Demystifying Data Center CLOS networks

Overview of Yahoo's Data Center Fabric v3

(2012-2018)

Igor Gashinsky, Yihua He

June 2019

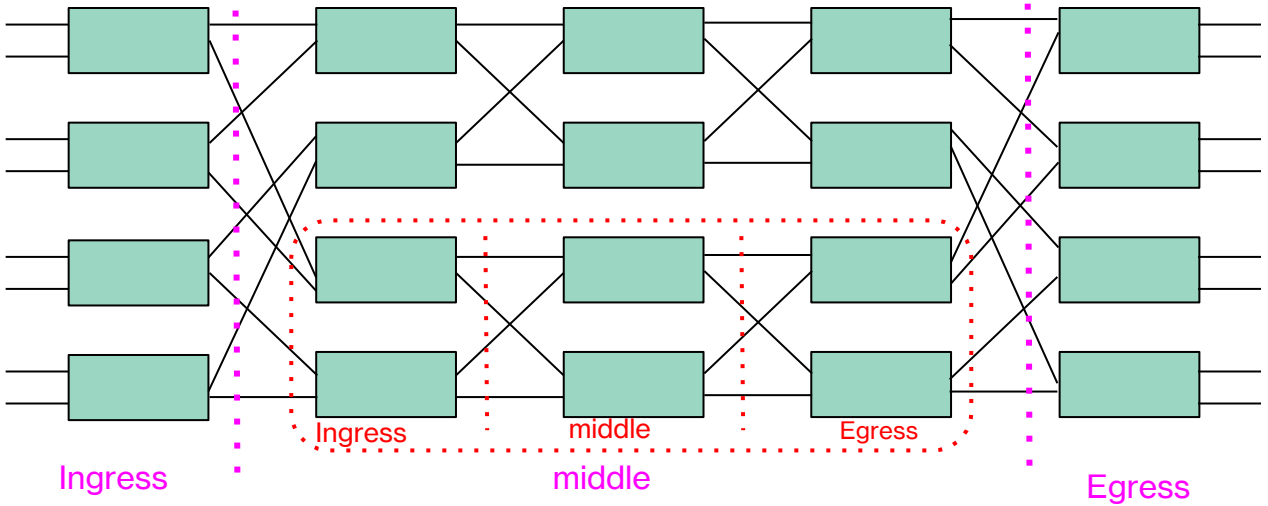
verizon media

Agenda

1. What's CLOS?
2. Different CLOS topologies
3. Gory Details of our topology
4. Scaling
5. Automation
6. Operations

What's a CLOS?

- "A Study of Non-blocking Switching Networks" Bell System Technical Journal, 1953
  - multistage switching network
  - ingress stage, the middle stage, and the egress stage
  - can be recursive!
- What everyone calls a "Leaf-Spine" design
- Beneš network



