August 11th 2014, APAN38

Network Performance Tutorial

John Hicks – Internet2

XSEDE Performance Use Cases

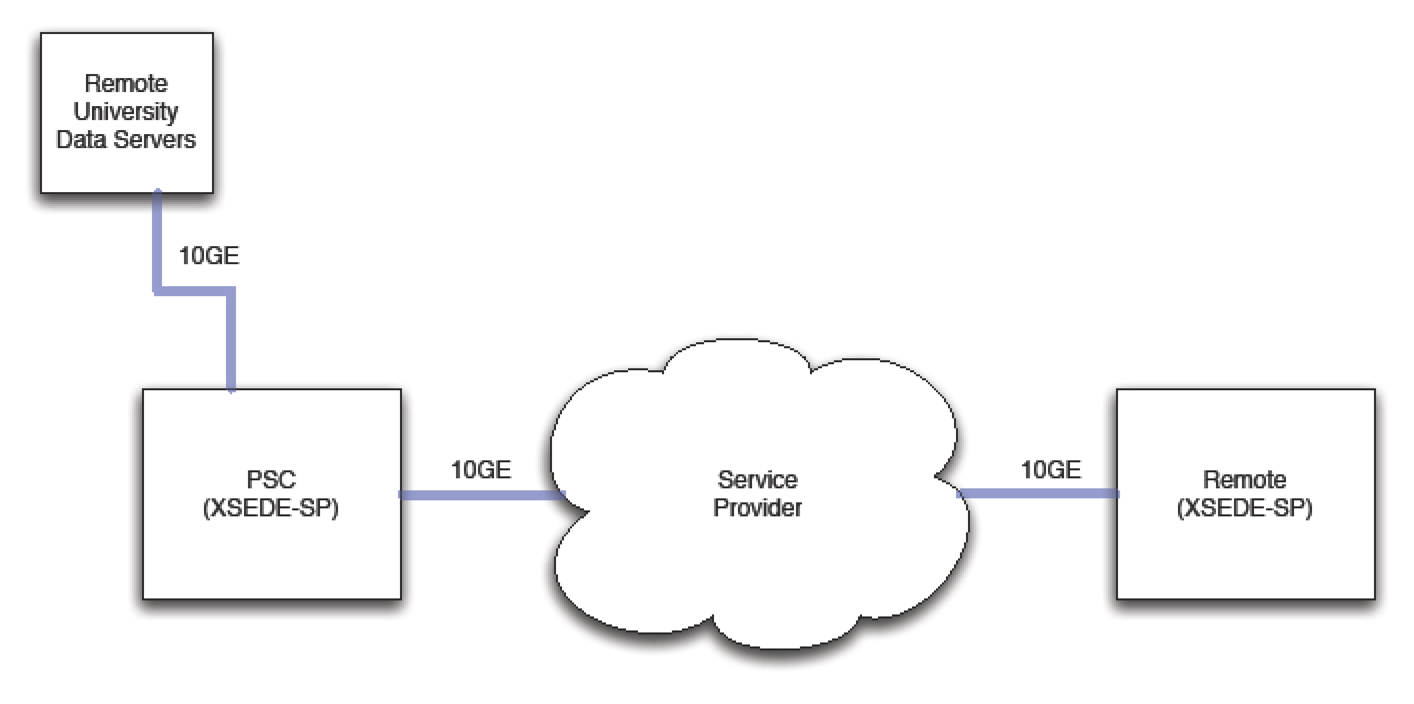
XSEDE Use Case Examples

- Debugging is simplified by the limited number of domains and the ongoing working relationship between network engineers at all sites
- XSEDE perfSONARs are not set up to alarm on conditions so current usage mode is primarily as a debugging resource when problems are noted
- Use case examples from XSEDE:
  - Campus integration case study
  - Jumbo frame MTU issues
  - Impact of small router buffers
  - Route changes

Campus integration case study

- The “case study” example describes debugging steps based on a true story
- Three institutions were involved: one University, PSC, and an additional XSEDE Service Provider (XSEDE-SP2)
- Initial network testing throughput was much lower than expected
- Debugging approach
  - Step by step
  - What to do without perfSONAR
  - How to take advantage of perfSONAR

