US LHC Tier-1
WAN Data Movement
Security Architectures

Phil DeMar (FNAL); Scott Bradley (BNL)

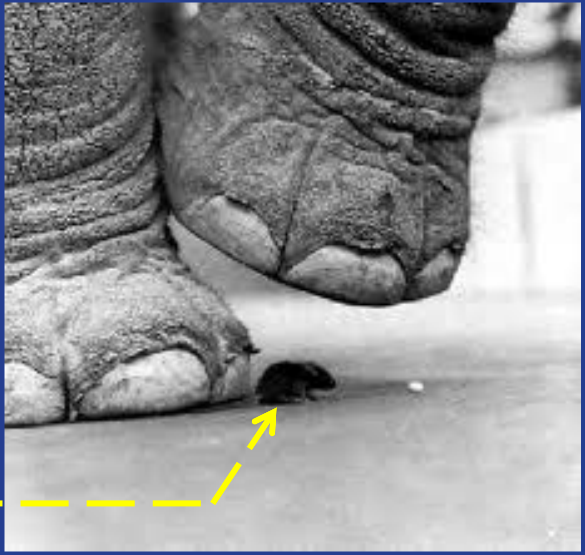
demar@fnal.gov; bradley@bnl.gov

CHEP 2013

October 17, 2013

The Challenges with High Impact Data Movement

- Network Bandwidth:
  - General R&E networks may not provide enough
- Site Perimeter Security Obstacles:
  - Firewall performance isn't keeping up
- Intermingling Bulk Data Traffic with Interactive Applications:
  - Don't want this to be your users audio/video apps

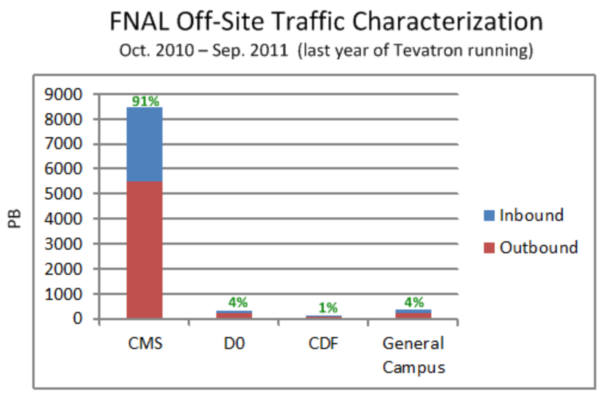


- Optimal performance may require use of alternate network paths
  - Means using non-default routing mechanisms

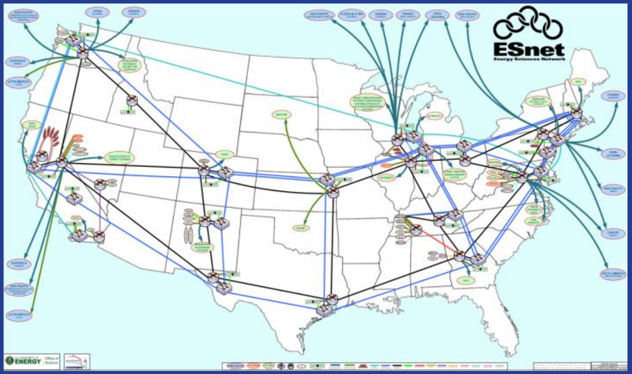
Fermilab

Long Term Trend Toward Isolating Wide-Area Science Data Movement

- At Tier-1 Centers, LHC WAN data dwarfs everything else:
  - (Final year of Tevatron operations)



- Hybrid R&E Networks to service high impact science data:
  - Isolated network paths
  - Potentially with B/W guarantees

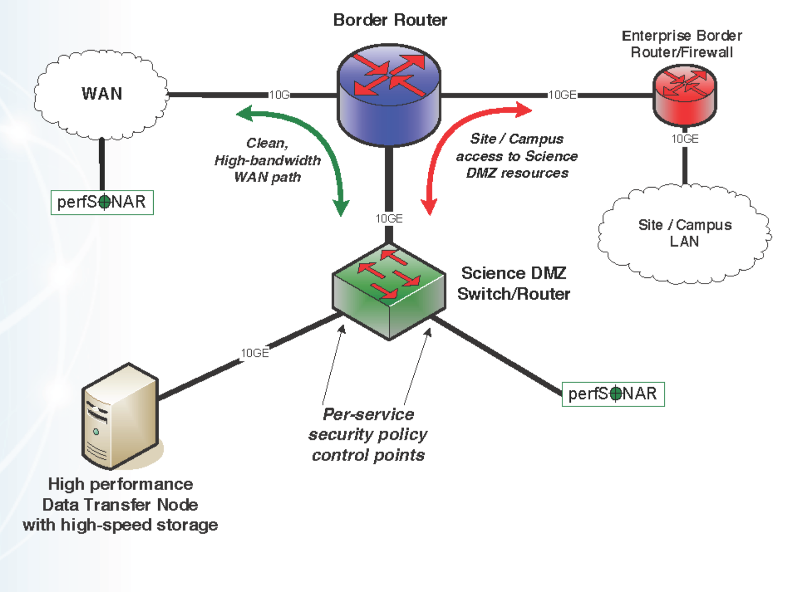


ESnet4 circa 2009

- Data circuit technologies enable virtual pt-to-pt connections
  - More secure “private” network paths
- Discipline-specific networks appear

