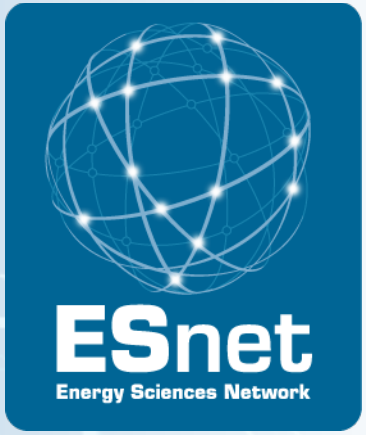


ESnet, the Science DMZ, and the role of Globus Online

Brian Tierney and Eli Dart

ESnet

Globus World

Argonne, IL

April 12, 2012

Overview

- What is ESnet?
- The Science Data Deluge
- What is a Science DMZ?
- Globus Online and the Science DMZ

ESnet **Supports DOE Office of Science**

SC provides broad support for Labs, Facilities, people

- almost $5B/year in funding
- 45% of Federal support for physical sciences research
- key funding for basic research in biology, computing, energy, climate
- over 100 Nobel Prizes in past 60 years

Supporting > 27,000 PhDs, grad students, engineers at >300 institutions.

Provides world's largest collection (32) of scientific user facilities:

- supercomputer centers, accelerators, light sources, neutron sources, electron microscopes, nano-scale centers, a sequencing center, fusion facilities.

ESnet is one of them – connecting DOE sites, facilities, scientists and collaborators.

- optimized for science data transport
- every service exists to support scientific discovery

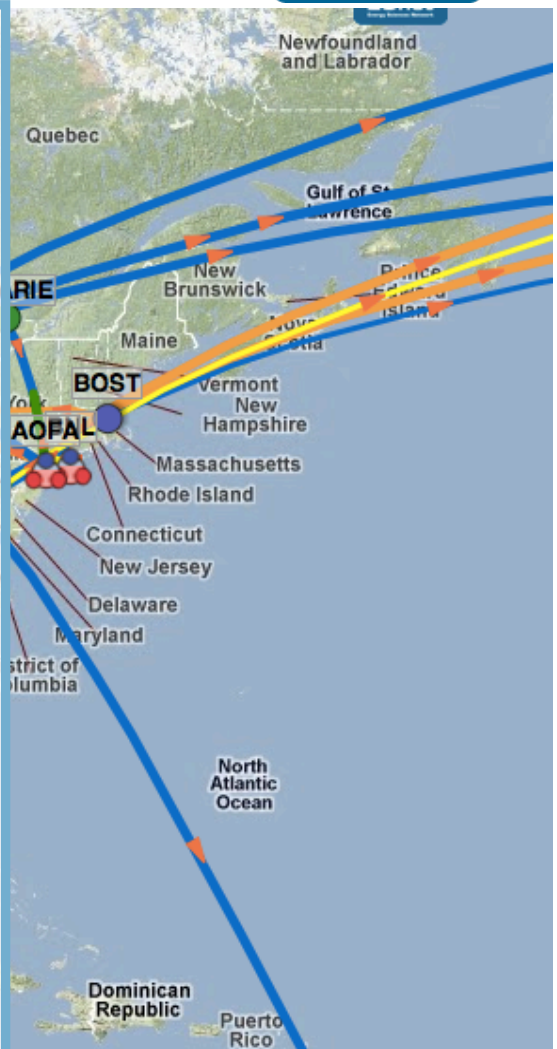
ESnet **Supports DOE Office of Science**

The Office of Science supports:

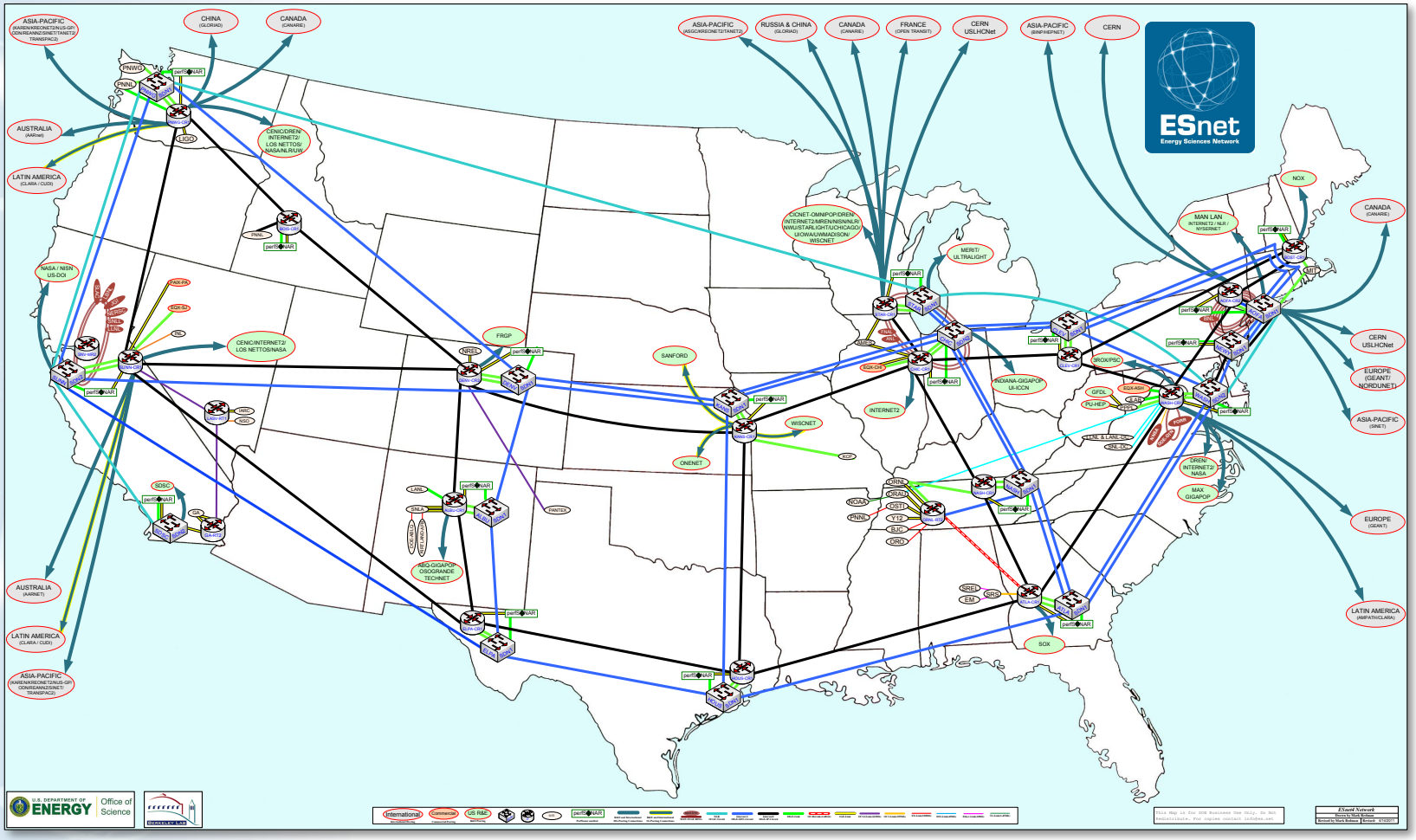
- 27,000 Ph.D.s, graduate students, undergraduates, engineers, and technicians
- 26,000 users of open-access facilities
- 300 leading academic institutions
- 17 DOE laboratories

What is ESnet?

