Simple Leaf Spine with a Touch of ToR

Network Designs for the Modern Data Center

Brenden Buresh
Distinguished Solutions Engineer

BRKDCN-2967

Cisco Webex App

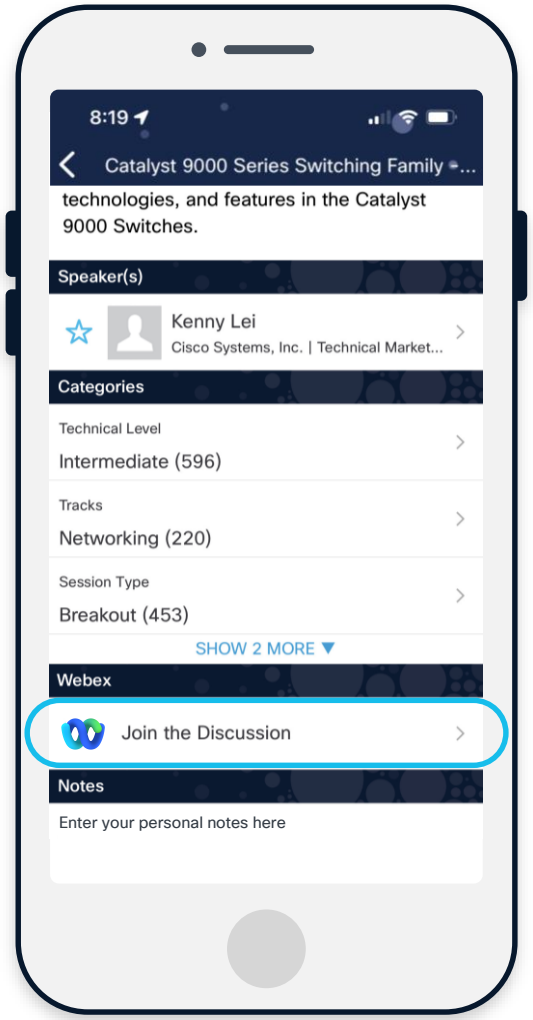
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

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

Webex spaces will be moderated by the speaker until June 13, 2025.



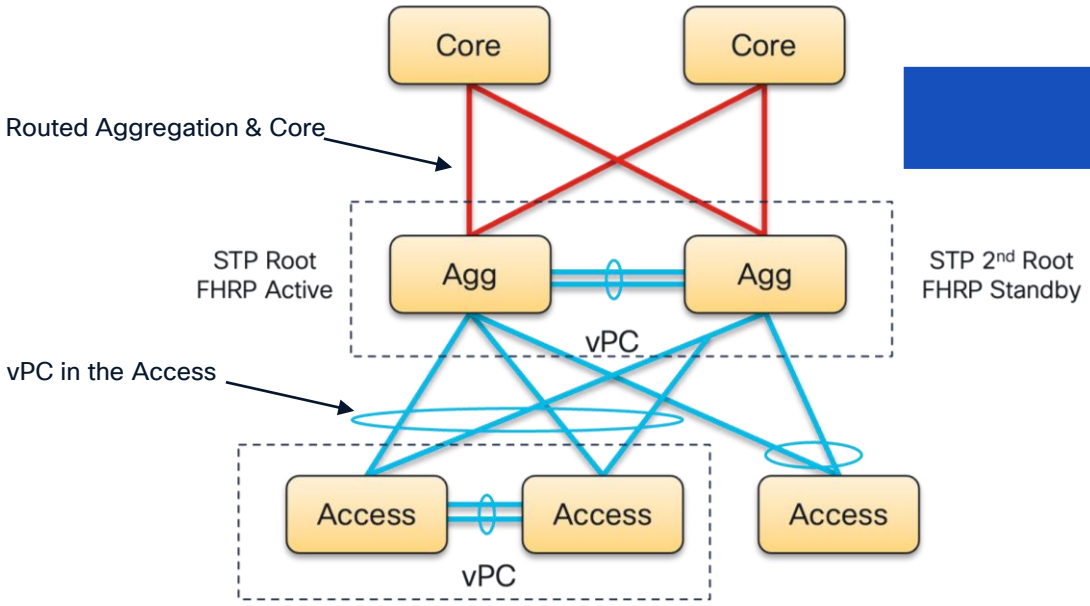
Agenda

Introduction

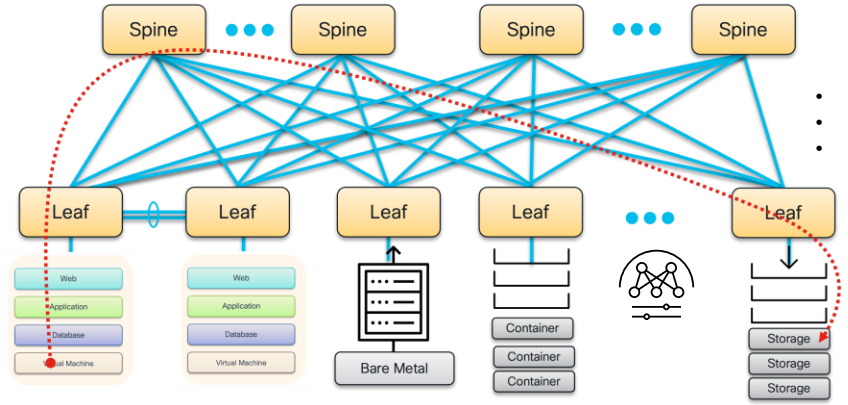
Evolving Network Designs

Traditional 3-Tier DC Network Design

vPC and Spanning-Tree

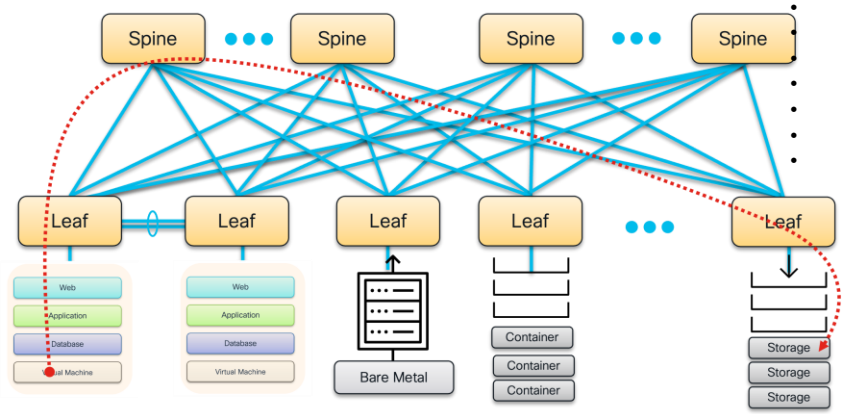


VXLAN Bridging/Routing Services



- VXLAN Flood & Learn
- VXLAN EVPN
- Separate Management Tools (e.g. Nexus Fabric Manager)

VXLAN Routing / Segmentation



- VXLAN Routing
- Fabric Controller (NDFC)
- Automation
- Management
- Visibility
- Multi-Site
- Security (GPO)

The New Cisco Nexus Dashboard

Simple, Intuitive, Powered by Cisco AI

Your key to a simple operational experience

Unify data center network architectures

Accelerate business innovation, minimize risk

Unleash the power of AI for networking

Reactive, proactive and predictive insights

Leverage OPEN protocols and standards

Provision
Once, deploy anywhere

Built-In automation | Infra-as-code

Secure
One source of truth

Segmentation | Zero-Trust

Manage
One logical network

Visibility | Maintenance

Analyze
Identify, troubleshoot, and fix faster

Tools | Reports | Prediction

The Evolution of DC Network Designs

Data Center Network Challenges

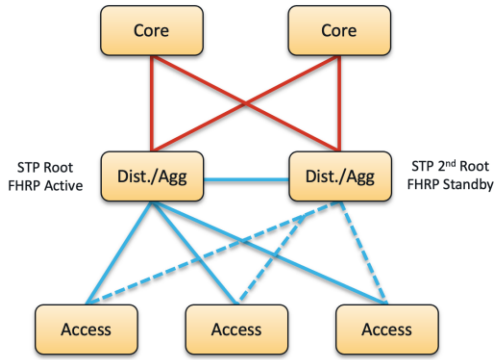
Legacy Architectures

Convergence Time:
STP high convergence time
TCN MAC Flush

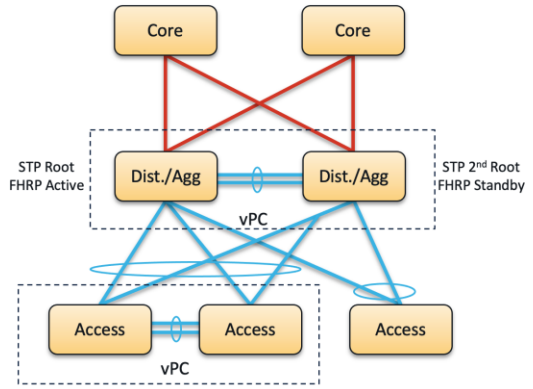
Unused Links:
STP Redundant links in blocked state

Rigid Network Services:
L4-L7 Services placed at Distribution Layer

Classic Spanning-Tree

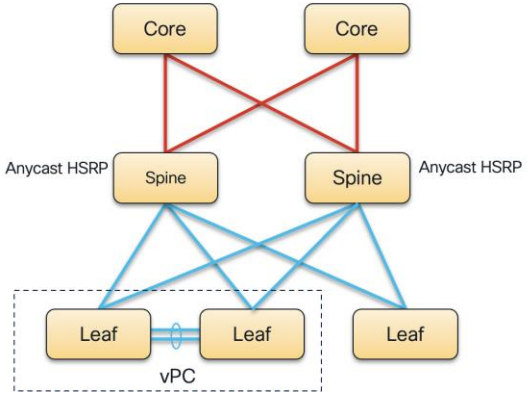


vPC and Spanning-Tree

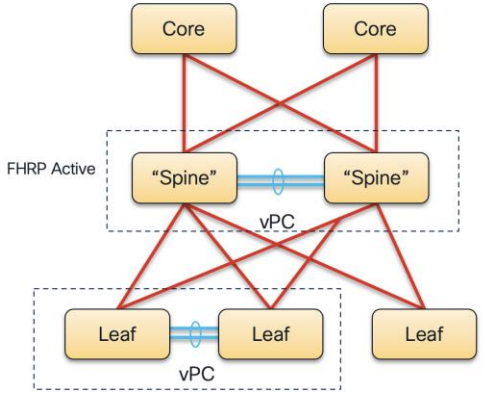


Suboptimal Forwarding:
Path defined by
Root Switch or FHRP Active

FabricPath (MAC-in-MAC)



VXLAN Flood & Learn (MAC-in-IP)



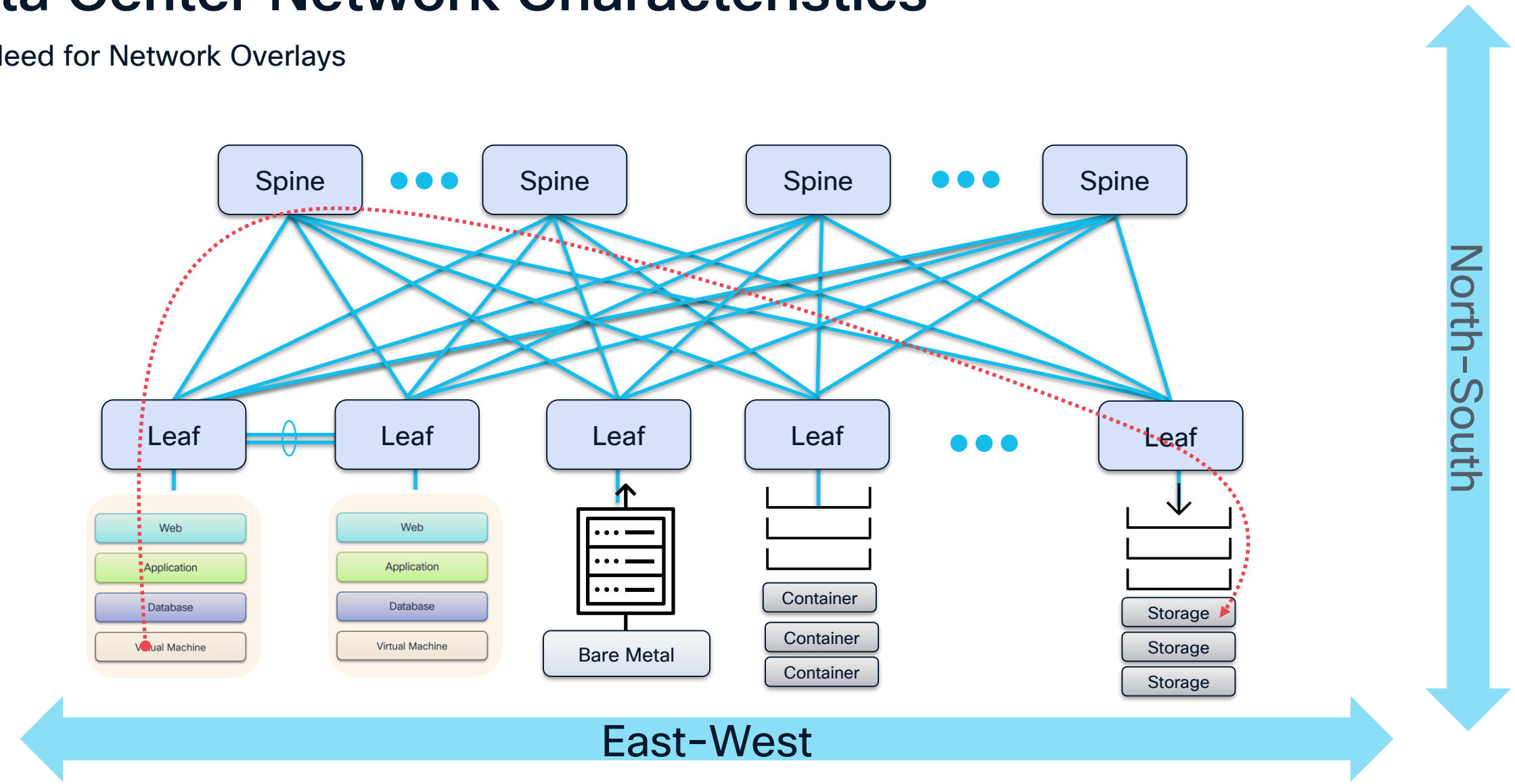
OpEx/CapEx:
Expensive to upgrade
Scale up with large Chassis

Limited Scale:
No Control Plane
Limited Workload Mobility

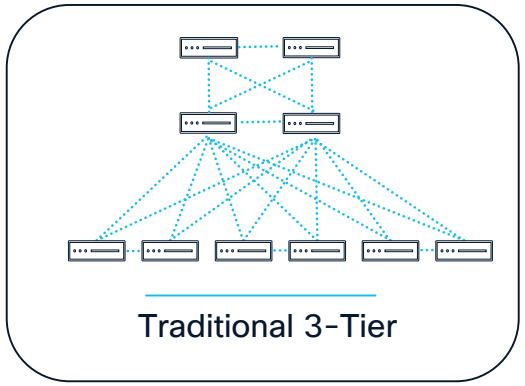
"Spine" = Not Really a Spine

Data Center Network Characteristics

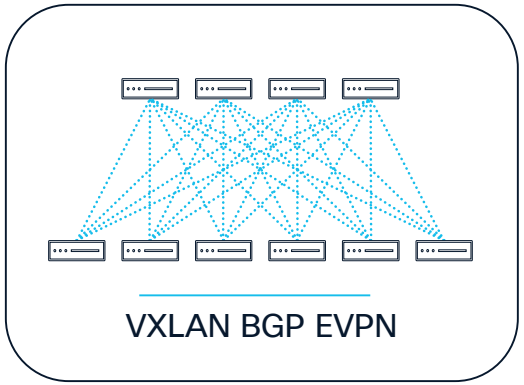
The Need for Network Overlays



Customer Use Cases for Data Center Fabrics



- ✓ Large broadcast/fault domains to support VM mobility
- ✓ Potential Suboptimal Forwarding
- ✓ Scale up behavior
- ✓ Relies on STP to ensure loop-free topology
- ✓ Disjointed Data Center Interconnect (DCI)



- ✓ Any workload anywhere across L3 boundaries
- ✓ Seamless VM Mobility (intra and inter-DC)
- ✓ Scale out behavior aligned to DC traffic patterns
- ✓ Leverages ECMP for optimal path over L3 network
- ✓ Next Gen Data Center Interconnect (DCI)

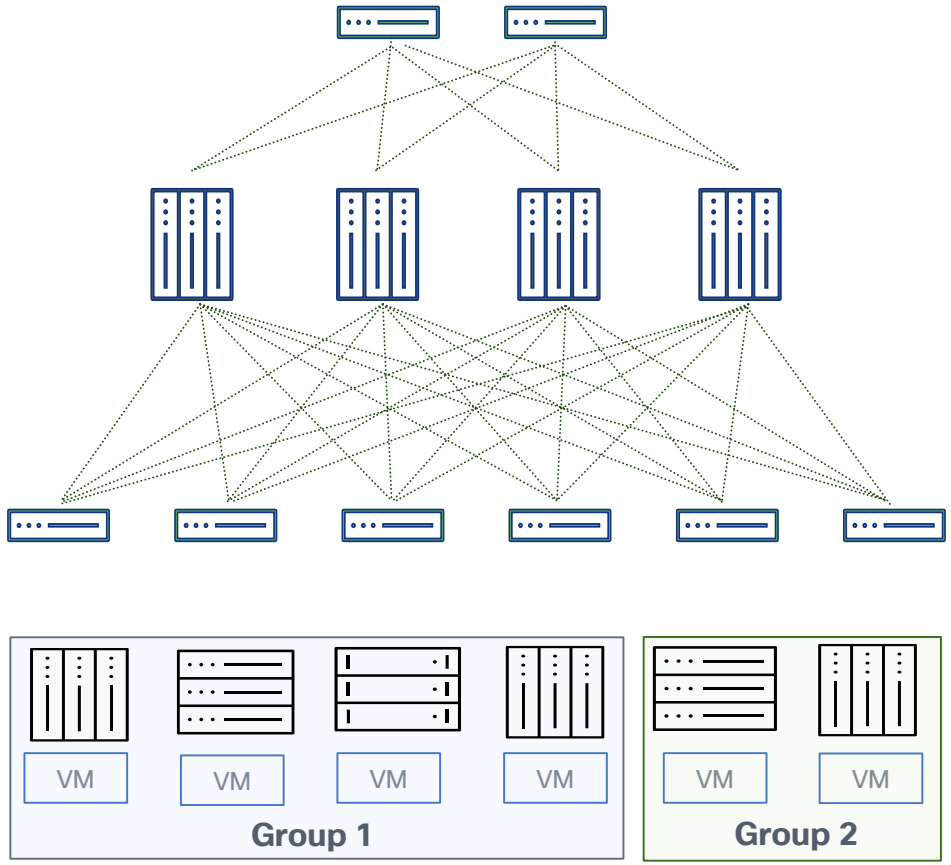
Design Building Blocks for VXLAN Fabrics

BGW

Spine

Leaf

Security



Border Gateways

- Dedicated or Shared Devices
- External Connectivity (WAN)
- East-West DCI Connectivity

Spines

- Fabric Bandwidth (i.e. 400G)
- High Availability (# of Spines)
- Scalability (Super Spines)

Leafs

- Fabric Capacity (# of Ports)
- Fabric Connectivity (i.e. 25G)
- Host Redundancy (VPC/ESI)

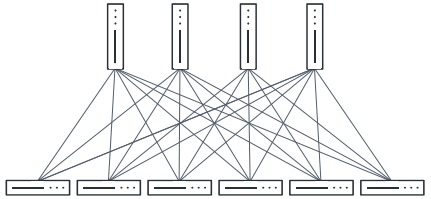
Security

- Network Segmentation (GPO)
- Device Segmentation (FW)
- Host Segmentation (Agent/DPU)

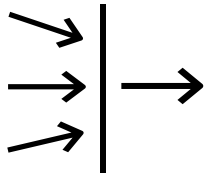
Leaf-Spine ”Fat Tree” Network Topologies

Key Requirements

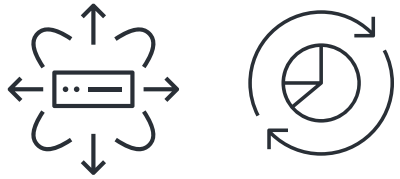
Topology
(Resiliency)



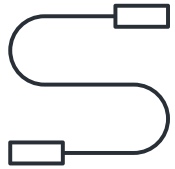
Bandwidth
(Oversubscription)



Platform Hardware
(Cost)



Optics
(Cabling)

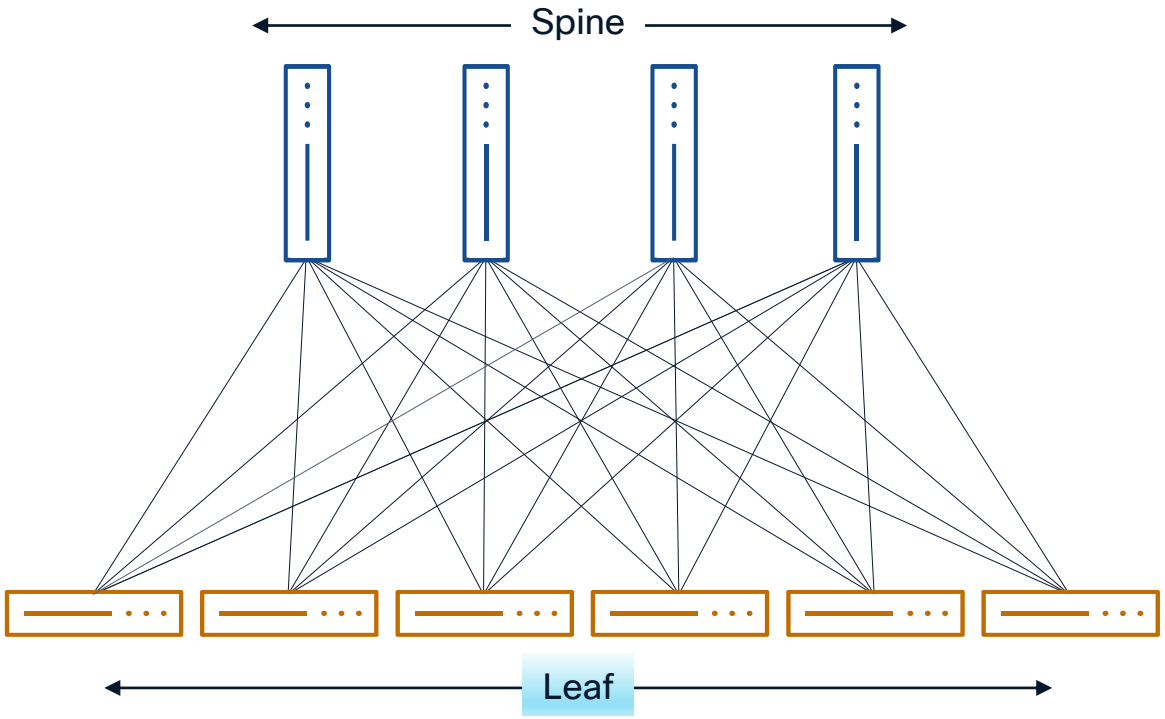


Define Requirements Before Designing Solution

Clos Topology

A Leaf and Spine Topology

- Variations and Names of the Same
  - 2-Tier
  - Fat Tree
  - Folded Clos
  - 3-Stage Clos
  - Butterfly (5-Stage | POD)

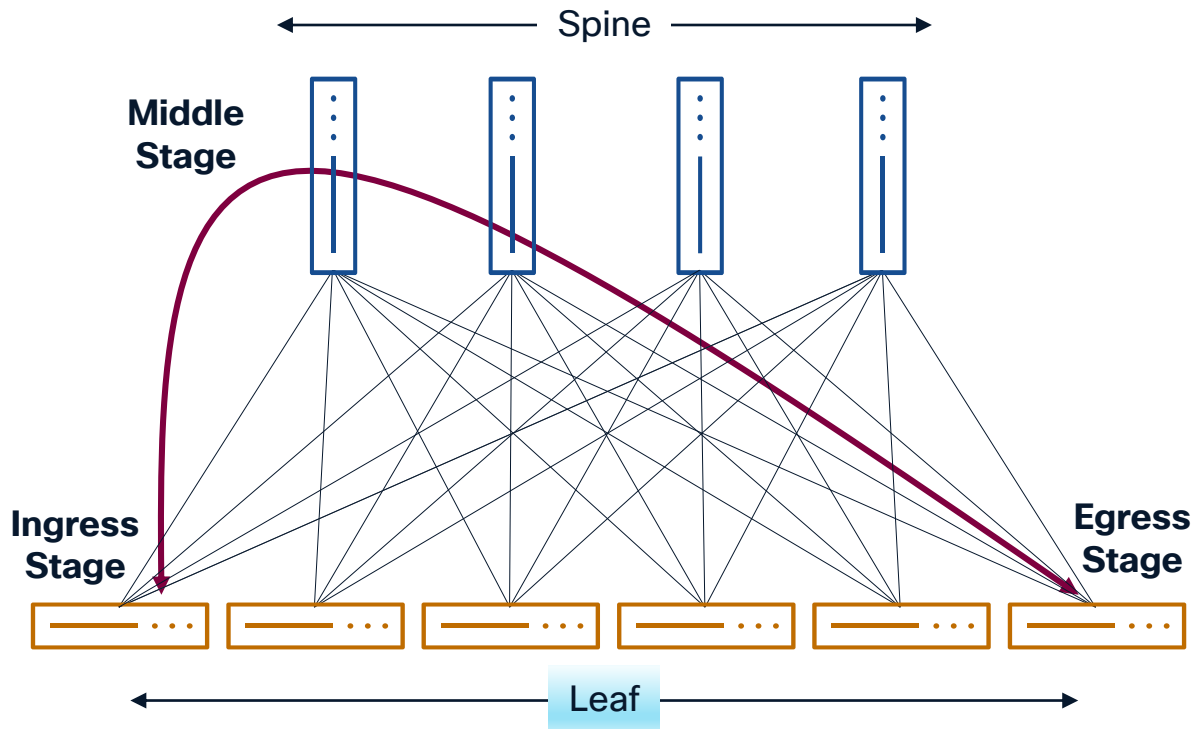


Clos Topology

Advantages

A Leaf and Spine Topology

- 3 Stages
- 2 Tiers

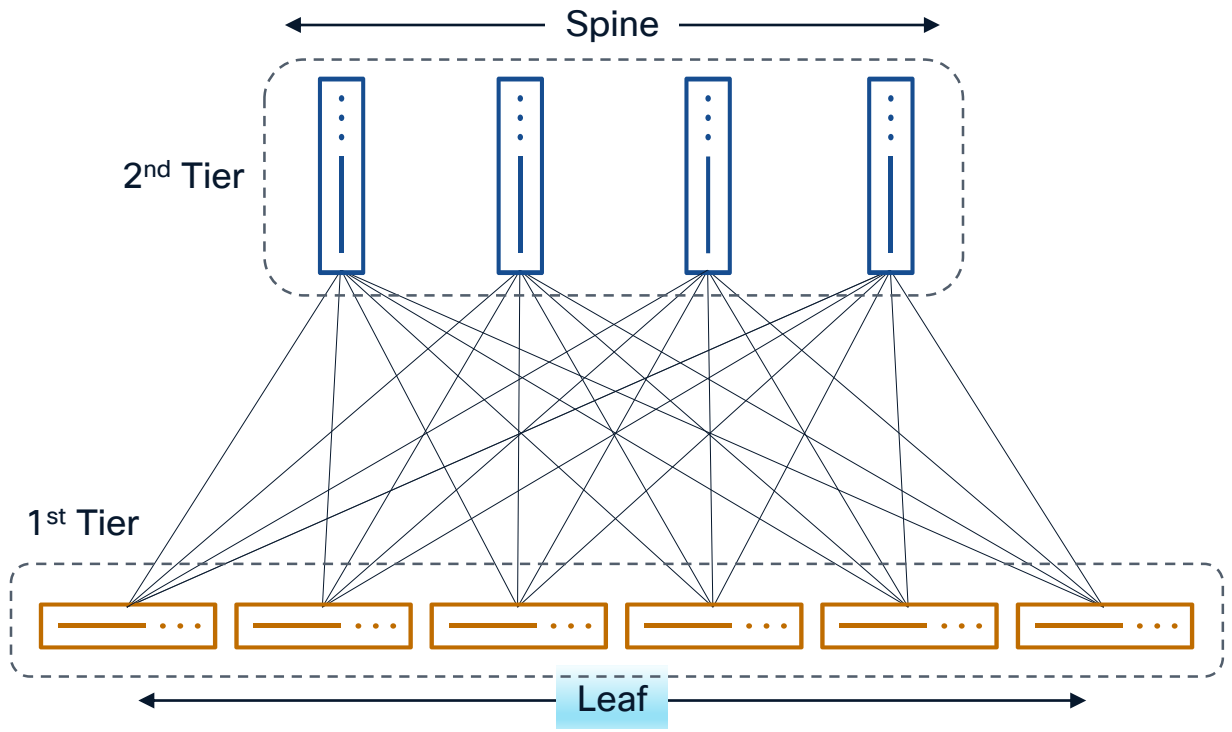


Clos Topology

Advantages

A Leaf and Spine Topology

- 3 Stages
- 2 Tiers



Clos Topology

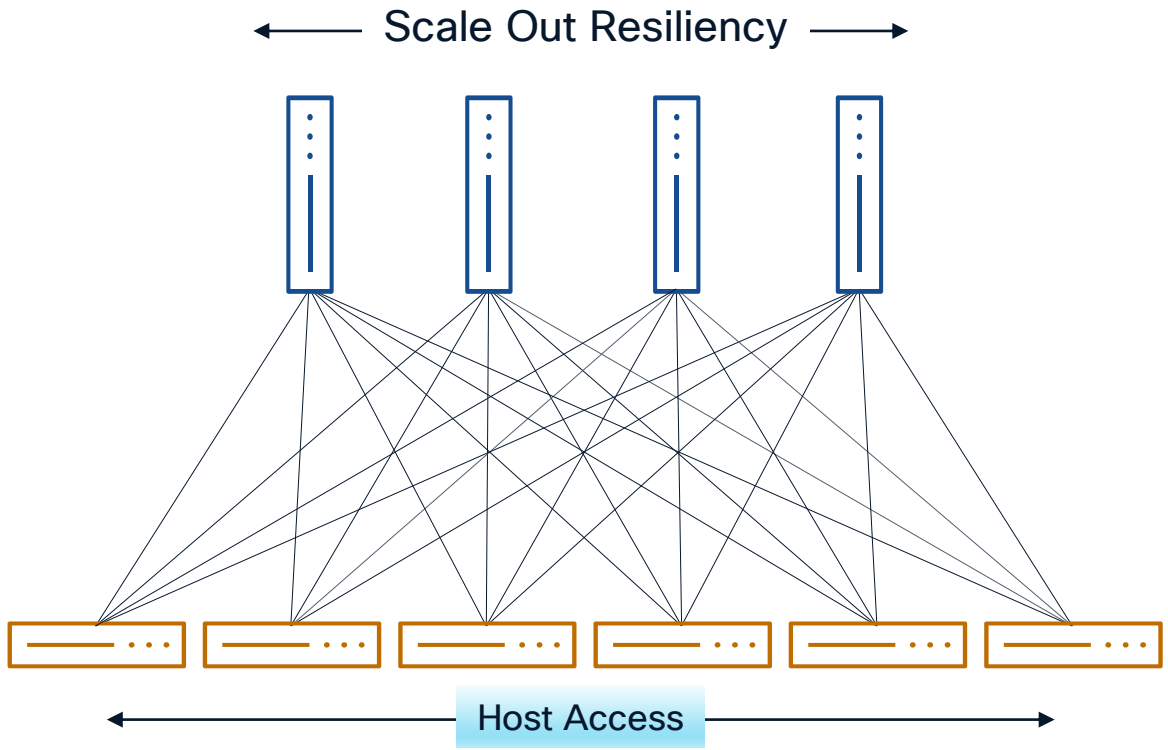
Scale Out Architecture

- Adding Spines Results in:
  - Increased physical path redundancy

    -> Greater Resiliency
- Adding Leafs Results in:
  - Increased port capacity *without increasing oversubscription ratio(s)*

    -> More Access Ports

Advantages

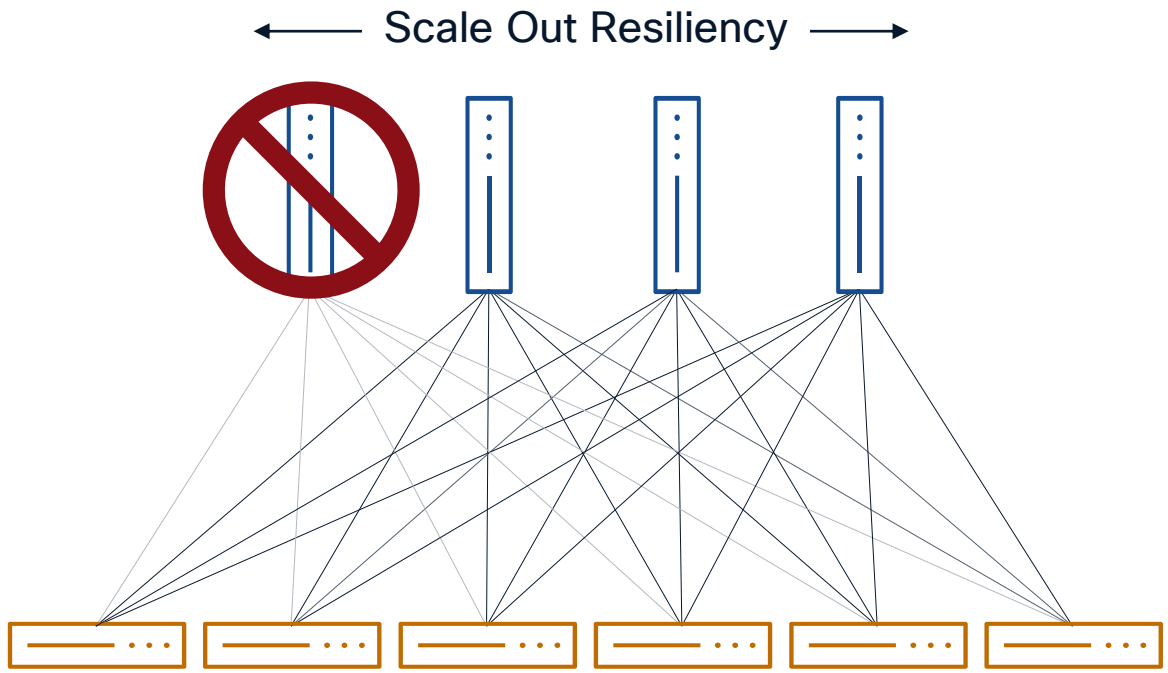


Clos Topology

Advantages

N+1 Redundancy

- Resiliency increases when adding redundancy to the topology
- Single spine failure
  - 4 spines = 25% impact
  - 8 spines = 12.5% impact
  - 12 spines = 8.3% impact

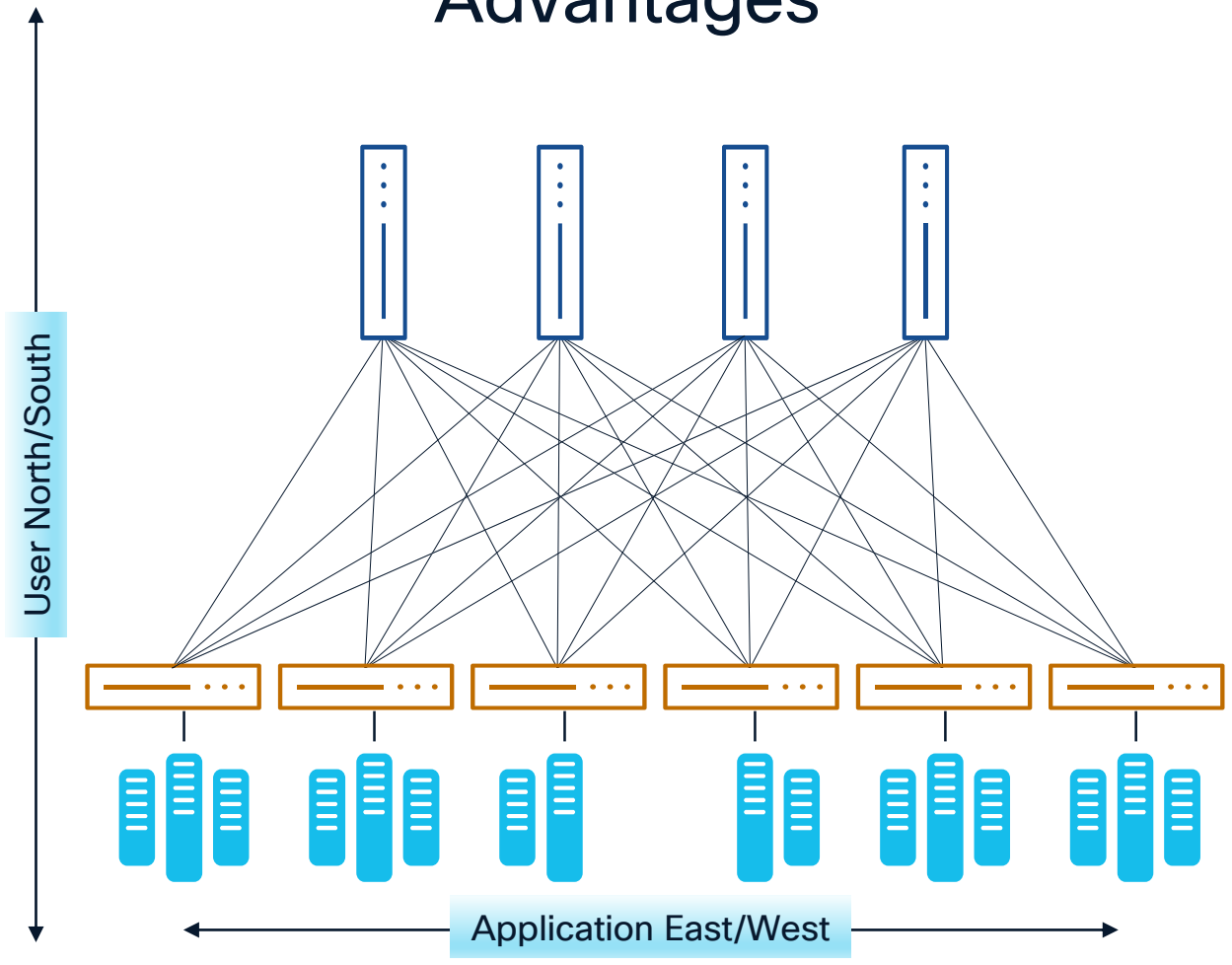


Clos Topology

Modern Application Needs

- Every (1) North South connection requires (8) East West connections
  - Application front end (user access)
  - Application back end (database, storage, etc.)

