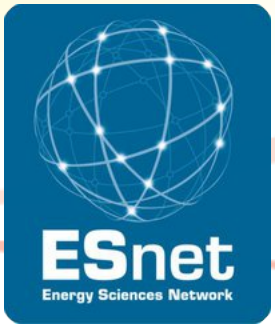


November 18th 2013, SC13 Network Performance Tutorial

Jason Zurawski – Internet2/ESnet

Welcome & Performance Primer

Who am I, Who are you?

Your Goals?

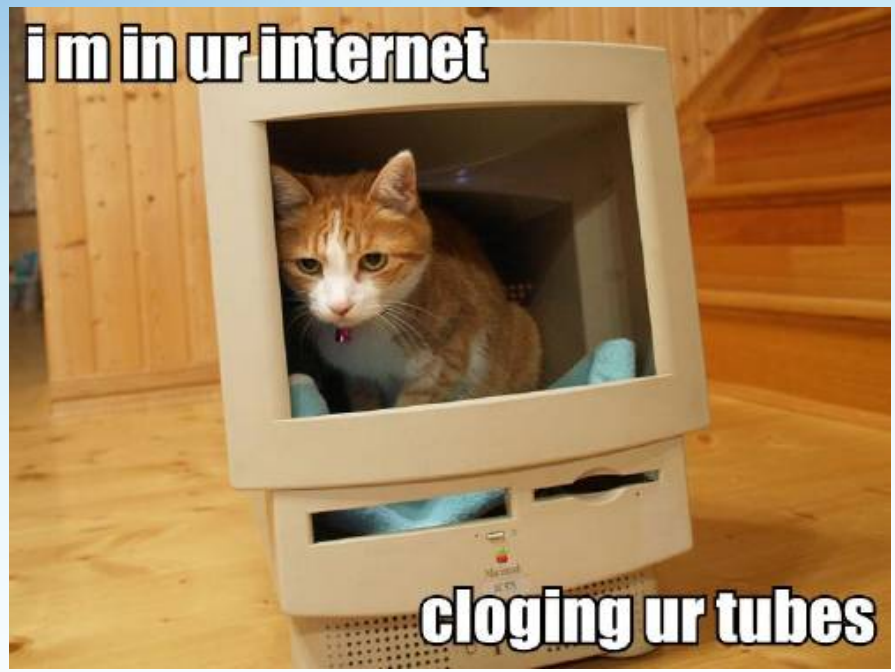
- What are your goals for this workshop?
  - Experiencing performance problems?
  - Responsible for the campus/lab network?
  - Learning about state of the art, e.g. 'What is perfSONAR'?
  - Developing or researching performance tools?
- Is there a Magic Bullet?
  - No, but we can give you access to strategies and tools that will help
  - Patience and diligence will get you to most goals
- This workshop is as much a learning experience for me as it is for you
  - What problem/problems need to be solved
  - What will make networking a less painful experience
  - How can we improve our goods/services

Problem: "The Network Is Broken"

- How can your users effectively report problems?
  - And how can you learn to take them seriously…
- How can users and the local administrators effectively solve multi-domain problems?
  - Eliminate the 'who you know' network to finding resources
  - Automate things when applicable
- Network as an instrument – should be as easy to use as possible
  - Smarter applications
  - Less 'friction'
- Components:
  - Tools to use
  - Questions to ask
  - Methodology to follow
  - How to ask for (and receive) help

Current State

- Traditional networking:
  - R&E or Commodity "TCP/IP" connectivity is subject to congestion by other users
    - TCP is sensitive to network use as well as physical flaws
    - Primary choice for application developers (reliability)
  - Supporting large sporadic flows is challenging for engineers



  - Need to worry about your network, as well as the networks of others (e.g. the end-to-end problem)
    - Can we 'see' how a network (or networks) are performing?
    - Can we dynamically change behavior and patterns?

The View From The Ivory Tower

- The End Game?
  - Many disciplines require a stable data transfer mechanism
  - Campuses/regionals have a duty to their customers to manage network traffic and deliver required bandwidth
- Recent calls to action
  - CC-NIE (NSF)
  - "Big Data" (NSF/NIH)
- End goal will be to make the campus and regional infrastructure ready for next generation of Networking
  - 100G
  - SDN
  - Science DMZ
  - ***Network Monitoring***

Science DMZ (in One Slide)

