August 9th 2011, OSG Site Admin Workshop

Jason Zurawski – Internet2 Research Liaison

Welcome & Performance Primer

Who are we, Who are you?

Agenda

- Welcome and Thanks
  - http://www.internet2.edu/workshops/npw/roster/neren.cfm
- Tutorial Agenda:
  - Network Performance Primer - Why Should We Care? (**30 Mins**)
  - Introduction to Measurement Tools (**20 Mins**)
  - Use of NTP for network measurements (**15 Mins**)
  - Use of the BWCTL Server and Client (**25 Mins**)
  - Use of the OWAMP Server and Client (**25 Mins**)
  - Use of the NDT Server and Client (**25 Mins**)
  - perfSONAR Topics (**30 Mins**)
  - Diagnostics vs Regular Monitoring (**20 Mins**)
  - Use Cases (**30 Mins**)
  - Exercises

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"

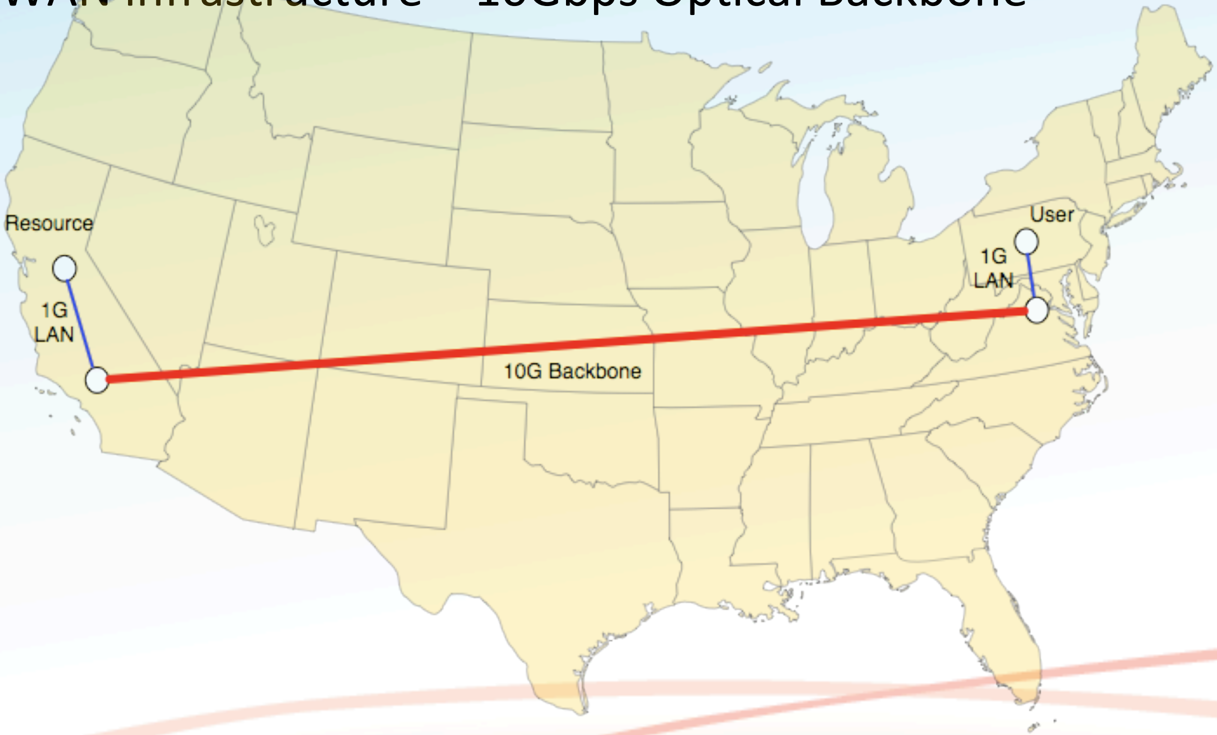
- How can your users effectively report problems?
  - And how can you learn to take them seriously…
- How can users and the local administrators effectively solve multi-domain problems?
  - Eliminate the 'who you know' network to finding resources
  - Automate things when applicable
- Components:
  - Tools to use
  - Questions to ask
  - Methodology to follow
  - How to ask for (and receive) help

