

# Globus Coping Skills: Taming the Instrument Monster and Cloud Storage



**Vas Vasiliadis**

vas@uchicago.edu, vasv@anl.gov

September 28, 2022





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



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

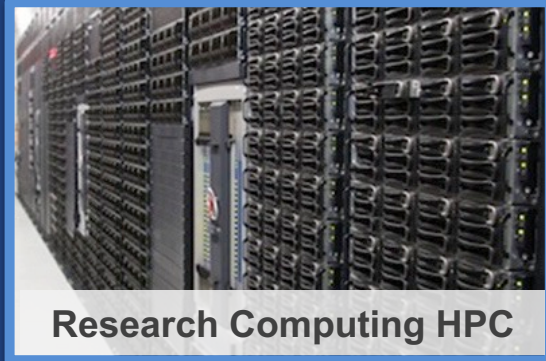
Argonne   
NATIONAL LABORATORY



# Operations are funded by subscribers



# We unify data access across disparate systems...



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



# Fast, reliable file transfer ...from any to any system

- Fire-and-forget transfers
- Optimized speed
- Assured reliability
- Unified view of storage
- Use existing credentials

Activity List ✓ RDA to ALCF noverify  
transfer completed

Overview Event Log

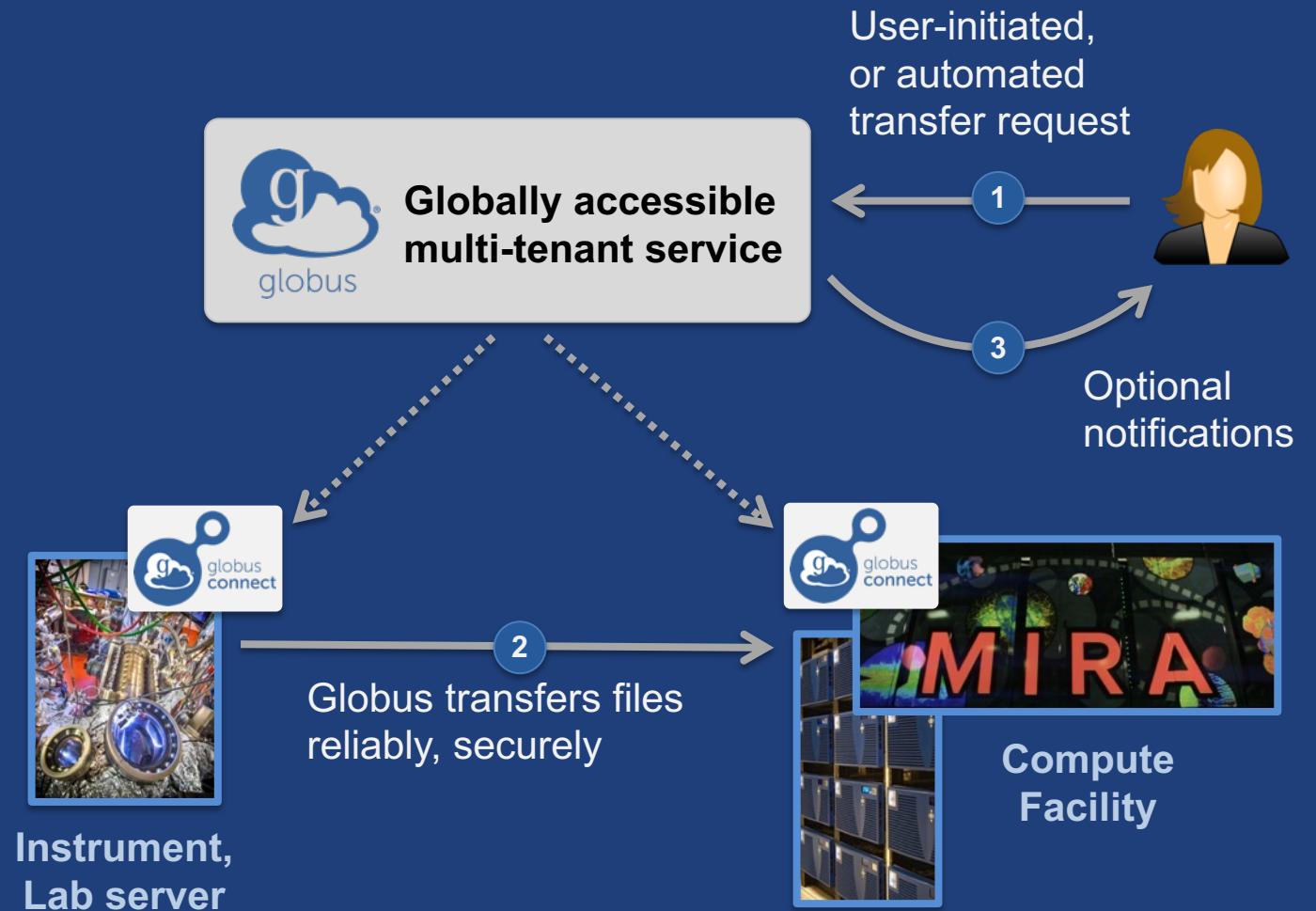
**72.8Gbps**

Task Label	RDA to ALCF noverify	6151	Files
Source	NCAR RDA Dataset Archive	2	Directories
Destination	DME PerfTest - Argonne	1.51 TB	Bytes Transferred
Task ID	20ebf766-a46d-11eb-8a95-d70d98a40c8d	9.10 GB/s	Effective Speed
Owner	Vas Vasiliadis (vas@globusid.org)	0	Skipped files on sync
Condition	SUCCEEDED	0	Skipped files on error
Requested	2021-04-23 02:50 pm		
Completed	2021-04-23 02:53 pm		
Duration	2 minutes 47 seconds		

