Globus: platform for data driven research

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

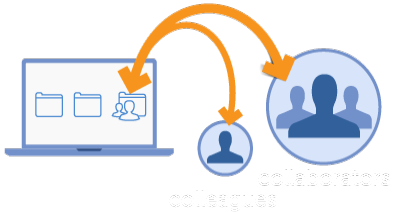
Globus is …

a non-profit service developed and operated

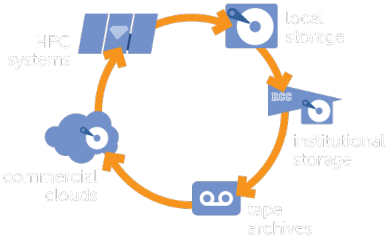
THE UNIVERSITY OF CHICAGO | RESEARCH

Globus Platform for Research IT

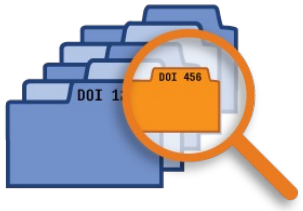
Managed transfer & sync



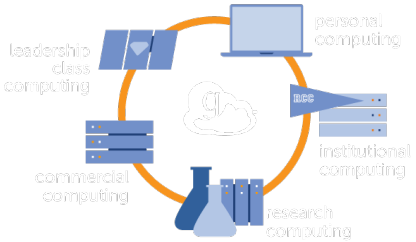
Collaborative data sharing



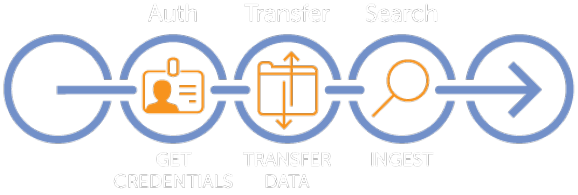
Unified data access



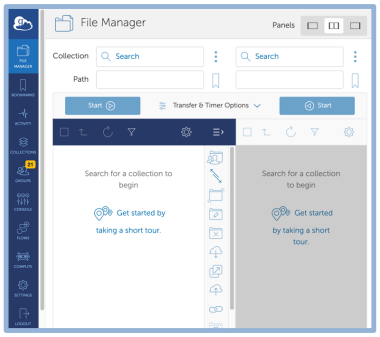
Publication & discovery



Managed remote execution



