Data Mobility Perspectives and the Modern Research Data Portal

Vas Vasiliadis
vas@uchicago.edu

Data Mobility Workshop
September 23, 2019
MRDP: Key elements

- **Science DMZ**: Fast, clean data path
- **Data Transfer Node**: Purpose-built data mover
- **Globus Platform**: Secure, reliable data orchestration
- **Globus Connect**: Storage system enabler
Who/What/Where is Globus?

Data mobility perspectives…
- Researcher
- Service provider/sysadmin
- Portal/app developer
...a brief detour on sustainability
...and THANK YOU, subscribers!
Globus sustainability model

- **Standard Subscription**
  - Sharing, data publication
  - HTTPS access
  - Console, usage reporting
  - Priority support
  - App integration support

- **High Assurance subscription**
  - App instance isolation
  - Additional authentication assurance
  - Audit logging
  - NIST 800-53, NIST 800-171 (+ BAA)

- **Branded Web Site**
- **Premium Storage Connectors**
- **Alternate Identity Provider** (InCommon is standard)
The Researcher Perspective
Unified data access across diverse storage systems
Sharing with collaborators, community

- Project repositories, replication stores
- Public repositories
- External campus storage
- Public / private cloud stores
  - red cloud, ceph, openstack
  - Google Drive, Amazon S3
Managing data from instruments

Next-Gen Sequencer
Cryo-EM
Advanced Light Source
MRI
Light Sheet Microscope
Analysis store
High-durability, low-cost store
Remote visualization
Personal system
Endpoints (Collections)

• **Software you deploy on your DTN – Globus Connect**
• **Storage abstraction**
  – All transfers happen between two endpoints
• **Collection ~= Endpoint**
• **Test / Demo Endpoints**
  – Globus Tutorial Endpoint 1/2
  – ESnet Read-Only *
  – DME Datasets *, DME PerfTest *
Globus Web App
File transfer
Data Sharing
Linked Identities
Groups
Globus security

- **Access control**
  - Identities provided and managed by institution
  - Globus as identity broker: no access to institutional user credentials
  - Institution controls all access policies
- **Researchers can overlay sharing permissions**
  - Data remain at institution, not hosted by Globus
- **Automated integrity checks of transferred data**
- **High service availability**
- **Monitoring**
- **Encryption: all communications, data in transit (optional)**
Globus for high assurance data management

• Restricted data handling: PHI, PII, CUI
• Security controls: NIST 800-53, 800-171 Low
• Business Associate Agreement (BAA) w/ UChicago
  – University of Chicago has a BAA with Amazon
• “Equivalent” UK/EU privacy contractual agreements
  – e.g. to cover Data Processor requirements under GDPR
High Assurance features

• **Additional authentication assurance**
  – Authenticate with a specific identity within a session
  – Reauthenticate after specified time period

• **Application instance isolation**
  – Authentication context is per application, per session

• **Forced encryption of data in transit**

• **Local audit logs (on data transfer nodes)**
Globus security: Operational security

- Intrusion detection and prevention
- Performance and health monitoring
- Logging
- Secure remote access, access control
- Uniform configuration management and change control
- Backups and disaster recovery
- Service data encrypted at rest
- AWS best practices (VPCs, IAM, Security Groups)
Protected Data Management

Authentication assurance

Session isolation
makes storage systems accessible via Globus
Globus Connect Personal

- Rapid installation/removal by non-privileged account
- Zero configuration; auto updating
- Handles NATs
The Service Provider and System Administrator Perspective
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/
Storage Connectors - globus.org/connectors

Current
- IBM Spectrum Scale
- S3
- ceph
- Google Drive
- Lustre
- box
- HPSS
- Western Digital
- ACTIVE SCALE
- hadoop

Planned
- Google Cloud
- Microsoft Azure
- wasabi
- hot cloud storage
Which version of Globus Connect Server do I use?

