cisco The bridge to possible

Simple Leaf/Spine with a Touch of ToR

Network Designs for the Modern Data Center

Brenden Buresh
Distinguished Architect
@BrendenBuresh

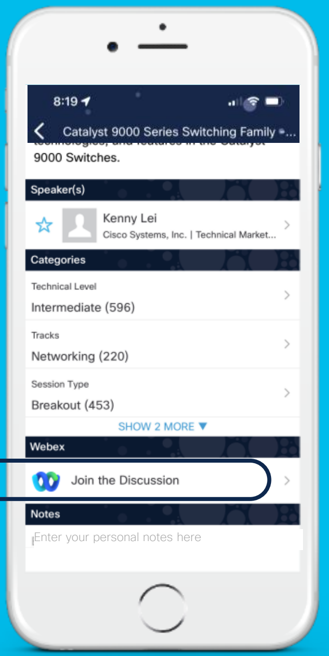
Cisco Webex App

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 until February 24, 2023.

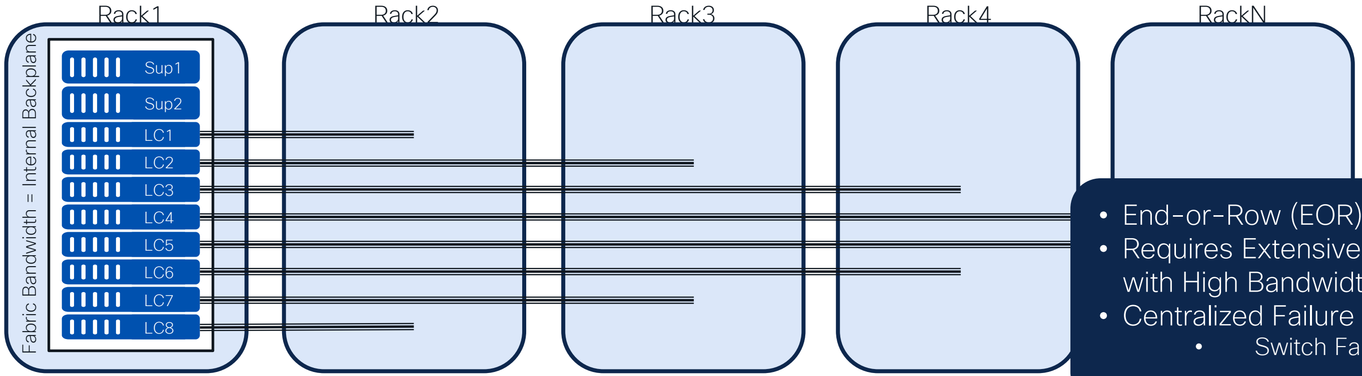
Agenda

- Why Did We Introduce FEX?
- The Evolution of DC Network Designs
- Fundamentals of VXLAN EVPN Design
- Bandwidth/Cost Evolution Over a Decade
- Migration Considerations
  - Migration with Rack Space Constraints
  - Migration without Rack Space Constraints
- Conclusion

Why Did We Introduce FEX?

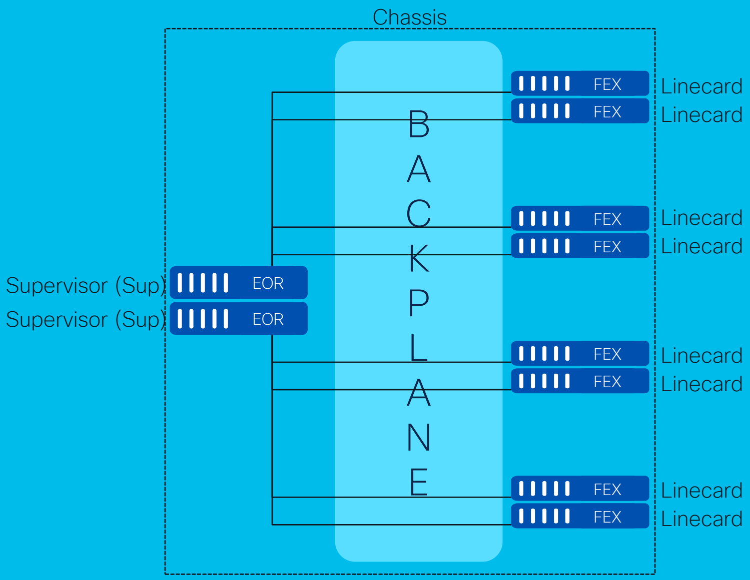
Middle of Row (MoR) and End of Row (EoR)

Big Centralized Chassis



- End-or-Row (EOR) or Mid-of-Row (MOR) Design
- Requires Extensive Structured Ethernet Cabling with High Bandwidth Switch
- Centralized Failure Domain
    - Switch Failure = Network Failure

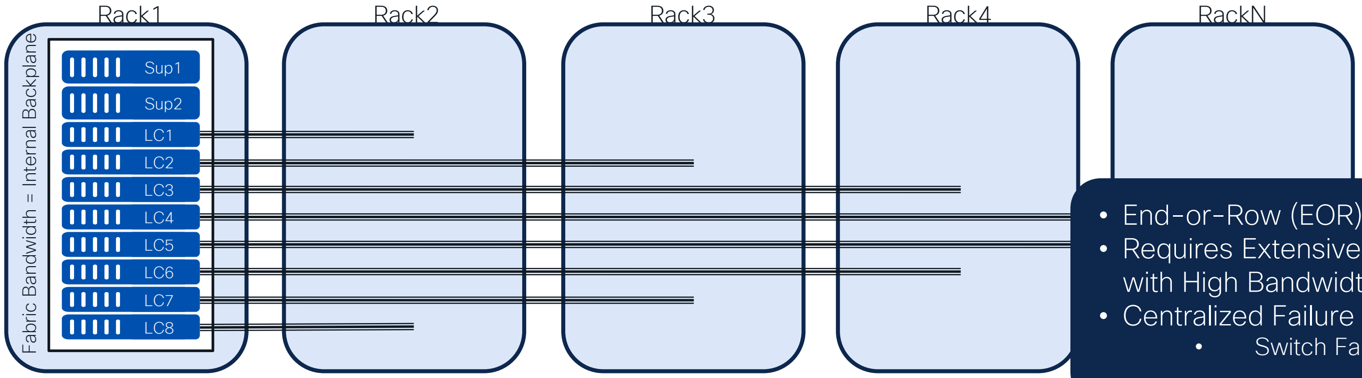
What is FEX?



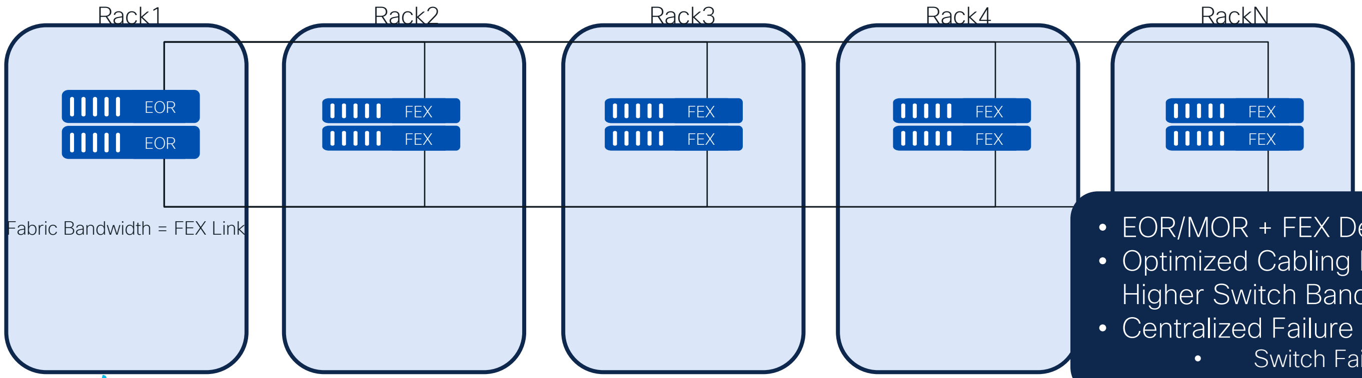
- A FEX can be seen as a way of “disaggregating” a traditional modular switch
- Enables the capability to build a centrally managed but highly distributed network design

Middle of Row (MoR) and End of Row (EoR)

Big Centralized Chassis

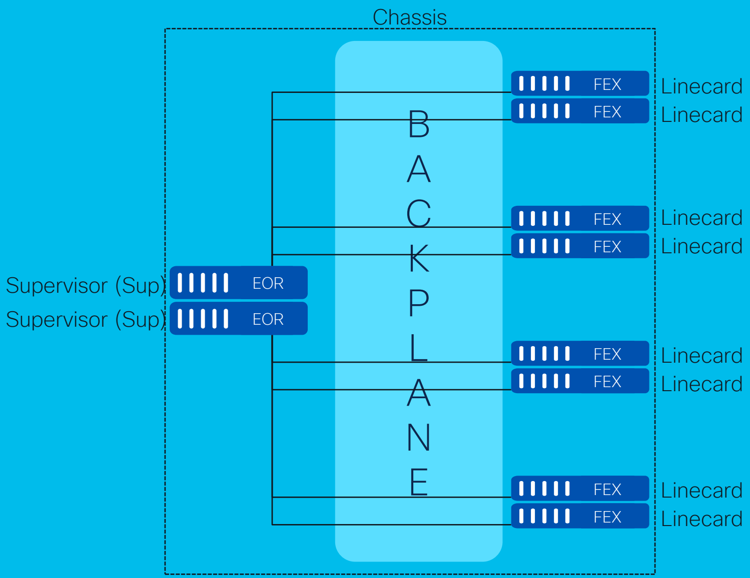


- End-or-Row (EOR) or Mid-of-Row (MOR) Design
- Requires Extensive Structured Ethernet Cabling with High Bandwidth Switch
- Centralized Failure Domain
    - Switch Failure = Network Failure



- EOR/MOR + FEX Design
- Optimized Cabling Requirement and Leverages Higher Switch Bandwidth
- Centralized Failure Domain
    - Switch Failure = Network Failure

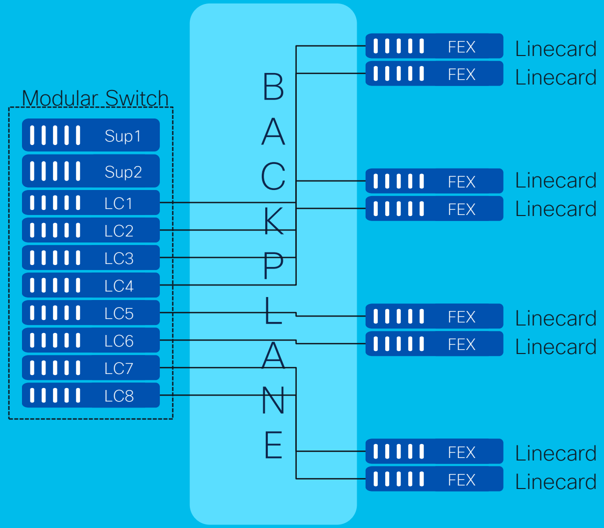
Why Did We Introduce FEX?



N5k: 24 FEX * 48 Host Ports = 1152 Host Ports (HIF)
N9k: 16 FEX * 48 Host Ports = 768 Host Ports (HIF)

- Centralized Management
- Modular Chassis Feeling
  - Unified CLI Structure for Config and Operation
- Capability of offering multiple port speeds (100M/1G/10G)
- Economics, Relative High Cost of Switch Ports or $ per Gbps

When to Avoid Leveraging FEX?



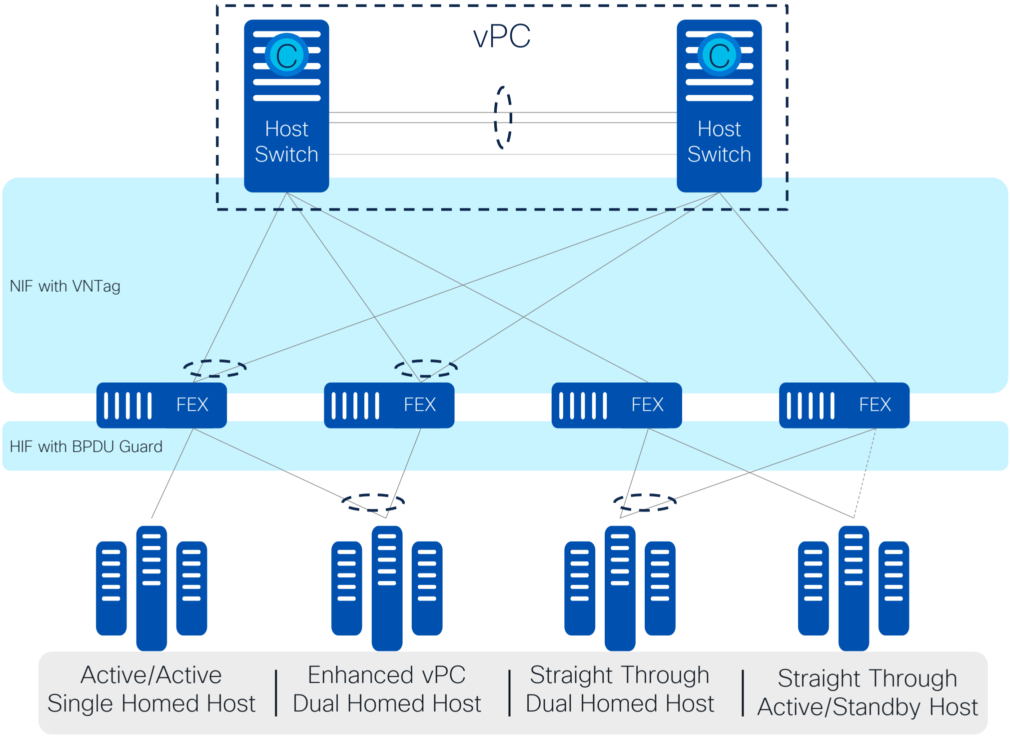
N7k: 64 FEX * 48 Host Ports = 3072 Host Ports (HIF)

- Extending Centralized Management (Beyond the Linecards)
- Increasing Modular Chassis Reach
  - Nested Linecard
  - Extending Failure Domain
- Giving Up the Benefits of a Distributed Fabric

A Data Center Fabric Prior to Data Center Fabrics

Cisco Fabric Extender (FEX) Overview

~14 Years ago
Around 2009



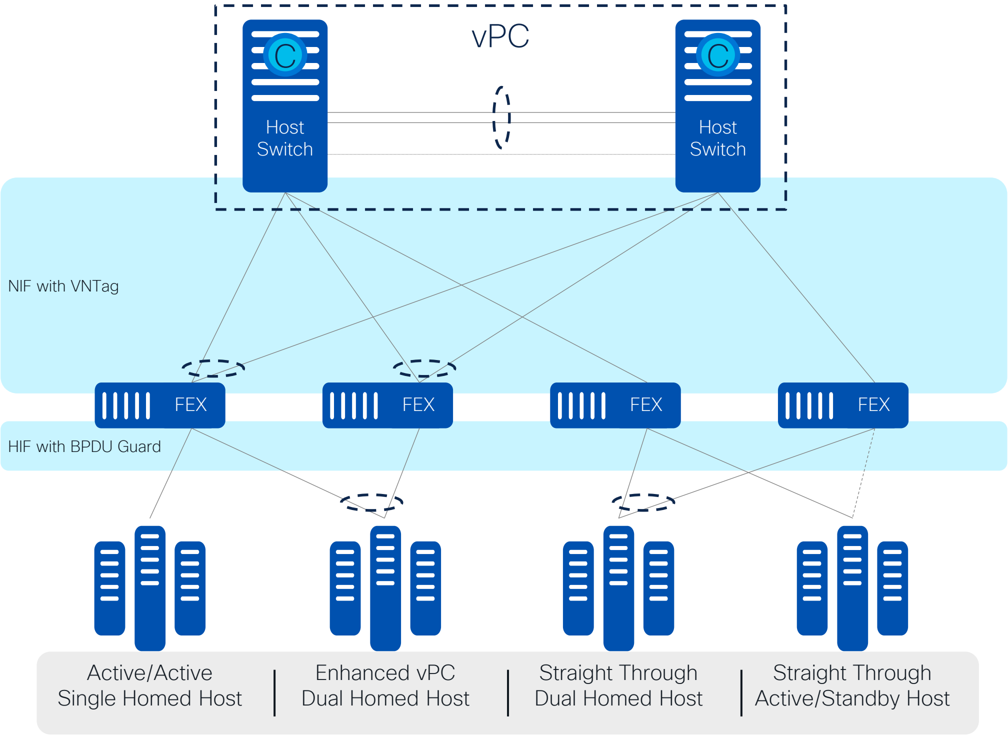
- Centralized Management
  - Co-located on the Switch
  - Limited to No Synchronization
  - Host Switch Operational Dependency
- Network Redundancy (NIF to NIF)
  - Uses VNTag (802.1BR / 802.1Qbh)
  - 1+1 Redundancy based on Layer-2 Port-Channel (vPC)
- Host Redundancy (Host to HIF)
  - Single Homed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

