cni

Coalition for Networked Information

globus

Scaling Instrument Science in the FAIR Age

Vas Vasiliadis – University of Chicago, Globus, ANL
vas@uchicago.edu

March 25, 2024

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

Globus is …

a non-profit service for research data management and advanced computing, developed and operated by

THE UNIVERSITY OF CHICAGO

THE UNIVERSITY OF CHICAGO | RESEARCH

Our mission is to…

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

Development is partly funded by...

U.S. DEPARTMENT OF
ENERGY

NSF

THE UNIVERSITY OF
CHICAGO

NIH

powered by
amazon
web services

ALFRED P. SLOAN FOUNDATION
1934

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

Argonne
NATIONAL LABORATORY

The last time we were together …Dec. 2018!

Protected Data Support

Working with PHI, PII, or CUI?
Need to manage HIPAA-regulated data?
Globus has you covered!

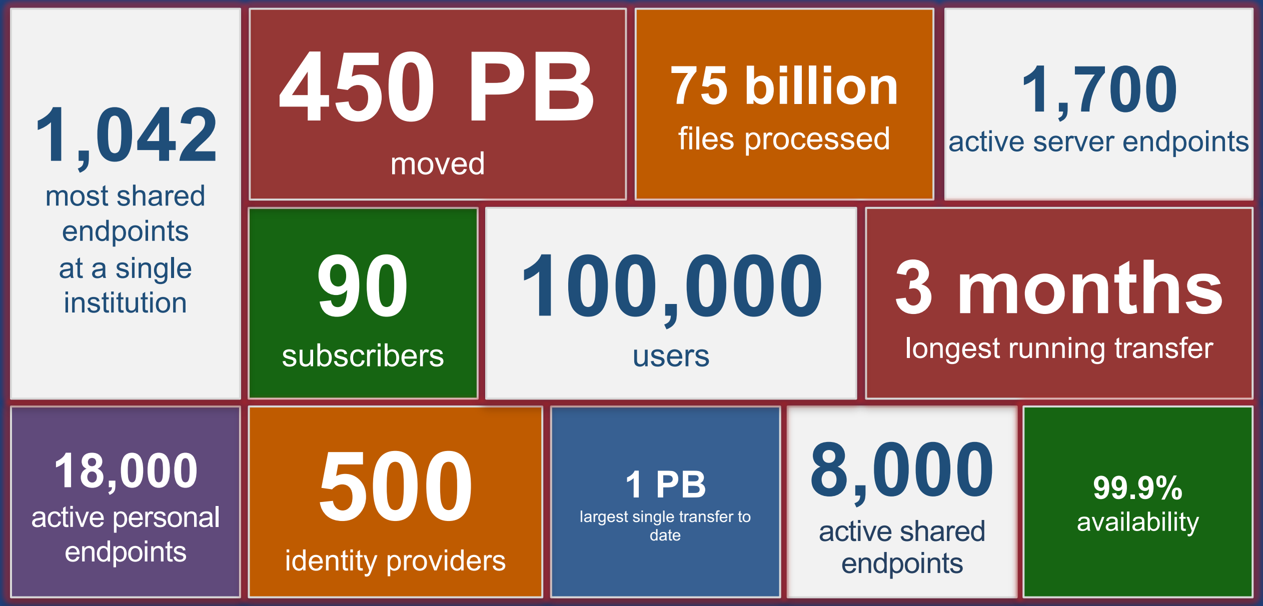
LEARN MORE 〉 REQUEST PRICING 〉

Research data management simplified.

TRANSFER SHARE PUBLISH BUILD

Our story back then…

- Oct. 1998: Globus Toolkit v1 released—The Grid
- Nov. 2010: Globus Online released—The Cloud
- Nov. 2013: Freemium model launched
- Jan. 2019 - 100th subscriber signed, >50% sustainable
- ??? – Globus becomes fully self-sustaining



1,042 most shared endpoints at a single institution

450 PB moved

75 billion files processed

1,700 active server endpoints

90 subscribers

100,000 users

3 months longest running transfer

18,000 active personal endpoints

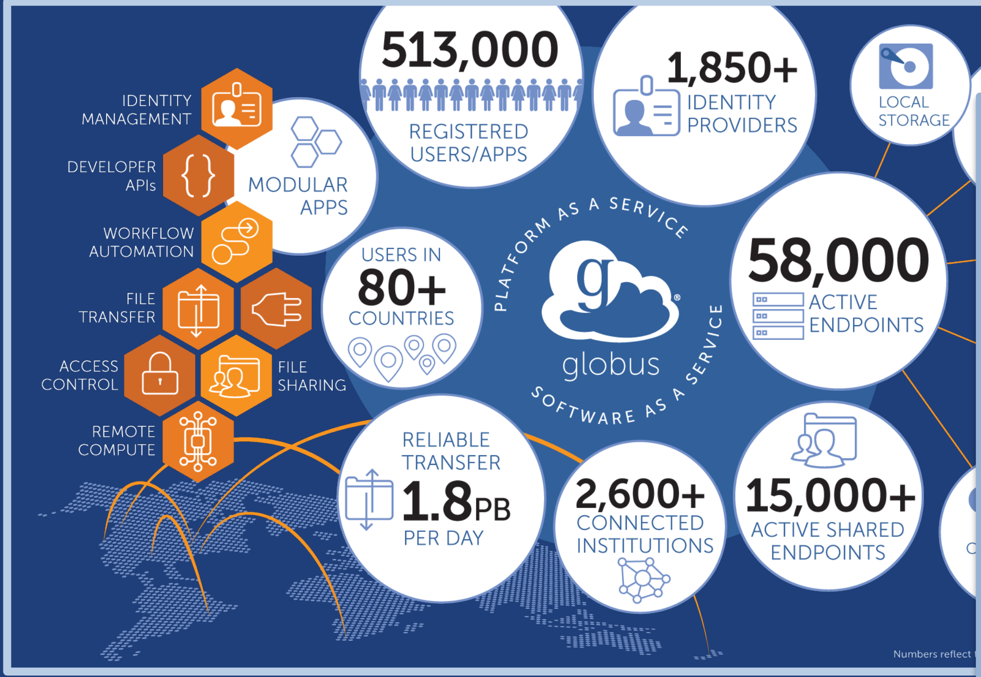
500 identity providers

1 PB largest single transfer to date

8,000 active shared endpoints

99.9% availability

What a difference 5 years make!