Campus integration case study

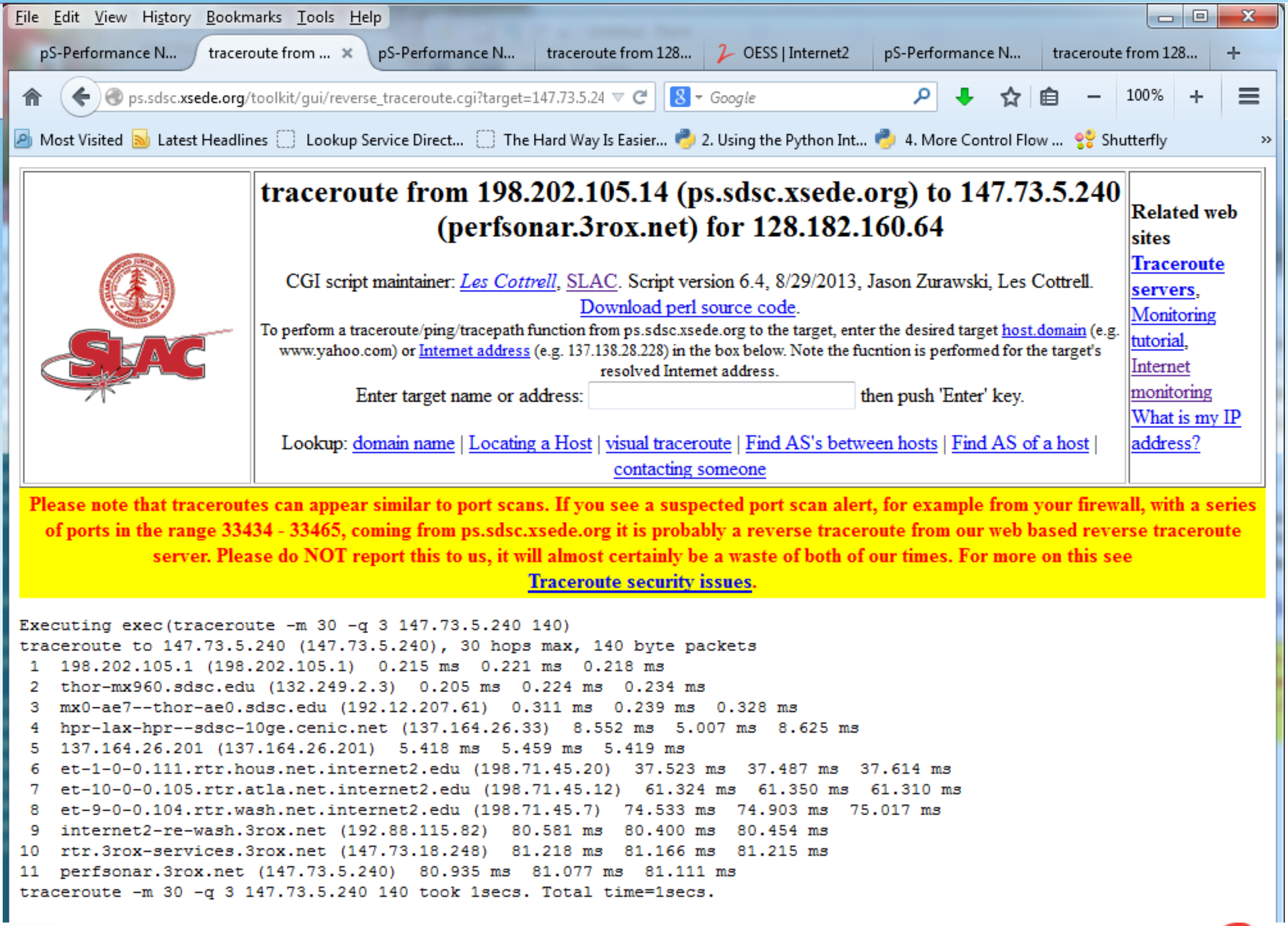


Initial conditions

- Primary direction of data flow was University -> PSC and University -> XSEDE-SP2
- Between 1 GbE connected hosts over a 10 GbE link:
  - University -> PSC maximum was 220 Mb/sec
  - University -> XSEDE-SP2 maximum was 13.8 Mb/sec

Check the Path

- Manually run traceroute from each end if no perfSONAR is available
  - Requires login access at both ends or a knowledgeable collaborator at the remote site
- With perfSONAR available and connected close to servers, use Reverse Traceroute
- In either case, traceroute is a necessary first step in debugging and access is usually available to the end user



traceroute from 198.202.105.14 (ps.sdsc.xsede.org) to 147.73.5.240 (perfsonar.3rox.net) for 128.182.160.64

CGI script maintainer: *Les Cottrell*, SLAC. Script version 6.4, 8/29/2013, Jason Zurawski, Les Cottrell.
Download perl source code.

To perform a traceroute/ping/tracepath function from ps.sdsc.xsede.org to the target, enter the desired target host.domain (e.g. www.yahoo.com) or Internet address (e.g. 137.138.28.228) in the box below. Note the fucntion is performed for the target's resolved Internet address.

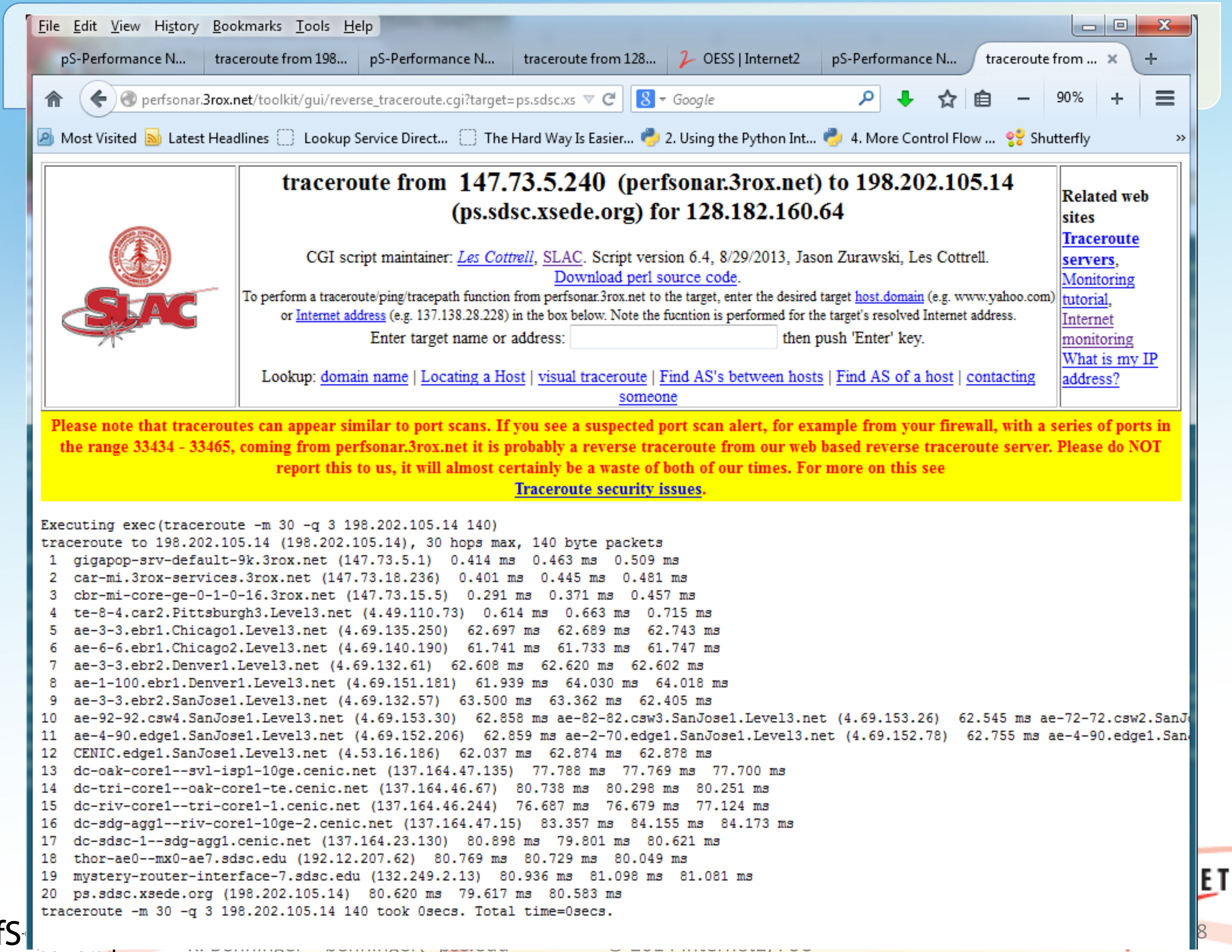
Enter target name or address: then push 'Enter' key.

Lookup: domain name | Locating a Host | visual traceroute | Find AS's between hosts | Find AS of a host | contacting someone

Related web sites
Traceroute servers,
Monitoring tutorial,
Internet monitoring
What is my IP address?

