FERMILAB-SLIDES-21-047-CCD
This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

Fermilab Network Architecture and Implementation Experience

Andrey Bobyshev, Workshop Datacentre Network Architectures

June 7 - 8, 2021

Outline

- Introduction of Fermilab's environment
- Modular Network Architecture
  - US CMS Tier 1/ End-to-End circuits
  - General Datacenter/FabricPath
  - Site Interconnect and other modules
- Current and future technologies in use
  - Classical Ethernet, Ethernet Fabrics/FabricPath, IP Fabrics/VXLANs
- Future plans and on-going upgrades

** Slides with the title ended with*  *are meant to be skipped*

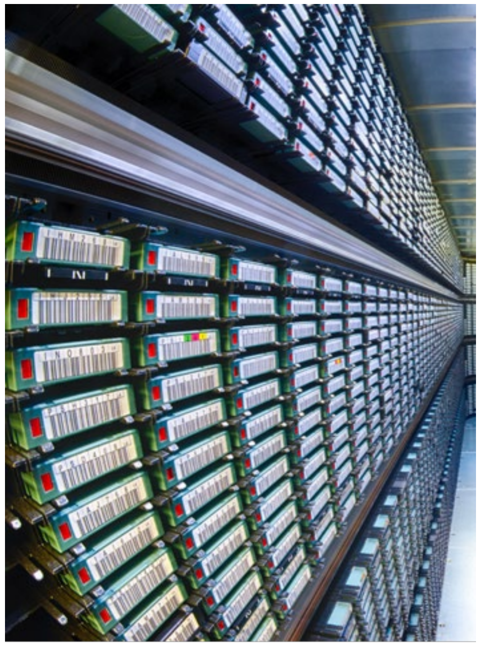
Fermilab’s environment

- Facility:
  - Multiple computing rooms/Two Blds. (1km apart)
  - Multiple stakeholders
  - Dispersed racks space in computing rooms
  - HPC Computing, Storage & Data Movement, Experiment Computing Facilities
- Multiple security zones
  - Two site-wide (Protected and Controlled)
  - Multiple zones for Business Applications module
  - Moving towards deeper networks segmentation for science traffic as well

