CISCO

You make **possible**

VXLAN vPC: Design and Best Practices

Nemanja Kamenica

BRKDCN-2249

Barcelona | January 27-31, 2020

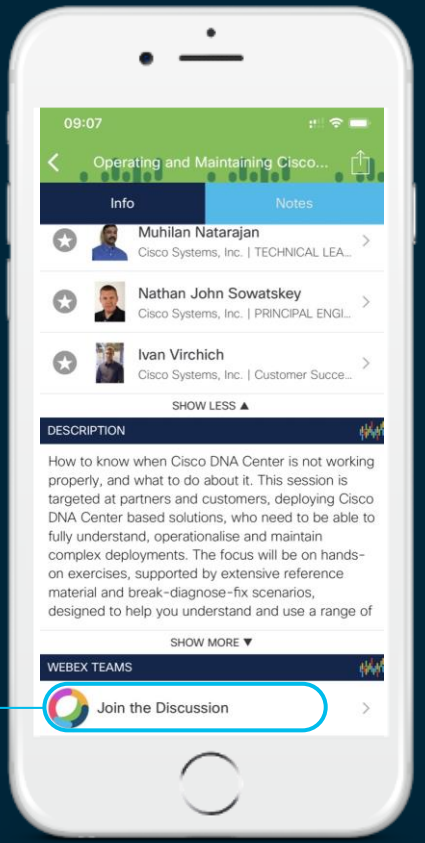
Cisco Webex Teams

Questions?

Use Cisco Webex Teams 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



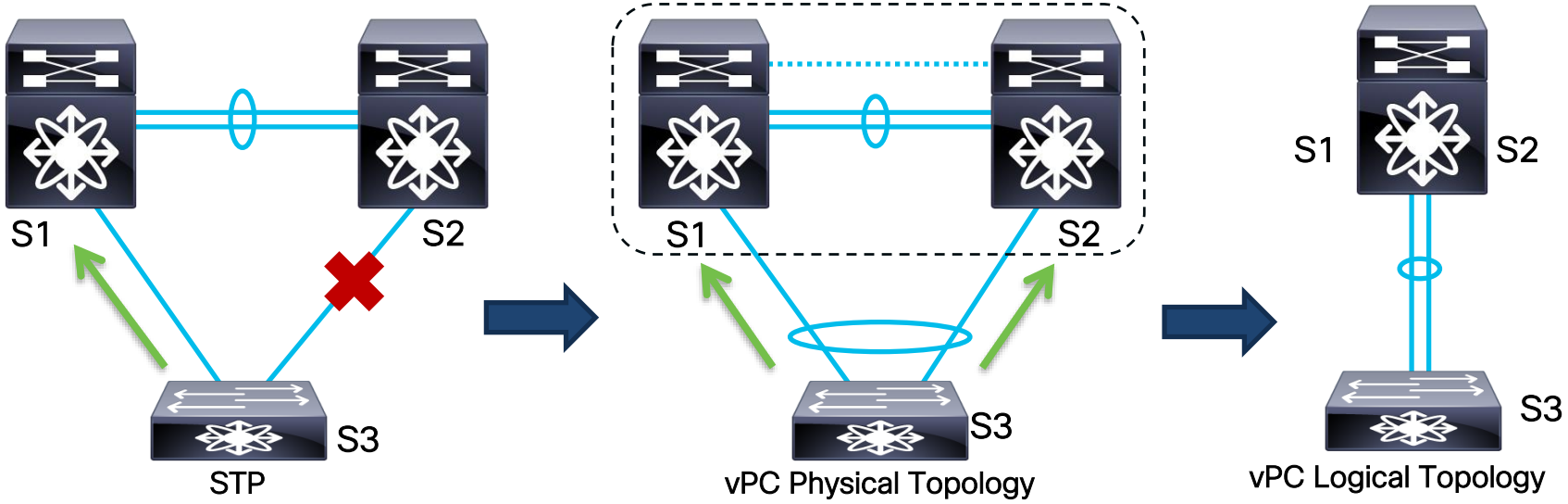
Session Objectives

- Provide refresh of vPC basics
- Understand the basic of VXLAN, including underlay, overlay and overlay control plane
- Provide a detailed understanding of vPC in VXLAN environment
- vPC as Border Gateway for Multi-Site deployments introduction
- Benefits of IP and MAC ECMP

Agenda

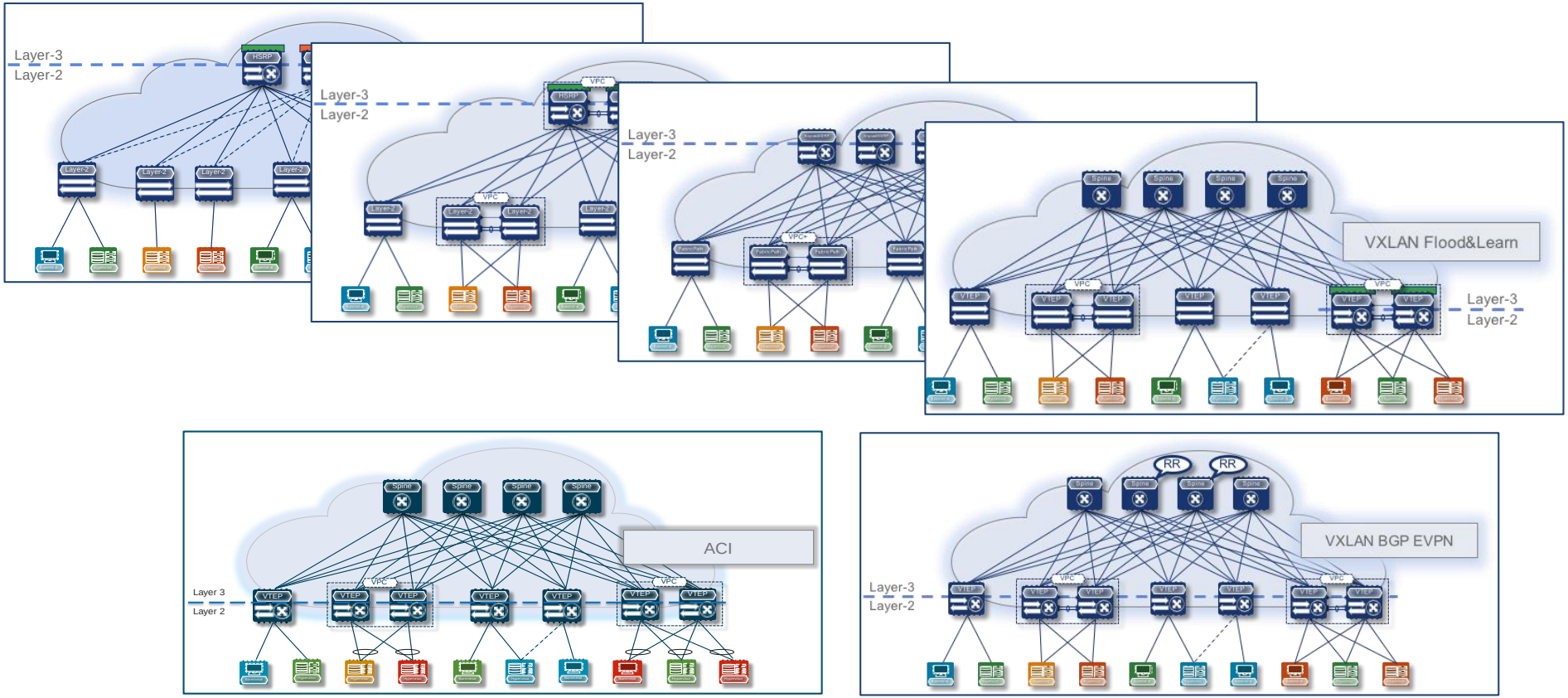
- Introduction
- vPC Basics
- vPC VXLAN Basics
- vPC Fabric Peering
- vPC Boarder Gateway
- MAC ECMP and IP ECMP
- Key Takeaways

Why vPC?



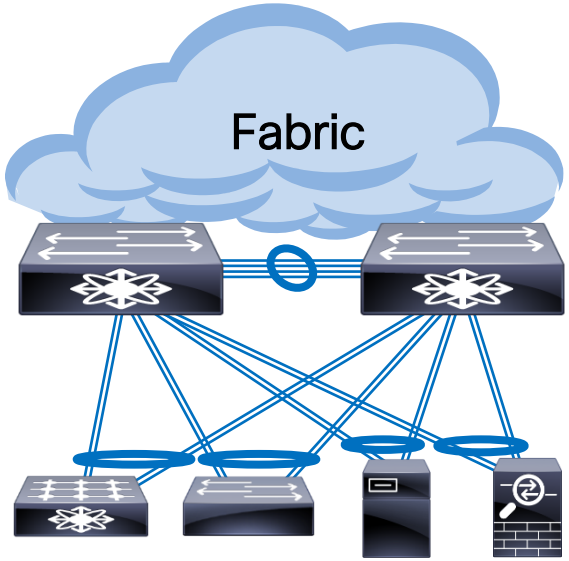
- No Blocked Ports
- More Usable Bandwidth

vPC everywhere!



Virtual Port Channel - vPC

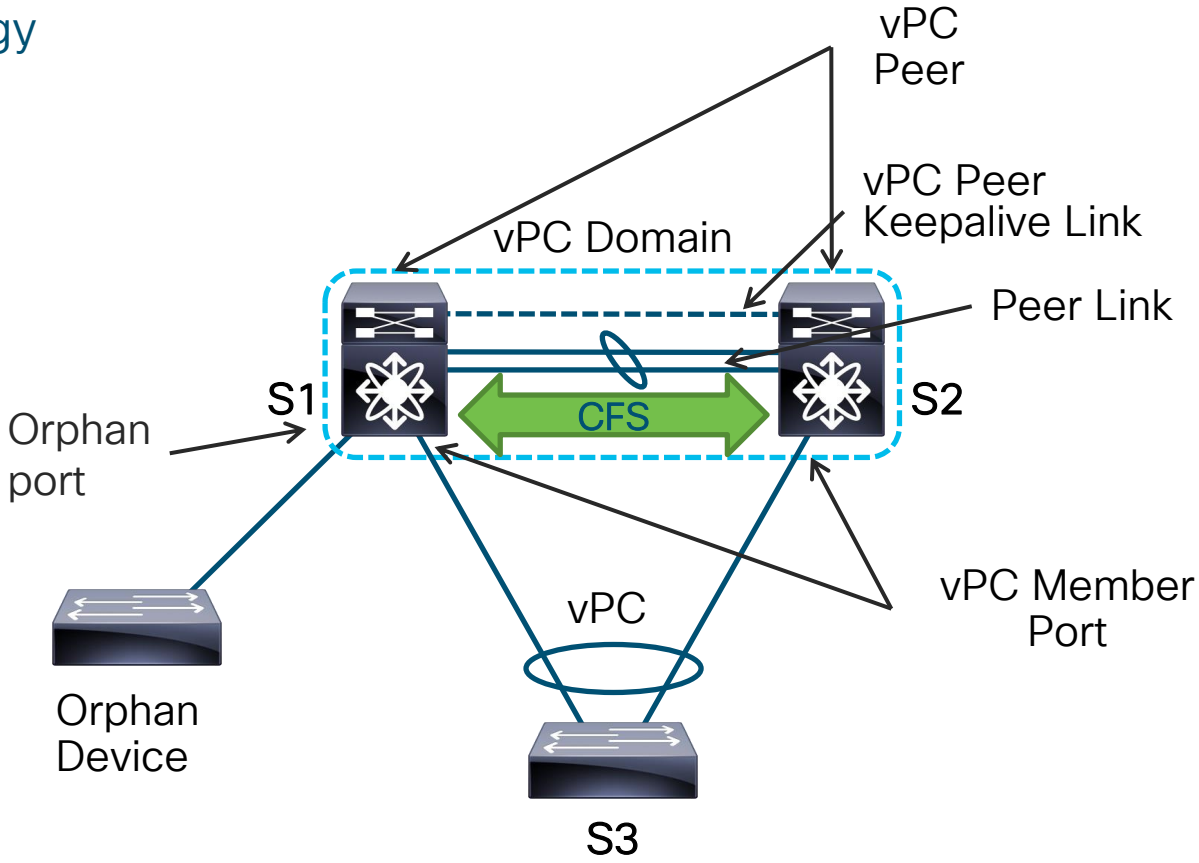
- Eliminates Spanning Tree blocked ports by providing loop-free topology
- Full link bandwidth utilization
- Provides device level redundancy
- Faster convergence over STP
- MC-LAG on Cisco Nexus Devices



vPC Basics

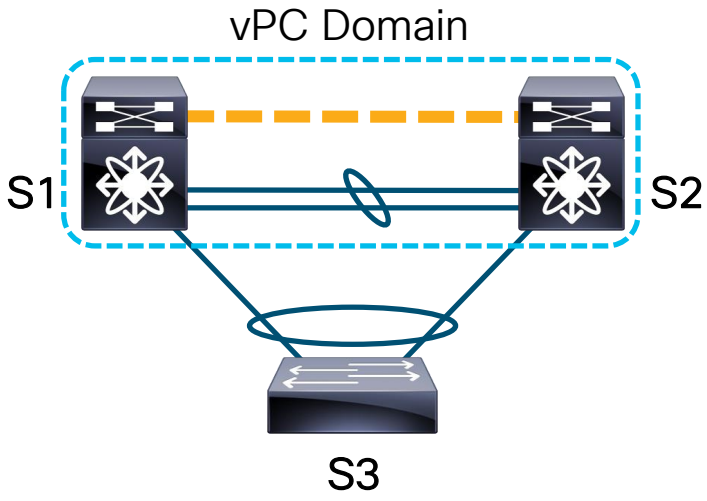
vPC Overview

Terminology



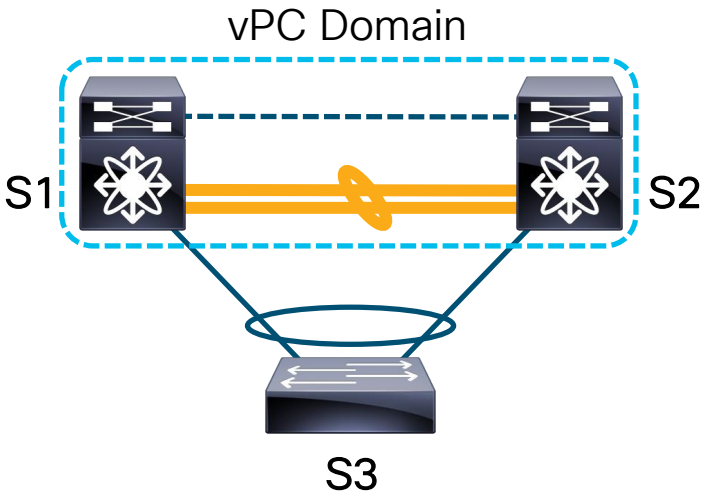
vPC Peer-keepalive link

- 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



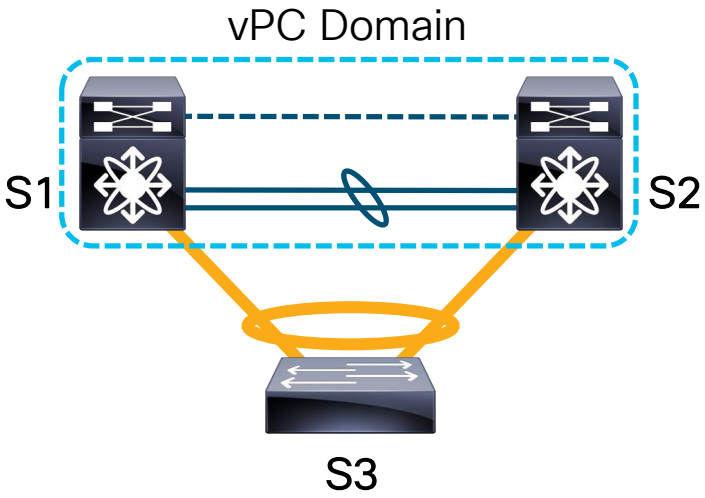
vPC Peer link

- 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



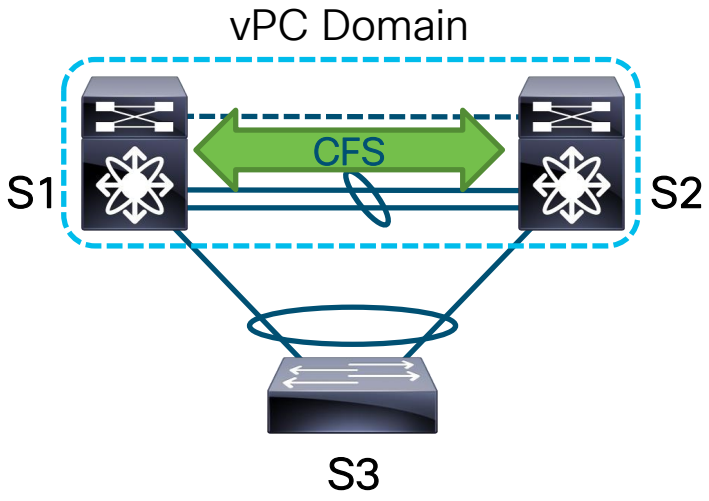
Virtual Port Channel – vPC

- 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



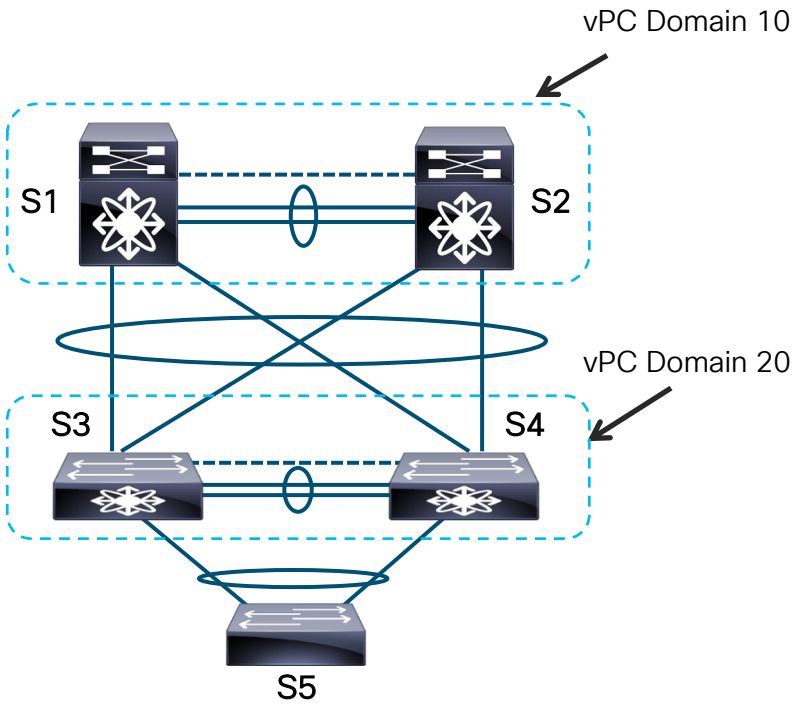
vPC Configuration Best Practices

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
```

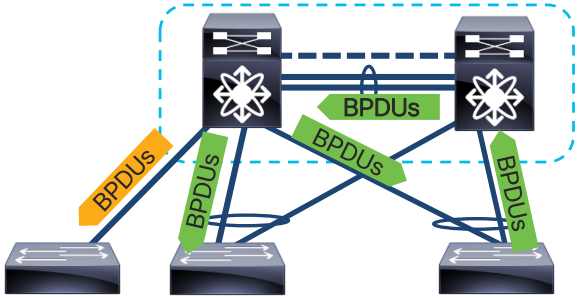


vPC Consistency check

- System configuration must be in sync
- Type 1 Consistency Check
  - Graceful Consistency check – suspends:
    - Per-interface inconsistent parameters – vPC member ports on secondary peer set to down state
    - Globally inconsistent parameters – misconfigured member ports on secondary peer suspended
  - Parameters: STP mode, STP VLAN state, STP global settings, LACP mode, MTU…
- Type 2 Consistency Check
  - Forwards traffic in case of inconsistency
  - Possible undesirable traffic forwarding behavior
  - Parameters: VLAN interface (SVI), ACL, QOS, IGMP snooping, HSRP…

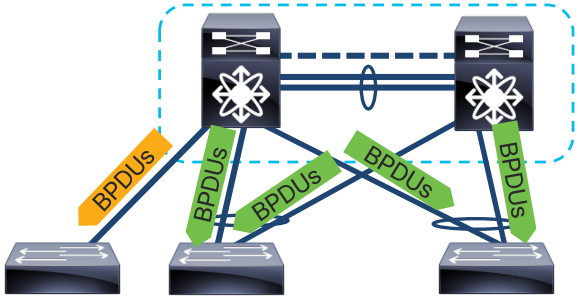
vPC Peer-Switch

- 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

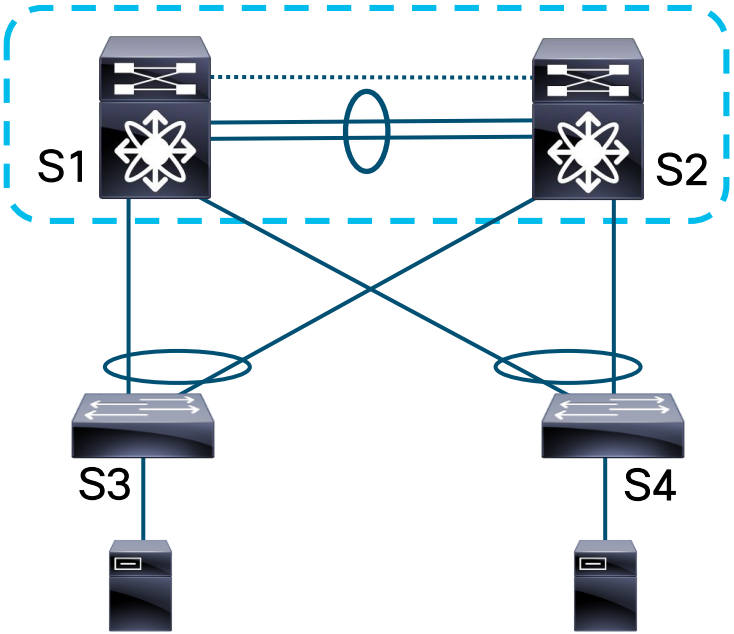
Nexus(config-vpc-domain)# peer-switch



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

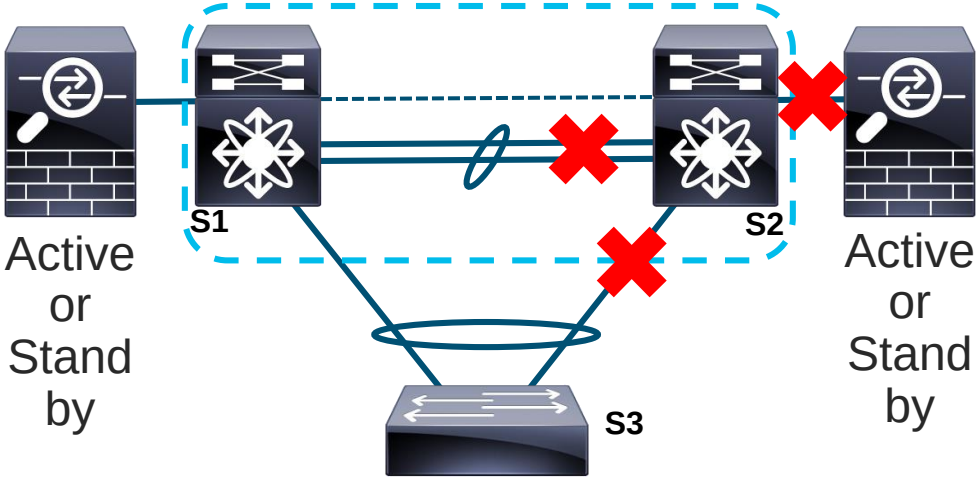
Nexus(config-vpc-domain)# peer-gateway



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

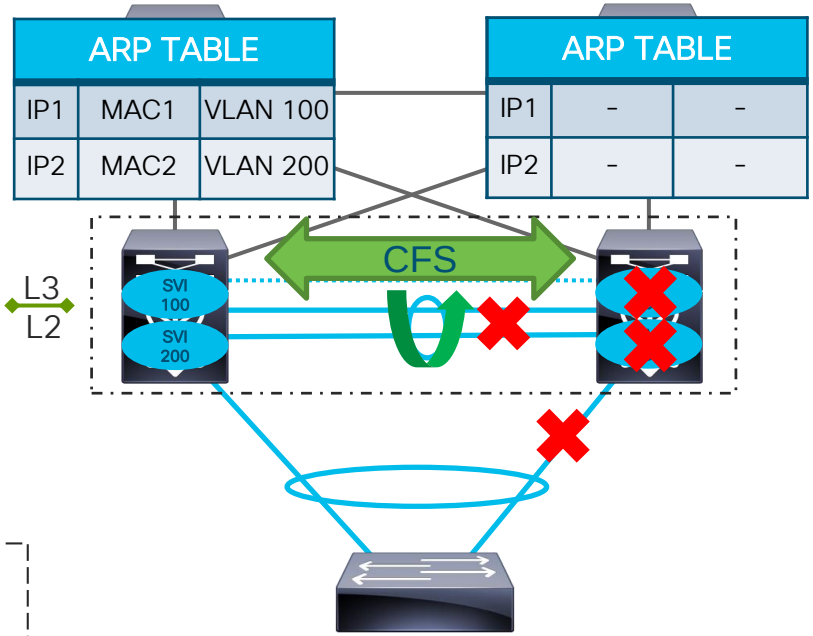
Nexus(config-if)# vpc orphan-ports suspend



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



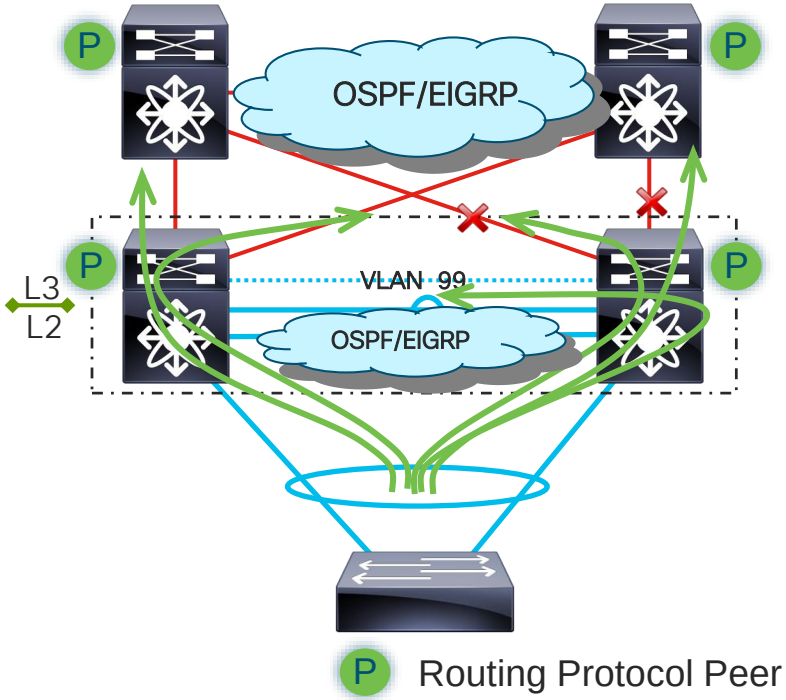

