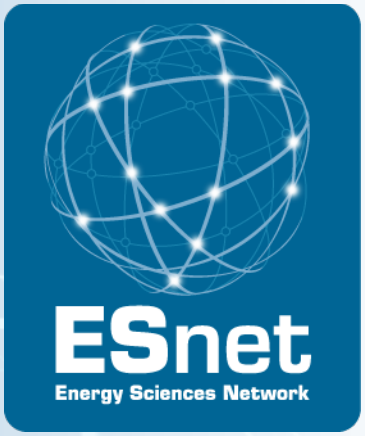


A Brief Overview of the Science DMZ

Jason Zurawski - **ESnet Engineering & Outreach**

engage@es.net

eduPERT Monthly Call

January 27th 2014

perfSONAR powered

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

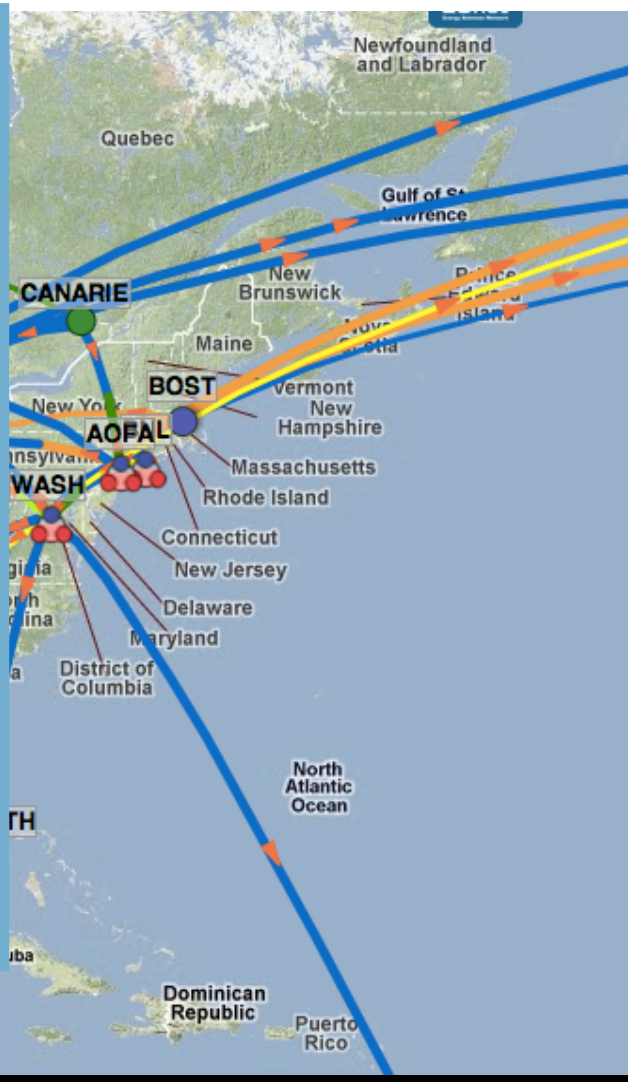
With contributions from S. Balasubramanian, G. Bell, E. Dart, M. Hester, B. Johnston, A. Lake, E. Pouyoul, L. Rotman, B. Tierney and others @ ESnet

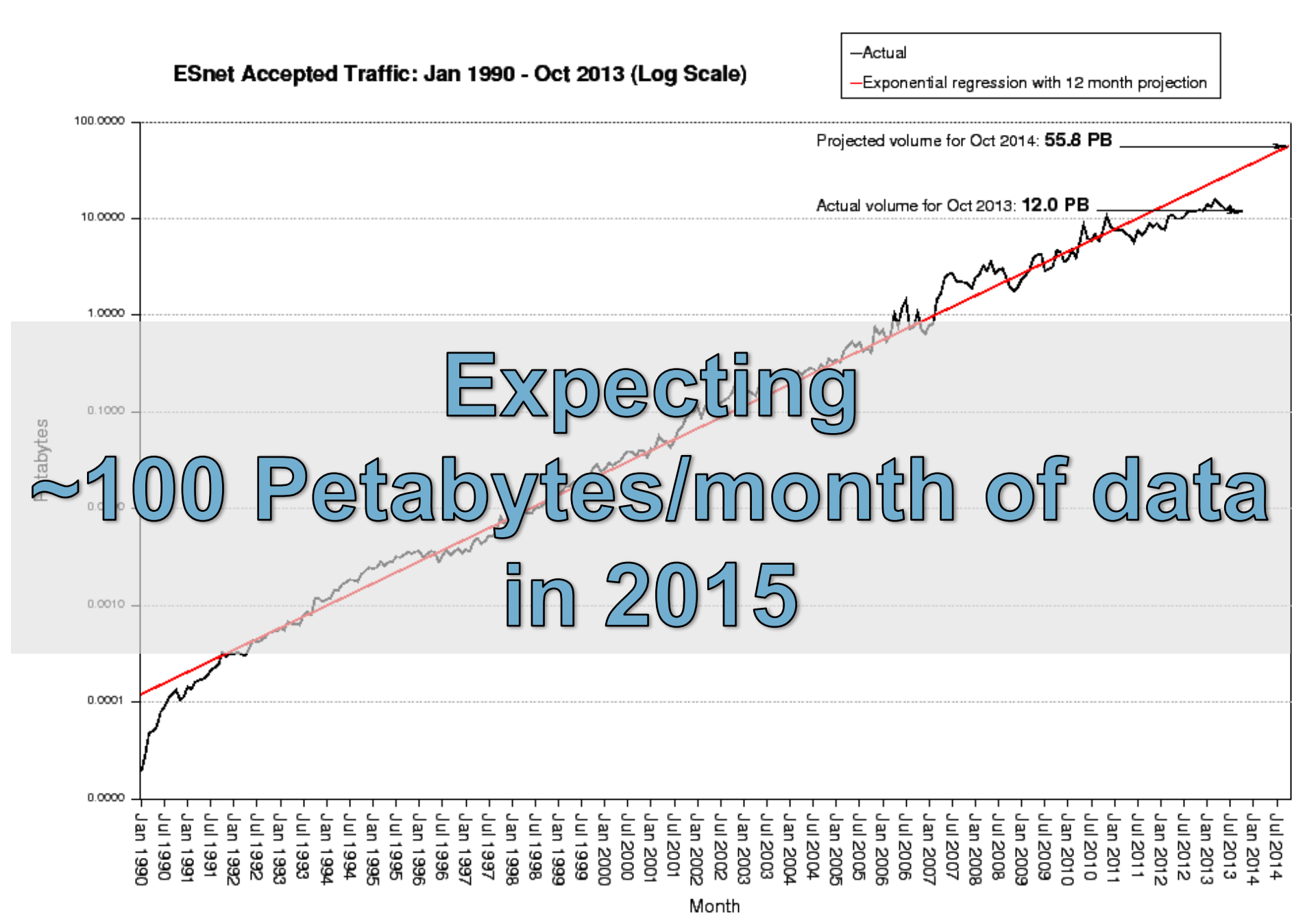
Overview

- What is ESnet?
- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- On the Topic of Security
- Wrap Up

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





Expecting ~100 Petabytes/month of data in 2015

Overview

- What is ESnet?
- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- On the Topic of Security
- Wrap Up

What is there to worry about?

- ***Genomics***
  - Sequencer data volume increase = 12x in 3 years
  - Sequencer cost decrease = 10x in 3 years
- ***High Energy Physics***
  - LHC experimental data = 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 changing operational model
- ***Common Threads***
  - Increased capability, greater need for data mobility due to span/depth of collaboration space
  - Global is the new local. Research is no longer done within a domain. End to end involves many fiefdoms to cross – and yes this becomes **your** problem when **your** users are impacted
  - The "***Campus Cyberinfrastructure - Network Infrastructure and Engineering (CC-NIE)***" program:

NSF 13-530: http://www.nsf.gov/pubs/2013/nsf13530/nsf13530.htm

perfSONAR powered

Motivation

Networks are an essential part of data-intensive science

- Connect data sources to data analysis
- Connect collaborators to each other
- Enable machine-consumable interfaces to data and analysis resources (e.g. portals), automation, scale

Performance is critical

- Exponential data growth
- Constant human factors
- Data movement and data analysis must keep up

Effective use of wide area (long-haul) networks by scientists has historically been difficult

perfSONAR powered

The Central Role of the Network

The very structure of modern science assumes science networks exist: high performance, feature rich, global scope

What is “The Network” anyway?

- “The Network” is the set of devices and applications involved in the use of a remote resource
  - This is not about supercomputer interconnects
  - This is about data flow from experiment to analysis, between facilities, etc.
- User interfaces for “The Network” – portal, data transfer tool, workflow engine
- Therefore, servers and applications must also be considered

What is important?

1. Correctness
2. Consistency
3. Performance

perfSONAR powered

TCP – Ubiquitous and Fragile

Networks provide connectivity between hosts – how do hosts see the network?

