
cisco Live!
Let's go

The bridge to possible

Multi-Tier Fabrics

Network Designs for the Modern Data Center

Max Ardica, Distinguished Engineer
@maxardica

cisco Live!

BRKDCN-2999

Abstract

Have you ever asked yourself what "Clos" is or where that Leaf/Spine thing comes from? If yes, this is the right session for you. We are going to cover Fat-Tree, Clos and Leaf/Spine designs and expand beyond just the Spine layer. We will spend some time on the Super-Spine and even Super-Spine fabrics. How you can cost effectively use 100G/400G and where fixed vs. modular Switches make sense.

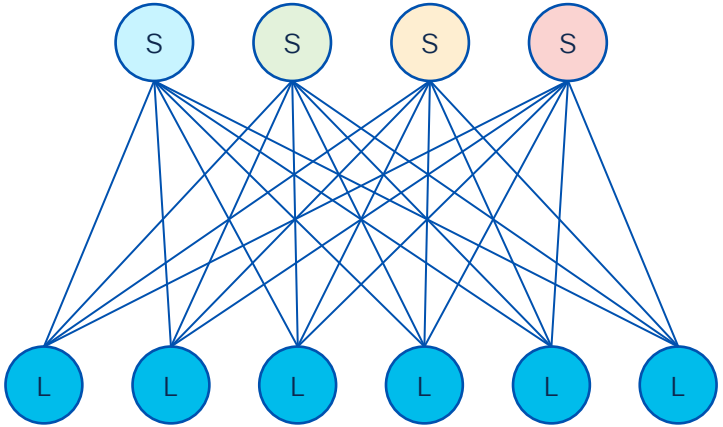
Agenda

- Introduction
- Paradigm and Fundamentals
- Data Center Design Evolution
  - ACI Fabrics
  - VXLAN EVPN Fabrics
  - Heterogeneous Fabrics
  - Routed Fabrics
- Conclusion

Paradigm and Fundamentals

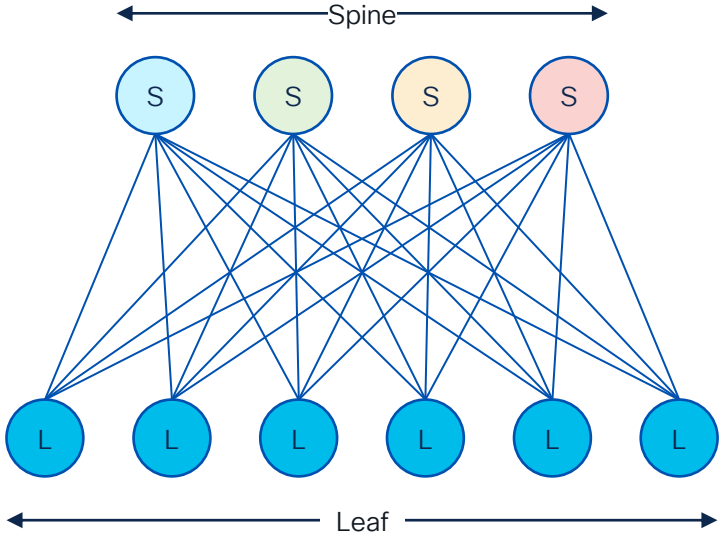
The Paradigm

- A Leaf and Spine Topology



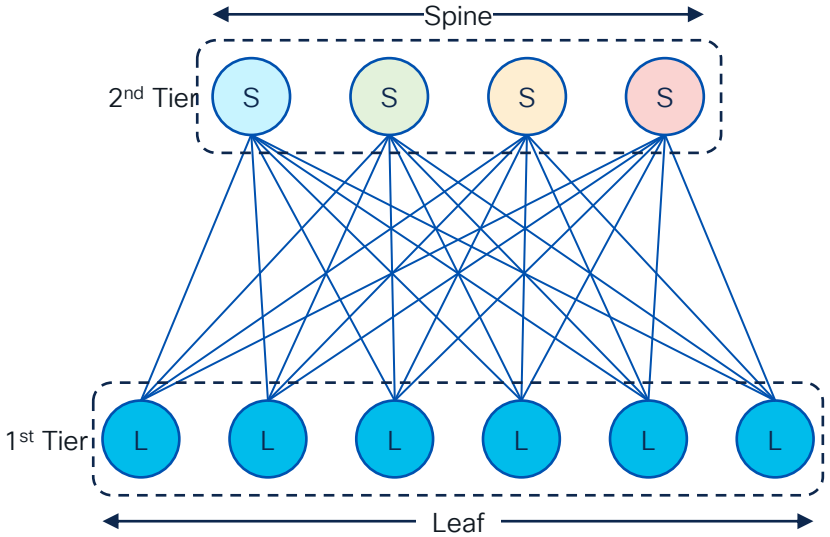
The Paradigm

- A Leaf and Spine Topology
- Variations or Names of the same:
  - Fat Tree
  - Folded Clos
  - 3 Stage Clos
  - 2 Tier Network



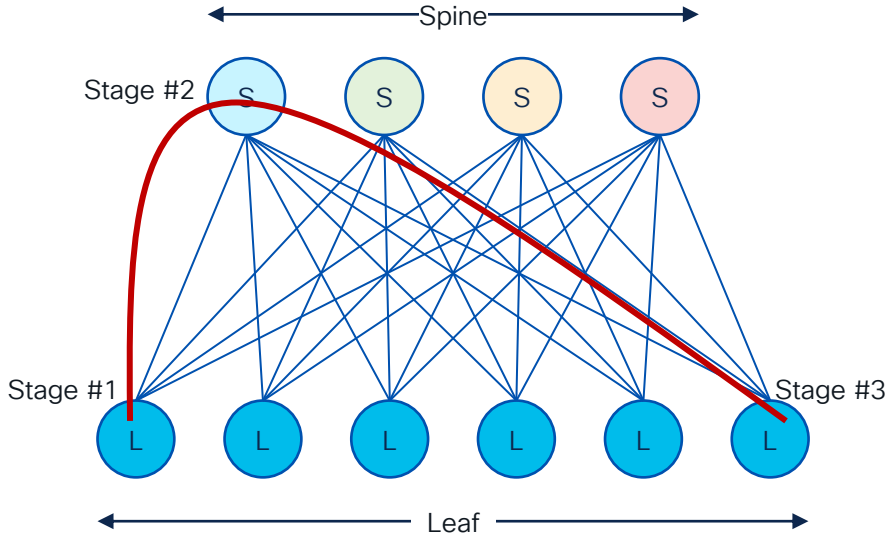
The Paradigm

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



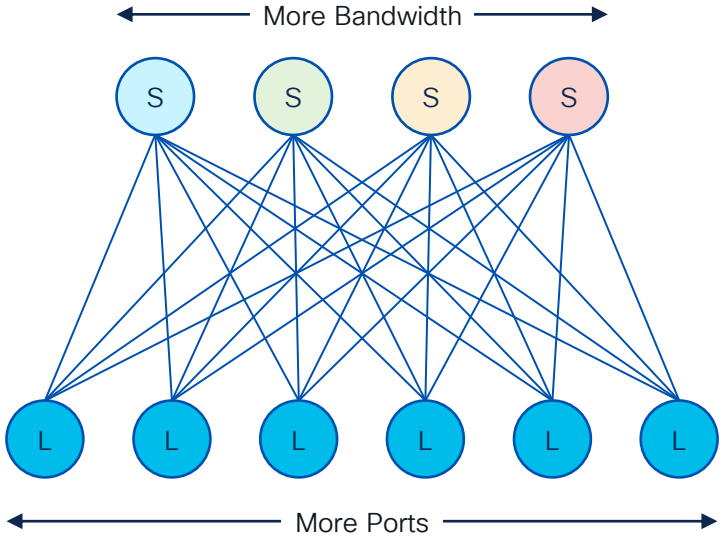
The Paradigm

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



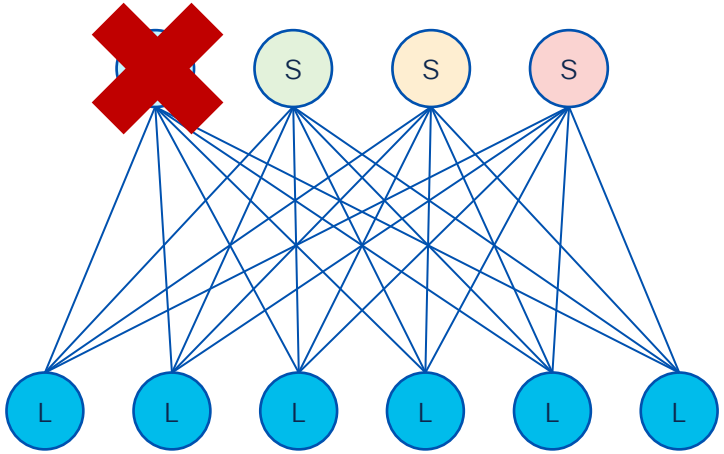
The Paradigm

- A Scale Out Architecture
  - More Leaf = More Ports
  - More Spine = More Bandwidth



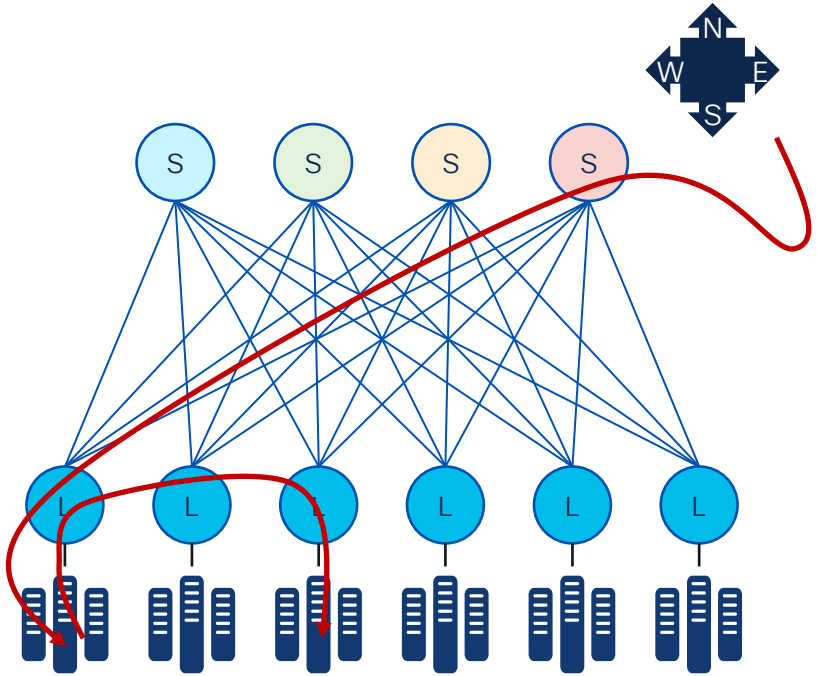
The Paradigm

- N+1 Redundancy
- Redundancy increases by Building out the Topology
- On Spine failure
  - 4 Spine = 25% impact
  - 8 Spine = 12.5% impact



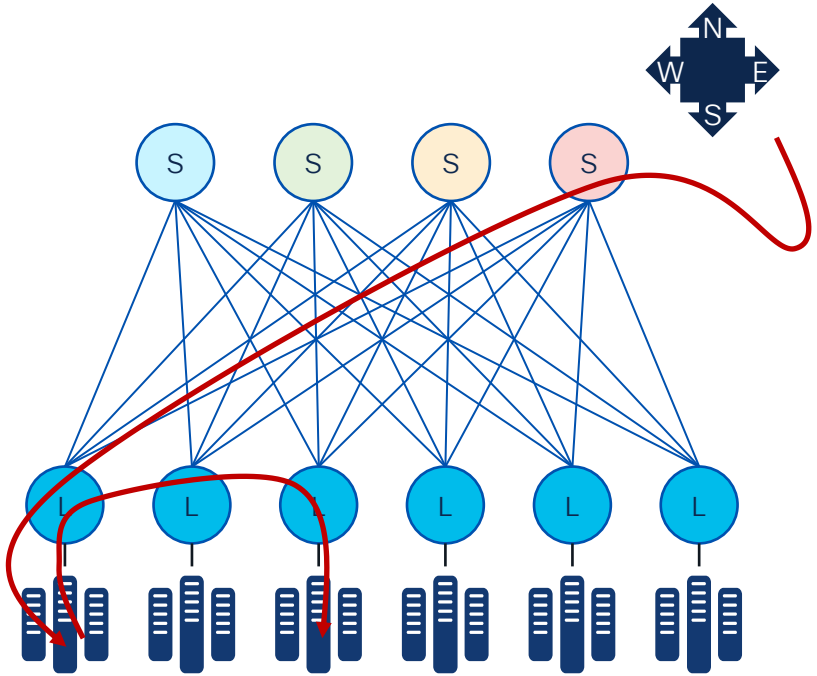
The Paradigm

- Modern Application Needs
  - Every (1) North to South Connection, requires eight (8) East to West
  - User Access the Frontend (Web)
  - Frontend connects to App, DB, Storage etc.



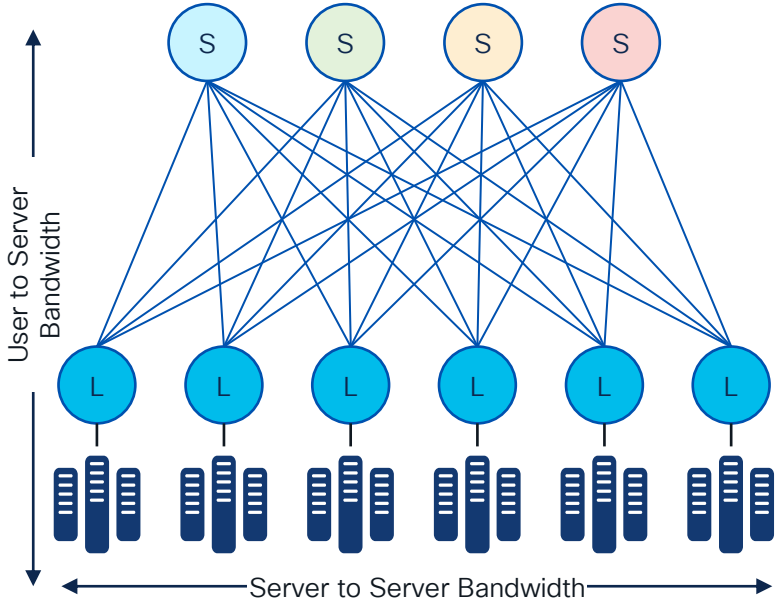
The Paradigm

- Optimized for East to West
  - Consistent Latency from Leaf to Leaf
  - Wide ECMP
- Flexibility for North to South
  - External Connectivity at Leaf or Spine layer



The Paradigm

- Bandwidth Requirements
- Oversubscription



‘How Many Spines do I need?’

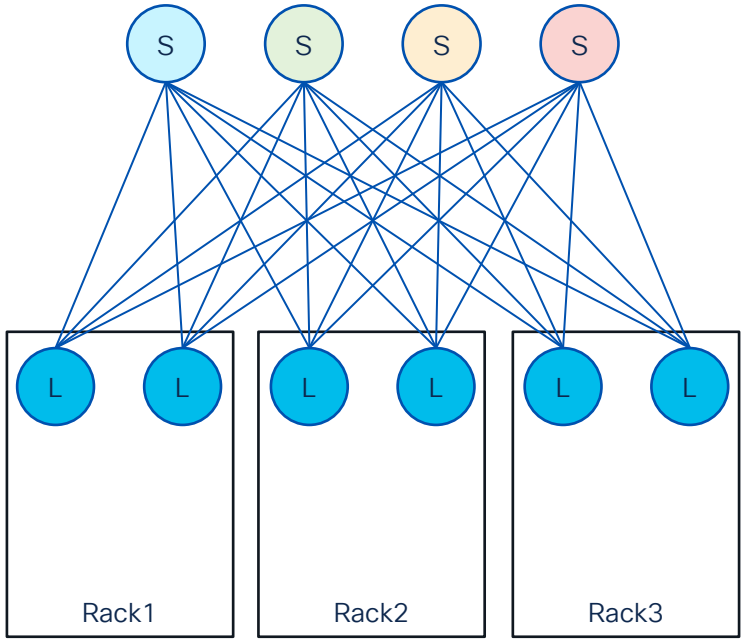
It Depends

How Many Spines do I need?

Oversubscription and Maximum Redundancy as the Criteria

Host Attachment Requirements

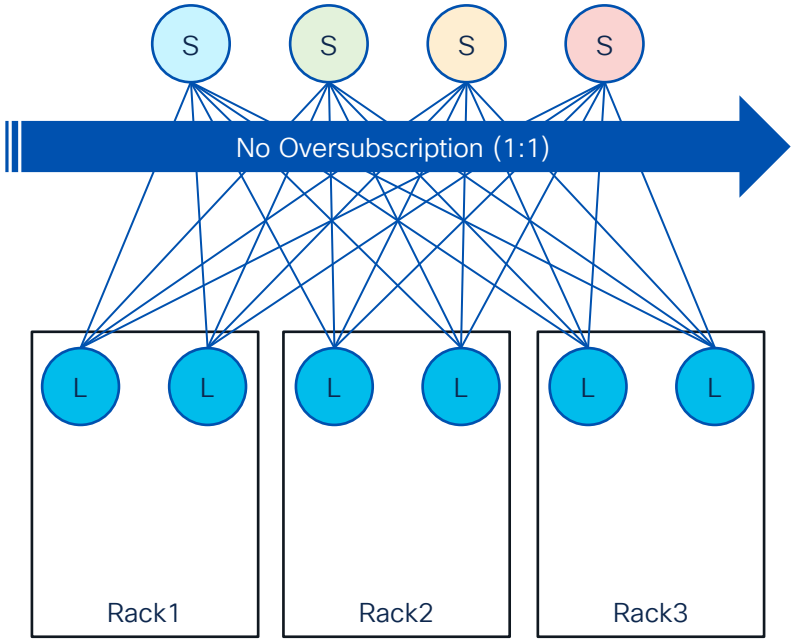
- 48 Server per Rack
- 2x 25Gbps NIC per Server
- 1x NIC per Switch



How Many Spines do I need?

Oversubscription and Maximum Redundancy as the Criteria

Host Attachment Requirements
- 48 Server per Rack
- 2x 25Gbps NIC per Server
- 1x NIC per Switch

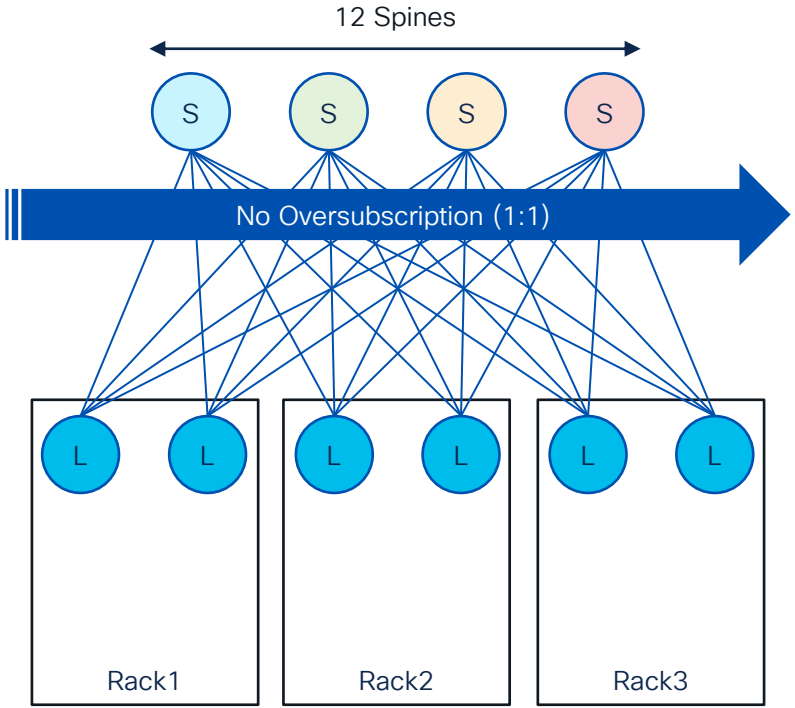


Resulting Uplink Requirements
- 48x 25Gbps per Leaf
- 1.2Tbps Uplink from Leaf to Spine
- 12x 100Gbps towards Spine

How Many Spines do I need?

Oversubscription and Maximum Redundancy as the Criteria

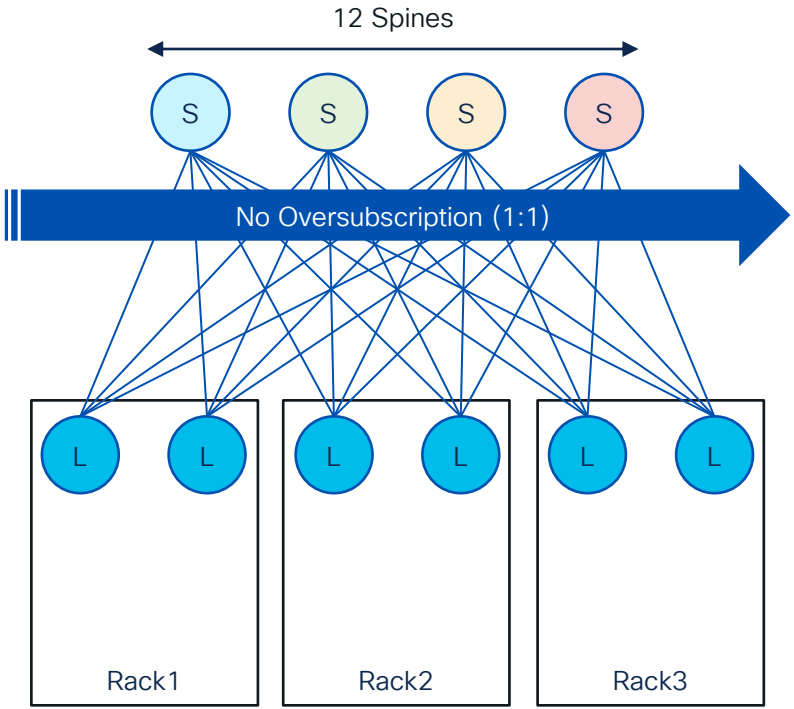
Host Attachment Requirements
- 48 Server per Rack
- 2x 25Gbps NIC per Server
- 1x NIC per Switch



Resulting Uplink Requirements
- 48x 25Gbps per Leaf
- 1.2Tbps Uplink from Leaf to Spine
- 12x 100Gbps towards Spine

Fabric Size – 12 Spine, 1:1 Oversubscription

Oversubscription and Maximum Redundancy as the Criteria



Let's Do some Math
Spine
8 Slot Modular Chassis
36x 100Gbps Port per Linecard
Total: 288 Spine Ports
Leaf
288 Spine Ports = 288 Leaf Switch
48x 25Gbps Host Ports Per Leaf
Total: 13'828 Host Ports
Fabric Bandwidth
1:1 Oversubscription
1.2Tbps Uplink * 288 Leaf
Total: 345.6Tbps

What if I Need to Connect another Server?

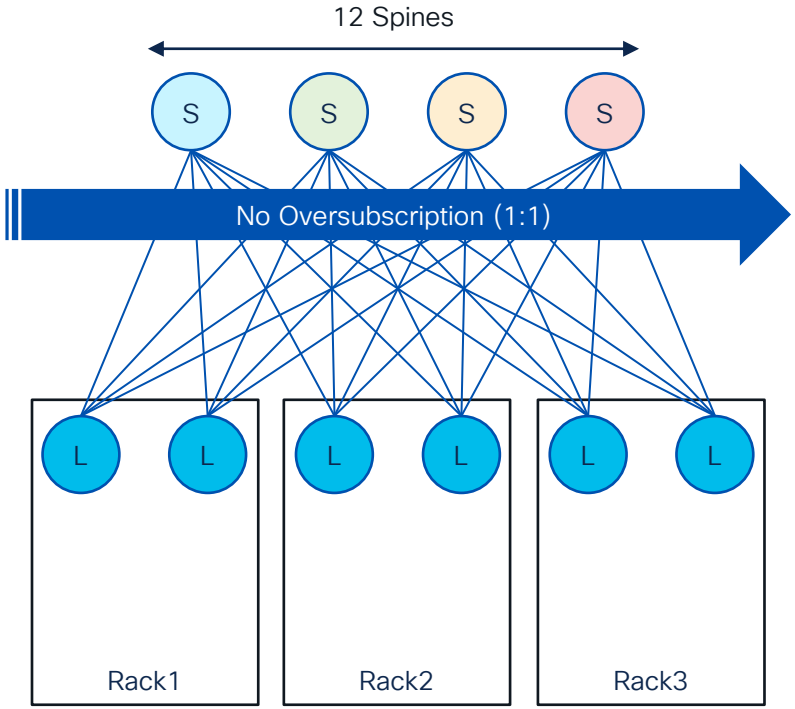
Is Scale Finite !?!?!?

‘Replacing the Chassis Size (8 Slot to 16 Slot Chassis)’

Is Scale Finite !?!?!?

