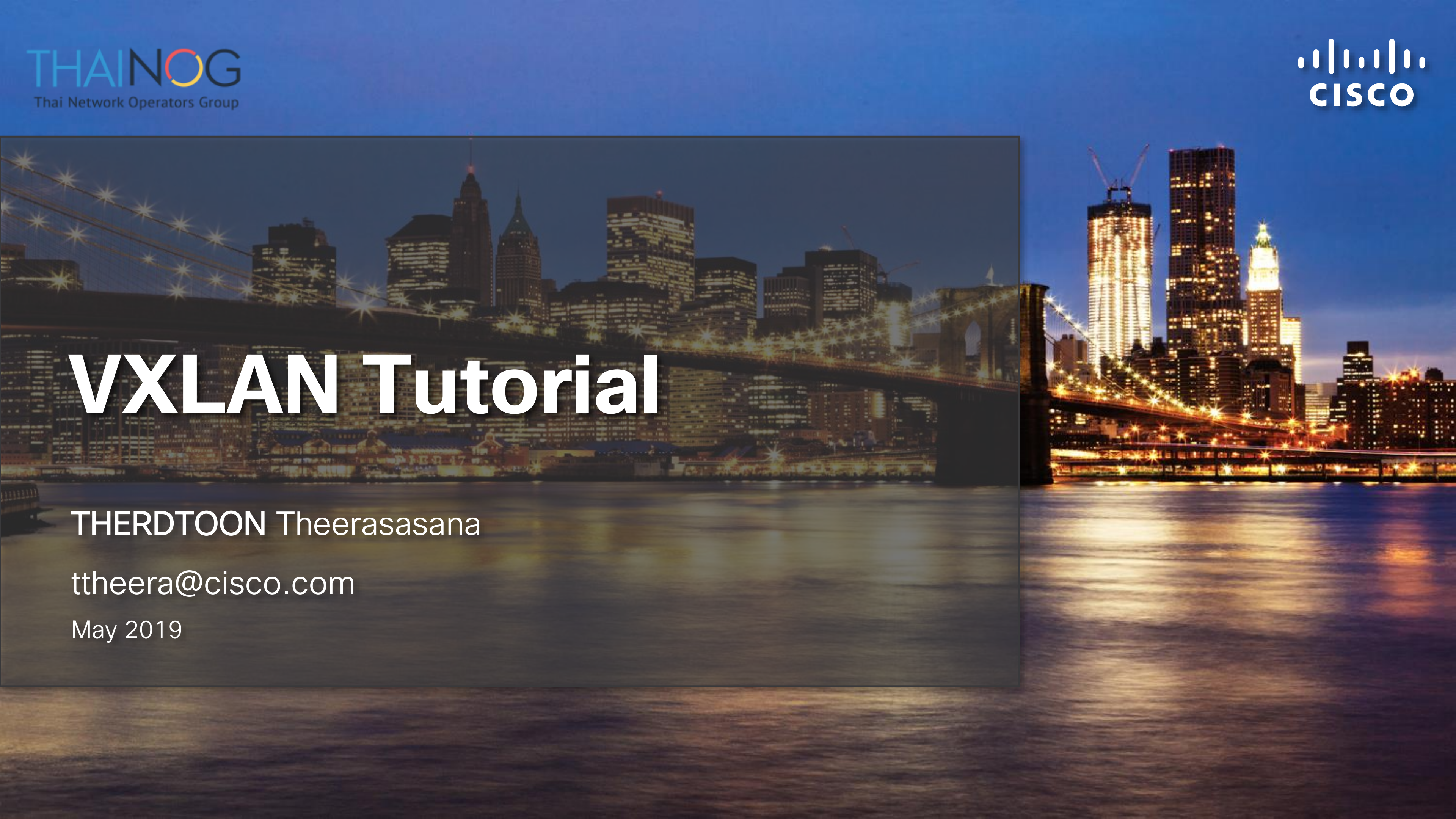


THAINOG
Thai Network Operators Group

cisco

VXLAN Tutorial

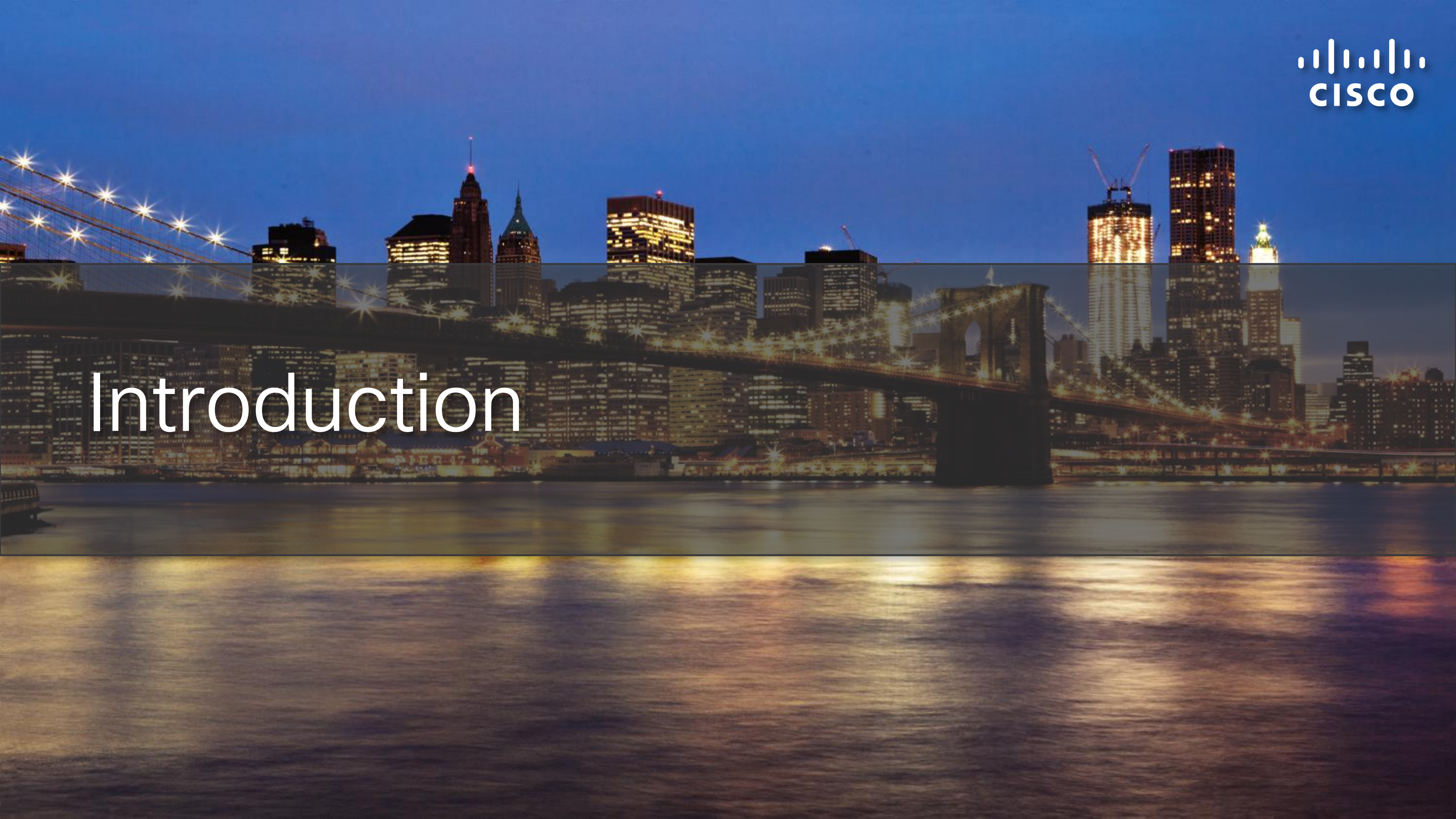
THERDTOON Theerasasana

ttheera@cisco.com

May 2019

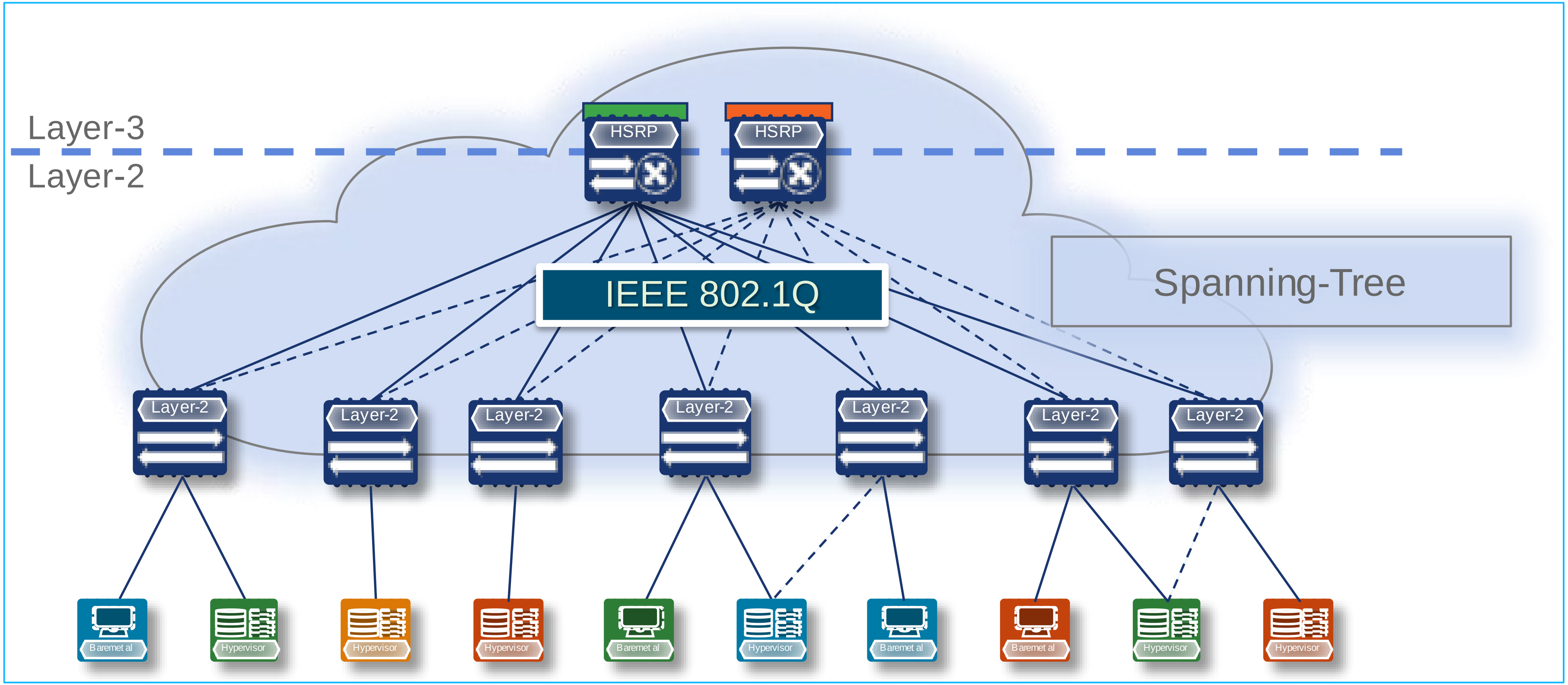
Agenda

- Introduction
- VXLAN – Flood and Learn
- VXLAN – BGP EVPN
- Summary

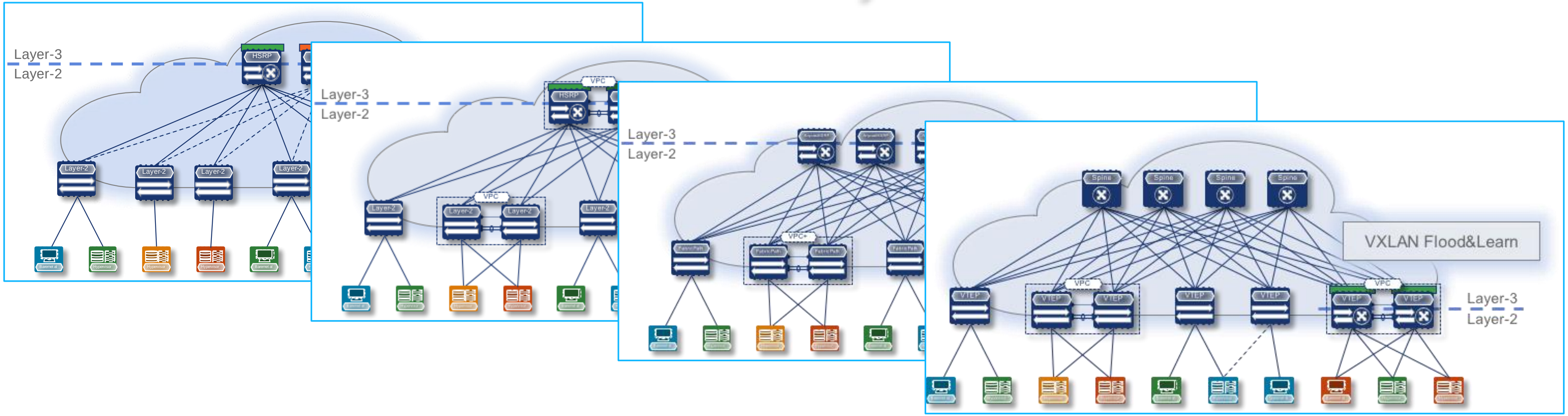


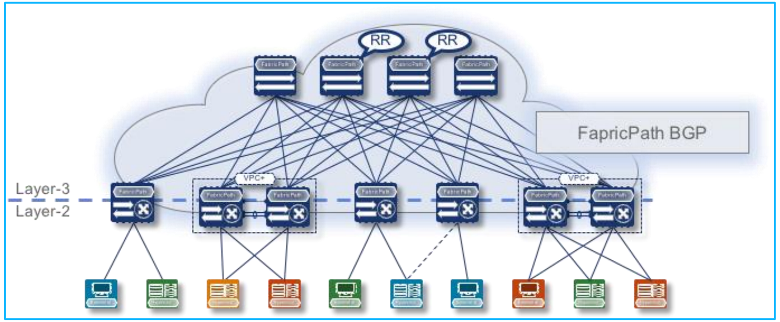
Introduction

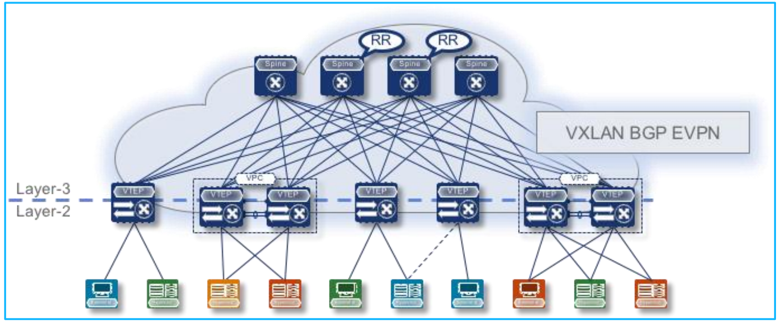
Traditional Data Center Networking



Data Center “Fabric” Journey



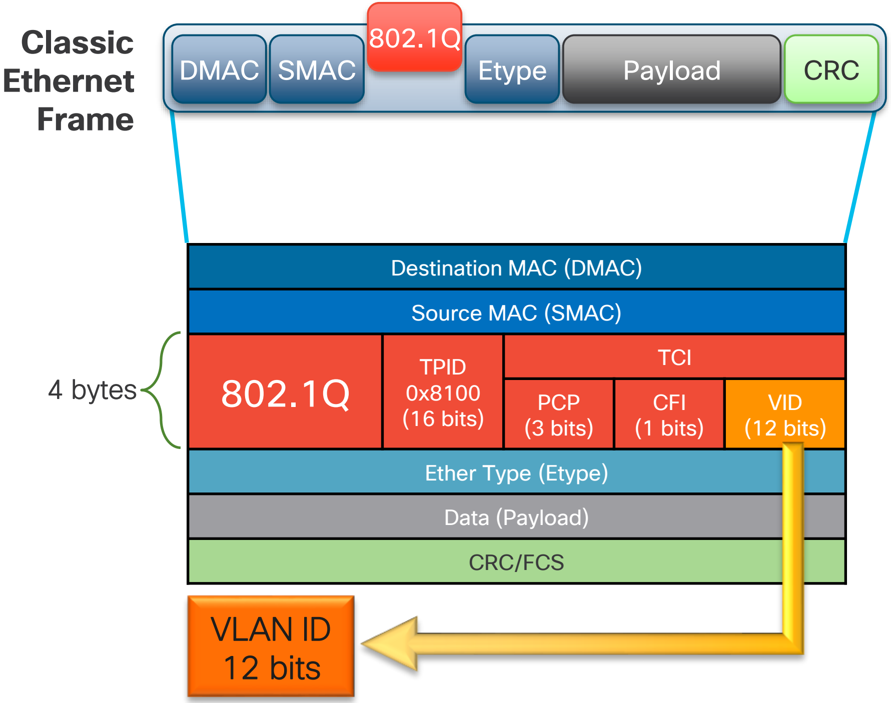




IEEE 802.1Q

Classic Ethernet IEEE 802.1Q Frame Format

- Traditionally VLAN is expressed over 12 bits (802.1Q tag)