verizon media

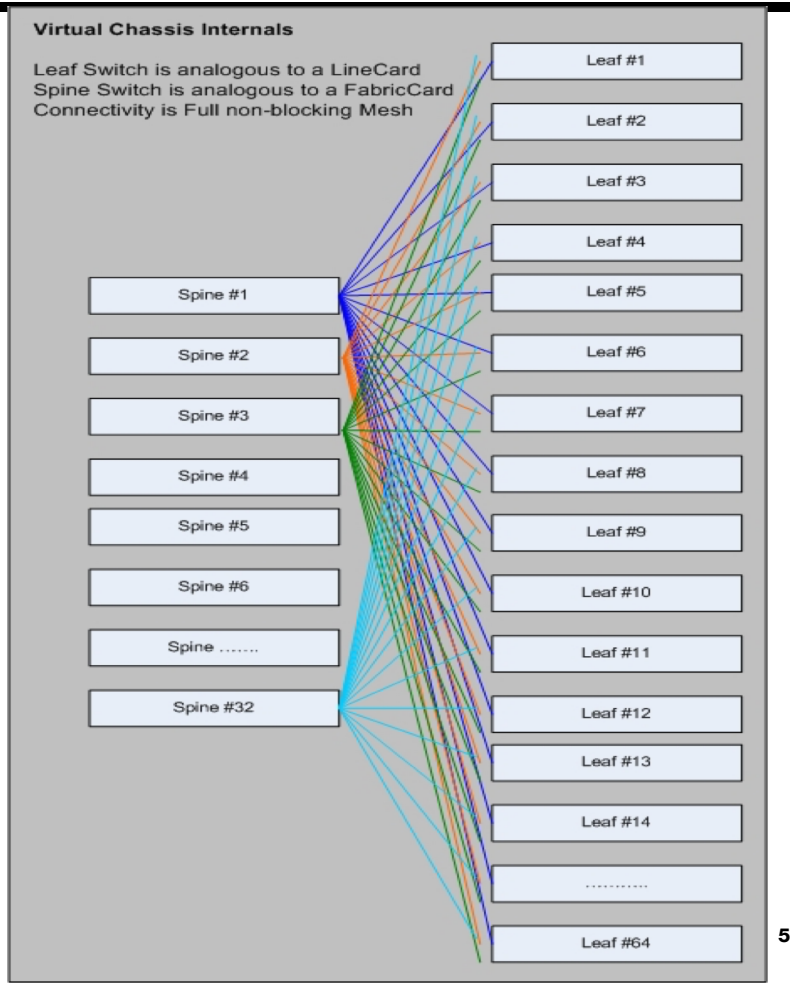
Different CLOS topologies

- **Not surprisingly, most modern network hardware uses some type of a CLOS already**
  - Chassis Routers/Switches
  - Linecards

verizon media

Virtual Chassis

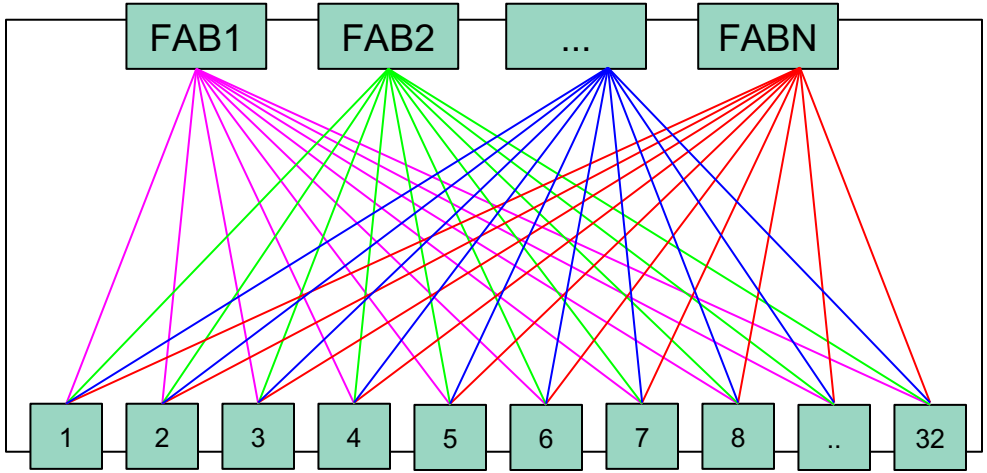
- Spine = Fabric Card
- Leaf = Linecard
- Internal interconnect = PCB x-bar



verizon media

Virtual Linecard

- 32 TOR's = 32 ports
- FAB(ric) switches = Fabric Chips
- Connect to a common fabric



