



Empowering Secure Data-Driven
Research: A Workshop on Science DMZ,
Globus, and InCommon Federation

Globus: platform for data driven research

Rachana Ananthakrishnan
Executive Director, Globus
University of Chicago
ranantha@uchicago.edu



Globus is ...

a non-profit service
developed and operated

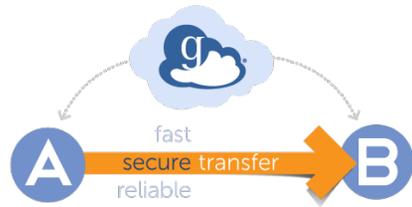


THE UNIVERSITY OF
CHICAGO

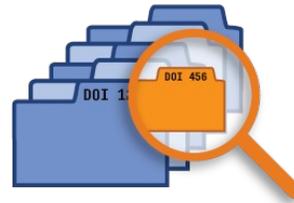
RESEARCH



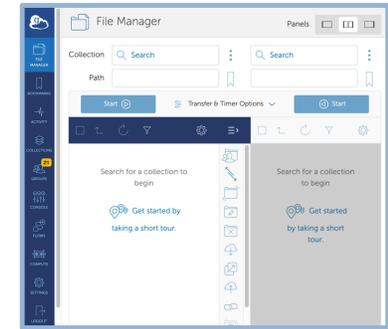
Globus Platform for Research IT



Managed transfer & sync



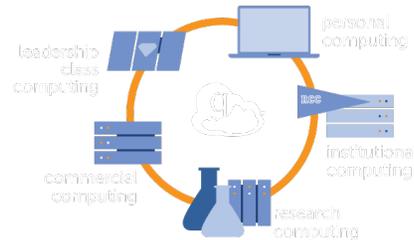
Publication & discovery



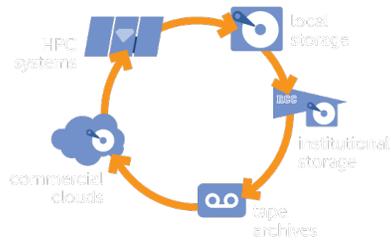
Software-as-a-Service



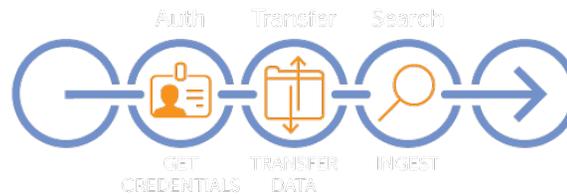
Collaborative data sharing



Managed remote execution



Unified data access



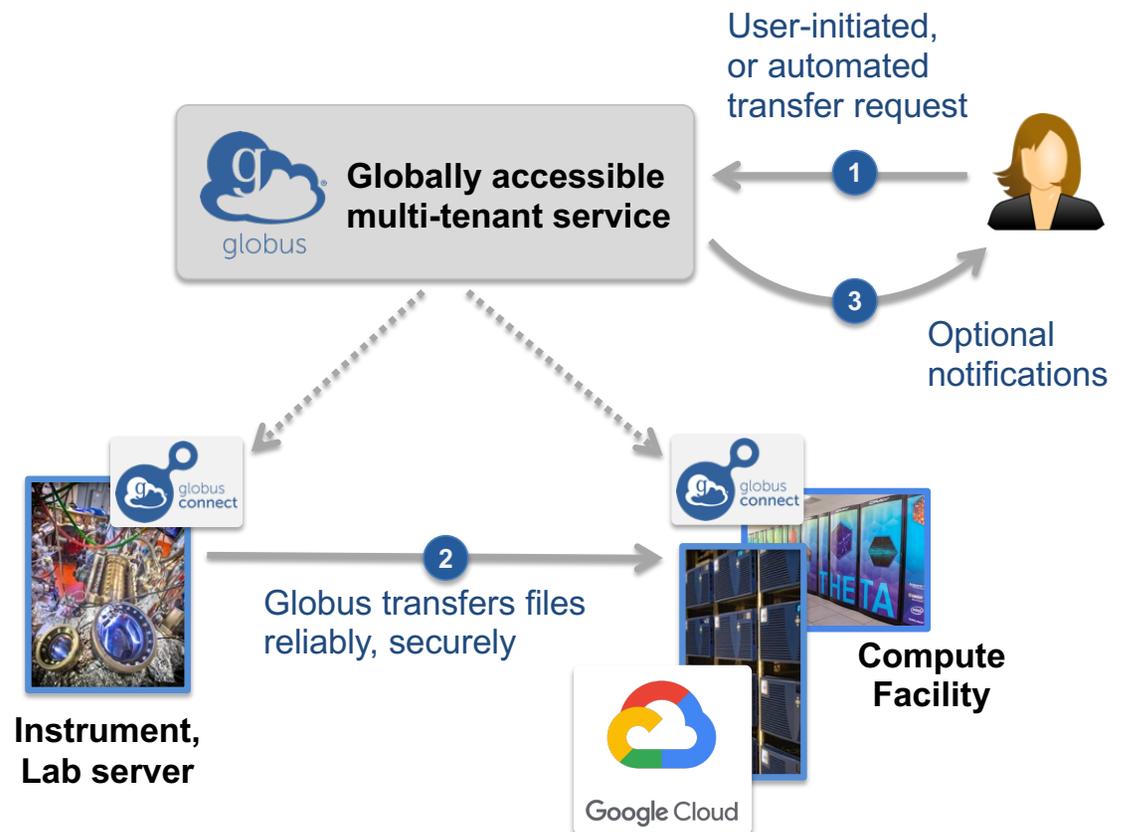
Reliable automation



Platform-as-a-Service

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

- Fire-and-forget transfers/sync
- Optimized speed
- Assured reliability
- Unified view of storage
- HTTP/S access to data





Intuitive web application interface

File Manager

Collection: UChicago RCC Midway3

Path: /~/

Transfer & Timer Options

NAME	LAST MODIFIED	SIZE
E099_HPPG_100_55_025C_att06_...	3/17/2023, 11:2...	110.45 KB
E099_HPPG_100_55_025C_att06_...	3/17/2023, 11:2...	113.91 KB
esgf_demo	3/11/2023, 12:1...	-
GW_Demo	4/18/2023, 02:...	-
TestFolder	9/30/2022, 12:...	-
TestUser1	3/20/2023, 05:...	-

ALCF Username

Password

Cryptocard or Mobile token password

SIGN IN

This is a Federal computer system and is the property of the United States Government. It is for authorized use only. Users (authorized or unauthorized) have no explicit or implicit expectation of privacy.

Any or all uses of this system and all files on this system may be intercepted, monitored, recorded, copied, audited, inspected, and disclosed to authorized site, Department of Energy, and law enforcement personnel, as well as authorized officials of other agencies, both domestic and foreign. By using this system, the user consents to such interception, monitoring, recording, copying, auditing, inspection, and disclosure at the discretion of authorized site or Department of Energy personnel.



Transfer/sync options

Start ▶ 1 Transfer & Timer Options ^ Start ◀

Label This Transfer

Transfer Settings

NOTE: These settings will persist during this session unless changed.

sync - only transfer new or changed files ⓘ

where the modification time is newer
 file size is different
 file does not exist on destination
 checksum is different

Files which are overwritten by this option.