- From an application's perspective, the interface to "the other end" is a socket
- Communication is between applications – mostly over TCP

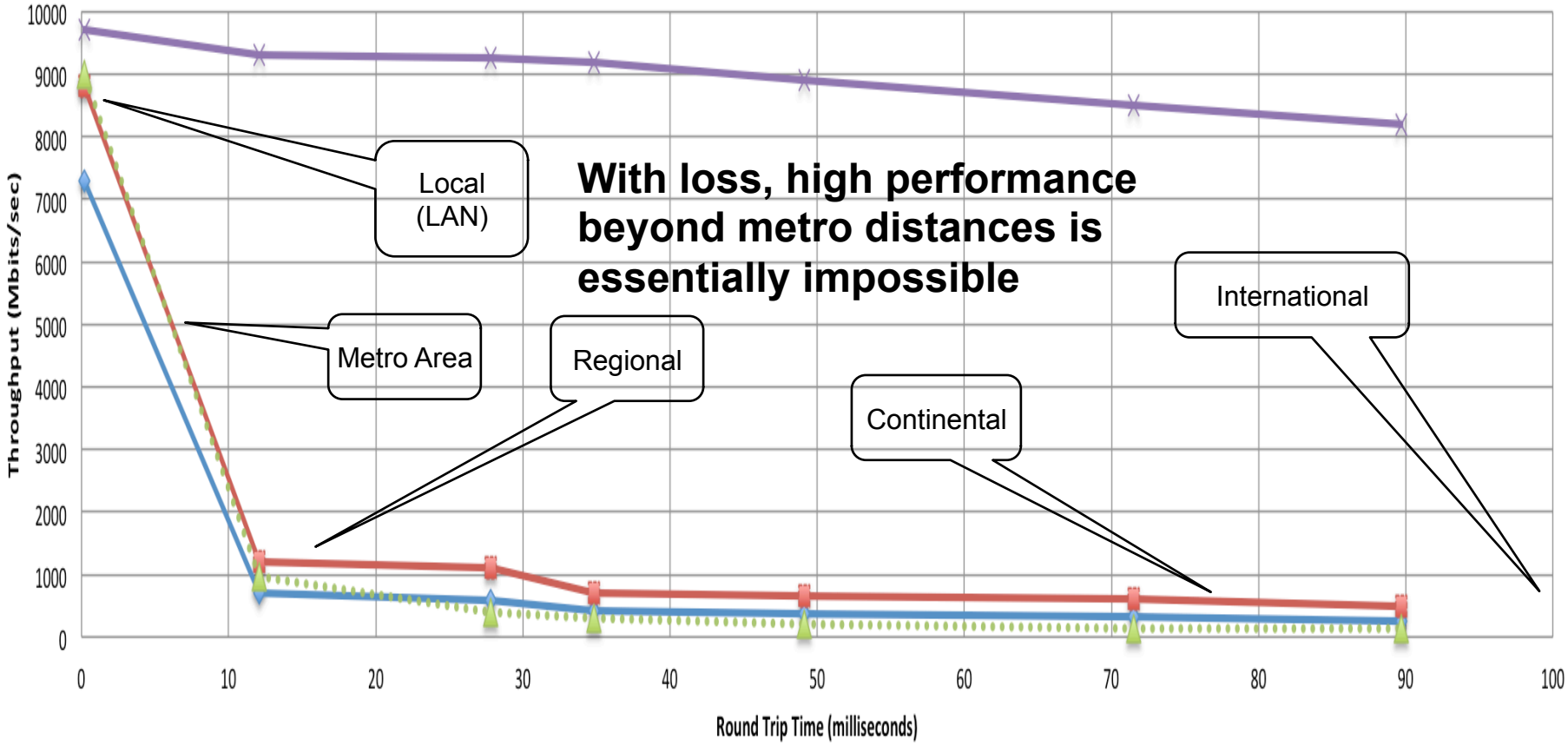
TCP – the fragile workhorse

- TCP is (for very good reasons) timid – packet loss is interpreted as congestion
- Packet loss in conjunction with latency is a performance killer
- Like it or not, TCP is used for the vast majority of data transfer applications (more than 95% of ESnet traffic is TCP)

perfSONAR powered

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

Throughput vs. increasing latency on a 10Gb/s link with *0.0046%* packet loss



Bandwidth Requirements to move Y Bytes of data in Time X

	Bits per Second Requirements				
10PB	25,020.0 Gbps	3,127.5 Gbps	1,042.5 Gbps	148.9 Gbps	34.7 Gbps
1PB	2,502.0 Gbps	312.7 Gbps	104.2 Gbps	14.9 Gbps	3.5 Gbps
100TB	244.3 Gbps	30.5 Gbps	10.2 Gbps	1.5 Gbps	339.4 Mbps
10TB	24.4 Gbps	3.1 Gbps	1.0 Gbps	145.4 Mbps	33.9 Mbps
1TB	2.4 Gbps	305.4 Mbps	101.8 Mbps	14.5 Mbps	3.4 Mbps
100GB	238.6 Mbps	29.8 Mbps	9.9 Mbps	1.4 Mbps	331.4 Kbps
10GB	23.9 Mbps	3.0 Mbps	994.2 Kbps	142.0 Kbps	33.1 Kbps
1GB	2.4 Mbps	298.3 Kbps	99.4 Kbps	14.2 Kbps	3.3 Kbps
100MB	233.0 Kbps	29.1 Kbps	9.7 Kbps	1.4 Kbps	0.3 Kbps
	1H	**8H**	**24H**	**7Days**	**30Days**

This table available at http://fasterdata.es.net

perfSONAR powered

Working With TCP In Practice

Far easier to support TCP than to fix TCP

- People have been trying to fix TCP for years – limited success
- Like it or not we're stuck with TCP in the general case

Pragmatically speaking, we must accommodate TCP

- Sufficient bandwidth to avoid congestion
- Zero packet loss
- Verifiable infrastructure
  - Networks are complex
  - Must be able to locate problems quickly
  - Small footprint is a huge win – small number of devices so that problem isolation is tractable

© Dog Shaming 2012

perfSONAR powered

Putting A Solution Together

Effective support for TCP-based data transfer

- Design for correct, consistent, high-performance operation
- Design for ease of troubleshooting

Easy adoption is critical

- Large laboratories and universities have extensive IT deployments
- Drastic change is prohibitively difficult

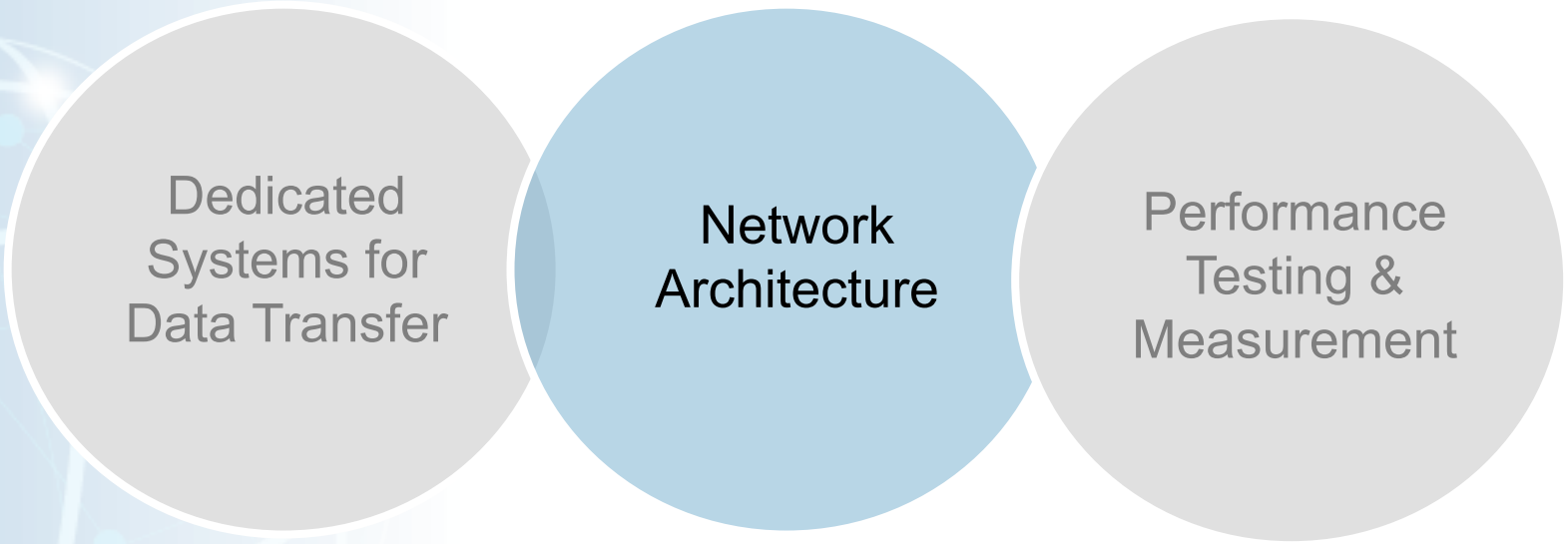
Cybersecurity – defensible without compromising performance

Borrow ideas from traditional network security

- Traditional DMZ – separate enclave at network perimeter ("Demilitarized Zone")
  - Specific location for external-facing services
  - Clean separation from internal network
- Do the same thing for science – ***Science DMZ***

perfSONAR powered

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
- Proper security
- 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

perfSONAR powered

Ad Hoc DTN Deployment

This is often what gets tried first

Data transfer node deployed where the owner has space

- This is often the easiest thing to do at the time
- Straightforward to turn on, hard to achieve performance

