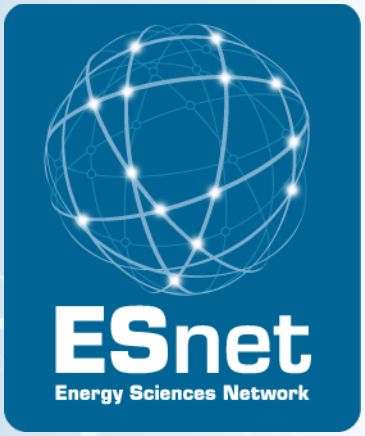


The Science DMZ

Brian Tierney, Eli Dart, Eric Pouyoul, Jason Zurawski

ESnet

Supporting Data-Intensive Research Workshop

QuestNet 2013

Gold Coast, Australia

July 2, 2013

What's there to worry about?

© Owen Humphreys/National Geographic Traveler Photo Contest 2013

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/

Overview

Part 1:

- **What is ESnet?**
- **Science DMZ Motivation**
- **Science DMZ Architecture**

Part 2:

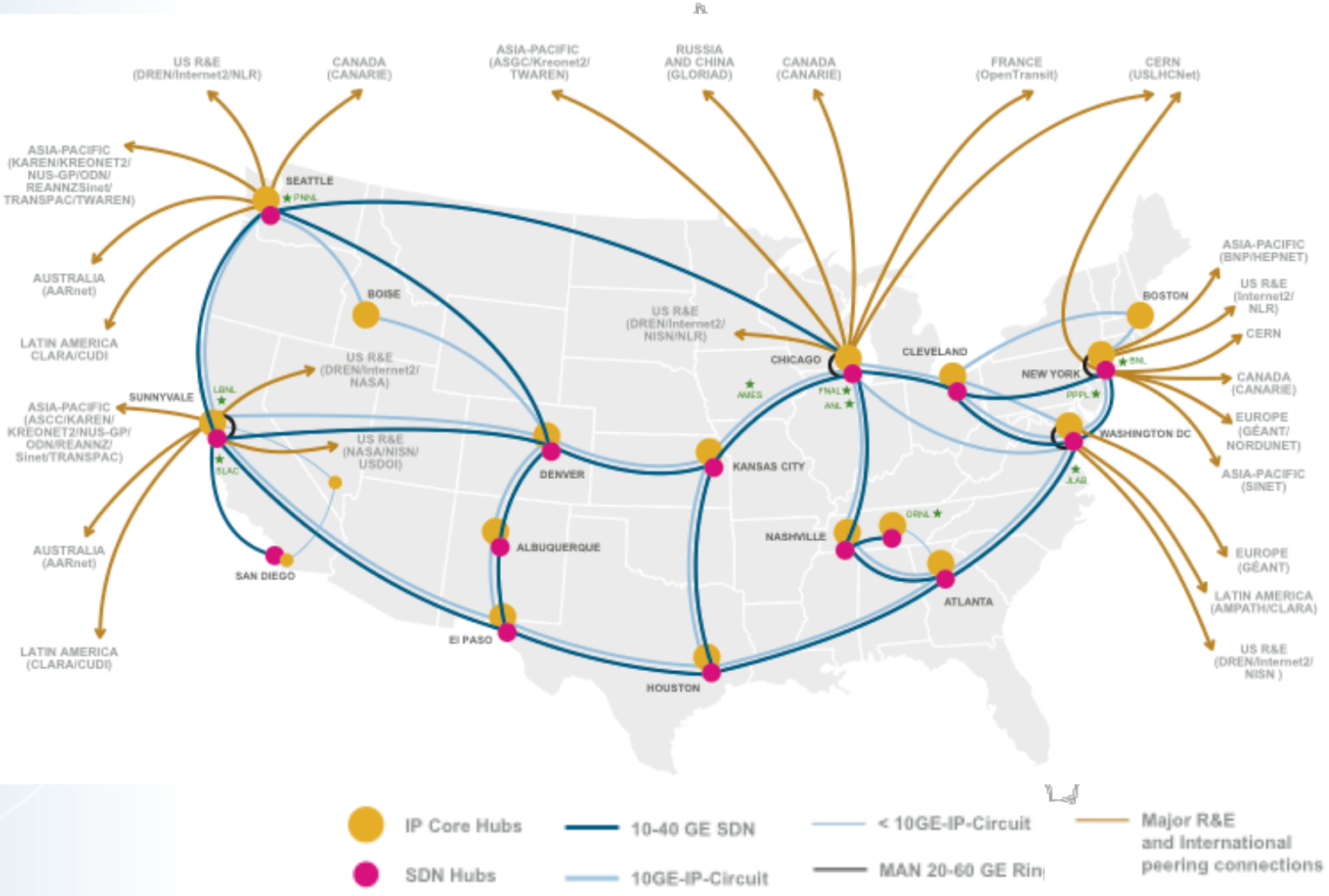
- PerfSONAR
- The Data Transfer Node
- Data Transfer Tools

Part 3:

- Science DMZ Security Best Practices
- Conclusions

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

- National Fiber footprint
- Distributed Team of 35
- Science Data Network
- International Collaborations
- Multiple 10G waves

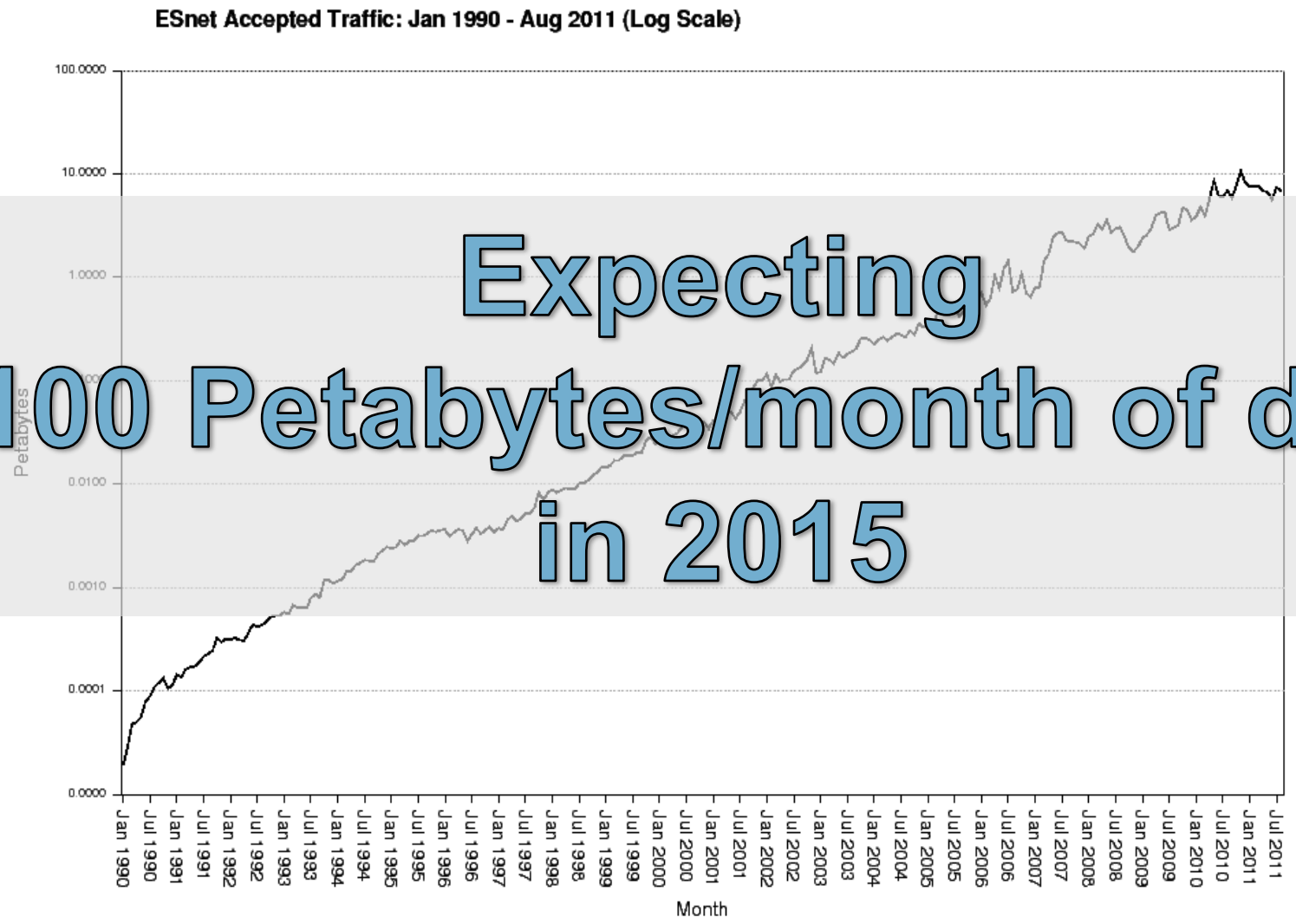


ESnet **Supports DOE Office of Science**

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

The Science Data Explosion

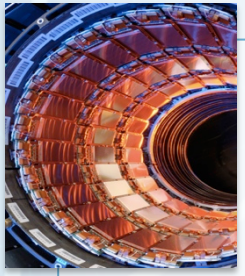


Expecting
100 Petabytes/month of data
in 2015

Data Explosion is Occurring Everywhere

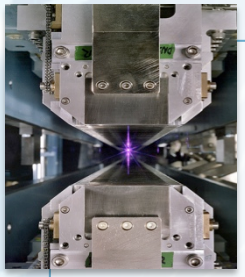
Genomics

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



High Energy Physics

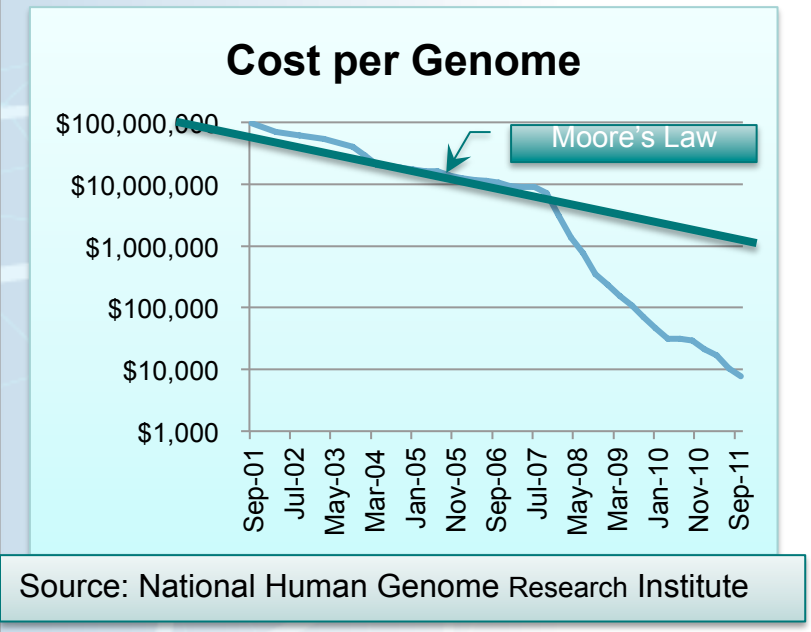
- LHC experiments produce & distribute petabytes of data/year
- Peak data rates increase 3-5x over 5 years



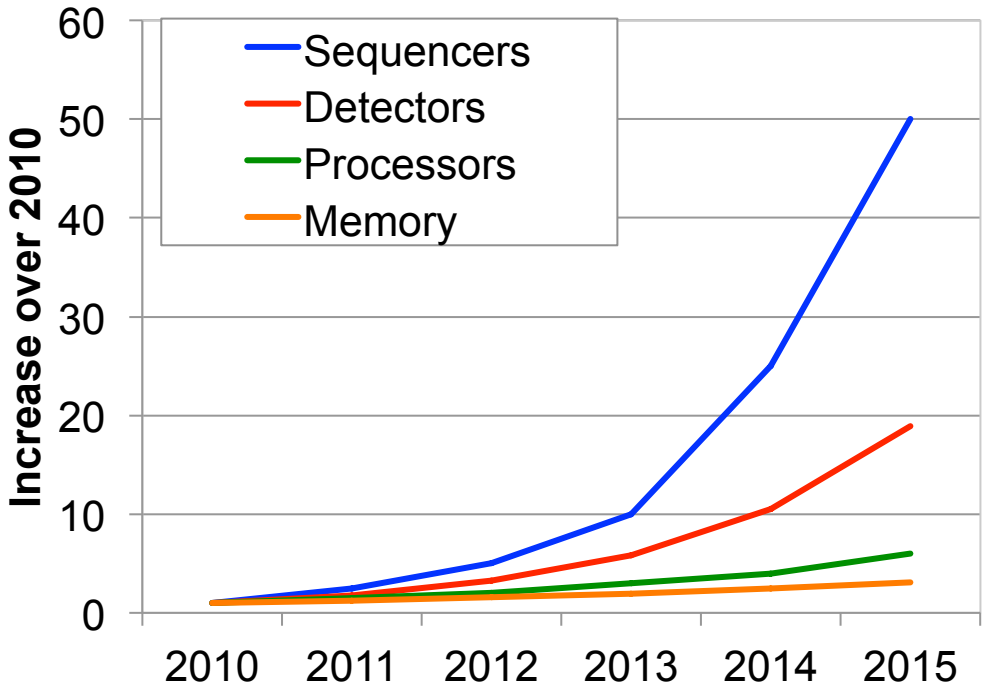
Light Sources

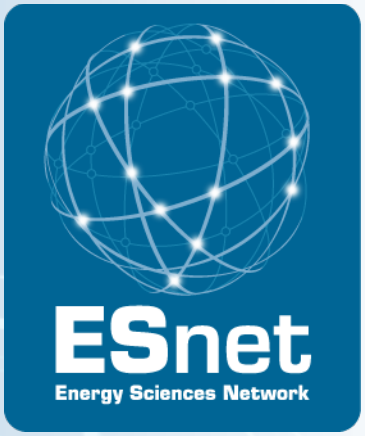
- Many detectors on a Moore's Law curve
- Data volumes rendering previous operational models obsolete

Sample Data Growth



Source: National Human Genome Research Institute





Motivation for the Science DMZ

Setting 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 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

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 and disks, packet loss issues, security devices
- A successful model has emerged – the Science DMZ
  - This model successfully in use by HEP (CMS/Atlas), Climate (ESG), several Supercomputer Centers, and others

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 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

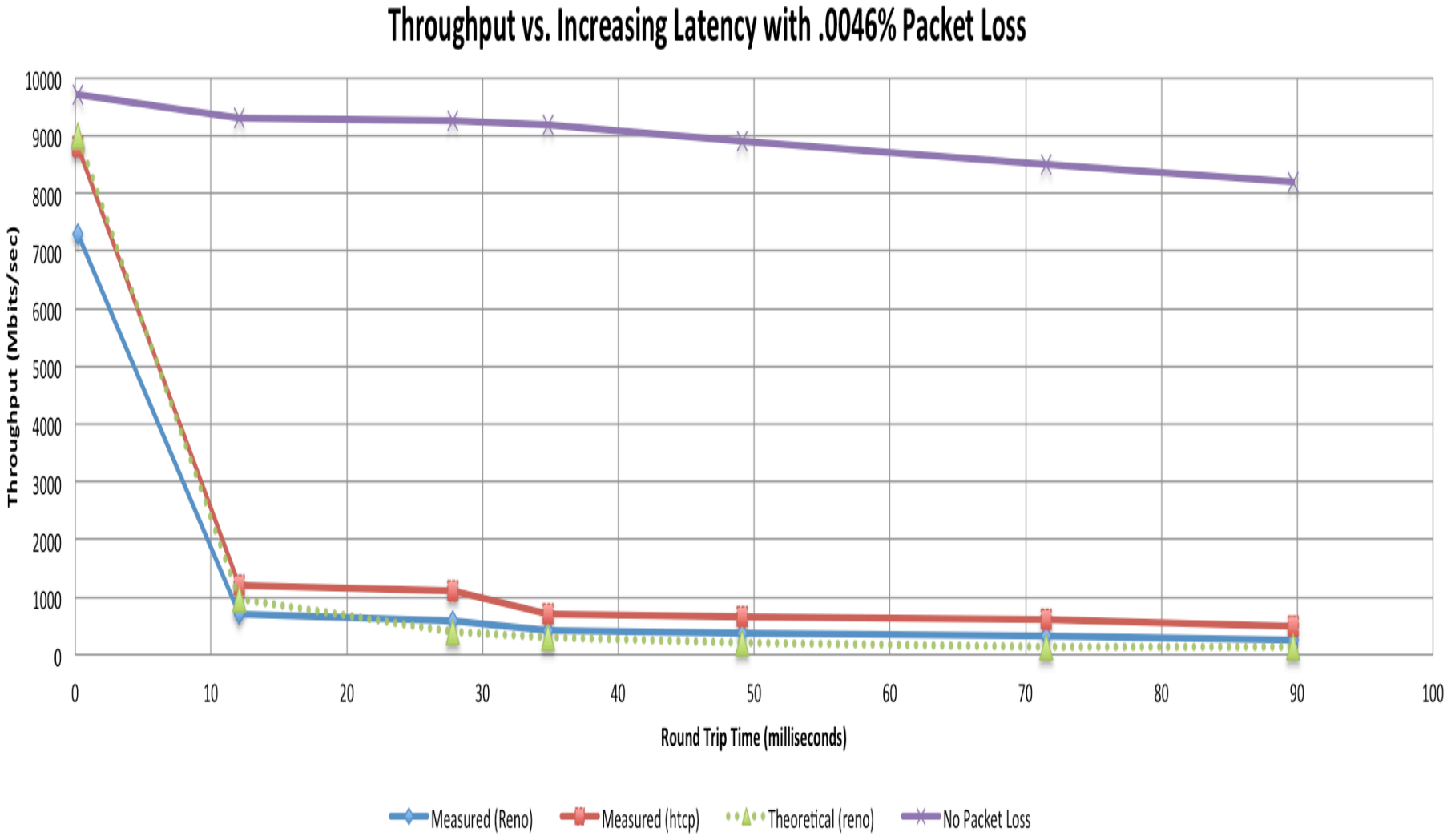
Small Router/Switch Buffers

- 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

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



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

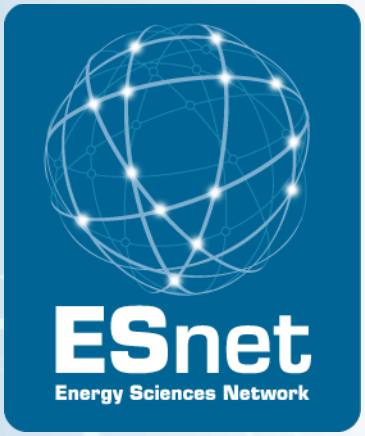
How Do We Accommodate TCP?

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

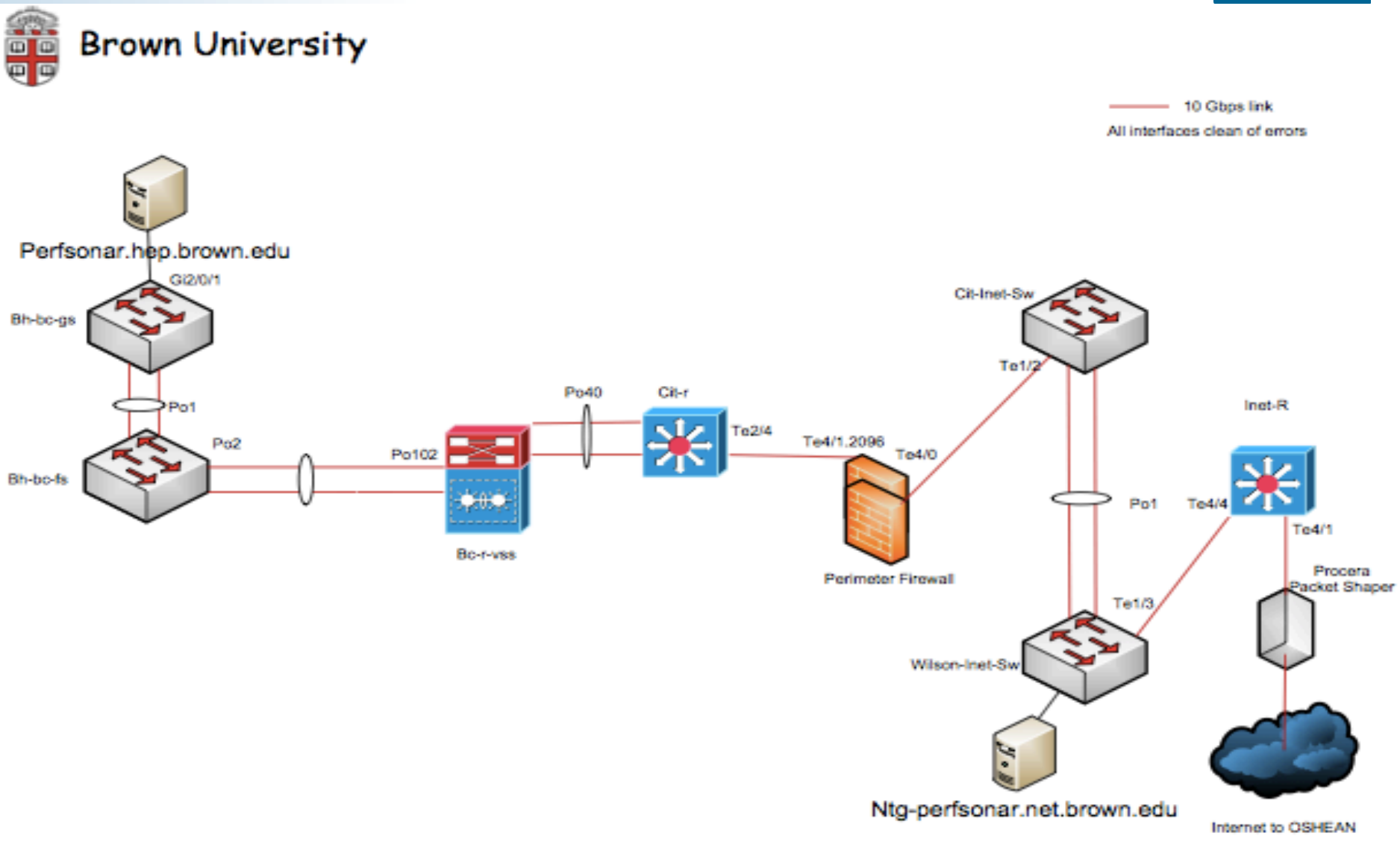
Equally important – the infrastructure must be verifiable so that clean service can be provided

