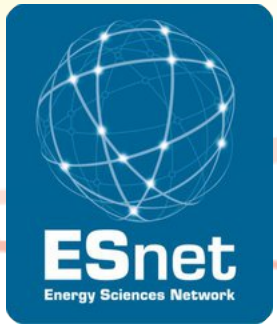


July 22nd 2013, XSEDE Network Performance Tutorial

Jason Zurawski – Internet2/ESnet

Kathy Benninger - Pittsburgh Supercomputing Center

Performance Use Cases

Use Cases

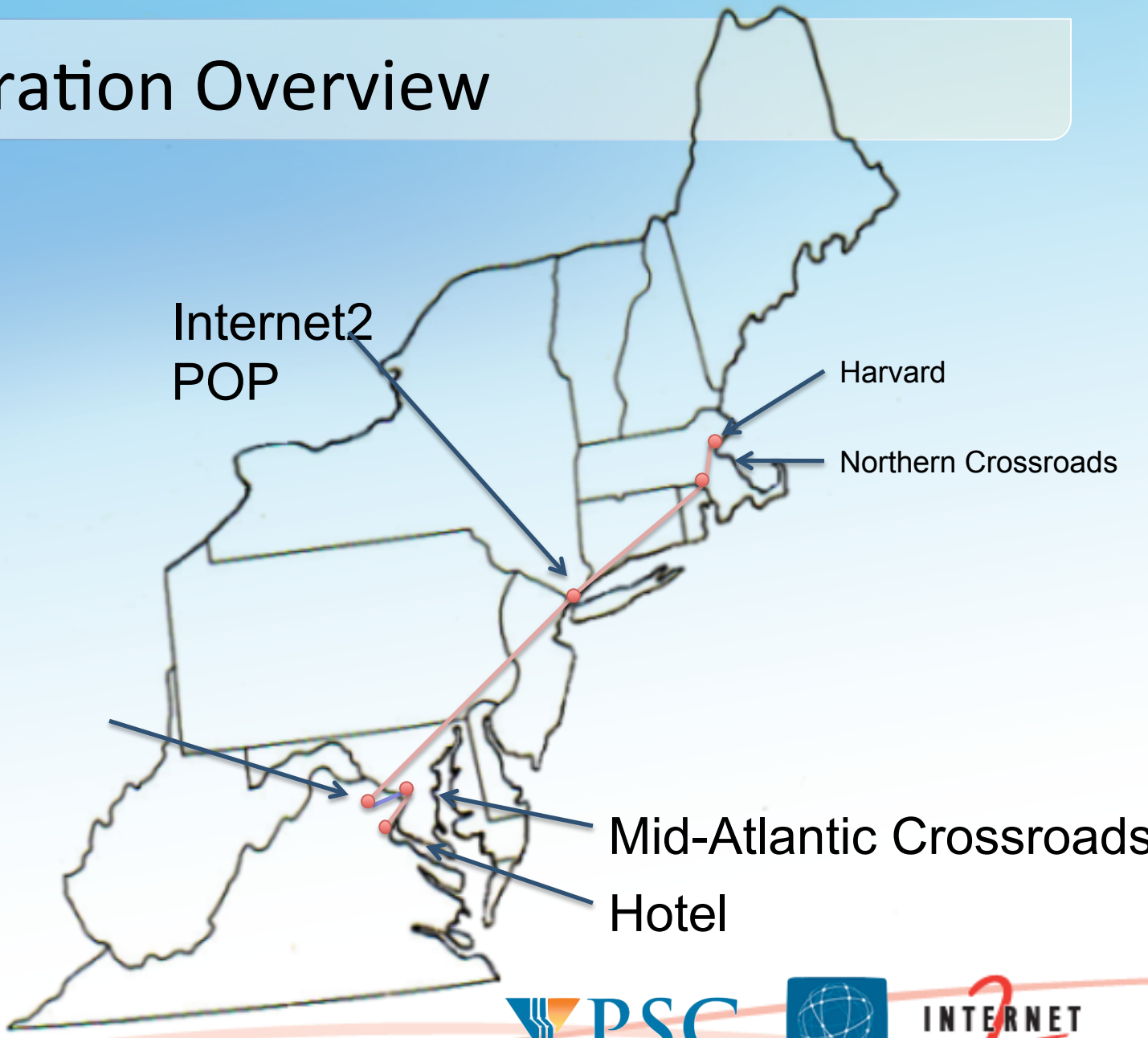
- The following use cases demonstrate use of perfSONAR tools to solve sometimes complex performance problems
  - Cisco Telepresence
    - Multi-domain path where performance guarantees dictate use of a specific application
  - Internet2 Backbone Incident
    - Learning the value of trusting the measurement tools
  - Three use case examples from XSEDE
    - Jumbo frame MTU issues
    - Impact of small router buffers
    - Route changes

J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Cisco TelePresence Demo

- 2 Locations
  - Harvard University (Boston, MA)
  - Spring Member Meeting (Arlington, VA)
- Must meet or exceed performance expectations
  - < 10 ms Jitter (Packet Arrival Variation)
  - < 160 ms End-to-End Delay
  - < 0.05% Packet Loss
- Network Path spanned:
  - ~450 Miles
  - 4 Distinct Domains
    - Internet2
    - Mid Atlantic Crossroads (MAX)
    - Northern Crossroads (NOX)
    - Harvard University

Demonstration Overview

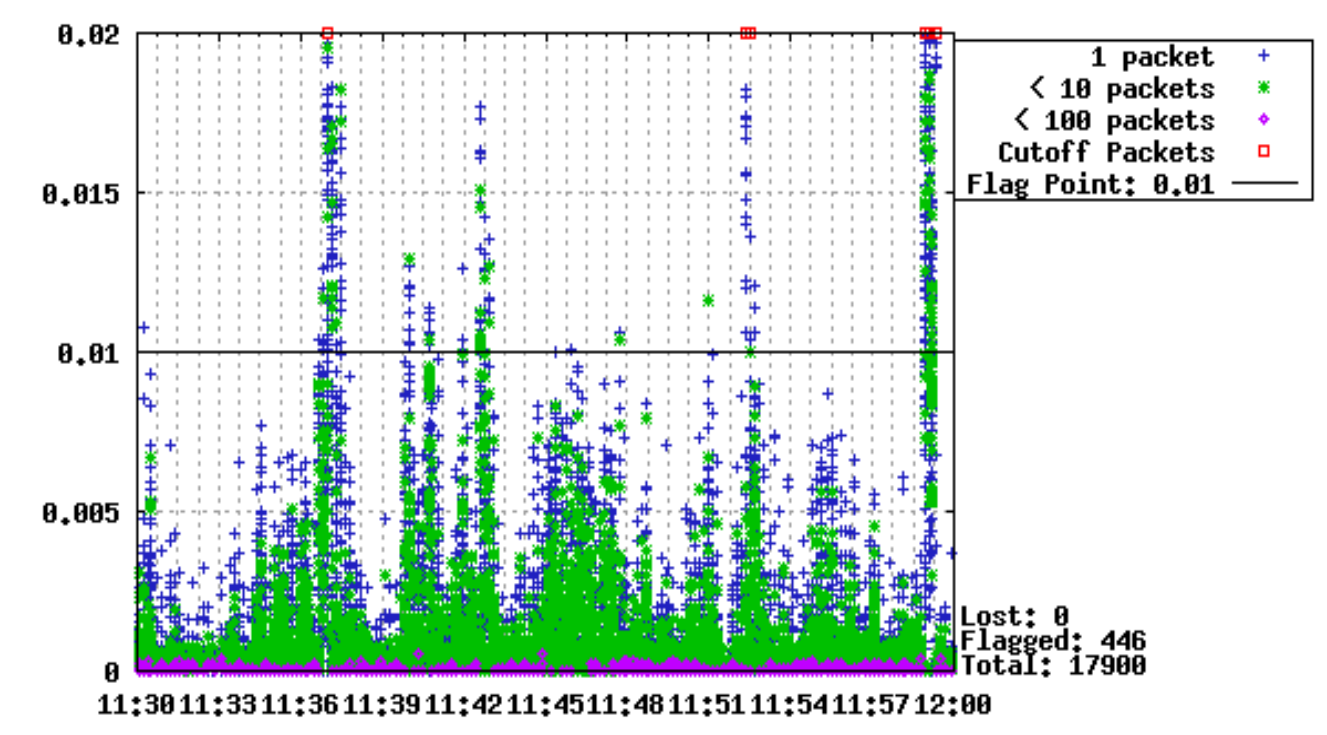


Cisco TelePresence Demo

- Performance Monitoring
  - Tools installed within each domain
  - Interested in several ‘metrics’
    - One Way Delay – OWAMP
    - Network Utilization – SNMP
- Several Problems Found (And Corrected)
  - Over-utilized Link
  - Traffic Spikes from Cross Traffic

Over-utilized Link

- Tools indicated high amounts of end-to-end Jitter:



- Goal: Isolate which segment (or segments) to examine further.

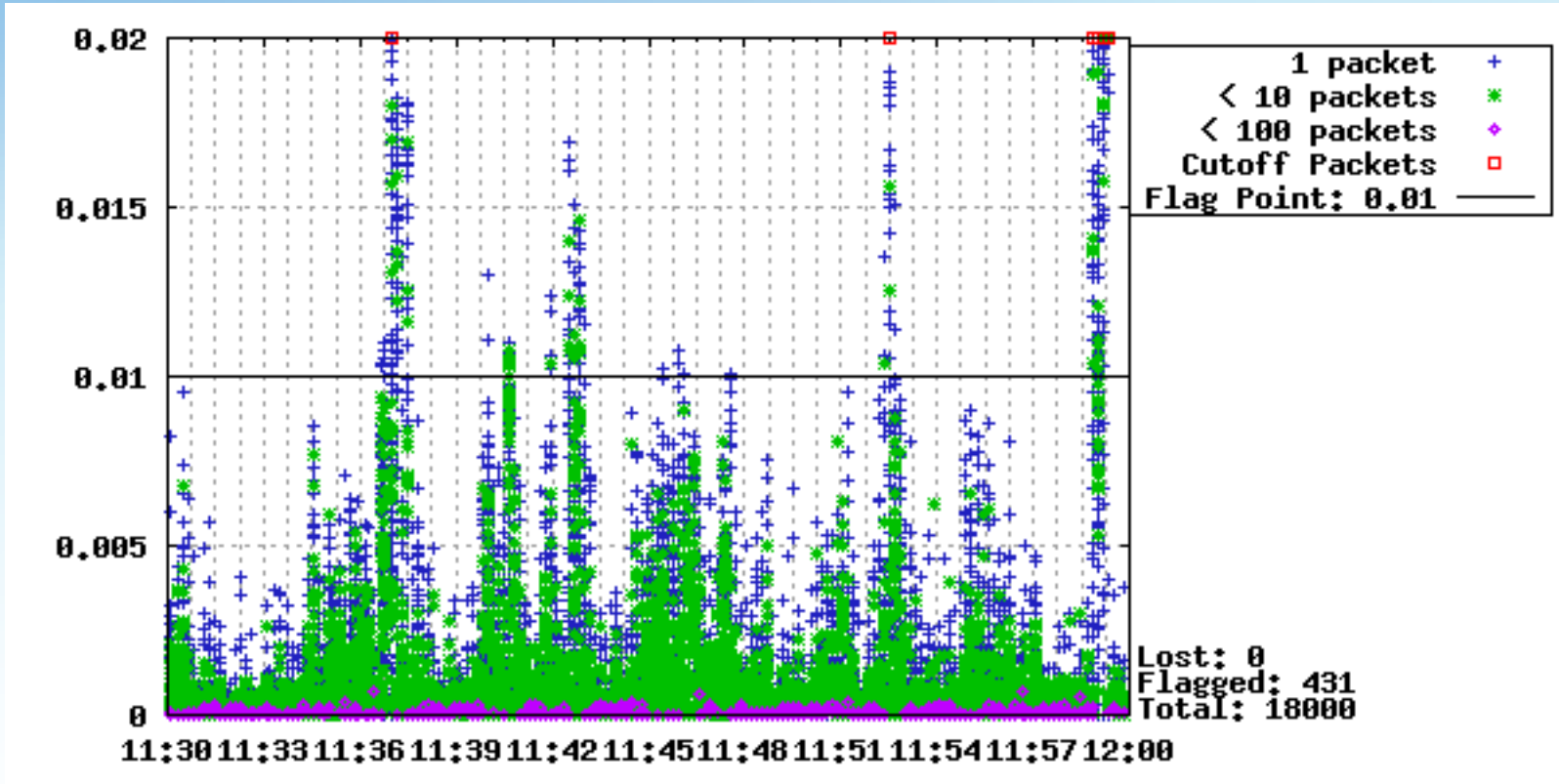
High Jitter – But Where?