The Evolution of DC Network Designs

A Data Center Fabric Prior to Data Center Fabrics

Cisco Fabric Extender (FEX) Overview

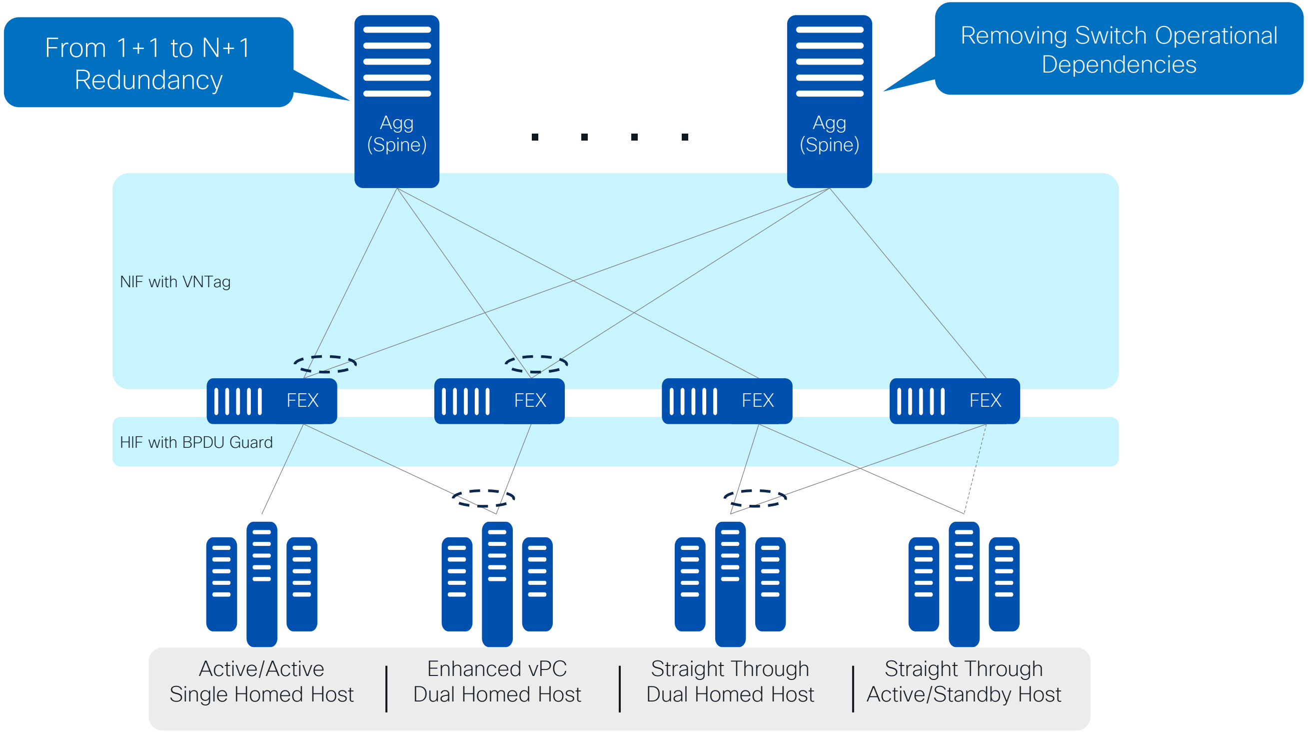
~14 Years ago
Around 2009



A Data Center Fabric Prior to Data Center Fabrics

Cisco Fabric Extender (FEX) Overview

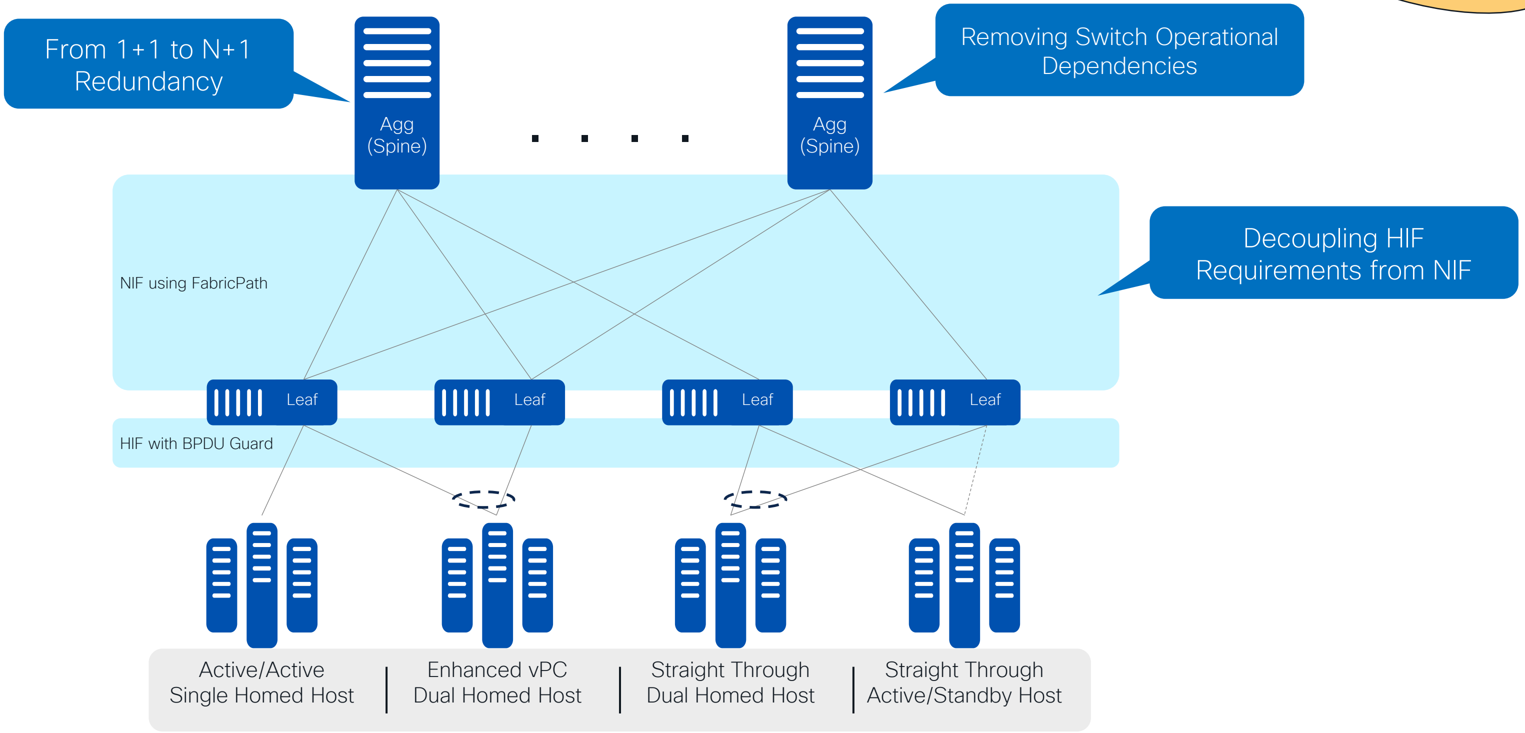
~12 Years ago
Around 2011



Early Steps in the Data Center Fabric Evolution

Evolution to a Fabric

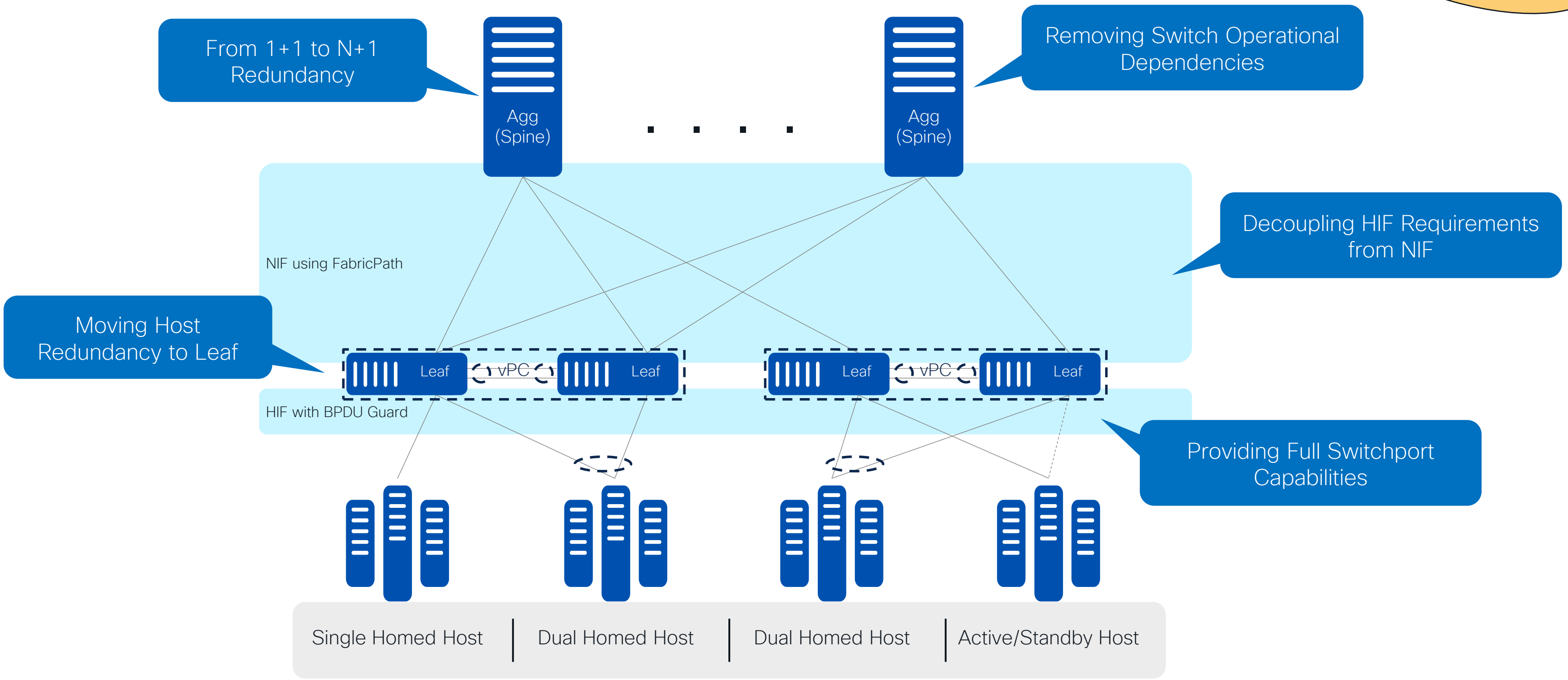
~12 Years ago
Around 2011



Early Steps in the Data Center Fabric Evolution

Evolution to a Fabric

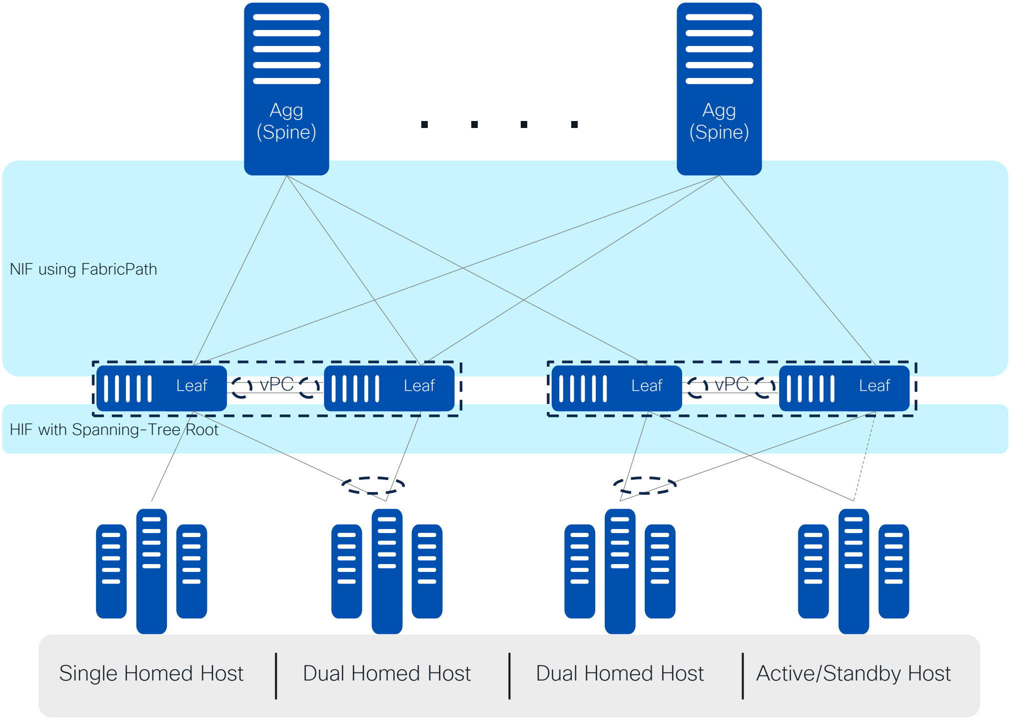
~12 Years ago
Around 2011



Early Steps in the Data Center Fabric Evolution

Cisco FabricPath Overview

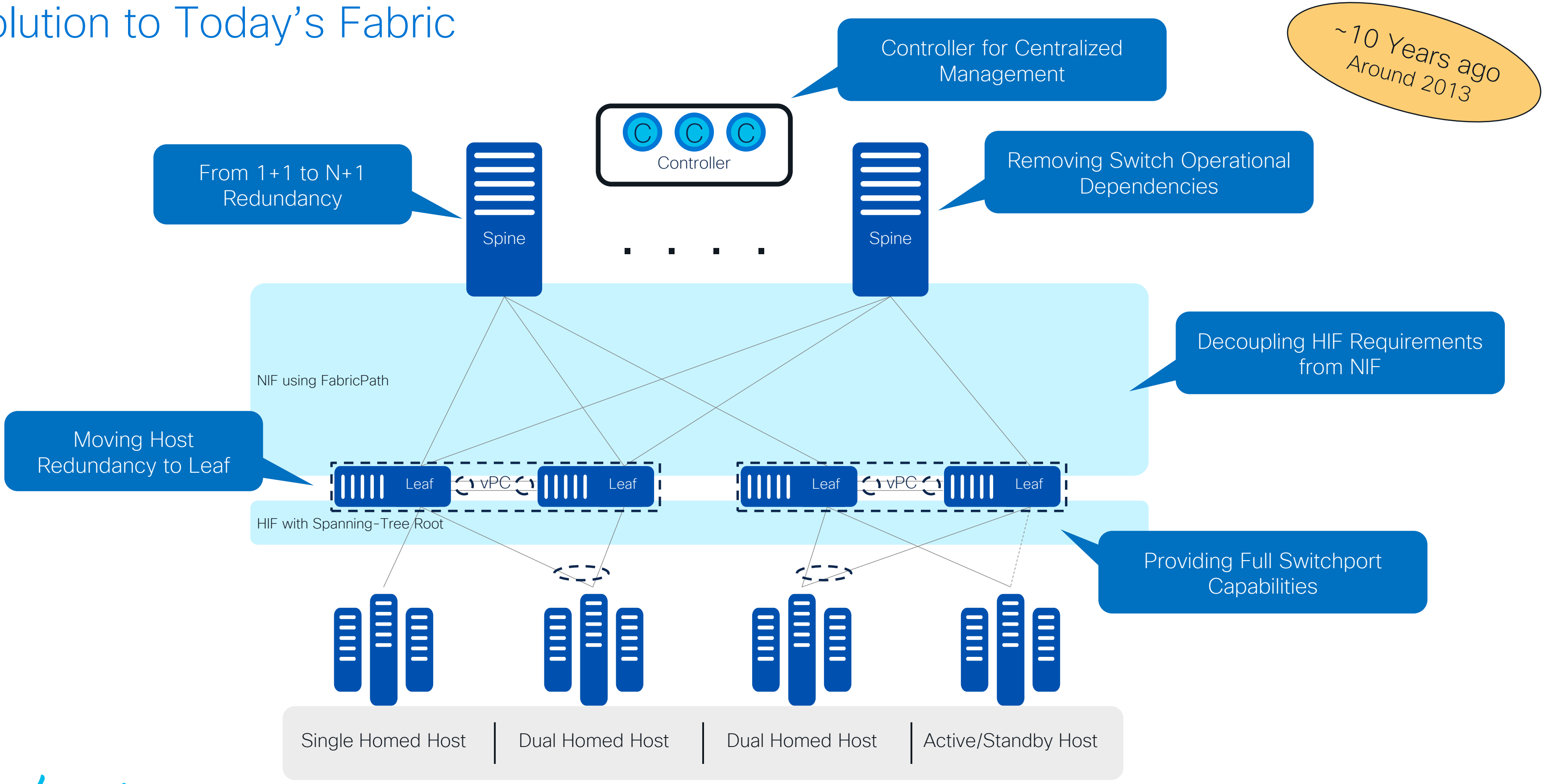
~12 Years ago
Around 2011



- 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

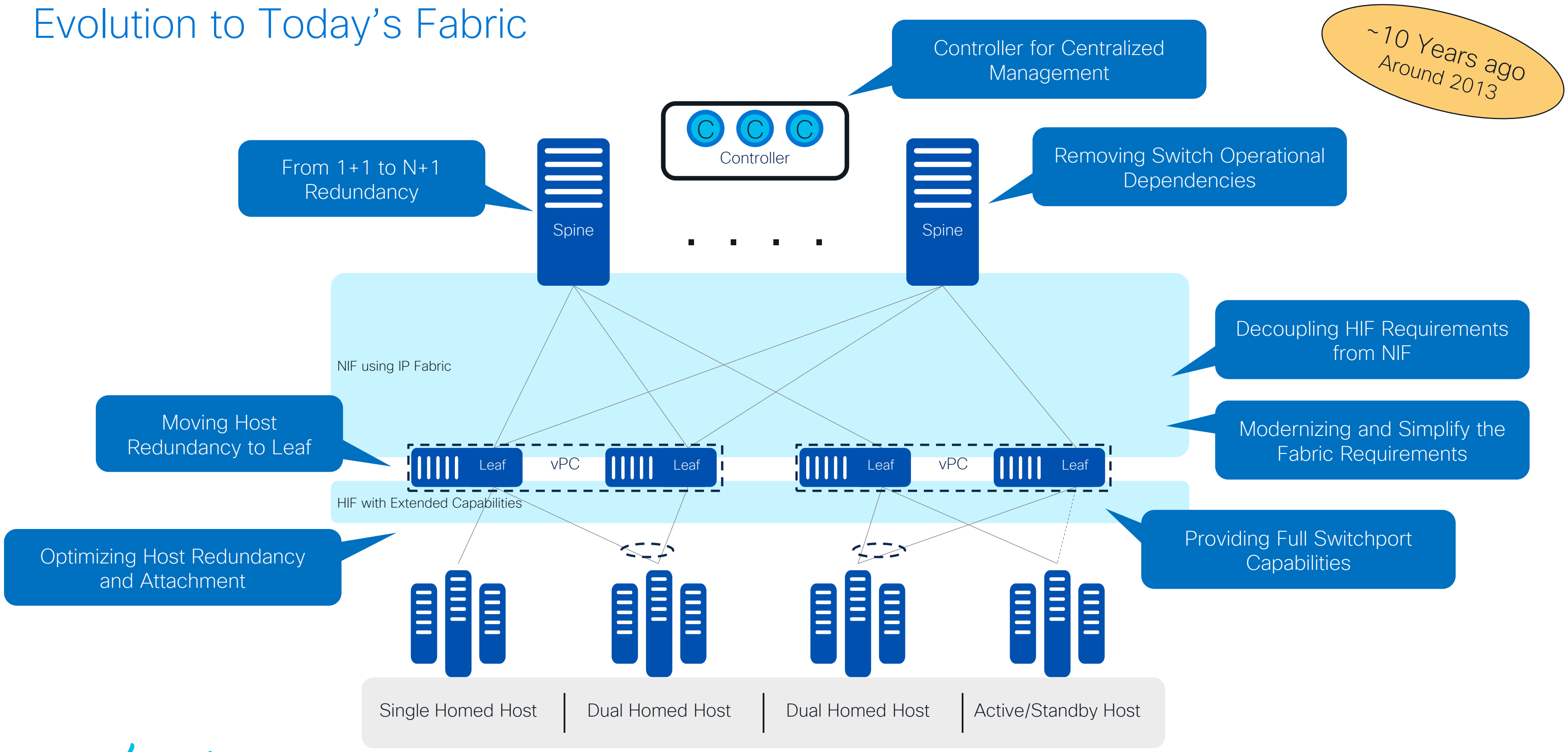
Using Mature SDN for Data Center Fabrics