Advantages

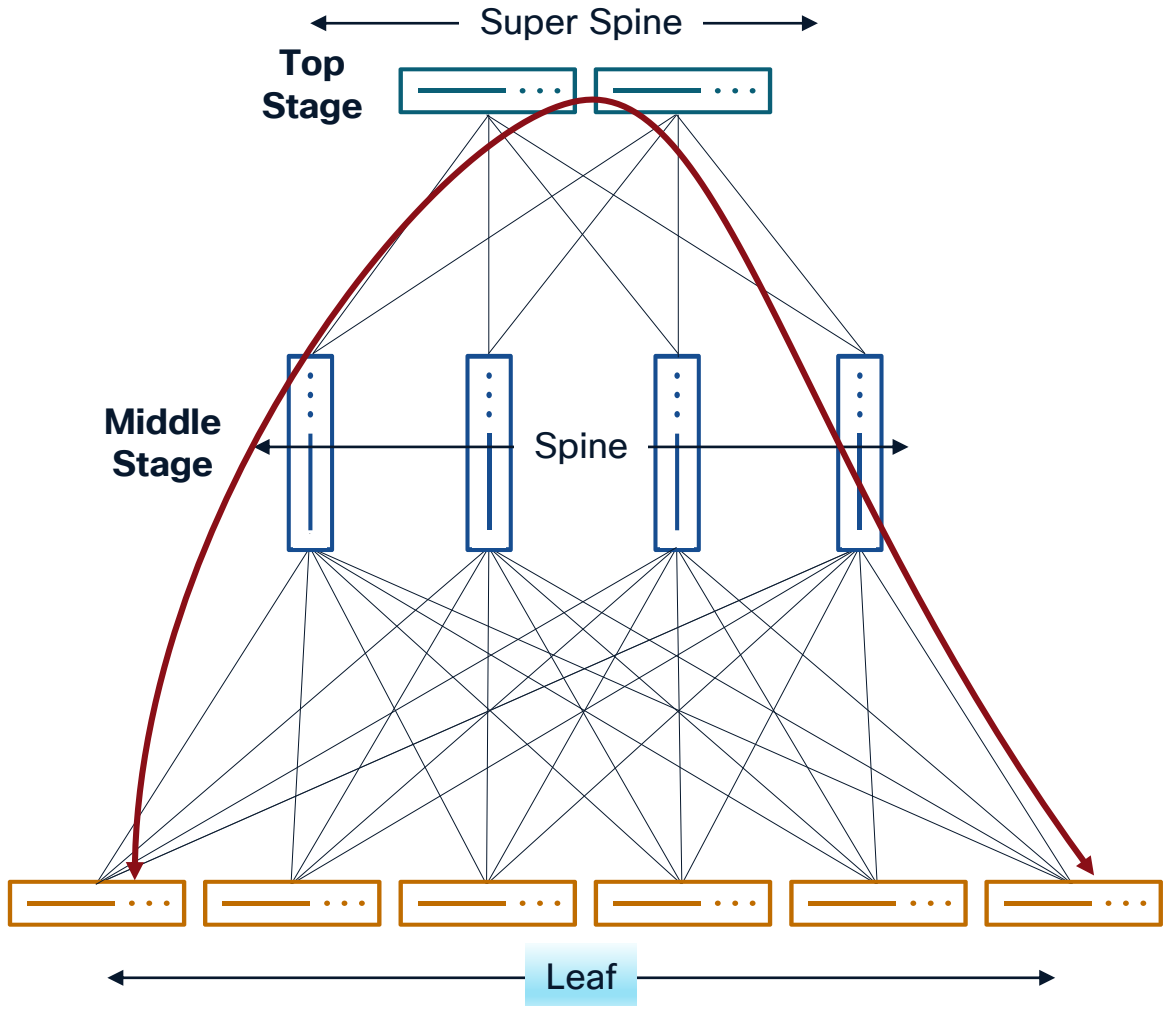


Clos Topology

A Leaf and Spine Topology

- 5 Stages
- 3 Tiers

Advantages



Oversubscription

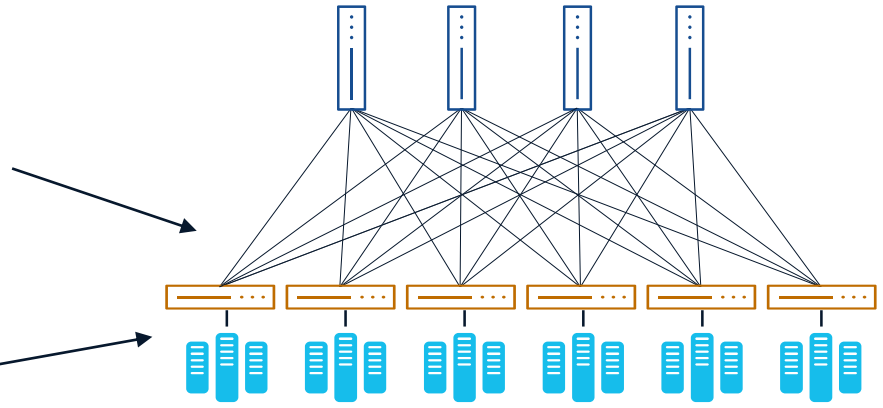
Oversubscription happens when an intermediate network device (switch) or link doesn't have enough capacity (bandwidth) to allow line rate communication between the two devices

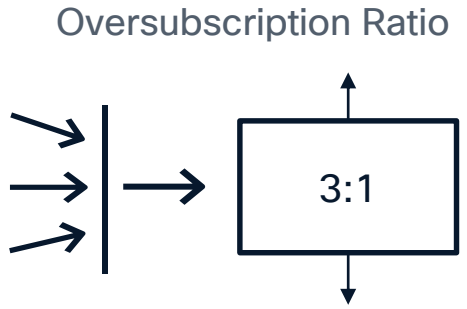
Oversubscription Ratio

[Number of Downlink Ports * Speed] ÷ [Number of Uplink to Spine * Speed] : 1

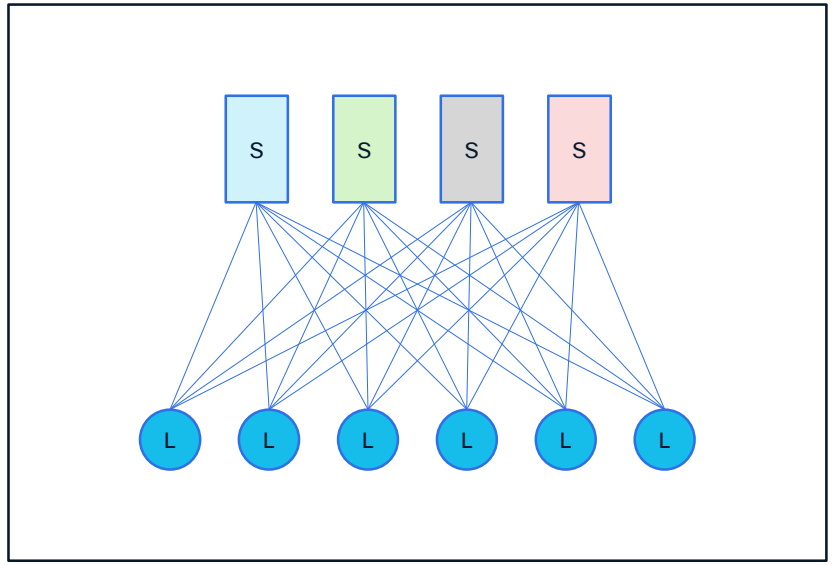
Uplink Ports
4x100Gbps = 400 Gbps

Downlink Ports
48x25Gbps = 1.2Tbps

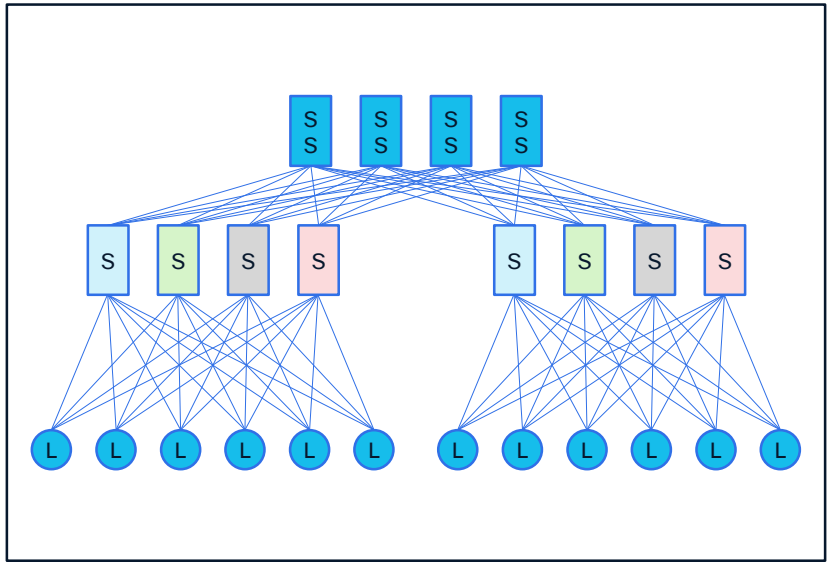




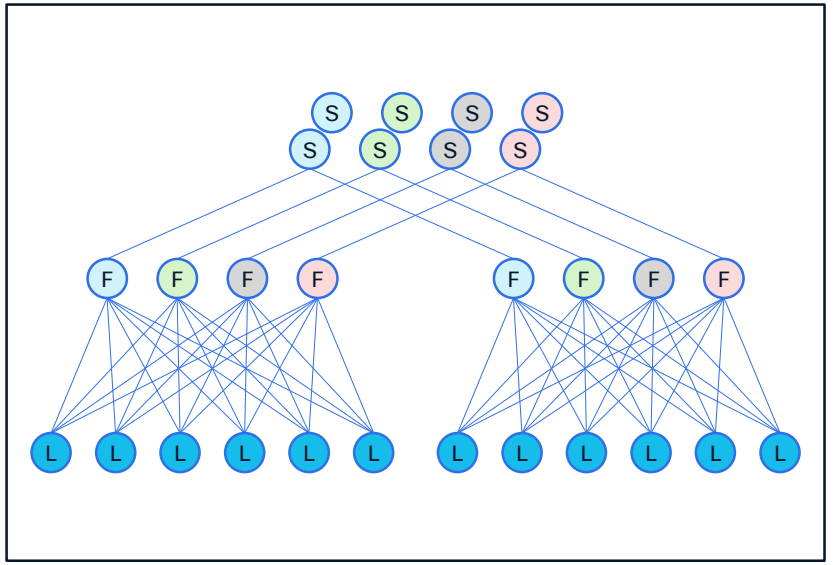
The Journey to Build Better and Further



2 Tier Leaf Spine



3 Tier Leaf-Spine-SuperSpine

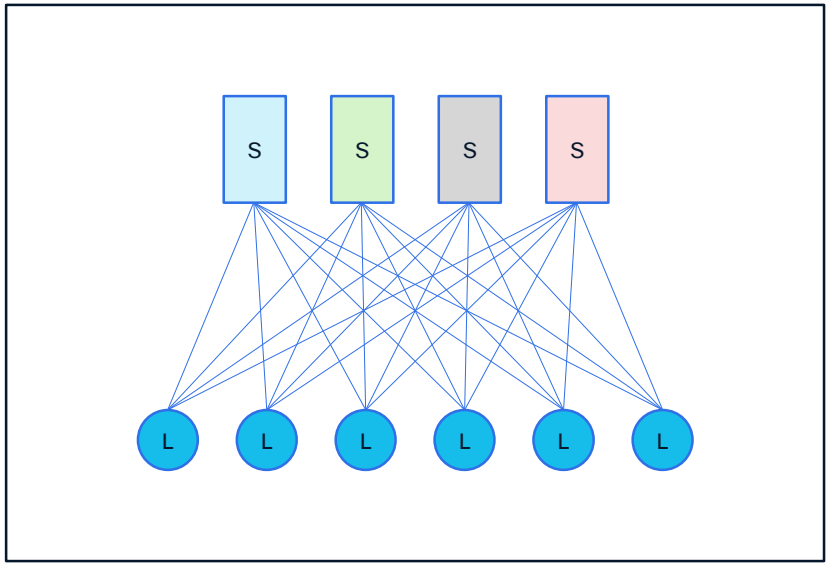


3 Tier Leaf-Fabric-Spine

Standard Design

2-Tier Leaf Spine

- A Perfectly Valid Way
- Tends to have “Finite Scale”
  - Maximum Chassis Capacity
  - Maximum Speed per Port
- Many Locations of Redundancy
  - Redundant Chassis Components
- Condensed Link and Bandwidth Presence
  - Aggregated within a Chassis



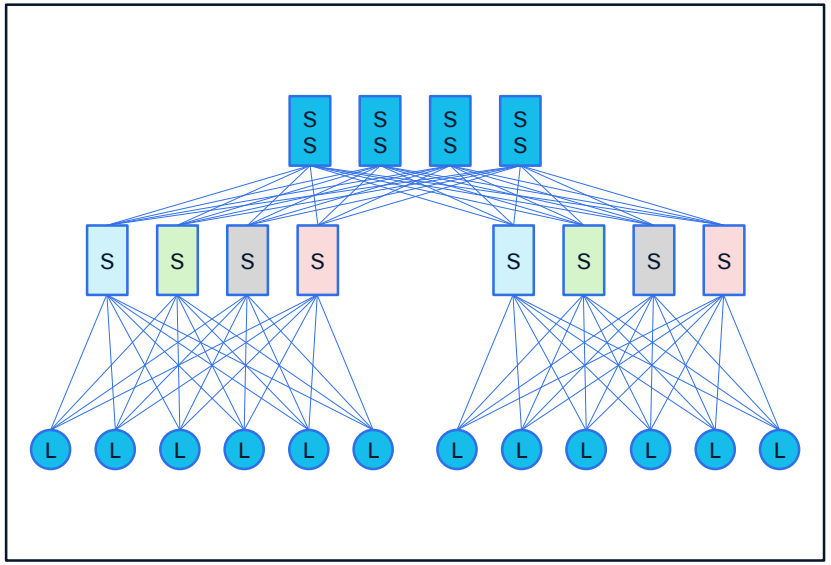
2 Tier Leaf Spine

- Use Modular Chassis at Spine
- Use Higher Density Linecards
- Use Higher Bandwidth per Port

Expanding Scale

3-Tier Multi Site

- Avoiding Scale-Up with Another Tier
- Distributed Link and Bandwidth Presence
  - Disaggregated Across Tiers
- Increases the “Finite Scale”
  - No Dependency on Chassis Capacity or Speed per Port
- Many Locations of Redundancy
  - Redundant Chassis Components
- Allows for Cost Optimization



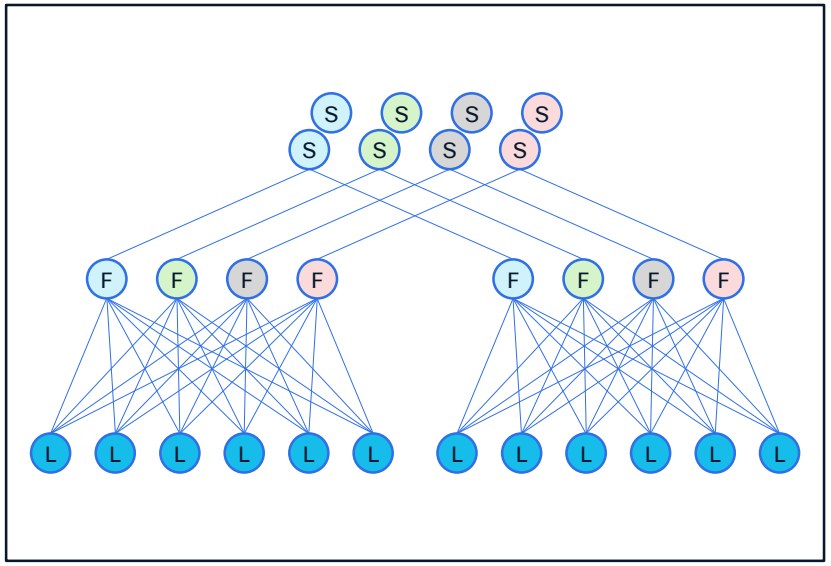
3 Tier Leaf-Spine-SuperSpine

- Scale-Out; Introduce a 3rd Tier
- Interconnect multiple 2 Tier “PODs”
- Use Modular or Fixed Spine & SuperSpine
- Use High Port Density
- Use High Bandwidth per Port

Increasing Resiliency

Multiplanar Architecture

- Increasing Scale-Out in All Tiers
- Reduce to the Max
  - Simple Design Principles
- Increases the “Finite Scale”
  - Scale as You Go
- Disaggregated Redundancy
- Flexible Link and Bandwidth Distribution
- Further Opportunity for Cost Optimization



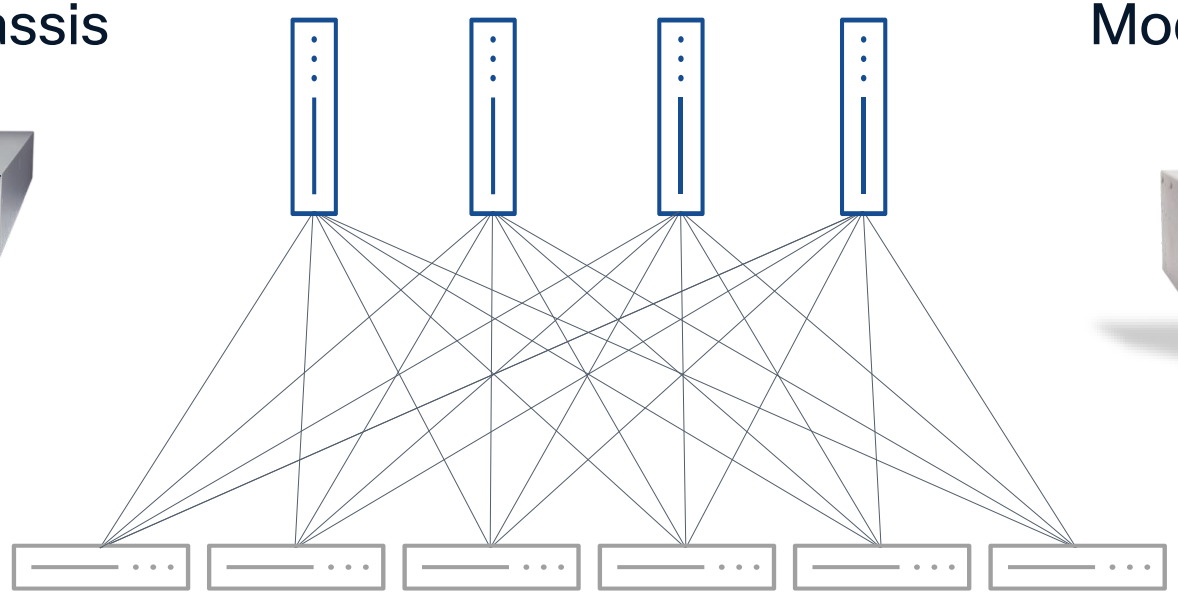
3 Tier Leaf-Fabric-Spine

- To Infinity and Beyond

Hardware Deployment Considerations

Hardware Components

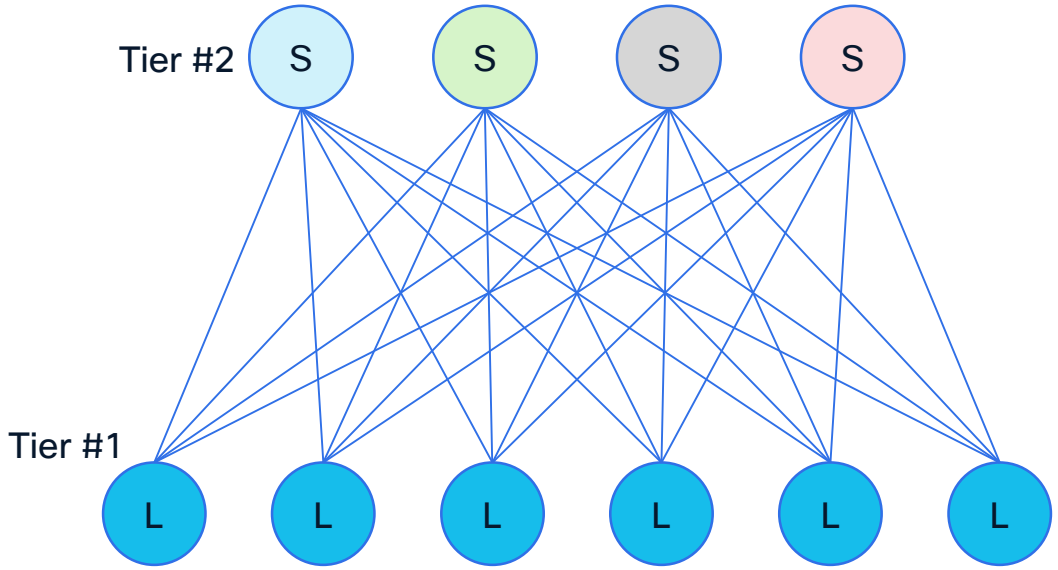
What should I use for **Spines**?
Modular or Fixed Chassis



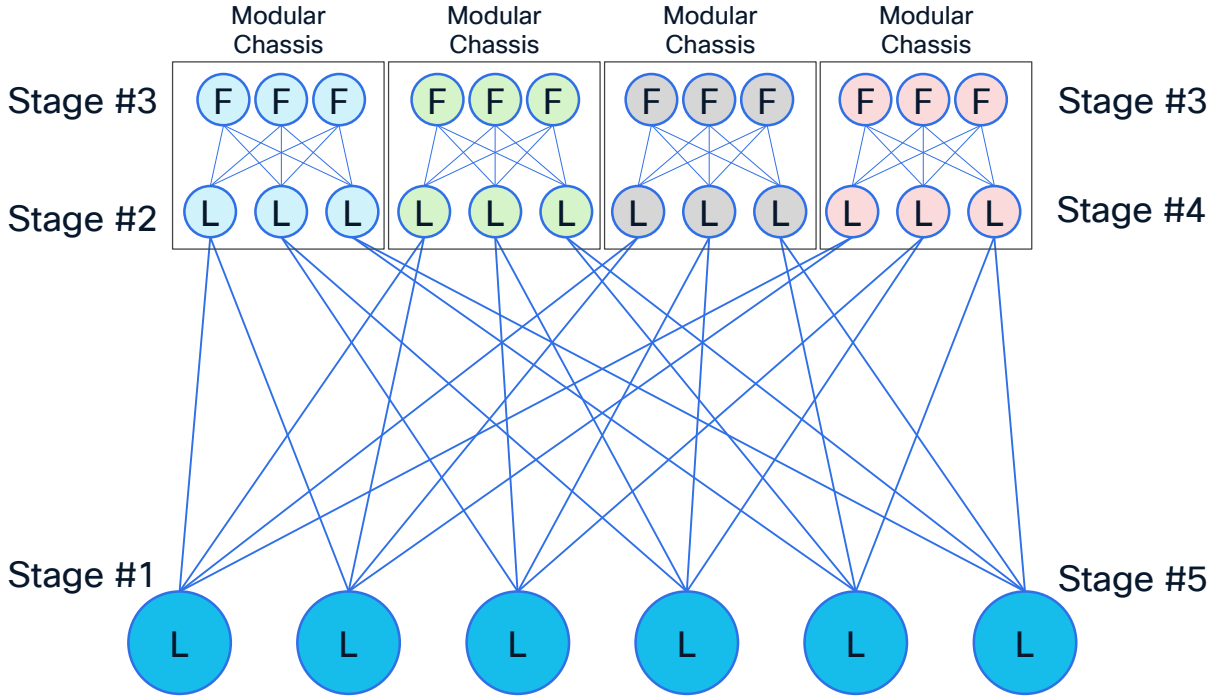
What should I use for **Leafs**?
Modular or Fixed Chassis

Resiliency	Oversubscription	Cost	Cabling	Power

Attributes to Scale



What you think you Built
2 Tier Leaf and Spine Network (3 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)



What you really Built
2 Tier Leaf and Spine Network (5 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)

