Achieving the Science DMZ

Eli Dart, Brian Tierney and Eric Pouyoul, ESnet

Joe Breen: University of Utah

Joint Techs, Baton Rouge, LA, January, 2012

Section 3: Bulk Data Transfer Tools

Section Outline

Setting expectations

What makes a fast data transfer tool

Just say no to scp

GridFTP

Commercial Tools

Tool Tuning

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?)

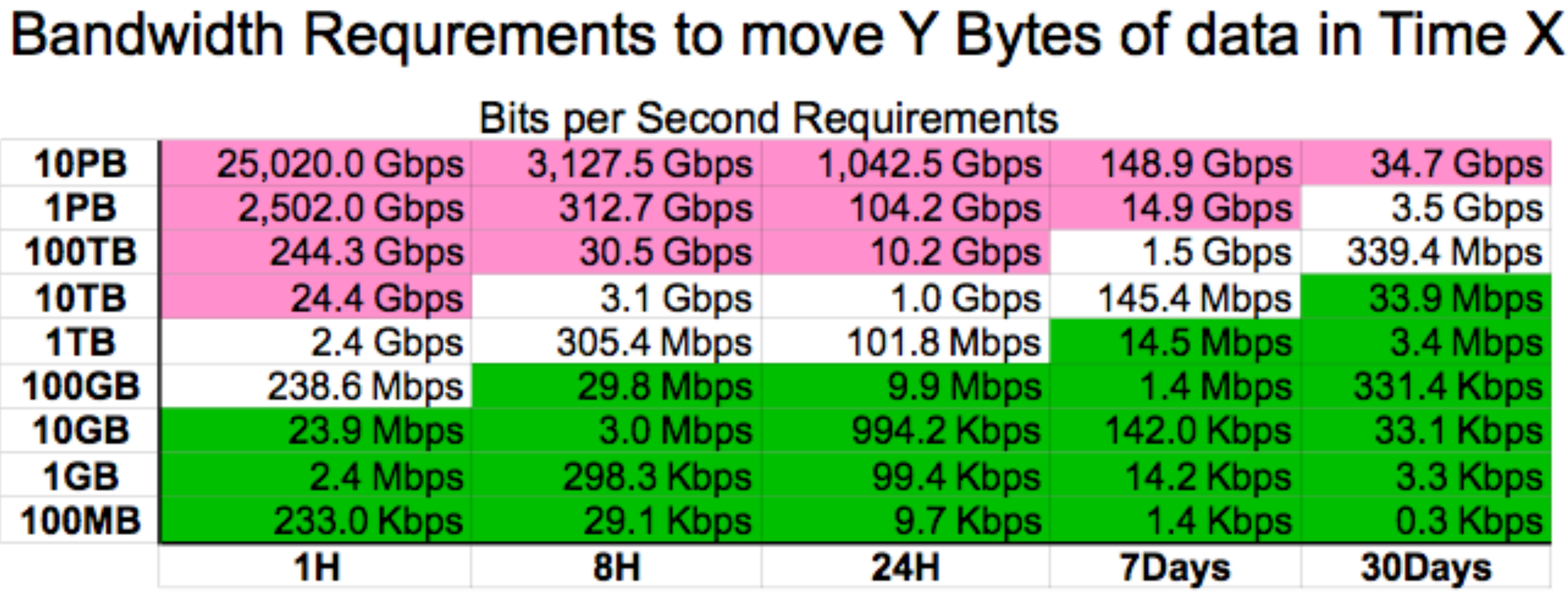
10 Gbps network : 20 minutes (need really fast disks and filesystem)

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

Bandwidth Requirements



Bandwidth Requrements to move Y Bytes of data in Time X

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

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

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	5.4 Gbps
– GridFTP, 8 streams	6.6 Gbps

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

Data Transfer Tools

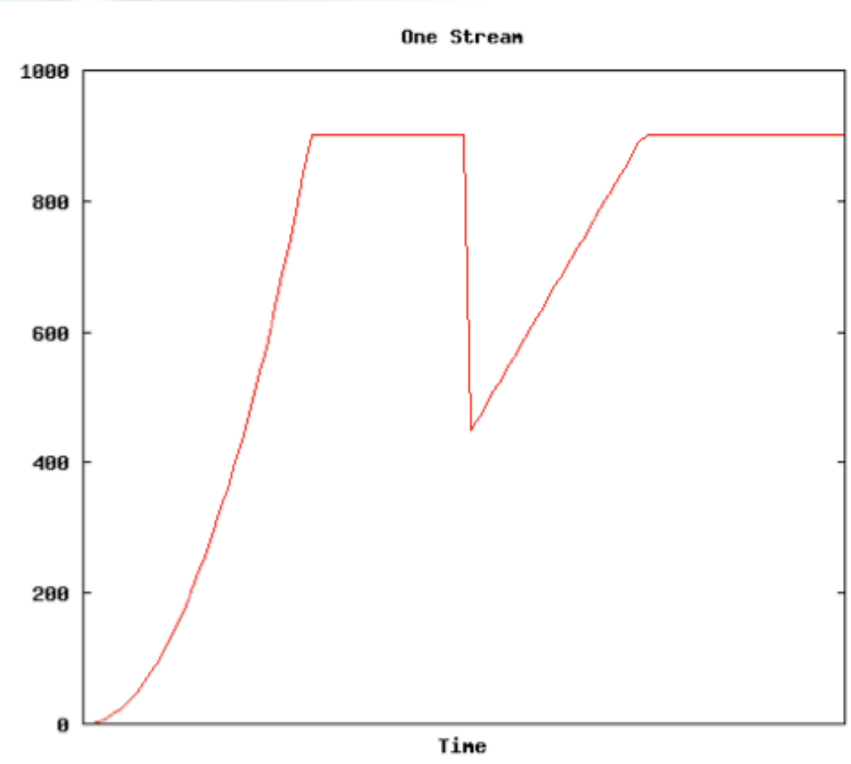
Parallelism is key

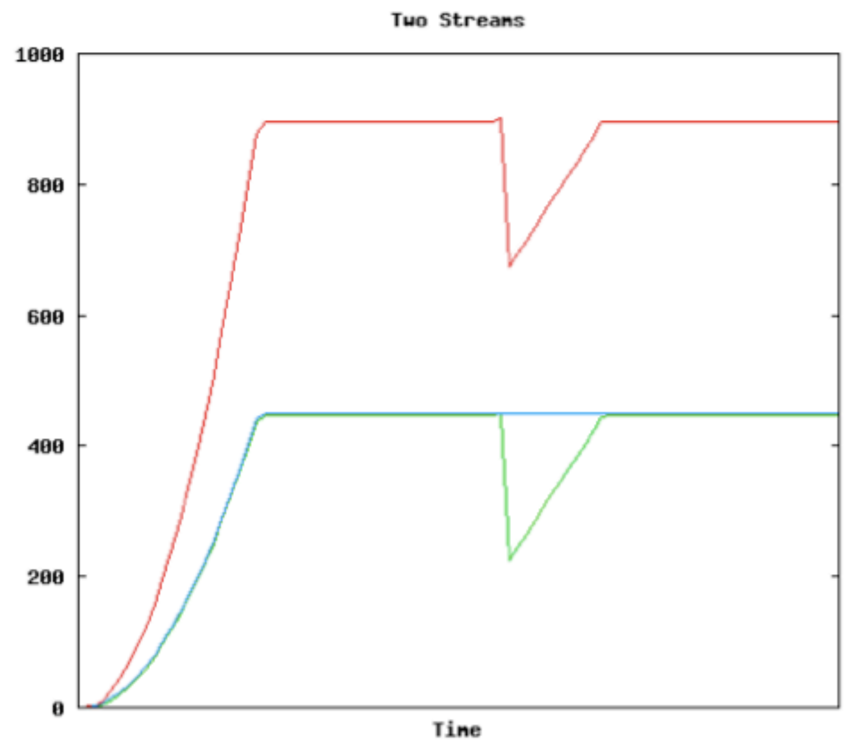
- It is much easier to achieve a given performance level with four parallel connections than one connection
- Several tools offer parallel transfers

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

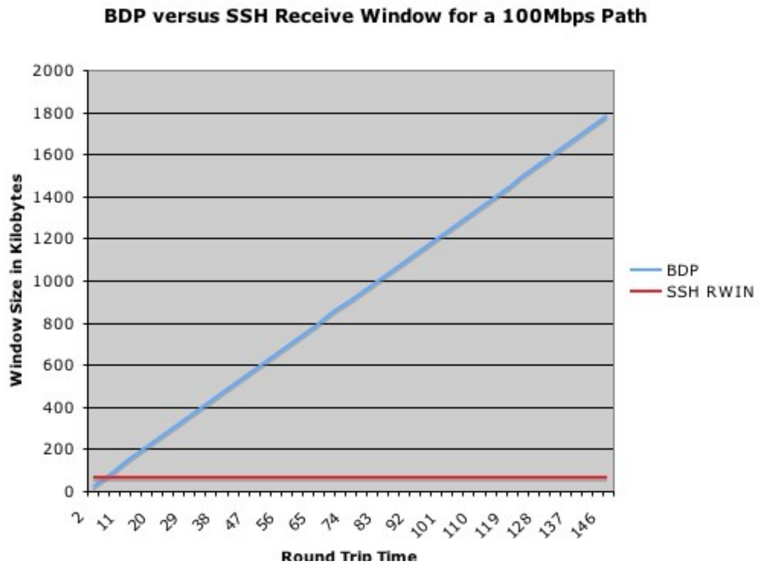
Cons:

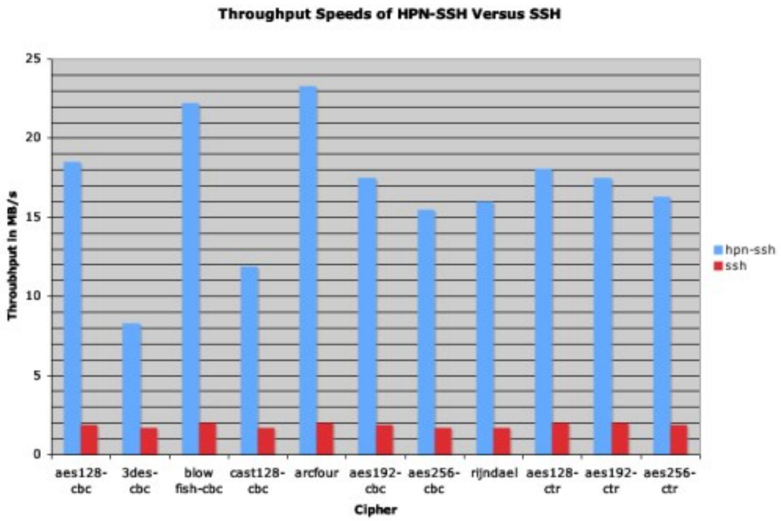
- The protocol used by SCP/SFTP has a fundamental flaw that limits WAN performance
- CPU speed doesn't matter – latency matters
- Fixed-size buffers reduce performance as latency increases
- It doesn't matter how easy it is to use SCP and SFTP – they simply do not perform

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
- being deployed by the DYNES 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

Sample Use:

- globus-url-copy -p 4 sshftp://data.lbl.gov/home/mydata/myfile file://home/mydir/myfile