  - Limits the maximum number of segments in a Data Center to 4096 VLANs



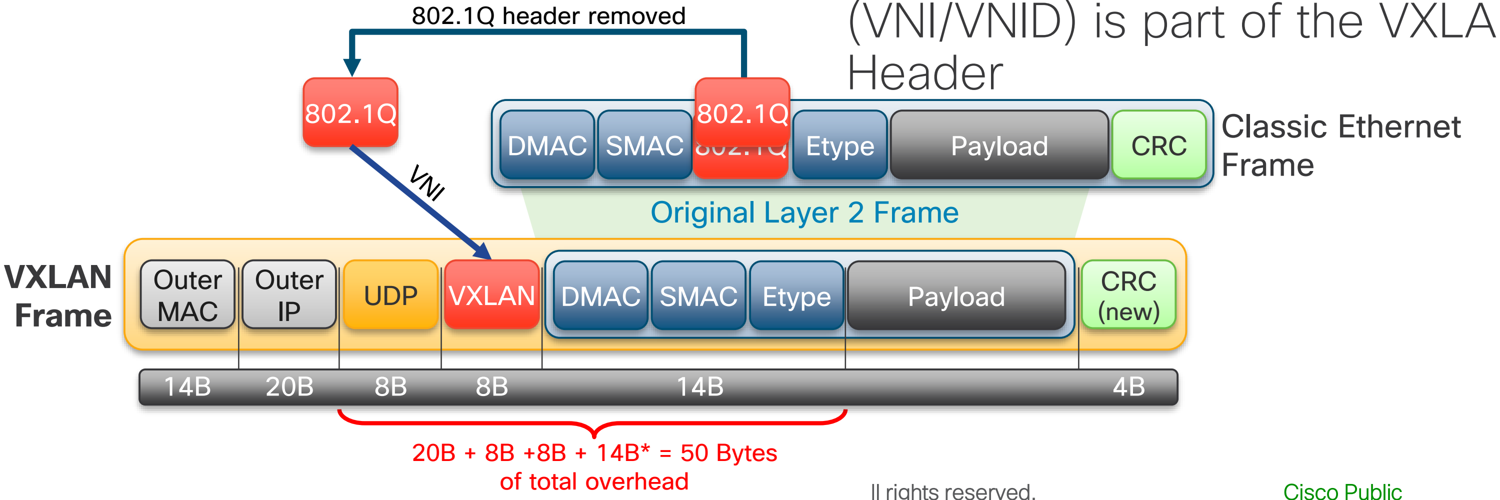
TPID = Tag Protocol Identifier, TCI = Tag Control Information,
PCP = Priority Code Point,
CFI = Canonical Format Indicator, VID = VLAN Identifier

VXLAN Overview

- Virtual eXtensible LAN
- Standards based Encapsulation
  - RFC 7348
  - Uses UDP-Encapsulation
- Transport Independent
  - Layer-3 Transport (Underlay)

- Transport Independent
  - Layer-3 Transport (Underlay)

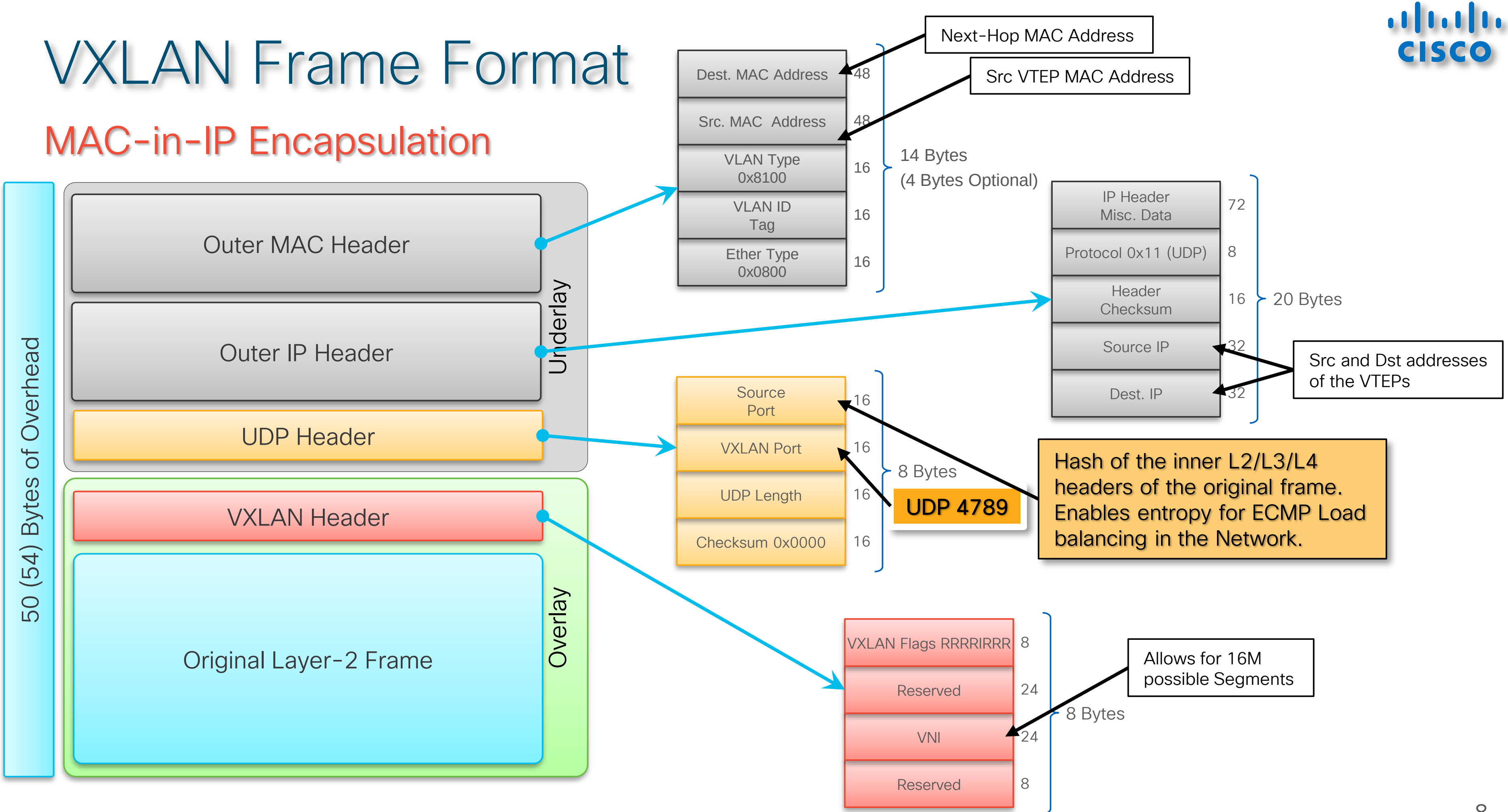
- VXLAN leverages the VNI field with a total address space of 24 bits
  - Support of ~16M segments
- The VXLAN Network Identifier (VNI/VNID) is part of the VXLAN Header



