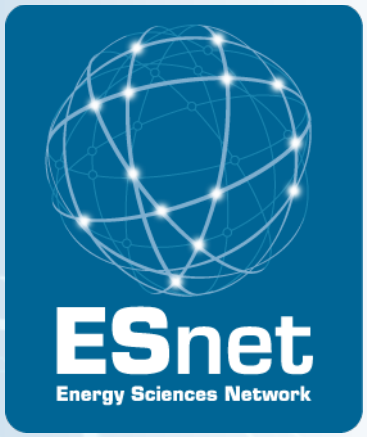


Overview of the Science DMZ

Brian Tierney and Eli Dart, ESnet

QuestNet 2013

Gold Coast, Australia

July 4, 2013

What's there to worry about?

Overview

- What is ESnet?
- What is a Science DMZ?
  - Architecture
  - perfSONAR
  - Data Transfer Nodes
  - Security issue for a Science DMZ

The Energy Sciences Network (ESnet)
A Department of Energy Facility

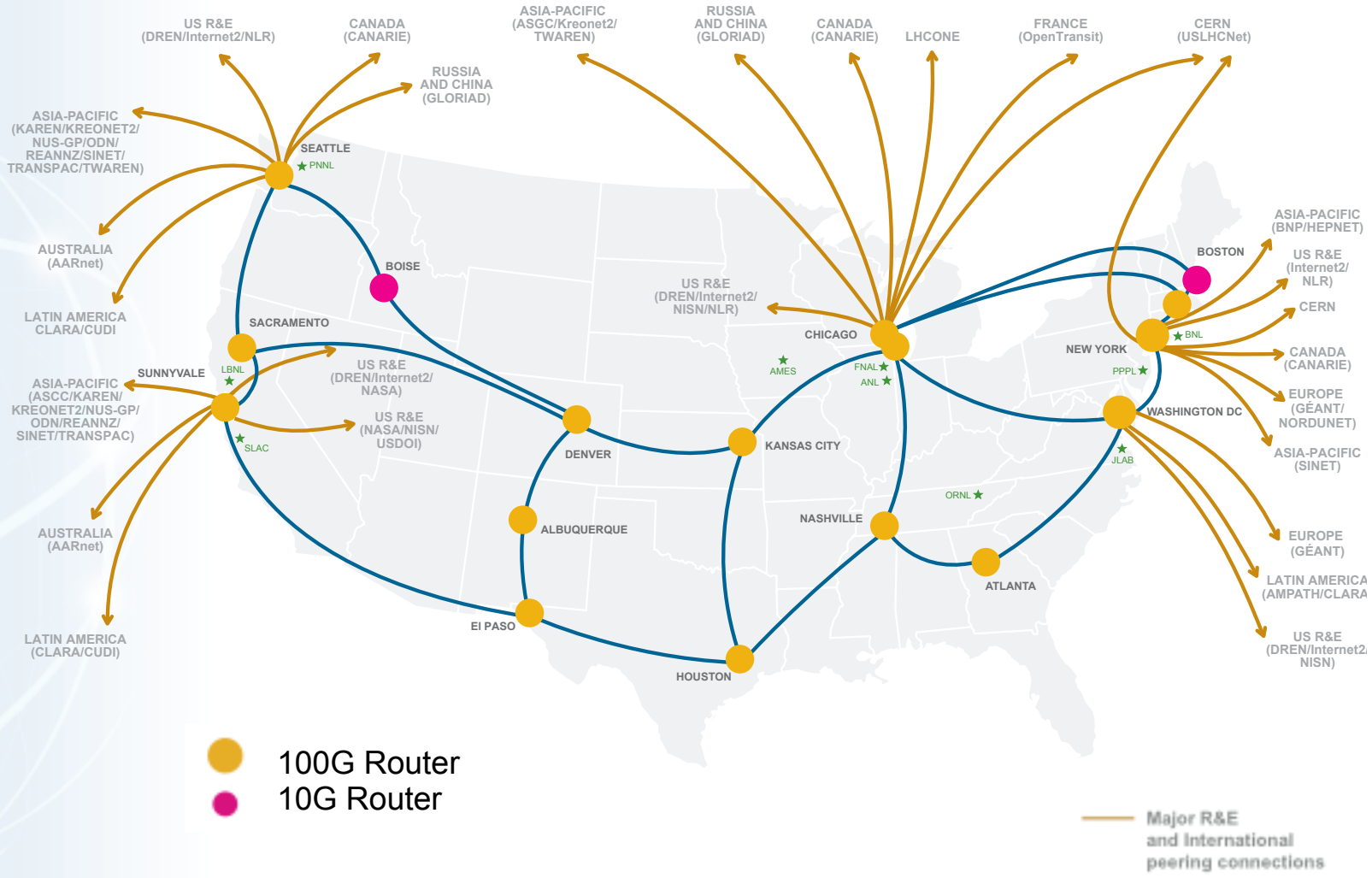
National Fiber footprint

Distributed Team of 40

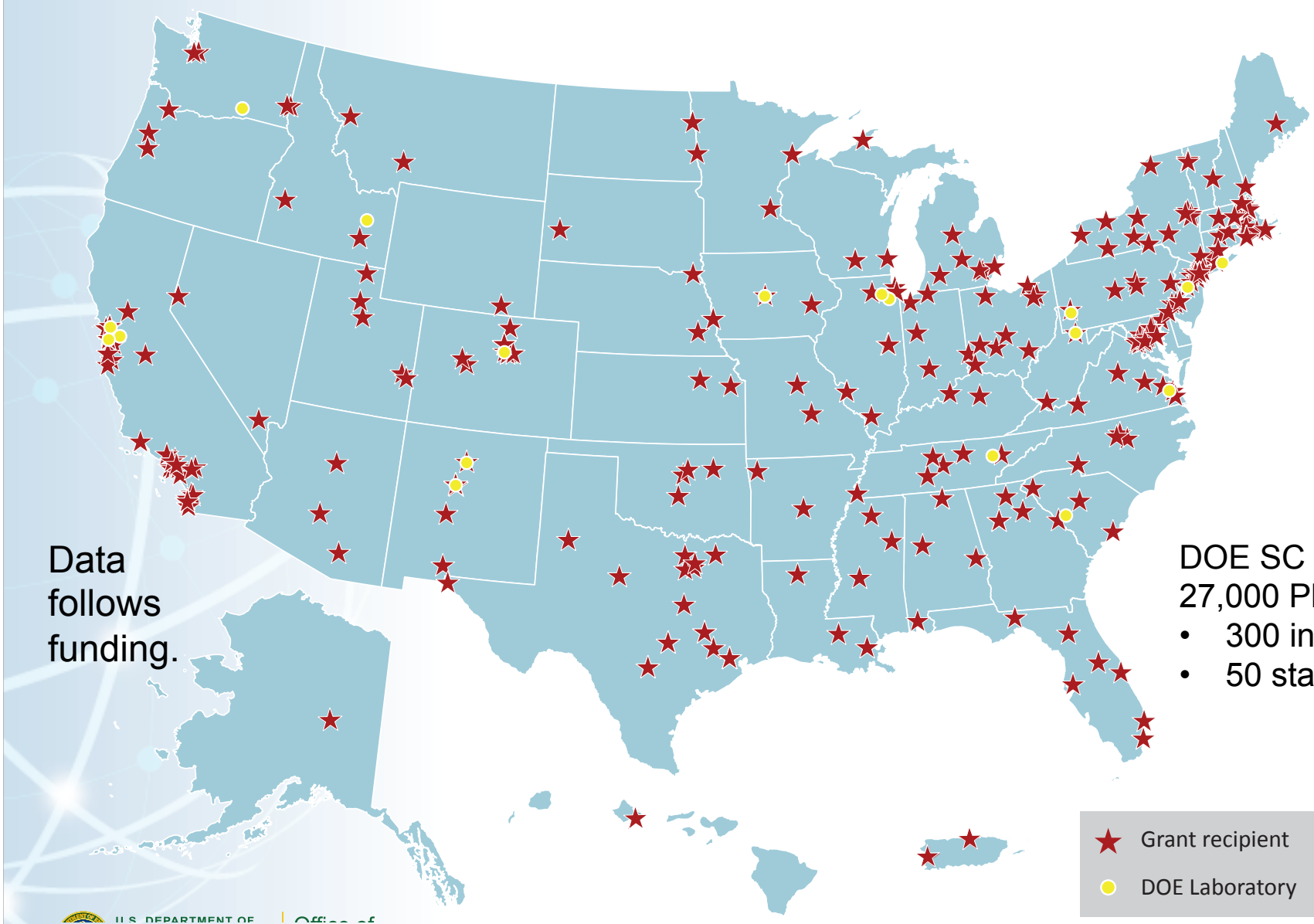
Science Data Network

International Collaborations

100G Backbone



Most ESnet Traffic Leaves DOE Complex

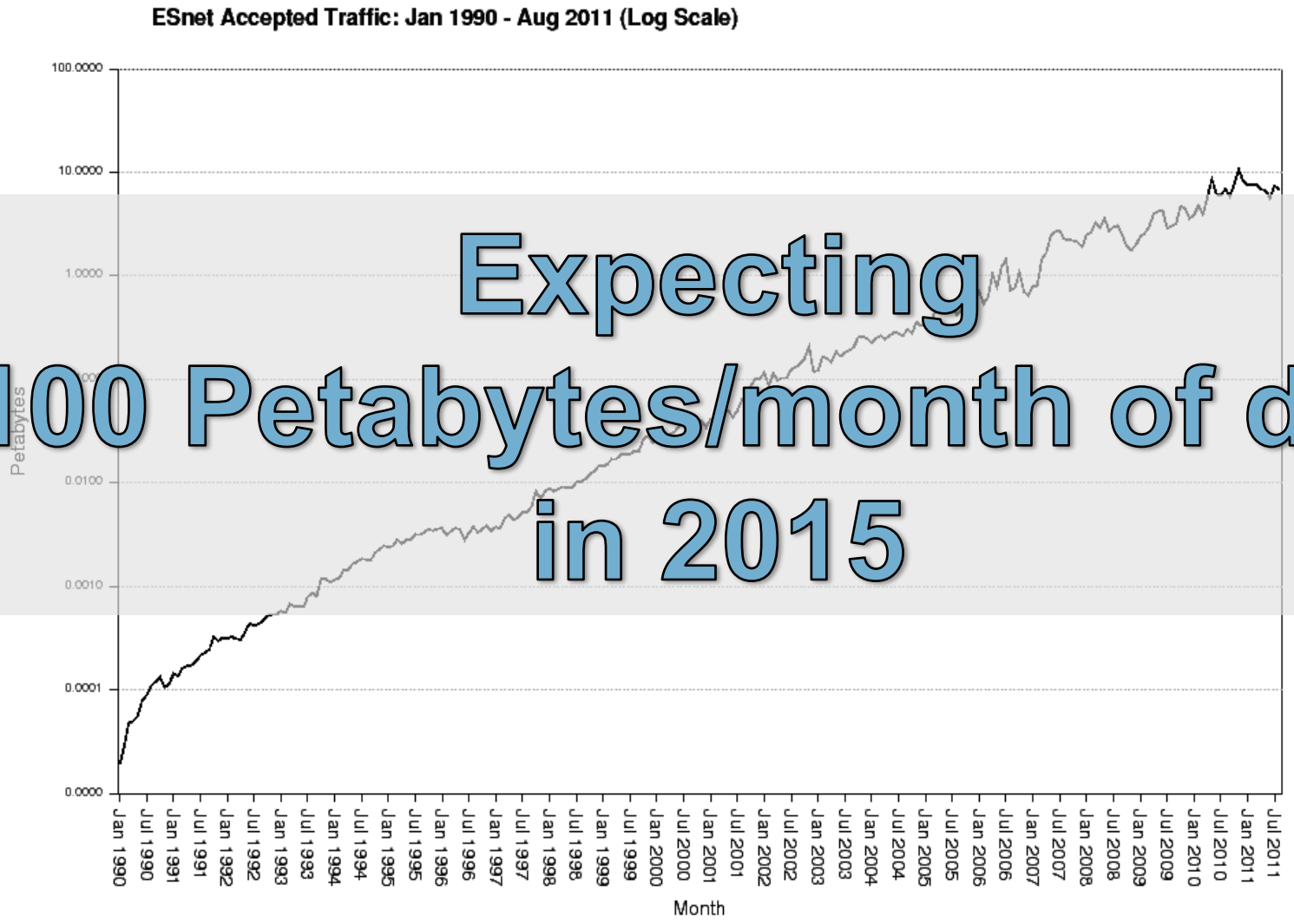


Data follows funding.

DOE SC supports 27,000 Ph.D.s

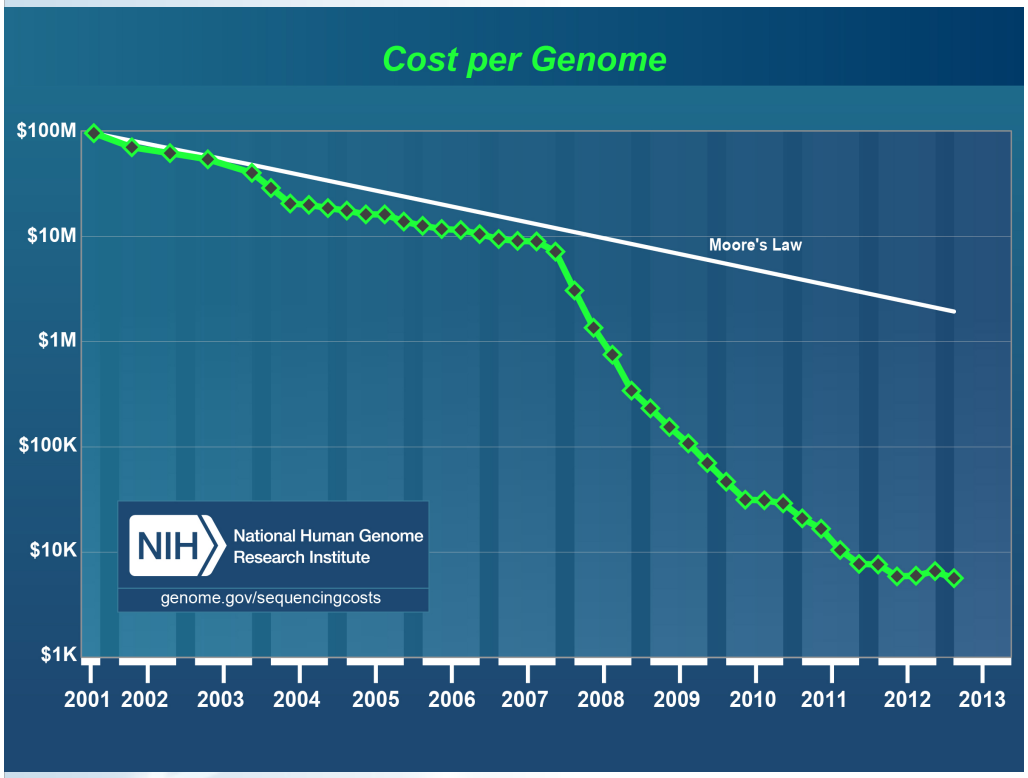
- 300 institutions
- 50 states