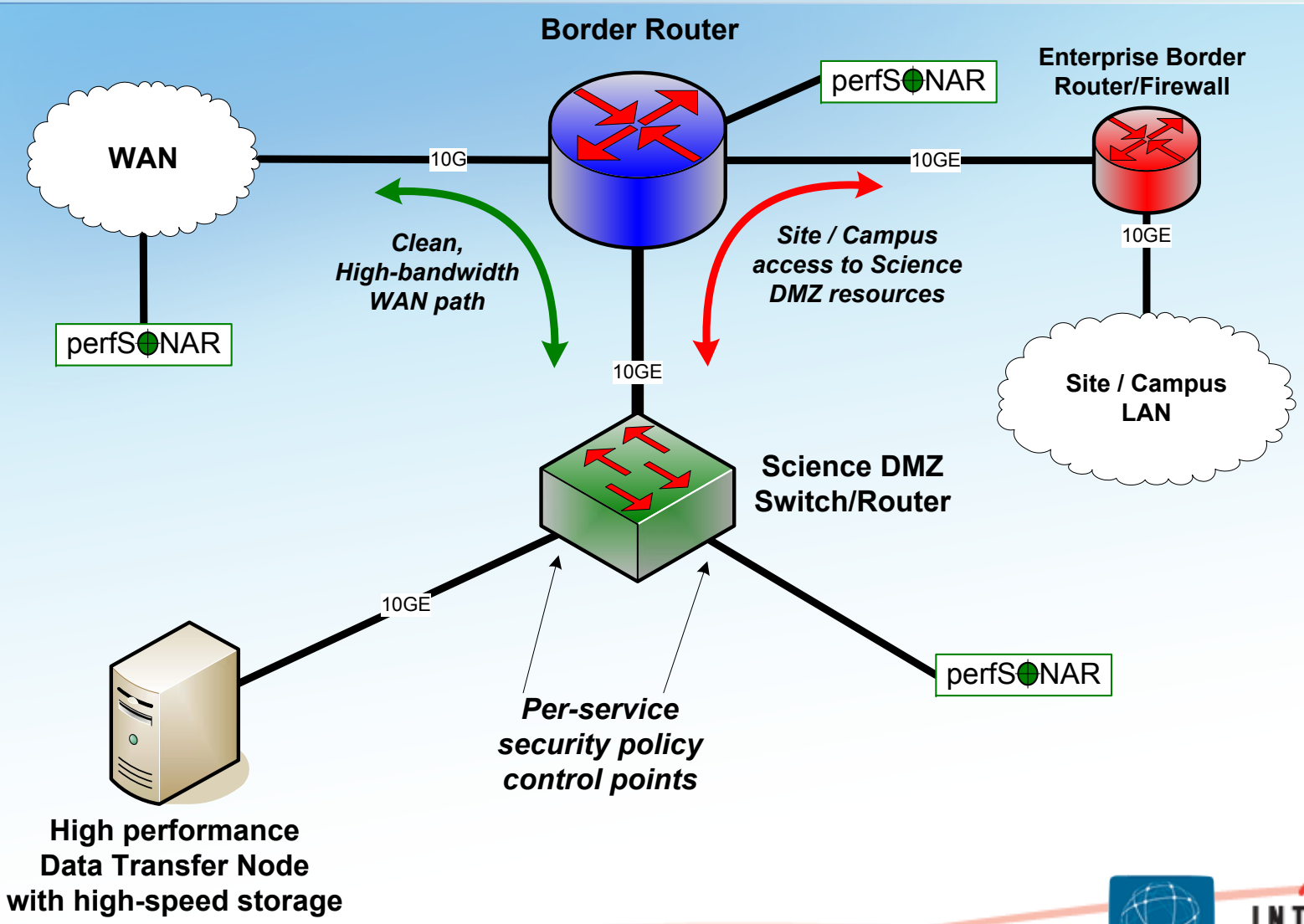
Consists of 3 key components, all required:

- "friction free" network path
  - Highly capable network devices (wire-speed, deep queues)
  - Virtual circuit connectivity option
  - Security policy and enforcement specific to science workflows
  - Located at or near site perimeter if possible
- Dedicated, high-performance data movers
  - a.k.a.: Data Transfer Node (DTN)
  - Optimized bulk data transfer tools such as GlobusOnline/GridFTP
- Performance measurement/test node
  - perfSONAR

Source: B. Tierney @ ESnet

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

Science DMZ Overview



The Problem Statement

- Data movement to support ~~science~~ advanced use cases:
  - Increasing in size (100s of TBs in the LHC World, approaching PB sizes)
  - Becoming more frequent (multiple times per day)
  - Reaching more consumers (VO sizes stand to increase, more VOs)
  - Time sensitivity (data may grow "stale" if not processed immediately)
  - Almost always "multi-domain"

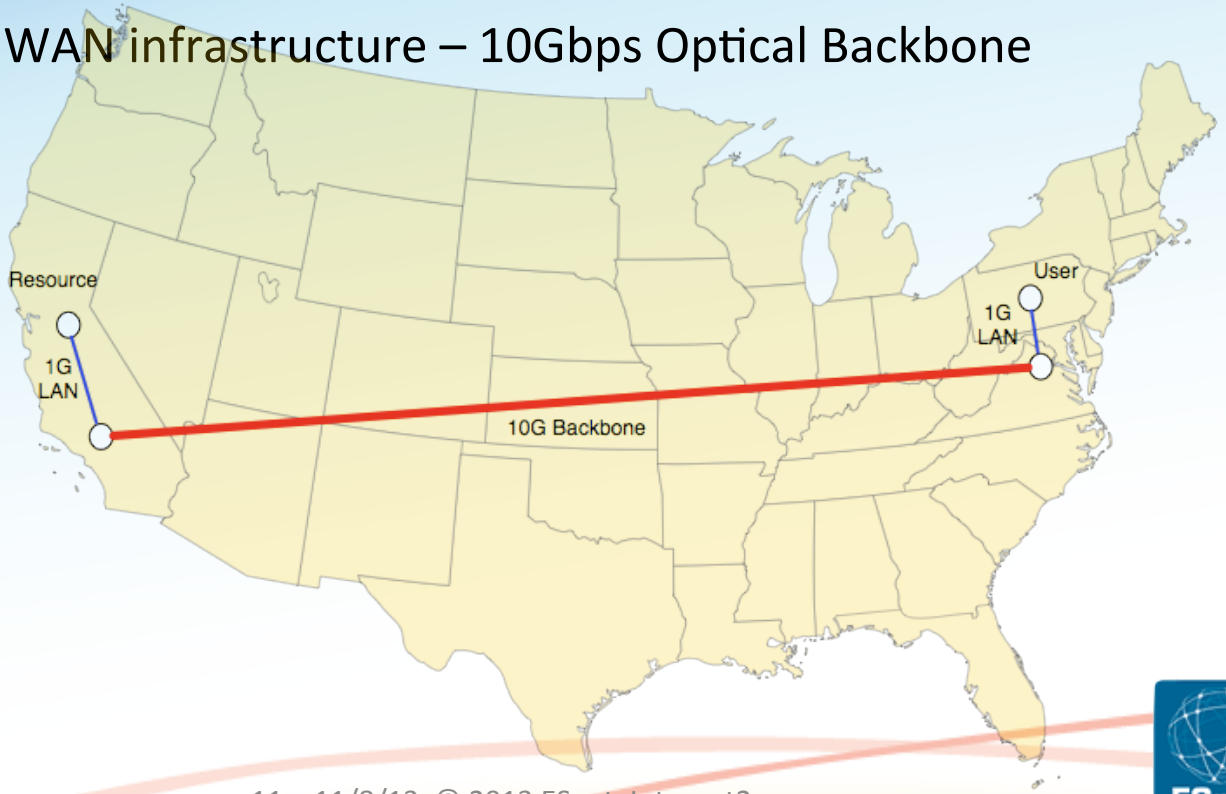
[1]

