Globus – A Platform for Research Data Management Challenges

ARGONNE ATPESC2025 EXTREME-SCALE COMPUTING

Lev Gorenstein
lev@globus.org

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

globus

Globus is …

a non-profit service developed and operated by

THE UNIVERSITY OF CHICAGO

Our mission is to…

increase the efficiency and effectiveness of researchers engaged in data-driven science and scholarship through *sustainable* software.

Our team comprises …

professional software developers and business operators with extensive experience in industry and academia

globus

a uchicago non-profit service

globus labs

Our freemium sustainability model

- Basic capabilities are available free of charge to anyone engaged in non-profit research
- Subscriptions enable multiple enhanced features for both researchers and system administrators
- Subscription pricing model based on fairness and equity

globus.org/subscriptions

Development is funded by...

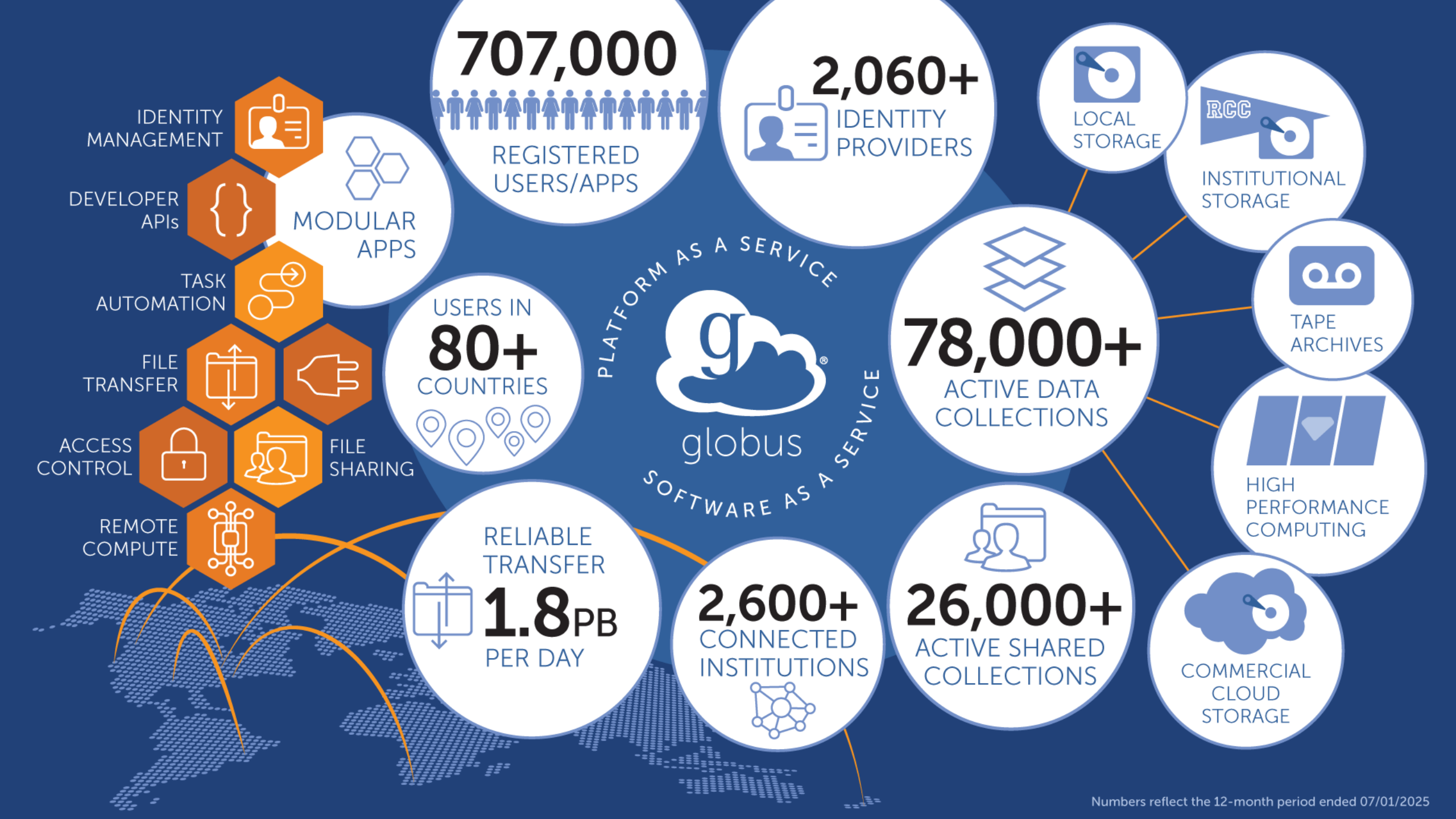
U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

Argonne
NATIONAL LABORATORY

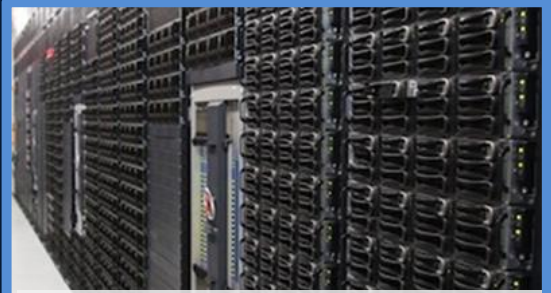
Operations are funded by subscribers


707,000
REGISTERED USERS/APPS
2,060+
IDENTITY PROVIDERS
IDENTITY MANAGEMENT
DEVELOPER APIs
MODULAR APPS
TASK AUTOMATION
FILE TRANSFER
ACCESS CONTROL
FILE SHARING
REMOTE COMPUTE
USERS IN 80+ COUNTRIES
PLATFORM AS A SERVICE
globus
SOFTWARE AS A SERVICE
LOCAL STORAGE
RCC
INSTITUTIONAL STORAGE
TAPE ARCHIVES
78,000+
ACTIVE DATA COLLECTIONS
HIGH PERFORMANCE COMPUTING
RELIABLE TRANSFER
1.8PB
PER DAY
2,600+
CONNECTED INSTITUTIONS
26,000+
ACTIVE SHARED COLLECTIONS
COMMERCIAL CLOUD STORAGE
Numbers reflect the 12-month period ended 07/01/2025

3,360,498,816,205 MB TRANSFERRED

Globus is a data management platform that delivers fast and reliable data transfer, collaboration, and automation services, directly from your own storage systems via software-as-a-service using existing identities.