20B + 8B +8B + 14B* = 50 Bytes of total overhead

VXLAN Frame Format

MAC-in-IP Encapsulation



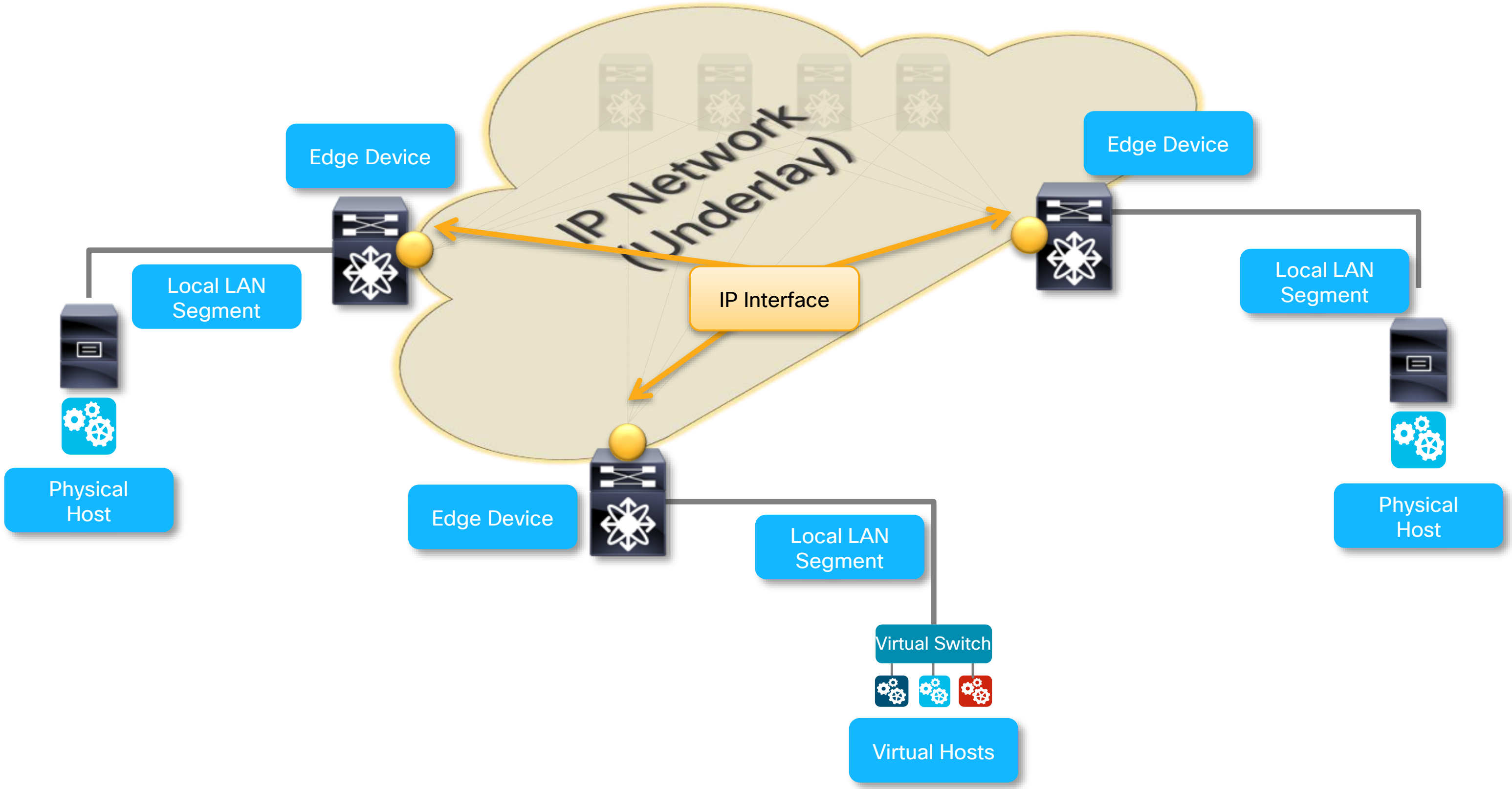
Why VXLAN?

VXLAN provides a Network with Segmentation, IP Mobility, and Scale

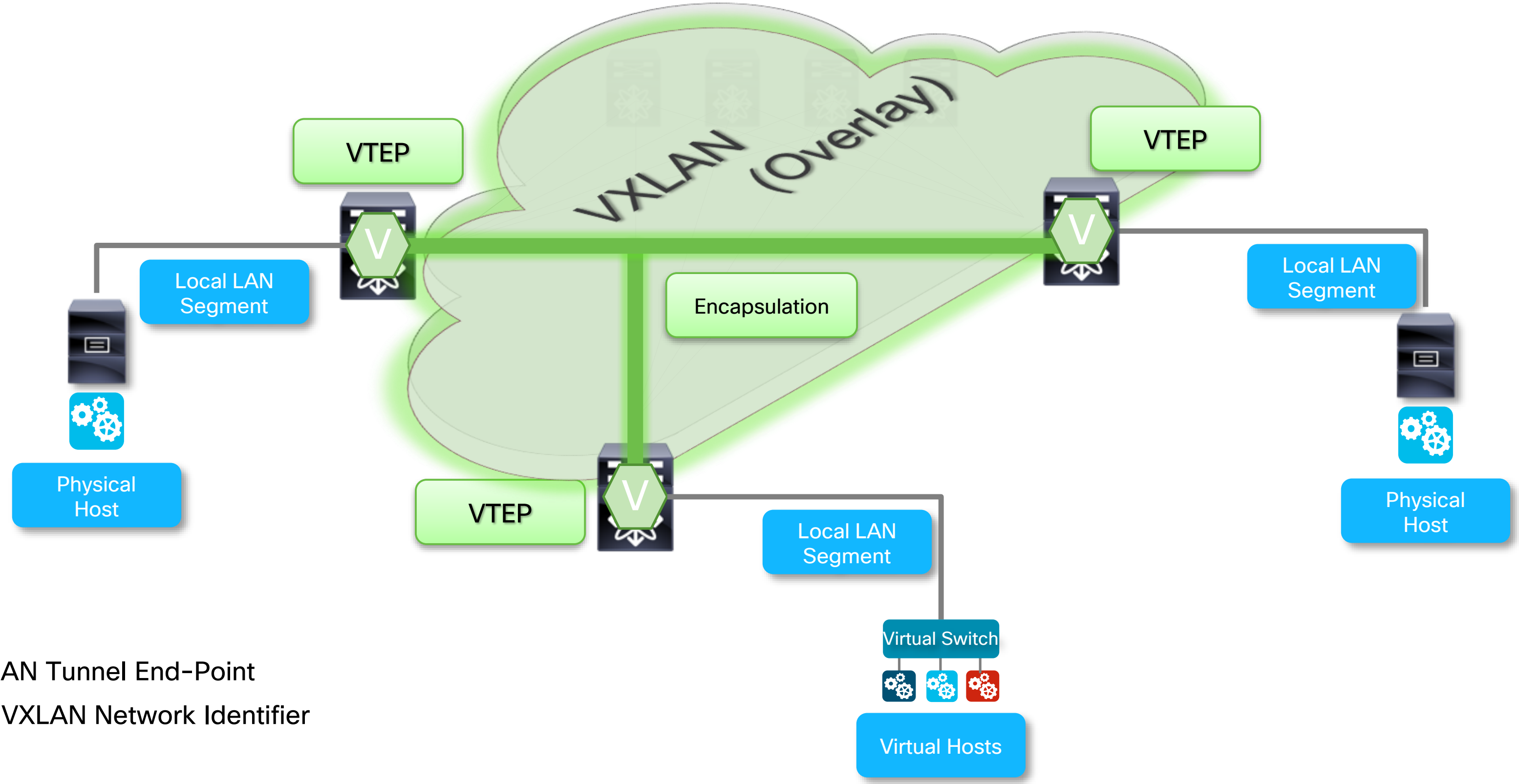
- “Standards” based Overlay
- Leverages Layer-3 ECMP – All links forwarding
- Increased Name-Space to 16M identifier
- Segmentation and Multi-Tenancy
- Integration of Physical and Virtual
- It’s SDN ☺

VXLAN – Flood and Learn

VXLAN Overview (1)



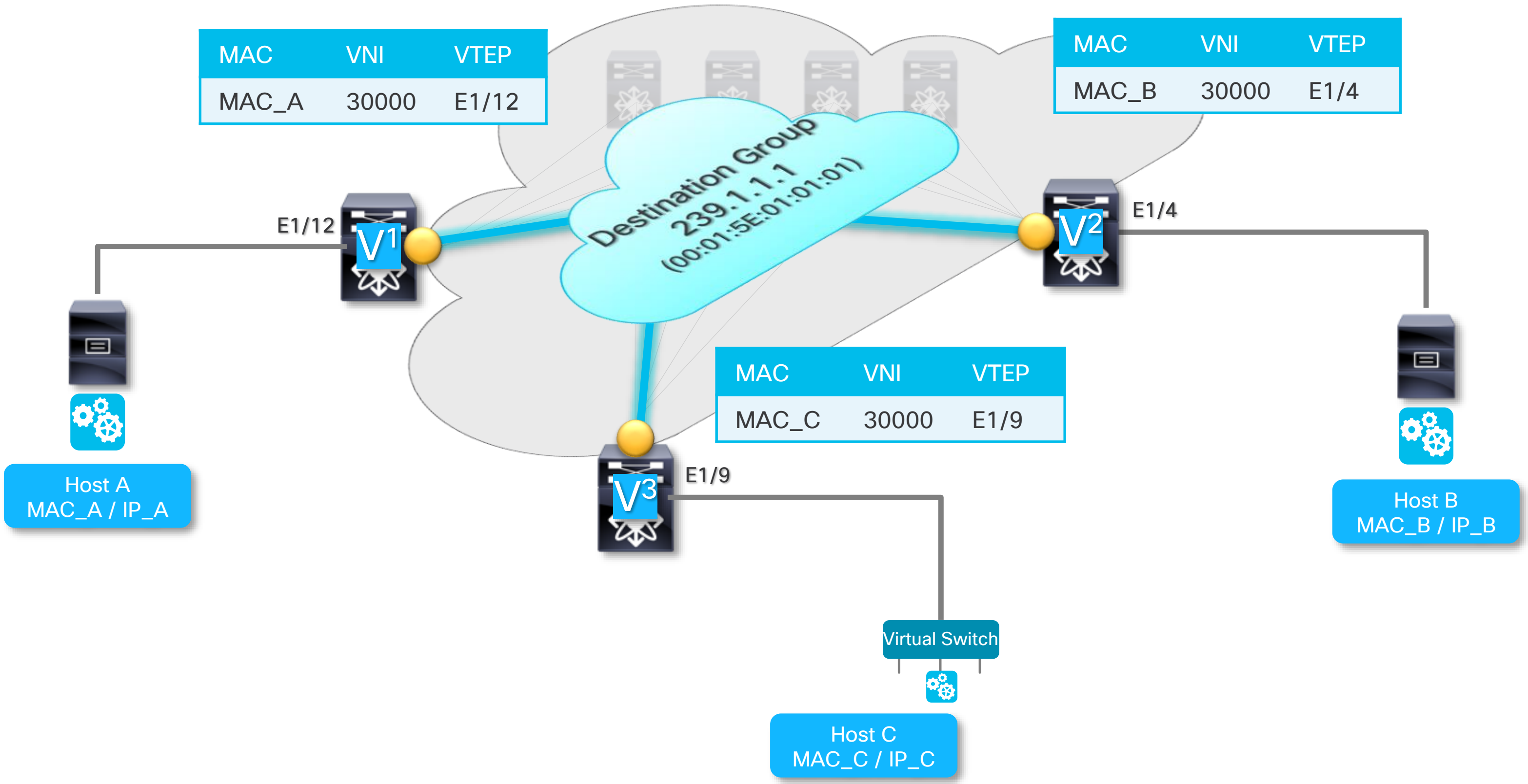
VXLAN Overview (2)