Evolution to Today's Fabric



Using Mature SDN for Data Center Fabrics

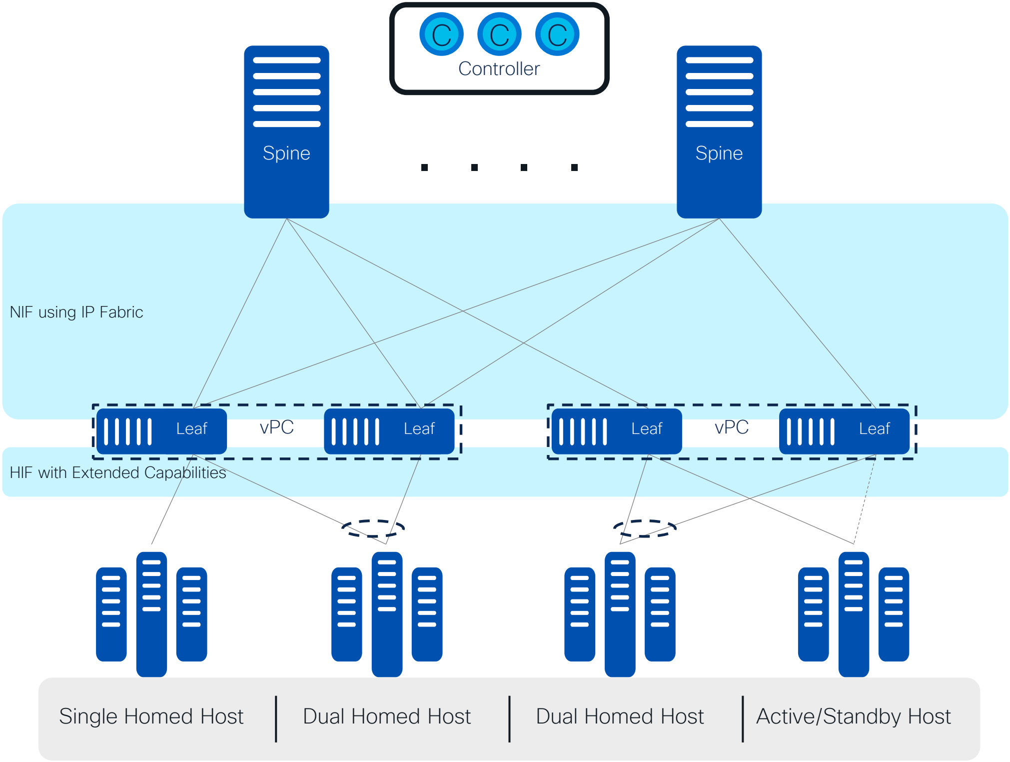
Evolution to Today's Fabric



Using Mature SDN for Data Center Fabrics

Cisco ACI and VXLAN EVPN Fabric Overview

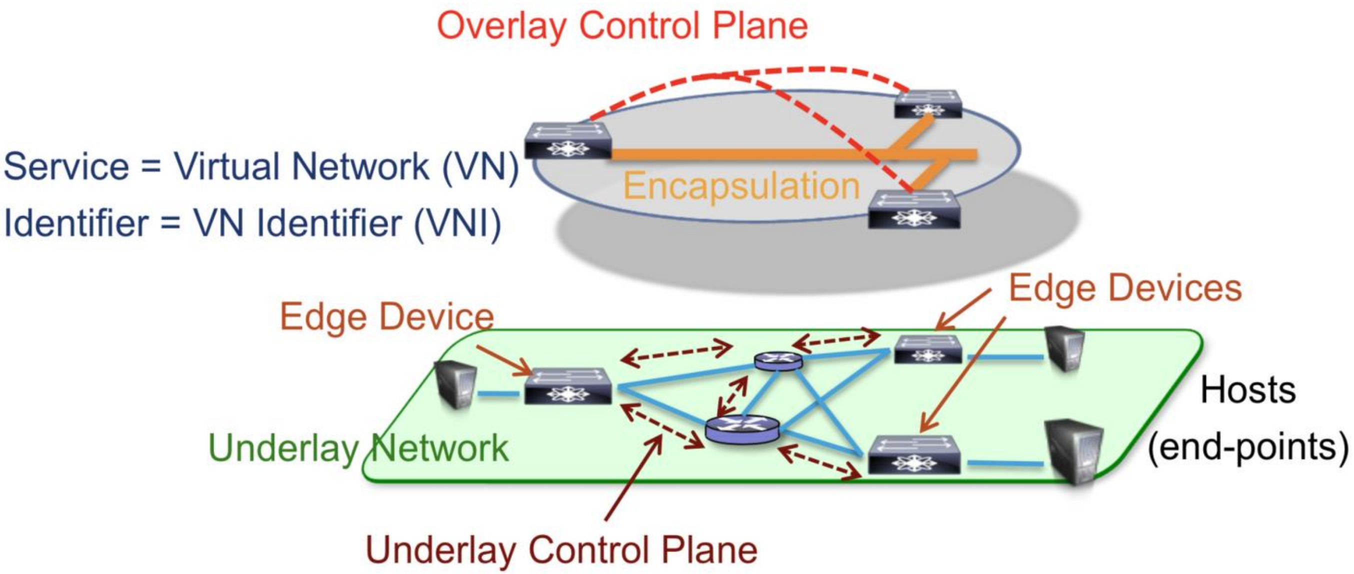
~10 Years ago
Around 2013



- 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

Fundamentals of VXLAN EVPN Design

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 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)

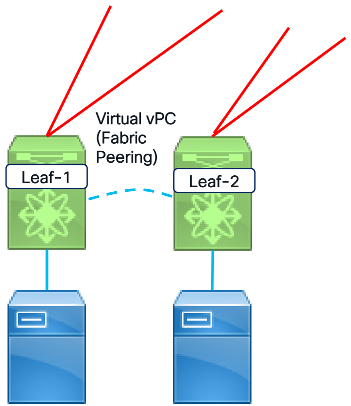
Network Components of VXLAN Overlays

- 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

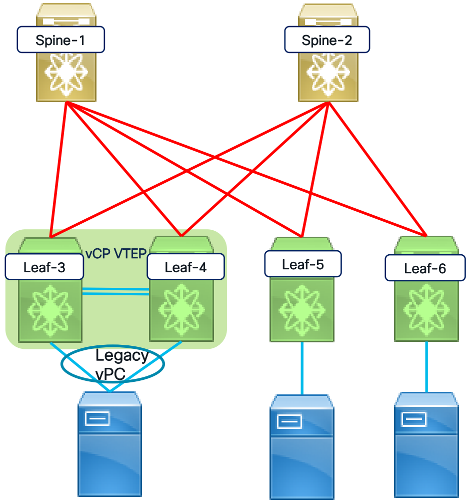
- Forms Routing Protocol adjacencies for underlay with Spines (OSPF, IS-IS, BGP)
- MP-BGP L2VPN EVPN neighborships with spines to exchange routes
- Performs VXLAN encapsulation and decapsulation
- Default Gateway Services for hosts using Distributed Anycast Gateway
- BUM replication/processing
- Connect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