By default, assume you should use GCS v4
GCS: Common configuration options

- **Endpoints page**
  - Display Name
  - Visibility
  - Encryption

- **DTN configuration file**
  - RestrictPaths
  - Sharing
  - SharingRestrictPaths
  - IdentityMethod (MyProxy, CILogon, Oauth)
Ports needed for Globus Connect Server v4

- Inbound: 2811 (control channel)
- Inbound: 7512 (MyProxy), 443 (OAuth)
- Inbound: 50000-51000 (data channel)
- If restricting outbound connections, allow connections on:
  - 80, 2223 (used during install/config)
  - 50000-51000 (GridFTP data channel)
- GCSv5 removes some of these requirements – yay!
Endpoint activation using MyProxy

Default configuration
(avoid if at all possible)
Endpoint activation using MyProxy OAuth

Best practice configuration
Just do it!
Please…
Alternative authentication methods and SSO

• **InCommon, EduGAIN, and other federations…**
  – Release R&S attributes to CILogon (especially ePPN)
  – Local account must match InCommon ID
  – In /etc/globus-connect-server.conf set:
    AuthorizationMethod = CILogon
    CILogonIdentityProvider = <institution_name>
  – Local account must match InCommon ID

• **Alternate identity providers**
  – Globus can add alternate IdPs to trusted list
  – Requires add-on subscription
Subscription configuration

- Subscription manager can create managed endpoints
- ...Required for sharing, management console, usage reporting, ...
- Configurable via web app or CLI
Visibility and Control
Endpoint roles
Management console
Usage reporting
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 Command Line Interface (CLI)

- Native application: docs.globus.org/cli
- Open source, uses Python SDK
- `globus login` – get access and refresh tokens
  - Tokens stored locally in ~/.globus.cfg
- Service (transfer/auth) invocation uses tokens
- `globus logout` – delete tokens
  
  docs.globus.org/cli/examples
UUIDs everywhere

- UUIDs for endpoint, task, user identity, groups…
- Use search/list options
- `get-identities` for identity username to UUID

```
$ globus endpoint search 'Globus Tutorial'
$ globus task list
$ globus get-identities vas@globus.org 14bf3755-6267-42f2-9e9c-ad324de4a1fb
```
Batch Transfers

- Transfer tasks have one source/destination, but can have any number of files
- Provide input source-dest pairs via local file
- e.g. move files listed in files.txt from $ep1 to $ep2

```
$ ep1=e261ffb8-6d04-11e5-ba46-22000b92c6ec
$ ep2=af7bda53-6d04-11e5-ba46-22000b92c6ec
$ globus transfer $ep1:/share/godata/ $ep2:/~/ --batch --label 'CLI Batch' < files.txt
```
 Parsing CLI output

- **Default output is text; for JSON output use** `--format json`

  ```
  $ globus endpoint search --filter-scope my-endpoints
  $ globus endpoint search --filter-scope my-endpoints --format json
  ```

- **Extract specific attributes using** `--jmespath <expression>`

  ```
  $ globus endpoint search --filter-scope my-endpoints --jmespath 'DATA[].[id, display_name]'
  ```
Permission management

- Set and manage permissions on shared endpoint
- Requires access manager role

```bash
$ share=<shared_endpoint_UUID>
$ globus endpoint permission create --permissions r --identity tuecke@globus.org $share:/NCARTest/
$ globus endpoint permission list $share
$ globus endpoint permission delete $share <perm_UUID>
```
Automation Examples

• Syncing a directory
  – bash script; calls the Globus CLI
  – Python module; run as script or import as module

• Staging data in a shared directory
  – bash and Python variants

• Removing directories after files are transferred
  – Python script

github.com/globus/automation-examples
The (Portal/Gateways/Apps/...) Developer Perspective
How can I (more tightly) integrate Globus into my research workflows?
Globus serves as...
A platform for building science gateways, web portals and other applications in support of research and education
Example: Data repositories
Example: Analysis services

Sanger Imputation Service
This is a free genotype imputation and phasing service provided by the Wellcome Trust Sanger Institute. You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genomes back. Click here to learn more and follow us on Twitter.

Before you start
Be sure to read through the instructions. You will need to set up a free account with Globus and have Globus Connect running at your institute or on your computer to transfer files to and from the service.

Ready to start?
If you are ready to upload in the details below to request an imputation and/or phasing job. If you have any questions, see the about page.

Full name
Organisation
Email address

DLHub
Data and Learning Hub for Science
A simple way to find, share, publish, and run machine learning models and discover training data for science

Documentation
Examples Read the Docs Python SDK CLI

Papers and Presentations
DLHub on ArXiv DLHub Slides
Example: Instruments

DMagic a Globus implementation at the APS

http://dmagic.readthedocs.org

http://tompy.readthedocs.org
Globus Platform-as-a-Service

Globus APIs
(Transfer, Search, Identifiers, ...)

