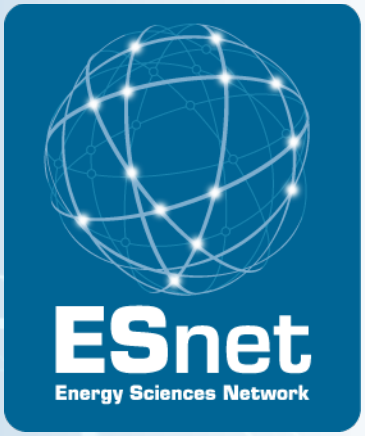


The Science DMZ:
Solving Data Mobility Challenges with the right architecture and the right tools

Jason Zurawski

Eli Dart

Mary Hester

Lauren Rotman

Brian Tierney

ESnet Science Engagement - engage@es.net

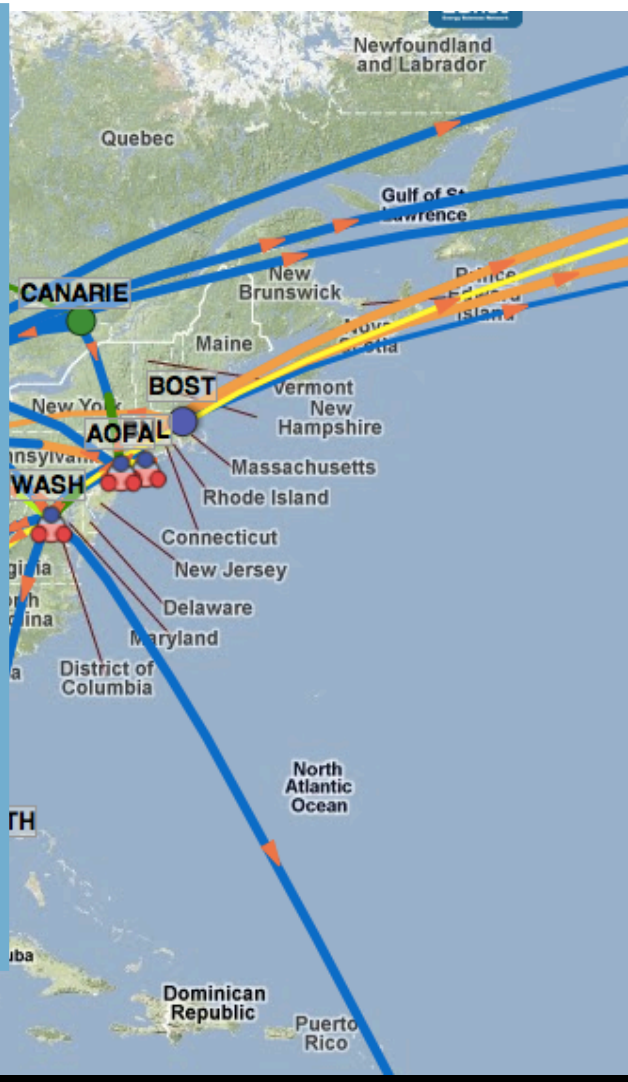
November 7, 2013

Overview

- What is ESnet?
- Why Use the Network (vs. hard drive, FedEx)?
- Understanding Network Performance and Expectations
- Science DMZ Overview and Architecture
- Integration with Globus Online
- Benefits to Science

What is ESnet?