```
Nexus(config-vpc-domain)# ip arp synchronize
Nexus(config-vpc-domain)# ipv6 nd synchronize
```

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

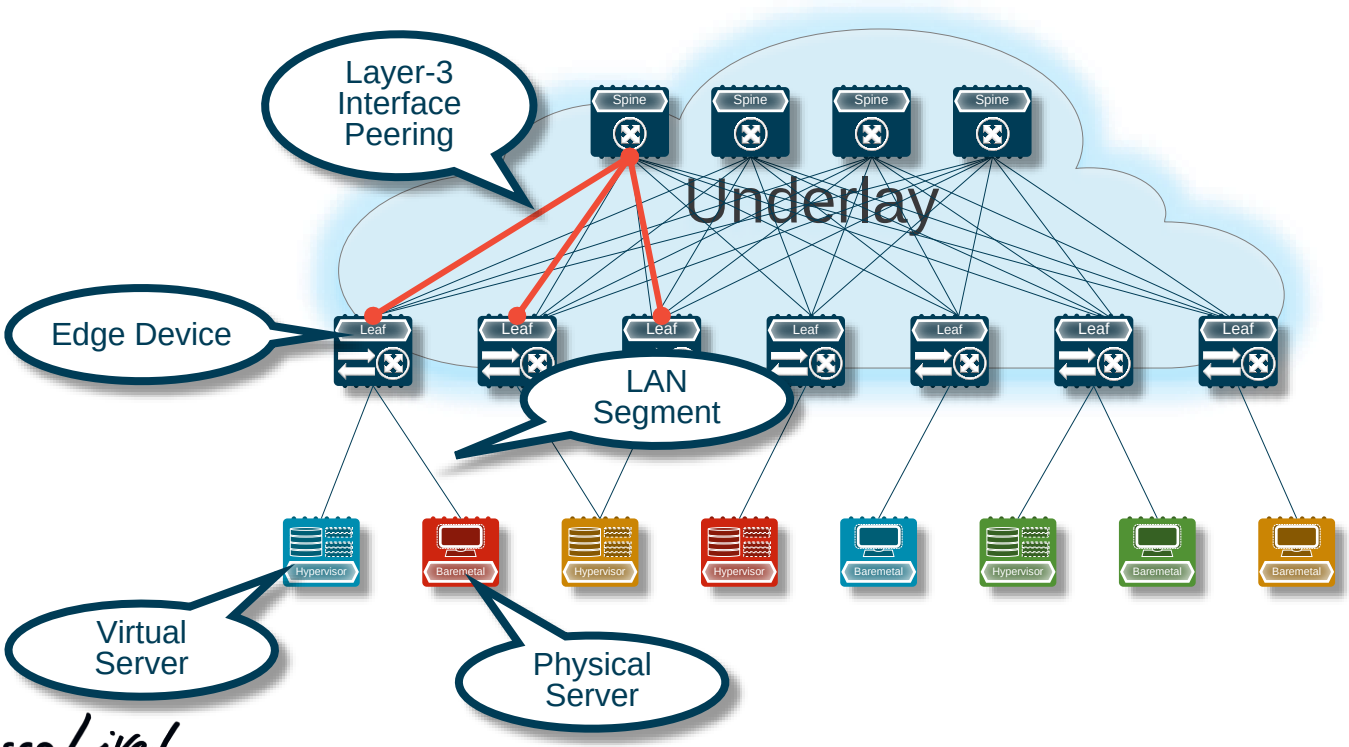


vPC VXLAN Basics

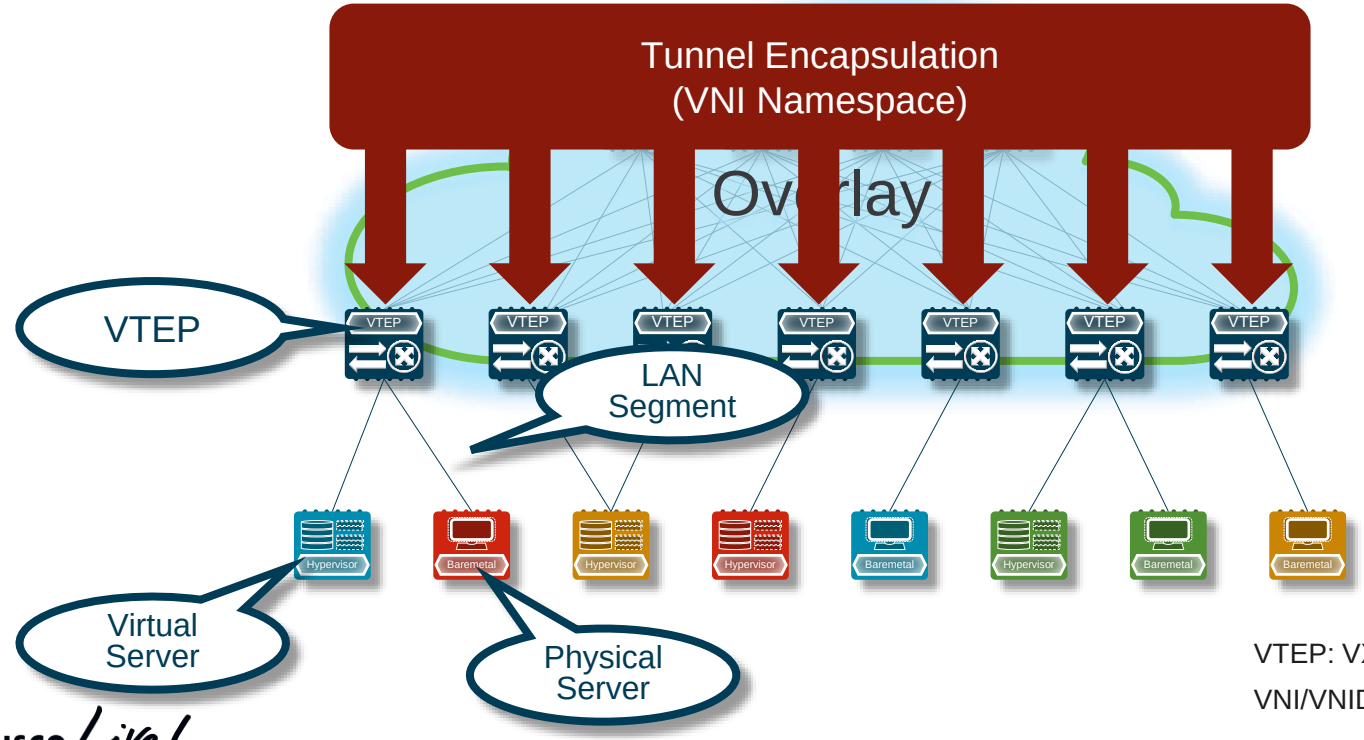
VXLAN Basics

VXLAN – Underlay

Taxonomy



VXLAN – Overlay



VTEP: VXLAN Tunnel End-Point

VNI/VNID: VXLAN Network Identifier

Overlay Technologies

Overlay Services
- Layer-2
- Layer-3
- Layer-2 and Layer-3

Tunnel Encapsulation

Underlay Transport Network

Control-Plane
- Peer-Discovery
- Route Learning and Distribution
  - Local Learning
  - Remote Learning

Data-Plane
- Overlay Layer-2/Layer-3 Unicast Traffic
- Overlay Broadcast, Unknown Unicast, Multicast traffic (BUM traffic) forwarding
  - Ingress Replication (Unicast)
  - Multicast

VXLAN – the Nuts and Bolts

What is VXLAN?

- ✓ VXLAN is a network overlay technology
- ✓ VXLAN builds Layer-2 & Layer-3 overlay networks on top of an IP routed network
- ✓ VXLAN uses MAC in UDP encapsulation (UDP destination port 4789)

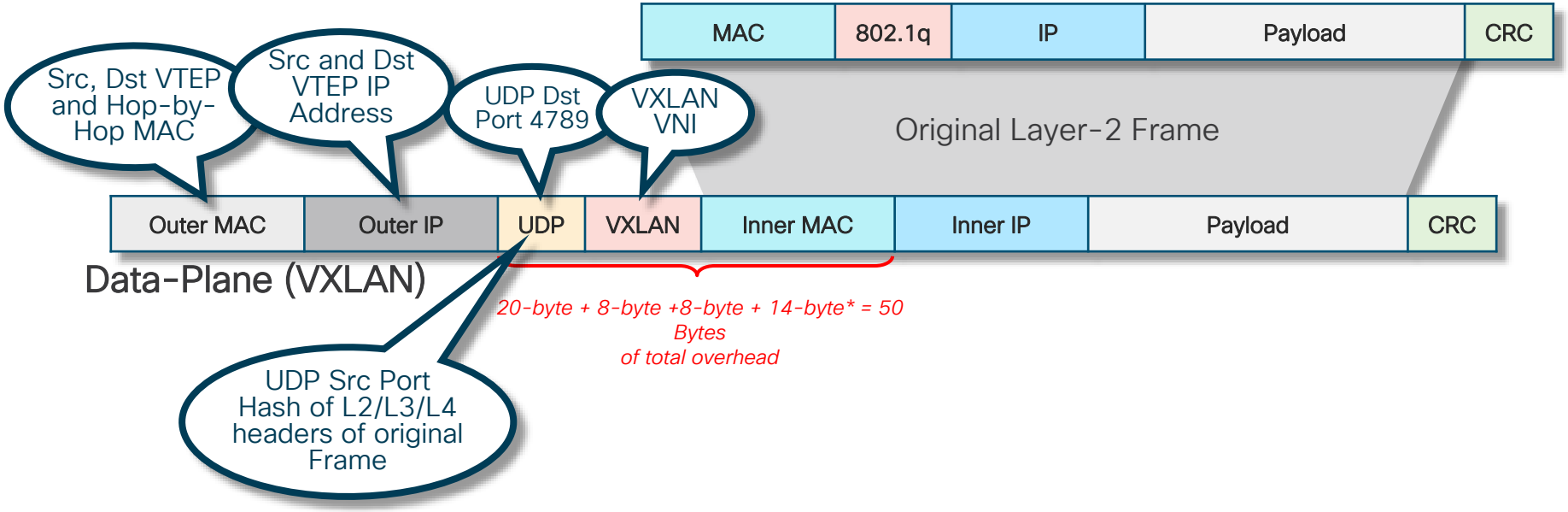
VXLAN Use Cases

- ✓ Extended Namespace
- ✓ Scalable L2 Domains
- ✓ Integrated Router and Switch
- ✓ Hybrid Overlays
- ✓ DCI Connectivity
- ✓ Multi-Tenancy

Why VXLAN?

- ✓ VXLAN provides a Network with Segmentation, IP Mobility, and Scale
- ✓ “Standards” based Overlay
- ✓ Leverages Layer-3 ECMP – all links forwarding
- ✓ Increased Name-Space to 16M identifier
- ✓ Segmentation and Multi-Tenancy
- ✓ Integration of Physical and Virtual

VXLAN – Data Plane

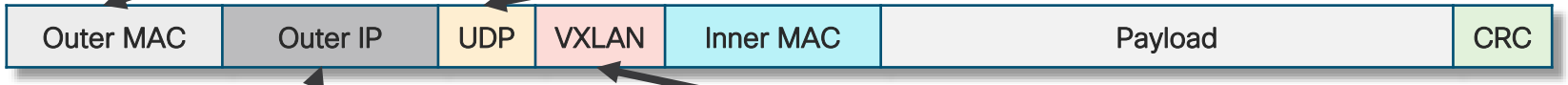


*plus 4-byte if IEEE 802.1q exists as part of Inner MAC Header

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	

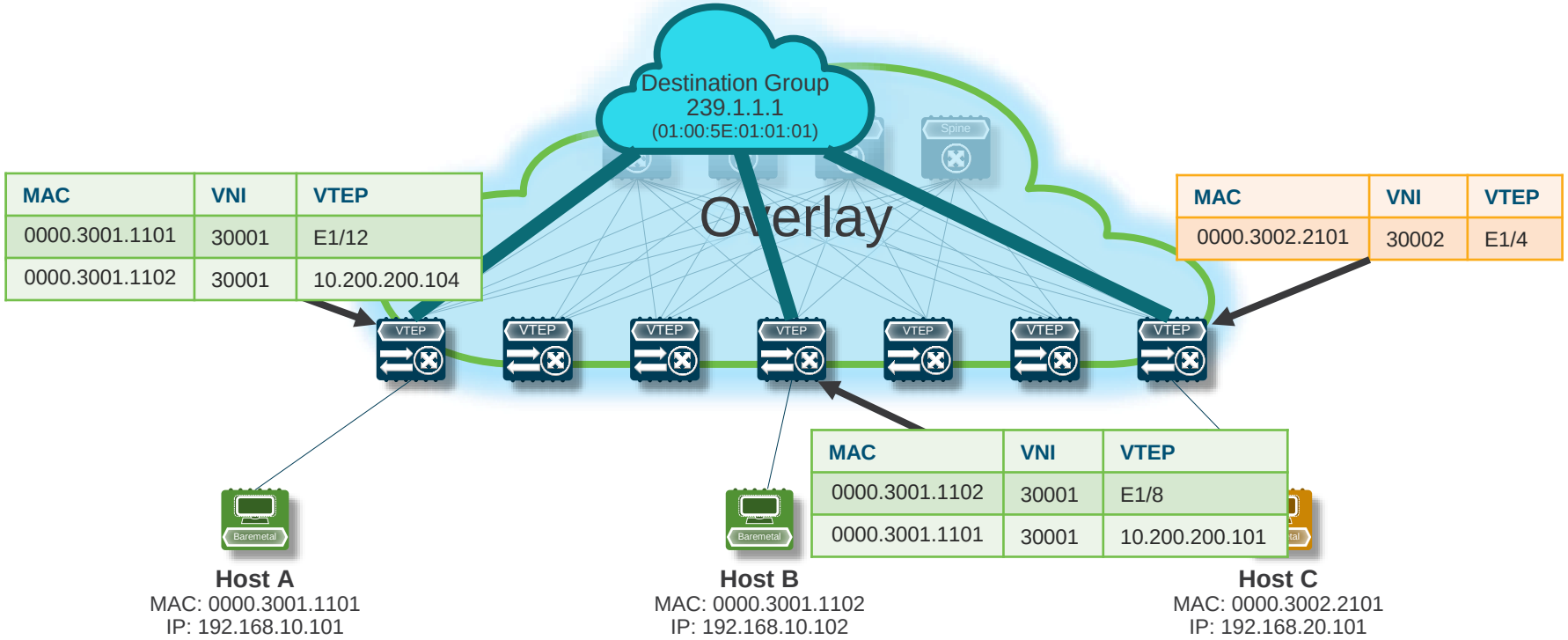
Field	Value	Bites	Total
Source Port	L2/L3/L4 Hash	16	8 Bytes
Destination Port	4789 (UDP)	16	
UDP Length		16	
Checksum	0x0000	16	



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	

VXLAN Flood & Learn

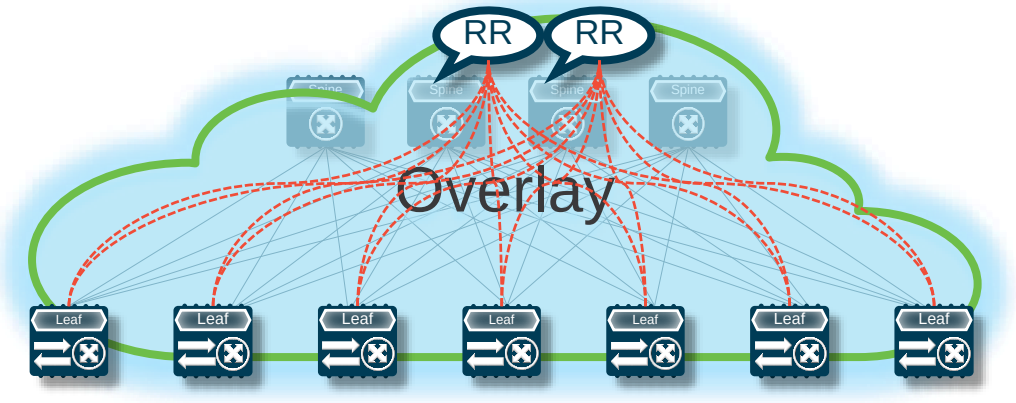


MAC	VNI	VTEP
0000.3001.1101	30001	E1/12
0000.3001.1102	30001	10.200.200.104

MAC	VNI	VTEP
0000.3002.2101	30002	E1/4

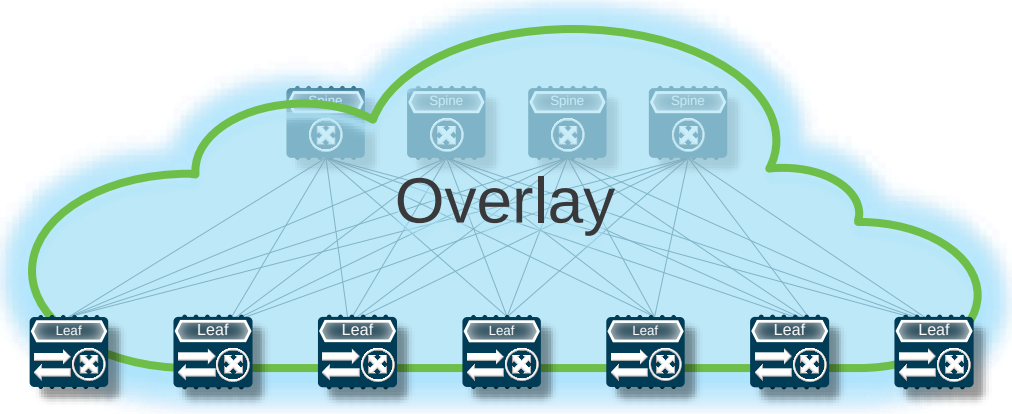
MAC	VNI	VTEP
0000.3001.1102	30001	E1/8
0000.3001.1101	30001	10.200.200.101

EVPN – Host and Subnet Route Distribution



- Host Route Distribution decoupled from the Underlay protocol
- Use MultiProtocol-BGP (MP-BGP) on the Leaf nodes to distribute internal Host/Subnet Routes and external reachability information
- Route-Reflectors (RR) deployed for scaling purposes

EVPN Control Plane – Host and Subnet Routes

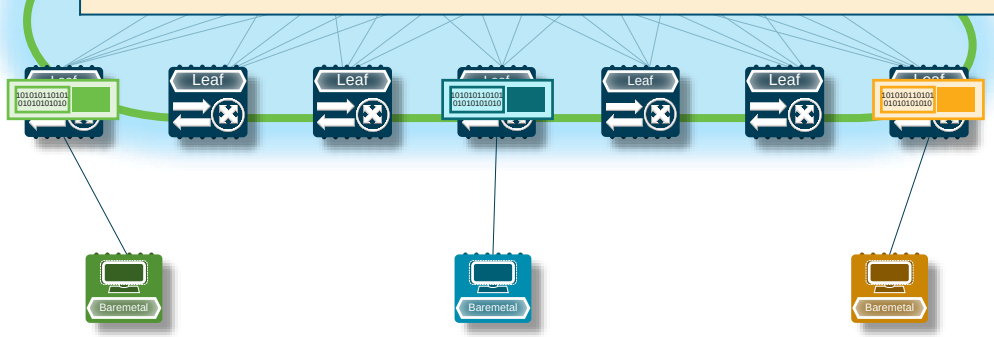


- BGP EVPN NLRI*
- Host MAC (Route Type 2)
  - MAC only, Single VNI, Single Route Target
- Host MAC+IP (Route Type 2)
  - MAC and IP, Two VNI, Two Route Target, Router MAC
- Internal and External Subnet Prefixes (Route Type 5)
  - IP Subnet Prefix, Single VNI, Single Route Target, Router MAC

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

Host Advertisements

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI /
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 A
MAC: 0000.3001.1101

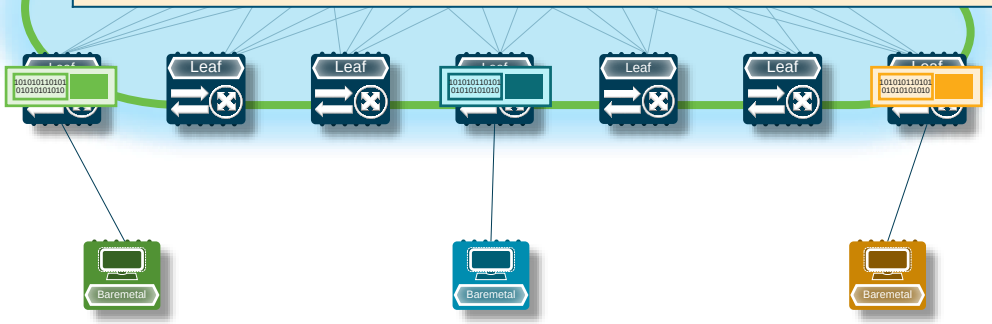
Host B
MAC: 0000.3001.1102

Host C
MAC: 0000.3002.2101

- Host MAC (Route Type 2)
  - MAC
  - L2VNI
  - Route Target for MAC-VRF
- MAC attributes are Mandatory

Host Advertisements

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

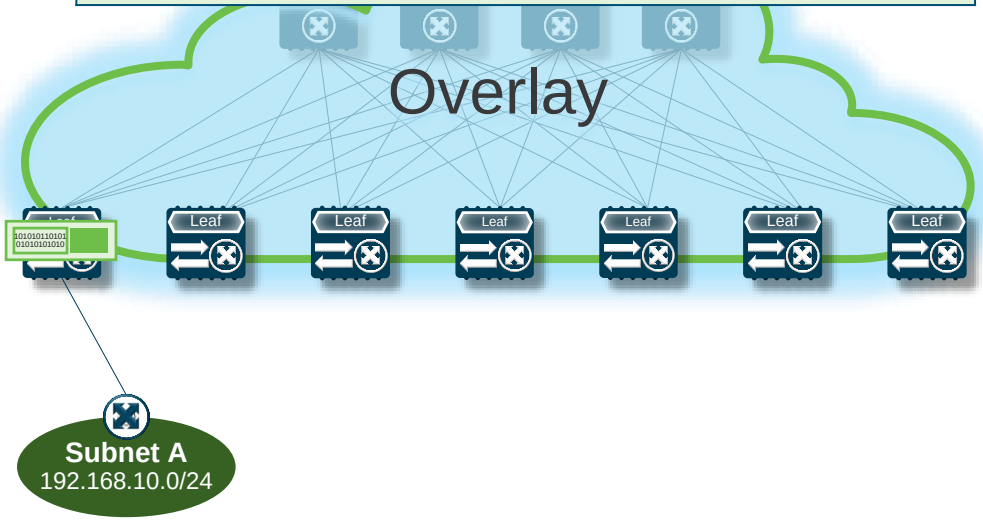
Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

- Host MAC+IP (Route Type 2)
  - MAC and IP
  - L2VNI
  - Route Target for MAC-VRF
  - L3VNI
  - Route Target for IP-VRF
  - Router MAC
- IP Attributes are Optional
- Populated through ARP/ND

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

Subnet Route Advertisements

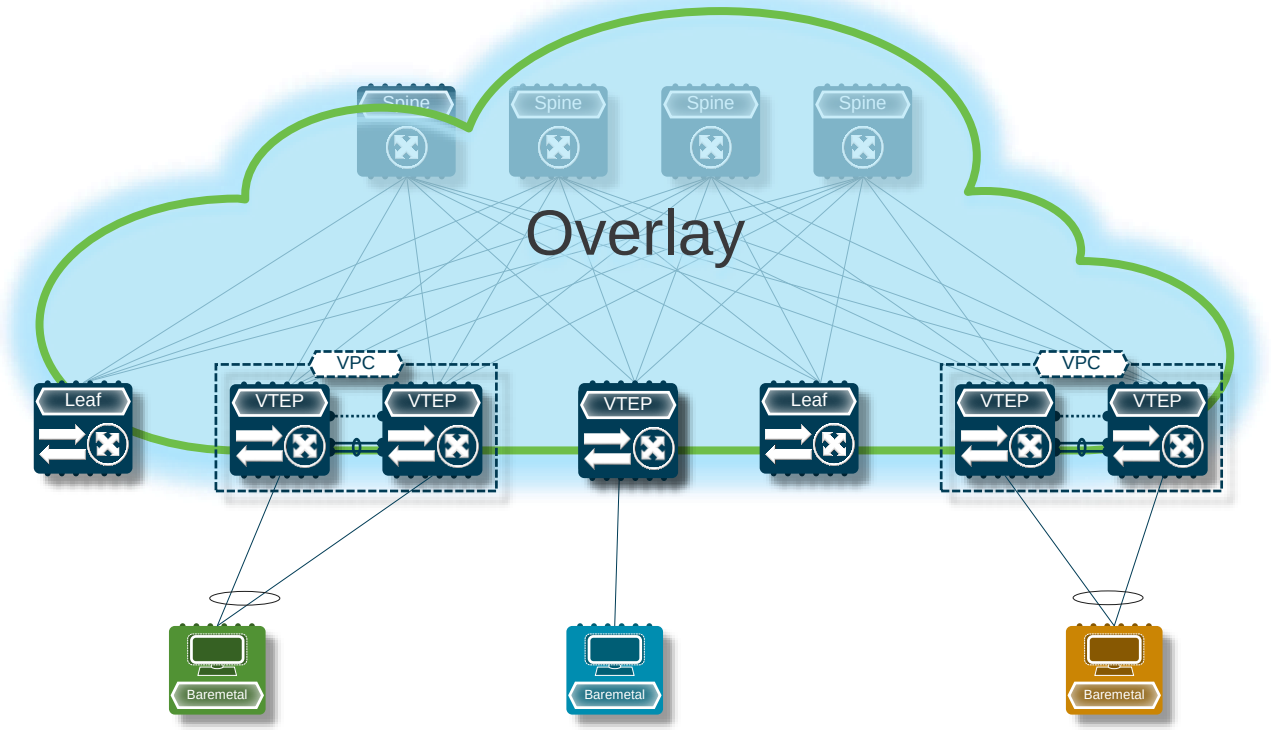
Type	IP / Length	L3VNI / RT	Next-Hop	Seq.
5	192.168.10.0 /24	5000, 65500:5000	10.200.200.101	



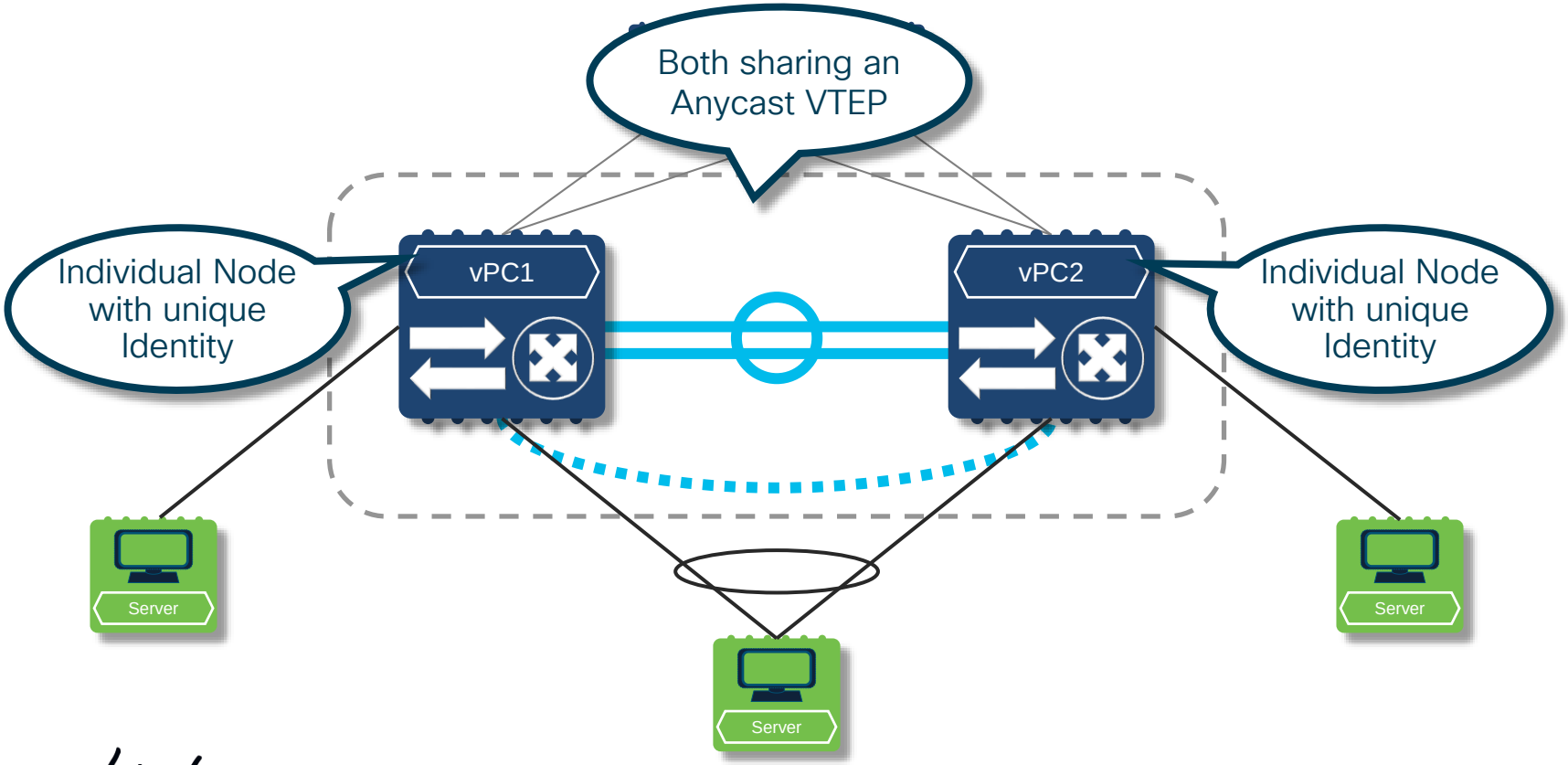
- Internal and External Subnet Prefixes (Route Type 5)
  - IP Prefix
  - L3VNI
  - Route Target for IP-VRF
  - Router MAC
- Populated through External Routing Protocol

vPC in VXLAN

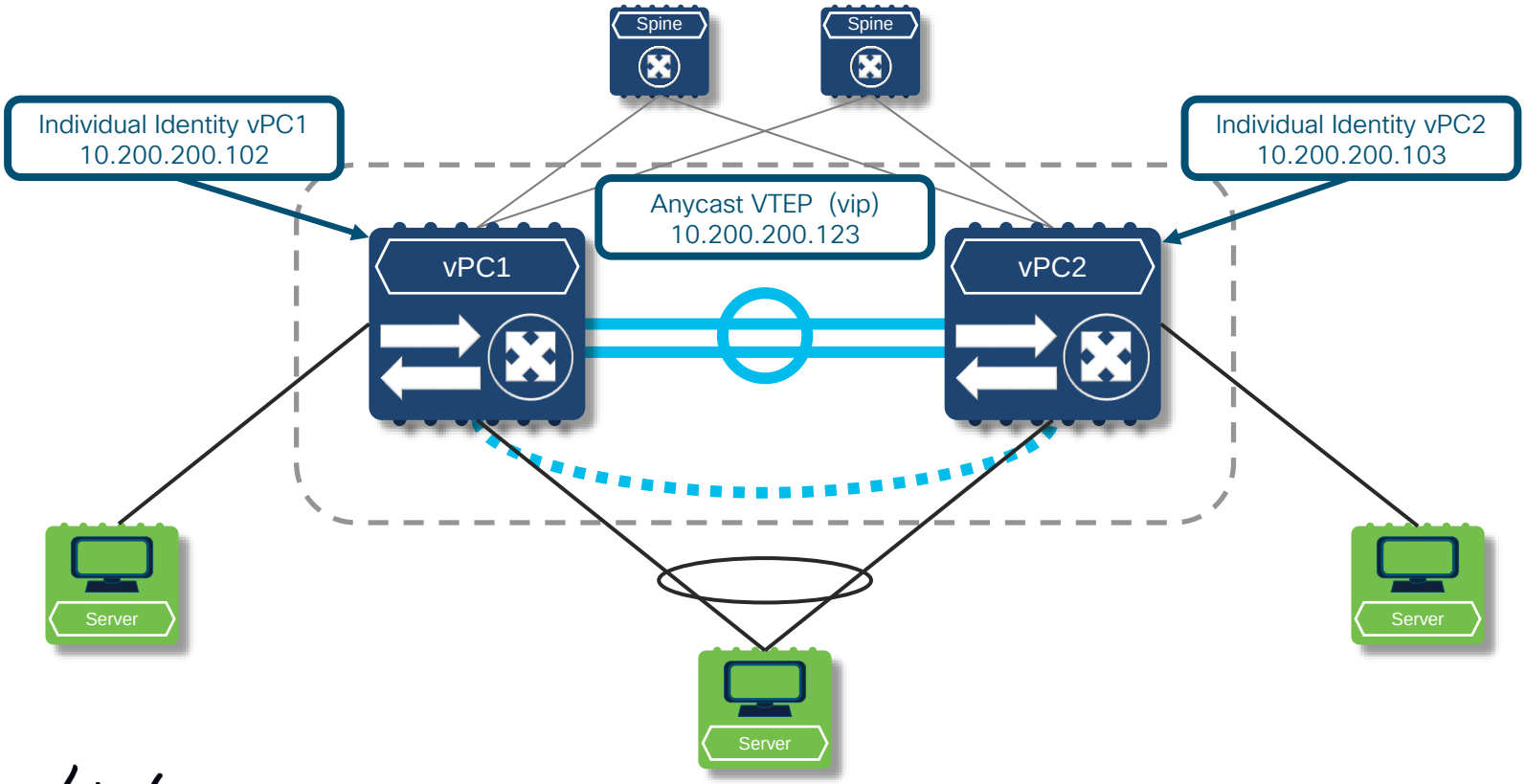
VXLAN – vPC



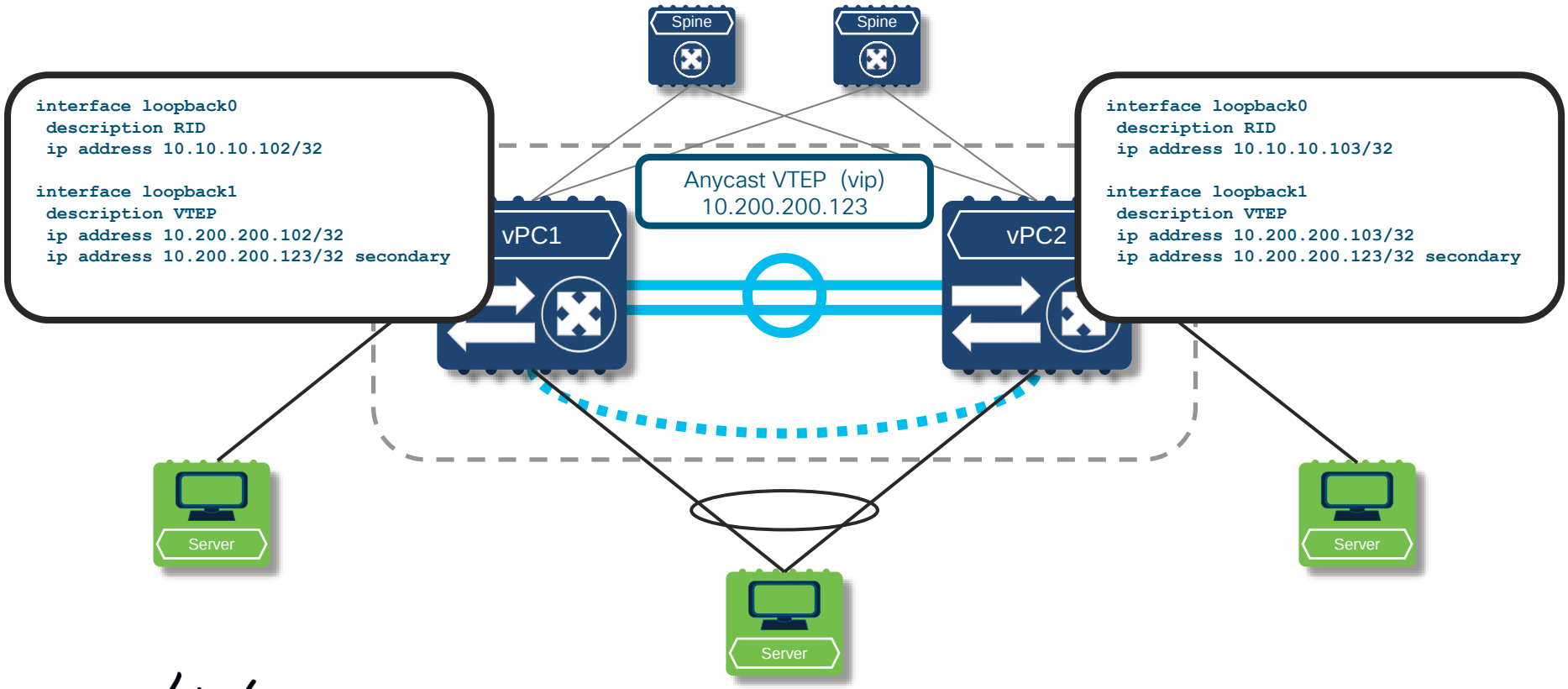
VXLAN – Anycast VTEP



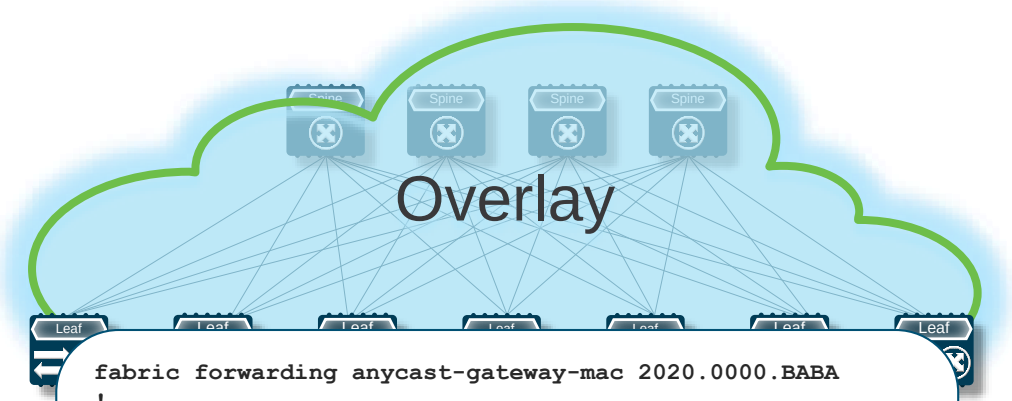
Anycast VTEP and Individual VTEPs



Anycast VTEP – Virtual IP Address



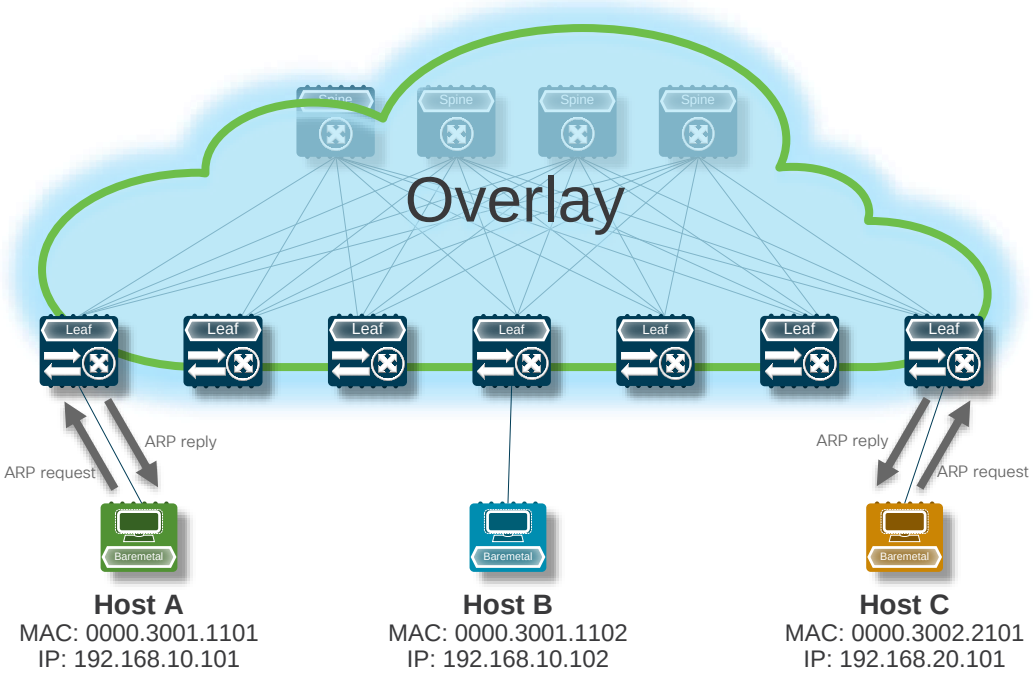
Distributed Anycast Gateway




