GLOBUS OVERVIEW

Lev Gorenstein, Senior Computational Scientist
Globus Overview

Outline
What to Expect From This Talk

Objectives

- What is Globus – history, capabilities, strengths
- What Globus is not
- Globus terminology and concepts
- Use cases – data transfers and sharing
- Pitfalls
- Demo
Globus Overview

Storage and Data Management
There is more to Research Computing than just computing!

- We are best known for our supercomputing clusters – but if it wasn’t for storage and other cyberinfrastructure, where would you put all those nice things you’ve just calculated?

- See [www.rcac.purdue.edu/storage](http://www.rcac.purdue.edu/storage) for all our storage options

- An interactive storage solutions finder: [www.rcac.purdue.edu/storage/solutions/](http://www.rcac.purdue.edu/storage/solutions/)

- Also check out [www.rcac.purdue.edu/services](http://www.rcac.purdue.edu/services) for other services we provide
There is more to Research Storage than just storage!

**Locations**
- Home directory
- Cluster scratch
- Data Depot
- Fortress
- Lab instrument
- Office workstation
- Laptop
- Cloud services
- PURR

**Actions**
- Generate
- Process/analyze
- Transfer
- Share
- Publish
Globus Overview
Non-profit service for secure, reliable research data management

- A platform and a service for moving, sharing and discovering data via a single interface
- A team at the University of Chicago and Argonne National Laboratory
- Funded by NSF, DoE, NIH and institutional subscriptions (freemium model)
  - Purdue is subscribed, RCAC pays
What is Globus?

History

- Stems from GridFTP and high energy physics community
- Started as a pure transfer tool with two strengths:
  - Fast transfers over good networks
  - Robust transfers over flaky networks
- Added functionality:
  - Data sharing and flexible access control
  - Identity management
  - Web GUI, scriptable command line tool, and powerful API with a Python SDK
  - Cross-platform
  - Great support
Globus is not your typical network drive!

What happens when you double-click on a Word document on a network drive?

- A copy of the document is transparently downloaded by the system
- You edit local temporary copy in Word
- A saved version is transparently uploaded back to the network drive at the end

What happens when you double-click on a Word document in Globus?

- **Nothing. It's a transfer tool!**
- The "download", "edit", "upload" steps are fully decoupled and have to be done **explicitly** by you

- On some modern endpoints, you may be taken to the browser's "Open/download" dialog, but still no automatic back and forth
- Negishi and Anvil have this feature, more to come
Globus Data Transfers

Globus transfers overview

- Secure unified interface to your data
- “Fire and forget” (Globus monitors the transfer, auto-resumes on errors, sends an email at the end)

- **Note:** the data channel is directly between A and B

- Your computer is only used for the command channel (dispatch a terabyte transfer using your phone!)
  - If the transfer is not from your computer, your computer does not have to stay on

https://www.globus.org/data-transfer
Globus sharing overview

- Easy (all you need is recipient’s email)
- Secure
- Flexible access control (user(s), groups, world, read, write)
  - Premium feature – subscription benefit
- No more “Hey, I uploaded a terabyte to Google Drive, what’s your Gmail?”

**Note: User B does not need to have an account on your storage system!**

- Approved for HIPAA data, too.
  - Purdue does not have a HIPAA-compliant endpoint yet, but we will!

https://www.globus.org/data-sharing
Vocabulary: Collections, Shares and Endpoints

“A named location containing data you can access with Globus”

Historic terminology:

- “Endpoints” (or “primary endpoints”) – the main location itself (e.g. “Purdue Data Depot”).
- “Shares” (or “shared endpoints”) – parts of the primary endpoint that have been given their own names and shared via Globus (e.g. “My subfolder for User B”).
- “A share off of a primary endpoint”

Globus recently adopted new terminology:

- “Endpoint” refers to hardware/software/system component (what admins deal with)
- “Collections” refers to the named location components (what users deal with)
- “Mapped collections” – where a Globus identity is mapped to a local user (== old “primary endpoints”)
- “Guest collections” – parts of the mapped collection that have been given their own names and shared via Globus (== old “shares”)

User adoption is slow, people often still use “Endpoints” in the sense of “Collections”
Vocabulary: Globus Account and Identities

“*You and the hats you wear*”

- Globus needs a handle to know you by (and to authenticate you) – a Globus account.
- In the simplest form, this is your organizational login, but there are many more Identity Providers that Globus recognizes (e.g. Gmail, ORCiD, ACCESS ID, etc).
  - Purdue is recognized – “Purdue University Main Campus”