Facility resources – the total picture

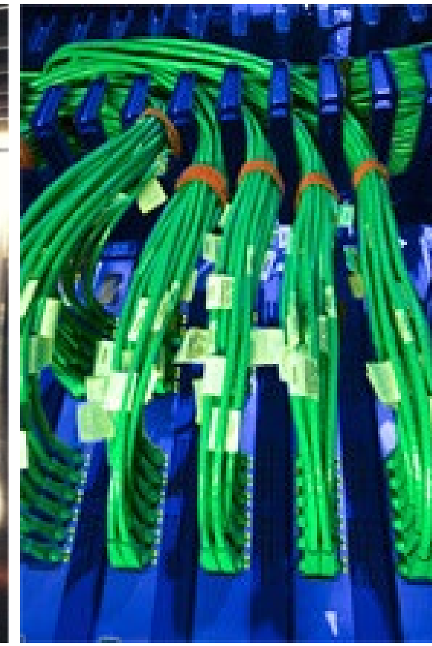
Processing

Tape Storage



Disk Systems

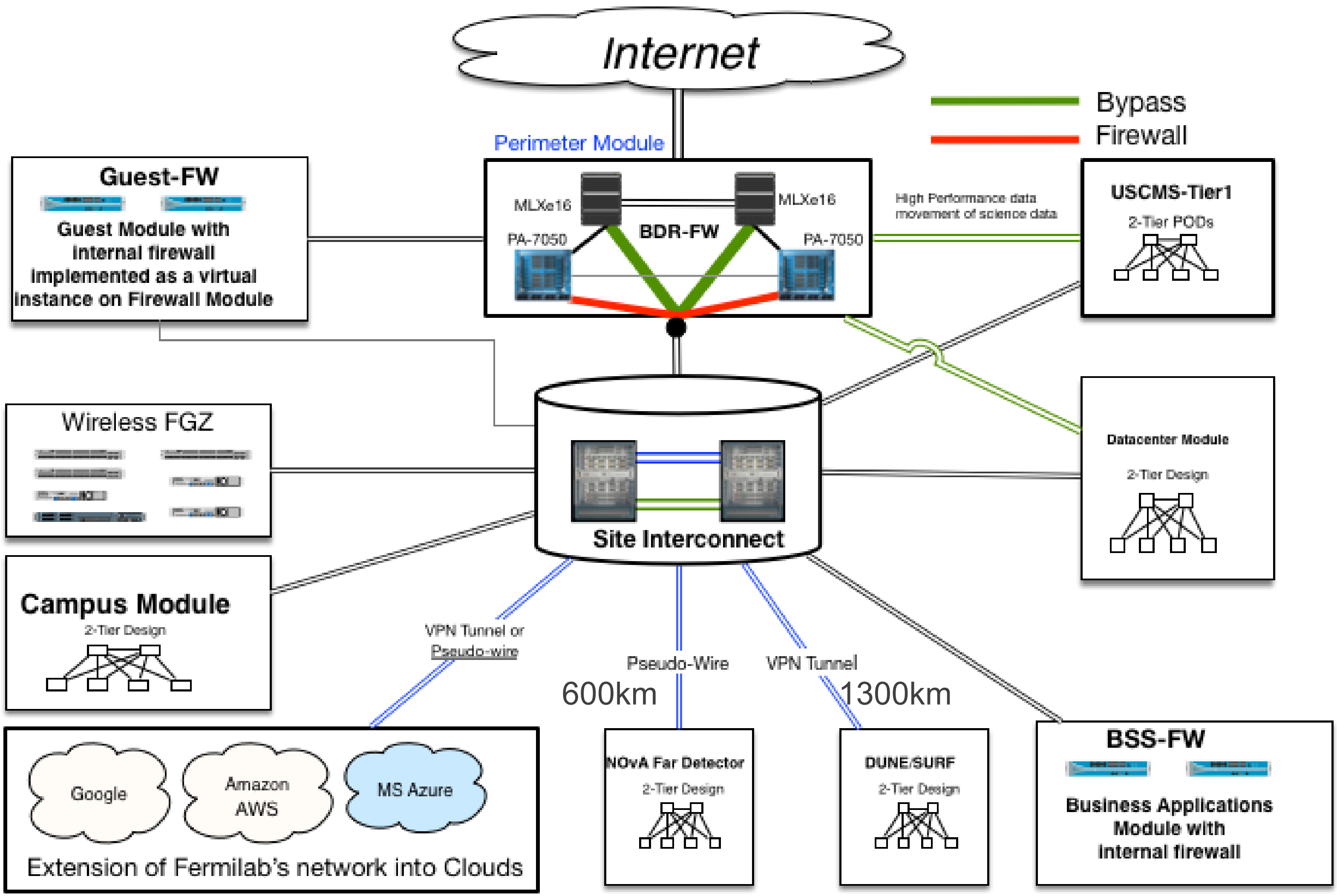
Networks



Network Architectures Overview

- 2-/3- Tier Network Architectures
- Spine and Leaf/ Clos-based networks
- Flat with Any-to-Any connections
- Technologies
  - Classical Ethernet/Spanning Tree Protocol
  - Ethernet Fabrics: FabricPath/TRILL/SPB
  - IP Fabrics/VXLAN
- Fermilab's Network Architecture
  - Modularity: Each of the modules is based on the architecture most suitable for its purpose
  - Separation of Science traffic from other types of traffic
    - Open nature of science data vs business traffic
    - High volume & high data rates

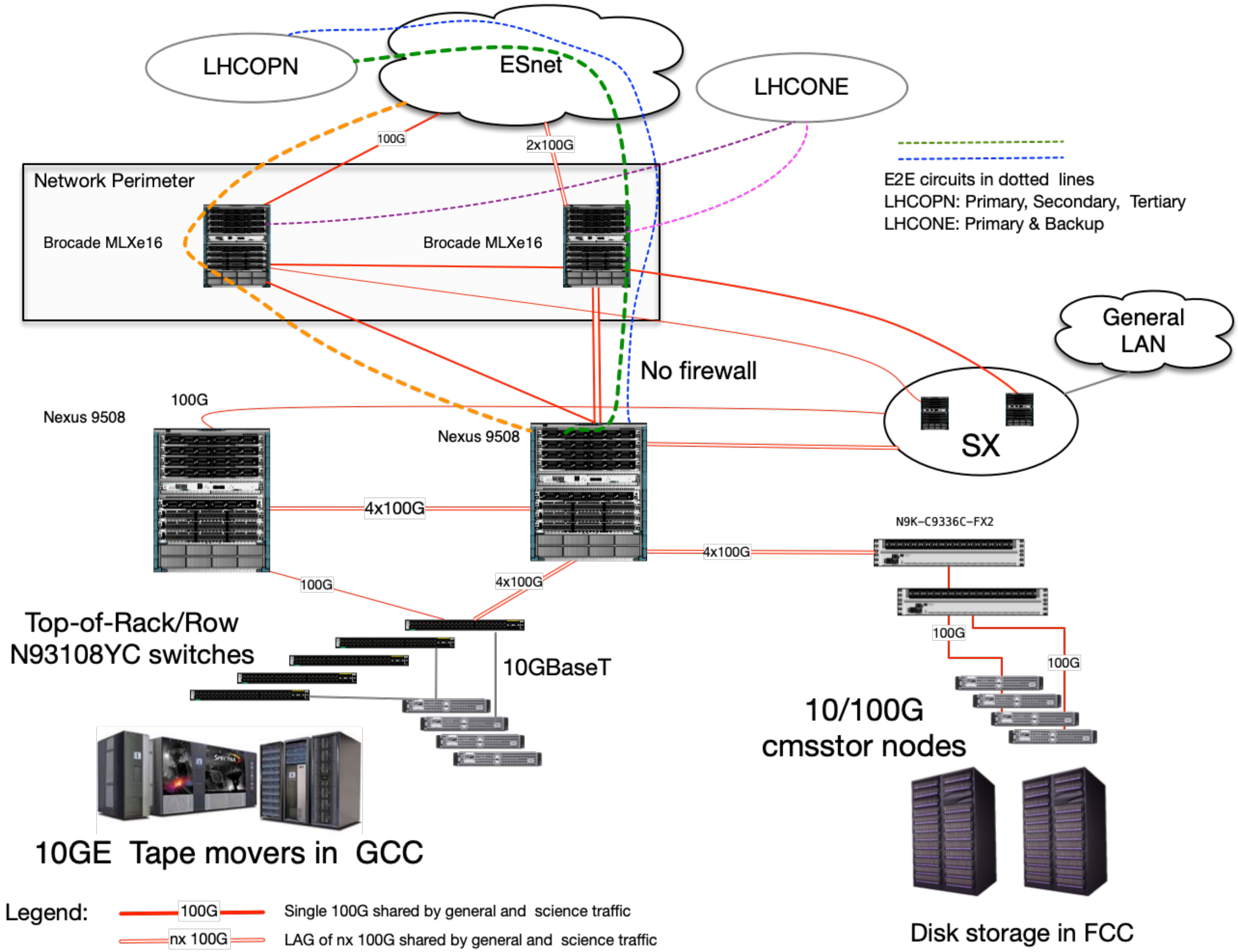
Modular Network Architecture (cont.)