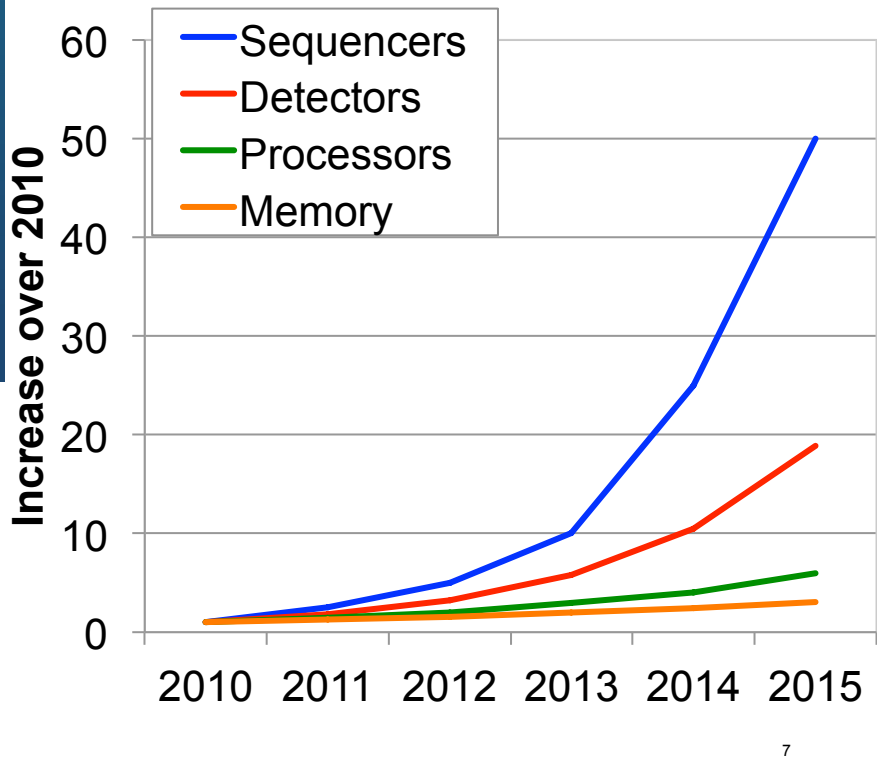
The Science Data Explosion



Expecting 100 Petabytes/month of data in 2015

Sample Data Growth





Raising Expectations: Time to Copy 1 Terabyte

10 Mbps network : 300 hrs (12.5 days)

100 Mbps network : 30 hrs

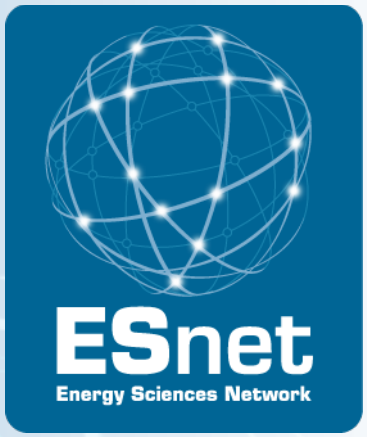
1 Gbps network : 3 hrs (are your disks fast enough?)

10 Gbps network : 20 minutes (requires RAID disk array)

These figures assume some headroom left for other users

Compare these speeds to:

- USB portable disk
  - 20-25 hours to load 1 Terabyte



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

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 RFC793, 1981)

Some soft failures only affect high bandwidth long RTT flows.

Hard failures are easy to detect & fix

- soft failures can lie hidden for many months!

