amsix
amsterdam internet exchange

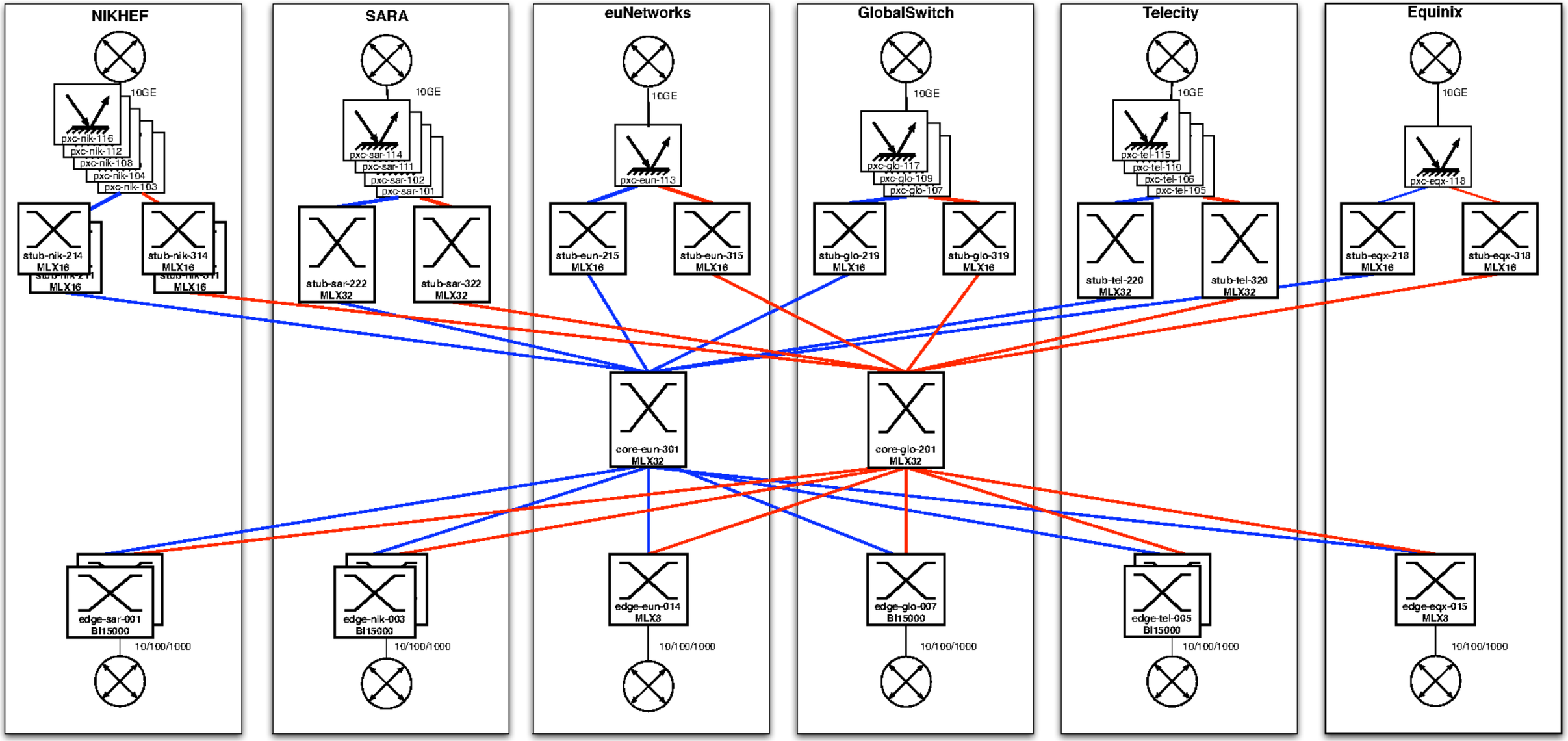
AMS-IX version 4

Details and operational experience

Martin Pels
<martin.pels@ams-ix.net>

5 October 2009

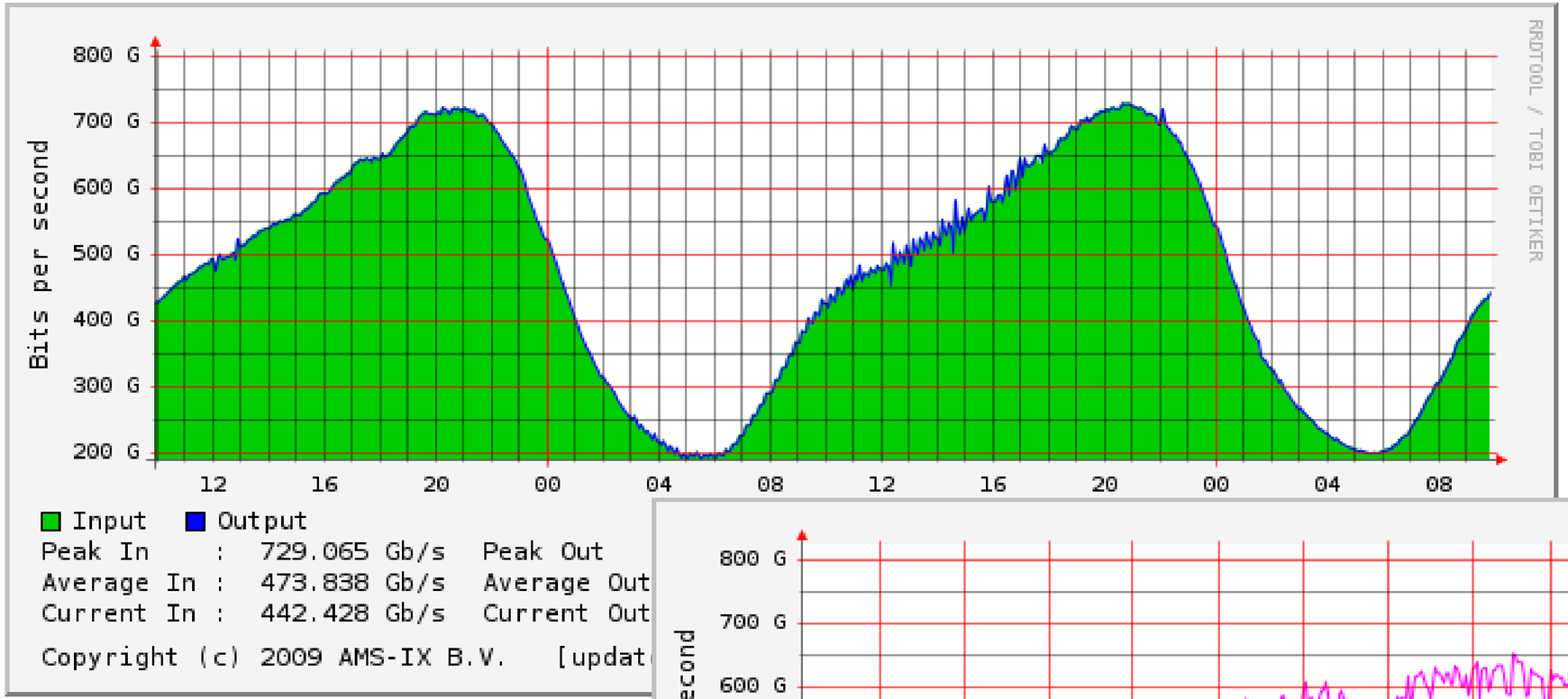
v1.2

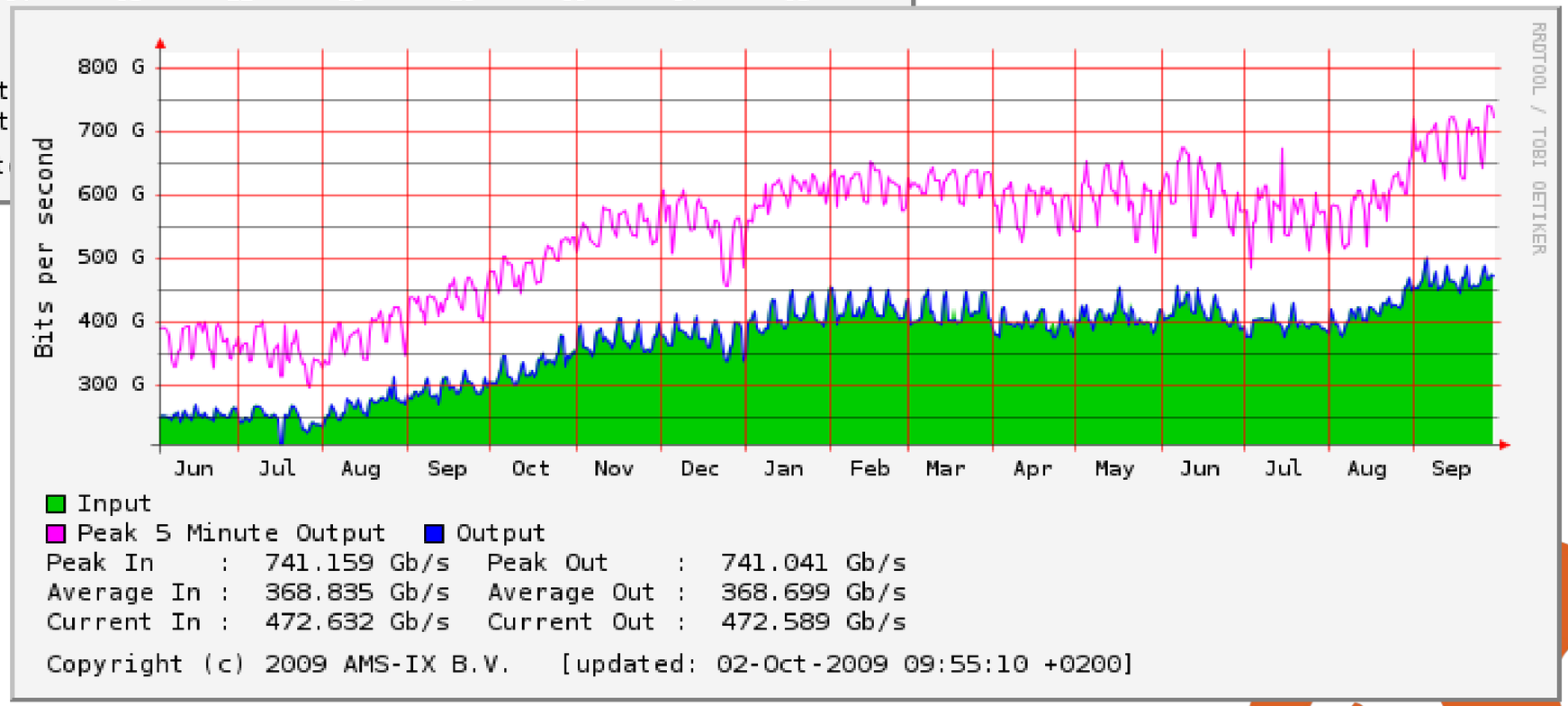


AMS-IX version 3

AMS-IX version 3

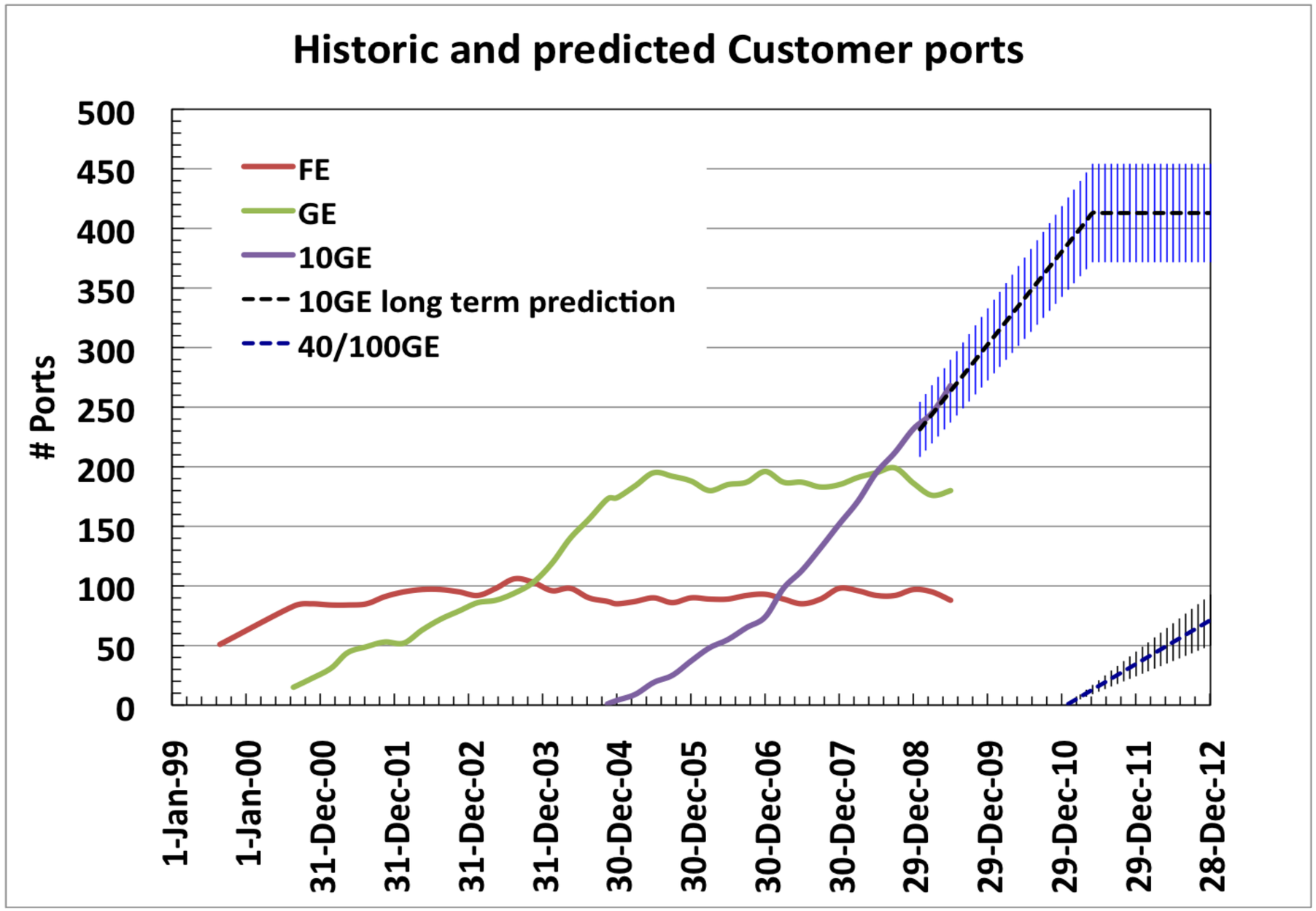
- E, FE and (N *) GE connections on BI-15k or RX8 switches
