The bridge to possible

Simple Leaf Spine with a Touch of ToR

Network Designs for the Modern Data Center

Brenden Buresh, Distinguished Solutions Engineer

BRKDCN-2667

cisco Live!

#CiscoLive

Cisco Webex App

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2667

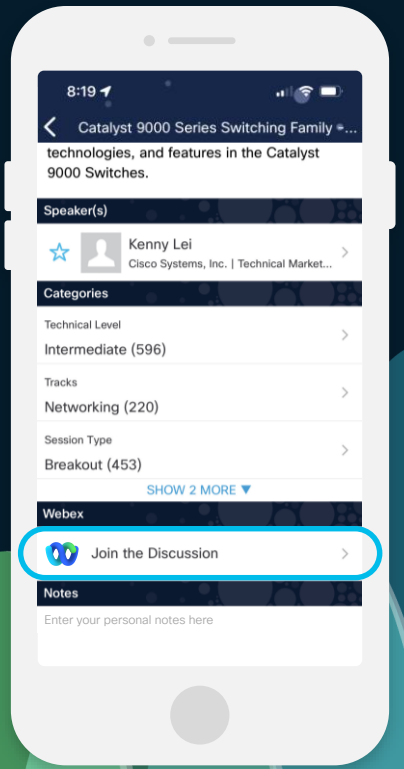
Questions?

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

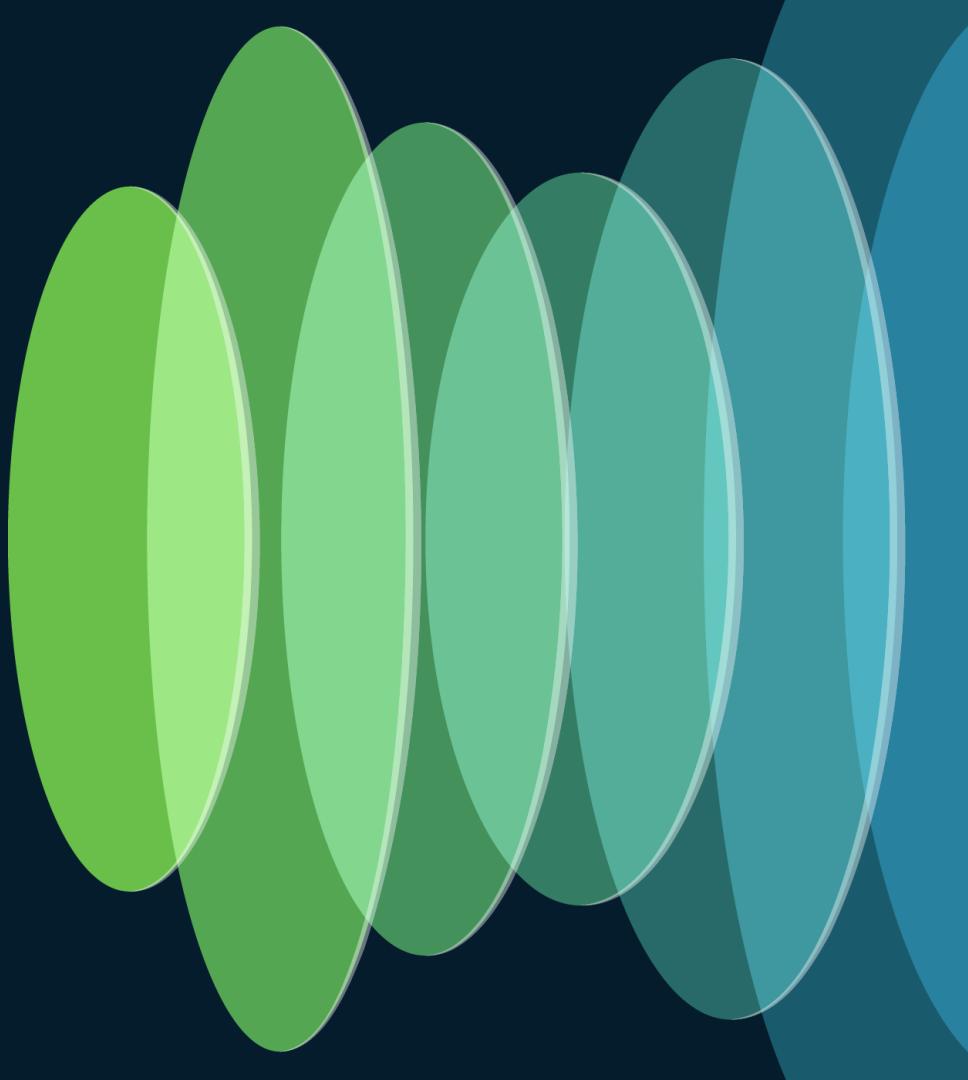
Webex spaces will be moderated by the speaker until June 7, 2024.



Agenda

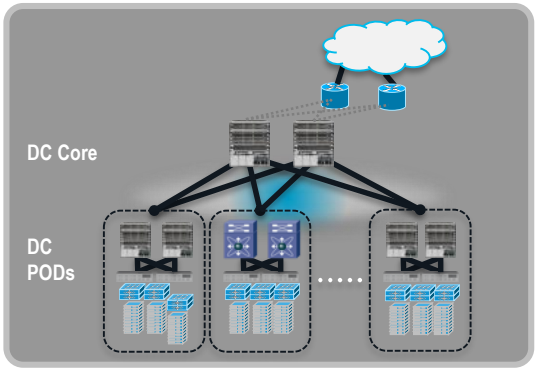
- Introduction
- The Evolution of DC Network Designs
- Leaf-Spine “Fat Tree” Network Topologies
- Hardware Deployment Considerations
- Fundamentals of VXLAN EVPN Design
- Multi Fabric Designs – VXLAN Multi Site
- Migration Considerations
  - Migration with Rack Space Constraints
  - Migration without Rack Space Constraints
- Conclusion

Introduction



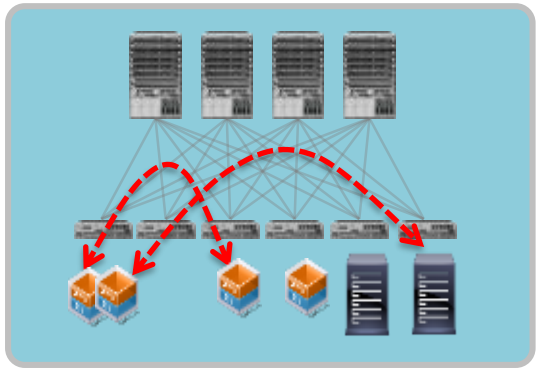
Evolving Network Designs

Traditional 3 Tier DC Network Design



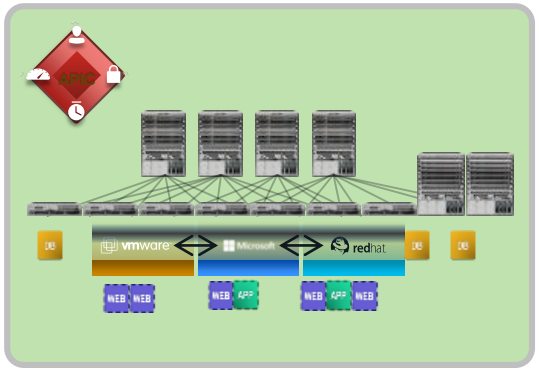
VPC in Access
Routed Aggregation & Core

Routed Fabric



VXLAN Bridging / Routing

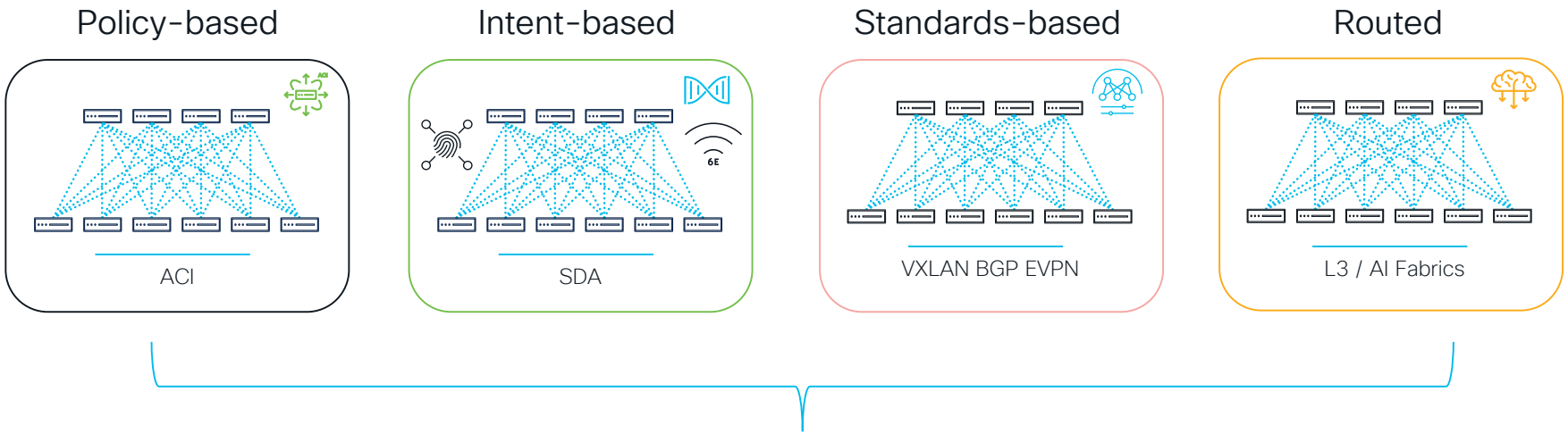
- VXLAN Flood & Learn
- VXLAN EVPN
- Separate Management Tools (e.g. Nexus Fabric Manager)



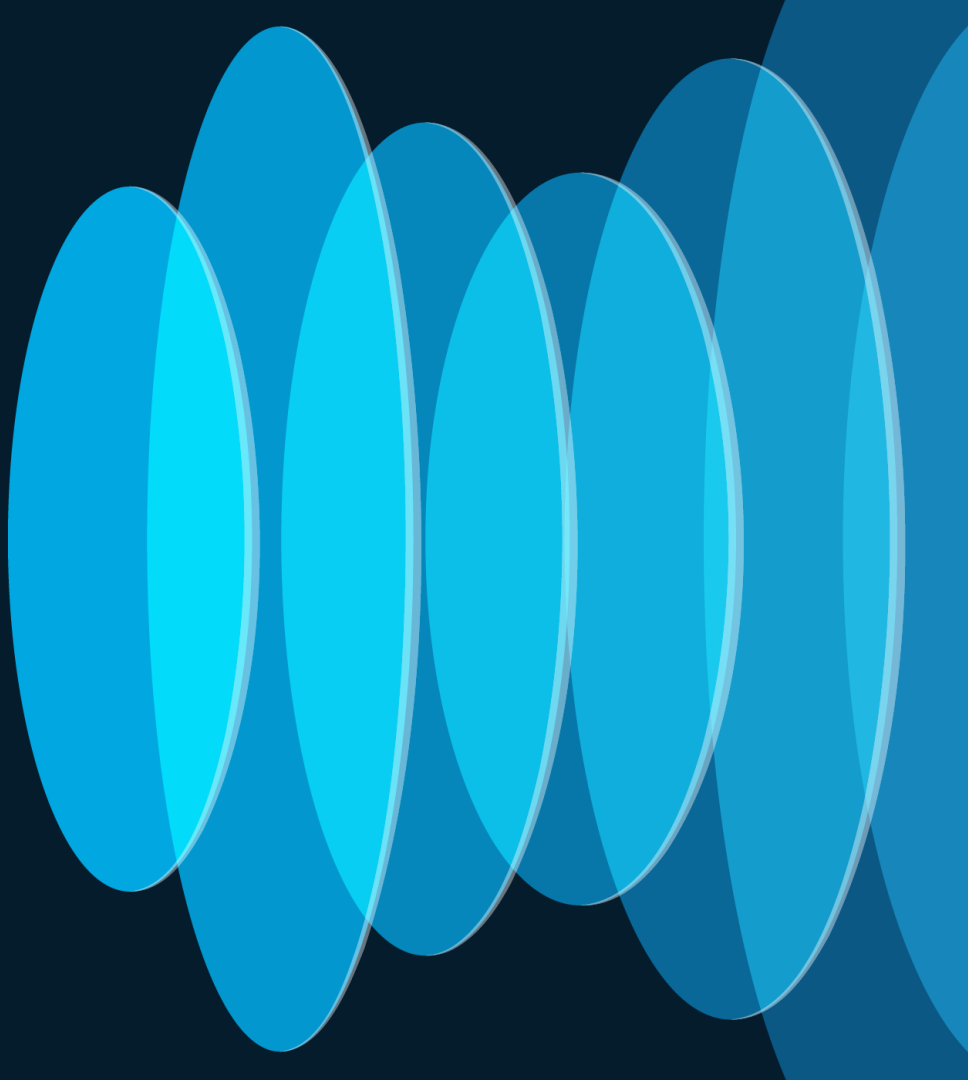
VXLAN Routing / Segmentation

- VXLAN Routing
- Policy Controller (APIC)
- Consistent policy across physical and virtual network
- Multi-hypervisor (VMware, MSFT, OVS)
- Endpoint agnostic (bare metal, VM, container)

Cisco Fabrics (Data Center & Campus)



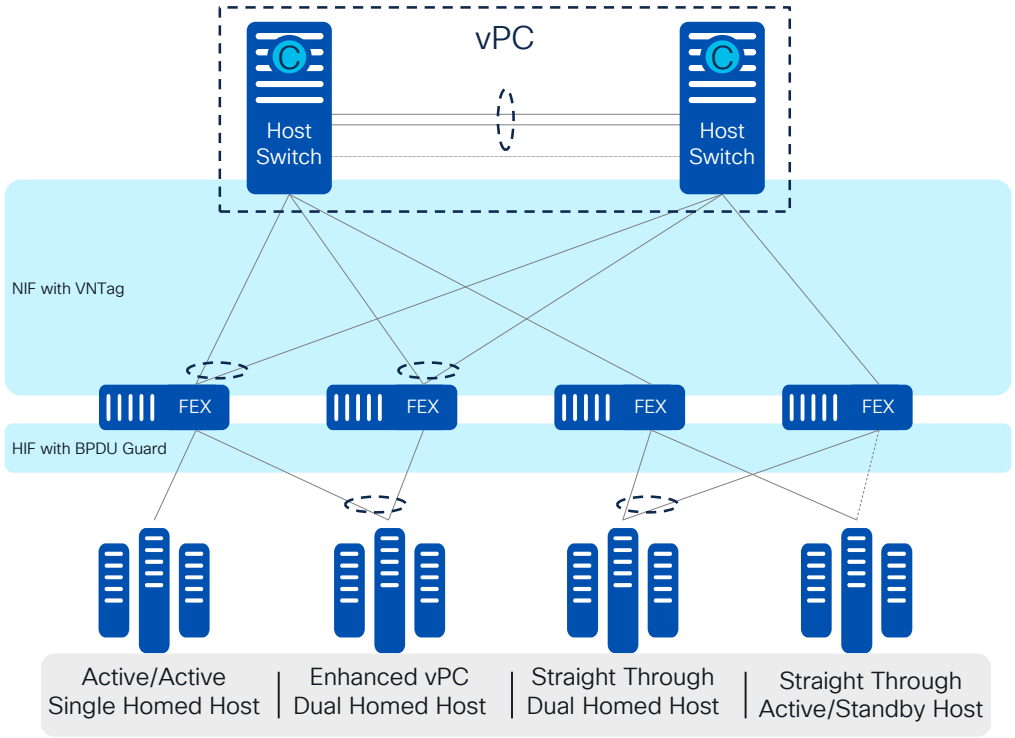
The Evolution of DC Network Designs



cisco Live!

A Data Center Fabric Prior to Data Center Fabrics

Cisco Fabric Extender (FEX) Overview



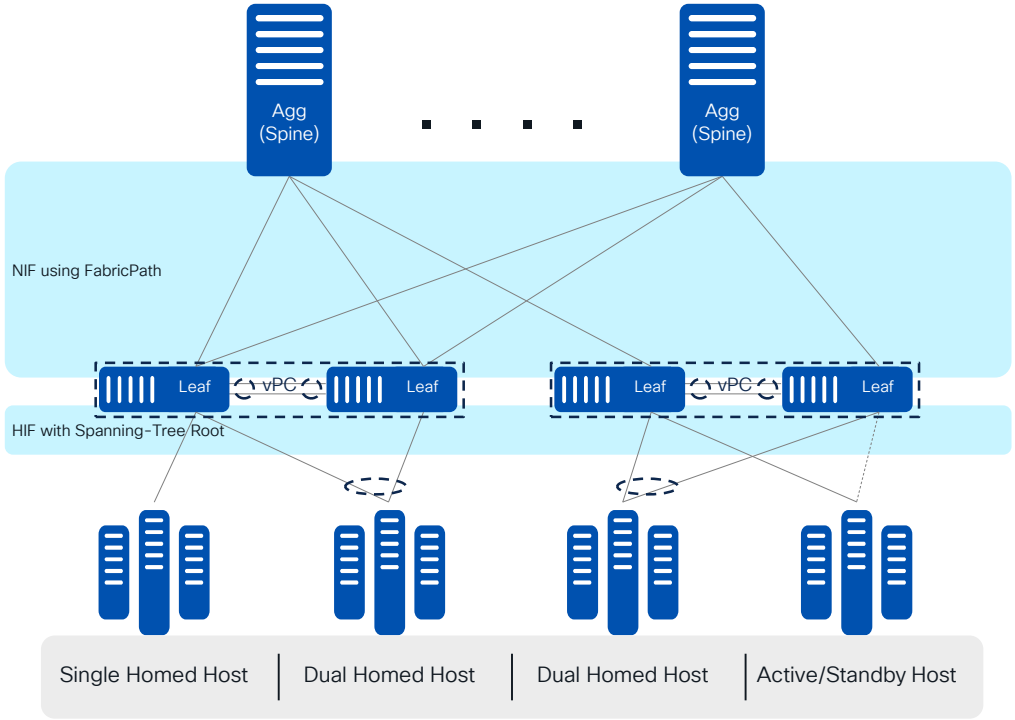
~15 Years ago
Around 2009

- Centralized Management
  - Co-located on the Switch
  - Limited to No Synchronization
  - Host Switch Operational Dependency
- Network Redundancy (NIF to NIF)
  - Uses VNTag (802.1BR / 802.1Qbh)
  - 1+1 Redundancy based on Layer-2 Port-Channel (vPC)
- Host Redundancy (Host to HIF)
  - Single Homed or Dual Homed Hosts (vPC, A/S)
  - Spanning-Tree BPDU Guard
  - Subset of HIF Capabilities (Dependent on Host Switch)

Early Steps in the Data Center Fabric Evolution

Cisco FabricPath Overview

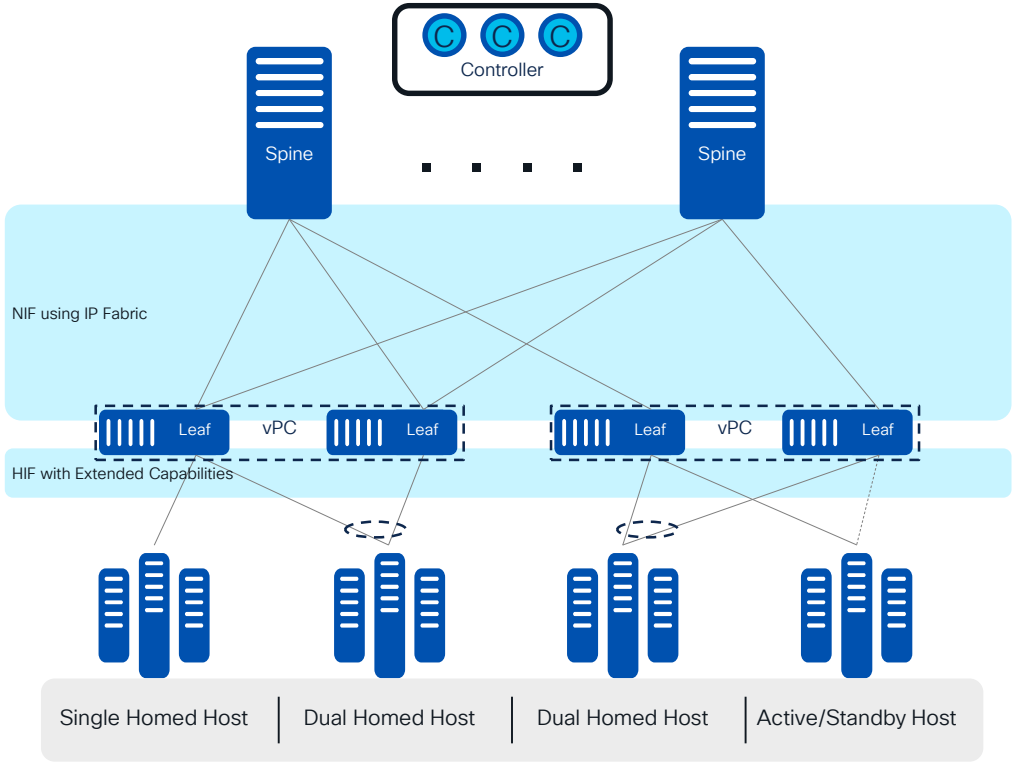
~12 Years ago
Around 2012



- Centralized Management
  - Nothing Really There
- Network Redundancy (Leaf to Spine)
  - FabricPath (MAC-in-MAC), requires Agg/Spine Support
  - N+1 Redundancy with ECMP
- Host Redundancy (Host to Leaf)
  - Single Homed or Dual Homed Hosts (vPC, A/S)
  - Full HIF Capabilities at Leaf with Spanning-Tree Root

