

Globus Overview

April 23, 2021



This material is based upon work supported by the National Science Foundation under Grant No. 1827184.

Lev Gorenstein
Senior Computational Scientist
ITaP Research Computing
lev@purdue.edu



There's more to Research Computing than just computing!

- Sure, 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 for all our storage options
- An interactive storage solutions finder: www.rcac.purdue.edu/storage/solutions/
- Also check out www.rcac.purdue.edu/services for other services we provide

There's more to Research Storage than just storage!

Locations

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

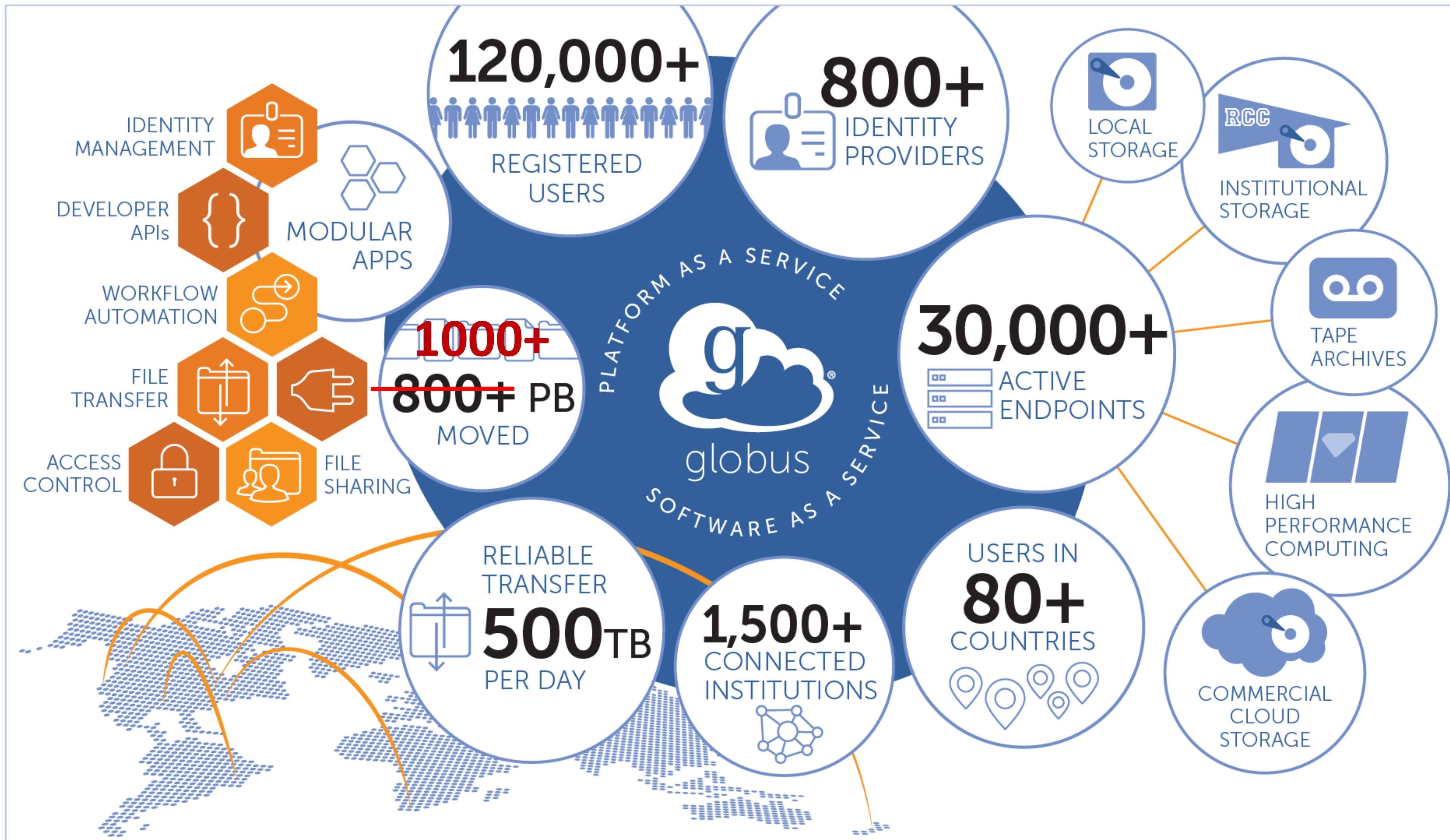
Actions

- Generate
- Process/analyze
- **Transfer**
- **Share**
- **Publish**



What is Globus?

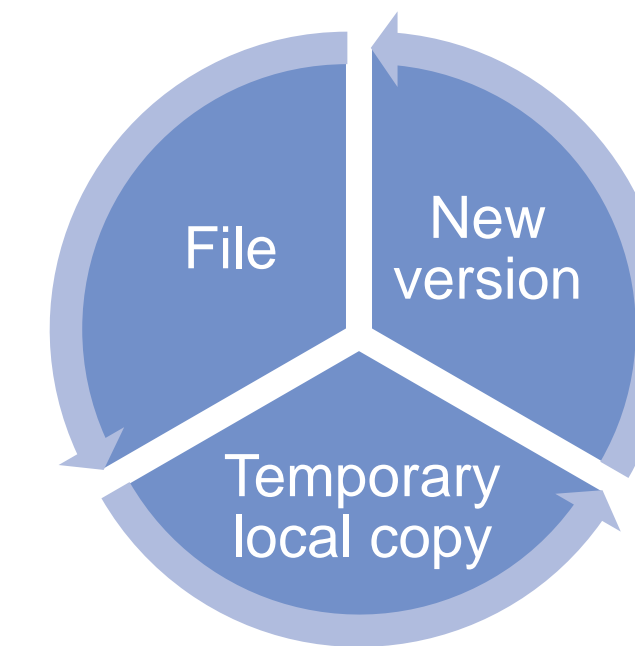
- Globus is a non-profit service for secure, reliable research data management
- A team at the University of Chicago and Argonne National Laboratory
- Funded by NSF, DoE, NIH and institutional subscriptions (freemium model)
 - Purdue is subscribed
- Stems from GridFTP and high energy physics community, but grew much beyond that
- 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



What Globus is not?

