
cisco Live!
ALL IN

The bridge to possible

Multi-Tier Fabrics

Network Designs for the Modern Data Center

Lukas Krattiger, Distinguished Engineer
@CCIE21921

BRKDCN-2099

cisco Live!

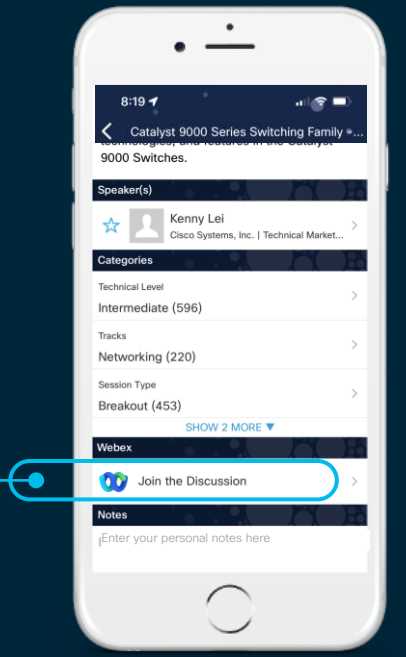
Cisco Webex App

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



https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2099

Webex spaces will be moderated by the speaker until June 17, 2022.

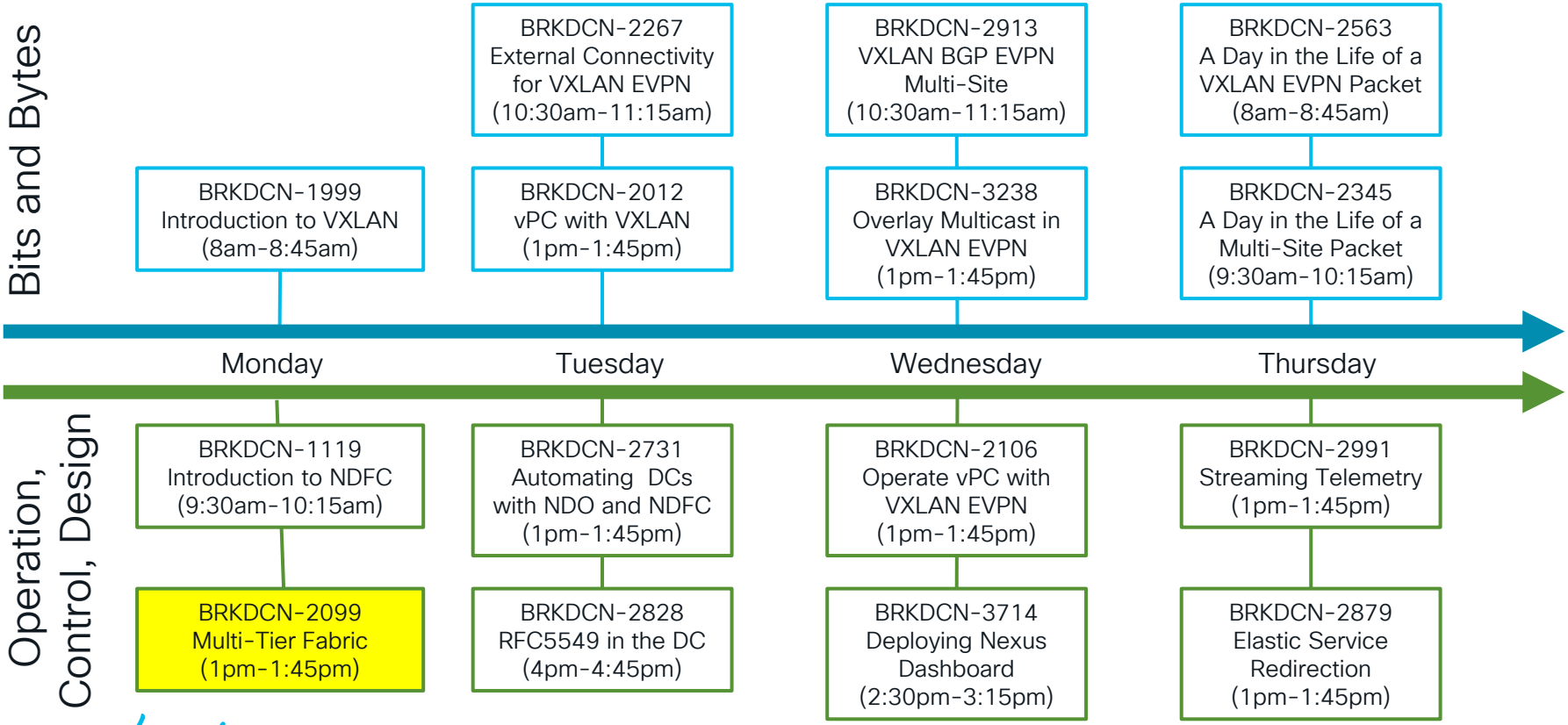
Abstract

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.

Introduction

- A brief Overview on Leaf and Spine Topologies
- Some terms and Nomenclature
- Design and Sizing considerations
- The 3-Stage and 5-Stage Clos
  - Or how we build beyond the Leaf and Spine Topology
  - Super-Spine and Spine Planes
- How does this fit into your DataCenter