delete files on destination that do not exist on source ⓘ

preserve source file modification times ⓘ

do NOT verify file integrity after transfer ⓘ

encrypt transfer ⓘ

Skip files on source with errors ⓘ

Fail on quota errors ⓘ

Notification Settings

Disable success notification ⓘ

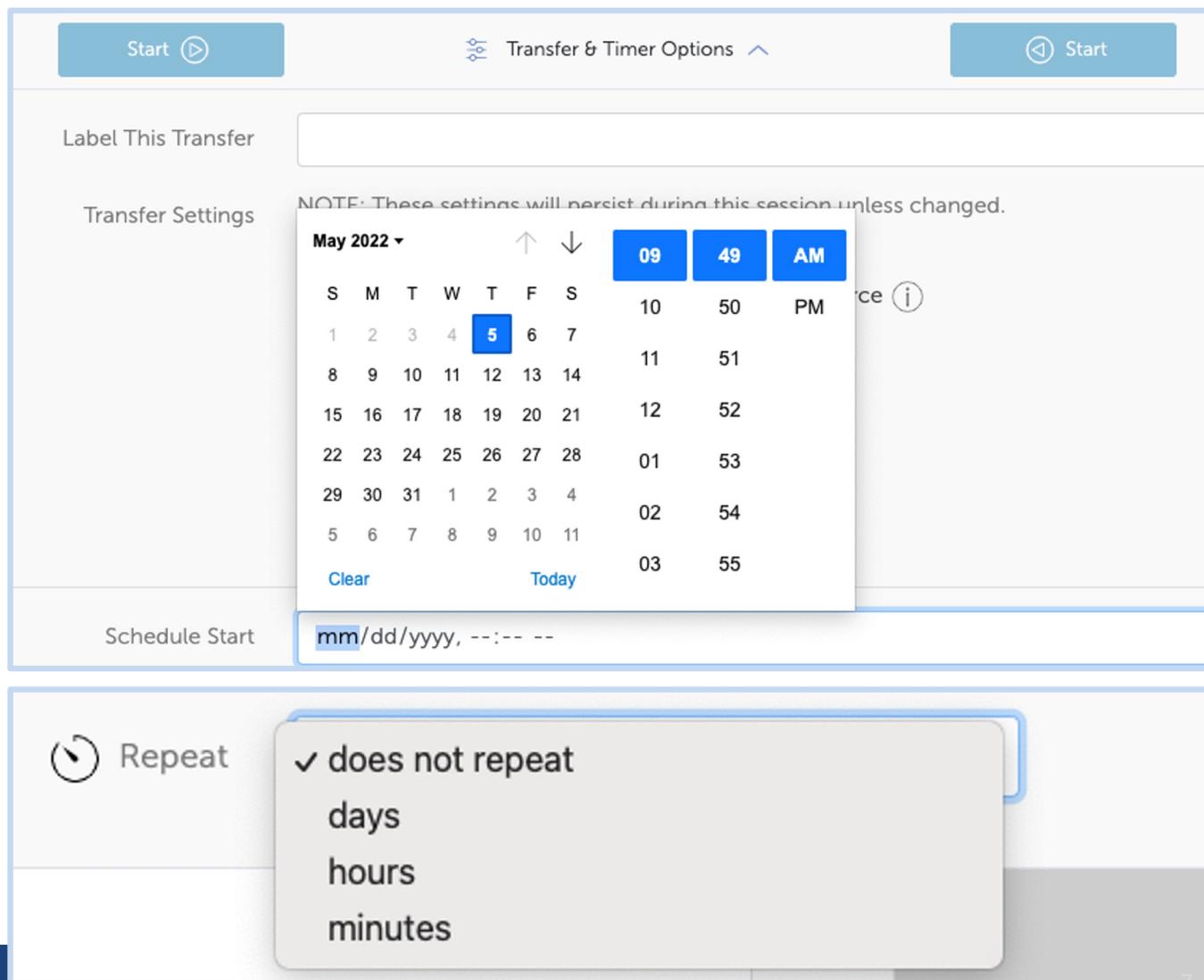
Disable failure notification ⓘ

Disable inactive notification ⓘ

Timers

**Scheduled
and/or recurring
file transfers**

**All Globus
transfer and
sync options**

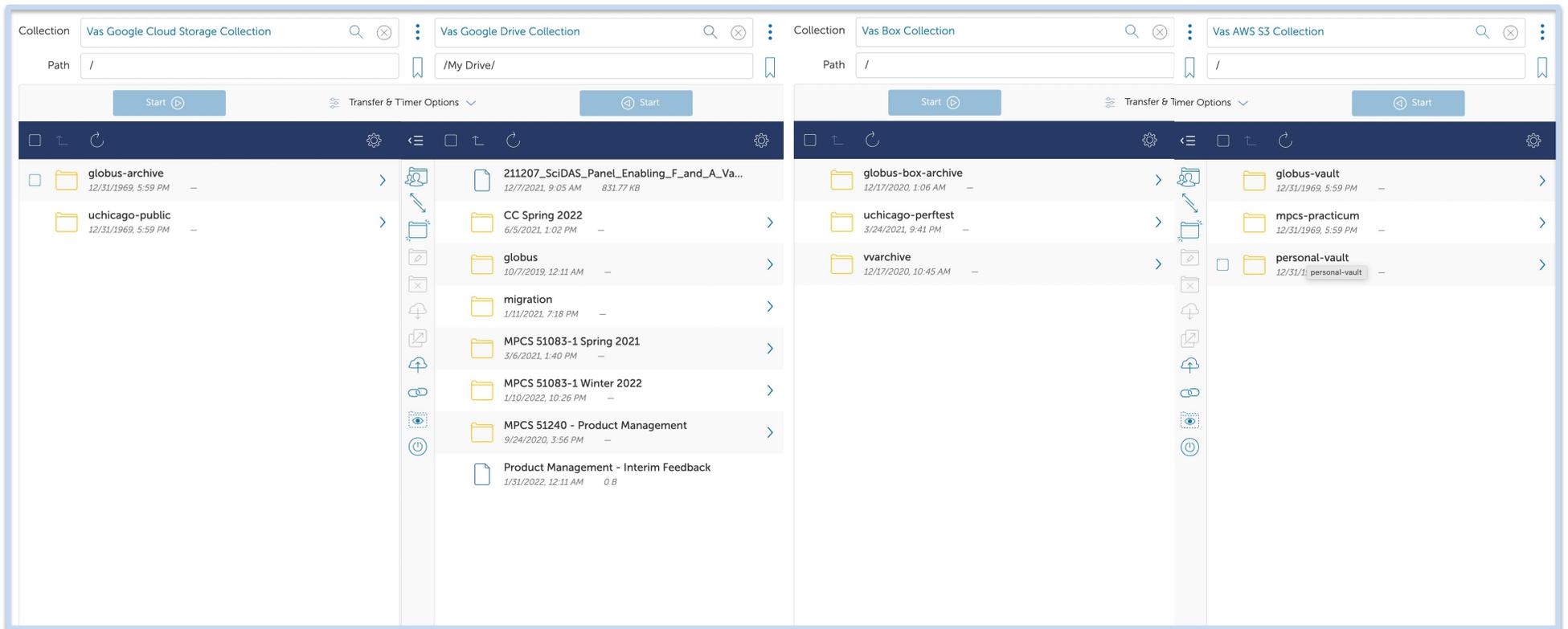


The screenshot displays the Globus Timers configuration interface. At the top, there are 'Start' buttons with play and back icons, and a 'Transfer & Timer Options' menu. Below this is a 'Label This Transfer' text input field. The 'Transfer Settings' section includes a note: 'NOTE: These settings will persist during this session unless changed.' A calendar pop-up is open, showing 'May 2022' with the 5th of the month selected. To the right of the calendar are time selection buttons for '09', '49', and 'AM', with 'PM' also visible. Below the calendar are 'Clear' and 'Today' links. The 'Schedule Start' field contains the placeholder text 'mm/dd/yyyy, --:-- --'. At the bottom, the 'Repeat' section is visible, with a dropdown menu open showing options: '✓ does not repeat', 'days', 'hours', and 'minutes'.

Supported storage systems



Uniform interface, consistent user experience



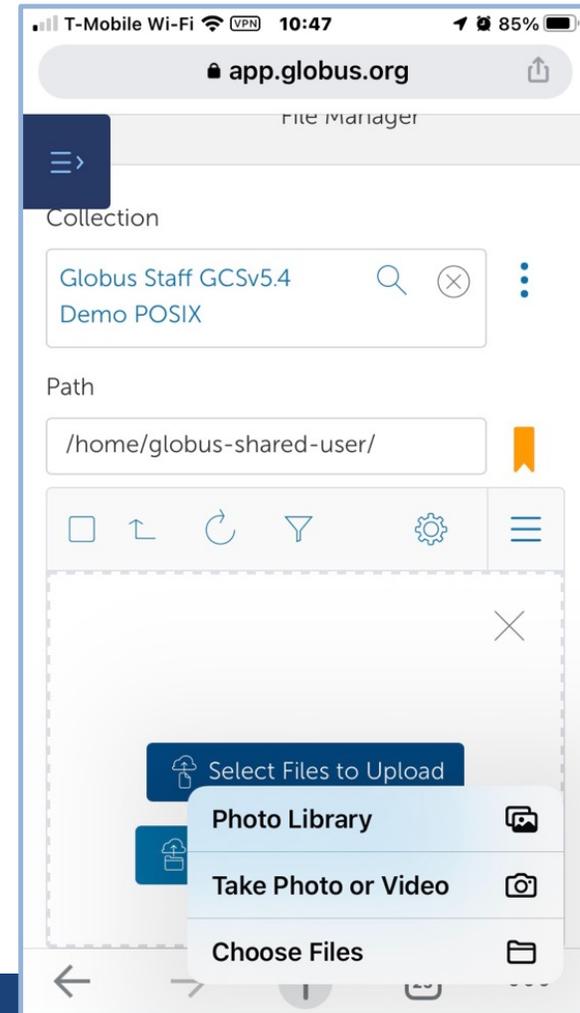
The screenshot displays a uniform file management interface across four different storage collections. Each pane includes a search bar, a path field, and a 'Start' button. The interface is consistent in layout and iconography, providing a consistent user experience across different data sources.

Collection	Path	Item Name	Item Type	Metadata
Vas Google Cloud Storage Collection	/	globus-archive	Folder	12/31/1969, 5:59 PM
		uchicago-public	Folder	12/31/1969, 5:59 PM
Vas Google Drive Collection	/My Drive/	211207_SciDAS_Panel_Enabling_F_and_A_Va...	File	12/7/2021, 9:05 AM 831.77 KB
		CC Spring 2022	Folder	6/5/2021, 1:02 PM
		globus	Folder	10/7/2019, 12:11 AM
		migration	Folder	1/11/2021, 7:18 PM
		MPCS 51083-1 Spring 2021	Folder	3/6/2021, 1:40 PM
		MPCS 51083-1 Winter 2022	Folder	1/10/2022, 10:26 PM
		MPCS 51240 - Product Management	Folder	9/24/2020, 3:56 PM
Product Management - Interim Feedback	File	1/31/2022, 12:11 AM 0 B		
Vas Box Collection	/	globus-box-archive	Folder	12/17/2020, 1:06 AM
		uchicago-perftest	Folder	3/24/2021, 9:41 PM
		wvarchive	Folder	12/17/2020, 10:45 AM
Vas AWS S3 Collection	/	globus-vault	Folder	12/31/1969, 5:59 PM
		mpcs-practicum	Folder	12/31/1969, 5:59 PM
		personal-vault	Folder	12/31/1/ personal-vault



Globus goes on the road

- **Upload photos from mobile device**
- **Leverages HTTP/S upload and responsive web application**