Transfer Settings

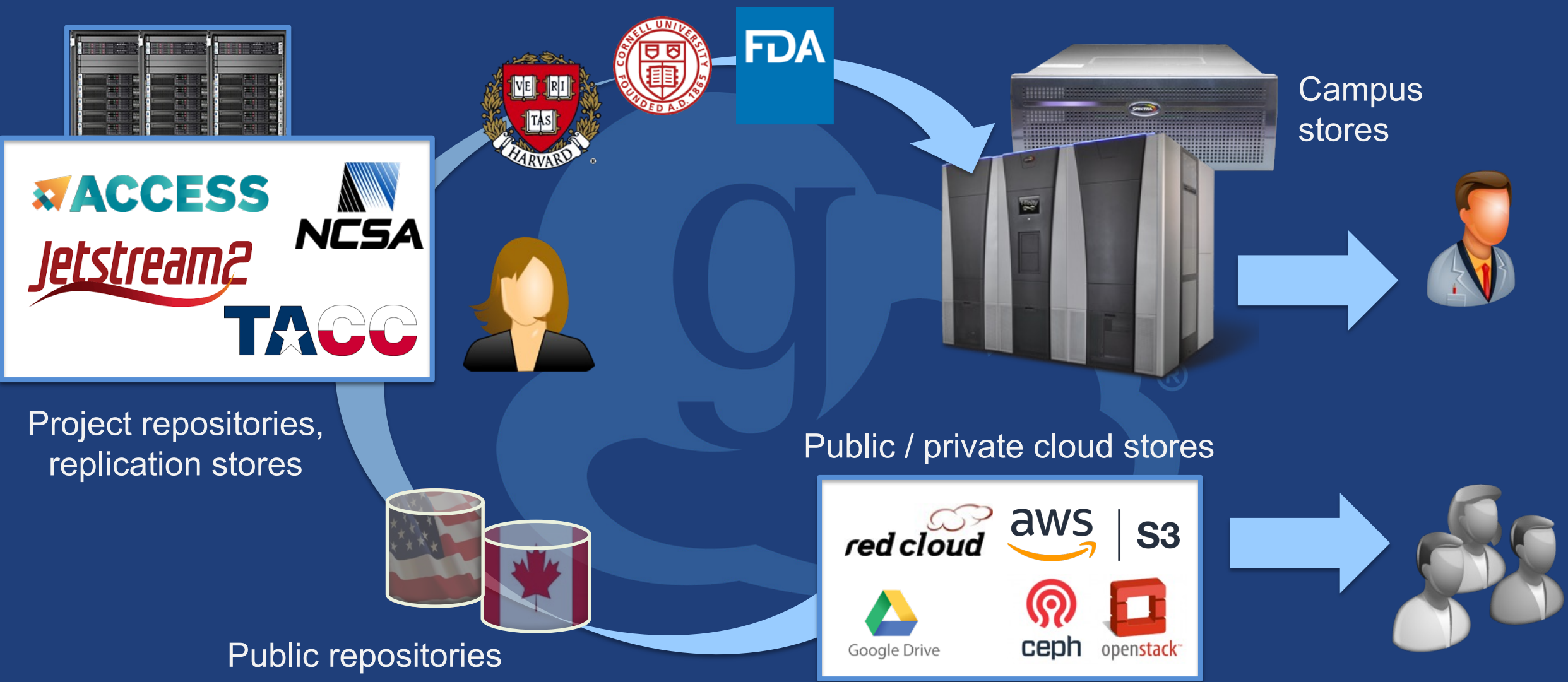
- transfer is not encrypted
- overwriting all files on destination

[View debug data](#)





