Cisco live!

6-9 March 2018 · Melbourne, Australia

Building Data Centre Networks with VXLAN BPG-EVPN

Lukas Krattiger
Principal Engineer


@CCIE21921

BRKDCN-3378

Cisco live!

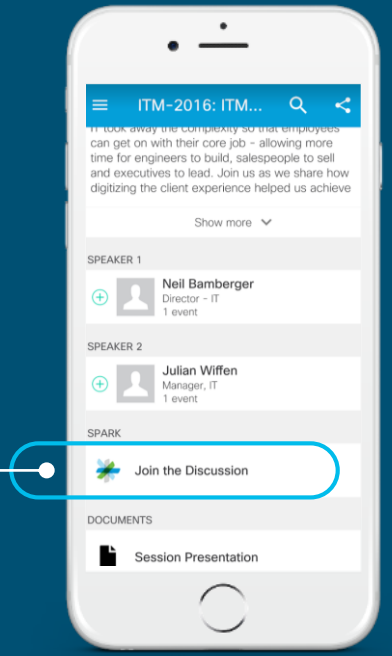
Cisco Spark

Questions?

Use Cisco Spark to communicate with the speaker after the session

How

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



Session Objective

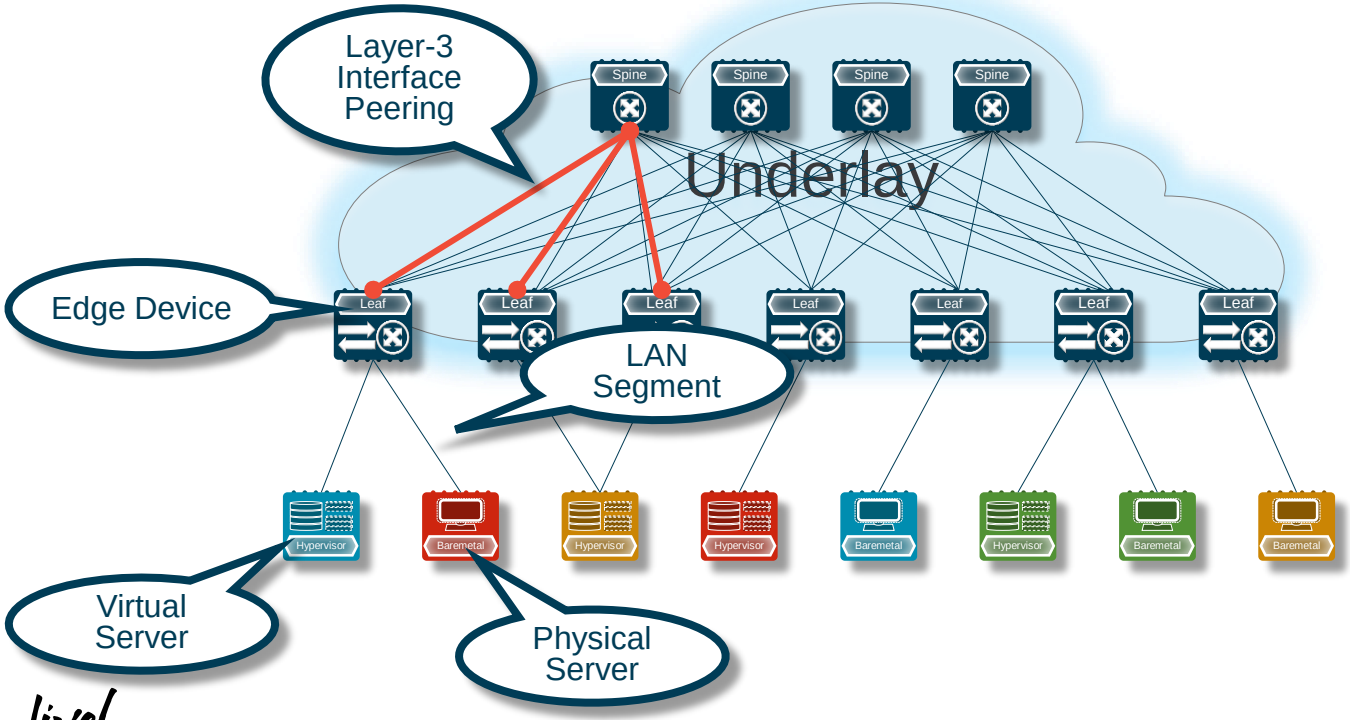
- A short Overview on **Overlays**
- **Standards and Implementation** on VXLAN BGP EVPN
- A walk-thru on **Control- & Data-Plane**
- Details around **Tenant Routed Multicast (TRM)**
- Overview and Details around EVPN **Multi-Site**
- **VXLAN OAM** – Operation, Administration and Management

Agenda

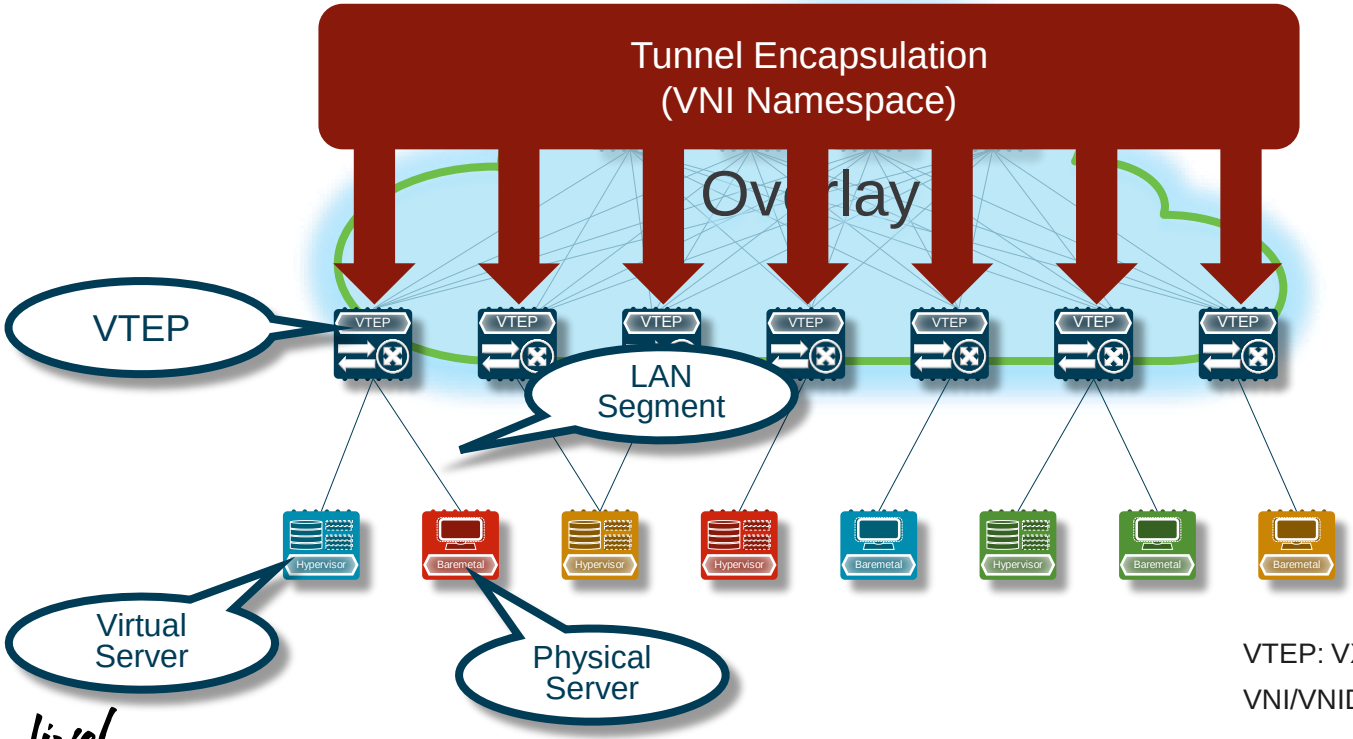
- Introduction to Overlays
- VXLAN with BGP EVPN
  - Standards and Implementation
  - Control & Data Plane
- Tenant Routed Multicast (TRM)
- Multi-Site
- VXLAN OAM

Introduction to Overlays

Overlay Taxonomy - Underlay



Overlay Taxonomy - Overlay



VTEP: VXLAN Tunnel End-Point

VNI/VNID: VXLAN Network Identifier

Understanding 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

Agenda

- Introduction to Overlays
- VXLAN with BGP EVPN
  - Standards and Implementation
  - Control & Data Plane
- Tenant Routed Multicast (TRM)
- Multi-Site
- VXLAN OAM

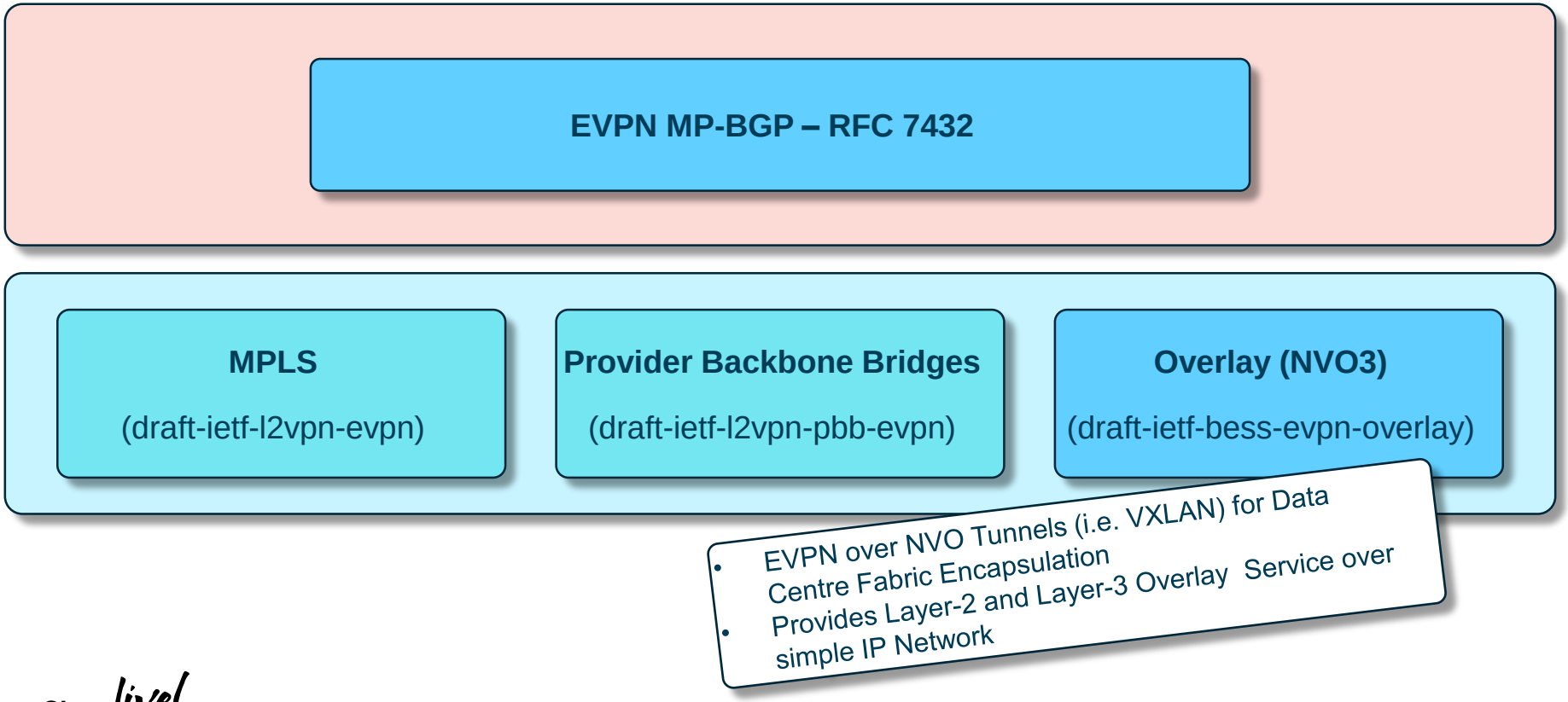
Standards and Implementation

What is … ?

- VXLAN
- Standards based Encapsulation
  - RFC 7348
  - Uses UDP-Encapsulation
- Transport Independent
  - Layer-3 Transport (Underlay)
- Flexible Namespace
  - 24-bit field (VNID) provides ~16M unique identifier
  - Allows Segmentations

- EVPN
- Standards based Control-Plane
  - RFC 7432
  - Uses Multiprotocol BGP
- Uses Various Data-Planes
  - VXLAN (EVPN-Overlay), MPLS, Provider Backbone (PBB)
- Many Use-Cases Covered
  - Bridging, MAC Mobility, First-Hop & Prefix Routing, Multi-Tenancy (VPN)

Introducing Ethernet VPN (EVPN)

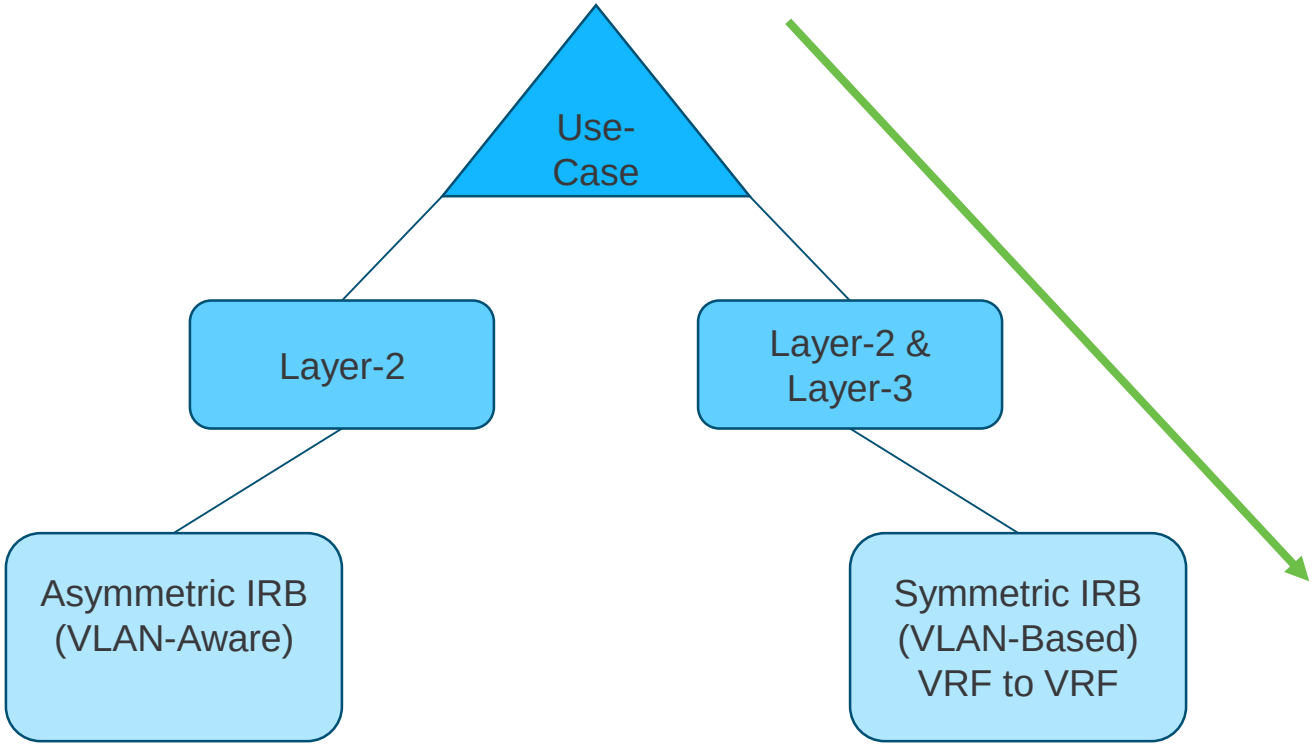


FYI

VXLAN and EVPN related RFCs & Drafts (IETF)

ID	Title	Category
RFC 7348	Virtual Extensible Local Area Network	Data Plane
RFC 7432	BGP MPLS based Ethernet VPNs	Control Plane
draft-ietf-bess-evpn-overlay	A Network Virtualisation Overlay Solution using EVPN	Control Plane
draft-ietf-bess-evpn-inter-subnet-forwarding	Integrated Routing and Bridging in EVPN	Control Plane
draft-ietf-bess-l2vpn-evpn-prefix-advertisement	IP Prefix Advertisement in E-VPN	Control Plane
draft-tissa-nvo3-oam-fm	NVO3 Fault Management / OAM	Management Plane

Different Type of EVPN Use-Cases



FYI

VXLAN and EVPN related RFCs & Drafts (IETF)

ID	Title	Category
RFC 7348	Virtual Extensible Local Area Network	Data Plane
RFC 7432	BGP MPLS based Ethernet VPNs	Control Plane
draft-ietf-bess-evpn-overlay	A Network Virtualisation Overlay Solution using EVPN	Control Plane
draft-ietf-bess-evpn-inter-subnet-forwarding	Integrated Routing and Bridging in EVPN	Control Plane
draft-ietf-bess-l2vpn-evpn-prefix-advertisement	IP Prefix Advertisement in E-VPN	Control Plane
draft-tissa-nvo3-oam-fm	NVO3 Fault Management / OAM	Management Plane

Integrated Routing and Bridging in EVPN

- Symmetric Inter-Subnet Forwarding
  - Bridge->Route/Route->Bridge
  - Symmetric VNI in both directions
  - Adjacency contains Remote VTEP,VRF
    - Optimal for Scale
  - Flexible Configuration

VTEP = VXLAN Tunnel End-Point
VRF = Virtual Routing and Forwarding
VNI = VXLAN Network Identifier

FYI

VXLAN and EVPN related RFCs & Drafts (IETF)

ID	Title	Category
RFC 7348	Virtual Extensible Local Area Network	Data Plane
RFC 7432	BGP MPLS based Ethernet VPNs	Control Plane
draft-ietf-bess-evpn-overlay	A Network Virtualisation Overlay Solution using EVPN	Control Plane
draft-ietf-bess-evpn-inter-subnet-forwarding	Integrated Routing and Bridging in EVPN	Control Plane
draft-ietf-bess-l2vpn-evpn-prefix-advertisement	IP Prefix Advertisement in E-VPN	Control Plane
draft-tissa-nvo3-oam-fm	NVO3 Fault Management / OAM	Management Plane

EVPN Layer-2 Service Interface

- Single Subnet per EVI
  - VLAN-based

- Per EVI BGP Route Distinguisher / Router Target per EVI / VNI
  - BGP Route-Target constrain mechanism to limit propagation (import/export)

- 1:1 mapping
  - EVI to Single Broadcast Domain (Bridge Domain)
- Ethernet Tag ID must be 0

VID = VLAN ID
VNI = VXLAN Network Identifier
EVI = EVPN Virtual Instance

FYI

VXLAN and EVPN related RFCs & Drafts (IETF)

ID	Title	Category
RFC 7348	Virtual Extensible Local Area Network	Data Plane
RFC 7432	BGP MPLS based Ethernet VPNs	Control Plane
draft-ietf-bess-evpn-overlay	A Network Virtualisation Overlay Solution using EVPN	Control Plane
draft-ietf-bess-evpn-inter-subnet-forwarding	Integrated Routing and Bridging in EVPN	Control Plane
draft-ietf-bess-l2vpn-evpn-prefix-advertisement	IP Prefix Advertisement in E-VPN	Control Plane
draft-tissa-nvo3-oam-fm	NVO3 Fault Management / OAM	Management Plane

IP-VRF-to-IP-VRF Model in EVPN

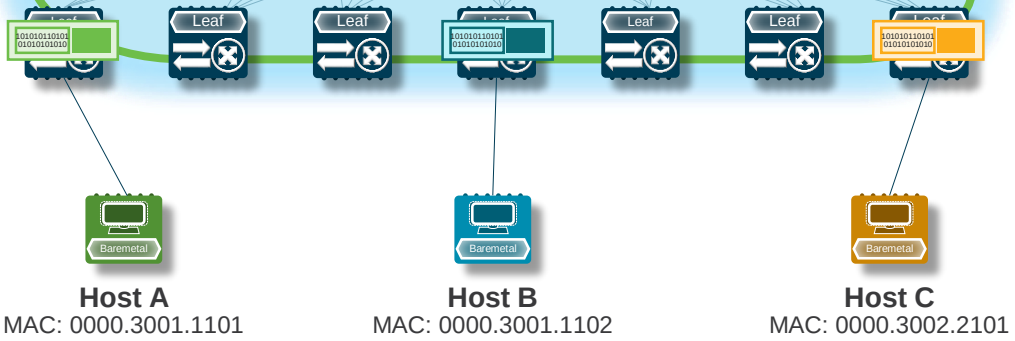
- Interface-Less Model
  - Route-Type 5 only
    - Next-Hop is remote VTEP
    - Two extended communities
      - Encapsulation Extended Community
      - Router's MAC Address (remote VTEP)

Route Type 2 = MAC/IP Route
Route Type 5 = IP Prefix Route

Control- & Data-Plane

Host Advertisements

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



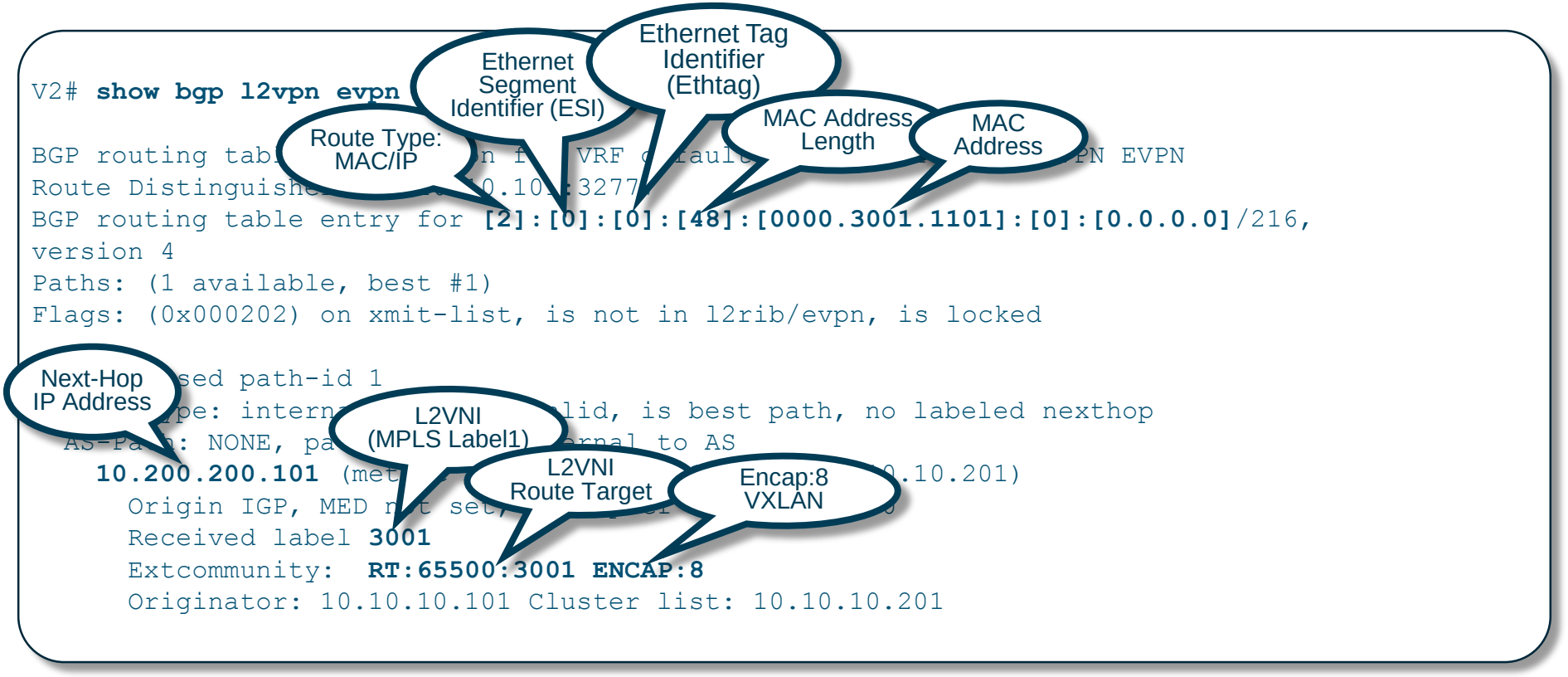
Host A
MAC: 0000.3001.1101

Host B
MAC: 0000.3001.1102

Host C
MAC: 0000.3002.2101

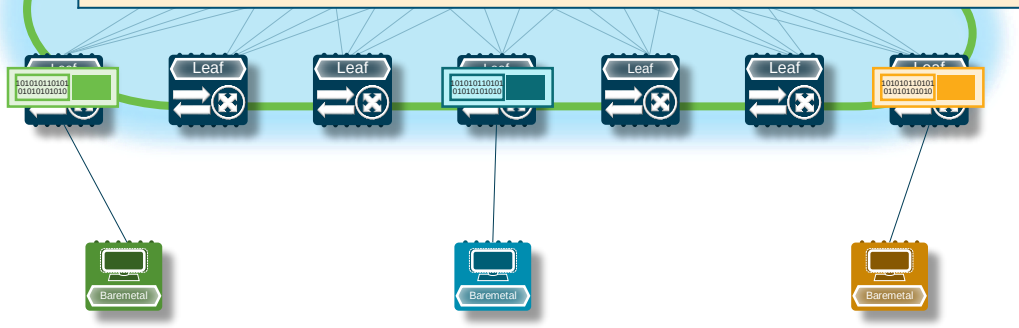
- Host MAC (Route Type 2)
  - MAC
  - MPLS Label1 (L2VNI*)
  - Route Target for MAC-VRF
- MAC attributes are Mandatory

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



Host Advertisements

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT
2	0000.3001.1101 / 48	3001, 65500:3001	192.168.10.101 /32	5000, 655
2	0000.3001.1102 / 48	3001, 65500:3001	192.168.10.102 /32	5000, 655
2	0000.3002.2101 / 48	3002, 65500:3002	192.168.20.101 /32	5000, 655



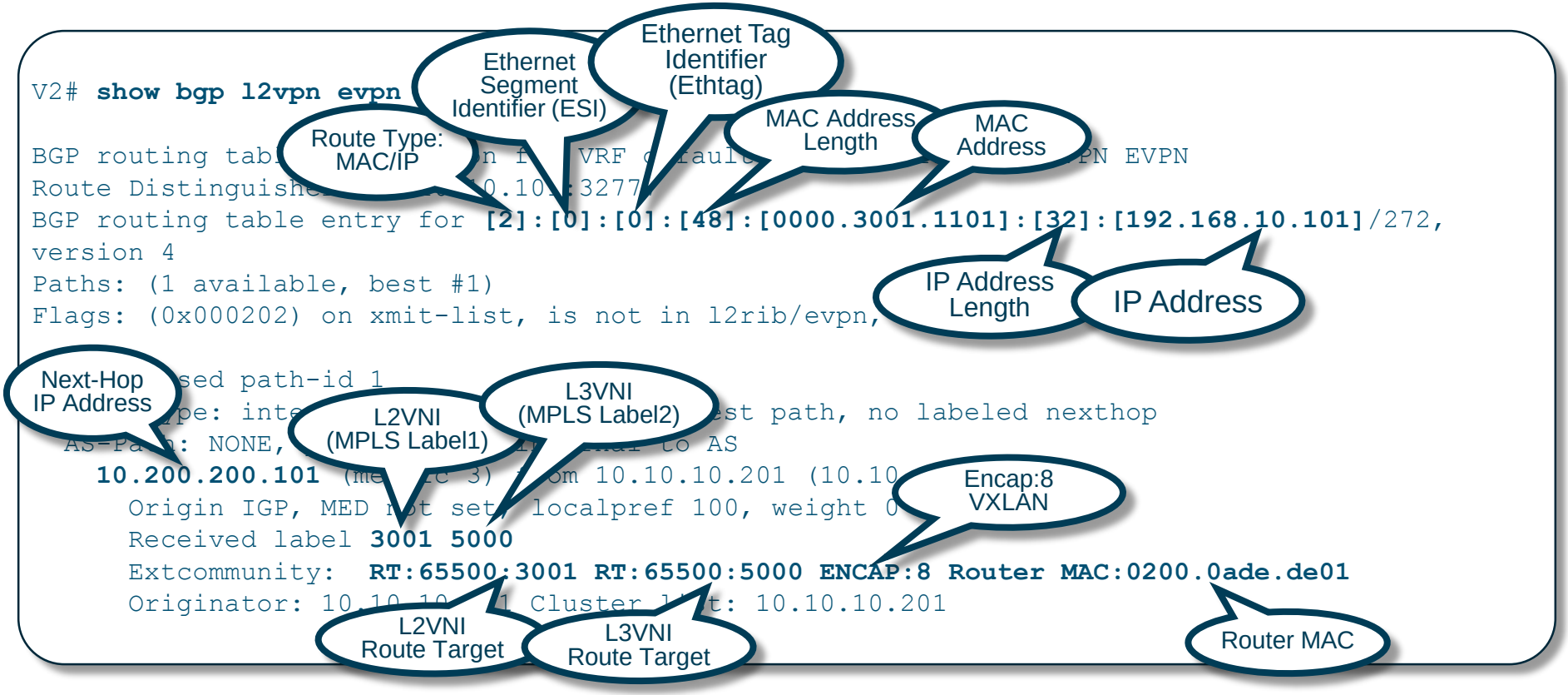
Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

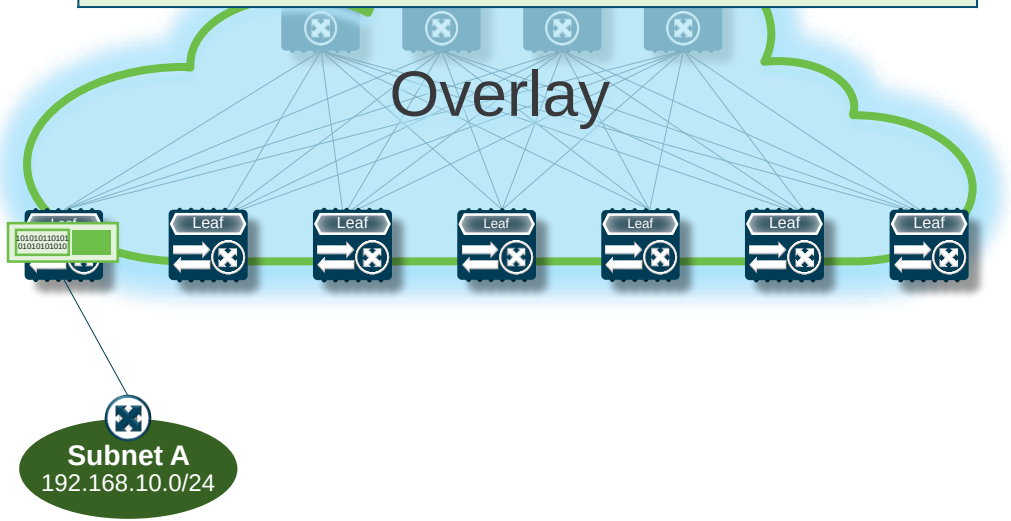
- Host MAC+IP (Route Type 2)
  - MAC and IP
  - MPLS Label1 (L2VNI)
  - Route Target for MAC-VRF
  - MPLS Label2 (L3VNI*)
  - Route Target for IP-VRF
  - Router MAC
- IP Attributes are Optional
- Populated through ARP/ND

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

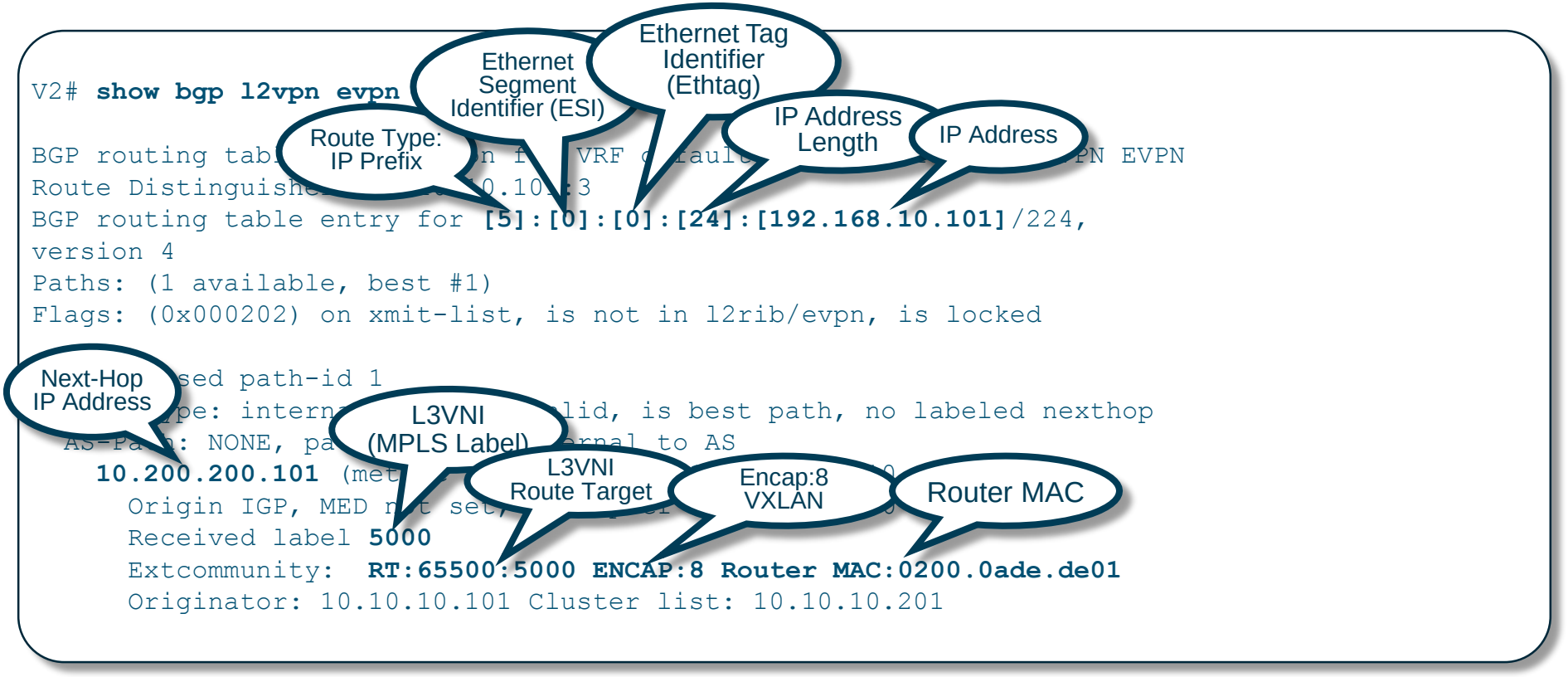


Subnet Route Advertisements

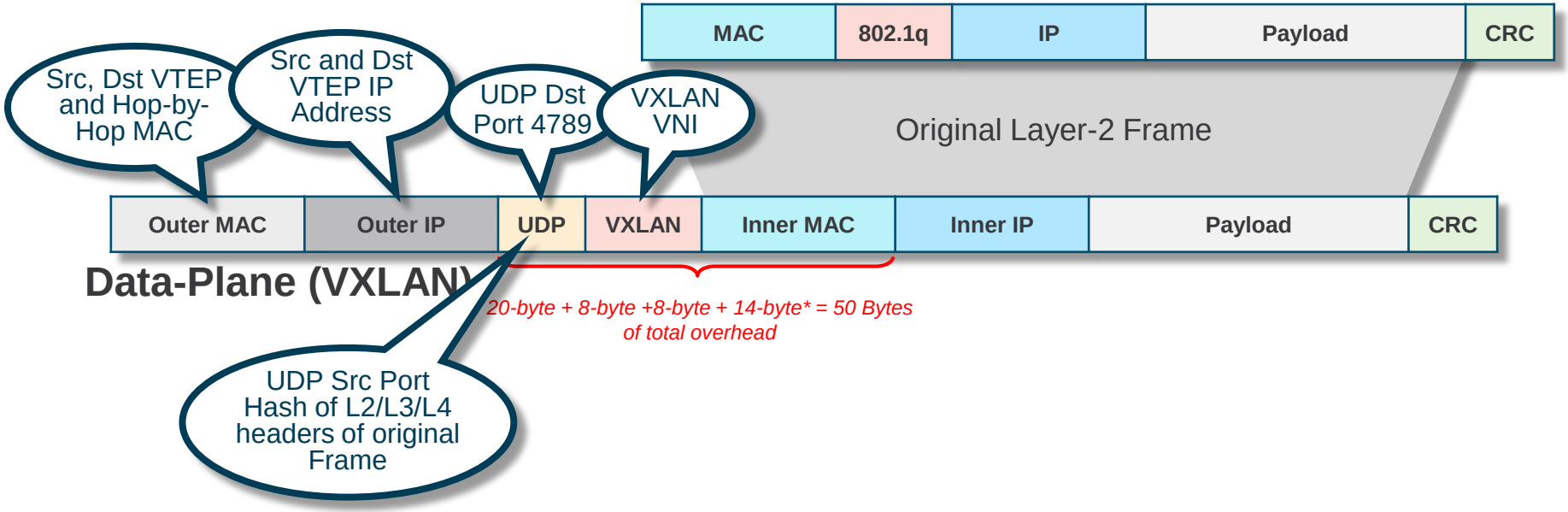
Type	IP / Length	L3VNI / RT	Next-Hop	Seq.
5	192.168.10.0 /24	5000, 65500:5000	10.200.200.101	