Fabric Size – 12 Spine, 1:1 Oversubscription

Oversubscription and Maximum Redundancy as the Criteria



Let's Do some Math	
Spine	
8 Slot Modular Chassis	16 Slot Modular
36x 100Gbps Port per Linecard	36x 100Gbps Port per Linecard
Total: 288 Spine Ports	Total: 576 Spine Ports
Leaf	
288 Spine Ports = 288 Leaf Switch	576 Spine Ports = 576 Leaf Switch
48x 25Gbps Host Ports Per Leaf	48x 25Gbps Host Ports Per Leaf
Total: 13'828 Host Ports	Total: 27'648 Host Ports
Fabric Bandwidth	
1:1 Oversubscription	1:1 Oversubscription
1.2Tbps Uplink * 288 Leaf	1.2Tbps Uplink * 576 Leaf
Total: 345.6Tbps	Total: 691.2Tbps

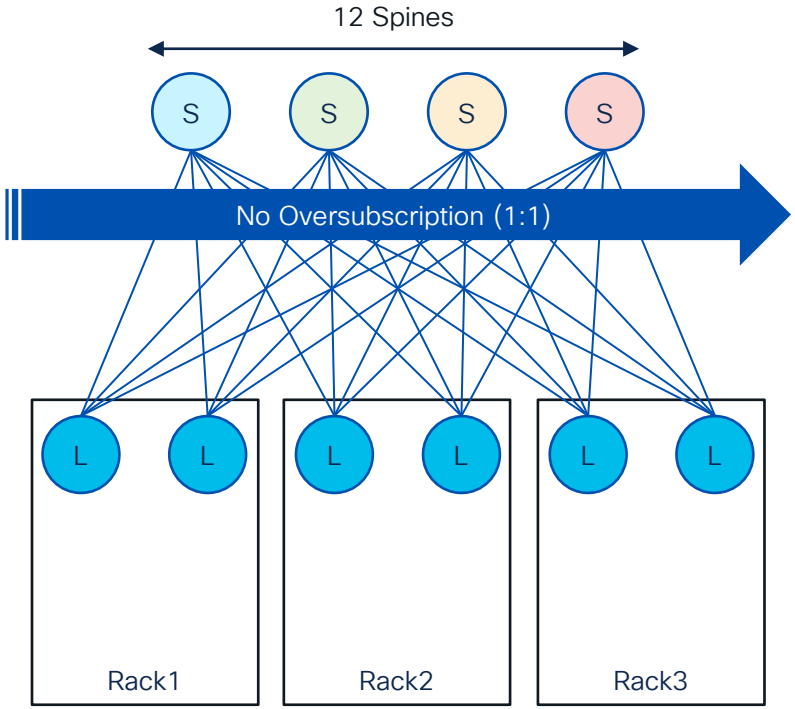
Doubling the Host Port Scale

‘Replacing the Spine Port Speed (100Gbps to 400Gbps)’

Is Scale Finite !?!?!?

Fabric Size – 12 Spine, 1:1 Oversubscription

Oversubscription and Maximum Redundancy as the Criteria



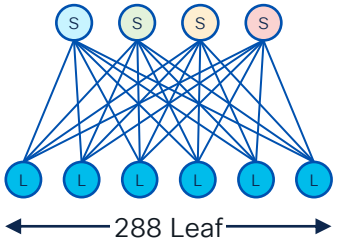
Let's Do some Math	
Spine	
8 Slot Modular Chassis	8 Slot Modular Chassis
36x 100Gbps Port per Linecard	36x 400Gbps Port per Linecard
Total: 288 Spine Ports	Total: 1152 Spine Ports
Leaf	
288 Spine Ports = 288 Leaf Switch	1152 Spine Ports = 1152 Leaf Switch
48x 25Gbps Host Ports Per Leaf	48x 25Gbps Host Ports Per Leaf
Total: 13'828 Host Ports	Total: 55'296 Host Ports
Fabric Bandwidth	
1:1 Oversubscription	1:1 Oversubscription
1.2Tbps Uplink * 288 Leaf	1.2Tbps Uplink * 1152 Leaf
Total: 345.6Tbps	Total: 1'382.4Tbps

Quadrupling the Host Port Scale (Breakout 4x 100Gbps at Spine)

‘Scale is very Linear in 2 Tier Networks’

More Spine Ports Results in More … Fabric Bandwidth, Leaf Count, Host Ports

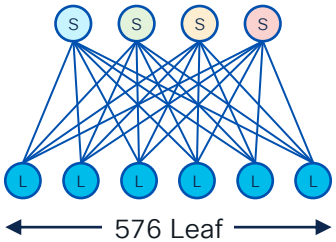
Attributes to Scale



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

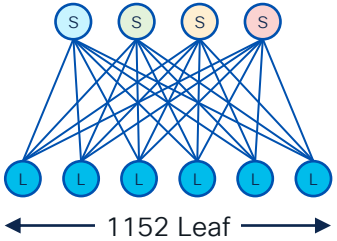
Scale-Up to Fill Chassis

Scale-Up to Bigger Chassis



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

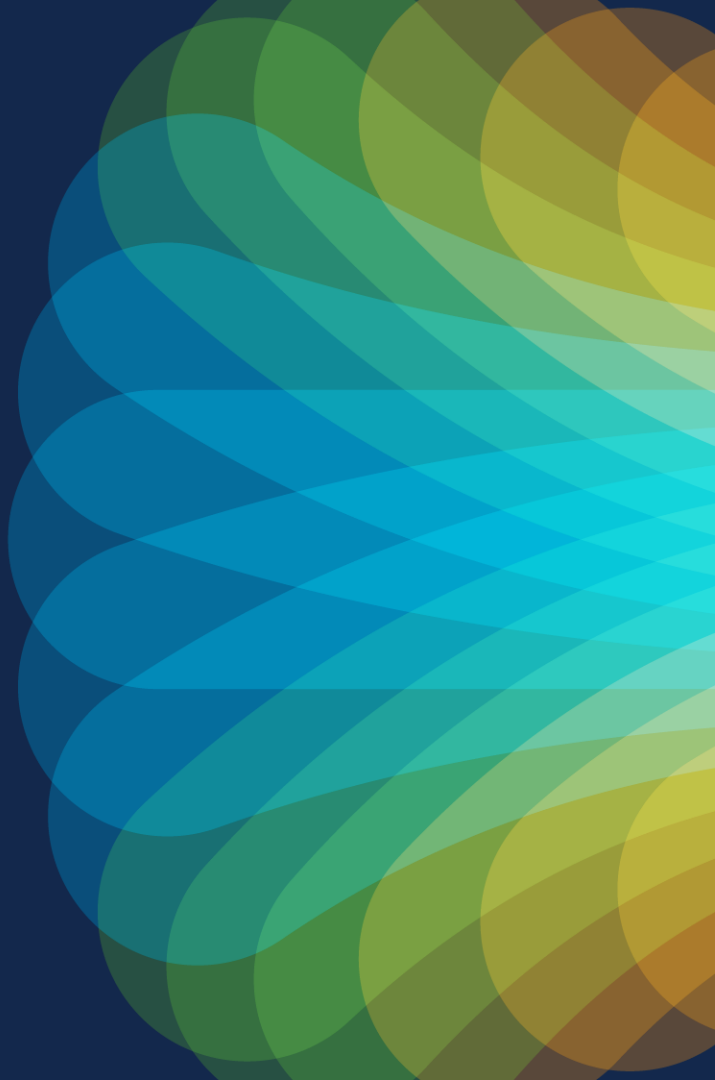
Scale-Up to Faster Linecards



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

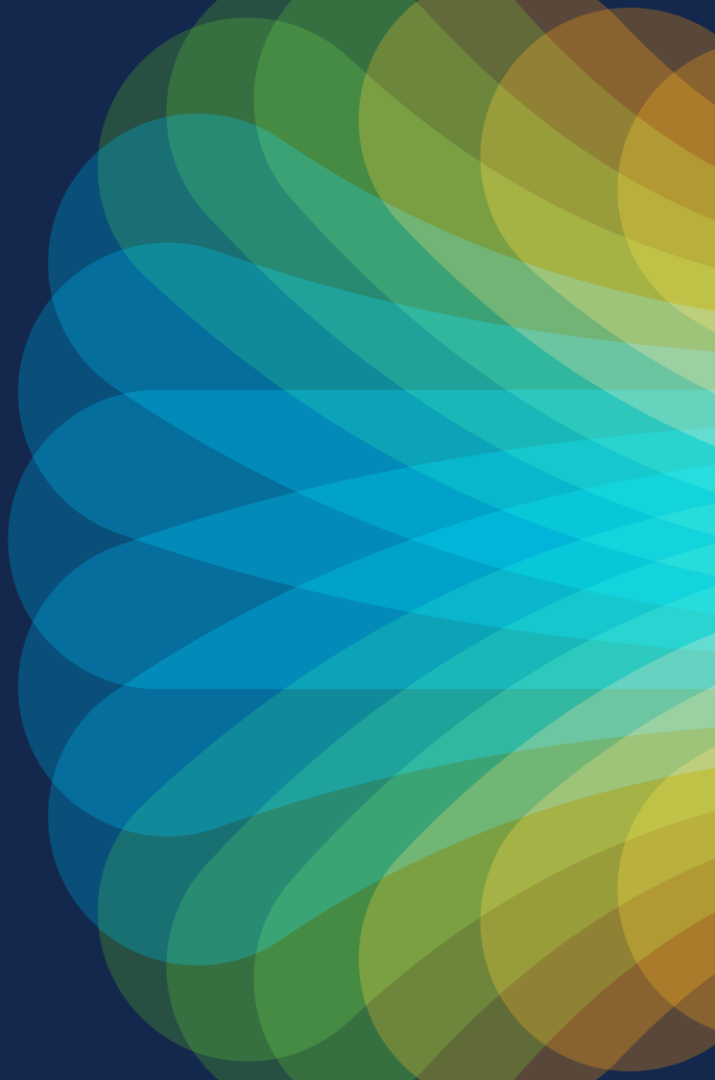
Design Evolution

1. Adopting Non-Modular Spines
2. Building a Distributed Architecture



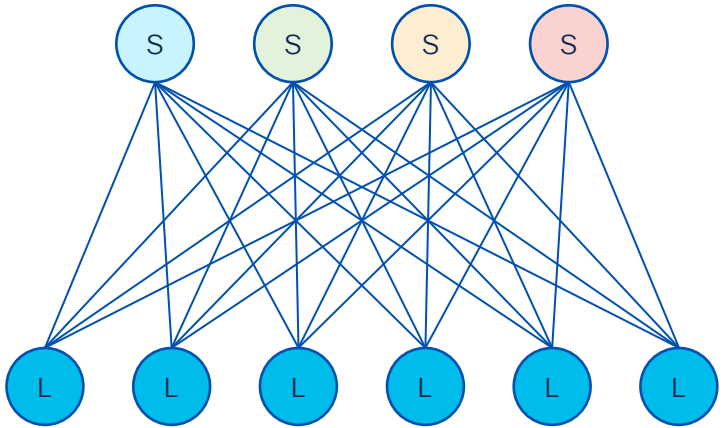
Design Evolution

Why Adopting Non-Modular Spines?

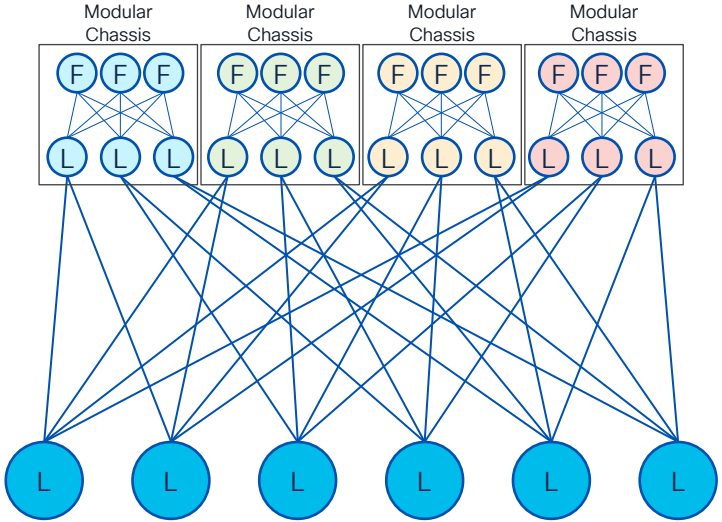


Why Adopting Non-Modular Spines?

2 Tier / 5 Stage Network with Modular Spine



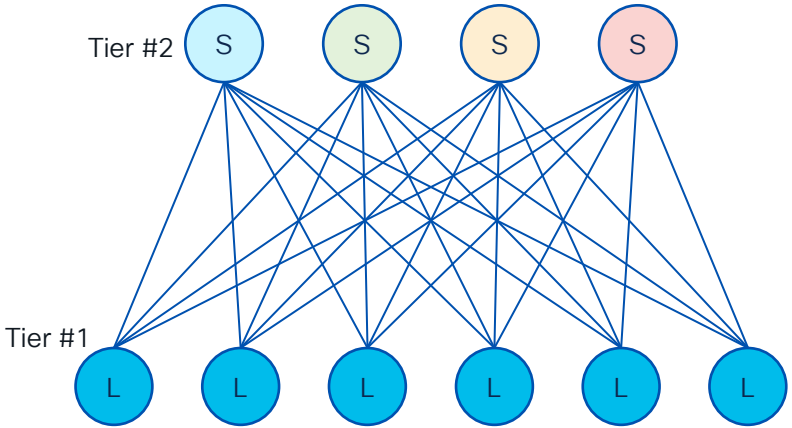
What you think you Built
2 Tier Leaf and Spine Network (3 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)



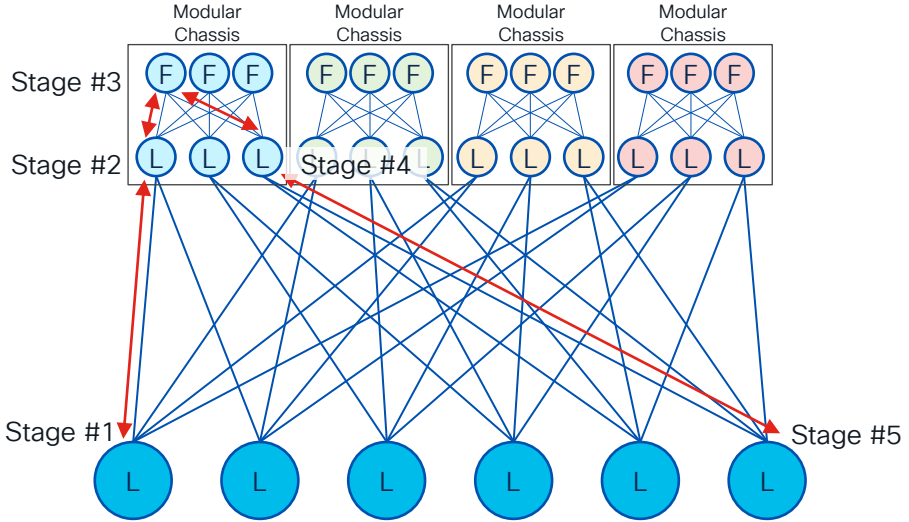
What you really Built
2 Tier Leaf and Spine Network (5 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)

Why Adopting Non-Modular Spines?

2 Tier / 5 Stage Network with Modular Spine



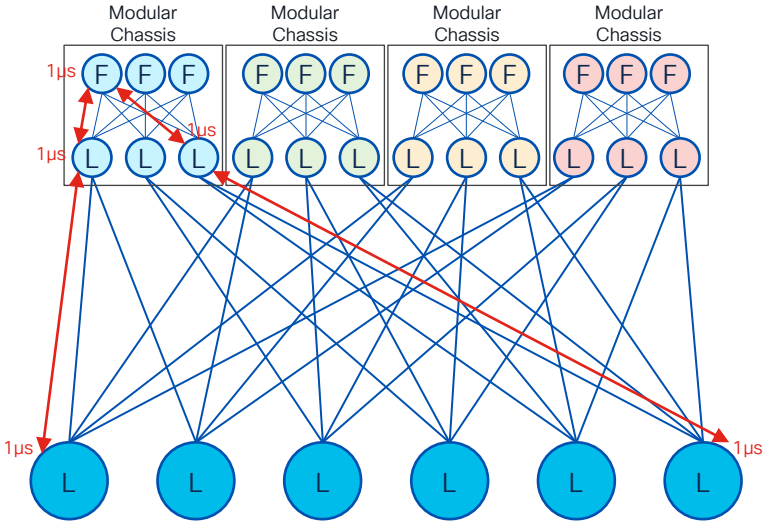
What you think you Built
2 Tier Leaf and Spine Network (3 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)



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)

Why Adopting Non-Modular Spines?

Latency Considerations with Modular Spines

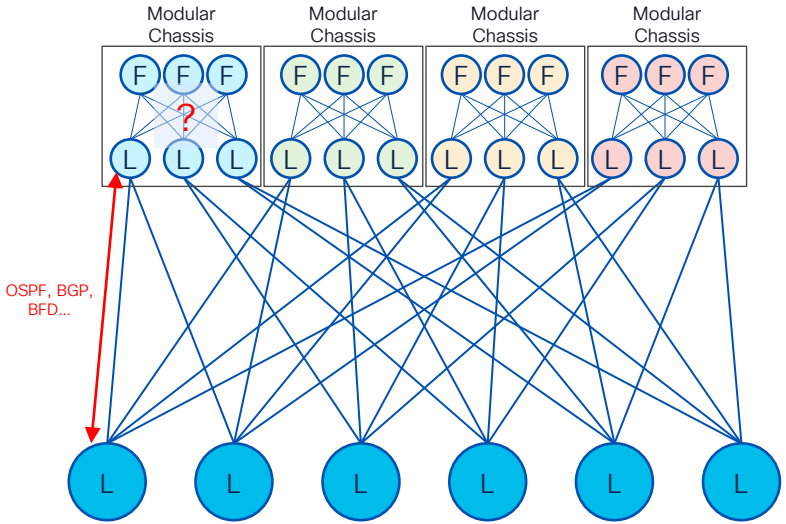


What you really Built
2 Tier Leaf and Spine Network (5 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)

- Generally, all Modular Switches operate in Store-and-Forward (SnF)
  - Packet Size dependent Latency
- Without Speed Change, Leaf operates in Cut-Through
  - Packet Size independent Latency
- Normalized, difference in Latency from Spine (Modular) to Leaf (Fixed) is 3:1

Why Adopting Non-Modular Spines?

Operational Considerations with Modular Spines

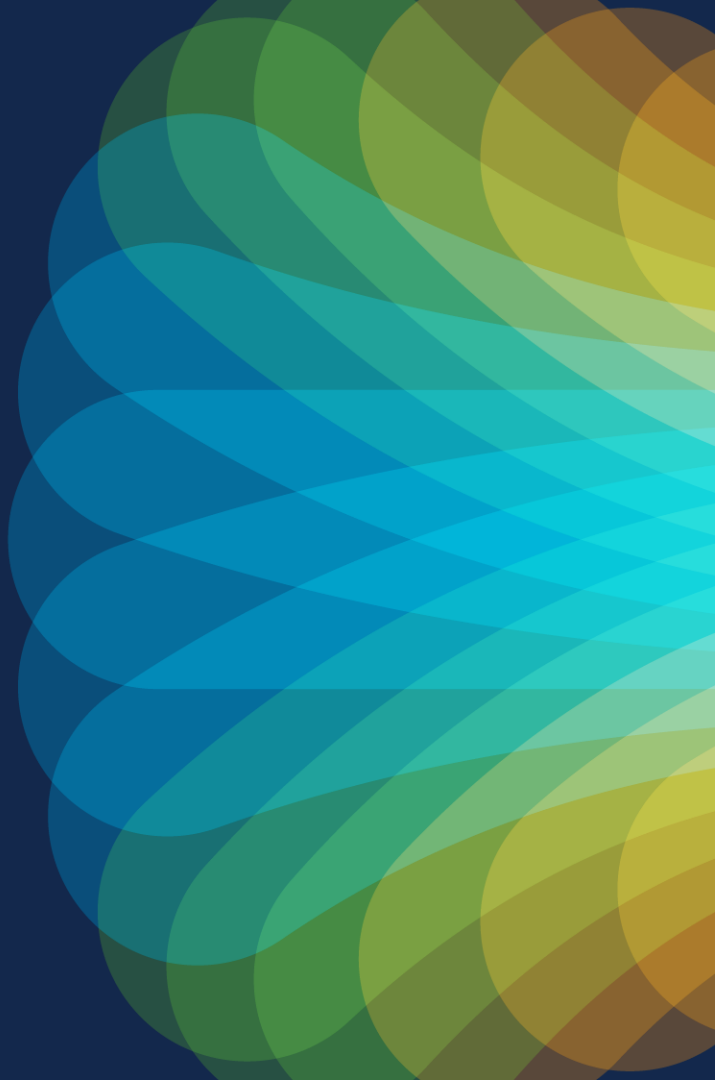


What you really Built
2 Tier Leaf and Spine Network (5 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)

- Within Leaf Tier and Between Leaf and Spine Tier
  - Full Behavior / Protocol Control
  - Layer-3 ECMP Load Balancing
  - Standards-based Routing Protocols
  - BFD for Fast Failure Detection
  - Minimal Exposure for Brownout
- Within Spine Tier
  - Intra-Chassis Load Balancing
  - Intra-Chassis Protocol
  - Intra-Chassis Failure Detection
  - Fully Redundant Components

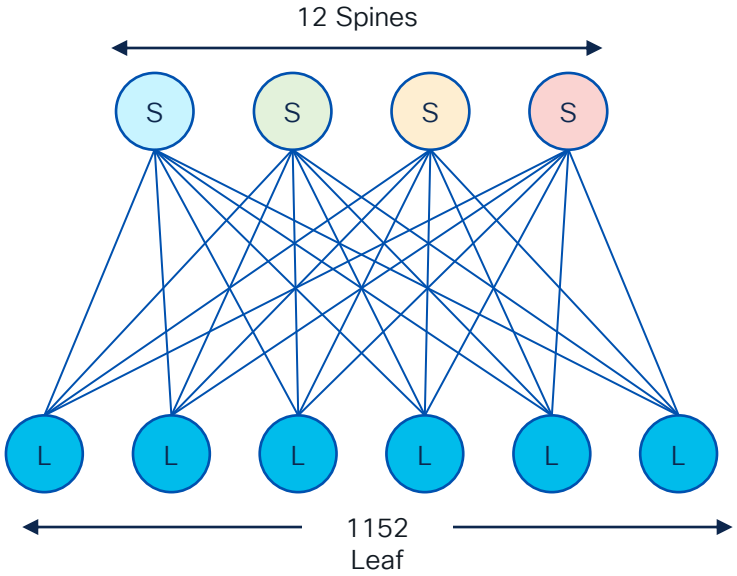
Design Evolution

Why Building a Distributed Architecture?



Design Evolution

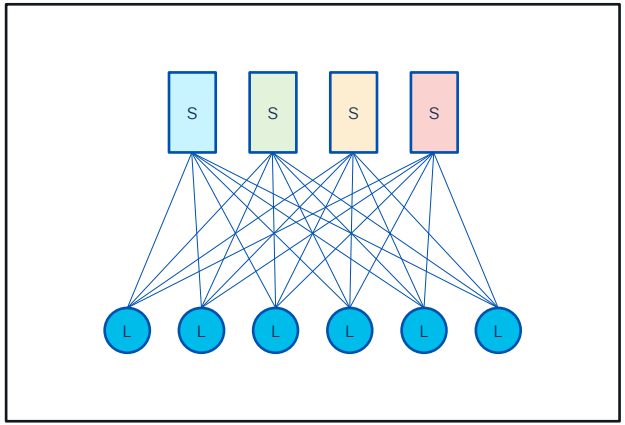
Various Considerations to Reduce the Fabric's Size



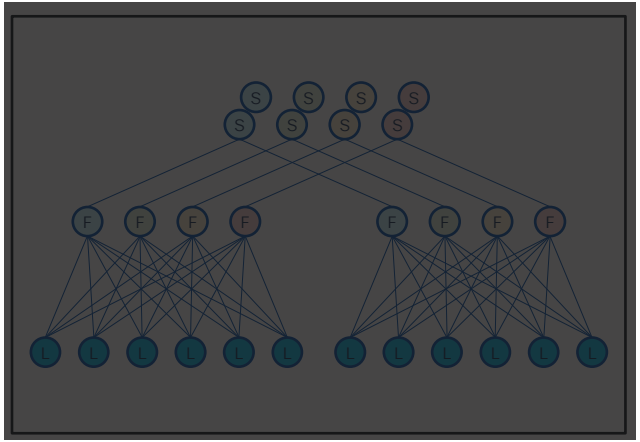
- What is my Failure Domain?
- What is my Change Domain?
- What is my Overall Scale?
- What is my Fabric Solution Scale?
- What is my Fabric SLA?
- What is my Maximum Downtime?

