

Globus: The Platform for Research Data Management

Vas Vasiliadis — vas@uchicago.edu





Globus is ...

a non-profit service
developed and operated by



THE UNIVERSITY OF
CHICAGO



Our team comprises ...

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



Our mission is to...

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

300,000
REGISTERED
USERS

1,400+
IDENTITY
PROVIDERS

LOCAL
STORAGE

RCC
INSTITUTIONAL
STORAGE



MODULAR
APPS

USERS IN
80+
COUNTRIES

SOFTWARE AS A SERVICE
PLATFORM AS A SERVICE



globus

39,000
ACTIVE
ENDPOINTS

TAPE
ARCHIVES

HIGH
PERFORMANCE
COMPUTING

RELIABLE
TRANSFER
1PB
PER DAY

1,600+
CONNECTED
INSTITUTIONS

10,000+
ACTIVE SHARED
ENDPOINTS

COMMERCIAL
CLOUD
STORAGE

Numbers reflect the 12-month period ended 3/31/2022



Development is funded by...



U.S. DEPARTMENT OF
ENERGY



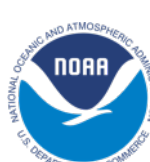
THE UNIVERSITY OF
CHICAGO



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

Argonne 
NATIONAL LABORATORY

Operations are funded by subscribers



Sustainability enabled by freemium model

- Basic file transfer free to non-profit research institutions
- Annual subscriptions provide advanced capabilities to researchers and visibility/control to administrators



80% US Institutions



20% International Institutions



13% Federal Agencies/
National Facilities



46% R1 Universities



13% Independent
Research Institutes



4% Hospital/Health
Care Systems



20% Other
Universities



4% Commercial



Globus is known for reliable, secure file transfer

Activity List ✓ RDA to ALCF noverify
transfer completed

Overview Event Log

72.8Gbps

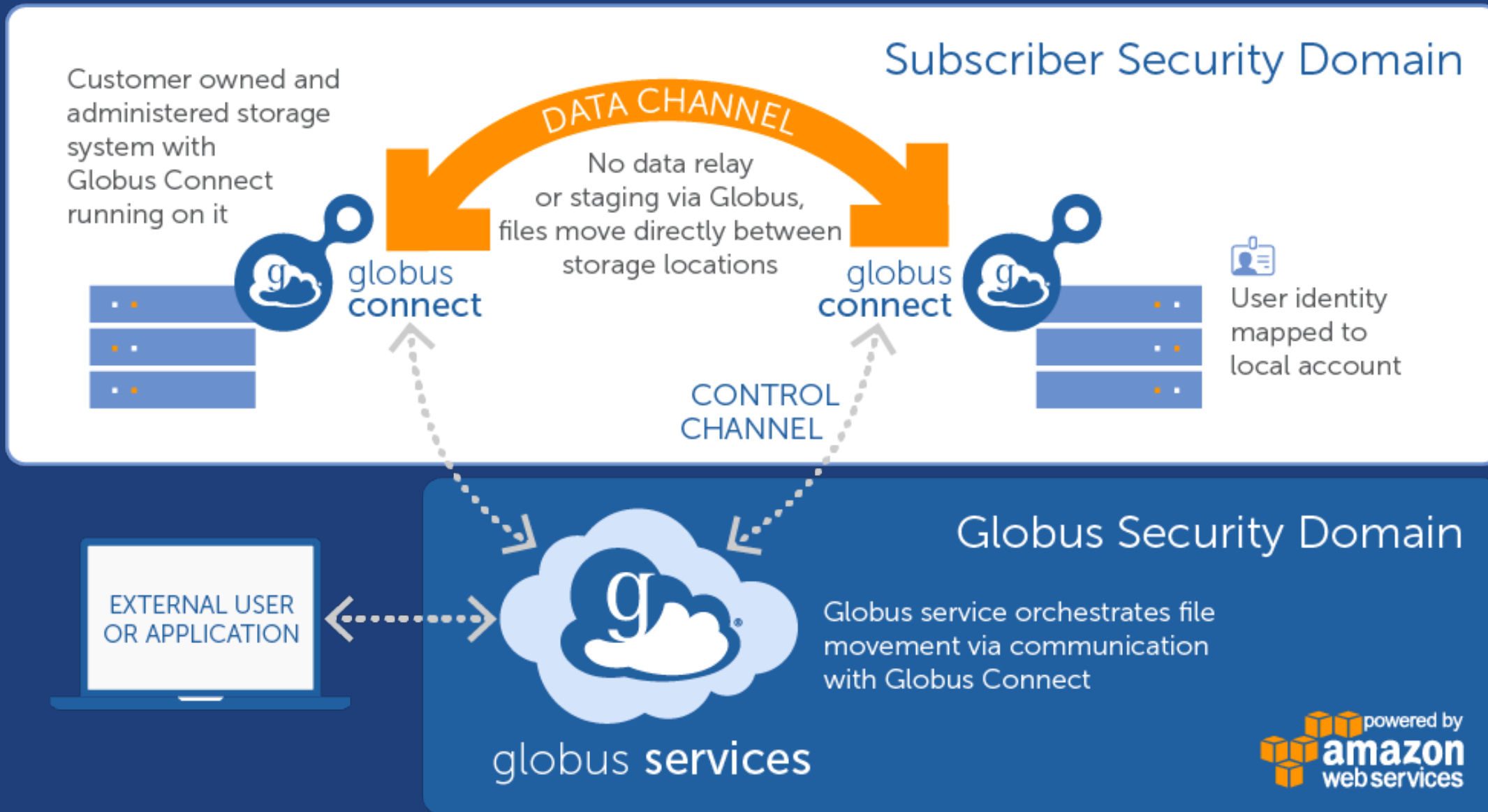
Task Label	RDA to ALCF noverify
Source	NCAR RDA Dataset Archive ⓘ
Destination	DME PerfTest - Argonne ⓘ
Task ID	20ebf766-a46d-11eb-8a95-d70d98a40c8d
Owner	Vas Vasiliadis (vas@globusid.org)
Condition	SUCCEEDED
Requested	2021-04-23 02:50 pm
Completed	2021-04-23 02:53 pm
Duration	2 minutes 47 seconds
Transfer Settings	<ul style="list-style-type: none">transfer is not encryptedoverwriting all files on destination