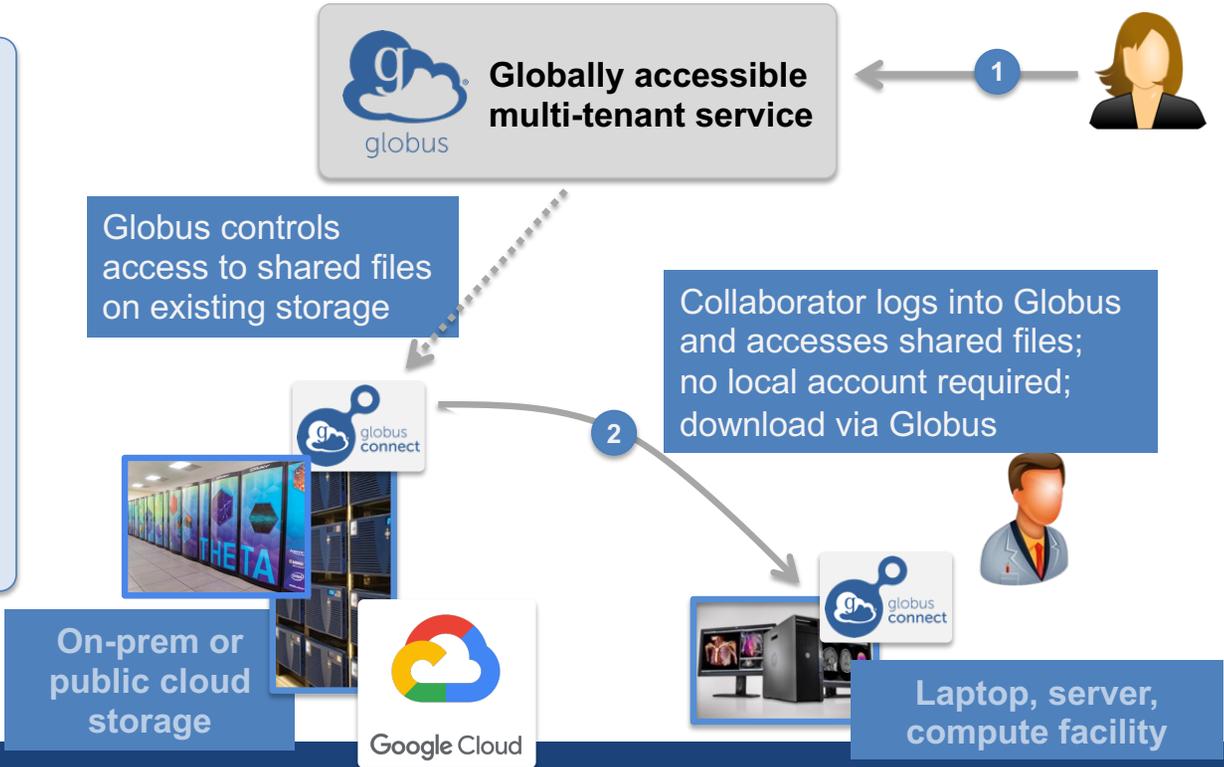




Secure data sharing ...from any storage

Select files to share, select user or group, and set access permissions

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





Data sharing – permissions & roles

Presentation Materials - RA

Overview Permissions Roles

USER OR GROUP	CREATED	EXPIRATION	READ	WRITE
Permissions granted by role				
Path: / Show link for sharing				
Brigitte Raumann (braumann@uchicago.edu)	3/8/2024, 04:33 PM	never expires	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Identity 6afa2dc5-d219-4078-9a06-ba37aa32c739 (6afa2dc5-d219-4078-9a06-ba37aa32c739@clients.auth.globus.org)	12/11/2023, 06:29 PM	never expires	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Path: /RPI/ Show link for sharing				
Public	2/27/2024, 03:00 PM			
Identity a7a2bd06-919b-478e-9477-447f14198a63 (foster@anl.gov)	5/3/2024, 04:20 PM			
Path: /Stanford/				
All Users	12/11/2023, 06:29 PM			

Overview Permissions Roles

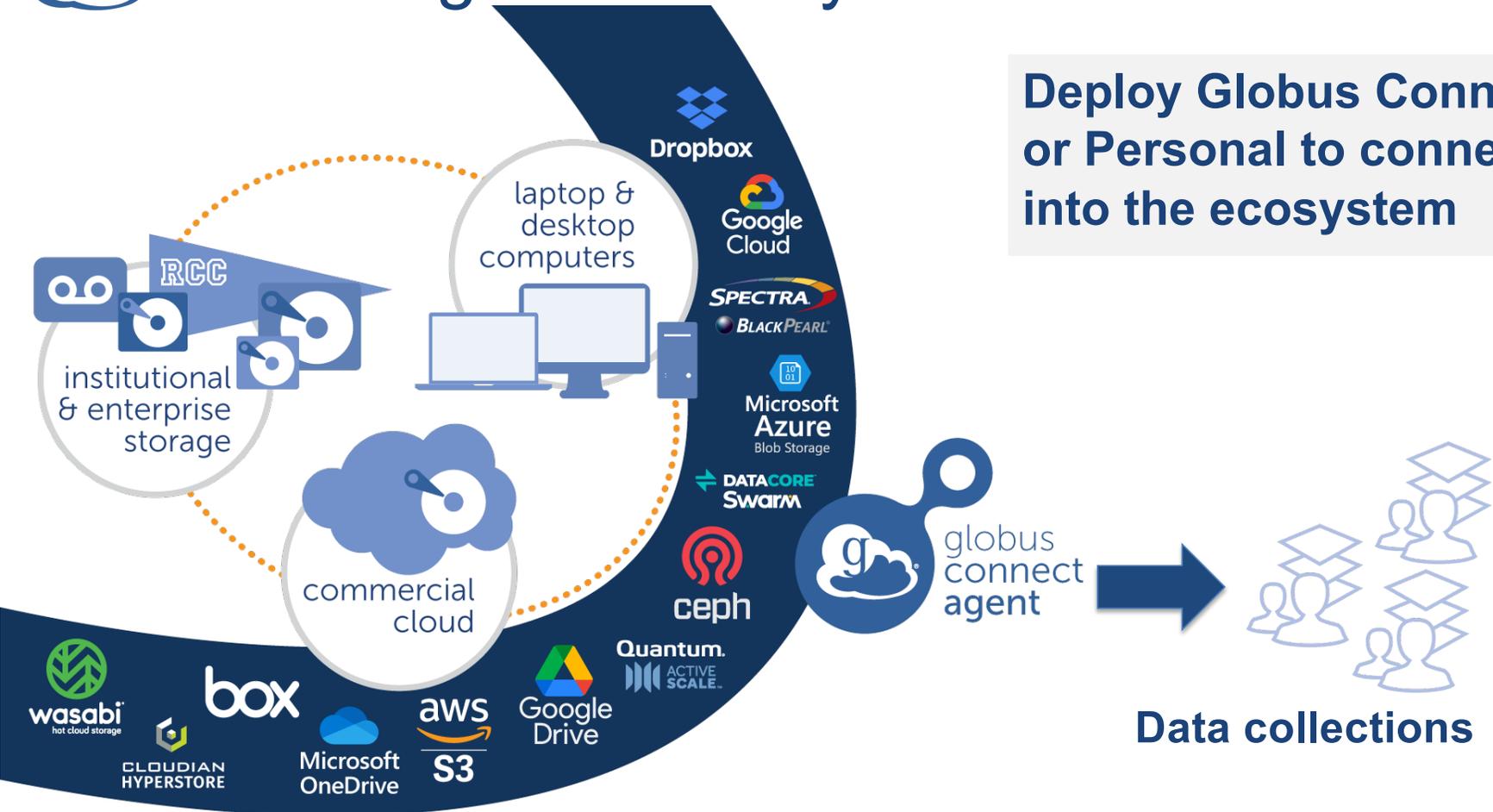
Assigned Roles

[Assign New Role](#)

	Rachana Ananthkrishnan	ranantha@uchicago.edu	Owner
	Rachana Ananthkrishnan	ranantha@uchicago.edu	Administrator
	7d69a95e-48a8-491e-8e72-ed1631e81954	7d69a95e-48a8-491e-8e72-ed1631e81954@clients.auth.globus.org	Access Manager

Enabling data ecosystem

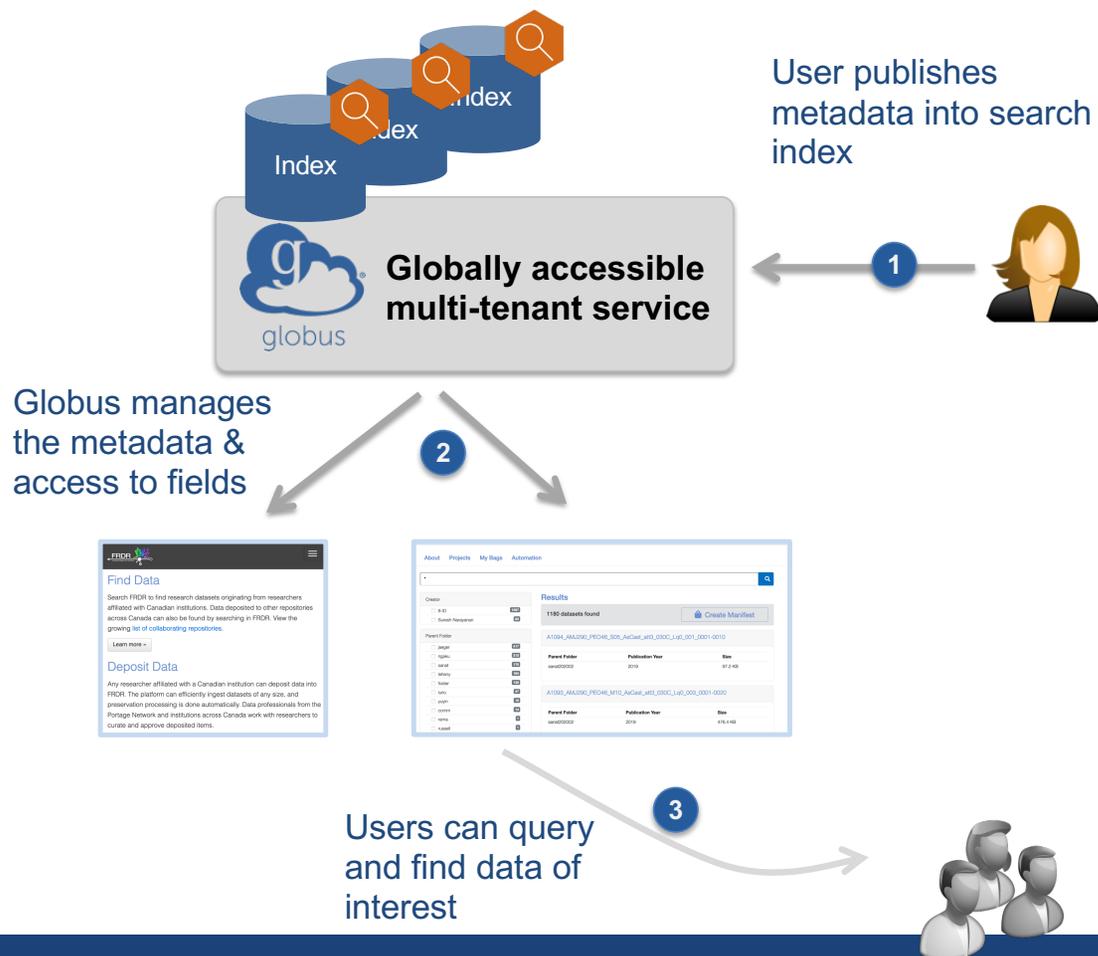
Deploy Globus Connect Server or Personal to connect storage into the ecosystem





