Demystifying Clos Fabrics

Chris Woodfield

2/12/2024

Who Am I?

My non-existent beard is quite grey.

25+ years in networking

Mostly specialized in backbone/edge networking until the datacenter team needed some extra BGP know-how. But that's another talk ☺

Severe Imposter Syndrome sufferer. Not going to claim to know it all, please let me know what I'm missing!

NANOG™

So you want to build a Clos fabric...

- Evolution of DC designs (L2 Fat Tree -> Clos)
- Where To Start?
  - Design Inputs
  - Flexible Outputs
- Overview of Design Options
- Gory Details

Background – L2 to L3

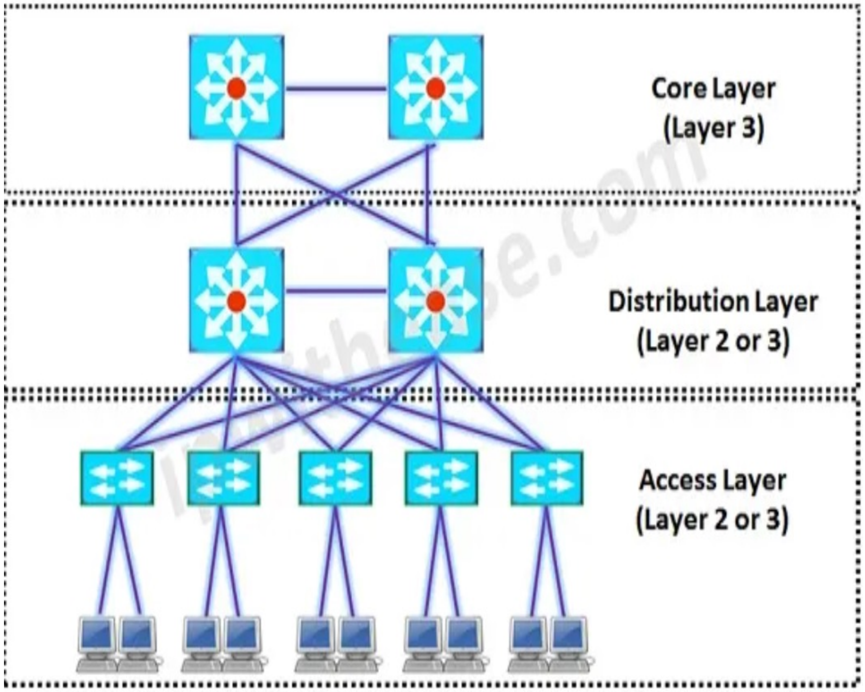
In The Beginning…

- Access, Distribution, Core
- Spanning Tree limited failure models/domains (generally Active/Standby)
- Load sharing at L2 achievable via LACP but few options beyond that
- Even with L3 replacing L2, topologies rarely changed unless a new DC fabric was deployed.

Background – L2 to L3

In The Beginning…

- No overlay/underlays (pre-VXLAN)
- L2 connectivity requirements called for VLAN trunks, with STP to handle redundancy
- Link failovers via “horizontal” handoffs
- Catastrophic failure modes in large L2 domains



Looking Backward to Look Forward

- First things first – it's a name, not an acronym
- First patents by Edward Irwin in 1938
- Charles Clos gets credit for the first production design in 1952.

June 3, 1941.

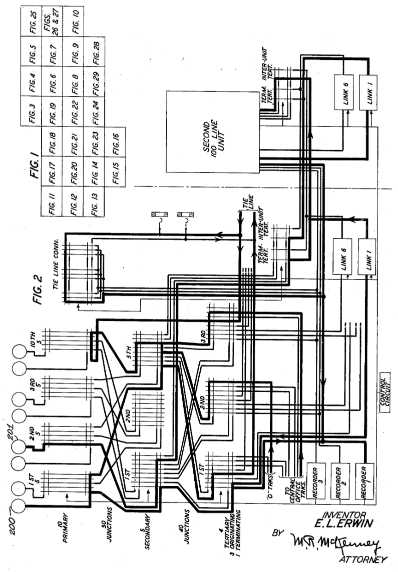
E. L. ERWIN

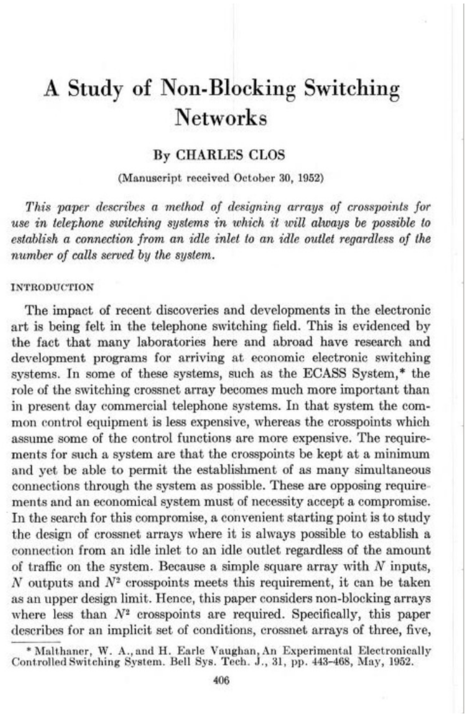
TELEPHONE SYSTEM

2,244,004

Filed Feb. 11, 1938

27 Sheets-Sheet 1





A Study of Non-Blocking Switching Networks

By CHARLES CLOS

(Manuscript received October 30, 1952)

This paper describes a method of designing arrays of crosspoints for use in telephone switching systems in which it will always be possible to establish a connection from an idle inlet to an idle outlet regardless of the number of calls served by the system.