6151	Files
2	Directories
1.51 TB	Bytes Transferred
9.10 GB/s	Effective Speed
0	Skipped files on sync
0	Skipped files on error

[View debug data](#)



Conceptual Architecture: Hybrid SaaS





Globus Conenct unifies data access across diverse landscape of storage systems...



“I need to easily, securely and reliably move or replicate my data between systems.”





Currently supported systems



Google Cloud

Microsoft Azure
Blob Storage



SCALITY



IBM Cloud



Google Drive



IBM Spectrum Scale



iRODS®

Quantum® ACTIVE SCALE™



HPSS



wasabi®
hot cloud storage

lustre™



We enable a “data appropriate” storage strategy

- **Access frequency (data temperature)**
- **Access modality(ies)**
 - Ad hoc, via web browser
 - Scripted, via CLI tools
 - Programmatic, via APIs
- **Uniform interface and user experience**



Google
Cloud Storage



Google Drive

Microsoft Azure
Blob Storage



Microsoft OneDrive



amazon
S3



Uniform interface, consistent user experience

The screenshot displays the Globus interface with four collection panes, each showing a list of folders and files. The panes are:

- Vas Google Cloud Storage Collection**: Shows folders like 'globus-archive' and 'uchicago-public'.
- Vas Google Drive Collection**: Shows folders like '211207_SciDAS_Panel_Enabling_F_and_A_Va...', 'CC Spring 2022', 'globus', 'migration', 'MPCS 51083-1 Spring 2021', 'MPCS 51083-1 Winter 2022', 'MPCS 51240 - Product Management', and 'Product Management - Interim Feedback'.
- Vas Box Collection**: Shows folders like 'globus-box-archive', 'uchicago-perftest', and 'warchive'.
- Vas AWS S3 Collection**: Shows folders like 'globus-vault', 'mpcs-practicum', and 'personal-vault'.