“your own storage” can mean different things



Research Computing HPC

ACCESS

NICIS NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM

aarnet Australia's Academic and Research Network

National Resources

Personal Systems

Globus unifies data across disparate systems ®

Desktop Workstations

Microsoft Azure Blob Storage

aws | S3

Google Cloud

red cloud

Public Cloud Storage

Archives

Instruments

Globus Connectors support diverse systems

aws | S3

Google Cloud

Microsoft Azure Blob Storage

DATACORE | Swarm

SCALITY

IBM Cloud

Microsoft OneDrive

Google Drive

IBM Spectrum Scale

box

iRODS

Quantum

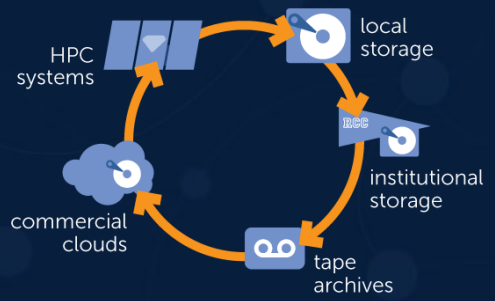
HPSS

wasabi
hot cloud storage

lustre

The Globus Web App and file transfer made simple

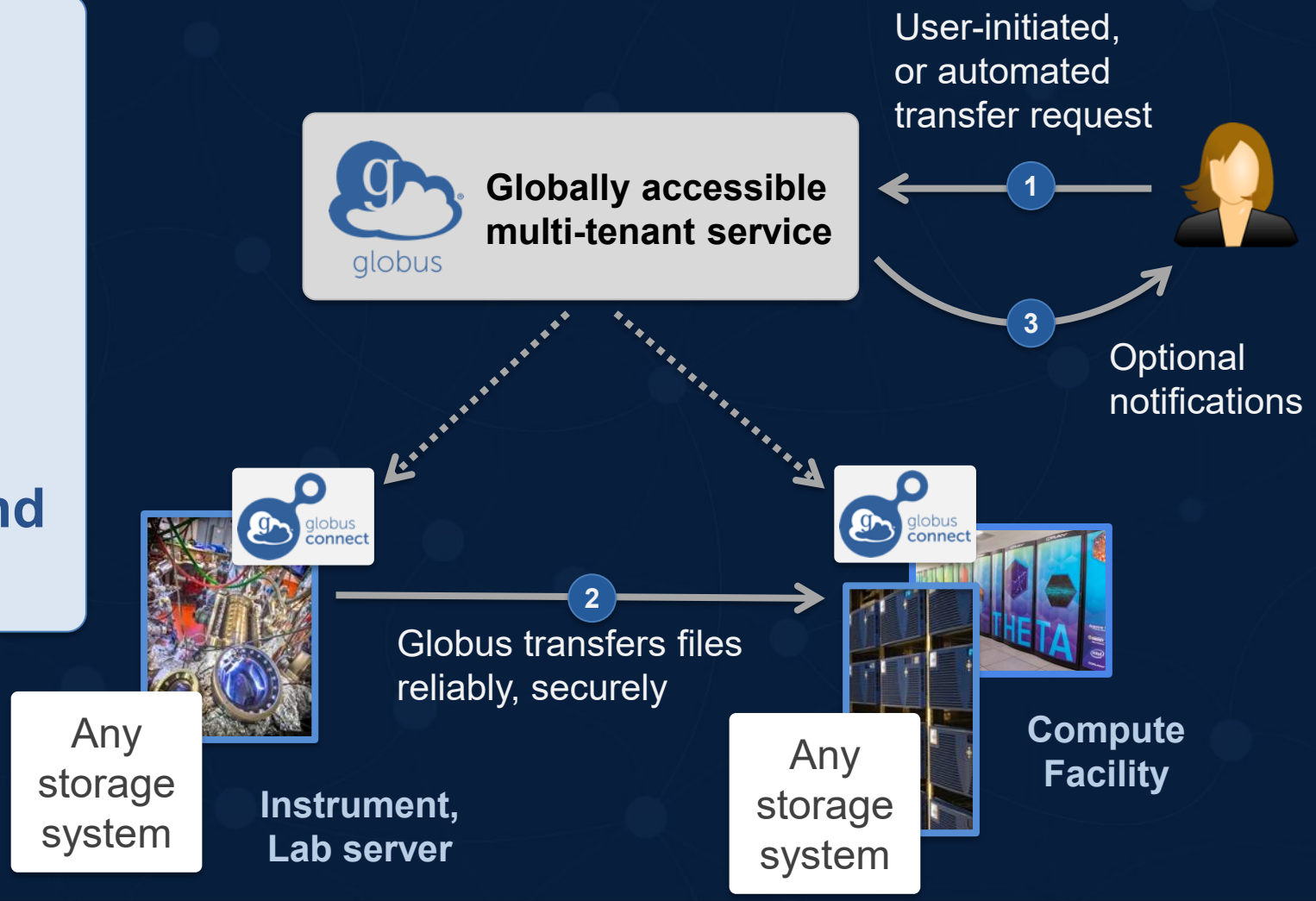
Managed transfer & sync



Unified data access

Assured data transfer …no babysitter required.

- **Fire-and-forget file transfer and synchronization**
- **Automated resolution of transient errors**
- **Browser-based access to "small" data**
- **Support for on-premises and cloud storage systems**



Endpoints, Collections and Globus Connect

- Globus Connect is installed to instantiate an Endpoint
- Endpoints host Collections
  - Collections abstract storage
  - Users interact with Collections
  - There are two types of Collections
    - Mapped Collections – Accessed by a user that has a local account on the storage system that collection abstracts.
    - Guest Collections – A collection that uses an existing mapped collection and adds the ability of a user to share access to their data on that collection.

Globus Connect

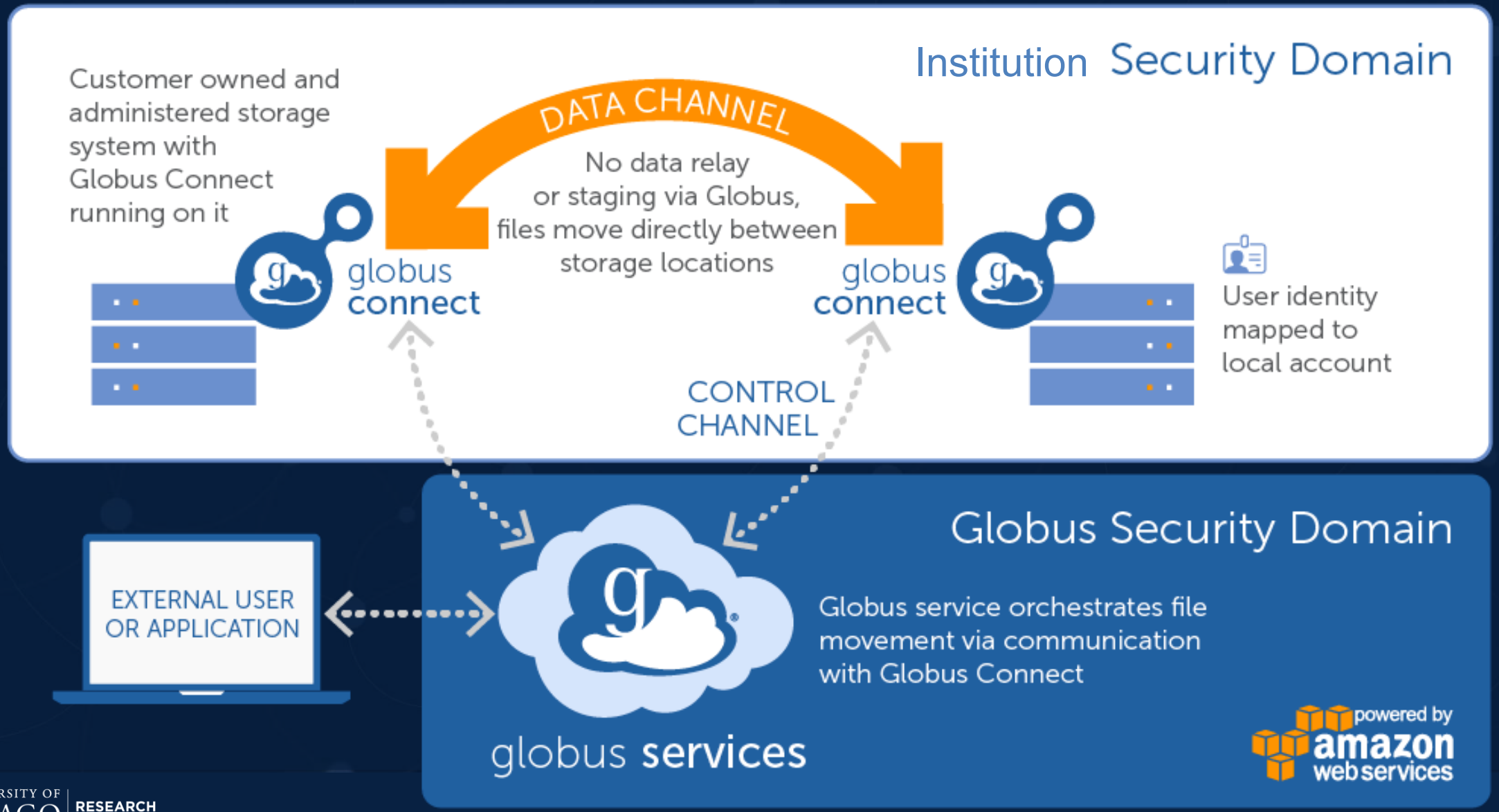
- **Globus Connect Server (GCS)**
  - Multi user Linux systems
  - May have different storage types
  - Installed by an administrator
  - Multiple mapped collections
- **Globus Connect Personal (GCP)**
  - Personal workstations and laptops (single user)
  - Installed by the user
  - Single mapped collection