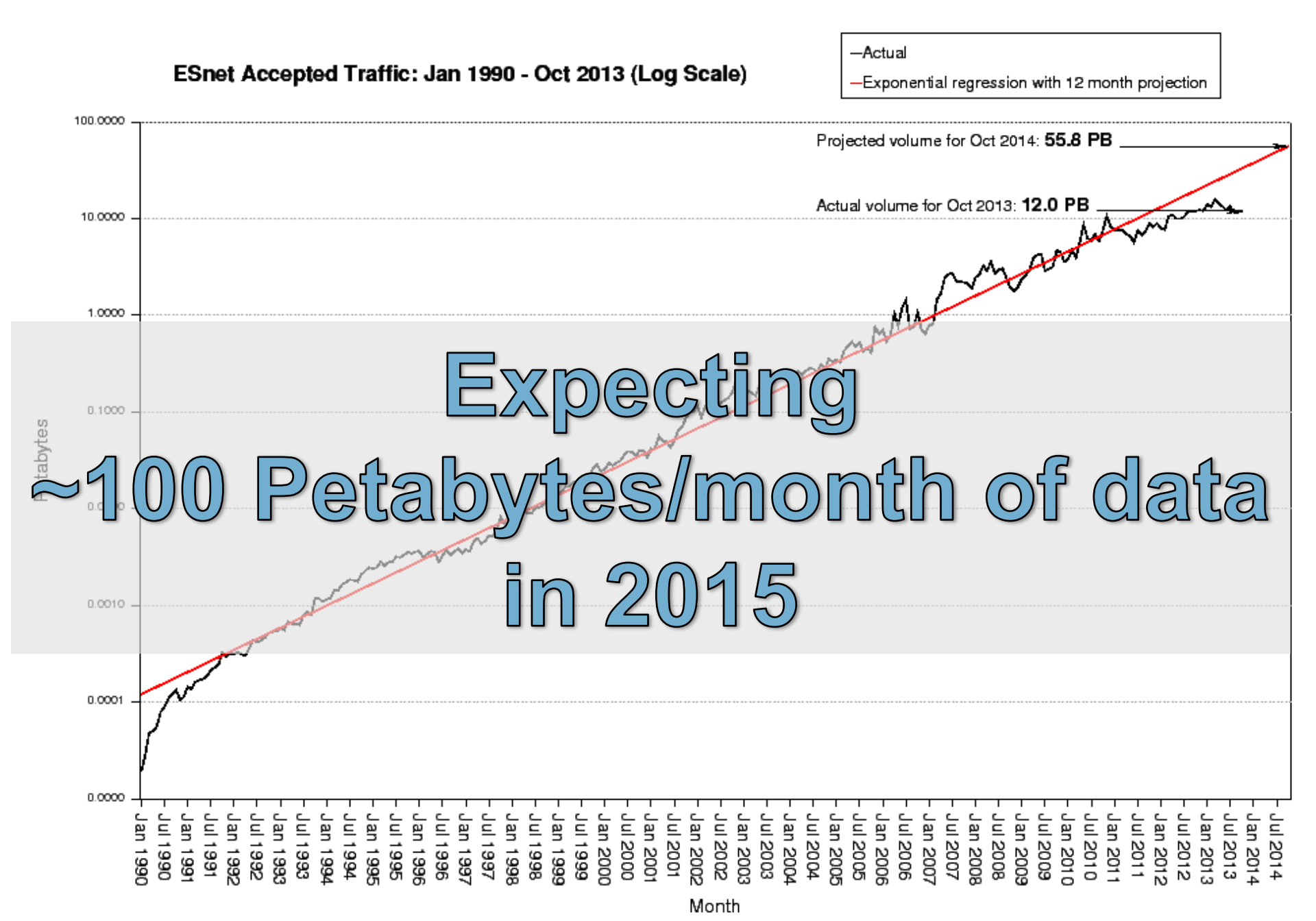
- A high-performance network linking DOE Office of Science researchers to global collaborators and resources around the world, including:
  - Supercomputer centers
  - User Facilities
  - Multi-program labs
  - Universities
  - Connectivity to Internet and Cloud providers
- A national DOE user facility providing:
  - Tailored data mobility solutions for science
  - *Dedicated Science Engagement team to support researchers*
  - Collaboration services e.g. audio/video conferencing



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



Expecting ~100 Petabytes/month of data in 2015

The Future: Data Mobility Services Over Networks

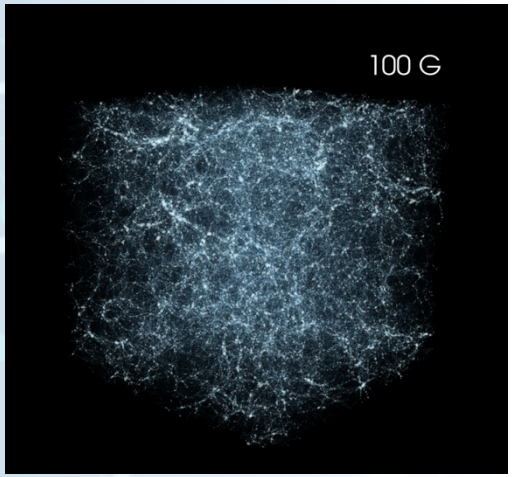
Data transfer between facilities is best achieved using networks

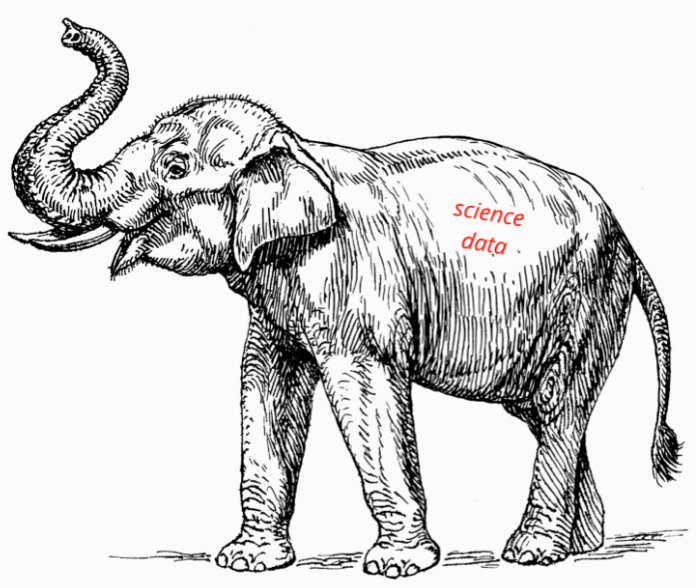
- Physical media transport is time-consuming and error-prone
- People physically transporting disks are a gating factor and reduce science productivity and output
- Multi-facility workflows using data mobility services have already proven successful
  - ✓ *The structure of modern science assumes that global research networks exist and function properly*

