Embracing Open:
The AMS-IX Journey to Open Networking

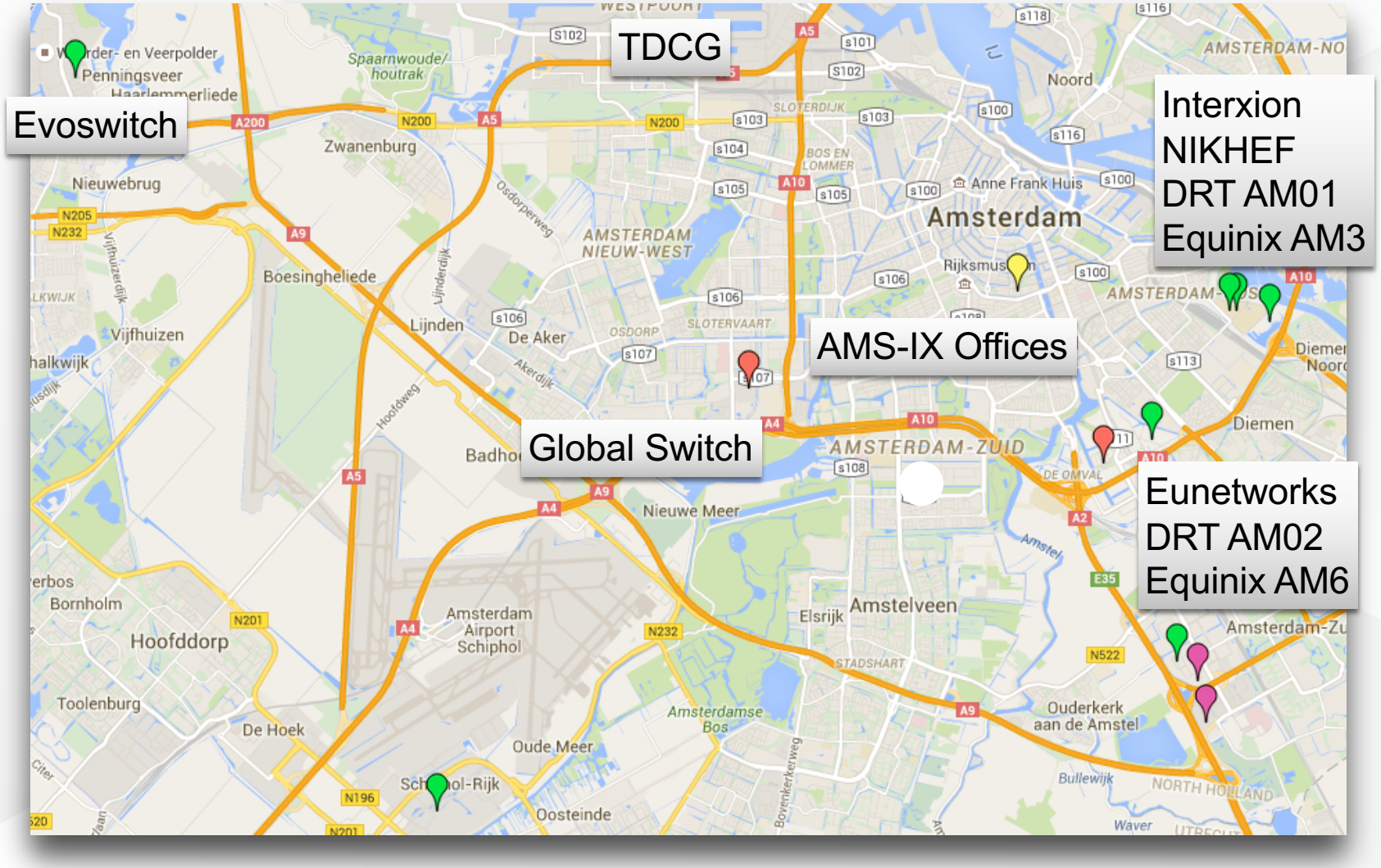
Bart Myszkowski
Stavros Konstantaras

vGRNOG 10
24-10-2020

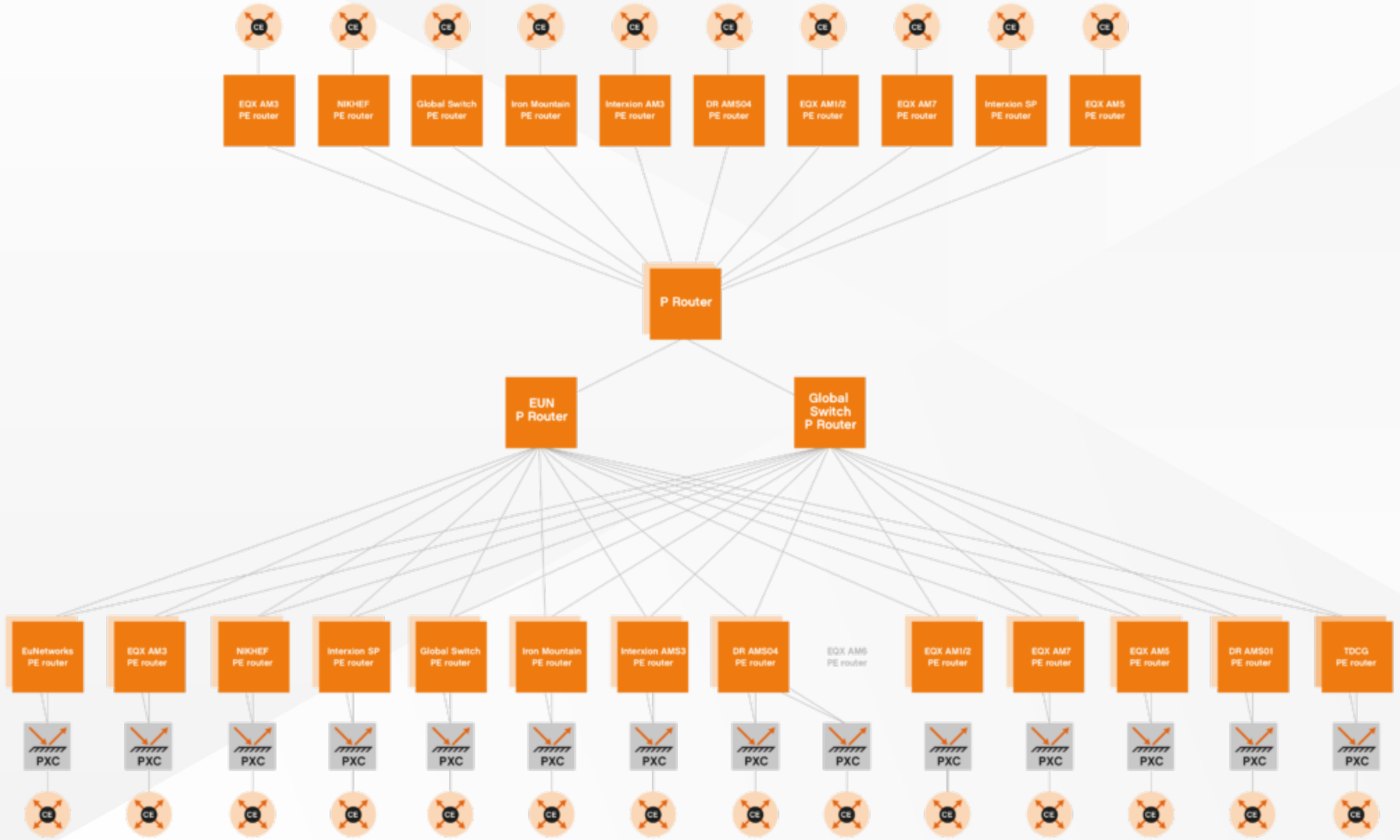
Embracing Open Networking Outline

- AMS-IX introduction
- Network overview and “before” state
- Upgrade motivations, options
- Why we chose open networking
- Open network fabric technology
- Network “after” state
- Experience and lessons learned

AMS-IX in Amsterdam:



AMS-IX Amsterdam Platform



<= Customer routers

<= Low Speed access

<= Core / Spine

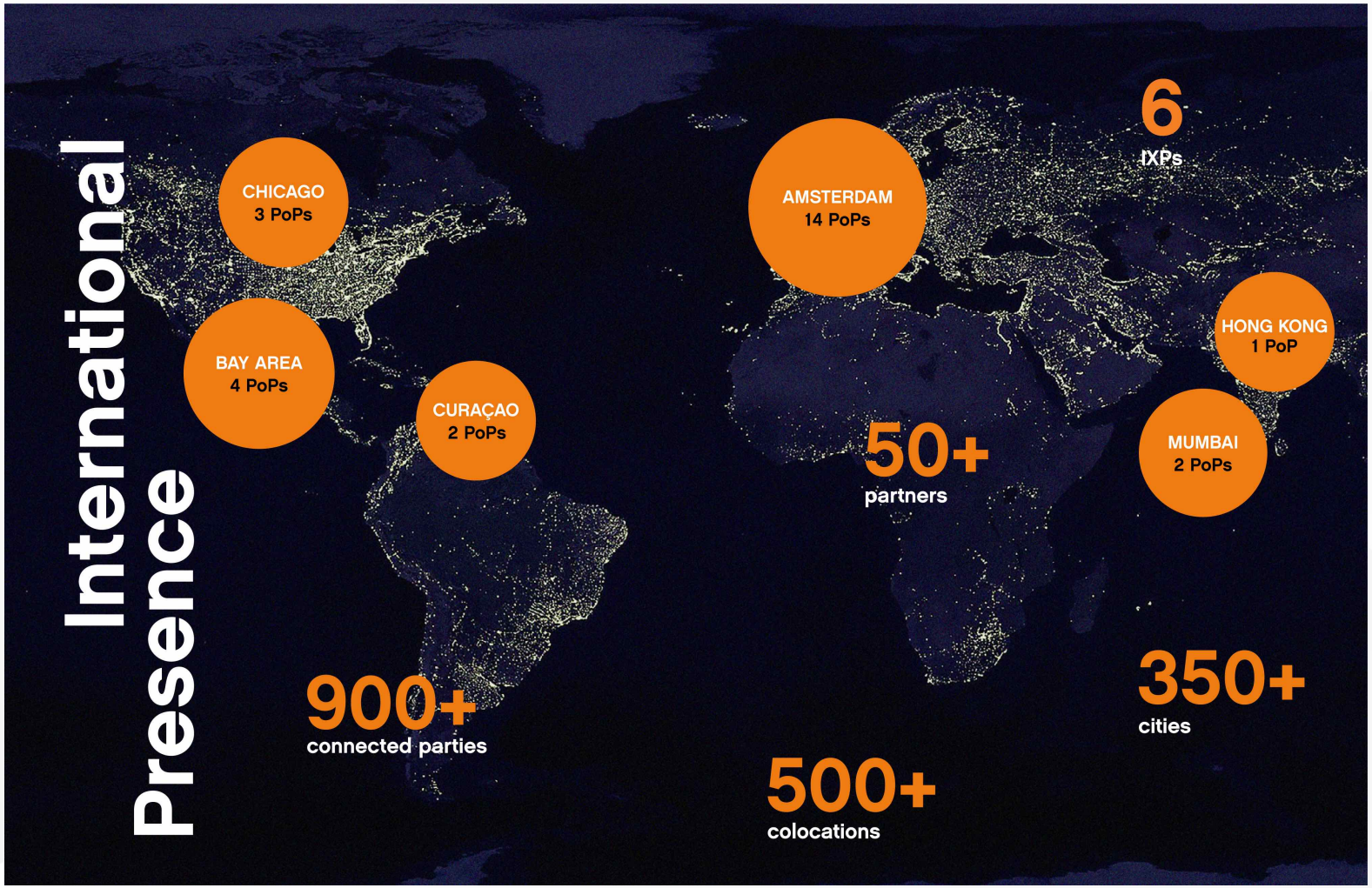
<= High Speed access

<= Optical access

<= Customer routers

https://www.ams-ix.net/ams/documentation/ams-ix-topology

AMS-IX Around the world

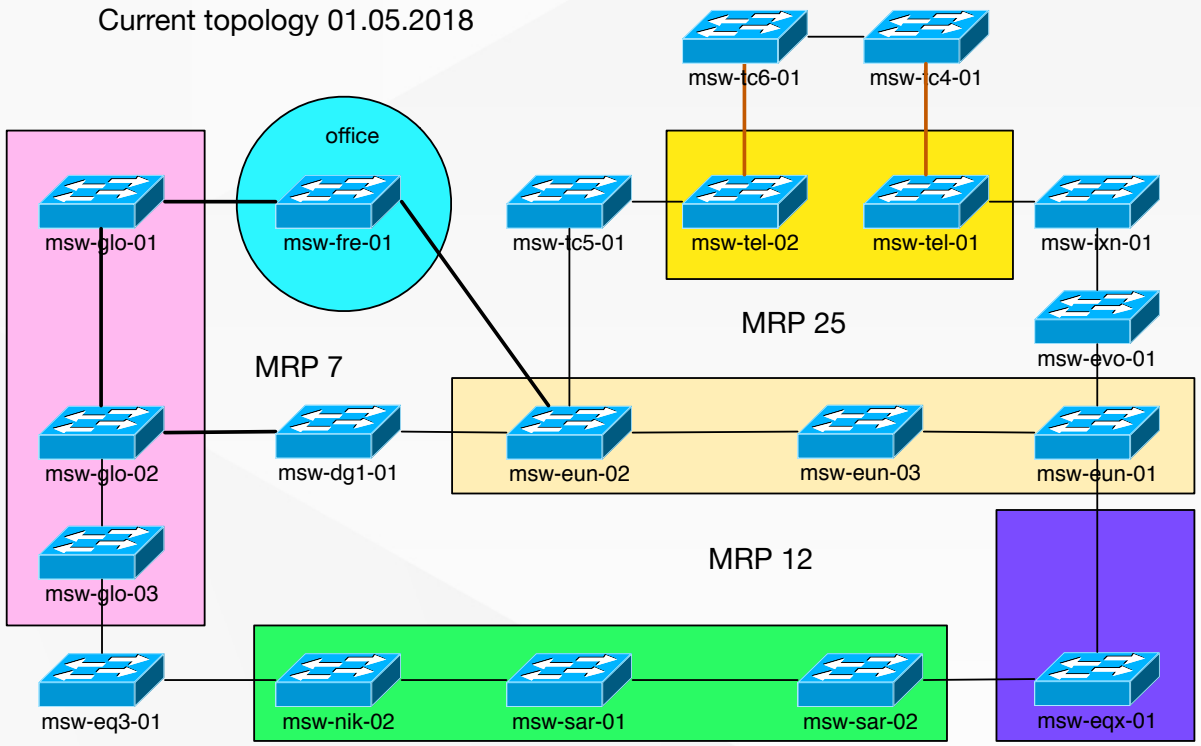


AMS-IX management network

- Gives us access to our production equipment (SLX, MLX, DWDMs, PXCs, TS etc.)
- Servers, load-balancers, firewalls, PTP devices, NIDs
- VM/SAN replication
- Monitoring system relies on management network
- Access to the Internet from office/sites

“Before” network set-up

- **Scale**
  - 22 switches, 15 geographically separate locations, 463 ports in use in NL
  - 10 switches on remote locations (CHI, BAY, HK, CW, NY)
- **Equipment in use:**
  - Foundry/Brocade FCX, FES, FGS, ICX (Ruckus)
- **Topology/protocol:**
  - Ring topology: 3 rings connected by 17 dark fibers
  - MRP (metro ring protocol) L2 resilience protocol