Optics

Direct Attach Cables (**DAC**) for Server Connectivity

Active Optical Cables (**AOC**) for Server Connectivity

Multimode Fiber (MMF) Optics for **Fiber Re-Use** <100m

Single Mode Fiber (SMF) Optics for **New Deployment** >100m

Cable Assembly

1:1 and Assembly Breakout

Individual Optics

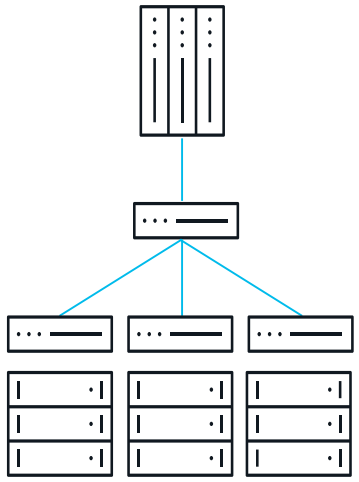
1:1 and Panel Breakout

Resiliency | Oversubscription | Cost | Cabling | Power

Optics – Panel Breakout

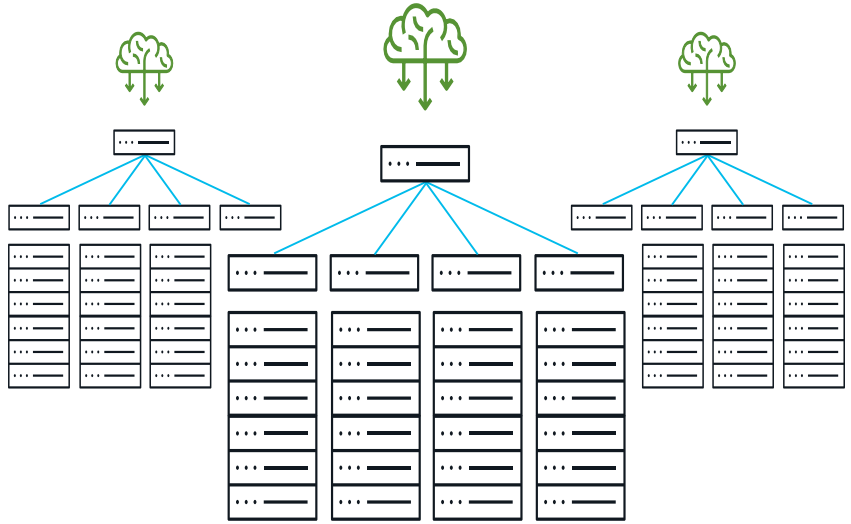
Multi-Tier Fabrics

Leaf | Spine | Super Spine



Large Scale AI/ML Clusters

IP Routed Fabric | Non-Oversubscribed

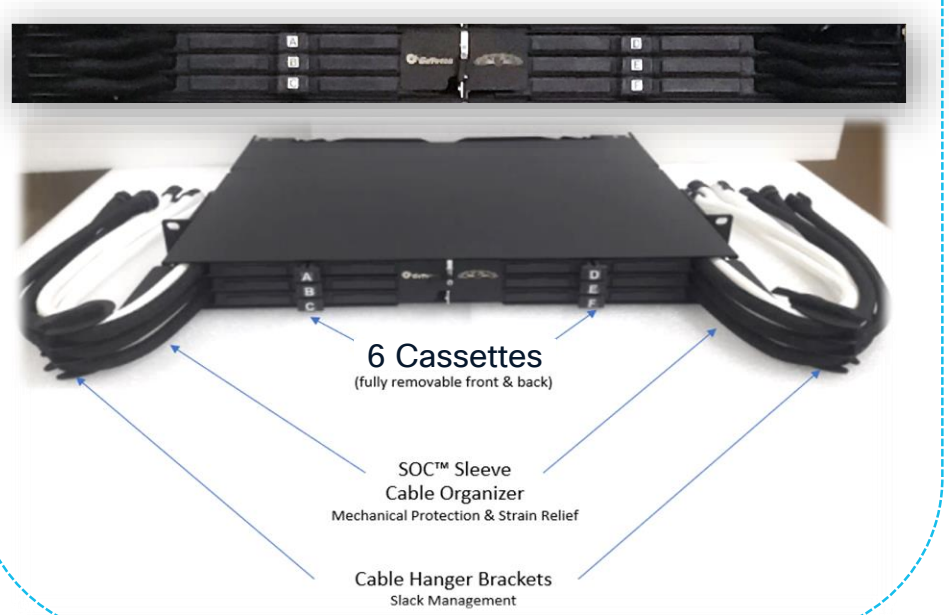


Cisco Patch Panel

400G MPO -> 4x 100G LC Breakout

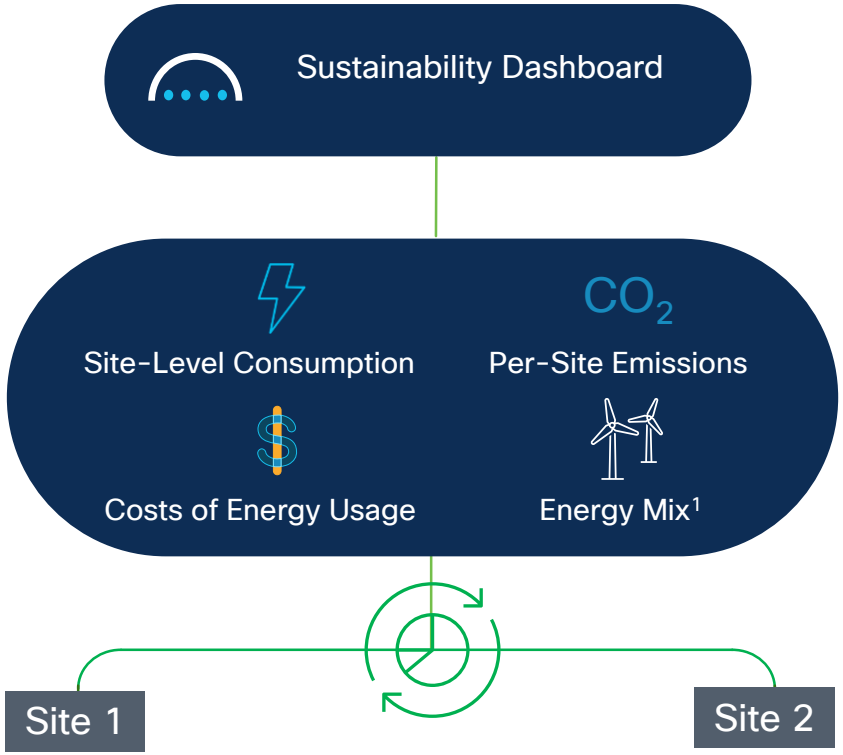
72x 100G Ports Per Rack Unit

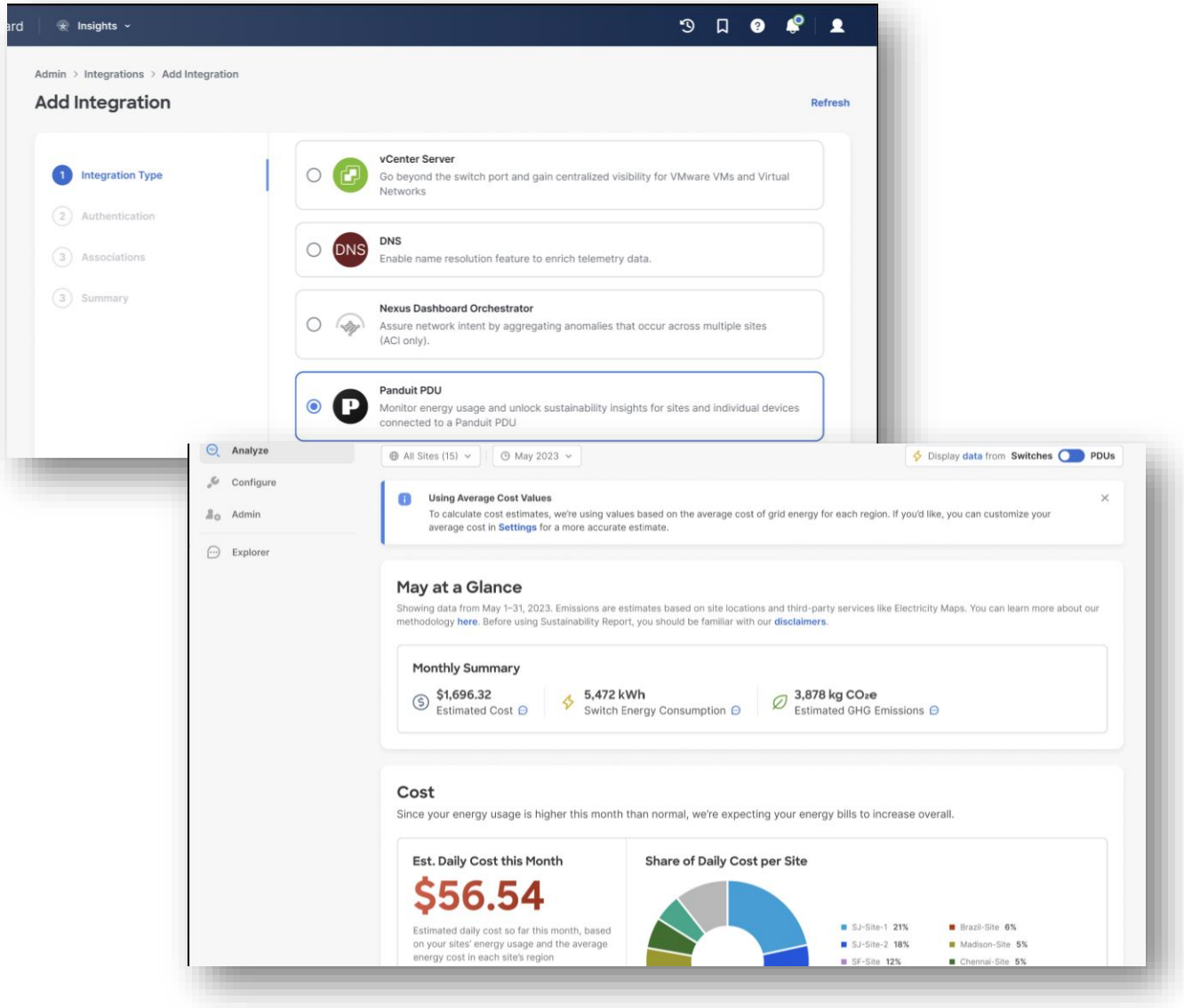
Available in 1RU / 2RU / 3RU



Resiliency | Oversubscription | Cost | Cabling | Power

Sustainability





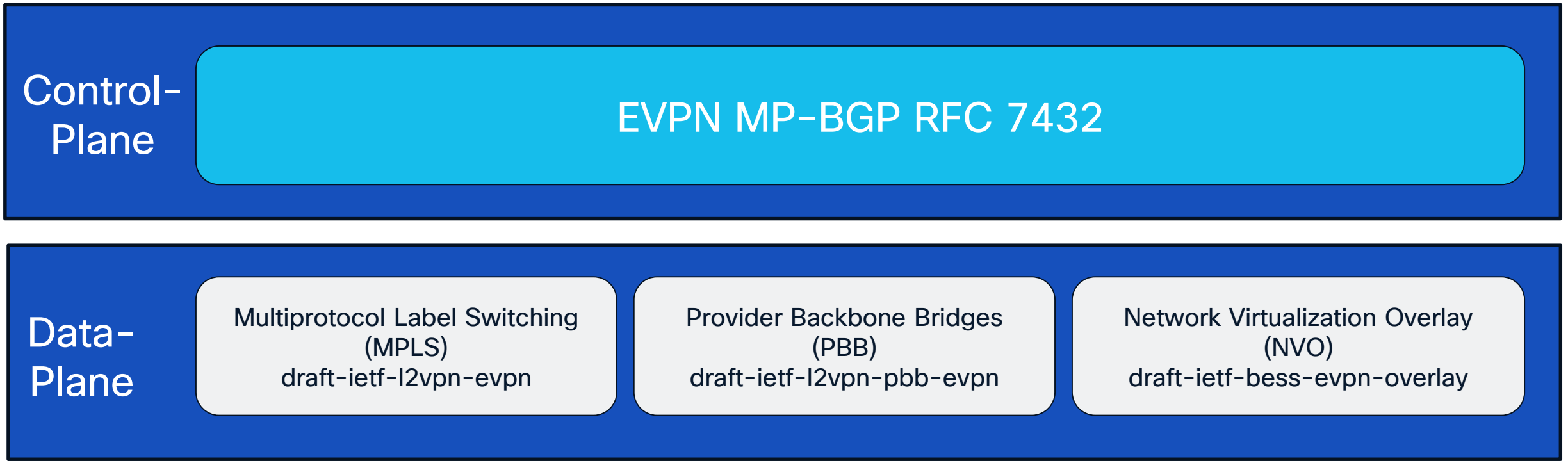
Key Sustainability Metrics

Emissions, Carbon Footprint by Switch or PDU

Fundamentals of VXLAN EVPN Design

EVPN in the Data Center

IETF/RFC Draft for Control and Data Plane



- EVPN over NVO Tunnels (i.e., VXLAN) for Data Center Fabric Encapsulations
- Provides Layer-2 and Layer-3 Overlays over IP Networks

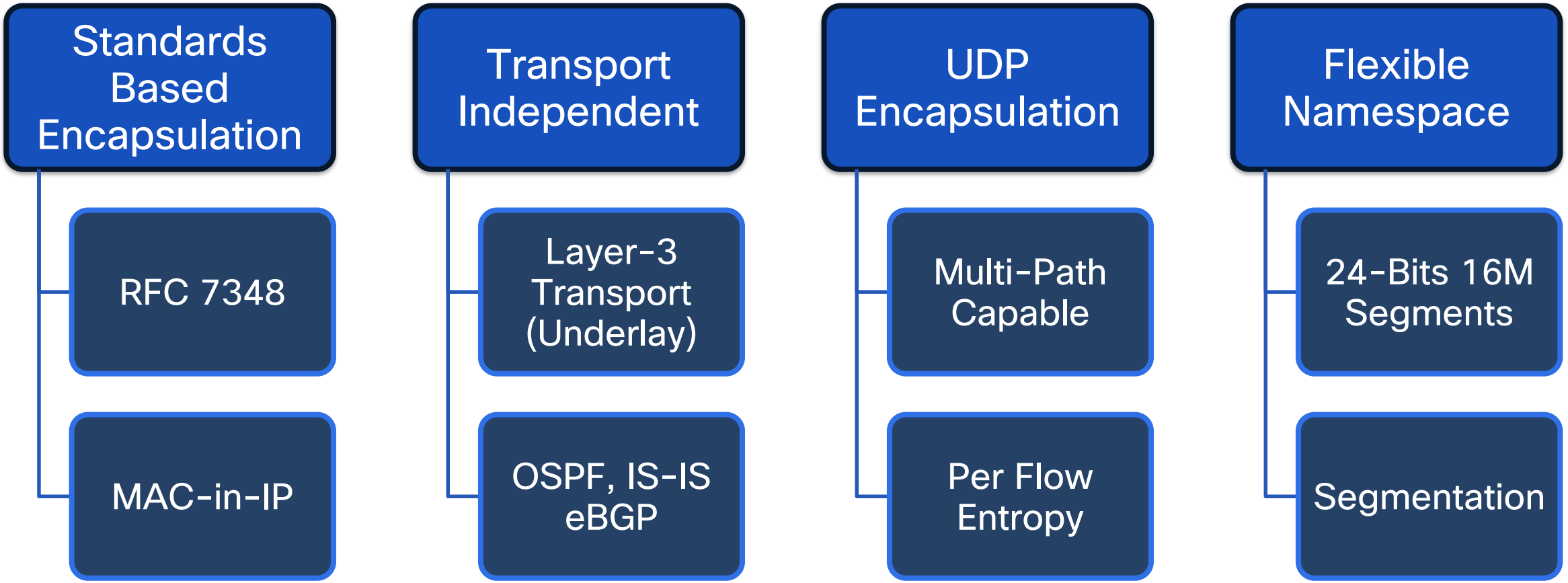
EVPN in the Data Center

For your reference

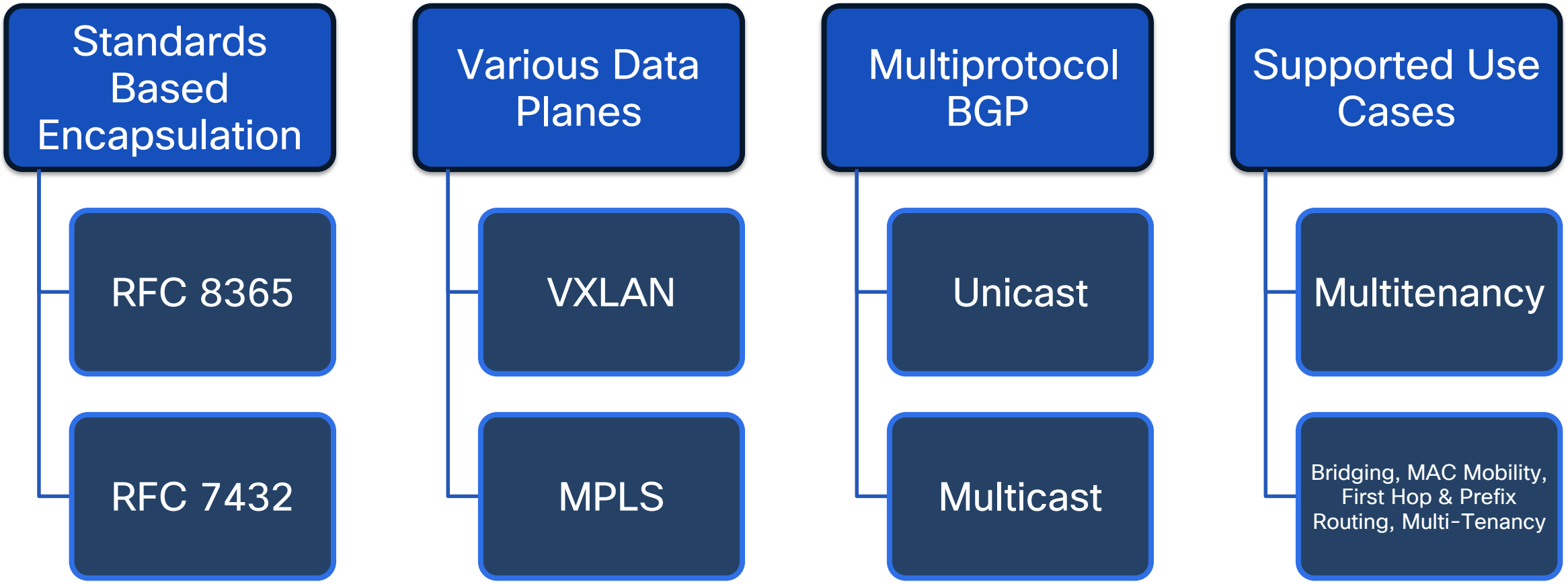
IETF/RFC Draft for Control and Data Plane

RFC / Draft	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	Network Virtualization Overlay Solution using EVPN	Control Plane
draft-ietf-bess-evpn-inter-subnet-forwarding	Integrated Routing and Bridging in EVPN	Control Plane
draft-ietf-bess-evpn-prefix-advertisement	IP Prefix Advertisement in EVPN	Control Plane
draft-tissa-nvo3-oam-fm	NVO3 Fault Management/OAM	Management Plane

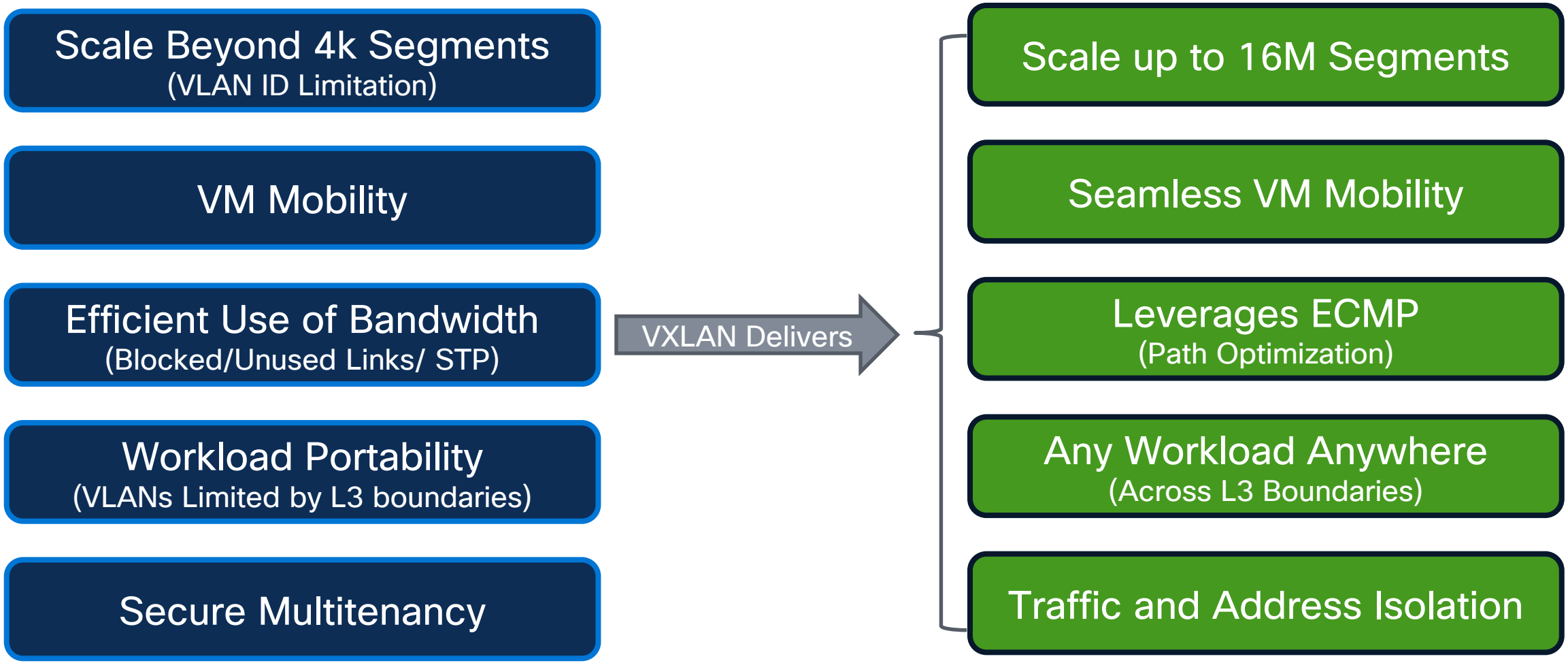
What is VXLAN?



What is EVPN?



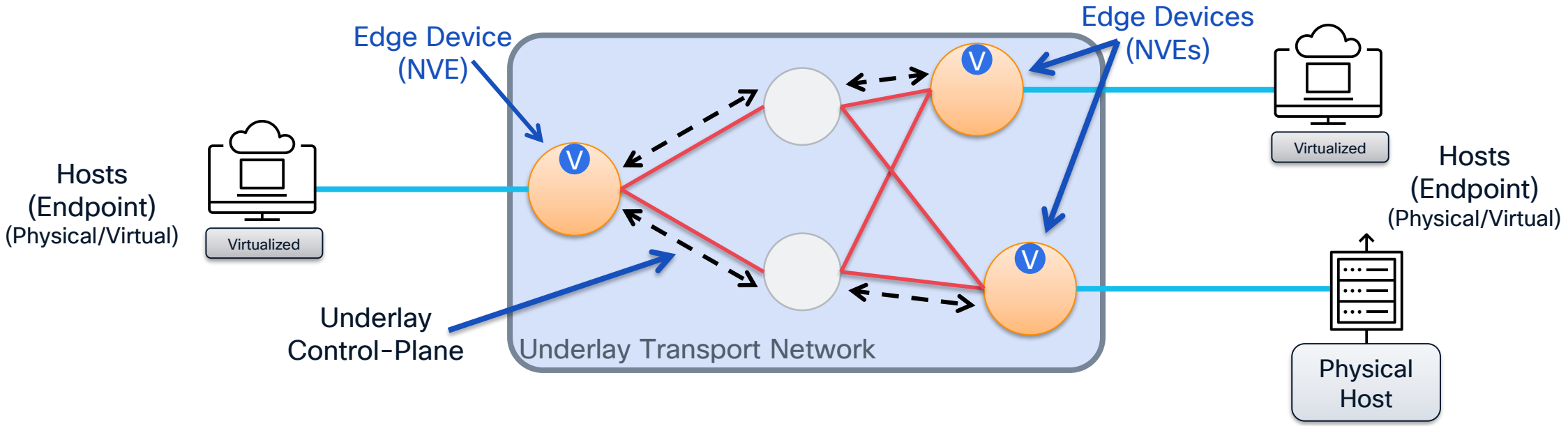
Why VXLAN – How Did We Get Here?



Underlay Taxonomy

- Edge Devices host the VTEP
- Responsible for the encapsulation and decapsulation of the VXLAN Header

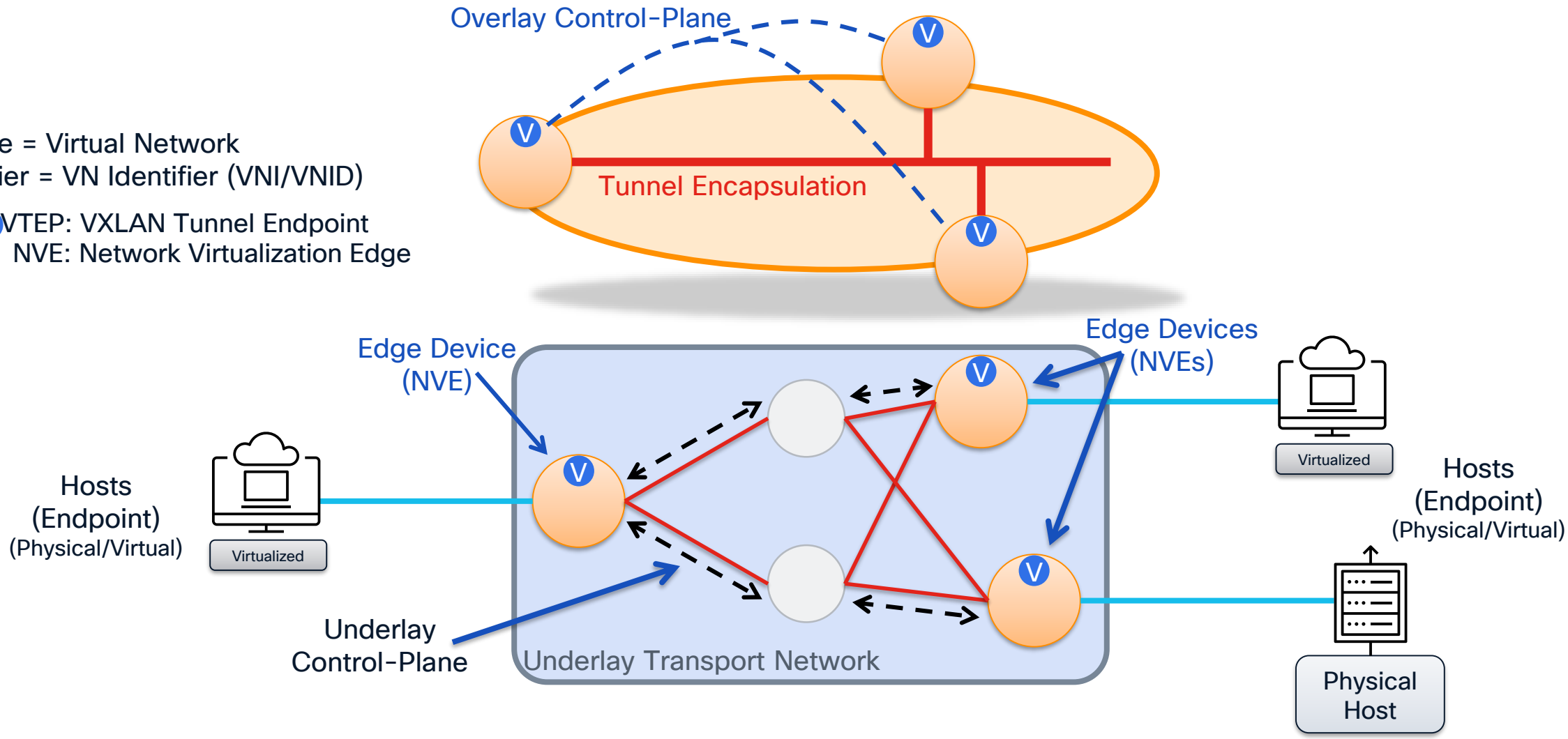
V VTEP: VXLAN Tunnel Endpoint
NVE: Network Virtualization Edge



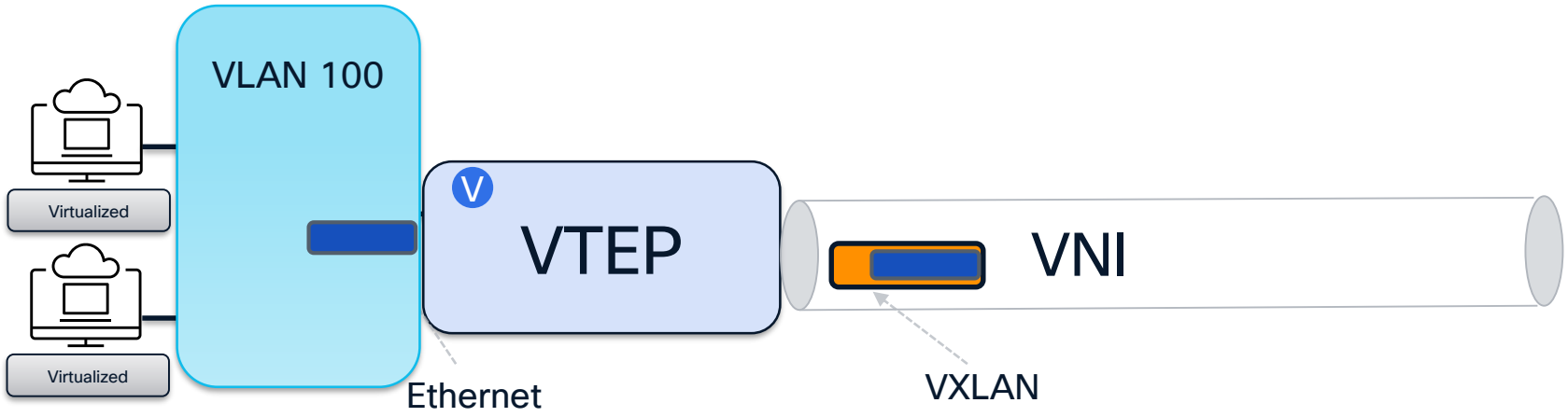
Overlay Taxonomy

Service = Virtual Network
Identifier = VN Identifier (VNI/VNID)

VTEP: VXLAN Tunnel Endpoint
NVE: Network Virtualization Edge

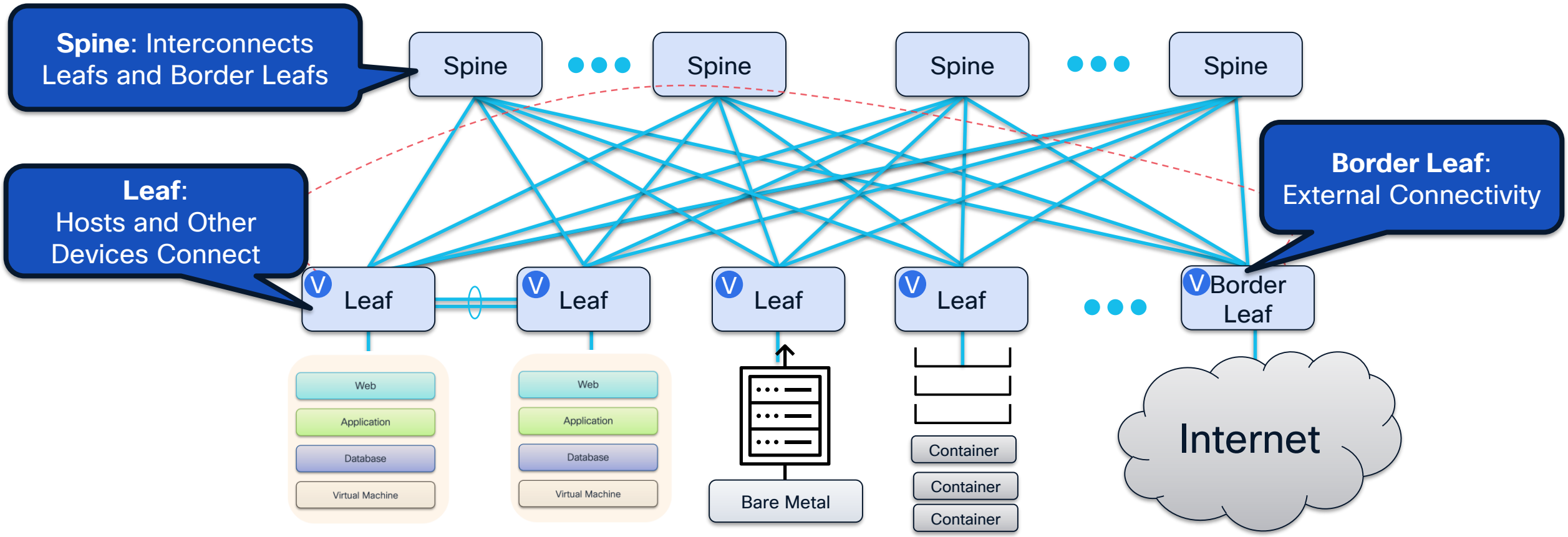


VXLAN (or Virtual) Tunnel Endpoint – VTEP



- VXLAN Tunnel Endpoint – Network Virtualization Edge
- Each VTEP is uniquely identified by an IP Address
- VTEP Discovers or learns remote VTEPs, and end hosts attached to them
- VTEP bridges when forwarding packets within the same VNI and Routes for Inter-VNI traffic

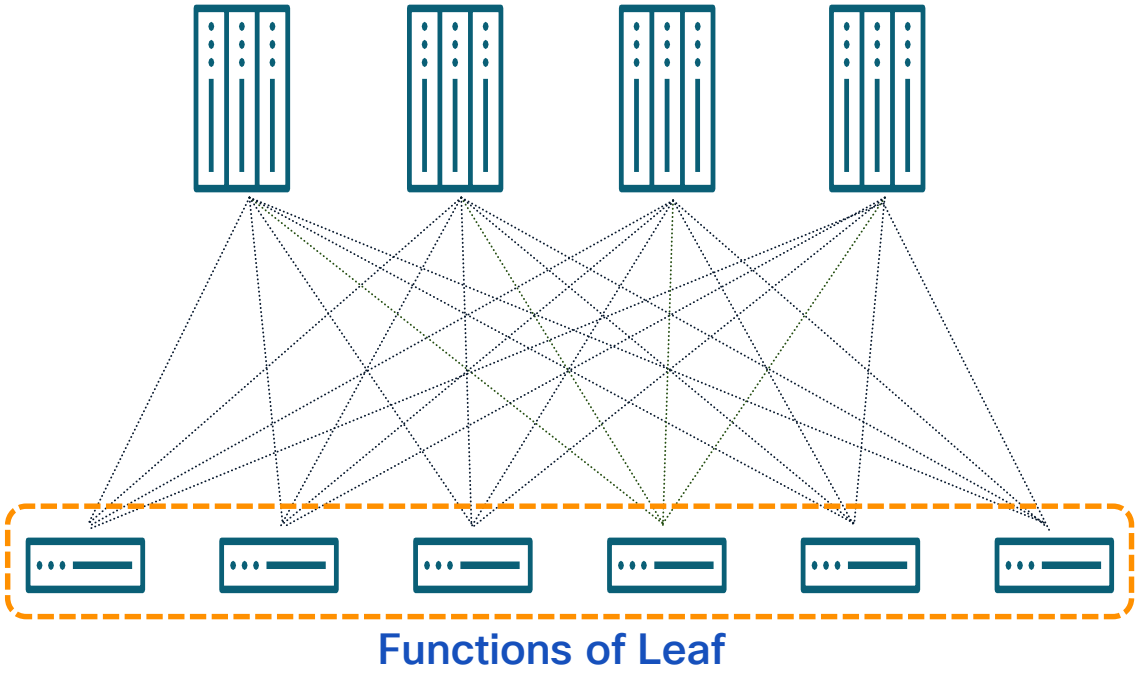
VTEP Device Roles



Components of VXLAN EVPN

Functions of Leaf

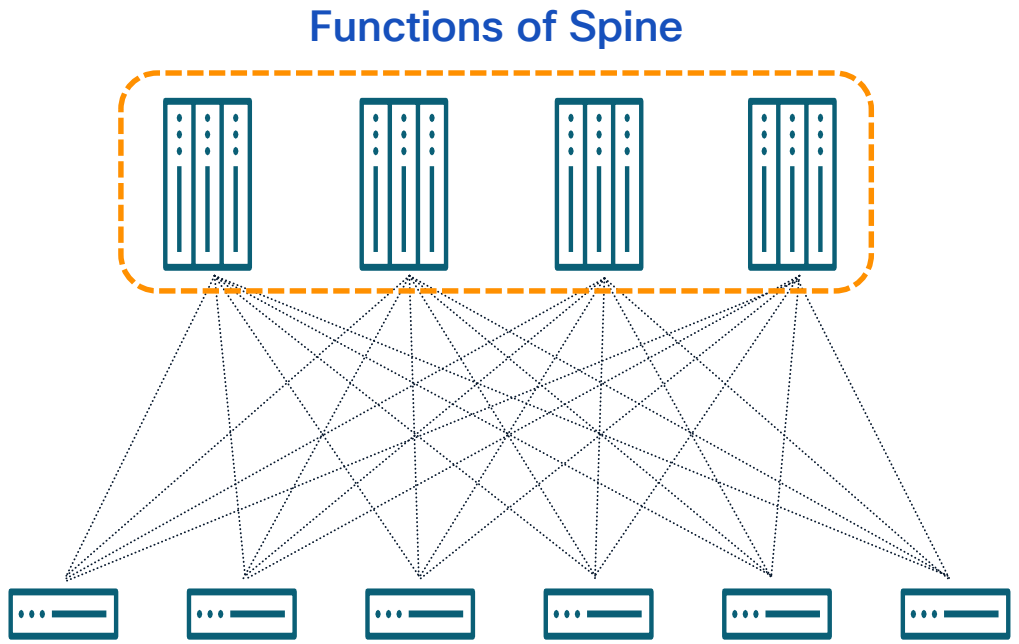
- Forms Routing Protocol adjacencies for underlay with Spines (OSPF, IS-IS, BGP)
- MP-BGP L2VPN EVPN neighborships with spines to exchange routes
- Performs VXLAN encapsulation and decapsulation
- Default Gateway Services for hosts using Distributed Anycast Gateway
- BUM replication/processing
- **Connect to Non-VXLAN segments using VRF-Lite extension (Typically done on Border Leaf)**