Using Mature SDN for Data Center Fabrics

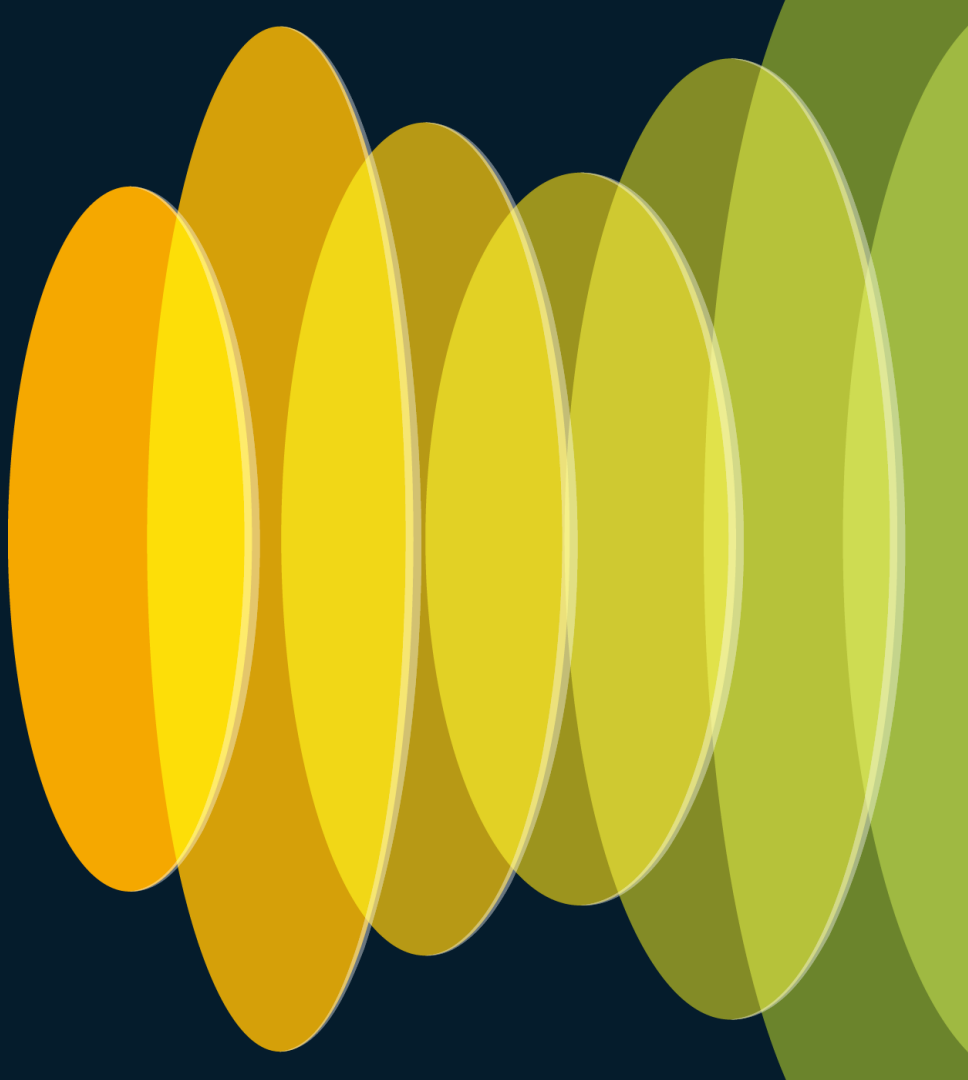
Cisco VXLAN EVPN Fabric Overview



~10 Years ago
Around 2014

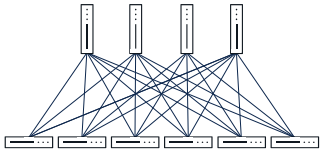
- Centralized Management
  - Independent to Switch Operating System
  - Full Config Synchronization
  - N+1 Cluster or High-Availability
- Network Redundancy (Leaf to Spine)
  - Uses VXLAN (RFC7348), the Spine is just an IP Router
  - N+1 Redundancy based on IP Fabric (ECMP)
- Host Redundancy (Host to Leaf)
  - Single Homed or Dual Homed Hosts (vPC, A/S)
  - Full HIF Capabilities

Leaf-Spine ”Fat Tree” Network Topologies

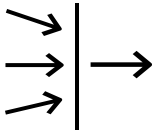


Key Requirements

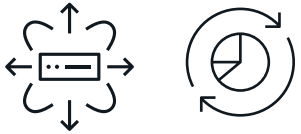
Topology
(Resiliency)



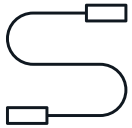
Bandwidth
(Oversubscription)



Platform Hardware
(Cost)



Optics
(Cabling)

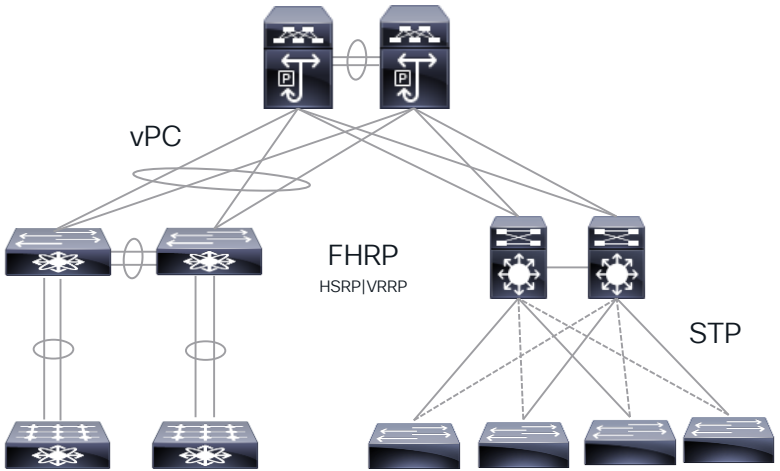


Define Requirements Before Designing Solution

Common Topologies

Three-Tier Hierarchical

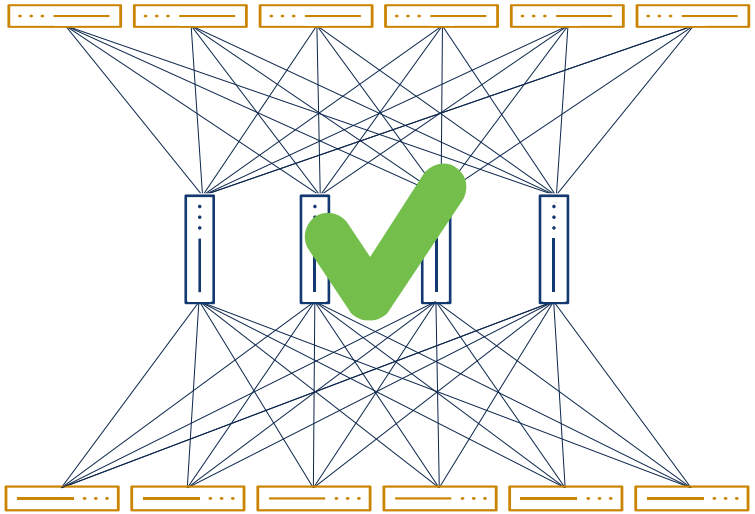
Inherent scale up topology building blocks
Requires multiple independent protocols configured in concert to be able to use all network links



"I always thought Ethernet forwarding with STP was a kludge, and the right solution was to do layer 3 forwarding..."
~Radia Perlman

Clos Fabric

Fully Interconnected 'fabric' with multiple design variations
Inherent modular scale out design (intra and inter-fabric)

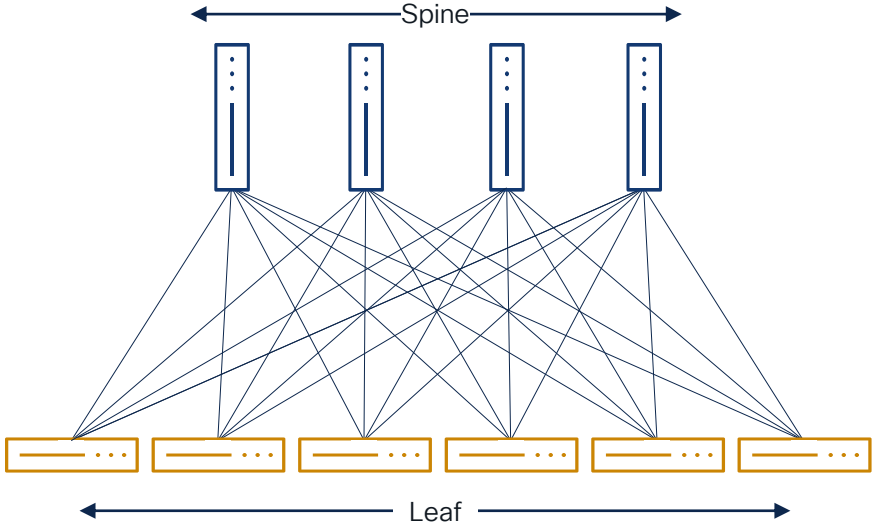


Clos Topology

A Leaf and Spine Topology

- Variations or Names of the same:
  - 2 Tier
  - Fat Tree
  - Folded Clos
  - 3 Stage Clos
  - Butterfly (5 Stage | PoD)

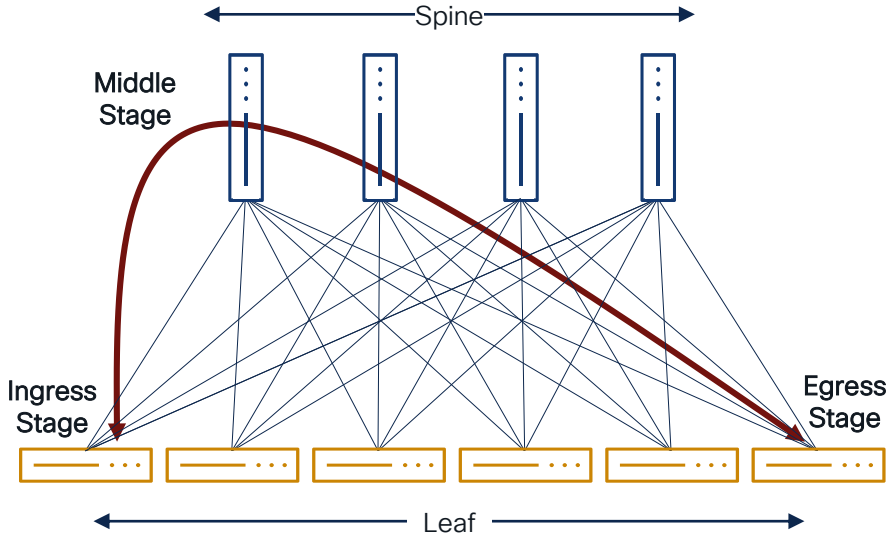
The Standard



Clos Topology

- A Leaf and Spine Topology
  - 3 Stages
  - 2 Tiers

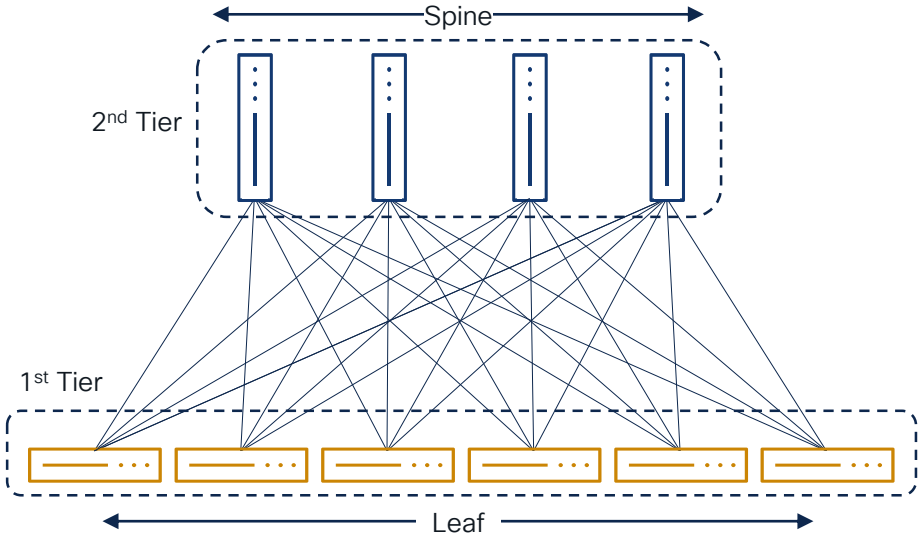
Advantages



Clos Topology

- A Leaf and Spine Topology
  - 3 Stages
  - 2 Tiers

Advantages



Clos Topology

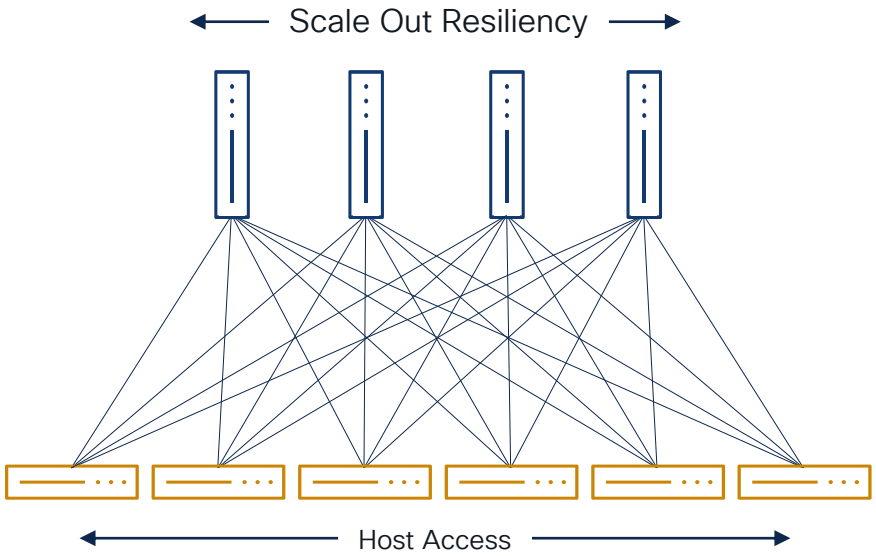
Scale Out Architecture

- Adding Spines results in:
  - Increased physical path redundancy

    -> Greater Resiliency
  - Increased leaf count *without increasing oversubscription ratio(s)*

    -> More Access Ports

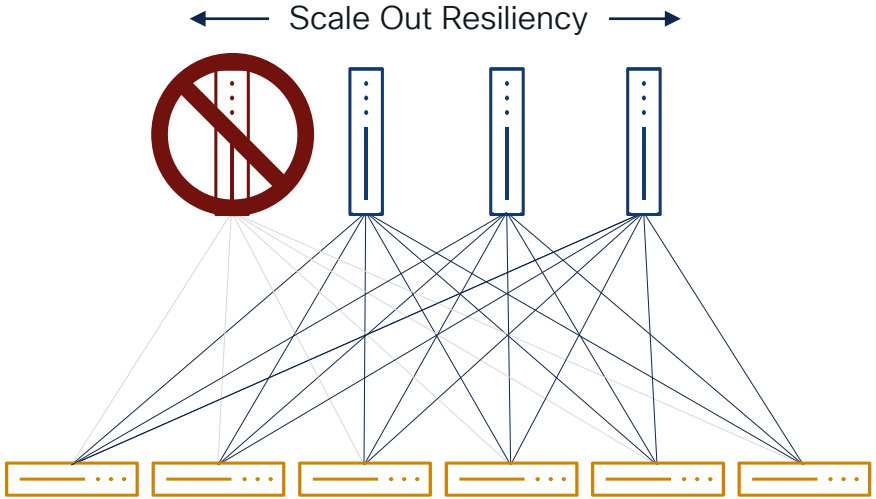
Advantages



Clos Topology

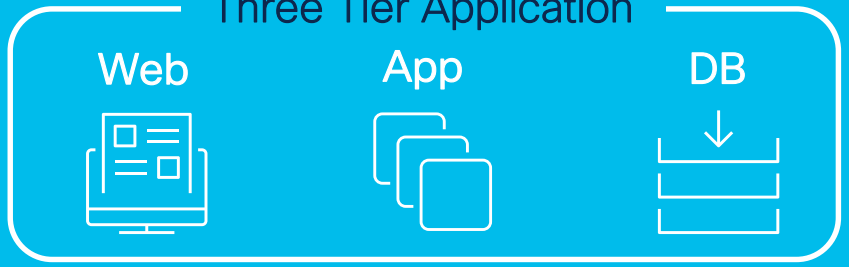
- N+1 redundancy
- **Resiliency increases when adding redundancy** to the topology
- Single spine failure
  - 4 Spine = 25% impact
  - 8 Spine = 12.5% impact
  - 12 Spine = 8.3% impact

Advantages

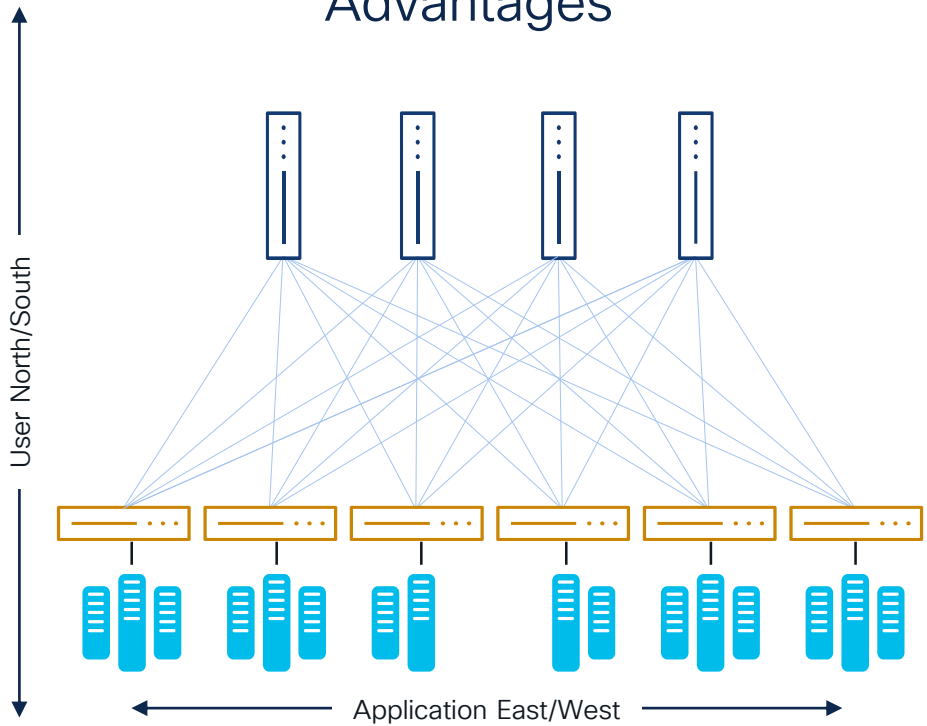


Clos Topology

- Modern Application Needs
  - Every (1) North to South Connection requires eight (8) East to West
  - Application frontend (User access)
  - Application backend, Database, Storage etc...



Advantages

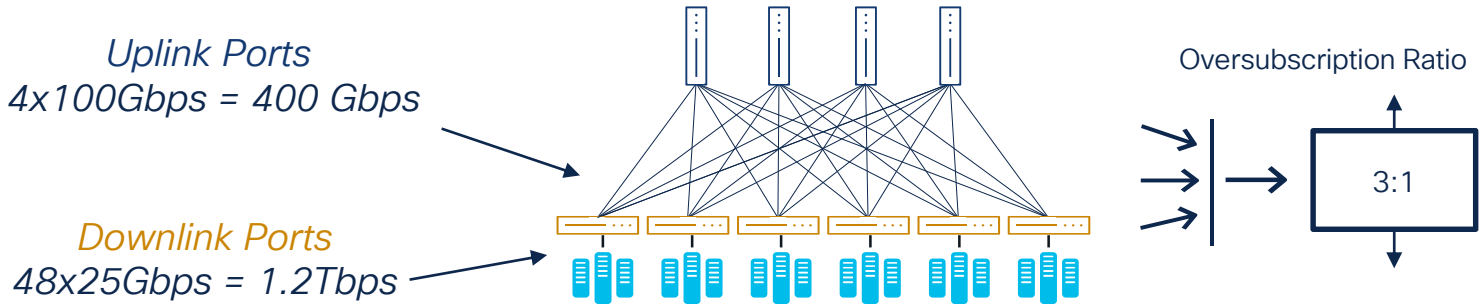


Oversubscription

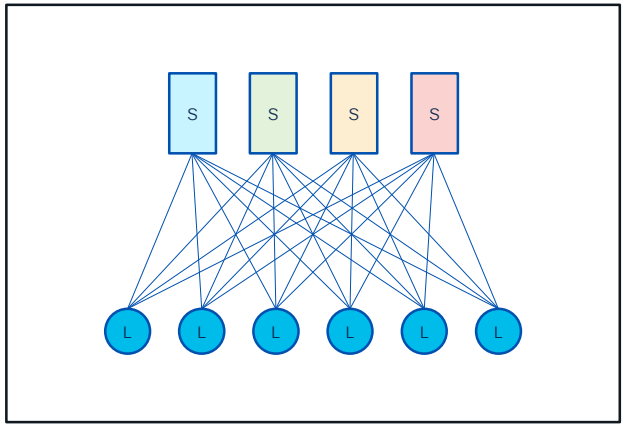
Oversubscription happens when an intermediate network device (switch) or link doesn't have enough capacity (bandwidth) to allow line rate communication between the two devices

Oversubscription Ratio

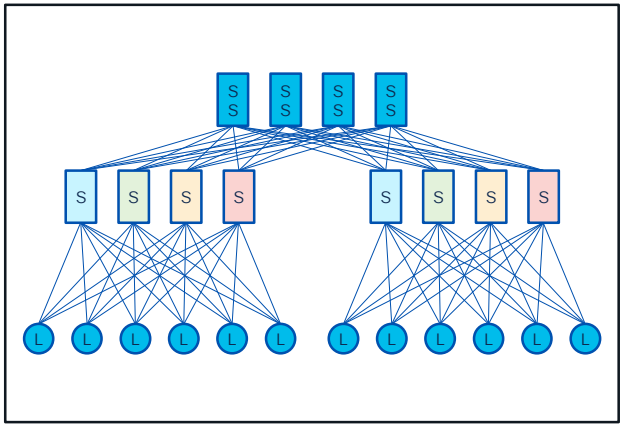
[Number of Downlink Ports * Speed] ÷ [Number of Uplink to Spine * Speed] : 1