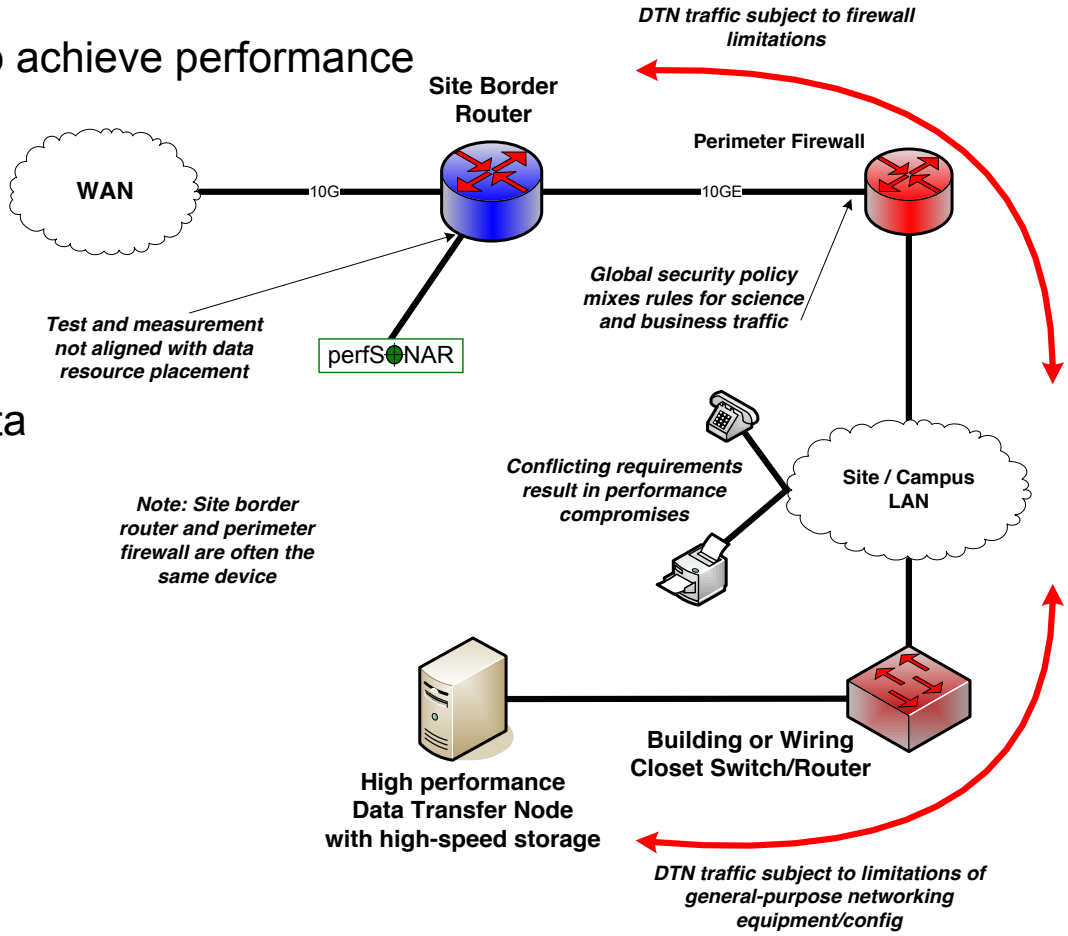
If present, perfSONAR is at the border

- This is a good start
- Need a second one next to the DTN

Entire LAN path has to be sized for data flows

Entire LAN path is part of any troubleshooting exercise

This usually fails to provide the necessary performance.



Small-scale Science DMZ Deployment

Add-on to existing network infrastructure

- All that is required is a port on the border router
- Small footprint, pre-production commitment

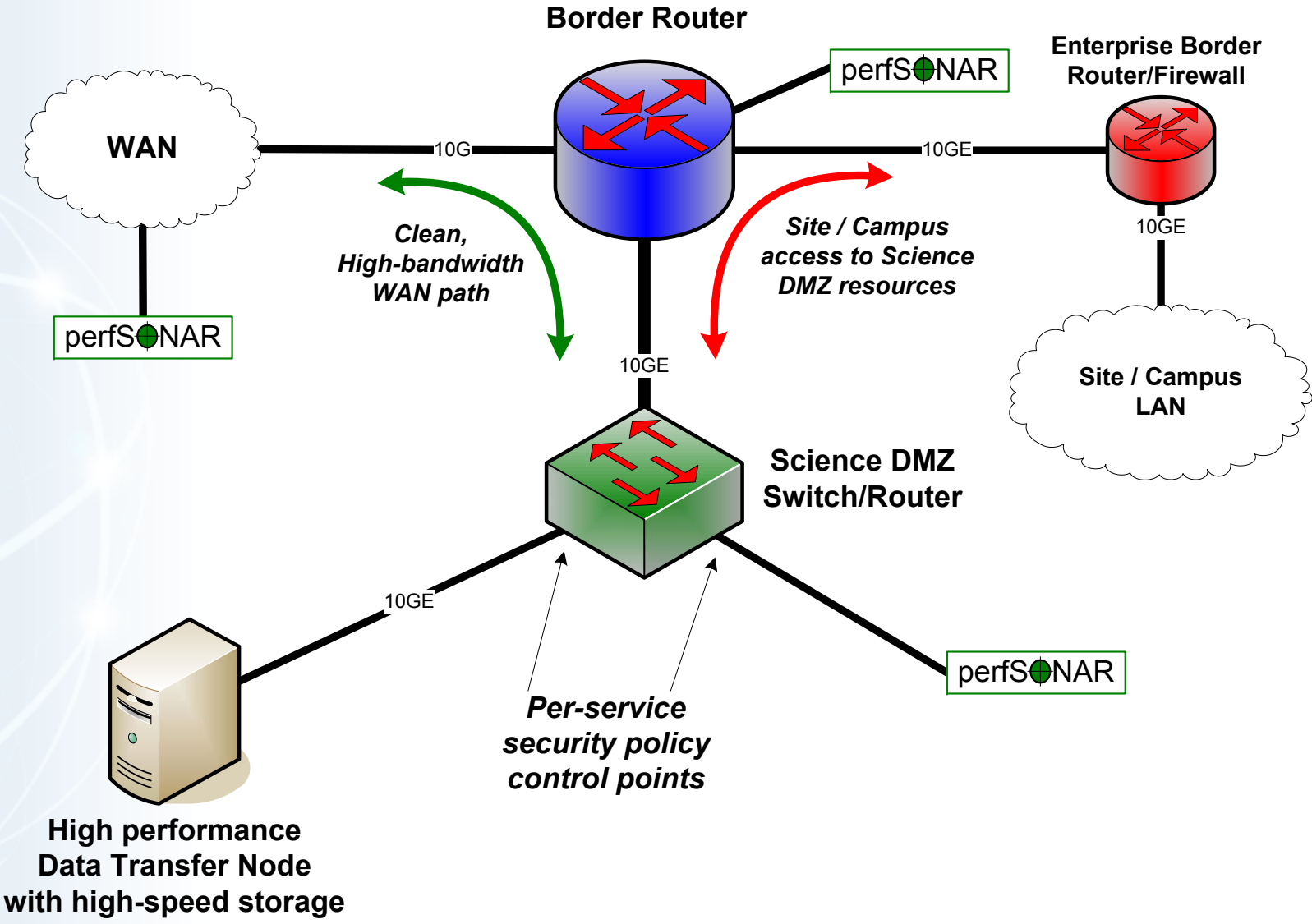
Easy to experiment with components and technologies

- DTN prototyping
- perfSONAR testing

Limited scope makes security policy exceptions easy

- Only allow traffic from partners
- Add-on to production infrastructure – lower risk