Components of VXLAN EVPN

Functions of Spine

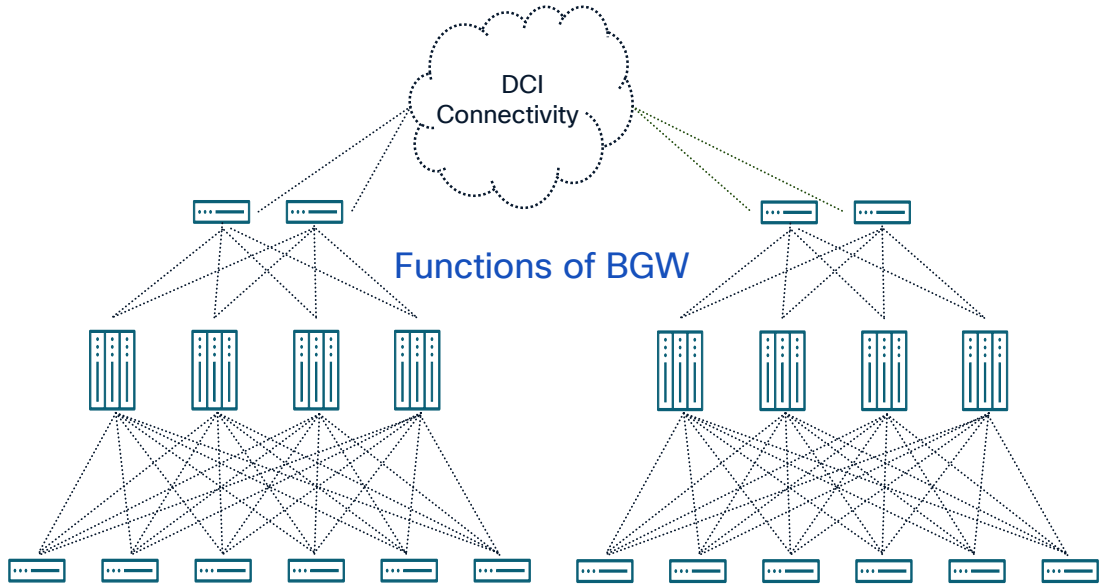
- Forms Routing Protocol adjacencies for underlay with Leaf (OSPF, IS-IS, BGP)
- MP-BGP L2VPN EVPN neighborships with Leaf switches to exchange routes
- Route Reflector for iBGP deployments
- PIM Anycast RP
- **Do NOT typically do VXLAN encapsulation and decapsulation (unless it is a border spine or border gateway spine)**



Components of VXLAN EVPN

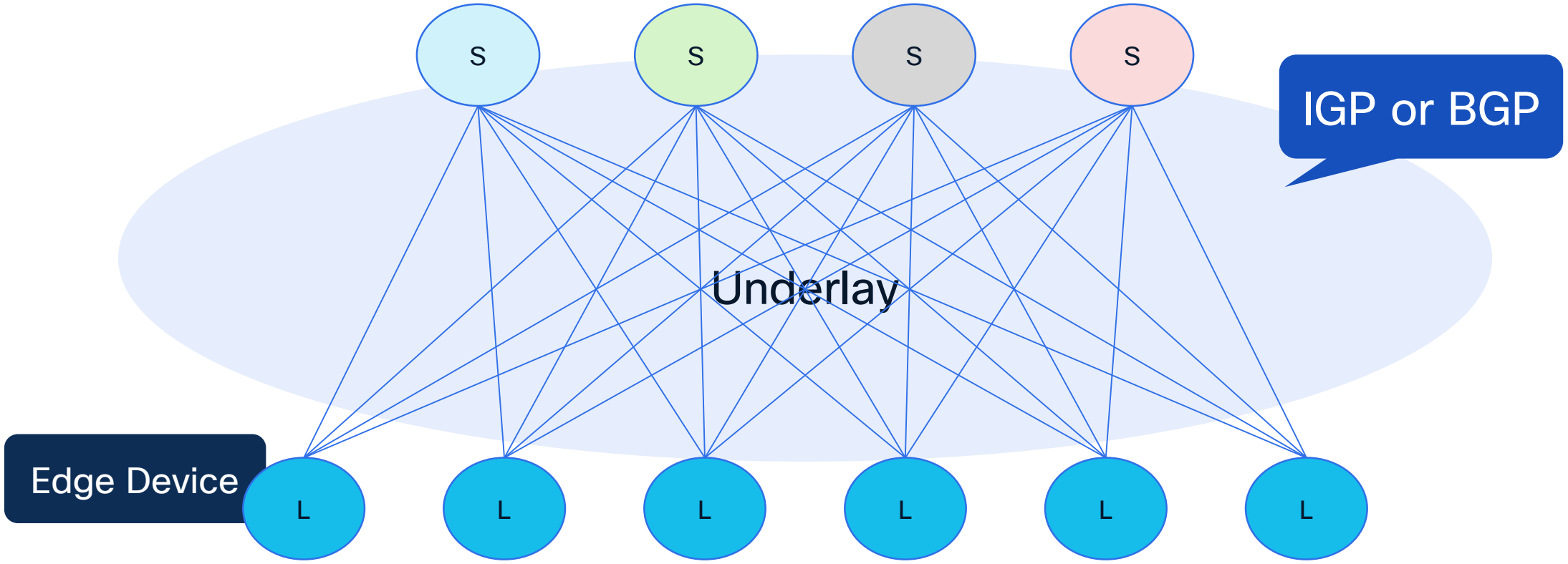
Functions of Border Gateway (BGW)

- Gateway (GW) to stitch multiple VXLAN BGP EVPN domains
- Provides Control- and Data-Plane separation
- Extends Layer-2 and Layer-3 with Control
- Allows to Scale beyond any Fabric Scale
- Facilitates Multi-DC and Multi-Pod Use-Cases
- More than just a Data Center Interconnect (DCI)



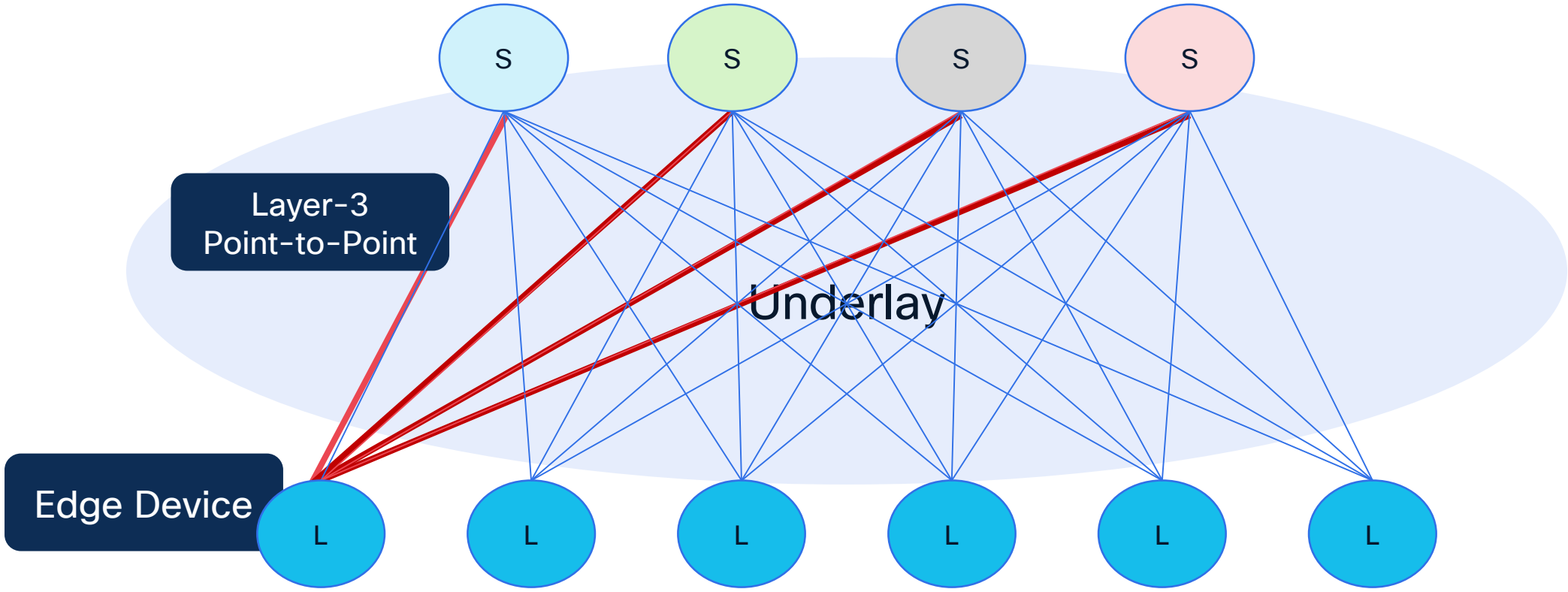
VXLAN: The Building Blocks

Underlay



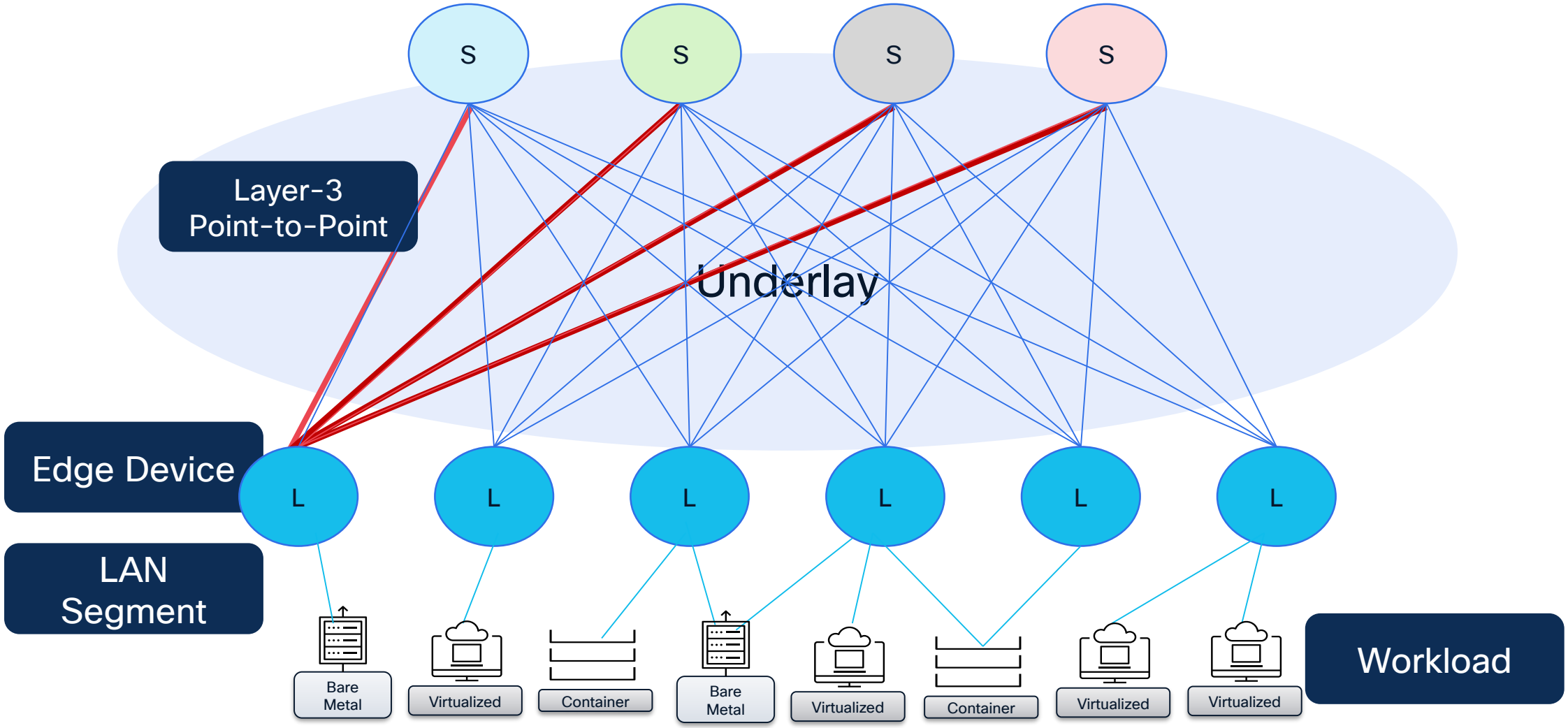
VXLAN: The Building Blocks

Underlay



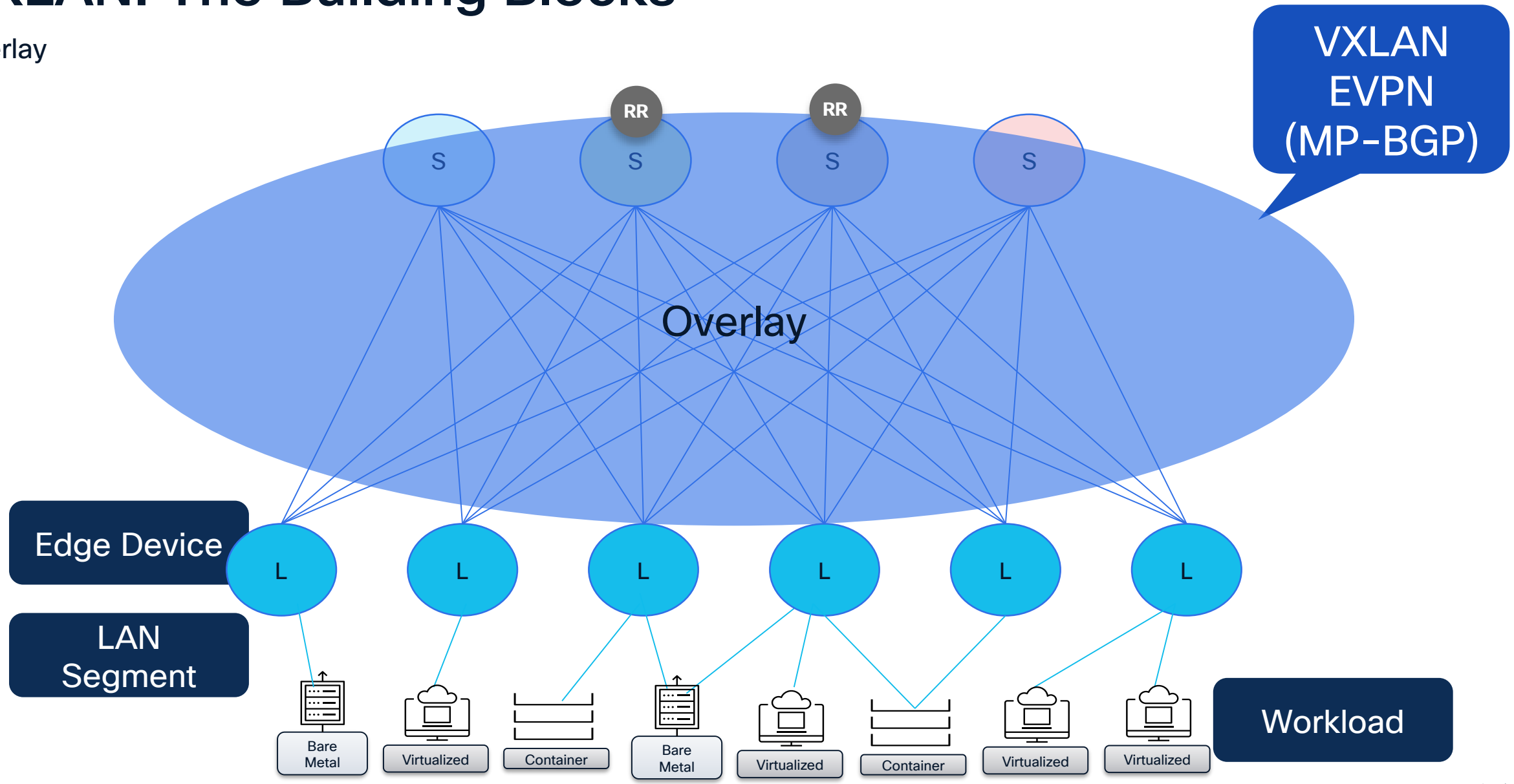
VXLAN: The Building Blocks

Underlay



VXLAN: The Building Blocks

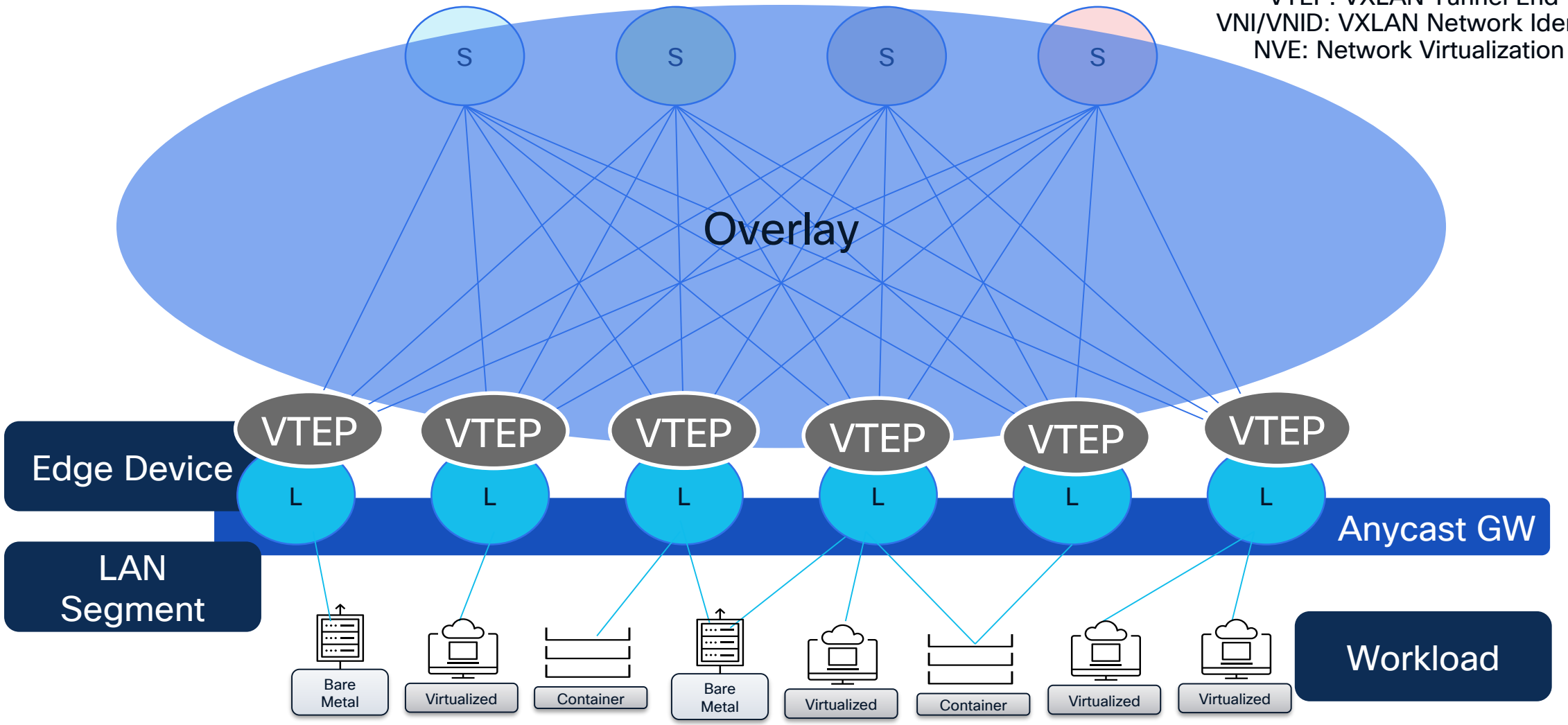
Overlay



VXLAN: The Building Blocks

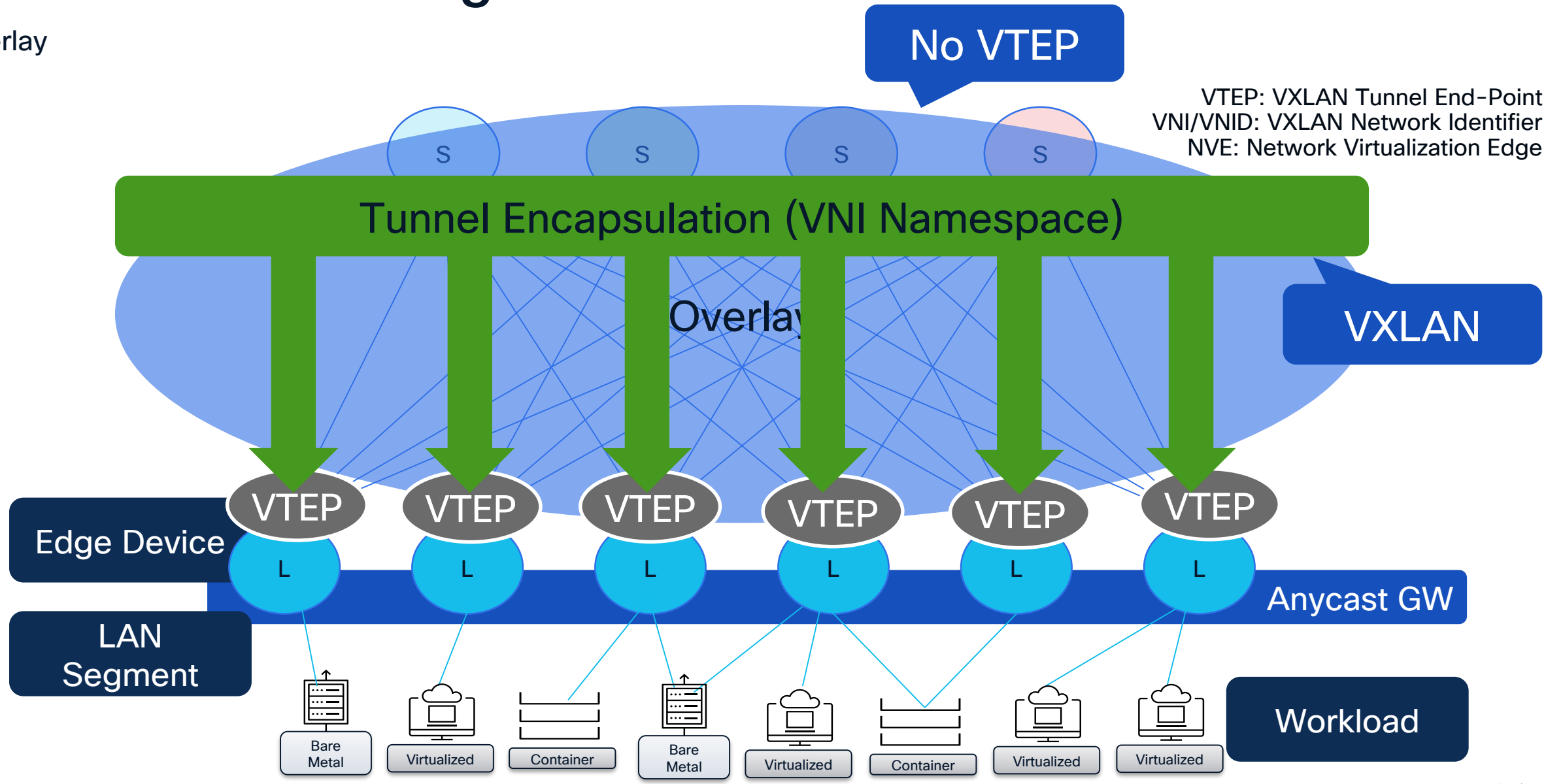
Overlay

VTEP: VXLAN Tunnel End-Point
VNI/VNID: VXLAN Network Identifier
NVE: Network Virtualization Edge



VXLAN: The Building Blocks

Overlay



BGP EVPN *Route Types – RFC 7432

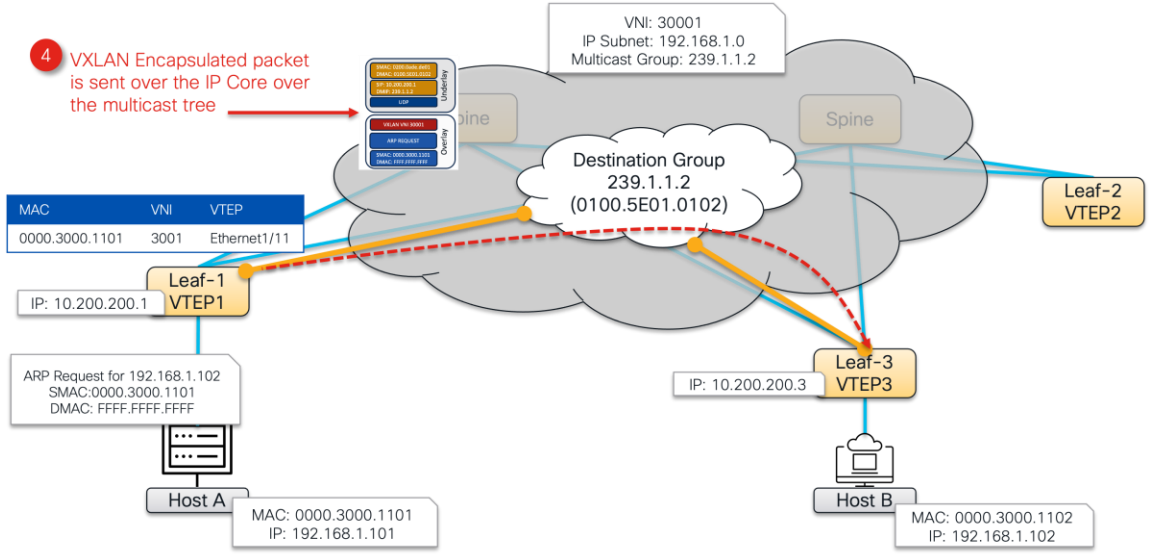
For your reference

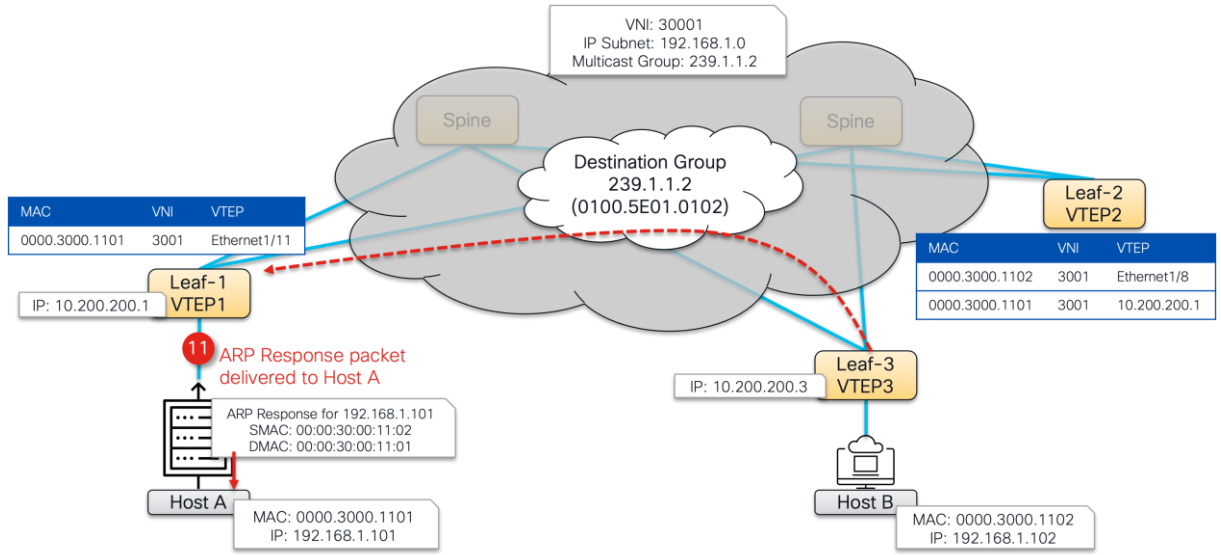
RFC/Draft	Route Type	Description	
RFC 7432	1	Ethernet Auto-Discovery EAD Route	BGP Based Multi-Homing Mass Withdrawal/Aliasing
	2	MAC/IP Advertisement Route	L2 VNI MAC or MAC-IP from L2 MAC Learning or ARP
	3	Inclusive Multicast Ethernet Tag Route	Dynamic Peer Discovery for EVPN Ingress Replication
	4	Ethernet Segment Route	BGP Based Multi-Homing BUM DF Election/Split-Horizon
draft-ietf-bess-evpn-prefix-advertisement	5	IP Prefix Route	IETF Draft, Advertise IP Prefixes

VXLAN Flood & Learn Mechanism

Multicast Based – PIM

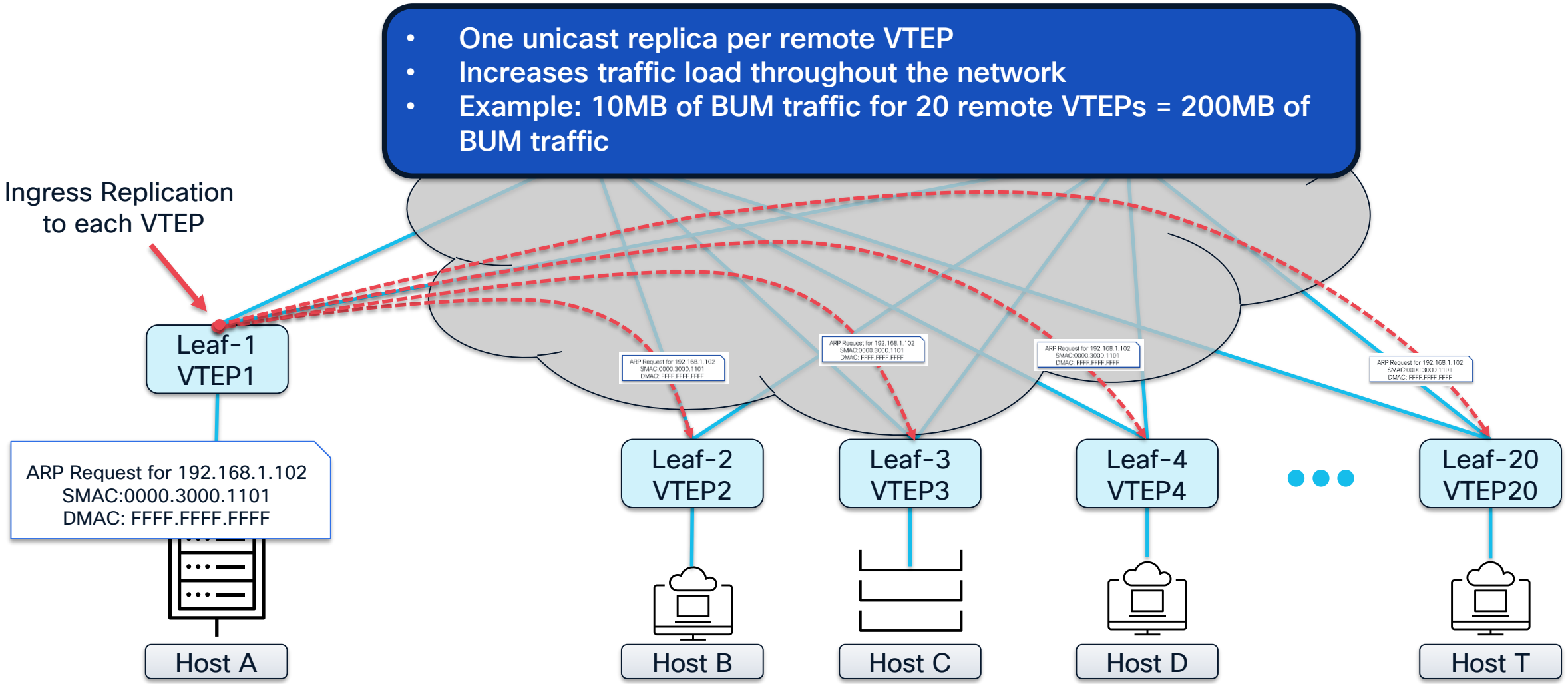
Multidestination Traffic is "Flooded" over the VXLAN Tunnel between VTEPs to "Learn" about the Host MAC addresses located behind the VTEPs so subsequent communication is delivered via Unicast