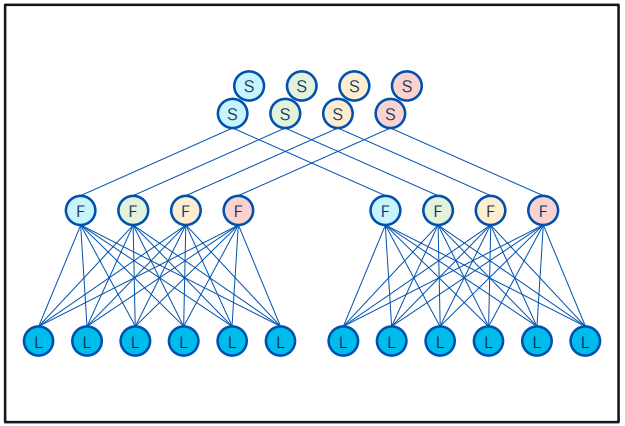
The Journey to Build Better and Further



2 Tier Leaf Spine



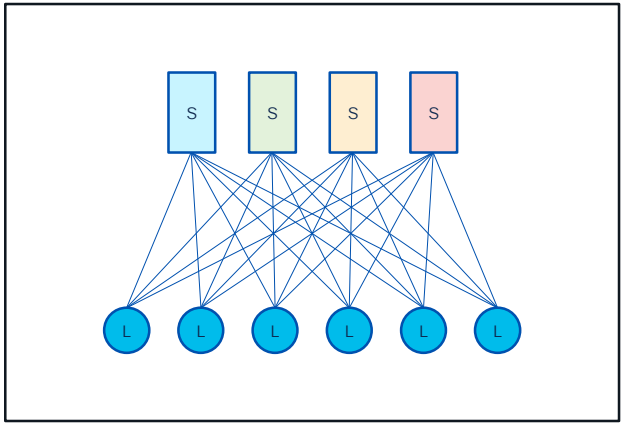
3 Tier Leaf-Spine-SuperSpine



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

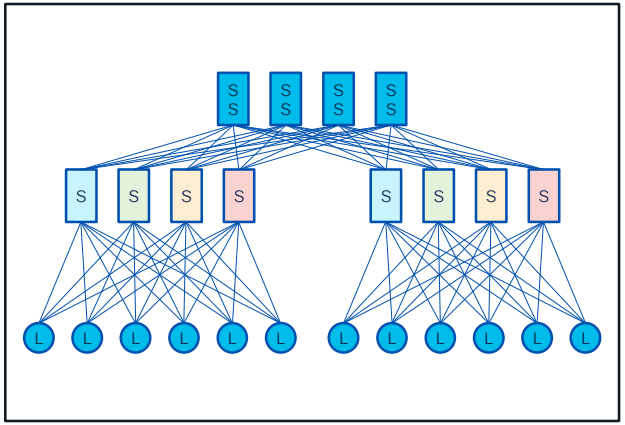


2 Tier Leaf Spine

- 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

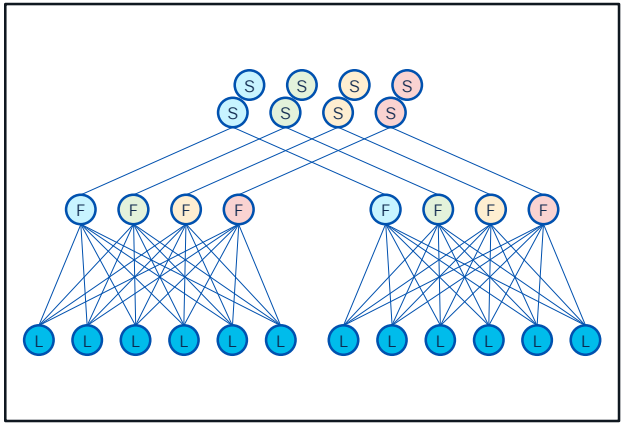


3 Tier Leaf-Spine-SuperSpine

- 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

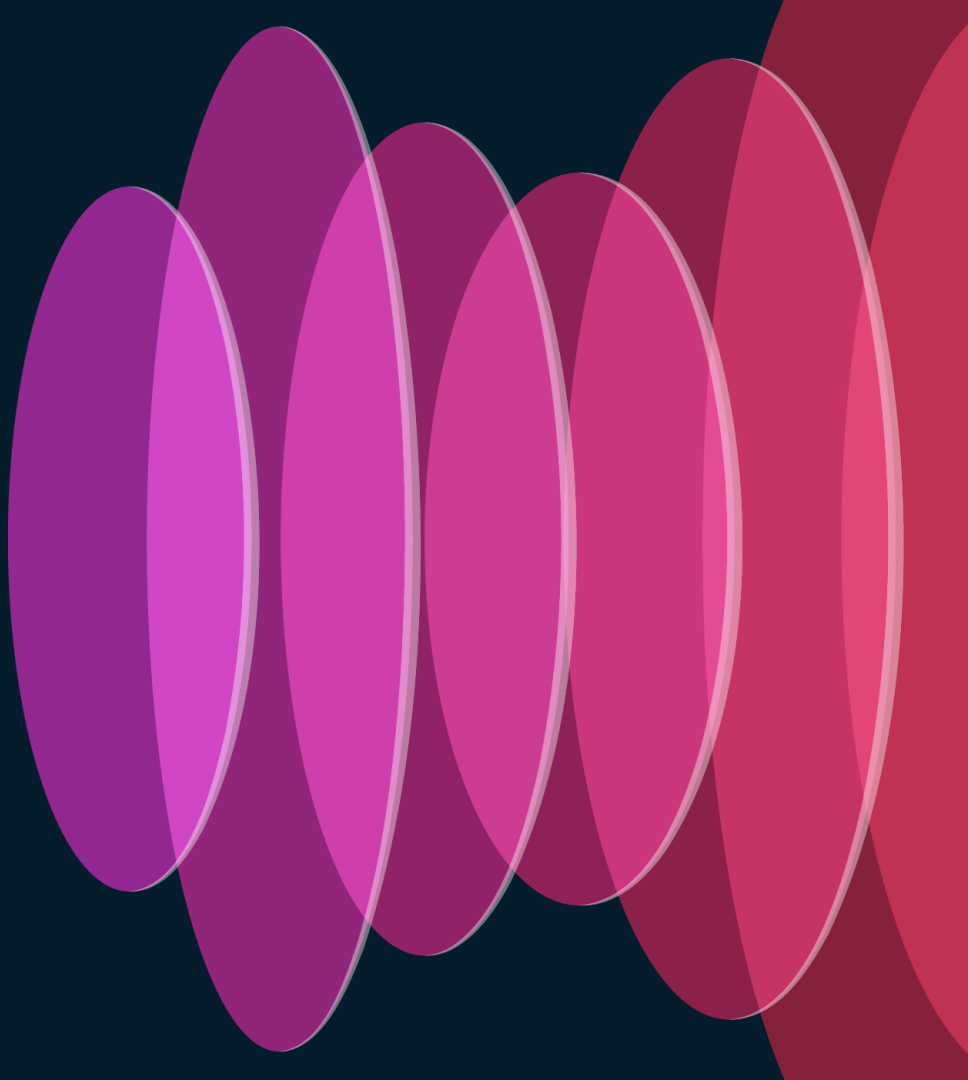
- Increasing Scale-Out in all Tiers
- Reduce to the Max
  - Simple Design Principles
- Increases the “Finite Scale”
  - Scale as You Go
- Disaggregated Redundancy
- Flexible Link and Bandwidth Distribution
- Further Possibility for Cost Optimization



3 Tier Leaf-Fabric-Spine

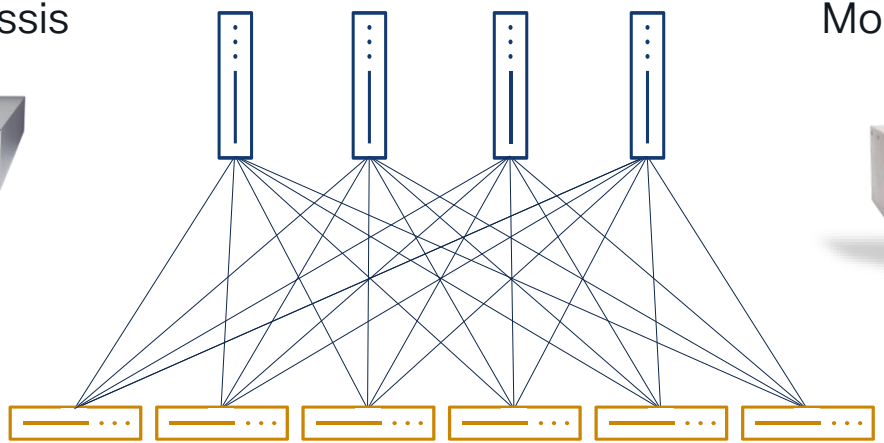
- To Infinity and the Beyond

Hardware Deployment Considerations



Hardware Components

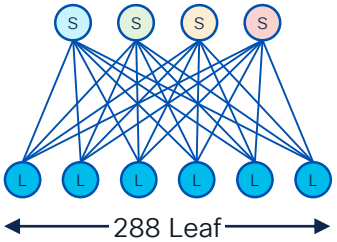
What should I use for **Spines**?
Modular or Fixed Chassis



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

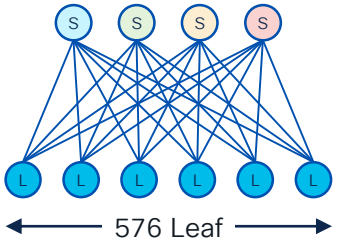
Resiliency	Oversubscription	Cost	Cabling

Attributes to Scale



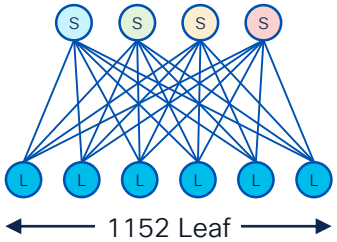
8 Slot Modular
36x 100Gbps
1:1 Oversubscription
13'828 Host Ports

Scale-Up to Fill Chassis



16 Slot Modular
36x 100Gbps
1:1 Oversubscription
27'648 Host Ports

Scale-Up to Bigger Chassis

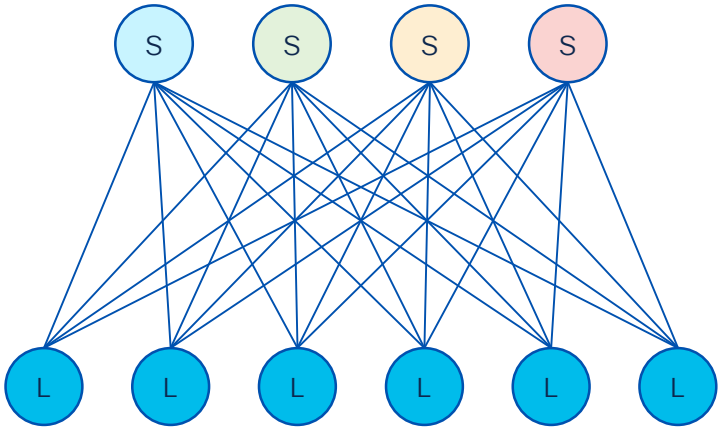


8 Slot Modular
36x 400Gbps
1:1 Oversubscription
55'296 Host Ports

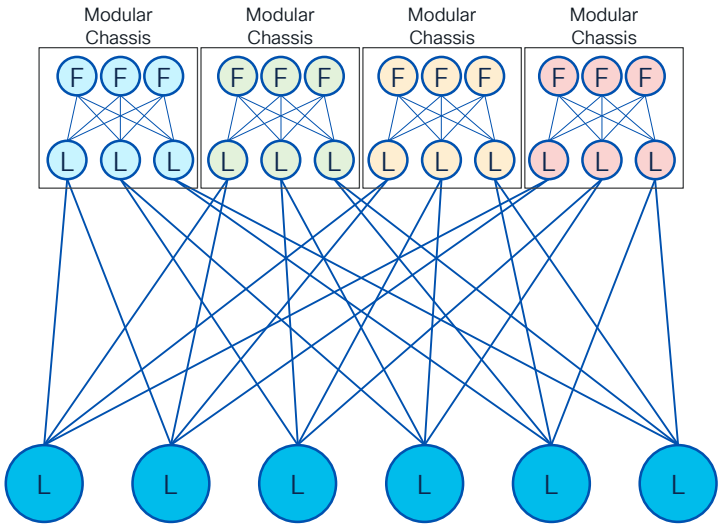
Scale-Up to Faster Linecards

Oversubscription Ratio doesn't influence Host Port scale

Fixed vs Modular – Details that Matter

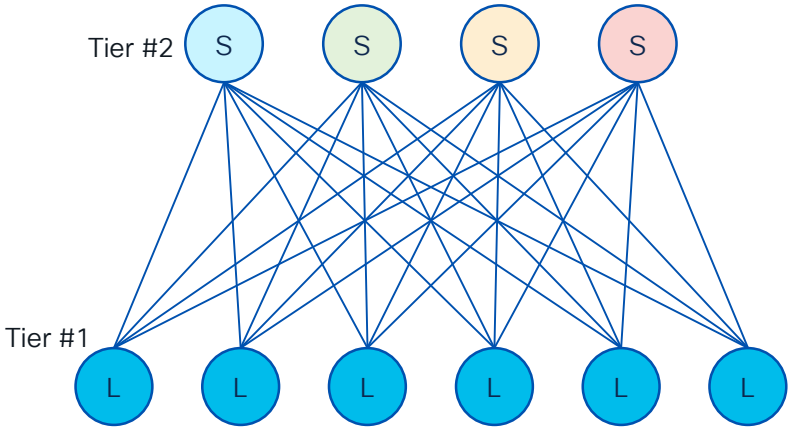


What you think you Built
2 Tier Leaf and Spine Network (3 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)

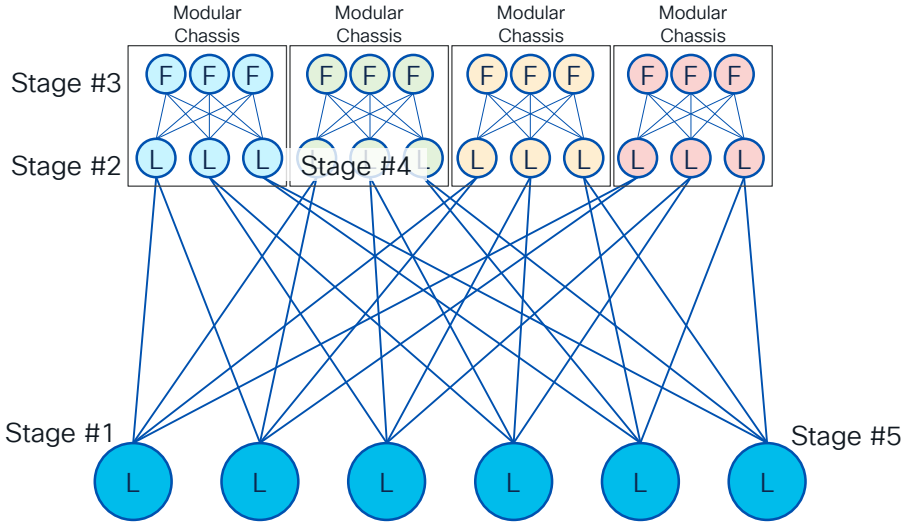


What you really Built
2 Tier Leaf and Spine Network (5 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)

Fixed vs Modular – Untangling the Details



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)

Hardware Components - Spines

Modular

Nexus 9408

Optimized for
100G/200G/400G

Nexus 9500

Nexus 9716D-GX
16p 400G

Nexus 9736C-FX
36p 100G

Nexus 9800

Nexus 9836DM-A
36p 400G **(w/ HBM)**

Nexus 98900CD-A
14p 400G & 24p 100G **(w/ HBM)**

Fixed

Nexus 9364D-GX2A
64p 400G

Nexus 9332D-H2R
32p 400G **(w/ HBM)**

Nexus 9348-GX2A
48p 400G

Nexus 93600CD-GX
28p 40/100G & 8p 400G

Resiliency | Oversubscription | Cost | Cabling

Hardware Components - Leafs

Modular

Nexus 9408

Optimized for
100G/200G/400G

Nexus 9500

Nexus 9716D-GX
16p 400G

Nexus 9736C-FX
36p 100G

Nexus 9800

Nexus 9836DM-A
36p 400G **(w/ HBM)**

Nexus 98900CD-A
14p 400G & 24p 100G **(w/ HBM)**

Fixed

Nexus 93240YC-FX2
48p 10/25G & 12p 40/100G

Nexus 93400LD-H1
48p 10/25/50G & 4p 100/400G

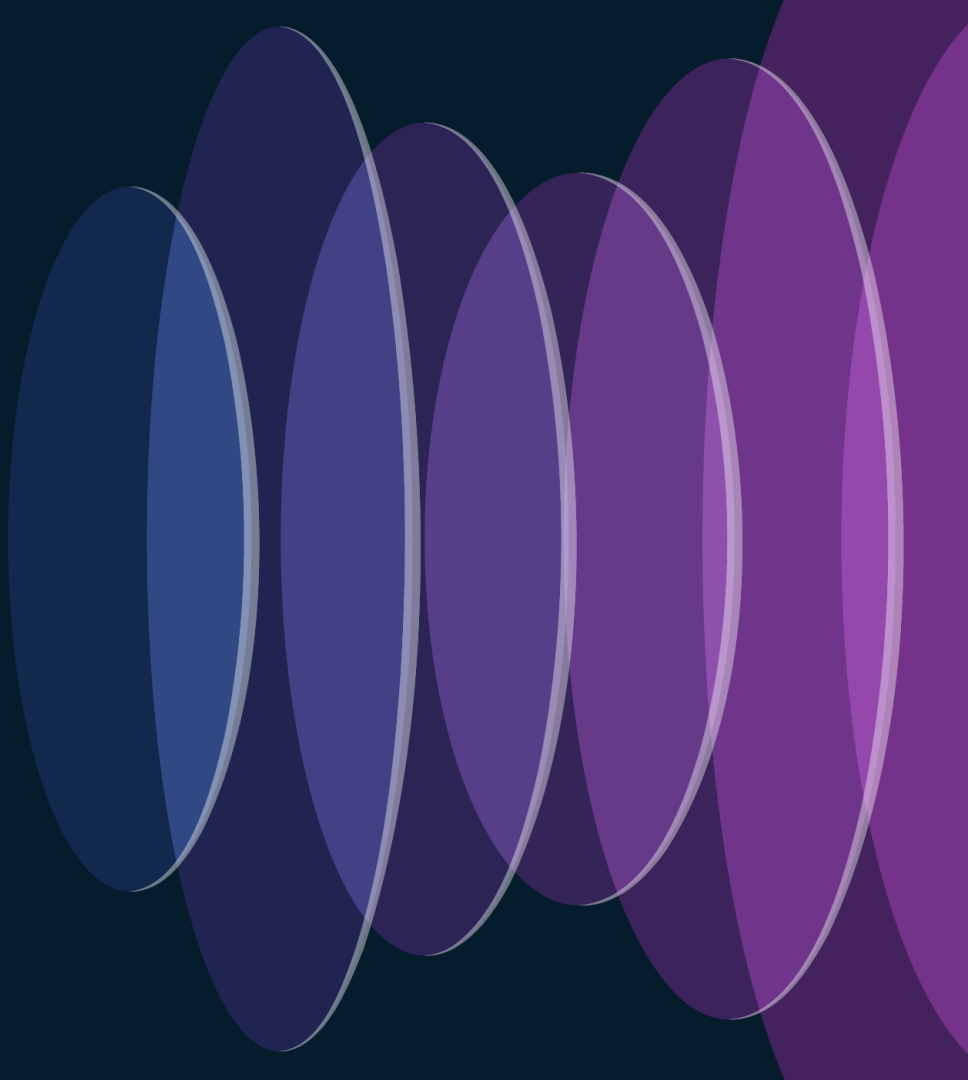
Nexus 93180YC-FX3
48p 10/25G & 6p 40/100G

Nexus 93360YC-FX2
96p 1/10/25G & 12p 40/100G

Nexus 93600CD-GX
28p 40/100G & 8p 400G

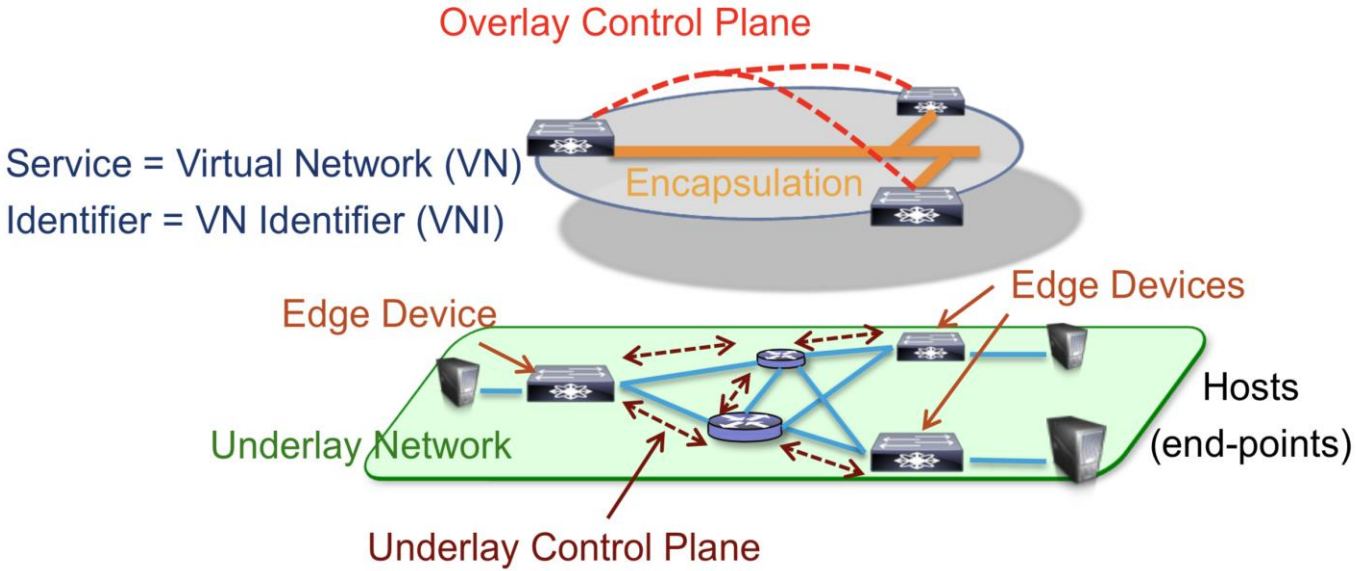
Resiliency | Oversubscription | Cost | Cabling

Fundamentals of VXLAN EVPN Design



cisco Live!

Underlay vs Overlay



Underlay is responsible for tunnel endpoint reachability while the management of virtual tunnels is handled by the overlay.

Network Overlay Services

L2 OVERLAYS

- MPLS L2 VPNs i.e. AToM, VPLS, PBB-EVPN
- Overlay Transport Virtualization OTV
- VXLAN Flood and Learn.
- VXLAN BGP EVPN (hybrid)
- L2TPv3
- Fabric Path/TRILL (MAC in MAC)
- ACI iVXLAN (hybrid)

L3 OVERLAYS

- MPLS L3 VPNs
- GRE
- LISP
- VXLAN BGP EVPN (hybrid)
- ACI iVXLAN (hybrid)

VXLAN BGP EVPN Provides Integrated Routing and Bridging (IRB) Fabric, best of L2 and L3 overlays with single overlay service.

VXLAN EVPN – What Is It?

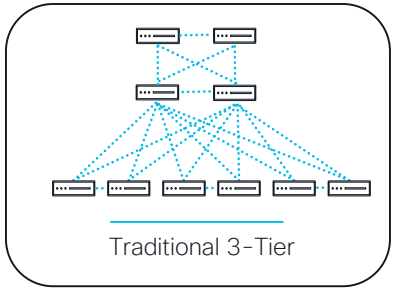
VXLAN

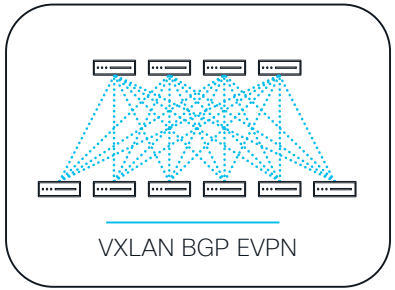
- Standards based Encapsulation
- RFC 7348
- Uses UDP-Encapsulation
- Transport Independent
- Layer-3 Transport (Underlay)
- Flexible Namespace
- 24-bit field (VNID) provides ~16M unique identifier
- Allows Segmentation

EVPN

- Standards based Control-Plane
- RFC 8365 (and 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)

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
- ✓ Over the top wireless

- ✓ Any workload anywhere across L3 boundaries
- ✓ Seamless VM Mobility (intra and inter-DC)
- ✓ Integrated wireless for seamless User Mobility
- ✓ Leverages ECMP for optimal path over L3 network
- ✓ Next Gen Data Center Interconnect (DCI)

VXLAN Benefits

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 Topology

- Typical Design used is Leaf/Spine Topology (CLOS based)
- Layer 3 Links between Leaf and Spines
- Unicast Packets are encapsulated within Unicast VXLAN Tunnels
- Broadcast Unknown Unicast and Multicast (BUM) traffic replication by Multicast or Ingress Replication (IR)

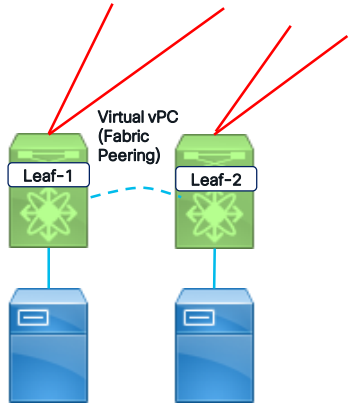
Network Components of VXLAN Overlays

- **VXLAN Segment**
  - VXLAN overlay network. Layer 2 Broadcast Domain.
- **VXLAN Network Identifier (VNID)**
  - Each VXLAN segment is identified by a 24-bit VNID.
- **VXLAN Tunnel Endpoint (VTEP)**
  - Tunnel Endpoint. RFC term Network Virtualization Edge.
  - Each VTEP is uniquely identified by an IP address.
  - VTEP switch when forwarding packets within the same VNID and route for inter-VNI traffic.