Companion Sessions – Week at a Glance



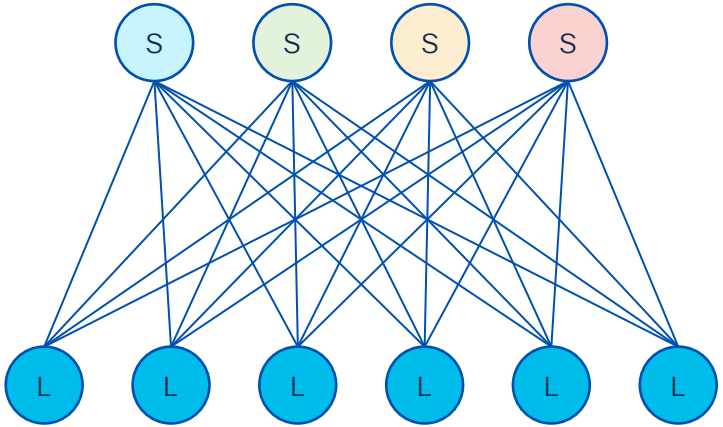
Agenda

- Introduction
- Paradigm and Fundamentals
- Design Evolution
- Conclusion

Paradigm and Fundamentals

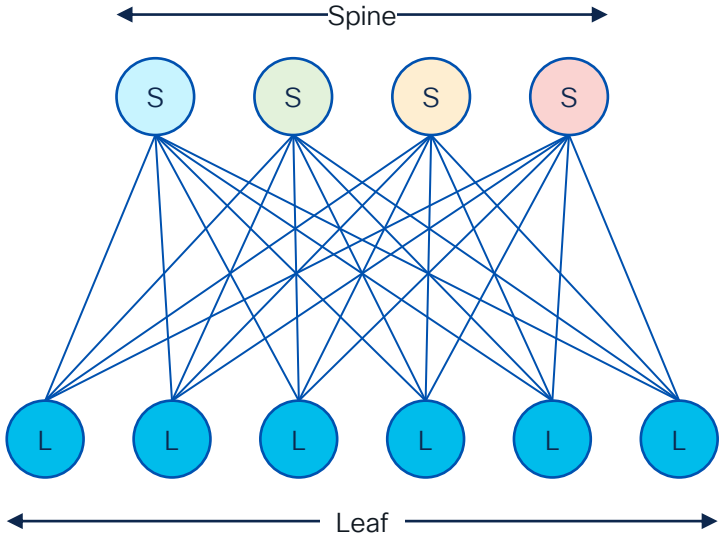
The Paradigm

- A Leaf and Spine Topology



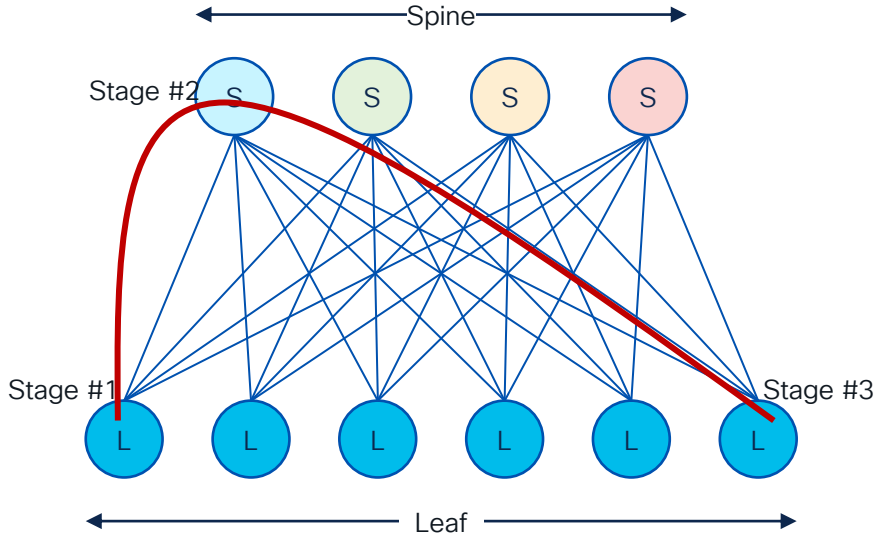
The Paradigm

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



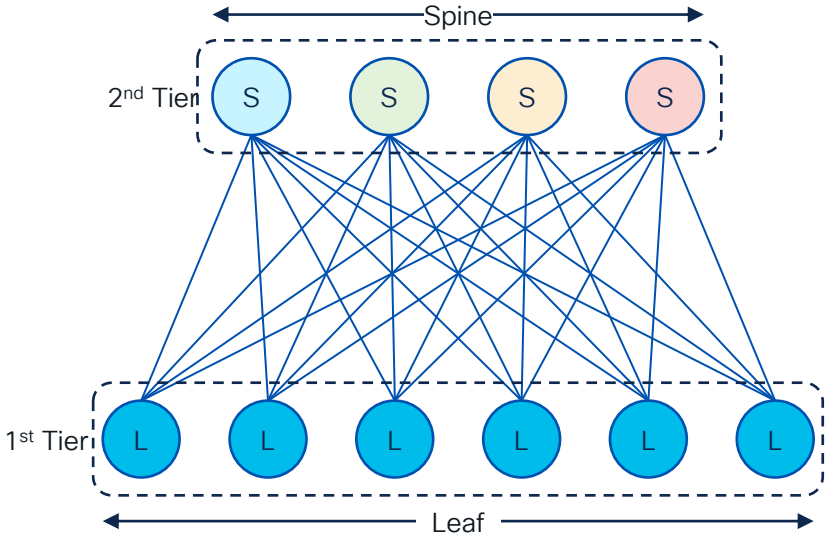
The Paradigm

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



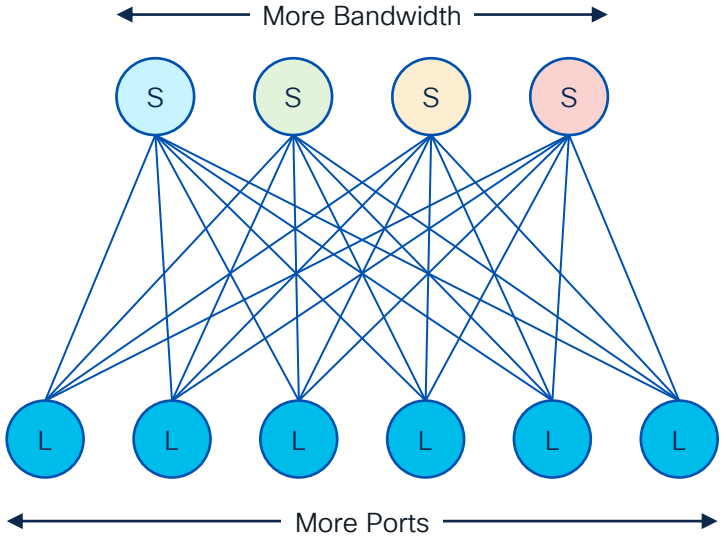
The Paradigm

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



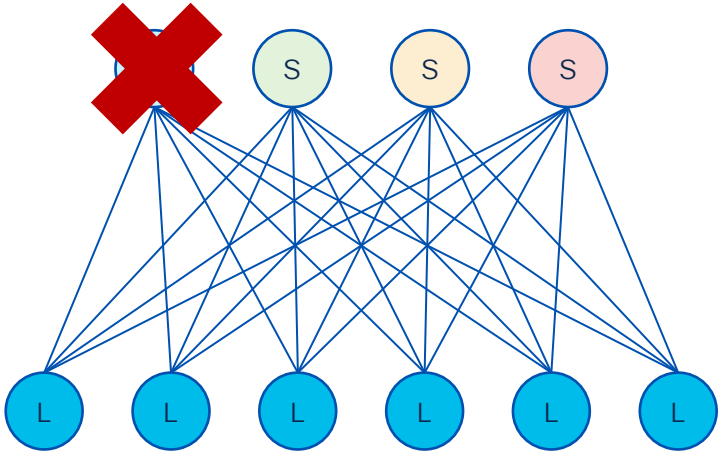
The Paradigm

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



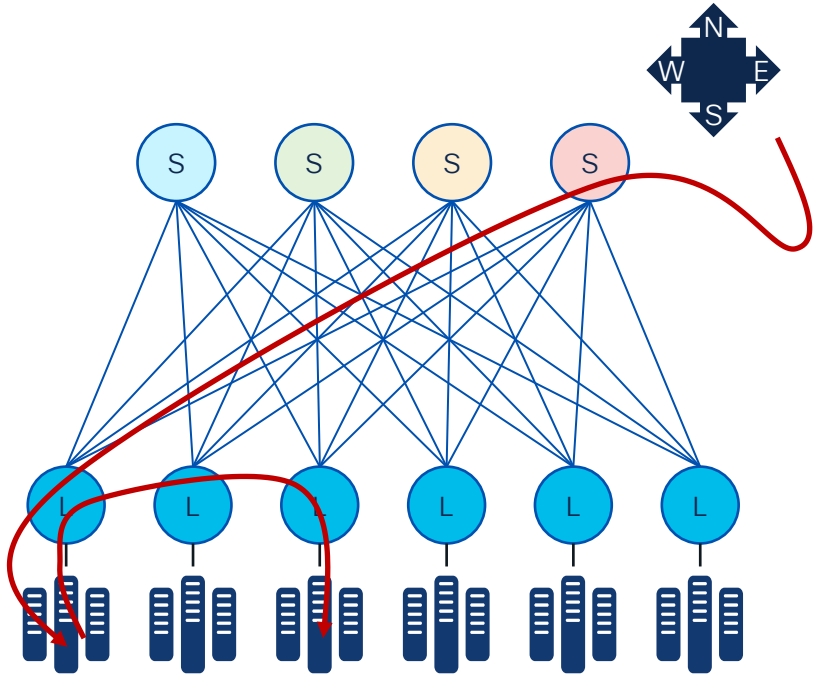
The Paradigm

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



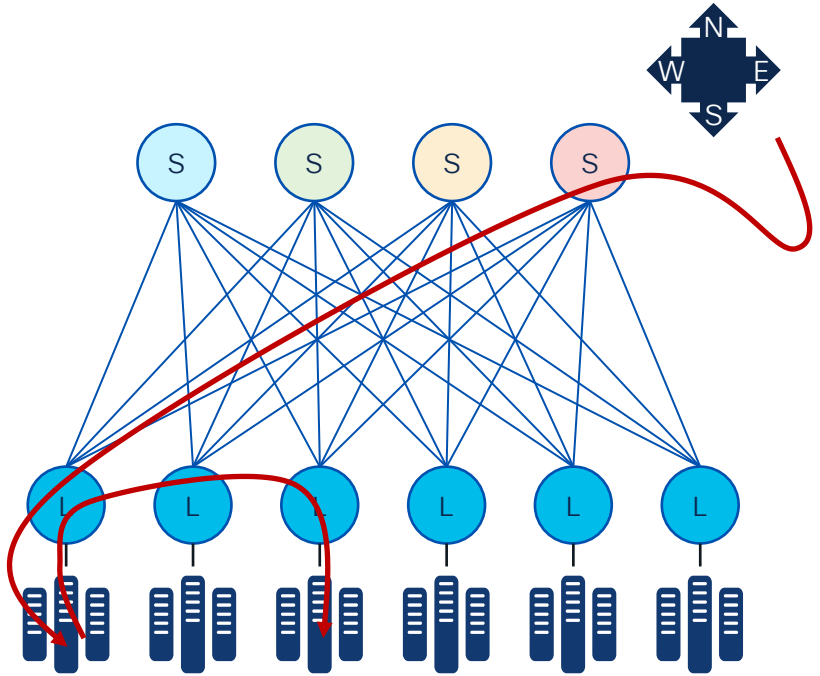
The Paradigm

- 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.



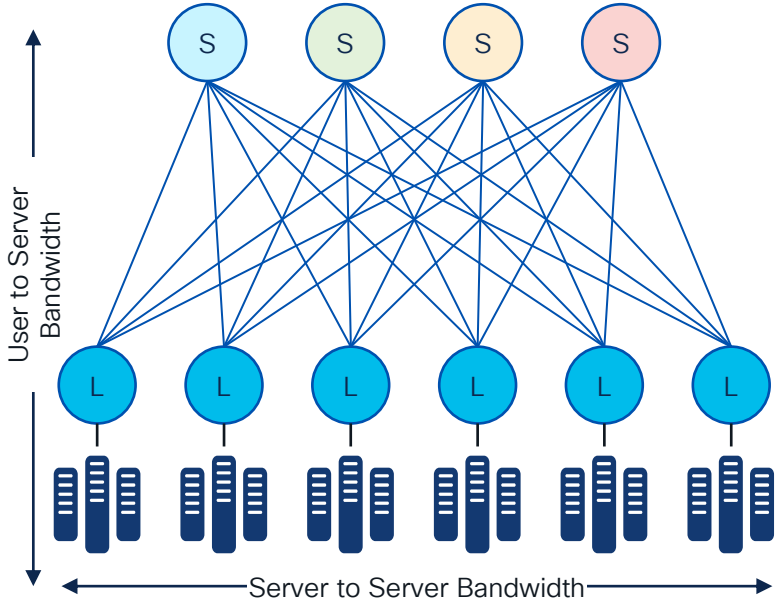
The Paradigm

- 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

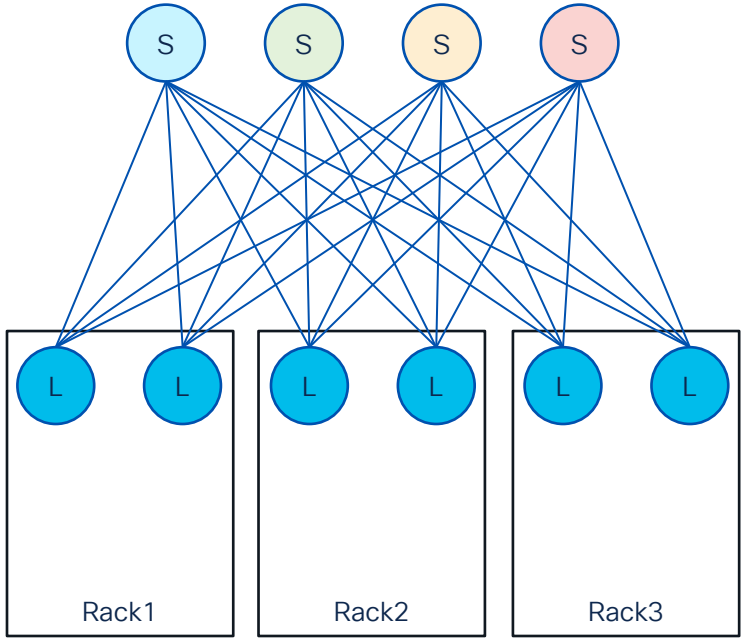


How Many Spines do I need?

It Depends

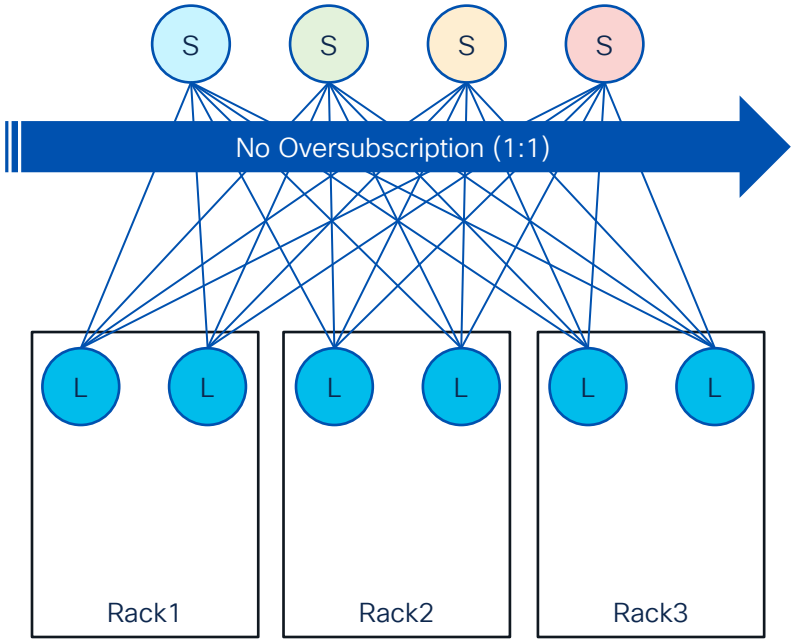
How Many Spines do I need?
Oversubscription and Maximum Redundancy as the Criteria

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

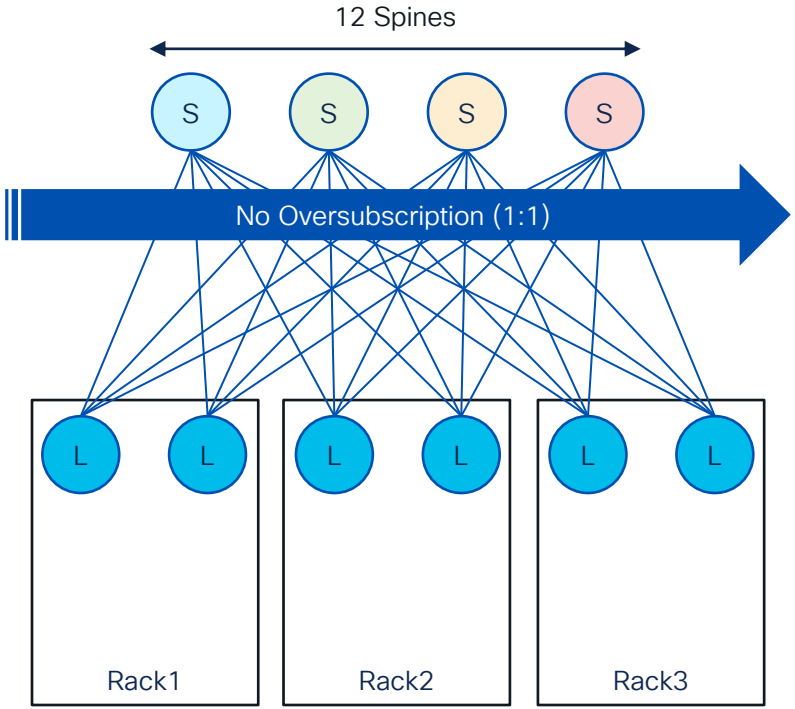
Host Attachment Requirements
- 48 Server per Rack
- 2x 25Gbps NIC per Server
- 1x NIC per Switch