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

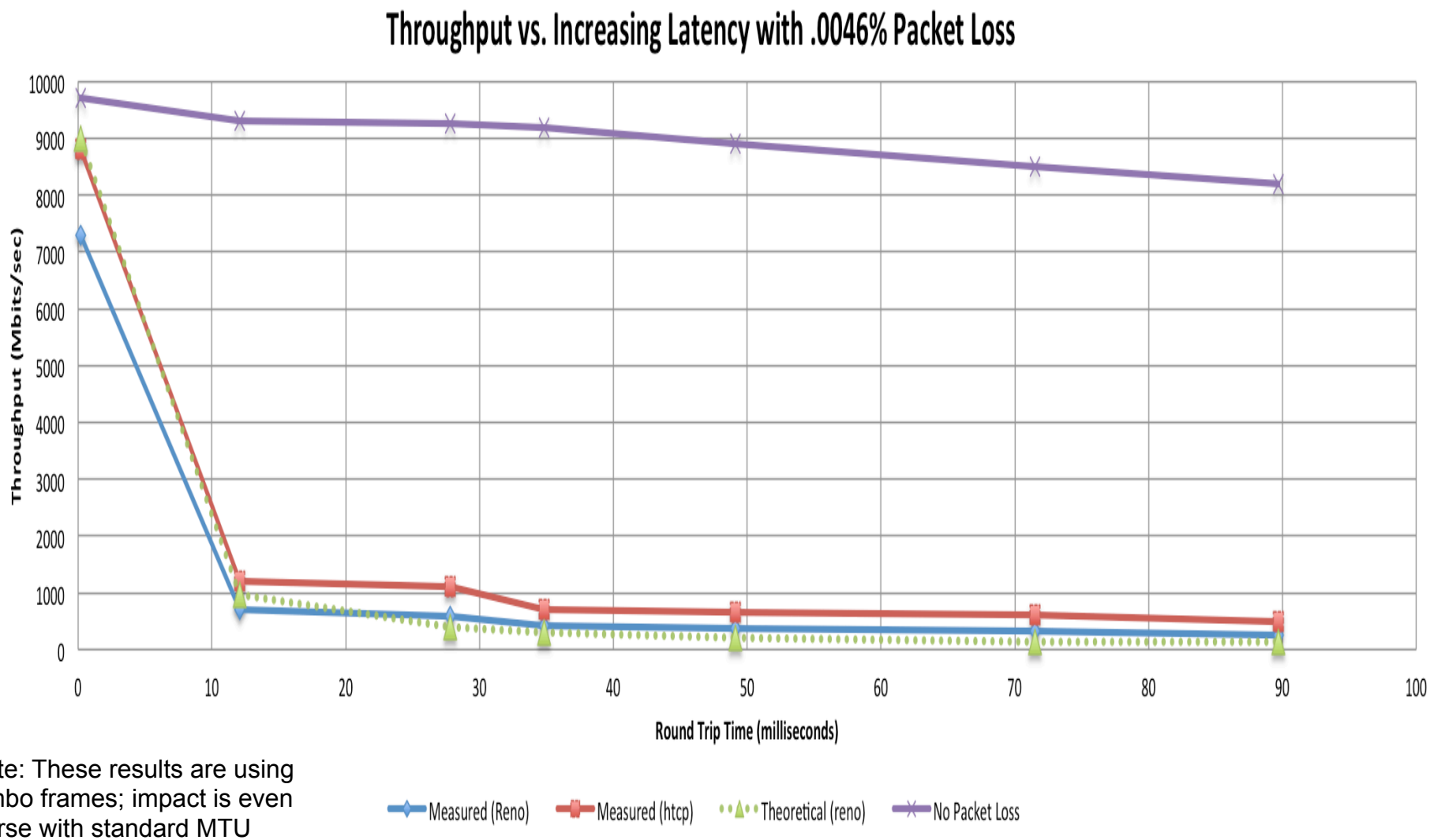
Small Router/Switch Buffers

- 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

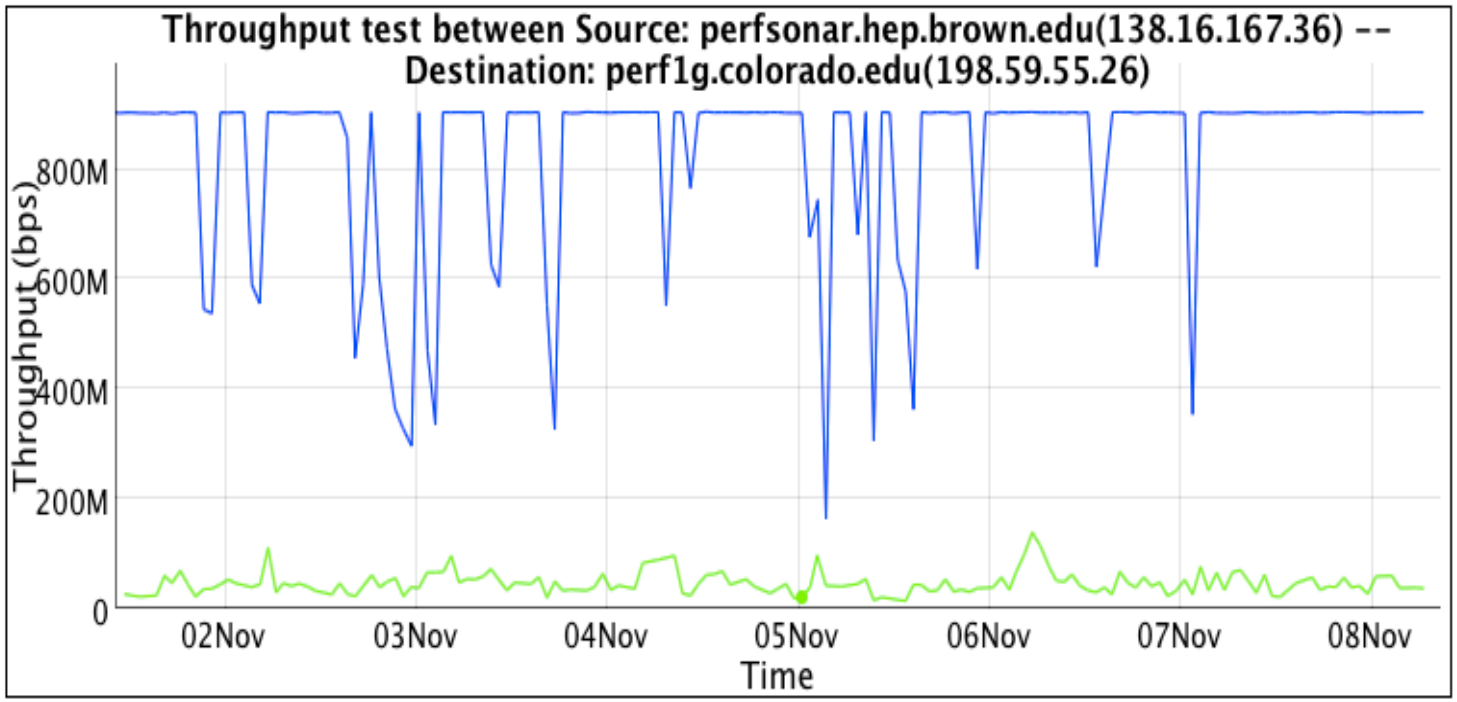
- Processor-based switching on routers due to faults, acl's, or mis-configuration

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



Note: These results are using jumbo frames; impact is even worse with standard MTU