Modular Network Architecture

- Module is a collection of network resources to perform a specific task or tasks, identifiable by groups of VLANs, IP subnets. Network modules are normally based on organizational boundaries, experiment's affiliation and/or geographical locations
- All network modules are connected to a special module called the Site Interconnect which provides network connectivity between all modules, and where all inter-module security policies are implemented
- Modules can also be directly connected to each other, allowing certain traffic to bypass the site-interconnect for more efficient routing and performance

The USCMS-T1 Facility



Fermilab

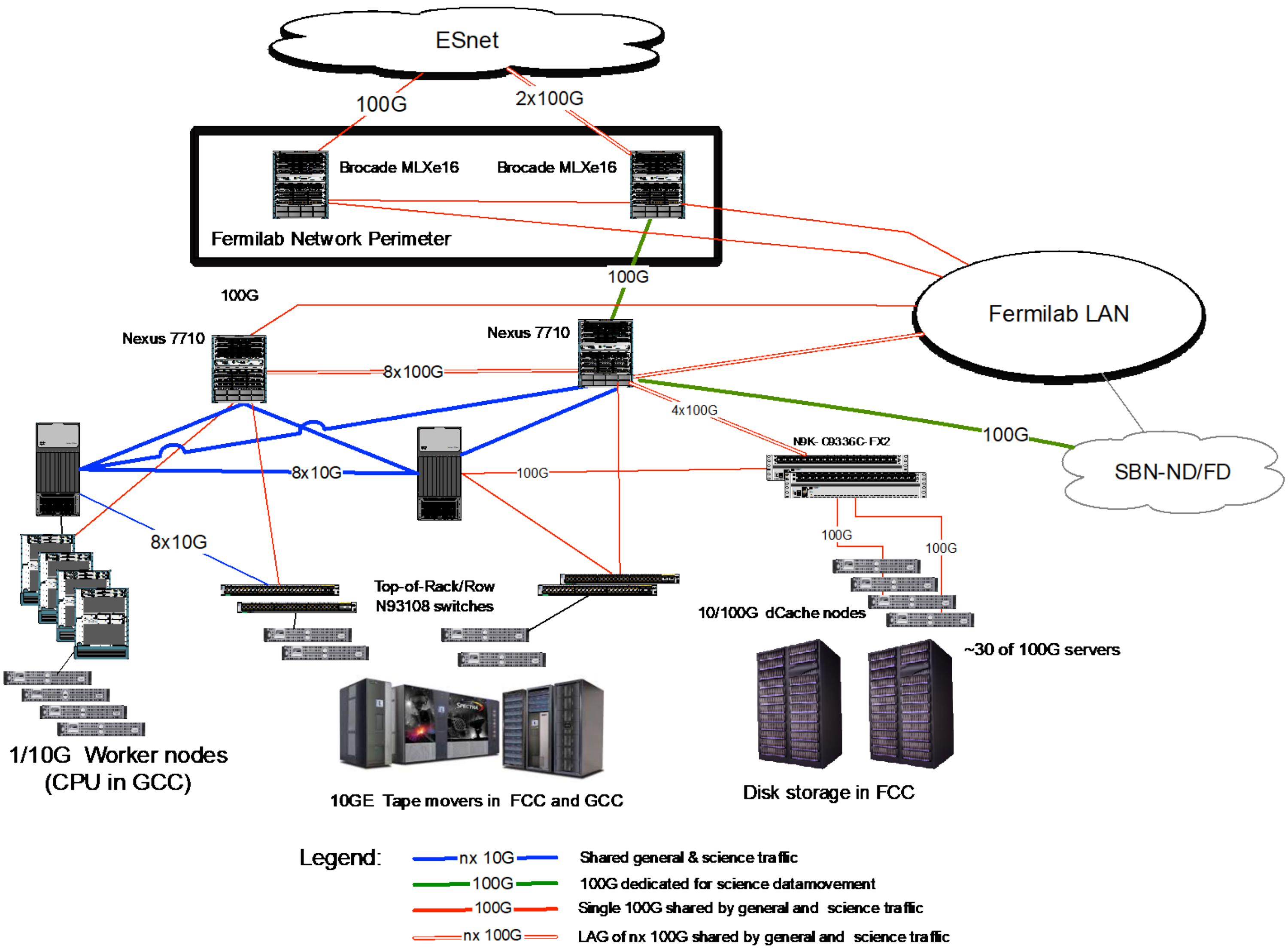
The USCMS Tier1 Facility

- 2-Tier Architecture (Aggregation/Access)
- 2 Cisco Nexus 9508 aggregation switches (10/100G LCs)
  - 4x100G between aggregation switches
  - 2x100G direct link to the Network Perimeter
  - 80G to the Site Interconnect (Why 80G?)
- Access ToR switches (10G & 100G for servers' side)
  - 6x100G uplinks (4:2) & 48-port 10G-BaseT
  - 8x100G uplinks, 24x QSFP-100G-SR4 (server's side)
- End Systems
  - 10G copper
  - 100G QSFP-100G-SX4
  - 1G Legacy servers (4x 6509E with 8x10G, C4948E with 4x 10G)

The USCMS Tier1 Facility/ Technology

- Currently based on classical Ethernet/STP / GLBP/HSRP
- Primary/Secondary Aggregation model /Asymmetric BW
- Dual-protocol: IPv4/IPv6
- Multiple VRFs
  - IPv6-only
  - In-band and out-of-band management
- In-Service Software upgrades
- Past Experience
  - vPC (Virtual Port-Channel): Our experience is not very good
    - Random pair-wise connectivity problems, usually after a long period of normal work
- Future/On-going upgrades
  - Native 400G Ethernet
  - ToR 8x400G, 28-port 100G/ 93600CD-GX
  - ToR 6x100G, 48-port 10G copper

General Datacenter / CE & FabricPath



General Datacenter

- Shared Aggregation for Science and Business applications
- Aggregation switches in four computing rooms (two buildings)/Nexus-7010/7710/9508 aggregation switches
- Primary/Secondary model / Asymmetric BW provisioning
- Access layer (per stakeholder)
  - Chassis-based Nexus 7Ks, Catalyst 6509Es
  - 10/100/400G ToR switches
- Per VLAN Classical Ethernet & FabricPath
  - Evolving from nx10G inter-switch uplinks to nx100G
  - Initially expensive 100G optics made FabricPath in Full-Mesh topology very efficient
  - With 100G optics becoming much cheaper -moving towards to more deterministic paths between switches in datacenter
  - E2E Circuits (NOvA, legacy CDF) from DC directly to the perimeter

General Datacenter (cont.)