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



Introducing VXLAN



*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	

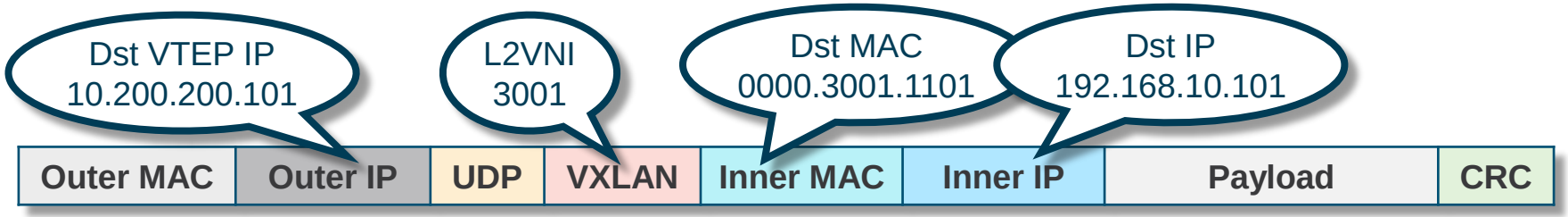
Field	Value	Bites	Total
IP Header	Misc. Data	72	20 Bytes
Protocol	0x11 (UDP)	8	
Header Checksum	Various	16	
Source IP	**Src, VTEP IP**	32	
Destination IP	**Dest. VTEP IP**	32	

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

VXLAN and BGP EVPN – Putting it Together

Control-Plane (BGP EVPN)

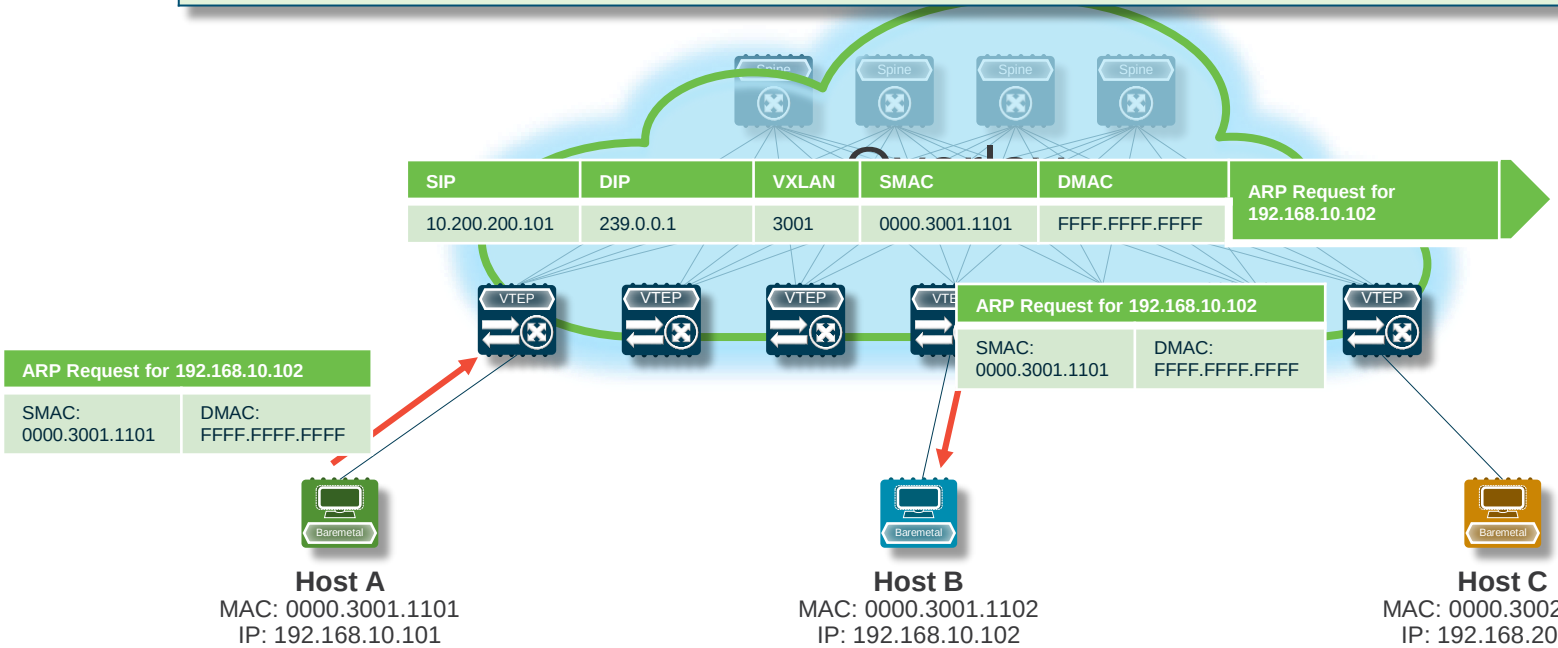
Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101/48	3001 65500:3001	192.168.10.101/32	5000 65500:5000	10.200.200.101	



Data-Plane (VXLAN)

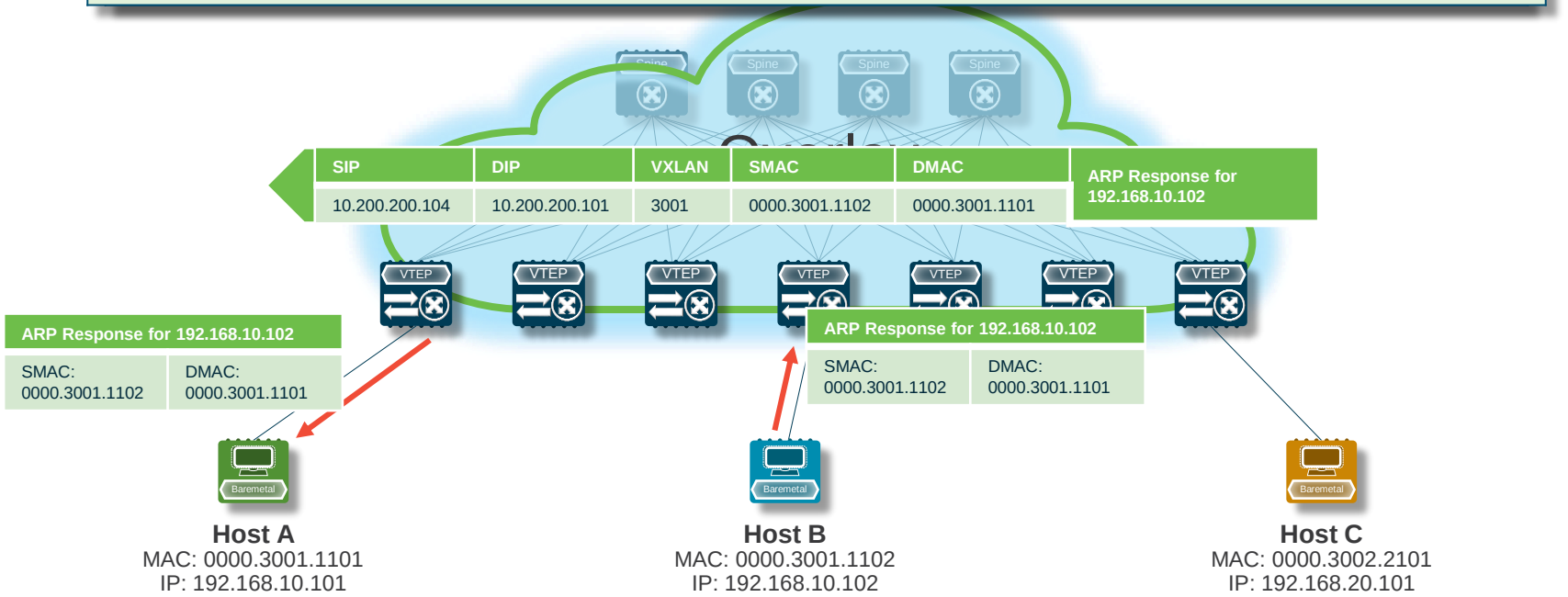
Packet Walk – ARP Request

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	3001, 65500:3001			10.200.200.101	



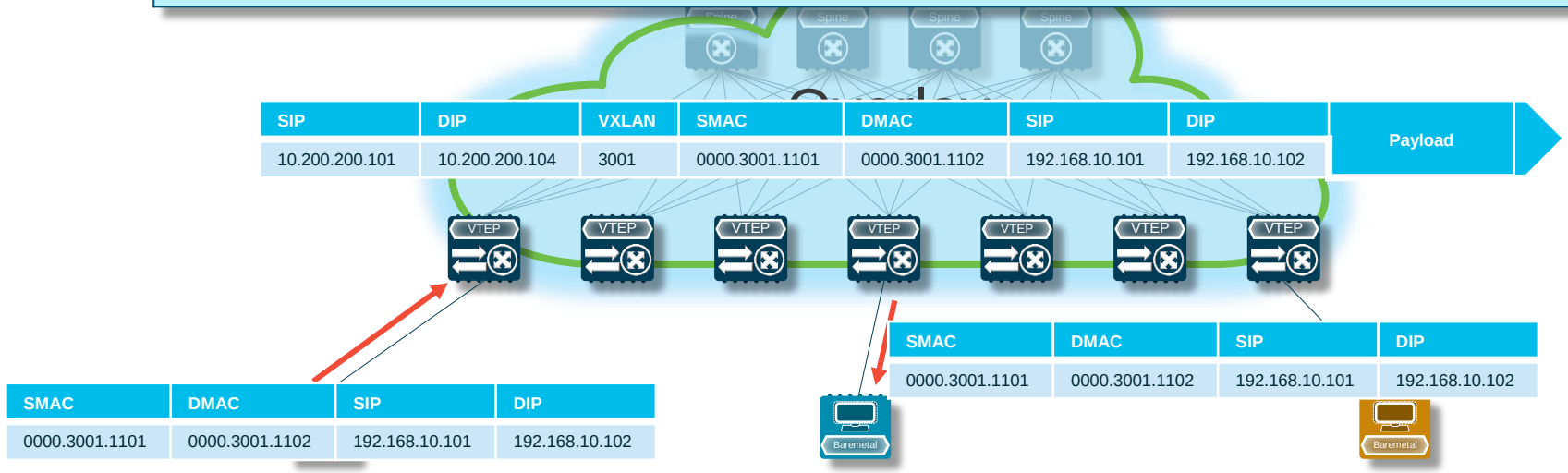
Packet Walk – ARP Response

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	



Packet Walk – Bridging

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	0000.3001.1102 / 48	3001, 65500:3001	192.168.10.102/32	5000, 65500:5000	10.200.200.104	



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

Integrated Routing and Bridging in EVPN

- Symmetric Inter-Subnet Forwarding
  - Bridge->Route/Route->Bridge
  - Symmetric VNI in both directions
  - Adjacency contains Remote VTEP,VRF
    - Optimal for Scale
  - Flexible Configuration

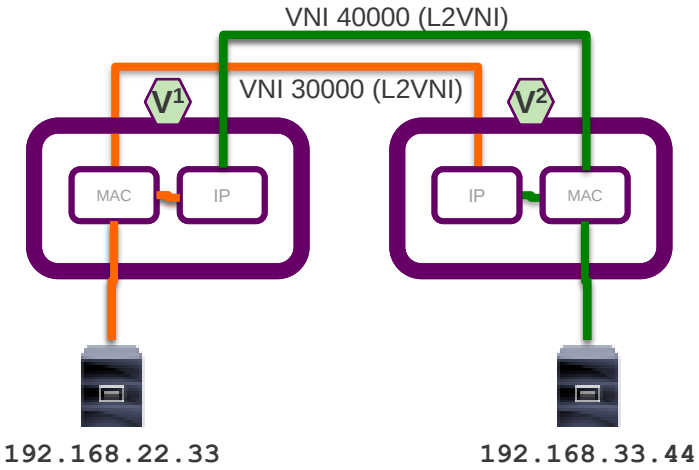
- Asymmetric Inter-Subnet Forwarding
  - Bridge->Route->Bridge
  - Different (Asymmetric) VNI depending on directions
  - Adjacency contains Remote VTEP,VRF and End-Points
    - Potential Sub-Optimal for Scale
  - Consistent Configuration

VTEP = VXLAN Tunnel End-Point
VRF = Virtual Routing and Forwarding
VNI = VXLAN Network Identifier

Operational Models for Asymmetric Inter-Subnet Forwarding

(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 4)

- Asymmetric IRB

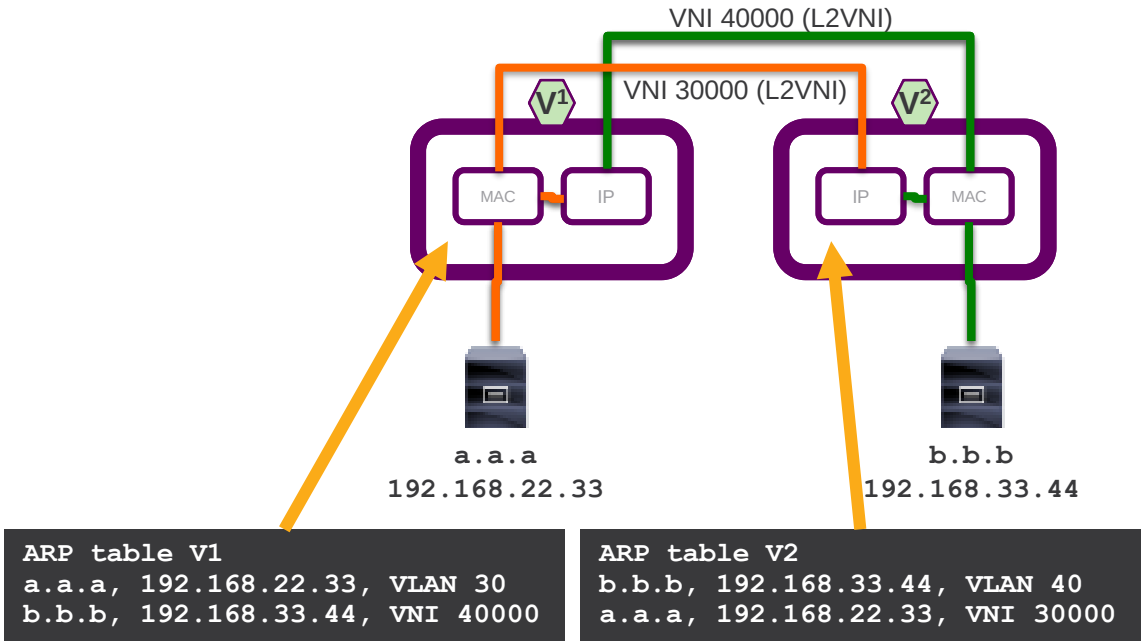


Bridge -> Route -> Bridge

Operational Models for Asymmetric Inter-Subnet Forwarding

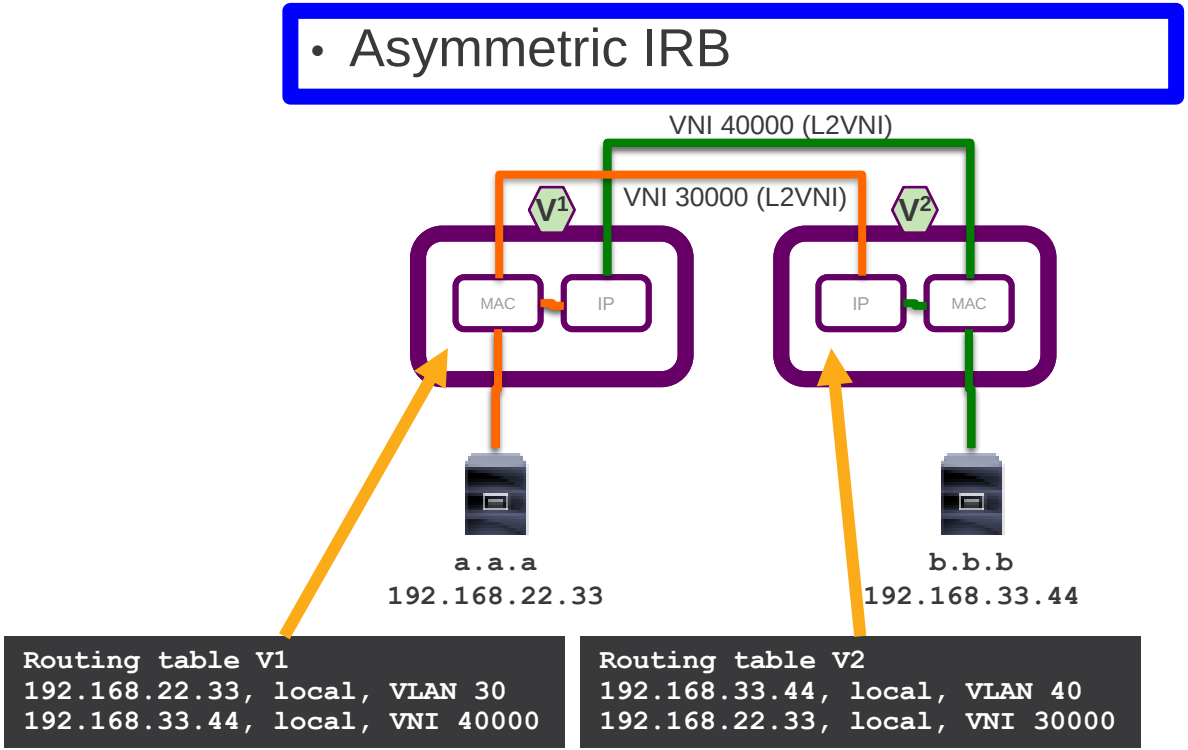
ARP and Adjacency Table

- Asymmetric IRB



Operational Models for Asymmetric Inter-Subnet Forwarding

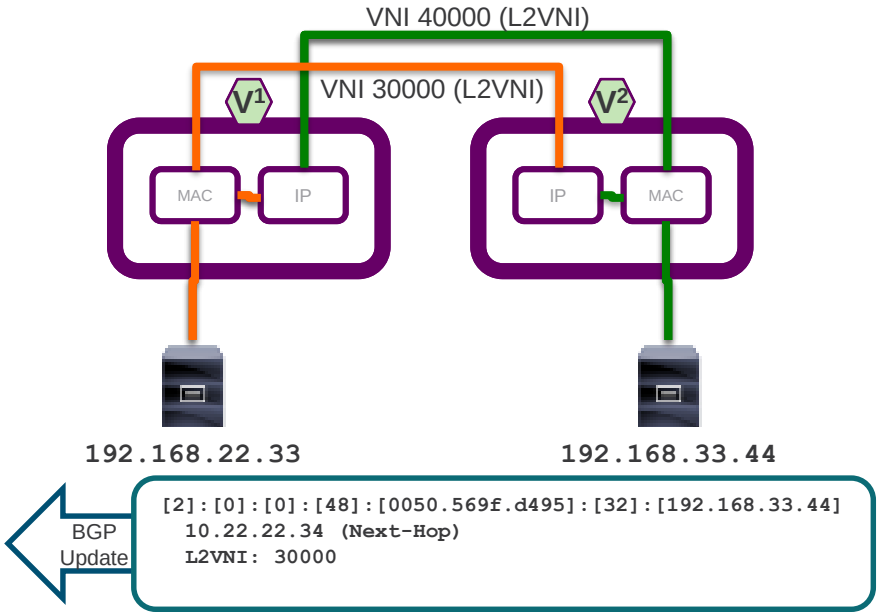
Routing Table



Operational Models for Asymmetric Inter-Subnet Forwarding

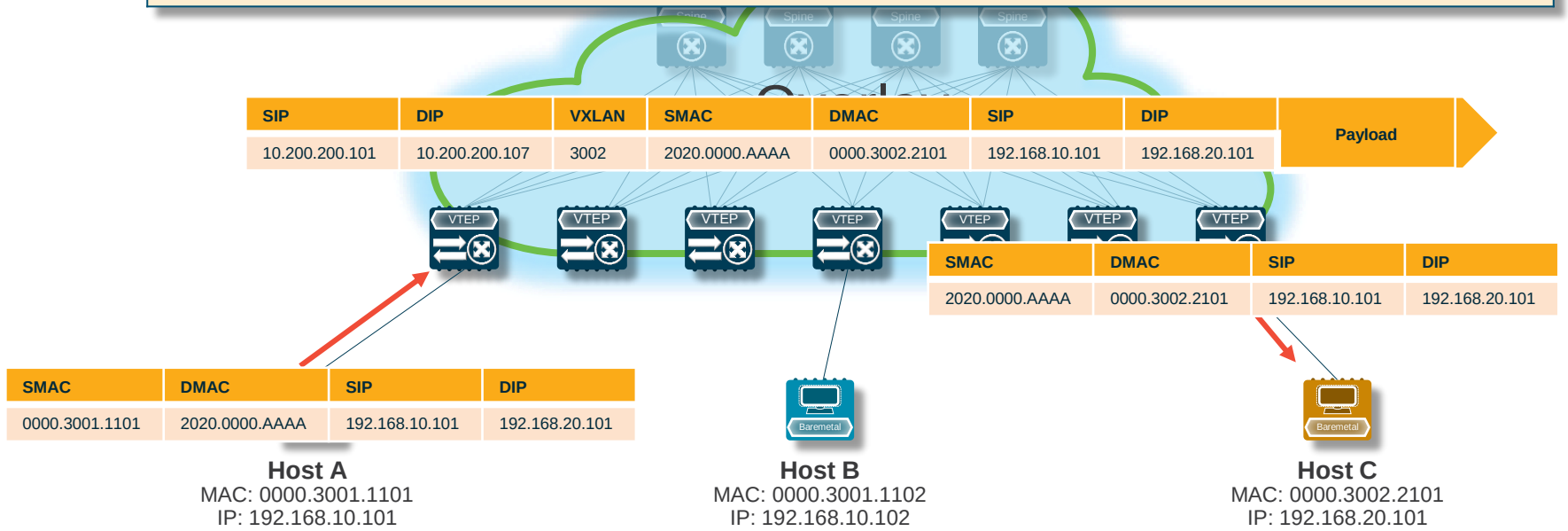
(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 4.1)

- Asymmetric IRB



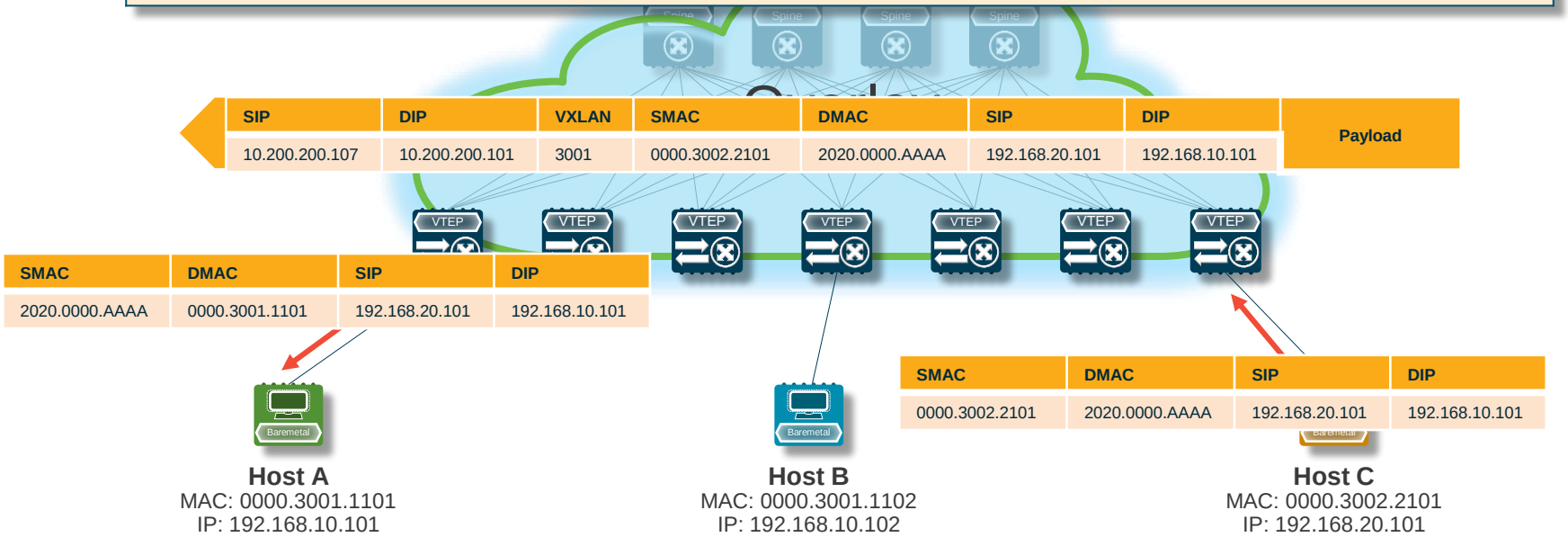
Packet Walk – Asymmetric IRB (A to C)

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		10.200.200.101	
2	0000.3002.2102 / 48	3002, 65500:3002	192.168.20.101/32		10.200.200.107	



Packet Walk – Asymmetric IRB (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		10.200.200.101	
2	0000.3002.2102 / 48	3002, 65500:3002	192.168.20.101/32		10.200.200.107	



Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

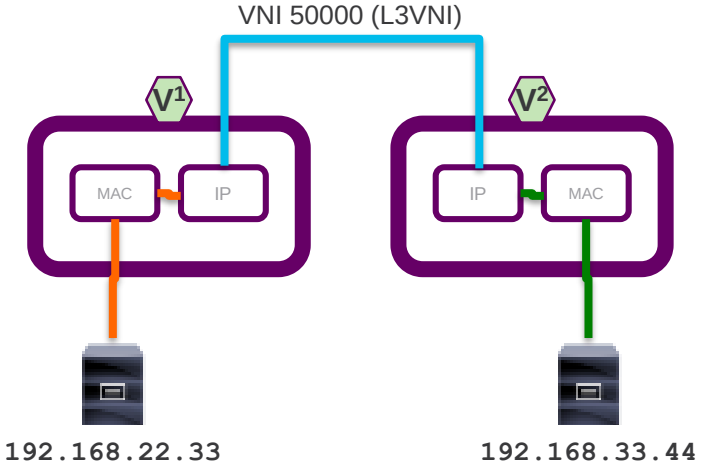
Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

Operational Models for Symmetric Inter-Subnet Forwarding

(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 5)

- Symmetric IRB

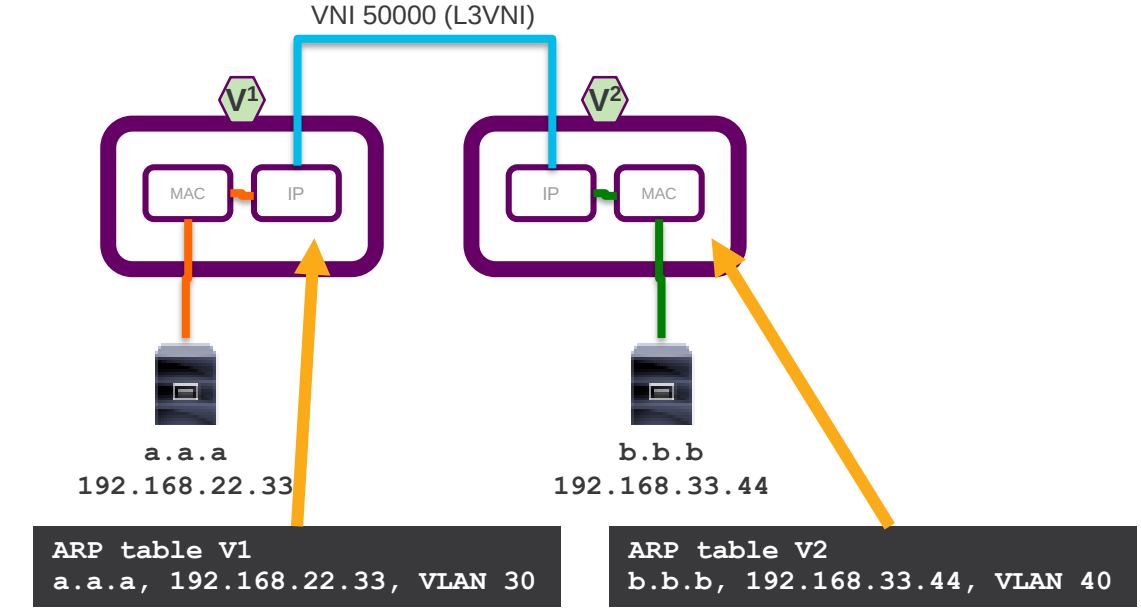


Bridge -> Route -> Route -> Bridge

Operational Models for Symmetric Inter-Subnet Forwarding

ARP and Adjacency Table

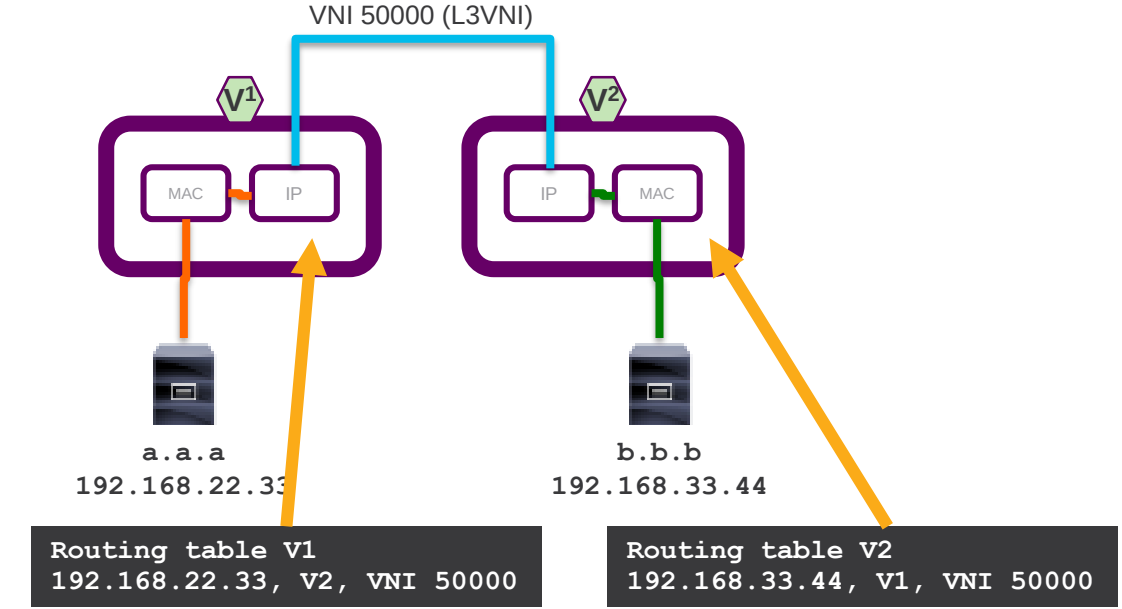
- Symmetric IRB



Operational Models for Symmetric Inter-Subnet Forwarding

Routing Table

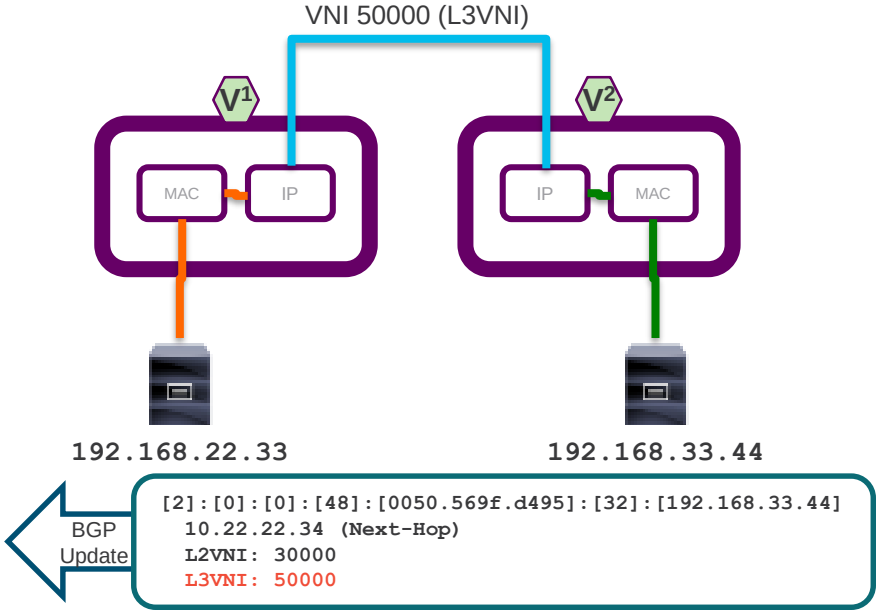
- Symmetric IRB



Operational Models for Symmetric Inter-Subnet Forwarding

(draft-ietf-bess-evpn-inter-subnet-forwarding – Section 5.1.1)