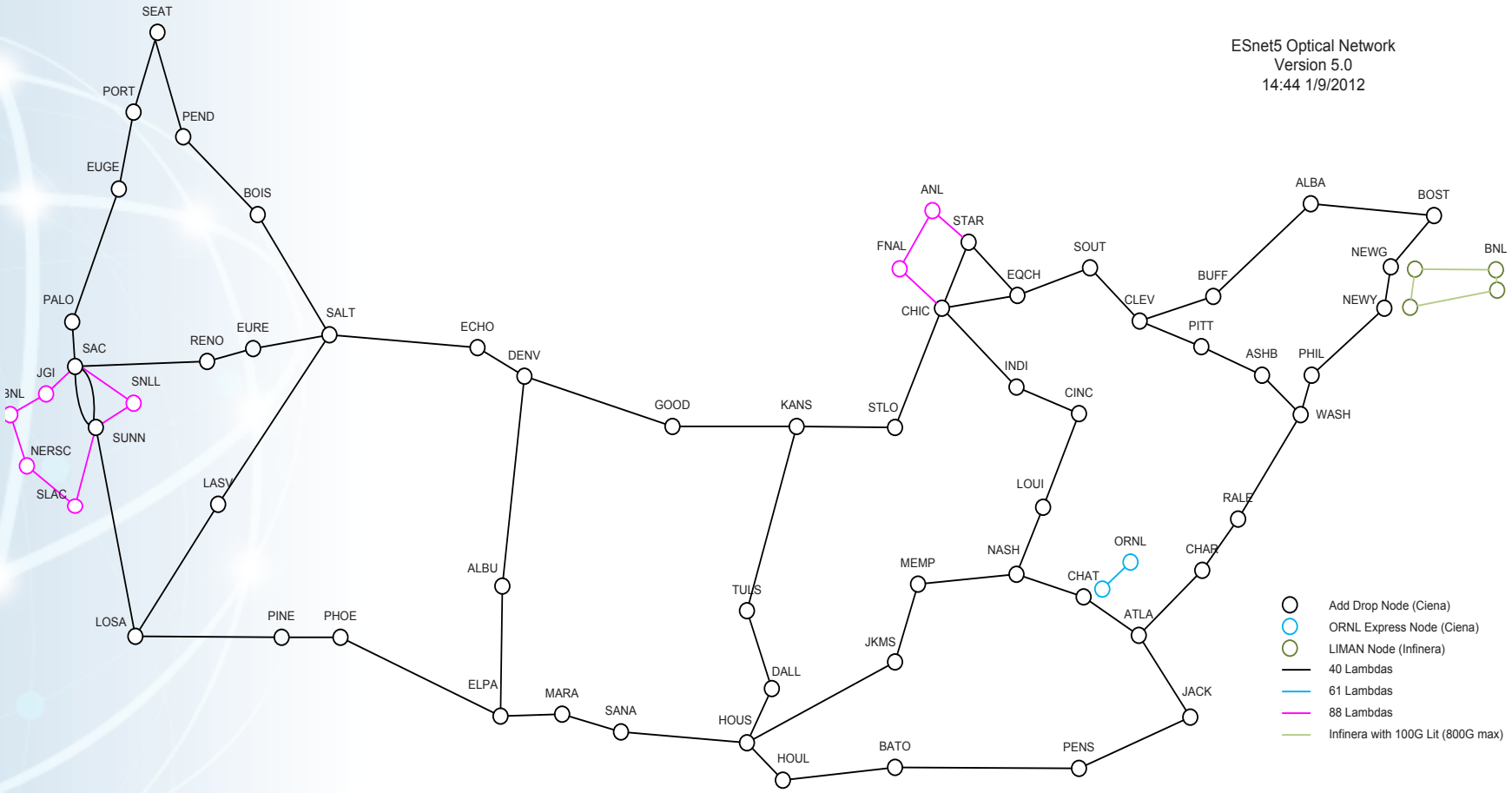
- A national network dedicated to and tailored specifically for science research and applications
  - 20-50Gbps backbone today
  - 100Gbps backbone very soon, easily upgradeable to Nx100Gbps
- A high-performance infrastructure linking DOE Office of Science researchers to global collaborators and resources at sites, including:
  - Supercomputer centers
  - User Facilities
  - Multi-program labs
  - Universities
  - Connectivity to the Internet and Cloud providers
- A national DOE user facility providing:
  - Tailored solutions for science data transfers
  - *Dedicated outreach team to support users*
  - Collaboration services including audio/video conferencing and federated trust