- 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



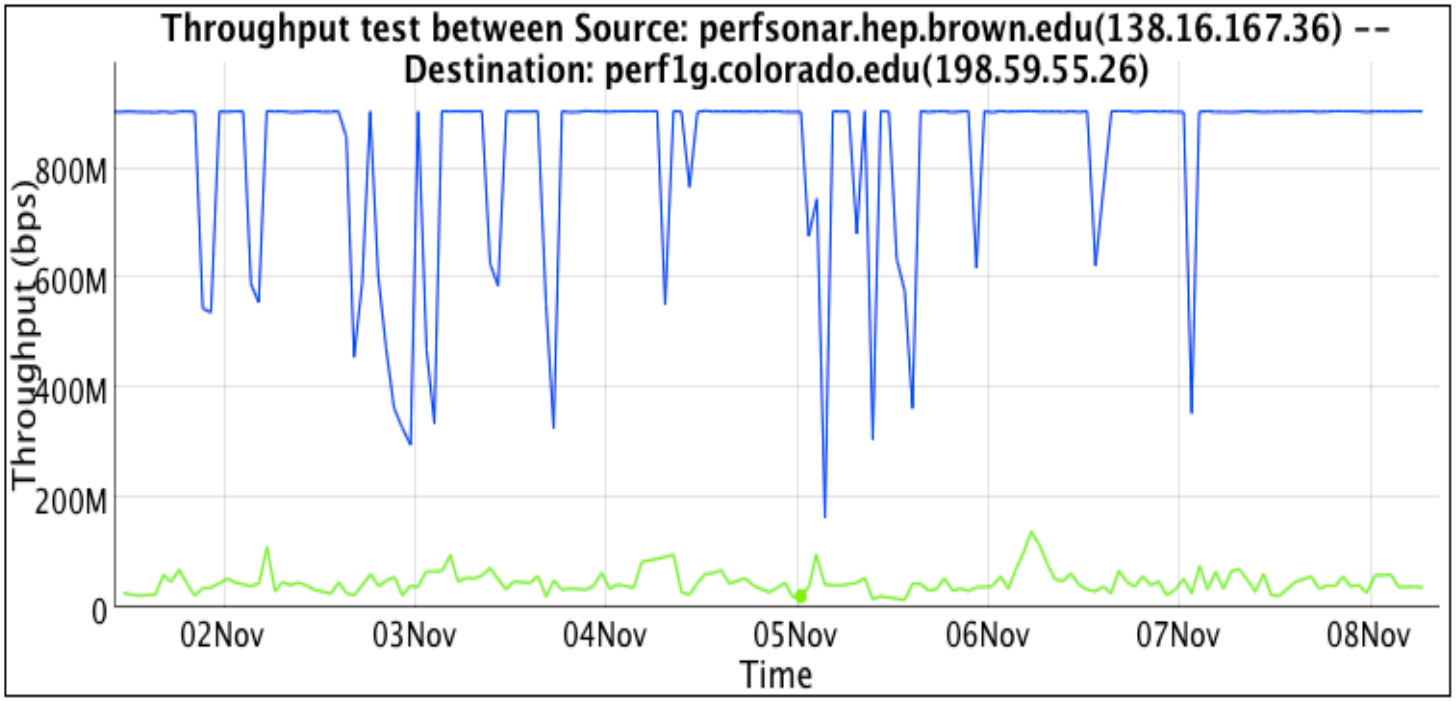
Performance Issues caused by Security Devices

Firewall Performance Example From Brown University



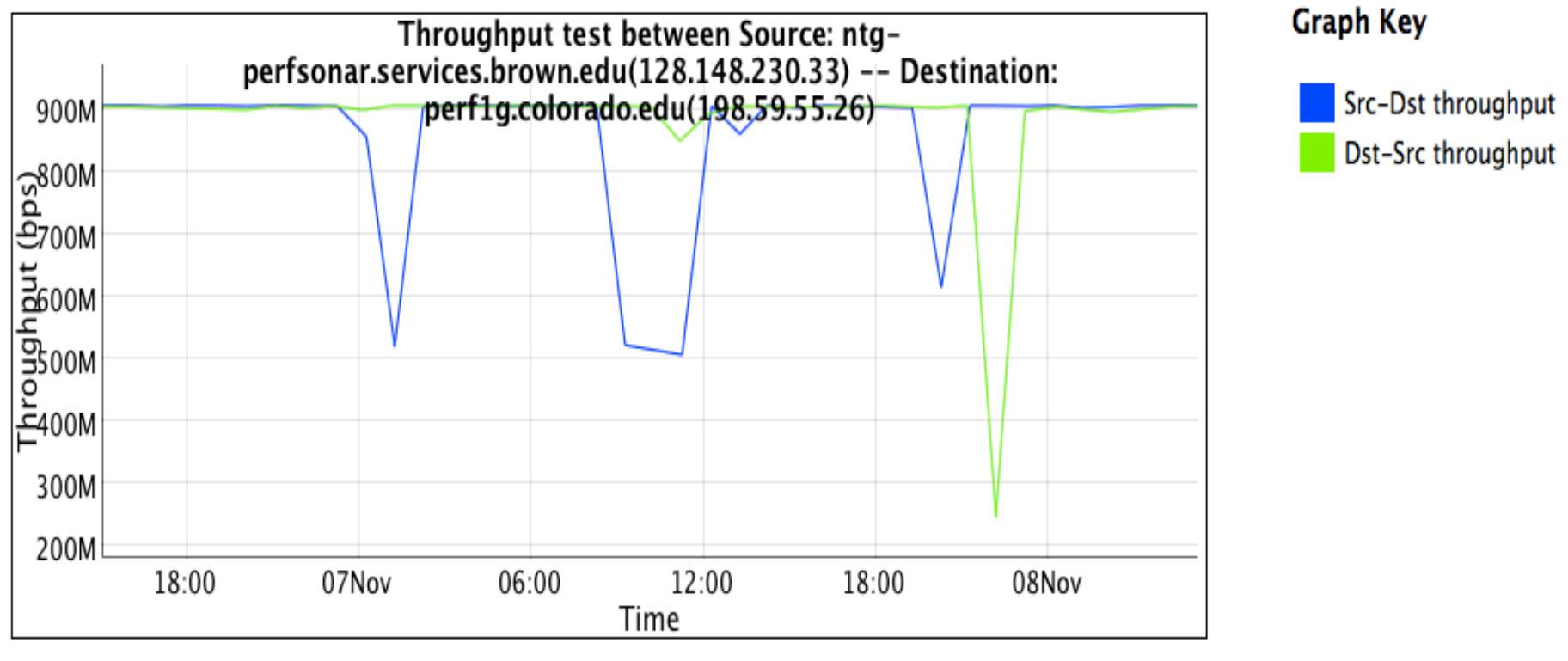
Brown University Example

Results to host behind the firewall:



Brown University Example

In front of the firewall:



Brown Univ. Example – TCP Dynamics

Want more proof – lets look at a measurement tool through the firewall.

- Measurement tools emulate a well behaved application

'Outbound', not filtered:

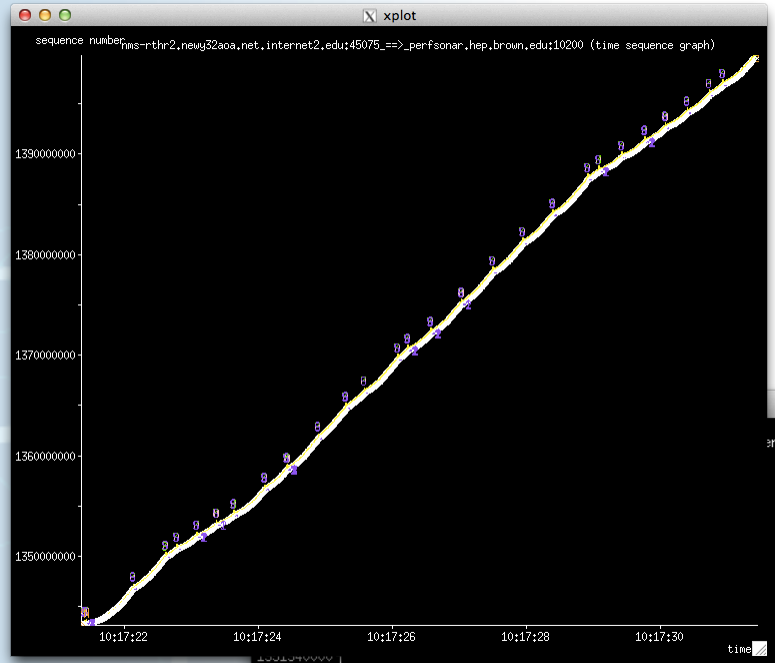
```
nuttcp -T 10 -i 1 -p 10200 bwctl.newy.net.internet2.edu
   92.3750 MB /   1.00 sec =  774.3069 Mbps     0 retrans
  111.8750 MB /   1.00 sec =  938.2879 Mbps     0 retrans
  111.8750 MB /   1.00 sec =  938.3019 Mbps     0 retrans
  111.7500 MB /   1.00 sec =  938.1606 Mbps     0 retrans
  111.8750 MB /   1.00 sec =  938.3198 Mbps     0 retrans
  111.8750 MB /   1.00 sec =  938.2653 Mbps     0 retrans
  111.8750 MB /   1.00 sec =  938.1931 Mbps     0 retrans
  111.9375 MB /   1.00 sec =  938.4808 Mbps     0 retrans
  111.6875 MB /   1.00 sec =  937.6941 Mbps     0 retrans
  111.8750 MB /   1.00 sec =  938.3610 Mbps     0 retrans

 1107.9867 MB /  10.13 sec =  917.2914 Mbps 13 %TX 11 %RX 0
retrans 8.38 msRTT
```

Thru the firewall

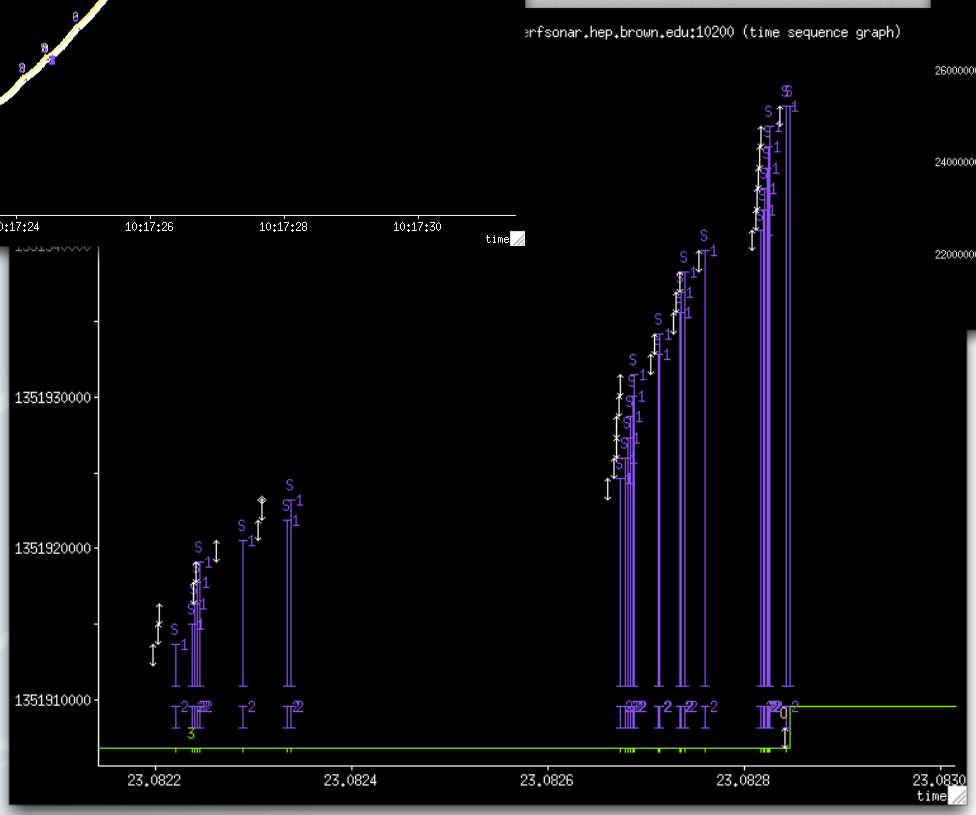
'Inbound', filtered:

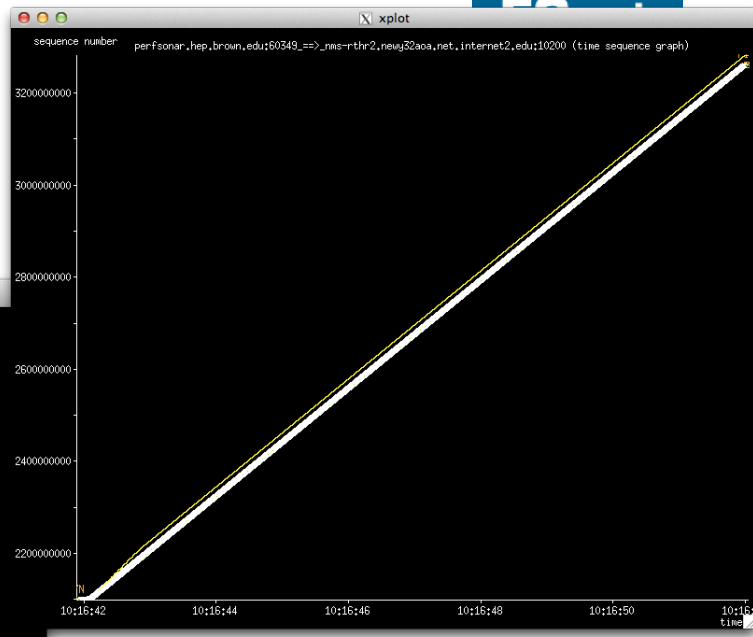
```
nuttcp -r -T 10 -i 1 -p 10200 bwctl.newy.net.internet2.edu
    4.5625 MB /   1.00 sec =   38.1995 Mbps    13 retrans
    4.8750 MB /   1.00 sec =   40.8956 Mbps     4 retrans
    4.8750 MB /   1.00 sec =   40.8954 Mbps     6 retrans
    6.4375 MB /   1.00 sec =   54.0024 Mbps     9 retrans
    5.7500 MB /   1.00 sec =   48.2310 Mbps     8 retrans
    5.8750 MB /   1.00 sec =   49.2880 Mbps     5 retrans
    6.3125 MB /   1.00 sec =   52.9006 Mbps     3 retrans
    5.3125 MB /   1.00 sec =   44.5653 Mbps     7 retrans
    4.3125 MB /   1.00 sec =   36.2108 Mbps     7 retrans
    5.1875 MB /   1.00 sec =   43.5186 Mbps     8 retrans

   53.7519 MB /  10.07 sec =   44.7577 Mbps 0 %TX 1 %RX 70
retrans 8.29 msRTT
```

tcptrace output: with and without a firewall



firewall





No firewall

Security Without Firewalls

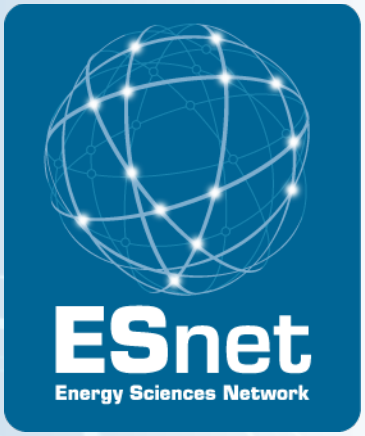
Does this mean we ignore security? ***NO!***

- 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 solutions

- Host-based IDS and firewalls
- Intrusion detection (Bro, Snort, others), flow analysis, …
- Tight ACLs reduce attack surface (possible in many but not all cases)
- ***Key point – performance is a mission requirement, and the security policies and mechanisms that protect the Science DMZ should be architected so that they serve the mission***

Much more on this topic in part 3 of this session.



Enter the Science DMZ

Traditional DMZ

DMZ – “Demilitarized Zone”

- Network segment near the site perimeter with different security policy
- Commonly used architectural element for deploying WAN-facing services (e.g. email, DNS, web)

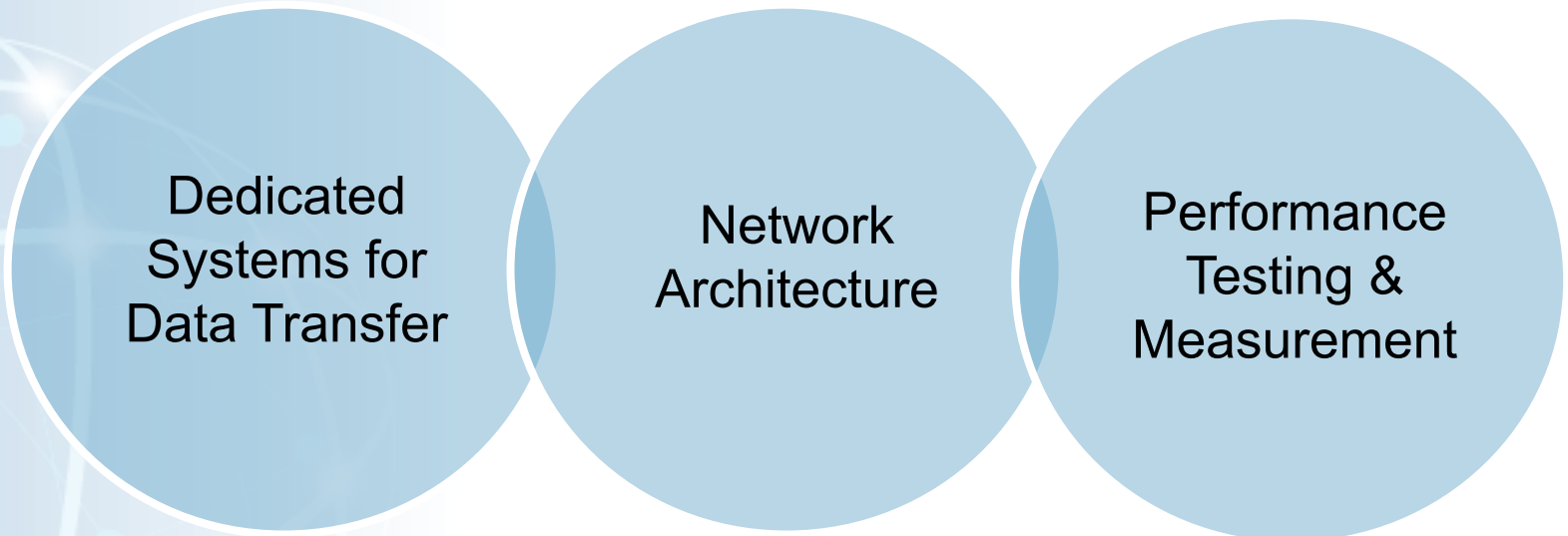
Traffic for WAN-facing services does not traverse the LAN

- WAN flows are isolated from LAN traffic
- Infrastructure for WAN services is specifically configured for WAN

Separation of security policy improves both LAN and WAN

- No conflation of security policy between LAN hosts and WAN services
- DMZ hosts provide specific services
- LAN hosts must traverse the same ACLs as WAN hosts to access DMZ

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

Science DMZ Takes Many Forms

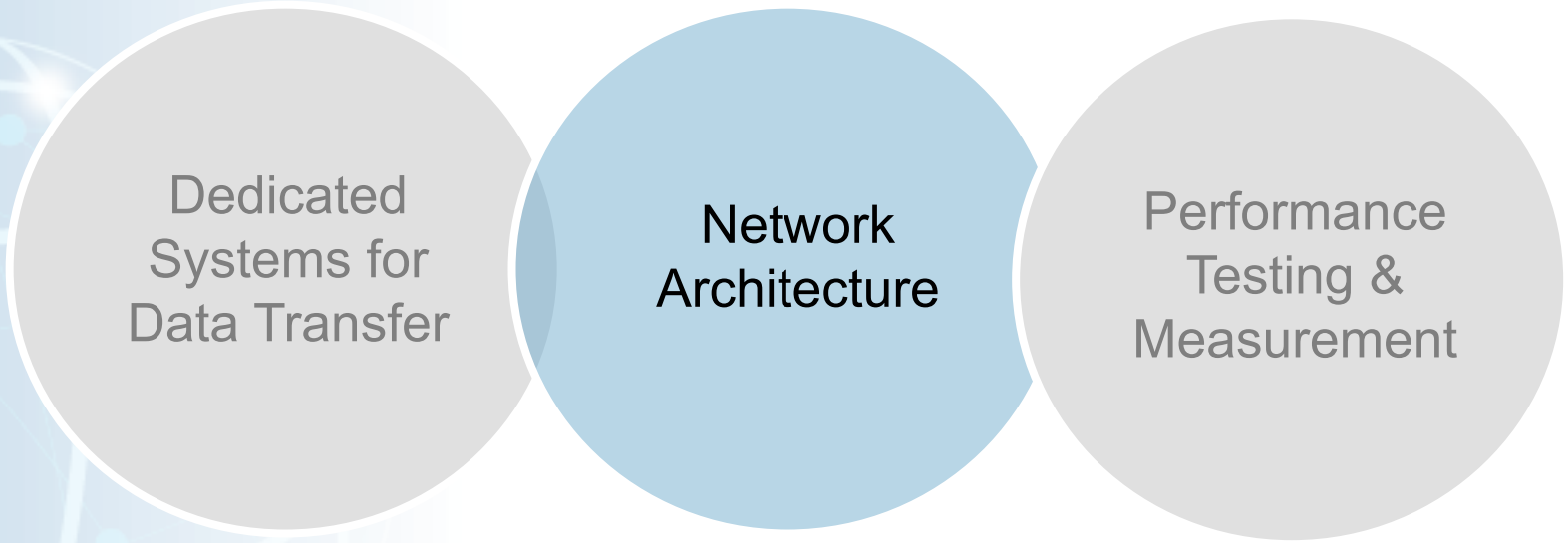
There are a lot of ways to combine these things – it all depends on what you need to do

- 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

Key point of concentration: eliminate sources of packet loss / packet friction

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

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

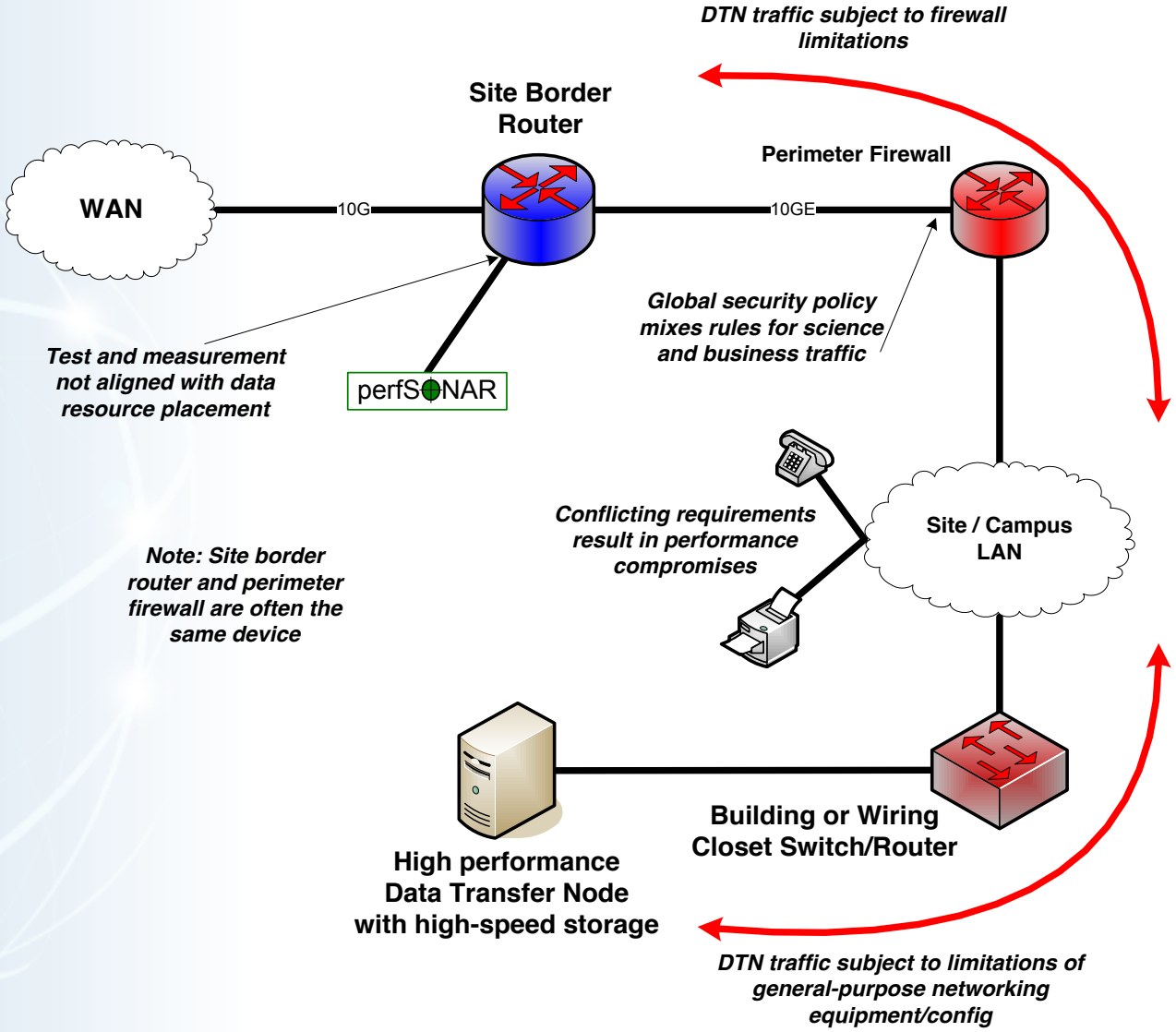
- This is a good start
- Need a second one next to the DTN

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.