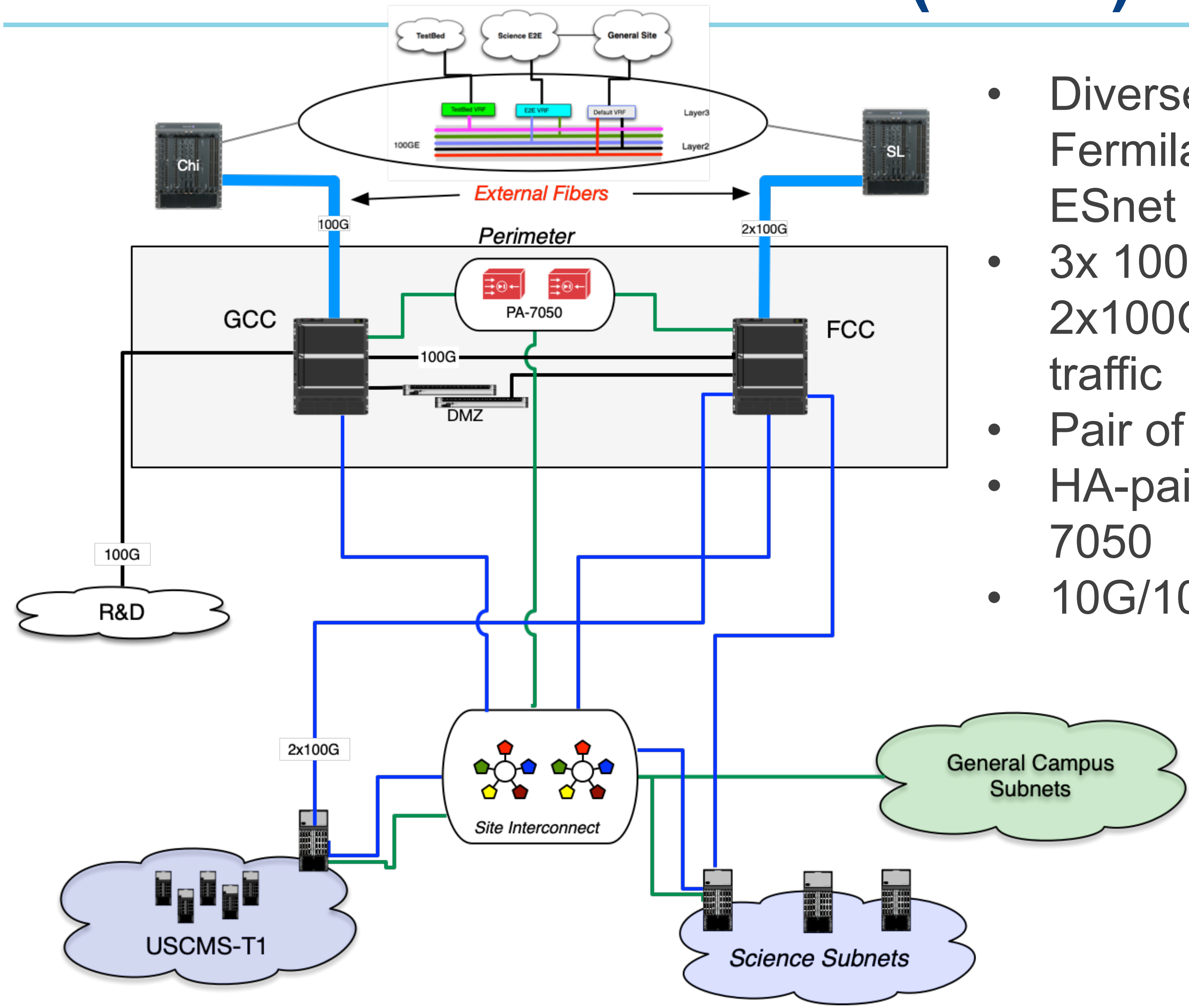
- Selected science traffic bypasses the Site Interconnect and the perimeter firewall
- On-going upgrade to native 400G
  - Two new Nexus 9508 with 100/400G line cards
  - Access switches with 400G uplinks
  - Discontinue FabricPath due to unavailability on the Nexus 95K platform
- In-Service Software upgrades
- The future plans
  - Stay with Classical Ethernet – Not expensive optics
  - VXLAN – No immediate plans to deploy in production. Potentially – for the purpose of extending LAN to/from remote locations such as SURF for DUNE

Network Perimeter Module

- Pair of Brocade MLXe16 routers in different buildings
- HA-pair of PA-7050 firewalls in different buildings
- Redundant access switches with 48x10G, 6x100G for DMZ
- Three 100G offsite circuits via two diverse fiber links to ESnet
- VRF separation of science from all other traffic
- Distribution of traffic between 100G circuits:

Circuit	Primary Function	Secondary Function
1x 100G	General Routed IP traffic	LHCONE Backup, LHCOPN E2E Tertiary
2x 100G	LHCOPN, LHCONE, E2E Circuits	General Routed IP traffic

100G Network Perimeter (cont.)

