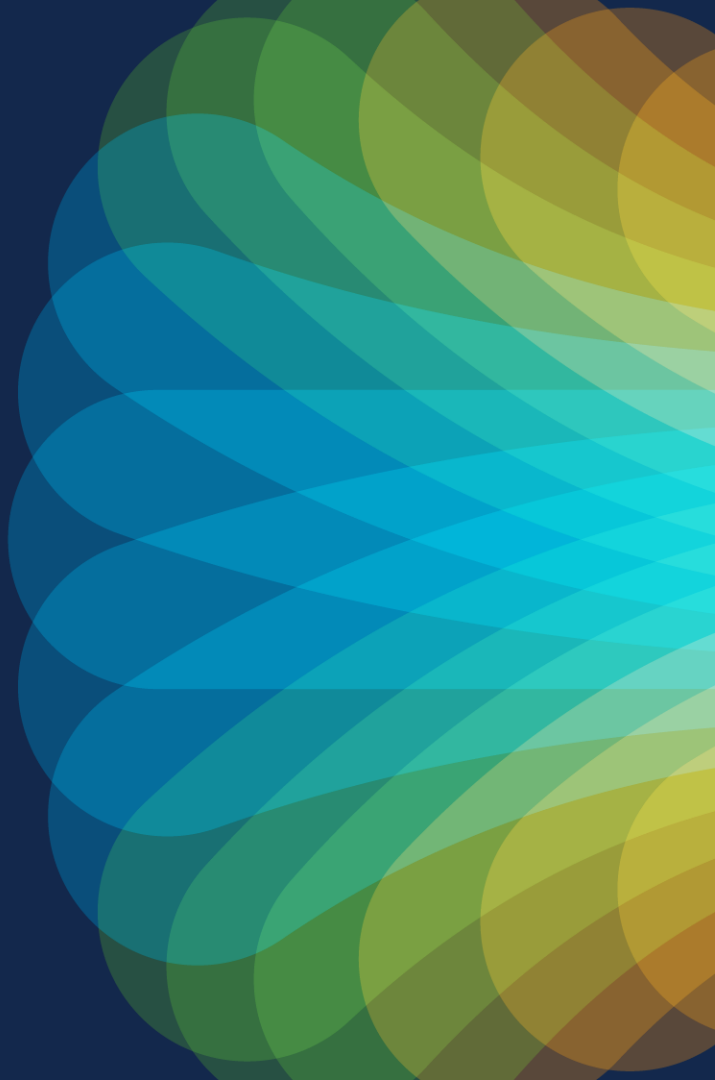
feature vpc
vpc domain 1
  ip arp synchronize
  peer-gateway
  peer-switch
  delay restore 150
  peer-keepalive destination 3.36.0.115 source 3.36.0.114
  auto-recovery reload-delay 360
  ipv6 nd synchronize
interface port-channel500
  switchport
  switchport mode trunk
  spanning-tree port type network
  description "vpc-peer-link Site1-Leaf1--Site1-Leaf2"
  no shutdown
  vpc peer-link
interface ethernet1/3
  channel-group 500 force mode active
  description "PO 500 (vpc-peer-link) member Site1-Leaf1-Ethernet1/3 to Site1-Leaf2-Ethernet1/3"
  no shutdown
interface ethernet1/4
  channel-group 500 force mode active
  description "PO 500 (vpc-peer-link) member Site1-Leaf1-Ethernet1/4 to Site1-Leaf2-Ethernet1/4"
  no shutdown
	feature vpc
	vpc domain 1
	  ip arp synchronize
	  peer-gateway
	  peer-switch
	  delay restore 150
	  peer-keepalive destination 3.36.0.114 source 3.36.0.115
	  auto-recovery reload-delay 360
	  ipv6 nd synchronize
	interface port-channel500
	  switchport
	  switchport mode trunk
	  spanning-tree port type network
	  description "vpc-peer-link Site1-Leaf2--Site1-Leaf1"
	  no shutdown
	  vpc peer-link
	interface ethernet1/3
	  channel-group 500 force mode active
	  description "PO 500 (vpc-peer-link) member Site1-Leaf2-Ethernet1/3 to Site1-Leaf1-Ethernet1/3"
	  no shutdown
	interface ethernet1/4
	  channel-group 500 force mode active
	  description "PO 500 (vpc-peer-link) member Site1-Leaf2-Ethernet1/4 to Site1-Leaf1-Ethernet1/4"
	  no shutdown