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
- Automatic user response: “The Network is broken”
  - Is this justified behavior?
    - 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

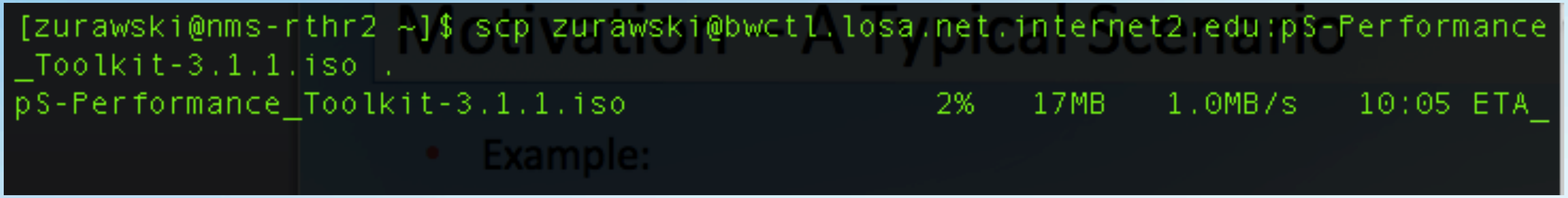


Motivation – A Typical Scenario

- 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

- 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

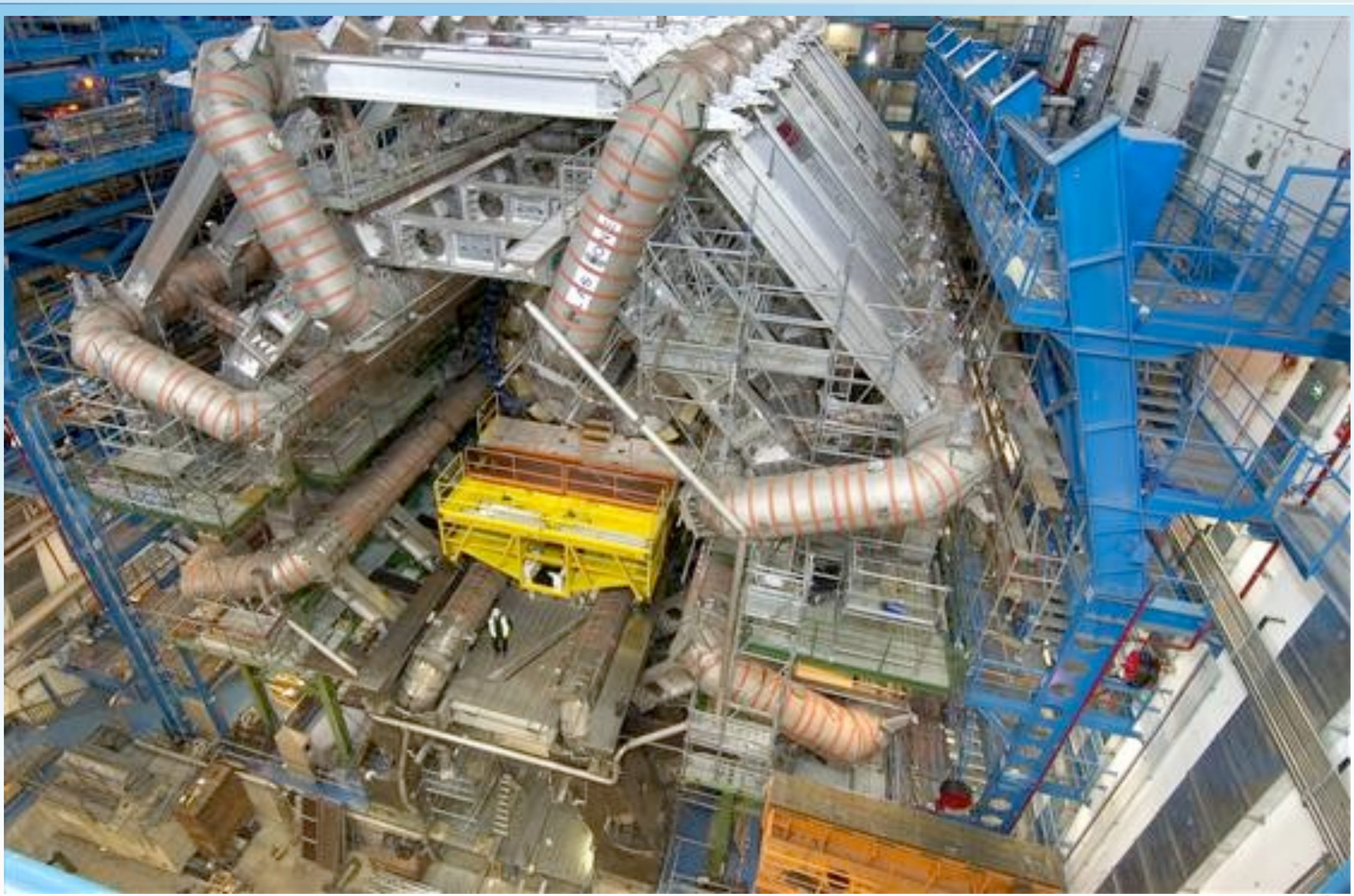
Why Worry About Network Performance?

