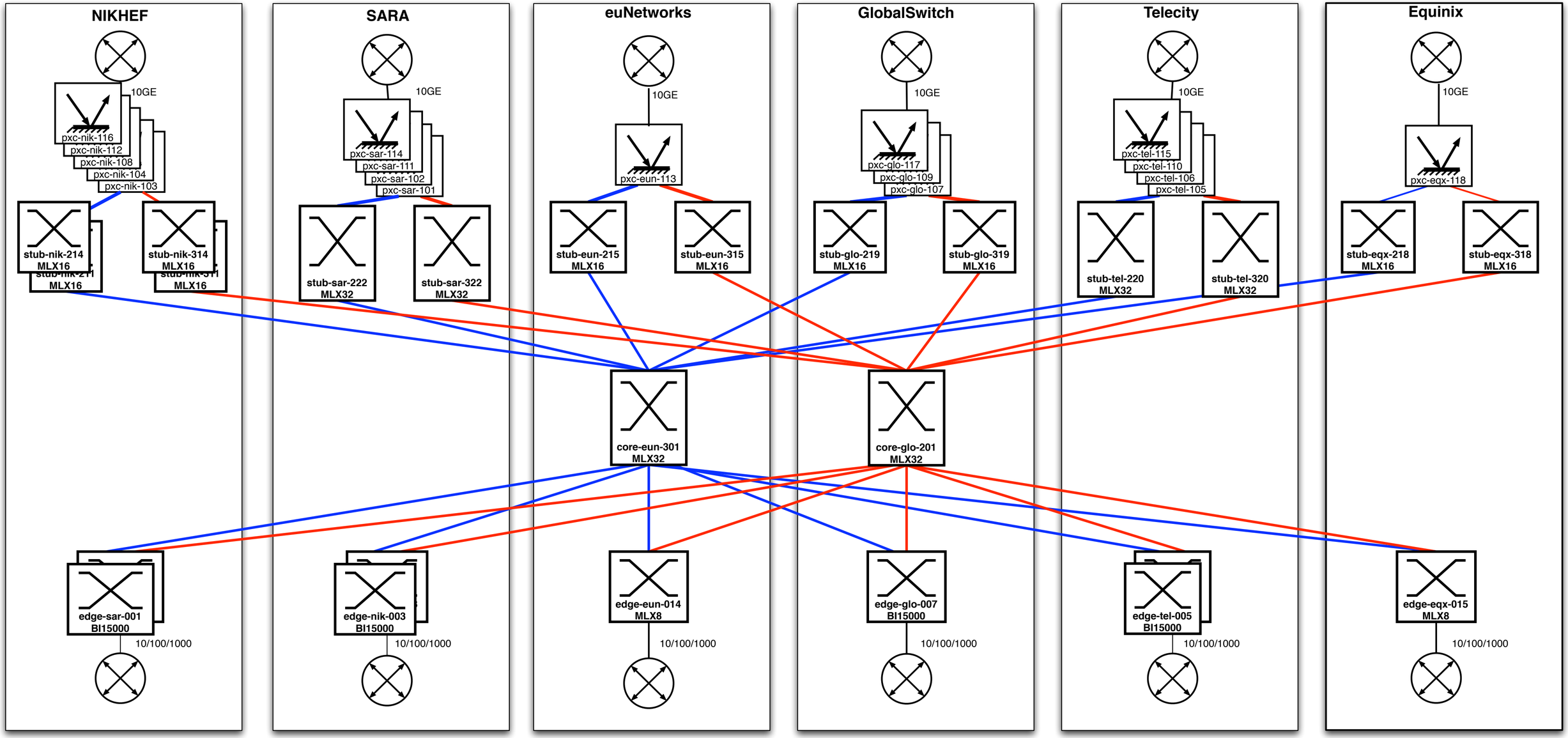
amsix
amsterdam internet exchange

AMS-IX version 4

an MPLS/VPLS based internet exchange

Overview

- AMS-IX version 3
  - Short overview
  - Bottlenecks and limitations
- AMS-IX version 4
  - The MPLS/VPLS platform
- AMS-IX v3 to v4 migration
- Operational Experience



June 2009 situation before start of migration

AMS-IX version 3

AMS-IX version 3

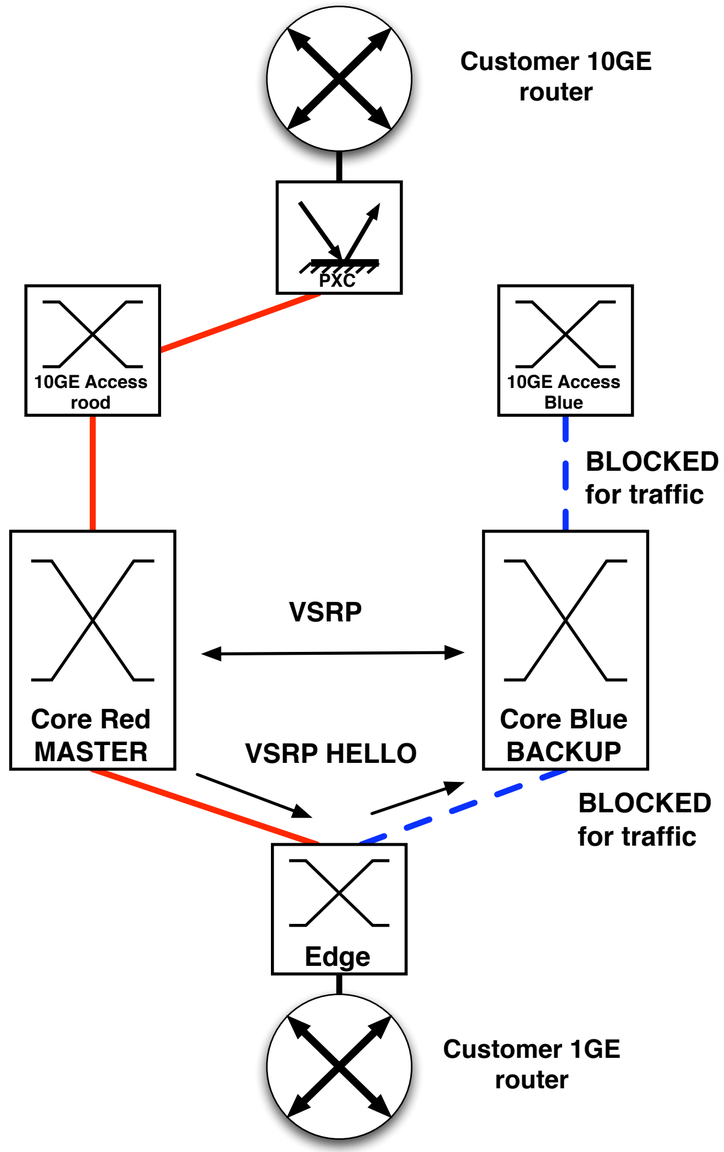
Characterization

- 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

AMS-IX version 3

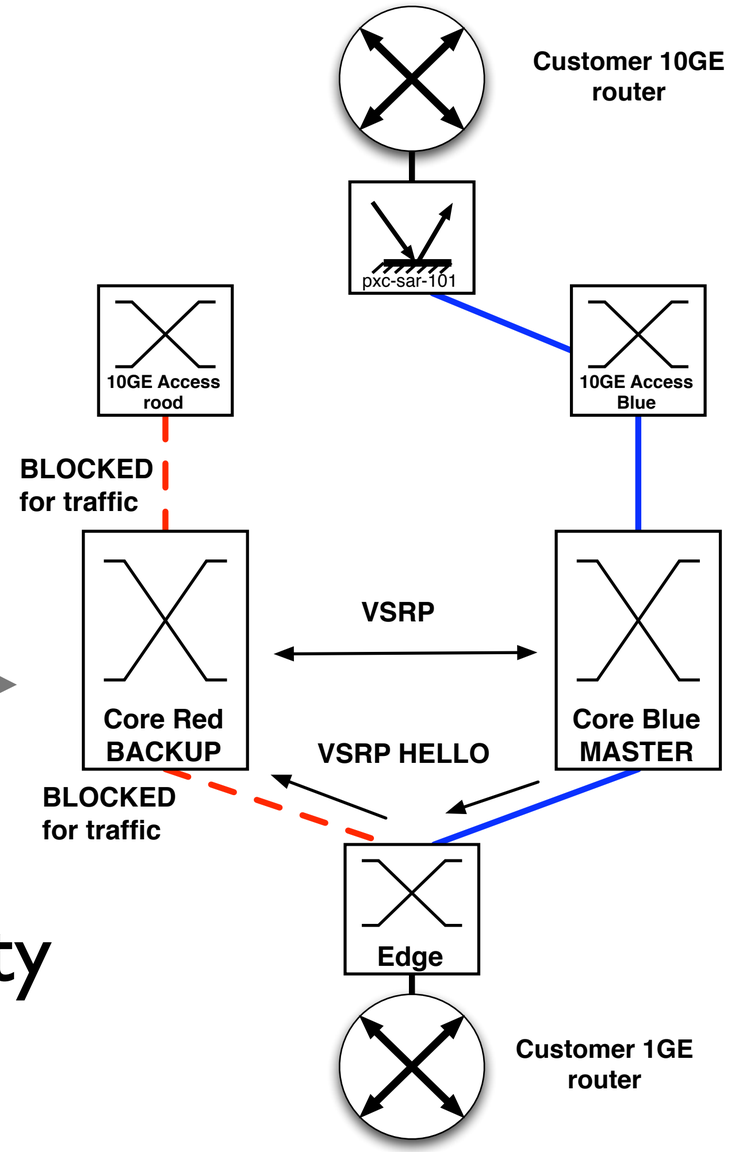
Characterization

- Two networks: one active at any moment in time
  - Selection of active network by VSRP
    - Inactive network switch blocks ports to prevent loops
  - PSCD, photonic switch control daemon
    - AMS-IX developed software to act on VSRP traps and manage PXCs