Please note that traceroutes can appear similar to port scans. If you see a suspected port scan alert, for example from your firewall, with a series of ports in the range 33434 - 33465, coming from ps.sdsc.xsede.org it is probably a reverse traceroute from our web based reverse traceroute server. Please do NOT report this to us, it will almost certainly be a waste of both of our times. For more on this see Traceroute security issues.

```
Executing exec(traceroute -m 30 -q 3 147.73.5.240 140)
traceroute to 147.73.5.240 (147.73.5.240), 30 hops max, 140 byte packets
 1  198.202.105.1 (198.202.105.1)  0.215 ms  0.221 ms  0.218 ms
 2  thor-mx960.sdsc.edu (132.249.2.3)  0.205 ms  0.224 ms  0.234 ms
 3  mx0-ae7--thor-ae0.sdsc.edu (192.12.207.61)  0.311 ms  0.239 ms  0.328 ms
 4  hpr-lax-hpr--sdsc-10ge.cenic.net (137.164.26.33)  8.552 ms  5.007 ms  8.625 ms
 5  137.164.26.201 (137.164.26.201)  5.418 ms  5.459 ms  5.419 ms
 6  et-1-0-0.111.rtr.hous.net.internet2.edu (198.71.45.20)  37.523 ms  37.487 ms  37.614 ms
 7  et-10-0-0.105.rtr.atla.net.internet2.edu (198.71.45.12)  61.324 ms  61.350 ms  61.310 ms
 8  et-9-0-0.104.rtr.wash.net.internet2.edu (198.71.45.7)  74.533 ms  74.903 ms  75.017 ms
 9  internet2-re-wash.3rox.net (192.88.115.82)  80.581 ms  80.400 ms  80.454 ms
10  rtr.3rox-services.3rox.net (147.73.18.248)  81.218 ms  81.166 ms  81.215 ms
11  perfsonar.3rox.net (147.73.5.240)  80.935 ms  81.077 ms  81.111 ms
traceroute -m 30 -q 3 147.73.5.240 140 took 1secs. Total time=1secs.
```



traceroute from 147.73.5.240 (perfsonar.3rox.net) to 198.202.105.14 (ps.sdsc.xsede.org) for 128.182.160.64

CGI script maintainer: Les Cottrell, SLAC. Script version 6.4, 8/29/2013, Jason Zurawski, Les Cottrell.
Download perl source code.

To perform a traceroute/ping/tracepath function from perfsonar.3rox.net to the target, enter the desired target host.domain (e.g. www.yahoo.com) or Internet address (e.g. 137.138.28.228) in the box below. Note the fucntion is performed for the target's resolved Internet address.

Enter target name or address: ______ then push 'Enter' key.

Lookup: domain name | Locating a Host | visual traceroute | Find AS's between hosts | Find AS of a host | contacting someone

Related web sites
Traceroute servers, Monitoring tutorial, Internet monitoring What is my IP address?

Please note that traceroutes can appear similar to port scans. If you see a suspected port scan alert, for example from your firewall, with a series of ports in the range 33434 - 33465, coming from perfsonar.3rox.net it is probably a reverse traceroute from our web based reverse traceroute server. Please do NOT report this to us, it will almost certainly be a waste of both of our times. For more on this see Traceroute security issues.

```
Executing exec(traceroute -m 30 -q 3 198.202.105.14 140)
traceroute to 198.202.105.14 (198.202.105.14), 30 hops max, 140 byte packets
 1  gigapop-srv-default-9k.3rox.net (147.73.5.1)  0.414 ms  0.463 ms  0.509 ms
 2  car-mi.3rox-services.3rox.net (147.73.18.236)  0.401 ms  0.445 ms  0.481 ms
 3  cbr-mi-core-ge-0-1-0-16.3rox.net (147.73.15.5)  0.291 ms  0.371 ms  0.457 ms
 4  te-8-4.car2.Pittsburgh3.Level3.net (4.49.110.73)  0.614 ms  0.663 ms  0.715 ms
 5  ae-3-3.ebr1.Chicago1.Level3.net (4.69.135.250)  62.697 ms  62.689 ms  62.743 ms
 6  ae-6-6.ebr1.Chicago2.Level3.net (4.69.140.190)  61.741 ms  61.733 ms  61.747 ms
 7  ae-3-3.ebr2.Denver1.Level3.net (4.69.132.61)  62.608 ms  62.620 ms  62.602 ms
 8  ae-1-100.ebr1.Denver1.Level3.net (4.69.151.181)  61.939 ms  64.030 ms  64.018 ms
 9  ae-3-3.ebr2.SanJose1.Level3.net (4.69.132.57)  63.500 ms  63.362 ms  62.405 ms
10  ae-92-92.csw4.SanJose1.Level3.net (4.69.153.30)  62.858 ms ae-82-82.csw3.SanJose1.Level3.net (4.69.153.26)  62.545 ms ae-72-72.csw2.SanJ
11  ae-4-90.edge1.SanJose1.Level3.net (4.69.152.206)  62.859 ms ae-2-70.edge1.SanJose1.Level3.net (4.69.152.78)  62.755 ms ae-4-90.edge1.San
12  CENIC.edge1.SanJose1.Level3.net (4.53.16.186)  62.037 ms  62.874 ms  62.878 ms
13  dc-oak-core1--svl-isp1-10ge.cenic.net (137.164.47.135)  77.788 ms  77.769 ms  77.700 ms
14  dc-tri-core1--oak-core1-te.cenic.net (137.164.46.67)  80.738 ms  80.298 ms  80.251 ms
15  dc-riv-core1--tri-core1-1.cenic.net (137.164.46.244)  76.687 ms  76.679 ms  77.124 ms
16  dc-sdg-agg1--riv-core1-10ge-2.cenic.net (137.164.47.15)  83.357 ms  84.155 ms  84.173 ms
17  dc-sdsc-1--sdg-agg1.cenic.net (137.164.23.130)  80.898 ms  79.801 ms  80.621 ms
18  thor-ae0--mx0-ae7.sdsc.edu (192.12.207.62)  80.769 ms  80.729 ms  80.049 ms
19  mystery-router-interface-7.sdsc.edu (132.249.2.13)  80.936 ms  81.098 ms  81.081 ms
20  ps.sdsc.xsede.org (198.202.105.14)  80.620 ms  79.617 ms  80.583 ms
traceroute -m 30 -q 3 198.202.105.14 140 took 0secs. Total time=0secs.
```

Network throughput testing

- iperf or nuttcp
- If no perfSONARs or BWCTL servers, testing requires host access at each end
- Requires some specialized knowledge to identify endpoints and set appropriate parameters
- Typically used by network engineers for performance diagnosis

