Globus for Research Data Management

Presented to
ATPESC 2017 Participants

Greg Nawrocki
University of Chicago - Globus
Q Center, St. Charles, IL (USA)
Date 08/04/2017
Best-practice deployment

- **Border Router**
  - 10GE connections to WAN and Science DMZ Switch/Router
  - Clean, High-bandwidth WAN path
  - Site / Campus access to Science DMZ resources

- **Science DMZ Switch/Router**
  - 10GE connections to Border Router and Enterprise Border Router/Firewall
  - Per-service security policy control points

- **Enterprise Border Router/Firewall**
  - 10GE connection to Science DMZ Switch/Router

- **WAN**
  - 10GE connection to Border Router

- **High performance Data Transfer Node with high-speed storage**
  - 10GE connection to Border Router

Details at: fasterdata.es.net
Best-practice deployment

- **Border Router**
  - Clean, High-bandwidth WAN path
  - Site / Campus access to Science DMZ resources
  - Per-service security policy control points

- **Enterprise Border Router/Firewall**

- **Science DMZ Switch/Router**

- **WAN**

- **High performance Data Transfer Node with high-speed storage**

Details at: fastedata.es.net
Research data management today

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

Index?
Globus delivers...
Fast and reliable big data transfer, sharing, publication, and discovery...
...directly from your own storage systems...
...via software-as-a-service using existing identities.
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...
Bridge to Instruments

Pre-processed Data

Analysis store

High durability, low cost store

Amazon Glacier

Raw Source Data
Bridge to collaborators

External Campus Storage

Public/Private Cloud stores

- XSEDE
- Jetstream
- NERSC

- EC2

- Globus

- Cornell University
  - Founded A.D. 1865
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**
Harness the power of the Globus research data management cloud.

Transfer API
The Transfer API provides a REST-style interface to the Globus reliable file transfer service. The API can be used to monitor the progress of file transfers, manage file transfer endpoints, list remote directories, and submit new transfer and delete tasks.

Resource Providers
Globus allows you, as a resource provider, to easily offer reliable, secure, high-performance research data management capabilities to your users and their collaborators, directly from your own storage infrastructure.

Toolkit
The open source Globus Toolkit is a fundamental enabling technology for the "Grid," allowing users to access high-performance computing resources securely across corporate, institutional, and geographic boundaries without sacrificing local autonomy.
Endpoints

• **Storage abstraction**
  – All transfers happen between endpoints
  – Globus Connect Server instantiates endpoints
  – [https://docs.globus.org/faq/globus-connect-endpoints/](https://docs.globus.org/faq/globus-connect-endpoints/)

• **Test / Demo Endpoints**
  – Globus Tutorial Endpoint 1
  – Globus Tutorial Endpoint 2
  – ESnet Test Endpoints
    o Read Only & Read / Write
    o Some contain file samples of various sizes

• **Globus Connect Personal**
  – Now your laptop is an endpoint
  – [https://www.globus.org/globus-connect-personal](https://www.globus.org/globus-connect-personal)
Globus Connect Personal (GCP)

- Installers do not require admin access
- Zero configuration; auto updating
- Handles NATs
• **Greg at Globus**  
  – Globus ID: nawrocki@globusid.org  
  – Email: greg@globus.org

• **Greg at University of Chicago**  
  – CILogon: nawrocki  
  – Email: nawrocki@uchicago.edu

• **Greg at home**  
  – Globus ID: nawrockipersonal@globusid.org  
  – Email: greg@nawrockinet.com
Globus SaaS Demo Identities

- **Greg at Globus**
  - Globus ID: nawrocki@globusid.org
  - Email: greg@globus.org

- **Greg at University of Chicago**
  - CILogon: nawrocki
  - Email: nawrocki@uchicago.edu

- **Greg at home**
  - Globus ID: nawrockipersonal@globusid.org
  - Email: greg@nawrockinet.com

Linked Identities

Primary Identity
Research data management simplified.

TRANSFER  SHARE  PUBLISH  BUILD

Get unified access to your research data, across all systems, using any existing identity.

Data stored at a different institution? At a
Use(r)-appropriate interfaces

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
...
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.
Command Line Interface

- Transfer and Auth
- Replaces old SSH-based command line shell
- Uses Python SDK
- Open source
  github.com/globus/globus-cli  
  docs.globus.org/cli
Enable existing institutional ID systems to be used in external web applications.

Integrate file transfer and sharing capabilities into scientific web apps, portals, gateways, etc.
## Data App: NCAR RDA

### NCEP Climate Forecast System Version 2 (CFSv2) Monthly Products

**ds094.2**

For assistance, contact [Bob Dattore](mailto: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>
<th>NCAR-Only Access</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Web Server Holdings</td>
<td>Subsetting</td>
<td>TDS Access</td>
<td>GLADE File Listing</td>
</tr>
<tr>
<td>Union of Available Products</td>
<td>Web File Listing</td>
<td>Request Globus Invitation</td>
<td></td>
<td>GLADE File Listing</td>
</tr>
<tr>
<td>Diurnal monthly means</td>
<td>Web File Listing</td>
<td>Get a Subset</td>
<td></td>
<td>GLADE File Listing</td>
</tr>
<tr>
<td>Regular monthly means</td>
<td>Web File Listing</td>
<td>Get a Subset</td>
<td></td>
<td>GLADE File Listing</td>
</tr>
</tbody>
</table>