NERSC
PURDUE UNIVERSITY
NCAR
Georgia Tech
USGS science for a changing world
Northeastern University
UCLA
UNIVERSITY OF MICHIGAN
JOHNS HOPKINS UNIVERSITY
Imperial College London
lrz
NOAA
Yale
SIMONS FOUNDATION
RIT
NSO
The Jackson Laboratory
FRED HUTCH CURES START HERE
NIH
Stanford University
Los Alamos NATIONAL LABORATORY
Duke UNIVERSITY
CGIAR
NeSI New Zealand eScience Infrastructure
ICM
INDIANA UNIVERSITY
wellcome trust sanger institute
UCSF
ESnet ENERGY SCIENCES NETWORK
KISTI www.kisti.re.kr
NEW YORK GENOME CENTER
NICIS NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM CHPC
syngenta
PRINCETON UNIVERSITY
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
aarnet Australia's Academic and Research Network
PennState
NCSA
UF UNIVERSITY of FLORIDA
KU LEUVEN
OAK RIDGE National Laboratory
CORNELL UNIVERSITY
MICHIGAN STATE UNIVERSITY
RICE
UCONN
Tufts UNIVERSITY
La Jolla Institute FOR IMMUNOLOGY
Dartmouth
University of Pittsburgh
UNIVERSITY OF DELAWARE
THE UNIVERSITY OF TENNESSEE KNOXVILLE
Washington University in St. Louis
THE UNIVERSITY OF ALABAMA
UNIVERSITÉ LAVAL
ATM
NIST
Queen Mary University of London
UNIVERSITY OF NOTRE DAME
UNIVERSITY OF BIRMINGHAM
THE ROCKEFELLER UNIVERSITY

- Network effects driving growth
- Asymptotically approaching sustainability :-)

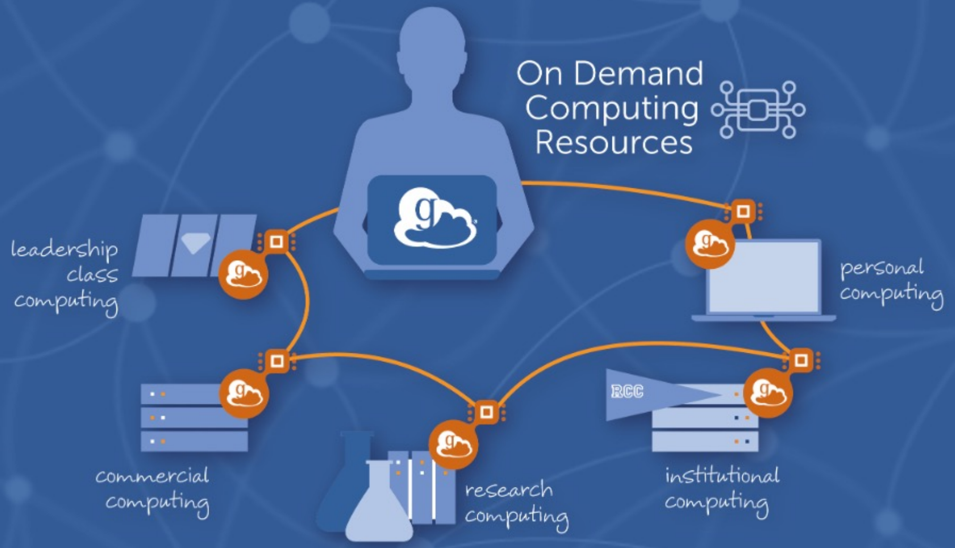
globus

a uchicago non-profit service

GET STARTED

LOG IN

Solutions ▾ Resources ▾ Pricing ▾ Newsroom ▾ Developers ▾ About ▾



Go beyond data

Globus Compute

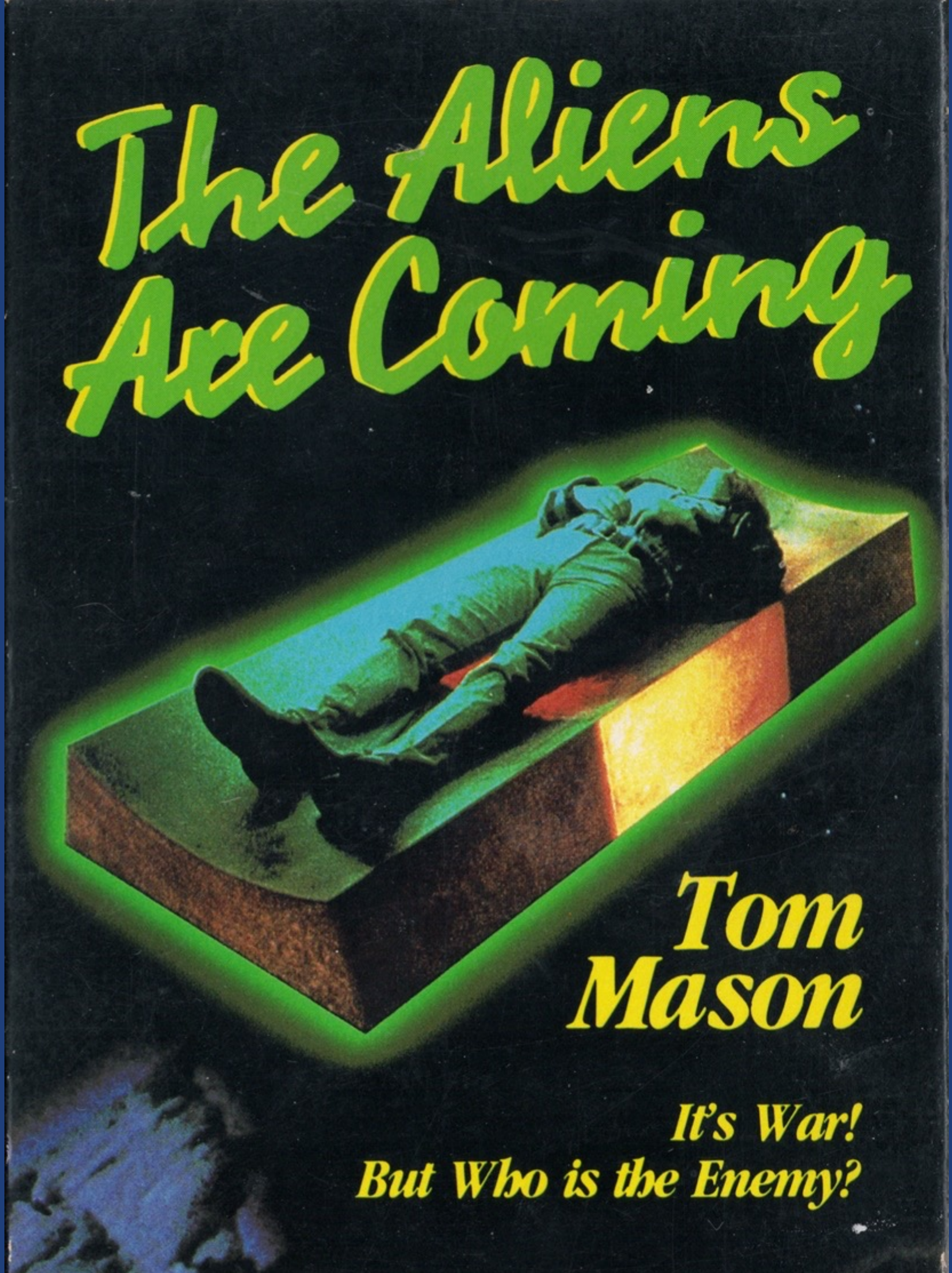
Reliable, distributed Function-as-a-Service

COMPUTE ANYWHERE: EDGE TO SUPERCOMPUTER ›

Research IT. Reimagined.

What’s driving the need to reimagine research IT?

Circa 1980…



Circa 2020: The Instruments are Coming!

