Globus for Research Data Management

Stuart Martin
Software Development Manager, Globus
sjmartin@uchicago.edu

August 11 2016
Globus delivers...

(big) data transfer, sharing, publication, and discovery...

...directly from your own storage systems
“I need to easily, quickly, & reliably move or replicate my data to other systems.”
“I need to get data from a scientific instrument to my analysis system.”
“I need to easily and securely share my data with my colleagues at other institutions.”
“I need a good place to store / backup / archive my (big) research data, at a reasonable price.”
“I need to publish my data so that others can find it and use it.”
Reliable, secure, high-performance file transfer and replication

- "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
Storage connectors

- **Standard supported storage connectors (Posix)**
  - Linux, Windows, MacOS
  - Lustre, GPFS, OrangeFS, etc.

- **Premium storage connectors**
  - HPSS
  - HDFS
  - S3
  - Ceph RadosGW (S3 API)
  - Spectra Logic BlackPearl
  - Google Drive
Demonstration:
Sign Up
File Transfer
Globus Transfer API

Nearly all Globus Web App functionality implemented via public Transfer API

This API provides a REST-style interface to the Globus reliable file transfer service. The Transfer API supports monitoring the progress of a user’s file transfer tasks, managing file transfer endpoints, listing remote directories, and submitting new transfer and delete tasks. The API is ideal for integration into Portals or Gateways to provide complex reliable file transfer capabilities without having to develop and support these features on your own. It is also easy to use for scripting, using any standard HTTPS or REST client library in scripting languages like Python and Ruby.

Contents

- API Overview - overview of API with authentication instructions and examples
- Endpoint Activation - associate user credentials with an endpoint
- Task Submission - submit transfer and delete tasks
- Task Management - monitor and cancel background transfer and delete tasks
- File Operations - foreground filesystem operations, including directory listing (ls), creating directories (mkdir), and renaming files (rename)
- Endpoint Management - create, update, and delete endpoint definitions and servers
- Endpoint Search - find endpoints matching a search query, by display_name and other endpoint fields

https://docs.globus.org/api/transfer/
New Globus CLI
(coming soon)

$ globus transfer
Usage: globus transfer [OPTIONS] COMMAND [ARGS]...

Interact with Globus Transfer API. Transfer, Delete, List, and Rename files on Endpoints, manage your Endpoints and Shares, and monitor your ongoing Transfer Tasks

Options:
-F, --format [json|text]  Output format for stdout. Defaults to text
-h, --help                 Show this message and exit.
--version                 Show the version and exit.

Commands:
  acl                       Manage Endpoint Access Control Lists
  async-delete             Submit a Delete Task
  async-transfer           Submit a Transfer Task
  bookmark                 Manage Endpoint Bookmarks
  endpoint                 Manage Globus Endpoint definitions
  ls                       List Endpoint directory contents
  mkdir                    Make a directory on an Endpoint
  rename                   Rename a file or directory on an Endpoint
  task                     Manage asynchronous Tasks
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
Demonstration: Globus sharing Group Management
Curated *publication* of data, with relevant metadata for *discovery*

- Identify
- Describe
- Curate
- Verify
- Access
- Preserve

1. Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)
2. Curator reviews and approves; data set published on campus or other storage
3. Peers, public search and discover data sets; transfer using Globus
Typical publication workflow

1. Publish Data
2. Describe Submission
3. Assemble Dataset (Transfer Data)
4. Curate Dataset
5. Search
6. Download

Argonne Storage System

Shared Endpoint

Typical publication workflow
## Publication GA release

### Identifier
- **Supported in GA release**
- **Consulting support**
- **Planned**

<table>
<thead>
<tr>
<th>Identifier</th>
<th>URL</th>
<th>Handle</th>
<th>DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Standard</td>
<td>Domain-specific</td>
<td>Custom</td>
</tr>
</tbody>
</table>

### Description
- None
- Standard
- Domain-specific
- Custom

### Curation
- None
- Acceptance
- Human-validated
- Machine-validated

### Access
- Anonymous
- Public
- Embargoed
- Collaborators

### Preservation
- Transient
- Project Lifetime
- Archive
- “forever”
Platform Questions

• How do you leverage Globus services in your own applications?

• How do you extend Globus with your own services?

• How do we empower the research community to create an integrated ecosystem of services and applications?
**NCEP Climate Forecast System Version 2 (CFSv2) Monthly Products**

ds094.2

For assistance, contact Bob Dattore (303-497-1825).