Problem or maintenance in red network

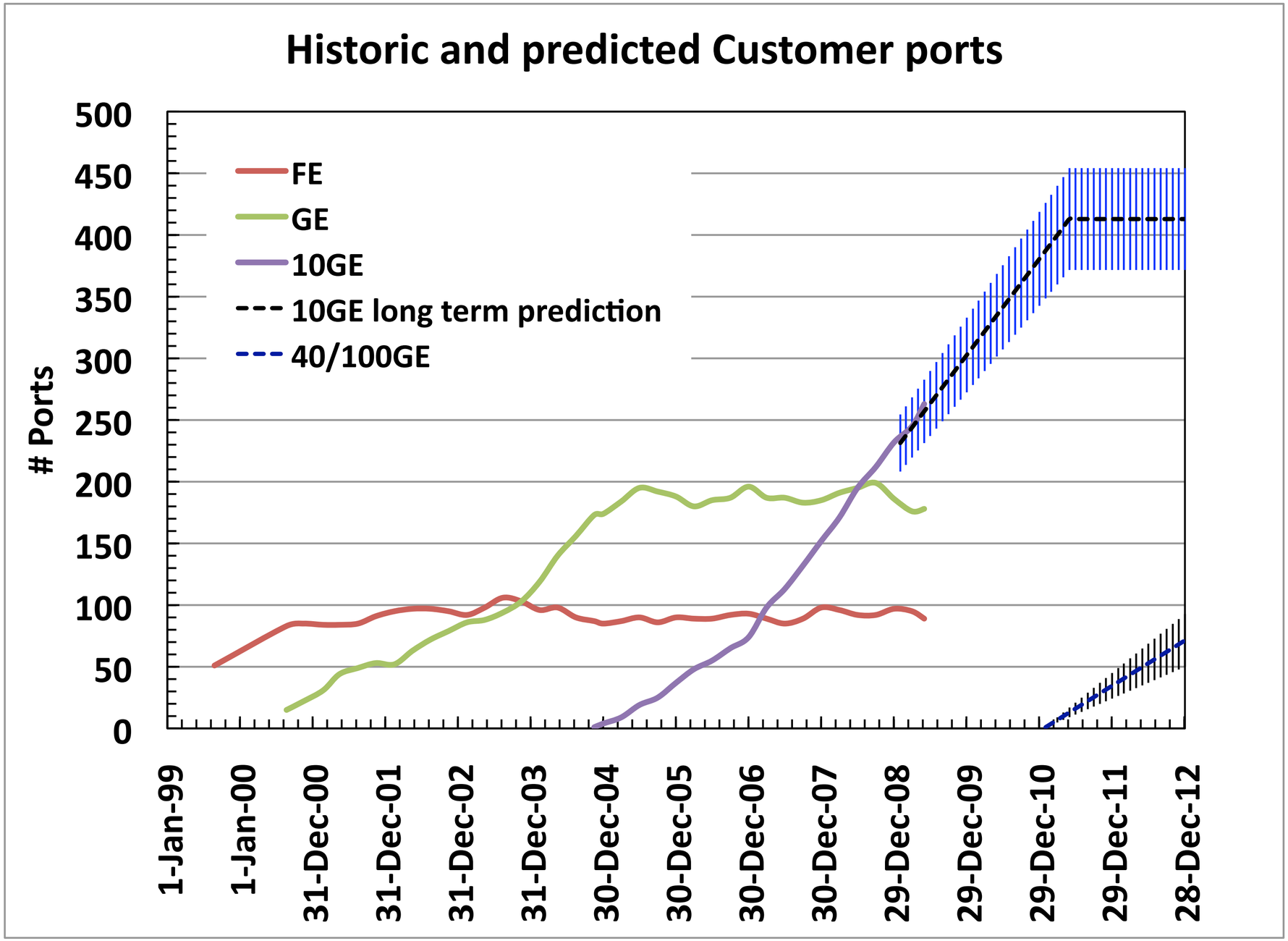
VSRP Priority Red Master lower than the Blue priority



AMS-IX Version 3 Platform

Topology Failover

Traffic and Port Prognoses



AMS-IX version 3

Bottlenecks and Limitations

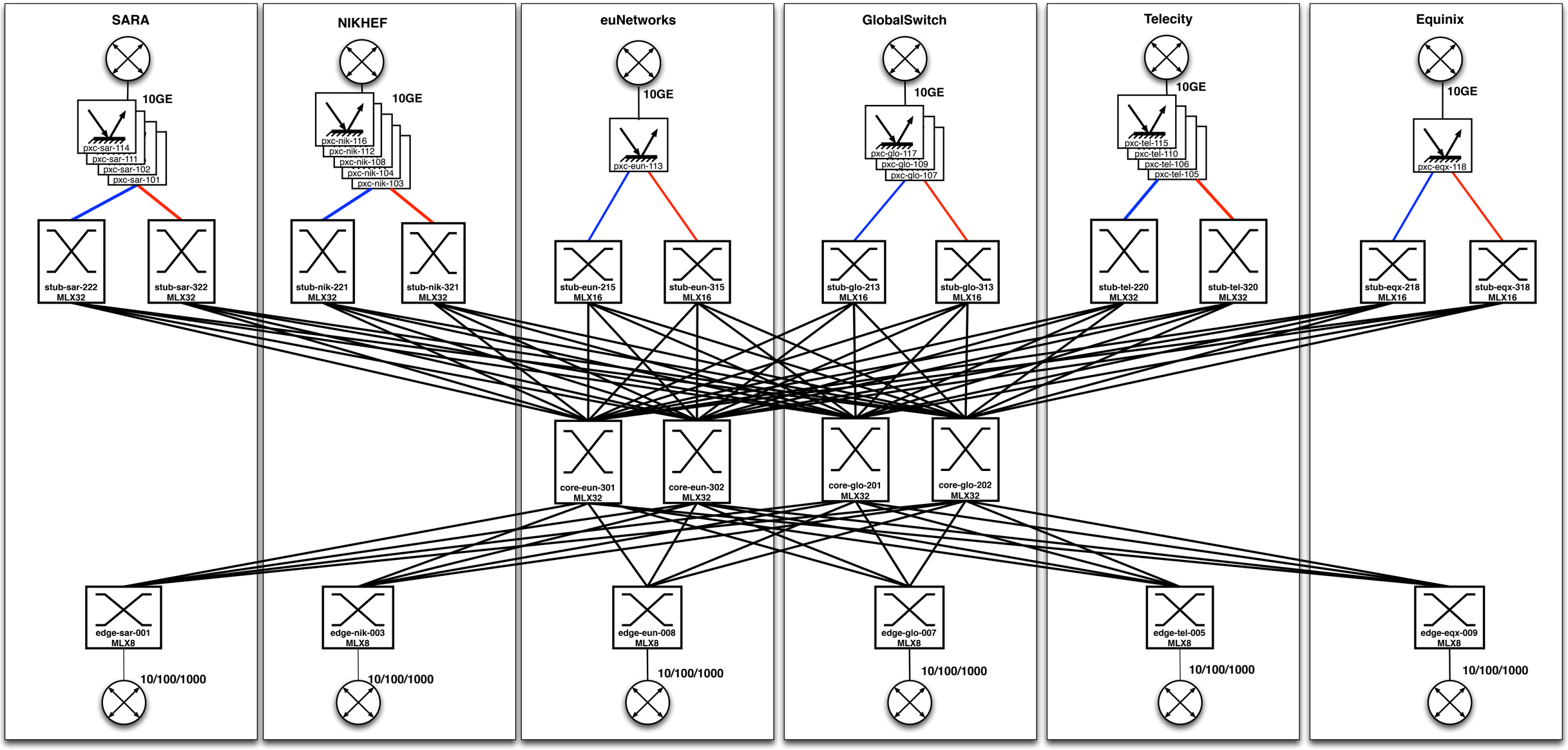
- Core switches (MLX32, 128 10GE line rate) fully utilized
  - Limits ISL upgrade
  - Summer 2009 no substantial bigger switches on the market
- Platform failover introduces short link-flap on **all** 10GE customer ports. In few (but increasing) cases this leads to BGP flapping
  - With more and more 10GE customer ports impact on overall platform stability becomes larger and larger
- 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