“Before” network issues

- Easy to create a loop/outage
- Inefficient link utilization, some bandwidth bottlenecks
- Ring isolation in case of double fiber cut or issue with MRP
- Different switches with different software versions, challenging to manage
- Some of the switches will be end-of-life soon
- Fiber cost: Management network (17 dark fibers) was completely separate from production network (30 dark fibers + DWDM)

Switching upgrade goals

- Make environment homogeneous (same HW/SW)
- Higher speed for VM moving, NAS/SAN cluster replication
- More redundant topology
- Easier management
- Better visibility

Where to go?

Technology? Pure L2, TRILL, eVPN, VxLAN etc.

Brand? Cisco, Juniper, Brocade, Arista, Huawei etc.

Hardware? Branded or baremetal

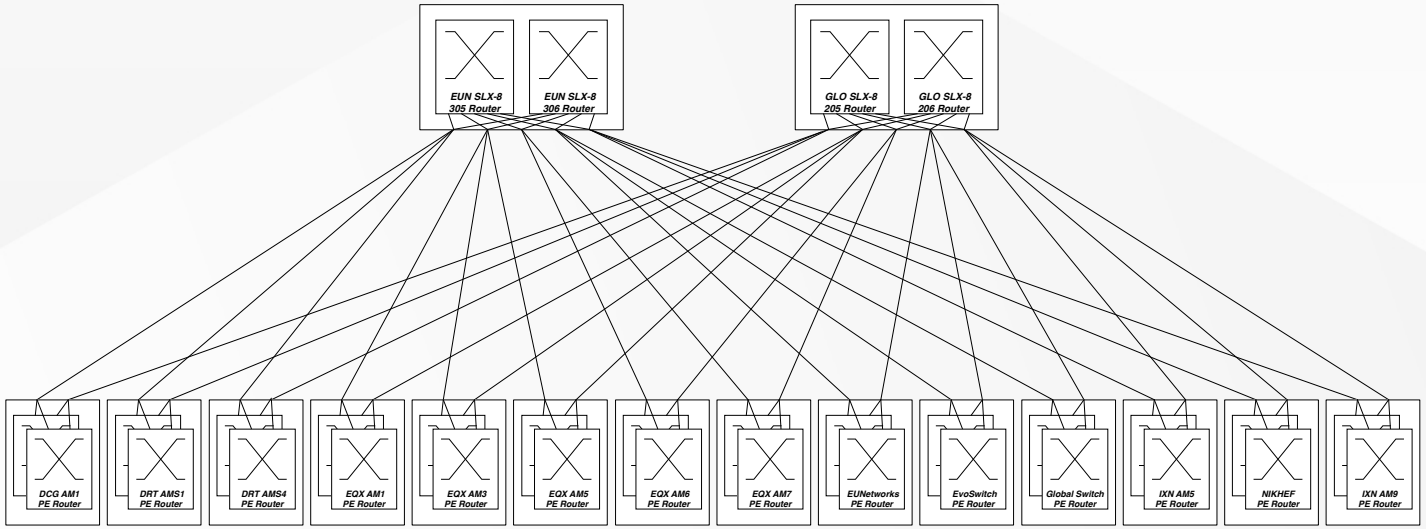
Software? Open source or branded

Decision #1

Re-design the management network with spine-leaf topology
- **A design that we know and trust**