Reliable automation

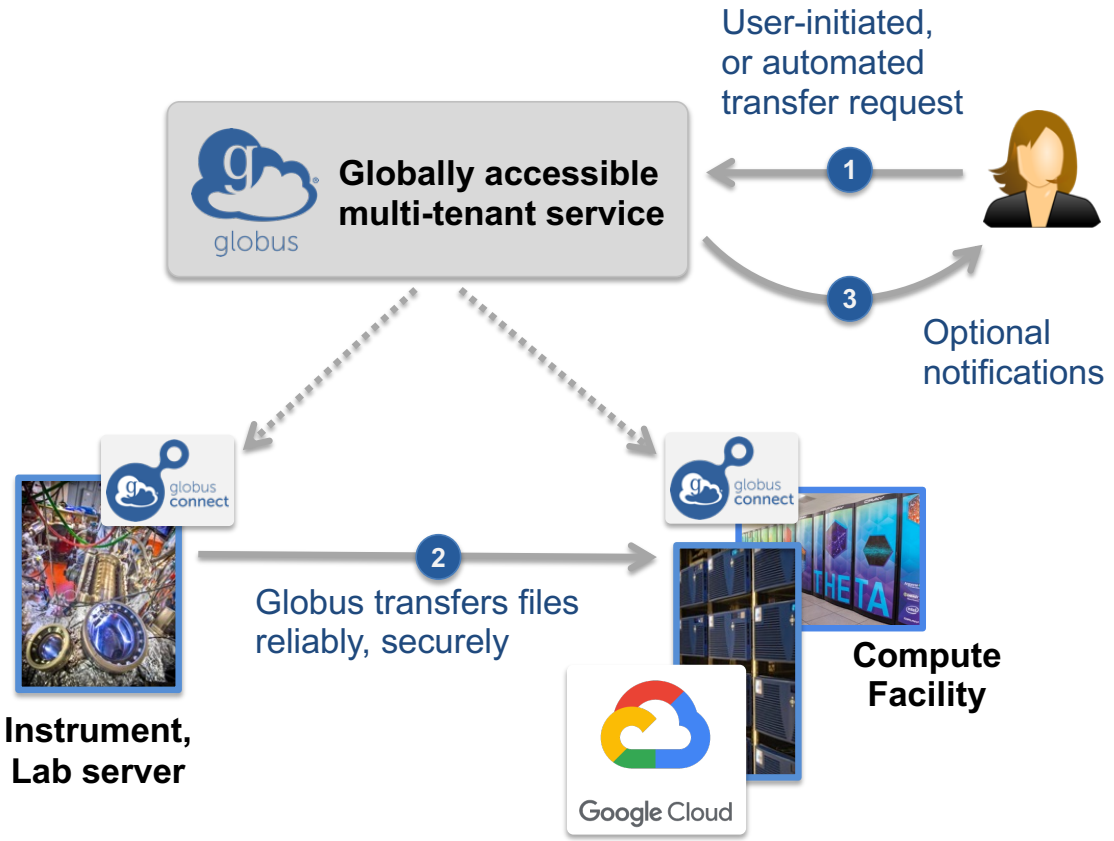


Software-as-a-Service

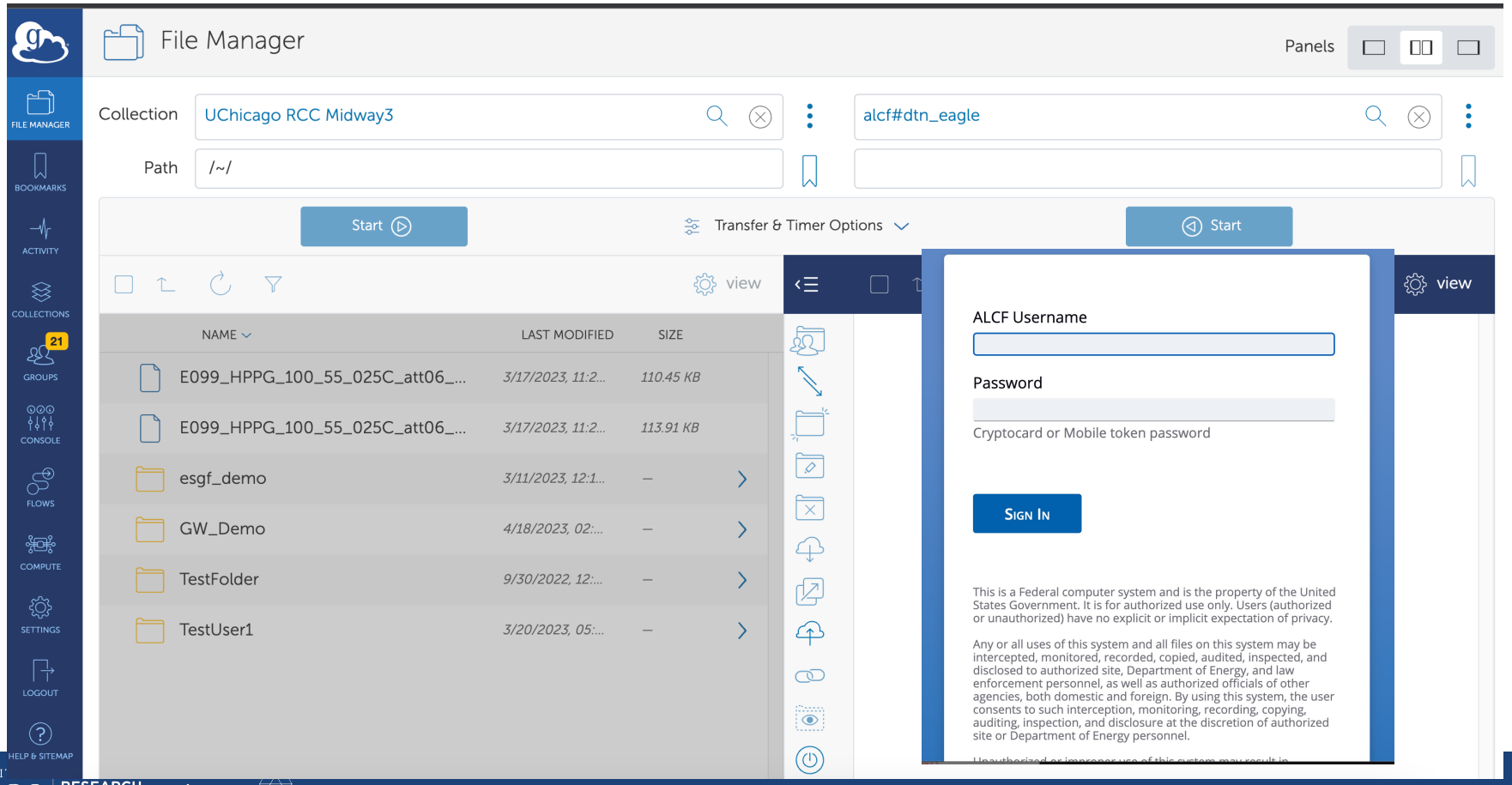
Platform-as-a-Service

Fast, reliable file transfer …from any to any system

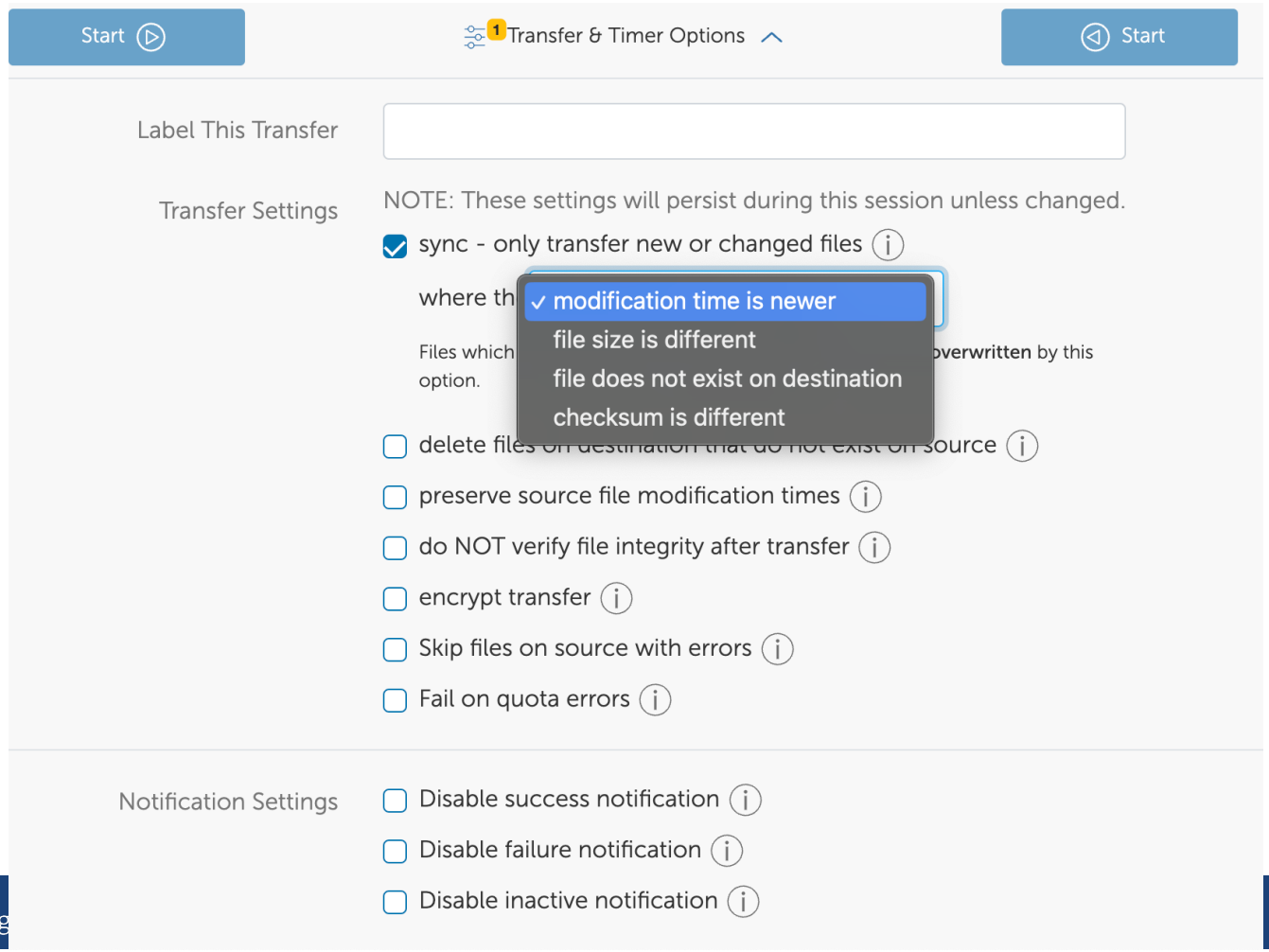
- **Fire-and-forget transfers/sync**
- **Optimized speed**
- **Assured reliability**
- **Unified view of storage**
- **HTTP/S access to data**



Intuitive web application interface



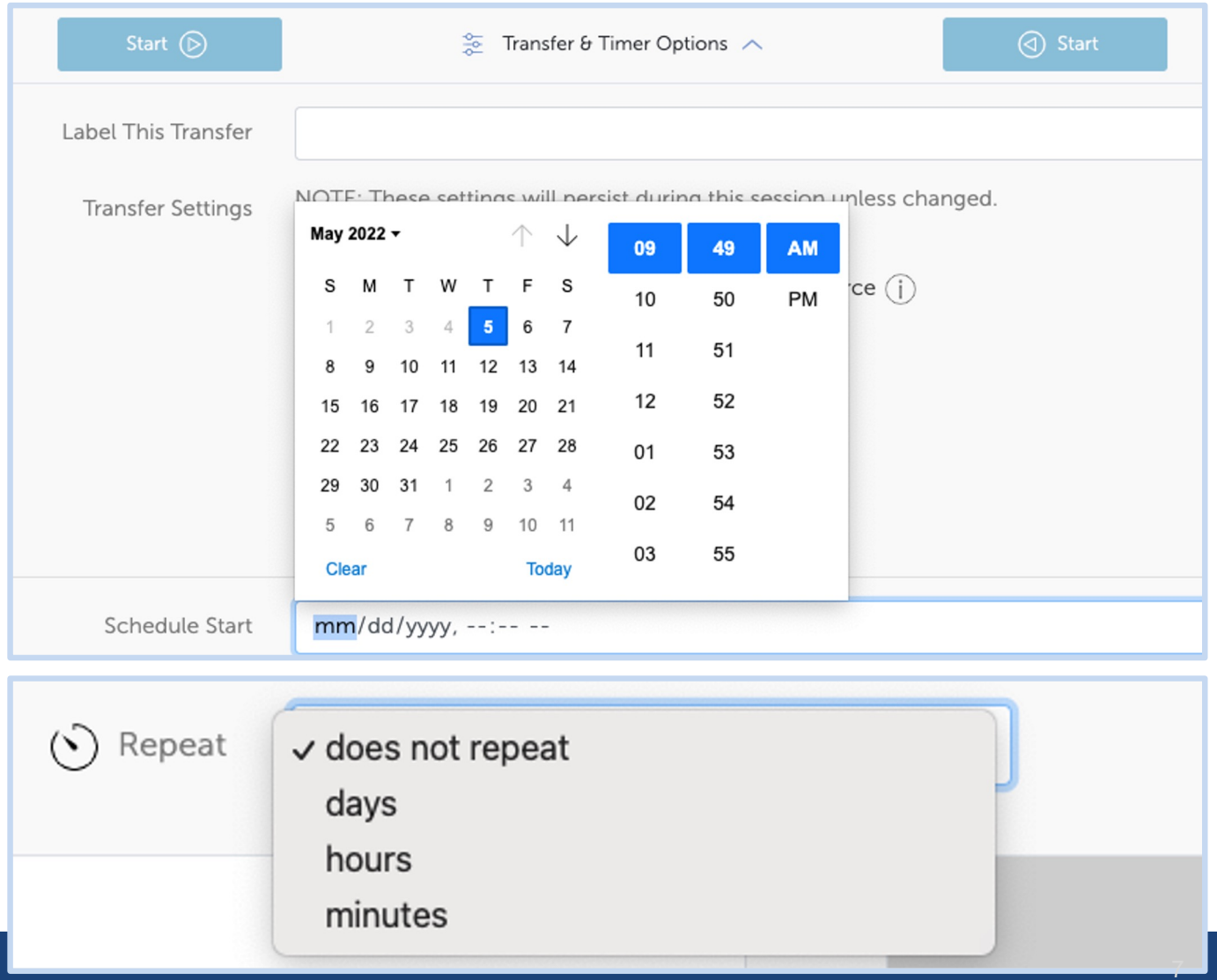
Transfer/sync options



Timers

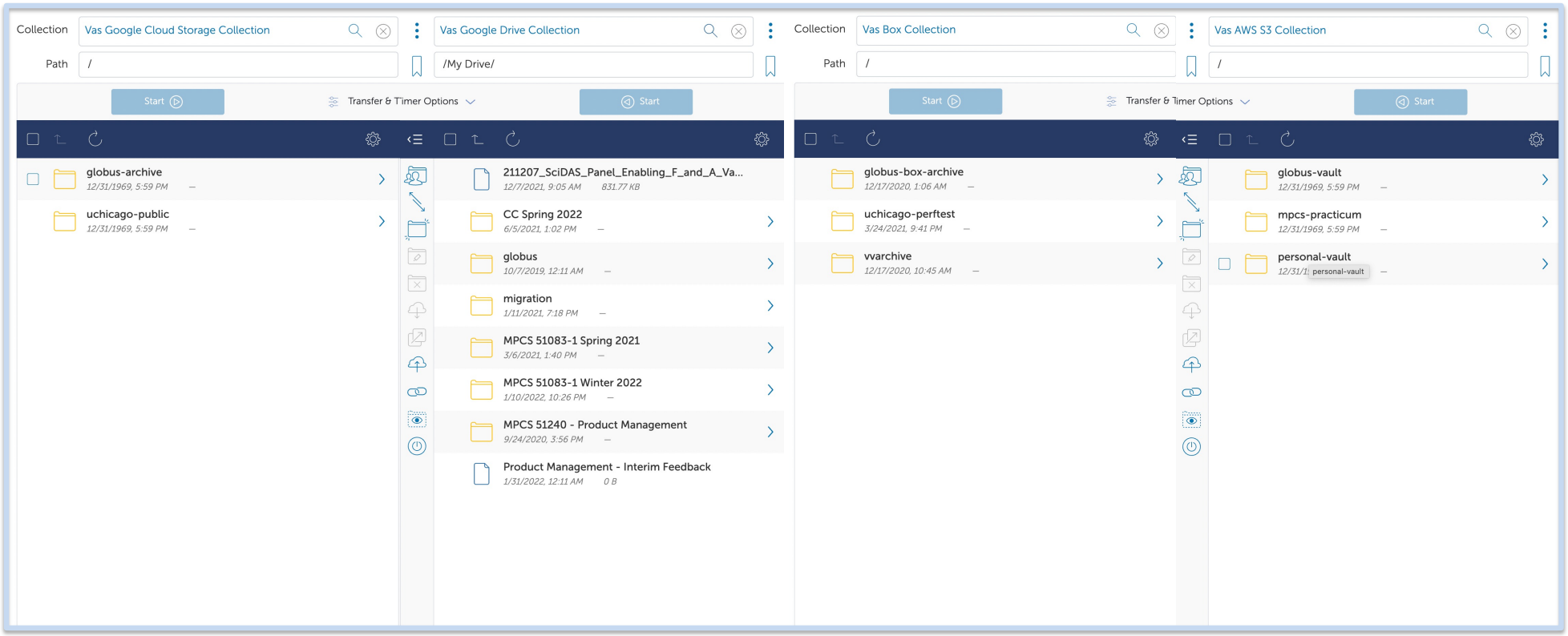
Scheduled and/or recurring file transfers