Available from: http://www.globus.org/toolkit/downloads/

Some newer GridFTP Features

ssh authentication option

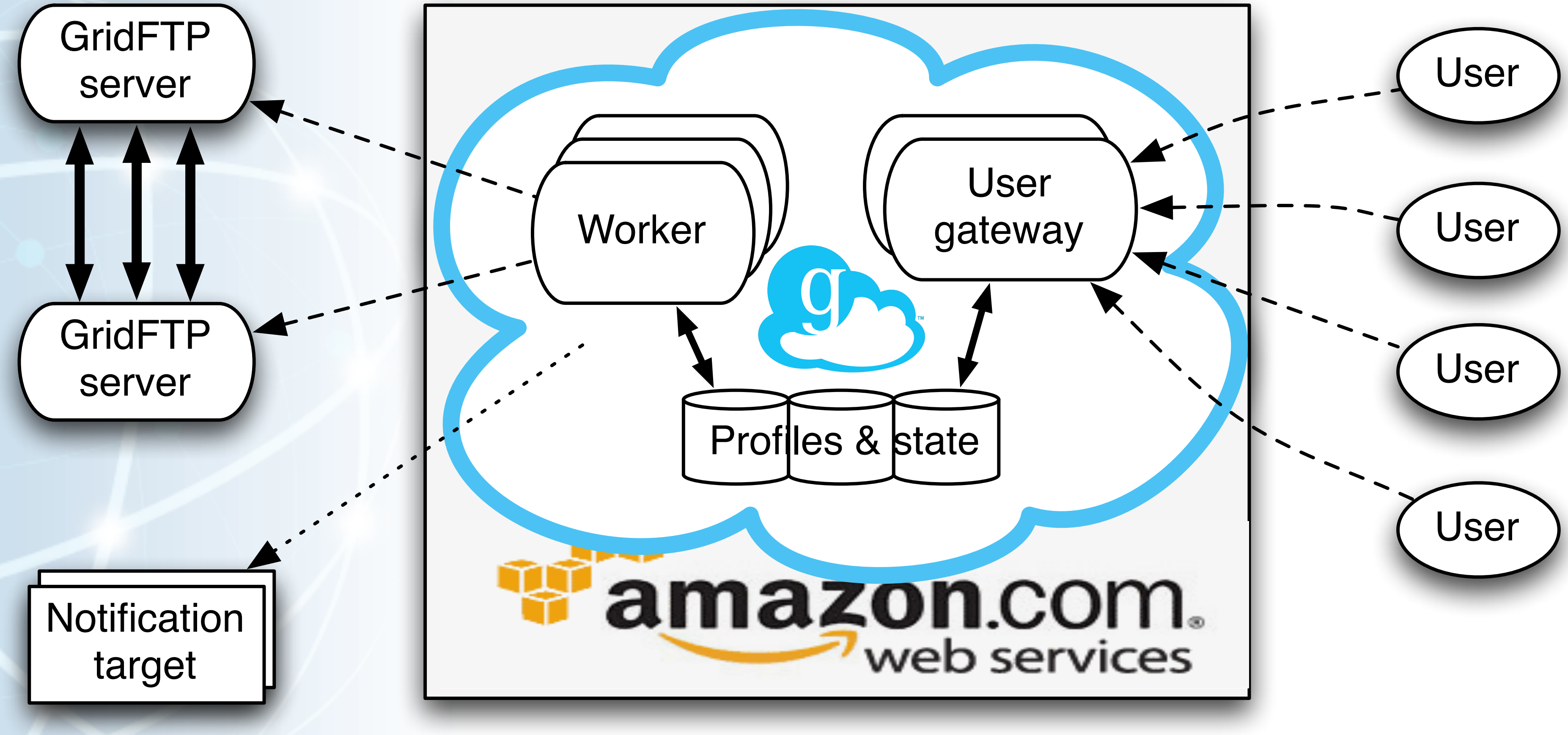
- Not all users need or want to deal with X.509 certificates
- Solution: Use SSH for Control Channel
  - Data channel remains as is, so performance is the same
- see http://fasterdata.es.net/gridftp.html for a quick start guide

Optimizations for small files

- Concurrency option (-cc)
  - establishes multiple control channel connections and transfer multiple files simultaneously.
- Pipelining option for multi-file transfers (-pp):
  - Client sends next request before the current completes
- Cached Data channel connections
  - Reuse established data channels (Mode E)
  - No additional TCP or GSI connect overhead

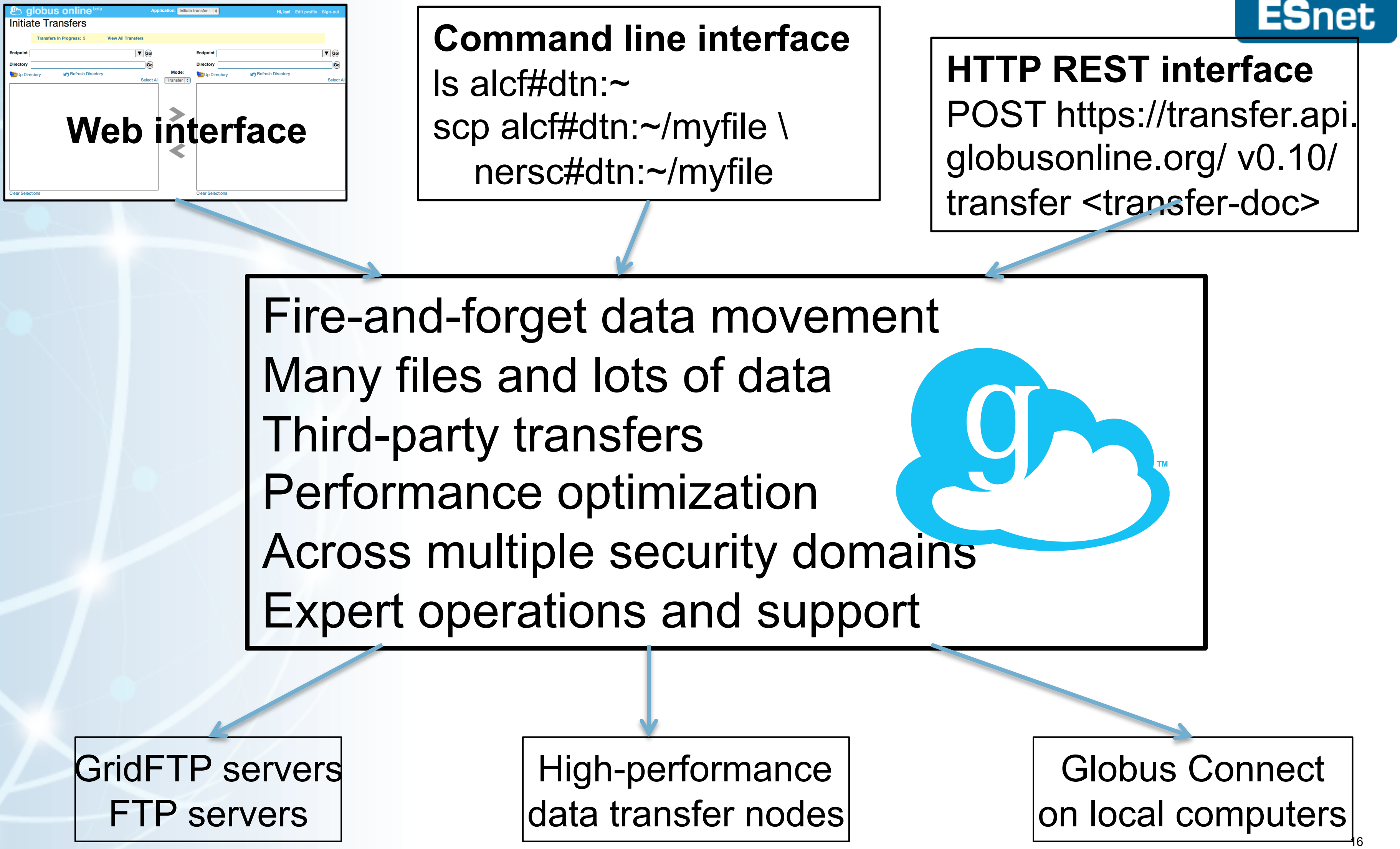
Support for UDT protocol

Globus Online: An easy to use wrapper for GridFTP:



Slide from Steve Tueke, ANL

Globus Online highlights



Web interface

Command line interface

```
ls alcf#dtn:~
scp alcf#dtn:~/myfile \
    nersc#dtn:~/myfile
```

HTTP REST interface

POST https://transfer.api.globusonline.org/ v0.10/ transfer <transfer-doc>

Fire-and-forget data movement
Many files and lots of data
Third-party transfers
Performance optimization
Across multiple security domains
Expert operations and support

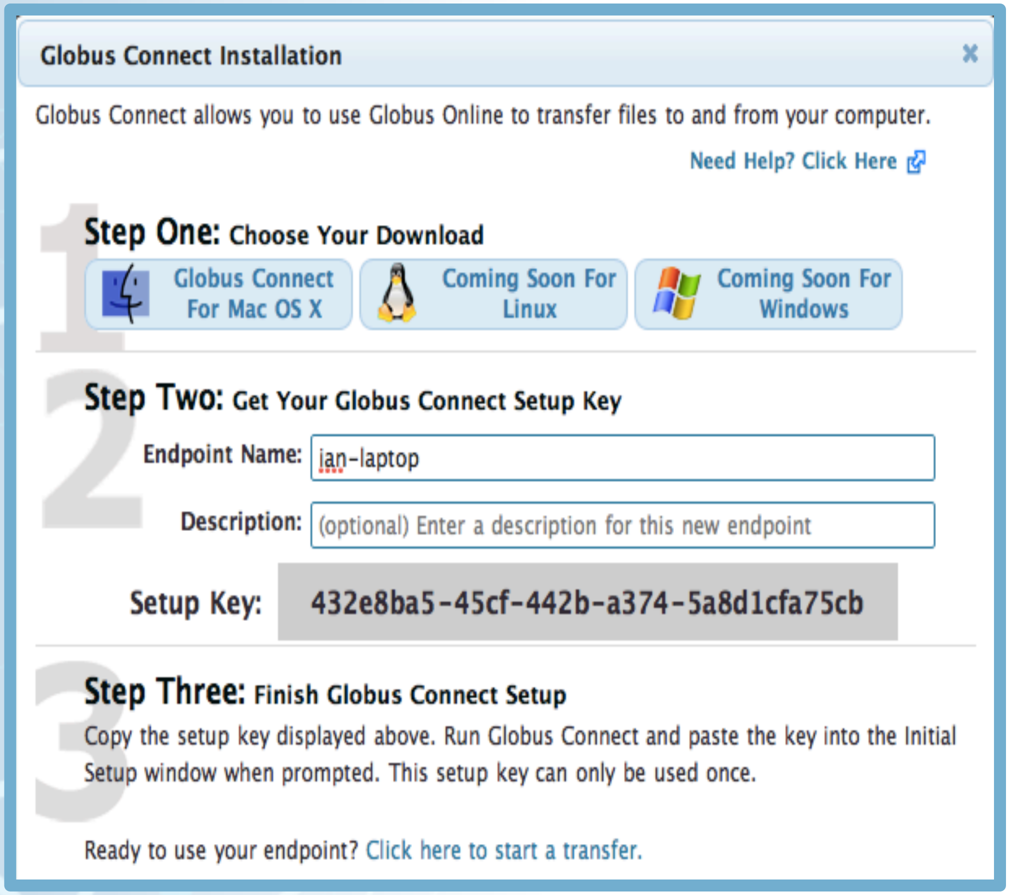
GridFTP servers
FTP servers

High-performance
data transfer nodes

Globus Connect
on local computers

Slide from Steve Tueke, ANL

Globus Connect to/from your laptop



Globus Connect Installation ×

Globus Connect allows you to use Globus Online to transfer files to and from your computer.

