Software-as-a-Service for Research Data Management

Steve Tuecke
Deputy Director, Computation Institute
University of Chicago & Argonne National Laboratory
Big Science built on Globus Toolkit

Earth System Grid

Cancer Biology Informatics Grid

LIGO data grid
Thinking about “small and medium labs”

- Big projects like LHC, LIGO, ESG, etc., can run resource-level services reliably—and build and operate effective collective services
- Small labs and collaborations have problems with both
- They need solutions, not toolkits—ideally outsourced solutions

Can we harness the power of the cloud to scale access to the grid?
Time-consuming tasks in science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature

- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
- …
• Most research performed in small laboratories
• Researchers are trained in their field, not in IT
  – They are not experts in collecting, moving, storing, indexing, analyzing, mining, sharing, updating, publishing, and archiving massive amounts of data
• Only limited capital is available for them to spend on data and IT support
• Investment is spent on traditional research tools (e.g., microscopes)—but the world is changing
  – Now need substantial and sophisticated IT to perform research, data manipulation, data mining, collaboration
Globus Toolkit
Build the Grid
Components for building custom grid solutions
globustoolkit.org

Globus Online
Use the Grid
Reliable file transfer
Software-as-a-Service
globusonline.org
Time-consuming tasks in science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature
- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
- …
The Challenge: Moving Big Data

• What should be trivial…

“I need my data over there – at my _____” (supercomputing center, campus server, etc.)

• … can be painfully tedious and time-consuming

“GAAAAH! %&@#&”

“Config issues”

“Firewall issues”

“Unexpected failure = manual retry”

Data Source ——— Data Destination
What is Globus Online?

• **Reliable file transfer.**
  – Easy “fire and forget” file transfers
  – Automatic fault recovery
  – High performance
  – Across multiple security domains

• **No IT required.**
  – Software as a Service (SaaS)
  – No client software installation
  – New features automatically available
  – Consolidated support and troubleshooting
  – Works with existing GridFTP servers
  – Globus Connect solves “last mile problem”

“I moved 400 GB of files and didn’t even have to think about it.”
– Lawrence Berkeley National Lab

“It’s just not a big deal to move big data anymore.”
– Initiative for Biomedical Informatics

“Fantastic! I have started using globus connect to transfer data, and it only took me 5 minutes to set up. Thank you!”
– NERSC user
Case Study: Lattice QCD

- **Fast**: Reduced transfer times
- **Easy**: Fire-and-forget transfers
  - Automated retry
  - No file pre-staging
  - No complex infrastructure
  - Convenient CLI or GUI interfaces

“Globus Online frees up my time to do more creative work than typing `scp` commands or devising scripts to initiate and monitor progress to move many files.”

Indiana University researcher moved ~6 TB from Oak Ridge to TACC in 2 days

“I moved 100 7.3 GB files tonight in about 1.5 hours. I am very impressed. I also like the new commands and help system.”

“Globus Online frees up my time to do more creative work than typing `scp` commands or devising scripts to initiate and monitor progress to move many files.”
Case Study: Enabling Users @ NERSC

• **Challenge:** Providing easy web-based access to data and compute resources
  – “We need to provide web-based ways to accomplish computing tasks – it’s what our scientists expect, and it’s also what will make them more productive.”

• **Solution:** GO enabled for all NERSC users
  – GO endpoints maintained by NERSC
  – GO is recommended method of transfer for NERSC users

• **Benefits for NERSC users**
  – Drag and drop archiving / syncing
  – Reliable, secure and “amazingly easy to use”
  – Researchers can focus on their work, not on IT

• **Benefits for NERSC**
  – Operations and support outsourced to GO
  – Fast and easy to make endpoints available
  – Automated authentication
  – Reliable performance and support

“Fantastic! I have already started using Globus Connect to transfer data, and it only took me 5 minutes to set up. Thank you!” – NERSC user

Hopper, Franklin and HPSS are among the NERSC resources leveraged by Globus Online.

www.globusonline.org
How It Works

1. User initiates transfer request

2. Globus Online moves files

3. Globus Online notifies user
Globus Online Highlights

Web interface

Command line interface

```
ls alcf#dtn:~
scp alcf#dtn:~/myfile
nersc#dtn:~/myfile
```

HTTP REST interface

```
POST https://transfer.api.globusonline.org/ v0.10/
transfer <transfer-doc>
```

- Fire-and-forget data movement
- Many files and lots of data
- Third-party transfers
- Performance optimization
- Across multiple security domains
- Expert operations and support

GridFTP servers
FTP servers

High-performance data transfer nodes

Globus Connect on local computers
Logging into the CLI

• Interactive login to command line interface:
  
  
  $ ssh tuecke11@cli.globusonline.org
  
  • Running commands remotely:
  
  
  $ ssh tuecke11@cli.globusonline.org <command>
  
  $ ssh tuecke11@cli.globusonline.org scp -r -s 3 -D olcf#/~/myfile* mylaptop:/~/projects/p1
  
  Task ID: 4a3c471e-edef-11df-aa30-1231350018b1
  
  $ _
  
  • Using CLI with gsssh:
  
  
  $ gsssh tuecke11@cli.globusonline.org <command>
Globus Connect to/from your laptop

Step One: Choose Your Download
- Globus Connect for Mac OS X
- Globus Connect for Linux
- Globus Connect for Windows

Step Two: Get Your Globus Connect Setup Key
- Endpoint Name: rachanaLaptop
- Description: Argonne Laptop

Step Three: Finish Globus Connect Setup
Copy the setup key displayed above. Run Globus Connect and paste the key into the Initial Setup window when prompted. This setup key can only be used once.

Initial Setup
Please type or paste your Globus Connect setup key into the field below and click 'OK' when finished.

Setup Key: 432e8ba5-45cf-442b-a374-5a8d1cfa75cb

www.globusonline.org
What’s next?

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature

- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports

...
Our goal: To accelerate the pace of discovery and innovation worldwide, by providing millions of researchers with unprecedented access to powerful research tools via software as a service (SaaS).

“Civilization advances by extending the number of important operations which we can perform without thinking of them”

Alfred North Whitehead, 1911
For More Information

• Visit [https://www.globusonline.org/signup](https://www.globusonline.org/signup) to:
  – Get a free account and start moving files

• Visit [www.globusonline.org](www.globusonline.org) for:
  – Tutorials
  – FAQs
  – Pro Tips
  – Troubleshooting

• Contact [support@globusonline.org](mailto:support@globusonline.org) for:
  – Help getting started
  – Help using the service
Reliable, high-performance, secure file transfer

Move files fast. No IT required.

Globus Online makes robust file transfer capabilities, traditionally available only on expensive, special-purpose software systems, accessible to everyone.

Why Use Globus Online?
See how easy file transfer can

For HPC Resource Owners
Enable Globus Online

For Developers
Integrate with Globus Online