Design Evolution

From a Single Large Fabric to a Distributed Architecture



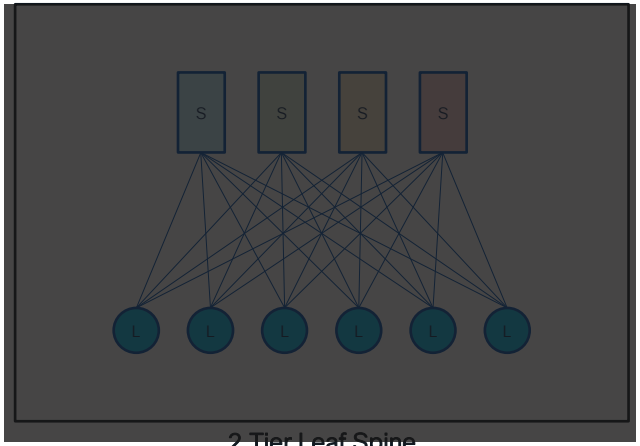
2 Tier Leaf Spine
(5 Stages)



3 Tier Leaf-Fabric-Spine
(5 Stages)

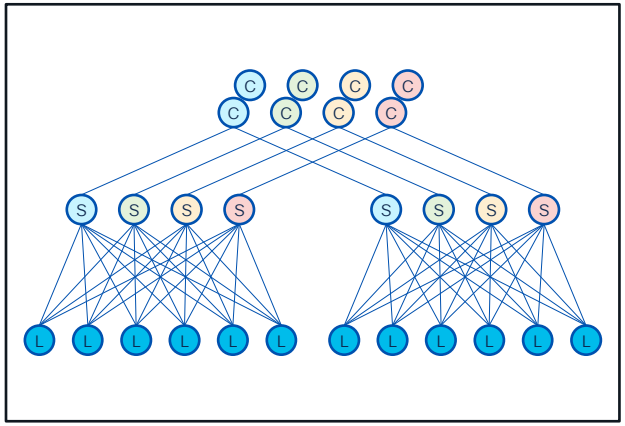
Design Evolution

From a Single Large Fabric to a Distributed Architecture



2 Tier Leaf Spine
(5 Stages)

Let's Move Forward



3 Tier Leaf-Fabric-Spine
(5 Stages)

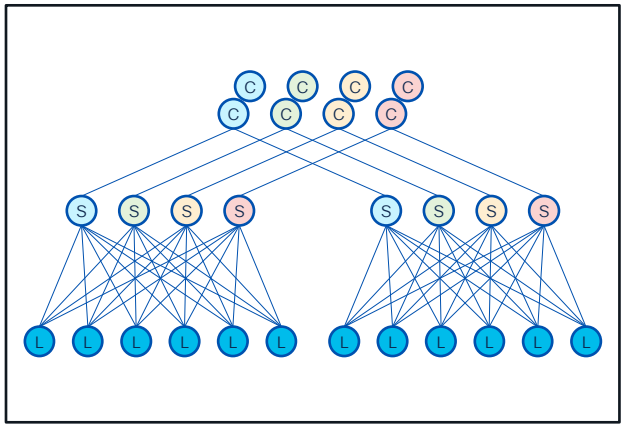
What we learned from the Cloud Titans

Building Scalable DataCenter Networks

#1	#2	#3	#4
Simplicity is Key Simple Design Principals	Scale as you Go Scale is Never Finite	Fail but Fail Fast Reduce Brown-Out Exposure	Redundant and Repeatable Risk is Never an Option

How the Cloud Titans Build

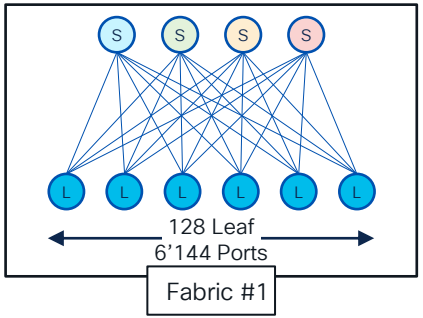
- Increasing Scale-Out in all Tiers
- Reducing Complexity
  - Simple Design Principles
- Increasing the “Finite Scale”
  - Scale as You Go
- Disaggregated Redundancy
- Flexible Link and Bandwidth Distribution
- Further Possibility for Cost Optimization



3 Tier Leaf-Fabric-Spine
(5 Stages)

Step #1 – Don’t Build Fabric for Maximum Leaf

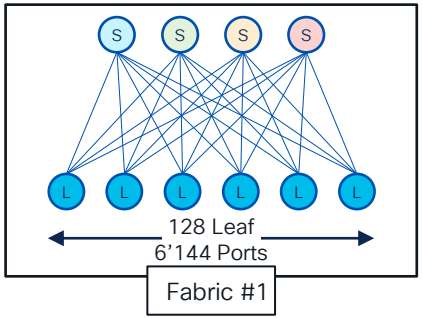
- Fixed Switch at the Fabric (Tier #2)
  - Depending on Oversubscription Ratio, reserve Ports
    - 1:1 Oversubscription
    - Reserve 50% from Tier #2 to Tier #3
- Common Fixed Spine Options
  - 64x 100Gbps (6.4Tbps Single ASIC)
  - 32x 400Gbps (12.8Tbps Single ASIC)
  - 64x 400Gbps (25.6Tbps Single ASIC)



Step #1 – Don't Build Fabric for Maximum Leaf

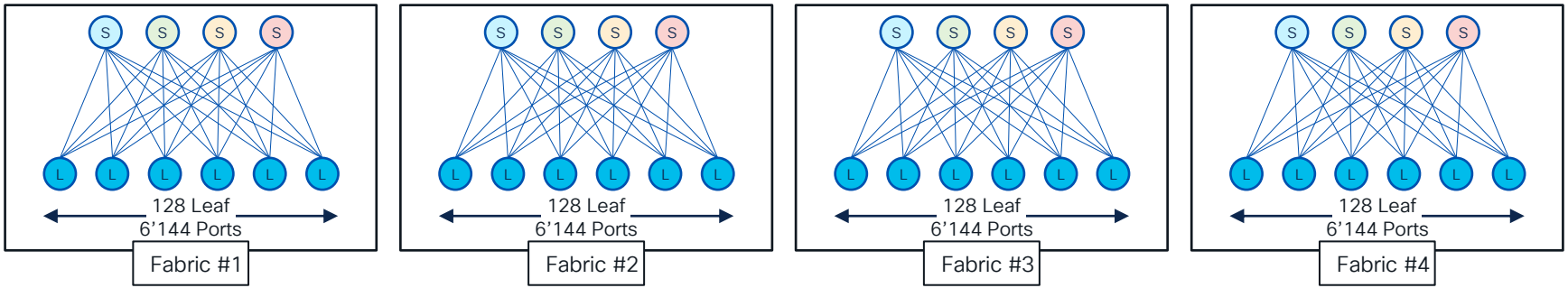
- Fixed Switch at the Fabric (Tier #2)
  - Depending on Oversubscription Ratio, reserve Ports
    - 1:1 Oversubscription
    - Reserve 50% from Tier #2 to Tier #3
- Common Fixed Spine Options
  - 64x 100Gbps (6.4Tbps Single ASIC)
  - 32x 400Gbps (12.8Tbps Single ASIC)
  - 64x 400Gbps (25.6Tbps Single ASIC)

Tier #2: Nexus 9364D-GX2B – 64x Ports 400Gbps
50% Uplink to 3rd Tier (32x 400Gbps)
50% Downlink for Leaf (128x 100Gbps)
Breakout: 32x 400Gbps = 128x 100Gbps



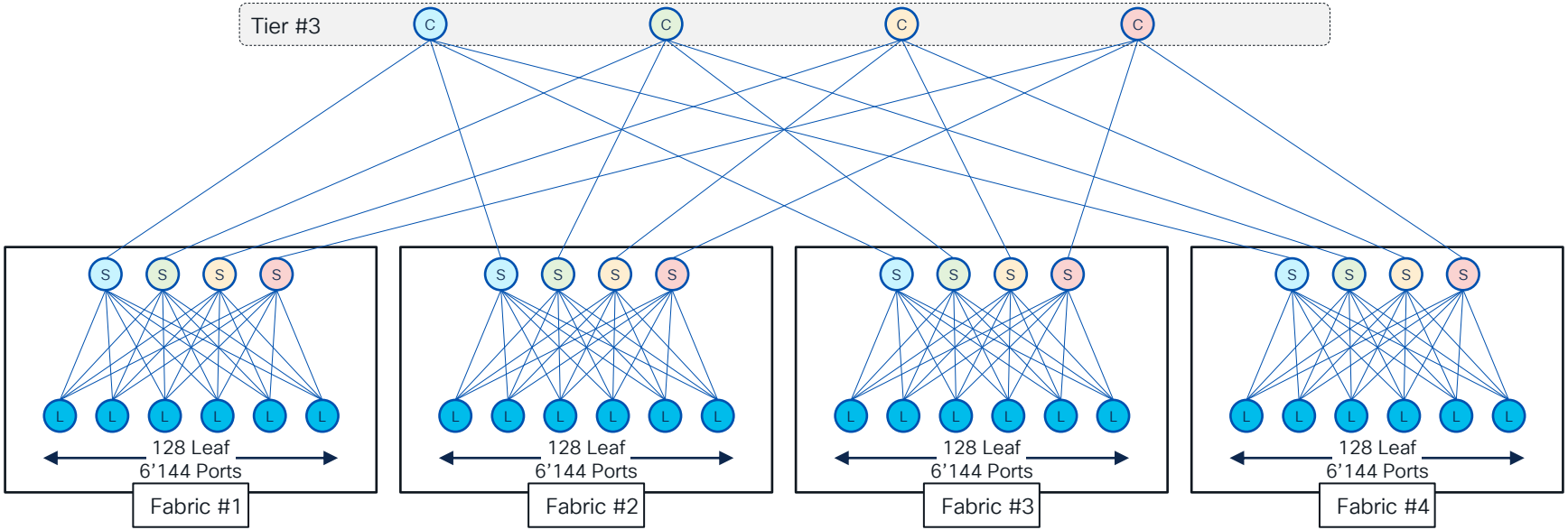
Step #2 – Repeat for Host Port Scale (Scale Out)

- Increasing Fabrics at need
  - of Host Port
  - of Oversubscription between Tier #2 and #3
- Result Defines Tier #2 to Tier #3 Uplinks
  - and respectively Tier #3 Requirements



Step #3 – Designing Tier #3

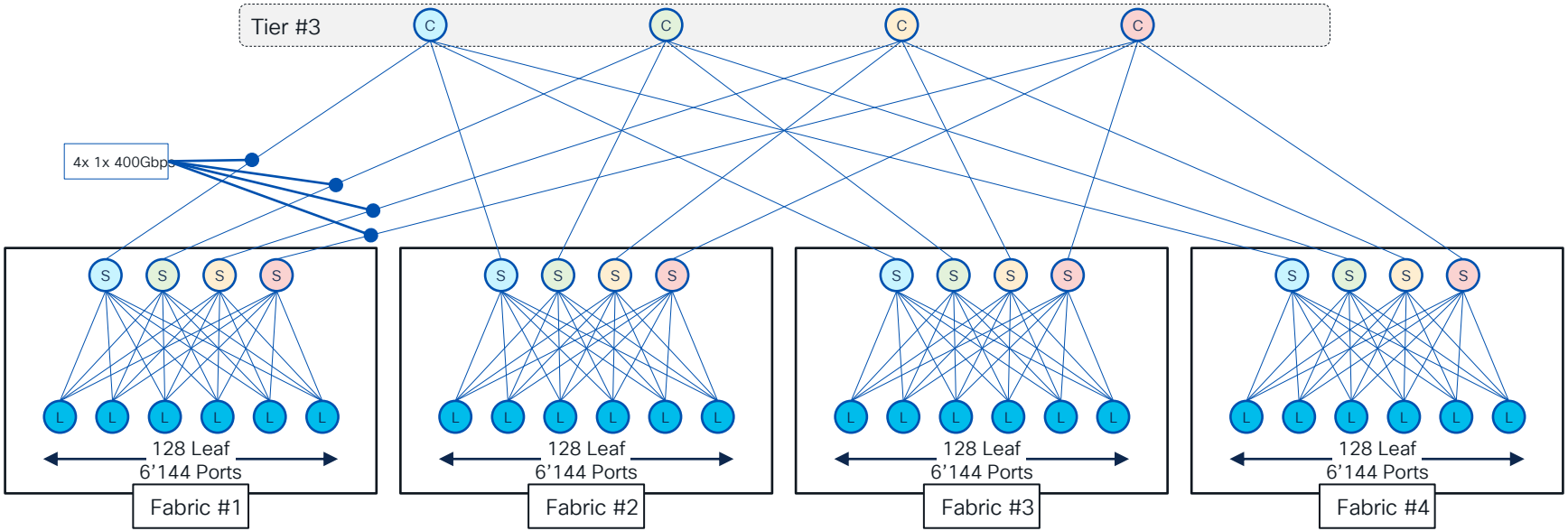
- Introducing Tier #3 Planes
  - Blue, Green, Yellow, Red
- Rule: Tier #2 Blue only connects to Tier#3 Blue
- Rule: Once entered a Plane, you stay in the Plane



Step #3 – Designing Tier #3 (Single Link)

- This Setup gets you all the way to 32 Fabrics
  - 32x 6'144 Host Ports = 196'608 Host Ports
  - Tier #3 is Layer-3 transport Network

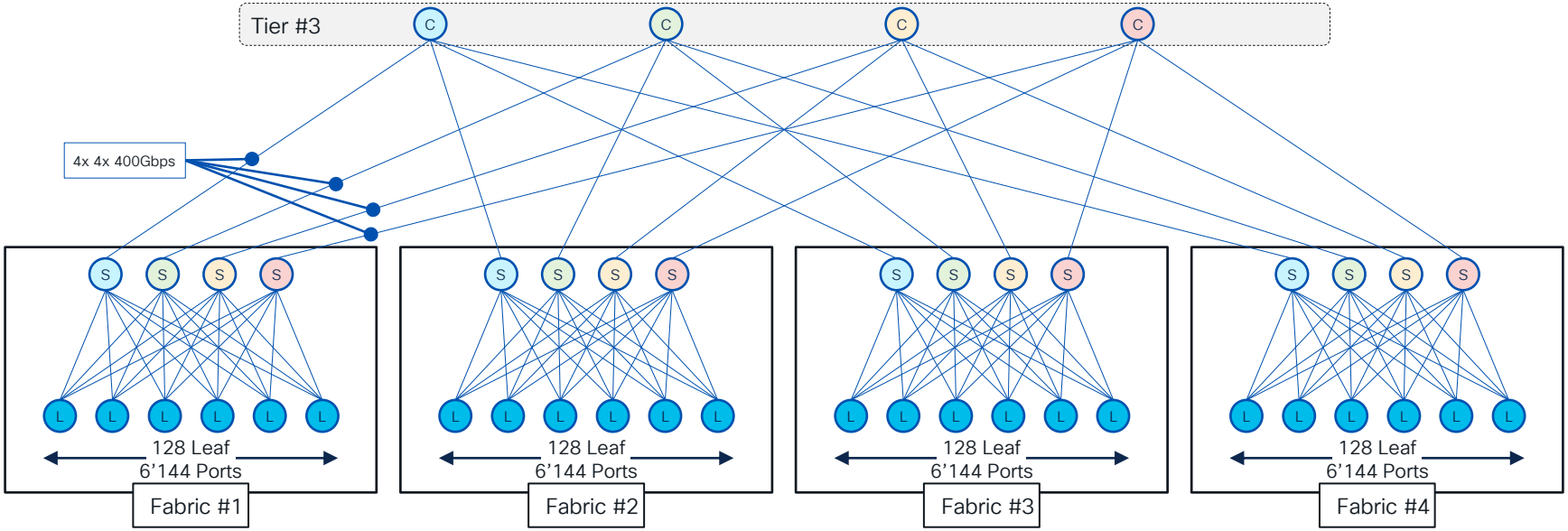
Tier #3: Nexus 9332D-GX2B (4 per Plane) - 32x Ports 400Gbps
4x 4x 400Gbps = 6.4Tbps inter-Fabric Bandwidth



Step #3 – Designing Tier #3 (Single Link)

- This Setup gets you all the way to 8 Fabrics
  - 32x 6’144 Host Ports = 49’152 Host Ports
  - Tier #3 is Layer-3 transport Network

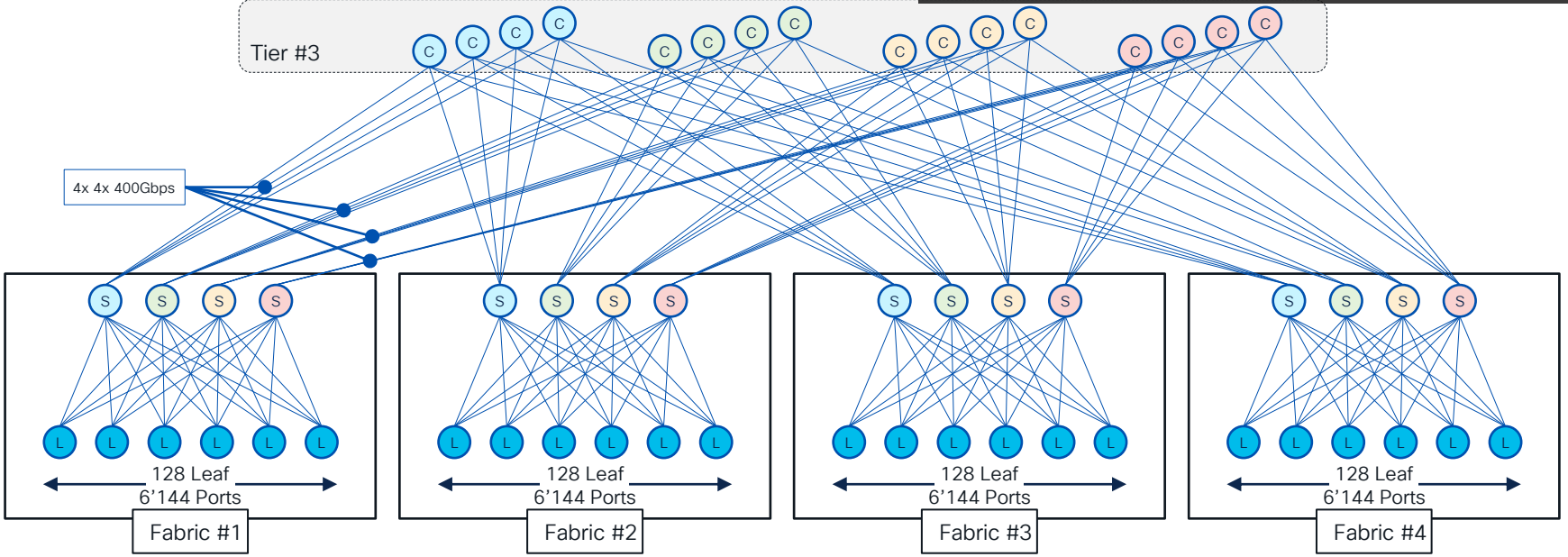
Tier #3: Nexus 9332D-GX2B (4 per Plane) - 32x Ports 400Gbps
4x 4x 400Gbps = 6.4Tbps inter-Fabric Bandwidth



Step #4 – Increasing the Tier #3 Planes

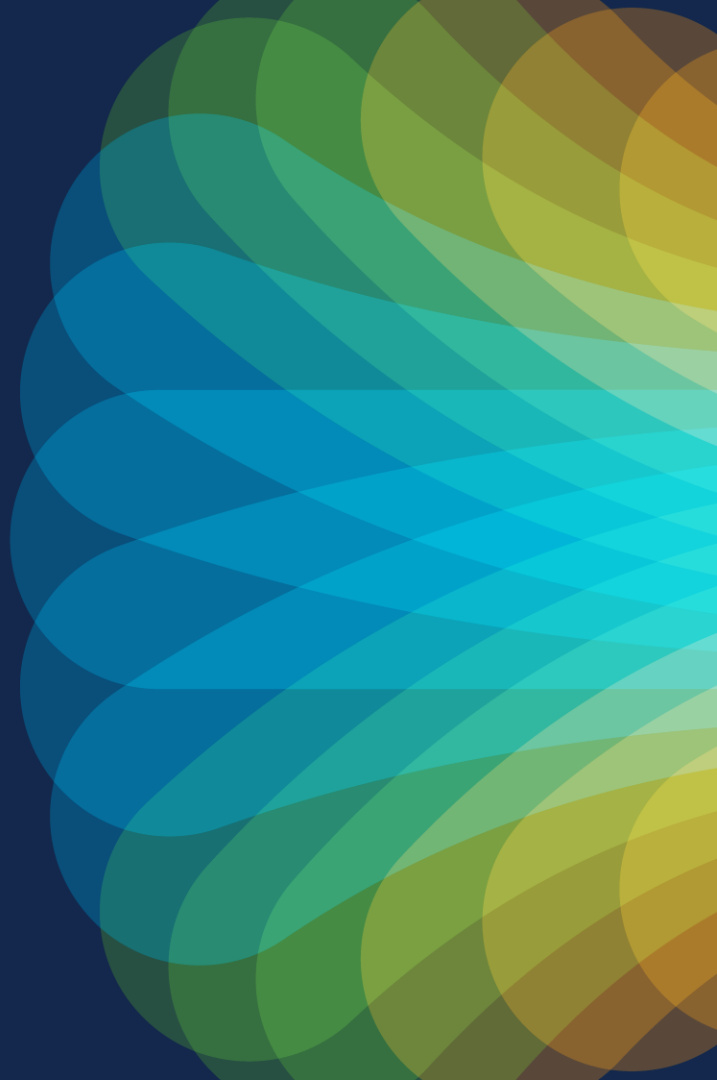
- This Setup gets you all the way to 32 Fabrics
  - 32x 6'144 Host Ports = 196'608 Host Ports
  - Tier #3 is Layer-3 transport Network

Tier #3: Nexus 9332D-GX2B (4 per Plane) - 32x Ports 400Gbps
4x 4x 400Gbps = 6.4Tbps inter-Fabric Bandwidth



Design Evolution and Its Mapping to Different Architectural Options

(ACI Fabrics, VXLAN EVPN Fabrics, Heterogeneous Fabrics, Routed Fabrics)

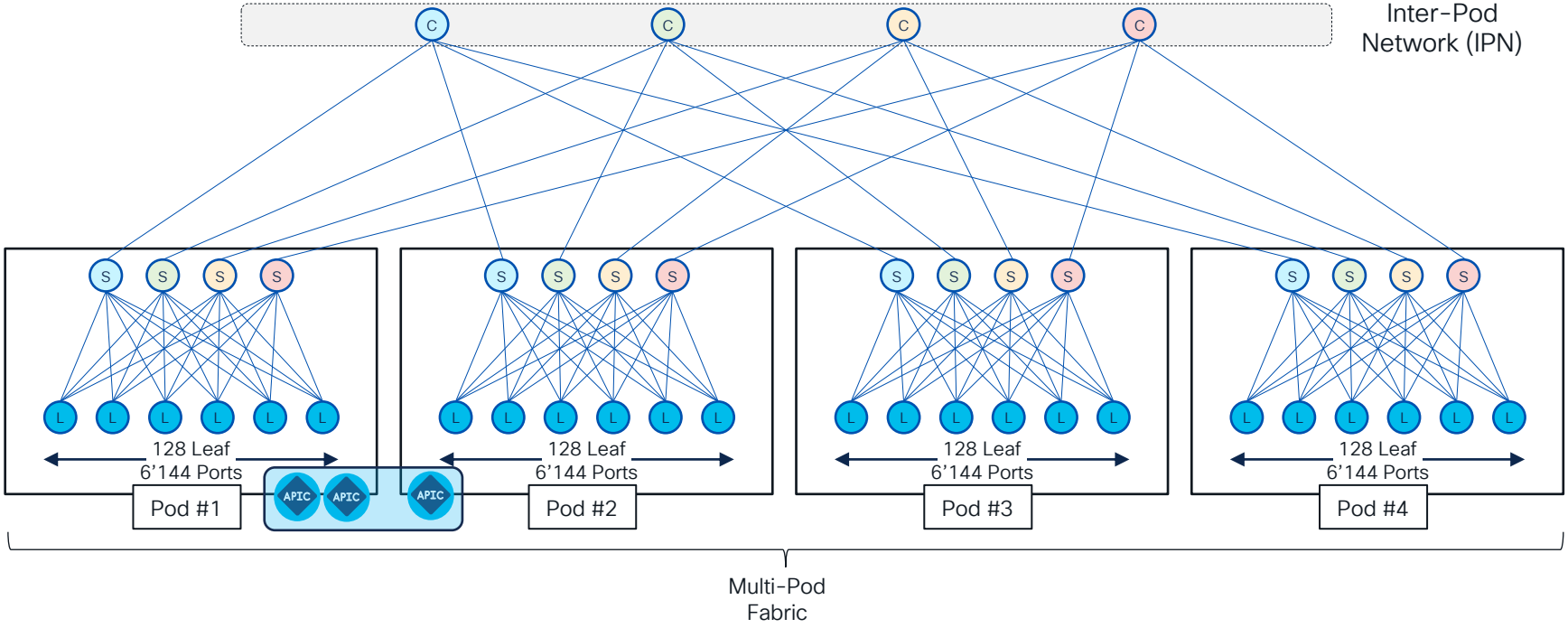


cisco Live!