Motivation

- Proactive vs Reactive Positions
  - Do you want to find problems before the users do?
  - Can monitoring tools help in other aspects of operations?
    - Capacity Planning
    - Scheduling Maintenance
    - Traffic Engineering
- "The Network is broken", Is this justifable?
  - In actuality, there is a lot of "network" between the applications
  - What about those applications?
  - What about the host itself?
- Lets try to put this into an example …

Motivation – A Typical Scenario

- User and resource are geographically separated
  - Common case: Remote instrument + distributed users
- Both have access to high speed communication network
  - LAN infrastructure - 1Gbps Ethernet
  - WAN infrastructure – 10Gbps Optical Backbone



Motivation – A Typical Scenario

- User wants to access a file at the resource (e.g. ~600MB)
- Plans to use COTS tools (e.g. "scp", but could easily be something scientific like "GridFTP" or simple like a web browser)
- What are the expectations?
  - 1Gbps network (e.g. *bottleneck* speed on the LAN)
  - 600MB * 8 = 4,800 Mb file
  - User expects *line rate*, e.g. 4,800 Mb / 1000 Mbps = 4.8 Seconds
  - Audience Poll: ***Is this expectation too high?***
- What are the realities?
  - Congestion and other network performance factors
  - Host performance
  - Protocol Performance
  - Application performance

Motivation – A Typical Scenario

- Real Example (New York USA to Los Angeles USA):

```
[zurawski@nms-rthr2 ~]$ scp zurawski@bwctl.losa.net.internet2.edu:pS-Performance
_Toolkit-3.1.1.iso .
pS-Performance_Toolkit-3.1.1.iso              2%   17MB   1.0MB/s   10:05 ETA_
```

- 1MB/s (8Mb/s) ??? 10 Minutes to transfer???
- Seems unreasonable given the investment in technology
  - Backbone network
  - High speed LAN
  - Capable hosts
- Performance realities as network speed decreases:
  - 100 Mbps Speed – 48 Seconds
  - 10 Mbps Speed – 8 Minutes
  - 1 Mbps Speed – 80 Minutes
- How could this happen? More importantly, why are there not more complaints?
- Audience Poll: Would you complain? If so, to whom?
- Brainstorming the above – where should we look to fix this?

Motivation – A Typical Scenario

- Expectation does not even come close to experience, time to debug. Where to start though?
  - Application
    - Have other users reported problems? Is this the most up to date version?
  - Protocol
    - Protocols typically can be tuned on an individual basis, consult your operating system.
  - Host
    - Are the hardware components (network card, system internals) and software (drivers, operating system) functioning as they should be?
  - LAN Networks
    - Consult with the local administrators on status and potential choke points
  - Backbone Network
    - Consult the administrators at remote locations on status and potential choke points (Caveat – do you [should you] know who they are?)

Motivation – A Typical Scenario (cont.)

- Following through on the previous, what normally happens …
  - Application
    - This step is normally skipped, the application designer will *blame the network*
  - Protocol
    - These settings may not be explored. Shouldn't this be automatic (e.g. autotuning)?
  - Host
    - Checking and diagnostic steps normally stop after establishing connectivity. E.g. "can I ping the other side"
  - LAN Networks
    - Will assure "internal" performance, but LAN administrators will ignore most user complaints and shift blame to upstream sources. E.g. "our network is fine, there are no complaints"
  - Backbone Network
    - Will assure "internal" performance, but Backbone responsibilities normally stop at the demarcation point, blame is shifted to other networks up and down stream

* Denotes Problem Areas from Example

Why Worry About Network Performance?

- Most network design lends itself to the introduction of flaws:
  - Heterogeneous equipment
  - Cost factors heavily into design – e.g. *Get what you pay for*
  - Design heavily favors ***protection*** and ***availability*** over performance
- Communication protocols are not advancing as fast as networks
  - *TCP/IP* is the king of the protocol stack
    - Guarantees reliable transfers
    - Adjusts to failures in the network
    - Adjusts speed to be *fair* for all
- User Expectations
  - ***Big Science*** is prevalent globally
  - "The Network is Slow/Broken" – is this the response to almost any problem? Hardware? Software?
  - Empower users to be more informed/more helpful

Why is Science Data Movement Different?

- Different Requirements
  - Campus network is not designed for large flows
    - ***Enterprise*** requirements
    - 100s of Mbits is common, any more is rare (or viewed as *strange*)
    - Firewalls
    - Network is designed to mitigate the risks since the common hardware (e.g. Desktops and Laptops) are un-trusted
  - Science is different
    - Network needs to be robust and stable (e.g. predictable performance)
    - 10s of Gbits of traffic (N.B. that its probably not sustained – but could be)
    - Sensitive to enterprise protections (e.g. firewalls, LAN design)
- ***Fixing*** is not easy
  - Design the base network for science, attach the enterprise on the side (expensive, time consuming, and good luck convincing your campus this is necessary…)
  - Mitigate the problems by moving your science equipment to the edge
    - Try to bypass that firewall at all costs
    - Get as close to the WAN connection as you can

Identifying Common Network Problems

- Internet2/ESnet engineers will help members and customers debug problems if they are escalated to us
  - Goal is to solve the entire problem – end to end
  - Involves many parties (typical: End users as well as Campus, Regional, Backbone staff)
  - Slow process of locating and testing each segment in the path
  - Have tools to make our job easier (more on this later)
- Common themes and patterns for ***almost every*** debugging exercise emerge
  - Architecture (e.g. LAN design, Equipment Choice, Firewalls)
  - Configuration
  - "Soft Failures", e.g. something that doesn't sever connectivity, but makes the experience unpleasant

Stumbling Blocks – The Concerns

- Network Design
  - Balancing the needs of all users (e.g. how does video differ from bulk data transfer)
  - An ounce of prevention (e.g. configuration, monitoring)
  - You care about your network, is it your job to care about the network of your peers?
- Packet Loss
  - "Congestive"; the realities of a general purpose network
  - "Non-Congestive"; fixable, if you can find it
    - Clean your fibers!
    - Throw away the crimped cable!
    - Increase your buffers!