Components of VXLAN EVPN

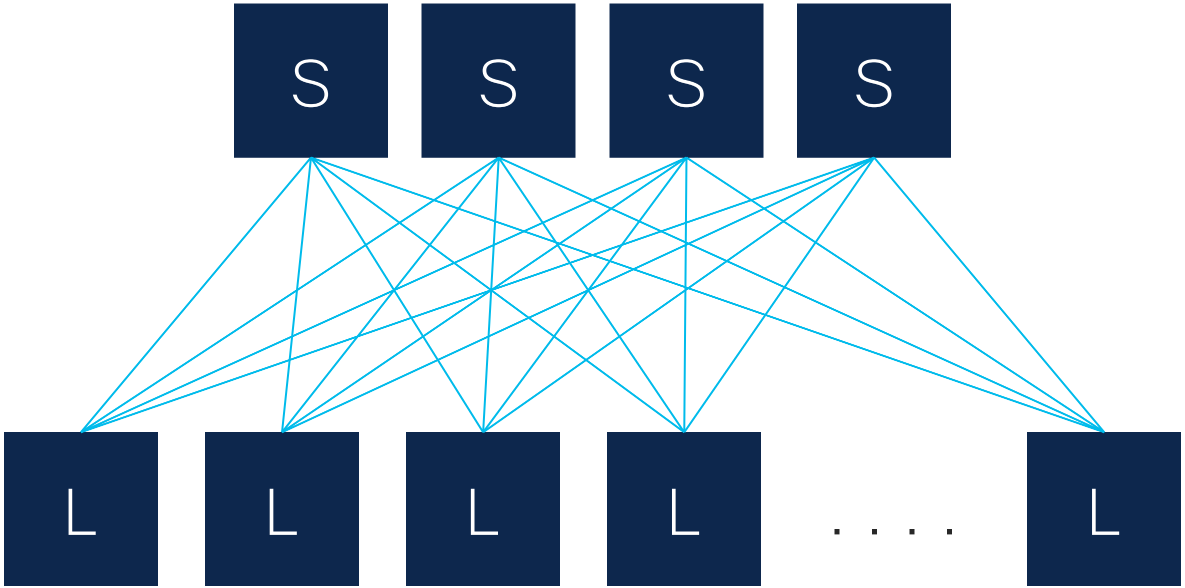
Functions of Spine

- 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



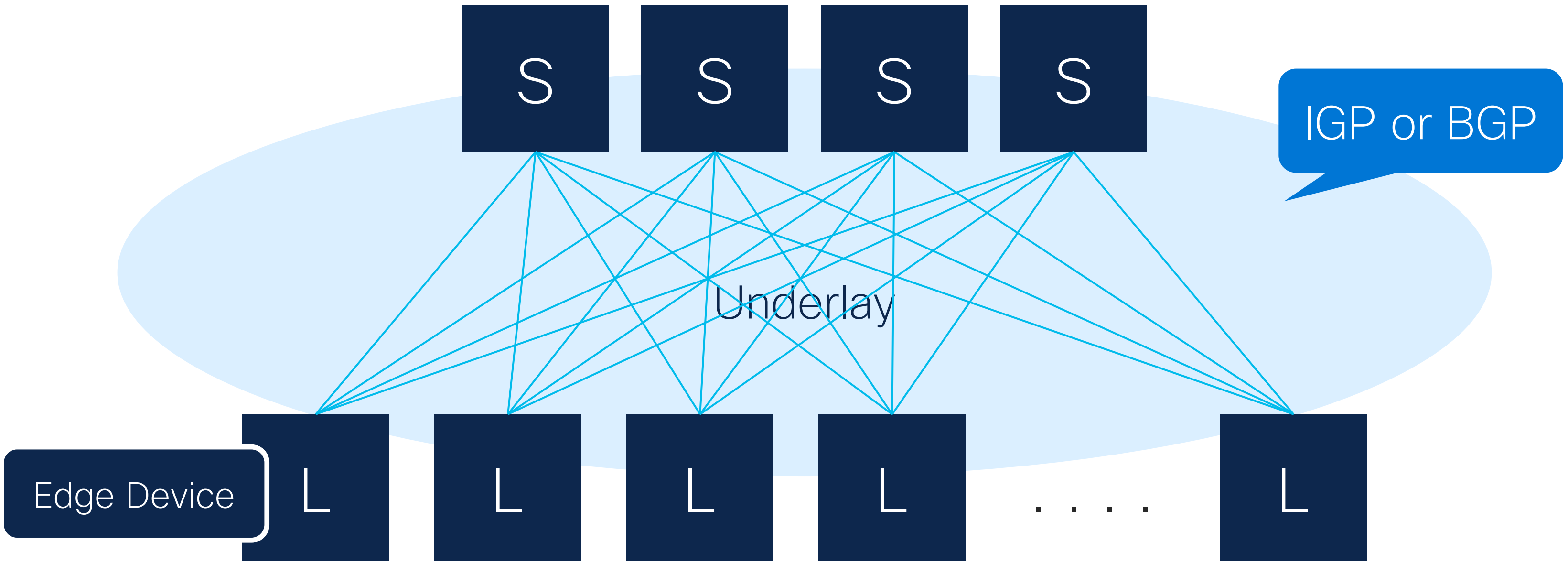
VXLAN: The Building Blocks

Underlay



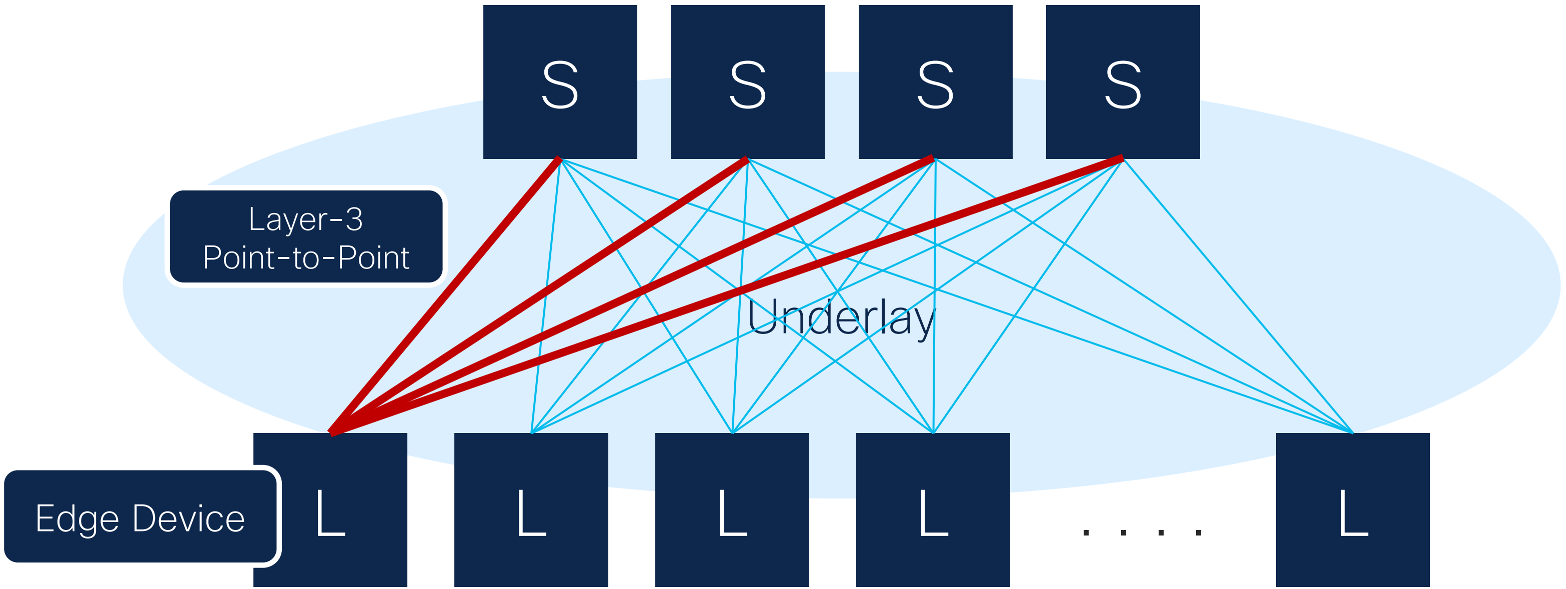
VXLAN: The Building Blocks

Underlay



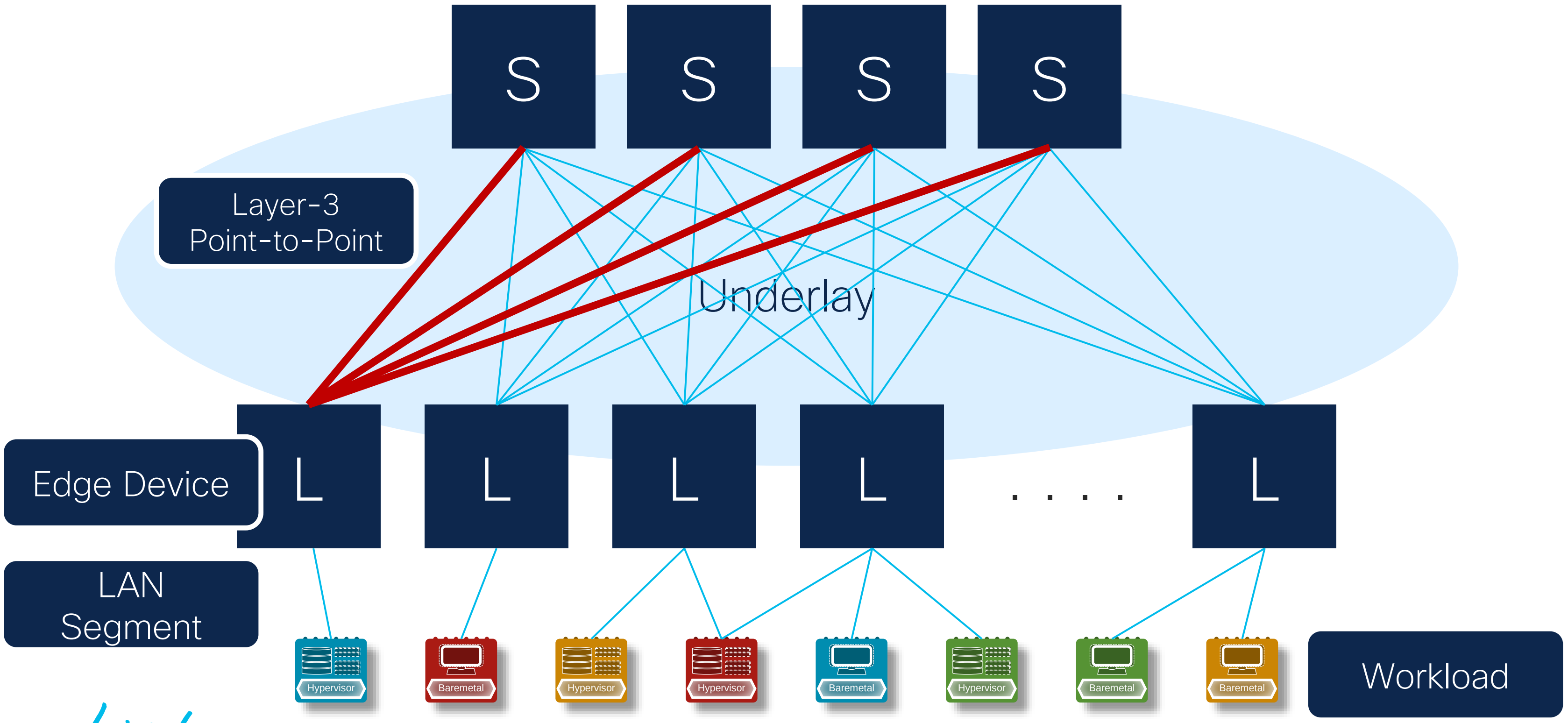
VXLAN: The Building Blocks

Underlay



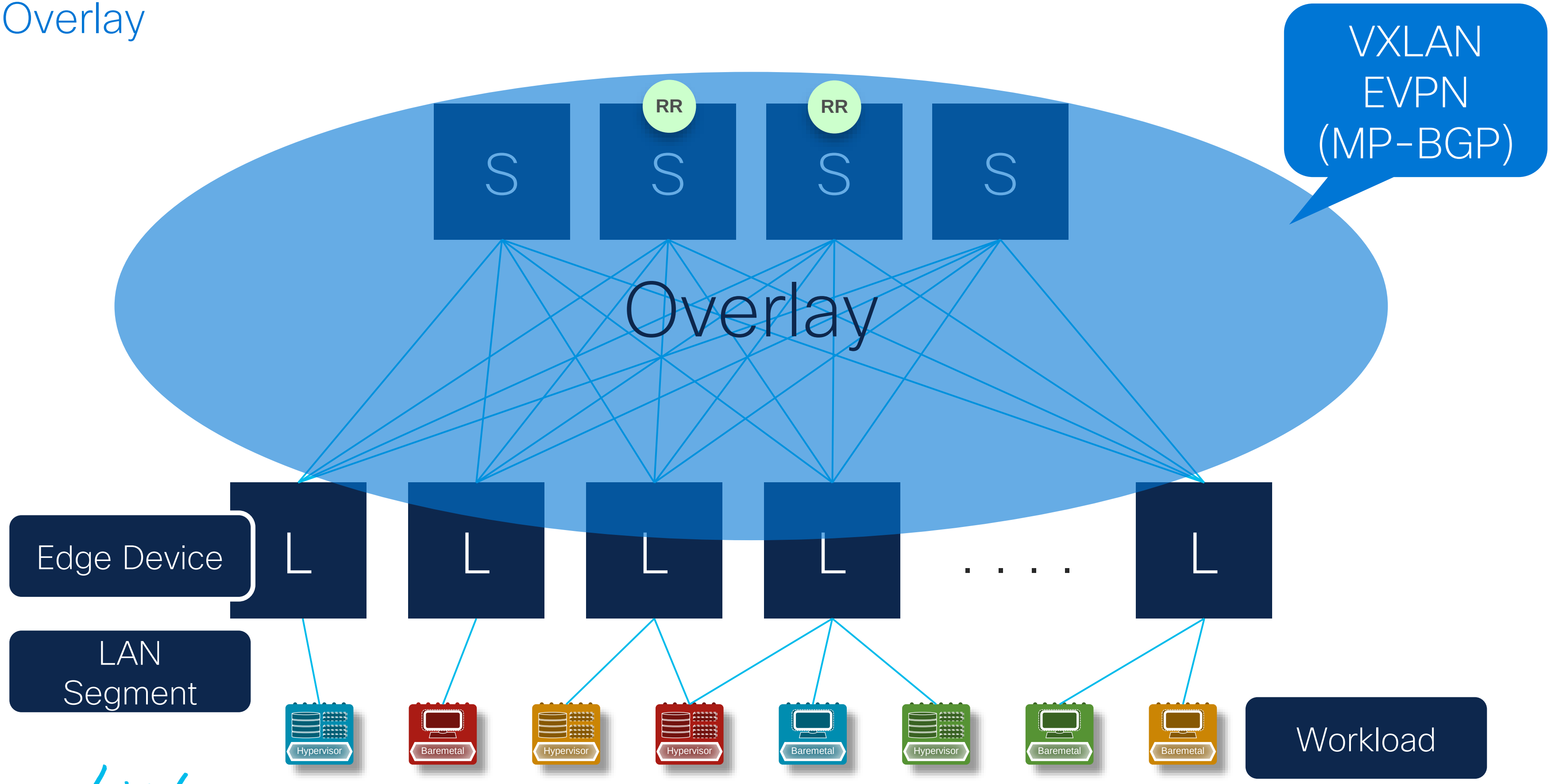
VXLAN: The Building Blocks

Underlay



VXLAN: The Building Blocks

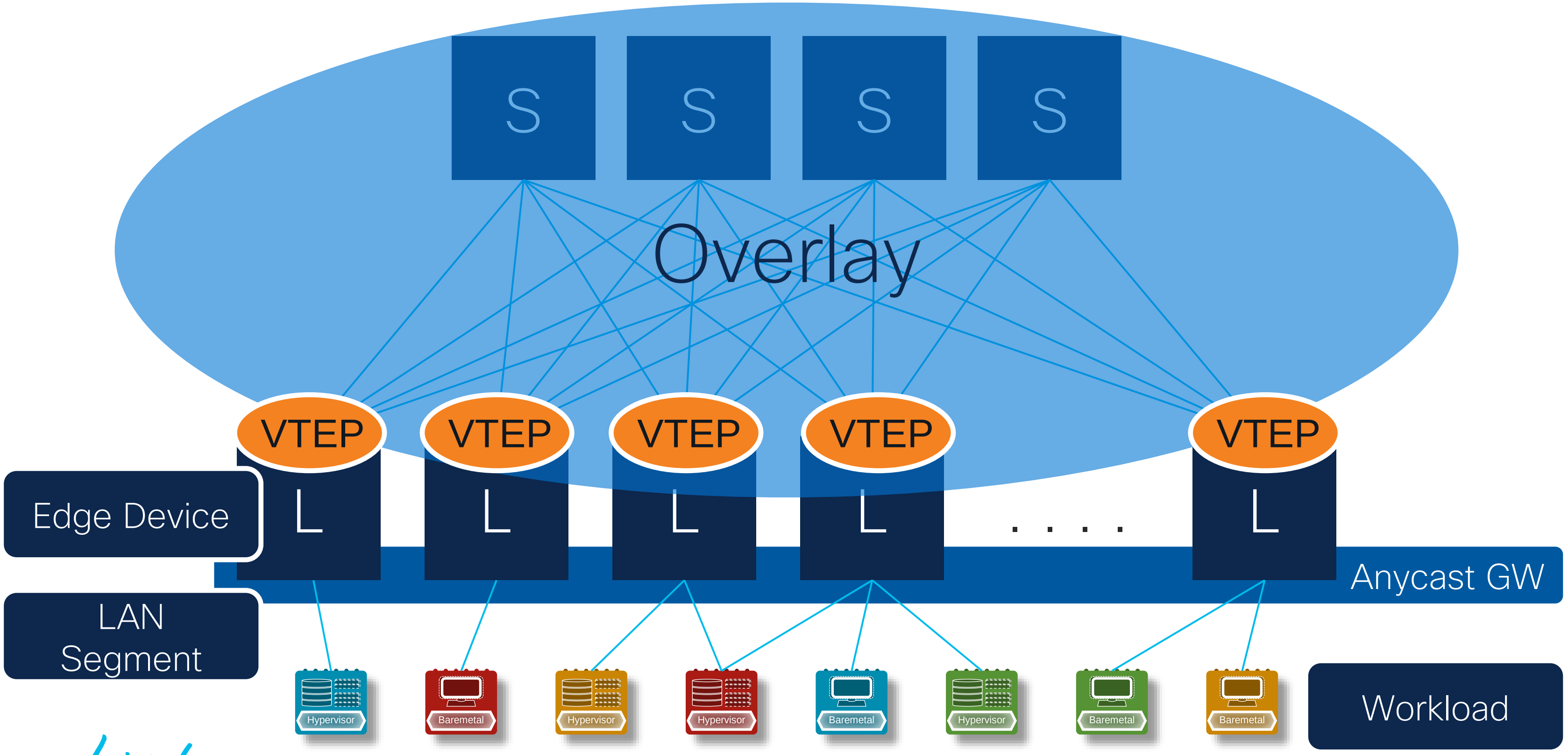
Overlay



VXLAN: The Building Blocks

Overlay

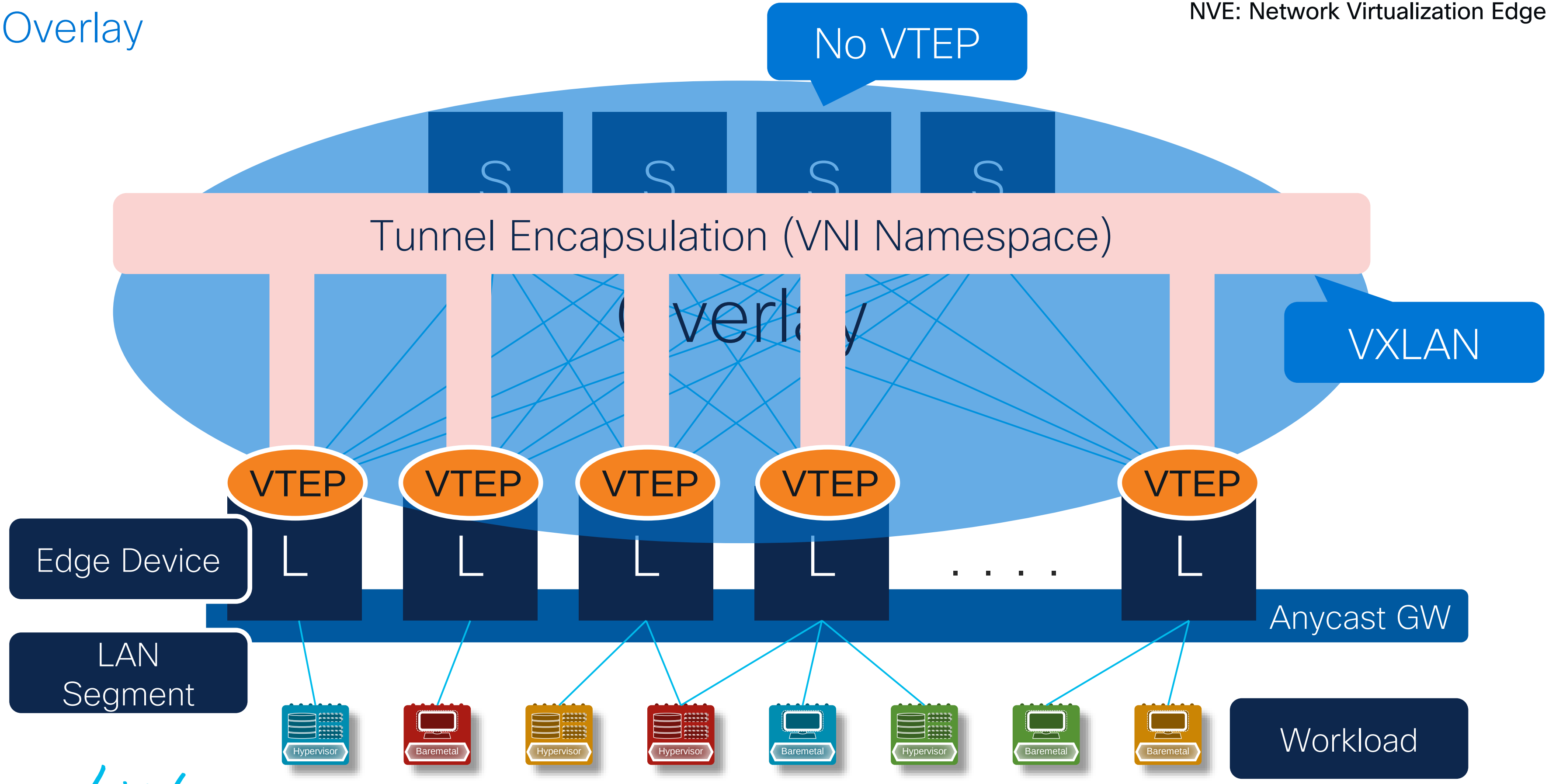
VTEP: VXLAN Tunnel End-Point
VNI/VNID: VXLAN Network Identifier
NVE: Network Virtualization Edge



VXLAN: The Building Blocks

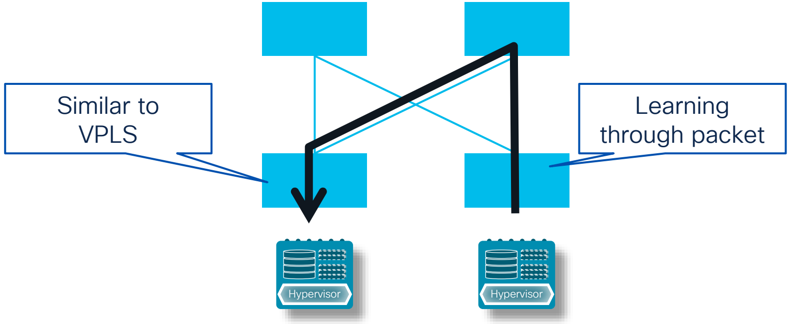
Overlay

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



What is Ethernet VPN?

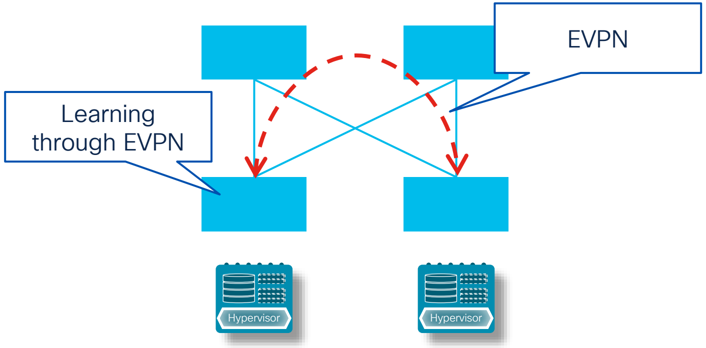
VXLAN only
No Control plane



- 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

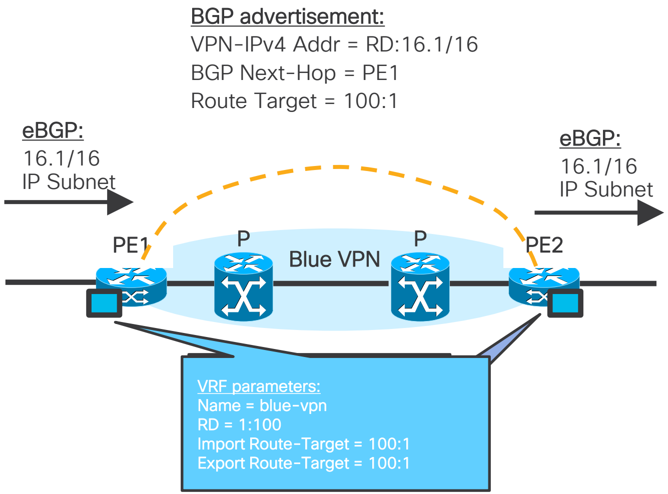
BGP EVPN Address Family

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

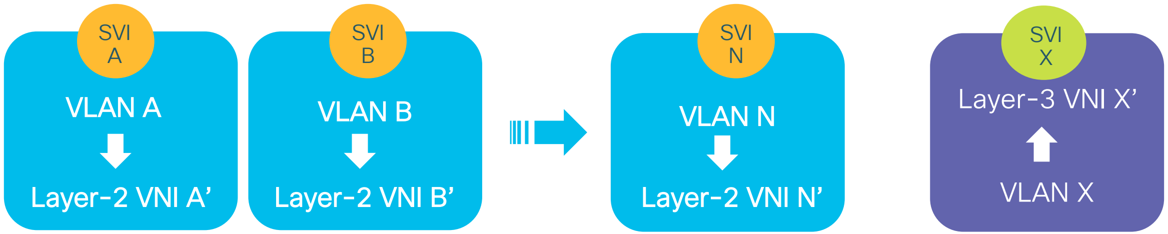
Route Target (RT): 8-byte field, VRF parameter, unique value to define the import/export rules for VPNv4 routes