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

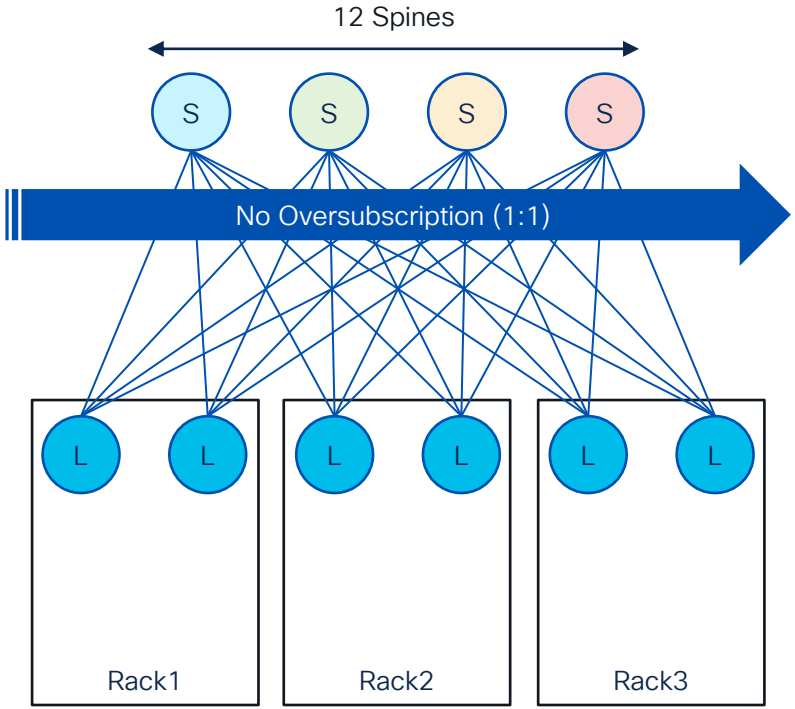
Host Attachment Requirements
- 48 Server per Rack
- 2x 25Gbps NIC per Server
- 1x NIC per Switch



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

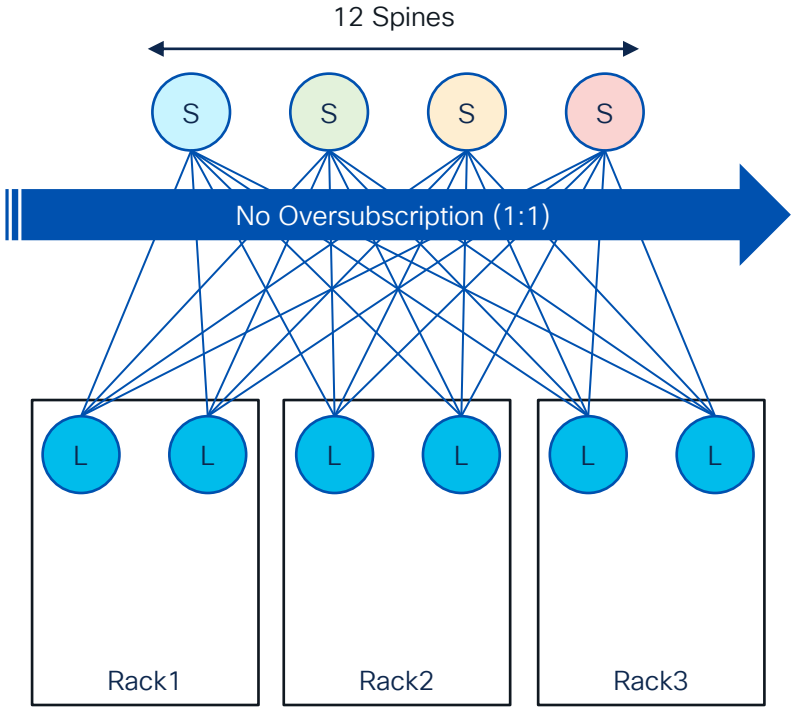
Fabric Size – 12 Spine, 1:1 Oversubscription

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

Modular vs. Fixed for Single Fabric (2 Tier)
Fabric Size – 12 Spine, 1:1 Oversubscription



Let's Do some Math	
Spine	
8 Slot Modular Chassis	Fixed Spine
36x 100Gbps Port per Linecard	64x 400Gbs
Total: 288 Spine Ports	Total: 256 Spine Ports
Leaf	
288 Spine Ports = 288 Leaf Switch	256 Spine Ports = 256 Leaf Switch
48x 25Gbps Host Ports Per Leaf	48x 25Gbps Host Ports Per Leaf
Total: 13'828 Host Ports	Total: 12'288 Host Ports
Fabric Bandwidth	
1:1 Oversubscription	1:1 Oversubscription
1.2Tbps Uplink * 288 Leaf	1.2Tbps Uplink * 256 Leaf
Total: 345.6Tbps	Total: 307.2Tbps

PRO
More Leaf
More Ports
More Bandwidth

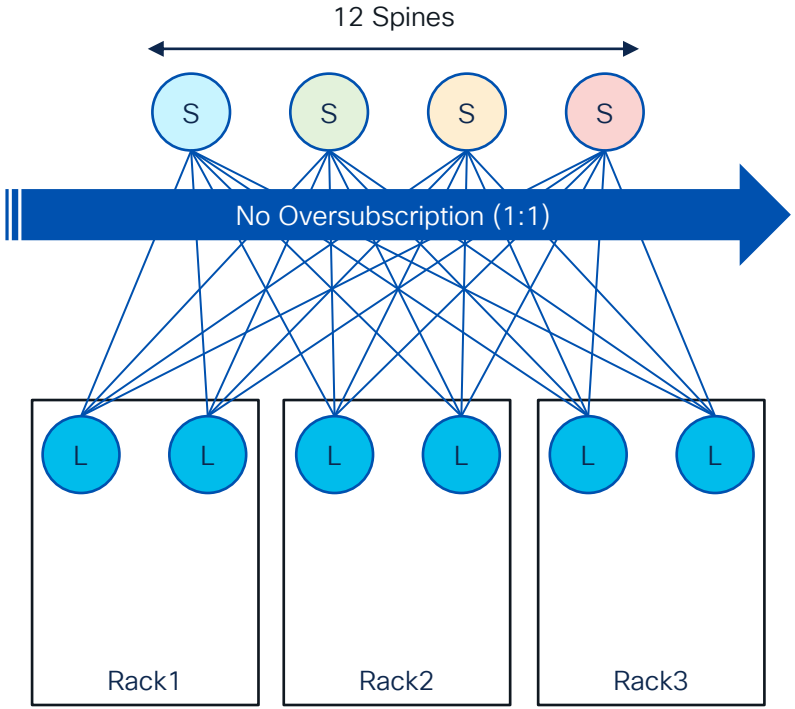
PRO
Less Latency
Less Power
All Single ASIC

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



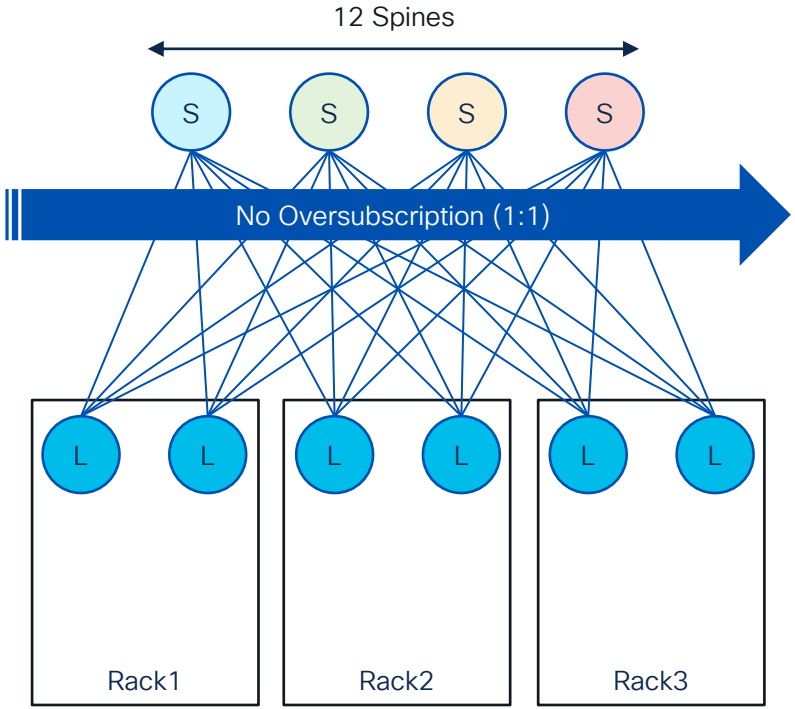
Let's Do some Math	
Spine	
8 Slot Modular Chassis	16 Slot Modular
36x 100Gbps Port per Linecard	36x 100Gbps Port per Linecard
Total: 288 Spine Ports	Total: 576 Spine Ports
Leaf	
288 Spine Ports = 288 Leaf Switch	576 Spine Ports = 576 Leaf Switch
48x 25Gbps Host Ports Per Leaf	48x 25Gbps Host Ports Per 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



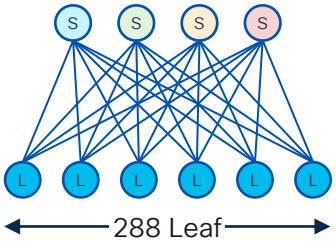
Let's Do some Math	
Spine	
8 Slot Modular Chassis	8 Slot Modular Chassis
36x 100Gbps Port per Linecard	36x 400Gbps Port per Linecard
Total: 288 Spine Ports	Total: 1152 Spine Ports
Leaf	
288 Spine Ports = 288 Leaf Switch	1152 Spine Ports = 1152 Leaf Switch
48x 25Gbps Host Ports Per Leaf	48x 25Gbps Host 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

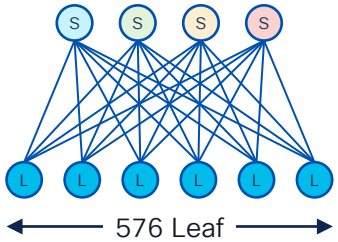
More Spine Ports Results in More … Fabric Bandwidth, Leaf Count, Host Ports

Attributes to Scale



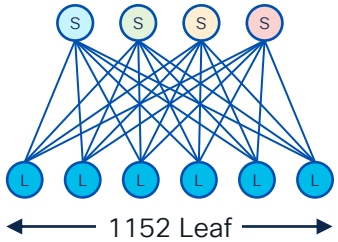
8 Slot Modular
36x 100Gbps
1:1 Oversubscription
13'828 Host Ports