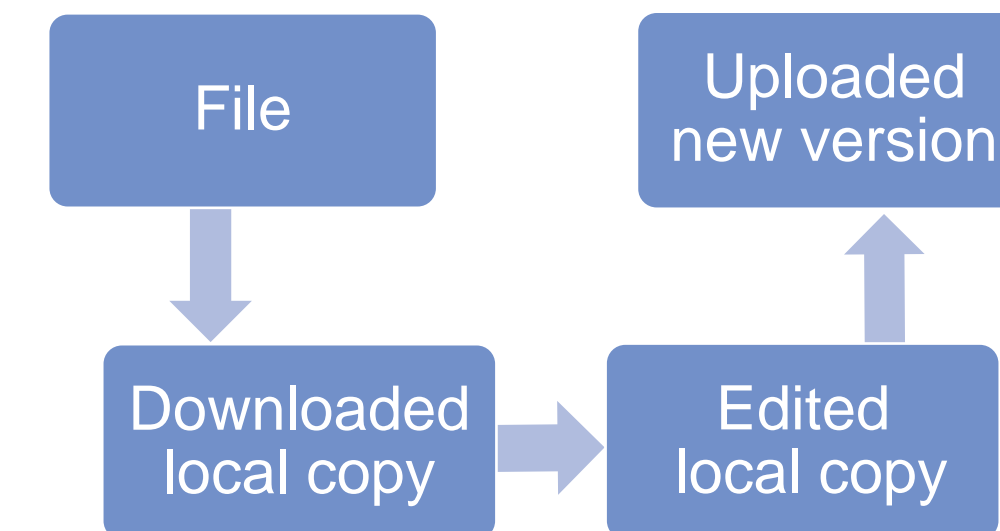
- **Globus is not your typical network drive!**

- 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?

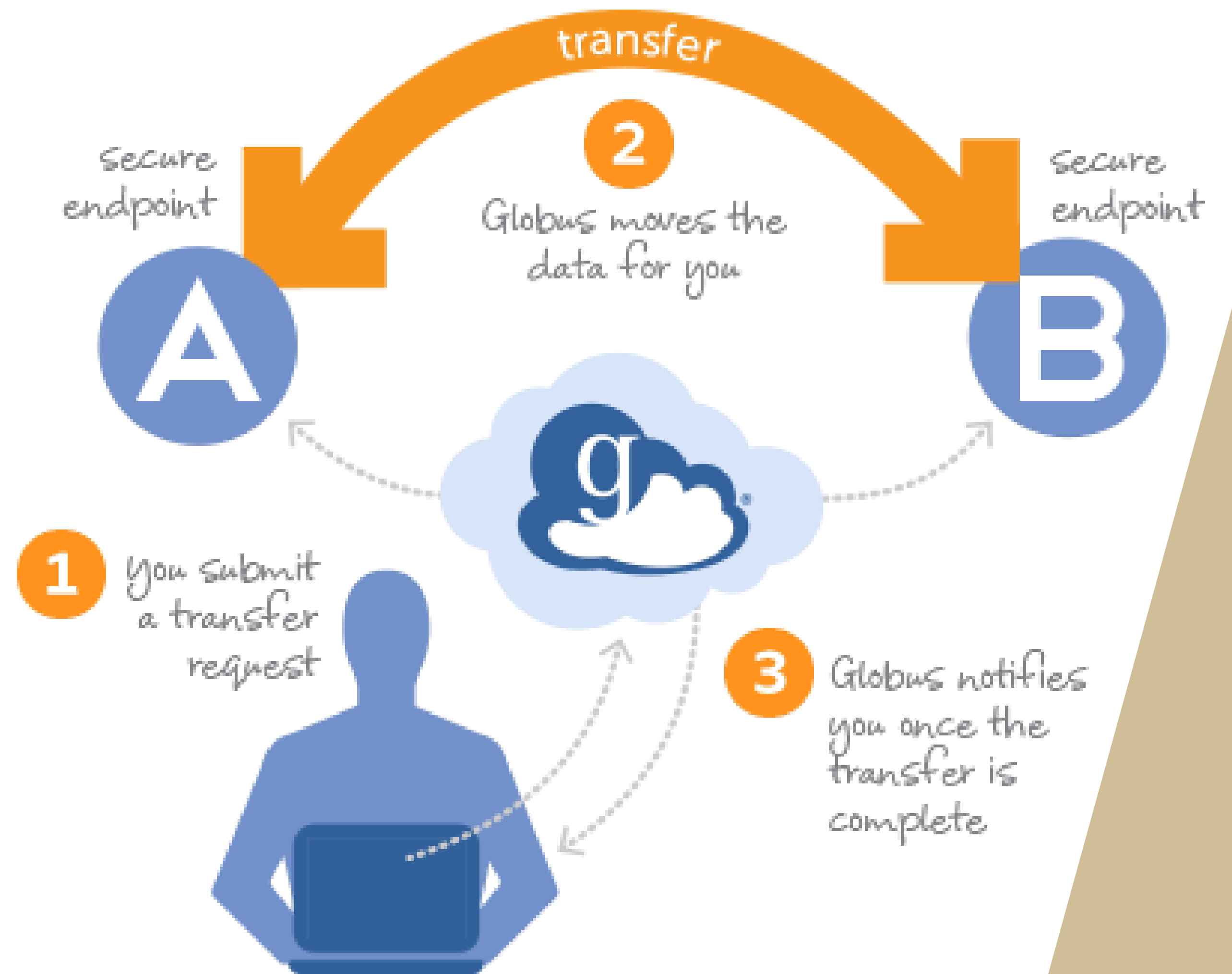
- **Nothing**
- **It's a transfer tool!**
The “*download*”, “*edit*”, “*upload*” steps are fully decoupled and have to be done explicitly by the user



- On some modern endpoints, you may be taken to the browser's “Open/download” dialog, but still no automatic back and forth
 - None of the RCAC endpoints have this feature yet