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



Logical Construct of Multi-Tenant VXLAN EVPN

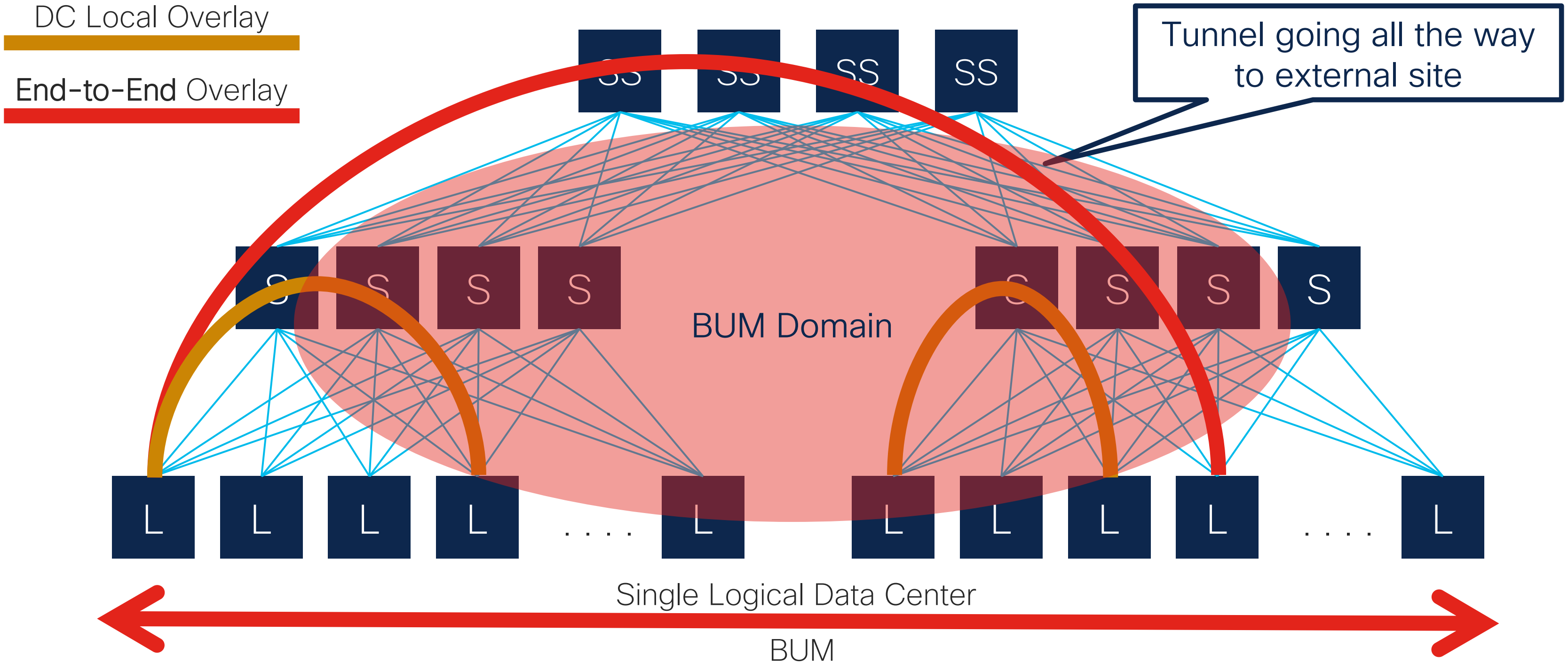
Tenant A (VRF A)



- 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

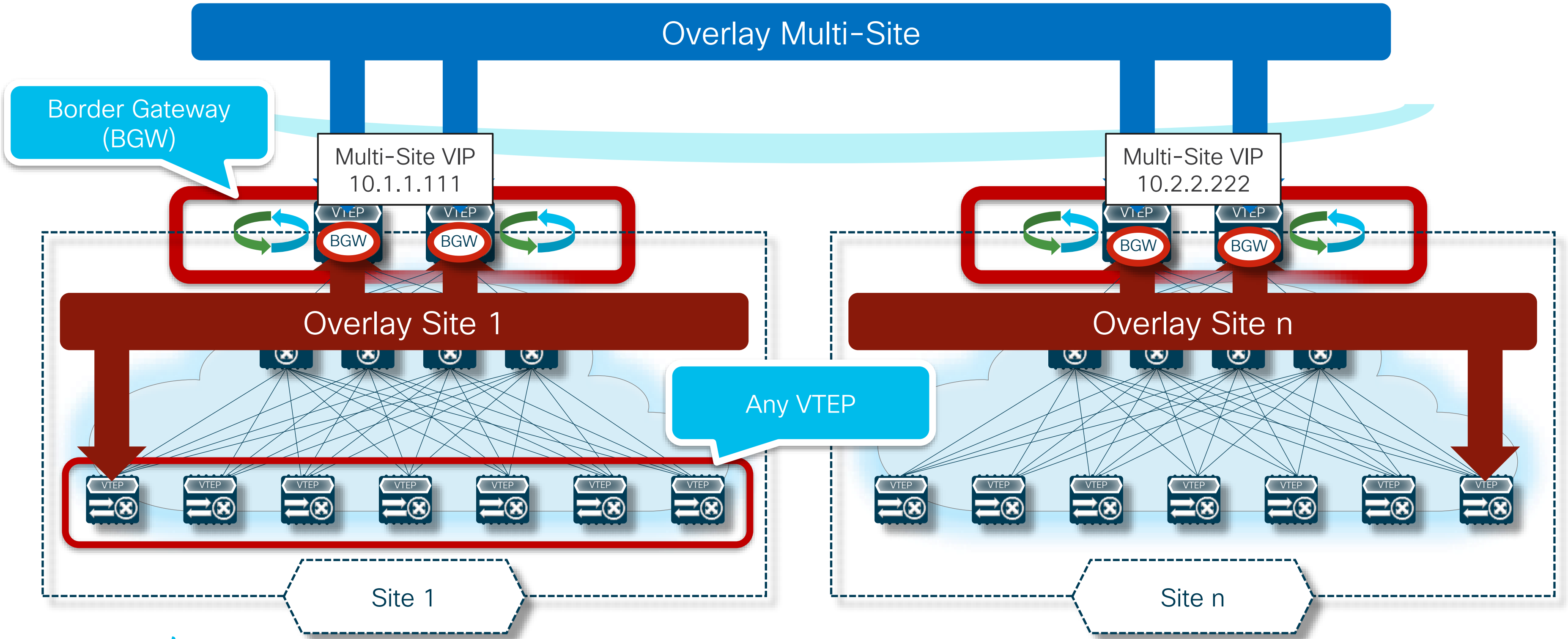
- 1 Layer-3 VNI per Tenant (VRF) for routing
- VNI X’ is used for routed packets

VXLAN Multi-Pod: Overlay Spread and Extend



VXLAN: The Building Blocks

Hierarchical VXLAN



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

Bandwidth/Cost Evolution Over a Decade

Pricing Economics

Nexus 5000 + FEX

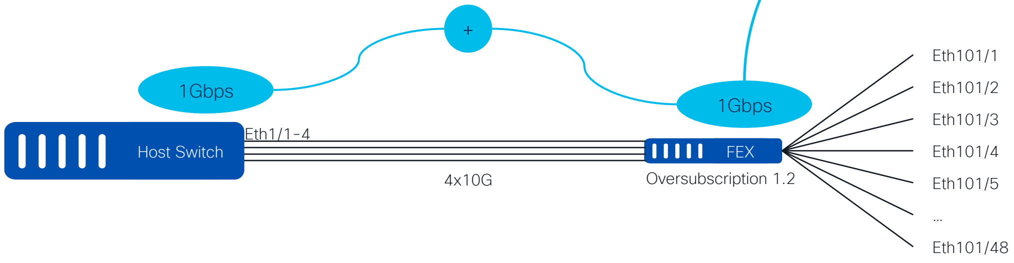
Before 2013

Host Switch	Price USD$	$ per Gbps
Cisco Nexus 5672UP-16G • 48 x 10G + 6 x 40G	USD $27,600.00	USD ~$38.00/Gbps
(48*10Gbps)+(6*40Gbps) =720Gbps/Switch	USD $27,600.00/720Gbps	

FEX	Price USD$	$ per Gbps
Cisco Nexus 2348UPQF-QSA • 48 x 1G + 6 x 10G	USD $24,300.00	USD $506.00/Gbps
48*1Gbps =48Gbps HIF per FEX	USD $24,300.00/48Gbps	

FEX Port $ per Gbps (1.2 Oversubscription)

USD ~$544.00/Gbps



Pricing Economics

Nexus 9000 + FEX

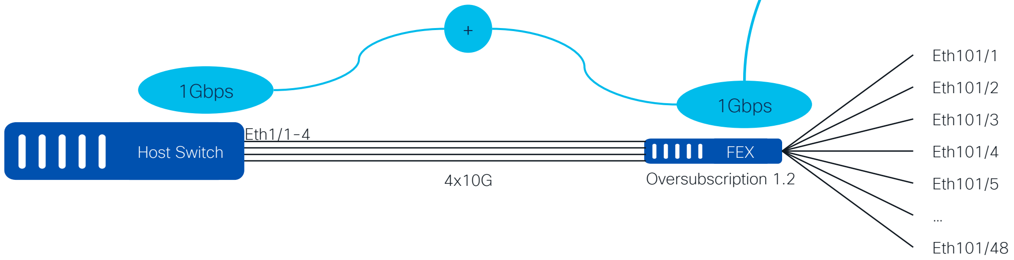
Around 2013

Host Switch	Price USD$	$ per Gbps
Cisco Nexus 93180YC-FX • 48 x 10G + 6 x 100G	USD $30,400.00	USD ~$28.00/Gbps
(48*10Gbps)+(6*100Gbps) =1080Gbps/Switch	USD $30,400.00/1080Gbps	

FEX	Price USD$	$ per Gbps
Cisco Nexus 2348UPQF-QSA • 48 x 1G + 6 x 10G	USD $24,300.00	USD $506.00/Gbps
48*1Gbps =48Gbps HIF per FEX	USD $24,300.00/48Gbps	

FEX Port $ per Gbps (1.2 Oversubscription)

USD ~$534.00/Gbps



Pricing Economics

Nexus 9000 + ToR

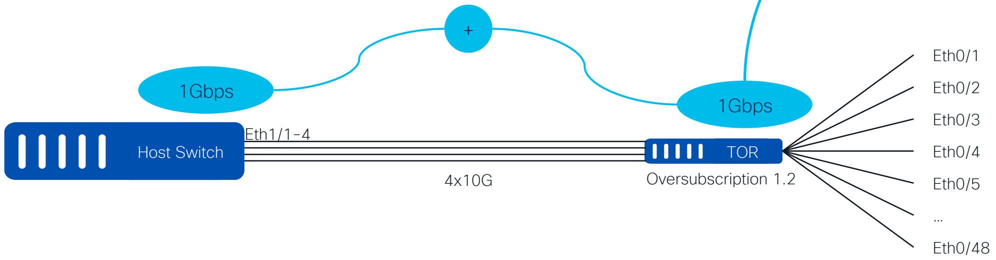
Today

Host Switch	Price USD$	$ per Gbps
Cisco Nexus 93180YC-FX • 48 x 10G + 6 x 100G	USD $30,400.00	USD ~$28.00/Gbps
(48*10Gbps)+(6*100Gbps) =1080Gbps/Switch	USD $30,400.00/1080Gbps	

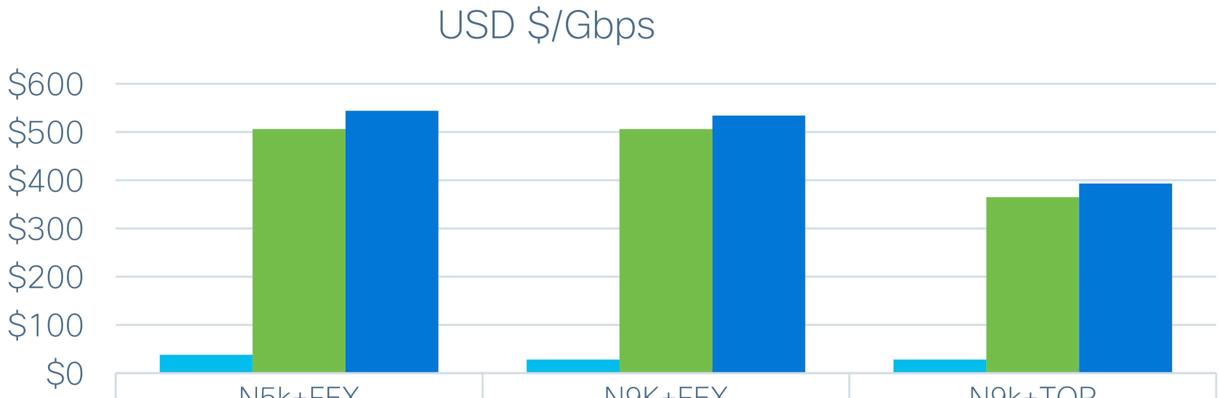
TOR	Price USD$	$ per Gbps
Cisco Nexus 9348GC-FXP • 48 x 1G + 2 x 100G	USD $17,500.00	USD ~$365.00/Gbps
48*1Gbps =48Gbps HIF per TOR	USD $17,500.00/48Gbps	

TOR Port $ per Gbps (1.2 Oversubscription)

USD ~$393.00/Gbps



Pricing Economics Comparison



	N5k+FEX	N9K+FEX	N9k+TOR
Switch Port	$38	$28	$28
FEX/TOR Port	$506	$506	$365
Total	$544	$534	$393

Bandwidth/Cost Change over a Decade

Pricing Economics

Today

Nexus 9000 + FEX with 1536 Ports @1Gbps

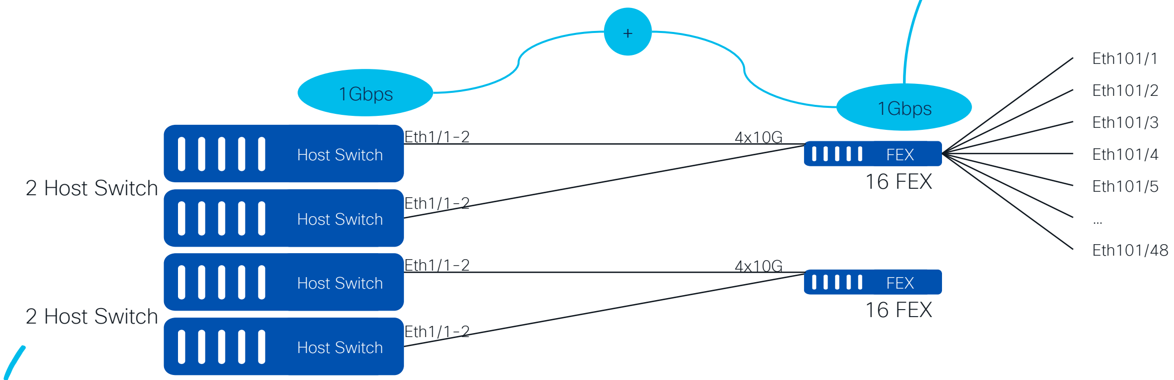
Host Switch	Price USD$ * 4 Host Switch	$ per Gbps
Cisco Nexus 93180YC-FX • 48 x 10G + 6 x 100G	USD $30,400.00	USD ~$28.00/Gbps
(48*10Gbps)+(6*100Gbps) =1080Gbps/Switch	USD $30,400.00/1080Gbps	

FEX	Price USD$ * 32 FEX	$ per Gbps
Cisco Nexus 2348UPQF-QSA • 48 x 1G + 6 x 10G	USD $24,300.00	USD $506.00/Gbps
48*1Gbps =48Gbps HIF per FEX	USD $24,300.00/48Gbps	

FEX Port $ per Gbps

USD ~$534.00/Gbps

USD ~$585.00/Host Port



Pricing Economics

Nexus 9000 Fabric with 1536 Ports @1Gbps

Today

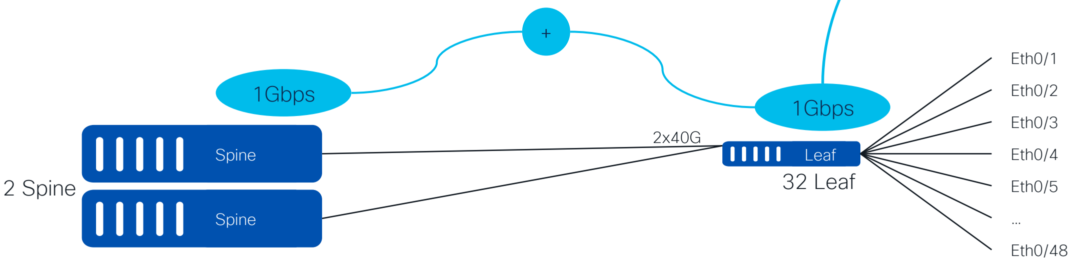
Spine	Price USD$ * 2 Spine	$ per Gbps
Cisco Nexus 9336C-FX2 • 36 x 100G	USD $38,800.00	USD ~$11.00/Gbps
36*100Gbps =3'600Gbps/Switch	USD $38,800.00/3600Gbps	

Leaf	Price USD$ * 32 Leaf	$ per Gbps
Cisco Nexus 9348GC-FXP • 48 x 1G + 2x 100G	USD $17,500.00	USD $365.00/Gbps
48*1Gbps =48Gbps HIF per TOR	USD $12,000.00/48Gbps	