Components of VXLAN EVPN

Functions of Leaf

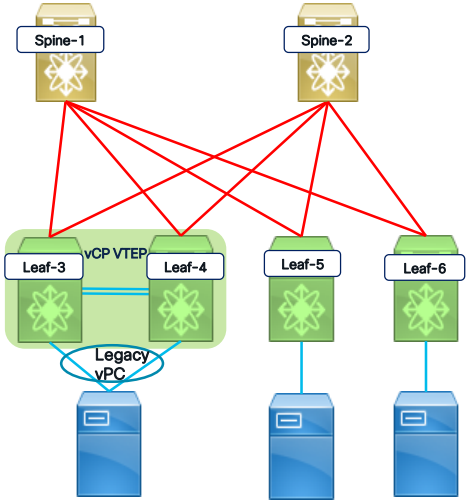
- Forms Routing Protocol adjacencies for underlay with Spines (OSPF, IS-IS, BGP)
- MP-BGP L2VPN EVPN neighborships with spines to exchange routes
- Performs VXLAN encapsulation and decapsulation
- Default Gateway Services for hosts using Distributed Anycast Gateway
- BUM replication/processing
- Connect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



Components of VXLAN EVPN

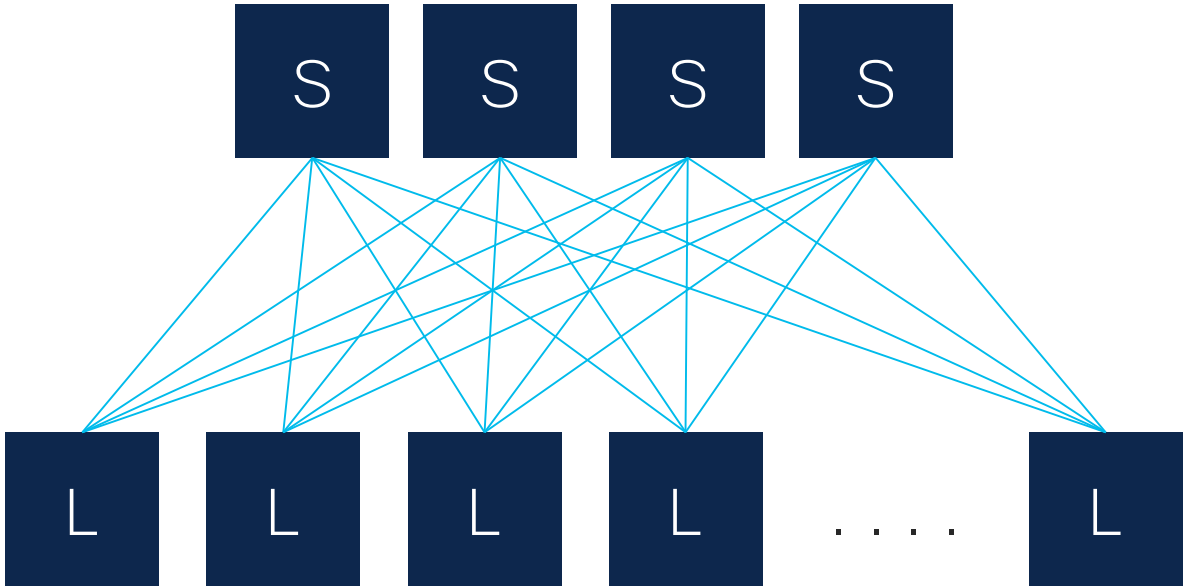
Functions of Spine

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



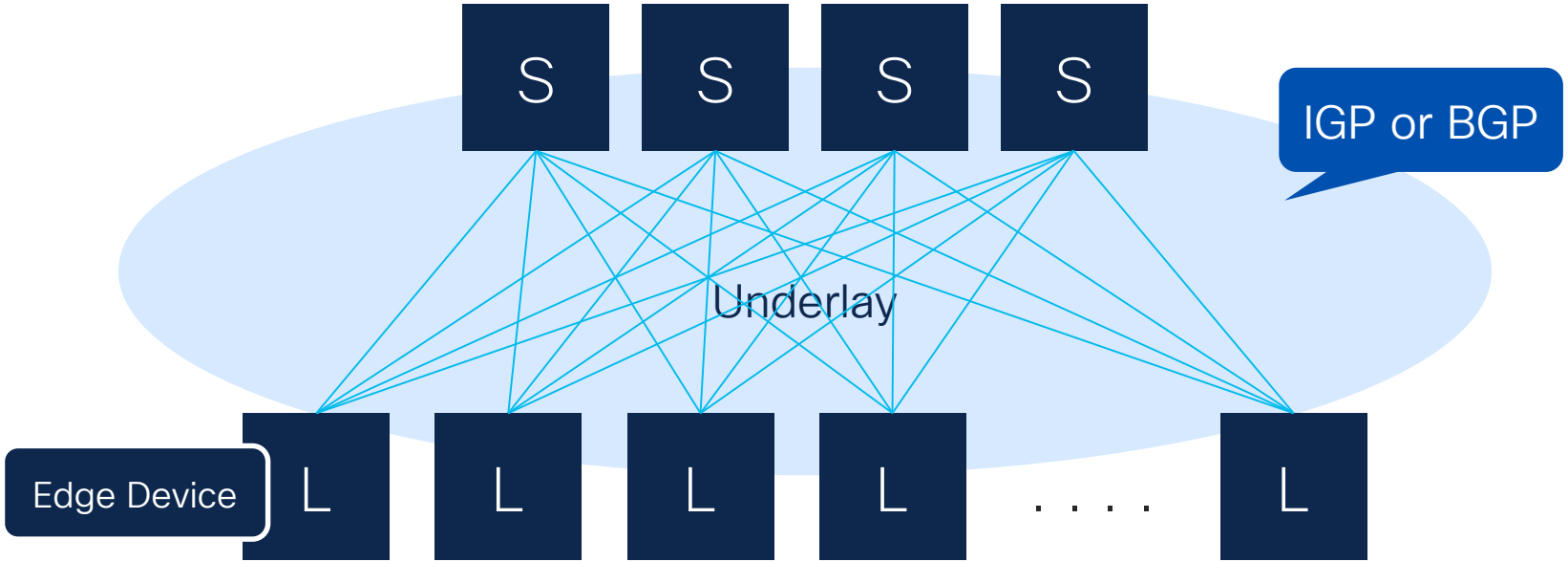
VXLAN: The Building Blocks

Underlay



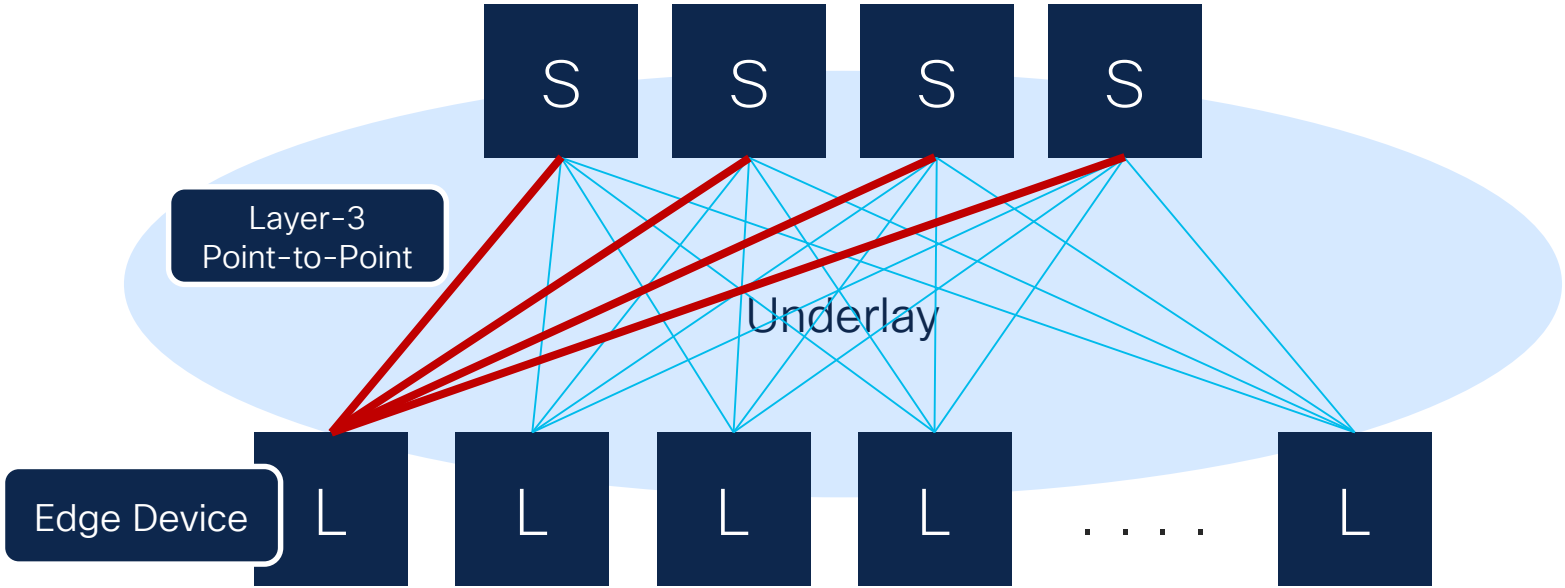
VXLAN: The Building Blocks

Underlay



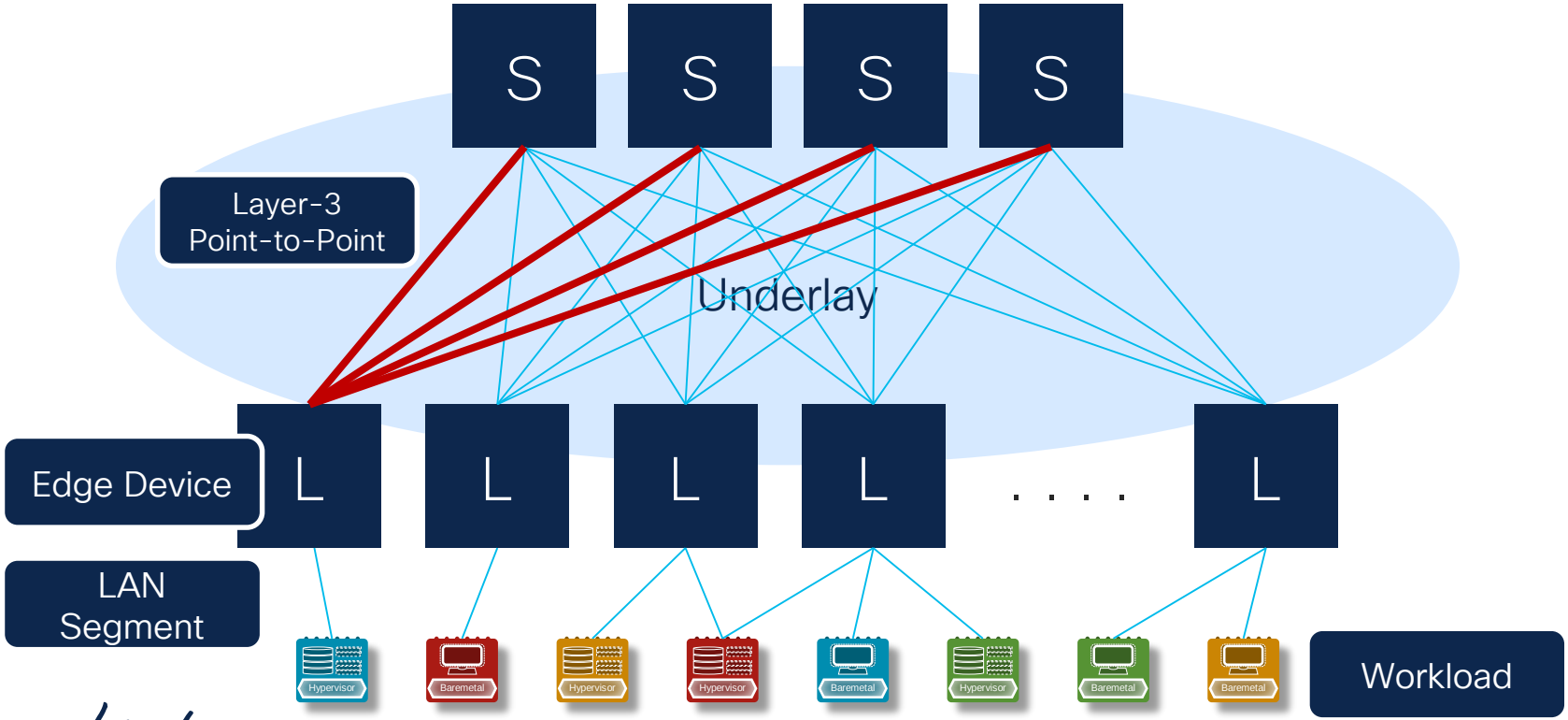
VXLAN: The Building Blocks

Underlay



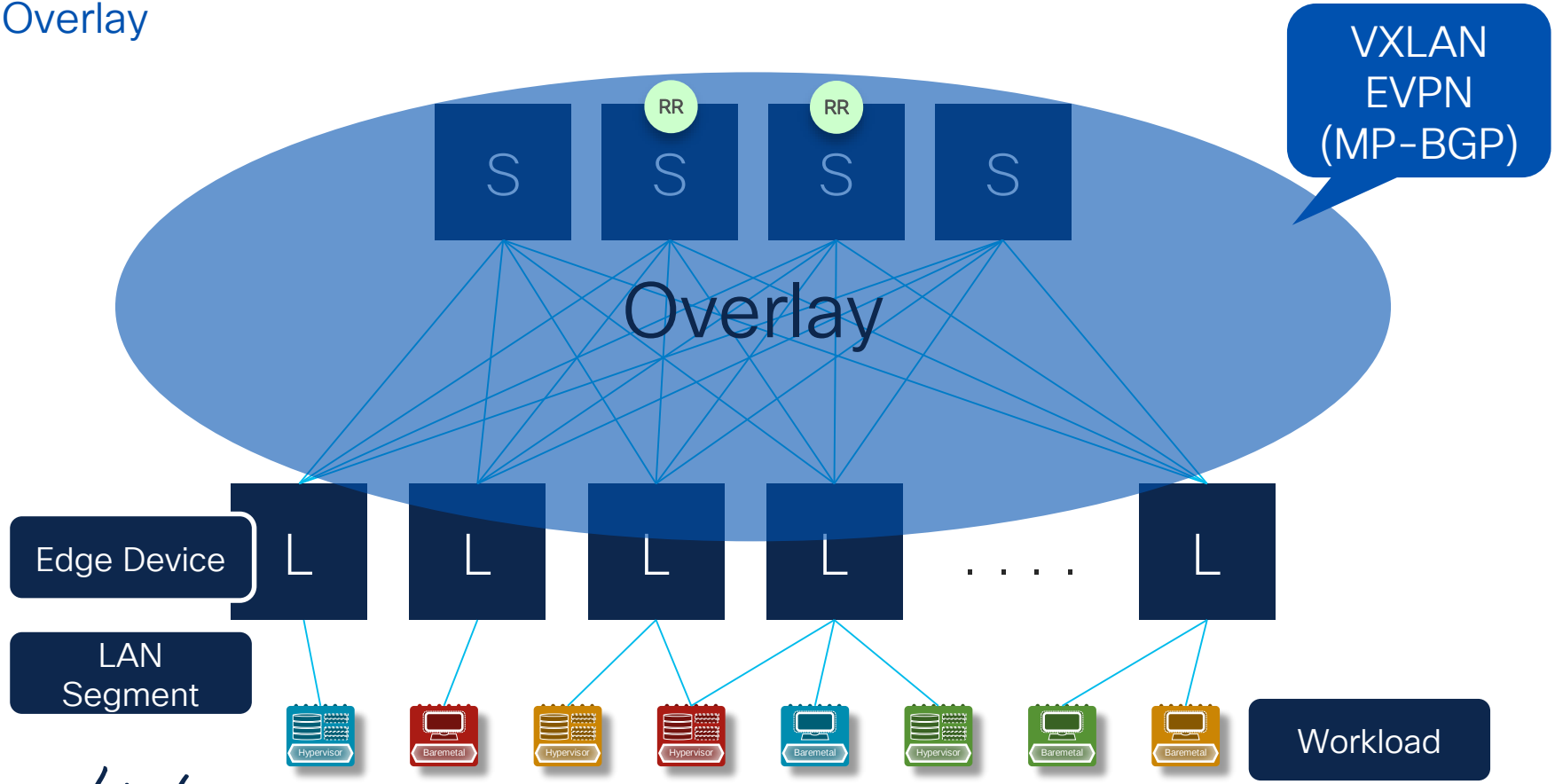
VXLAN: The Building Blocks

Underlay



VXLAN: The Building Blocks

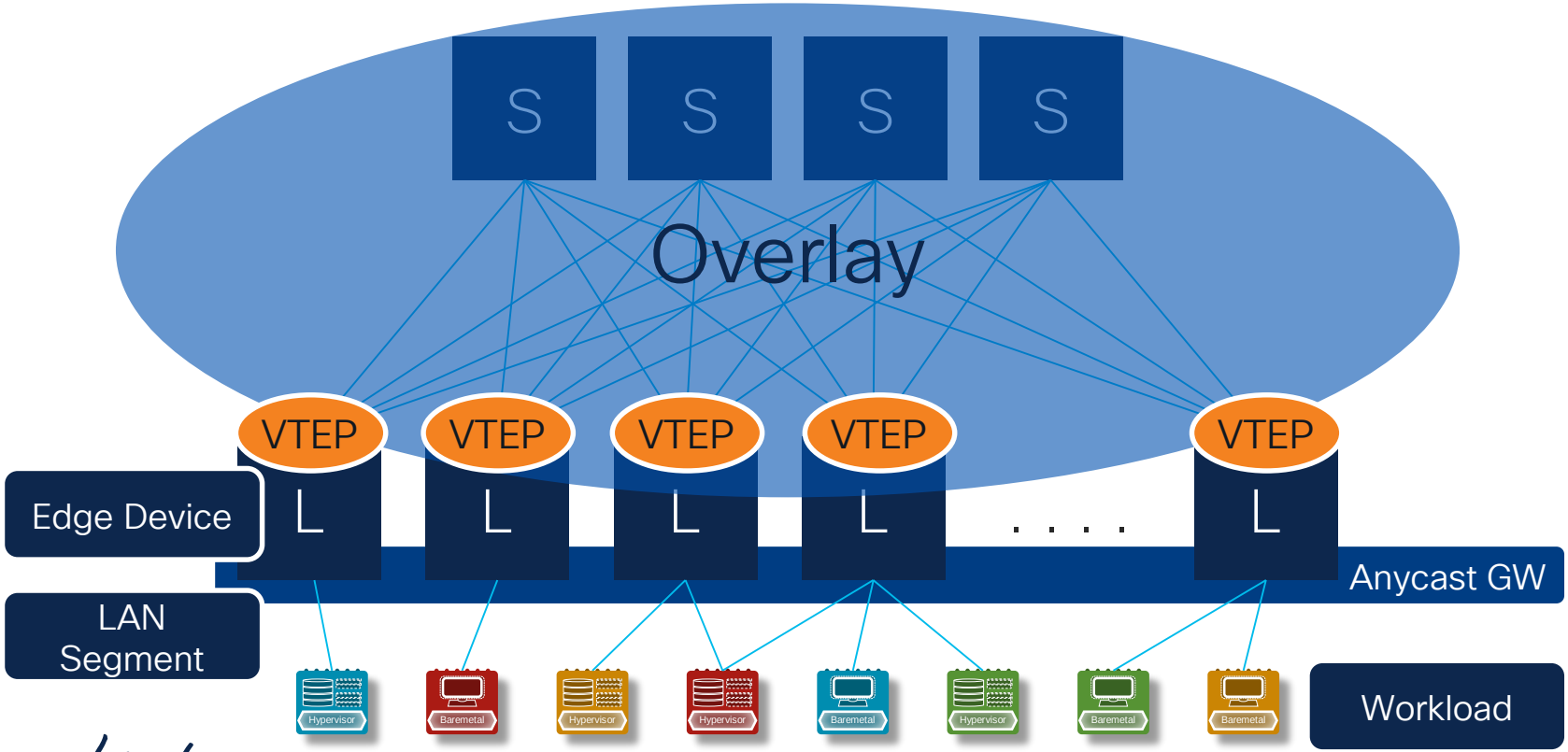
Overlay



VXLAN: The Building Blocks

Overlay

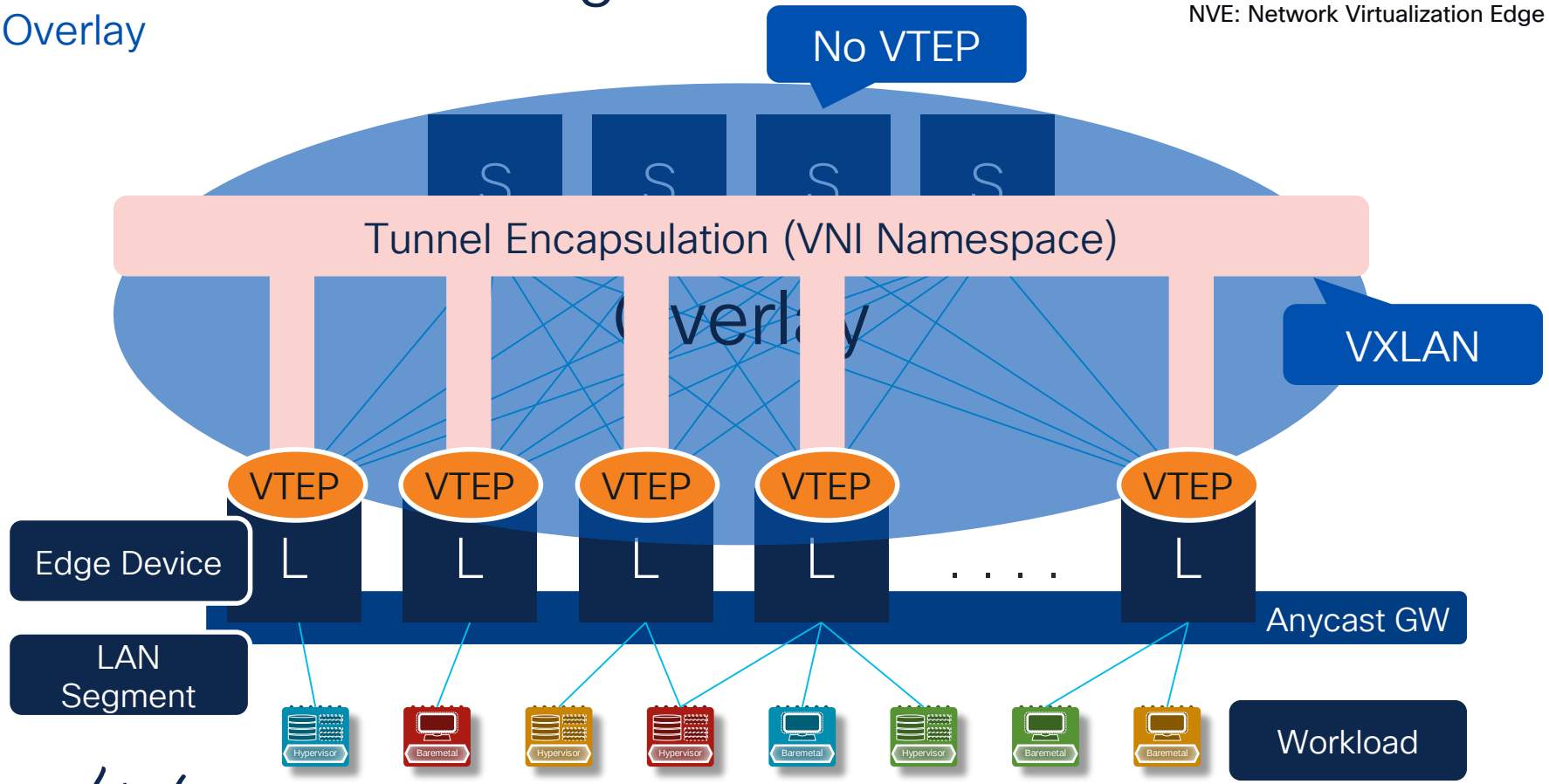
VTEP: VXLAN Tunnel End-Point
VNI/VNID: VXLAN Network Identifier
NVE: Network Virtualization Edge



VXLAN: The Building Blocks

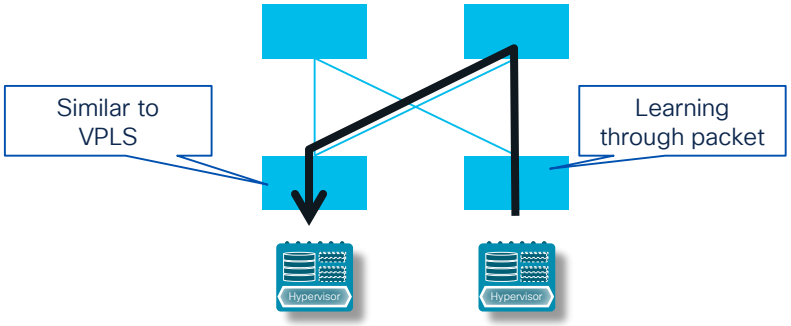
Overlay

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



What is Ethernet VPN?