A better approach: simple Science DMZ



Science DMZ – Flexible Design Pattern

The Science DMZ design pattern is highly adaptable to research

Deploying a research Science DMZ is straightforward

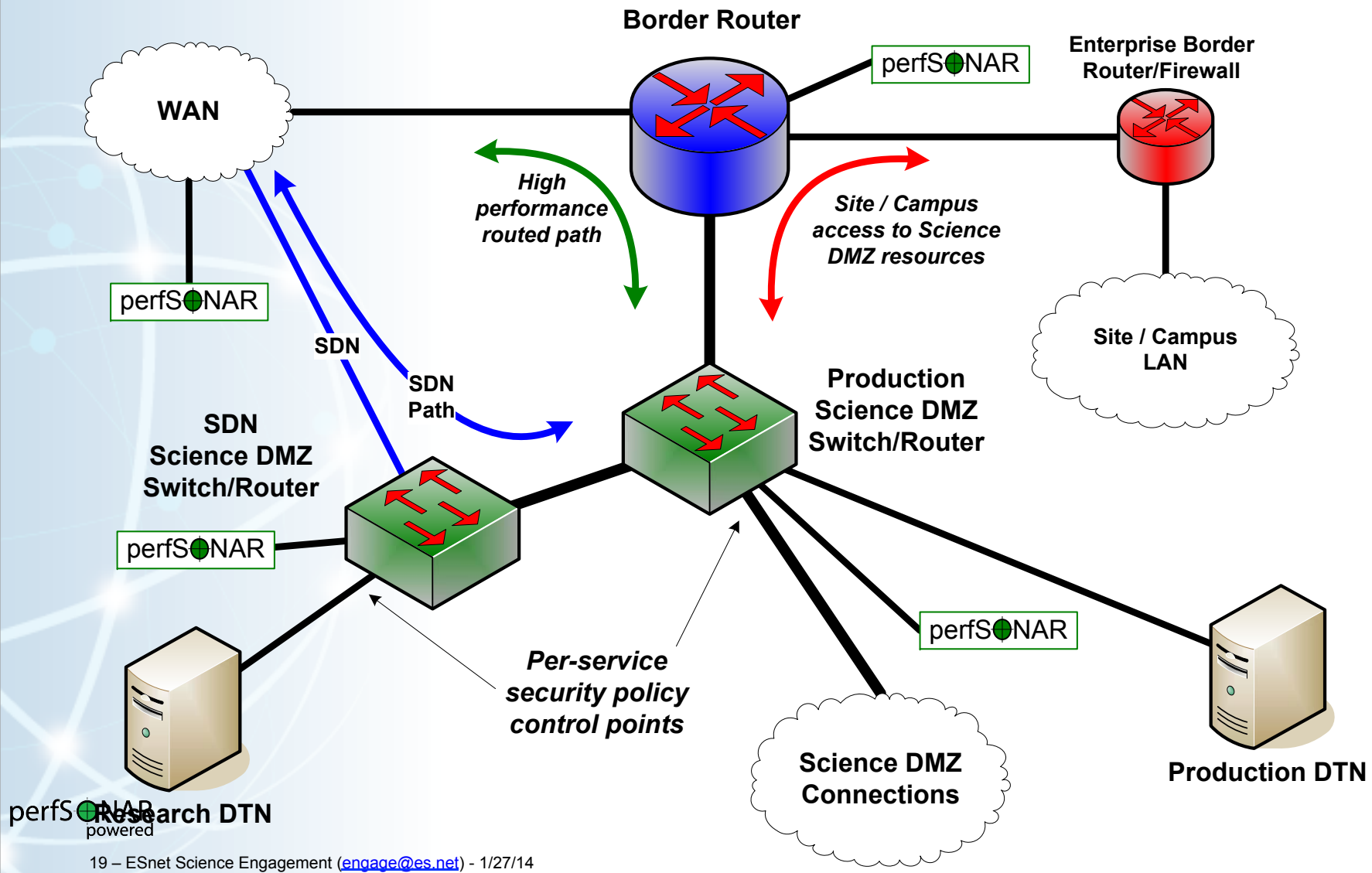
- The basic elements are the same
  - Capable infrastructure designed for the task
  - Test and measurement to verify correct operation
  - Security policy well-matched to the environment, application set is strictly limited to reduce risk
- Connect the research DMZ to other resources as appropriate

The same ideas apply to supporting an SDN effort

- Test/research areas for development
- Transition to production as technology matures and need dictates
- One possible trajectory follows…

perfSONAR powered

Science DMZ – Separate SDN Connection



Common Threads

Two common threads exist in all these examples

Accommodation of TCP

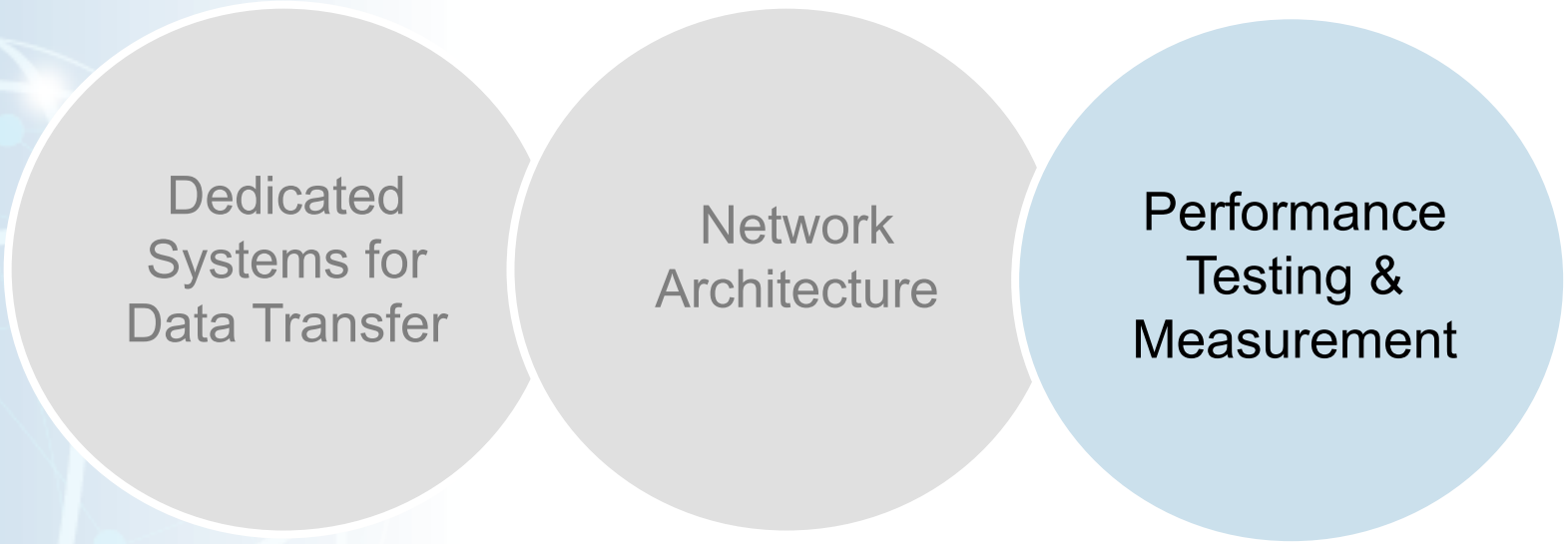
- Wide area portion of data transfers traverses purpose-built path
- High performance devices that don't drop packets

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

perfSONAR powered

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
- Proper security
- 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

perfSONAR powered

Performance Monitoring

Everything may function perfectly when it is deployed

Eventually something is going to break

- Networks and systems are complex
- Bugs, mistakes, …
- Sometimes things just break – this is why we buy support contracts

Must be able to find and fix problems when they occur

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)

perfSONAR powered

Soft Network Failures – Hidden Problems

"Soft failures" result in degraded capability

- Connectivity exists
- Performance impacted
- Typically something in the path is functioning, but not well

Hard failures are easy to detect

- Link down, system crash, software crash
- Traditional network/system monitoring tools designed to quickly find hard failures

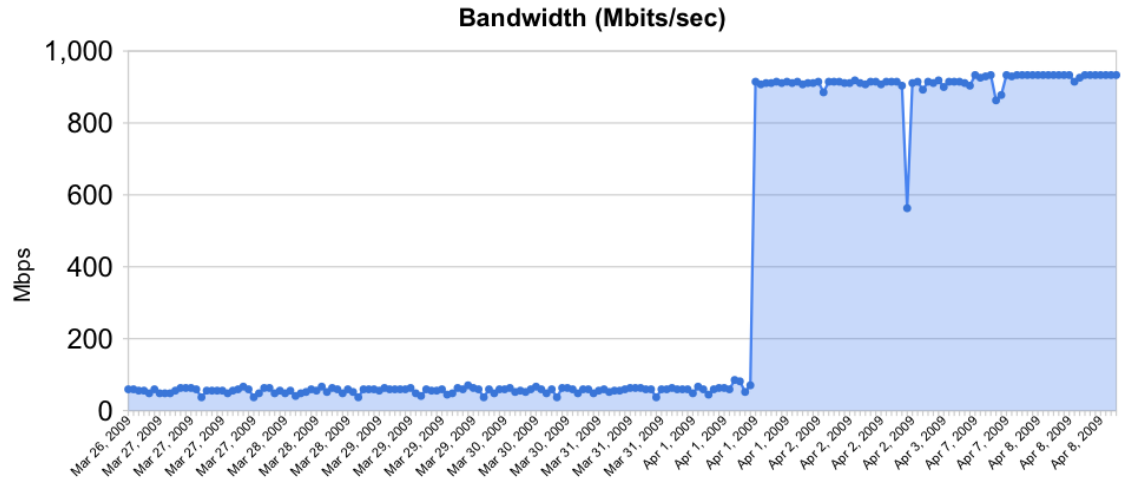
Soft failures are hard to detect with traditional methods