Scale-Up to Fill Chassis



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

Scale-Up to Bigger Chassis



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

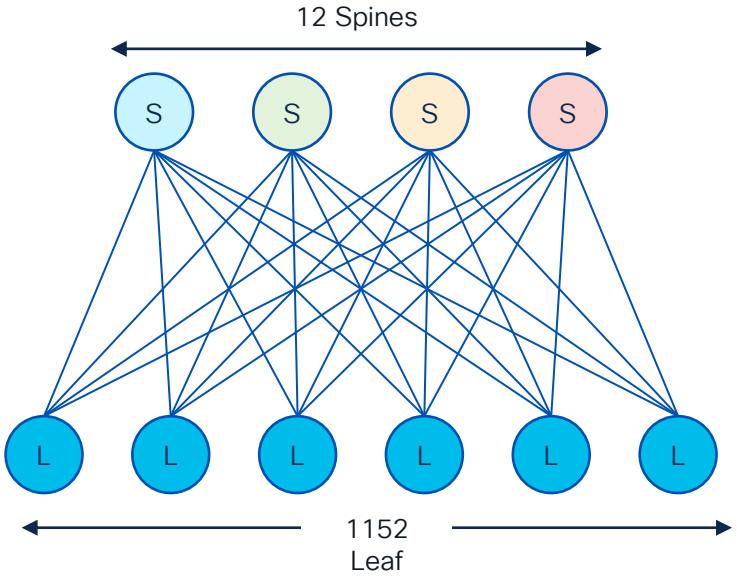
Scale-Up to Faster Linecards

Oversubscription Ratio doesn't influence Host Port scale

Points To Think About

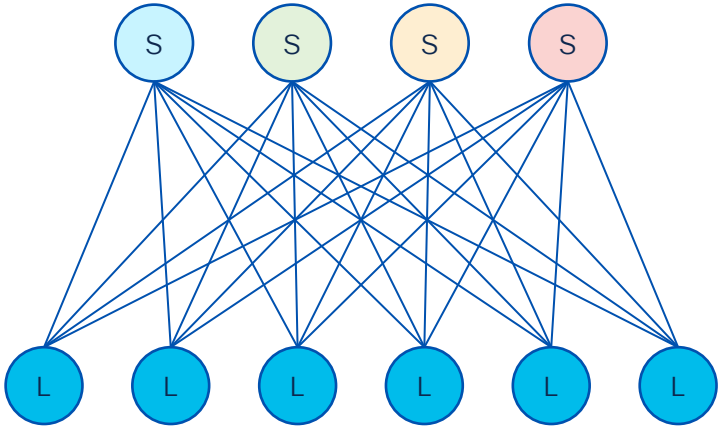
- How to Scale 2 Tier Networks
- Every Spine scaling involves Scale-Up
  - Initial Chassis Build Out (4)
    - Add Linecards (up to 4)
  - Increase Chassis Size (4 > 8)
    - Add more Linecards (up to 8)
  - Increase Chassis Size (8 > 16)
    - Add more Linecards (up to 16)
  - Increase Chassis Speed (4)
    - Add Faster Linecards (up to 4)
  - Increase Chassis Size (4 > 8)
    - Add more Linecards (up to 8)
  - Increase Chassis Size (8 > 16)
    - Add more Linecards (up to 16)

What Else to Think About?

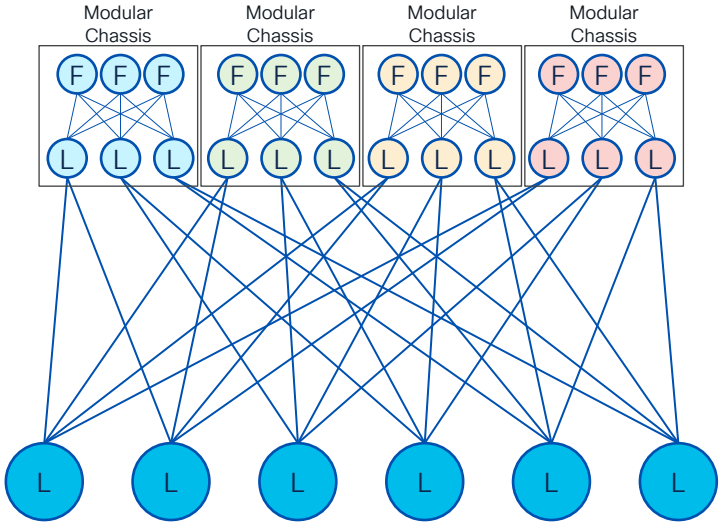


- What is my Failure Domain?
- What is my Change Domain?
- What is my Overall Scale?
- What is my Fabric Solution Scale?
- What is my Fabric SLA?
- What is my Maximum Downtime?

What did I really Build – Untangling the Details
2 Tier / 5 Stage Network with Modular Spine

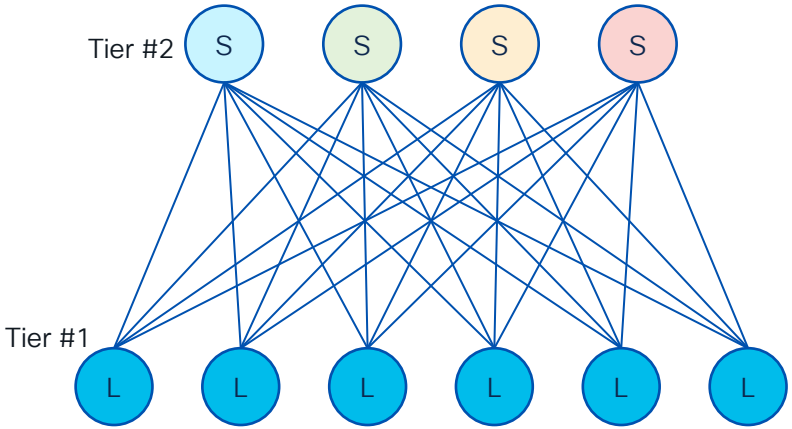


What you think you Built
2 Tier Leaf and Spine Network (3 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)

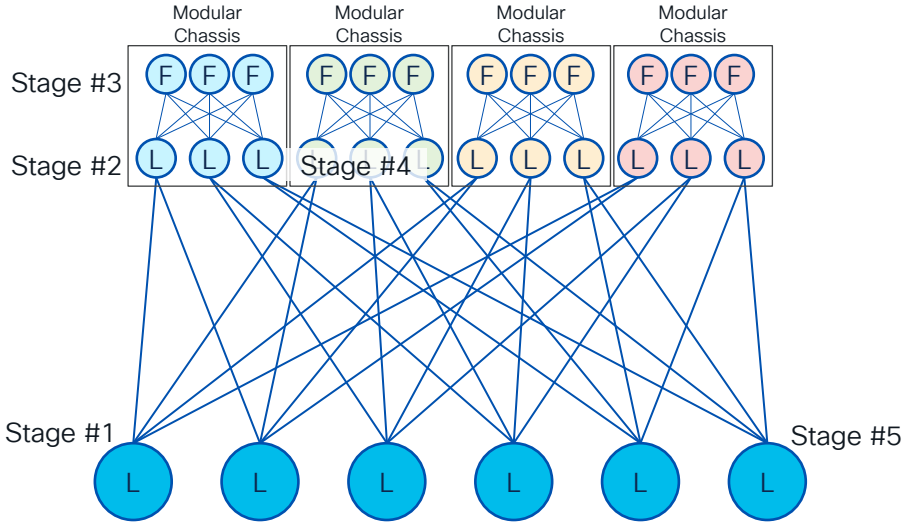


What you really Built
2 Tier Leaf and Spine Network (5 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)

What did I really Build – Untangling the Details
2 Tier / 5 Stage Network with Modular Spine

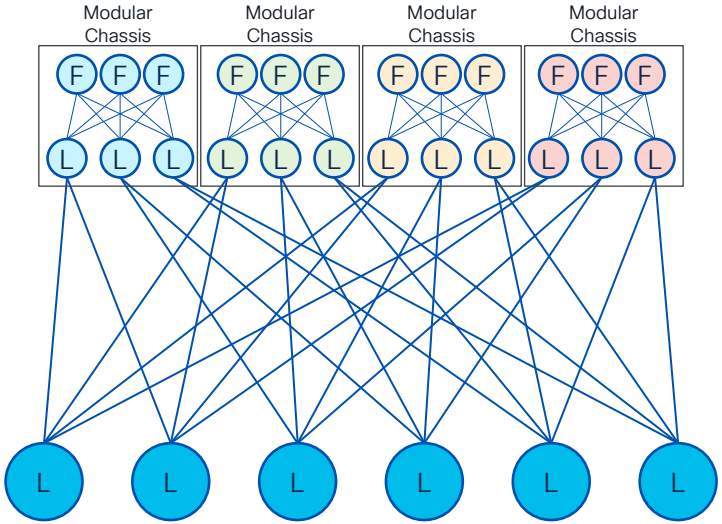


What you think you Built
2 Tier Leaf and Spine Network (3 Stage)
Spine: Modular Chassis (4 Slot, 8, Slot, 16 Slot)
Leaf: Fixed Switch (single ASIC)



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)

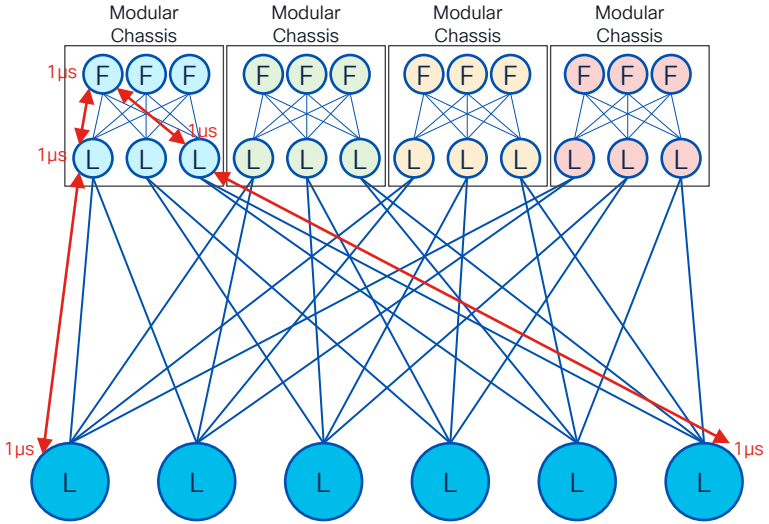
Forwarding Behavior
2 Tier / 5 Stage Network with Modular Spine



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

- Fixed Leaf represents 1 Stage
- Modular Spine represents 3 Stages

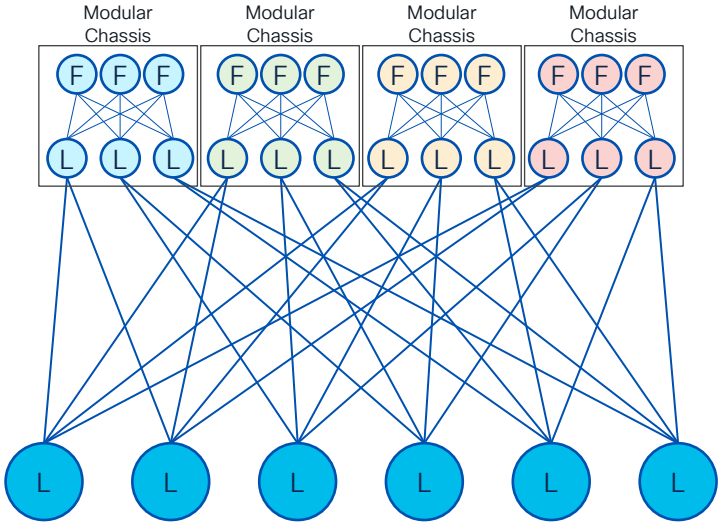
Latency Behavior
2 Tier / 5 Stage Network with Modular Spine



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

Intra-Chassis Behavior
2 Tier / 5 Stage Network with Modular Spine



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)

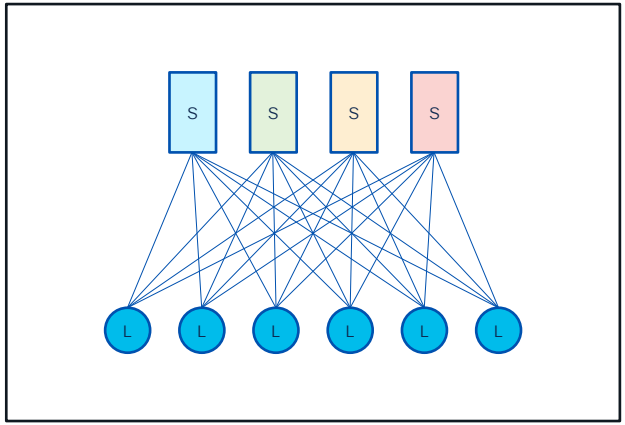
- Within Leaf Tier and Between Leaf and Spine Tier
  - Full Behavior / Protocol Control
  - Layer-3 ECMP Load Balancing
  - Standards-based Routing Protocols
  - BFD for Fast Failure Detection
  - Minimal Exposure for Brownout
- Within Spine Tier
  - Intra-Chassis Load Balancing
  - Intra-Chassis Protocol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

With all this Knowledge, why do we still build 2 Tier Scale Up Networks with Big-Fat Spines?