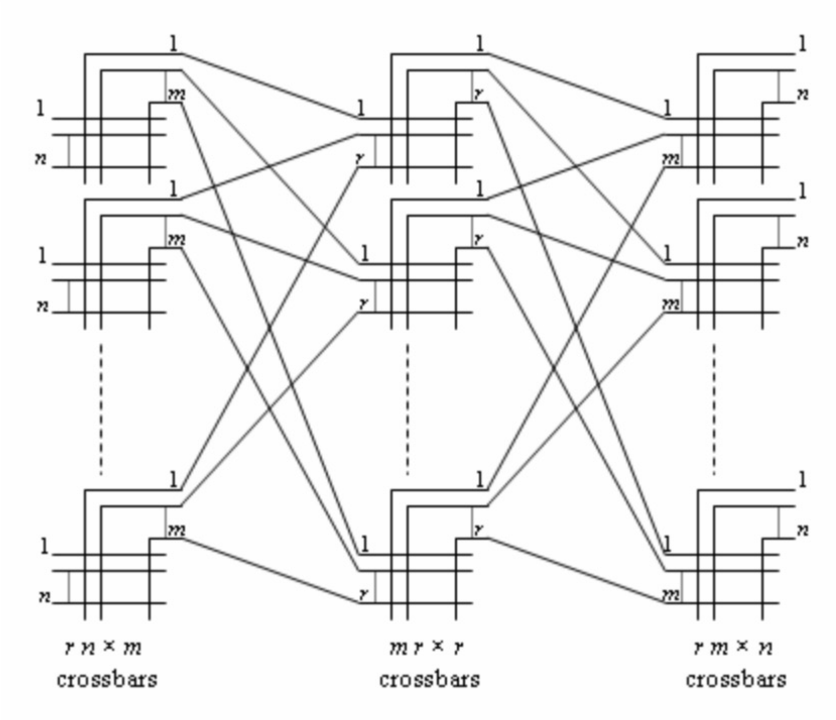
INTRODUCTION

The impact of recent discoveries and developments in the electronic art is being felt in the telephone switching field. This is evidenced by the fact that many laboratories here and abroad have research and development programs for arriving at economic electronic switching systems. In some of these systems, such as the ECASS System,* the role of the switching crossnet array becomes much more important than in present day commercial telephone systems. In that system the common control equipment is less expensive, whereas the crosspoints which assume some of the control functions are more expensive. The requirements for such a system are that the crosspoints be kept at a minimum and yet be able to permit the establishment of as many simultaneous connections through the system as possible. These are opposing requirements and an economical system must of necessity accept a compromise. In the search for this compromise, a convenient starting point is to study the design of crossnet arrays where it is always possible to establish a connection from an idle inlet to an idle outlet regardless of the amount of traffic on the system. Because a simple square array with N inputs, N outputs and N^2 crosspoints meets this requirement, it can be taken as an upper design limit. Hence, this paper considers non-blocking arrays where less than N^2 crosspoints are required. Specifically, this paper describes for an implicit set of conditions, crossnet arrays of three, five,

* Malthaner, W. A., and H. Earle Vaughan, An Experimental Electronically Controlled Switching System. Bell Sys. Tech. J., 31, pp. 443–468, May, 1952.

What *IS* A Clos Fabric?

- Focus on horizontal scaling (more devices) vs vertical (more bandwidth between devices)
- Devices use ECMP to make use of all available paths
- Available ports and ECMP width are primary scalability constraints, not link speeds
- No need for intra-layer links

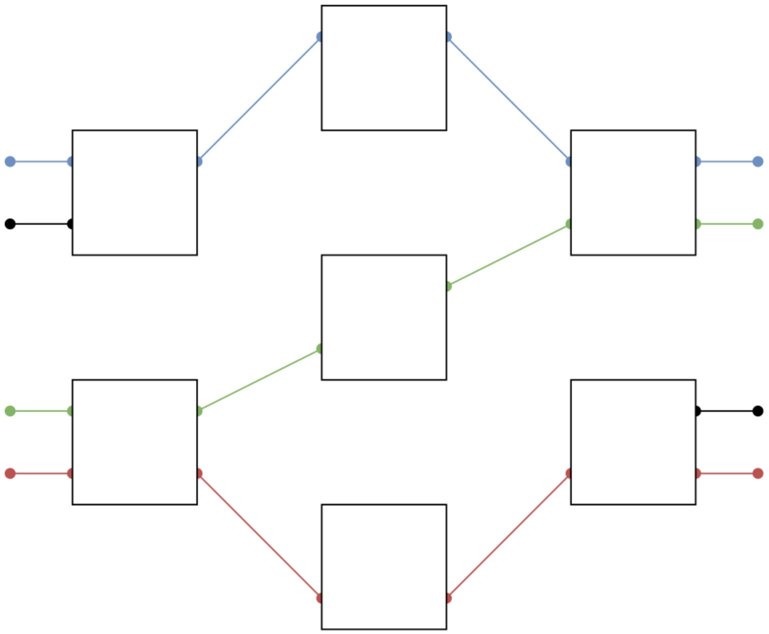


Key Developments

- Improvements in ECMP features/algorithms were key enabler of Clos designs for IP networks
  - 2000s-era hardware only supported ~8 way ECMP, and many products did it *very* poorly.
- L2 to L2.5 to L3 designs (required L2 encapsulation solutions)
- IGP (OSPFv2/v3, IS-IS) link-state complexity gives rise to BGP-based DC designs (RFC7398)
  - IGP still fine for small-to-medium-size fabrics
  - BGP Equal Cost Multipath