Stumbling Blocks – Network Design

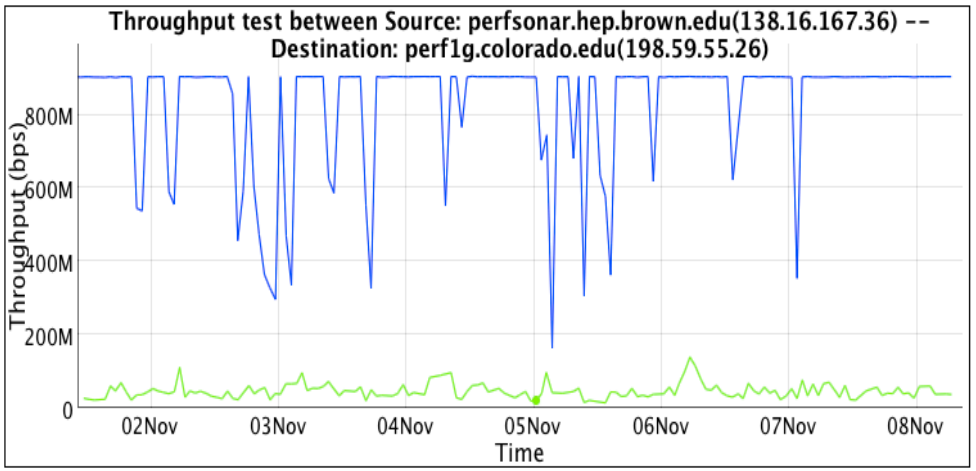
- LAN vs WAN Design
  - Multiple Gbit flows [to the outside] should be close to the WAN connection
  - Eliminate the number of hops/devices/physical wires that may slow you down and add delay (buffering)
  - Great performance on the LAN != Great performance on the WAN
    - Think about how TCP works – latency plays a big role in recovering from loss
- *You Get What you Pay For*
  - Inexpensive equipment will let you down
  - What could go wrong?
    - Small buffers, potentially shared, creates questionable performance (e.g. internal switching fabric can't keep up demands)
    - Lack of diagnostic tools (SNMP, etc.)
- Default configurations are (***always***) bad
  - Hosts, Switches/Routers

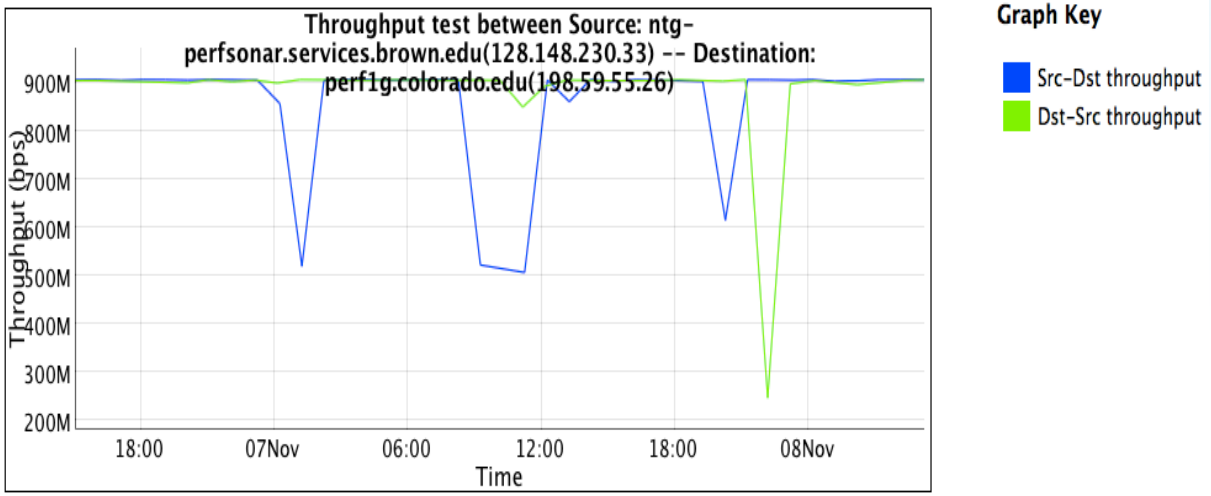
Stumbling Blocks – Firewalls/Shapers

- Designed to stop 'traffic'
  - Read this slowly a couple of times…
  - Performing a read of headers and/ or data. Matching signatures
- Contain small buffers
  - Concerned with protecting the network, not impacting your performance

- Will be ***a lot*** slower than the original wire speed
  - A "***10G Firewall***" may handle 1 flow close to 10G, doubtful that it can handle a couple.
- If *firewall-like* functionality is a must – consider using router filters instead
  - Or per host firewall configurations …

Stumbling Blocks – Firewalls/Shapers





Sensible Security

- Security must be viewed as a 'system'
  - Component based:
    - A firewall (hopefully one that is updated and monitored)
    - Federated identity
    - Etc.
  - System based:
    - Comprehensive Campus CI Plan
    - Identification of data risks (PHI, users, etc.)
    - Identification of hardware risks (its not just servers, HVAC, Phones, Printers, etc. are on the net too…)
- "You're doing it wrong"
  - Its true having a firewall ensures that if something goes wrong, you still have a job the next day
  - It's a greater sin to install a firewall, learn little about it, lapse in software updates, and stand behind it as the law of the land
    - E.g. network attacks favor the attacker, once they figure out vulnerable existing system software or hardware must be updated.

