



U.S. DEPARTMENT OF
ENERGY

Office of
Science

100GE Upgrades at FNAL

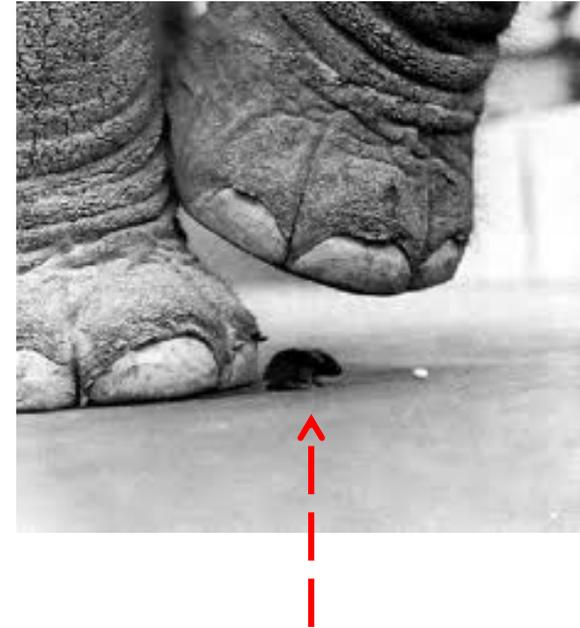
Phil DeMar; Andrey Bobyshev

CHEP 2015

April 14, 2015

FNAL High-Impact Traffic Isolation Philosophy

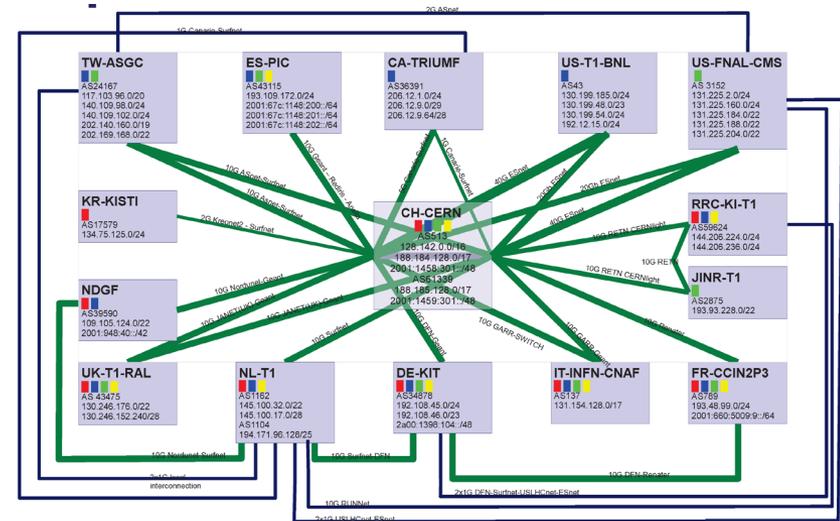
- If feasible, science data traffic kept logically separate:
 - Optimal performance likely over WAN science data network paths
 - Easier to target internal LAN upgrades on high-impact science needs
 - May facilitate more flexible perimeter security models
 - More limited interaction with sensitive or interactive traffic:
 - Don't want this to be your users audio/video applications
- Routed IP network path still supports our general internet traffic
 - Also a significant percentage of our science data traffic as well



FNAL's Science Data WAN Network Paths

- LHCOPN:

- T0 <-> T1 data
- (Most T1 <-> T1 data movement
- Supported by three virtual circuits:
 - Primary – 40Gb/s guaranteed
 - Secondary – 30Gb/s (failover)
 - Tertiary – 10Gb/s (failover II)



(Obligatory LHCOPN figure)

- LHCONE

- Private routed virtual network
- FNAL LHCONE traffic via ESnet, including trans-Atlantic stuff
- US LHC universities starting to migrate to ESnet for LHCONE

- Legacy pt-to-pt virtual circuits (static)

