HPC’s Pivot to Data
Suzanne Parete-Koon (OLCF)
What is HPC’s Pivot to Data?

• The boundaries of the scientific computing ecosystem are being pushed beyond single centers providing HPC resources or any single experimental facility.

• Increasing data volume, variety, and velocity require additional resources to foster a productive environment for scientific discovery. This focus on data does not imply a reduction in importance of large-scale simulation capabilities.

• Predicting how and why scientific workflows achieve their observed end-to-end performance is a growing challenge for scientists.

• This talk will focus on data movement- because it is an excellent example of a task that requires multi-facility collaboration to enable multi-facility scientific workflows.
Data Growth

NCCS/OLCF HPSS Usage

NERSC HPSS Usage

Petabytes

2010 2011 2012 2013 2014


Petabytes

0.1 1 10 100 1000
Data Movement

ESnet Accepted Traffic: Jan 1990-Mar. 2014

Petabytes

100000
10000
1000
0.100
0.010
0.001
0.0001
0.00

1990
1995
2000
2005
2010
2014
Even projects with workflows designed to utilize only one HPC center need efficient data transfer, because data is typically moved to more permanent storage at the end of a project.
Intra-workflows: Simulation

Intra-workflows: Experiment

Single particle X-ray Crystallography
10E6 diffractive images reconstructed to reproduce angstrom scale 3-D structure.
Enabling Intra-workflow: Remote Transfers

Centers

• How can we enable secure data movement?
• What transfer tools are best for the system?
• How can we efficiently monitor our network?

Users

• Can I script the transfer? Click and forget?
• How long will the transfer take?
• What tools do you recommend?
The Science DMZ

• A network architecture explicitly designed for high-performance applications, where the science network is distinct from the general-purpose network

• The use of dedicated systems for data transfer

• Performance measurement and network testing systems that are regularly used to characterize the network and are available for troubleshooting

• Security policies and enforcement mechanisms that are tailored for high performance science environments
OLCF Infrastructure

Titan, 27PF, 299,008 Cores
18,866 Cray XK7 Nodes

Eos, 200 TF, 11,904 Cores
744 Cray XC30 Nodes

Rhea
Focus
EVEREST

Infiniband Core (SION)

18 SFA12K
500GB/s

Atlas1
144 FDR IB Ports

Atlas2
144 FDR IB Ports

Data Transfer Nodes

18 SFA12K
500GB/s

HPSS
35PB, 5GB/s ingest

Ethernet Core

1 x 100GbE
2 x 10GbE
NERSC Infrastructure

NERSC Facility: Computing, Data and Interconnect

**Hopper: 1.3PF, 212 TB RAM**
- Cray XE6, 150,000 Cores
- 2.2 PB Local Scratch, 70 GB/s

**Edison: >2PF, 333 TB RAM**
- Cray XC30
- 6.4 PB Local Scratch, 140 GB/s

**Production Clusters**
- Carver, PDSF, JGI, KBASE, HEP, BES
- 14x QDR, 2GB/s, 15 TB

**Vis & Analytics**
- Data Transfer Nodes
- Adv. Arch. Testbeds
- Science Gateways

**Ethernet & IB Fabric**
- Science Friendly Security
- Production Monitoring
- Power Efficiency
- WAN

**1.1 PB**
- 5 x DDN9900

**3.8 PB**
- 4 x DDN9900 NexSAN

**250 TB**
- 2 x NetApp 5460

**40 PB stored, 240 PB capacity, 20 years of community data**

**Global Scratch**
- 15 GB/s

**/project**
- 36 GB/s

**/home**
- 10 GB/s

**HPSS**
- 12 GB/s

**2 x 10 Gb**

**1 x 100 Gb**

ESnet

Software Defined Networking
Perfsonar: How well is the transfer performing?

http://fasterdata.es.net/performance-testing/perfsonar/perfsonar-dashboard/
Perfsonar: How is my transfer performing?
The Pros and Cons of SCP and Rsync

SCP Common on all Unix-like systems
Can be scripted into a workflow when the transfer destination allows for passwordless SSH logins.

Single stream: therefore slow and take poor advantage of WAN.
The common version do not allow much control over buffer size and fault checking.
The Pros and Cons Parallel Streams: GridFTP, globusonline, bbcp

Multi-streams allow fast transfers on WAN.
Many Options for customization.
Globus is extremely user-friendly once it is set up.

The software must be available at both end of the transfer and the availability varies.
Setup required.
Security policies impact ease of use.

Activation Energy

$E_a = 134 \text{ kJ}$
$\Delta H = -226 \text{ kJ}$
What rates can I expect?

<table>
<thead>
<tr>
<th>Service</th>
<th>Rate Mbits</th>
<th>Average Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUC</td>
<td>1 TB</td>
<td>1 TB</td>
</tr>
<tr>
<td>GUC (stripe)</td>
<td>1000 GB</td>
<td>1000 GB</td>
</tr>
<tr>
<td>Globus</td>
<td>2000 GB</td>
<td>2000 GB</td>
</tr>
<tr>
<td>BBCP</td>
<td>3000 GB</td>
<td>3000 GB</td>
</tr>
<tr>
<td>RSYNC</td>
<td>4000 GB</td>
<td>4000 GB</td>
</tr>
</tbody>
</table>

For 1 TB:
- GUC: 1500 Mbits
- GUC (stripe): 2000 Mbits
- Globus: 3000 Mbits
- BBCP: 3500 Mbits
- RSYNC: 4000 Mbits

For 10 100 GB files:
- GUC: 1000 Mbits
- GUC (stripe): 1500 Mbits
- Globus: 2000 Mbits
- BBCP: 2500 Mbits
- RSYNC: 3000 Mbits
Best tool features for the system?

