#ThaiNOG

Extending VXLAN EVPN to the Campus and Data Center Interconnect (DCI)

Rev. 1

THERDTOON Theerasasana
ttheera@cisco.com

May 2022

THAINOG
Thai Network Operators Group

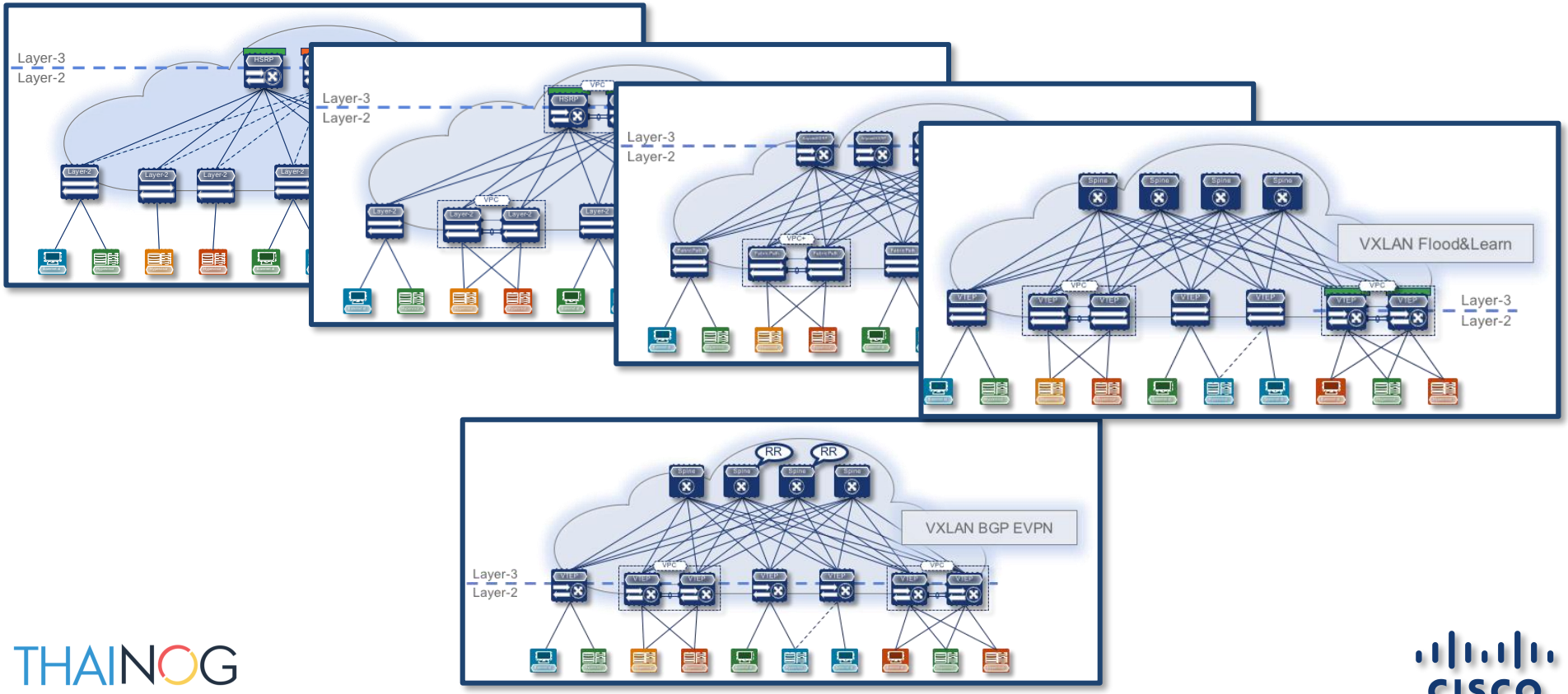
CISCO

Agenda

- VXLAN and EVPN Overview
- VXLAN EVPN for Campus
- Data Center Interconnect (DCI) using VXLAN EVPN

VXLAN and EVPN Overview

Data Center “Fabric” Journey



Why VXLAN Overlay

Customer Needs	VXLAN Delivered
Any workload anywhere – VLANs limited by L3 boundaries	Any Workload anywhere- across Layer 3 boundaries
VM Mobility	Seamless VM Mobility
Scale above 4k Segments (VLAN limitation)	Scale up to 16M segments
Efficient use of bandwidth	Leverages ECMP for optimal path usage over the transport network
Secure Multi-tenancy	Traffic & Address Isolation

VXLAN and EVPN

VXLAN

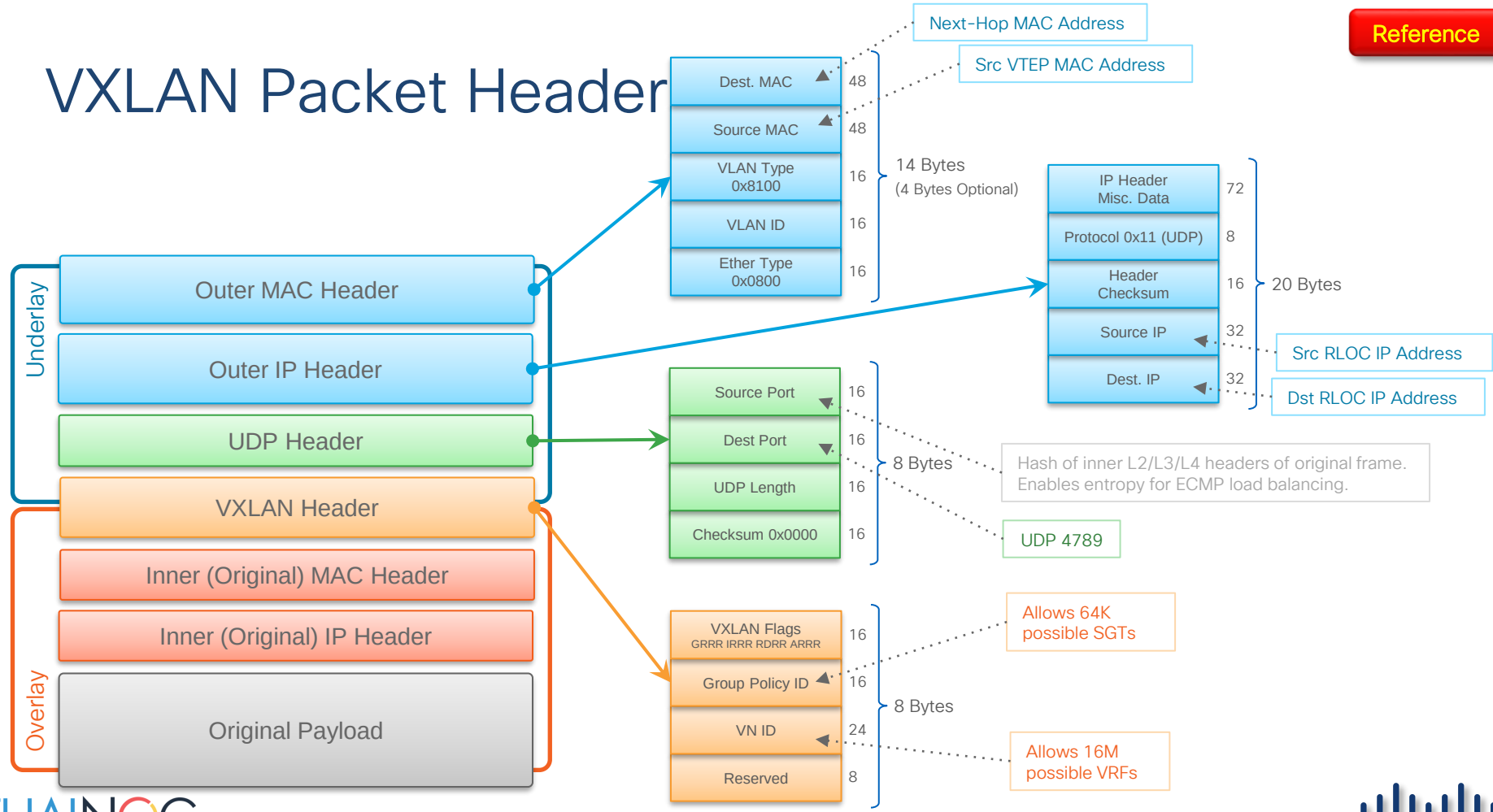
- Standards based Encapsulation
  - RFC 7348
  - Uses UDP-Encapsulation
- Transport Independent
  - Layer-3 Transport (Underlay)
- Flexible Namespace
  - 24-bit field (VNID) provides ~16M unique identifier
  - Allows Segmentations

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)

VXLAN Packet Header

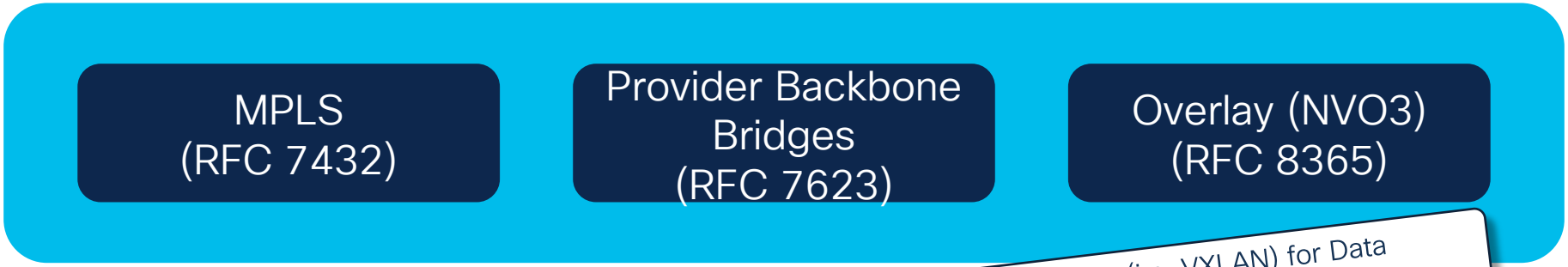
Reference



THAINOG
Thai Network Operators Group

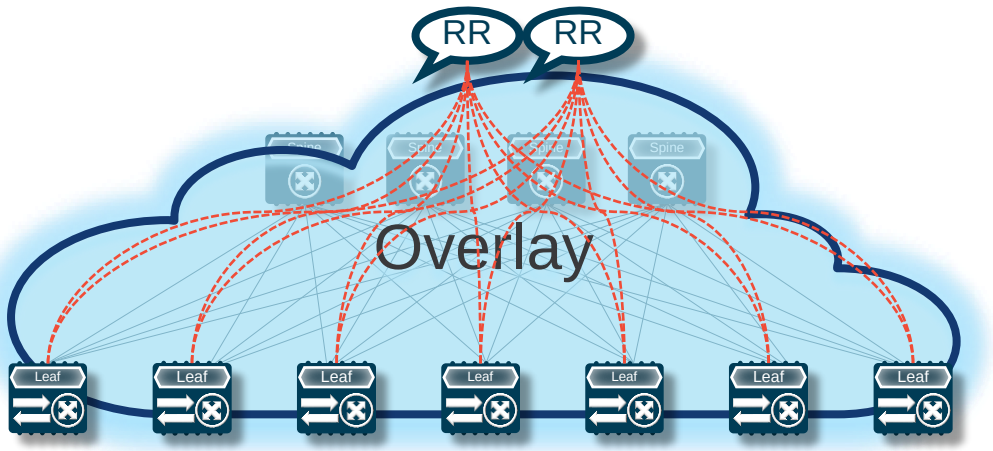
cisco

Ethernet VPN (EVPN)



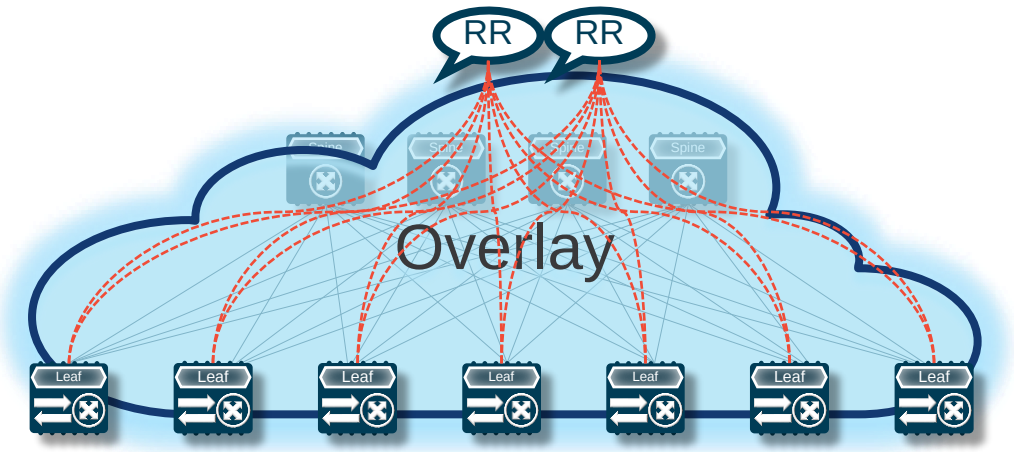
- EVPN over NVO Tunnels (i.e. VXLAN) for Data Center Fabric Encapsulation
- Provides Layer-2 and Layer-3 Overlay Service over simple IP Network