verizon media

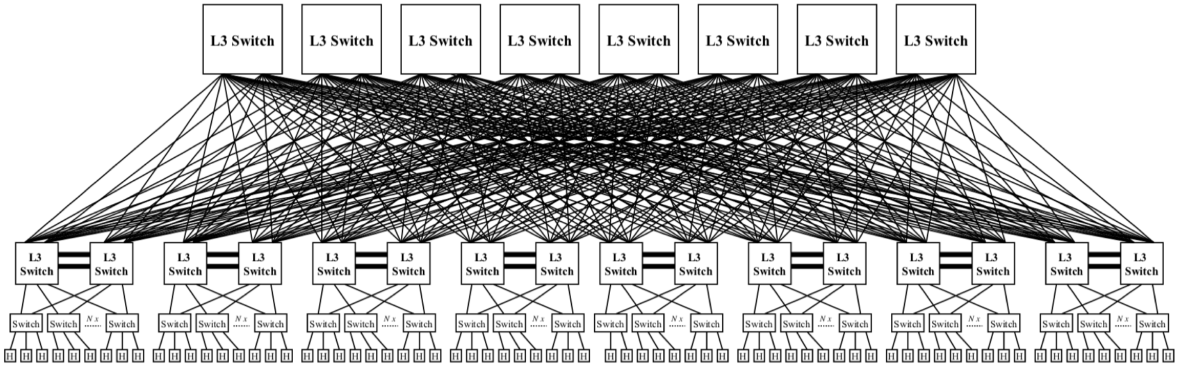
NANOG 40 - Feb 2008

We are currently building this…

** 8 way ECMP w/ 2x10GE LAGs**

** Way too many paths **

** Way too many cables **



L3 Switch <10GE> L3 Switch
L3 Switch <GE> Switch
Host <GE> Switch

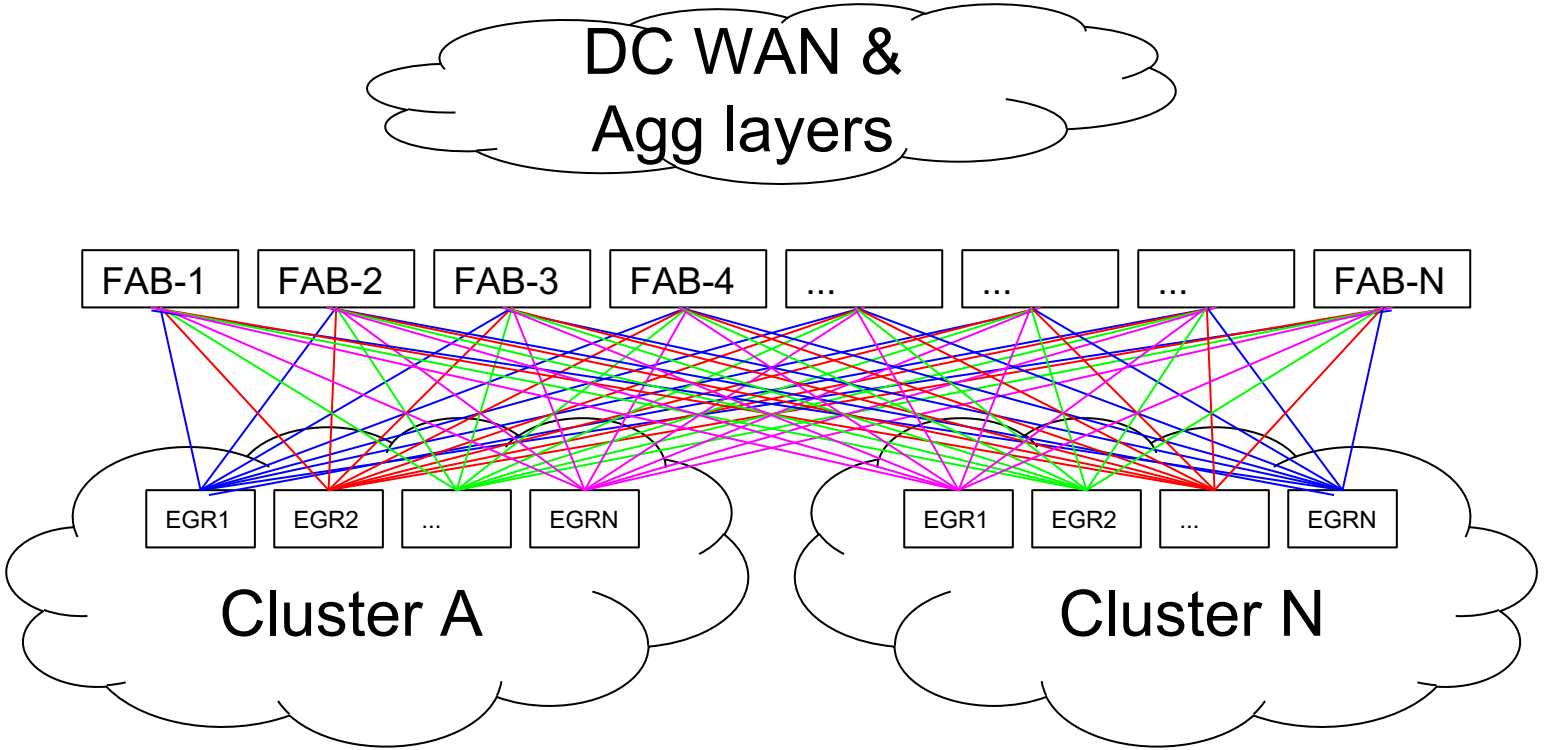
Fast Forward to 2012

Multidimensional Folded Clos Fabric

- First deployed in 2012
- 1k, 5k, 10k, 20k Node Cluster Sizes in production
  - scales to 40k, 80k, 160k, 320k nodes in a single cluster
  - blast domain vs scaling size tradeoff
- Clusters interconnected with a common East-West fabric layer
- Old: 10G to Server, 40G Core
- New: 25G to Server, 100G Core
- Layer 3, dual stack IPv4/IPv6, BGP-based
- Fully Automated provisioning, self healing system