Ad Hoc DTN Deployment



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

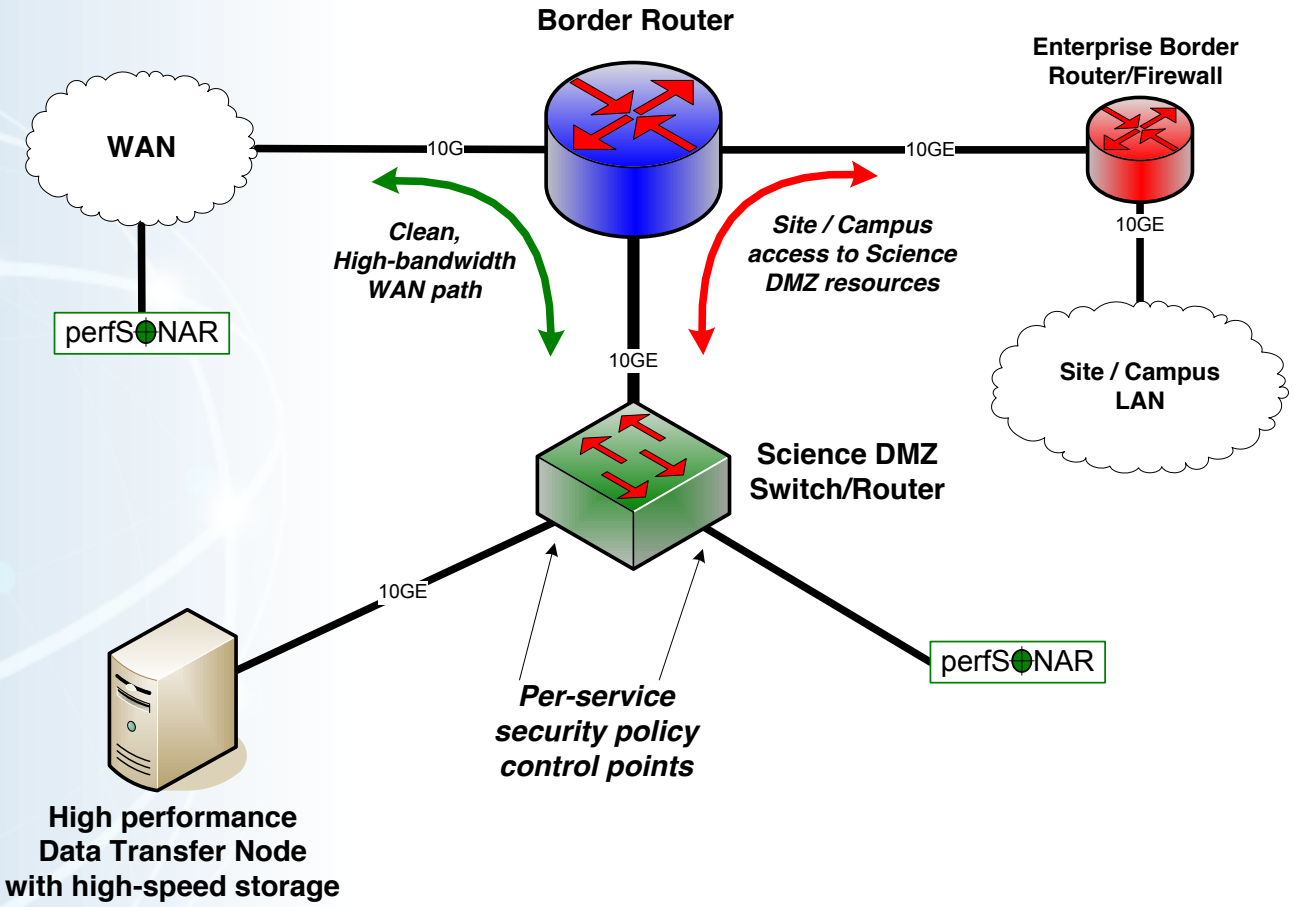
Easy to experiment with components and technologies

- DTN prototyping
- perfSONAR testing

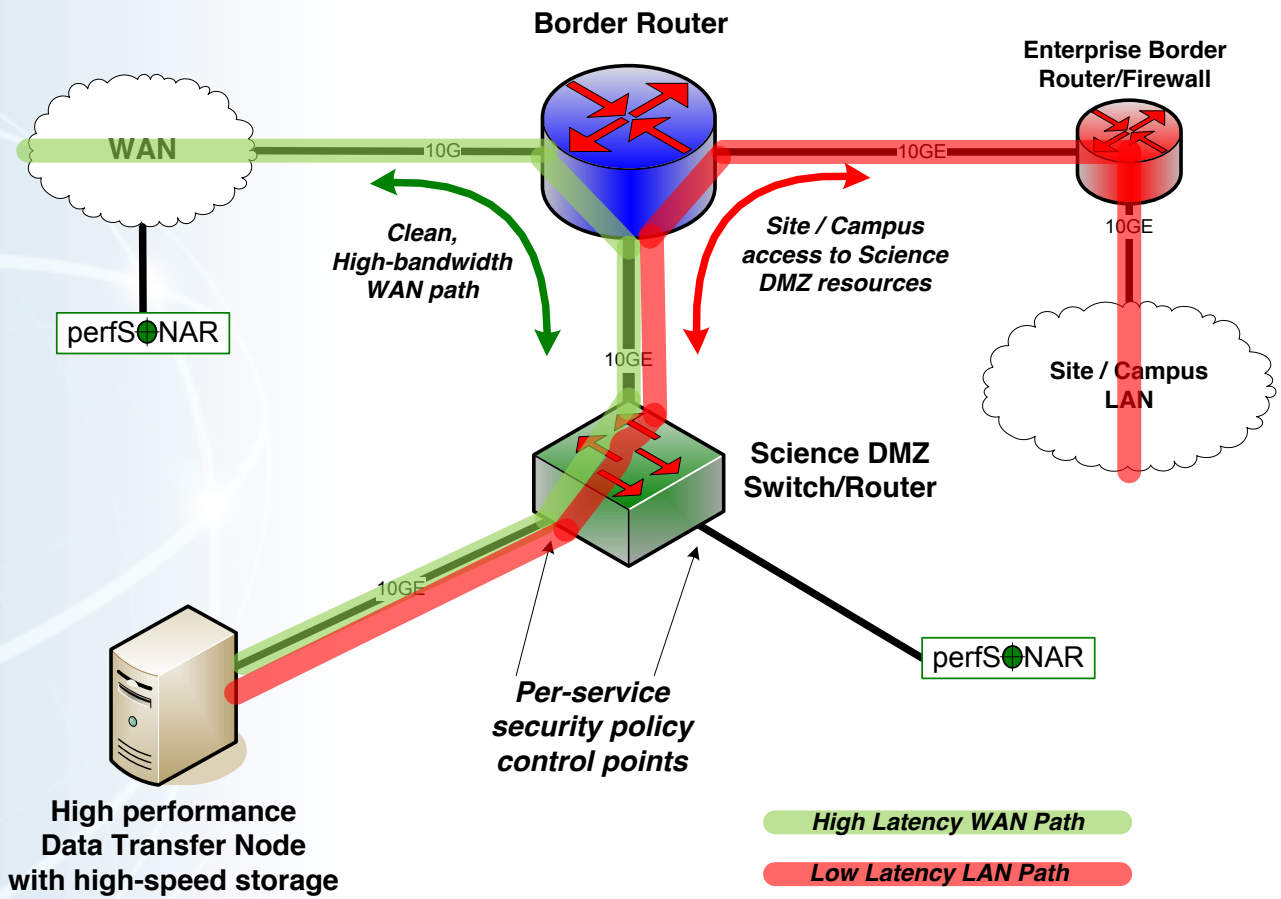
Limited scope makes security policy exceptions easy

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

A better approach: simple Science DMZ



Prototype Science DMZ Data Path



Prototype With Virtual Circuits

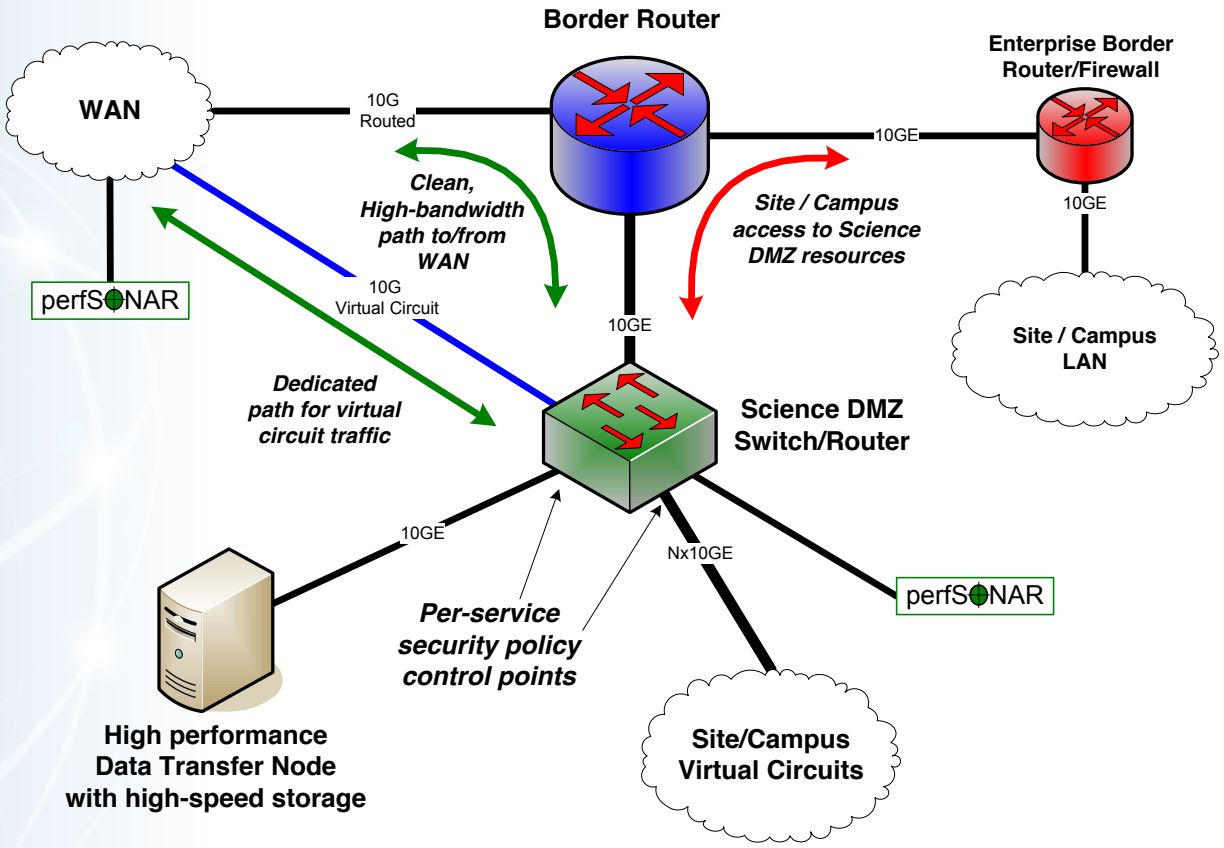
Small virtual circuit prototype can be done in a small Science DMZ

- Perfect example is a Software Defined Networking (SDN) testbed
- Virtual circuit connection may or may not traverse border router

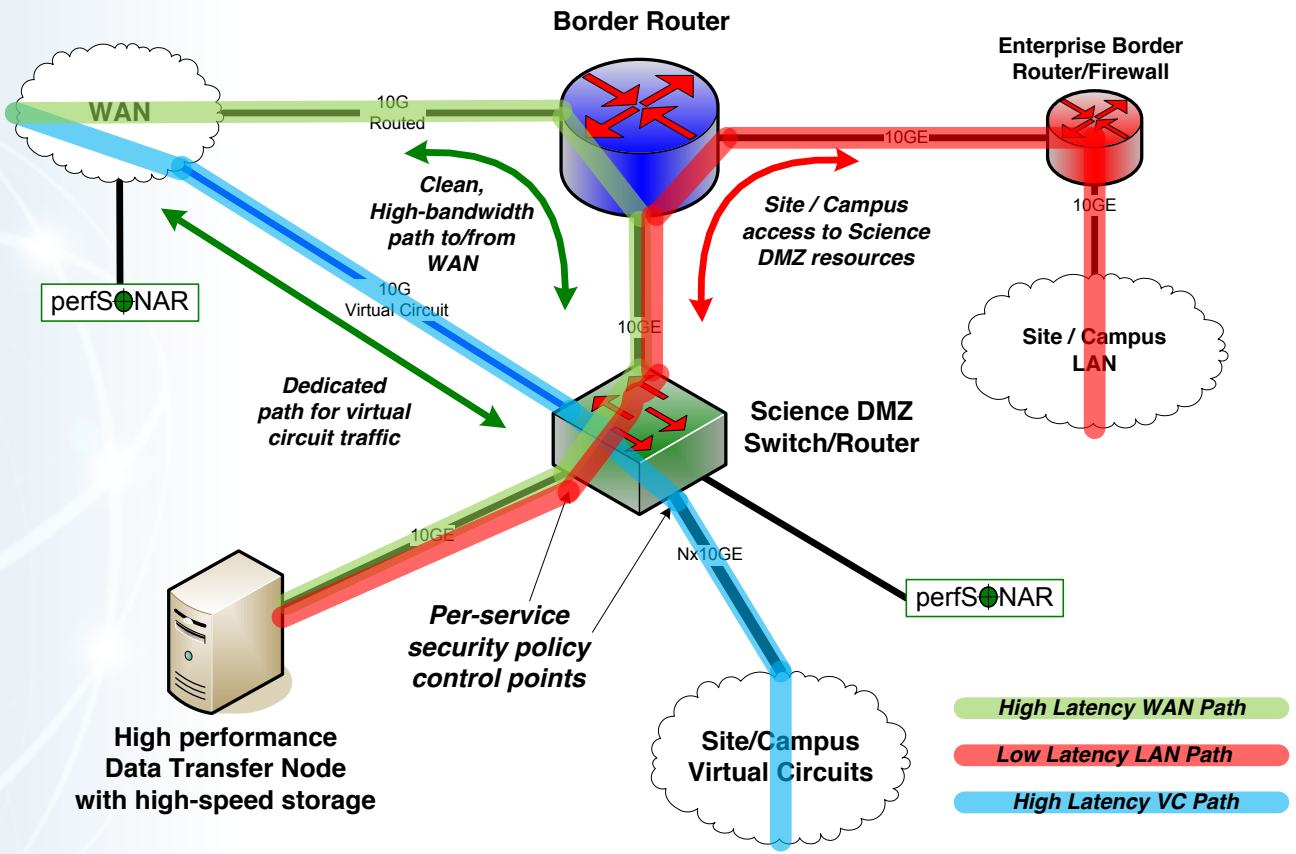
As with any Science DMZ deployment, this can be expanded as need grows

In this particular diagram, Science DMZ hosts can use either the routed or the circuit connection

Virtual Circuit Prototype Deployment



Virtual Circuit Prototype Data Path



Support For Multiple Projects

Science DMZ architecture allows multiple projects to put DTNs in place

- Modular architecture
- Centralized location for data servers

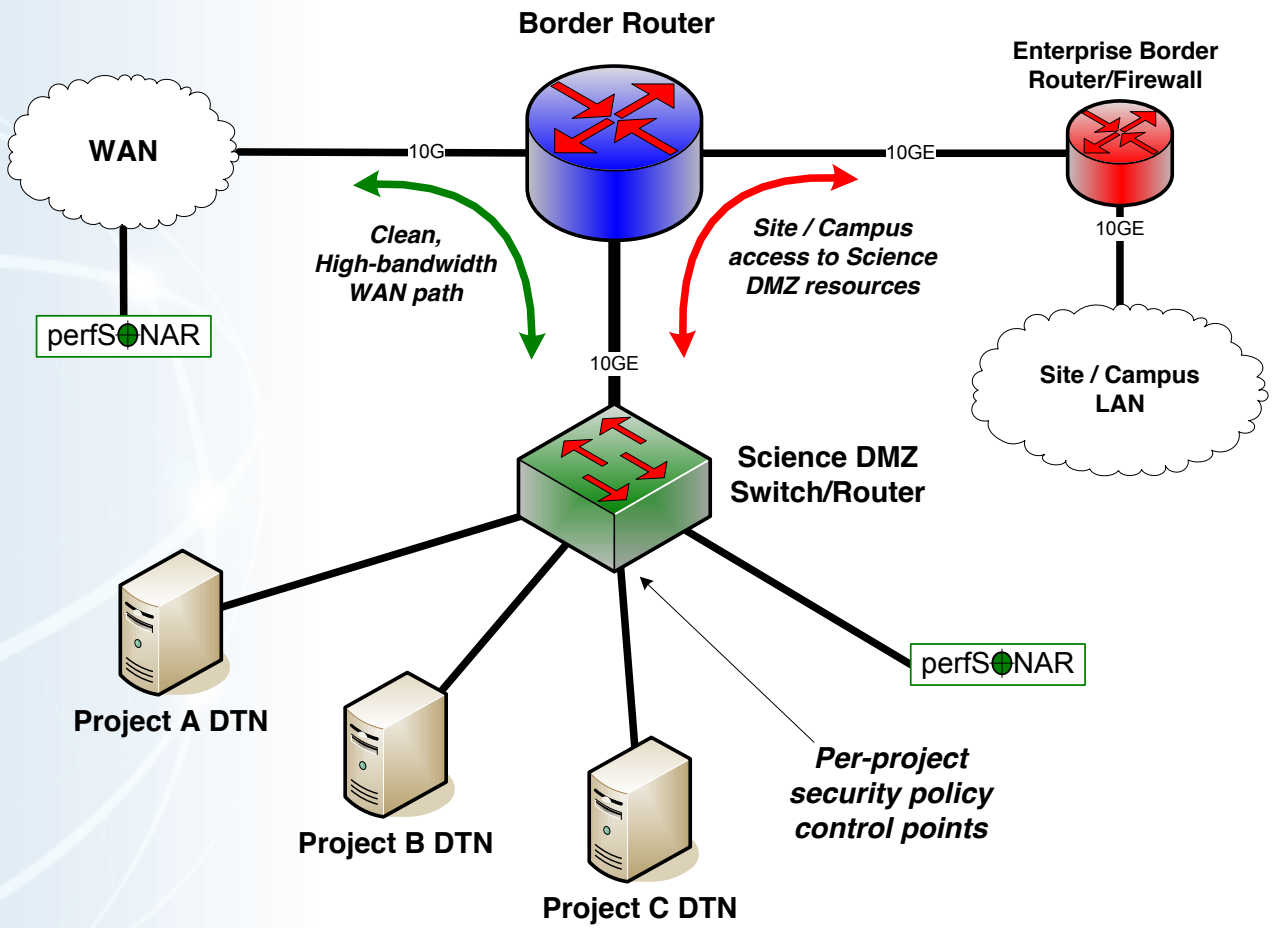
This may or may not work well depending on institutional politics

- Issues such as physical security can make this a non-starter
- On the other hand, some shops already have service models in place

On balance, this can provide a cost savings – it depends

- Central support for data servers vs. carrying data flows
- How far do the data flows have to go?

Multiple Projects



Supercomputer Center Deployment

High-performance networking is assumed in this environment

- Data flows between systems, between systems and storage, wide area, etc.
- Global filesystem often ties resources together
  - Portions of this may not run over Ethernet (e.g. IB)
  - Implications for Data Transfer Nodes

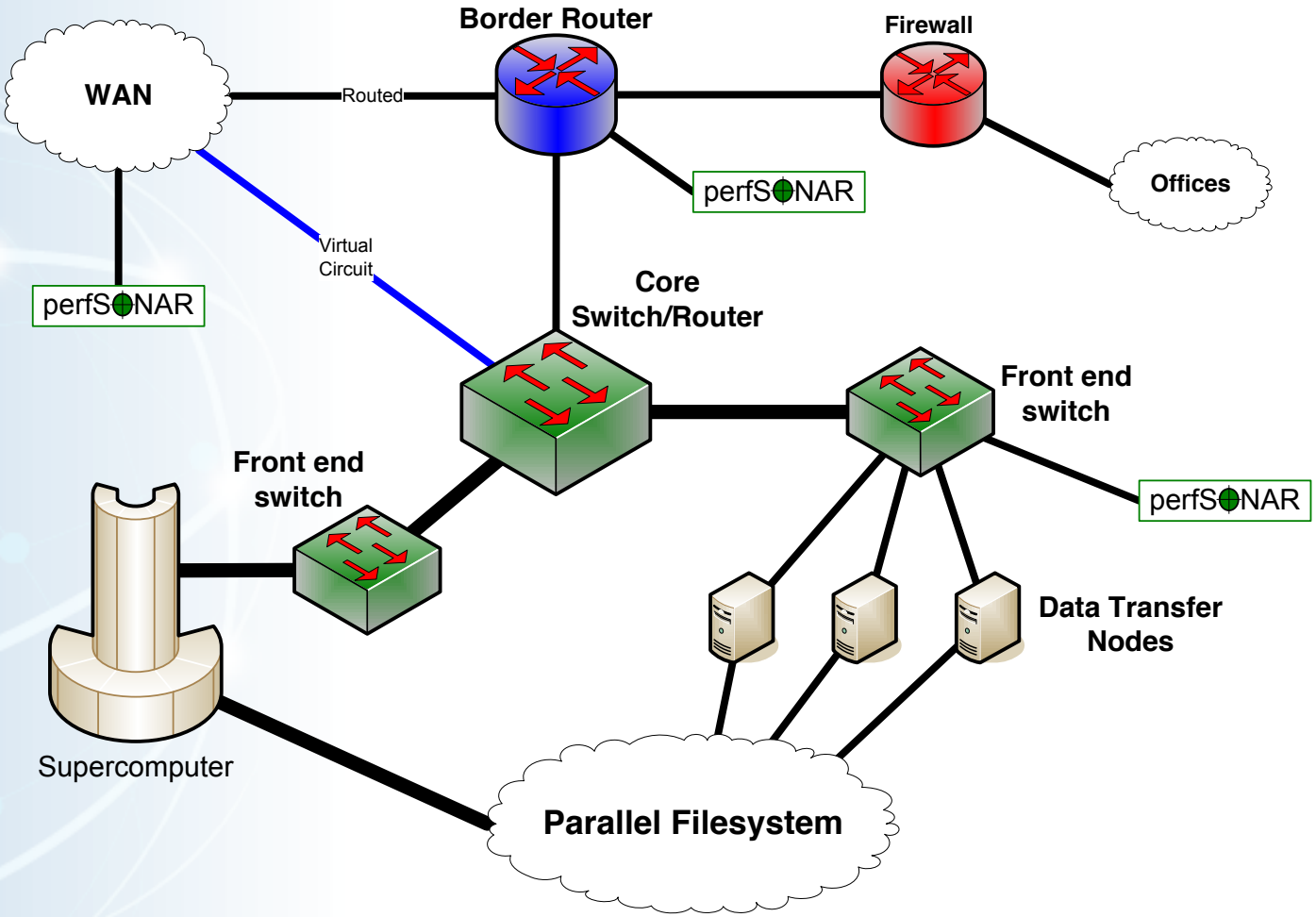
“Science DMZ” may not look like a discrete entity here

- By the time you get through interconnecting all the resources, you end up with most of the network in the Science DMZ
- This is as it should be – the point is appropriate deployment of tools, configuration, policy control, etc.