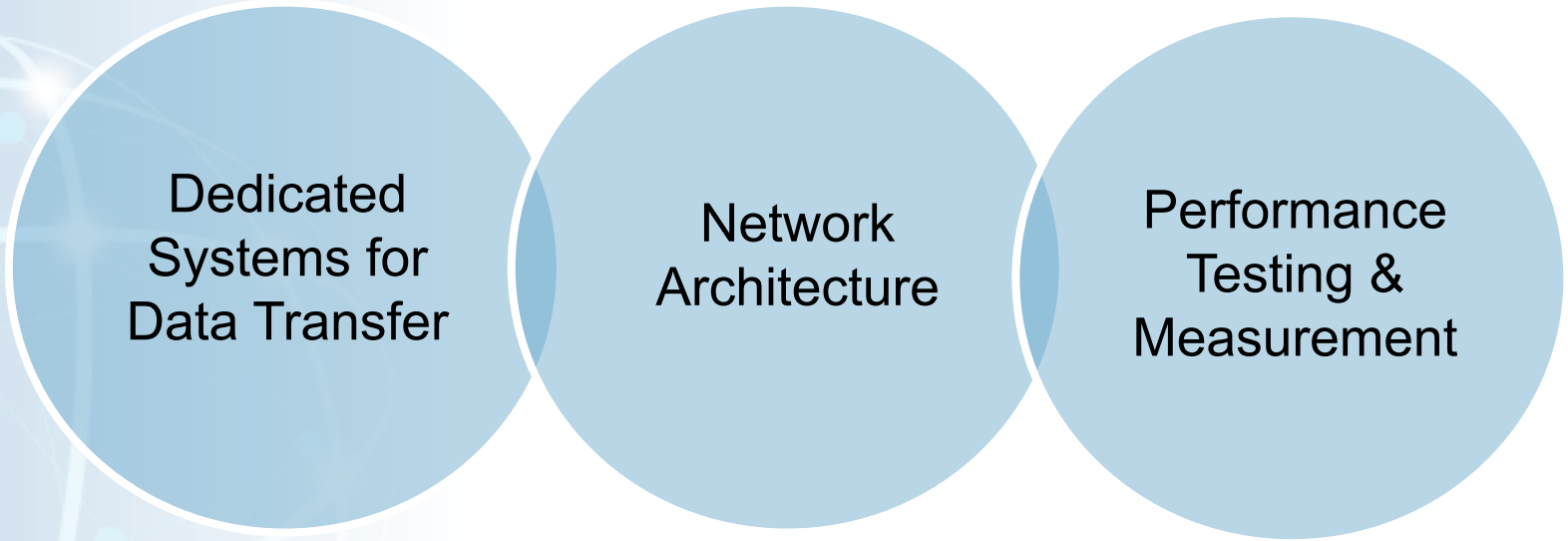
Another Common Problem: the Firewall



Blue = outgoing, green = incoming

This flow is between two 1G hosts on a 10G network

The Data Transfer Trifecta: The "Science DMZ" Model



Data Transfer Node

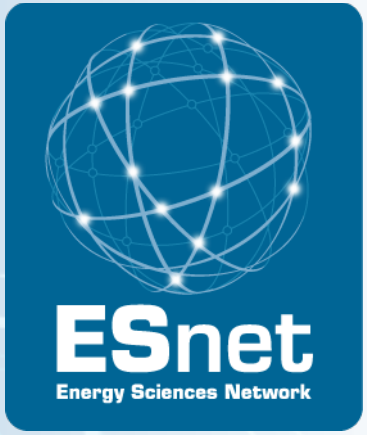
- High performance
- Configured for data transfer
- Proper tools

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



Science DMZ Architecture

Science DMZ Takes Many Forms

Goal of the Science DMZ Architecture:

- Deploy resources as close to the border router as possible
  - Fewer network hops = fewer potential sources of problems
- Bypass inline packet inspection devices (firewalls, IPS, shapers) them impede high bandwidth flows

There are many ways to combine the Science DMZ elements – it all depends on site requirements

- Small installation for a project or two
- Facility inside a larger institution
- Institutional capability serving multiple departments/divisions

Many Universities and Labs are now creating Science DMZs

- In the US, a NSF program (CC-NIE) is funding several campuses to build a Science DMZ
- In Australia, the RDSI is creating a Science DMZ at several large campuses

Science DMZ Security

Goal – disentangle security policy and enforcement for science flows from that of business systems

Rationale

- Science flows are relatively simple from a security perspective
- Narrow application set on Science DMZ hosts
  - Data transfer, data streaming packages
  - Performance / packet loss monitoring tools
  - No printers, document readers, web browsers, building control systems, staff desktops, etc.
- Security controls that are typically implemented to protect business resources often cause performance problems
- Sizing security infrastructure on designed for business networks to handle large science flows is *expensive*

In Big Data Science, Performance Is a Core Requirement Too

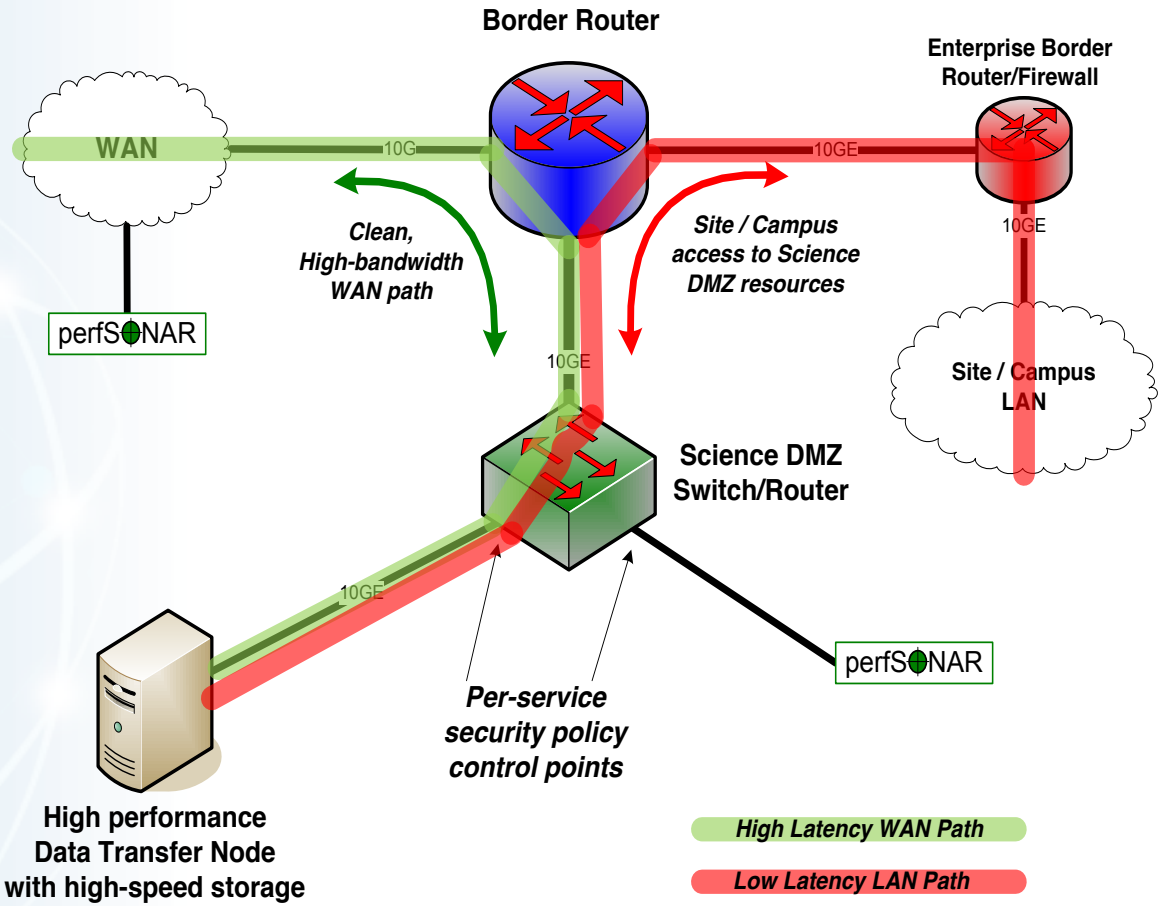
Core information security principles:

- Confidentiality, Integrity, Availability (CIA)

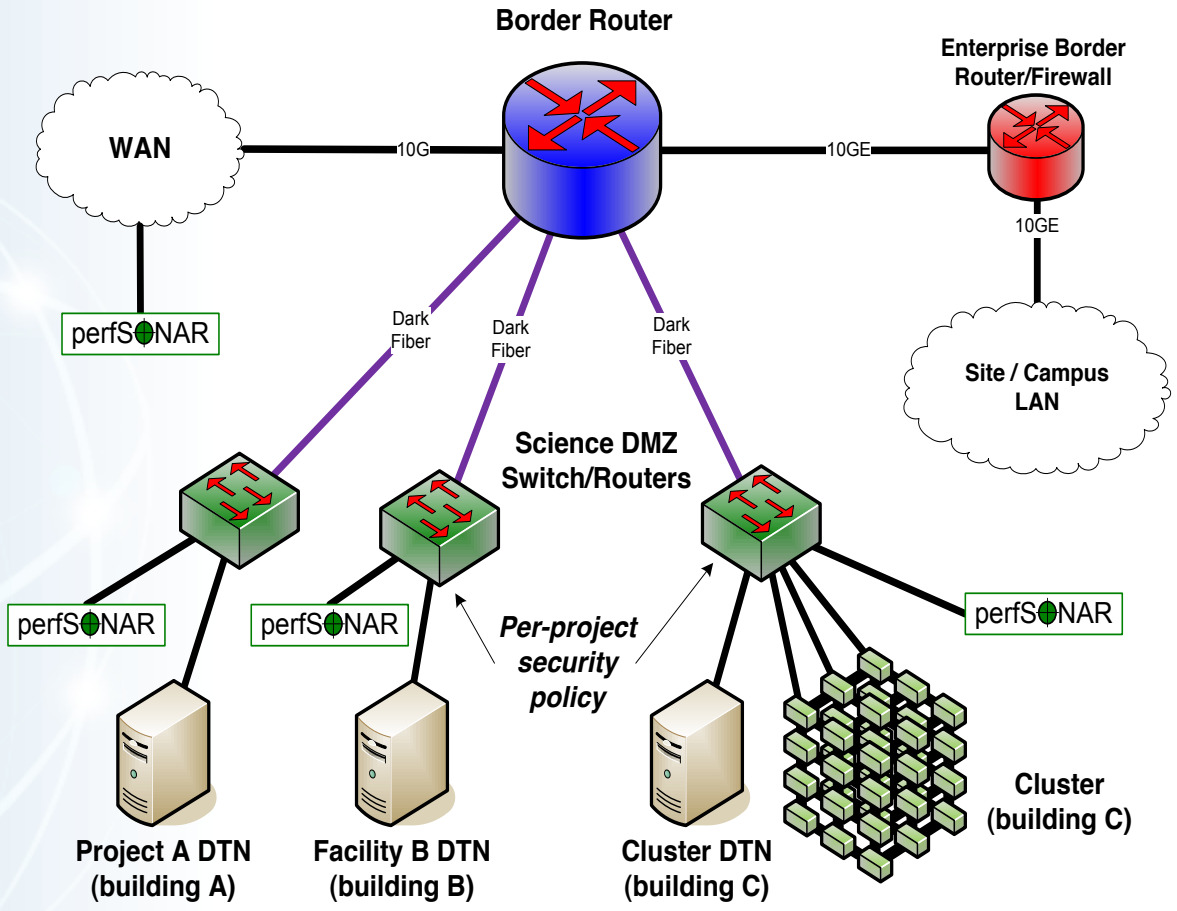
In data-intensive science, **performance** is an additional core mission requirement (**CIAP**)