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

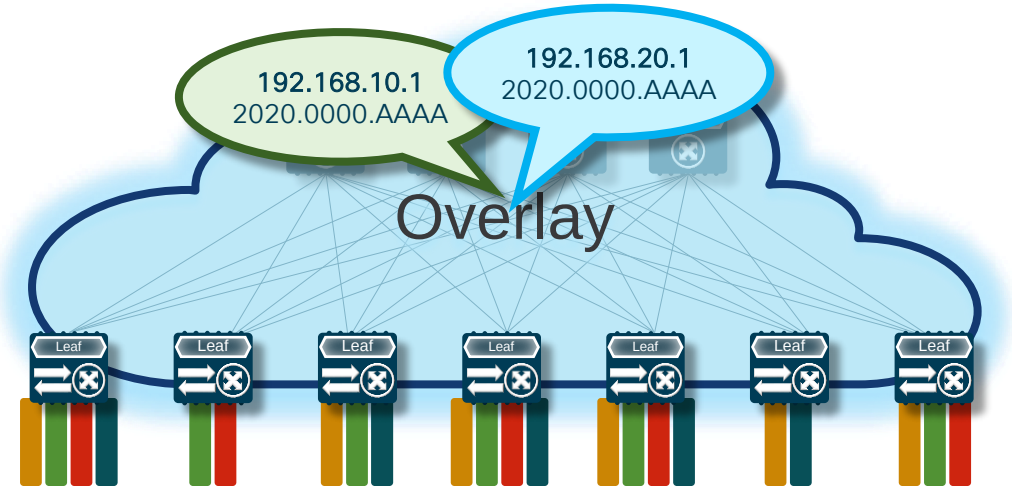
EVPN Control Plane – Host and Subnet Routes



- 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)*

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

VXLAN EVPN for Campus

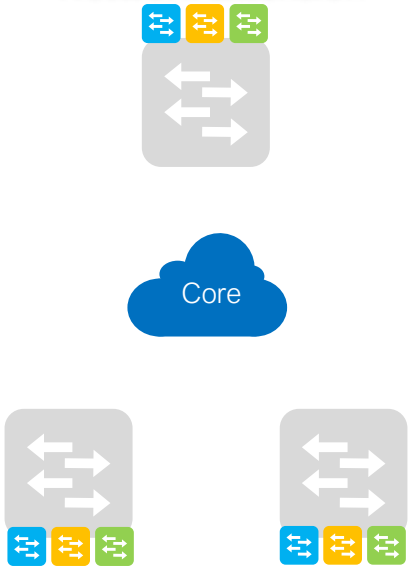
THAINOG
Thai Network Operators Group

EVPN Requirements & Drivers

Requirements	Drivers
Industry-standard	Multi-vendor IT strategy
One Fabric Architecture	Unified operation across – Campus | DC | WAN
Proven and Scalable	BGP Protocol History. Minimum new learning curve
Hierarchical Fabric Domain	Multi-tier Overlay network architecture
Flexible Overlay	Use-case driven customize Overlay networks Types and Topologies

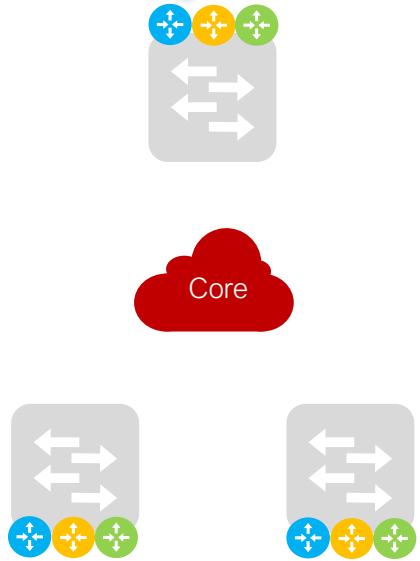
BGP EVPN Drivers in Enterprise

Network Extension



- **Bridge** connection between across Core network
- User devices are virtually in common L2 segment
- Logical topologies with deterministic overlay Layer 2 network infrastructure.

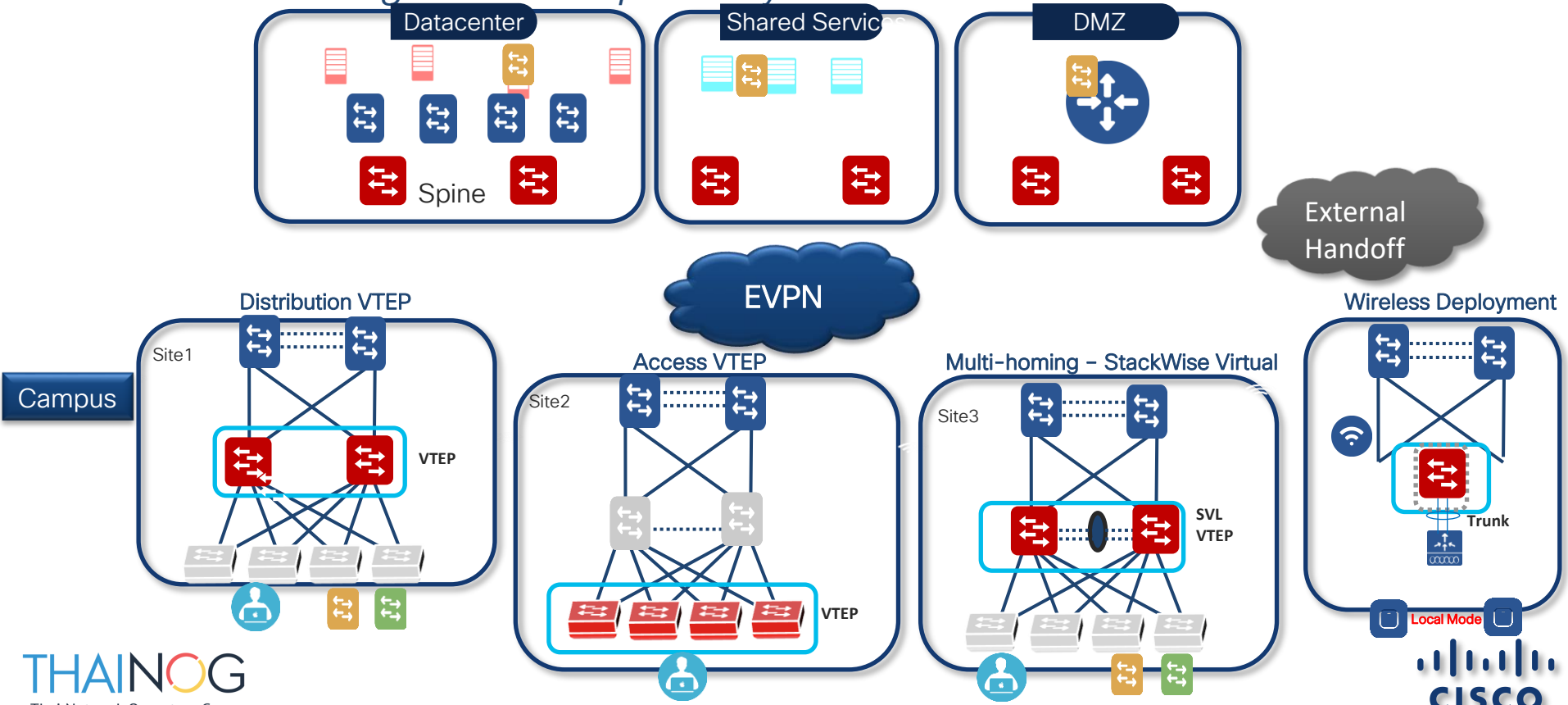
Network Segmentation



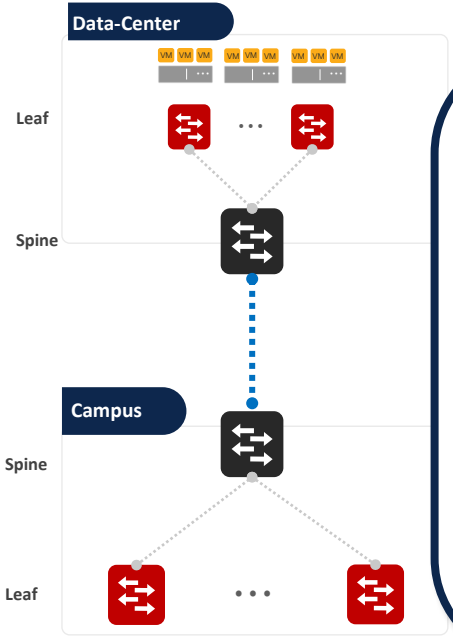
- **Routed** connection at first-hop gateway
- User devices are segmented across Core network
- Logical overlay IP routed network providing flexible topology support

VXLAN BGP EVPN Solution

End-to-End Design and Interoperability



BGP EVPN Inter-Domain Routing



Single Site – Inter-domain EVPN fabric

Hierarchical control-plane for better scale

Spine and Leaf BGP peering in separate domain

Spine-to-Spine BGP Peering

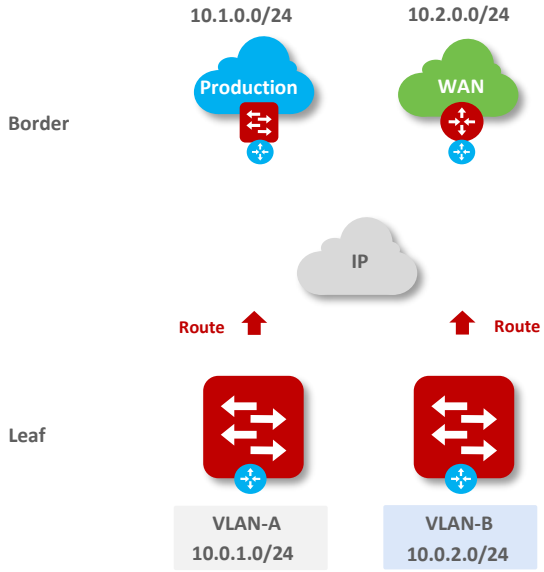
EVPN prefix exchange Next-Hop-Unchanged

Extended

Segmented

BGP EVPN Peers

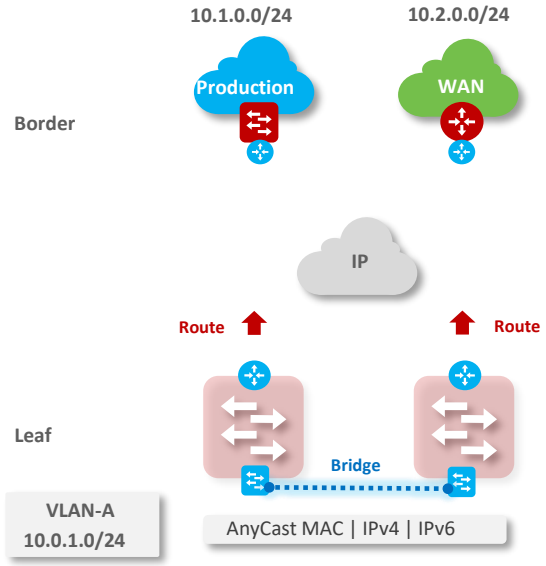
L3VNI – Network Segmentation and Routing



Routing

- First-Hop Distributed Gateway at Access
- Access Network policy enforcement point
- Network address routing across fabric
- Data plane segmentation thru VXLAN
- IPv4 / v6 support

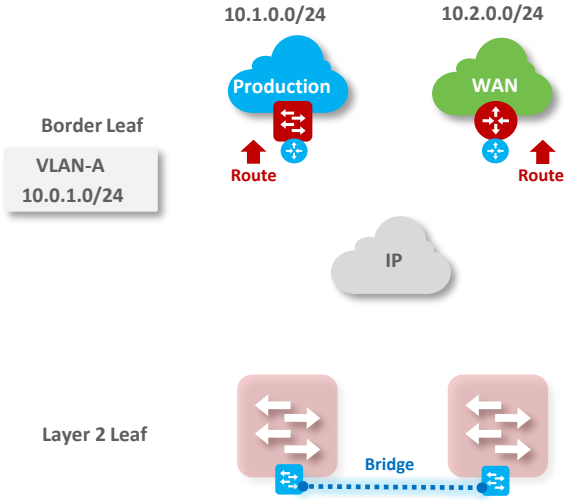
Distributed AnyCast Gateway



Routing + Bridging

- First-Hop Distributed Anycast Gateway at Access
- Bridge in same VLAN across Leaf's in fabric
- Route locally based on local routing policy
- Access Network policy enforcement point
- Host + Network address routing across fabric
- Data plane segmentation thru VXLAN
- IPv4 / v6 support