NDFC : vPC Config Generated

Site1-Leaf-1	Site1-Leaf-2

vlan 3600
interface loopback1
  ip address 10.3.0.3/32
  ip address 10.3.0.4/32 secondary
  description VTEP loopback interface
  ip router ospf UNDERLAY area 0.0.0.0
  no shutdown
  ip pim sparse-mode
interface Vlan3600
  ip address 10.4.0.26/30
  no ip redirects
  no ipv6 redirects
  mtu 9216
  ip router ospf UNDERLAY area 0.0.0.0
  ip ospf network point-to-point
  ip pim sparse-mode
  description VPC-Peer-Link SVI
  no shutdown
exit
	vlan 3600
	interface loopback1
	  ip address 10.3.0.2/32
	  ip address 10.3.0.4/32 secondary
	  ip router ospf UNDERLAY area 0.0.0.0
	  ip pim sparse-mode
	  description VTEP loopback interface
	  no shutdown
	interface Vlan3600
	  ip address 10.4.0.25/30
	  no ip redirects
	  no ipv6 redirects
	  mtu 9216
	  ip router ospf UNDERLAY area 0.0.0.0
	  ip ospf network point-to-point
	  ip pim sparse-mode
	  description VPC-Peer-Link SVI
	  no shutdown
	exit

Key Takeaways



Key Takeaways

vPC is the best way to connect End-hosts and LAN Segments to Leafs with all active paths, and without spanning tree loops

vPC fabric peering is best way to build vPC in VXLAN fabric as it removes need for dedicating front panel ports for vPC peer-link and reduces traffic over peer-link

vPC BGW provides L2/L3 DCI by utilizing VXLAN multisite for greenfield, small sites and legacy sites

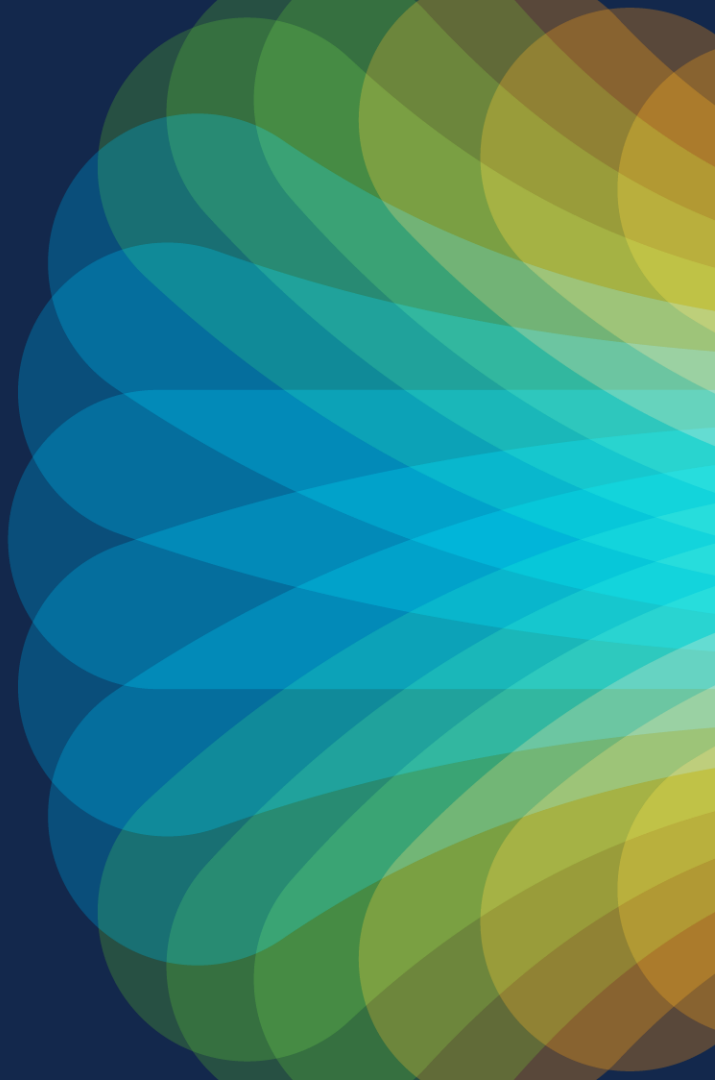
NDFC allows the automation of fabric and vPC configuration with all best practices included in configuration templates

Resources

- Understand and Configure Nexus 9000 vPC with Best Practices
- vPC Fabric Peering Configuration
- Configure VXLAN
- NextGen DCI with VXLAN EVPN Multi-Site Using vPC Border Gateways
- VXLAN EVPN Multi-Site Design and Deployment White Paper
- Migrating Classic Ethernet Environments to VXLAN BGP EVPN
- Nexus 9000 VXLAN Configuration Guide, Release 10.4(x)
- VXLAN Network with MP-BGP EVPN Control Plane Design Guide
- Build Hierarchical Fabrics with VXLAN EVPN Multi-Site
- Migrating Cisco FabricPath Environments to VXLAN BGP EVPN White Paper

The bridge to possible

Thank you



cisco Live!


cisco Live!
Let's go


cisco Live!
Let's go