NEW YORK
GENOME CENTER®

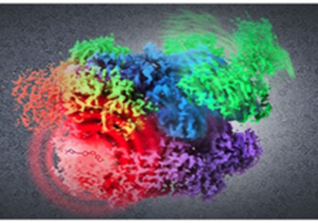
ADVANCED LIGHT SOURCE

NISC
NIH Intramural Sequencing Center

PURDUE
SCIENCE

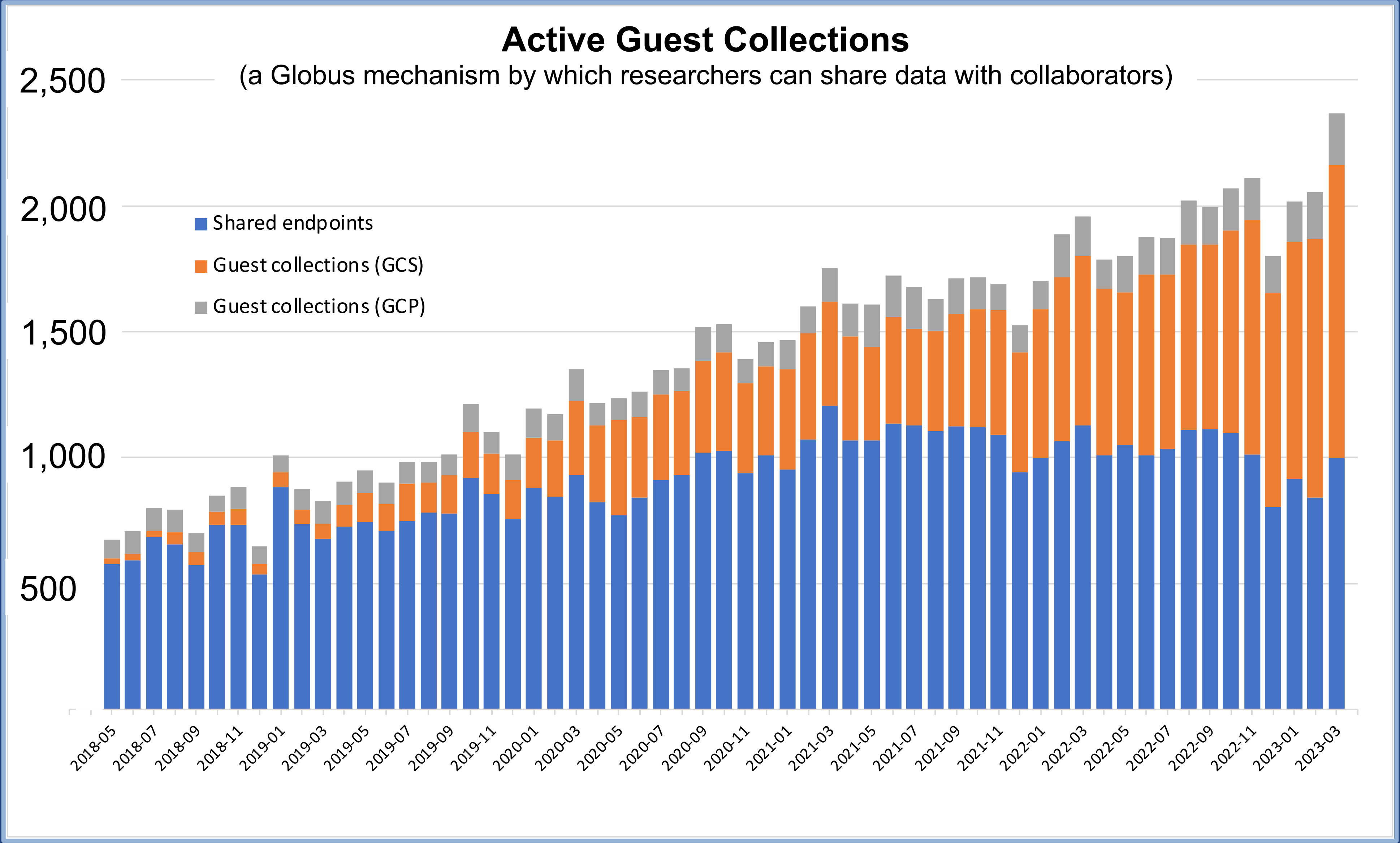
Purdue Cryo-EM Facility

High Resolution
Cryo-EM



National Cryo-Electron Microscopy Facility

More (and larger) collaborations

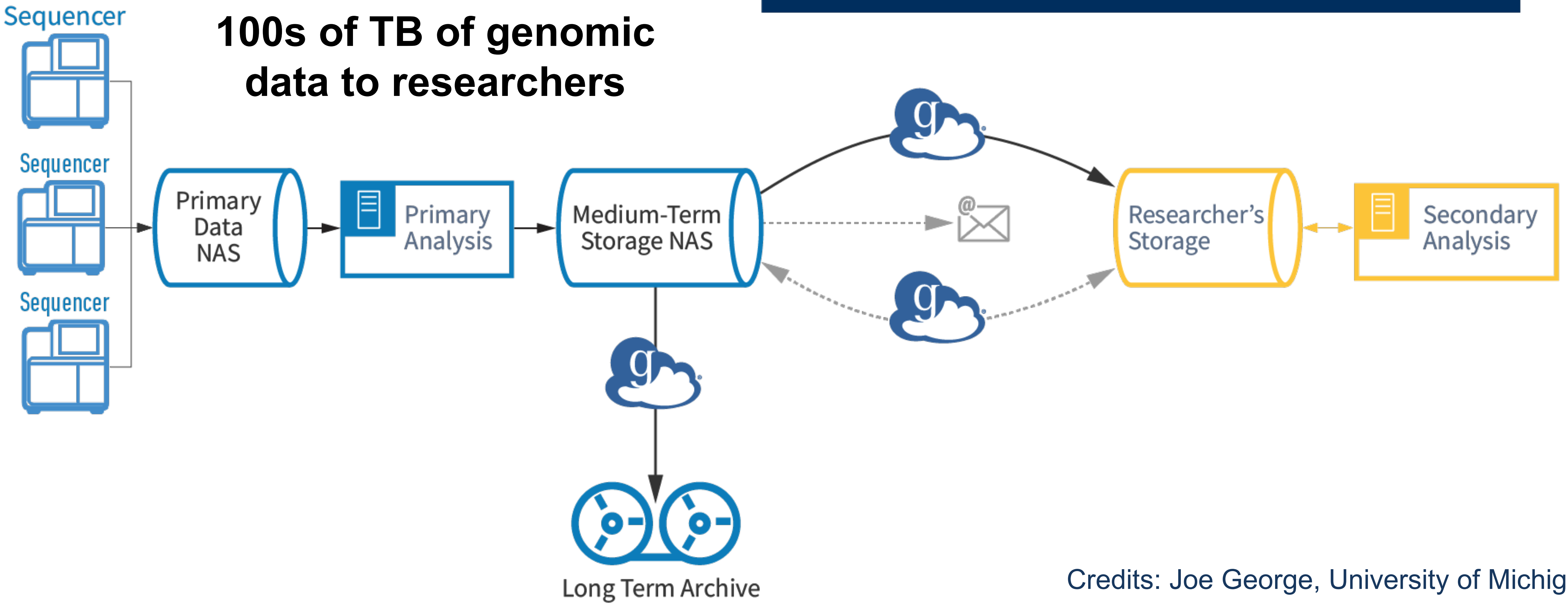


Dramatic changes in sensor data rates and broader collaborations demand **large-scale automation, from capture through publication**