```
fabric forwarding anycast-gateway-mac 2020.0000.BABA
!
interface Vlan100
vrf member myvrf_5000
ip address 192.168.10.1/24
fabric forwarding mode anycast-gateway
!
interface Vlan200
vrf member myvrf_5000
ip address 192.168.20.1/24
fabric forwarding mode anycast-gateway
```

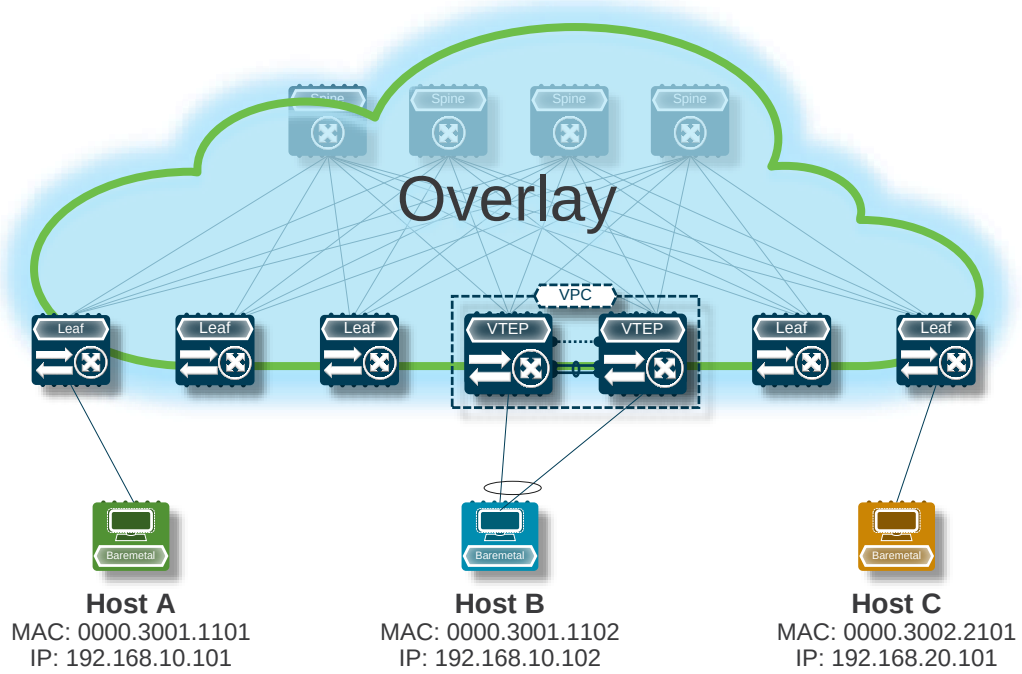
- Distributed First-Hop Routing on Edge Device
  - All Edge Device share same Gateway IP and MAC address
  - Pervasive Gateway approach
- Gateway is always active
  - No First-Hop redundancy protocol for hello or state exchange
- Distributed and smaller state
  - Only local End-Points ARP entries

Anycast – One to Nearest Association



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

Anycast Gateway - vPC

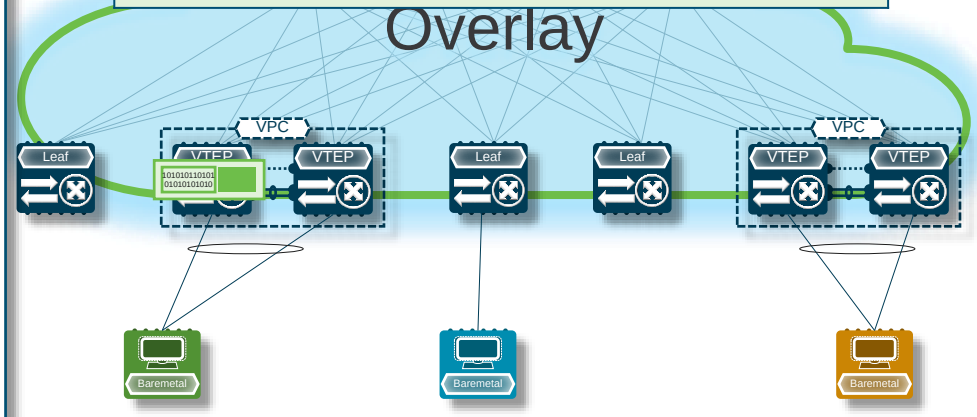


- Active / Active Forwarding with no First-Hop redundancy protocol (ie HSRP, VRRP)
  - ARP / ND Synchronized via vPC
  - Fast Failover
- Port-Channel Hashing avoids duplicate
  - Broadcast, Unknown Unicast, Multicast (BUM) hashed on single link

Host Advertisements with vPC – Flood & Learn

- Anycast VTEP for vPC domain
  - Both peers represented by VIP
  - Underlay ECMP Load Share to Anycast VTEP
- Hosts advertised by VIP
  - Dual-homed hosts
  - Orphan hosts

MAC Address	L2VNI	Next-Hop
0000.3001.1101	30001	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

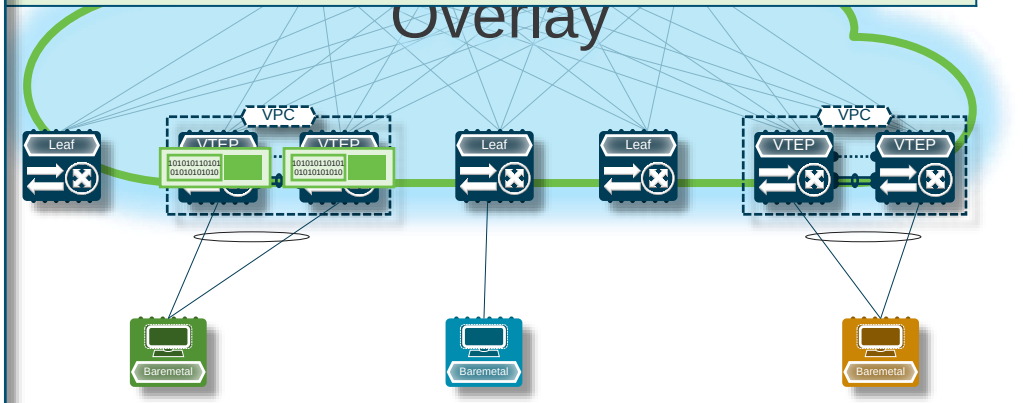
Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

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	

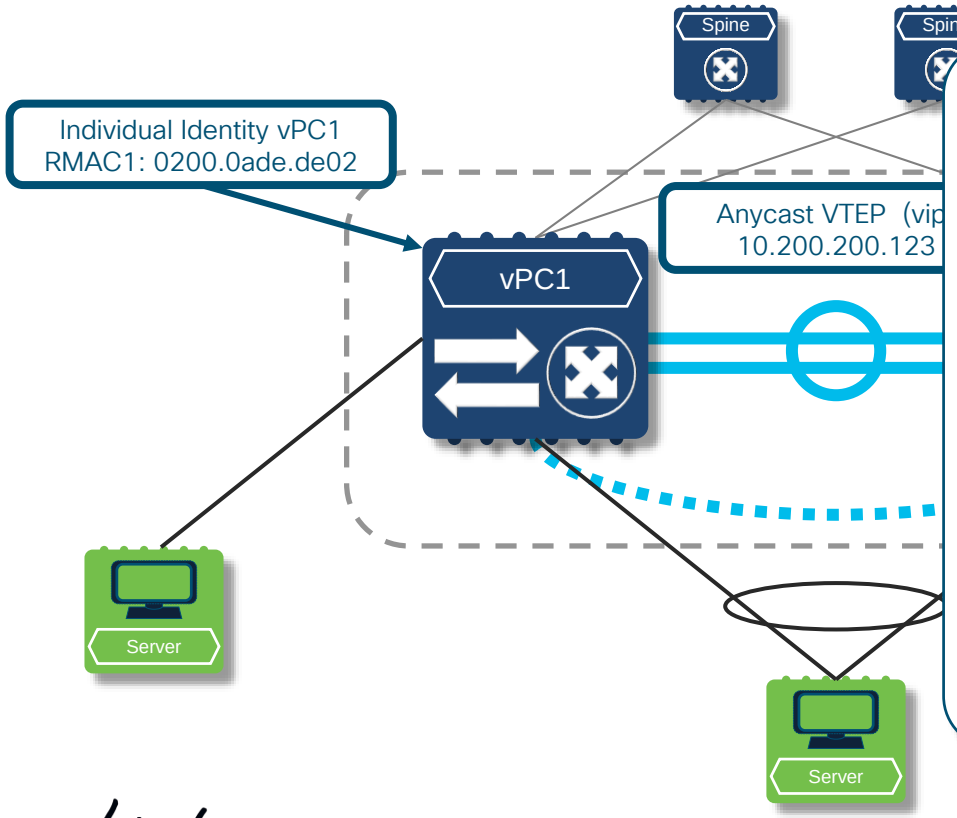


Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

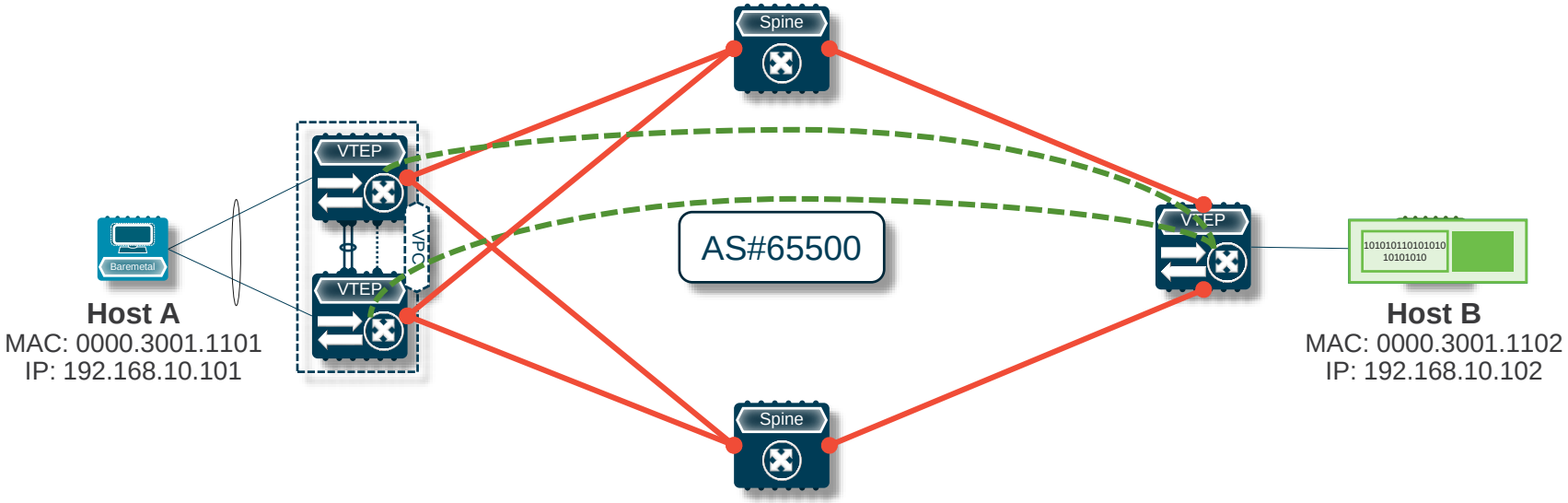
Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

Anycast VTEP – Virtual IP Address

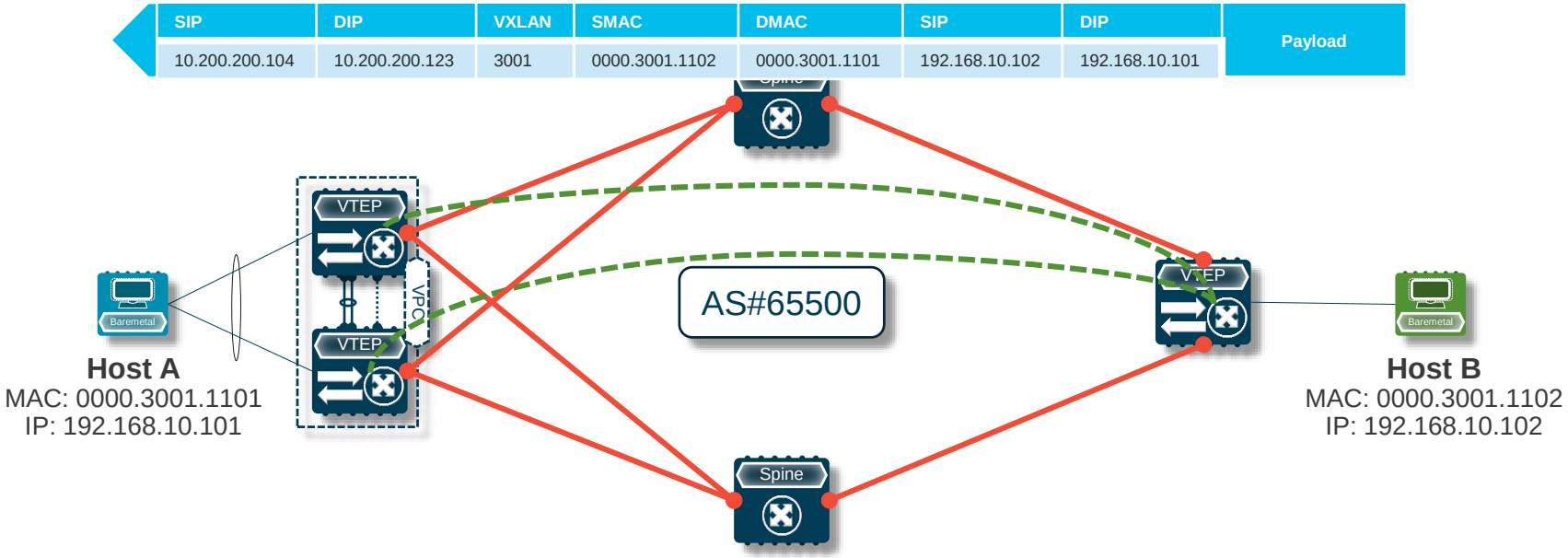


- Router MAC (RMAC) individual peer MAC address (Switch System MAC)
- vPC Type 2 and Type 5 routes will be advertise with (VIP, RMAC)
- Orphan Type 2 and Type 5 routes will be advertised with (VIP, RMAC)

ECMP to the Anycast VTEP – Underlay

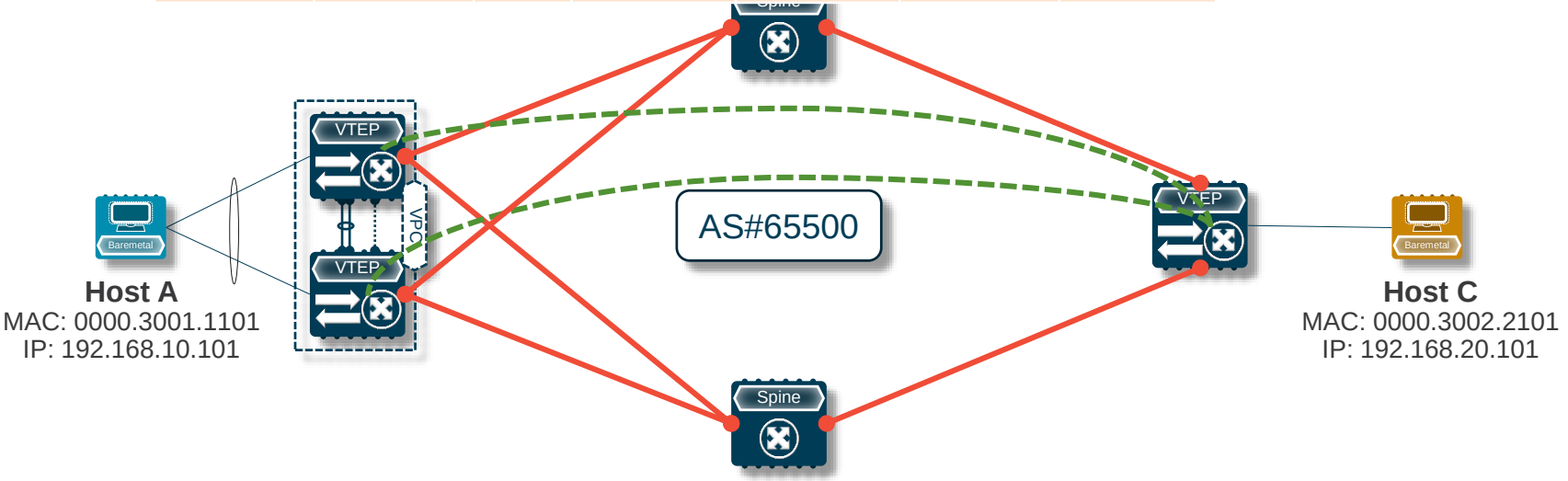


Bridging to a vPC Domain – VXLAN



Routing to a vPC Domain – VXLAN

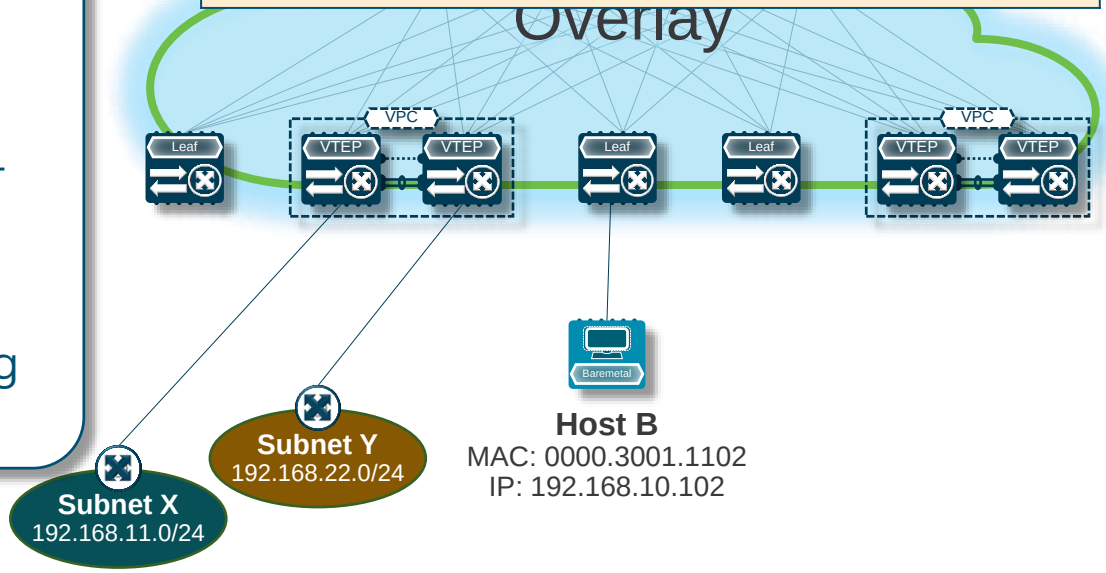
SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
10.200.200.107	10.200.200.123	5000	0200.0ade.de01	0200.0ade.de03	192.168.20.101	192.168.10.101	



Subnet Route Advertisement with vPC

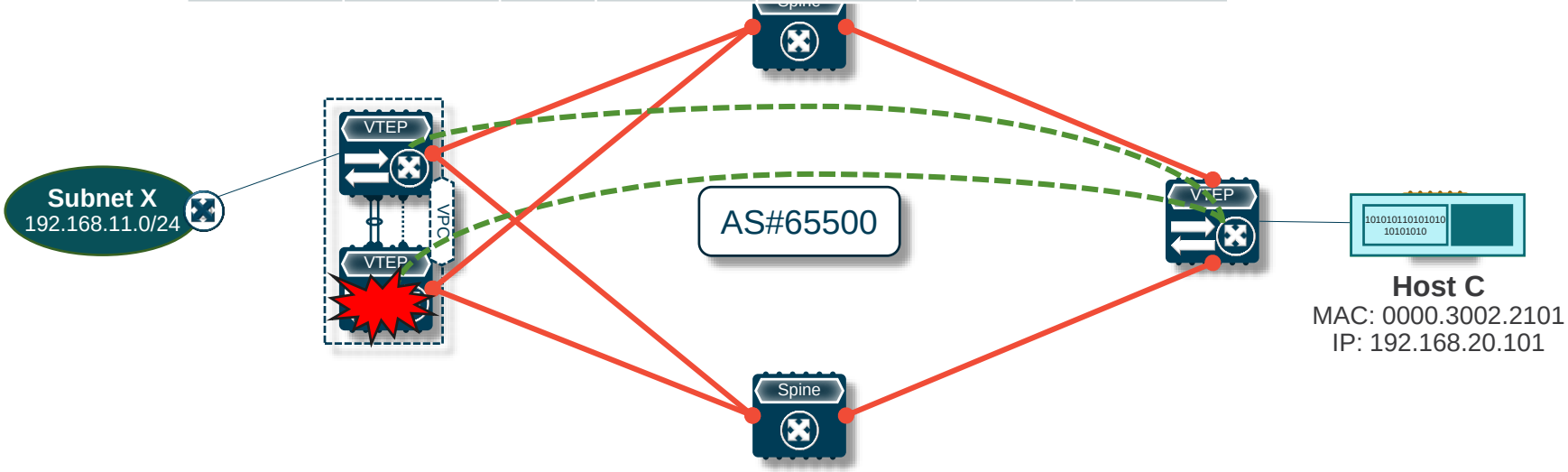
- Subnet Route Advertisement
  - Route Type 5
  - Next-Hop is Anycast VTEP
- Ensure Sync of Subnet
  - Dual-Connect Networks (Point-2-Point not Layer-3 over vPC)
  - Synchronize Routing Table
  - Can cause inefficient forwarding

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	

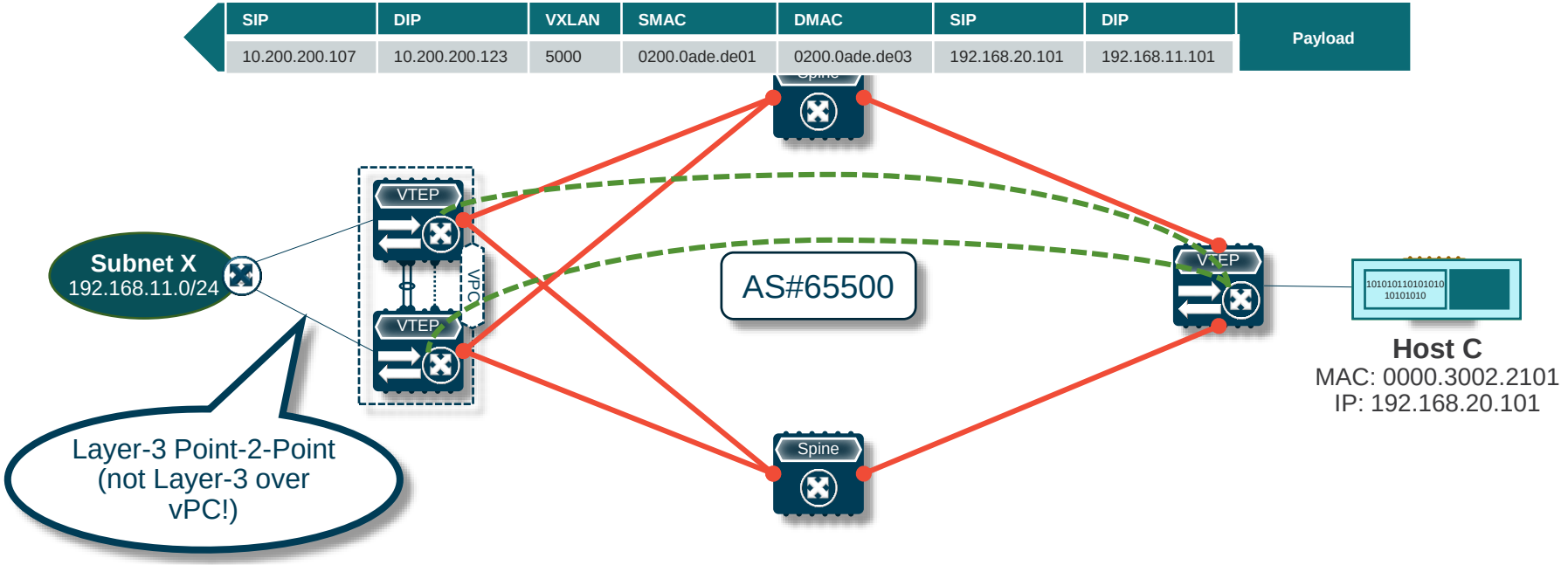


Subnet Route Advertisement with vPC

SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
10.200.200.107	10.200.200.123	5000	0200.0ade.de01	0200.0ade.de03	192.168.20.101	192.168.11.101	

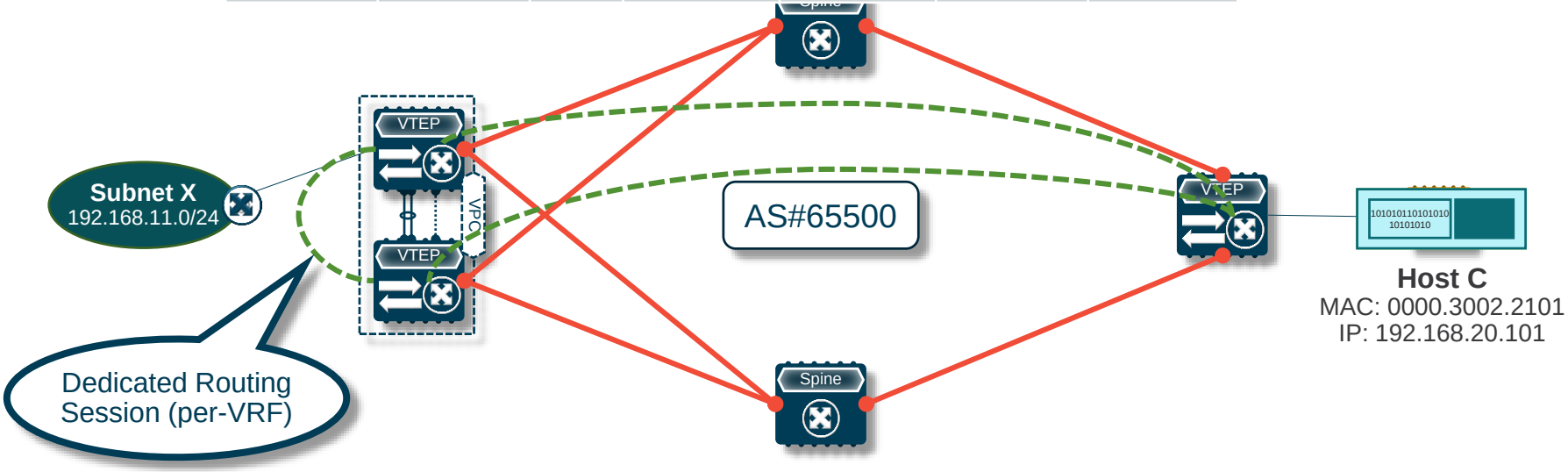


Dual-Attach Networks to vPC Domain



vPC – Synchronizing the Routing

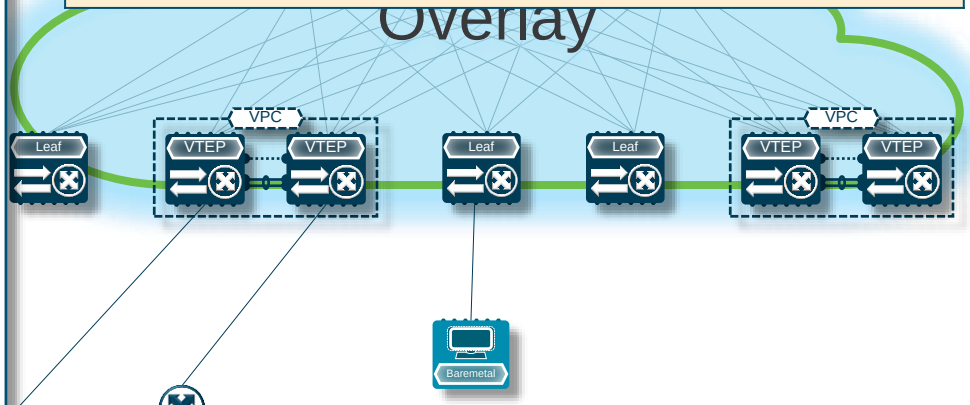
SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
10.200.200.107	10.200.200.123	5000	0200.0ade.de01	0200.0ade.de03	192.168.20.101	192.168.11.101	



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	



Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

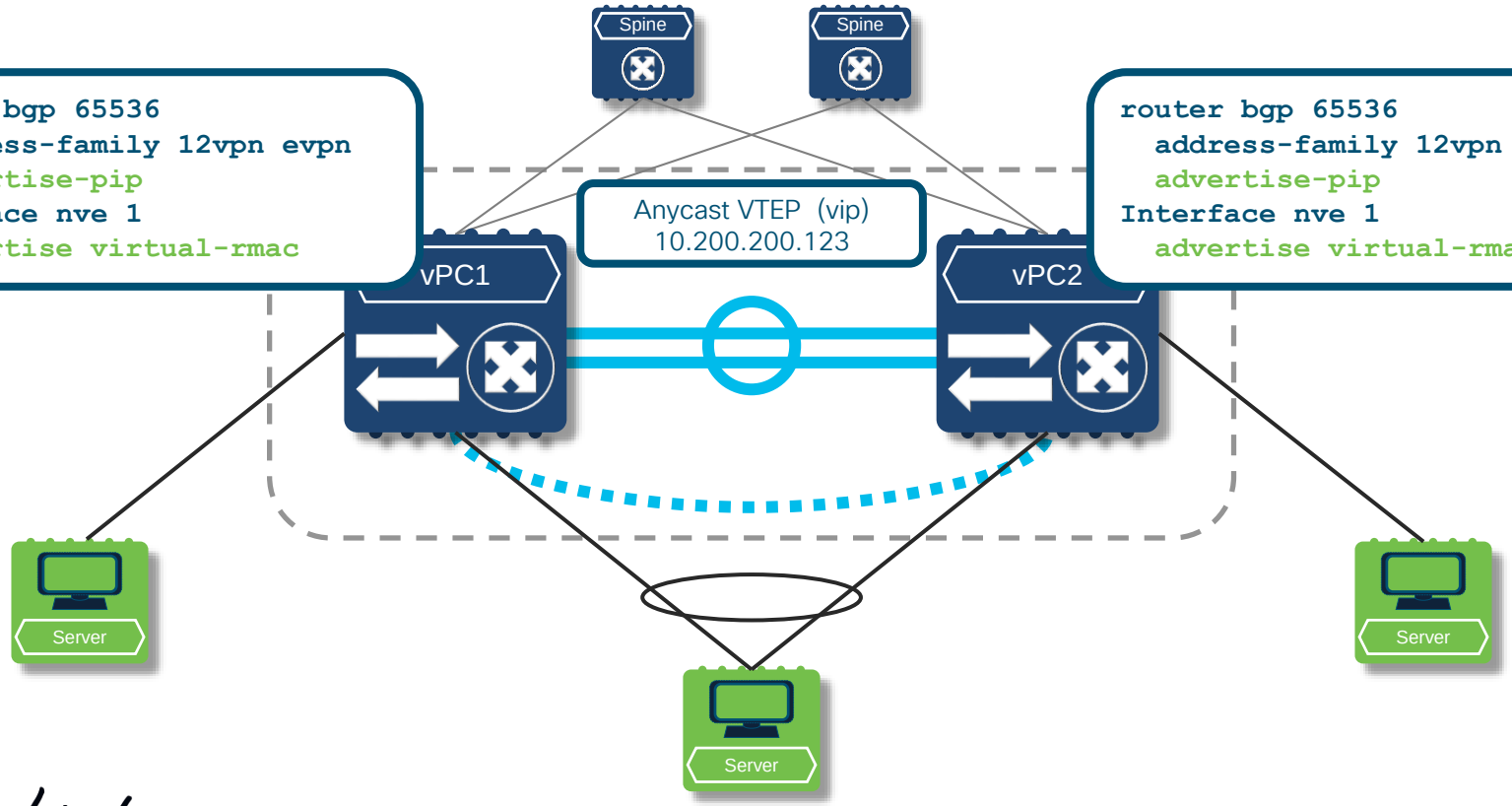
Subnet Y
192.168.22.0/24

Subnet X
192.168.11.0/24

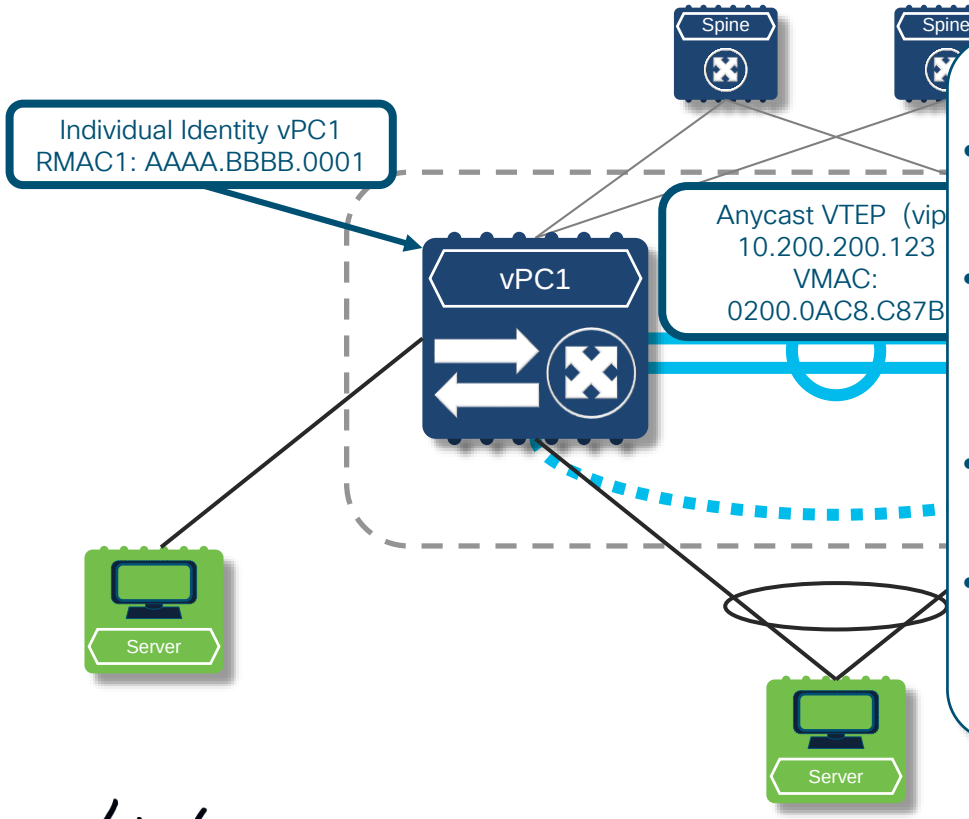
Advertise Primary IP Address