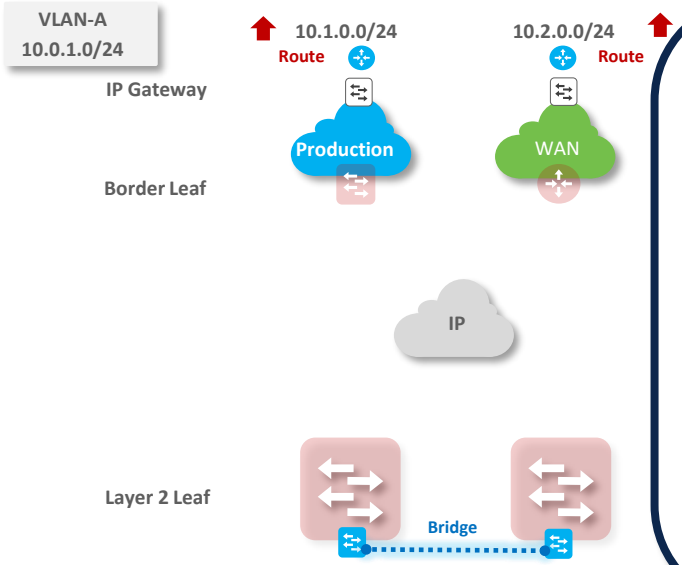
Centralized Gateway



Routing + Bridging

- Multi-Hop Centralized Gateway
- Bridge in same VLAN across Leaf's in fabric
- Route remotely based on remote routing policy
- Access Network policy enforcement point
- Host address routing across fabric
- Data plane segmentation thru VXLAN
- IPv4 / v6 support

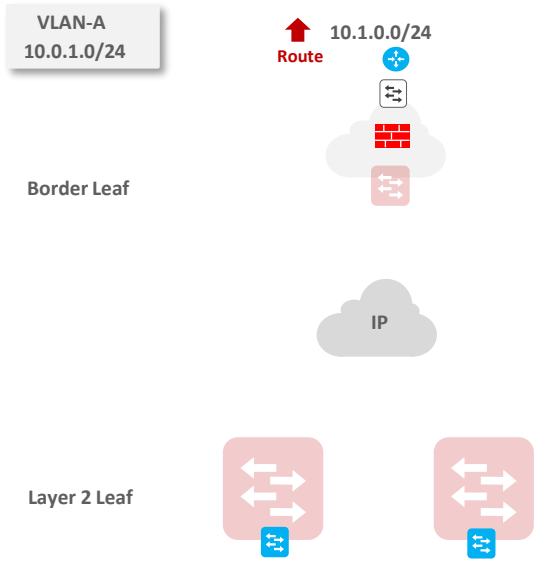
Layer 2 Network Extensions



Bridging

- IP Gateway beyond EVPN fabric
- Bridge in same VLAN across Leaf's in fabric
- Route outside fabric based on remote routing policy
- Access Network policy enforcement point
- Host address routing across fabric
- Data plane segmentation thru VXLAN
- IPv4 / v6 support

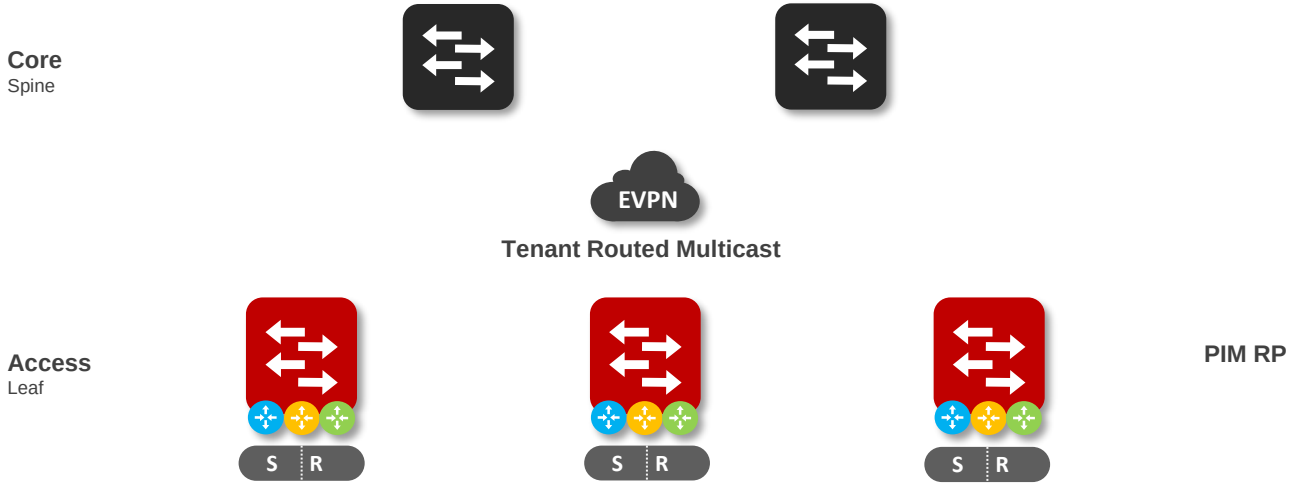
L2 – Hub-n-Spoke Network Extension



Bridging

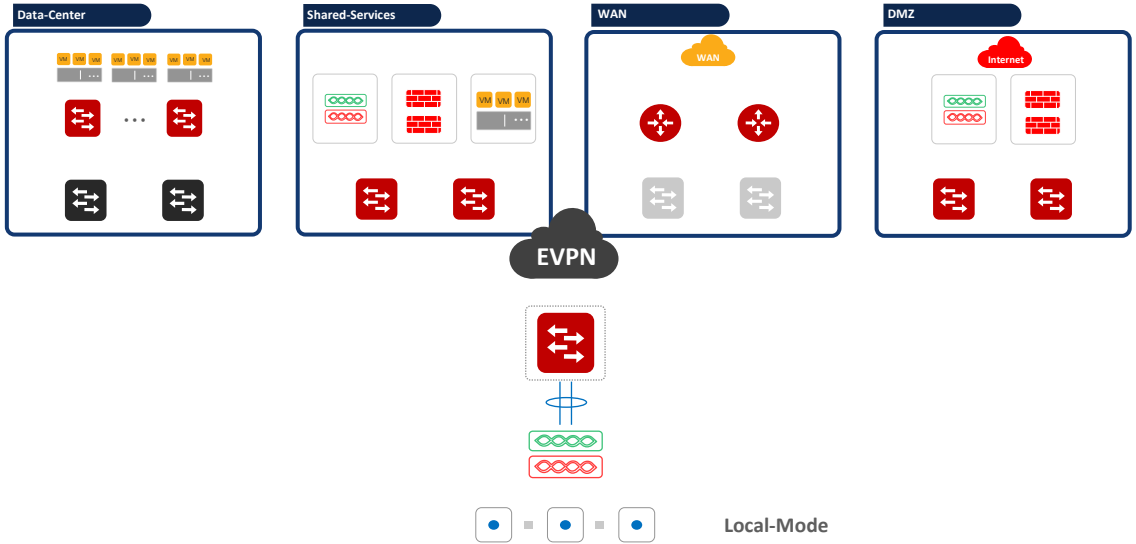
- IP Gateway beyond EVPN fabric
- Border L2 Leaf Hub. Layer 2 Leaf Spokes.
- Point-to-Point L2VNIs to Hub
- Route outside fabric based on remote routing policy
- Access Network policy enforcement point
- Host address routing across fabric
- Data plane segmentation thru VXLAN
- IPv4 / v6 support

Tenant Routed Multicast Architecture



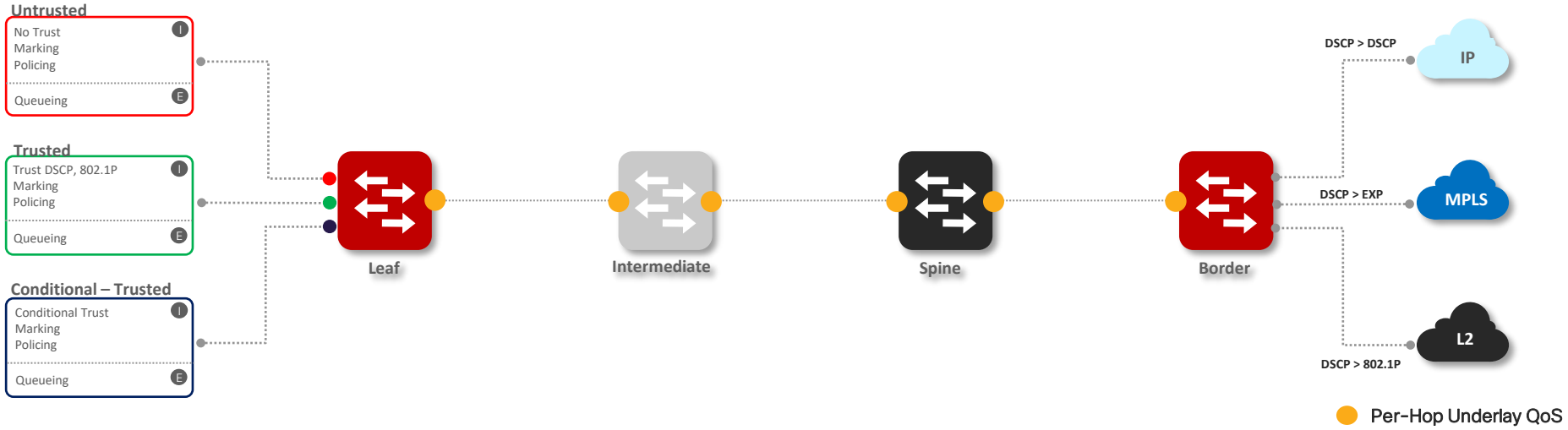
- TRM enables Multicast over VXLAN enabled network for Layer 3 network segments.
- Integrated PIM RP and PIM-SM in Underlay support enables fabric-enabled source and receivers Multicast forwarding topologies.
- VTEP provides integrated PIM RP function

Wireless Integration in EVPN Networks



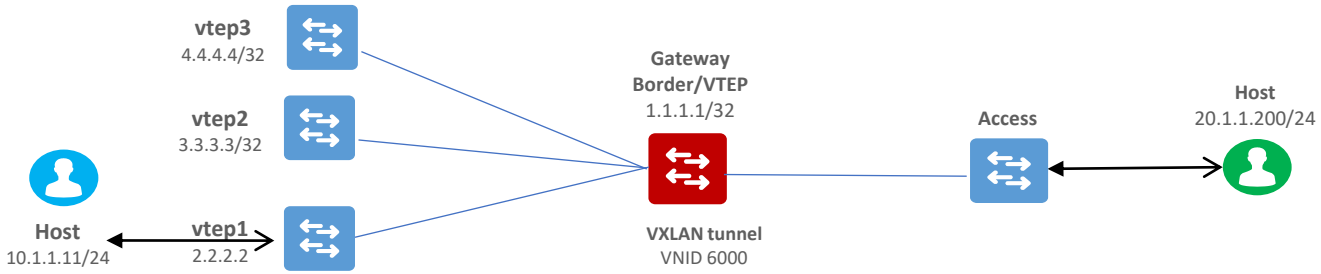
- **Transparent Wireless network integration into BGP EVPN fabric network**
- **Underlay CAPWAP communication between AP and WLC. User Policy enforcement maintains at WLC.**
- **VTEP in Wireless aggregation can overlay network traffic based on routing policy.**

VXLAN QoS Management



- **Trust-Boundary and policy enforcement at network edge.**
- **Per-hop Underlay QoS policy provides differentiated service treatment for combined underlay and overlay traffic class**
- **QoS policy and marking at Border supports default or user-defined policy with interworking external network domain**

VXLAN Aware Flexible NetFlow



- **Supports v4 and v6 protocols**
- **Bi-directional flow detection over the NVE interface**

Flow record fields	Packet data
Source Address	10.1.1.11
Destination Address	20.1.1.200
Source Port	47321
Destination Port	80
IP Protocol	6
TCP Flags	0x1A
Source SGT	0
Interface	nve10
Flow direction	input
VNID	**6000**
VXLAN Flags	**I**
VXLAN SRC VTEP	**1.1.1.1**
VXLAN DST VTEP	**2.1.1.1**

Data Center Interconnect (DCI) with VXLAN EVPN

THAINOG
Thai Network Operators Group

VXLAN Evolves as the Control Plane Evolves!