Scalable data discovery ...for any domain

- Metadata store with fine grained visibility controls
- Schema agnostic, with dynamic schema
- Federated authentication integration
- Query and discovery API with facets





Data ingest with Globus Search

POST /index/{index_id}/ingest'

```
{
  "ingest_type": "GMetaList",
  "ingest_data": {
    "gmeta": [
      {
        "id": "filetype",
        "subject": "https://search.api.globus.org/abc.txt",
        "visible_to": ["public"],
        "content": {
          "metadata-schema/file#type": "file"
        }
      },
      ...
    ]
  }
}
```



- Bulk create and update
- Task model for ingest at scale



Data ingest with Globus Search

POST /index/{index_id}/ingest'

```
{
  "ingest_type": "GMetaList",
  "ingest_data": {
    "gmeta": [
      {
        "id": "weight",
        "subject": "https://search.api.globus.org/abc.txt",
        "visible_to": ["urn:globus:auth:identity:46bd0f56-
          e24f-11e5-a510-131bef46955c"],
        "content": {
          "metadata-schema/file#size": "37.6",
          "metadata-schema/file#size_human": "<501b"
        }
      },
      ...
    ]
  }
}
```



Search
Index



Visibility limited to Globus Auth identity

- Single user
- Globus Group
- Registered client application



Data discovery with Globus Search

GET /index/{index_id}/search?q=type%3Ahdf5

Simple query

```
{
  "@datatype": "GSearchResult",
  "@version": "2017-09-01",
  "count": 1,
  "gmeta": [
    {
      "@datatype": "GMetaResult",
      "@version": "2019-08-27",
      "entries": [
        { ... }
      ],
      "subject": "https://..."
    }
  ],
  "offset": 0,
  "total": 1
}
```





Data discovery with Globus Search

/index/{index_id}/search

```
{
  "filters": [
    {
      "type": "range",
      "field_name": "pubdate",
      "values": [
        {
          "from": "*",
          "to": "2020-12-31"
        }
      ]
    }
  ],
  "facets": [
    {
      "name": "Publication Date",
      "field_name": "pubdate",
      ...
    }
  ]
}
```

Complex query

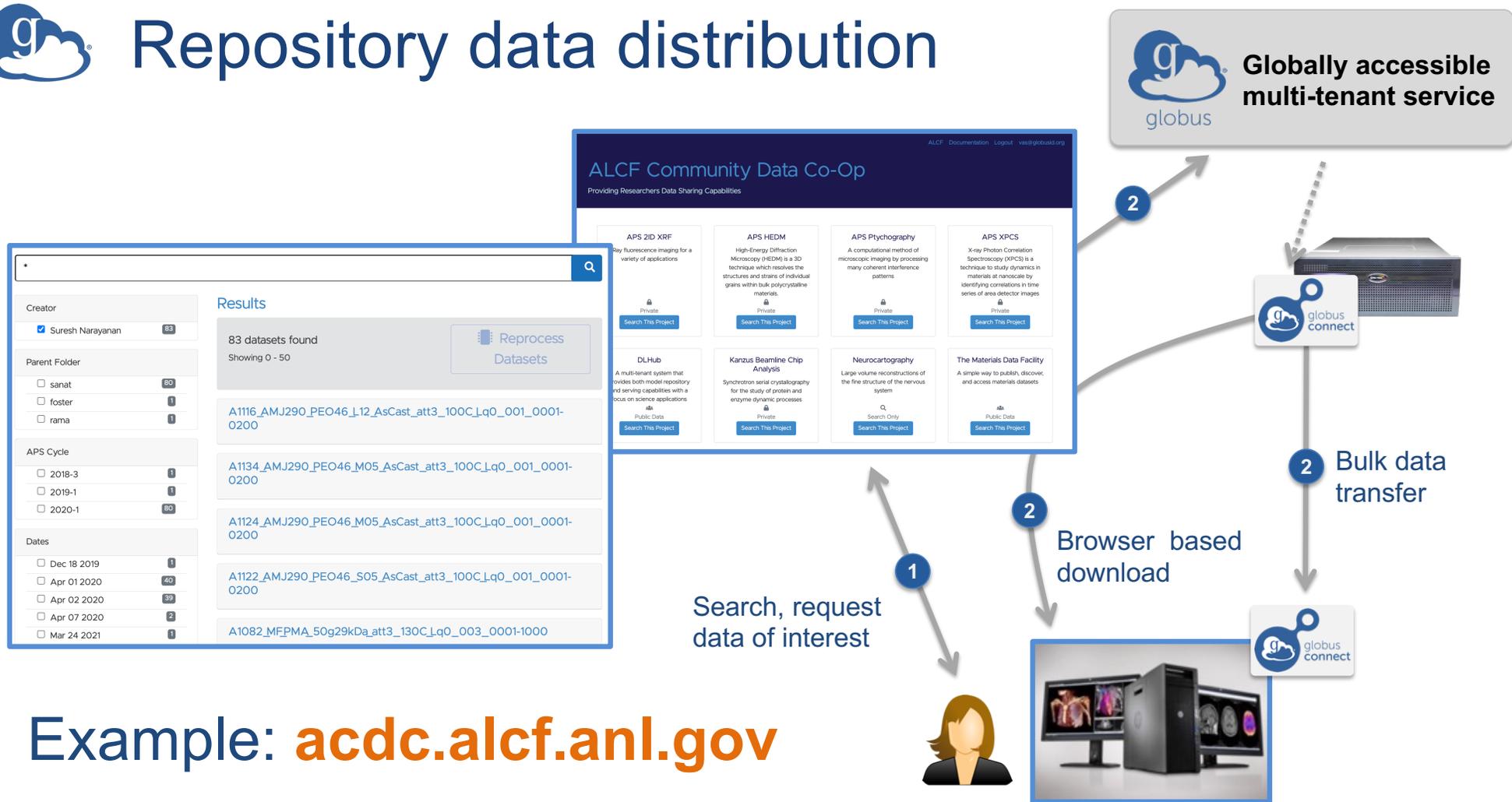
Filter
Facets
Boosts
Sort
Limit



Search
Index

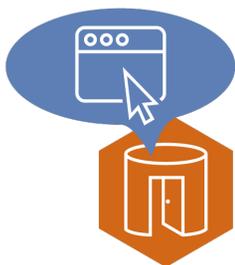


Repository data distribution



Example: acdc.alcf.anl.gov

Portals/Science Gateways/Data Commons



Serverless Portals

A JavaScript-based client-side data portal with no managed infrastructure requirements. Use our template repository to deploy a portal in just a few clicks.



Django-based Portals

A collection of tools that enable you to rapidly create an easily accessible portal for your data using Python and Django.



Custom Integrations

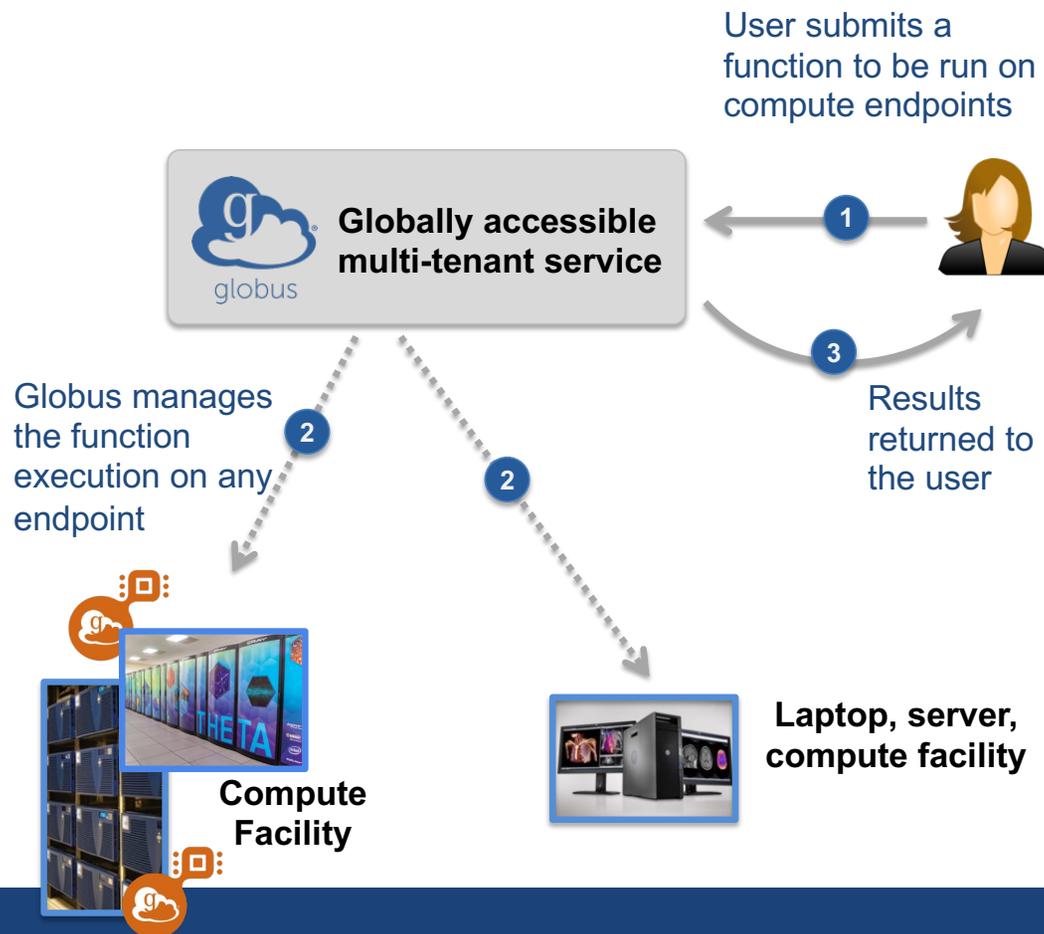
Use our first-class SDKs and service APIs to integrate Globus capabilities into your own application.