BWCTL for baseline bandwidth

- If you have login on a perfSONAR or a server running BWCTL (and if the relevant perfSONARs are not access restricted) **you** can manually run a BWCTL command:

```
[benninge@ps ~]$ bwctl -s ps.nics.xsede.org -t 30 -i 5 –f
```

- BWCTL supports third party test initiation

BWCTL example

```
[benninge@ps ~]$ bwctl -s ps.nics.xsede.org -t 30 -i 5 -f m
bwctl: Using tool: iperf
bwctl: 35 seconds until test results available

RECEIVER START
------------------------------------------------------------
Server listening on TCP port 5128
Binding to local address 128.182.112.220
TCP window size: 0.08 MByte (default)
------------------------------------------------------------
[ 15] local 128.182.112.220 port 5128 connected with 192.249.6.3 port 5128
[ ID] Interval       Transfer     Bandwidth
[ 15]  0.0- 5.0 sec  2877 MBytes  4827 Mbits/sec
[ 15]  5.0-10.0 sec  3226 MBytes  5412 Mbits/sec
[ 15] 10.0-15.0 sec  2519 MBytes  4227 Mbits/sec
[ 15] 15.0-20.0 sec  1780 MBytes  2987 Mbits/sec
[ 15] 20.0-25.0 sec  1897 MBytes  3183 Mbits/sec
[ 15] 25.0-30.0 sec  2010 MBytes  3372 Mbits/sec
[ 15]  0.0-30.1 sec  14355 MBytes  3999 Mbits/sec
[ 15] MSS size 8948 bytes (MTU 8988 bytes, unknown interface)

RECEIVER END
```

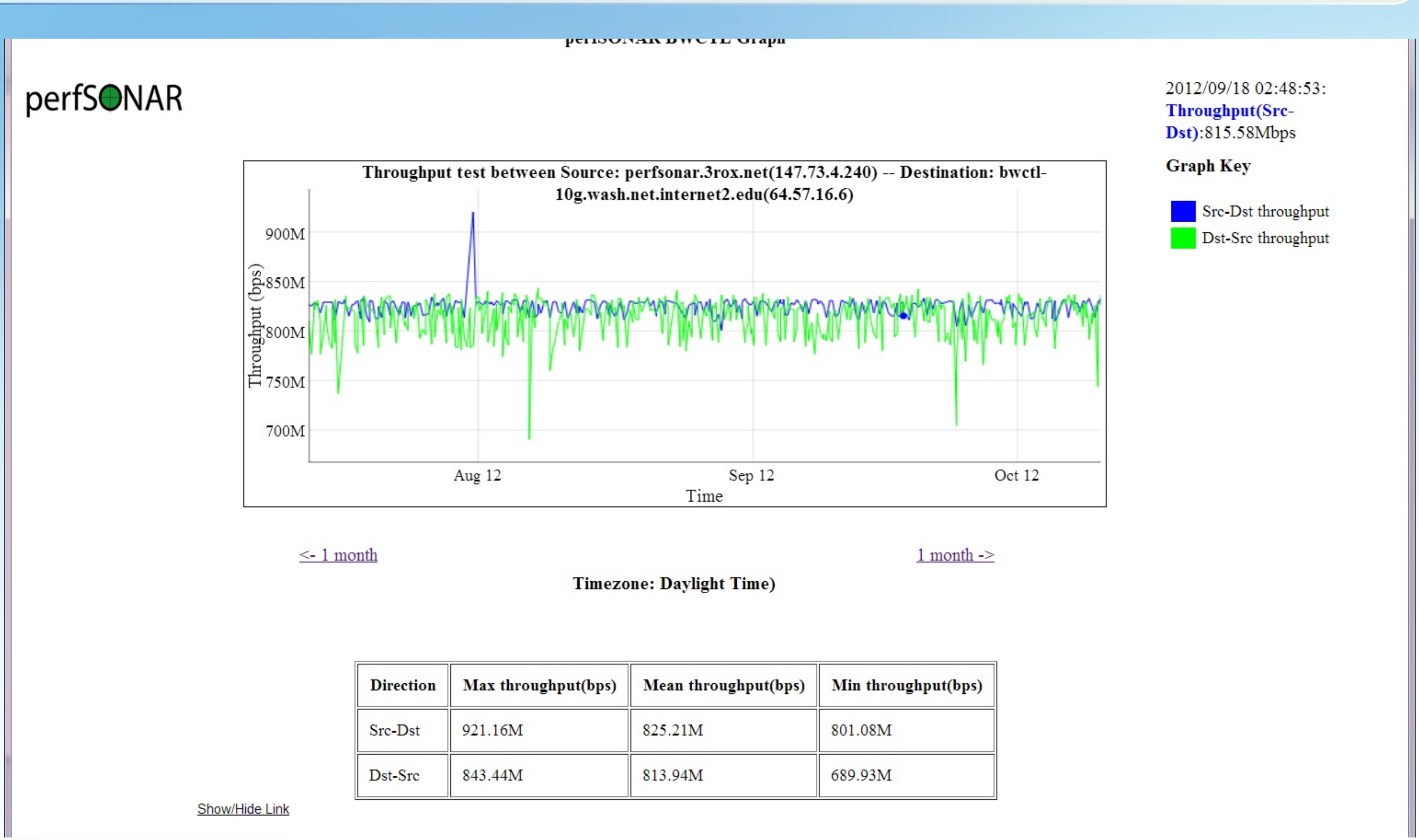
3rd party BWCTL example

```
[benninge@ps ~]$ bwctl -s ps.nics.xsede.org -c ps.iu.xsede.org -t 30 -i 5 -f m
bwctl: Using tool: iperf
bwctl: 37 seconds until test results available

RECEIVER START
------------------------------------------------------------
Server listening on TCP port 5047
Binding to local address 149.165.227.125
TCP window size: 0.08 MByte (default)
------------------------------------------------------------
[ 15] local 149.165.227.125 port 5047 connected with 192.249.6.3 port 5047
[ ID] Interval       Transfer     Bandwidth
[ 15]  0.0- 5.0 sec  5480 MBytes  9193 Mbits/sec
[ 15]  5.0-10.0 sec  5892 MBytes  9886 Mbits/sec
[ 15] 10.0-15.0 sec  5898 MBytes  9896 Mbits/sec
[ 15] 15.0-20.0 sec  5898 MBytes  9896 Mbits/sec
[ 15] 20.0-25.0 sec  5898 MBytes  9896 Mbits/sec
[ 15] 25.0-30.0 sec  5898 MBytes  9896 Mbits/sec
[ 15]  0.0-30.0 sec  34989 MBytes  9777 Mbits/sec
[ 15] MSS size 8948 bytes (MTU 8988 bytes, unknown interface)

RECEIVER END
```