Over-utilized Link

- Process:
  - Tools are installed and available in each domain
  - 'Decompose' the entire end-to-end path, and examine the performance between testing points:
    - Meeting Hotel to NOX
    - Meeting Hotel to Internet2 (New York)
    - Meeting Hotel to Internet2 (Washington)
    - Meeting Hotel to MAX

Over-utilized Link

- Meeting Hotel to NOX



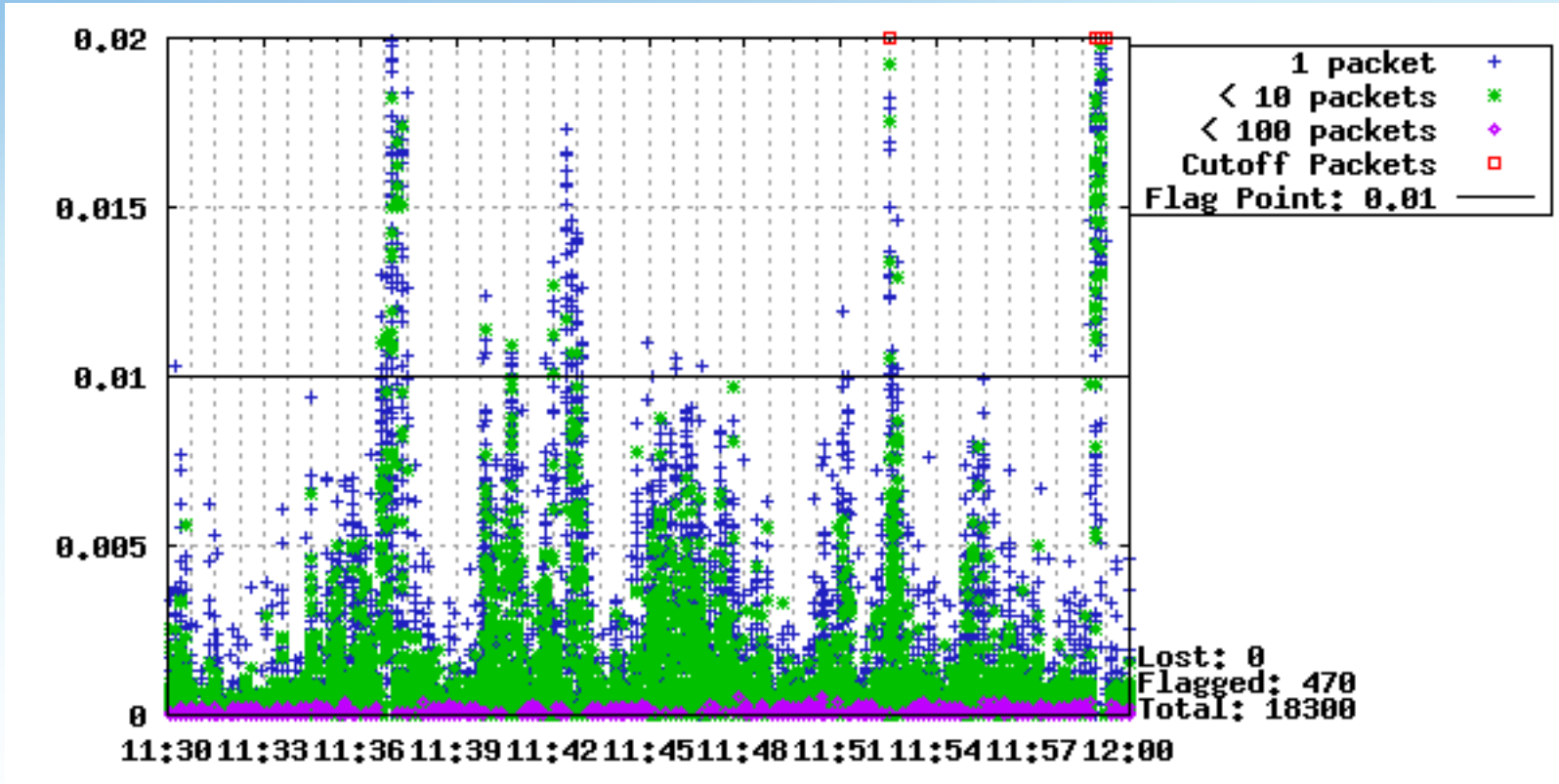
J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Still Seen on Shorter Path

J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

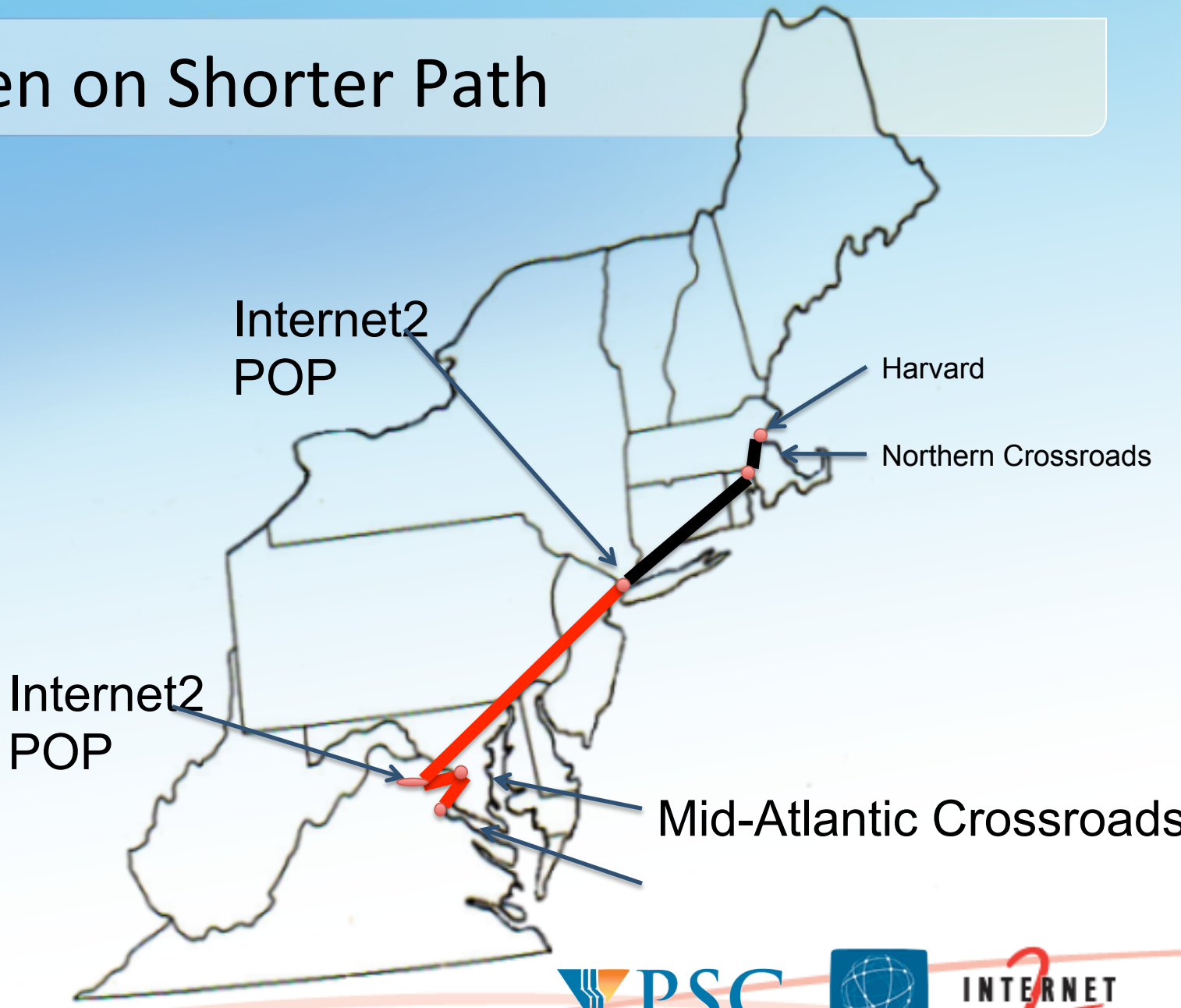
Over-utilized Link

- Meeting Hotel to Internet2 (New York)



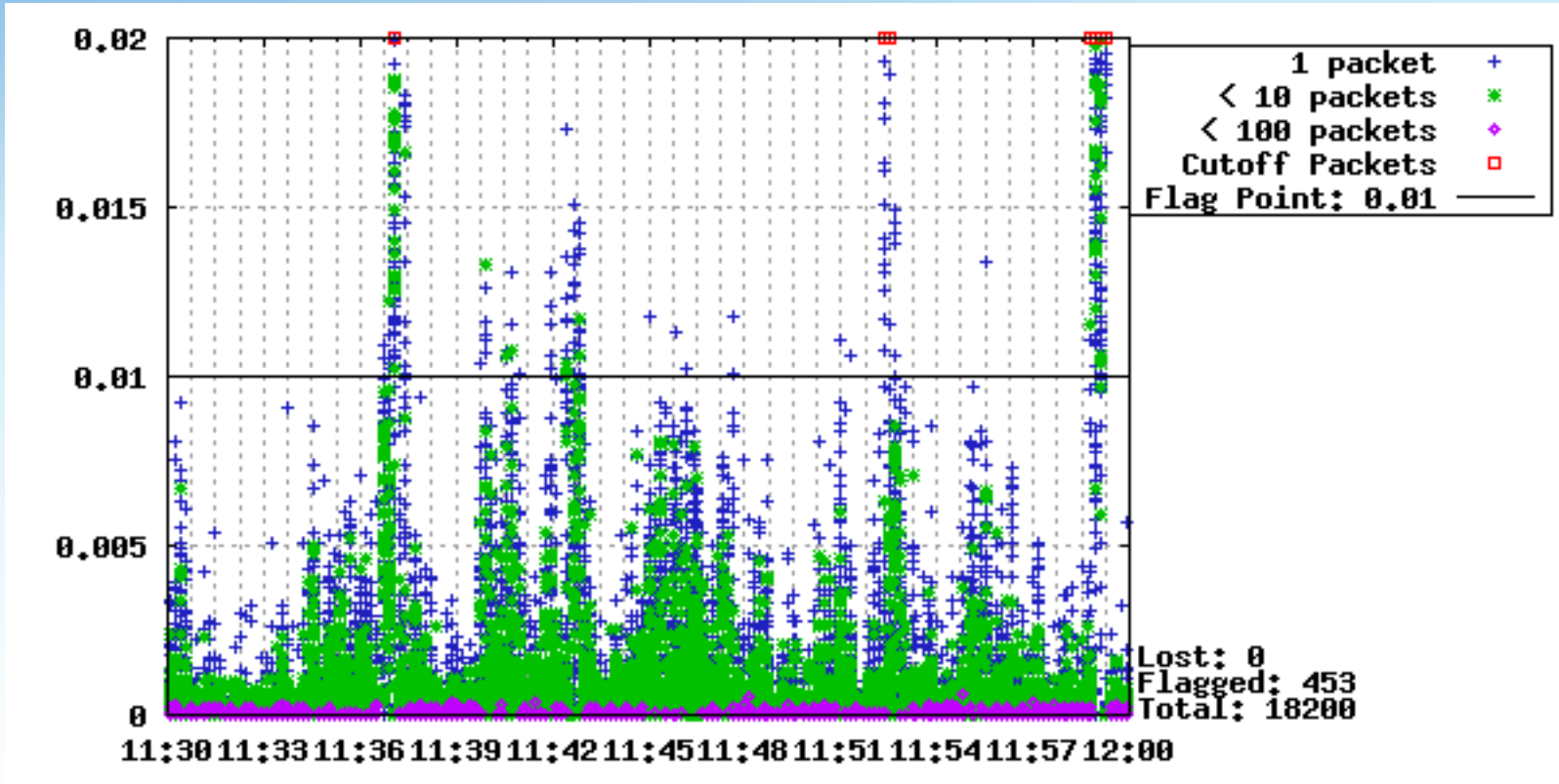
J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Still Seen on Shorter Path



Over-utilized Link

- Meeting Hotel to Internet2 (Washington)