- 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

“Big” Science

- A Few words on the LHC
  - 17 Mile Circumference “ring” in Switzerland/France
  - Collide opposing beams of particles (3.5 TeV each – 7TeV collision)
  - “Detectors” are present to gather data on the collision (ALICE, ATLAS, CMS, LHCb)
  - Data is stored at CERN (Tier0), and distributed world wide to other Tiers (1, 2, 3) for processing and analysis
    - Different types of data, Raw + several kinds of processed data to find areas of interest.
    - N.B. even the raw data doesn’t capture anything – the machine would produce ***1PB*** (!) of data, ***per second*** (!!), if it was unfiltered
    - Typical processed data set (2011) = ***10 – 100 TB***.
  - Tier1s receive and distribute data to Tier2s, Tier2s do the same for Tier3s
  - Each Tier contains storage and processing software/hardware.
    - Goal is to get the data to the lowest tier ***within 4 hours*** (!)

“Big” Science – ATLAS Detector



Why is Science Data Movement Different?

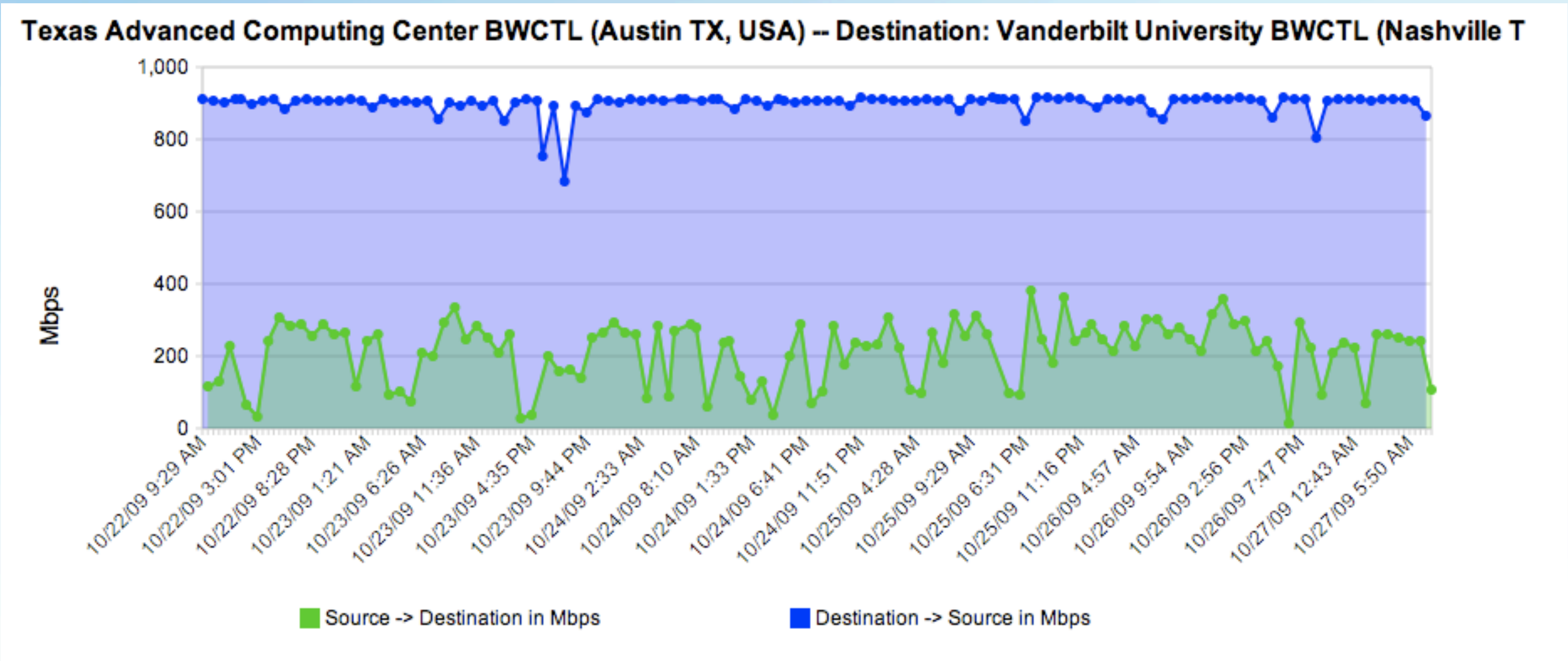
- Different Requirements
  - Campus network is not designed for large flows
    - ***Enterprise*** requirements
    - 100s of Mbits is common, any more is rare (or viewed as *strange*)
    - Firewalls
    - Network is designed to mitigate the risks since the common hardware (e.g. Desktops and Laptops) 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

- The above examples paint a broad picture: there is a problem, *somewhere*, that needs to be fixed
- What could be out there?
  - Architecture
  - Common Problems, e.g. "Soft Failures"
- Myths and Pitfalls
  - Getting trapped is easy
  - Following a bad lead is easy too

Identifying Common Network Problems

- Audience Question: Would you complain if you knew what you were getting was not correct?



- N.B. Actual performance between Vanderbilt University and TACC – Should be about 1Gbps in both directions.

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 severe connectivity, but makes the experience unpleasant

perfSONAR powered

INTERNET2

Architectural Considerations

- 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
  - Great performance on the LAN != Great performance on the WAN
- *You Get What you Pay For*
  - Cheap equipment will let you down
  - Network
    - Small Buffers, questionable performance (e.g. internal switching fabric can't keep up w/ LAN demand let alone WAN)
    - Lack of diagnostic tools (SNMP, etc.)
  - Storage
    - Disk throughput needs to be high enough to get everything on to the network
    - Plunking a load of disk into an incapable server is not great either
      - Bus performance
      - Network Card(s)