- No obvious single event
- Sometimes no indication at all of any errors

Independent testing is the only way to reliably find soft failures

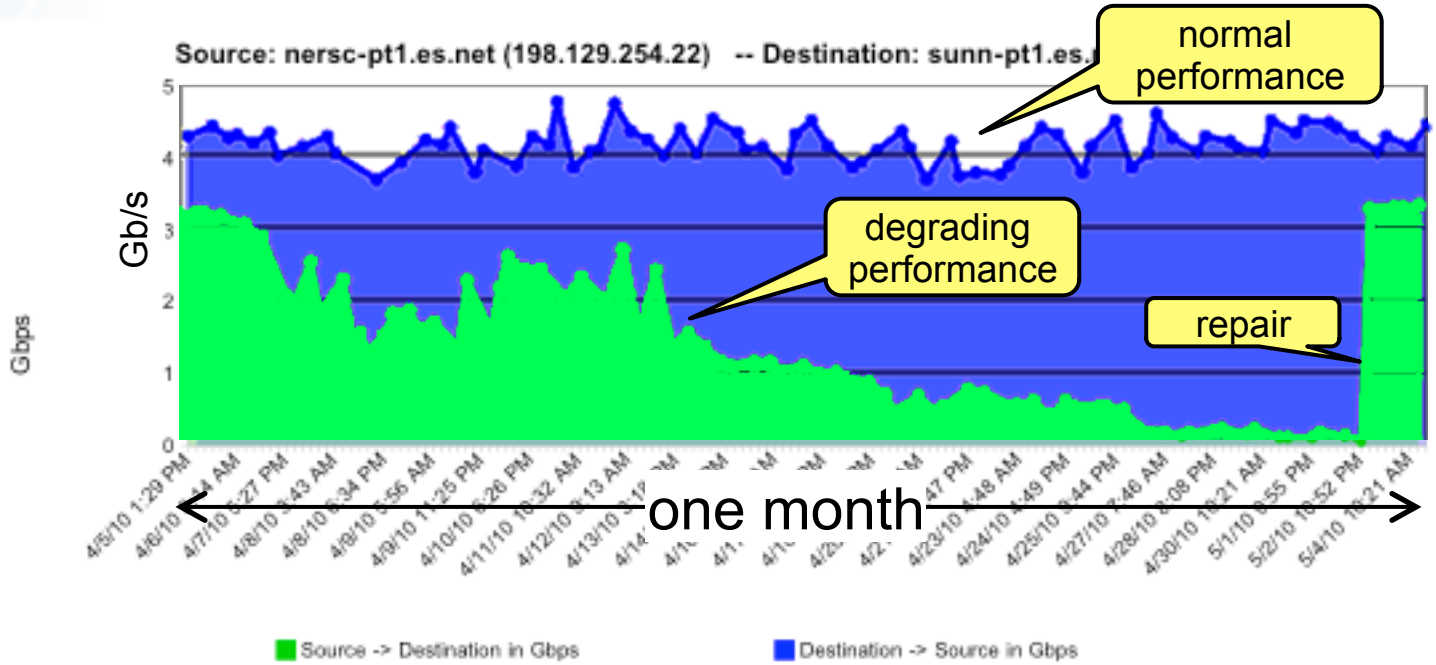
perfSONAR powered

Sample Soft Failures



Rebooted router
with full route table

Gradual failure
of optical line
card



Testing Infrastructure – perfSONAR

perfSONAR is:

- A widely-deployed test and measurement infrastructure
  - ESnet, Internet2, US regional networks, international networks
  - Laboratories, supercomputer centers, universities
- A suite of test and measurement tools
- A collaboration that builds and maintains the toolkit

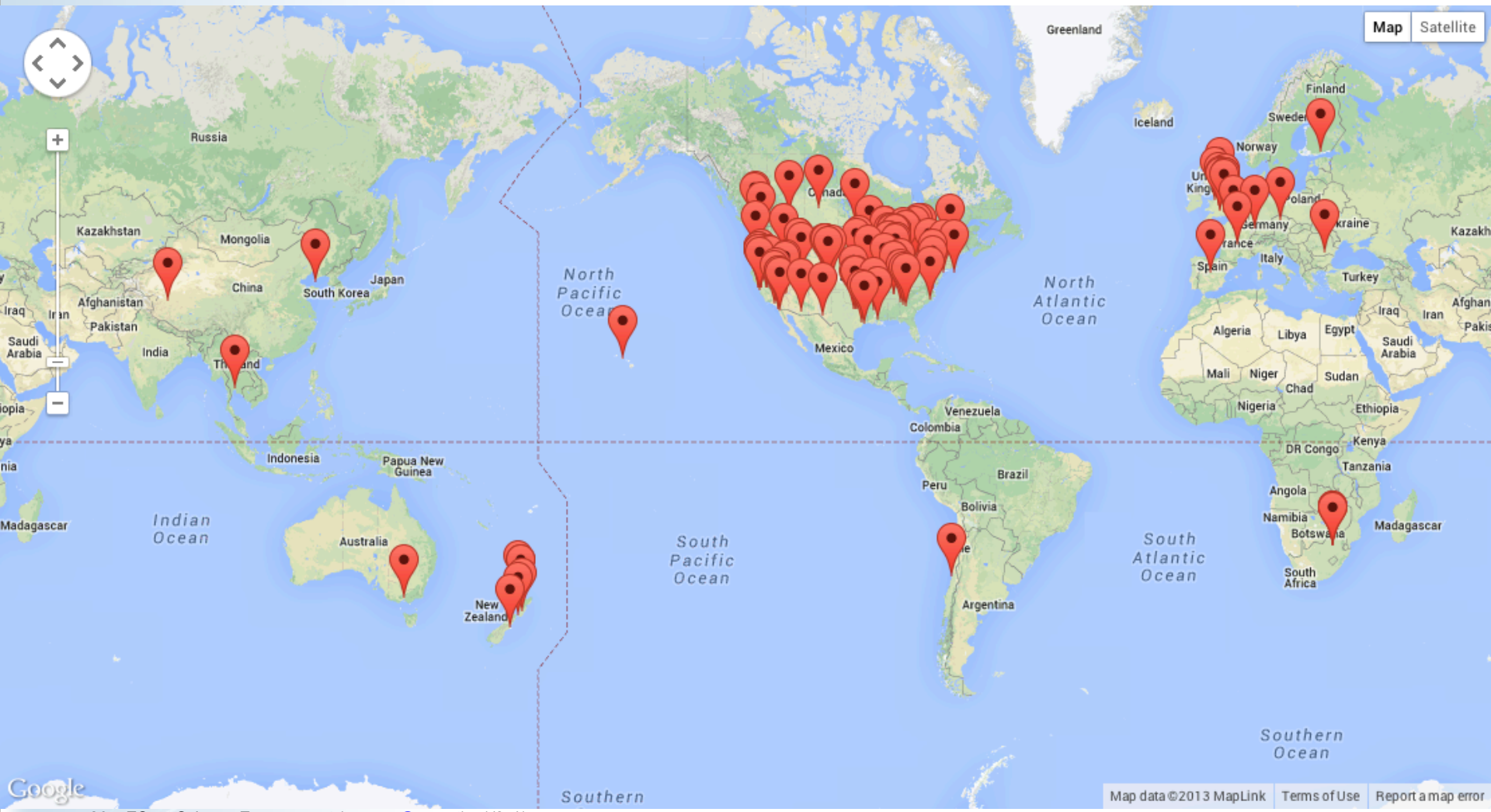
By installing perfSONAR, a site can leverage over 900 test servers deployed around the world

perfSONAR is ideal for finding soft failures

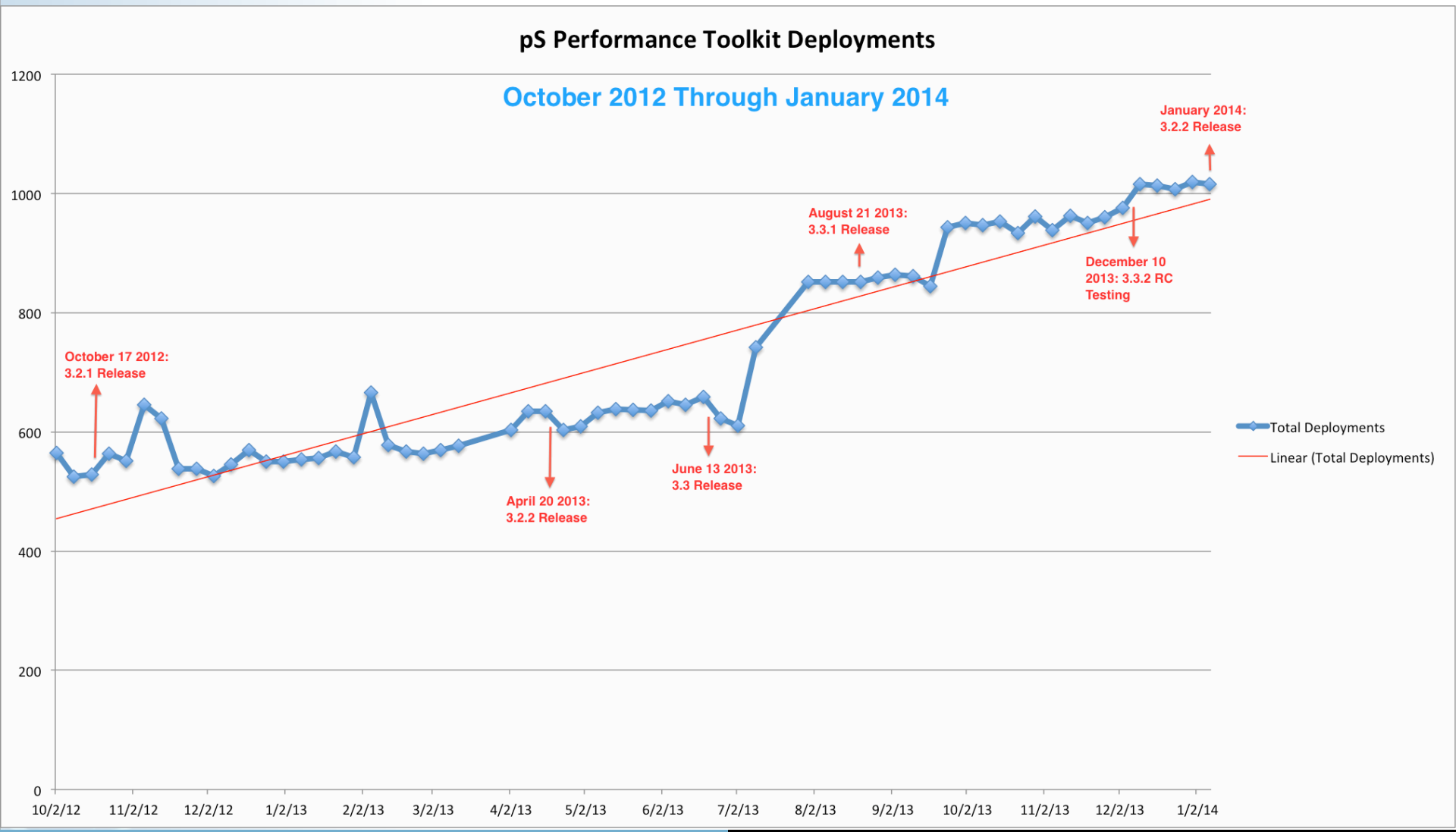
- Alert to existence of problems
- Fault isolation
- Verification of correct operation