ESnet-Developed Model for a Prototype Site Science DMZ

- General idea: separate science data from general campus network infrastructure
- Components:
  - Data Transfer Node (DTN):
    - Optimized for WAN transfers
  - Bypass routing around site firewalls
  - Network measurement infrastructure
    - PerfSONAR


Border Router
Enterprise Border Router/Firewall
WAN
10G
10GE
10GE
Clean, High-bandwidth WAN path
Site / Campus access to Science DMZ resources
perfSONAR
10GE
Site / Campus LAN
Science DMZ Switch/Router
10GE
perfSONAR
Per-service security policy control points
High performance Data Transfer Node with high-speed storage

ESnet prototype Science DMZ figure

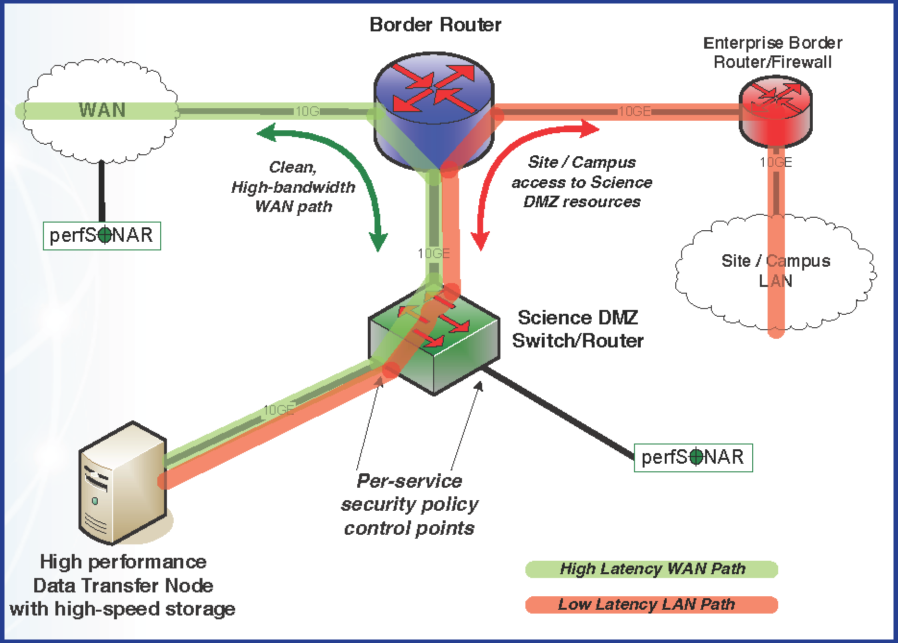
- An architecture, not implementation

Prototype Science DMZ Data Paths

- Custom WAN path(s) for science DMZ systems:
  - Appropriate bandwidth provisioning is primary goal
  - Optimizing latency isn't a goal

- Specific security policies & tools for science data traffic

- Science DMZ resides outside of site perimeter



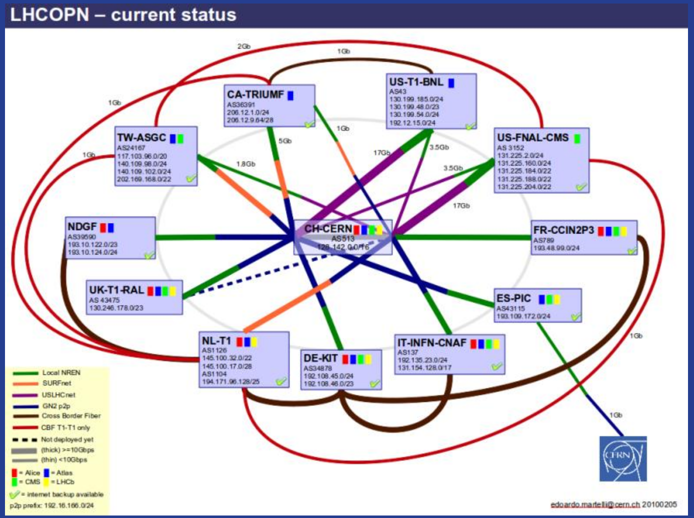
Border Router
Enterprise Border Router/Firewall
WAN
10G
10GE
perfSONAR
Clean, High-bandwidth WAN path
Site / Campus access to Science DMZ resources
10GE
10GE
Site / Campus LAN
Science DMZ Switch/Router
10GE
perfSONAR
Per-service security policy control points
High performance Data Transfer Node with high-speed storage
High Latency WAN Path
Low Latency LAN Path