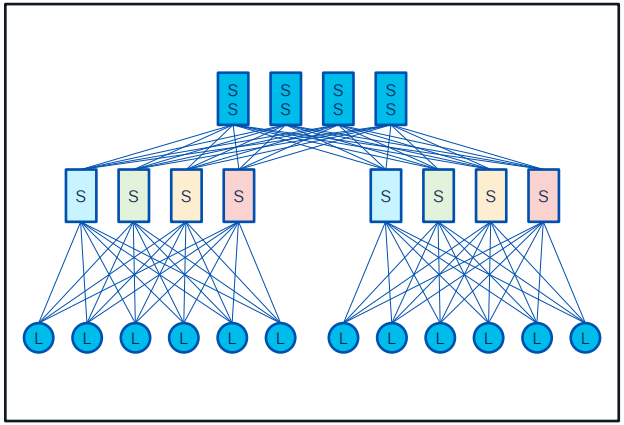
The Elephant in the Room

Design Evolution

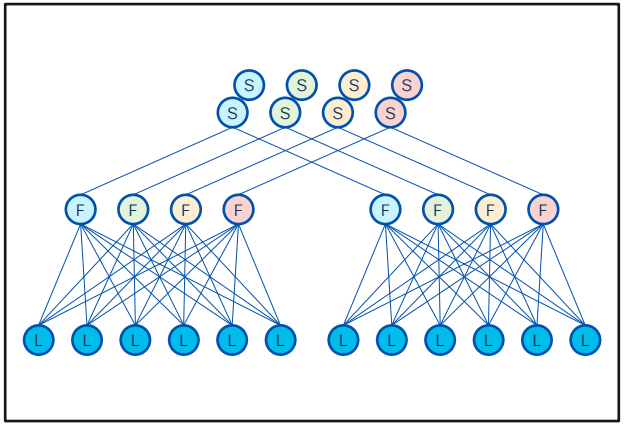
The Journey to Build Better and Further



2 Tier Leaf Spine
(5 Stages)



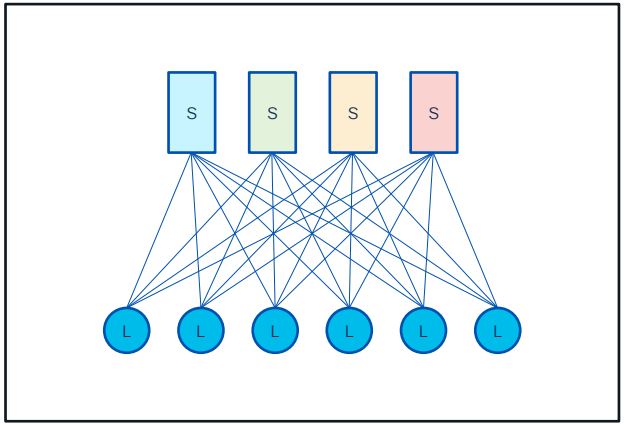
3 Tier Leaf-Spine-SuperSpine
(11 Stages)



3 Tier Leaf-Fabric-Spine
(5 Stages)

The Status Quo
Discussed at Length

- A perfectly valid way
- Tends to have “Finite Scale”
  - Maximum Chassis capacity
  - Maximum Speed per Port
- Many Locations of Redundancy
  - Redundant Chassis Components
- Condensed Link and Bandwidth Presence
  - Aggregated within a Chassis



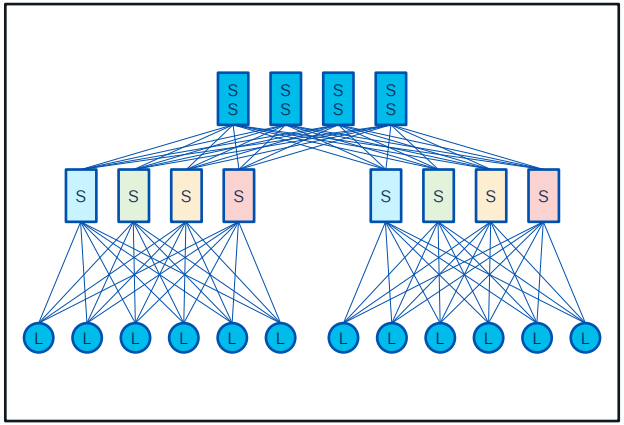
2 Tier Leaf Spine
(5 Stages)

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

The Journey to Build Better and Further

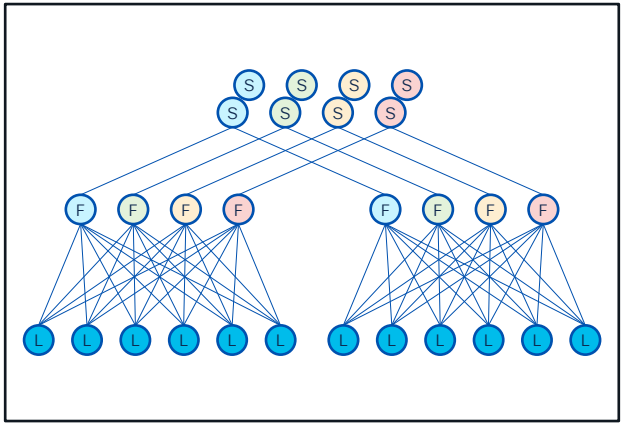
2 Tier Leaf Spine
(5 Stages)

Let's Move Forward



3 Tier Leaf-Spine-SuperSpine
(11 Stages)

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

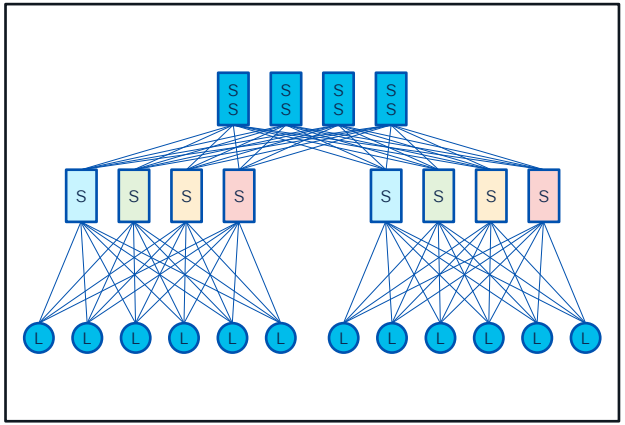


3 Tier Leaf-Fabric-Spine
(5 Stages)

- To Infinity and the Beyond

Nothing New
Let's Not Stop Here

- Avoiding Scale-Up with another Tier
- Distributed Link and Bandwidth Presence
  - Disaggregated across Tiers
- Increases the "Finite Scale"
  - No Dependency on Chassis capacity or Speed per Port
- Many Locations of Redundancy
  - Redundant Chassis Components
- Allows for Cost Optimization



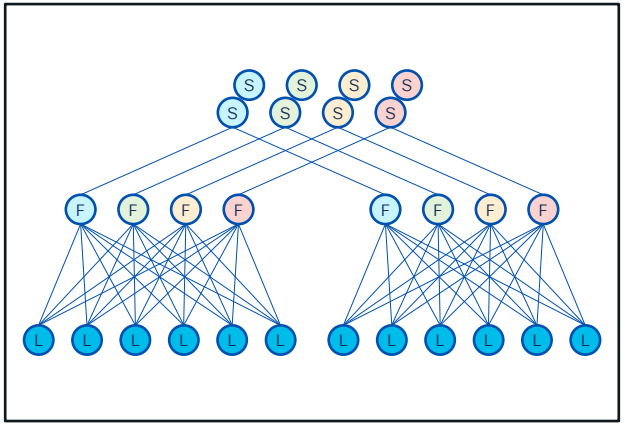
3 Tier Leaf-Spine-SuperSpine
(11 Stages)

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

The Journey to Build Better and Further

2 Tier Leaf Spine
(5 Stages)

3 Tier Leaf-Spine-SuperSpine
(11 Stages)



3 Tier Leaf-Fabric-Spine
(5 Stages)

- To Infinity and the Beyond

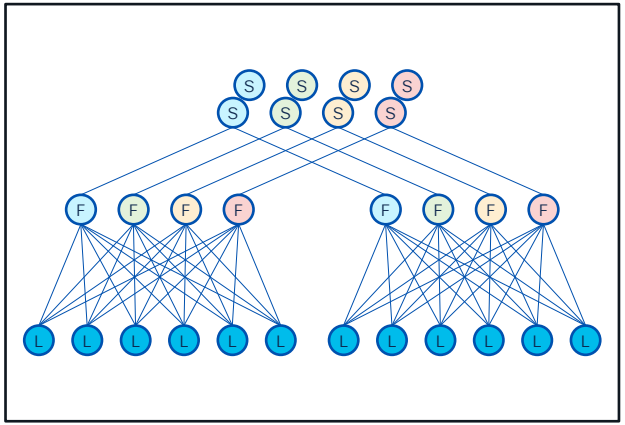
Let's Move Forward

What we learned from the Cloud Titans
Building Scalable DataCenter Networks

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

How the Cloud Titans Build

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

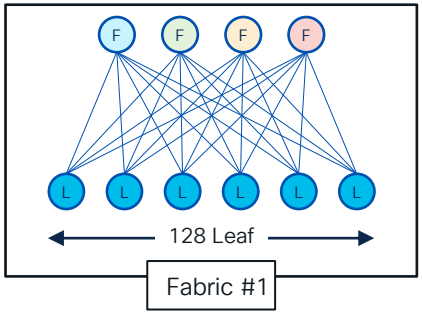


3 Tier Leaf-Fabric-Spine
(5 Stages)

- To Infinity and the Beyond

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

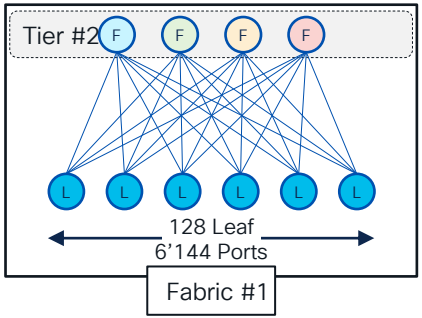
- Fixed Switch at the Fabric (Tier #2)
  - Depending on Oversubscription Ratio, reserve Ports
    - 1:1 Oversubscription
    - Reserve 50% from Tier #2 to Tier #3
- Common Fixed Spine Options
  - 64x 100Gbps (6.4Tbps Single ASIC)
  - 32x 400Gbps (12.8Tbps Single ASIC)
  - 64x 400Gbps (25.6Tbps Single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