Before Yesterday

Yet Another Encapsulation

- Flood & Learn (Multic
- Data-Plane only

Yesterday

VXLAN for the Data Center – Intra-DC

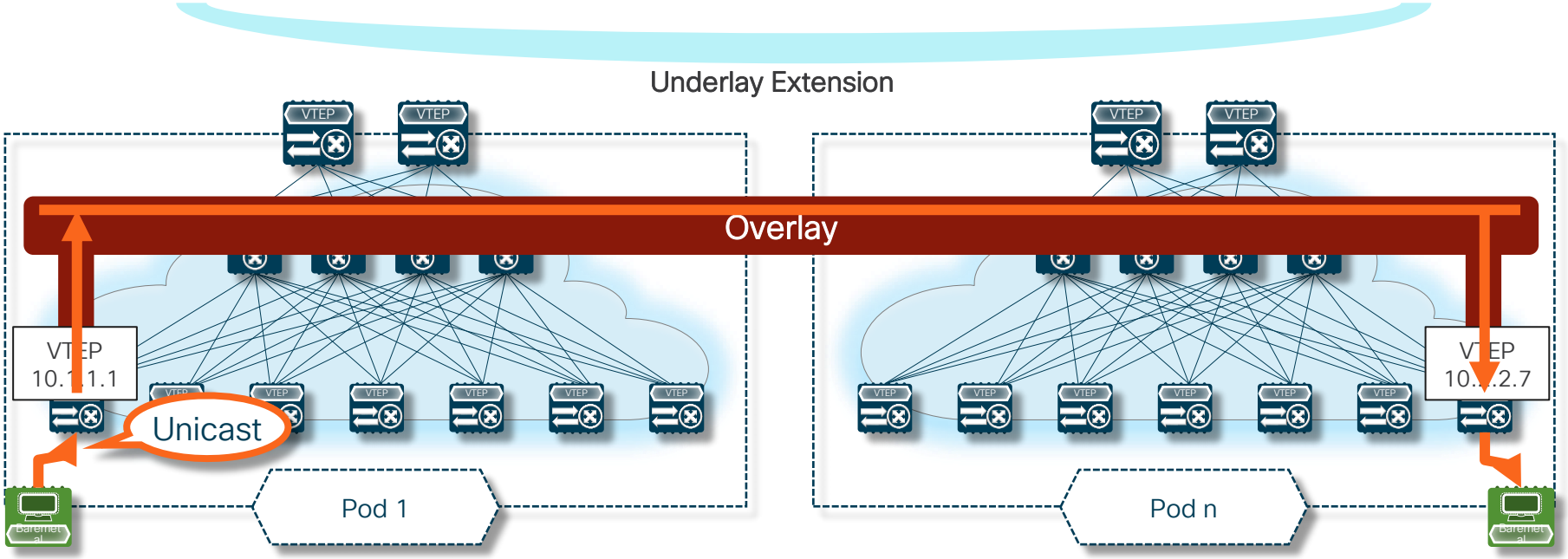
- Control-Plane
- Active VTEP Discove
- Multicast and Unicas

Today

VXLAN for DCI – Inter-DC

- DCI Ready
- ARP/ND caching/suppress
- Multi-Homing
- Failure Domain Isolation
- Loop Protection

Multi-Pod End-to-End Encapsulation



Multi-Pod Characteristics – ”The Single”

- **Single** Overlay Domain – End-to-End Encapsulation
- **Single** Overlay Control-Plane Domain – End-to-End EVPN Updates
- **Single** Underlay Domain End-to-End
- **Single** Replication Domain for BUM
- **Single** VNI Administrative Domain

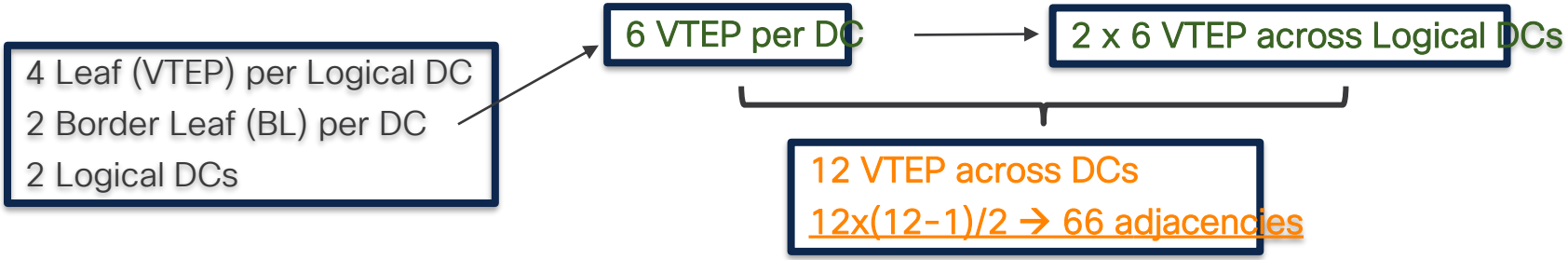
Building Underlay Hierarchies – Non Hierarchical Overlay

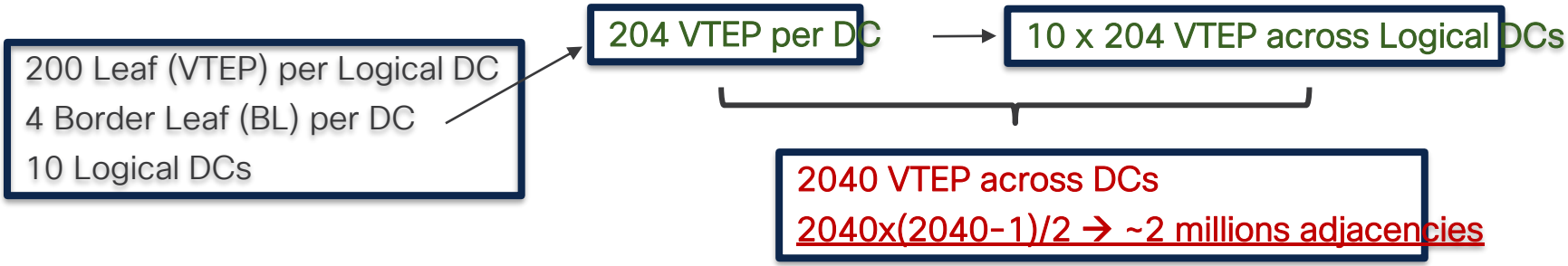
The Ugly Multi-Pod Truth

What about the Required VXLAN Tunnel Adjacencies?

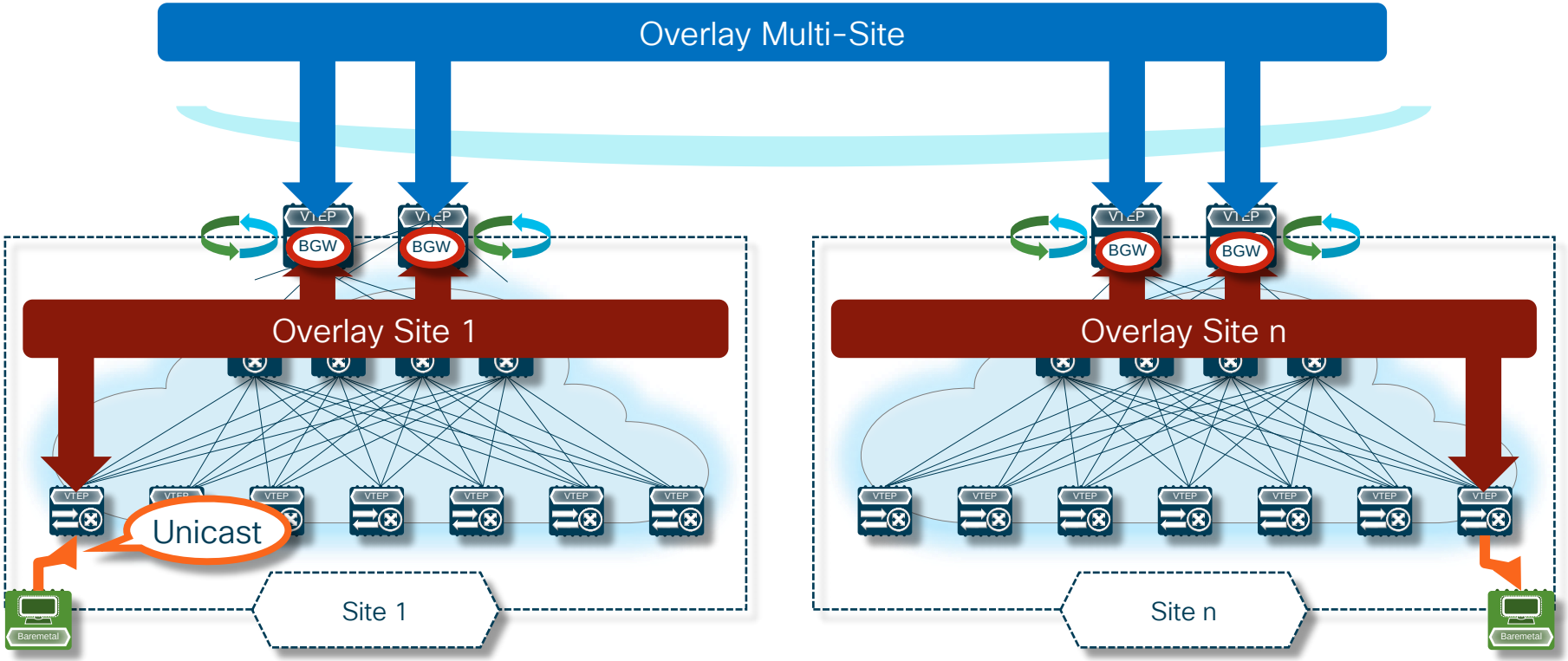
Tunnel adjacencies

$$\frac{N * (N-1)}{2}$$





VXLAN Multi-Site
Hierarchical Overlay Domains



VXLAN Multi-Site Characteristics – "The Multiple"

- **Multiple** Overlay Domains – Interconnected & Controlled
- **Multiple** Overlay Control-Plane Domains – Interconnected & Controlled
- **Multiple** Underlay Domains – Isolated
- **Multiple** Replication Domains for BUM – Interconnected & Controlled
- **Multiple** VNI Administrative Domains – Downstream VNI

Underlay Isolation – Overlay Hierarchies

The Multi-Site Truth

What about the Required VXLAN Tunnel Adjacencies?

Tunnel adjacencies

$$\frac{N * (N-1)}{2}$$

4 Leaf (VTEP) per Site
2 Border Leaf (BL) per Site
2 Logical Sites

→ 6 VTEP per Site → 2 x 6 VTEP across Sites

12 VTEP across Sites
6x(6-1)/2 → 15 adjacencies per Site

---

200 Leaf (VTEP) per Site
4 Border Leaf (BL) per Site
10 Logical Sites

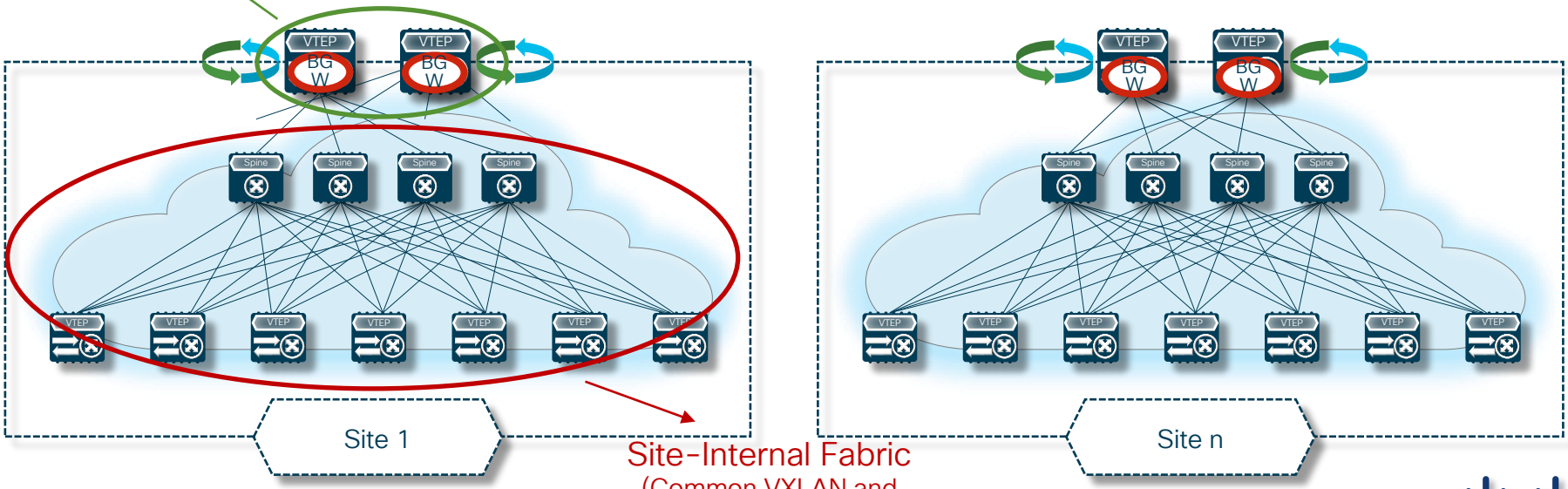
→ 204 VTEP per Site → 10 x 204 VTEP across Sites