Requirements

- Scale the core to at least double amount of ports (Q2/3 2009)
- Keep resilience in platform and 10GE access but reduce impact on failover.
- Increase amount of 10GE customer ports on access switches
  - More local switching
- Migrate to single architecture platform
  - Reduce management overhead
- Use future proof (3 to 5 years) hardware that allows upscaling to high-density 10GE (2010) and 40/100GE (end 2010, early 2011)



Complete MPLS/VPLS topology

AMS-IX version 4

AMS-IX version 4

Overview

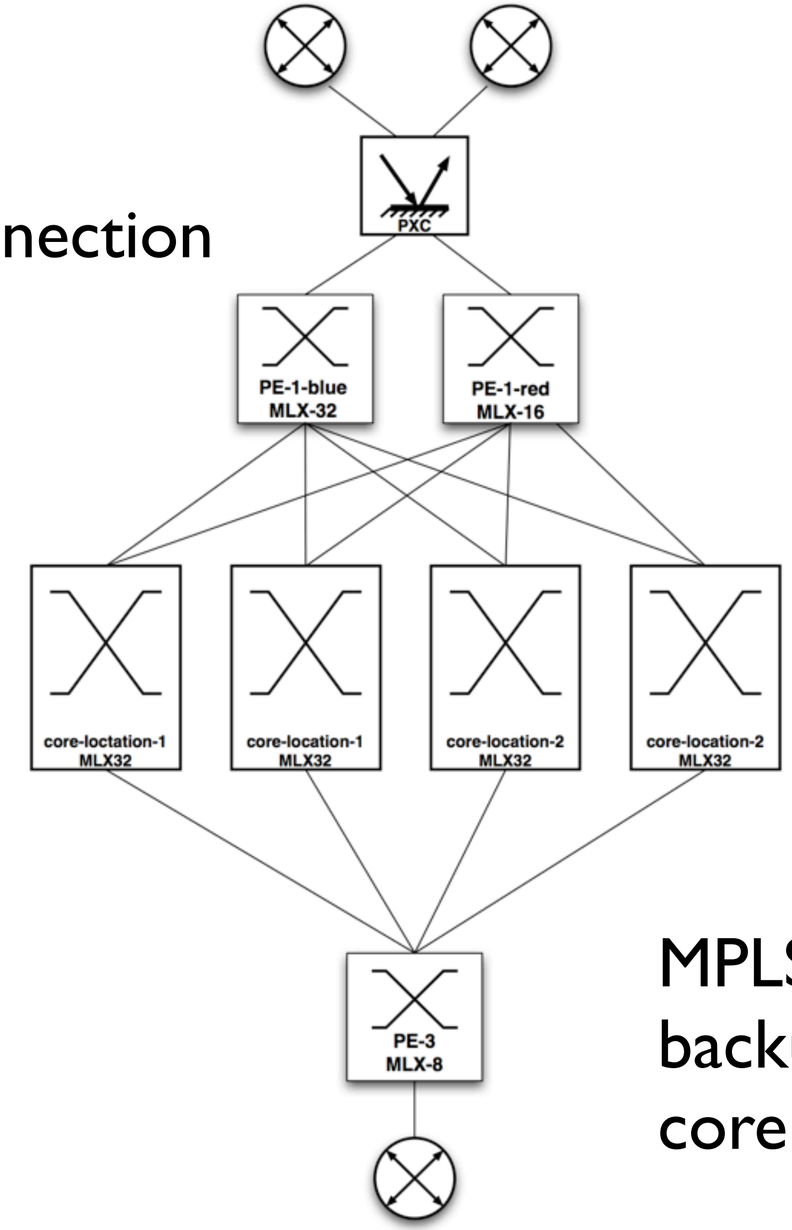
- MPLS/VPLS-based peering platform
- Scaling of core switches by adding extra switches in parallel
  - 4 LSPs between each pair of access switches
    - Load balancing of traffic over 4 LSPs between each pair of access switches
- Retain 10GE access switch resilience
  - Keep 10GE customer connection on PXC
  - No need for complete platform failover anymore
    - Local impact only (single pair of access switches on a site)

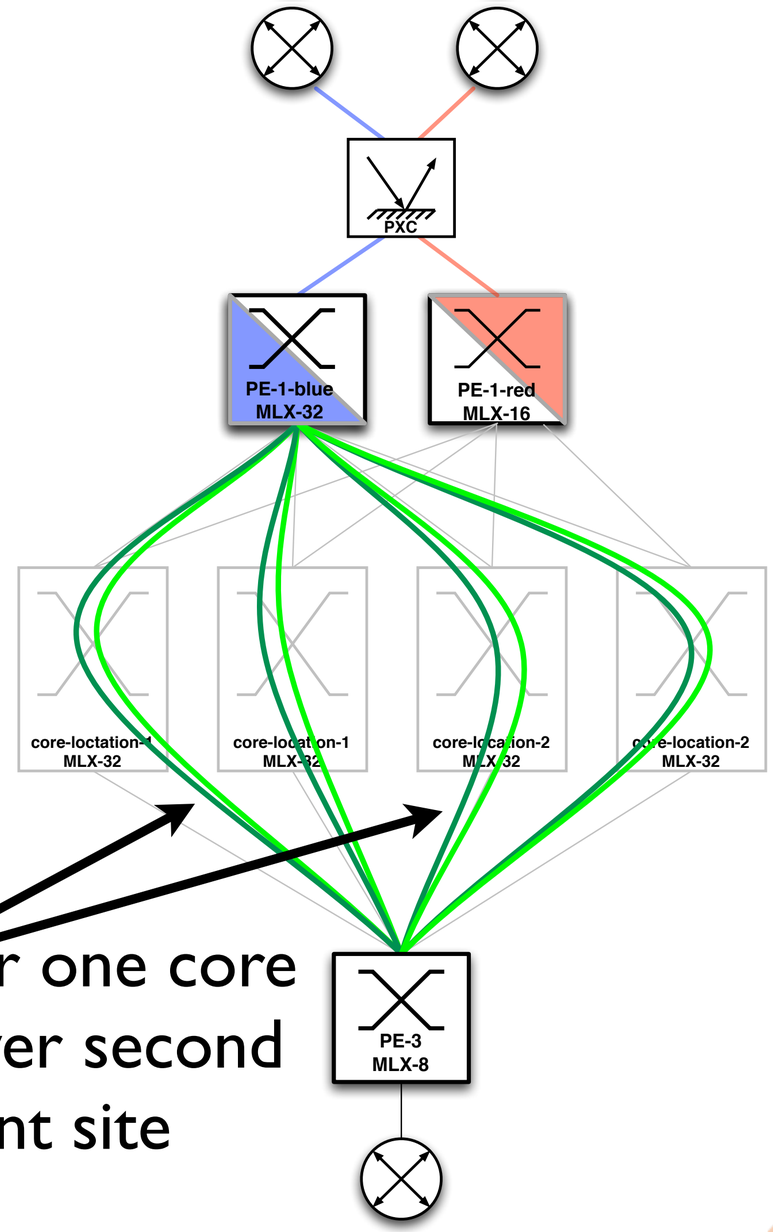
AMS-IX version 4

Characterization

- OSPF
  - BFD for fast detection of link failures
- RSVP-TE signalled LSPs over predefined paths
  - primary and secondary (backup) paths defined
- VPLS instance per VLAN
  - Static defined VPLS peers (LDP signalled)
  - Load balanced over parallel LSPs over all core routers
- Layer 2 ACLs instead of Port Security
  - Manual adjustment for now