ESnet figure on Science DMZ data paths

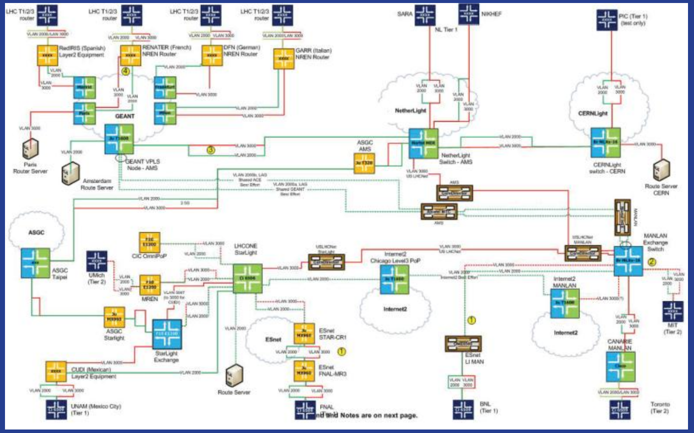
Diverse Set of LHC WAN Paths

Controlled-access traffic reduces security risks

- LHC Optical Private Network (**LHCOPN**) for T0 <-> T1 data
  - Tightly-controlled access
  - Dedicated bandwidth
- LHC Open Network Exchange (**LHCONE**) for T1/T2 <-> T1/T2
  - Loosely-controlled access
  - Mix of routed & end-to-end traffic
- End-to-end circuits:
  - Sometimes over private links
  - Sometimes over R&E networks
- Routed IP path over general R&E network infrastructure
  - When no other paths are available



LHCOPN



LHCONE

Policy-Based Routing (PBR)

- PBR = forwarding of specific packets based on administrative decision or policy:
  - In contrast to following normal (dynamic…) routing tables
  - Manually configured on routers
- Cisco implementation is Route-Map command:
  - Up to 5-tuple mapping (src/dest IP, src/dest port, protocol)
  - Basic components are “mapping” & “action taken if matched”
- Similar capabilities available from other vendors
  - Generic name = ACL-based forwarding

IPSLA & Object Tracking

- PBR is a form of static routing:
  - Potential for black-holing traffic is path is down
- IPSLA and Object Tracking can be used to avoid black-holing
  - IPSLA continuously checks to make sure path is up:
    - ICMP ping to remote end of path
    - Also used for SNMP monitoring of status for end-to-end circuits
  - Object Tracking maintains status of path as an object
    - If IPSLA fails, object (path) is marked as down
  - Route-maps (PBR) configured with tracking object identifier
    - If object is down, PBR forwarding is not implemented

FNAL Implementation

Site Perimeter Basics at FNAL

- FNAL does not have a site firewall
  - OK, we have one but science data systems aren't behind it
  - Firewalls aren't a factor for our science data movement
- Site security based on wide spectrum of controls
  - Strong authentication mandated
  - Onus on sysadmins to secure their systems
  - Intense vulnerability scanning
  - Perimeter controls (ACLs), IPS, web proxy, etc
- By default, science data must pass thru perimeter controls
  - Bypass exception:

    "Known traffic from well-managed systems at trusted sites"