- Symmetric IRB

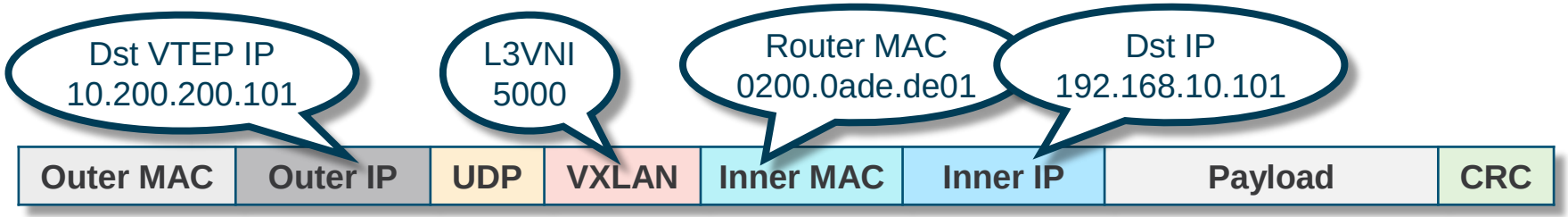


VXLAN and BGP EVPN – Putting it Together

Control-Plane (BGP EVPN)

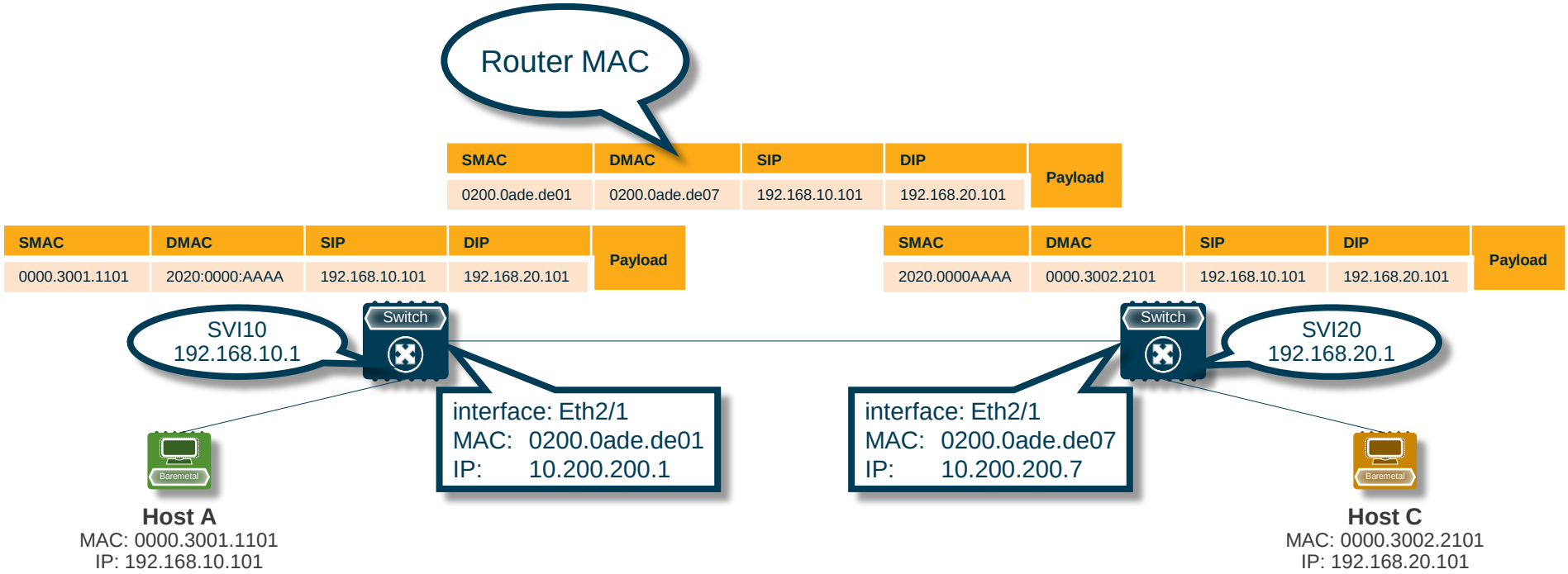
Extended Community
Router MAC
0200.0ade.de01

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	

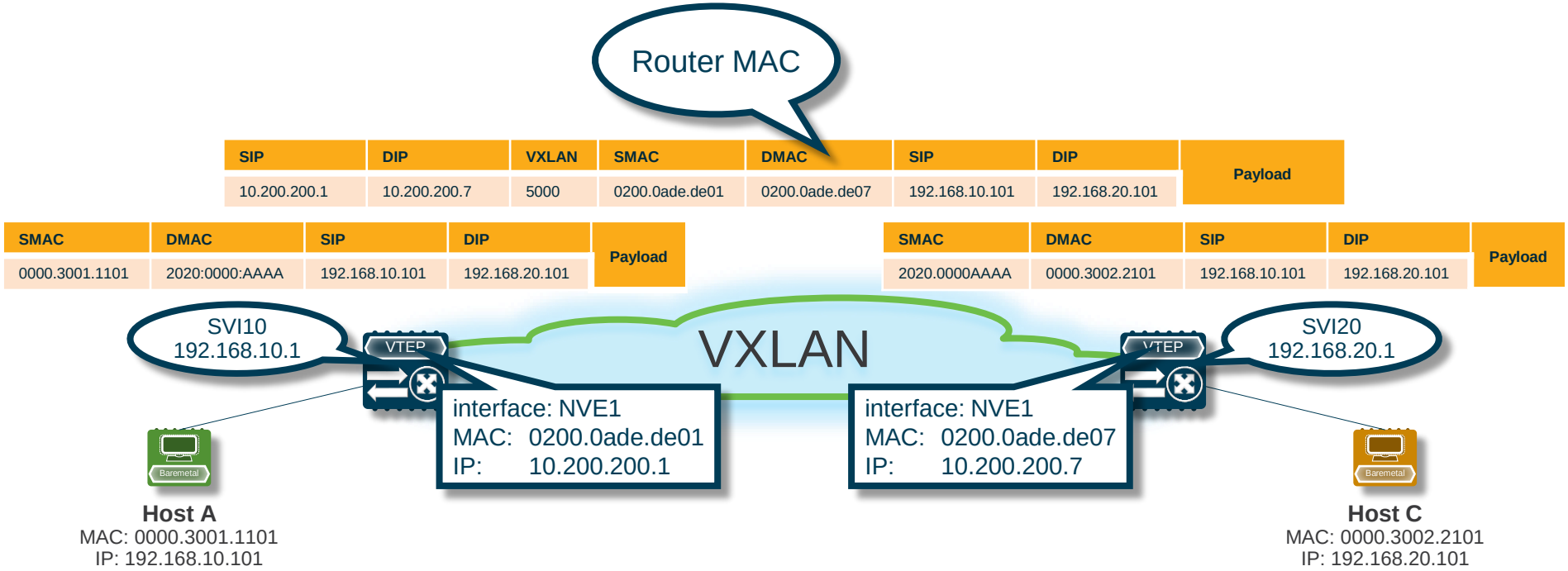


Data-Plane (VXLAN)

Routing and the Router MAC – Ethernet

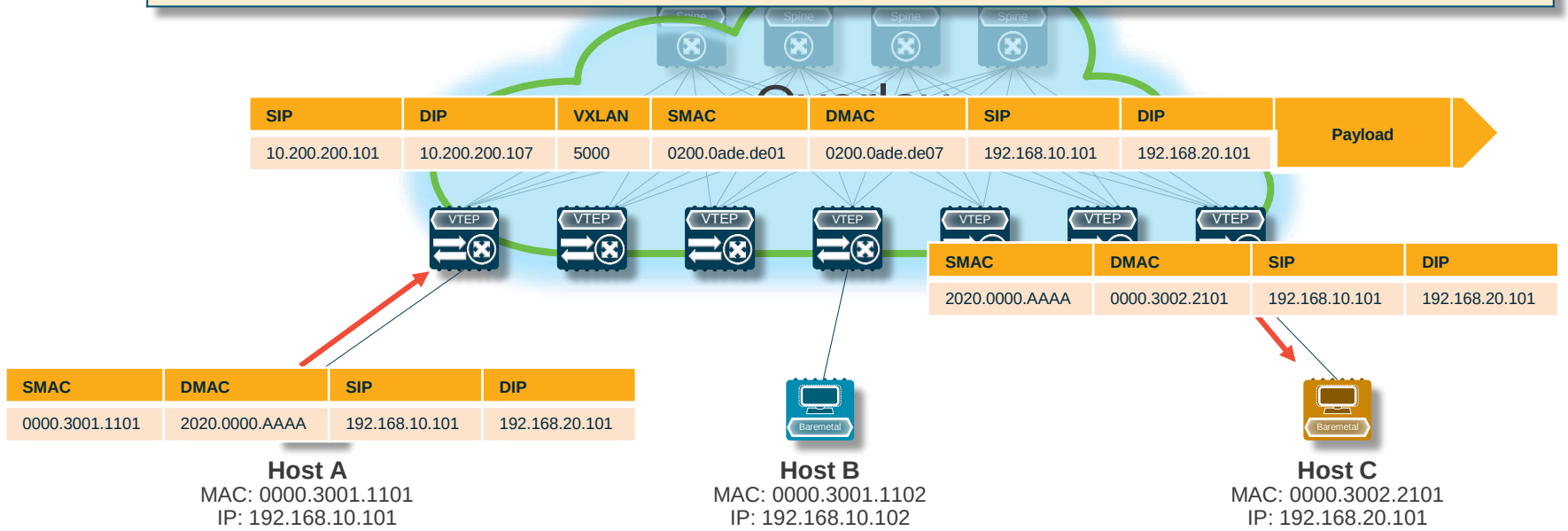


Routing and the Router MAC – VXLAN



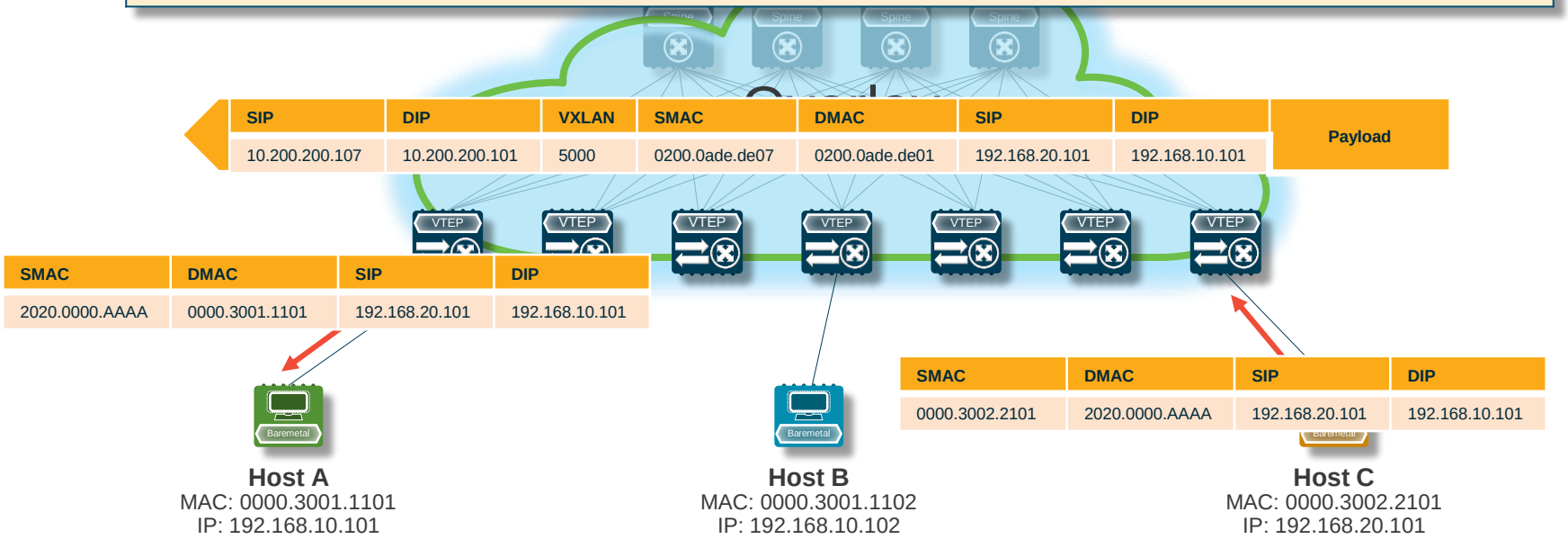
Packet Walk – Symmetric IRB (A to C)

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	



Packet Walk – Symmetric IRB (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	



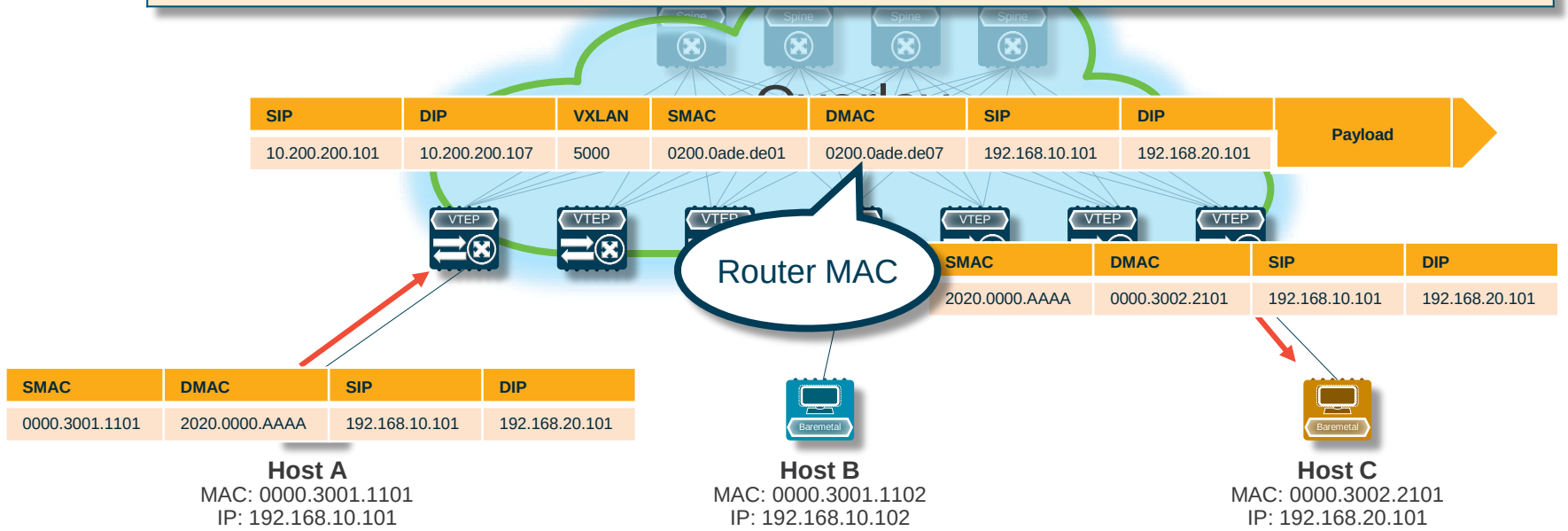
Host A
MAC: 0000.3001.1101
IP: 192.168.10.101

Host B
MAC: 0000.3001.1102
IP: 192.168.10.102

Host C
MAC: 0000.3002.2101
IP: 192.168.20.101

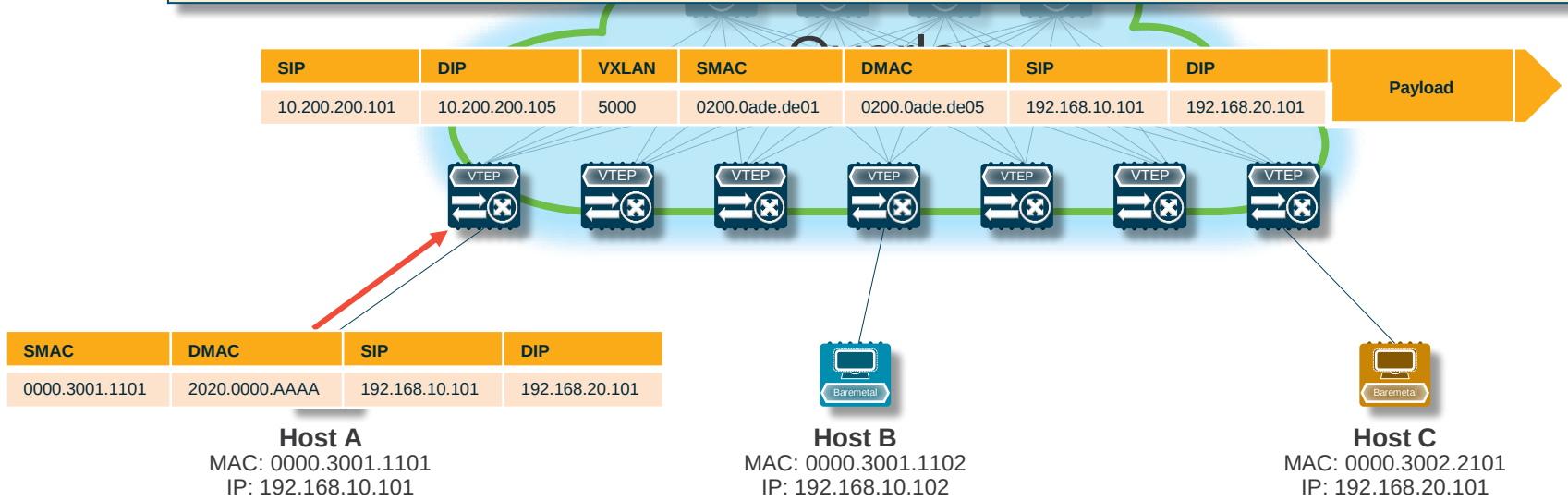
Packet Walk – Routing

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	



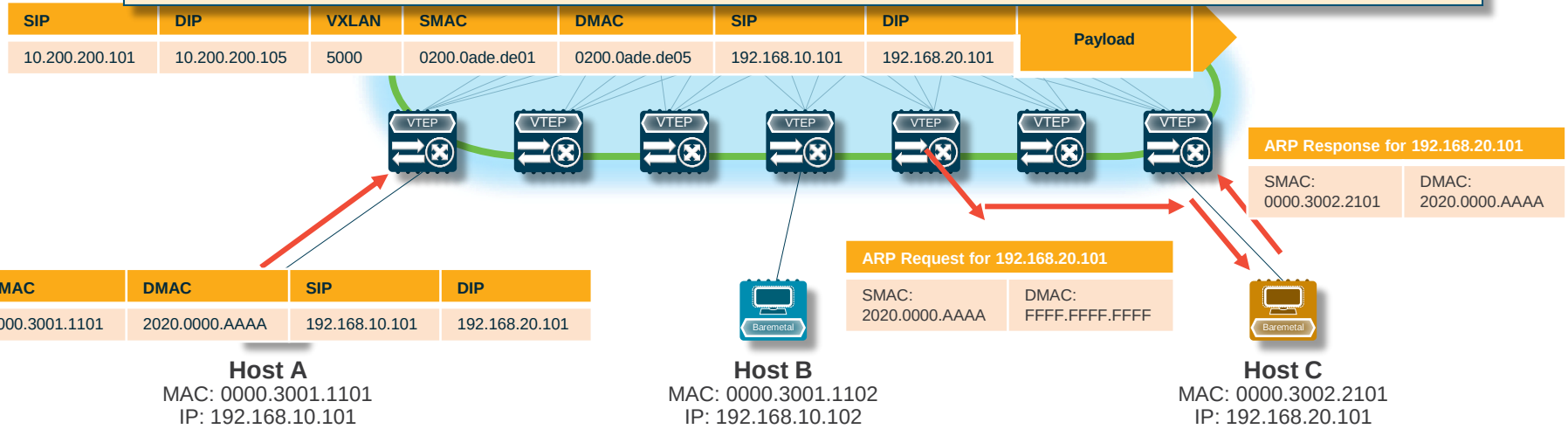
Packet Walk – Routing (Silent Host)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	3001, 65500:3001	192.168.10.101	5000, 65500:5000	10.200.200.101	
5			192.168.20.0/24	5000, 65500:5000	10.200.200.105	
5			192.168.20.0/24	5000, 65500:5000	10.200.200.107	



Packet Walk – Routing (Silent Host)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101 / 48	3001, 65500:3001	192.168.10.101	5000, 65500:5000	10.200.200.101	
5			192.168.20.0/24	5000, 65500:5000	10.200.200.105	
5			192.168.20.0/24	5000, 65500:5000	10.200.200.107	
2	0000.3002.21o1 / 48	3002, 65500:3002	192.168.20.101	5000, 65500:5000	10.200.200.107	



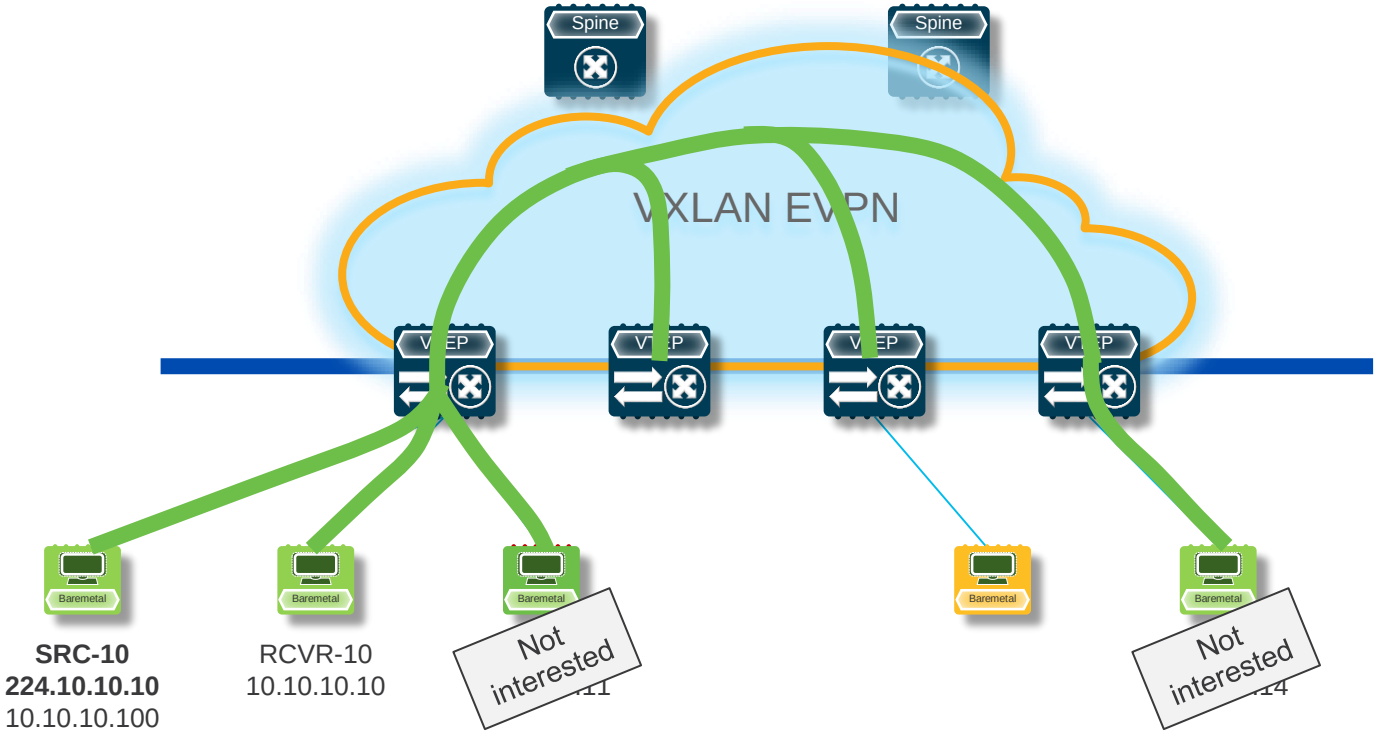
Agenda

- Introduction to Overlays
- VXLAN with BGP EVPN
  - Standards and Implementation
  - Control & Data Plane
- Tenant Routed Multicast (TRM)
- Multi-Site
- VXLAN OAM

Multicast Forwarding
Tenant Routed Multicast (TRM)

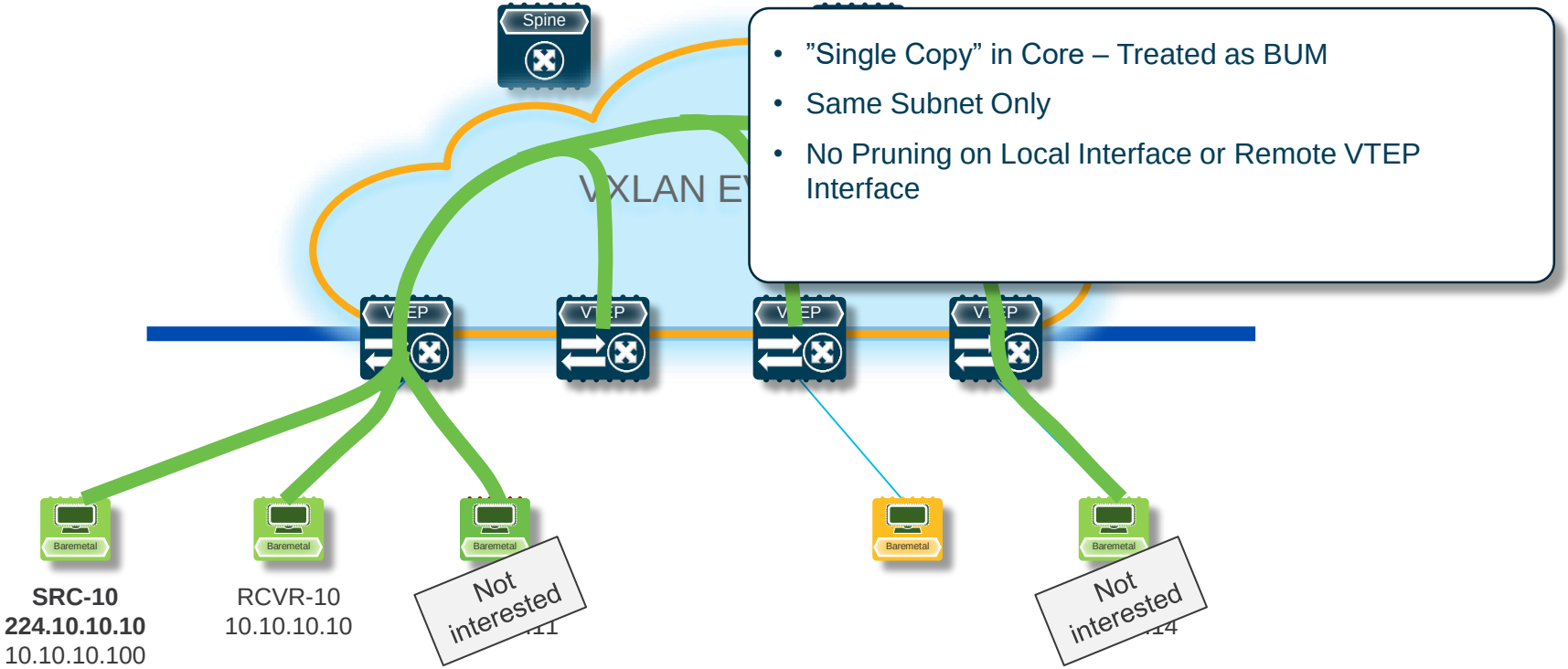
Same Subnet Forwarding **no** IGMP Snooping

Traditional Forwarding in VXLAN Overlays



Same Subnet Forwarding **no** IGMP Snooping

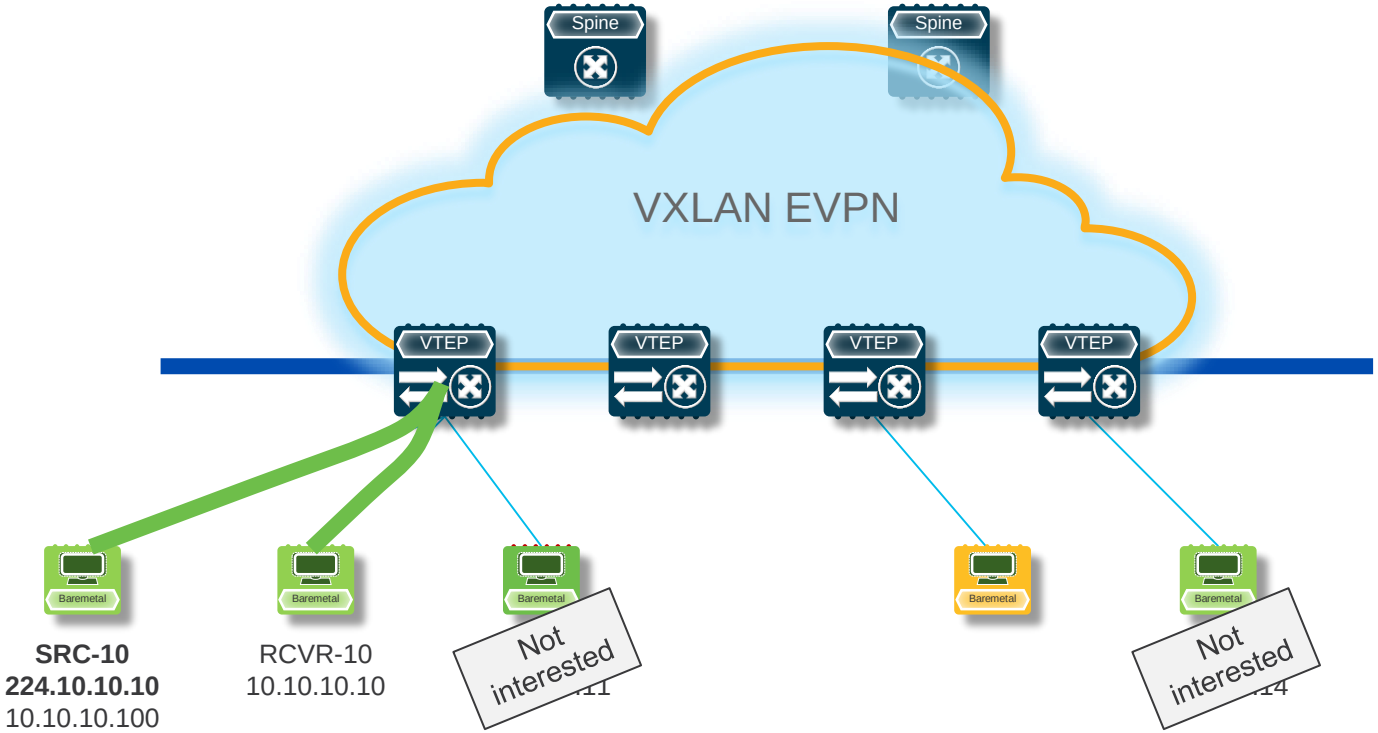
Traditional Forwarding in VXLAN Overlays



- ”Single Copy” in Core – Treated as BUM
- Same Subnet Only
- No Pruning on Local Interface or Remote VTEP Interface

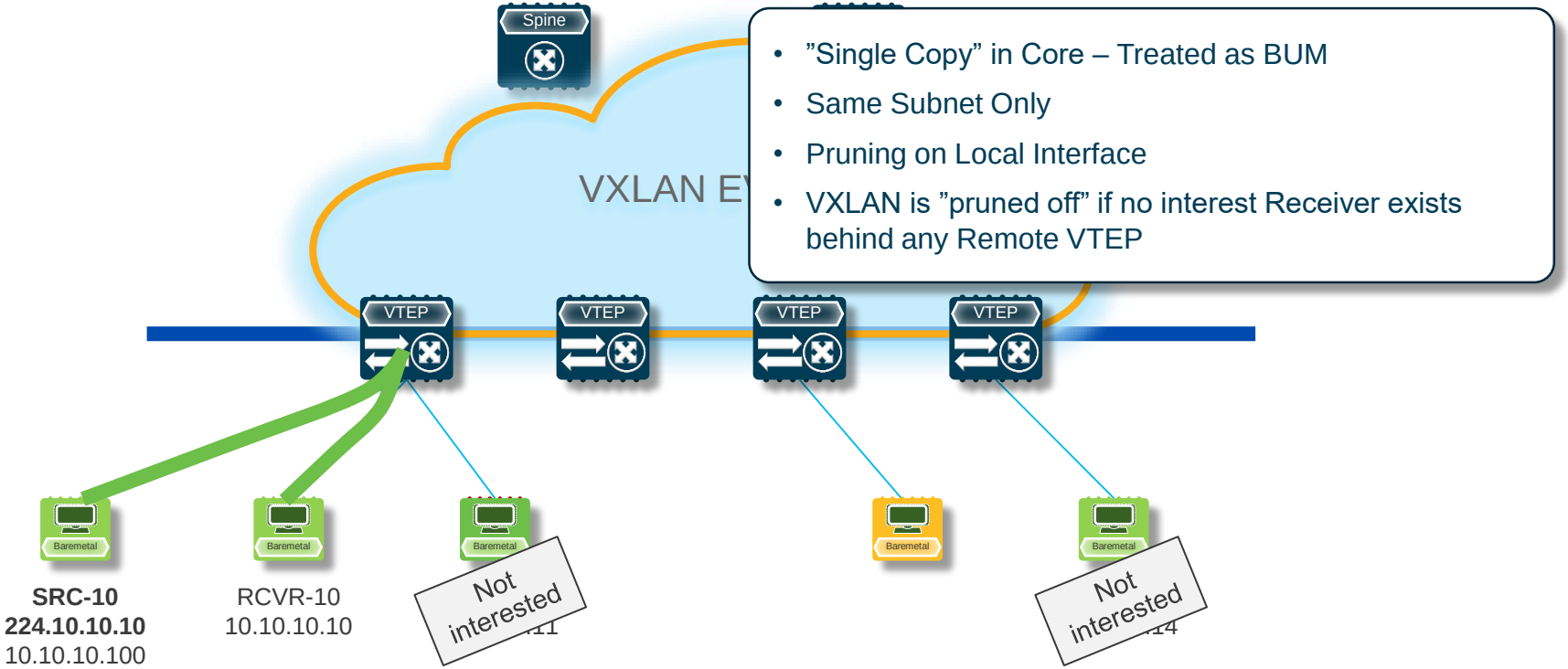
Same Subnet Forwarding **with** IGMP Snooping