Start: Questions To Ask Yourself

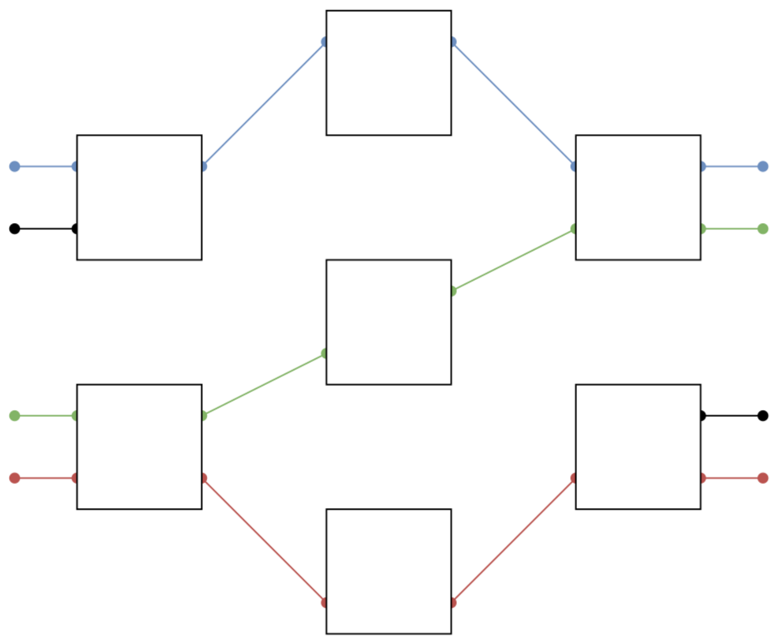
- Will this be a Layer 3 only network, or underlay-overlay?
- Dual-stack or IPv4/v6 only?
- What are your scaling constraints beyond the network itself?
  - Cage/room size, power budget
  - Max expected compute needs



NANOG™

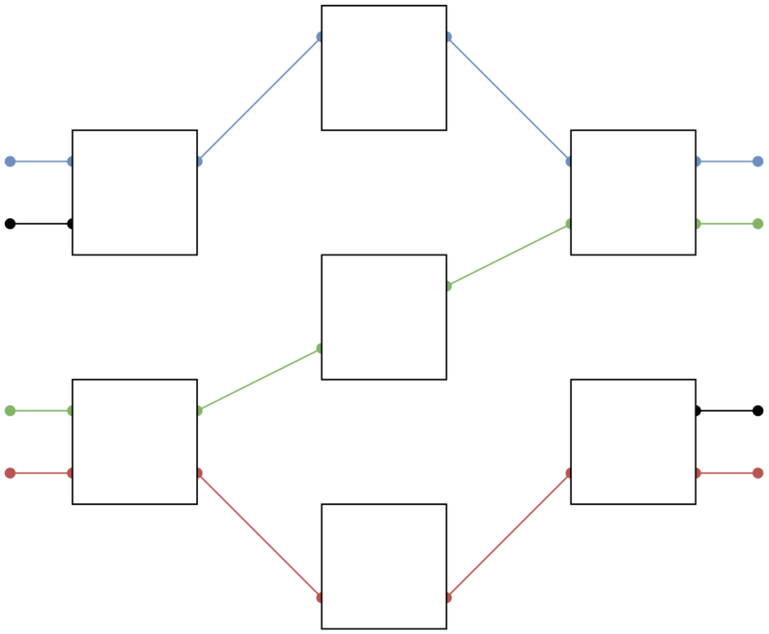
Questions To Ask Yourself

- Per-Rack bandwidth requirements
  - 10/25/100G to host?
  - Hosts per rack?
- Oversubscription Requirements
- Failure Tolerance (50% is decent rule of thumb)
  - Consider MTTR of failed links/devices



Golden Rule Of DC Design

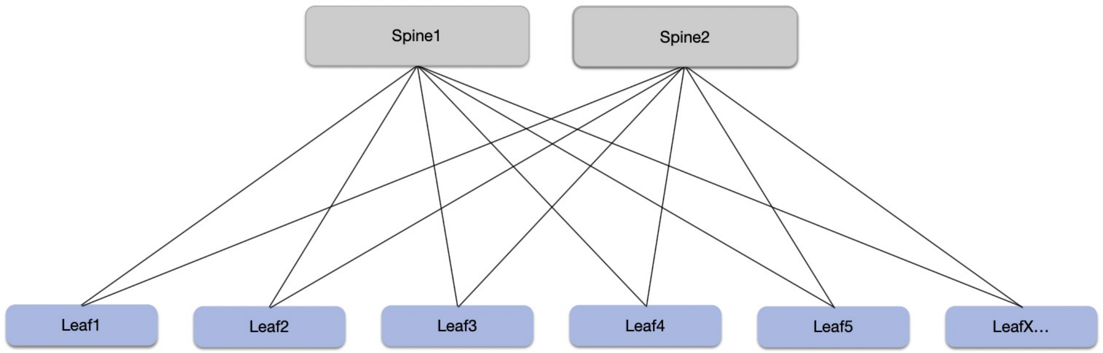
- *Design* the maximum scale you know you will need/can build in the space.
- *Implement* the design organically.
- *Single Source Of Truth* for provisioning data.
  - *Automate* provisioning to make capacity adds safe and routine.



Simple Leaf/Spine Clos Network 101

3-stage – Leaf -> Spine -> Leaf path

More complex designs are out there, but will you need them?

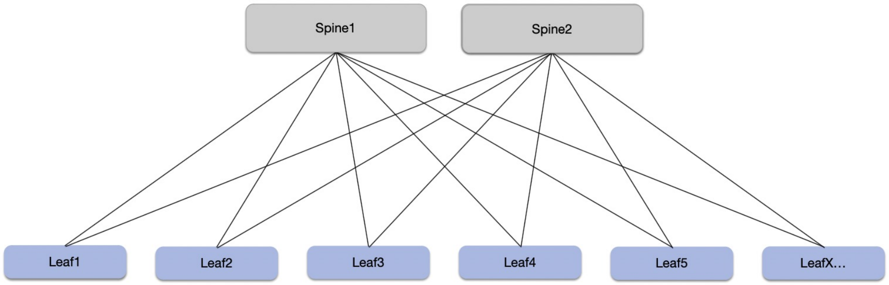


Simple Leaf/Spine Clos Network

Hardware assumptions (not-latest-generation):