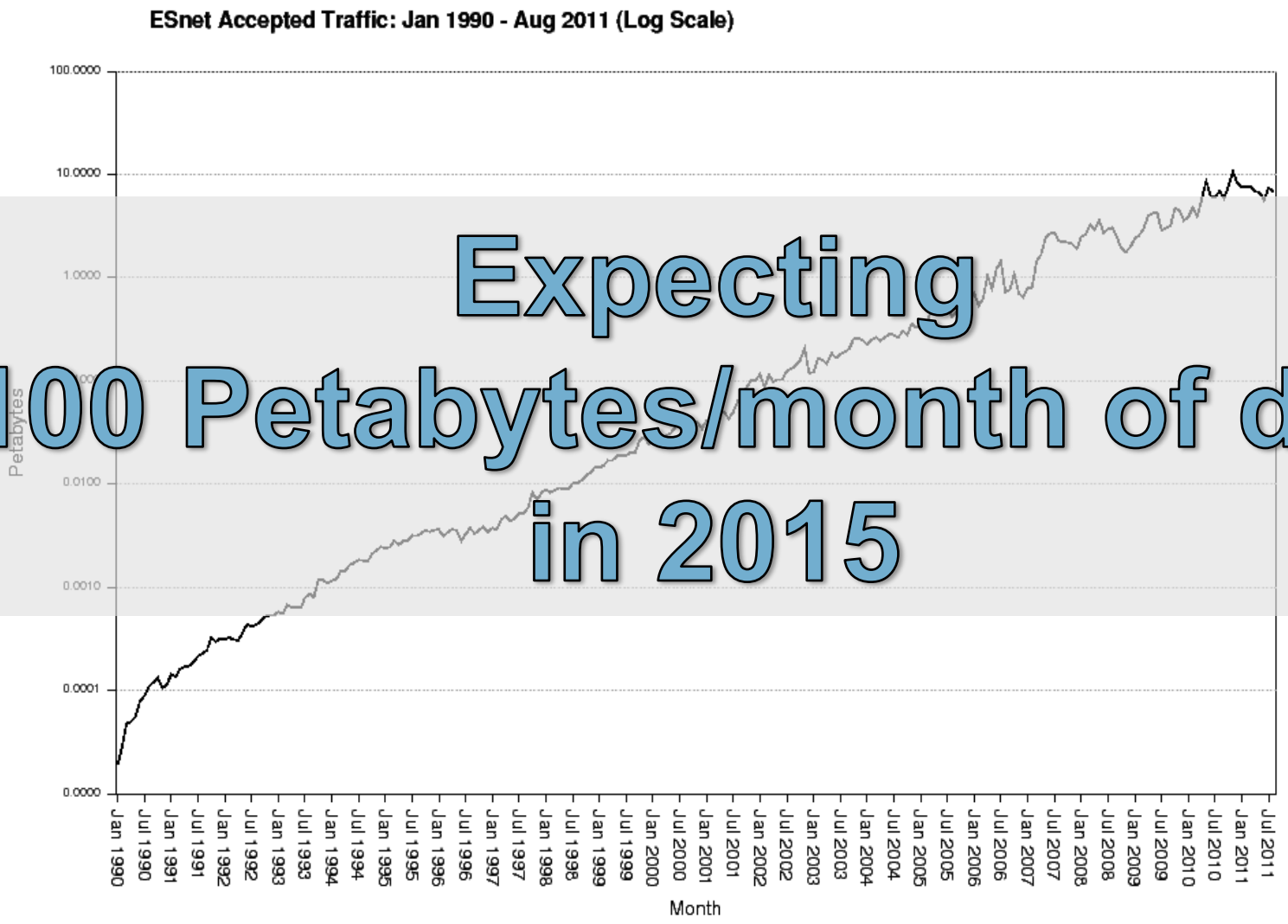
ESnet4 Topology (n x 10G core)



ESnet5 Optical Infrastructure Footprint



The Science Data Explosion

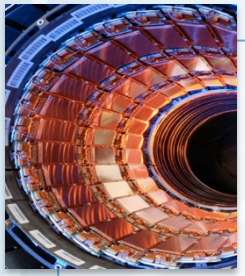


Expecting 100 Petabytes/month of data in 2015

Data Explosion is Occurring Everywhere

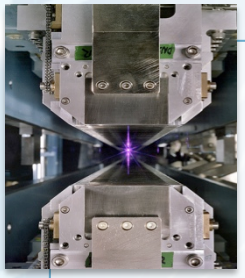
Genomics

- Sequencer data volume increasing 12x over the next 3 years
- Sequencer cost decreasing by 10x over same time period



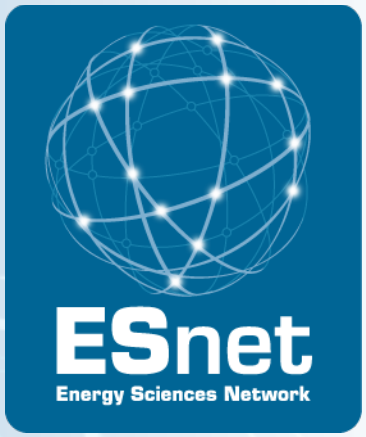
High Energy Physics

- LHC experiments produce & distribute petabytes of data/year
- Peak data rates increase 3-5x over 5 years