globus.org/portals



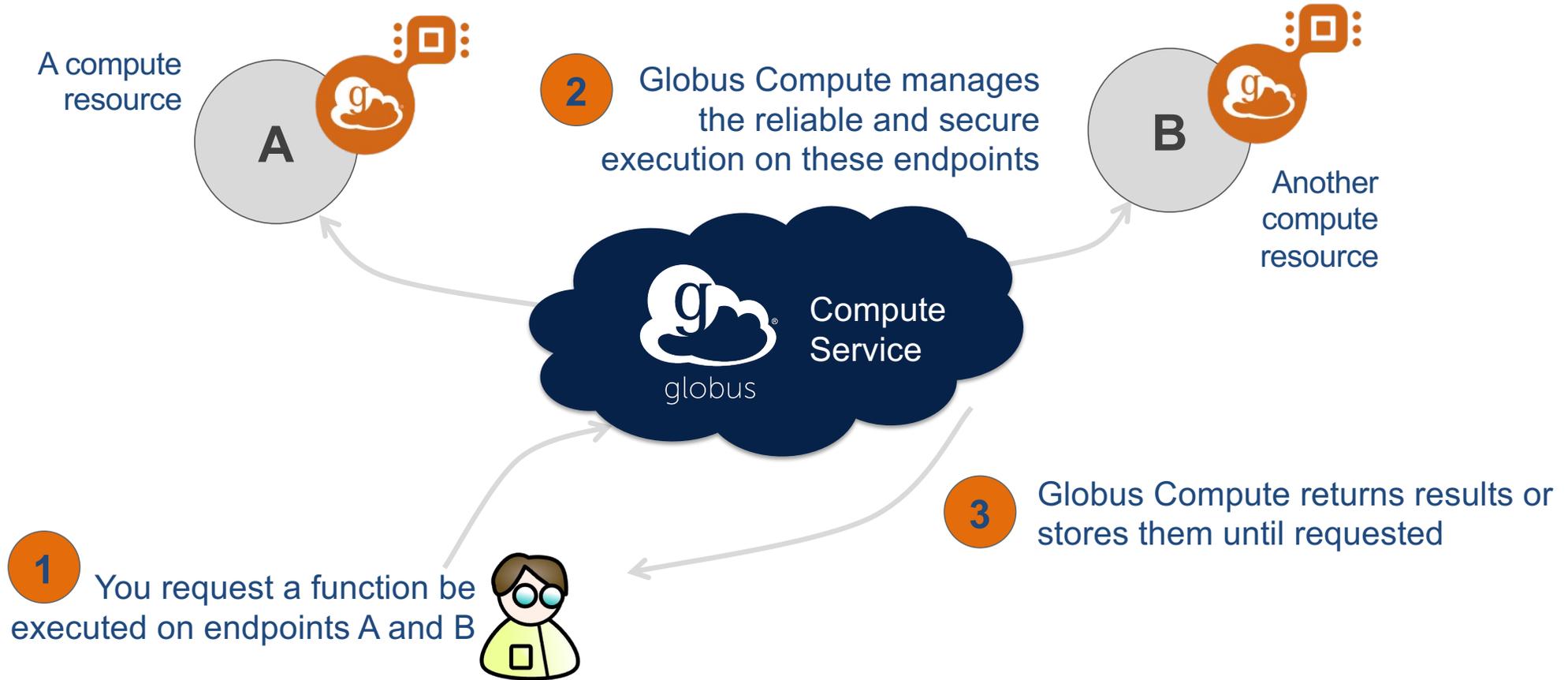
Managed compute ...on any system

- Fire and forget function execution
- Federated authentication, and local access control
- Uniform interface to various compute resources
- Support use of Python for functions

