Introduction to Globus for New Users

SaaS for Research Data Management

Vas Vasiliadis
vas@uchicago.edu

NYSERNet– May 1, 2018
Research data management today

How do we...
...move?
...share?
...discover?
...reproduce?
Globus delivers...

Secure, reliable, data transfer, sharing, publication, and discovery...

...directly from your own storage systems...

...via software-as-a-service
Globus enables...
Campus Bridging
...within and beyond campus boundaries
Bridge to campus HPC

Move datasets to campus research computing center

Move results to laptop, department, lab, ...
Bridge to national cyberinfrastructure

Move datasets to supercomputer, national facility

Move results to campus (...)

MIRA
Bridge to instruments

Pre-processed Data

Analysis store

High durability, low cost store

Raw Source Data

Amazon Glacier

High durability, low cost store

Analysis store

Pre-processed Data
Bridge to collaborators

- XSEDE
- JetStream
- NERSC
- EC2
- External Campus Storage
- Public/Private Cloud stores

NEW YORK UNIVERSITY

CORNELL UNIVERSITY

red cloud
ceph
openstack
amazon S3
Google Drive

Globus Connect

Cloud stores
Bridge to community/public

Project Repositories, Replication Stores

Public Repositories
Globus SaaS: Research data lifecycle

1. Researcher initiates transfer request; or requested automatically by script, science gateway.

2. Globus transfers files reliably, securely.

3. Researcher selects files to share, selects user or group, and sets access permissions.

4. Globus controls access to shared files on existing storage; no need to move files to cloud storage!

5. Collaborator logs in to Globus and accesses shared files; no local account required; download via Globus.

6. Researcher assembles data set; describes it using metadata (Dublin core and domain-specific).

7. Curator reviews and approves; data set published on campus or other system.

8. Peers, collaborators search and discover datasets; transfer and share using Globus.

- Access any storage...
- ...with an existing identity
- ...using a Web browser
Conceptual architecture: Hybrid SaaS

DATA Channel

Source Endpoint

Globus "client" software

Destination Endpoint

No data relay or staging via Globus

Source owned and administered storage system

CONTROL Channel

Subscriber Control Domain

Globus Control Domain

Single, globally accessible multi-tenant service
Endpoints: A Storage Abstraction

• Endpoint: Storage abstraction
• All transfers happen between two endpoints
• Testing endpoints
  – Globus Tutorial Endpoints (initial endpoint validation)
  – ESnet Test Endpoints (diverse dataset samples)

...makes your storage system a Globus endpoint
Globus Connect Personal

- Installers do not require admin access
- Zero configuration; auto updating
- Handles NATs
Globus Connect Server

- Makes your storage accessible via Globus
- Multi-user server, installed and managed by sysadmin
- Default access for all local accounts
- Native packaging
  - Linux: DEB, RPM

docs.globus.org/globus-connect-server-installation-guide/
Demonstration
File Transfer
Data Sharing

- **Select the directory and create a “shared endpoint”**
  - …just another type of endpoint
  - …but does not require activation

- **Share with user/group/all Globus users**

- **Roles delegate control/monitor rights to endpoints**
Conceptual architecture: Sharing

- **External User Control Domain**
- **Managed Endpoint**
- **Subscriber Control Domain**
- **Globus Control Domain**

**DATA Channel**
- **CONTROL Channel**

**Managed Endpoint**
- Administrator managed filesystem permissions

**Shared Endpoint**
- User managed "overlay" permissions

**Globus**
- **globus connect**

---

**User managed**
- "overlay" permissions

---

**External User Control Domain**
- A person's avatar with the role of an external user.
Demonstration

File Sharing
Groups

- **Sharing**: Access permissions for multiple people
- **Roles**: Endpoint management and monitoring
- **Key concepts**
  - Visibility
  - Membership model
  - Subgroups
  - Settings (policies, additional data, terms)
  - Roles (delegating authority to others)
Demonstration
Group Management
The Globus Web App - Hidden in Plain Sight

• The Hamburger Menu
  – Varies by endpoint/storage type
  – A great place to get the link to a share

• Transfer Settings
  – label – see recognizable names in activity monitor
  – sync - only transfer new or changed files
  – delete files on destination that do not exist on source
  – preserve source file modification times
  – verify file integrity after transfer **
  – encrypt transfer

• Unified search: Endpoints, users, groups
Bookmarks

• Just like browser bookmarks – frequently used, or maybe not used frequently enough!