perfSONAR powered

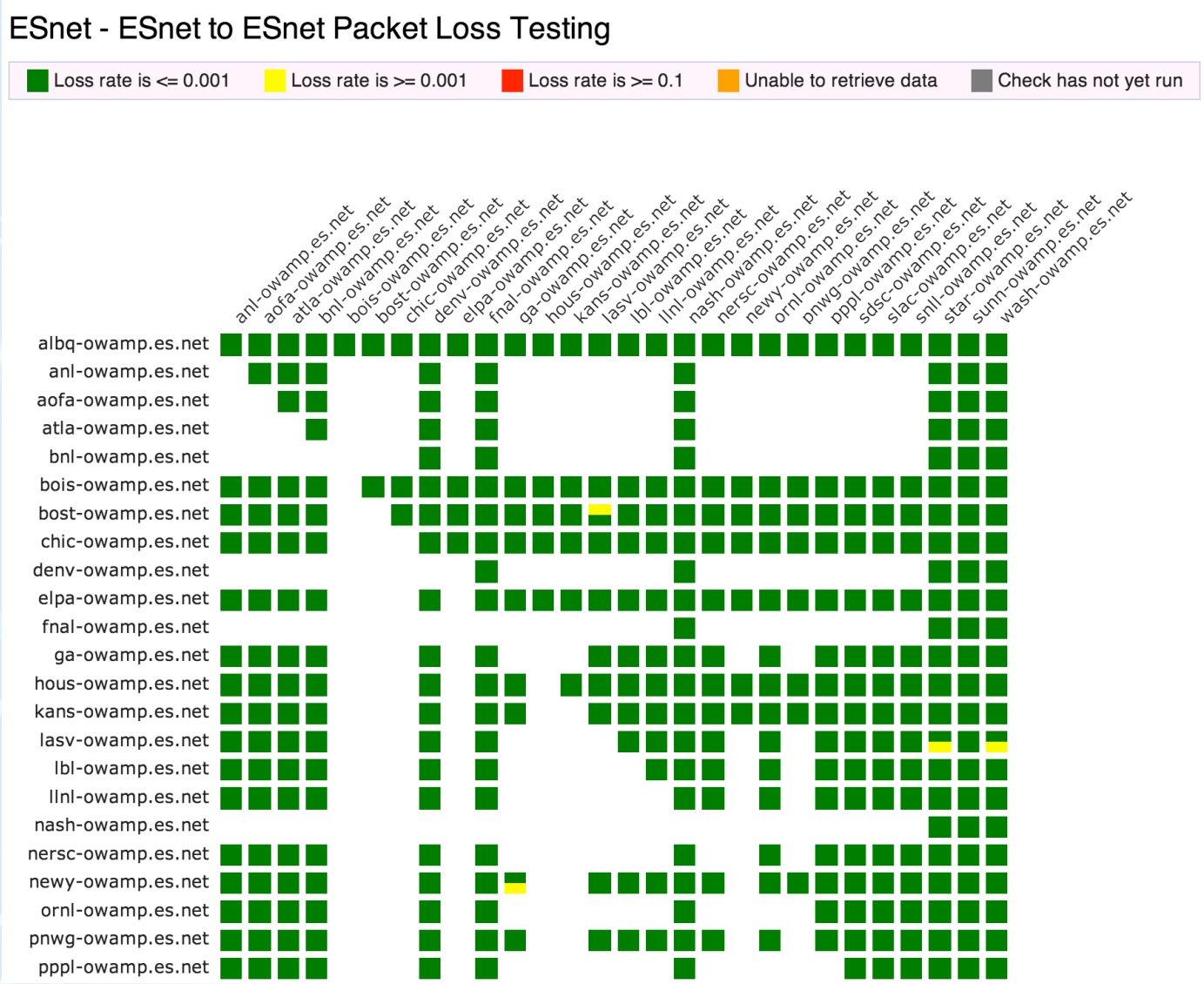
World-Wide perfSONAR-PS Deployments: 1000+ as of January 2014



World-Wide perfSONAR-PS Deployments: 1000+ as of January 2014

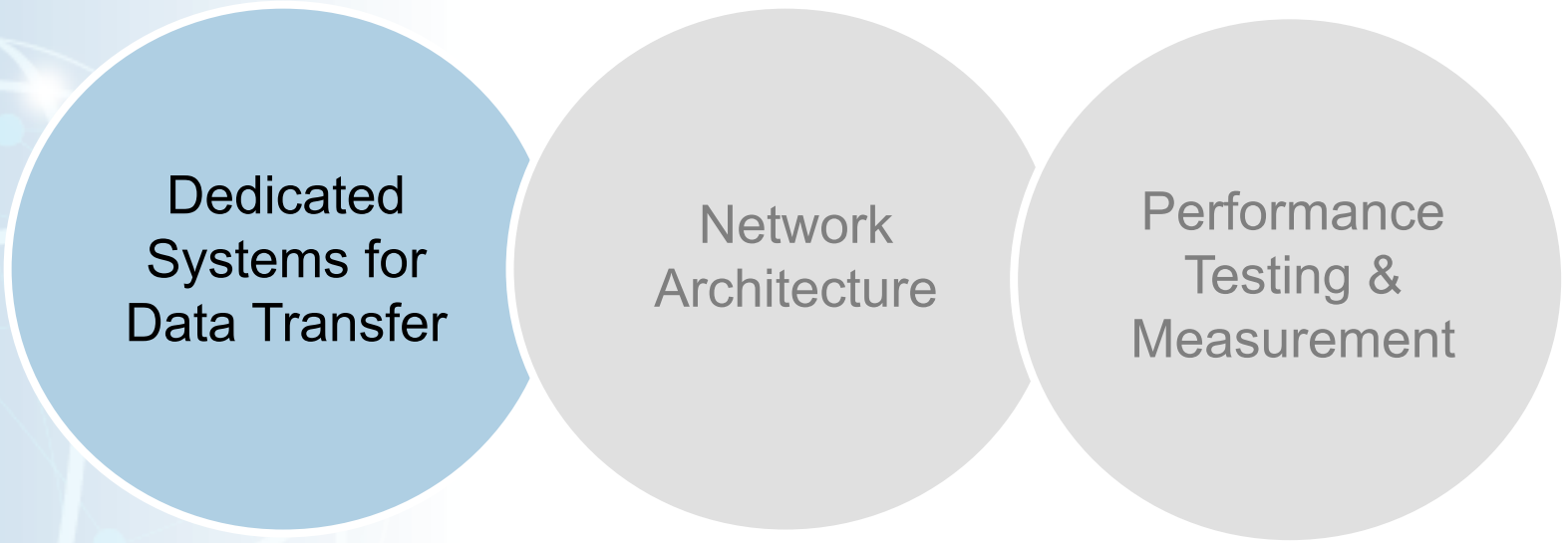


perfSONAR Dashboard: http://ps-dashboard.es.net



perfSONAR powered

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
- Proper security
- 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

perfSONAR powered

Dedicated Systems – The Data Transfer Node

The DTN is dedicated to data transfer

Set up specifically for high-performance data movement

- System internals (BIOS, firmware, interrupts, etc.)
- Network stack
- Storage (global filesystem, Fibrechannel, local RAID, etc.)
- High performance tools
- No extraneous software

Limitation of scope and function is actually powerful

- No conflicts with configuration for other tasks
- Small application set makes cybersecurity easier

Data Transfer Tools For DTNs

Parallelism is important

- It is often easier to achieve a given performance level with four parallel connections than one connection
- Several tools offer parallel transfers, including Globus/GridFTP

Latency interaction is critical

- Wide area data transfers have much higher latency than LAN transfers
- Many tools and protocols assume a LAN

Workflow integration is important

Key tools: Globus Online, HPN-SSH

perfSONAR powered

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).

perfSONAR powered

Overview

- What is ESnet?
- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- On the Topic of Security
- Wrap Up

perfSONAR powered

Science DMZ Security

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

Rationale

- Science data traffic is simple from a security perspective
- Narrow application set on Science DMZ
  - Data transfer, data streaming packages
  - No printers, document readers, web browsers, building control systems, financial databases, staff desktops, etc.
- Security controls that are typically implemented to protect business resources often cause performance problems

