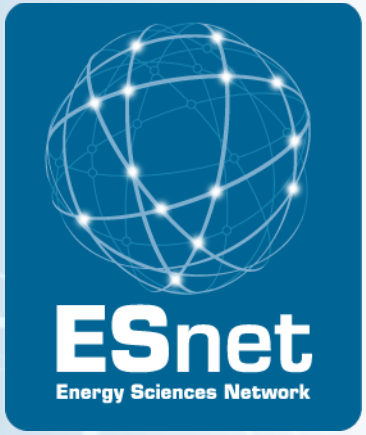


Science DMZs

Understanding their role in high-performance data transfers

Chris Tracy, Network Engineer

Eli Dart, Network Engineer

ESnet Engineering Group

Overview

Bulk Data Movement – a common task

Pieces of the puzzle

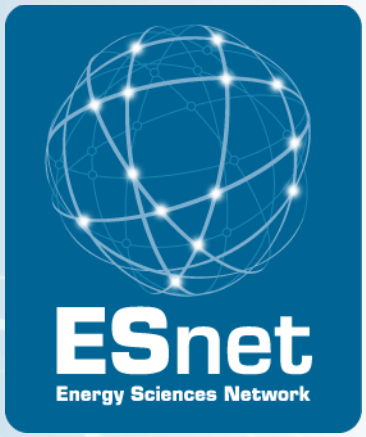
- Network Architecture
- Dedicated Hosts

Case Study

- International Fusion Research Collaboration

Discussion Topics

References



Bulk Data Movement

Bulk Data Movement

Common task at all data scales

Driven by collaboration, distributed resources

- Computing centers
- Facilities
- Major instruments (e.g. LHC)

Fundamental to the conduct of science (scientific productivity follows data locality)

Data sets of 200GB to 5TB are now common

Often a difficult task for various reasons

Time to Copy 1 Terabyte

These figures assume some headroom left for other users:

- 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)

Compare these speeds to:

- USB 2.0 portable disk
  - 60 MB/sec (480 Mbps) peak
  - 20 MB/sec (160 Mbps) reported online
  - 5-10 MB/sec reported by colleagues
  - 15-40 hours to load 1 Terabyte

Data Throughput – Transfer Times

Throughput required to move Y bytes in X time

File size	8 Hours	24 Hours	7 Days	30 Days
10PB	2,777.8 Gbps	925.9 Gbps	132.3 Gbps	30.9 Gbps
1PB	277.8 Gbps	92.6 Gbps	13.2 Gbps	3.1 Gbps
100TB	27.8 Gbps	9.3 Gbps	1.3 Gbps	308.6 Mbps
10TB	2.8 Gbps	925.9 Mbps	132.3 Mbps	30.9 Mbps
1TB	277.8 Mbps	92.6 Mbps	13.2 Mbps	3.1 Mbps
100GB	27.8 Mbps	9.3 Mbps	1.3 Mbps	308.6 Kbps
10GB	2.8 Mbps	925.9 Kbps	132.3 Kbps	30.9 Kbps
1GB	277.8 Kbps	92.6 Kbps	13.2 Kbps	3.1 Kbps
100MB	27.8 Kbps	9.3 Kbps	1.3 Kbps	0.3 Kbps
Time to transfer				

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

Data Transfer Is A Tractable Problem

In many cases, data sets are less than 10TB or even 1TB

- Using the previous table, transferring 1TB per day requires:
  - 92.6 Mbps for 24 hours
  - 277.8 Mbps for 8 hours
- A well-configured infrastructure can do this quickly and easily
  - We can transfer 10GB files from LBL to the Data Transfer Nodes at NERSC and ORNL in less than a minute
  - One recent measurements was 1.6 Gbps disk-to-disk using COTS hardware and GridFTP transfer tools
- A single 1 Gbps data transfer host should be able to meet the needs of a great many scientists

Many Users Have Difficulty

Bulk Data Movement is Tractable

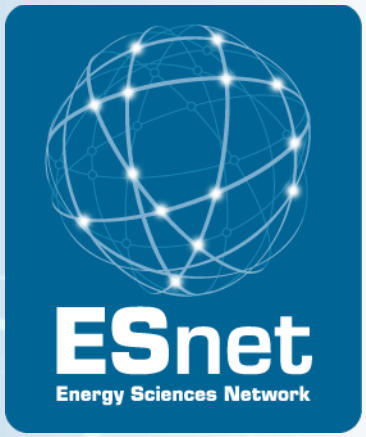
- OK, then why is it so hard?