```
router bgp 65536
  address-family 12vpn evpn
  advertise-pip
Interface nve 1
  advertise virtual-rmac
```

```
router bgp 65536
  address-family 12vpn evpn
  advertise-pip
Interface nve 1
  advertise virtual-rmac
```



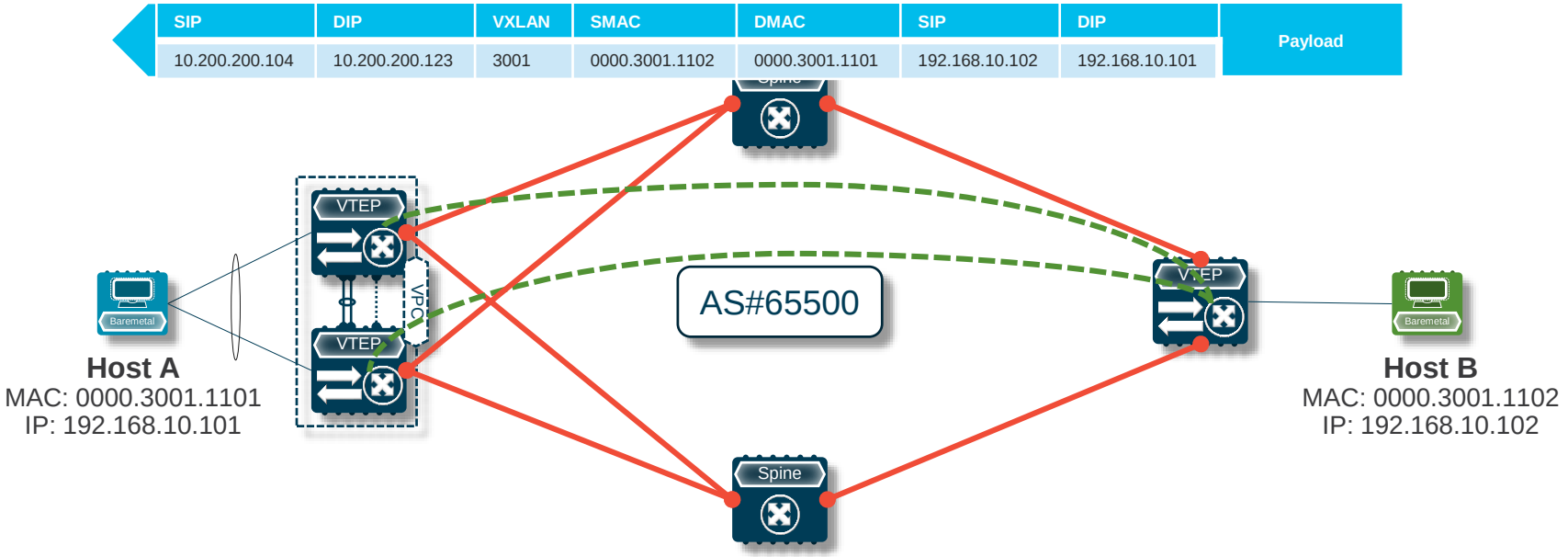
Anycast Primary IP – MAC addressing



- Virtual MAC (VMAC) common MAC address for VPC domain
- VMAC is derived from VIP
  - 02.00. + 4 Bytes of VIP converted in HEX
- vPC and Orphan Type 2 routes will be advertised with (VIP,VMAC)
- vPC and Orphan Type 5 routes will be advertised with (PIP,RMAC)

Bridging to a vPC Domain – VXLAN

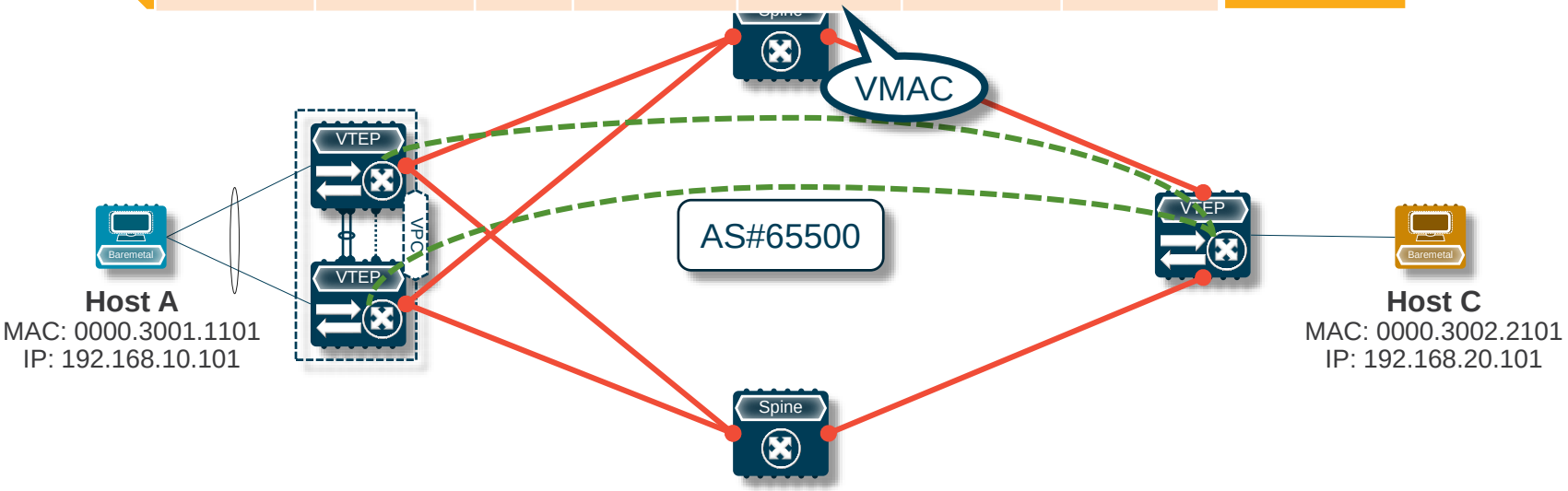
Advertise-PIP



Routing to a vPC Domain – VXLAN

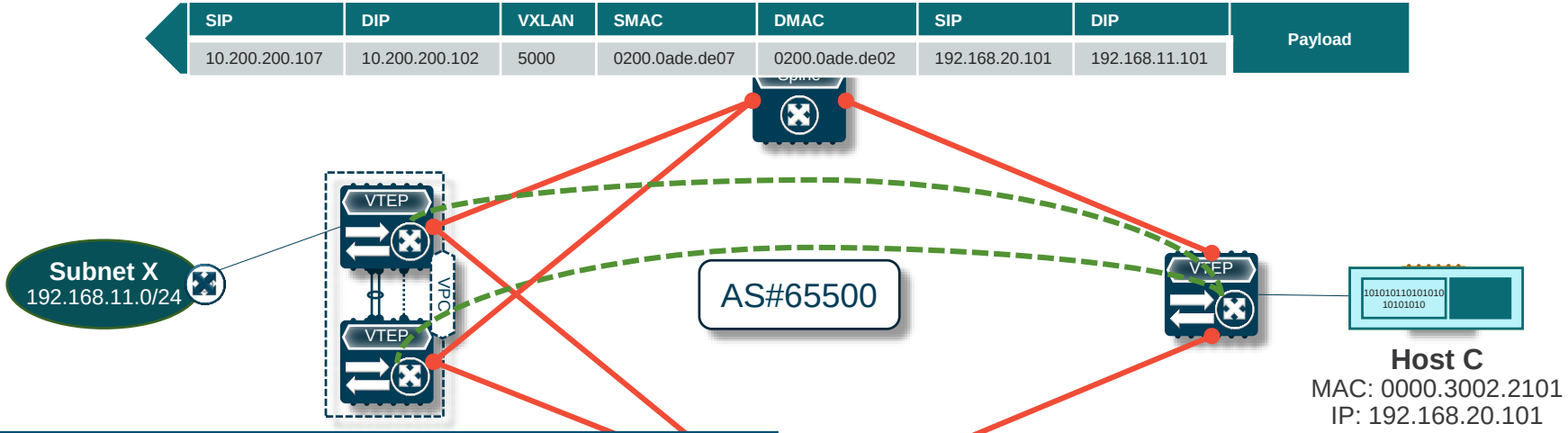
Advertise-PIP

SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
10.200.200.107	10.200.200.123	5000	0200.0ade.de01	0200.0ac8.c87b	192.168.20.101	192.168.10.101	



vPC – Advertise Subnet Individually

Advertise-PIP

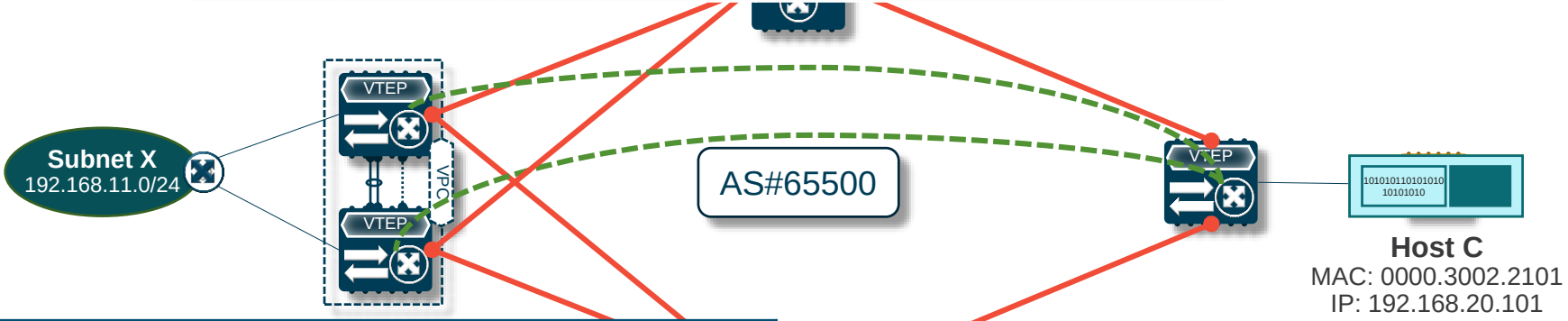


Type	IP / Length	L3VNI / RT	Next-Hop	Seq.
5	192.168.11.0 /24	5000, 65500:5000	10.200.200.102	

VPC – Advertise Subnet Individually

Advertise-PIP

SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
10.200.200.107	10.200.200.102	5000	0200.0ade.de07	0200.0ade.de02	192.168.20.101	192.168.11.101	
10.200.200.107	10.200.200.103	5000	0200.0ade.de07	0200.0ade.de03	192.168.20.101	192.168.11.101	



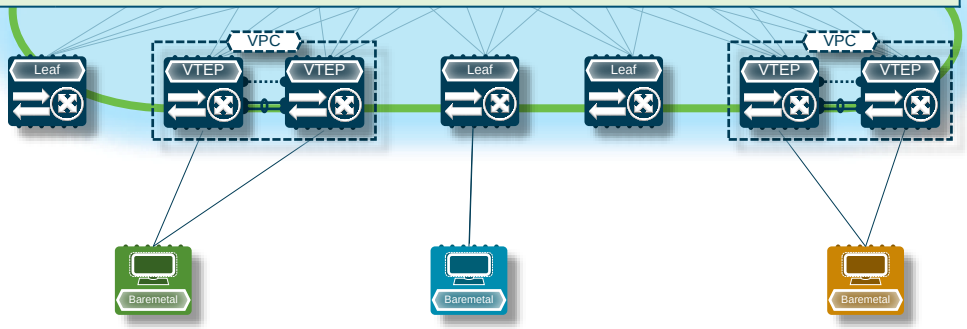
Type	IP / Length	L3VNI / RT	Next-Hop	Seq.
5	192.168.11.0 /24	5000, 65500:5000	10.200.200.102	
5	192.168.11.0/24	5000, 65500:5000	10.200.200.103	

BUM traffic in VPC

Ingress Replication

- Dynamic List of VNI to VTPE paring
  - Uses EVPN Type 3 routes
- Traffic is replicated at ingress VTEP
- Can be inefficient for large scale deployments

Type	IP / Length	L2VNI / RT	Next-Hop	Seq.
3	192.168.10.101 /32	3001, 65500:3001	10.200.200.123	
3	192.168.10.102 /32	3001, 65500:3001	10.200.200.104	
3	192.168.20.101 /32	3002, 65500:3002	10.200.200.124	



Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

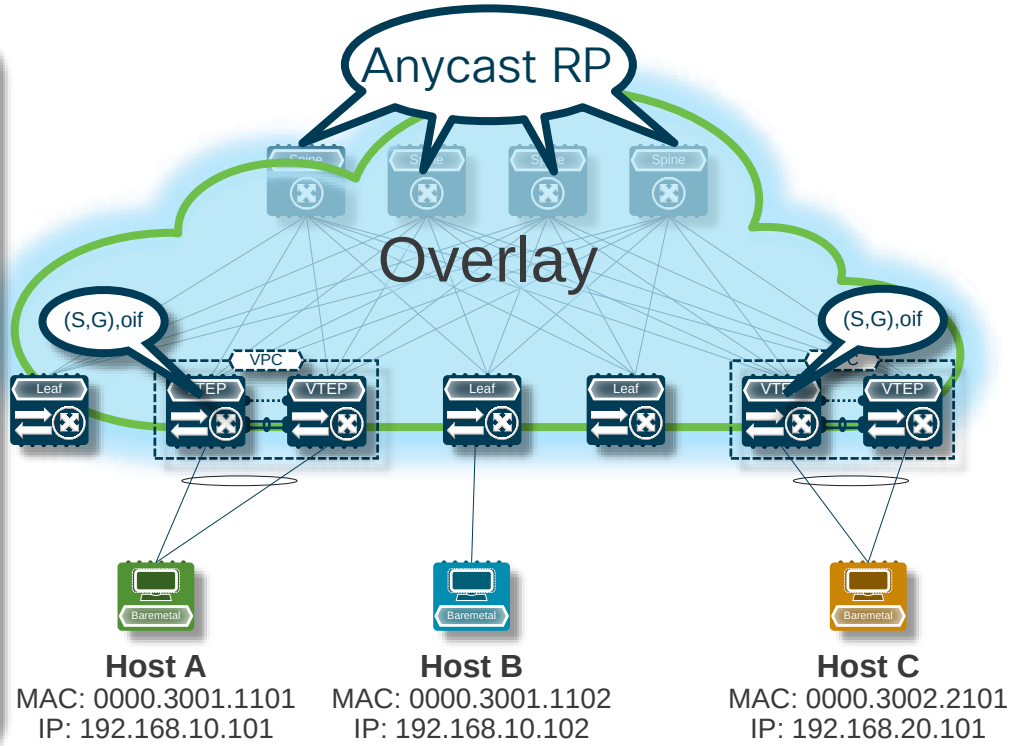
Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

BUM traffic in VPC

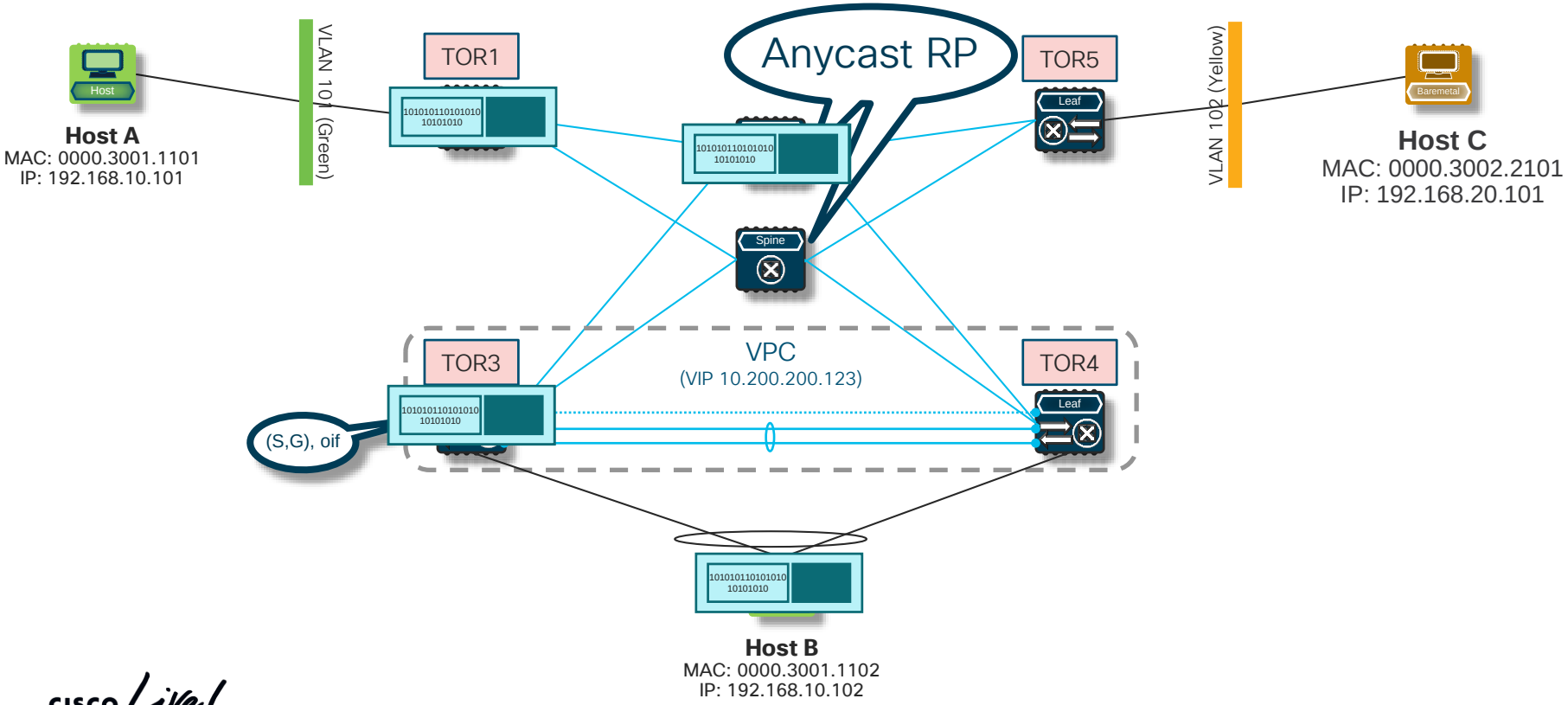
Multicast Mode

- Decapsulation Node, decapsulates all BUM traffic
  - Peer with lower cost to RP will be elected as decapsulation node
  - Same cost to RP, vPC Primary will be elected
- Decapsulation node have (S,G) and OIF entries for VNI associated multicast group
  - Source is Anycast VTP address
  - Multicast Group associated to the VNI



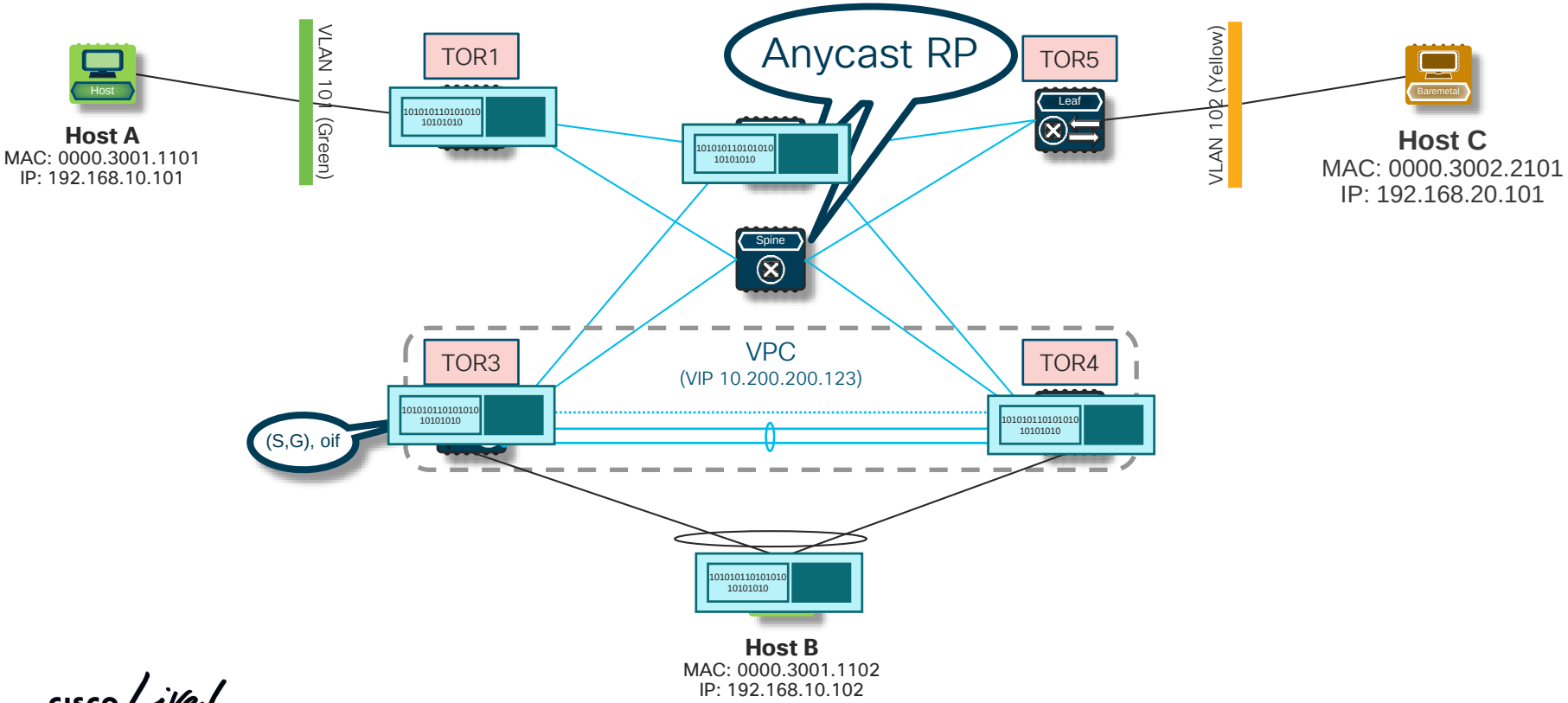
BUM traffic – Sender in VPC

Multicast Mode



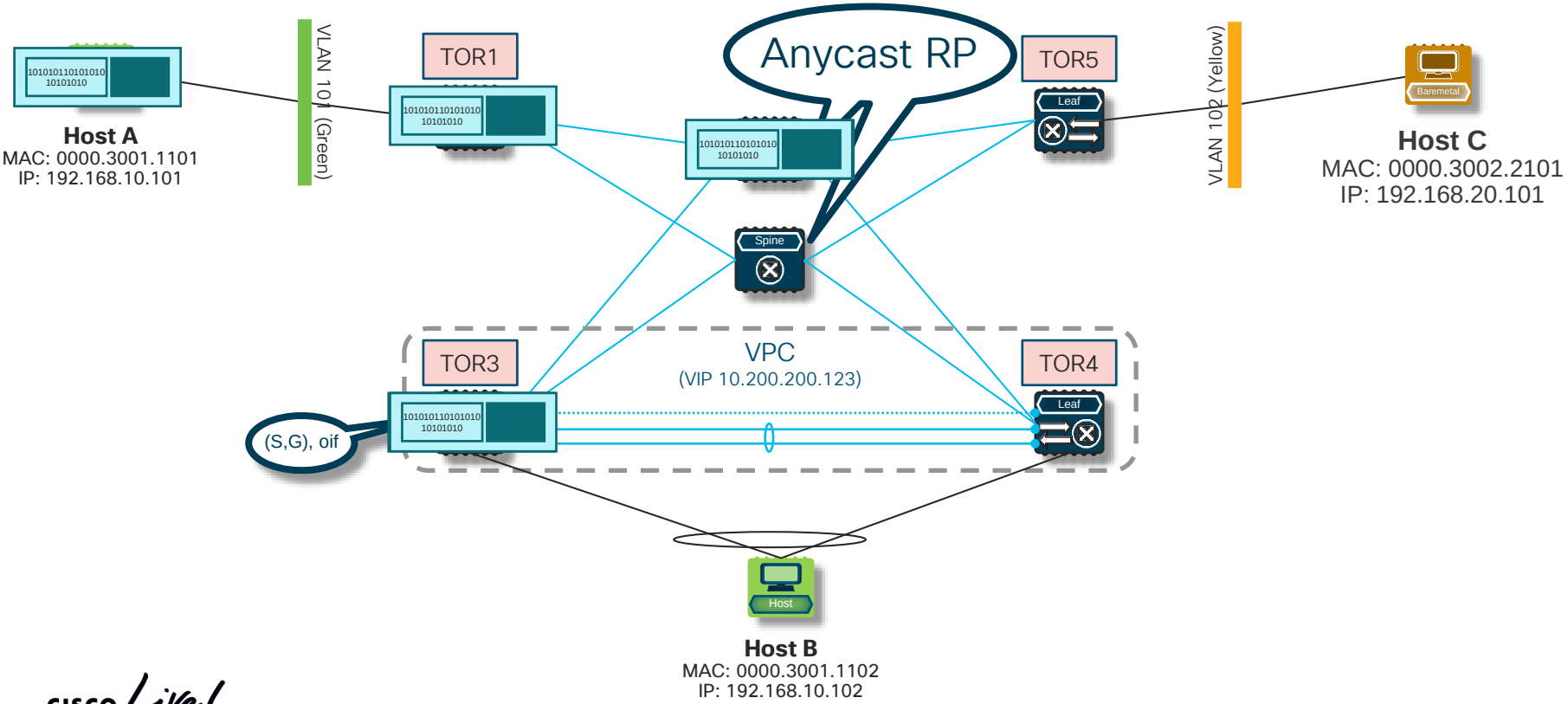
BUM traffic – Sender in VPC

Multicast Mode



BUM traffic – Receiver in VPC

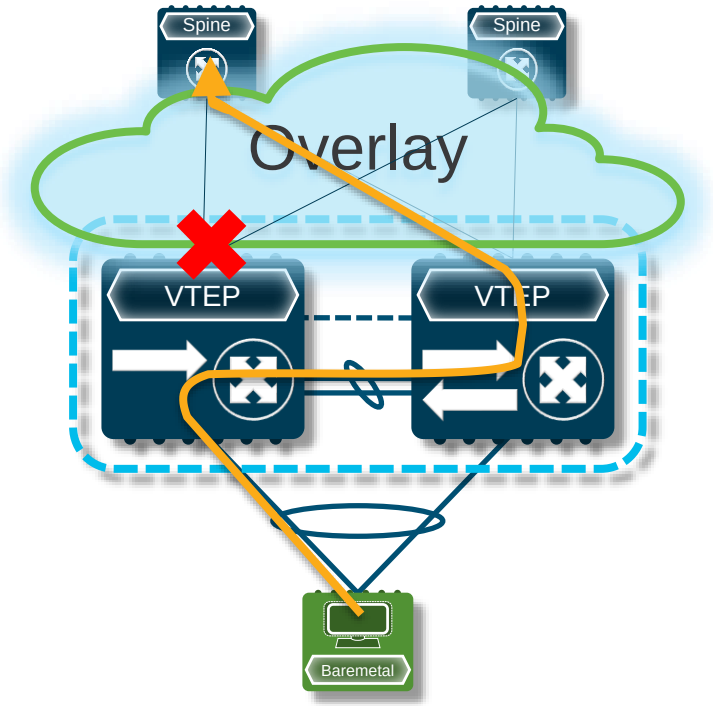
Multicast Mode



Additional Features

vPC Infrastructure VLANs

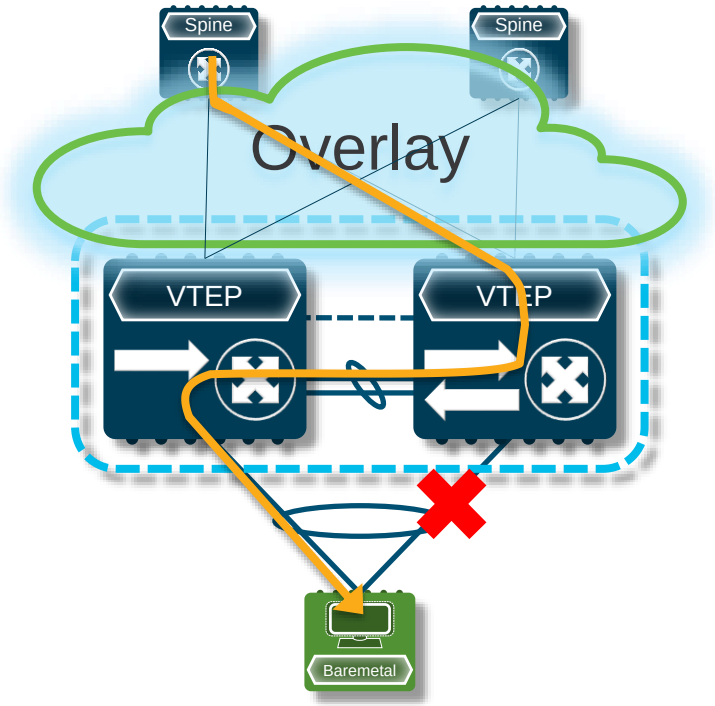
- Infrastructure VLANs are used for Backup Routing Path
- Infrastructure VLAN is present on Peer Link
- Used in case of failure of uplinks on a vPC peer
- Required for BUM traffic transfer



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

vPC Infrastructure VLANs

- Infrastructure VLANs are used for Backup Routing Path
- Infrastructure VLAN is present on Peer Link
- Used in case of failure of uplinks on a vPC peer
- Required for BUM traffic transfer

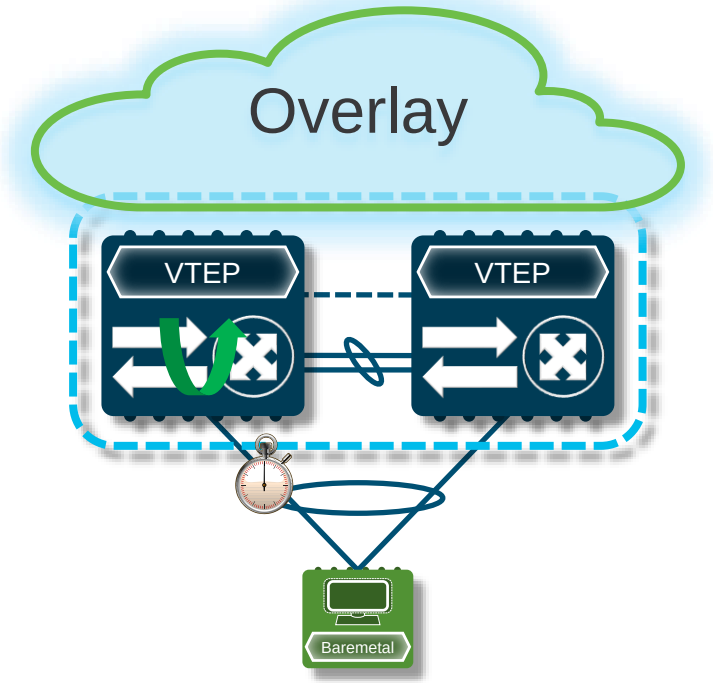


Nexus (config)# system nve infra-vlans <1-3967>

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

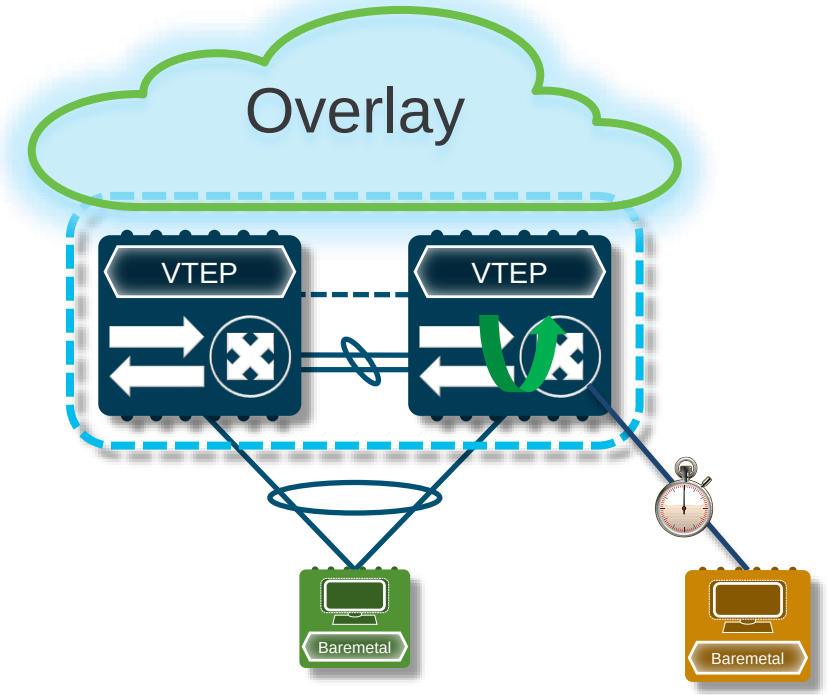


Nexus(config-vpc-domain)# delay restore <1-3600 sec>

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

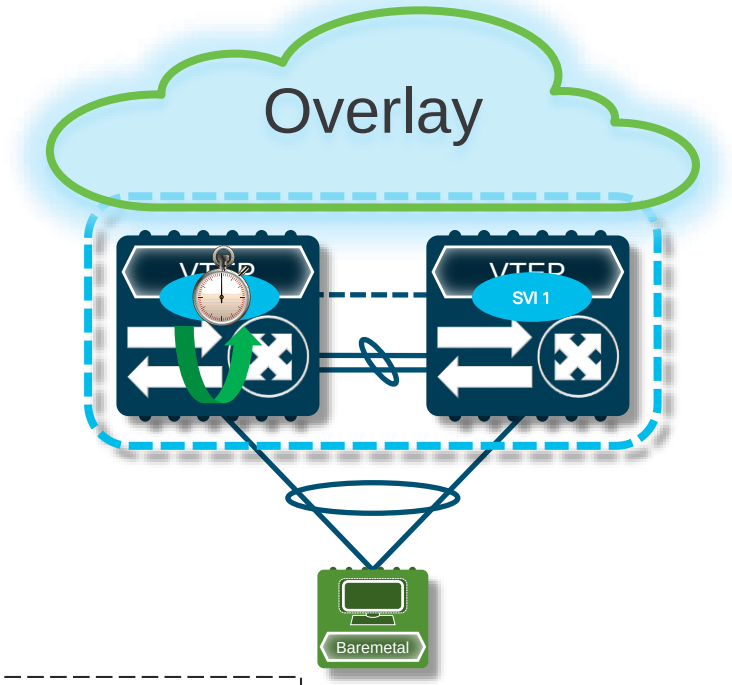


Nexus(config-vpc-domain)# delay restore orphan-port <0-300 sec>

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

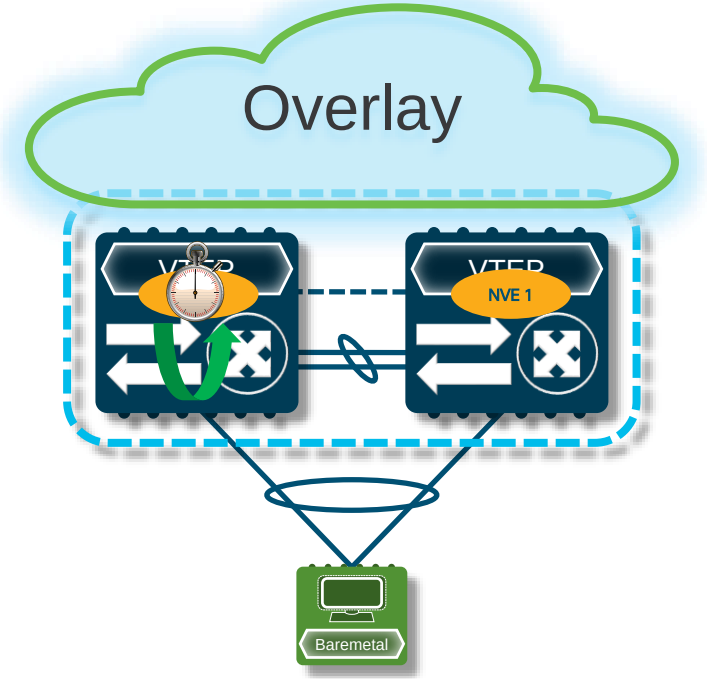


Nexus(config-vpc-domain)# delay restore interface-vlan <1-3600 sec>

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>

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

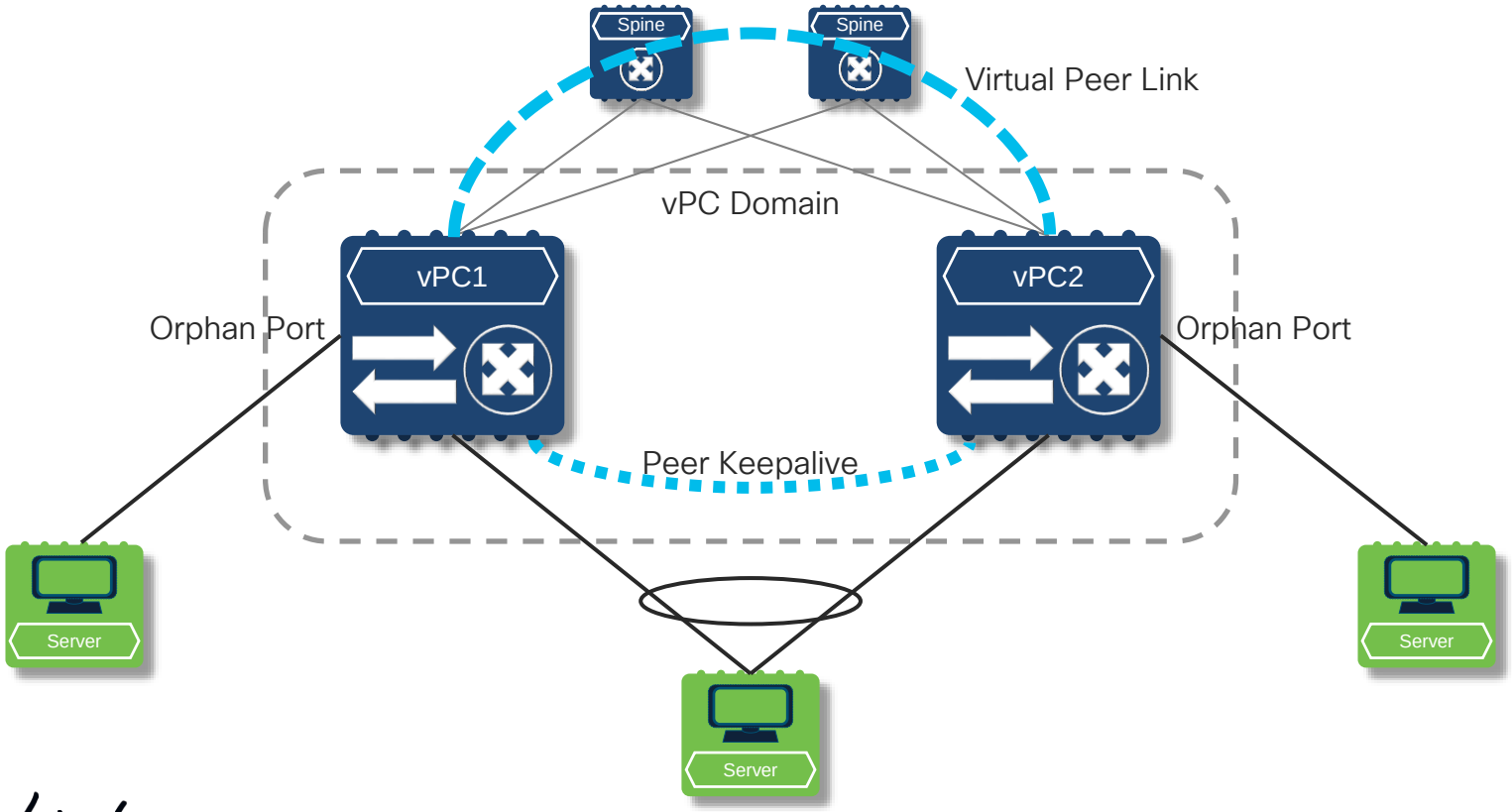
When vPC VTEP consistency check failed:

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

vPC Fabric Peering

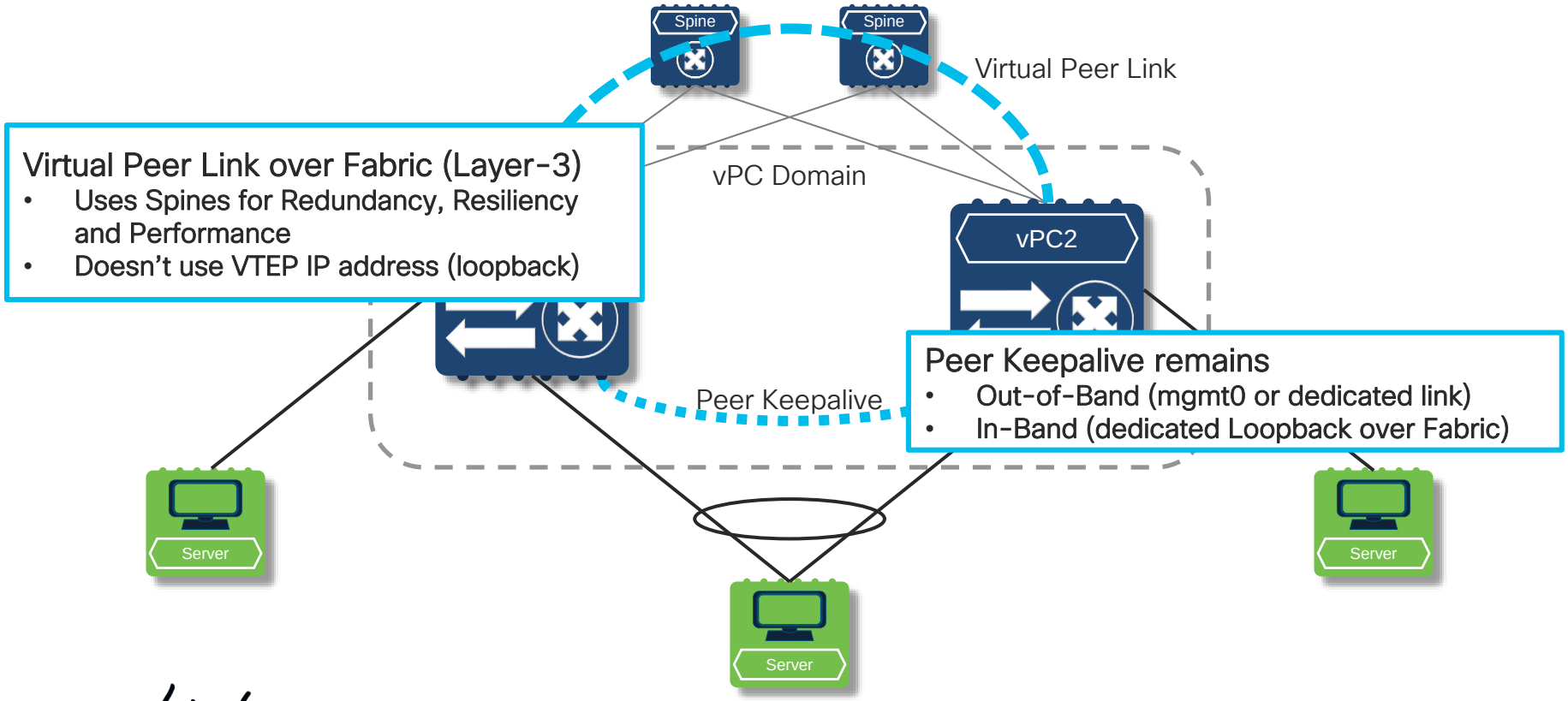
vPC with Fabric Peering

for VXLAN BGP EVPN



vPC with Fabric Peering

for VXLAN BGP EVPN



Virtual Peer Link over Fabric (Layer-3)

- Uses Spines for Redundancy, Resiliency and Performance
- Doesn't use VTEP IP address (loopback)

Peer Keepalive remains

- Out-of-Band (mgmt0 or dedicated link)
- In-Band (dedicated Loopback over Fabric)

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
```