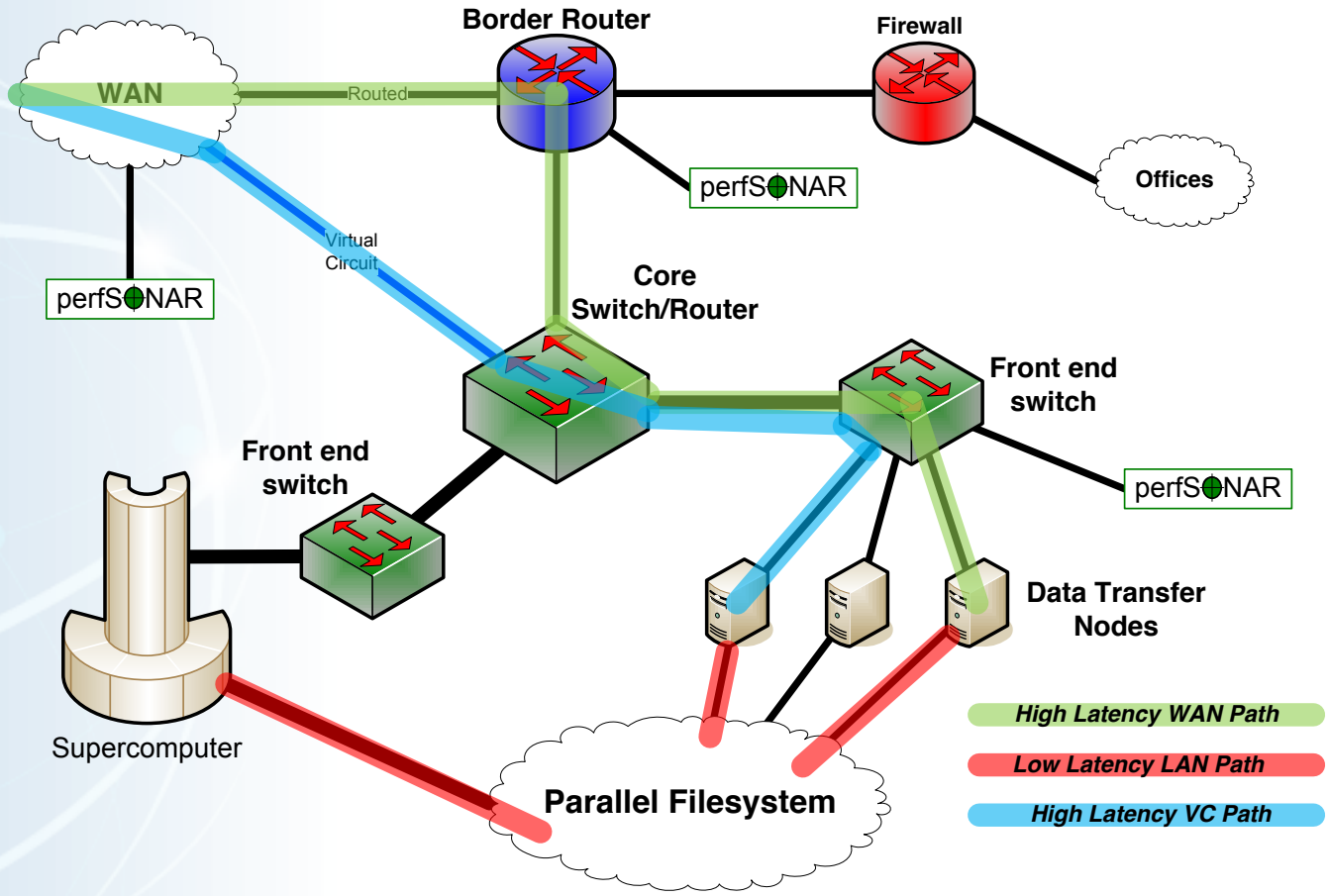
Office networks can look like an afterthought, but they aren’t

- Deployed with appropriate security controls
- Office infrastructure need not be sized for science traffic

Supercomputer Center



Supercomputer Center Data Path



Major Data Site Deployment

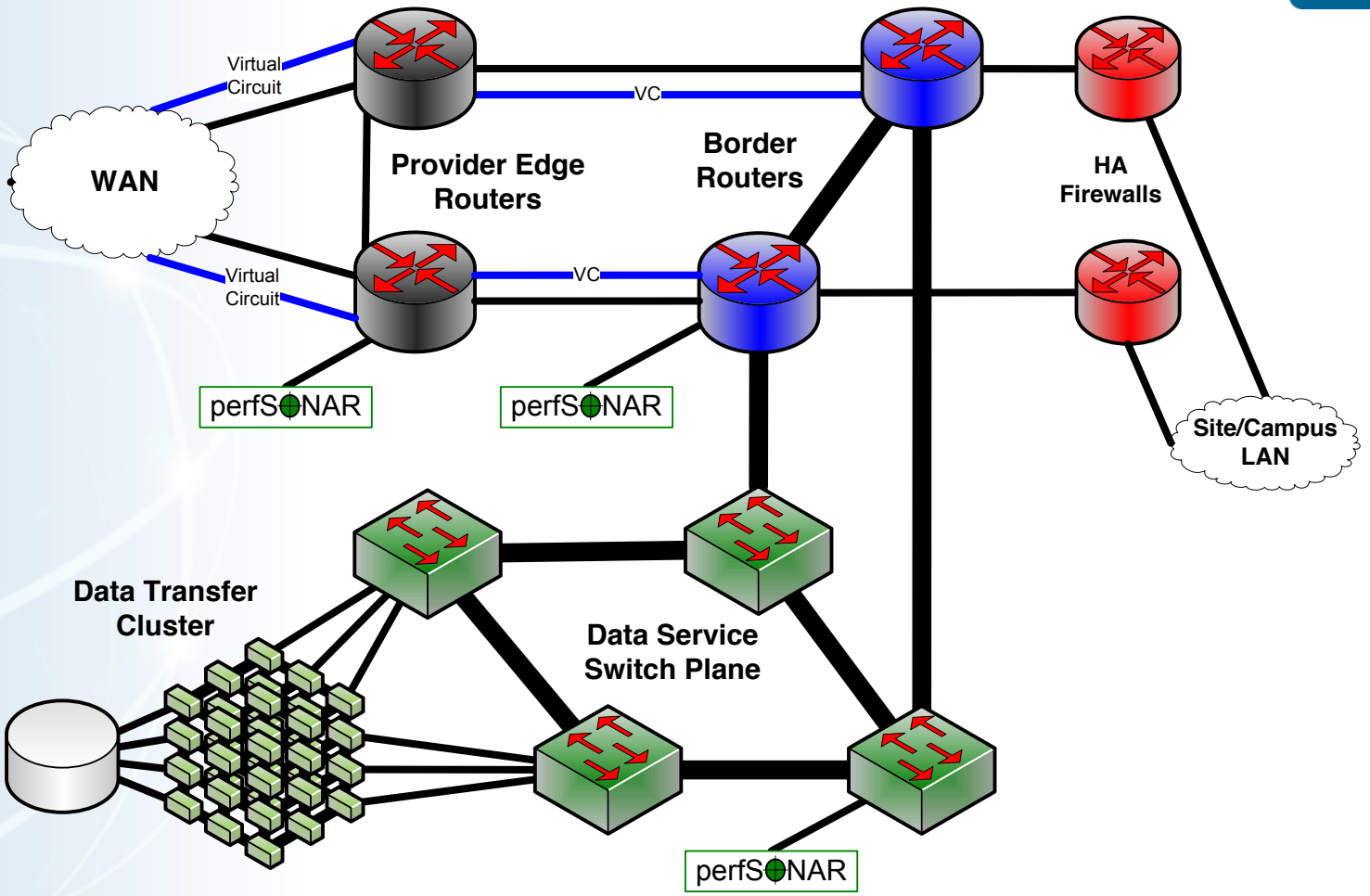
In some cases, large scale data service is the major driver

- Huge volumes of data – ingest, export
- Individual DTNs don't exist here – data transfer clusters

Single-pipe deployments don't work

- Everything is parallel
  - Networks (Nx10G LAGs, soon to be Nx100G)
  - Hosts – data transfer clusters, no individual DTNs
  - WAN connections – multiple entry, redundant equipment
- Choke points (e.g. firewalls) cause problems

Data Site – Architecture



Distributed Science DMZ

Fiber-rich environment enables distributed Science DMZ

- No need to accommodate all equipment in one location
- Allows the deployment of institutional science service

WAN services arrive at the site in the normal way

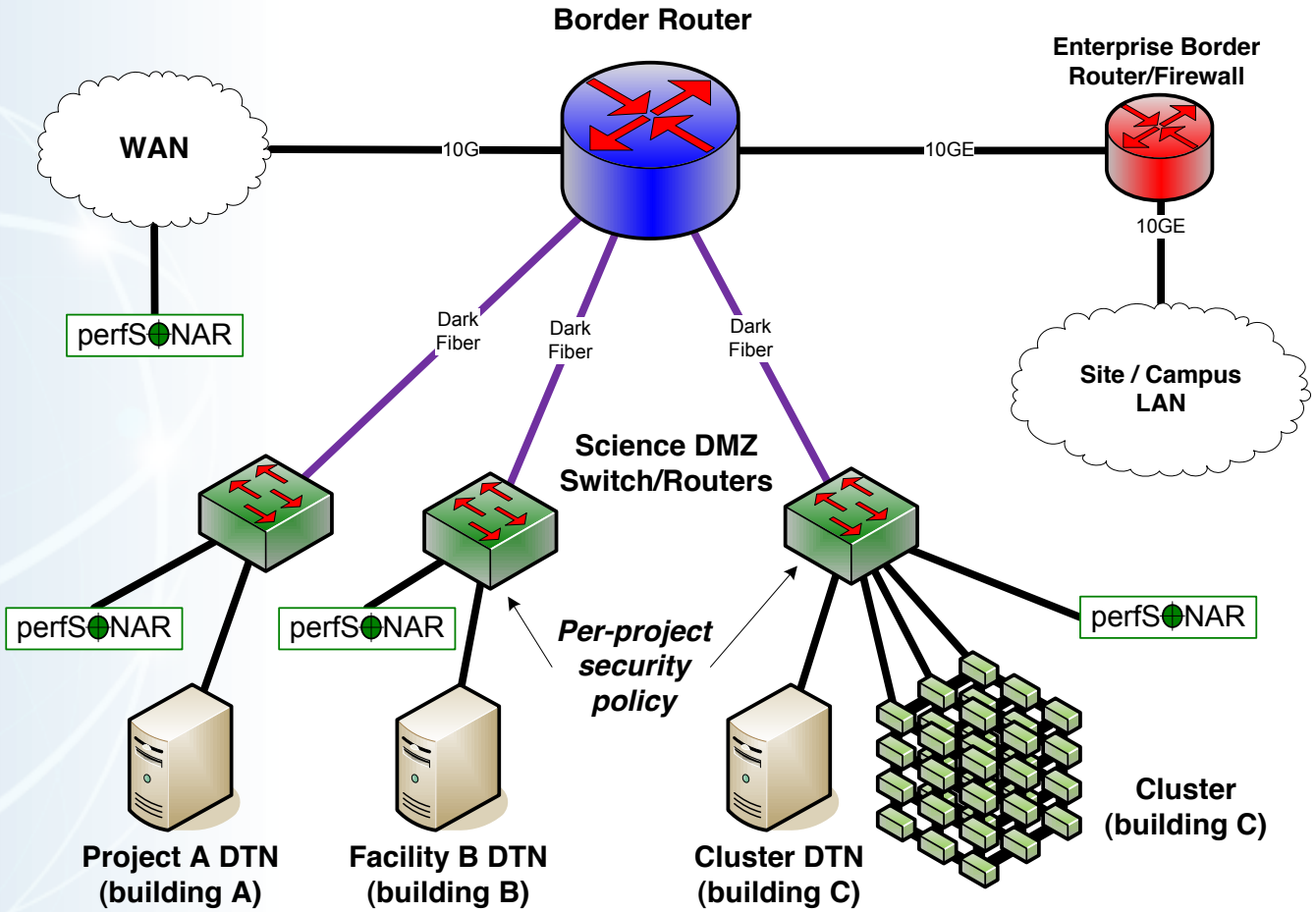
Dark fiber distributes connectivity to Science DMZ services throughout the site

- Departments with their own networking groups can manage their own local Science DMZ infrastructure
- Facilities or buildings can be served without building up the business network to support those flows

Security is potentially more complex

- Remote infrastructure must be monitored
- Several technical remedies exist (arpwatch, no DHCP, separate address space, etc)
- Solutions depend on relationships with security groups

Multiple Science DMZs – Dark Fiber



Summary So Far

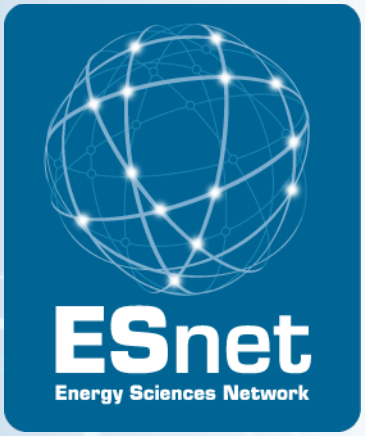
There is no single “correct” way architect a Science DMZ

- these are a few “design patterns”

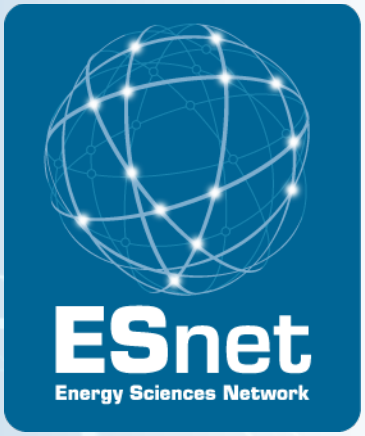
It depends on things like:

- site requirements
- existing resources
- availability of dark fiber
- budget

The main point is to reduce the opportunities for packet loss



End of Part 1



The Science DMZ, Part 2

Brian Tierney, Eli Dart, Eric Pouyoul, Jason Zurawski

ESnet

Supporting Data-Intensive Research Workshop

QuestNet 2013

Gold Coast, Australia

July 2, 2013

Overview

Part 1:

- What is ESnet?
- Science DMZ Motivation
- Science DMZ Architecture

Part 2:

- **PerfSONAR**
- **The Data Transfer Node**
- **Data Transfer Tools**

Part 3:

- Science DMZ Security Best Practices
- Conclusions

Common Themes

Two common threads exist in all the examples in part 1:

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
- 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

The Data Transfer Trifecta: The "Science DMZ" Model

Dedicated Systems for Data Transfer

Network Architecture

Performance Testing & Measurement

Data Transfer Node

- High performance
- Configured for data transfer
- Proper tools

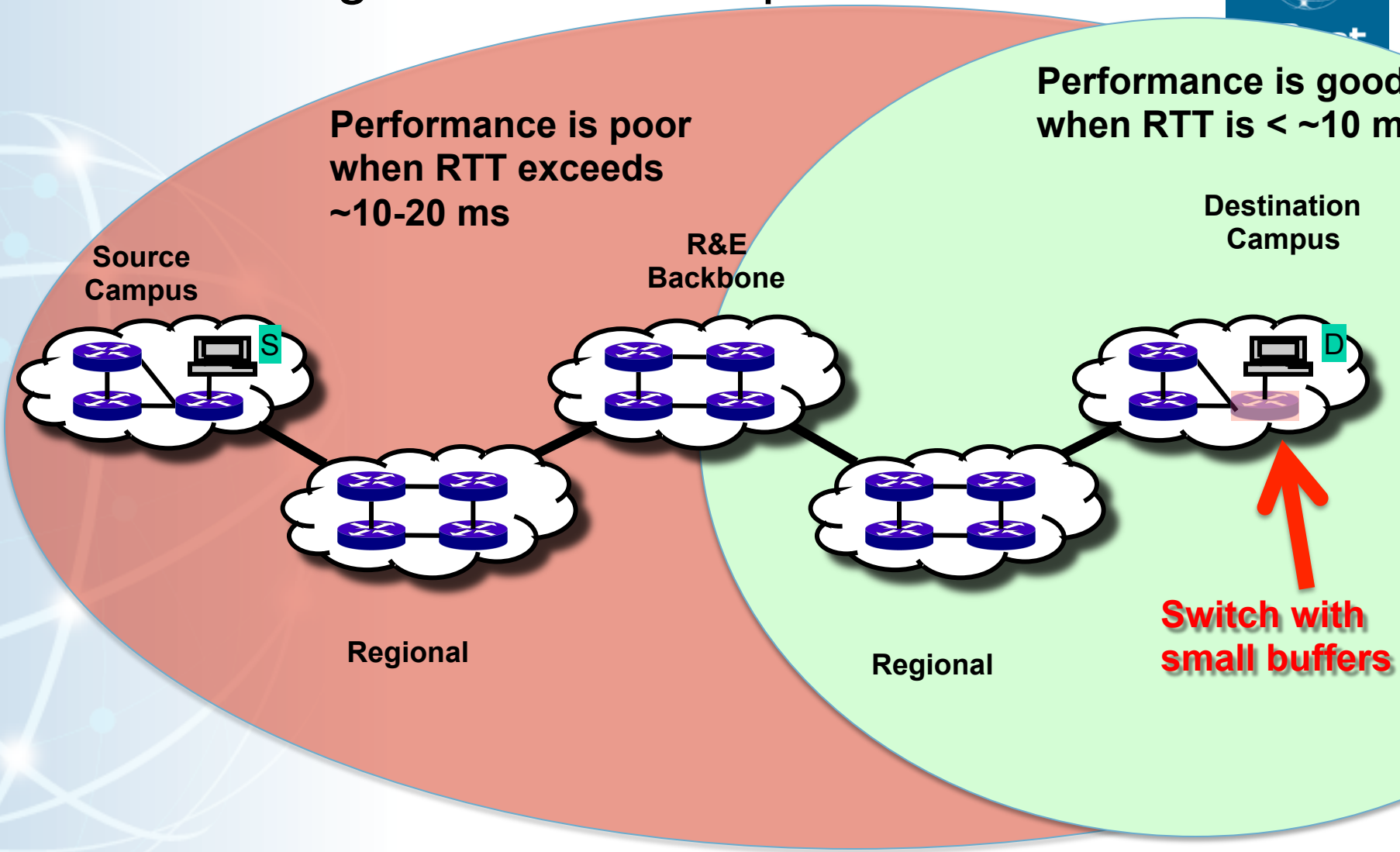
Science DMZ

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

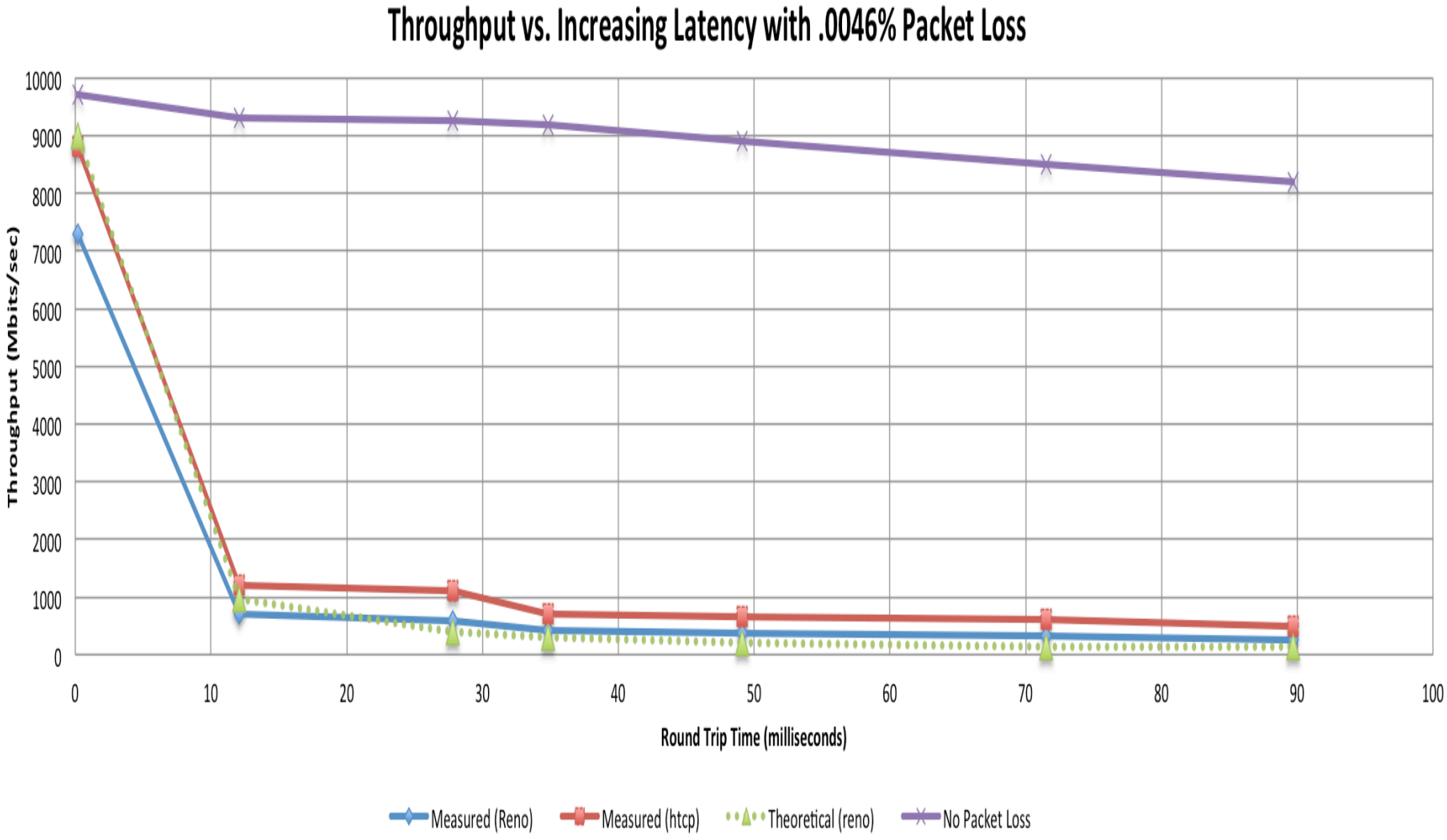
perfSONAR

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

Local testing will not find all problems



Remember this plot?