- When you first login to Globus, your Globus account will be established. You will be asked to chose your Organization (a.k.a. Identity Provider).
  - If it is one of the 1400+ Globus recognizes, it’ll send you to the organizational login page (like BoilerKey)
  - Otherwise, Globus can serve as its own identity provider (the Globus ID)

**Note:** anyone can use Globus! You do not have to be in one of the recognized organizations!
Vocabulary: Globus Account and Identities

"You and the hats you wear"

- You have a Purdue career account, a Gmail, another university account, an ORCiD, an ACCESS account, etc., but this is still the same you.

- A Globus account is a set of linked identities that you have used to login to Globus.
  - You don't have to link them, but it is handy.
  - [https://app.globus.org/account/](https://app.globus.org/account/) or [https://transferrcac.purdue.edu/account/](https://transferrcac.purdue.edu/account/)
Globus Overview

Globus Login and Demo
Login to Globus

transfer.rcac.purdue.edu or globus.org

- Purdue people: select “Purdue University Main Campus” as Organization. Will be taken to the BoilerKey 2FA page.

- Non-Purdue people
  - From organizations known to Globus: search for their institution in the Organization drop-down menu. Will be taken to their institution’s login page.
  - From organizations not known to Globus: “use Globus ID to sign in”

- You will land in “File Manager” tool
- Docs: docs.globus.org/how-to/get-started/
Data transfers

- Go to transfer.rcac.purdue.edu or globus.org and login using “Purdue University Main Campus” as organization from drop-down menu. Use BoilerKey 2FA.

- Globus transfers:
  - In the “File Manager” tool, search for source collection in one panel, destination collection in another panel... highlight files, tweak options, hit “Start!”
  - Can be scheduled (repeated on a timer!)
  - Globus getting started guide: docs.globus.org/how-to/get-started/ (also in every RCAC resource’s User Guide under “File Storage and Transfer” section)
Data Sharing

- Globus can be used for sharing – **even when recipient(s) do not have account on our system!**

- Go to transferr.rcac.purdue.edu or globus.org and login using “Purdue University Main Campus” as organization from drop-down menu. Use BoilerKey 2FA.

- Globus sharing:
  - “European colleague needs to get (or put) a terabyte of data in my scratch space”
  - In “File Manager,” navigate to files you want to make available, click “Share” to create a share, then select people/groups to grant access. Read-only or read-write.
  - Globus sharing guide: docs.globus.org/how-to/share-files
Globus Connect Personal: make your computer an endpoint

- Not needed to transfer between *existing* endpoints
- Needed to teach your computer speak Globus

- Download: [app.globus.org/file-manager/gcp](http://app.globus.org/file-manager/gcp) or from the “Collections” section inside Globus:

- Versions for all major OS: [www.globus.org/globus-connect-personal](http://www.globus.org/globus-connect-personal)
- Does not require administrator privileges

- Detailed docs:
  - For Windows
  - For Mac OS X
  - For Linux
Example installation of Globus Connect Personal

Globus Connect Personal Setup would like to:

- View information about your linked identities
- View identity details
- View your identity
- Create Globus Connect Personal collections in the Globus Transfer service

Provide a label for future reference:
My Laptop

You can rescan this consent at any time by visiting the Manage Collections page.

By clicking “Allow”, you allow Globus Connect Personal Setup, in accordance with its terms of service and privacy policy, to use the above listed information and services.

Owner Identity: lew@purdue.edu
Collection Name: My new laptop endpoint
Description: My office laptop

High Assurance: [ ]

Choose this option only if your computer stores sensitive information such as Protected Health Information or Controlled Unclassified Information.

You will now see Globus Connect Personal running in the notification area of the taskbar. You can access your collection using the links below.

Save

Runs in the taskbar
Globus Overview

Globus Tips and Tricks

Do’s, don’t’s and usage scenarios
A note on VPN and network paths

- You do not have to be on VPN to use Globus
- For Globus transfers to/from your computer, **VPN will slow you down**
- **Common mistake**
  - “I have Data Depot mounted on my computer as a network drive, I will use Globus Connect Personal on my computer and share/transfer off of that drive”
  - **Painfully slow and flaky** (data travels from Depot, through Purdue VPN, down to your PC, and then up from PC to the destination)
  - Share/transfer from the **main Depot collection instead** (direct flow Depot -> destination!)
Fortress... bundle up!