# g ...simplify secure sharing with collaborators...

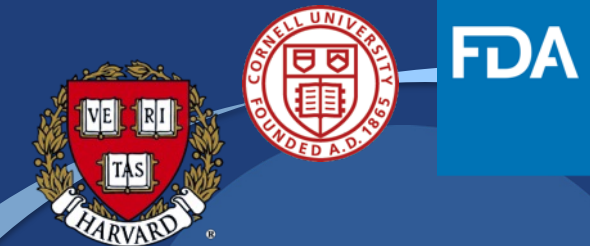


Project repositories, replication stores

ACCESS  
Jetstream2  
NCSA  
TACC



Public repositories



Campus stores

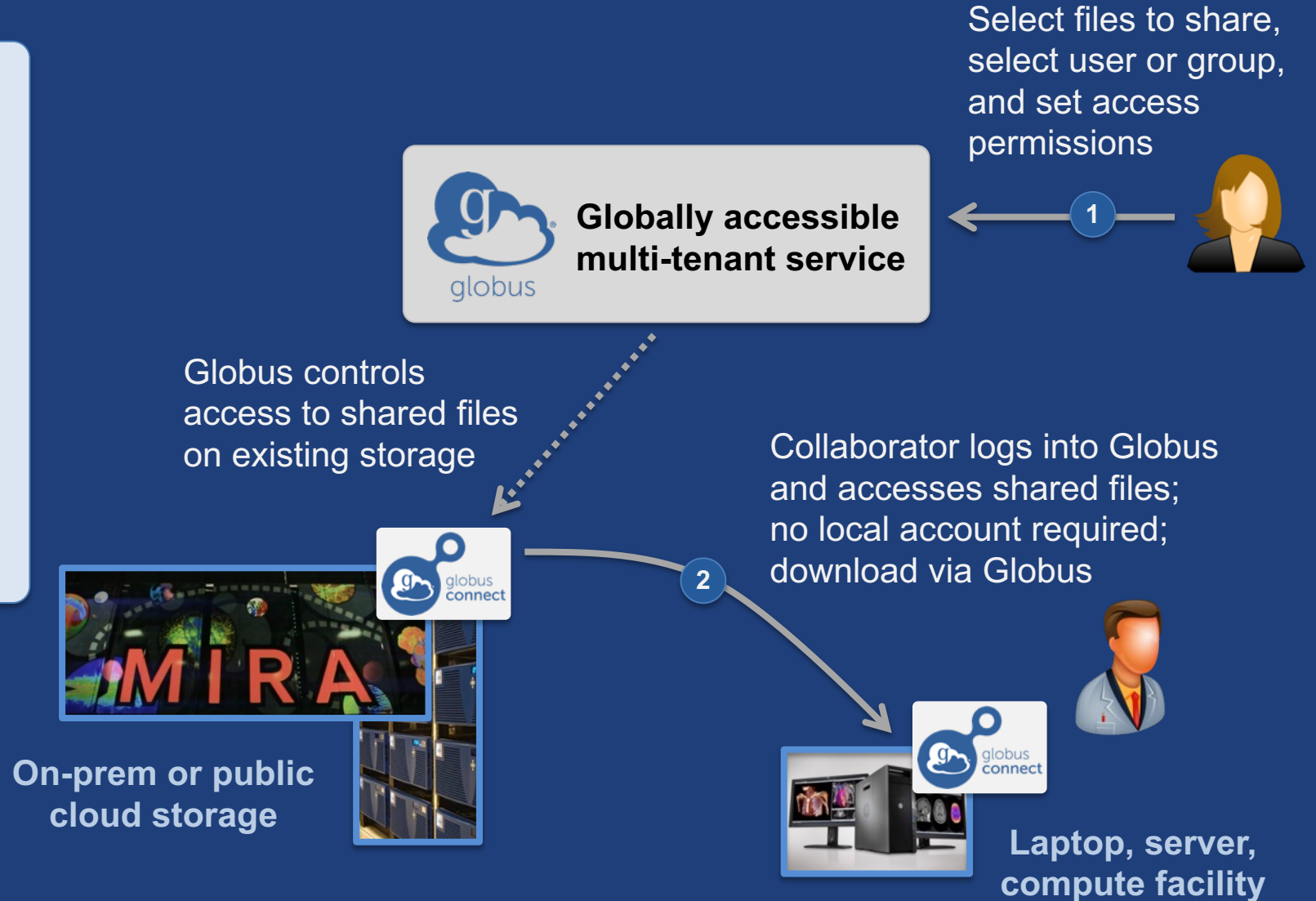
Public / private cloud stores

red cloud | aws | S3  
Google Drive | ceph | openstack



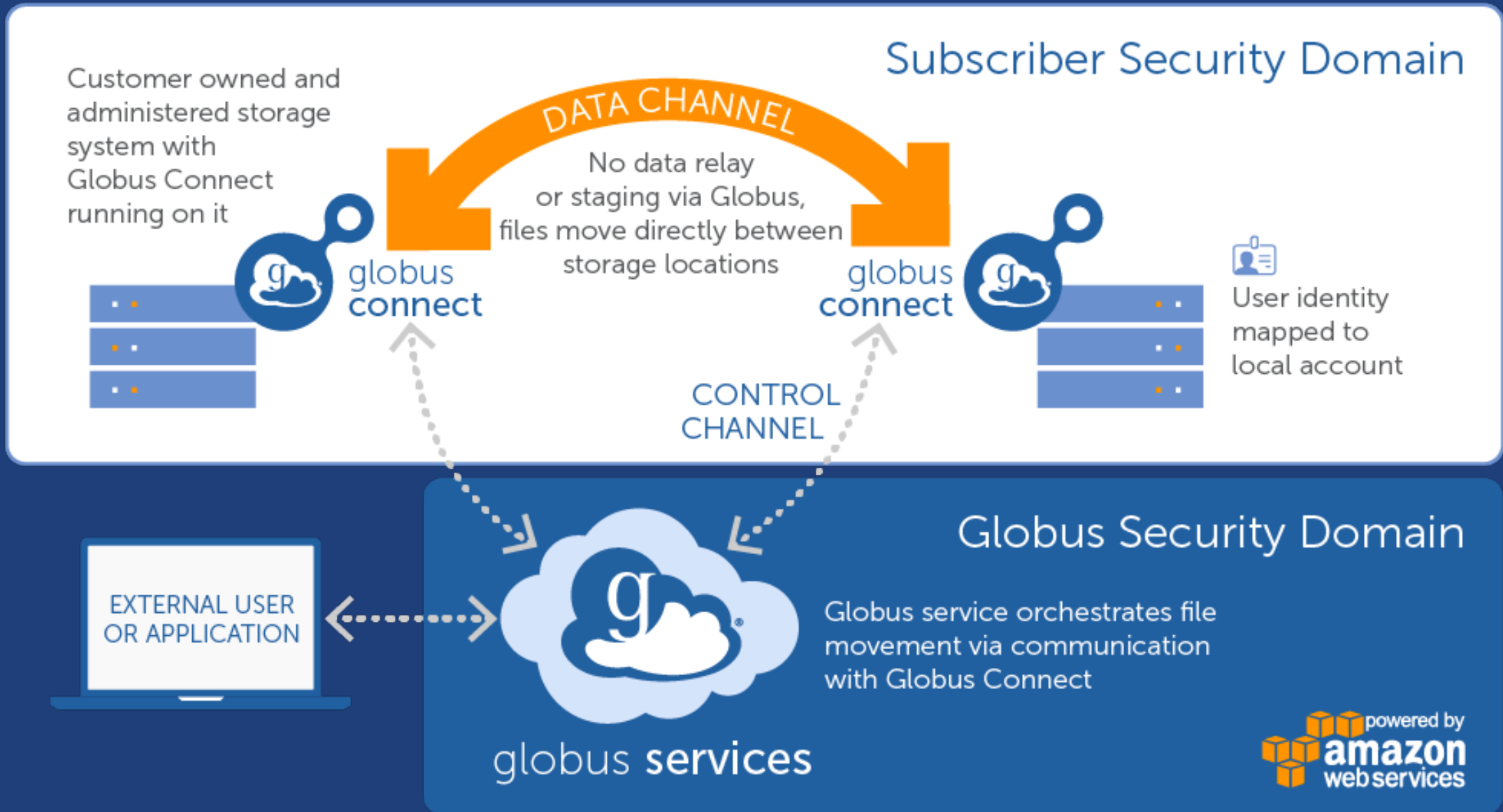
# Secure data sharing ...from any storage

- Fine-grained access control “overlay” on storage system
- Share with any identity, email, group
- No need to stage data just for sharing



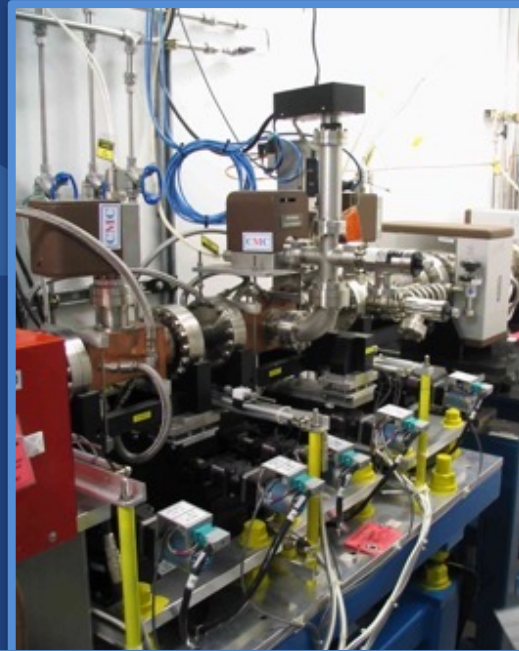


# Conceptual Architecture: Hybrid SaaS



 Seen any of these monsters on campus?

