Managing Research Data with Globus Software-as-a-Service

ACI-REF Virtual Residency 2017
Vas Vasiliadis
vas@uchicago.edu
Research data management today

How do we...
...move?
...share?
...discover?
...reproduce?
Globus delivers...
(Big) 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 research computing/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

Amazon Glacier

Raw Source Data
Bridge to collaborators

- XSEDE
- Jetstream
- NERSC
- EC2
- Cornell University
- University of Michigan

External Campus Storage

Public/Private Cloud stores

- Red Cloud
- Ceph
- OpenStack
- Google Drive
- Amazon Web Services
- S3
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.

- Use a Web browser
- Access any storage
- Use an existing identity
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**
Thank you to our users!

- 290 PB transferred
- 50 billion tasks processed
- 62,000 registered users
- 350+ federated identities
- 10,000 active endpoints
- 350+ 100TB+ users
- 1 PB largest single transfer to date
- 10,000 active users
- 5,119 active shared endpoints
- 99.5% uptime
- 3 months longest running managed transfer
- 48 most server endpoints on one campus
- 1 PB largest single transfer to date
Demonstration
File Transfer
How can I use Globus on my computer?
...makes your storage system a Globus endpoint
Globus Connect Personal

- Installers do not require admin access
- Zero configuration; auto updating
- Handles NATs
Demonstration
File Sharing
Federated Identity
Data Publication and Discovery

https://publish.globus.org
Demonstration

Data Publication
How do I do that on my research storage system(s)?
Globus Connect Server

- Create endpoint on practically any filesystem
- Enable access for all users with local accounts
- Native packages: RPMs and DEBs
Demonstration

1. Install Globus Connect Server
   - Access server as root-capable user
   - Update repo
   - Install packages
   - Setup Globus Connect Server

2. Log into Globus

3. Access the newly created endpoint

4. Transfer a file
Installing Globus Connect Server

$ sudo su
$ curl -LOs http://toolkit.globus.org/ftppub/globus-connect-server/globus-connect-server-repo_latest_all.deb
$ dpkg -i globus-connect-server-repo_latest_all.deb
$ apt-get update
$ apt-get -y install globus-connect-server
$ globus-connect-server-setup
$ _

You have a working Globus endpoint!
Common configuration options

- Manage Endpoints page
  - Display Name
  - Visibility
  - Encryption

- DTN local config file (.ini format)
  - /etc/globus-connect-server.conf
  - RestrictPaths
  - IdentityMethod (CILogon, OAuth)
  - Sharing
  - SharingRestrictPaths
Storage connectors

• **Standard storage connectors (POSIX)**
  – Linux, Windows, MacOS
  – Lustre, GPFS, OrangeFS, etc.

• **Premium storage connectors**

  AWS S3
  Ceph RadosGW (S3 API)
  Spectra Logic BlackPearl
  HPSS

  Google Drive (beta)
  Box (in progress)
  HDFS (in progress)
  iRODS (in progress)
  HGST Active Archive (in progress)

[docs.globus.org/premium-storage-connectors](https://docs.globus.org/premium-storage-connectors)
Deployment
Best Practices
Data Transfer Node in a Science DMZ

Data Transfer Node (DTN)

* Please see TCP ports reference: https://docs.globus.org/resource-provider-guide/#open-tcp-ports_section
Endpoint activation using MyProxy

1. Access endpoint
2. Username and password
3. Campus username and password
4. TLS handshake
5. Transfer request
6. Certificate
7. Authorization (resolve local user)
8. Access files
9. Control channel authorization
10. Data transfer

DON’T LEAVE IT LIKE THIS!

Remote cluster with Globus Connect Server or laptop/PC with Globus Connect Personal
Endpoint activation using MyProxy OAuth

1. Access endpoint
2. OAuth redirect
3. Campus username password
4. OAuth Server
5. Username password
6. MyProxy Online CA
7. Certificate
8. MyProxy Server
9. Username password
10. Authorization (resolve local user)
11. Globus transfer and sharing hosted service
12. GridFTP Server

Remote cluster with Globus Connect Server or laptop/PC with Globus Connect Personal

Yes, please do this!
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

Web

CLI

Globus service

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

- Transfer and Auth
- Uses Python SDK
- Open source

[GitHub Repository](github.com/globus/globus-cli)
[Docs](docs.globus.org/cli)
Demonstration
Command Line Interface
Globus as PaaS

Integrate file transfer and sharing capabilities into scientific web apps, portals, gateways, etc.