- Leaf: 32x100G (w/ 25/10G breakout capabilities)
- Spine: 64x100G
- 1RU Fixed Form Factor for leaf, 2RU for spine

- 2x Spine delivers:
  - 100Gbps per leaf at 50% redundancy
  - Reserve leaf ports for additional spines

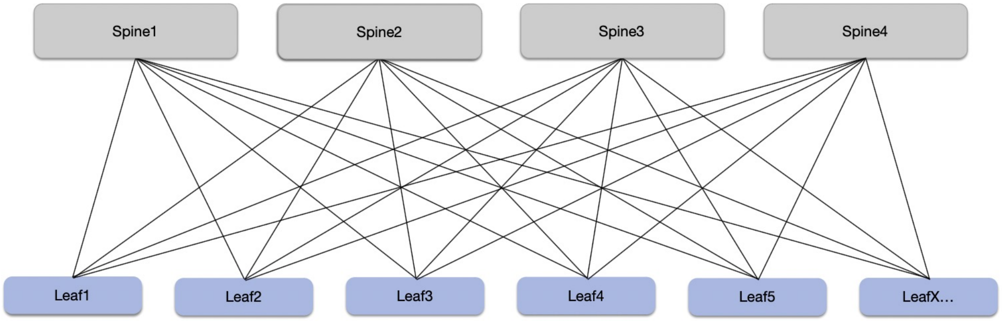


Simple Leaf/Spine Clos Network

Hardware assumptions (Tomahawk3 or similar):

- Leaf: 32x100G (w/ 25/10G breakout capabilities)
- Spine: 64x100G
- 1RU Fixed Form Factor for leaf, 2RU for spine

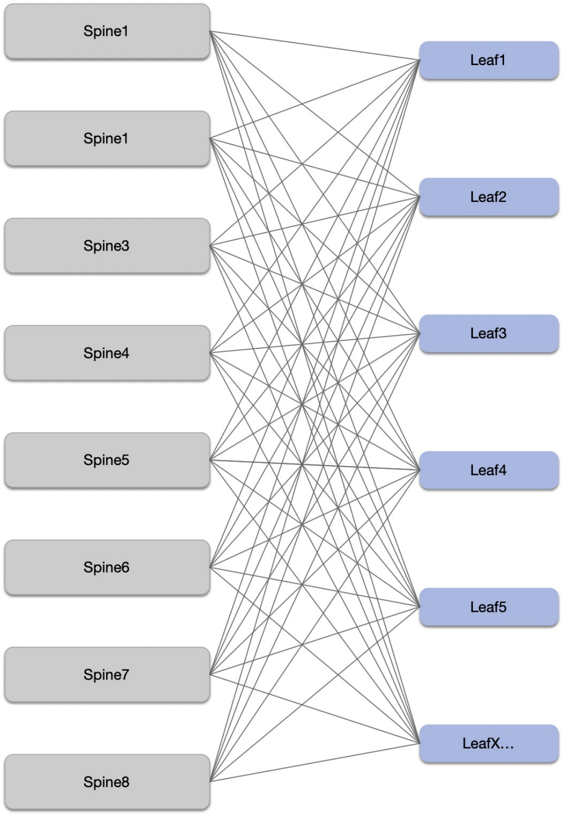
- 4x Spine delivers:
  - 200Gbps per leaf at 50% redundancy



Simple Leaf/Spine Clos Fabric

With 8x Spine Count:

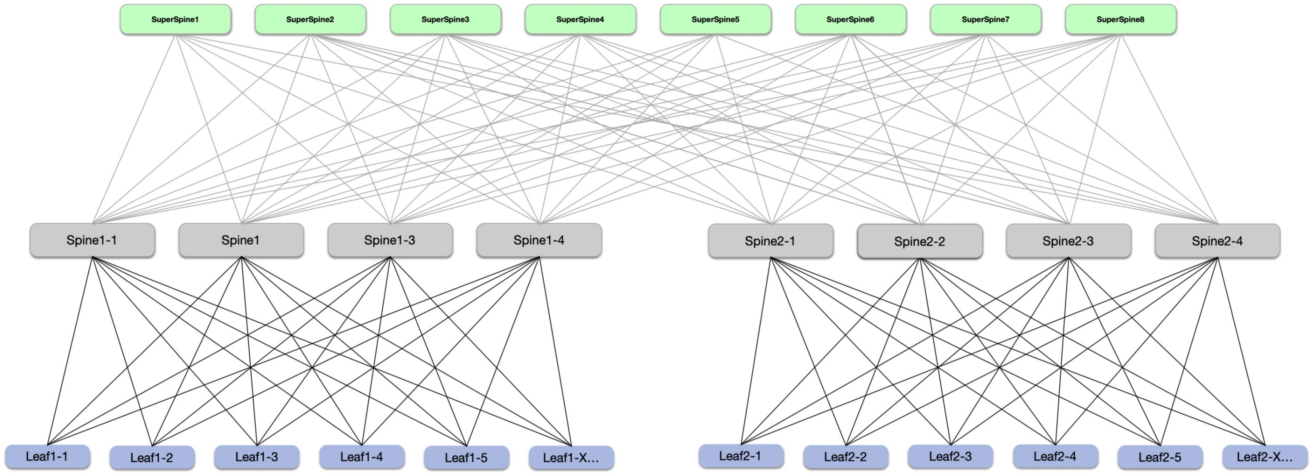
- 40x 1RU hosts per rack, 25G per host = 1Tbps per rack
  - 1.66:1 oversubscription at 75% capacity
- 20x 2RU hosts per rack, 100G per host = 2Tbps per rack
  - 3.33:1 oversubscription at 75% capacity
  - Need 2xRU leaf switch (64x100G)