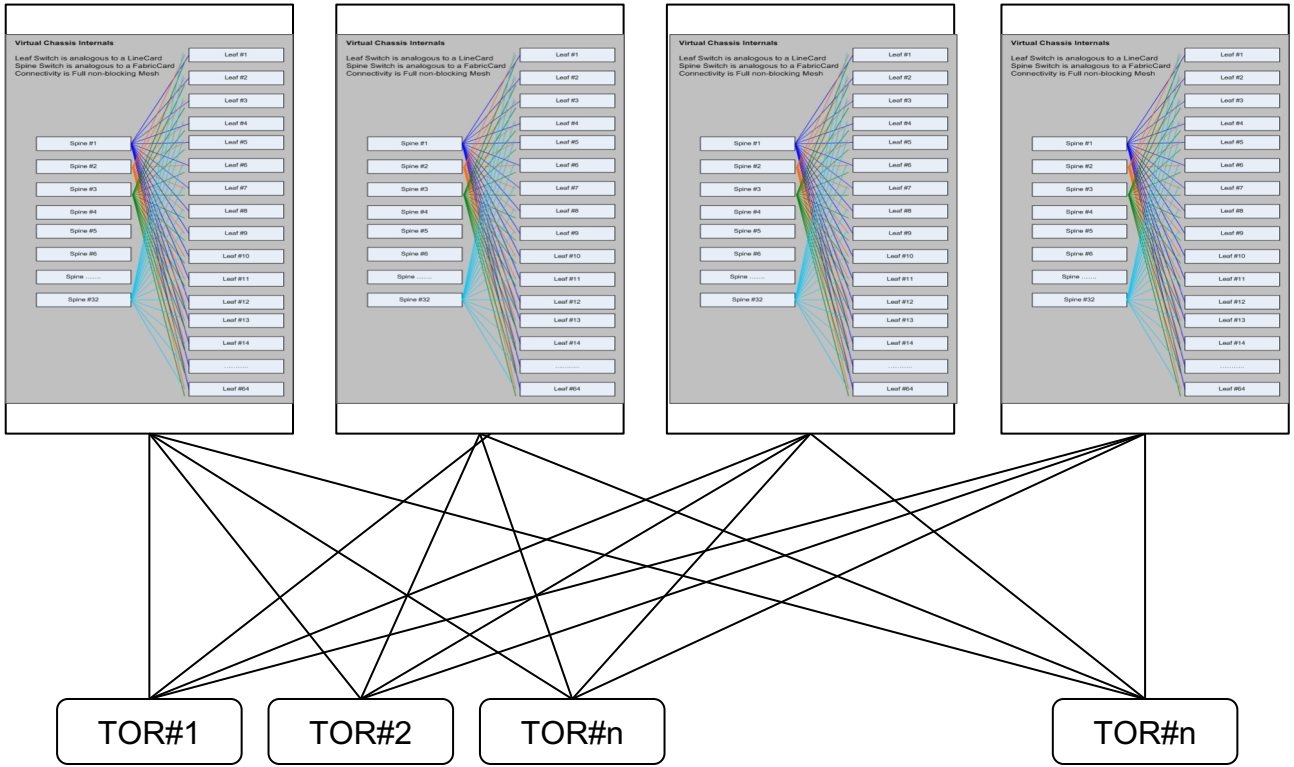
verizon media

High-level Topology



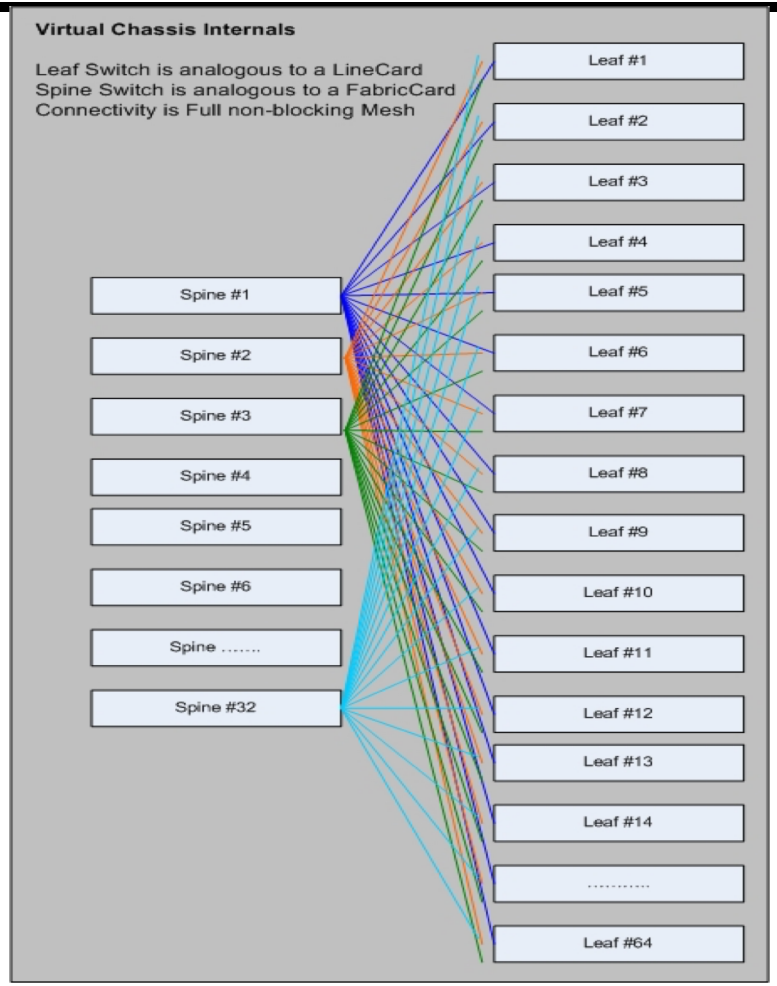