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

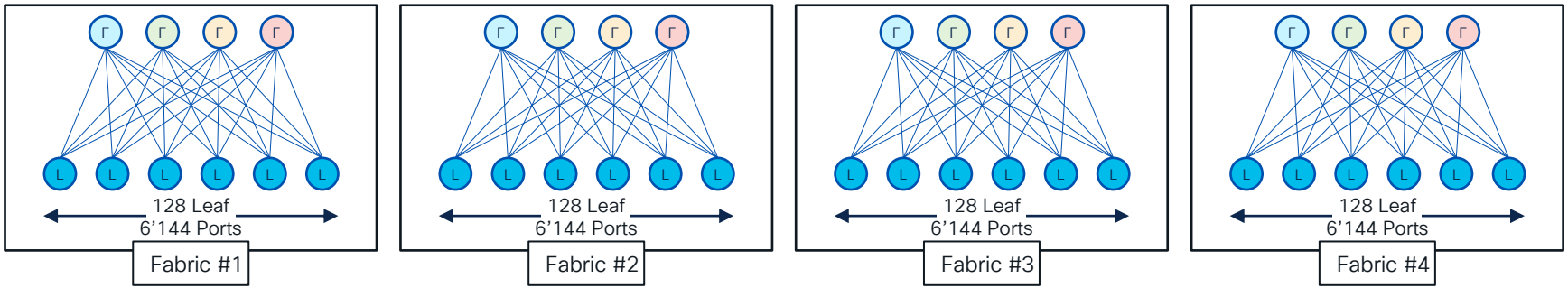
- Fixed Switch at the Fabric (Tier #2)
  - Depending on Oversubscription Ratio, reserve Ports
    - 1:1 Oversubscription
    - Reserve 50% from Tier #2 to Tier #3
- Common Fixed Spine Options
  - 64x 100Gbps (6.4Tbps Single ASIC)
  - 32x 400Gbps (12.8Tbps Single ASIC)
  - 64x 400Gbps (25.6Tbps Single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

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



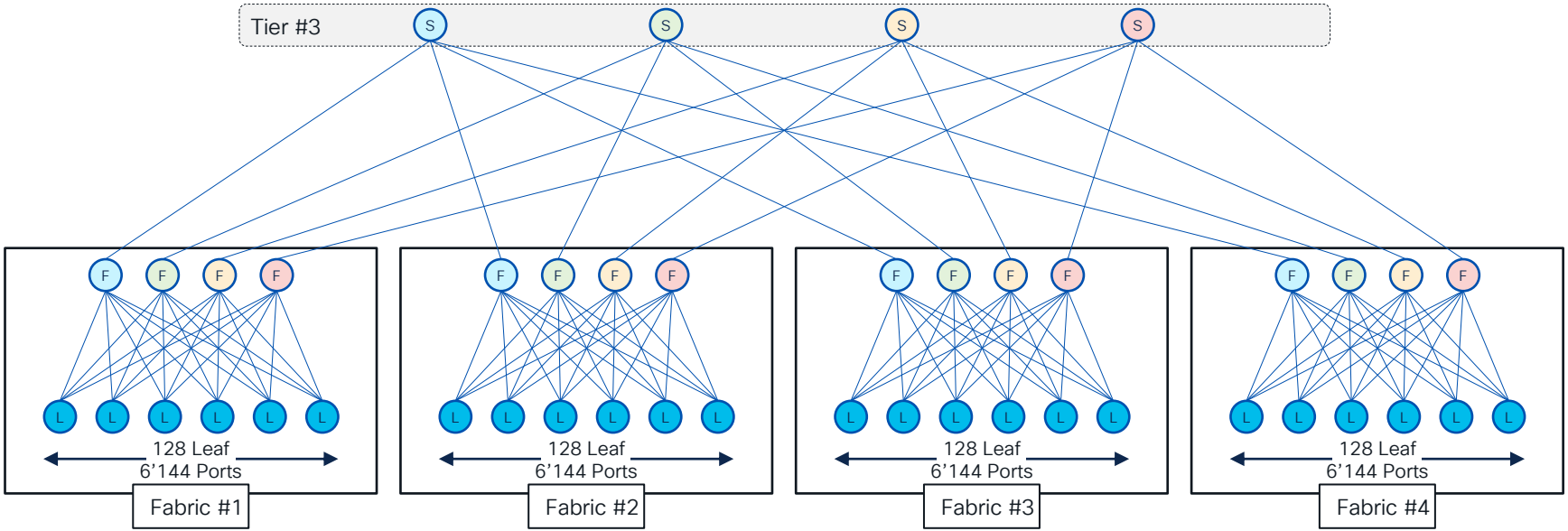
Step #2 – Repeat for Host Port Scale (Scale Out)

- 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



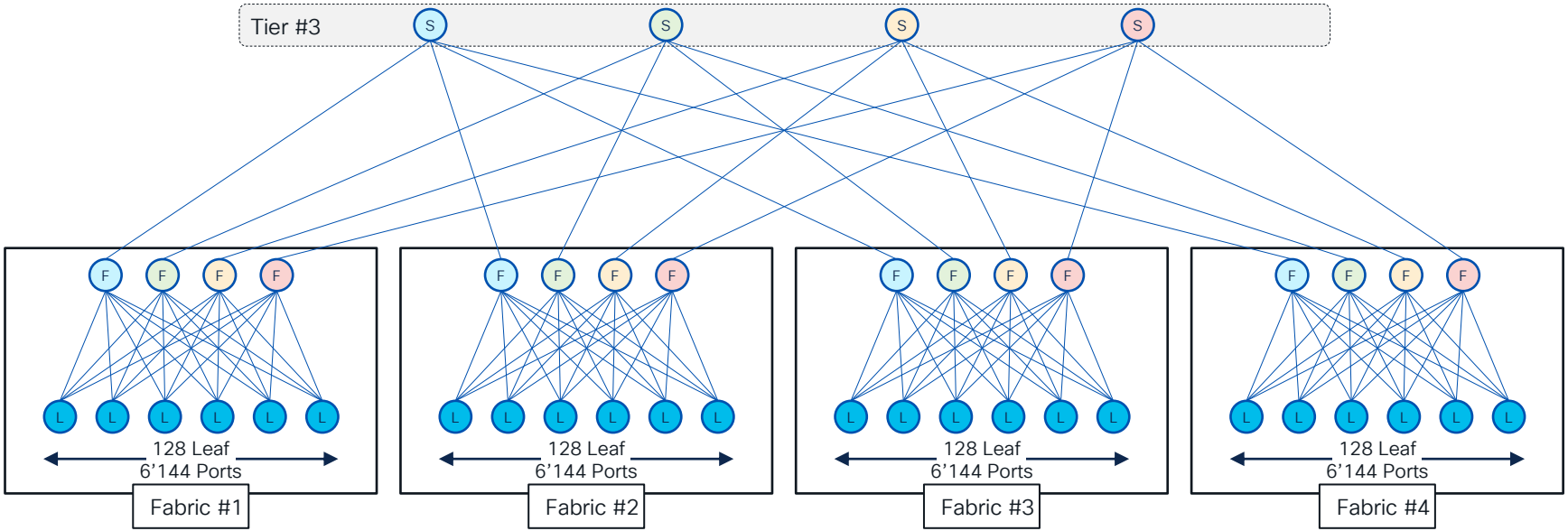
Step #3 – Designing Tier #3

- Introducing Tier #3 Planes
  - Blue, Green, Yellow, Red
- Rule: Tier #2 Blue only connects to Tier#3 Blue

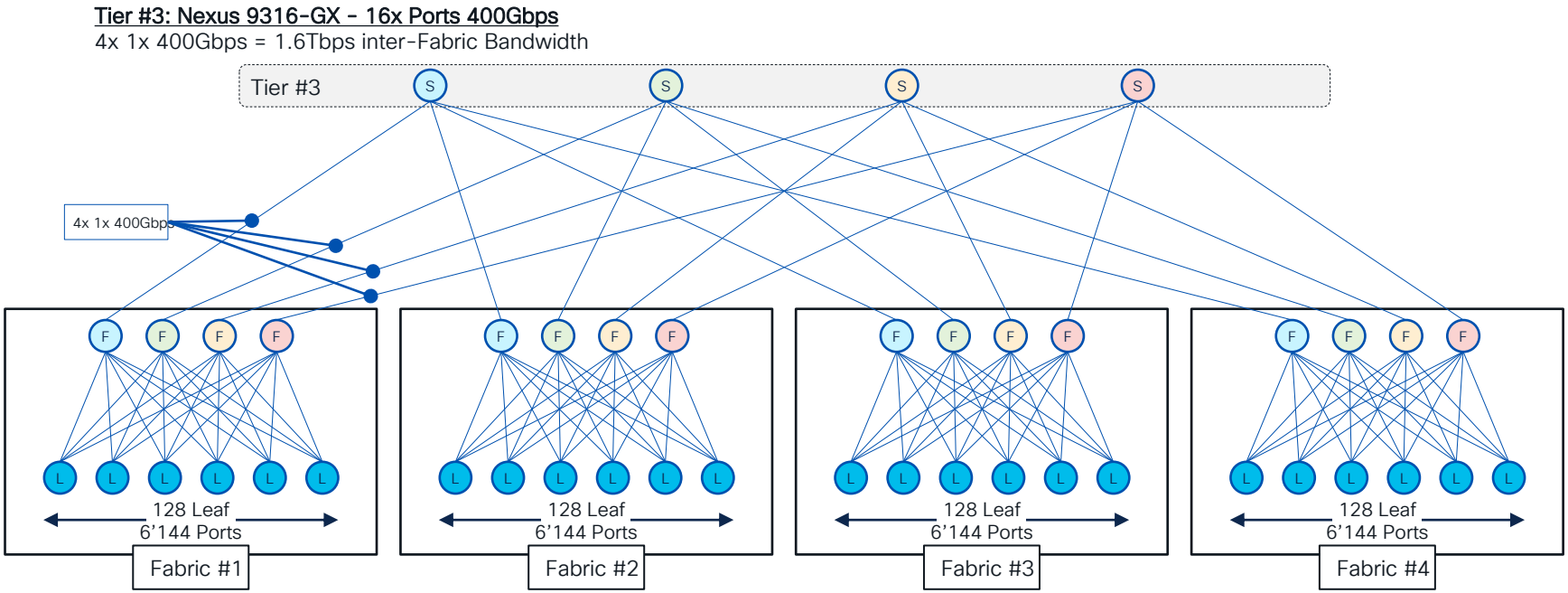


Step #3 – Designing Tier #3

- Inter-Fabric Decision is made at Leaf Layer
  - Deterministic Path from Tier #2 to Tier #3
- Rule: Once entered a Plane, you stay in the Plane



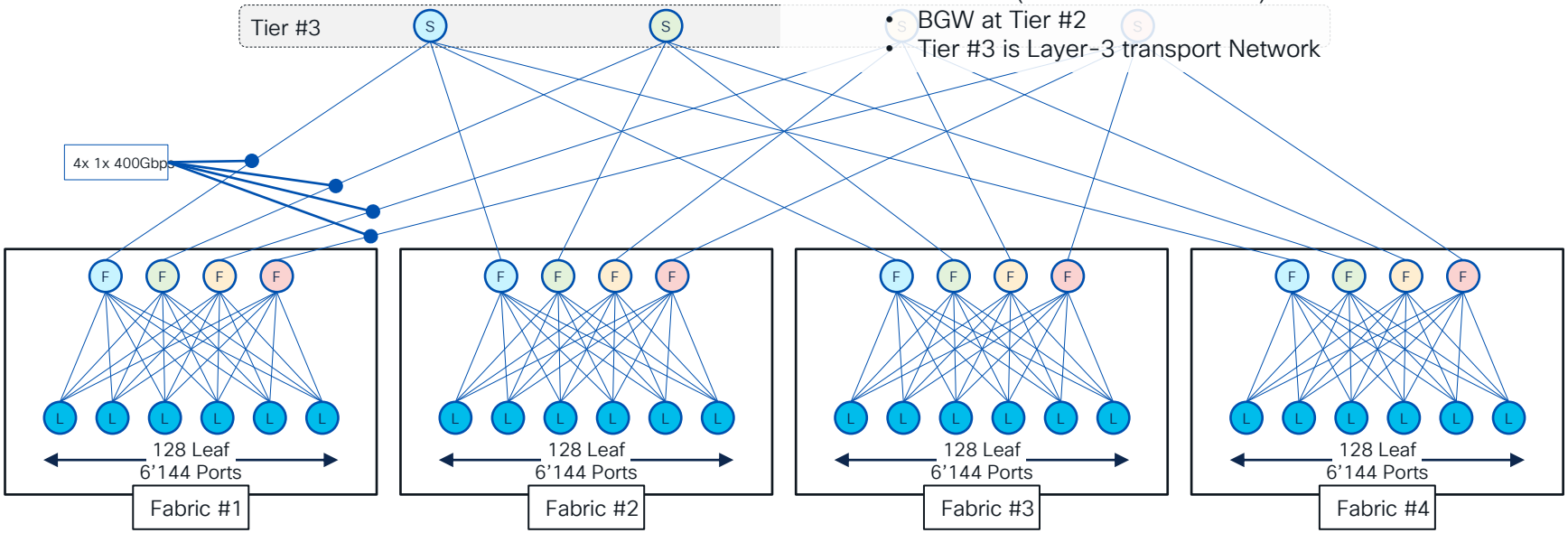
Step #3 – Designing Tier #3 (Single Link)



