Moving Data Over Networks
Network-Based Data Transfer at NERSC

Eli Dart, Network Engineer
ESnet Science Engagement
Lawrence Berkeley National Laboratory

NERSC Users Group Training
Berkeley, CA
February 24, 2016
Outline

• Context
• Science DMZ overview
• Data Transfer Nodes
• Handoff to Shreyas Cholia
Science Networks for Science

- The global Research & Education (R&E) network ecosystem is comprised of hundreds of international, national, regional and local-scale networks – each independently owned and operated.

- These networks are part of and connected to the Internet, but are engineered specifically for high-performance scientific applications.
Effective High Performance Data Transfer

- Data transfers between resources connected to R&E networks can do much better than data transfers which use the commodity Internet
  - Terabytes are no problem
  - Petabytes are feasible
- Just need to make sure we do a couple of things
  - Long distance portions work well in general
  - Large-scale computing centers work well in general
  - Local configuration is really important
- NERSC has high-performance data resources
  - Fast networks
  - Fast systems and filesystems
- This talk will describe what you can do to interface with NERSC effectively
Motivation

• Networks are an essential part of data-intensive science
  – Connect data sources to data analysis
  – Connect collaborators to each other
  – Enable machine-consumable interfaces to data and analysis resources (e.g. portals), automation, scale

• Performance is critical
  – Exponential data growth
  – Constant human factors
  – Data movement and data analysis must keep up

• Effective use of wide area (long-haul) networks by scientists has historically been difficult

• Some of this is for your system administrator
  – Point your sysadmin to http://fasterdata.es.net/ for more info
  – Feel free to follow up with me later – engage@es.net
The Central Role of the Network

• The very structure of modern science assumes science networks exist: high performance, feature rich, global scope

• What is “The Network” anyway?
  – “The Network” is the set of devices and applications involved in the use of a remote resource
    • This is not about supercomputer interconnects
    • This is about data flow from experiment to analysis, between facilities, etc.
  – User interfaces for “The Network” – portal, data transfer tool, workflow engine
  – Therefore, servers and applications must also be considered

• What is important? Ordered list:
  1. Correctness
  2. Consistency
  3. Performance
TCP – Ubiquitous and Fragile

• Networks provide connectivity between applications running on hosts
  – From an application’s perspective, the interface to “the other end” is a socket
  – Host operating system kernel provides socket interface, kernel implements TCP where the application can’t see
  – Communication is between applications – mostly over TCP

• TCP – the fragile workhorse
  – TCP is (for very good reasons) timid – packet loss is interpreted as congestion
  – Like it or not, TCP is used for the vast majority of data transfer applications (more than 95% of ESnet traffic is TCP)
  – Packet loss in conjunction with latency is a performance killer
A small amount of packet loss makes a huge difference in TCP performance.

Throughput vs. Increasing Latency with .0046% Packet Loss

With loss, high performance beyond metro distances is essentially impossible.
Working With TCP In Practice

- Far easier to support TCP than to fix TCP
  - People have been trying to fix TCP for years – limited success
  - Like it or not we’re stuck with TCP in the general case
- Pragmatically speaking, we must accommodate TCP
  - Sufficient bandwidth to avoid congestion
  - Zero packet loss
  - Verifiable infrastructure
    - Networks are complex
    - Must be able to locate problems quickly
    - Small footprint is a huge win – small number of devices so that problem isolation is tractable
Putting A Solution Together

• Effective support for TCP-based data transfer
  – Design for correct, consistent, high-performance operation
  – Design for ease of troubleshooting

• Easy adoption is critical
  – Large laboratories and universities have extensive IT deployments
  – Drastic change is prohibitively difficult

• Cybersecurity – defensible without compromising performance

• Borrow ideas from traditional network security
  – Traditional DMZ
    • Separate enclave at network perimeter ("Demilitarized Zone")
    • Specific location for external-facing services
    • Clean separation from internal network
  – Do the same thing for science – Science DMZ
The Science DMZ Design Pattern

- **Data Transfer Node**
  - High performance
  - Configured specifically for data transfer
  - Proper tools

- **Network Architecture**
  - Dedicated network location for high-speed data resources
  - Appropriate security
  - Easy to deploy - no need to redesign the whole network

- **Performance Testing & Measurement**
  - Enables fault isolation
  - Verify correct operation
  - Widely deployed in ESnet and other networks, as well as sites and facilities

© 2015, Energy Sciences Network
Abstract or Prototype Deployment

- (This section is for your system administrator – send them to me, use engage@es.net)
- Add-on to existing network infrastructure
  - All that is required is a port on the border router
  - Small footprint, pre-production commitment
- Easy to experiment with components and technologies
  - DTN prototyping
  - perfSONAR testing
- Limited scope makes security policy exceptions easy
  - Only allow traffic from partners
  - Add-on to production infrastructure – lower risk than rebuilding existing infrastructure
Science DMZ Design Pattern (Abstract)

- **WAN**
  - perfSONAR

- **Border Router**
  - 10GE
  - Clean, High-bandwidth WAN path
  - Site / Campus access to Science DMZ resources
  - 10GE

- **Science DMZ Switch/Router**
  - Per-service security policy control points
  - 10GE

- **Enterprise Border Router/Firewall**
  - 10GE

- **Site / Campus LAN**

- **High performance Data Transfer Node**
  - with high-speed storage
  - perfSONAR