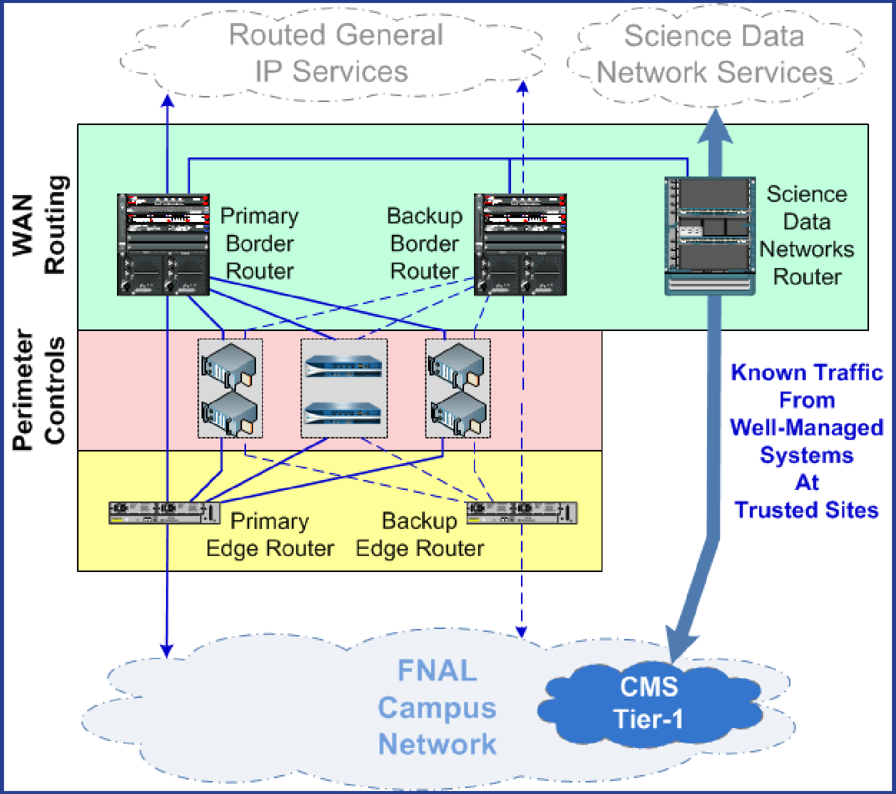
    - Exception based on risk analysis and acceptable residual risk

FNAL's PBR Implementation

- Forwards science traffic to alternate network paths
- 2-tuple, based on src/dest netblocks
  - CMS Tier1 is always one netblock
  - Remote T0/T1/T2 netblock is always the other
- All FNAL PBR is **internal**:
  - No PBR forwarding into WAN
- Perimeter security control mechanism for bypass traffic:
  - Only PBR-validated traffic can use the bypass route

FNAL Network Perimeter versus the Science DMZ Model

- CMS Tier-1 integrated into campus network:
  - No special DTN nodes
  - Tier-1 dCache servers are equivalent to DTNs
- Separate border router for bypass traffic:
  - Consistent with bypass traffic security policy

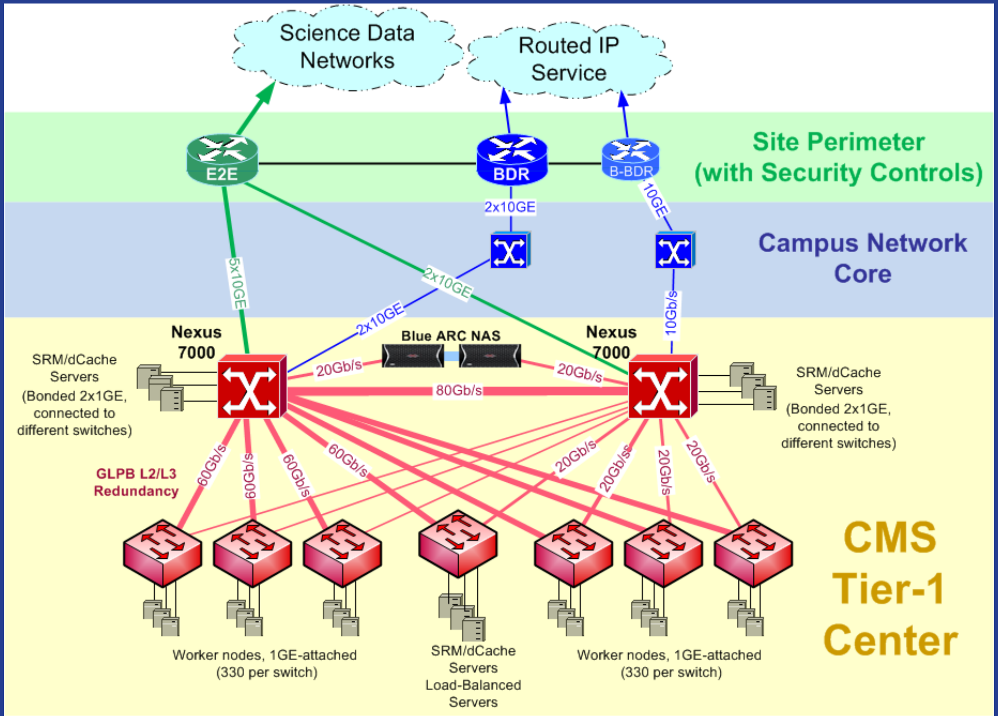


Routed General IP Services
Science Data Network Services
WAN Routing
Primary Border Router
Backup Border Router
Science Data Networks Router
Perimeter Controls
Primary Edge Router
Backup Edge Router
Known Traffic From Well-Managed Systems At Trusted Sites
FNAL Campus Network
CMS Tier-1

- Non-bypass traffic to/from Tier-1 passes through normal perimeter security controls