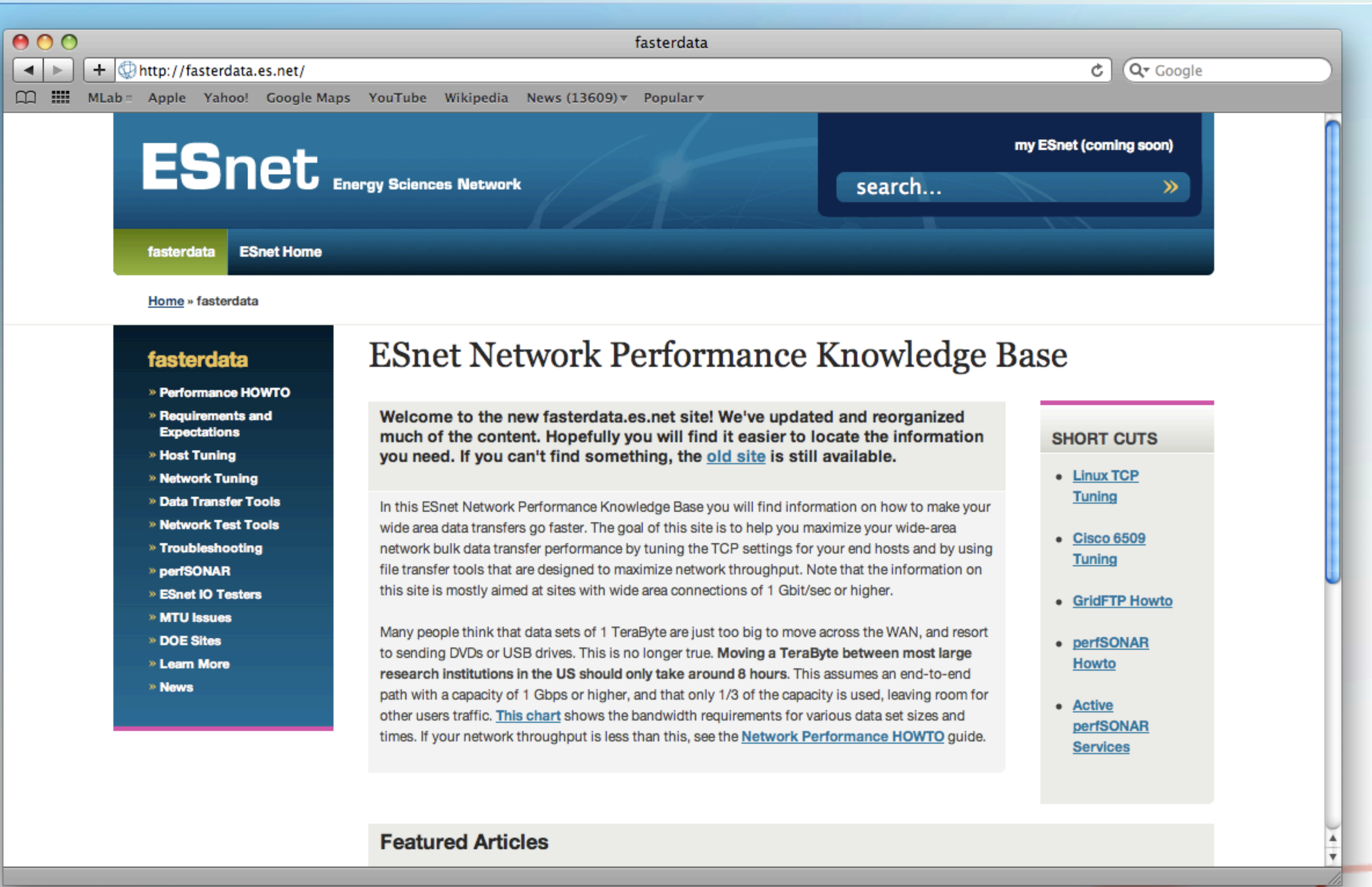
Architectural Considerations – cont.

- Firewalls
  - Designed to stop traffic
    - read this slowly a couple of times…
  - 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

Configuration

- Host Configuration
  - Tune your hosts (especially compute/storage!)
  - Changes to several parameters can yield 4 – 10X 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).