BWCTL Scheduled Testing

- With login access on one of the perfSONARs at an end site, you can schedule testing to gather a performance picture throughout the day and across several days.
- Scheduled testing to intermediate hops will offer view of path segments
- Test scheduling will typically be done by network engineering staff who admin the perfSONAR systems

Scheduled throughput testing



perfSONAR BWCTL Graph

perfSONAR

2012/09/18 02:48:53: **Throughput(Src-Dst)**:815.58Mbps

Graph Key

🟦 Src-Dst throughput

🟩 Dst-Src throughput

Throughput test between Source: perfsonar.3rox.net(147.73.4.240) -- Destination: bwctl-10g.wash.net.internet2.edu(64.57.16.6)

<- 1 month

1 month ->

Timezone: Daylight Time)

Direction	Max throughput(bps)	Mean throughput(bps)	Min throughput(bps)
Src-Dst	921.16M	825.21M	801.08M
Dst-Src	843.44M	813.94M	689.93M

Show/Hide Link

If performance is as expected…

- Declare victory and celebrate!

If performance needs improvement…

- Check end host tuning
- NDT/NPAD
- Linux script to gather OS version, sysctl, lspci, and ifconfig parameters
  - http://staff.psc.edu/benninge/networking/check_net_config.html
- May be complicated by login access issues
- Knowledge of TCP tuning, NIC configuration, and system hardware along with admin access will be needed to interpret the results and implement corrections.

MTU discovery and MTU mismatch

- Potential issue between XSEDE and non-XSEDE sites
- XSEDE network standard is 9000 byte MTU
- Non-XSEDE sites often use 1500 byte MTU
  - Implementation of Science DMZs doesn't guarantee 9000 byte MTU support throughout a site
- MTU discovery may not work correctly
  - Broken – network infrastructure does not handle jumbo frames correctly
  - Firewalls blocking or limiting ICMP packets

MTU testing - tracepath

```
[benninge@perfsonar ~]$ tracepath www.iup.edu
 1:  perfsonar.3rox.net (147.73.5.240)                  0.126ms pmtu 9000
 1:  gigapop-srv-default-9k.3rox.net (147.73.5.1)       1.065ms asymm  2
 1:  gigapop-srv-default-9k.3rox.net (147.73.5.1)       1.018ms asymm  2
 2:  re-rtr.3rox-services.3rox.net (147.73.18.225)     13.035ms
 3:  internet2-wash-3rox.net.internet2.edu (192.88.115.83)  7.494ms
 4:  204.238.76.65 (204.238.76.65)                     10.061ms
 5:  204.238.76.65 (204.238.76.65)                     10.117ms pmtu 1500
 5:  204.238.76.58 (204.238.76.58)                     11.266ms
 6:  172.28.82.1 (172.28.82.1)                         36.637ms
 7:  dmz-hub.net.iup.edu (192.231.220.1)               36.448ms
 8:  no reply
```

- Commonly available for the end user to run

MTU testing – ping with varying packet size

```
[benninge@perfsonar ~]$ ping -s 1472 -M do www.sru.edu -c 5
PING www.sru.edu (205.149.70.100) 1472(1500) bytes of data.
1480 bytes from blog.sru.edu (205.149.70.100): icmp_seq=1 ttl=248 time=23.5 ms
1480 bytes from blog.sru.edu (205.149.70.100): icmp_seq=2 ttl=248 time=23.3 ms
1480 bytes from blog.sru.edu (205.149.70.100): icmp_seq=3 ttl=248 time=23.4 ms
1480 bytes from blog.sru.edu (205.149.70.100): icmp_seq=4 ttl=248 time=23.2 ms
1480 bytes from blog.sru.edu (205.149.70.100): icmp_seq=5 ttl=248 time=23.1 ms

--- www.sru.edu ping statistics ---
5 packets transmitted, 5 received, 0% packet loss, time 4028ms
rtt min/avg/max/mdev = 23.148/23.348/23.583/0.153 ms
[benninge@perfsonar ~]$ ping -s 1473 -M do www.sru.edu -c 5
PING www.sru.edu (205.149.70.100) 1473(1501) bytes of data.
From 204.238.76.65 icmp_seq=1 Frag needed and DF set (mtu = 1500)
From perfsonar.3rox.net (147.73.5.240) icmp_seq=2 Frag needed and DF set (mtu = 1500)
From perfsonar.3rox.net (147.73.5.240) icmp_seq=2 Frag needed and DF set (mtu = 1500)
From perfsonar.3rox.net (147.73.5.240) icmp_seq=2 Frag needed and DF set (mtu = 1500)
From perfsonar.3rox.net (147.73.5.240) icmp_seq=2 Frag needed and DF set (mtu = 1500)

--- www.sru.edu ping statistics ---
1 packets transmitted, 0 received, +5 errors, 100% packet loss, time 1002ms
```

- ping can typically be run by end user

MTU testing - BWCTL connects but fails

- Site network configuration does not handle jumbo frames correctly:
  - bwctl testing connects, but subsequently fails to run
  - Manual bwctl testing fails and reports. Example:

```
[benninge@perfsonar ~]$ bwctl -t 10 -i 2 -f m -L 300 -c net-test.univ.edu
bwctl: Using tool: iperf
bwctl: 17 seconds until test results available

RECEIVER START
bwctl: exec_line: iperf -B net-test.univ.edu -s -f m -m -p 5293 -t 10 -i 2
bwctl: start_tool: 3582477743.167692
------------------------------------------------------------
Server listening on TCP port 5293
Binding to local address net-test.univ.edu
TCP window size: 0.08 MByte (default)
------------------------------------------------------------
[ 15] local 111.222.33.44 port 5293 connected with 55.66.7.89 port 5293
bwctl: local tool did not complete in allocated time frame and was killed
bwctl: stop_exec: 3582477759.069982

RECEIVER END
```

Additional checks

- Check for firewalls or intentional rate limiting
- perfSONAR can work within a firewall but requires:
  - http://psps.perfsonar.net/toolkit/FAQs.html#Q6
  - http://fasterdata.es.net/performance-testing/perfsonar/ps-howto/perfsonar-firewall-requirements/
- Check network equipment counters to verify traffic volume
- Note that link aggregation of multiple 1 GbEs or 10 GbEs still only support a single flow maximum of 1 Gbps or 10 Gbps
- Verify that the file transfer software is the best choice among the available options
- Check end system performance characteristics
- Often requires consultation with network engineering staff and computer systems staff

Outcomes

- Significantly improved the end-to-end network throughput
- Initial single stream iperf testing between the University and the XSEDE-SP2 site was 13.8 Mb/sec
- Tuning and reconfiguration increased the achievable throughput to 807 Mb/sec (1GbE connected hosts)