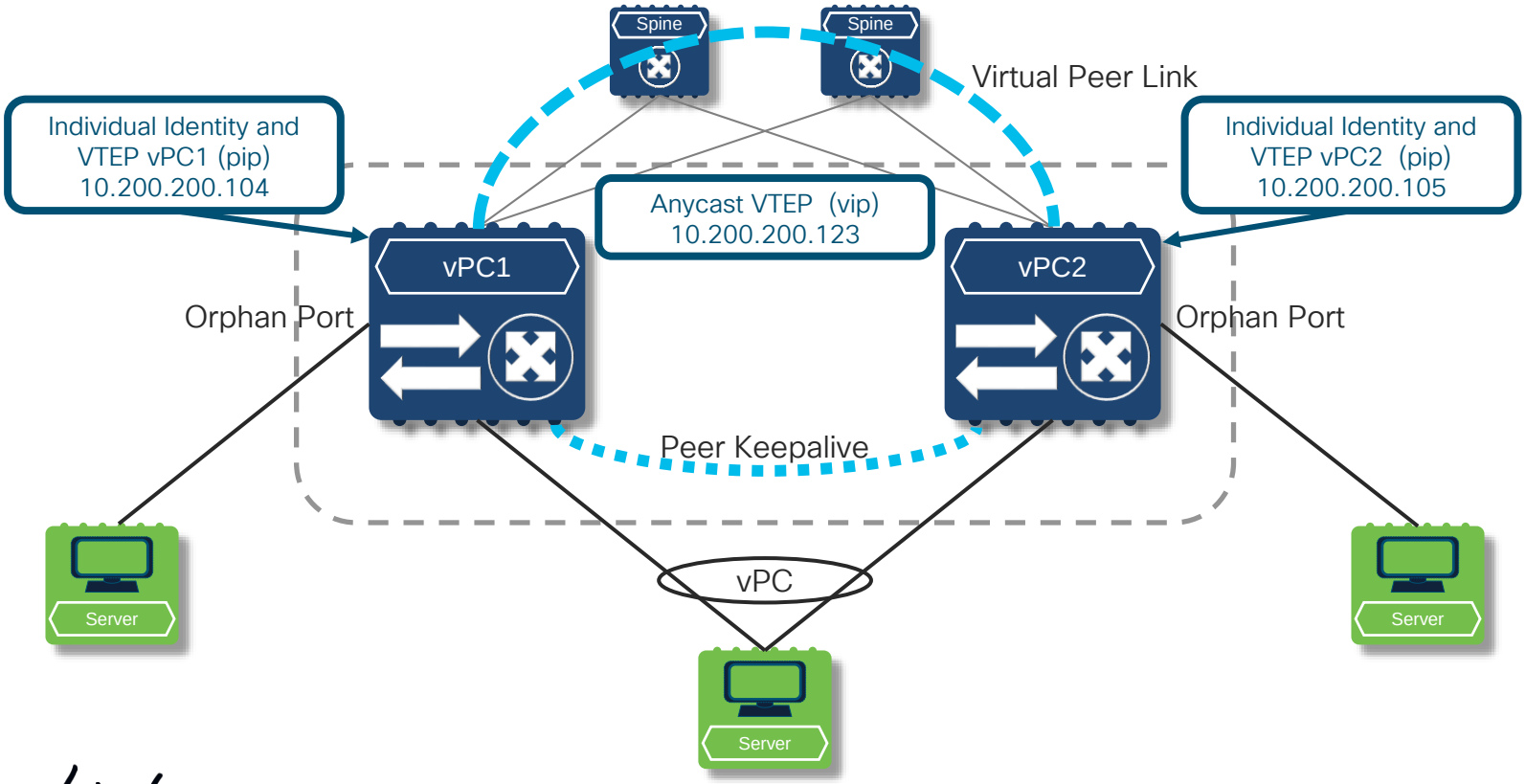
Define Port-Type Fabric

All Interface to Spine

Ensure appropriate MTU

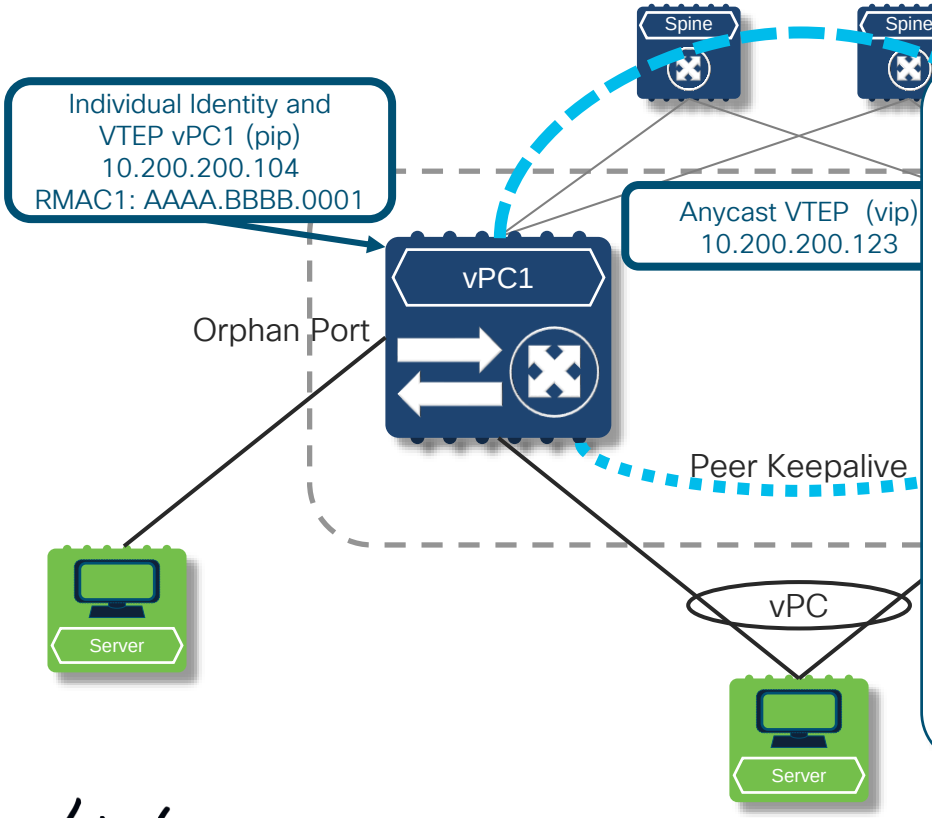
Anycast VTEP and Individual VTEPs

vPC with Fabric Peering



Anycast VTEP and Individual VTEPs

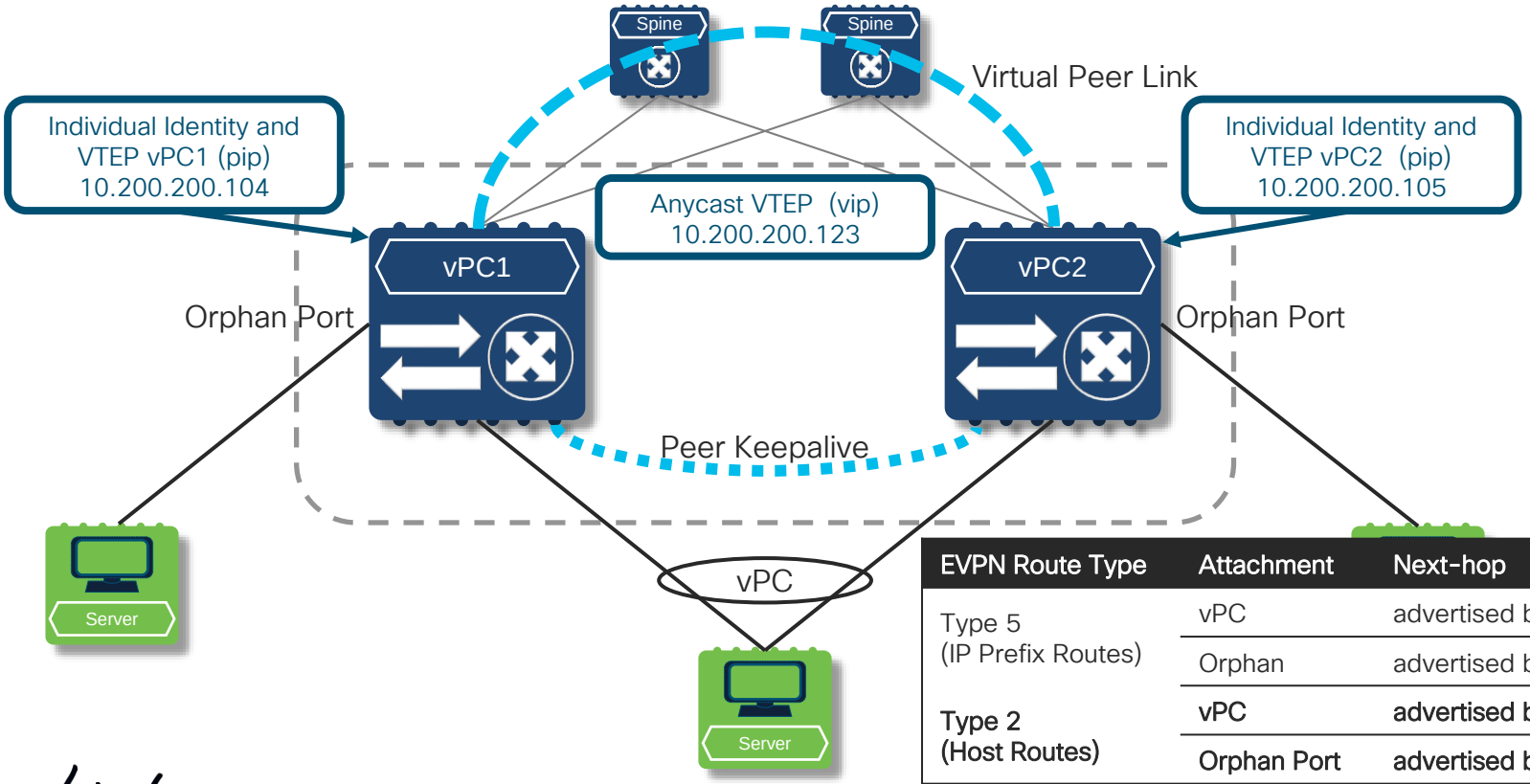
vPC with Fabric Peering



- Advertise PIP and Advertise VMAC are mandatory in vPC Fabric peering
- vPC Type 2 routes will be advertise with (VIP,VMAC)
- Orphan Type 2 routes will be advertised with (PIP,RMAC)
- vPC and Orphan Type 5 routes will be advertised with (PIP,RMAC)

Host Attachment

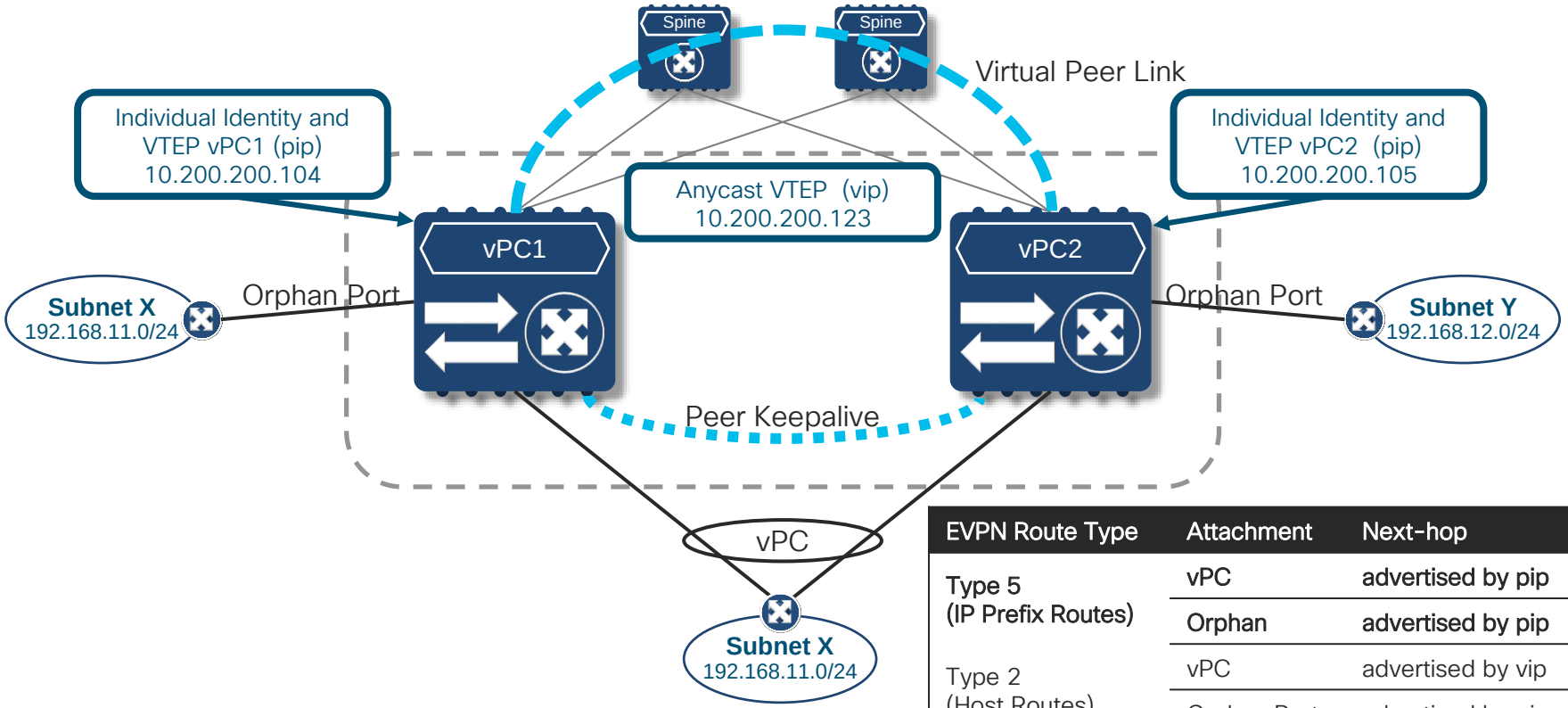
vPC with Fabric Peering



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

Network Attachment

vPC with Fabric Peering

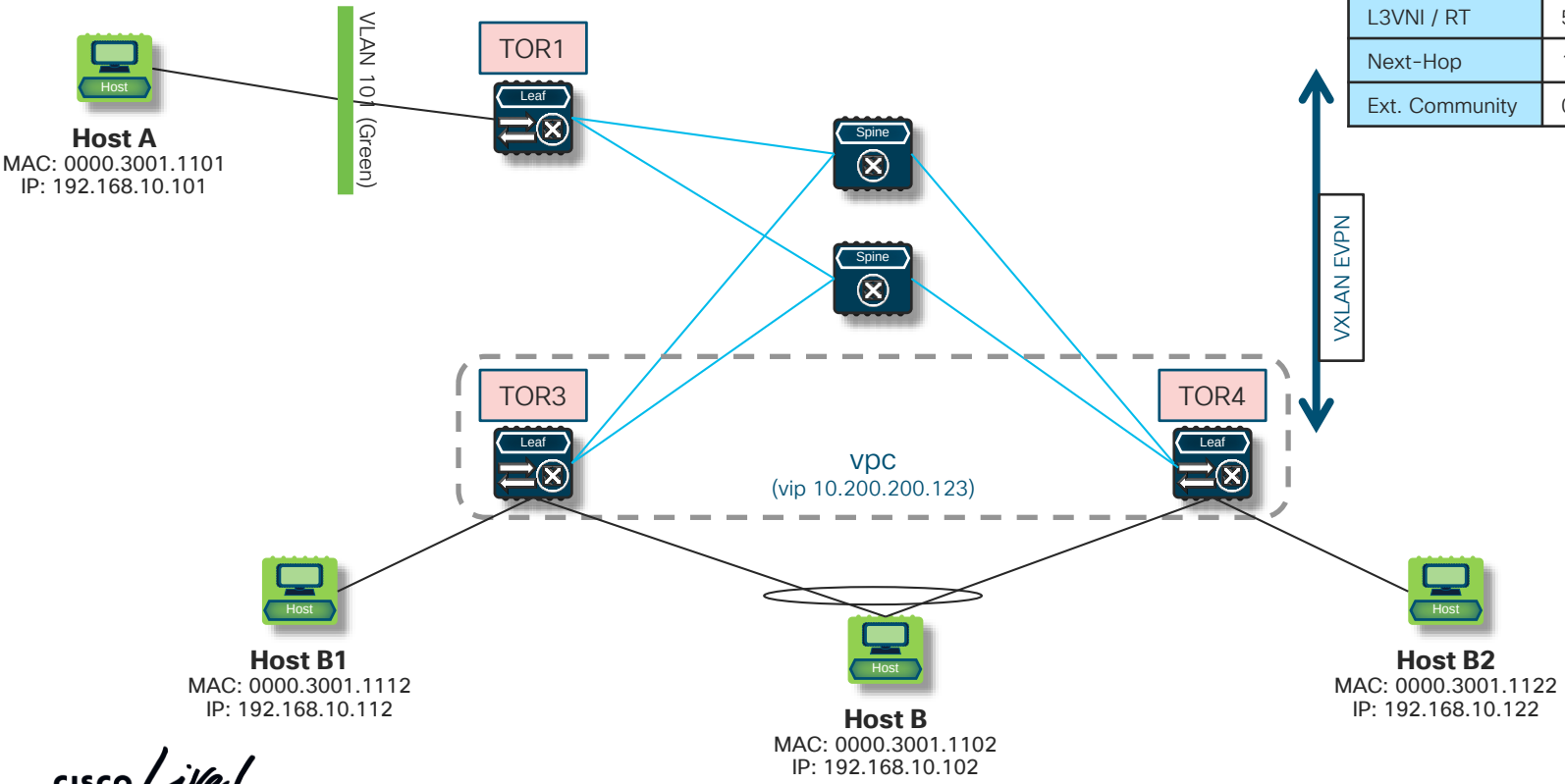


EVPN Route Type	Attachment	Next-hop
Type 5 (IP Prefix Routes)	vPC	advertised by pip
	Orphan	advertised by pip
Type 2 (Host Routes)	vPC	advertised by vip
	Orphan Port	advertised by pip

Packet Walk – vPC with Fabric Peering

vPC and Orphan Host

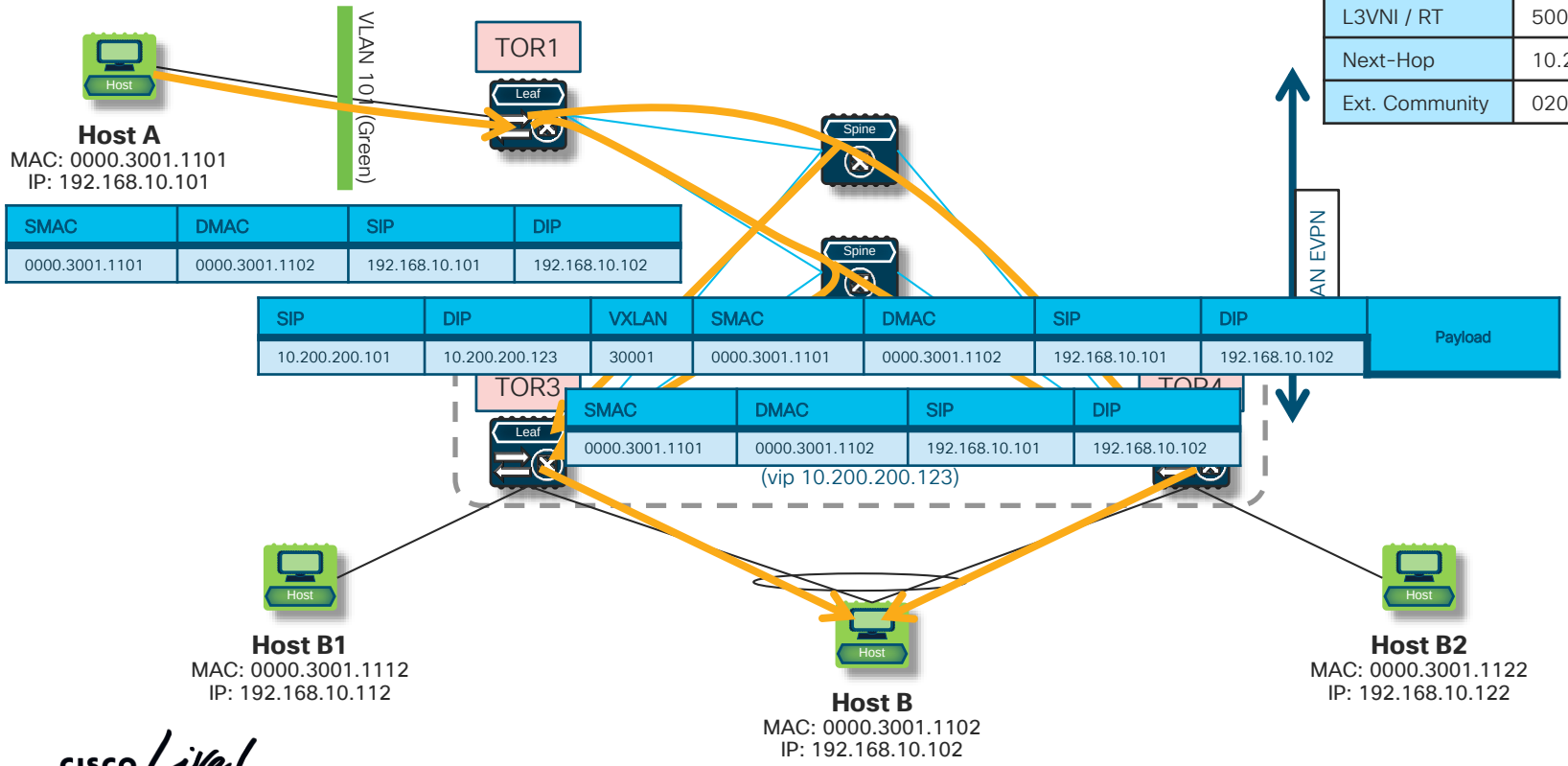
EVPN Control-Plane	
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ac8.c87b