What is perfSONAR?

perfSONAR is a tool to:

- Set network performance expectations
- Find network problems (“soft failures”)
- Help fix these problems

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

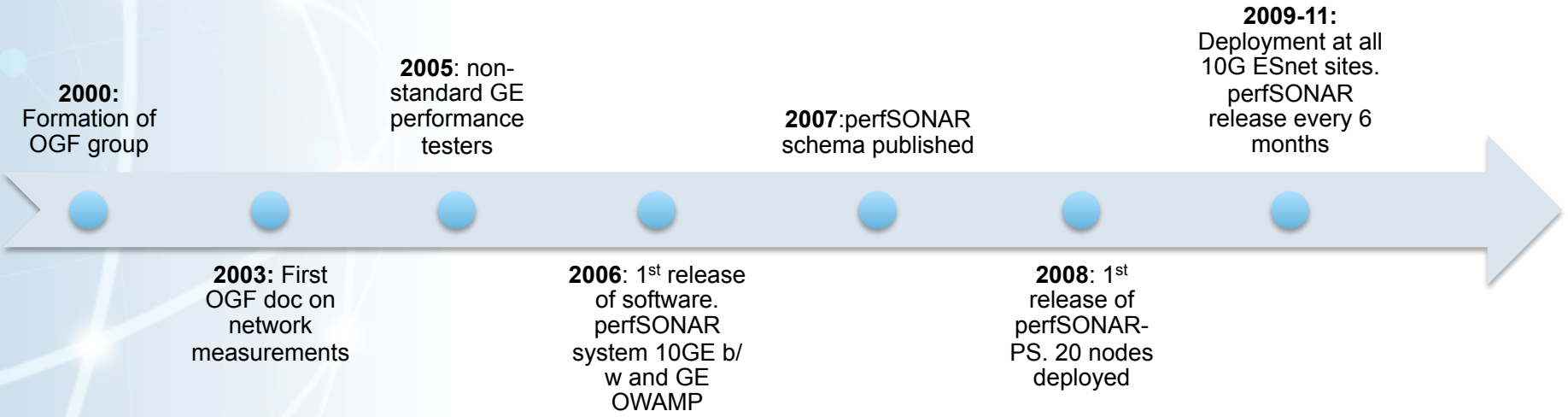
perfSONAR Motivation and Timeline

Motivation

- Help set accurate performance expectations
- Make it easier to troubleshoot performance issues

Objective

- Provide an infrastructure that identifies and isolates performance problems in a multi-domain environment.



Impact

- Almost every ESnet site has fixed network performance problems discovered by perfSONAR
- LHC collaboration has deployed extensively
- 580+ perfSONAR hosts deployed on 100 networks and in 14 countries

perfSONAR-PS Software

perfSONAR-PS is an open source implementation of the perfSONAR measurement infrastructure and protocols

- written in the perl programming language

http://psps.perfsonar.net

All products are available as RPMs.

The perfSONAR-PS consortium supports CentOS 6.

RPMs are compiled for i386 and x86 64

Functionality on other platforms and architectures is possible, but not supported.

- Should work: Red Hat Enterprise Linux and Scientific Linux (v5)
- Harder, but possible:
  - Fedora Linux, SuSE, Debian Variants

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 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

Lookup Service Directory Search

https://stats.es.net/perfSONAR/directorySearch.html?match=.au

Calendar | LBL Gmail | Google Docs | Google Code | Google Sites | Gmail | Facebook | NESG | ESnet | LBNL | More | WV

perfSONAR

perfSONAR Global Service and Data View

wser

ject(s):

- All (9395)
- AARNet (50)
- ACORN (16)
- ALICE (100)

ect one or more projects, and hit update.

vice information match string:

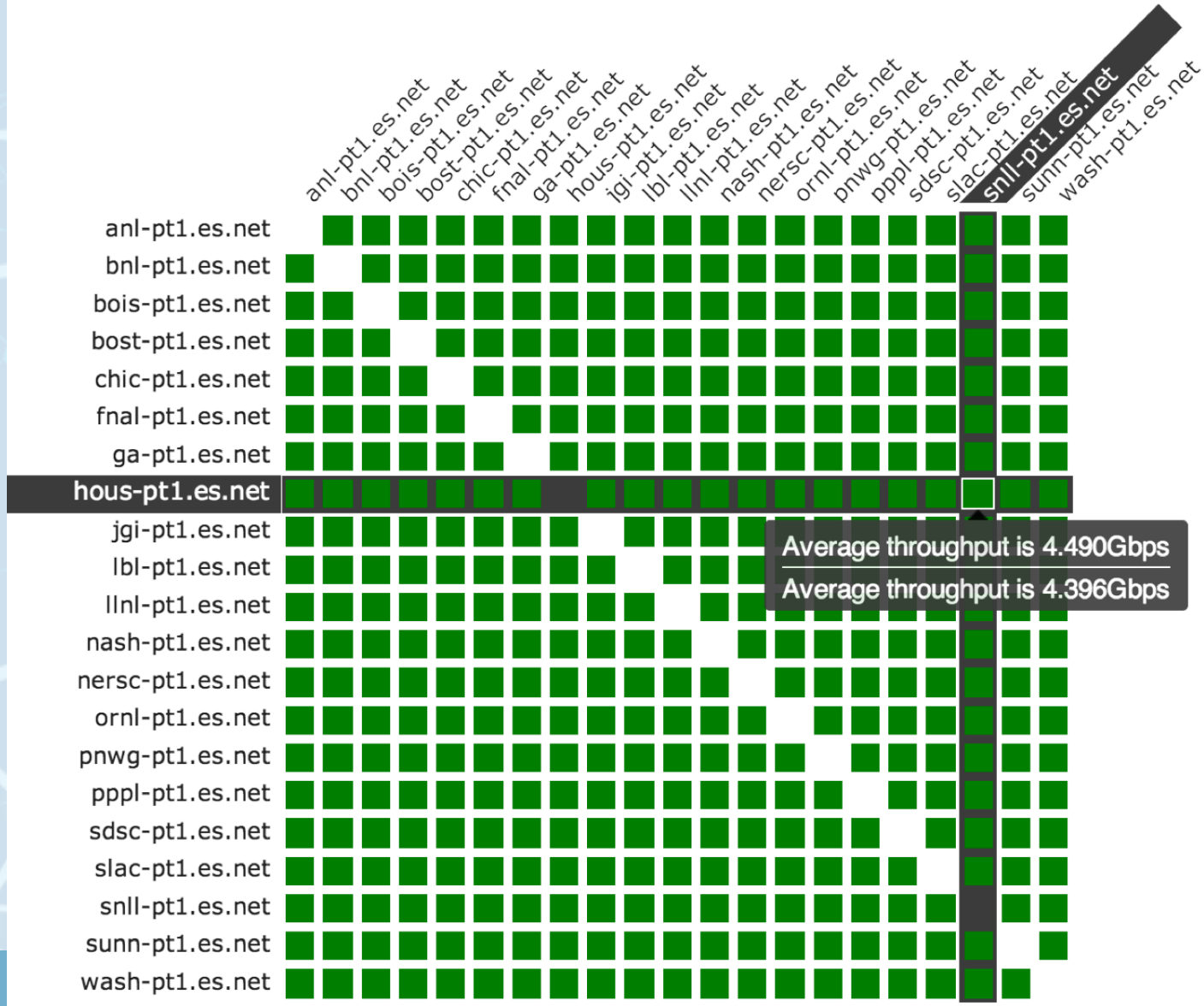
Update

- Measurement Tools (30)
- BWCTL Servers (11)
  - AARNet, drw-a-ps1.aarnet.net.au
  - AARNet, mel-a-ps1.aarnet.net.au
  - AARNet, syd-a-ps1.aarnet.net.au
  - Australia-ATLAS, ps-bandwidth.atlas.uni
  - Australian National University
  - Monash University
  - ORAU, ndt.orau.org
  - RAU, gt-medicionesuy.rau.edu.uy
  - RDSI, 1g.bne-qcif-s-ps1.aarnet.net.au
  - RDSI, bne-qcif-s-ps1.aarnet.net.au
  - SPRACE, perfsonar-bw.sprace.org.br
- NDT Servers (2)
- NPAD Servers (1)
- OWAMP Servers (12)

Service Information

Address	Location	Project(s)	Example Command Line
tcp://drw-a-ps1.aarnet.net.au:4823	at AARNet in AARNet Darwin	TLPW : AARNet	bwctl -T iperf -t 20 -i 1 -f -m -c drw-a-ps1.aarnet.net.au:4823

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



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

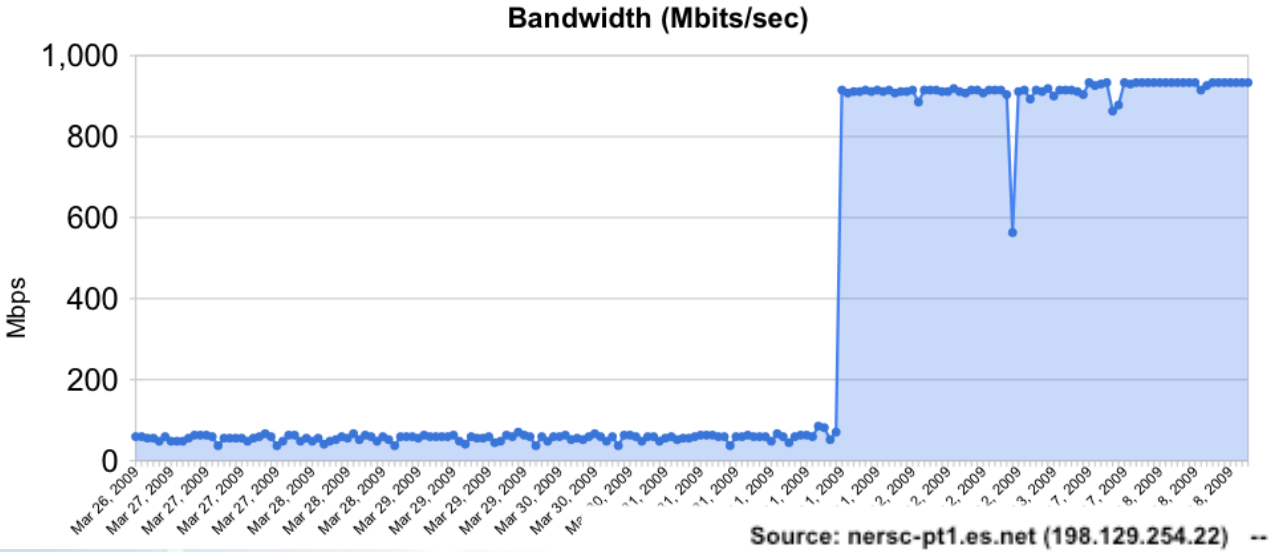
- Failing optics
- Somebody messed around in a patch panel and kinked a fiber
- Hardware goes bad

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

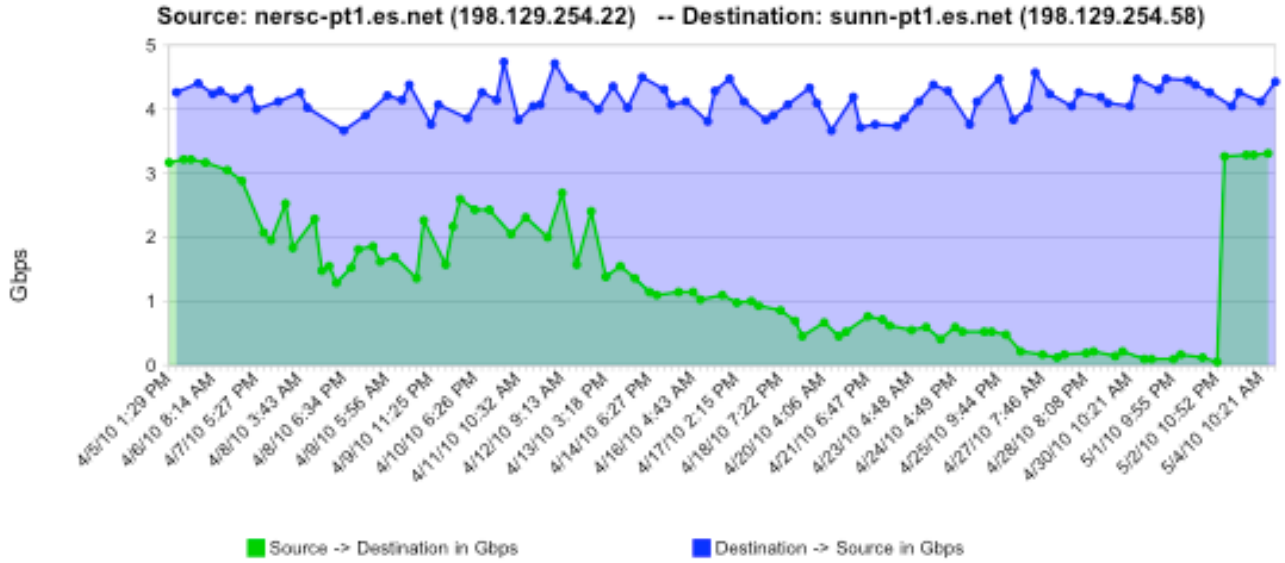
Important to continually collect, archive, and alert on active throughput test results

perfSONAR Results: Sample Soft Failures as seen by perfSONAR



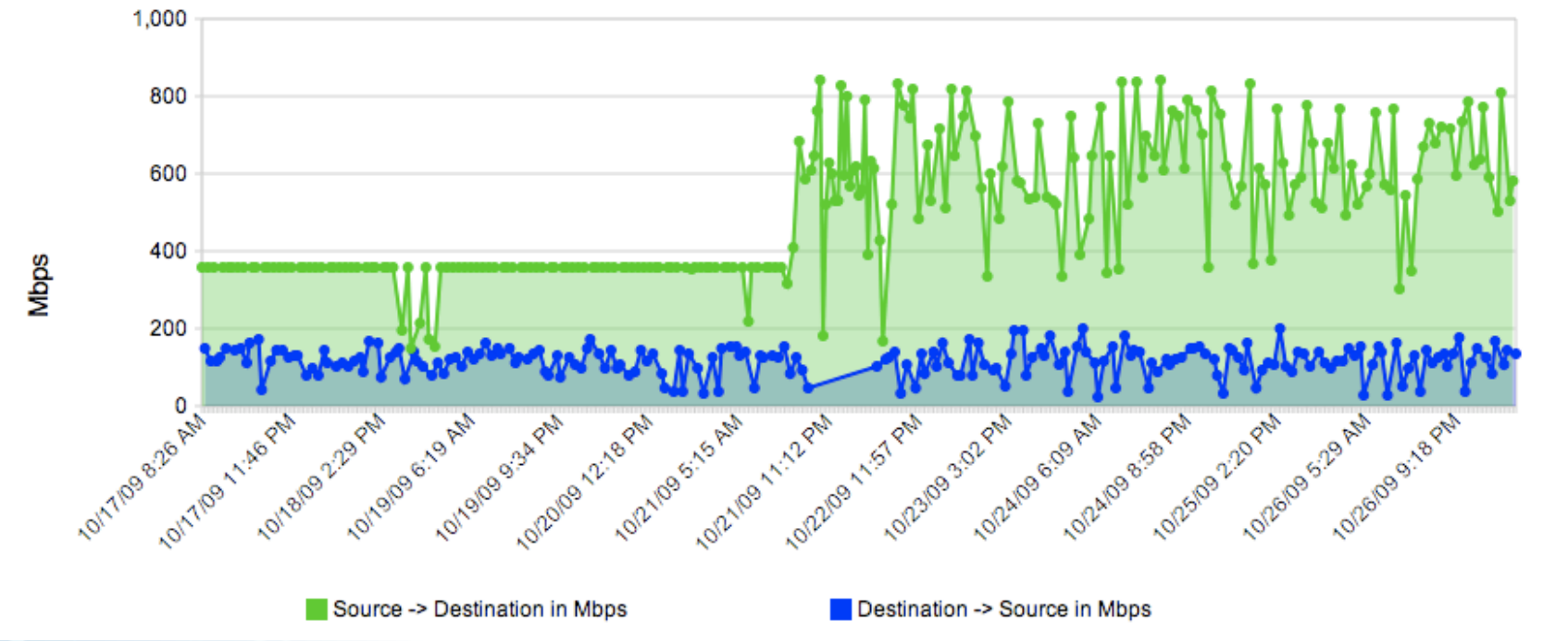
Rebooted router with full route table

Gradual failure of optical line card



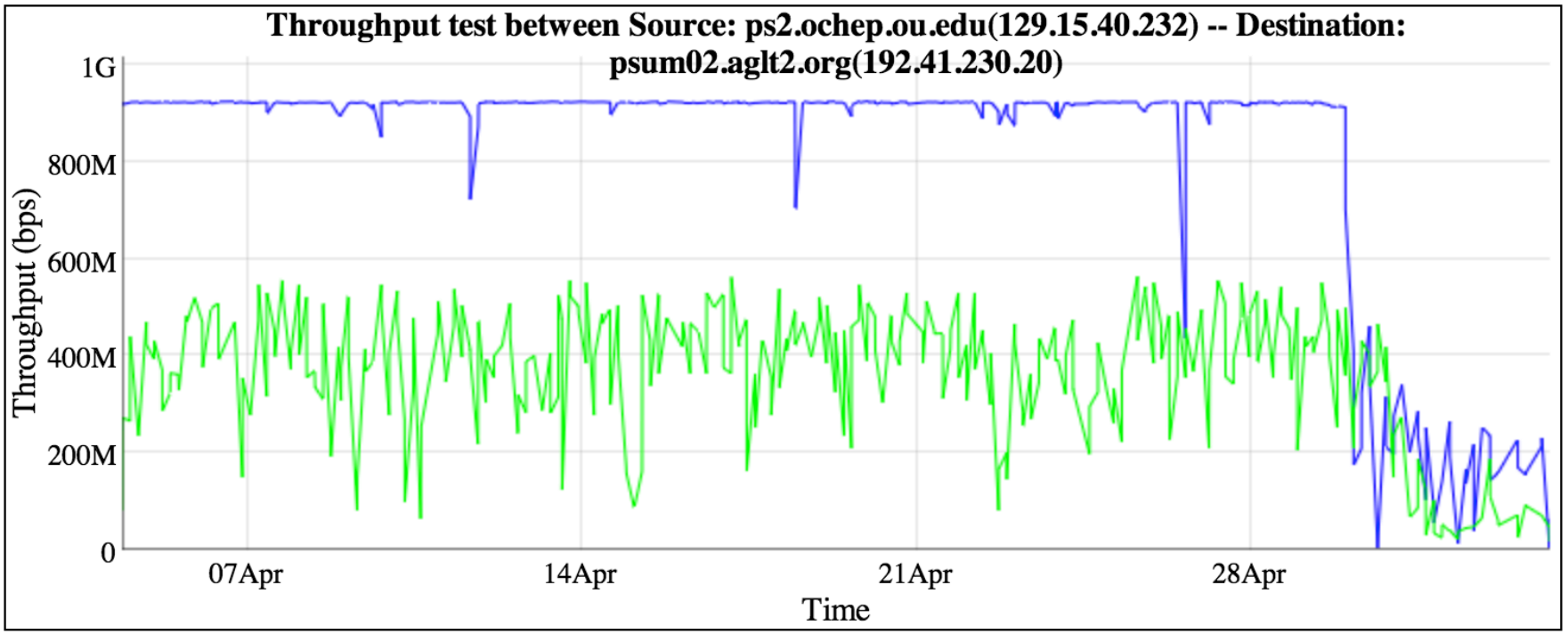
Host Tuning Example

- 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?

Develop a Test Plan

What are you going to measure?

- Achievable bandwidth
  - 2-3 regional destinations
  - 4-8 important collaborators
  - 4-8 times per day to each destination
  - 20 second tests within a region, longer across oceans and continents
- Loss/Availability/Latency
  - OWAMP: ~10 collaborators over diverse paths
- Interface Utilization & Errors (via SNMP)

What are you going to do with the results?