Need Help? Click Here

Step One: Choose Your Download

Step Two: Get Your Globus Connect Setup Key

Endpoint Name: ian-laptop

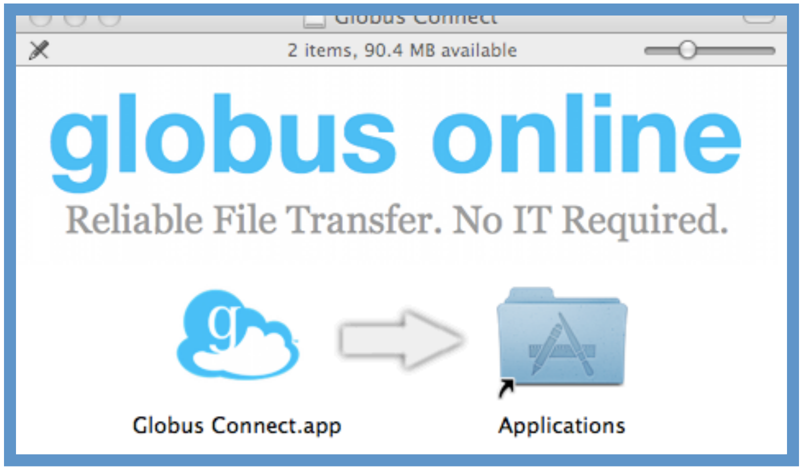
Description: (optional) Enter a description for this new endpoint

Setup Key: 432e8ba5-45cf-442b-a374-5a8d1cfa75cb

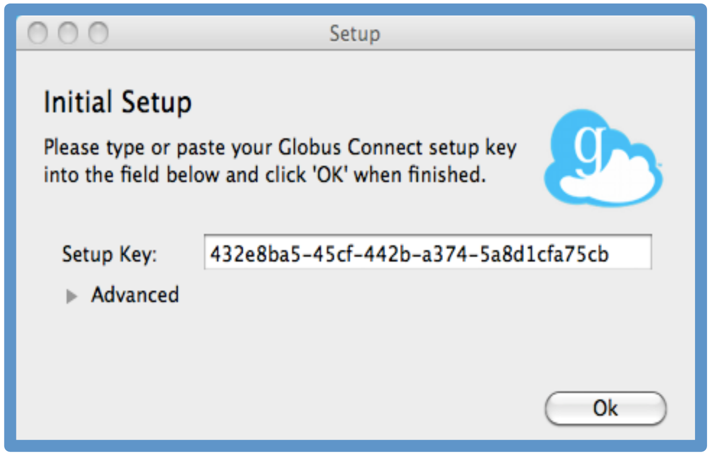
Step Three: Finish Globus Connect Setup

Copy the setup key displayed above. Run Globus Connect and paste the key into the Initial Setup window when prompted. This setup key can only be used once.

Ready to use your endpoint? Click here to start a transfer.



Globus Connect
2 items, 90.4 MB available
globus online
Reliable File Transfer. No IT Required.
Globus Connect.app
Applications



Initial Setup

Please type or paste your Globus Connect setup key into the field below and click 'OK' when finished.

Setup Key: 432e8ba5-45cf-442b-a374-5a8d1cfa75cb

Slide from Steve Tueke, ANL

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
  - Avoids the fairly complex GridFTP server installation process

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

Globus Connect Multi-User Installation

GridFTP finally comes as an easy to install RPM wrapped in a shell script

Installation steps:

```
wget http://connect.globusonline.org/linux/stable/globusconnect-multiuser-latest.tgz

tar -xvzf globusconnect-multiuser-latest.tgz

cd gcmu*

sudo ./install

  (And answer a couple simple questions)
```

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 congested, high-latency path

- advantage of these tools less clear on an uncongested path

They all have different, fairly complicated pricing models

Next Generation Tools/Protocols

RDMA-based tools:

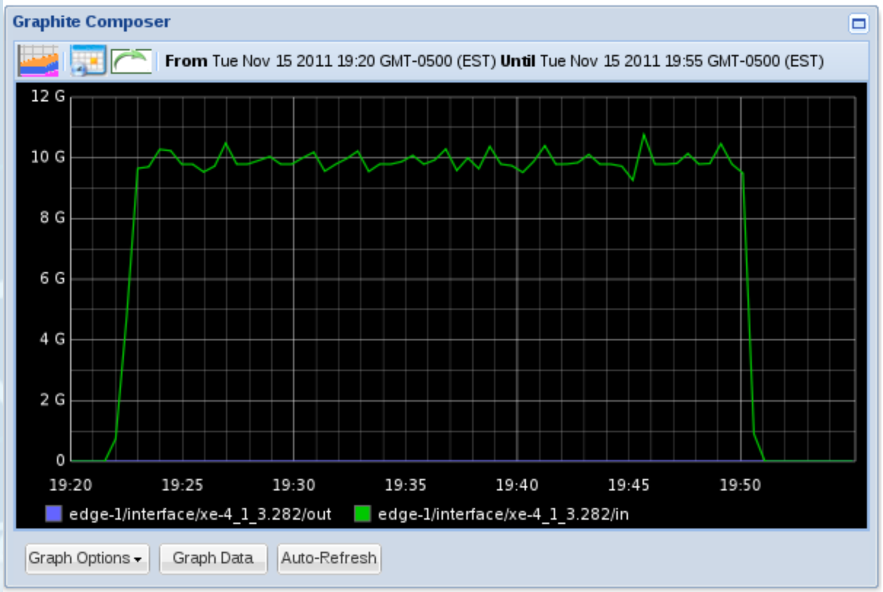
- Several groups have been experimenting with RDMA over the WAN
  - XIO driver for GridFTP (UDEL, OSU)
  - RFTP: BNL
- Over a dedicated layer-2 circuit, performance is the same as TCP, with **much** less CPU
- Requires hardware support on the NIC (e.g.: Mellanox)
  - Software version exists, but requires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

Session Layer Networking / Phoebus:

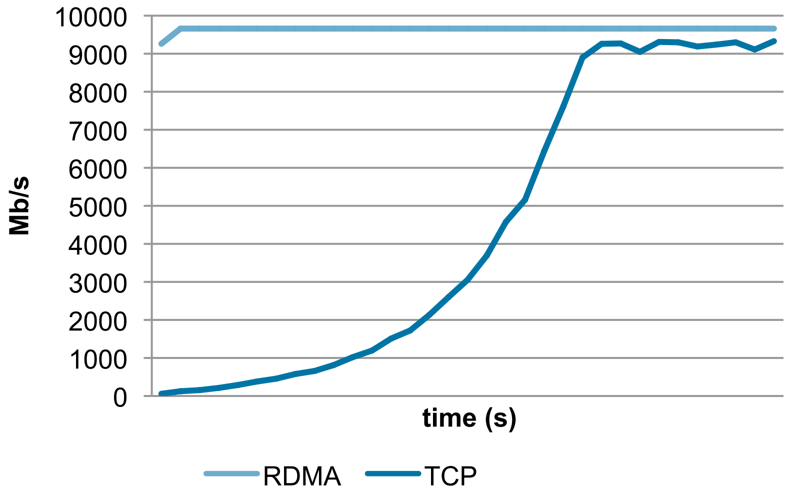
Phoebus Gateway can by used to translate being the LAN protocol (e.g. TCP) and a more efficient WAN protocol (e.g.: RDMA)

Sample RDMA Results: 10G dedicated layer-2 circuit, Long Island NY to Seattle

- 9.9G for both TCP and RDMA
  - 80% CPU for TCP
  - 3-4% CPU load for RDMA
- RDMA ramps up much faster than TCP



From Tue Nov 15 2011 19:20 GMT-0500 (EST) **Until** Tue Nov 15 2011 19:55 GMT-0500 (EST)



Tuning your Data Transfer Tools

Be sure to check the following:

- What is your host's maximum TCP window size?
  - 32M is good for most many environments
  - More for jumbo frames or very long RTT paths
- Which TCP congestion algorithm are you using?
  - Cubic or HTCP are usually best

- How many parallel streams are you using?
  - Use as few as possible that fill the pipe, usually 2-4 streams
  - Too many streams usually end up stepping on each other
    - May need more streams in cases of:
      - Very high RTT paths
      - Traversing slow firewalls
      - Paths without enough switch buffering

Section 4: Network Performance Monitoring and Troubleshooting using perfSONAR

Section Outline

Problem definition

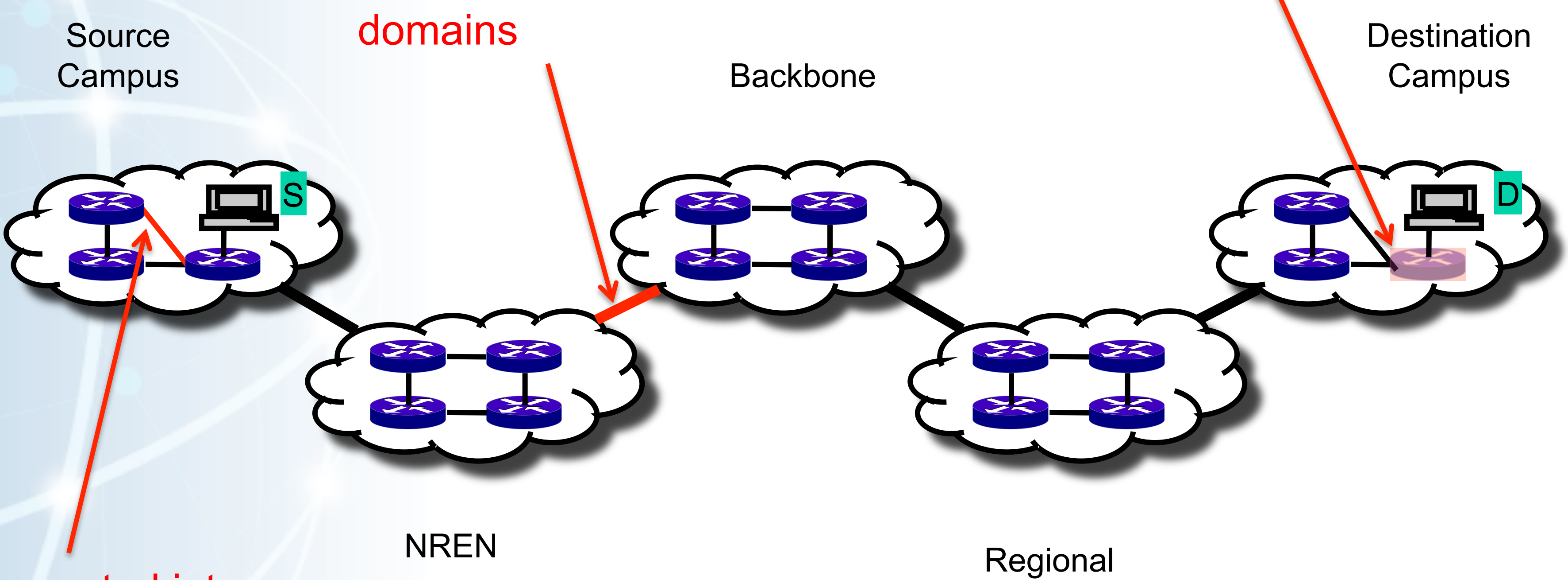
perfSONAR overview

Case studies

Site deployment recommendations

perfSONAR host recommendations

Where are common problems?

