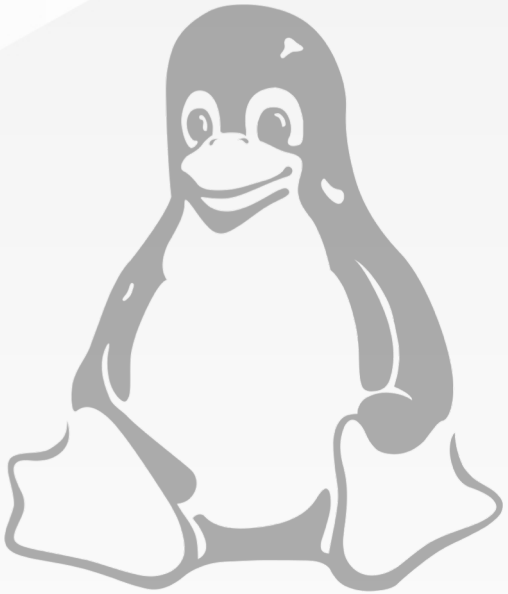


Embracing Open:
The AMS-IX Journey to Open Networking

Maxx Cherevko
Tiago Felipe Gonçalves
IX Fórum 13
São Paulo, BR
10-11 December 2019

Embracing Open Networking
Outline

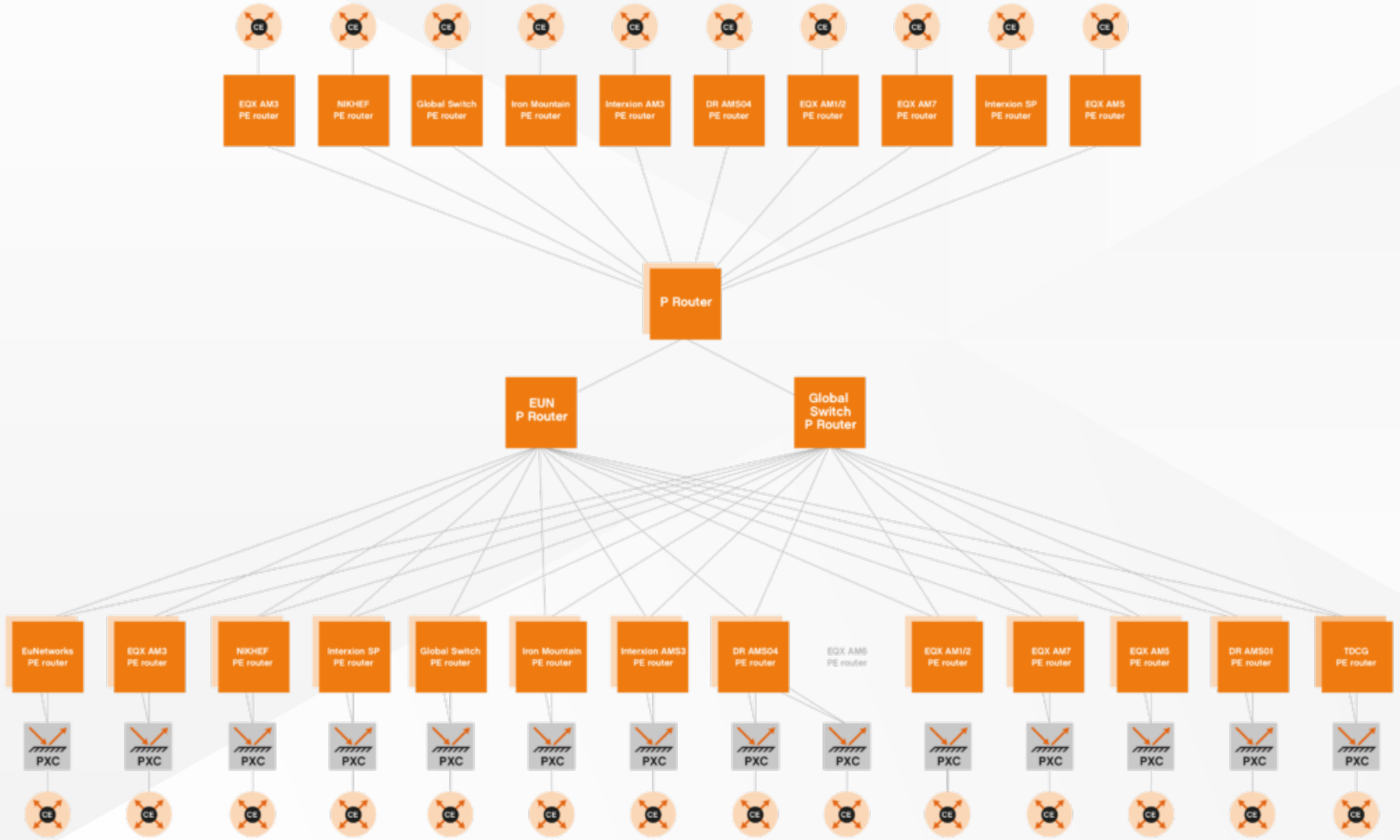
- AMS-IX introduction
- Network overview and “before” state
- Upgrade motivations and options
- Why we chose open networking
- Open network fabric technology
- Network “after” state
- Experience and lessons learned

AMS-IX in Amsterdam:

https://www.ams-ix.net/ams/colocations

1 - The Datacenter Group
2 - Digital Realty AMS01
3 - Digital Realty AMS04
4 - Equinix AM1/2
5 - Equinix AM3
6 - Equinix AM5
7 - Equinix AM6
8 - Equinix AM7
9 - EuNetworks
10 - Iron Mountain
11 - Global Switch
12 - Interxion
13 - Interxion
14 – Nikhef