ACI Fabrics

ACI Fabrics

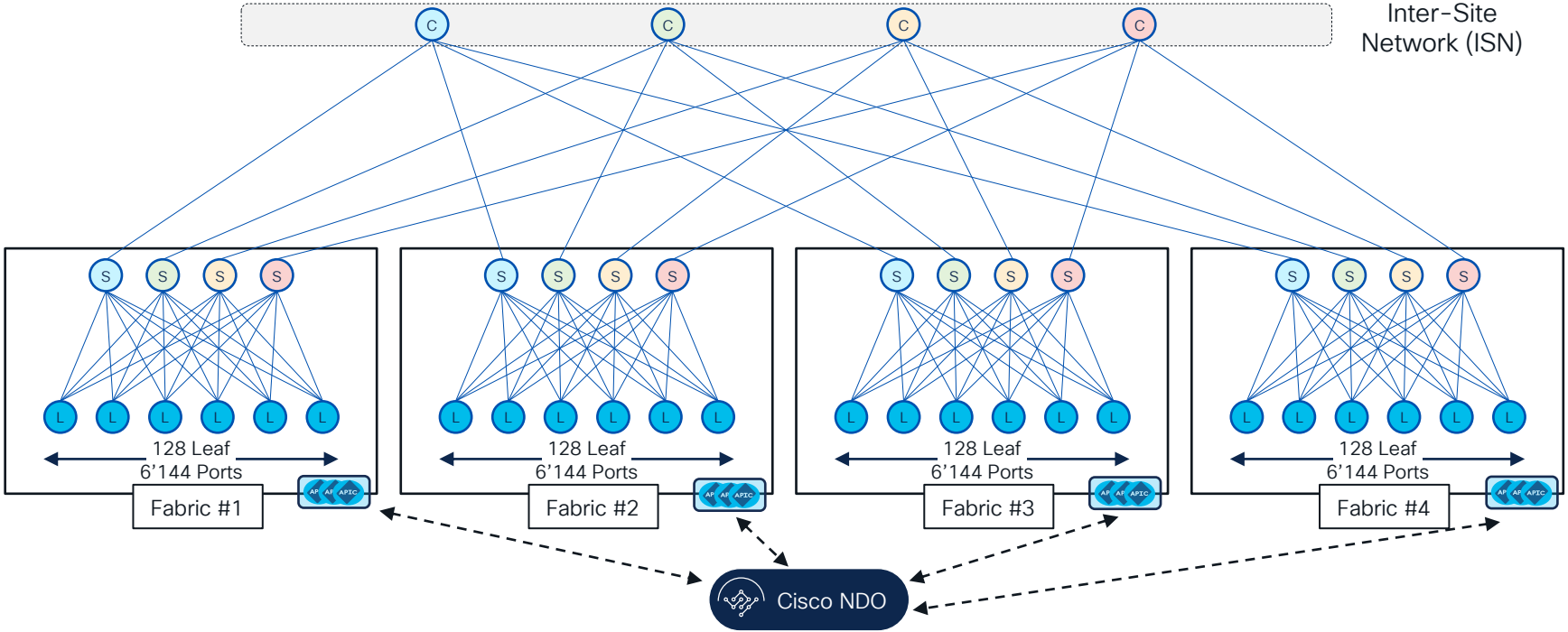
ACI Multi-Pod (Up to 25 Pods in the Same Fabric)



For More Information on ACI Multi-Pod please Refer to BRKDCN-2949

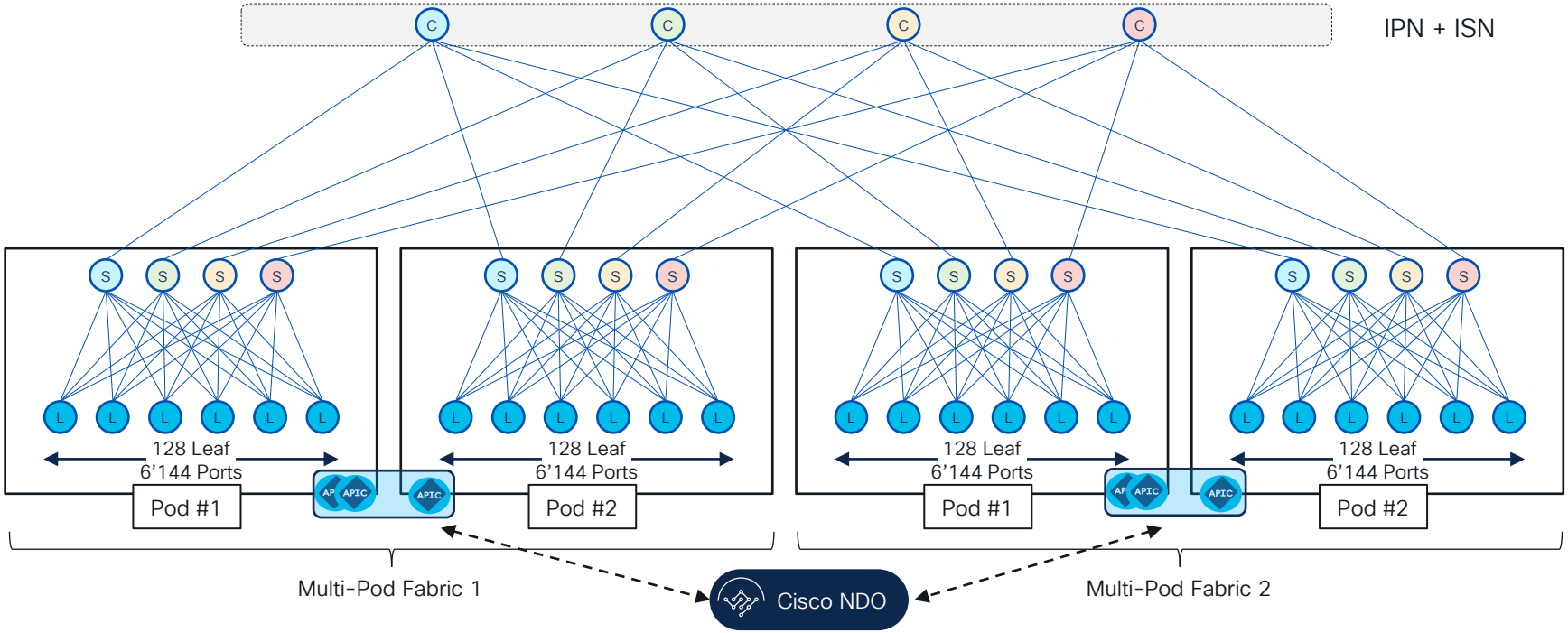
ACI Fabrics

ACI Multi-Site (Up to 14 Fabrics in the Same Multi-Site Domain)



ACI Fabrics

Multi-Pod + Multi-Site

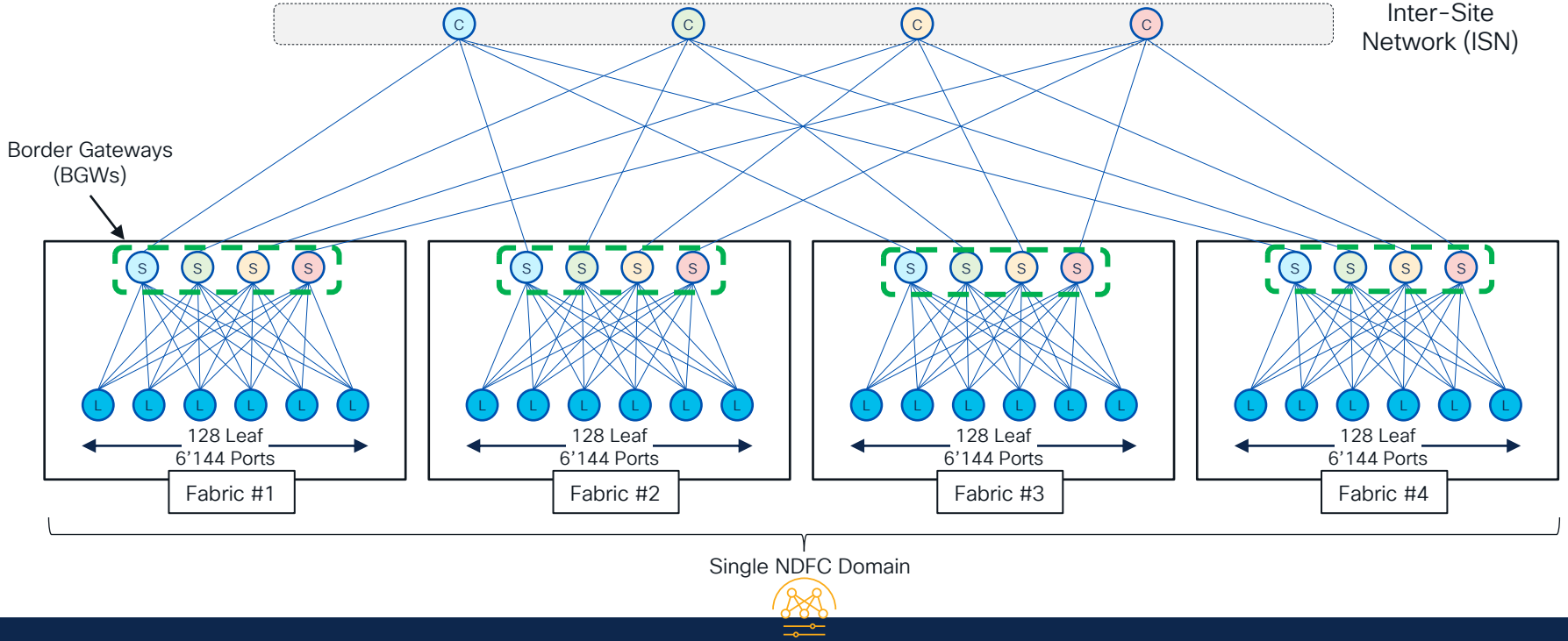


Want to know how to provision Multi-Pod and Multi-Site from scratch? Refer to BRKDCN-2919

VXLAN EVPN Fabrics

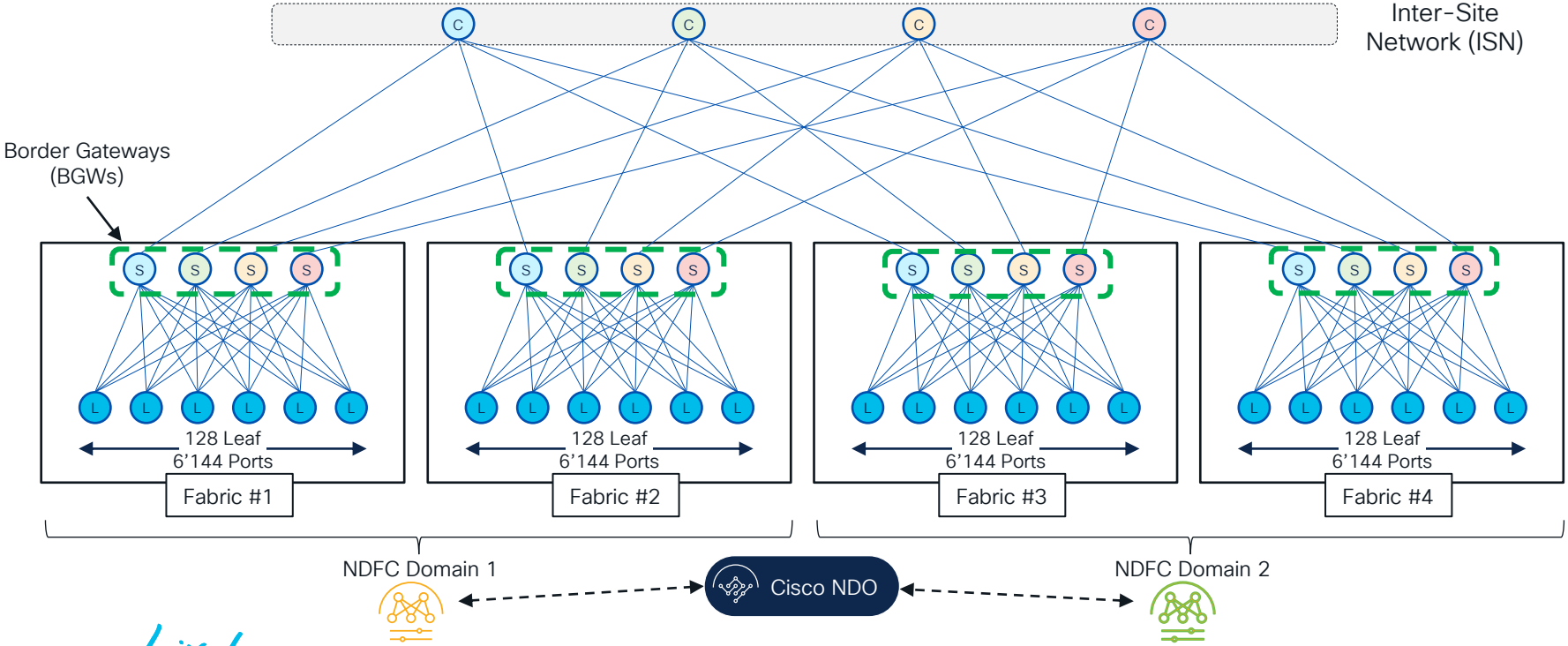
VXLAN EVPN Fabrics

Multi-Site, Single NDFC Domain



VXLAN EVPN Fabrics

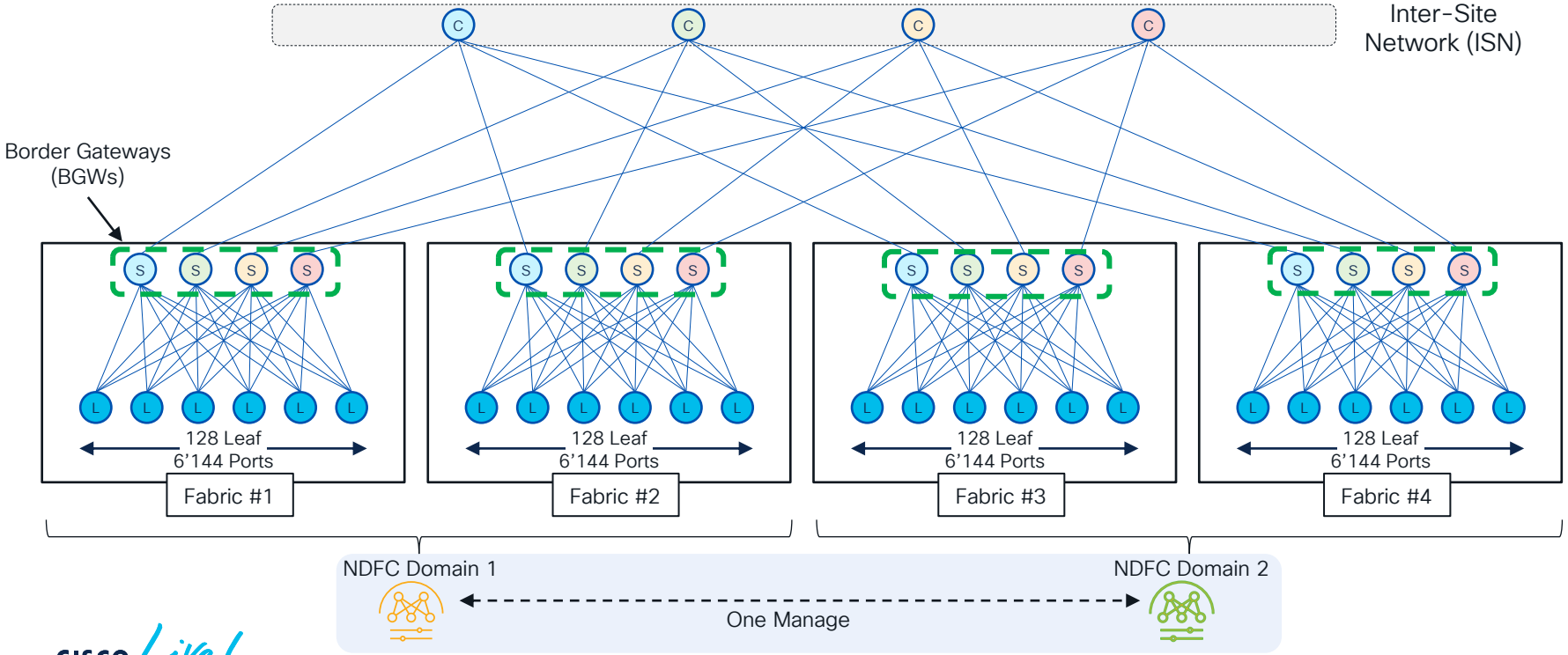
Multi-Site, Single NDFC Domain



VXLAN EVPN Fabrics

Multi-Site, Single NDFC Domain

Future



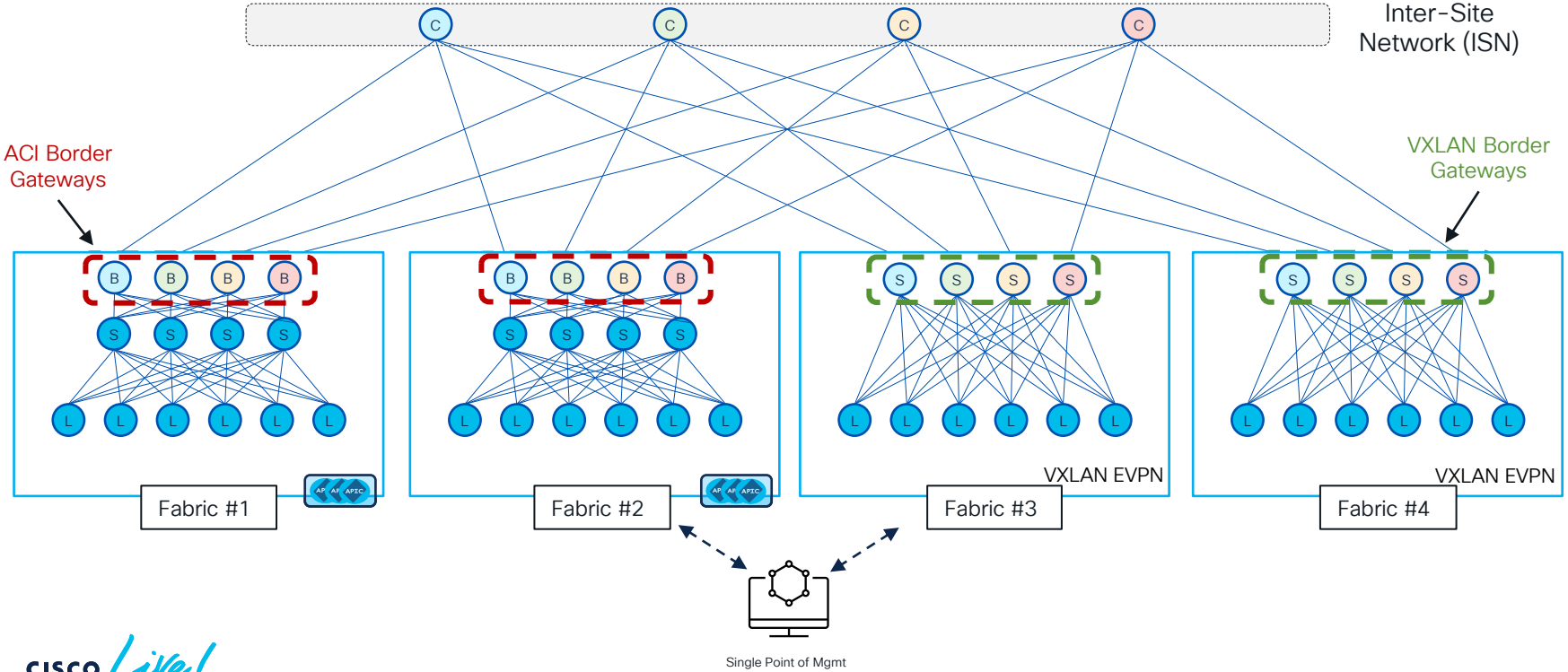
Heterogeneous Fabrics

Heterogeneous Fabrics

Introducing ACI Border Gateways

Future

"Opening Up" L2/L3 Connectivity between ACI and VXLAN EVPN Fabrics



Routed Fabrics with RFC 7938 and RFC 5549

Routed Fabrics

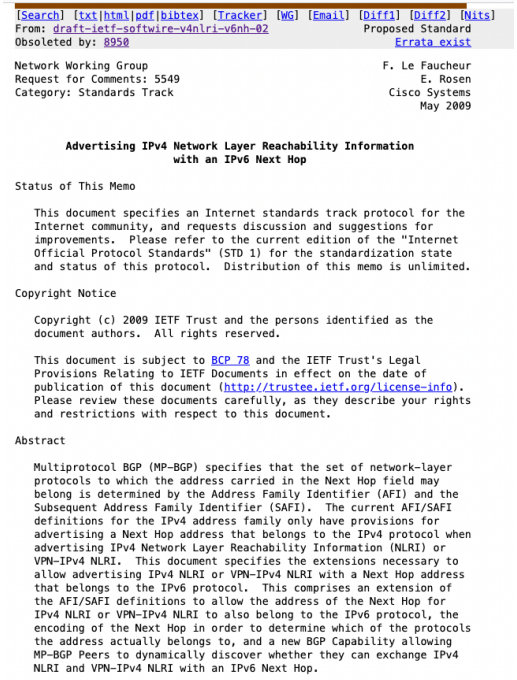
- A brief touchpoint of the work at the IETF (Internet Engineering Task Force) and what RFC (Request for Comment) are Standard and what Informational
- What is this RFC 5549 about – why do we have it and what is it good for
- Deployment Scenarios in Service Provider (SP) and Data Center (DC)
- How to make Layer-3 BGP Fabric deployments even simpler
- Addressing modern Cloud Native Applications needs

‘Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop’

https://datatracker.ietf.org/doc/html/rfc5549

What is RFC 5549?

By the Standards Body



[Search] [txt|html|pdf|bibtex] [Tracker] [WG] [Email] [Diff1] [Diff2] [Nits]
From: draft-ietf-softwire-v4nlri-v6nh-02 Proposed Standard
Obsoleted by: 8950 Errata exist

Network Working Group F. Le Faucheur
Request for Comments: 5549 E. Rosen
Category: Standards Track Cisco Systems
May 2009

Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop

Status of This Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents in effect on the date of publication of this document (http://trustee.ietf.org/license-info). Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

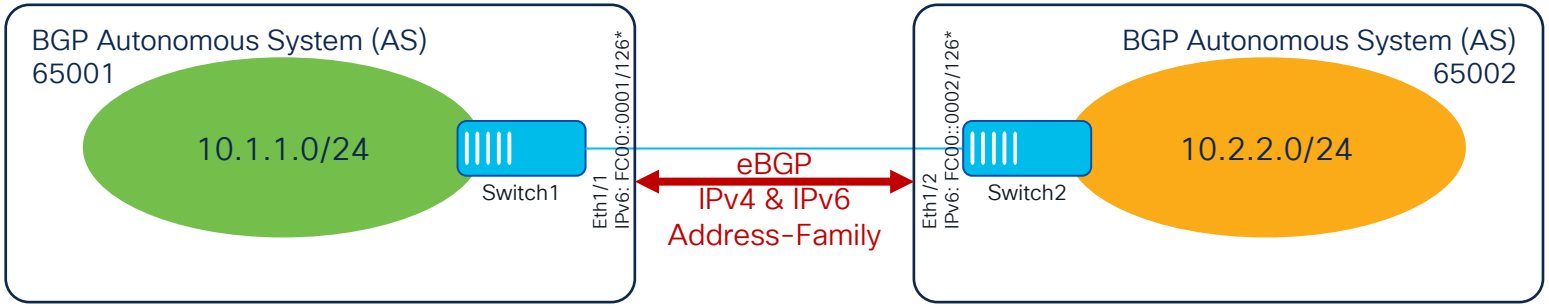
Multiprotocol BGP (MP-BGP) specifies that the set of network-layer protocols to which the address carried in the Next Hop field may belong is determined by the Address Family Identifier (AFI) and the Subsequent Address Family Identifier (SAFI). The current AFI/SAFI definitions for the IPv4 address family only have provisions for advertising a Next Hop address that belongs to the IPv4 protocol when advertising IPv4 Network Layer Reachability Information (NLRI) or VPN-IPv4 NLRI. This document specifies the extensions necessary to allow advertising IPv4 NLRI or VPN-IPv4 NLRI with a Next Hop address that belongs to the IPv6 protocol. This comprises an extension of the AFI/SAFI definitions to allow the address of the Next Hop for IPv4 NLRI or VPN-IPv4 NLRI to also belong to the IPv6 protocol, the encoding of the Next Hop in order to determine which of the protocols the address actually belongs to, and a new BGP Capability allowing MP-BGP Peers to dynamically discover whether they can exchange IPv4 NLRI and VPN-IPv4 NLRI with an IPv6 Next Hop.

- Internet Engineering Task Force (IETF) Request for Comment (RFC)
- Categorized for Standards Track
- Internet Standard since 2009
- Updated by RFC 8950
  - aka RFC 5549bis
- Industry wide adoption for more than 10 years
- Invented and Authored by Cisco

- RFC 5549
  - https://datatracker.ietf.org/doc/html/rfc5549
- RFC 8950
  - https://datatracker.ietf.org/doc/html/rfc8950

What is RFC 5549 for?