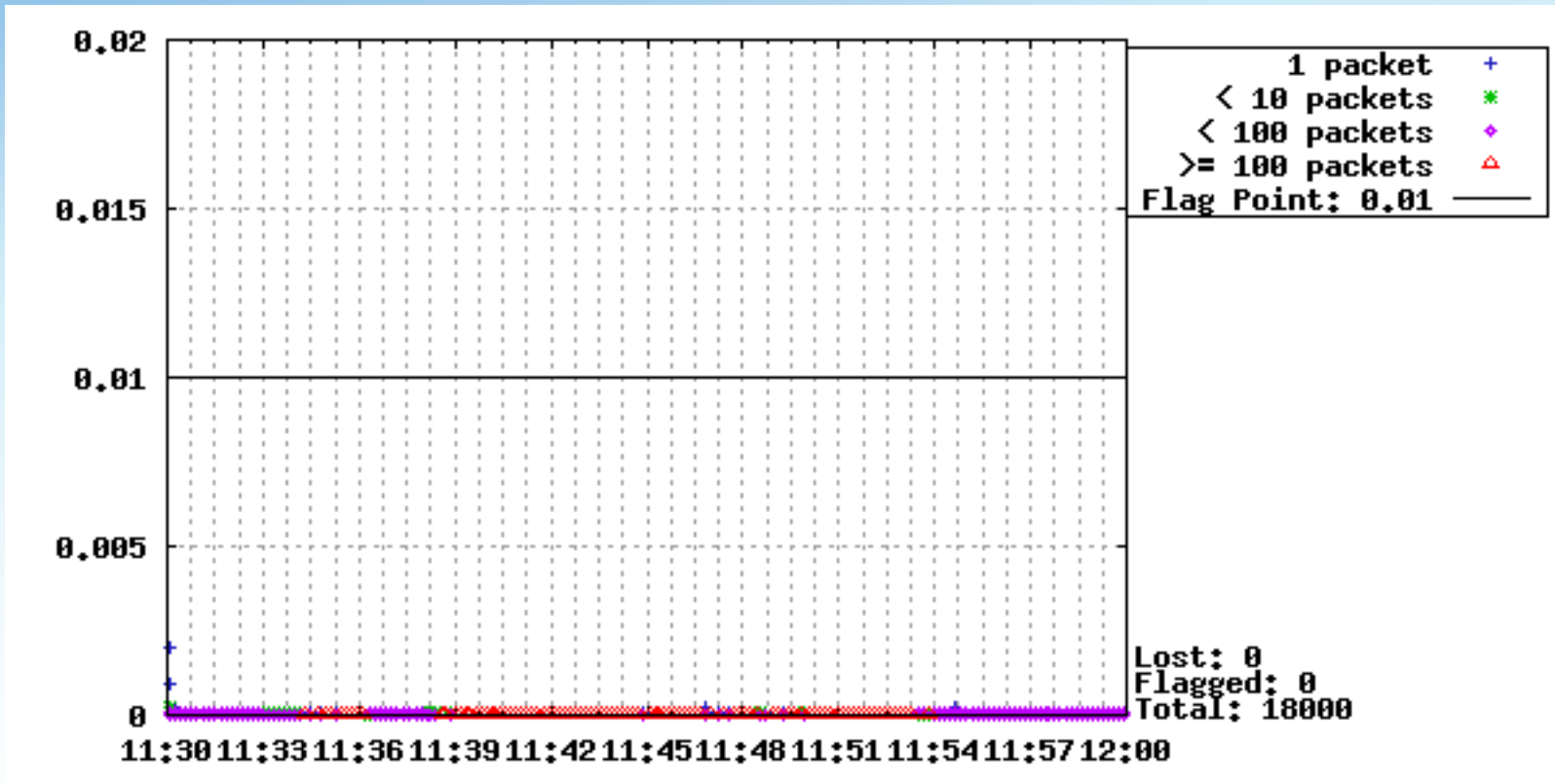
J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Still Seen on Shorter Path

J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Over-utilized Link

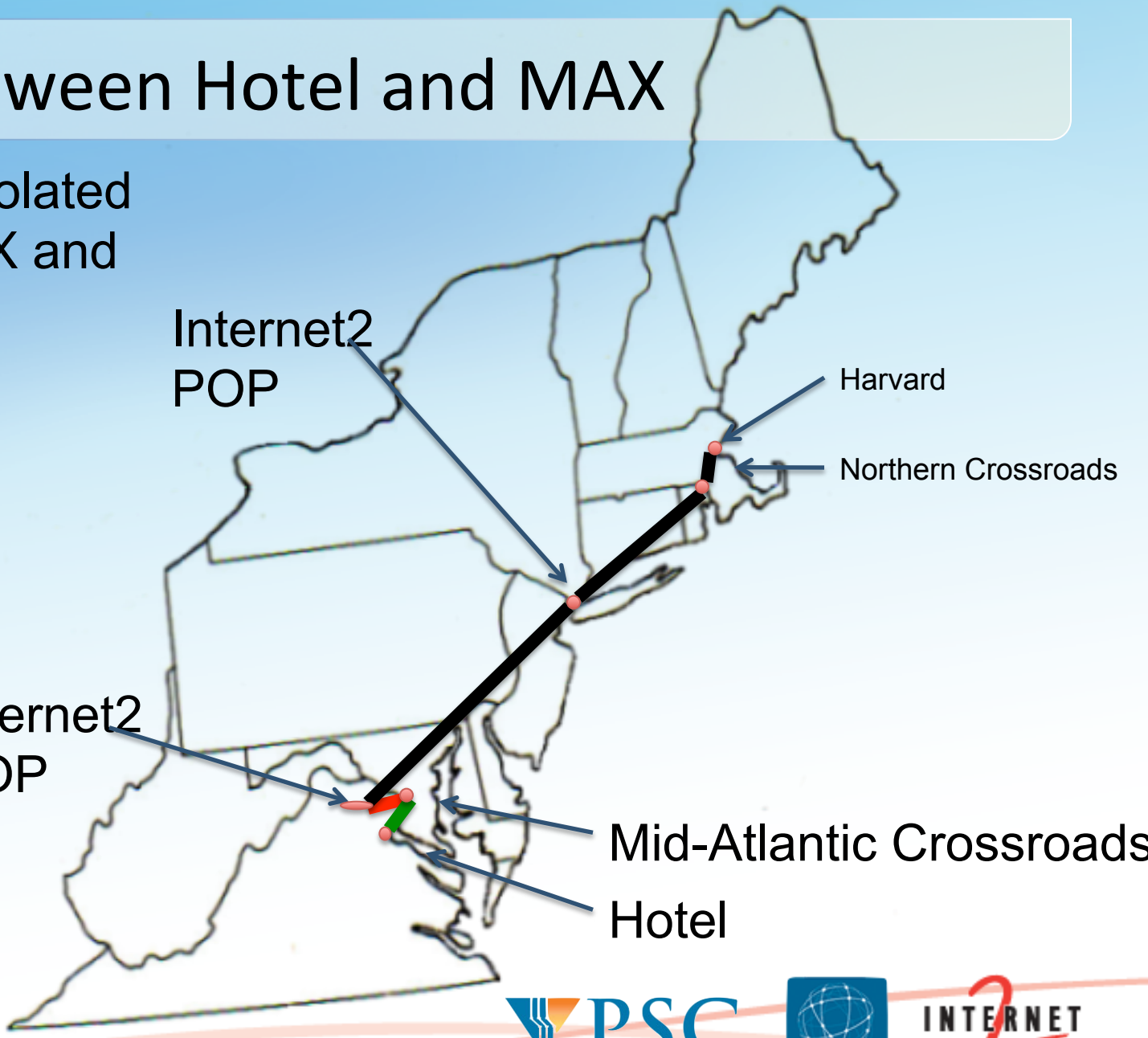
- Meeting Hotel to MAX



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Clean Between Hotel and MAX

Problem is isolated between MAX and Harvard



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Over-utilized Link

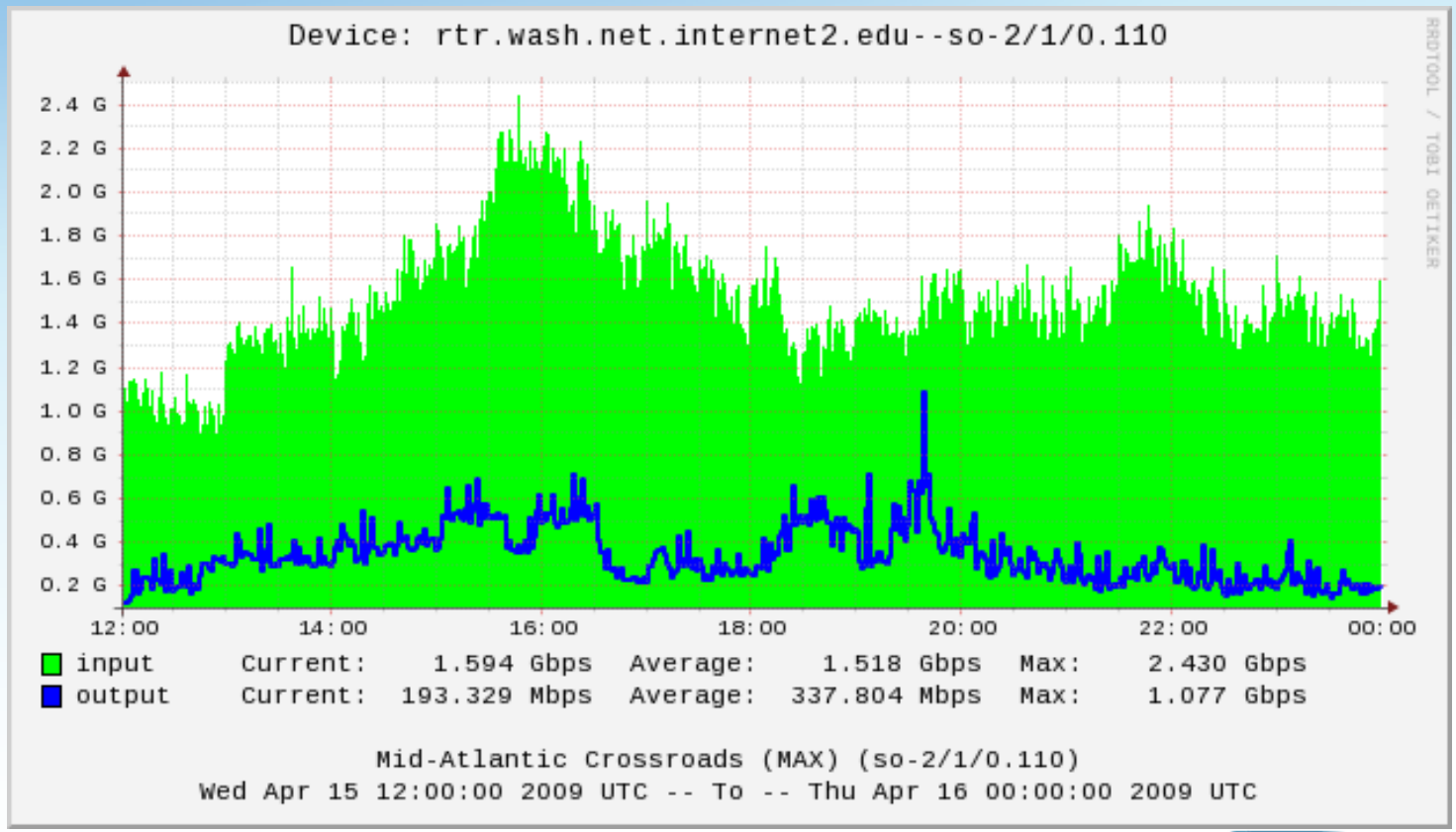
- Results of 1st Round of Debugging
  - OWAMP Confirms that the path is ‘clean’ between the Hotel and MAX.
  - The path is ‘noisy’ between MAX and Harvard (could be anywhere – we only know where it isn’t)
- Action Plan
  - Use other resource available, Utilization, to see if there is a ‘pinch point’ on one of the links.
  - Isolate our search to areas between MAX and Harvard
  - Start at MAX

Over-utilized Link

- Starting in the MAX domain, we know of 4 links:
  - Hotel to College Park MD
  - MAX Core in College Park MD
  - College Park MD to McLean VA
  - Internet2 Uplink in McLean VA
- Get information on each link:
  - 1G from Hotel to College Park MD
  - 10G MAX Core and transit to McLean VA
  - 2.5G Uplink to Internet2 in McLean VA

Over-utilized Link

- Utilization on Internet2 Uplink from MAX:



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Over-utilized Link

- 2nd Round Debugging Results:
  - 'Pinch Point' found: traffic was coming very close to 2.5G limit
  - Not constant – but noticeable during network busy hours
  - 'Pinch Point' corrected (e.g. 2.5G uplink replaced with 10G uplink)
  - All other segments of the path appeared clean
  - Further end-to-end testing after upgrade revealed no additional problems.

Internet2 Backbone Incident

Original Report – The Network is Broken!