Scientists are asked to integrate their own data transfer infrastructure

- They don't have the funding for systems support
- They don't have funding for infrastructure
- "Why would I buy a new switch? $50k is another postdoc!"

No guarantee that the people "at the other end" have a well-configured infrastructure – why build my own?

- If the barriers to use of the network are too high, it doesn't matter if ESnet or CSTnet builds cutting edge network infrastructures



Pieces of the Puzzle

Network Architecture

Network Architecture

Most LANs are not purpose-built for science traffic

- Often carry many types of traffic:
  - desktop machines, laptops, wireless
  - VoIP
  - HVAC control systems
  - Financial systems, HR
  - *some science data coming from some place*

Bulk data transfer traffic is typically very different than enterprise traffic

Architecture – Enterprise Networks

Provide access to commercial Internet

Business continuity

- Risk management
  - Personally Identifiable Information (PII), Financial information
  - Embarrassment due to security incidents
- Relatively low bandwidth unless there are a lot of users

Unsophisticated user base (from a computer security perspective)

- Lots of desktop boxes
- Laptops, visitors (hosts that visit other networks)

Need network-level policy controls to mitigate risk

- Firewalls, Management of file sharing traffic (e.g. BitTorrent), etc

Architecture – Science Networks

High bandwidth Requirement (10s of Gbps)

- Not just in connection speed, but in delivered performance to computational, visualization and storage resources
- Different tool set and traffic profile
  - This isn't for desktop boxes
  - Built for special-purpose hosts, e.g. data movers

Relatively sophisticated users

Sensitive to perturbations caused by security devices

- Numerous cases of firewalls causing problems
- Often difficult to diagnose
- Router filters can often provide equivalent security without the performance impact