Managing Instrument Data Delivery

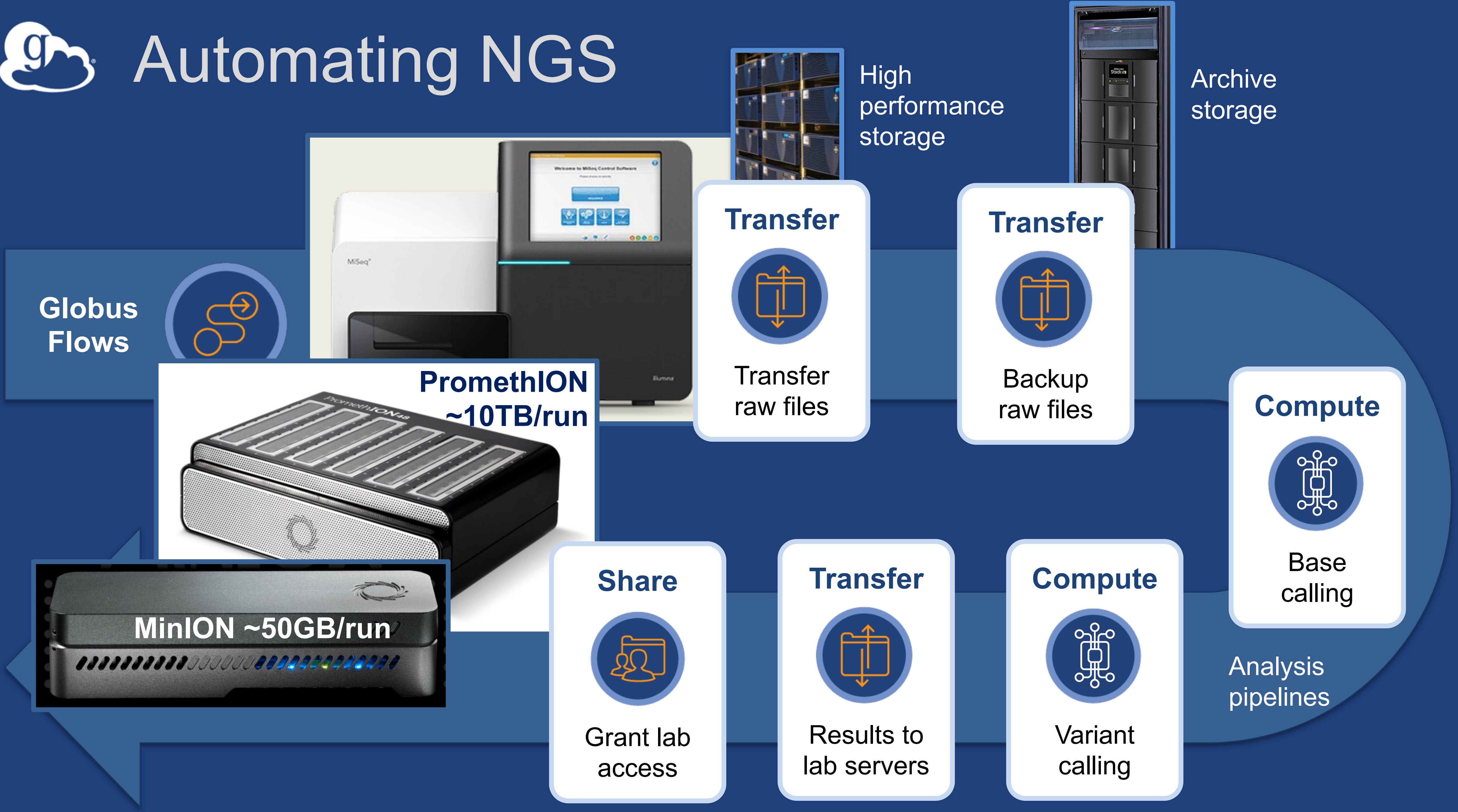
BIOMEDICAL RESEARCH CORE FACILITIES
ADVANCED GENOMICS CORE
UNIVERSITY OF MICHIGAN