verizon media

High-level - Cluster



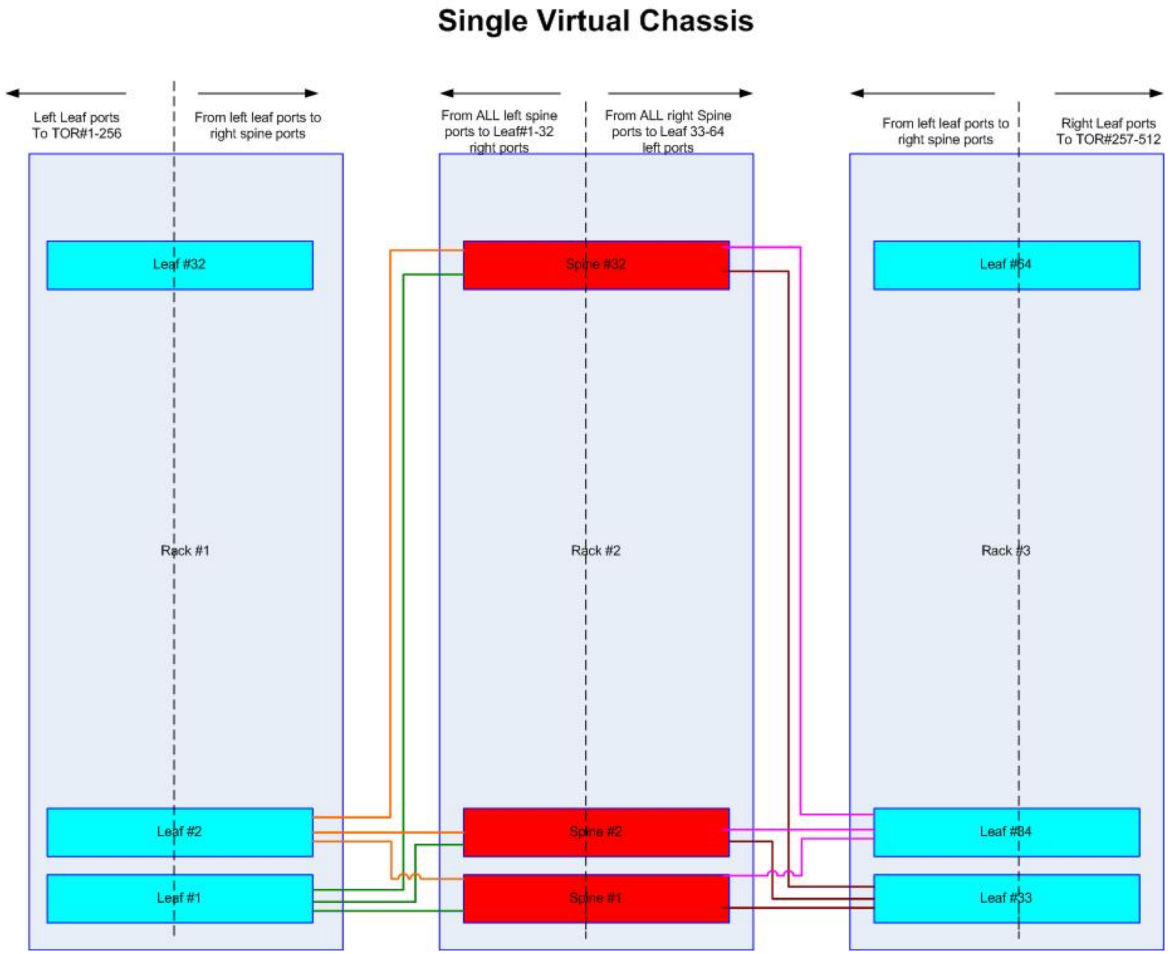
verizon media

Looks familiar?