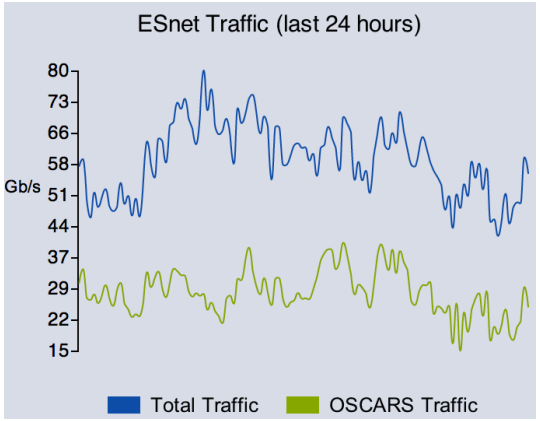
Reasons for use of long distance data services include:

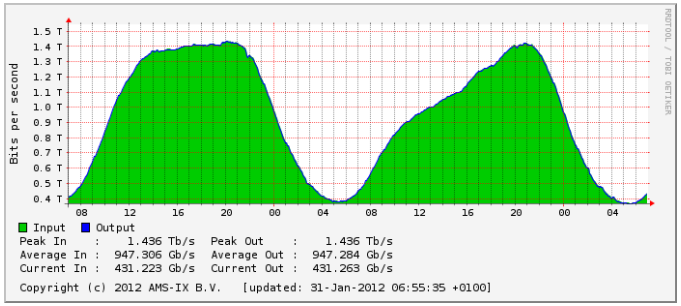
- Movement of data sets between HPC facilities
- Download historical data for model input/reference (e.g. weather)
- Ingest of experimental data for analysis
- Serve data sets or results to remote users

The Wide Area is Engineered for Elephants









Your Campus Network Should Be Engineered for Elephants too!!

Common Issues Impeding Network Adoption and Performance

Local Networks are typically built for business use cases, not science – which leads to:

Hardware:

- Firewalls, naively configured or deployed
- "Lossy" Networks
- Systems that aren't well configured, default settings

Software:

- Legacy data transfer tools (scp)

Culture:

- "We've always shipped disks"

Time to Raise Your Network Expectations:
Time to Copy 1 Terabyte

On a…

10 Mbps network: 300 hrs (12.5 days)

100 Mbps network: 30 hrs

1 Gbps network: 3 hrs (are your disks fast enough?)

10 Gbps network: 20 minutes (fast disks and fast filesystems)

These figures assume some headroom left for other users

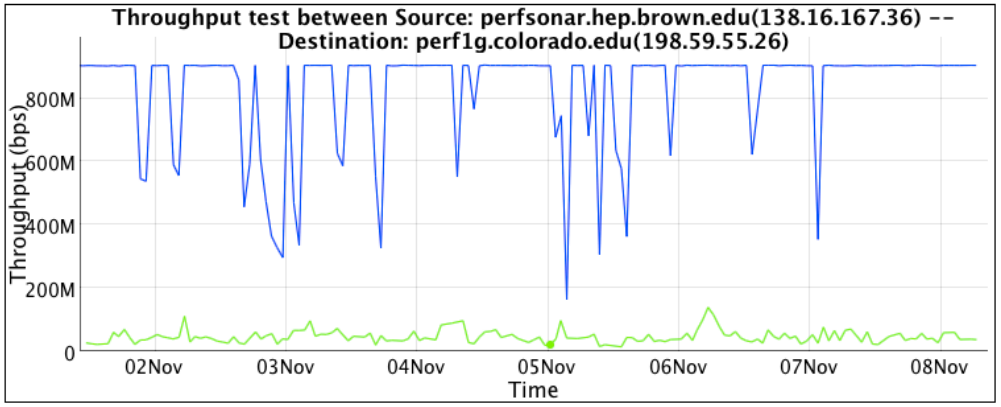
Compare these speeds to:

- USB 2.0 portable disk
  - 60 MB/sec (480 Mbps) peak
  - 20 MB/sec (160 Mbps) reported on line
  - 5-10 MB/sec reported by colleagues
  - 15-40 hours to load 1 Terabyte

https://fasterdata.es.net/home/requirements-and-expectations/

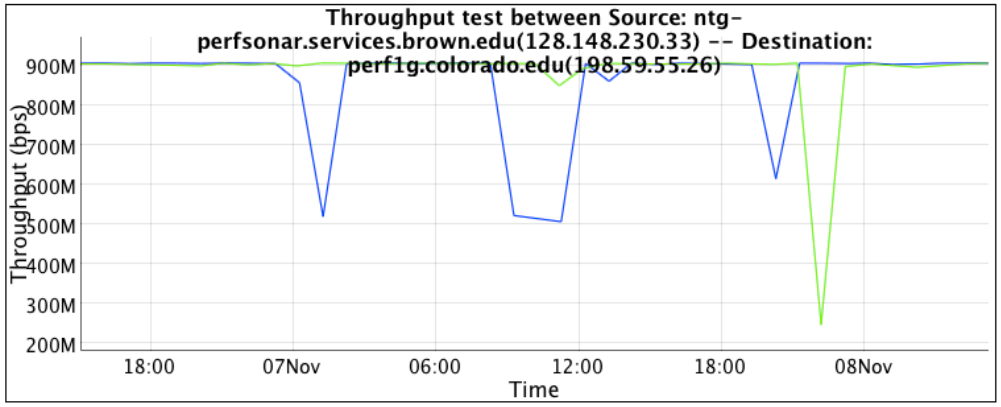
Firewall Performance Example

Observed performance, via perfSONAR, through a firewall:



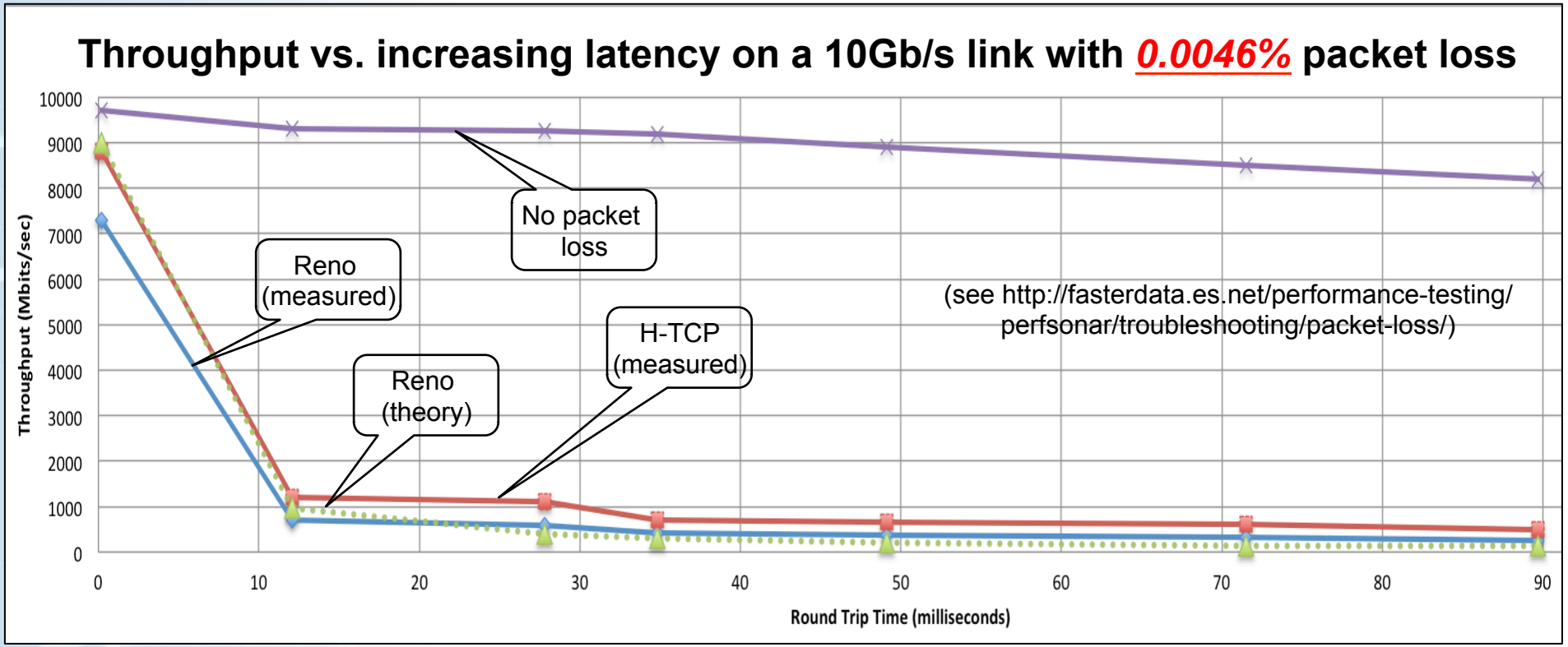
Almost 20 times slower!