All Globus transfer and sync options



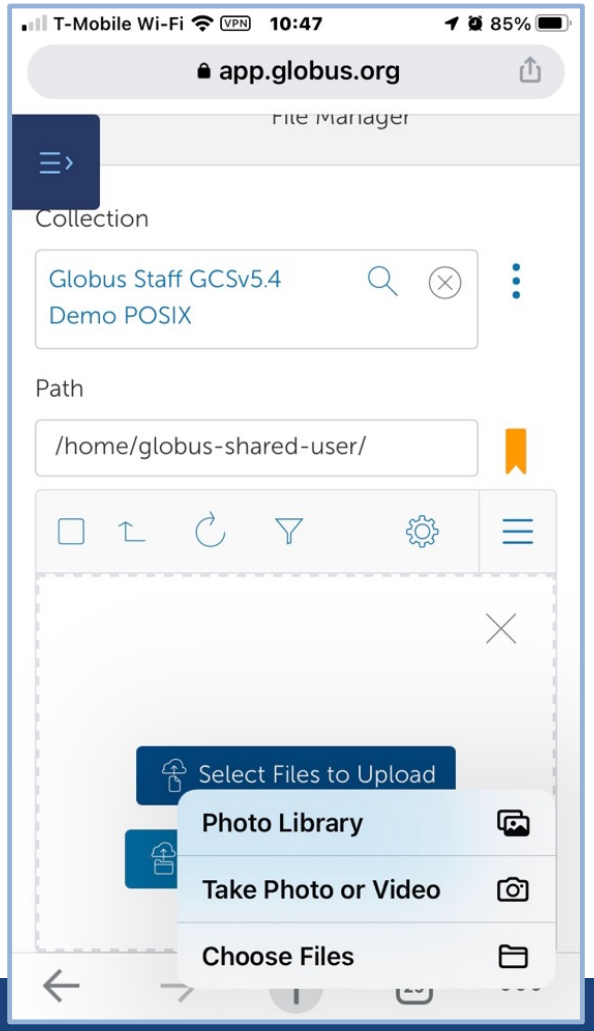
Supported storage systems

Uniform interface, consistent user experience



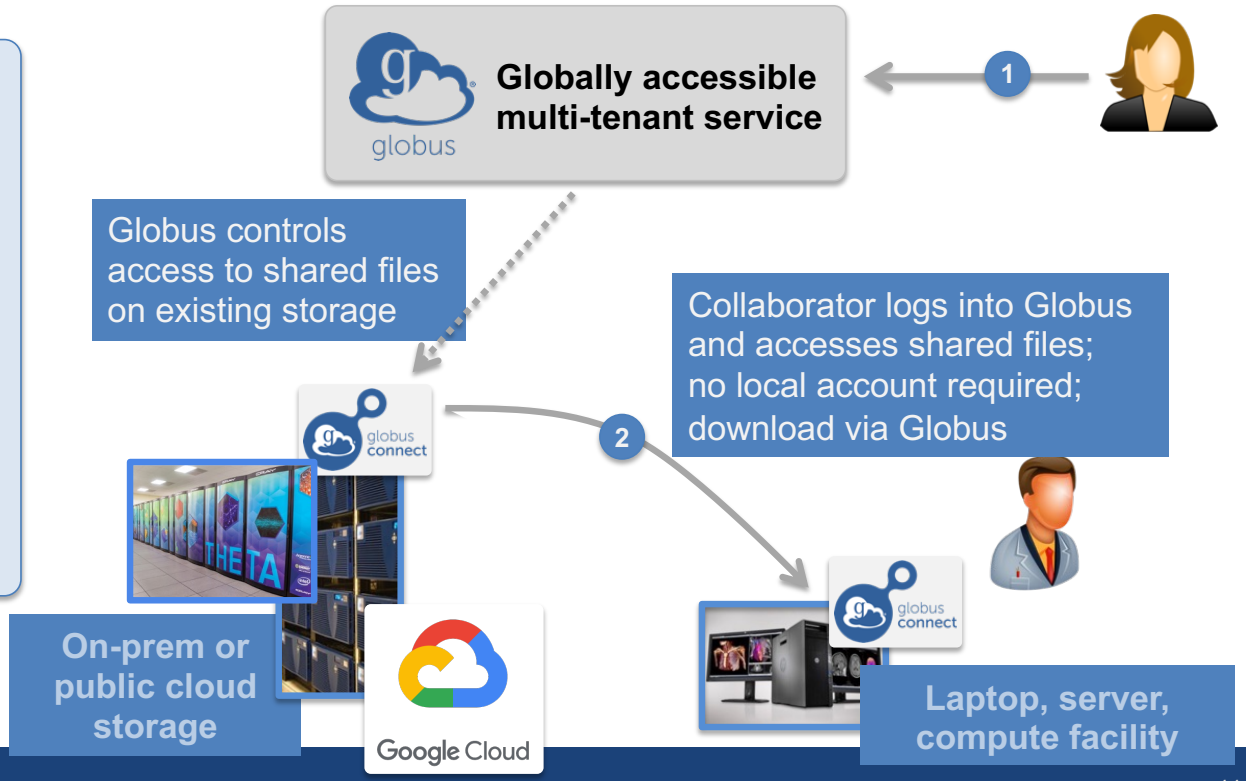
Globus goes on the road

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

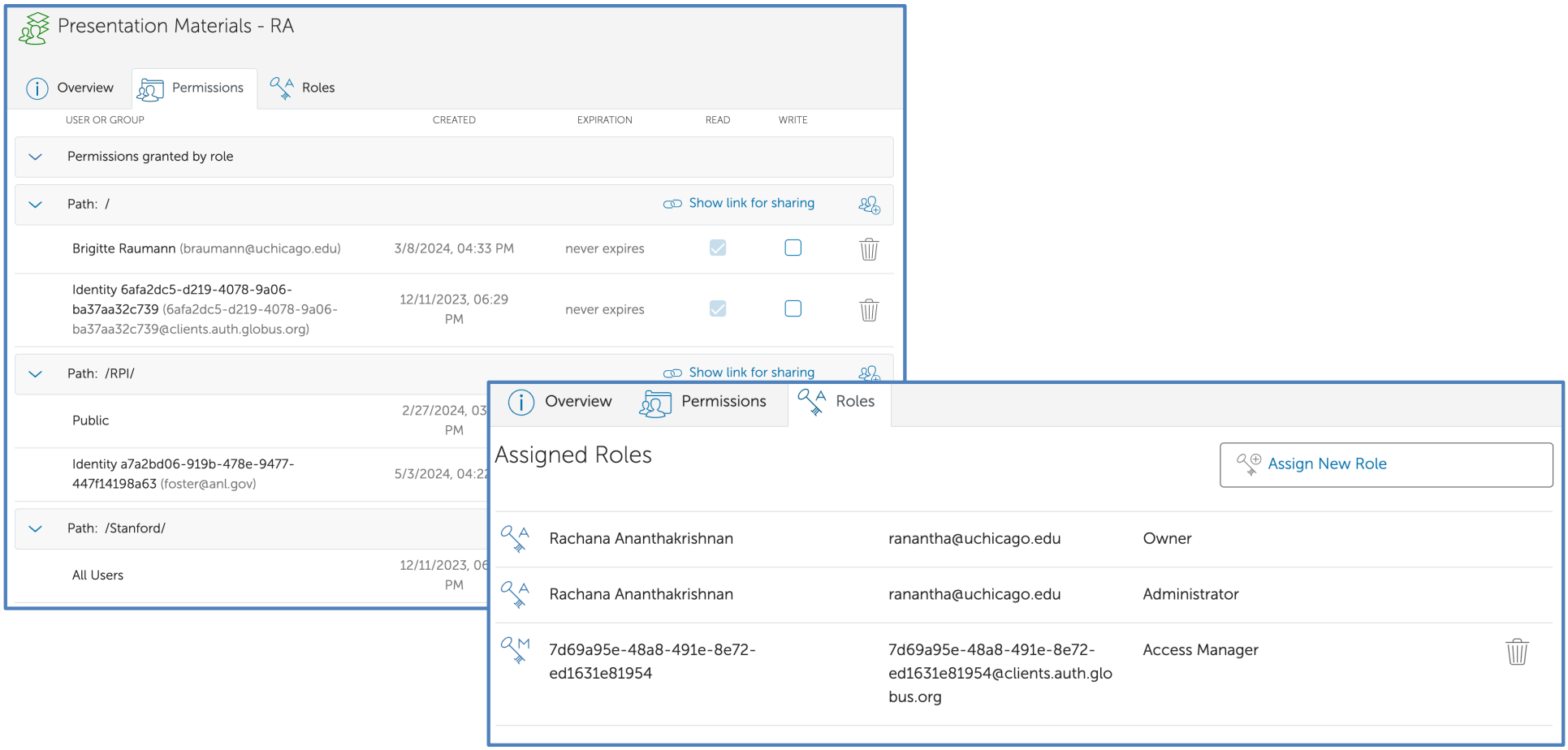


Secure data sharing …from any storage

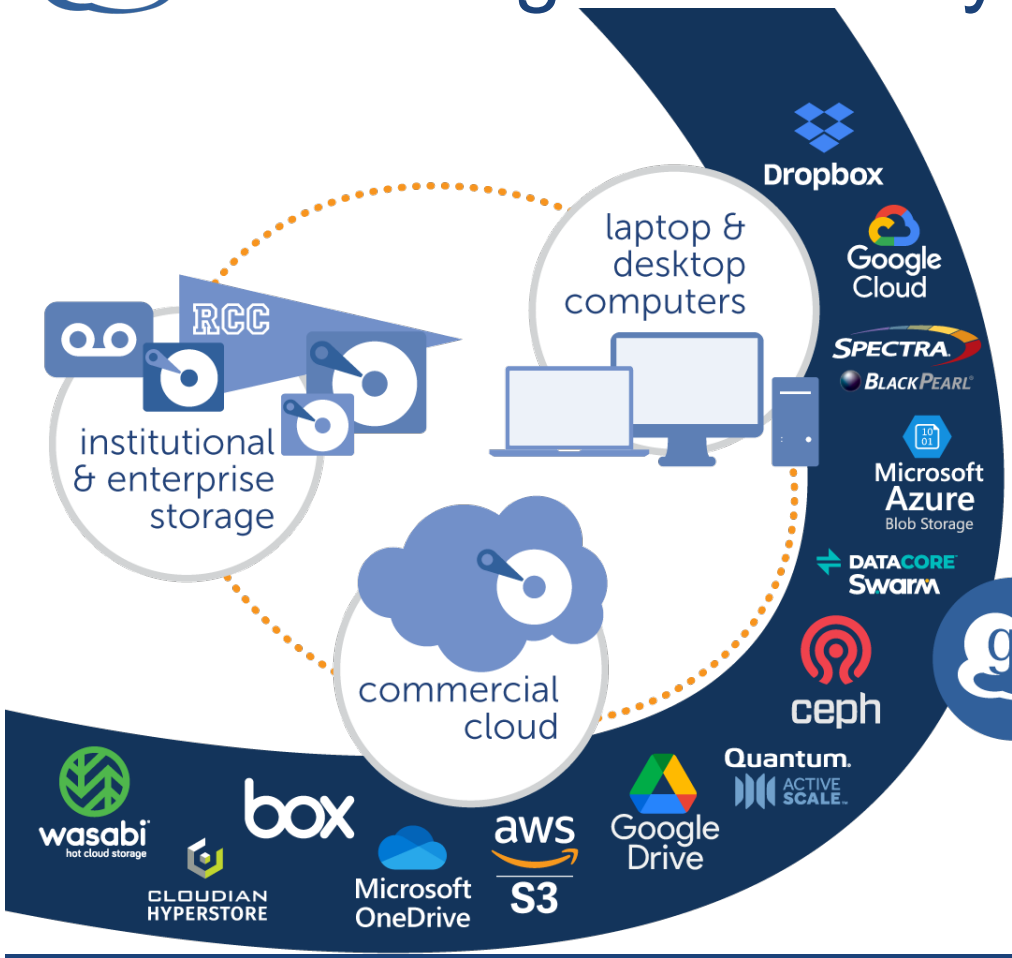
- **Fine-grained access control "overlay" on storage system**
- **Share with any identity, email, group**
- **No need to stage data just for sharing**