- Six US Tier-2s have circuits to FNAL

Current Perimeter Architecture

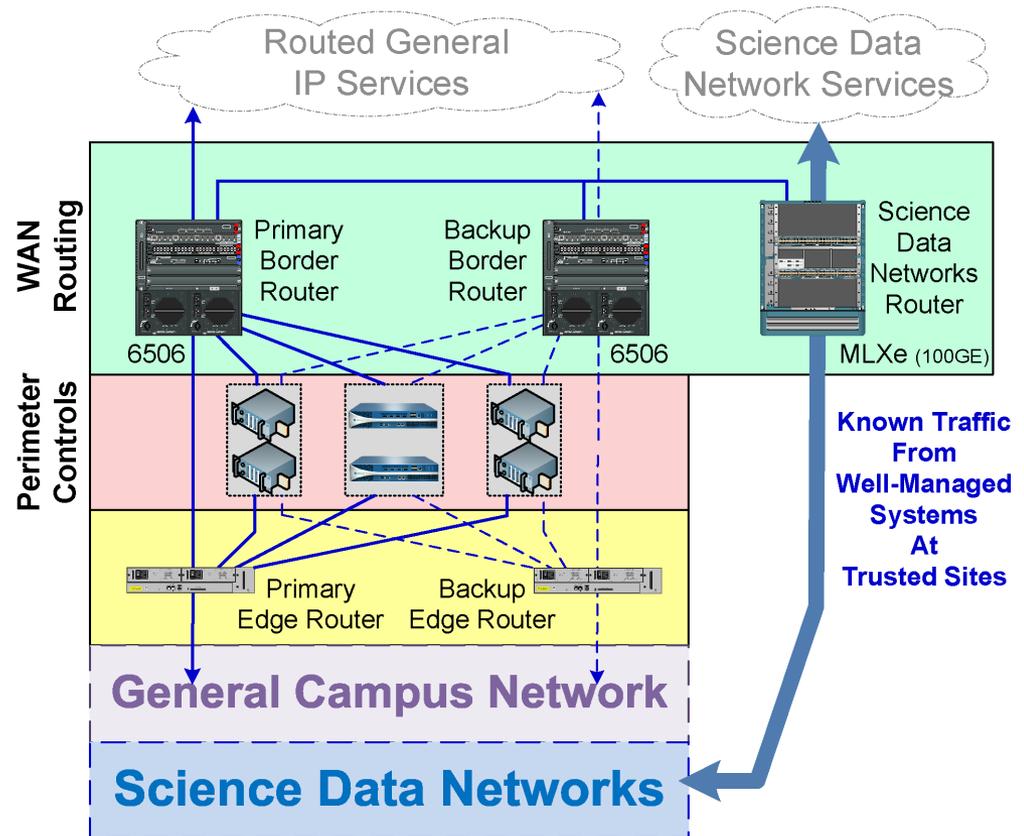
- Science data network services supported with separate infrastructure:

- Separate border router
- Bypass path(s)
- Security model:

Known Traffic from well-known systems at trusted sites

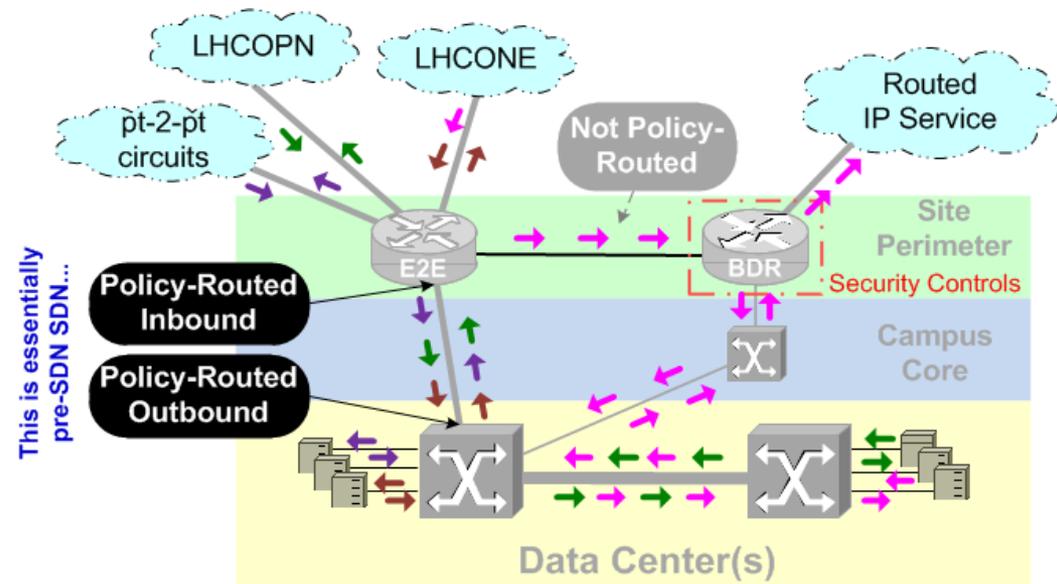
- General routed internet traffic via primary border router:

- Fails over to secondary (not load-balanced...)
- Some science data via routed IP



How Science Network Traffic Is Isolated

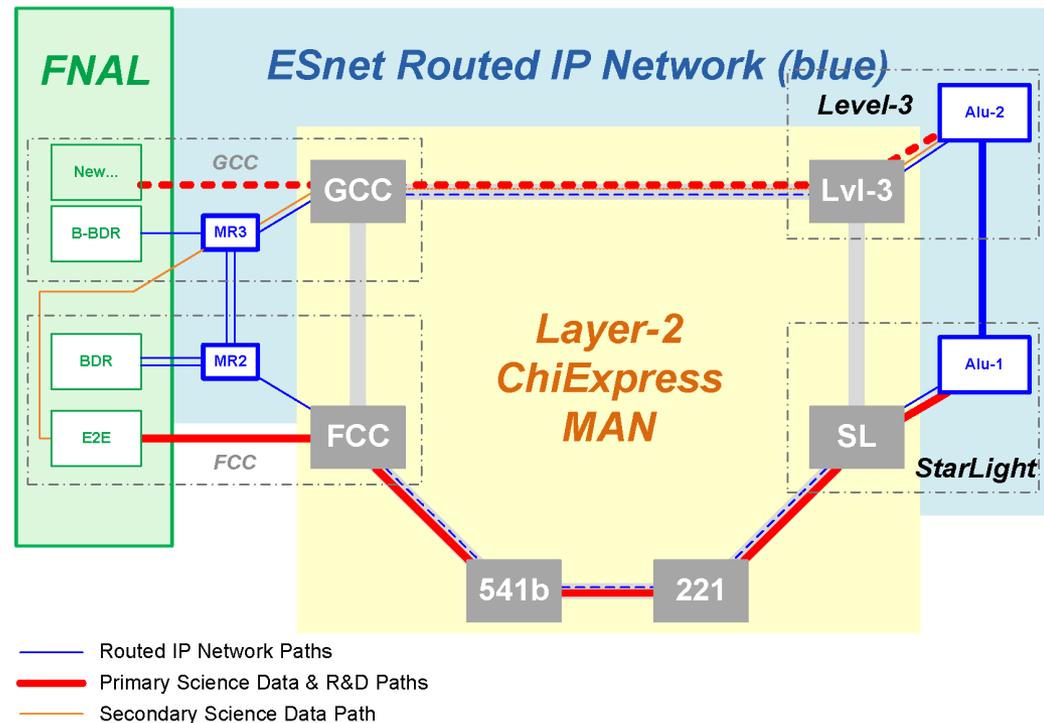
- Keyed on Policy-Based Routing (PBR)
 - Essentially source/destination ACLs
 - Satisfies security policy for bypass traffic
 - Ingress on science data network border router
 - Egress for bypass from data center LANs (ie., CMS Tier-1)
- Object tracking & SLA monitoring check that PBR path is functional
- Science network non-PBR traffic forwarded to routed IP path
 - May create path asymmetry issues