### Description

<table>
<thead>
<tr>
<th>Data Description</th>
<th>Data File Downloads</th>
<th>Customizable Data Requests</th>
<th>Other Access Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Web Server Holdings</td>
<td>Globus Transfer Service (GridFTP)</td>
<td>Subsetting</td>
</tr>
</tbody>
</table>

### Union of Available Products

- **Diurnal monthly means**
- **Regular monthly means**
- **Selected Parameter/Level Time Series**

**NCAR-Only Access**

- **Central File System (GLADE) Holdings**
- **Tape Archive (HPSS) Holdings**

---

**Hello tuecke@uchicago.edu**
Globus at a glance

- 4 major services
- 100 PB transferred
- 20 billion files processed
- 25,000 registered users
- 13 national labs use Globus services
- 8,000 active endpoints
- >300 users are active daily
- 99.95% uptime over the past two years
- >30 institutions subscribed
- 1 PB largest single transfer to date
- 3 months longest continuously managed transfer
- 157 federated campus identities
Globus endpoints worldwide
Globus is enabling…

Study of the structure and evolution of galaxies, the nature of dark energy, and cosmological history of the universe

Joel Brownstein
University of Utah

Source: University of Utah
Globus is enabling…

Development of numerical simulations of severe storms for improved responsiveness to weather events

Ann Syrowoski
University of Illinois
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
Globus Subscriptions

Globus file transfer is free to use for non-profit research and educational purposes. Subscriptions offer researchers access to enhanced features such as file sharing and data publication, and provide system administrators with a comprehensive view into how storage systems are being utilized. Your subscription helps us make Globus self-sustaining so we can continue to offer high-quality data management infrastructure to all researchers.

Click here to start your 90-day free trial of Globus premium features

Subscriptions for Non-Profit Research and Education

If you wish to use Globus in a commercial setting, you must have a commercial subscription.

<table>
<thead>
<tr>
<th>Features (click  for description)</th>
<th>Basic - Free</th>
<th>Starter Subscription</th>
<th>Standard Subscription</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer level</td>
<td>Unlimited</td>
<td>Unlimited</td>
<td>Unlimited</td>
</tr>
<tr>
<td>User level</td>
<td>Unlimited</td>
<td>Unlimited</td>
<td>Unlimited</td>
</tr>
<tr>
<td>Managed endpoints</td>
<td>None</td>
<td>1</td>
<td>Unlimited</td>
</tr>
<tr>
<td>Management console</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Usage reports</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Support for Globus Connect, Web, CLI</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Shared endpoints</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Globus Plus users</td>
<td>—</td>
<td>—</td>
<td>✓</td>
</tr>
</tbody>
</table>

https://www.globus.org/pricing
Globus Developer Workshops

We’re presenting a series of Globus developer workshops across the US, building on the success of the workshop held at GlobusWorld 2016. These workshops are made possible by the various hosting institutions that generously provide meeting space and other financial support.

The following workshops are currently scheduled:
- September 13-14, 2016 - Lawrence Berkeley National Laboratory
- October 12-13, 2016 - Yale University

If you would like to host a workshop at your institution please contact us.

Workshops are free to attend and open to all, but we do require registration since attendance will be limited by venue size.

Why Attend?
- See how to easily provide Globus services on existing storage systems
- Hear how other institutions are using Globus
- Learn from the experts on the best ways to apply Globus technologies
- Connect with colleagues and Globus developers

https://www.globusworld.org/workshops/
Sample python data portal

Modern Research Data Portal
It's how research data management is done!

**Globus Transfer API**
API reference for transfer and sharing functions.

**Globus Auth API**
API reference for authentication and authorization.

**Frequently Asked Questions**
When all else fails...

Next-Generation Portal Leverages Science DMZ

Portal server applications:
- web server
- search
- database
- authentication

API DTNs (data access governed by portal)

Data Transfer Path
Portal Query/Browse Path

https://fasterdata.es.net/
Move portal storage into Science DMZ, with Globus endpoint

Leave Portal Web server behind firewall

Globus handles the security and data heavy lifting
Developer Workshops scheduled

• September 13-14, 2016 - LBNL, Berkeley, CA
• October 12-13, 2016 - Yale University, New Haven, CT
• Date TBD - NCAR, Boulder, CO
Questions

More information:
www.globus.org
www.globus.org/data-publication

Contact:
support@globus.org
Hands-on exercise

http://tinyurl.com/globusatpesc