US-CMS Tier-1 – Alternate WAN Paths

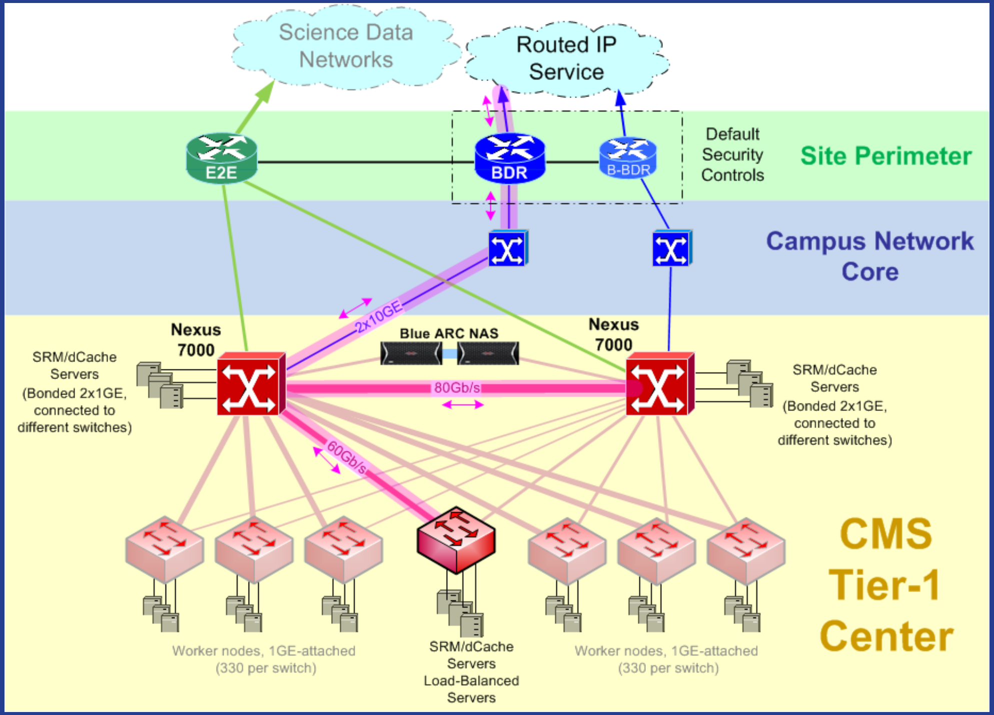
- US-CMS Tier-1:
  - ~1600 systems
  - Distributed across 4 computer rooms
  - dCache servers distributed as well


Science Data Networks
Routed IP Service
E2E
BDR
B-BDR
Site Perimeter (with Security Controls)
Campus Network Core
2x10GE
10GE
5x10GE
2x10GE
2x10GE
10Gb/s
Nexus 7000
Nexus 7000
Blue ARC NAS
20Gb/s
80Gb/s
20Gb/s
SRM/dCache Servers (Bonded 2x1GE, connected to different switches)
SRM/dCache Servers (Bonded 2x1GE, connected to different switches)
GLPB L2/L3 Redundancy
60Gb/s
60Gb/s
60Gb/s
60Gb/s
20Gb/s
20Gb/s
20Gb/s
20Gb/s
Worker nodes, 1GE-attached (330 per switch)
SRM/dCache Servers Load-Balanced Servers
Worker nodes, 1GE-attached (330 per switch)
CMS Tier-1 Center

- Primary & secondary Tier-1 LAN switches
  - Connections to campus core
  - Also to bypass perimeter router (E2E) for WAN science data
    - Higher bandwidth connection for science data movement