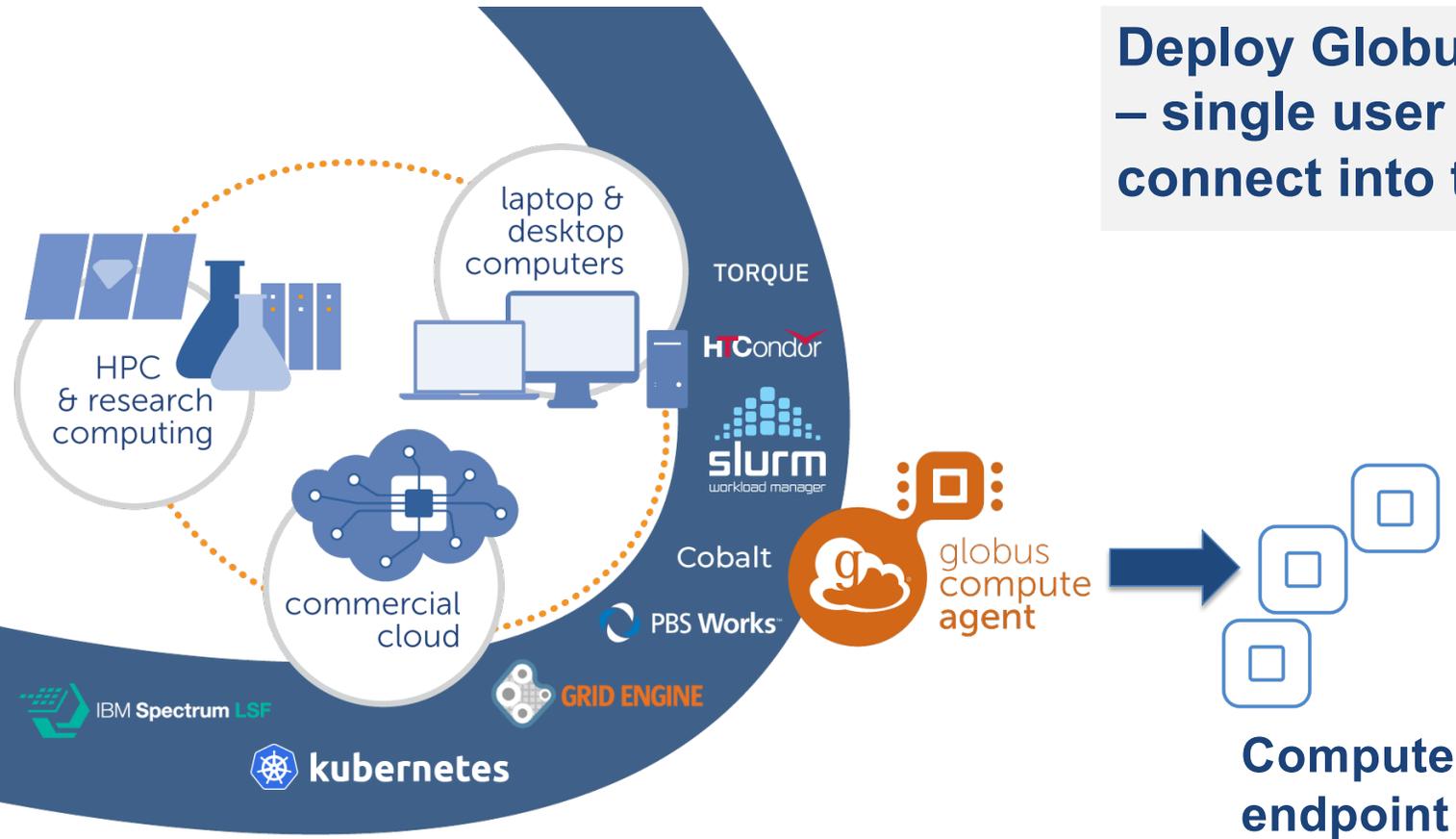




User interaction with Globus Compute



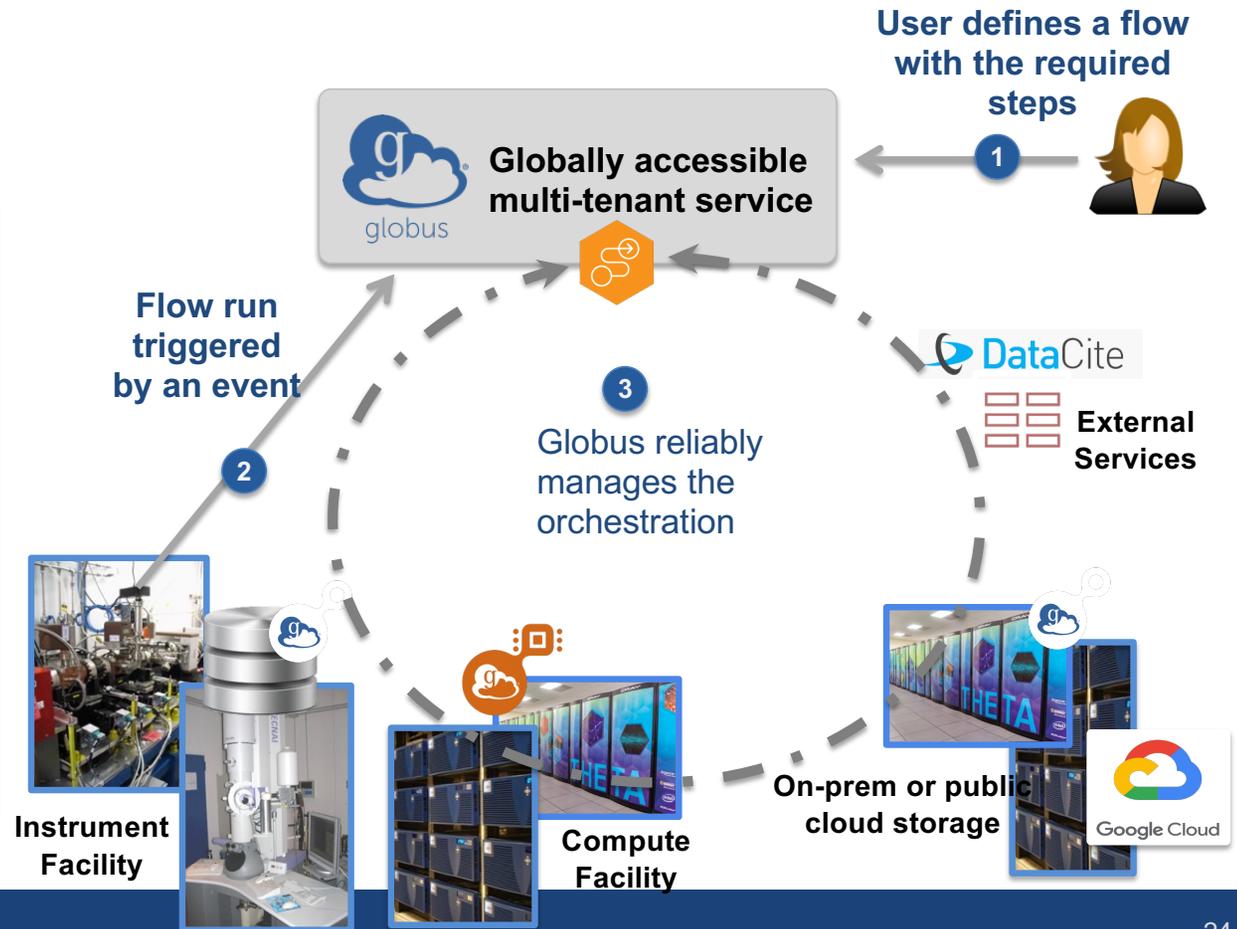
Compute ecosystem



**Deploy Globus Compute agent
– single user or multi user to
connect into the ecosystem**

Reliable automation ...spanning all resources

- Managed reliable task orchestration
- Declarative language for flow definition
- Event driven execution model
- Extensible to integrate external services





Flow lifecycle

- Define flow and input schema using JSON

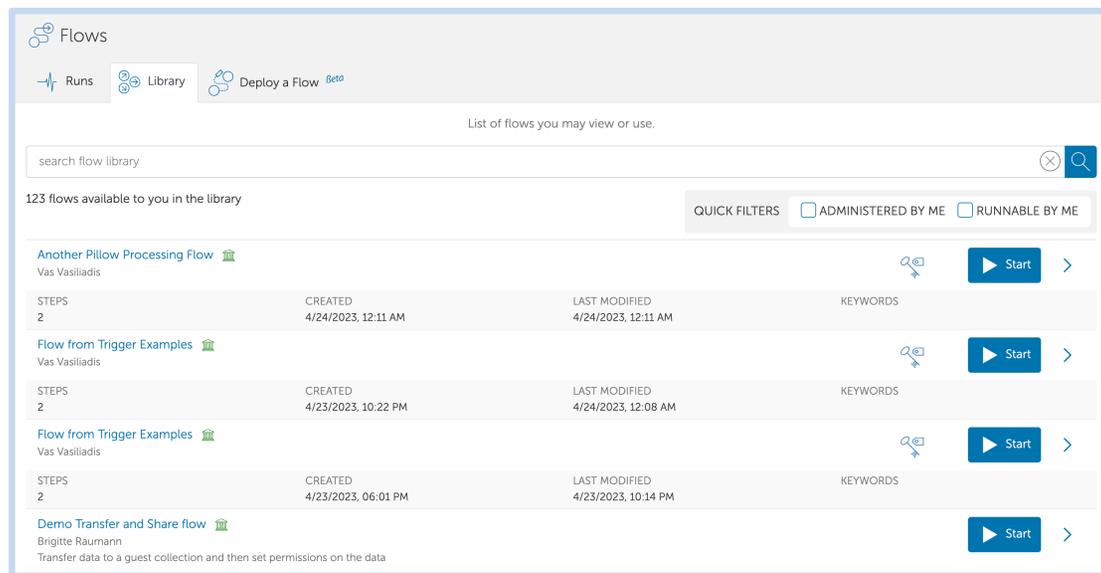
globus.github.io/flows-ide/

```
{
  "States": {
    "ProcessFiles": {
      "End": true,
      "Type": "Action",
      "Comment": "Process files - generate thumbnails",
      "WaitTime": 180,
      "ActionUrl": "https://compute.actions.globus.org/fxap",
      "Parameters": {
        "kwargs.$": "$.input.compute_function_kwargs",
        "endpoint.$": "$.input.compute_endpoint_id",
        "function.$": "$.input.compute_function_id"
      },
      "ResultPath": "$.ProcessFiles"
    },
    "TransferFiles": {
      "Next": "ProcessFiles",
      "Type": "Action",
      "Comment": "Transfer to a guest collection",
      "WaitTime": 60,
      "ActionUrl": "https://actions.automate.globus.org/transfer/transfer",
      "Parameters": {
        "transfer_items": [
          {
            "recursive.$": "$.input.recursive_tx",
            "source_path.$": "$.input.source.path",
            "destination_path.$": "$.input.destination.path"
          }
        ],
        "source_endpoint_id.$": "$.input.source.id",
        "destination_endpoint_id.$": "$.input.destination.id"
      },
      "ResultPath": "$.TransferFiles"
    }
  },
  "Comment": "Transfer and process files by invoking a funcX function",
  "StartAt": "TransferFiles"
}
```

Flow lifecycle



- Define flow and input schema JSON
- **Deploy to Flows service**



The screenshot displays the Google Cloud Flows service interface. At the top, there are navigation tabs for 'Runs', 'Library', and 'Deploy a Flow ^{Beta}'. Below the navigation is a search bar labeled 'search flow library' and a 'List of flows you may view or use.' section. A filter section shows '123 flows available to you in the library' and 'QUICK FILTERS' with options for 'ADMINISTERED BY ME' and 'RUNNABLE BY ME'. The main content area lists several flows, each with a 'Start' button and a chevron icon.

Flow Name	Author	Steps	Created	Last Modified	Keywords	Actions
Another Pillow Processing Flow	Vas Vasiladis	2	4/24/2023, 12:11 AM	4/24/2023, 12:11 AM		Start >
Flow from Trigger Examples	Vas Vasiladis	2	4/23/2023, 10:22 PM	4/24/2023, 12:08 AM		Start >
Flow from Trigger Examples	Vas Vasiladis	2	4/23/2023, 06:01 PM	4/23/2023, 10:14 PM		Start >
Demo Transfer and Share flow	Brigitte Raumann				Transfer data to a guest collection and then set permissions on the data	Start >

Flow lifecycle



- Define flow and input schema JSON
- Deploy to Flows service
- **Set access policy for visibility and execution**

 Assign New Role

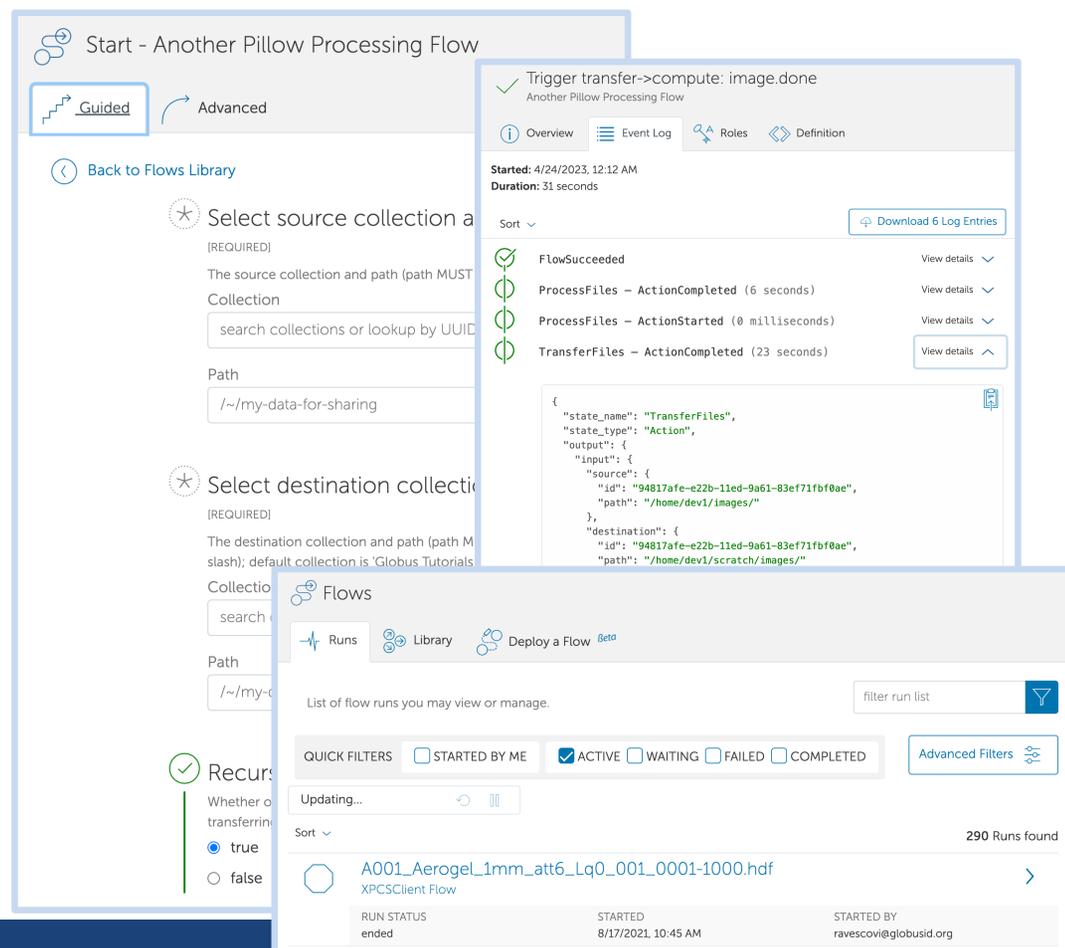
Assign To User
 Group
 All Logged In Users
 Public (anonymous users)