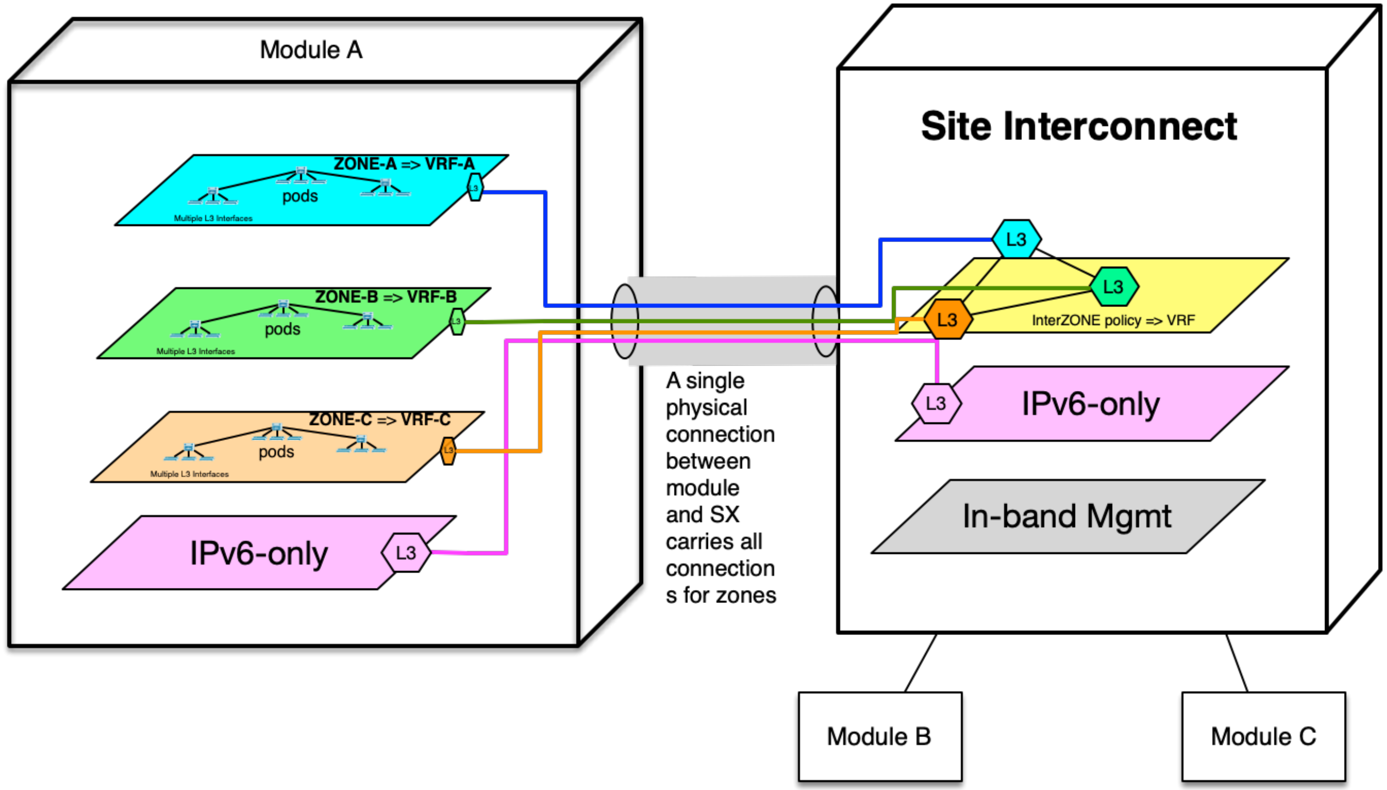


- Diverse fibers from Fermilab to two diverse ESnet PoPs at Chicago
- 3x 100G circuits with 2x100G LAG for science traffic
- Pair of Brocade MLXe16
- HA-pair of PA firewall 7050
- 10G/100G DMZ

Site Interconnect Module

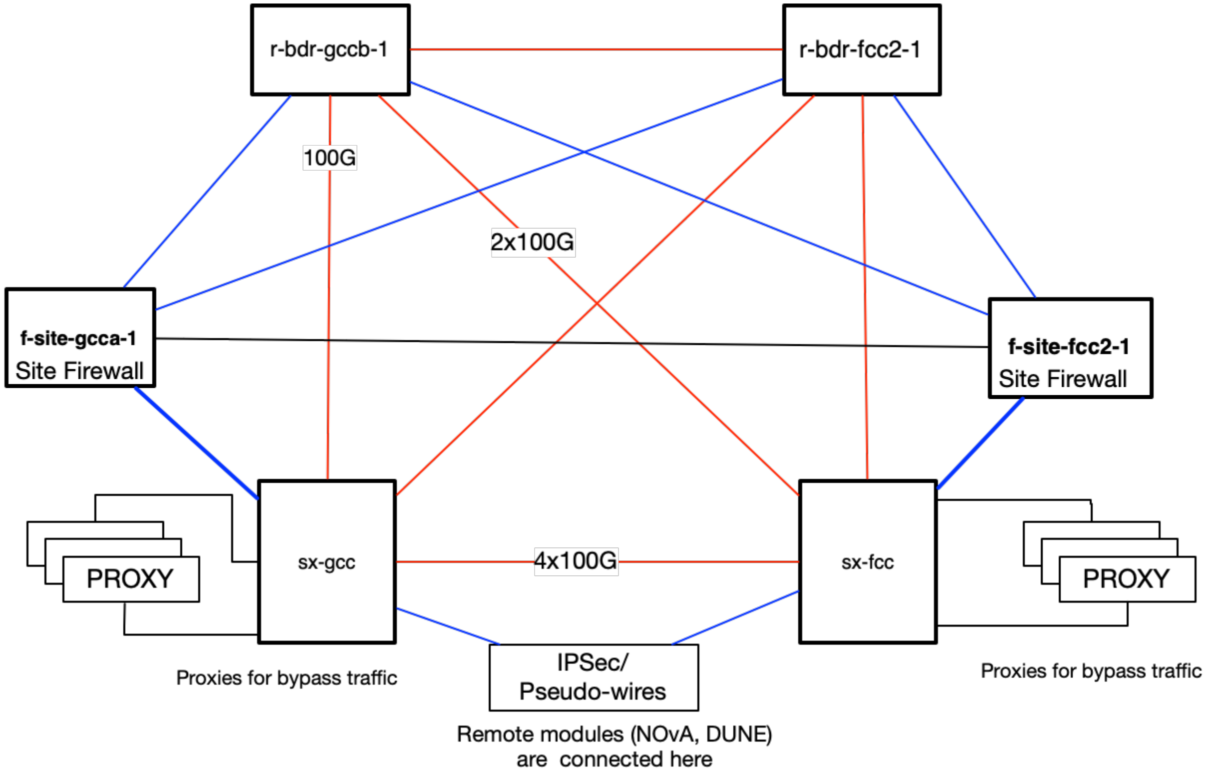
- A module that interconnects all other modules
- Vision: interconnects could be at
  - Layer 3 (Typical)
  - Layer 2 (Separating SPT domains)
  - Layer 1 (Optical switching)
- Common security inter-module policies are applied (in addition to module specific security protections)
- Multiple VRFs for Layer3 connections
- Currently: nx10G & nx100G connections
- Implementation:
  - Two Cisco Nexus 7710s in different buildings
  - Admin VDC
  - 2 x M3 12-port 100G modules, 2x F3 48-port 10G modules