- Feb 10th 2011 – Original report from Vanderbilt University (US CMS Heavy ION Tier2 Facility, Nashville TN) noting problems to Port d'Informació Científica (PIC – Barcelona Spain)
- Observation from users:
  - *We are having trouble (slow transfers) with transfers from the CMS T1 sites in Spain (PIC). Here are traceroutes ... who can I talk to about this? Are we at least going along reasonable routes?*
- Quick mental triage on my part:
  - **Users are sharp, they have done this sort of thing before**
  - **They know the value of monitoring, and know when they are in over their head**
  - **Traceroutes are good, some real measurements would be better**
  - **Will require allocation of resources to address, coordinated by me now ☺**

Resource Allocation & Instrumentation

- *"I wish someone would develop a framework to make this easier"*
  - Yes, perfSONAR works well – ***when it is deployed***.
  - We still don't have universal deployment, so the backchannel network of emails to "people you know" is still required
- Coordination in domains, need to talk to people in each and allocate testers (if they don't exist yet)
  - PIC*
  - CESCA
  - RedIRIS
  - GÉANT
  - Internet2*
  - SOX
  - Vanderbilt*

*** Started with these for simplicity**

J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Resource Allocation & Instrumentation

- End Systems @ PIC and Vanderbilt
  - pS Performance Toolkit on a spare server
  - Racked next to the data movement tools
  - Benefits:
    - The similar OS and performance settings on each end "*levels the playing field*"
    - All tools are now available, if we want to run an NDT we can, if we need regular BWCTL, we have it.
  - Cost to me and remote hands = < 1hr of installation/configuration
- Internet2
  - Regular BWCTL, OWAMP testing in place.
  - Interface Utilization and Errors available for all links
  - Web100 enabled services for NDT and NPAD

Structured Debugging Methodology

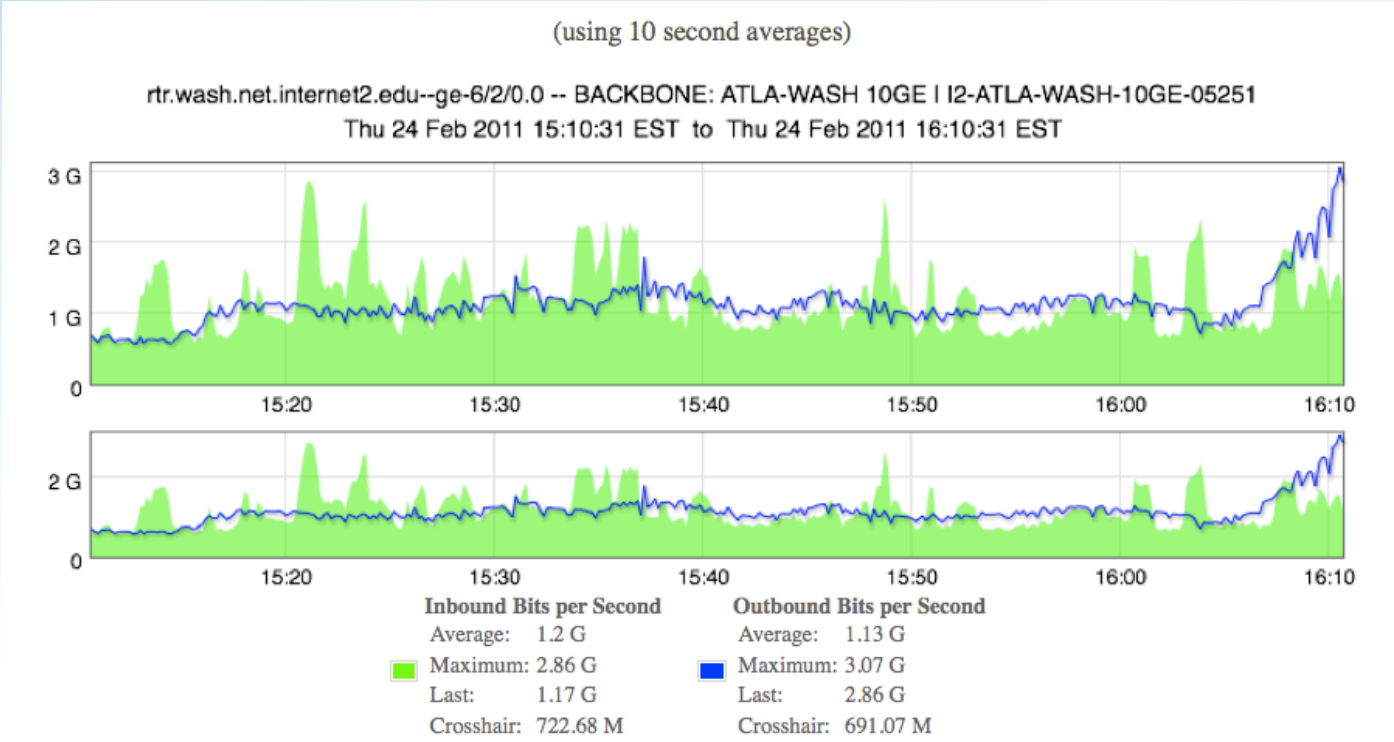
- Divide and Conquer
  - Bisect the path and test the segments individually
  - Rule out paths that are doing well, subdivide those that aren't again and again
- Use of one tool a time
  - Collect as much as you can with each tool
  - Move to the next to gather different metrics
- Patience
  - Its not hard, but it is time consuming

Real Debugging – Results (Traceroutes)

- Methodology
  - GÉANT Circuit from Frankfurt terminates at Internet2 Washington DC. Use test points here.
  - Vanderbilt connects through SOX, which connects to Internet2 in Atlanta GA. Use test points here too.
  - 2 10G backbone links separate Atlanta and Washington.
- Between PIC and Vanderbilt were asymmetric
  - PIC->CESCA->RedIRIS->GEANT->Internet2->SOX->Vanderbilt
  - Vanderbilt->SOX->NLR->GEANT->RedIRIS->CESCA->PIC
- Focus on the US connectivity:
  - Between Vanderbilt and 2 Internet2 hosts, no asymmetry was observed
  - Path:
    - Vanderbilt->SOX->Internet2 (ATLA)->Internet2 (WASH)

Real Debugging – Results (Utilization)

- In the Internet2 case, utilization and errors are available.
- There are two backbone links between ATLA and WASH
  - 10G CPS Link – ruled this out of the process
  - 10G R&E Link



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Real Debugging – Results (NDT)

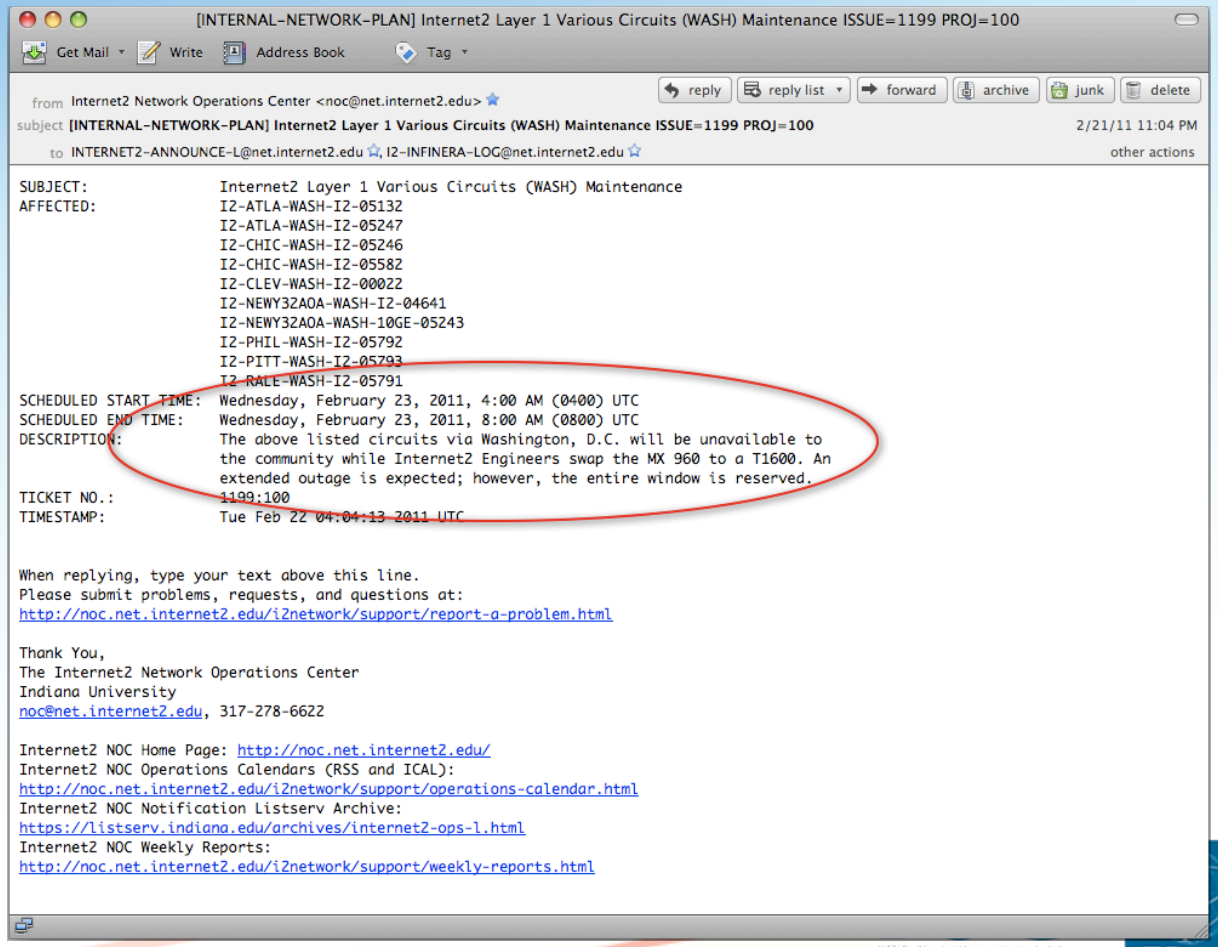
- NDT is not run "*regularly*", so our use will strictly be diagnostic.
- Vanderbilt (client) -> PIC (server)
  - running 10s outbound test (client to server) . . . . . 522.24 Mb/s
  - running 10s inbound test (server to client) . . . . . 169.89 kb/s
- Vanderbilt (client) -> WASH (server)
  - running 10s outbound test (client to server) . . . . . 922.47 Mb/s
  - running 10s inbound test (server to client) . . . . . 1.35 Mb/s
- Vanderbilt (client) -> ATLA (server)
  - running 10s outbound test (client to server) . . . . . 935.98 Mb/s
  - running 10s inbound test (server to client) . . . . . 933.82 Mb/s

Real Debugging – Results (NDT, cont.)

- We now have a minor result
  - Performance on a shorter path to from Vanderbilt to ATLA seems expected.
  - Can we use this to our advantage?
- Internet2 Atlanta (client) -> Internet2 Washington (server)
  - running 10s outbound test (client to server) . . . . . 978.44 Mb/s
  - running 10s inbound test (server to client) . . . . . 251.95 kb/s
- Very promising result … but we aren't done!
  - Can't declare victory with just this
  - Use other tools as much as we can
  - See if we can confirm that this segment is a problem

Real Debugging – Side Bar

- Related information is a good thing. There is a trouble ticket system that alerts to changes in the network:



[INTERNAL-NETWORK-PLAN] Internet2 Layer 1 Various Circuits (WASH) Maintenance ISSUE=1199 PROJ=100

Get Mail · Write · Address Book · Tag

reply · reply list · forward · archive · junk · delete

from Internet2 Network Operations Center <noc@net.internet2.edu>

subject [INTERNAL-NETWORK-PLAN] Internet2 Layer 1 Various Circuits (WASH) Maintenance ISSUE=1199 PROJ=100 — 2/21/11 11:04 PM

to INTERNET2-ANNOUNCE-L@net.internet2.edu, I2-INFINERA-LOG@net.internet2.edu — other actions

```
SUBJECT:                Internet2 Layer 1 Various Circuits (WASH) Maintenance
AFFECTED:               I2-ATLA-WASH-I2-05132
                        I2-ATLA-WASH-I2-05247
                        I2-CHIC-WASH-I2-05246
                        I2-CHIC-WASH-I2-05582
                        I2-CLEV-WASH-I2-00022
                        I2-NEWY32AOA-WASH-I2-04641
                        I2-NEWY32AOA-WASH-10GE-05243
                        I2-PHIL-WASH-I2-05792
                        I2-PITT-WASH-I2-05793
                        I2-RALE-WASH-I2-05791
SCHEDULED START TIME:   Wednesday, February 23, 2011, 4:00 AM (0400) UTC
SCHEDULED END TIME:     Wednesday, February 23, 2011, 8:00 AM (0800) UTC
DESCRIPTION:            The above listed circuits via Washington, D.C. will be unavailable to
                        the community while Internet2 Engineers swap the MX 960 to a T1600. An
                        extended outage is expected; however, the entire window is reserved.
TICKET NO.:             1199:100
TIMESTAMP:              Tue Feb 22 04:04:13 2011 UTC


When replying, type your text above this line.
Please submit problems, requests, and questions at:
http://noc.net.internet2.edu/i2network/support/report-a-problem.html

Thank You,
The Internet2 Network Operations Center
Indiana University
noc@net.internet2.edu, 317-278-6622

Internet2 NOC Home Page: http://noc.net.internet2.edu/
Internet2 NOC Operations Calendars (RSS and ICAL):
http://noc.net.internet2.edu/i2network/support/operations-calendar.html
Internet2 NOC Notification Listserv Archive:
https://listserv.indiana.edu/archives/internet2-ops-l.html
Internet2 NOC Weekly Reports:
http://noc.net.internet2.edu/i2network/support/weekly-reports.html
```

Real Debugging – Results (BWCTL)

- Regular monitoring is your friend … ***WHEN YOU USE IT PROPERLY*** ☺
  - Internet2 has lots of fancy GUIs that expose the BWCTL data, these should be viewed every now and then
  - We even have plugins for NAGIOS developed by perfSONAR-PS to alarm when performance dips below expectations
  - We did neither of these properly ☺