• Creating a bookmark

• Using a bookmark

• Sorting and filtering

• Editing and deleting
Data Publication and Discovery

Materials Data Facility Community home page

The Materials Data Facility (MDF) is a scalable repository where materials scientists can publish, preserve, and share research data. The repository provides a focal point for the materials community, enabling publication and discovery of materials data of all sizes.

MDF is a pilot project funded by NIST, and serves as the first pilot community of the National Data Service.

Contact Ben Blaiszik (blaiszik@uchicago.edu) to begin publishing your data

https://publish.globus.org
## Globus data publication framework

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

<table>
<thead>
<tr>
<th>Description</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Anonymous</td>
</tr>
<tr>
<td>Acceptance</td>
<td>Public</td>
</tr>
<tr>
<td>Human-validated</td>
<td>Embargoed</td>
</tr>
<tr>
<td>Machine-validated</td>
<td>Collaborators</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Curation</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
</tr>
<tr>
<td>Acceptance</td>
</tr>
<tr>
<td>Human-validated</td>
</tr>
<tr>
<td>Machine-validated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anonymous</td>
</tr>
<tr>
<td>Public</td>
</tr>
<tr>
<td>Embargoed</td>
</tr>
<tr>
<td>Collaborators</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transient</td>
</tr>
<tr>
<td>Project Lifetime</td>
</tr>
<tr>
<td>Archive</td>
</tr>
<tr>
<td>“forever”</td>
</tr>
</tbody>
</table>
Demonstration
Data Publication
Why use Globus?

• Simplicity
  – Consistent UI across systems
  – Easy access to collaborators

• Reliability and performance
  – “Fire-and-forget” file transfer
  – Maximized WAN throughput

• Operational efficiency
  – Low overhead SaaS model
  – Highly automatable: CLI, RESTful API

• Access to a large and growing community
How can I integrate Globus into my research workflows?
Globus serves as...

...a platform for building science gateways, portals, and other web applications in support of research and education
Use(r)-appropriate interfaces

Globus service

Web

CLI

Rest API

GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
...
Globus as PaaS

- Use existing institutional ID systems in external web applications
- Integrate file transfer and sharing capabilities into scientific web apps, portals, gateways, etc.
Globus PaaS developer resources

Python SDK

Requirements
- You need to be in the tutorial users group for sharing: https://www.globus.org/app/groups/50b6a29c-63ac-11e4-8062-22000b92c687
- Installed Globus Python SDK

```
from future import print

tutorial_endpoint_1 = "ddb59af0-6d04-11e5-ba46-22000b92c687" # endpoint "Globus CLI"

In [15]:
tutorial_endpoint_2 = "ddb59af0-6d04-11e5-ba46-22000b92c687" # endpoint "Globus CLI"

In [16]:
tutorial_users_group = "50b6a29c-63ac-11e4-8062-22000b92c687" # group "Tutorial Users"
```

Configuration

First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you can use the following website. Click the "Jupyter Notebook" option and copy the resulting text below, or click on "Globus CLI" and use the following command:

```
In [16]:
transfer_token = None # if None, tries to get token from ~/.globus.cfg file
```
Thank you to our sponsors…

U.S. Department of Energy

NSF

The University of Chicago

NIH

National Institute of Standards and Technology
U.S. Department of Commerce

Argonne National Laboratory

powered by Amazon Web Services
Thanks to our workshop sponsor!

Western Digital®
Globus sustainability model

- **Standard Subscription**
  - Shared endpoints
  - Data publication
  - Management console
  - Usage reporting
  - Priority support
  - Application integration
  - HTTPS support (coming soon)

- **Branded Web Site**

- **Premium Storage Connectors**

- **Alternate Identity Provider (InCommon is standard)**
Globus by the numbers

- 1,042 most shared endpoints at a single institution
- 400 PB transferred
- 66 billion files processed
- 100,000 users
- 24 Petabyte+ institutions
- 15,000 active transfer users
- 3 months longest running transfer
- 20,000 active endpoints
- 500+ identity providers
- 1 PB largest single transfer to date
- 8,000 active shared endpoints
- 99.9%+ availability
Our supporters
Join the Globus community

- Access the service: globus.org/login
- Create a personal endpoint: globus.org/app/endpoints/create-gcp
- Documentation: docs.globus.org
- Engage: globus.org/mailing-lists
- Subscribe: globus.org/subscriptions
- Need help? support@globus.org
- Follow us: @globusonline