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

- This Setup gets you all the way to 16 Fabrics
  - 16x 6'144 Host Ports = 98'304 Host Ports
- Valid with VXLAN EVPN
  - 16 Sites (max Site scale is 128)
  - BGW at Tier #2
  - Tier #3 is Layer-3 transport Network

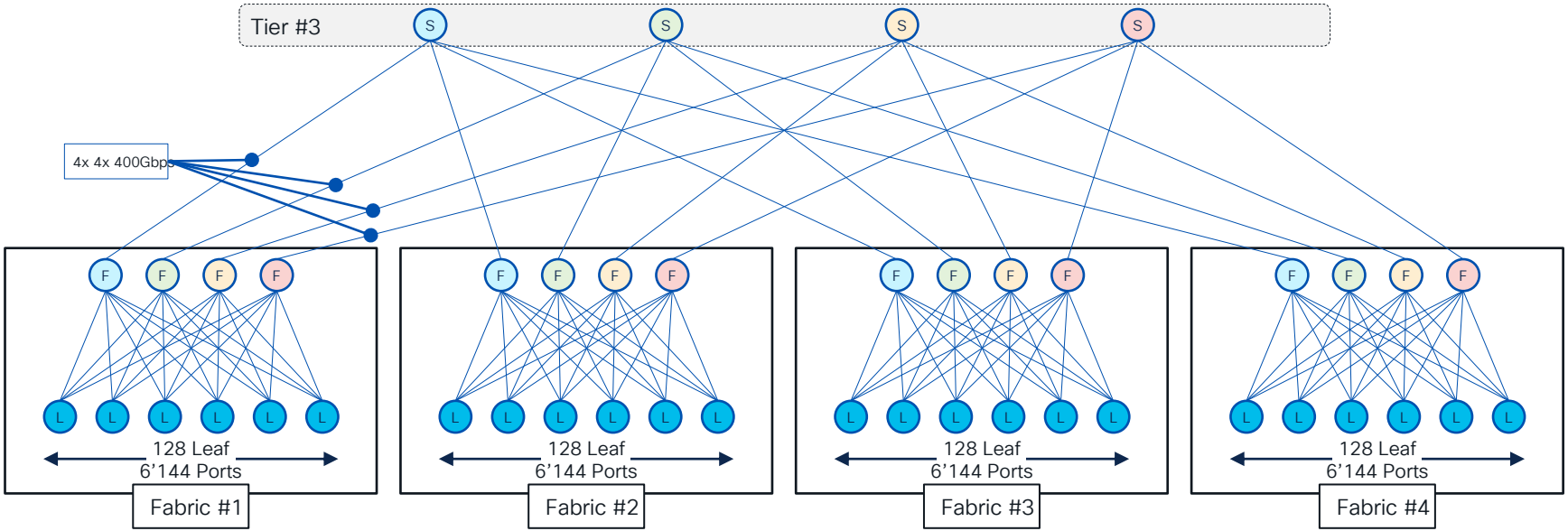
Tier #3: Nexus 9316-GX – 16x Ports 400Gbps
4x 1x 400Gbps = 1.6Tbps inter-Fabric Bandwidth



Step #3 – Designing Tier #3 (Multi-Link)

- Or increase Bandwidth to 6.4Tbps
  - Multiple Links (4) between Tier #2 and Tier #3

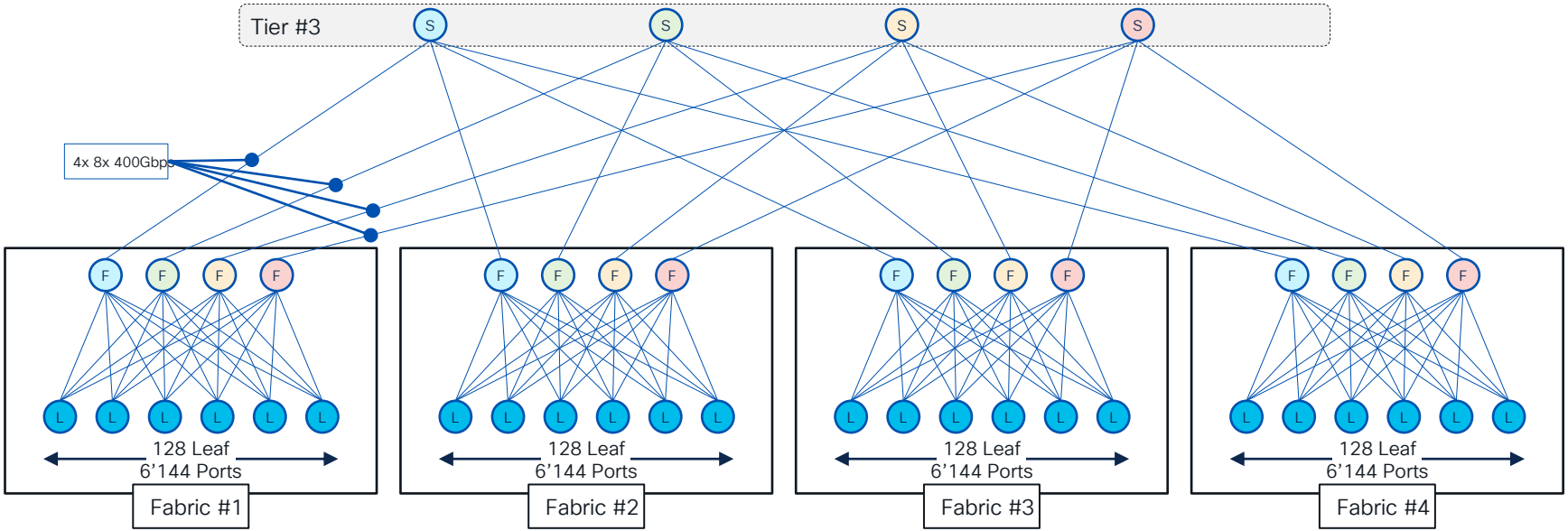
Tier #3: Nexus 9316-GX – 16x Ports 400Gbps
4x 4x 400Gbps = 6.4Tbps inter-Fabric Bandwidth



Step #3 – Designing Tier #3 (Multi-Link)

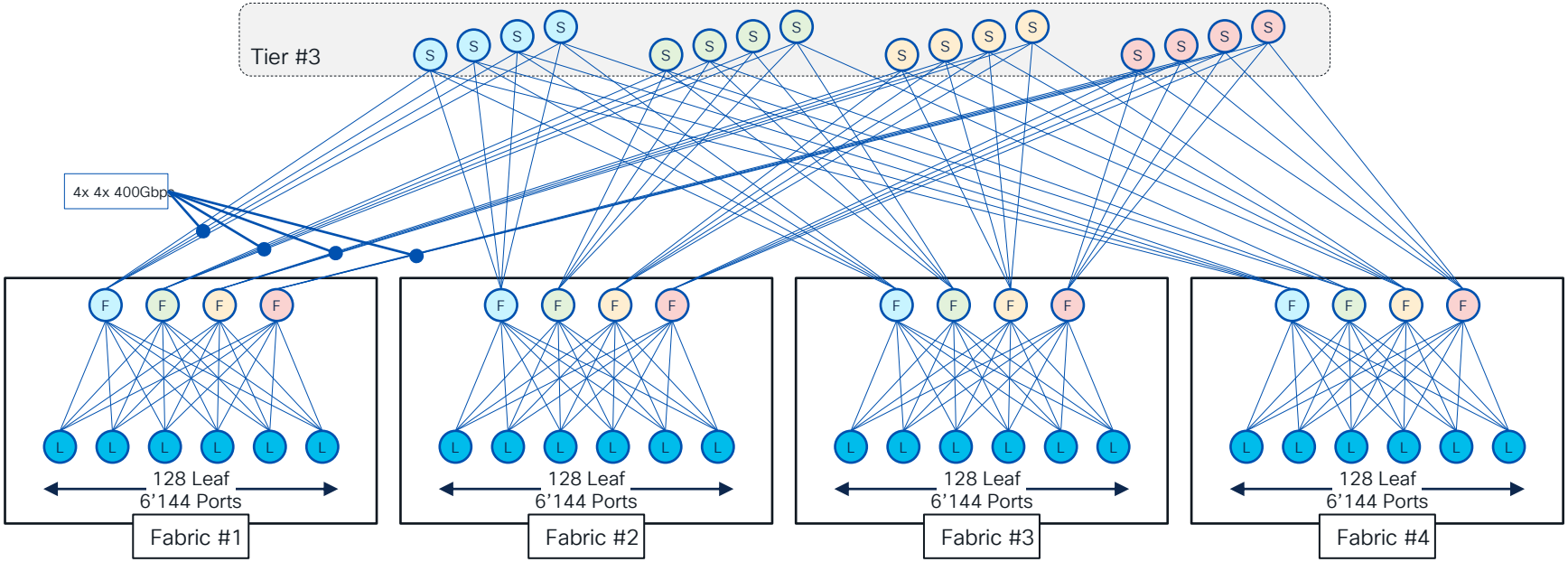
- Or increase Bandwidth to 12.8Tbps
  - Multiple Links (8) between Tier #2 and Tier #3

Tier #3: Nexus 9332-GX – 32x Ports 400Gbps
4x 8x 400Gbps = 12.8Tbps inter-Fabric Bandwidth