Traditional Forwarding in VXLAN Overlays



Same Subnet Forwarding **with** IGMP Snooping

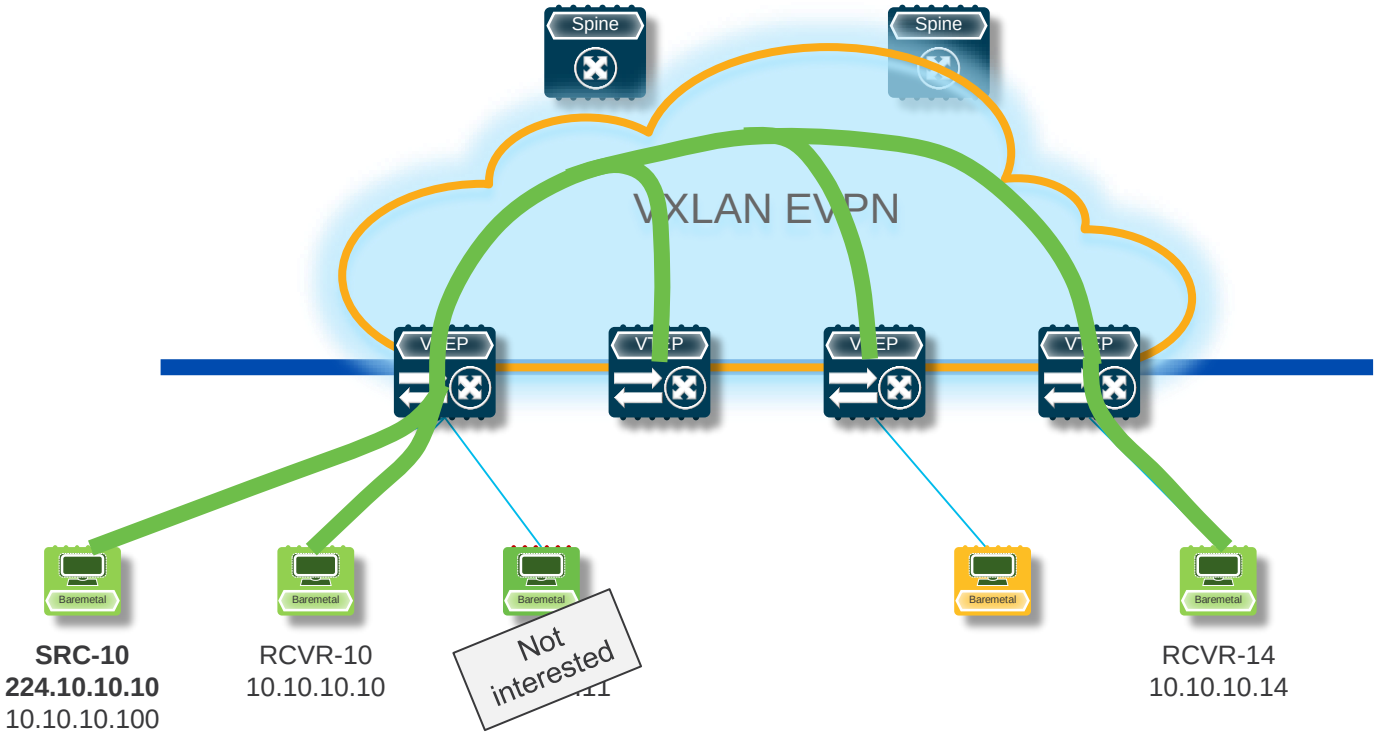
Traditional Forwarding in VXLAN Overlays



- ”Single Copy” in Core – Treated as BUM
- Same Subnet Only
- Pruning on Local Interface
- VXLAN is ”pruned off” if no interest Receiver exists behind any Remote VTEP

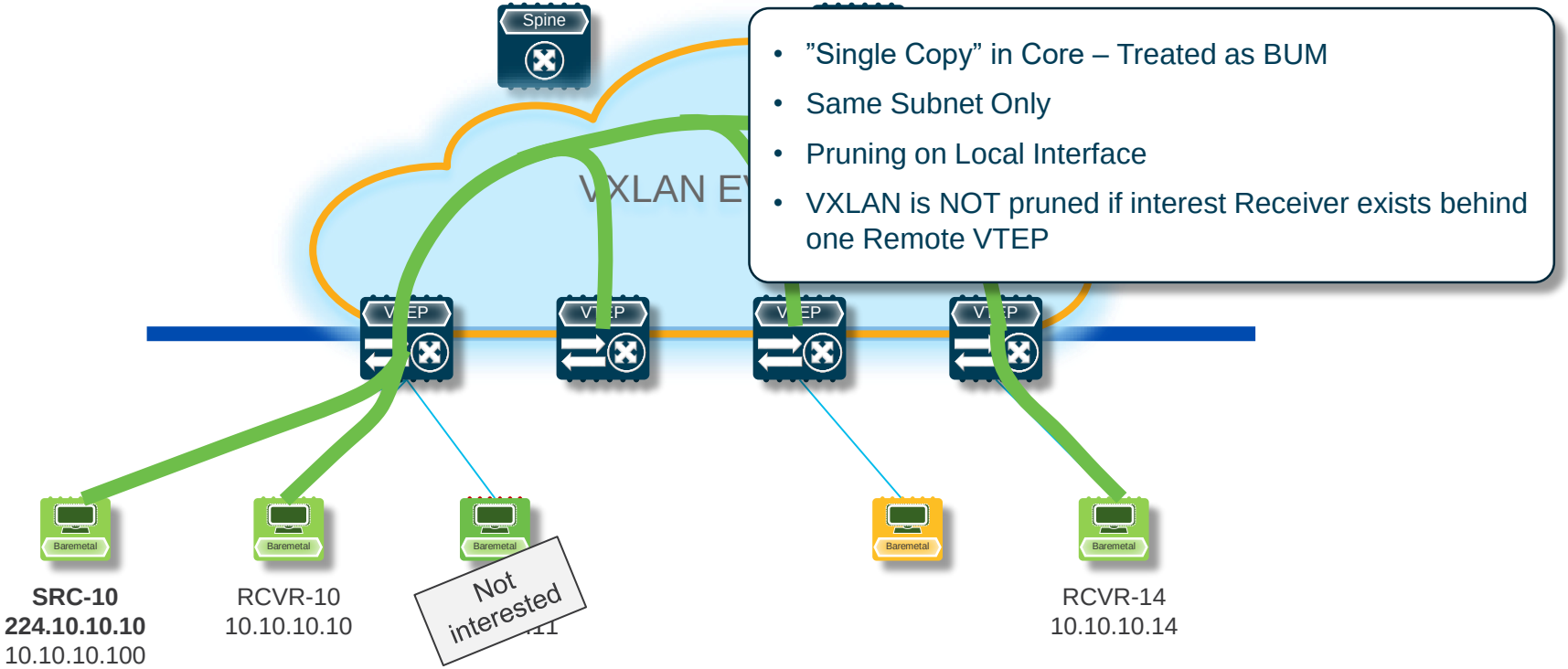
Same Subnet Forwarding **with** IGMP Snooping

Traditional Forwarding in VXLAN Overlays



Same Subnet Forwarding **with** IGMP Snooping

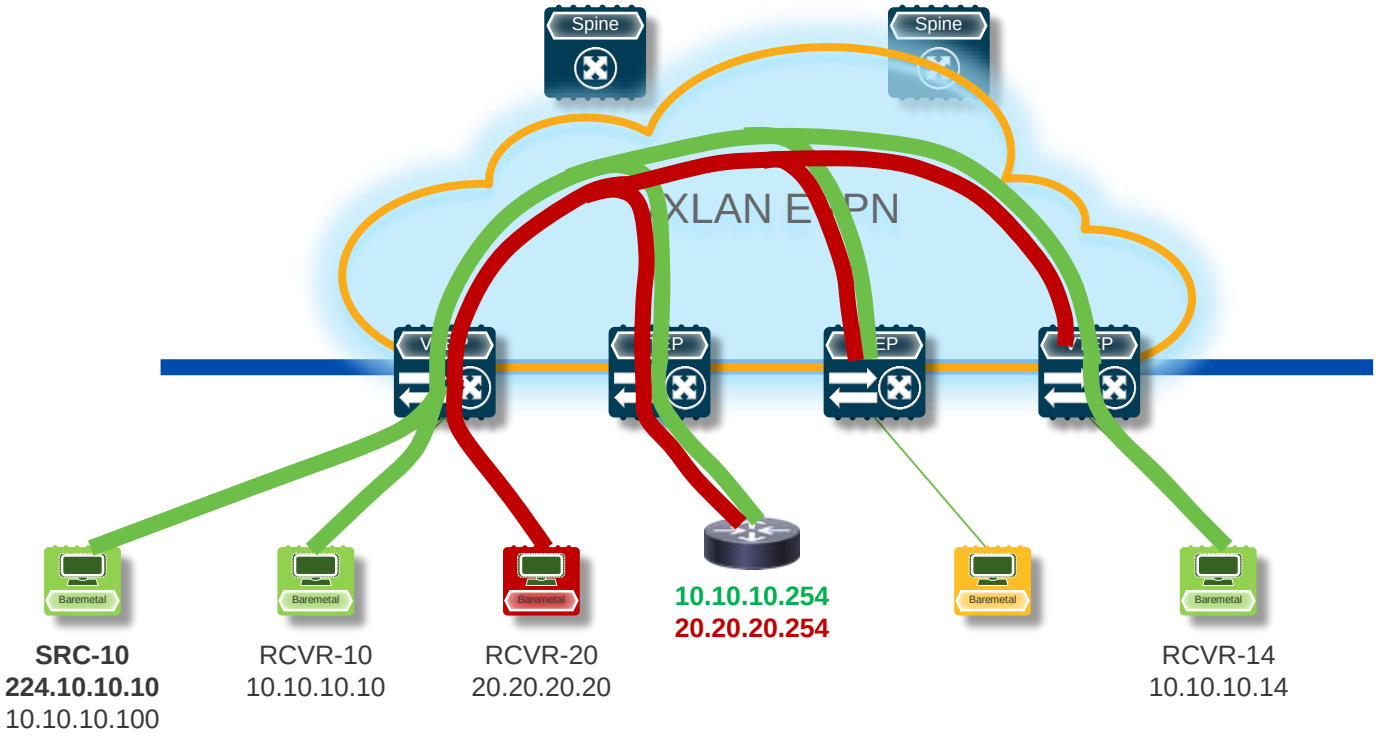
Traditional Forwarding in VXLAN Overlays



- ”Single Copy” in Core – Treated as BUM
- Same Subnet Only
- Pruning on Local Interface
- VXLAN is NOT pruned if interest Receiver exists behind one Remote VTEP

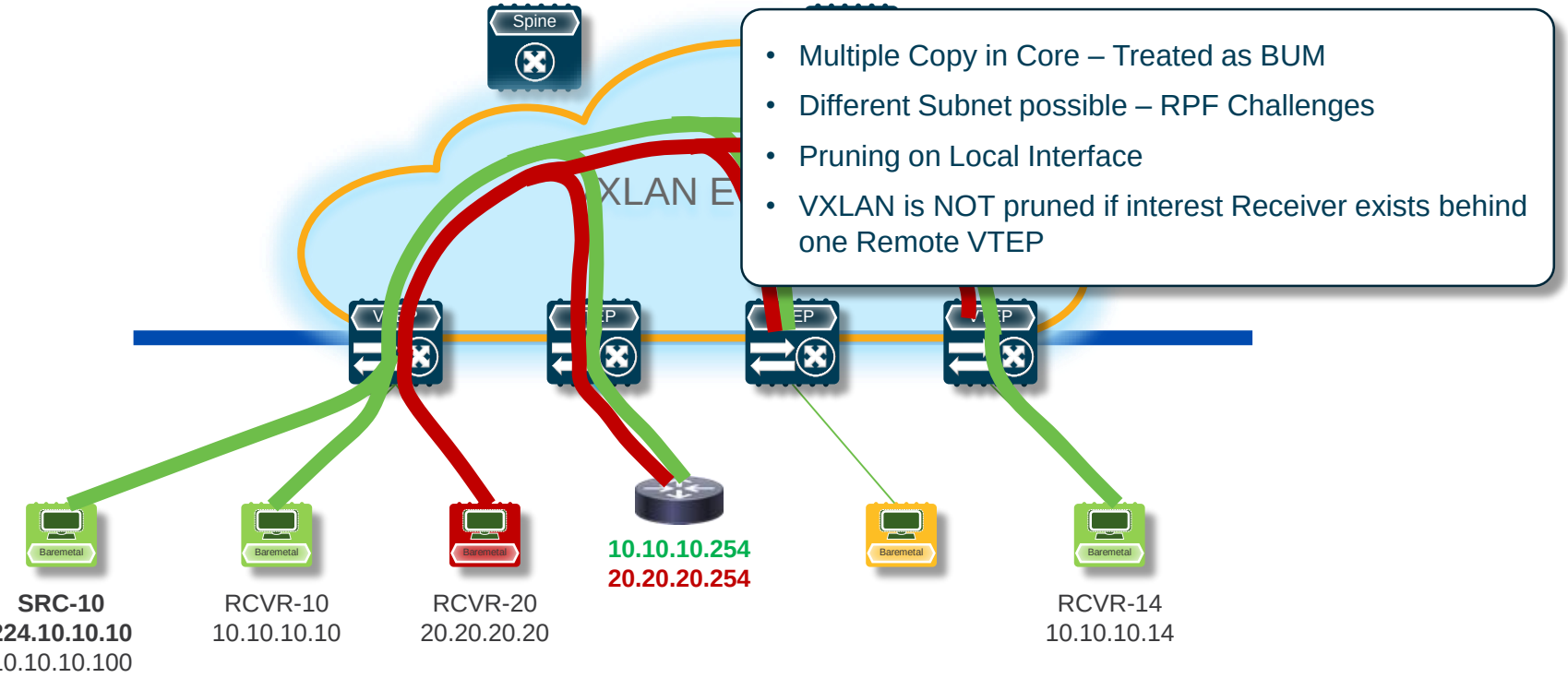
Different Subnet Forwarding – Router on-a-Stick

Traditional Forwarding in VXLAN Overlays



Different Subnet Forwarding – Router on-a-Stick

Traditional Forwarding in VXLAN Overlays

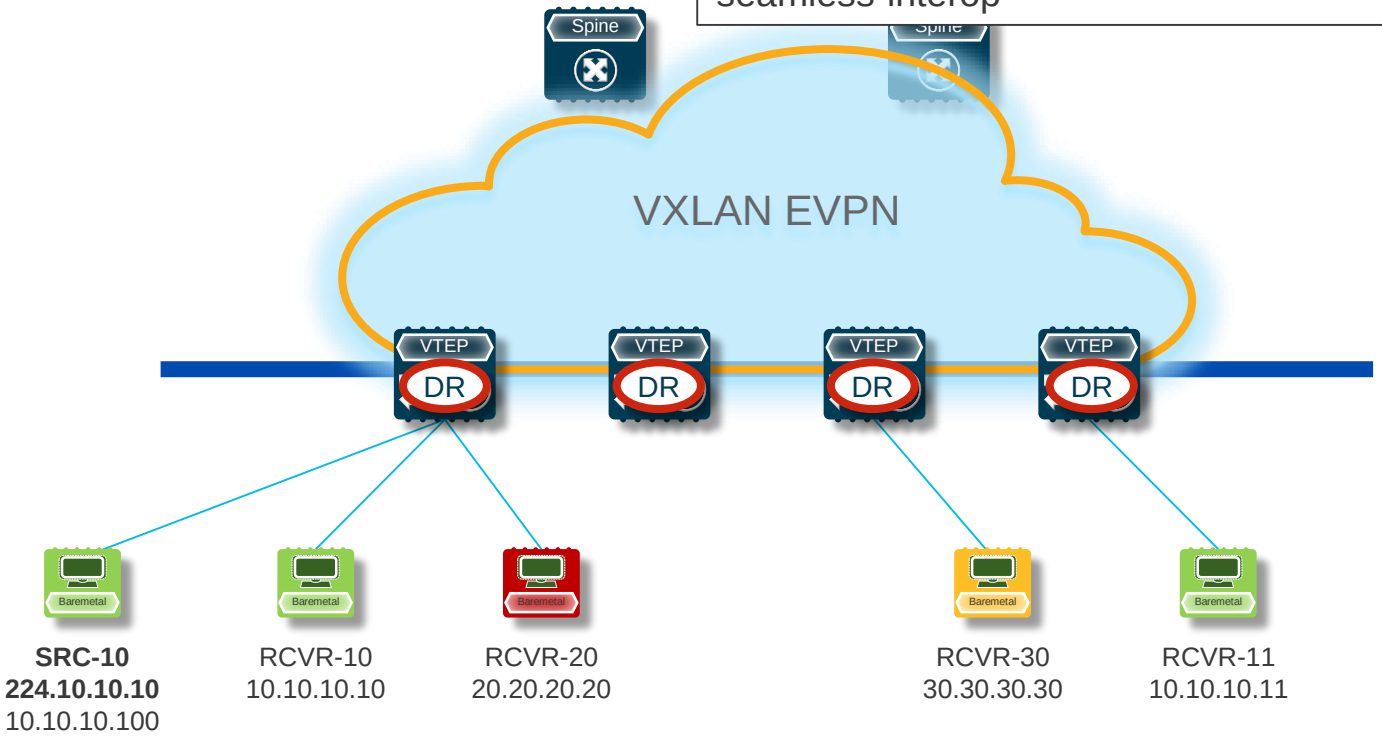


- Multiple Copy in Core – Treated as BUM
- Different Subnet possible – RPF Challenges
- Pruning on Local Interface
- VXLAN is NOT pruned if interest Receiver exists behind one Remote VTEP

Functional Components

Tenant Routed Multicast (TRM)

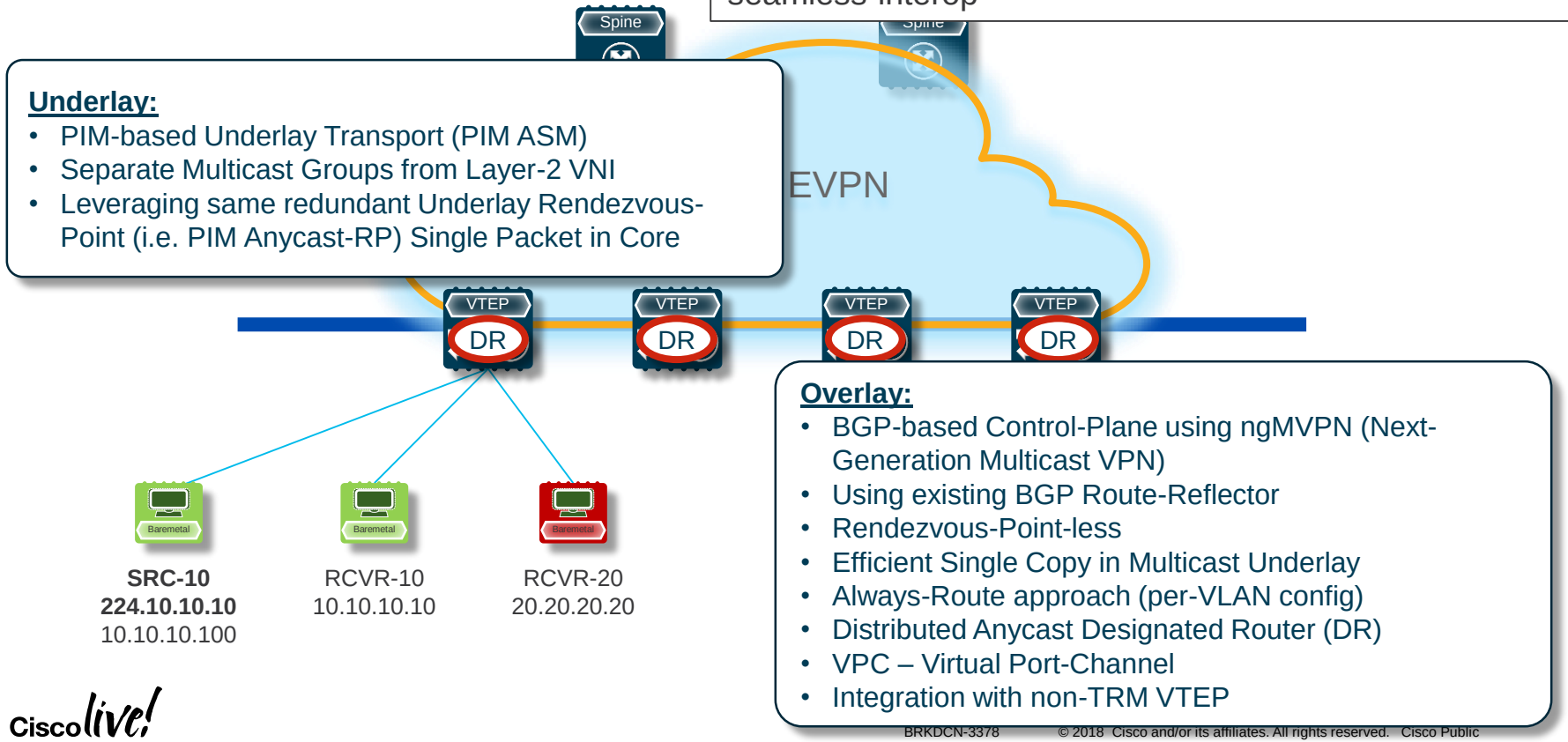
https://tools.ietf.org/html/draft-sajassi-bess-evpn-mvpn-seamless-interop



Functional Components

Tenant Routed Multicast (TRM)

https://tools.ietf.org/html/draft-sajassi-bess-evpn-mvpn-seamless-interop



Underlay:

- PIM-based Underlay Transport (PIM ASM)
- Separate Multicast Groups from Layer-2 VNI
- Leveraging same redundant Underlay Rendezvous-Point (i.e. PIM Anycast-RP) Single Packet in Core

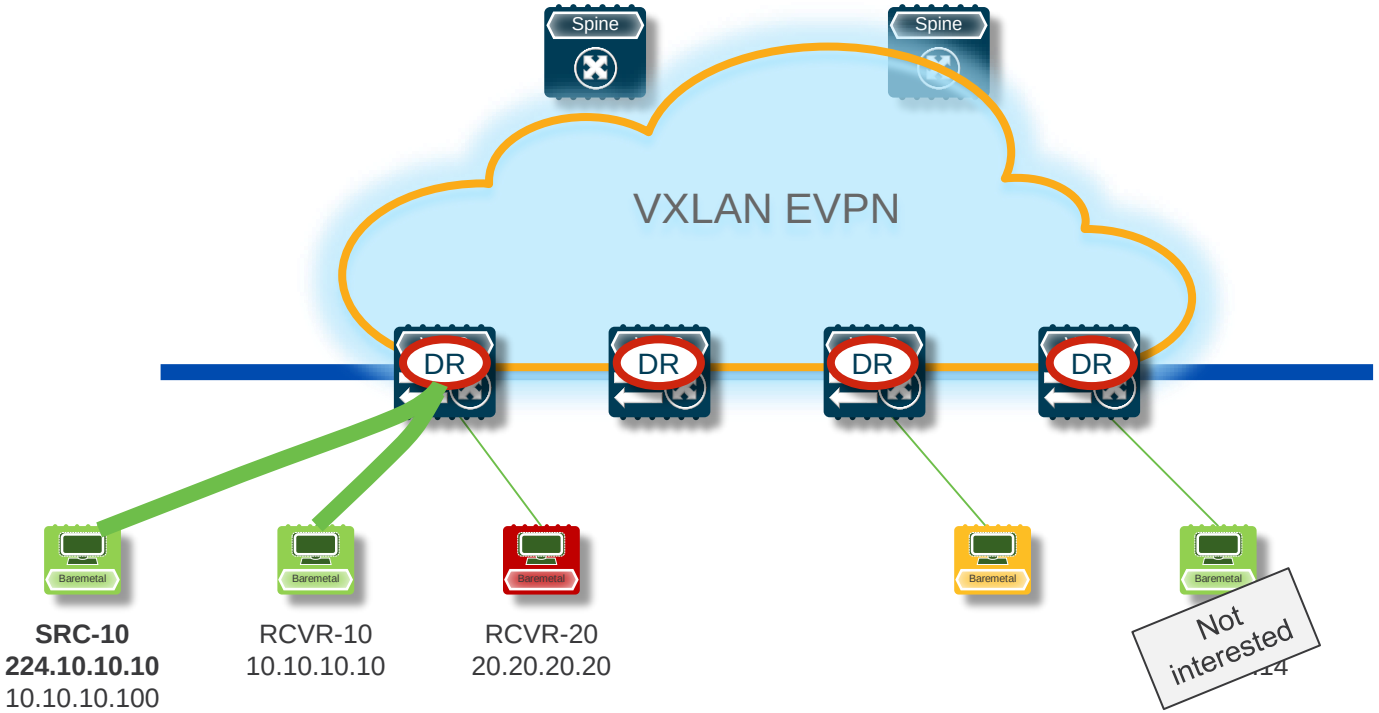
Overlay:

- BGP-based Control-Plane using ngMVPN (Next-Generation Multicast VPN)
- Using existing BGP Route-Reflector
- Rendezvous-Point-less
- Efficient Single Copy in Multicast Underlay
- Always-Route approach (per-VLAN config)
- Distributed Anycast Designated Router (DR)
- VPC – Virtual Port-Channel
- Integration with non-TRM VTEP

Forwarding Behaviour
Tenant Routed Multicast (TRM)

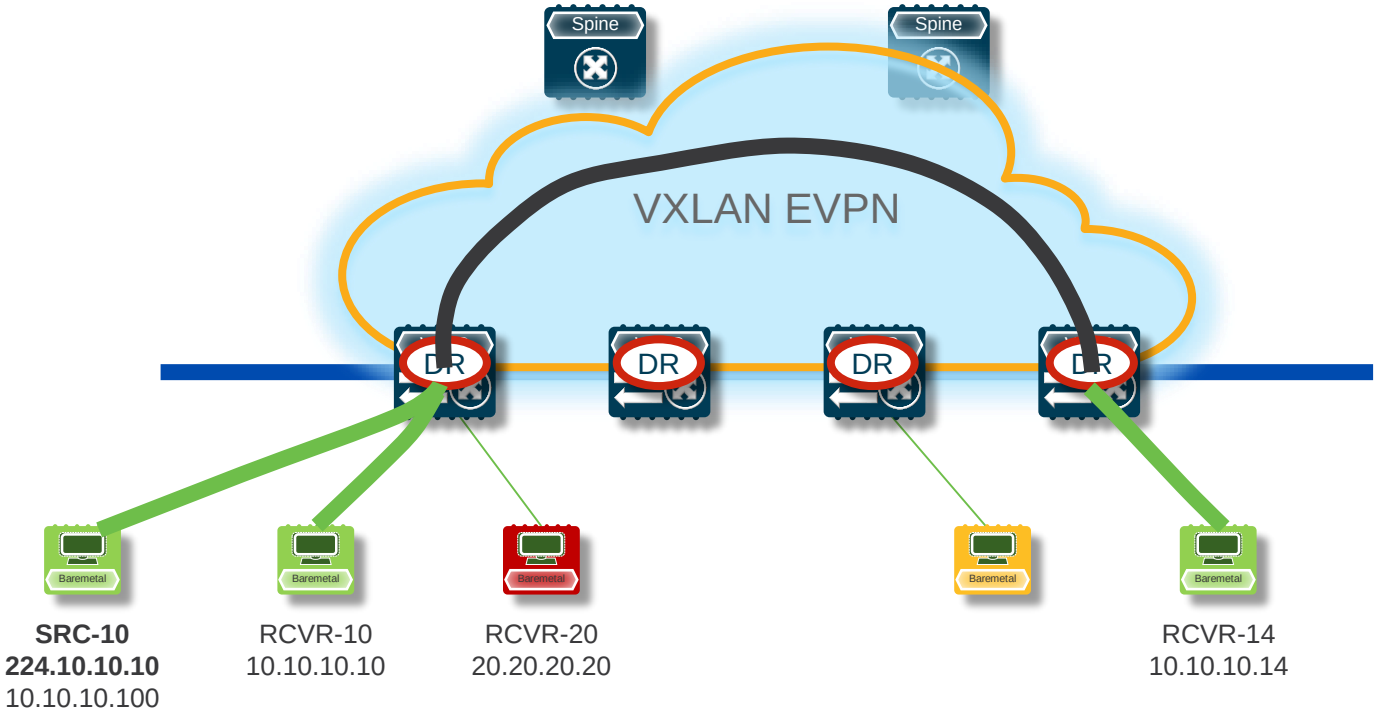
Same Subnet Forwarding – Local and Remote Snooping

TRM Forwarding (Layer-2 only Mode)



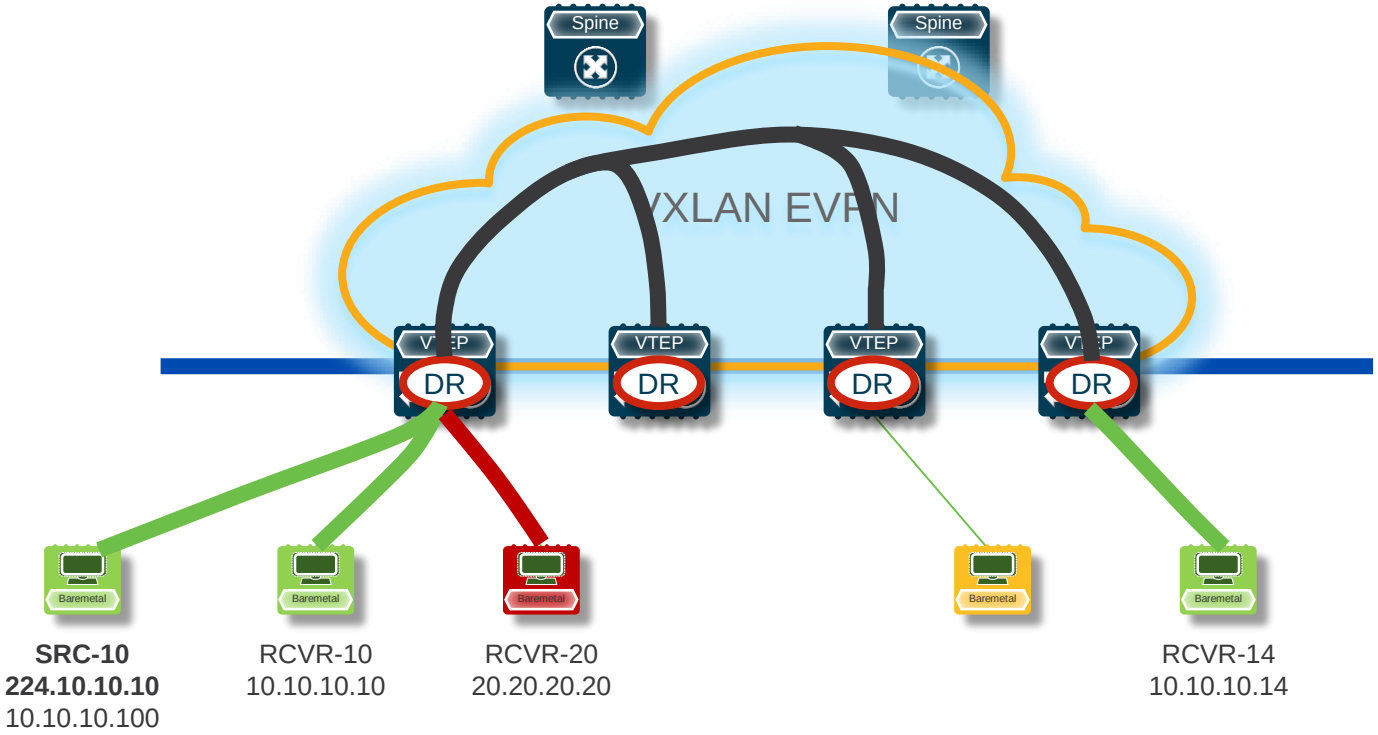
Same Subnet Forwarding – Local and Remote Snooping

TRM Forwarding (Layer-2 only mode)



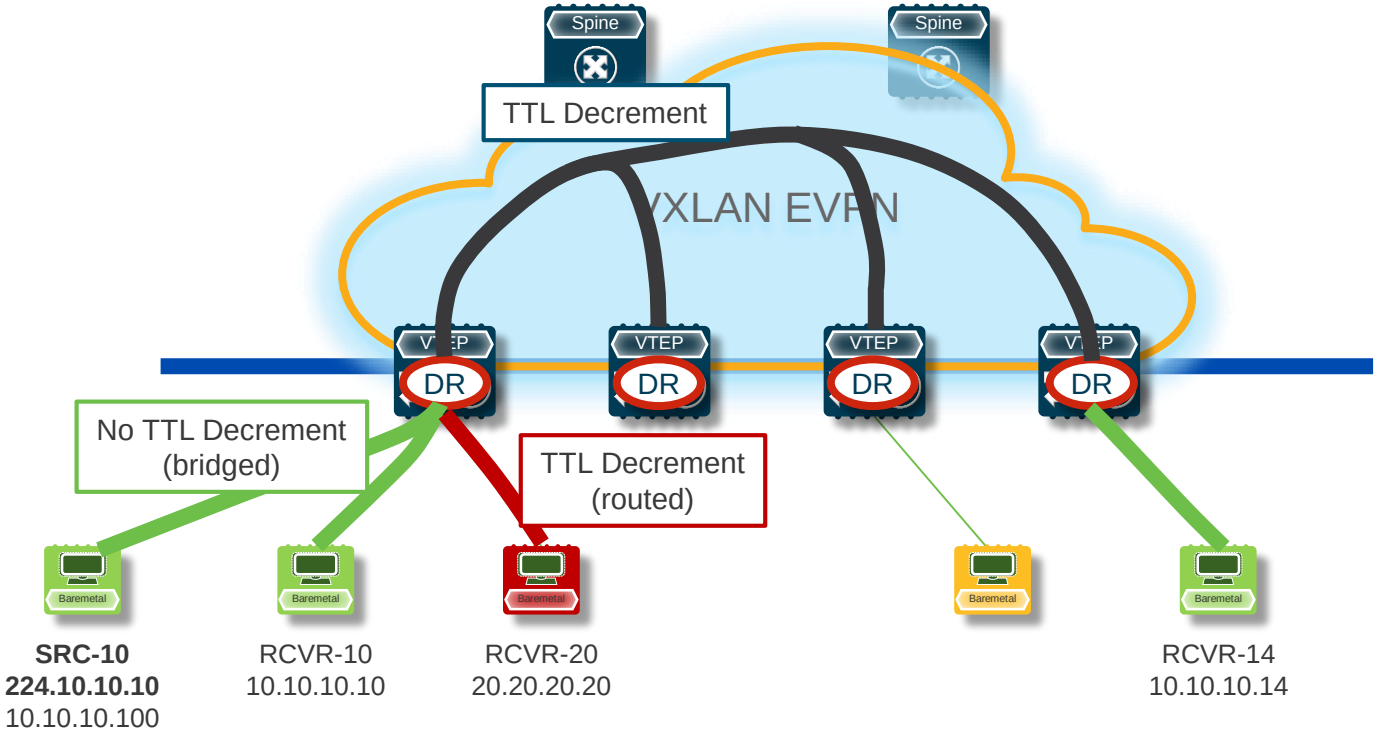
Different and Same Subnet Forwarding

TRM Forwarding (Layer-3 Mode)