Next-Gen Sequencers



Advanced Light Source



fMRI



Light Sheet Microscope



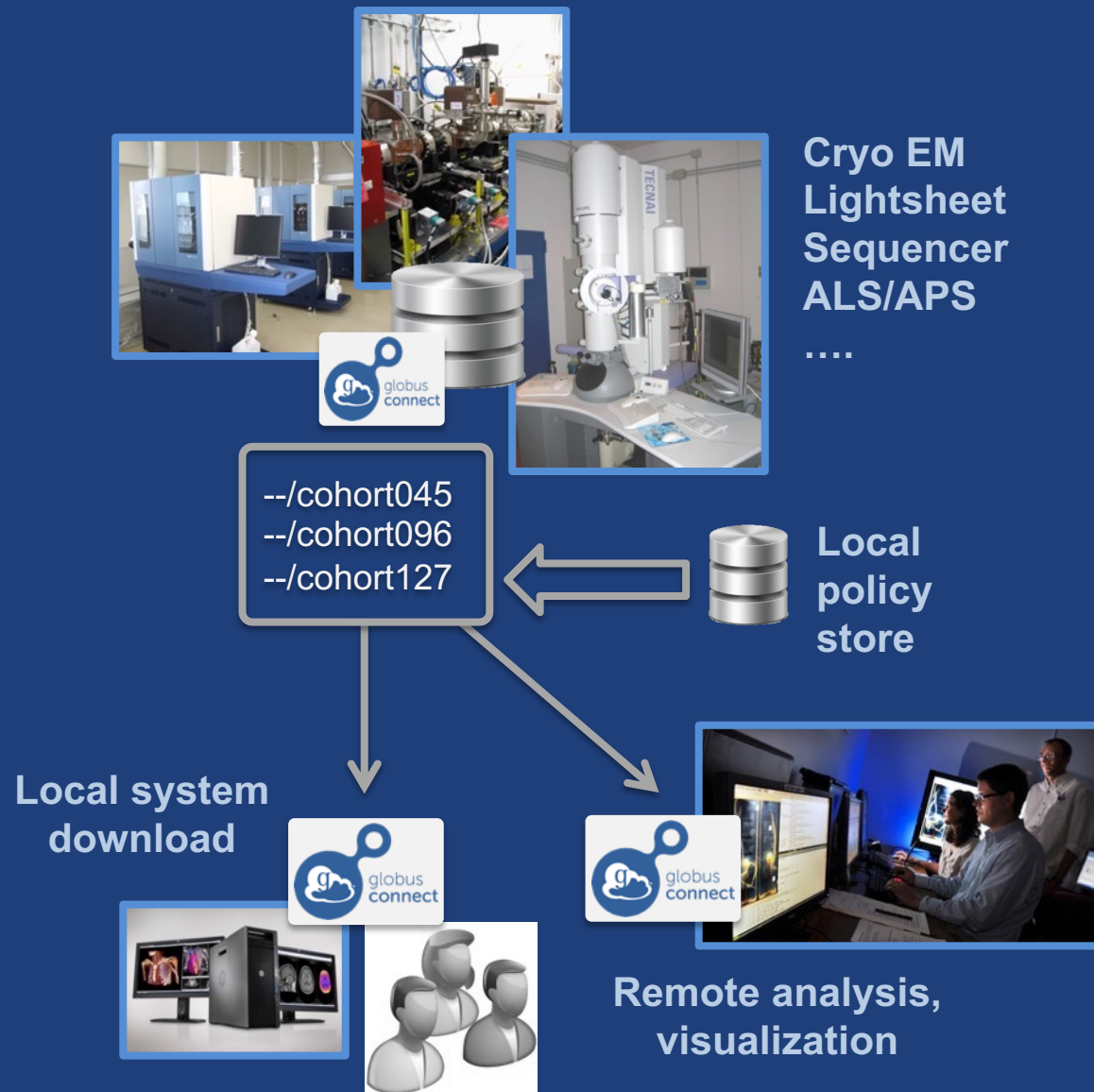
Cryo-EM





# Instrument data management needs

- **Reliable, near-real time data access**
- **Self-service access control, management**
- **Grant data access to collaborators**
- **Compute on data across storage classes**
- **Do it all at SCALE**





# What is “SCALE”?

- ~5TB/day
- >10TB/day at full cycle
- **Constraints**
  - Local storage
  - Time!



Krios CryoEM  
Courtesy of Thermo Fisher

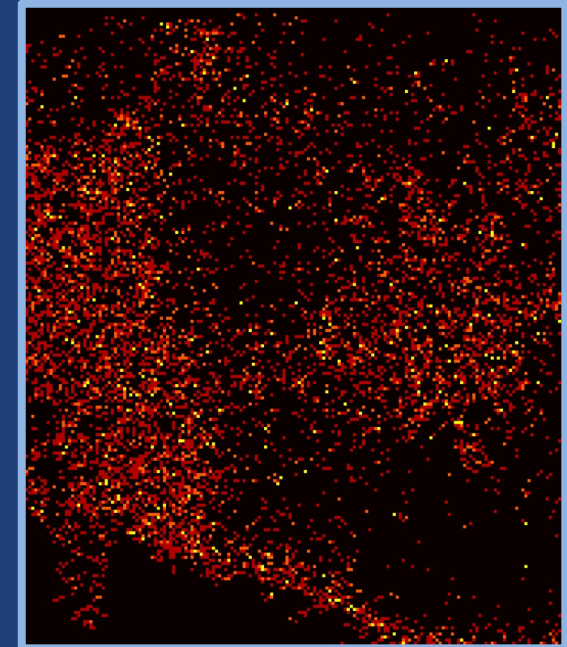
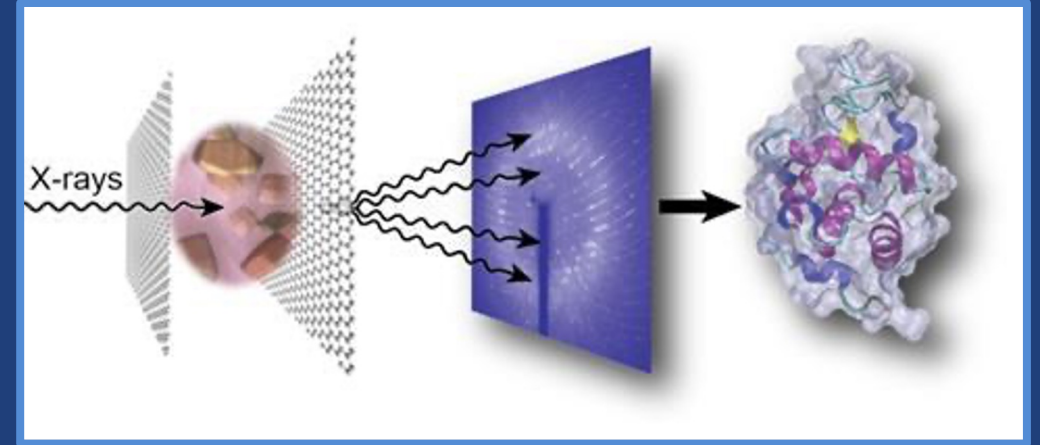
# Scale ...and then some

- 2-D, 3-D imaging
- 2016: ~112TB/month
- 100x – 1,000x growth



# Enabling serial crystallography at scale

- **Serially image chips with 000's of embedded crystals**
- **Quality control first 1,000 to report failures**
- **Analyze batches of images as they are collected**
- **Report statistics and images during experiment**
- **Return structure to scientist**





# Globus Automation Capabilities



## 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





# “Simple” Automation Use Cases

- **Data backup – as user, as system**
- **Stage data in or out as part of a compute job**
- **Data portal/gateway submits a transfer of compute results as the user**
- **Data portal/gateway monitors transfer, initiates processing or backup of data**



Recurring transfers with sync option



Copy /ingest  
Daily @ 3:30am





# Scheduled transfers using Globus timers (globus 'cron')





# Globus Command Line Interface (CLI)



# Globus Command Line Interface

```
(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.
  delete        Submit a Delete Task
  endpoint      Manage Globus Endpoint definitions
  get-identities Lookup Globus Auth Identities
  list-commands List all CLI Commands
  login         Login to 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
  task          Manage asynchronous Tasks
  transfer      Submit a Transfer Task
  version       Show the version and exit
  whoami        Show the currently logged-in identity.
```

**Automation of  
simple data  
management tasks**

**Integration with  
existing scripts  
(job submission ...)**

**Open source, uses  
the Python SDK**



# Automation using Globus Flows

- **Managed, secure, reliable task orchestration**
- **Support for heterogenous resources**
- **Event driven execution model**
- **Extensible via custom actions**

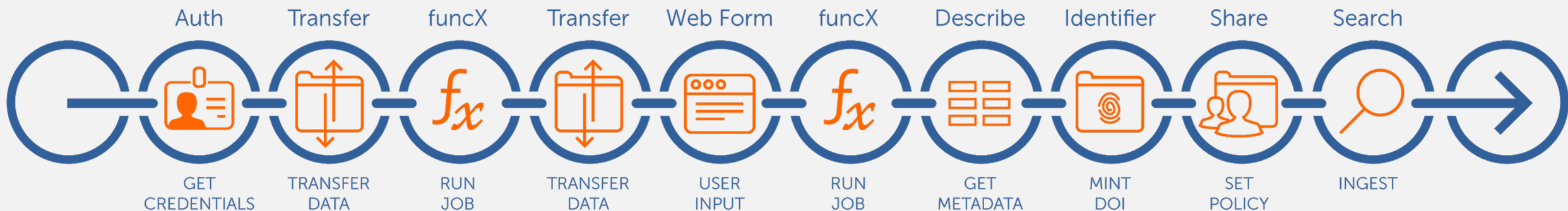




# The Globus Flows service

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

\* Coming soon





# SSX Automation

## Data capture

Globus  
Flows



funcX



Launch QA  
job



Carbon!