Light Sources

- Many detectors on a Moore's Law curve
- Data volumes rendering previous operational models obsolete



The Science DMZ

Science DMZ Origins

ESnet has a lot of experience with different scientific communities at multiple data scales

Significant commonality in the issues encountered, and solution set

- The causes of poor data transfer performance fit into a few categories with similar solutions
  - Un-tuned/under-powered hosts, packet loss issues, security devices
- A successful model has emerged – the Science DMZ
  - This model successfully in use by CMS/Atlas, ESG, NERSC, ORNL, ALS, GA, and others
- New NSF Campus Cyberinfrastructure Solicitation provides funding for Campuses to deploy a Science DMZ
    - http://www.nsf.gov/pubs/2012/nsf12541/nsf12541.htm

One motivation for Science DMZ model: Soft Network Failures

Soft failures are where basic connectivity functions, but high performance is not possible.

TCP was intentionally designed to hide all transmission errors from the user:

- "As long as the TCPs continue to function properly and the internet system does not become completely partitioned, no transmission errors will affect the users." (From IEN 129, RFC 716)

Some soft failures only affect high bandwidth long RTT flows.

Hard failures are easy to detect & fix

- soft failures can lie hidden for years!

One network problem can often mask others

Common Soft Failures

Random Packet Loss

- Bad/dirty fibers or connectors
- Low light levels due to amps/interfaces failing
- Duplex mismatch

Small Router/Switch Buffers

- Switches not able to handle the long packet trains prevalent in long RTT sessions and local cross traffic at the same time

Un-intentional Rate Limiting

- Processor-based switching on routers due to faults, acl's, or mis-configuration
- Security Devices
  - E.g.: 10X improvement by turning off Cisco Reflexive ACL

A small amount of packet loss makes a huge difference in TCP performance

A Nagios alert based on ESnet's regular throughput testing between one site and ESnet core alerted us to poor performance on high latency paths

No errors or drops reported by routers on either side of problem link

- only active testing using perfSONAR bwctl tests caught this problem

Using packet filter counters, we saw 0.0046% loss in one direction

- 1 packet in 22000 packets

Performance impact of this: (outbound/inbound)

- To/from test host 1 ms RTT : 7.3 Gbps out / 9.8 Gbps in
- To/from test host 11 ms RTT: 1 Gbps out / 9.5 Gbps in
- To/from test host 51ms RTT: 122 Mbps out / 7 Gbps in
- To/from test host 88 ms RTT: 60 Mbps out / 5 Gbps in
  - More than 80 times slower!

How To Accommodate TCP?

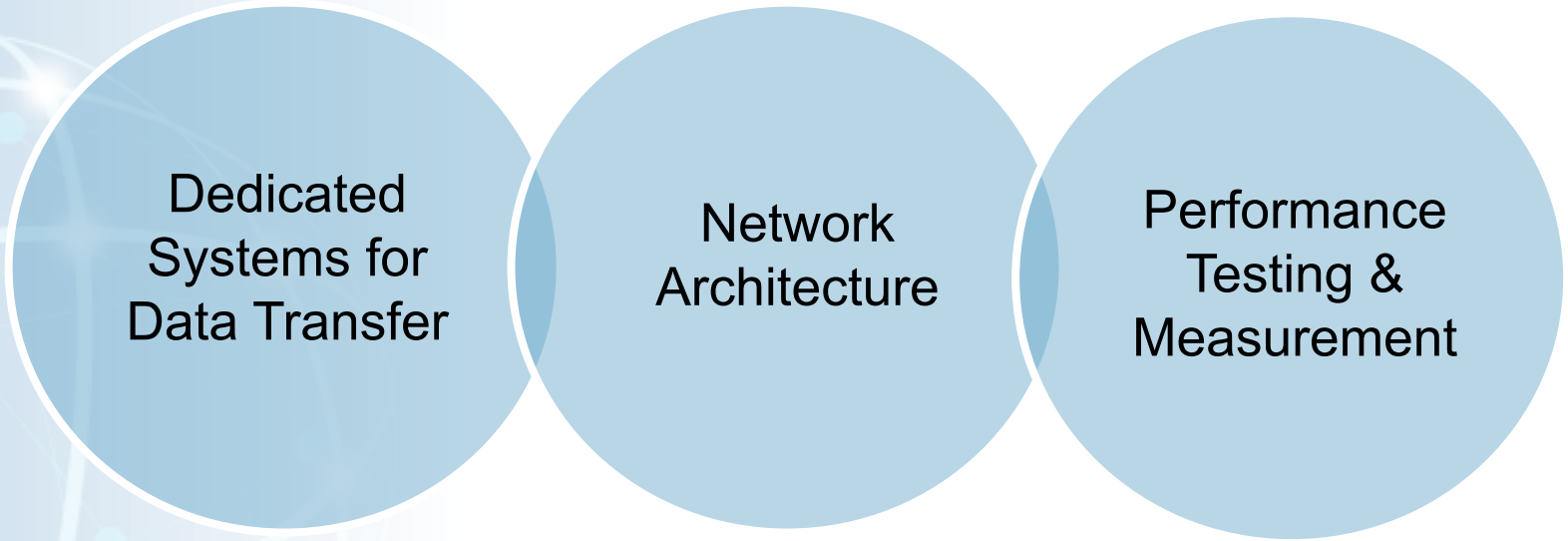
High-performance wide area TCP flows must get loss-free service