AMS-IX Amsterdam Platform



<= Customer routers

<= Low Speed access

<= Core / Spine

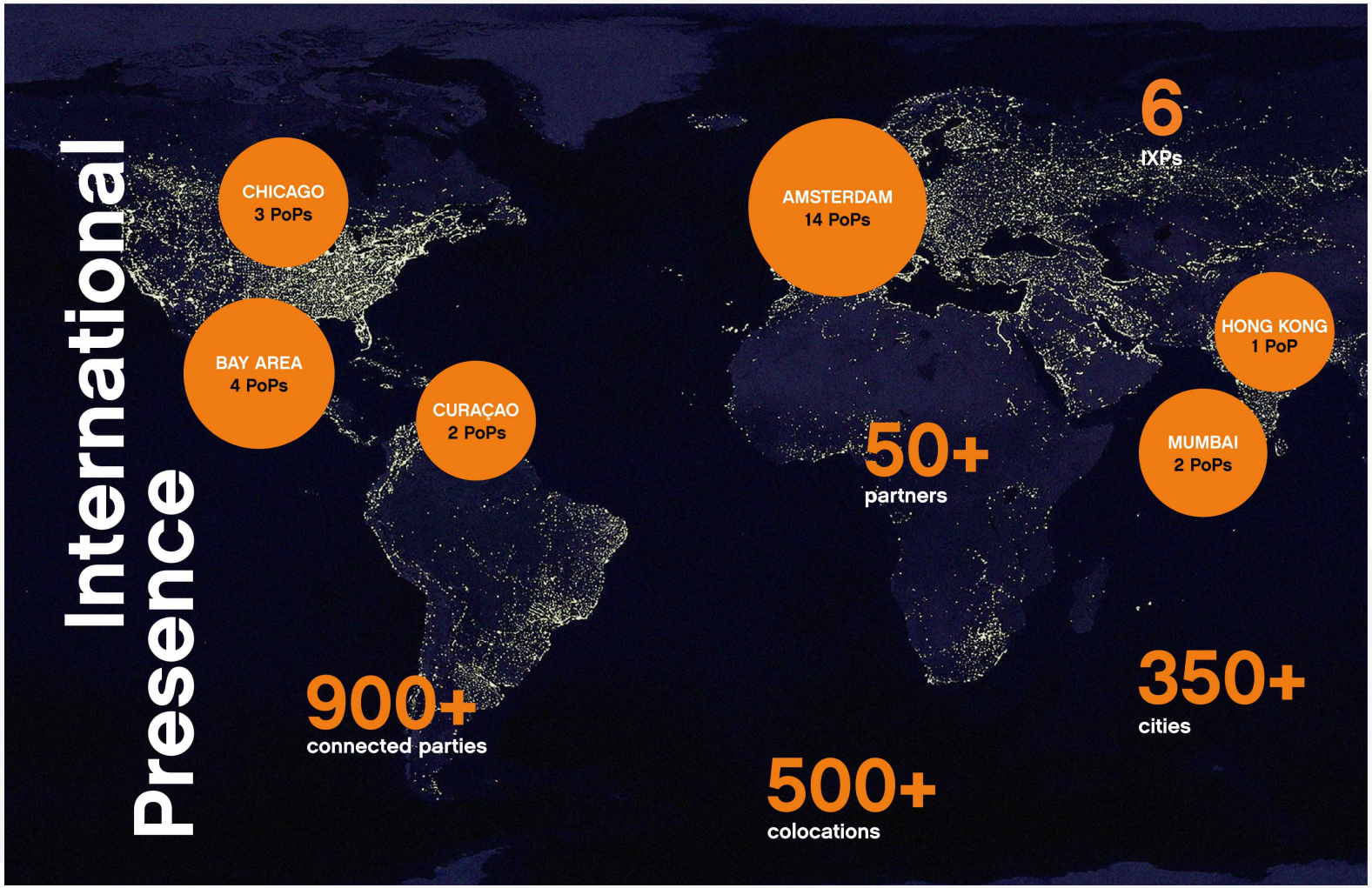
<= High Speed access

<= Optical access

<= Customer routers

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

AMS-IX Around the world



We are present globally with AMS-IX Internet Exchanges and Internet Exchanges powered and managed by AMS-IX trough IXaaS – IX-as-a-Service that's integrated with IX-API, where our customers can use the standard API to manage their links.

If you need more information, please check:
- https://www.ams-ix.net/ams/service/ix-as-a-service
- https://ix-api.net/

AMS-IX management network

- The management network purpose is to have a different network segment(physical or logical), isolating the traffic from the production platform.

- **A common practice is to use in-band management, to reduce cost!!! Please, don't!**

AMS-IX management network

- Since the beginning, we decided to use an OOB (out-of-band) management, because it needs to be reliable, especially during crisis moments!

- Our management network nowadays is being used to:

- Gives us access to our production equipment (SLX, MLX, DWDMs, PXCs, TS etc.), Servers, load-balancers, firewalls, PTP devices, Terminal Servers, NIDs, …