- CIA principles are important, but ***if the performance isn't there the science mission fails***
  - This isn't about "how much" security you have, but how the security is implemented
  - We need to be able to appropriately secure systems in a way that does not compromise performance

Simple Science DMZ



Multiple Science DMZs – Dark Fiber



Security Without Firewalls

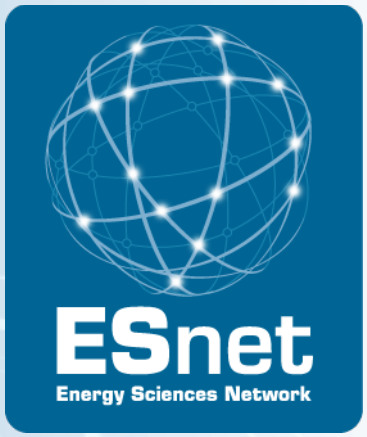
Does this mean we ignore security? ***NO!***

- We **must** protect our systems
- We just need to find a way to do security that does not prevent us from getting the science done

Lots of other security tools are still available:

- Host-based IDS and firewalls
- Intrusion detection (Bro, Snort, etc.), flow analysis, …
- Tight ACLs to reduce attack surface (possible in many but not all cases)
- Blackhole routing

Since performance is a mission requirement, the security policies and mechanisms that protect the Science DMZ should be architected so that they serve the mission



perfSONAR

What is perfSONAR?

perfSONAR is a set of tools to:

- Set network performance expectations
- Find network problems (“soft failures”)
  - Isolate the cause of the problems
  - Make sure the problem stays fixed

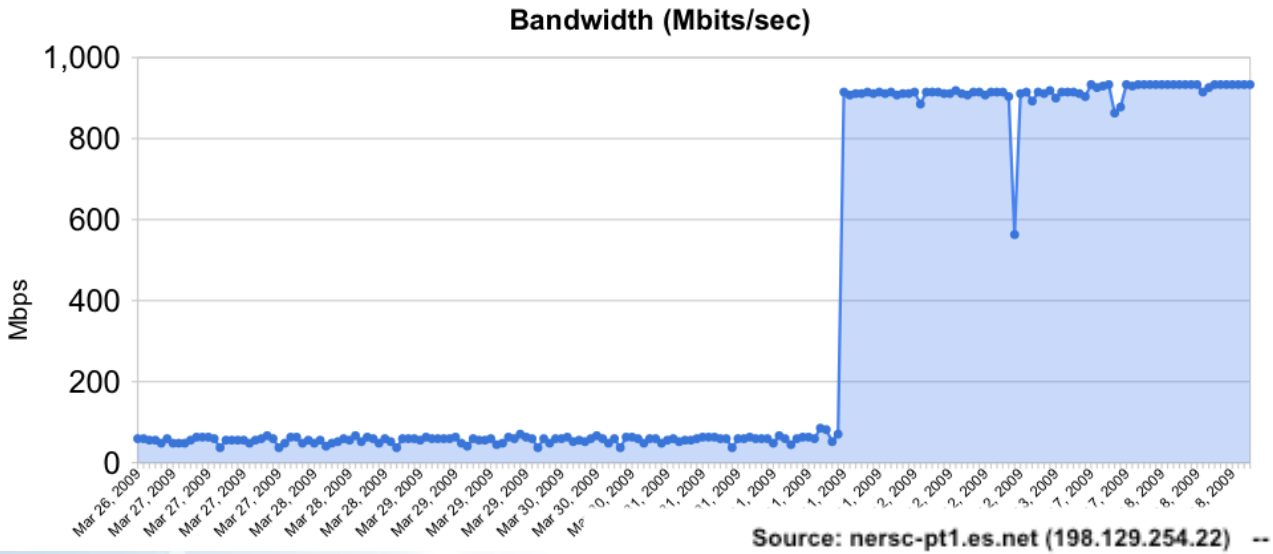
All in multi-domain environments

- These problems are all harder when multiple networks are involved

perfSONAR is provides a standard way to publish active and passive monitoring data

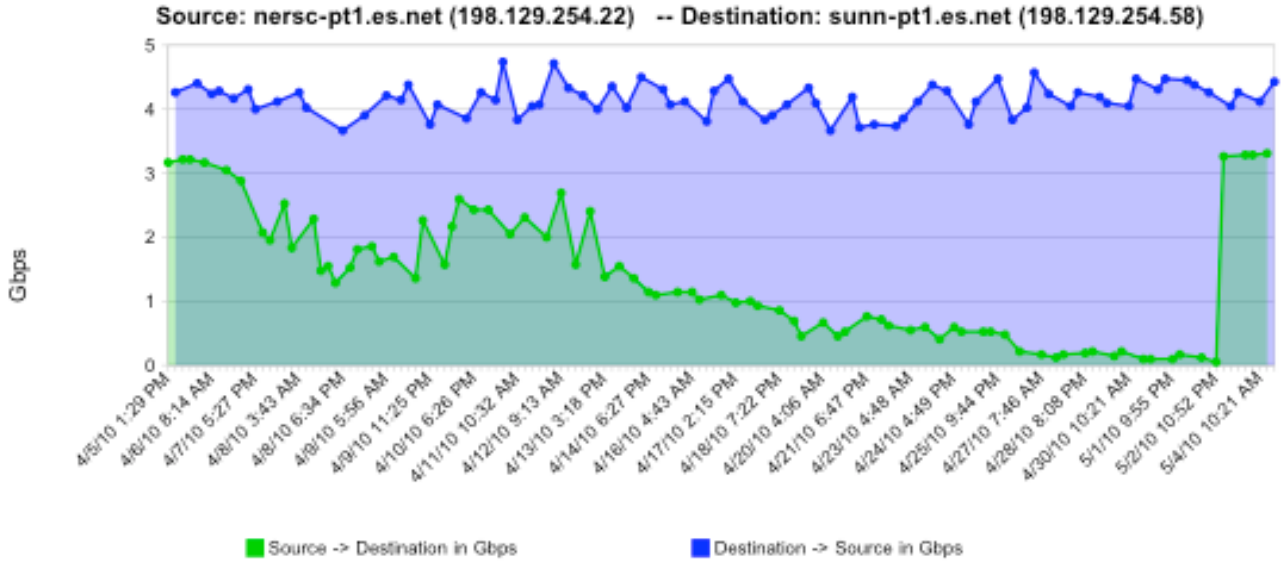
- This data is interesting to network researchers as well as network operators