VXLAN Flood & Learn Mechanism

Unicast Based - Ingress Replication or Head-End Replication



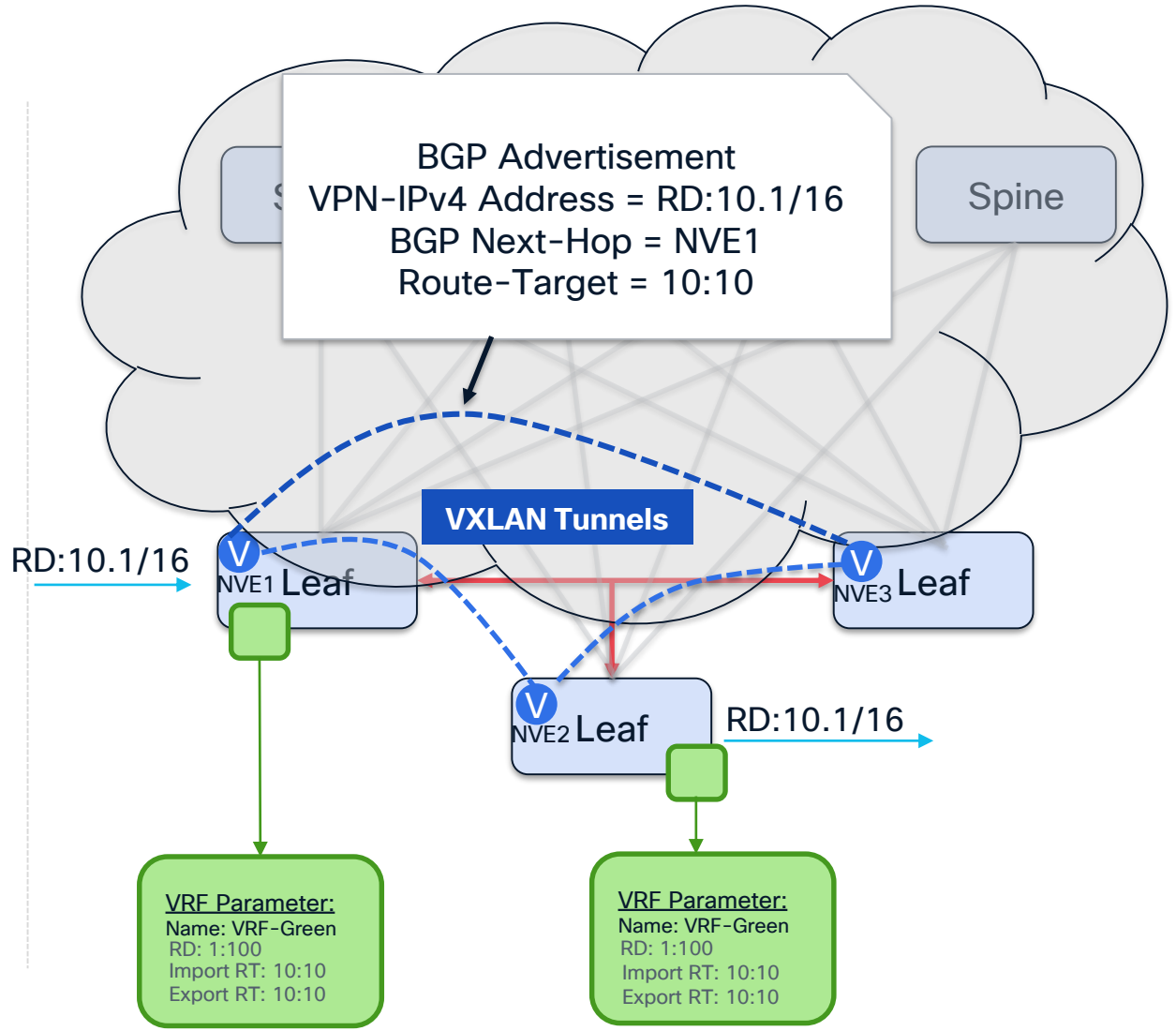
MP-BGP VPN Terminology

VPN Address Family
A Multi-Protocol BGP Extension to Distribute VPN Routes

Virtual Routing & Forwarding (VRF)
Overlay MAC or IP Routes are stored in dedicated VRF tables (MAC or IP VRF)

Route Distinguisher (RD)
8-byte Field, VRF Parameter
Value to make the MAC or IP Routes unique (RD + VPN prefix)

Route Target (RT)
8-byte Field, VRF Parameter
Value for the Import/Export Rules of VPN Routes (MAC or IP) into the VRF



EVPN in the Data Center

MP-BGP for EVPN

- MP-BGP is the routing protocol for EVPN
- Multi-tenancy construct using VRF (Virtual Route & Forwarding)
  - Route Distinguisher, and Route Targets
- New Address Family “l2vpn evpn” for distributing EVPN Routes
- EVPN Routes = [MAC] + [IP]
- iBGP or eBGP Support

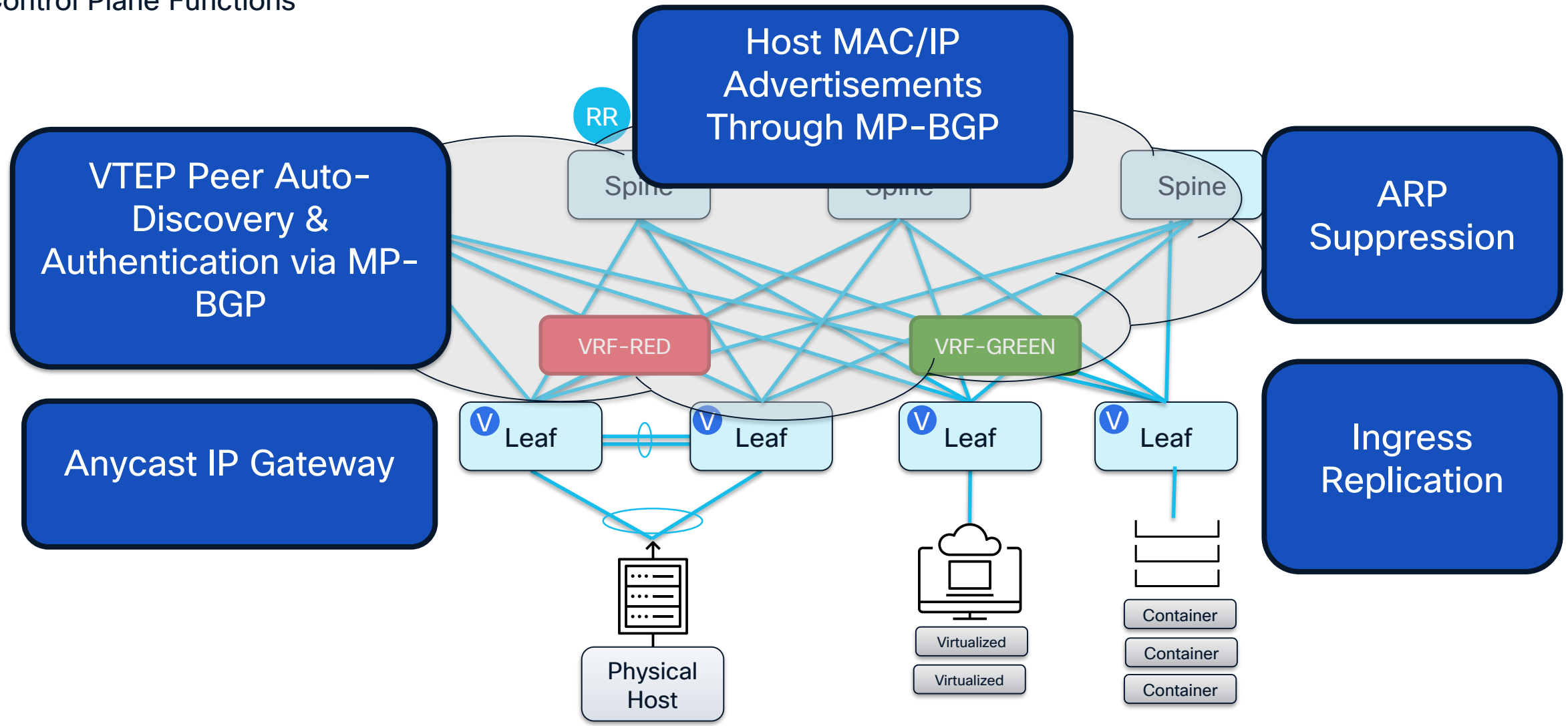
```
vrf context myvrf_50000
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
```

```
evpn
  vni 30000 l2
    rd auto
    route-target import auto
    route-target export auto
```

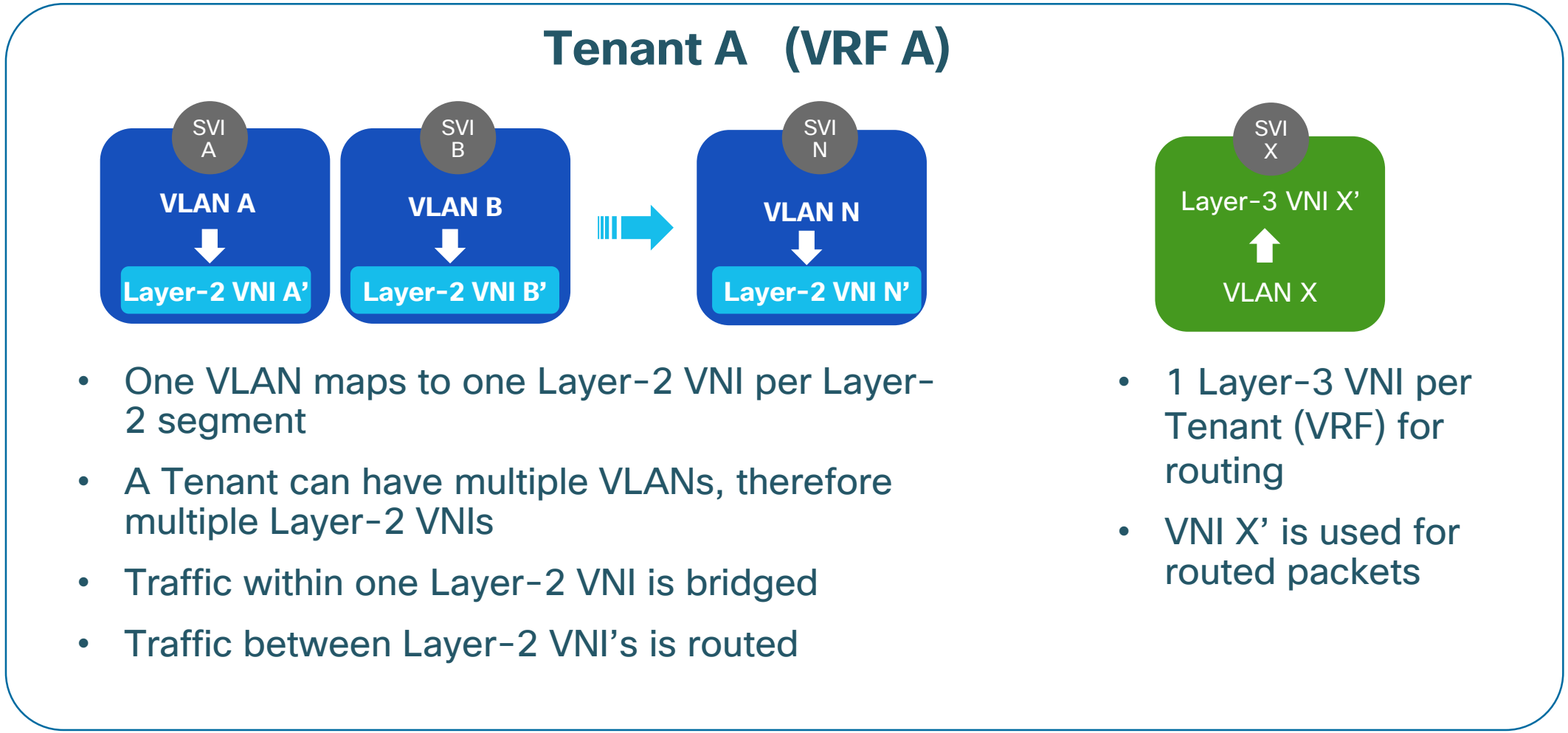
```
router bgp 65101
  router-id 10.2.0.1
  neighbor 10.2.0.3
    remote-as 65101
    update-source loopback0
    address-family l2vpn evpn
 vrf myvrf_50000
    address-family ipv4 unicast
      advertise l2vpn evpn
```

EVPN in the Data Center

Control Plane Functions



Logical Construct of Multi-Tenant VXLAN EVPN



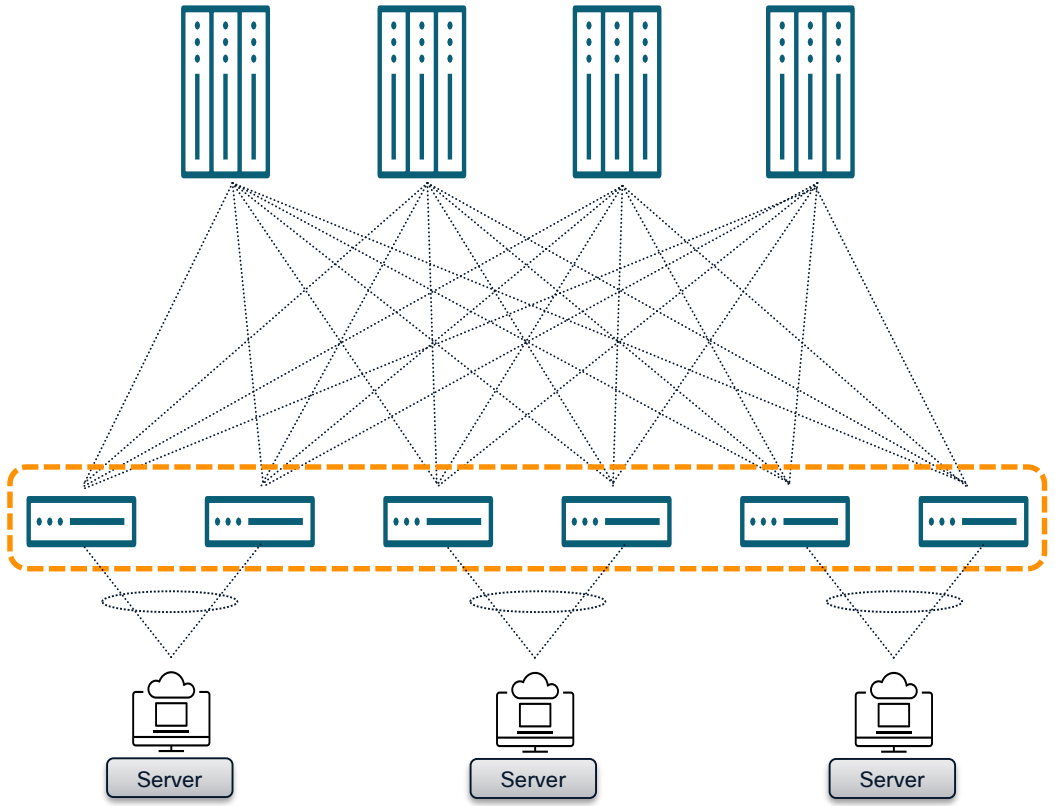
- One VLAN maps to one Layer-2 VNI per Layer-2 segment
- A Tenant can have multiple VLANs, therefore multiple Layer-2 VNIs
- Traffic within one Layer-2 VNI is bridged
- Traffic between Layer-2 VNI's is routed

- 1 Layer-3 VNI per Tenant (VRF) for routing
- VNI X' is used for routed packets

Host Connectivity in VXLAN Fabric

Virtual Port Channel (VPC)

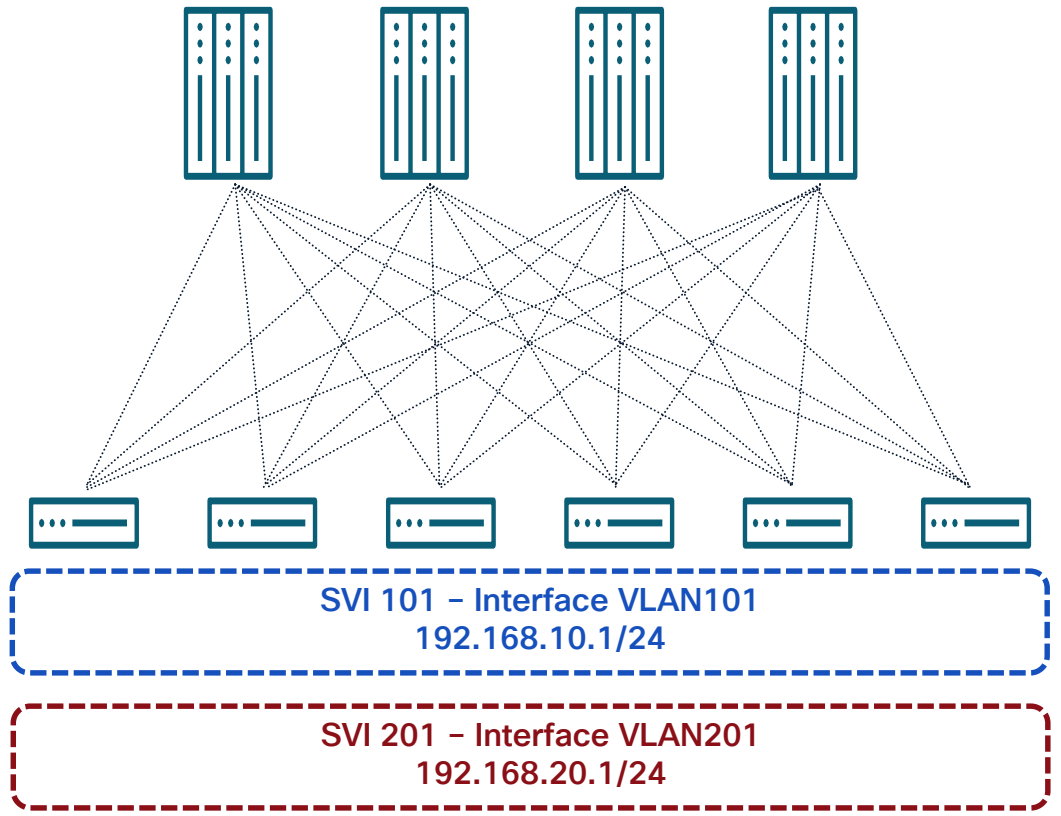
- Virtual Port Channel
  - Multi Chassis Link Aggregation
  - Layer-2 Multihoming
- Host Side Connectivity
  - Dual Connect Hosts Using Port Channels
- Fabric Side Connectivity
  - Individual VTEPs
  - Uses a Common Anycast VTEP
  - Seen as Remote VTEP from Remote Nodes
- VPC is Lone Multihoming Technology Supported by NDFC
- Standards Based ESI Support Coming in NX-OS 10.6(1)
- Physical Peer-Link is Optional, starting from NX-OS 9.2(3) Switches can be Configured with VPC Fabric Peering



First Hop Routing in VXLAN Fabric

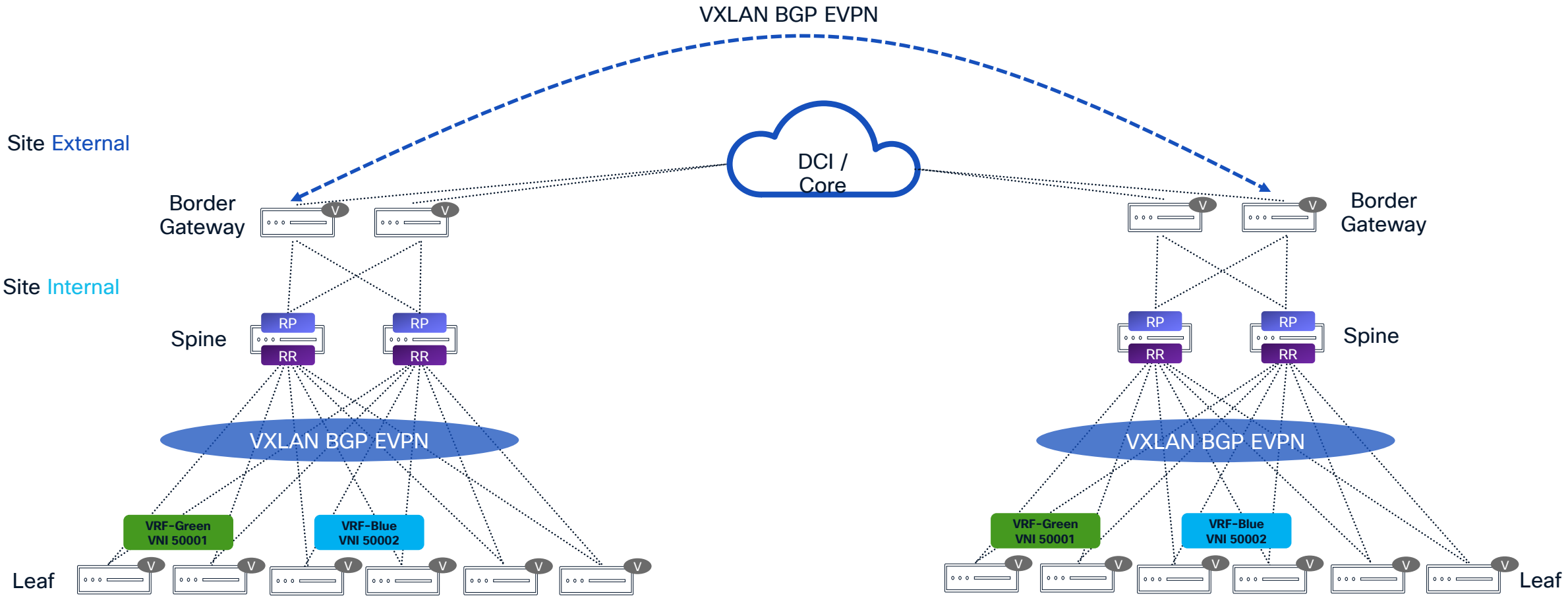
Distributed Anycast Gateway (DAG)

- Distributed First Hop routing on Edge Device
  - All Edge Device Share Same Gateway IP, MAC Address
  - Pervasive Gateway Approach
- Gateway is Always Active
  - No First Hop Redundancy Protocol for Hello or State Exchange
- Associated to a Unique MAC address to Facilitate Host Mobility
- DAG is close to the End-Hosts, hair-pinning in the fabric is almost eliminated
- It scales out with the fabric, new switches will be configured with the DAG automatically, if required
- Distributed and Smaller State
  - Only Local Endpoints ARP Entries



Multi Fabric Designs – VXLAN Multi Site

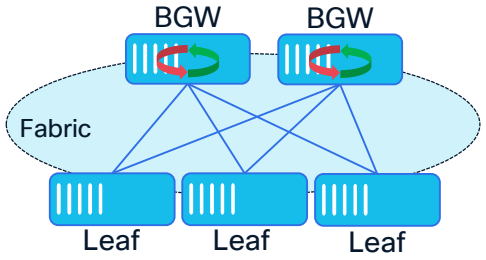
Connecting Multiple VXLAN EVPN Sites



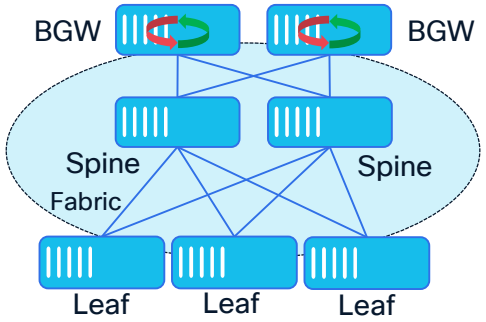
Border Gateways

Deployment Considerations

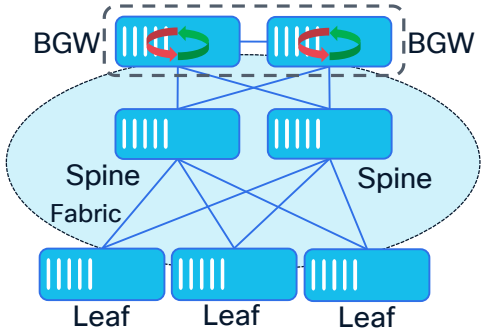
Shared Anycast Border Gateway



Dedicated Anycast Border Gateway

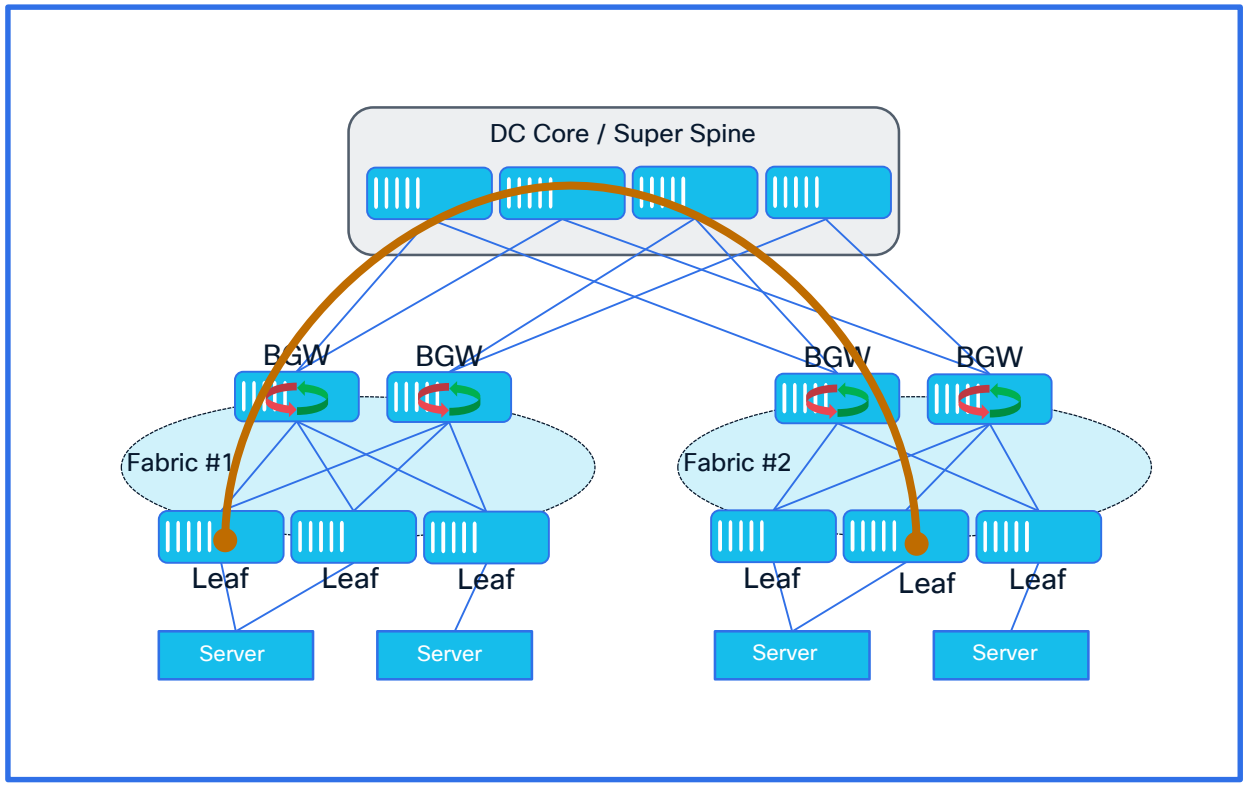


Dedicated vPC Border Gateway



- Border Gateways used for two main functions:
  - Interconnecting each site to the Inter-Site network (for East-West traffic flows)
  - Connecting each site to the external Layer 3 domain (for North-South traffic flows)
- Possible deployment models:
  - Anycast Border Gateways
  - vPC Border Gateways – The only supported method to connect end-hosts at the BGW level
- BGW function enablement in the VXLAN EVPN fabric:
  - BGW is a logical role dictated by features and configurations. Devices can either be placed at the leaf or spine level

Use Case #1 – Compartmentalization



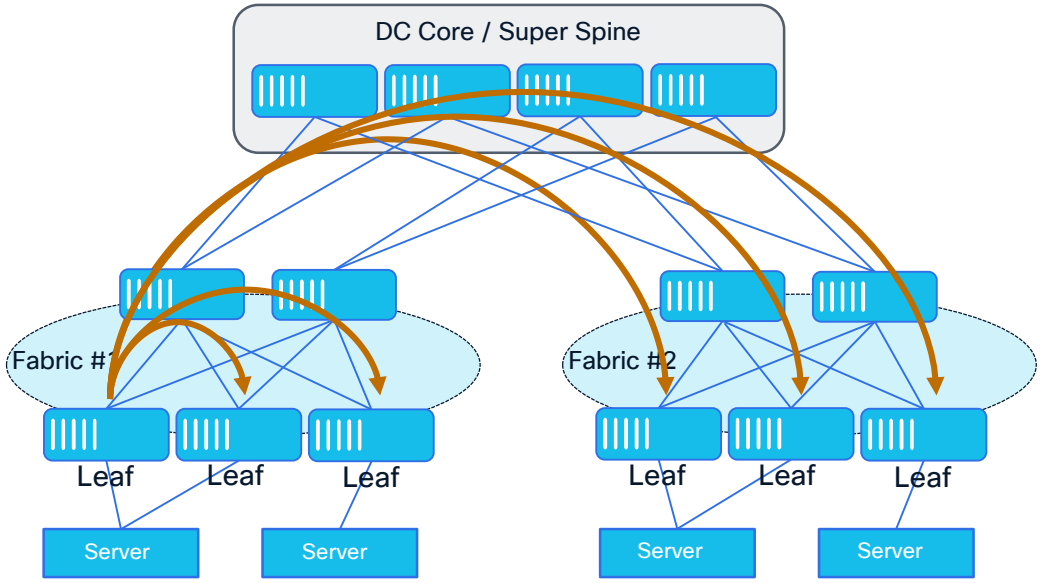
- Multiple Fabrics, single Data Center
  - Single or Multiple Data Halls
  - Within a Geographic Locations
- Control at BGW (Border Gateway)
  - Allows Layer-2 and Layer-3 Extension
  - VNIs are selectively advertised
  - Allows Traffic Control (BUM*)
  - Optimizes BUM* Replication

*BUM – Broadcast, Unknown Unicast, Multicast

BUM Optimization

Use Case #1 Compartmentalization

Single Fabric BUM with Ingress Replication

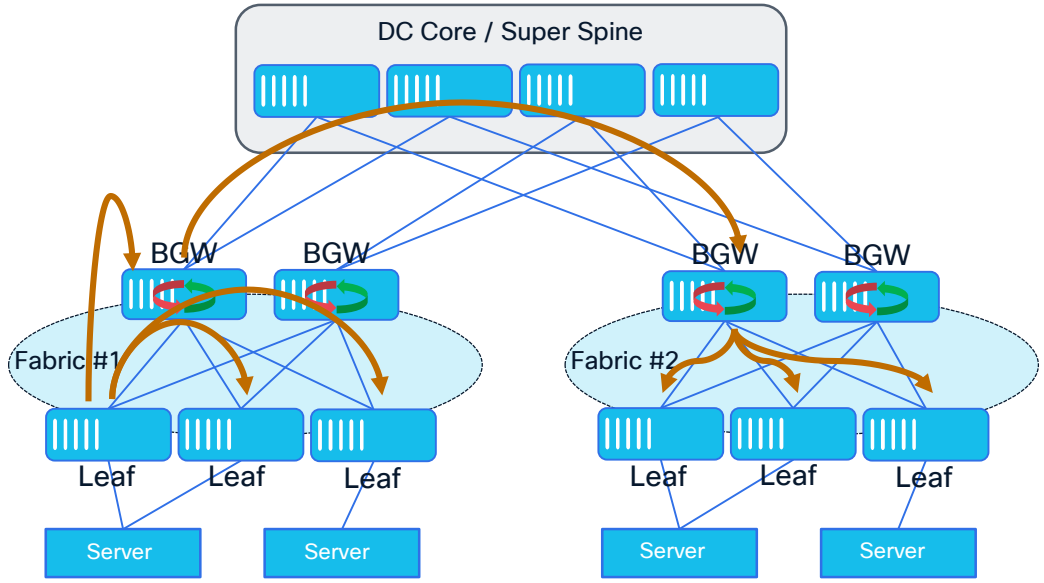


Single BUM Packet, 5x Replicated

3 Replication over DC Core / Super Spine (Between)

2 Replication for Fabric #1 (Local)

Multi-Site BUM with Ingress Replication



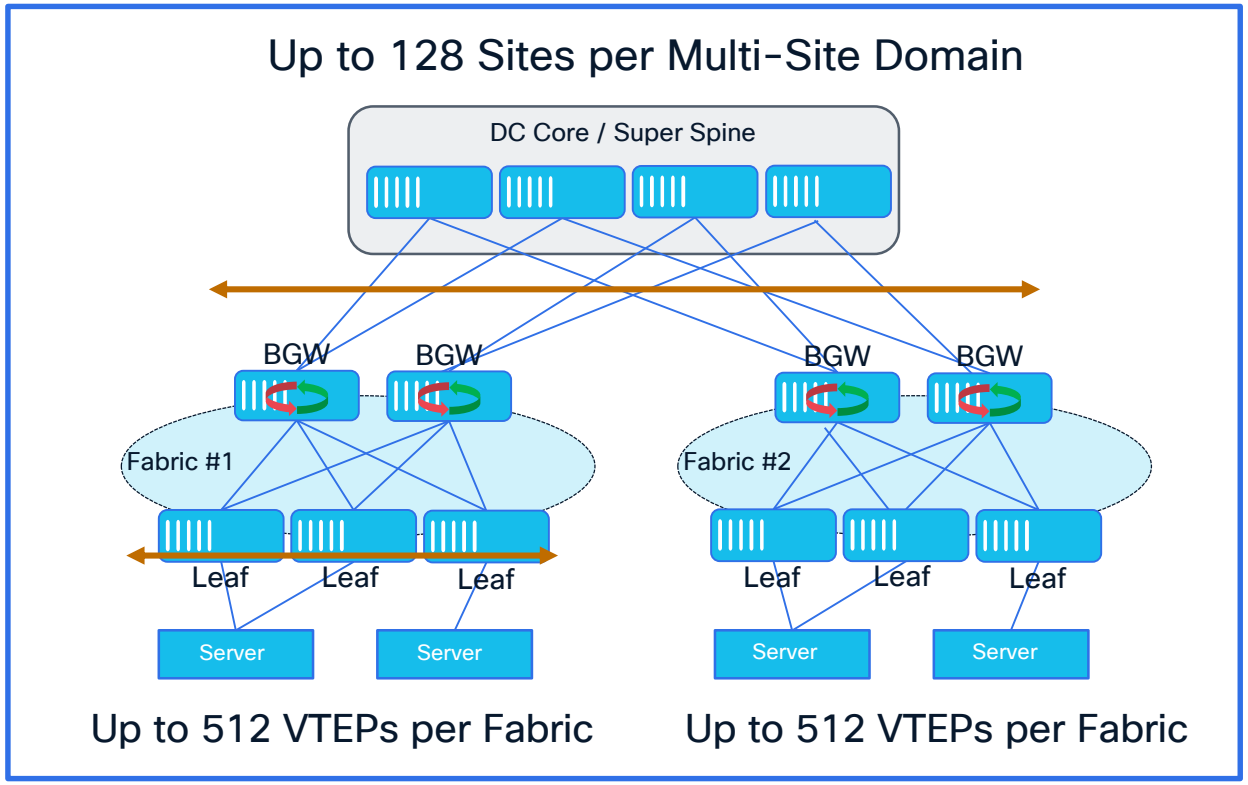
Single BUM Packet, 3x Replicated

1 Replication over DC Core / Super Spine (Between)

3 Replication for Fabric #1 (Local)

3 Replication for Fabric #2 (Local)

Use Case #2 – Building Scalable Networks



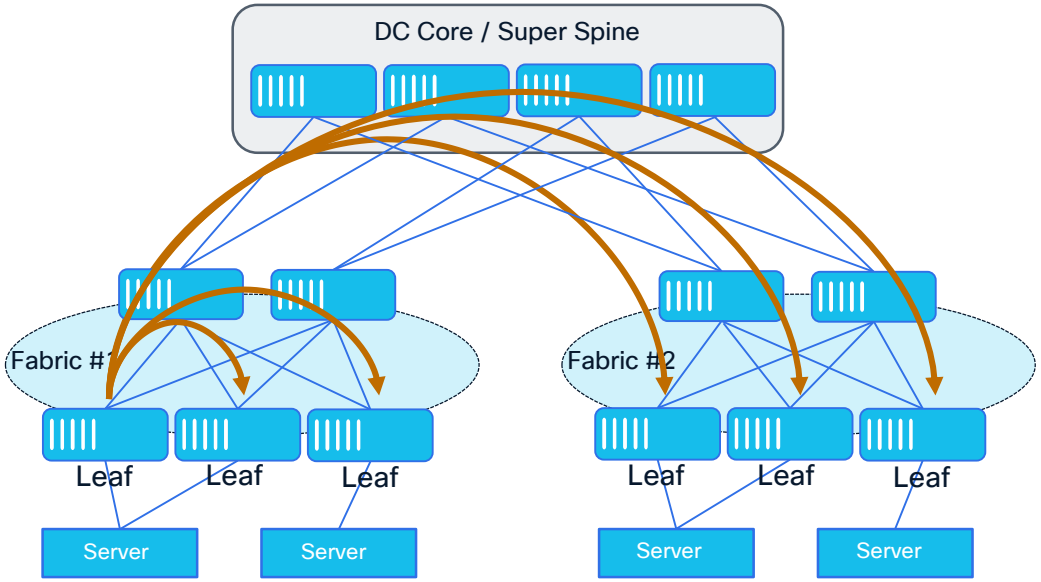
- Multiple Fabrics , single or multiple Data Center
  - Single or Multiple Data Halls
  - Within or between Geographic Locations
- Control at BGW (Border Gateway)
  - Reduces Remote VTEP Count
  - Expands VTEP scale
- Scale through Hierarchy
  - Multiply VTEP with Sites