- VM/SAN replication

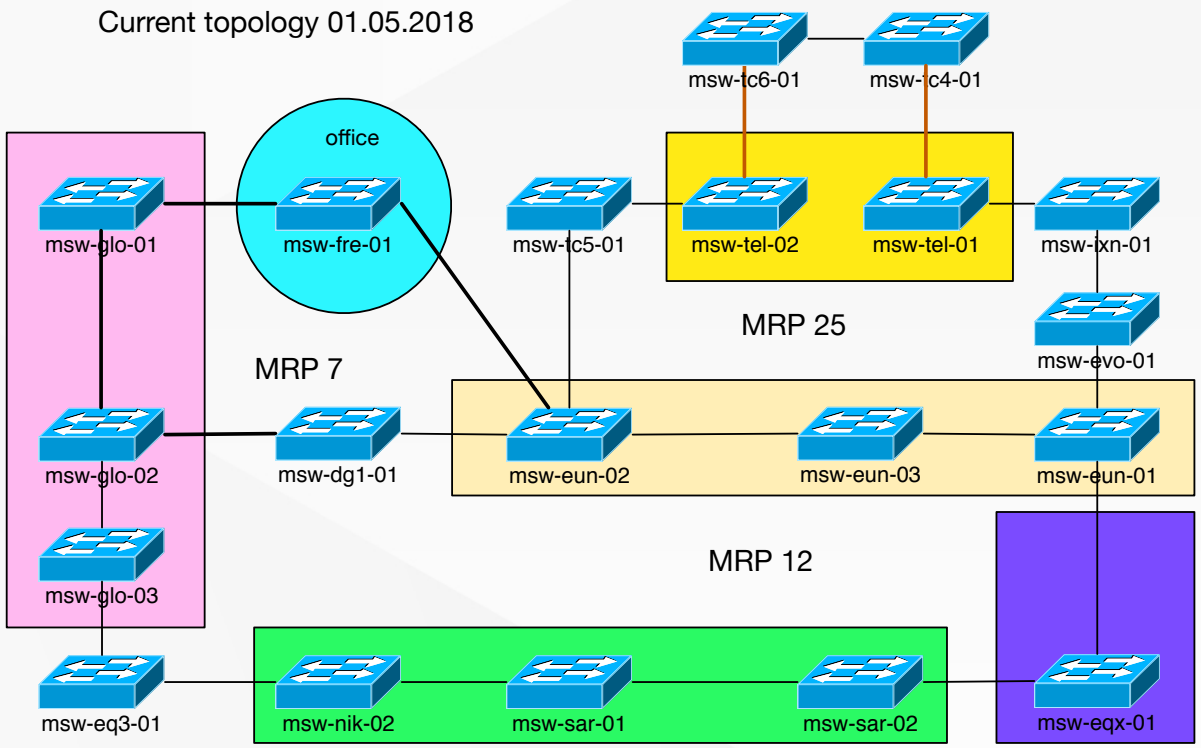
- NMS/Monitoring system relies on management network

- Internet access for office/sites

- Transport any kind of critical/confidential traffic

“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

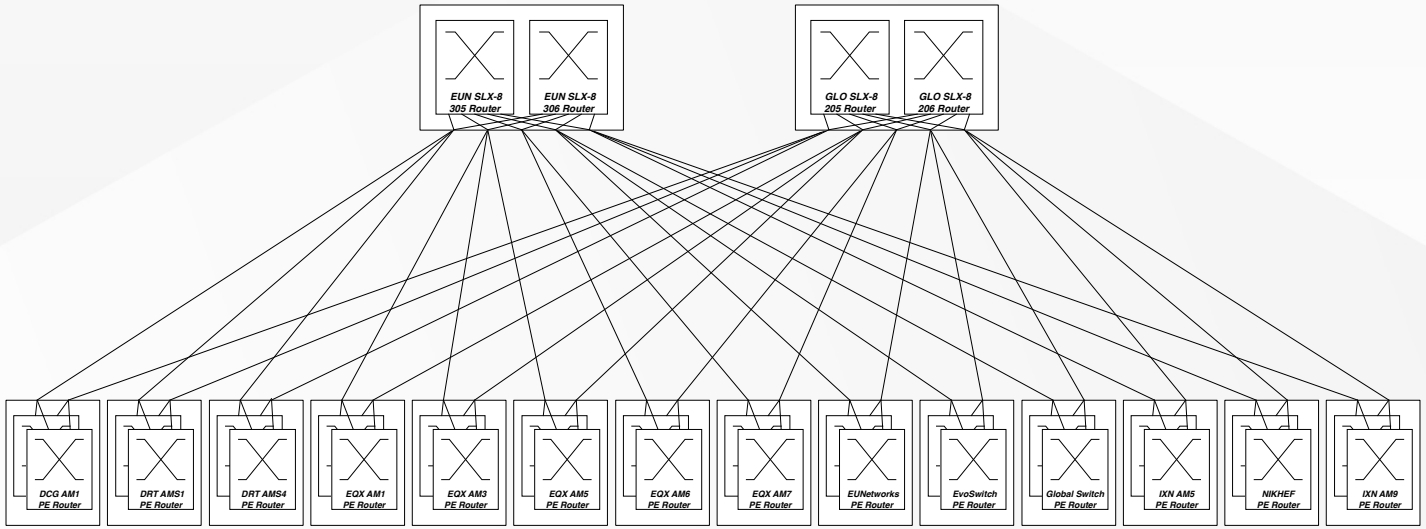
- Easy to create a loop/outage
- Inefficient link utilization, and bandwidth bottlenecks
- Ring isolation in case of double fiber cut or issue with MRP
- Different switches with different software versions, challenging to manage
- Some of the switches would be end-of-life
- Fiber cost: Management network completely separate from production network

Switching upgrade goals

- Make environment homogeneous (same HW/SW)
- A better approach for VM moving and NAS/SAN cluster replication
- Redundant and reliable topology
- Easier management
- Better visibility

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



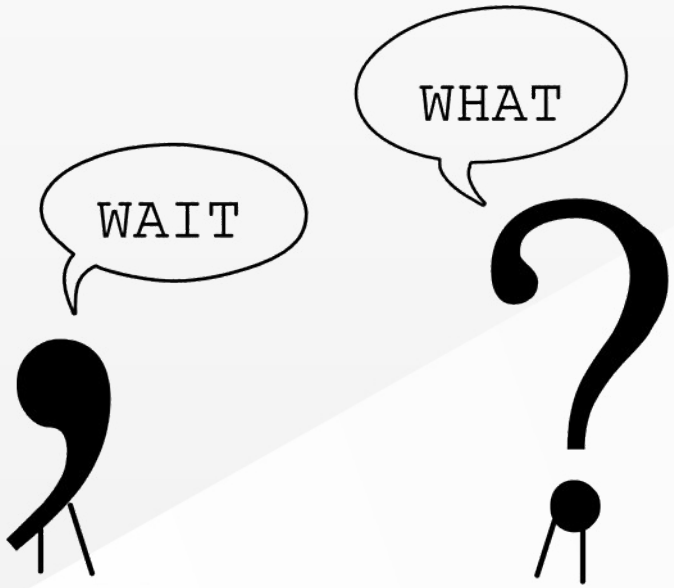
Where to go?