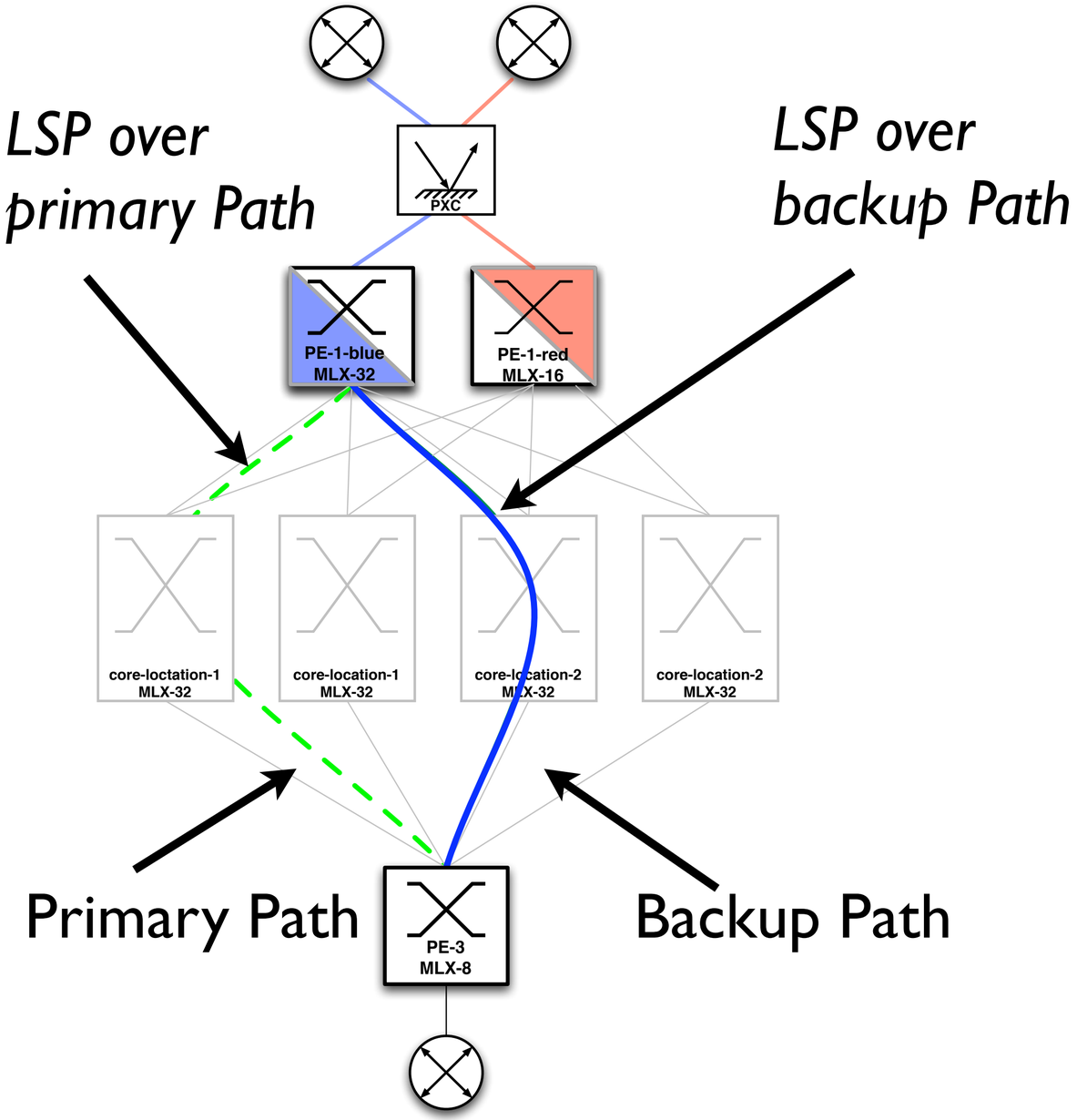
Physical Interconnection

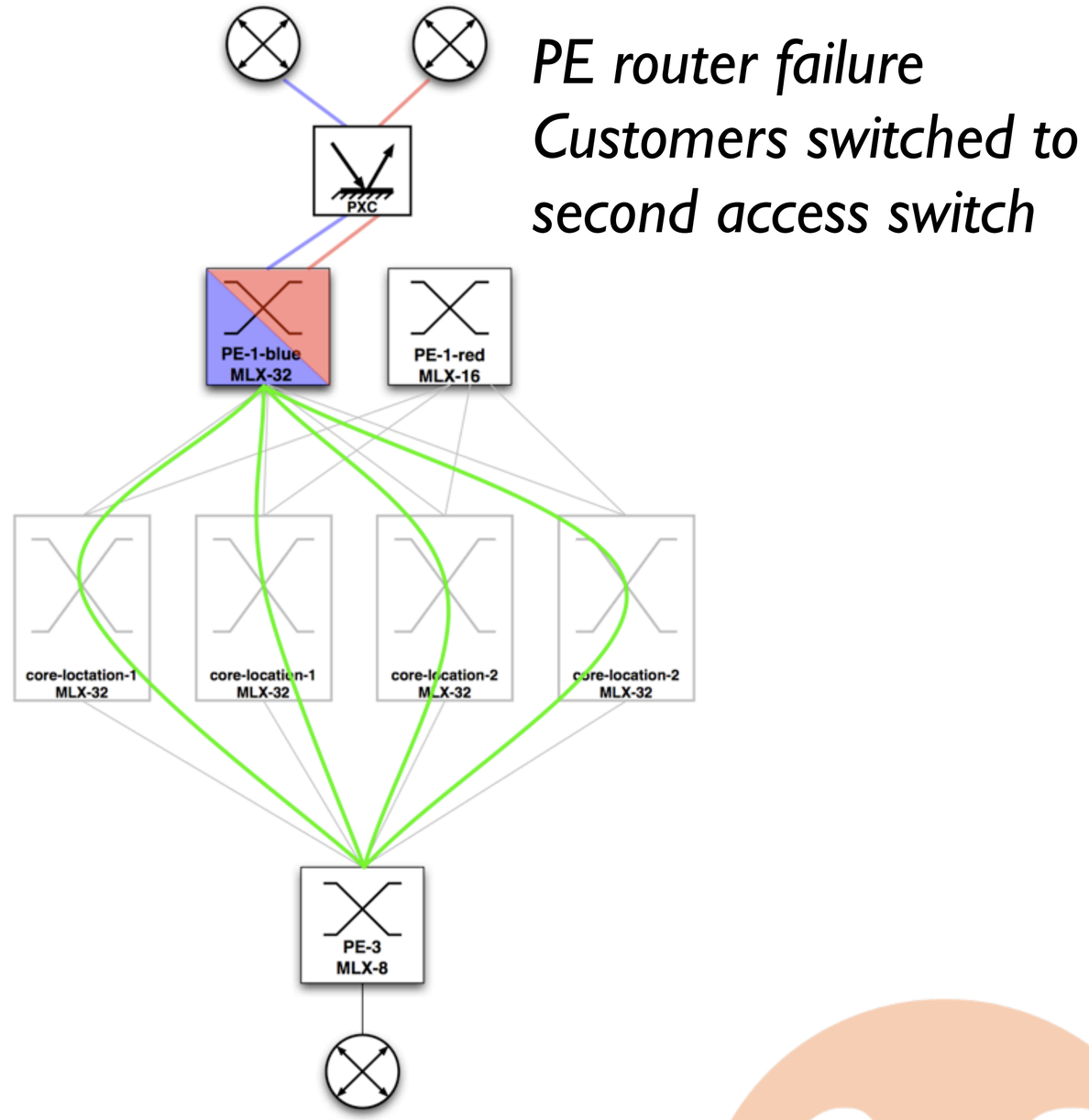




MPLS path over one core
backup path over second
core on different site

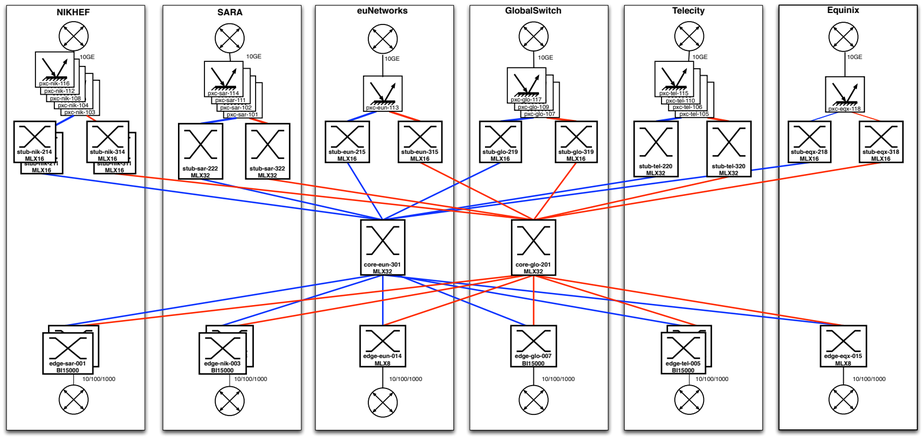
MPLS/VPLS setup



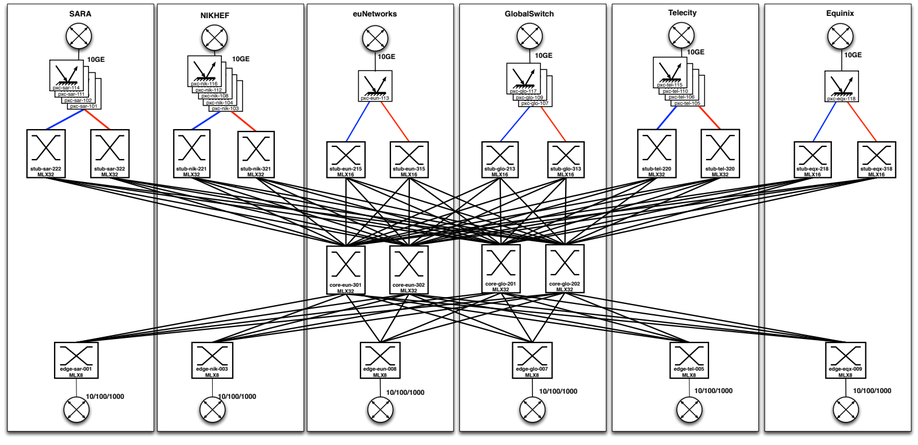


MPLS/VPLS setup

Resilience

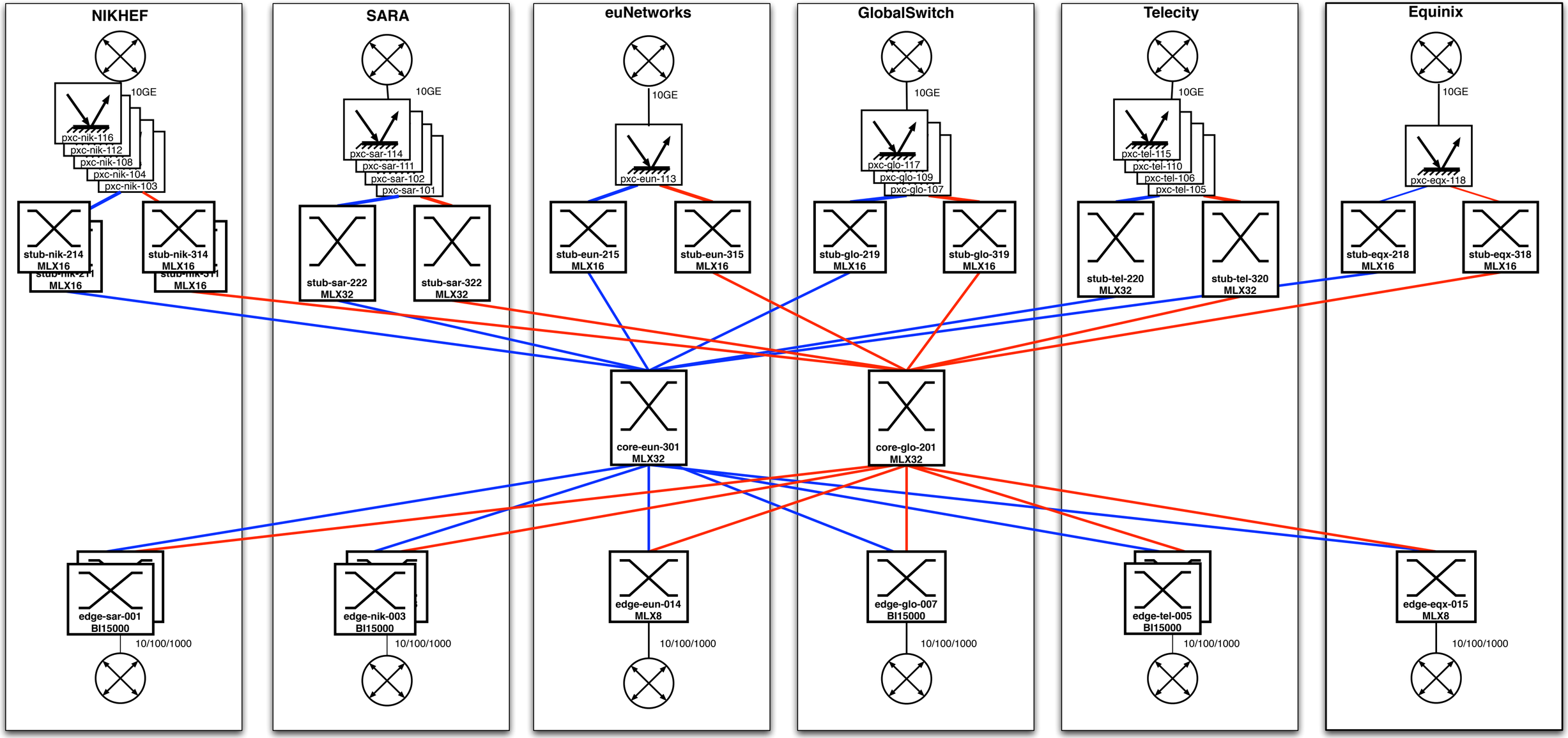


?



AMS-IX v3 to v4 migration

How did we do the platform migration?



Migration steps: Initial situation

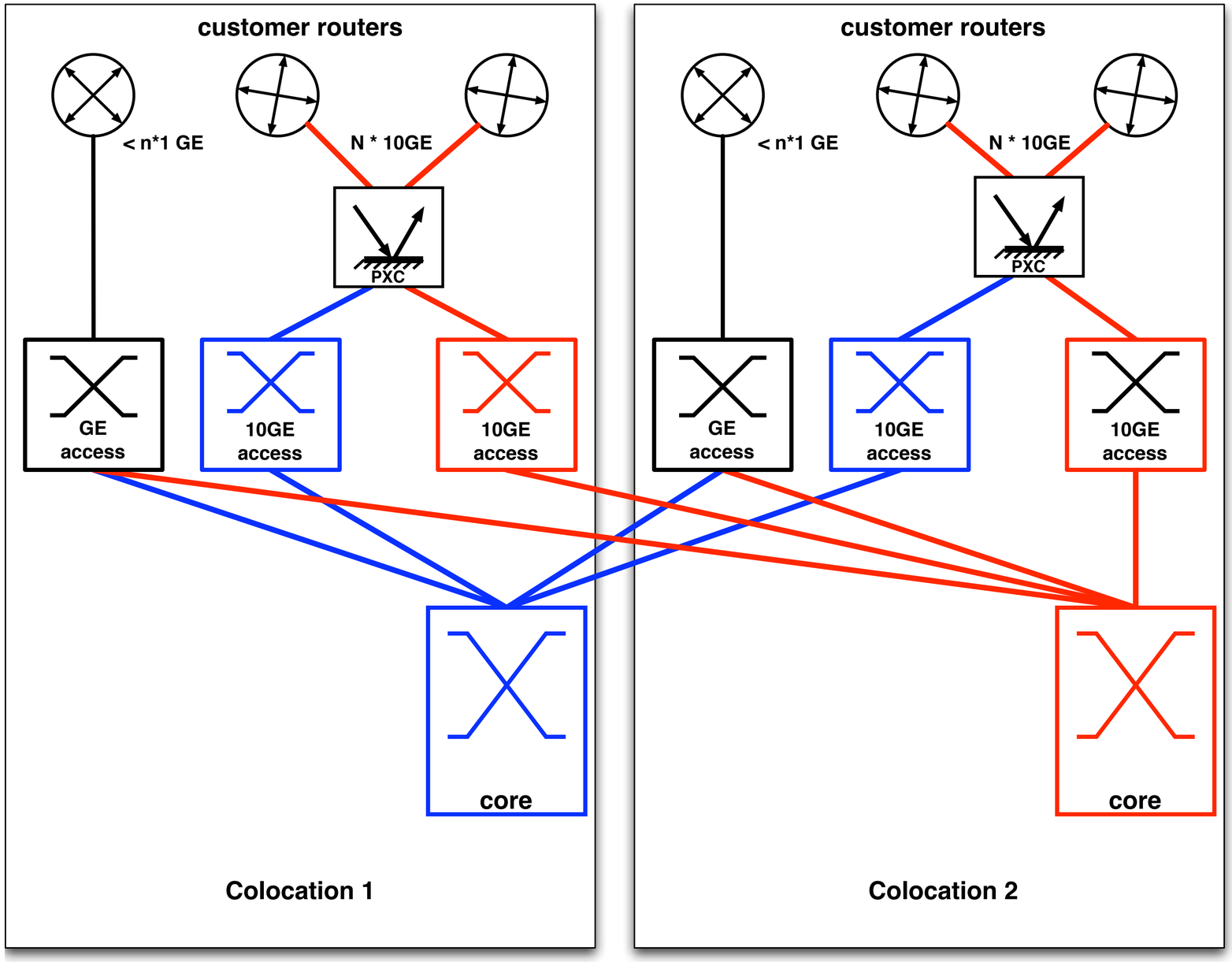
AMS-IX v3 to v4 migration

Platform Migration

Preparation

- Build new version of PSCD (Photonic Switch Control Deamon)
  - No VSRP traps but LSP state in MPLS cloud
- Develop configuration automation
  - Describe network in XML, generate configurations from this
- Move non MPLS capable access switches behind MPLS routers and PXC as a 10GE customer connection
- Upgrade all non MPLS capable 10GE access switches to Brocade MLX hardware
- Define migration scenario that would have no customer impact