- (N *) 10GE connections resilient connected on switching platform (MLX16 or MLX32) via PXCs
- Brocade “port security” on customer interface to enforce one MAC per port rule for loop prevention
- VSRP (Brocade proprietary) between core switches for failovers of complete platform





AMS-IX customer traffic

daily and yearly traffic



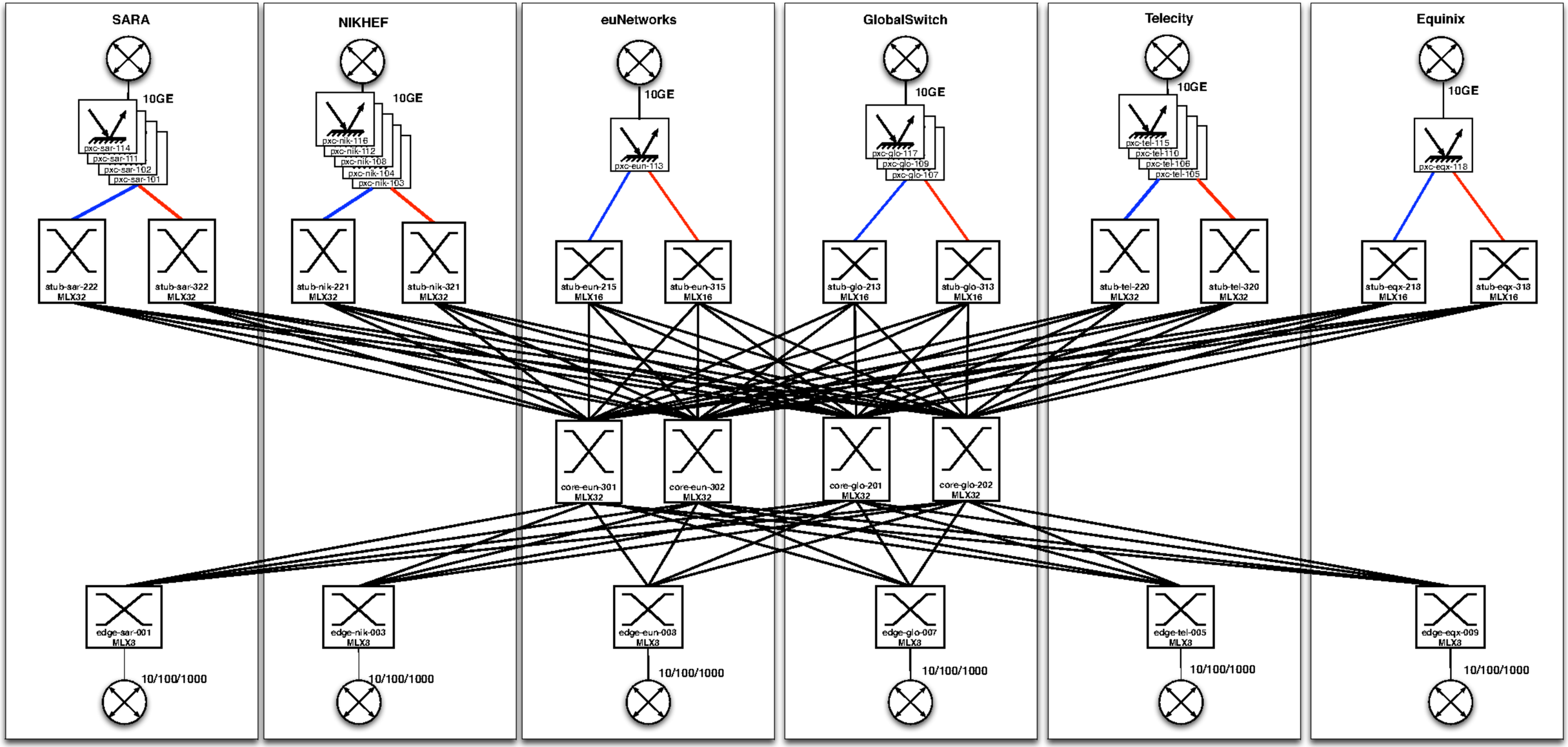
Traffic and port prognoses

Longterm 10G and 40G/100G customer port predictions

AMS-IX version 3

Bottlenecks and limitations

- Core switches (MLX32, 128x 10GE line rate) fully utilized
  - No substantially bigger switches on the market
- Platform failover introduces short link-flap on **all** 10GE customer ports
  - In few (but increasing) cases this leads to BGP flapping
- Growth of number of 10G connections and 10GE customer LAG size requires larger 10GE access switches
  - Smaller switches => less local switching => larger ISL trunks