Data sharing – permissions & roles



Enabling data ecosystem

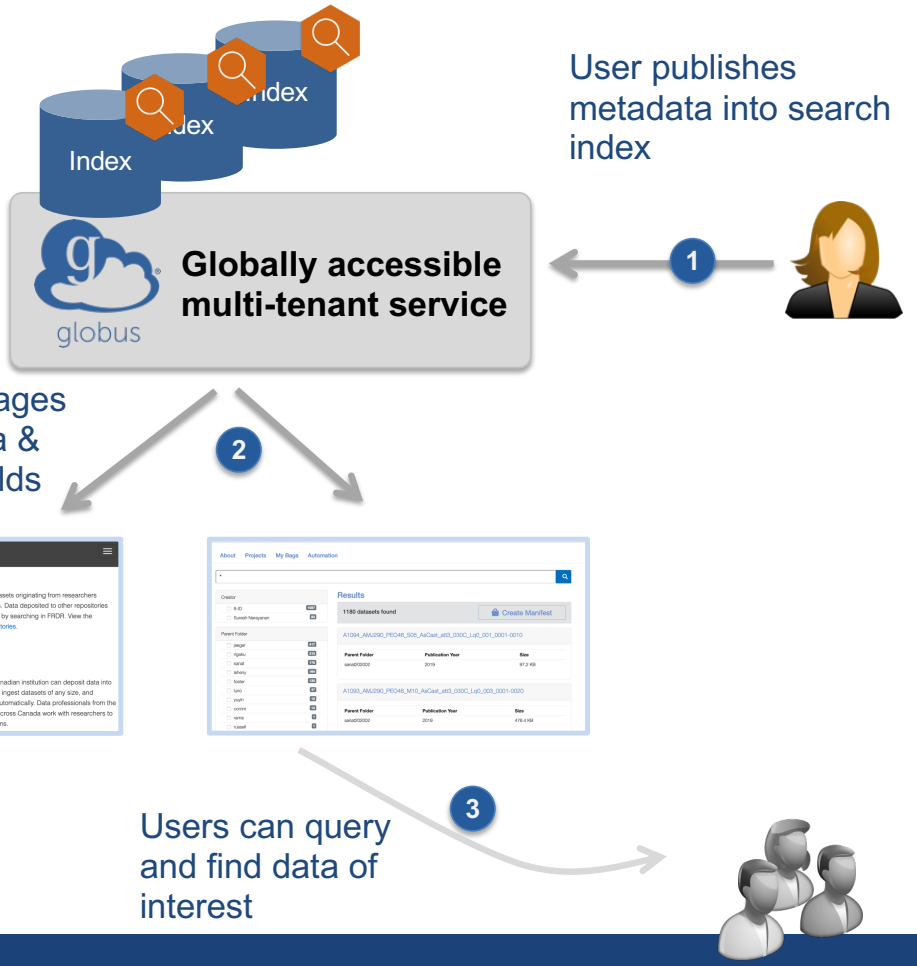


Data collections

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": ”<50lb”
      }
    },
    ...
  ]
}
```

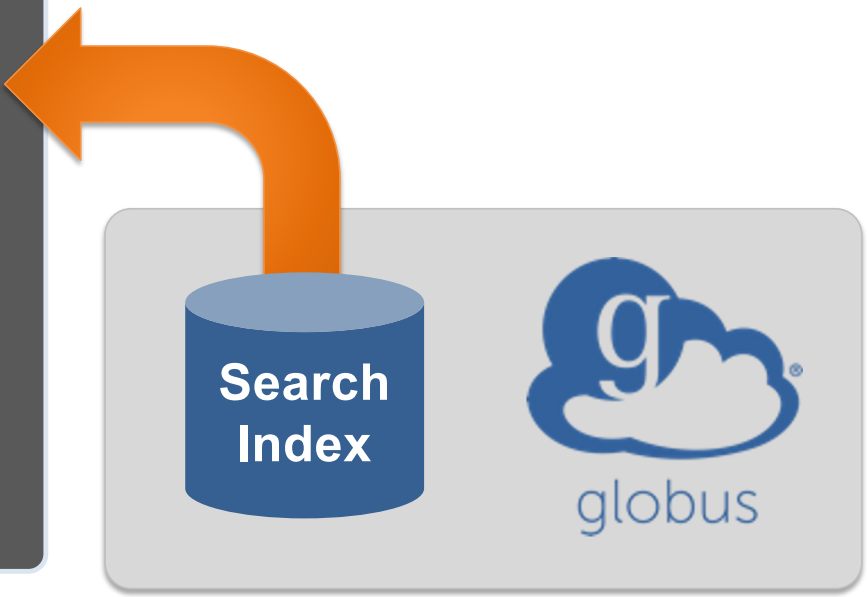
Visibility limited to Globus Auth identity
- Single user
- Globus Group
- Registered client application

Data discovery with Globus Search

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

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

Simple query



Data discovery with Globus Search