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

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

Hardware? Branded, whiteboxes or baremetal

Software? Open source or branded



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 anytime (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 actually provides better manageability than traditional systems

Classic switch design

sw1

Management plane

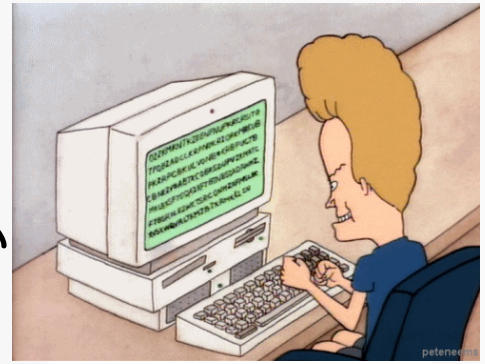
User tools for managing infrastructure (CLI,REST-API, SNMP etc.)

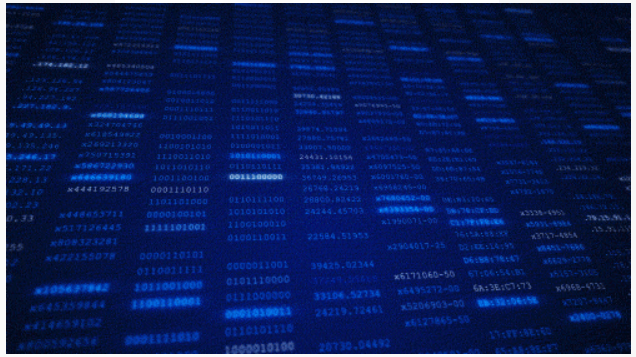
Control plane

Signaling between network entities

Data plane

Movement of application data packets





sw2

Management plane

User tools for managing infrastructure (CLI,REST-API, SNMP etc.)