Each pane has a search bar, a path field, and a 'Start' button. The interface is consistent across all panes, with a dark navigation bar at the top of each pane containing icons for home, back, forward, refresh, and settings.



Move without (worrying about) limits

- **API request rates**
- **Data volume**
- **Third-party tools cannot circumvent these...**
- **...but you can use Globus to “fire-and-forget”**
- **→ it will (eventually) be done**
- **File size ...see “Data appropriate” above :-)**

```
3/9/2022, 08:33 PM endpoint too busy View details ^  
  
Error (transfer)  
Endpoint: Vas Google Drive Collection (d6d62391-fdda-4ba5-ac78-6523f806ea79)  
Server: m-422a8b.d8b83.36fe.data.globus.org:443  
File: /My%20Drive/migration/uchicago-perftest/cc32-16p32-16/test1185  
Command: STOR /My Drive/migration/uchicago-perftest/cc32-16p32-16/test1185  
Message: Fatal FTP response  
---  
Details: 451-GlobusError: v=1 c=TOO_BUSY\r\n451-GlobusError: v=1 c=INTERNAL_ERROR\r\n451-\r\n451-GD-Method: "PATCH"\r\n451-GD-URI:  
"https://www.googleapis.com/upload/drive/v3/files/11h9NBppR7qG7YaZDqyWi-8w9rzw9gGj"\r\n451-GD-Response-Code: "403"\r\n451-GD-Response: "User rate limit  
exceeded."\r\n451 End.\r\n
```

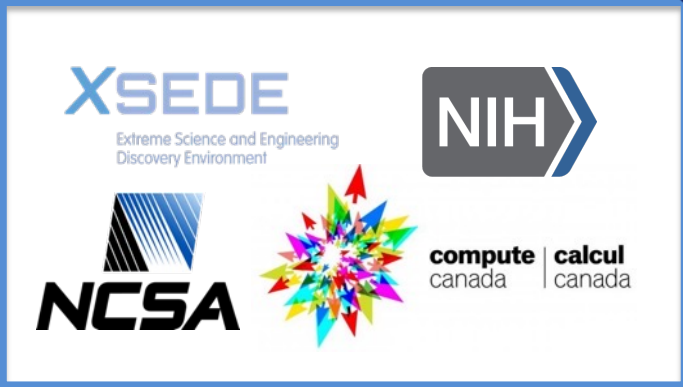
We simplify secure sharing with collaborators...



UCSF



On-premises stores



XSEDE
Extreme Science and Engineering
Discovery Environment

NIH

NCSA

compute canada | calcul canada



Public / private cloud stores

Project repositories,
replication stores



Public repositories



Microsoft Azure
Blob Storage

aws | S3 | ceph | openstack

Google Cloud





...help researchers manage instrument data...



Next-Gen Sequencer



Advanced Light Source



Cryo-EM



MRI



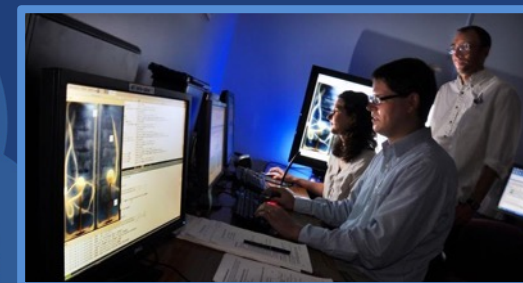
Light Sheet Microscope



Analysis store



High-durability, low-cost store



Remote visualization



Personal system





...and build data-centric applications

Sanger Imputation Service **Beta** Home About Instructions Resources Status

Sanger Imputation Service

This is a free genotype **imputation** and **phasing** service provided by the Wellcome Trust. You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genotypes. To learn more and follow us on Twitter.