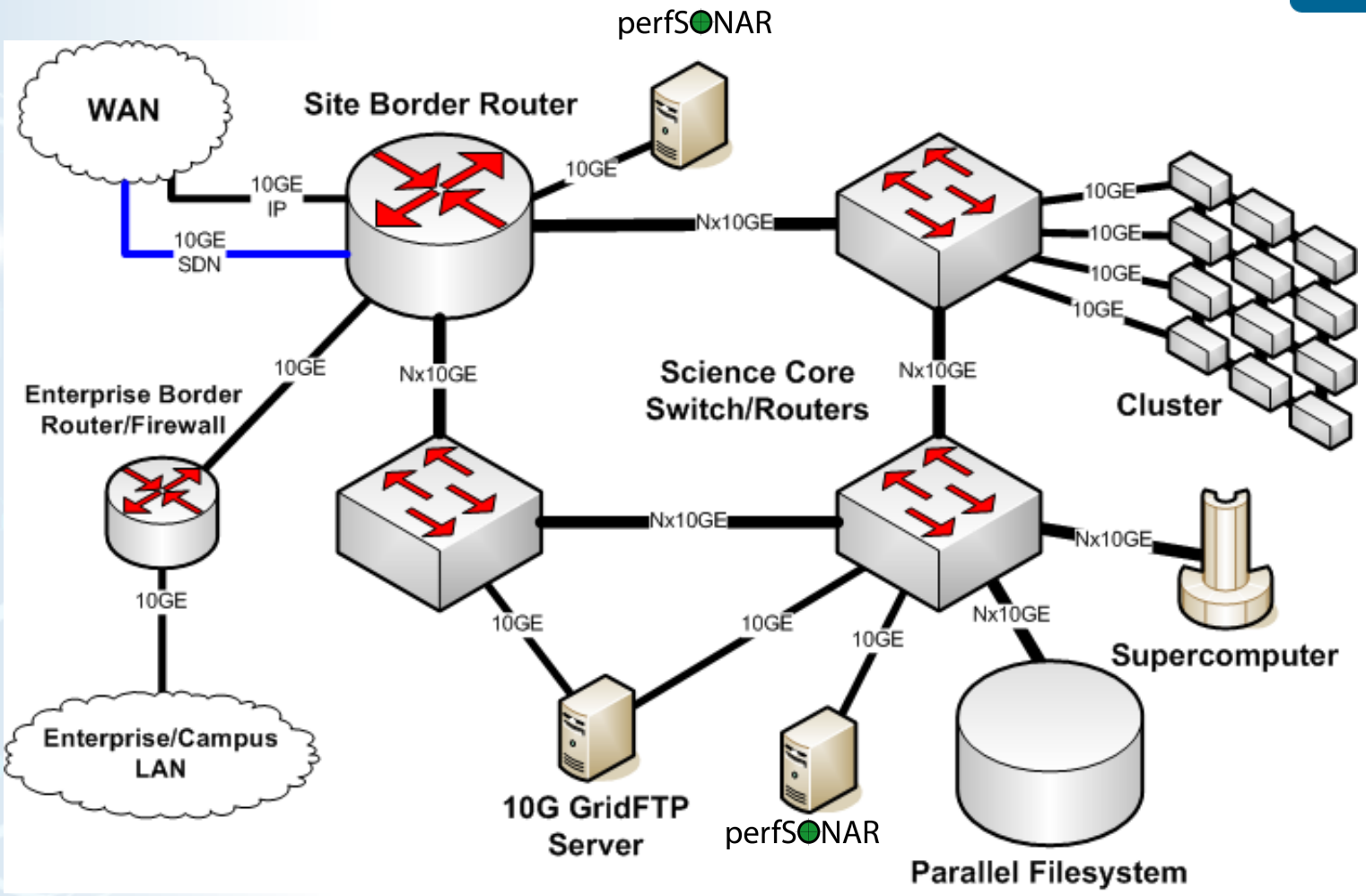
Enterprise vs. Science Networks

Science and Enterprise network requirements are in conflict

One possible solution:

- Build a science network for the science and attach the enterprise network to the science network
- Put the Enterprise security perimeter at the edge of the enterprise network, not at the site border
- Science resources are not burdened by Enterprise firewall configuration

Separate Enterprise and Science Networks



Network Pitfalls – Soft Failures

"Soft Failures"

- network problems that don't result in total loss of connectivity
- network (or a particular router or link) is up, but does not perform well
- problem often goes unnoticed until someone tries to use the WAN for high throughput

Examples

- process switching ("punting")
- dirty fiber
- failing optics
- misconfigured (or hardware lacking) buffers/queues
- routing table overflow in Cisco devices (causes punting)

Network Architecture Summary

Build a network to support bulk data transfers with data transfer in mind

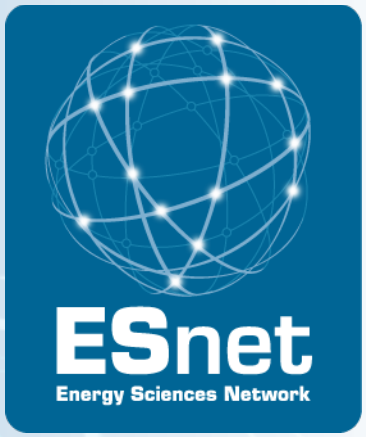
- Don't just plug it in anyplace
- Avoid traversing Enterprise infrastructure if you can
- Connect your bulk transfer resources as close to the perimeter as you can

Configure routers and switches for adequate buffering

- Watch drop counters (e.g. "sho int sum" or "sho int queue" in IOS)
- Watch error counters

If you have to, collocate your data server near the border router

- On-site transfers to your server in another building will usually be high performance due to low latency
- WAN transfers bypass the Enterprise infrastructure