64000+ VTEP to extend Layer-2 or/and Layer-3 segments

VTEP Scalability

Use Case #2 Building Scalable Networks

Single Fabric (or Multi-Pod)

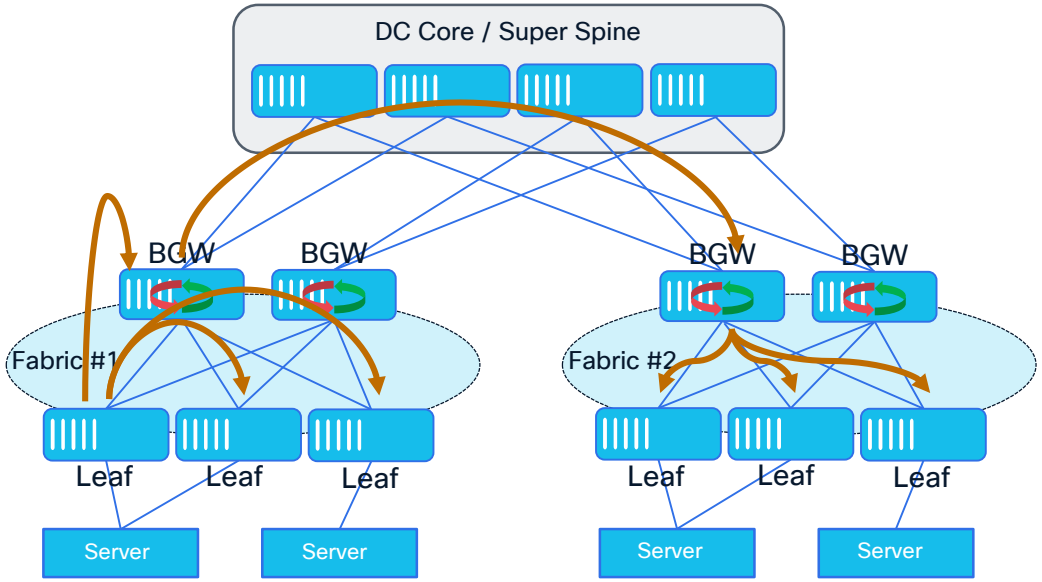


Leaf #1 sees every VTEP, 5 VTEP Peer

3 VTEP Peer for Fabric #2 (Between)

2 VTEP Peer for Fabric #1 (Local)

Multiple Fabric with Multi-Site

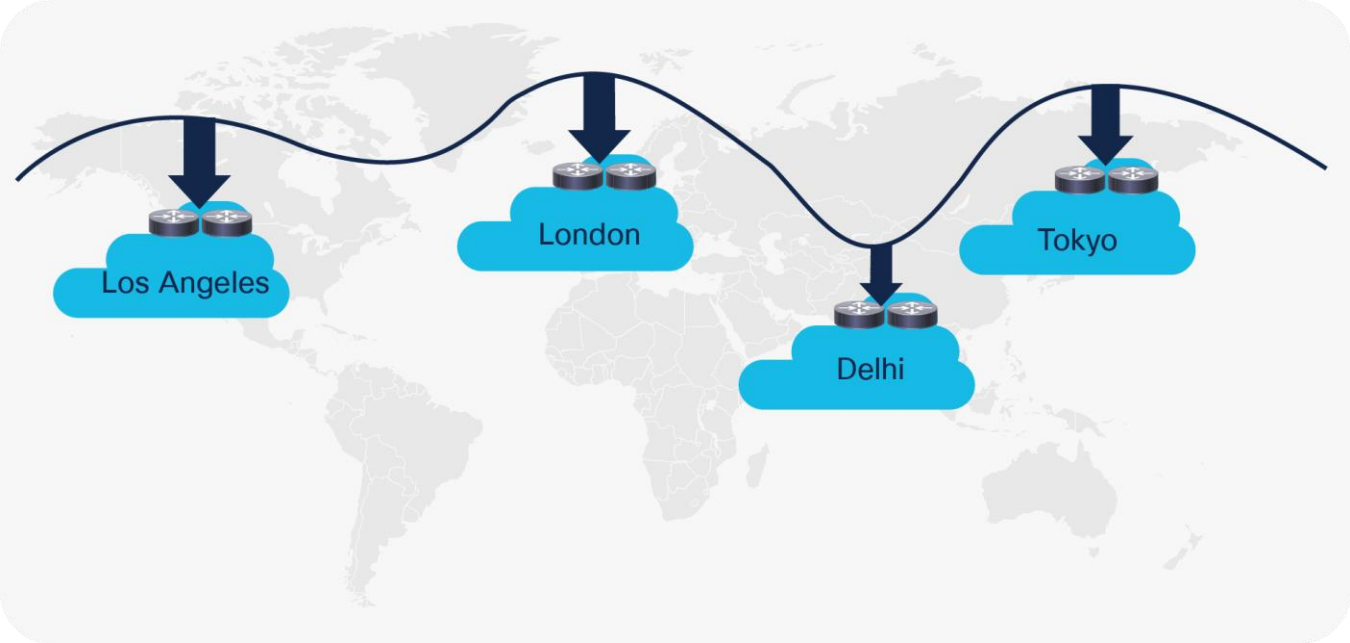


Leaf #1 sees only local VTEP, 3 VTEP Peer

1 VTEP Peer for Exit, BGW (Between)

2 VTEP Peer for Fabric #1 (Local)

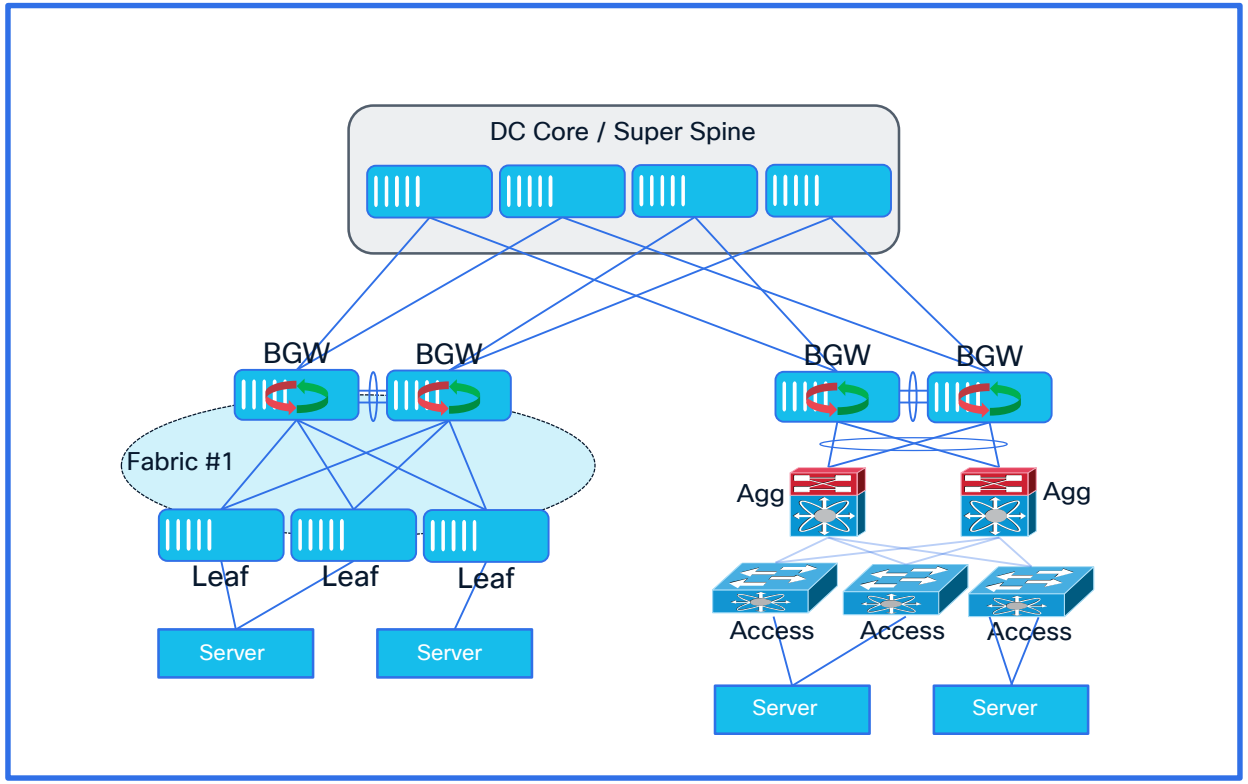
Use Case #3 – Data Center Interconnect (DCI)



- Multiple Fabrics, Geographically Dispersed
- Classic DCI Use Case
  - Allows Extension of Layer-2
  - Allows Extension of Layer-3
  - Allows Traffic Control (BUM)
  - Defines VNI allocation and stitching
  - Optimizes BUM* Replication

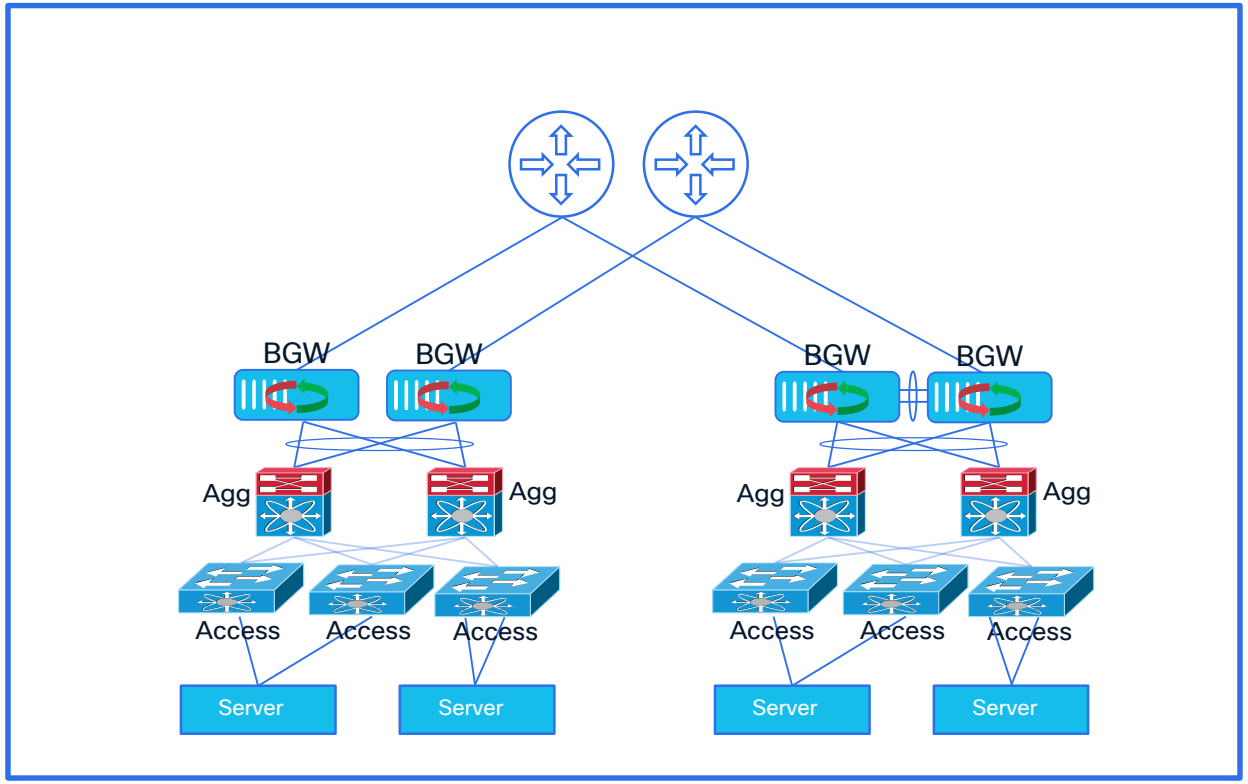
Works Within a Geographic Location – Works Between Geographic Locations

Use Case #4 – Integration with Legacy Networks



- Integrating Fabrics with Legacy Networks
  - BGW Frontends Legacy Network
  - BGW Frontends New Network
- Host Mobility and Migration
  - Provides Distributed Default Gateway
  - Allows Layer-2 Extension where needed
- Benefits from all Multi-Site functions
  - Layer-2, Layer-3 Multicast and Unicast VPNs between different Networks for Migration or Co-Existence
- Distributed Anycast Gateway and/or HSRP coexistence

Use Case #5 – Next-Gen DCI



- vPC Border Gateways used to provide a modern DCI architecture to two or more traditional fabrics
- Endpoint Default Gateways can be deployed on the vPC BGWs to provide the anycast functionality
- Single technology for a secure L2 and L3 extension

Multi Site Architecture Summary

#1

Border Gateway (BGW)

A Gateway (GW) to stitch multiple VXLAN BGP EVPN domains
Provides Control- and Data-Plane separation
Extends Layer-2 and Layer-3 with Control
Allows to Scale beyond any Fabric Scale
Facilitates Multi-DC and Multi-Pod Use-Cases
More than just a Data Center Interconnect (DCI)

#2

VXLAN BGP EVPN Multi-Site

A Simple add or drop-in
First introduced in September 2017 – proven and deployed
A Solution combining EVPN DCI Overlay (RFC9014) and IPVPN-EVPN interworking (draft-ietf-evpn-ipvpn)
Provides Layer-2 and Layer-3 extension
Wide Hardware Support
Flexible Deployment Option – Not just for VXLAN Fabrics

Nexus VXLAN Fabric Security

Security for Open and Programmable Fabric

Secured Traffic

- VXLAN IP First-Hop-Security FHS 10.4(1)F
- Multi-Site Storm Control
- Private VLANs 10.2(3)F
- Hardware Line-rate encryption (MACSec)
- Multi-hop VTEP encryption (CloudSec)
- Group Policy Option 10.4(3)F

Visibility and Proactive Threat Response

- Netflow w/ Stealthwatch
- Nexus Dashboard- Proactive Advisories
- GNMI Streaming Telemetry

Secured Services

- RADSec (RFC 6614)
- Secure NTP
- RADIUS over DTLS (10.4.3)

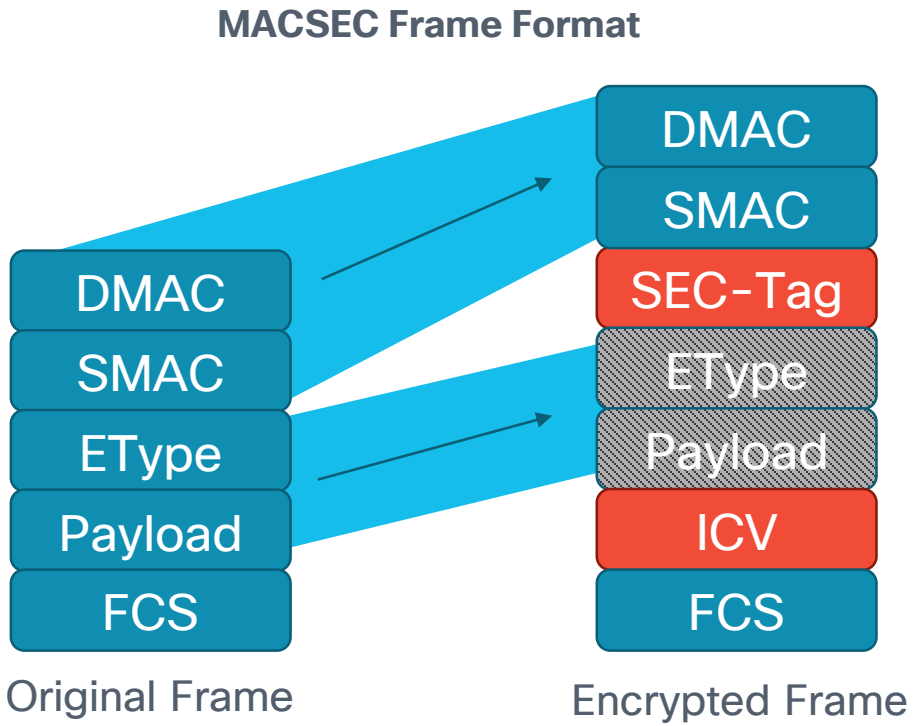
Security Compliance

- FIPS
- Common Criteria
- NDcPP
- UC-APL

VXLAN Fabric Security

Hardware Line-Rate Encryption

- Industry Standard security technology
- Provides confidentiality and integrity for Ethernet/IP traffic
- Multiple Options available
  - MACSec secures traffic on ethernet links on a hop-by-hop basis. Both sides of the ethernet link must be MACSec capable
    - Extensible Authentication Protocol (EAP)
    - Postquantum Cryptography Ready. QKD protocols available
  - CloudSec secures the traffic on a multi-hop path between two VTEPs (BGWs In our case).
- They work at line rate, solves (relatively) low IPSEC throughput
- Look at products datasheet for details on the support



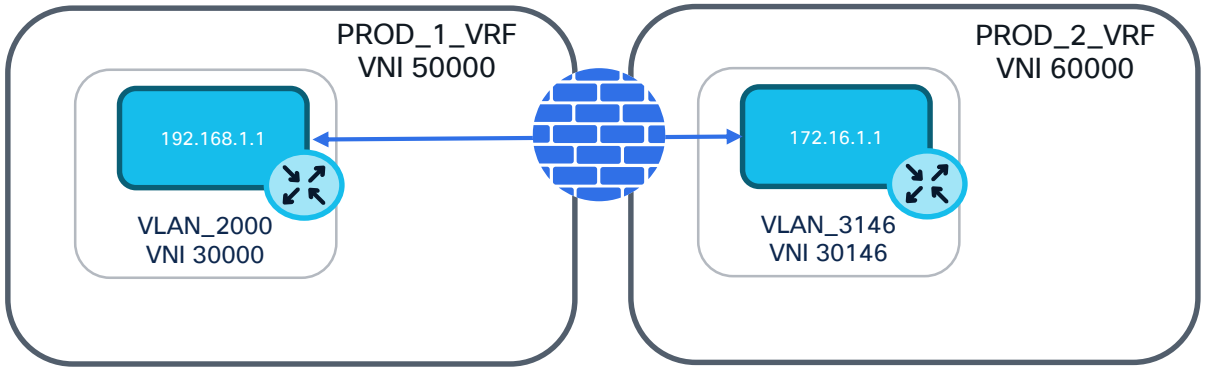
Segmentation

- Segmentation enables the isolation of network-attached resources through a combination of device capabilities plus control and data plane functions
- Segmentation often works in parallel with traffic enforcement mechanisms, which enable selective permit/deny actions across different segmentation zones
- Data Center security has a lot of traction nowadays
  - Attacks number are increasing and harder to detect and mitigate
- Administrators need to design secure DC networks considering several aspects?
  - What segmentation level is required in my fabric?
  - Should my VTEPs control inter-VRF, intra-subnet or inter-subnet traffic?
  - Are VTEPs stateless capabilities sufficient? Or should the traffic be offloaded to some external security devices?
  - How can I insert an external security device in the traffic path?

Segmentation

Classic Service Insertion Methods

- Classic methods require traffic to be forwarded to the firewall device by trusting the next hops present in the Layer-2 and Layer-3 RIBs/FIBs. IP/MAC lookup only
- Different traditional flavours:
  - Firewall As Default Gateway
  - Firewall As Perimeter Device
- Introduce configuration complexities and/or sub-optimal performances
- Administrators are forced to disable or work around key fabric functions.
  - DAG when firewall is used as default gateway
    - Although functionality is preserved, traffic experiences significantly more hairpinning.

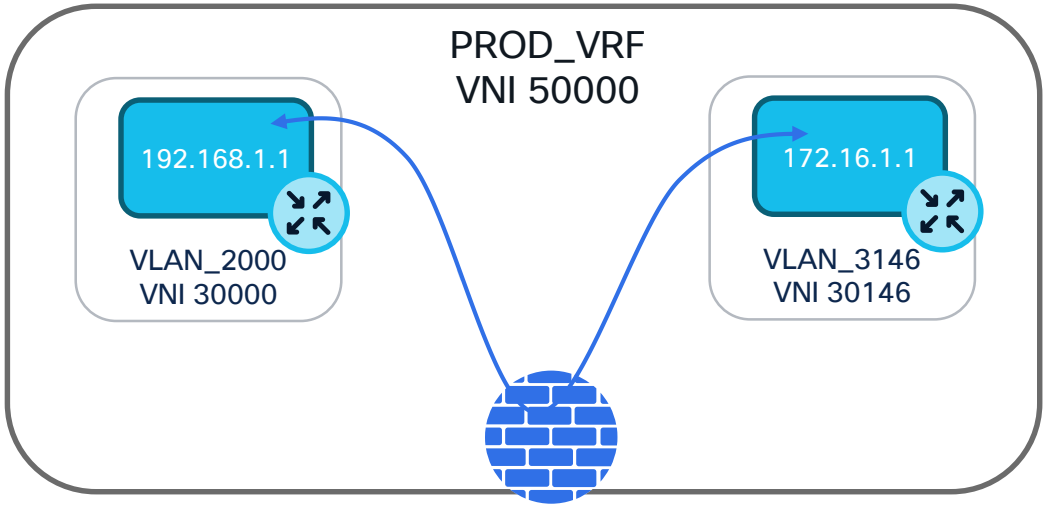


Classic VRF Sandwich

Segmentation

Service Redirection Methods

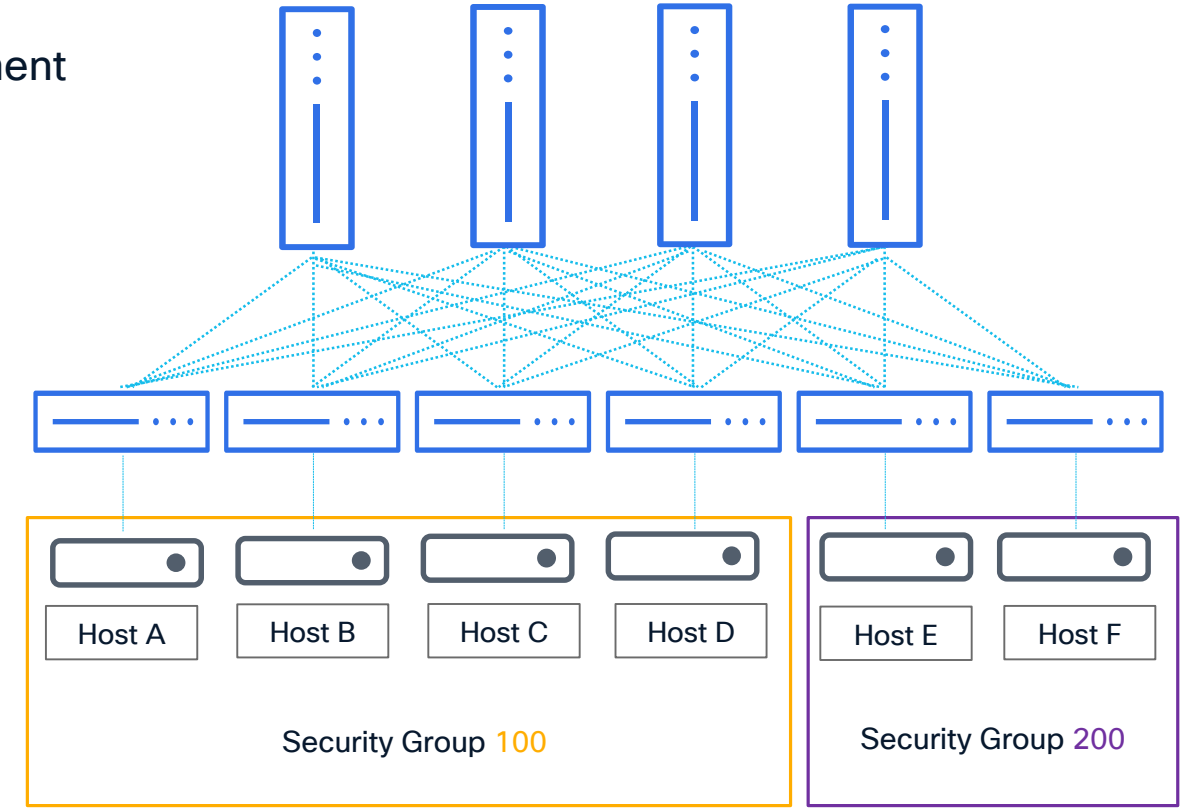
- Redirection/Stitching methods reroute traffic through security devices using redirection techniques, there is no need to influence Layer-2 and Layer-3 RIBs/FIBs.
- NX-OS has multiple features that helps administrators under the ePBR umbrella term
  - Available since NX-OS 9.3(5)
  - Automated and simplified configuration
  - Ensures the symmetry will be maintained
  - Integrates with IP SLA
  - Supports IP access-list and now GPO



Group Policy Option

Enhancing Data Center Segmentation with Simplicity and Granularity

- VXLAN GPO Standard
  - Extension to VXLAN and EVPN that allows policy enforcement
  - Backward Compatible
- Grouping and Classification
  - Classify endpoints into security-groups
  - Based on IP, VLANs, Port
- Enforcement for inter-group traffic
  - Permit
  - Deny
  - Redirect

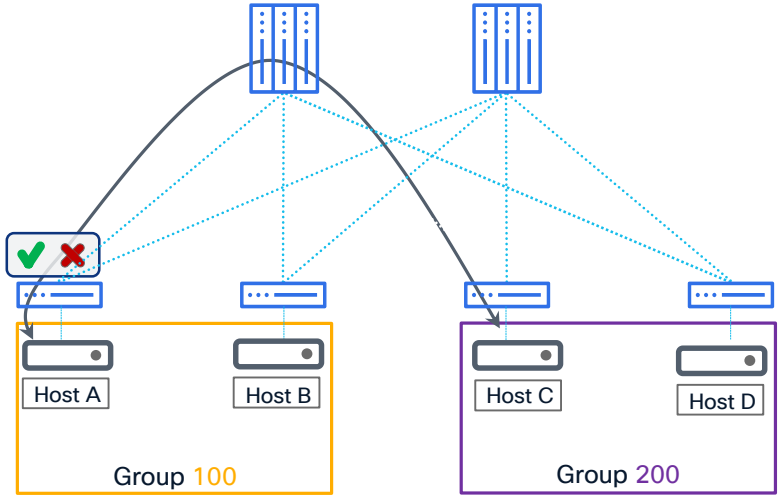


Group Policy Option

Main Use Cases

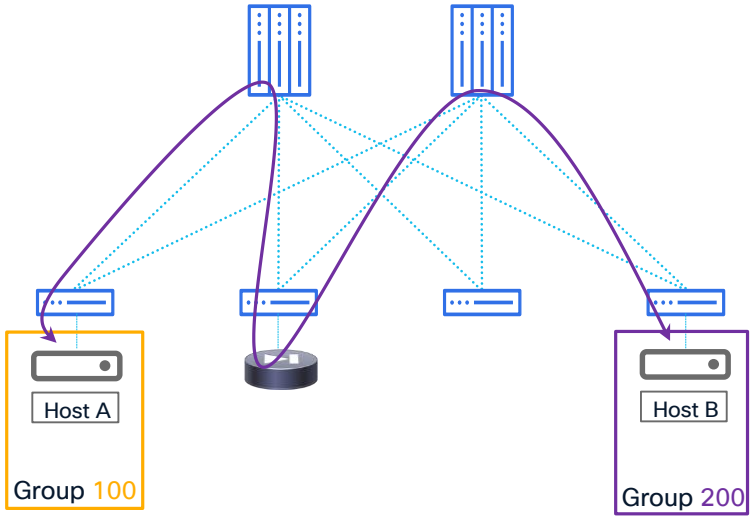
Segmentation

- Policies for macro and micro segmentation
- Control N/S and E/W traffic
- Granular filtering



Service Redirection

- Easily insert service devices in the traffic path
- Traffic can be steered w/o touching routing tables or creating complex network designs



VXLAN GPO – Standards Based Segmentation

Data Plane
VLAN Group Policy Option

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: April 25, 2019

M. Smith
Cisco Systems, Inc.
L. Kreeger
Arrcus, Inc.
October 22, 2018

VXLAN Group Policy Option
draft-smith-vxlan-group-policy-05

Abstract

This document defines a backward compatible extension to Virtual eXtensible Local Area Network (VXLAN) that allows a Tenant System Interface (TSI) Group Identifier to be carried for the purposes of policy enforcement.

Control Plane
Group Policy ID BGP Extended Community

bess
Internet-Draft
Intended status: Standards Track
Expires: 22 April 2024

W. Lin
Juniper Networks
J. Drake
Individual
D. Rao
Cisco Systems
20 October 2023

Group Policy ID BGP Extended Community
draft-wlin-bess-group-policy-id-extended-community-03

Abstract

Group Based Policy can be used to achieve micro or macro segmentation of user traffic. For Group Based Policy, a Group Policy ID, also known as Group Policy Tag, is used to represent a logical group that shares the same policy and access privilege. This specification defines a new BGP extended community that can be used to propagate Group Policy ID through a BGP route advertisement in the control plane. This is to facilitate policy enforcement at the ingress node when the optimization of network bandwidth is desired.

Data Plane and Control Plane
EVPN Group Policy

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: 5 September 2024

W. Lin
Juniper
D. Rao
A. Sajassi
M. Smith
Cisco
L. Kreeger
Arrcus
4 March 2024

EVPN Group Policy
draft-lrss-bess-evpn-group-policy-00

Abstract

Group Based Policy can be used to achieve micro or macro segmentation of user traffic. For Group Based Policy, a Group Policy ID, also known as Group Policy Tag, is used to represent a logical group that shares the same policy and access privilege. This document defines a backward compatible extension to Virtual eXtensible Local Area Network (VXLAN) that allows a Group Policy ID to be carried for the purposes of policy enforcement at the egress Network Virtualization Edge (NVE). It also defines a new BGP Extended Community that can be used to propagate Group Policy ID through a BGP route advertisement in the control plane. This is to facilitate policy enforcement at the ingress NVE when feasible.