Check  
threshold

Transfer



Transfer  
raw files

funcX

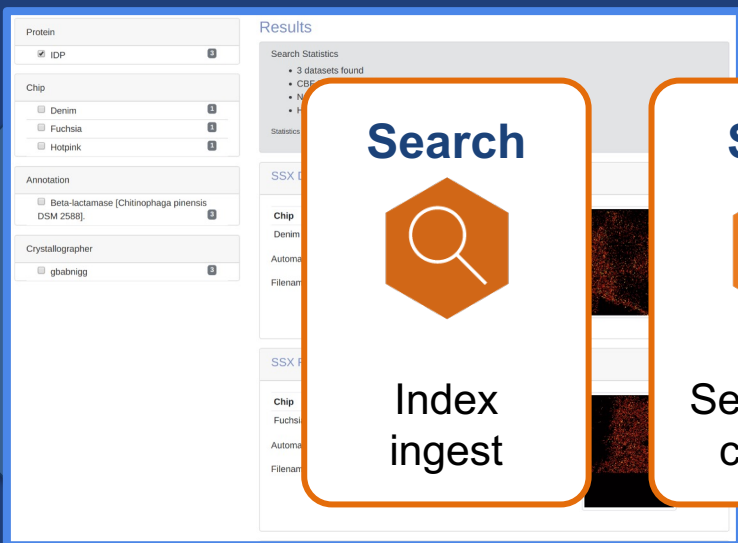


Analyze  
images

Image  
processing



## Data publication



Search



Index  
ingest

Share



Set access  
controls

Transfer



Move results  
to repo

funcX

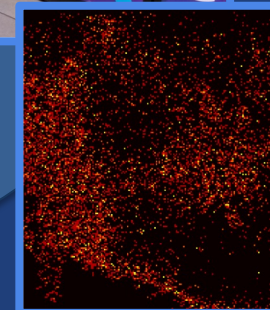


Gather  
metadata

funcX

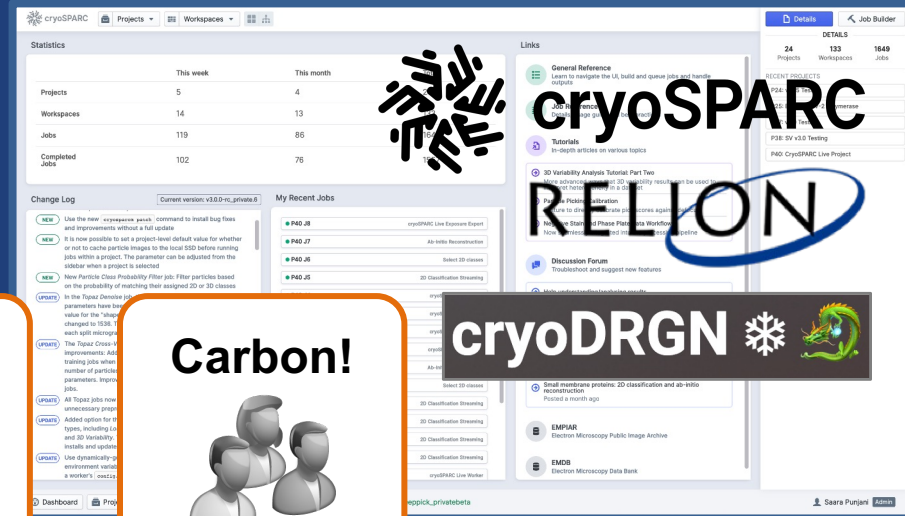


Visualize





# Automating cryoEM flows



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

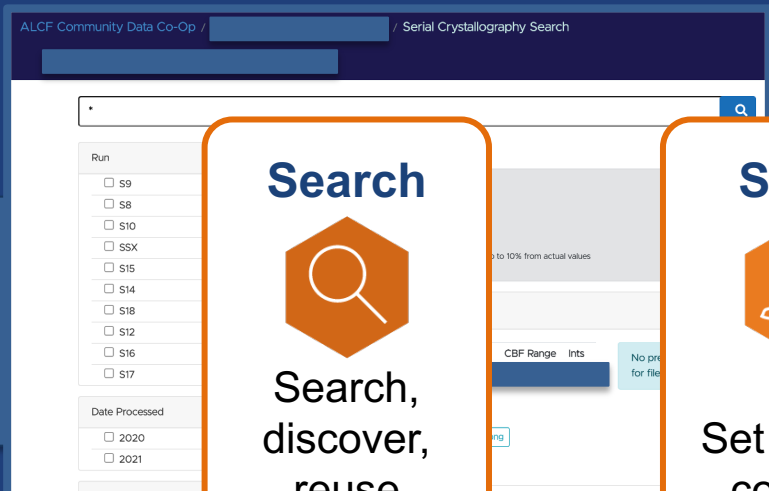


Move final  
files to repo

Search



Index  
ingest

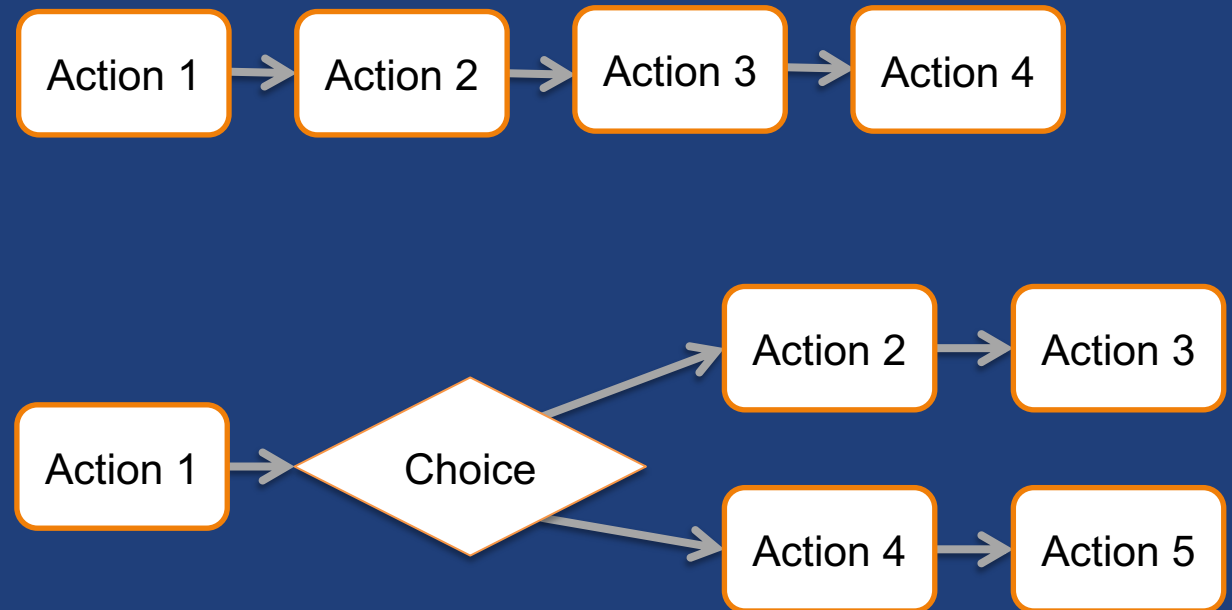




# Creating and deploying flows



- Define flows using a declarative language (JSON or YAML)
- Deploy flows to the Globus Flows service
- Set access policy for visibility and execution