Stumbling Blocks – Packet Loss

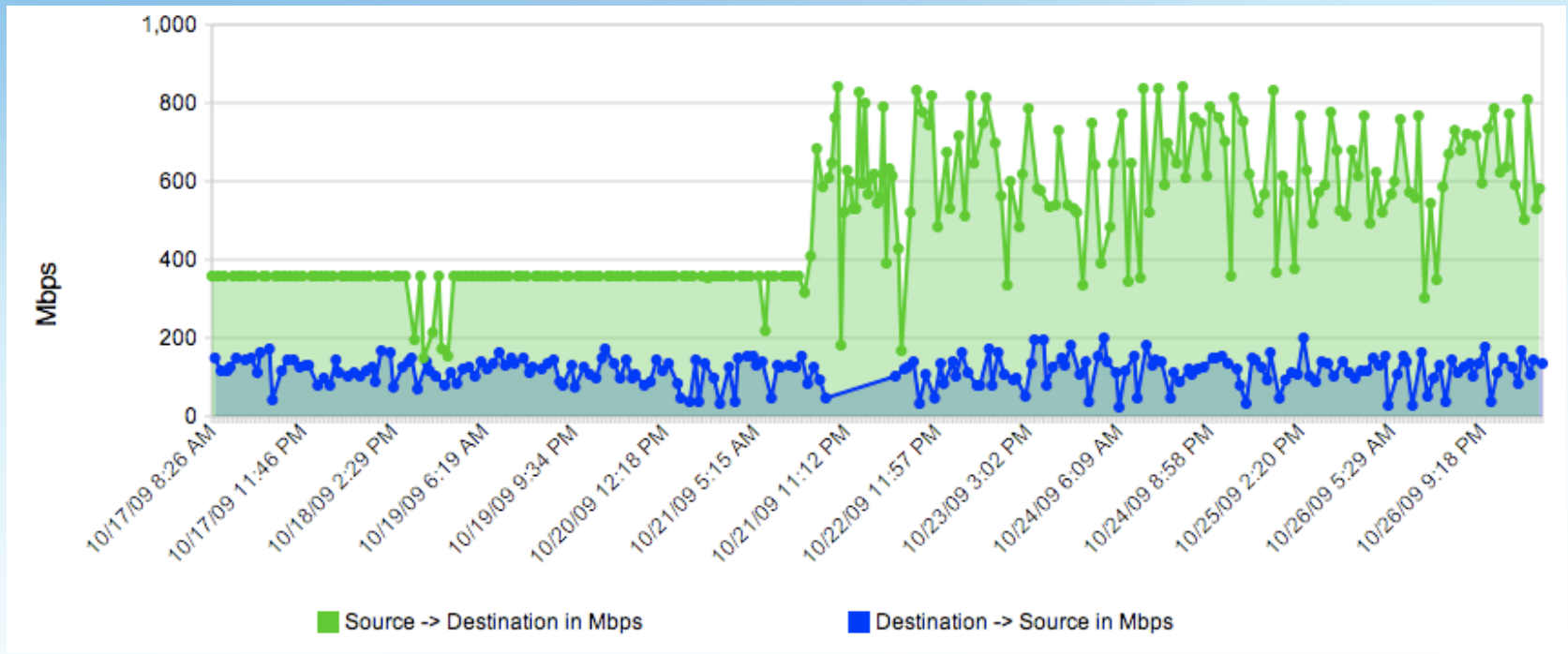
- Bandwidth Delay Product & Buffering
  - The amount of "in flight" data allowed for a TCP connection
  - BDP = bandwidth * round trip time
  - Example: 1Gb/s cross country, ~100ms
    - 1,000,000,000 b/s * .1 s = 100,000,000 bits
    - 100,000,000 / 8 = 12,500,000 bytes
    - 12,500,000 bytes / (1024*1024) ~ 12MB
- "Buffer Bloat"
  - Less of a concern in the R&E community; the added delay you get with too much buffering on a (low speed) connection
- TCP Dynamics (e.g. congestion control algorithms)
  - Additive-increase/Multiplicative-decrease [AIMD]
  - E.g. You cut your speed in half (sometimes less) with each loss.
  - Slowly increase to your prior speed and hope you don't take more loss.
    - Think about a short path with a lot of loss
    - Think about a long path with little loss

Stumbling Blocks – Configuration

- Host Configuration
  - Tune your hosts (especially compute/storage!)
  - Changes to several parameters can yield 4 – 10 x improvement
  - Takes minutes to implement/test
  - Instructions: http://fasterdata.es.net/tuning.html
- Network Switch/Router Configuration
  - ***Out of the box*** configuration may include small buffers
  - Competing Goals: Video/Audio etc. needs small buffers to remain responsive. Science flows need large buffers to push more data into the network.
  - Read your manuals and test LAN host to a WAN host to verify (not LAN to LAN).

Stumbling Blocks – Configuration – cont.

- Host Configuration – spot when the settings were tweaked…



- N.B. Example Taken from REDDnet (UMich to TACC), using BWCTL measurement)