AMS-IX version 4

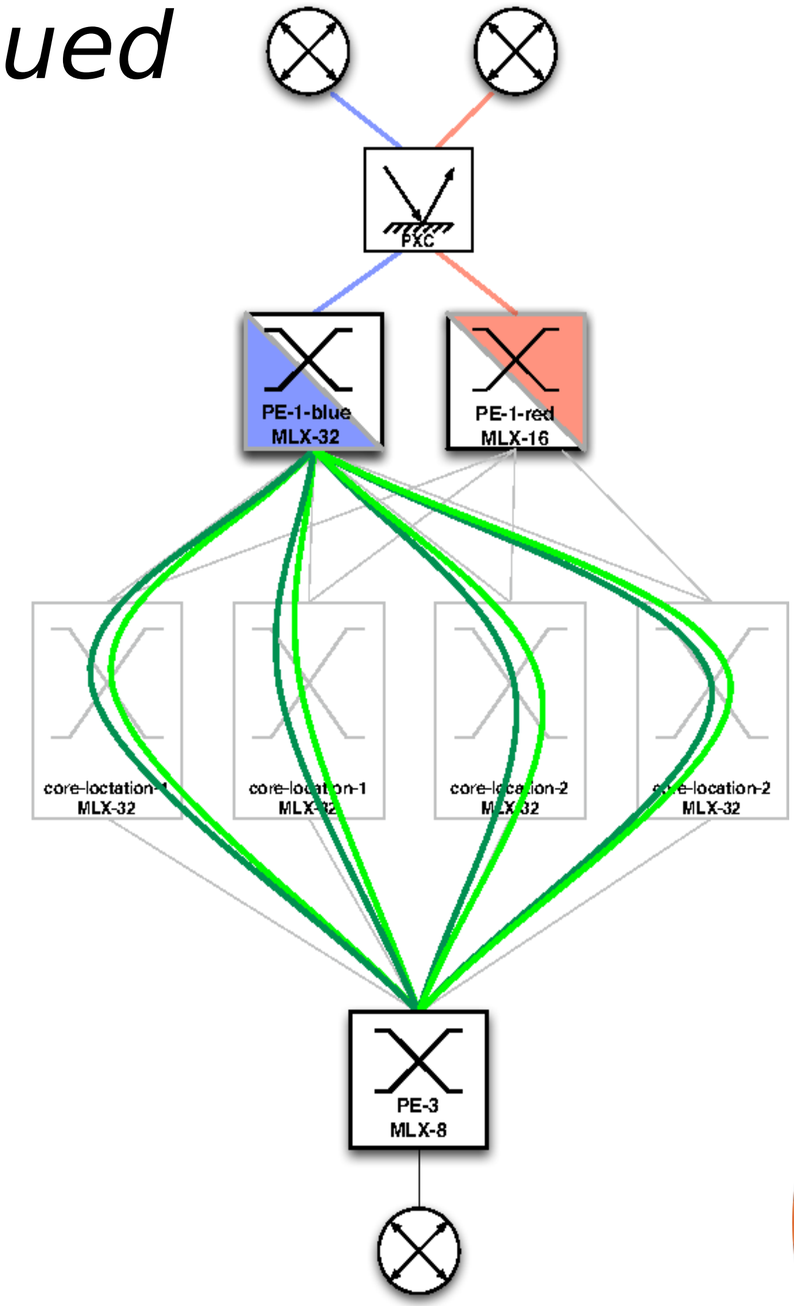
AMS-IX version 4

- Single hardware platform: Brocade MLX
- Upscaling of access switches to Brocade MLX32
- MPLS/VPLS-based peering platform
- Physical star system, logical full mesh
- VPLS instance per VLAN-service
- Port security replaced by L2 ACLs