Before you start **Ready to start?** **News**

Cancer Registry Records for Research (CR3)

federated network of cancer registry data

Home Make a Request Sign Out traumann@uchicago.edu

Search All X Q Advanced

Filters [0] Clear

Registry

- SEER Registry (173,646)
- Medical Center Registry (108,515)
- State Registry (151,989)

Reporting Source

- Autopsy only (201)
- Death certificate only (3,853)
- Hospital inpatient/outpatient or clinic (388,460)
- Laboratory only (hospital or private) (10,707)
- Nursing/convalescent home/hospice (3,408)
- Other hospital outpatient unit or surge... (1,754)
- Physicians office/private medical practice... (17,023)
- Radiation treatment or medical oncology... (8,744)

Age Group at Diagnosis

- 40-44 (16,291)
- 45-49 (25,314)

434,150 available records

Registry

Registry	Percentage
SEER Registry	40%
Medical Center Registry	35%
State Registry	25%

Age Group at Diagnosis

Diagnosis per Year

AJCC Stage/Best CS

HuBMAP

Donors Samples Datasets

Datasets

433 results found

Dataset	Group	Data Types
HBM454.RMLS.428	Vanderbilt TMC	MALDI IMS positive [Pyramid]
HBM638.GFJG.839	University of California San Diego TMC	scATAC-seq (SNARE-seq) [Processed]
HBM437.KPNV.984	University of California San Diego TMC	scATAC-seq (SNARE-seq) [Lab Processed] Kidney (Right) Published 2020-09-12 15:48:09
HBM595.QDQD.996	University of California San Diego TMC	scRNA-seq (SNARE-seq) [Lab Processed] Kidney (Right) Published 2020-09-12 15:48:09

Dataset Metadata

Data Type

- SNARE-seq 72
- Untargeted LC-MS 50
- CODEX 26
- CODEX [Cytokit + SPRM] 26
- Autofluorescence Microscopy 19

Organ

- Kidney (Left) 120
- Kidney (Right) 94
- Small Intestine 52

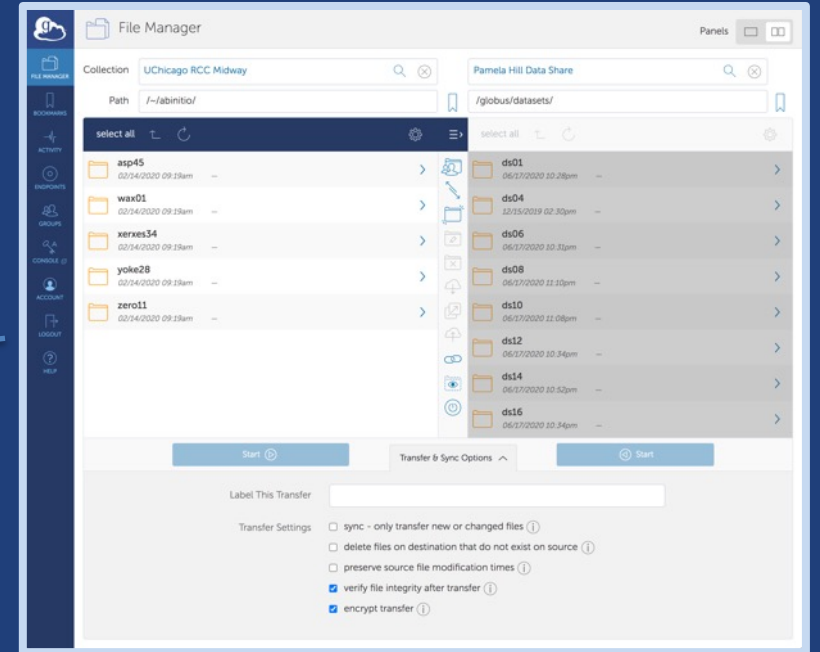


Use(r)-appropriate interfaces

Globus service



Web



CLI

Platform
(RESTful APIs)

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

```
Usage: globus [OPTIONS] COMMAND [ARGS]...

Options:
  -v, --verbose           Control level of output
  -h, --help             Show this message and exit.
  -F, --format [unix|json|text] Output format for stdout. Defaults to text
  --jmespath, --jq TEXT  A JMESPath expression to apply to json
                        output. Takes precedence over any specified '
                        --format' and forces the format to be json
                        processed by this expression
  --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        Manage your Globus config file. (Advanced Users)
  delete        Submit a delete task (asynchronous)
  endpoint      Manage Globus endpoint definitions
  get-identities Lookup Globus Auth Identities
  list-commands List all CLI Commands
  login         Log into Globus to get credentials for the Globus CLI
  logout        Logout of the Globus CLI
  ls            List endpoint directory contents
  mkdir         Make a directory on an endpoint
  rename        Rename a file or directory on an endpoint
  rm            Delete a single path; wait for it to complete
  session       Manage your CLI auth session
  task          Manage asynchronous tasks
  transfer      Submit a transfer task (asynchronous)
  update        Update the Globus CLI to its latest version
  version       Show the version and exit
  whoami        Show the currently logged-in primary identity.
```