Packet Walk – vPC with Fabric Peering

vPC Host

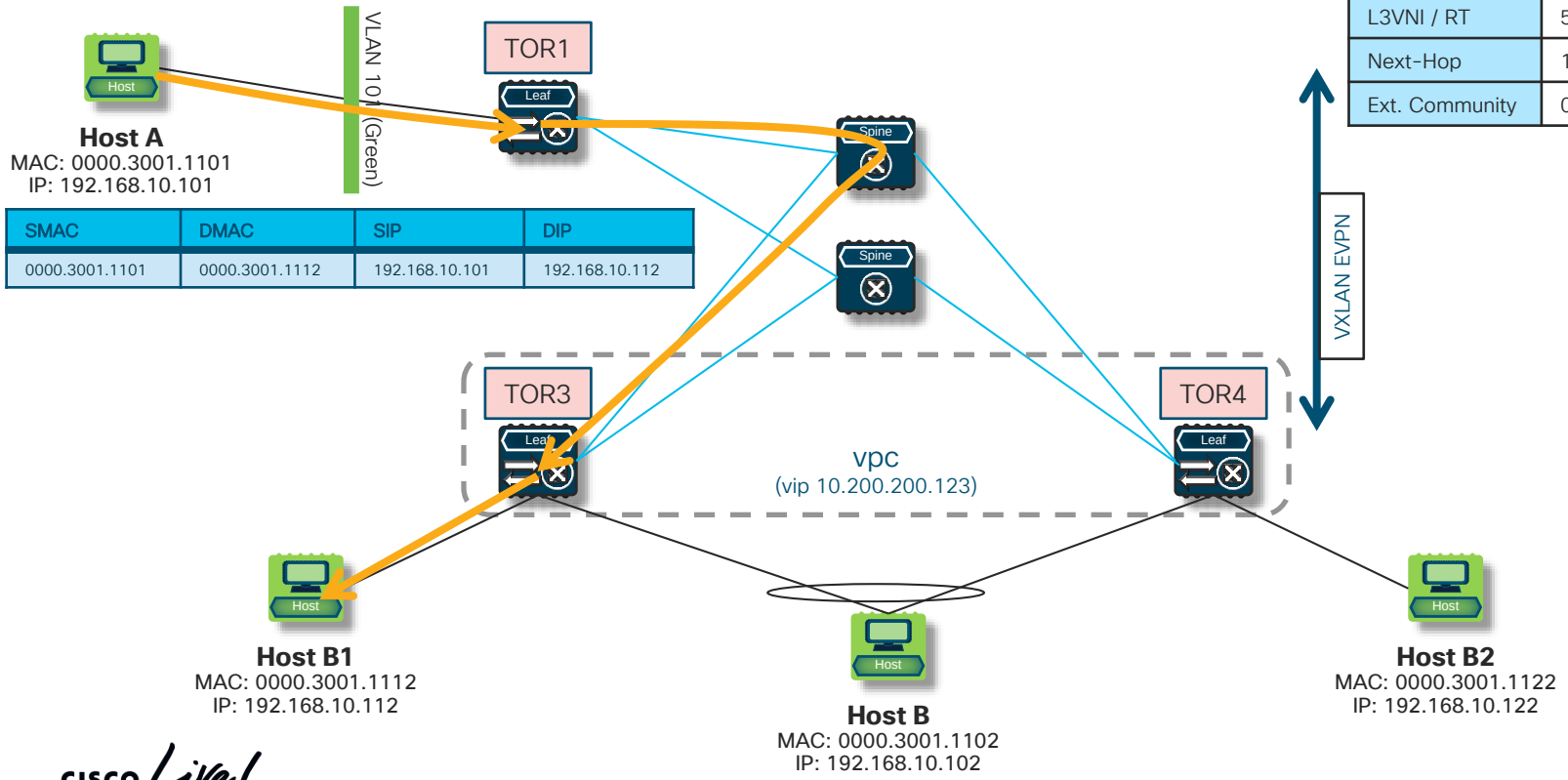
EVPN Control-Plane	
Type	2
MAC / Length	0000.3001.1102 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.102 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.123
Ext. Community	0200.0ac8.c87b



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	10.200.200.103
Ext. Community	0200.0ade.de03

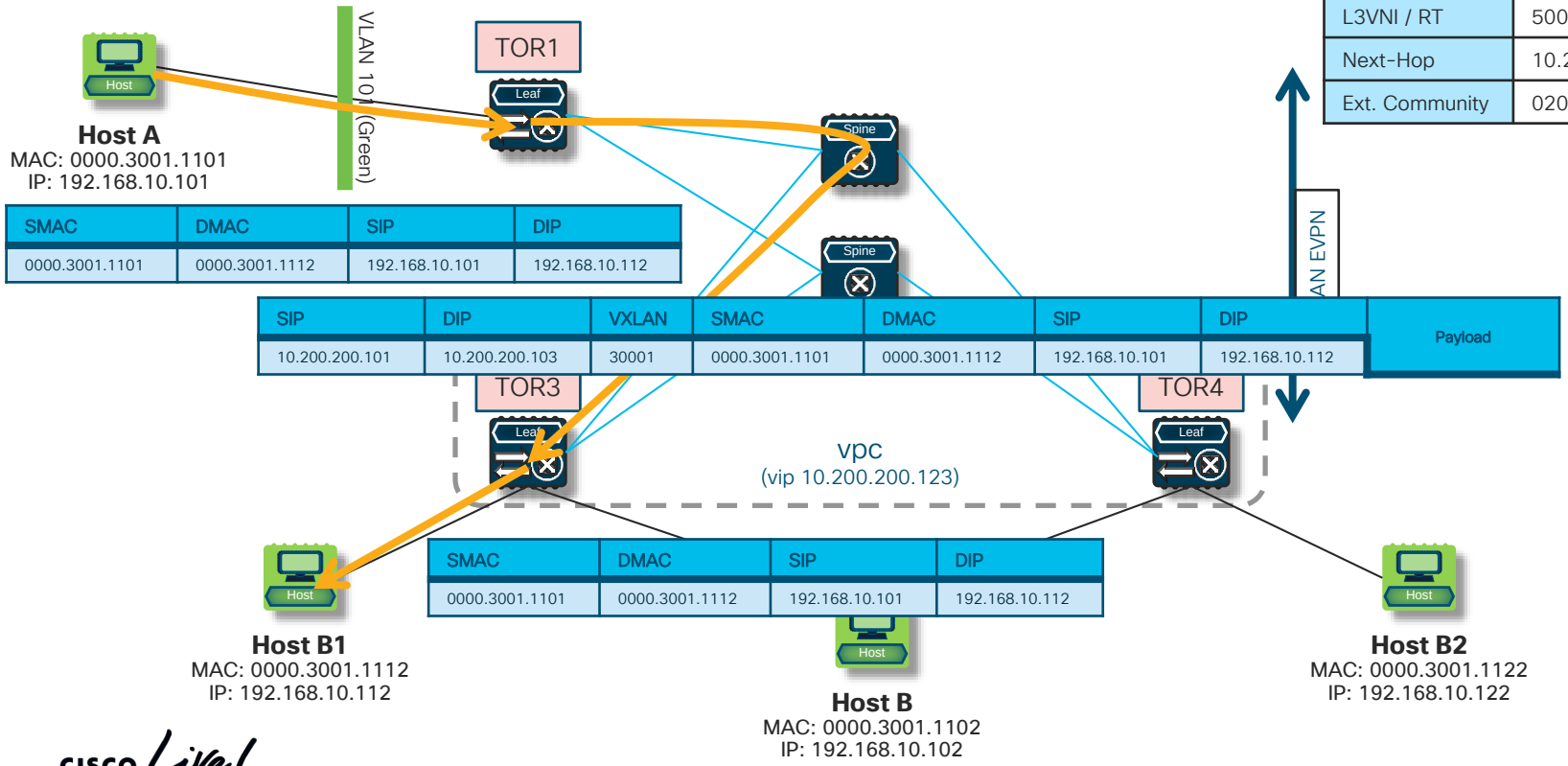


SMAC	DMAC	SIP	DIP
0000.3001.1101	0000.3001.1112	192.168.10.101	192.168.10.112

Packet Walk – vPC with Fabric Peering

Orphan Host

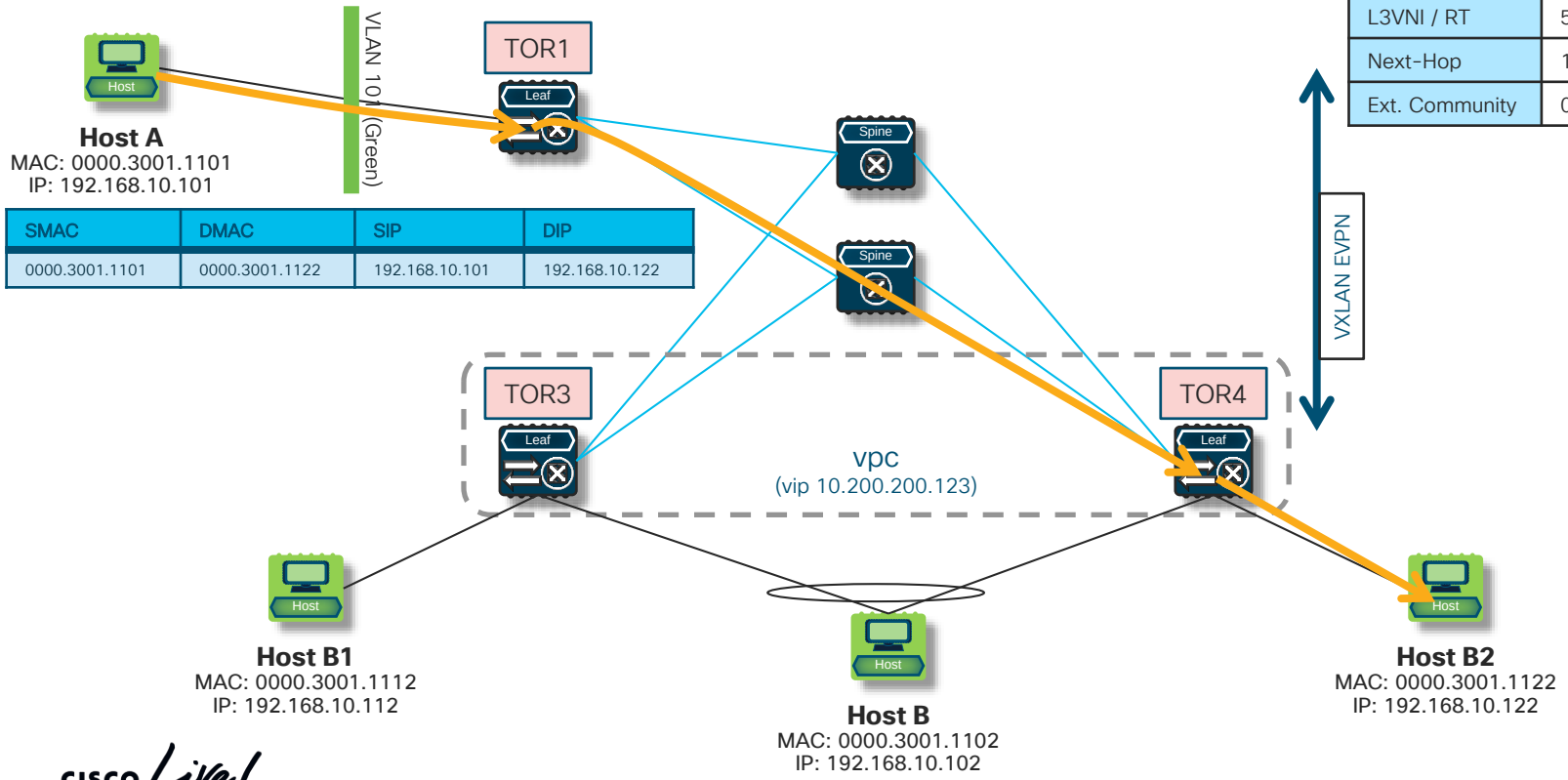
EVPN Control-Plane	
Type	2
MAC / Length	0000.3001.1112 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.112 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.103
Ext. Community	0200.0ade.de03



Packet Walk – vPC with Fabric Peering

Orphan Host

EVPN Control-Plane	
Type	2
MAC / Length	0000.3001.1122 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.122 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.104
Ext. Community	0200.0ade.de04

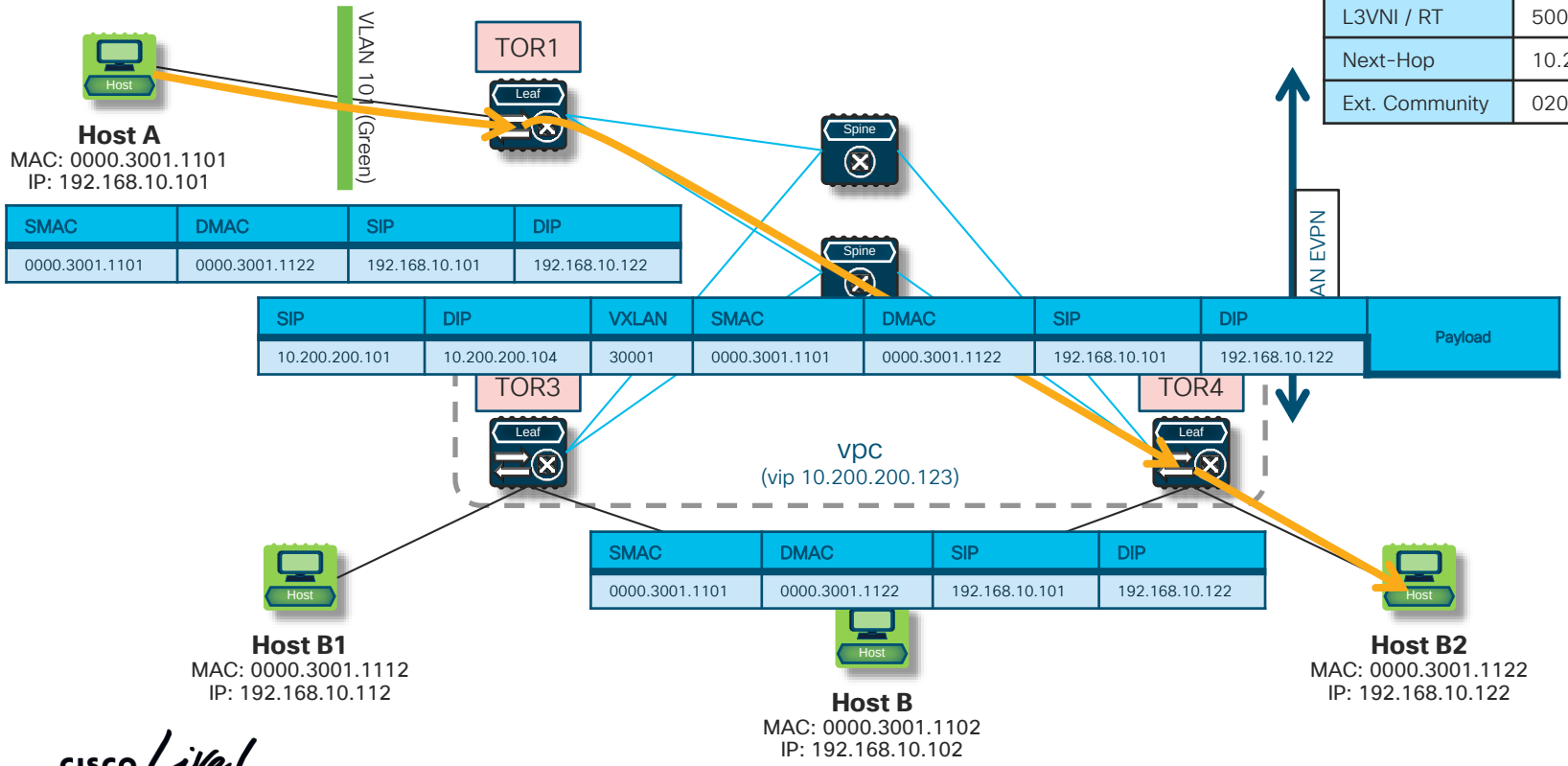


SMAC	DMAC	SIP	DIP
0000.3001.1101	0000.3001.1122	192.168.10.101	192.168.10.122

Packet Walk – vPC with Fabric Peering

Orphan Host

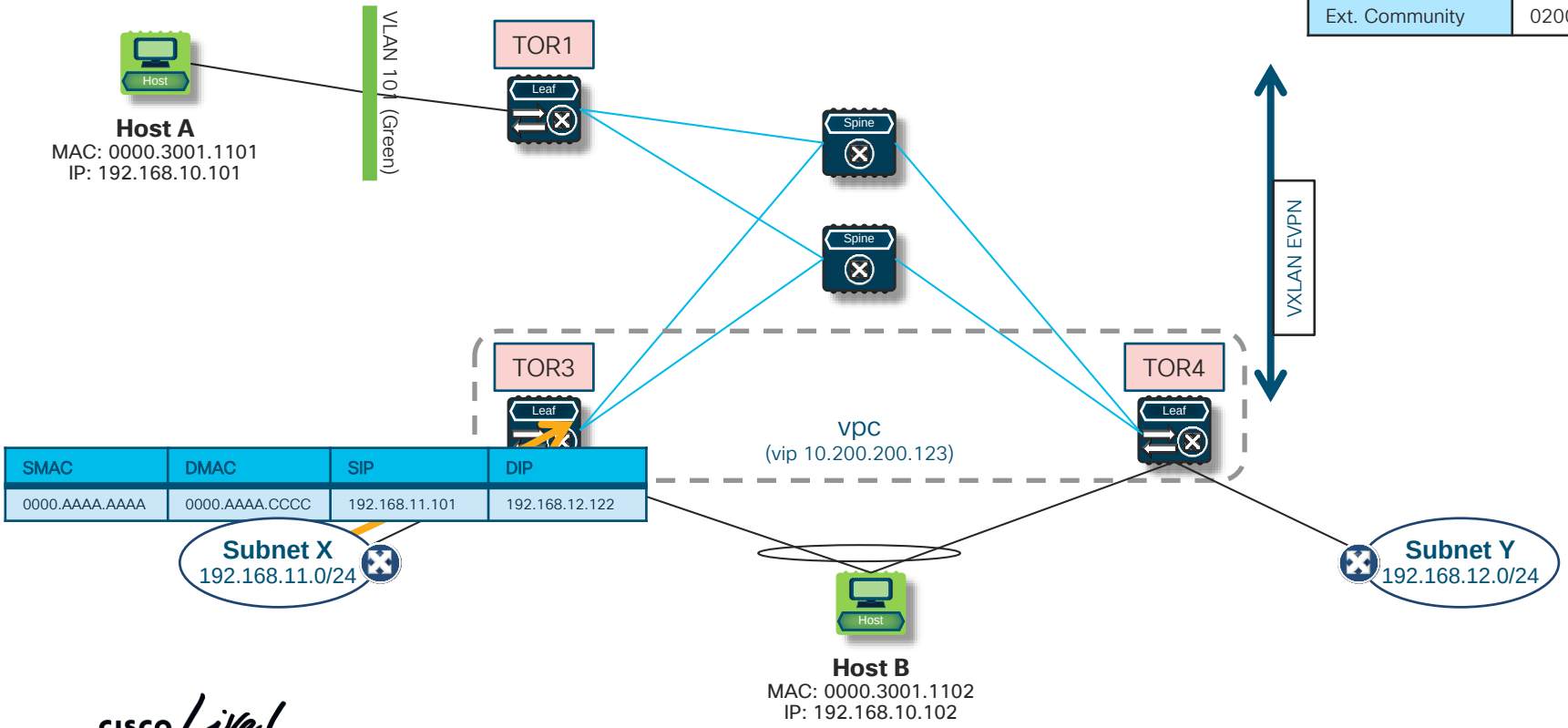
EVPN Control-Plane	
Type	2
MAC / Length	0000.3001.1122 / 48
L2VNI / RT	3001 / 65500:3001
IP / Length	192.168.10.122 / 32
L3VNI / RT	5000 / 65500:5000
Next-Hop	10.200.200.104
Ext. Community	0200.0ade.de04



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	10.200.200.104
Ext. Community	0200.0ade.de04

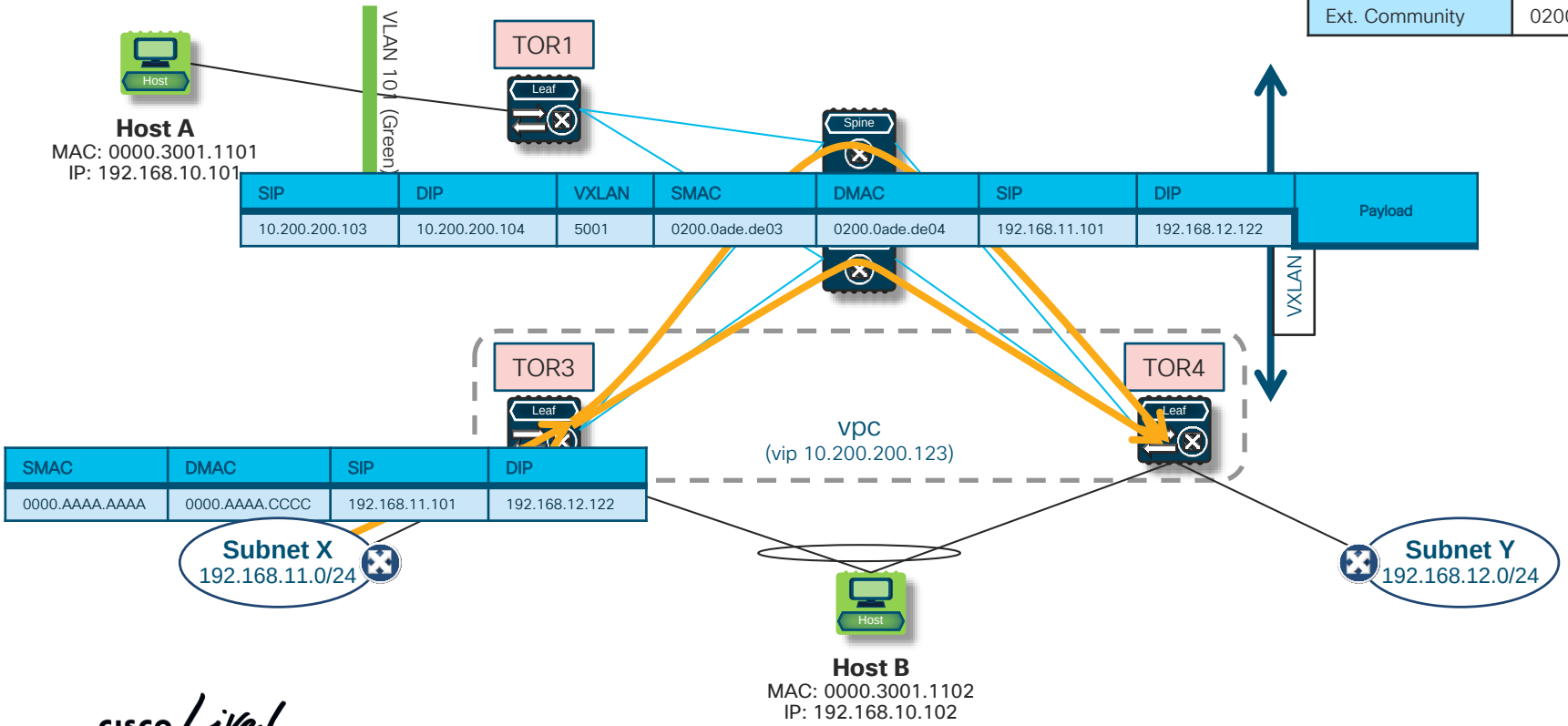


SMAC	DMAC	SIP	DIP
0000.AAAA.AAAA	0000.AAAA.CCCC	192.168.11.101	192.168.12.122

Packet Walk – vPC Fabric Peering

Orphan Networks

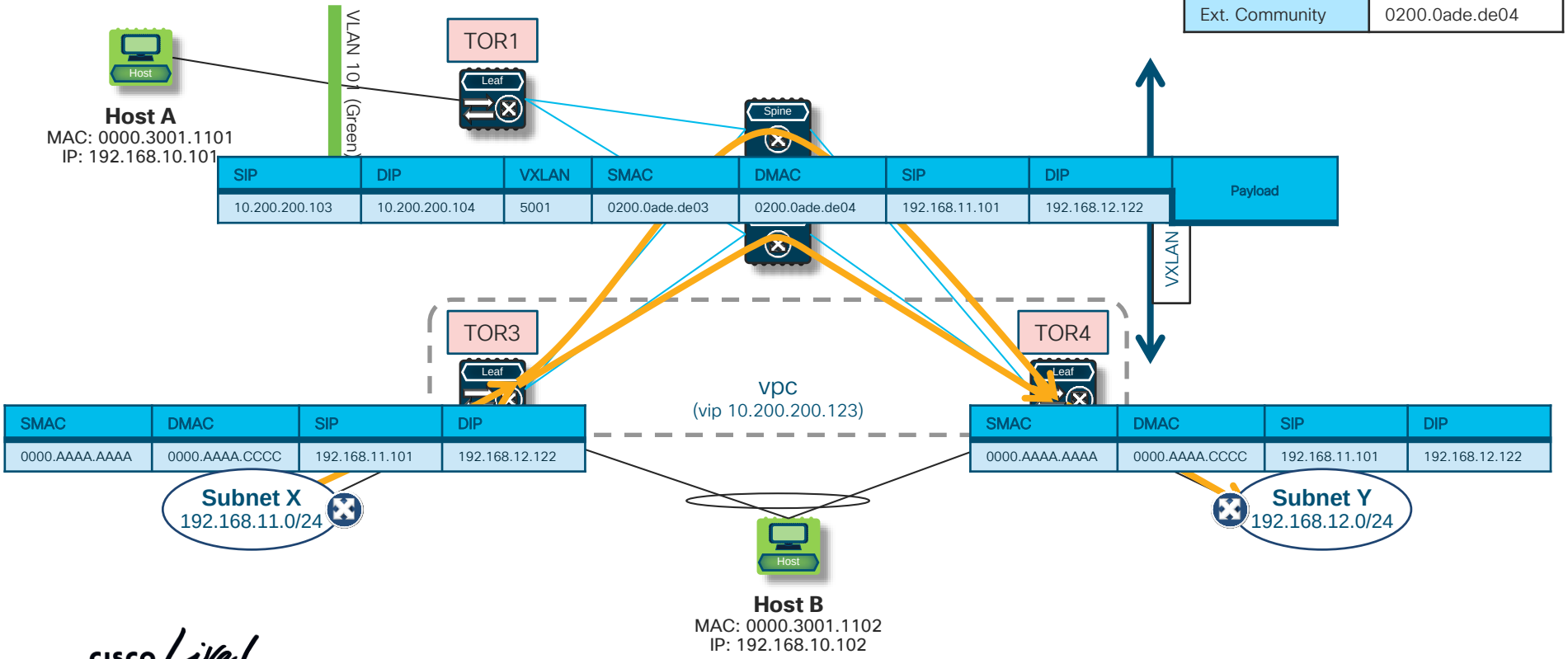
EVPN Control-Plane	
Type	5
IP / Length	192.168.12.0 / 24
L3VNI / RT	5001 / 65500:5000
Next-Hop	10.200.200.104
Ext. Community	0200.0ade.de04



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	10.200.200.104
Ext. Community	0200.0ade.de04

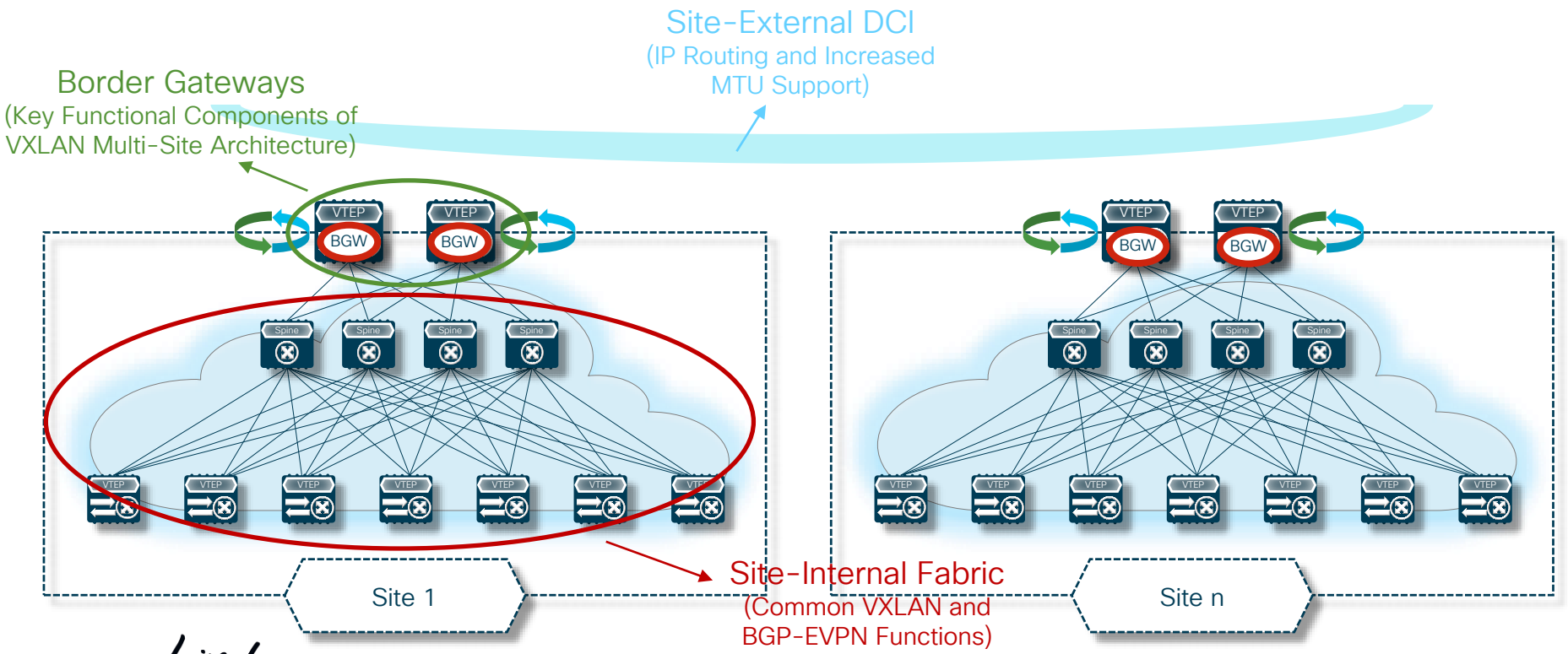


VPC Border Gateway

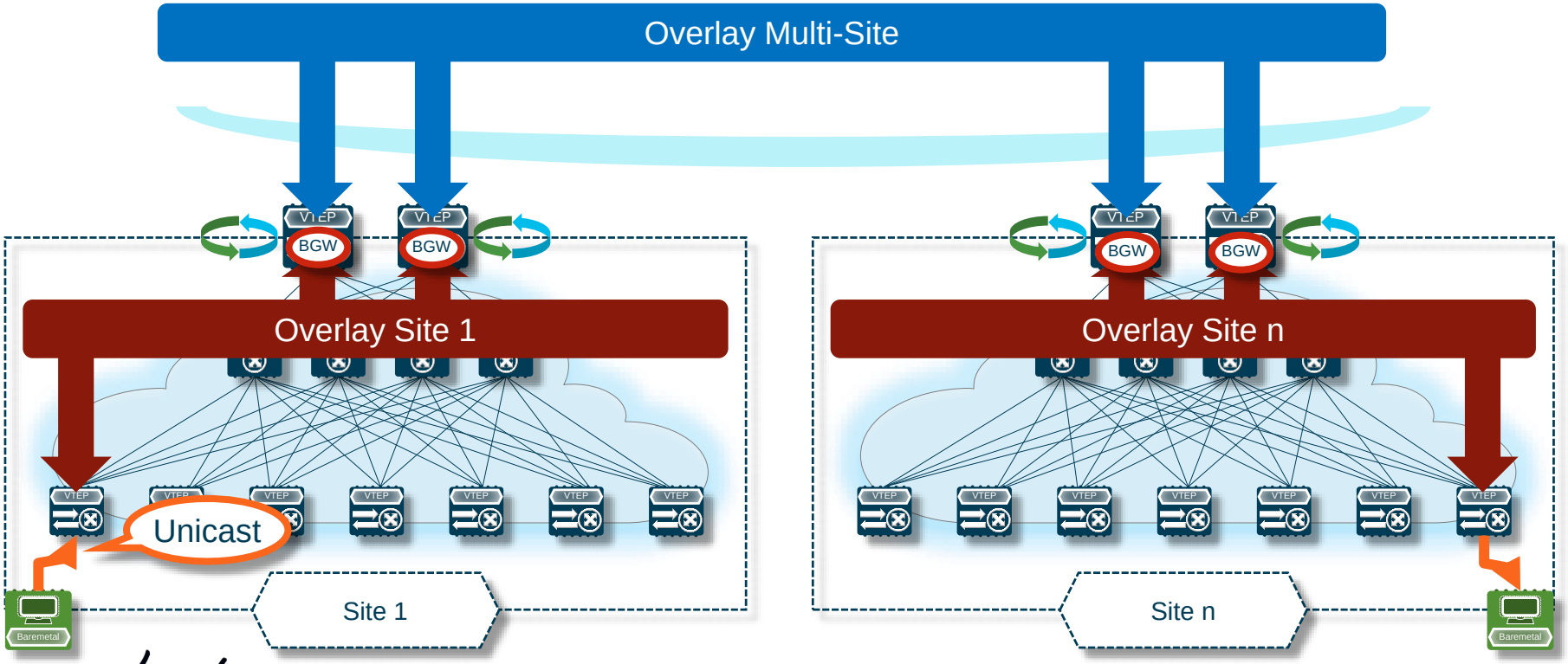
VXLAN Multi-Site

Functional Components

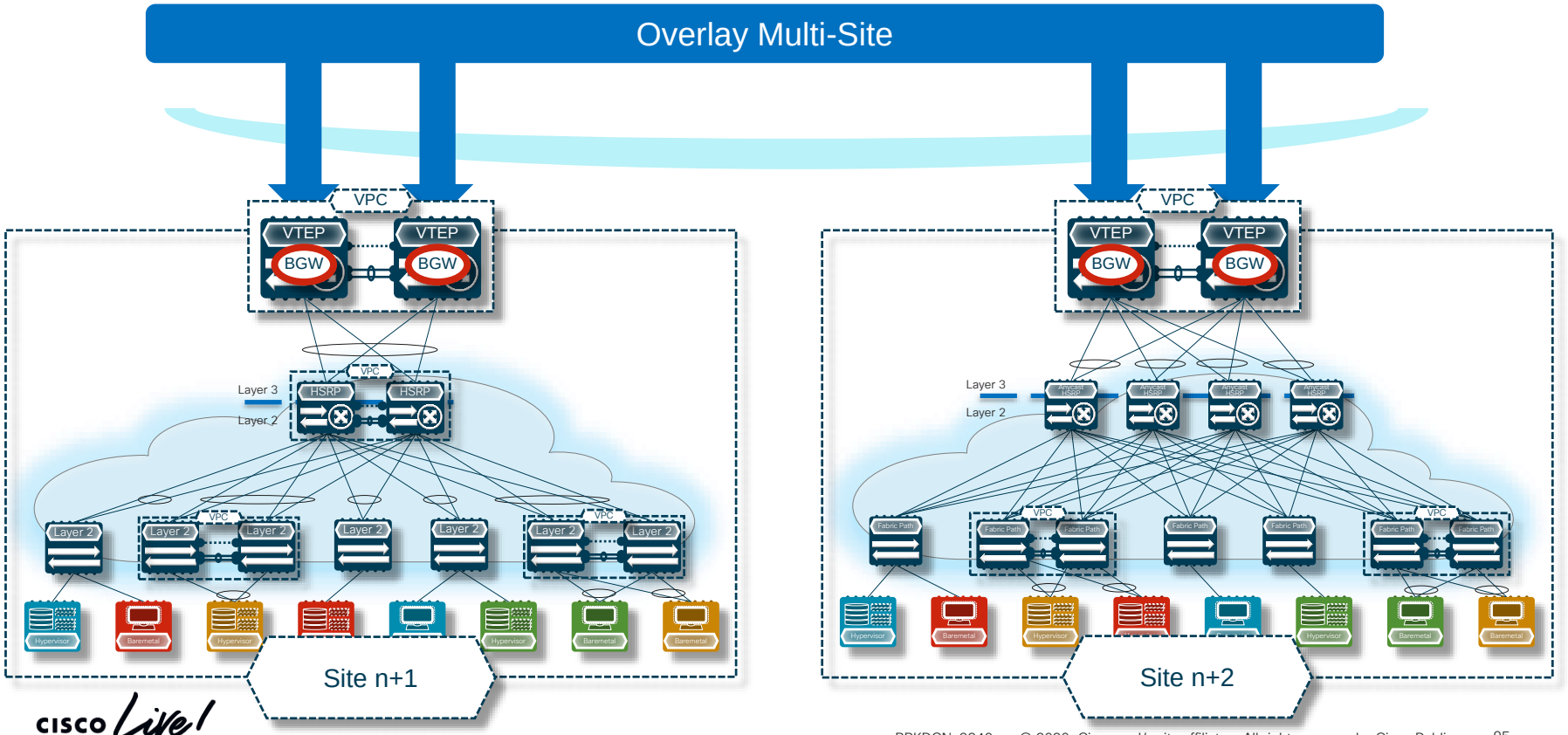
https://tools.ietf.org/html/draft-sharma-multi-site-evpn