Observed performance, via perfSONAR, bypassing firewall:



Traffic was unimpeded by additional processing or resource constraints

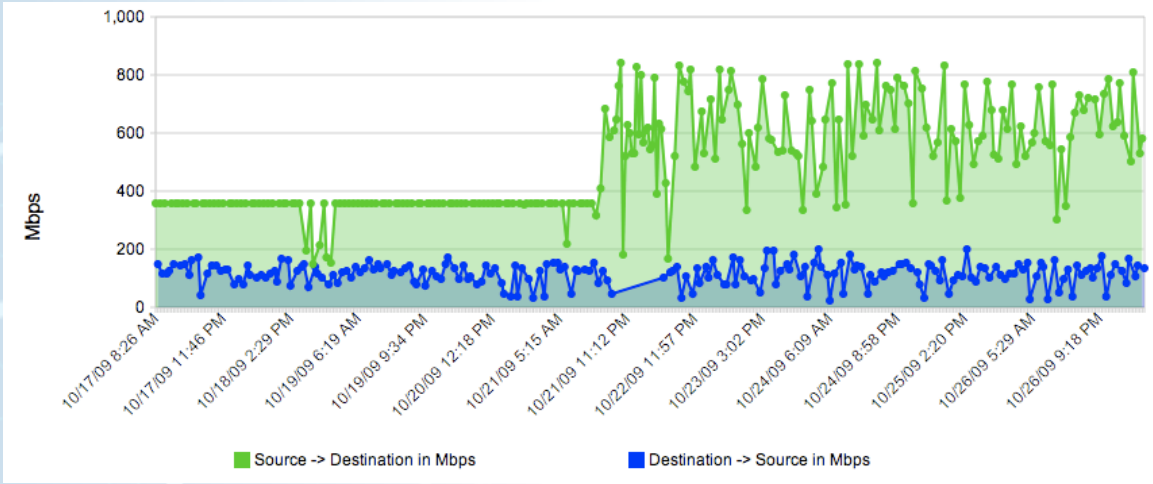
A small amount of packet loss makes a huge difference in TCP performance



- On a 10 Gb/s LAN path the impact of low packet loss rates is minimal
- On a 10Gb/s WAN path the impact of low packet loss rates is enormous

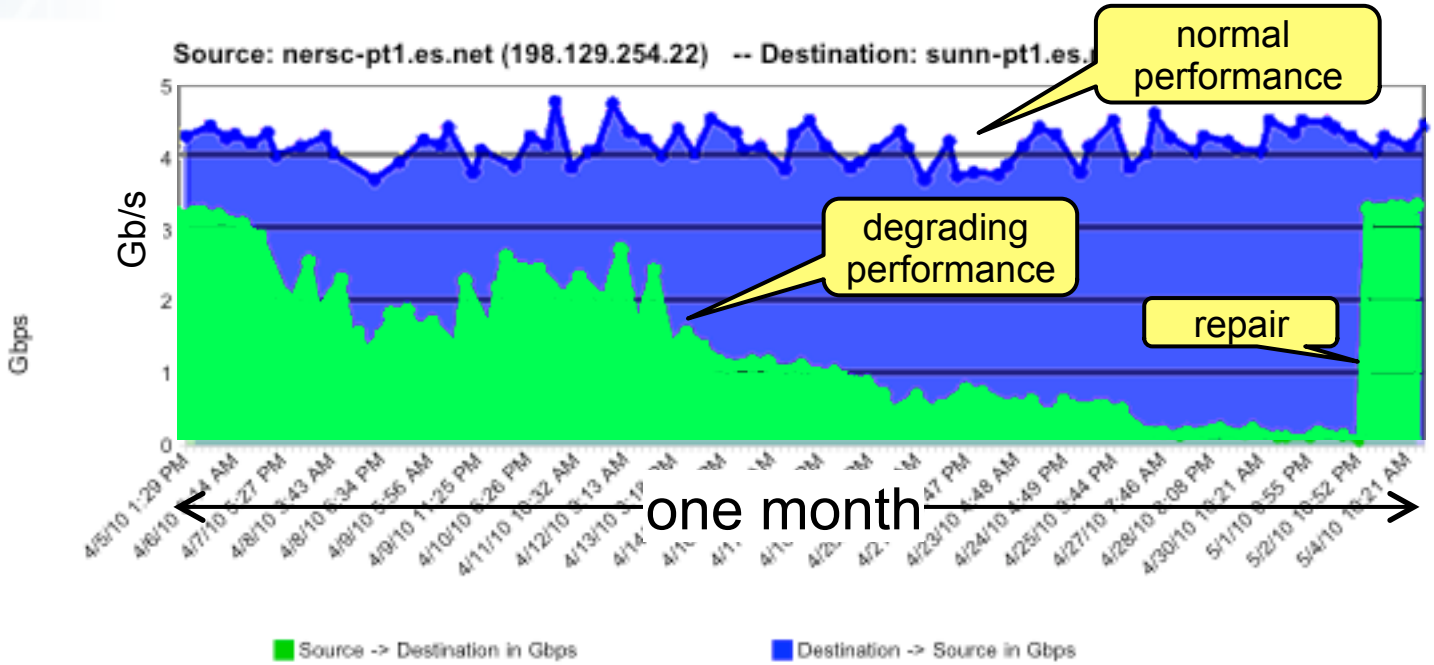
- ***Implications***: error-free paths are essential for high-volume data transfers