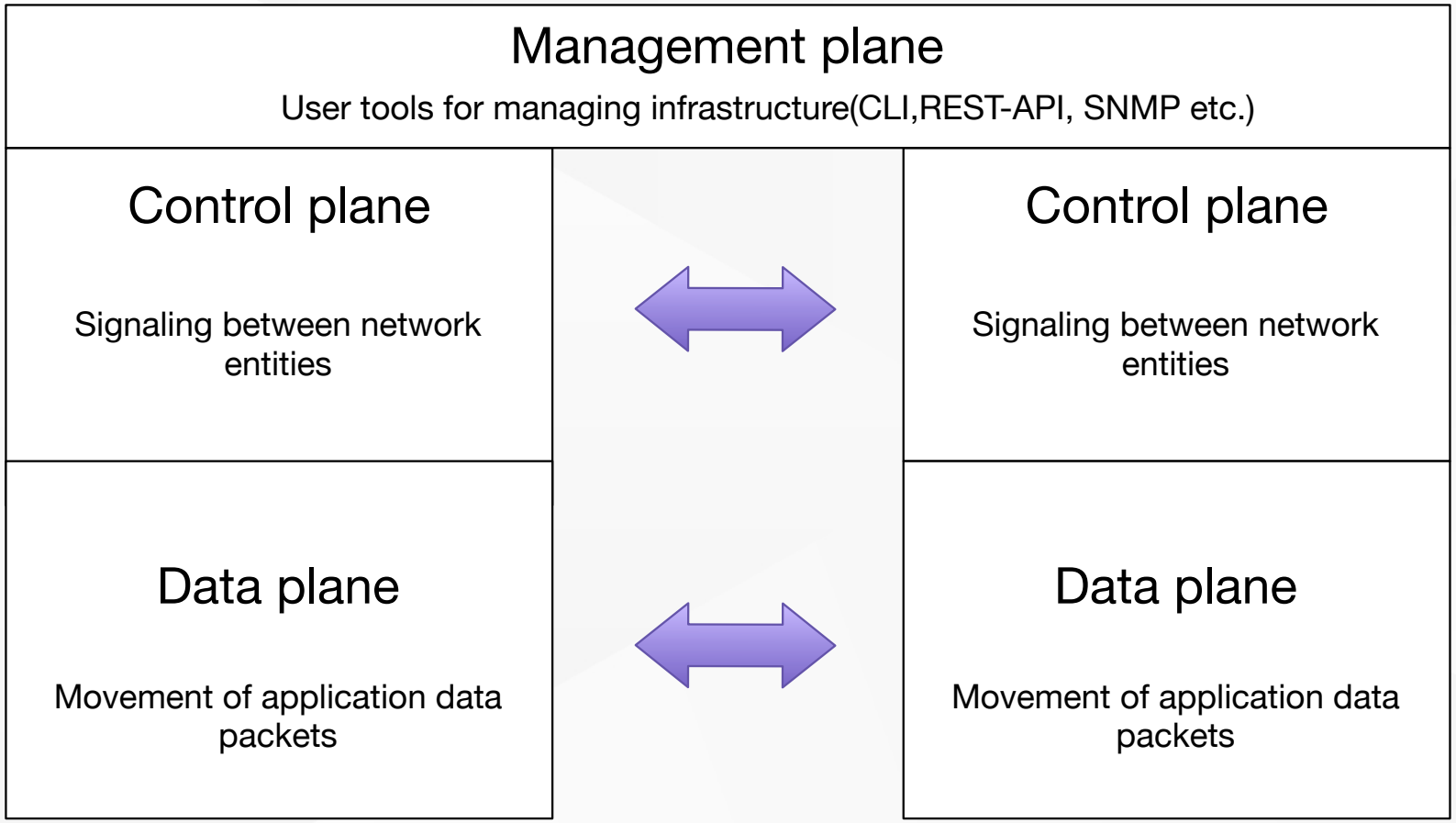
Control plane

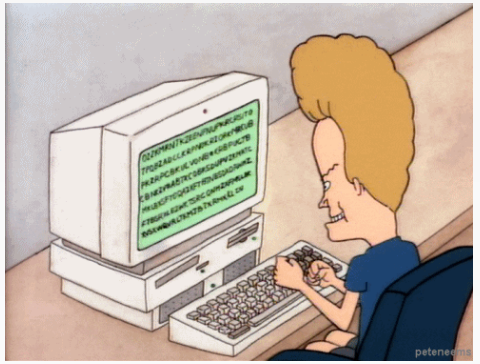
Signaling between network entities

Data plane

Movement of application data packets

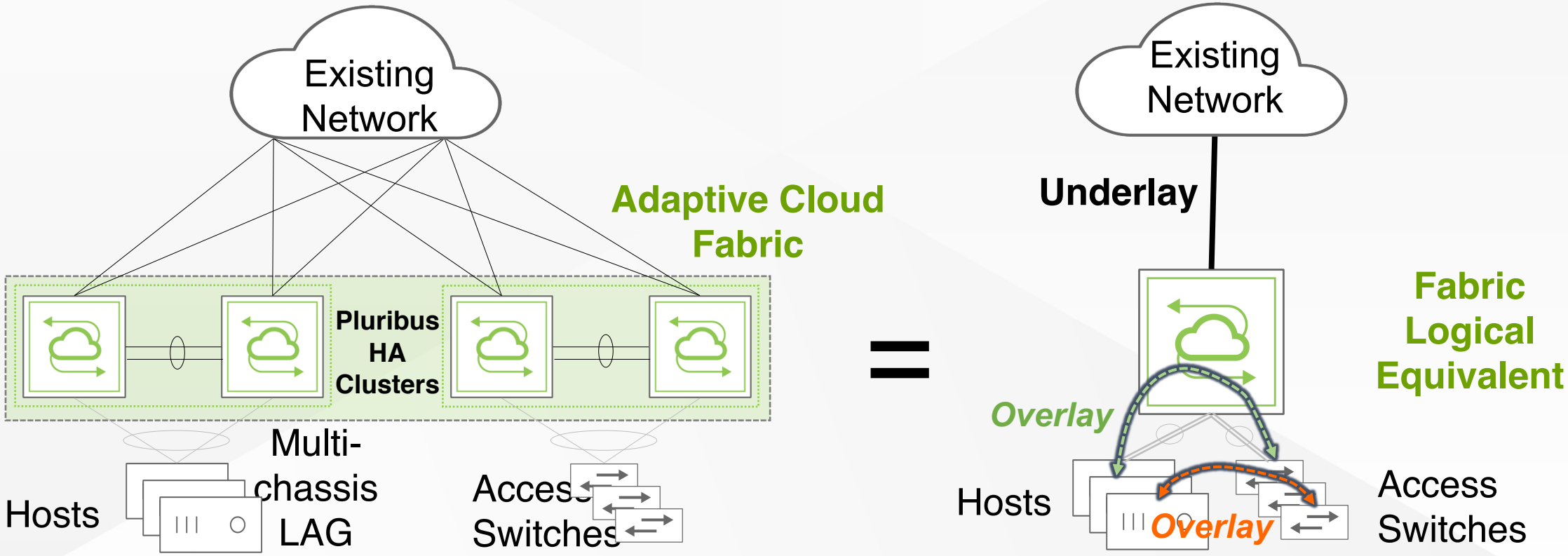
Pluribus distributed SDN fabric concept