Use Globus to deliver 100s of TB of genomic data to researchers



Credits: Joe George, University of Michigan

Automating NGS



Globus Flows

PromethION ~10TB/run

MinION ~50GB/run

High performance storage

Archive storage

Transfer

Transfer raw files

Transfer

Backup raw files

Compute

Base calling

Analysis pipelines

Compute

Variant calling

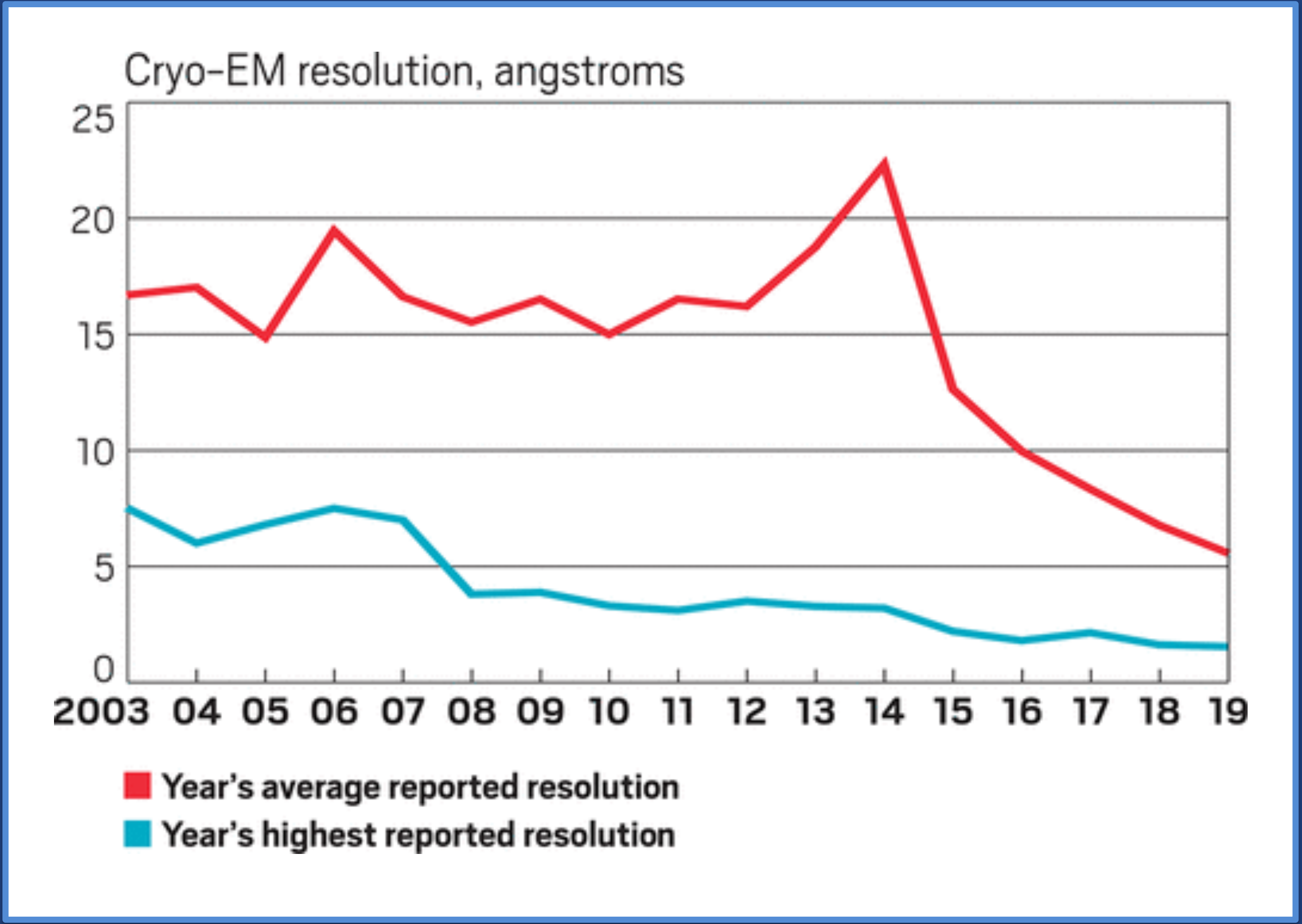
Transfer

Results to lab servers

Share

Grant lab access

Cryo-EM will eat (all) your resources



Automating cryo-EM flows

Globus Flows

Transfer

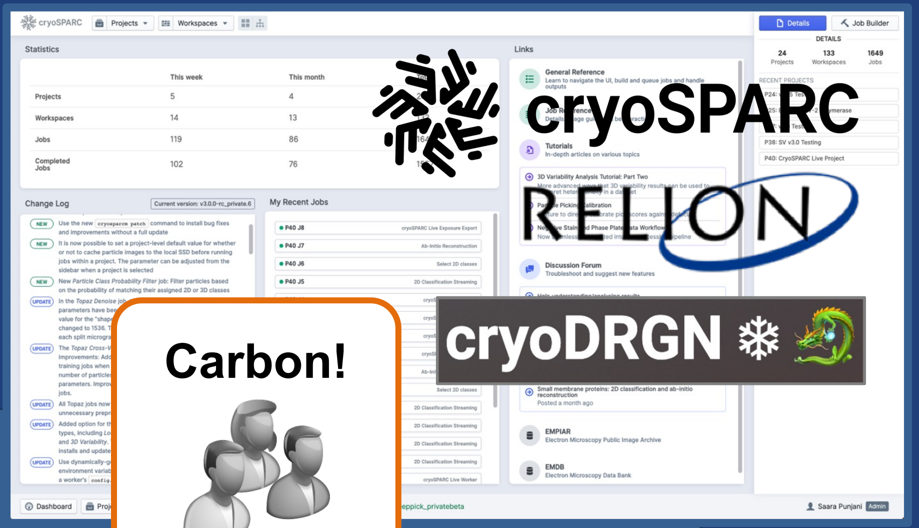
Transfer raw files

Compute

Launch analysis job

Carbon!

Correct, classify, …



cryoSPARC | Projects ▾ | Workspaces ▾

Statistics

	This week	This month
Projects	5	4
Workspaces	14	13
Jobs	119	86
Completed Jobs	102	76

Change Log — Current version: v3.0.0-rc_private.6

- NEW Use the new `cryosparcm patch` command to install bug fixes and improvements without a full update
- NEW It is now possible to set a project-level default value for whether or not to cache particle images to the local SSD before running jobs within a project. The parameter can be adjusted from the sidebar when a project is selected
- NEW New *Particle Class Probability Filter* job: Filter particles based on the probability of matching their assigned 2D or 3D classes
- UPDATE In the *Topaz Denoise* job … parameters have bee… value for the "shape… changed to 1536. T… each split microgra…
- UPDATE The *Topaz Cross-V*… improvements: Add… training jobs when… number of particles… parameters. Improv… jobs.
- UPDATE All Topaz jobs now… unnecessary prepr…
- UPDATE Added option for th… types, including *Lo*… and *3D Variability*. … installs and update…
- UPDATE Use dynamically-g… environment variab… a worker's `config.`…

My Recent Jobs

- P40 J8 — cryoSPARC Live Exposure Export
- P40 J7 — Ab-Initio Reconstruction
- P40 J6 — Select 2D classes
- P40 J5 — 2D Classification Streaming
- … — Select 2D classes
- … — 2D Classification Streaming
- … — 2D Classification Streaming
- … — 2D Classification Streaming
- … — 2D Classification Streaming
- … — cryoSPARC Live Worker

Links

- General Reference — Learn to navigate the UI, build and queue jobs and handle outputs
- Job Reference — Details…
- Tutorials — In-depth articles on various topics
- 3D Variability Analysis Tutorial: Part Two — More advanced ways that 3D variability results can be used to interpret heterogeneity in a dataset
- Particle Picking Calibration
- Negative Stain and Phase Plate Data Workflow — Now seamlessly … pipeline
- Discussion Forum — Troubleshoot and suggest new features
- Small membrane proteins: 2D classification and ab-initio reconstruction — Posted a month ago
- EMPIAR — Electron Microscopy Public Image Archive
- EMDB — Electron Microscopy Data Bank

Details | Job Builder

DETAILS

24	133	1649
Projects	Workspaces	Jobs

RECENT PROJECTS

- P24: …
- P25: …merase
- …: … Testing
- P38: SV v3.0 Testing
- P40: CryoSPARC Live Project

Dashboard | Proj… | …eppick_privatebeta | Saara Punjani Admin

Compute

Extract metadata

Search

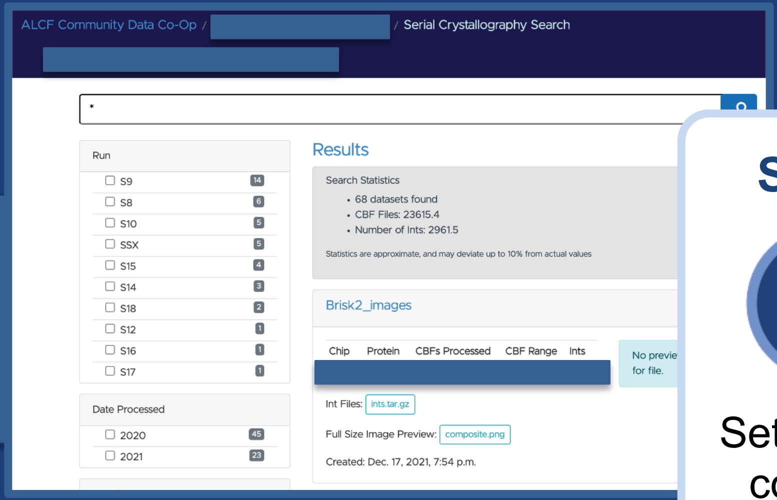
Ingest to index

Transfer

Move final files to repo

Share

Set access controls



ALCF Community Data Co-Op / [redacted] / Serial Crystallography Search

[redacted]