- Sufficient bandwidth to avoid congestion
- Deep enough buffers in routers and switches to handle bursts
  - Especially true for long-distance flows due to packet behavior
  - No, this isn't buffer bloat

Equally important – the infrastructure must be verifiable so that clean service can be provided

- Stuff breaks
  - Hardware, software, optics, bugs, …
  - How do we deal with it in a production environment?
- Must be able to prove a network device or path is functioning correctly
  - Accurate counters must exist and be monitored
  - Need ability to run tests - perfSONAR
- Small footprint is a huge win – small number of devices so that problem isolation is tractable

The Data Transfer Trifecta: The "Science DMZ" Model



Data Transfer Node

- High performance
- Configured for data transfer
- Proper tools

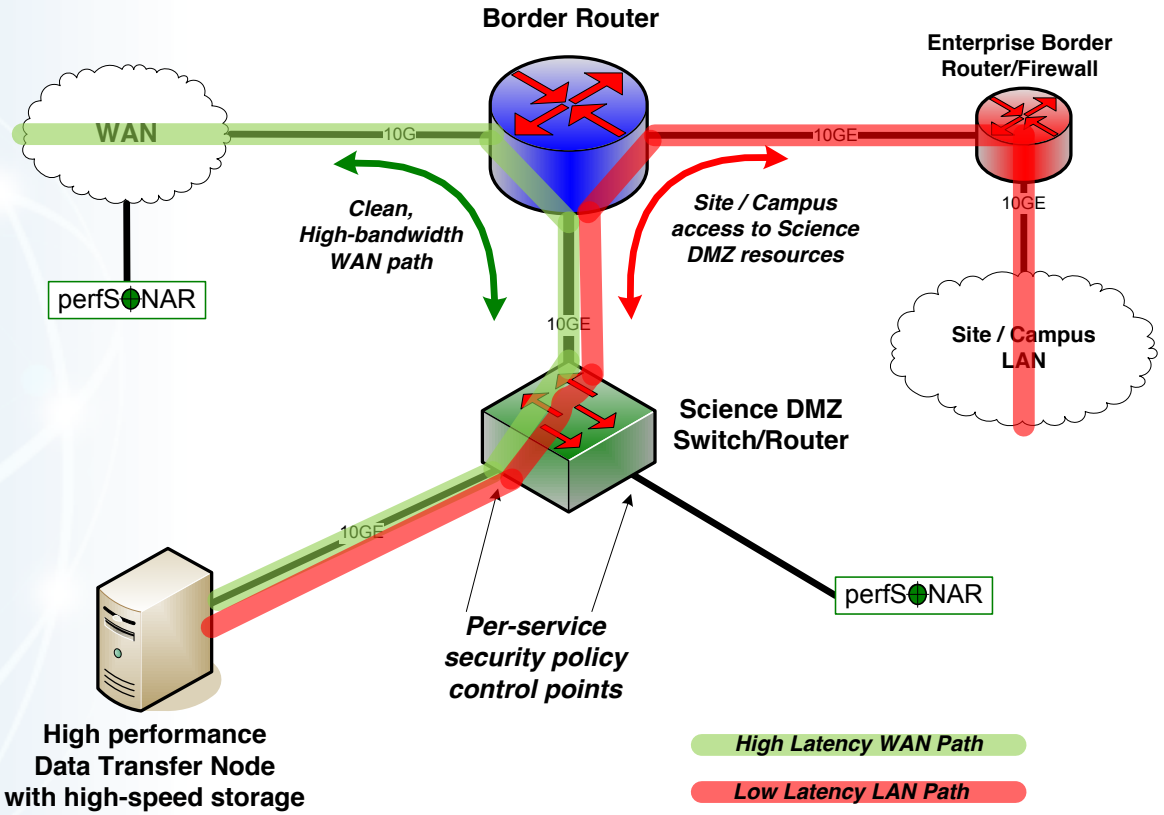
Science DMZ

- Dedicated location for DTN
- Easy to deploy - no need to redesign the whole network

perfSONAR

- Enables fault isolation
- Verify correct operation
- Widely deployed in ESnet and other networks, as well as sites and facilities

Simple Science DMZ



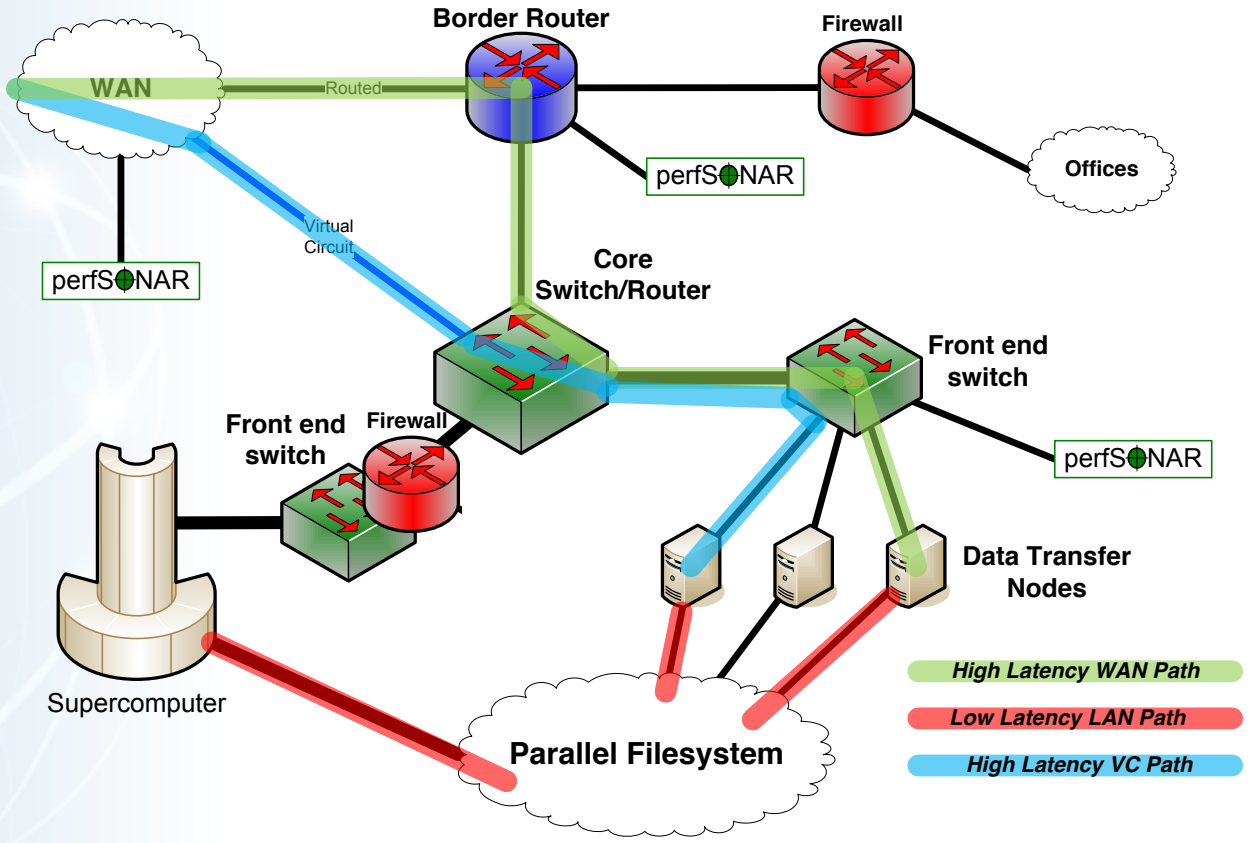
Science DMZ Takes Many Forms

There are many ways to combine the Science DMZ elements – it all depends on what you need to do