- Defines a specific behavior in Border Gateway Protocol (BGP)
- Allows IPv4 Network Layer Reachability via a IPv6 Next-Hop

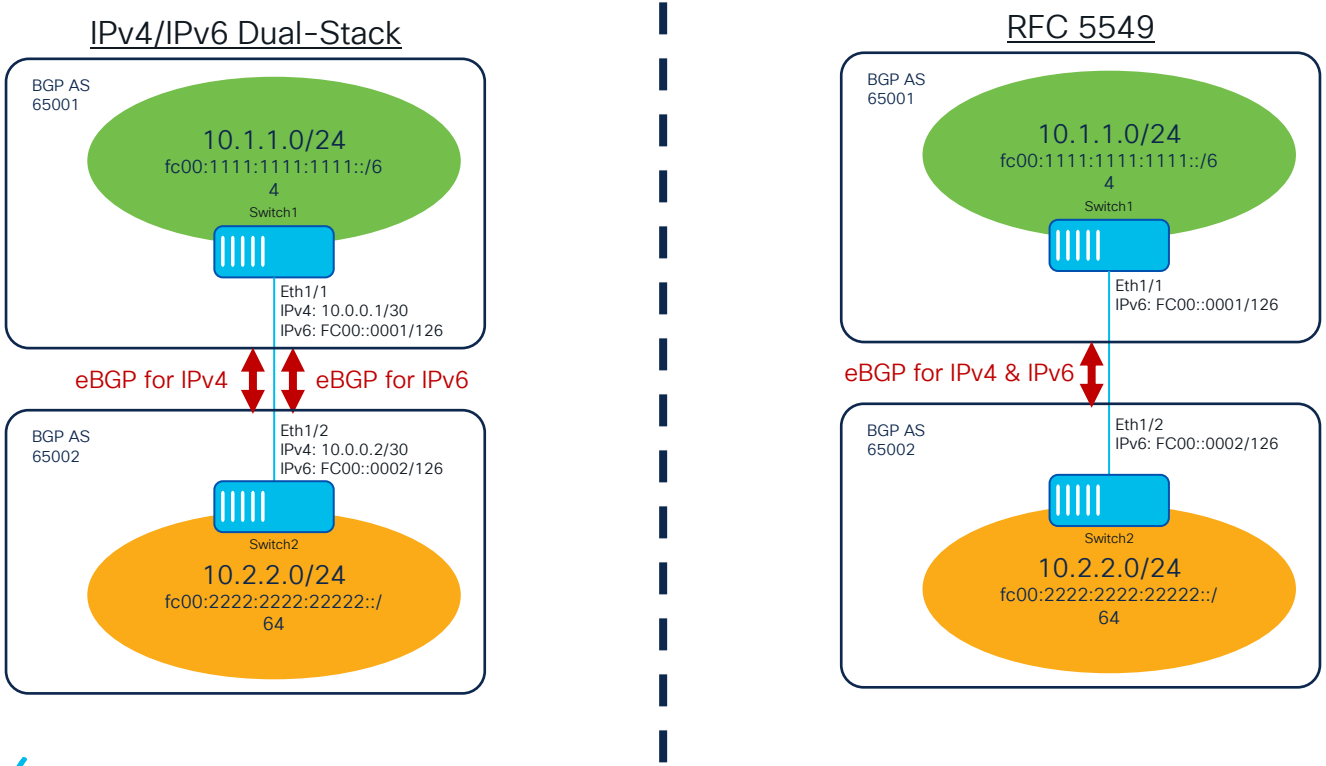



```
[Output]
Switch1# show ip bgp
BGP routing table information for VRF default, address family IPv4 Unicast
BGP table version is 7, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2

   Network          Next Hop            Metric     LocPrf     Weight Path
*>e10.2.2.0/24      fc00::0002               0                     0 65002 ?
```

*I don't think you will see /126 in real world, more likely this is going to be a /64 or better /127

Side-by-Side

IPv4/IPv6 Dual-Stack and RFC 5549



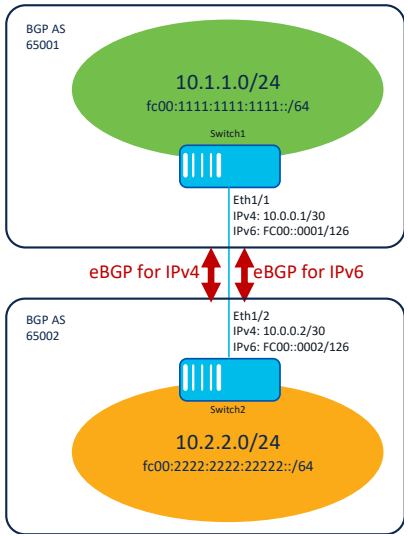
Side-by-Side – Config with IPv6 Numbered

IPv4/IPv6 Dual-Stack and RFC 5549

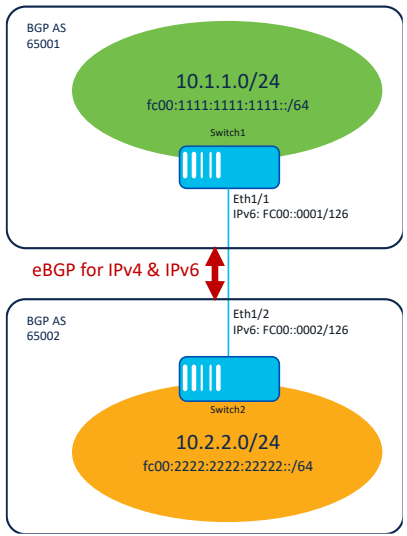
Per-Address-Family Peering

```
[Config]
router bgp 65001
  neighbor 10.0.0.2
    remote-as 65001
    address-family ipv4 unicast
  neighbor FC00:0001
    remote-as 65001
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 FC00:0001/126
 ip address 10.0.0.1/30
```

IPv4/IPv6 Dual-Stack



RFC 5549



Per-Neighbor Peering

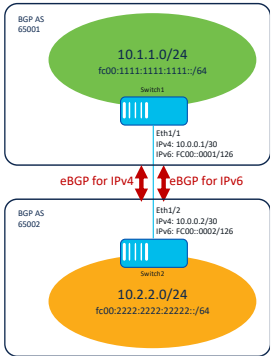
```
[Config]
router bgp 65001
  neighbor FC00:0001
    remote-as 65001
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 FC00:0001/126
 ip forward
```

*IPv6 GUA – IPv6 Global Unicast Address

Side-by-Side – Oper with IPv6 Numbered

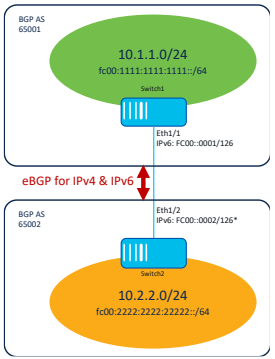
IPv4/IPv6 Dual-Stack and RFC 5549

IPv4/IPv6 Dual-Stack




```
[Output]
Switch1# show ip bgp

Network            Next Hop            Metric     LocPrf     Weight Path
*>e10.2.2.0/24     10.0.0.2                 0                     0 65002 ?


Switch1# show ipv6 bgp

Network            Next Hop            Metric     LocPrf     Weight Path
*>efc00:2222:2222:2222::/64
                   fc00::0002               0                     0 65002 ?
```

RFC 5549



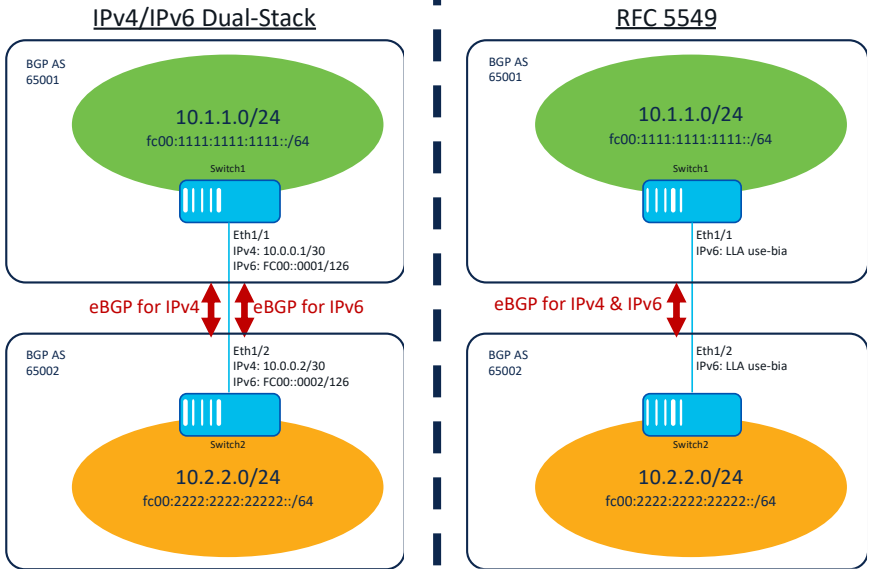

```
[Output]
Switch1# show ip bgp

Network            Next Hop            Metric     LocPrf     Weight Path
*>e10.2.2.0/24     fc00::0002               0                     0 65002 ?


Switch1# show ipv6 bgp

Network            Next Hop            Metric     LocPrf     Weight Path
*>efc00:2222:2222:2222::/64
                   fc00::0002               0                     0 65002 ?
```

Side-by-Side – Config with Unnumbered (LLA)

IPv4/IPv6 Dual-Stack and RFC 5549

Per-Address-Family Peering

```
[Config]
router bgp 65001
  neighbor 10.0.0.2
    remote-as 65001
    address-family ipv4 unicast
  neighbor FC00:0001
    remote-as 65001
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 FC00:0001/126
 ip address 10.0.0.1/30
```



Per-Neighbor **Unnumbered** Peering

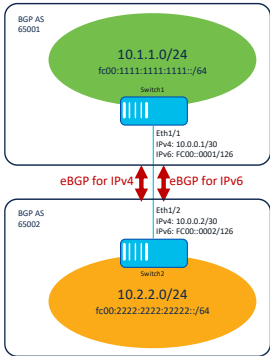
```
[Config]
router bgp 65001
  neighbor Ethernet1/1
    remote-as 65001
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

Removing the need for Interface IP Addressing or BGP Peer Configuration with IPv6 Link-Local Addressing and BGP interface peering (unnumbered)

Side-by-Side – Oper with Unnumbered (LLA)

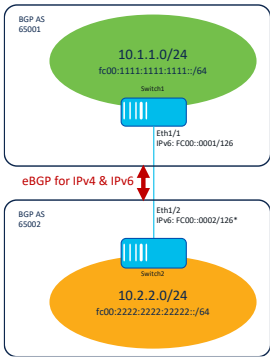
IPv4/IPv6 Dual-Stack and RFC 5549

IPv4/IPv6 Dual-Stack




```
[Output]
Switch1# show ip bgp

Network            Next Hop            Metric     LocPrf     Weight Path
*>e10.2.2.0/24     10.0.0.2                 0                     0 65002 ?


Switch1# show ipv6 bgp

Network            Next Hop            Metric     LocPrf     Weight Path
*>efc00:2222:2222:2222::/64
                   fc00::0002               0                     0 65002 ?
```

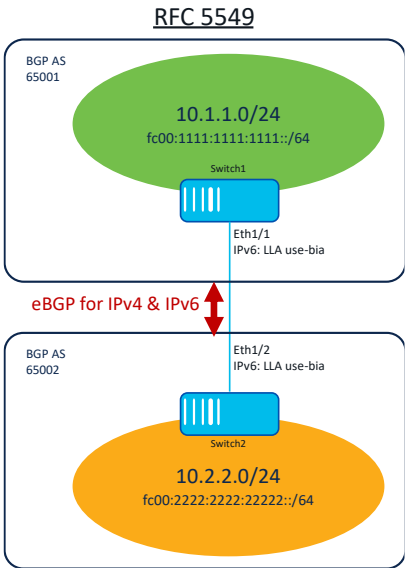
RFC 5549




```
[Output]
Switch1# show ip bgp

Network            Next Hop            Metric     LocPrf     Weight Path
*>e10.2.2.0/24     fe80::720f:6aff:fe4d:a7f0
                                             0                     0 65002 ?


Switch1# show ipv6 bgp

Network            Next Hop            Metric     LocPrf     Weight Path
*>efc00:2222:2222:2222::/64
                   fe80::720f:6aff:fe4d:a7f0
                                             0                     0 65002 ?
```

Deployment Simplification IPv6 Link-Local and BGP Interface Peering

For the rest of this Presentation, we are using IPv6 Link-Local and BGP Interface Peering



Per-Neighbor **Unnumbered** Peering

```
[Config]
router bgp 65001
  neighbor Ethernet1/1
    remote-as 65001
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

Removing the need for Interface IP Addressing or BGP Peer Configuration with IPv6 Link-Local Addressing and BGP interface peering

For the rest of this Presentation, we are using IPv6 Link-Local and BGP Interface Peering

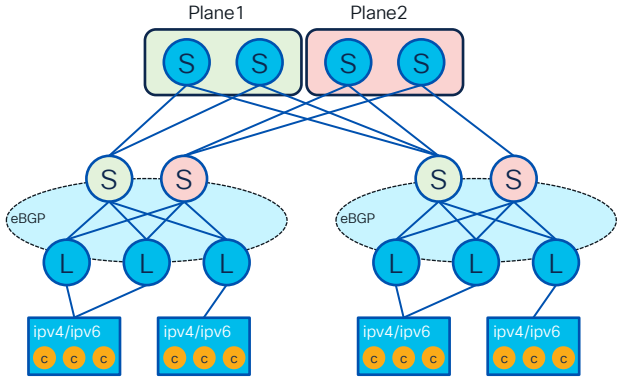
‘How does this fit in the Data Center?’

RFC 5549 Use Cases

RFC 5549 Use Cases?

From the Data Center Playbook

- For Example, Use of BGP for Routing in Large-Scale Data Centers
  - RFC 7938 https://datatracker.ietf.org/doc/html/rfc7938
- Used as Routing Protocol between Leaf, Spine and other Tiers (ie Super-Spine)
  - IPv4 and IPv6 Prefix with a single Routing Protocol Session – No VRF, VPNs or Overlays
  - Ready for “Cloud Native Applications”* – no need for Layer-2
  - Better BGP Session Scale on Leaf to Server (Fan Out) - Less Point-to-Point IP Addressing



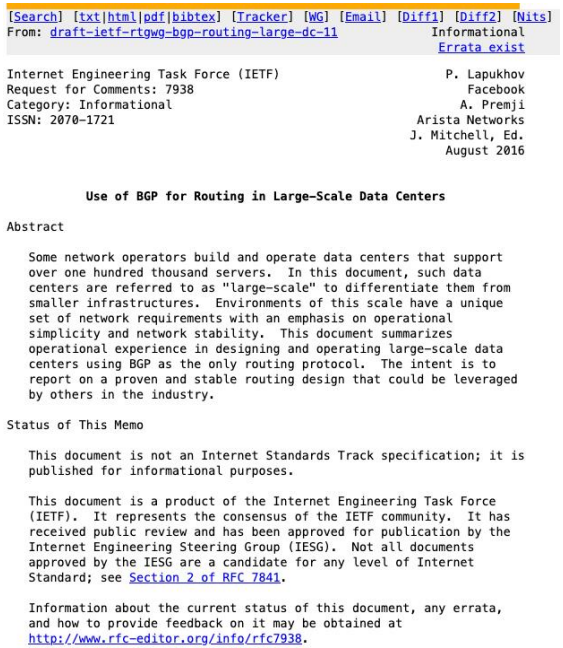
*”Cloud Native Applications” are generally defined as Container- or Kubernetes-based Applications

‘Reducing the number of Control- and Data-Plane Protocols in the Data Center’

Building for the “Cloud Native Application” … and other use cases

What is RFC 7938?

By the Standards Body



[Search] [txt|html|pdf|bibtex] [Tracker] [WG] [Email] [Diff1] [Diff2] [Nits]
From: draft-ietf-rtgwg-bgp-routing-large-dc-11 Informational
Errata exist

Internet Engineering Task Force (IETF) P. Lapukhov
Request for Comments: 7938 Facebook
Category: Informational A. Premji
ISSN: 2070-1721 Arista Networks
J. Mitchell, Ed.
August 2016

Use of BGP for Routing in Large-Scale Data Centers

Abstract

Some network operators build and operate data centers that support over one hundred thousand servers. In this document, such data centers are referred to as "large-scale" to differentiate them from smaller infrastructures. Environments of this scale have a unique set of network requirements with an emphasis on operational simplicity and network stability. This document summarizes operational experience in designing and operating large-scale data centers using BGP as the only routing protocol. The intent is to report on a proven and stable routing design that could be leveraged by others in the industry.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7938.

- Categorized as Informational RFC
- Basically, a Design Guide for Leaf/Spine Topologies
  - Checkout my Multi-Tier session
- Chooses EBGP as Routing Protocol for the Data Center
  - A flat Layer-3 only approach
  - No Network Overlays considered
- Is RFC 7938 dated?
  - No specific reference to IPv6
  - Only 2-Byte ASN reference
  - Talks about TRILL for Layer-2

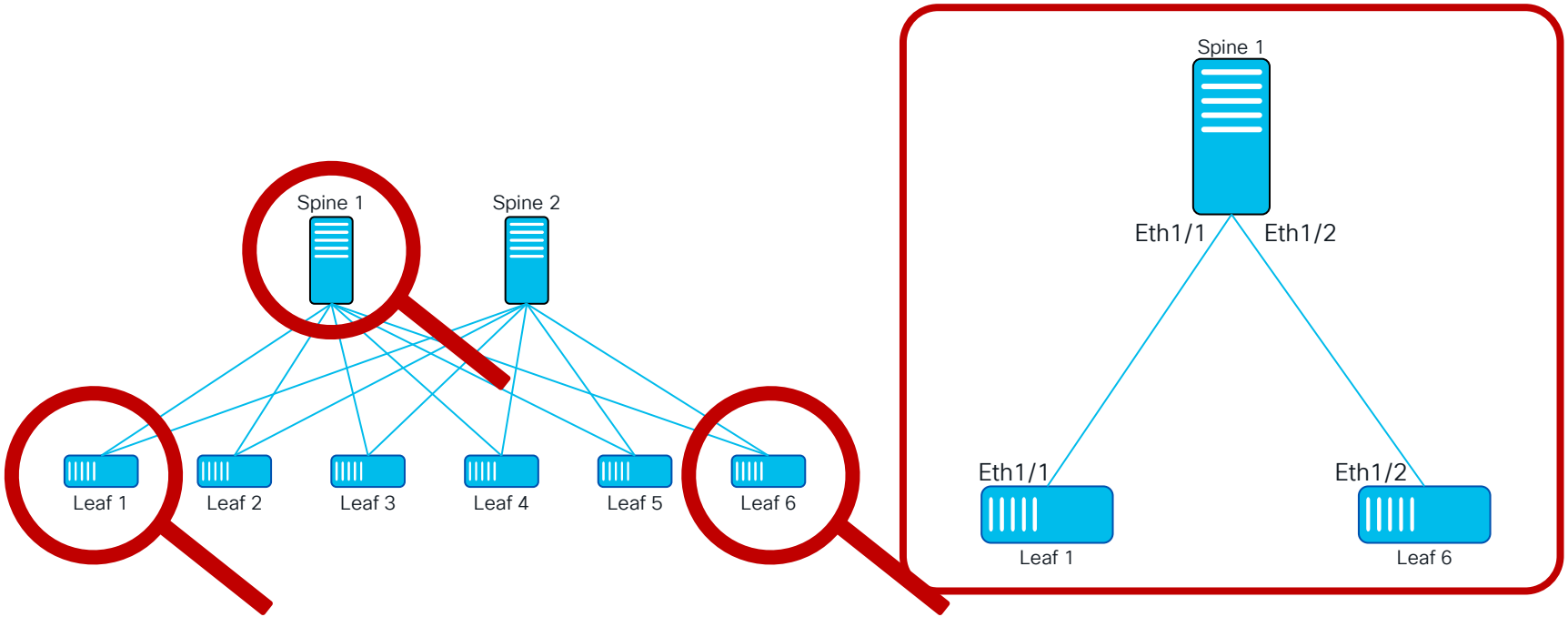
- RFC 7938
  - https://datatracker.ietf.org/doc/html/rfc7938

'Advertising IPv4 & IPv6 Prefix Information with an IPv6 Next Hop enables BGP for Routing in Large-Scale Data Centers to carry IPv4 & IPv6 Address-Family'

How RFC 7938 can leverage RFC 5549 (RFC 8950)

Deployments with RFC 5549 at a glance

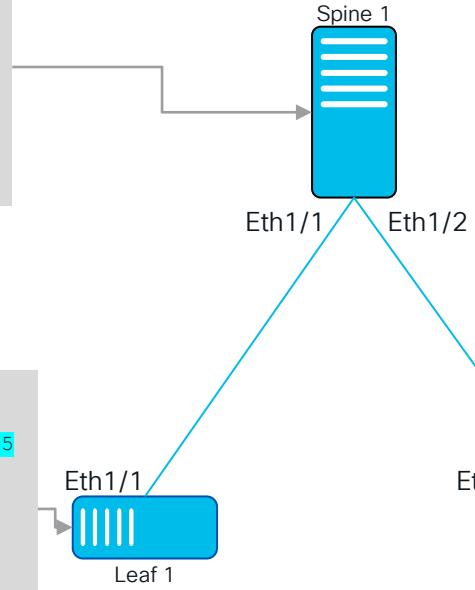
Magnifying some Nodes



local bgp ASN
local bgp RID
bgp neighbor / next-hop / if
peer bgp ASN
IPv6 Link-Local

Adding some Loopacks

[Config]
```
interface loopback0
  ip address 10.51.51.51/32 tag 12345
  ipv6 address fc00::51/128 tag 12345
!
router bgp 65111
 address-family ipv4 unicast
  redistribute direct route-map TAG
 address-family ipv6 unicast
  redistribute direct route-map TAG
```



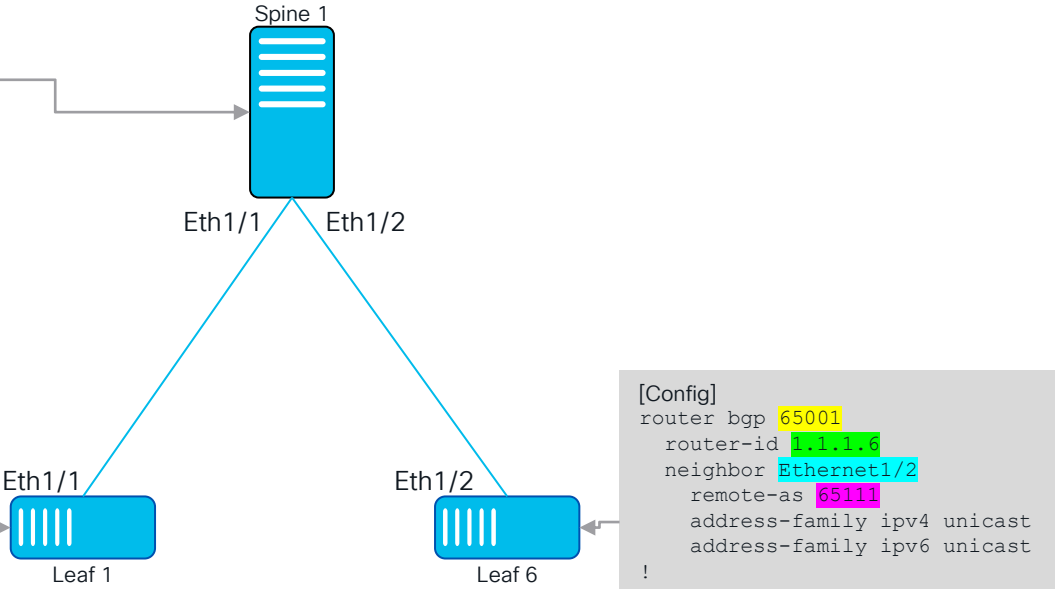
[Config]
```
route-map TAG permit 10
  match tag 12345
```

[Config]
```
interface loopback0
  ip address 10.131.131.131/32 tag 12345
  ipv6 address fc00::131/128 tag 12345
!
router bgp 65001
 address-family ipv4 unicast
  redistribute direct route-map TAG
 address-family ipv6 unicast
  redistribute direct route-map TAG
```

[Config]
```
interface loopback0
  ip address 10.132.132.132/32 tag 12345
  ipv6 address fc00::132/128 tag 12345
!
router bgp 65001
 address-family ipv4 unicast
  redistribute direct route-map TAG
 address-family ipv6 unicast
  redistribute direct route-map TAG
```

Config per RFC7938 (Dual-AS)

```
[Config]
router bgp 65111
  router-id 111.1.1.1
  neighbor Ethernet1/1-2
    remote-as 65001
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1-2
 ipv6 link-local use-bia
 ip forward
```




```
[Config]
router bgp 65001
  router-id 1.1.1.1
  neighbor Ethernet1/1
    remote-as 65111
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

```
[Config]
router bgp 65001
  router-id 1.1.1.6
  neighbor Ethernet1/2
    remote-as 65111
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/2
 ipv6 link-local use-bia
 ip forward
```

This is NOT just going to work (Source AS = Destination AS) – 2 Different Ways to Remediate

Oper IPv4 per RFC7938 (Dual-AS)

[Config]
```
router bgp 65111
  router-id 111.1.1.1
  neighbor Ethernet1/1-2
    remote-as 65001
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1-2
 ipv6 link-local use-bia
 ip forward
```

[Config]
```
router bgp 65001
  router-id 1.1.1.1
  neighbor Ethernet1/1
    remote-as 65111
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

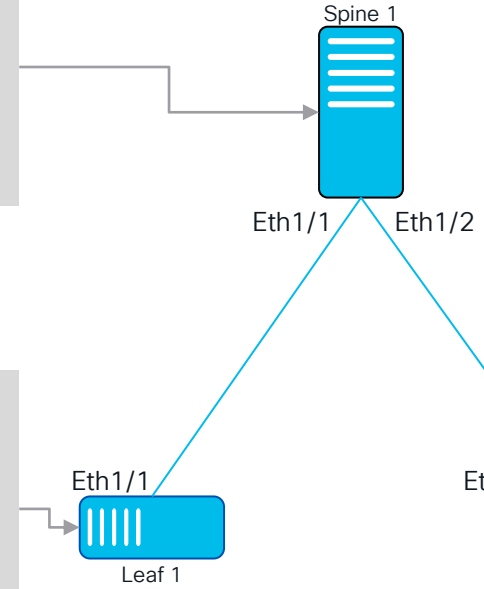
[Output]
```
Leaf1# show ip route
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.131.131.131/32, ubest/mbest: 2/0, attached
    *via 10.131.131.131, Lo0, [0/0], 00:19:19, local, tag 12345
    *via 10.131.131.131, Lo0, [0/0], 00:19:19, direct, tag 12345

Leaf1# show ip bgp
BGP routing table information for VRF default, address family IPv4 Unicast
BGP table version is 8, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2


   Network            Next Hop            Metric     LocPrf     Weight Path
*>r10.131.131.131/32  0.0.0.0                  0        100      32768 ?
```

No Routes from the other Leaf (same ASN)

This is NOT just going to work (Source AS = Destination AS) – 2 Different Ways to Remediate

Config per RFC7938 (Dual-AS with knobs)

```
[Config]
router bgp 65111
  router-id 111.1.1.1
  neighbor Ethernet1/1-2
    remote-as 65001
    inherit peer DISABLE-AS-PATH-CHECK
!
interface Ethernet1/1-2
 ipv6 link-local use-bia
 ip forward
```




```
[Config]
template peer DISABLE-AS-PATH-CHECK
  address-family ipv4 unicast
    allowas-in
    disable-peer-as-check
```

```
[Config]
router bgp 65001
  router-id 1.1.1.1
  neighbor Ethernet1/1
    remote-as 65111
    inherit peer DISABLE-AS-PATH-CHECK
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

```
[Config]
router bgp 65001
  router-id 1.1.1.6
  neighbor Ethernet1/2
    remote-as 65111
    inherit peer DISABLE-AS-PATH-CHECK
!
interface Ethernet1/2
 ipv6 link-local use-bia
 ip forward
```

Option #1 – Dual-AS; Let's turn some BGP knobs

Oper IPv4 per RFC7938 (Dual-AS with knobs)

[Config]
```
router bgp 65111
  router-id 111.1.1.1
  neighbor Ethernet1/1-2
    remote-as 65001
    inherit peer DISABLE-AS-PATH-CHECK
!
interface Ethernet1/1-2
 ipv6 link-local use-bia
 ip forward
```

[Config]
```
router bgp 65001
  router-id 1.1.1.1
  neighbor Ethernet1/1
    remote-as 65111
    inherit peer DISABLE-AS-PATH-CHECK
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

[Output]
```
Leaf1# show ip route
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.131.131.131/32, ubest/mbest: 2/0, attached
    *via 10.131.131.131, Lo0, [0/0], 00:19:19, local, tag 12345
    *via 10.131.131.131, Lo0, [0/0], 00:19:19, direct, tag 12345
10.132.132.132/32, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe0b:6196%default, Eth1/1, [20/0], 00:00:23, bgp-65001, external, tag 65111

Leaf1# show ip bgp
BGP routing table information for VRF default, address family IPv4 Unicast
BGP table version is 6, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2

   Network            Next Hop            Metric     LocPrf     Weight Path
*>r10.131.131.131/32  0.0.0.0                  0        100      32768 ?
*>e10.132.132.132/32  fe80::720f:6aff:fe0b:6196
                                                                      0 65111 65001 ?
```

Option #1 – Dual-AS; Let's turn some BGP knobs

Oper IPv6 per RFC7938 Dual-AS

[Config]
```
router bgp 65111
  router-id 111.1.1.1
  neighbor Ethernet1/1-2
    remote-as 65001
    inherit peer DISABLE-AS-PATH-CHECK
!
interface Ethernet1/1-2
 ipv6 link-local use-bia
 ip forward
```