TOR Port $ per Gbps

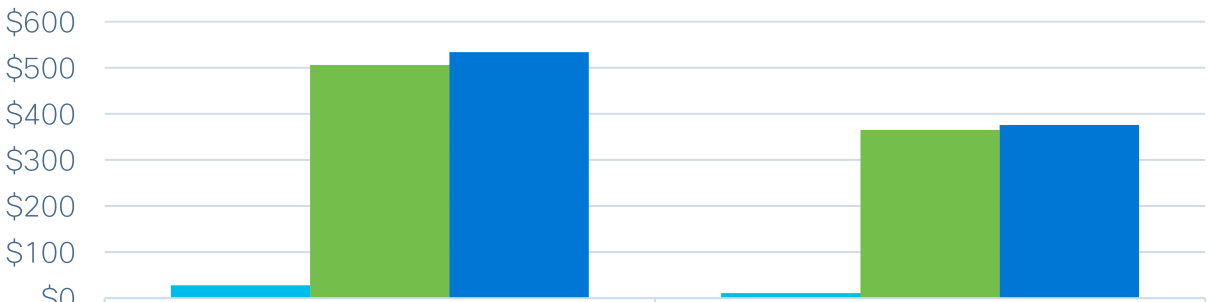
USD ~$376.00/Gbps

USD ~$415.00/Host Port



Pricing Economics Comparison with 1536 Ports @1Gbps

USD $/Gbps



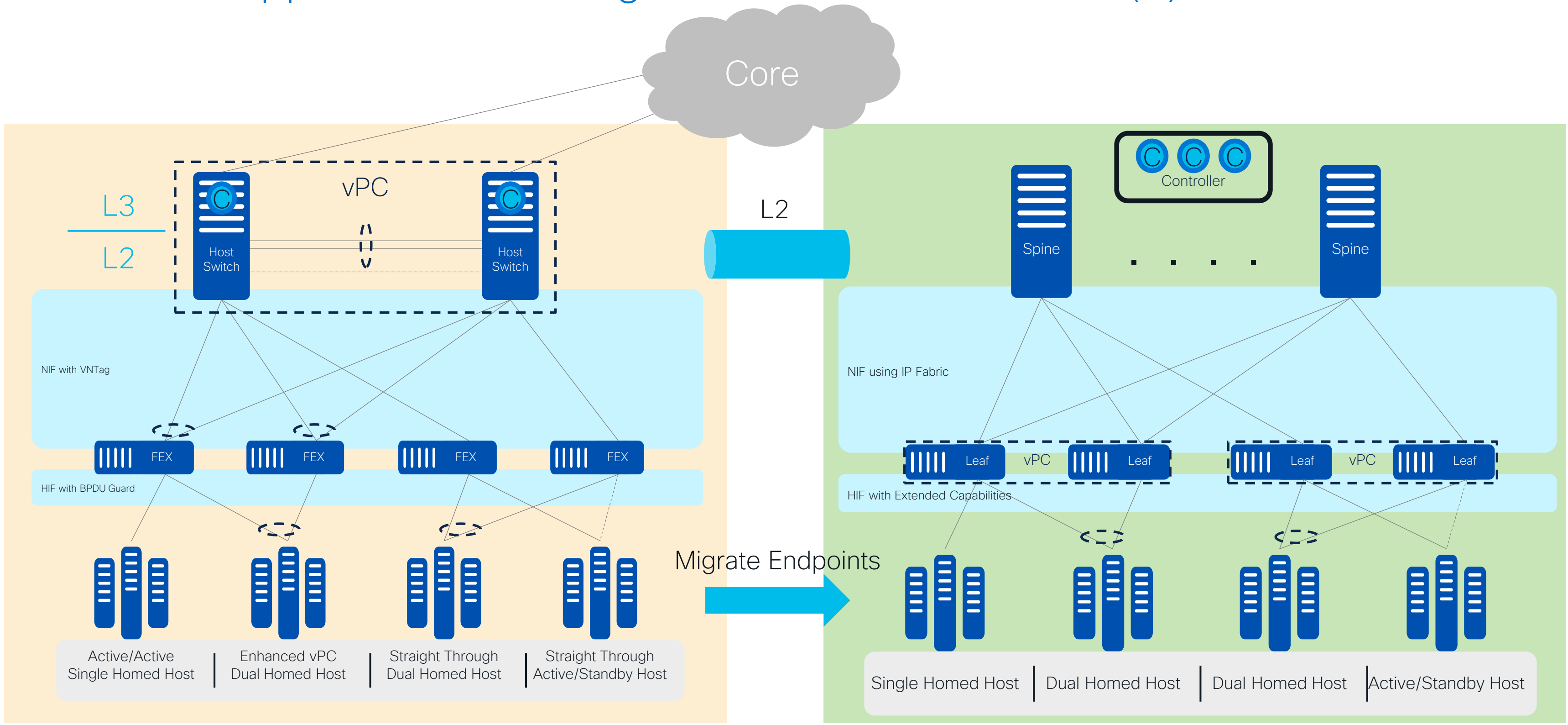
	N9k+FEX	N9k+TOR
Switch Port	$28	$11
FEX/TOR Port	$506	$365
Total	$534	$376

Optimizing Further with Port Count

Migration Considerations

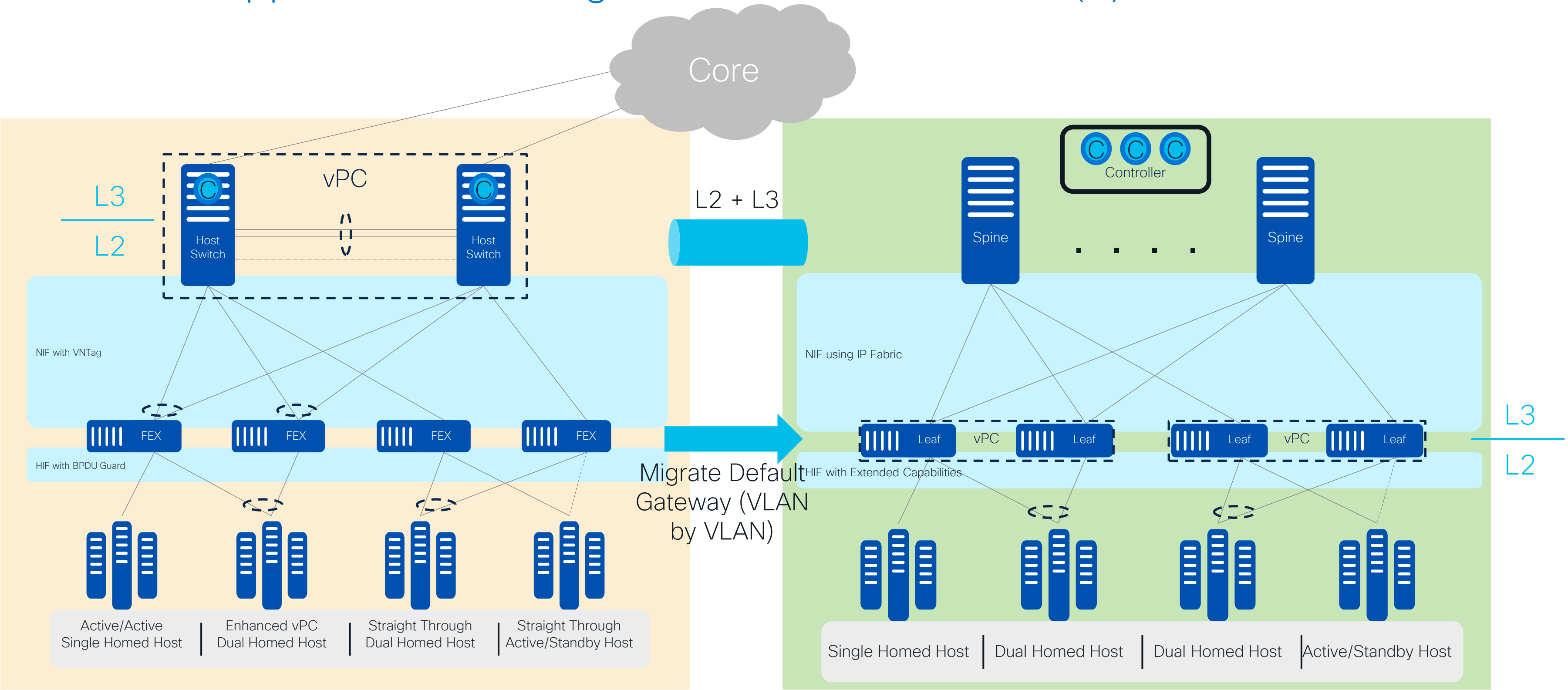
Migration Considerations

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



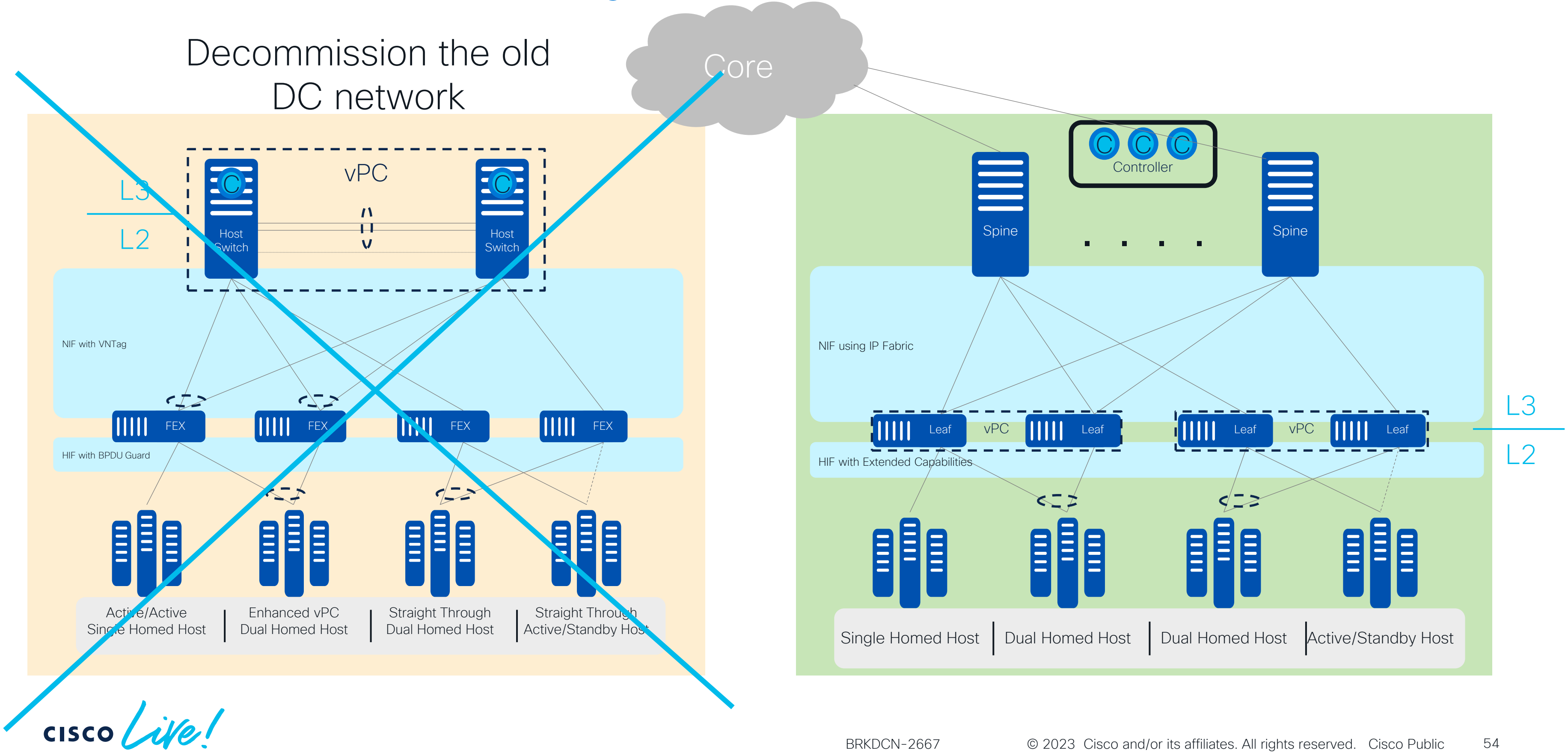
Migration Considerations

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



Migration Considerations

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

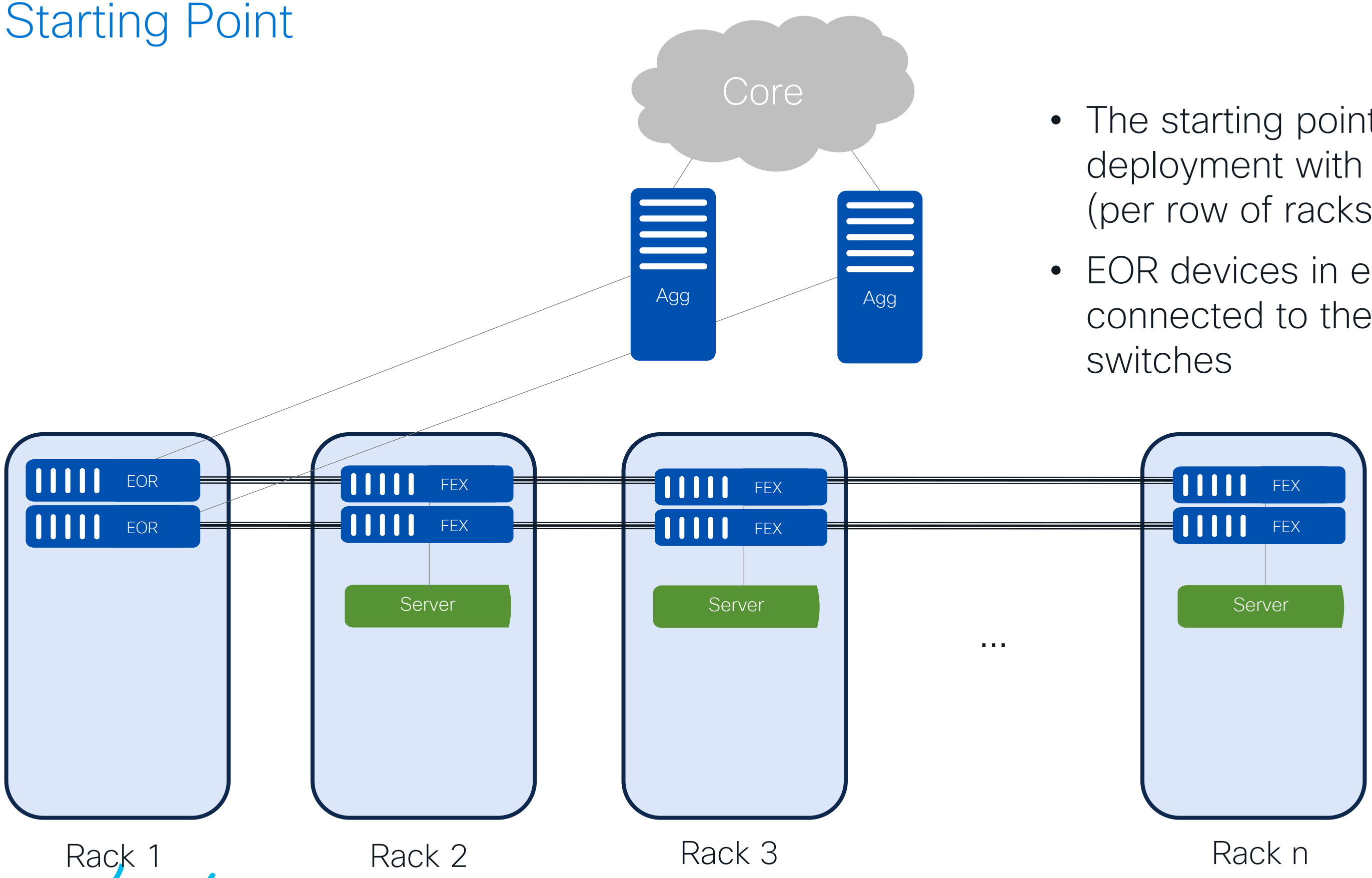


Migration with Rack Space Constraints

Migration with Rack Space Constraints

1

Starting Point

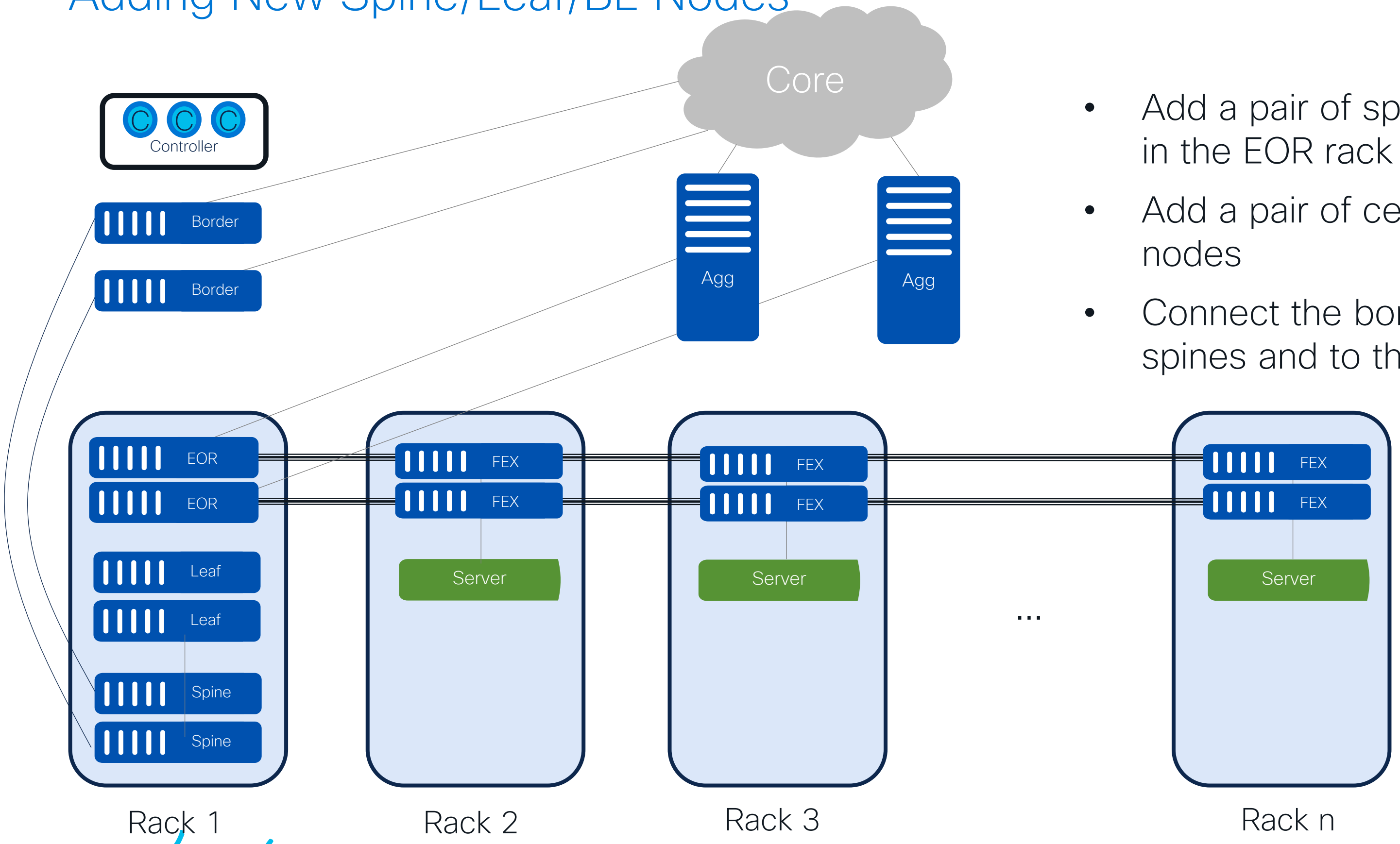


- The starting point is the traditional FEX deployment with a pair of EOR devices (per row of racks)
- EOR devices in each row are connected to the centralized Agg switches