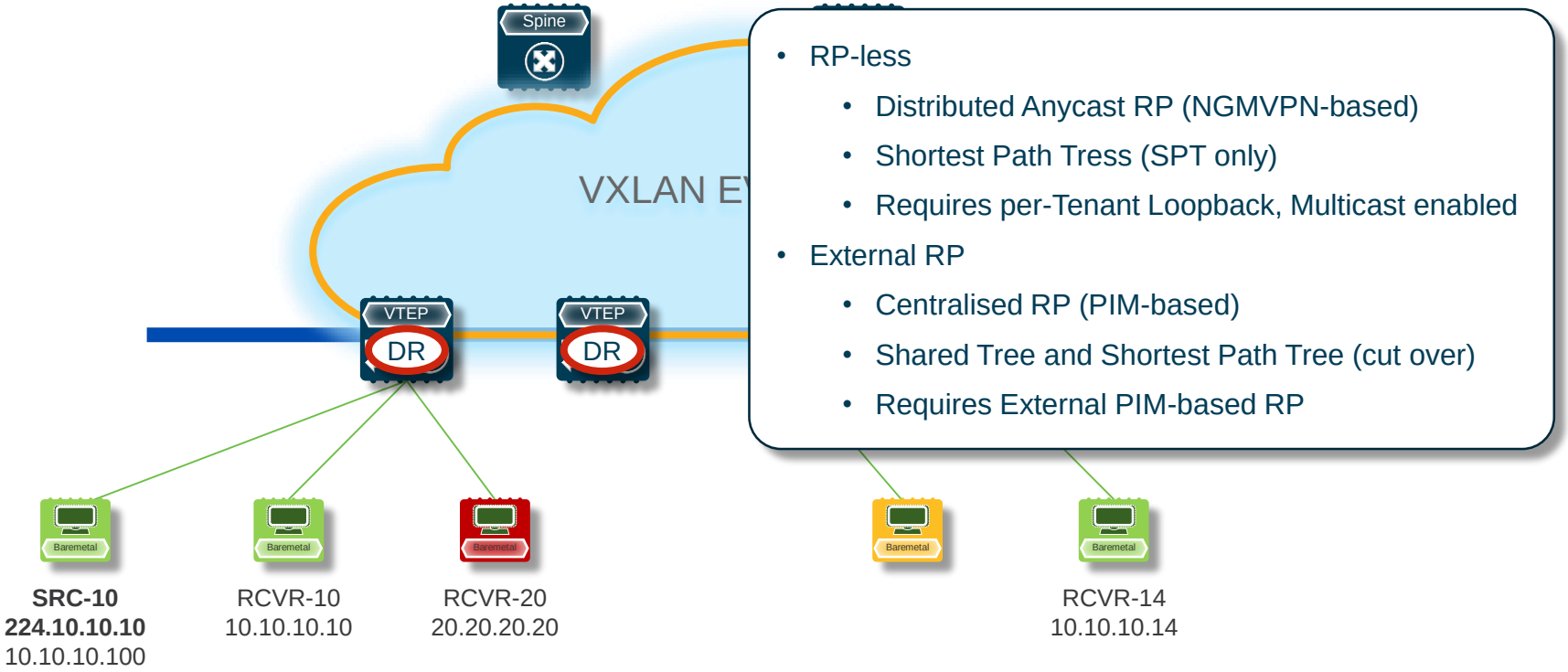


Local and Remote Forwarding

TRM Forwarding (Layer-3 Mode)

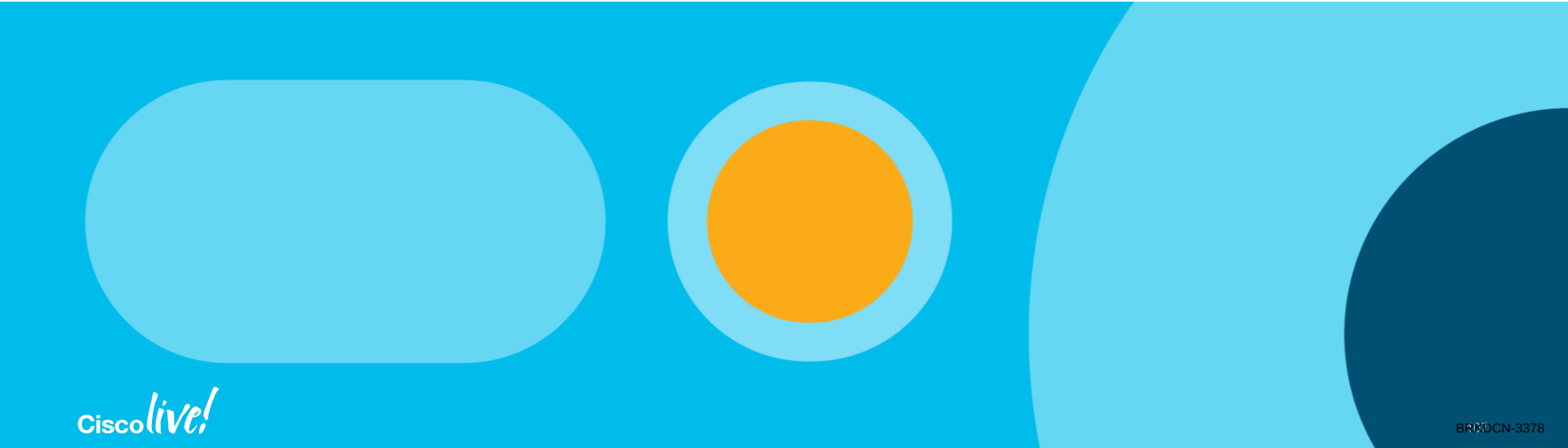


Overlay Rendezvous Point

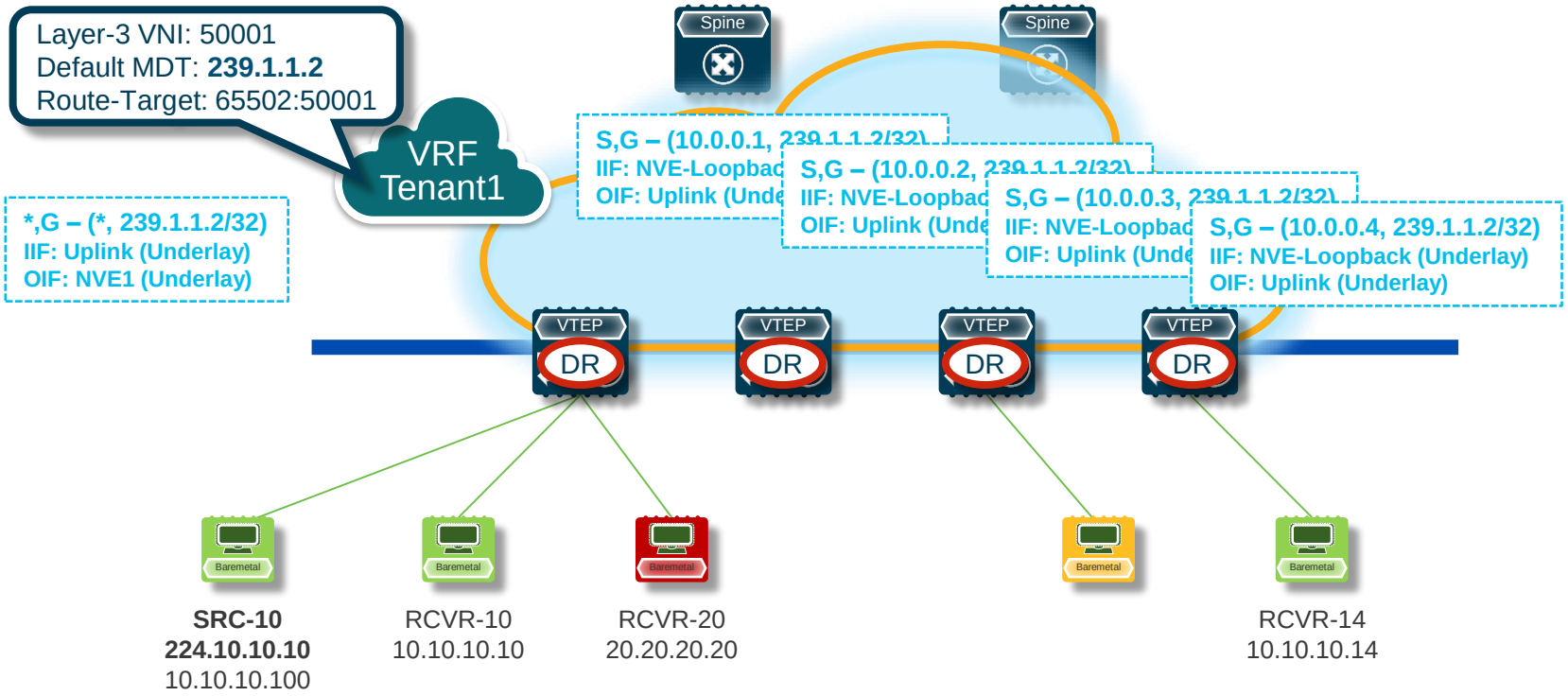


- RP-less
  - Distributed Anycast RP (NGMVPN-based)
  - Shortest Path Tress (SPT only)
  - Requires per-Tenant Loopback, Multicast enabled
- External RP
  - Centralised RP (PIM-based)
  - Shared Tree and Shortest Path Tree (cut over)
  - Requires External PIM-based RP

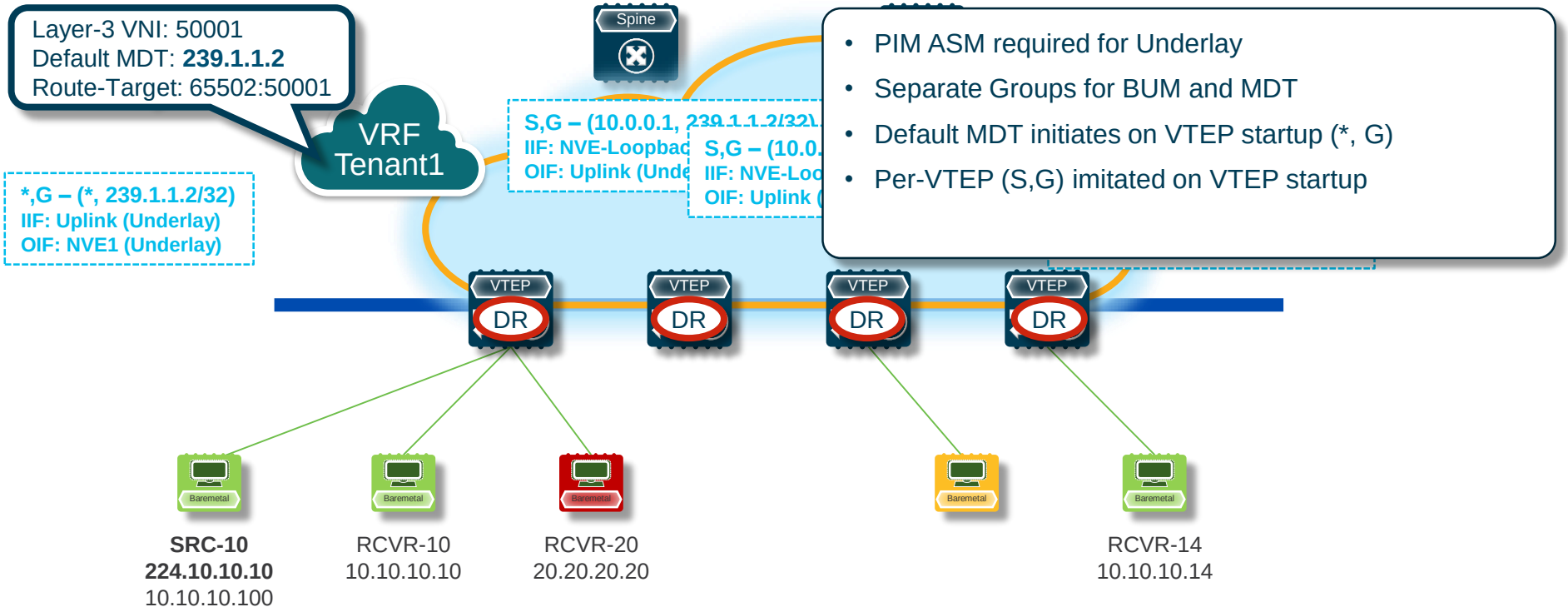
TRM Control- & Data-Plane



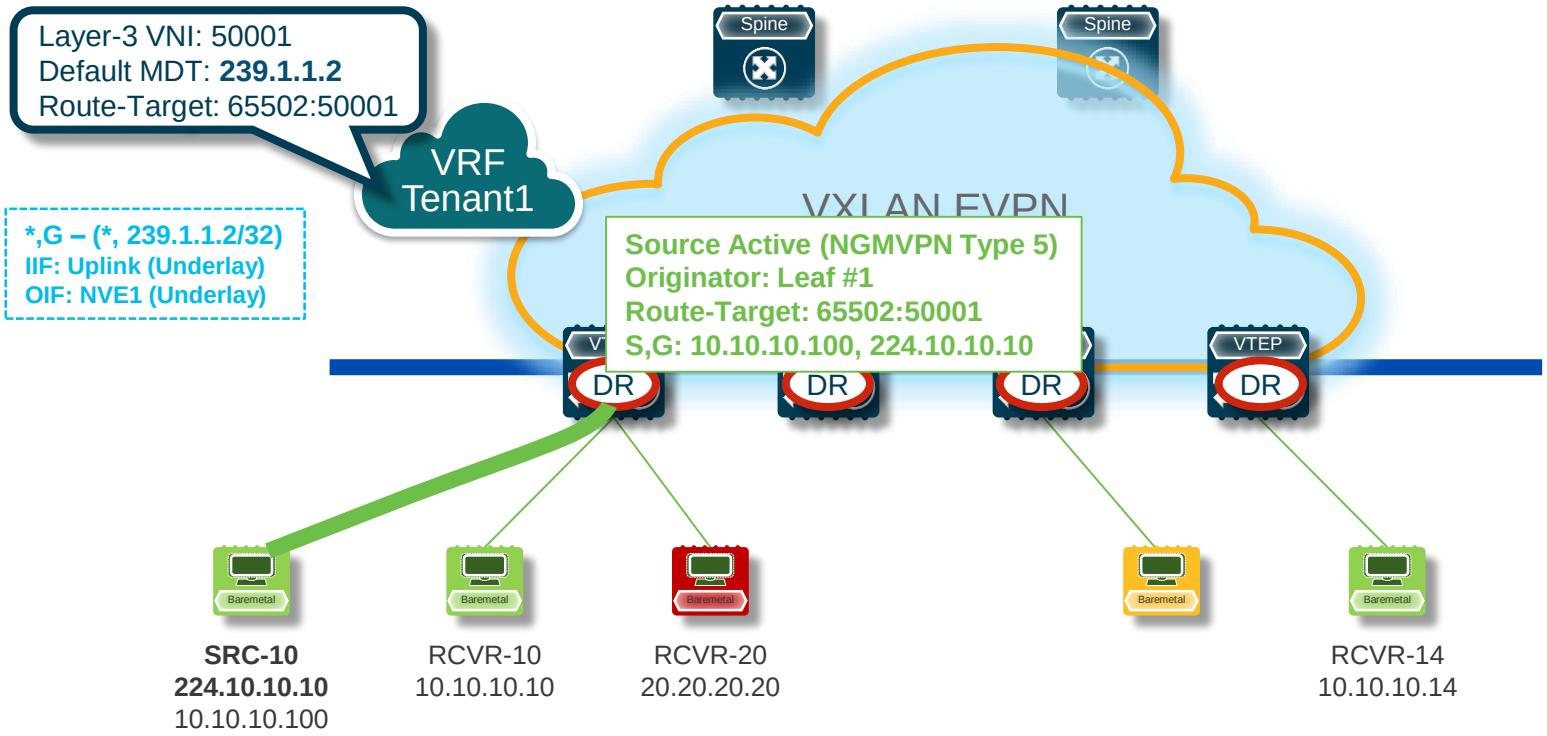
Underlay Multicast Tree – PIM ASM



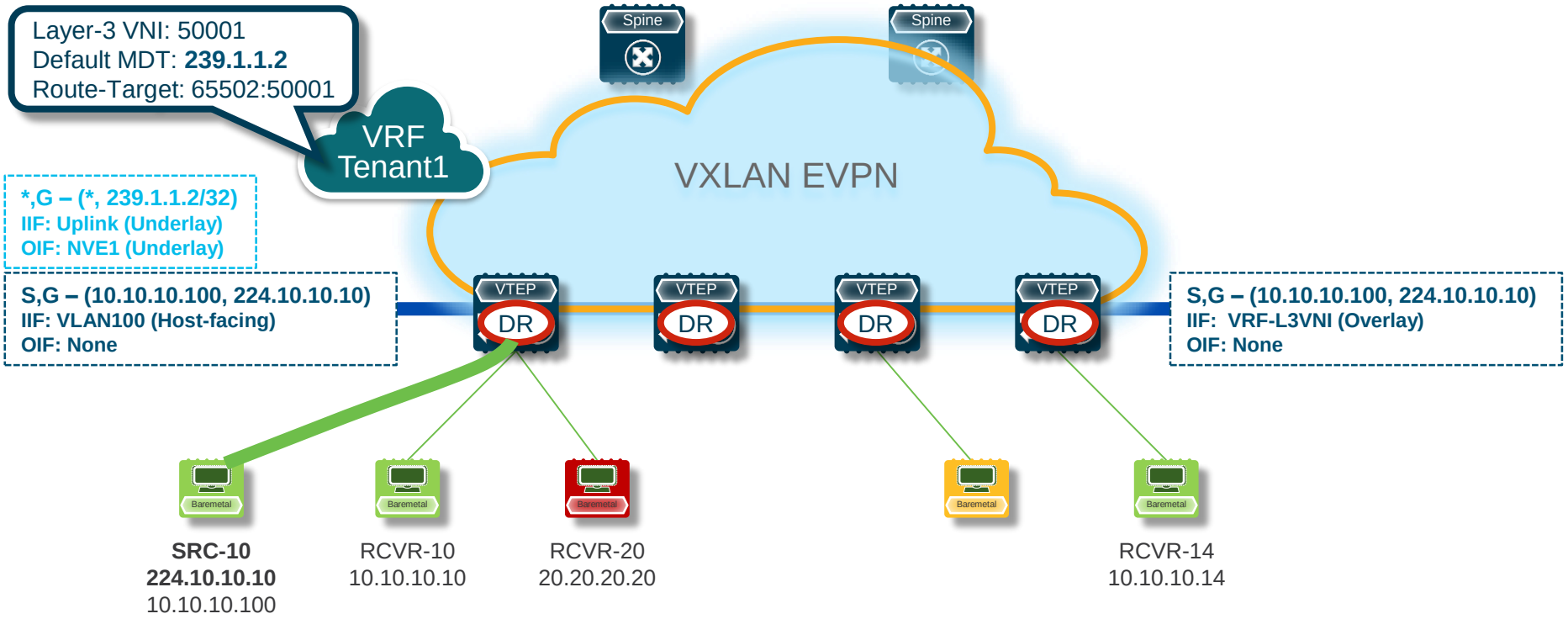
Underlay Multicast Tree – PIM ASM



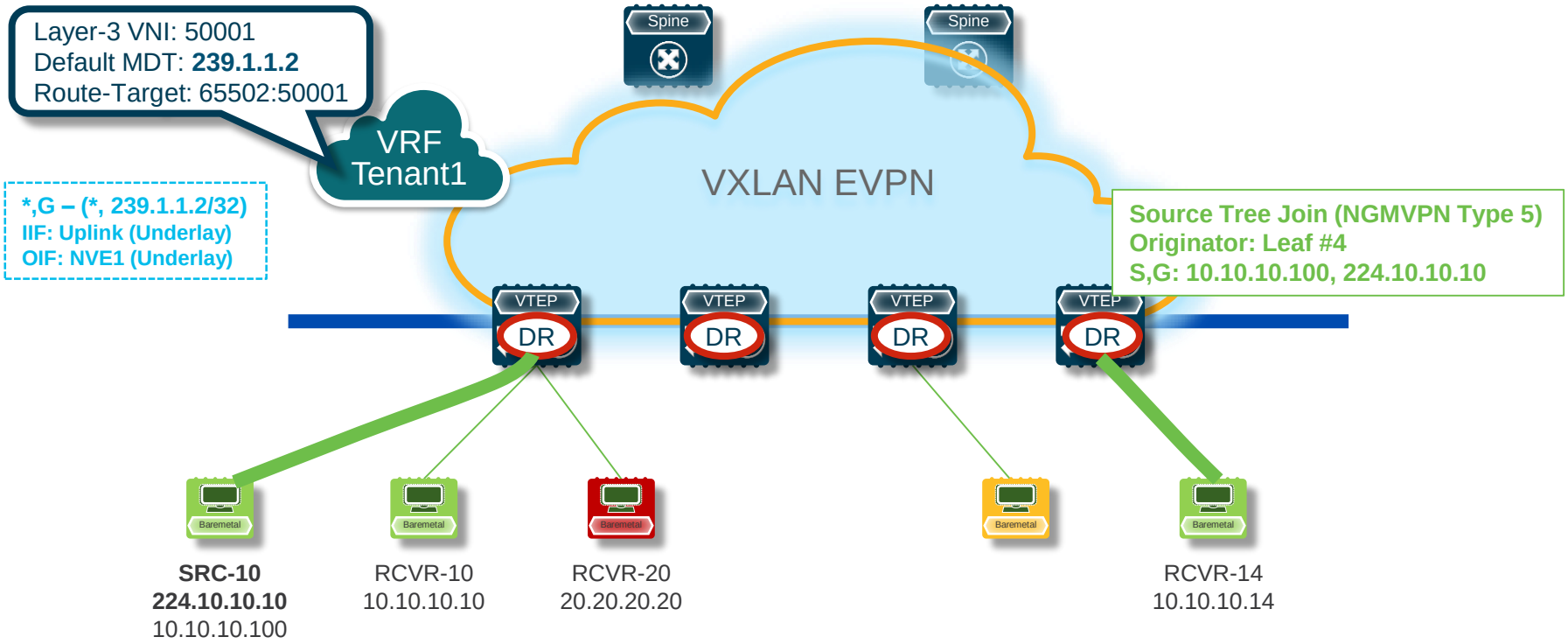
NGMVPN – Source Active Advertisement (MVPN Type 5)



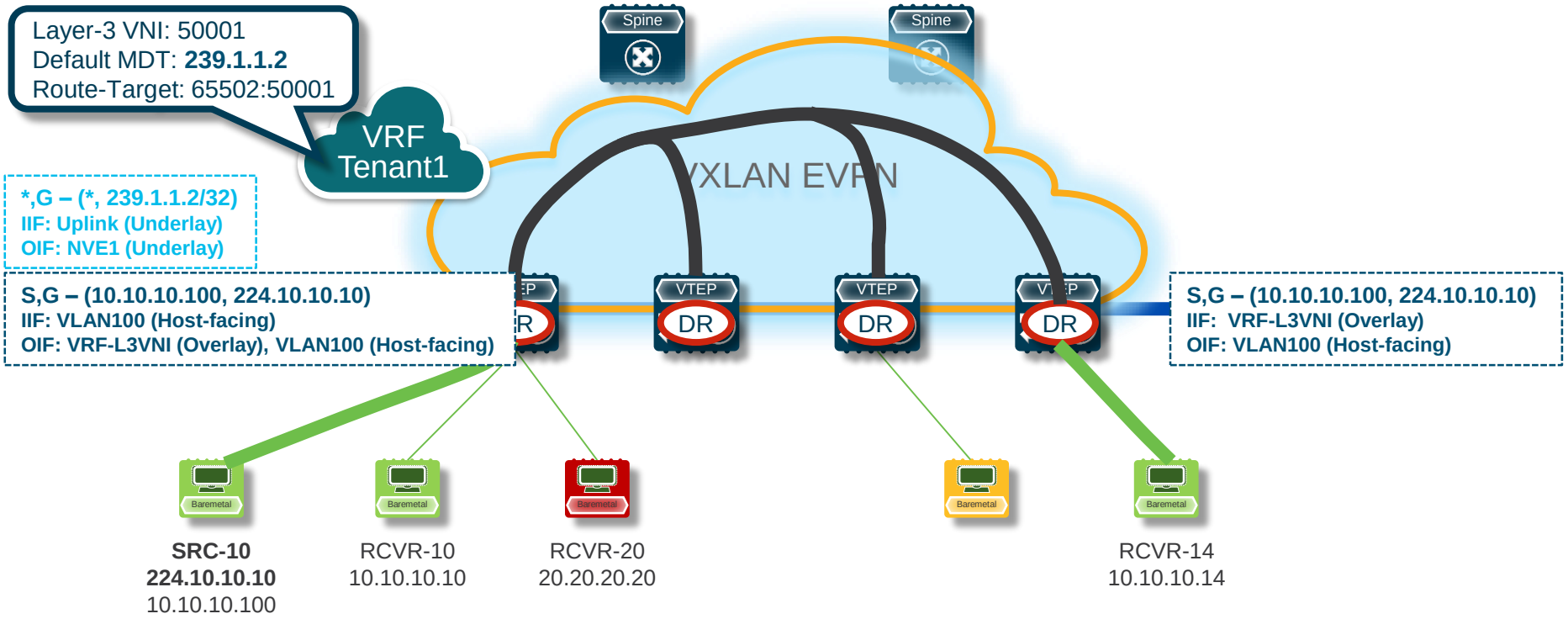
MRoute – Overlay Multicast Tree (Source Join)



NGMVPN – Source Tree Join (MVPN Type 7)



MRoute – Overlay Multicast Tree (Receiver Join)



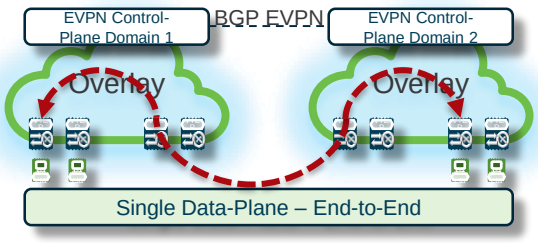
Agenda

- Introduction to Overlays
- VXLAN with BGP EVPN
  - Standards and Implementation
  - Control & Data Plane
- Tenant Routed Multicast (TRM)
- Multi-Site
- VXLAN OAM

VXLAN BGP EVPN Multi-Site

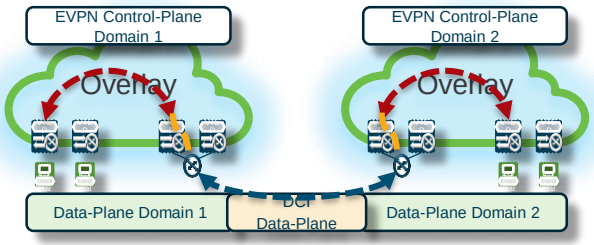
Inter-X Connectivity

VXLAN Multi-Pod



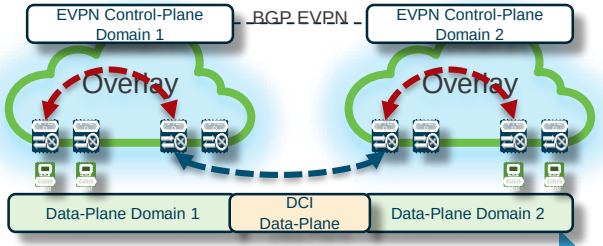
- Single Fabric with End-to-End Encapsulation
- Build Hierarchy in the Underlay – Flatten it in the Overlay

VXLAN Multi-Fabric



- Multiple Fabrics – Normalised through Ethernet
- Multiple Fabrics Interconnect using DCI (Layer 2 and Layer 3)

VXLAN Multi-Site

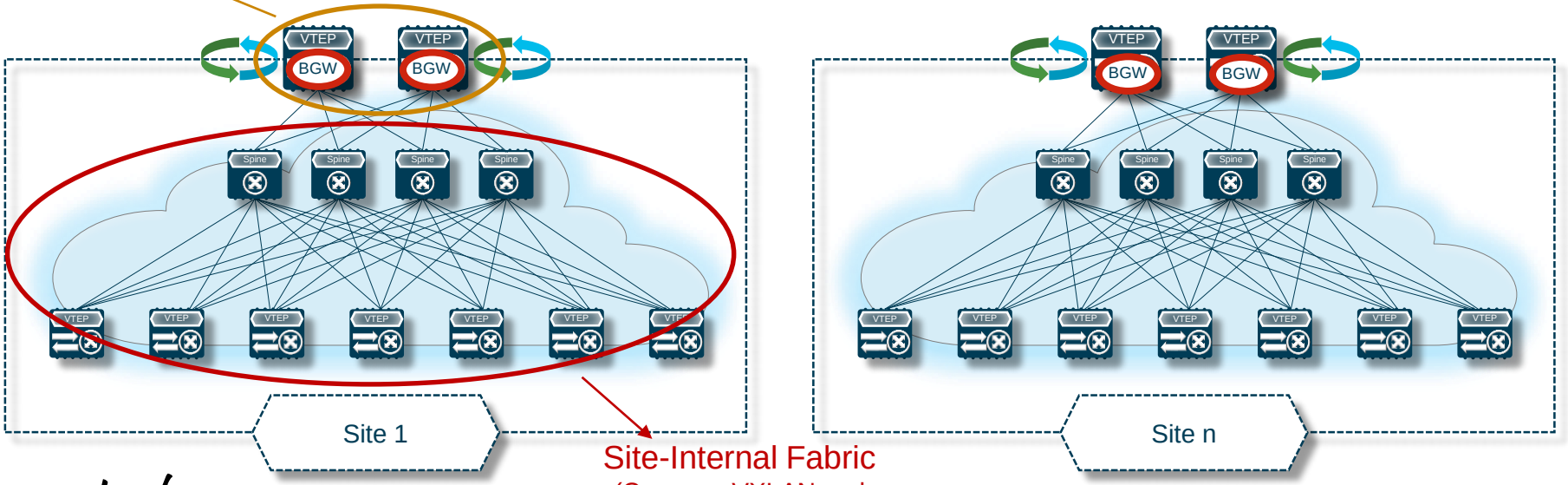


- Multiple Fabrics with Integrated DCI
- Integrated DCI – Scaling within and between Fabrics

Functional Components

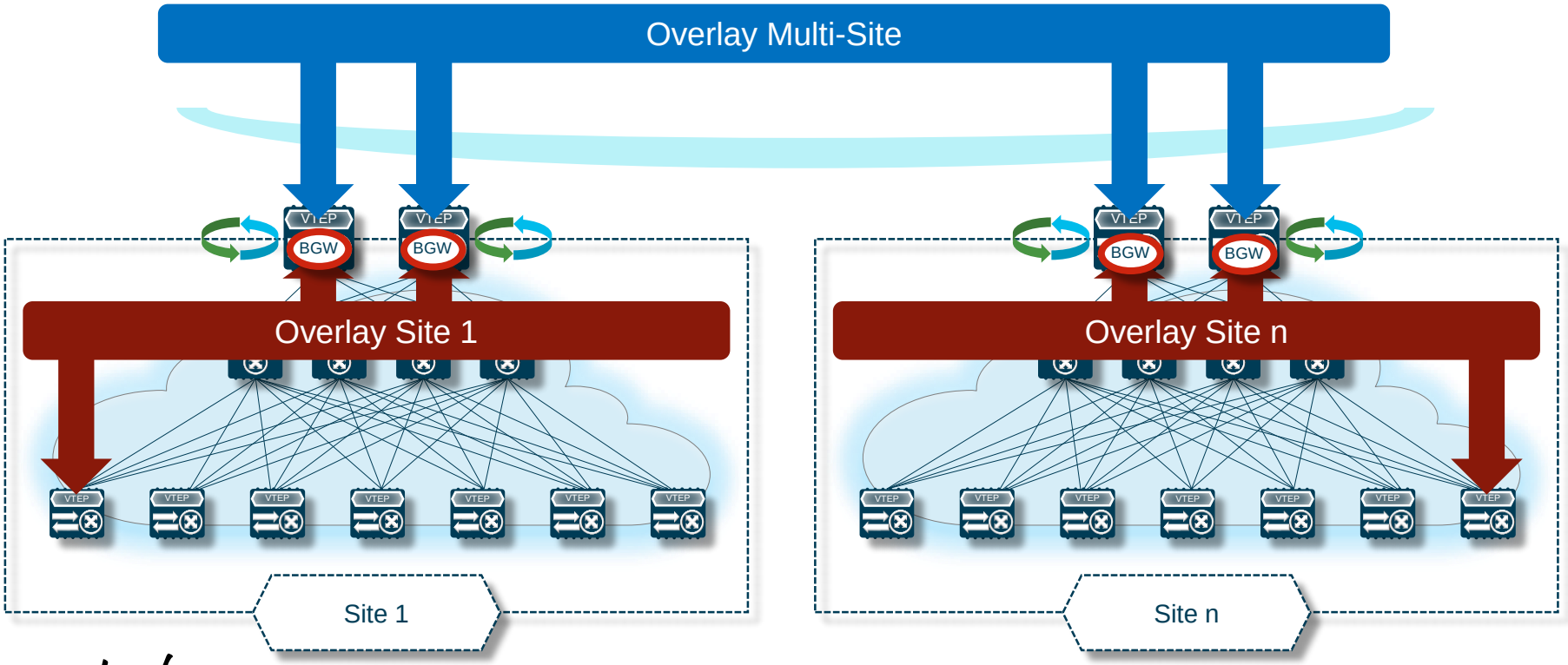
https://tools.ietf.org/html/draft-sharma-multi-site-evpn

Border Gateways
(Key Functional Components of VXLAN Multi-Site Architecture)