Globus Mapped Collections at the ALCF

- **Home**
  - /home via alcf#dtn_home
- **Project based filesystems**
  - Grand
  - Eagle
- **HPSS**
- **https://docs.alcf.anl.gov/data-management/data-transfer/using-globus/**

Installing Globus Connect Personal

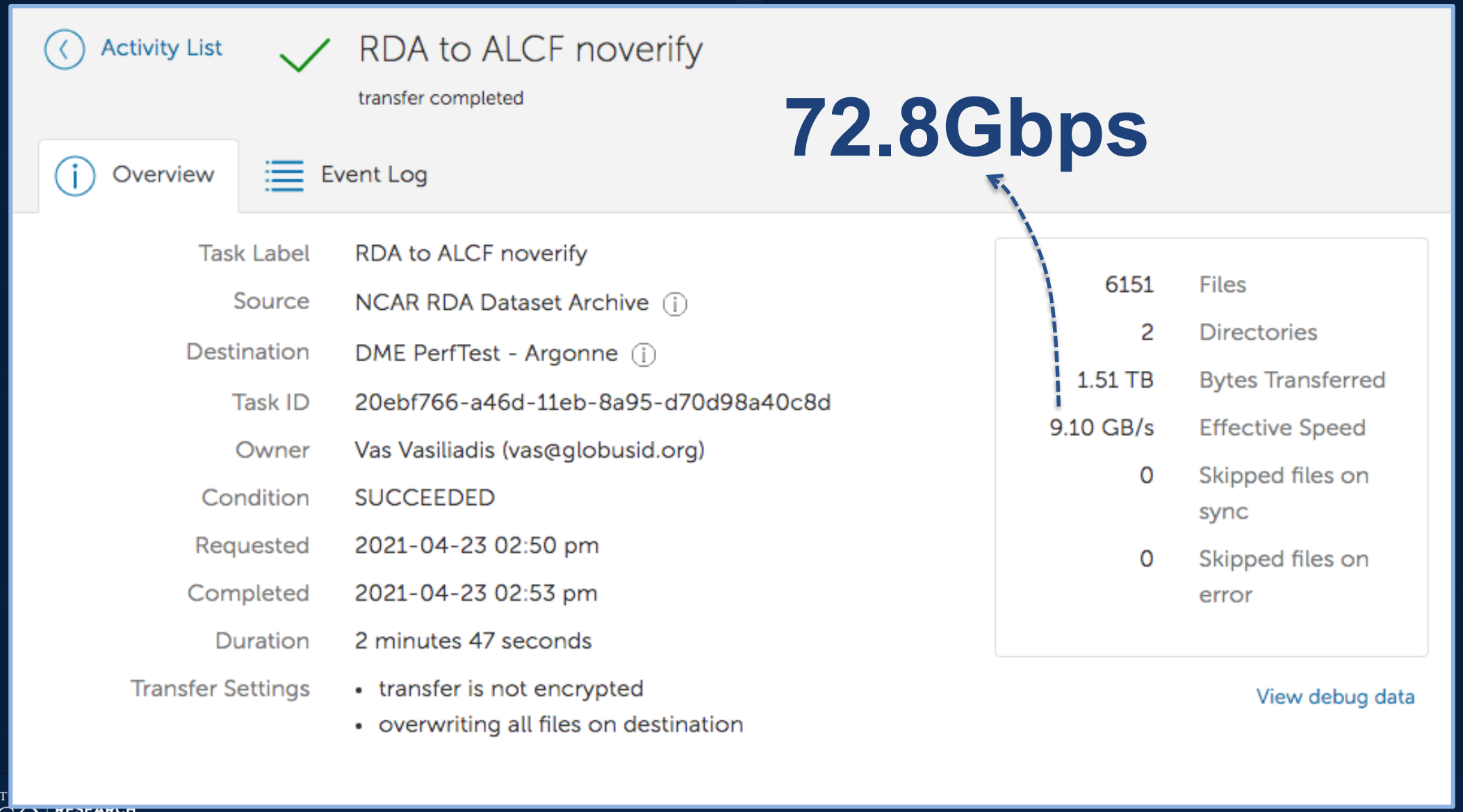
Quick review



Globus data movement, a tale of two protocols

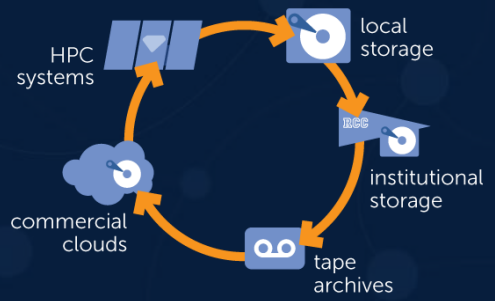
- Globus Transfer
  - GridFTP
  - Fast
  - Reliable (network back off, fire and forget, retries on failures)
  - A remote control for data
- Globus Downloads
  - HTTPS
  - Intended for situations where users don’t have the option of a destination endpoint
  - Direct to the machine doing the pulling
  - Likely to bring great sadness for large collections of files

GridFTP is reliable, secure …and fast!