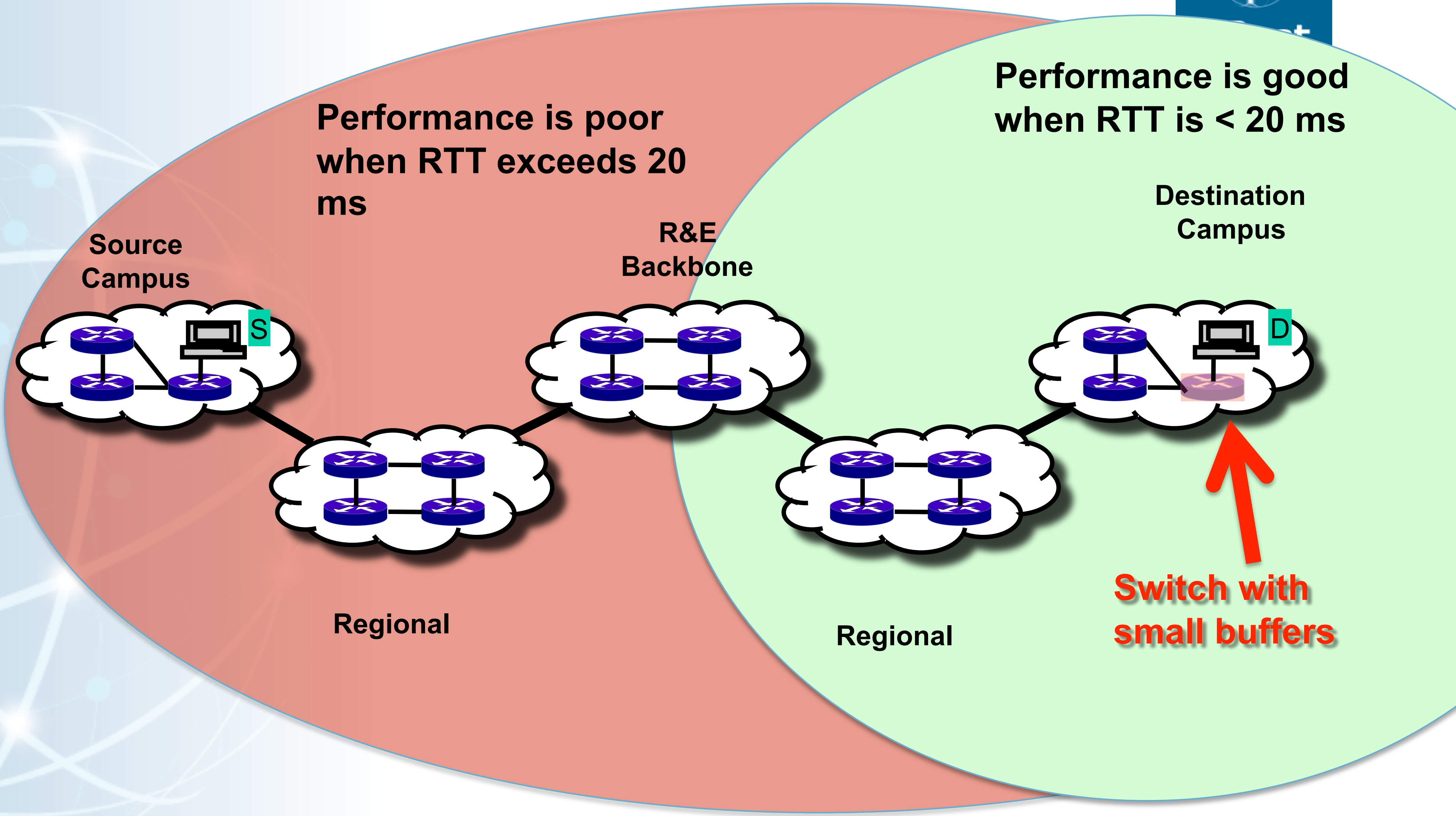


Congested or faulty links between domains

Latency dependant problems inside domains with small RTT

Congested intra-campus links

Local testing will not find all problems



Performance is poor when RTT exceeds 20 ms

Performance is good when RTT is < 20 ms

Source Campus

R&E Backbone

Destination Campus

Regional

Regional

Switch with small buffers

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

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

A Nagios alert based on our regular throughput testing between one site and ESnet core alerted us to poor performance on high latency paths

No errors or drops reported by routers on either side of problem link

- only perfSONAR bwctl tests caught this problem

Using packet filter counters, we saw 0.0046% loss in one direction

- 1 packets out of 22000 packets

Performance impact of this: (outbound/inbound)

- To/from test host 1 ms RTT : 7.3 Gbps out / 9.8 Gbps in
- To/from test host 11 ms RTT: 1 Gbps out / 9.5 Gbps in
- To/from test host 51ms RTT: 122 Mbps out / 7 Gbps in
- To/from test host 88 ms RTT: 60 Mbps out / 5 Gbps in
  - More than 80 times slower!

Common Soft Failures

Random Packet Loss

- Bad/dirty fibers or connectors
- Low light levels due to amps/interfaces failing
- Duplex mismatch

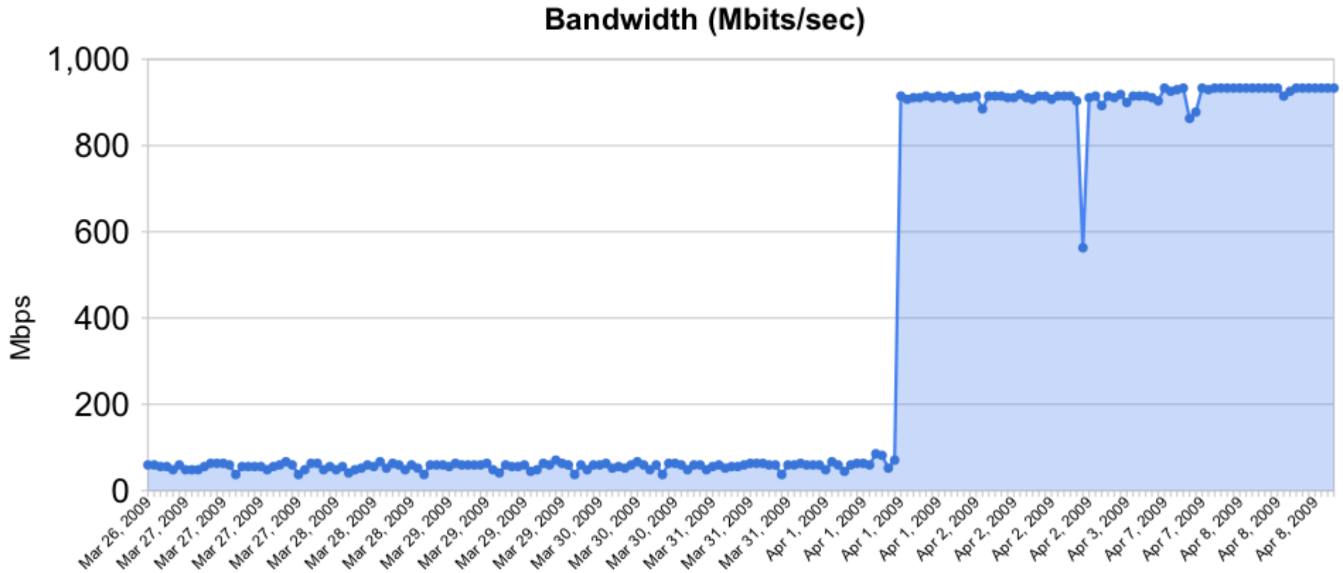
Small Queue Tail Drop

- 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

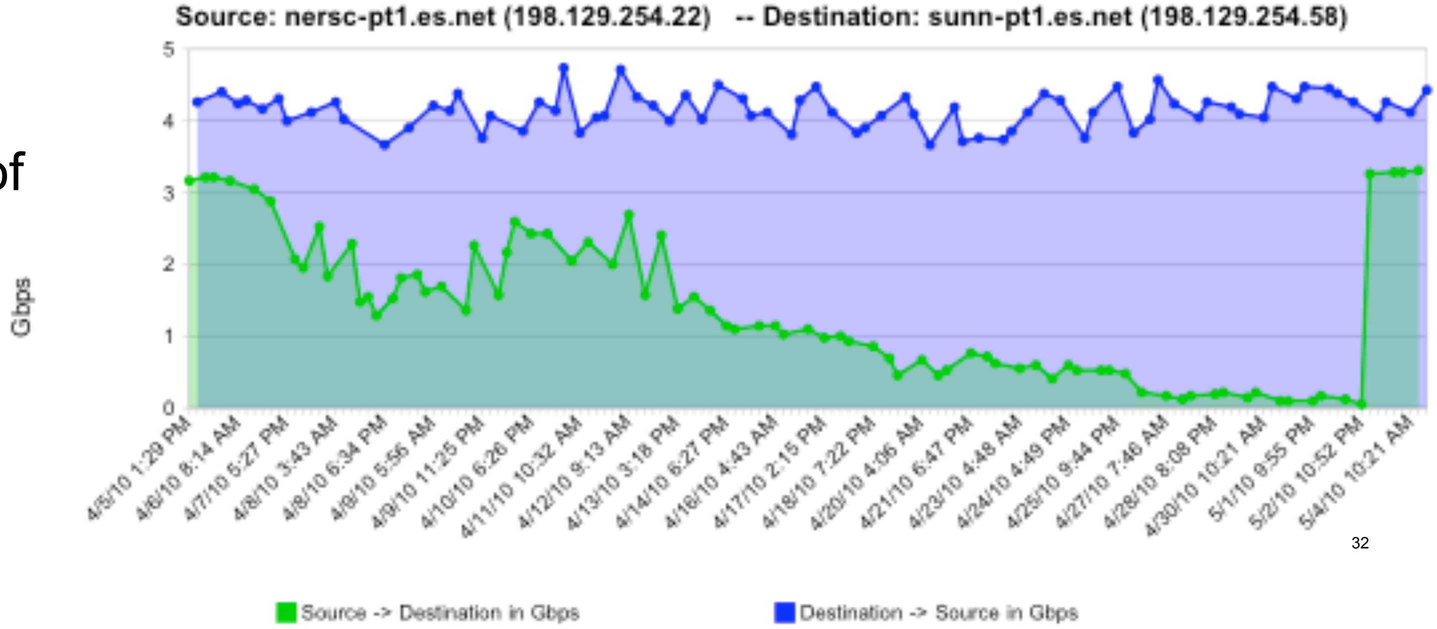
- Processor-based switching on routers due to faults, acl's, or mis-configuration
- Security Devices
  - E.g.: 10X improvement by turning off Cisco Reflexive ACL

Sample Results: Finding/Fixing soft failures



Rebooted router with full route table

Gradual failure of optical line card



perfSONAR Overview

Addressing the Problem: perfSONAR

perfSONAR - an open, web-services-based framework for:

- running network tests
- collecting and publishing measurement results

ESnet and Internet2 are:

- Deploying the framework across the science community
- Encouraging people to deploy ‘known good’ measurement points near domain boundaries
  - “known good” = hosts that are well configured, enough memory and CPU to drive the network, proper TCP tuning, clean path, etc.
- Using the framework to find and correct soft network failures.