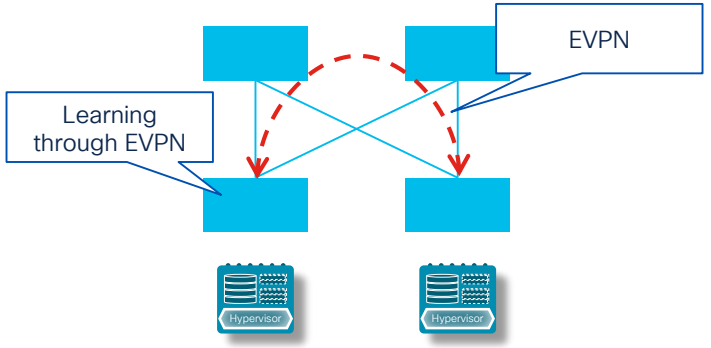
VXLAN only
No Control plane



- L2+L3 multipath
- Huge address space

+And some risks

VXLAN with EVPN



- BUM optimization
- L3 optimization
- Multi homing
- Peer auto discovery
- Peer authentication
- Host mobility etc

EVPN can bring intelligence

BGP EVPN Overview

- MP-BGP EVPN AF carries following information: MAC, IP and network prefix, VRF/VNID and VTEP IP (NLRI Next Hop).
- BGP EVPN distributes MAC,IP info avoiding flooding.
- VXLAN BGP AFI=25 (Layer 2 VPN) and SAFI = 70 (EVPN).
- VXLAN is the Tunnel Encapsulation Protocol and MP-BGP EVPN is the Control Plane for overlay distributing Layer 2 and Layer 3 routing information (MAC,IP).

***NLRI**: Network Layer Reachability Information (NLRI) is exchanged between BGP peers, indicating how to reach prefixes.*

***AFI and SAFI**: AFI means Address Family Indicator and SAFI is the Subsequent Address Family Indicator. They are used in the Multiprotocol Extensions to BGP and are exchanged during neighbor capability exchange during the process for loading the peers.*

MP-BGP EVPN Advertisements

EVPN Prefix Types

- BGP EVPN uses 5 different route types for IP prefixes and advertisement
  - Type 1 - Ethernet Auto-Discovery (A-D) route
  - **Type 2 - MAC advertisement route → L2 VNI MAC/MAC-IP**
  - Type 3 - Inclusive Multicast Route → EVPN IR, Peer Discovery
  - Type 4 - Ethernet Segment Route
  - **Type 5 - IP Prefix Route → L3 VNI Route**

- Route type 2 or MAC Advertisement route is for MAC and ARP resolution advertisement, MAC or MAC-IP
- Route type 5 or IP Prefix route will be used for the advertisement of prefixes, IP only

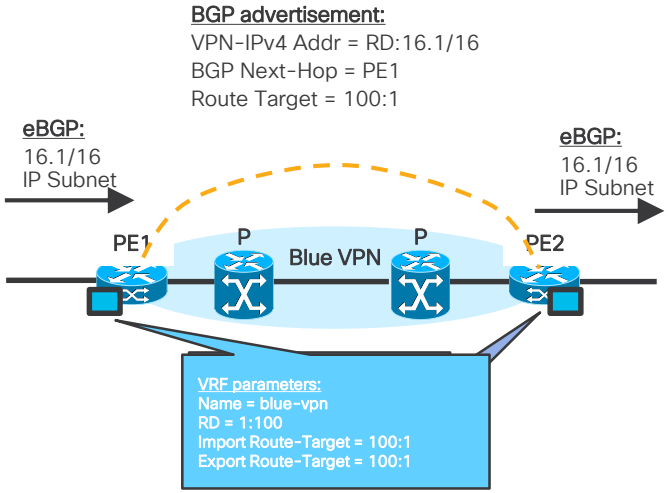
BGP EVPN Address Family

Virtual Routing and Forwarding (VRF)
Layer-3 segmentation for tenants' routing space

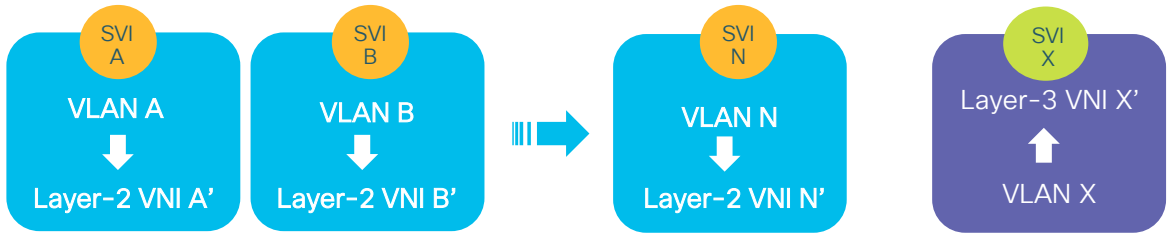
Route Distinguisher (RD):
8-byte field, VRF parameters; unique value to make VPN IP routes unique: RD + VPN IP prefix

Route Target (RT): 8-byte field, VRF parameter, unique value to define the import/export rules for VPNv4 routes

VPN Address-Family:
Distribute the MP-BGP VPN routes



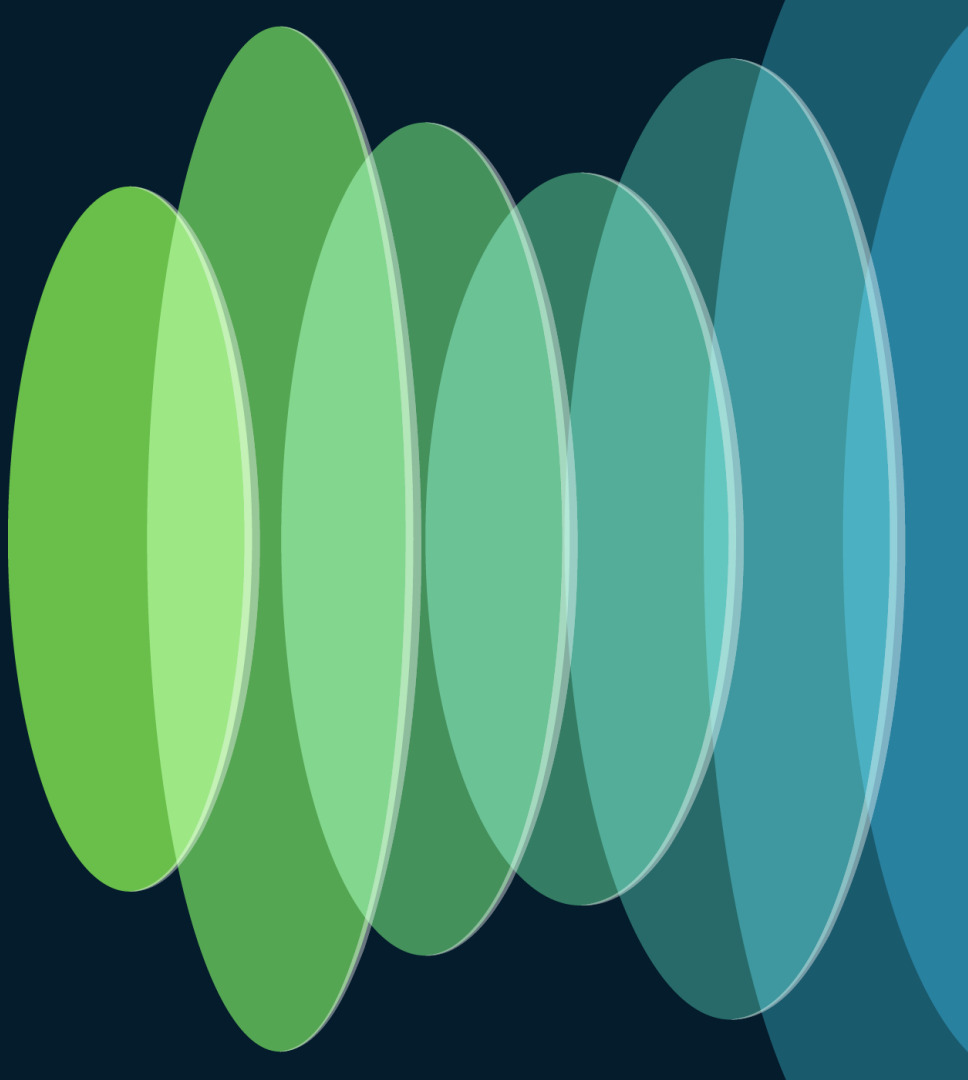
Logical Construct of Multi-Tenant VXLAN EVPN



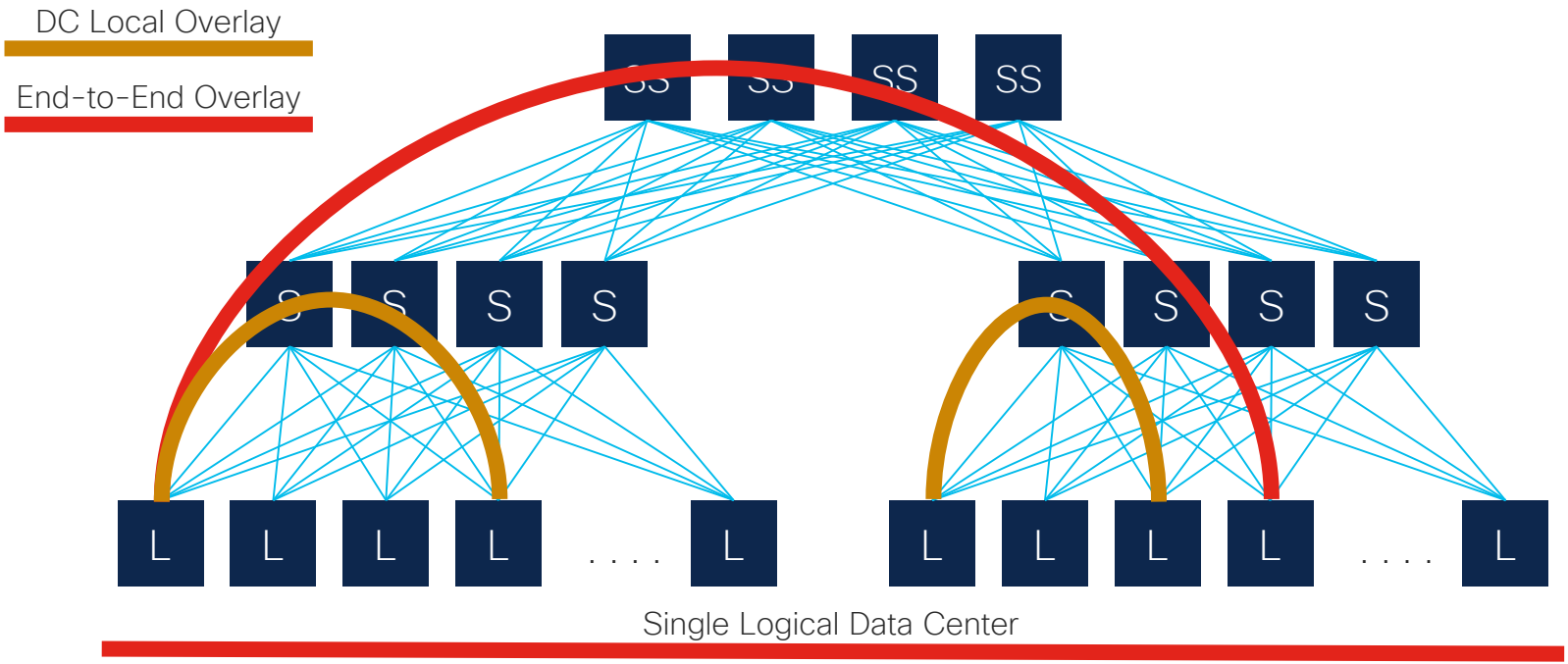
- One VLAN maps to one Layer-2 VNI per Layer-2 segment
- A Tenant can have multiple VLANs, therefore multiple Layer-2 VNIs
- Traffic within one Layer-2 VNI is bridged
- Traffic between Layer-2 VNI's is routed