Site Interconnect (Multiple VRFs)



Multiple VRFs to separate traffic of virtual networks, security zones or between different network modules

Site Interconnect topology

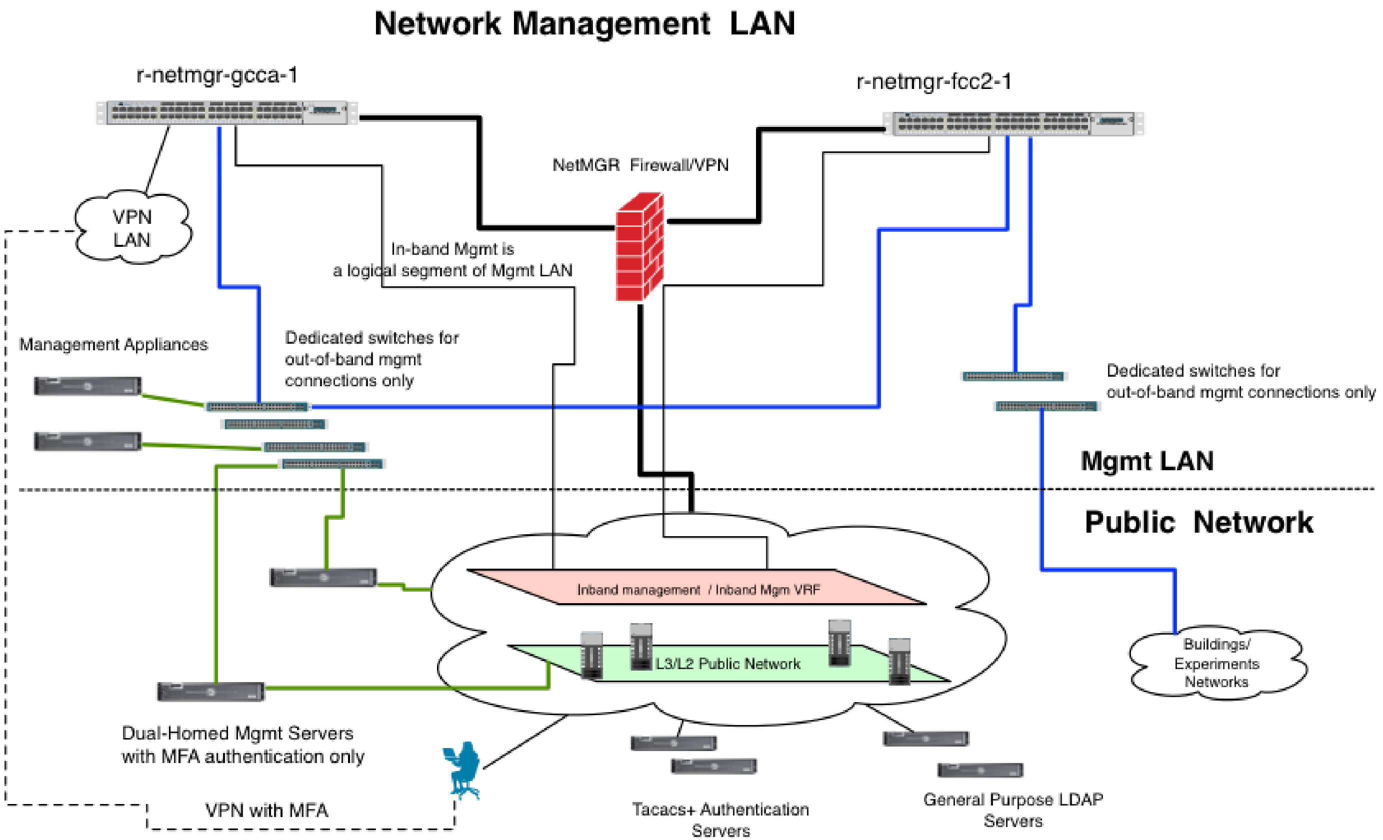


All SX- connections are LAGs:

- nx100G
- nx10G

At the Site Interconnect traffic can be going directly to the Perimeter (bypass traffic) or forwarded to the Site firewall. Selected bypass traffic is hidden by proxies. Remote sites are terminated at the SX via IPSec tunnels or pseudo-wires.

Network Mgmt LAN & MFA



- A separate physical infrastructure for Out-of-Band bgmt
- A separate VRF for In-band mgmt.
- For critical devices both, In-band and Out-of-Band

Summary

- Modular Fermilab Network Architecture
- The science facilities are based on two-tier model with aggregation and access layer
- Standard Spine/Leaf to build the Clos architecture – impractical in our DCs
- Classical Ethernet/SPT, currently FabricPath
- IP Fabric/VXLAN, no plans to deploy within DC but to be used to extend DC VLANs to/from remote locations
- Nx100/400G LAG uplinks between switches
- 10G copper is default server connections
- Growing deployment of 100G servers (~ 100 now)
- Optics:
  - 100G - QSFP28 CWDM4/LR4 – 2km/10km - $200/$600
  - 100G Servers – QSFP28 SR4/MPO within same row of racks, $80
  - 400G - QSP28-DD FR4 - 2km / $3000

Questions ?