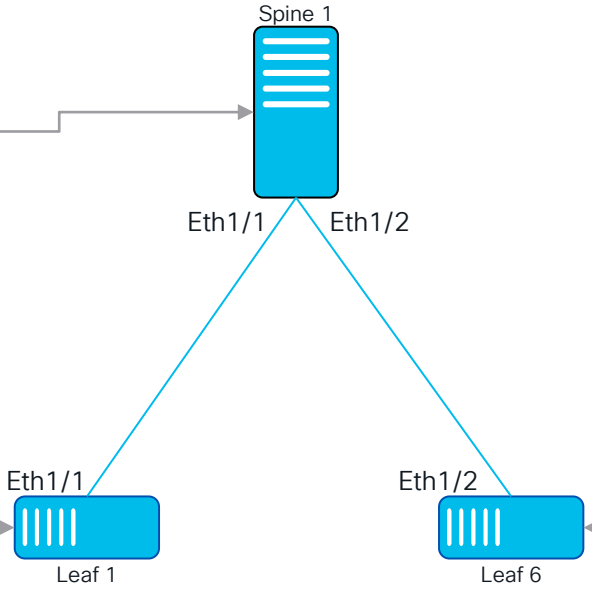
[Config]
```
router bgp 65001
  router-id 1.1.1.1
  neighbor Ethernet1/1
    remote-as 65111
    inherit peer DISABLE-AS-PATH-CHECK
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

[Output]
```
Leaf1# show ipv6 route
IPv6 Routing Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

fc00::131/128, ubest/mbest: 2/0, attached
    *via fc00::131, Lo0, [0/0], 09:48:14, direct, , tag 12345
    *via fc00::131, Lo0, [0/0], 09:48:14, local, tag 12345
fc00::132/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe0b:6196, Eth1/1, [20/0], 00:00:25, bgp-65001, external, tag 65111

Leaf1# show ipv6 bgp
BGP routing table information for VRF default, address family IPv6 Unicast
BGP table version is 11, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2

   Network            Next Hop            Metric     LocPrf     Weight Path
*>rfc00::131/128      0::                      0        100      32768 ?
*>efc00::132/128      fe80::720f:6aff:fe0b:6196
                                                                      0 65111 65001 ?
```

Option #1 – Dual-AS; Let's turn some BGP knobs

Config of Multi-AS

```
[Config]
router bgp 65111
  router-id 111.1.1.1
  neighbor Ethernet1/1
    remote-as 65001
    address-family ipv4 unicast
    address-family ipv6 unicast
 neighbor Ethernet1/2
    remote-as 65006
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```




```
[Config]
router bgp 65001
  router-id 1.1.1.1
  neighbor Ethernet1/1
    remote-as 65111
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

```
[Config]
router bgp 65006
  router-id 1.1.1.6
  neighbor Ethernet1/2
    remote-as 65111
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/2
 ipv6 link-local use-bia
 ip forward
```

Option #2 - Multi-AS; each Switch will get its own AS (more in BGP Auto-Fabric)

Oper of IPv4 in Multi-AS

[Config]
```
router bgp 65111
  router-id 111.1.1.1
  neighbor Ethernet1/1
    remote-as 65001
    address-family ipv4 unicast
    address-family ipv6 unicast
 neighbor Ethernet1/2
    remote-as 65006
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

[Config]
```
router bgp 65001
  router-id 1.1.1.1
  neighbor Ethernet1/1
    remote-as 65111
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

[Output]
```
Leaf1# show ip route
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.131.131.131/32, ubest/mbest: 2/0, attached
    *via 10.131.131.131, Lo0, [0/0], 00:19:19, local, tag 12345
    *via 10.131.131.131, Lo0, [0/0], 00:19:19, direct, tag 12345
10.132.132.132/32, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe0b:6196%default, Eth1/1, [20/0], 00:00:23, bgp-65001, external, tag 65111

Leaf1# show ip bgp
BGP routing table information for VRF default, address family IPv4 Unicast
BGP table version is 6, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2

   Network            Next Hop            Metric     LocPrf     Weight Path
*>r10.131.131.131/32  0.0.0.0                  0        100      32768 ?
*>e10.132.132.132/32  fe80::720f:6aff:fe0b:6196
                                                                      0 65111 65006 ?
```

Option #2 - Multi-AS; each Switch will get its own AS (more in BGP Auto-Fabric)

Oper of IPv6 in Multi-AS

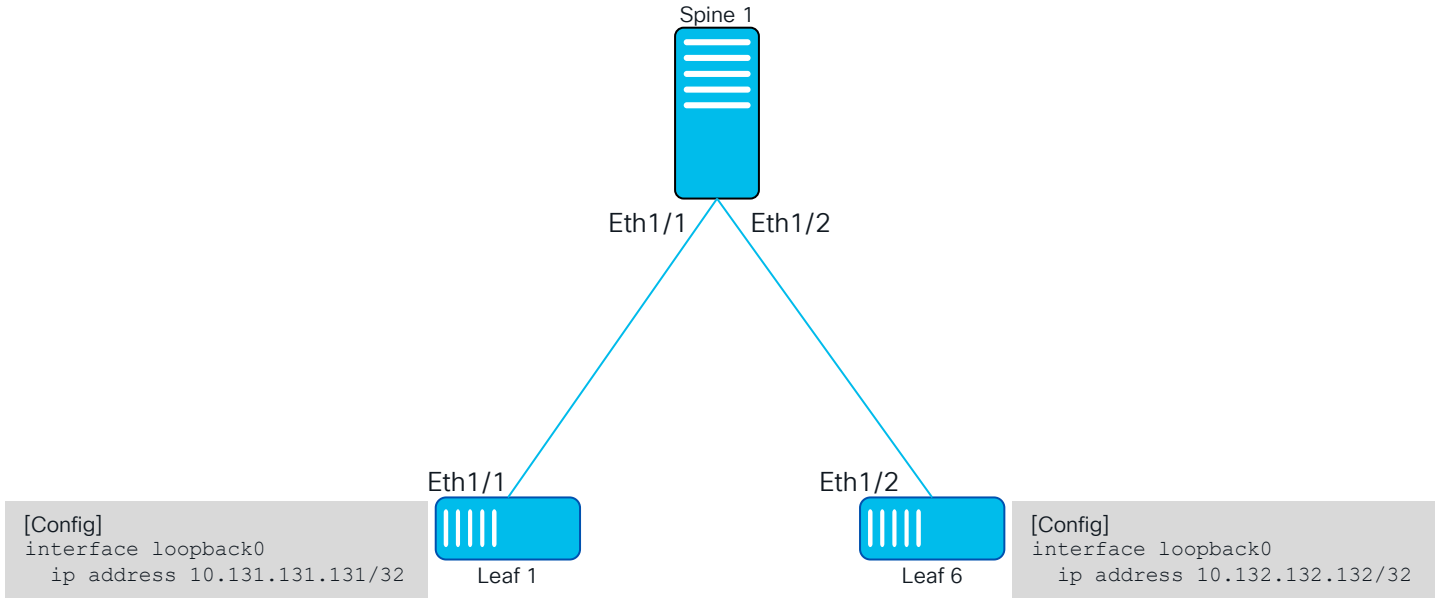
[Config]
```
router bgp 65111
  router-id 111.1.1.1
  neighbor Ethernet1/1
    remote-as 65001
    address-family ipv4 unicast
    address-family ipv6 unicast
 neighbor Ethernet1/2
    remote-as 65006
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

[Config]
```
router bgp 65001
  router-id 1.1.1.1
  neighbor Ethernet1/1
    remote-as 65111
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

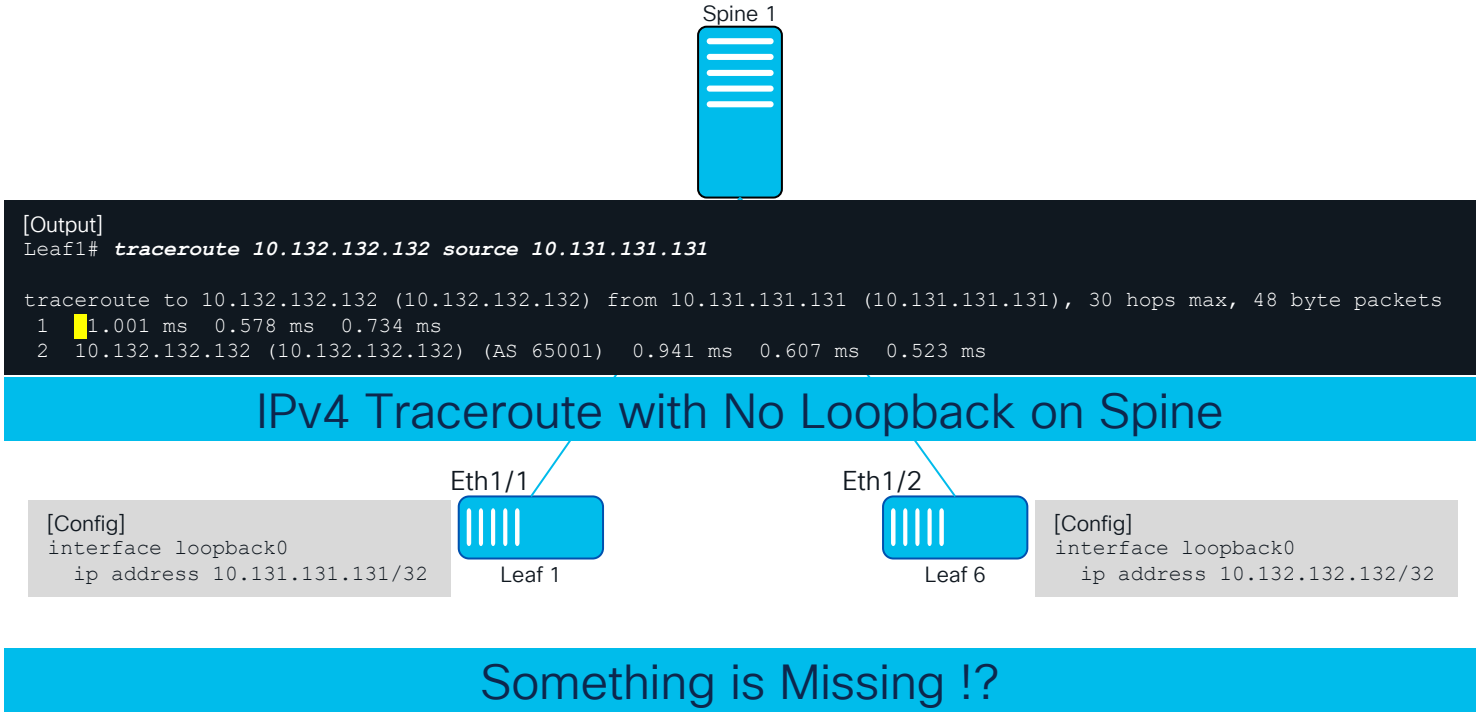
[Output]
```
Leaf1# show ipv6 route
IPv6 Routing Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

fc00::131/128, ubest/mbest: 2/0, attached
    *via fc00::131, Lo0, [0/0], 09:48:14, direct, , tag 12345
    *via fc00::131, Lo0, [0/0], 09:48:14, local, tag 12345
fc00::132/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe0b:6196, Eth1/1, [20/0], 00:00:25, bgp-65001, external, tag 65111

Leaf1# show ipv6 bgp
BGP routing table information for VRF default, address family IPv6 Unicast
BGP table version is 11, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2

   Network            Next Hop            Metric     LocPrf     Weight Path
*>rfc00::131/128      0::                      0        100      32768 ?
*>efc00::132/128      fe80::720f:6aff:fe0b:6196
                                                                      0 65111 65006 ?
```

Option #2 - Multi-AS; each Switch will get its own AS (more in BGP Auto-Fabric)

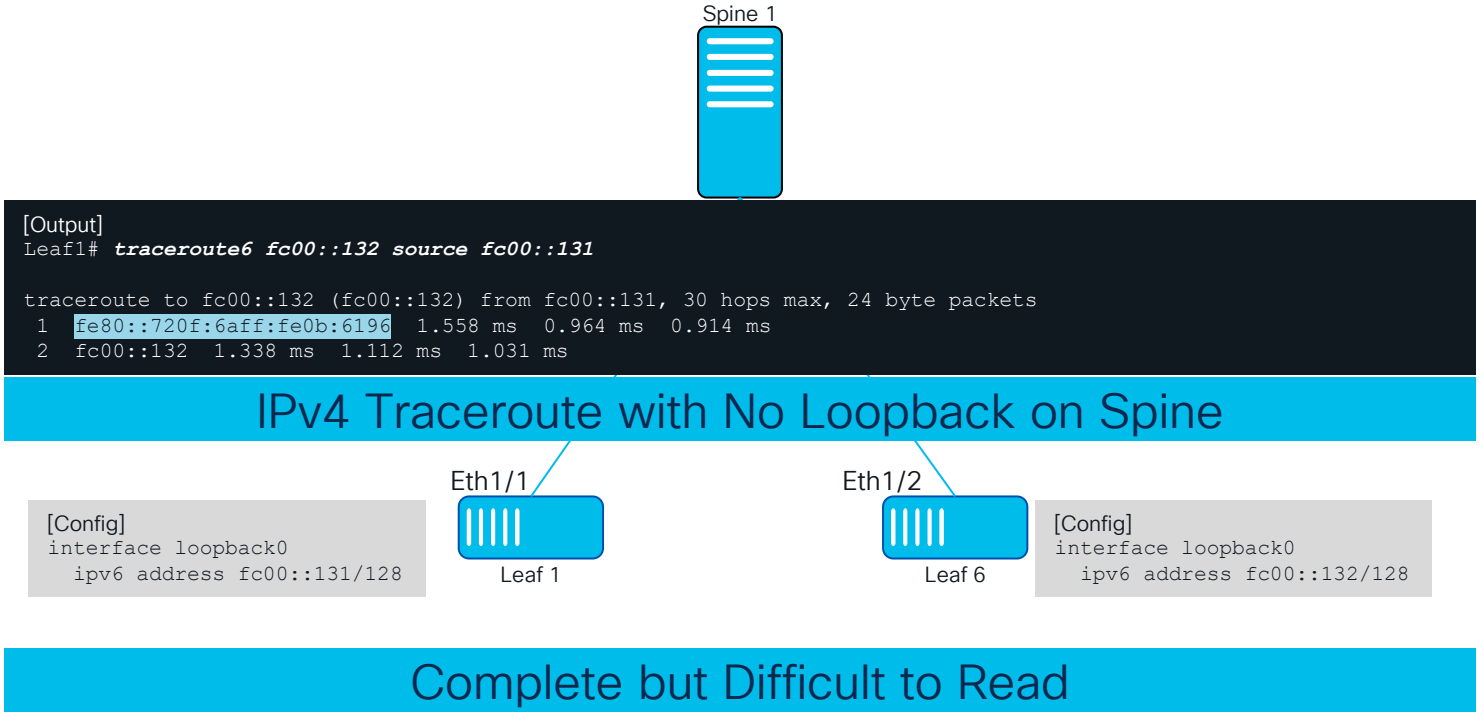
Ping and Traceroute – we need some Loopbacks



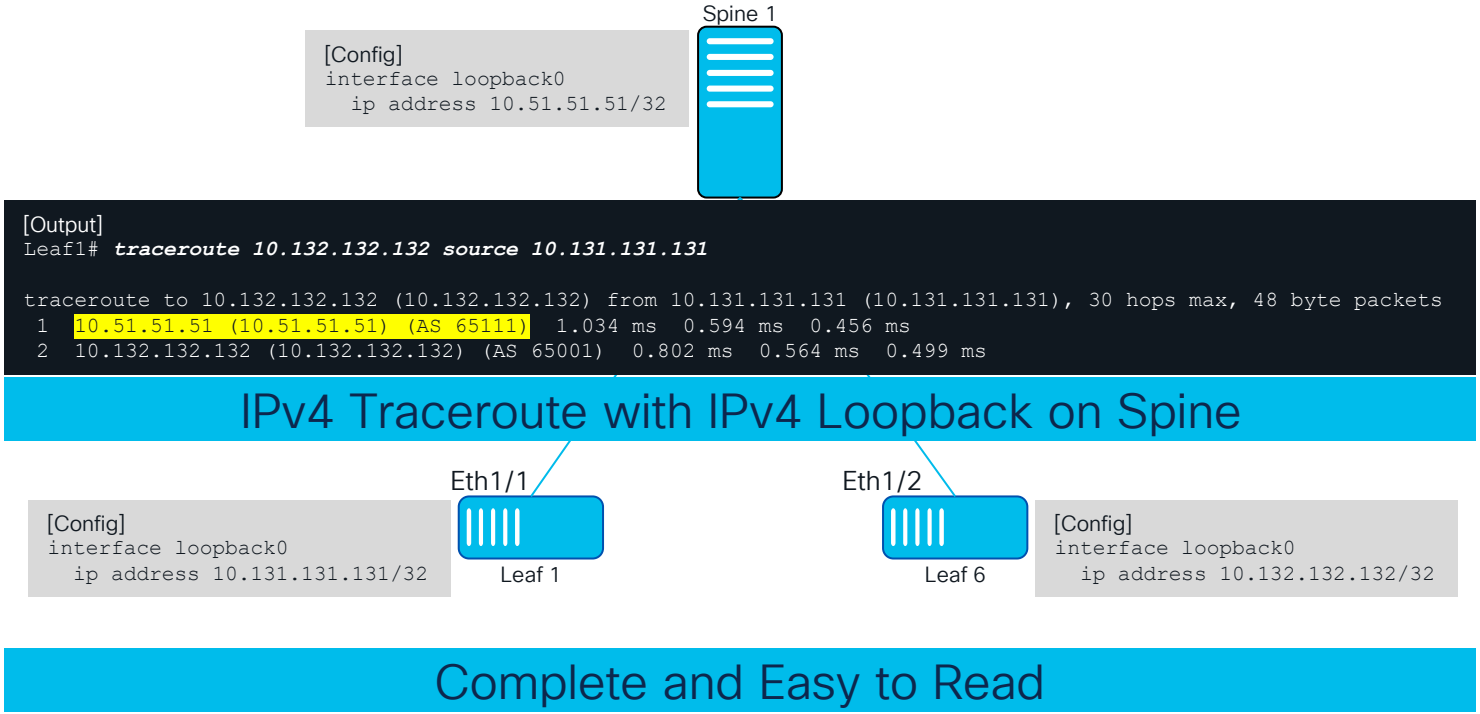
Ping and Traceroute – we need some Loopbacks



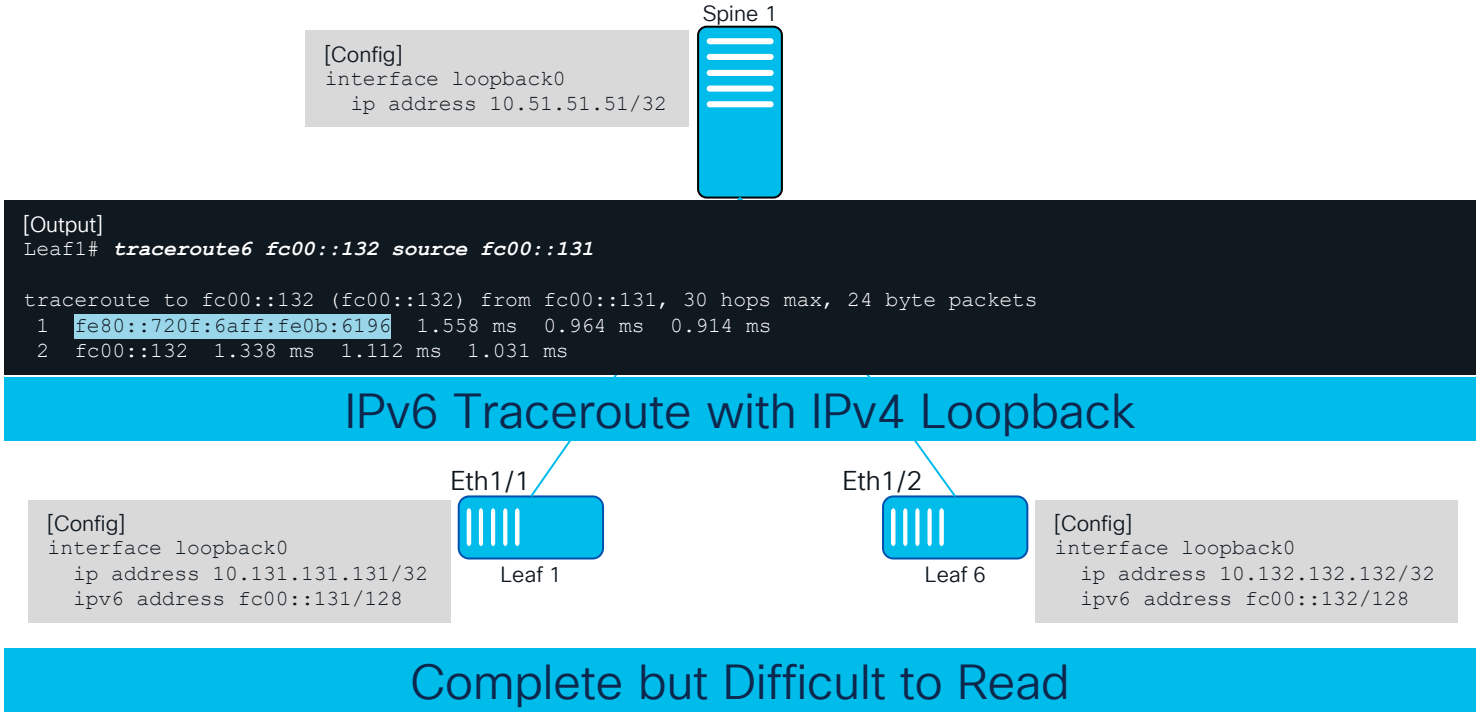
Ping and Traceroute – we need some Loopbacks



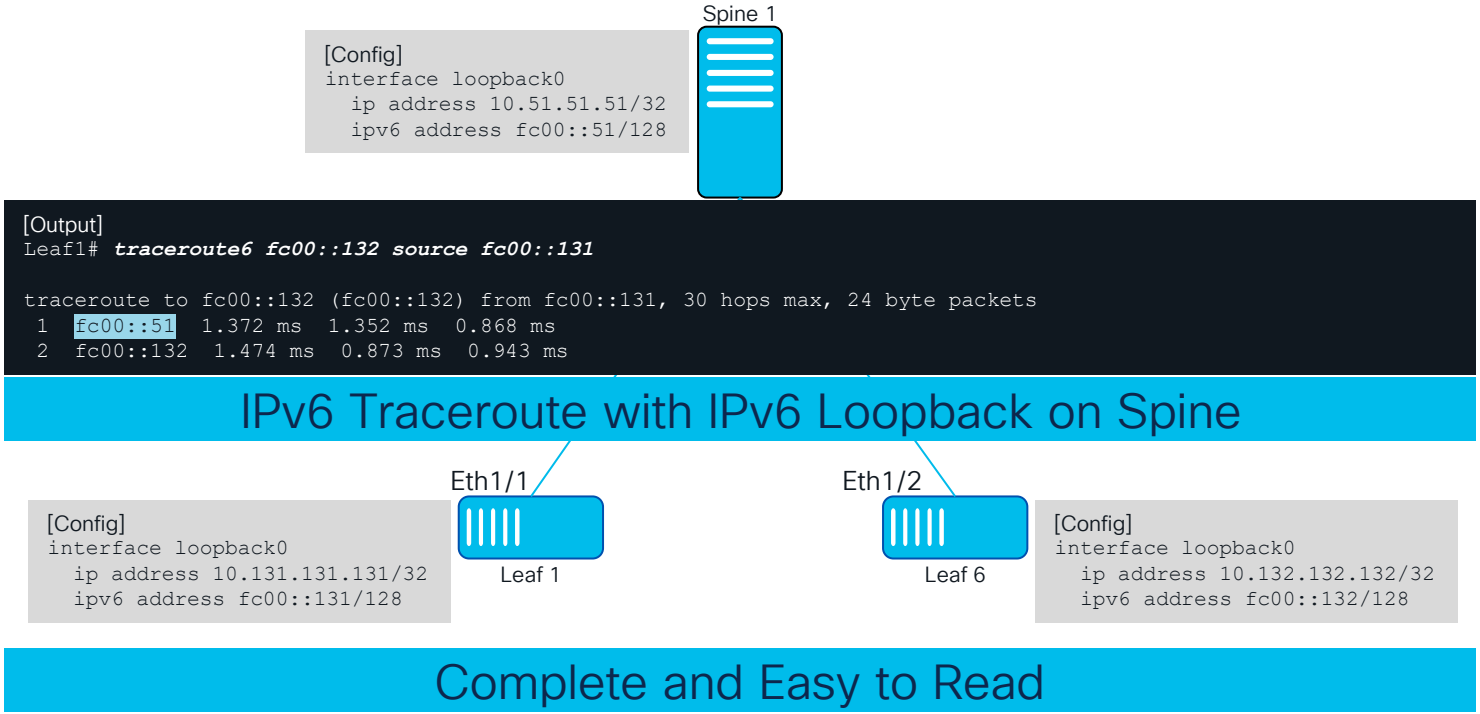
Ping and Traceroute – we need some Loopbacks



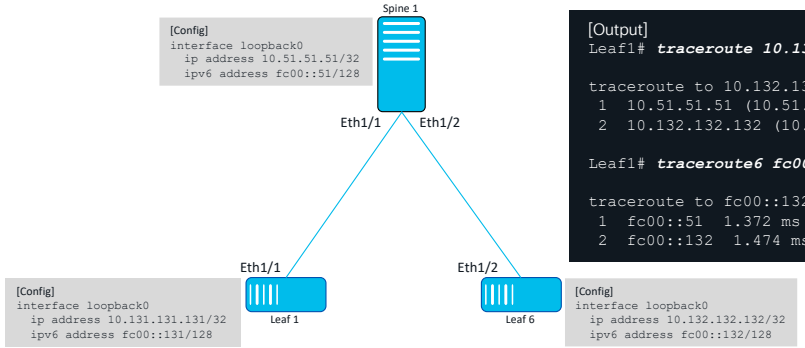
Ping and Traceroute – we need some Loopbacks



Ping and Traceroute – we need some Loopbacks



Ping and Traceroute – we need some Loopbacks

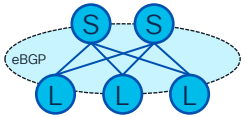