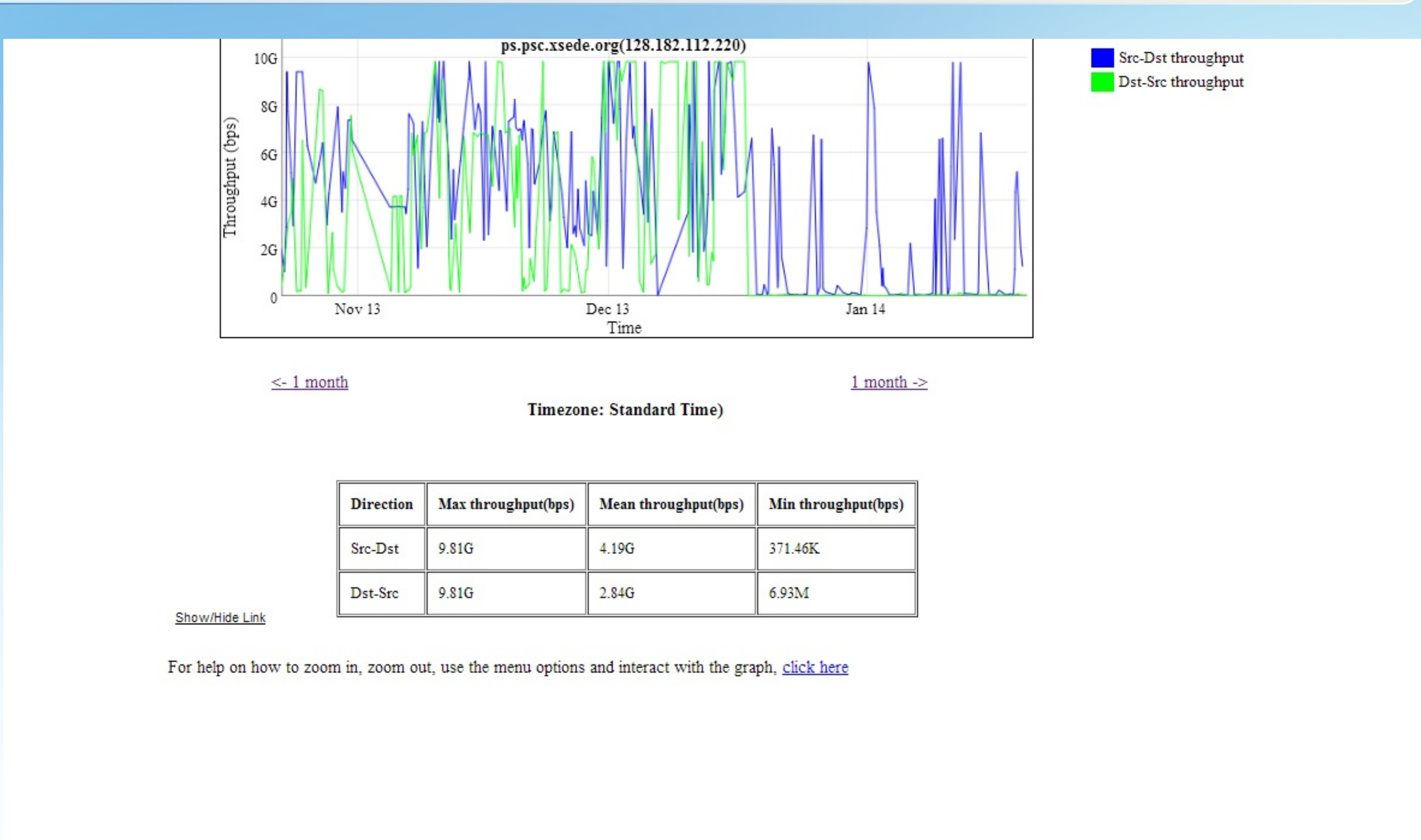
Outcomes

- Initial single stream iperf tests between the University and PSC was 220Mb/s
- Throughput improved to over 990 Mb/s on each parallel 1 GbE stream following steps presented
- Transfers could completely consume the available 5 Gb/s bandwidth between the University site and PSC in testing
- Transferring 470 TB of data in 22 days yielded an overall average of 21.4 TB/day with a daily average of 2.0 Gb/s and a daily maximum of 4.2 Gb/s.

perfSONAR view of network problems

- The following slides represent problems identified from perfSONAR test results
- Observations of scheduled testing over time
  - Network engineer initially scheduled the tests
  - Users can view the test results graphed on the perfSONAR Measurement Archive

perfSONAR view of router/switch buffering



<- 1 month

1 month ->

Timezone: Standard Time)

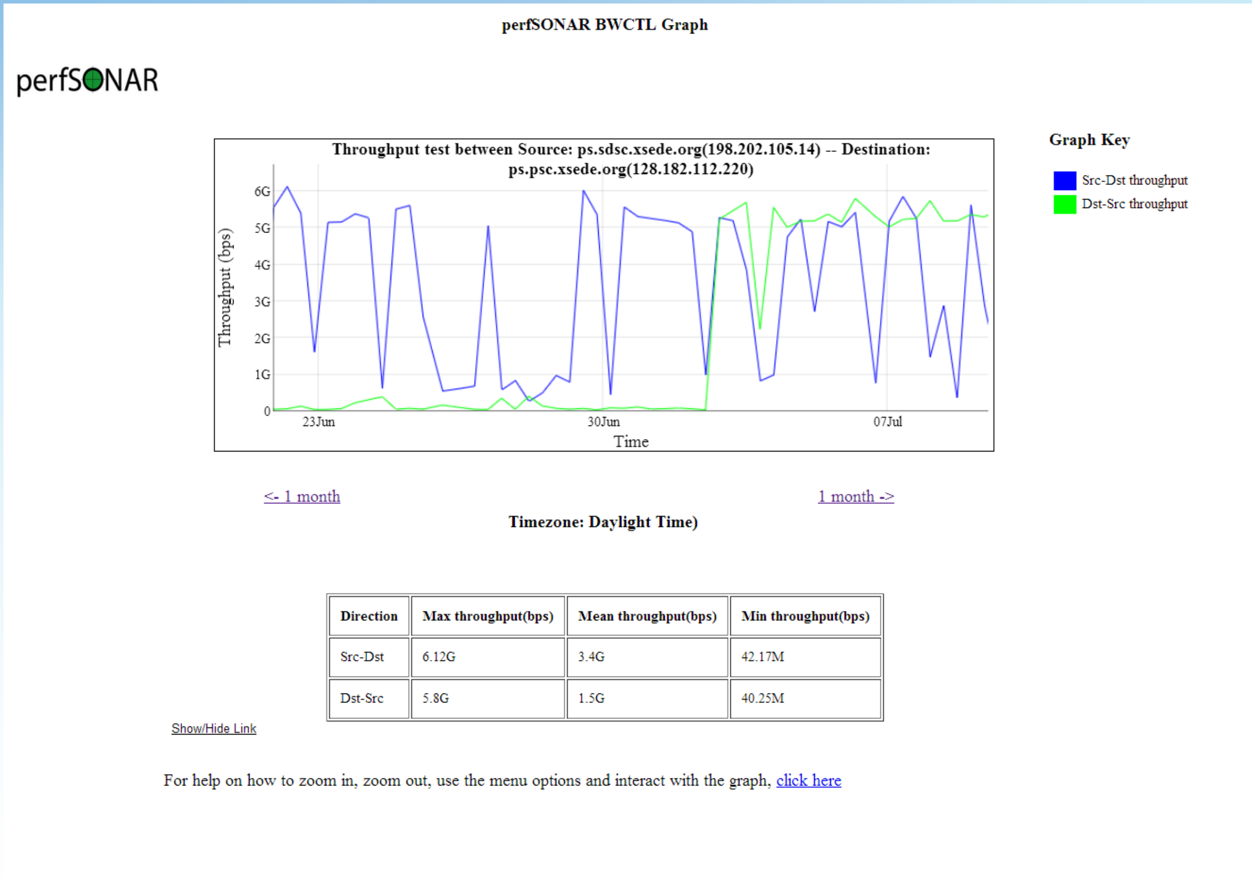
Direction	Max throughput(bps)	Mean throughput(bps)	Min throughput(bps)
Src-Dst	9.81G	4.19G	371.46K
Dst-Src	9.81G	2.84G	6.93M