# Globus-provided flows

## Two Stage Globus Transfer

kurt@globus.org

This flow requires at least one collection to be managed under a Globus subscription. The flow will perform a data transfer between source and destination collections in two stages. The first stage transfers from the source collection to an intermediate collection, and the second stage transfers from the intermediate collection to the destination collection. Data used in this flow are deleted from the intermediate collection after the final transfer is complete. Transferring data through an intermediate location can enable or improve performance in some firewalled or other network configurations.

 Start



STEPS  
25

CREATED  
2022-03-30 11:24

LAST MODIFIED  
2022-03-30 11:24

KEYWORDS  
Two Stage,Two Hop,Intermediate,Globus  
Transfer,Transfer,Globus  
Production,Production

## Move (copy and delete) files using Globus

This flow requires at least one collection to be managed under a Globus subscription. Following the transfer operation, data in the source collection will be deleted if the transfer to the destination collection is successful.

 Start



STEPS  
23


CREATED  
2021-10-21 13:53

LAST MODIFIED  
2022-03-30 11:20



KEYWORDS  
Move,Data Move,Globus  
Transfer,Transfer,Globus  
Production,Production




# Search and discover available flows















## Flows

 Runs  Library

List of flows you may view or use.



QUICK FILTERS  ADMINISTERED BY ME  RUNNABLE BY ME

<b>UMich cryoEM Flow</b> Vas Vasiliadis Moves files from instrument to a compute node for processing, runs image processing code using funcX, moves processed image files to S3 endpoint and shares images with the specified group of users.		 Start 
<b>Trigger Sample Flow</b> Vas Vasiliadis Moves files to the specified guest collection and grants read-only permissions to the specified user or group.		 Start 
<b>EGI Conference Flow</b> Vas Vasiliadis		 Start 
<b>XPCSBost Flow</b> Hannah Parraga		 Start 

STEPS	CREATED	LAST MODIFIED	KEYWORDS
4	9/22/2022, 12:44 PM	9/25/2022, 05:37 PM	transfer, compute, share, trigger,funcX, tutorial
2	9/14/2022, 02:14 PM	9/25/2022, 05:37 PM	transfer, share, tigger, tutorial
2	9/20/2022, 11:11 PM	9/20/2022, 11:11 PM	
9	9/13/2022, 03:04 PM	9/13/2022, 03:04 PM	



# Running Globus flows

- **Start and manage runs via web app, API**
- **Globus Automate CLI and SDK:**  
**[globus-automate-client.readthedocs.io](https://globus-automate-client.readthedocs.io)**
- **Event driven execution of flows: Triggers**
  - e.g., when a file of specific type is created
  - e.g., after a file has not been modified for a defined period





# Managing flow execution



- Provide inputs and run an instance of the flow
- Check run progress/status, cancel run
- Set access policy for monitoring, managing

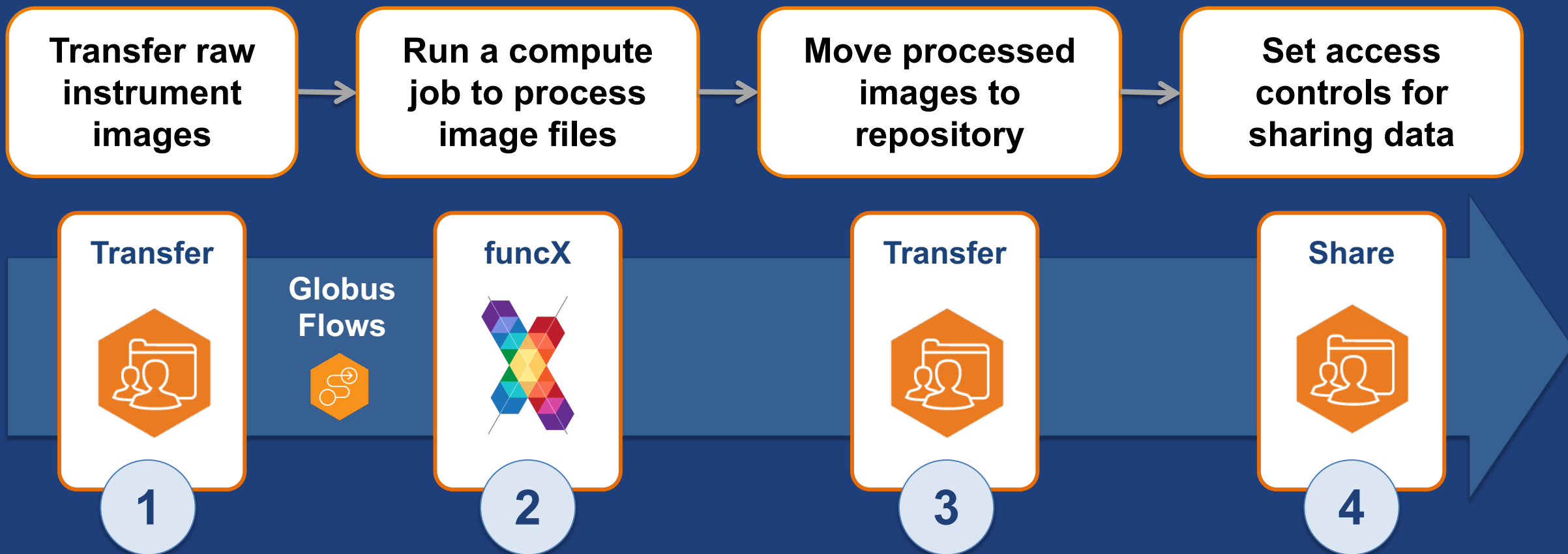
The screenshot displays the GigaFlow web interface. The main view is titled 'Flows' and shows a list of flow runs. The left sidebar contains navigation icons for File Manager, Bookmarks, Activity, Collections, Groups, Account, Logout, and Help. The main content area shows a list of flow runs with columns for status, name, and completion time.