Group Policy Option

Compatibility and Licensing

Licensing: Essentials

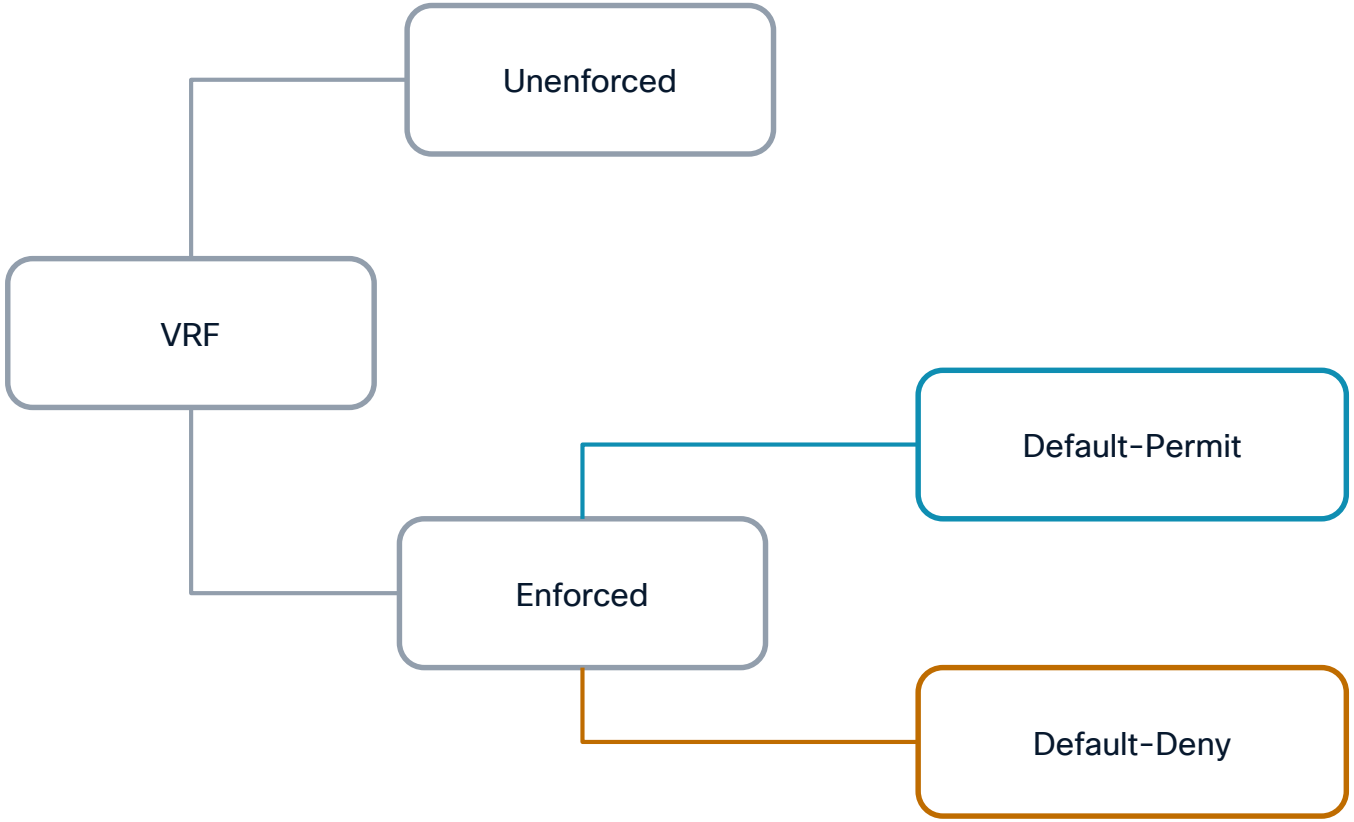
- GPO introduced in NX-OS 10.4(3)F
  - FX3, GX and GX2 platforms*
  - Permit/Deny enforcement options
  - Single and Multi Site** Fabrics
- NX-OS 10.5(1)F introduced
  - 9408 platform*
  - Single Service Function Redirection in Single Fabrics
- NX-OS 10.5(2)F introduced
  - FX, FX2 and HX platforms*
  - Multi Service Function Redirection in Single and Multi-Site** Fabrics
  - IPv6 in Underlay

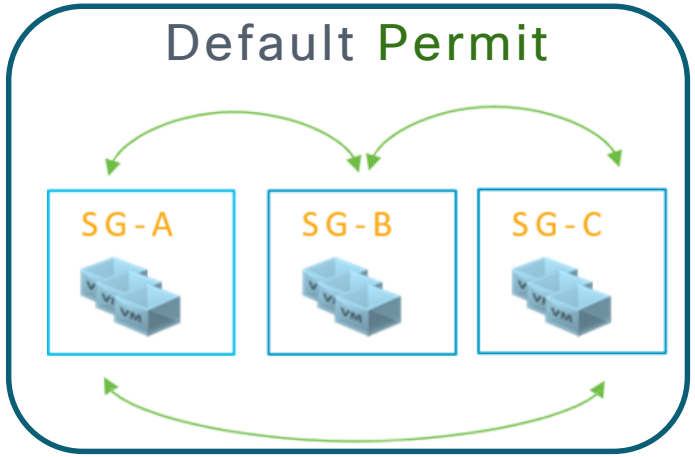
*Min 24G RAM

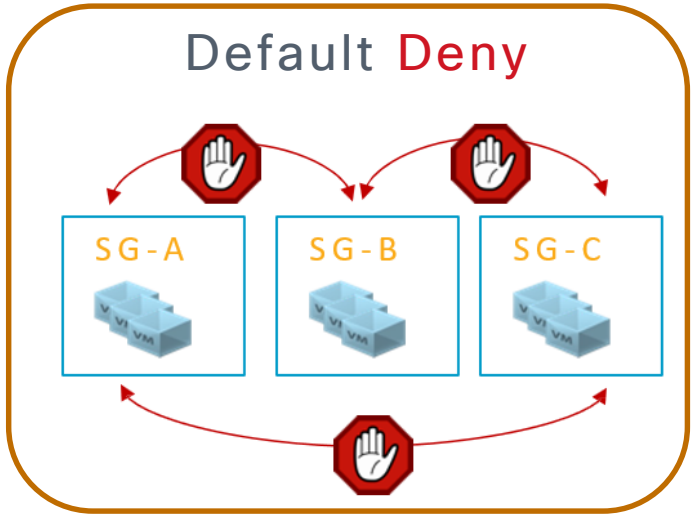
**Multi-Site requires Advantage on BGWs

Group Policy Option

VRF Options



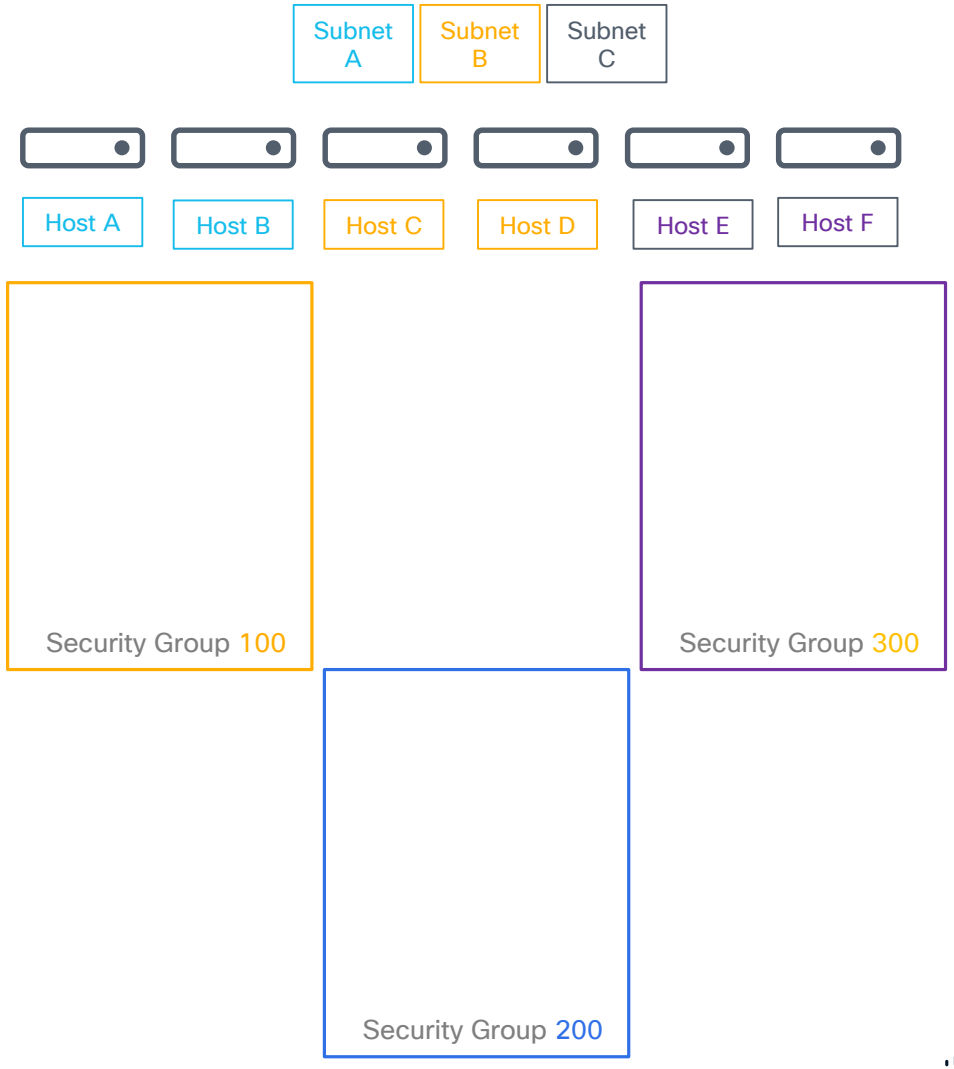




Group Policy Option Fundamentals

Endpoint Classification

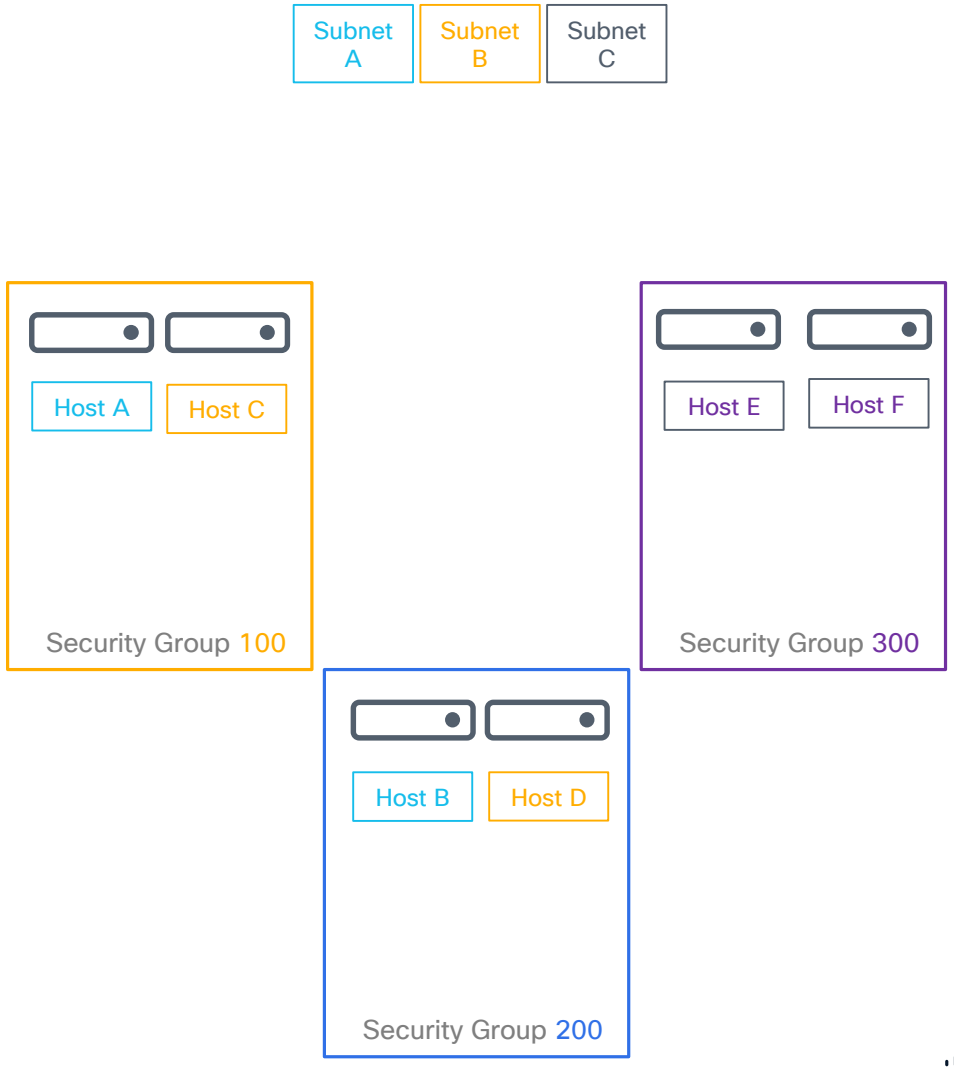
- Classification method depends on resource type
  - Connected Endpoints:
    - IP based classification
    - LPMs (including host-routes)
  - VLAN based classification
  - External Networks
    - IP Based Classification
      - LPMs



Group Policy Option Fundamentals

Endpoint Classification

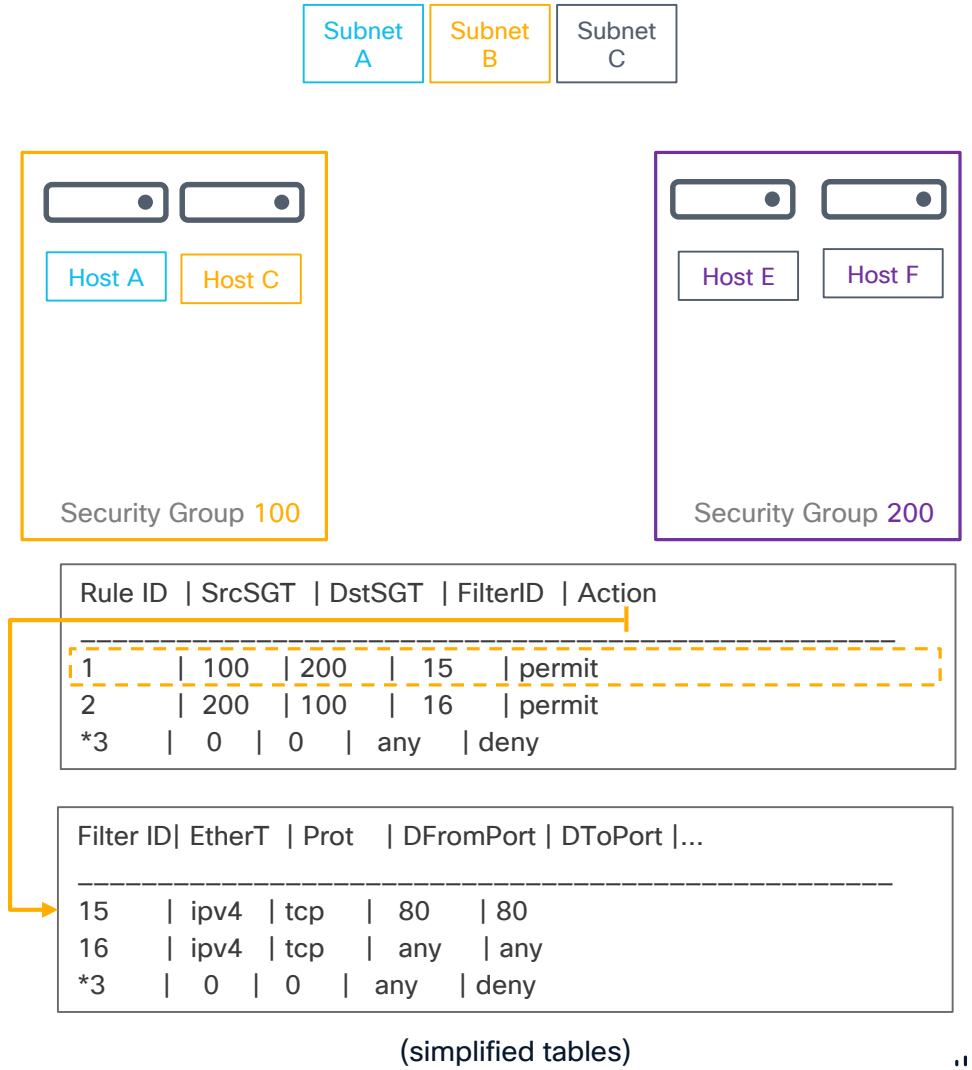
- Classification method depends on resource type
  - Connected Endpoints:
    - IP based classification
    - LPMs (including host-routes)
  - VLAN based classification
  - External Networks
    - IP Based Classification
      - LPMs
- Endpoints in the same subnet can be mapped to different or same SGs
- Endpoint in different subnets can be mapped to the same SG



Group Policy Option Fundamentals

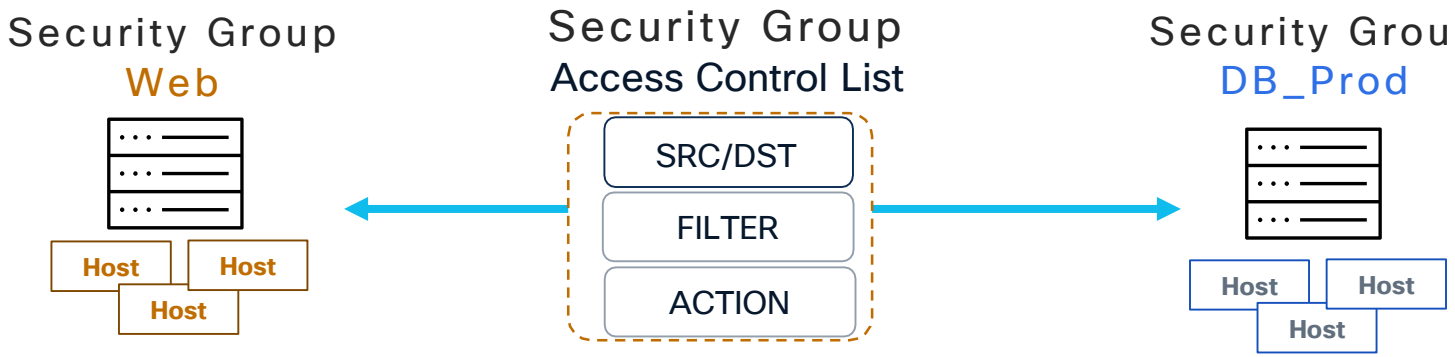
Traffic Enforcement

- Traffic enforcement is done by the VTEPs
- Every packet is evaluated in a security-rule table build based on configurations that contains records(SGACLs) to identify source and destination SGT and the conditions (protocol, ports etc)
- When a match is found then the correspondent action is applied
- When multiple matches are configured then a priority logic will be used for the tie break
- If a 1:1 match is not found then a catch–all rule will be hit, the action in that case depends on configurations*



(simplified tables)

VXLAN GPO – Configuration Steps



Step 1: Enable Security Group Feature

```
feature security-group
```

Step 2: Create Security Group (SG)

```
security-group 100 name DB_Prod
   match connected-endpoints vrf vrf_a ipv4 10.1.1.2/32
   match connected-endpoints vrf vrf_a ipv4 10.1.1.3/32
   match connected-endpoints vrf vrf_a ipv4 30.1.1.2/32


security-group 200 name Web
   match connected-endpoints vrf vrf_a ipv4 20.1.1.2/32
   match connected-endpoints vrf vrf_a ipv4 20.1.1.3/32
```

Step 3: Define Protocols and Corresponding Security Action

```
class-map type security match-any db_ports
   match ipv4 tcp stateful dport 3306
   match ipv4 tcp stateful dport 1433

policy-map type security web_db
   class db_ports
     permit
```

Step 4: Apply VRF Enforcement & Establish Contracts

```
vrf context vrf_a
   security enforce tag 10000 default deny
   security contract source 100 destination 200 policy web_db
```

Group Policy Option Fundamentals

What about Multi-Site

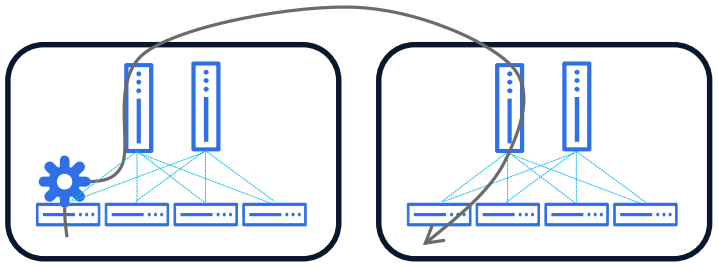
- There are no differences in terms of EP classification, security enforcement and configurations
- Control Plane:
  - No granularity for prefixes received from policy-unaware sites, they will be marked with SGT 15
  - Policy-aware BGWs will re-originate the remote type-2 and type-5 in the local fabric with SGT 15
  - Contracts associations must be defined using this SGT when enforcing traffic to/from these sites
- Data Plane:
  - Traffic can always be enforced based on the above assumptions
  - Enforcement will happen either at the policy-aware ingress leaf or at the BGW, depending on the traffic flow

Group Policy Option Fundamentals

What about Multi-Site

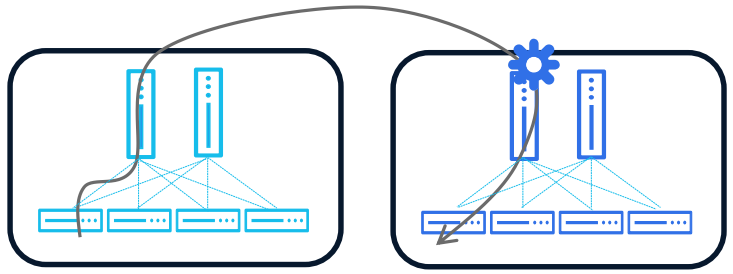
GPO is supported in VXLAN EVPN Multi-Site in these scenarios

Policy Aware → Policy Aware



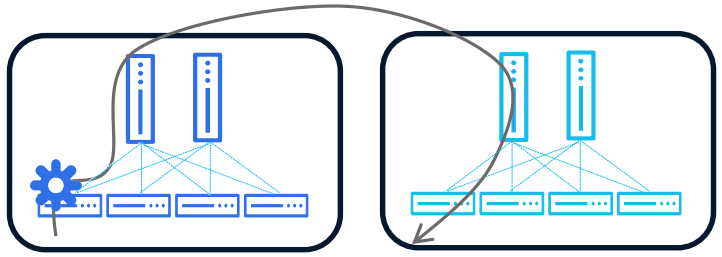
Policy applied on Ingress Leaf

Policy Unaware → Policy Aware



Policy applied on BGW* in policy aware Site

Policy aware → Policy Unaware



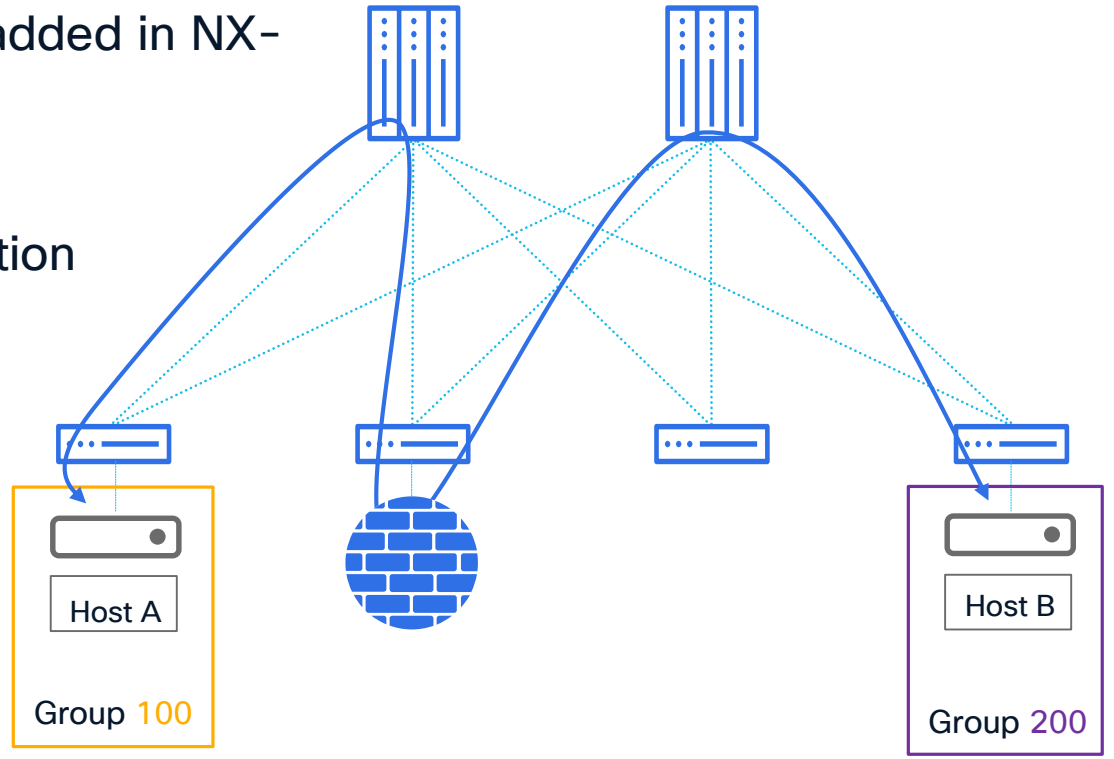
Policy applied on Ingress Leaf

* BGW-Spines for diagram simplicity

GPO Based Service Redirection

Highlights

- GPO Service Redirection support introduced in NX-OS 10.5(1)F
- Multi-Site and multi-node service-chains support added in NX-OS 10.5(2)F
  - Maximum 5 service-functions
- Extends SGACL capabilities with the redirection action
- Redirection can leverage ePBR functionalities
  - Simplified configuration workflow
  - IP SLA service endpoints health checks
  - Symmetric traffic forwarding
  - Various load balancing algorithms

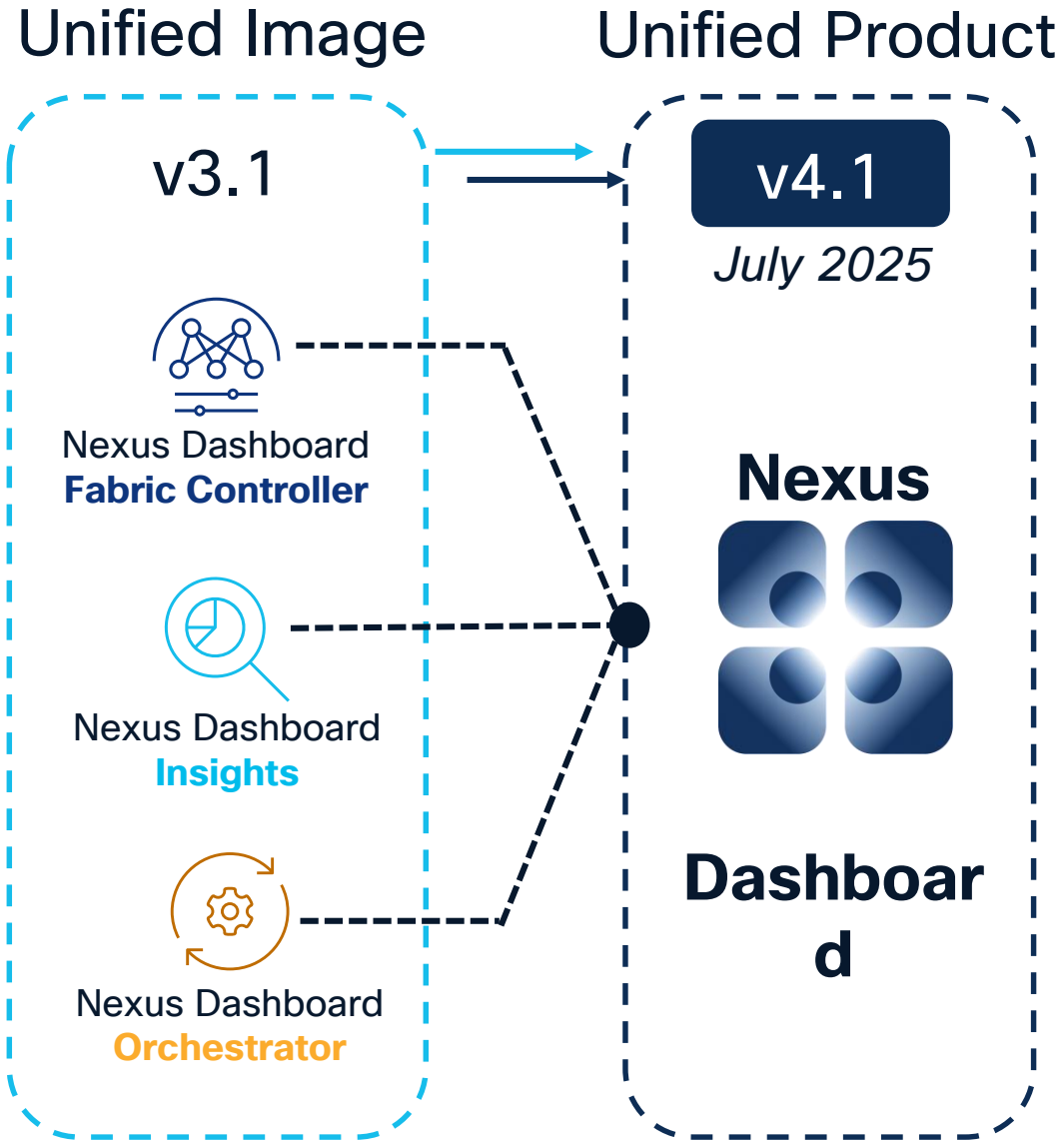


Nexus VXLAN Fabric Operations

Cisco Nexus Dashboard – Product Unification

Unified Image

- Single Image Installation
- All Services pre-installed (NDO/NDI/NDFC)
- Service versions bound to ND version



Unified Product

- Single image installation
- Services merged into base OS
- Services == ND Capabilities
  - Automation *[Controller]*
  - Telemetry
  - Orchestration

Cisco Nexus Dashboard Vision

Your key to a simple operational experience

Unify data center network architectures

Accelerate business innovation, minimize risk

Unleash the power of AI for networking

Reactive, proactive and predictive insights

Leverage open protocols and standards

Provision

Once, deploy anywhere

Built-In automation | Infra-as-code

Secure

One source of truth

Segmentation | Zero-Trust

Manage

One logical network

Visibility | Maintenance

Analyze

Identify, troubleshoot, and fix faster

Tools | Reports | Prediction

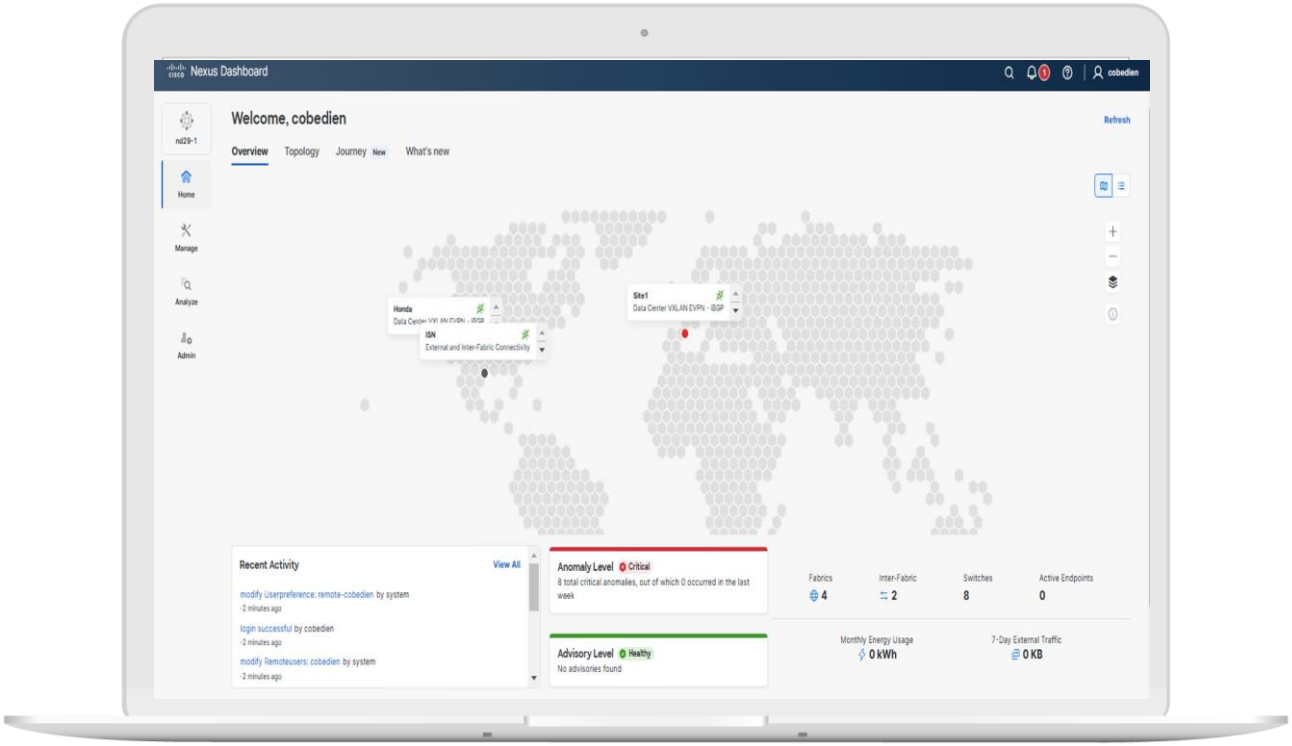
Cisco Nexus Dashboard Architecture

Microservices Architecture

Both Virtual and Physical ND Supported

One Manage | Multi-Cluster Support

3-Node Active-Active High Availability

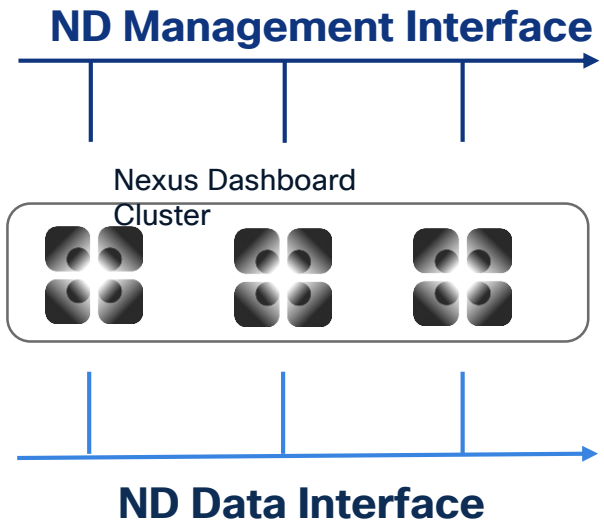


Benefits

Flexibility and Scalability | Highly Customizable | Single Vantage Point

Cisco Nexus Dashboard Clustering

Each Nexus Dashboard Node in a cluster has two interfaces, each in a *different subnet*