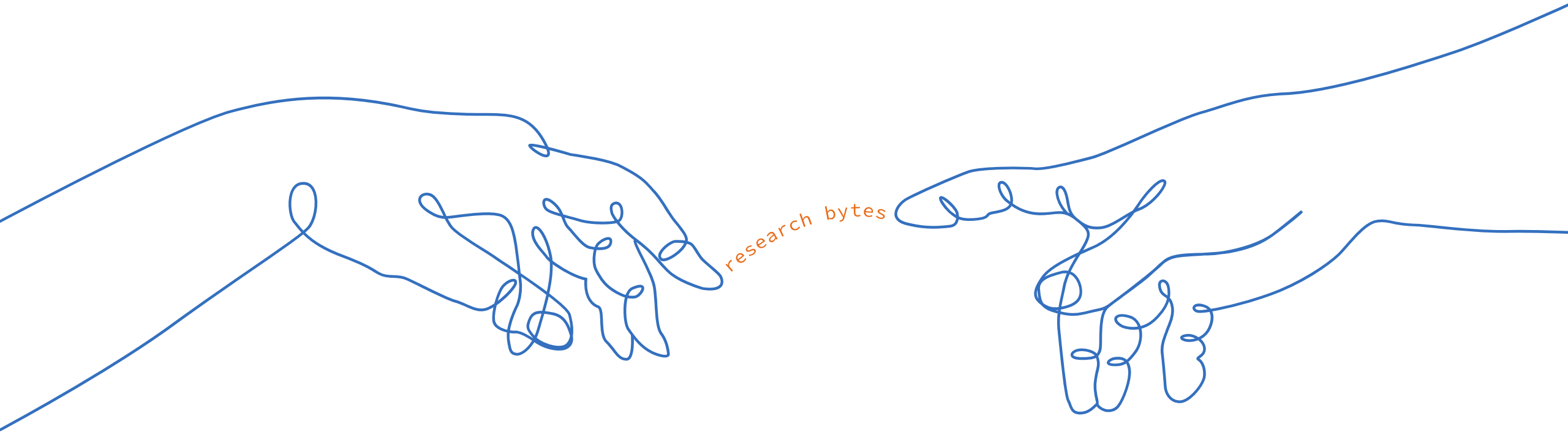


The Globus Web App and two common data transfer use cases

Managed transfer & sync



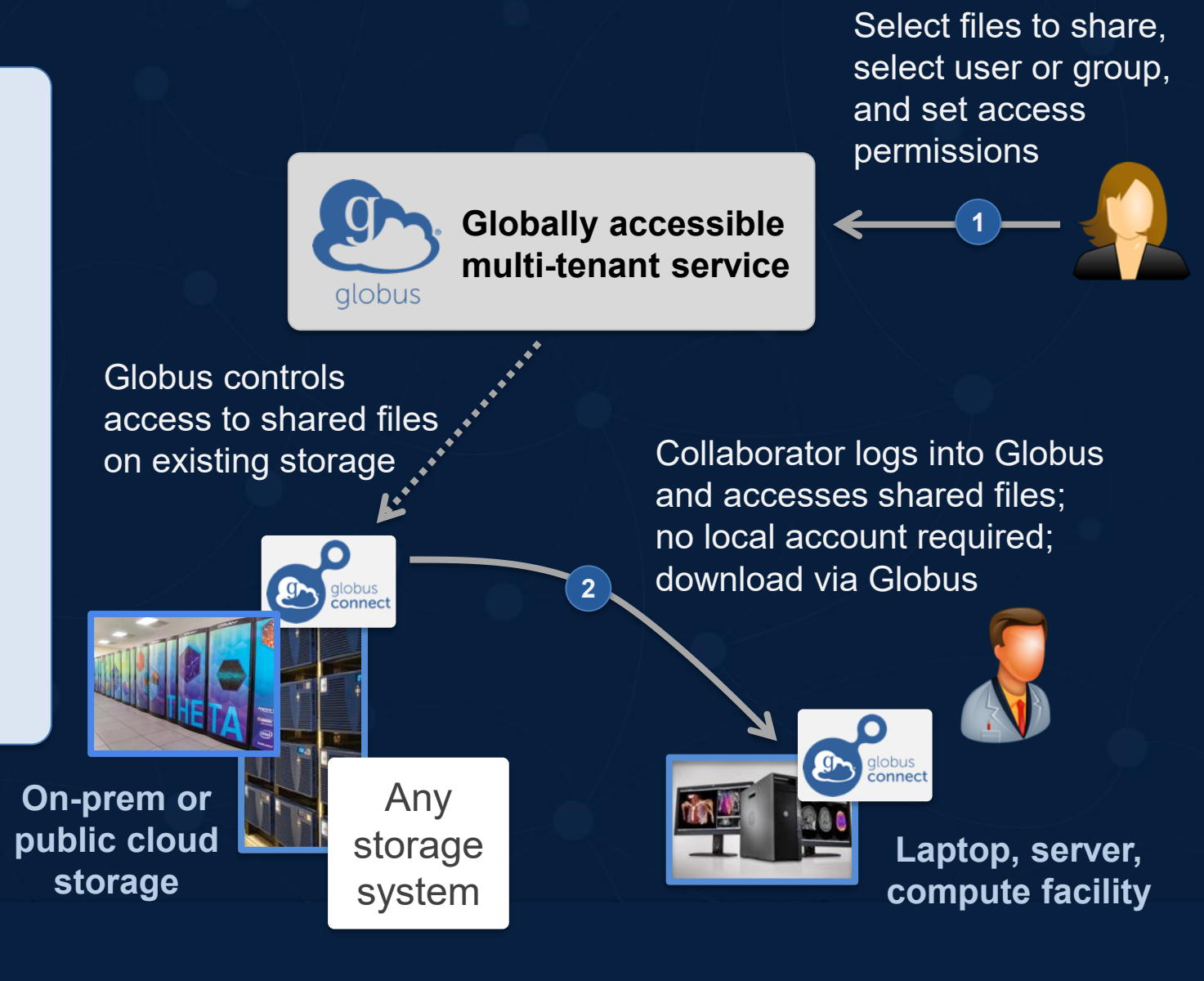
Unified data access



Research is inherently collaborative …but how do we share, distribute data easily yet securely?

Easy data sharing …no “special” treatment required

- **Share with collaborators at any institution, company**
- **Share from current storage - no data staging required**
- **Storage system administrators can enforce compliance policies**
- **Fine-grained access control**



The Globus Web App and a data sharing use case

Data sharing

Globus Sharing (Guest Collections) at the ALCF

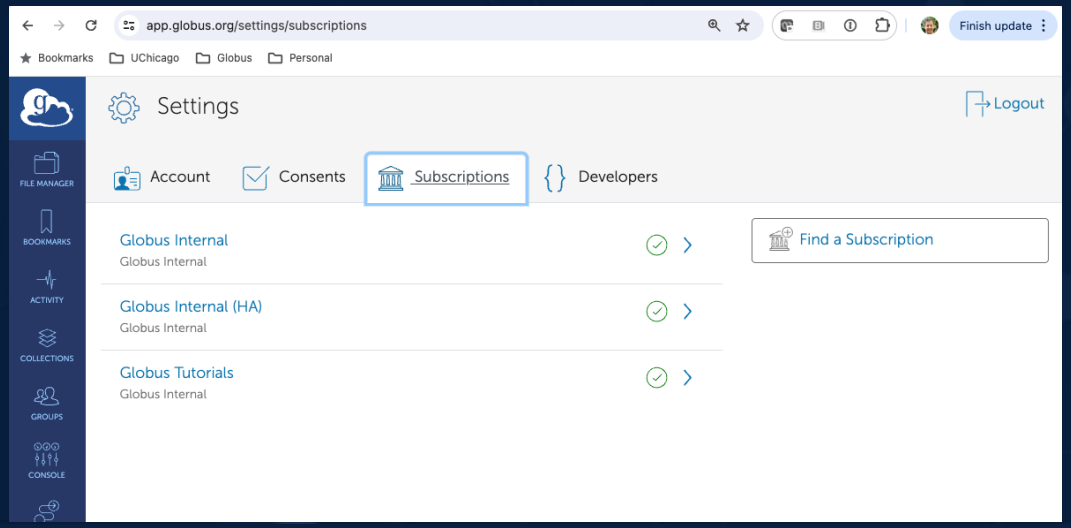
- **Grand / Eagle**
- **https://docs.alcf.anl.gov/data-management/acdc/eagle-data-sharing/**
- **Just Google “ALCF Globus Sharing”**

Globus Subscriptions

- By having a role in a Subscription Management Group and / or subscribing resources, Subscribers can enable
  - File Sharing: The ability to create Guest Collections
  - Premium Connectors: beyond POSIX
  - Administrative Oversight: The Globus Console
  - Usage Reports: Monthly and “since the beginning” stats
  - Data Automation Services: The ability to deploy more than one Flow
  - HTTPS: Add upload / download to collections
  - Manage protected data: Subscription options for protected data
    - High Assurance: PII, CUI
    - BAA: PHI

  … and more!
  - globus.org/why-subscribe

How can I find out if my institution is a Globus Subscriber and how can I join our Subscription Group?



Globus core security features

- **Access Control**
  - Identities provided and managed by institution
  - Institution controls all access policies
  - Globus is identity broker; no access to/storage of user credentials
- **Data remain at institutions, not stored by Globus**
- **Integrity checks of transferred data**
- **High availability and redundancy**
- **Encryption of user files and Globus control data**

Globus for protected data management

Security controls
→ NIST 800-53
→ 800-171 Medium

Restricted data handling
→ PHI, PII, CUI
→ Compliant data sharing

BAA w/Uchicago
→ UChicago BAA with Amazon

Globus High Assurance features

- **Additional authentication assurance**
  - Authenticate with a *specific identity* within session
  - Reauthenticate after specified time period
- **Session/device isolation**
  - Authentication context is per application, per session
- **Enforces encryption of all user data in transit**
- **Audit logging**

With Globus, it's not a tradeoff you need to make…

Access
Sharing
Collaboration

Security
Privacy
Compliance

Globus is SaaS and PaaS

Web app facilitates *ad hoc* data management

Platform services simplify creation of portals, gateways and applications for automating data management in instrument cores and other facilities