Now On To 100GE Upgrades

FNAL WAN Access – ChiExpress MAN

- ChiExpress:
 - ESnet local loop access for FNAL (and neighboring ANL)
 - 100GE-based technology
 - Full geographic redundancy
 - Reasonable channel reconfiguration (~1hr)
- FNAL channels:
 - One 100GE
 - 2nd 100GE in place but not in use yet
 - Four 10GE channels
 - Routed IP
 - Backup science data path



New Perimeter Architecture

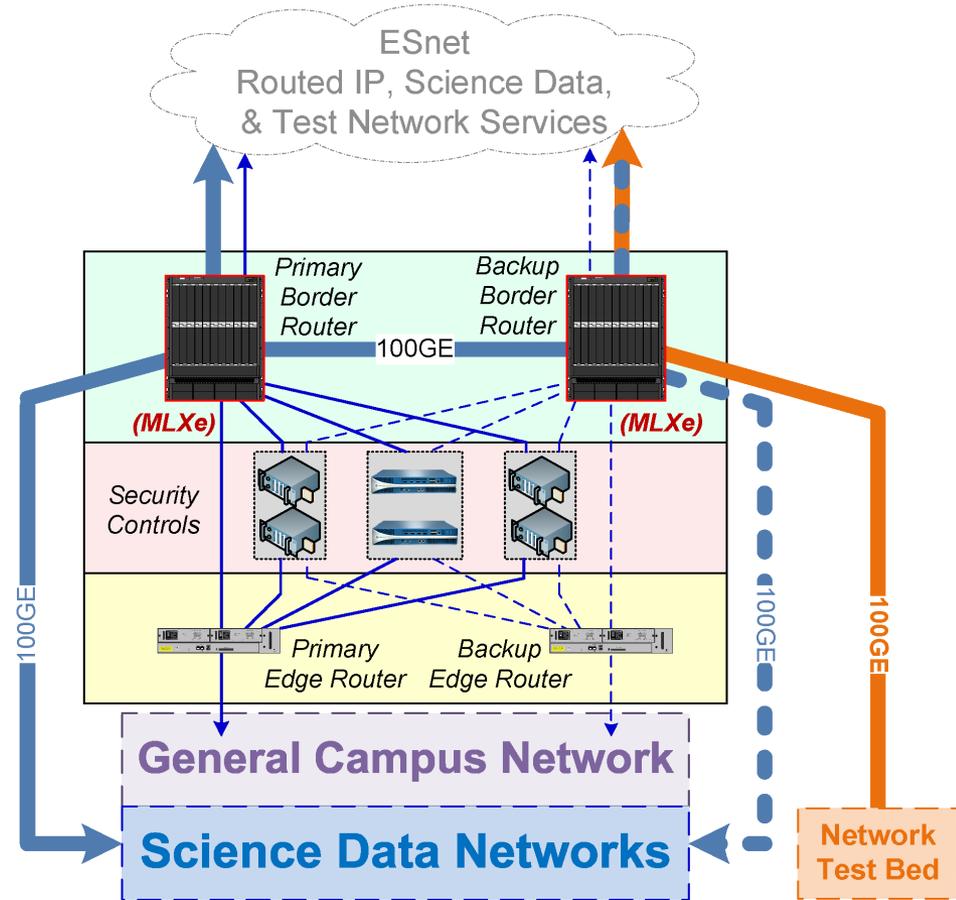
- Consolidation to two border devices:

- Cost-driven
- Based on Brocade MLXe's
- In progress:
 - 4-6 month cutover

- Science traffic still kept logically separate

- Routed IP traffic still based on default & failover

- Network R&D becoming a high-impact traffic factor

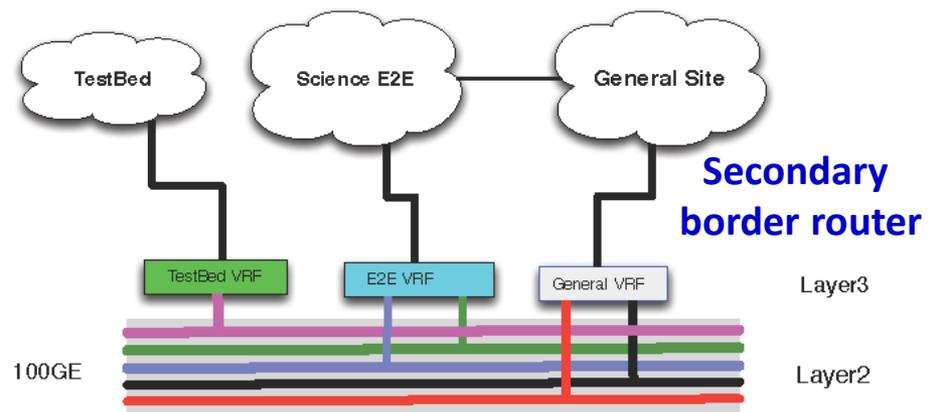
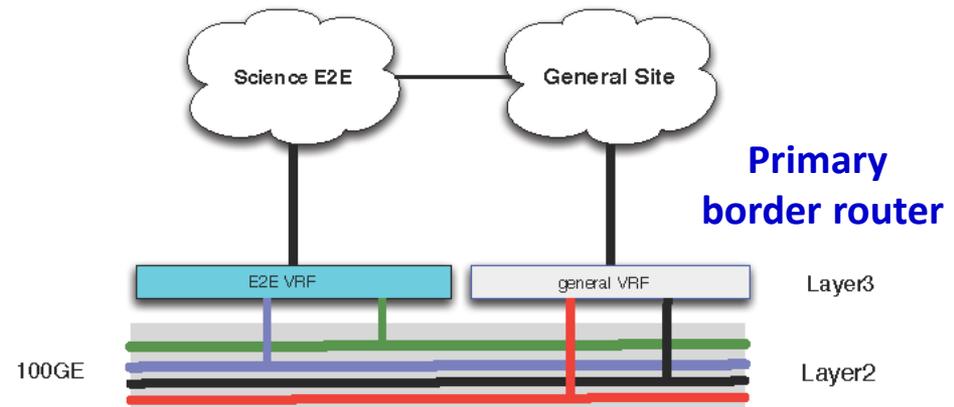


WAN Traffic Distribution Strategy

- One 100GE channel for production science data traffic
 - Circuits (including LHCOPN) and LHCONE traffic...
- Second 100GE for network research & failover of science data traffic (LHCONE)
 - Will reexamine/readjust this strategy when science data traffic starts pushing up toward 100GE:
 - Optimally, just upgrade that channel to 2x100GE
 - Or load-balance LHCONE traffic across existing 100GE channels
 - Generally, circuit-based traffic falls back on routed IP path
- Migrating routed IP traffic over to 100GE channel(s) is a goal
 - Starting to saturate current 2x10GE capacity
 - Support for >10GE flows
 - But no specific time table yet