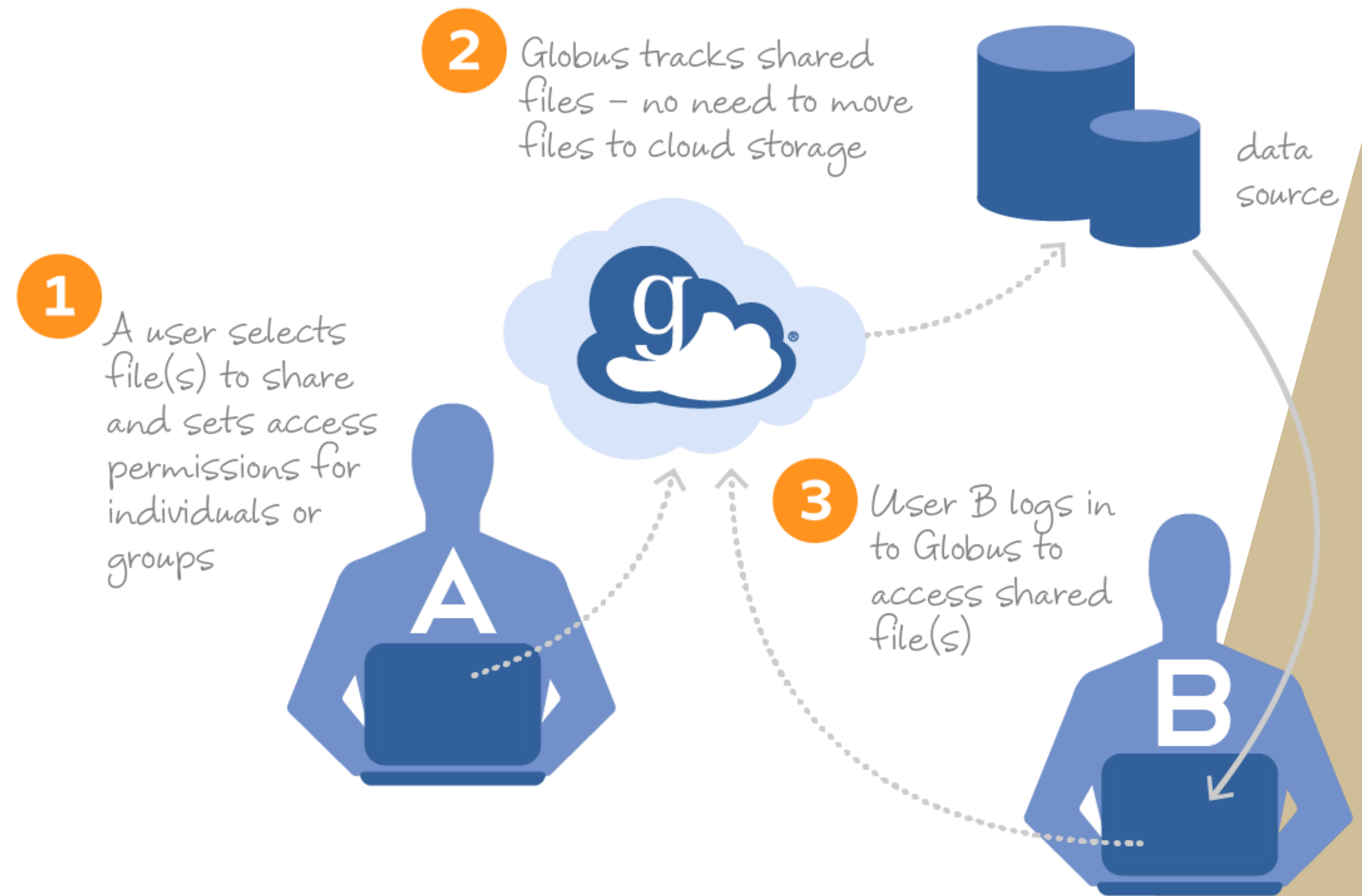
Globus transfers overview

- 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



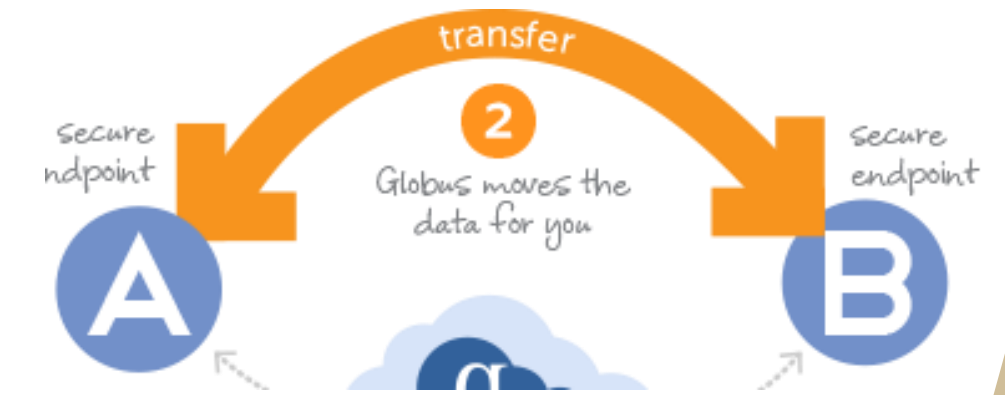
Globus sharing overview

- 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!



Vocabulary: Endpoints (a.k.a. Collections) and Shares

- “A named location containing data you can access with Globus”
- Historic terminology:
 - “**Endpoint**” – the main location itself (e.g. “*Purdue Data Depot*”).
 - “**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*”).
- Globus recently introduced new terminology:
 - “**Endpoint**” refers to hardware/software/system component (what admins deal with)
 - “**Collections**” refers to the named location components (what users deal with)
 - “**Shares**” – parts of the main collection that have been given their own names and shared via Globus
- New adoption is slow, people often still use “Endpoints” in the sense of “Collections”
- Shares (shared endpoints) are named locations, so they are Endpoints (Collections), too



