



**EPOC**

Engagement and Performance  
Operations Center

# perfSONAR Topics

Ken Miller, Jason Zurawski

[ken@es.net](mailto:ken@es.net), [zurawski@es.net](mailto:zurawski@es.net)

ESnet / Lawrence Berkeley National Laboratory

***Modern Cyberinfrastructure for Research Data  
Management Workshop  
University of Central Florida  
February 16-17, 2023***

**TACC**  
TEXAS ADVANCED COMPUTING CENTER



**ESnet**  
ENERGY SCIENCES NETWORK

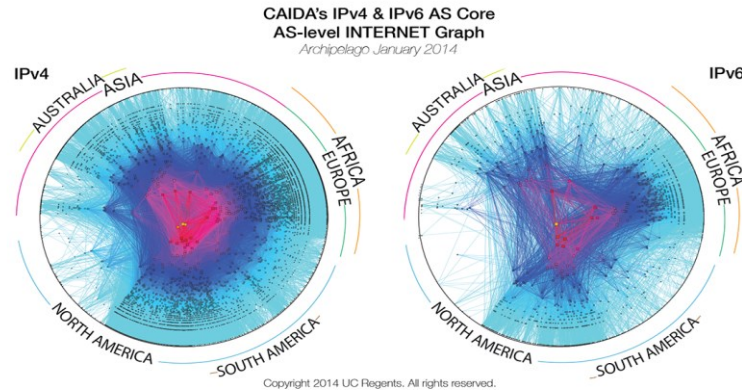
# Outline

- Problem Statement on Network Connectivity
- Supporting Scientific Users
- Network Performance & TCP Behaviors w/ Packet Loss
- What is perfSONAR
- Deployment Overview
- Command's & Syntax
- Examples
- Conclusions



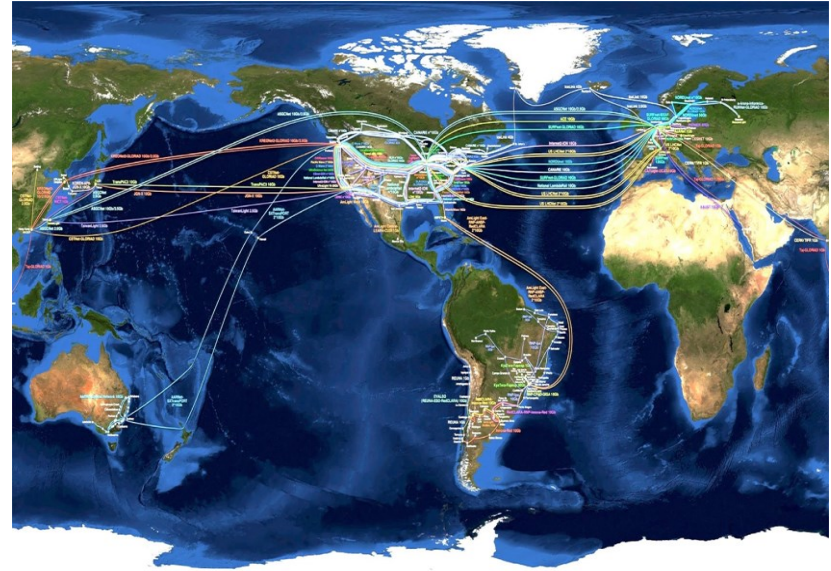
# Problem Statement

- The global Research & Education network ecosystem is comprised of hundreds of international, national, regional and local-scale networks.



# Problem Statement

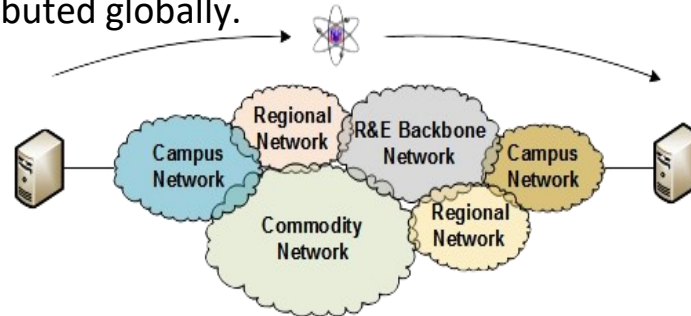
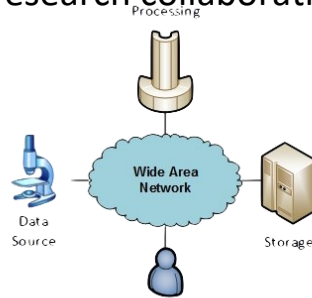
- While these networks all interconnect, each network is owned and operated by separate organizations (called “domains”) with different policies, customers, funding models, hardware, bandwidth and configurations.



©2011 Global Lambda Integrated Facility Visualization by Robert Peterson, NCSA, University of Illinois at Urbana-Champaign Data Compilation by Barrie D. Brown, University of Illinois at Chicago Texture Retouch by Jeff Cooper, NCSA Earth Texture: webdewberry.com www.gli-f.net

# The R&E Community

- The global Research & Education network ecosystem is comprised of hundreds of international, national, regional and local-scale resources – each independently owned and operated.
- This complex, heterogeneous set of networks ***must*** operate seamlessly from “end to end” to support science and research collaborations that are distributed globally.



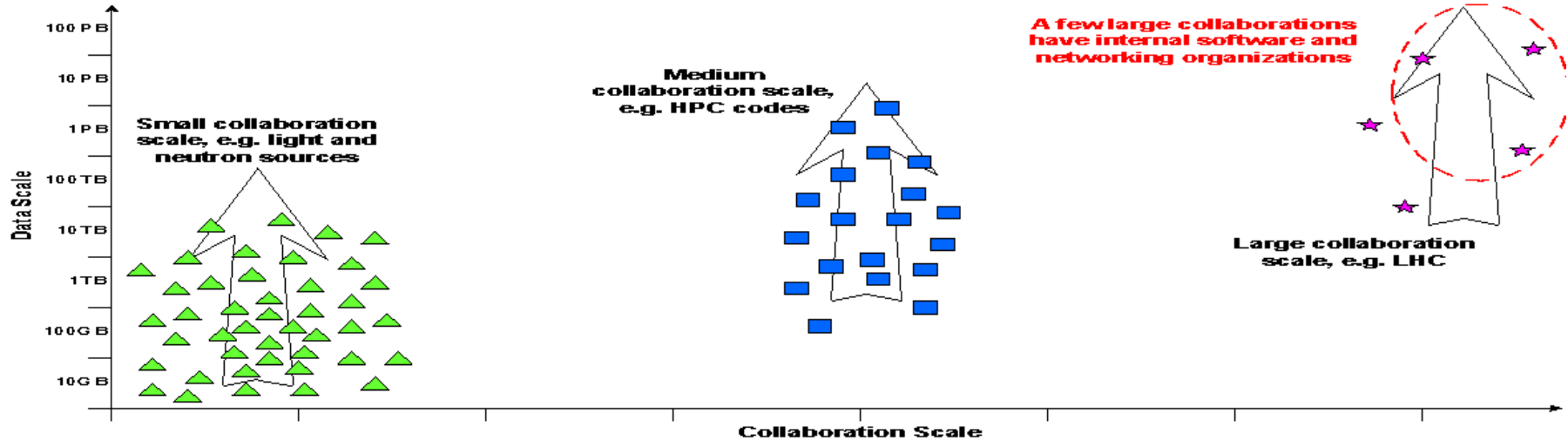
- Data mobility is required; there is no liquid market for HPC resources (people use what they can get – DOE, XSEDE, NOAA, etc. etc.)
  - To stay competitive, we must learn the use patterns, and support them
  - This may mean making sure your network, and the networks of others, are functional

# Outline

- Problem Statement on Network Connectivity
- **Supporting Scientific Users**
- Network Performance & TCP Behaviors w/ Packet Loss
- What is perfSONAR
- Deployment Overview
- Command's & Syntax
- Examples
- Conclusions