By Default, Tier-1 Traffic Follows General IP (Routed) Path

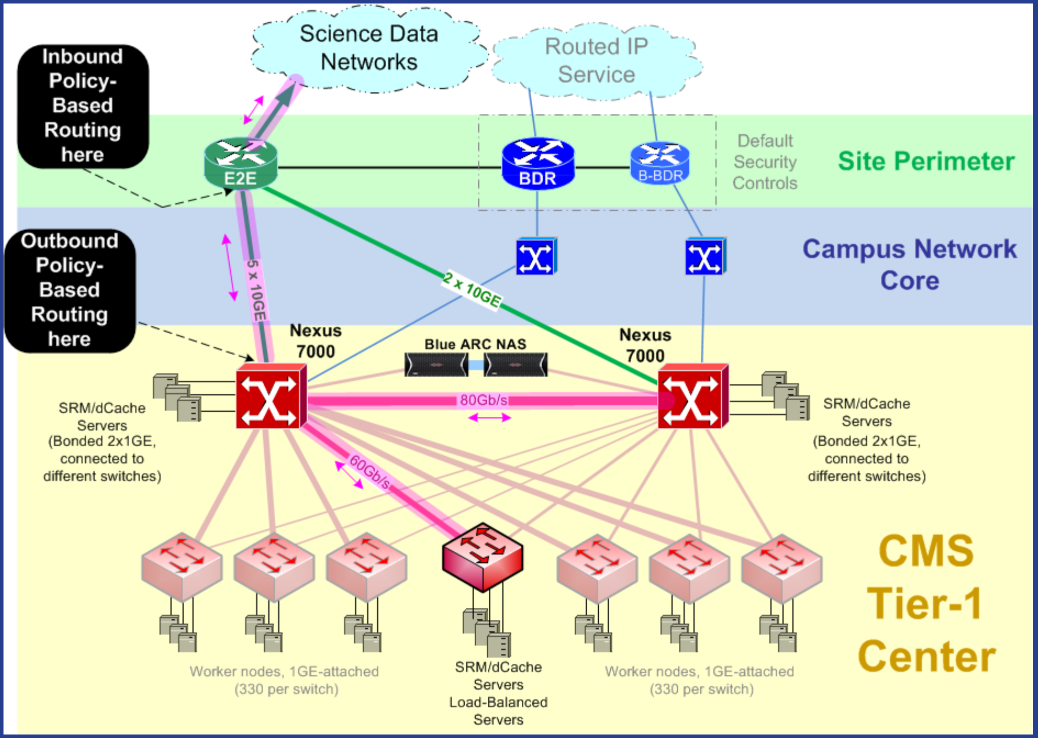
- Tier-1 WAN traffic forwarded through primary Tier-1 switch
  - Path symmetry more important than traffic load balancing
  - Layer -2 traffic within LAN distributed across links via VPC


Science Data Networks
Routed IP Service
E2E
BDR
B-BDR
Default Security Controls
Site Perimeter
Campus Network Core
2x10GE
Nexus 7000
Blue ARC NAS
Nexus 7000
SRM/dCache Servers (Bonded 2x1GE, connected to different switches)
80Gb/s
SRM/dCache Servers (Bonded 2x1GE, connected to different switches)
60Gb/s
CMS Tier-1 Center
Worker nodes, 1GE-attached (330 per switch)
SRM/dCache Servers Load-Balanced Servers
Worker nodes, 1GE-attached (330 per switch)

- Unless bypass routed, traffic will pass through campus core & border router
  - This includes perimeter security controls

Alternate Network Paths via PBR

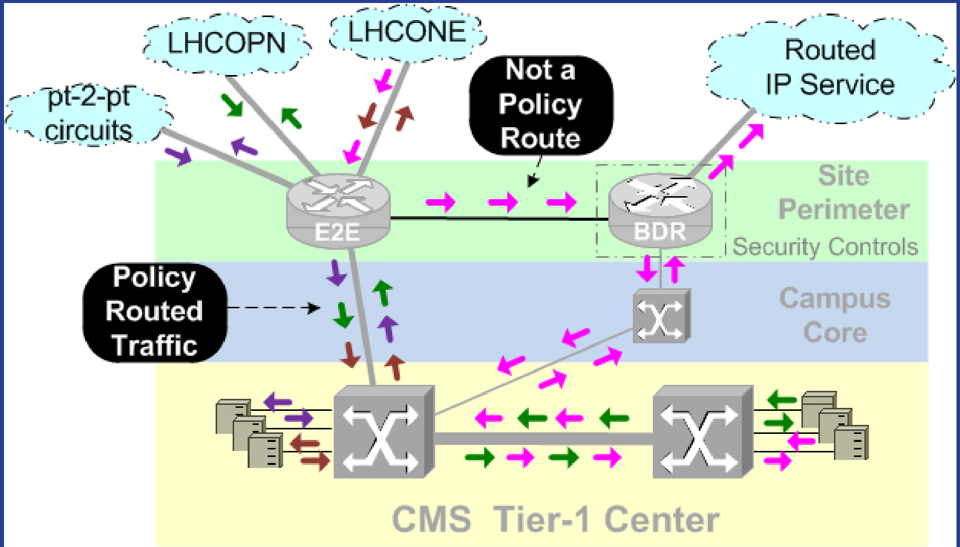
- PBR steers select traffic to/from bypass router
  - Based on src/dst address blocks
  - Our Tier-1 netblock is always one of the tuples
    - Remote Tier-0, Tier-1, or Tier-2 is the other


Science Data Networks
Routed IP Service
Inbound Policy-Based Routing here
E2E
BDR
B-BDR
Default Security Controls
Site Perimeter
Outbound Policy-Based Routing here
5 x 10GE
2 x 10GE
Campus Network Core
Nexus 7000
Blue ARC NAS
Nexus 7000
80Gb/s
60Gb/s
SRM/dCache Servers (Bonded 2x1GE, connected to different switches)
SRM/dCache Servers (Bonded 2x1GE, connected to different switches)
CMS Tier-1 Center
Worker nodes, 1GE-attached (330 per switch)
SRM/dCache Servers Load-Balanced Servers
Worker nodes, 1GE-attached (330 per switch)