- Small installation for a project or two
- Facility inside a larger institution
- Institutional capability serving multiple departments/divisions
- Science capability that consumes a majority of the infrastructure

Some of these are straightforward, others are less obvious

Supercomputer Center Example



Globus Online / GridFTP and the Science DMZ

ESnet recommends Globus Online / GridFTP for data transfers to/from the Science DMZ

Key features needed by a Science DMZ

- High Performance: parallel streams, small file optimization
- Reliability: auto-restart, user-level checksum
- Multiple security models: ssh key, X509, Open ID, Shibboleth, etc.
- Firewall/NAT traversal support
- Easy to install and configure

Globus Online has all these features

Globus Online Wish List Item

New service to help with performance troubleshooting

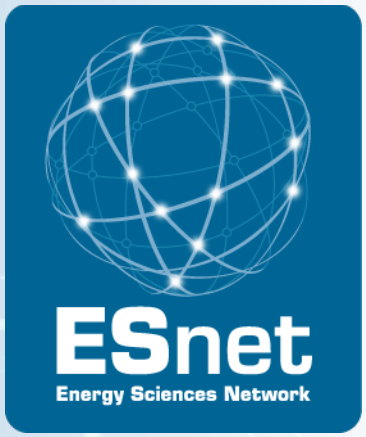
- Ability to query the transfer logs for history of performance
  - e.g.: give me the disk-to-disk throughput from server A to server B for the last 30 days
    - Include file size, # streams, block size, etc.
- This would allow us to compare network testing results to disk-to-disk results
  - Ability to know for certain if the bottleneck is the disk or the network
  - Even Better: include XIO instrumentation summary of network vs disk throughput

Questions?

Email: BLTierney@es.net

Learn more at: http://fasterdata.es.net/fasterdata/science-dmz/

- Includes a full day tutorial slides and video



Extra Slides

Common Threads

Two common threads exist in all these examples

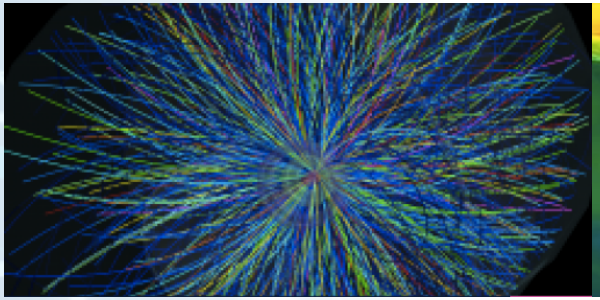
Accommodation of TCP

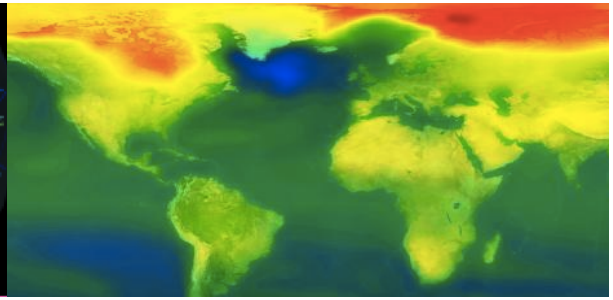
- Wide area portion of data transfers traverses purpose-built path
- High performance devices that don't drop packets

Ability to test and verify

- When problems arise (and they always will), they can be solved if the infrastructure is built correctly
- Small device count makes it easier to find issues
- Multiple test and measurement hosts provide multiple views of the data path
  - perfSONAR nodes at the site and in the WAN
  - perfSONAR nodes at the remote site

Success Stories - Communities





Large Hadron Collider

- Multi-continent data distribution model
- Service interface to the network dramatically increases productivity
- Multiple petabytes per year

Climate Science

- Earth System grid spans multiple nations, institutions
- Automated data replication – terabytes to petabytes

Success Stories - Facilities

Supercomputer center Data Transfer Nodes

- Argonne, NERSC, Oak Ridge supercomputer centers
- Dedicated systems for high-speed data transfer

Per-project data transfer infrastructures

- Fusion data – China → US
- ALS experiment data → Canada