Show/Hide Link

For help on how to zoom in, zoom out, use the menu options and interact with the graph, click here

perfSONAR view of router/switch buffering

- Outbound bwctl multi-Gb/s; inbound << 1 Gb/s



perfSONAR BWCTL Graph

perfSONAR
Throughput test between Source: ps.sdsc.xsede.org(198.202.105.14) -- Destination: ps.psc.xsede.org(128.182.112.220)
Throughput (bps)
6G
5G
4G
3G
2G
1G
0
23Jun
30Jun
07Jul
Time
Graph Key
Src-Dst throughput
Dst-Src throughput

<- 1 month

1 month ->

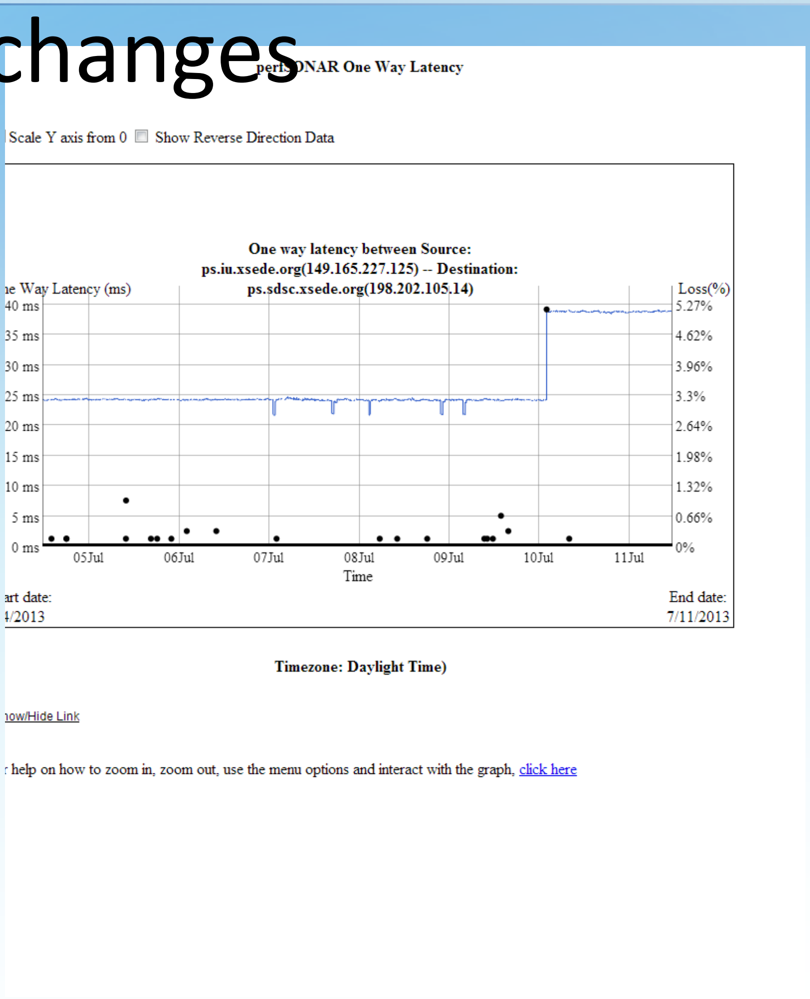
Timezone: Daylight Time)

Direction	Max throughput(bps)	Mean throughput(bps)	Min throughput(bps)
Src-Dst	6.12G	3.4G	42.17M
Dst-Src	5.8G	1.5G	40.25M

Show/Hide Link

For help on how to zoom in, zoom out, use the menu options and interact with the graph, click here

perfSONAR OWAMP view of route changes



perfSONAR One Way Latency

Scale Y axis from 0 □ Show Reverse Direction Data

One way latency between Source: ps.iu.xsede.org(149.165.227.125) -- Destination: ps.sdsc.xsede.org(198.202.105.14)

art date:
/2013

End date:
7/11/2013

Timezone: Daylight Time)