AMS-IX version 4

continued

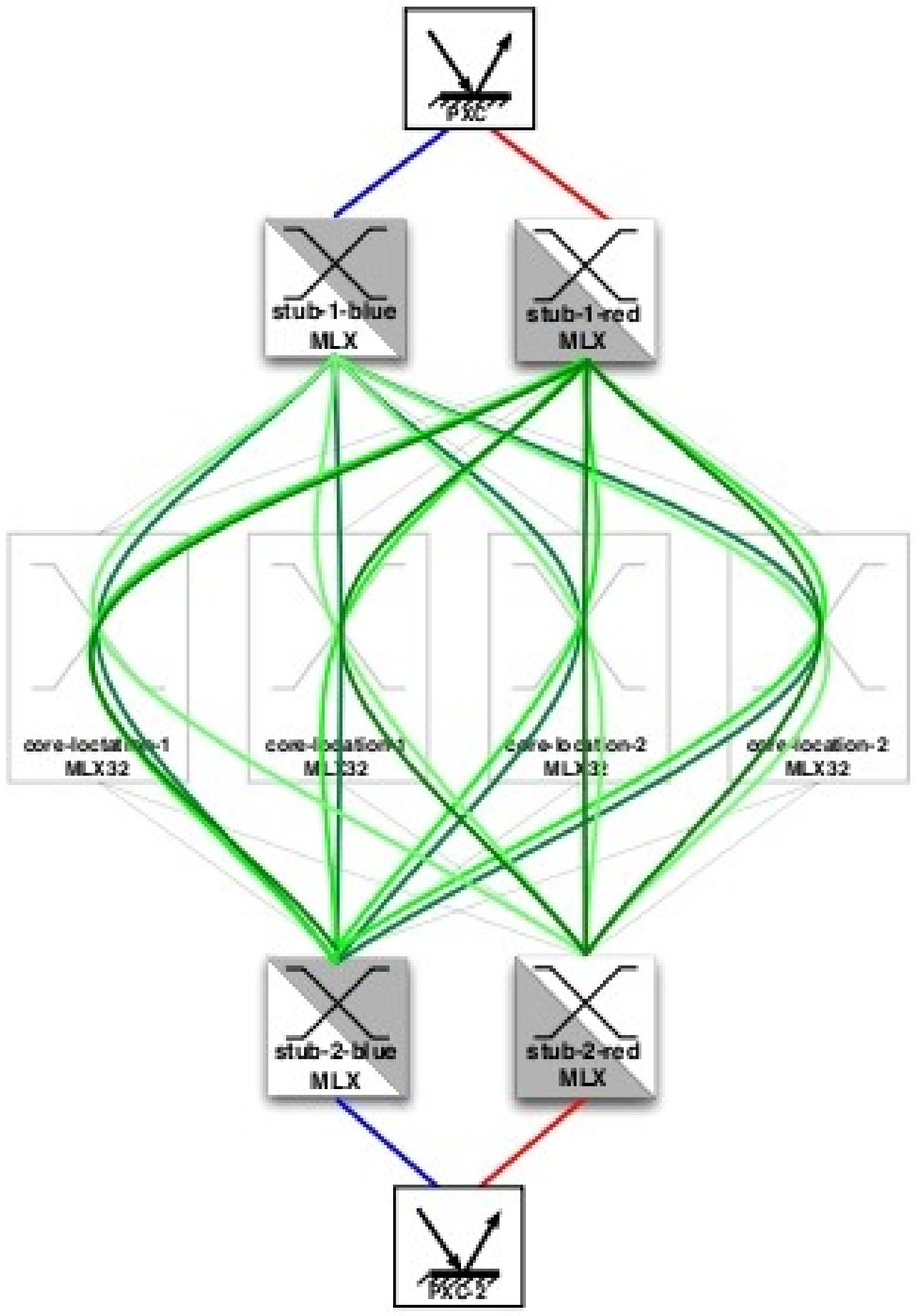
- Active/active, load-sharing over 4 cores
  - 4x2 LSPs between each pair of access switches
- Core redundancy (50% backbone load)
- Access switch redundancy (50% of ports active, PXC failovers)

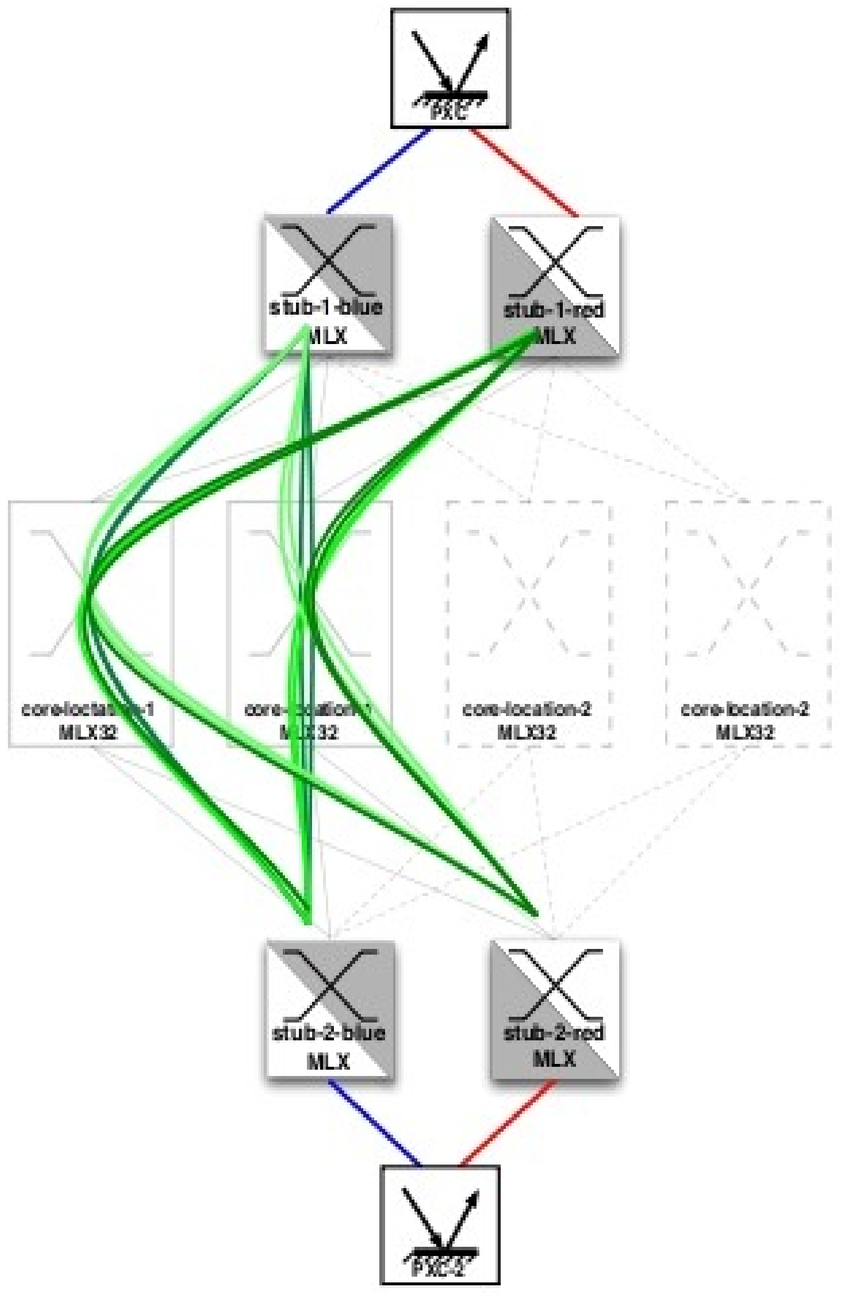


Resilience

core

- LSP path change
  - on backbone link failure
  - on core switch failure
  - service interruption: ~20ms
  - **no link flaps!**