Group Tutorial Users

Role Administered By
can start this flow, view this flow and associated activity, and modify this flow
 Runnable By
can start this flow and view associated activity
 Visible To
can view this flow and associated activity

Flow lifecycle

- Define flow and input schema JSON
- Deploy to Flows service
- Set access policy for visibility and execution
- **Run (debug) and monitor**



The screenshot displays the 'Start - Another Pillow Processing Flow' configuration page. It includes a 'Guided' tab, a 'Back to Flows Library' button, and sections for selecting source and destination collections. The source collection is set to 'Collection' with path '/~/my-data-for-sharing'. The destination collection is also 'Collection' with path '/~/my-c'. A 'Recursive' checkbox is checked.

An execution log window is open, showing the following details:
Trigger: transfer->compute: image.done
Started: 4/24/2023, 12:12 AM
Duration: 31 seconds
Log entries:
- FlowSucceeded (View details)
- ProcessFiles - ActionCompleted (6 seconds) (View details)
- ProcessFiles - ActionStarted (0 milliseconds) (View details)
- TransferFiles - ActionCompleted (23 seconds) (View details)
JSON output:

```
{
  "state_name": "TransferFiles",
  "state_type": "Action",
  "output": {
    "input": {
      "source": {
        "id": "94817afe-e22b-11ed-9a61-83ef71fb0ae",
        "path": "/home/dev1/images/"
      },
      "destination": {
        "id": "94817afe-e22b-11ed-9a61-83ef71fb0ae",
        "path": "/home/dev1/scratch/images/"
      }
    }
  }
}
```

The 'Flows' section shows a 'Runs' tab with a list of 290 runs. A quick filter is set to 'ACTIVE'. One run is highlighted:

Run ID	Run Status	Started	Started By
A001_Aerogel1mm_att6_Lq0_001_0001-1000.hdf	ended	8/17/2021 10:45 AM	ravescov@globusid.org



Flow lifecycle: Write once, run many



- Define flow and input schema JSON
- Deploy to Flows service
- Set access policy for visibility and execution
- Run (debug) and monitor
- **...and run again!**

The screenshot displays the 'Flows' management interface. At the top, there are navigation tabs for 'Runs', 'Library', and 'Deploy a Flow beta'. Below this, a search bar labeled 'filter run list' is present. A 'QUICK FILTERS' section includes checkboxes for 'STARTED BY ME', 'ACTIVE', 'WAITING' (which is checked), 'FAILED', and 'COMPLETED'. An 'Advanced Filters' button is also visible. A 'List Update in 14' indicator with refresh and pause icons is shown. The main area displays a table of flow runs, with a 'Sort' dropdown and '81 Runs found' indicator. The table lists five runs, each with a unique ID, flow name, status, start time, and user.

Flow ID	Flow Name	Run Status	Started	Started By
A019_00003_Vol20_quenchT102p7ohms_att1_Rq0_0001-100000.hdf	XPCSCClient Flow	ended	8/24/2021, 06:23 PM	ravescovi@globusid.org
A019_00001_Vol20_quenchT102p7ohms_att1_Rq0_0001-100000.hdf	XPCSCClient Flow	ended	8/24/2021, 06:23 PM	ravescovi@globusid.org
A018_00002_Vol20_quenchT102p7ohms_att1_Rq0_0001-100000.hdf	XPCSCClient Flow	ended	8/24/2021, 06:23 PM	ravescovi@globusid.org
A016_00004_Vol20_quenchT102p7ohms_att1_Rq0_0001-100000.hdf	XPCSCClient Flow	ended	8/24/2021, 06:23 PM	ravescovi@globusid.org
A016_00003_Vol20_quenchT102p7ohms_att1_Rq0_0001-100000.hdf	XPCSCClient Flow	ended	8/24/2021, 06:23 PM	ravescovi@globusid.org

Run a flow

- **From the Globus web app**
 - Guided run with custom Globus construct support
- **Using CLI**
- **Using SDK or API**

 **Source** [REQUIRED]
Globus-provided flows require that at least one collection is managed under a subscription.

Collection

 **Intermediate** [REQUIRED]
Globus-provided flows require that at least one collection is managed under a subscription.

Collection

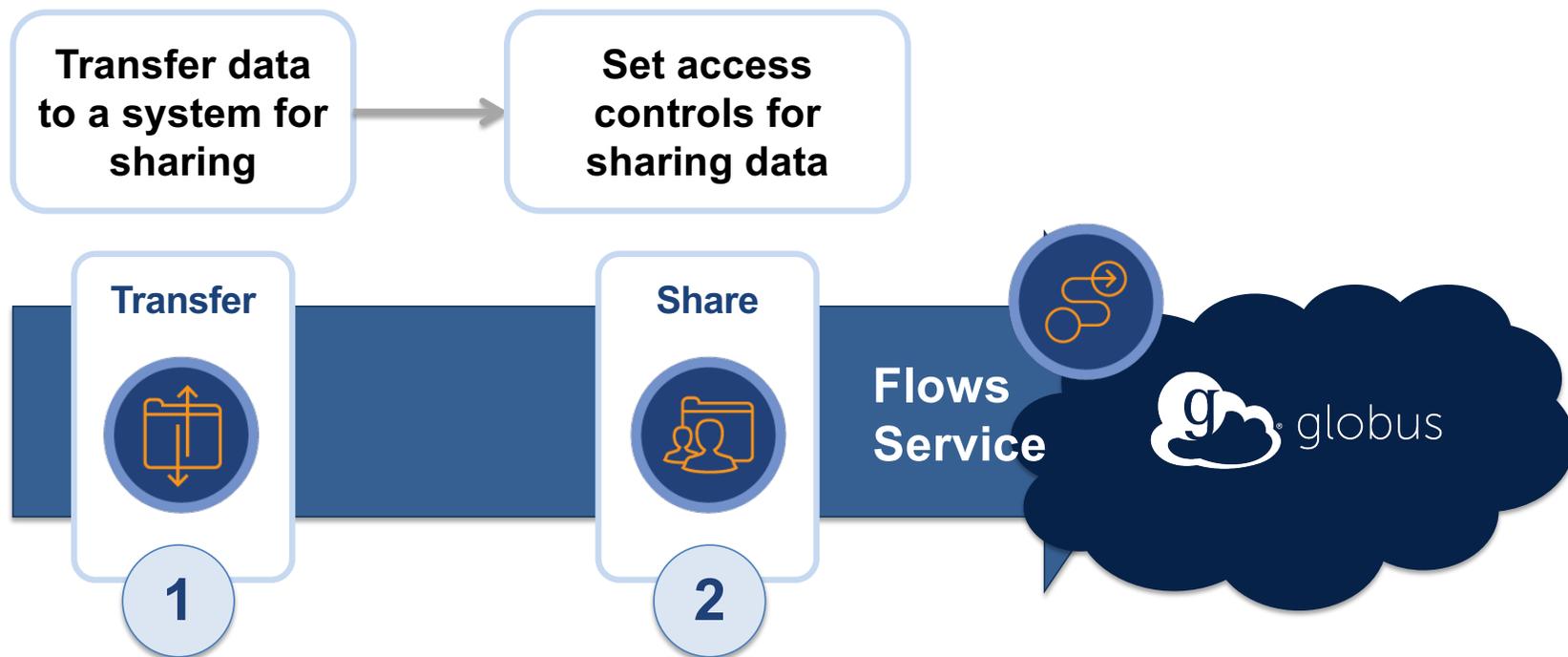
 **Destination** [REQUIRED]
Globus-provided flows require that at least one collection is managed under a subscription.