VTEP – VXLAN Tunnel End-Point

VNI/VNID – VXLAN Network Identifier

VXLAN Flood & Learn



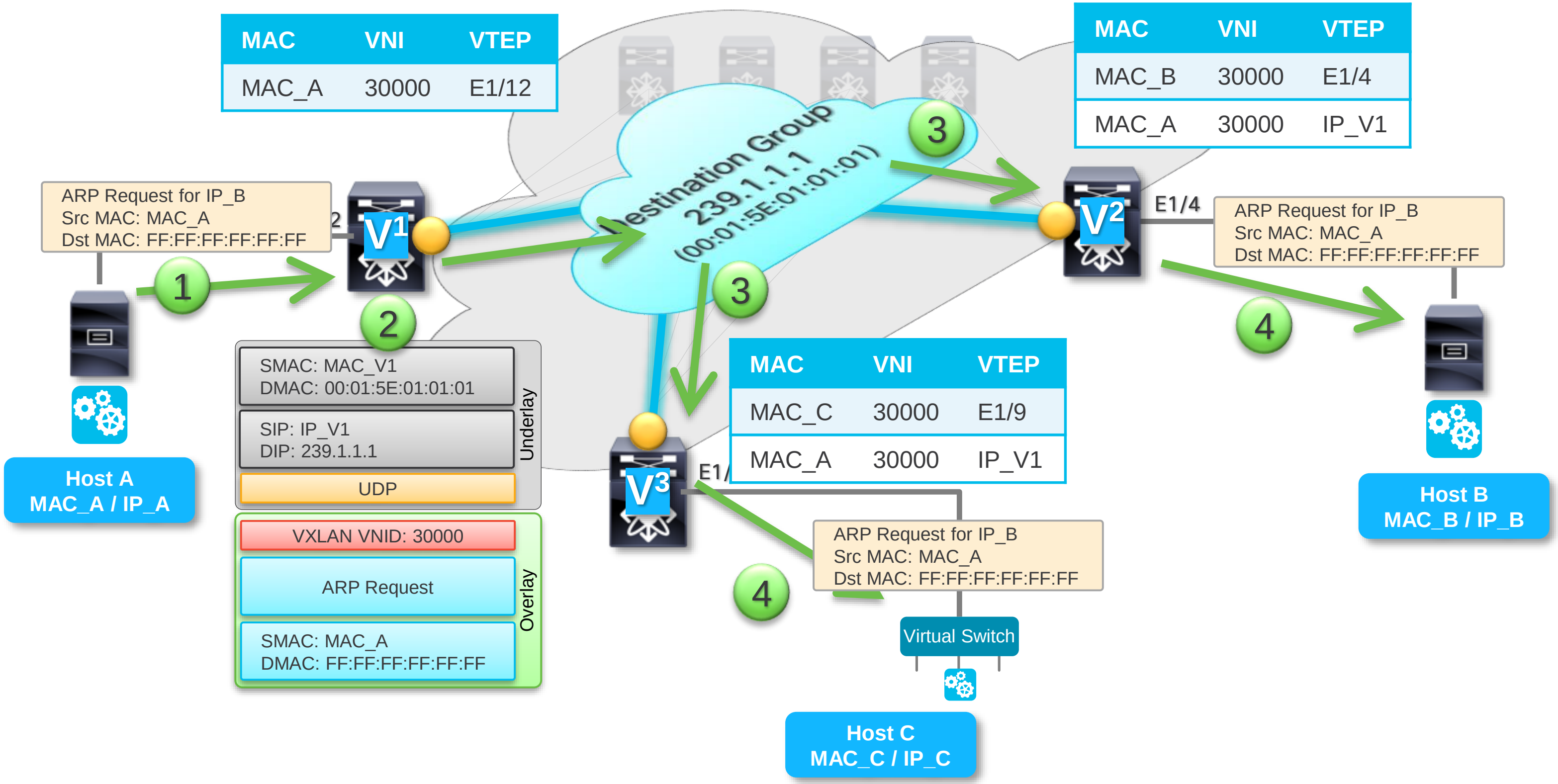
MAC	VNI	VTEP
MAC_A	30000	E1/12

MAC	VNI	VTEP
MAC_B	30000	E1/4

MAC	VNI	VTEP
MAC_C	30000	E1/9

VTEP Peer Discovery & Address Learning (1)

VXLAN Flood & Learn



MAC	VNI	VTEP
MAC_A	30000	E1/12

MAC	VNI	VTEP
MAC_B	30000	E1/4
MAC_A	30000	IP_V1

MAC	VNI	VTEP
MAC_C	30000	E1/9
MAC_A	30000	IP_V1

ARP Request for IP_B
Src MAC: MAC_A
Dst MAC: FF:FF:FF:FF:FF:FF

Underlay
SMAC: MAC_V1
DMAC: 00:01:5E:01:01:01
SIP: IP_V1
DIP: 239.1.1.1
UDP

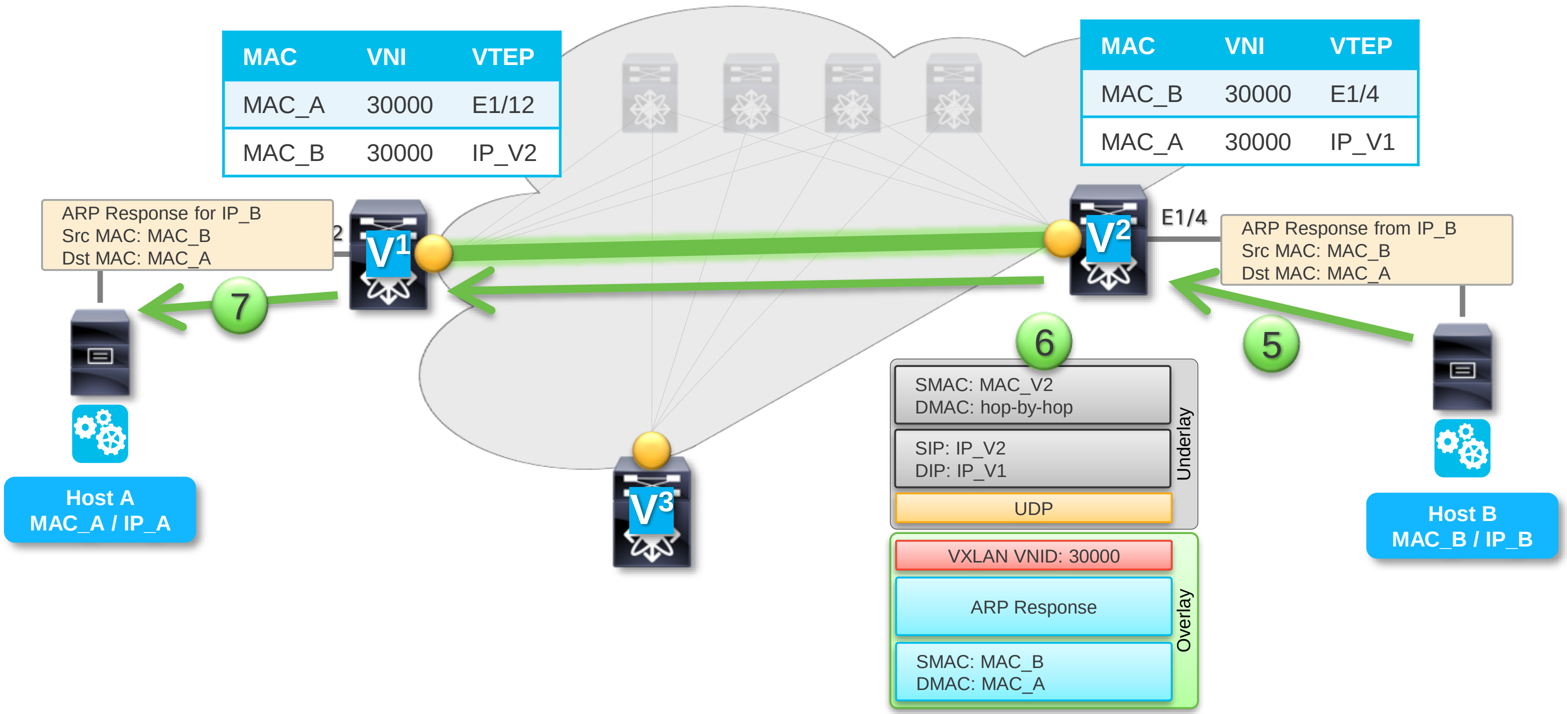
Overlay
VXLAN VNID: 30000
ARP Request
SMAC: MAC_A
DMAC: FF:FF:FF:FF:FF:FF

ARP Request for IP_B
Src MAC: MAC_A
Dst MAC: FF:FF:FF:FF:FF:FF

ARP Request for IP_B
Src MAC: MAC_A
Dst MAC: FF:FF:FF:FF:FF:FF

VTEP Peer Discovery & Address Learning (2)