perfSONAR Results: Sample Soft Failures as seen by perfSONAR



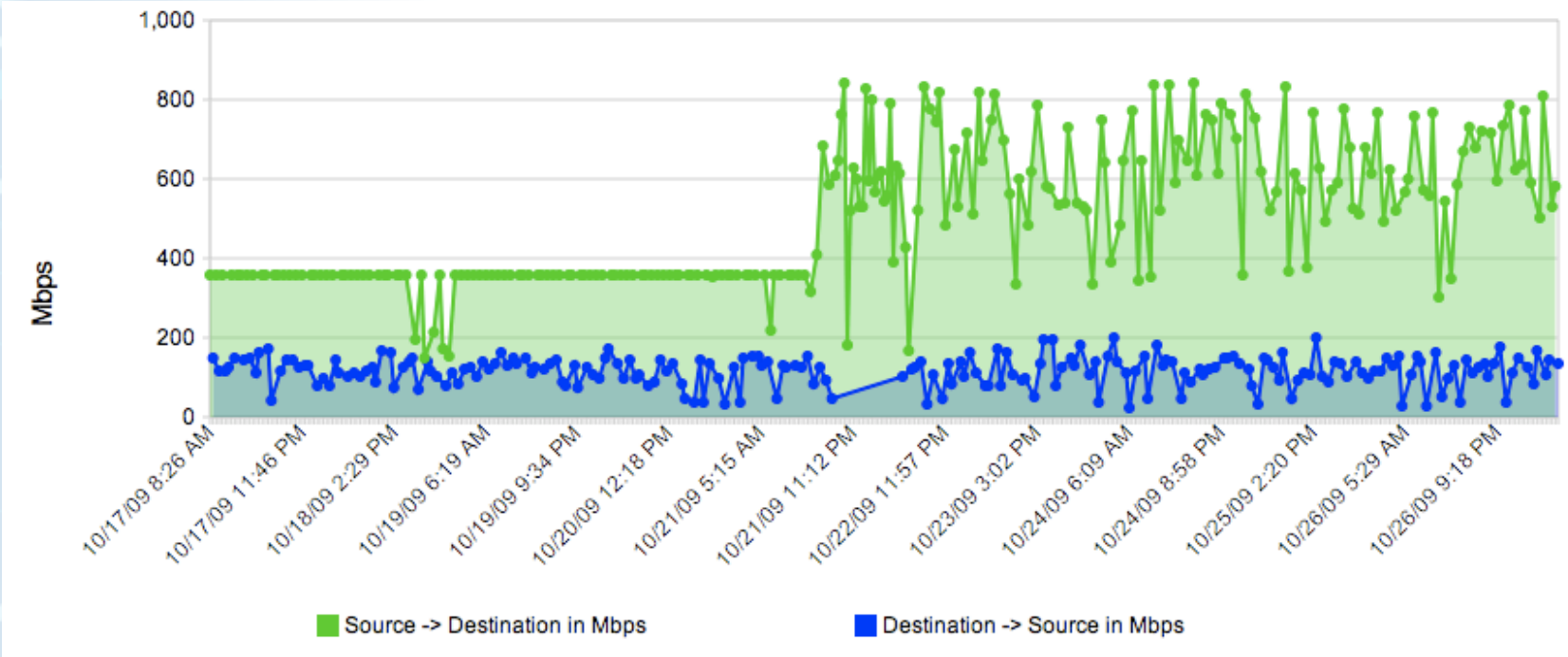
Rebooted router with full route table

Gradual failure of optical line card



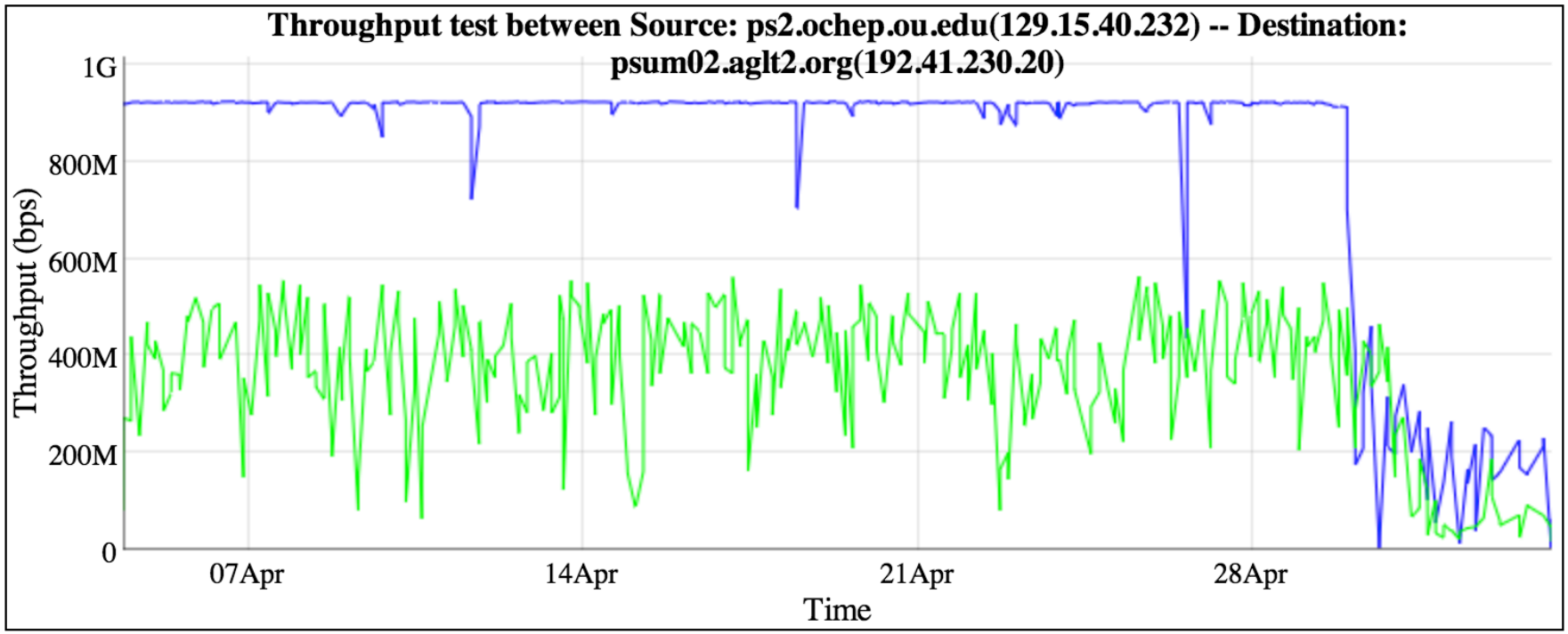
Host Tuning Example

- Host Configuration – spot when the TCP settings were tweaked…



- Example Taken from REDDnet (UMich to TACC, using BWCTL measurement)
- Host Tuning: http://fasterdata.es.net/fasterdata/host-tuning/linux/

What Monitoring Can (and Cannot) Tell You



Can you tell what is going on here?

perfSONAR Toolkit

The “perfSONAR Toolkit” is an open source implementation and packaging of the perfSONAR measurement infrastructure and protocols from ESnet and Internet2

http://psps.perfsonar.net

All components are available as RPMs, and bundled into a CentOS 6-based “netinstall” and a “Live CD”

- perfSONAR tools are much more accurate if run on a dedicated perfSONAR host, not on the DTN

Very easy to install and configure

- Usually takes less than 30 minutes

perfSONAR Toolkit Services

PS-Toolkit includes these measurement tools:

- BWCTL: network throughput
- OWAMP: network loss, delay, and jitter
- traceroute

Test scheduler:

- runs bwctl, traceroute, and owamp tests on a regular interval

Measurement Archives (data publication)

- SNMP MA – router interface Data
- pSB MA -- results of bwctl, owamp, and traceroute tests

Lookup Service: used to find services

PS-Toolkit includes these web100-based Troubleshooting Tools

- NDT (TCP analysis, duplex mismatch, etc.)
- NPAD (TCP analysis, router queuing analysis, etc)

Toolkit Web Configuration Interface

User Tools

- Local Performance Services
- Global Performance Services
- Java OWAMP Client
- Reverse Traceroute
- Reverse Ping
- Reverse Tracepath

Service Graphs

- Throughput
- One-Way Latency
- Traceroute
- Ping Latency
- SNMP Utilization
- Cacti Graphs

Toolkit Administration

- Administrative Information
- External BWCTL Limits
- External OWAMP Limits
- Enabled Services
- NTP
- Scheduled Tests
- Cacti SNMP Monitoring
- perfSONAR Logs

Performance Toolkit

pS-Performance Node For LBNL In Berkeley , CA , US

Host Information	
Organization Name	LBNL
City, State, Country	Berkeley, CA, US
Zip Code	94720
Latitude,Longitude	37.875985,-122.250014
Administrator Name	Brian Tierney
Administrator Email	bltierney@lbl.gov

Communities This Host Participates In

pS-NPToolkit-3.3

Host Status	
Primary Address	nettest.lbl.gov
MTU	1500
NTP Status	Synced
Globally registered	Yes

Services Offered

Bandwidth Test Controller (BWCTL)[1] — Running

- tcp://nettest.lbl.gov:4823

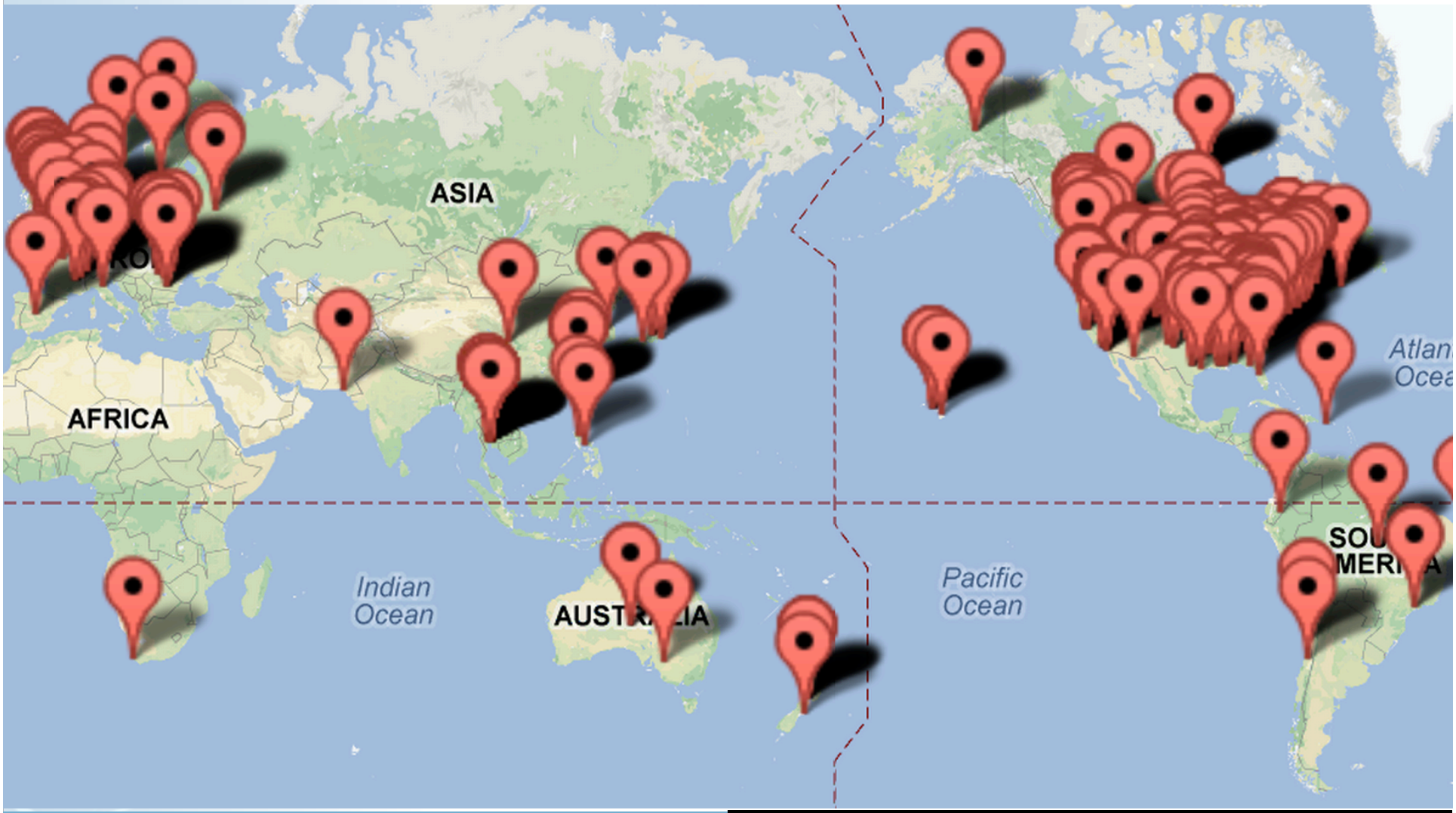
Network Diagnostic Tester (NDT)[1] — Running