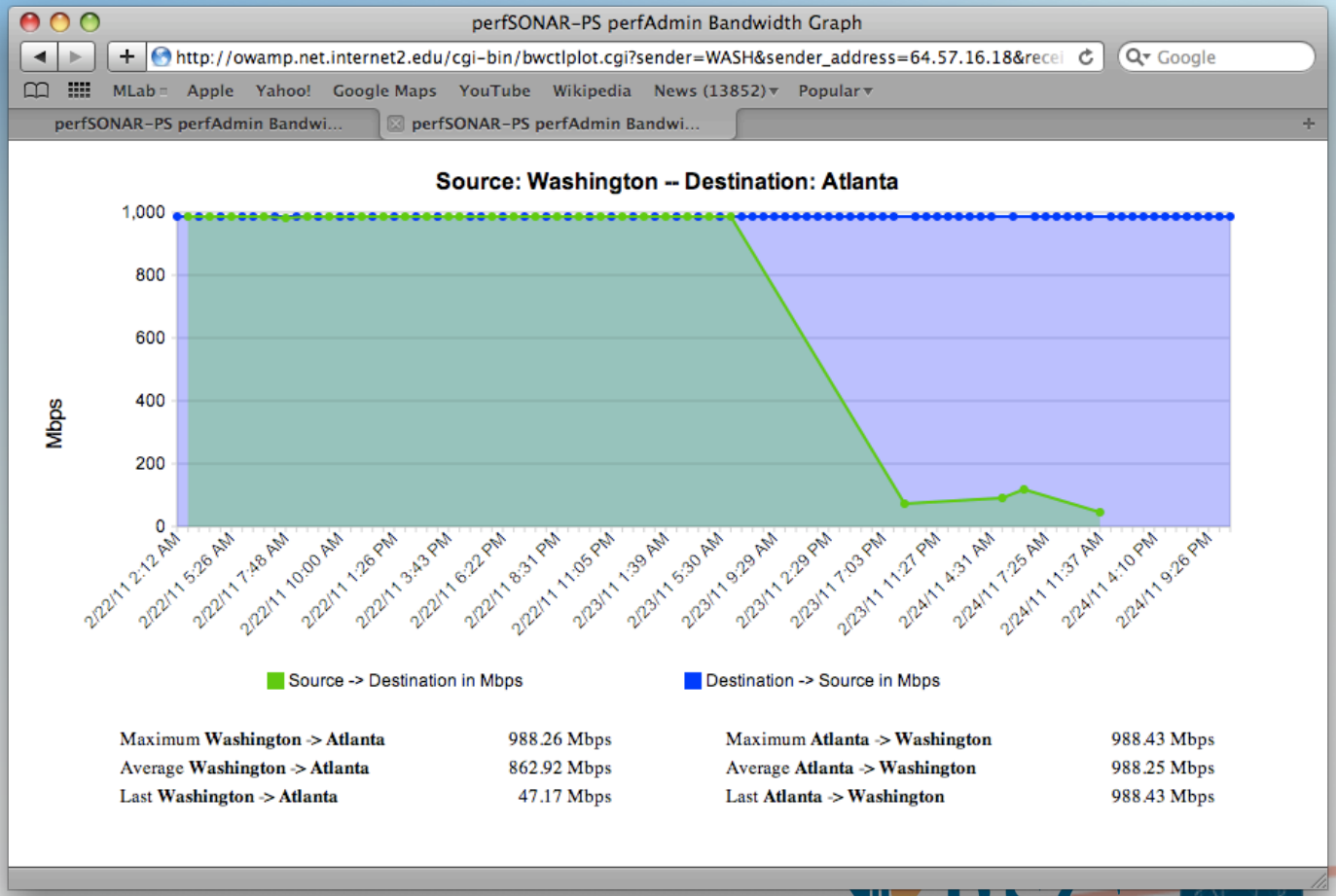
BWCTL - Internet2 Network IPv4 TCP Throughput

bwctl/iperf	Senders								
		Atlanta	Chicago	Houston	KansasCity	LosAngeles	NewYorkCity	SaltLakeCity	Seattle	Washington
Atlanta			942.06 Mbps / 2011-02-24 20:29:23UTC	941.72 Mbps / 2011-02-24 19:00:00UTC	940.73 Mbps / 2011-02-24 20:43:25UTC	739.75 Mbps / 2011-02-24 20:06:08UTC	132.37 Mbps / 2011-02-24 16:32:41UTC	751.79 Mbps / 2011-02-24 20:00:55UTC	584.60 Mbps / 2011-02-24 20:21:30UTC	44.99 Mbps / 2011-02-24 11:37:04UTC

J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

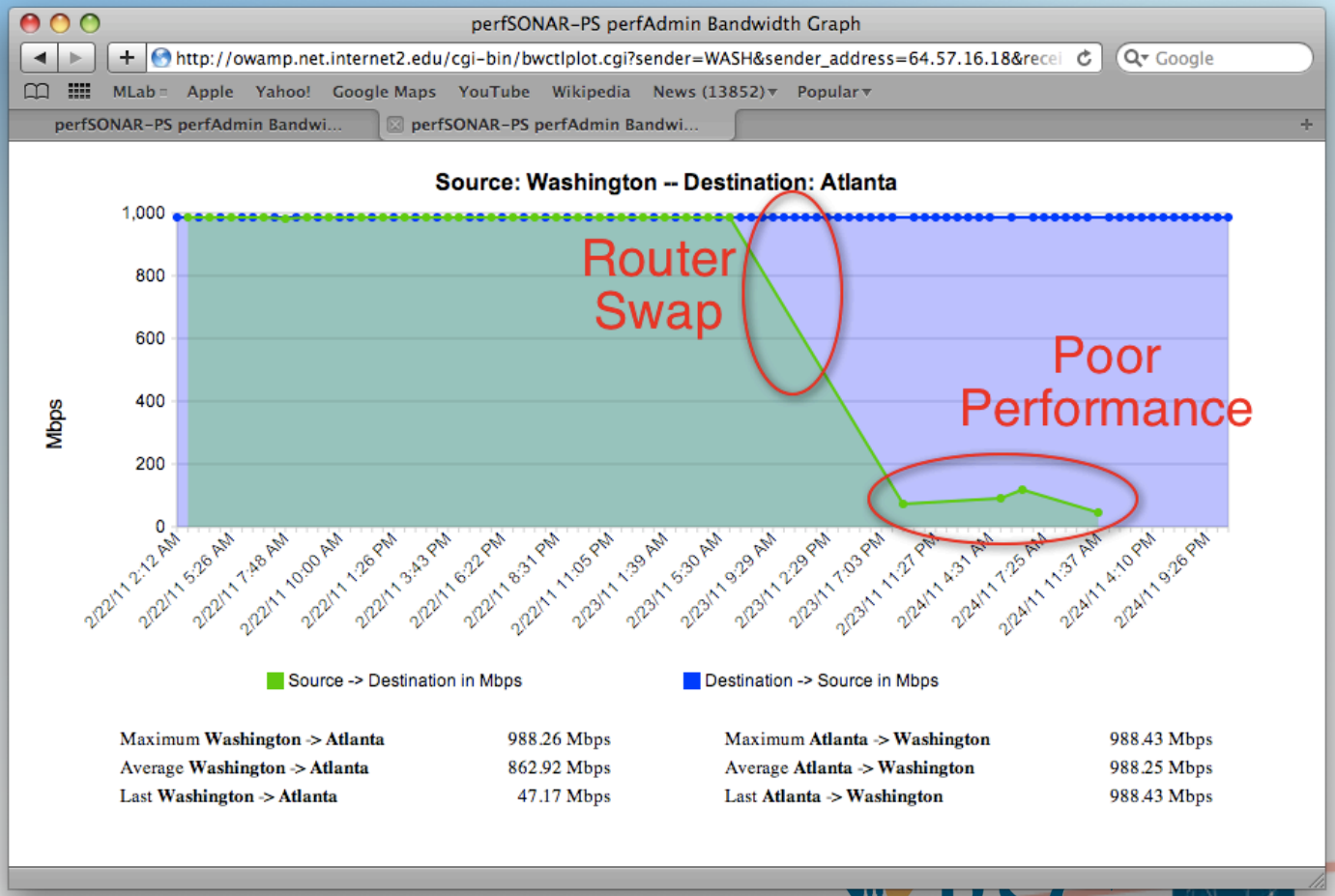
Real Debugging – Results (BWCTL)

- Digging Deeper on WASH:



Real Debugging – Results (BWCTL)

- Remember that trouble ticket …



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Real Debugging – Results Review

- Now we have several results
  - NDT diagnostics show poor results
    - PIC->Vanderbilt
    - WASH->Vanderbilt
    - WASH->ATLA
  - NDT diagnostics show good results
    - ATLA->Vanderbilt
  - BWCTL regular monitoring shows poor results
    - ATLA to WASH
    - ATLA to NEWY (which goes over the WASH path), we can ignore further debugging for here for now
  - BWCTL regular monitoring shows good results
    - Everywhere else
- Don't call it a day yet! One more tool too look at.

J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Real Debugging – Results (OWAMP)

- Much like BWCTL, we keep this going all the time. Also like BWCTL, we don't have alarms to tell us things are bad ☺

Real Debugging – Results (OWAMP)

- Interpreting the graph shows a pretty constant stream of loss (WASH -> ATLA). Note this is a "soft failure", not loss of connectivity

J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Problem Location and Reporting

- At this stage we have our evidence from all of the tools.
- Time to escalate – this is why we have 24/7/365 NOCs after all
- Problem reported Feb 24th 2011 @ 4pm EST
  - Evidence from tests above provided, lots of detail!
- First response from operations:

> Jason,
>
> I'm not seeing any degradation over that backbone link. Could you
> provide me with a trace? Perhaps there is loss occurring elsewhere along
> the path?
>
> Thanks,
>
> Greg

Denial?

- This first response should make anyone upset, especially after providing evidence from four (4!) tools
- To be fair … operations may have a different set of tools they are working with:
  - Monitoring of the Interface counters is something most are taught to watch – we revealed on Slide 11 that there was no evidence of errors. Utilization looked "ok"
  - Can't speak for the regular monitoring – these have been in place on the Internet2 observatory for around 6 years. Alarming is not in place at a minimum.

Ok, Not Complete Denial …

- Bringing in more eyes sometimes gets results, especially when they have looked at the evidence and can agree something doesn't smell right…

Entered on 02/24/2011 at 22:32:07 UTC (GMT+0000) by Tom Knoeller:
We are seeing errors increasing on the ATLA side. But light levels look good on both sides. Probably need to do some emergency work to throw some loops in the circuit to see where the problem is. My guess is going to be the XENPAK PIC on the WASH router as that is what changed 2 days ago, but testing will confirm that.

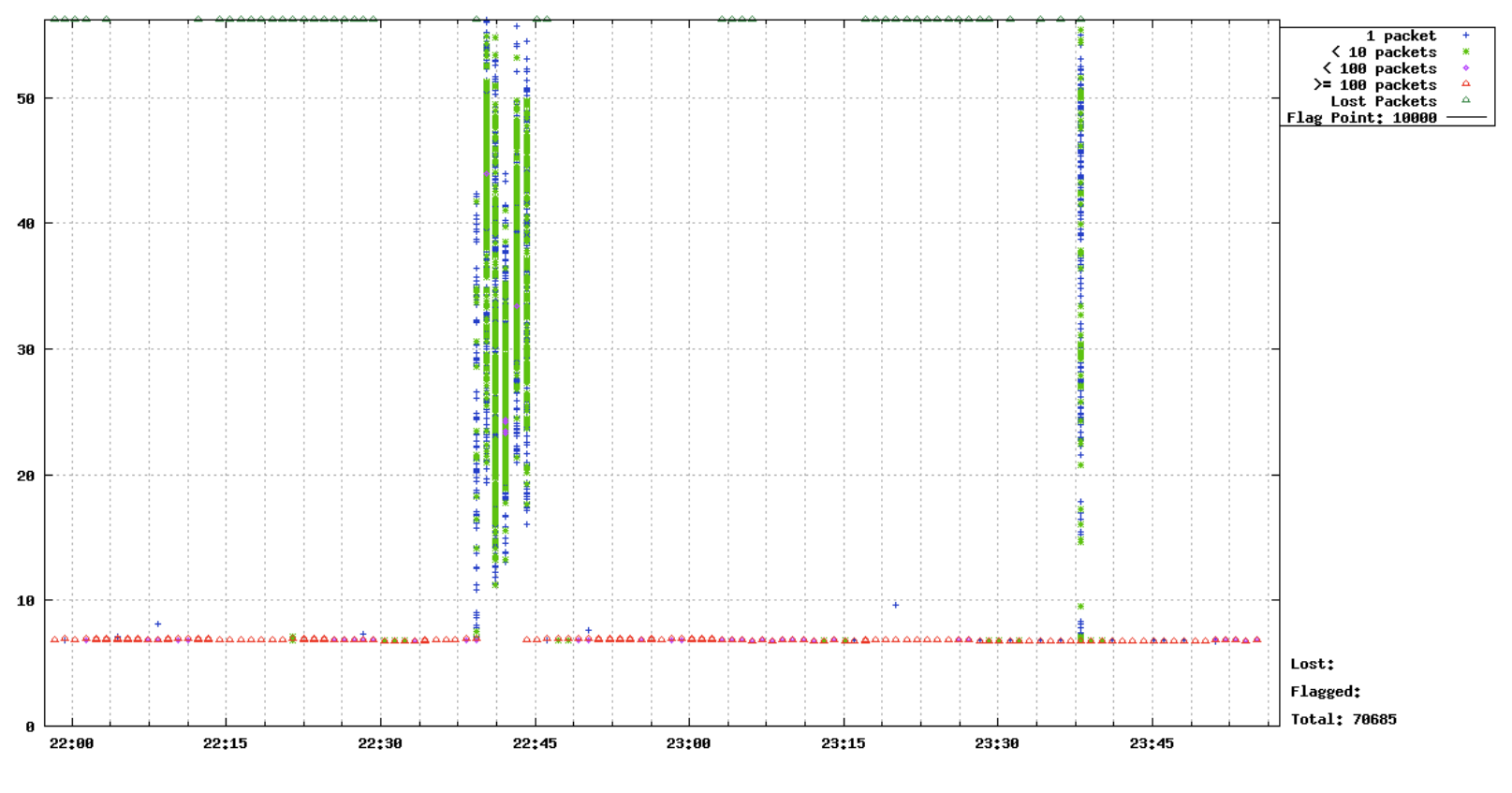
Traffic on the link is light, so I think we can turn off ISIS to divert the traffic without to much pain.