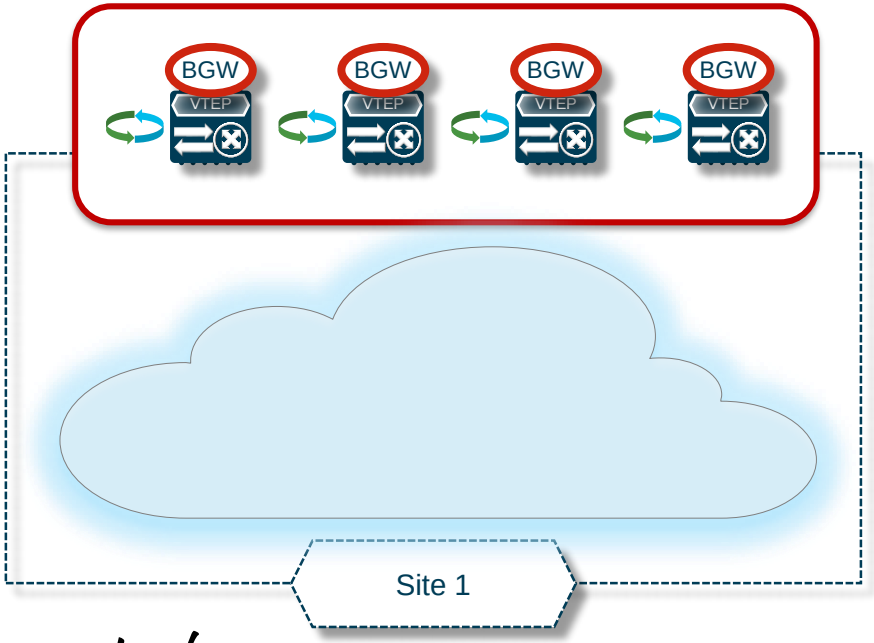


Site-Internal Fabric
(Common VXLAN and BGP-EVPN Functions)

Hierarchical Overlay Domains



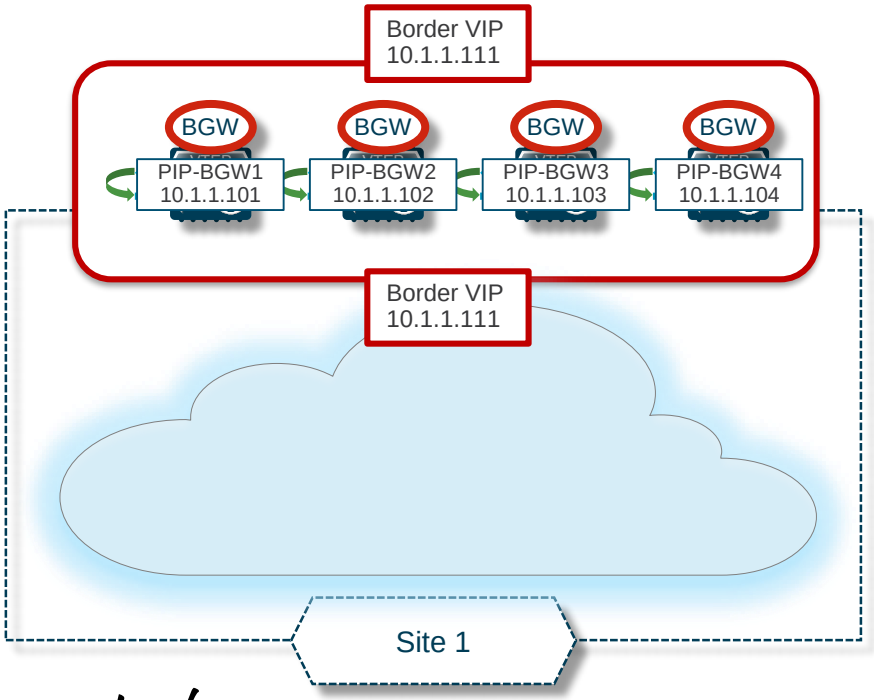
Anycast Border Gateway (1)



Anycast Border Gateway

- Up to 4 Border Gateways
- Border Gateway
  - Deploying at Leaf – 7.0(3)I7(1)
  - Deploying at Spine – 7.0(3)I7(2)

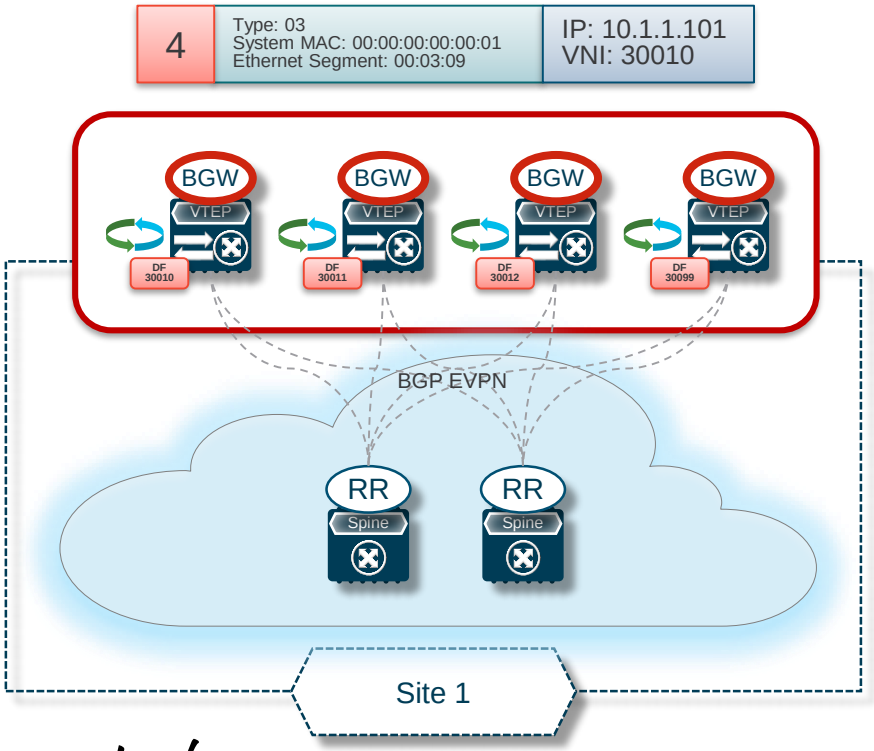
Anycast Border Gateway (2)



Anycast Border Gateway

- Common Virtual IP (VIP) across BGW
  - VIP for communication between the Border Gateways in **different Sites**
  - VIP for communication between Border Gateway and Leaf **within a Site**
- Individual Primary IP (PIP) per BGW
  - Used for Broadcast, Unknown Unicast and Multicast (BUM) replication
  - PIP for communication with Single-Homed endpoints (routed only), intra- and inter-Site

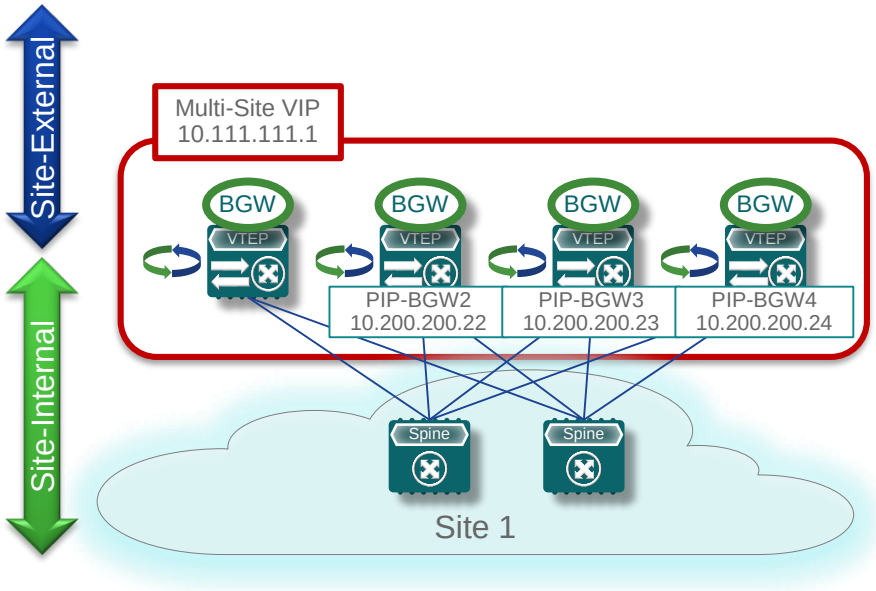
Anycast Border Gateway (3)



Anycast Border Gateway

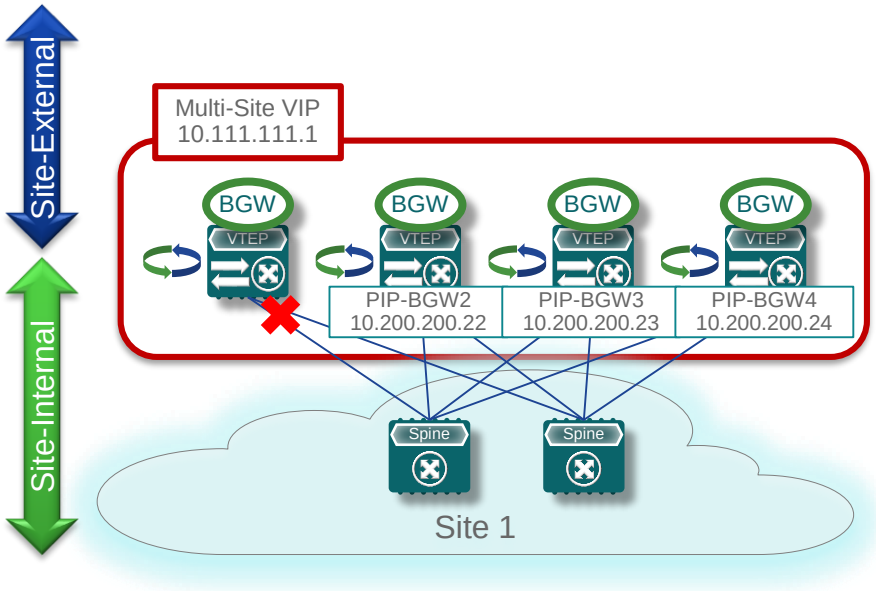
- Per-VNI Designated Forwarder (DF) election
  - Each BGW can serve as DF for a single or a set of Layer-2 VNIs
  - DF election and assignment is automatic
- Using BGP EVPN Route Type 4 for DF election
  - Operator Managed Assignment (Type: 03)
  - Six Octet Site Identifier (System MAC: 00:00:00:00:00:01)
  - Multi-Site Discriminator (Ethernet-Segment: 00:03:09)
  - Originators IP Address (PIP): 10.1.1.101
  - Layer-2 VNI: 30010

Failure Detection on BGWs – Fabric Isolation (1)



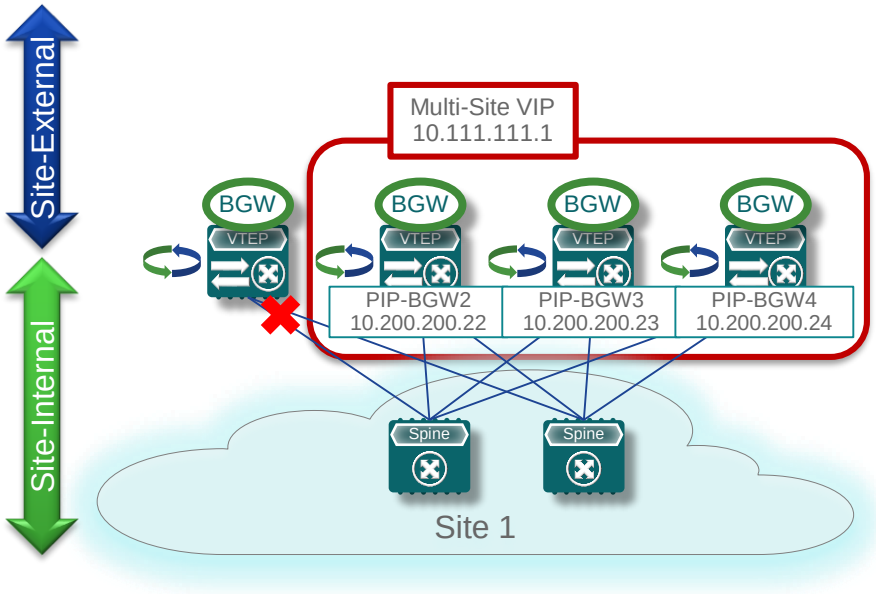
- The Site-Internal interfaces on BGW nodes are constantly tracked to determine their status ('**evpn multisite fabric-tracking**' command)

Failure Detection on BGWs – Fabric Isolation (2)



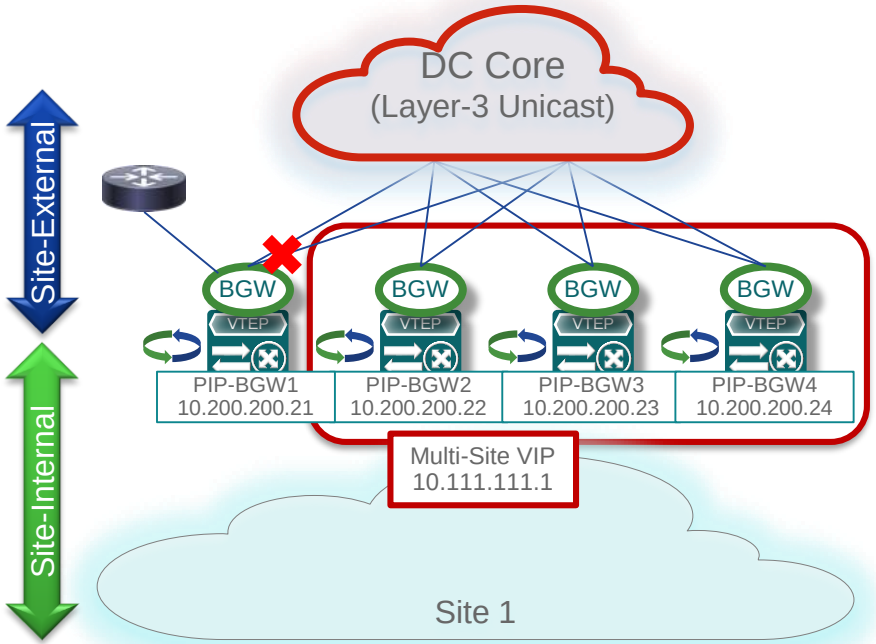
- The Site-Internal interfaces on BGW nodes are constantly tracked to determine their status ('**evpn multisite fabric-tracking**' command)
- If all the Site-Internal interfaces are detected as down:
  1. The isolated BGW stops advertising PIP/VIP addresses toward the Site-External network
  2. The remaining BGWs perform new DF elections for the L2VNIs owned by the isolated BGW

Failure Detection on BGWs – Fabric Isolation (3)



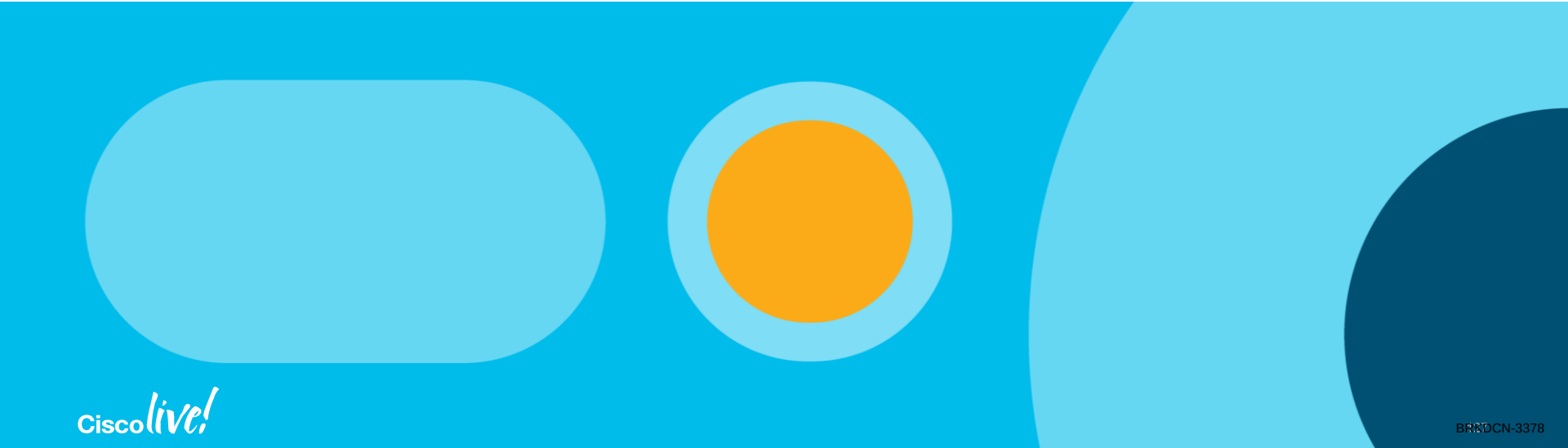
- The Site-Internal interfaces on BGW nodes are constantly tracked to determine their status ('**evpn multisite fabric-tracking**' command)
- If all the Site-Internal interfaces are detected as down:
  1. The isolated BGW stops advertising PIP/VIP addresses toward the Site-External network
  2. The remaining BGWs perform new DF elections for the L2VNIs owned by the isolated BGW
- As a result, the BGW becomes isolated from both the Site-Internal and Site-External networks
- Seamless BGW node reinsertion using a "delay-restore" timer for the VIP address

Failure Detection on BGWs – DCI Isolation



- The Site-External interfaces on BGW nodes are also tracked to determine their status ('**evpn multisite dci-tracking**' command)
- If all the Site-External interfaces are detected as down, the isolated BGW node:
  1. Stops advertising VIP VTEP address toward the Site-Internal network
  2. Withdraws BGP EVPN Type-4 advertisements (triggering a new DF election between other BGWs)
  3. Starts functioning as a regular VTEP (PIP still up)
- As a result, the BGW continues to operate as a Site-Internal VTEP
- Seamless BGW node reinsertion using a "delay-restore" timer for the VIP address

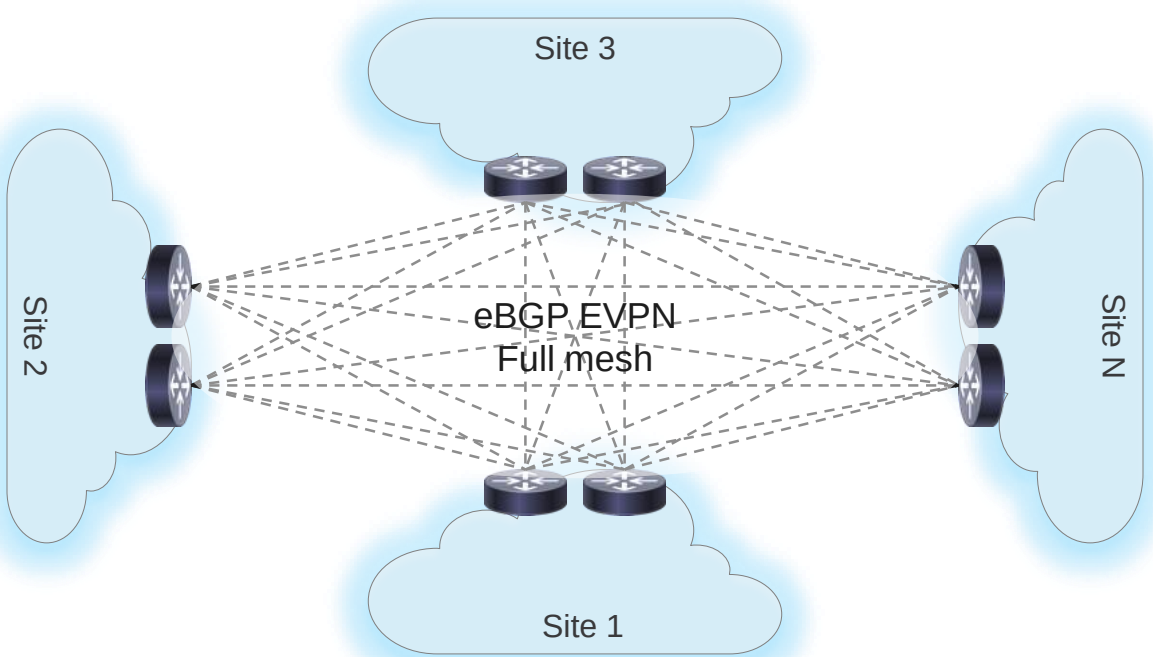
Multi-Site Control- & Data-Plane



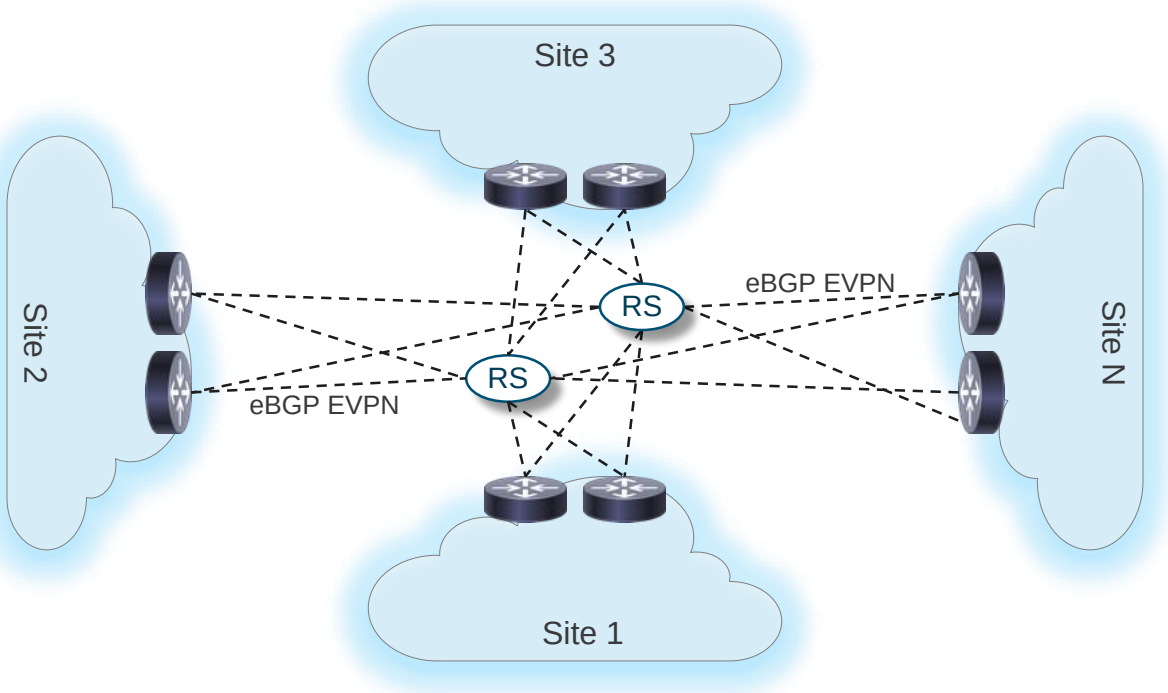
Multi-Site Control Plane Deployment Considerations

- MP-eBGP EVPN only inter-Sites
  - Next-hop behaviour (VXLAN tunnel termination and reorigination) and loop protection (as-path attribute)
- Two main options for underlay and overlay control plane deployment
  1. I-E-I (Recommended)
     - Intra-Site: IGP (OSPF, IS-IS) as underlay CP, iBGP as overlay CP
     - Inter-Sites: eBGP for both underlay and overlay CPs
  2. E-E-E
     - Intra-Site and Inter-Sites: eBGP for both underlay and overlay CPs
- Full mesh of MP-eBGP EVPN adjacencies across sites
  - Recommended to deploy a couple of **Route-Servers** with 3 or more sites
  - RS in a separate AS only perform control plane functions (“eBGP Route-Reflectors”, IETF RFC 7947)
  - RS functions: EVPN routes reflection, next-hop-unchanged, route-target rewrite

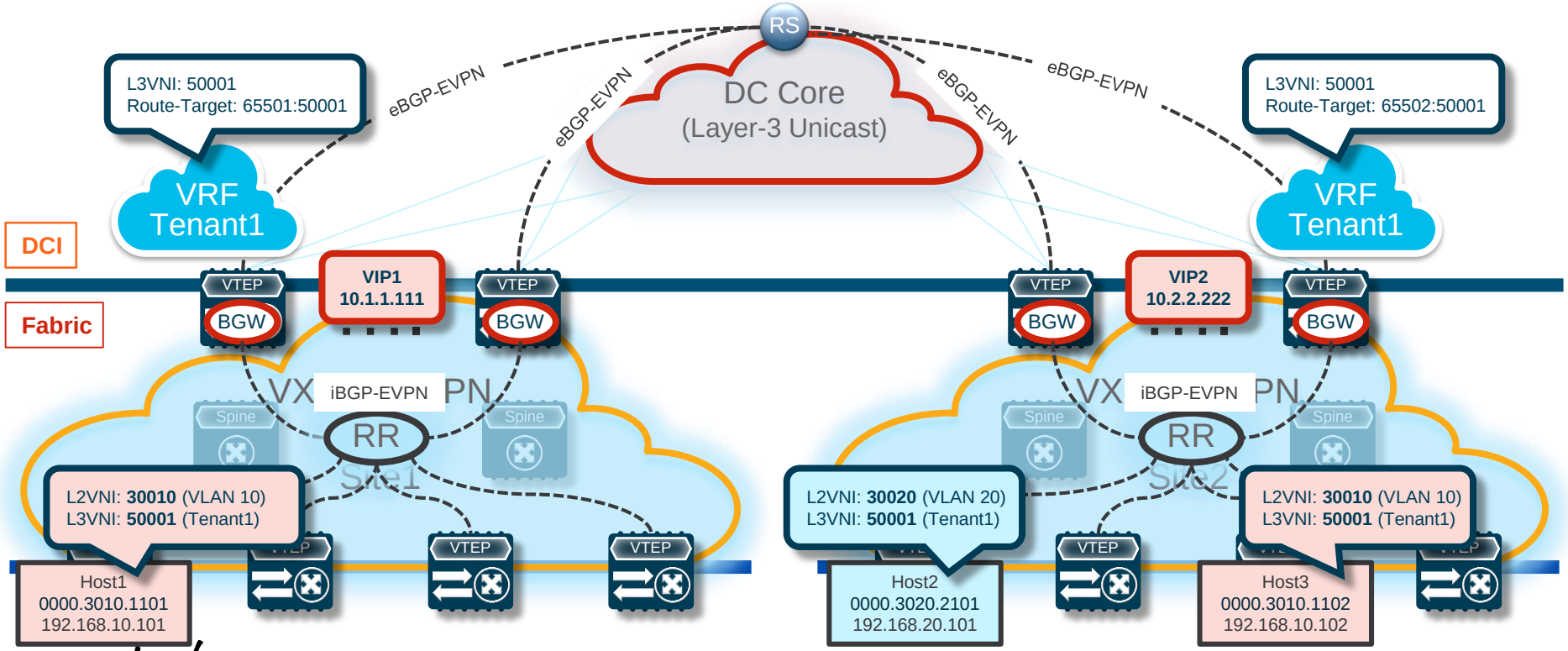
Multi-Site Overlay Control Plane – back-to-back



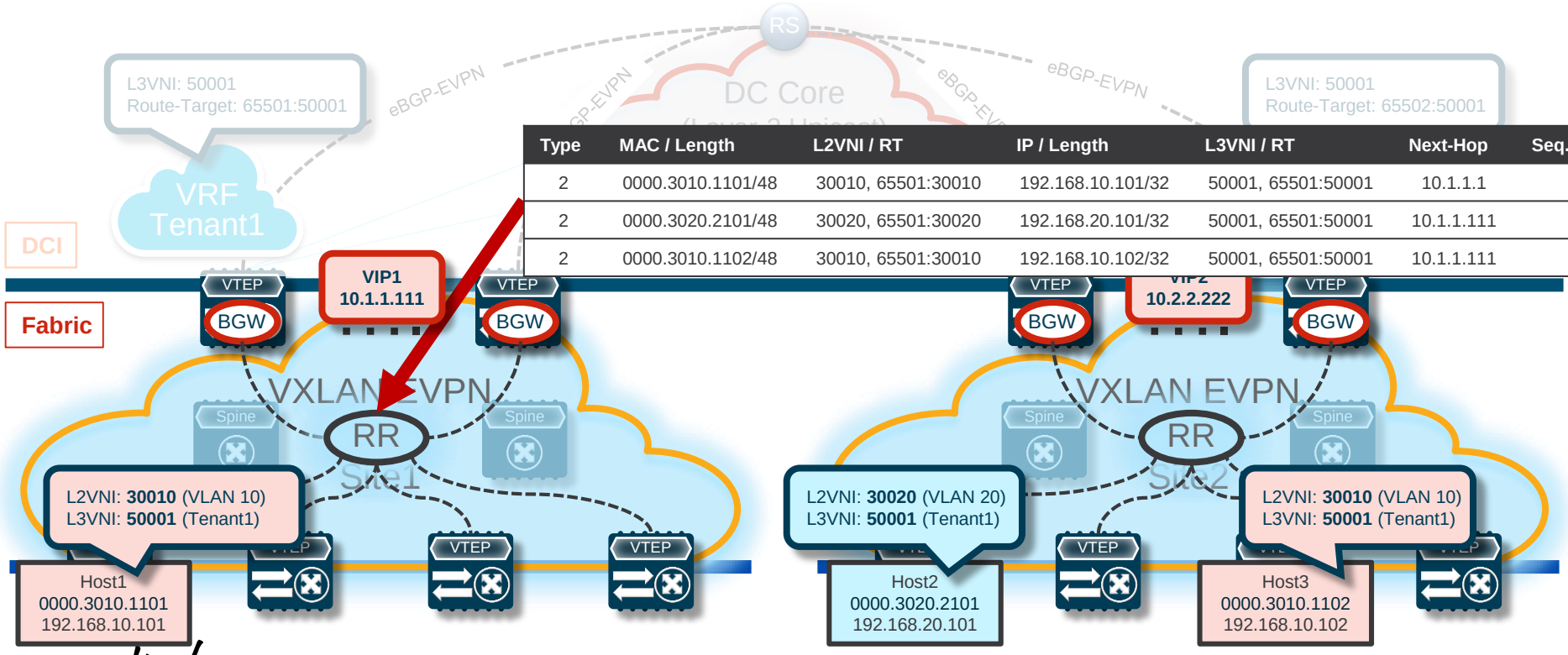
Multi-Site Overlay Control Plane – Route-Server



Multi-Site Overlay Control Plane – Tenants



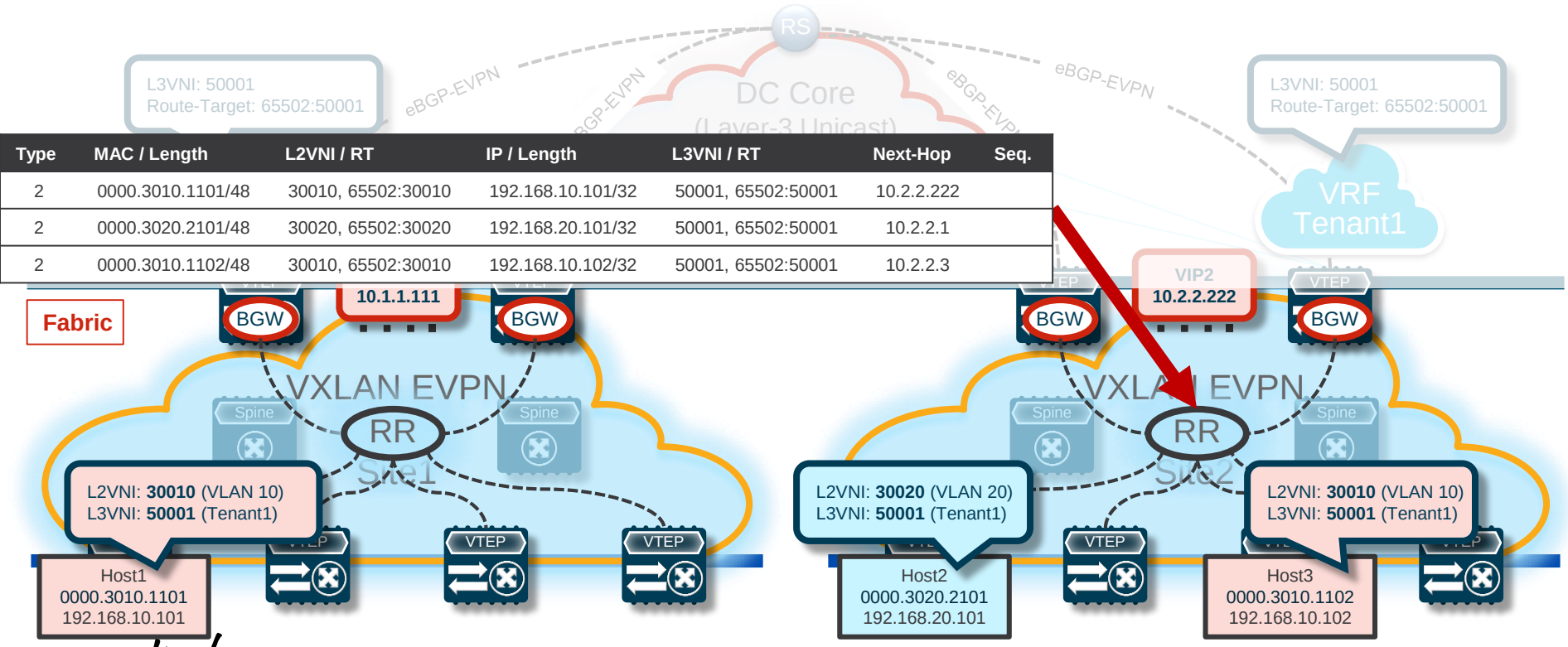
Multi-Site Overlay Control Plane – Site1