- NAGIOS Alerts
- Reports to user community
- Dashboard

Host Considerations

Dedicated perfSONAR hardware is best

- Server class is a good choice

Other applications will perturb results

Separate hosts for throughput tests and latency/loss tests is preferred

- Throughput tests can cause increased latency and loss
  - Latency tests on a throughput host are still useful however
- Dual-homed host is a possibility, but not currently supported by the Toolkit

1Gbps vs 10Gbps testers

- There are a number of problem that only show up at speeds above 1Gbps

Sample Host configs at:

- http://psps.perfsonar.net/toolkit/hardware.html

VM Considerations

Virtual Machines do not work as well for perfSONAR hosts

- Clock sync issues are a bit of a factor
- Throughput is reduced for 10G hosts
- NDT, SNMP archive, 1G BWCTL are fine on a VM
  - OWAMP, 10G BWCTL are not

Common perfSONAR Use Case

Trouble ticket comes in:

"I'm getting terrible performance from site A to site B"

If there is a perfSONAR node at each site border:

- Run tests between perfSONAR nodes
  - performance is often clean
- Run tests from end hosts to perfSONAR host at site border
  - Often find packet loss (using owamp tool)
  - If not, problem is often the host tuning or the disk
  - If not that, suspect a switch buffer overflow problem
    - These are the hardest to prove

If there is not a perfSONAR node at each site border

- Try to get one deployed
- Run tests to other nearby perfSONAR nodes

WAN Test Methodology – Problem Isolation

Segment-to-segment testing is unlikely to be helpful

- TCP dynamics will be different
- Problem links can test clean over short distances
- An exception to this is hops that go thru a firewall

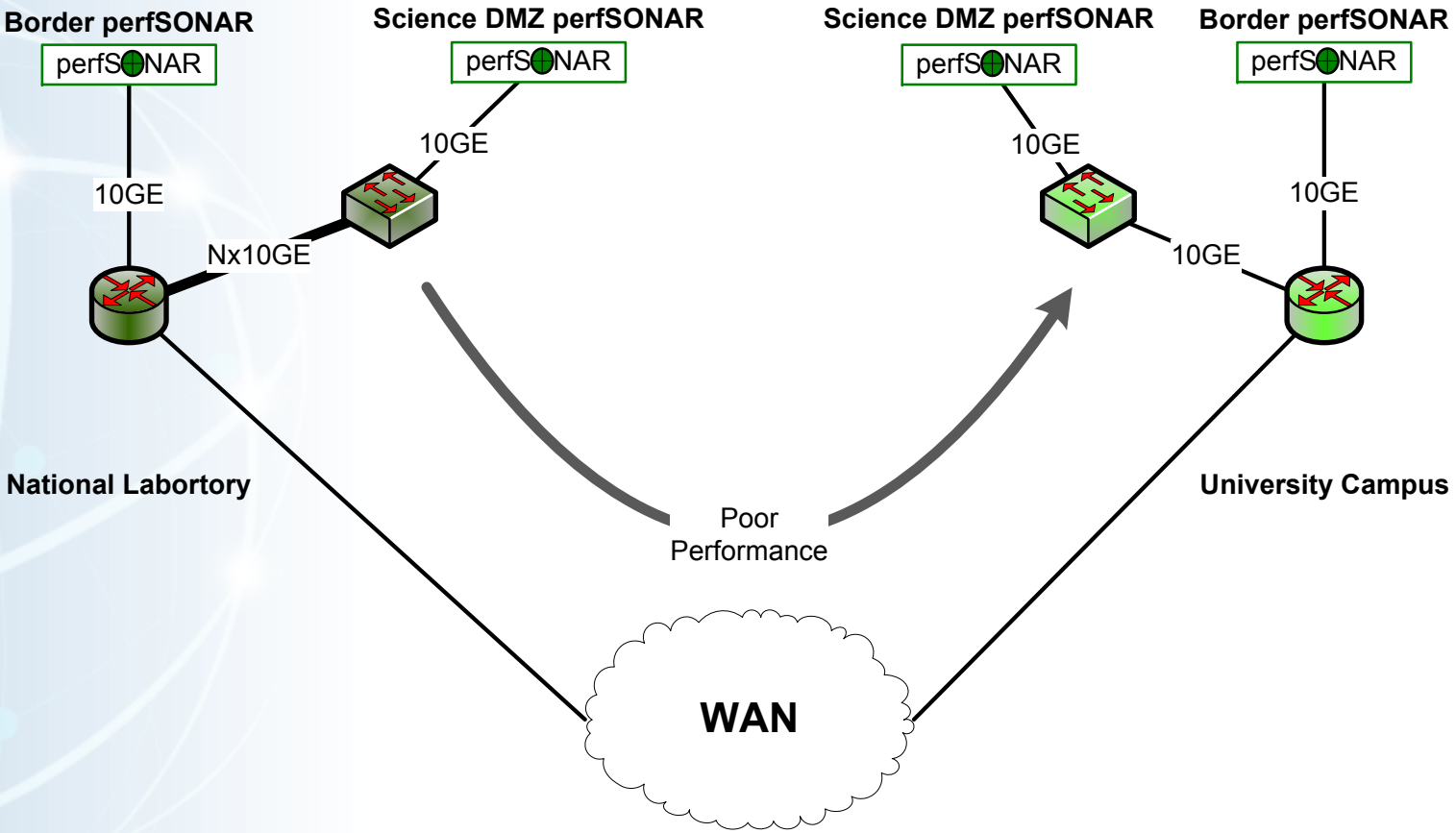
Run long-distance tests

- Run the longest clean test you can, then look for the shortest dirty test that includes the path of the clean test

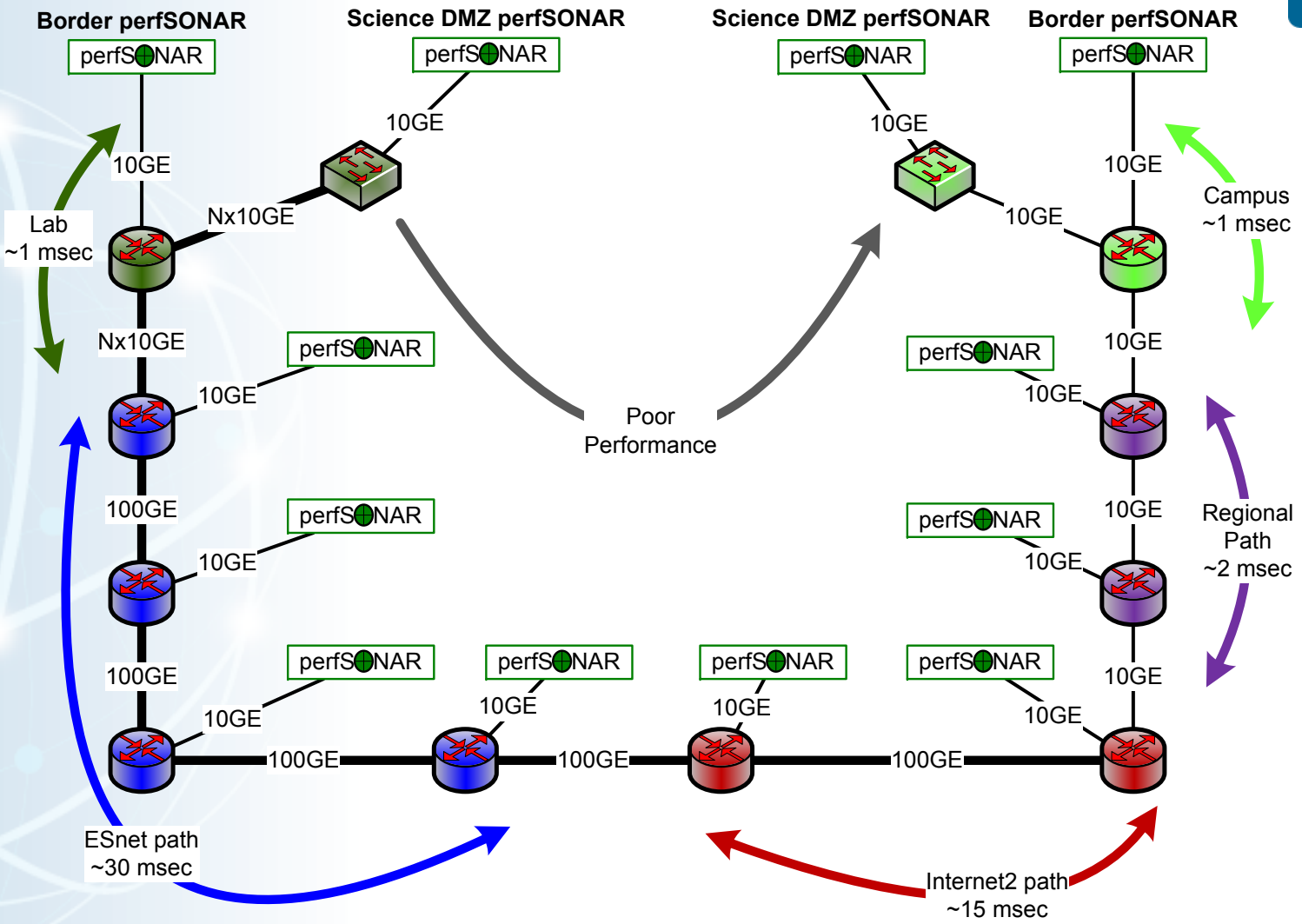
In order for this to work, the testers need to be already deployed when you start troubleshooting

- ESnet has at least one perfSONAR host at each hub location.
  - Many (most?) R&E providers in the world have deployed at least 1
- If your provider does not have perfSONAR deployed ask them why, and then ask when they will have it done

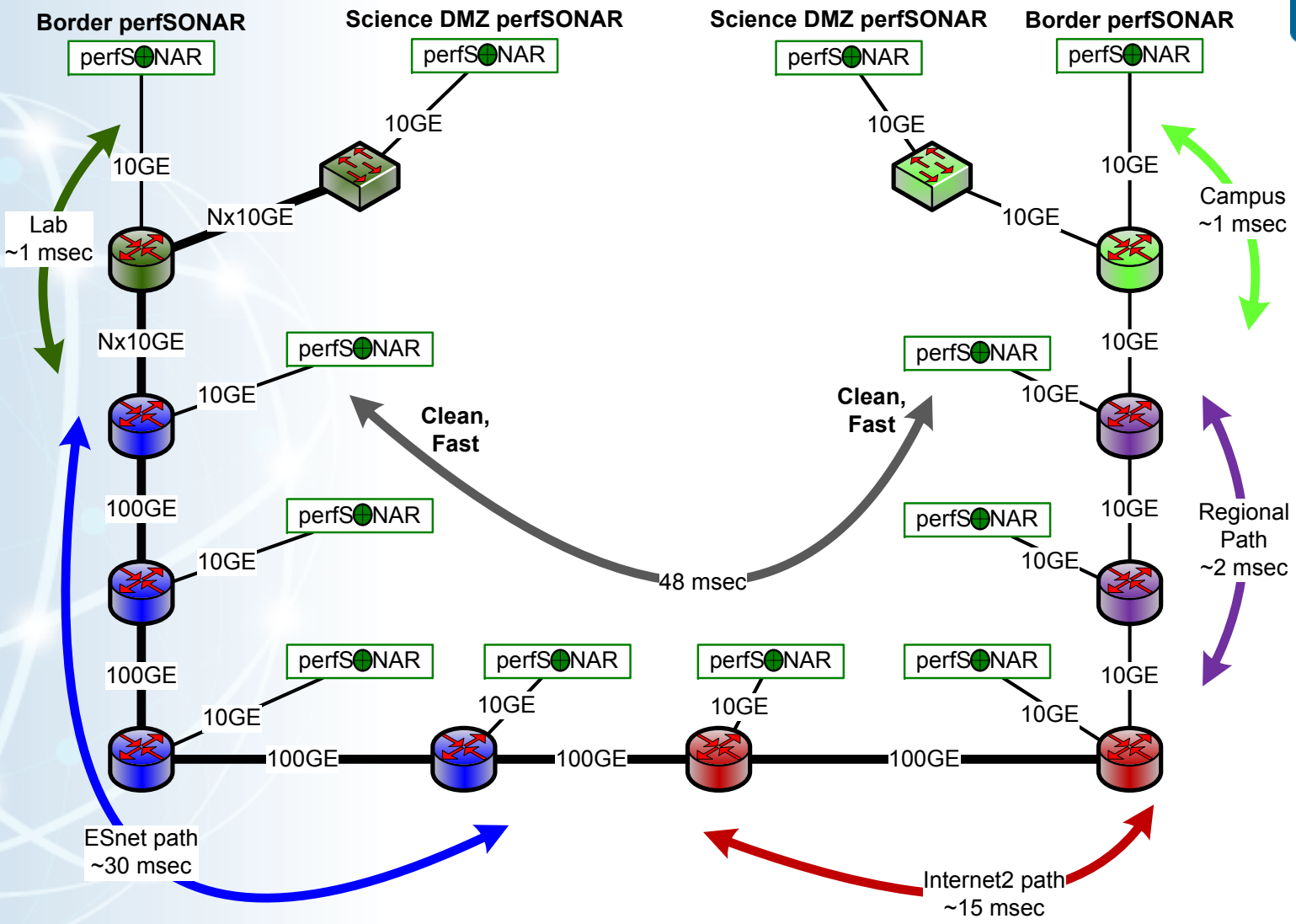
Network Performance Troubleshooting Example



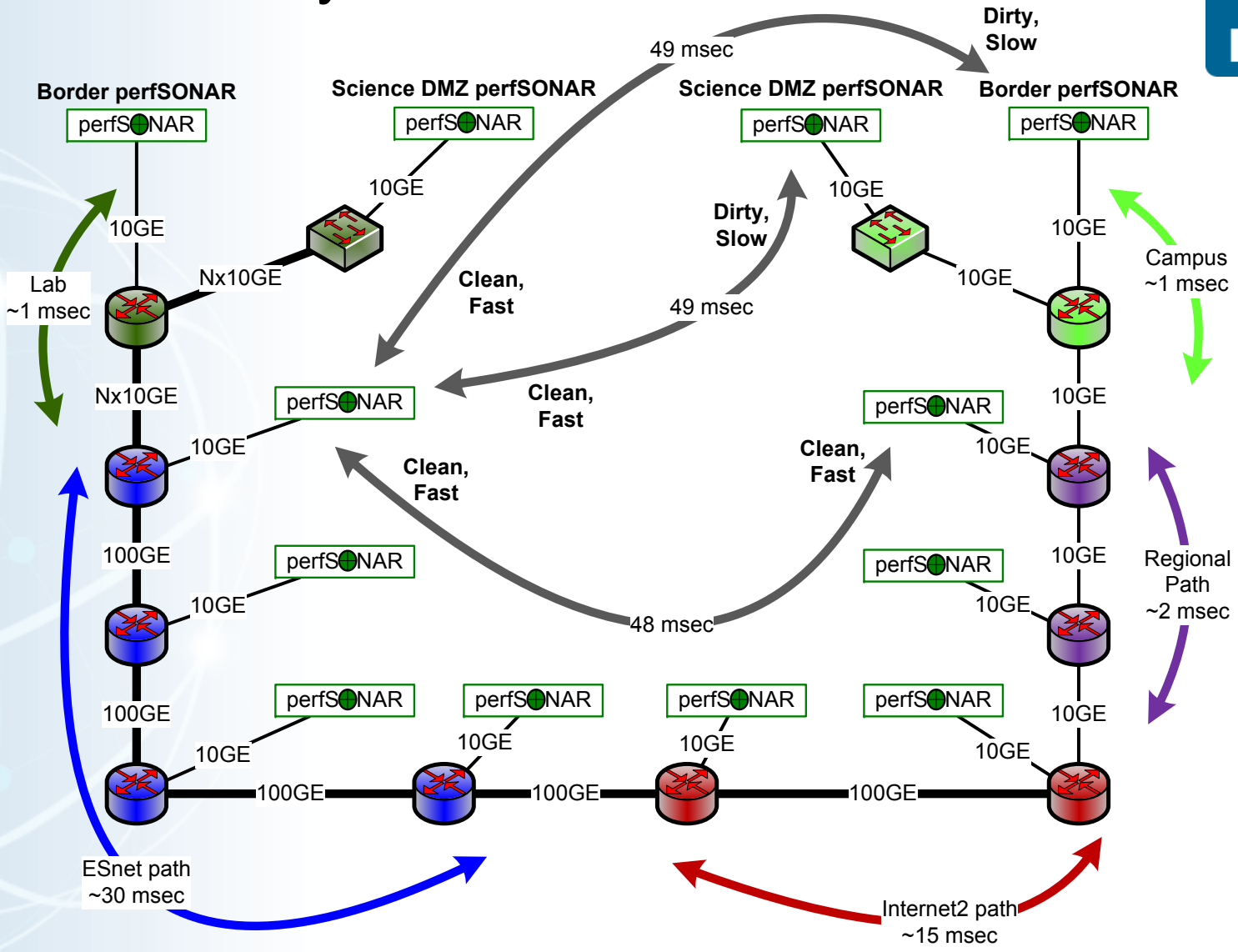
Wide Area Testing – Full Context



Wide Area Testing – Long Clean Test



Wide Area Testing – Poorly Performing Tests Illustrate Likely Problem Areas



Lessons From This Example

This testing can be done quickly if perfSONAR is already deployed

Huge productivity

- Reasonable hypothesis developed quickly
- Probable administrative domain identified
- Testing time can be short – an hour or so at most

Without perfSONAR cases like this are very challenging

Time to resolution measured in months

In order to be useful for data-intensive science, the network must be fixable quickly, because it ***will*** break

The Science DMZ model allows high-performance use of the network, but perfSONAR is necessary to ensure the whole kit functions well

perfSONAR Community

perfSONAR-PS is working to build a strong user community to support the use and development of the software.

perfSONAR-PS Mailing Lists

- Announcement Lists:
  - https://mail.internet2.edu/wws/subrequest/perfsonar-ps-announce
  - https://mail.internet2.edu/wws/subrequest/performance-node-announce
- Users List:
  - https://mail.internet2.edu/wws/subrequest/performance-node-users

More on perfSONAR

http://psps.perfsonar.net/

https://code.google.com/p/perfsonar-ps/

Designing and Building a Data Transfer Node

Eric Pouyoul and Brian Tierney, ESnet

Section Outline

Designing and Building a Data Transfer Nodes

- DTN Hardware Selection
- DTN Tuning

Data Transfer Tools

- Setting expectations
- What makes a fast data transfer tool
- Just say no to scp
- Open Source Tools
- Commercial Tools
- Tool Tuning

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

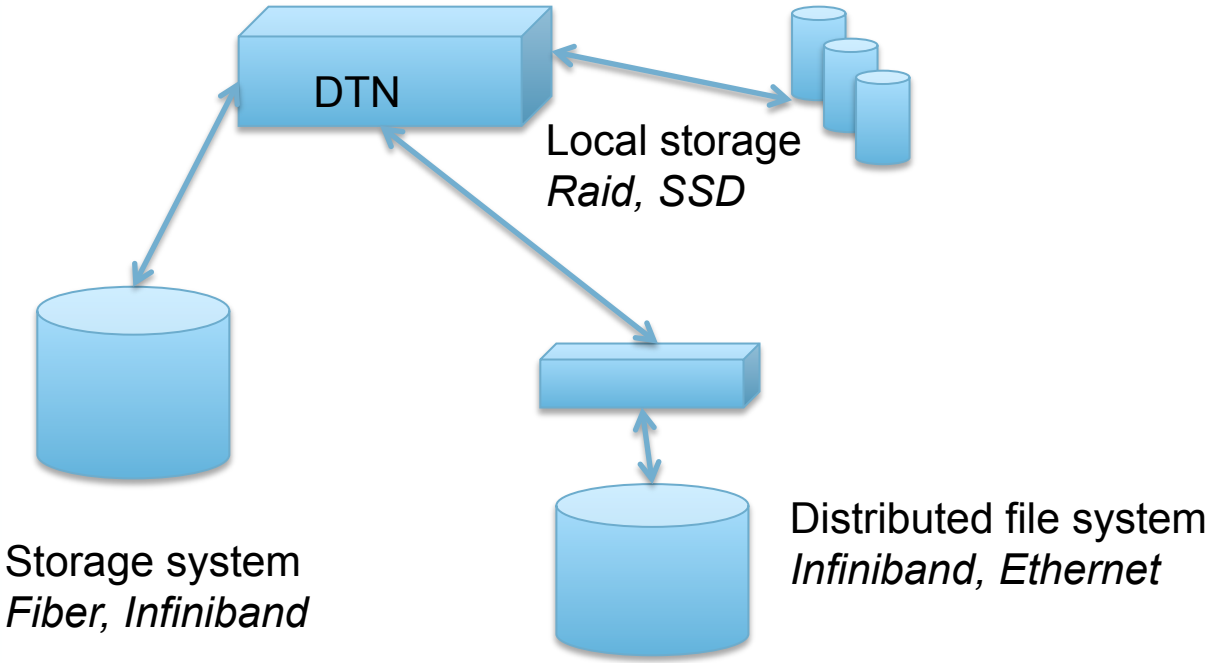
Note: the workflow we are optimizing for here is sequential reads/ write of large files, and a moderate number of high bandwidth flows.

We assume this host is dedicated to data transfer, and not doing data analysis/manipulation

Storage Architectures

There are multiple options for DTN Storage

- this does not really impact DTN node design



DTN Hardware Selection

The typical engineering trade-offs between cost, redundancy, performance, and so forth apply when deciding on what hardware to use for a DTN node

- e.g. redundant power supplies, SATA or SAS disks, cost/ performance for RAID controllers, quality of NICs, etc.

We recommend getting a system that can be expanded to meet future storage/bandwidth requirements.

- science data is growing faster than Moore Law, so might as well get a host that can support 40GE now, even if you only have a 10G network

IMPORTANT: Buying the right hardware is not enough. Extensive tuning is needed for a well optimized DTN.

Motherboard and Chassis selection

Full 40GE requires a PCI Express gen 3 (aka PCIe gen3) motherboard

- this means are limited to a Intel Sandy Bridge or Ivy Bridge host.

Other considerations are memory speed, number of PCI slots, extra cooling, and an adequate power supply.

Intel's Sandy/Ivy Bridge (Ivy Bridge is about 20% faster) CPU architecture provides these features, which benefit a DTN:

- PCIe Gen3 support (up to 16 GB/sec)
- Turbo boost (up to 3.9 Ghz for the i7)
- Faster QPIC for communication between processors

Memory and PCI Slot Considerations

- Memory
  - We recommend 32GB of RAM for a DTN node. More is better.
- PCI Slots
  - Be sure to get the right number of the right type of PCI slots for your needs.
  - PCI slots are defined by:
    - Form factor: This is the length of the slots, referred as the number of PCI lanes it can support. A 16 lane controller is twice as long as a 8 lane controller.
    - Number of wired lanes: not all lanes of the slot may be wired. Some 8 lanes controller may only have 4 lanes wired.
    - PCIe 2.0 is 500 MB/sec per lane. A typical host supports 8 lane (x8) cards, or up to 4 GB/sec. A high-end host might have 16 lane (x16) slots, or up to 8 GB/sec.
    - PCIe 3.0 doubles this bandwidth.

PCI Bus Considerations

Make sure the motherboard you select has the right number of slots with the right number of lanes for you planned usage. For example:

- 10GE NICs require a 8 lane PCIe-2 slot
- 40G/QDR NICs require a 8 lane PCIe-3 slot
- HotLava has a 6x10GE NIC that requires a 16 lane PCIe-2 slot
- Most RAID controllers require 8 lane PCIe-2 slot
- Very high-end RAID controllers for SSD might require a 16 lane PCIe-2 slot or a 8 line PCIe-3 slot
- A high-end Fusion IO ioDrive requires a 16 lane PCIe-2 slot

Some motherboards to look at include the following:

- SuperMicro X9DR3-F
- Sample Dell Server (Poweredge r320-r720)
- Sample HP Server (ProLiant DL380p gen8 High Performance model)

Storage Subsystem Selection

- Deciding what storage to use in your DTN is based on what you are optimizing for:
  - performance, reliability, and capacity, and cost.
- SATA disks historically have been cheaper and higher capacity, while SAS disks typically have been the fastest.
  - However these technologies have been converging, and with SATA 3.1 is less true.

SSD Storage: on its way to becoming mainstream.

SSD

SSD storage costs much more than traditional hard drives (HD), but are much faster. They come in different styles:

- PCIe card: some vendors (Fusion I/O) build PCI cards with SSD.
  - These are the fastest type of SSD: up to several GBytes/sec per card.
    - Note that this type of SSD is typically not hot-swapable.
- HD replacement: several vendors now sell SSD-based drives that have the same form factor as traditional drives such as SAS and SATA.
  - The downside to this approach is that performance is limited by the RAID controller, and not all controllers work well with SSD.
    - Be sure that your RAID controller is “SSD capable”.
- Note that the price of SSD is coming down quickly, so a SSD-based solution may be worth considering for your DTNs.

RAID Controllers

- Often optimized for a given workload, rarely for performance.
- RAID0 is the fastest of all RAID levels but is also the least reliable.
- The performance of the RAID controller is a factor of the number of drives and its own processing engine.