Other Soft Failures



Host Configuration – spot when the TCP settings were tweaked…

Gradual failure of optical line card



Legacy Data Transfer Tools

In addition to the network, using the right data transfer tool is critical

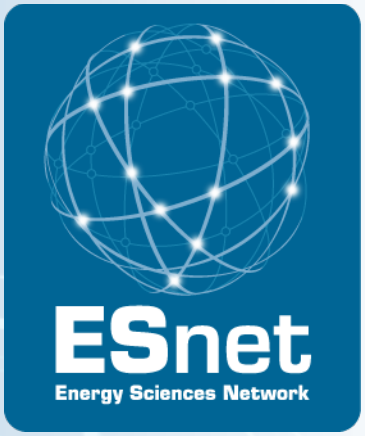
Sample Results:

Data transfer from Berkeley, CA to Argonne, IL (near Chicago).
RTT = 53 ms, network capacity = 10Gbps.

Tool	Throughput
scp:	140 Mbps
HPN patched scp:	1.2 Gbps
ftp	1.4 Gbps
GridFTP, 4 streams	5.4 Gbps
GridFTP, 8 streams	6.6 Gbps

globus online

Note that to get more than 1 Gbps (125 MB/s) disk to disk requires RAID (e.g. data distributed over multiple disks and accessed in parallel).



Achieving these results:

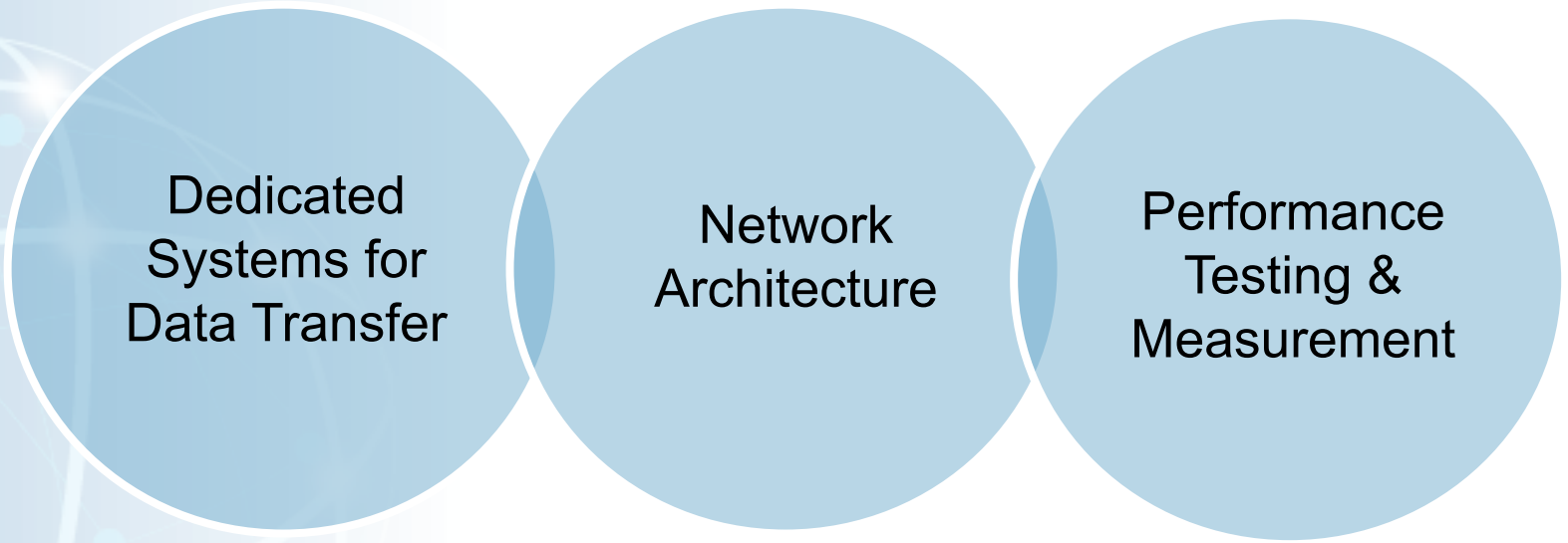
Globus Online-enabled The Science DMZ

What is the Science DMZ?

The Science DMZ is a model for network architectures, system configurations, cybersecurity policies, and performance tools that together create an optimized environment for data-intensive science.

The Science DMZ meets the specific requirements of science applications – “the elephants,” which have entirely unique characteristics compared to “general purpose” or business-oriented applications.