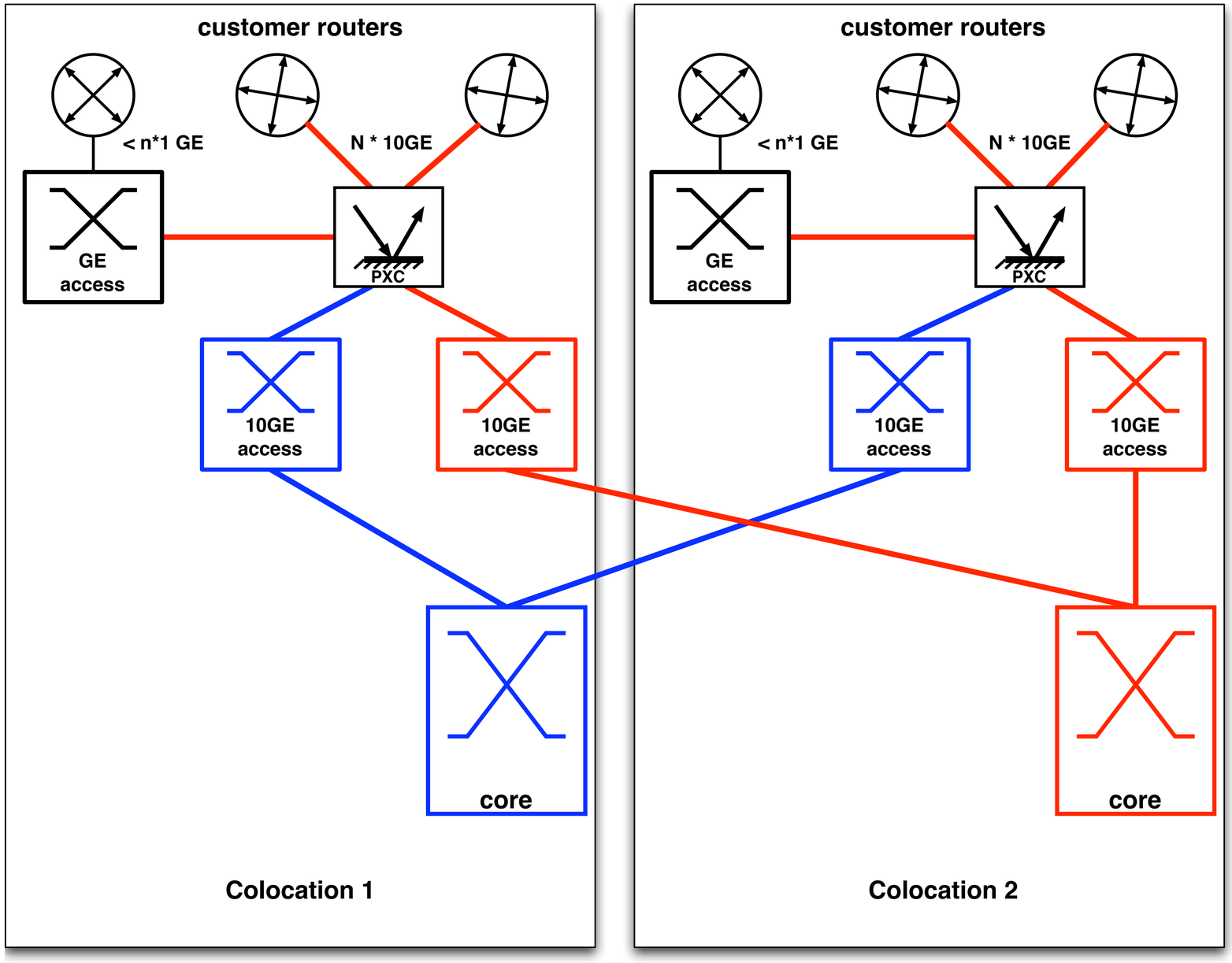
- 2 Co-location sites only for simplicity
- Double L2 network
- VSRP for master slave selection and loop protection



Migration steps: Initial situation simplified

AMS-IX v3 to v4 migration

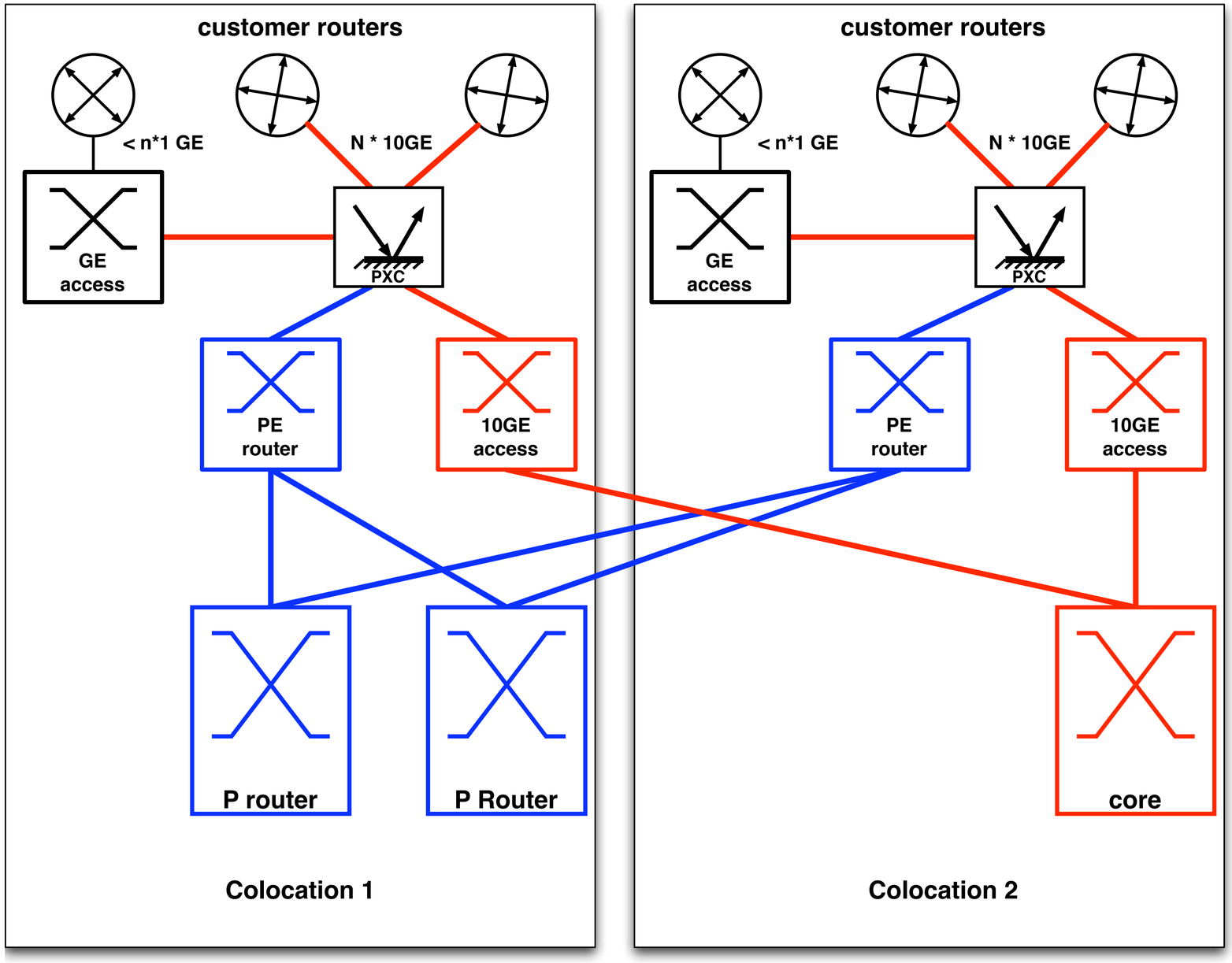
- Not possible to connect GE access switch to both MPLS/VPLS cloud and basic L2 network



Migration steps: move GE access behind PXC

AMS-IX v3 to v4 migration

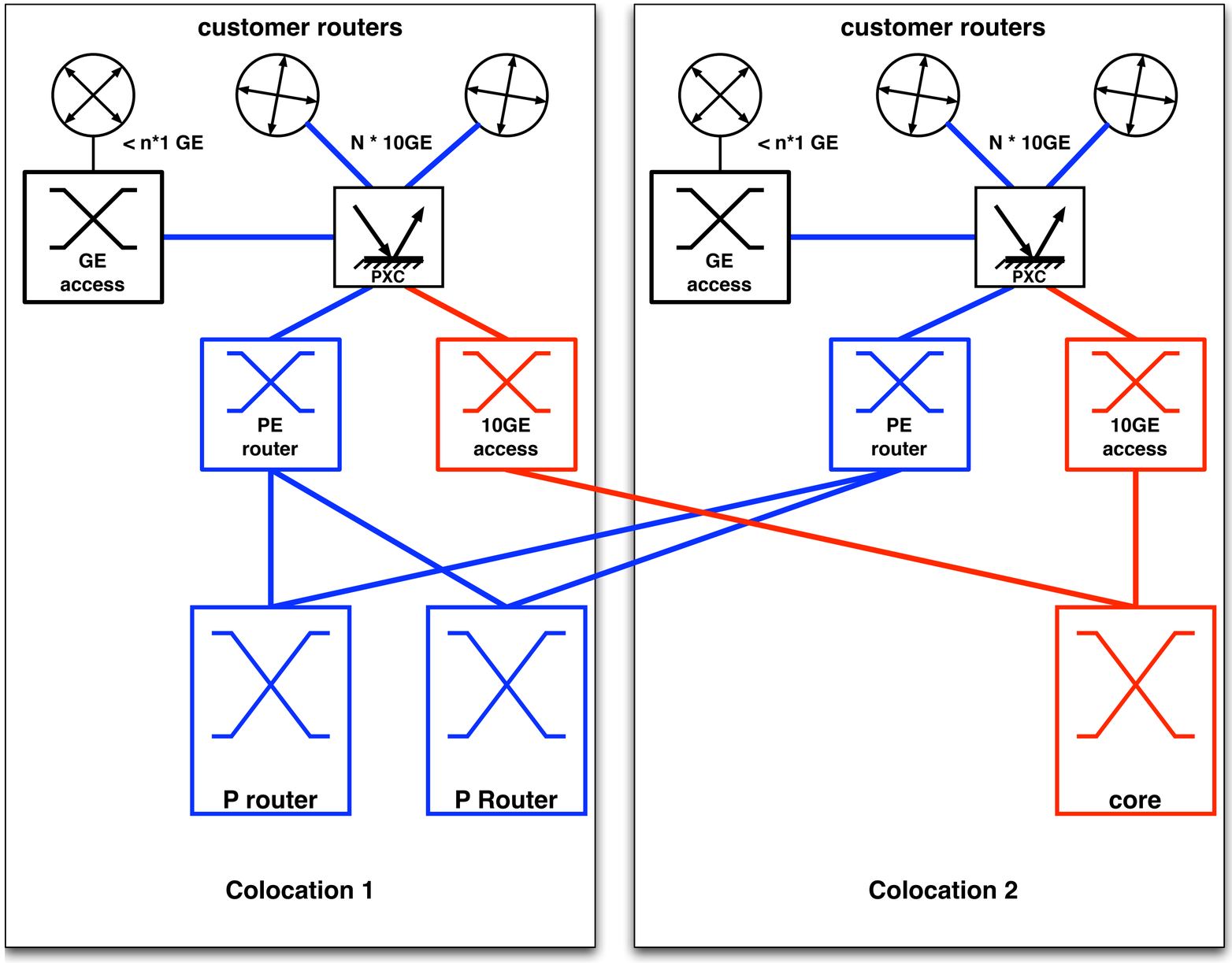
- Production on L2 network (red)
- Migrate blue network to MPLS/VPLS
  - Traffic between two PE routers load balanced over 2 LSPs, one over each P router
- Test functionality and connections using test traffic sent by Anritsu traffic generators