Fiber connectivity solution: re-use current production DWDM set-up

- Use existing DWDM muxes on production fibers to support new channels/wavelengths to connect the management network
- Eliminate rings, move to fully redundant leaf-spine topology
- Eliminate separate management network fibers, reduce cost



Advantages of open network: bare metal + software

- Decoupling hardware from software on network equipment (same as we have on servers now)
- Ability to change OS or hardware any point of time (like we do with Linux Debian → CentOS)
- New players appeared on the market with newest software features (Pluribus Networks, Cumulus, BigSwitch, IPinfusion etc.)
- Ability to use free OPX (openswitch.net) project

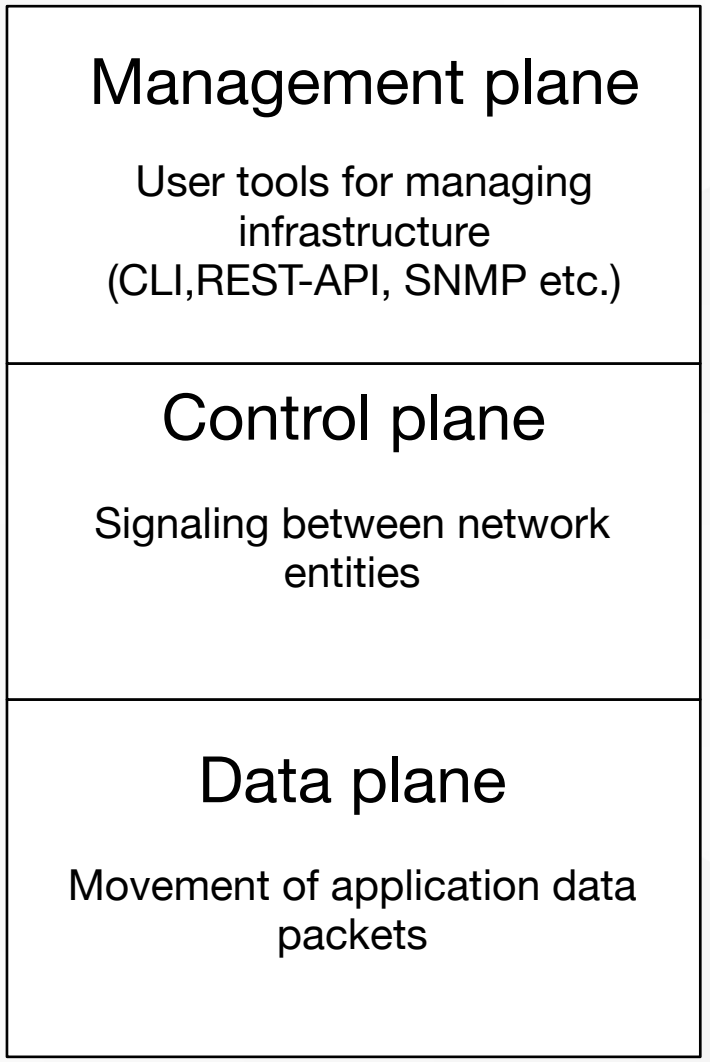
Other decision considerations for open network

- **HW/SW maturity**
  - White box HW standardized in OCP, used for years in hyperscale DCs
  - NOS SW also in wide use, supports all the L2/L3 protocols and features that we need
- **Support**
  - Larger vendors now offering open networking with full support
- **Manageability**
  - Newer SDN approach provide better manageability than traditional systems