- **ESnet Science Engagement**
  - engage@es.net
  - 2/24/17

© 2015, Energy Sciences Network
Local And Wide Area Data Flows

Border Router

WAN

10GE

Clean, High-bandwidth WAN path

High performance Data Transfer Node with high-speed storage

Site / Campus access to Science DMZ resources

Science DMZ Switch/Router

10GE

Site / Campus LAN

Enterprise Border Router/Firewall

10GE

Per-service security policy control points

High Latency WAN Path

Low Latency LAN Path
Modular Architecture – Multiple Science DMZs

Border Router

Enterprise Border Router/Firewall

WAN

10G

10GE

10GE

Science DMZ Switch/Routers

Per-project security policy

Project A DTN (building A)

Facility B DTN (building B)

Cluster DTN (building C)

Cluster (building C)

Site / Campus LAN

Dark Fiber

Dark Fiber

Dark Fiber

perfSONAR

perfSONAR

perfSONAR

perfSONAR

15 – ESnet Science Engagement (engage@es.net) - 2/24/17

© 2015, Energy Sciences Network
Supercomputer Center Deployment

• High-performance networking is assumed in this environment
  – Data flows between systems, between systems and storage, wide area, etc.
  – Global filesystem often ties resources together
    • Portions of this may not run over Ethernet (e.g. IB)
    • Implications for Data Transfer Nodes

• “Science DMZ” may not look like a discrete entity here
  – By the time you get through interconnecting all the resources, you end up with most of the network in the Science DMZ
  – This is as it should be – the point is appropriate deployment of tools, configuration, policy control, etc.

• Office networks can look like an afterthought, but they aren’t
  – Deployed with appropriate security controls
  – Office infrastructure need not be sized for science traffic
HPC Center Data Path

Diagram showing the architecture of a high-performance computing (HPC) center data path. The diagram includes:
- WAN
- Border Router
- Firewall
- Offices
- PerfSONAR
- Front end switch
- Data Transfer Nodes
- Supercomputer
- Parallel Filesystem

The diagram illustrates the flow of data through different components, highlighting the High Latency WAN Path and Low Latency LAN Path.
Common Threads

• Two common threads exist in all these examples
• Accommodation of TCP
  – Wide area portion of data transfers traverses purpose-built path
  – High performance devices that don’t drop packets
• Ability to test and verify
  – When problems arise (and they always will), they can be solved if the infrastructure is built correctly
  – Small device count makes it easier to find issues
  – Multiple test and measurement hosts provide multiple views of the data path
    • perfSONAR nodes at the site and in the WAN
    • perfSONAR nodes at the remote site
Dedicated Systems – Data Transfer Node

• The DTN is dedicated to data transfer
• Set up specifically for high-performance data movement
  – System internals (BIOS, firmware, interrupts, etc.)
  – Network stack
  – Storage (global filesystem, Fibrechannel, local RAID, etc.)
  – High performance tools
  – No extraneous software
• Limitation of scope and function is powerful
  – No conflicts with configuration for other tasks
  – Small application set makes cybersecurity easier – key point
Data Transfer Tools For DTNs

• Parallelism is important
  – It is often easier to achieve a given performance level with four parallel connections than one connection
  – Several tools offer parallel transfers, including Globus/GridFTP

• Latency interaction is critical
  – Wide area data transfers have much higher latency than LAN transfers
  – Many tools and protocols assume a LAN

• Workflow integration is important

• Key tools: Globus Online, HPN-SSH

• ESnet test DTNs: http://fasterdata.es.net/performance-testing/DTNs/
Data Transfer Tool Comparison

• In addition to the network, using the right data transfer tool is critical
• Data transfer test from Berkeley, CA to Argonne, IL (near Chicago). RTT = 53 ms, network capacity = 10Gbps.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCP:</td>
<td>140 Mbps</td>
</tr>
<tr>
<td>HPN patched SCP:</td>
<td>1.2 Gbps</td>
</tr>
<tr>
<td>FTP</td>
<td>1.4 Gbps</td>
</tr>
<tr>
<td>GridFTP, 4 streams</td>
<td>5.4 Gbps</td>
</tr>
<tr>
<td>GridFTP, 8 streams</td>
<td>6.6 Gbps</td>
</tr>
</tbody>
</table>

• NERSC DTNs have both HPN-SSH and Globus
• Key point – your local DTN and network connection significantly affect your ability to move data in and out of NERSC
Performance Between Computing Facilities

October 2016
L380 Data Set

Data set: L380
Files: 19260
Directories: 211
Other files: 0
Total bytes: 4442781786482 (4.4T bytes)
Smallest file: 0 bytes (0 bytes)
Largest file: 11313896248 bytes (11G bytes)
Size distribution:
1 - 10 bytes: 7 files
10 - 100 bytes: 1 files
100 - 1K bytes: 59 files
1K - 10K bytes: 3170 files
10K - 100K bytes: 1560 files
100K - 1M bytes: 2817 files
1M - 10M bytes: 3901 files
10M - 100M bytes: 3800 files
100M - 1G bytes: 2295 files
1G - 10G bytes: 1647 files
10G - 100G bytes: 3 files

23 – ESnet Science Engagement (engage@es.net) - 2/24/17
Handoff to Shreyas Cholia

- Thanks!