VXLAN Flood & Learn



MAC	VNI	VTEP
MAC_A	30000	E1/12
MAC_B	30000	IP_V2

MAC	VNI	VTEP
MAC_B	30000	E1/4
MAC_A	30000	IP_V1

ARP Response for IP_B
Src MAC: MAC_B
Dst MAC: MAC_A

ARP Response from IP_B
Src MAC: MAC_B
Dst MAC: MAC_A

SMAC: MAC_V2
DMAC: hop-by-hop

SIP: IP_V2
DIP: IP_V1

UDP

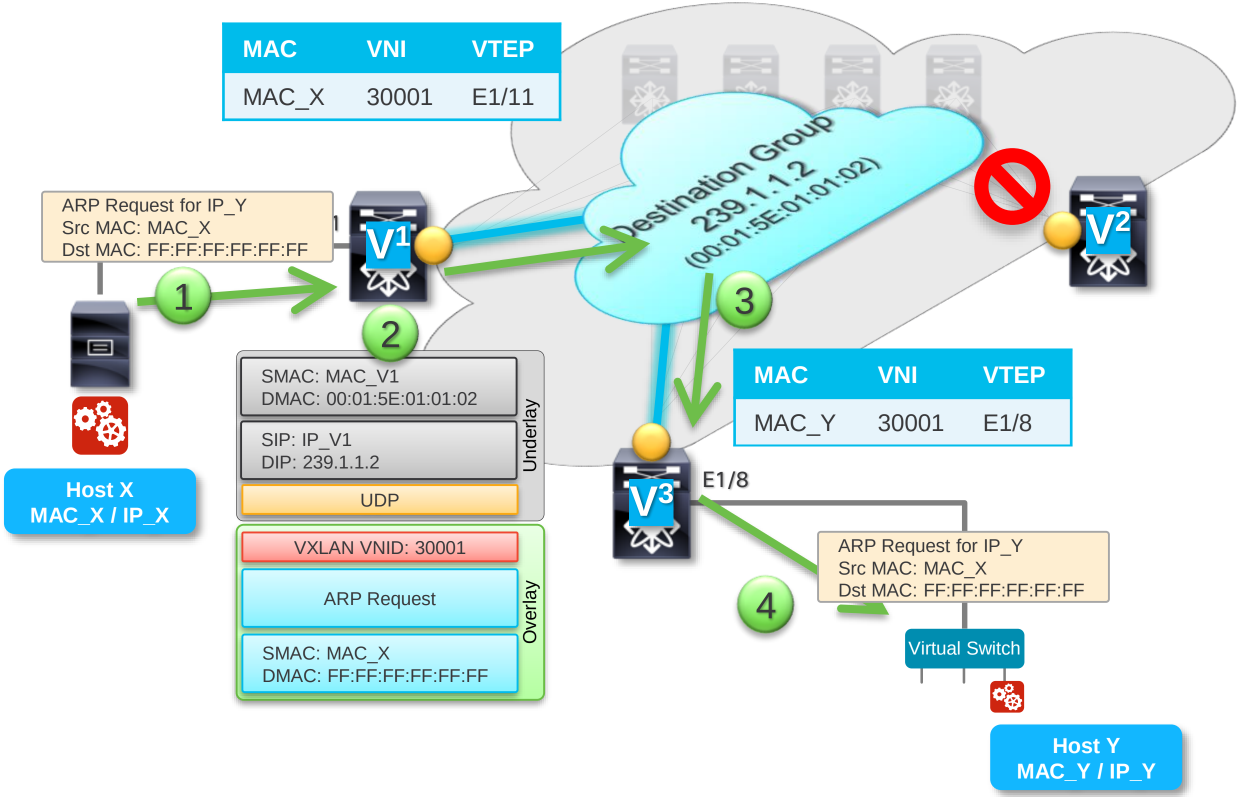
VXLAN VNID: 30000

ARP Response

SMAC: MAC_B
DMAC: MAC_A

VTEP Peer Discovery & Address Learning (3)

VXLAN Flood & Learn



MAC	VNI	VTEP
MAC_X	30001	E1/11

ARP Request for IP_Y
Src MAC: MAC_X
Dst MAC: FF:FF:FF:FF:FF:FF

SMAC: MAC_V1
DMAC: 00:01:5E:01:01:02

SIP: IP_V1
DIP: 239.1.1.2

UDP

VXLAN VNID: 30001

ARP Request

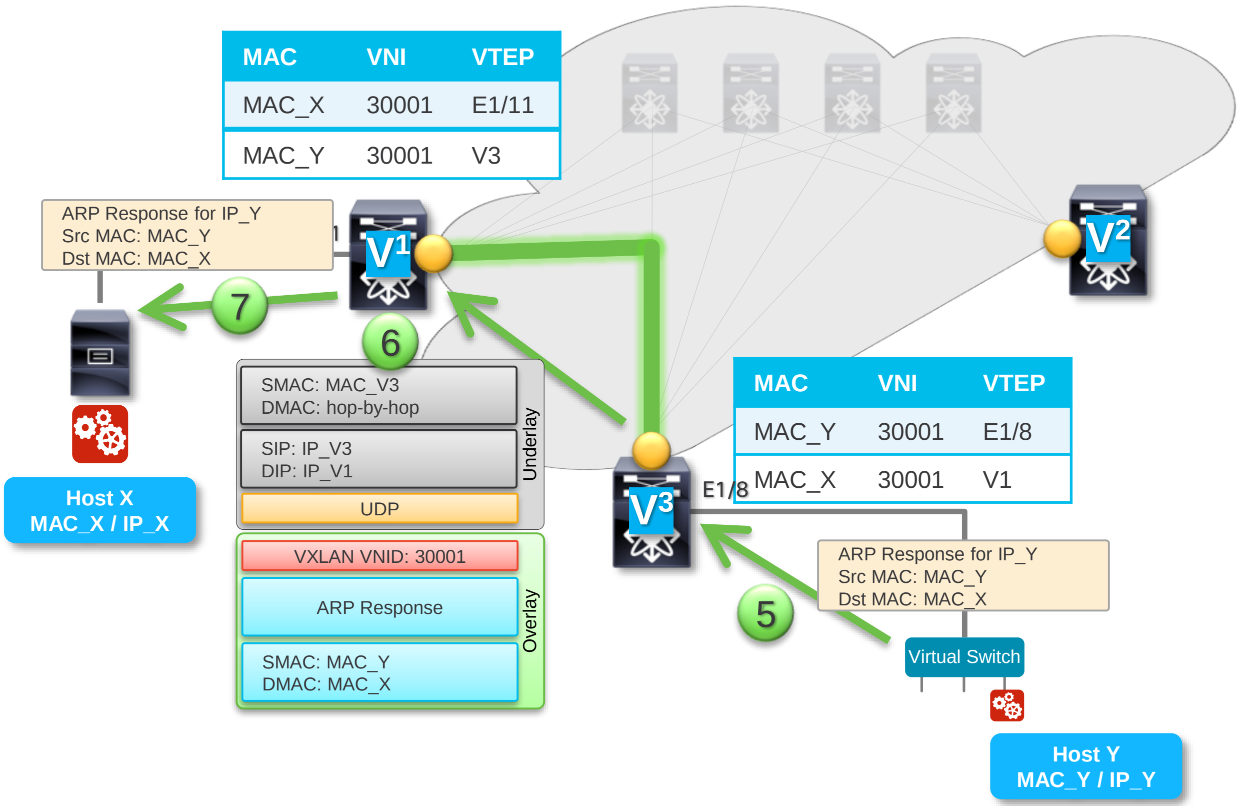
SMAC: MAC_X
DMAC: FF:FF:FF:FF:FF:FF

MAC	VNI	VTEP
MAC_Y	30001	E1/8

ARP Request for IP_Y
Src MAC: MAC_X
Dst MAC: FF:FF:FF:FF:FF:FF

VTEP Peer Discovery & Address Learning (4)

VXLAN Flood & Learn



MAC	VNI	VTEP
MAC_X	30001	E1/11
MAC_Y	30001	V3

MAC	VNI	VTEP
MAC_Y	30001	E1/8
MAC_X	30001	V1

ARP Response for IP_Y
Src MAC: MAC_Y
Dst MAC: MAC_X

SMAC: MAC_V3
DMAC: hop-by-hop

SIP: IP_V3
DIP: IP_V1

UDP

VXLAN VNID: 30001

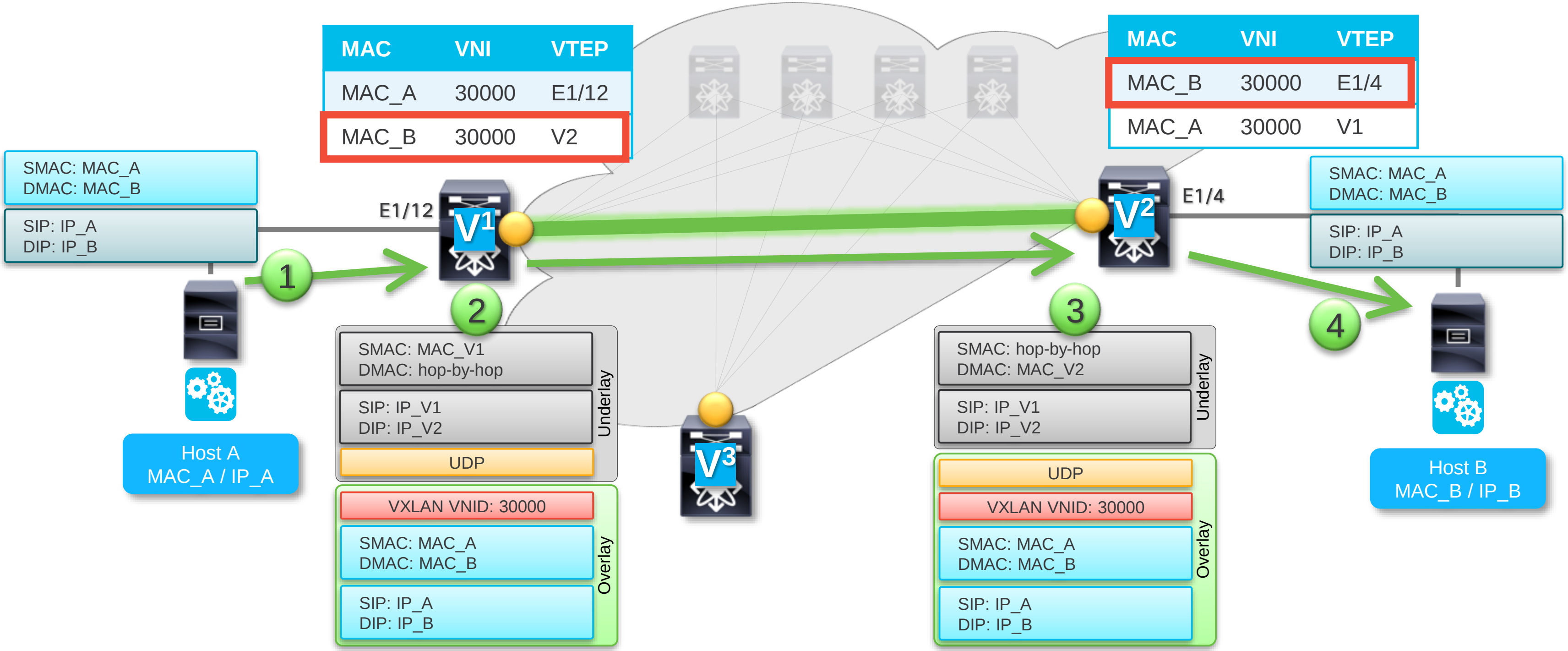
ARP Response

SMAC: MAC_Y
DMAC: MAC_X

ARP Response for IP_Y
Src MAC: MAC_Y
Dst MAC: MAC_X

VXLAN Packet Forwarding (1)