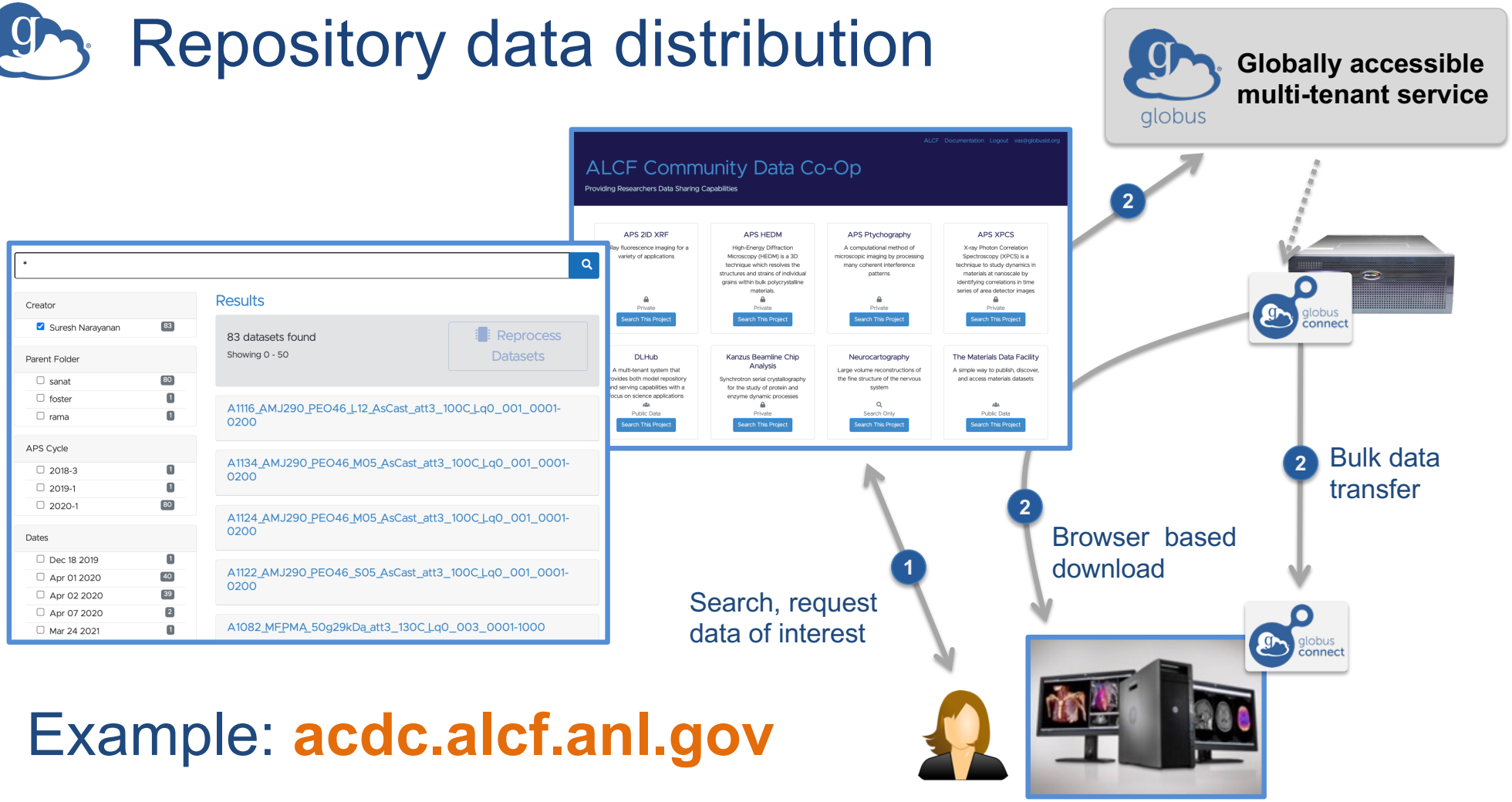
/index/{index_id}/search

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

Complex query

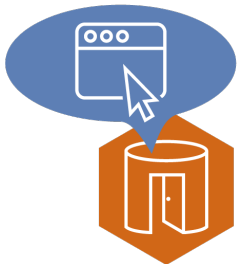
- Filter
- Facets
- Boosts
- Sort
- Limit

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.

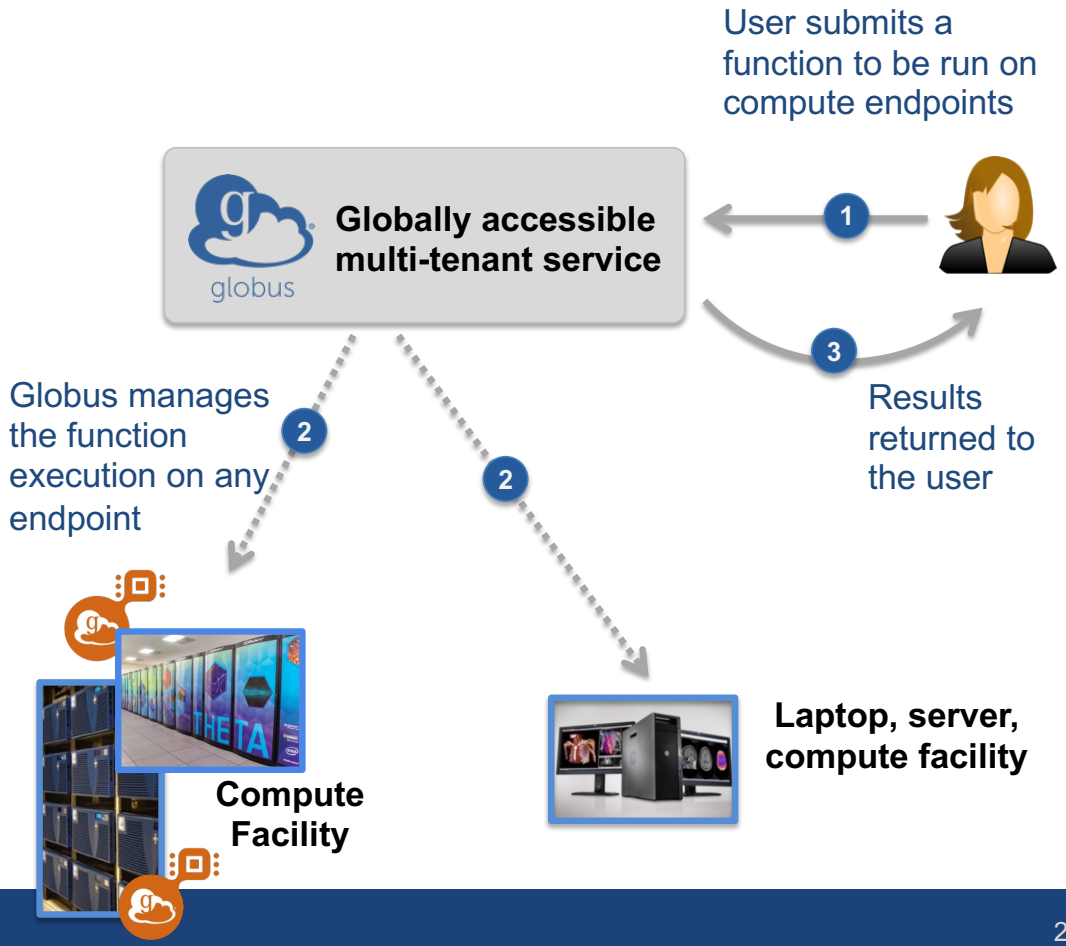
Custom Integrations

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

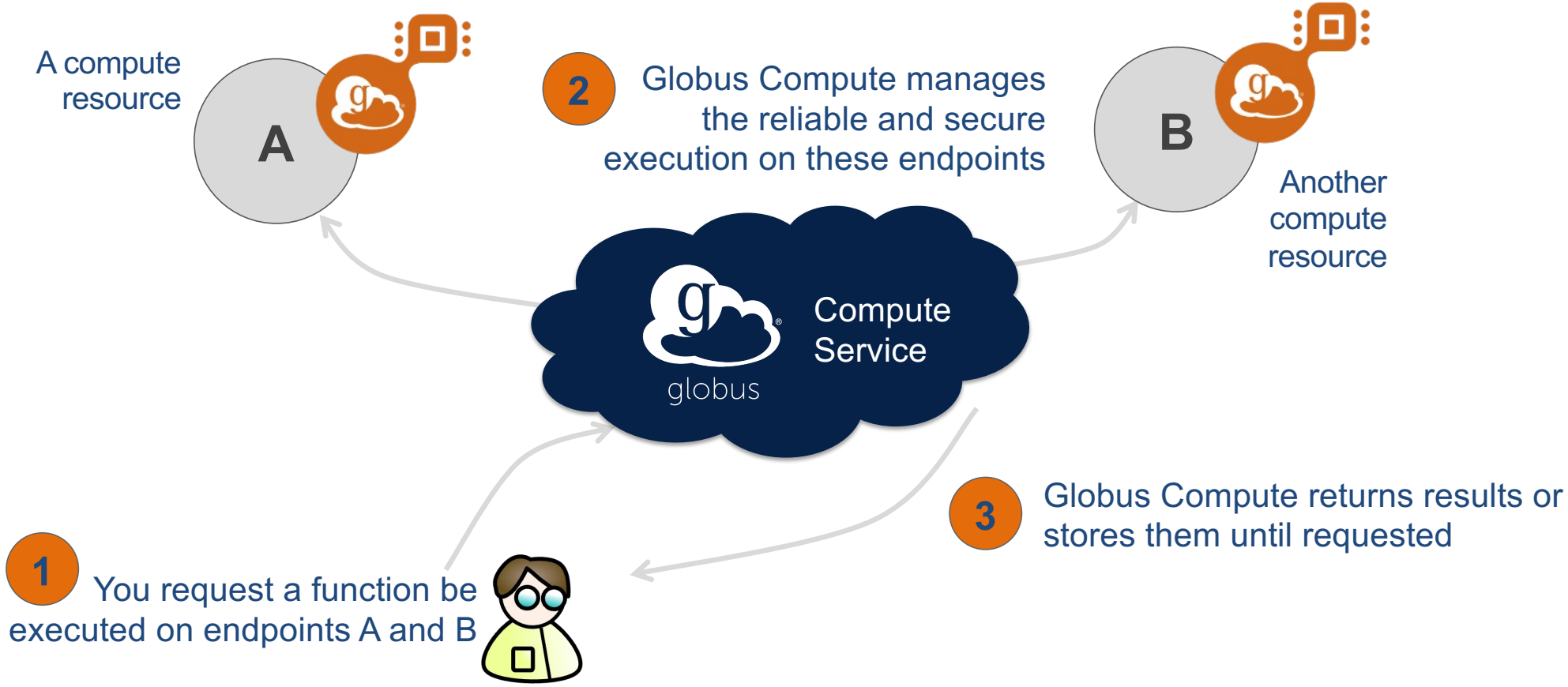
globus.org/portals

Managed compute …on any system

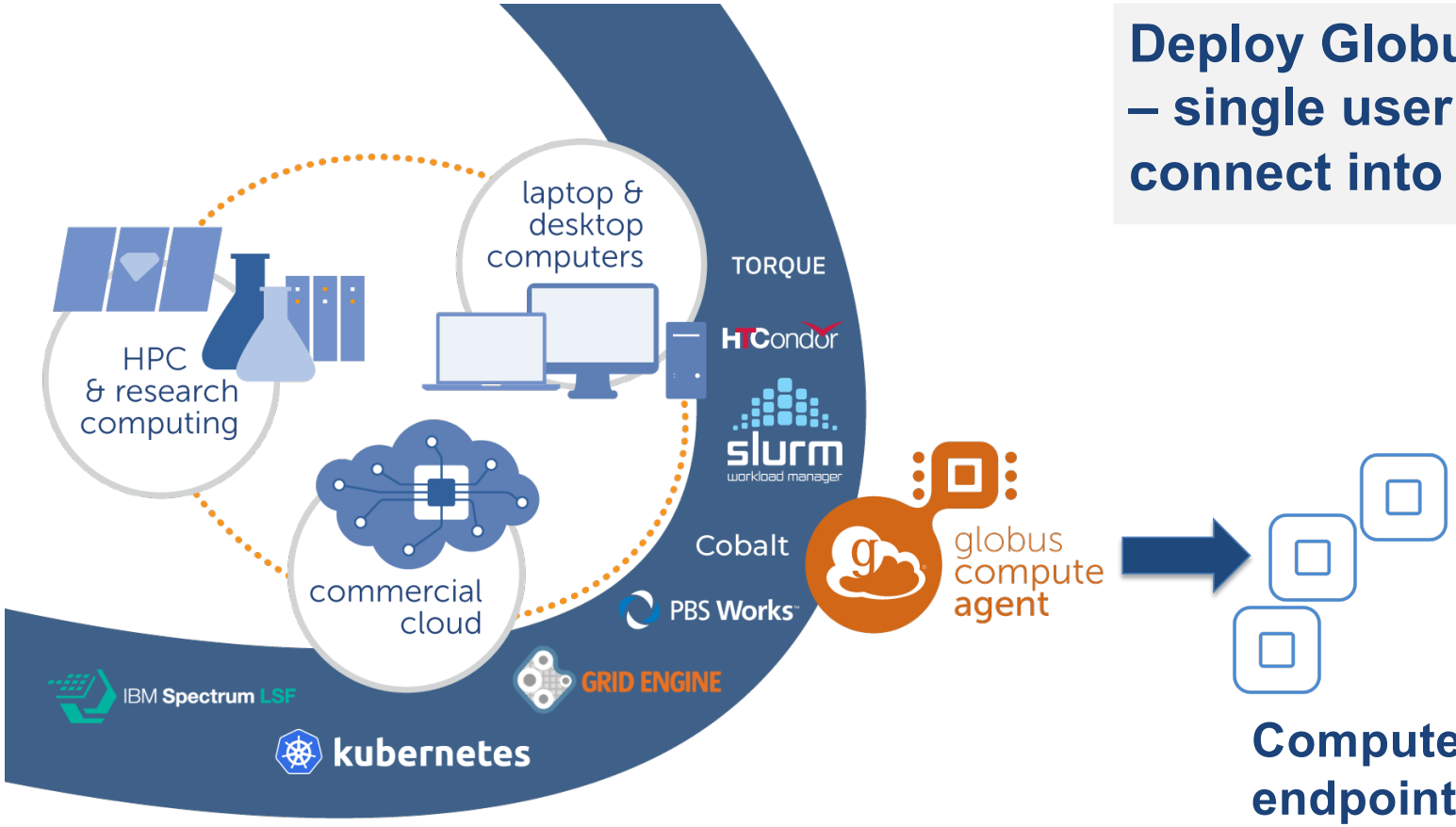
- **Fire and forget function execution**