- PBR is manually configured:
  - A bit of a pain, but scalable to level of CMS collaboration
  - Dealing with address changes at remote sites also an issue

Non-PBR Traffic on Bypass Connections

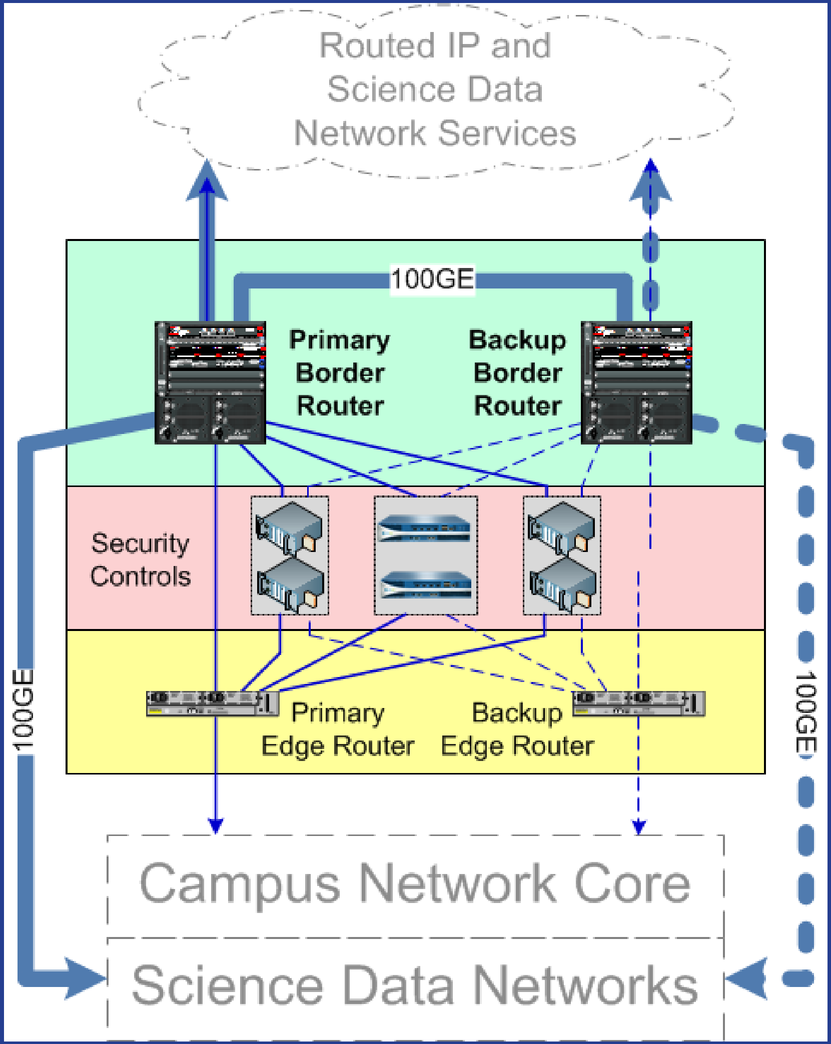
- Incoming traffic not in PBR tables forwarded to border router:
  - Still gets into the Tier1
  - But passes through security controls
  - Also creates WAN path asymmetry
    - May cause firewall problem on remote end



- We monitor flow data between bypass and border router for this type of traffic
  - Will add to PBR tables if its valid CMS traffic
  - But this is still a manual process

FNAL Future Directions – Perimeter Architecture

- Costs of 100GE will necessitate consolidating bypass router functions into border & backup border routers:
    - Consistent with general technology trend to consolidate network hardware & virtualize network functions