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

2

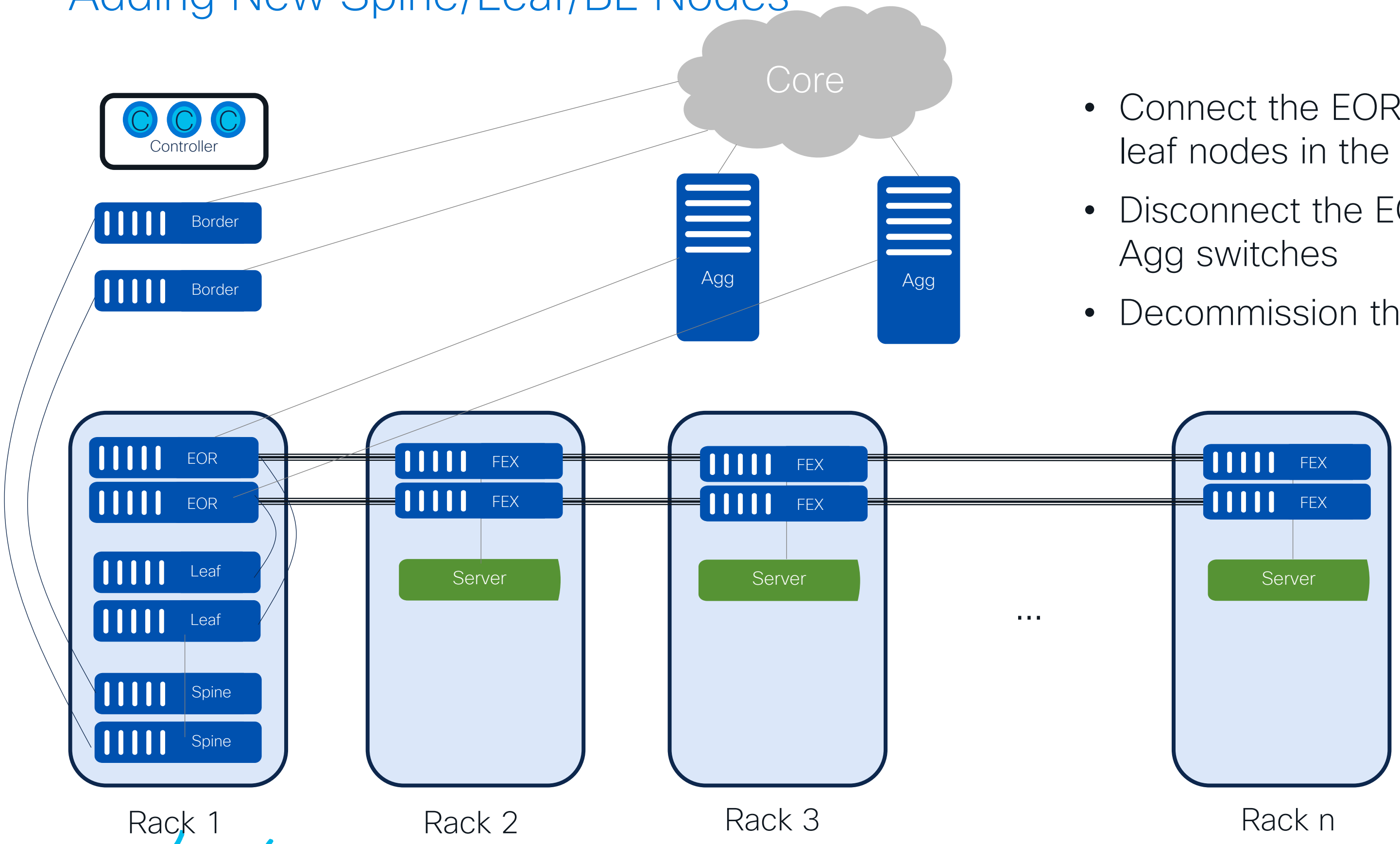


- Add a pair of spine and leaf switches in the EOR rack
- Add a pair of centralized border nodes
- Connect the border nodes to the spines and to the core

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

3

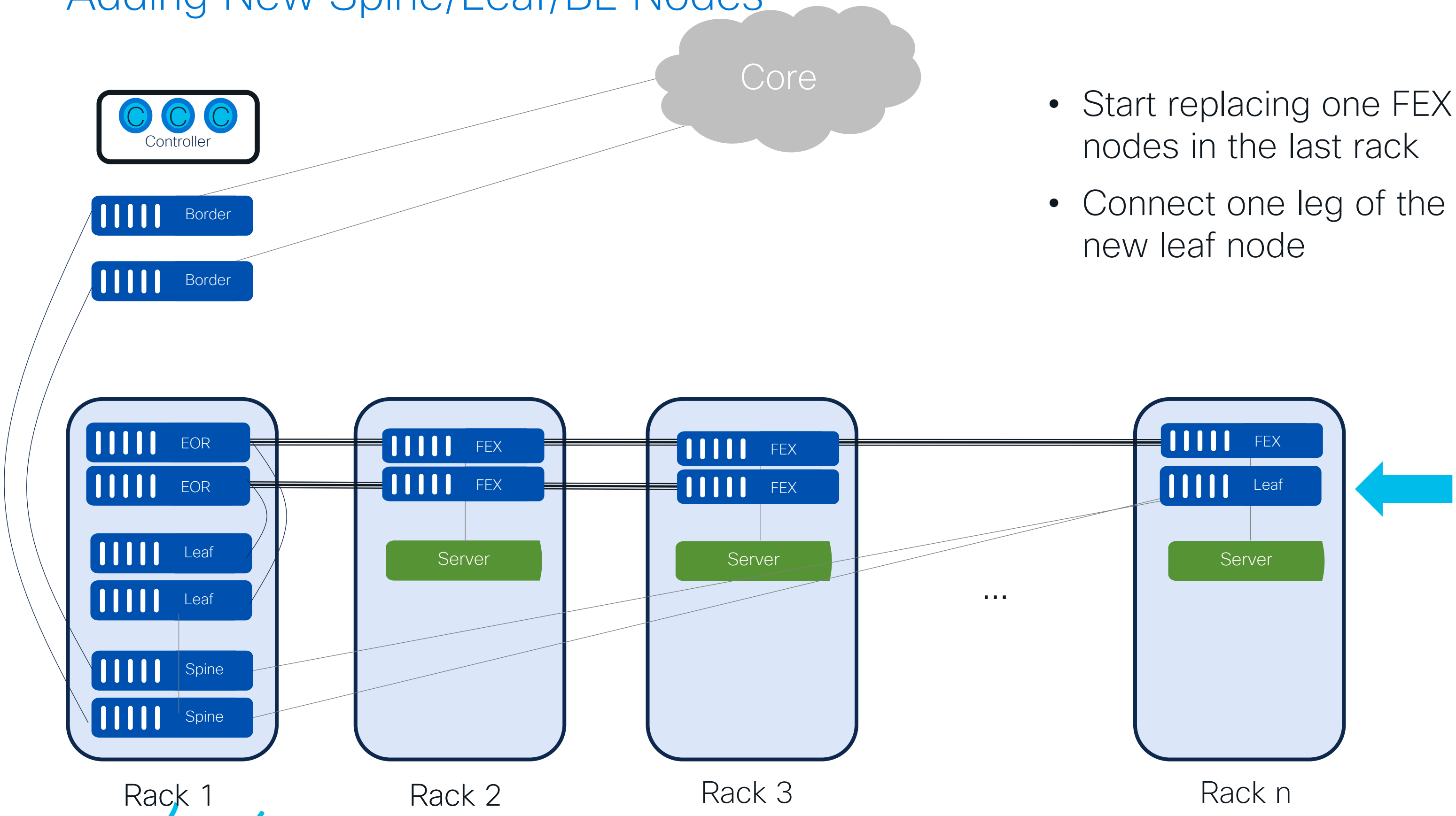


- Connect the EOR devices to the pair of leaf nodes in the EOR rack (L2 + L3)
- Disconnect the EOR devices from the Agg switches
- Decommission the Agg switches

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

4

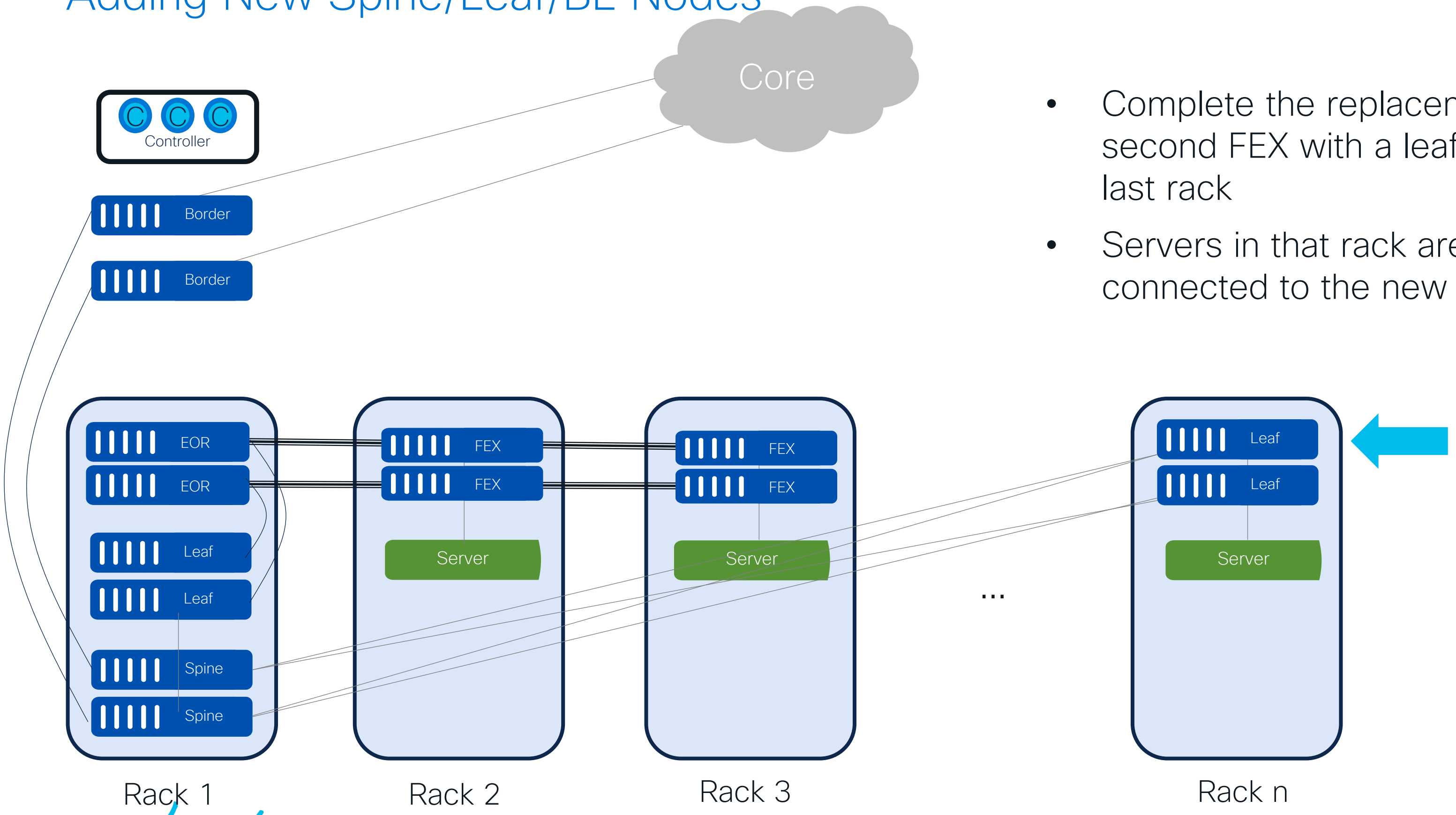


- Start replacing one FEX with a new leaf nodes in the last rack
- Connect one leg of the servers to the new leaf node

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

5

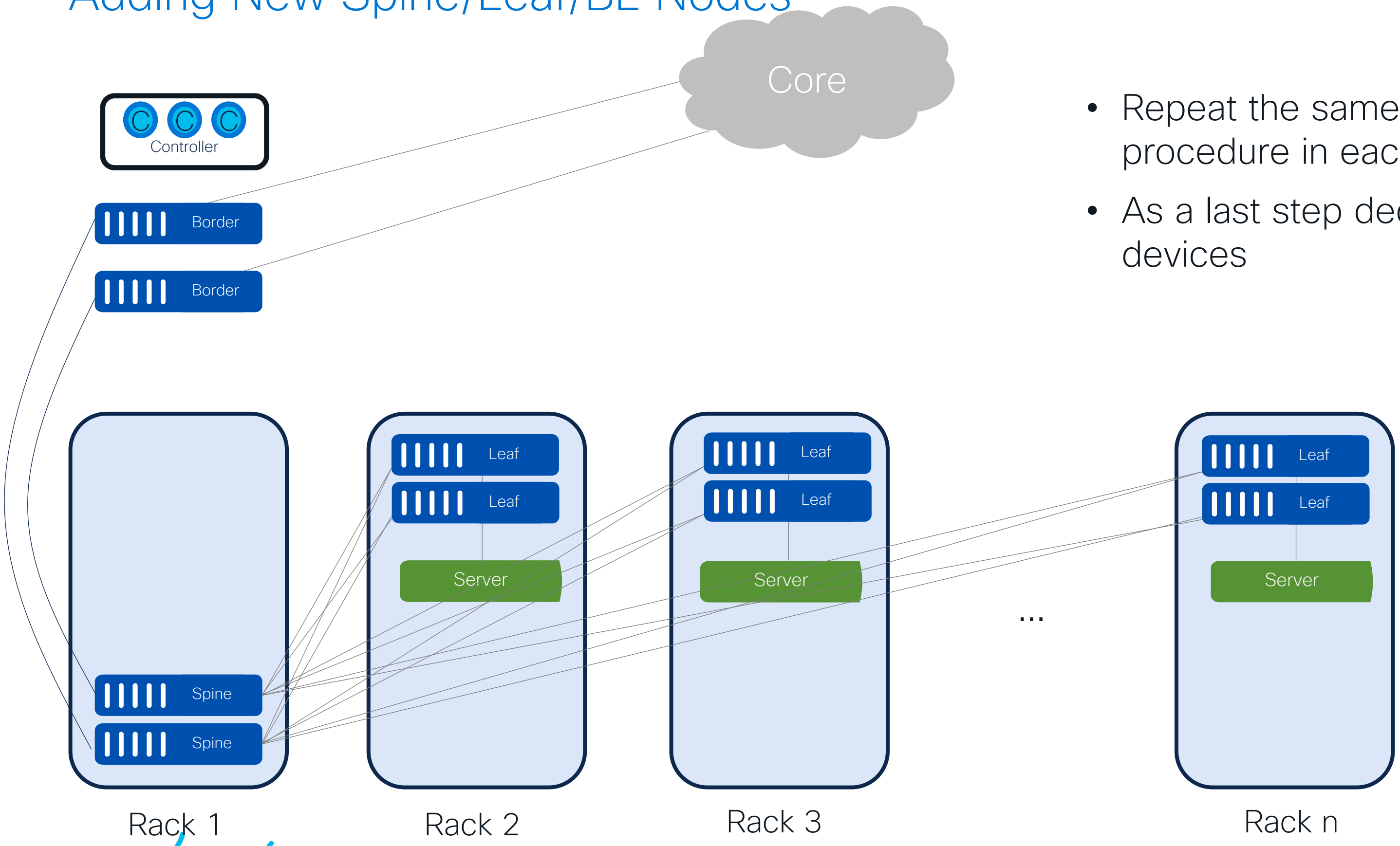


- Complete the replacement of the second FEX with a leaf node in the last rack
- Servers in that rack are now only connected to the new fabric