verizon media

Physical layout



verizon media

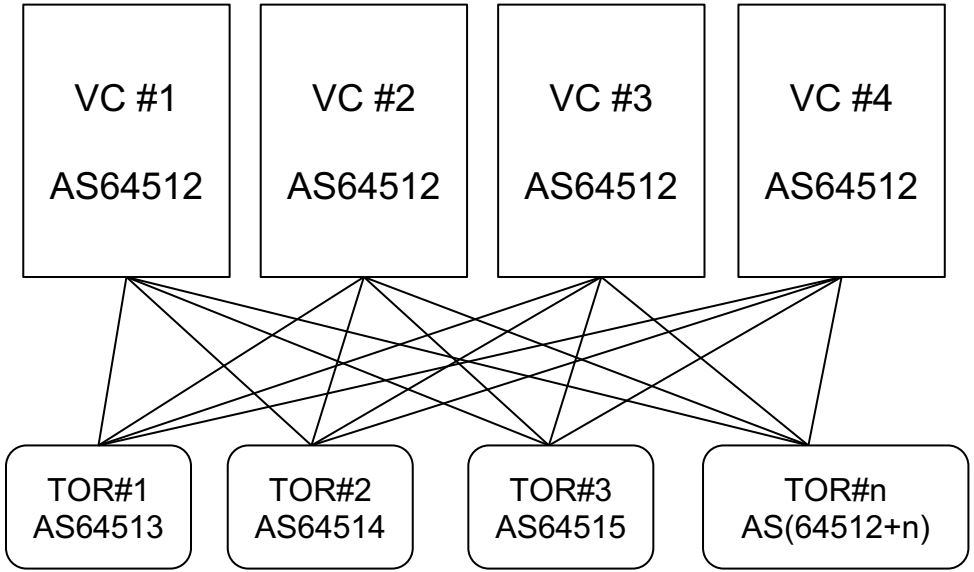
First Prototype - 2011

verizon media

Actual picture

verizon media

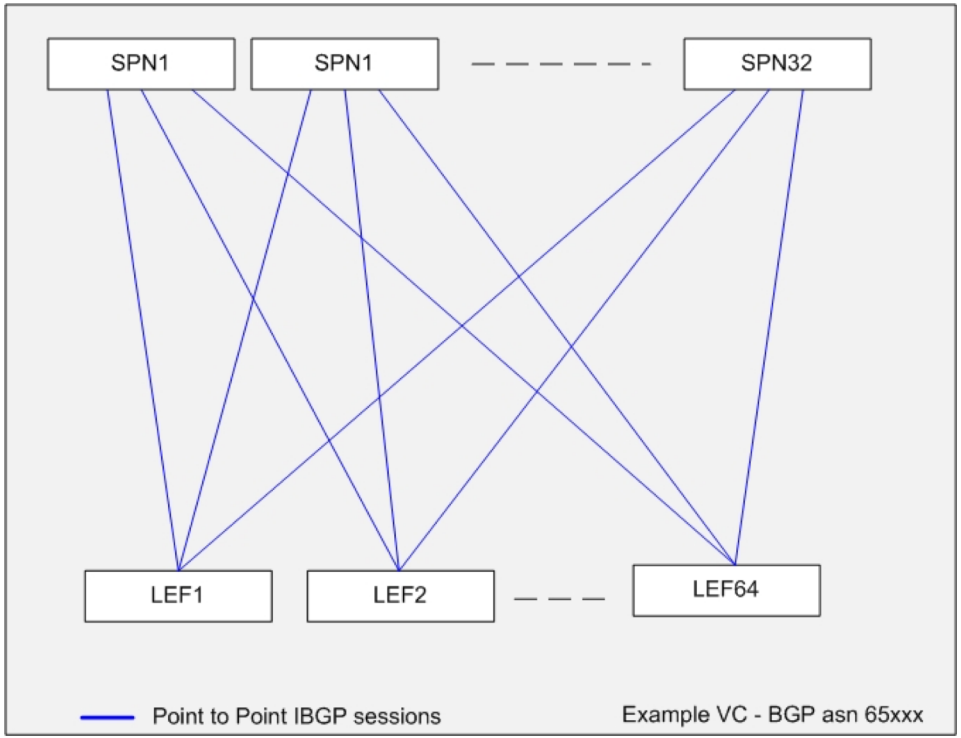
TOR perspective



- **Each TOR uses a unique private ASN**
- **TOR EBGP peers to a single LEF on each of the N Virtual Chassis's**
- **TORs have network statements for Lo0 and all host subnets**

