Globus: The Platform for Research Data Management

Vas Vasiliadis
ATPESC – August 6, 2021
Warning:
We may hear from unregistered attendees!
Globus is …

a non-profit service developed and operated by

THE UNIVERSITY OF CHICAGO
Our mission is to...

increase the efficiency and effectiveness of researchers engaged in data-driven science and scholarship through sustainable software
180,000
REGISTERED USERS

1,200+
IDENTITY PROVIDERS

35,000
ACTIVE ENDPOINTS

35,000
ACTIVE ENDPOINTS

1PB
RELIABLE TRANSFER PER DAY

1,600+
CONNECTED INSTITUTIONS

9,500+
ACTIVE SHARED ENDPOINTS

Users in 80+ COUNTRIES

Globus
PLATFORM AS A SERVICE
SOFTWARE AS A SERVICE

Developer APIs
Workflow Automation
File Transfer
Access Control
Modular Apps
Identity Management
Local Storage
Institutional Storage
Tape Archives
High Performance Computing
Commercial Cloud Storage
Globus and the 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. The Globus Command Line Interface, API sets, Python SDK and Action Providers give you a platform…

7. … for building science gateways, portals and automations.

8. Streamlining research workflows and ensuring those that need access to the data have it.

- Use a Web browser, CLI or platform services
- Access any storage
- Use an existing identity
Hybrid SaaS Architecture

Customer owned and administered storage system with Globus Connect running on it.

DATA CHANNEL
No data relay or staging via Globus, files move directly between storage locations.

CONTROL CHANNEL

Subscriber Security Domain
User identity mapped to local account.

Globus Security Domain
Globus service orchestrates file movement via communication with Globus Connect.

EXTERNAL USER OR APPLICATION

globus services

powered by Amazon Web Services
Endpoints, Collections and Globus Connect

- **Globus Connect Server**
  - Multi-user Linux Systems
  
  docs.globus.org/globus-connect-server

- **Globus Connect Personal**
  - Personal workstations and laptops
  - OS specific instructions: docs.globus.org/how-to
The *ad hoc* user’s perspective..
Use(r)-appropriate interfaces

Globus service

Web

Platform (RESTful APIs)

CLI

GET /endpoint/go%23ep1
PUT /endpoint/demodoc#my_endpt

200 OK

X-Transfer-API-Version: 0.10
Content-Type: application/json

Web interface

CLI interface

Globus service interface

Platform (RESTful APIs) and interfaces:

- **Web**: Utilizes URLs for accessing endpoints.
- **CLI**: Provides commands for managing endpoints and services.
- **Platform (RESTful APIs)**: Use HTTP methods like GET and PUT for interacting with the service.

Example URLs:

- GET /endpoint/go%23ep1
- PUT /endpoint/demodoc#my_endpt

Response:

- 200 OK
- X-Transfer-API-Version: 0.10
- Content-Type: application/json
Globus CLI enables simple automation...

- Open source native app
- Uses Python SDK
- Consistent access control model
  - Access and refresh tokens
  - Tokens stored locally
- docs.globus.org/cli
...such as data staging for compute

- User securely uploads data for analysis
- Results available with fine-grained permissions
- Automated setup/tear down of folders, permissions
BUT, the instruments are coming!
Automating XPCS at scale

1. Imaging
2. Acquisition
3. XPCS-Eigen
4. Plot results
5. Publication
6. Discovery
7. Science!
Making data discoverable

petreldata.alcf.anl.gov
Instrument data orchestration

• Authentication and Authorization
• Data transfer and sharing
• Data description and discovery
• Data (and compute) orchestration
Globus Auth: Foundational IAM service

Brokers authentication and authorization among...

- End-users
- Identity providers: enterprise, external (federated identities)
- Services: resource servers with REST APIs
- Apps: web, mobile, desktop, command line clients
- Services acting as clients to other services

- OAuth 2.0 Authorization Framework (a.k.a. OAuth2)
- OpenID Connect Core 1.0 (a.k.a. OIDC)
Step 0: Application registration

- Set desired scopes
- Set callback URL
- Get client ID and secret
- Consents implement least privileges principle

developers.globus.org
Authorization Code Grant

1. Access portal

2. Redirect user

3. User authenticates and consents

4. Authorization code

5. Authenticate using client id and secret, send authorization code

6. Access token(s)

7. Authenticate with access token(s), giving client authority to invoke the requested service
Globus Python SDK

• Python client library for accessing Globus APIs
• For example, `globus_sdk.TransferClient` class handles connection management, security, framing, marshaling when accessing the Transfer service

```python
from globus_sdk import TransferClient, AuthClient
tc = TransferClient()
```

globus.github.io/globus-sdk-python
Experimenting with Globus platform services

jupyter.demo.globus.org
Data transfer and sharing

- Move data to collection → Submit Transfer task
- Make data accessible → Set guest collection access rule
- Grant user(s) access → Add/confirm Group membership

POST /endpoint/{endpoint_id}/access
POST /transfer
GET /groups/my_groups

Transfer service
Groups service
Data description and discovery

• Metadata store with fine-grained visibility controls
• Schema agnostic ➔ dynamic schemas
• Simple search using URL query parameters
• Complex search using search request document

docs.globus.org/api/search
Data ingest with Globus Search

POST /index/{index_id}/ingest'

```json
{
    "ingest_type": "GMetaList",
    "ingest_data": {
        "gmeta": [
            {
                "id": "filetype",
                "subject": "https://search.api.globus.org/abc.txt",
                "visible_to": ["public"],
                "content": {
                    "metadata-schema/file#type": "file"
                }
            },
            ...
        ]
    }
}
```
Data ingest with Globus Search

**POST /index/{index_id}/ingest**

```json
{
    "ingest_type": "GMetaList",
    "ingest_data": {
        "gmeta": [
            {
                "id": "size",
                "subject": "https://search.api.globus.org/abc.txt",
                "visible_to": ["urn:globus:auth:identity:46bd0f56-e24f-11e5-a510-131bef46955c"],
                "content": {
                    "metadata-schema/file#size": "1000000",
                    "metadata-schema/file#size_human": "1MB"
                }
            }
        ]
    }
}
```

Visibility limited to Globus Auth identity
- Single user
- Globus Group
- Registered client application
Data discovery with Globus Search

GET /index/{index_id}/search?q=type%3Ahdf5

```json
{
  "@datatype": "GSearchResult",
  "@version": "2017-09-01",
  "count": 1,
  "gmeta": [
    {
      "@datatype": "GMetaResult",
      "@version": "2019-08-27",
      "entries": [
        { ... }
      ],
      "subject": "https://...
    }
  ],
  "offset": 0,
  "total": 1
}
```
POST /index/{index_id}/search

```
{
  "filters": [
    {
      "type": "range",
      "field_name": "pubdate",
      "values": [
        {
          "from": "*",
          "to": "2020-12-31"
        }
      ]
    }
  ],
  "facets": [
    {
      "name": "Publication Date",
      "field_name": "pubdate",
      ...
    }
  ]
}
```
Data (and compute) automation

- **Flows**: A platform service for defining, applying, and sharing distributed research automation flows
- Flows comprise **Actions**
- **Action Providers**: Called by Flows to perform tasks
- **Triggers***: Start flows based on events

* In development
globus.org

docs.globus.org

outreach@globus.org

support@globus.org