how/Hide Link

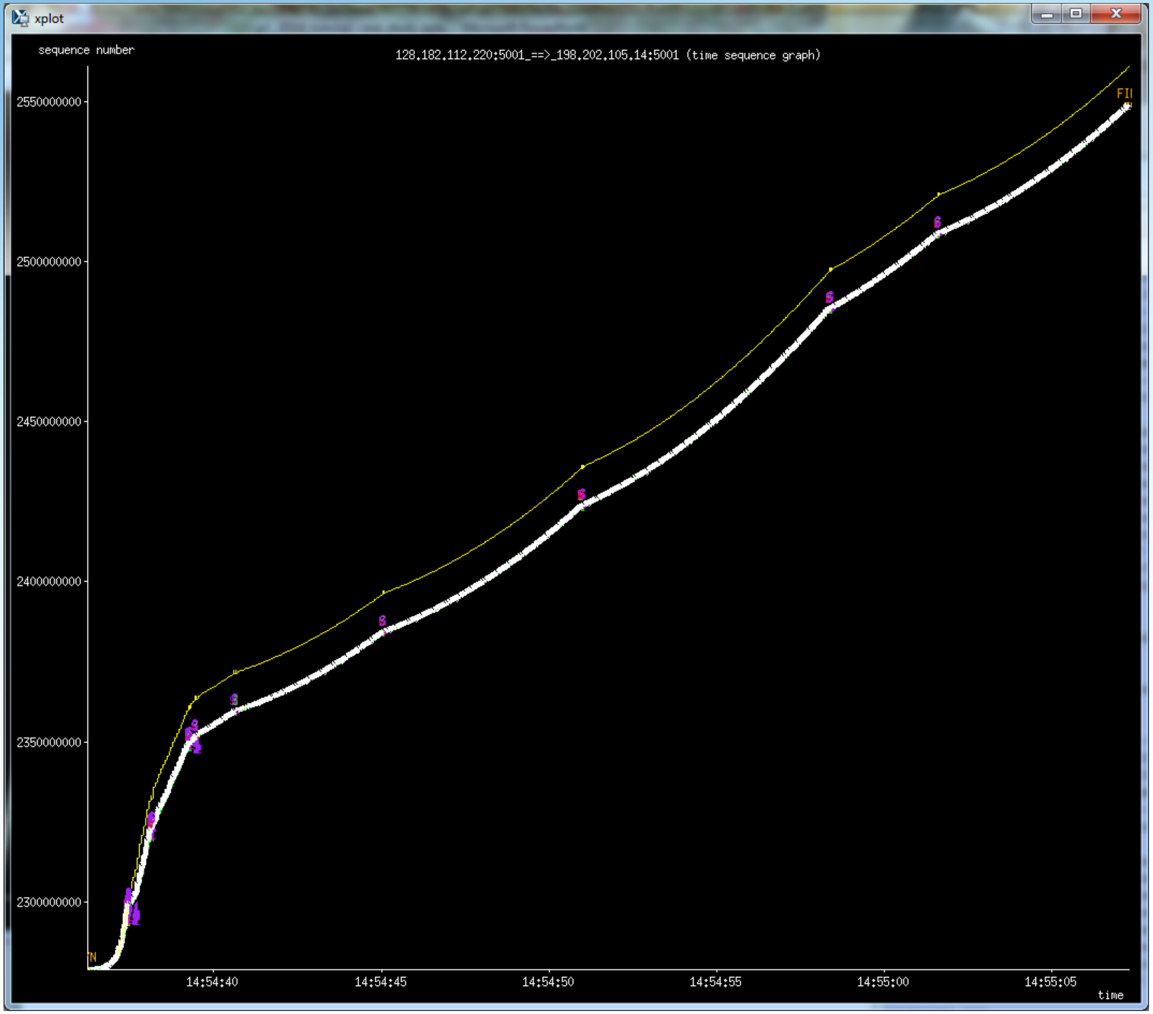
help on how to zoom in, zoom out, use the menu options and interact with the graph, click here

perfSONAR view of network problems

- Graph of BWCTL iperf traffic
  - Use tcpdump to collect the packet headers
  - tcptrace to process the tcpdump data
  - xplot to graph
- Generated and interpreted by network engineer

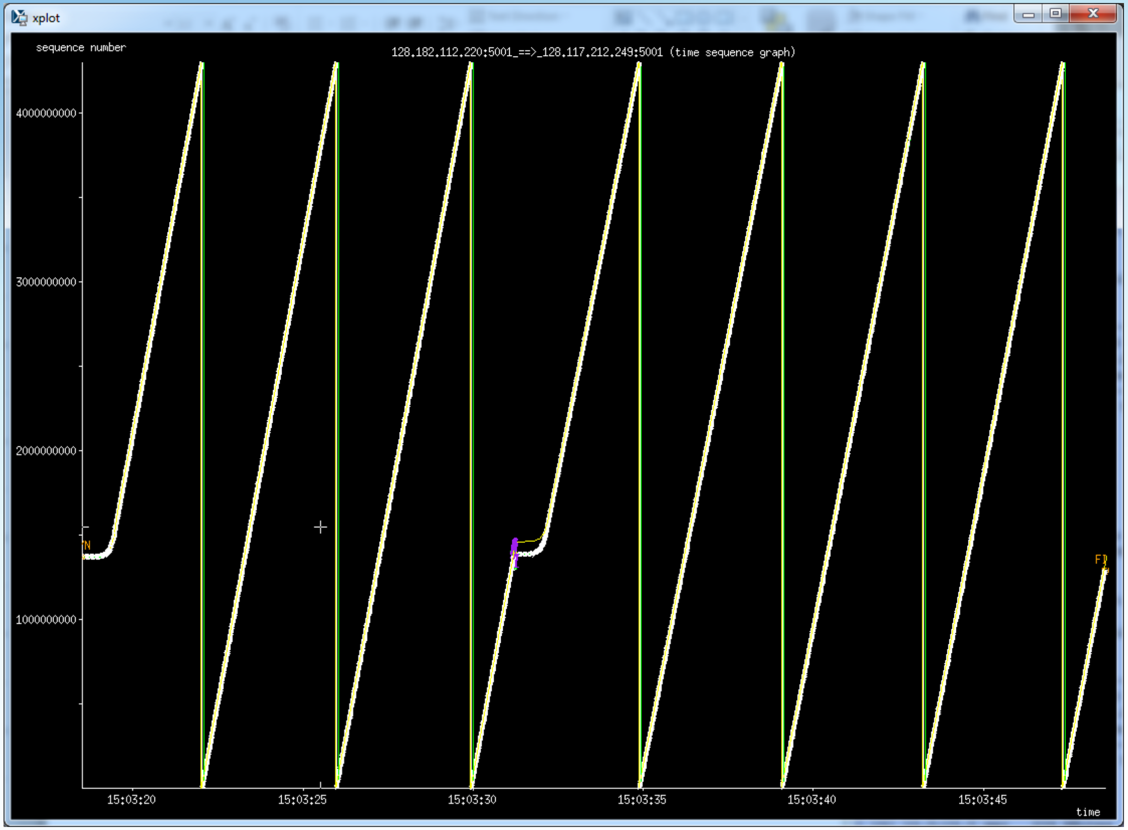
tcpdump/tcptrace/xplot view

TCP buffer size is too small to support bandwidth and RTT



tcpdump/tcptrace/xplot view

Sufficient TCP buffer size to support full 10 Gbps at the RTT



XSEDE Performance Use Cases

August 11th 2014, APAN38

Network Performance Tutorial

John Hicks – Internet2

Special thanks to perfSONAR partners for assistance in lesson material