Status	Flow Name	Run Status
✓	CryoGN 045 Abbvie Instrument Flow	completed at 8/23/2022, 10:21 AM
✓	I030_Gel5d_A03_5_att0_Rq0_ XPCSBoost Flow	completed at 8/22/2022, 04:47 AM
✓	I030_Gel5d_A03_5_att0_Rq0_ XPCSBoost Flow	completed at 8/22/2022, 04:47 AM

Overlaid on the right is a configuration window for 'Start - UMich cryoEM Flow'. It includes a 'Back to Flows Library' link and three sections for selecting source and destination collections and paths. Each section has a 'Collection' search field and a 'Path' field with a 'Browse' button. At the bottom are 'Start Flow' and 'Cancel' buttons.



# A simplified example



# Our environment

```
def process_images(input_path=None, result_path=None):
    import os
    import glob
    from PIL import Image

    files = (file for file in glob.glob(os.path.join(input_path, '*')) \
            if os.path.isfile(os.path.join(input_path, file)))

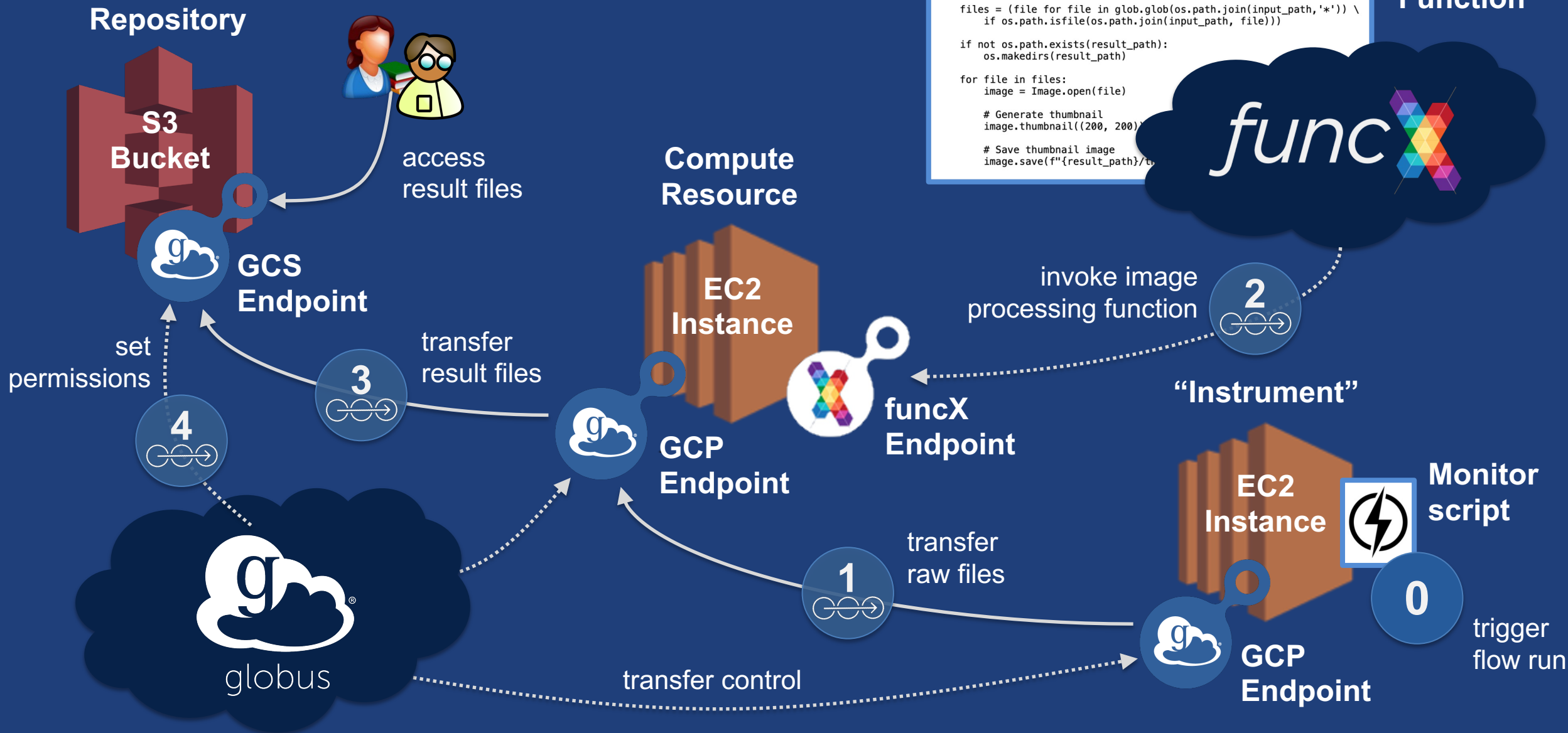
    if not os.path.exists(result_path):
        os.makedirs(result_path)

    for file in files:
        image = Image.open(file)

        # Generate thumbnail
        image.thumbnail((200, 200))

        # Save thumbnail image
        image.save(f'{result_path}/t')
```

Registered Function





# Taming Cloud Storage



# “Data appropriate” storage

- **Often driven by institution...**
- **...and sometimes by vendor policy!**
- **Decision based on**
  - Access frequency (data temperature)
  - Access modality(ies)
    - Ad hoc, via web browser
    - Scripted, via CLI tools
    - Programmatic, via APIs
  - Storage system limits
  - ...yeah, and cost



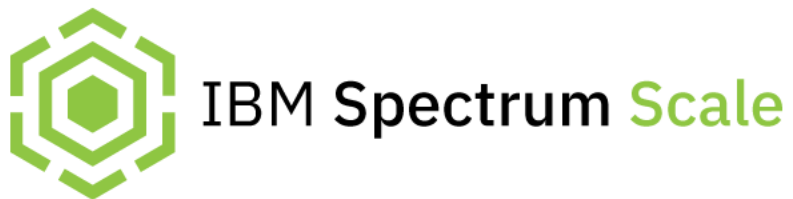
# Move without (worrying about) limits

- API request rates
- File size
- Data volume
- Third-party tools cannot circumvent...
- ...but Globus lets you “fire-and-forget”
- → it will (eventually) be done