Who is perfSONAR?

The perfSONAR Consortium is a joint collaboration between

- ESnet
- Géant
- Internet2
- Rede Nacional de Ensino e Pesquisa (RNP)

Decisions regarding protocol development, software branding, and interoperability are handled at this organization level

There are at least two independent efforts to develop software frameworks that are perfSONAR compatible.

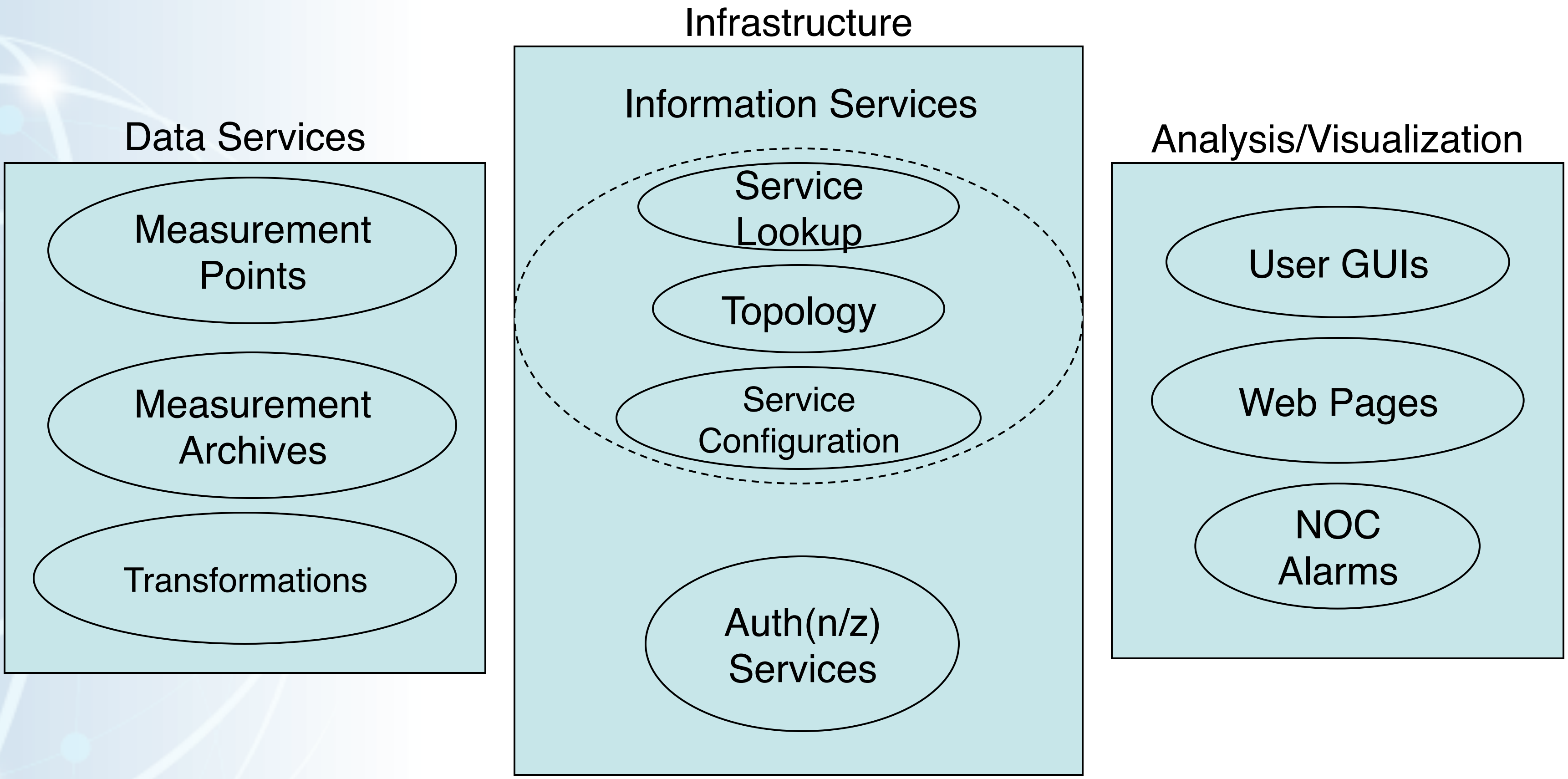
- perfSONAR-MDM
- perfSONAR-PS

Each project works on an individual development roadmap and works with the consortium to further protocol development and insure compatibility

perfSONAR Terminology

- perfSONAR: standardized schema, protocols, APIs
- perfSONAR-MDM: GÉANT Implementation and deployment
  - aimed at NRENS
- perfSONAR-PS: ESnet/Internet2 implementation and deployment
  - aimed at end-users and network admins (site and backbone)
- perfSONAR Performance Toolkit
  - Easy to install Packaging of perfSONAR-PS
  - “network install” and “LiveCD” versions

perfSONAR Architecture Overview



perfSONAR Services

PS-Toolkit includes these measurement tools:

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

Measurement Archives (data publication)

- SNMP MA – Interface Data
- pSB MA -- Scheduled bandwidth and latency data

Lookup Service

- gLS – Global lookup service used to find services
- hLS – Home lookup service for registering local perfSONAR metadata

PS-Toolkit includes these Troubleshooting Tools

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

perfSONAR-PS Utility

perfSONAR-PS appeals to both network users and operators:

- Operators:
  - Easy deployment
  - Minimal maintenance
  - Results relevant to common problems (e.g. connectivity loss, equipment failure, performance problems)
- Users:
  - Immediate access to network data
  - Cross domain capabilities

Adoption is spreading to networks of all sizes

The perfSONAR-PS framework has two primary high level use cases:

- Diagnostic (e.g. on-demand) use
- Monitoring Infrastructure

perfSONAR-PS Utility - Diagnostics

The pS Performance Toolkit was designed for diagnostic use and regular monitoring

- All tools preconfigured
- Minimal installation requirements
- Can deploy multiple instances for short periods of time in a domain

perfSONAR-PS Utility - Monitoring

Regular monitoring is an important design consideration for perfSONAR-PS tools

- perfSONAR-BUOY and PingER provide scheduling infrastructure to create regular latency and bandwidth tests
- The SNMP MA integrates with COTS SNMP monitoring solutions

The pSPT is capable of organizing and visualizing regularly scheduled tests

NAGIOS can be integrated with perfSONAR-PS tools to facilitate alerting to potential network performance degradation

Global PerfSONAR-PS Deployments

Based on “global lookup service” (gLS) registration, Dec 2011: currently deployed in over 150 locations

- ~ 275 bwctl and owamp servers
- ~ 230 active probe measurement archives
- ~ 25 SNMP measurement archives
- Countries include: USA, Australia, Hong Kong, Argentina, Brazil, Japan, China, Canada, Netherlands, Switzerland, Finland, Sweden, Italy, France, Pakistan

US Atlas Deployment

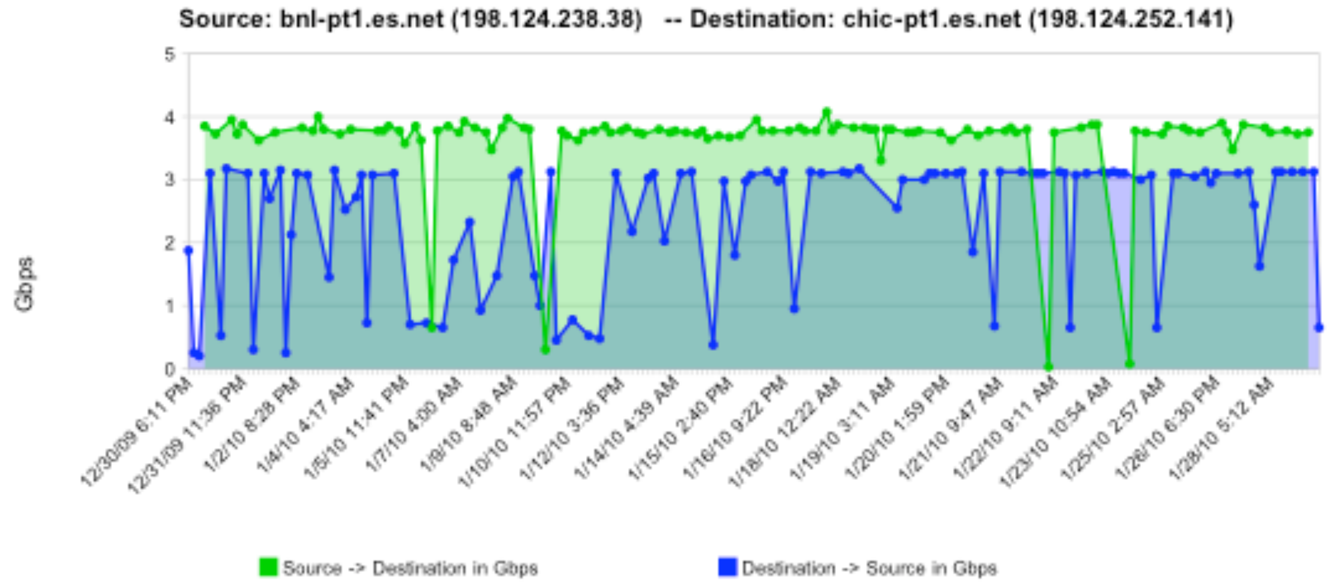
- Monitoring all “Tier 1 to Tier 2” connections

For current list of public services, see:

- http://stats.es.net/perfSONAR/directorySearch.html
- Many more “private” perfSONAR nodes deployed

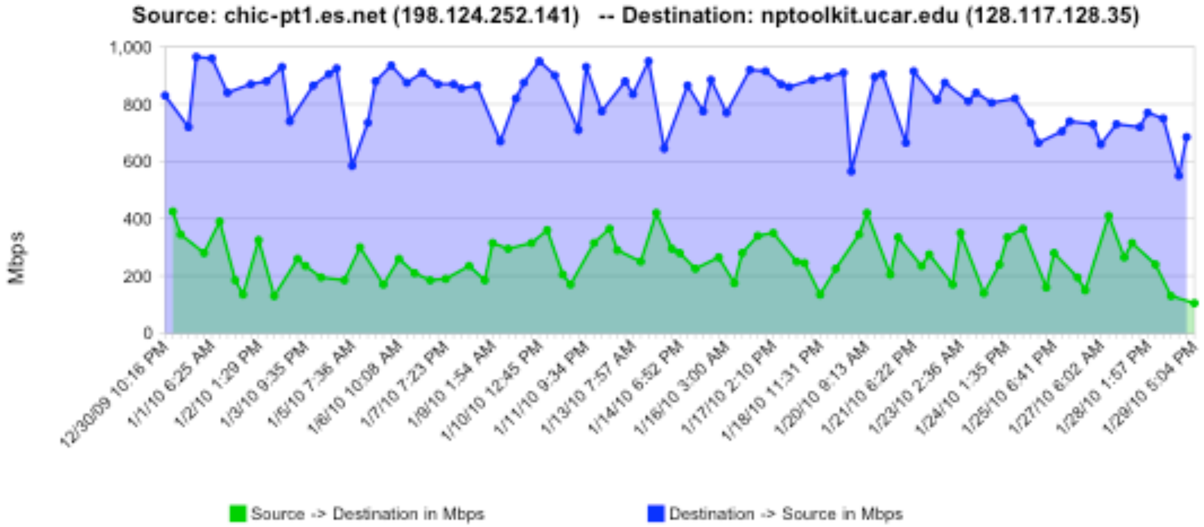
perfSONAR Case Studies

Sample Results: Throughput tests



Heavily used path: probe traffic is “scavenger service”