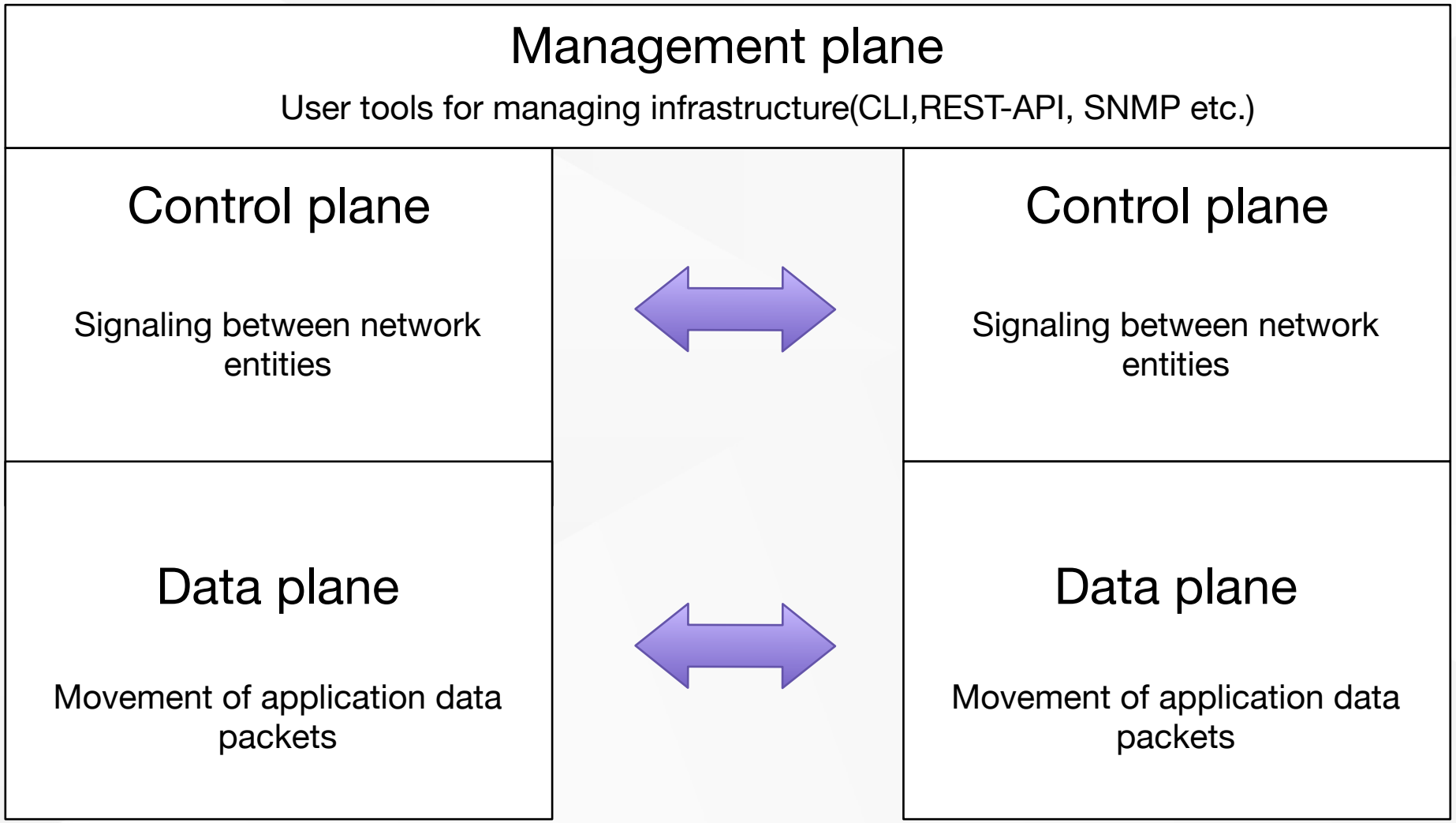
Decision #2

We selected Dell for HW switches and Pluribus for the NOS

Classic switch design

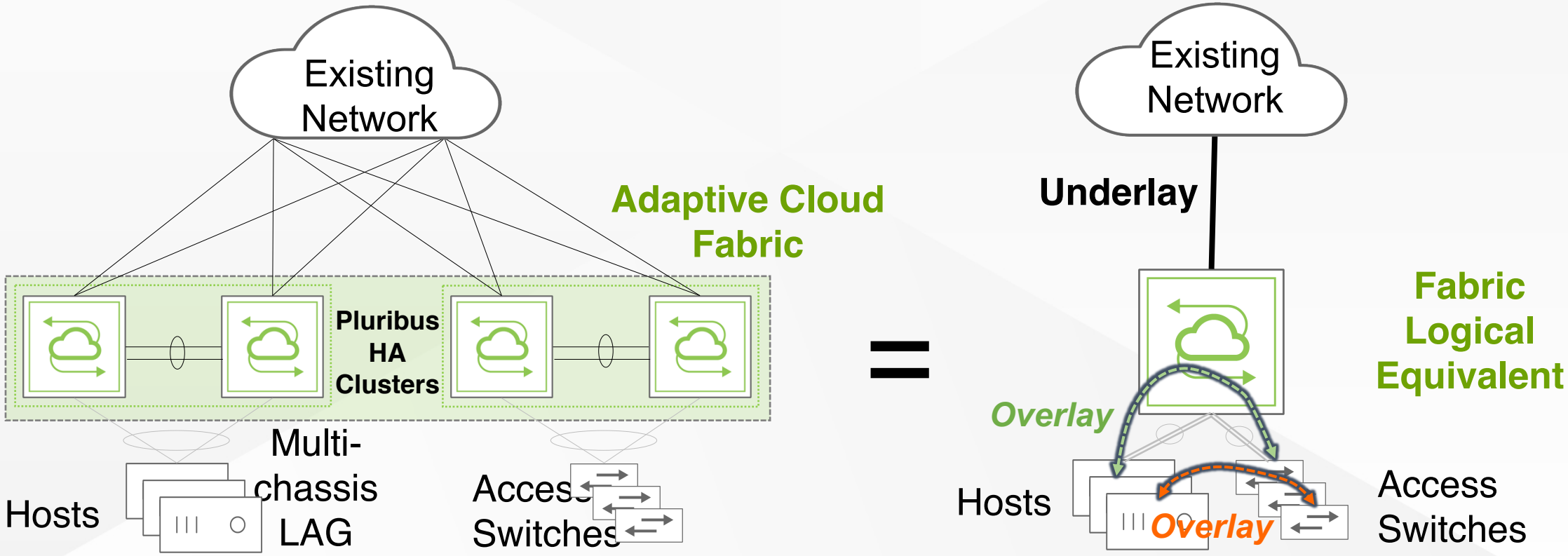


Pluribus distributed SDN fabric concept



Fabric logical view

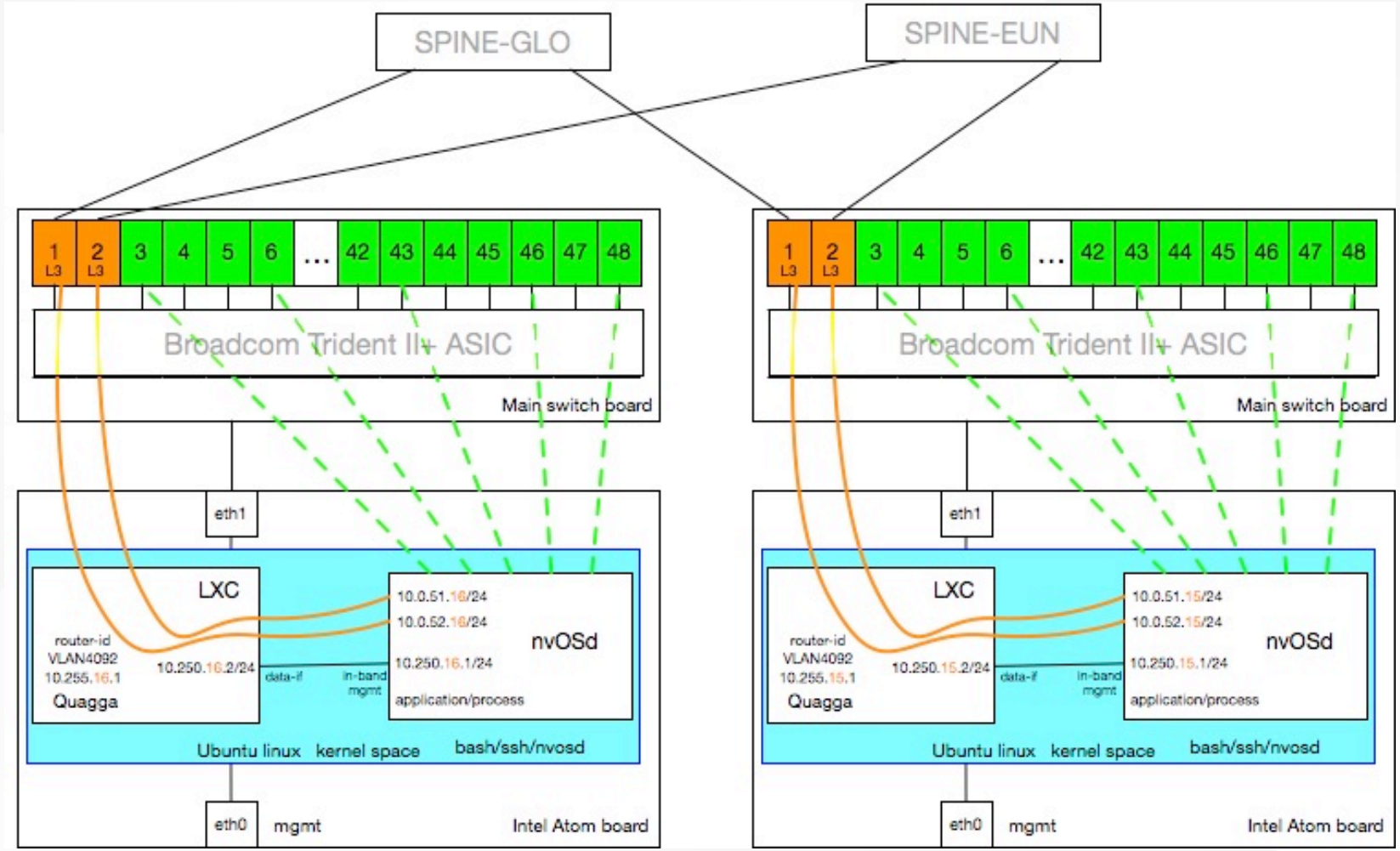
- Multiple geographically distributed sites act as one programmable entity
- Deploy network services as “fabric object” which updates all switches in fabric



Geographically dispersed sites

Open switch configuration

- **Switching ASIC connects at high speed to CPU (e.g. Intel)**
- **L2/L3 protocols run in Linux containers**