The Data Transfer Trifecta: The "Science DMZ" Model



Data Transfer Node

- High performance
- Configured for data transfer
- Proper tools like Globus Online

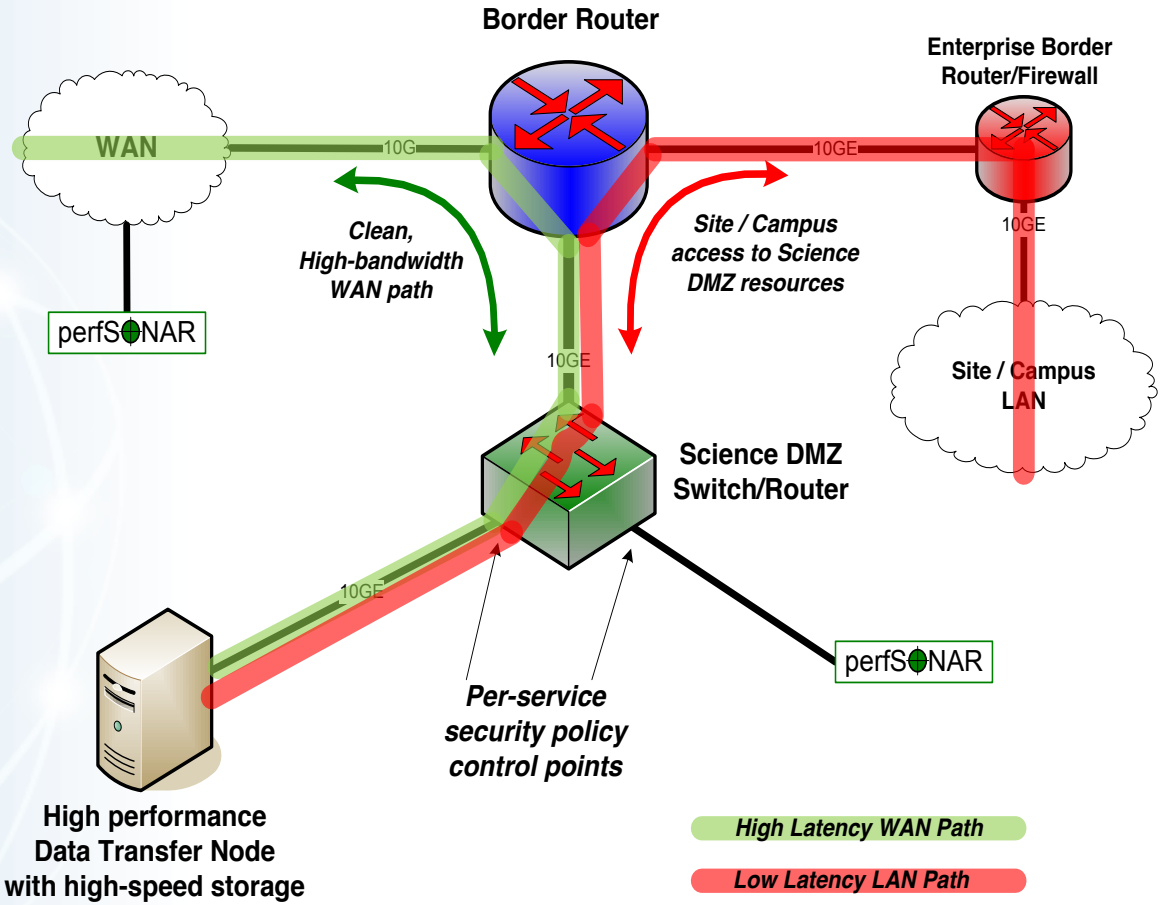
Implementation Location

- 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

Many Universities and Labs are now creating Science DMZs

- NSF programs (CC-NIE) are funding campuses to build a Science DMZ
- Joining the Internet2 “innovation platform” requires having a Science DMZ (http://www.internet2.edu/vision-initiatives/initiatives/innovation-platform/)

Globus Online 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

Dedicated Hardware

- Highly capable hardware and tuned software
- http://fasterdata.es.net/science-dmz/DTN/

For More Information

http://fasterdata.es.net/science-dmz/

Operating Innovative Networks (OIN)

- http://www.oinworkshop.com

Globus Tutorial @ SC13:

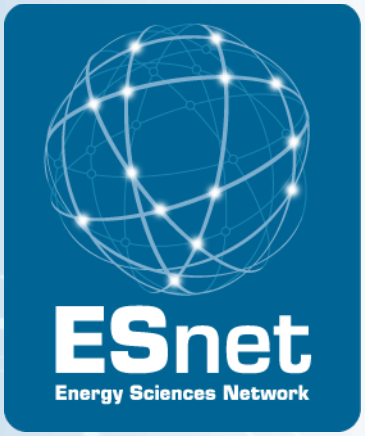
- http://sc13.supercomputing.org/schedule/event_detail.php?evid=tut170

Paper on the Science DMZ to be presented at SC13:

- http://www.es.net/assets/pubs_presos/sc13sciDMZ-final.pdf

Questions? email engage@es.net

Future Event Alerts: https://gab.es.net/mailman/listinfo/fasterdata-events



Extra Slides

Photon Science Data Increase

Many detectors are semiconductors

- Similar technology to digital cameras
- Exponential growth
- Increase in sensor area (512x512, 1024x1024, 2048x2048, …)
- Increase in readout rate (1Hz, 10Hz, 100Hz, 1kHz, 1MHz, …)