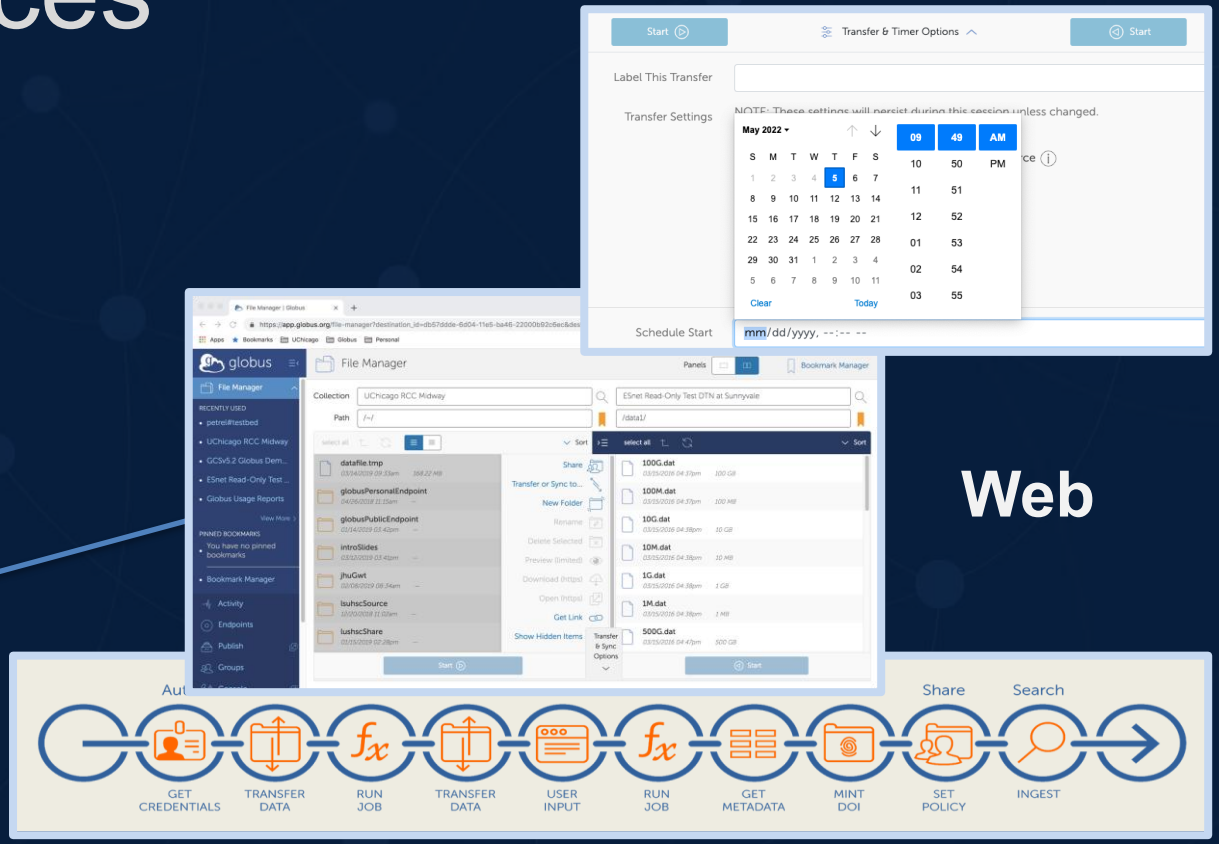
Services and Interfaces

```
(globus-cli) jupiter:~ vas$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

Options:
  -v, --verbose                 Control level of output
  -h, --help                    Show this message and exit.
  -F, --format [json|text]      Output format for stdout. Defaults to text
  --map-http-status TEXT        Map HTTP statuses to any of these exit codes:
                                0,1,50-99. e.g. "404=50,403=51"

Commands:
  bookmark        Manage Endpoint Bookmarks
  config          Modify, view, and manage your Globus CLI config.
```

CLI



Web

Globus services

Transfer, Auth,
Flows, Group,
Search, Compute

```
GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
...
```

REST API
Python and
JavaScript SDK

Simple Globus Automation Capabilities

Timed transfer is a pretty common automation use case for backup and replication. The Globus WebApp allows you to integrate this type of automation quickly and intuitively.

Timer Service

The Globus WebApp supports recurring and scheduled transfers.

Code your own Automation

Command Line Interface

The CLI provides an interface to Globus services from the shell and is suited to both interactive and scripting use cases.

Globus API / SDK

Our open REST APIs, Python SDK and JavaScript SDK empower you to create an integrated ecosystem of research data services and applications. Harness the power of the Globus platform so you can focus on building your application.

Build automations …with a Managed Service

- **Managed reliable task orchestration**
- **Spanning all resources**
- **Event driven execution model – state machine**
- **Declarative language for flow definition**
- **Reuseable and sharable**
- **Definable input schema**
- **Extensible for integrating external services**

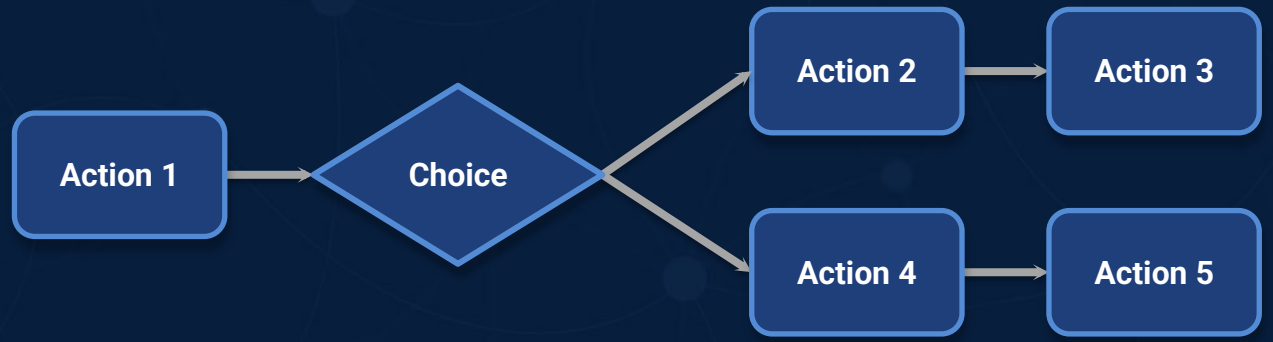
Globus Flows

- **Managed, secure (Globus Auth), service for reliable task orchestration**
- **Support for heterogenous resources**
- **Extensible and authorable event driven execution model**
  - Flow Definition (JSON)
  - Input Schema (JSON)
  - Deployment
- **Extensible via custom actions**

What's a Flow?

A sequence of steps…

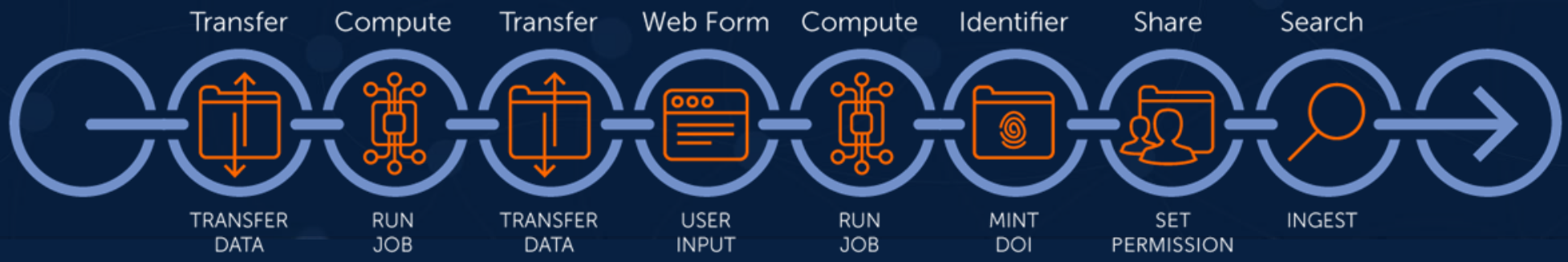
- **Hosted**
- **Reusable**
- **Flexible**
- **Shareable**



How Do You Use a Flow?

Each time a flow is started, it's called a ***run***

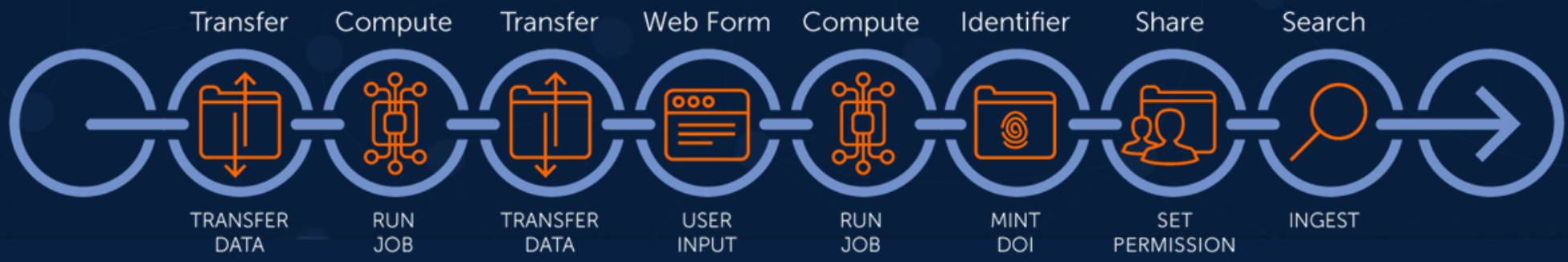
- Start from Web App, CLI, API, Python SDK
- Provide input
- Performs a series of ***actions*** with that input
- Manage the run (Evaluate; Share)



What's an Action?