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

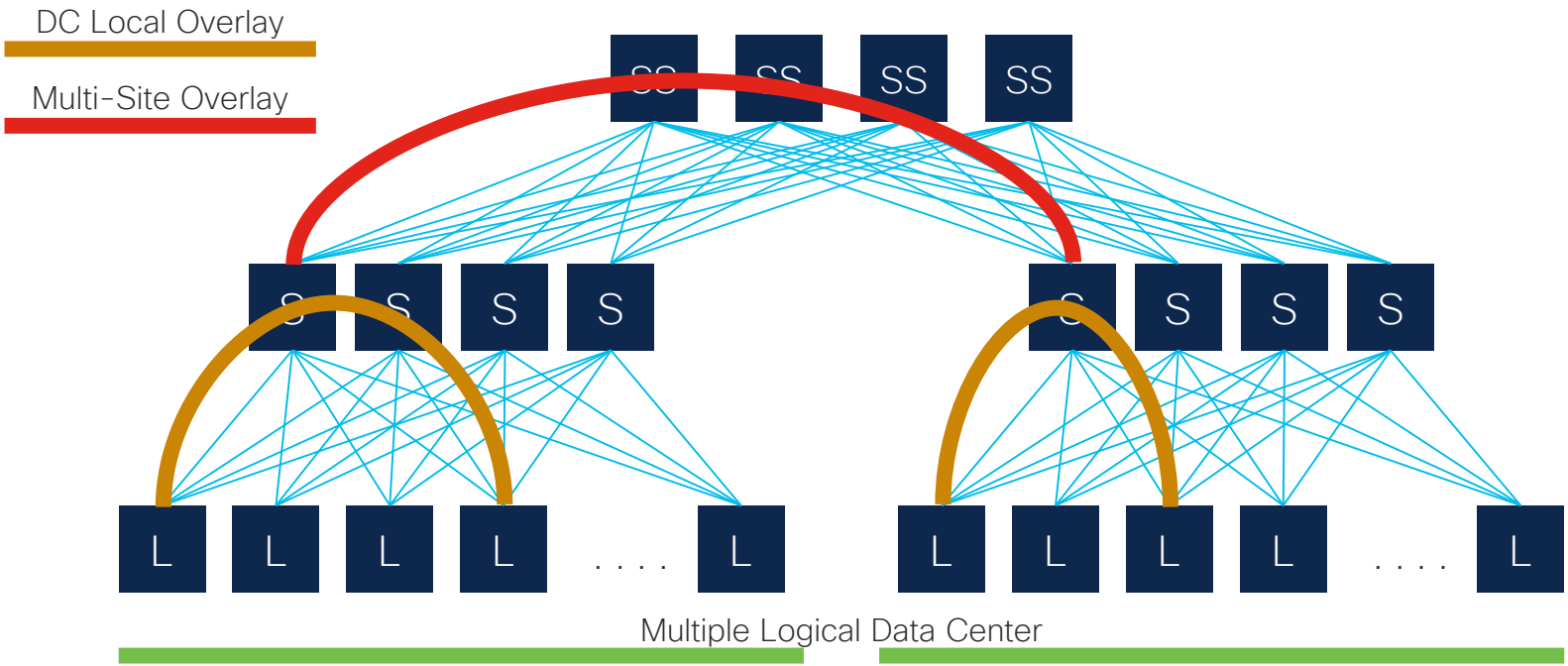
Multi Fabric Designs – VXLAN Multi Site



Multi-Pod: Overlay Spread and Extend

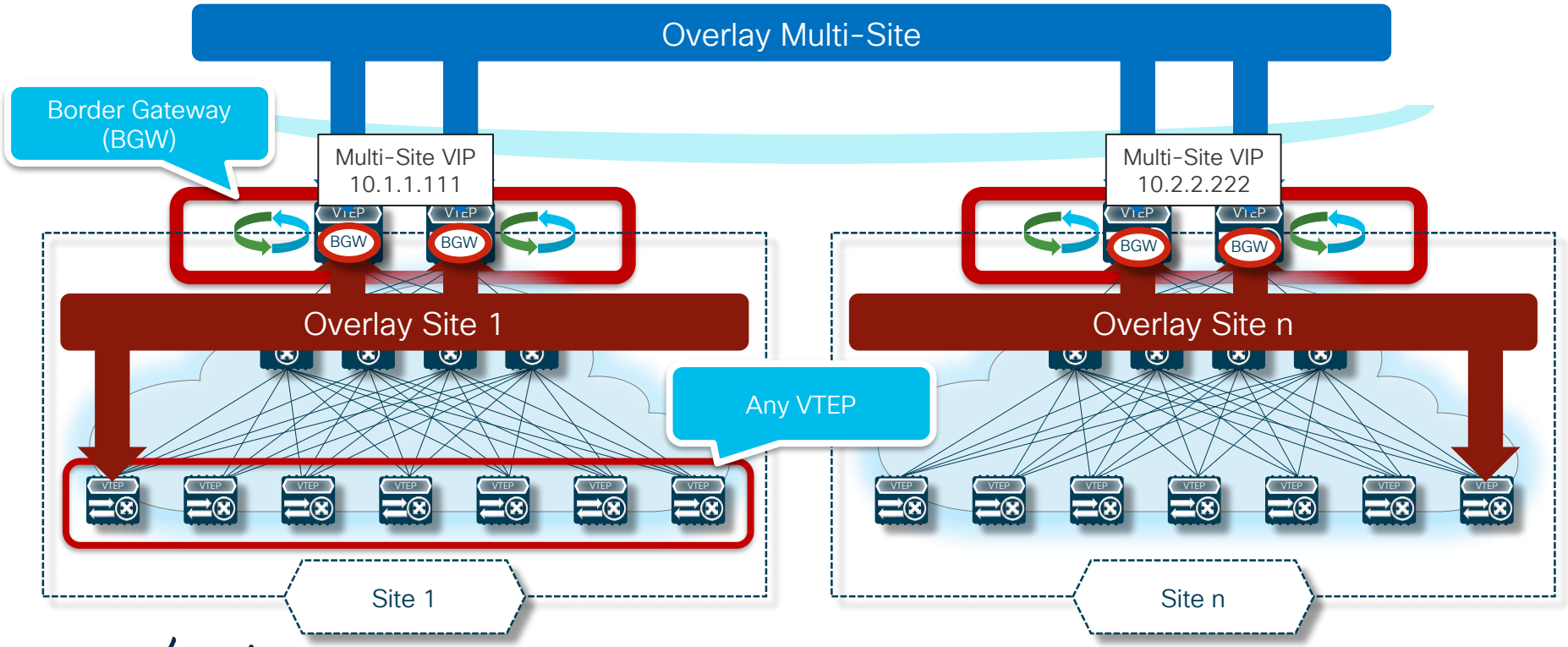


Multi-Site: A Change to Hierarchical Overlays

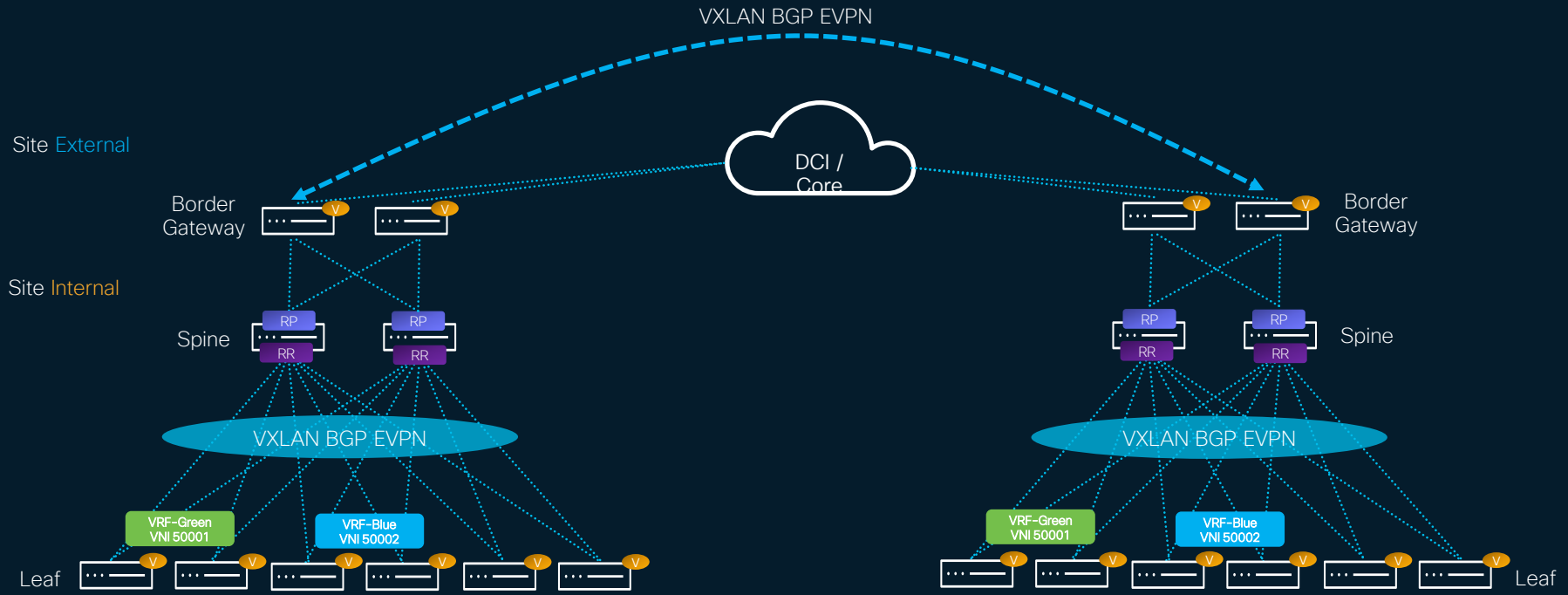


VXLAN: The Building Blocks

Hierarchical VXLAN



Connecting Multiple VXLAN EVPN Sites



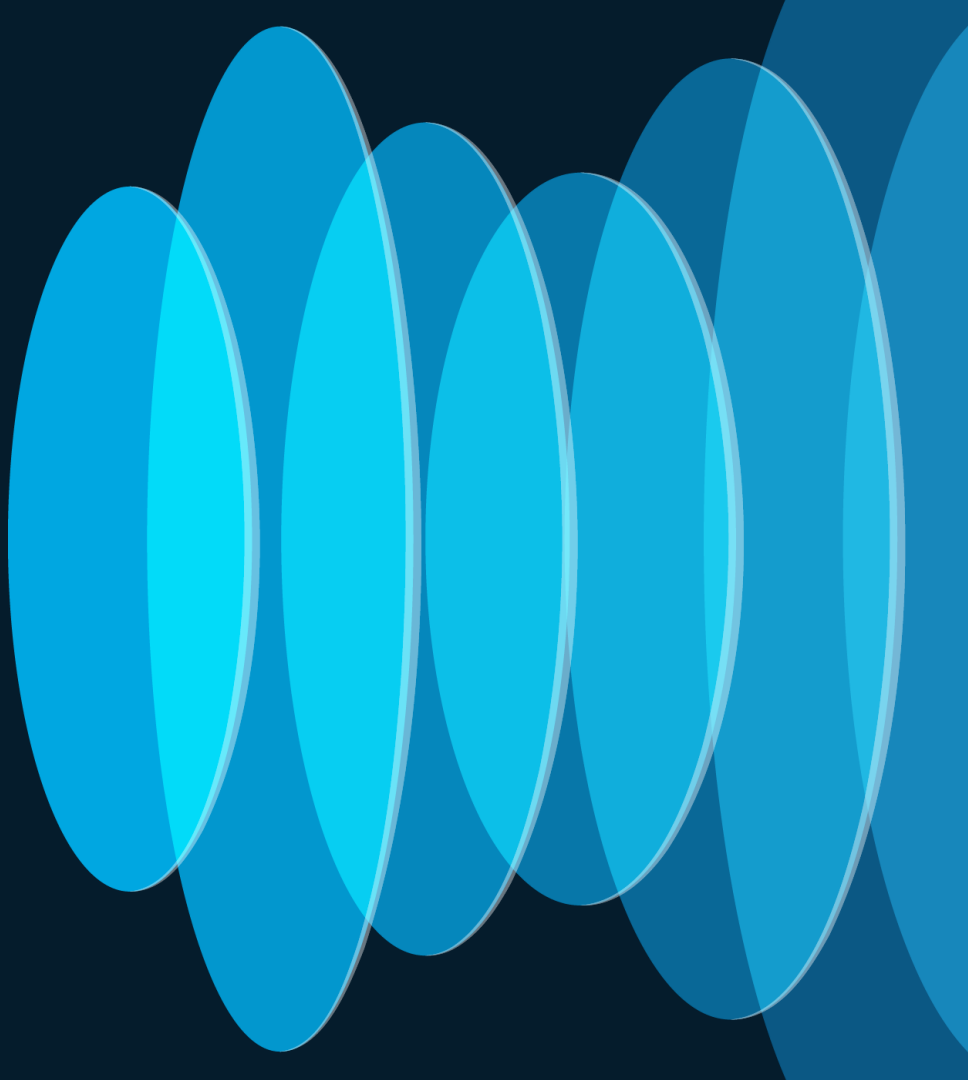
VXLAN Multi-Site Characteristics

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

Underlay Isolation – Overlay Hierarchies

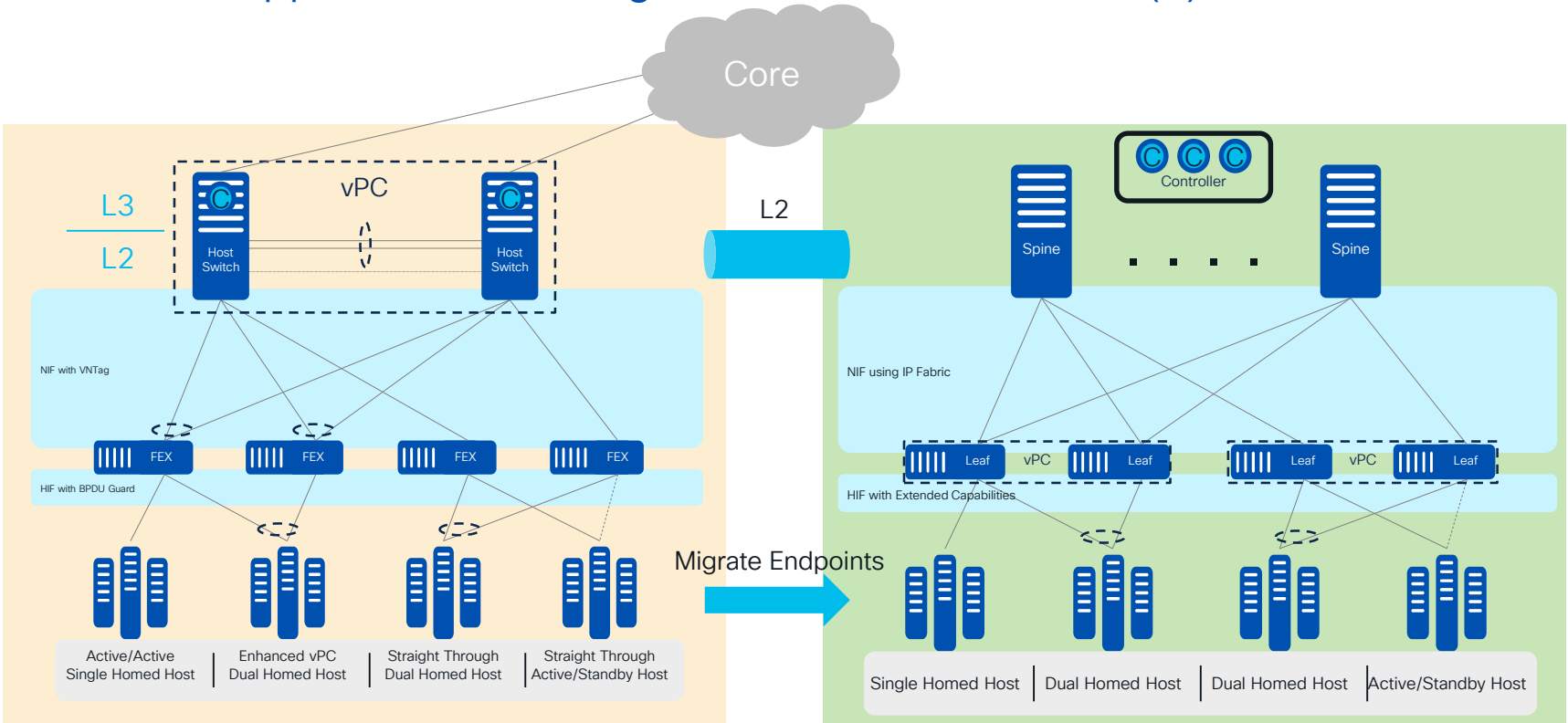
More information available at the VXLAN Multi-Site White paper page:
https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-739942.html

Migration Considerations



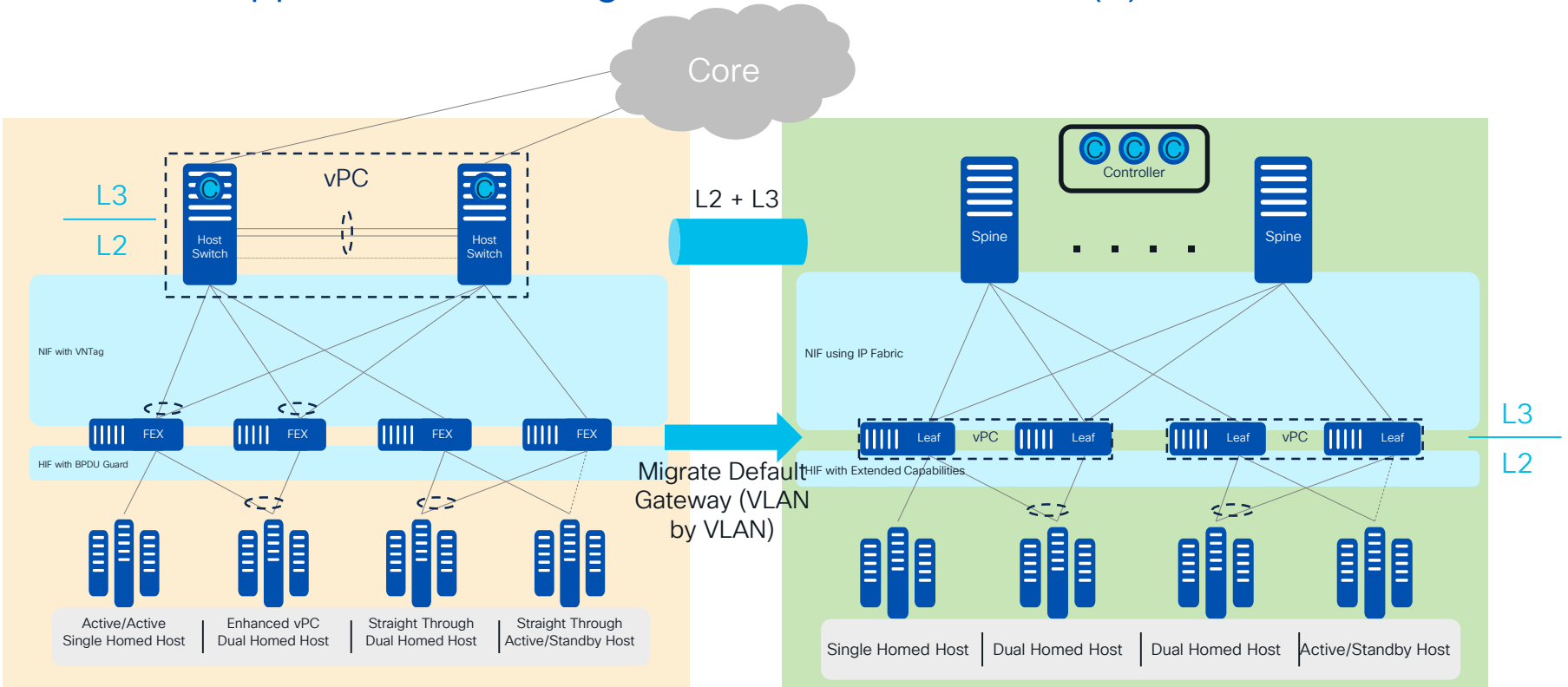
Migration Considerations

The Usual Approach of Building a New Parallel Network (1)



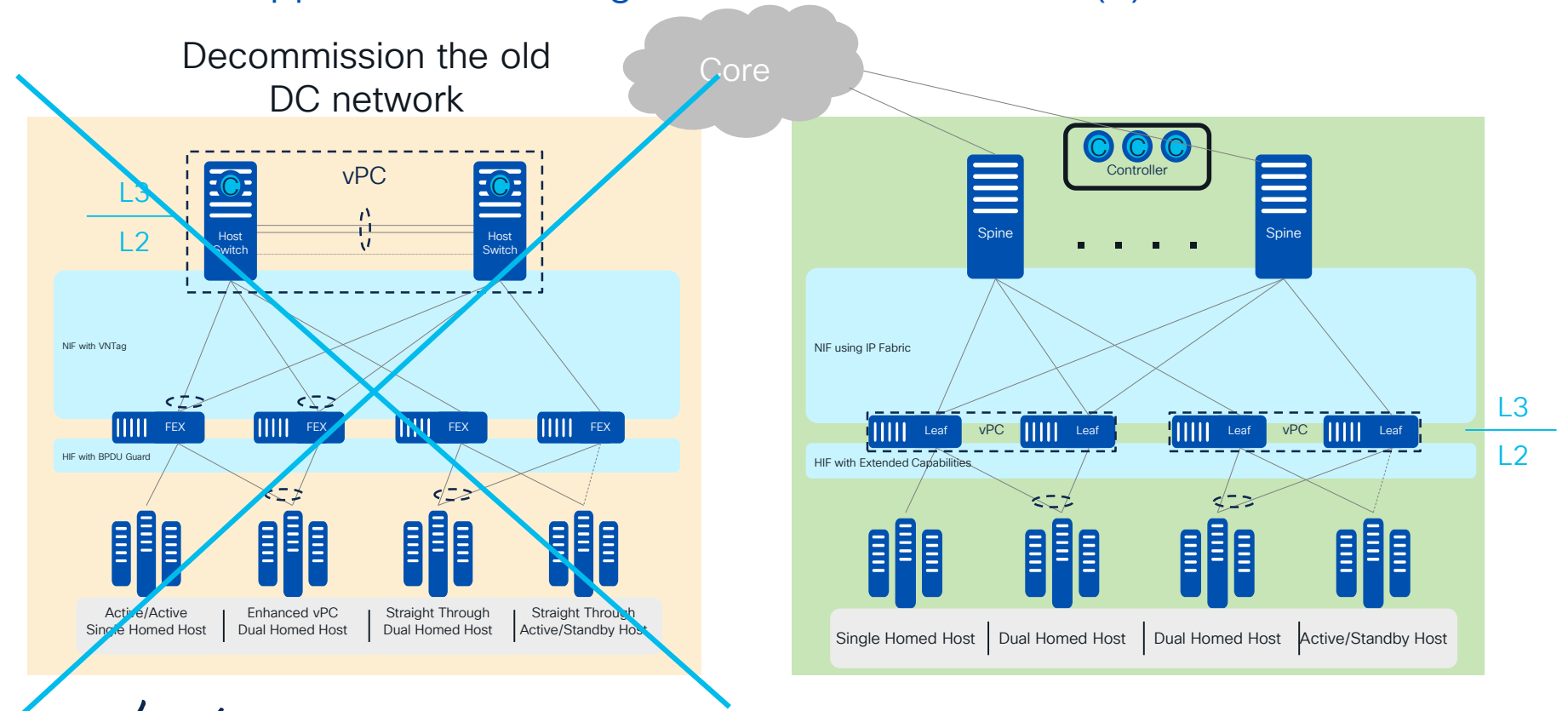
Migration Considerations

The Usual Approach of Building a New Parallel Network (2)

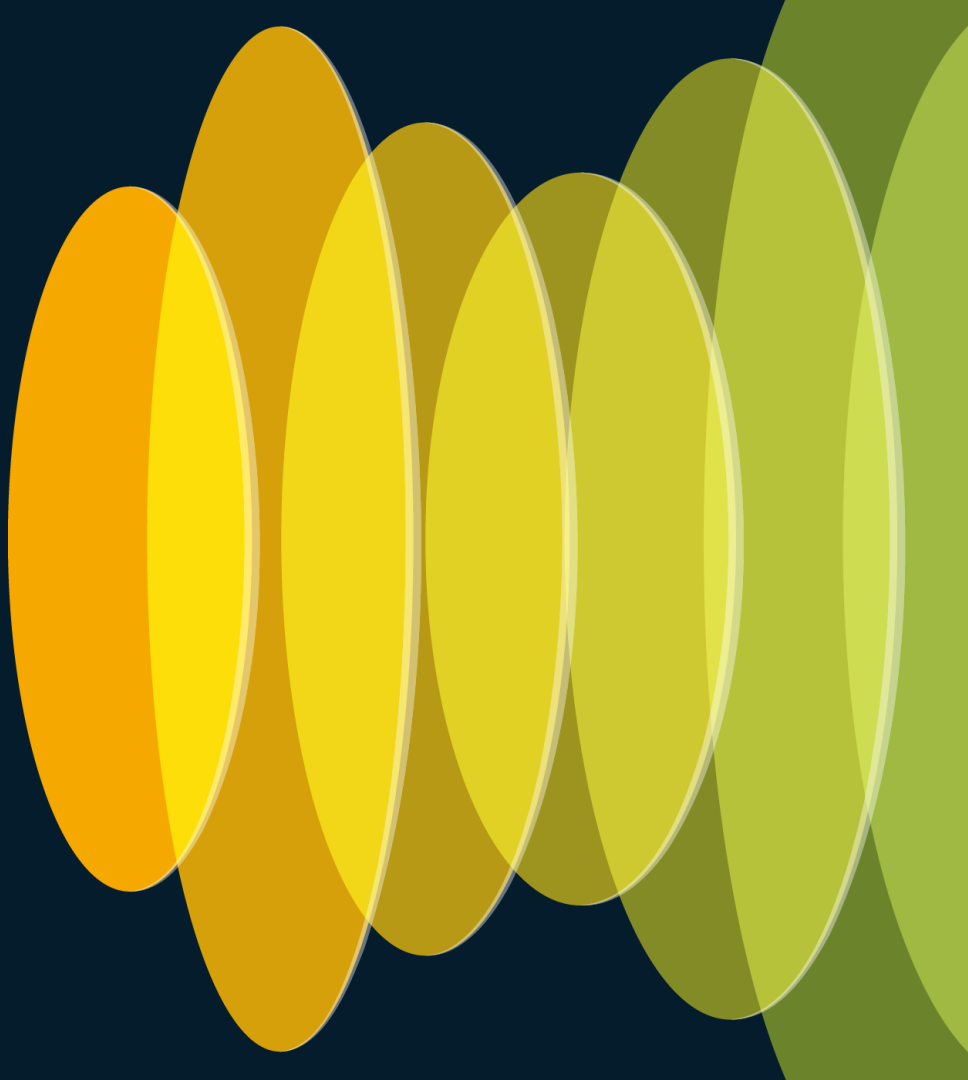


Migration Considerations

The Usual Approach of Building a New Parallel Network (3)



Migration With Rack Space Constraints

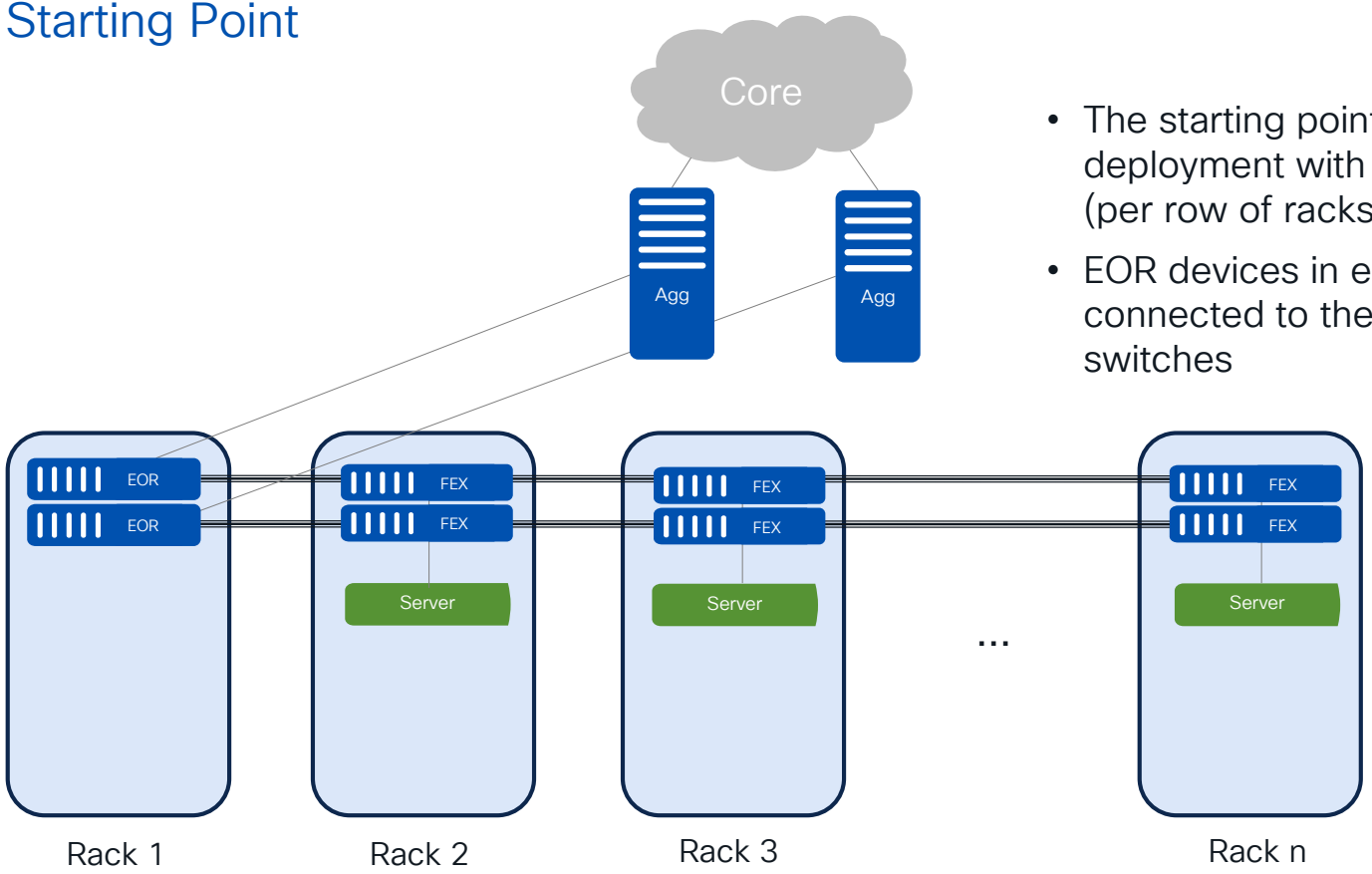


cisco Live!