**Globus Transfer Service (GridFTP)**

Home | Find Data | Ancillary Services | About/Contact | Data Citation | Web Services | For Staff
Storage connectors

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

• **Premium storage connectors**
  
<table>
<thead>
<tr>
<th>AWS S3</th>
<th>HPSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceph RadosGW (S3 API)</td>
<td>HDFS (in progress)</td>
</tr>
<tr>
<td>Spectra Logic BlackPearl</td>
<td>iRODS (in progress)</td>
</tr>
<tr>
<td>Google Drive</td>
<td>HGST Active Archive (in progress)</td>
</tr>
</tbody>
</table>

[docs.globus.org/premium-storage-connectors](docs.globus.org/premium-storage-connectors)
Globus Connect Server

- Runs on Linux
  - CentOS 5, 6, and 7
  - Debian 7 and 8
  - Fedora 23 and 24
  - Red Hat Enterprise Linux 5, 6, and 7
  - Scientific Linux 5, 6, and 7
  - SuSE Linux Enterprise Server 11sp3
  - Ubuntu 12.04 LTS, 14.04 LTS, 15.10, and 16.04 LTS

https://docs.globus.org/globus-connect-server-installation-guide/
Performance and Reliability

- Multiple DTNs per endpoint
- Network Use Tuning
  - Concurrency
  - Parallelism
- Network Use Options
  - Minimal
  - Normal
  - Aggressive
  - Custom

https://docs.globus.org/globus-connect-server-installation-guide/#setting_endpoint_network_use_options
Illustrative performance

- 20x scp throughput (typical)
  - >100x demonstrated

- On par/faster than UDP based tools (NASA JPL study and anecdotal)

- Capable of saturating “any” WAN link
  - Demonstrated 85Gbps sustained disk-to-disk
  - Typically require throttling for QoS
**Disk-to-Disk Throughput**

- **GridFTP (4 streams)**
- **GridFTP (1 stream)**
- **sftp**
- **scp (w/HPN)**
- **scp**

**Bar Chart**

- Disk-to-Disk Throughput (Mbps)

**Key Points**

- Berkeley, CA to Argonne, IL (RTT: 53 ms, Capacity: 10Gbps)
- scp is 24x slower than GridFTP on this path
- >1 Gbps (125 MB/s) disk-to-disk requires RAID array

Source: ESnet (2016)
Thank you to our sponsors
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**
  - Amazon S3, Ceph, HPSS, Spectra, Google Drive, Box*, HDFS*

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

*Coming soon
Globus by the numbers

48
most server endpoints on one campus

280 PB
transferred

350
100TB+ users

10,000
active users

10,000
active endpoints

300+
federated identities

1 PB
largest single transfer to date

5,119
active shared endpoints

60,000
registered users

47 billion
tasks processed

3 months
longest running managed transfer

99.5%
uptime
THANK YOU, subscribers!
Join the Globus Community

- **Documentation**
  - docs.globus.org

- **Join the mailing lists**
  - globus.org/mailing-lists

- **Lots of good open source examples**
  - github.com/globus/
  - github.com/globus/globus-sdk-python
  - Discussions on developer-discuss@globus.org

- **When all else fails**
  - https://www.globus.org/contact-us
Globus Admin

- **Globus Connect Server (GCS) Installation**
  - [https://docs.globus.org/globus-connect-server-installation-guide/](https://docs.globus.org/globus-connect-server-installation-guide/)

- **Globus Connect Server Installation on the EC2 Tutorial Server**
  - [https://www.globusworld.org/tutorials](https://www.globusworld.org/tutorials)
  - You’ll need your own EC2 instance
  - When we do the tour we supply temporary instances

- **Helpful slides**

- **Configuration options**
  - `/etc/globus-connect-server.conf`
Automation Examples

• Syncing a directory
  – Bash script that calls the Globus CLI and a Python module that can be run as a script or imported as a module.

• Staging data in a shared directory
  – Bash / Python

• Removing directories after files are transferred
  – Python script

• Simple code examples for various use cases using Globus
  – https://github.com/globus/automation-examples
Globus Transfer API Set

- **Helpful slides**
  - Both transfer and auth covered

- **Doc**
  - [https://docs.globus.org/api/transfer/](https://docs.globus.org/api/transfer/)

- **Sample data portal**

- **Jupyter notebook**
  - [https://github.com/globus/globus-jupyter-notebooks](https://github.com/globus/globus-jupyter-notebooks)
Globus Auth API Set

• **Helpful slides**  
  – Both transfer and auth covered

• **Doc**  
  – [https://docs.globus.org/api/auth/](https://docs.globus.org/api/auth/)

• **Sample data portal**  

• **Native app examples**  
  – [https://github.com/globus/native-app-examples](https://github.com/globus/native-app-examples)
Globus on your Campus

• Webinars
• Programs
  – Helping you evangelize Globus within your institution.
• Professional Services
• Globus World Tour
  – Taking the show on the road.