**Throughput test using GUC**

<table>
<thead>
<tr>
<th>Test</th>
<th>Average Rate Mbit/s</th>
<th>% Network Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLCF Mem to NERSC Mem</td>
<td>4144</td>
<td>100</td>
</tr>
<tr>
<td>NERSC Disk to ORNL Mem</td>
<td>3335</td>
<td>80</td>
</tr>
<tr>
<td>OLCF Disk to NERSC Mem</td>
<td>2078</td>
<td>50</td>
</tr>
<tr>
<td>1 1TB file</td>
<td>1932</td>
<td>46</td>
</tr>
</tbody>
</table>

**Throughput test using GUC stripe: utilizing two DTNs**

<table>
<thead>
<tr>
<th>Test</th>
<th>Average Rate Mbit/s</th>
<th>% Network Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLCF Mem to NERSC Mem</td>
<td>8874</td>
<td>100</td>
</tr>
<tr>
<td>NERSC Disk to ORNL Mem</td>
<td>5339</td>
<td>60</td>
</tr>
<tr>
<td>ORNL Disk to NERSC Mem</td>
<td>4646</td>
<td>52</td>
</tr>
<tr>
<td>1 1TB file</td>
<td>2965</td>
<td>33</td>
</tr>
</tbody>
</table>

**Disk Throughput Measurement with fio**

<table>
<thead>
<tr>
<th>Site</th>
<th>IO Operation</th>
<th>Number of Processes</th>
<th>Average Rate Mbit/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>NERSC</td>
<td>read</td>
<td>1</td>
<td>14859</td>
</tr>
<tr>
<td>NERSC</td>
<td>read</td>
<td>2</td>
<td>13193</td>
</tr>
<tr>
<td>ORNL</td>
<td>write</td>
<td>1</td>
<td>2113</td>
</tr>
<tr>
<td>ORNL</td>
<td>write</td>
<td>2</td>
<td>3684</td>
</tr>
<tr>
<td>ORNL</td>
<td>write</td>
<td>8</td>
<td>11642</td>
</tr>
</tbody>
</table>

This filesystem benchmarking with fio shows that the performance improvement from GUC to GUC with striping comes from parallelism at the Lustre client.
## Case Studies

<table>
<thead>
<tr>
<th></th>
<th>Year</th>
<th>Methods</th>
<th>Data</th>
<th>Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combustion research challenge for multi-lab workflow</td>
<td>2007-2009</td>
<td>dtns/SCP/ Kepler workflow</td>
<td>10TB</td>
<td>OLCF⇒NERSC</td>
</tr>
<tr>
<td>20th Century Climate Reanalysis</td>
<td>2011</td>
<td>HPSS Direct</td>
<td>40 TB</td>
<td>OLCF⇒NERSC</td>
</tr>
<tr>
<td>DNS Combustion</td>
<td>2013</td>
<td>dtns/ Globus</td>
<td>80TB</td>
<td>OLCF⇒ALCF</td>
</tr>
<tr>
<td>LSCS</td>
<td>2013-14</td>
<td>dtns / bbcp Globus</td>
<td>130 TB</td>
<td>SLAC⇒NERSC</td>
</tr>
</tbody>
</table>
Road Maps: Data-Centric Services

- ORNL is establishing a Compute & Data Environment for Science (CADES). OLCF is a key partner.
- CADES will be a HUB to share data infrastructure and compute & data science capabilities with and among many projects.
Conclusion

- Success for workflows with large data movement depends critically on a "Science DMZ" like approach to connectivity.

- HPC centers and their users benefit from collaboration between different centers and facilities and collaboration between centers and projects.

- Computing centers traditionally focused on large-scale simulation are expanding their repertoire to include user-facing data services.

- Data services are built around a long-lived resources, the data itself. Both HPC architecture and allocation of will need to adapt to this longevity.
# Data Transfer Working Group

<table>
<thead>
<tr>
<th>CADES</th>
<th>ESnet</th>
<th>OLCF</th>
<th>NERSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. Shipman</td>
<td>E. Dart</td>
<td>B. Caldwell</td>
<td>S. Cannon</td>
</tr>
<tr>
<td>J. Zurawski</td>
<td>J. Hill</td>
<td>C. Layton</td>
<td>D. Hazen</td>
</tr>
<tr>
<td></td>
<td>S. Parete-Koon</td>
<td>D. Pelfrey</td>
<td>J. Hick</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H. Nam</td>
<td>D. Skinner</td>
</tr>
<tr>
<td></td>
<td></td>
<td>J. Wells</td>
<td></td>
</tr>
</tbody>
</table>

This manuscript has been authored by an author at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. This work used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.