- Custom portals
- Science Gateways
- Unique workflows
- Any research application...





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 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**

Globus for protected data management

Security controls

- NIST 800-53
- 800-171 Low+



Restricted data handling

- PHI, PII, CUI
- Compliant data sharing

BAA w/Uchicago

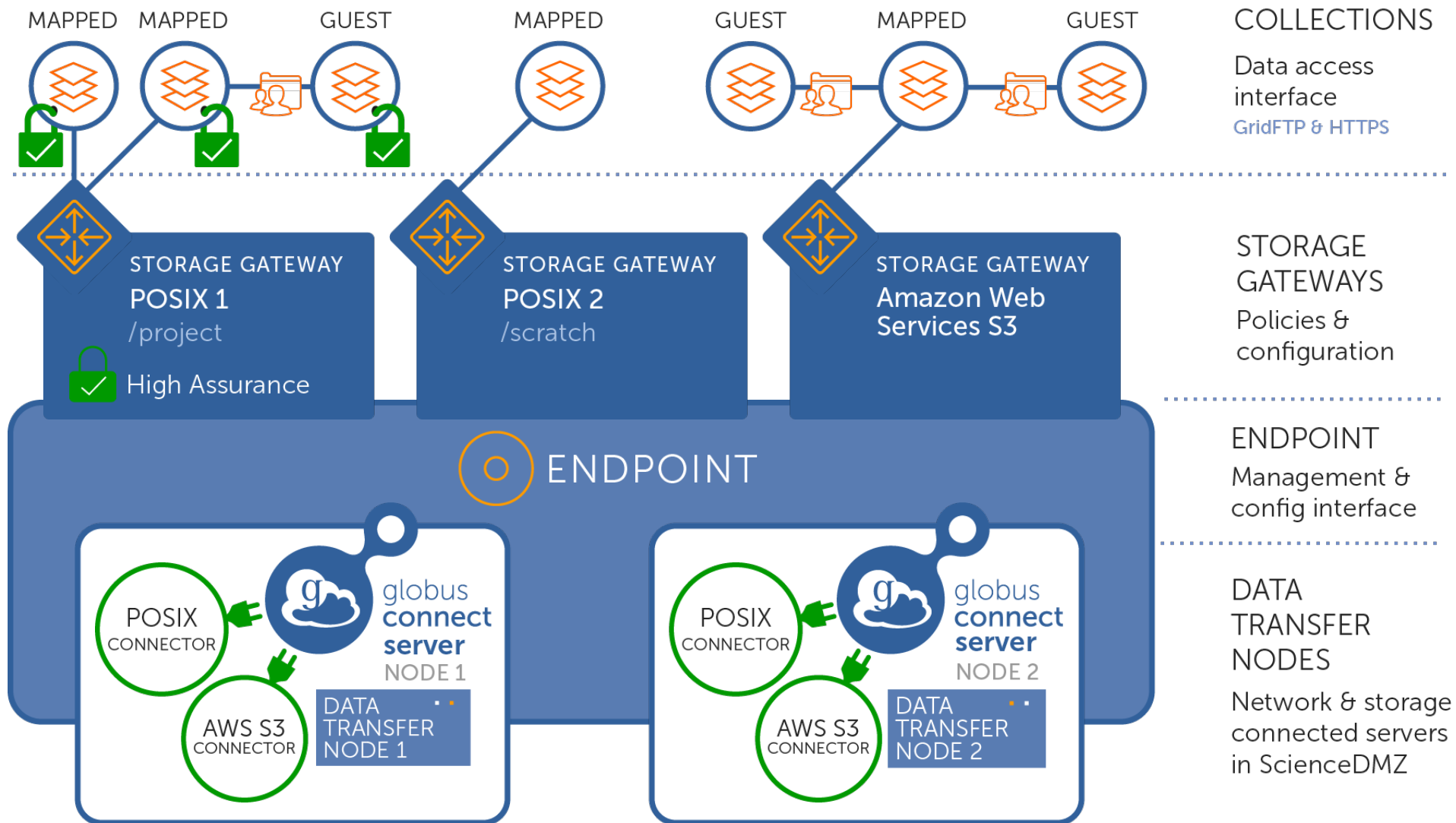
- UChicago BAA with Amazon



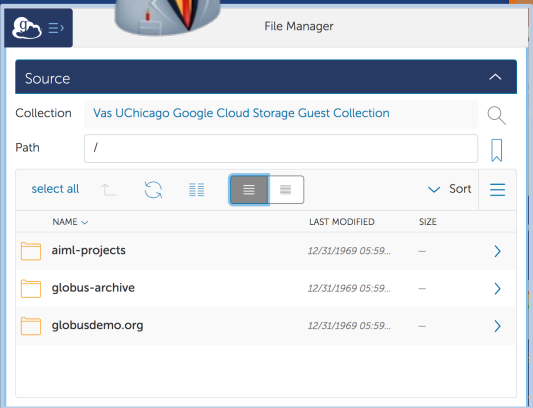
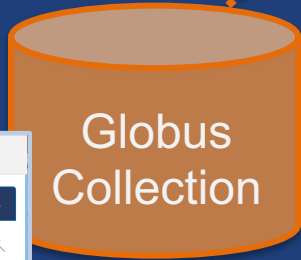
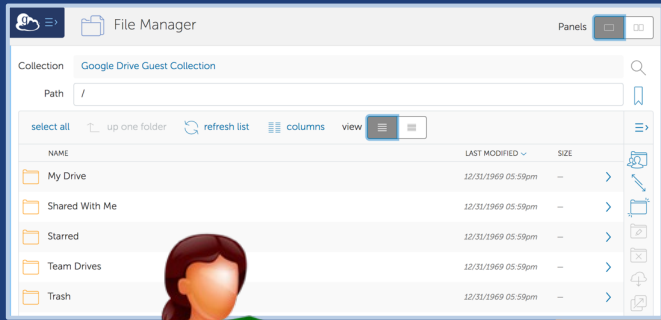
Let's take a look...



Globus Connect Server architecture



Connectors/gateways for cloud storage



Globus endpoint



Data



Other Globus endpoints



Cloud Storage Gateways (using Globus connectors)



Control





Deployment varies by cloud provider, but has common elements

- **Provider configuration**
 - Register an application
 - Get app credentials
- **Storage gateway creation**
 - Use app credentials
 - Specify access policy
- **Mapped collection creation**