Host Configuration



Home » fasterdata

fasterdata

- » Performance HOWTO
- » Requirements and Expectations
- » Host Tuning
- » Network Tuning
- » Data Transfer Tools
- » Network Test Tools
- » Troubleshooting
- » perfSONAR
- » ESnet IO Testers
- » MTU Issues
- » DOE Sites
- » Learn More
- » News

ESnet Network Performance Knowledge Base

Welcome to the new fasterdata.es.net site! We've updated and reorganized much of the content. Hopefully you will find it easier to locate the information you need. If you can't find something, the old site is still available.

In this ESnet Network Performance Knowledge Base you will find information on how to make your wide area data transfers go faster. The goal of this site is to help you maximize your wide-area network bulk data transfer performance by tuning the TCP settings for your end hosts and by using file transfer tools that are designed to maximize network throughput. Note that the information on this site is mostly aimed at sites with wide area connections of 1 Gbit/sec or higher.

Many people think that data sets of 1 TeraByte are just too big to move across the WAN, and resort to sending DVDs or USB drives. This is no longer true. **Moving a TeraByte between most large research institutions in the US should only take around 8 hours.** This assumes an end-to-end path with a capacity of 1 Gbps or higher, and that only 1/3 of the capacity is used, leaving room for other users traffic. This chart shows the bandwidth requirements for various data set sizes and times. If your network throughput is less than this, see the Network Performance HOWTO guide.

SHORT CUTS

- Linux TCP Tuning
- Cisco 6509 Tuning
- GridFTP Howto
- perfSONAR Howto
- Active perfSONAR Services