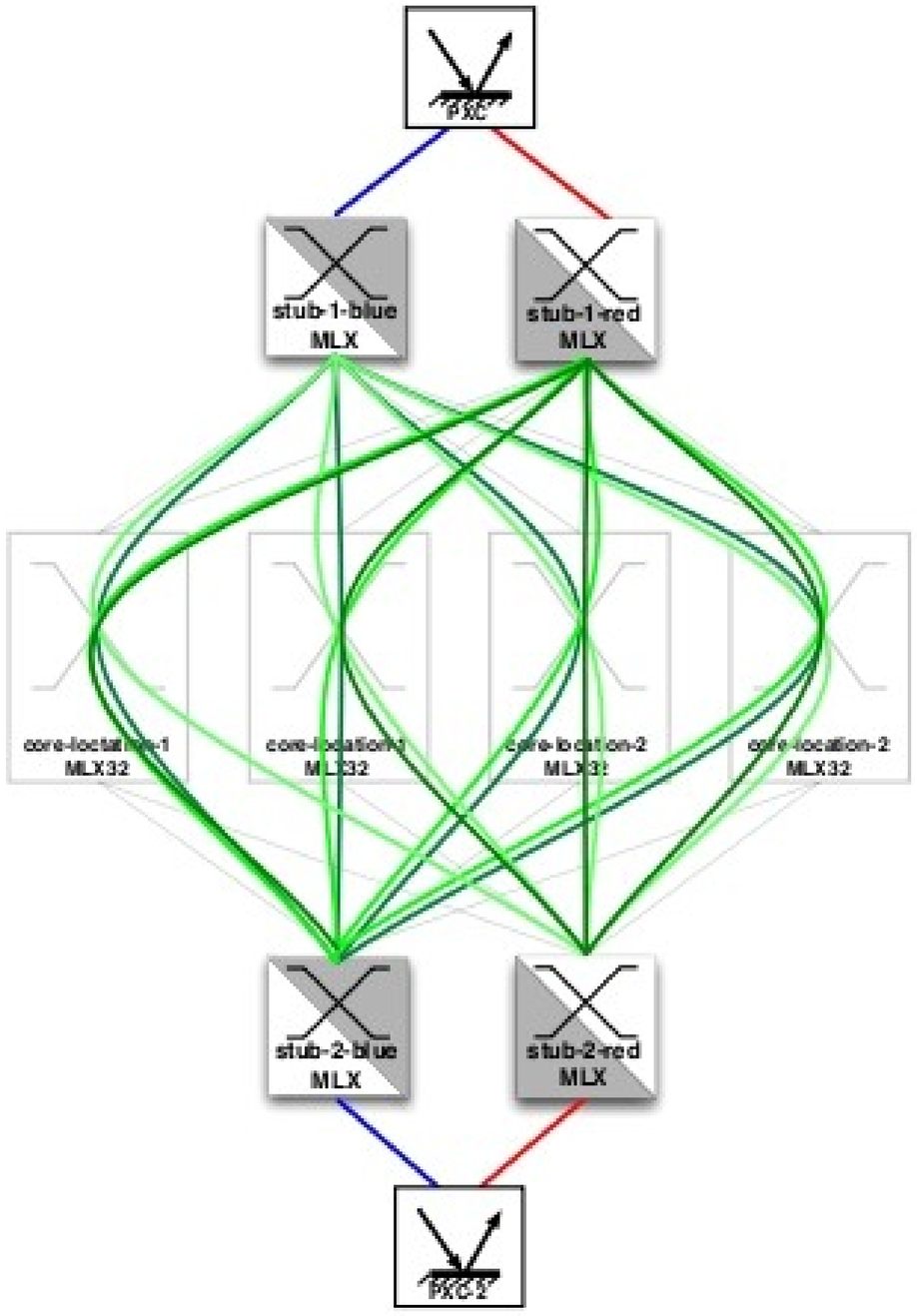


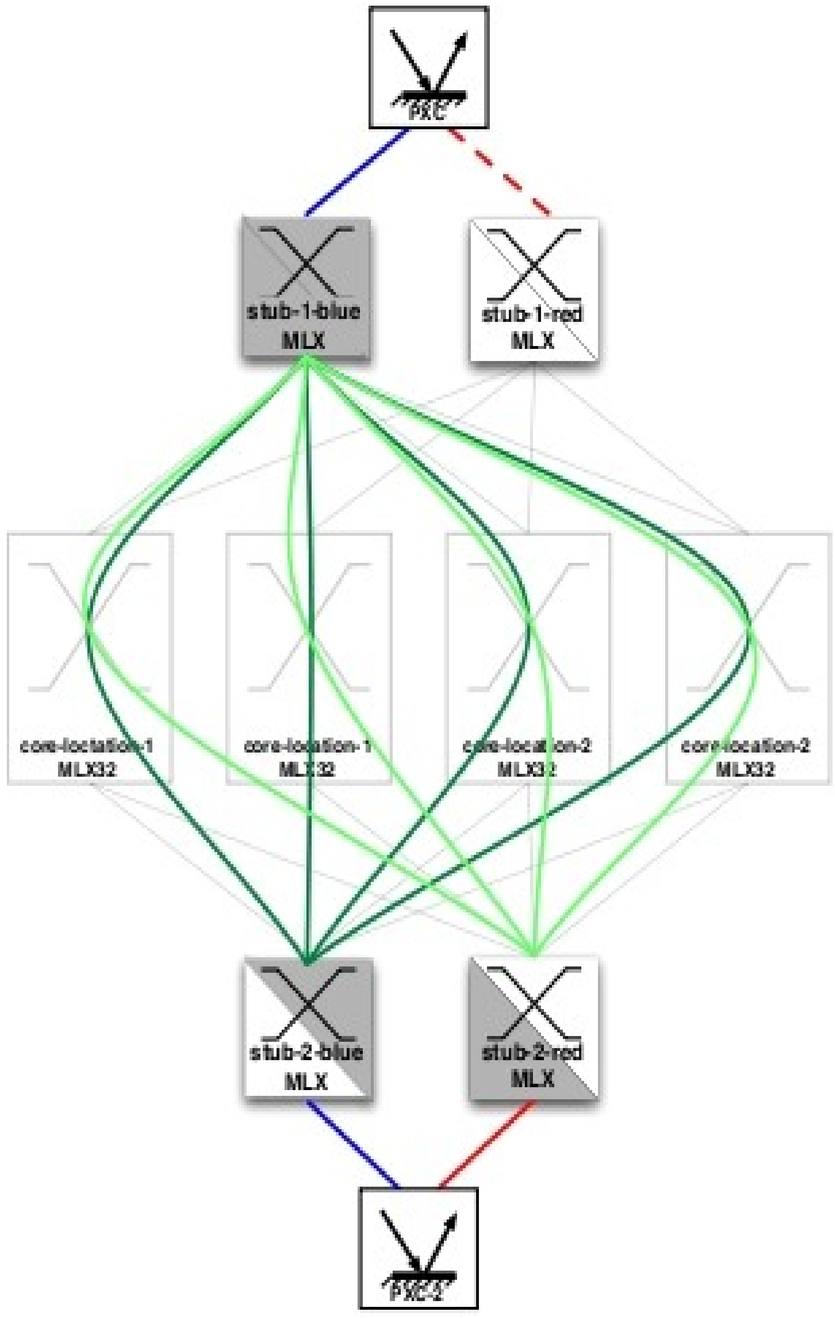
LSP path change

Resilience

PEs

- PXC failovers
  - on access switch failure
  - triggered by LSP failure
  - service interruption: ~250ms
  - **localized to one set of PEs**





PXC failover

Platform migration

in a nutshell

- Move 1G access switches behind PXCs
  - Customer ports cannot be L2 and VPLS concurrently
- Migrate one half of platform to VPLS
- Migrate second half of platform to VPLS
- Merge both halves into a single active/active platform
- Connect 1G access switches directly to cores
- *Details in EIX session on Thursday*

Operational experience

Issues

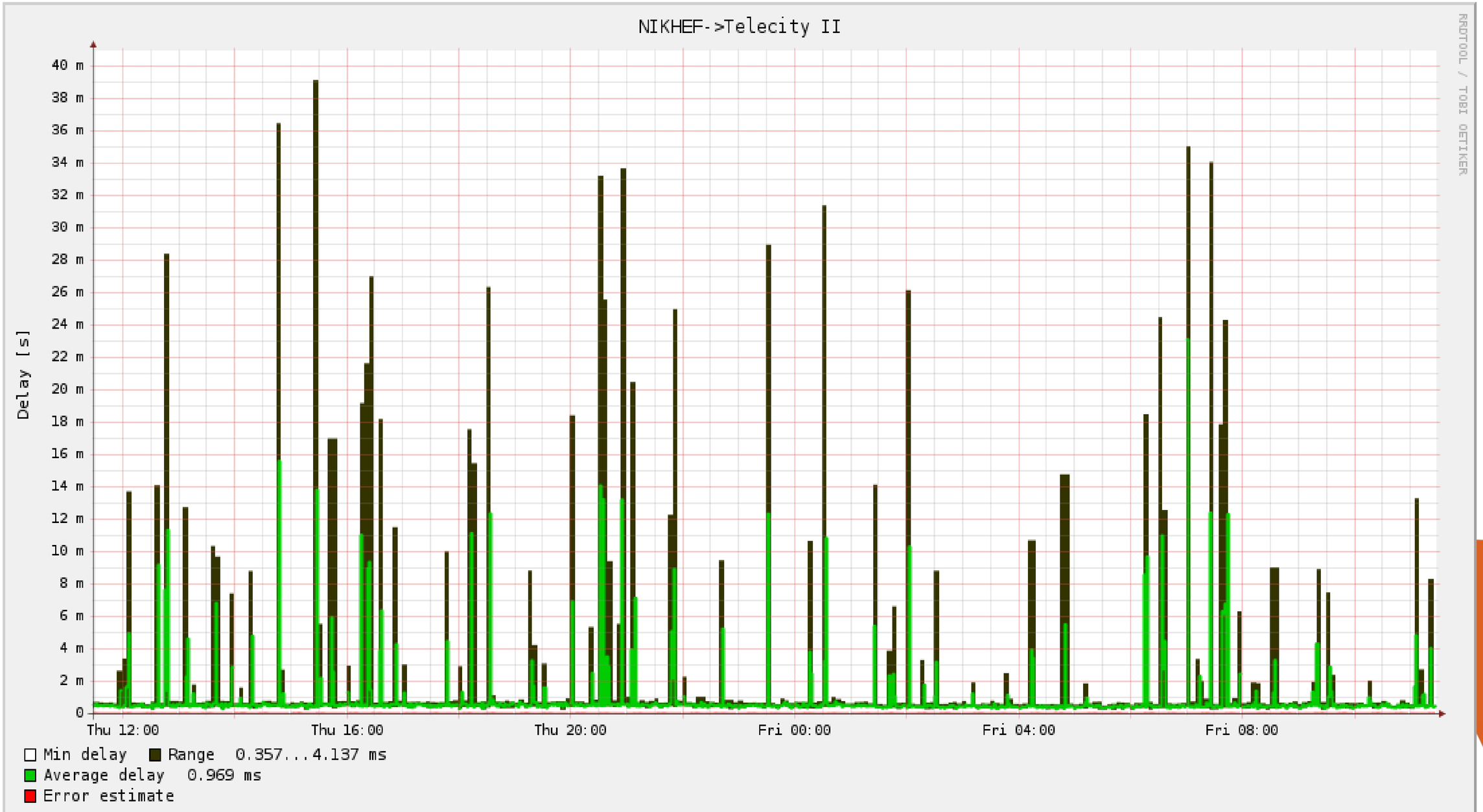
- BFD instability
  - High LP CPU load caused BFD timeouts
  - Resolved by increasing timers
- Bug: ghost tunnels
  - Double “Up” event for LSP path
  - Results in unequal load-balancing
  - Scheduled to be fixed in next patch release

Operational experience
Issues (2)

- Multicast replication
  - Replication done on ingress PE, not on core
  - Only uses 1st link of aggregate of 1st LSP
  - With PIM-SM snooping traffic is balanced over multiple links, but this has some serious bugs
  - Bugfixes and load-sharing of multicast traffic over multiple LSPs scheduled for next major release

Operational experience

Issues (3)



Operational experience
Issues (3)

- Delay spikes in RIPE TTM graphs
  - TTM datagrams have high interval (2 packets per minute), with some entropy (source port changes)
  - Brocade VPLS CAM: Entries programmed individually for each backbone port, age out after 60s
    - For 24-port aggregates, traffic often passes port without programming => CPU learning => high delay
- Does not affect real-world traffic
  - Much lower interval between frames
- Looking into changing/disabling CAM aging

Operational experience

Issues (4)