An ***action*** is a special type of ***state***

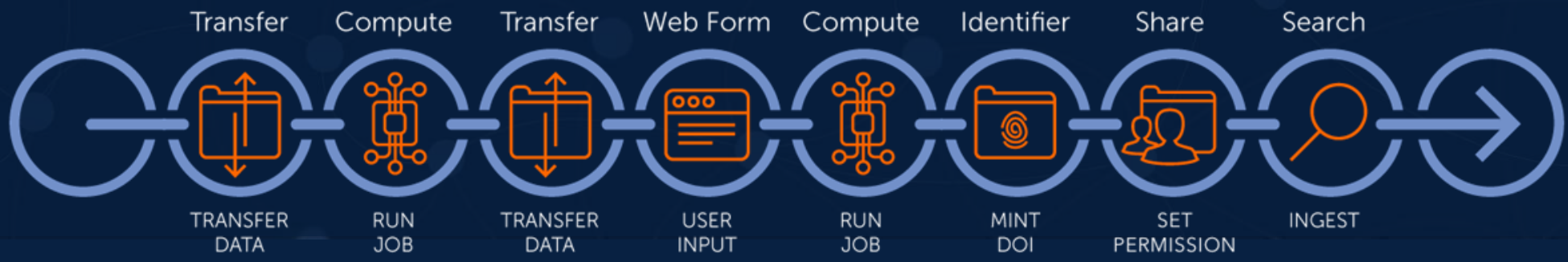
- State Machine
- AWS Step Functions
- Step Function Language
- Basic states like "Wait", "Pass", "Choice"



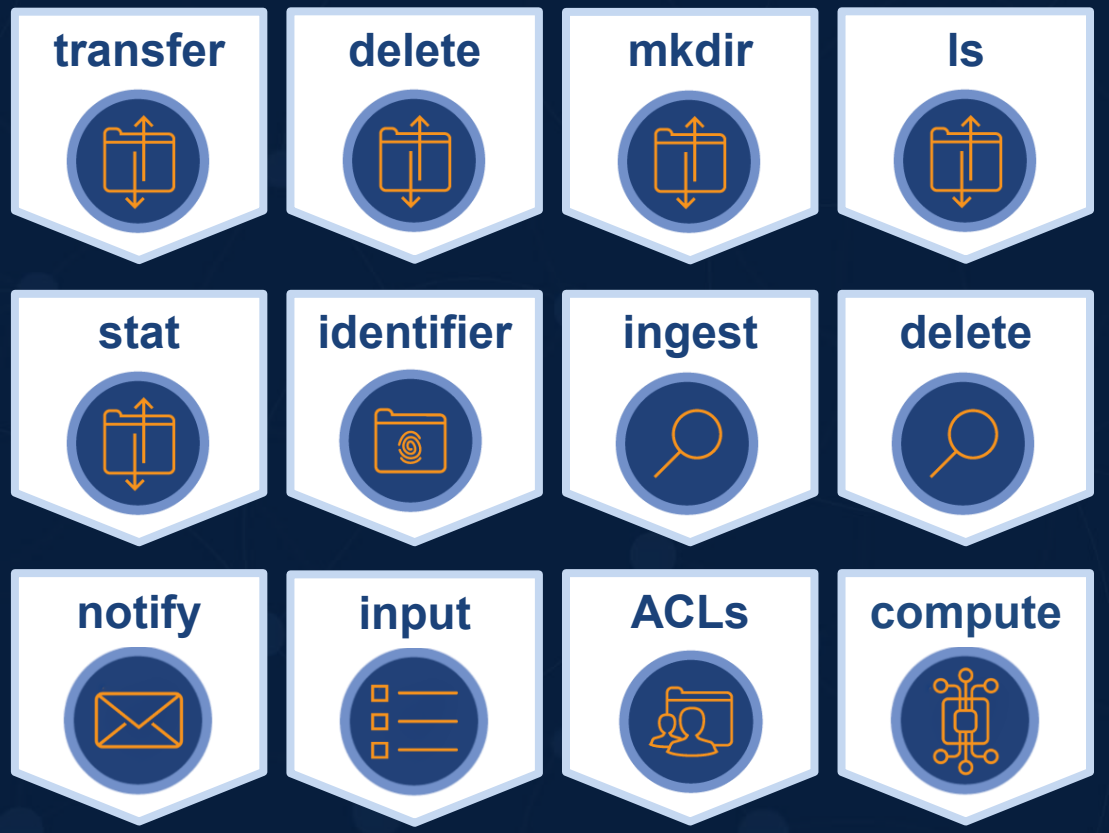
What's an Action?

…So what's the ***action*** state type?

- Operations you can perform on other resources and services
- *Managed* interactions
  - Authenticate, authorize, validate, store, auto-retry

