ESnet – Science Perspective and Network Services

Eli Dart, Network Engineer

ESnet Network Engineering Group

NERSC Oakland Scientific Facility

Oakland, CA

June 21, 2011

Outline

Network Services

- Routed network
- Science Data Network (SDN)
- Network performance
- ECS

Science Requirements

- Selected highlights → new usage paradigms
- User taxonomy

End to end performance

- Components of high performance data transfer
- Common problems and how to fix them
- http://fasterdata.es.net/

Network Services – Routed IP

Best-effort IP – the “standard” network service

Scavenger service – less-than-best-effort

- Blast away without fear of hurting anything
- Mostly used inside ESnet (e.g. non-conforming traffic)

Lots of effort expended to make the IP network loss-free

- Deep queues for burst tolerance
- Test and measurement
- ***Verification of performance is critical*** *(see performance section)*

Network Services – SDN

Science Data Network – SDN

- Dynamic virtual circuits
- Carried over separate infrastructure (separate circuits, and in most cases separate routers)
- Traffic engineering, bandwidth guarantees

What is a virtual circuit?

- Path through the network with customized behavior
  - Bandwidth guarantee
  - Explicit path
  - Edge to edge integrity (no injection of traffic)
- Layer2 or layer3

SDN circuits – example uses

Deliver traffic to a particular destination in the network

- Match on layer3 information (src/dst IP address, port, etc.)
- Deliver traffic to a particular ESnet router
- Traffic is then routed normally
- Example – deliver science data to a particular peering exchange to take advantage of high-bandwidth cloud provider infrastructure

Provide dedicated VLAN service between sites

- Match on layer2 information (VLAN tag)
- Hosts or routers at the ends are in a common broadcast domain
- Examples include remote filesystem mount, cluster to cluster transfers, point to point link between sites (use controlled by BGP between sites), etc.

Explicit paths for diversity

- Used extensively for LHCOPN
- Multiple paths between sites traverse different physical infrastructure (no single outage severs all connectivity)

Network Services – Performance

perfSONAR – deployed at all 10G ESnet locations, most lower-speed locations

- E.g. nersc-pt1.es.net, chic-pt1.es.net, sunn-pt1.es.net, etc.
- https://stats.es.net/perfSONAR/directorySearch.html (type ‘ESnet’ in search box at upper left, or select ESnet from top left and click Update)
- All *-pt1 hosts accept bwctl/iperf tests from R&E hosts worldwide

Disk performance testers (DiskPT hosts)

- Well-configured Linux machines with fast disk running anonymous GridFTP servers
- Download data files to test disk-to-disk performance
- http://fasterdata.es.net/fasterdata/esnet-io-testers/

Consulting / troubleshooting – trouble@es.net

- ESnet routinely helps with network architecture, host configuration, and troubleshooting
- Use of perfSONAR and other tools to isolate problems

ECS - Video, audio and web collaboration services

For organizations and projects funded by the DOE Office of Science

- Information and registration: **http://www.ecs.es.net/**

H323 Reservationless Videoconference Service

- Supports HiDef and StdDef room and desktop endpoints
- Endpoint configuration allows easy conference IDs
  - Select a HD or SD prefix, add any combination of #s, all sites dial same conference ID. Example: 7563772 for 75Nersc
- Supports phone participants
- Streaming available for all meetings

ECS - Video, audio and web collaboration services

Audio/Web Collaboration Service

- Supports toll free domestic and international phone participants
  - 96 participants by default, more by special arrangement
- Chairperson conference activation required
  - Has control of meeting to dial out, record meeting, mute/unmute audio
- Web collaboration allows participants to view chairperson or co-presenter’s desktop or application
  - Allows application sharing
- Available on-demand or by email invitations
- Phone and web can be used separately or simultaneously

Science Requirements

Sources

- Requirements workshops
  - 2 workshops per year, each program office every 3 years
  - http://www.es.net/about/science-requirements/network-requirements-workshops/
  - http://www.es.net/about/science-requirements/reports/
- Program and other directives
- Network observation and operational experience

Science mission drives ESnet

- ESnet is the high-performance networking facility of the DOE Office of Science
- We devote significant effort to understanding the science so as to better serve the scientists

Science Requirements – coming changes

Many collaborations and disciplines are re-thinking the ways in which they use scientific infrastructure

- Dramatic changes in costs for some components
- Significant increase in data intensity across many scientific disciplines
- New paradigms are increasing the need for network services

What does this mean for infrastructure providers?

- Data transfer must be simple to use, reliable, and consistent
  - Usability, reliability, and consistency trump performance
  - Dedicated infrastructure (e.g. NERSC DTNs) is very important
- Multi-site workflows will become increasingly common

Light and Neutron Sources

ALS at LBL, APS at ANL, LCLS at SLAC, NSLS at BNL, SNS at ORNL, etc.