BNL Implementation

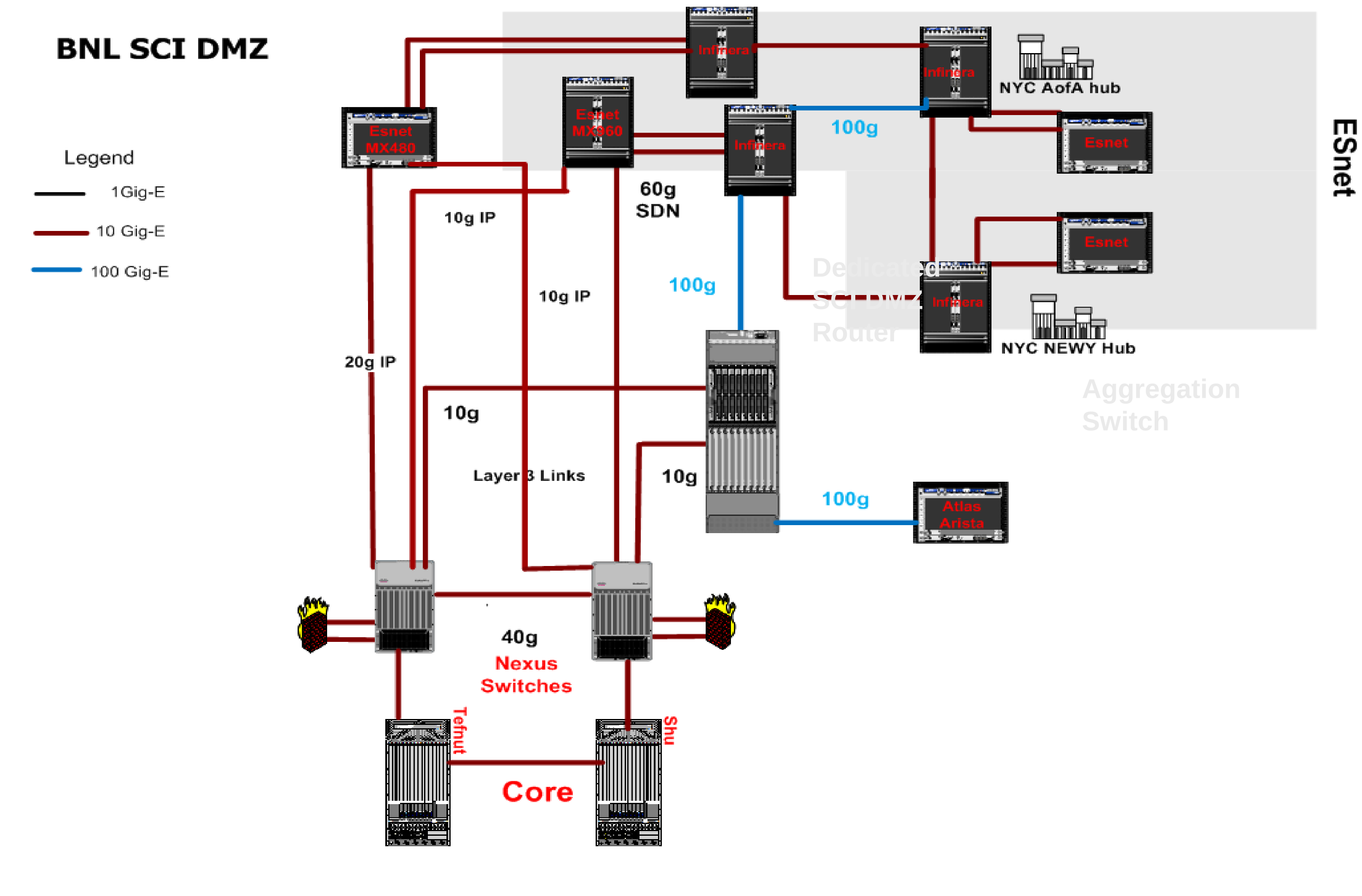
BNL Science DMZ

- 100gb/sec WAN Connectivity:
  - Provides native 100gb/sec. interfaces
  - Will interface to Testing and Production 100g waves
  - Supports multiple 10gb/sec. and 40gb/sec. connections
  - Initially 2 attachment ports at 100gb/sec.
  - Dedicated CIDR block for IP addressing
  - Will have limited Etherchannel to BNL campus
  - Dedicated routing platform – Juniper MX2010

BNL Science DMZ cont

- Current Status
  - First 100g wave in testing phase
  - Will participate in TA100 testing with CERN
  - Currently evaluating an Arista 7508E switch for aggregation, others to follow
    - High port density and types are key requirements

BNL Science DMZ Topology



General Future Directions - OpenFlow

- PBR has worked very well to isolate & control our science data traffic, but:
  - Manual configuration is a pain
  - Adds complexity to site routing & troubleshooting
  - Keeping up with address changes/additions is difficult
- OpenFlow - emerging standard for flow-based forwarding:
  - PBR is essentially flow-based forwarding, too
  - We’re investigating OpenFlow to replace current PBR
    - Long term vision - end-to-end forwarding based on OpenFlow
    - Short term goal - replace PBR within the US-CMS Tier-1

Summary

- Separating science data movement from general network traffic has worked well at US-LHC Tier-1s
  - Enabled us to meet needs of both LHC stakeholders & general users, but not at each other's expense
  - Science DMZ architectures based around PBR for LHC traffic:
    - Avoids performance issues with overloading perimeter security tools
- Our implementations work well for us because:
  - We are dealing with established traffic characteristics
  - Our stakeholders are well-organized & long-lived
  - May not translate well to other disciplines
- Looking toward OpenFlow as a more standard approach to separate out our science data movement

Fermilab

?

Questions