VXLAN Flood & Learn



MAC	VNI	VTEP
MAC_A	30000	E1/12
MAC_B	30000	V2

MAC	VNI	VTEP
MAC_B	30000	E1/4
MAC_A	30000	V1

SMAC: MAC_A
DMAC: MAC_B

SIP: IP_A
DIP: IP_B

Host A
MAC_A / IP_A

SMAC: MAC_V1
DMAC: hop-by-hop

SIP: IP_V1
DIP: IP_V2

UDP

VXLAN VNID: 30000

SMAC: MAC_A
DMAC: MAC_B

SIP: IP_A
DIP: IP_B

SMAC: hop-by-hop
DMAC: MAC_V2

SIP: IP_V1
DIP: IP_V2

UDP

VXLAN VNID: 30000

SMAC: MAC_A
DMAC: MAC_B

SIP: IP_A
DIP: IP_B

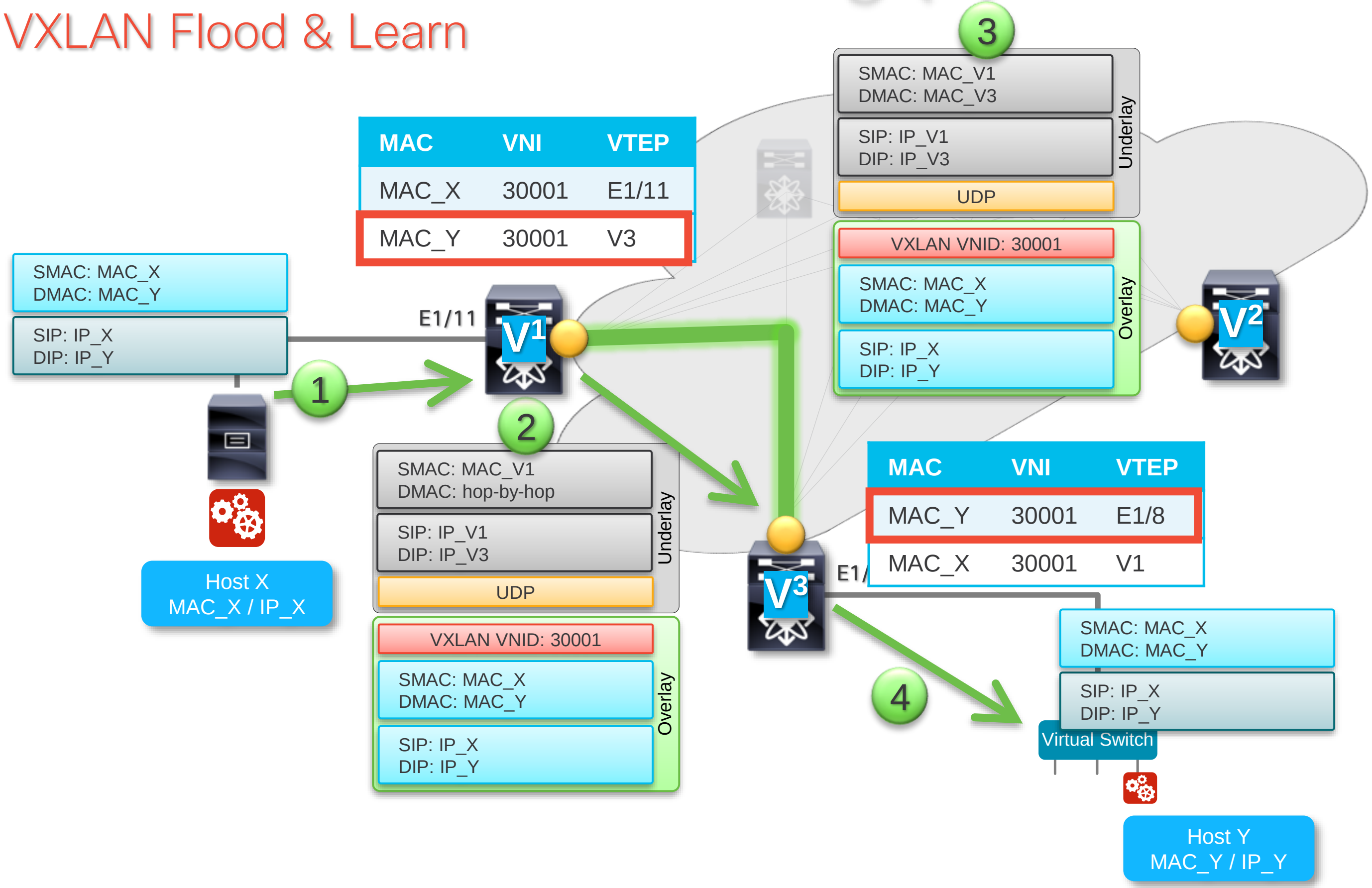
SMAC: MAC_A
DMAC: MAC_B

SIP: IP_A
DIP: IP_B

Host B
MAC_B / IP_B

VXLAN Packet Forwarding (2)

VXLAN Flood & Learn



MAC	VNI	VTEP
MAC_X	30001	E1/11
MAC_Y	30001	V3

SMAC: MAC_X
DMAC: MAC_Y

SIP: IP_X
DIP: IP_Y

SMAC: MAC_V1
DMAC: hop-by-hop

SIP: IP_V1
DIP: IP_V3

UDP

VXLAN VNID: 30001

SMAC: MAC_X
DMAC: MAC_Y

SIP: IP_X
DIP: IP_Y

SMAC: MAC_V1
DMAC: MAC_V3

SIP: IP_V1
DIP: IP_V3

UDP

VXLAN VNID: 30001

SMAC: MAC_X
DMAC: MAC_Y

SIP: IP_X
DIP: IP_Y

MAC	VNI	VTEP
MAC_Y	30001	E1/8
MAC_X	30001	V1

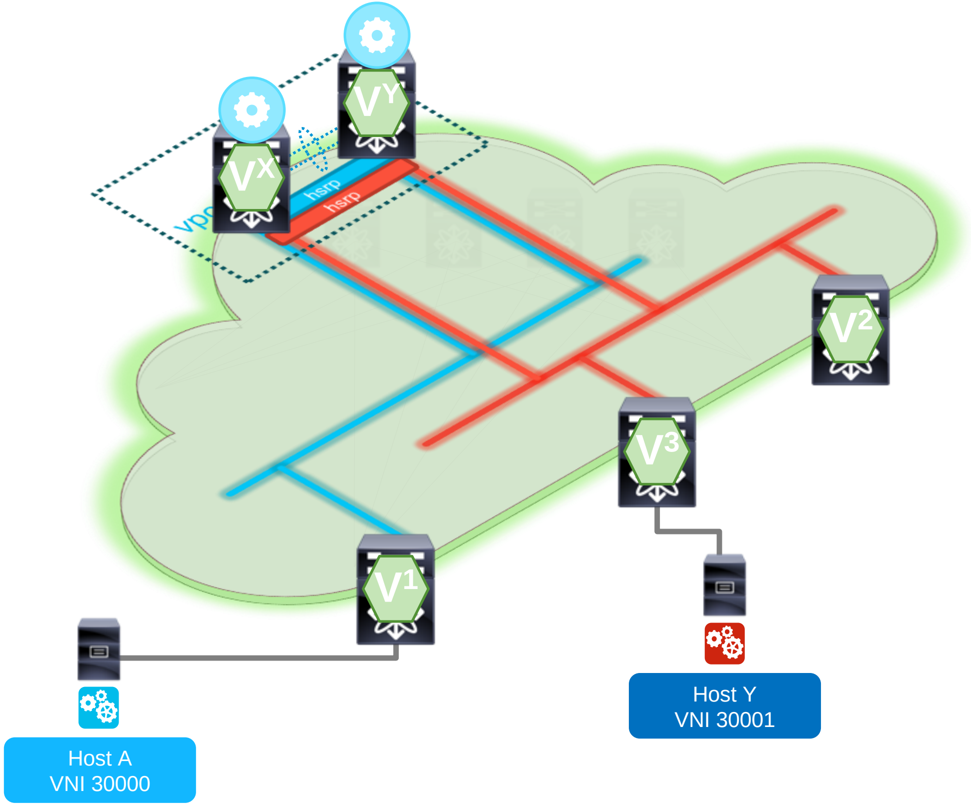
SMAC: MAC_X
DMAC: MAC_Y

SIP: IP_X
DIP: IP_Y

Centralized Gateway (FHRP)

VXLAN Routing

- Centralized Routing in a Layer-2 VXLAN Network
  - Routing between VNI (Different Subnet)
  - Bridging within VNI (Same Subnet)
- Inter-VXLAN Routing at Core/Aggregation Layer
- vPC provides MAC state synchronization and HSRP peering
  - Redundant VTEPs share Anycast VTEP IP address in the Underlay
- Bottleneck for throughput

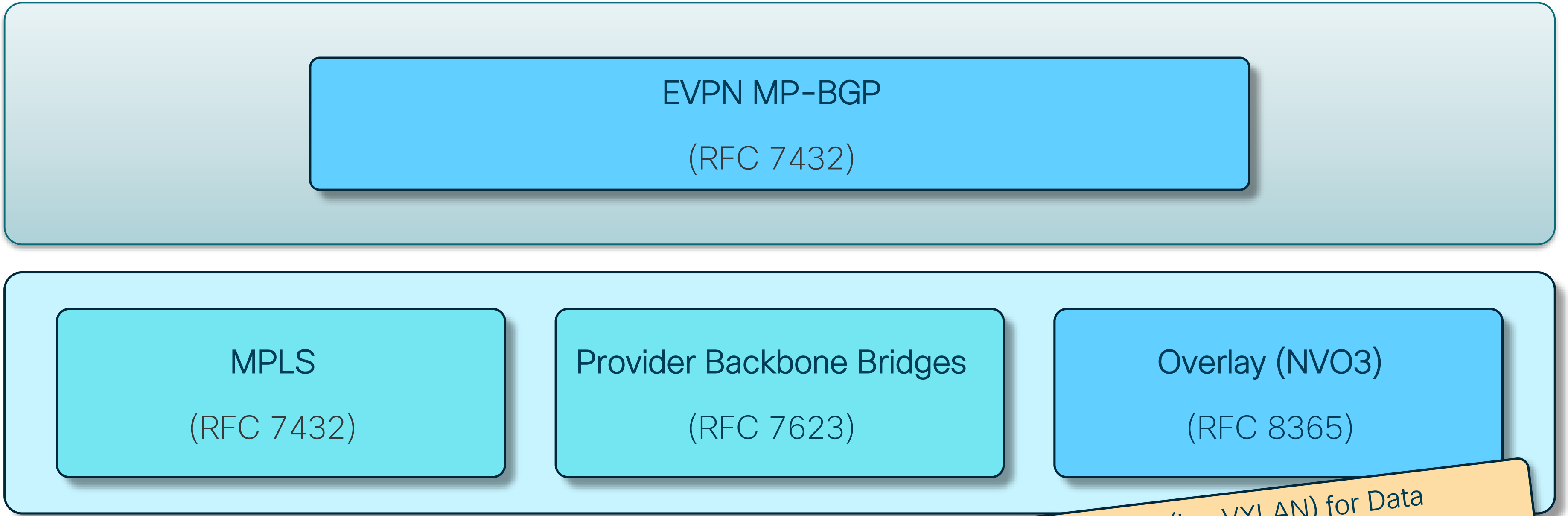


VXLAN Benefits

- Flexible placement of any workload in any rack throughout and between data centers
- Decoupling between physical and virtual networks
- Large Layer 2 network to provide work load mobility
- Centralized Management, provisioning, and automation, from a controller
- Scale, performance, agility and stream lined operations
- Better utilization of available network paths in the underlying infrastructure

VXLAN with BGP EVPN

Ethernet VPN (EVPN)



EVPN MP-BGP

(RFC 7432)

MPLS

(RFC 7432)

Provider Backbone Bridges

(RFC 7623)