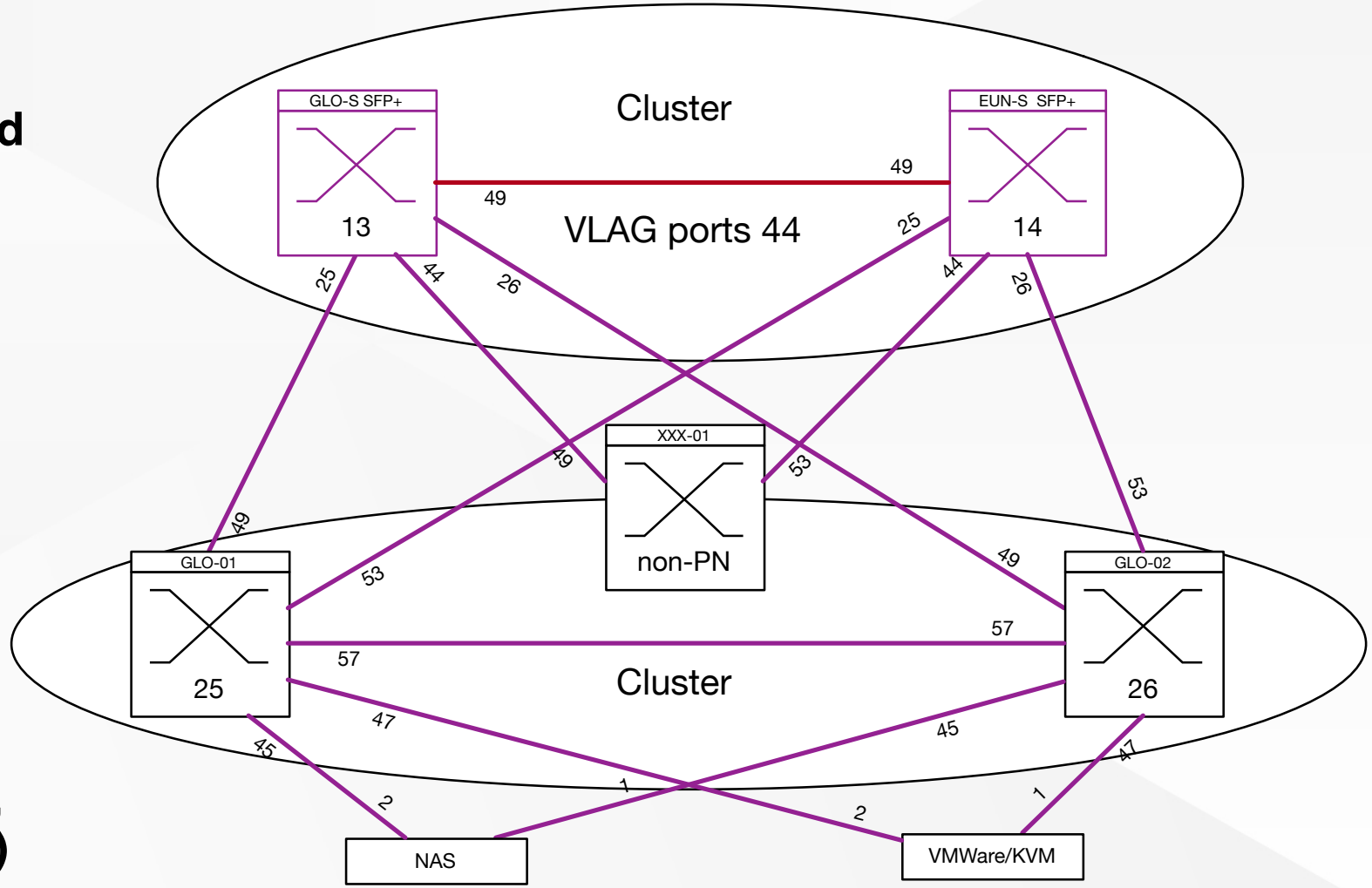


Building a fabric with VxLAN

- VxLAN enables L2 network over L3 underlay (with OSPF)
- Use all available links
- Traffic is load balanced using ECMP over all backbone links
- MC-LAG for critical servers/NAS
- Loop-free
- Enables network segmentation for application isolation

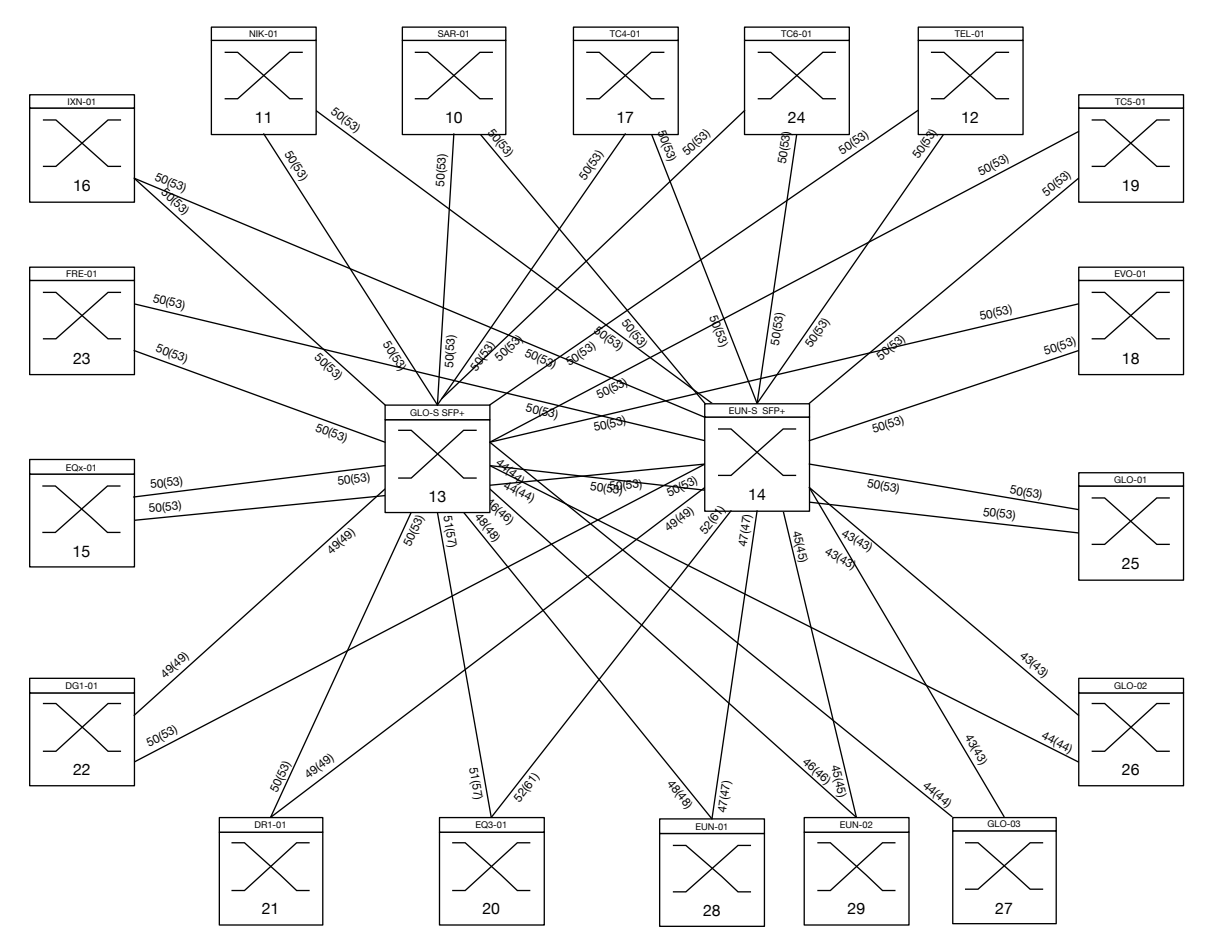
MC-LAG redundant connections

- **Two switches configured as a cluster support redundant connections to avoid downtime during maintenance or device/link failure**
- **Spine cluster enables redundant leaf connections**
- **Leaf cluster used where needed for critical infrastructure (e.g. NAS, production web servers)**



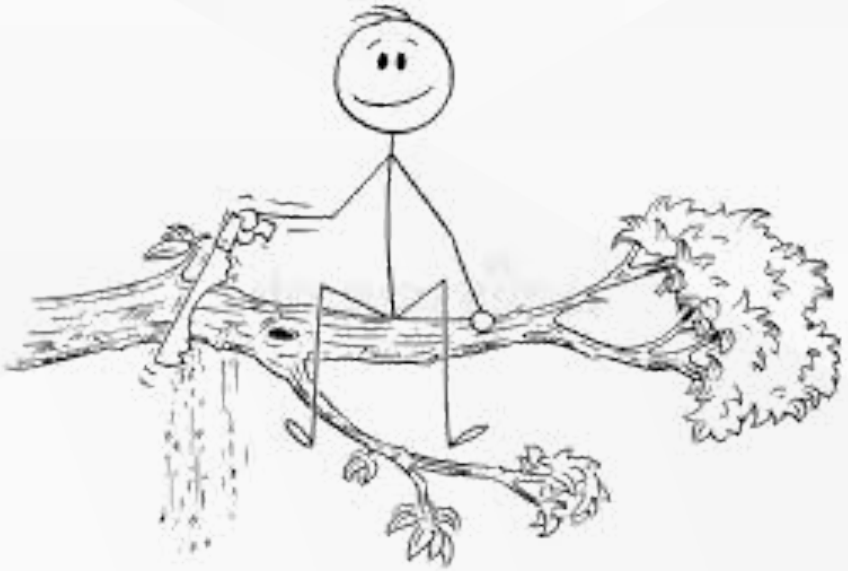
New AMS-IX management network (“after”)