verizon media

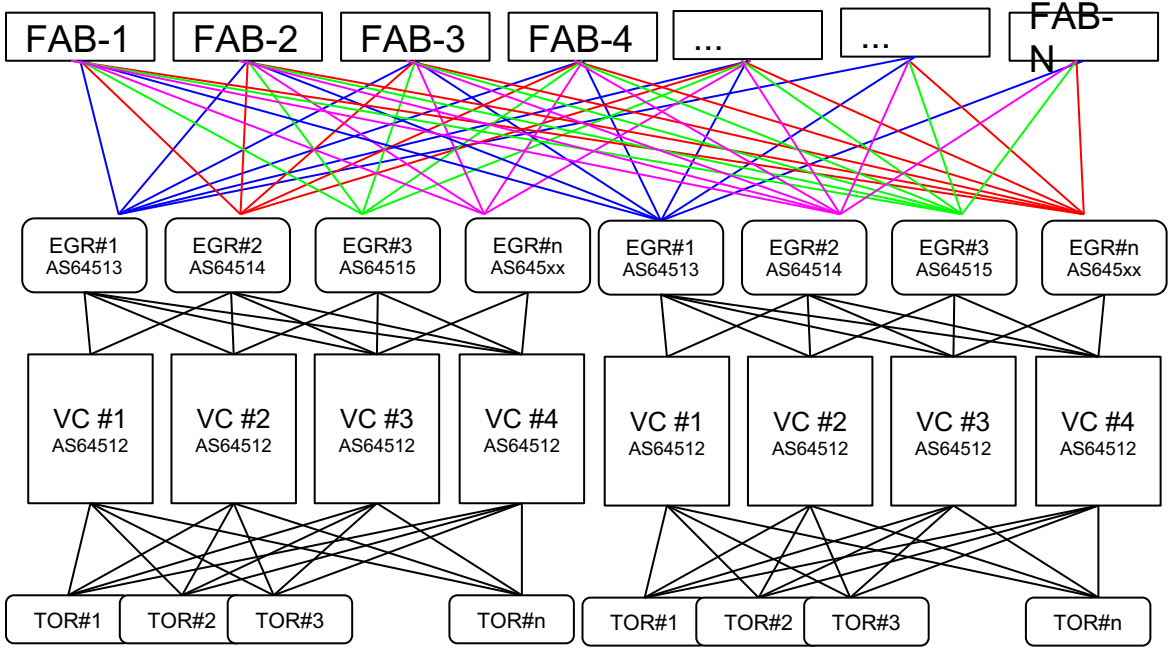
Inside the Virtual Chassis



- **Each Virtual Chassis uses a private ASN**
- **SPN is a route reflectors to the LEF**
- **SPN-LEF ibgp sessions are pt-2-pt**
  - update src local-interface
- **SPN has network statements for all SPN-LEF interconnects**
- **LEF has network statements for LEF-TOR interconnects**
- **LEF uses next-hop-self for SPN-LEF ibgp sessions**
- **All BGP next hop addresses are learned via bgp**
  - There is no IGP inside of the VC
- **All Virtual Chassis's use the same ASN**

verizon media

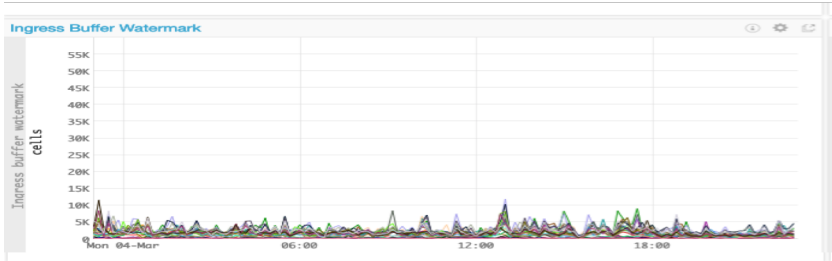
Cluster to fabric connectivity

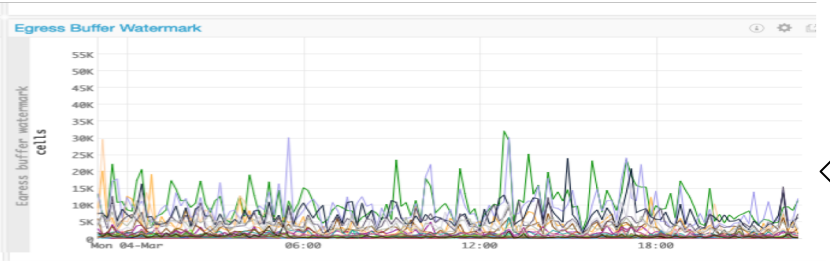


- **EGR = EGress Router,**
  - Same class of devices as SPN/LEF
  - EGRs are like TORs
- **EGRs are special**
  - Each EGRs can connect to multiple LEFs in a single VC and multiple VCs
  - EGR can aggregate cluster subnets
  - variation on traditional CLOS architecture
- **FABs**
  - connect to EGRs in multiple clusters
  - Use “remove-private” so that multiple clusters can use the same set of ASNs
  - speak eBGP to EGRs and OSPF to higher level of devices
  - redistribute aggregated subnets from each cluster to rest of DC topology