Fabric logical view

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



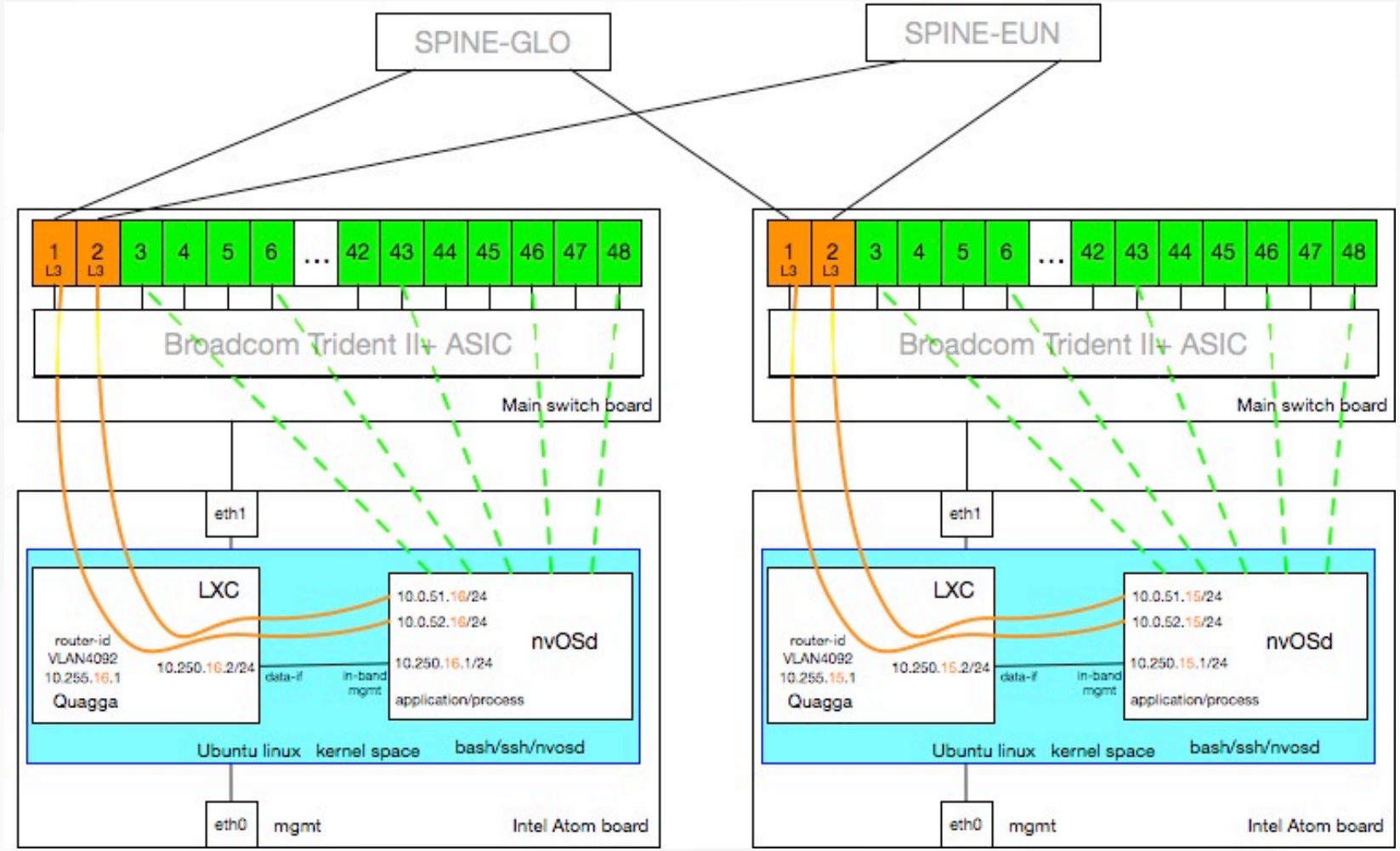
Geographically dispersed sites

Building a fabric with VxLAN

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

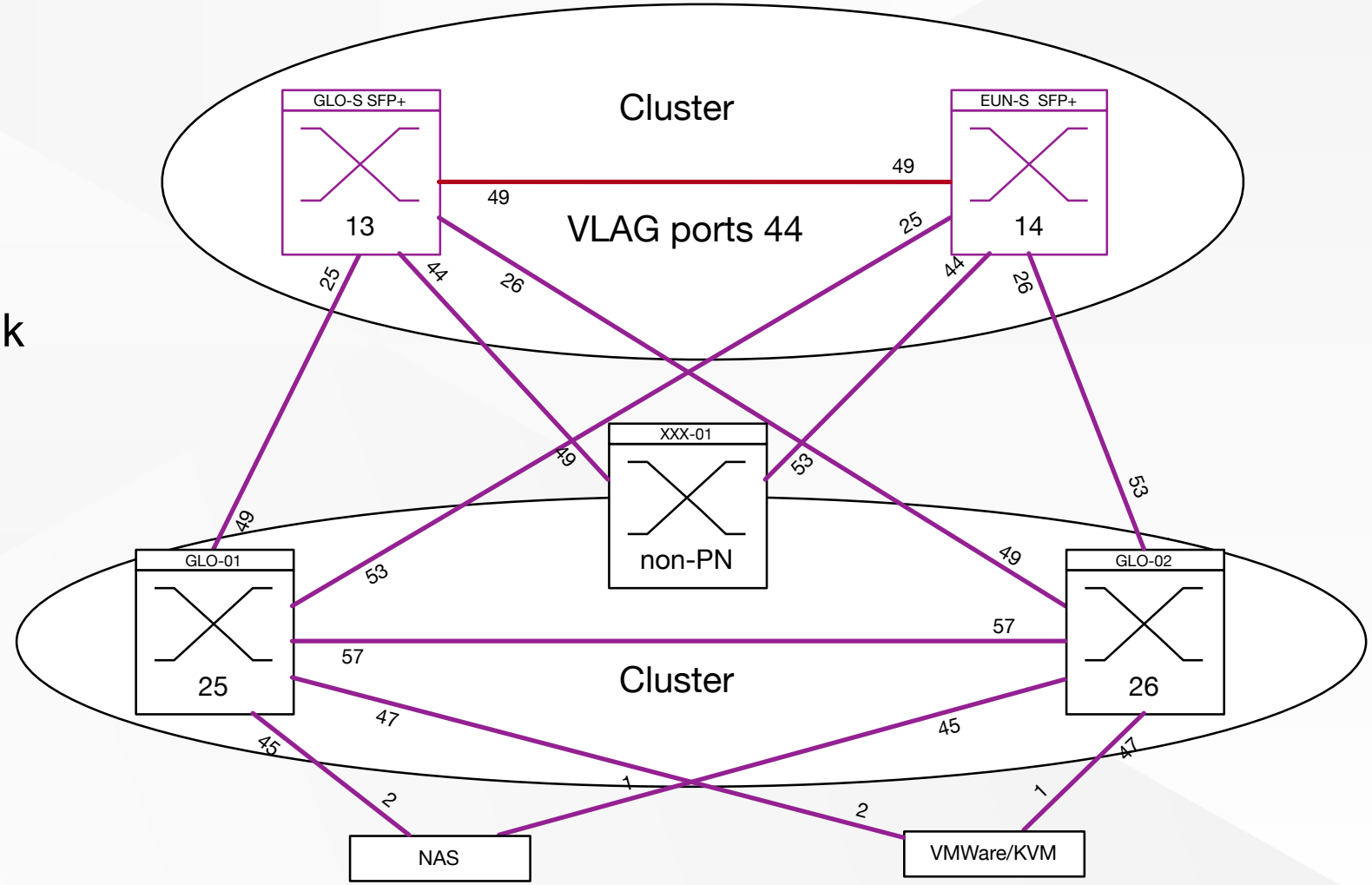