Vocabulary: Globus Account and Identities



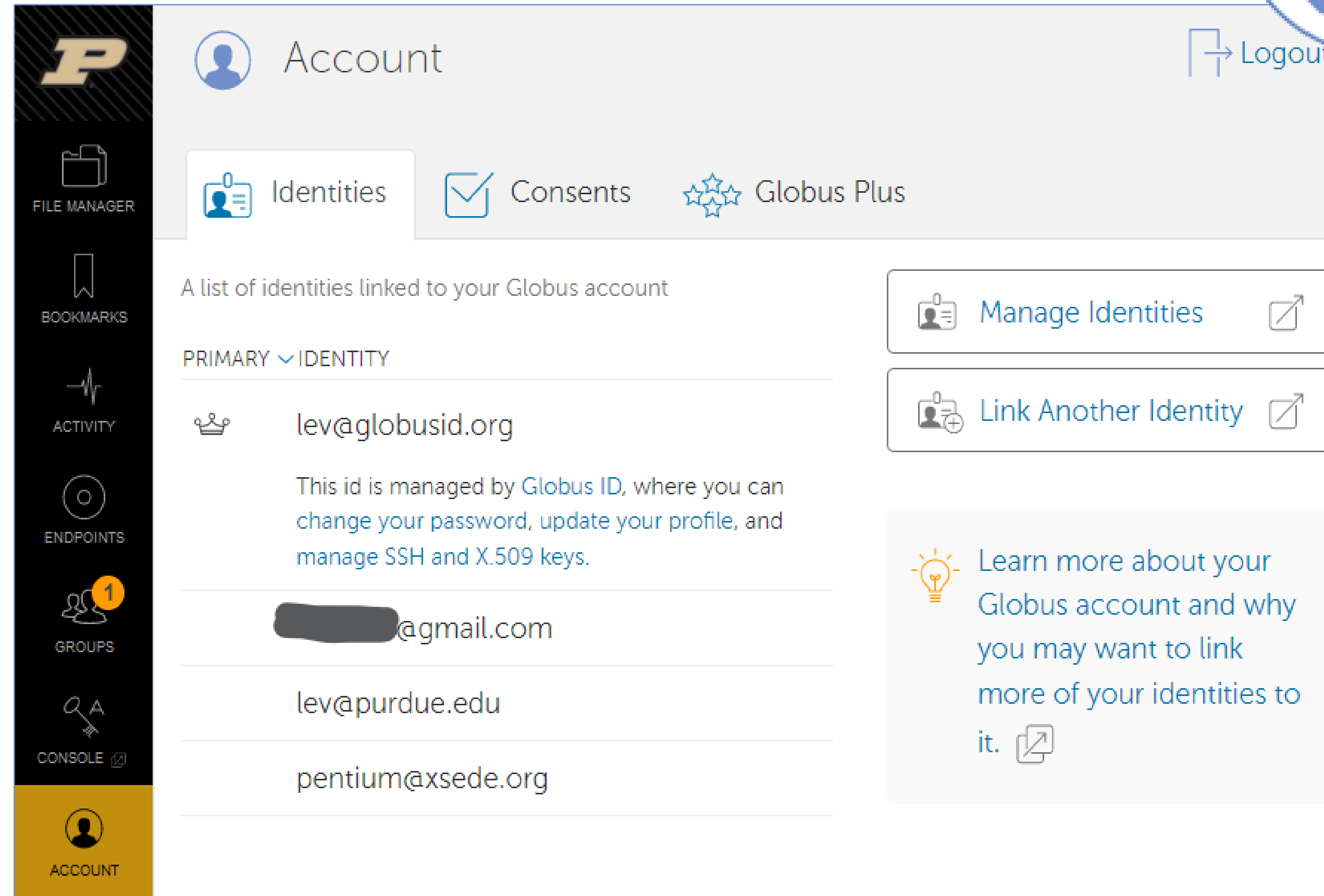
- ***“You and the hats you wear”***
- Globus needs a handle to know you by (and to authenticate you) – a Globus account
- In the simplest form, this is your organizational login, but there are be many more *Identity Providers* that Globus recognizes (e.g. Gmail, ORCiD, etc).
 - Purdue is recognized – “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).
 - It’s the one of the 800+ 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, a.k.a. “I’ll just let them create an account for themselves”)

Note: anyone can use Globus! You don’t 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 XSEDE 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



The screenshot shows the 'Account' page in the Globus interface. On the left is a navigation sidebar with icons for File Manager, Bookmarks, Activity, Endpoints, Groups (with a '1' notification), Console, and Account. The main content area is titled 'Account' and includes a 'Logout' button. Below the title are tabs for 'Identities', 'Consents', and 'Globus Plus'. A message states: 'A list of identities linked to your Globus account'. Under the 'PRIMARY IDENTITY' section, the following identities are listed:

- lev@globusid.org (marked as primary with a crown icon): This id is managed by Globus ID, where you can change your password, update your profile, and manage SSH and X.509 keys.
- [Redacted]@gmail.com
- lev@purdue.edu
- pentium@xsede.org

 To the right of the identity list are two buttons: 'Manage Identities' and 'Link Another Identity'. At the bottom right, there is a lightbulb icon and a text box that says: 'Learn more about your Globus account and why you may want to link more of your identities to it.'

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 **GlobusID** to sign in”.
- Docs: docs.globus.org/how-to/get-started/

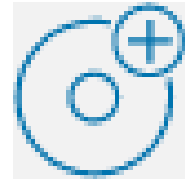


The screenshot shows the login interface for the Purdue Globus Web App. At the top left is the Purdue University logo. At the top right is the Globus logo. The main heading is "Log in to use Purdue Globus Web App". Below this is the instruction "Use your existing organizational login" with examples: "e.g., university, national lab, facility, project". A dropdown menu is set to "Purdue University Main Campus". A link says "Didn't find your organization? Then use **Globus ID** to sign in. (What's this?)". A blue "Continue" button is present. A grey box contains the CILOGON logo and text: "Globus uses CILogon to enable you to Log In from this organization. By clicking Continue, you agree to the [CILogon privacy policy](#) and you agree to share your username, email address, and affiliation with CILogon and Globus. You also agree for CILogon to issue a certificate that allows Globus to act on your behalf." Below this is an "Or" separator. At the bottom are two buttons: "Sign in with Google" and "Sign in with ORCID iD".