- **Geographically distributed fabric built on standard OSPF underlay**
- **Loop-free ECMP/BFD for efficient multi-pathing**
- **No STP, fast re-convergence**
- **No controller = no split brain, resilient**
- **vLAG for critical servers & NAS**
- **Improved visibility**

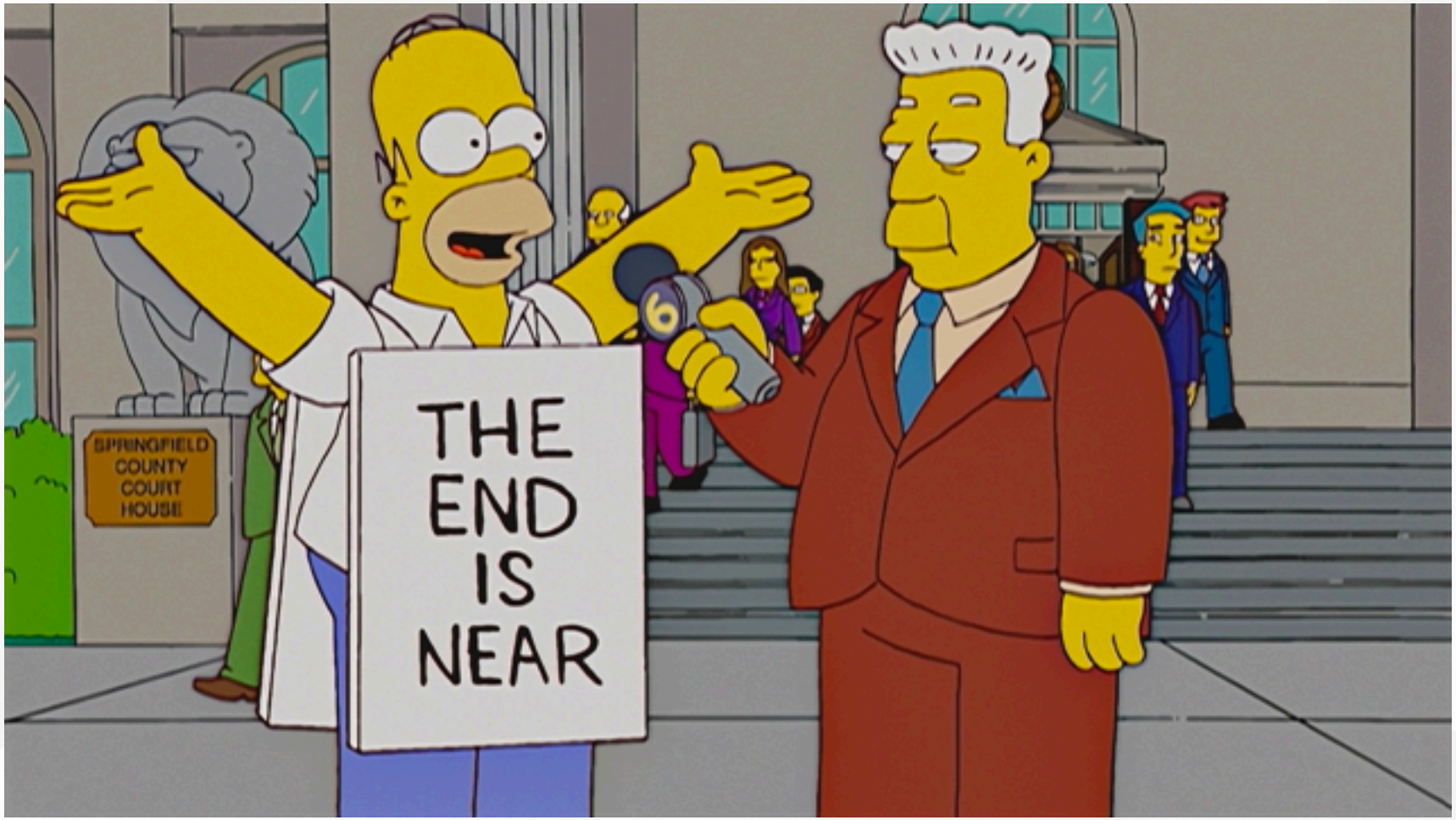


The unified fabric is amazing!!!

But it’s also (very) dangerous! ☠️



Apocalypse?!? Management unreachable?



Software stack broken?!?

What if (for any reason like upgrade, fail, bug, ...) the software stack on the management switch is not working and we need to manage???

We have a backup plan before start to run!

Normal operation

- All management traffic flow normal
- The green line is a passive channel directly connected to the terminal server, a backdoor segment
- Very useful for maintenance window, firmware upgrades, critical events on the management network, …

Management switch failure on Leaf B

- If the management switch fail, the site still working, but you will lose all the management capability
- On this kind of event, we can reach the terminal server via passive channel and connect to the devices via console for troubleshooting

Spines failure

- If the Spines fail, normally the sites will be isolated
- On this kind of event, we can reach the terminal server via passive channel and connect to the devices via console for troubleshooting

Experience to date

- **Best result of adopting new open network approach with fabric concept = simpler management**
  - Whole network visibility and monitoring
  - Automation / reduced manual operations steps, e.g. one step to configure new L2VPN across multiple sites
  - Segmentation / isolation of different applications is built in, managed at fabric level
- **Lower HW costs also a plus**

Thank you!

Questions, suggestions or remarks?

Fabric failure

- If the fabric fail (fat fingers, upgrade or real failure), the site will be isolated
- On this kind of event, we can reach the terminal server via passive channel and connect to the devices via console for troubleshooting
- You're such a lucky guy, I recommend buying lots of lucky charms!!!