 - User account \leftrightarrow Globus identity mapping
- **Guest collections shield users from complexity**

Google Cloud Platform GCS Connector Demonstration Search Products, resources, docs (/)

API APIs & Services Client ID for Web application DOWNLOAD JSON RESET SECRET

Enabled APIs & services
Library
Credentials
OAuth consent screen
Domain verification
Page usage agreements

Name *
Globus Connect Server v5.4 Gateways

The name of your OAuth 2.0 client. This name is only used to identify the client in the console and will not be shown to end users.

The domains of the URIs you add below will be automatically added to your OAuth consent screen as authorized domains.

Authorized JavaScript origins ⓘ
For use with requests from a browser

URIs 1 *
https://75869.7567.data.globus.org

URIs 2 *
https://d65b7.ac45.data.globus.org

+ ADD URI



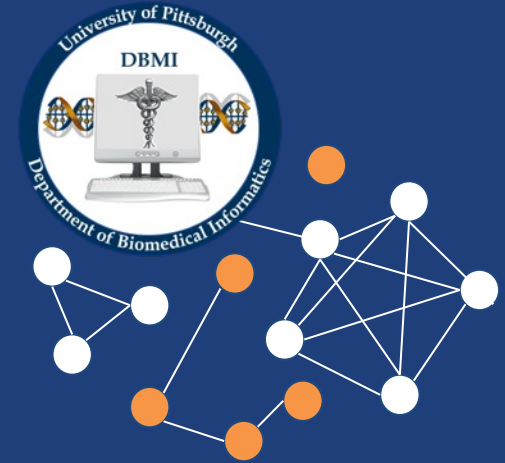
Globus Search and Portal Framework

Example: Advanced Photon Source

acdc.alcf.anl.gov

Cancer Registry Records for Research (CR3)