# Understanding Data Trends

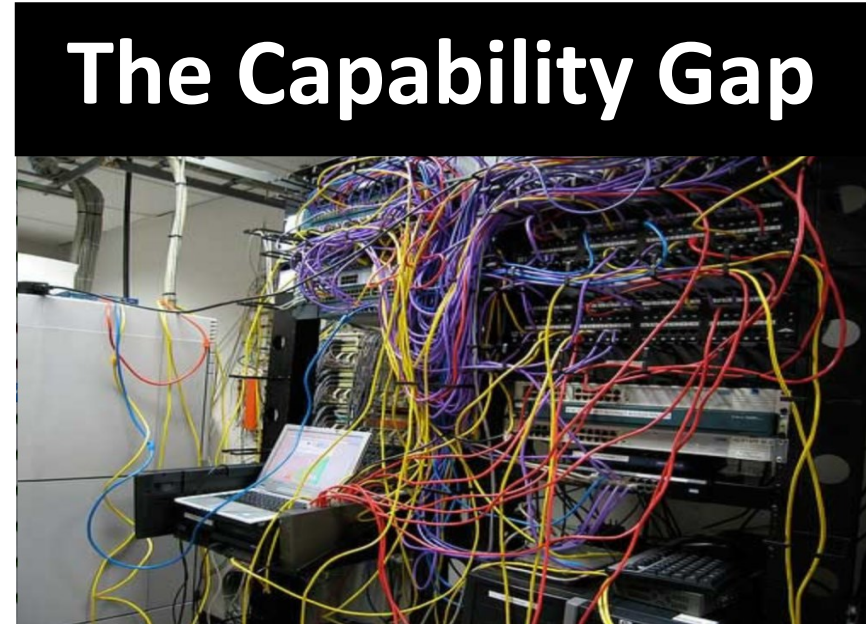


<http://www.es.net/science-engagement/science-requirements-reviews/>



# Challenges to Network Adoption

- Causes of performance issues are complicated for users.
- Lack of communication and collaboration between the CIO's office and researchers on campus.
- Lack of IT expertise within a science collaboration or experimental facility
- User's performance expectations are low ("The network is too slow", "I tried it and it didn't work").
- Cultural change is hard ("we've always shipped disks!").
- Scientists want to do science not IT support





# Outline

- Problem Statement on Network Connectivity
- Supporting Scientific Users
- **Network Performance & TCP Behaviors w/ Packet Loss**
- What is perfSONAR
- Deployment Overview
- Command's & Syntax
- Examples
- Conclusions



# Let's Talk Performance ...

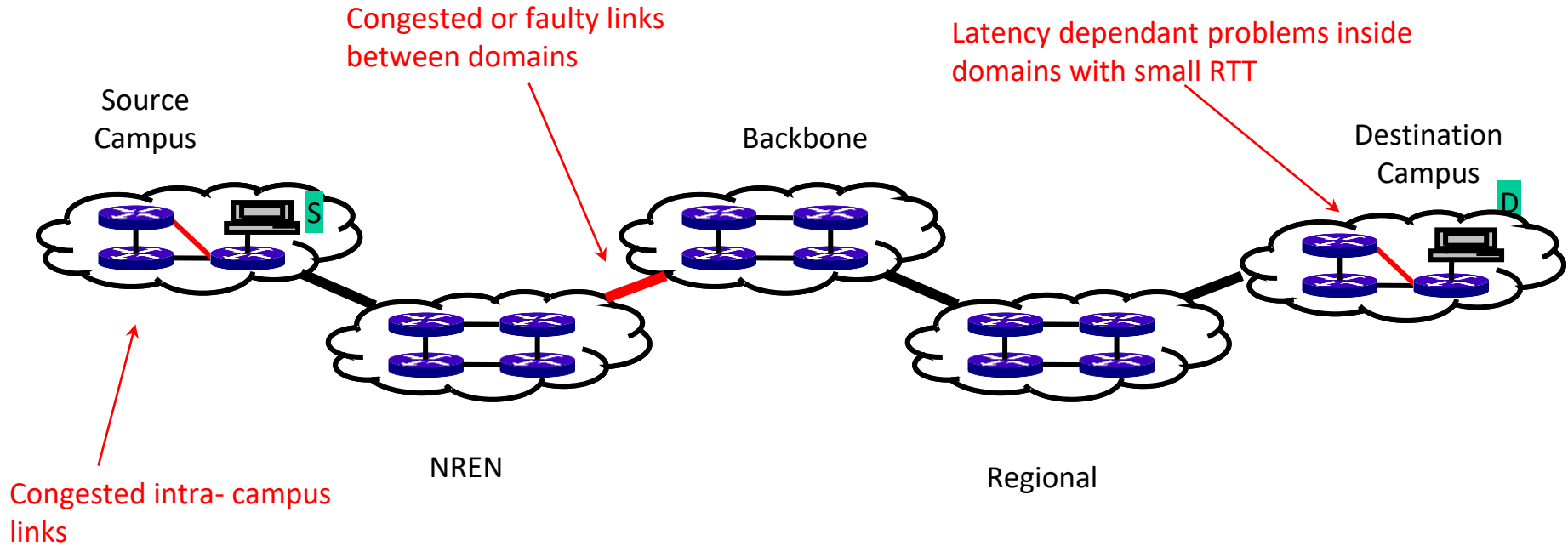
"In any large system, there is always something broken."

- *Jon Postel*

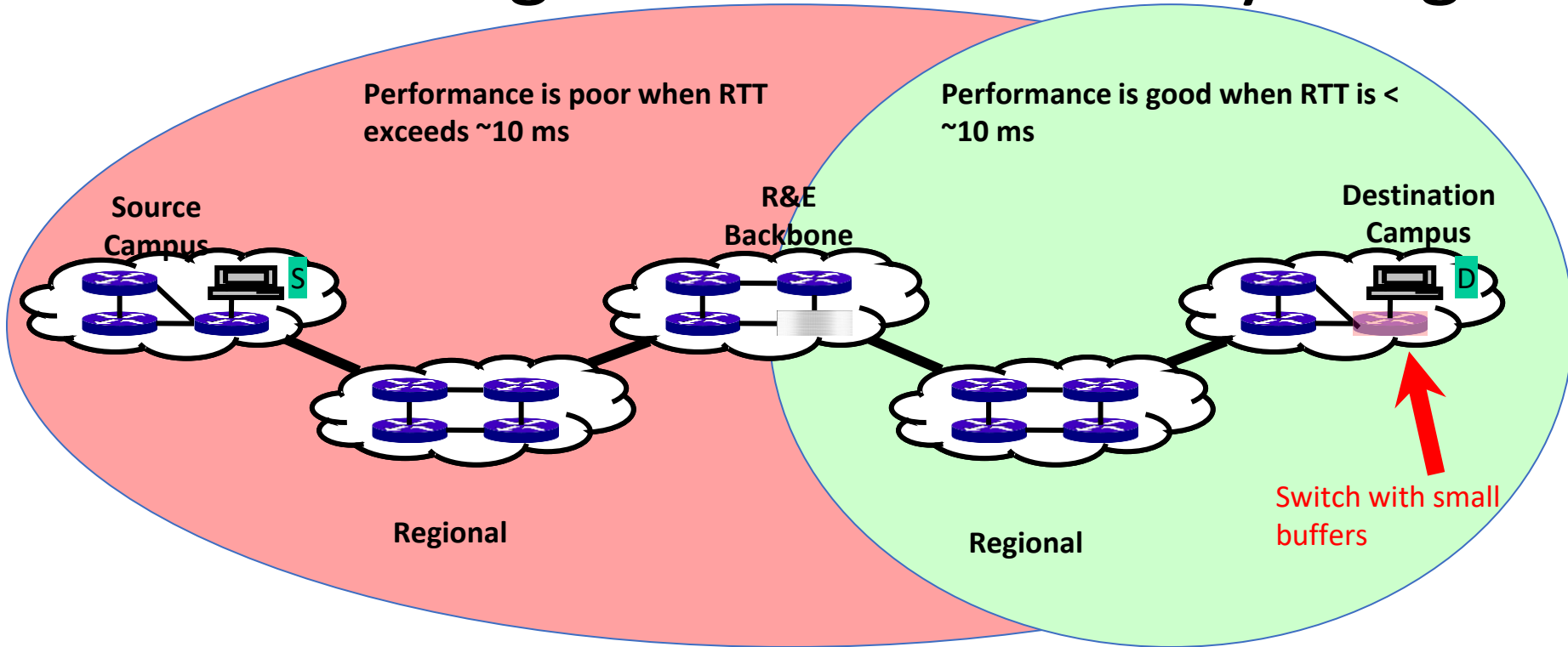
- Modern networks are occasionally designed to be *one-size-fits-most*
  - e.g. if you have ever heard the phrase "converged network", the design is to facilitate CIA (Confidentiality, Integrity, Availability)
- It's all TCP
  - Bulk data movement is a common thread (move the data from the microscope, to the storage, to the processing, to the people – and they are all sitting in different facilities)
  - This fails when TCP suffers due to path problems (ANYWHERE in the path)
  - It's easier to work with TCP than to fix it (20+ years of trying...)
- TCP suffers the most from unpredictability; Packet loss/delays are the enemy
  - Small buffers on the network gear and hosts
  - Incorrect application choice
  - Packet disruption caused by overzealous security
  - Congestion from herds of mice
- It all starts with knowing your users, and knowing your network