NANOG™

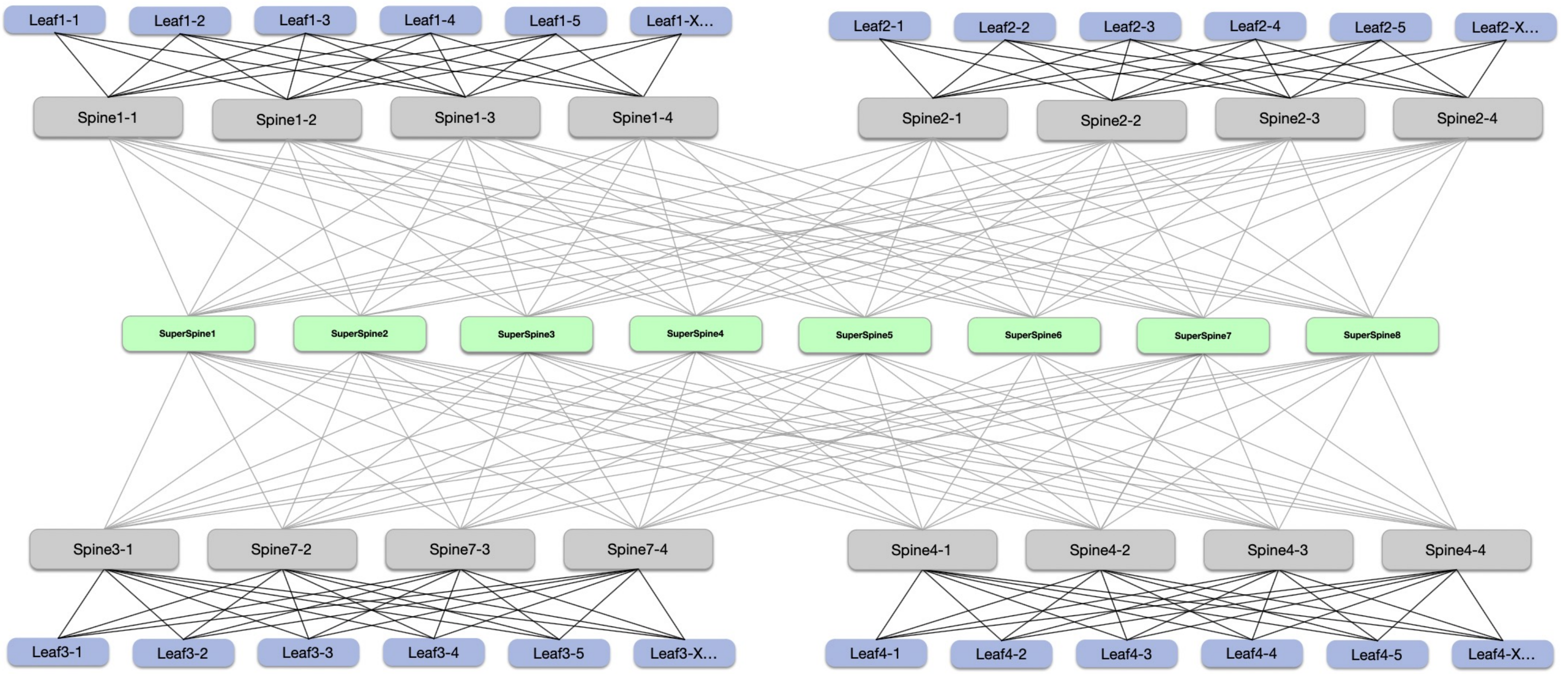
Going Bigger?

- Add a “Super Spine”, serving multiple clusters.
- 5-stage Clos (Leaf1 -> Spine1 -> SuperSpine -> Spine2 -> Leaf2)