&

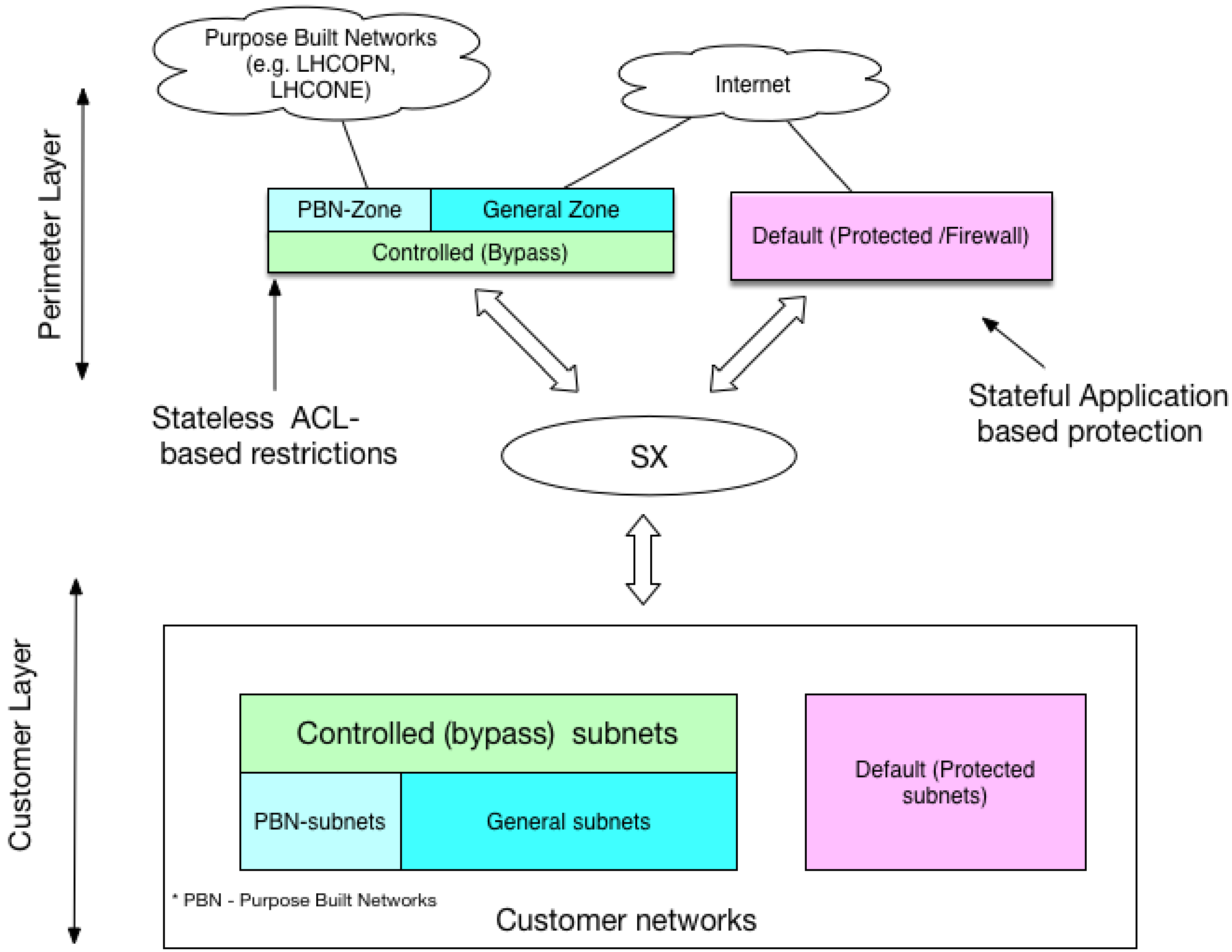
Optional Slides

FNAL E2E circuits (from my.es.net)

NAME ▴	DESCRIPTION	CAPACITY	TYPE	FROM	TO
ZWWY	FNAL - UFL, VLAN 1999, 10G	10.0Gbps	n/a	FNAL	INTERNET2
YN7K	Wenji:20,000 Mbps:vlan=1662	20.0Gbps	n/a	STARLIGHT	FNAL
TJ33	FNAL-TIFR v2499	1.00Mbps	n/a	FNAL	CERN
MYC2	FNAL LHCOPN TERTIARY	10.0Gbps	n/a	FNAL	CERN
MKYZ	FNAL - UFL (via AL2S CHIC), VLAN 1304, 1G	1.00Gbps	n/a	FNAL	INTERNET2
MKDR	FNAL - PURDUE, VLAN 2549, 1G	1.00Gbps	n/a	FNAL	BTAA
J99C	FNAL Secondary LHCOPN	0.00bps	n/a	CERN	FNAL
G6D4	FNAL - ASGC, VLAN 3120, 1G	1.00Mbps	n/a	STARLIGHT	FNAL
F7RR	FNAL- UMN-NOVA-B VLAN 203	500Mbps	n/a	FNAL	BTAA
F39R	FNAL LHCOPN PRIMARY	30.0Gbps	n/a	FNAL	CERN
EYK6	FNAL- UMN-NOVA-A VLAN 201	1.00Gbps	n/a	BTAA	FNAL
DTRD	FNAL - UCSD, VLAN 3021, 1G	1.00Gbps	n/a	FNAL	PWAVE-SUNN