# Where Are The Problems?

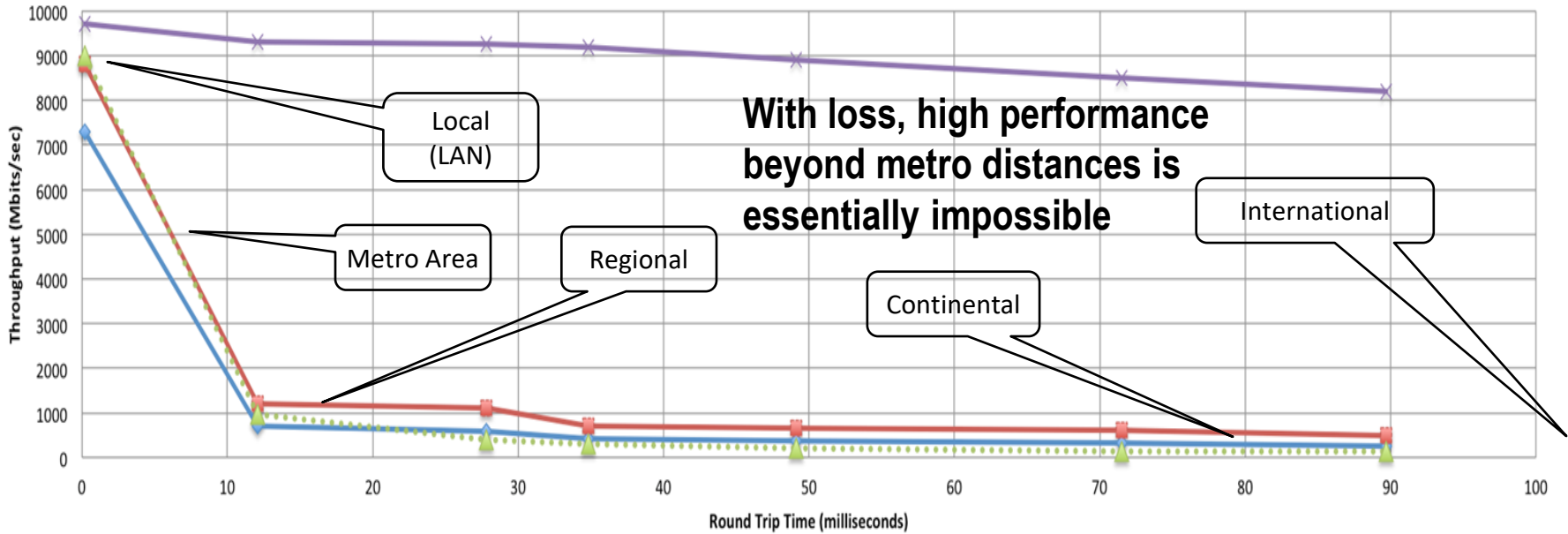


# Local Testing Will Not Find Everything



# Soft Failures Cause Packet Loss and Degraded TCP Performance

Throughput vs. Increasing Latency with .0046% Packet Loss



With loss, high performance beyond metro distances is essentially impossible

Measured (TCP Reno)

Measured (HTCP)

Theoretical (TCP Reno)

Measured (no loss)



# Soft Network Failures

- **Soft failures are where basic connectivity functions, but high performance is not possible.**
- **TCP was intentionally designed to hide all transmission errors from the user:**
  - “As long as the TCPs continue to function properly and the internet system does not become completely partitioned, no transmission errors will affect the users.” (From IEN 129, RFC 716)
- **Some soft failures only affect high bandwidth long RTT flows.**
- **Hard failures are easy to detect & fix**
  - **soft failures can lie hidden for years!**
- **One network problem can often mask others**



# Problem Statement: Hard vs. Soft Failures

- **“Hard failures” are the kind of problems every organization understands**
  - Fiber cut
  - Power failure takes down routers
  - Hardware ceases to function
- **Classic monitoring systems are good at alerting hard failures**
  - i.e., NOC sees something turn red on their screen
  - Engineers paged by monitoring systems
- **“Soft failures” are different and often go undetected**
  - Basic connectivity (ping, traceroute, web pages, email) works
  - Performance is just poor
- **How much should we care about soft failures?**

# Causes of Packet Loss

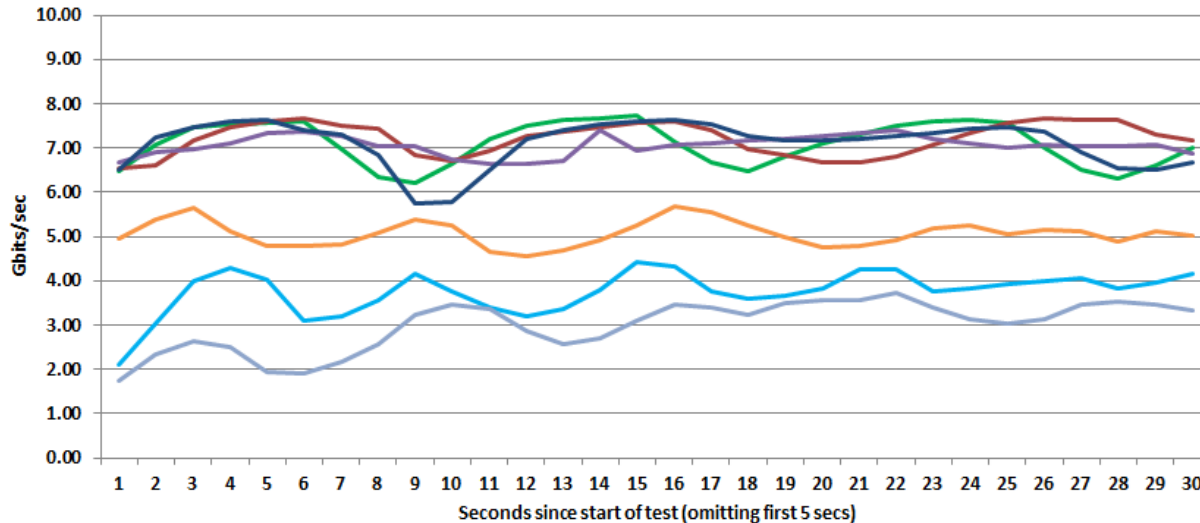
- **Network Congestion**
  - Easy to confirm via SNMP, easy to fix with \$\$
  - This is not a 'soft failure', but just a network capacity issue
  - Often people assume congestion is the issue when in fact it is not.
- **Under-buffered switch dropping packets**
  - Hard to confirm
- **Under-powered firewall dropping packets**
  - Hard to confirm
- **Dirty fibers or connectors, failing optics/light levels**
  - Sometimes easy to confirm by looking at error counters in the routers
- **Overloaded or slow receive host dropping packets**
  - Easy to confirm by looking at CPU load on the host



# Under-buffered Switches are probably our biggest problem today

**Comparison of Linecards & Devices**  
 Averages of 15 tests, 30 seconds each  
 with 50ms simulated RTT + 2Gbps UDP Background Traffic

- Arista 7500E-48S-LC
- Cisco WS-X6716-10GE Performance Mode
- Brocade NI-MLX-10Gx8-M (64M max-queue-size)
- Cisco WS-X6716-10GE Oversubscription Mode
- Cisco WS-X6704-10GE
- Arista 7150
- Brocade NI-MLX-10Gx8-M (default 1M max-queue-size)



# Outline

- Problem Statement on Network Connectivity
- Supporting Scientific Users
- Network Performance & TCP Behaviors w/ Packet Loss
- **What is perfSONAR**
- Deployment Overview
- Command's & Syntax
- Examples
- Conclusions