Separation allows each to be optimized

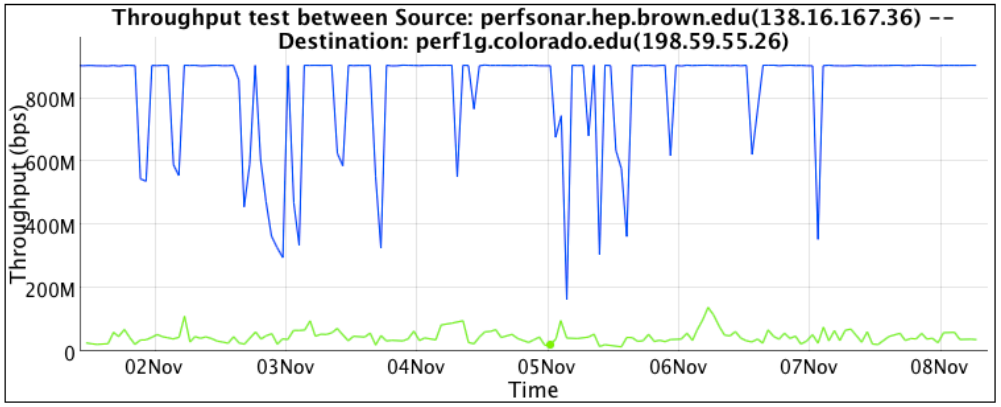
Not "how much" security, but how the security is implemented

perfSONAR powered

Firewall Performance Example

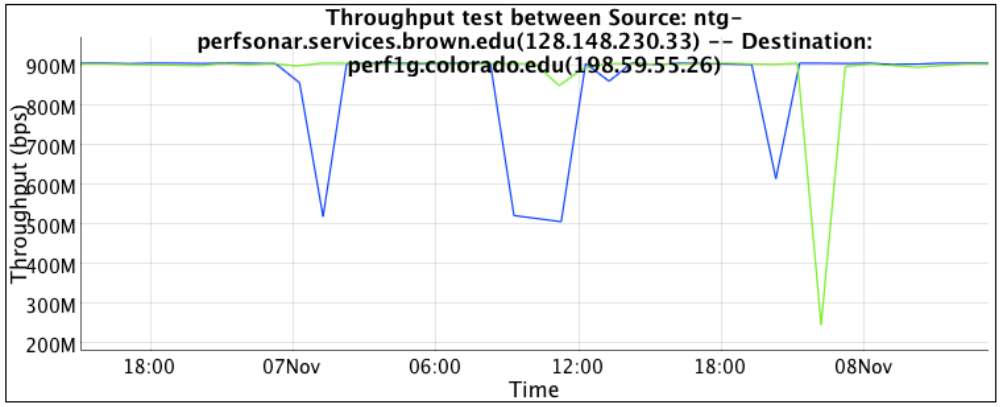
Observed performance, via perfSONAR, through a firewall:

Almost 20 times slower!



Observed performance, via perfSONAR, bypassing firewall:

Traffic was unimpeded by additional processing or resource constraints



perfSONAR powered

Placement Outside the Firewall

The Science DMZ resources are placed outside the enterprise firewall for performance reasons

- The meaning of this is specific – ***Science DMZ traffic does not traverse the firewall***
- Packet filtering is fine – just don't do it with a firewall. Consider ACLs (paths of science data are well known)

Lots of heartburn over this, especially from the perspective of a conventional firewall manager.

- Lots of policy directives mandating firewalls
- Firewalls are designed to protect converged enterprise networks
- Why would you put critical assets outside the firewall???

perfSONAR powered

Security Without Firewalls

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

- We **must** protect our systems
- Just do security without preventing science
- You can do packet filtering without using a firewall

Example – firewall rules for science traffic use address/port

- Instead implement filtering on Science DMZ router. Protect hosts and services on a direct basis (instead of blanket policy)
- Science wins – increased performance
- Business network wins – no need to size the firewall for science data deluge (10G firewalls are more expensive than 1G)

Key point – security policies and mechanisms that protect the Science DMZ should be implemented so that they do not compromise performance

perfSONAR powered

Overview

- What is ESnet?
- Science DMZ Motivation and Introduction
- Science DMZ Architecture
- Network Monitoring
- Data Transfer Nodes & Applications
- On the Topic of Security
- Wrap Up

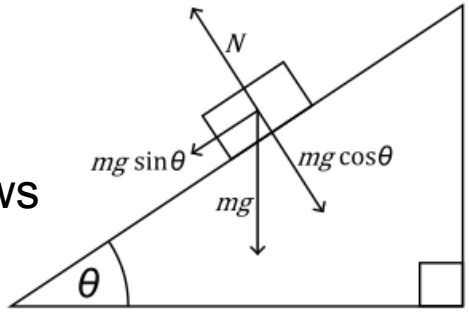
perfSONAR powered

The Science DMZ in 1 Slide

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

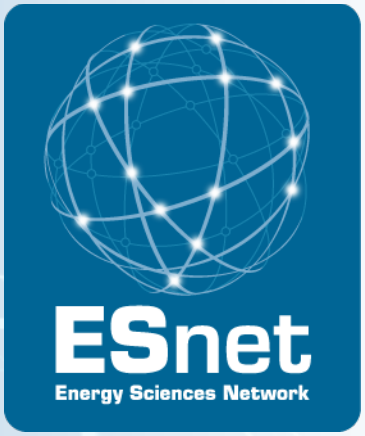
Performance measurement/test node

- perfSONAR

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

Links

- ESnet fasterdata knowledge base
  - http://fasterdata.es.net/
- Science DMZ paper from SC13
  - http://www.es.net/assets/pubs_presos/sc13sciDMZ-final.pdf
- Science DMZ email list
  - https://gab.es.net/mailman/listinfo/sciencedmz
- perfSONAR
  - http://fasterdata.es.net/performance-testing/perfsonar/
  - http://psps.perfsonar.net/
- Additional material
  - http://fasterdata.es.net/science-dmz/
  - http://fasterdata.es.net/host-tuning/



A Brief Overview of the Science DMZ

Questions/Comments/Criticisms?

Jason Zurawski - zurawski@es.net

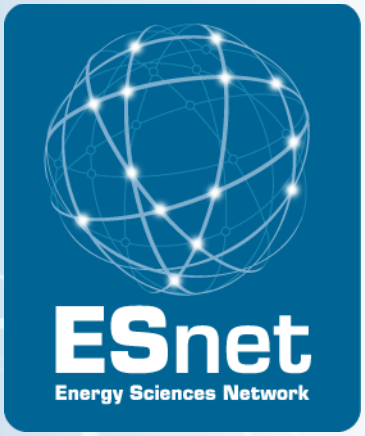
ESnet Science Engagement – engage@es.net

http://fasterdata.es.net

perfSONAR powered

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB



Extra Slides

Router Buffering

perfSONAR powered

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

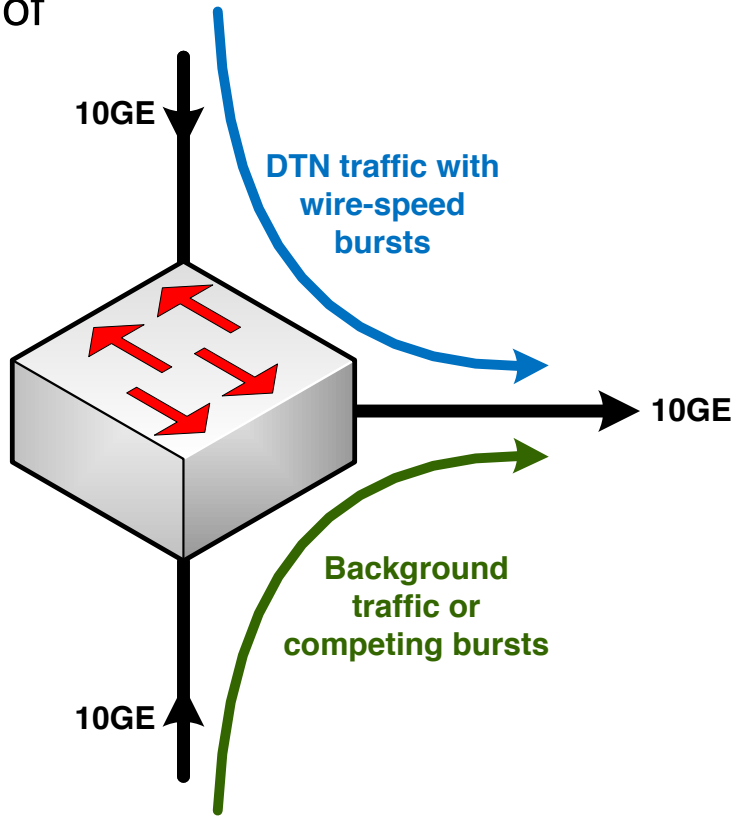
Multiple Ingress Data Flows, Common Egress

Hosts will typically send packets at the speed of their interface (1G, 10G, etc.)