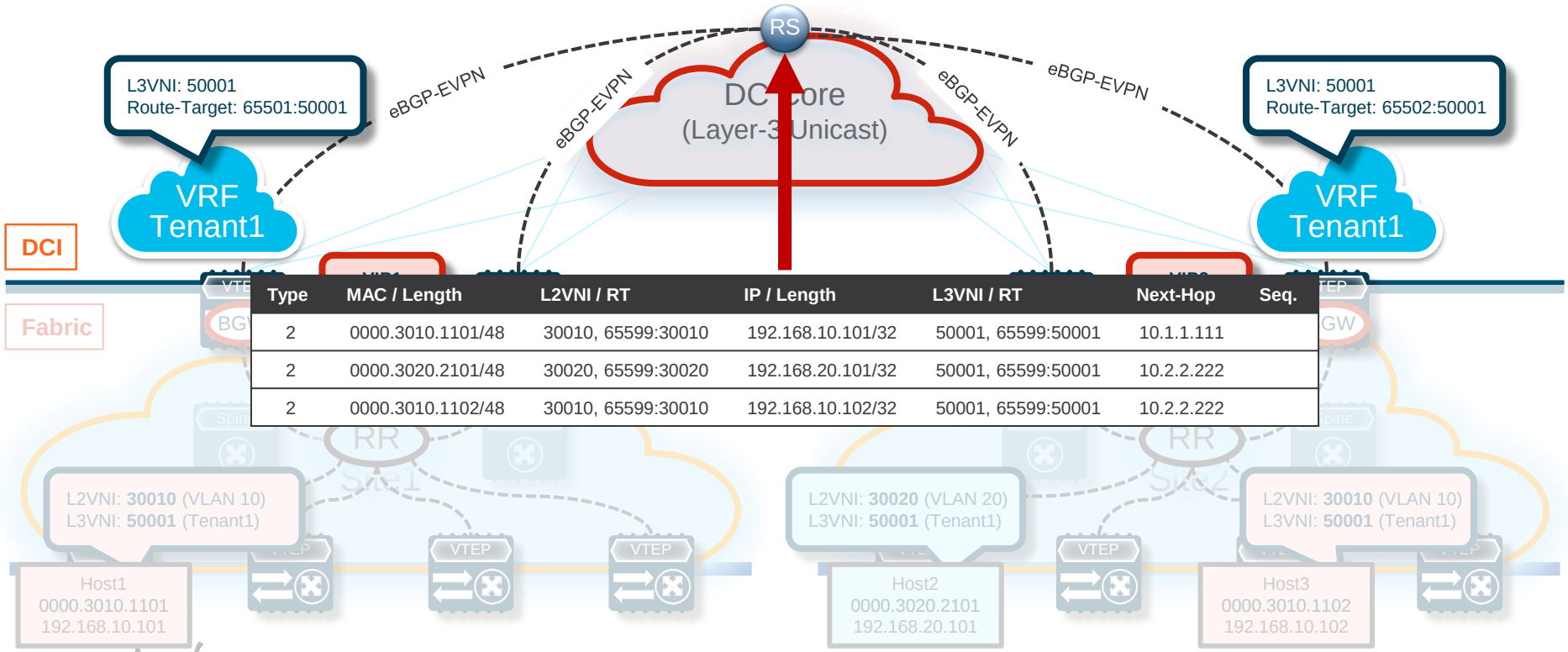
Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3010.1101/48	30010, 65501:30010	192.168.10.101/32	50001, 65501:50001	10.1.1.1	
2	0000.3020.2101/48	30020, 65501:30020	192.168.20.101/32	50001, 65501:50001	10.1.1.111	
2	0000.3010.1102/48	30010, 65501:30010	192.168.10.102/32	50001, 65501:50001	10.1.1.111	

Multi-Site Overlay Control Plane – Site2

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3010.1101/48	30010, 65502:30010	192.168.10.101/32	50001, 65502:50001	10.2.2.222	
2	0000.3020.2101/48	30020, 65502:30020	192.168.20.101/32	50001, 65502:50001	10.2.2.1	
2	0000.3010.1102/48	30010, 65502:30010	192.168.10.102/32	50001, 65502:50001	10.2.2.3	

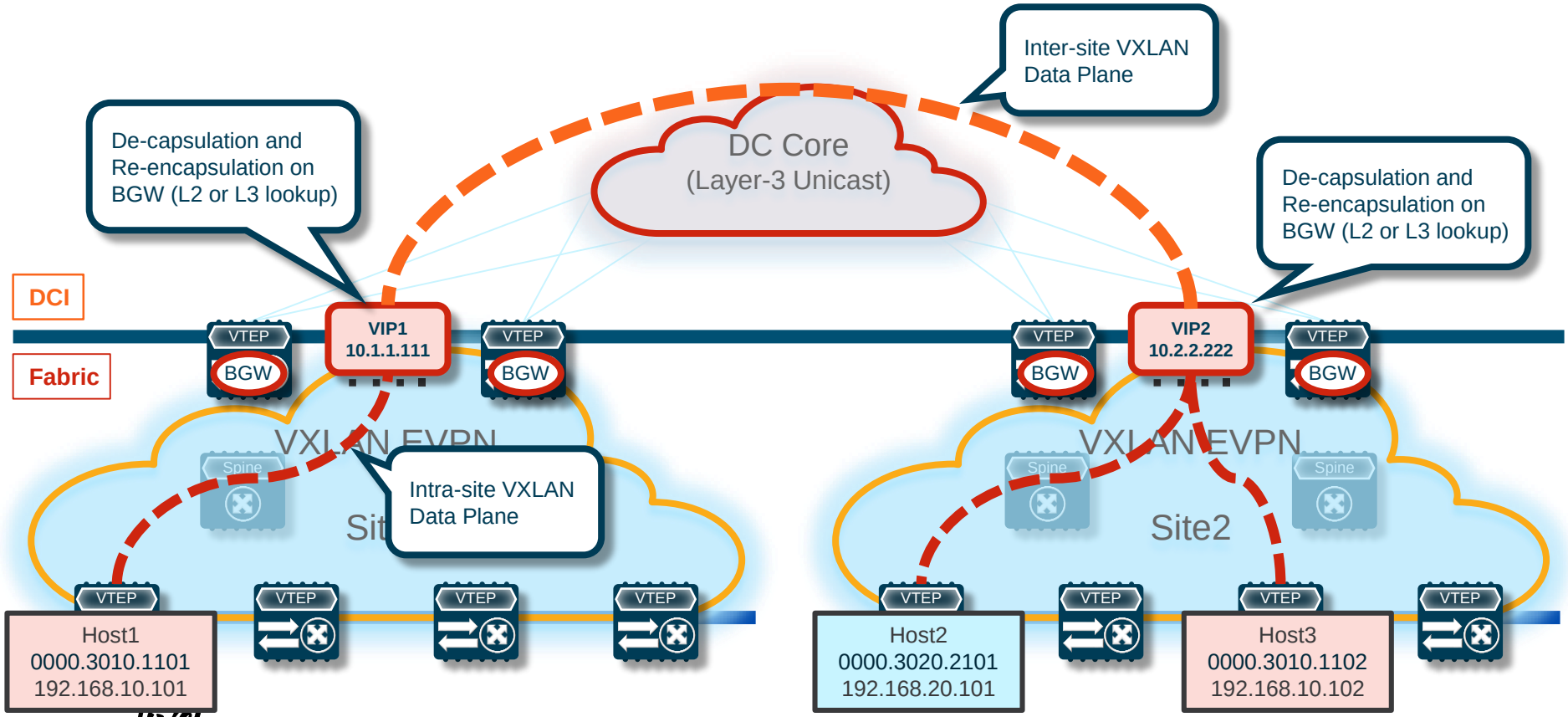


Multi-Site Overlay Control Plane – Between Sites

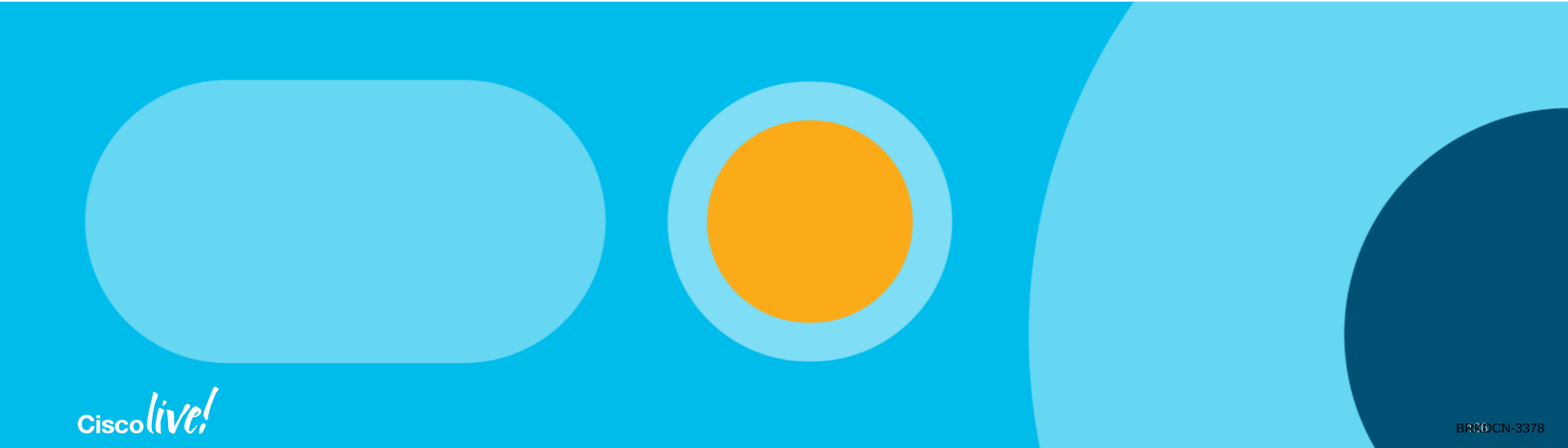


Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3010.1101/48	30010, 65599:30010	192.168.10.101/32	50001, 65599:50001	10.1.1.111	
2	0000.3020.2101/48	30020, 65599:30020	192.168.20.101/32	50001, 65599:50001	10.2.2.222	
2	0000.3010.1102/48	30010, 65599:30010	192.168.10.102/32	50001, 65599:50001	10.2.2.222	

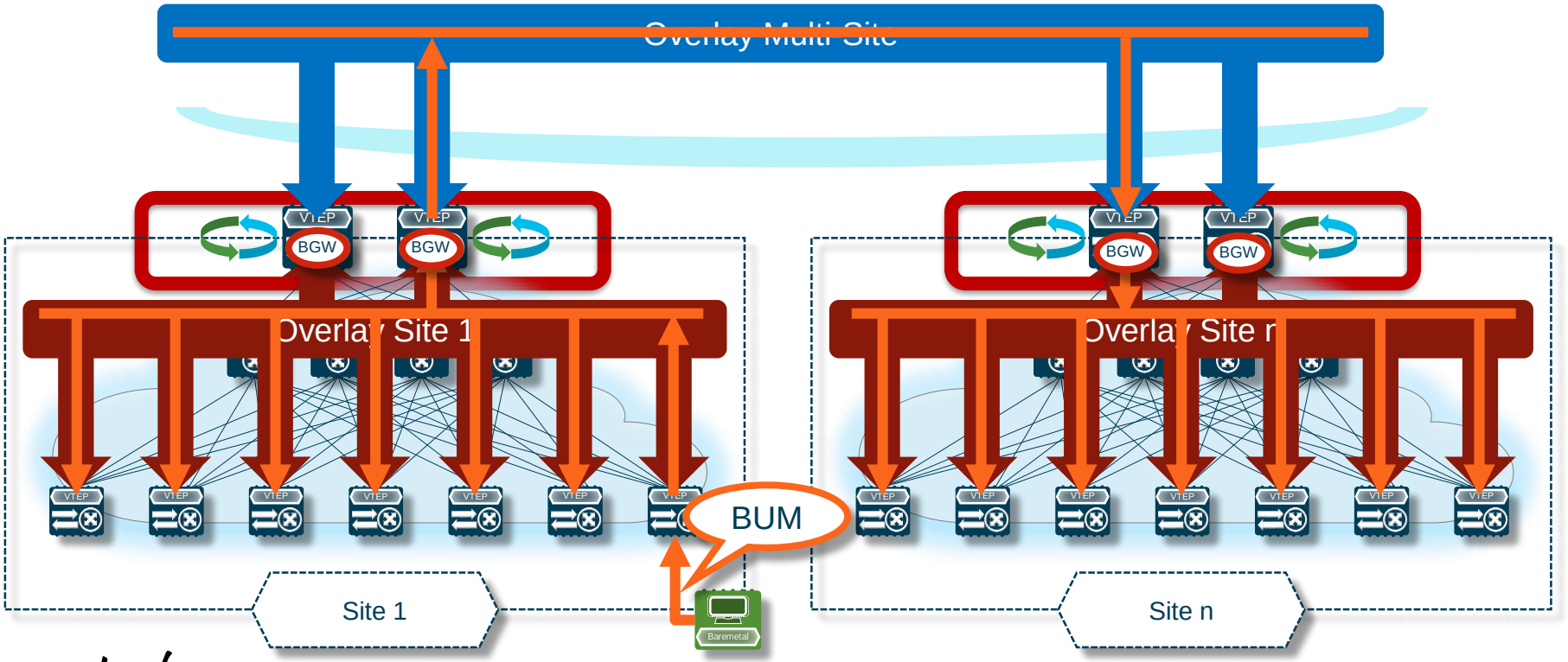
Multi-Site Overlay Data Plane – Overview



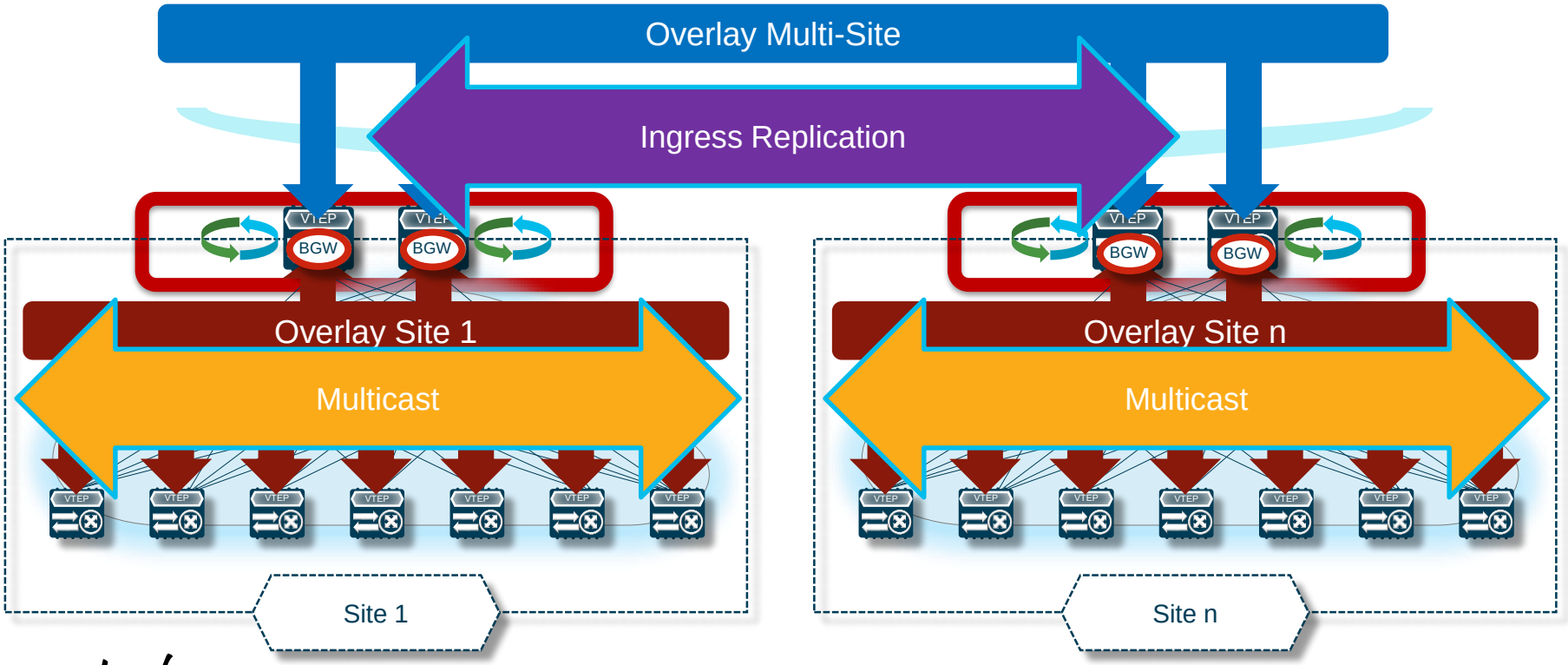
Multi-Site Packet Walk (BUM)



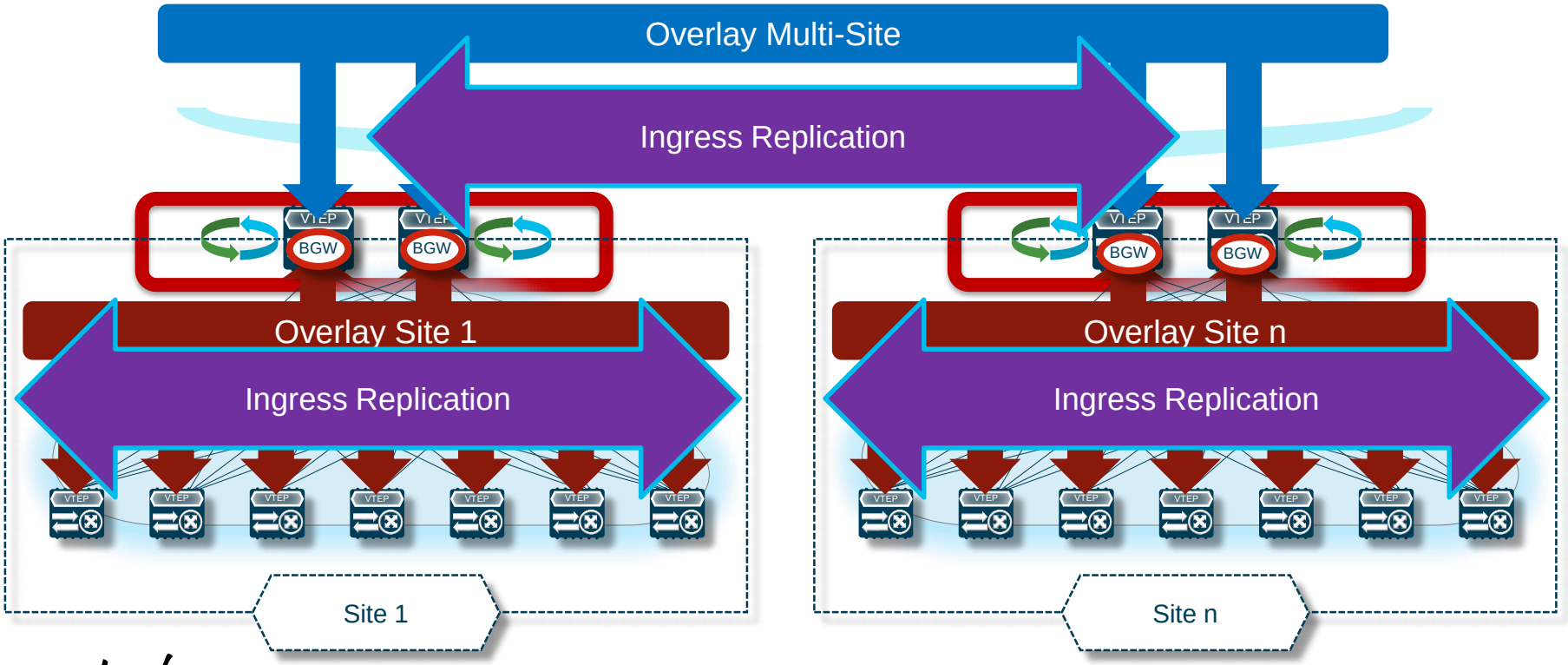
Multi-Site – BUM Traffic Distribution



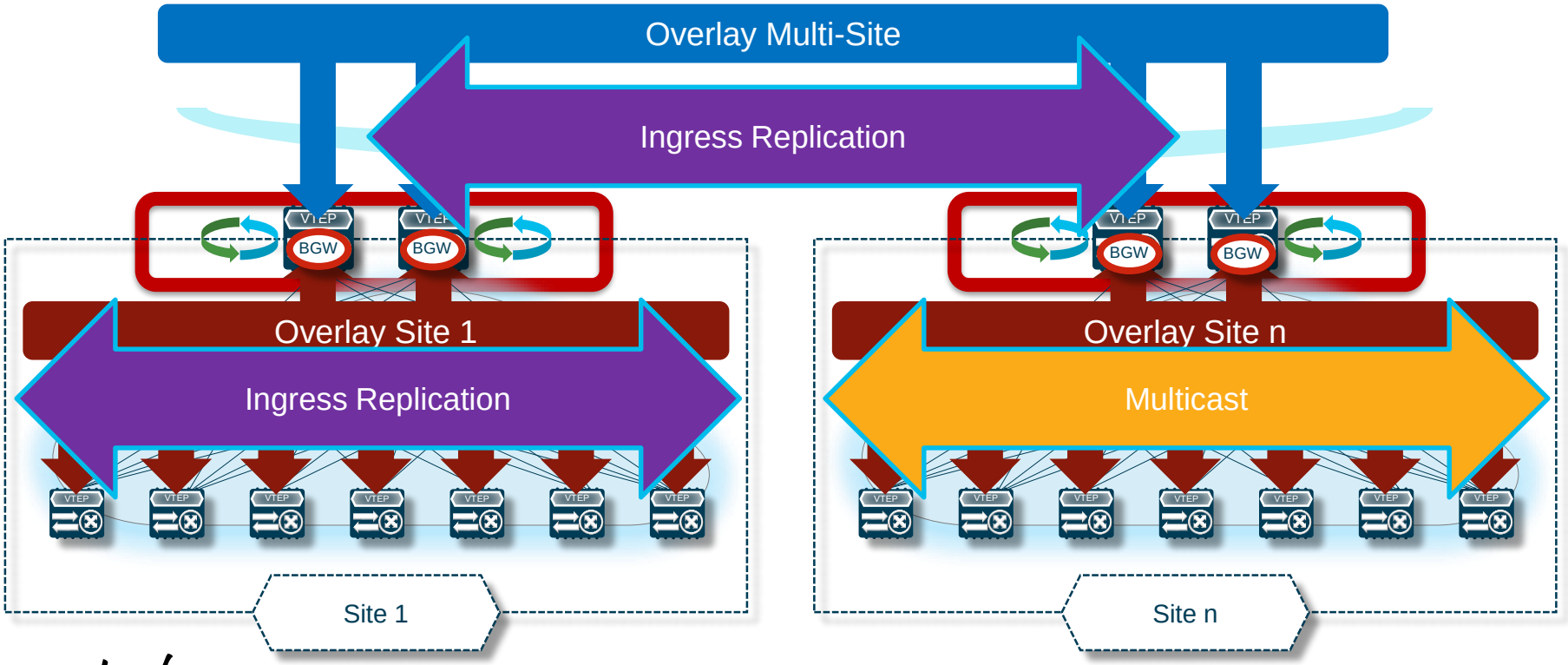
Multi-Site – BUM Replication Modes (Multicast Sites)



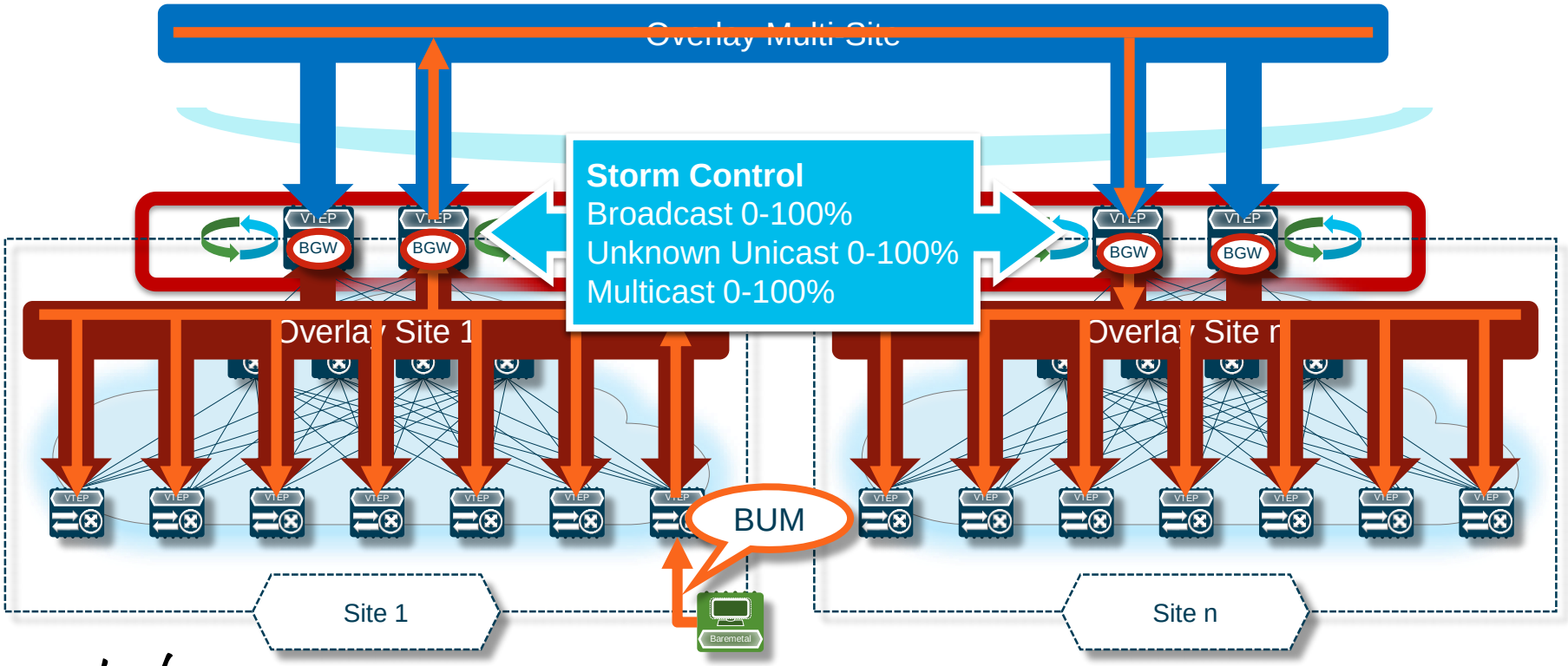
Multi-Site – BUM Replication Modes (All Ingress Replication)



Multi-Site – BUM Replication Modes (Mixed Site)



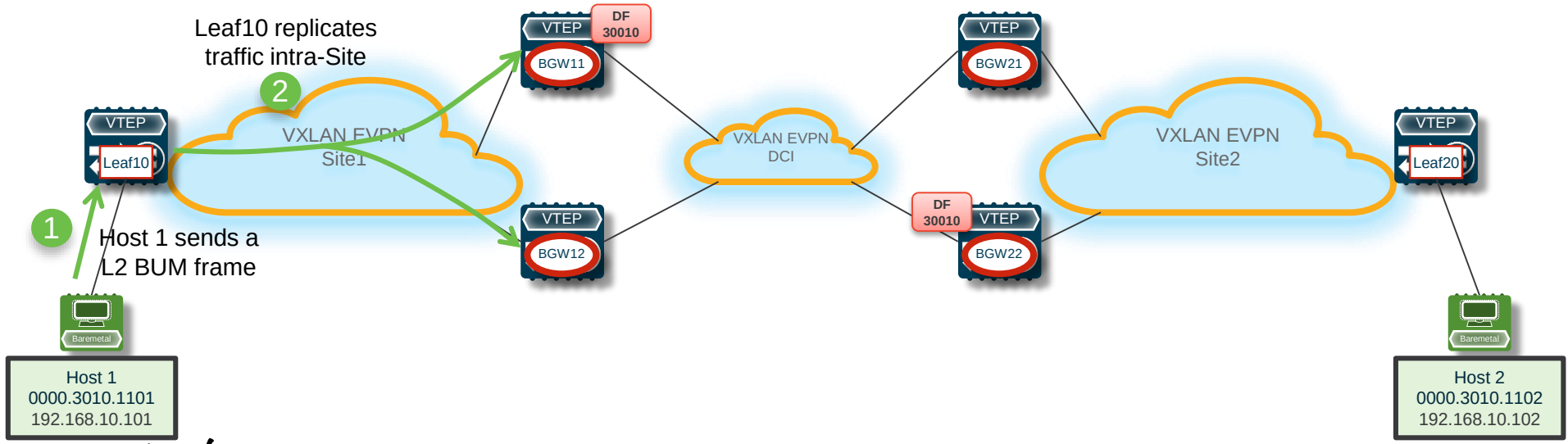
Multi-Site – BUM Traffic Enforcement



Layer 2 (BUM) – Site 1

Bridge

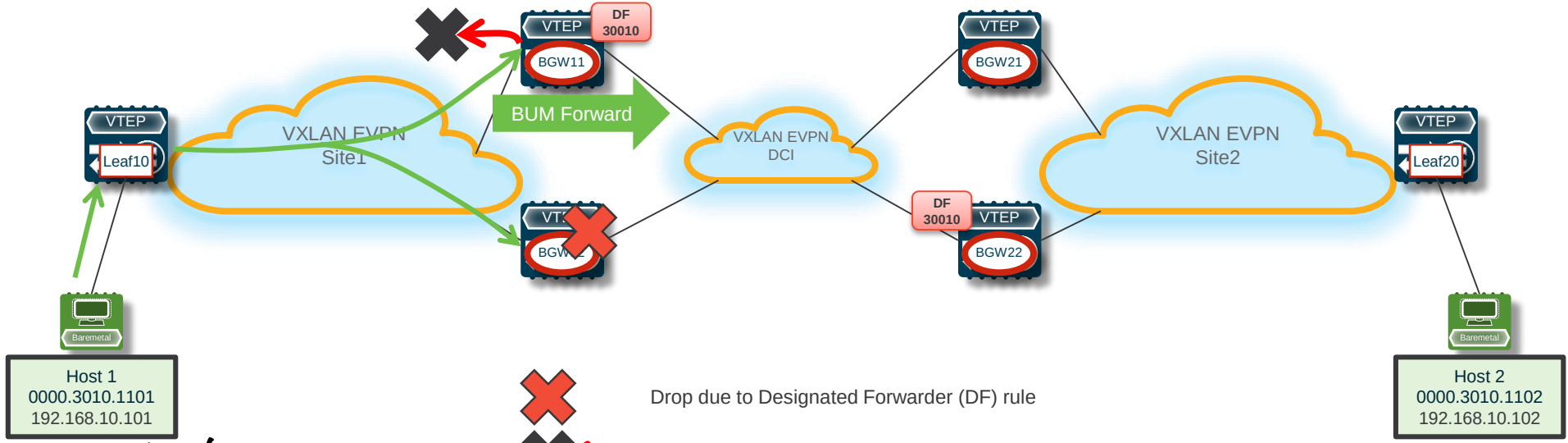
SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
L10	DGROUP	30010	H1-MAC	ALL-F	H1-IP	ALL-255	



Layer 2 (DF and Split Horizon) – Site 1

Bridge

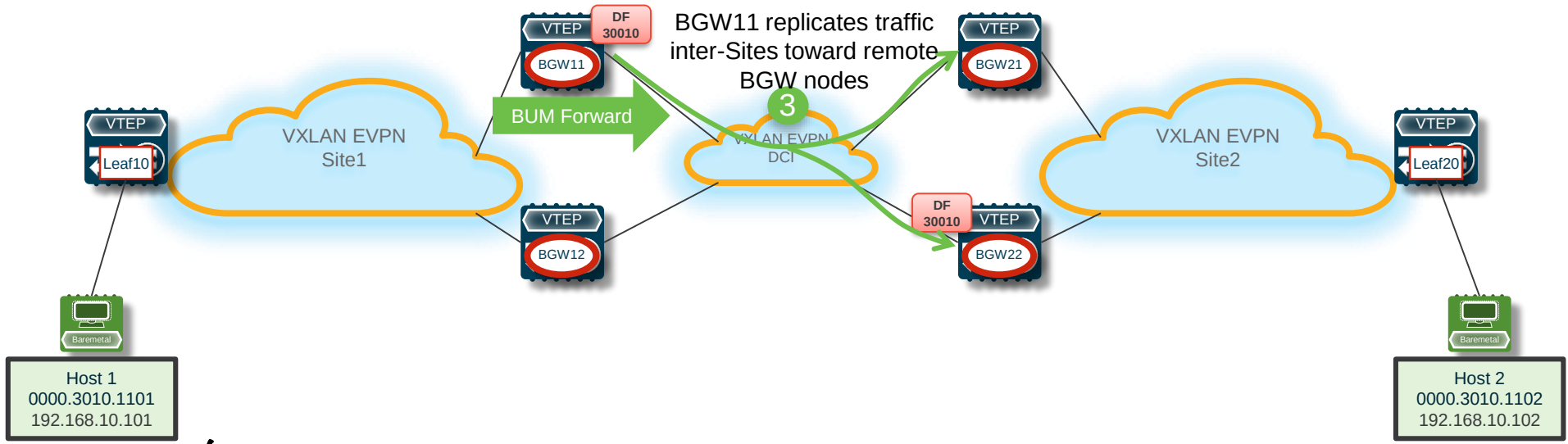
SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
L10	DGROUP	30010	H1-MAC	ALL-F	H1-IP	ALL-255	



Layer 2 (BUM) – DCI

Bridge

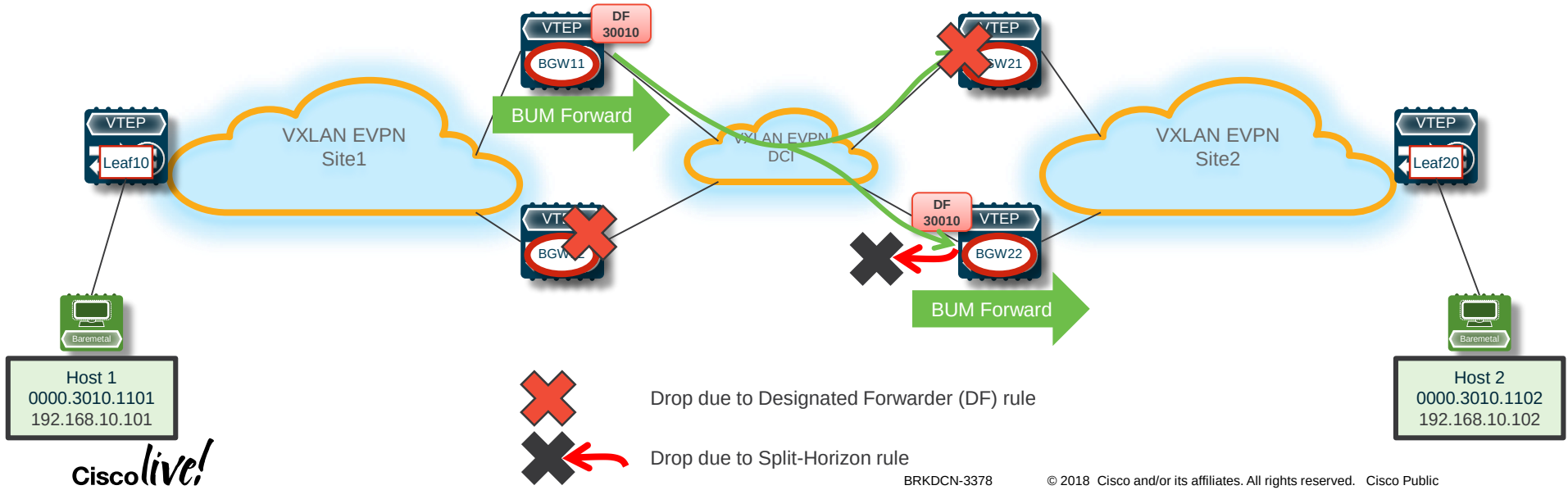
SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
BGW11	BGW21	30010	H1-MAC	ALL-F	H1-IP	ALL-255	
BGW11	BGW22	30010	H1-MAC	ALL-F	H1-IP	ALL-255	