# But ... It's Not Just the Network

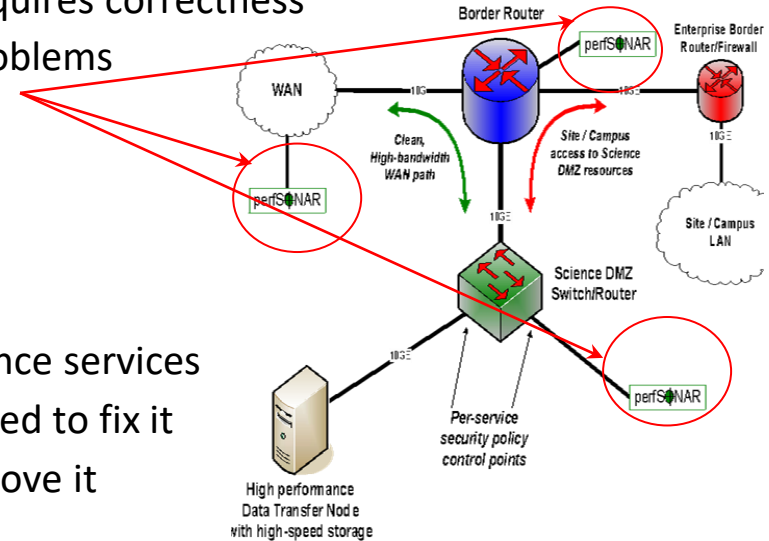
- **Perhaps you are saying to yourself “I have no control over parts of my campus, let alone the 5 networks that sit between me and my collaborators”**
  - Significant gains are possible in isolated areas of the OSI Stack
- **Things “you” control:**
  - Choice of data movement applications (say no to SCP and RSYNC)
  - Configuration of local gear (hosts, network devices)
  - Placement and configuration of diagnostic tools, e.g. *perfSONAR*
  - Use of the diagnostic tools
- **Things that need some help:**
  - Configuration of remote gear
  - Addressing issues when the diagnostic tools alarm
  - Getting someone to “care”

# Network Monitoring

- **All networks do some form monitoring.**
  - Addresses needs of local staff for understanding state of the network
    - Would this information be useful to external users?
    - Can these tools function on a multi-domain basis?
- **Beyond passive methods, there are active tools.**
  - E.g. often we want a ‘throughput’ number. Can we automate that idea?
  - Wouldn’t it be nice to get some sort of plot of performance over the course of a day? Week? Year? Multiple endpoints?
- **perfSONAR = Measurement Middleware**

# perfSONAR

- All the previous Science DMZ network diagrams have little perfSONAR boxes everywhere
  - The reason for this is that consistent behavior requires correctness
  - Correctness requires the ability to find and fix problems
    - ***You can't fix what you can't find***
    - ***You can't find what you can't see***
    - ***perfSONAR lets you see***
- Especially important when deploying high performance services
  - If there is a problem with the infrastructure, need to fix it
  - If the problem is not with your stuff, need to prove it
    - Many players in an end to end path
    - Ability to show correct behavior aids in problem localization



# What is perfSONAR?

- perfSONAR is a tool coordinated suite of tools to:
  - Set network performance expectations
  - Find network problems (“soft failures”)
  - Helps coordinate fixing these problems  
... all in multi-domain environments
- These problems are all harder when multiple networks are involved
- perfSONAR provides a standardized way to publish active and passive monitoring data
  - This data is interesting to network researchers as well as network operators



# perfSONAR History

- perfSONAR can trace its origin to the Internet2 “End 2 End performance Initiative” from the year 2000. What has changed since then?
  - The Good News:
    - TCP is much less fragile; Cubic is the default CC alg, autotuning is and larger TCP buffers are everywhere
    - Reliable parallel transfers via tools like Globus Online
    - High-performance UDP-based commercial tools like Aspera
  - The Bad News:
    - The **wizard gap** is still large
    - Jumbo frame use is still small
    - Under-buffered and switches and routers are still common
    - Under-powered/misconfigured firewalls are common
    - Soft failures still go undetected for months
    - Users still might not know or understand what kind of performance to expect

# Simulating Performance

- It's infeasible to perform at-scale data movement all the time – as we see in other forms of science, we need to rely on simulations
- Network performance comes down to a couple of key metrics:
  - Throughput (e.g. “how much can I get out of the network”)
  - Latency (time it takes to get to/from a destination)
  - Packet loss/duplication/ordering (for some sampling of packets, do they all make it to the other side without serious abnormalities occurring?)
  - Network utilization (the opposite of “throughput” for a moment in time)
- We can get many of these from a selection of active and passive measurement tools – enter the perfSONAR Toolkit



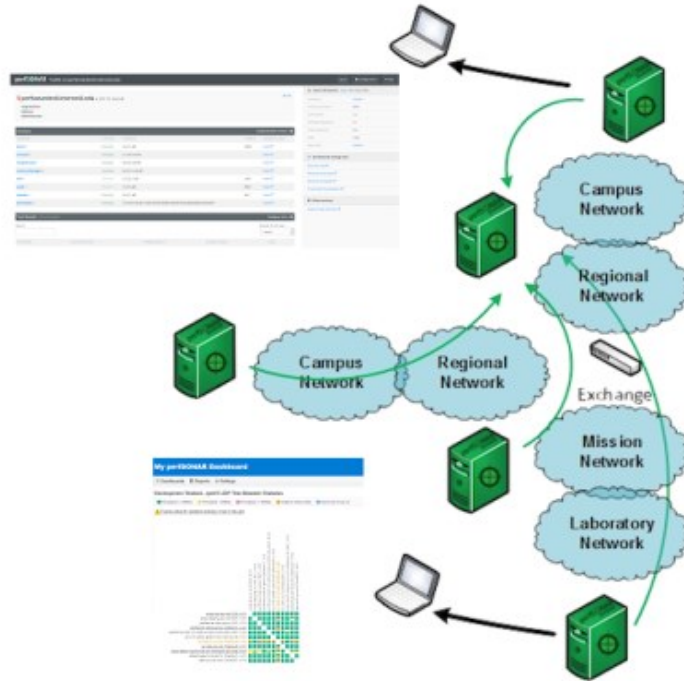
# perfSONAR Toolkit

- The “perfSONAR Toolkit” is an open source implementation and packaging of the perfSONAR measurement infrastructure and protocols
  - [http://docs.perfsonar.net/install\\_getting.html](http://docs.perfsonar.net/install_getting.html)
- All components are available as RPMs, DEBs, and bundled as CentOS 7, Debian 9 or Ubuntu 16 and 18-based packages (as of perfSONAR v. 4.2.4)
  - perfSONAR tools are much more accurate if run on a dedicated perfSONAR host
- Very easy to install and configure
  - Usually takes less than 30 minutes

# Outline

- Problem Statement on Network Connectivity
- Supporting Scientific Users
- Network Performance & TCP Behaviors w/ Packet Loss
- What is perfSONAR
- **Deployment Overview**
- Command's & Syntax
- Examples
- Conclusions

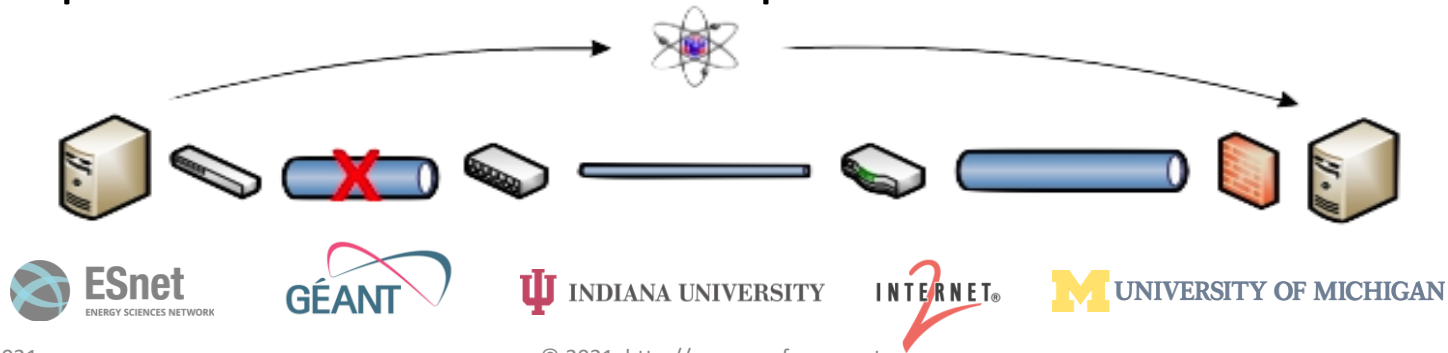
# Toolkit “Beacon” Use Case



- The general use case is to establish some set of tests to other locations/facilities
- To answer the what/why questions:
  - Regular testing with select tools helps to establish patterns – how much bandwidth we would see during the course of the day – or when packet loss appears
  - We do this to ‘points of interest’ to see how well a real activity (e.g. Globus transfer) would do.
- If performance is ‘bad’, don’t expect much from the data movement tool

# Benefits: Finding needles in haystacks

- Above all, perfSONAR allows you to maintain a healthy, high-performing network because it helps identify the “soft failures” in the network path.
  - Classical monitoring systems have limitations
    - Performance problems are often only visible at the ends
    - Individual network components (e.g. routers) have no knowledge of end host state
  - perfSONAR tests the network in ways that classical monitoring systems do not
- More perfSONAR distributions equal better network visibility.