2040 VTEP across Sites
204x(204-1)/2 → ~20K adjacencies per site

VXLAN Multi-Site Functional Components

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

Site-External DCI
(IP Routing and Increased MTU Support)

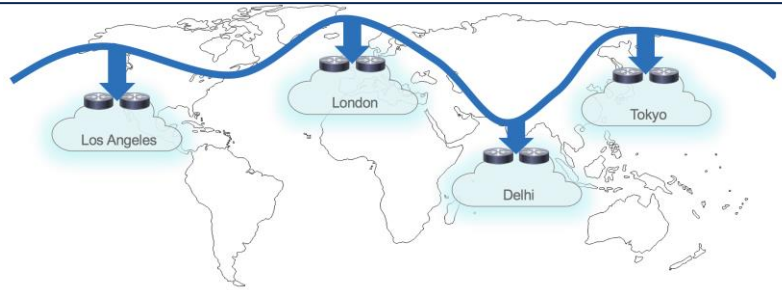


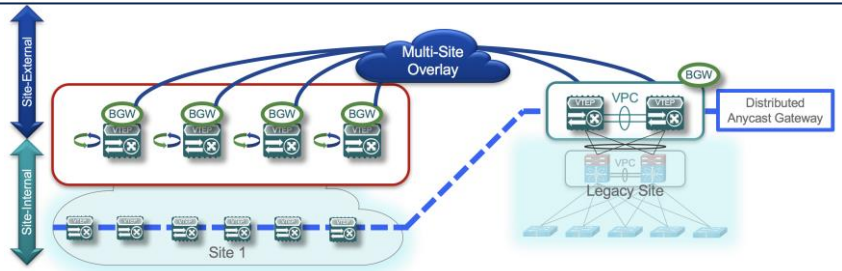
Site-Internal Fabric
(Common VXLAN and BGP-EVPN Functions)

VXLAN Multi-Site Main Use Cases

Scale-Up Model to Build a Large Intra-DC Network

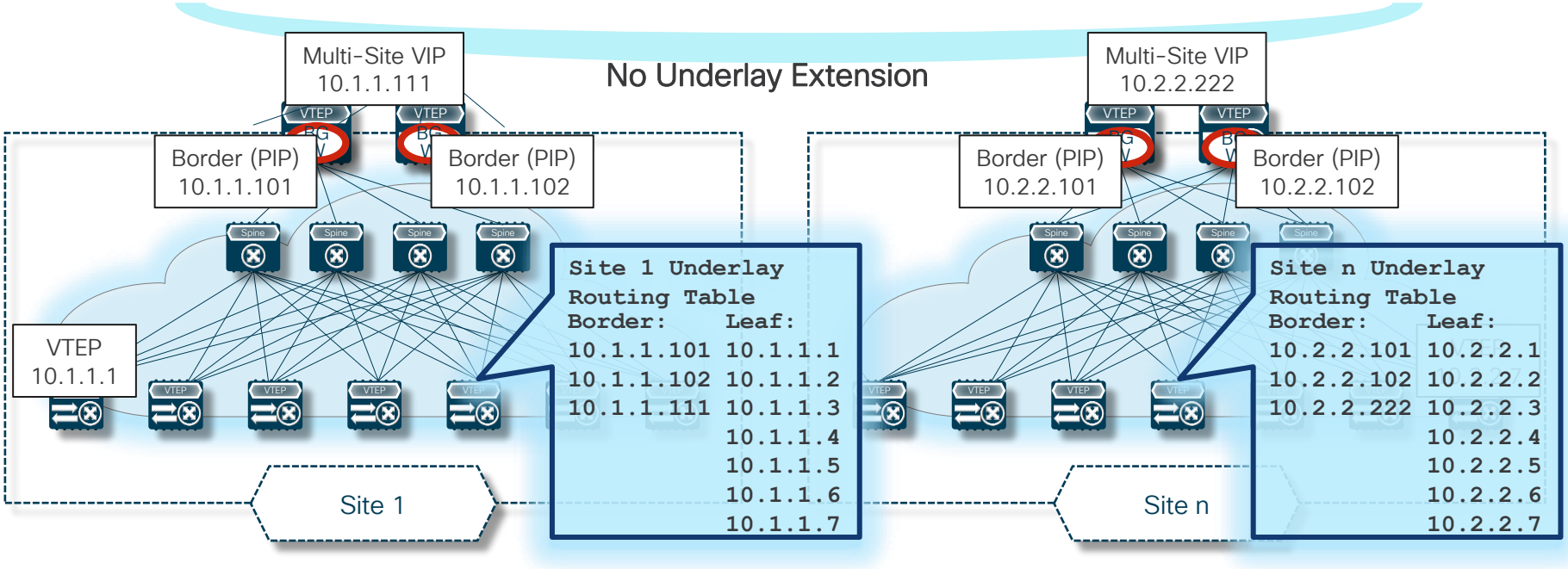
Data Center Interconnect (DCI)





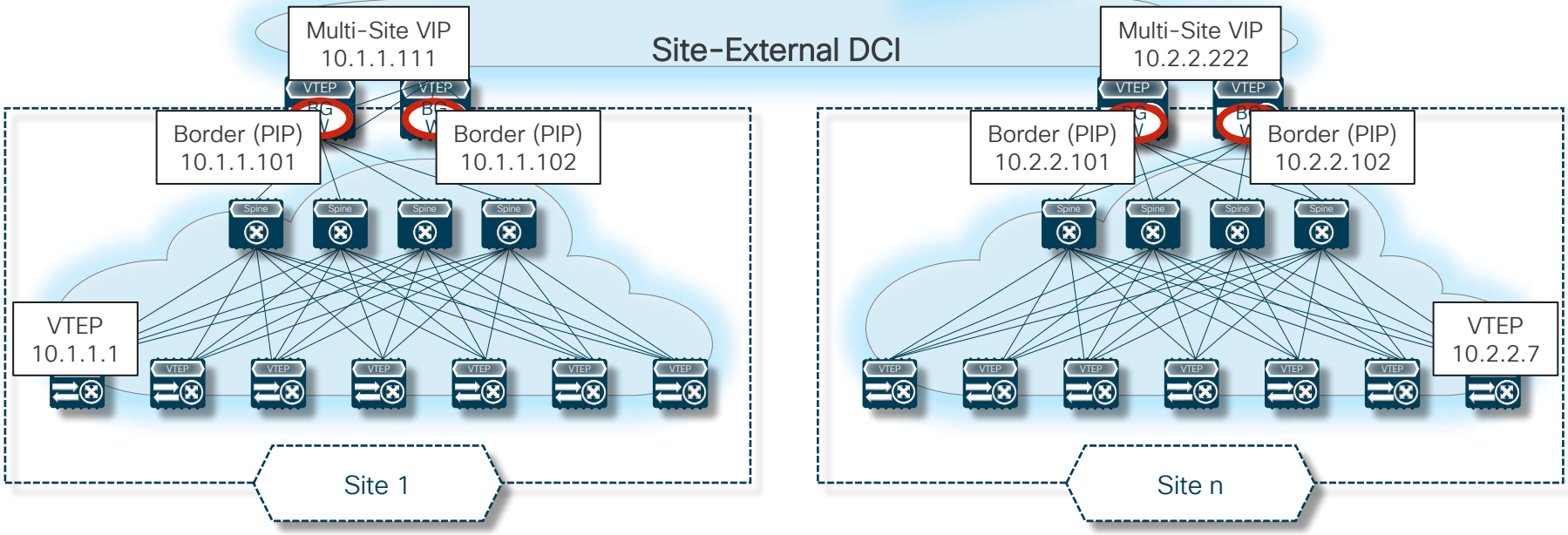
Integration with Legacy Networks (Coexistence and/or Migration)