*

Run

- □ S9 14
- □ S8 6
- □ S10 5
- □ SSX 5
- □ S15 4
- □ S14 3
- □ S18 2
- □ S12 1
- □ S16 1
- □ S17 1

Date Processed

- □ 2020 45
- □ 2021 23

Results

Search Statistics

- 68 datasets found
- CBF Files: 23615.4
- Number of Ints: 2961.5

Statistics are approximate, and may deviate up to 10% from actual values

Brisk2_images

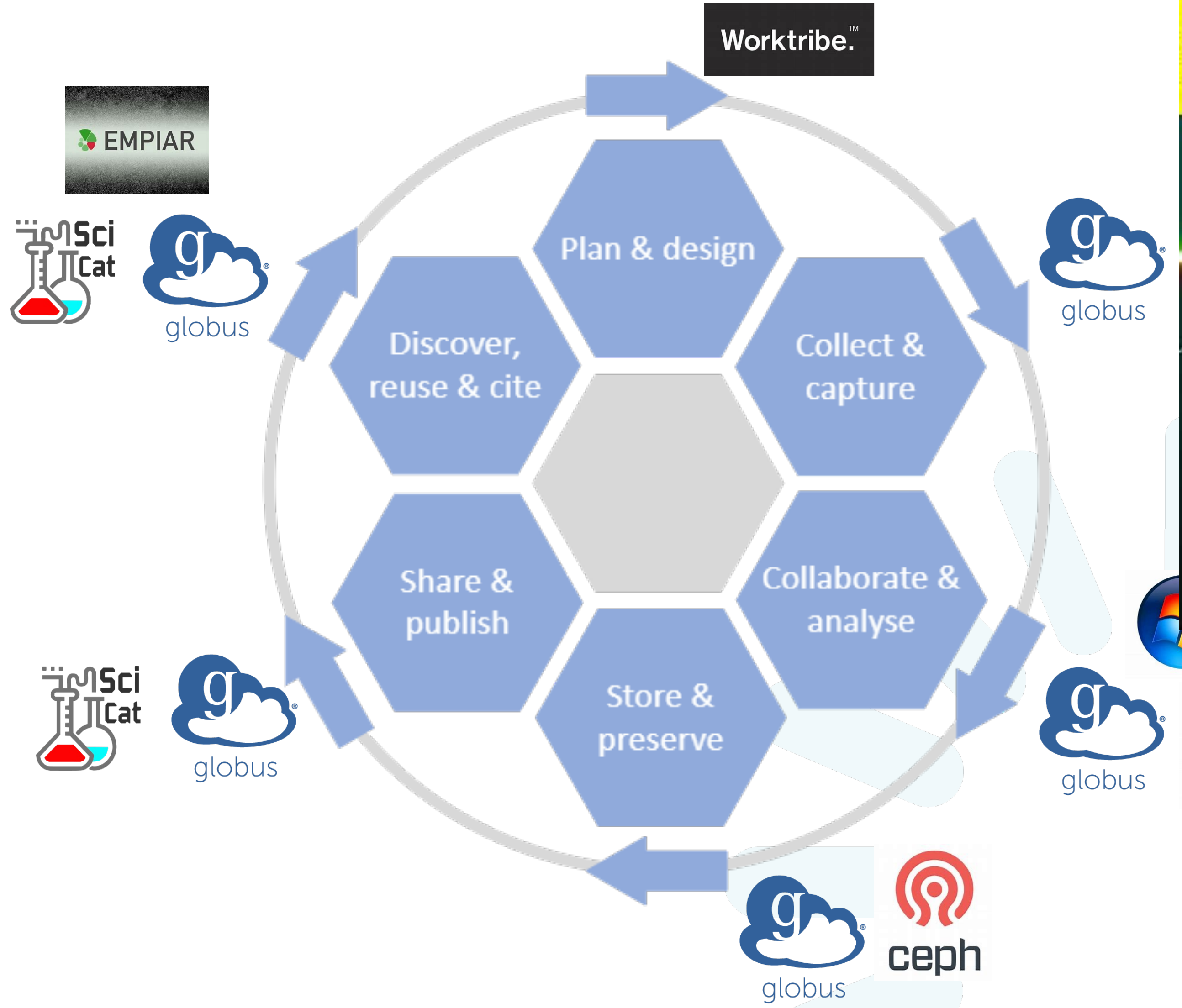
Chip	Protein	CBFs Processed	CBF Range	Ints
[redacted]				

No previe… for file.

Int Files: ints.tar.gz

Full Size Image Preview: composite.png

Created: Dec. 17, 2021, 7:54 p.m.



Plan & design

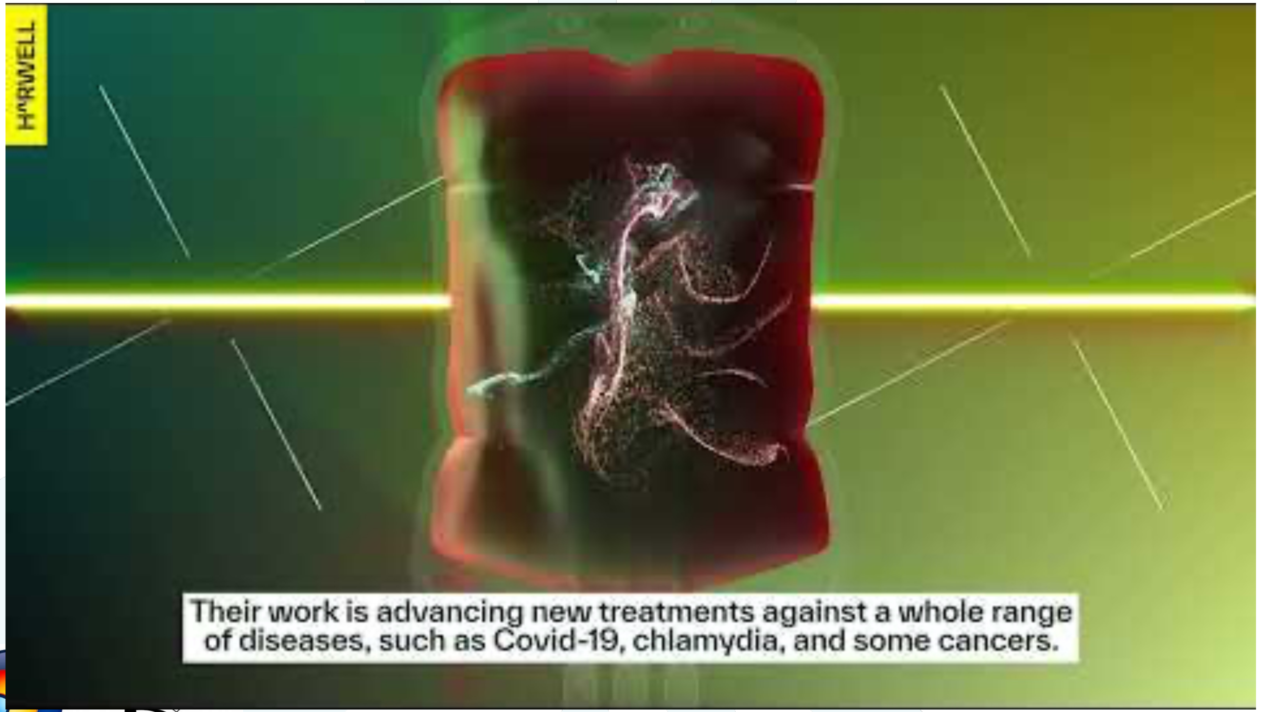
Collect & capture

Collaborate & analyse

Store & preserve

Share & publish

Discover, reuse & cite



Their work is advancing new treatments against a whole range of diseases, such as Covid-19, chlamydia, and some cancers.