Pieces of the Puzzle

Dedicated Hosts

Data Transfer Nodes

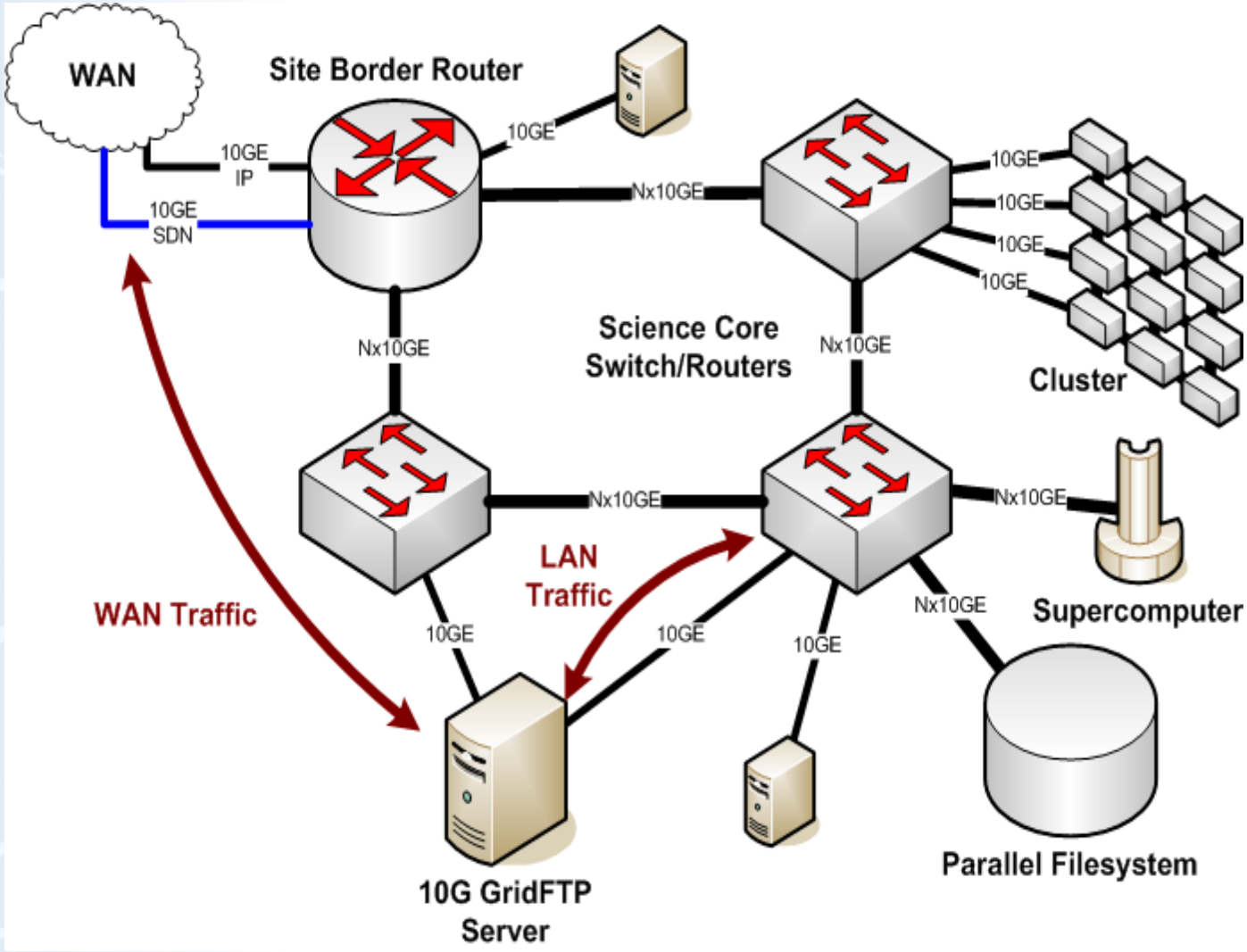
Reasons for dedicated hosts

- One thing to test and tune
- One place for large WAN flows to go
  - easier to give one host a special configuration than to do this for all workstations
- One set of firewall exceptions

If you can, use different network connections for LAN and WAN flows

- LAN flows can easily saturate network interfaces, especially 1Gbps interfaces
- LAN flows recover quickly, unlike WAN flows

Internal / External Traffic Separation



Data Transfer Nodes

- Large physics experiments (BaBar, LHC, RHIC, Tevatron, etc) already do this
- Recent success story – Fusion Research
  - Two systems – one at GA and one at EAST in China
  - Data transfers now keep up with instrument duty cycle
  - More information about this example in the case study
- Additional success stories – Data Transfer Nodes
  - DOE Supercomputing centers at Argonne, Oak Ridge, NERSC
  - Dedicated hosts with access to shared global filesystems

Where To Deploy Dedicated Systems?