Service Desk: Lets get this into the hands of the oncall to work tonight.

Thanks,
-Tom

Testing Hypothesis

- When operations tweaks things, the tools know:



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Testing Hypothesis

- Explanation from the trouble ticket:

Entered on 02/24/2011 at 23:35:07 UTC (GMT+0000) by Tom Knoeller:

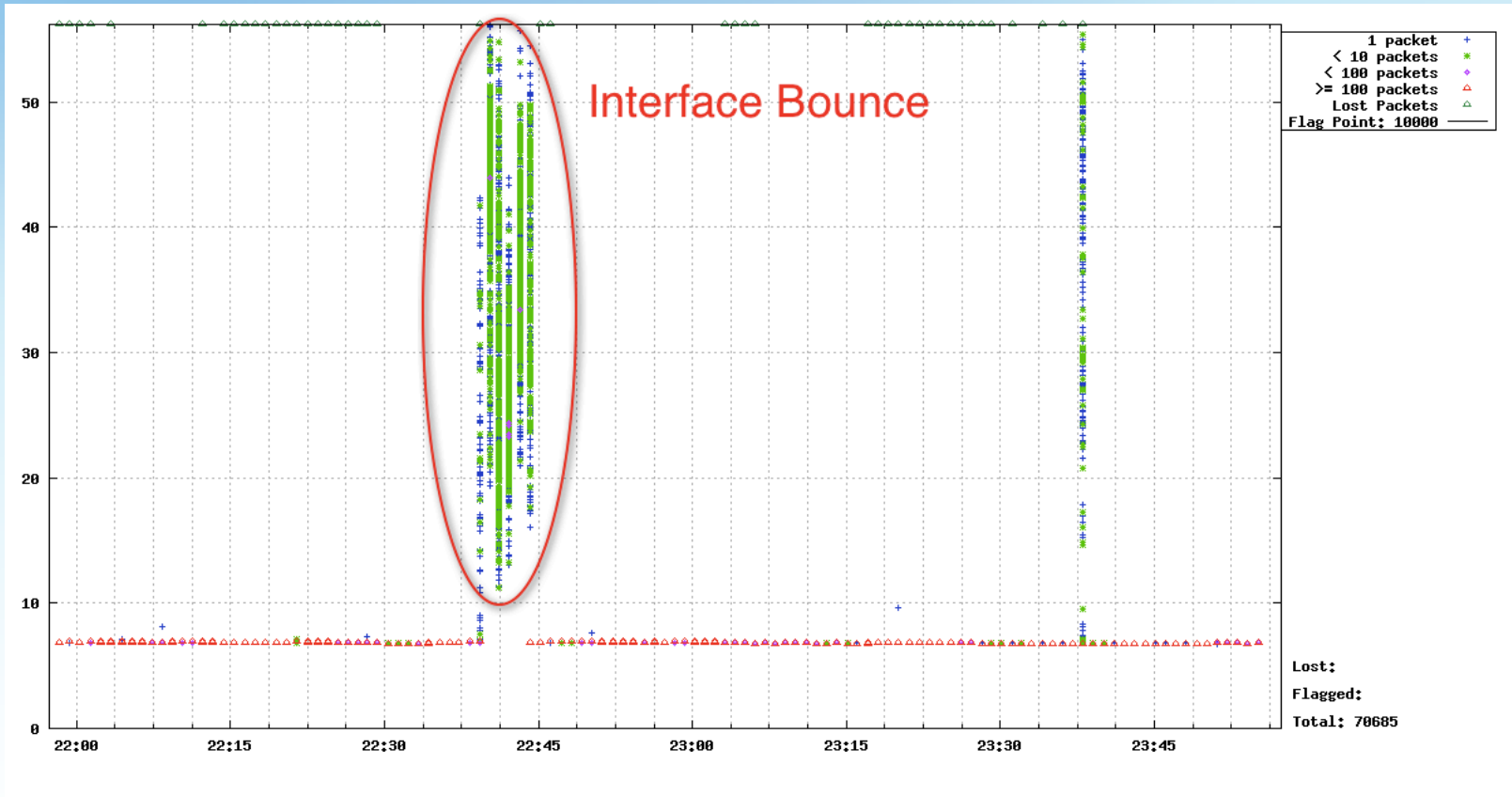
> Traffic on the link is light, so I think we can turn off
> ISIS to divert the traffic without to much pain.

And for those playing the home game, we tried to turn off the link, but I did not think about the offered load being higher with no packet loss, so it overloaded the other backbone link. At this point, the interface is turned on and running in a degraded state until a emergency FSR can be done to move to a new PIC.

-Tom

Testing Hypothesis

- Interpretation:



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Testing Hypothesis

- Next step:

Entered on 02/24/2011 at 23:47:08 UTC (GMT+0000) by Hans Addleman:
Tom suspected and I agreed that this might be the XENPAK optic failing in the T1600. I had a tech run a new fiber over to a new port (1/1/3) and the errrors are still being observed on the ATLA side.

Next course is going to be terminal looping the circuit and doing some testing to see if this is a layer 1 issue perhaps.

Hans Addleman
IU Global NOC Engineer
addlema@grnoc.iu.edu

- Maintenance was scheduled for Feb 24th 2011 @ 6:30PM EDT
  - If you are keeping track, this is only 2.5 hours since the ticket was opened

Solution In Place … Will It Hold?

- Not longer after swapping to a different interface:

Entered on 02/25/2011 at 00:02:08 UTC (GMT+0000) by Hans Addleman:
Okay.. it just took a minute for the counters to settle down.

The swap of interfaces fixed the problem! Traffic on that link jumped up by almost 3gig and the link looks healthy again.

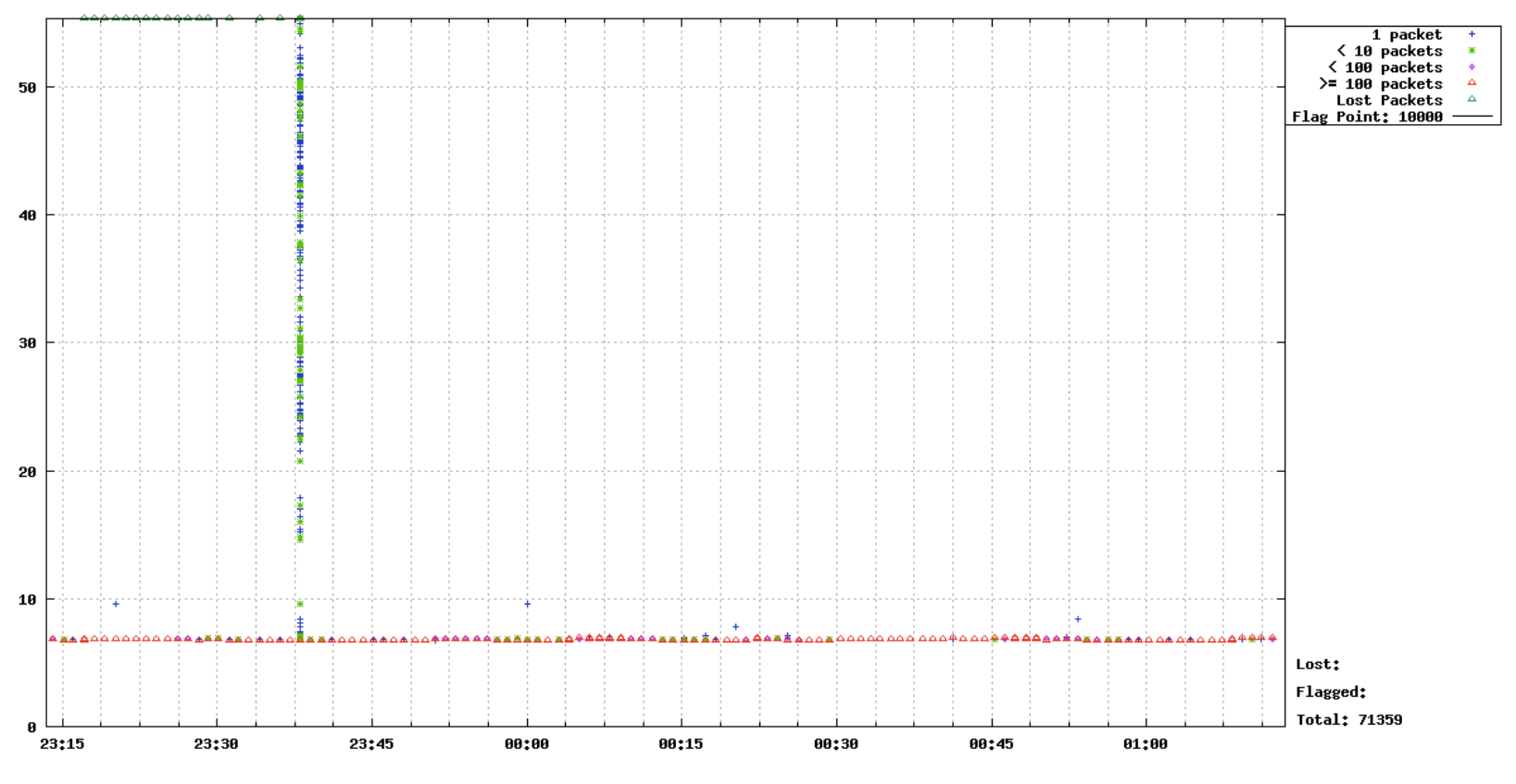
So we have a bad xenpak in WASH that we can worry about in the morning. Tom is going to work with Ross to start sending out spares to the sites.

Thanks to Tom for all the initial leg work on this.. Made my part of it this evening very easy.
Hans Addleman
IU Global NOC Engineer
addlema@grnoc.iu.edu

- And what do the tools say …

Solution In Place … Will It Hold?

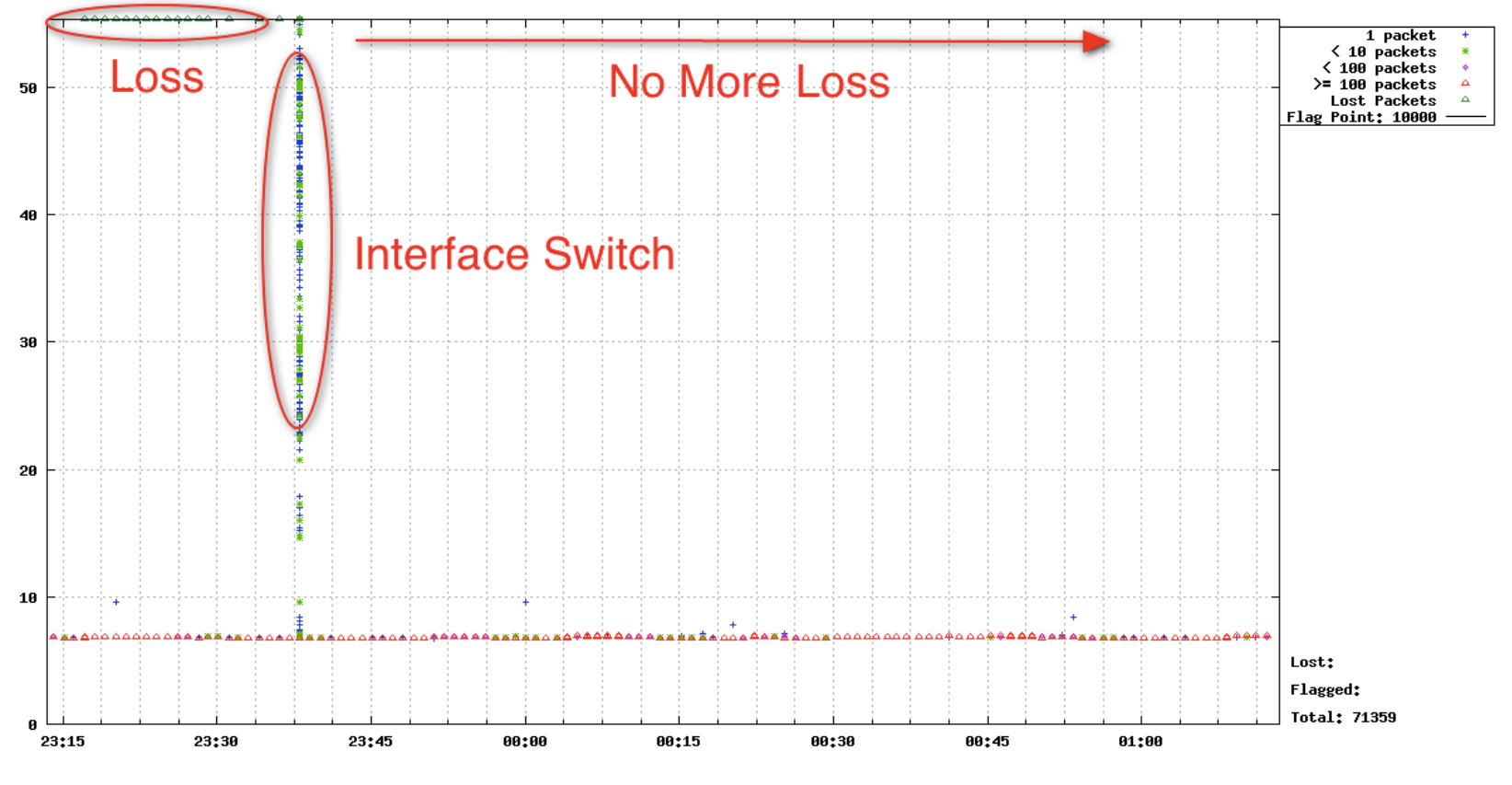
- OWAMP is sensitive, so lets go back to it:



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Solution In Place … Will It Hold?