- tcp://nettest.lbl.gov:3001
- http://nettest.lbl.gov:7123

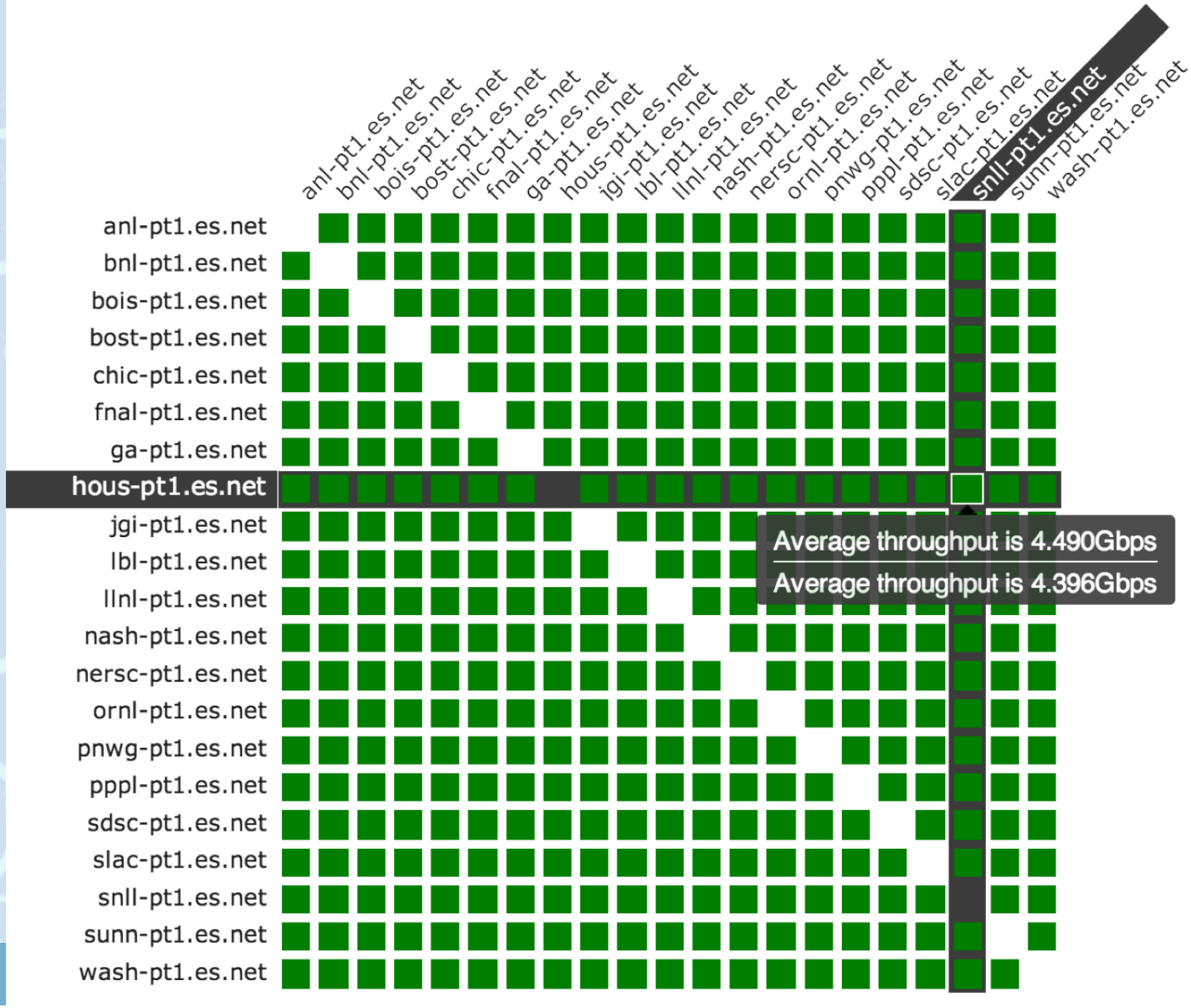
Network Path and Application Diagnosis (NPAD)[1] — Running

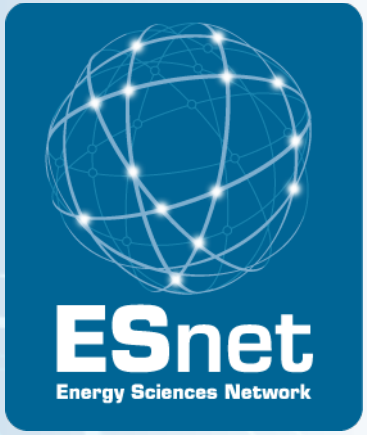
- tcp://nettest.lbl.gov:8001
- http://nettest.lbl.gov:8000

World-Wide perfSONAR-PS Deployments: 535 bwctl nodes, 533 owamp nodes as of June ‘13



perfSONAR Dashboard: http://ps-dashboard.es.net





Data Transfer Node and Data Transfer Tools

7/26/13

The Data Transfer Node

A DTN server is made of several subsystems. Each needs to perform optimally for the DTN workflow:

Storage: capacity, performance, reliability, physical footprint

Networking: protocol support, optimization, reliability

Motherboard: I/O paths, PCIe subsystem, IPMI

Chassis: adequate power supply, extra cooling

DTNs are usually optimized for sequential reads/write of large files, and a moderate number of high bandwidth flows.

The DTN is dedicated to data transfer, and not doing data analysis/ manipulation

DTN Tuning
http://fasterdata.es.net/science-dmz/DTN/tuning/

Defaults are usually not appropriate for high throughput.

What needs to be tuned:

- BIOS
- Firmware
- Device Drivers
- Networking
- File System
- Application

Network Tuning


```
# add to /etc/sysctl.conf
net.core.rmem_max = 33554432
net.core.wmem_max = 33554432
net.ipv4.tcp_rmem = 4096 87380 33554432
net.ipv4.tcp_wmem = 4096 65536 33554432
net.core.netdev_max_backlog = 250000

Add to /etc/rc.local
# increase txqueuelen
/sbin/ifconfig eth2 txqueuelen 10000
/sbin/ifconfig eth3 txqueuelen 10000
```

```
# make sure cubic and/or htcp are loaded
/sbin/modprobe tcp_htcp
/sbin/modprobe tcp_cubic
# set default to CC alg to htcp
net.ipv4.tcp_congestion_control=htcp

# with some 10G NICs increasing interrupt
    coalescing helps a lot:
/usr/sbin/ethtool -C ethN rx-usecs 75

And use Jumbo Frames!
```

Using the right tool is Critical!

Sample Results: 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	6.6 Gbps	(disk limited)

– Note that to get more than 1 Gbps (125 MB/s) disk to disk requires RAID.

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
- 3rd party transfer support

Globus Online has all these features

Commercial Data Transfer Tools

There are several commercial UDP-based tools

- Aspera: http://www.asperasoft.com/
  - RDSI project is using Aspera
- Data Expedition: http://www.dataexpedition.com/
- TIXstream: http://www.tixeltec.com/tixstream_en.html

These should all do better than TCP on a lossy path

- UDP is less sensitive to soft failures
- Advantage of these tools less clear on an clean path

They all have different, fairly complicated pricing models

Just how far can this scale?

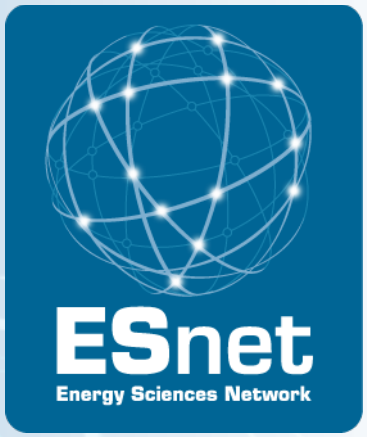
We recently did a demonstration of 2 DTN nodes in Chicago, USA transferring data to 2 DTN nodes in Maastricht, NL (RTT = 115 ms)

Using 4 10G NICS per host, and 2 TCP flows per NIC, we achieved a total of 78 Gbps memory to memory.