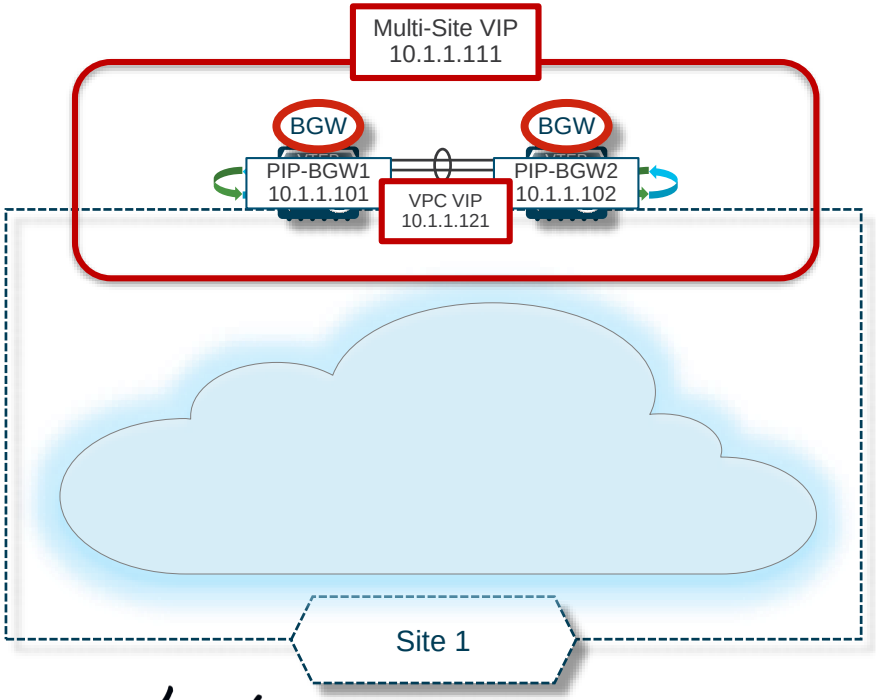
Multi-Site – Hierarchical Overlay Domains



vPC Border Gateway



vPC Border Gateway and Transit Traffic

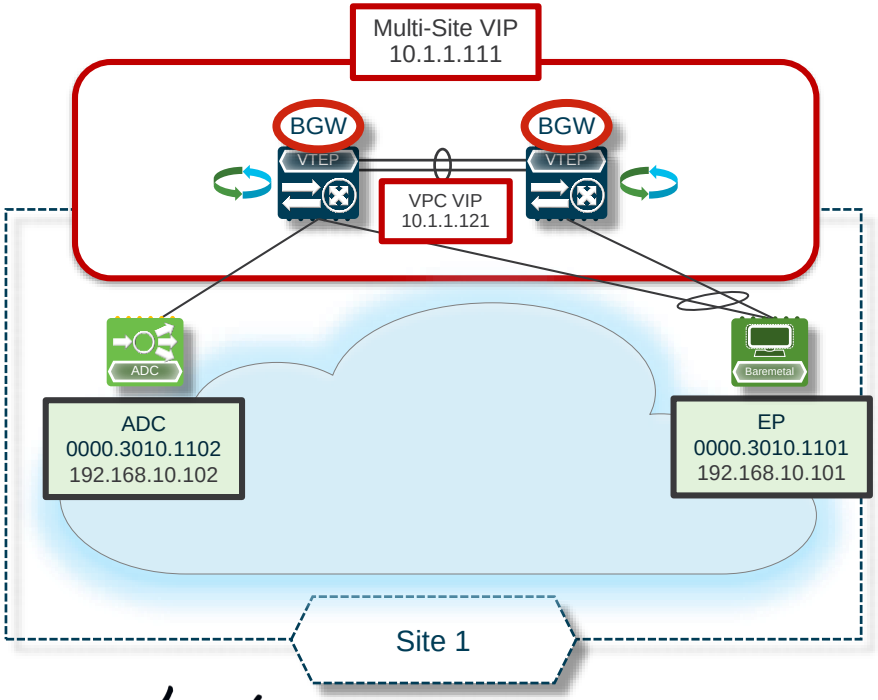


vPC Border Gateway

- Common Multi-Site Virtual IP (Multi-Site VIP) across BGWs
  - Multi-Site VIP for Inter-Site transit communication (transit)
- Common VPC Virtual IP (vPC VIP) across BGWs
  - Used by default for external communication
  - Used for Broadcast, Unknown Unicast and Multicast (BUM) replication
- Individual Primary IP (PIP) per BGW
  - Used for external communication with “advertised-pip”

VPC Border Gateway

Locally Attached End-Points

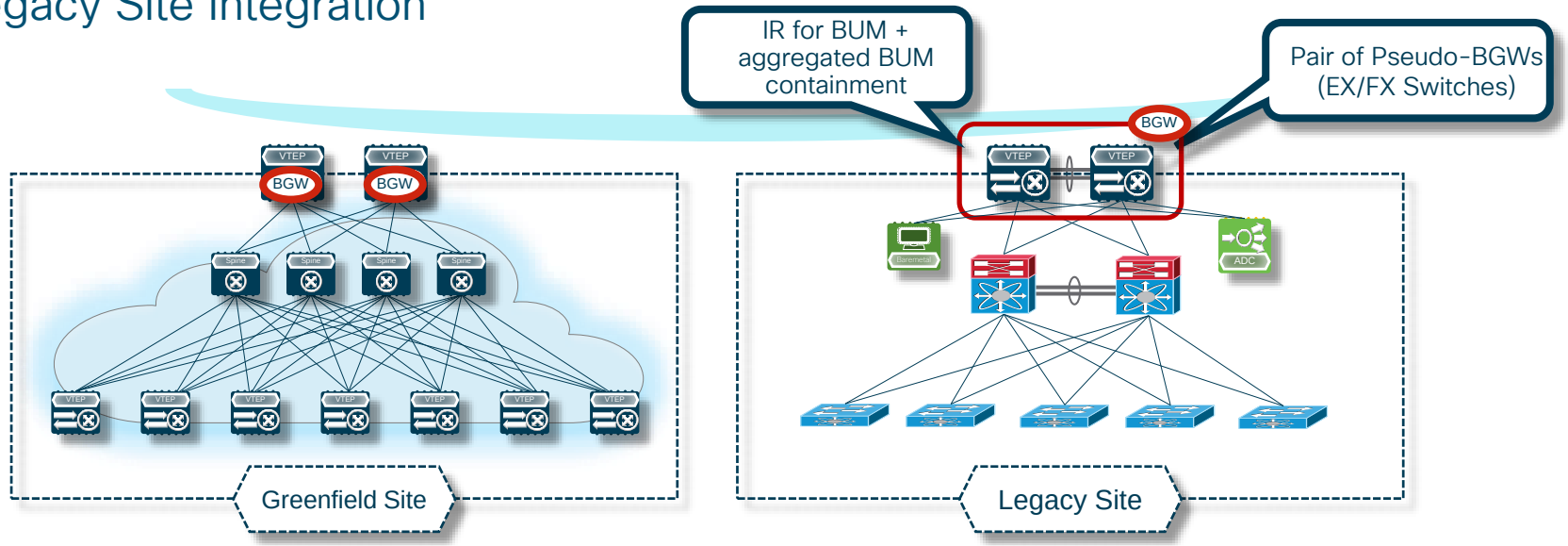


vPC Border Gateway

- Single- or Dual-Homed End-Points
  - Services Appliance (i.e. Firewall, ADC etc.)
  - Physical or Virtual Servers
- Advertised and Reachable through vPC Virtual IP Address (vPC VIP)
  - Intra-Site: Leaf nodes use vPC VIP to reach End-Points connected to Border Gateways
  - Inter-Site: Remote Border Gateways use vPC VIP to reach End-Points connected to Border Gateways
  - Traffic potentially traverses vPC Peer-Link

VXLAN Multi-Site

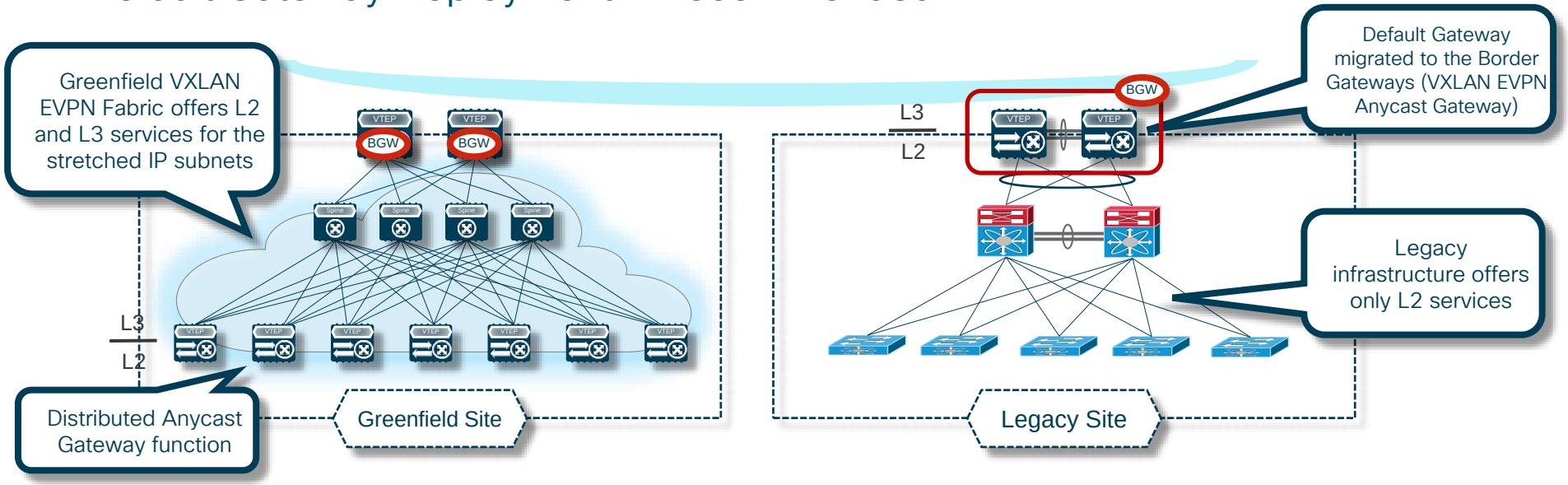
Legacy Site Integration



- Coexistence and/or migration use cases
  - Extend Layer-2 and Layer-3 multi-tenant connectivity across sites
- Deploy a pair of Pseudo-BGWs in the legacy site
  - Simplified configuration required on Pseudo-BGWs nodes
  - Still offering native Multi-Site functions (Ingress Replication for BUM, BUM containment, etc.)

Multi-Site and Legacy Site Integration

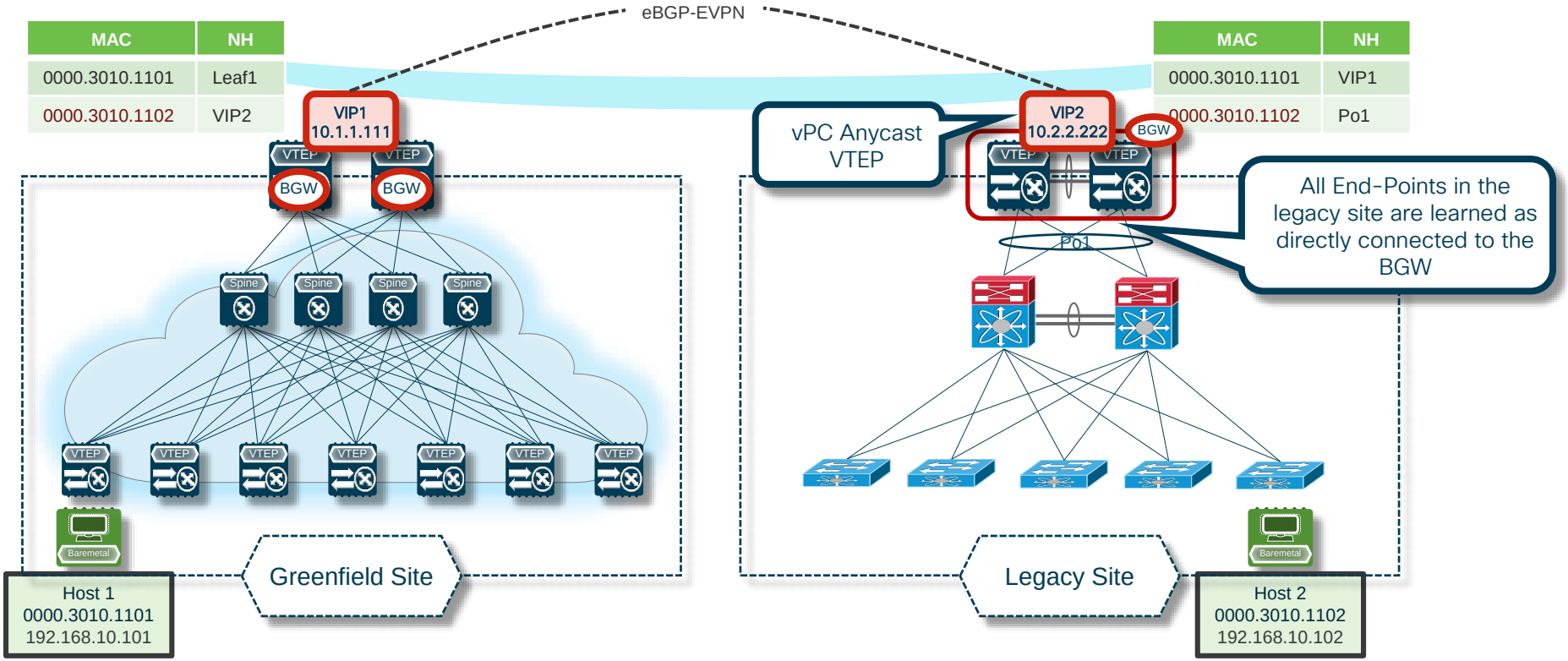
Default Gateway Deployment – Recommended



- Recommended approach is to migrate the default gateway from the legacy aggregation devices to the Border Gateways (VXLAN EVPN Anycast Gateway)
- Optimize routing between endpoints deployed across sites

Multi-Site and Legacy Site Integration

Layer-2 Control Plane Exchange across Sites

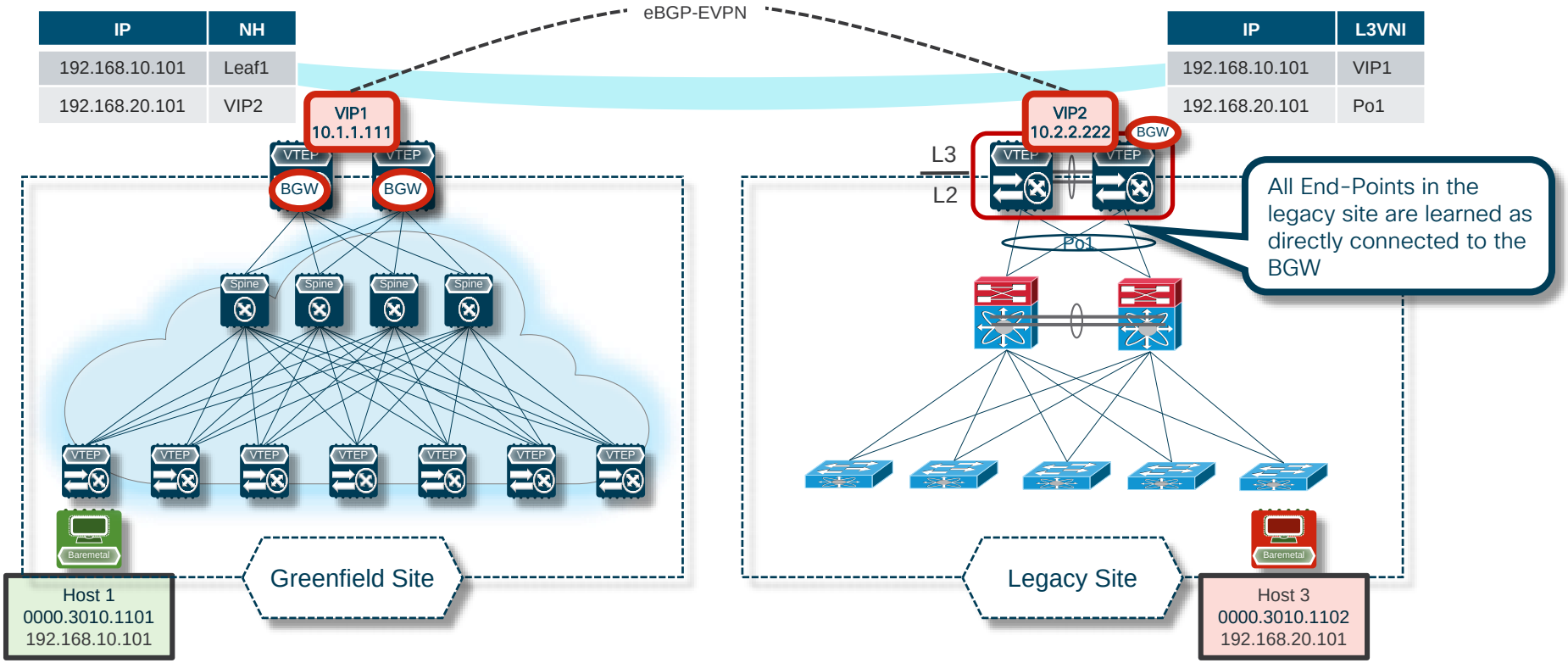


Multi-Site and Legacy Site Integration

Layer-3 Control Plane Exchange across Sites

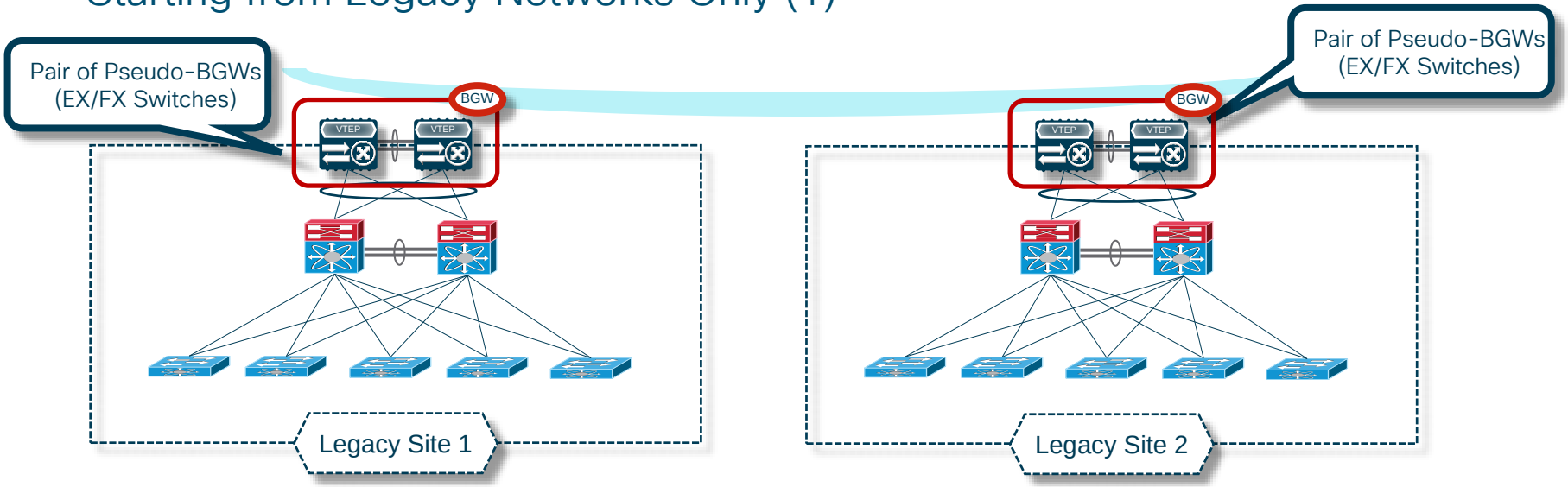
IP	NH
192.168.10.101	Leaf1
192.168.20.101	VIP2

IP	L3VNI
192.168.10.101	VIP1
192.168.20.101	Po1



VXLAN Multi-Site and Legacy Site Integration

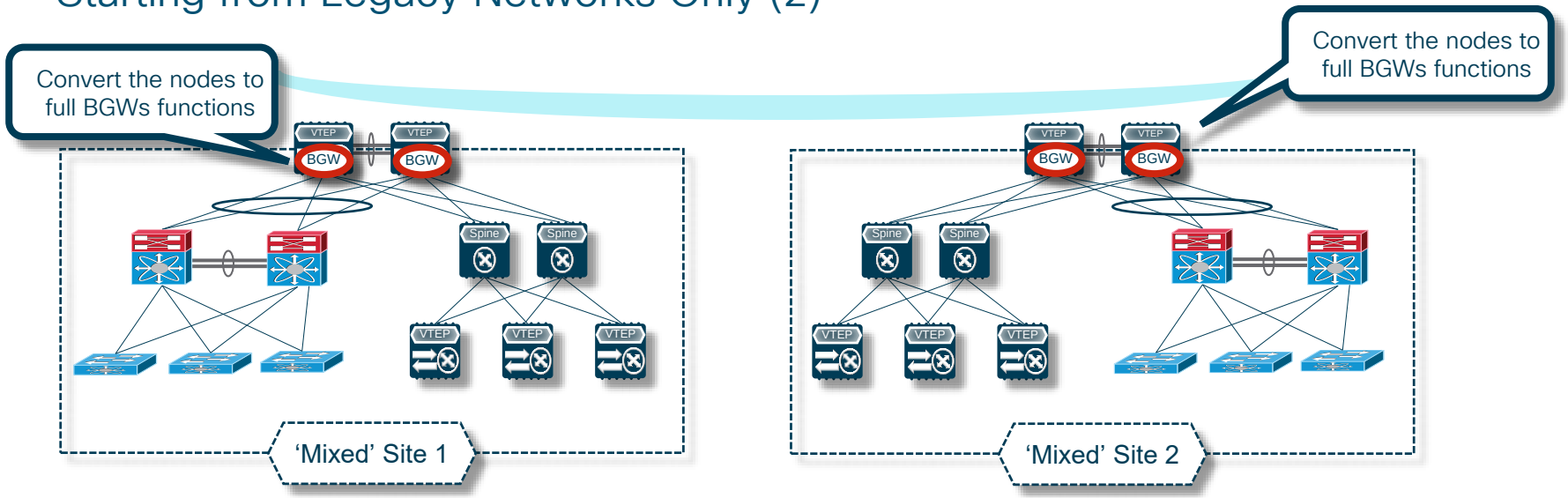
Starting from Legacy Networks Only (1)



- A pair of Pseudo-BGWs inserted in each legacy site to extend Layer-2 and Layer-3 connectivity between sites
  - Replacement of traditional DCI technologies (EoMPLS, VPLS, OTV, …)
- Slowly phase out the legacy networks and replace them with VXLAN EVPN fabrics

VXLAN Multi-Site and Legacy Site Integration

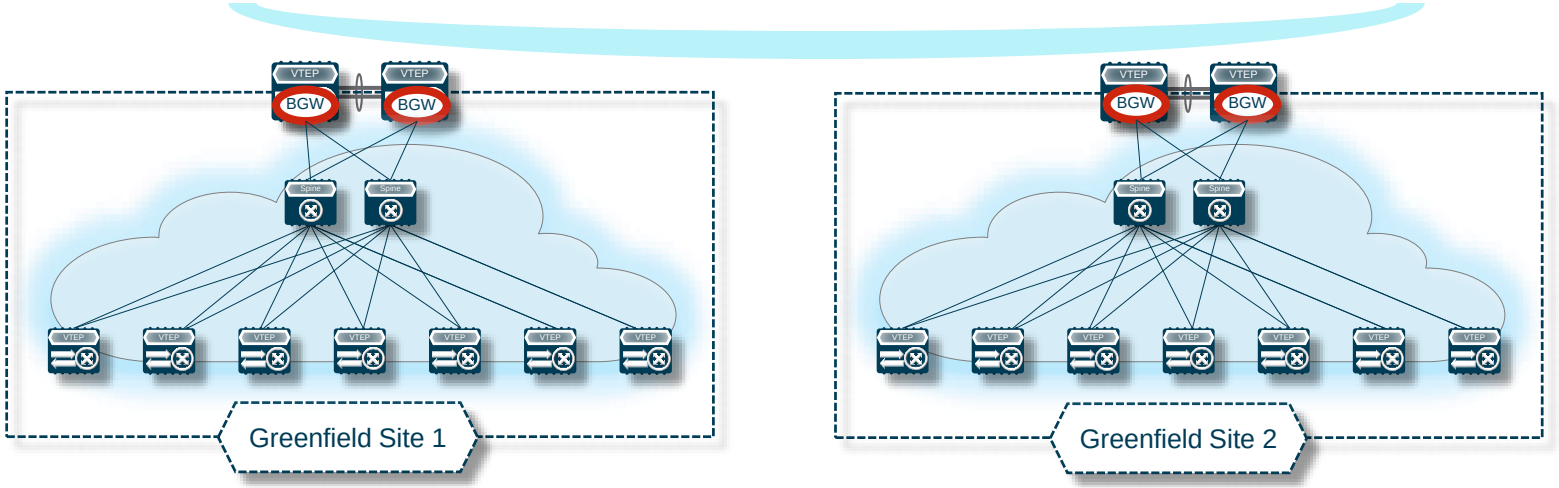
Starting from Legacy Networks Only (2)



- Introduce VXLAN EVPN spines and additional VTEPs in each site
- Convert the Pseudo-BGWs to full BGW (may require vPC support on BGWs)
- Migrate endpoints between the legacy network and the new VXLAN EVPN fabric

VXLAN Multi-Site and Legacy Site Integration

Starting from Legacy Networks Only (3)



- Decommission the legacy networks and leave only the VXLAN EVPN fabrics in place

MAC ECMP and IP ECMP

MAC ECMP and IP ECMP – Characteristics

MAC ECMP

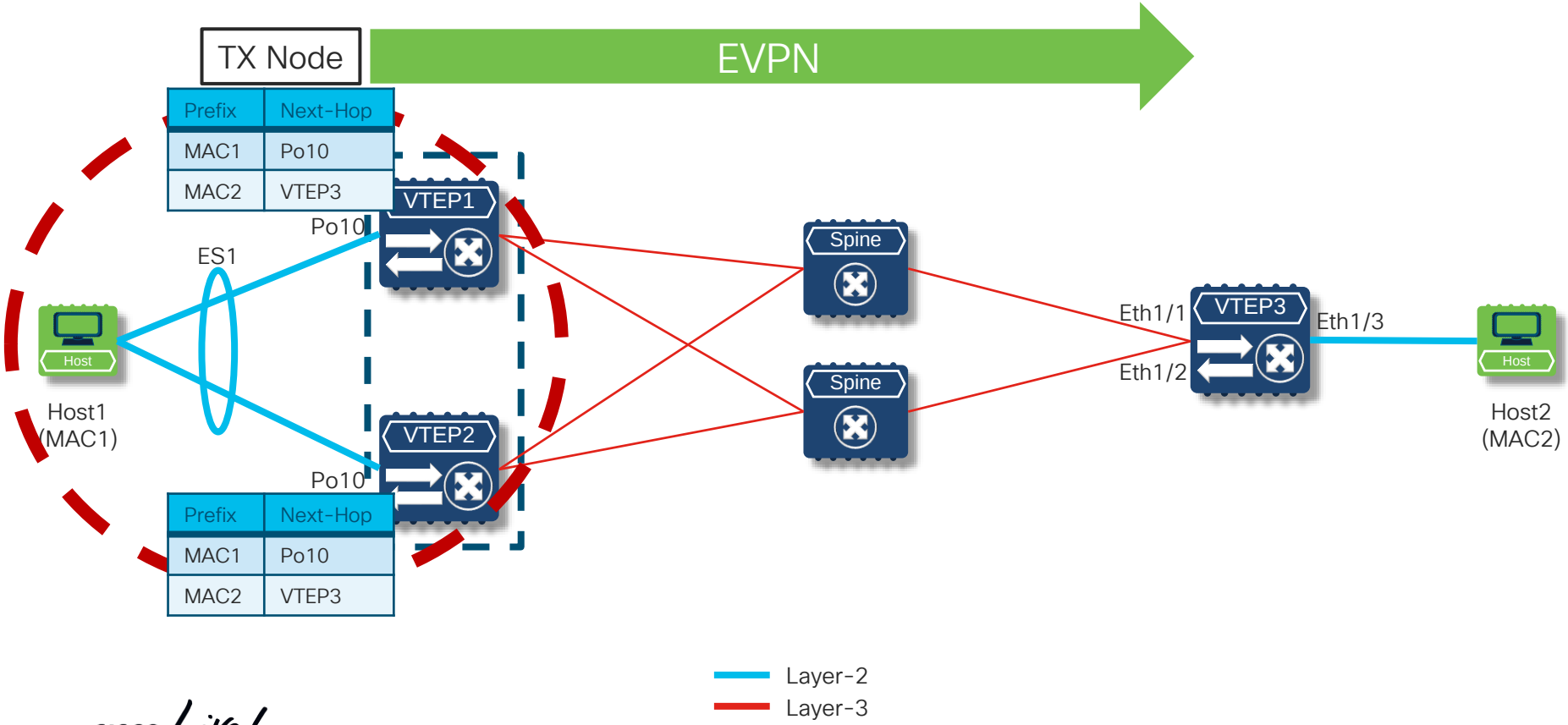
- MAC Multipathing – ESI
- Overlay ECMP
- Using two Next-Hops for a given MAC address
- Needs a Level of MAC-based indirection on the RX Node
- Specific Hardware capability needed

IP ECMP

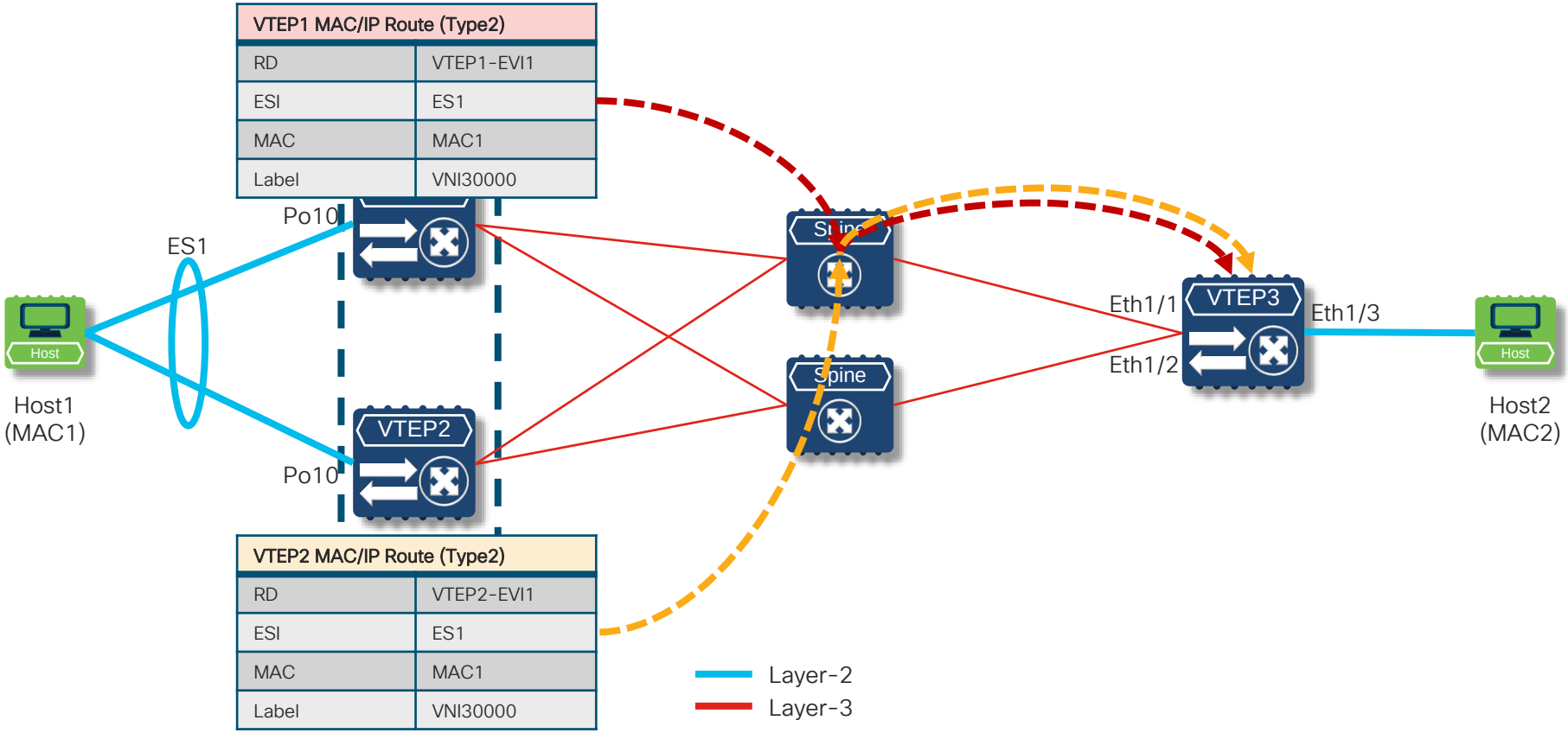
- IP Multipathing – VPC Anycast VTEP
- Underlay ECMP
- Using a single Next-Hop for a given MAC address
- Needs an IP-Based Level of indirection on the RX Node