- **Federated authentication, and local access control**
- **Uniform interface to various compute resources**
- **Support use of Python for functions**



User interaction with Globus Compute



Compute ecosystem

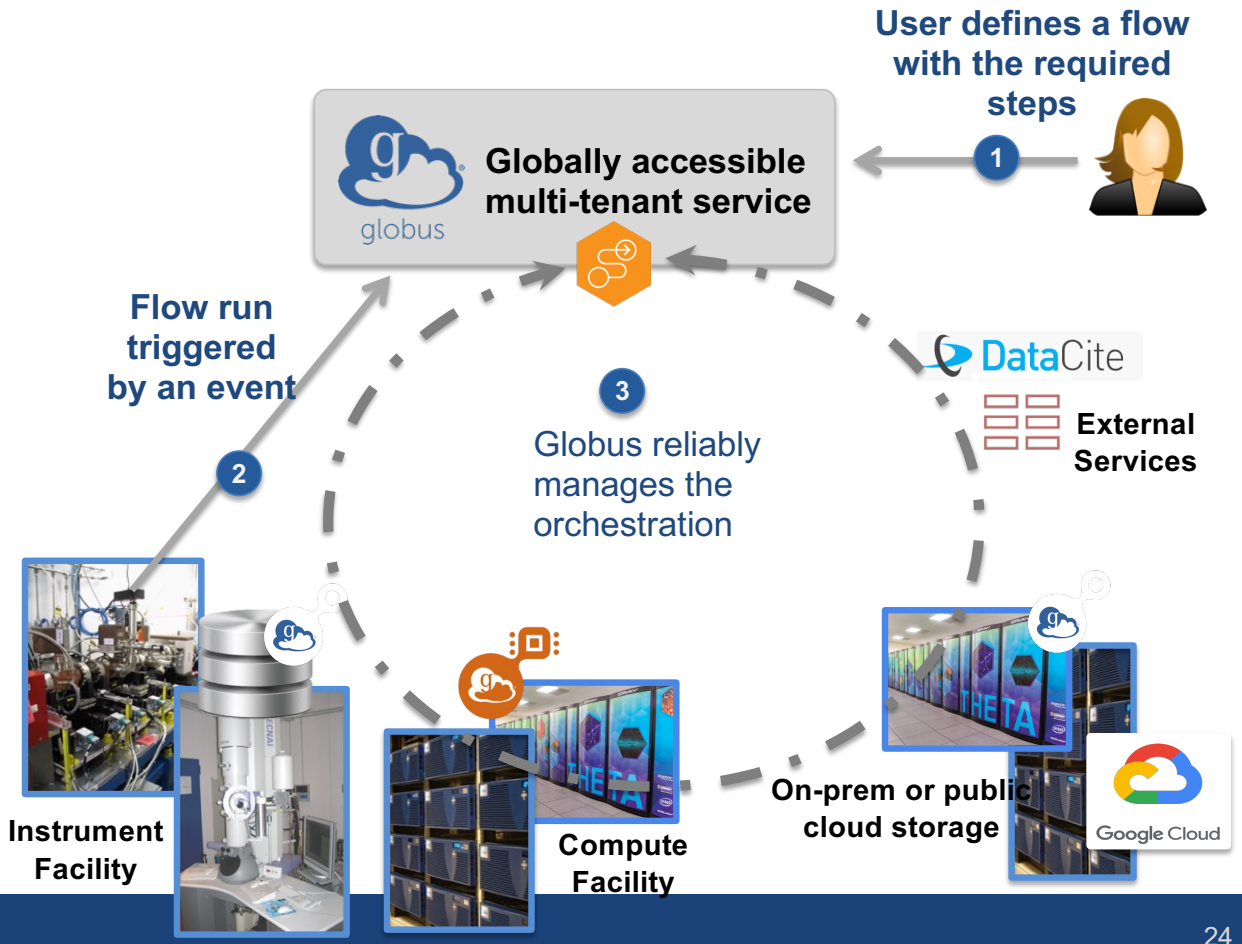


Deploy Globus Compute agent – single user or multi user to connect into the ecosystem

Compute endpoint

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

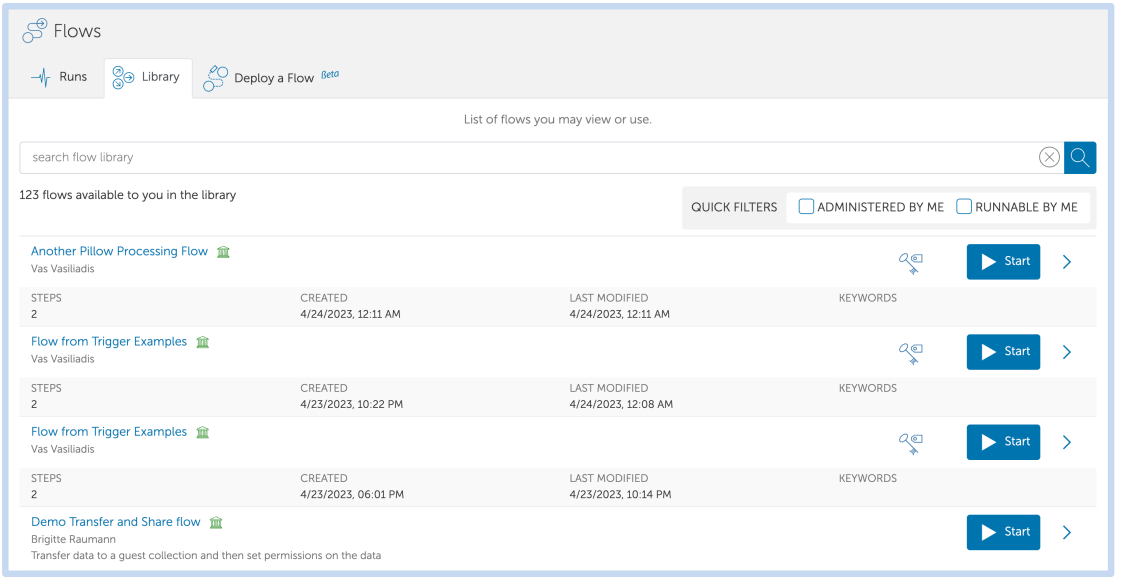
- **Define flow and input schema using JSON**

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"
}
```