VRFs

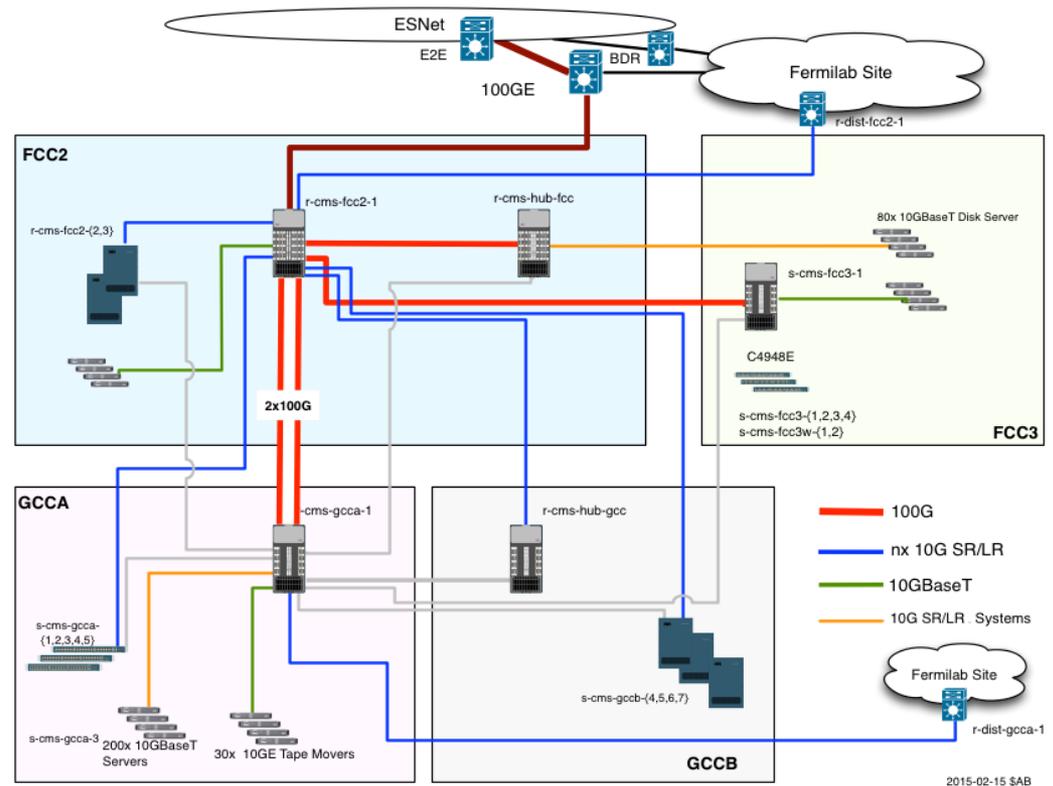
- Separation of science data & general IP network routing done by virtual routing & forwarding (VRF) technology:
 - Shared layer-2 (SVIs)
 - Separate layer-3
 - Preserves current security model
- Current implementation problem:
 - Only default MLXe VRF has PBR capability
 - Looking into work-arounds