This required considerable DTN tuning, and a loss-free path.

What this demonstration shows:

- It is possible to create a loss-free network over very long distances
- TCP works fine under the right conditions.
- 40Gbps per DTN is very doable.



Summary

Science DMZ Summary

Consists of **three key components**, all required:

"Friction free" network path

- Highly capable network devices (wire-speed, deep queues)
- Virtual circuit connectivity option
- Security policy and enforcement specific to science workflows
- Located at or near site perimeter if possible

Dedicated, high-performance Data Transfer Nodes (DTNs)

- Hardware, operating system, libraries all optimized for transfer
- Includes optimized data transfer tools such as Globus Online and GridFTP

Performance measurement/test node

- perfSONAR

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

Science DMZ Community

The Science DMZ community is growing as well. We would encourage everyone to join the conversation as you implement your networks:

- Mailing List
  - https://gab.es.net/mailman/listinfo/sciencedmz

- Forums:
  - http://fasterdata.es.net/forums/

Questions?

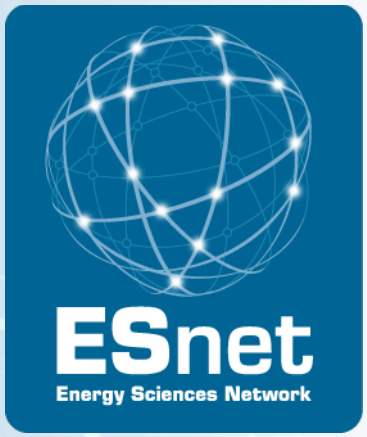
Email: BLTierney@es.net, engage@es.net

http://fasterdata.es.net: ESnet “knowledge base” of tips and tricks for obtaining maximum WAN throughput

Learn more about Science DMZs at:

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

- Includes a full day tutorial slides and video



Extra Slides

fasterdata.es.net

ESnet maintains a “knowledge base” of tips and tricks for obtaining maximum WAN throughput

Lots of useful stuff there, including:

- TCP tuning information (in cut and paste friendly form)
- Data Transfer Node (DTN) tuning information
  - Also in cut and paste friendly form
- DTN reference designs
- Science DMZ information
- perfSONAR information

Energy Sciences Network Overview

A national network, optimized for science:

- connecting 40 labs, facilities with >100 networks
- optimized for massive science data flows
- offering capabilities not available commercially
- $34.5M in FY12 (40 staff)

$62M ARRA stimulus grant funding:

- optical fiber assets
- access to spectrum
- world's first 100G network at scale
- easier (and cheaper) scaling

On the web:

- www.es.net
- fasterdata.es.net
- my.es.net

TCP Issues

It is far easier to architect the network to support TCP than it is to fix TCP

- People have been trying to fix TCP for years – limited success
- Packet loss is still the number one performance killer in long distance high performance environments

Pragmatically speaking, we must accommodate TCP

- Implications for equipment selection
  - Equipment must be able to accurately account for packets
- Implications for network architecture, deployment models
  - Infrastructure must be designed to allow easy troubleshooting
  - Test and measurement tools are critical

How Do We 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

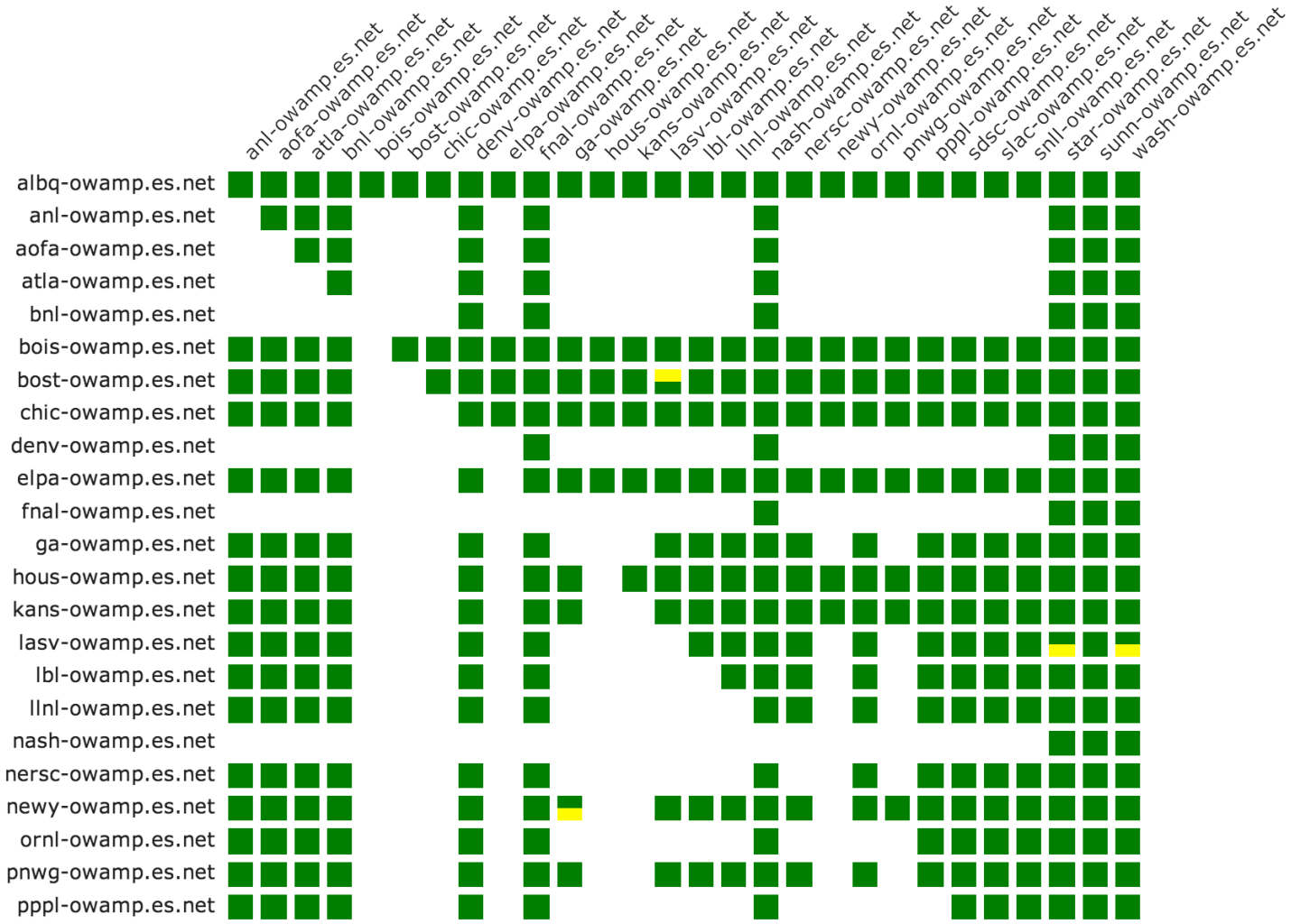
Equally important – the infrastructure must be verifiable so that clean service can be provided

- 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
  - Regular active test should be run - perfSONAR
- Small footprint is a huge win
  - Fewer the number of devices = easier to locate the source of packet loss

perfSONAR Dashboard: http://ps-dashboard.es.net

ESnet - ESnet to ESnet Packet Loss Testing

Loss rate is <= 0.001 | Loss rate is >= 0.001 | Loss rate is >= 0.1 | Unable to retrieve data | Check has not yet run



Importance of Regular Testing

You can't wait for users to report problems and then fix them (soft failures can go unreported for many months!)

Things just break sometimes

- Failing optics
- Somebody messed around in a patch panel and kinked a fiber
- Hardware goes bad

Problems that get fixed have a way of coming back

- System defaults come back after hardware/software upgrades
- New employees may not know why the previous employee set things up a certain way and back out fixes

Important to continually collect, archive, and alert on active throughput test results

Science DMZ Benefits

Better access to remote facilities by local users

Local facilities provide better service to remote users

Ability to support science that might otherwise be impossible

Metcalf's Law – value increases as the square of connected devices

- Communication between institutions with functional Science DMZs is greatly facilitated
- Increased ability to collaborate in a data-intensive world

Cost/Effort benefits also

- Shorter time to fix performance problems – less staff effort
- Appropriate implementation of security policy – lower risk
- No need to drag high-speed flows across business network → lower IT infrastructure costs

If Not Firewalls, Then What?

Remember – the goal is to protect systems in a way that allows the science mission to succeed

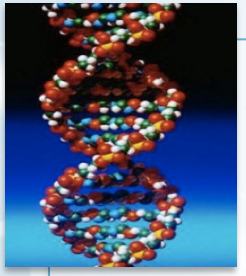
There are multiple ways to solve this – some are technical, and some are organizational/sociological

Note: this is harder than just putting up a firewall and thinking you are done

You need some combination of:

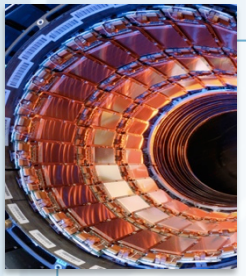
- Aggressive Router ACLs
- Network and Host IDS
- Blackhole routing
- As few services on the DTNs as possible

Data Explosion is Occurring Everywhere



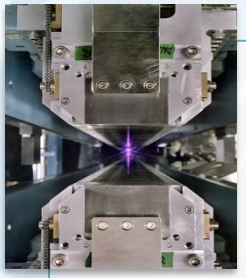
Genomics

- 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