- *From 213.136.17.28: icmp_seq=1 Packet is claustrophobic*
- Limited to single user
- Suspecting problem caused by protocol-stack on client ;-)

Operational experience

The good stuff

- Increased stability
  - Backbone failures handled by MPLS (not seen by customers)
  - Access switch failures handled for a single pair of switches
  - Phased relocation of traffic streams
  - Looped traffic filtered by L2 ACL => No effect on linecard CPU

Operational experience

The good stuff (2)

- Easier debugging of customer ports
  - Simply swap to different, active switch using Glimmerglass PXC
- Config generation
  - Absolute necessity due to size of MPLS/VPLS configuration
  - Fairly simple because of single hardware platform

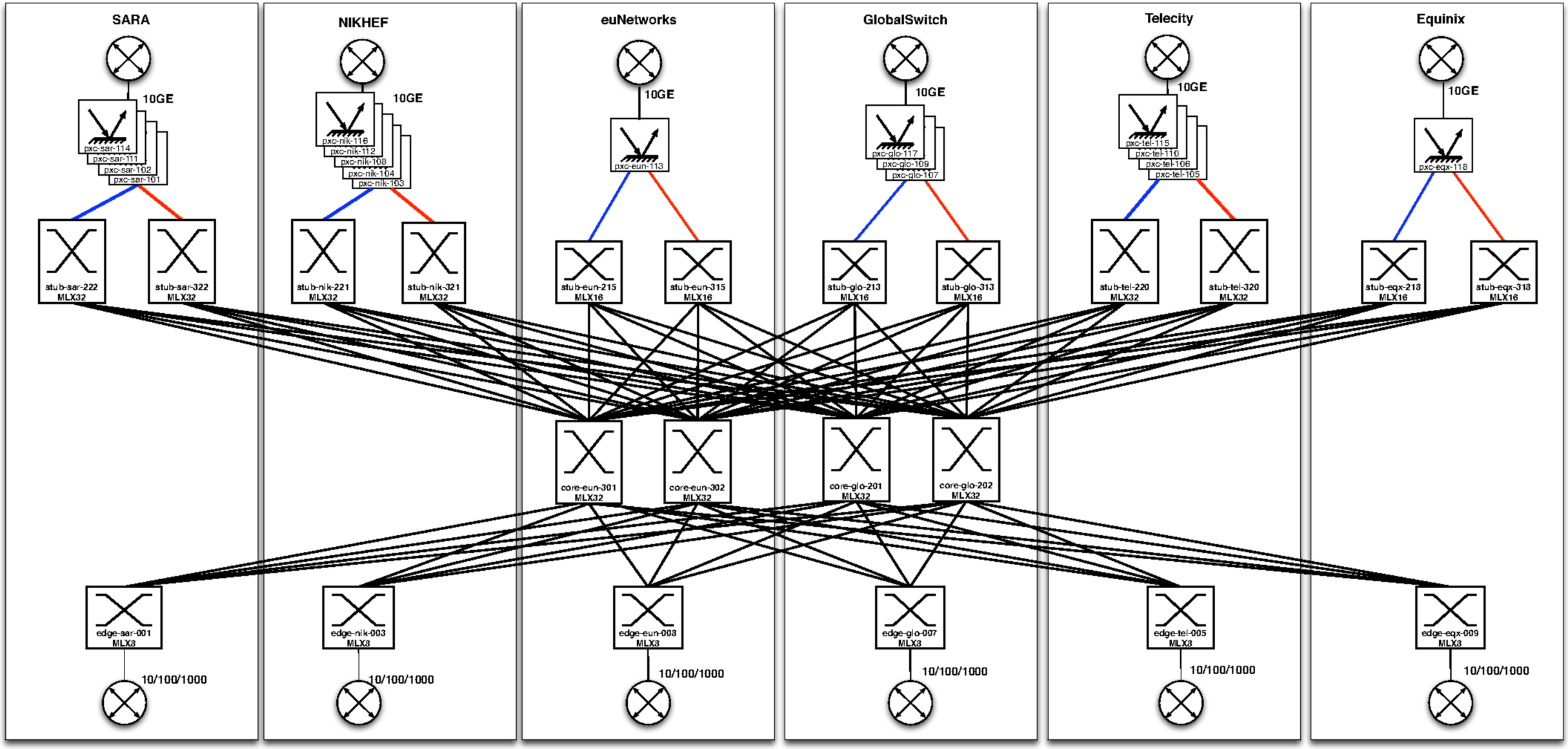
Operational experience

The good stuff (3)

- Scalability (future options)
  - Bigger core devices
    - Do not need to be MPLS-capable
  - Load-sharing over > 4 cores
    - Pending feature request
  - Use of different cores for sets of PEs
  - Multiple layers of P-routers

Conclusions

- Some issues found
  - Nothing with impact on customer traffic
- Traffic load-sharing over multiple devices solves scaling issues in the core
- Increased stability of the platform
  - Backbone failures not seen at the access level
  - Access switch failures trigger failover for corresponding Glimmerglass PXCs only
- Upscaling access switches allows for higher access port density
- Single hardware platform simplifies configuration generation


SARA
NIKHEF
euNetworks
GlobalSwitch
Telecity
Equinix
10GE
pxc-sar-114
pxc-sar-111
pxc-sar-102
pxc-sar-101
pxc-nik-116
pxc-nik-112
pxc-nik-108
pxc-nik-104
pxc-nik-103
pxc-eun-113
pxc-glo-117
pxc-glo-109
pxc-glo-107
pxc-tel-115
pxc-tel-110
pxc-tel-106
pxc-tel-105
pxc-eqx-118
stub-sar-222 MLX32
stub-sar-322 MLX32
stub-nik-221 MLX32
stub-nik-321 MLX32
stub-eun-215 MLX16
stub-eun-315 MLX16
stub-glo-213 MLX16
stub-glo-313 MLX16
stub-tel-220 MLX32
stub-tel-320 MLX32
stub-eqx-218 MLX16
stub-eqx-318 MLX16
core-eun-301 MLX32
core-eun-302 MLX32
core-glo-201 MLX32
core-glo-202 MLX32
edge-sar-001 MLX8
edge-nik-003 MLX8
edge-eun-003 MLX8
edge-glo-007 MLX8
edge-tel-005 MLX8
edge-eqx-009 MLX8
10/100/1000

Questions?