100GE In the Data Centers

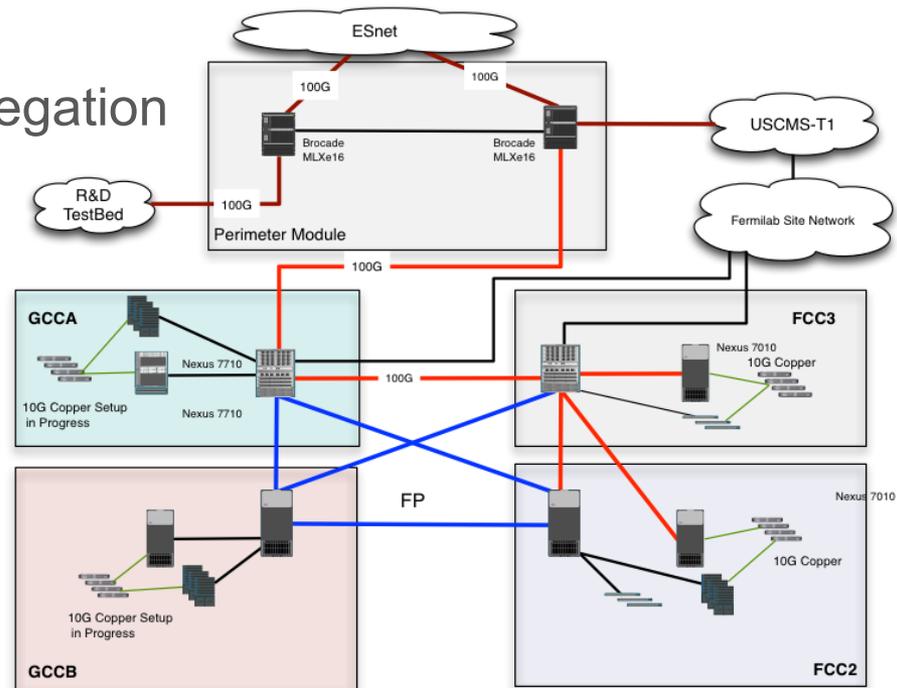
100GE Upgrades – CMS Tier-1

- Interconnect between dual core Nexus 7000's upgraded to 2x100GE:
 - Replaces 16 x 10GE
 - Addtl 100GE links to aggregation switches
- WAN connection upgraded to 100GE
 - Replaces 4 x 10GE
- Wide-scale 10GB-T deployment
 - ~200 systems this year
 - EOR/MOR, not TOR



100GE Upgrades – General Data Center(s)

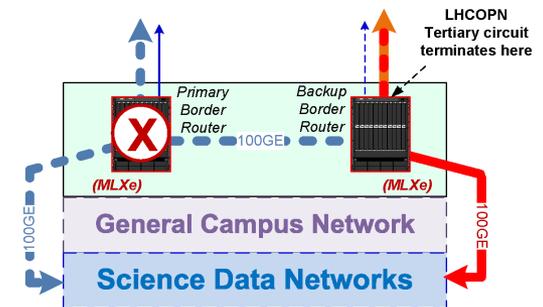
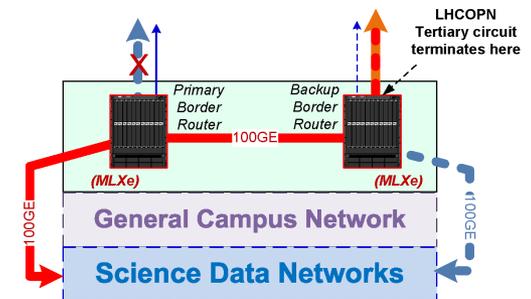
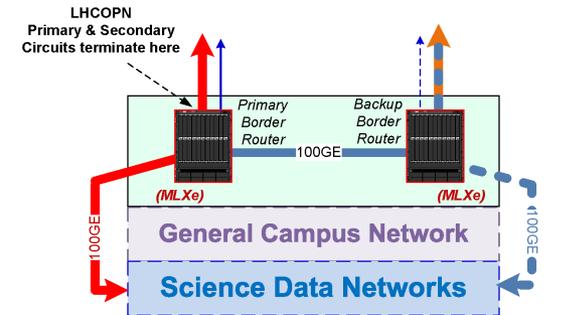
- Interconnect between dual core Nexus 7710's upgraded to 100GE:
 - Replaces 8 x 10GE
 - 100GE link to one aggregation switch as well
- WAN connection upgraded to 100GE
 - Replaces 2 x 10GE
- 10GB-T deployment scaling up
 - ~50 systems this year



SAB 2015-02-09

Resiliency

- 100GE MAN channel up:
 - LHCOPN & LHCONE follow usual paths
 - No PBR changes
- 100GE MAN channel down:
 - LHC tertiary & LHCONE thru other 100GE
 - No PBR changes
- Primary border rtr down:
 - LHC tertiary & LHCONE thru other LAN path
 - PBR multihop tracking reroutes to backup bypass



Future Directions

- Immediate term – finish consolidation into two border devices
- Short term – consolidate science data network traffic:
 - Migrate T1 <->T1 traffic over to LHCONE
 - Phase out legacy static circuits in favor of LHCONE:
 - Obviously requires coordinated actions with remote sites
- Long term – evaluate SDN-based technologies (OpenFlow) to isolate on-site science data traffic
 - Evolution of inter-domain SDN is too unclear now to start planning SDN beyond our border

Questions?