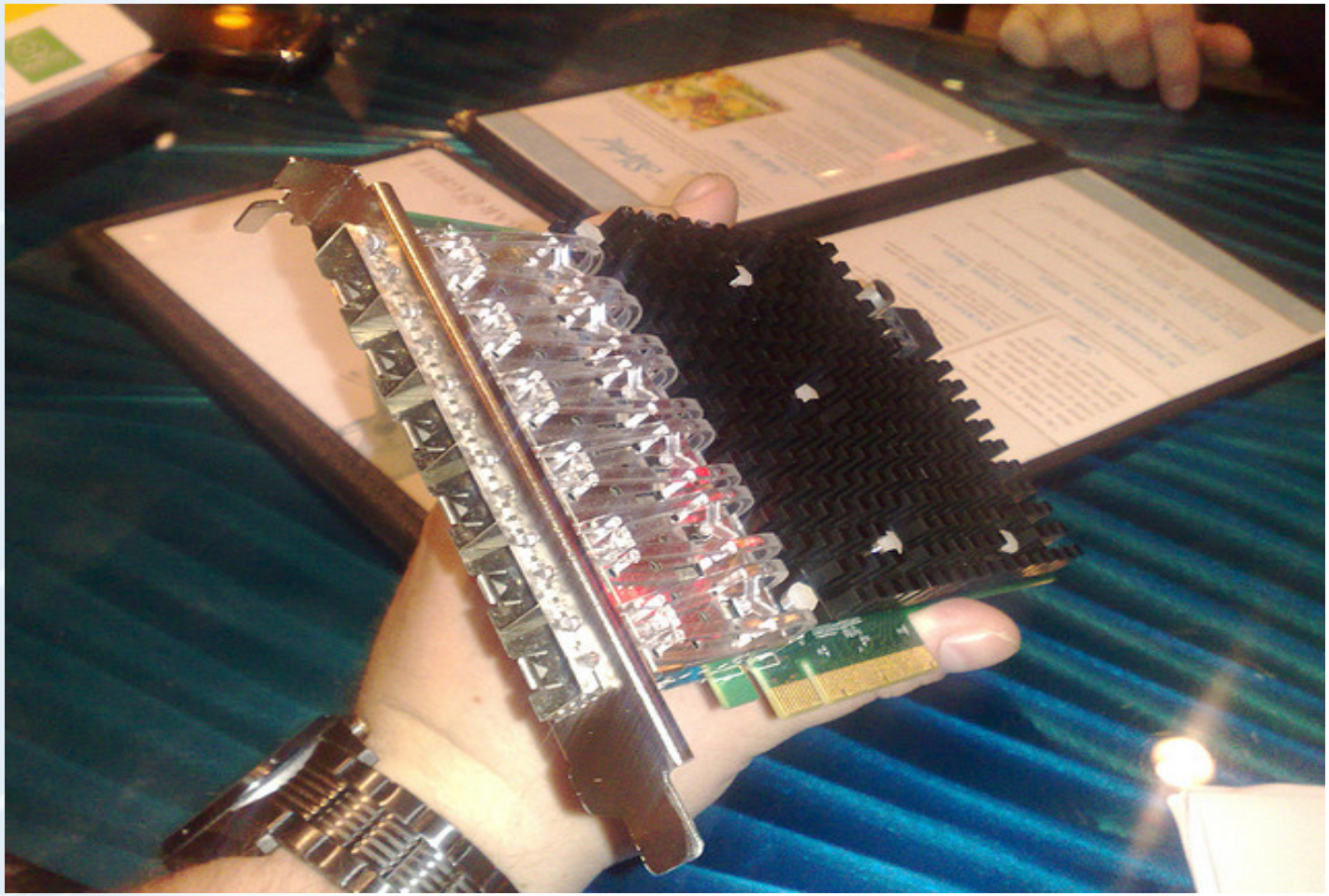
RAID Controller

Be sure your RAID controller has the following:

- 1GB of on-board cache
- PCIe Gen3 support
- dual-core RAID-on-Chip (ROC) processor if you will have more than 8 drives

One example of a RAID card that satisfies these criteria is the Areca ARC-1882i.

Networking Subsystem



Network Subsystem Selection

There is a huge performance difference between cheap and expensive 10G NICs.

- You should not go cheap with the NIC, as a high quality NIC is important for an optimized DTN host.

NIC features to look for include:

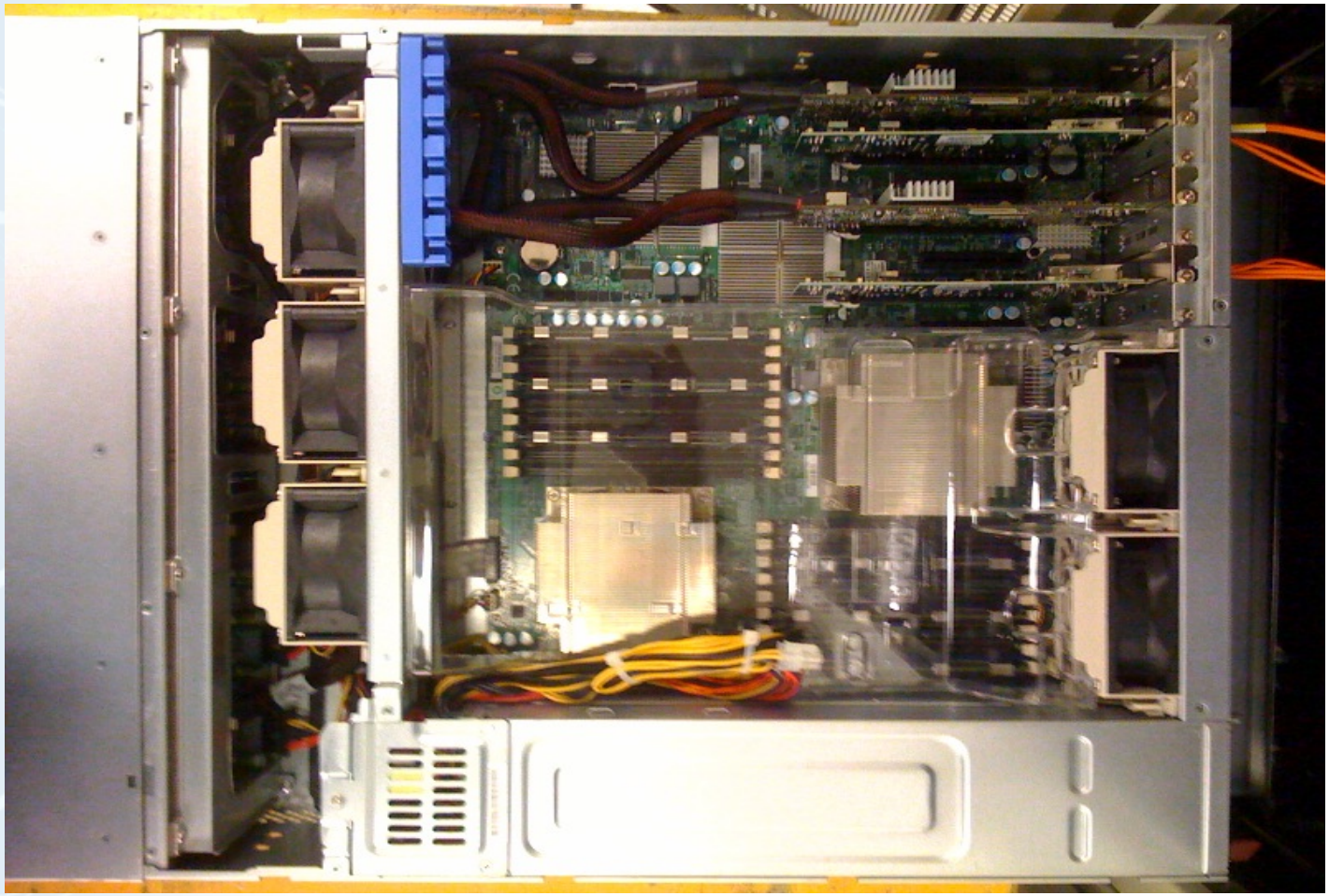
- support for interrupt coalescing
- support for MSI-X
- TCP Offload Engine (TOE)
- support for zero-copy protocols such as RDMA (RoCE or iWARP)

Note that many 10G and 40G NICs come in dual ports, but that does not mean if you use both ports at the same time you get double the performance. Often the second port is meant to be used as a backup port.

- True 2x10G capable cards include the Myricom 10G-PCIE2-8C2-2S and the Mellanox MCX312A-XCBT.

Several vendors now support iWarp, but currently Mellanox is the only vender that supports RoCE in hardware.

Tuning the Data Transfer Host



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

Defaults are usually not appropriate for performance.

What needs to be tuned:

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

DTN Tuning

Tuning your DTN host is extremely important. We have seen overall IO throughput of a DTN more than double with proper tuning.

Tuning can be as much art as a science. Due to differences in hardware, its hard to give concrete advice.

Here are some tuning settings that we have found do make a difference.

- This tutorial assumes you are running a Redhat-based Linux system, but other types of Unix should have similar tuning nobs.

Note that you should always use the most recent version of the OS, as performance optimizations for new hardware are added to every release.

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 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 the Myricom 10G NIC increasing
    interrupt coalencing helps a lot:
/usr/sbin/ethtool -C ethN rx-usecs 75


And use Jumbo Frames!
```

BIOS Tuning

For PCI gen3-based hosts, you should

- enable “turbo boost”,
- disable hyperthreading and node interleaving.

More information on BIOS tuning is described in this document:

http://www.dellhpcsolutions.com/assets/pdfs/Optimal_BIOS_HPC_Dell_12G.v1.0.pdf

I/O Scheduler

The default Linux scheduler is the "fair" scheduler. For a DTN node, we recommend using the "deadline" scheduler instead.

To enable deadline scheduling, add "elevator=deadline" to the end of the "kernel' line in your /boot/grub/grub.conf file, similar to this:

```
kernel /vmlinuz-2.6.35.7 ro root=/dev/VolGroup00/
LogVol00 rhgb quiet elevator=deadline
```

Interrupt Affinity

- Interrupts are triggered by I/O cards (storage, network). High performance means lot of interrupts per seconds
- Interrupt handlers are executed on a core
- Depending on the scheduler, core 0 gets all the interrupts, or interrupts are dispatched in a round-robin fashion among the cores: both are bad for performance:
  - Core 0 get all interrupts: with very fast I/O, the core is overwhelmed and becomes a bottleneck
  - Round-robin dispatch: very likely the core that executes the interrupt handler will not have the code in its L1 cache.
  - Two different I/O channels may end up on the same core.

A simple solution: interrupt binding

- Each interrupt is statically bound to a given core (network -> core 1, disk -> core 2)
- Works well, but can become an headache and does not fully solve the problem: one very fast card can still overwhelm the core.
- Needs to bind application to the same cores for best optimization: what about multi-threaded applications, for which we want one thread = one core ?

Interrupt Binding Considerations

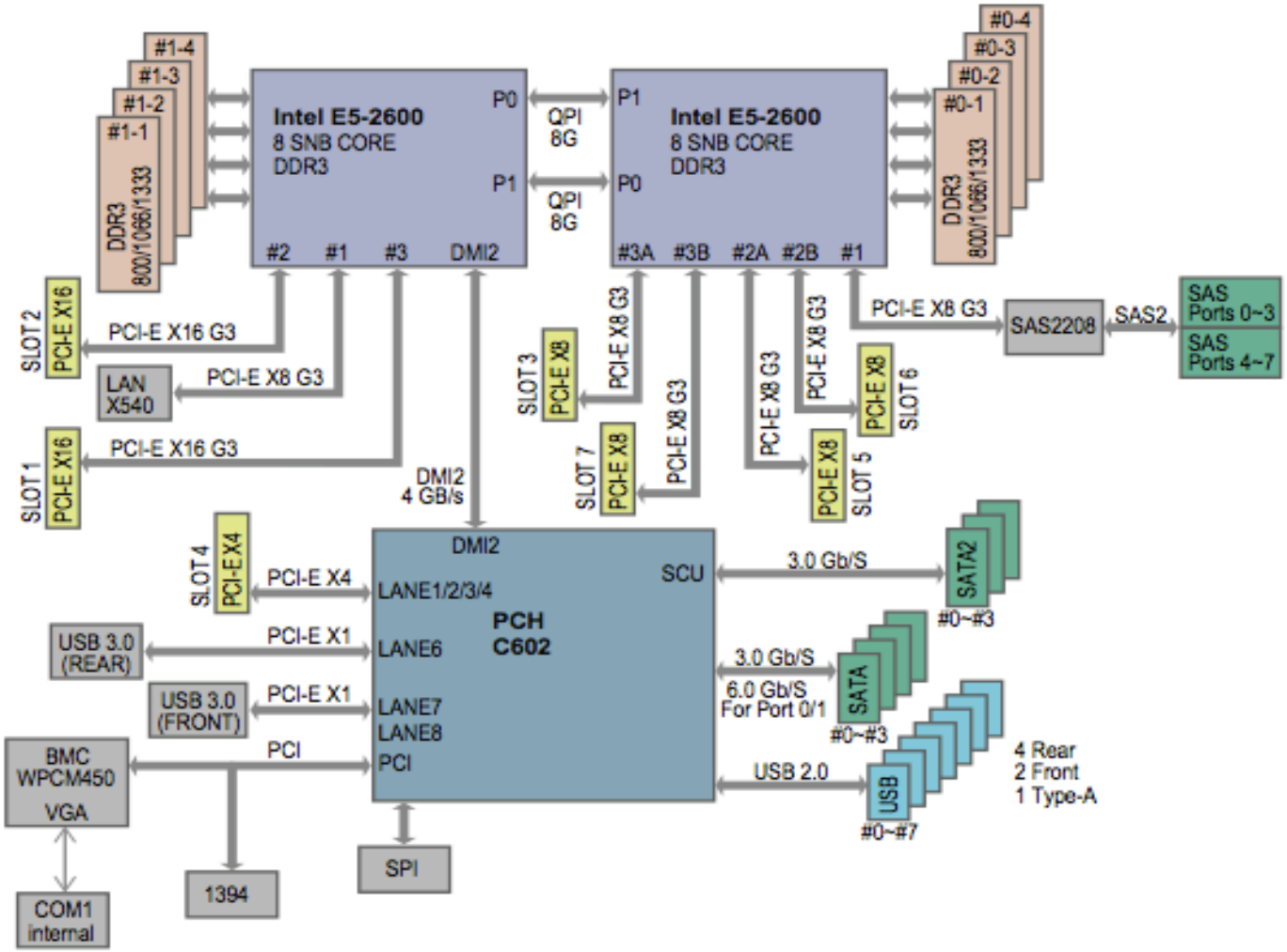
On a multi-processor host, your process might run on one processor, but your I/O interrupts on another processor.

- When this happens there will be a huge performance hit.
  - a single 10G NIC can saturate the QPIC that connects the two processors.
- You may want to consider getting a single CPU system, and avoid dual-processor motherboards if you only need 3 PCIe slots.
- If you need to optimize for a small number of very fast (> 6Gbps) flows, this means you will need to manage IQR bindings by hand.
  - If you are optimizing for many 500Mbps-1Gbps flows, this will be less of an issue.
  - You may be better off doing 4x10GE instead of 1x40GE, as you will have more control mapping IRQs to processors.

The impact of this is even greater with Sandy/Ivy Bridge Hosts, as the PCI bus slots are connected directly to a processor.

Intel Sandy/Ivy Bridge

SUPERMICRO X9DAX-iF/-7F/-iTF/-7TF Motherboard User's Manual



File System Tuning

We recommend using the ext4 file system in Linux for DTN nodes.

Increasing the amount of "readahead" usually helps on DTN nodes where the workflow is mostly sequential reads.

- However you should test this, as some RAID controllers do this already, and changing this may have adverse affects.

Setting readahead should be done at system boot time. For example, add something like this to /etc/rc.local:

```
/sbin/blockdev --setra 262144 /dev/sdb
```

More information on readahead:

http://www.kernel.org/doc/ols/2004/ols2004v2-pages-105-116.pdf

EXT4 Tuning

The file system should be tuned to the physical layout of the drives.

- Stride and stripe-width are used to align the volume according to the stripe-size of the RAID.
  - stride is calculated as Stripe Size / Block Size.
  - stripe-width is calculated as Stride * Number of Disks Providing Capacity.

Disabling journaling will also improve performance, but reduces reliability.

Sample mkfs command:

```
/sbin/mkfs.ext4  /dev/sdb1 –b 4096 –E stride=64 \
stripe-width=768 –O ^has_journal
```

File System Tuning (cont.)

There are also tuning settings that are done at mount time. Here are the ones that we have found improve DTN performance:

- data=writeback
  - this option forces ext4 to use journaling only for metadata. This gives a huge improvement in write performance
- inode_readahead_blks=64
  - this specifies the number of inode blocks to be read ahead by ext4's readahead algorithm. Default is 32.
- Commit=300
  - this parameter tells ext4 to sync its metadata and data every 300s. This reduces the reliability of data writes, but increases performance.
- noatime,nodiratime
  - these parameters tells ext4 not to write the file and directory access timestamps.

Sample fstab entry:

```
/dev/sdb1 /storage/data1 ext4
inode_readahead_blks=64,data=writeback,barrier=0,commit=300,noa
time,nodiratime
```

For more info see: http://kernel.org/doc/Documentation/filesystems/ext4.txt

RAID Controller

Different RAID controllers provide different tuning controls. Check the documentation for your controller and use the settings recommended to optimize for large file reading.

- You will usually want to disable any “smart” controller built-in options, as they are typically designed for different workflows.

Here are some settings that we found increase performance on a 3ware RAID controller. These settings are in the BIOS, and can be entered by pressing Alt+3 when the system boots up.

- Write cache – Enabled
- Read cache – Enabled
- Continue on Error – Disabled
- Drive Queuing – Enabled
- StorSave Profile – Performance
- Auto-verify – Disabled
- Rapid RAID recovery – Disabled

Virtual memory Subsystem

Setting dirty_background_bytes and dirty_bytes improves write performance.

For our system, the settings that gave best performance was:

```
echo 1000000000 > /proc/sys/vm/dirty_bytes
echo 1000000000 > /proc/sys/vm/dirty_background_bytes
```

For more information see:

http://www.kernel.org/doc/Documentation/sysctl/vm.txt

SSD Issues

Tuning your SSD is more about reliability and longevity than performance, as each flash memory cell has a finite lifespan that is determined by the number of "program and erase (P/E)" cycles. Without proper tuning, SSD can die within months.

- **never** do "write" benchmarks on SSD: this will damage your SSD quickly.

Modern SSD drives and modern OSes should all includes TRIM support, which is important to prolong the live of your SSD. Only the newest RAID controllers include TRIM support (late 2012).

Swap

To prolong SSD lifespan, do not swap on an SSD. In Linux you can control this using the sysctl variable vm.swappiness. E.g.: add this to /etc/sysctl.conf:

- vm.swappiness=1

This tells the kernel to avoid unmapping mapped pages whenever possible.

Avoid frequent re-writing of files (for example during compiling code from source), use a ramdisk file system (tmpfs) for /tmp /usr/tmp, etc.

ext4 file system tuning for SSD

These mount flags should be used for SSD partitions.

- noatime: Reading accesses to the file system will no longer result in an update to the atime information associated with the file. This eliminates the need for the system to make writes to the file system for files which are simply being read.
- discard: This enables the benefits of the TRIM command as long for kernel version >=2.6.33.

Sample /etc/fstab entry:

```
/dev/sda1 /home/data ext4 defaults,noatime,discard 0 1
```

For more information see:

http://wiki.archlinux.org/index.php/Solid_State_Drives

Benchmarking

Simulating a single thread writing sequentially a file can be done using dd as:

```
$ dd if=/dev/zero of=/storage/data1/file1 bs=4k
  count=33554432
```

Simulating sequentially reading a file is done by:

```
$ dd if=/storage/data1/file1 of=/dev/null bs=4k
```

Sample Hardware configuration

Motherboard: SuperMicro X9DR3-F

CPU: 2 x Intel(R) Xeon(R) CPU E5-2667 0 @ 2.90GHz

Memory: 64G (8 x 8GiB Kingston DDR3 1600 MHz / 0.6 ns)

RAID: 2 x 3Ware 9750SA-24i (24 ports SAS) connected to CPU 0

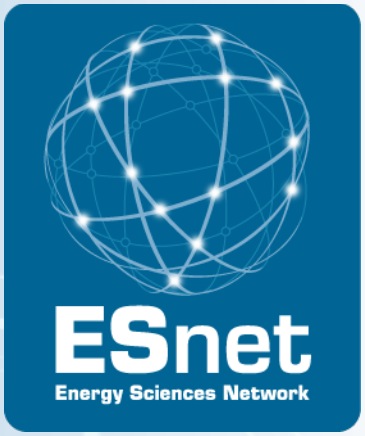
- Areca ARC-1882i should be even better, but we have not tested it

Network Controller: Myricom 10G-PCIE2-8C2-2S

- Use Mellanox cards if you are interested in RDMA

Disks: 16 x Seagate 1TB SAS HDD 7,200 RPM drives

Total cost: $12K in mid 2012



Bulk Data Transfer Tools

Sample Data Transfer Results

Using the right tool is very important

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.

Data Transfer Tools

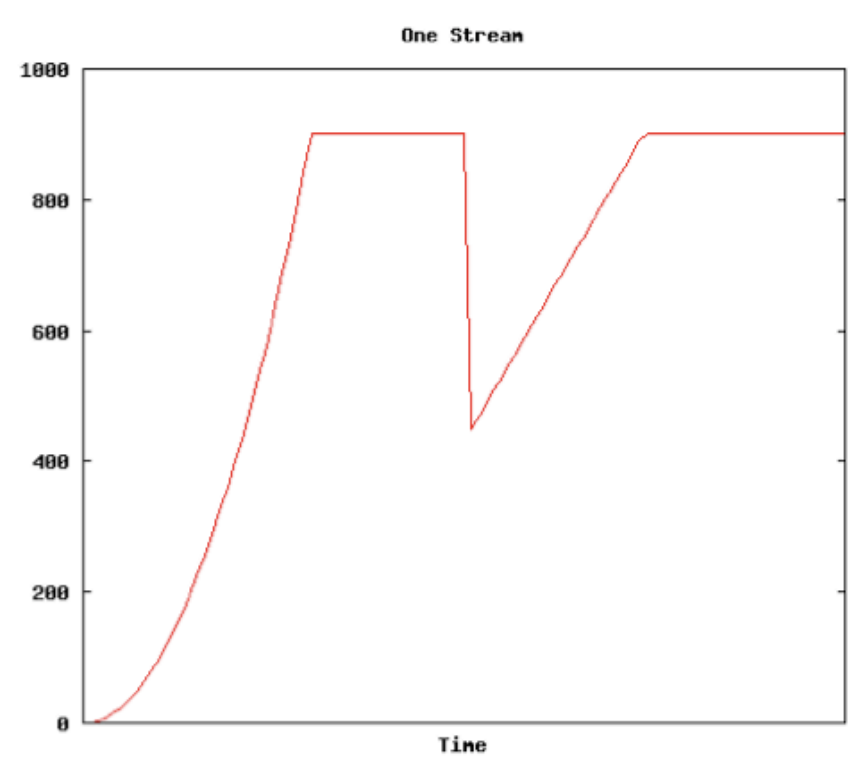
Parallelism is key

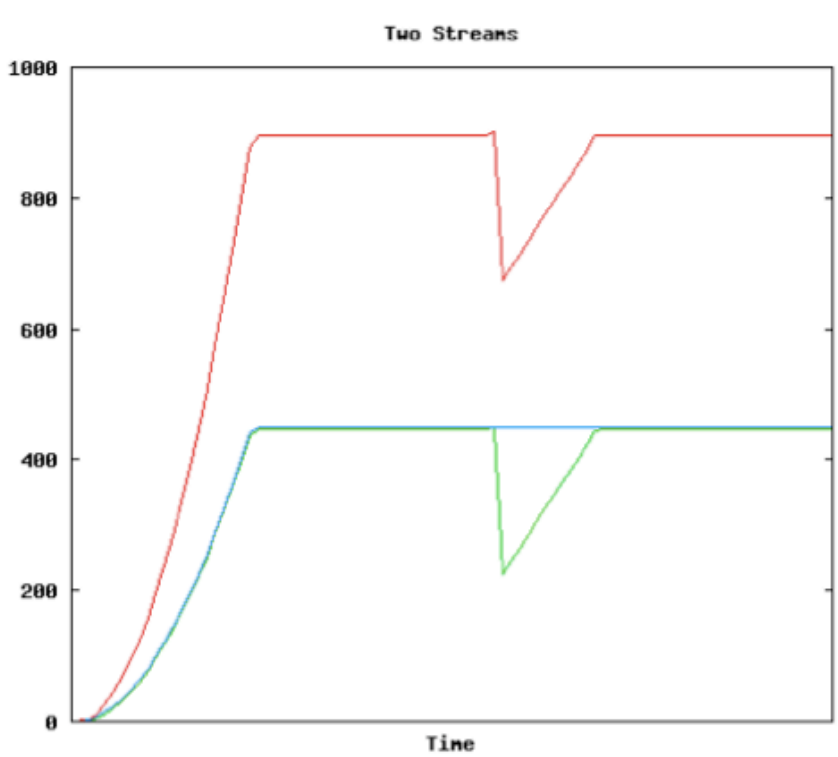
- It is much easier to achieve a given performance level with four parallel connections than one connection
- Several tools offer parallel transfers

Latency interaction is critical

- Wide area data transfers have much higher latency than LAN transfers
- Many tools and protocols assume a LAN
- Examples: SCP/SFTP, HPSS mover protocol

Parallel Streams Help With TCP Congestion Control Recovery Time





Why Not Use SCP or SFTP?

Pros:

- Most scientific systems are accessed via OpenSSH
- SCP/SFTP are therefore installed by default
- Modern CPUs encrypt and decrypt well enough for small to medium scale transfers
- Credentials for system access and credentials for data transfer are the same

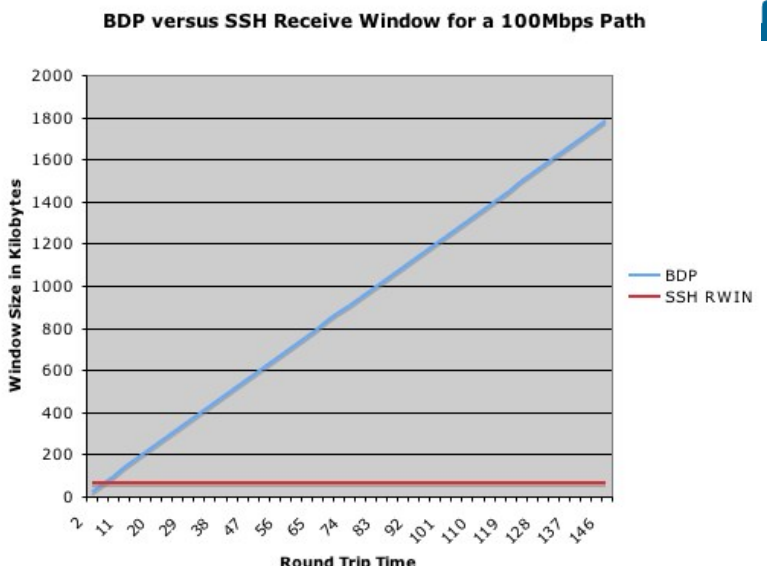
Cons:

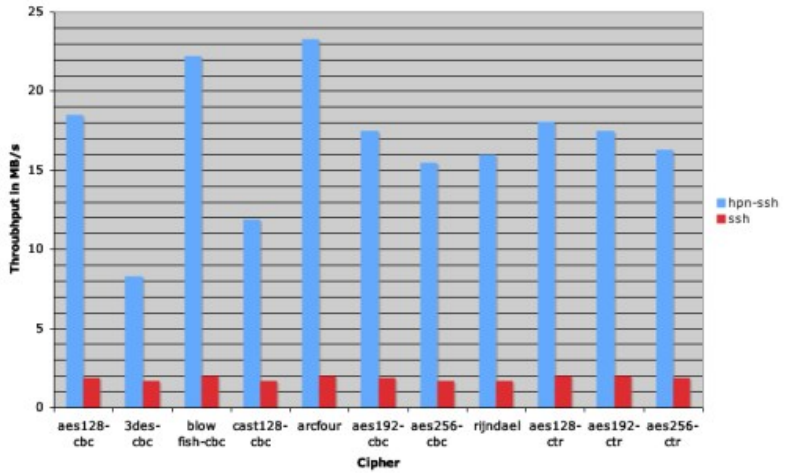
- The protocol used by SCP/SFTP has a fundamental flaw that limits WAN performance
- CPU speed doesn't matter – latency matters
- Fixed-size buffers reduce performance as latency increases
- It doesn't matter how easy it is to use SCP and SFTP – they simply do not perform

Verdict: Do Not Use Without Performance Patches

A Fix For scp/sftp

- PSC has a patch set that fixes problems with SSH
  - http://www.psc.edu/networking/projects/hpn-ssh/
- Significant performance increase
- Advantage – this helps rsync too





sftp

Uses same code as scp, so don't use sftp WAN transfers unless you have installed the HPN patch from PSC

But even with the patch, SFTP has yet another flow control mechanism

- By default, sftp limits the total number of outstanding messages to 16 32KB messages.
- Since each datagram is a distinct message you end up with a 512KB outstanding data limit.
- You can increase both the number of outstanding messages ('-R') and the size of the message ('-B') from the command line though.

Sample command for a 128MB window:

- sftp -R 512 -B 262144 user@host:/path/to/file outfile

FDT

FDT = Fast Data Transfer tool from Caltech

- http://monalisa.cern.ch/FDT/
- Java-based, easy to install
- used by US-CMS project

GridFTP

GridFTP from ANL has features needed to fill the network pipe

- Buffer Tuning
- Parallel Streams

Supports multiple authentication options

- Anonymous
- ssh
- X509

Ability to define a range of data ports

- helpful to get through firewalls

New Partnership with ESnet and Globus Online to support Globus Online for use in Science DMZs

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

What is Globus Online?

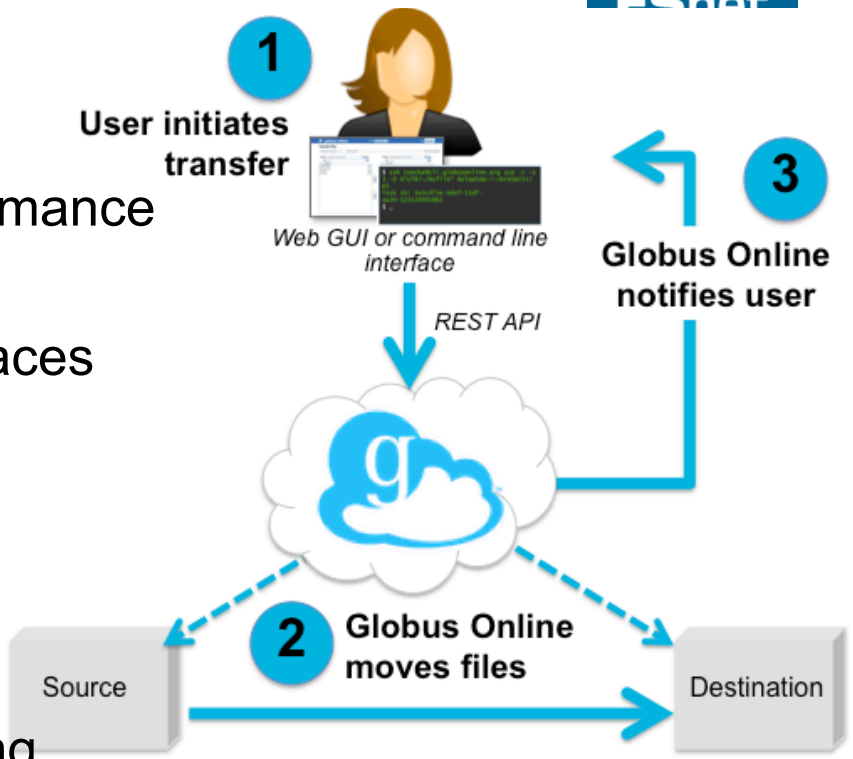
Move, sync files

- Easy “fire-and-forget” transfers
- Automatic fault recovery & high performance
- Across multiple security domains
- Web, command line, and REST interfaces

Minimize IT costs

- Software as a Service (SaaS)
  - No client software installation
  - New features automatically available
- Consolidated support & troubleshooting
- Simple endpoint installation with Globus Connect and GridFTP

Recommended by XSEDE, Blue Waters, NERSC, ALCF, NCAR, ESnet, many Universities



Globus Connect Multi-User

Use Globus Connect Multi-User (GCMU) to:

- Create transfer endpoints in minutes
- Enable multi-user GridFTP access for a resource
- GCMU packages a GridFTP server, MyProxy server and MyProxy Online CA pre-configured for use with Globus Online
  - Simplifies the fairly complex GridFTP server installation process

See: http://www.globusonline.org/gcmu/

Other Data Transfer Tools

bbcp: http://www.slac.stanford.edu/~abh/bbcp/

- supports parallel transfers and socket tuning
- bbcp -P 4 -v -w 2M myfile remotehost:filename

lftp: http://lftp.yar.ru/

- parallel file transfer, socket tuning, HTTP transfers, and more.
- lftp -e 'set net:socket-buffer 4000000; pget -n 4 [http|ftp]://site/path/file; quit'

axel: http://axel.alioth.debian.org/

- simple parallel accelerator for HTTP and FTP.
- axel -n 4 [http|ftp]://site/file

Commercial Data Transfer Tools

There are several commercial UDP-based tools

- Aspera: http://www.asperasoft.com/
- 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

- advantage of these tools less clear on an clean path

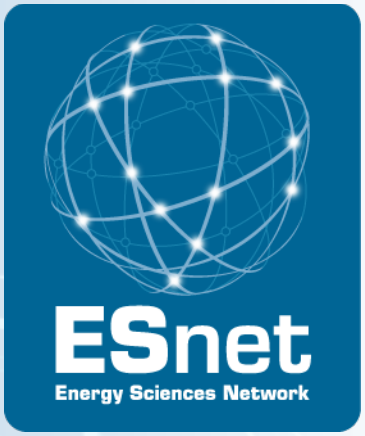
They all have different, fairly complicated pricing models

http://fasterdata.es.net

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

Lots of useful stuff there, including:

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



Advanced Transfer Technologies

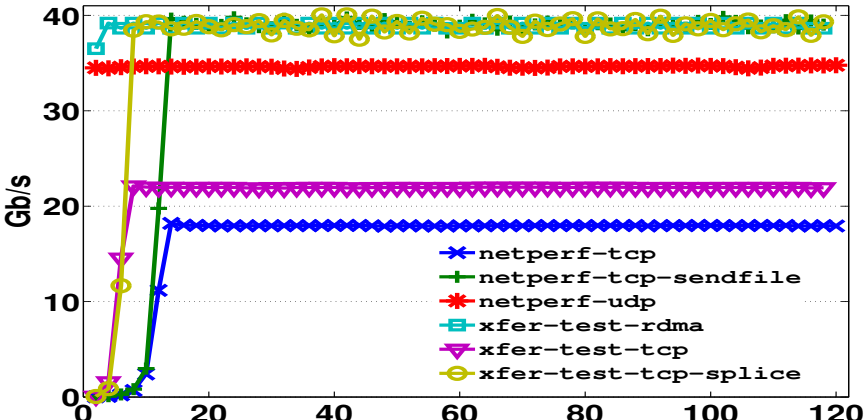
RDMA over Converged Ethernet (RoCE)

RDMA-based tools:

- Several groups have been experimenting with RDMA over the WAN
  - XIO driver for GridFTP (Indiana Univ, Ohio State Univ.)
  - RFTP: Brookhaven National Lab (BNL)
- Over a dedicated layer-2 circuit, performance is the same as TCP, with **much** less CPU
- Requires hardware support on the NIC (e.g.: Mellanox)
  - Software version exists, but requires custom kernel and is slower
- RDMA tuning can be quite tricky to get right

Recent Testbed Results: Single flow 40G

Tool	Protocol	Gbps	Send CPU	Recv CPU
netperf	TCP TCP-sendfile UDP	17.9 39.5 34.7	100% 34% 100%	87% 94% 95%
xfer_test	TCP TCP-splice RoCE	22 39.5 39.2	100% 43% 2%	91% 91% 1%
GridFTP	TCP UDT RoCE	13.3 3.6 13	100% 100% 100%	94% 100% 150%



Just how far can this scale?

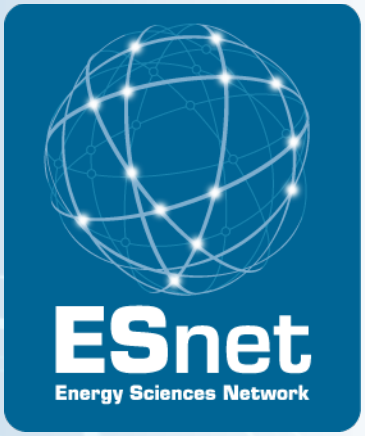
We recently did a demonstration of 2 DTN nodes in Chicago transferring data to 2 DTN nodes in Maastricht, NL (RTT = 115 ms)

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

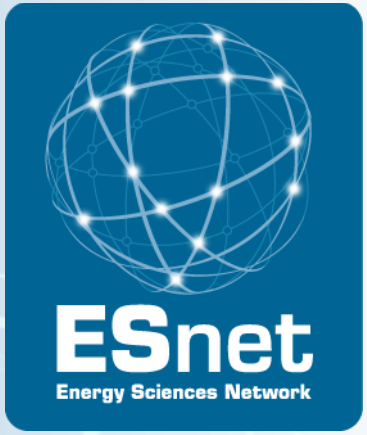
This required all the tuning mentioned above, and a loss-free path.

What this 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.



End of Part 2



The Science DMZ, Part 3

Brian Tierney, Eli Dart, Eric Pouyoul, Jason Zurawski

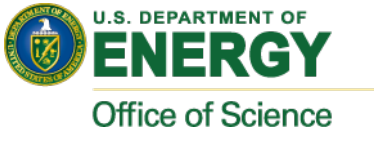
ESnet

Supporting Data-Intensive Research Workshop

QuestNet 2013

Gold Coast, Australia

July 2, 2013



Overview

Part 1:

- What is ESnet?
- Science DMZ Motivation
- Science DMZ Architecture

Part 2:

- PerfSONAR
- The Data Transfer Node
- Data Transfer Tools

Part 3:

- **Science DMZ Security Best Practices**
- **Conclusions**

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

Science DMZ 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 data plane***
- This has nothing to do with whether packet filtering is part of the security enforcement toolkit

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

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

The answer is that firewalls are typically a poor fit for high-performance science applications

The Ubiquitous Firewall

The workhorse device of network security – the firewall – has a poor track record in high-performance contexts

- Firewalls are typically designed to support a large number of users/devices, each with low throughput requirements
  - Data intensive science typically generates a much smaller number of connections that are much higher throughput

Modern firewalls are far more than a packet filter:

- Decode certain application protocols (IDS/IPS functionality, URL filter, etc.)
- Rewrite headers (e.g. NAT)
- VPN Gateway

None of these are relevant to Science DMZ applications

Firewalls vs Router Access Control Lists

When you ask a firewall administrator to allow data transfers through the firewall, what do they ask for?

- IP address of your host
- IP address of the remote host
- Port range
- ***That looks like an ACL to me – I can do that on the router!***

Firewalls make expensive, low-performance ACL filters compared to the ACL capabilities are typically built into the router

Security Without Firewalls

Data intensive science traffic interacts poorly with 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

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

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

Other Security Tools

Intrusion Detection Systems (IDS)

- One example is Bro – http://bro-ids.org/
- Bro is high-performance and battle-tested
  - Bro protects several high-performance national assets
  - Bro can be scaled with clustering: http://www.bro-ids.org/documentation/cluster.html
- Other IDS solutions are available also

Blackhole Routing to block attacks

Netflow, IPFIX, sflow, etc. can provide visibility

Other Security Tools (2)

Aggressive access lists

- More useful with project-specific DTNs
- If the purpose of the DTN is to exchange data with a small set of remote collaborators, the ACL is pretty easy to write
- Large-scale data distribution servers are hard to handle this way (but then, the firewall ruleset for such a service would be pretty open too)

Limitation of the application set

- One of the reasons to limit the application set in the Science DMZ is to make it easier to protect
- Keep unnecessary applications off the DTN (and watch for them anyway using a host IDS – take violations seriously)

Other Security Tools (3)

Using a Host IDS is recommended for hosts in a Science DMZ

There are several open source solutions that have been recommended:

- OSSec: http://www.ossec.net/
- Rkhunter: http://rkhunter.sourceforge.net (rootkit detection + FIM)
- chkrootkit: http://chkrootkit.org/
- Logcheck: http://logcheck.org (log monitoring)
- Fail2ban: http://www.fail2ban.org/wiki/index.php/Main_Page
- denyhosts: http://denyhosts.sourceforge.net/

Using OpenFlow to help secure the Science DMZ

Using OpenFlow to control access to a network-based service seems promising

- E.G.: Sam Russell's work at REANNZ:
  - http://pieknywidok.blogspot.com.au/2013/01/thimble-secure-high-speed-connectivity.html
- This could significantly reduce the attack surface for any authenticated network service

Collaboration Within The Organization

All stakeholders should collaborate on Science DMZ design, policy, and enforcement

The security people have to be on board

- Remember: in some organizations security people already have political cover – it's called the firewall
- If a host gets compromised, the security officer can say they did their due diligence because there was a firewall in place
- If the deployment of a Science DMZ is going to jeopardize the job of the security officer, expect pushback

The Science DMZ is a strategic asset, and should be understood by the strategic thinkers in the organization

- Changes in security models
- Changes in operational models
- Enhanced ability to compete for funding
- Increased institutional capability – greater science output

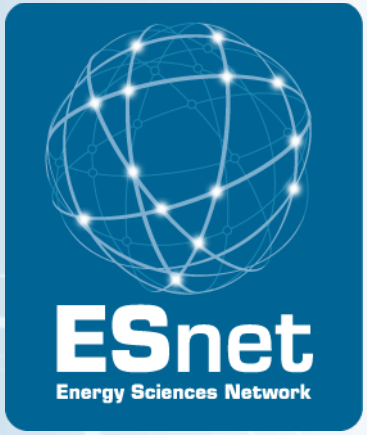
Is it possible to get a firewall that can handle 10G flows?

Yes, but just barely, and it will cost around $500K.

- Will this $500K give you any added security over router ACLs?

10G host interfaces have been around for 10 years, and true 10G firewalls for only a couple years

How long will it take for there to be a true 40G firewall? Or 100G?



Case Study

Sample Security Analysis from the University of Illinois (Nick Buraglio)

How do we security now?

Firewalls

IPS

ACLs

Black hole routing

IDS

Host IDS

SNMP collection

The first 2 (Firewalls and IPS) are the only ones with performance implications. Can we create a secure environment without them?

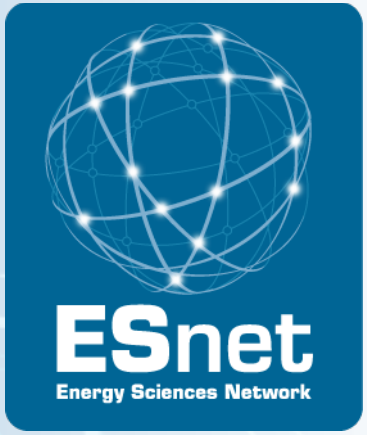
Management and Security Concerns

- "Adding visibility is essential for accountability"
- "Timely mitigation of issues is required"
- "Automated mitigation is highly desirable"*
- "Once you've broken into a DMZ host you have an outpost in enemy territory"

Final Requirements

University of Illinois management, network engineers, and security staff decided on the following for their Science DMZ:

- Flow Data for accountability (netflow/sflow/jflow)
- SNMP collection for baseline creation and capacity planning
- Router ACLs for best practice ingress blocks
- Passive network IDS for monitoring (Bro)
- Host IDS on all hosts outside the firewall (OSSec)
- IDS triggered black hole routing for mitigation
  - Triggers from both network and host IDS
- Bogon (bogus IP address) filtering



Summary and Conclusions

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

perfSONAR

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

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

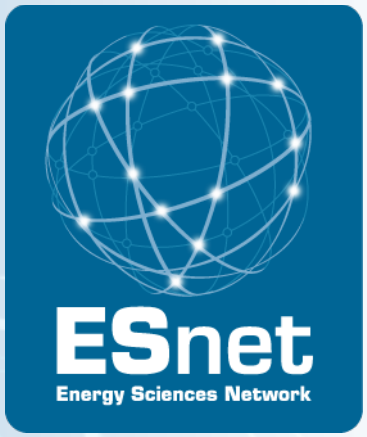
- Shorter time to fix performance problems – less staff effort
- Appropriate implementation of security policy – lower risk
- No need to drag high-speed flows across business network → lower IT infrastructure costs

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?

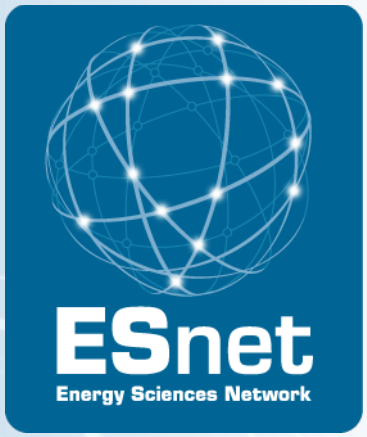
Brian Tierney– bltierney@es.net

http://www.es.net/

http:data.es.net/

These slides available at: http://fasterdata.es.net/science-dmz/learn-more/

Thanks for listening!



Extra Slides

State of the Campus

Show of hands – is there a firewall on your campus?

- Do you know who 'owns' it? Maintains it? Is it being maintained?
- Have you ever asked for a 'port' to be opened? White list a host? Does this involve an email to 'a guy' you happen to know?
- Has it prevented you from being 'productive'?

In General …

- Yes, they exist.
- Someone owns them, and probably knows how to add rules – but the 'maintenance' question is harder to answer.
  - Like a router/switch, they need firmware updates too…
- Will it impact you – 'it depends'. Yes, it will have an effect on your traffic at all times, but will you notice?
  - Small streams (HTTP, Mail, etc.) – you won't notice slowdowns, but you will notice blockages
  - Larger streams (Data movement, Video, Audio) – you will notice slowdowns

The Firewall

The firewall is a useful tool:

- A layer or protection that is based on allowed, and disallowed, behaviors
- One stop location to install instructions (vs. implementing in multiple locations)
- Very necessary for things that need ‘assurance’ (e.g. student records, medical data, protecting the HVAC system, IP Phones, and printers from bad people, etc.)

But, the firewall delivers functionality that can be implemented in different ways:

- Filtering ranges can be implemented via ACLs
- Port/Host blocking can be done on a host by host basis
- IDS tools can implement near real-time blocking of ongoing attacks that match heuristics

The role of Campus Firewalls

I am not here to make you throw away the Firewall

- The firewall has a role; it's time to define what that role ***is***, and ***is not***
- Policy may need to be altered (pull out the quill pens and parchment)
- Minds may need to be changed

I am here to make you think critically about campus security as a system. That requires:

- Knowledge of the risks and mitigation strategies
- Knowing what the components do, and do not do
- Humans to implement and manage certain features – this may be a shock to some (lunch is never free)

When Security and Performance Clash

What does a firewall do?

- Streams of packets enter into an ingress port – there is some buffering
- Packet headers are examined. Have I seen a packet like this before?
  - Yes – If I like it, let it through, if I didn't like it, goodbye.
  - No - Who sent this packet? Are they allowed to send me packets? What port did it come from, and what port does it want to go to?
- Packet makes it through processing and switching fabric to some egress port. Sent on its way to the final destination.

Where are the bottlenecks?

- Ingress buffering – can we tune this? Will it support a 10G flow, let alone multiple 10G flows?
- Processing speed – being able to verify quickly is good. Verifying slowly will make TCP sad
- Switching fabric/egress ports. Not a huge concern, but these can drop packets too
- Is the firewall instrumented to know how well it is doing? Could I ask it?

When Security and Performance Clash

Lets look at two examples, that highlight two primary network architecture use cases:

- Totally protected campus, with a border firewall
  - Central networking maintains the device, and protects all in/ outbound traffic
  - Pro: end of the line customers don't need to worry (as much) about security
  - Con: end of the line customers ***must*** be sent through the disruptive device

- Unprotected campus, protection is the job of network customers
  - Central networking gives you a wire and wishes you best of luck
  - Pro: nothing in the path to disrupt traffic, unless you put it there
  - Con: Security becomes an exercise that is implemented by all end customers