# <http://stats.es.net/ServicesDirectory/>

perfSONAR
Lookup Service Directory

**Search**

Filter results by searching for specific terms: 🔍

Search
Show All

**Browser**

- ▶ pScheduler Server (1575)
- ▶ BWCTL Server (1893)
- ▶ OWAMP Server (1882)
- ▶ NDT Server (412)
- ▶ NPAD Server (136)
- ▶ Ping Responder (2078)
- ▶ Traceroute Responder (2080)
- ▶ MA (1933)
- ▶ BWCTL MP (1883)
- ▶ OWAMP MP (1881)
- ▶ bwctl10g (7)

Showing 15719 of 15719 services on 2048 hosts.

**Communities**

**Developer**

**Service Information**

Service Name	Addresses	Geographic Location	Communities	Version	Custom

**Host Information**

Host Name	Hardware	System Info	Toolkit Version	Communities

**Service Map**

Dane do Mapy ©2017 | Warunki korzystania z programu



# perfSONAR Dashboard: Raising Expectations and improving network visibility

## Status at-a-glance

- Packet loss
- Throughput
- Correctness

## Current live instances at

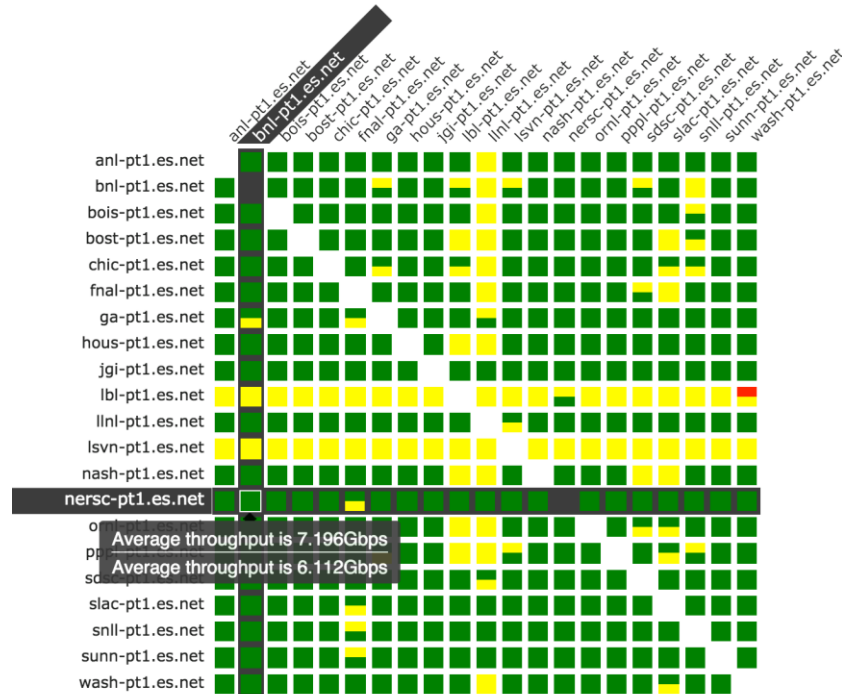
<http://pas.net.internet2.edu/>  
<http://ps-dashboard.es.net/>

## Drill-down capabilities:

- Test history between hosts
- Ability to correlate with other events
- Very valuable for fault localization and isolation

## ESnet - ESnet Hub to Large DOE Site Border Throughput Testing

■ Throughput  $\geq$  5000Mbps
 ■ Throughput  $<$  5000Mbps
 ■ Throughput  $\leq$  1000Mbps
 ■ Unable to



# Outline

- Problem Statement on Network Connectivity
- Supporting Scientific Users
- Network Performance & TCP Behaviors w/ Packet Loss
- What is perfSONAR
- Deployment Overview
- **Command's & Syntax**
- Examples
- Conclusions

# pscheduler – The Secret Sauce

- pscheduler is the engine that drives perfSONAR
  - Coordinates timeslots and schedules tests between nodes
  - Creates a common syntax that all tools use
  - Handles the storage of results
- This approach to test management ensures that:
  - Tests that could impact each other's performance, like throughput, are never run simultaneously
  - Simplified coordination, where you don't need an engineer at the other end to start a daemon, open a port, etc.
  - Access control is maintained and test limits are enforced



# Basic syntax

*pscheduler task [options] test-type [test-options]*

- *task* is what you want pscheduler to do
- *test-type* is how you want pscheduler to do it

There's more than one tool to run many of these tests, and pscheduler gives you the option to choose that tool:

- *iperf2/iperf3/nuttcp* for bandwidth
- *traceroute/tracepath/paris-traceroute* for routing
- *--help is your friend!*

# Remote Commands, AKA, Your Best Friends

- --source and --dest flags do what you expect they would, but neither end has to be your node
  - You can run most tests between two remote nodes
  - This includes perfSONAR's built-in self-diagnostic tools
- This core piece of functionality is what makes perfSONAR useful in day-to-day troubleshooting activities and makes it more than just a simple performance data collector

```
[ps-iniu@thrpt10ge-1 ~]$ pscheduler task throughput --source perf.newy32aoa.neaar.net --dest test.seat.transpac.org  
Submitting task...
```

Task URL:

<https://perf.newy32aoa.neaar.net/pscheduler/tasks/666b16fc-c32d-4960-a8ac-ef66b2eca183>

Running with tool 'iperf3'

Fetching first run...

Next scheduled run:

<https://perf.newy32aoa.neaar.net/pscheduler/tasks/666b16fc-c32d-4960-a8ac-ef66b2eca183/runs/31b57ed0-c39c-4533-8654-60e8a2137d64>

Starts 2021-05-26T14:18:53Z (~6 seconds)

Ends 2021-05-26T14:19:12Z (~18 seconds)

Waiting for result...

\* Stream ID 5

Interval	Throughput	Retransmits	Current Window
0.0 - 1.0	4.67 Gbps	1	148.25 MBytes
1.0 - 2.0	9.89 Gbps	0	148.39 MBytes
2.0 - 3.0	9.90 Gbps	0	148.39 MBytes
3.0 - 4.0	9.90 Gbps	2	152.96 MBytes
4.0 - 5.0	9.90 Gbps	0	152.96 MBytes
5.0 - 6.0	9.89 Gbps	0	152.96 MBytes
6.0 - 7.0	9.90 Gbps	0	152.96 MBytes
7.0 - 8.0	9.90 Gbps	0	152.96 MBytes
8.0 - 9.0	9.90 Gbps	0	152.96 MBytes
9.0 - 10.0	9.89 Gbps	0	152.96 MBytes

Summary

Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 10.0	9.37 Gbps	3	9.20 Gbps

No further runs scheduled.

```
[ps-iniu@thrpt10ge-1 ~]$ pscheduler task --tool tracepath trace --source perf.newy32aoa.neaar.net --dest test.seat.transpac.org
Submitting task...
Task URL:
https://perf.newy32aoa.neaar.net/pscheduler/tasks/a447400a-125a-46cb-9e78-020ab537cca1
Running with tool 'tracepath'
Fetching first run...
```

Next scheduled run:

```
https://perf.newy32aoa.neaar.net/pscheduler/tasks/a447400a-125a-46cb-9e78-020ab537cca1/runs/19416479-28b0-40a3-9844-2bd9c4f93a70
```

Starts 2021-05-26T14:30:25Z (~1 seconds)

Ends 2021-05-26T14:32:06Z (~100 seconds)

Waiting for result...

```
1      vlan-150.rtr.newy32aoa.neaar.net (192.203.116.32) AS396390 0.142 ms mtu 9000 bytes
      INDIANA-UNIVERSITY-NEAAR, US
2      et-2-1-5.127.rtsw.newy32aoa.net.internet2.edu (198.71.45.192) AS11537 0.789 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
3      ae-3.4079.rtsw.wash.net.internet2.edu (162.252.70.138) AS11537 6.192 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
4      ae-0.4079.rtsw2.ashb.net.internet2.edu (162.252.70.137) AS11537 6.62 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
5      ae-2.4079.rtsw.ashb.net.internet2.edu (162.252.70.74) AS11537 6.54 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
6      ae-20.4079.rtsw.clev.net.internet2.edu (162.252.70.129) AS11537 13.544 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
7      ae-3.4079.rtsw3.eqch.net.internet2.edu (162.252.70.131) AS11537 20.07 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
8      ae-5.4079.rtsw.eqch.net.internet2.edu (162.252.70.162) AS11537 26.237 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
9      ae-0.4079.rtsw.minn.net.internet2.edu (162.252.70.107) AS11537 27.898 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
10     ae-1.4079.rtsw.seat.net.internet2.edu (162.252.70.172) AS11537 59.964 ms mtu 9000 bytes
      INTERNET2-RESEARCH-EDU, US
11     207.231.240.24 AS53965 59.659 ms mtu 9000 bytes
      CCSEBGP, US
12     test.seat.transpac.org (192.203.115.2) AS22388 59.635 ms mtu 9000 bytes
      TRANSPAC, US
```