Large number of beamlines, instruments

- Hundreds to thousands of scientists per facility
- Academia, Government, Industry

Data rates have historically been small

- Hand-carry of data on physical media has been the norm for a very long time: CDs → DVDs → USB drives
- Scientists typically do not use the network for data transfer today

Near future: much higher data rates/volumes

- Next round of instrument upgrades will increase data volumes by 10x or even 100x, e.g. from 700GB/day to 70TB/day
- *Network-based data transport is going to be necessary for thousands of scientists that will be doing this for the first time in their careers*

Light and Neutron Sources

New science architectures coming

- Experiment automation leads to the need for near-real-time health checks
  - Stream sample experiment output to remote location for verification of experiment setup
  - Significant efficiencies of automation are driving this
- Multi-site dependencies (e.g. need for analysis at supercomputer centers)
  - Need a general model for streaming from detectors to supercomputer centers
  - Supercomputer centers often say that allocations change from year to year, therefore significant effort to support one particular scientist may not be wise resource allocation
  - However, many light source users will need to stream data to supercomputer centers – generalized support for this use model will result in significantly increased scientific productivity

Light and Neutron Sources

Some of these data increases have already taken place

Dedicated data transfer hardware and perfSONAR have been used to fix performance problems

- Networks must be loss free
- Networks must be monitored to ensure that they stay clean

These solutions will need to be generalized

- Science DMZs and/or Data Transfer Nodes (DTNs) for light sources
- Assist users with figuring out the “other end” (e.g. suggestions for common architectures such as DTN or Science DMZ)
- Requiring that every collaboration implement their own solution (as many light sources do currently) will result in tens of one-offs over the next few years
  - Difficult to troubleshoot
  - High support load for facility, system and network support staff
  - Therefore, a systematic approach must be developed for large-scale science infrastructure

Common Themes – Science Requirements

New science processes such as remote instrument control, experiment health monitoring, etc will place new demands on networks

- Multi-site near-real-time or real-time network interaction
- Need expressed by multiple science communities (light sources, biology, HPC users, etc)
- Many of these communities are not network experts, and will need help from networking organizations in order to progress

Increasing data intensity of science across many disciplines

- Many collaborations that have historically not used the network for data transport must begin soon – ‘sneakernet’ will no longer be practical
- Many collaborations that have gotten by with using SCP/rsync/etc for WAN transfers will no longer be able to do so – must change to GridFTP or something similar to increase performance
- Collaborations that require >10Gbps connectivity today will need >100Gbps connectivity by 2015 – 10x increase every 4 years