- **Create network of federated cancer registries**
 - Deploy similar infrastructure at other cancer registries
 - Enable queries across multiple registries
- **Federation via Globus: network scale \leftrightarrow local control**
 - Data owners input/export data sets, apply QC, set access policies
 - Registry data remain at the institution where they were generated
 - Identities are provided/authenticated by the institution, not Globus
 - System scale depends on data owners providing storage resources



Federated logon using Globus Auth with Pitt/UPMC as identity providers

Cancer Registry Records for Research (CR3)
federated network of cancer registry data

Home Make a Request Sign Out braumann@uchicago.edu

Search All X Q Advanced

Filters [0] Clear

Registry

- kentucky (173,646)
- pitt (108,515)
- umich (151,989)

Reporting Source

- Autopsy only (201)
- Death certificate only (3,853)
- Hospital inpatient/outpatient or clinic (388,460)
- Laboratory only (hospital or private) (10,707)
- Nursing/convalescent home/hospice (3,408)
- Other hospital outpatient unit or surge... (1,754)
- Physicians office/private medical pract... (17,023)
- Radiation treatment or medical oncology... (8,744)

Age Group at Diagnosis

- 40-44 (16,291)
- 45-49 (25,314)
- 50-54 (38,330)
- 55-59 (49,893)
- 60-64 (57,927)
- 65-69 (61,450)
- 70-74 (54,744)
- 75-79 (45,330)
- 80-84 (31,959)
- 85+ (26,458)

Registry

Registry	Percentage
kentucky	40%
pitt	25%
umich	35%

Age Group at Diagnosis

Age Group	Count
40-44	16,291
45-49	25,314
50-54	38,330
55-59	49,893
60-64	57,927
65-69	61,450
70-74	54,744
75-79	45,330
80-84	31,959
85+	26,458

Diagnosis per Year

Year	Count
2007	24,000
2008	10,000
2009	15,000
2010	24,000
2011	24,000
2012	24,000
2013	24,000
2014	24,000
2015	24,000
2016	24,000
2017	24,000

AJCC Stage/Best CS

Stage/Best CS	Count
0	20,000
IA	35,000
II	25,000
IIIB	15,000
NA	35,000

Race

Race	Percentage
White	92.6%
Black	~7.4%
Unknown	~0%
Other Asian...	~0%
American I...	~0%
Vietnamese...	~0%

Gender

Gender	Percentage
Female	50.3%
Male	49.7%

Sex

- Female (218,515)
- Male (215,635)

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/>



Current Focus...

Automating data management at scale



Timer Service

Scheduled and recurring transfers
(*a.k.a. Globus cron*)

Command Line Interface

Ad hoc scripting and integration



Globus Flows service

Comprehensive task (data and compute) orchestration with human in the loop interactions