Migration steps: Migrate one half to MPLS/VPLS

AMS-IX v3 to v4 migration

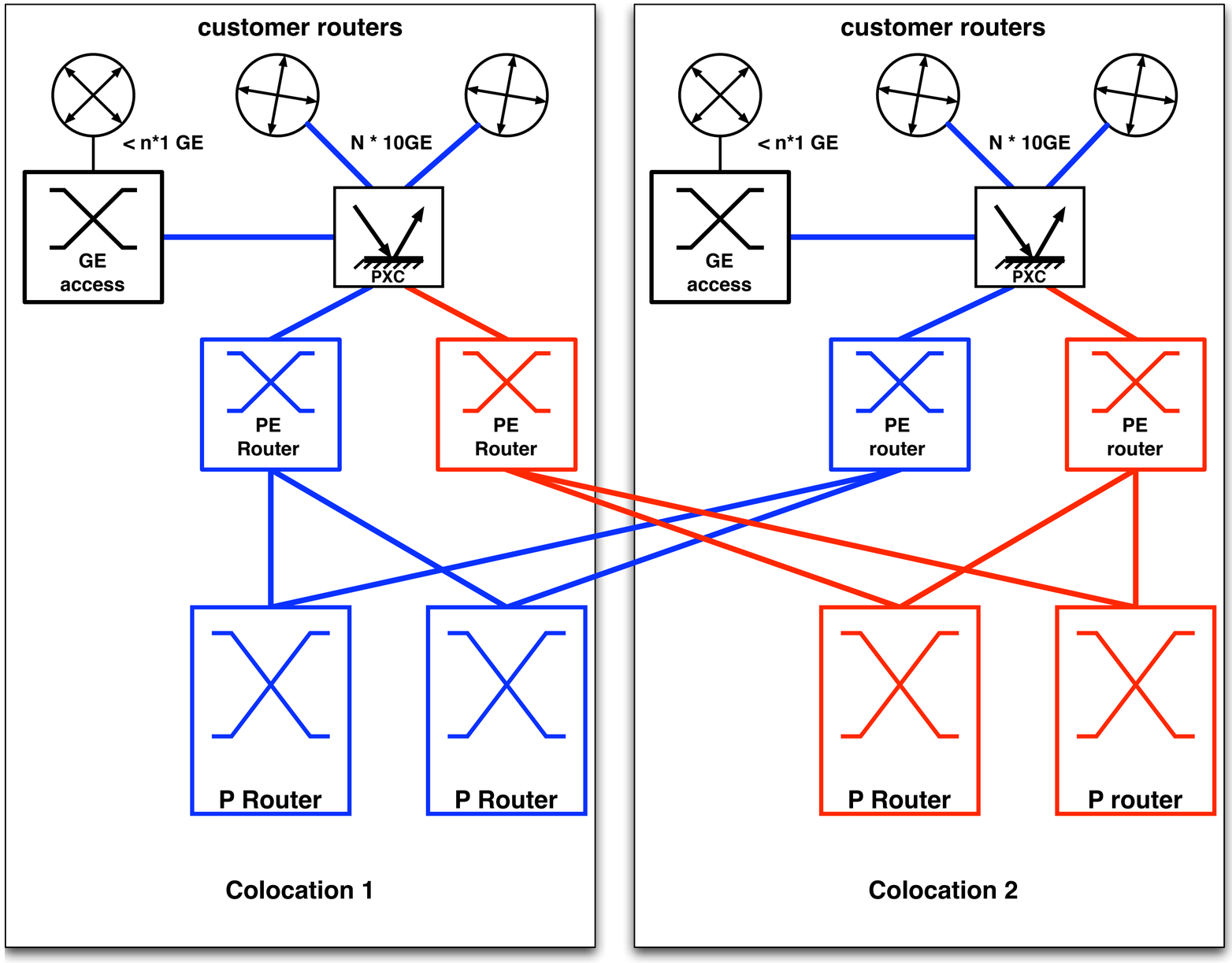
- Move production traffic to MPLS/VPLS cloud
  - Use PXCs for failover
  - New PSCD
- Run production on MPLS/ VPLS cloud for 6 weeks



Migration steps: Production on MPLS/VPLS, L2 backup

AMS-IX v3 to v4 migration

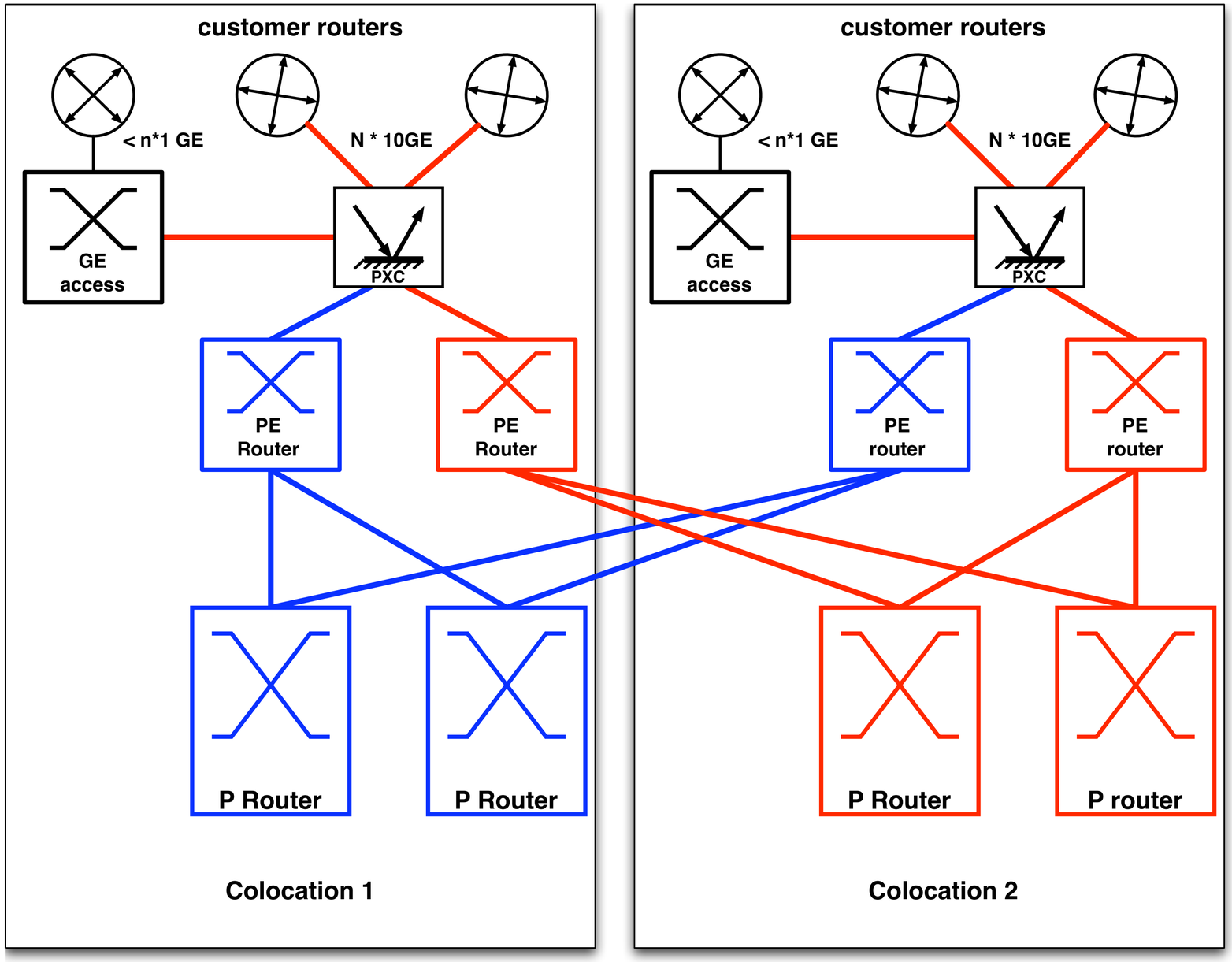
- Migrate second half of the platform to MPLS/VPLS
- Test functionality and connections using test traffic sent by Anritsu traffic generators