- You will see this warning on Fortress collections:

- Fortress is a tape archive
  - Give it 1 GB in one file, and it will fly through it happily
  - Give it the same 1 GB in a million of 1 KB files, and storage admins will not like you (and you will not like yourself when it comes to extraction)

- Globus makes it way too easy to “just drop” a million of small files, so please be aware (and don’t)!
  - A “small file” on Fortress scale is something under 30-50MB per file
  - This applies to HSI and SFTP usage, too
Fortress... try not to retry

- On most endpoints, Globus automagically resumes interrupted transfers from the offset
- Fortress is an exception – Globus restarts the entire file from the beginning
- Getting a large file from Fortress over poor network... good luck

- Two-step to the rescue:
  - Transfer from Fortress to any “normal” filesystem (Depot or cluster scratch)
  - Transfer from this intermediate stop to the final destination

```plaintext
Fortress ➔ Fast RCAC Network ➔ Depot or Scratch

- frequent drops, keeps restarting from zero, never finishes
- very rare drops
- may restart from zero, but finishes fast
```

frequent drops, but resumes from offset keeps progressing
Fortress... sync smart!

- Do not use “sync by checksum” on Fortress!
  - Any other method is ok.

- Remember, Fortress is a tape archive
- Existence, size, modification time are metadata (i.e. quick lookup)
- But checksum needs to be computed (by physically reading the file). I.e. every archived file needs to be staged from tape to disk cache and then check-summed to compare.
- I.e. a lot of tape read overhead, no win.
Globus Tips and Tricks

Command-line and developer friendly!

- Command line utility: docs.globus.org/cli/
  - Installed on all RCAC systems
  - Cross-platform, easily installable anywhere
  - Can do anything web GUI does, and more
  - Scriptable transfers and workflows (examples at github.com/globus/automation-examples)

- Command line utility for scheduled transfers: pypi.org/project/globus-timer-cli/

- Also has an API and a full-blown Python SDK
  - Can use to build CLI and web applications, gateways and portals

$ globus --help
$ globus list-commands

$ globus-timer --help
Globus Tips and Tricks

Globus use case scenarios for researchers and facilities

- **Unattended transfers** between internal or external storage resources
- **Share data** with collaborators
- **Publish data** (a.k.a “share with the world”)
- Deliver to customers
- Transfer from an instrument PC
- Send to home base from the field
- Make “incoming/outgoing” boxes
- **Fortress made easy!**
  - Individual backup subfolders in the lab Fortress space (more flexible and easy to use permissions than with hsi/htar/Unix groups!)

Tell us your needs - we are very interested in working with you!
New capability: Globus Flows

- A service for defining and executing secure reliable automated data flows at scale
- A flow (workflow) is made of series of steps and a state machine description ("if this then that")
- Steps (actions) are carried out by “action providers” – and can be either in the Globus ecosystem, or outside. E.g.:
  - Data replications
  - Landing zone two-step transfer
  - Acquire – process – archive cycles
Benefits of Purdue Globus subscriptions

Anyone can use Globus’ free tier:

- Unlimited transfers
- Unlimited un-managed endpoints
- “All-or-nothing” sharing (can chose to make things either fully private or fully public)
- Web and CLI access

Purdue subscription adds:

- Flexible file sharing (private, public, and anything in between)
- Unlimited managed shareable endpoints on all RCAC filesystems
- Ability to grant managed shareable status to endpoints operated by other Purdue units
- Globus Plus (extras to enable GCP-to-GCP transfers and sharing from GCP endpoints)
- Globus Console for IT staff
- Globus Support for IT staff
Globus Overview

What Comes Next?
Upcoming Seminars

- Research Storage 101: March 10
- Software Installation 101: March 3
- Open OnDemand 101: March 24
- Workflow Automation Tools for Many-Task Computing: March 30
- Running Bioinformatics Analysis in HPC: March 9
- Containerized Bioinformatics Applications for HPC: March 29
- NLP 101: March 31
- Time Series Forecasting 101: April 7

https://www.rcac.purdue.edu/news/events
THANK YOU

Feel free to reach out to lev@purdue.edu with questions.

Slides and recording are posted at:
https://www.rcac.purdue.edu/training/globus

General help: rcac-help@purdue.edu

Coffee Hour: https://www.rcac.purdue.edu/coffee