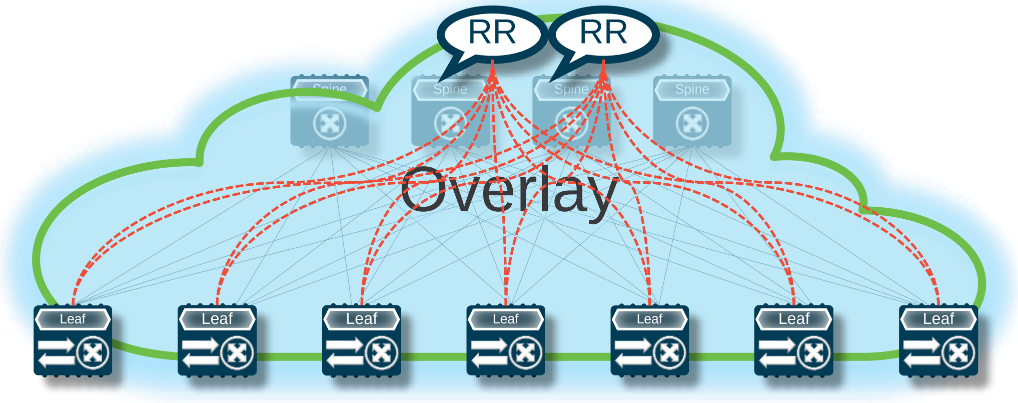
Overlay (NVO3)

(RFC 8365)

- Standards based Control-Plane
  - RFC 7432
  - Uses Multiprotocol BGP

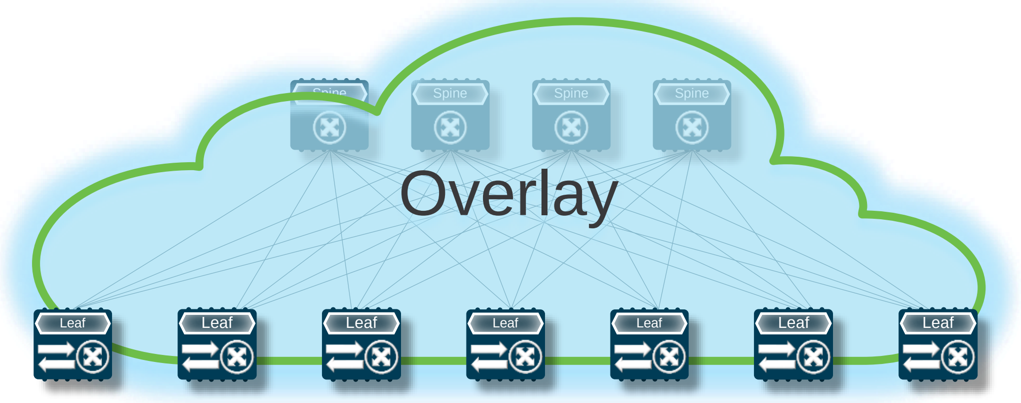
- EVPN over NVO Tunnels (i.e. VXLAN) for Data Center Fabric Encapsulation
- Provides Layer-2 and Layer-3 Overlay Service over simple IP Network

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

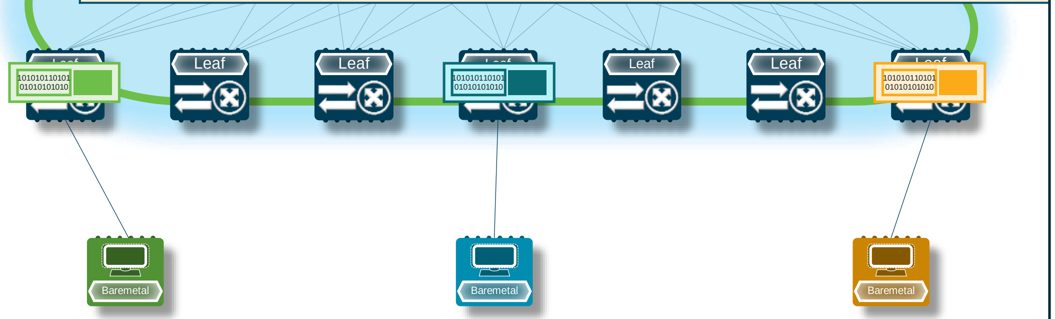


- BGP EVPN NLRI*
- Host MAC (Route Type 2)
  - MAC only, Single VNI, Single Route Target
- Host MAC+IP (Route Type 2)
  - MAC and IP, Two VNI, Two Route Target, Router MAC
- Internal and External Subnet Prefixes (Route Type 5)
  - IP Subnet Prefix, Single VNI, Single Route Target

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

Host Advertisements (L2VNI)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / I
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

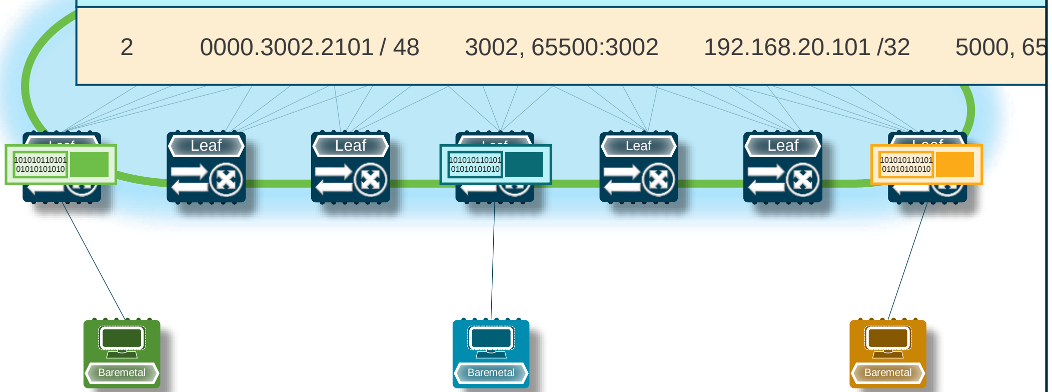
Host C
MAC: 0000.3002.2101

- Host MAC (Route Type 2)
  - MAC
  - MPLS Label1 (L2VNI*)
  - Route Target for MAC-VRF
- MAC attributes are Mandatory

**L2VNI*: *VNI for all Bridging operation ("VLAN-VNI")*

Host Advertisements (L3VNI)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / I
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

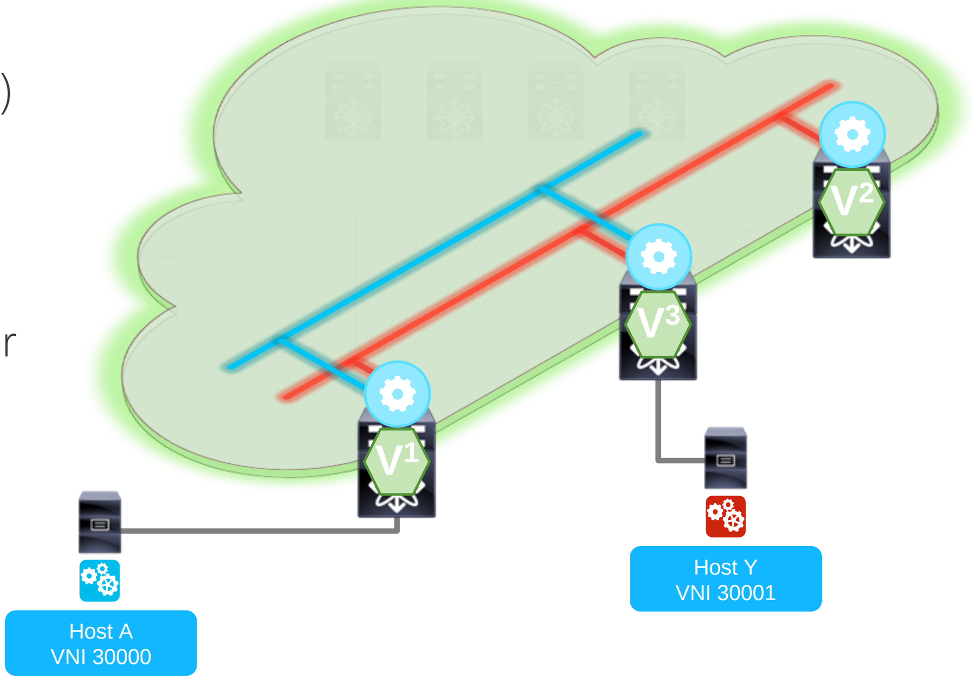
- Host MAC+IP (Route Type 2)
  - MAC and IP
  - MPLS Label1 (L2VNI)
  - Route Target for MAC-VRF
  - MPLS Label2 (L3VNI*)
  - Route Target for IP-VRF
  - Router MAC
- IP Attributes are Optional
- Populated through ARP/ND

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

Distributed IP Anycast Gateway*

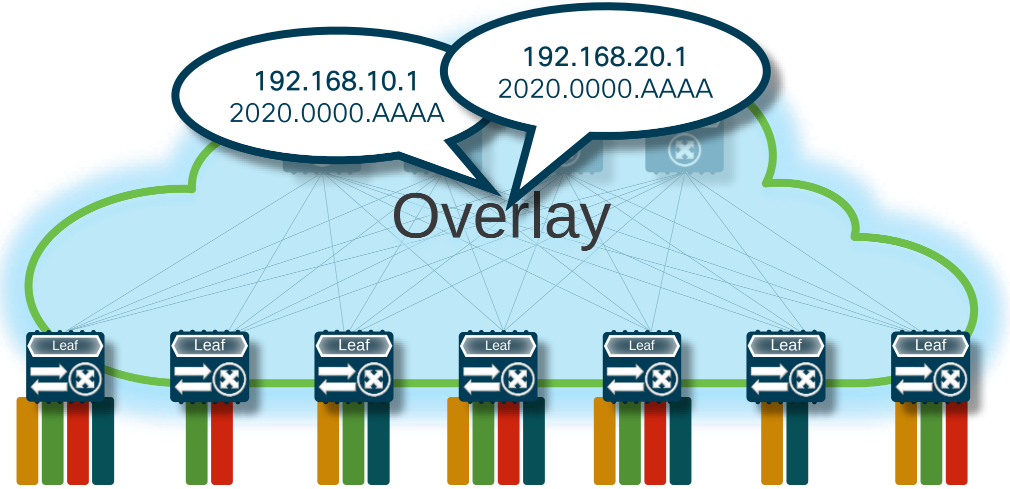
VXLAN/EVPN

- Distributed Routing with IP Anycast Gateway (Integrated Route/Bridge IRB)
  - Routing between VNI (Different Subnet)
  - Bridging within VNI (Same Subnet)
- Inter-VXLAN Routing Leaf/Access Layer
  - All Leafs share gateway IP and MAC for a Subnet (No HSRP)
  - A Host will always find its Gateway directly attached anywhere it moves



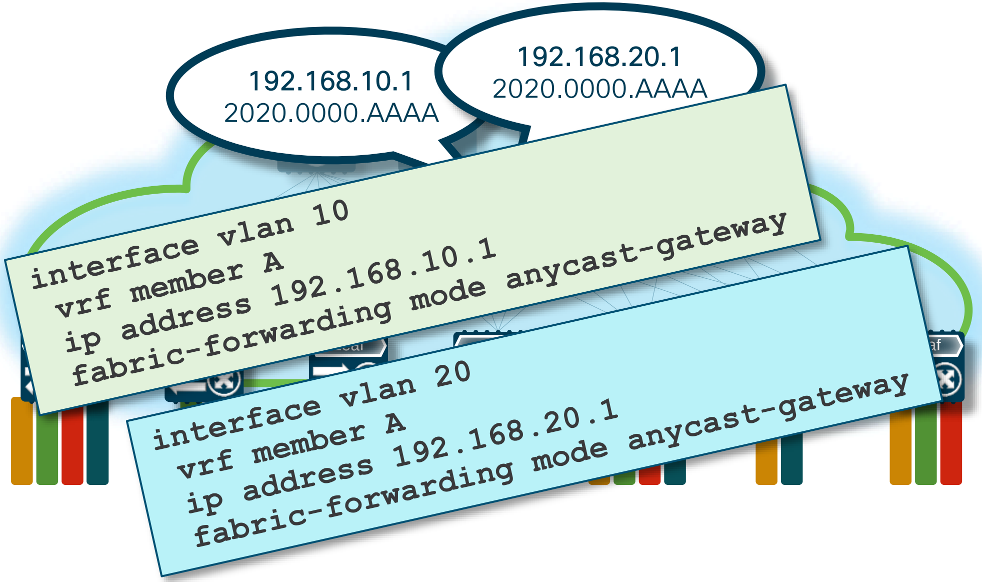
*Requires EVPN Control-Plane.