- Instantaneous rate, not average rate
- If TCP has window available and data to send, host sends until there is either no data or no window

Hosts moving big data (e.g. DTNs) can send large bursts of back-to-back packets

- This is true even if the average rate as measured over seconds is slower (e.g. 4Gbps)
- On microsecond time scales, there is often congestion
- Router or switch must queue packets or drop them



perfSONAR powered

Router and Switch Output Queues

Interface output queue allows the router or switch to avoid causing packet loss in cases of momentary congestion

In network devices, queue depth (or 'buffer') is often a function of cost

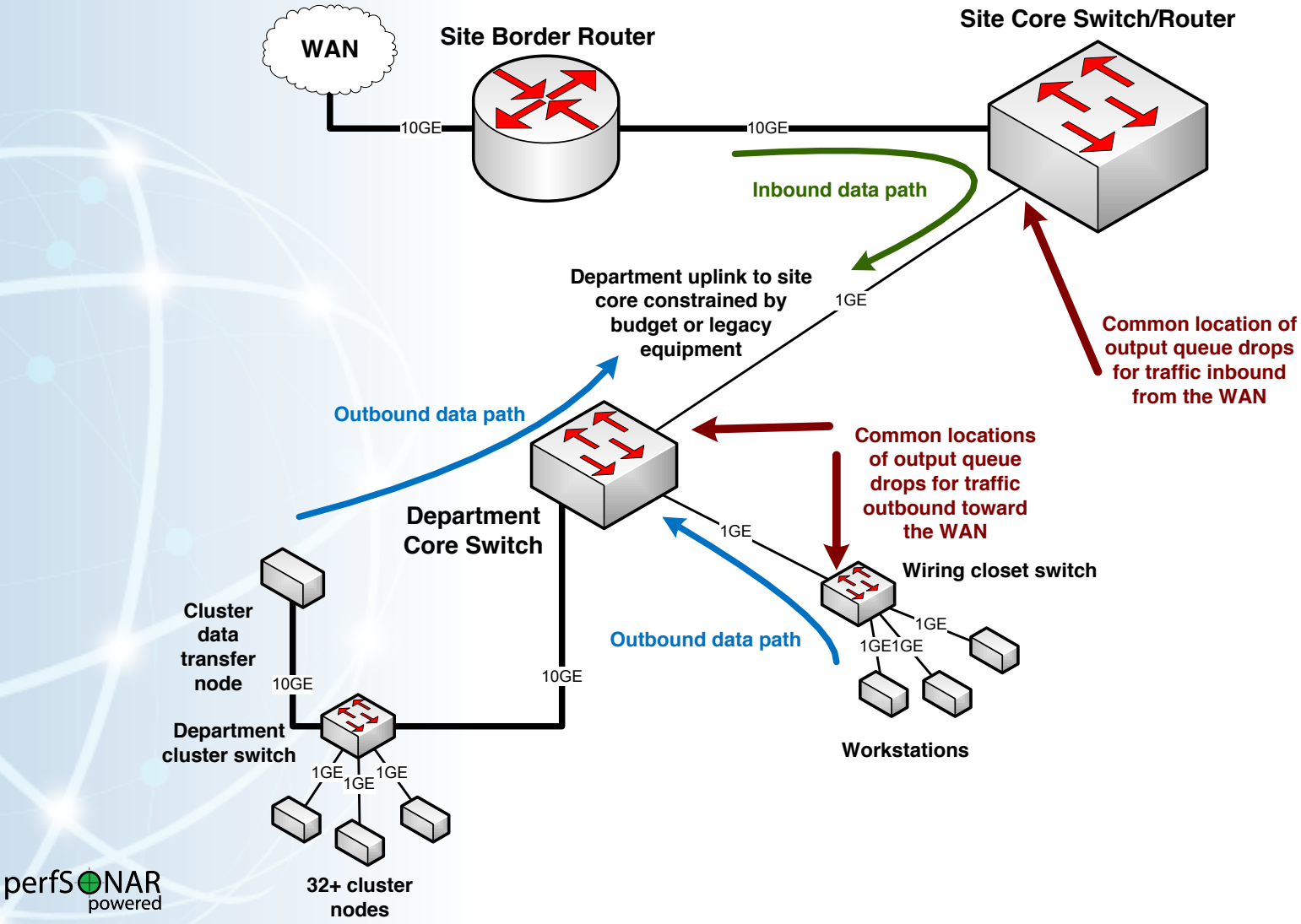
- Cheap, fixed-config LAN switches (especially in the 10G space) have inadequate buffering. Imagine a 10G 'data center' switch as the guilty party
- Cut-through or low-latency Ethernet switches typically have inadequate buffering (the whole point is to avoid queuing!)

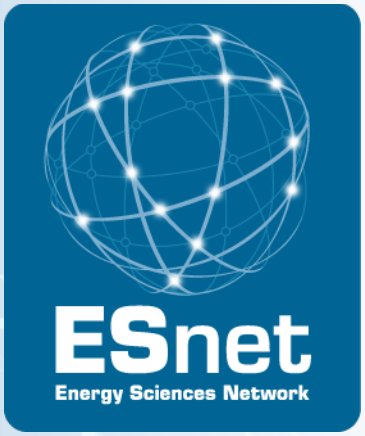
Expensive, chassis-based devices are more likely to have deep enough queues

- Juniper MX and Alcatel-Lucent 7750 used in ESnet backbone
- Other vendors make such devices as well - details are important
- Thx to Jim: http://people.ucsc.edu/~warner/buffer.html

This expense is one driver for the Science DMZ architecture – only deploy the expensive features where necessary

Output Queue Drops – Common Locations





Extra Slides

ALS Use Case/Workflow Design

perfSONAR powered

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

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

perfSONAR powered

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

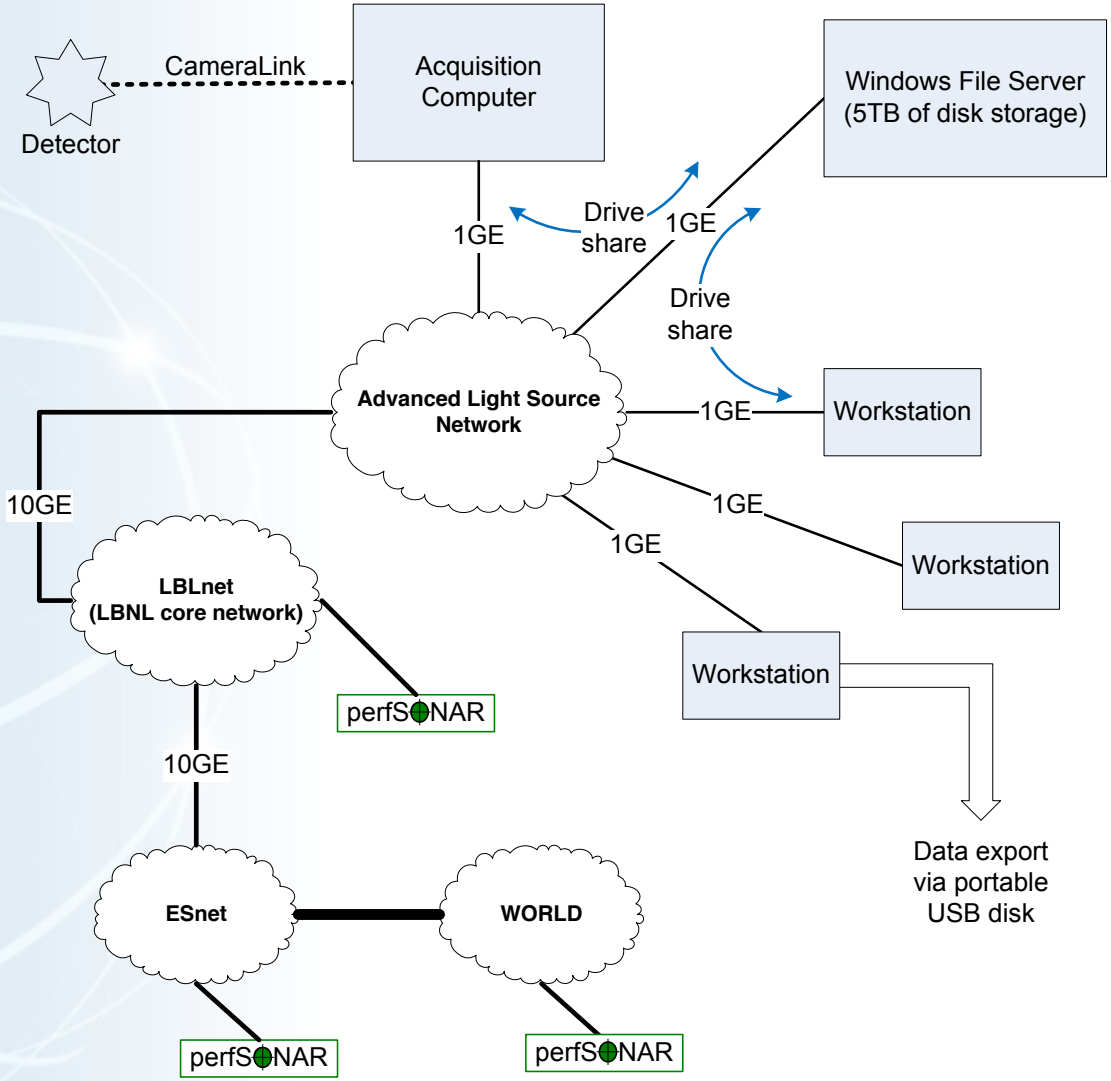
NERSC collaboration

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

Collaboration is ongoing

perfSONAR powered

Original Workflow Infrastructure



Improved Workflow Infrastructure