Collection

Automation by triggering a run when an event occurs

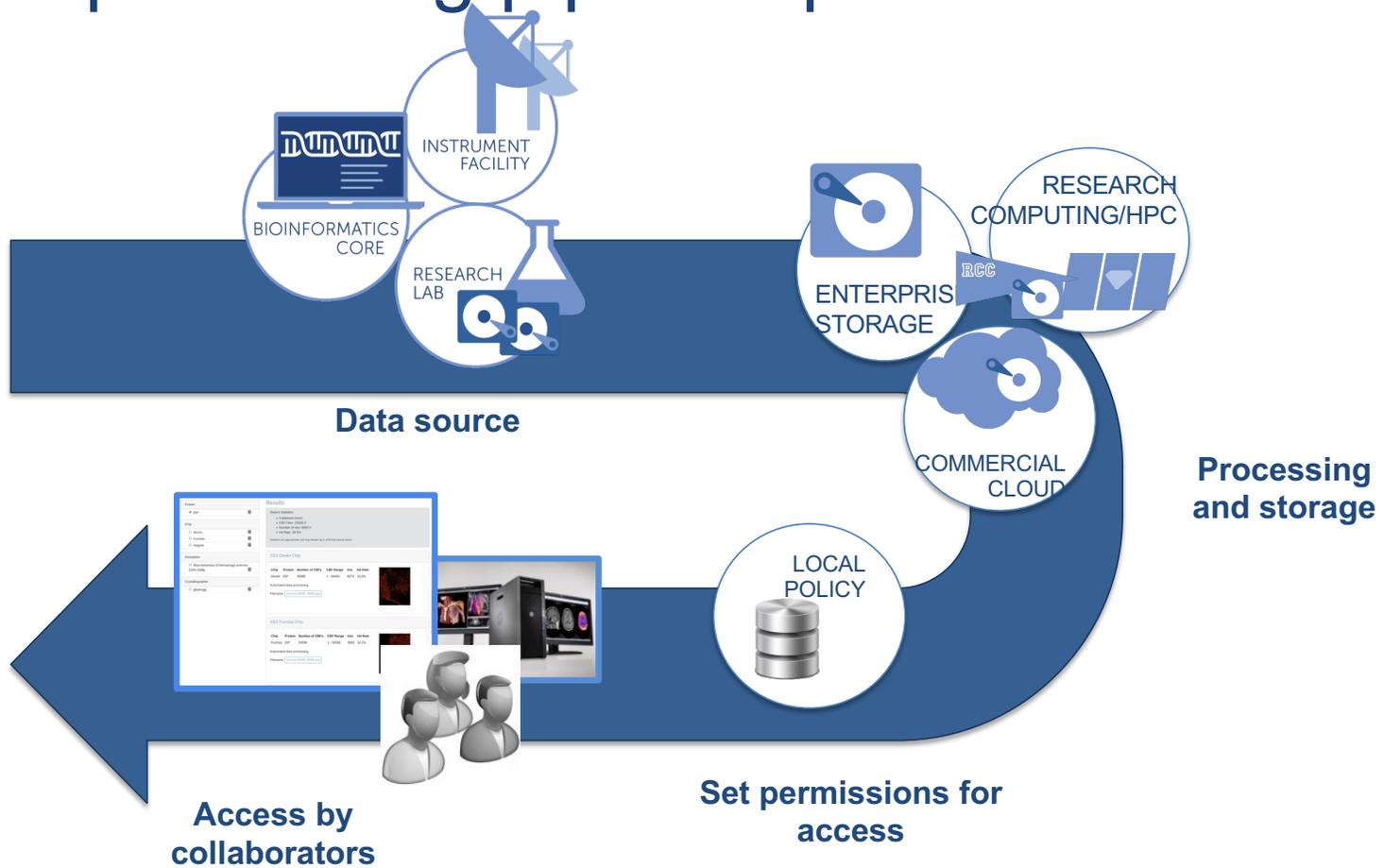


Transfer and share





Data processing pipeline pattern





SSX Automation



Compute



Launch QA job



Carbon!



Check threshold

Transfer



Transfer raw files

Compute



Analyze images

Image processing



Search



Ingest to index

Share



Set access controls

Transfer



Move results to repo

Compute

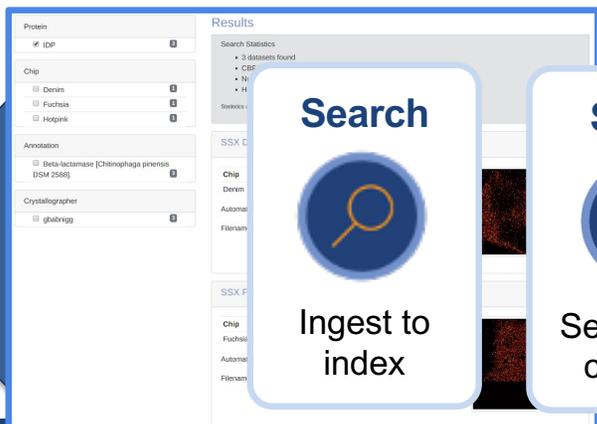
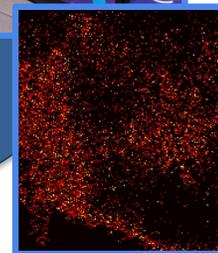


Gather metadata

Compute

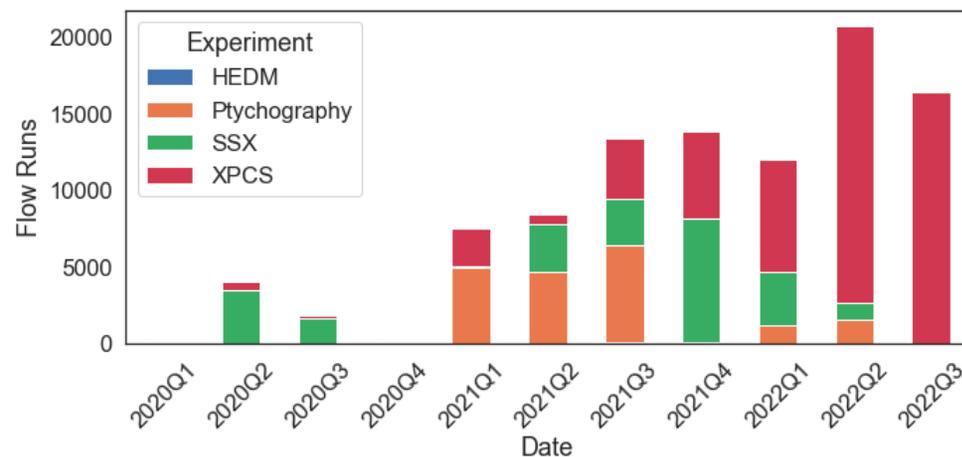
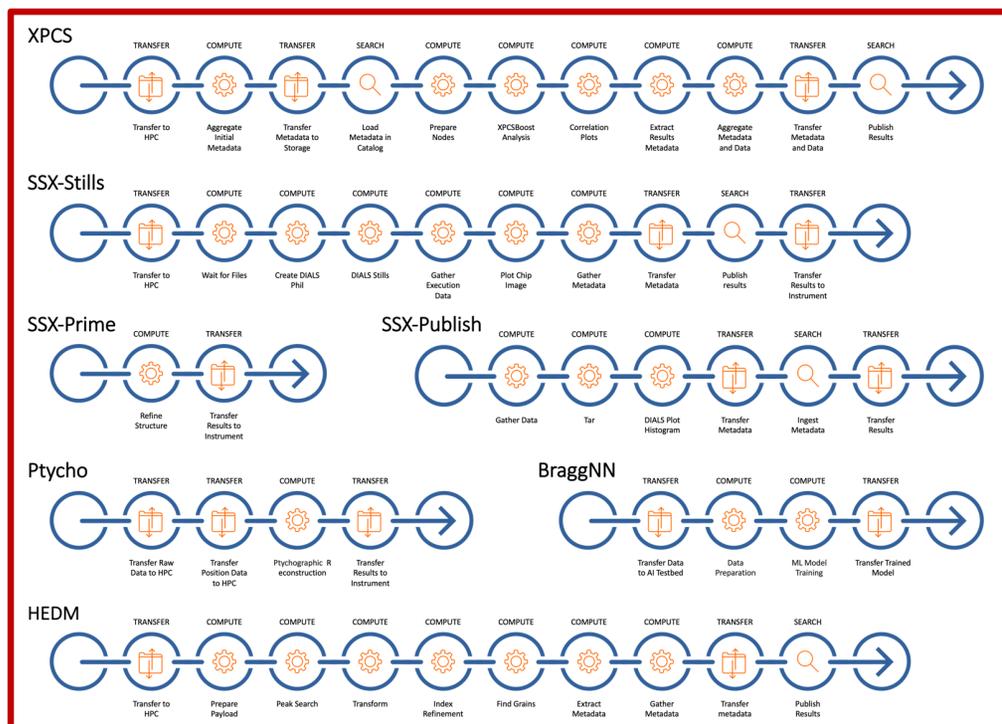


Visualize





Globus Flows at Advanced Photon Source



Patterns



Article

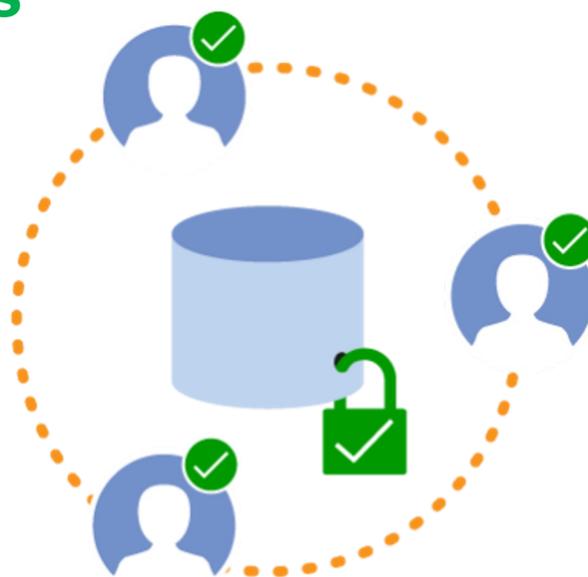
Linking scientific instruments and computation: Patterns, technologies, and experiences

Rafael Vescovi,¹ Ryan Chard,¹ Nickolaus D. Saint,⁶ Ben Blaiszik,^{1,6} Jim Pruyne,^{1,6} Tekin Bicer,^{1,3} Alex Lavens,⁴ Zhengchun Liu,¹ Michael E. Papka,^{2,7} Suresh Narayanan,³ Nicholas Schwarz,³ Kyle Chard,^{1,5} and Ian T. Foster^{1,5,*}

Globus High Assurance for managing protected data

Security controls

- NIST 800-53
- 800-171 Low+



Restricted data handling

- PHI, PII, CUI
- Compliant data sharing

BAA w/Uchicago

- UChicago BAA with Amazon

 Thank you!

- **Engage: ranantha@uchicago.edu**
- **Website: globus.org**
- **Documentation: docs.globus.org**
- **Support: support@globus.org**