Data Center Fabrics – What Really Matters

Ivan Pepelnjak (ip@ioshints.info)
NIL Data Communications

Who is Ivan Pepelnjak (@ioshints)

- Networking engineer since 1985
- Technical director, later Chief Technology Advisor @ NIL Data Communications
- Consultant, blogger (blog.ioshints.info), book and webinar author
- Currently teaching “Scalable Web Application Design” at University of Ljubljana

Focus:

- Large-scale data centers and network virtualization
- Networking solutions for cloud computing
- Scalable application design
- Core IP routing/MPLS, IPv6, VPN

Disclaimers

- This presentation describes real-life data center fabric architecture(s)
- It's not an endorsement or bashing of companies, solutions or products mentioned on the following slides
- It describes *features* not *futures*

Datacenter Networks are in my Way

James Hamilton (Amazon), October 2010

- Can we do any better?
- Will Data Center Fabrics solve the problem?

Why Does It Matter?

Cloud computing: large-scale elastic data centers

Hard to build them using the old tricks

Scale-out apps generate east-west (inter-server) traffic

Existing DC designs focused on north-south (server-to-user) traffic

IaaS requires flexible VM placement and VM mobility

Hard to implement with existing VLAN-based approaches

What Is a Fabric?

Juniper

- Any-to-any non-blocking connectivity
- Low latency and jitter
- No packet drops under congestion
- Linear cost and power scaling
- Support of virtual networks and services
- Modular distributed implementation
- Single logical device

Cisco

- Open (standards-based)
- Secure (isolation of virtual zones)
- Resilient (fault-tolerant, stateless)
- Scalable
- Flexible (incl. auto-provisioning)
- Integrated (compute, network & storage)

Brocade

- Flatter
- Intelligent (auto-discovery)
- Scalable (multi-pathing)
- Efficient (automatic shortest path fwd)
- Simple (single logical entity)

The answer seems to depend on the capabilities of your gear

What Do We Really Need?

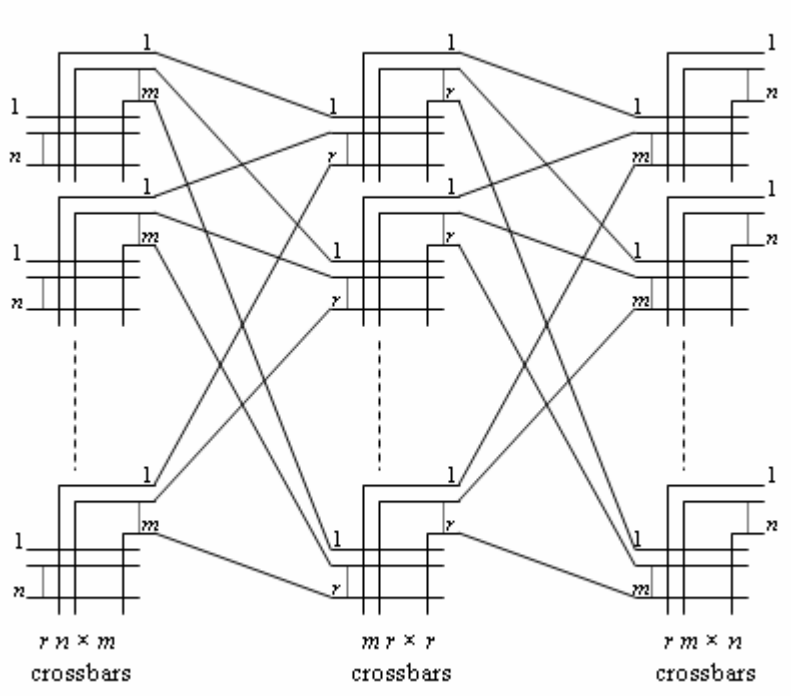
- Equidistant endpoints with non-blocking network core
- Unlimited workload mobility
- Lossless transport (storage and elephant flows)
- Simplified provisioning and management

Questions:

- L2 or L3? Does it really matter?
- Can we do it with existing gear?
- Do we need TRILL/SPB/DCB/FabricPath/QFabric...?