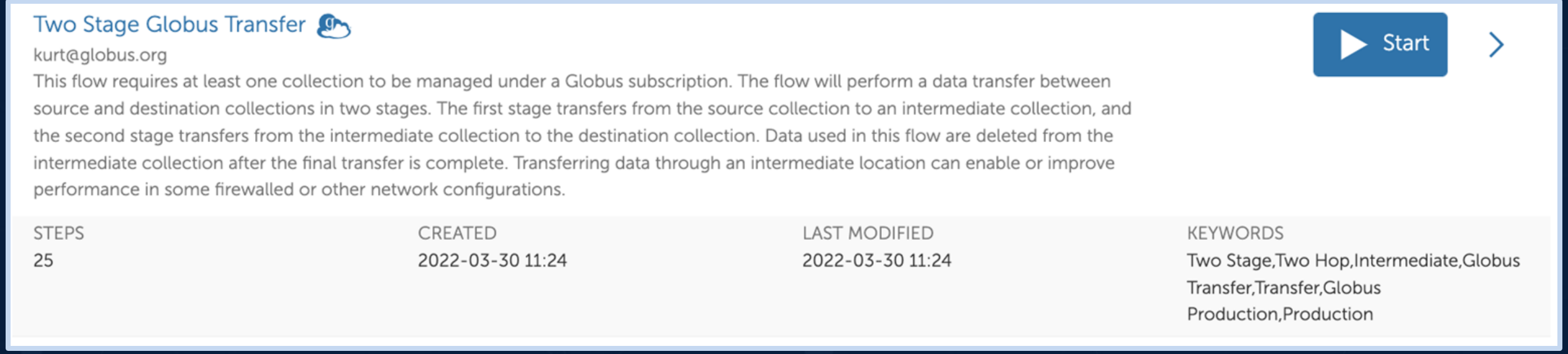


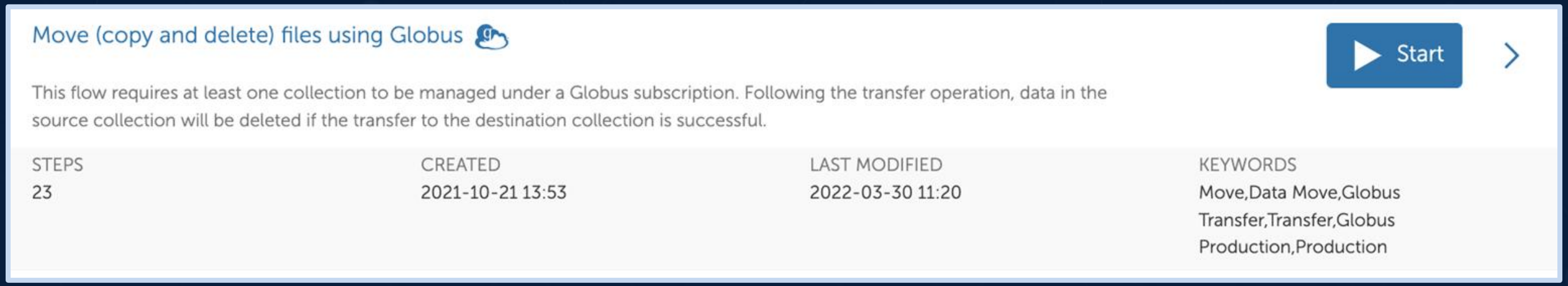
Globus-Provided Actions



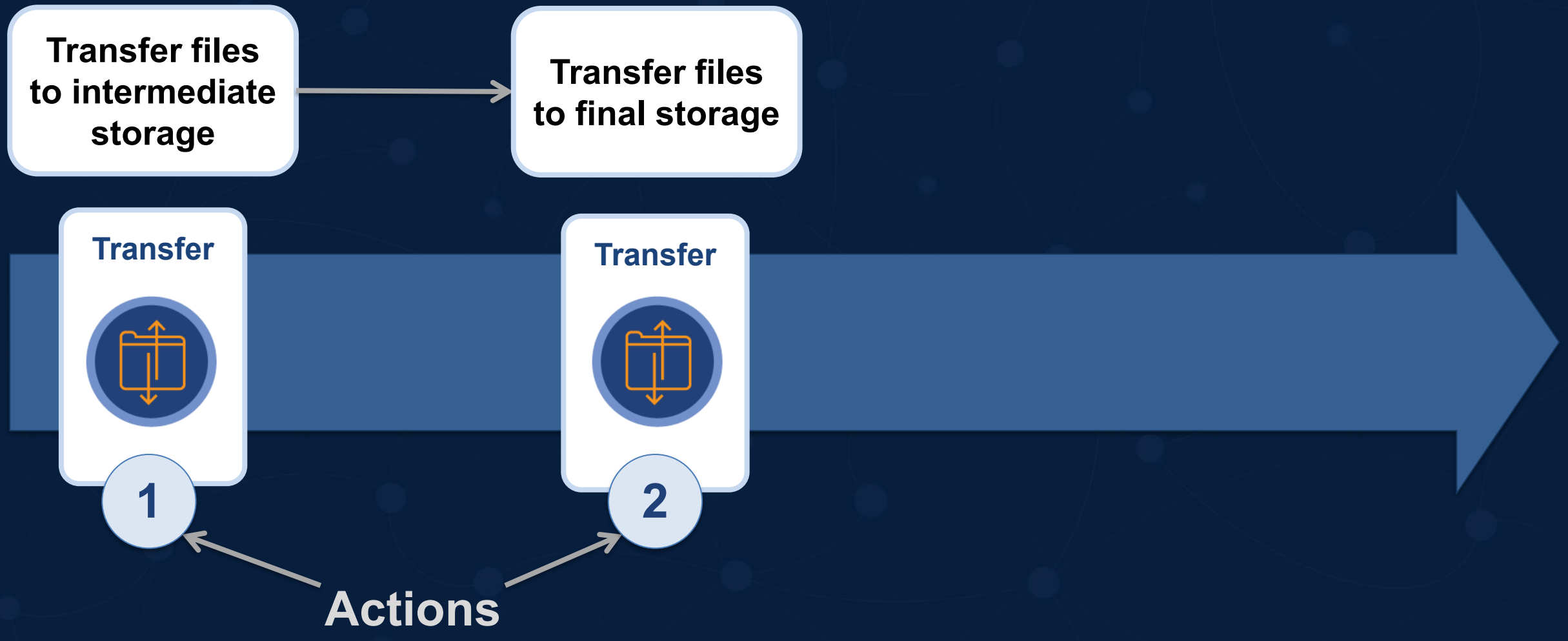
https://docs.globus.org/api/flows/hosted-action-providers/

Globus-provided flows





A simple, rather contrived, use case

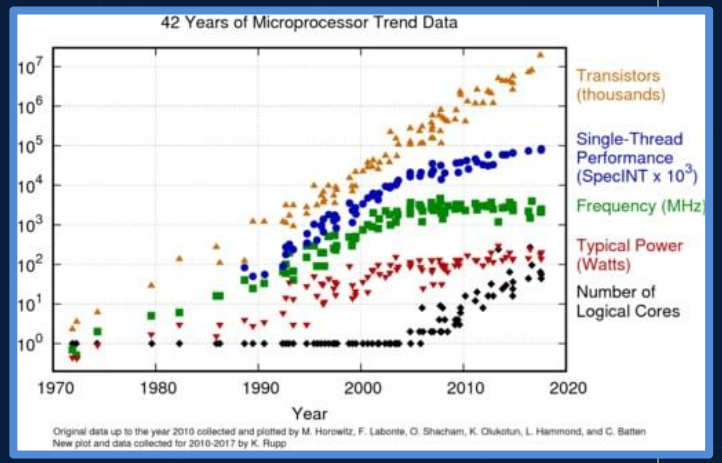


We need to do things with our data, we need computation

But research computing can be daunting

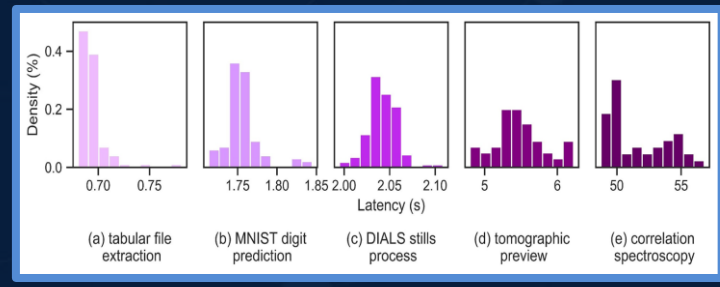
Resources

- Hardware specialization
- Specialization leads to distribution



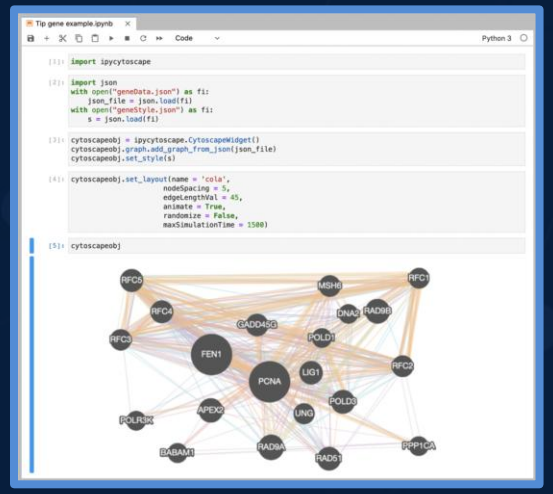
Workloads

- Interactive, real-time workloads
- Machine learning training and inference
- Components may best be executed in different places



Users

- Diverse backgrounds and expertise
- Different user interfaces (e.g., notebooks)



What does this mean for the scientist?

- **Remote computing is notoriously complicated**
  - Authentication
  - Network connections
  - Configuring/managing jobs
  - Interacting with resources (waiting in queues, scaling nodes)
  - Configuring execution environment
  - Getting results back again
- **Researchers need to overcome the same obstacles every time they move to a new resource**

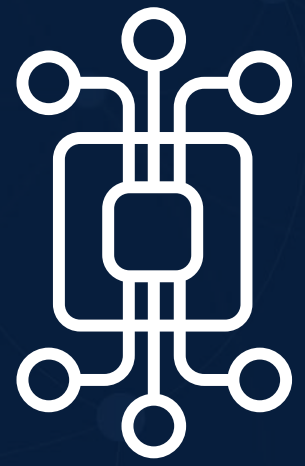
How can we bridge this gap?