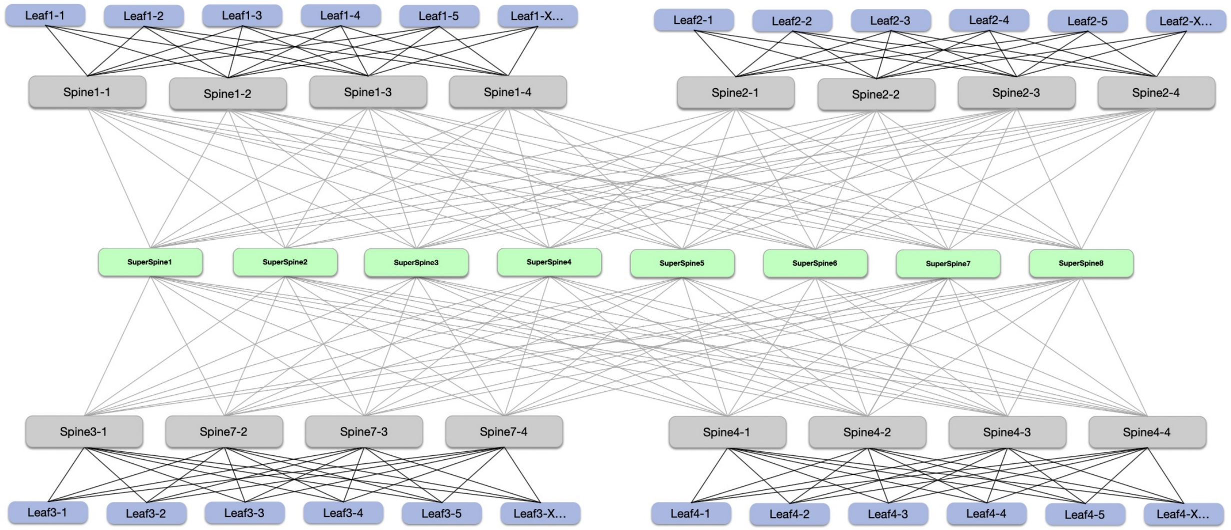


NANOG™

Going Bigger – 5-Stage Clos

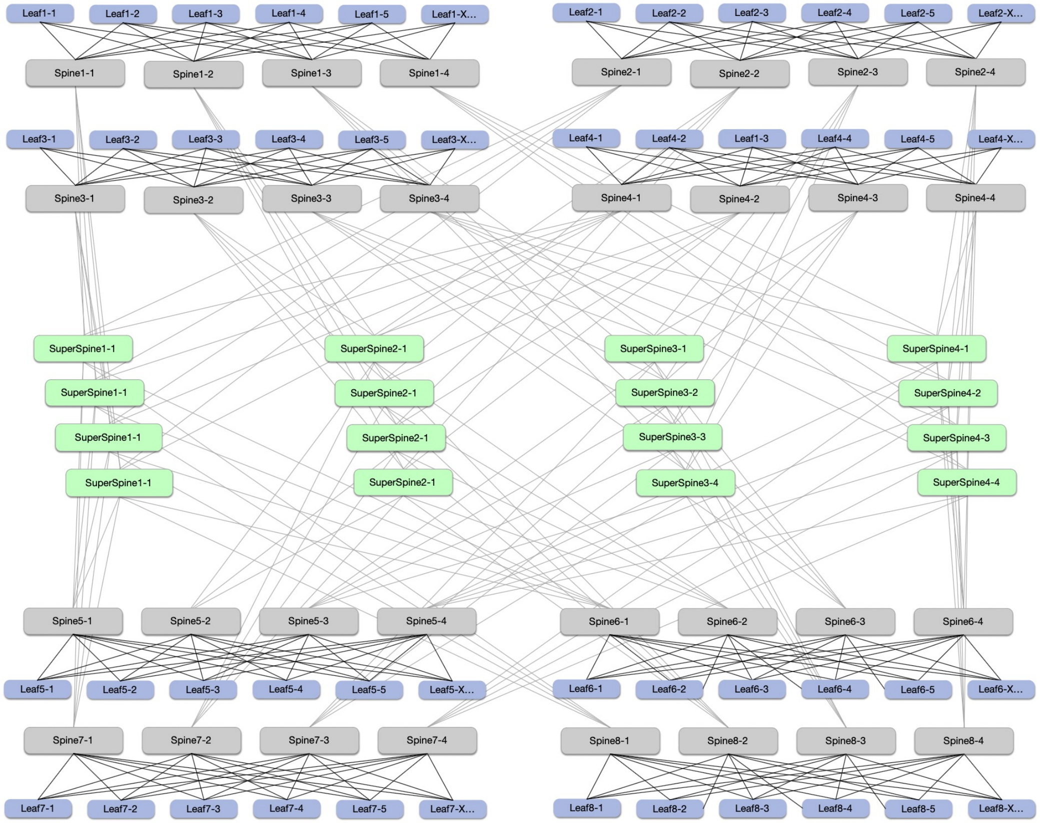


- 8x SuperSpines x 4 Spines/cluster –
  - 200Gbps per leaf @50% capacity
  - 1.6Tbps per cluster @50% - cluster count only limited by SS port capacity
  - 16 clusters w/ 64x100G devices, more with modular



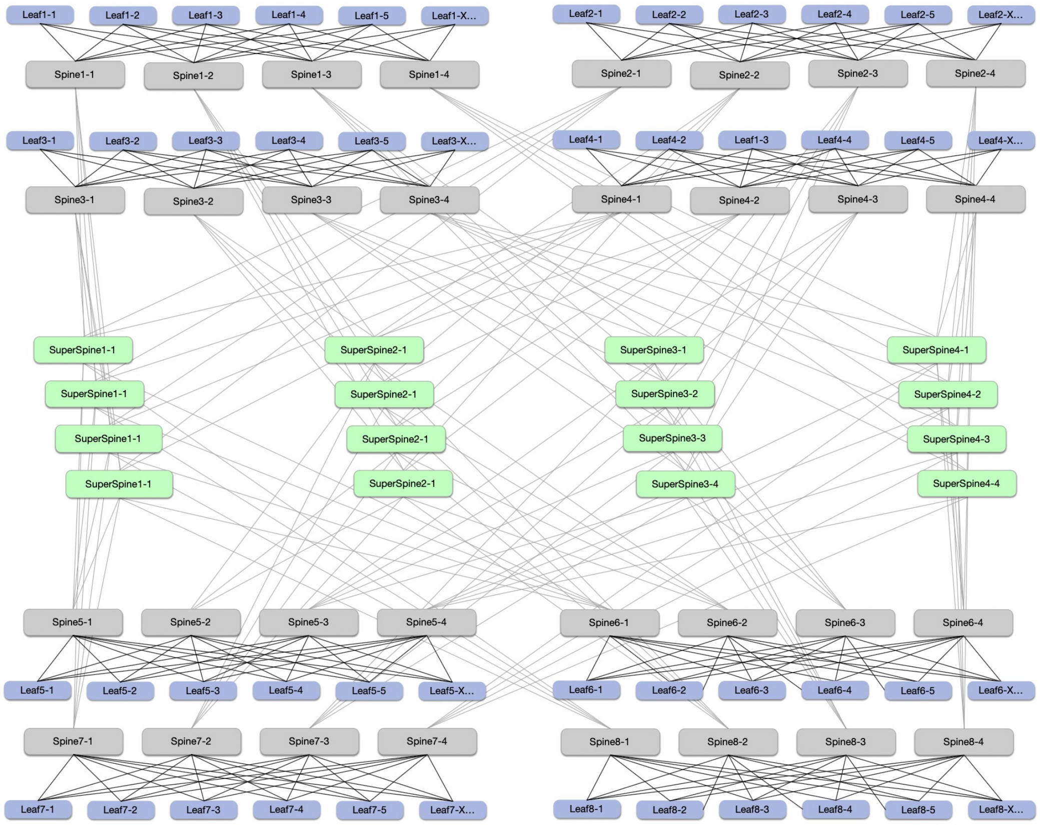
Going EVEN Bigger?