A History Lesson: Clos Networks

- Multistage switching network (Charles Clos, 1953)
- Used to build large voice switches



Data Center usage

- High-bandwidth fabrics from low-latency non-blocking switches

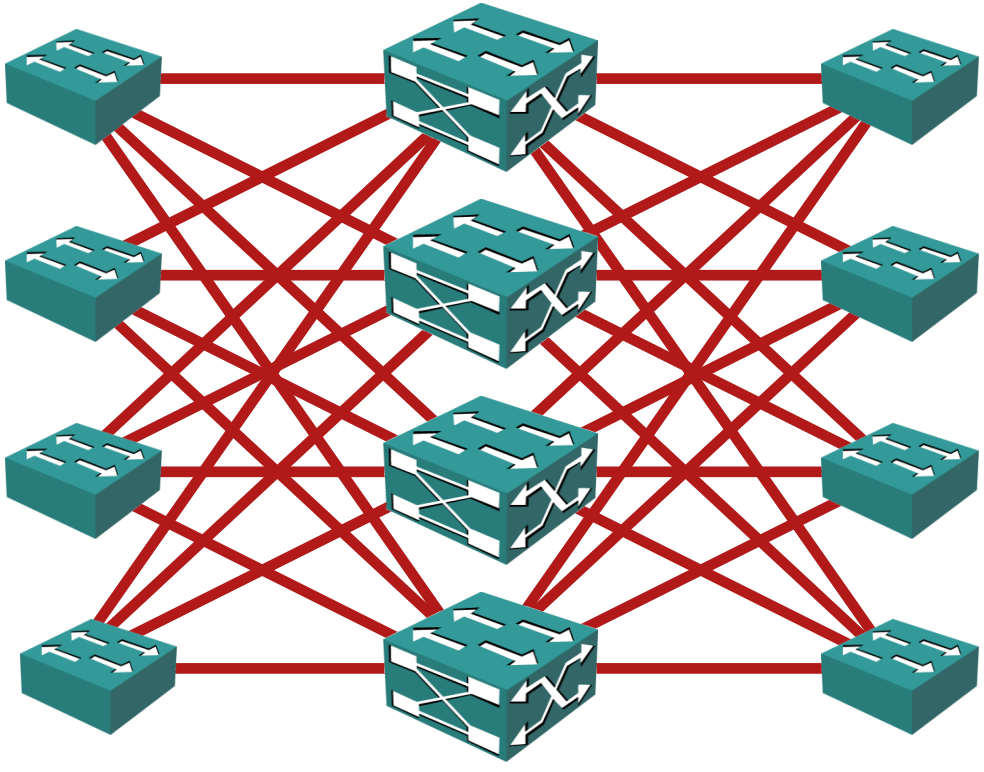
Caveats

- Non-blocking switching in the core
- Oversubscription at the edge (usual approach)
- Not a true non-blocking architecture, relies on statistics

3-Stage Clos Network: Leaf & Spine Architecture

Generic architecture:

- Every leaf node connected to every spine node
- Relies on ECMP on every hop
- L3 fabric or large-scale bridging (TRILL or SPB)