- Clearly dependent on network architecture – Know Your Network
- However, we have seen significant performance benefits when data transfer systems are moved near the site perimeter
- A DMZ network holds the external-facing servers that provide service to the Internet (e.g. DNS, Mail, Web)
- A “Science DMZ” could attach high-performance data servers to the site border router
  - This can be done with dark fiber if you’ve got the fiber – no need to move the machines to a different building, etc.
  - No need to drag large wide area data flows through the site network or the site firewall

Dedicated Host Summary

Operating System

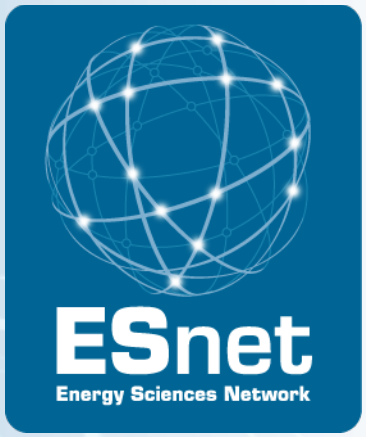
- Use a newer OS that supports TCP buffer autotuning and congestion recovery
- Increase the maximum TCP autotuning buffer size
- Use a modern congestion control algorithm

Network Connection

- Logically attach dedicated hosts as close to the site perimeter as possible

Data Transfer Tools

- Use tools which exploit parallelism (e.g., GridFTP)
- Don’t use tools for WAN transfer that assume a LAN (e.g., SCP/ SFTP)



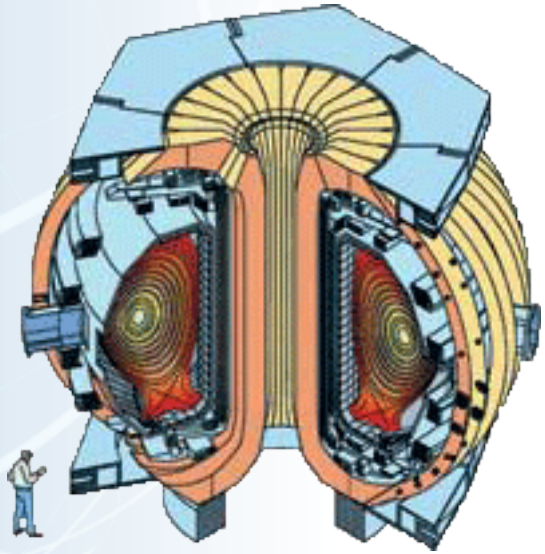
Case Study

International Fusion Research Collaboration

GA / DIII-D Collaboration with IPP / EAST

Collaboration between Chinese and US scientists

- EAST Tokamak at Institute of Plasma Physics in China
- Data generated by EAST at IPP, analyzed by DIII-D collaborators at General Atomics (GA) in San Diego, CA



Source: https://fusion.gat.com/global/DIII-D

Source: http://en.wikipedia.org/wiki/EAST

GA / DIII-D Collaboration with IPP / EAST

Semi-real-time data movement requirement

- Goal was to have data transfers keep up with instrument duty cycle
- Minimum data rate: 50MB in 2 minutes
- Transfers must be repeatable

Several difficulties encountered

- Firewall issues
- Packet loss

Solution: Dedicated Data Transfer Hosts

Both GA and IPP deployed a dedicated Linux host

- Properly tuned TCP configuration for WAN
- GridFTP for parallelized data transfer

WAN transfers happen between these dedicated hosts

- Data copied from EAST to local WAN transfer host
- Data transferred to WAN host at remote site (GA)
- Data copied to analysis machines at GA