- Multiple SuperSpine layers – optimal for fixed form factor devices at all layers
- Each Spine connects to one SuperSpine layer only



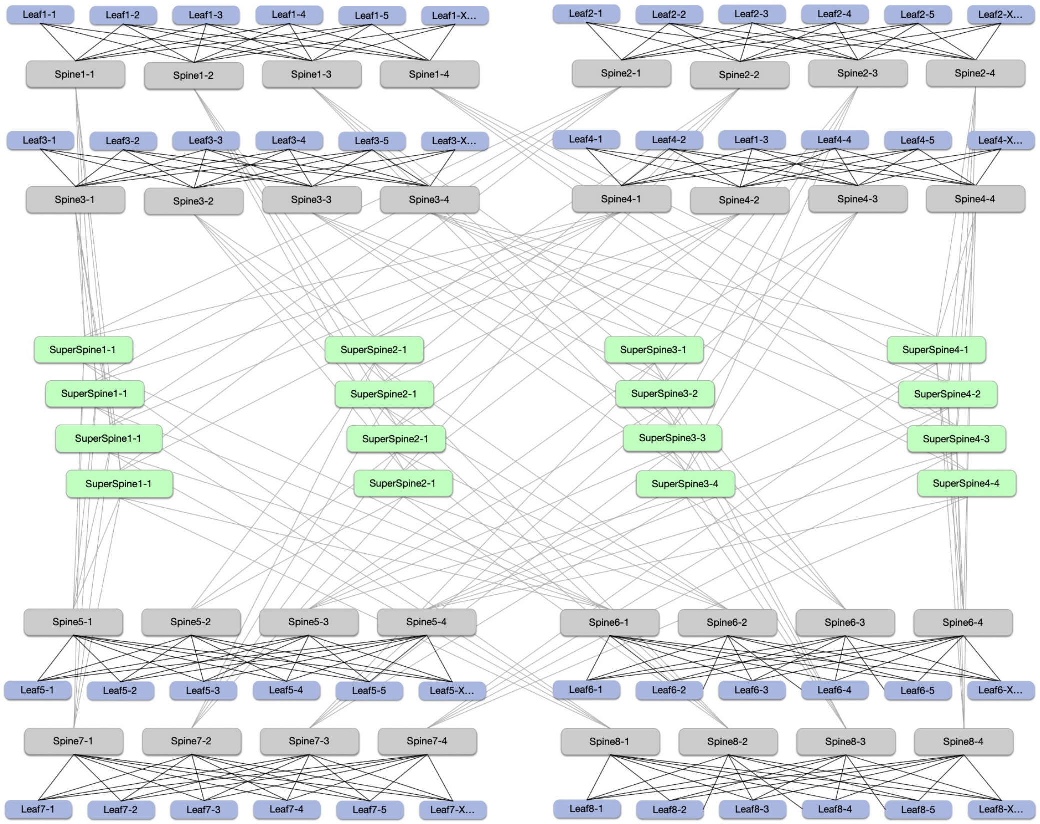
Going EVEN Bigger?

- Lower per-cluster bandwidth – fewer racks supported per cluster, in trade for wider cluster scalability.
- Can scale this by adding devices to each SuperSpine



Going EVEN Bigger?

- ECMP “spray” limited to each device’s uplinks, may help keep link capacity more uniform
- *Lots* more devices. Automation becomes a must have, not a nice-to-have.


Leaf1-1 Leaf1-2 Leaf1-3 Leaf1-4 Leaf1-5 Leaf1-X...
Spine1-1 Spine1-2 Spine1-3 Spine1-4
Leaf2-1 Leaf2-2 Leaf2-3 Leaf2-4 Leaf2-5 Leaf2-X...
Spine2-1 Spine2-2 Spine2-3 Spine2-4
Leaf3-1 Leaf3-2 Leaf3-3 Leaf3-4 Leaf3-5 Leaf3-X...
Spine3-1 Spine3-2 Spine3-3 Spine3-4
Leaf4-1 Leaf4-2 Leaf1-3 Leaf4-4 Leaf4-5 Leaf4-X...
Spine4-1 Spine4-2 Spine4-3 Spine4-4
SuperSpine1-1
SuperSpine1-1
SuperSpine1-1
SuperSpine1-1
SuperSpine2-1
SuperSpine2-1
SuperSpine2-1
SuperSpine2-1
SuperSpine3-1
SuperSpine3-2
SuperSpine3-3
SuperSpine3-4
SuperSpine4-1
SuperSpine4-2
SuperSpine4-3
SuperSpine4-4
Spine5-1 Spine5-2 Spine5-3 Spine5-4
Leaf5-1 Leaf5-2 Leaf5-3 Leaf5-4 Leaf5-5 Leaf5-X...
Spine6-1 Spine6-2 Spine6-3 Spine6-4
Leaf6-1 Leaf6-2 Leaf6-3 Leaf6-4 Leaf6-5 Leaf6-X...
Spine7-1 Spine7-2 Spine7-3 Spine7-4
Leaf7-1 Leaf7-2 Leaf7-3 Leaf7-4 Leaf7-5 Leaf7-X...
Spine8-1 Spine8-2 Spine8-3 Spine8-4
Leaf8-1 Leaf8-2 Leaf8-3 Leaf8-4 Leaf8-5 Leaf8-X...