Migration with Rack Space Constraints

Starting Point

1

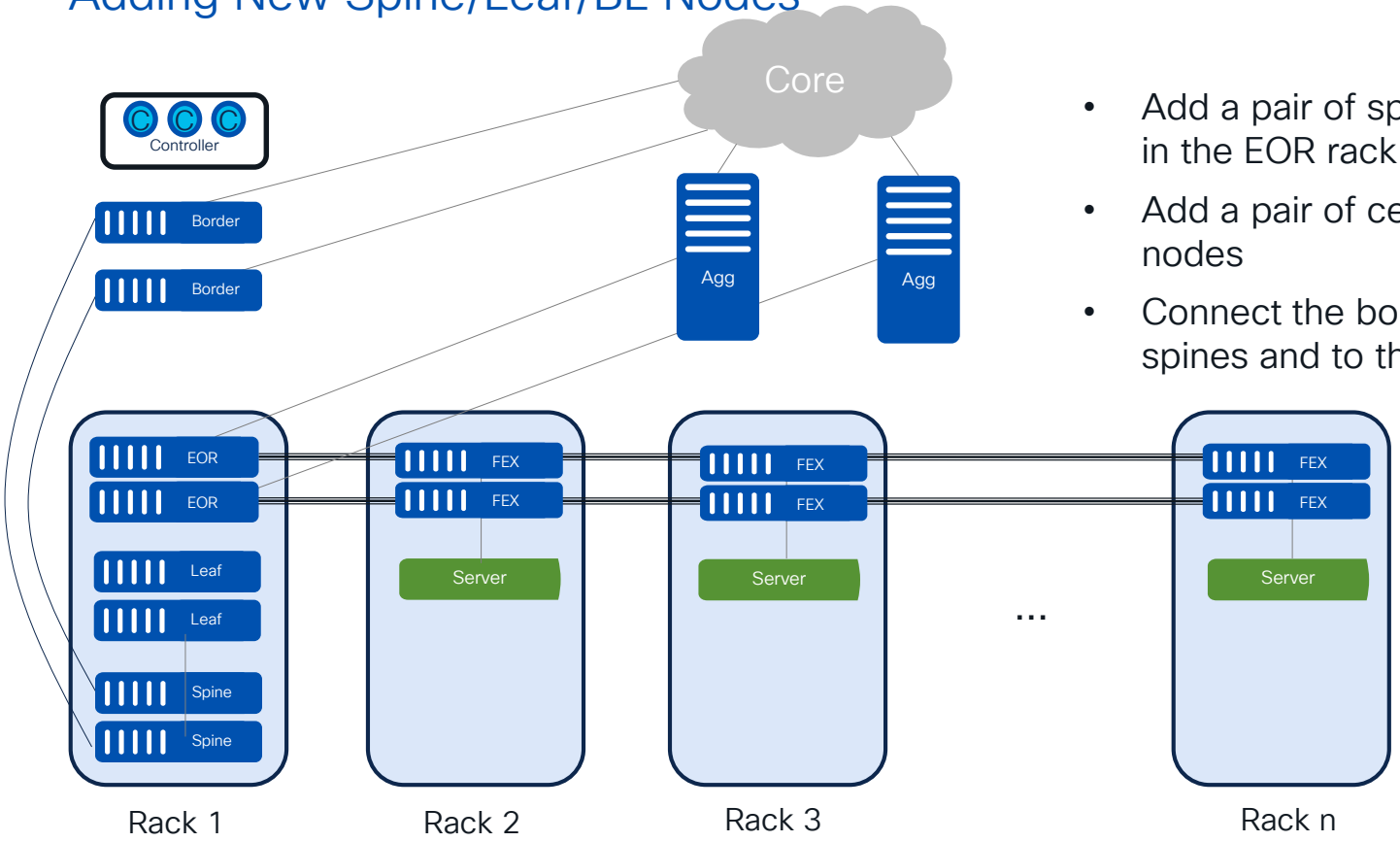


- The starting point is the traditional FEX deployment with a pair of EOR devices (per row of racks)
- EOR devices in each row are connected to the centralized Agg switches

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

2

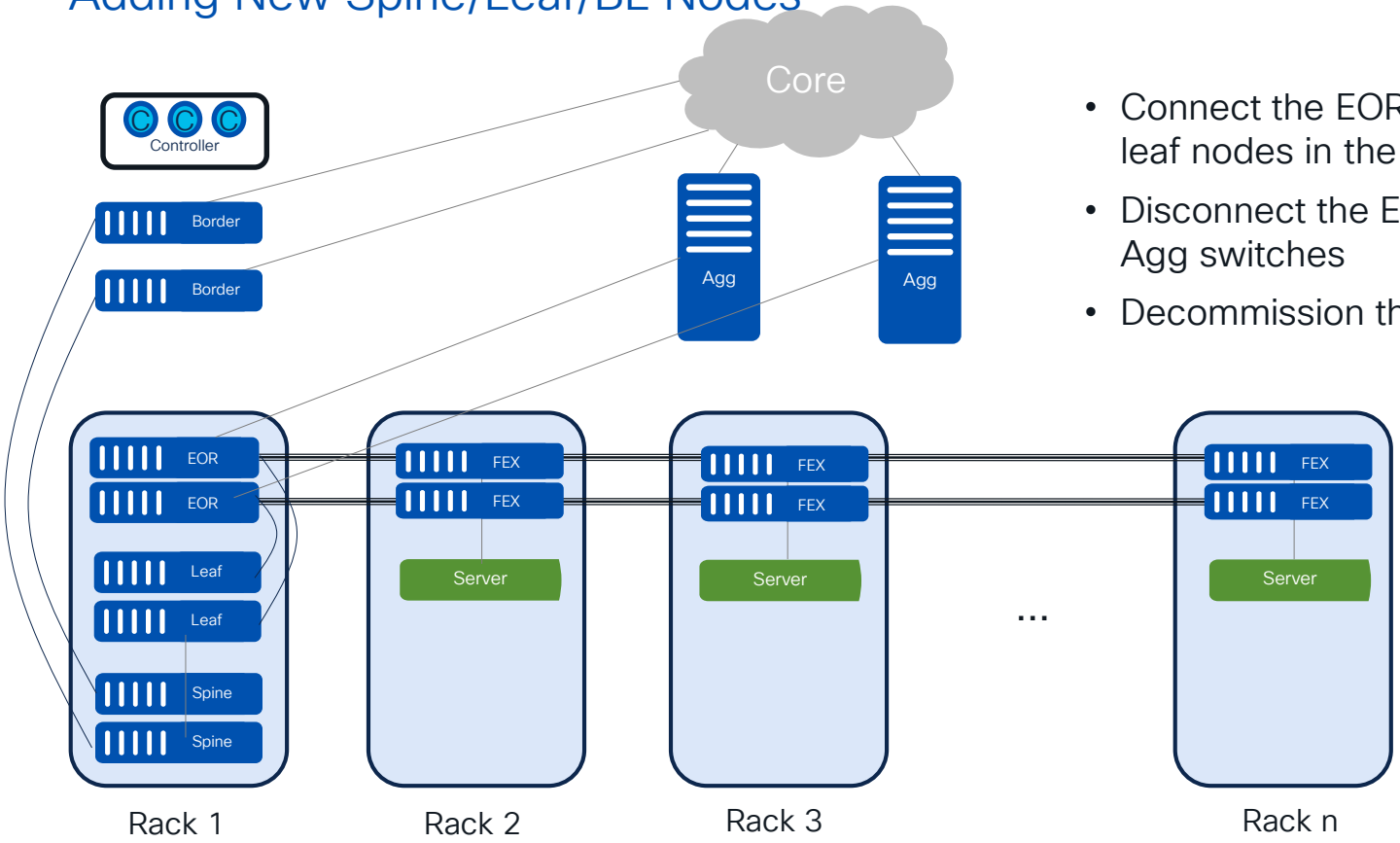


- Add a pair of spine and leaf switches in the EOR rack
- Add a pair of centralized border nodes
- Connect the border nodes to the spines and to the core

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

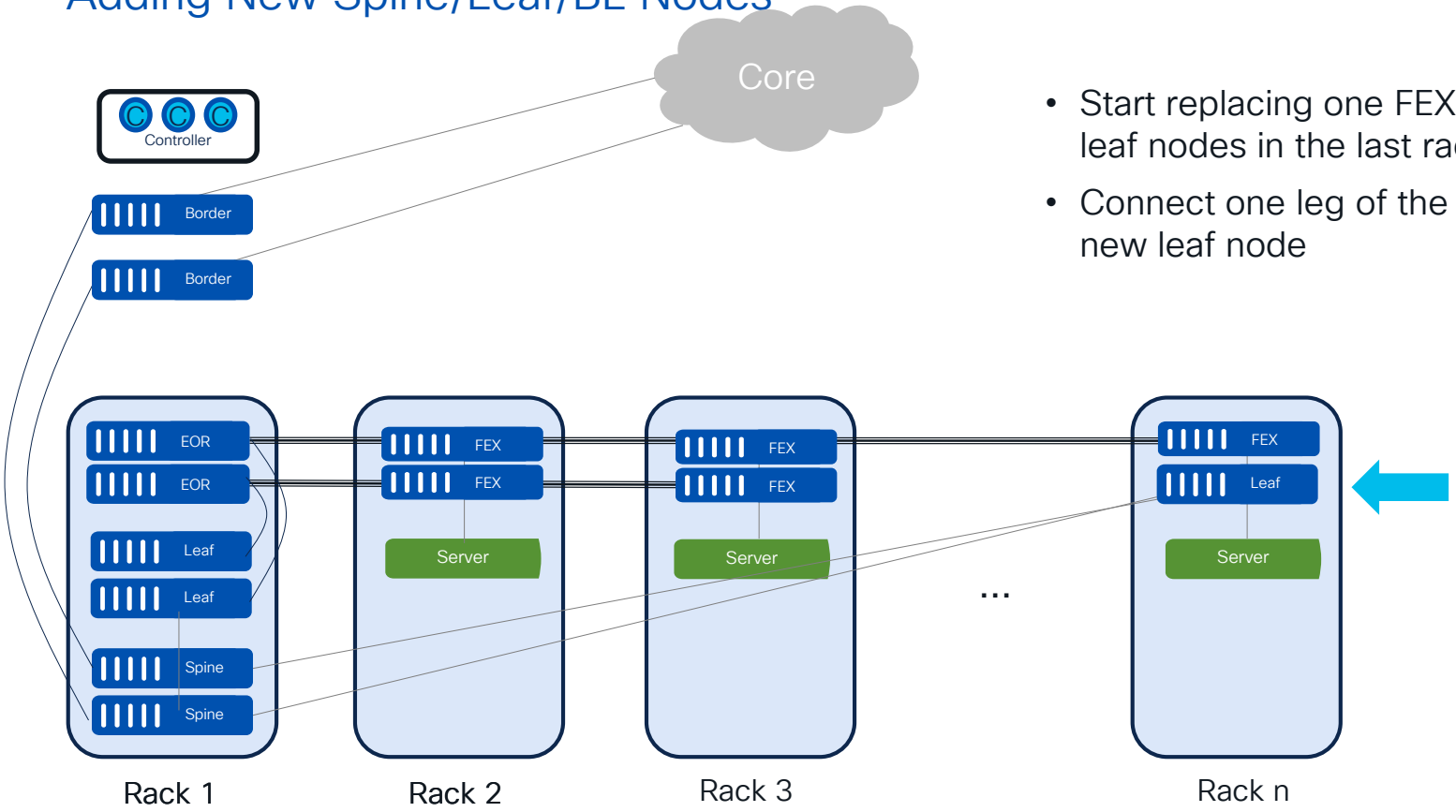
3



- Connect the EOR devices to the pair of leaf nodes in the EOR rack (L2 + L3)
- Disconnect the EOR devices from the Agg switches
- Decommission the Agg switches

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

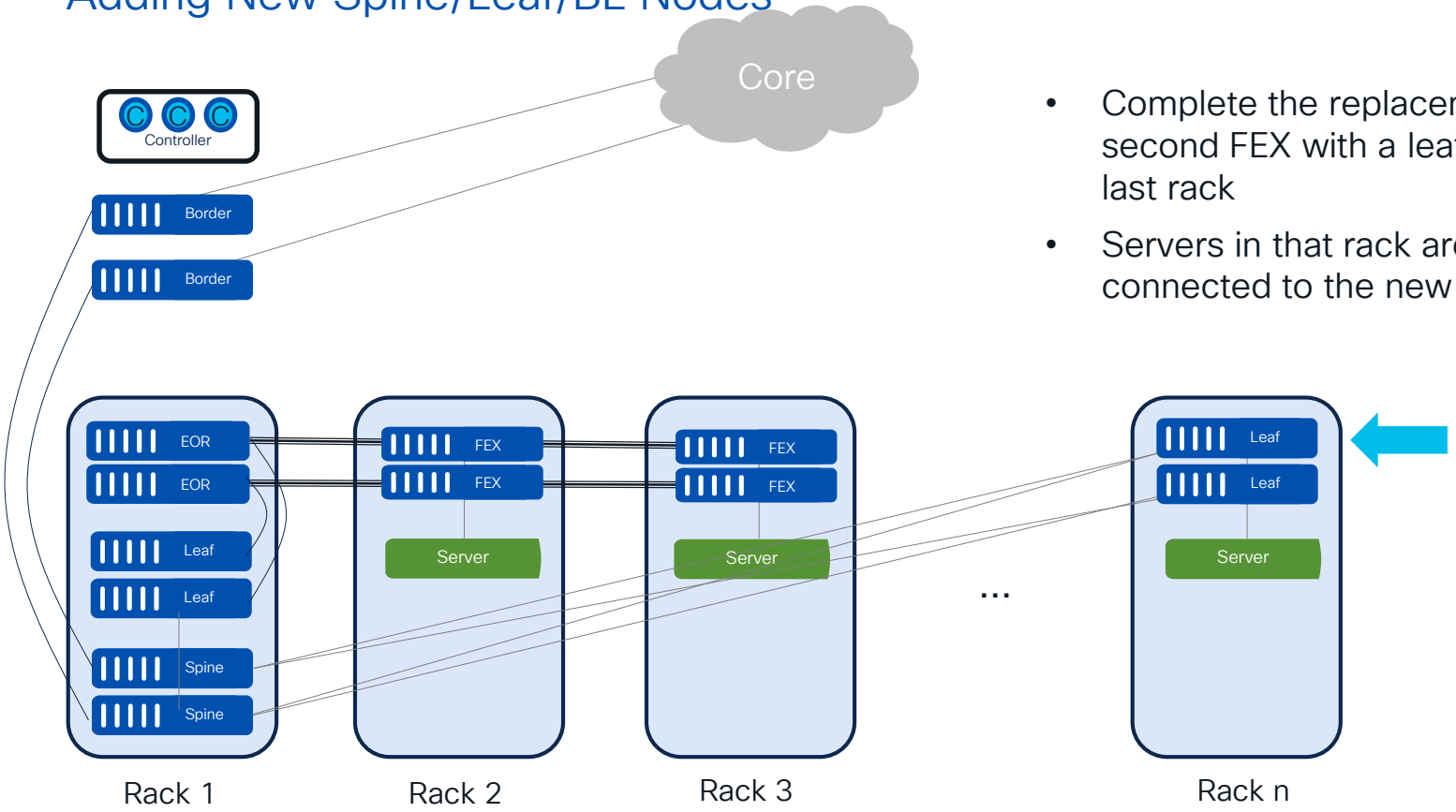


- Start replacing one FEX with a new leaf nodes in the last rack
- Connect one leg of the servers to the new leaf node

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

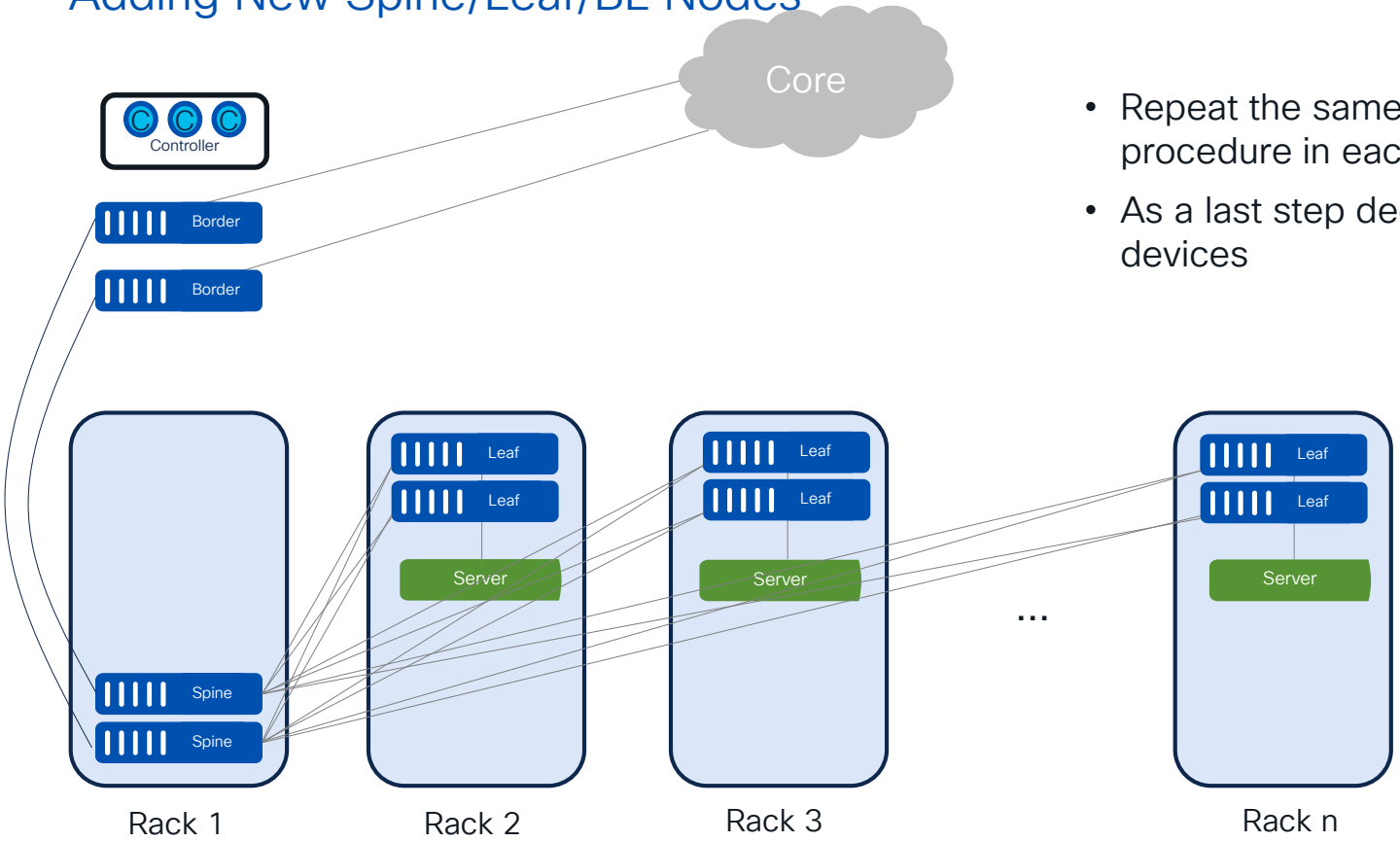
5



- Complete the replacement of the second FEX with a leaf node in the last rack
- Servers in that rack are now only connected to the new fabric

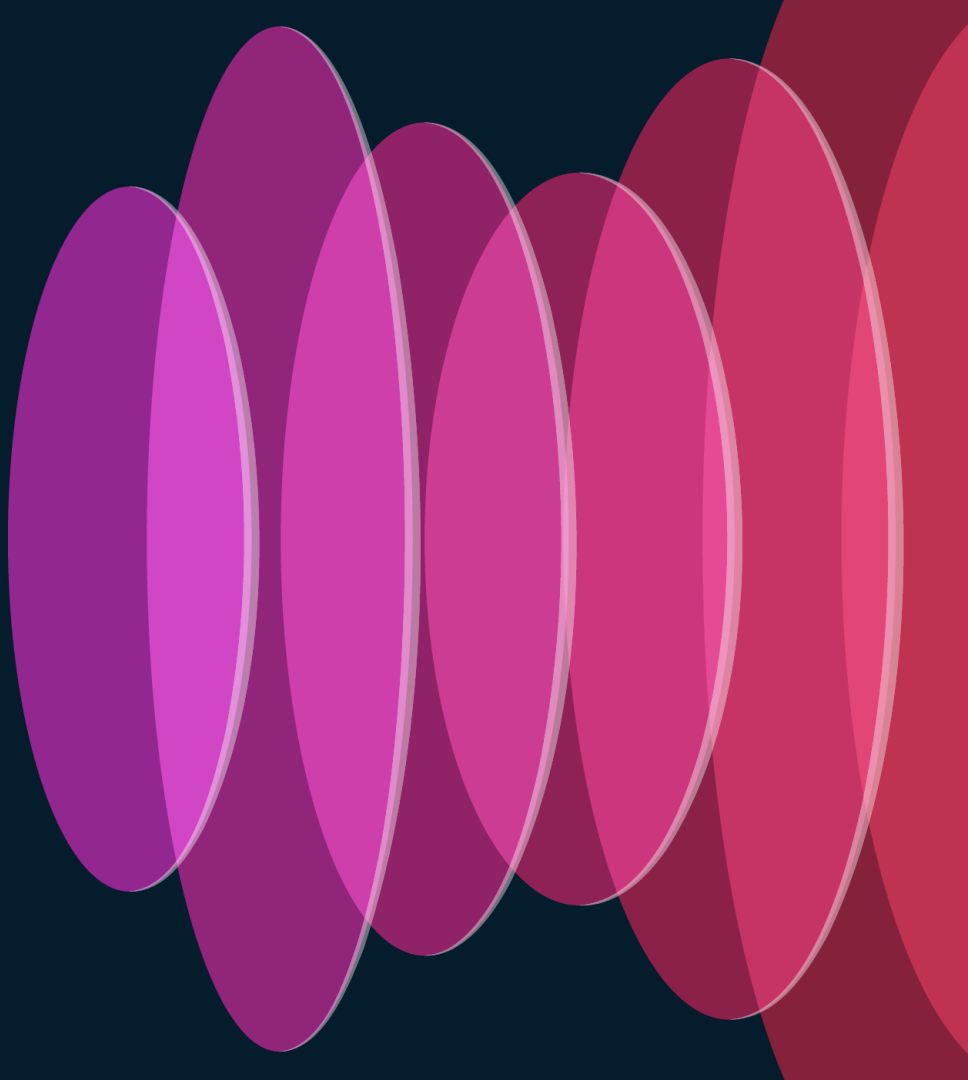
Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes



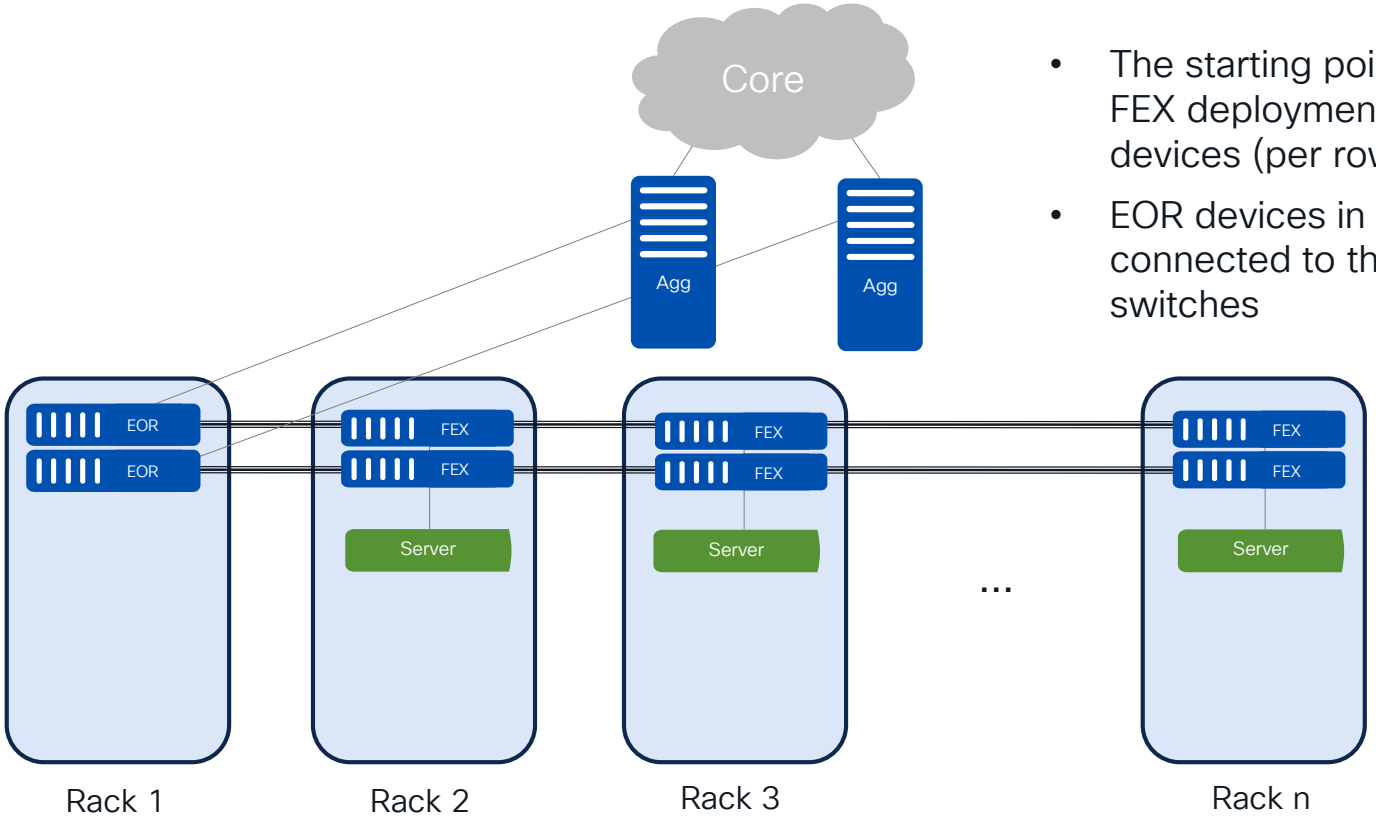
- Repeat the same FEX replacement procedure in each rack
- As a last step decommission the EOR devices