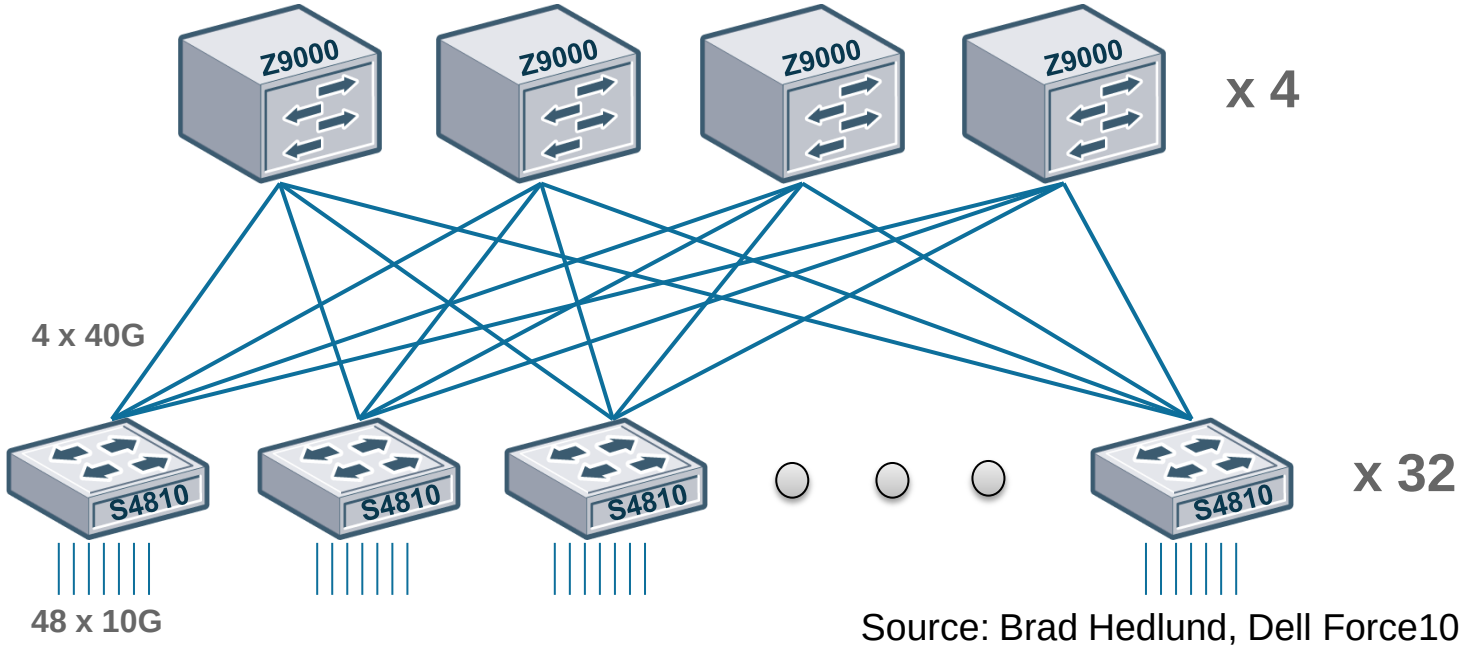
Forwarding decisions:

a) Full lookup at every hop

b) Default route @ leaf switches pointing to spine

c) Single L2/L3 lookup, label switching throughout the fabric (MPLS or QFabric)

Sample Leaf & Spine Designs



Source: Brad Hedlund, Dell Force10

Force10:

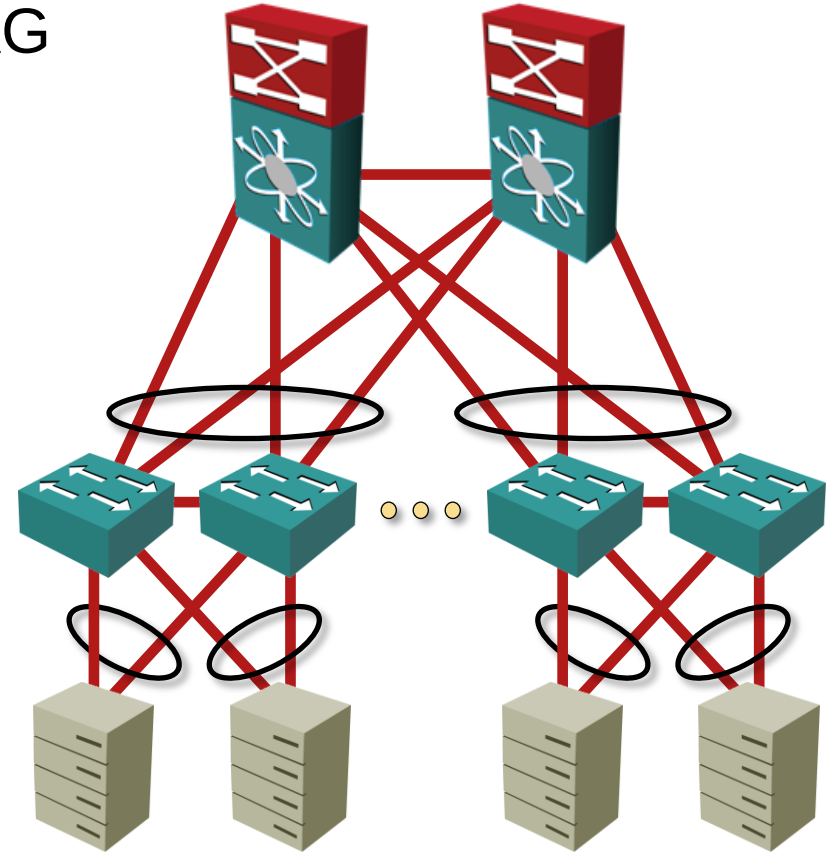
- 4 x Z9000 (32 x 40GE)
- 32 x S4810 (48 x 10GE, 4 x 40GE)
- 4-way ECMP, 40GE uplinks
- ~1500 10GE ports @ 3:1 oversubscription