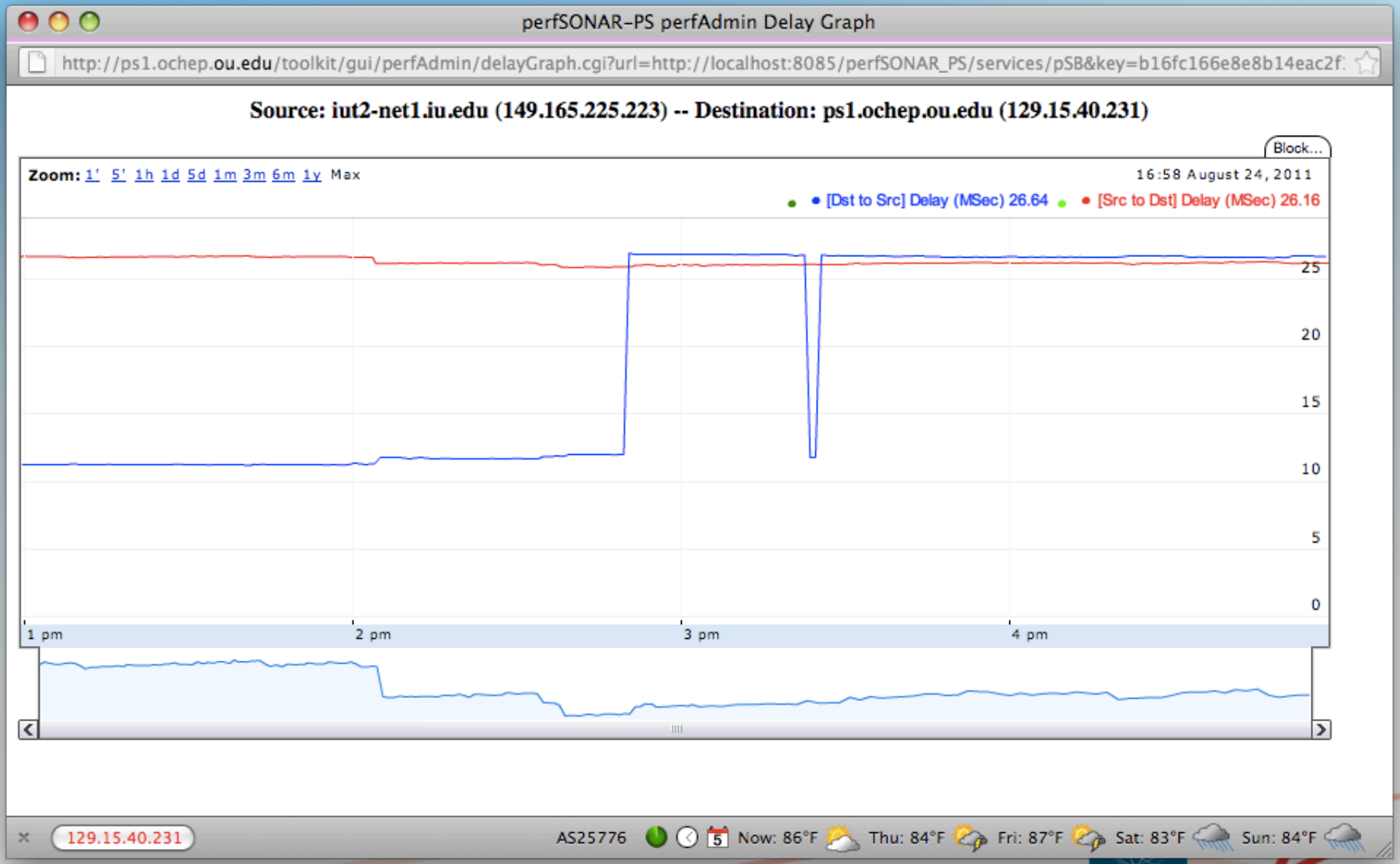
Stumbling Blocks - Congestion

- The end goal is enabling true R&E use of the network
  - Most research use follows the 'Elephant' Pattern. You can't stop the elephant and inspect it's hooves without causing a backup at the door to the circus tent
  - Regular campus patterns are often 'mice', small, fast, harder to track on an individual basis (e.g. we need big traps to catch the mice that are dangerous)
  - Security and performance can work well together – it requires critical thought (read that as ***time***, ***people***, and perhaps ***money***)
  - Easy economic observation – impacting your researchers with slower networks makes them less competitive, e.g. they are pulling in less research dollars vs. their peers