Protocol Choices

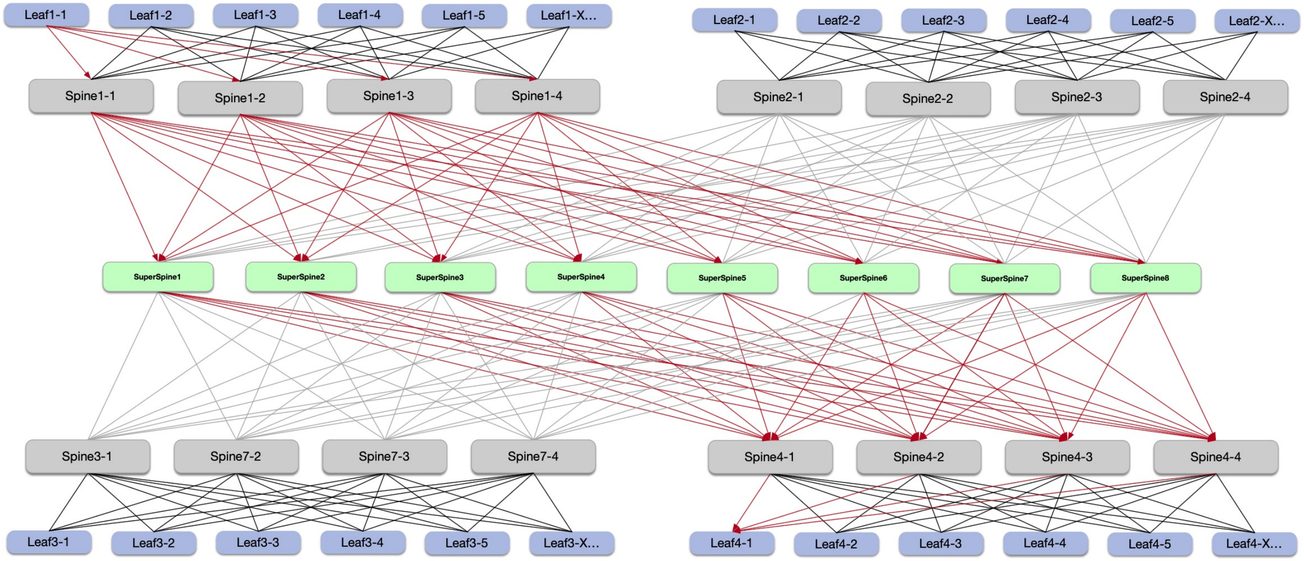
- EBGP tends to be most widely used, but is the most config-intensive (read: automate your peer configs)
- OSPF/OSPFv3 viable for pure-L3 routing (overlays can still be handled at edge w/ iBGP), but pay careful attention to route/LSA scale.
- BGP models:
  - Each device its own ASN (RFC7398) – use 32-bit ASN space
  - Can duplicate ASNs across layers (cluster spines, etc) – this will eliminate layer-level loops via BGP loop prevention
  - Be very careful if/where you aggregate
  - BFD? Link loss may be all the signal you need.

Addressing/policy Choices

- It is the Year Of Our Lord 2024. PLEASE run IPv6.
- If your prod traffic is overlay, consider an IPv6-only underlay if your hardware supports it (and it doesn’t, find a different vendor)
- Aggregate device links (loopback and interfaces) – cheap route optimization. Most TCAMs don’t have IPv6 exact-match FIB for /128s.
- Implement GSHUT (RFC8326) in your BGP policy if not already supported – makes it very easy to drain traffic from a device while keeping it on-net for troubleshooting.

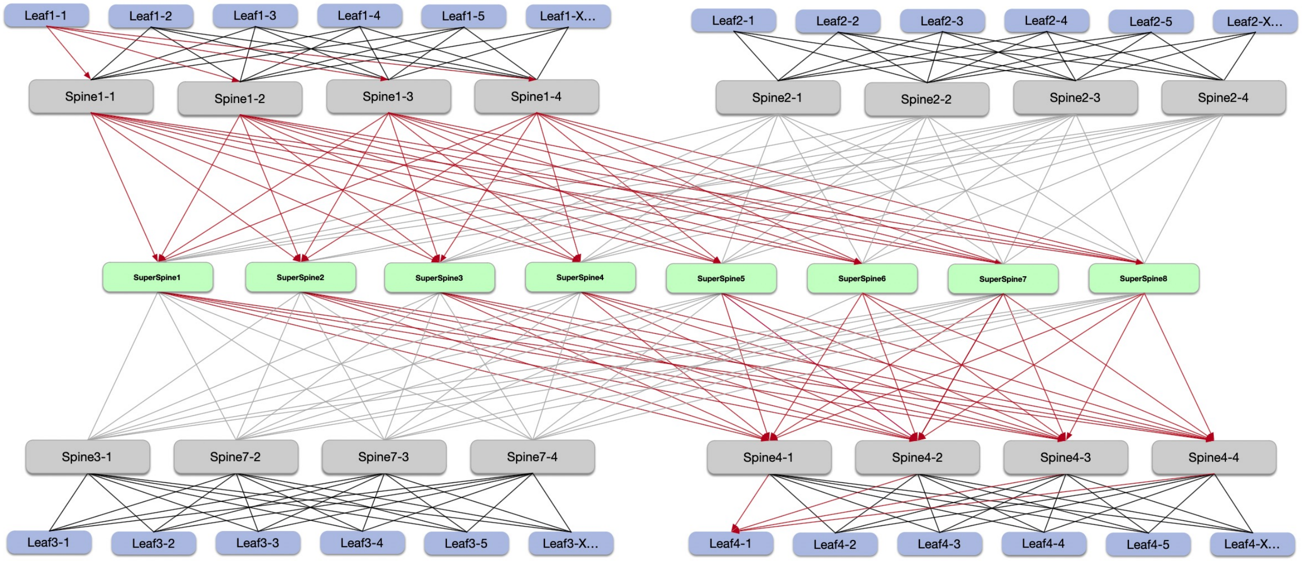
Monitoring/Path Explosion

- 4 * 8 * 4 = 128 possible paths between leaf devices (!)
- One lossy path can ruin your day



Monitoring/Path Explosion

- Flow-based traces necessary, to test all possible hashing combinations (pingmesh)
- Can *you* spot the bad path?



Thank you

2-12-2024

NANOG™