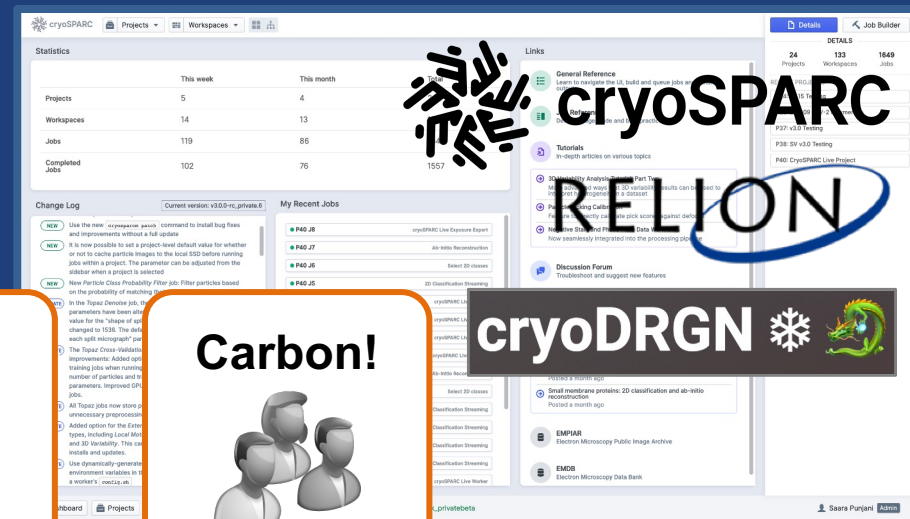
The Globus Timer service

- **Scheduled/recurring file transfers**
- **Supports all Globus transfer and sync options**
- **Service with a command line interface**
- **Example: NIH – hpc.nih.gov/storage/globus_cron.html**





Automating Cryo-EM Flows



cryoSPARC

RELION



**Globus
Flows**



Auth



Get
credentials

Transfer



Transfer
raw files

funcX



Launch
analysis job

Carbon!



Correct,
classify, ...

funcX



Extract
metadata

Search



Search,
discover,
reuse

Share



Set access
controls

Transfer

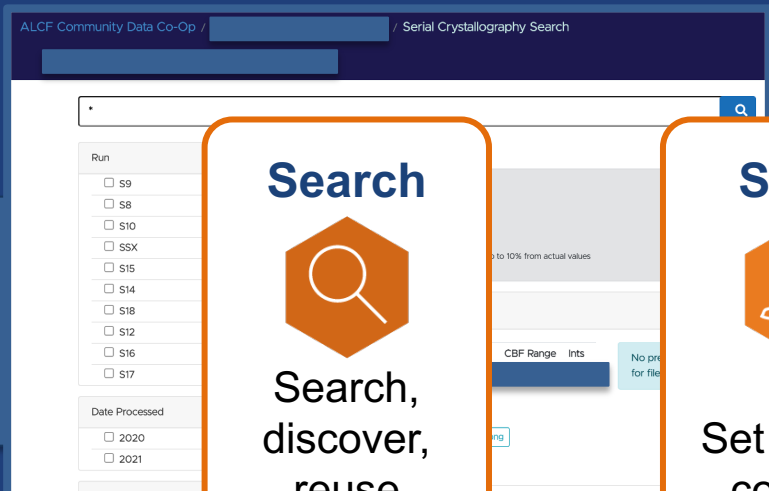


Move final
files to repo

Search



Index
ingest

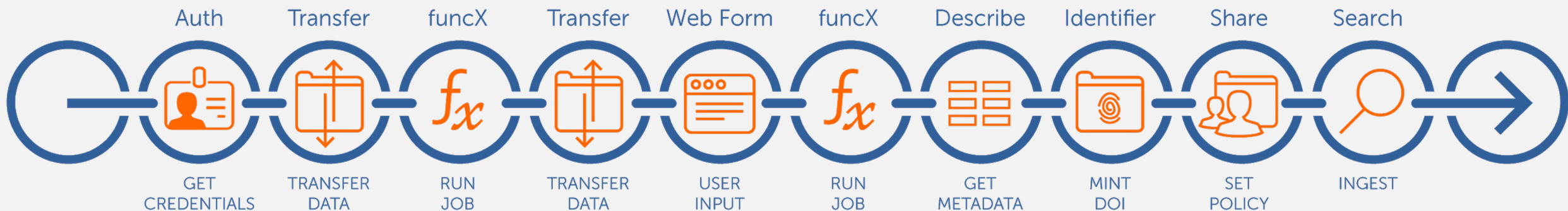




Managed automation of tasks

- **Flows:** A platform service for defining, applying, and sharing distributed research automation flows
- Flows comprise **Actions**
- **Action Providers:** Called by Flows to perform tasks
- **Triggers*:** Start flows based on events

* In development





docs.globus.org

globus.org/connectors

globus.org/subscriptions

outreach@globus.org

support@globus.org