Open switch configuration

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



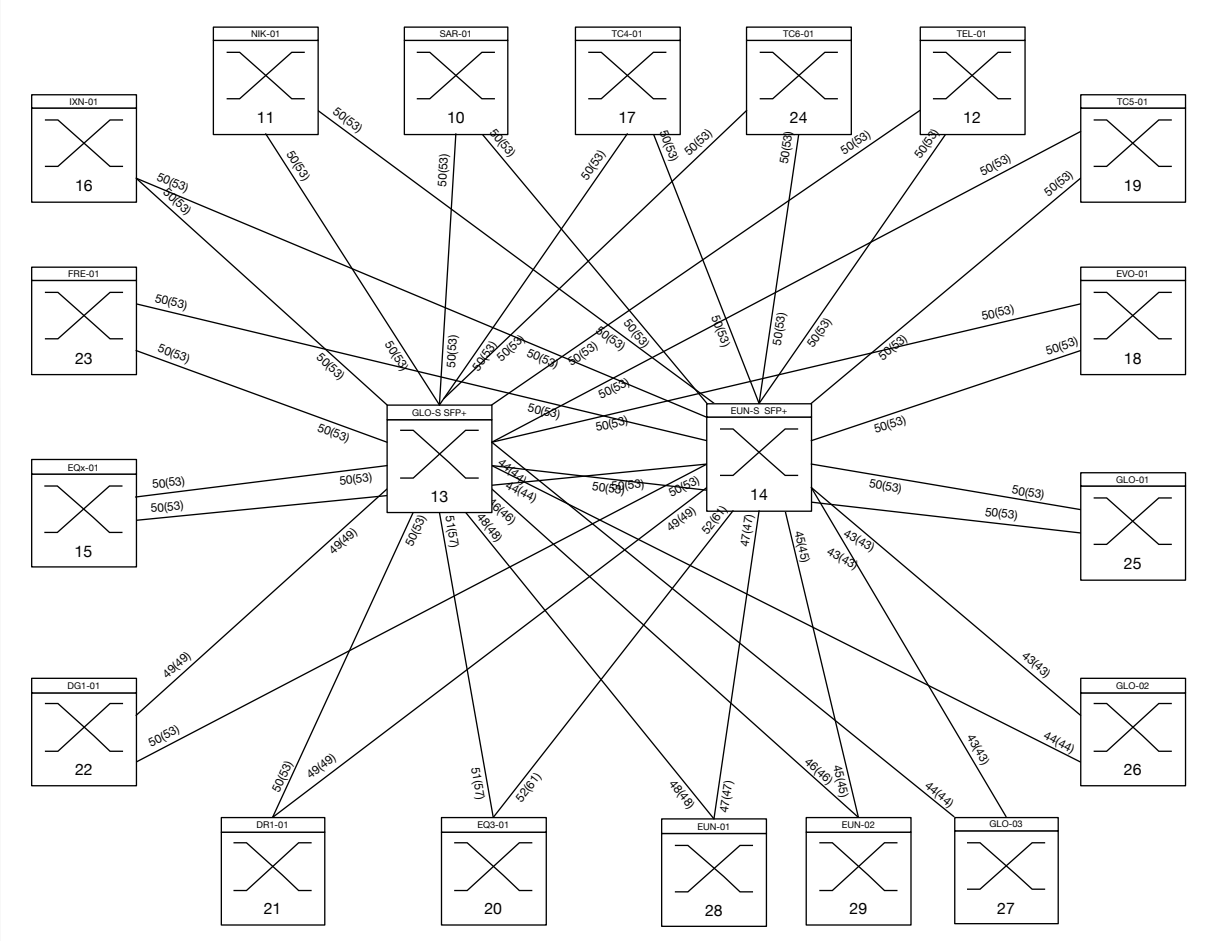
MC-LAG redundant connections

- 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
- 1077 mgmt ports in use in NL