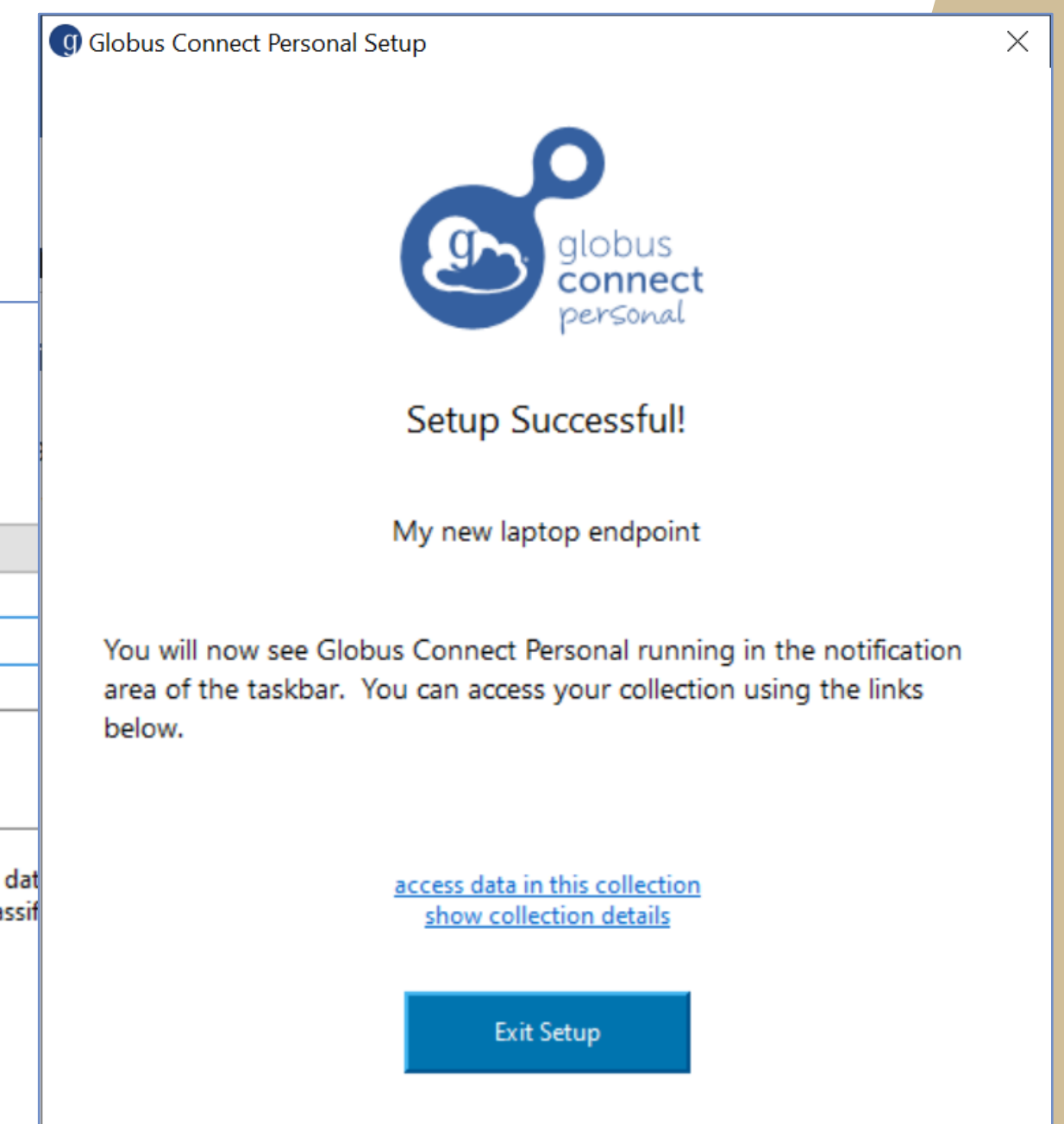
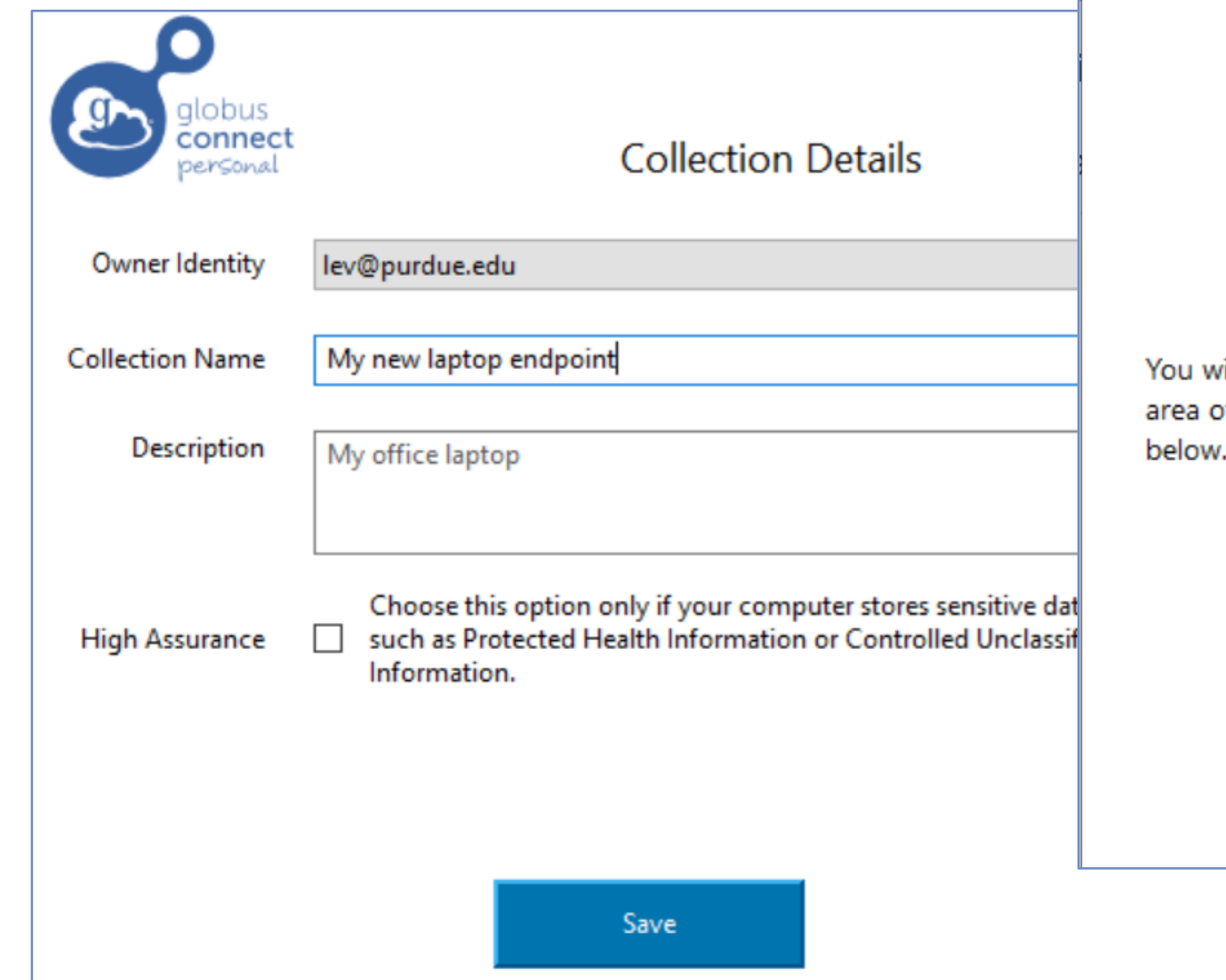