Soft Failures

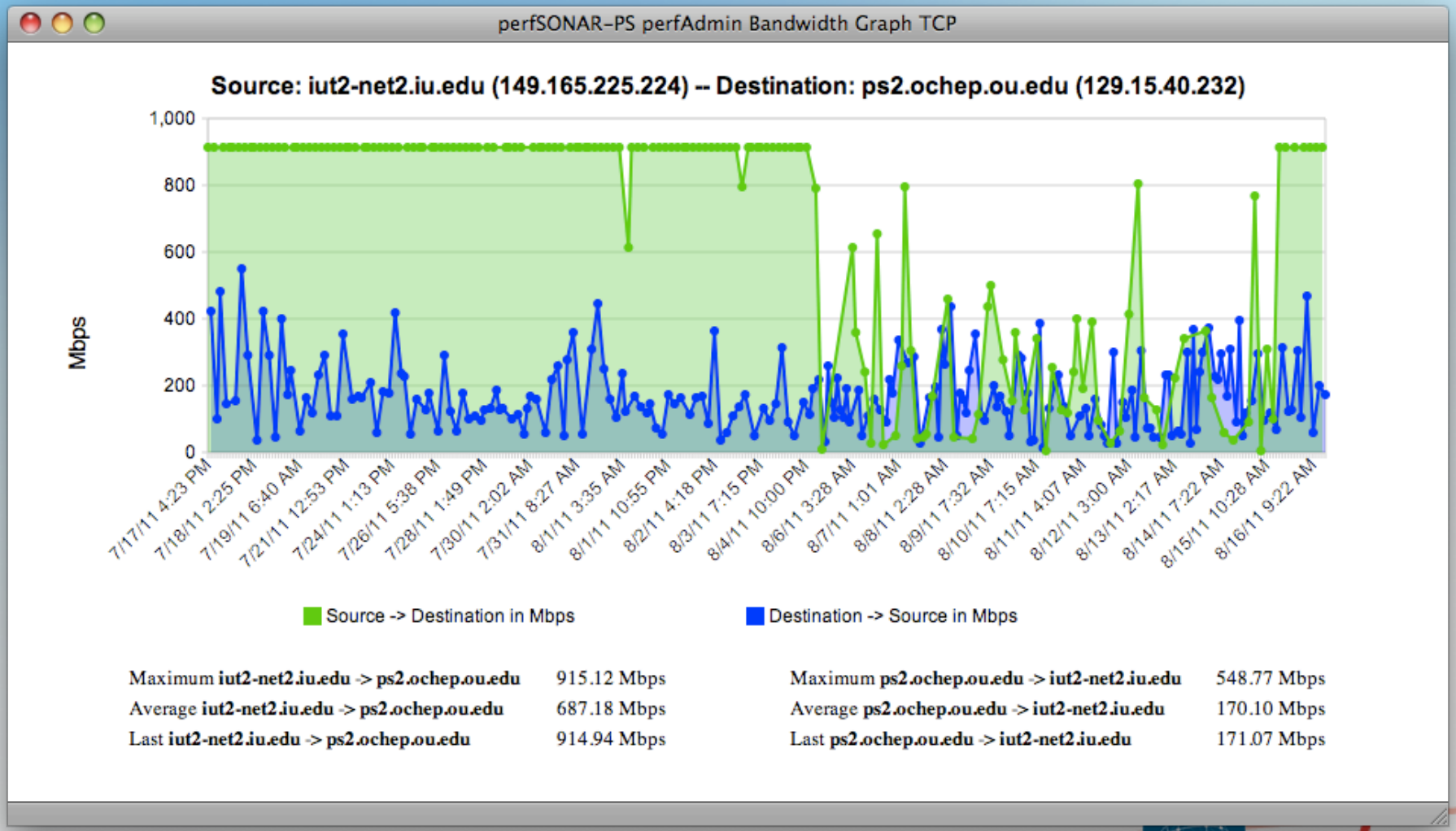
- ***Soft Failures*** are any network problem that does not result in a loss of connectivity
  - Slows down a connection
  - Hard to diagnose and find
  - May go unnoticed by LAN users in some cases, but remote users may be the ones complaining
    - Caveat – How much time/energy do you put into listing to complaints of remote users?
- Common:
  - Dirty or Crimped Cables
  - Failing Optics/Interfaces
  - [Router] Process Switching, aka "*Punting*"
  - Router Configuration (Buffers/Queues)