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**

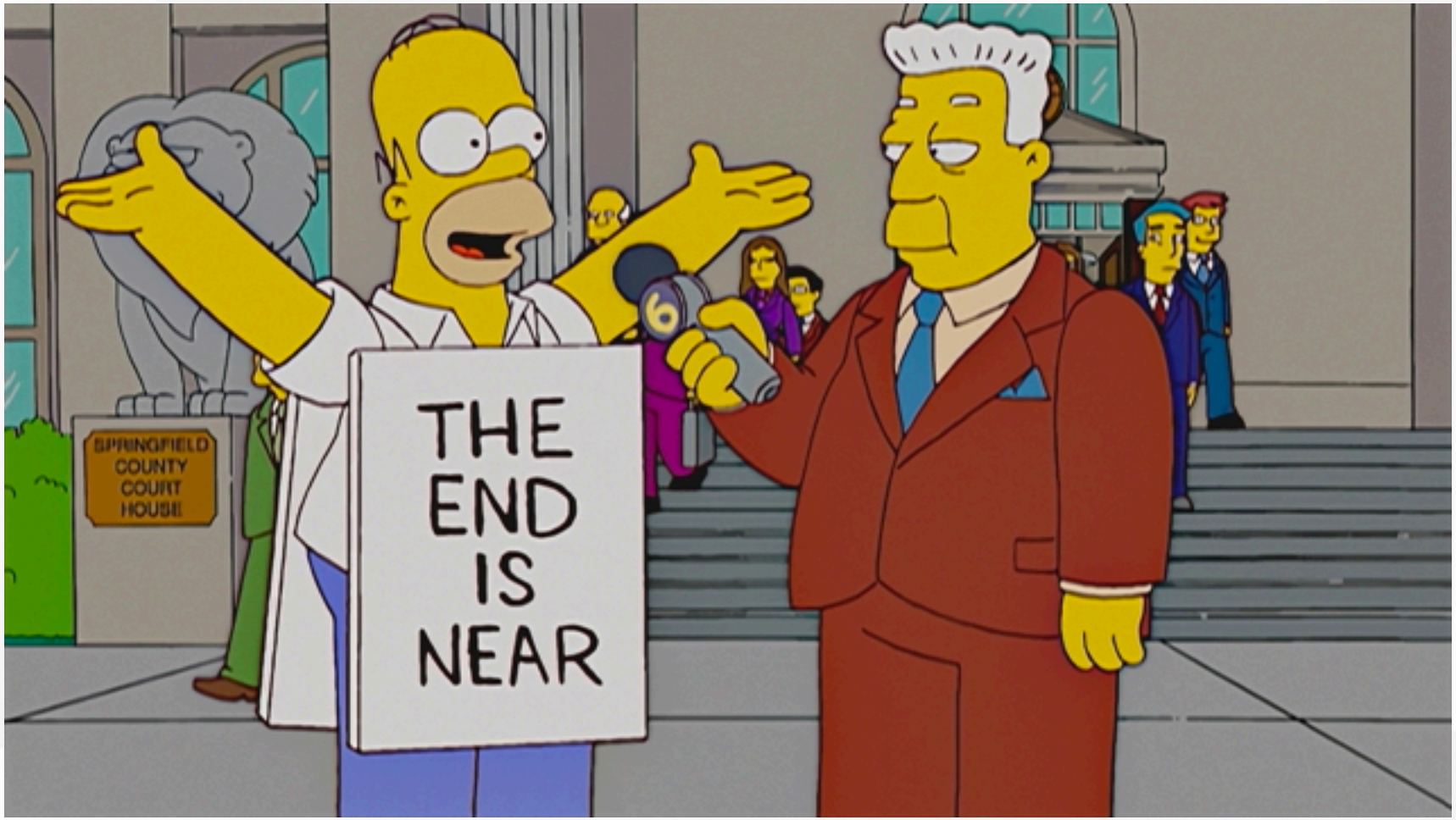
Unified fabric examples

```
CLI (noc@eun-spine-01) > port-show no-show-headers | grep demonix
eun-03       27   27        185.55.138.                    b0:83:fe:c1:b6:a0 109  109   eun-03       up,host,vlan-up                                fd,1g,10g,
autoneg       demonix-ipmi
eun-03       28   28        185.55.137.                    b0:83:fe:c1:b6:9e 105  105   eun-03       up,host,vlan-up                                fd,1g,10g,
autoneg       demonix-mng
```

```
root@eun-spine-01:/var/tmp# lsb_release -a
No LSB modules are available.
Distributor ID: Ubuntu
Description:    Ubuntu 14.04.5 LTS
Release:        14.04
Codename:       trusty
root@eun-spine-01:/var/tmp# uname -a
Linux eun-spine-01 4.4.0-31-generic #50~14.04.1-Ubuntu SMP Wed Jul 13 01:07:32 UTC 2016 x86_64 x86_64 x86_64 GNU/Linux
root@eun-spine-01:/var/tmp# lxc-ls
eun-spine-01-vrouter
root@eun-spine-01:/var/tmp# ps aux | grep quagga
root     13809  0.0  0.0  38648  2124 ?        Ss    2018  30:44 /usr/sbin/zebra --daemon --config_file /etc/quagga/zebra.conf -A 127.0.0.1
root     13825  0.4  0.0  38052  1692 ?        Ss    2018 3928:12 /usr/sbin/bfdd --daemon --config_file /etc/quagga/bfdd.conf -A 127.0.0.1
root     13829  0.1  0.1  42112  3928 ?        Ss    2018 1299:01 /usr/sbin/ospfd --daemon --config_file /etc/quagga/ospfd.conf -A 127.0.0.1
root     13833  0.0  0.0  35800   444 ?        Ss    2018  47:05 /usr/sbin/watchquagga --daemon zebra bfdd ospfd
root     28685  0.0  0.0  33352  3284 pts/2    S+   13:51   0:00 grep --color=auto quagga
root@eun-spine-01:/var/tmp#
```

```
CLI (noc@eun-spine-01*) > node-show format name,fab-name,state,device-state,firmware-upgrade,
name          fab-name state   device-state firmware-upgrade
------------ -------- ------- ------------ ----------------
eun-spine-01 AMS-IX   online  ok           not-required
glo-spine-01 AMS-IX   online  ok           not-required
glo-01       AMS-IX   online  ok           not-required
glo-03       AMS-IX   online  ok           not-required
glo-02       AMS-IX   online  ok           not-required
tc5-01       AMS-IX   online  ok           not-required
dr1-01       AMS-IX   online  ok           not-required
eun-02       AMS-IX   online  ok           not-required
eun-03       AMS-IX   online  ok           not-required
```

Apocalypse?!? Management unreachable?



Software stack broken?!?

=> For any reason (upgrade, fail, bug,..) the software stack on the management switch is not working and we need to manage?!? the management switch (or another devices on that location):

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

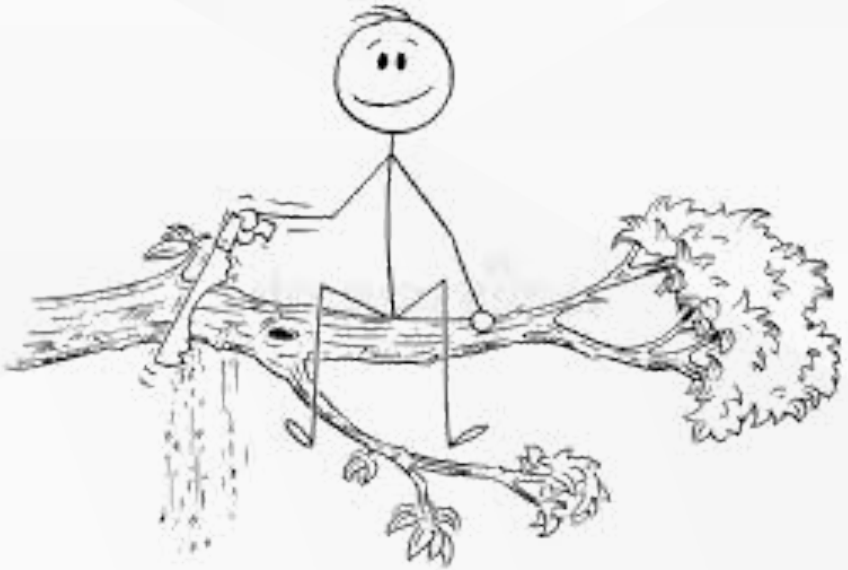
- If the management switch fail, the site still working, but you will lose all the management capability
- On this kind of event, we can reach the terminal server via passive channel and connect to the devices via console for troubleshooting

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

The unified fabric is amazing!!!

But it’s also really dangerous! ☠



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!!!

If you want more details or talk about AMS-IX, meet us for a beer at The Peering Coordination Forum =]

Thank you!

Questions, suggestions or remarks?

maxx.cherevko@ams-ix.net

tiago.goncalves@ams-ix.net