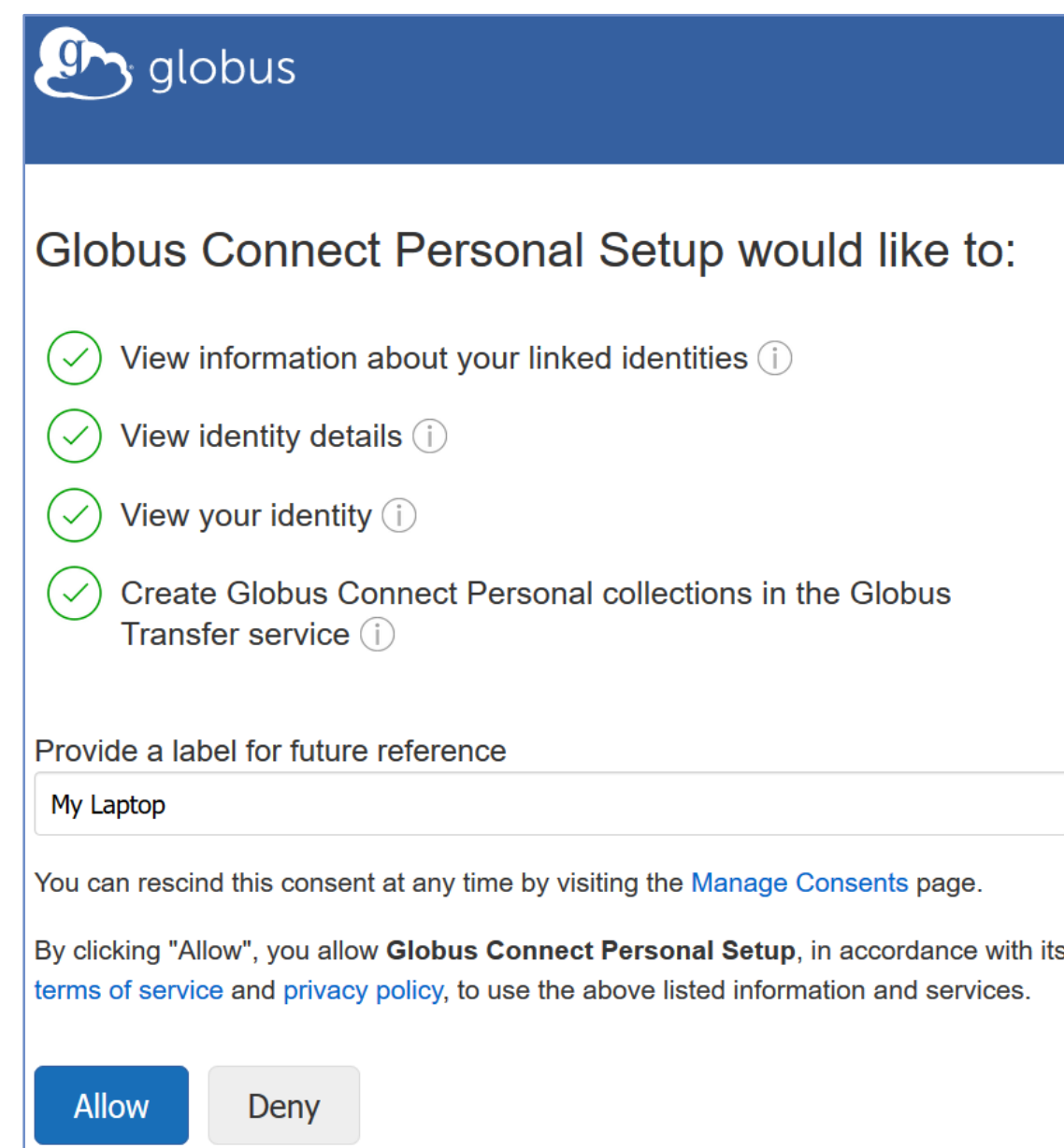
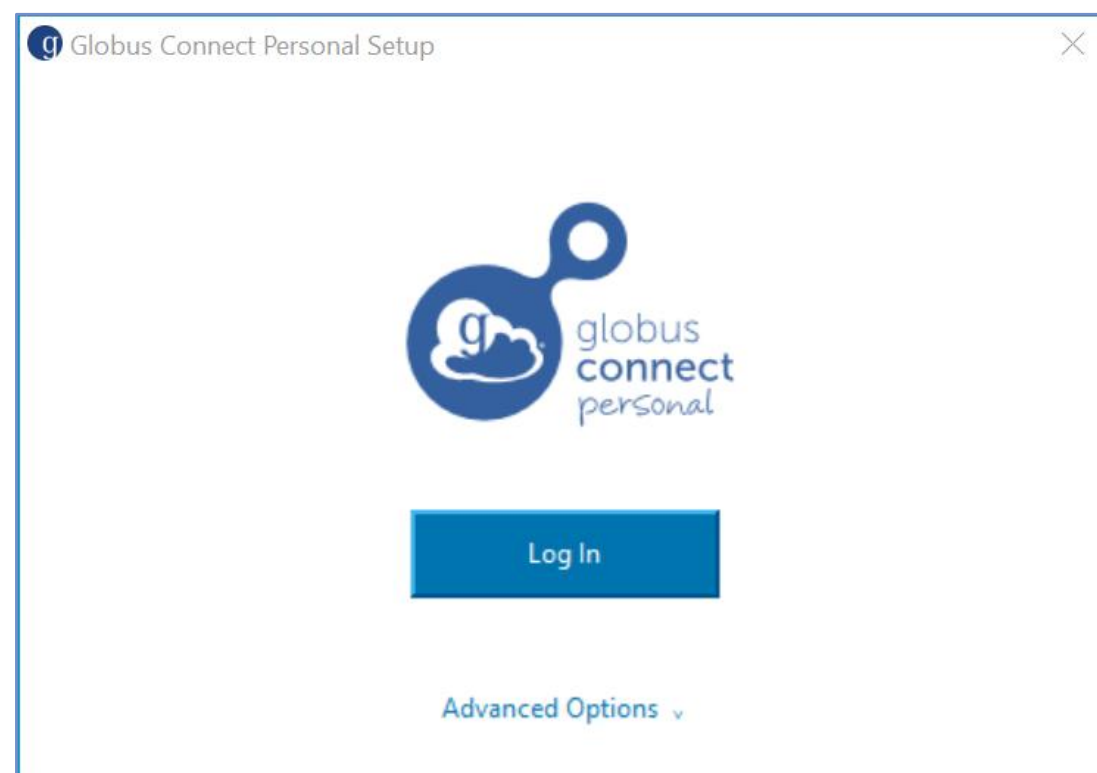
Demo – Globus