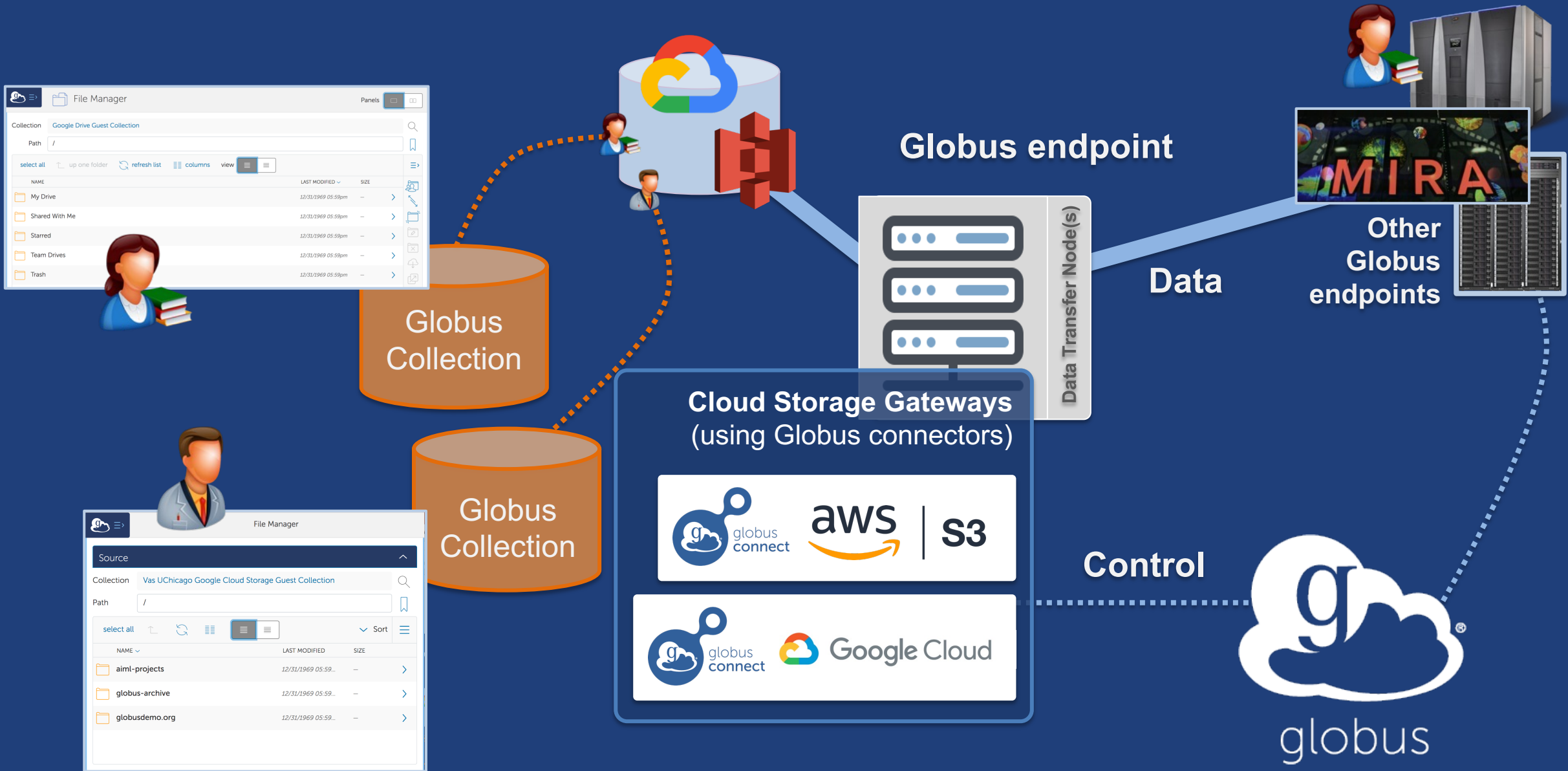
```
3/9/2022, 08:33 PM endpoint too busy View details ^  
  
Error (transfer)  
Endpoint: Vas Google Drive Collection (d6d62391-fdda-4ba5-ac78-6523f806ea79)  
Server: m-422a8b.d8b83.36fe.data.globus.org:443  
File: /My%20Drive/migration/uchicago-perftest/cc32-16p32-16/test1185  
Command: STOR /My Drive/migration/uchicago-perftest/cc32-16p32-16/test1185  
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: "PA  
"https://www.googleapis.com/upload/drive/v3/files/11h9NBppR7qG7YaZDqyWi-8w9rzW9gGj"\r\n451-GD-Response-Code: "  
exceeded."r\n451 End.r\n
```



# Globus connectors



# Connectors provide a uniform user experience



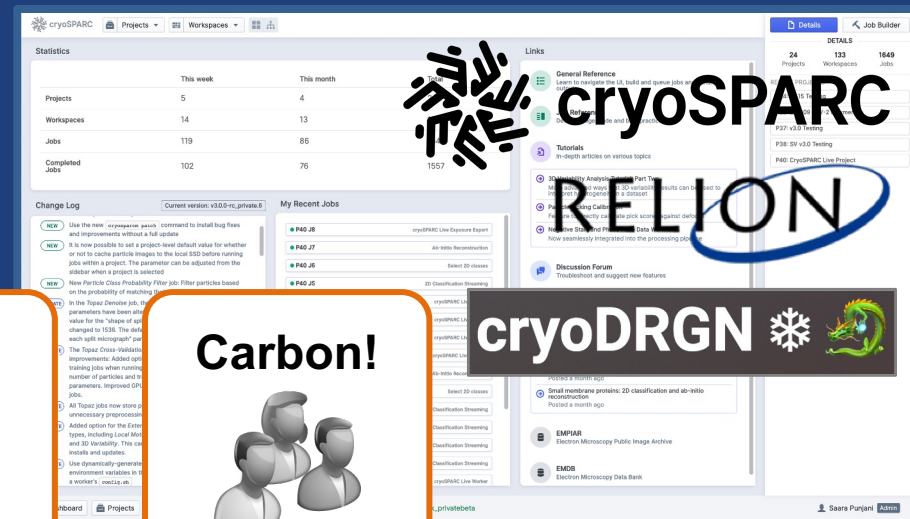




**Current areas of focus...**



# Automating Cryo-EM Flows



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

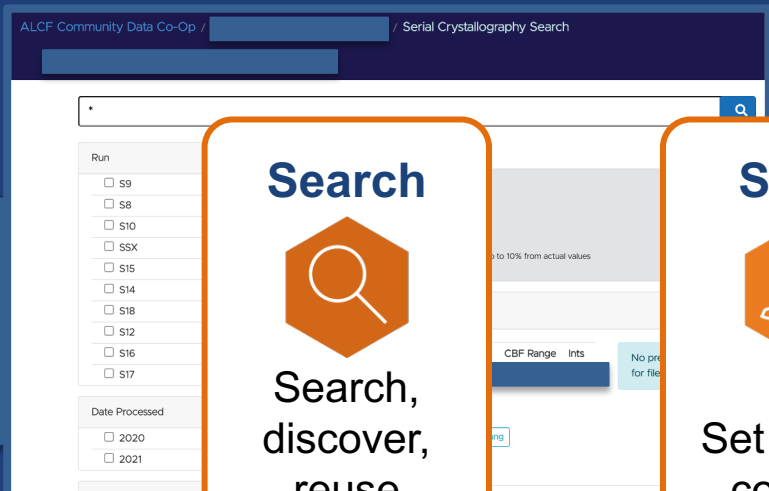


Move final  
files to repo

Search



Index  
ingest





# Incorporating compute into Flows





Automating computation with *funcX*\*

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



 THE UNIVERSITY OF  
CHICAGO

**I** ILLINOIS

Argonne   
NATIONAL LABORATORY

*\* funcX is in currently under development and in limited production use*



# Globus Search and Portal Framework

## Example: Advanced Photon Source

[acdc.alcf.anl.gov](https://acdc.alcf.anl.gov)

# Support resources

- **Globus documentation: [docs.globus.org](https://docs.globus.org)**
- **YouTube channel: [youtube.com/user/GlobusOnline](https://youtube.com/user/GlobusOnline)**
- **Helpdesk and issue escalation: [support@globus.org](mailto:support@globus.org)**
- **Mailing Lists**
  - [globus.org/mailing-lists](https://globus.org/mailing-lists)
- **Customer engagement team**
  - Office Hours