```
[Output]
Leaf1# traceroute 10.132.132.132 source 10.131.131.131

traceroute to 10.132.132.132 (10.132.132.132) from 10.131.131.131 (10.131.131.131), 30 hops max, 48 byte packets
 1  10.51.51.51 (10.51.51.51) (AS 65111)  1.034 ms  0.594 ms  0.456 ms
 2  10.132.132.132 (10.132.132.132) (AS 65001)  0.802 ms  0.564 ms  0.499 ms

Leaf1# traceroute6 fc00::132 source fc00::131

traceroute to fc00::132 (fc00::132) from fc00::131, 30 hops max, 24 byte packets
 1  fc00::51  1.372 ms  1.352 ms  0.868 ms
 2  fc00::132  1.474 ms  0.873 ms  0.943 ms
```

- A Loopback per Switch helps in Operational Tasks
  - For IPv4, add a IPv4 Loopback. For IPv6, add a IPv6 Loopback
- Ping for Connectivity Test
  - Loopback to Loopback
  - Physical Interface to Physical Interface (Link-Local Address)
- Traceroute becomes easy to Read
  - Each Hop clearly identified by the Loopback IP address (IPv4 or IPv6)
  - In Leaf/Spine, Loopback address is sufficient (there is no other path)
- In-band Management (Loopback to Loopback or LLA to LLA)

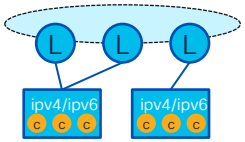
BGP Auto-Fabric

What is BGP Auto-Fabric?



Self Organized BGP Fabric

Autonomously Derives Key Values for BGP
Avoids Per-Interface IP Addressing
Automates BGP ASN and Router-ID
Simplifies BGP Peer Configuration



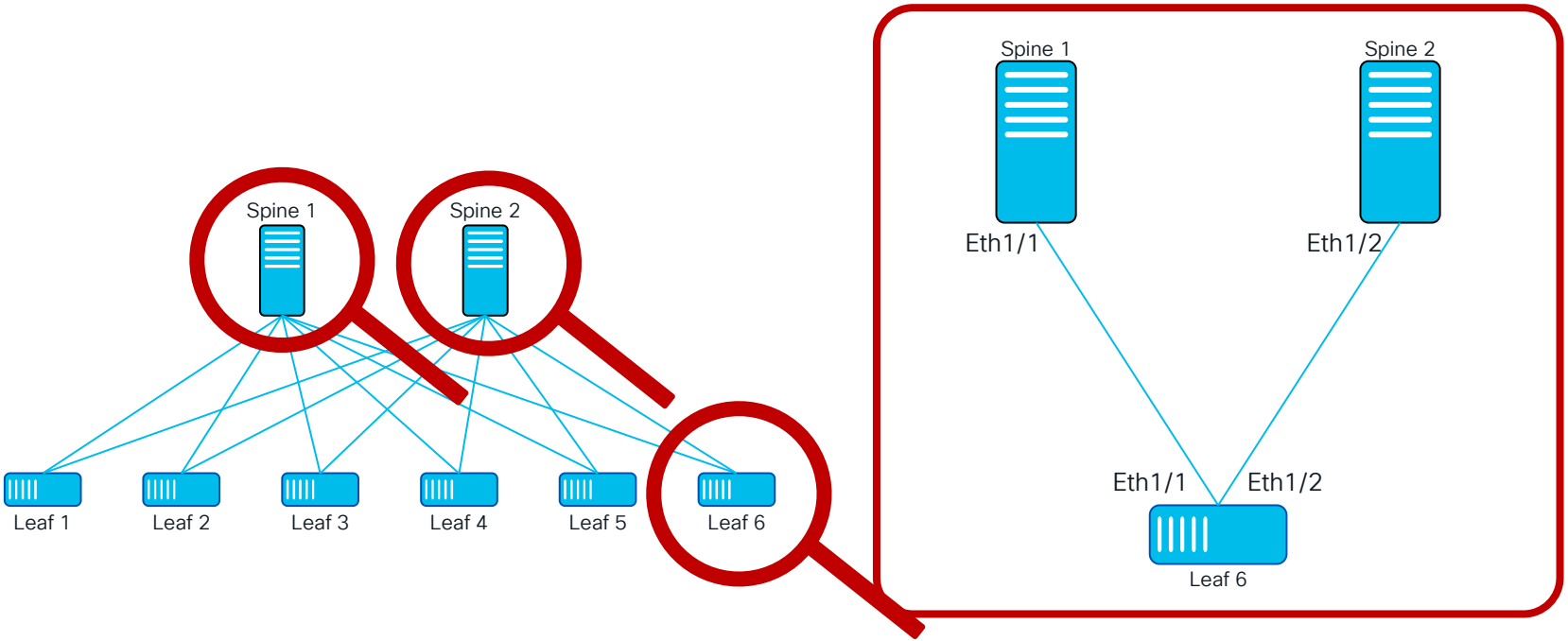
For IPv4 and IPv6 Needs

No Requirement for Dual-Stack Config
Simplifies BGP peering with End-Points
Autonomous Node IP Assignment
Ready for “Cloud Native Applications”*

*”Cloud Native Applications” are generally defined as Container- or Kubernetes-based Applications

BGP Auto-Fabric at a glance

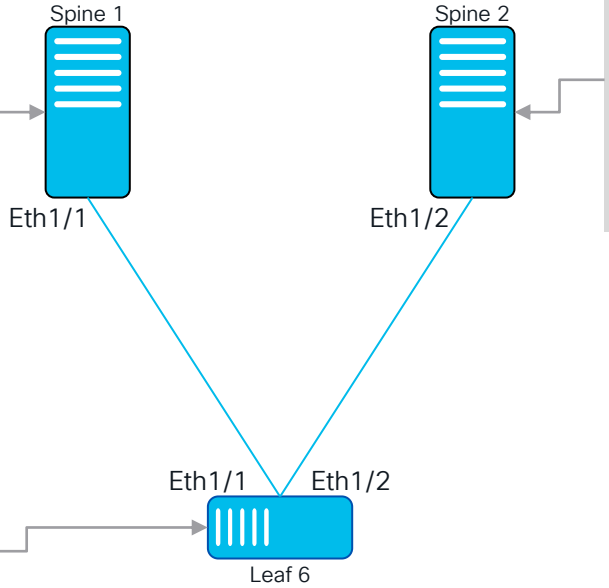
Magnifying some Nodes



Config - BGP Auto-Fabric at a glance

```
[Config]
router bgp auto
  router-id auto
  neighbor Ethernet1/1
    remote-as external
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

```
[Config]
router bgp auto
  router-id auto
  neighbor Ethernet1/2
    remote-as external
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/2
 ipv6 link-local use-bia
 ip forward
```




```
[Config]
router bgp auto
  router-id auto
  neighbor Ethernet1/1-2*
    remote-as external
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1-2
 ipv6 link-local use-bia
 ip forward
```

*BGP interface peering with interface range definition will be available in a subsequent release (roadmap)

Config Results - Auto Derived

```
[Config]
router bgp auto
  router-id auto
  neighbor Ethernet1/1
    remote-as external
    address-family ipv4 unicast
    address-family ipv6 unicast
!
interface Ethernet1/1
 ipv6 link-local use-bia
 ip forward
```

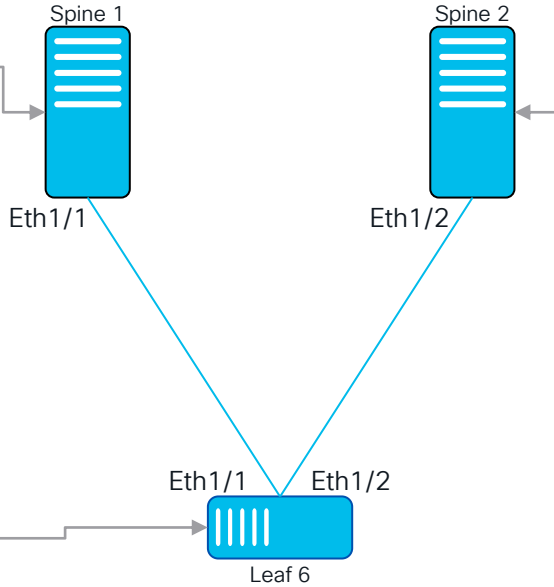
Auto/Peering Seed Values
Global: [System MAC]-#-#
Per-Interface: [Interface MAC]-#-#

```
[Output]
Spine1# show bgp sessions
Total peers 1, established peers 1
ASN 4272508914
VRF default, local ASN 4272508914
peers 1, established peers 1, local router-id 21.77.167.239
State: I-Idle, A-Active, O-Open, E-Established, C-Closing, S-Shutdown

Neighbor        ASN    Flaps LastUpDn|LastRead|LastWrit St Port(L/R)  Notif(S/R)
fe80::720f:6aff:fe0b:6196%Ethernet1/1
                4268165528 0     01:01:59|00:00:52|00:00:19 E   23388/179        0/0

Spine1#
Spine1# show ipv6 interface brief
IPv6 Interface Status for VRF "default"(1)
Interface        IPv6 Address/Link-local Address          Interface Status
                                                          prot/link/admin
Eth1/1           fe80::720f:6aff:fe4d:a7f0                up/up/up
                 fe80::720f:6aff:fe4d:a7f0

Spine1#
Spine1# show ip interface brief

IP Interface Status for VRF "default"(1)
Interface            IP Address      Interface Status
Eth1/1               forward-enabled protocol-up/link-up/admin-up
```

local bgp ASN
local bgp RID
bgp neighbor / next-hop / if
peer bgp ASN
IPv6 Link-Local

Adding some Loopacks

[Config]
```
interface loopback0
  ip address 10.131.131.131/32 tag 12345
  ipv6 address fc00::131/128 tag 12345
!
router bgp auto
 address-family ipv4 unicast
  redistribute direct route-map TAG
 address-family ipv6 unicast
  redistribute direct route-map TAG
```

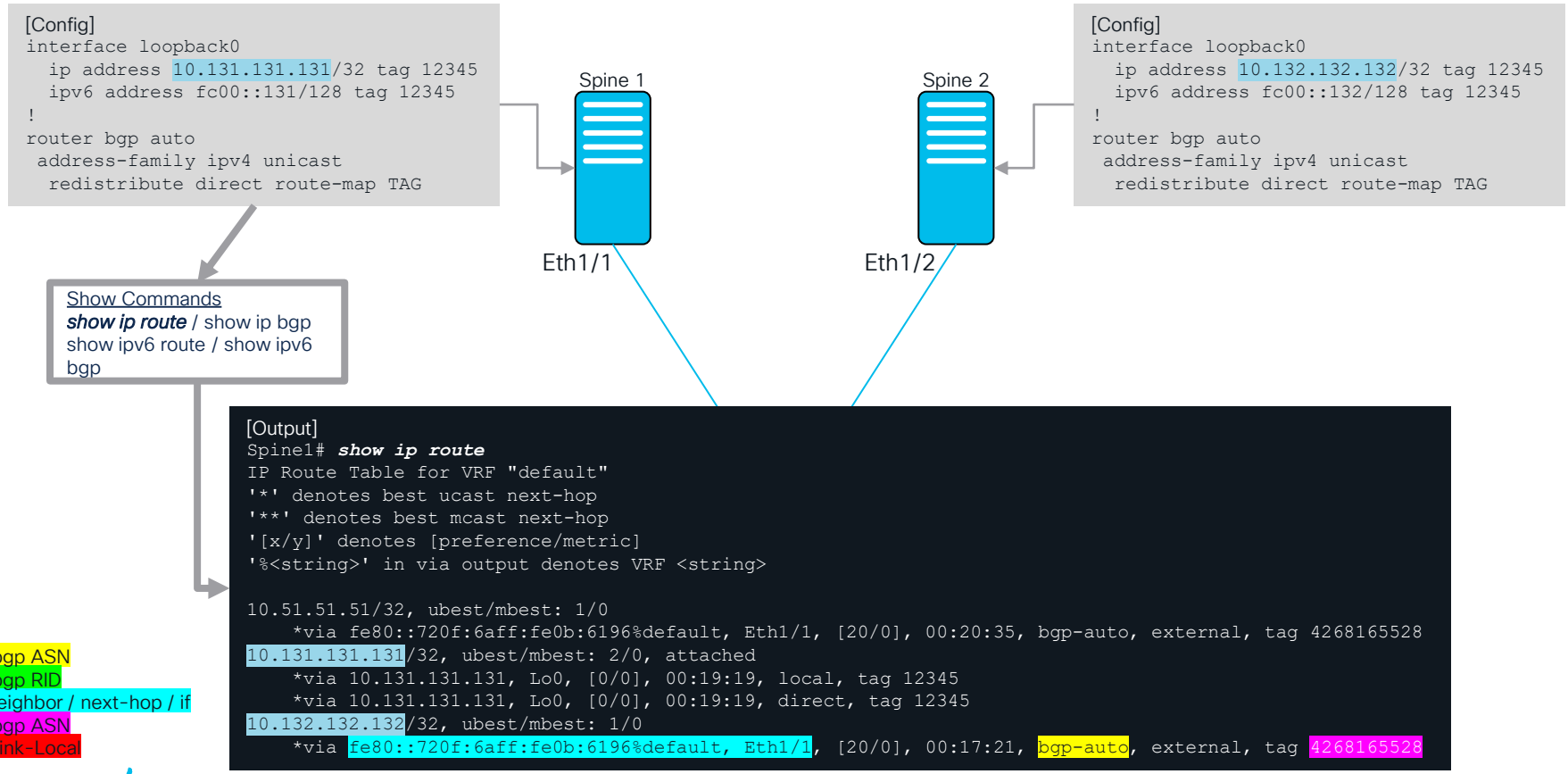


[Config]
```
interface loopback0
  ip address 10.132.132.132/32 tag 12345
  ipv6 address fc00::132/128 tag 12345
!
router bgp auto
 address-family ipv4 unicast
  redistribute direct route-map TAG
 address-family ipv6 unicast
  redistribute direct route-map TAG
```

[Config]
```
route-map TAG permit 10
  match tag 12345
```

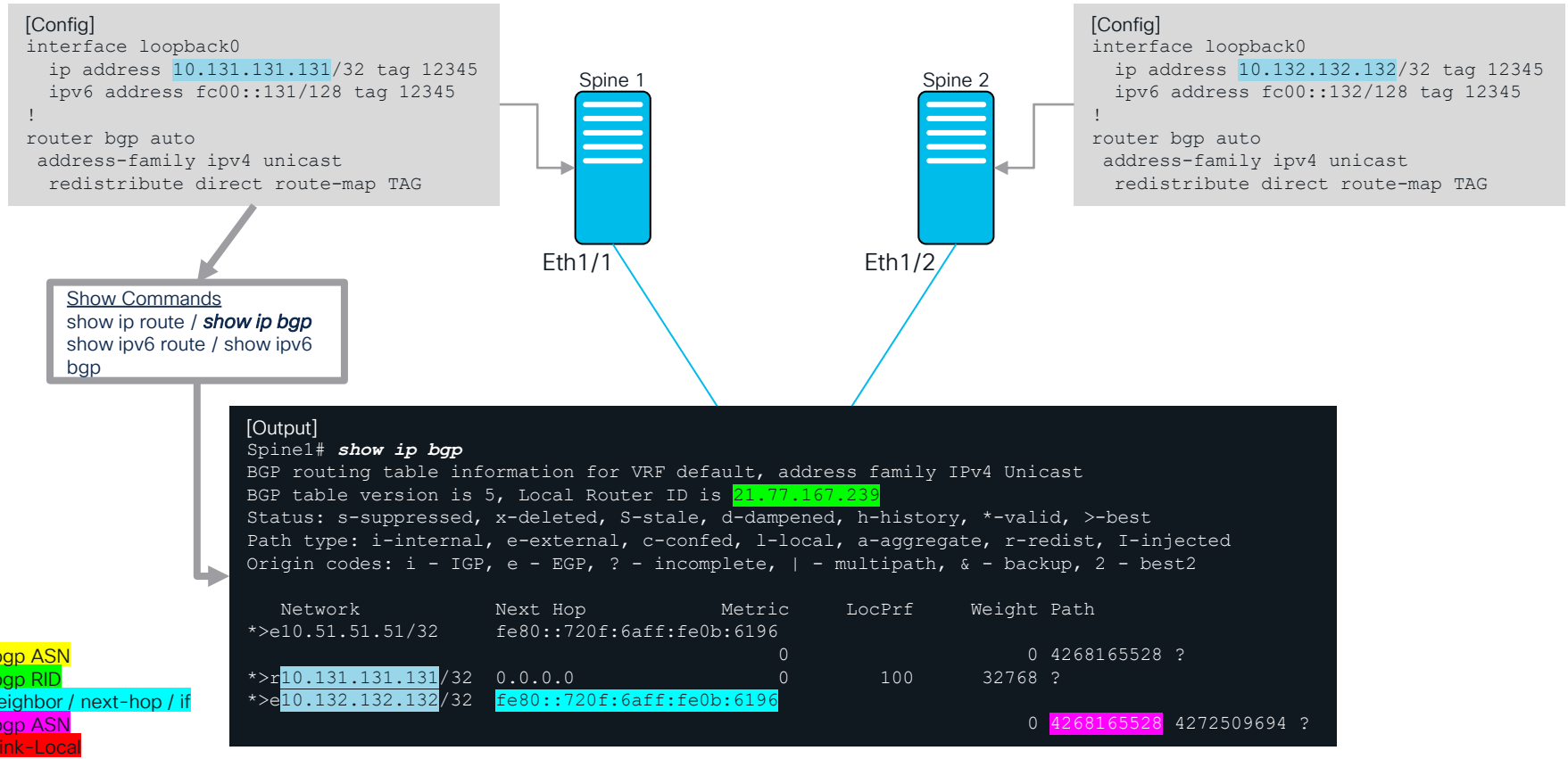
[Config]
```
interface loopback0
  ip address 10.51.51.51/32 tag 12345
  ipv6 address fc00::51/128 tag 12345
!
router bgp auto
 address-family ipv4 unicast
  redistribute direct route-map TAG
 address-family ipv6 unicast
  redistribute direct route-map TAG
```

IPv4 Routing Output Example

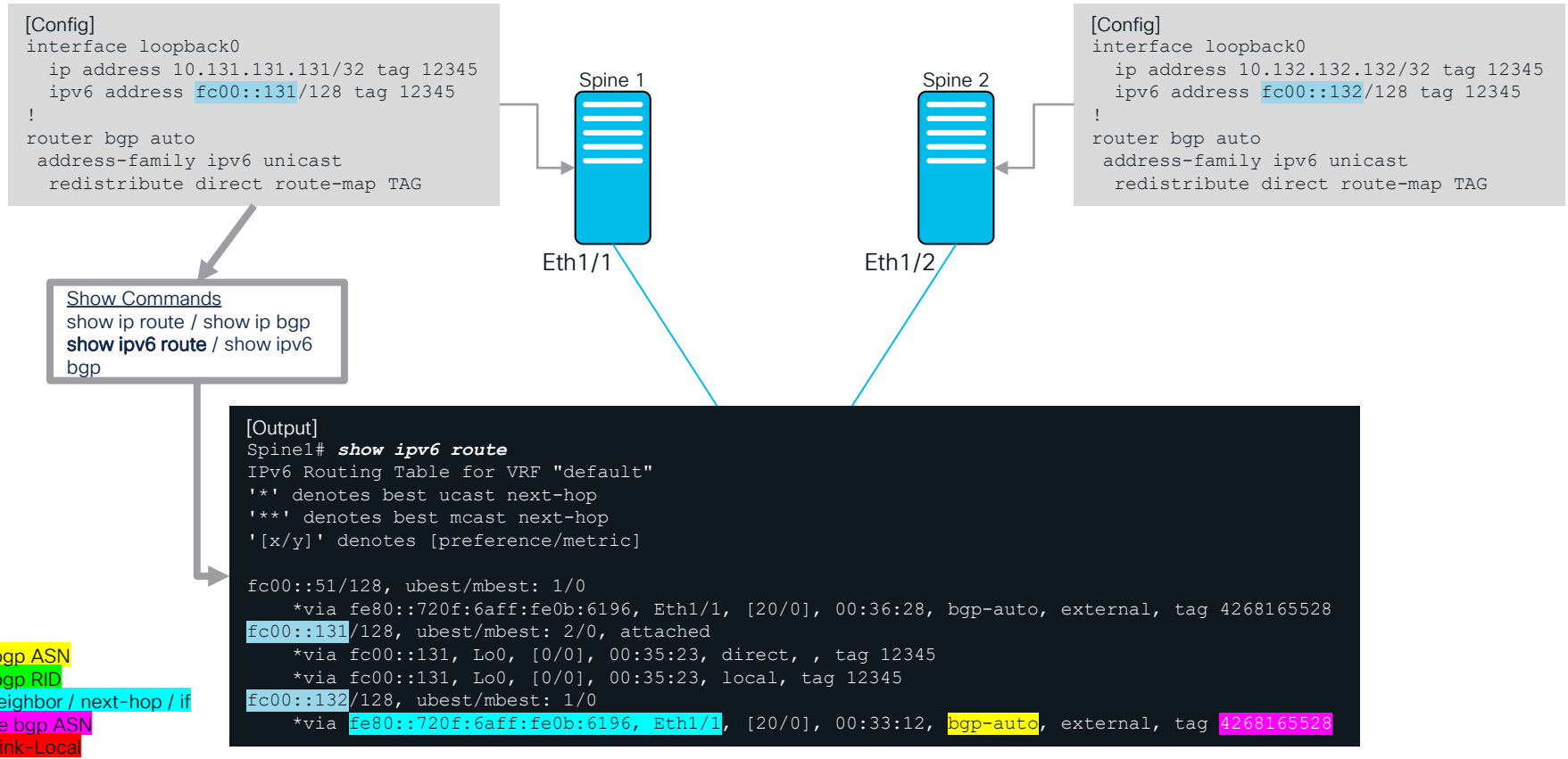


local bgp ASN
local bgp RID
bgp neighbor / next-hop / if
peer bgp ASN
IPv6 Link-Local

IPv4 Routing Output Example

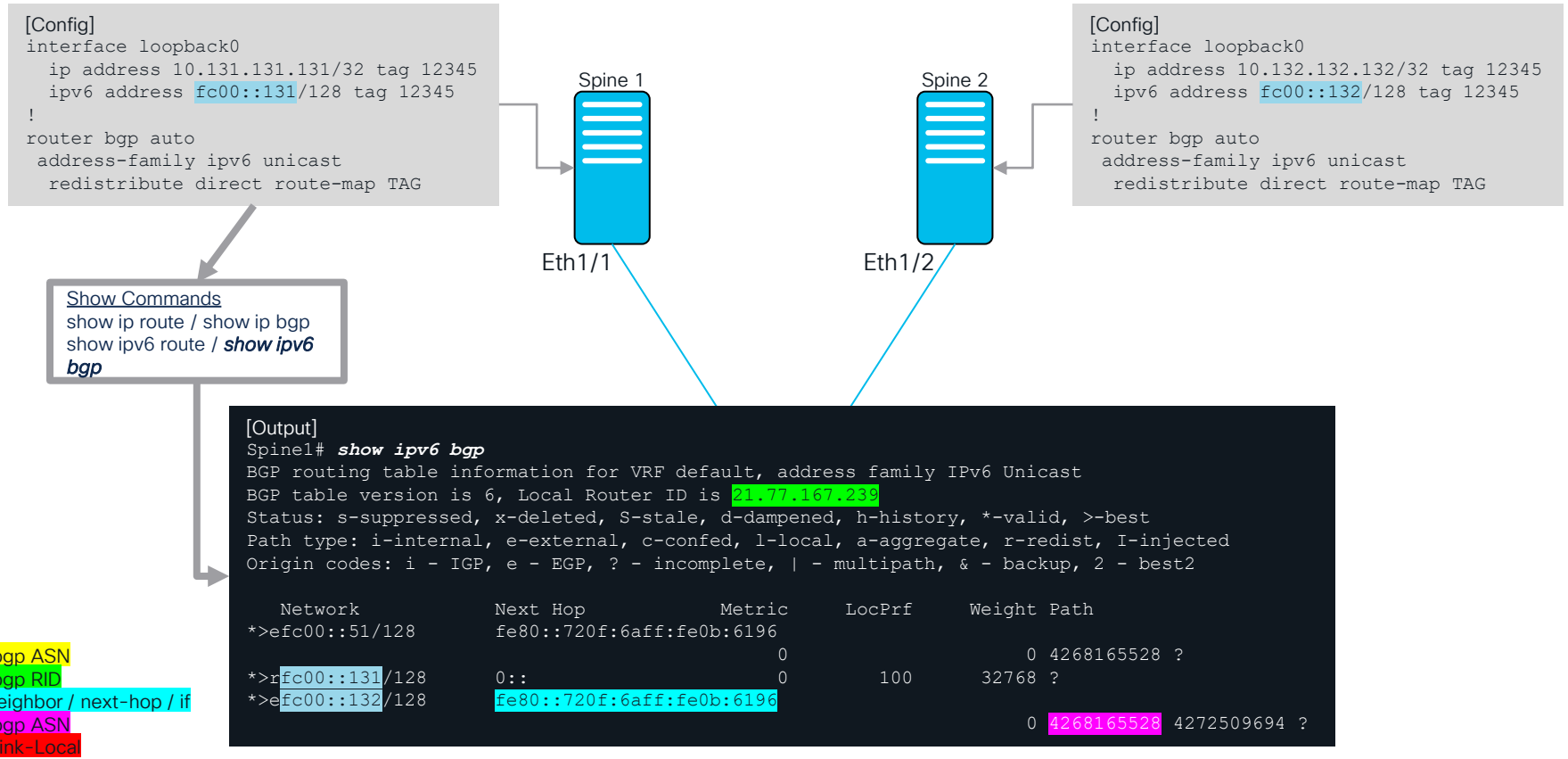


IPv6 Routing Output Example



local bgp ASN
local bgp RID
bgp neighbor / next-hop / if
remote bgp ASN
IPv6 Link-Local