Use existing institutional ID systems in external web applications
Simple Automation

Recurring transfers with sync option

Copy /ingest
Daily @ 3:30am

Data distribution

Staging area cleanup

1. Check if successful transfer
2. Delete data from staging area
Data Distribution; APS - DMagic

DMAGIC
a Globus implementation at the APS

 Courtesy of Francesco De Carlo, Argonne National Laboratory (2016)

dmagic.readthedocs.io
Analysis App: Wellcome Sanger

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 your data, please fill in the details below to register an imputation and/or phasing job. If you need more information, see the about page.

- Full name
- Organisation
- Email address

What is this
Globus user identity

Next

News
@sangerimpute

11/05/2016
Thanks to EAGLE, we can now return phased data. The HRC panel has been updated to r1.1 to fix a known issue. See ChangeLog for more details.

15/02/2016
Globus API changed, please see updated instructions.

17/12/2015
New status page and reworked internals. See ChangeLog.

09/11/2015
Pipeline updated to add some features requested by users. See ChangeLog.

See older news...
National Resource Access

Jetstream Web App would like to:

- Access all Jetstream resources

By clicking “Allow”, you allow Jetstream Web App to access information and services. You can rescind this permission at any time.

Compute Canada has partnered with Globus to offer this high performance file transfer service.

Calcium Canada s’est associé à Globus pour vous offrir ce service de transfert de fichier à haute performance.

Log in to use Compute Canada Globus Web App

Use your existing organizational login

e.g. university, national lab, facility, project, Google or Globus ID

(Your Globus username and password used prior to February 13, 2016 is now Globus ID)

WestGrid

Continue

Didn’t find your organization? Then use Globus ID to sign up.
Identity Management

The new Systems Biology Knowledgebase (KBase) is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It will be a community-driven, extensible and scalable open source software framework and application system. KBase will offer free and open access to data, models and simulations, enabling scientists and researchers to build, new knowledge and share their findings.

What can KBase do?
- Combine heterogenous data types
- Offer standardized access to bioinformatic and modeling analyses
- Use evidence-supported annotations of genome structure and genetic function
- Discover new associations and network structures in community and molecular networks
- Map genotype to complex organismal traits
- Design and refine experiments using models of metabolism, regulation and community function
- Enable sharing of data, hypotheses, and newly-generated knowledge

Accurate research on campus by providing connective services
for local, cloud and national cyberinfrastructure

Connecting Science to Cycles and Data
CI Connect offers campus IT organizations connective services for cyberinfrastructure. End-users quickly gain access to shared campus clusters while gaining access to distributed resources of national HPC or HTC facilities (such as the Open Science Grid) or cloud-based resources. CI Connect offers several capabilities based on Globus, HTCondor and other HTC and data sharing technologies that augment your vision of providing scientific computational resources to research communities you support.

Reliable, High Performance CI Services
CI Connect provides services to CI providers at US universities and laboratories following a simple set of principles:

- 2013-02-26
- Proposed: Genomic Science Contractors-Grantees Meeting

Connected environments from hosted services
Resources of a campus cluster (or campus grid) can be
Globus PaaS developer resources

Python SDK

Requirements
- You need to be in the tutorial users group for sharing: https://www.globus.org/app/groups/56b6a29c-63e5-46d4-8d04-11e5-ba46-22000b92c6ec
- Installed Globus Python SDK

Jupyter Notebook

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

```python
In [15]: from __future__ import print_  
tutorial_endpoint_1 = "dd59a5f0-d9b3-42e8-9f97-11e5-ba46-22000b92c6ec"  
tutorial_endpoint_2 = "dd59a5f0-6d04-11e5-ba46-22000b92c6ec"  
tutorial_users_group = "50b6a29c-63e5-46d4-8d04-11e5-ba46-22000b92c6ec"

In [16]: transfer_token = None # if None, tries to get token from ~/.globus.cfg file
```
...on sustainability
Thank you to our sponsors...
...and THANK YOU, subscribers!
Globus sustainability model

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

- **Branded Web Site**

- **Premium Storage Connectors**

- **Alternate Identity Provider (InCommon is standard)**

*Coming soon*
Demonstration Management Console
Facilitating adoption

- GlobusWorld Tour: 7 stops, over 250 attendees
- Lots of open source developer materials
- Contact us to host event

September 12-13
Future Directions

• New data management interface
• Collections: the evolution of endpoints
• HTTPS access to storage
• Enhanced sharing capabilities
• Metadata and search
• Data Publication PaaS

$= 1,500+$ IdPs
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