Arista:

- 16 x 7508 (384 x 10GE)
- 384 x 7050S-64 (48 x 10GE, 4 x 40GE)
- 16-way ECMP, 40GE → 4 x 10GE uplinks
- ~18000 10GE ports @ 3:1 oversubscription

Layer-2 Clos Network Without Shortest Path Bridging

Trick: use large core switches and MLAG



Sample design (Arista):

- 2 x 7508 (768 10GE ports)
- 45 x 7050S-64
- 1980 10GE server ports
- Oversubscription 2,75:1

Does it make sense?

- Single failure domain
- Broadcasts will kill you
- Check L2/L3/ARP table sizes in the core switch

Layer-2 Reality Checks

Use cases:

- Virtual network segments
- VM mobility

Flooding will kill you:

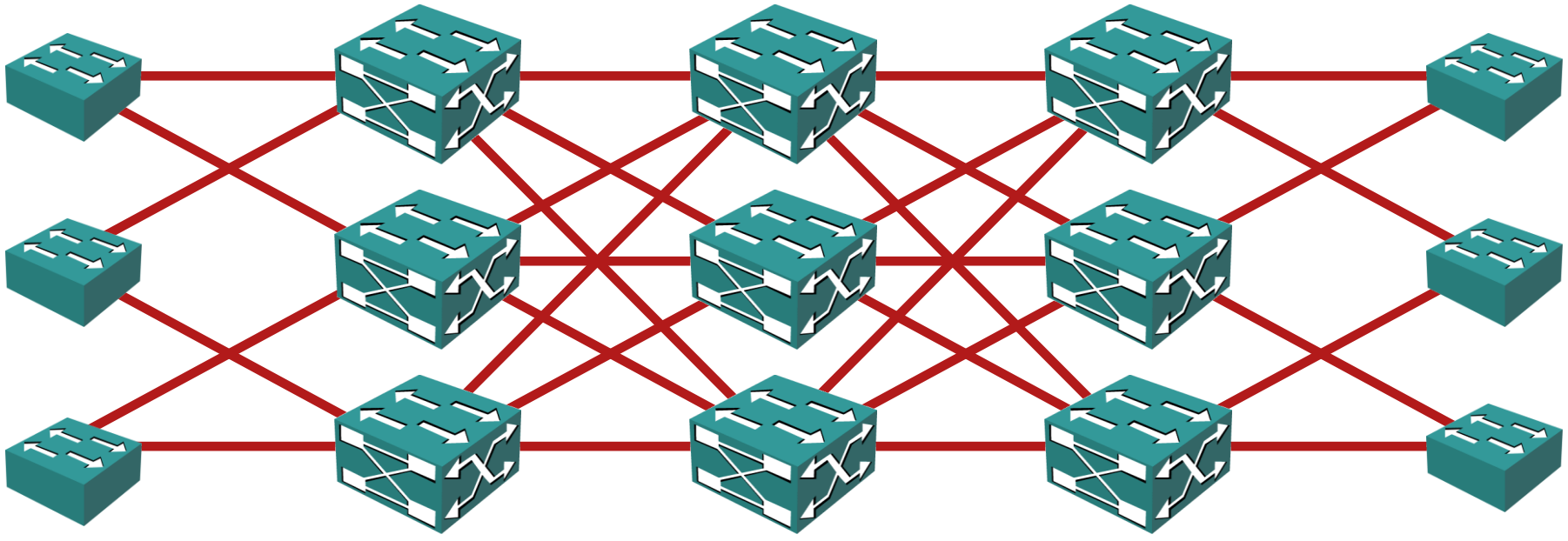
- Typical design: every VLAN on every access port
- Hypervisors put NICs into promiscuous mode