- Also in every resource's User Guide under *"File Storage and Transfer"* section
- 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:
 - Search for source endpoint in one panel, destination endpoint in another panel... highlight files, hit *"Start!"*
 - Globus getting started guide: docs.globus.org/how-to/get-started/
- Globus can be used for sharing – even when recipient(s) do not have account on our system!
 - *"European colleague needs to get (or put) a terabyte of data in my scratch space"*
 - Navigate to files you want to make available, click *"Share"* to create a share, then select people/groups to grant access to the share
 - Globus sharing guide: docs.globus.org/how-to/share-files

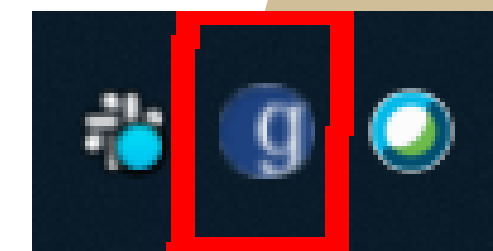
Demo of *Globus Connect Personal*: make your computer an endpoint

- Not needed to transfer between *existing* endpoints
- Needed to teach your computer speak Globus
- Download: app.globus.org/file-manager/gcp or from the “Endpoints” section inside Globus: 
- Docs: www.globus.org/globus-connect-personal

- Versions for major OS
- Example installation:

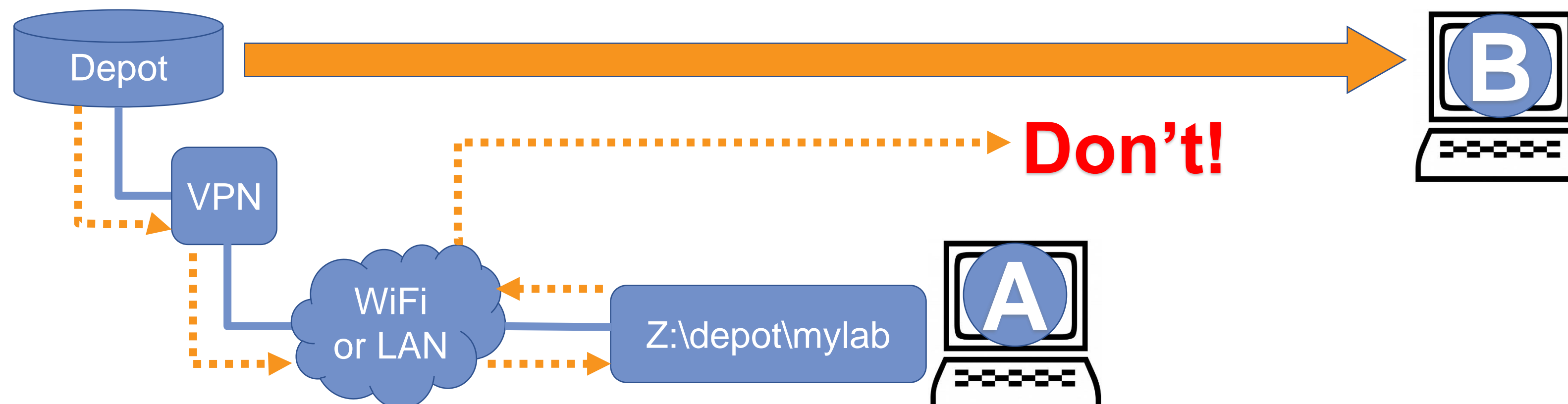


Runs in the taskbar



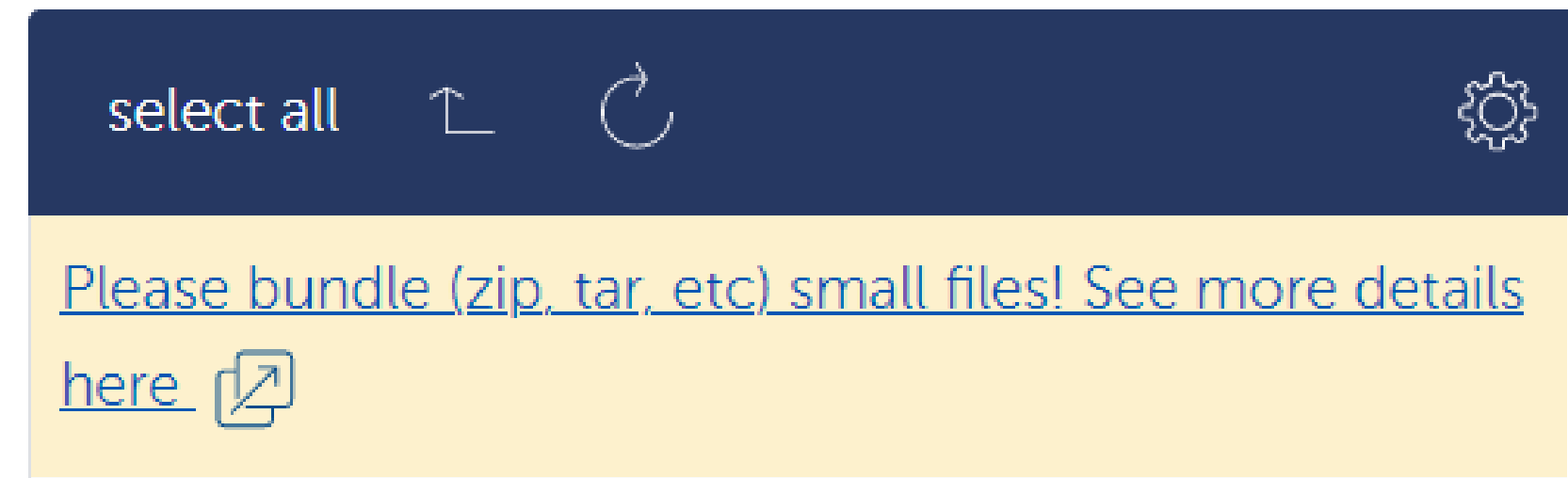
A note on VPN and network paths

- 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 endpoint instead** (direct flow Depot -> destination!)