# Outline

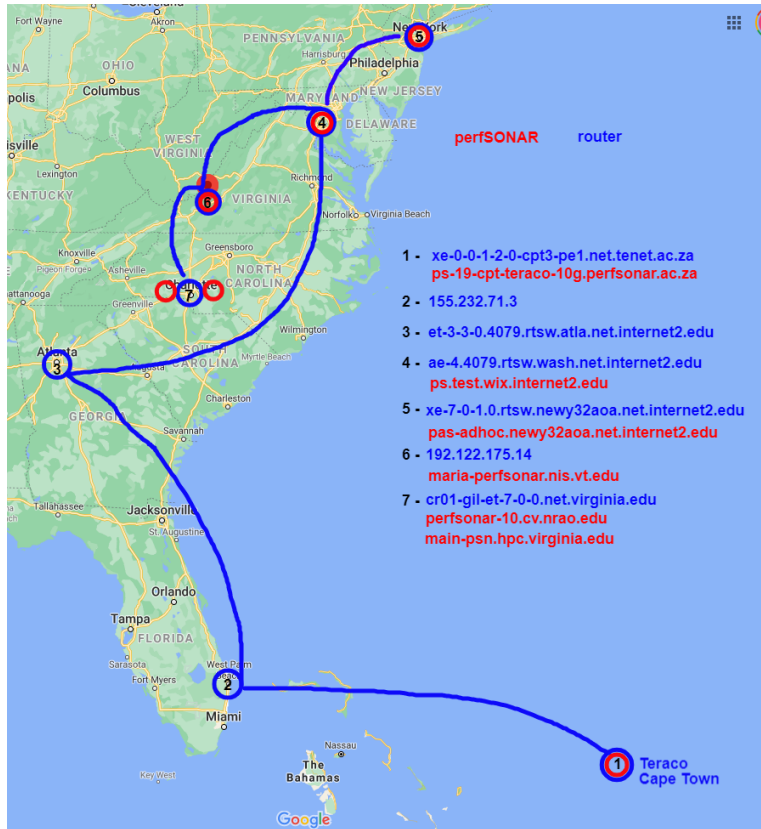
- Problem Statement on Network Connectivity
- Supporting Scientific Users
- Network Performance & TCP Behaviors w/ Packet Loss
- What is perfSONAR
- Deployment Overview
- Command's & Syntax
- **Examples**
- Conclusions



# NRAO/UVA <> SARAQ Performance Problem

- Data sharing from the National Radio Astronomy Observatory, located on the University of Virginia campus, to the South African Radio Astronomy Observatory
  - Low performance – 4.8Mbps
- Initial testing from the South African side revealed a few potential problems, such as asymmetric routing and paths with unnecessarily circuitous routes.
  - These were identified using normal traceroutes and quickly corrected
  - No appreciable change in performance

# Identification of possible pS toolkits



- 1 [te0-1-0-1-cpt2-pe1.net.tenet.ac.za](http://te0-1-0-1-cpt2-pe1.net.tenet.ac.za) (155.232.40.9) AS2018 0.703 ms
- ...
- 5 155.232.71.3 AS2018 166.6 ms  
TENET-1, ZA
- 6 [et-3-3-0.4079.rtsw.atla.net.internet2.edu](http://et-3-3-0.4079.rtsw.atla.net.internet2.edu) (162.252.70.42) AS11537 172.712 ms  
INTERNET2-RESEARCH-EDU, US
- 7 [ae-4.4079.rtsw.wash.net.internet2.edu](http://ae-4.4079.rtsw.wash.net.internet2.edu) (198.71.45.7) AS11537 185.775 ms  
INTERNET2-RESEARCH-EDU, US
- 8 [ae-0.4079.rtsw2.ashb.net.internet2.edu](http://ae-0.4079.rtsw2.ashb.net.internet2.edu) (162.252.70.137) AS11537 186.419 ms  
INTERNET2-RESEARCH-EDU, US
- 9 [ae-2.4079.rtsw.ashb.net.internet2.edu](http://ae-2.4079.rtsw.ashb.net.internet2.edu) (162.252.70.74) AS11537 185.845 ms  
INTERNET2-RESEARCH-EDU, US
- 10 192.122.175.14 AS40220 186.368 ms  
MARIA, US
- 11 [br01-udc-et-1-0-0-20.net.virginia.edu](http://br01-udc-et-1-0-0-20.net.virginia.edu) (192.35.48.33) AS225 188.065 ms  
VIRGINIA-AS, US
- 12 [cr01-udc-et-4-2-0.net.virginia.edu](http://cr01-udc-et-4-2-0.net.virginia.edu) (128.143.236.6) AS225 188.448 ms  
VIRGINIA-AS, US
- 13 [cr01-gil-et-7-0-0.net.virginia.edu](http://cr01-gil-et-7-0-0.net.virginia.edu) (128.143.236.89) AS225 203.281 ms  
VIRGINIA-AS, US
- 14 [perfonar-10.cv.nrao.edu](http://perfonar-10.cv.nrao.edu) (198.51.208.55) AS225 188.179 ms  
VIRGINIA-AS, US

# perfSONAR Lookup Service Directory

perfSONAR Lookup Service Directory

**Search**

Filter results by searching for specific terms:

**Browser**

- ▶ pScheduler Server (12)
- ▶ BWCTL Server (6)
- ▶ OWAMP Server (15)
- ▶ NDT Server (5)
- ▶ Ping Responder (1)
- ▶ Traceroute Responder (1)
- ▶ MA (11)
- ▶ BWCTL MP (6)
- ▶ OWAMP MP (4)
- ▶ twamp (3)

Showing: 69 of 7919 services on 14 hosts.

**Communities**

**Developer**

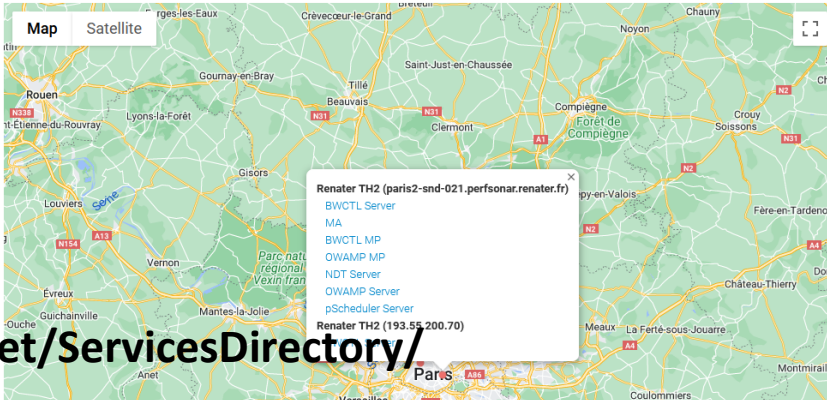
**Service Information**

Service Name	Addresses	Geographic Location	Communities	Version	Example Command-Line
Renater TH2 BWCTL Server	193.55.200.70	Renater TH2, Paris, France (48.8560, 2.3834)			<pre>bwctl -T iperf3 -l 30 -O 4 -c "193.55.200.70:4823" bwtracroute -T tracepath -c "193.55.200.70:4823" bwping -c "193.55.200.70:4823" bwctl -T iperf -l 30 -l 1 -f m -c</pre>

**Host Information**

Host Name	Hardware	System Info	Toolkit Version	Communities
<a href="http://paris2-snd-021.perfsonar.renater.fr">paris2-snd-021.perfsonar.renater.fr</a> 193.55.200.70	Processor #1: 3.50GHz (8 cores) Processor #2: 3.50GHz (8 cores) Memory: 64.18GB	Operating System: CentOS 6.10 (Final) Kernel: Linux 2.6.32-754.35.1.el6.x86_64	4.0.2.5-1.el6	