Migration with Rack Space Constraints

Adding New Spine/Leaf/BL Nodes

6



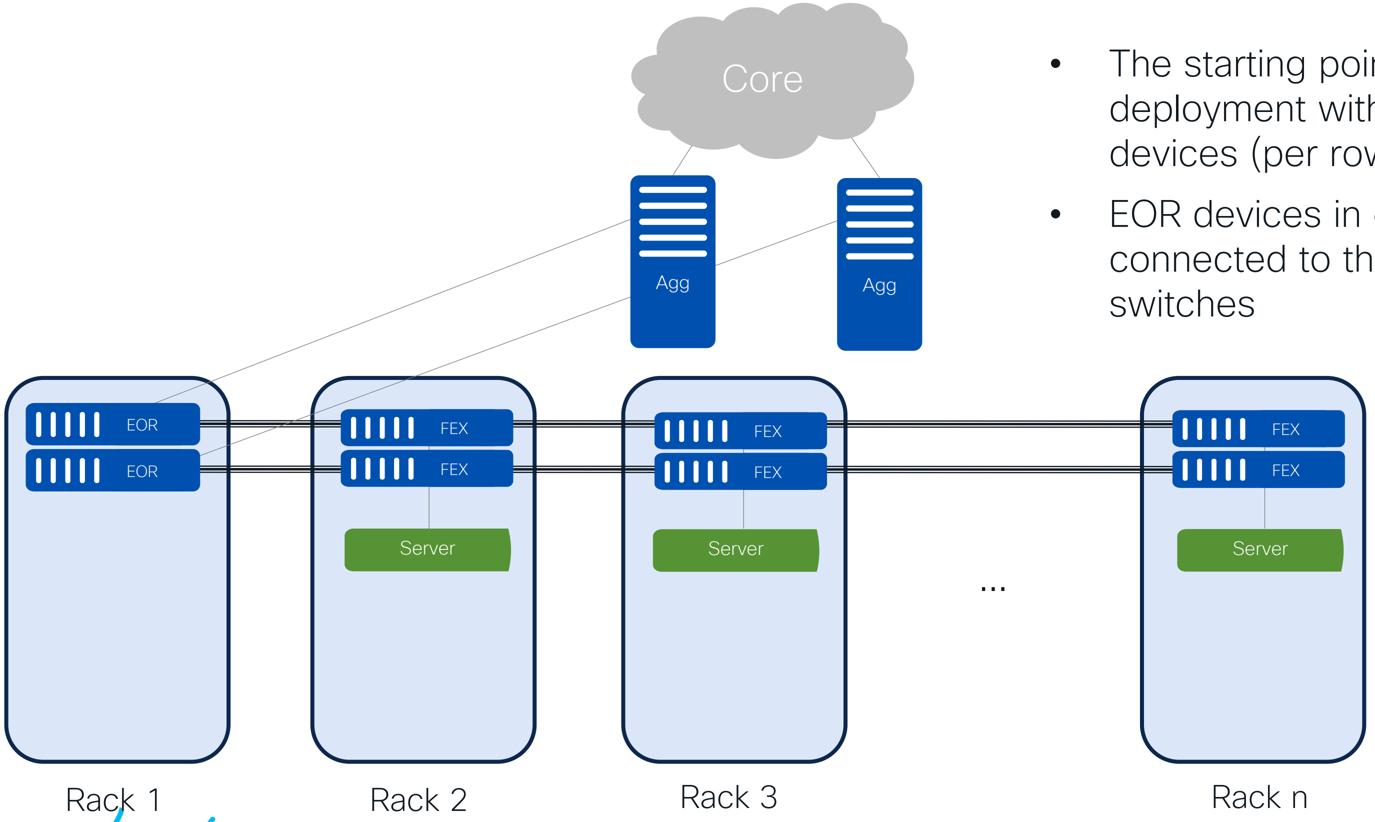
- Repeat the same FEX replacement procedure in each rack
- As a last step decommission the EOR devices

Migration without Rack Space Constraints

Migration without Rack Space Constraints

1

Starting Point

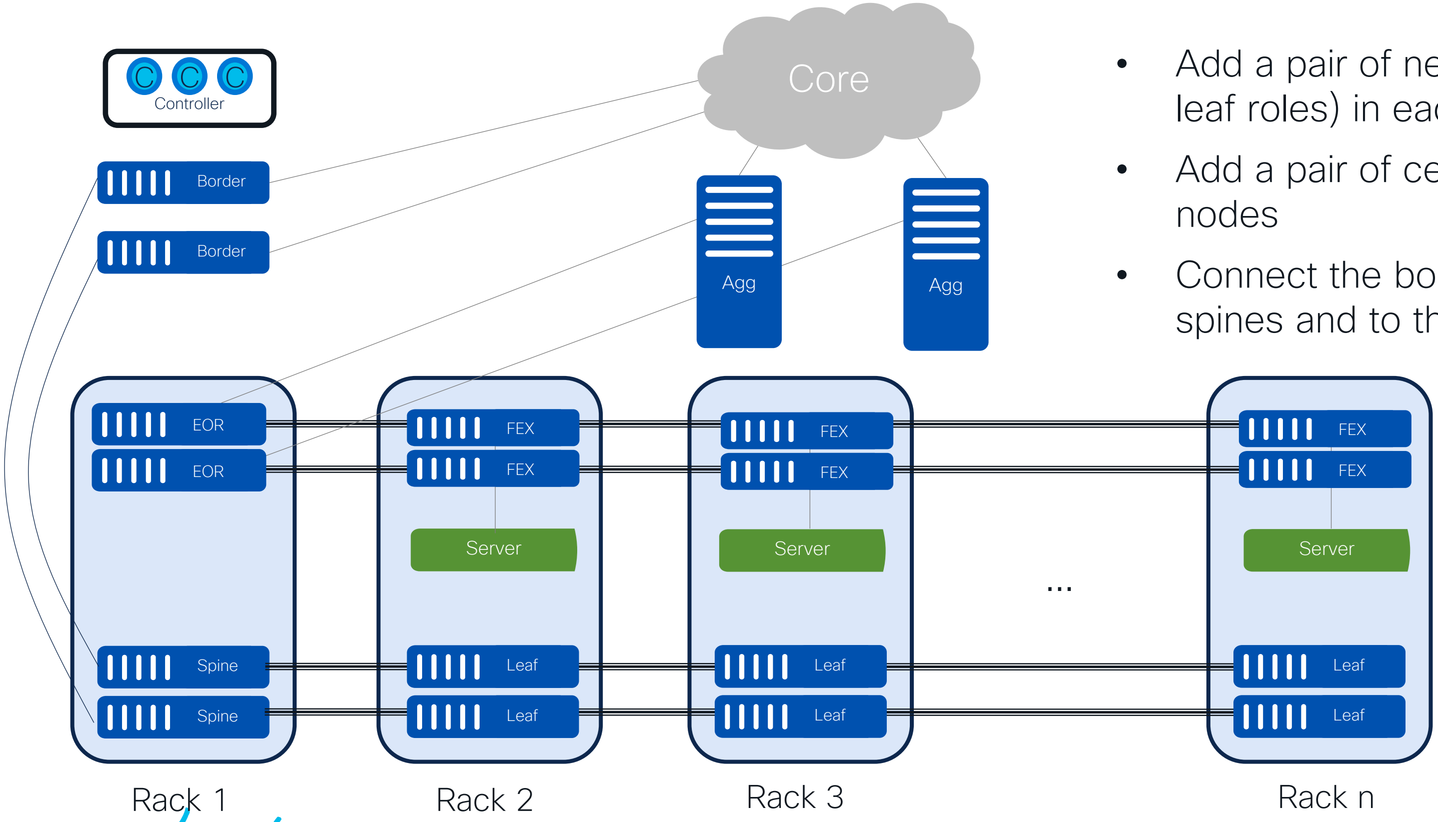


- The starting point is the traditional FEX deployment with a pair of EOR devices (per row of racks)
- EOR devices in each row are connected to the centralized Agg switches

Migration without Rack Space Constraints

2

Adding New Spine/Leaf/BL Nodes

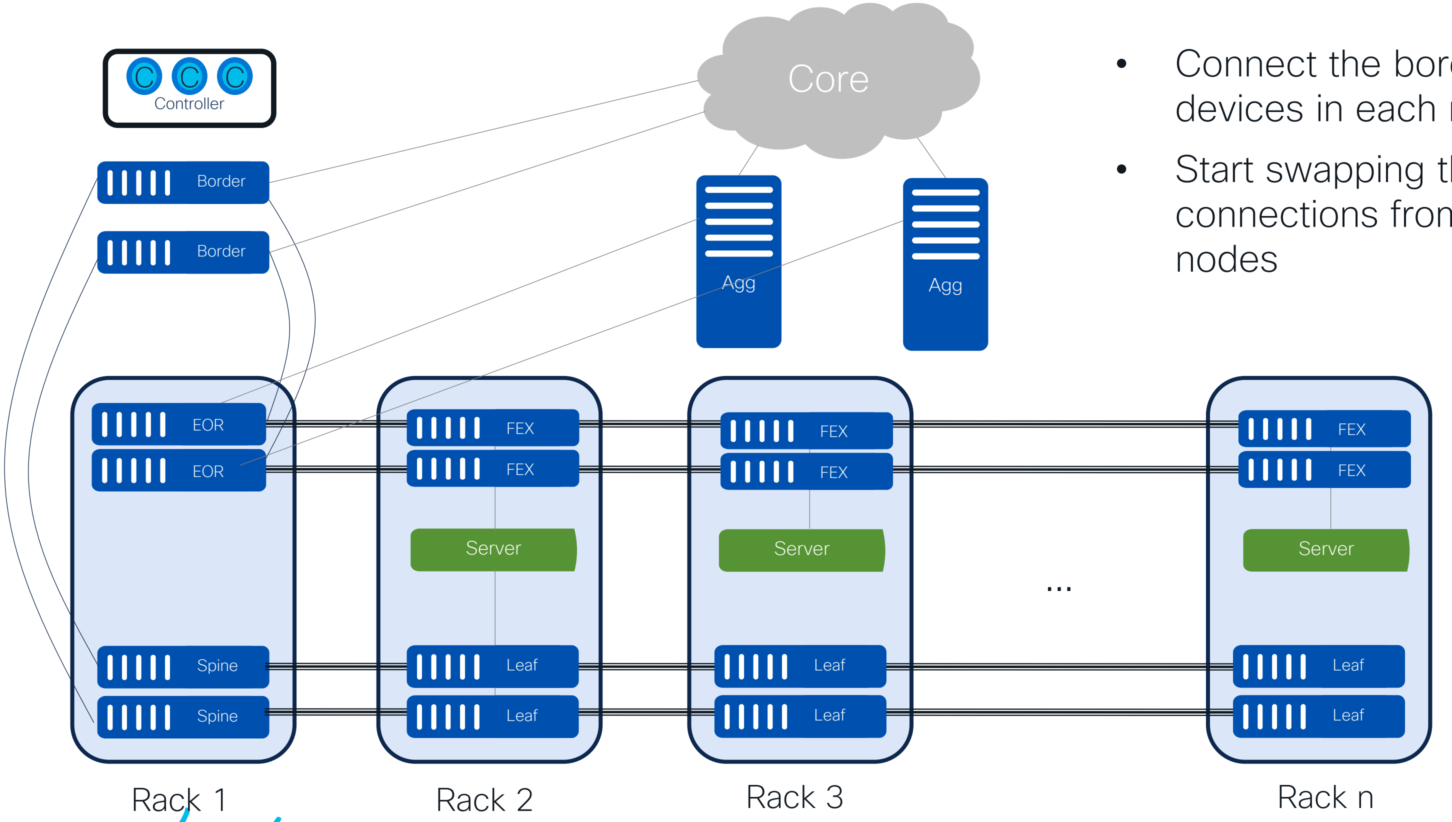


- Add a pair of new devices (spine and leaf roles) in each rack
- Add a pair of centralized border nodes
- Connect the border nodes to the spines and to the core

Migration without Rack Space Constraints

3

Connect Old and New Infrastructures and Migrate Endpoints

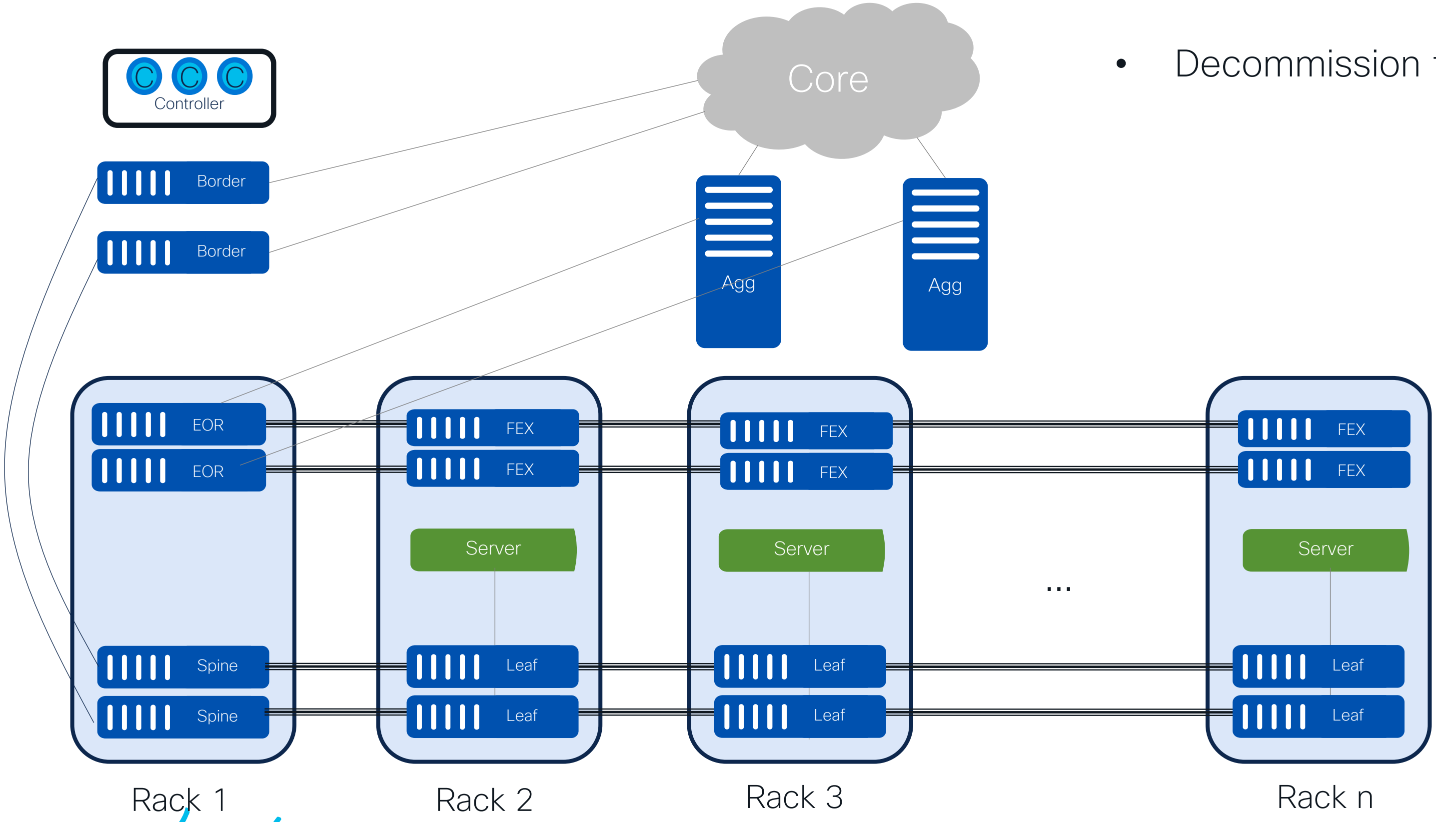


- Connect the border nodes to the EOR devices in each row (L2 and L3)
- Start swapping the servers connections from the FEX to the leaf nodes

Migration without Rack Space Constraints

4

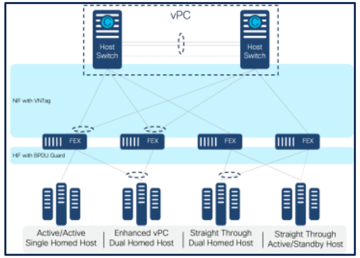
Decommission the Old Infrastructure



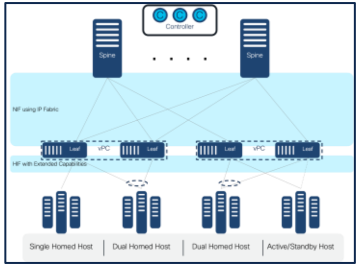
- Decommission the old infrastructure

Conclusion

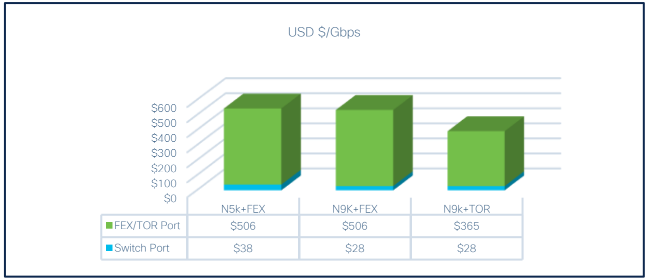
Conclusions



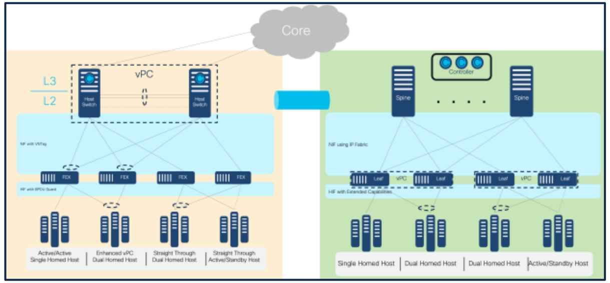
- FEX was the first attempt to build a fabric infrastructure
  - Centralized Management
  - Network and Host Redundancy



- 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

Complete your Session Survey

- Please complete your session survey after each session. Your feedback is very important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (open from Thursday) to receive your Cisco Live t-shirt.
- All surveys can be taken in the Cisco Events Mobile App or by logging in to the Session Catalog and clicking the "Attendee Dashboard" at https://www.ciscolive.com/emea/learn/sessions/session-catalog.html

Continue Your Education

Visit the Cisco Showcase for related demos.

Book your one-on-one Meet the Engineer meeting.

Attend any of the related sessions at the DevNet, Capture the Flag, and Walk-in Labs zones.

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

The bridge to possible

Thank you

cisco Live!

ALL IN