MAC ECMP

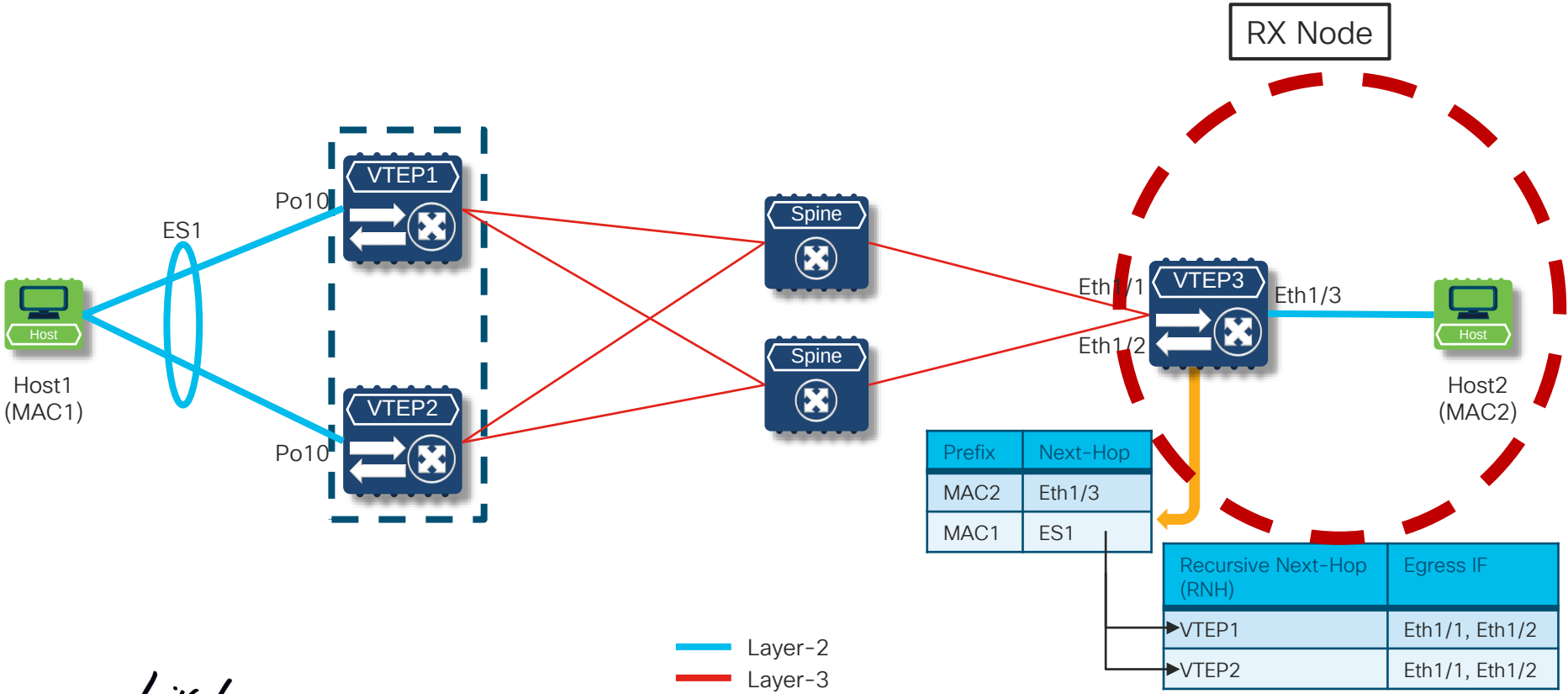
Multihomed Host – TX Node Bundling



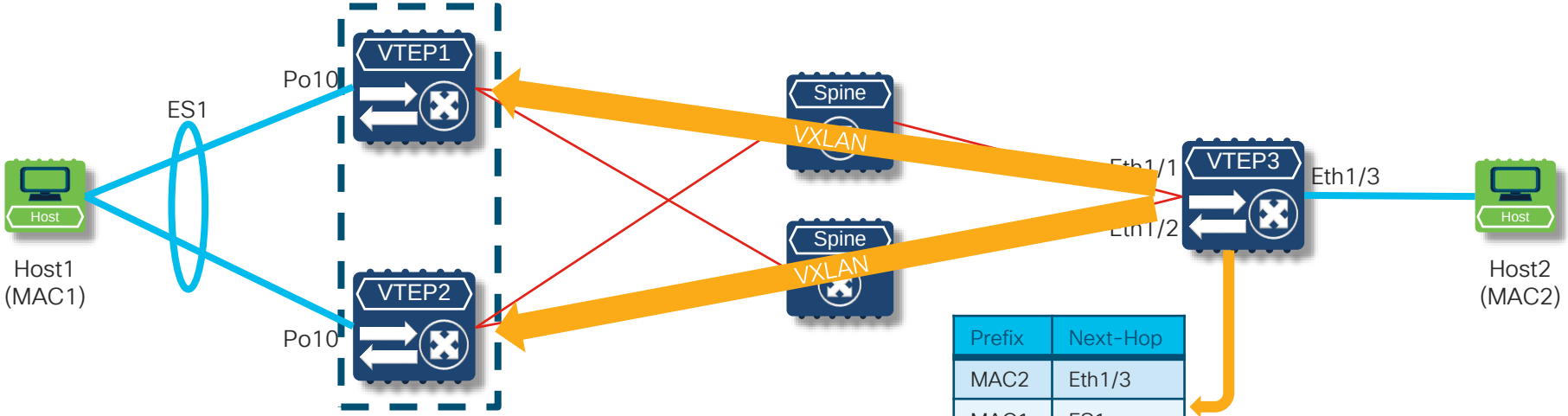
Multihomed Host – ECMP with MAC Multipath



Multihomed Host – RX Node View



Multihomed Host – RX Node Data Plane View



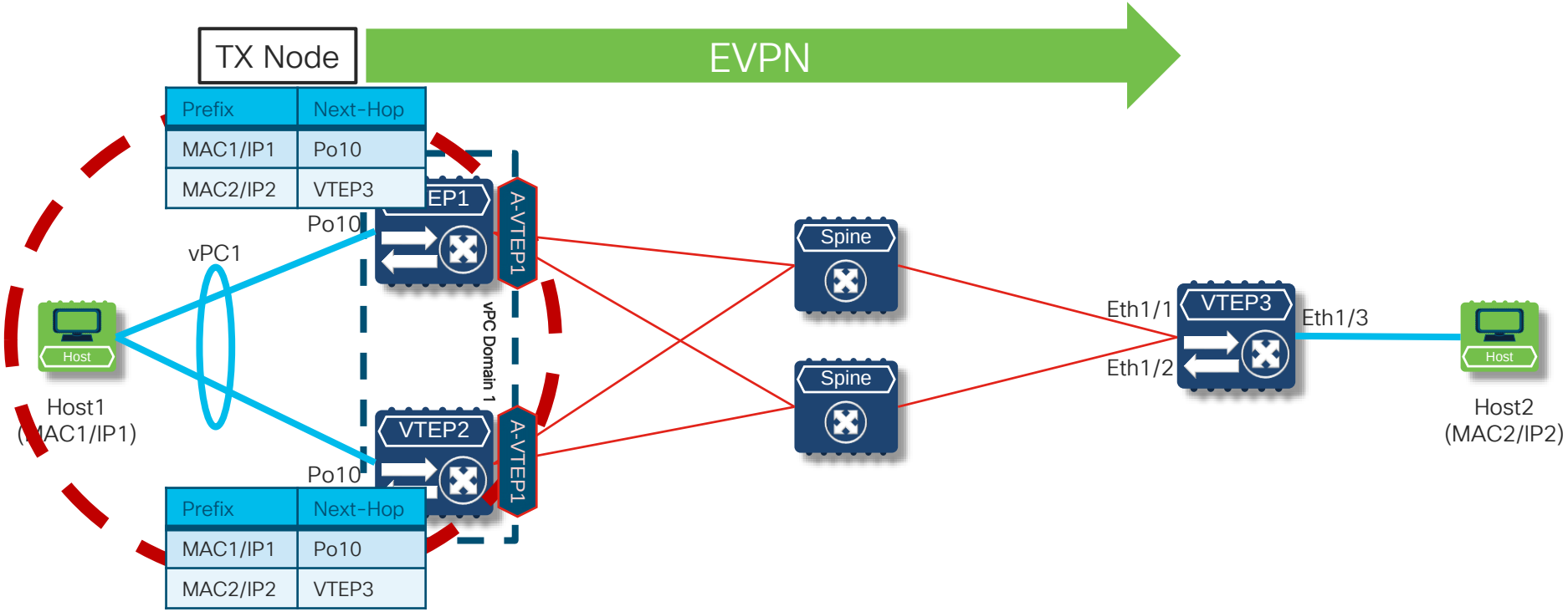
Prefix	Next-Hop
MAC2	Eth1/3
MAC1	ES1

Recursive Next-Hop (RNH)	Egress IF
VTEP1	Eth1/1, Eth1/2
VTEP2	Eth1/1, Eth1/2

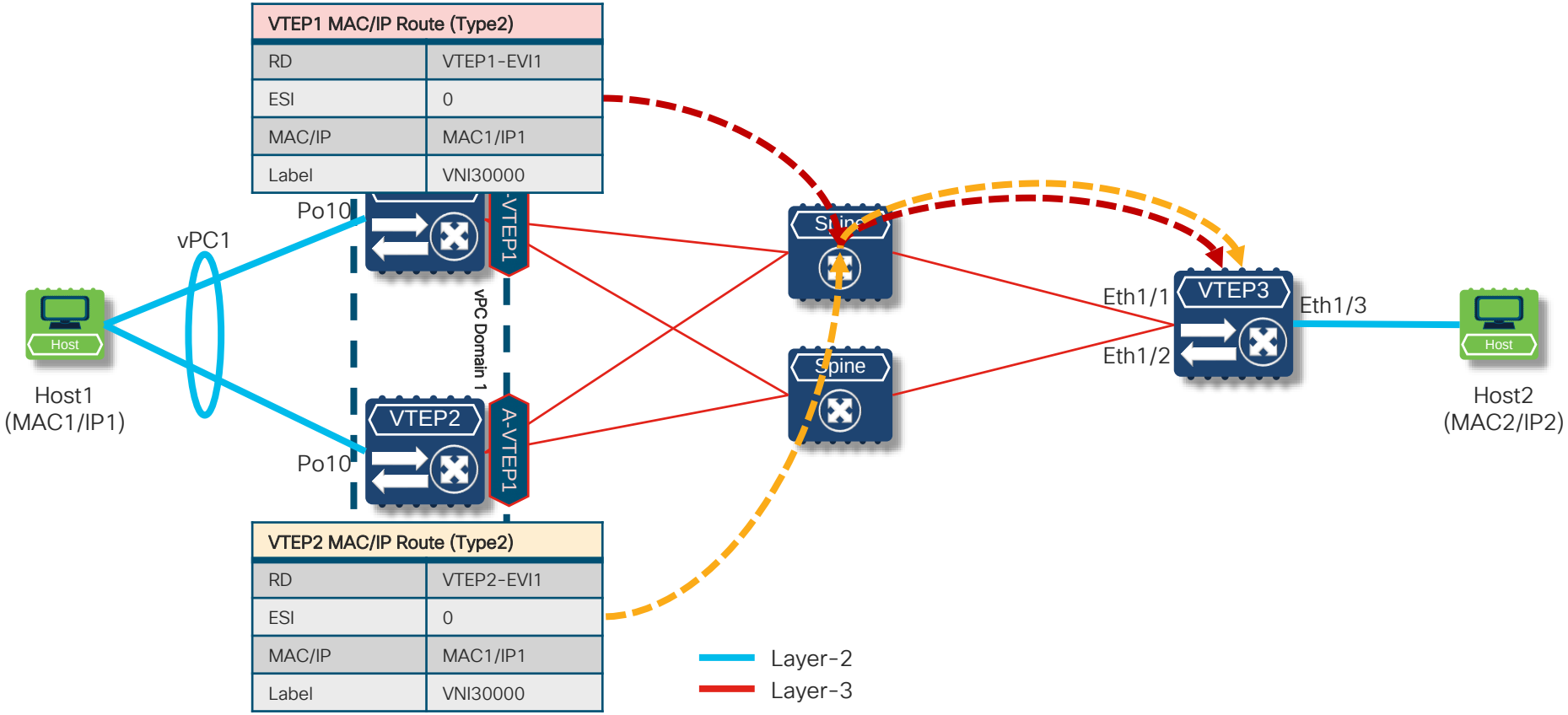
Layer-2
Layer-3

IP ECMP

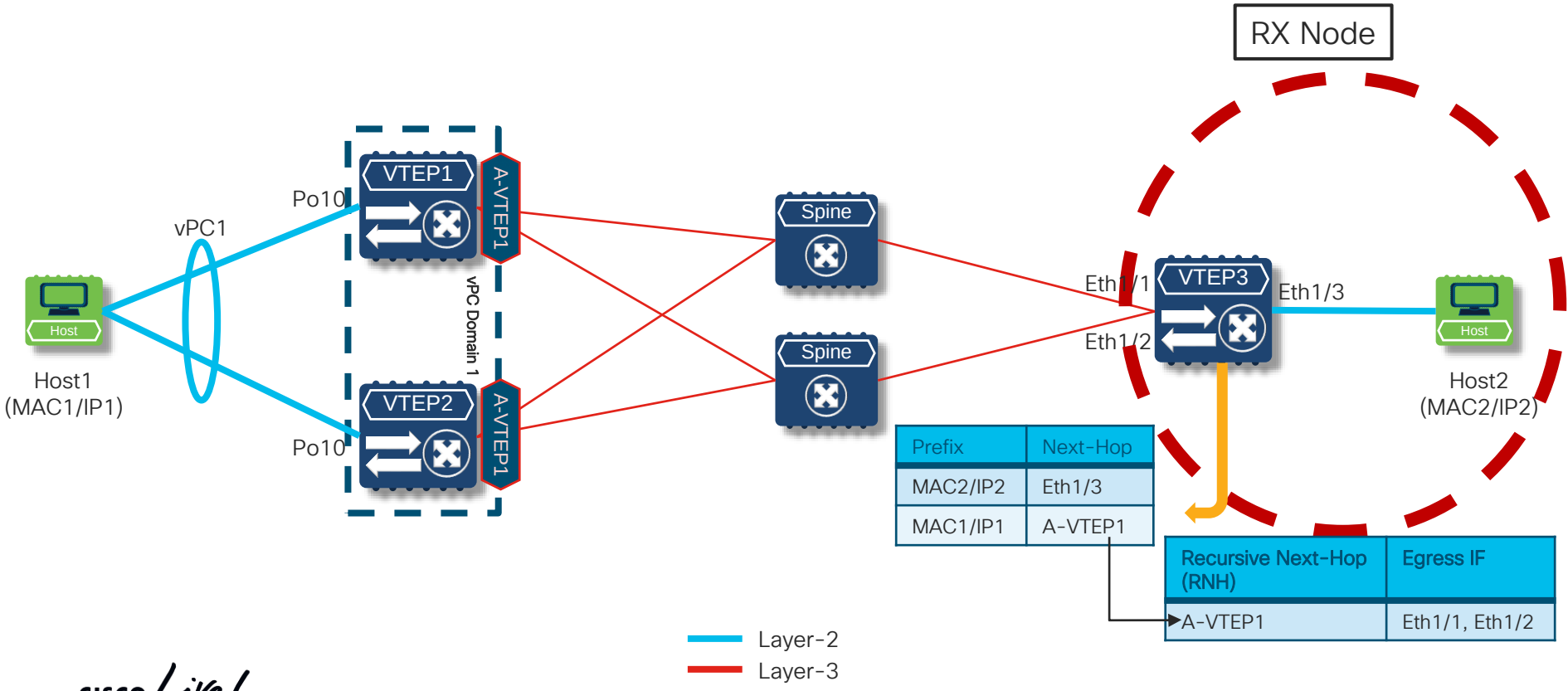
Multihomed Host – TX Node Bundling



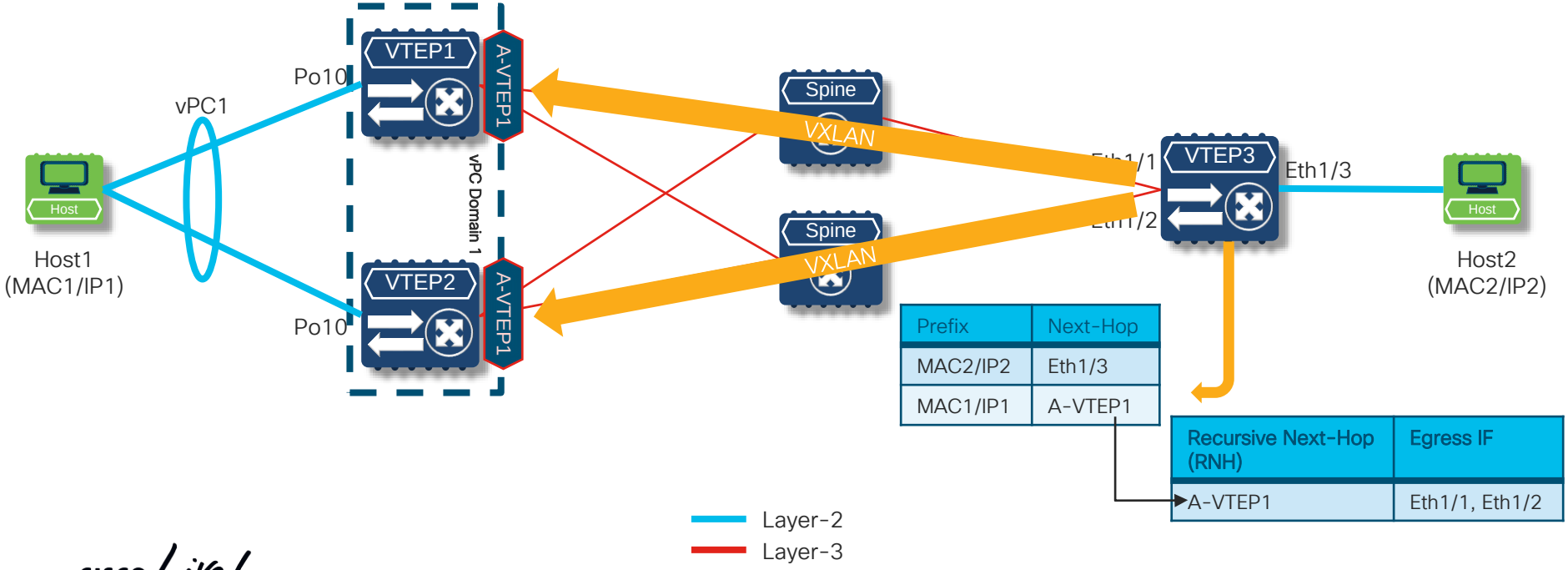
Multihomed Host – Anycast with ECMP



Multihomed Host – RX Node View

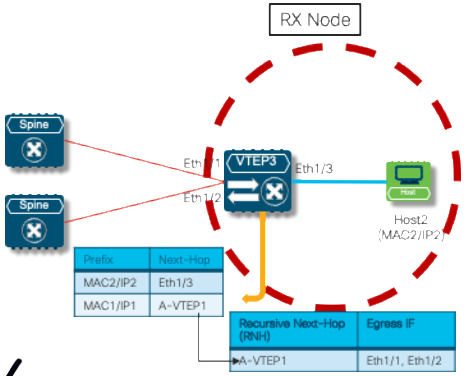


Multihomed Host – RX Node Data Plan View

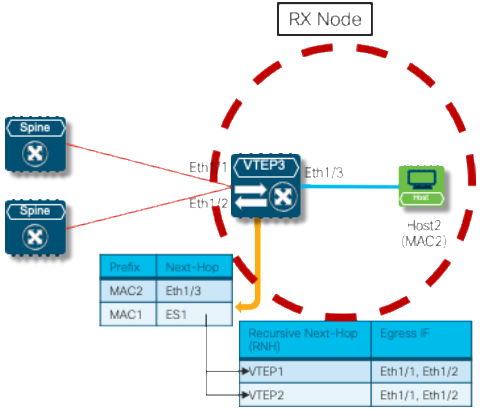


Anycast and MAC ECMP

- Anycast uses an IP-Based indirection to represent a single MAC Address behind multiple remote Nodes with a common IP Address
- ECMP is used for reaching the Anycast IP Address (Anycast VTEP)



- MAC ECMP uses an indirection to represent a single MAC Address behind multiple remote Nodes with individual IP Addresses
- ECMP is used to have common MAC behind multiple remote Nodes



Key Takeaways

Key Takeaways

- vPC provides full bandwidth usage without Spanning Three
- Dual-Home end point devices to VXLAN fabric
- vPC is better with vPC Fabric Peering
- Boarder Gateway for DCI and migration to VXLAN fabric

NXOS Fabric Operations

Keynote — Tuesday 09:30

BRKDCN-2458 — Tuesday 11:00
NXOS Operations & practices

BRKDCN-1645 — Tuesday 14:30
Introduction VXLAN

BRKDCN-1687 — Tuesday 15:45
Introduction DCNM

BRKDCN-2939 — Tuesday 17:00
Easy NXOS Fabric (VXLAN)

BRKDCN-2249 — Wednesday 08:30
VXLAN VPC

BRKDCN-2125 — Wednesday 14:45
Overlay Management w/DCNM

BRKDCN-2011 — Wednesday 16:45
Network DevOps in your DC

BRKDCN-3020 — Thursday 11:15
Visibility & Monitoring

BRKDCN-2712 — Thursday 14:45
Insights w/Analytics

Keynote — Thursday 17:00

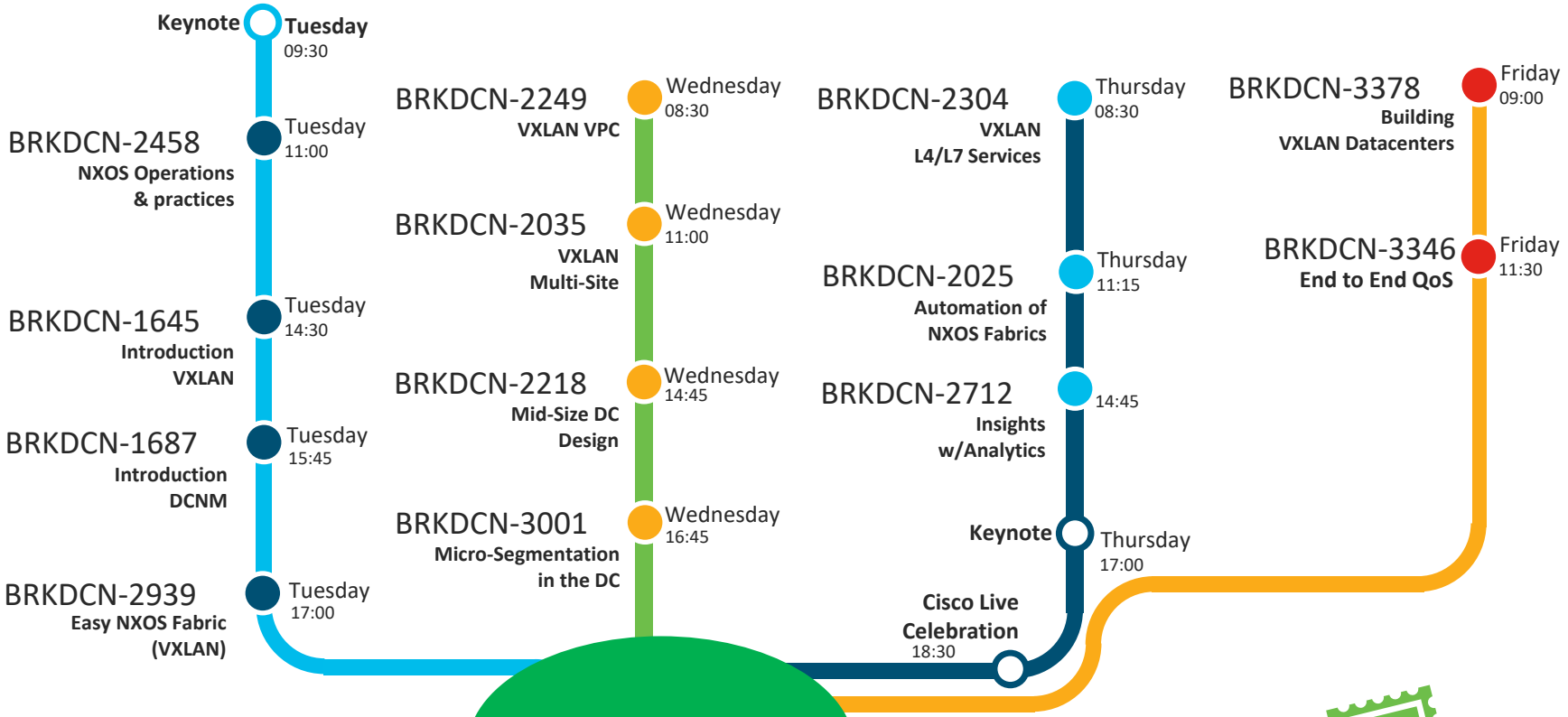
Cisco Live Celebration 18:30

BRKDCN-3378 — Friday 09:00
Building VXLAN Datacenters

BRKDCN-3040 — Friday 11:30
Troubleshooting VXLAN

CISCO Live!

GURU



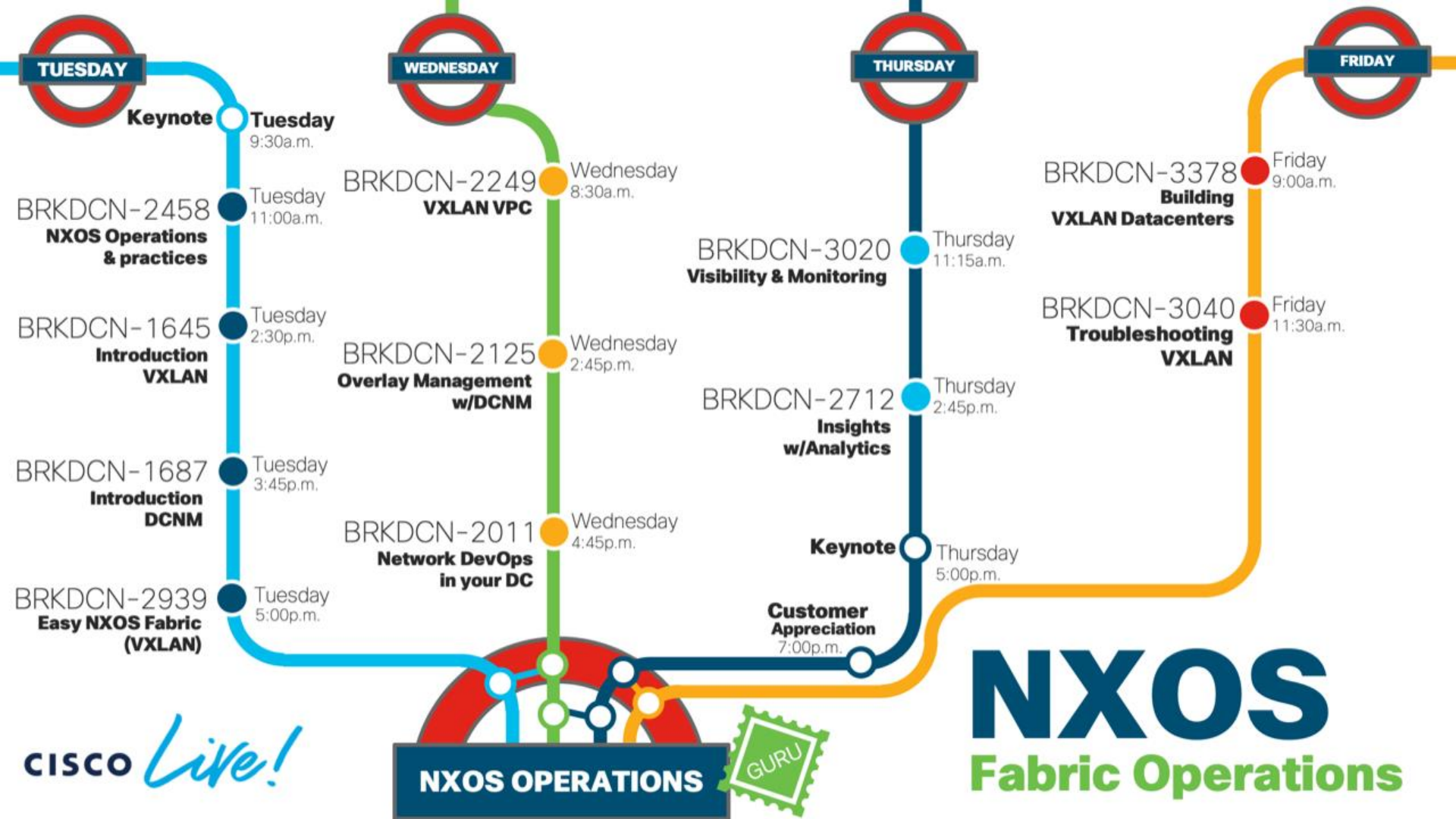
NXOS

Fabric Technology

GURU

cisco Live!


TUESDAY
Keynote
Tuesday 9:30a.m.
BRKDCN-2458 NXOS Operations & practices
Tuesday 11:00a.m.
BRKDCN-1645 Introduction VXLAN
Tuesday 2:30p.m.
BRKDCN-1687 Introduction DCNM
Tuesday 3:45p.m.
BRKDCN-2939 Easy NXOS Fabric (VXLAN)
Tuesday 5:00p.m.
WEDNESDAY
BRKDCN-2249 VXLAN VPC
Wednesday 8:30a.m.
BRKDCN-2035 VXLAN Multi-Site
Wednesday 11:00a.m.
BRKDCN-2218 Mid-Size DC Design
Wednesday 2:45p.m.
BRKDCN-3001 Micro-Segmentation in the DC
Wednesday 4:45p.m.
THURSDAY
BRKDCN-2304 VXLAN L4/L7 Services
Thursday 8:30a.m.
BRKDCN-2025 Automation of NXOS Fabrics
Thursday 11:15a.m.
BRKDCN-2712 Insights w/Analytics
2:45p.m.
Keynote
Thursday 5:00p.m.
Customer Appreciation
7:00p.m.
FRIDAY
BRKDCN-3378 Building VXLAN Datacenters
Friday 9:00a.m.
BRKDCN-3346 End to End QoS
Friday 11:30a.m.
CISCO Live!
NXOS TECHNOLOGY
GURU
NXOS
Fabric Technology


TUESDAY
WEDNESDAY
THURSDAY
FRIDAY
Keynote
Tuesday
9:30a.m.
BRKDCN-2458
NXOS Operations & practices
Tuesday
11:00a.m.
BRKDCN-1645
Introduction VXLAN
Tuesday
2:30p.m.
BRKDCN-1687
Introduction DCNM
Tuesday
3:45p.m.
BRKDCN-2939
Easy NXOS Fabric (VXLAN)
Tuesday
5:00p.m.
BRKDCN-2249
VXLAN VPC
Wednesday
8:30a.m.
BRKDCN-2125
Overlay Management w/DCNM
Wednesday
2:45p.m.
BRKDCN-2011
Network DevOps in your DC
Wednesday
4:45p.m.
BRKDCN-3020
Visibility & Monitoring
Thursday
11:15a.m.
BRKDCN-2712
Insights w/Analytics
Thursday
2:45p.m.
Keynote
Thursday
5:00p.m.
Customer Appreciation
7:00p.m.
BRKDCN-3378
Building VXLAN Datacenters
Friday
9:00a.m.
BRKDCN-3040
Troubleshooting VXLAN
Friday
11:30a.m.
NXOS OPERATIONS
GURU
CISCO Live!
NXOS
Fabric Operations

Complete your online session survey

- Please complete your session survey after each session. Your feedback is very important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (starting on Thursday) to receive your Cisco Live t-shirt.
- All surveys can be taken in the Cisco Events Mobile App or by logging in to the Content Catalog on ciscolive.com/emea.

Cisco Live sessions will be available for viewing on demand after the event at ciscolive.com.

Continue your education

Demos in the Cisco campus

Walk-in labs

Meet the engineer 1:1 meetings

Related sessions

Thank you

CISCO

You make **possible**