Globus for Data Management

Rachana Ananthakrishnan
(ranantha@uchicago.edu)
Data Management Challenges

• “Transfers often take longer than expected based on available network capacities”

• “Lack of an easy to use interface to some of the high-performance tools”

• “Tools [are] too difficult to install and use”

• “Time and interruption to other work required to supervise large data transfers”

• “Need data transfer tools that are easy to use, well-supported, and permitted by site and facility cybersecurity organizations”

• “No easy mechanism to share selected data with collaborators”

Credit: Some data points from ESNet reports
We envisage a world where data flows rapidly, reliably, and securely among:

- experimental facilities,
- online and archival storage,
- computing facilities,
- and remote institutions
We envisage a world where data ... 

... is easily and securely shared with collaborators and partners at other institutions
We envisage a world where data ... is readily discoverable and accessible to collaborators, regardless of their and the data’s location
Reliable, secure, high-performance file transfer and synchronization

- “Fire-and-forget” transfers
- Automatic fault recovery
- Seamless security integration
- Powerful GUI and APIs

1. User initiates transfer request
2. Globus moves and syncs files
3. Globus notifies user
Interactive login to command line interface:

```
$ ssh tuecke@cli.globusonline.org
```

Running commands remotely:

```
$ ssh tuecke@cli.globusonline.org <command>
```

```
$ ssh tuecke@cli.globusonline.org scp -r -s 3 -D \nnersc#dtn:~/myfile* mylaptop:~/projects/p1
Task ID: 4a3c471e-edef-11df-aa30-1231350018b1
$ _
```

Using CLI with gsissh:

```
$ gsissh tuecke@cli.globusonline.org <command>
```
Simple, secure *sharing* off existing storage systems

- Easily share large data with any user or group
- No cloud storage required

1. User A selects file(s) to share, selects user or group, and sets permissions
2. Globus tracks shared files; no need to move files to cloud storage!
3. User B logs in to Globus and accesses shared file

• Easily share large data with any user or group
• No cloud storage required
Share files/folders

Transfer Files

Please select an endpoint above.
Manage permissions

Manage Shared Endpoint

Manage Permissions For ranantha#testshare

Host: ucrcc#midway:~/.share/CSSEF/

<table>
<thead>
<tr>
<th>name</th>
<th>read</th>
<th>write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rachana Ananthakrishnan</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Ian Foster</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Globus Team</td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>

A group added successfully

ID (User or Group) [ ]
Path /
Permissions [ ] read [ ] write

Add Cancel
Endpoint Administrator Controls

- **Restricted paths**
  - Files and folders that can be used to transfer
- Enable or disable ability to share from an endpoint
- **Sharing restricted paths**
  - Files and folders that can be used to share
- Whitelist of users who are allowed to share from an endpoint
- Only allow read permissions on shared endpoints created by users
Amazon S3 Endpoints
Deploying Globus...
8,000 active endpoints (in the past year)
<table>
<thead>
<tr>
<th>Institution</th>
<th>Endpoints</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALCF</td>
<td>alcf#*</td>
<td></td>
</tr>
<tr>
<td>JGI</td>
<td>jgi#*</td>
<td></td>
</tr>
<tr>
<td>LANL</td>
<td>Globus Connect Personal</td>
<td>Updating to DTNs</td>
</tr>
<tr>
<td>LBNL</td>
<td>lbnl#*</td>
<td>Sharing enabled</td>
</tr>
<tr>
<td>NERSC</td>
<td>nersc#*</td>
<td>Custom site</td>
</tr>
<tr>
<td>OLCF</td>
<td>olcf#*</td>
<td></td>
</tr>
<tr>
<td>PNNL</td>
<td>pic#*</td>
<td></td>
</tr>
<tr>
<td>SDSC</td>
<td>sdsc#*</td>
<td>Sharing enabled</td>
</tr>
<tr>
<td>XSEDE</td>
<td>xsede#*</td>
<td>In process of enabling Sharing</td>
</tr>
</tbody>
</table>
85 U.S. campuses
Globus Connect Personal

Create Globus Connect Personal Endpoint

Step 1  Get Your Globus Connect Personal Setup Key
Please enter a unique name for your Globus Connect Personal endpoint to help you identify it.

Endpoint Name: rananthalaptop25

Setup Key: 5768f6dd-13b3-4ef5-933d-d43fb508a6a6
Copy the key above to your clipboard.

Step 2  Download & Install Globus Connect Personal
Click one of the buttons below to download and install Globus Connect Personal for your operating system.

- [for Mac OS X]
- [for Linux]
- [for Windows]

Once downloaded, run the installer. When prompted, paste in the Setup Key to complete the installation.

globus connect personal
Drag the Globus icon to the Applications folder.

Globus Connect Personal
Applications
Globus increasingly used to better utilize network

Enable computing facilities to better utilize high performance network infrastructure

Source: University of Nebraska Holland Computing Center
Using Globus...
Ann Syrowski (Illinois) moves data across XSEDE, NCSA MSS and PSC, to leverage HPC facilities across the country.
Dan Kozak (Caltech) replicates 1 PB LIGO astronomy data for resilience
Erin Miller (PNNL) collects data at Advanced Photon Source, renders at PNNL, and views at ANL

Credit: Kerstin Kleese-van Dam
Users upload data to their archives at University of Exeter: Globus API used for integration into their services
Earth System Grid Federation leverages Globus for data download: including the user interfaces
Using Globus web pages
Globus is moving beyond transfer and sharing to data publication and discovery
Curated **publishing** and rich **discovery** of research data.

- **Describe and publish data in hosted collections**
- **Leverage institutional storage**
- **Customizable publication and curation workflows**

1. User A chooses a collection to publish into, selects file(s) to publish, and associates descriptive metadata
2. Globus stores the data in collection’s external storage; no need to move files to cloud storage!
3. User B uses Globus to discover and download published datasets
Globus Data Publication

- SaaS for publishing large research data
- Bring your own storage
- Extensible metadata
- Publication and curation workflows
- Public and restricted collections
- Rich discovery model
We are a non-profit, delivering a production-grade service to the non-profit research community

Our challenge: Sustainability
Globus Provider Subscriptions

- Managed Endpoints
  - Priority support
  - Management console
  - Usage reports
  - Mass Storage System optimization
  - Host shared endpoints
  - Integration support

- Branded Web Site

- Alternate Identity Provider (InCommon is standard)

globus.org/provider-plans
Thank you to our sponsors!

U.S. DEPARTMENT OF ENERGY

NSF

NATIONAL INSTITUTES OF HEALTH

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

powered by amazon web services