VXLAN Multi-Site Underlay Isolation

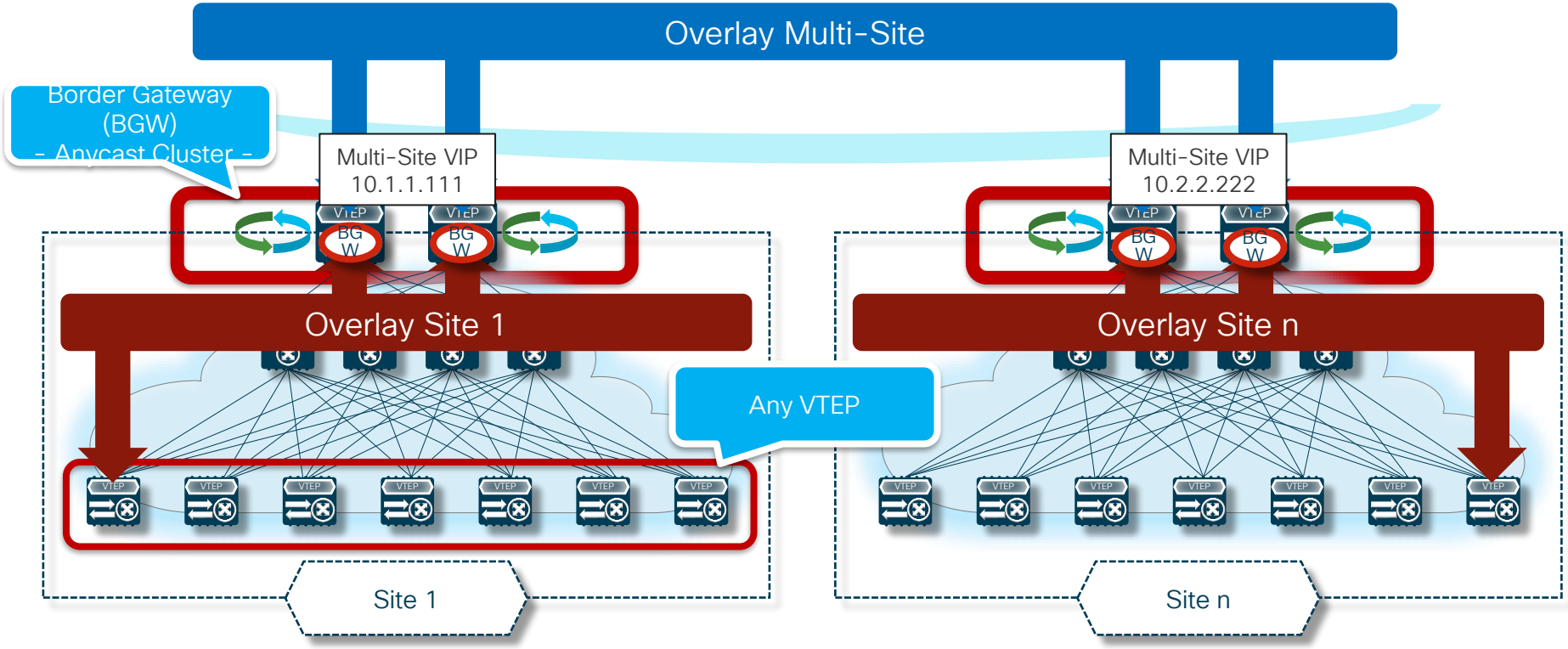


VXLAN Multi-Site Site-External DCI

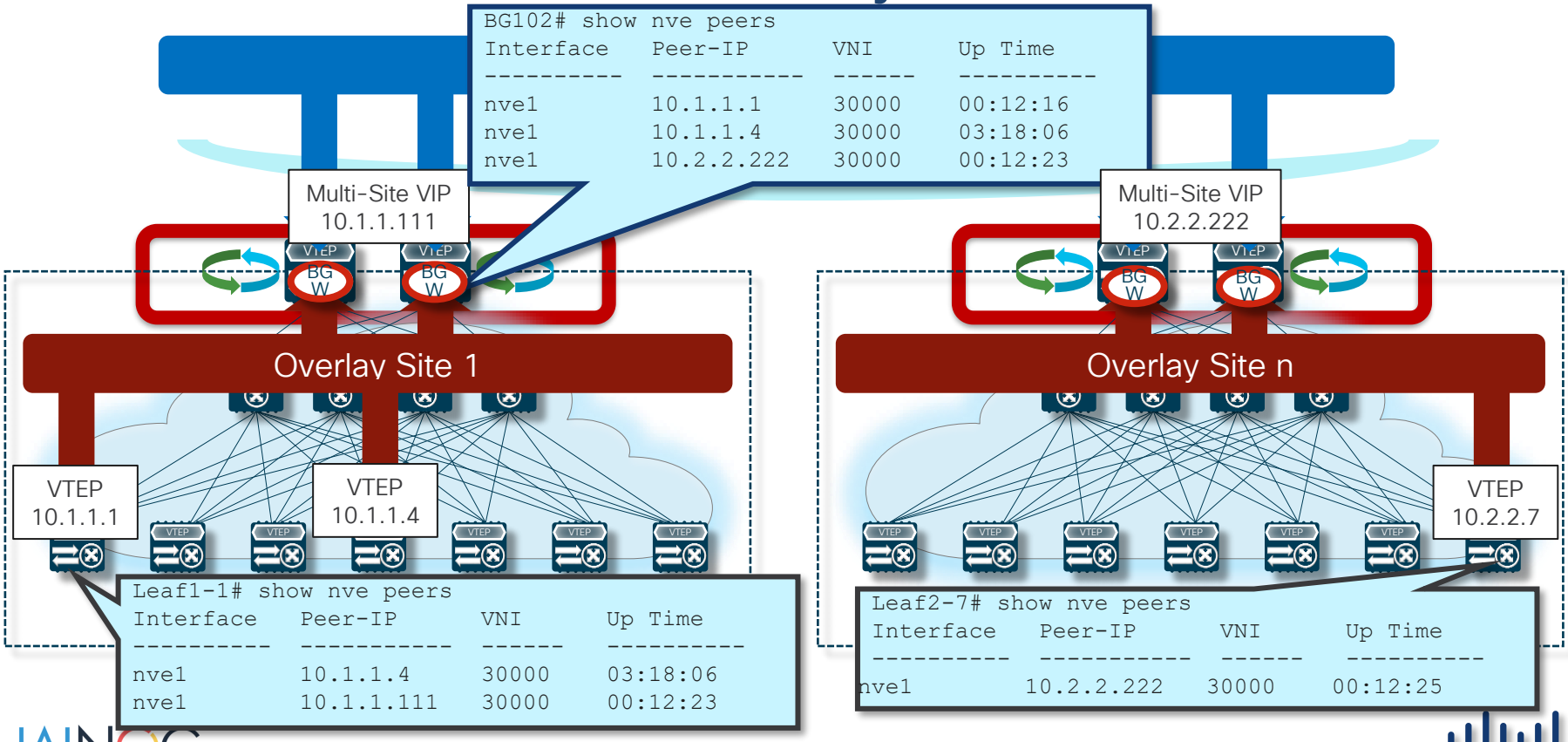
```
Inter-Site Network
Routing Table
 Border Site1:  Border Site2:
 10.1.1.101     10.2.2.101
 10.1.1.102     10.2.2.102
 10.1.1.111     10.2.2.222
```