TRILL (RFC 5556): “around 1,000 end-hosts in a single bridged LAN of 100 bridges”

Other reality checks:

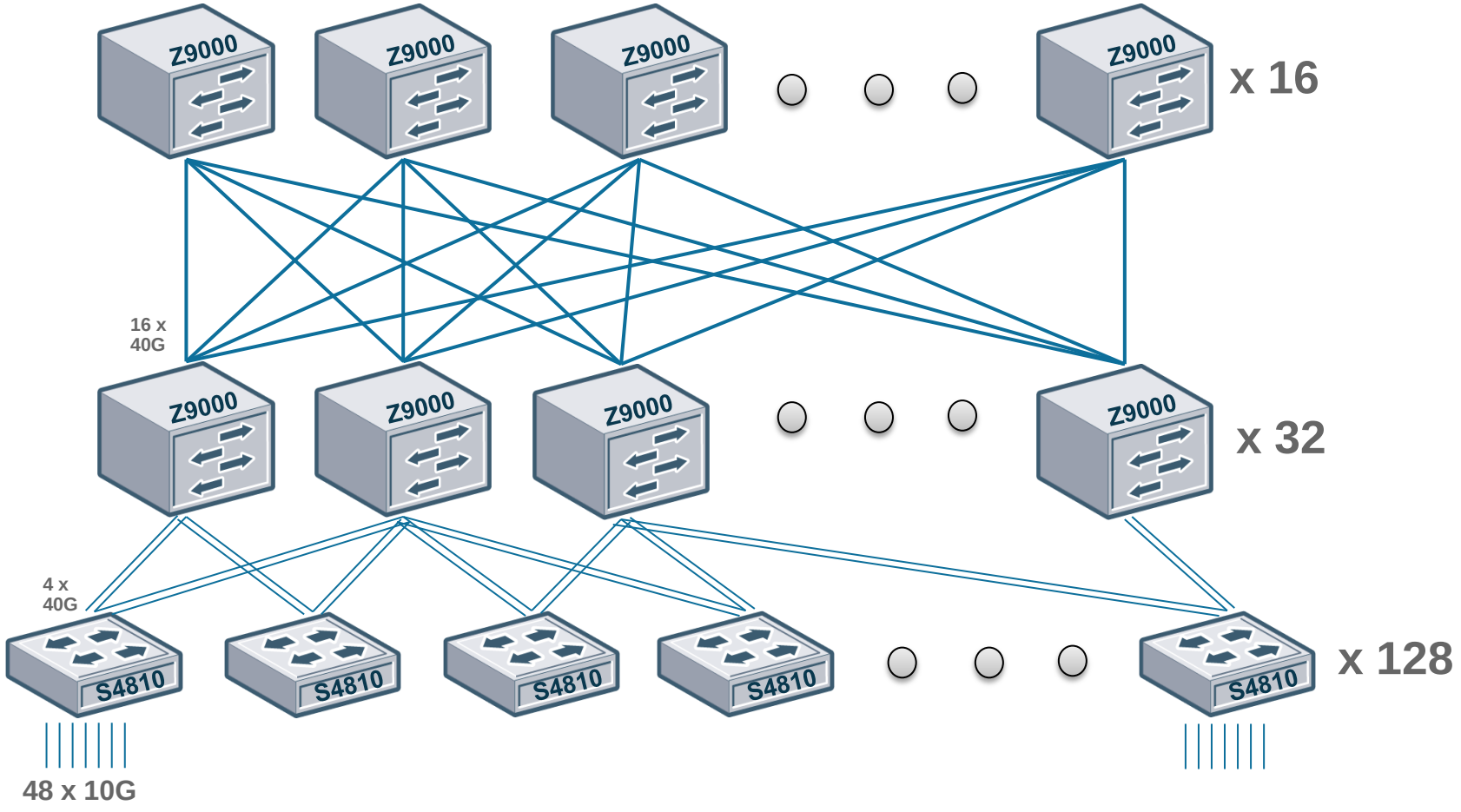
- VMware HA cluster can have up to 32 hosts
- VMware vDS can span 350 hosts (no vMotion between vDS)