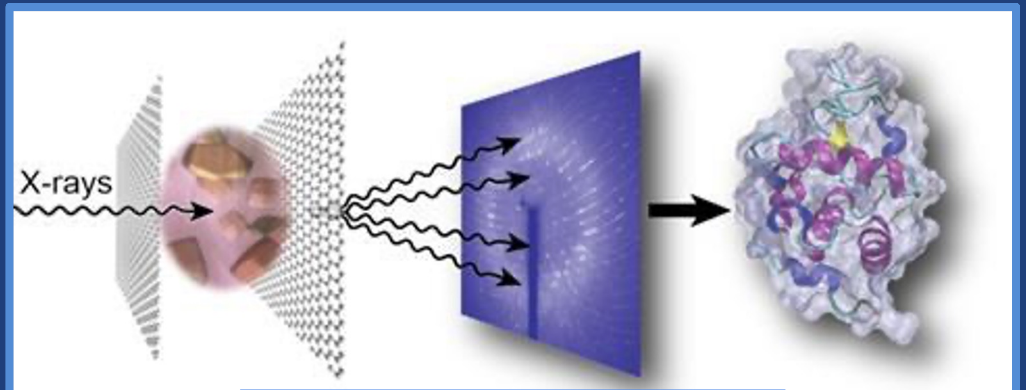
openstack ceph Sci Cat

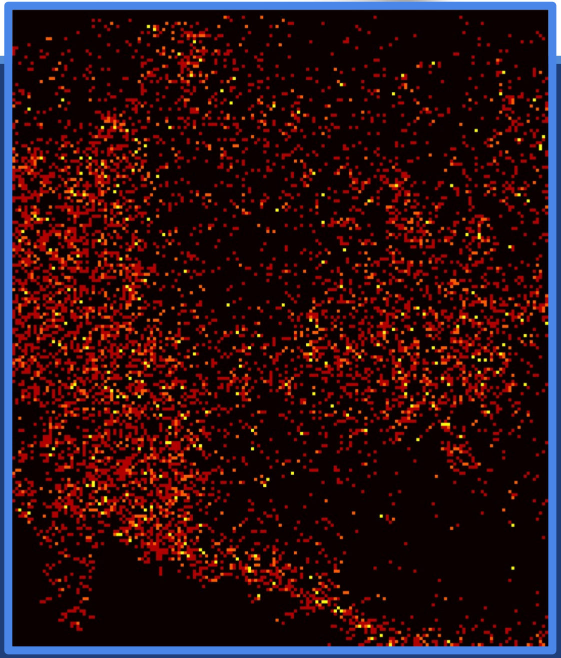
Ramos, S. GlobusWorld 2023
https://www.youtube.com/watch?v=q-o6VKIEPj4

The Rosalind Franklin Institute

Serial crystallography at the Argonne APS

- Serially image chips with thousands 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 crystal structure to scientist





Darren Sherrell, Gyorgy Babnigg, Andrzej Joachimiak

How Globus enables SSX automation at scale

Globus Flows

Data capture

Compute

Launch QA job

Carbon!