Distributed IP 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 redundancy protocol for hello or state exchange
- Distributed and smaller state
  - Only local End-Points ARP entries

Distributed IP Anycast Gateway

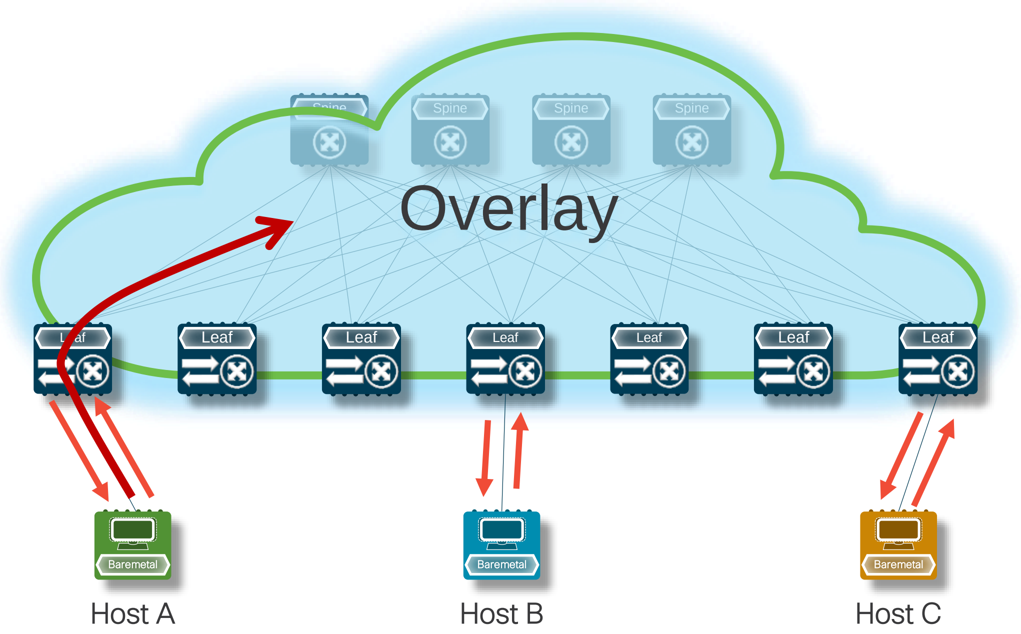


```
interface vlan 10
 vrf member A
  ip address 192.168.10.1
   fabric-forwarding mode anycast-gateway
```

```
interface vlan 20
 vrf member A
  ip address 192.168.20.1
   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 redundancy protocol for hello or state exchange
- Distributed and smaller state
  - Only local End-Points ARP entries

Anycast – One-to-Nearest Association

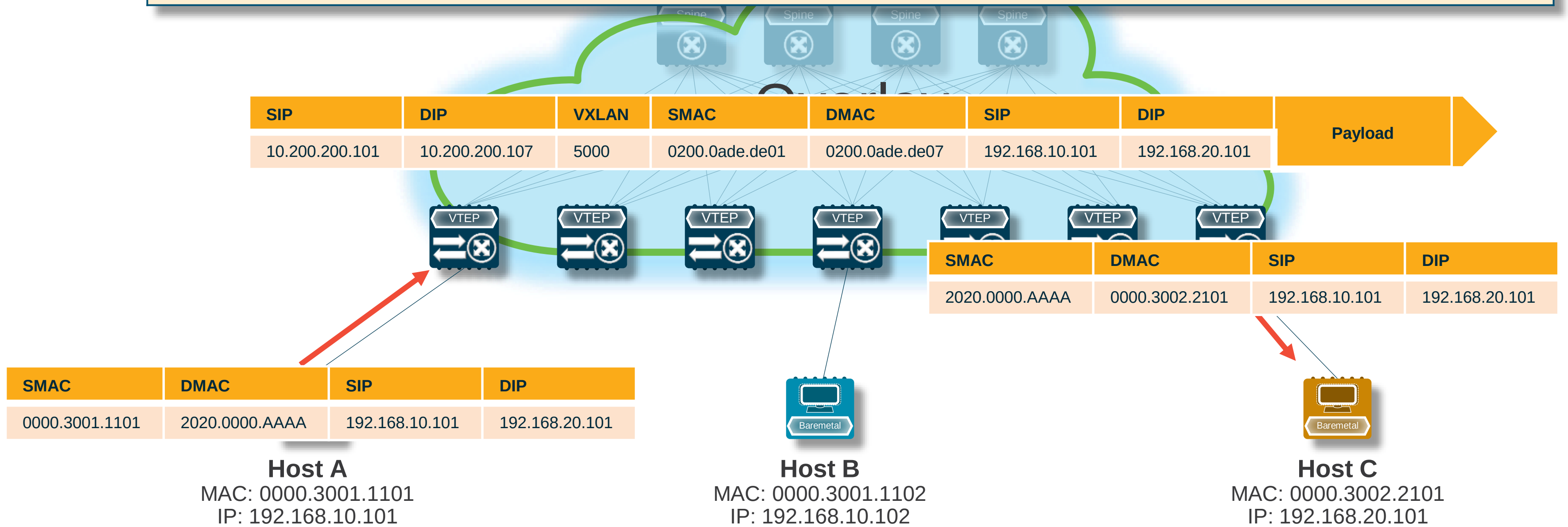


- Network Addressing and Routing Methodology
- Datagrams sent from a single Sender to the Topologically Nearest Node
- Group of potential Receivers, all identified by the same Destination Address

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

VXLAN Routing – Symmetric (A to C)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	3001, 65500:3001	192.168.10.101/32	5000, 65500:5000	10.200.200.101	
2	0000.3002.2102 / 48	3002, 65500:3002	192.168.20.101/32	5000, 65500:5000	10.200.200.107	



SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
10.200.200.101	10.200.200.107	5000	0200.0ade.de01	0200.0ade.de07	192.168.10.101	192.168.20.101	

SMAC	DMAC	SIP	DIP
2020.0000.AAAA	0000.3002.2101	192.168.10.101	192.168.20.101

SMAC	DMAC	SIP	DIP
0000.3001.1101	2020.0000.AAAA	192.168.10.101	192.168.20.101

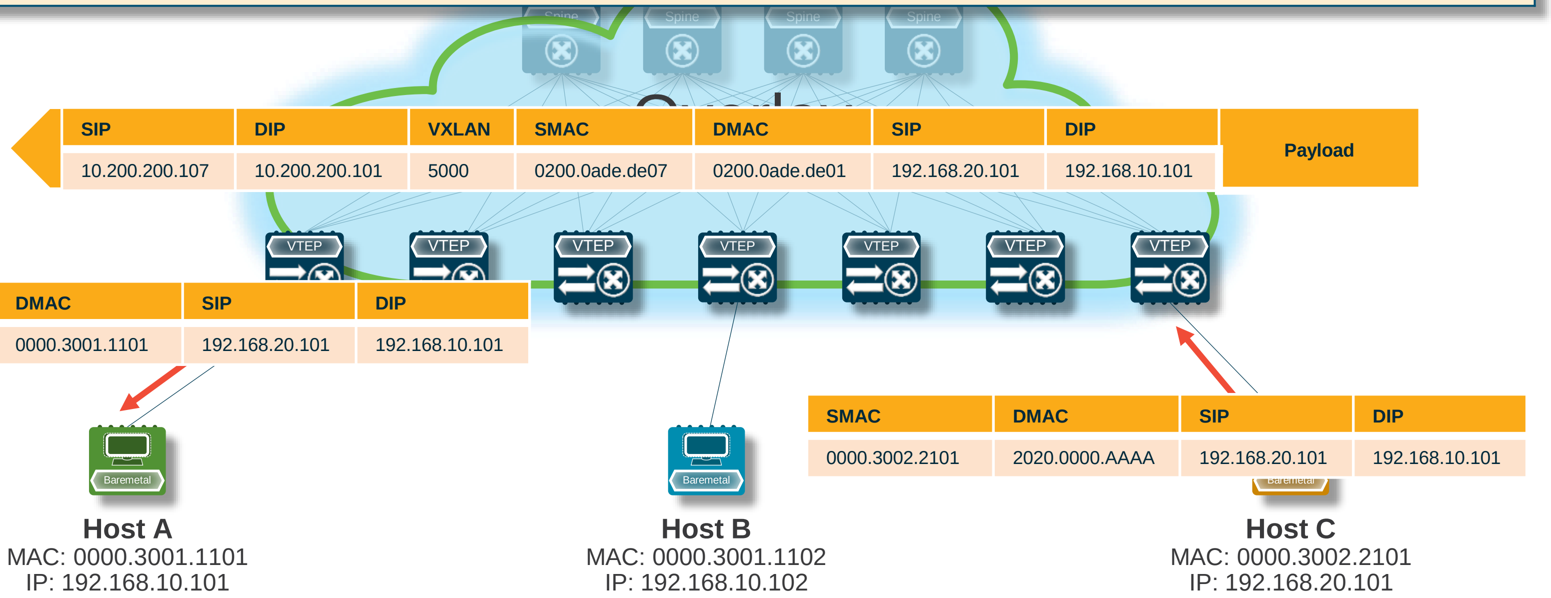
Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

VXLAN Routing – Symmetric (C to A)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	3001, 65500:3001	192.168.10.101/32	5000, 65500:5000	10.200.200.101	
2	0000.3002.2102 / 48	3002, 65500:3002	192.168.20.101/32	5000, 65500:5000	10.200.200.107	



SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
10.200.200.107	10.200.200.101	5000	0200.0ade.de07	0200.0ade.de01	192.168.20.101	192.168.10.101	

SMAC	DMAC	SIP	DIP
2020.0000.AAAA	0000.3001.1101	192.168.20.101	192.168.10.101

SMAC	DMAC	SIP	DIP
0000.3002.2101	2020.0000.AAAA	192.168.20.101	192.168.10.101

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

Summary

Summary

- Overlays (VXLAN) for Network Virtualization
  - Layer-3 in the Underlay – Defines the Topology
  - Layer-2 and Layer-3 in the Overlay – Defines the Services
  - End-Points State exists in the Overlay
- VXLAN evolved as the Control-Plane evolved
  - Applicability changes over time – VXLAN EVPN Multi-Site for DCI
- BGP EVPN for integrated Layer-2 and Layer-3 Services
  - Control-Plane driven
  - Optimal Routing and Bridging
  - Avoid hair pinning and reduced failure domains

cisco