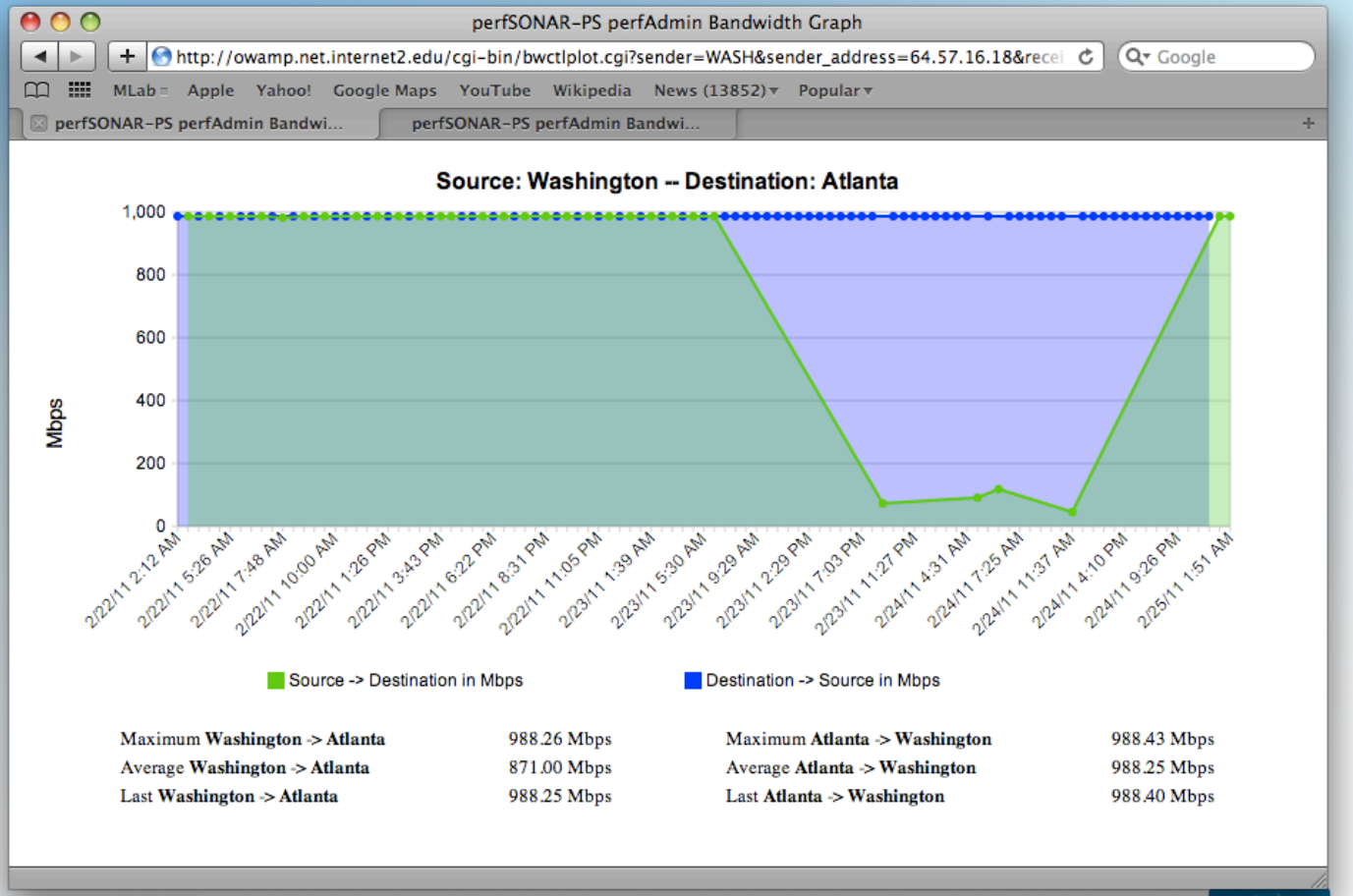
- Interpreting:



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Solution In Place … Will It Hold?

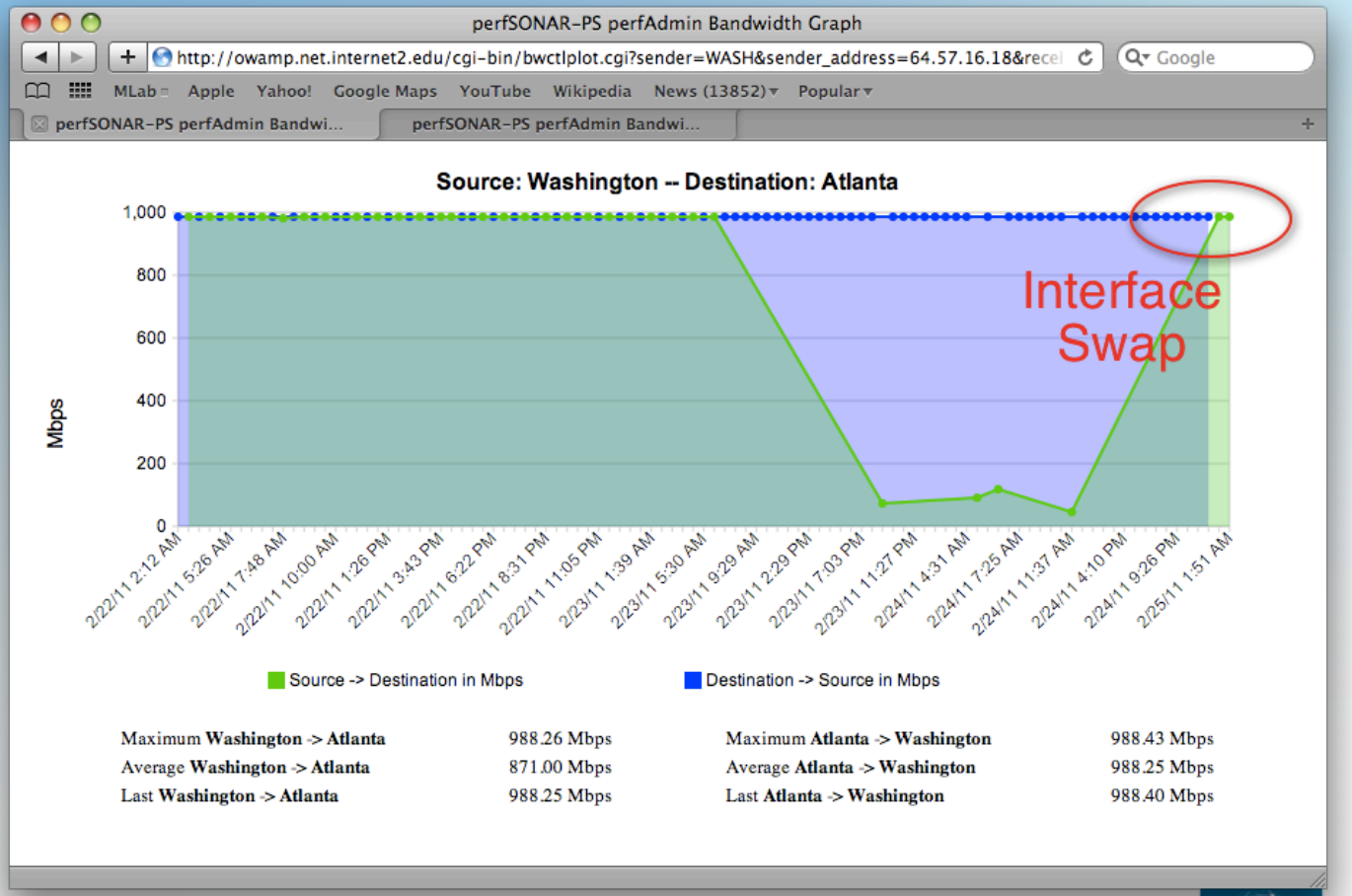
- What about BWCTL?



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Solution In Place … Will It Hold?

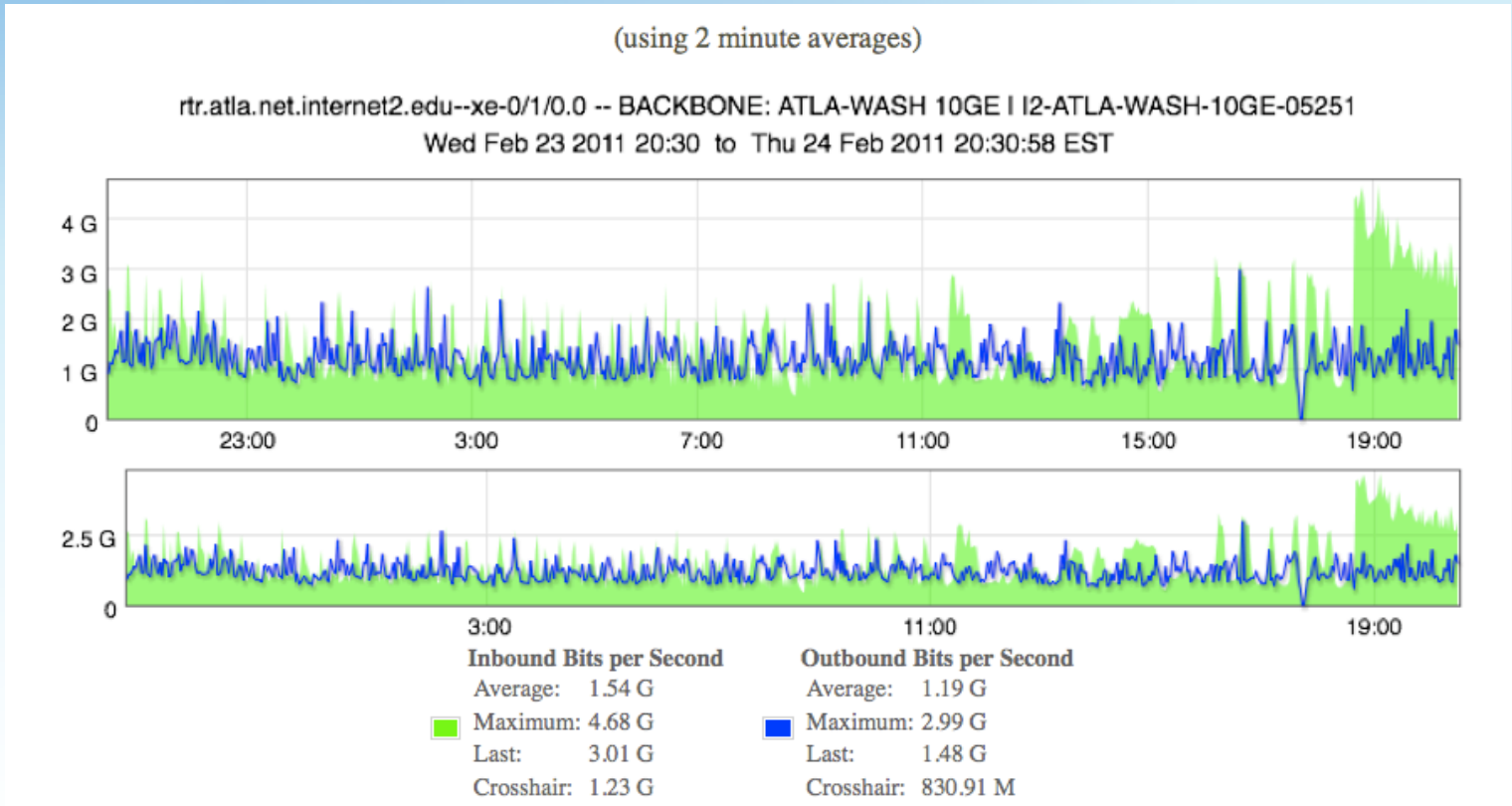
- Interpreting:



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Solution In Place … Will It Hold?