Data Automation

File Sharing

File Transfer, Sync

Globus Auth
(identity and access management)
Globus Auth addresses security challenges

• Make it easy for developers to provide login for their apps (web, mobile, desktop, command line)

• …and protect all REST API communications
  o App → Globus service (MRDP, Jupyter Notebook)
  o App → non-Globus service (graph service in MRDP)
  o Service → Service

• …while
  – Not introducing yet another identity
  – Providing a platform to consolidate existing identities
  – Providing a least privileges security model (via consents)
  – Being web friendly and language/framework agnostic
Based on widely used web standards

- OAuth 2.0 Authorization Framework (a.k.a. OAuth2)
- OpenID Connect Core 1.0 (a.k.a. OIDC)
- Access via OAuth2 and OIDC libraries of your choice
  - Google OAuth Client Libraries, Apache mod_auth_openidc, etc.
  - Globus Python SDK

[docs.globus.org/api/auth](docs.globus.org/api/auth)
Auth Example: Authorization Code Grant

1. Access portal

2. Redirects 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) to give the client the authority invoke the transfer service

Client (Web Portal, Application, Jupyter)

Browser (User)

Globus Auth (Authorization Server)

Globus Transfer (Resource Server)

Identity Provider
Globus Transfer API

- Globus Web App consumes public Transfer API
- Resource named by URL (standard REST approach)
  - Query params allow refinement (e.g., subset of fields)
- Globus APIs use JSON for documents and resource representations
- Requests authorized via OAuth2 access token
  - Authorization: Bearer asdflkqhafdsafeawk

[docs.globus.org/api/transfer](docs.globus.org/api/transfer)
Globus Python SDK

- Python client library for the Globus Auth and Transfer REST APIs
- `globus_sdk.TransferClient` class handles connection management, security, framing, marshaling

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

[globus.github.io/globus-sdk-python](globus.github.io/globus-sdk-python)
Globus Helper Pages

- Globus pages designed for use by your web apps
  - Browse Endpoint
  - Activate Endpoint
  - Select Group
  - Manage Identities
  - Manage Consents
  - Logout

docs.globus.org/api/helper-pages
Globus PaaS developer resources

Python SDK

Documentation:
docs.globus.org/api

github.com/globus

Example:

```
from __future__ import print_function

# Jupyter Notebook

# Configuration
First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you can use the one from your Globus CLI and
```

docs.globus.org/api
github.com/globus
Globus Auth: Native apps

• **Client that cannot keep a secret, e.g…**
  – Command line, desktop apps
  – Mobile apps
  – Jupyter notebooks

• **Native app is registered with Globus Auth**
  – Not a confidential client

• **Native App Grant is used**
  – Variation on the Authorization Code Grant

• **Globus SDK:**
  – To get tokens: NativeAppAuthClient
  – To use tokens: AccessTokenAuthorizer
Native App grant

1. Run application
2. URL to authenticate
3. Authenticate and consent
4. Auth code
5. Register auth code
6. Exchange code
7. Access tokens
8. Authenticate with access tokens to invoke transfer service as user
Refresh tokens

- **Common use cases**
  - Portal checking transfer status when user is not logged in
  - Running command line app from script

- **Refresh tokens issued to client, in particular scope**

- **Client uses refresh token to get access token**
  - Confidential client: client_id and client_secret required
  - Native app: client_secret not required

- **Refresh token good for 6 months after last use**

- **Consent rescindment revokes all tokens**
Refresh tokens

1. Run application
2. URL to authenticate
3. Authenticate and consent
4. Auth code
5. Register auth code
6. Exchange code, request refresh tokens
7. Access tokens and refresh tokens
8. Store refresh tokens
9. Exchange refresh token for new access tokens
10. Access tokens
11. Authenticate with access tokens to invoke service as user

Browser

Globus Auth (Authorization Server)

Native App (Client)

App/Service (Resource Server)
Native App/Refresh Tokens Sample Code

github.com/globus/native-app-examples

• ./example_copy_paste.py
  – User copies and pastes code to the app

• ./example_copy_paste_refresh_token.py
  – Stores refresh token locally, uses it to get new access tokens

• See README for installation
Automating data flows at scale
Instrument Use Cases

• Advanced Photon Source
  – Connectomics
  – Time series spectroscopy

• Scanning Electron Microscope
  – Materials science

• Cryo-electron Microscope
UChicago Kasthuri Lab: Brain aging and disease

- Construct connectomes—mapping of neuron connections
- Use APS synchrotron to rapidly image brains
  - Beam time available once every few months
  - ~20GB/minute for large (cm) unsectioned brains
- Generate segmented datasets/visualizations for the community
- Perform semi-standard reconstruction on all data across HPC resources
Building the connectome

1. Imaging
2. Acquisition
3. Pre-processing
4. Preview/Center
5. User validation
6. Reconstruction
7. Publication
8. Visualization
9. Science!