**Service Map**



Map Satellite

Renater TH2 (paris2-snd-021.perfsonar.renater.fr)  
BWCTL Server  
MA  
BWCTL MP  
OWAMP MP  
NDT Server  
OWAMP Server  
pScheduler Server  
Renater TH2 (193.55.200.70)

Lookup listed testpoints and toolkits by almost any criteria:

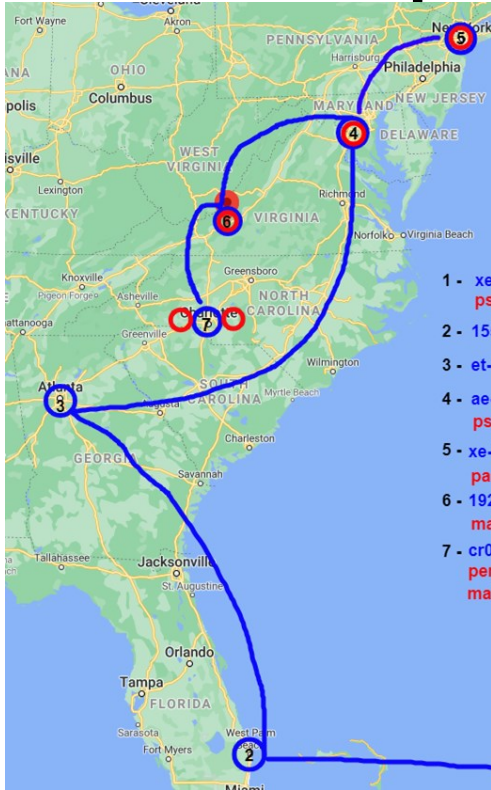
- Hostname
- IP address
- Institution
- City
- Country
- REN

pS instance must have commodity internet access to be listed.

<http://stats.es.net/ServicesDirectory/>

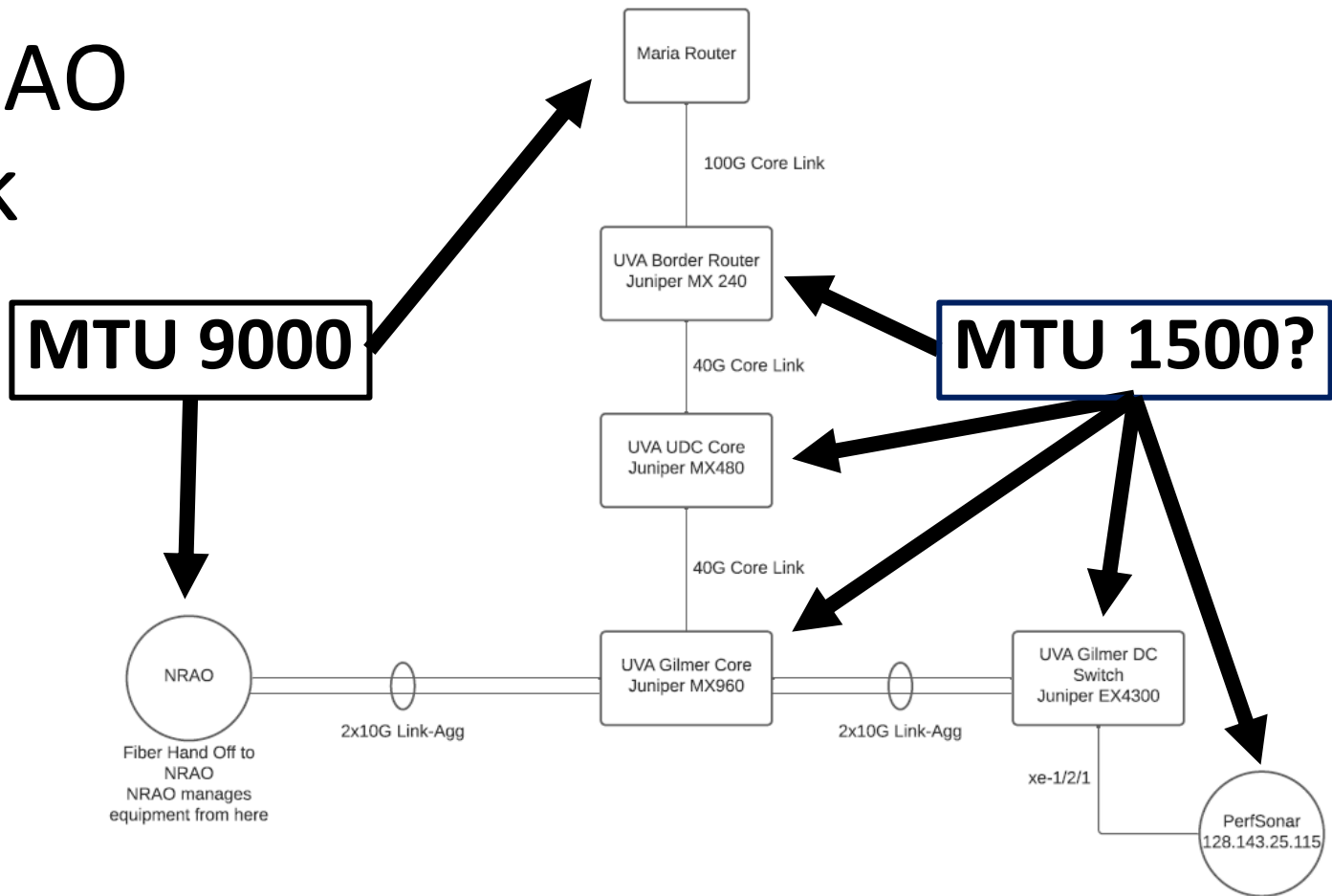


# Initial problem isolation



- Tests from various domestic and international perfSONAR nodes to UVAs campus were telling:
  - CHPC South Africa -> Internet2 Washington - 6.67 Gbps
  - Internet2 Albany -> Internet2 Washington - 9.893 Gbps
  - Internet2 Washington -> NRAO - **3.31 Gbps (lots of retries)**
  - Internet2 Washington -> HPC University Virginia - **2.21 Gbps (lots of retries)**
  - NRAO -> HPC University Virginia - **6.64 Gbps (lots of retries)**

# UVA/NRAO Network



# Path MTU Discovery (PMTUD)

- Is a layer 3 construct
- Requires UDP and ICMP to function
  - UDP packets larger than the MTU setting of the receiving router interface will trigger an ICMP “unreachable” message back to the sending router, which in turn causes a renegotiation to a lower MTU
- All is not lost if PMTUD doesn't work
  - Smart transfer tools can figure out a common MTU, at the cost of time
  - Packets sent at 9K can be fragmented to adhere to a smaller MTU, at the cost of performance...unless the no-fragment flag is set
  - Neither of these scenarios is good for high performance. PMTUD should be made to work and common MTUs enforced wherever possible

# Further isolation

Working inward from a known good ESnet perfSONAR node to UVA:

Interval	Throughput	Retransmits	Current Window
0.0 - 1.0	9.13 Gbps	22	33.17 MBytes
1.0 - 2.0	9.35 Gbps	0	33.58 MBytes
2.0 - 3.0	9.38 Gbps	0	33.58 MBytes
3.0 - 4.0	9.38 Gbps	0	33.58 MBytes
4.0 - 5.0	9.38 Gbps	0	33.58 MBytes
5.0 - 6.0	9.35 Gbps	0	33.58 MBytes
6.0 - 7.0	9.36 Gbps	0	33.58 MBytes
7.0 - 8.0	9.38 Gbps	0	33.58 MBytes
8.0 - 9.0	9.37 Gbps	0	33.58 MBytes
9.0 - 10.0	9.37 Gbps	0	33.58 MBytes

Summary Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 10.0	9.35 Gbps	22	9.25 Gbps

**This test looks good,  
because the hosts  
successfully negotiate  
1500 MTU**



# Negotiations break down

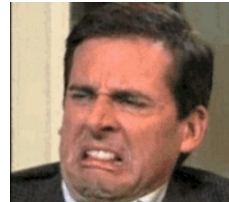
Working inward from a known good ESnet perfSONAR node to NRAO:  
(Keep in mind, we know MTU 9000 on both ends, but with a step down to 1500 in the middle of the UVA campus)