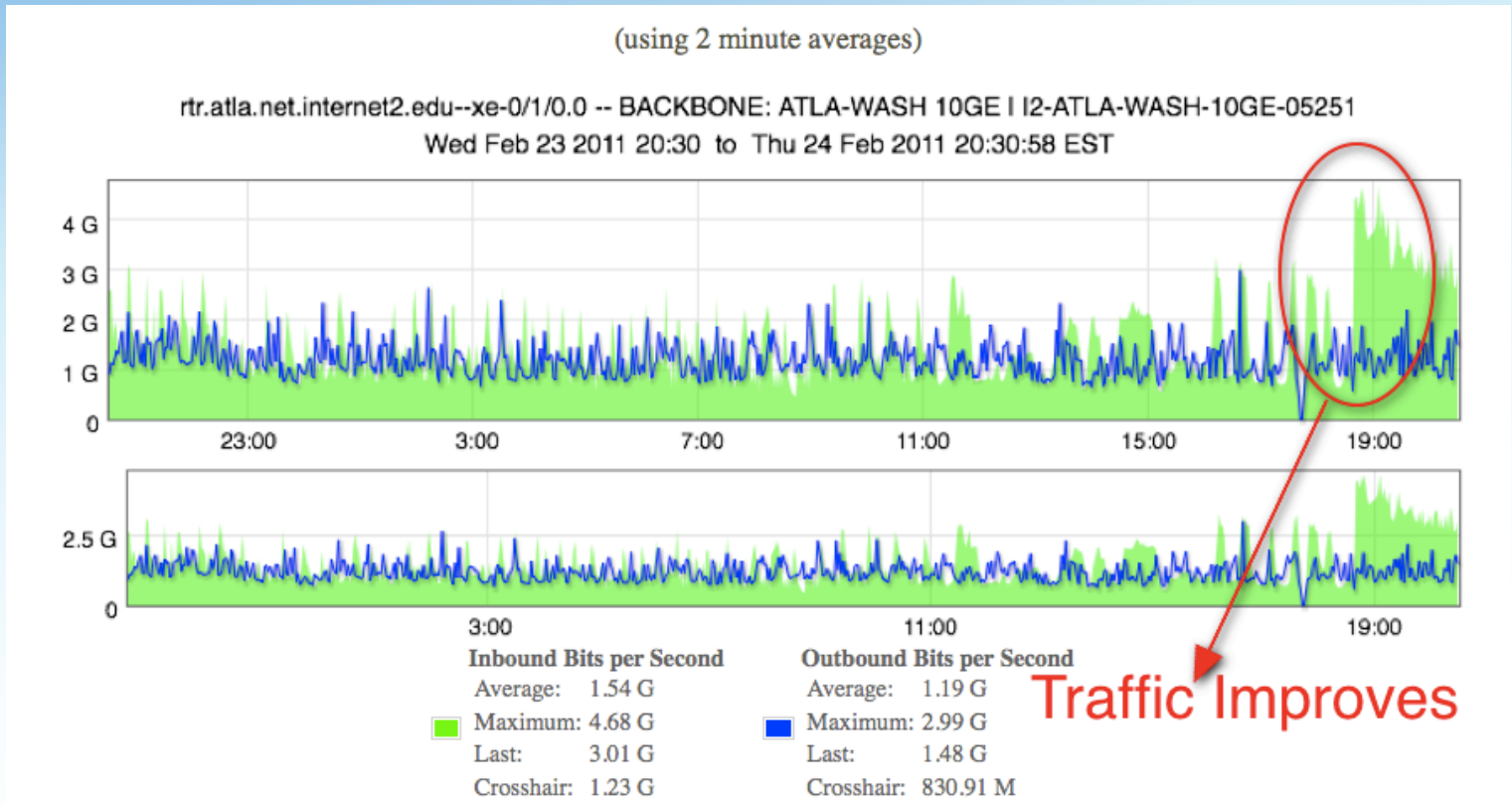
- Lastly, how about network utilization. In theory this should have limited all traffic…



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu

Solution In Place … Will It Hold?

- And it did …



Re-testing, Notification of Customer

- NDT is good for a one off, lets verify the paths again
- Vanderbilt (client) -> WASH (server)
  - running 10s outbound test (client to server) . . . . . 923.47 Mb/s
  - running 10s inbound test (server to client) . . . . . 914.02 Mb/s
- Vanderbilt (client) -> PIC (server)
  - running 10s outbound test (client to server) . . . . . 524.05 Mb/s
  - running 10s inbound test (server to client) . . . . . 550.64 Mb/s
- Not “perfect”, but closer
  - Client was asked to verify CMS applications
  - Debugging shouldn’t stop, there are more parts of the path to explore.

XSEDE Use Cases

XSEDE Use Cases

- 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
- Three use case examples from XSEDE
  - Jumbo frame MTU issues
  - Impact of small router buffers
  - Route changes

XSEDE Use Cases – MTU

- Site network configuration does not handle jumbo frames correctly.
  - bwctl testing connects, but subsequently fails to run
  - Manual bwctl testing fails and reports (example with names and addresses changed to protect the guilty):

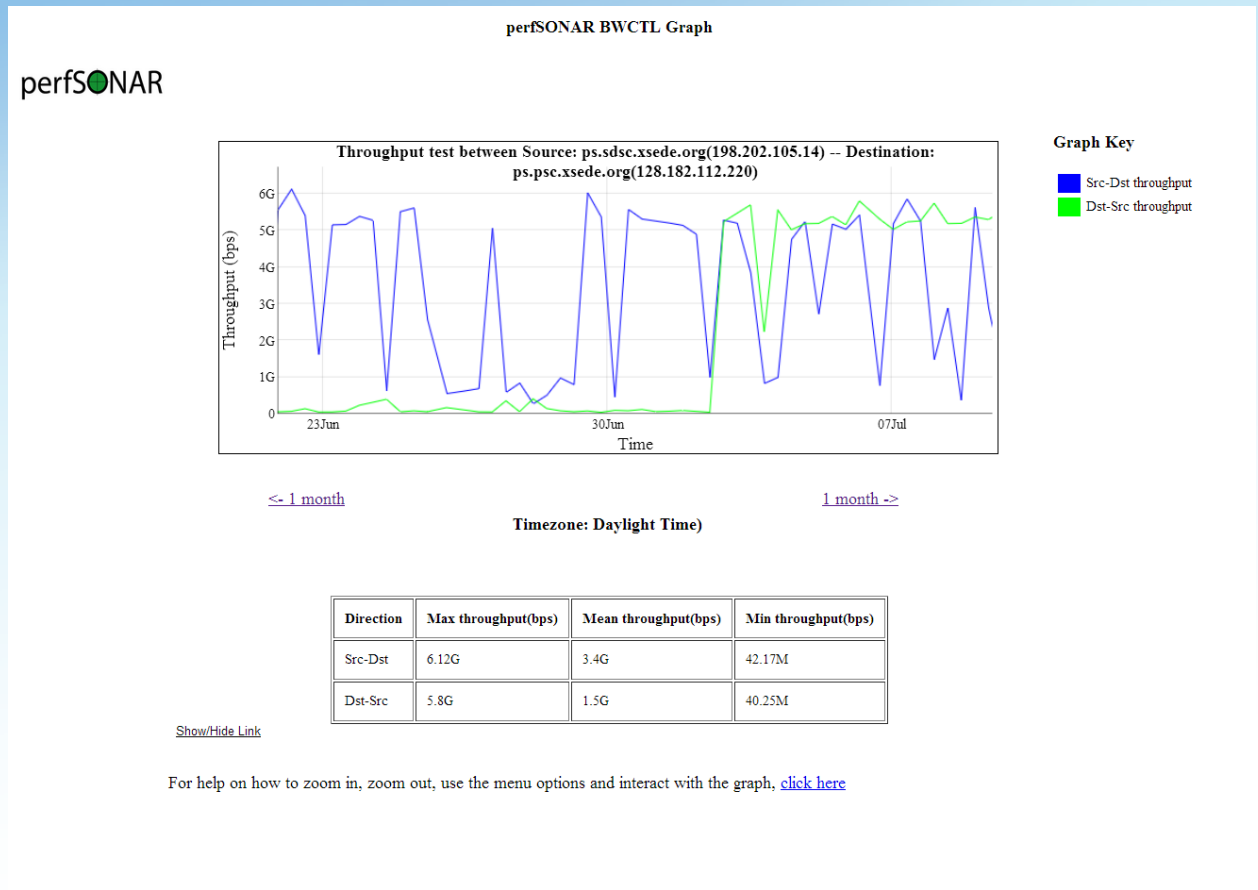
```
[benninge@perfsonar ~]$ bwctl -t 10 -i 2 -f m -L 300 -c net-test.WellKnownU.edu
bwctl: Using tool: iperf
bwctl: 17 seconds until test results available

RECEIVER START
bwctl: exec_line: iperf -B net-test.WellKnownU.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.WellKnownU.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
```

XSEDE Use Cases – Small router buffers

- 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 ->

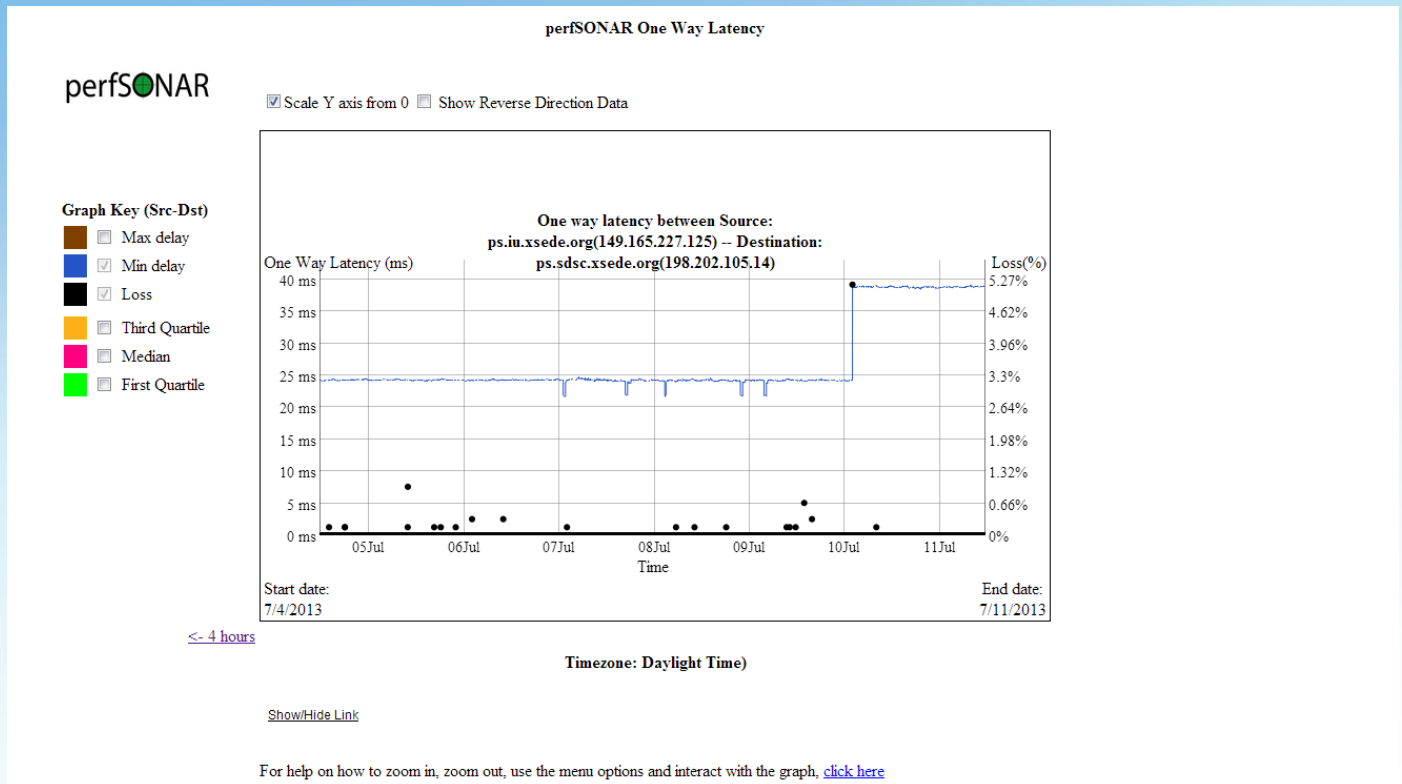
Timezone: Daylight Time)

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

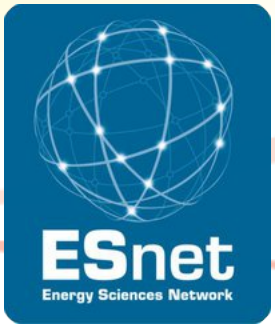
Show/Hide Link

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

XSEDE Use Cases – Route changes



J. Zurawski – zurawski@es.net & K. Benninger - benninger@psc.edu



Performance Use Cases

July 22nd 2013, XSEDE Network Performance Tutorial

Jason Zurawski – Internet2/ESnet
Kathy Benninger - Pittsburgh Supercomputing Center

For more information, visit http://www.internet2.edu/workshops/npw

perfSONAR powered