Beyond Leaf & Spine



- Leaf nodes connected to a few second-stage nodes
- Full mesh in the core
- Core stages perform ECMP traffic distribution
- 3-stage Clos fabric used internally by high-bandwidth switches

Sample Folded Clos Network



6144 10GE ports @ 3:1 oversubscription

128 x S48100 + 48 x Z9000 (Dell Force10)

Source: Brad Hedlund, Dell Force10

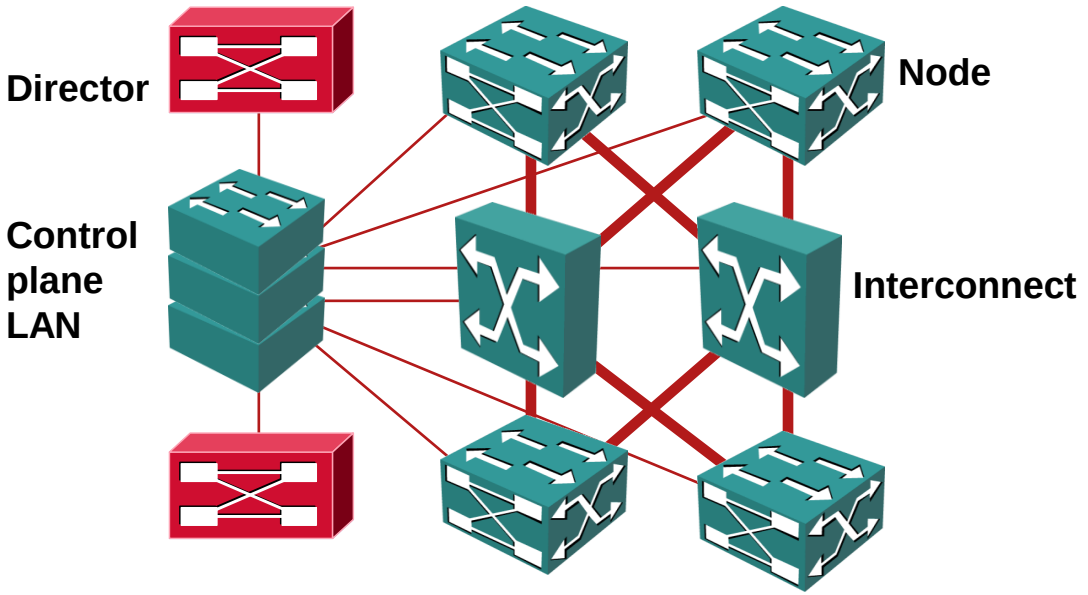
QFabric: Internal+External Clos Fabric

Director

- Compute resources, runs multiple *Routing Engines*

Interconnect

- High-speed 3-stage 10Tbps Clos network
- Up to four interconnects per QFabric



Node (QFX3500)

- Layer2/3 packet forwarding
- 3:1 oversubscription
- 160 Gbps (4 * 40 Gbps) between QFX3500 and QF/Interconnect
- Single (ingress node) packet lookup (sounds like MPLS/VPN) – 5 µs across the QFabric