Interval	Throughput	Retransmits	Current Window
0.0 - 1.0	2.54 Mbps	2	8.95 KBytes
1.0 - 2.0	0.00bps	1	8.95 KBytes
2.0 - 3.0	0.00bps	0	8.95 KBytes
3.0 - 4.0	0.00bps	31	3.07 KBytes
4.0 - 5.0	0.00bps	67	5.12 KBytes
5.0 - 6.0	4.45 Mbps	2	17.41 KBytes
6.0 - 7.0	8.26 Mbps	0	33.79 KBytes
7.0 - 8.0	23.28 Mbps	0	94.21 KBytes
8.0 - 9.0	51.75 Mbps	0	218.11 KBytes
9.0 - 10.0	83.88 Mbps	0	392.19 KBytes

**9000B packets failing**

**1500B packets after re-negotiation**

Summary Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 10.0	17.42 Mbps	103	10.29 Mbps



# Traceroute: ESnet to NRAO

```
traceroute to perfsonar-10.cv.nrao.edu (198.51.208.55), 30 hops max, 60 byte packets
 1 esneteastrt1-eastdcpt1.es.net (198.124.238.37) 0.549 ms 0.544 ms 0.547 ms
 2 newycr5-ip-a-esneteastrt1.es.net (198.124.218.17) 1.969 ms 1.963 ms 1.953 ms
 3 aofacr5-ip-a-newycr5.es.net (134.55.37.77) 2.330 ms 2.304 ms 2.313 ms
 4 et-2-1-5.197.rtsw.newy32aoa.net.internet2.edu (64.57.28.14) 2.323 ms 2.324 ms 2.327 ms
 5 ae-3.4079.rtsw.wash.net.internet2.edu (162.252.70.138) 7.571 ms 7.672 ms 7.528 ms
 6 ae-0.4079.rtsw2.ashb.net.internet2.edu (162.252.70.137) 8.095 ms 8.077 ms 8.061 ms
 7 ae-2.4079.rtsw.ashb.net.internet2.edu (162.252.70.74) 28.089 ms 18.414 ms 18.454 ms
 8 192.122.175.14 (192.122.175.14) 8.221 ms 8.179 ms 8.205 ms
 9 br01-udc-et-1-0-0-20.net.virginia.edu (192.35.48.33) 10.310 ms 10.310 ms 10.383 ms
10 cr01-udc-et-4-2-0.net.virginia.edu (128.143.236.6) 12.609 ms 12.603 ms 12.638 ms
11 cr01-gil-et-7-0-0.net.virginia.edu (128.143.236.89) 12.407 ms 12.403 ms 12.393 ms
12 perfsonar-10.cv.nrao.edu (198.51.208.55) 10.058 ms 10.032 ms 10.022 ms
```

Well, that looks good. Let's try tracepath and see where the MTU changes

# Tracepath: ESnet to NRAO

```
1?: [LOCALHOST] pmtu 9000
1: esneteastrt1-eastdcpt1.es.net 0.788ms
1: bnlmr2-bnlpt1.es.net 0.728ms
2: no reply
3: aofacr5-ip-b-newycr5.es.net 2.411ms asymm 2
4: et-2-1-5.197.rtsw.newy32aoa.net.internet2.edu 2.468ms asymm 3
5: ae-3.4079.rtsw.wash.net.internet2.edu 8.176ms asymm 4
6: ae-0.4079.rtsw2.ashb.net.internet2.edu 8.889ms asymm 5
7: ae-2.4079.rtsw.ashb.net.internet2.edu 8.242ms asymm 6
8: 192.122.175.14 8.522ms asymm 7
9: no reply
10: no reply
11: no reply
12: no reply
Traceroute works, but tracepath doesn't??
```

# Different Tools, Different Packets

- Traceroute uses small 60B UDP packets
- Tracepath uses larger 64KB UDP packets

So, somewhere we have a roadblock. Small packets can make it through, but larger ones are dropped (not fragmented).

How do we figure out the max size? Trial and error. Start at 9K and cut the size in half until you get a response, then sneak back up until the packets disappear again.



# Tracepath: ESnet to NRAO, 1509 bytes

1: esneteastrt1-eastdcpt1.es.net	0.340ms	
2: no reply		
3: aofacr5-ip-a-newycr5.es.net	2.279ms asymm	2
4: et-2-1-5.197.rtsw.newy32aoa.net.internet2.edu	2.310ms asymm	3
5: ae-3.4079.rtsw.wash.net.internet2.edu	7.574ms asymm	4
6: ae-0.4079.rtsw2.ashb.net.internet2.edu	9.422ms asymm	5
7: ae-2.4079.rtsw.ashb.net.internet2.edu	7.986ms asymm	6
8: 192.122.175.14	8.123ms asymm	7
9: no reply		

← MARIA  
← UVA

# Tracepath: ESnet to NRAO, 1508 bytes

1: bnlnr2-bnlpt1.es.net	0.327ms	
2: no reply		
3: aofacr5-ip-b-newycr5.es.net	2.332ms asymm 2	
4: et-2-1-5.197.rtsw.newy32aoa.net.internet2.edu	2.338ms asymm 3	
5: ae-3.4079.rtsw.wash.net.internet2.edu	7.668ms asymm 4	
6: ae-0.4079.rtsw2.ashb.net.internet2.edu	9.833ms asymm 5	
7: ae-2.4079.rtsw.ashb.net.internet2.edu	7.872ms asymm 6	
8: 192.122.175.14	8.166ms asymm 7	
9: br01-udc-et-1-0-0-20.net.virginia.edu	9.998ms asymm 7	
9?: br01-udc-et-1-0-0-20.net.virginia.edu	asymm 7	
10: cr01-udc-et-4-2-0.net.virginia.edu	10.470ms asymm 8	
11: cr01-gil-et-7-0-0.net.virginia.edu	10.208ms asymm 9	
12: cr01-gil-et-7-0-0.net.virginia.edu	10.253ms pmtu 1500	
12: perfsonar-10.cv.nrao.edu	10.154ms !H	
Resume: pmtu 1500		

← MARIA  
← UVA

# Problem located

- The issue was between the MARIA router and the UVA router
  - The MARIA interface was configured for MTU 9192
  - The UVA interface was configured for MTU 1518
- With PMTUD broken there was no hope for external MTU 9000 equipment to negotiate an appropriate MTU with the NRAO node
- UVA changed the MTU on their router interface to match that of MARIA, while keeping their downstream equipment at their campus standard MTU 1500

# Yeah, yeah, but what about performance??

## Before:

```
pscheduler task throughput --source cpt-chpc-10g.perfsonar.ac.za --dest perfsonar-10.cv.nrao.edu
```

### Summary

Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 10.0	380.37 Kbps	58	108.18 Kbps

## After:

```
pscheduler task throughput -t 30 --source cpt-chpc-10g.perfsonar.ac.za --dest perfsonar-10.cv.nrao.edu
```

### Summary

Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 30.0	2.67 Gbps	0	2.62 Gbps

# Outline

- Problem Statement on Network Connectivity
- Supporting Scientific Users
- Network Performance & TCP Behaviors w/ Packet Loss
- What is perfSONAR
- Deployment Overview
- **Conclusions**

# Benefit: Active and Growing Community

- Active email lists and forums provide:
  - Instant access to advice and expertise from the community.
  - Ability to share metrics, experience and findings with others to help debug issues on a global scale.
- Joining the community automatically increases the reach and power of perfSONAR
  - The more endpoints means exponentially more ways to test and discover issues, compare metrics



# perfSONAR Community

- The perfSONAR collaboration is working to build a strong user community to support the use and development of the software.
- perfSONAR Mailing Lists
  - Announcement Lists:
    - <https://lists.internet2.edu/sympa/subscribe/perfsonar-announce>
  - Users List:
    - <https://lists.internet2.edu/sympa/subscribe/perfsonar-user>

# Resources



- perfSONAR website
  - <http://www.perfsonar.net/>
- perfSONAR Documentation
  - <http://docs.perfsonar.net/>
- perfSONAR mailing lists
  - <http://www.perfsonar.net/about/getting-help/>
- perfSONAR directory
  - <http://stats.es.net/ServicesDirectory/>
- perfSONAR YouTube Channel
  - <https://www.youtube.com/channel/UCjK-P49pAKK9hUrrNbbe0Sg>
- FasterData Knowledgebase
  - <http://fasterdata.es.net/>





**EPOC**

Engagement and Performance  
Operations Center

# perfSONAR Topics

Ken Miller, Jason Zurawski

[ken@es.net](mailto:ken@es.net), [zurawski@es.net](mailto:zurawski@es.net)

ESnet / Lawrence Berkeley National Laboratory

***Modern Cyberinfrastructure for Research Data  
Management Workshop  
University of Central Florida  
February 16-17, 2023***

**TACC**  
TEXAS ADVANCED COMPUTING CENTER



**ESnet**  
ENERGY SCIENCES NETWORK