Scientists' data transfer requirements are met

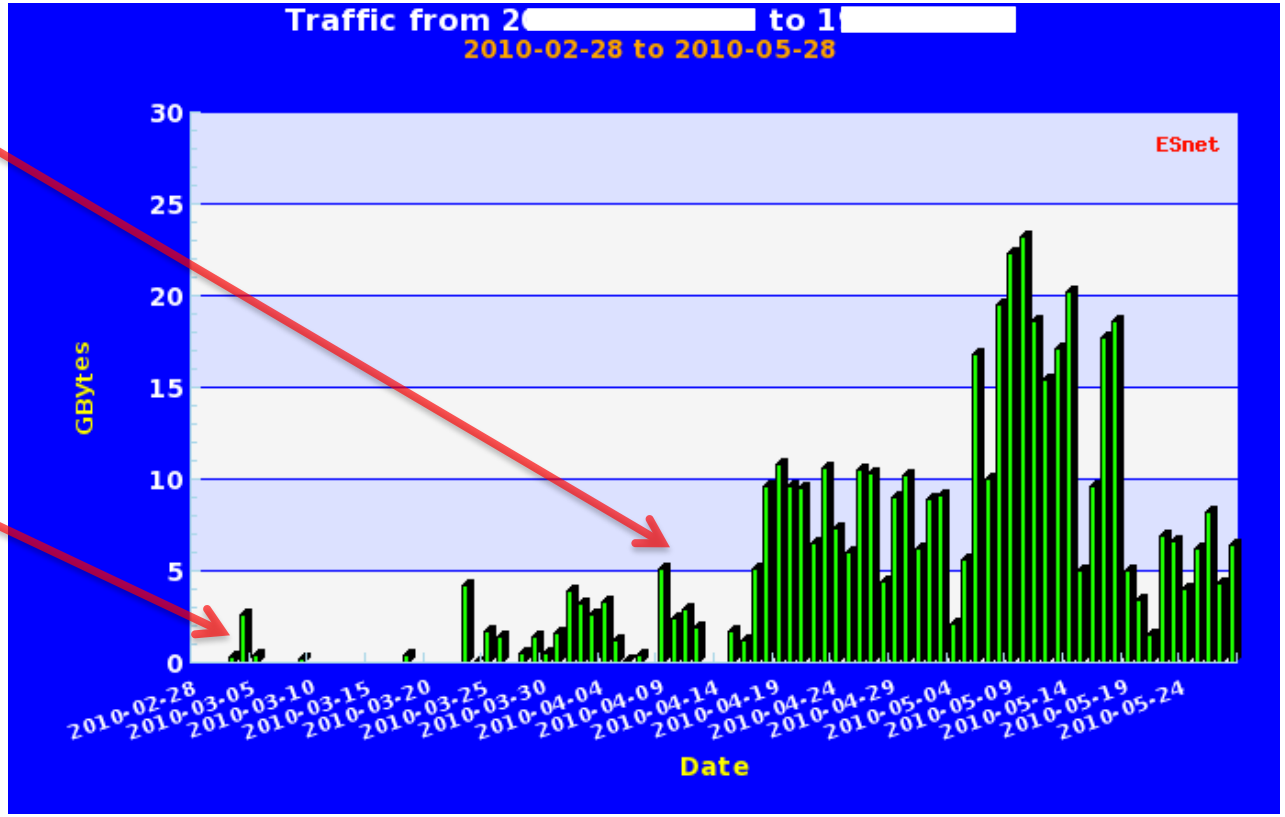
- Collaboration moves successfully forward

Transfer Statistics: EAST (at IPP) to GA

Path: CSTNET —> GLORIAD —> ESnet
Round Trip Time ~330ms, 19 hops

Went into production between March and May

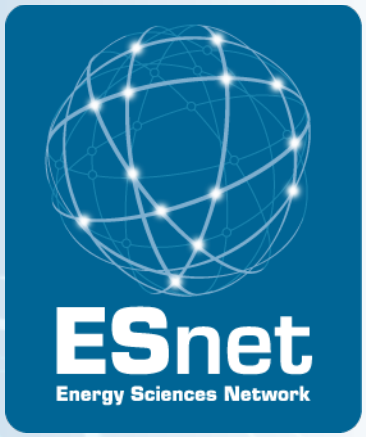
Dedicated transfer hosts deployed and initial test results appear promising



Source: ESnet NetFlow data, dedicated WAN hosts, 2010-02-28 to 2010-05-28

We're All In This Together

- It is our collective job to support science
- Science is increasingly data-intensive
- Scope of collaboration is regional to global
- Therefore, science requires data movement, now and into the future
- Our customers cannot succeed unless we work together
  - Well-configured end systems and high performance networks are both necessary
  - Neither is a solution in itself
- ESnet is willing to help with design, troubleshooting, etc., both for sites and for scientists



Discussion Topics

Discussion Topics

- Are there other obvious places to put dedicated systems?
- Pick the low-hanging fruit first!
- Science DMZs – issues for deployment
  - Firewalls (do we need have to have firewalls if there are no windows clients? Windows “no-fly” zones…)
  - If the site security policy treats the boxes on the Science DMZ as external, does that help?
  - Funding silos, territory/influence concerns, etc.

References

ESnet Network Performance Knowledge Base:

http://fasterdata.es.net

More in-depth tutorials / presentations on these topics at:

http://fasterdata.es.net/tutorials.html

Fusion research related:

http://en.wikipedia.org/wiki/Tokamak

http://en.wikipedia.org/wiki/EAST

https://fusion.gat.com/global/DIII-D

http://en.wikipedia.org/wiki/ITER

Questions?

Thanks!