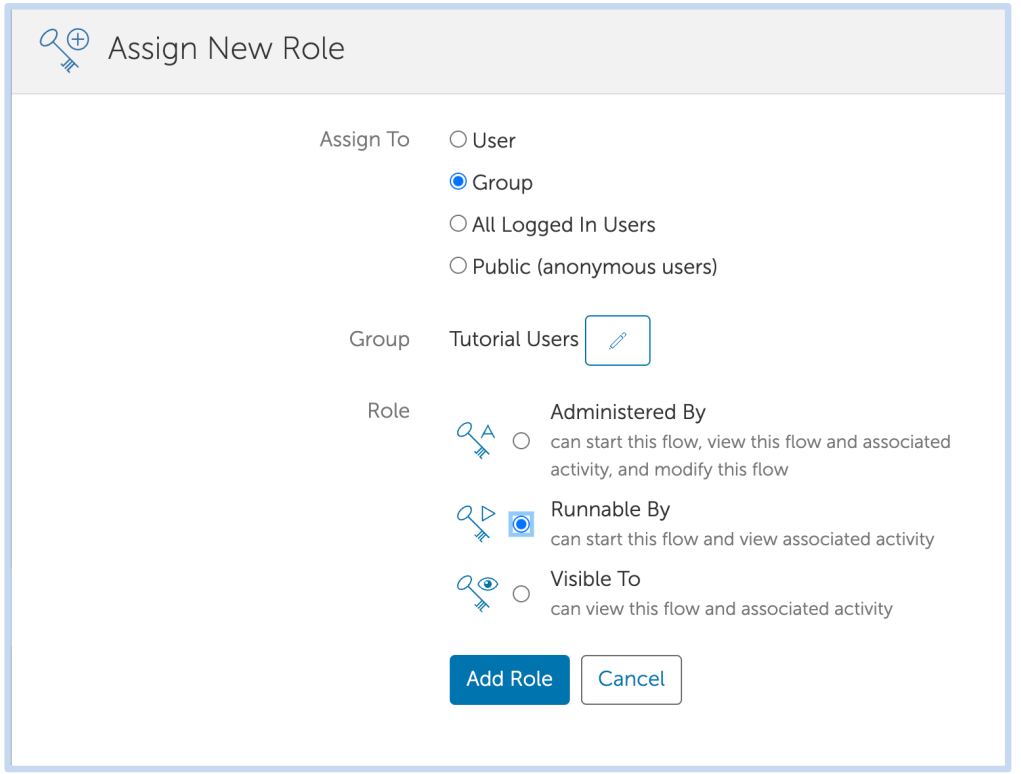
Flow lifecycle

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



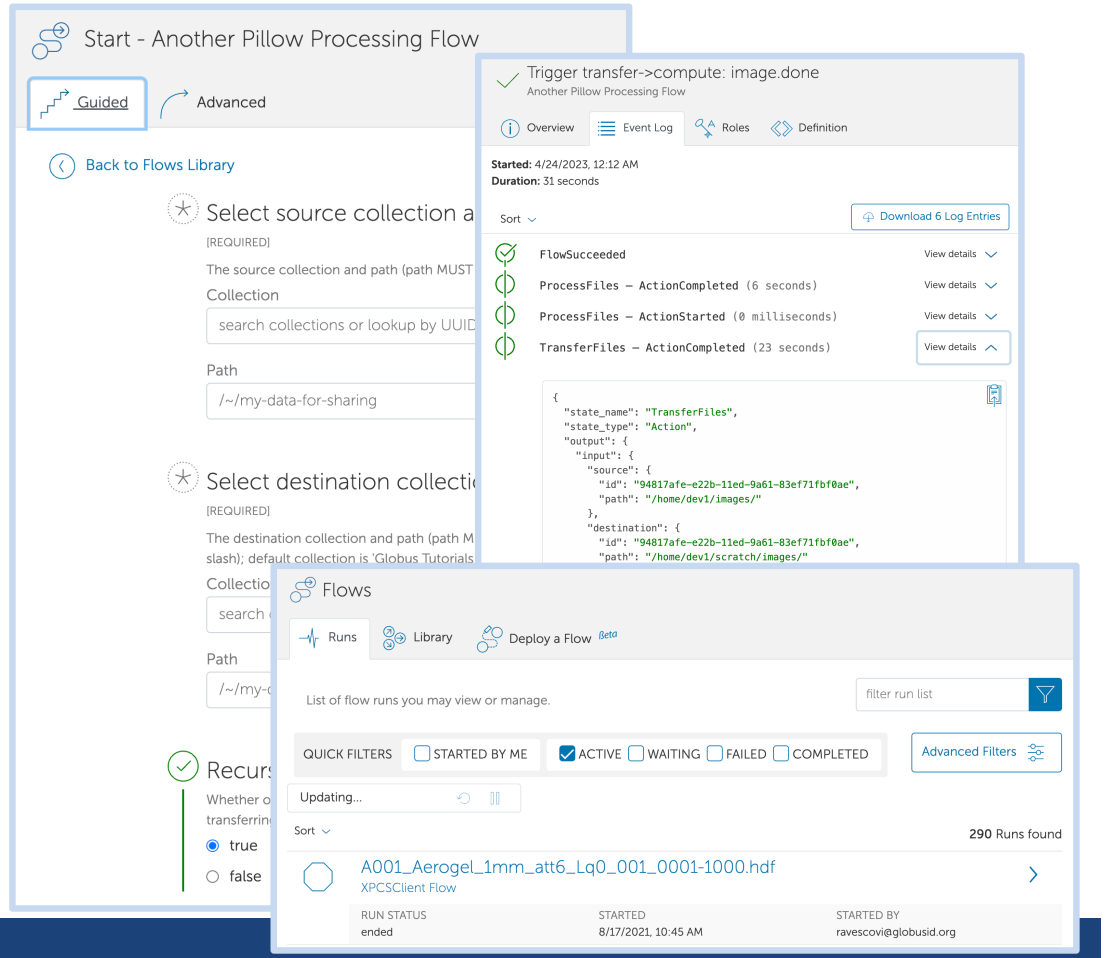
Flow lifecycle

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



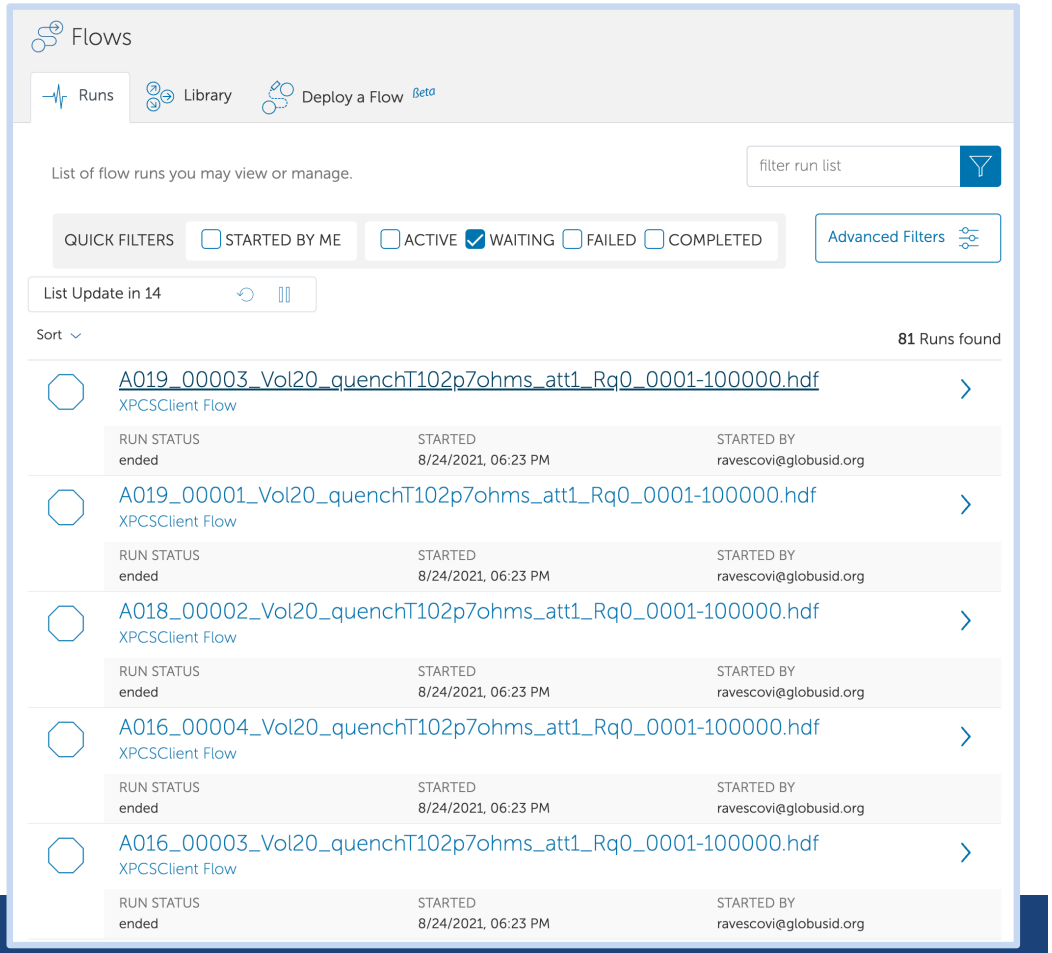
Flow lifecycle

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



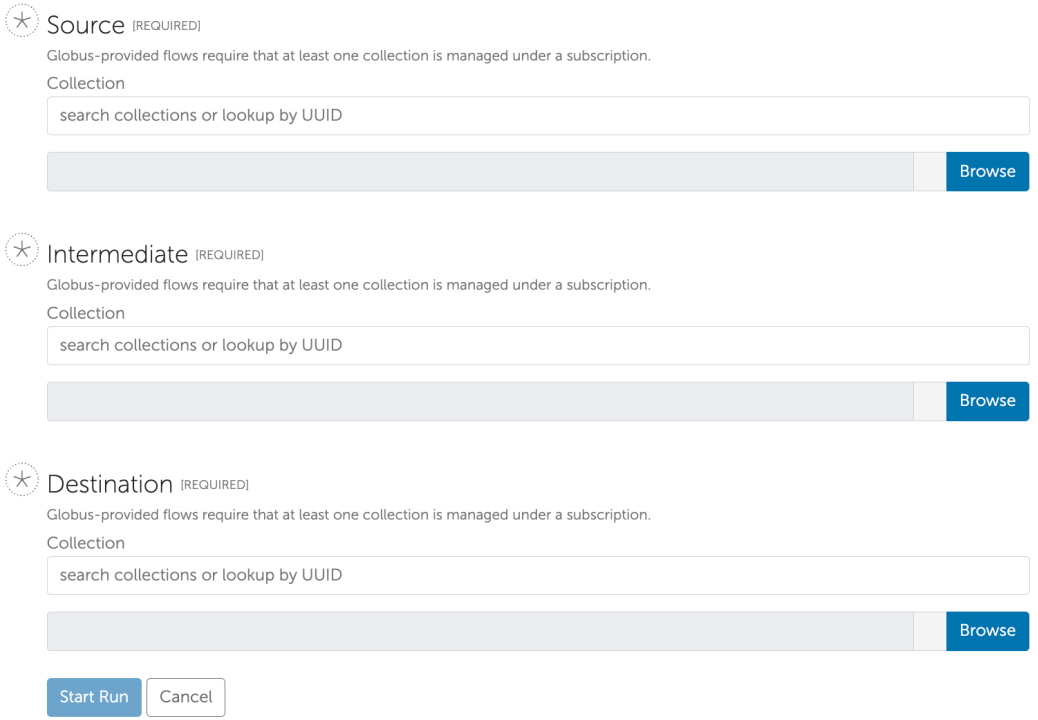
Flow lifecycle: Write once, run many

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



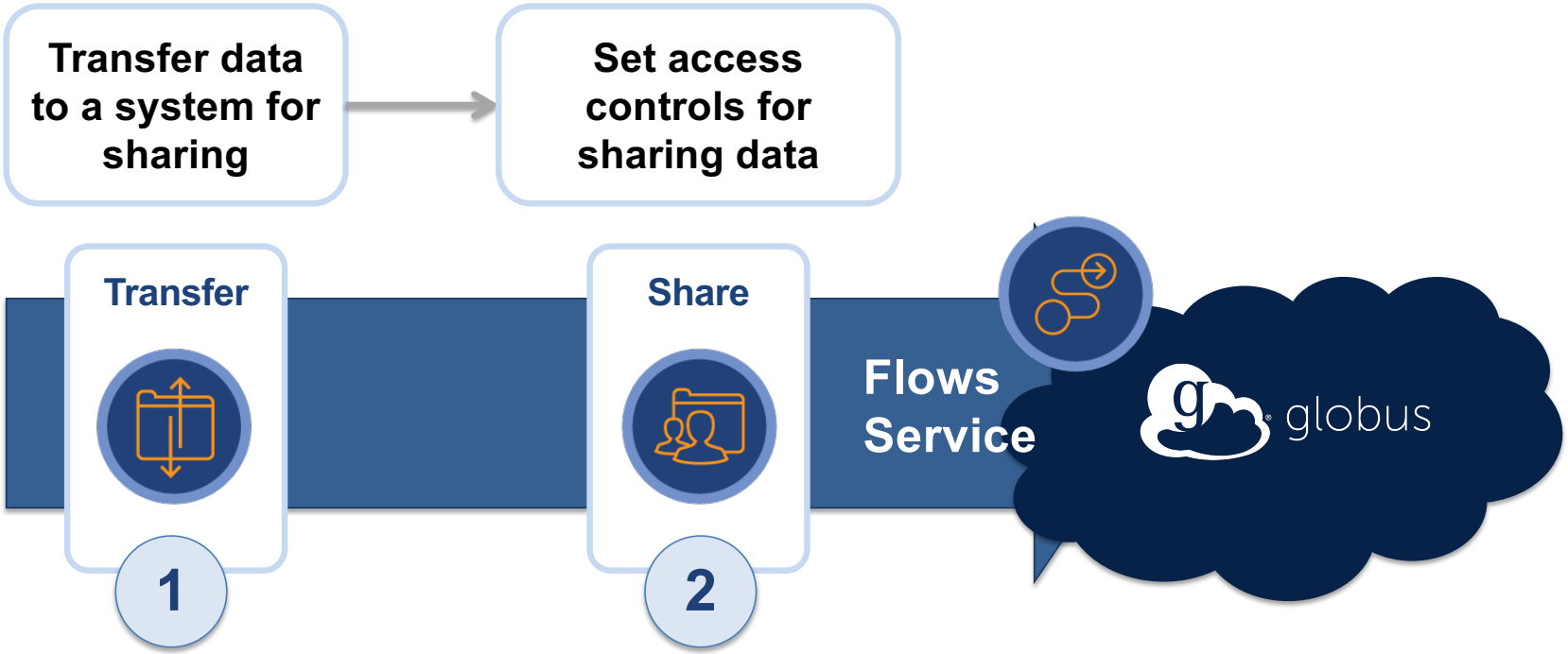
Run a flow

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

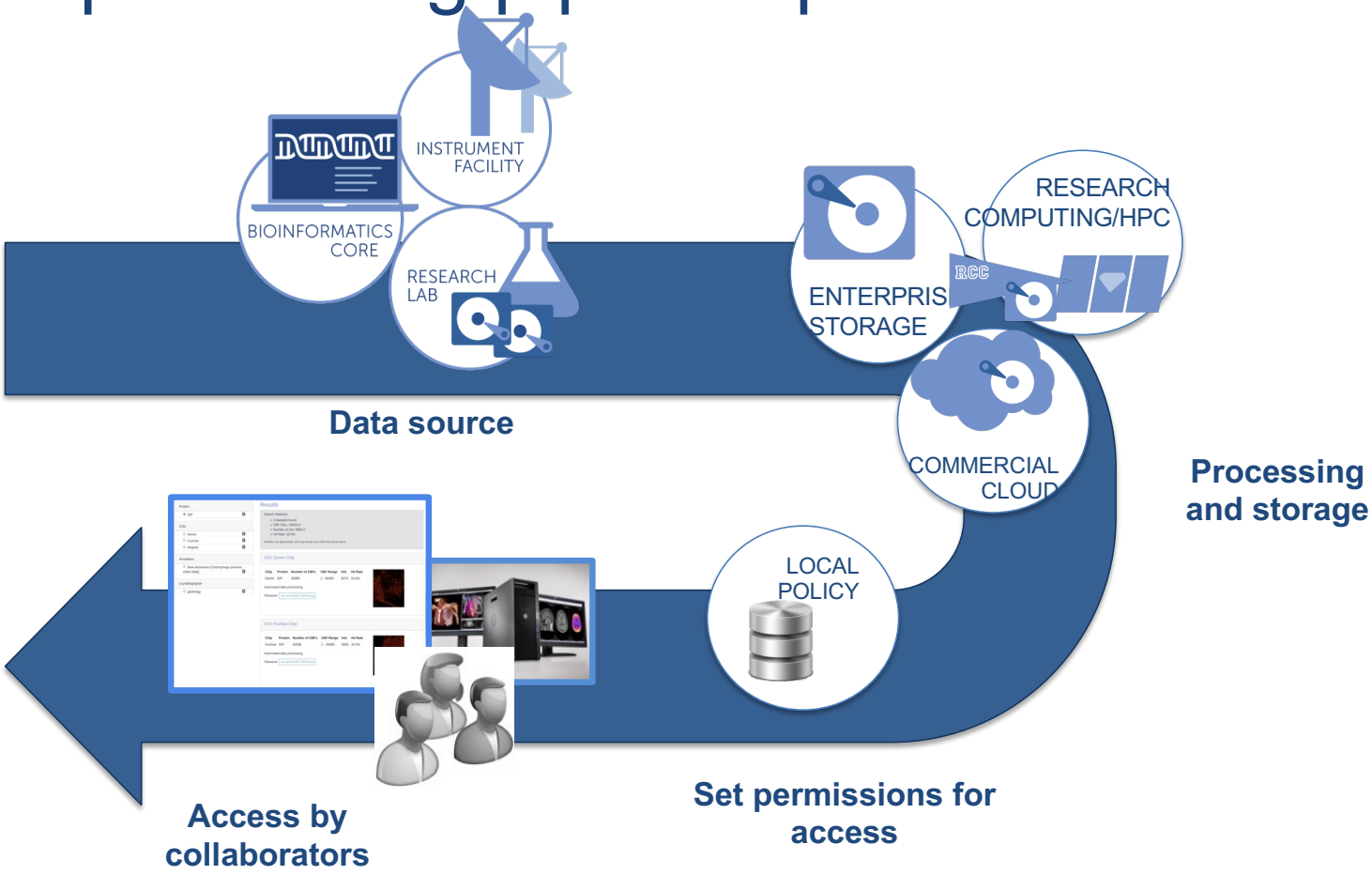