Asymmetric Routing - Latency

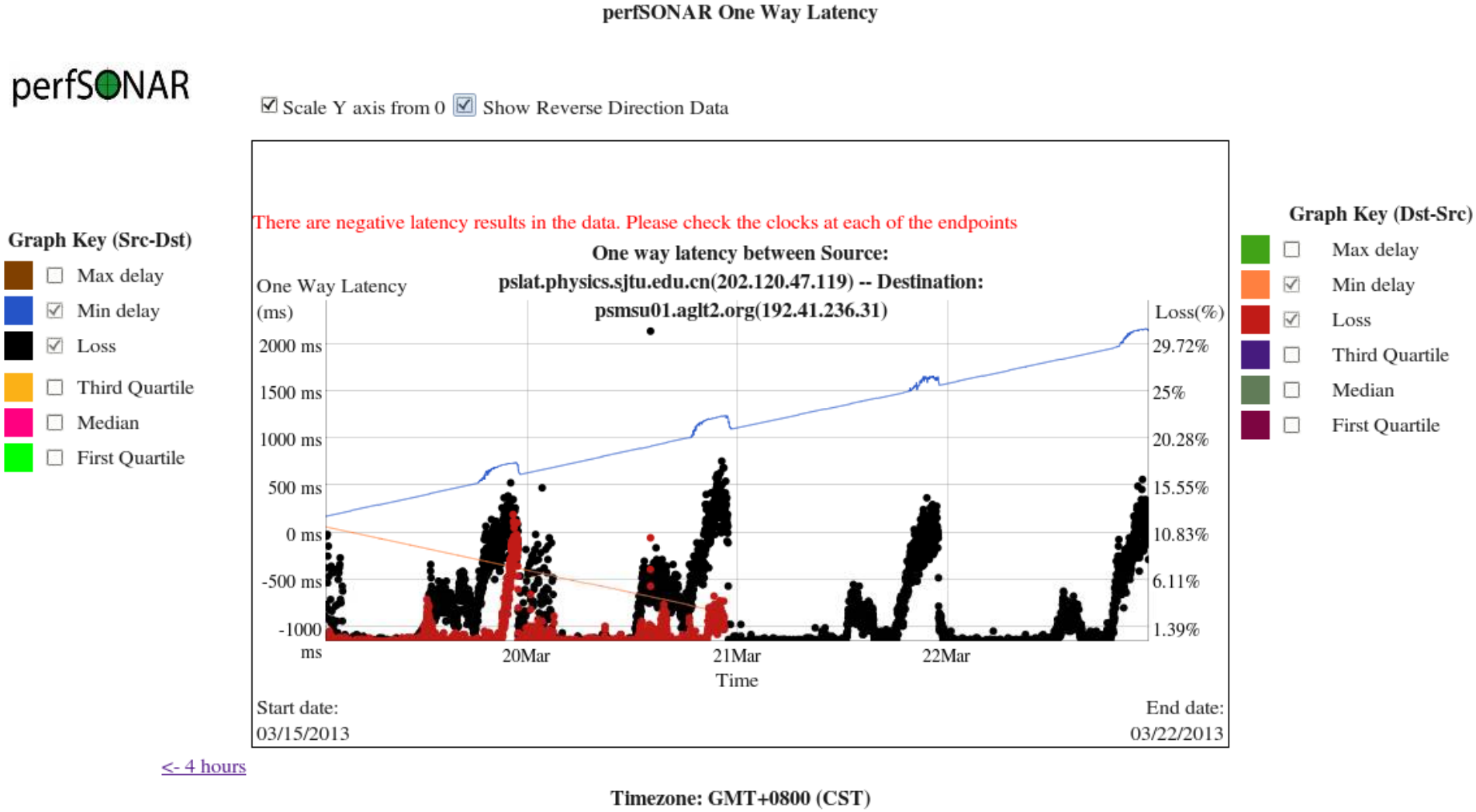


Asymmetric Routing – Loss on Commodity

Asymmetric Routing – Bandwidth

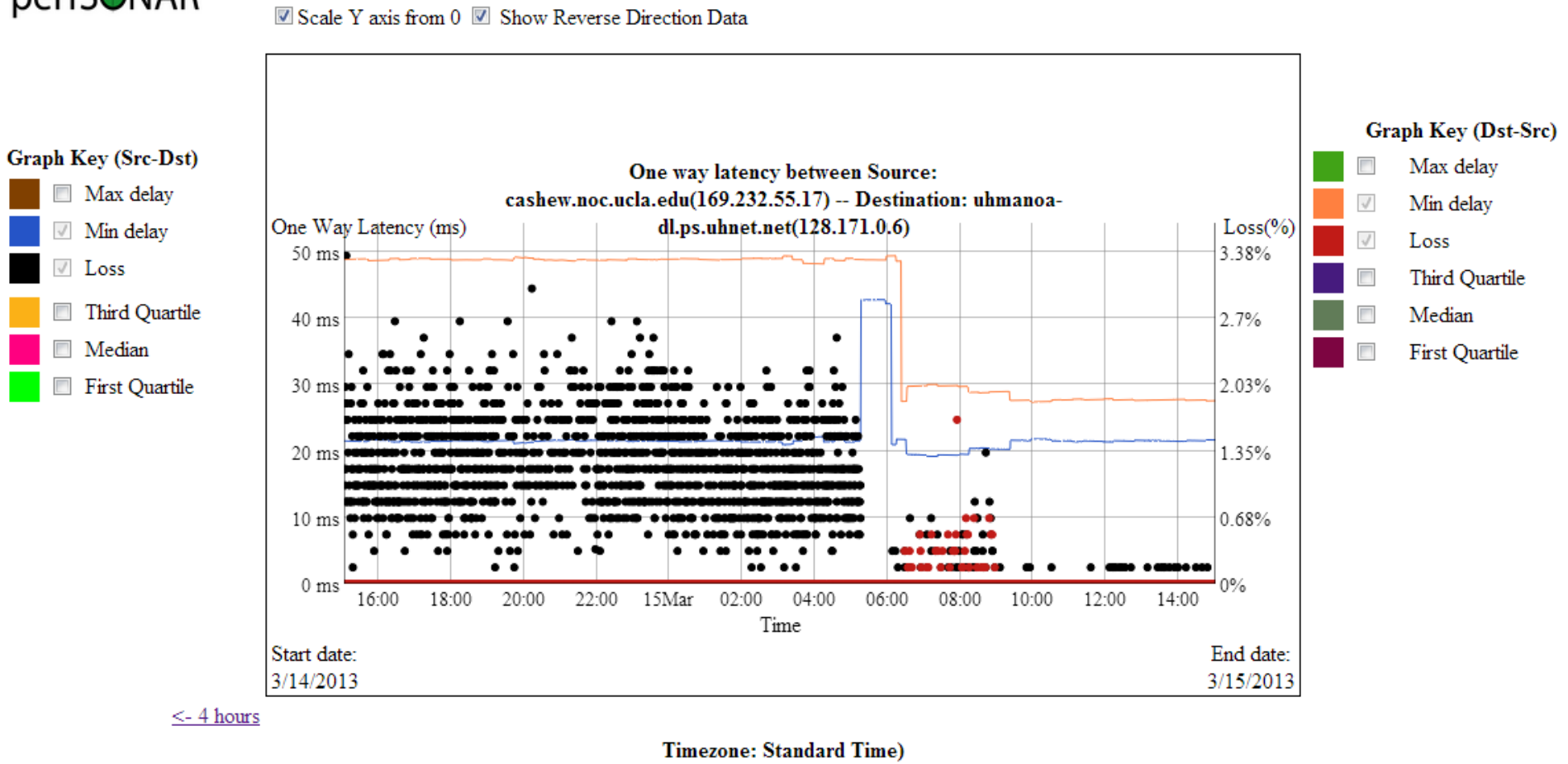


Congestion on Link + Drifting Clock



Adding Attenuator to Noisy Link

perfSONAR



Topics of Discussion

- Diagnosis Methodology
  - Find a measurement server "near me"
    - Why is this important?
    - How hard is this to do?
  - Encourage user to participate in diagnosis procedures
  - Detect and report common faults in a manner that can be shared with admins/NOC
    - 'Proof' goes a long way
  - Provide a mechanism for admins to review test results
  - Provide feedback to user to ensure problems are resolved

Topics of Discussion – cont.

- Partial Path Decomposition
  - Networking is increasingly:
    - Cross domain
    - Large scale
    - Data intensive
  - Identification of the end-to-end path is key (must solve the problem end to end…)
  - Discover measurement nodes that are “near” this path
  - Provide proper authentication or receive limited authority to run tests
    - No more conference calls between 5 networks, in the middle of the night
  - Initiate tests between various nodes
  - Retrieve and store test data for further analysis

Topics of Discussion – cont.

- Systematic Troubleshooting
  - Having tools deployed (along the entire path) to enable adequate troubleshooting
  - Getting end-users involved in the testing
  - Combining output from multiple tools to understand problem
    - Correlating diverse data sets – only way to understand complex problems.
  - Ensuring that results are adequately documented for later review
- On Demand vs Regular Testing
  - On-Demand testing can help solve existing problems once they occur
  - Regular performance monitoring can quickly identify and locate problems before users complain
    - Alarms
    - Anomaly detection
  - Testing and measuring performance increases the value of the network to all participants

Welcome & Performance Primer

November 18th 2013, SC13 Network Performance Tutorial

Jason Zurawski – Internet2/ESnet

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

https://submissions.supercomputing.org/?page=SessionEval&id=sess118

perfSONAR powered