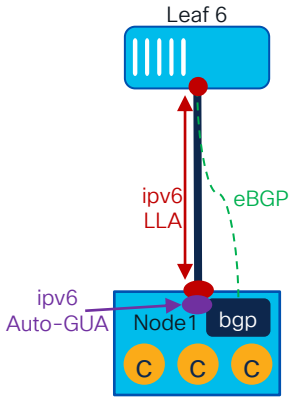
IPv6 Routing Output Example



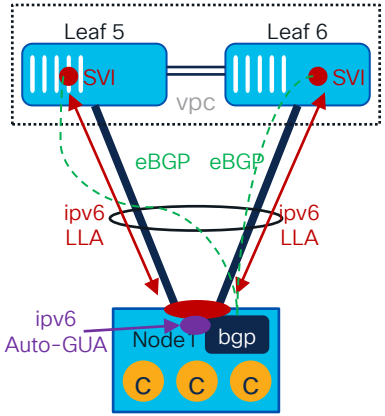
Host Attachments

BGP Auto-Fabric at a glance

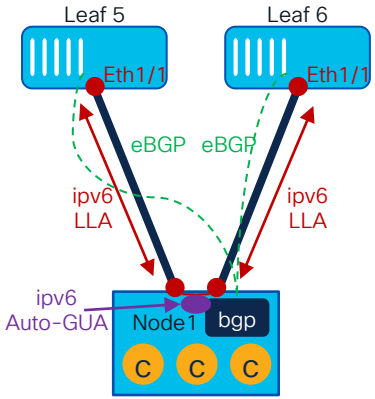
Single Attached Host



Dual Attached Host (Multi-Chassis LAG)

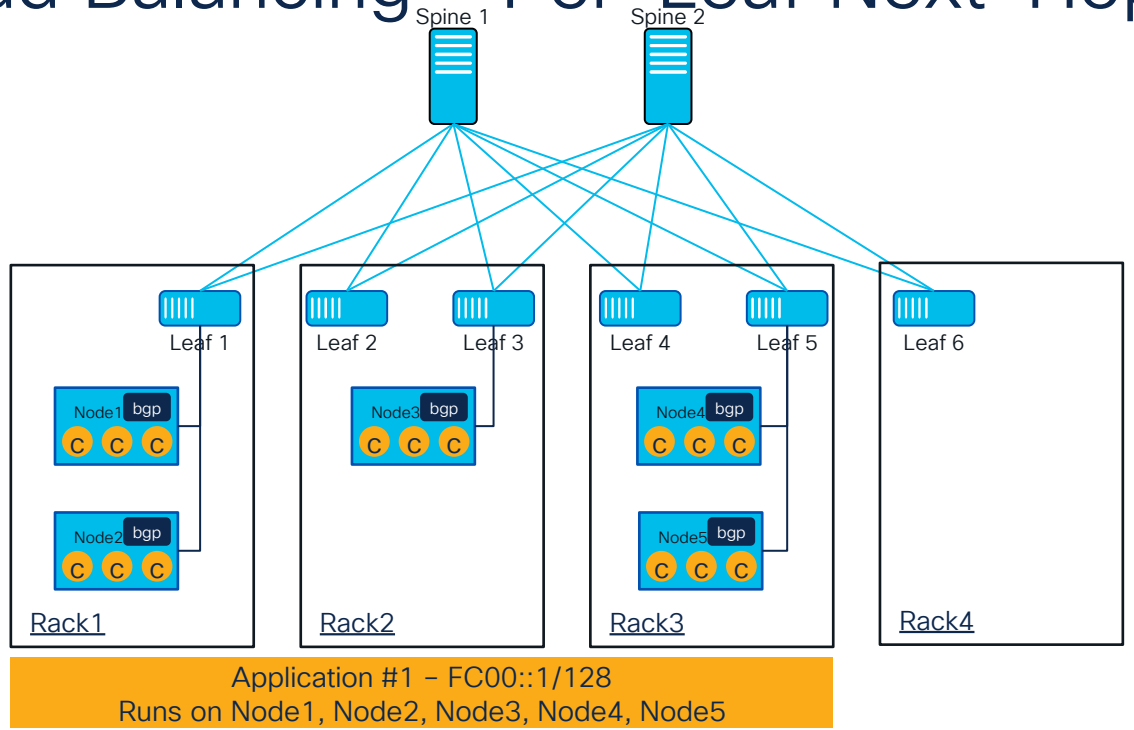


Dual Attached Host (Layer-3 ECMP)

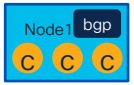


‘Kubernetes (K8s) Infrastructure Connectivity – Network Designs for the Modern Data Center’

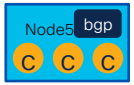
BGP Load Balancing – Per-Leaf Next-Hop



Node IP Addressing Example:
IPv6 LLA – FE80::MAC
IPv6 GUA – FC00::RackID:NodeID



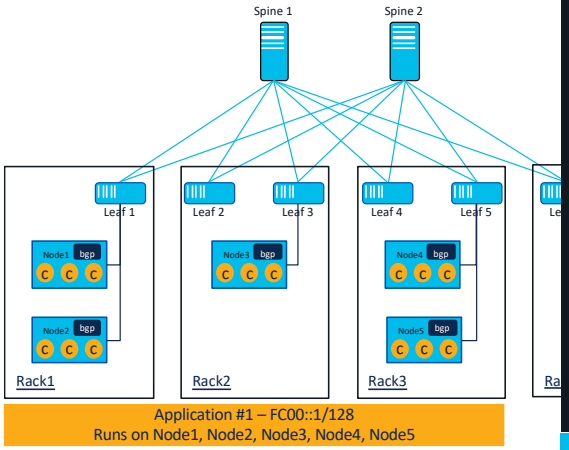
MAC Address – a1:b1:c1:d1:e1:11
IPv6 Link-Local Address – FE80::a1b1:c1d1:e111/64
IPv6 Global Unicast Address – FC00::0001:a1b1:c1d1:e111/64



MAC Address – a5:b5:c5:d5:e5:55
IPv6 Link-Local Address – FE80::a5b5:c5d5:e555/64
IPv6 Global Unicast Address – FC00::0003:a5b5:c5d5:e55

BGP Load Balancing – Per-Leaf Next-Hop

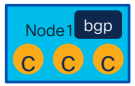
Per-Leaf Next-Hop



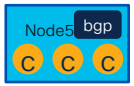

```
[Output]
Spine1# show ipv6 route
IPv6 Routing Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

fc00::1/128, ubest/mbest: 2/0
    *via fe80::720f:6aff:fe4d:a7f0, Eth1/1, [20/0], 00:08:08, bgp-auto, external, tag Rack1-4byteASN
    *via fe80::720f:6aff:fe4d:ab00, Eth1/3, [20/0], 00:17:22, bgp-auto, external, tag Rack2-4byteASN
    *via fe80::720f:6aff:fe4d:a766, Eth1/5, [20/0], 00:22:52, bgp-auto, external, tag Rack3-4byteASN

fc00::0001:a1b1:c1d1:e111/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:a7f0, Eth1/1, [20/0], 04:20:13, bgp-auto, external, tag Rack1-4byteASN
fc00::0001:a2b2:c2d2:e222/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:a7f0, Eth1/1, [20/0], 04:20:13, bgp-auto, external, tag Rack1-4byteASN
fc00::0002:a3b3:c3d3:e333/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:ab00, Eth1/3, [20/0], 04:18:47, bgp-auto, external, tag Rack2-4byteASN
fc00::0003:a4b4:c4d4:e444/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:a766, Eth1/5, [20/0], 04:18:47, bgp-auto, external, tag Rack3-4byteASN
fc00::0003:a5b5:c5d5:e555/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:a766, Eth1/5, [20/0], 04:18:47, bgp-auto, external, tag Rack3-4byteASN
```

Load Balancing to where the Server connects (Leaf)

Node IP Addressing Example:
IPv6 LLA – FE80::MAC
IPv6 GUA – FC00::RackID:NodeID

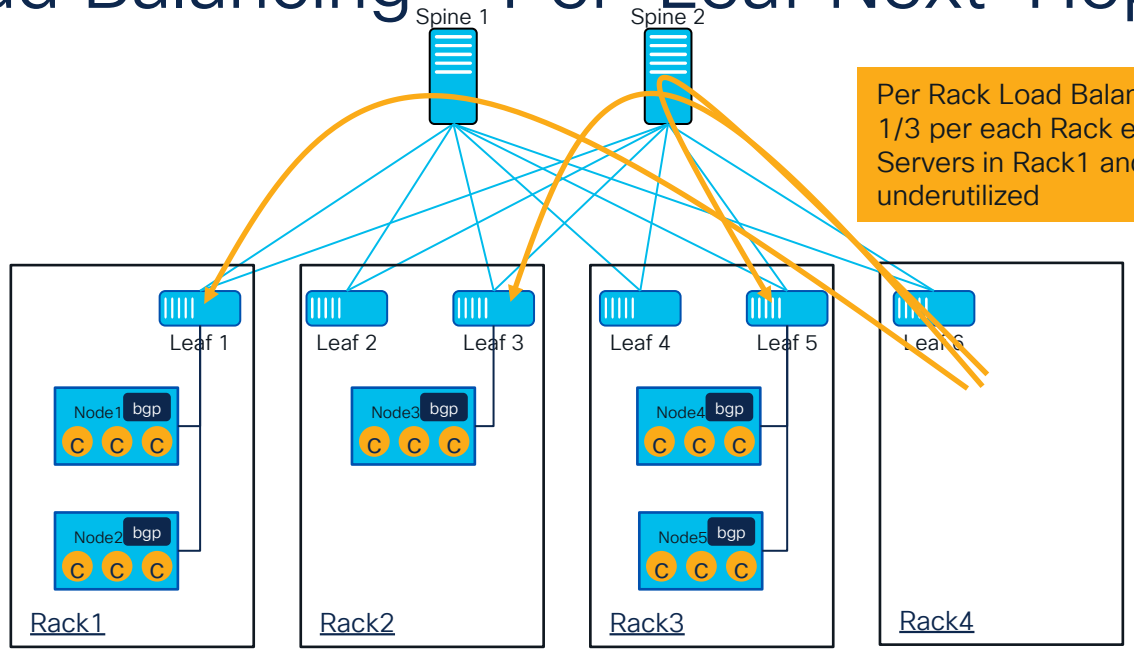


MAC Address – a1:b1:c1:d1:e1:11
IPv6 Link-Local Address – FE80::a1b1:c1d1:e111/64
IPv6 Global Unicast Address – FC00::0001:a1b1:c1d1:e111/64



MAC Address – a5:b5:c5:d5:e5:55
IPv6 Link-Local Address – FE80::a5b5:c5d5:e555/64
IPv6 Global Unicast Address – FC00::0003:a5b5:c5d5:e555/6

BGP Load Balancing – Per-Leaf Next-Hop



Per Rack Load Balancing
1/3 per each Rack even as 5 Server exists
Servers in Rack1 and Rack3 are probably underutilized

Application #1 – FC00::1/128
Runs on Node1, Node2, Node3, Node4, Node5

Node IP Addressing Example:
IPv6 LLA – FE80::MAC
IPv6 GUA – FC00::RackID:NodeID

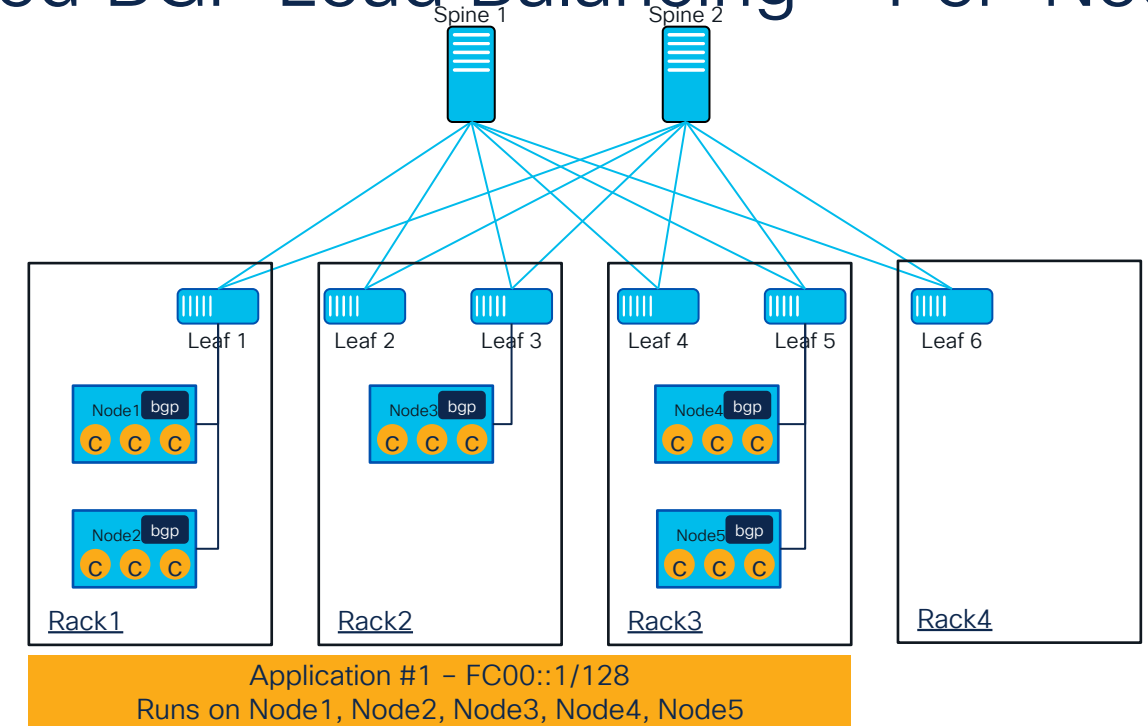
Node1 bgp

MAC Address – a1:b1:c1:d1:e1:11
IPv6 Link-Local Address – FE80::a1b1:c1d1:e111/64
IPv6 Global Unicast Address – FC00::0001:a1b1:c1d1:e111/64

Node5 bgp

MAC Address – a5:b5:c5:d5:e5:55
IPv6 Link-Local Address – FE80::a5b5:c5d5:e555/64
IPv6 Global Unicast Address – FC00::0003:a5b5:c5d5:e55

Optimized BGP Load Balancing – Per-Node NH



Node IP Addressing Example:
IPv6 LLA – FE80::MAC
IPv6 GUA – FC00::RackID:NodeID

Node1 bgp

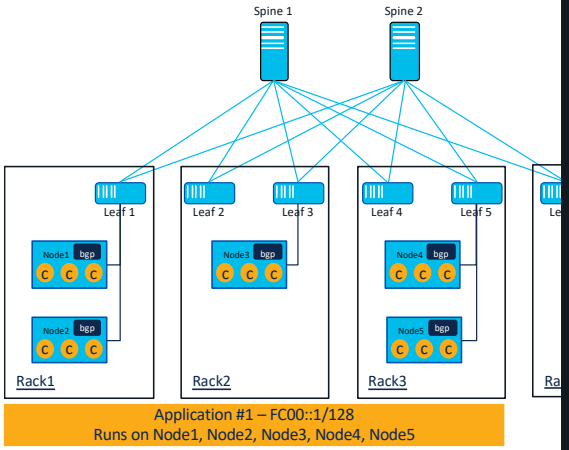
MAC Address – a1:b1:c1:d1:e1:11
IPv6 Link-Local Address – FE80::a1b1:c1d1:e111/64
IPv6 Global Unicast Address – FC00::0001:a1b1:c1d1:e111/64

Node5 bgp

MAC Address – a5:b5:c5:d5:e5:55
IPv6 Link-Local Address – FE80::a5b5:c5d5:e555/64
IPv6 Global Unicast Address – FC00::0003:a5b5:c5d5:e55

Optimized BGP Load Balancing – Per-Node NH

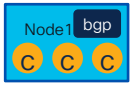
Per-Node Next-Hop



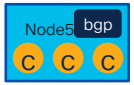

```
[Output]
Spine1# show ipv6 route
IPv6 Routing Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

fc00::1/128, ubest/mbest: 2/0
    *via fc00::0001:a1b1:c1d1:e111, Eth1/1, [20/0], 00:08:08, bgp-auto, external, tag Rack1-4byteASN
    *via fc00::0001:a2b2:c2d2:e222, Eth1/1, [20/0], 00:08:08, bgp-auto, external, tag Rack1-4byteASN
    *via fc00::0002:a3b3:c3d3:e333, Eth1/3, [20/0], 00:17:22, bgp-auto, external, tag Rack2-4byteASN
    *via fc00::0003:a4b4:c4d4:e444, Eth1/5, [20/0], 00:22:52, bgp-auto, external, tag Rack3-4byteASN
    *via fc00::0003:a5b5:c5d5:e555, Eth1/5, [20/0], 00:22:52, bgp-auto, external, tag Rack3-4byteASN

fc00::0001:a1b1:c1d1:e111/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:a7f0, Eth1/1, [20/0], 04:20:13, bgp-auto, external, tag Rack1-4byteASN
fc00::0001:a2b2:c2d2:e222/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:a7f0, Eth1/1, [20/0], 04:20:13, bgp-auto, external, tag Rack1-4byteASN
fc00::0002:a3b3:c3d3:e333/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:ab00, Eth1/3, [20/0], 04:18:47, bgp-auto, external, tag Rack2-4byteASN
fc00::0003:a4b4:c4d4:e444/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:a766, Eth1/5, [20/0], 04:18:47, bgp-auto, external, tag Rack3-4byteASN
fc00::0003:a5b5:c5d5:e555/128, ubest/mbest: 1/0
    *via fe80::720f:6aff:fe4d:a766, Eth1/5, [20/0], 04:18:47, bgp-auto, external, tag Rack3-4byteASN
```

Load Balancing to where the Application runs (Server)

Node IP Addressing Example:
IPv6 LLA – FE80::MAC
IPv6 GUA – FC00::RackID:NodeID

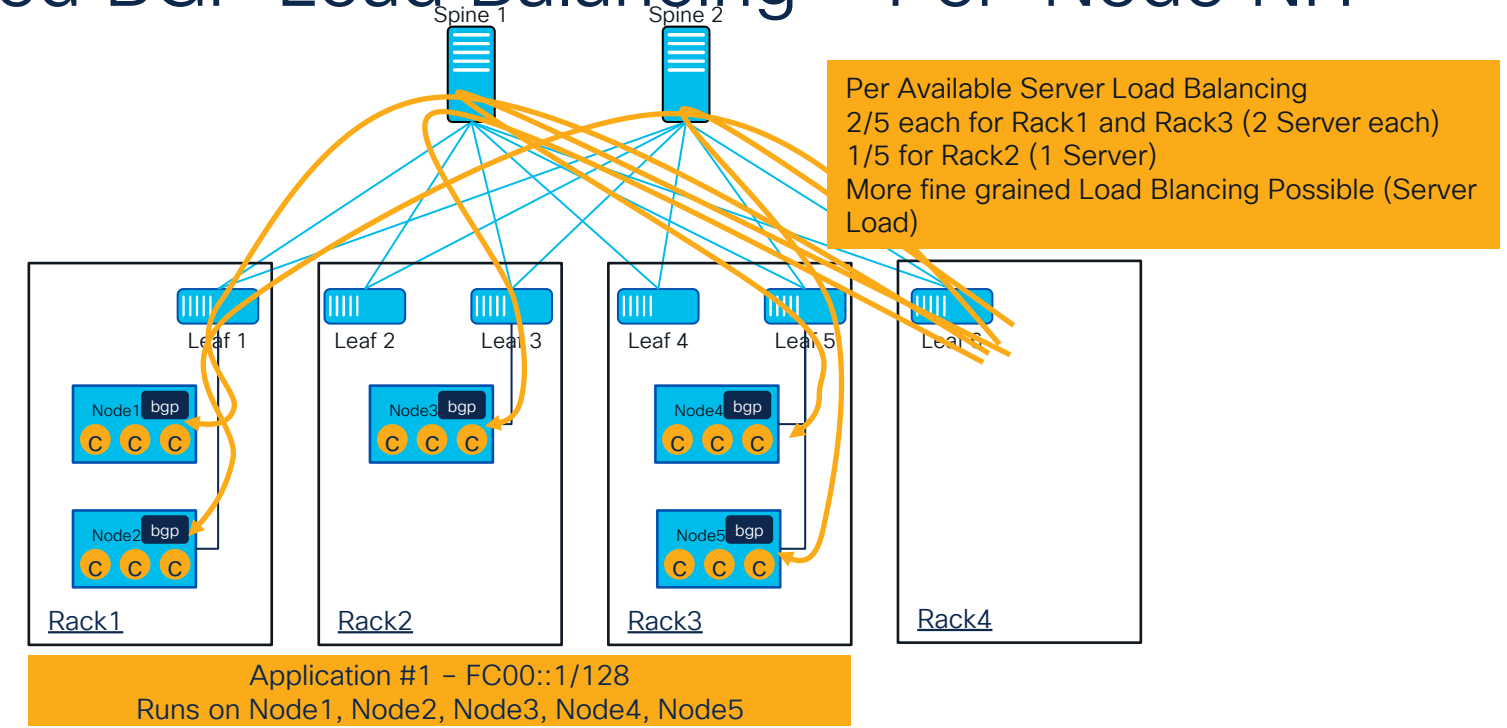


MAC Address – a1:b1:c1:d1:e1:11
IPv6 Link-Local Address – FE80::a1b1:c1d1:e111/64
IPv6 Global Unicast Address – FC00::0001:a1b1:c1d1:e111/64



MAC Address – a5:b5:c5:d5:e5:55
IPv6 Link-Local Address – FE80::a5b5:c5d5:e555/64
IPv6 Global Unicast Address – FC00::0003:a5b5:c5d5:e55

Optimized BGP Load Balancing – Per-Node NH



Node IP Addressing Example:
IPv6 LLA – FE80::MAC
IPv6 GUA – FC00::RackID:NodeID

Node1 bgp

MAC Address – a1:b1:c1:d1:e1:11
IPv6 Link-Local Address – FE80::a1b1:c1d1:e111/64
IPv6 Global Unicast Address – FC00::0001:a1b1:c1d1:e111/64

Node5 bgp

MAC Address – a5:b5:c5:d5:e5:55
IPv6 Link-Local Address – FE80::a5b5:c5d5:e555/64
IPv6 Global Unicast Address – FC00::0003:a5b5:c5d5:e55

Conclusion

‘We can scale from Small to Very Large – Don’t be shy starting with a Small Setup; we can Evolve!’

Key Takeaway #1

‘Bigger is not Always Better; Using Fixed-Form factor Switches is a Modern Practice’

Key Takeaway #2

‘More Switches != Higher Cost’

Key Takeaway #3

‘Routed Fabrics is a Real Thing ☺’

Key Takeaway #4

Resources – Cisco NX-OS

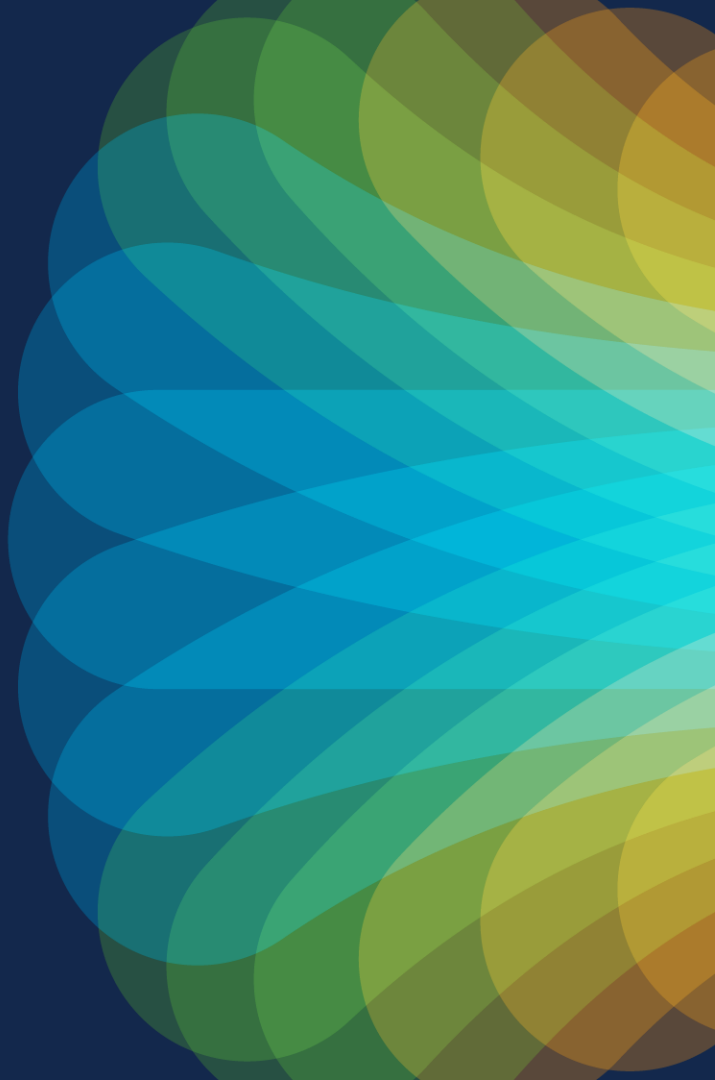
- RFC 5549
  - See Unicast Routing Configuration Guide – Advanced BGP
- BGP Auto-Fabric
  - Supported starting NX-OS 10.2(3)F
  - See Unicast Routing Configuration Guide – Advanced BGP

Resources – IETF

- RFC 5549 – Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop
  - https://datatracker.ietf.org/doc/html/rfc5549
- RFC 8950 – Advertising IPv4 Network Layer Reachability Information (NLRI) with an IPv6 Next Hop
  - https://datatracker.ietf.org/doc/html/rfc8950
- RFC 7938 – Use of BGP for Routing in Large-Scale Data Centers
  - https://datatracker.ietf.org/doc/html/rfc7938

The bridge to possible

Thank you



cisco Live!


cisco Live!
Let's go