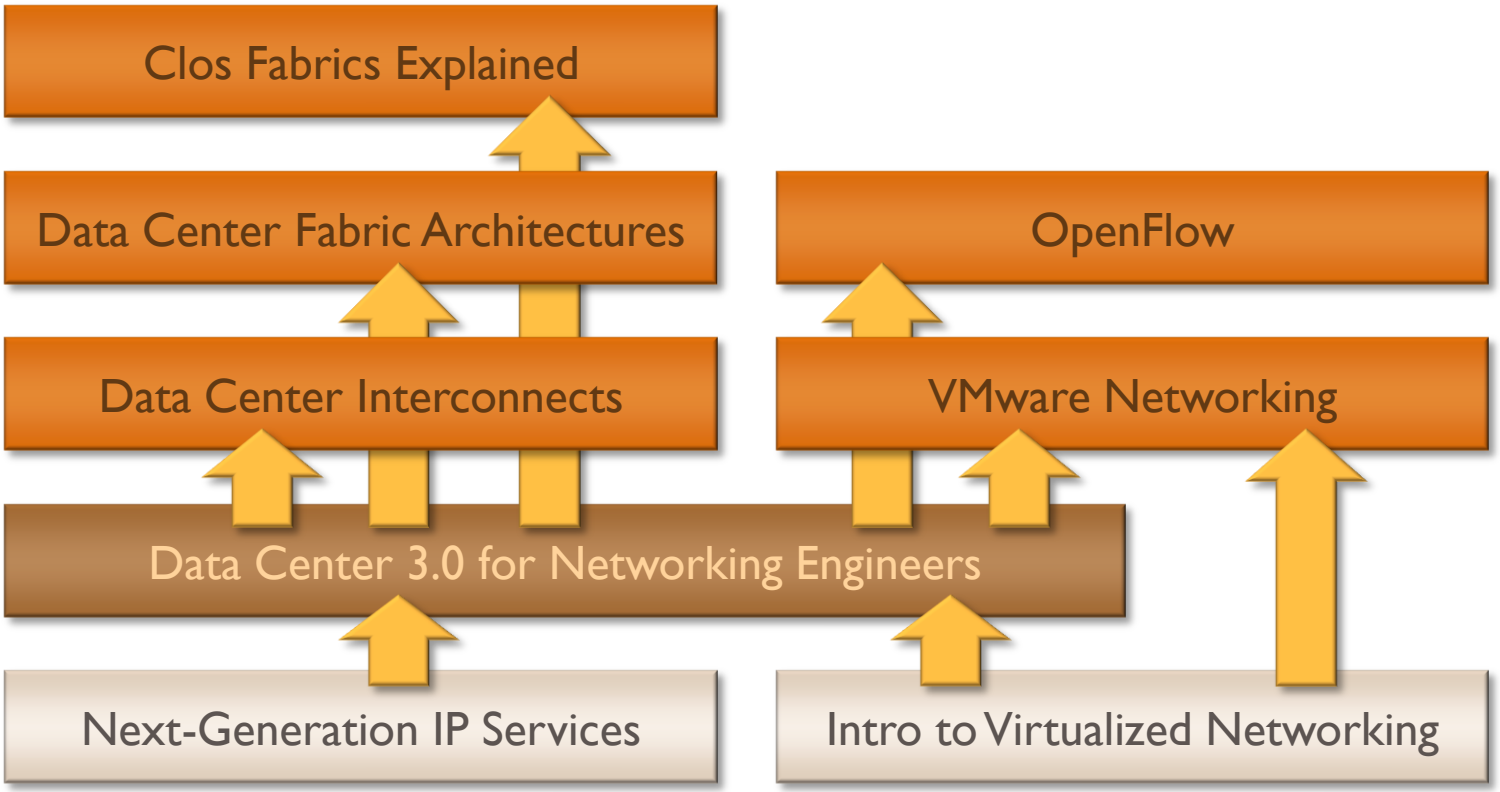
QF/Nodes and QF/Interconnects form a Clos fabric

Single configuration and management entity

Conclusions

- Large-scale DC fabrics with equidistant endpoints: no problem
- Use tested architectures (Clos) and technologies (IP)
- Keep your L2 domains small
- Use MAC-over-IP to build scalable virtual networks
- Not much need for emerging L2 technologies (TRILL, SPB ...)

Reference: Data Center Webinars



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

More information @ http://www.ipspace.net/Webinars

Reference: Blogs and Podcasts

- Packet Pushers Podcast & blog (packetpushers.net)
- Fragmentation Needed (Chris Marget)
- My Etherealmind (Greg Ferro)
- The Data Center Overlords (Tony Bourke)
- Majornetwork (Markku Leiniö)
- Telecom Occasionally (Dmitri Kalintsev)
- The Networking Nerd (Tom Hollingsworth)
- Static NAT (Josh O'Brien)
- BradHedlund.com (Brad Hedlund, Dell Force 10)
- NetworkJanitor.net (Kurt Bales)
- The Lone Sysadmin (Bob Plankers)
- High Scalability Blog (Todd Hoff)
- Twilight in the Valley of the Nerds (Brad Casemore)
- blog.ioshints.info & ipspace.net (yours truly)

Questions?