Move closer to researchers' environments

- ➔Researchers primarily work in high level languages
- ➔Functions are a natural unit of computation
- ➔The Function-as-a-Service (FaaS) model allows researchers to work in a familiar language (e.g., Python) using familiar interfaces (e.g., Jupyter)

Globus Compute helps bridge the gap

Managed, federated Function-as-a-Service for reliably, scaleably and securely executing functions on remote endpoints from laptops to supercomputers



THE UNIVERSITY OF CHICAGO
ILLINOIS
Argonne NATIONAL LABORATORY

Globus Compute …FaaS on any system

- **Fire-and-forget remote computation**
- **Uniform interface, from laptop to supercomputer**
- **Federated authentication, and local access control**

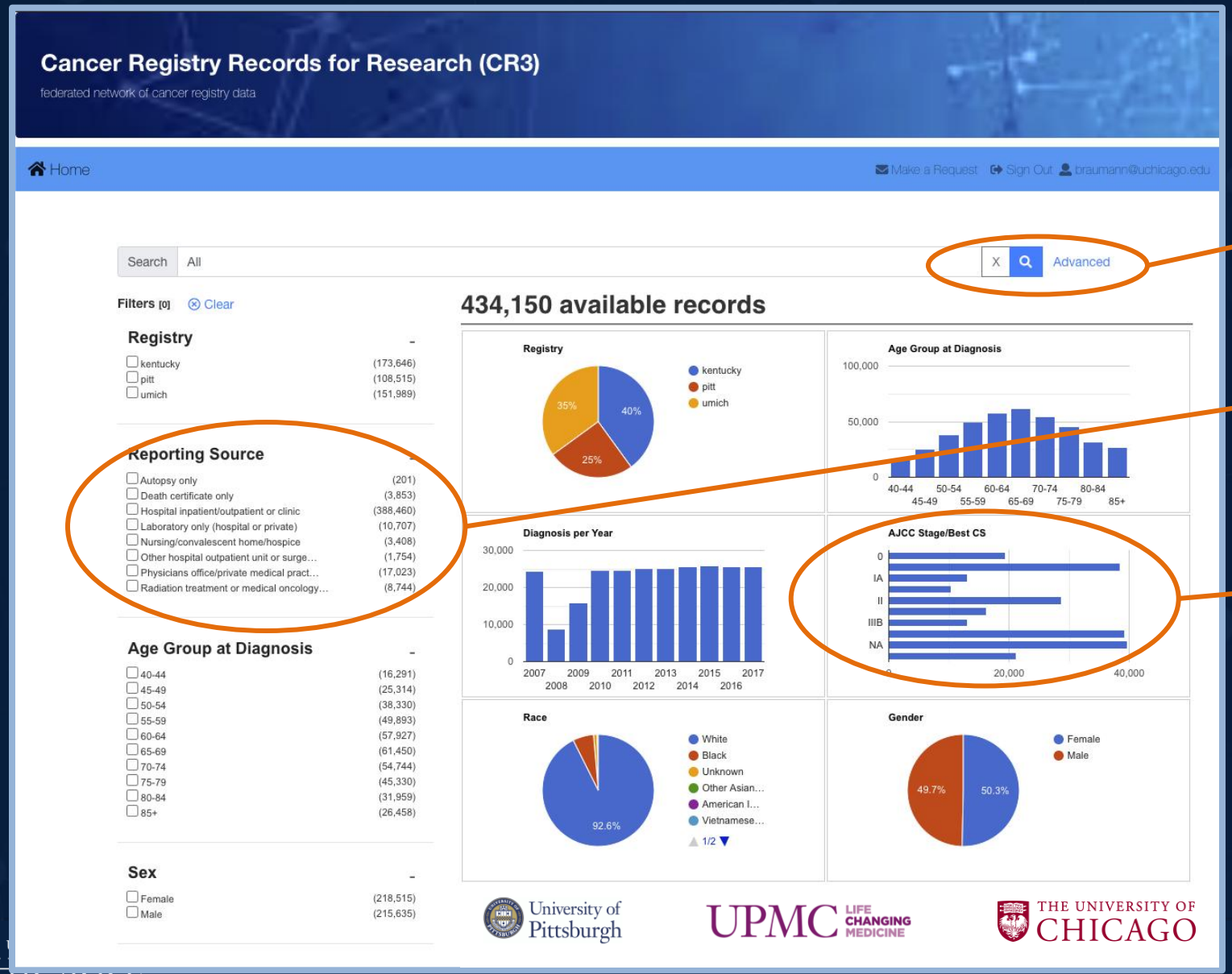
Enabling Interoperability with Globus Compute

- **Compute service — Highly available cloud-hosted service; managed fire-and-forget function execution**
- **Compute endpoint — Abstracts access to compute resources; edge device to supercomputer**
- **SDK — Python interface for interacting with the service; familiar Globus look and feel**
- **Security — Leverages Globus Auth; user authentication and identity mapped to local account**
- **Multi-user Compute endpoints (MEP) enable administrators to securely offer compute resources to users without mandating the typical shell access (i.e., ssh).**

Data description and discovery with Globus Search

- Metadata store with fine-grained visibility controls
- Schema agnostic - dynamic schemas
- Simple search using URL query parameters
- Complex search using search request document

Data Portals – Search and discovery in action



Google-like text search with facets for filtering

Variable facets based on source registry index

Dynamically updating charts as facets change

Developed using a framework based on the Globus Modern Research Data Portal* design pattern (**docs.globus.org/mrdp**)

* PeerJ Articles:cs-144 https://peerj.com/articles/cs-144/

Use case: cryoEM automation

Globus Flows

Transfer

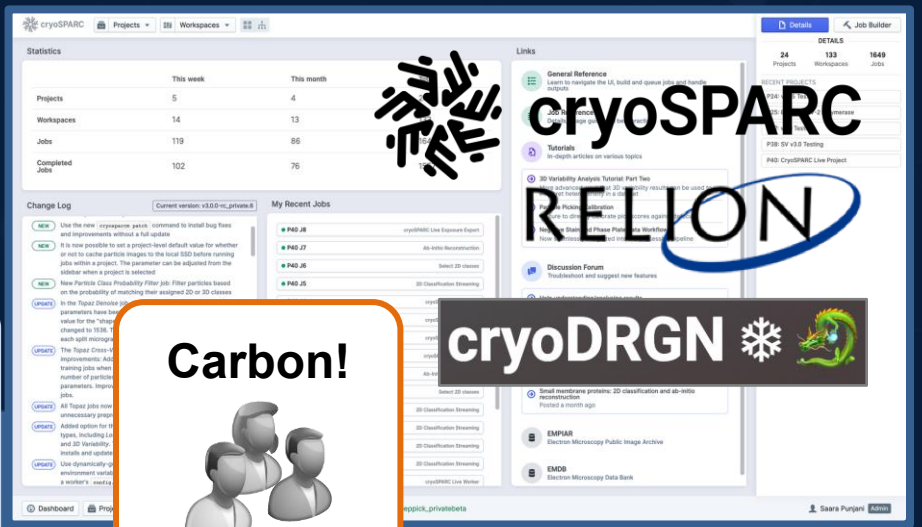
Transfer raw files

Compute

Launch analysis job

Carbon!

Correct, classify, …



Compute

Extract metadata

Search

Ingest to index

Transfer

Move final files to repo

Share

Set access controls

https://www.nlm.nih.gov/oet/ed/cde/tutorial/img/06aCDE.png

Globus – FAIR by default

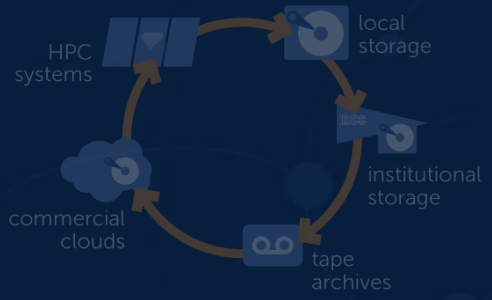
Reliable automation

Publication & discovery

Software-as-a-Service

Data sharing

Managed remote execution



Unified data access

Managed transfer & sync

Platform-as-a-Service

FAIR **Data Practices and Processes**

- Findability
  - Provide methods for metadata ingest, indexing, the ability to search quickly and a way to create lasting identifiers - Search
- Accessibility
  - Limit barriers to data but provide authentication and authorization when necessary – Transfer - Share
- Interoperability
  - Data processing operations should fit as universally as possible to workflows, analysis and storage – Flows - Compute
- Reusability
  - Curated from acquisition to publication to discovery

Support resources

- Globus documentation: docs.globus.org
- YouTube channel: youtube.com/GlobusOnline
- Helpdesk: support@globus.org
- Mailing Lists: globus.org/mailing-lists
- Customer engagement team (office hours)
- Professional services team (advisory, custom work)

Globus – A Platform for Research Data Management Challenges

ARGONNE ATPESC2025 EXTREME-SCALE COMPUTING

Lev Gorenstein
lev@globus.org

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

globus