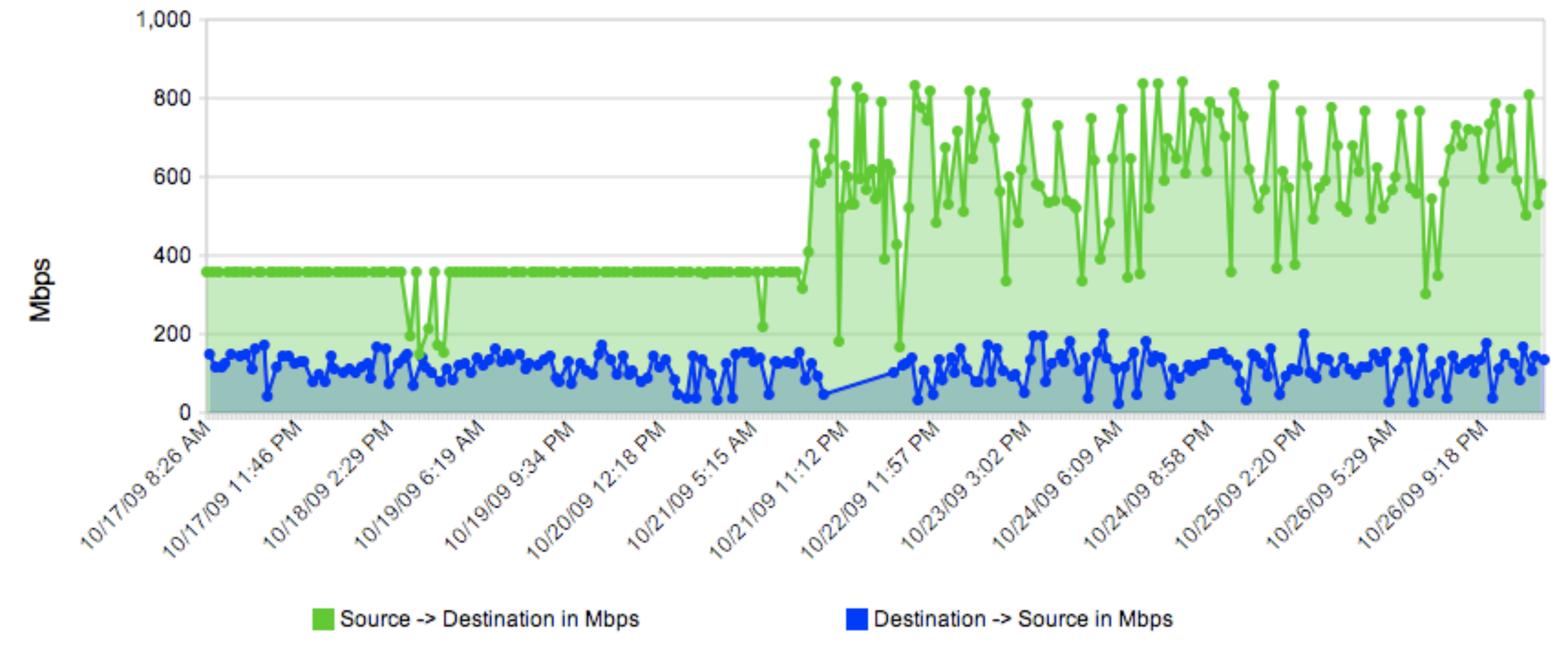
Featured Articles

perfSONAR powered

INTERNET2

Configuration – cont.

- Host Configuration – spot when the settings were tweaked…



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

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)

Soft Failures – cont.

- Dirty or Crimped Cables and Failing Optics/Interfaces
  - Throw off very low levels of loss – may not notice on a LAN, will notice on the WAN
  - Will be detected with passive tools (e.g. SNMP monitoring)
  - Question: Would you fix it if you knew it was broken?
- [Router] Process Switching
  - "Punt" traffic to a slow path
  - Duplicate traffic onto multiple paths
- Router Configuration (Buffers/Queues)
  - Need to be large enough to handle science flows
  - Routing table overflow (e.g. system crawls to a halt when memory is exhausted)

Myths and Pitfalls

- "My LAN performance is great, WAN is probably the same"
  - TCP recovers from loss/congestion quickly on the LAN (low RTT)
  - TCP will cut speed in half for every loss/discard on the WAN – will take a long time to recover for a large RTT/
  - Small levels of loss on the LAN (ex. 1/1000 packets) will go unnoticed, will be very noticeable on the WAN.
- "Ping is not showing loss/latency differences"
  - ICMP May be blocked/ignored by some sites
  - Routers process ICMP differently than other packets (e.g. may show phantom delay)
  - ICMP may hide some (not all) loss.
  - Will not show asymmetric routing delays (e.g. taking a different path on send vs receive)
- Our goal is to dispel these and others by educating the proper way to verify a network – we have lots of tools at our disposal but using these in the appropriate order is necessary too