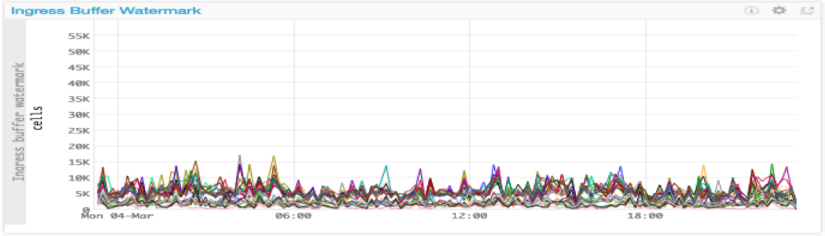
verizon media

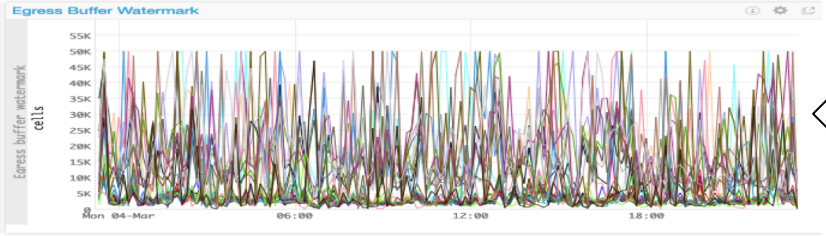
Incast & Buffer Pressure



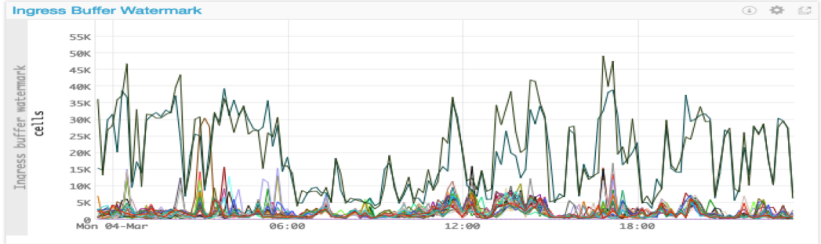


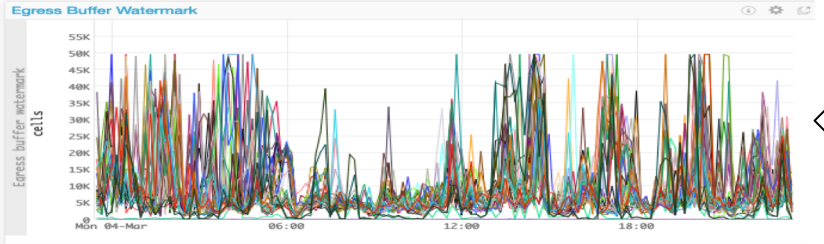
SPINE





LEAF





TOR

verizon media

Scaling

- **Single cluster capacity**
  - Number of edge positions (TOR or EGR) = R_{spn}* R_{lef}/2, where R is switch port radix
    - R_{spn}= 32, R_{lef}= 32 => 512 position = 20K nodes
    - R_{spn}= 64, R_{lef}= 32 => 1024 positions = 40K nodes
    - R_{spn}= 64, R_{lef}= 64 => 2048 positions = 80K nodes
    - R_{spn}= 128, R_{lef}= 128 => 8192 positions = 320k nodes
  - TOR oversubscription ratio is decided by number of VCs (or number of uplinks)
    - 2VC -- 1:6, 4 VC -- 1:3, 6 VC --- 1:2, 8 VC --- 1:1.5, 12 VC --- 1:1
- **Multiple clusters**
  - Multiple clusters can be connected thru N-way FABs
  - Additional east-west capacity between the clusters can be added by:
    - Horizontal scaling (additional EGR's, or additional FABs)
    - Multiple FAB planes (inc dedicated "internal" FABs for east-west traffic only)
    - Turning a FAB into a VC itself

Automation

- **Management complexity:**
  - Large number of devices, links and initial configurations
  - Dynamic environment, asynchronized LEF/TOR installations, image and config updates
  - Device/links failure detection and remediation
- **Automation is the solution**
  - Treat the network with CI/CD principles
  - Device, topology and config modelling abstracted by template and database
  - Integration of inventory/DNS with Zero Touch Provisioning for initial bootstrap and configuration
  - Separation of config intent and config state, complete control loop by state machines
  - Check out our NANOG 68 presentation “Network Automation with State Machines”

Operational Experience

- **Very stable protocol stack, fast convergence**
  - 2012 => 250ms end-to-end convergence
  - 2018 => 125ms end-to-end convergence
  - Even in 2018 some BGP stacks cause micro-blackholes - watch out!
- **Significantly lower hardware failure rate than expected**
- **Easy installation and continuous management**
- **Oversubscription ratios**
- **Buffer management techniques**

Questions?

igor@verizonmedia.com
hyihua@verizonmedia.com

verizon media