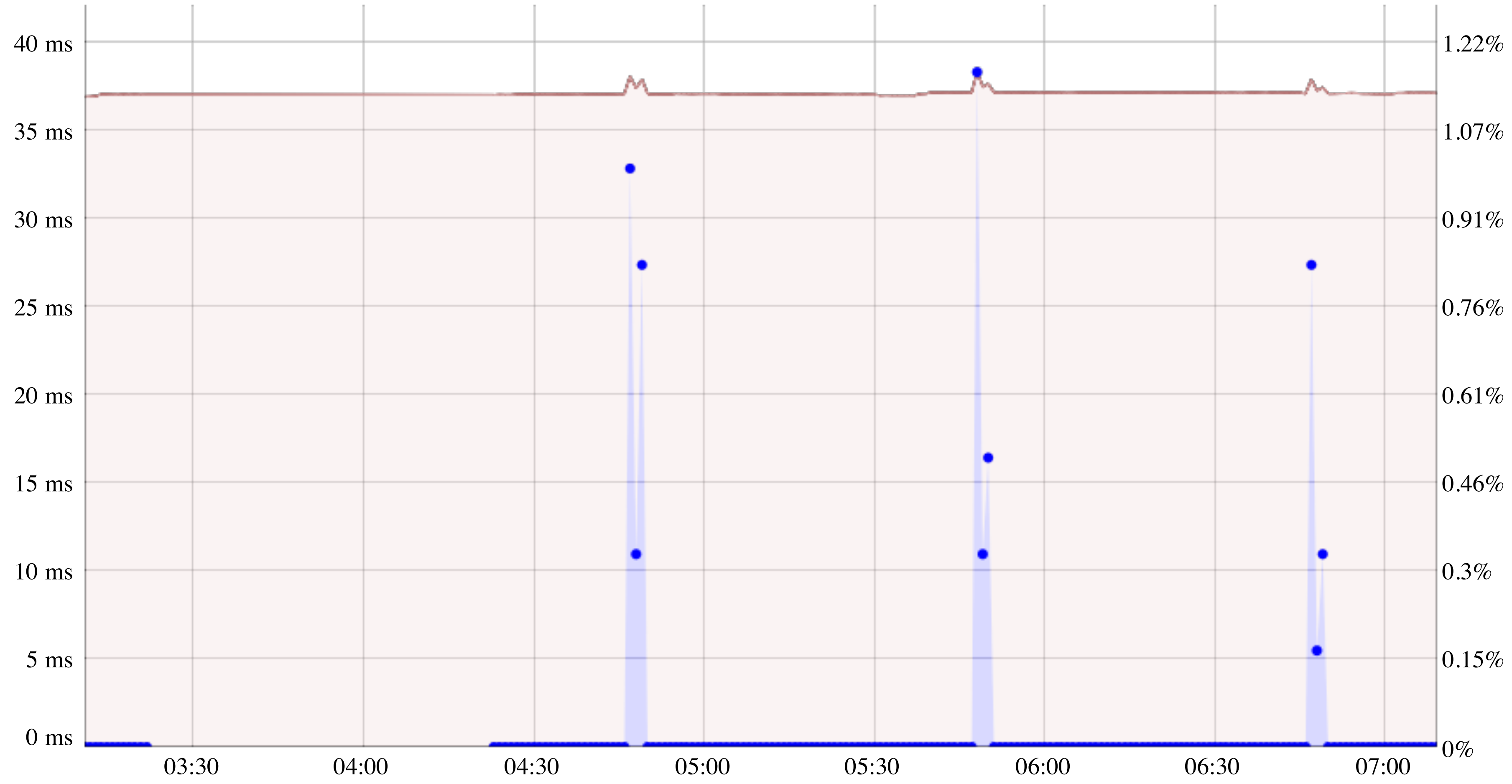
Asymmetric Results: different TCP stacks?



Sample Results: Latency/Loss Data

Source: ps-lat.es.net (198.129.254.187) -- Destination: bost-owamp.es.net (198.124.238.58)

One Way Delay



Timezone: PST

Common 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:

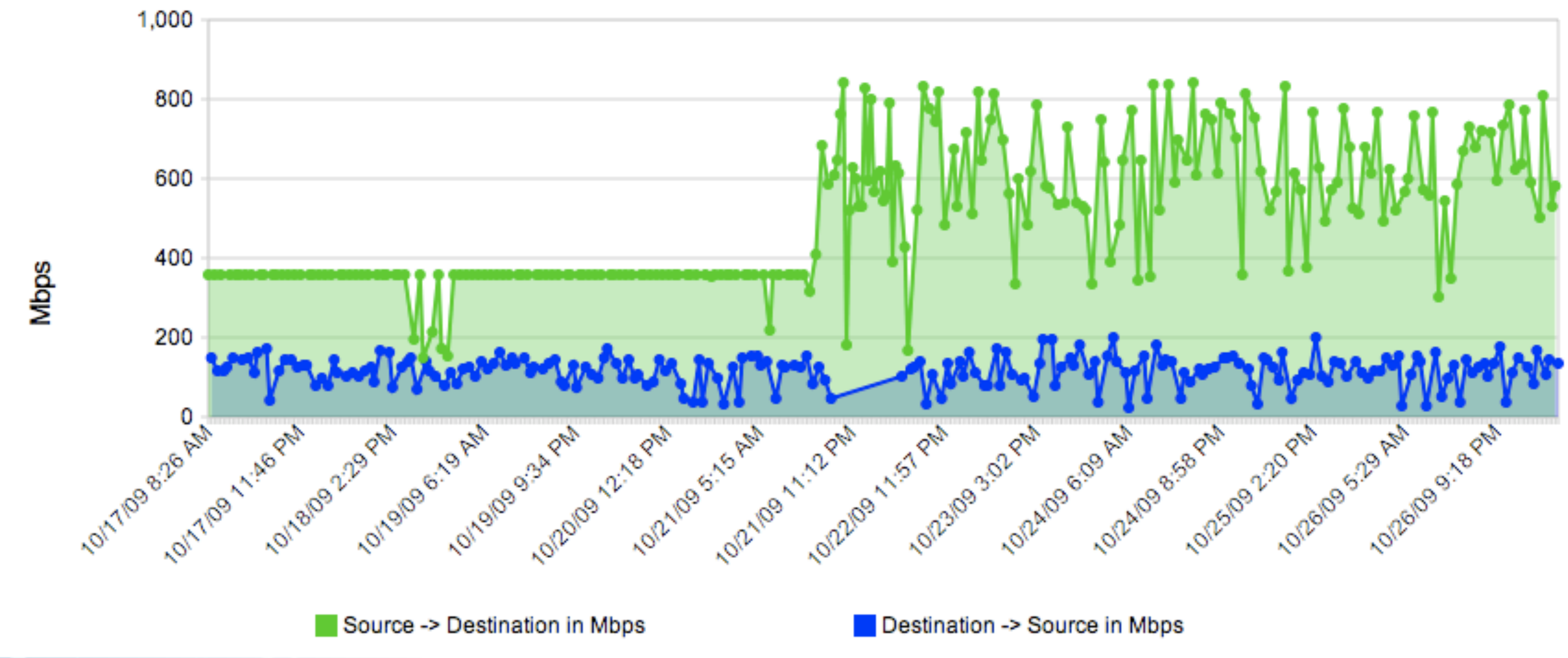
- Run tests between perfSONAR nodes
  - performance is often clean
- Run tests from end hosts to perfSONAR host at site border
  - Often find packet loss (using owamp tool)
  - If not, problem is often the host tuning or the disk

If there is not a perfSONAR node at each site border

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

REDDnet Use Case – Host Tuning

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



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

Troubleshooting Example: LLNL to BADC (Rutherford Lab, UK)

User trying to send climate data from LLNL (CA, USA) to BADC (U.K.) reports terrible performance (< 30 Mbps) in 1 direction, good performance (700 Mbps) in the other direction

Network Path used:

- ESnet to AofA (aofa-cr2.es.net): bwctl testing from llnl-pt1.es.net to aofa-pt1.es.net:
  - 5 Gbps both directions
- GÉANT2 to UK via Amsterdam: bwctl tests llnl-pt1.es.net to london.geant2.net:
  - 800 Mbps both directions
  - Testing to GÉANT perfSONAR node in London critical to rule out trans-Atlantic issues
- JANET to Rutherford lab
  - no bwctl host ☹, but used router filter packet counters to verify no packet loss in JANET

Suspect router buffer issue at RL, but very hard to prove without a perfSONAR hosts at Rutherford lab and in JANET

Problems finally solved once test hosts temporarily deployed in JANET and at RL (just-in-time deployment of test hosts makes troubleshooting *hard*)

Troubleshooting Example: Bulk Data Transfer between DOE Supercomputer Centers

Users were having problems moving data between supercomputer centers, NERSC and ORNL

- One user was: “waiting more than an entire workday for a 33 GB input file” (this should have taken < 15 min)