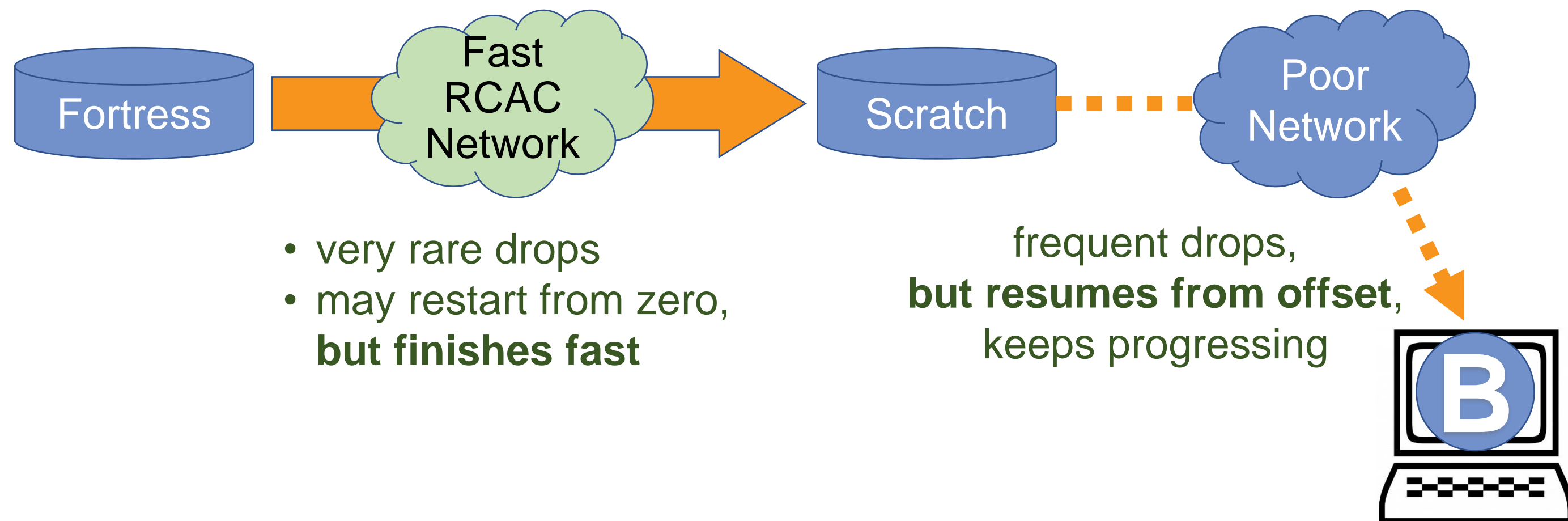
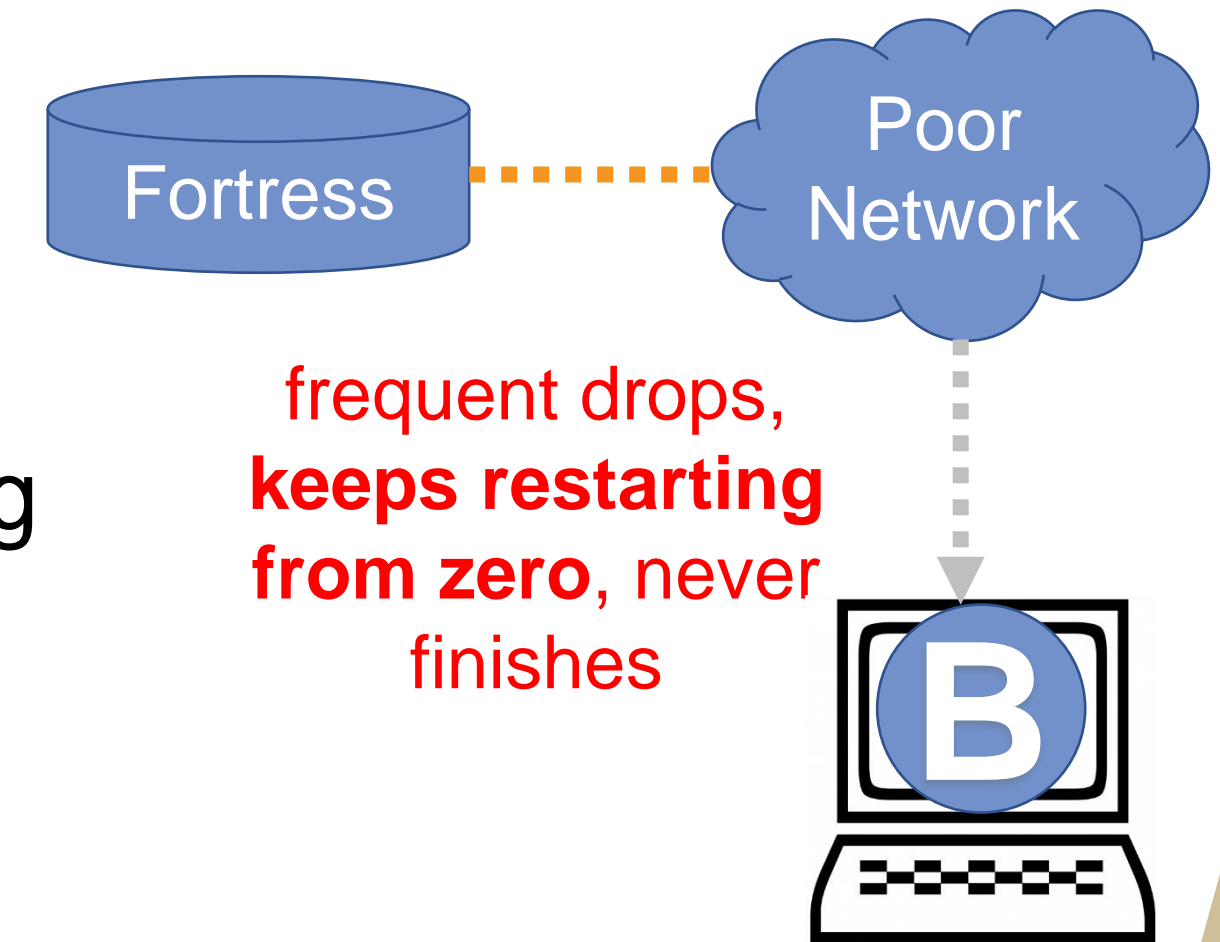
Fortress... bundle up!

- You will see this warning on Fortress endpoints:
- 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 you when it comes to extraction)
- Globus makes it way too easy to “just drop” a million of small files, so please be aware
 - A “*small file*” on Fortress scale is typically considered something under 30-50MB per file
 - This applies to HSI and SFTP usage, too



Fortress... try to not 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



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)

```
$ globus --help  
$ globus list-commands
```

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

```
$ globus-timer --help
```

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

Possible 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!**

Benefits of Purdue Globus subscription

- **Anyone can use Globus' free tier:**
 - **Unlimited transfers**
 - **Unlimited un-managed endpoints**
 - No sharing (but 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
- **More to come:**
 - Endpoints on Box Research Lab Folders and REED Folders

Thank You

Questions?

Slides: rcac.purdue.edu/training

Email Help: rcac-help@purdue.edu

Sign-in Coffee Hours: Monday–Thursday, 2:00 – 3:30pm

WebEx

rcac.purdue.edu/coffee

Friday Seminars: Fridays, 2:00pm

WebEx

rcac.purdue.edu/news/events