Primary LHCOPN circuit

Two Site-wide Security Zones



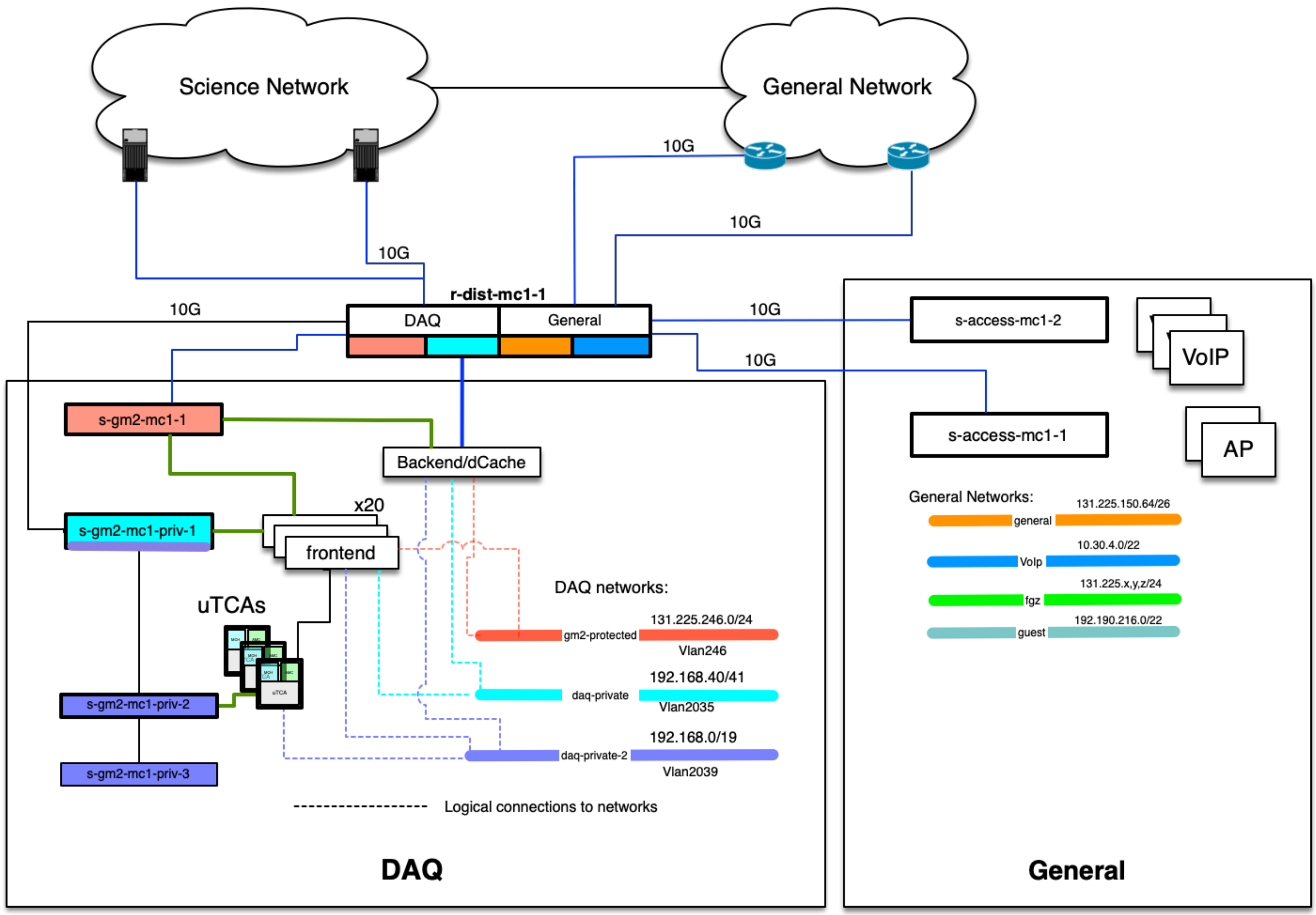
Network Architecture for Experiment DAQs

- Nowadays a typical DAQ system consists of custom built electronics and commodity servers. A number of servers could be just a few, or a few hundred (NOvA ~ 300)
- A typical DAQ now is built using 10G end systems and 100G uplinks to central computing and storage resources
- The Short Base Neutrino & Long Base Neutrino requested 100G uplinks from the detectors to datacenter - Growing demands for bandwidth within LAN

Network Architectures for small DAQs

- Physical separation does not make sense
- Two different VRFs
  - General Networking
  - DAQ itself
- Separate uplinks:
  - SX for general traffic
  - DC for Science data movement

Small DAQ network (G-2 example)



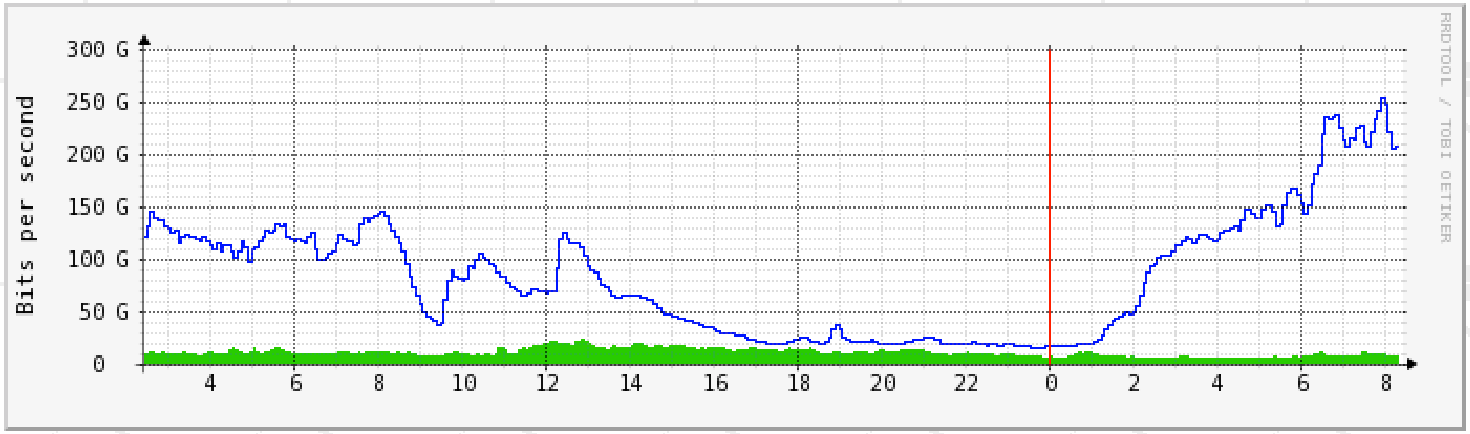
USCMS Utilization 400G Inter-building channel

Traffic Analysis for port-channel132 r-cms-gcca-1

Switch:	r-cms-fcc2-1
Location:	FCC2-1593
Maintainer:	
Interface Type:	propVirtual
Interface Name:	port-channel132
Connected To:	r-cms-gcca-1
Max Speed:	400.0 Gbits/s

The statistics were last updated **Saturday, 26 March, 08:22:06 CDT**

`Daily' Graph (5 Minute Average)



Max In:	23.2 Gb/s (5.8%)	Average In:	10.8 Gb/s (2.7%)	Current In:	7684.7 Mb/s (1.9%)
Max Out:	254.4 Gb/s (63.6%)	Average Out:	84.1 Gb/s (21.0%)	Current Out:	209.0 Gb/s (52.2%)