Topics of Discussion in this Workshop

- Diagnosis Methodology
- Partial Path Decomposition
- Systematic Troubleshooting
- On Demand vs Regular Testing

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

Our Goals

- To spread the word that today' s networks really can, do, and will support demanding applications
  - Science
    - Physics
      - LHC, LIGO
    - Astronomy
      - LSST, SDSS, eVLBI
    - Biology and Climate
      - Genome Sequencing, Weather simulations, remote senors
  - Arts and Humanities
    - Distance learning, synchronized performance
  - Computational and Network Research
    - DYNES, GENI, MeasurementLab, etc.
- To increase the number of test points
  - Instrumenting the end to end path is key
  - Spread the knowledge and encourage adoption

Final Thoughts

- See a talk from the recent Joint Techs Conference:
  - http://www.internet2.edu/presentations/jt2010july/20100714-metzger-whatnext.pdf
- Take home points:
  - Close to $1 Billion USD spent on networking at all levels (Campus, Regional, Backbone) in the next 2 years due to ARRA Funding
  - Unprecedented access and capacity for many people
  - Ideal View:
    - Changes will be seamless
    - Completed on time
    - Bandwidth will solve all performance problems
  - Realistic View:
    - Network 'breaks' when it is touched (e.g. new equipment, configs)
    - Optimization will not be done in a global fashion (e.g. backbone fixes performance, but what about regional and campus?)
    - Bandwidth means nothing when you have a serious performance problem

Welcome & Performance Primer

August 9th 2011, OSG Site Admin Workshop

Jason Zurawski – Internet2 Research Liaison

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