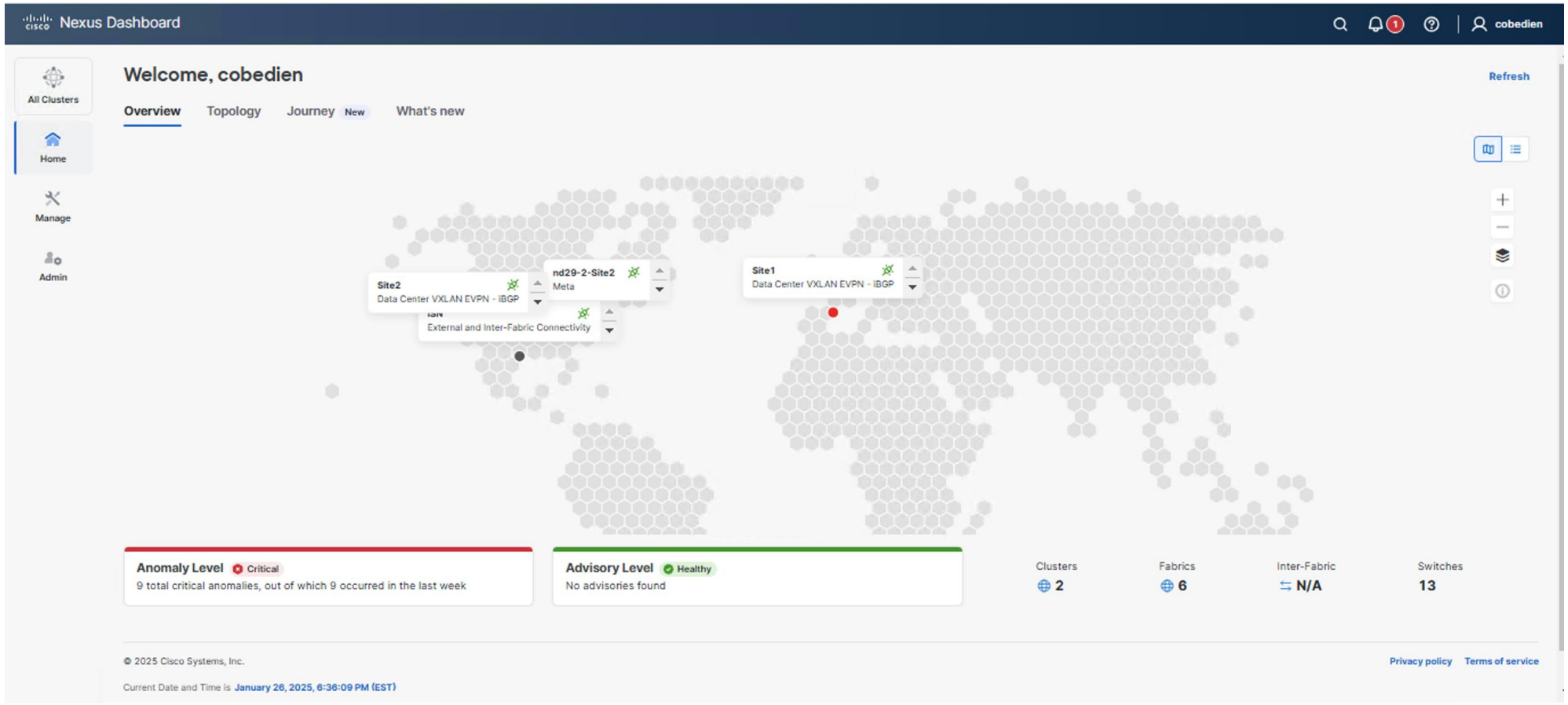
Nexus Dashboard Management Interface

Connects to the management network, and it provides web/API access to the Nexus Dashboard cluster

The Nexus Dashboard Data Interface

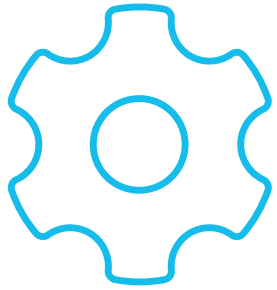
Connects and provides IP reachability to the physical data center network infrastructure, via In-Band or Mgmt0 (OOB)

Cisco Nexus Dashboard One-Manage



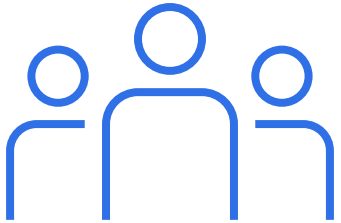
One-Manage View for Multi-Clusters

Cisco Nexus Dashboard Three Pillars



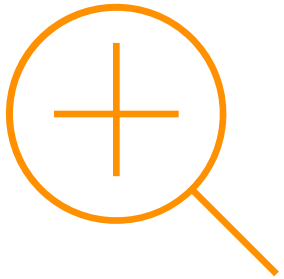
Automation

Accelerate provisioning and simplify deployments



Management

In depth Management and control for all network deployments



Visibility

Get Centralized Visibility and Monitoring views

Cisco Nexus Dashboard

One-Mange / Federation for Multi-Cluster Deployments

Single sign-on for clusters

(With remote authentication)

Single dashboard

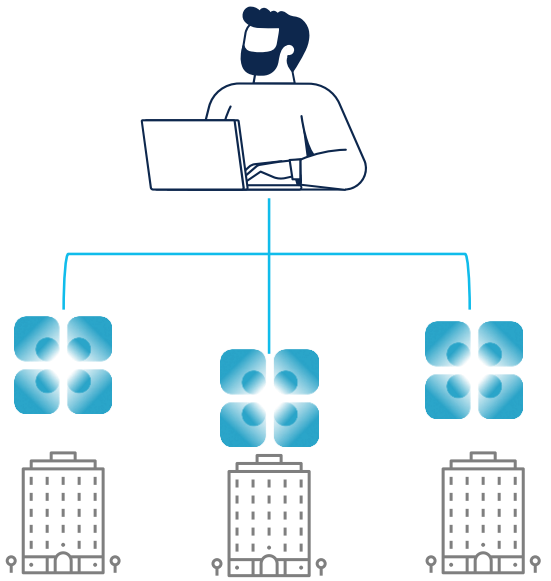
Including a summary of fabrics, switch details, etc.

VXLAN EVPN Fabric

- Unified view and automation workflows for Multi-Fabric deployments
- Federation of MSDs
- L2/L3 Network stretch

10 Maximum clusters in a federation

Based on 4.1 scale *(Increased from 4 in v3.2)*



Distributed Data Centers

Benefits

Scale Out Deployments	DCI Control Between DC and ND (Multi-Fabric)	DCI Connectivity Health, Tunnel Path and Statistics

Automation

Accelerate Provisioning From Days to Minutes

Day 0 – VXLAN EVPN Underlay Configuration

Define fabric settings (Underlay/Overlay), AS#, Replication Mode, IGP, IP Pools

Import Switches and Define Roles

Calculate and Deploy

Day 1 – VXLAN EVPN Overlay Configuration

VRF and Network Configuration and Attach

Interface Policy and Attach

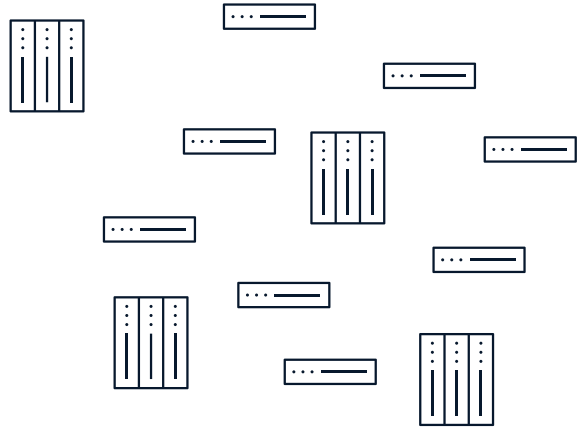
Calculate and Deployment

Benefits

Simplify Fabric Deployments	Developer Agility	VXLAN EVPN Multi-Site

Provision New Fabric in Minutes

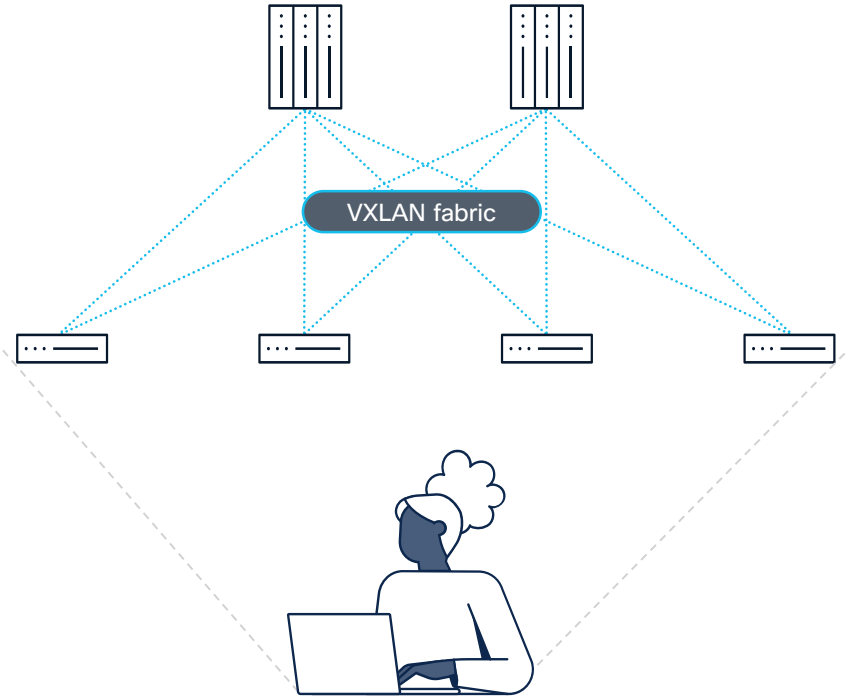
Un-provisioned switches



Fabric Builder

Cisco Best Practice Configuration Built-In

Support for Both Greenfield and Brownfield Deployments



Benefits

Accelerate Fabric Deployments	Automated Consistency	Minimize Risk

Data Center VXLAN EVPN – Day-0 Underlay

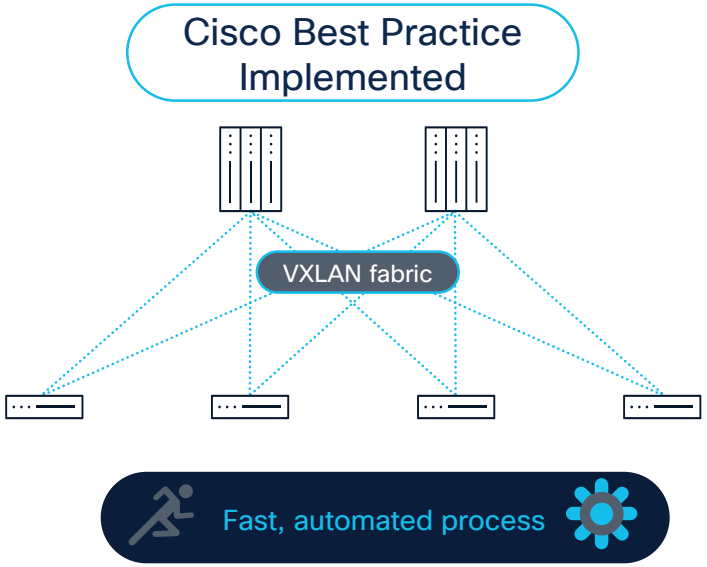
Not on VXLAN EVPN Today?

Nexus Dashboard Automation	Build VXLAN fabric in few minutes	Templates already embed best practices	IP addresses, overlay pool, routing profiles, replication attributes –all taken care by ND

Step 2
Discover
Import switches with POAP or Day-0 config
Define switch Roles (Border, Leaf, Spine, etc)
[Optional] Create vPC pairs

Step 1
Create
Define fabric settings (Underlay, Overlay) - AS#, Replication Mode, IGP, IP Pools, etc.

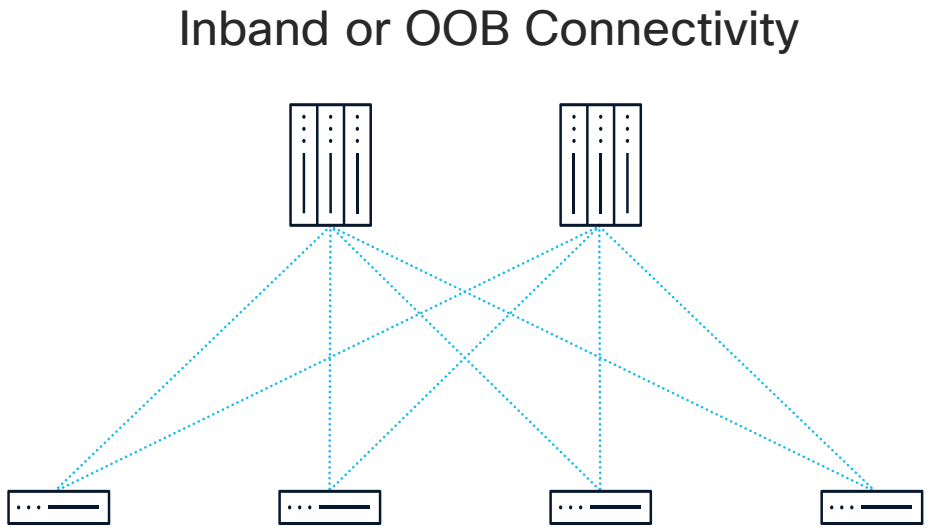
Step 3
Recalculate and Deploy
Generates config based on intent
Preview side by side diffs



Zero Touch Deployment

POAP – Power On Auto Provisioning

- Flexible bootstrap and management via in-band (front-panel port) or out of band port
- Supports VXLAN EVPN, classic LAN fabrics
- Convenient connectivity options for all device roles: leaf, spine, border leaf, border spine



Benefit

Zero Touch Fabric Onboarding and Management

Fabric Switch Role Example

Different Roles for Border Gateway (BGW)

vPC Border Gateway

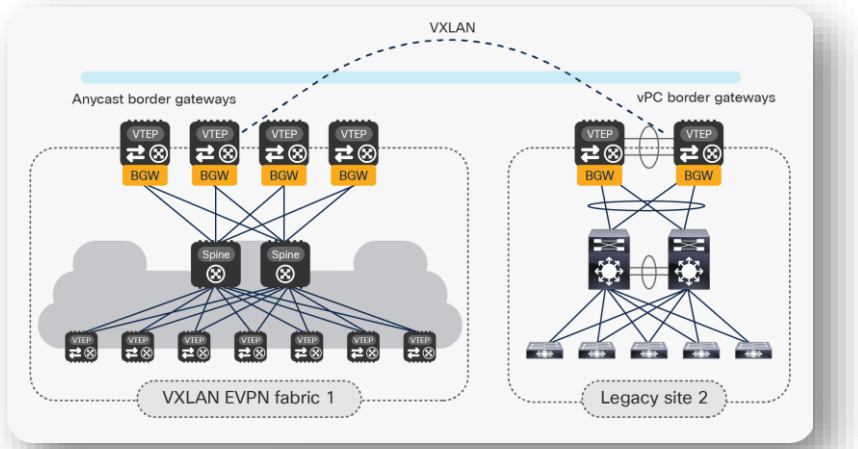
Used to locally dual-attach Layer 2 networks or Endpoints
Allows Distributed Anycast Gateway (DAG)

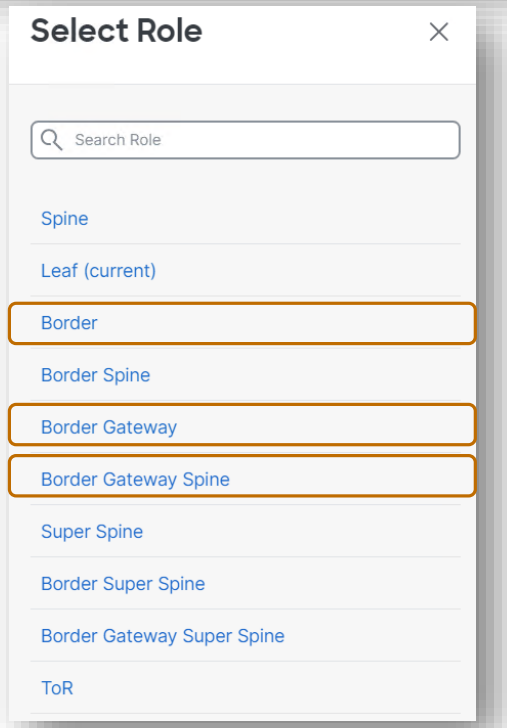
Border Gateway

Layer 3 based Anycast BGW deployed at the Leaf Layer

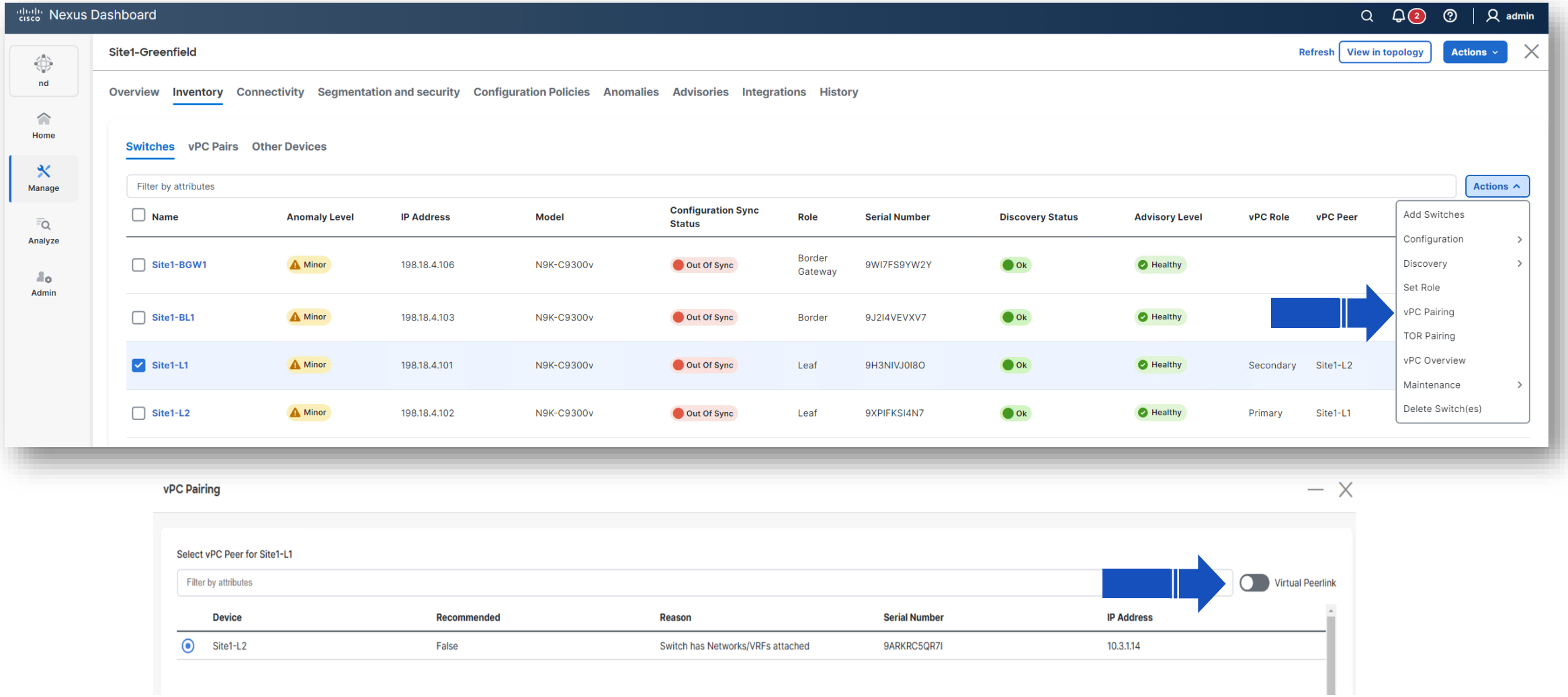
Border Gateway Spine

Layer 3 based Anycast BGW deployed at the Spine Layer

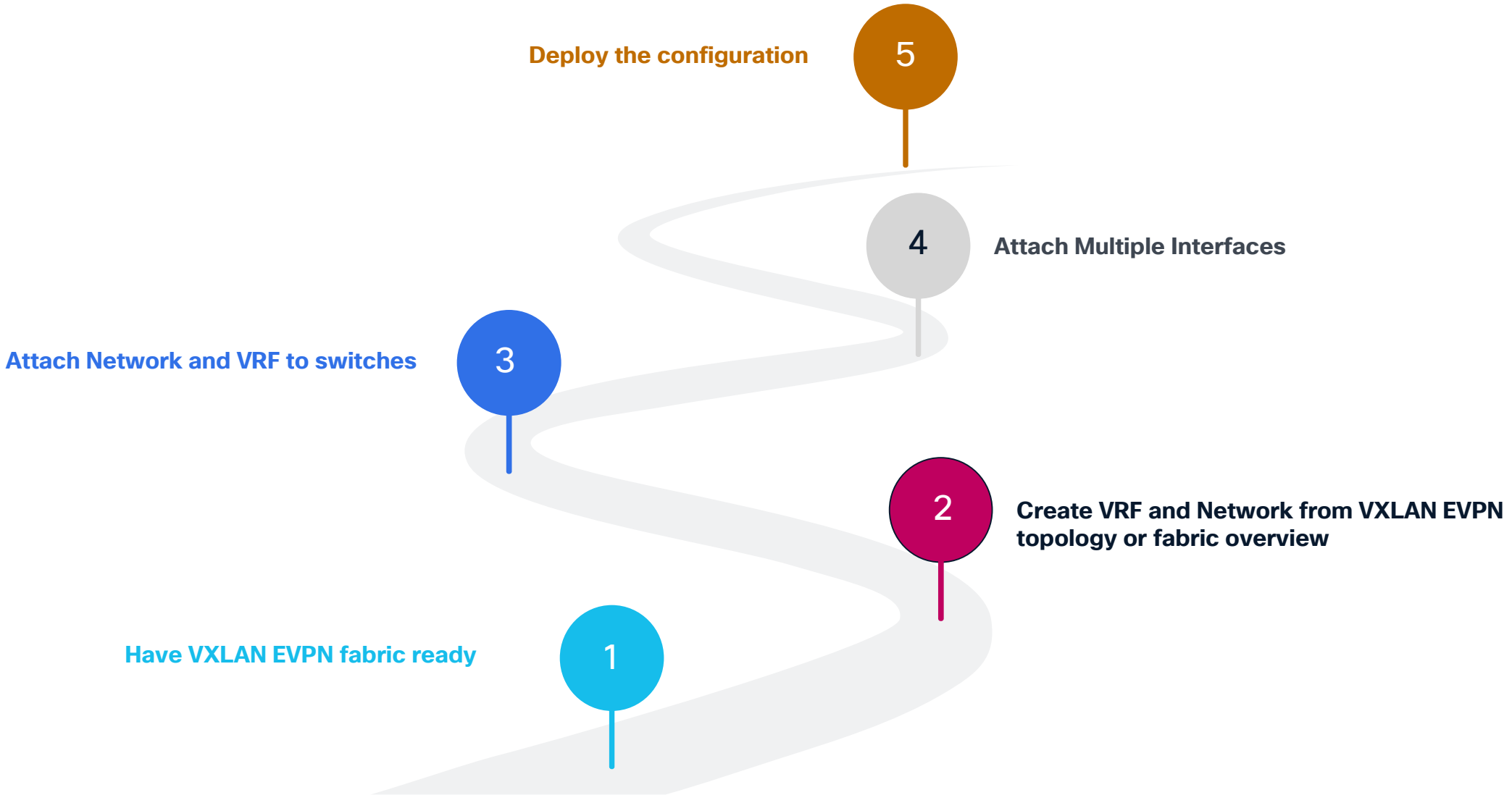




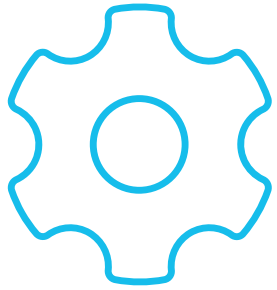
Virtual Port Channel (vPC)



Data Center VXLAN EVPN – Day-1 Overlay

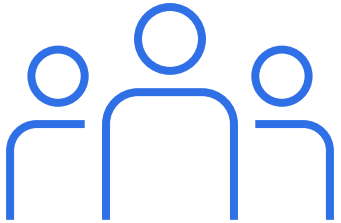


Cisco Nexus Dashboard



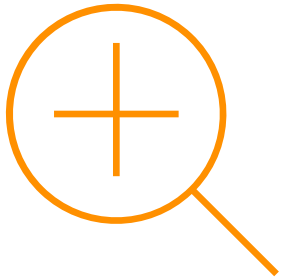
Automation

Accelerate provisioning
and simplify deployments



Management

In depth Management
and control for all
network deployments



Visibility

Get Centralized Visibility
and Monitoring views

Management

Single Point of Management for Data Center Operations

Configuration Compliance

Image and Patch Management

Fabric Backup and Restore

Hardware Replacement / RMA

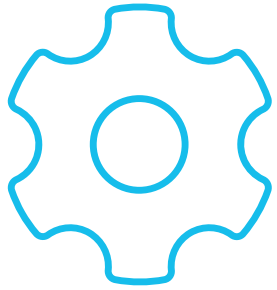
Role Based Access Control (RBAC) and Change Control

Management for non-Nexus platforms

Benefits

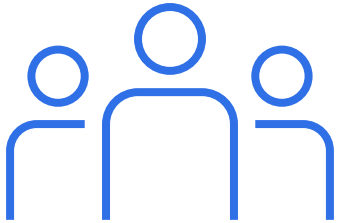
Reliability	Compliance	Secure

Cisco Nexus Dashboard



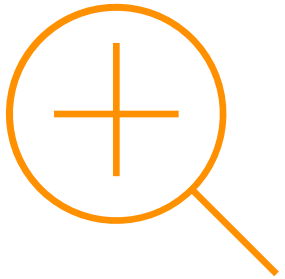
Automation

Accelerate provisioning
and simplify deployments



Management

In depth Management
and control for all
network deployments



Visibility

Get Centralized Visibility
and Monitoring views

Visibility and Monitoring

Comprehensive Monitoring

Enhanced Topology Views

Compute and Endpoint Visibility

OAM support with NDFC

Detailed Inventory, Health, Resource Consumption Information on Devices

End-to-End Visibility, Monitoring and Troubleshooting

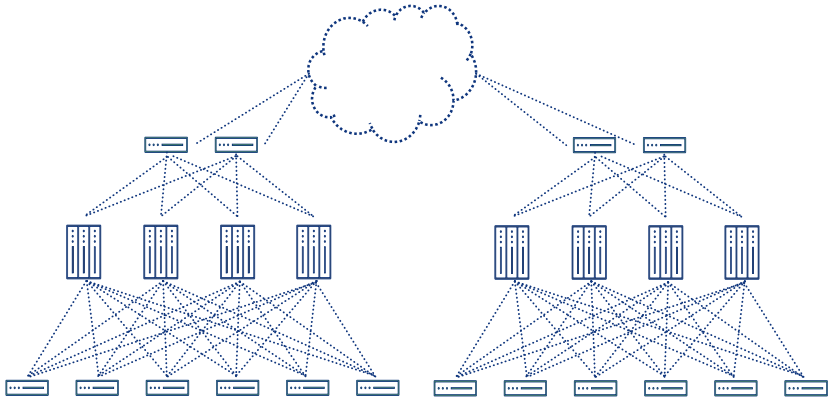
Integrate with Day 2 Operations

Benefits

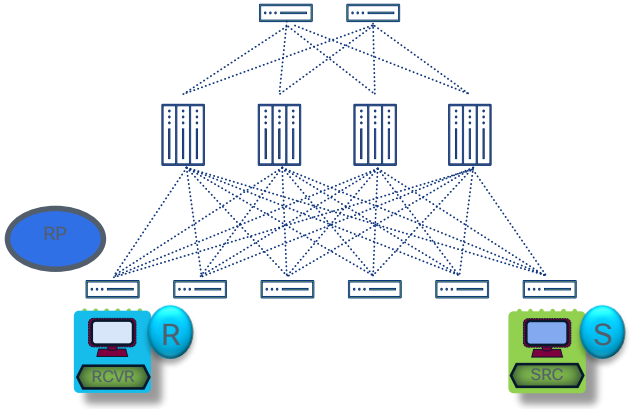
Reliability | Compliance | Secure

Advanced Features Support Pt 1

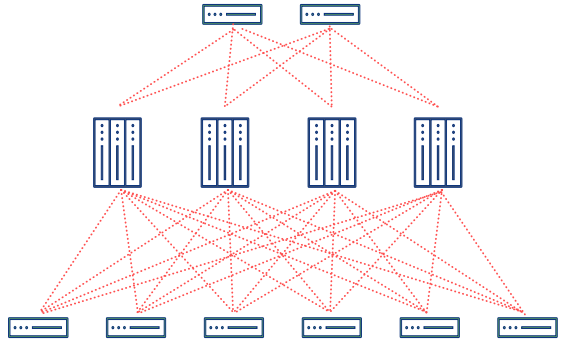
Muti-Site



Multicast



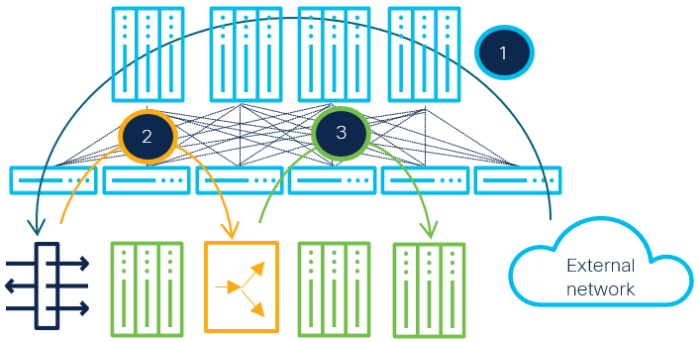
MacSec



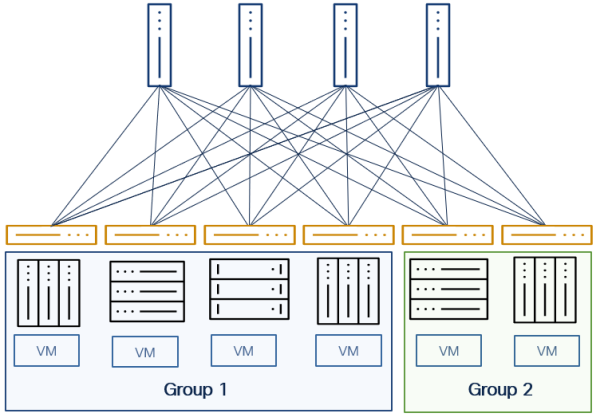
1000’s of CLI Configuration Lines

Advanced Features Support Pt 2

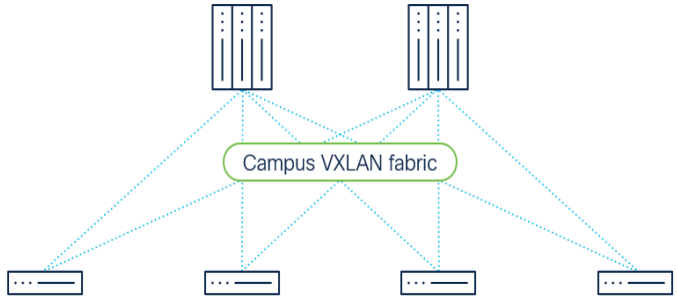
L4-7 Service Insertion



Segmentation / GPO



Campus VXLAN EVPN

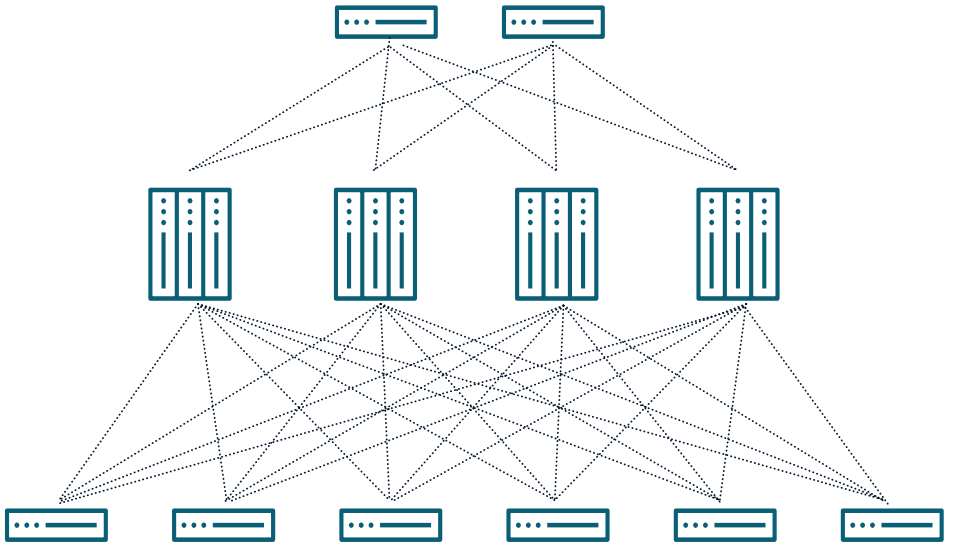


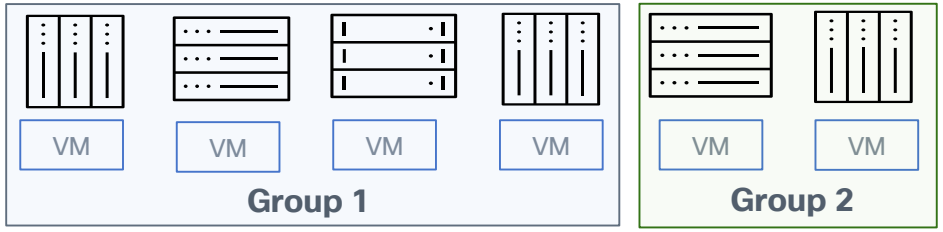
1000's of CLI Configuration Lines

Conclusion

Conclusion

- Requirements for Designing a Data Center Fabric Continue to Evolve
- Important Factors for Consideration Include:
  - Scalability
  - Performance
  - Oversubscription
  - Economics
  - Sustainability
- Build Modern Data Center Fabrics to Support:
  - Network Simplicity
  - Repeatability
  - Distributed Architectures (Minimize Failure Domains)
  - Better & Faster Troubleshooting
  - Automation and Lifecycle Management





Complete your session evaluations

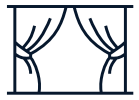
Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live mobile app.

Continue your education



Visit the Cisco Showcase for related demos

Book your one-on-one Meet the Engineer meeting

Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs

Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact me at: bburesh@cisco.com

Thank you

cisco Live !


CISCO