VXLAN Multi-Site Border Gateway



VXLAN Multi-Site Tunnel Adjacencies

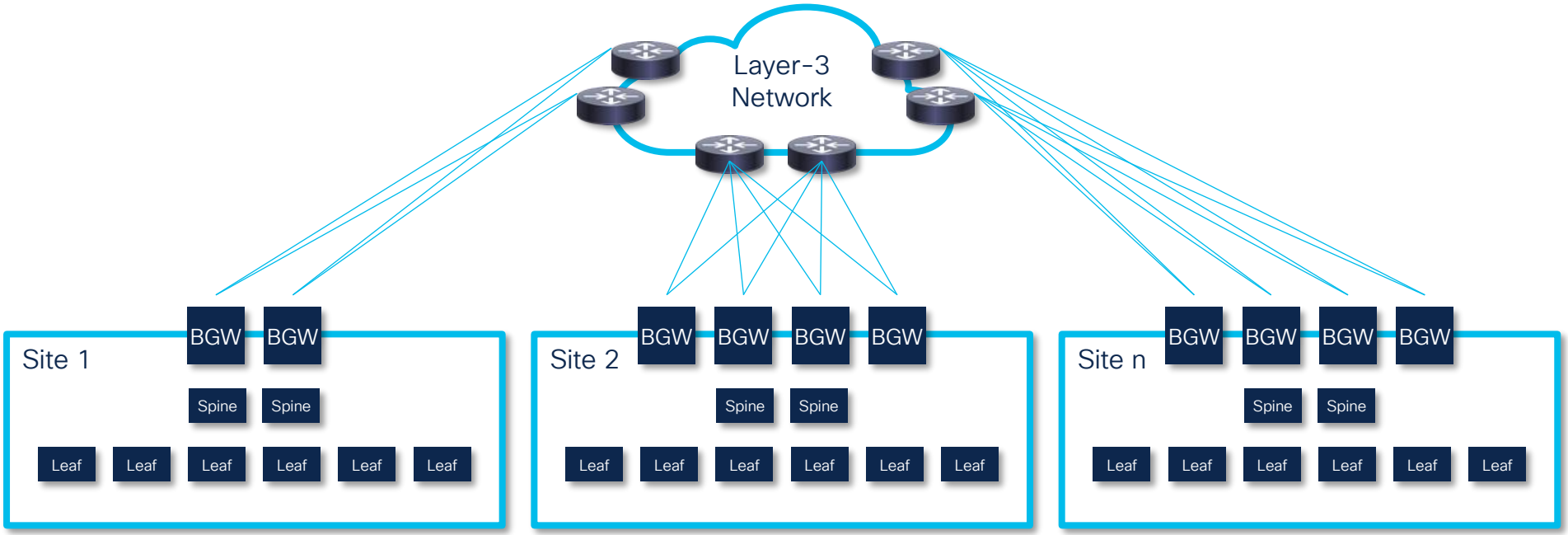


VXLAN EVPN Multi-Site Topologies

THAINOG
Thai Network Operators Group

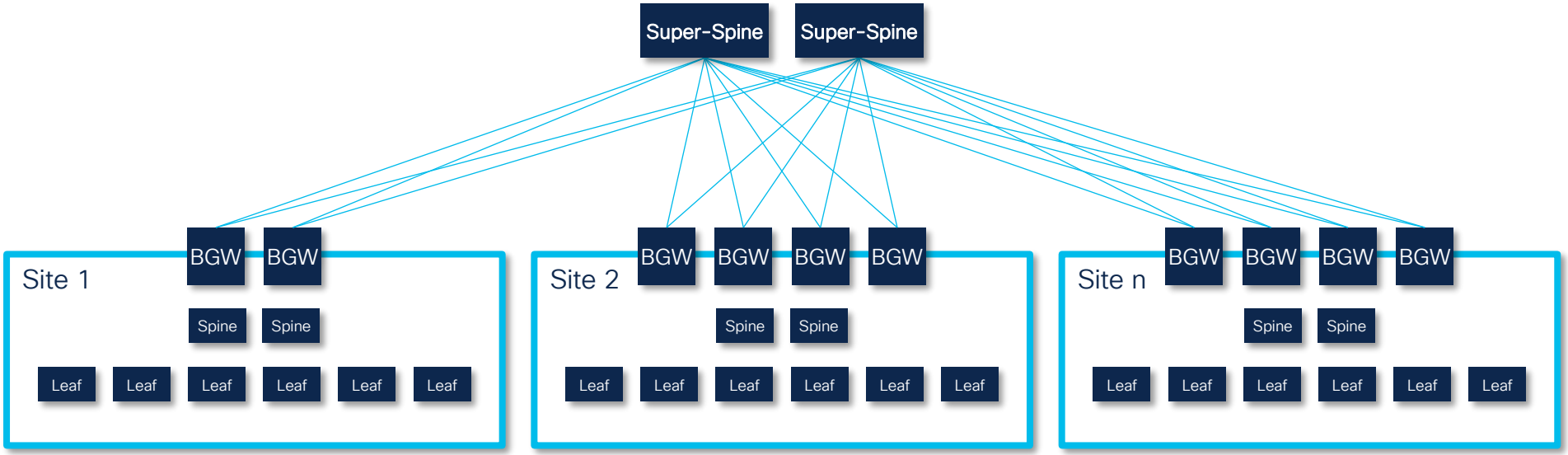
Reference

BGW-to-Cloud



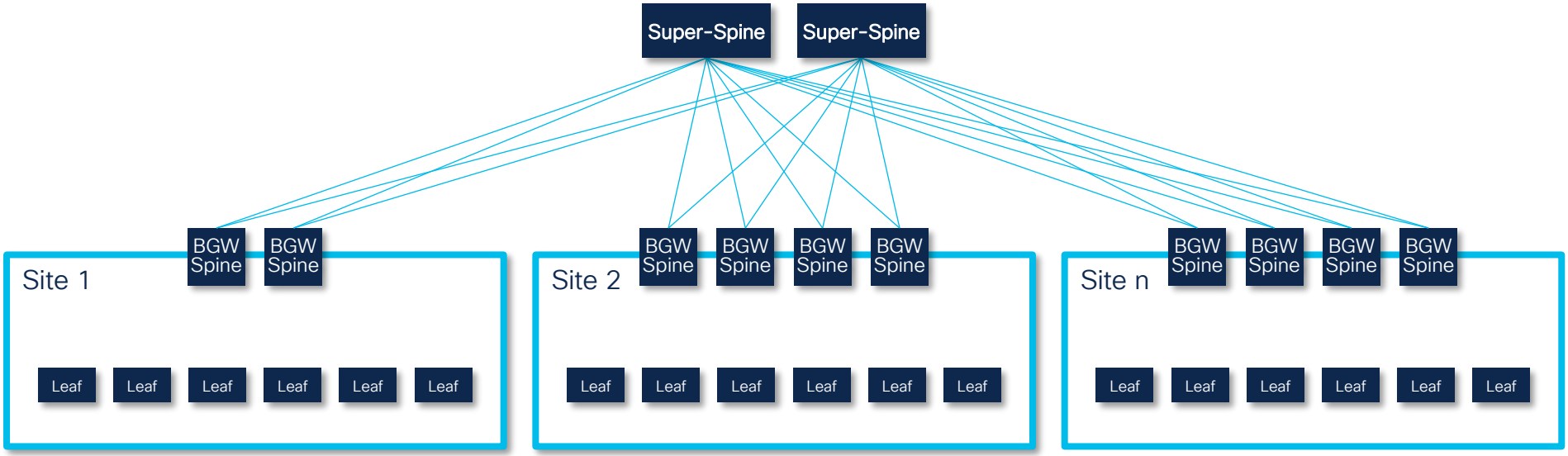
Reference

BGWs between Spine and Super-Spine



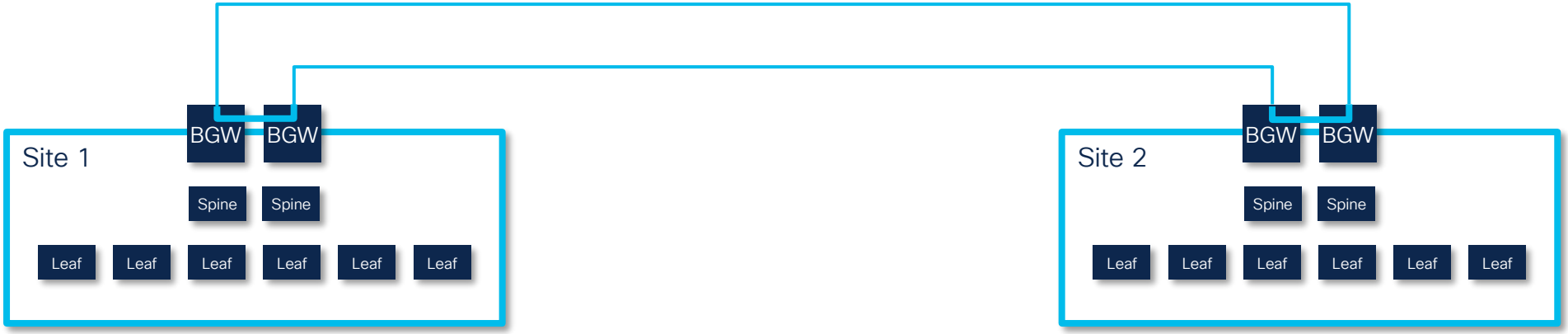
Reference

BGWs on Spine



Reference

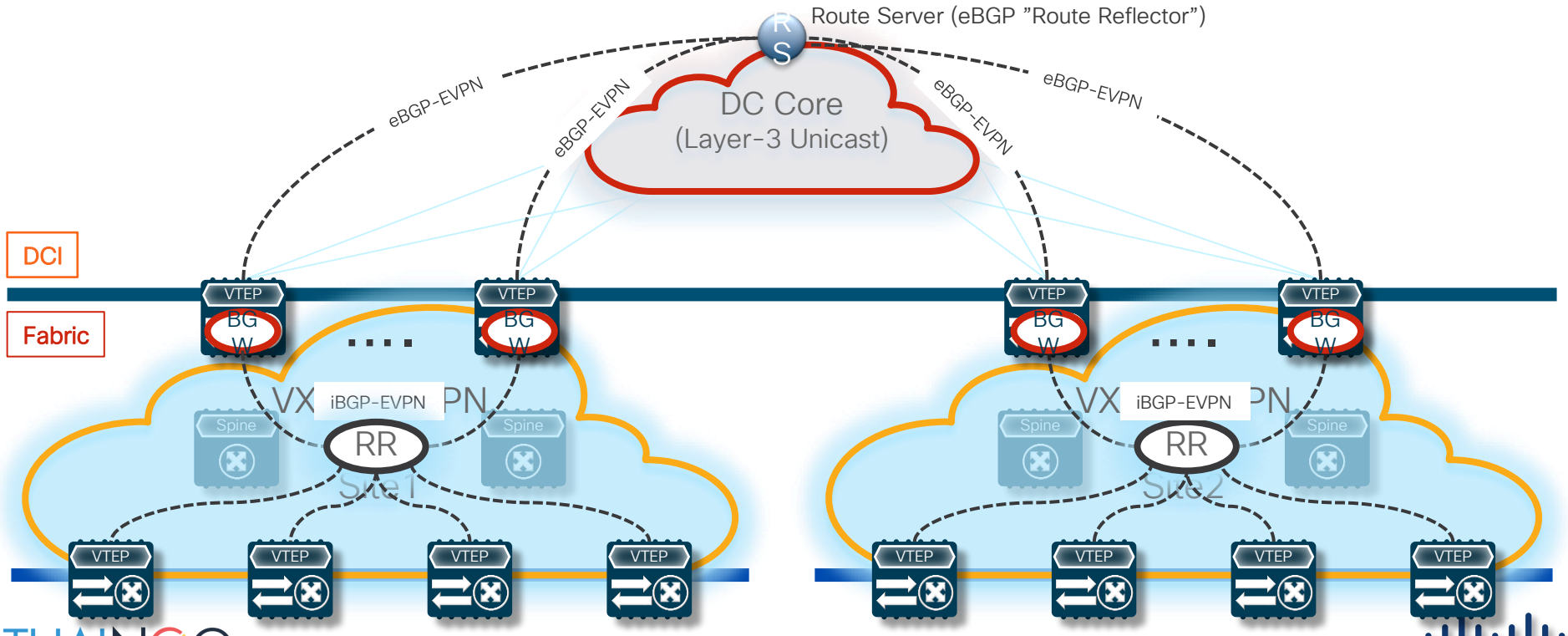
BGWs Back-to-Back



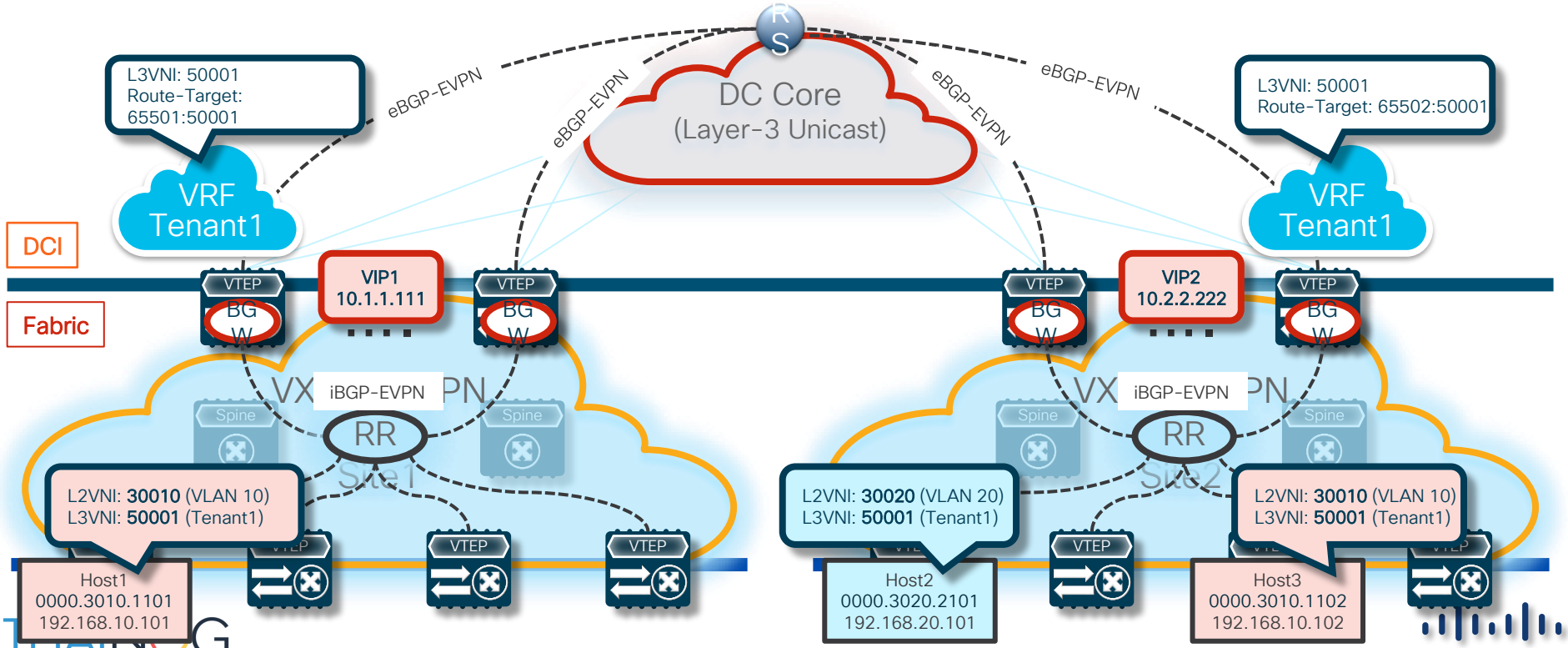
VXLAN EVPN Multi-Site Control Plane

THAINOG
Thai Network Operators Group

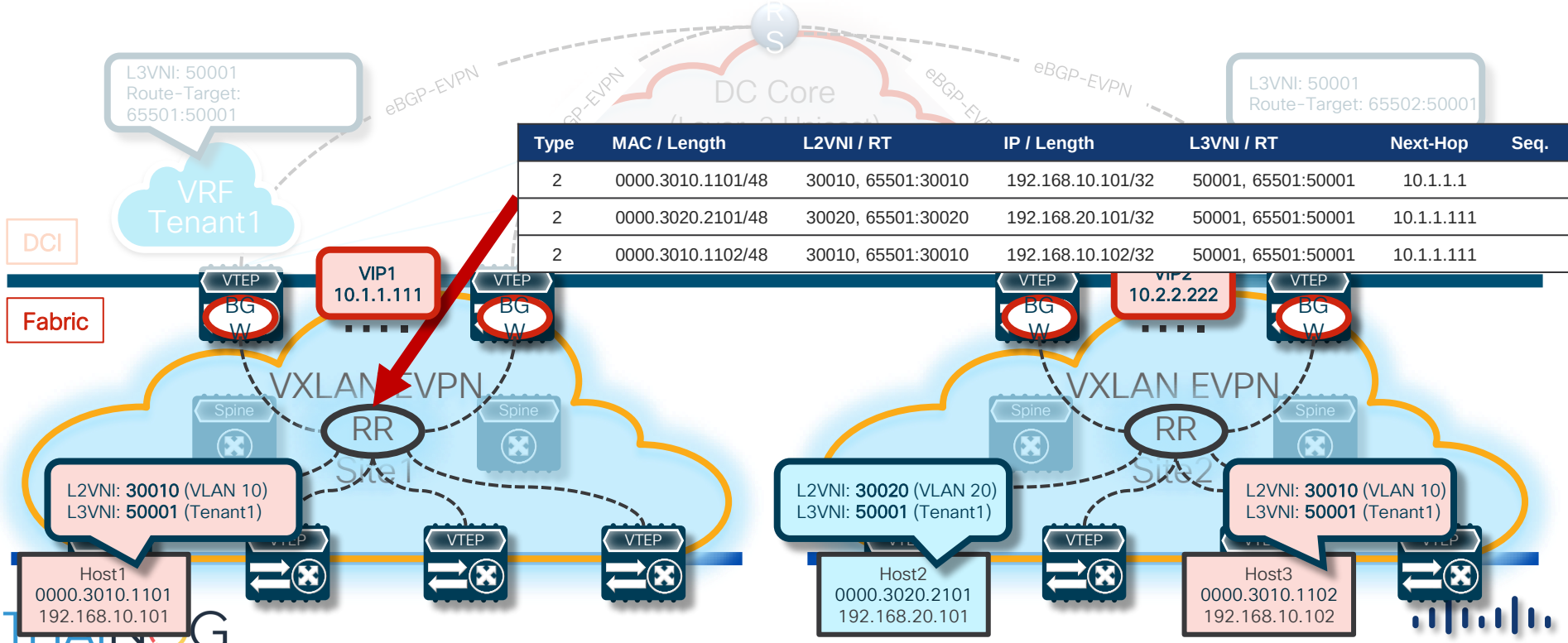
VXLAN Multi-Site
Overlay Control Plane (L3 Core)



VXLAN Multi-Site
Overlay Control Plane



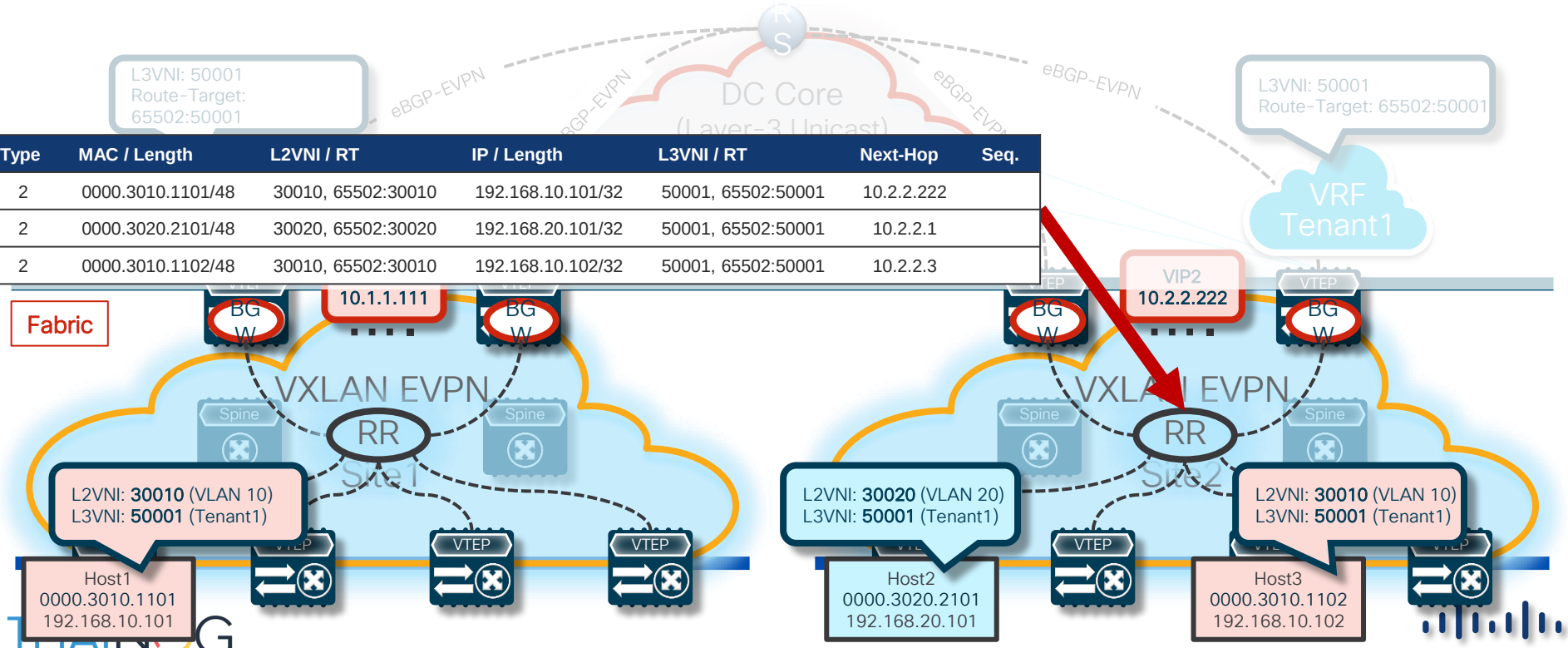
VXLAN Multi-Site
Overlay Control Plane (Site 1)