Layer 2 (DF and Split Horizon) – DCI

Bridge

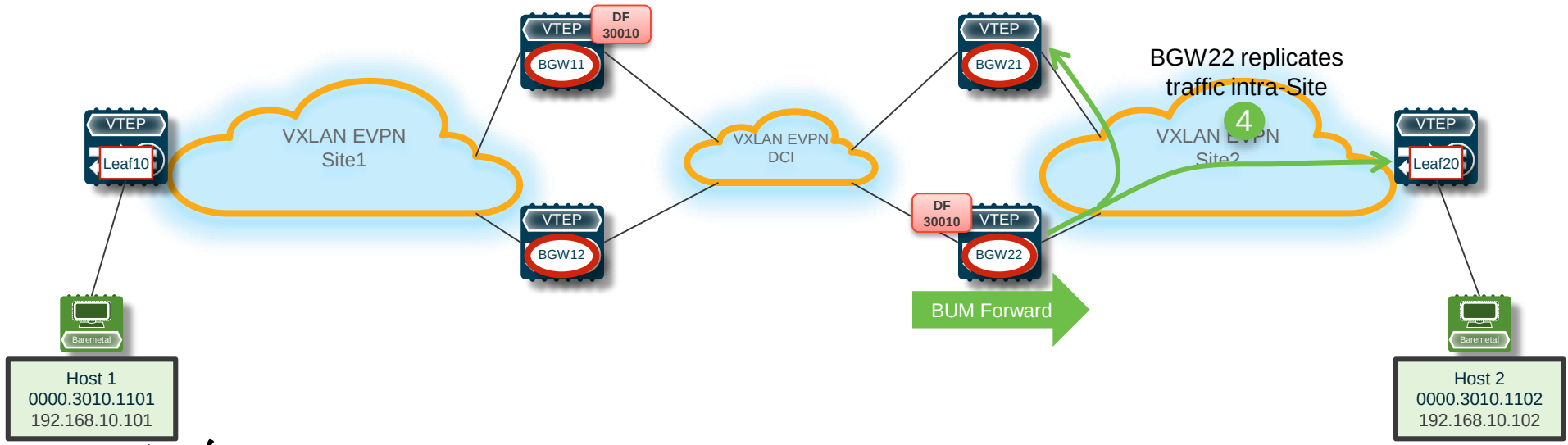
SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
BGW11	BGW21	30010	H1-MAC	ALL-F	H1-IP	ALL-255	
BGW11	BGW22	30010	H1-MAC	ALL-F	H1-IP	ALL-255	



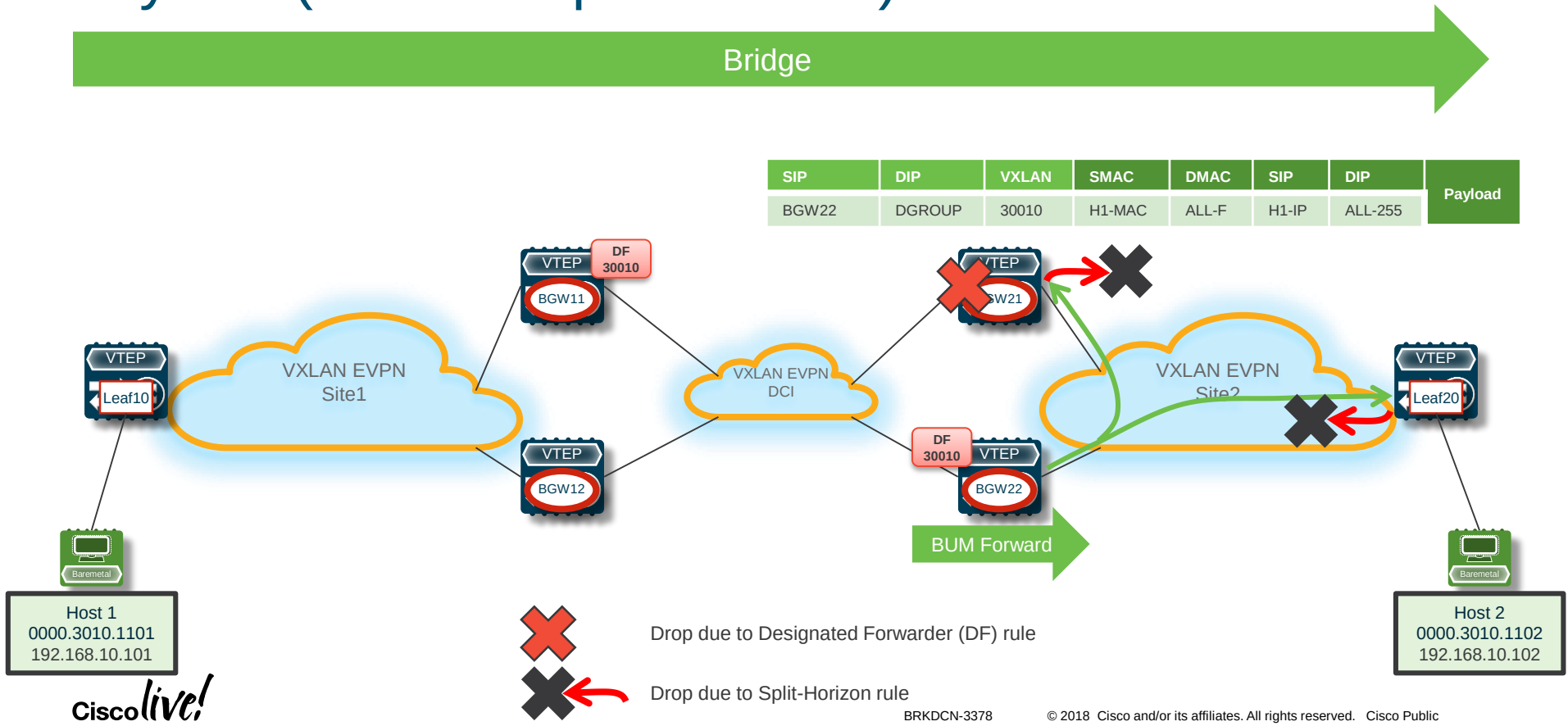
Layer 2 (BUM) – Site 2

Bridge

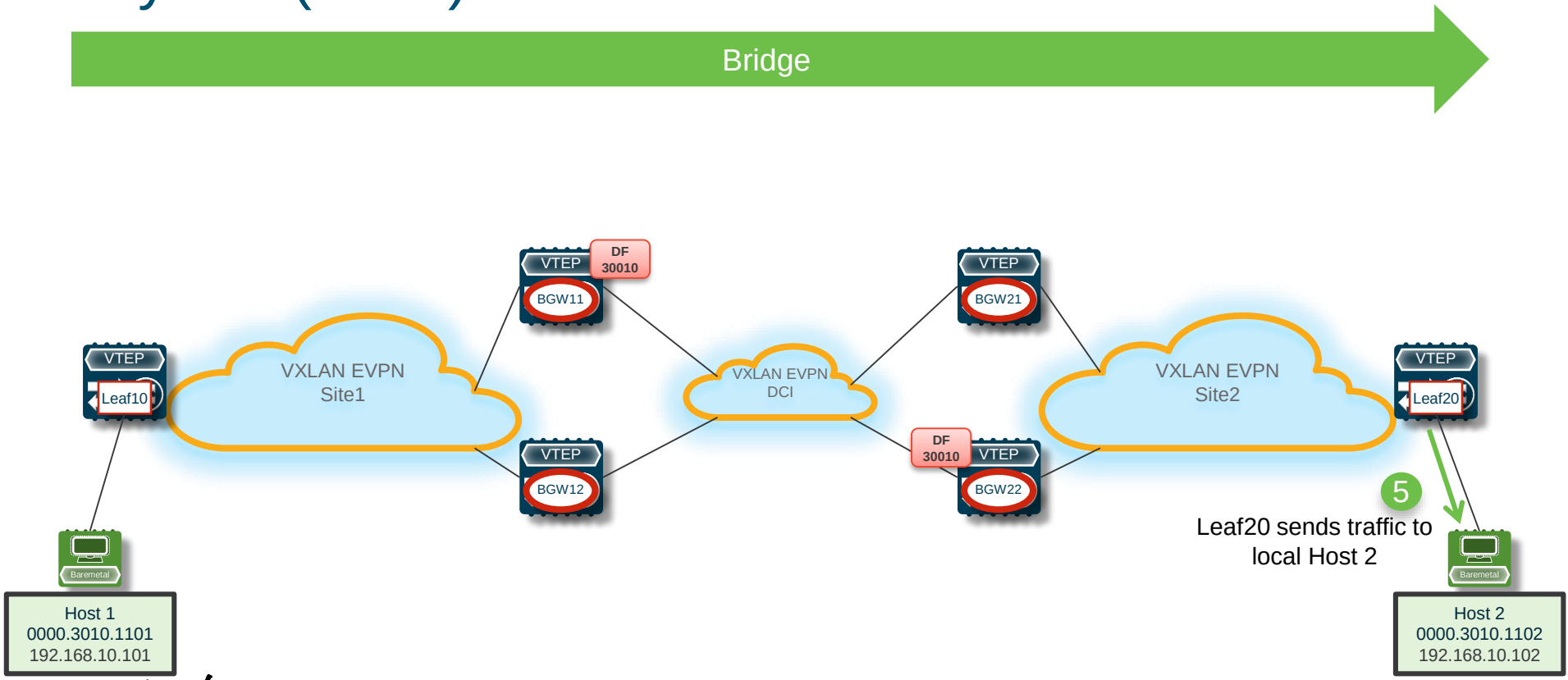
SIP	DIP	VXLAN	SMAC	DMAC	SIP	DIP	Payload
BGW22	DGROUP	30010	H1-MAC	ALL-F	H1-IP	ALL-255	



Layer 2 (DF and Split Horizon) – Site 2



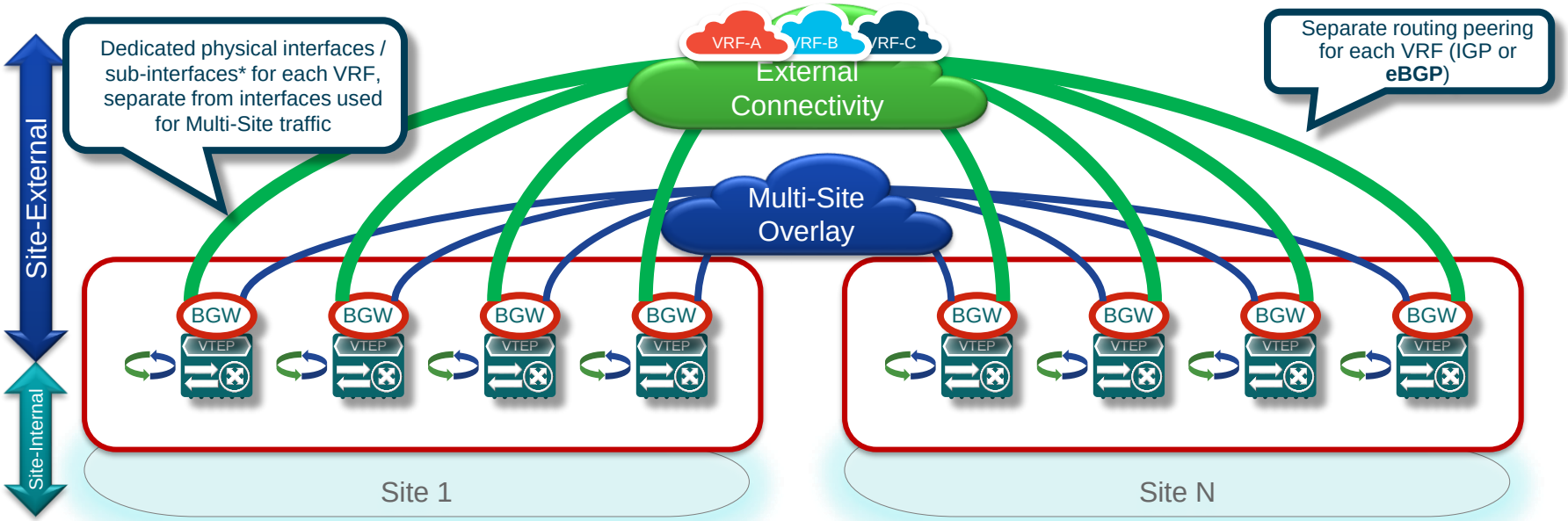
Layer 2 (BUM) – Site 2



Multi-Site and External Layer 3 Connectivity

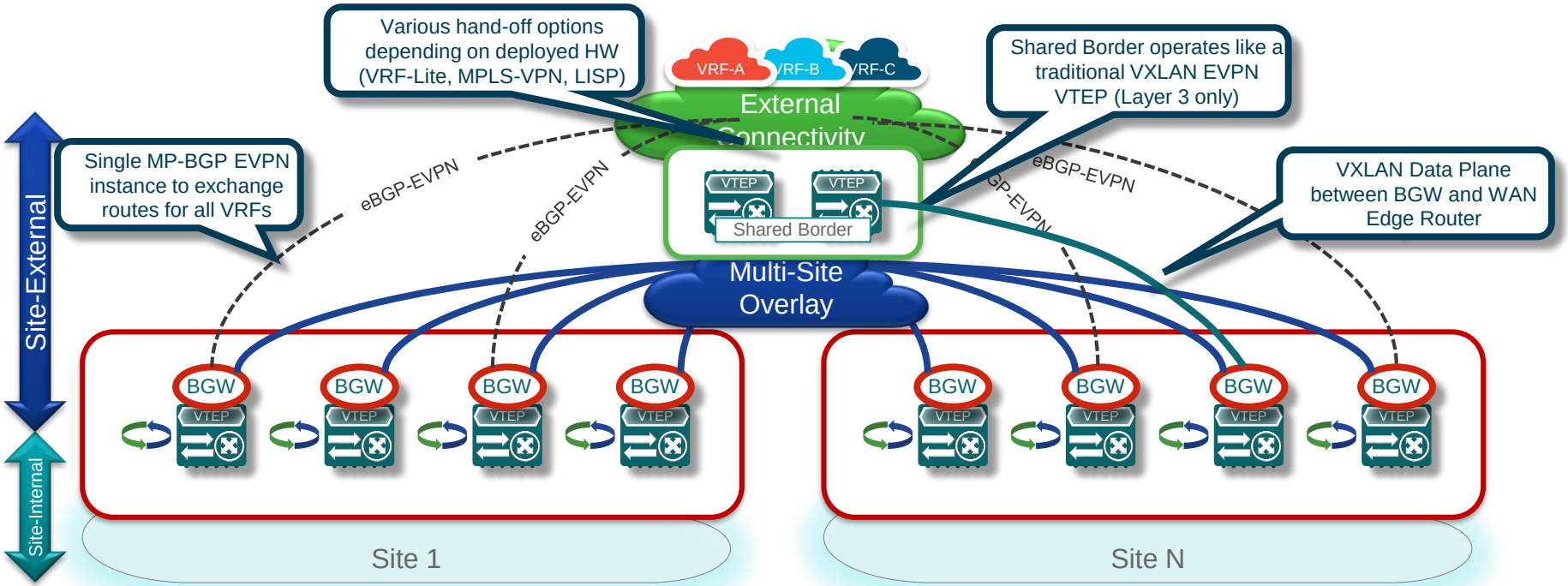
- The BGW nodes can also be used to provide Layer-3 external connectivity to each site
- Different connectivity models are supported
  - VRF-Lite peering with external WAN Edge routers
  - MP-BGP EVPN peering with external WAN Edge routers (Shared Border deployment model, aka GOLF)
  - Dedicated or shared pair of WAN Edge routers across sites
- External Layer-3 network may be different from the DCI network used for inter-site communication

Border Gateway and VRF-Lite



*No current SVIs support on BGWs

Border Gateway and Shared External Connectivity



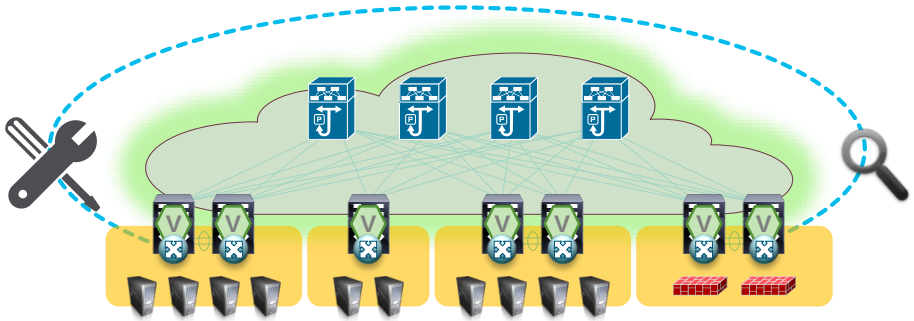
Agenda

- Introduction to Overlays
- Introduction to Overlays
- VXLAN with BGP EVPN
  - Standards and Implementation
  - Control & Data Plane
- Tenant Routed Multicast (TRM)
- Multi-Site
- VXLAN OAM

Operations, Administration and Management (OAM)

- OAM – processes, activities, tools and standards
- Various Mode of Operation
- Pro-Active
  - Controlling a Situation
- Re-Active
  - Responding to a Situation

VXLAN OAM - OAM Model of Operation

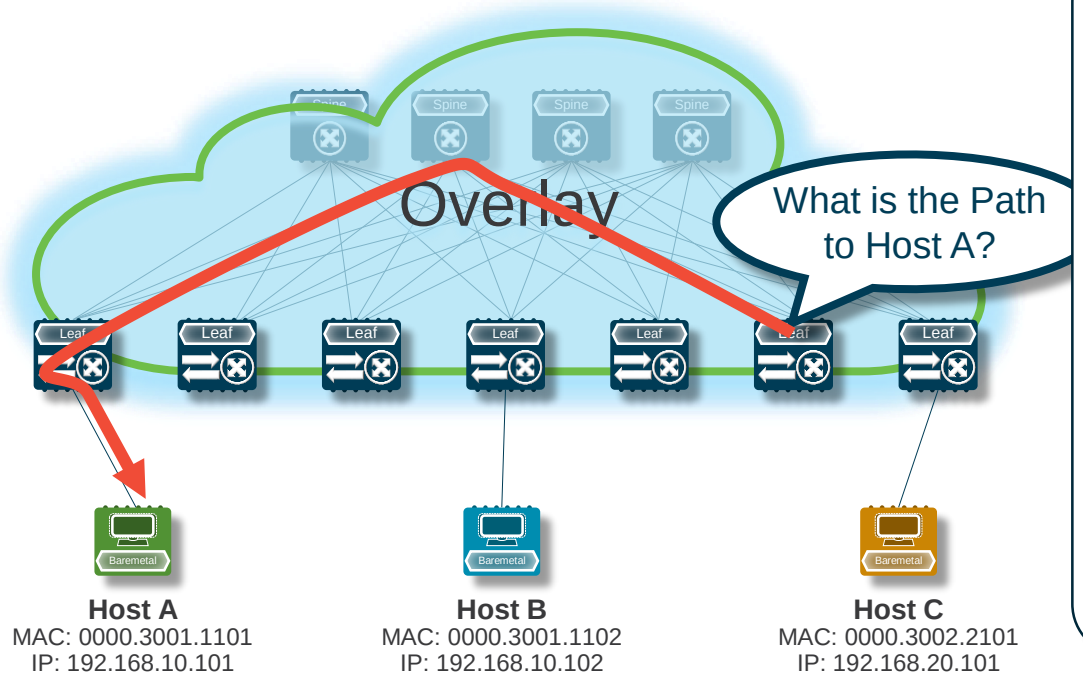


Endpoint Locator	Ping / Path MTU	Pathtrace	Pro-Active Monitoring
• Locate End-Host and Segment Identifier • Track History of End-Host • Provide Fabric Host-Count and Activity	• Check liveliness of End-Host • Option to specify Payload Parameters	• Trace paths to End-Host and Tunnel-Endpoint • Get Path, Interface and Error statistics along path • Specify Payload Parameters for Path Selection	• Proactive Monitoring with Threshold and State Notifications

NGOAM or VXLAN OAM

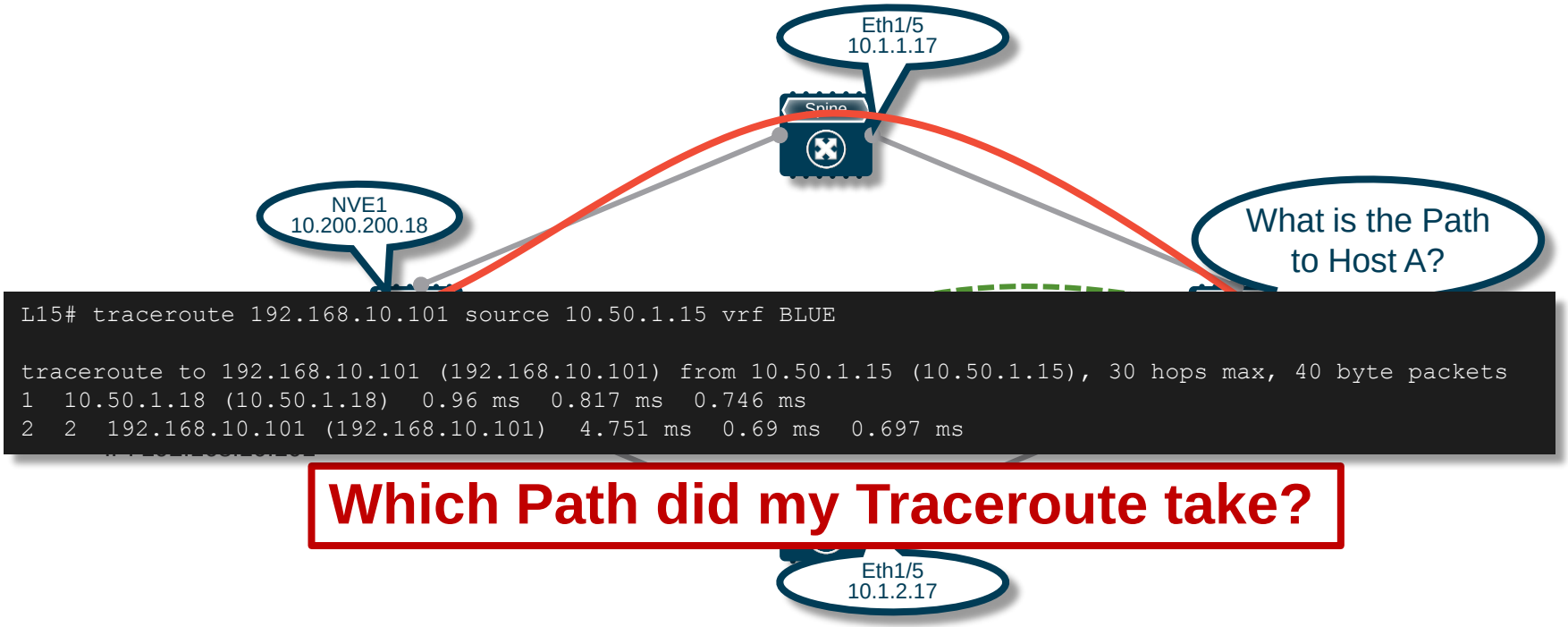
- Next Generation OAM for Data Centre Fabrics
- Running on Nexus 9000, Nexus 7000 and Nexus 5600
  - VXLAN Today
  - All IP Tomorrow
- Various Methods to Execute and Retrieve Data
  - Command Line Interface (CLI)
  - NX-API
  - DCNM (using NX-API)

Endpoint Traceroute – VXLAN OAM



- Endpoint Reachability
  - Uses ICMP
  - VTEP to Endpoint reachability
  - VTEP to VTEP reachability
- Validates Overlay Path
  - Single Specified Path
  - Multiple, Specified Path
- Provides Overlay to Underlay correlation

How Would a Normal Traceroute Look Alike?



Endpoint Traceroute – VXLAN OAM – Close-Up

```
L15# traceroute nve ip 192.168.10.101 vrf BLUE source 10.50.1.15 sport 35977 verbose

Codes: '!' - success, 'Q' - request not sent, '.' - timeout,
'D' - Destination Unreachable, 'X' - unknown return code,
'm' - malformed request(parameter problem),
'c' - Corrupted Data/Test, '#' - Duplicate response

Traceroute Request to peer ip 10.200.200.18 source ip 10.200.200.15
Sender handle: 94
  1 !Reply from 10.1.1.17,time = 1 ms
  2 !Reply from 10.200.200.18,time = 1 ms
  3 !Reply from 192.168.10.101,time = 4 ms
```

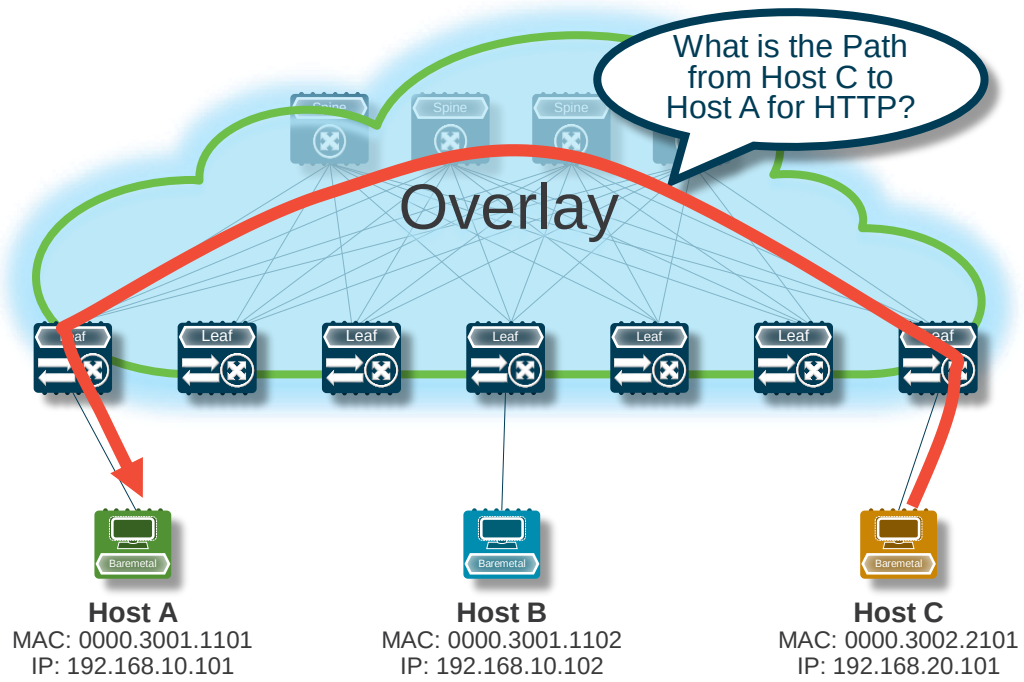
Spine Ingress Interface IP

Destination VTEP IP

Host A IP

Spine Ingress Interface and Destination VTEP IP Address are Underlay Information – additions vs. standard Traceroute

Pathtrace for Enhanced Network Visibility



- Application Specific Pathtrace
  - Uses “draft-tissa-nvo3-oam-fm”
  - Endpoint to Endpoint Pathtrace
  - Adds Interface Load and Error Statistic of the Path
  - Uses Protocol Information
- Validates Specific or All Path
- Provides Overlay to Underlay correlation
- Superset of NVE Ping/Traceroute

Pathtrace – VXLAN OAM – Close-Up

```
L15# pathtrace nve ip unknown vrf BLUE
       payload
       ip 192.168.10.101 192.168.20.101
       port 54321 80
       proto 6
       payload-end

Codes: '!' - success, 'Q' - request not sent, '.' - timeout,
'D' - Destination Unreachable, 'X' - unknown return code,
'm' - malformed request(parameter problem),
'c' - Corrupted Data/Test, '#' - Duplicate response

Path trace Request to peer ip 10.200.200.18 source ip 10.200.200.15
Sender handle: 142

Hop   Code   ReplyIP   IngressI/f  EgressI/f   State
=================================================
  1 !Reply from 10.1.1.17, Eth1/5  Eth1/8  UP / UP
  2 !Reply from 10.200.200.18, Eth1/54  Unknown  UP / DOWN
```

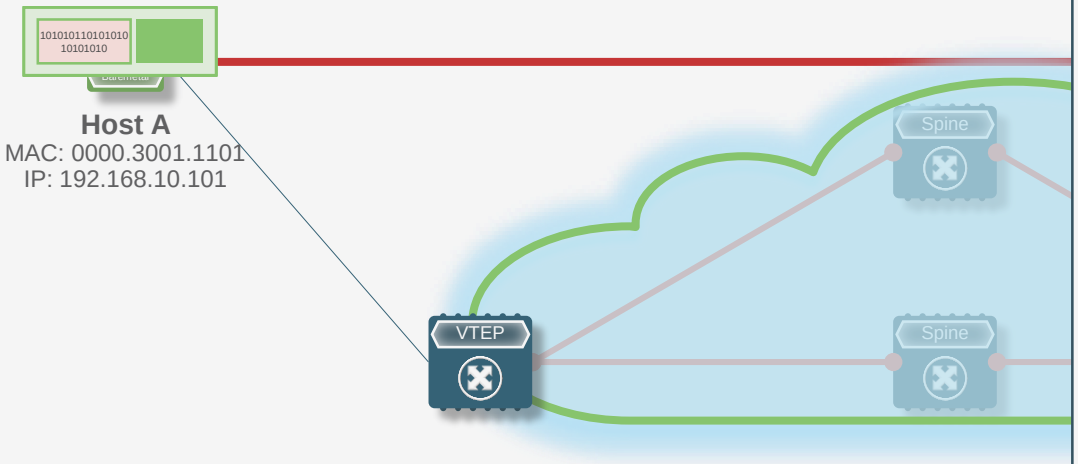
Known or Unknown VTEP IP Address

Dst Endpoint IP / Src Endpoint IP

Source Port / Destination Port

TCP (IANA Protocol Number 6)

Why are we Specifying Payload Information?



- VXLAN provides variable UDP Source Port in Outer Header
- Hash of the inner Layer-2/Layer-3/Layer-4 Headers of the original Ethernet Frame.
- Enables entropy for ECMP Load balancing in the Network

Which Path did your Application Traffic took?

Pathtrace – VXLAN OAM – Close-Up

```
L15# pathtrace nve ip unknown vrf BLUE payload ip 192.168.10.101 ...

Codes: '!' - success, 'Q' - request not sent, '.' - timeout,
'D' - [illegible]
'm' - [illegible]
'c' - [illegible]

Path trace Request to peer ip 10.200.200.18 source ip 10.200.200.15
Sender [illegible]

Hop   Code   ReplyIP   IngressI/f  EgressI/f   State
=============================================================
  1 !Reply from 10.1.1.17, Eth1/5  Eth1/8  UP / UP
  2 !Reply from 10.200.200.18, Eth1/54  Unknown  UP / DOWN
```

Spine Ingress Interface, Egress Interface and Destination VTEP IP Address are Underlay Information – additions vs. standard and NVE Traceroute

Spine Ingress Interface IP

Spine Ingress Interface

Spine Egress Interface

Interface Status

Destination VTEP IP

Destination Leaf Ingress Interface

Database Output – VXLAN OAM – Close-Up

```
L15# show ngoam pathtrace database session 168 detail

Pathtrace entry for session id 168
==================================

Start time: Tue Jun 13 01:18:39.710 PDT
End time: Tue Jun 13 01:18:39.735 PDT

Last Clear of Summary Statistics: Never
Pathtrace Requests: sent (2)/received (0)/timeout (0)/unsent (0)
Pathtrace Replies: sent (0)/received (2)/unsent (0)/Duplicate (0)

! Reply from 10.1.1.17 on Eth1/5, state  UP. Sent on Eth1/8, state  UP.
 Interface stats for interface: Eth1/5
------------------------------
Rx Len                : 84
Rx Bytes              : 66113123
Rx Pkt rate           : 0
Rx Byte rate          : 0
Rx Load               : 0
```

OAM Session ID

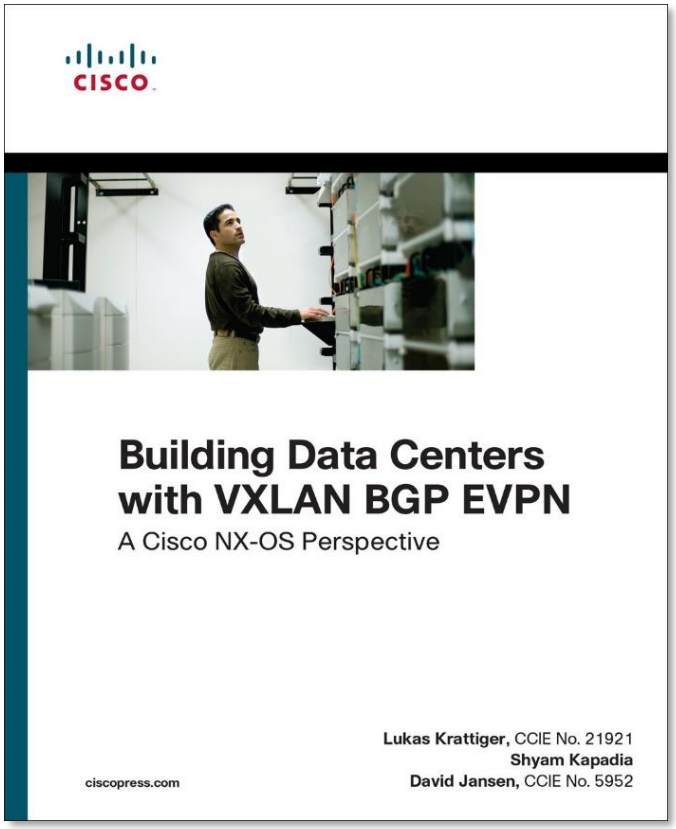
Interface Statistics

Summary

Summary

- Overview on VXLAN Overlay
- Standards and Implementation
- Control- and Data-Plane interactions
- Some info around Multicast forwarding
- How Multi-Site enhanced VXLAN EVPN
- Operations is key – VXLAN OAM

If you haven't had enough VXLAN BGP EVPN



Links & Resources

- VXLAN Multi-Site Intro
  - https://blogs.cisco.com/datacenter/vxlan-innovations-vxlan-evpn-multi-site-part-2-of-2
- VXLAN Multi-Site @ Cisco Live online
  - https://www.ciscolive.com/global/on-demand-library/?search=BRKDCN-2035#/
- ”eBGP” for EVPN
  - https://learningnetwork.cisco.com/blogs/community_cafe/2017/11/02/vxlan-ebgp-evpn-the-incarnation-of-a-hybrid-guest-post
- Configuration Example
  - https://communities.cisco.com/community/technology/datacenter/data-center-networking/blog/2015/05/19/vxlanevpn-configuration-example

Q & A

Complete Your Online Session Evaluation

- Give us your feedback and receive a **Cisco Live 2018 Cap** by completing the overall event evaluation and 5 session evaluations.
- All evaluations can be completed via the Cisco Live Mobile App.

Don’t forget: Cisco Live sessions will be available for viewing on demand after the event at www.CiscoLive.com/Global.

Ciscolive!

Thank you


cisco
You're
Cisco live!