Migration steps: Two MPLS/VPLS platforms

AMS-IX v3 to v4 migration

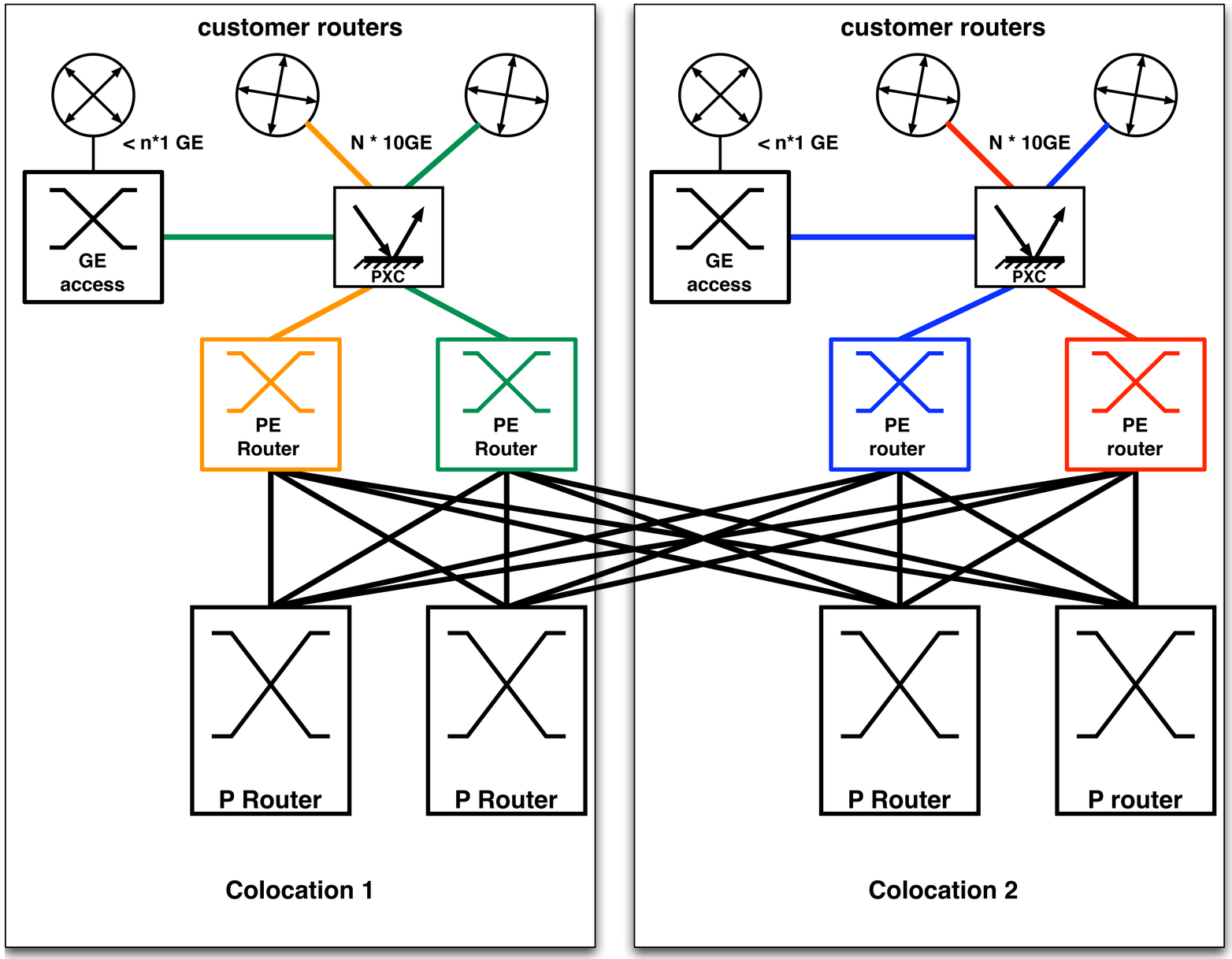
- Move production traffic to red MPLS/VPLS cloud using the newly developed version of PSCD to manage the PXCs
- Still two separate networks, both MPLS/ VPLS based



Migration steps: production on second MPLS/VPLS platform

AMS-IX v3 to v4 migration

- All PE routers connected to all P routers
  - Between each pair of PE routers, 4 LSPs. One over each P router
  - Traffic between each pair of PE routers load balanced over the 4 LSPs
- 10GE customer connections distributed over local PE routers
  - Resilience in 10GE customer connection to local PE router by means of PXCs



Migration steps: integration to single MPLS/VPLS cloud

AMS-IX v3 to v4 migration

Migration - Conclusion

- Traffic load balancing over multiple core switches solves scaling issues in the core
- Increased stability of the platform
  - Backbone failures are handled in the MPLS cloud and not seen at the access level.
  - Access switch failures are handled by PXC for a single pair of switches only and not the whole platform
- Upscaling access switches to Brocade MLX32 allows for higher access port density

Operational Experiences

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

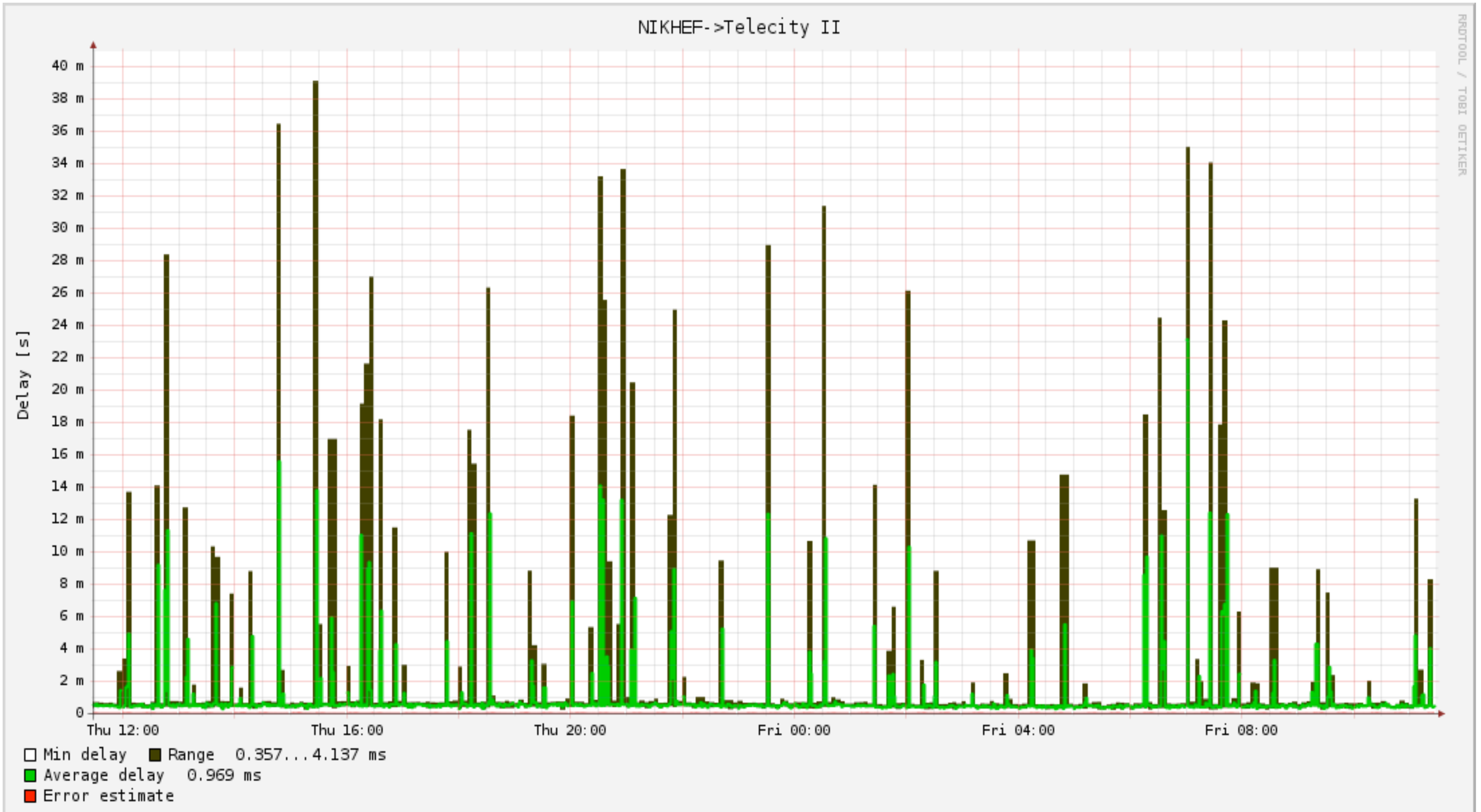
Operational experience

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

Questions ?