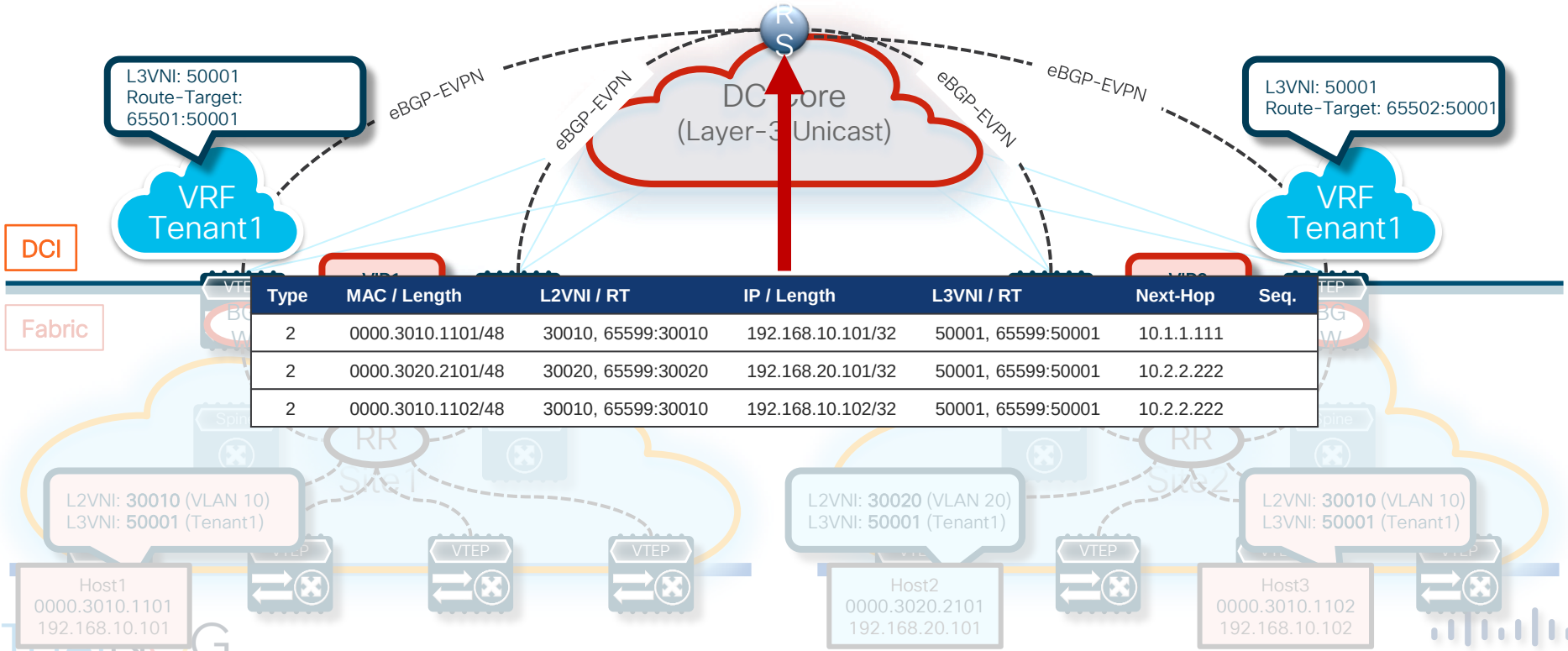
Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3010.1101/48	30010, 65501:30010	192.168.10.101/32	50001, 65501:50001	10.1.1.1	
2	0000.3020.2101/48	30020, 65501:30020	192.168.20.101/32	50001, 65501:50001	10.1.1.111	
2	0000.3010.1102/48	30010, 65501:30010	192.168.10.102/32	50001,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

VXLAN Multi-Site
Overlay Control Plane (Site 2)

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	



VXLAN Multi-Site
Overlay Control Plane (DCI)

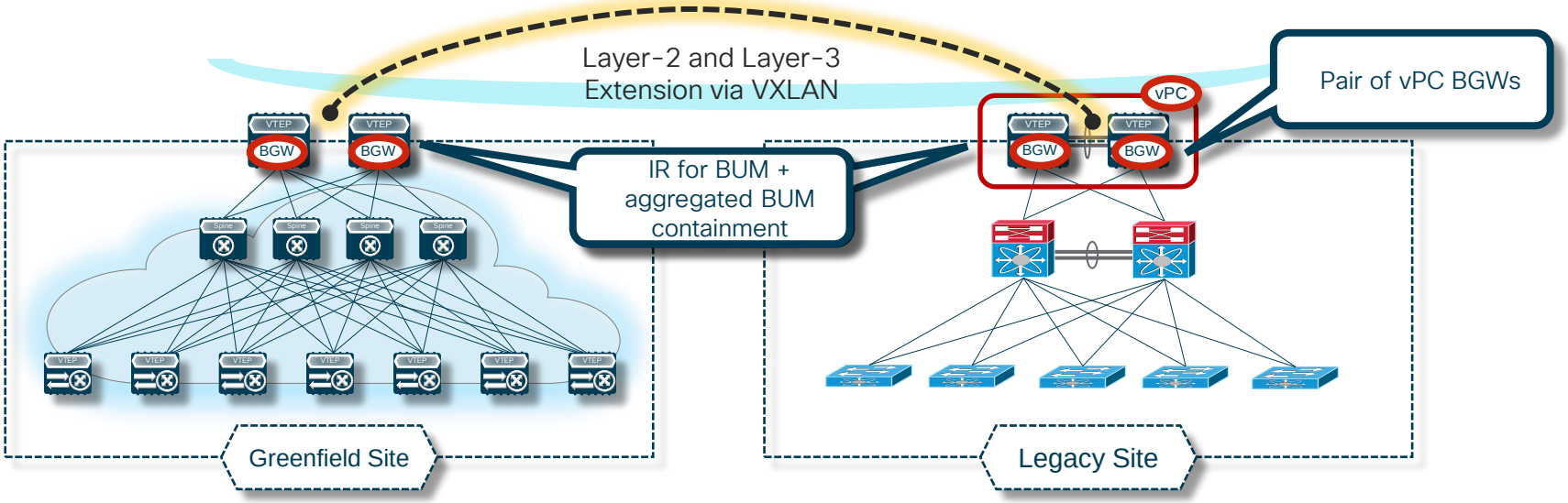


Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3010.1101/48	30010, 65599:30010	192.168.10.101/32	50001, 65599:50001	10.1.1.111	
2	0000.3020.2101/48	30020, 65599:30020	192.168.20.101/32	50001, 65599:50001	10.2.2.222	
2	0000.3010.1102/48	30010, 65599:30010	192.168.10.102/32	50001,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

Legacy Site Integration

VXLAN Multi-Site with vPC BGWs

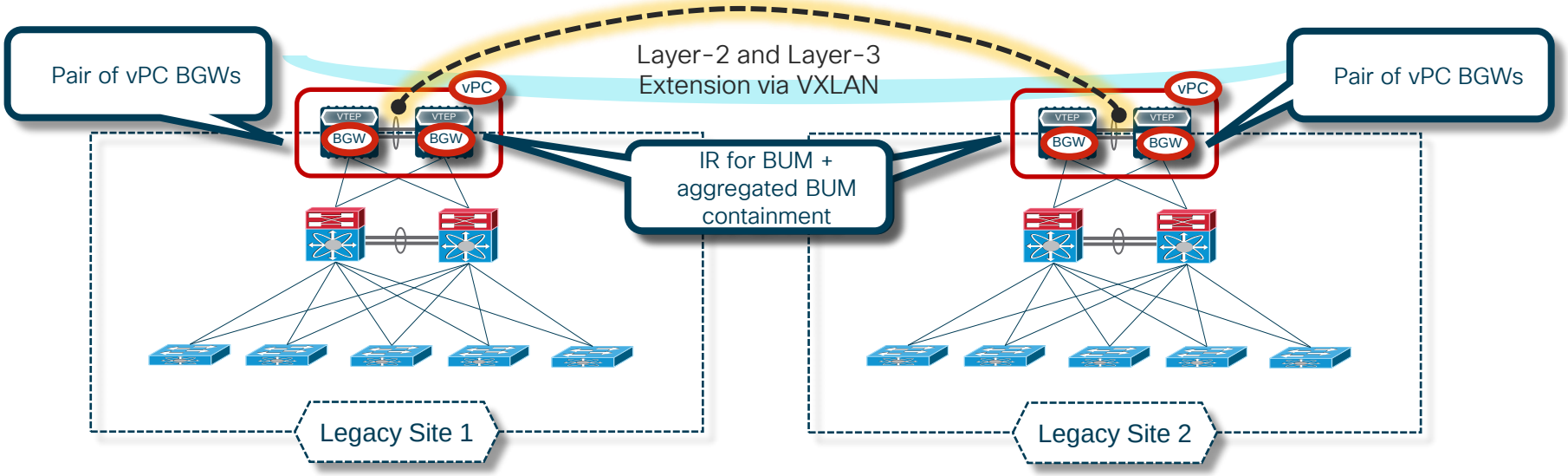
Migration/Coexistence Use Case



- Coexistence and/or migration use cases
  - Need to extend Layer-2 and Layer-3 multi-tenant connectivity across sites
- Deploy a pair of vPC BGWs in the legacy site
  - Seamless connectivity extension via VXLAN
  - Leveraging native Multi-Site functions (Ingress Replication for BUM, BUM containment, etc.)

VXLAN Multi-Site with vPC BGWs

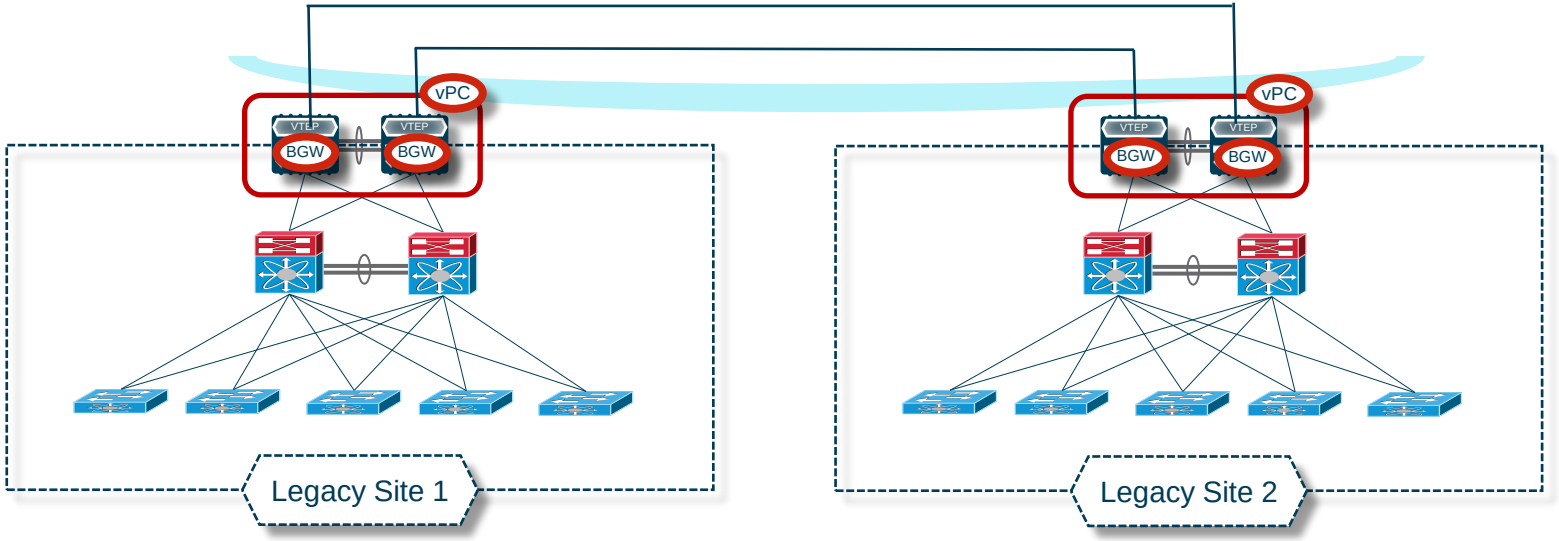
Next-Gen DCI to Interconnect Legacy Networks



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

VXLAN Multi-Site with vPC BGWs

Next-Gen DCI Use Case with Back-to-Back BGWs



- Typical topology leveraging dedicated dark fiber links or DWDM circuits
- 'Squared' and 'full mesh' topologies are both fully supported
- Recommended to limit the back-to-back deployment to two sites
  - Recommended to insert Layer 3 core network with 3+ sites

Thank you

THAINOG
Thai Network Operators Group

#ThaiNOG

#ThaiNOG