Rough User Grouping By Data Set Size

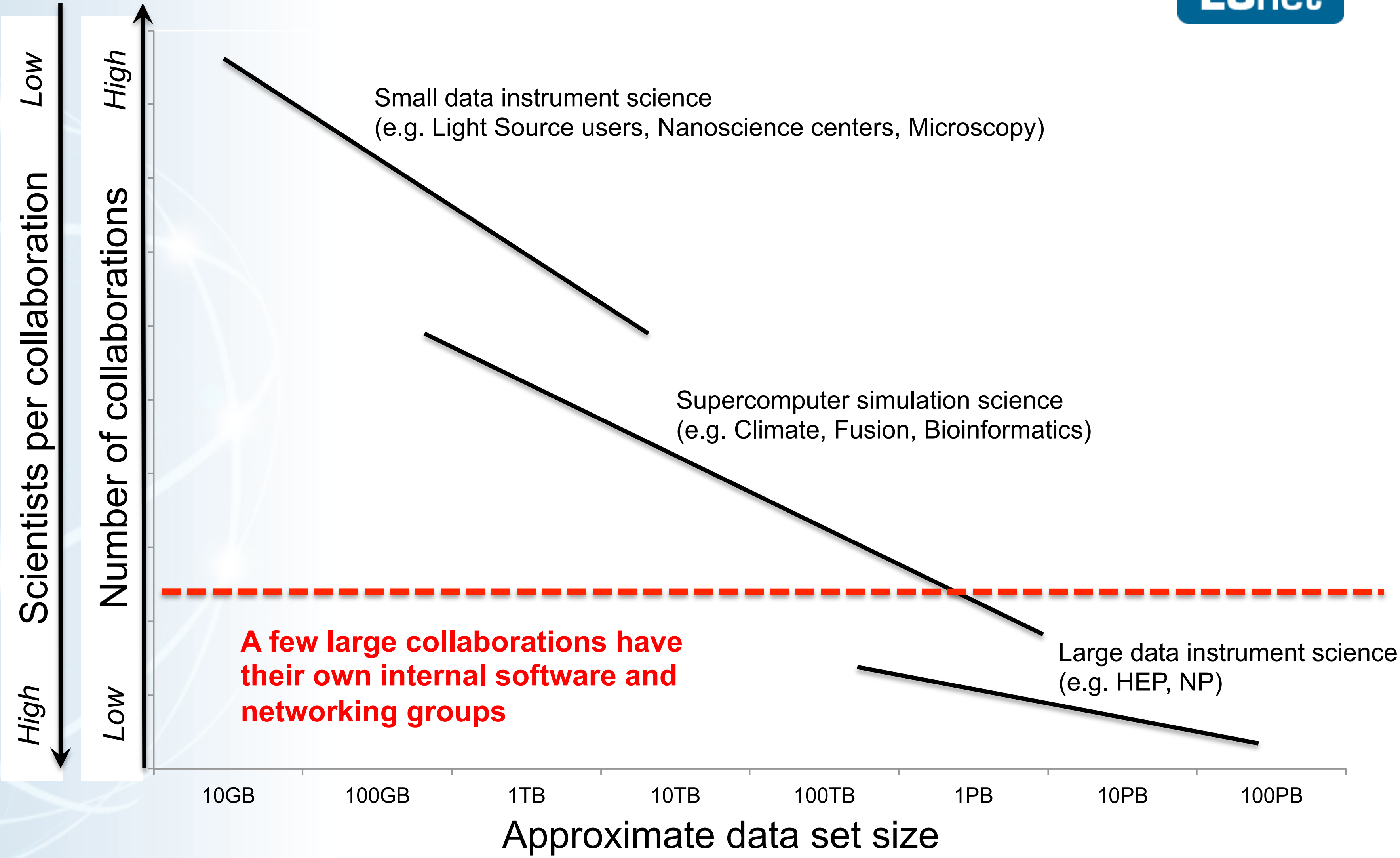

Chart Discussion (1)

The chart is a crude generalization – it is not meant to describe specific collaborations, but to illustrate some common aspects of many collaborations

Small data instrument science

- Light sources, microscopy, nanoscience centers, etc.
- Typically small number of scientists per collaboration, many many collaborations
- Individual collaborations typically rely on site support and grad students
- This group typically has difficulty moving data via the network

Chart Discussion (2)

Supercomputer simulation science

- Climate, fusion, bioinformatics, computational astrophysics, etc.
- Larger collaborations, often multi-site
- Reliant on supercomputer center staff for help with network issues, or on grad students
- This group typically has difficulty transferring data via the network
  - Data Transfer Nodes are starting to help
  - Many users still want to use HPSS directly (often performs poorly)

Large data instrument science (HEP, NP)

- Very large collaborations – multi-institution, multi-nation-state
- Collaborations have their own software and networking shops
- Typically able to use the network well, in some cases expert

End to end performance

High-performance data transfer components

- Local storage
- End systems and application software (e.g. GridFTP)
- Each network in the path
- ***All of these must function correctly!***

Data transfers are difficult and complex – therefore:

- Systems must be maintained in working order
  - Configured properly, tested regularly
  - NERSC is a leader here – DTNs are a perfect example
- Proper tools must be available
  - Unfortunately, each discipline seems to have their own pet tool
  - Therefore, multiple tools must be maintained (GridFTP, bbcp, etc.)
- Documentation is critical – most users are not experts

Common problems and their solutions

Many users are relying on SCP/SFTP or rsync over SSH

- These are essentially guaranteed to perform poorly
- http://fasterdata.es.net/fasterdata/say-no-to-scp/
- HPN-SSH patches help, but they are no longer under development (and most users don't have the clout to get HPN-SSH installed at both ends anyway)
- Solution: use a different tool (e.g. GridFTP via Globus Online, if there is a GridFTP server available at the remote end)
- If there is a need to get a GridFTP server deployed somewhere, ESnet can help advocate

Common problems and their solutions

General poor performance

- Ensure that the hosts involved have been set up correctly
- http://fasterdata.es.net/fasterdata/host-tuning/
- Note that TCP autotuning is available on almost all systems
- On a clean high-speed network, modern inexpensive hosts should be able to easily get 600Mbps of throughput (disk subsystem permitting)
- Multiple gigabits can be achieved with some effort (e.g. with DiskPT hosts as a model)
- http://fasterdata.es.net/fasterdata/esnet-io-testers/
- If the hosts are set up correctly and there are still problems, please involve ESnet (there may be network issues somewhere along the path)

Common problems and their solutions

Firewall issues

- Tools or protocols are blocked
- If the user is having trouble with their local security people, ESnet may be able to help (we often have other contacts at sites)

Shipping disks or other process issues

- In general, we are entering an era where dedicated data transfer hosts will be significant enablers for science
  - DTNs at beamlines or at other facilities
  - Globus Online endpoints
- Some collaborations want to ship disks because that is what they have done in the past
  - Either this was the only thing that worked before, or they don't want to fight the local battle, or something else
  - Strategically, this is not where we need to be since there are significant productivity gains to be had through use of networks
- ESnet will help where we can – please involve us

fasterdata.es.net

Network performance knowledge base

- Lots of documentation
- Configuration designed for cut and paste (e.g. into /etc/sysctl.conf for host tuning)
- More information added all the time

Suggestions for fasterdata are welcome

- This includes HOWTO-style documentation that NERSC staff or users would find helpful
- We would like to link to documentation at other sites (e.g. NERSC DTN documentation)
- Contributions are welcome

http://fasterdata.es.net/

Questions?

Thanks!

Questions?

Thanks!