APS

Argonne JLSE

User validation

Neuroanatomy reconstruction pipeline
Data Example

ALCF Data Discovery Portal

https://petreldata.net
Prototypical Globus Platform example:

Modern Research Data Portal

docs.globus.org/mrdp
Modern Research Data Portal Design Pattern

- **Science DMZ**
- **Firewall**
- **Researcher’s System**
  - **Identity Provider**
  - **Globus Web Helper Pages**
  - **Globus Auth**
  - **Portal Web Server (Client)**
- **Other Services**
- **HTTPS**
- **REST**
- **Other Endpoints**
- **GridFTP**
- **User Endpoint (optional)**
- **Portal Endpoint (DTNs)**
- **Globus PaaS**
- **Globus Transfer**

**Applications**
- **Browser**
- **Login**
- **Portal Web Server (Client)**
- **Portal Web Server (optional)**
- **Globus Auth**
- **Globus Transfer**
- **Globus Web Helper Pages**
- **Other Services**
- **GridFTP**
- **Other Endpoints**

**Science DMZ**
- **User Endpoint (optional)**
- **Portal Endpoint (DTNs)**
- **Globus PaaS**
- **Globus Transfer**
- **Globus Auth**
- **Globus Web Helper Pages**
- **Identity Provider**
- **HTTPS**
- **REST**
- **Other Services**
- **GridFTP**
- **Other Endpoints**

**Portal Web Server (Client)**
- **Globus Auth**
- **Globus Transfer**
- **Globus Web Helper Pages**
- **Identity Provider**
- **HTTPS**
- **REST**
- **Other Services**
- **GridFTP**
- **Other Endpoints**

**User Endpoint (optional)**
- **Portal Endpoint (DTNs)**
- **Globus PaaS**
- **Globus Transfer**
- **Globus Auth**
- **Globus Web Helper Pages**
- **Identity Provider**
- **HTTPS**
- **REST**
- **Other Services**
- **GridFTP**
- **Other Endpoints**

**Portal Endpoint (DTNs)**
- **Globus PaaS**
- **Globus Transfer**
- **Globus Auth**
- **Globus Web Helper Pages**
- **Identity Provider**
- **HTTPS**
- **REST**
- **Other Services**
- **GridFTP**
- **Other Endpoints**

**Globus PaaS**
- **Globus Transfer**
- **Globus Auth**
- **Globus Web Helper Pages**
- **Identity Provider**
- **HTTPS**
- **REST**
- **Other Services**
- **GridFTP**
- **Other Endpoints**
Legacy research data portal architecture

Portal server applications:
- web server
- search
- database
- authentication
- data service
MRDP Architecture

Data Path

Portal Query/Browse Path

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

API DTNs
(data access governed by portal)
Relevant data sharing elements

- One-time, manual creation of shared endpoint
- Permissions set per folder on shared endpoint
- Permissions management can be automated
  - User: researcher@uchicago.edu
  - Group: search for group to get Group UUID
  - Application …yes, apps are people too!
- Roles for management of endpoint and tasks
  - Grant rights to other users, groups or applications
- Access manager role for managing permissions
**Security elements: Confidential app**

- Uses client id and secret
- Ensure application is on a secure device
- Set up policy for rotation of secret (limited life tokens)
- Identity: `<app_client_id>@clients.auth.globus.org`
Client credential grant

1. Authenticate with app client id and secret

Web Application, Science Gateway, Data Portal (Client)

2. Access Tokens

Globus Auth (Authorization Server)

3. Authenticate as app with access tokens to invoke service (on behalf of authorized user, within a given scope)

Globus Transfer (Resource Server)
Confidential clients must be registered

- Register at developers.globus.org
  - Redirects; e.g. https://your_fqdn/app/authcallback
  - Scopes
    - e.g. globus:auth:scope:transfer.api.globus.org:all
    - e.g. profile, email, openid

- Get client id and secret
- Ensure secret is properly protected, rotated
Support resources

• **Globus documentation:** [docs.globus.org](https://docs.globus.org)
• **Sample code:** [github.com/globus](https://github.com/globus)
• **Helpdesk and issue escalation:** [support@globus.org](mailto:support@globus.org)
• **Customer engagement team**
• **Globus professional services team**
  – Assist with portal/gateway/app architecture and design
  – Develop custom applications that leverage the Globus platform
  – Advise on customized deployment and integration scenarios
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