Migration Without Rack Space Constraints



Migration without Rack Space Constraints

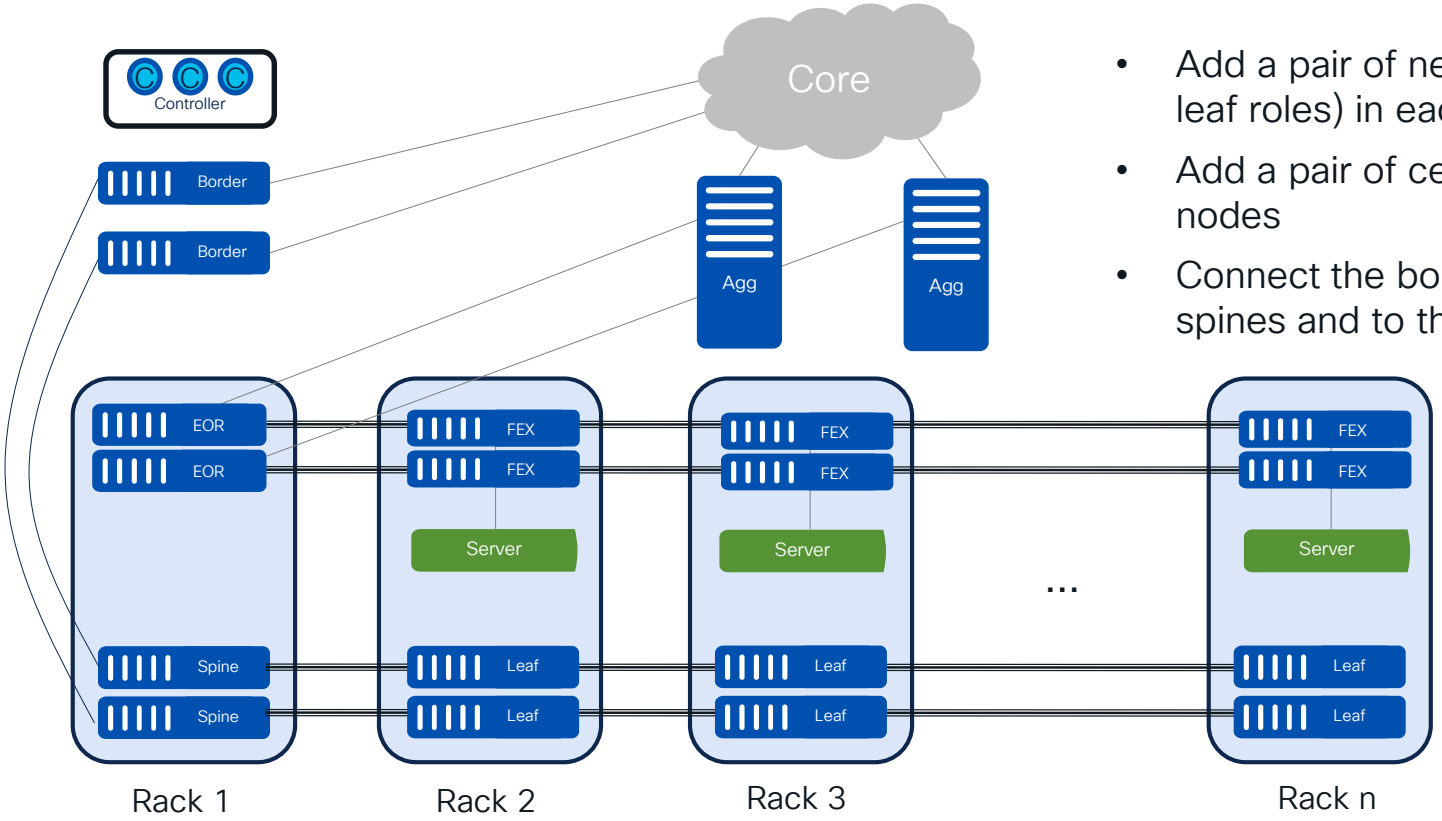
Starting Point



- The starting point is the traditional FEX deployment with a pair of EOR devices (per row of racks)
- EOR devices in each row are connected to the centralized Agg switches

Migration without Rack Space Constraints

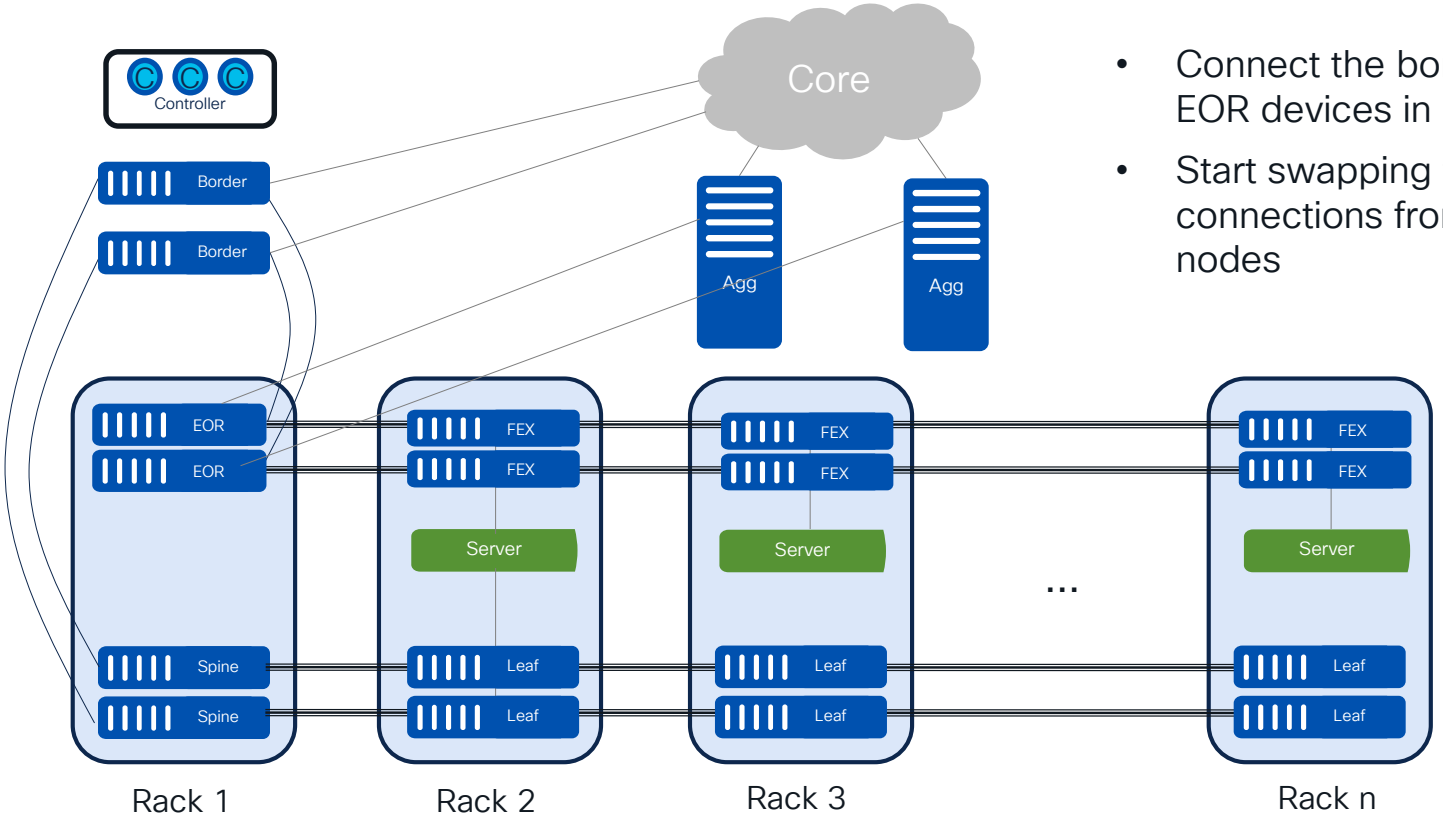
Adding New Spine/Leaf/BL Nodes



- Add a pair of new devices (spine and leaf roles) in each rack
- Add a pair of centralized border nodes
- Connect the border nodes to the spines and to the core

Migration without Rack Space Constraints

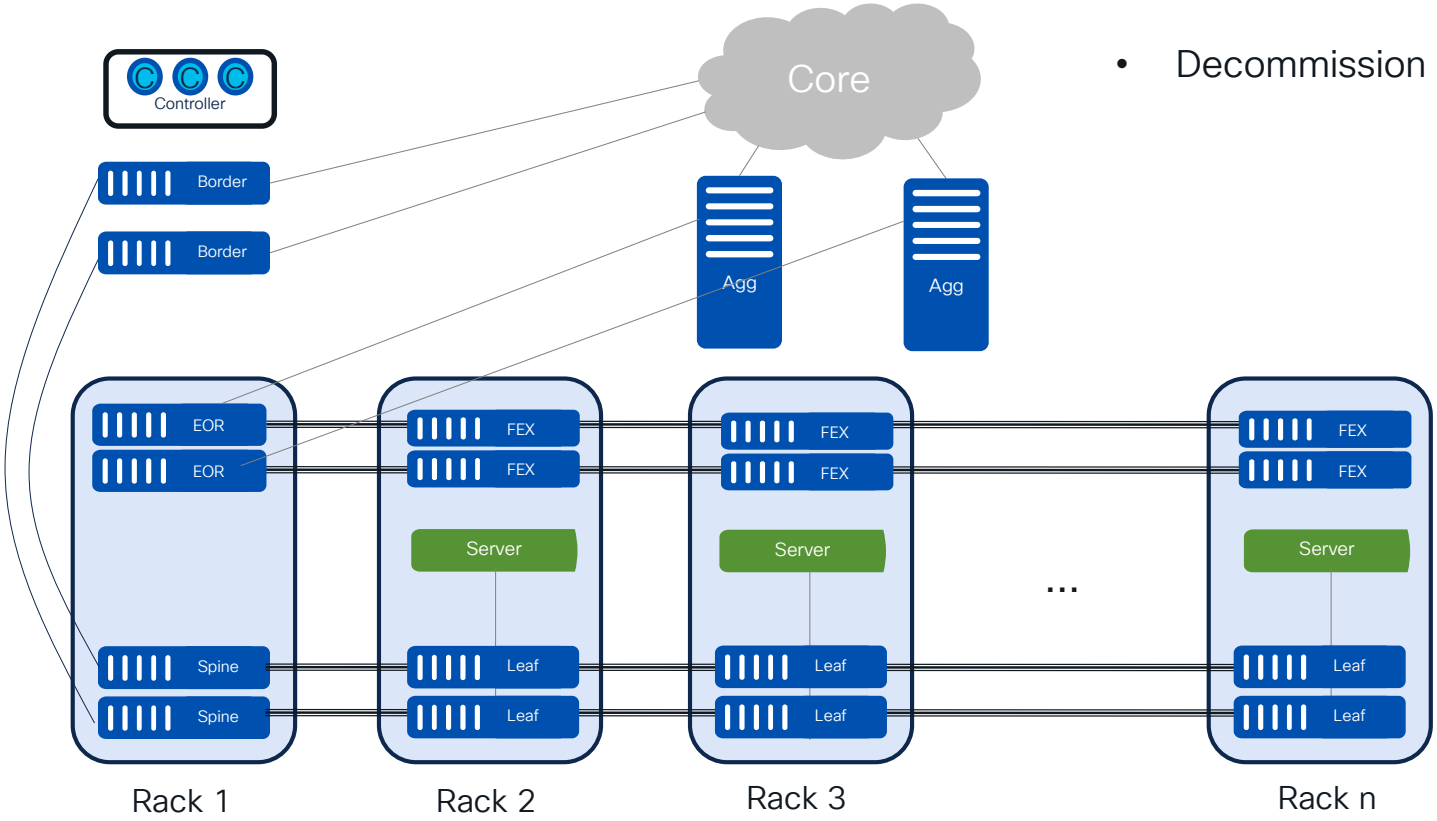
Connect Old and New Infrastructures and Migrate Endpoints



- Connect the border nodes to the EOR devices in each row (L2 and L3)
- Start swapping the servers connections from the FEX to the leaf nodes

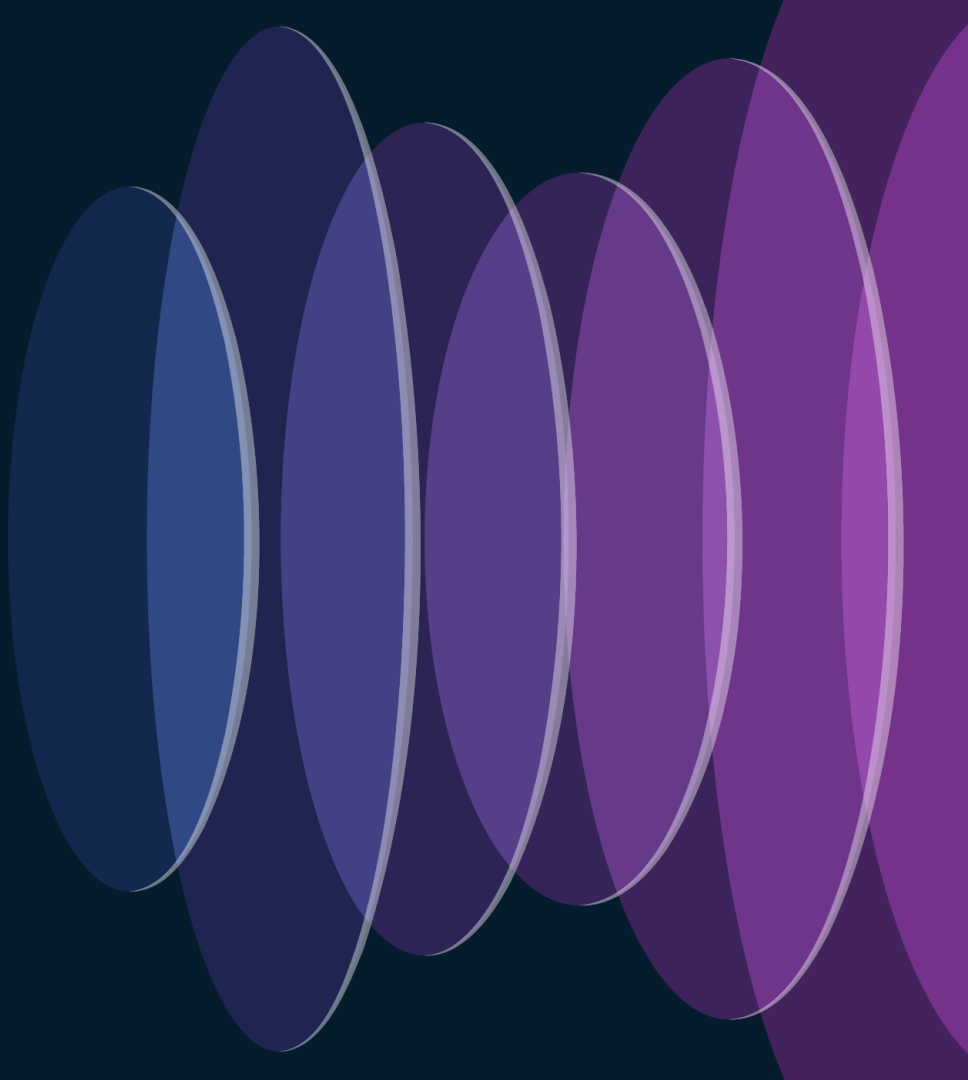
Migration without Rack Space Constraints

Decommission the Old Infrastructure

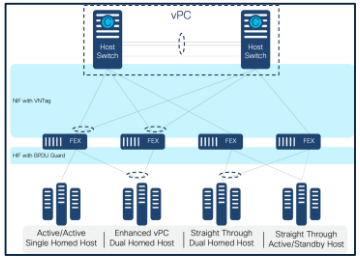


- Decommission the old infrastructure

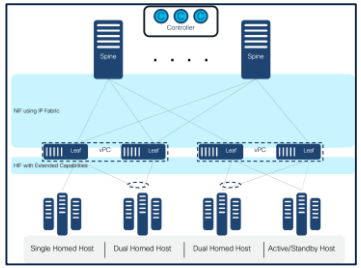
Conclusion



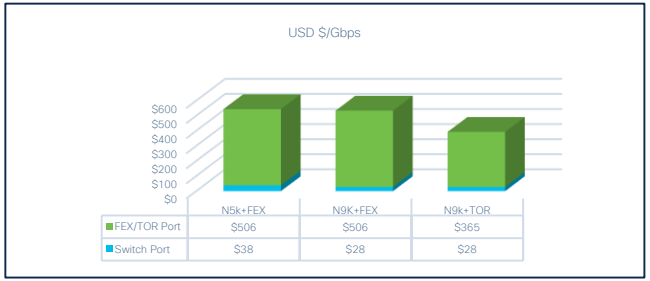
Conclusions



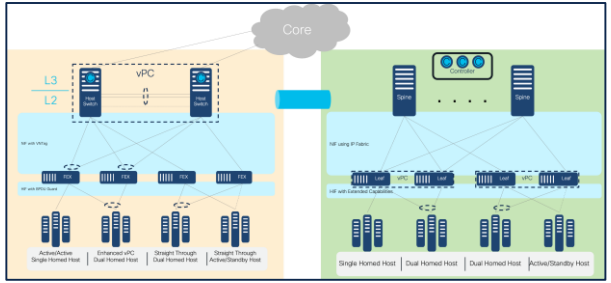
- FEX was the first attempt to build a fabric infrastructure
  - Centralized Management
  - Network and Host Redundancy



- Evolution of network architectures to deliver full fledge fabrics
  - Centralized Management with Controller
  - Fully distributed control and data planes



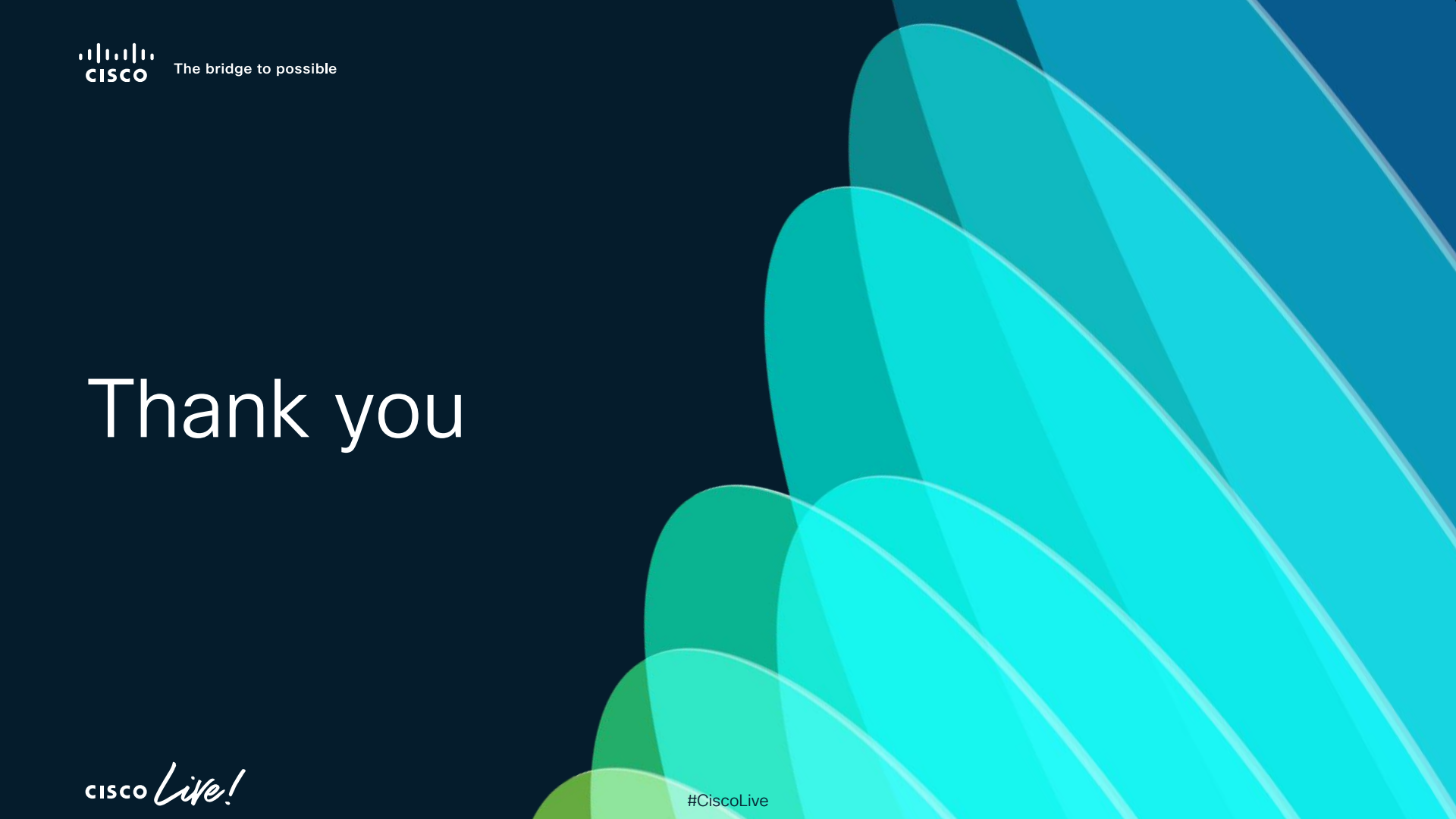
- Bandwidth/Cost Evolution over a Decade
- Economics started favoring deployment of switches as ToRs



- Usual migration approach of building a parallel network
- Couple options based on existence of rack space constraints

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


The bridge to possible
Thank you
cisco Live!
#CiscoLive