perfSONAR-PS measurement tools were installed

- Regularly scheduled measurements were started

Numerous choke points were identified & corrected

- Router tuning, host tuning, cluster file system tuning

Dedicated wide-area transfer nodes were setup

- Now moving 40 TB in less than 3 days

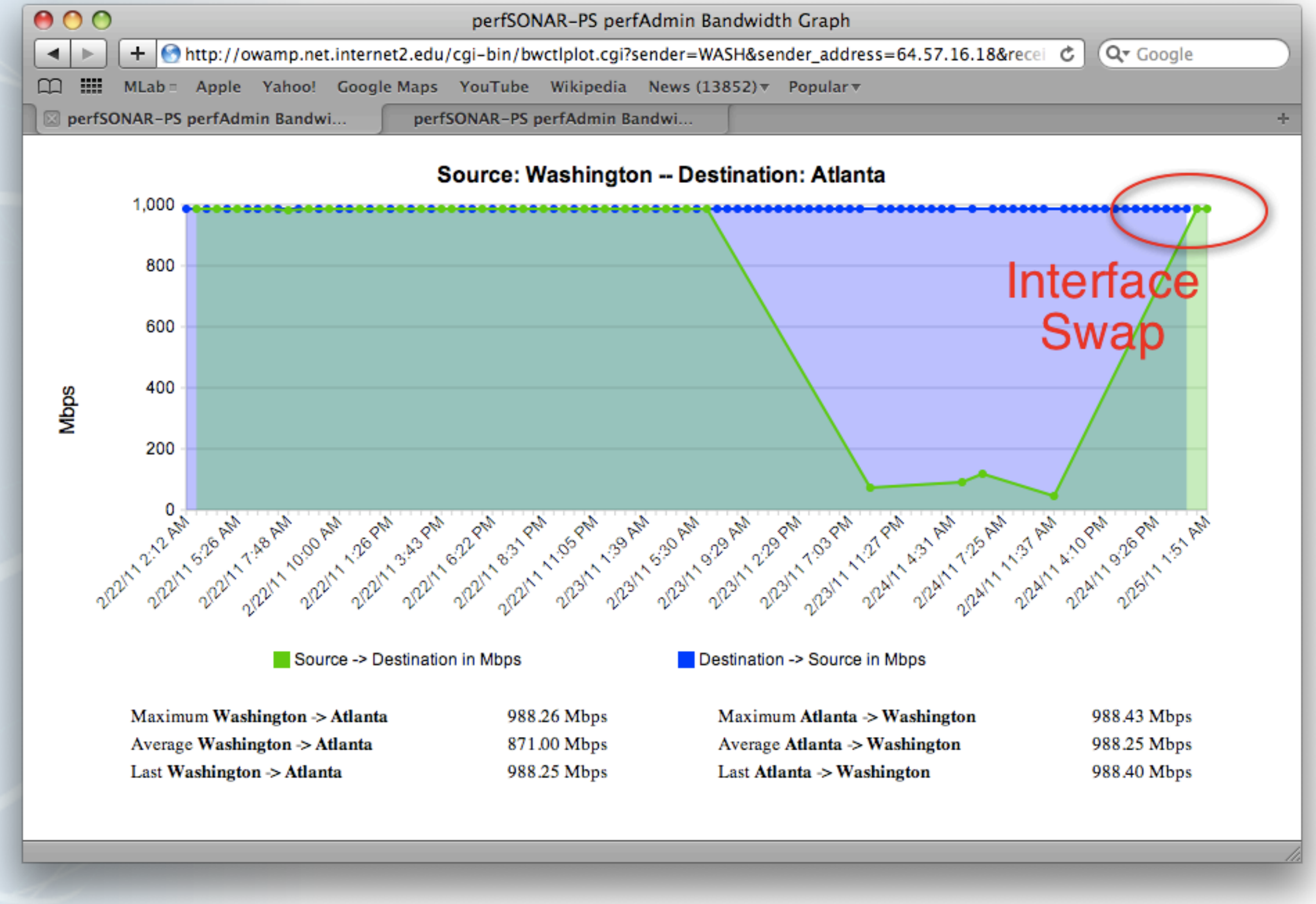
Troubleshooting Example: China to US

Difficulty getting science data moved from neutrino detectors in China to analysis in US

- Multiple difficulties (host config, packet loss, etc.)
- Installed perfSONAR-PS host in Hong Kong
  - Regular tests were started
  - Over time, multiple issues discovered and corrected, and performance improved
  - Performance went from 3Mbps to 200Mbps
- Automated testing over time provided several advantages
  - Performance problems can be correlated with network events
    - Path changes; Hardware failures; Host-level changes
  - Sometimes difficult to convince some entities that they have problems to fix without proof

Internet2 Backbone Example

bwctl results

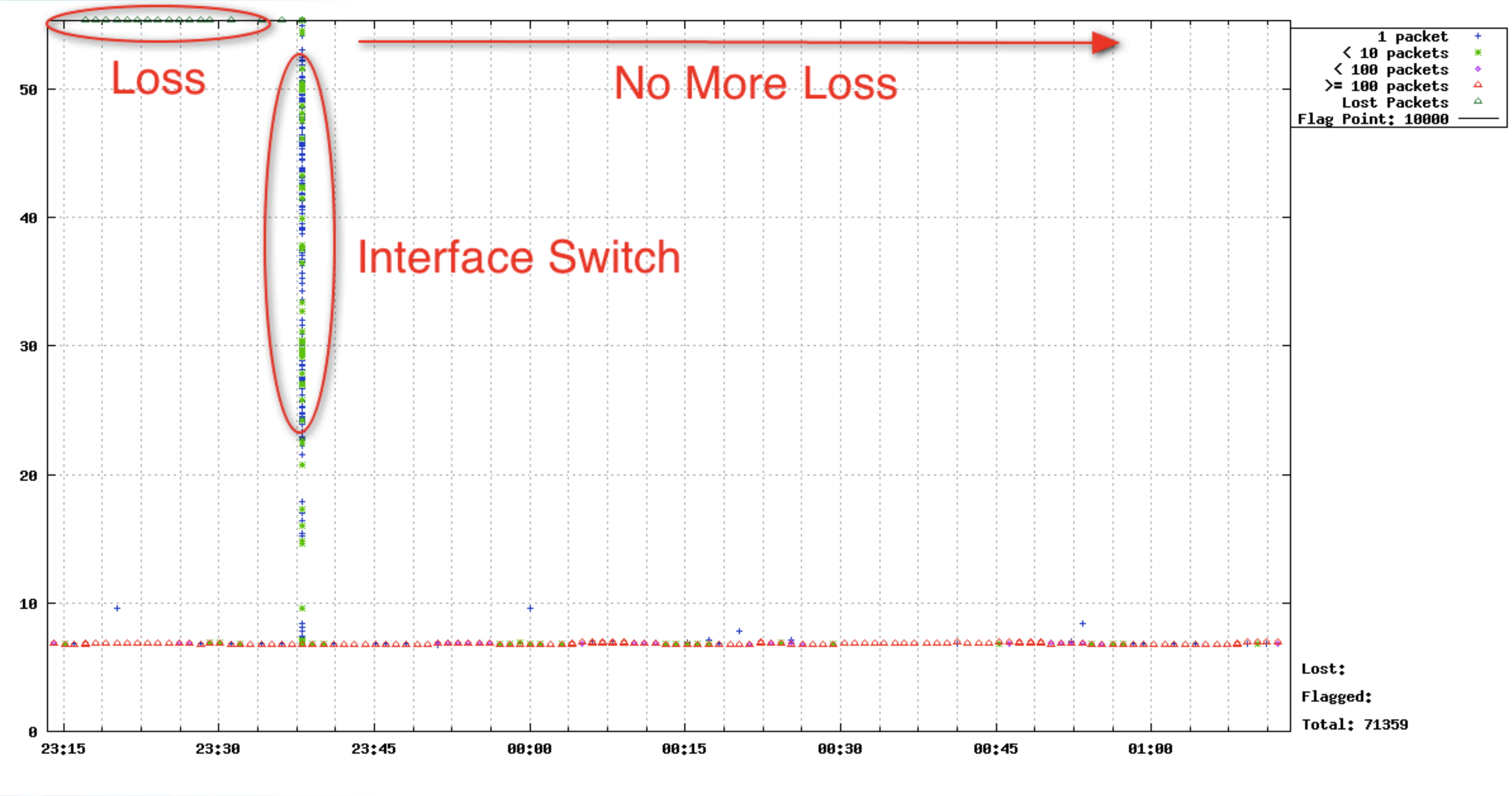


Interface Swap

Maximum **Washington -> Atlanta**	988.26 Mbps	Maximum **Atlanta -> Washington**	988.43 Mbps
Average **Washington -> Atlanta**	871.00 Mbps	Average **Atlanta -> Washington**	988.25 Mbps
Last **Washington -> Atlanta**	988.25 Mbps	Last **Atlanta -> Washington**	988.40 Mbps

Internet2 Backbone Example

Owamp data plot



Loss

No More Loss

Interface Switch

Effective perfSONAR Deployment Strategies

Levels of perfSONAR deployment

ESnet classifies perfSONAR deployments into 3 "levels":

Level 1: Run a bwctl server that is registered in the perfSONAR Lookup Service.

- This allows remote sites and ESnet engineers to run tests to your site.

Level 2: Configure "perfSONAR BOUY" to run regularly scheduled tests to/from your host.

- This allows you to establish a performance baseline, and to determine when performance changes.

Level 3: Full set of perfSONAR services deployed (everything on the PS Performance Toolkit)

perfSONAR-PS Software

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

- written in the perl programming language

http://software.internet2.edu/pS-Performance_Toolkit/

All products are available as RPMs.

The perfSONAR-PS consortium supports CentOS (version 5).

RPMs are compiled for i386 architecture, but work w/ x86 64 bit too

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

Deploying perfSONAR-PS Tools In Under 30 Minutes

There are two easy ways to deploy a perfSONAR-PS host

“Level 1” perfSONAR-PS install:

- Build a Linux machine as you normally would (configure TCP properly! See: http://fasterdata.es.net/TCP-tuning/)
- Go through the Level 1 HOWTO
- http://fasterdata.es.net/ps_level1_howto.html
  - Includes bwctl.limits file to restrict to R&E networks only
- Simple, fewer features, runs on a standard Linux build

Use the perfSONAR-PS Performance Toolkit netinstall CD

- Most of the configuration via Web GUI
- http://psps.perfsonar.net/toolkit/
- Includes more features (perfSONAR level 3)

Measurement Recommendations for end sites

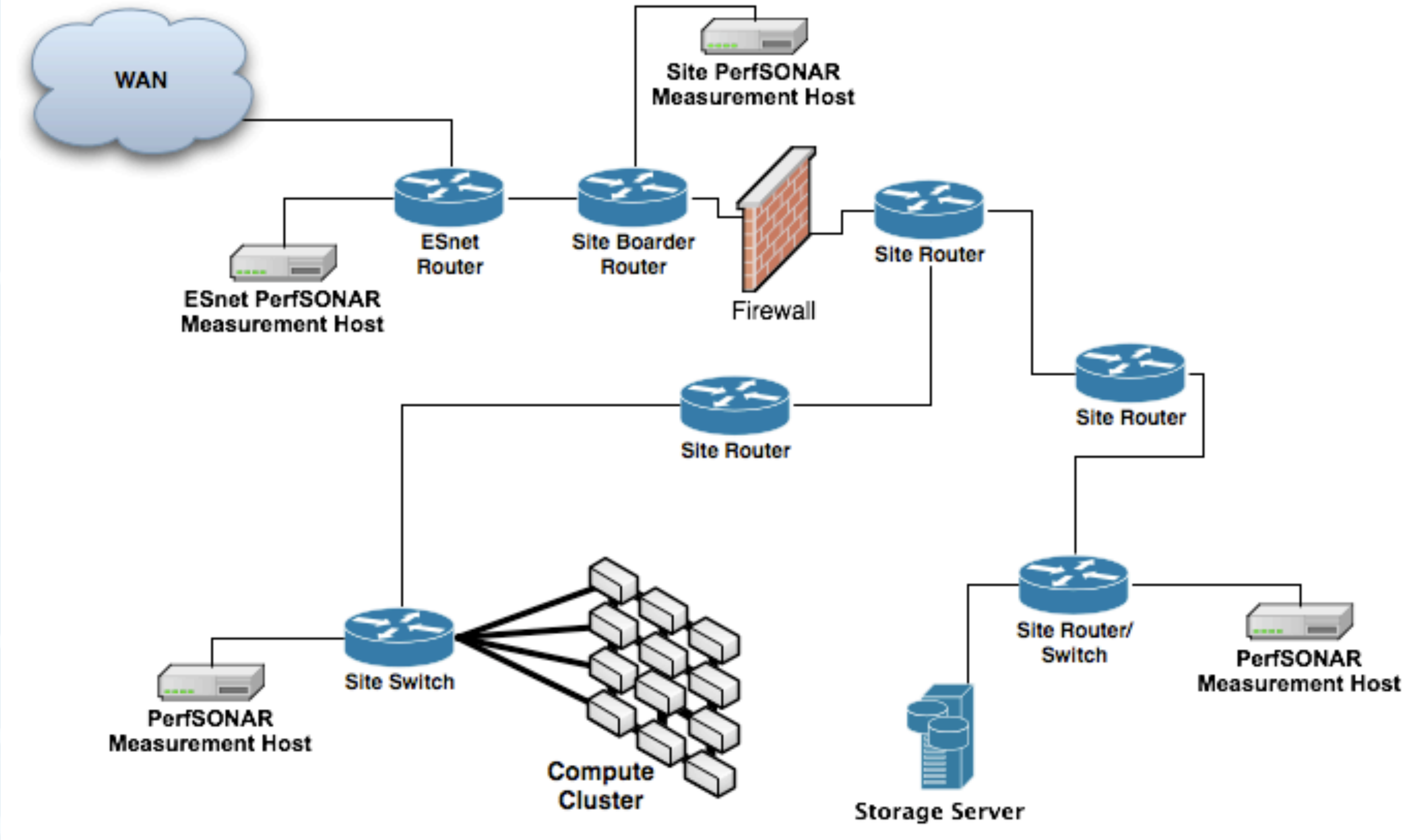
Deploy perfSONAR-PS based test tools

- At Site border
  - Use to rule out WAN issues
- Near important end systems and all DTNs
  - Use to rule out LAN issues