Step #4 – Increasing the Tier #3 Planes

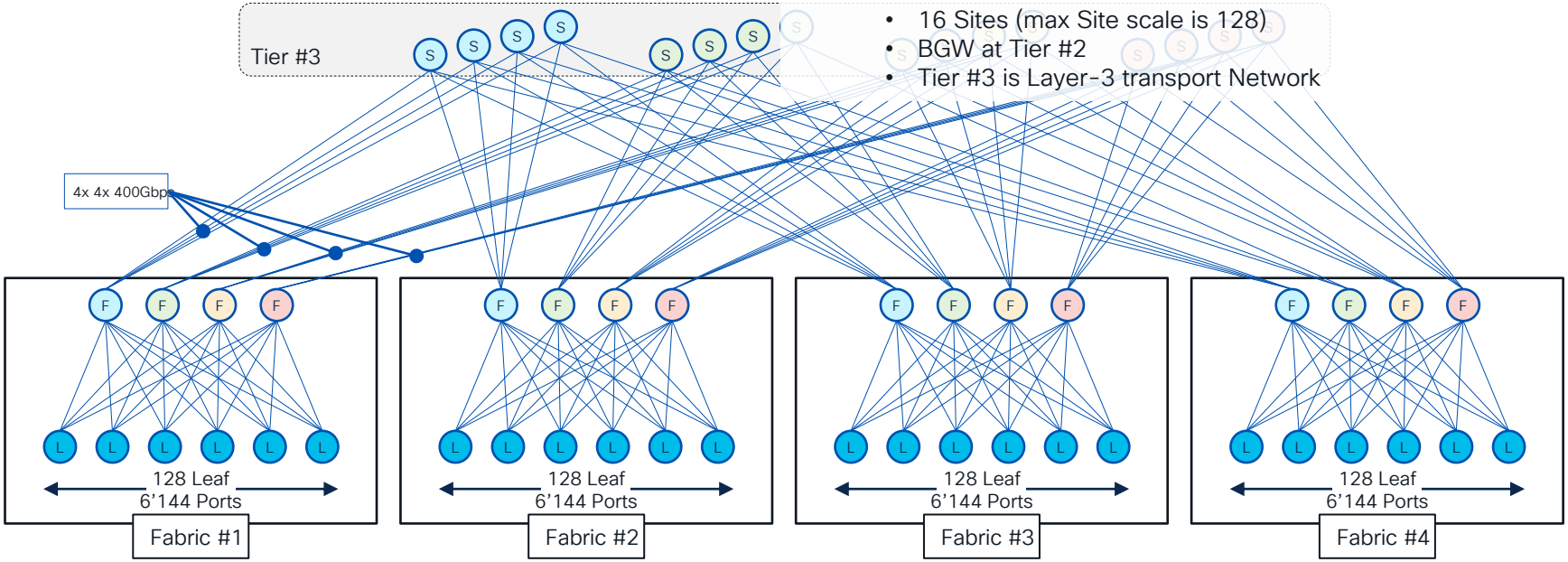
Tier #3: Nexus 9316-GX (4 per Plane) - 16x Ports 400Gbps
4x 4x 400Gbps = 6.4Tbps inter-Fabric Bandwidth



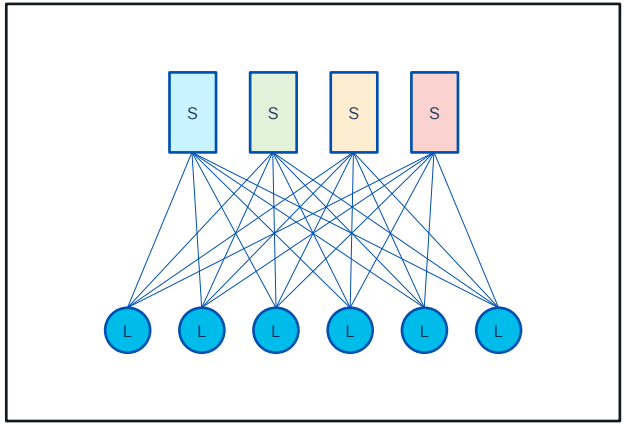
Step #4 – Increasing the Tier #3 Planes

<u>Tier #3: Nexus 9316-GX (4 per Plane) - 16x Ports 400Gbps</u>
4x 4x 400Gbps = 6.4Tbps inter-Fabric Bandwidth

- This Setup gets you all the way to 16 Fabrics
  - 16x 6'144 Host Ports = 98'304 Host Ports
- Valid with VXLAN EVPN
  - 16 Sites (max Site scale is 128)
  - BGW at Tier #2
  - Tier #3 is Layer-3 transport Network

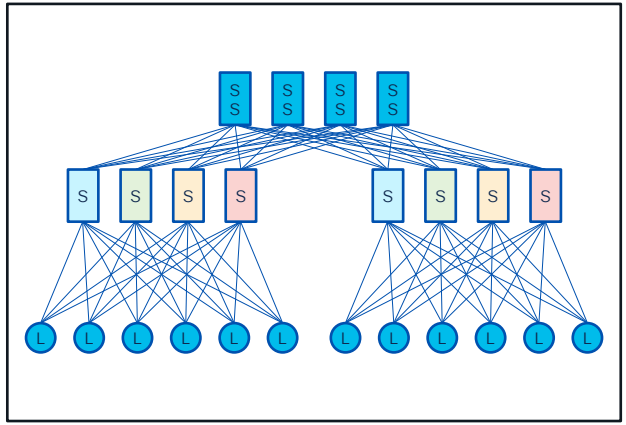


The Journey to Build Better and Further



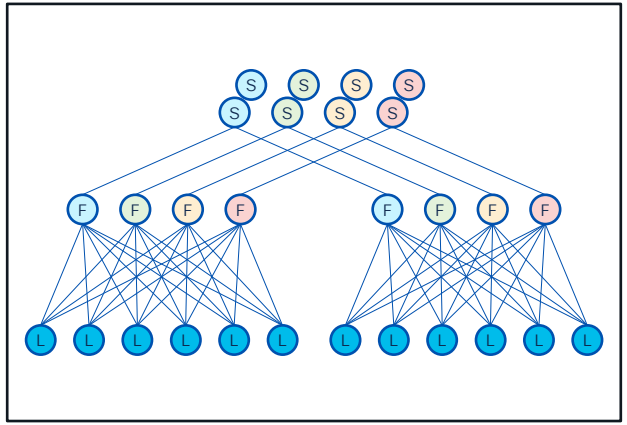
2 Tier Leaf Spine
(5 Stages)

- Use of Modular Spine
- 4-, 8- or 16-Slot Chassis
- 36x 100Gbps (or 36x 400Gbps) Linecard
- 100Gbps Uplink between Leaf and Spine
  - 144 Leaf (6'912 Host Ports)
  - 288 Leaf (13'824 Host Ports)
  - 576 Leaf (27'648 Host Ports)
  - 1152 Leaf (55'296 Host Ports



3 Tier Leaf-Spine-SuperSpine
(11 Stages)

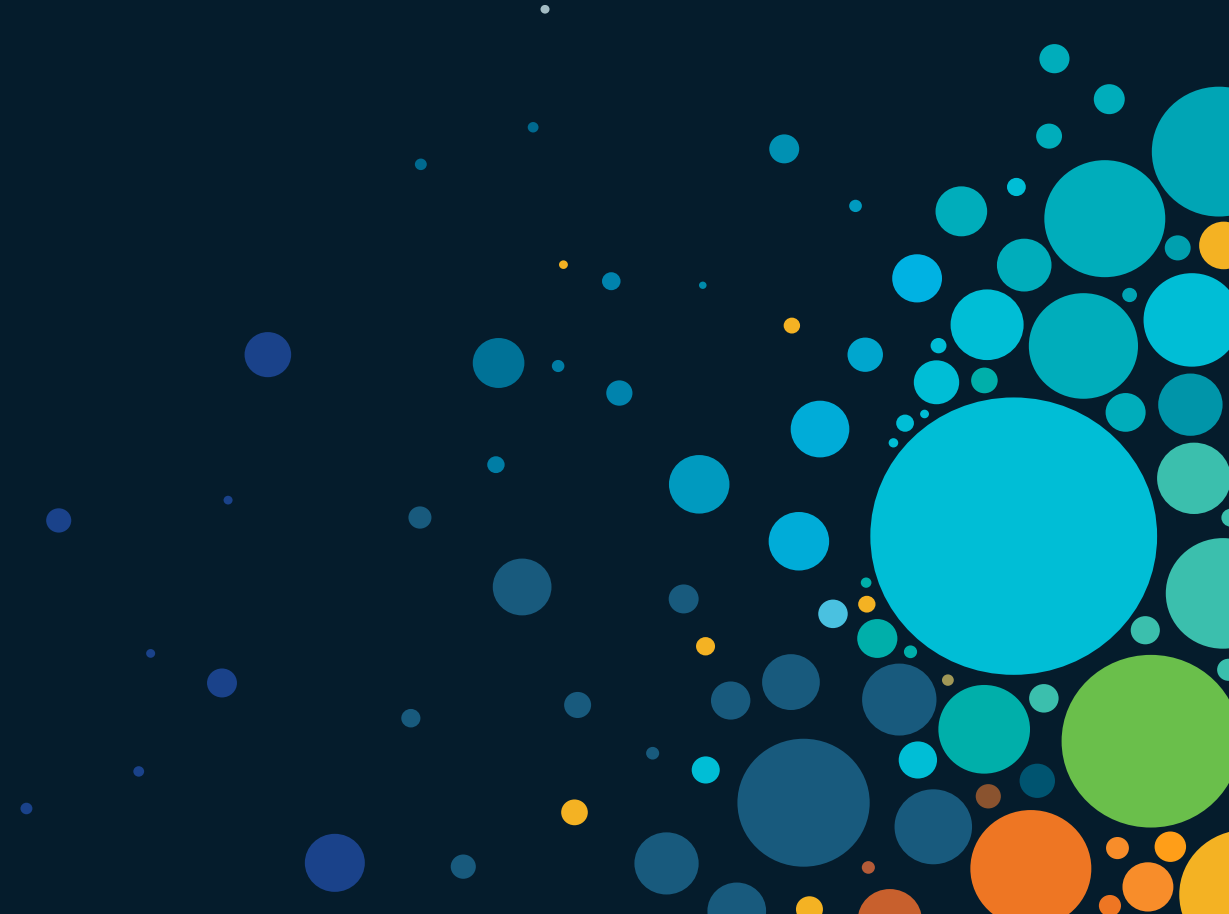
- Use of Modular Spine and SuperSpine
- 16-Slot Chassis
- 36x 100Gbps Linecard
- 100G Uplink Between Leaf, Spine and SuperSpine
  - 512 Leaf (24'576 Host Ports)



3 Tier Leaf-Fabric-Spine
(5 Stages)

- Use of Fixed Fabric and Spine
- 16x to 64x 400Gbps per Fabric (Tier #2)
- 16x to 64x 400Gbps per Spine (Tier #3)
- 100Gbps Uplink between Leaf and Spine
- 400Gbps Uplink between Tier #2 and Tier #3
  - 1 Fabric, 128 Leaf (6'144 Host Ports)
  - 4 Fabric, 128 Leaf (24' 576 Host Ports)
  - 16 Fabric, 128 Leaf (98'304 Host Ports)
  - 128 Fabric, 128 Leaf (786'432 Host Ports)

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

Technical Session Surveys

- Attendees who fill out a minimum of four session surveys and the overall event survey will get Cisco Live branded socks!
- Attendees will also earn 100 points in the Cisco Live Game for every survey completed.
- These points help you get on the leaderboard and increase your chances of winning daily and grand prizes.

Cisco Learning and Certifications

From technology training and team development to Cisco certifications and learning plans, let us help you empower your business and career. www.cisco.com/go/certs

Pay for Learning with Cisco Learning Credits

(CLCs) are prepaid training vouchers redeemed directly with Cisco.

Learn

Cisco U.
IT learning hub that guides teams and learners toward their goals

Cisco Digital Learning
Subscription-based product, technology, and certification training

Cisco Modeling Labs
Network simulation platform for design, testing, and troubleshooting

Cisco Learning Network
Resource community portal for certifications and learning

Train

Cisco Training Bootcamps
Intensive team & individual automation and technology training programs

Cisco Learning Partner Program
Authorized training partners supporting Cisco technology and career certifications

Cisco Instructor-led and Virtual Instructor-led training
Accelerated curriculum of product, technology, and certification courses

Certify

Cisco Certifications and Specialist Certifications
Award-winning certification program empowers students and IT Professionals to advance their technical careers

Cisco Guided Study Groups
180-day certification prep program with learning and support

Cisco Continuing Education Program
Recertification training options for Cisco certified individuals

Here at the event? Visit us at **The Learning and Certifications lounge at the World of Solutions**

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

cisco Live!

Thank you


cisco Live!
ALL IN