Automation by triggering a run when an event occurs

Transfer and share



Data processing pipeline pattern



SSX Automation

Globus Flows

Compute
Launch QA job

Carbon!
Check threshold

Transfer
Transfer raw files

Compute
Analyze images

Image processing

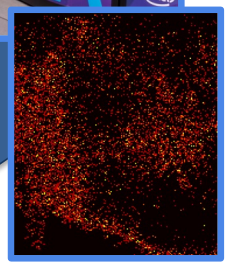
Search
Ingest to index

Share
Set access controls

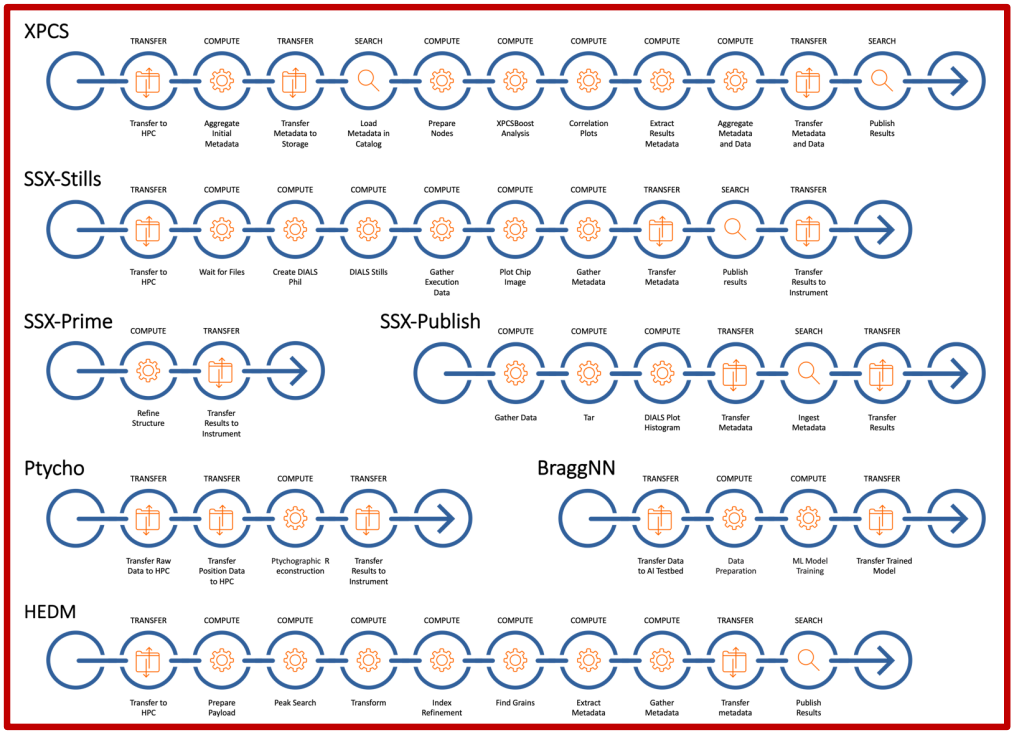
Transfer
Move results to repo

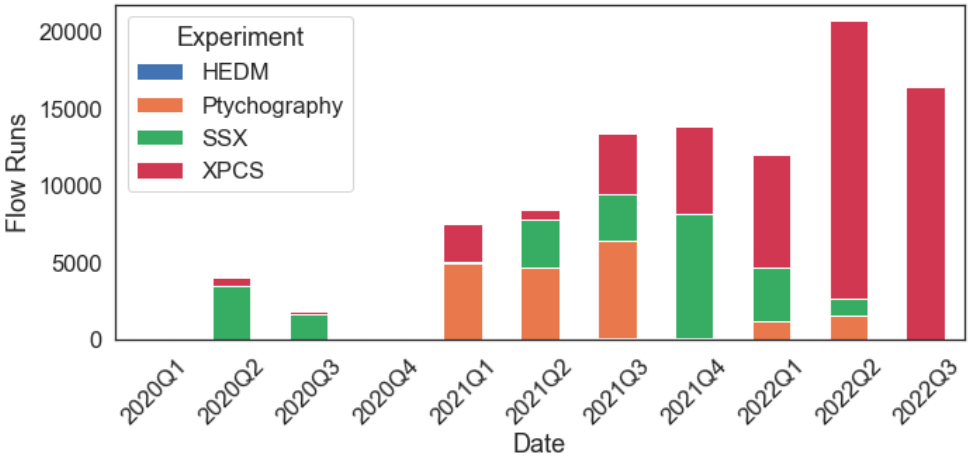
Compute
Gather metadata

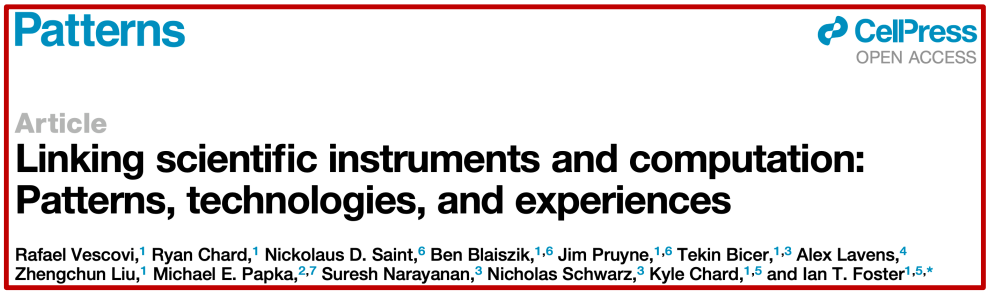
Compute
Visualize



Globus Flows at Advanced Photon Source







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**