Use it to:

- Find & fix current local problems
- Identify when they re-occur
- Set user expectations by quantifying your network services

Sample Site Deployment



Importance of Regular Testing

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

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

Developing a Measurement Plan

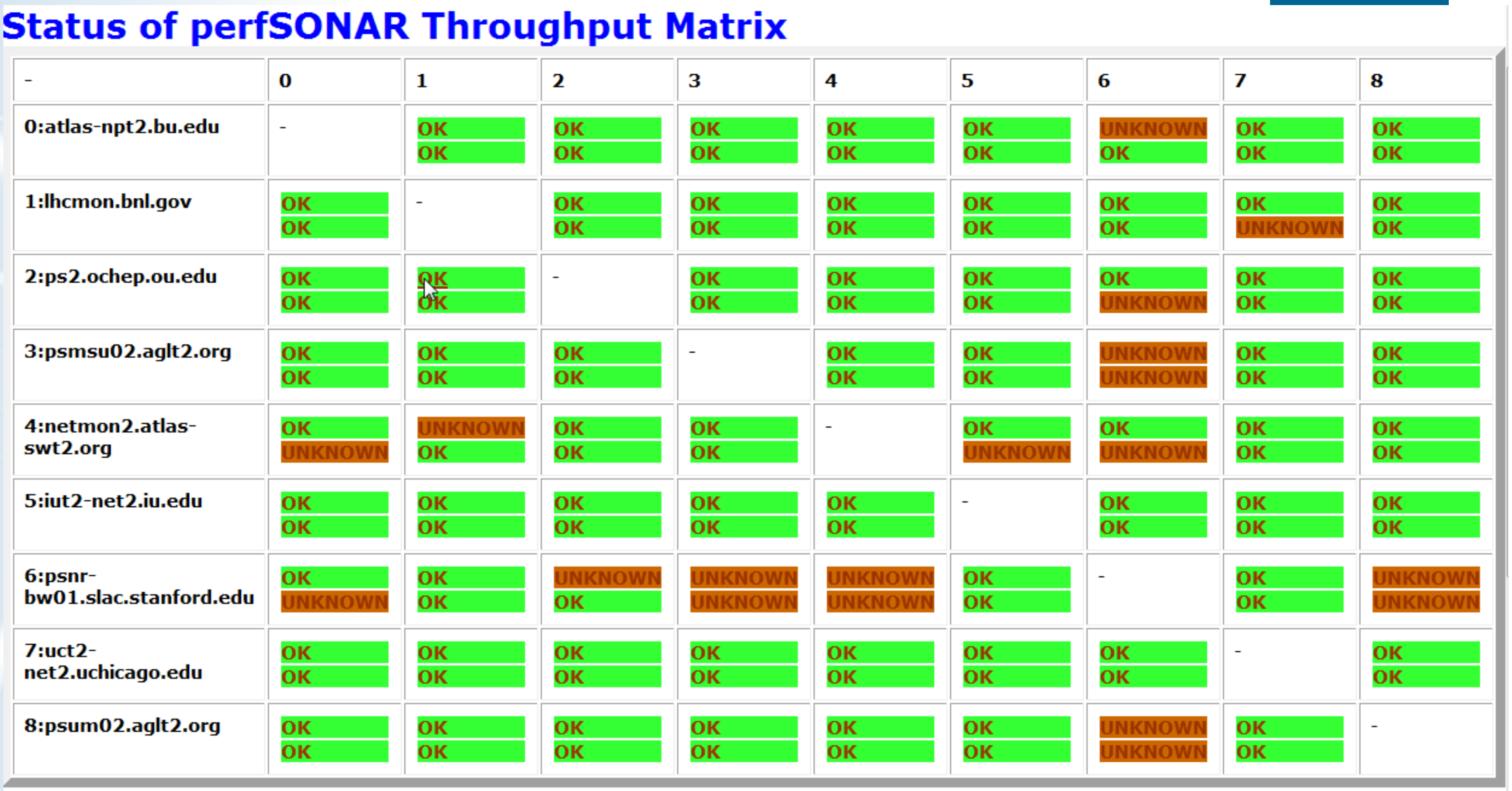
What are you going to measure?

- Achievable bandwidth
  - 2-3 regional destinations
  - 4-8 important collaborators
  - 4-10 times per day to each destination
  - 20 second tests within a region, longer across the Atlantic or Pacific
- Loss/Availability/Latency
  - OWAMP: ~10 collaborators over diverse paths
  - PingER: use to monitor paths to collaborators who don’t support owamp
- Interface Utilization & Errors

What are you going to do with the results?

- NAGIOS Alerts
- Reports to user community
- Post to Website

Sample tool: Atlas perfSONAR Dashboard



Status of perfSONAR Throughput Matrix

-	0	1	2	3	4	5	6	7	8
0:atlas-npt2.bu.edu	-	OK OK	OK OK	OK OK	OK OK	OK OK	UNKNOWN OK	OK OK	OK OK
1:lhcmon.bnl.gov	OK OK	-	OK OK	OK OK	OK OK	OK OK	OK OK	OK UNKNOWN	OK OK
2:ps2.ochep.ou.edu	OK OK	OK OK	-	OK OK	OK OK	OK OK	OK UNKNOWN	OK OK	OK OK
3:psmsu02.aglt2.org	OK OK	OK OK	OK OK	-	OK OK	OK OK	UNKNOWN UNKNOWN	OK OK	OK OK
4:netmon2.atlas-swt2.org	OK UNKNOWN	UNKNOWN OK	OK OK	OK OK	-	OK UNKNOWN	OK UNKNOWN	OK OK	OK OK
5:iut2-net2.iu.edu	OK OK	OK OK	OK OK	OK OK	OK OK	-	OK OK	OK OK	OK OK
6:psnr-bw01.slac.stanford.edu	OK UNKNOWN	OK OK	UNKNOWN OK	UNKNOWN UNKNOWN	UNKNOWN UNKNOWN	OK OK	-	OK OK	UNKNOWN UNKNOWN
7:uct2-net2.uchicago.edu	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	-	OK OK
8:psum02.aglt2.org	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	UNKNOWN UNKNOWN	OK OK	-

perfSONAR Security models

Security and Privacy Issues with perfSONAR

The ESnet viewpoint is that perfSONAR services should be as open as possible

We make all of the following publically accessible via perfSONAR:

- all SNMP data on utilization, errors, drops
- All topology data

Anyone from an R&E network anywhere in the world can run bwctl tests to our servers

- TCP tests limited to 120 seconds
- UDP tests limited to 200 Mbps, 600 seconds

ESnet has had no security related issues since we deployed perfSONAR 5 years ago.

Commonly heard Security Concerns

DDOS attack using bwctl:

- bwctl has controls to limit test duration, UDP rates, allow subnets
- ESnet provides a bwctl control file with only R&E networks, updated nightly

SNMP utilization data is sensitive information

- maybe for the military, but we don’t think so for R&E

perfSONAR Host Recommendations

Host Considerations

Dedicated perfSONAR hardware is best

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

1Gbps vs 10Gbps testers

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

Virtual Machines do not work well for perfSONAR hosts

- clock sync issues
- throughput is reduced significantly for 10G hosts
- caveat: this has not been tested recently, and VM technology and motherboard technology has come a long way

Sample Host Configuration #1

10G throughput host: 1U, RAID disk and dual power supplies for reliability, on board IPMI: ($3000 USD)

- Intel Xeon 2.66GHz 4 Cores Processor
- (2) 4GB Modules Kingston Brand DDRIII 1333 ECC
- (2) 500GB WD SATA II Drive Enterprises
- 3Ware 9650SE-4LP 4 Ports with BBU Installed
- Myricom 10G-PCIE-8B-S

Sample Host Configuration #2

1G Host deployed by the US Atlas project in 2008:

- Intel Pentium DC E2200 2.4GHz 1MB 800MHz Processor
- Intel 945GC/ICH7 Chipset Main Board
- Onboard Marvel 8056 GbE LAN Controller
- 2GB DDR2-5300 RAM 667MHz Non-ECC Unbuffered
- 160GB SATA 7200RPM Hard Drive
- $650 USD

Perfect for a latency host or a 1G tester, no redundancy however

perfSONAR Summary

Soft failures are everywhere

We all need to look for them, and not wait for users to complain

perfSONAR is MUCH more useful when its on every segment of the end-to-end path

Ideally all networks and high BW end sites to deploy at least a “level 1” host

10G test hosts are needed to troubleshoot 10G problems

perfSONAR is MUCH more useful when its open

locking it down behind firewalls/ACLs defeats the purpose

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

More Information

Download the perfSONAR performance Toolkit:

- http://software.internet2.edu/pS-Performance_Toolkit/

ESnet network performance troubleshooting guide:

- http://fasterdata.es.net/fasterdata/troubleshooting/overview/

Information on downloading/installing perfSONAR

- http://fasterdata.es.net/fasterdata/perfSONAR/

Graphs of ESnet perfSONAR data:

- http://stats.es.net/

Slides from recent full day perfSONAR workshop from Internet2

- http://www.internet2.edu/workshops/npw/roster/learn-11.cfm

email: BLTierney@es.net

Extra Slides

Socket Buffer Autotuning

To solve the buffer tuning problem, based on work at LANL and PSC, Linux OS added TCP Buffer autotuning

- Sender-side TCP buffer autotuning introduced in Linux 2.4
- Receiver-side autotuning added in Linux 2.6

Most OS's now include TCP autotuning

- TCP send buffer starts at 64 KB
- As the data transfer takes place, the buffer size is continuously re-adjusted up max autotune size

Current OS Autotuning default maximum buffers

- Linux 2.6: 256K to 4MB, depending on version
- Windows Vista: 16M
- Mac OSX 10.5-10.6: 4M
- FreeBSD 7 and 8: 256K