Data infrastructure needs significant change/upgrade

- Most photon scientists are not “computer people”
  - Different from HEP, HPC centers
  - They need data issues solved – they don’t want to solve them
  - ***They should not have to be come network experts!***
- Physical transport of portable media has reached a breaking point
- Default configs no longer perform well enough

ALS Beamline 8.3.2

Broad science portfolio: Applied science, biology, earth sciences, energy, environmental sciences, geology, cosmological chemistry

Detector upgrade → large increase in data rate/volume (50x)

Detector output: sets of large TIFF files

Beamline scientist Dula Parkinson reached out to LBLnet

LBLnet reached out to ESnet

Infrastructure improvements

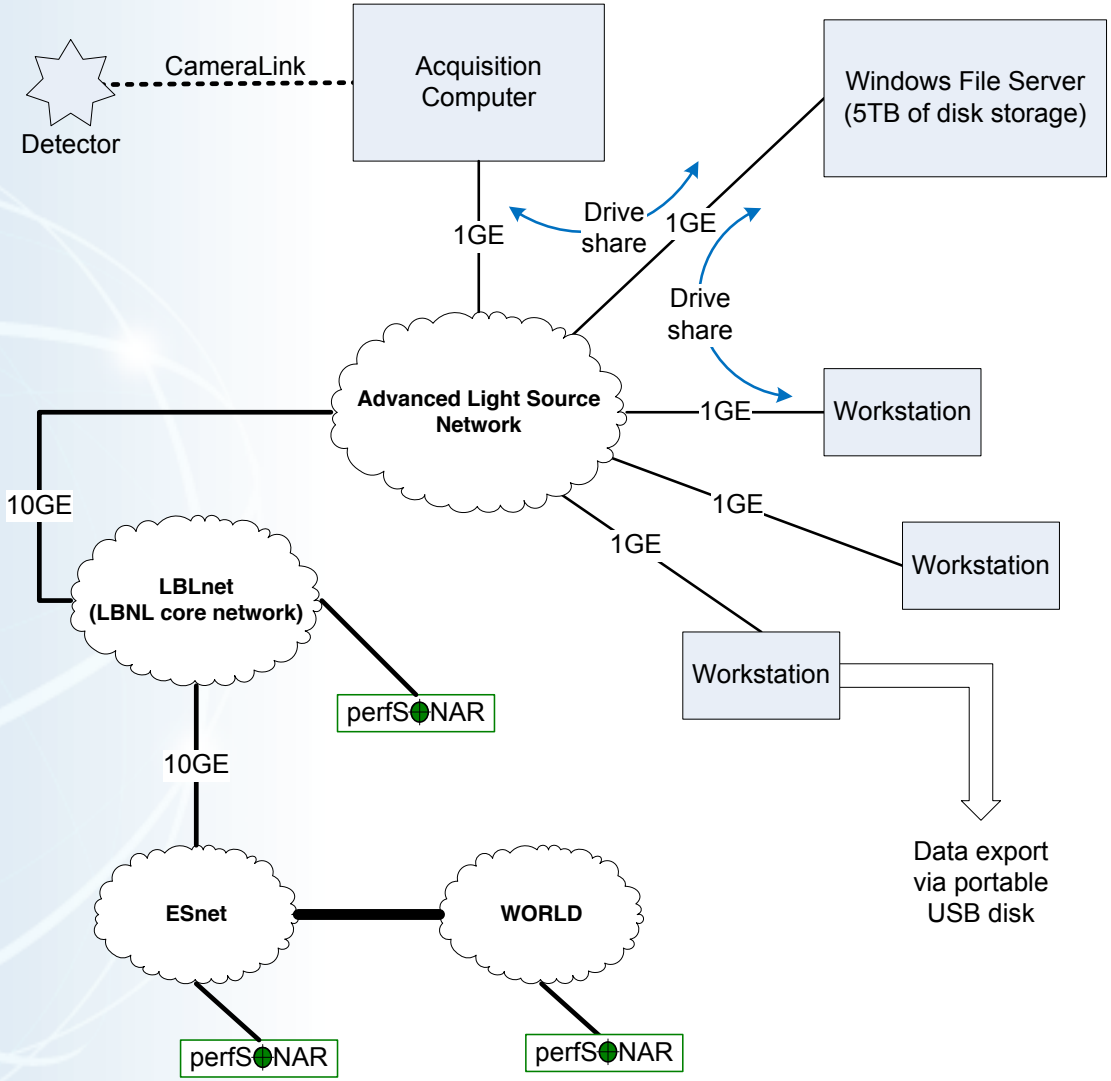
- Used perfSONAR to find failing router line card
- DTN built from Fasterdata reference design

NERSC collaboration

- Data workflow (python scripts, etc.)
- Data analysis

Collaboration is ongoing

Original Workflow Infrastructure



Improved Workflow Infrastructure