Check threshold

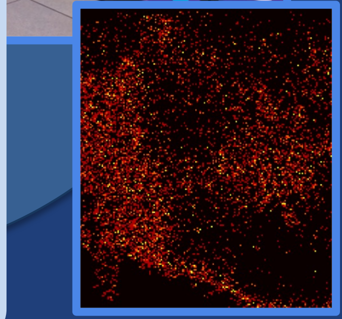
Transfer

Transfer raw files

Compute

Analyze images

Image processing



Compute

Visualize

Compute

Gather metadata

Transfer

Move results to repo

Share

Set access controls

Search

Ingest to index

Data publication

Protein

☑ IDP 3

Chip

□ Denim 1

□ Fuchsia 1

□ Hotpink 1

Annotation

□ Beta-lactamase [Chitinophaga pinensis DSM 2588]. 3

Crystallographer

□ gbabnigg 3

Results

Search Statistics

- 3 datasets found
- CBF
- N
- H

Statistics

SSX D

Chip

Denim

Automat

Filenam

SSX F

Chip

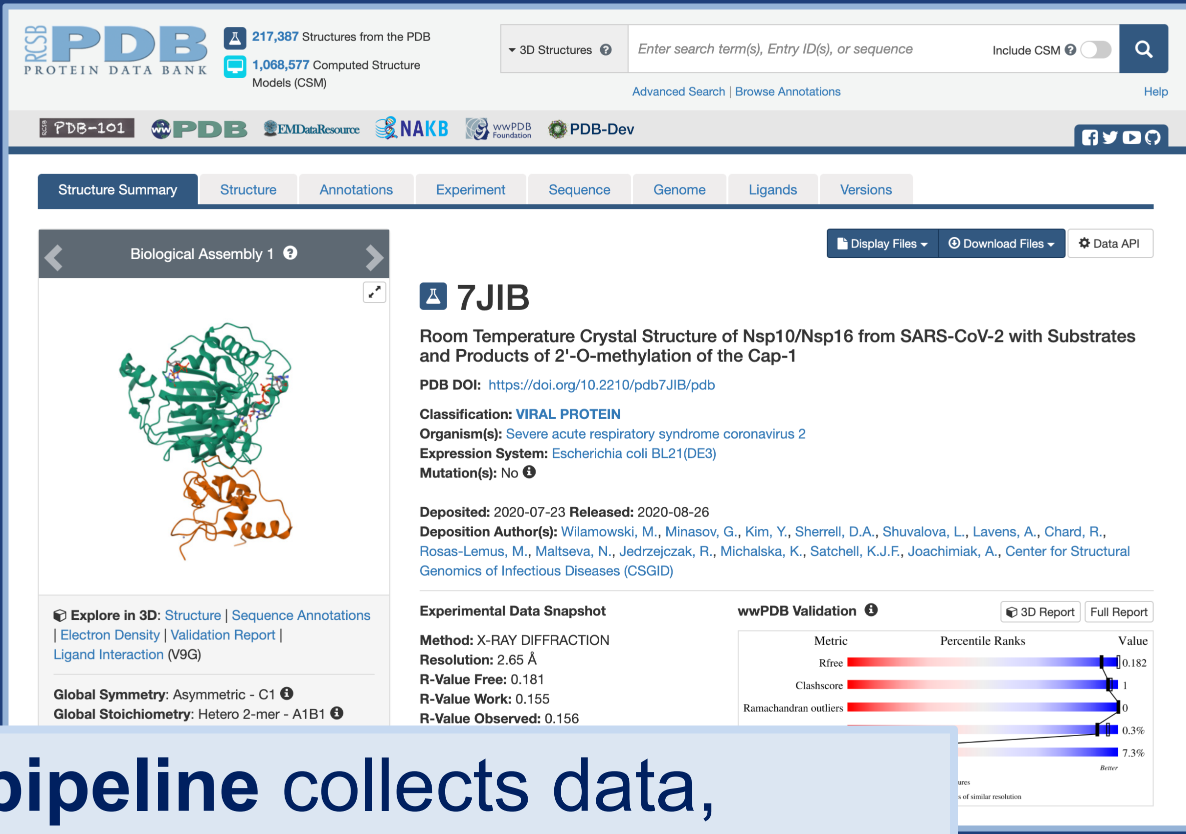
Fuchsia

Automat

Filenam

Outcomes and impact

> “These data services have taken the time to solve a structure from weeks to days and now to hours”
> *Darren Sherrell, SBC beamline scientist APS Sector 19*



RCSB PDB PROTEIN DATA BANK

217,387 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

7JIB

Room Temperature Crystal Structure of Nsp10/Nsp16 from SARS-CoV-2 with Substrates and Products of 2'-O-methylation of the Cap-1

PDB DOI: https://doi.org/10.2210/pdb7JIB/pdb

Classification: VIRAL PROTEIN
Organism(s): Severe acute respiratory syndrome coronavirus 2
Expression System: Escherichia coli BL21(DE3)
Mutation(s): No

Deposited: 2020-07-23 **Released:** 2020-08-26
Deposition Author(s): Wilamowski, M., Minasov, G., Kim, Y., Sherrell, D.A., Shuvalova, L., Lavens, A., Chard, R., Rosas-Lemus, M., Maltseva, N., Jedrzejczak, R., Michalska, K., Satchell, K.J.F., Joachimiak, A., Center for Structural Genomics of Infectious Diseases (CSGID)

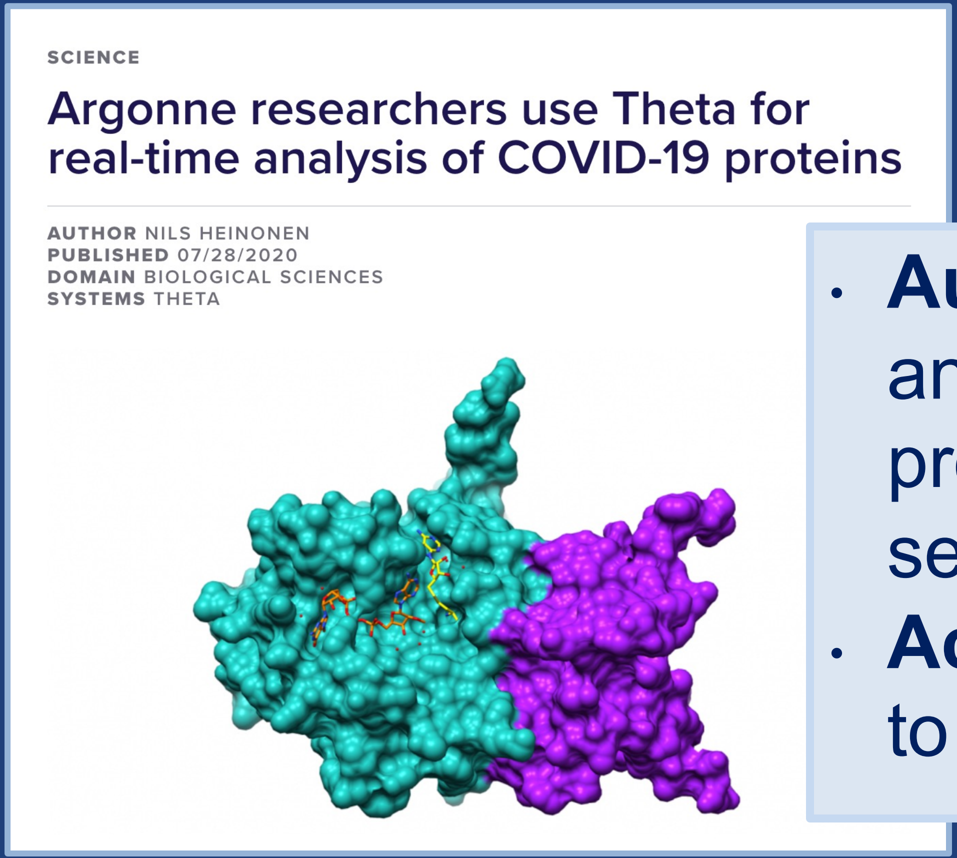
Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (V9G)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Hetero 2-mer - A1B1

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.65 Å
R-Value Free: 0.181
R-Value Work: 0.155
R-Value Observed: 0.156

wwPDB Validation



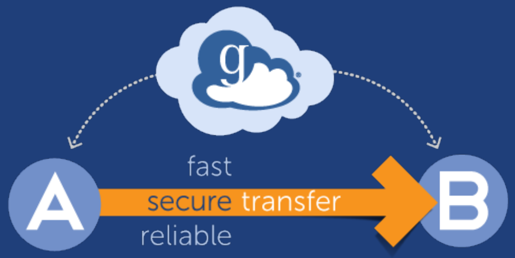
SCIENCE

Argonne researchers use Theta for real-time analysis of COVID-19 proteins

AUTHOR NILS HEINONEN
PUBLISHED 07/28/2020
DOMAIN BIOLOGICAL SCIENCES
SYSTEMS THETA

- **Automation pipeline** collects data, analyzes and visualizes the data, solves protein structure and loads results into a searchable portal for real-time feedback
- **Achieved over 10-100x speed up** in time to solution of protein structures at APS

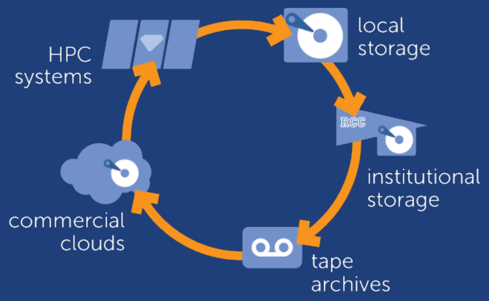
Where does Globus fit in this picture?



Managed transfer & sync

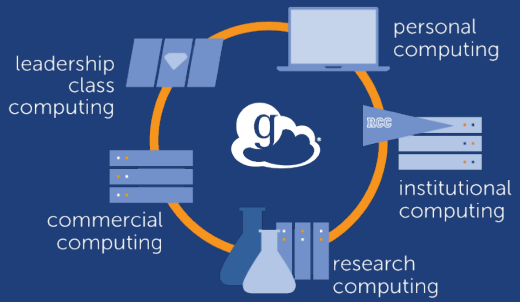
colleagues
collaborators

Collaborative data sharing

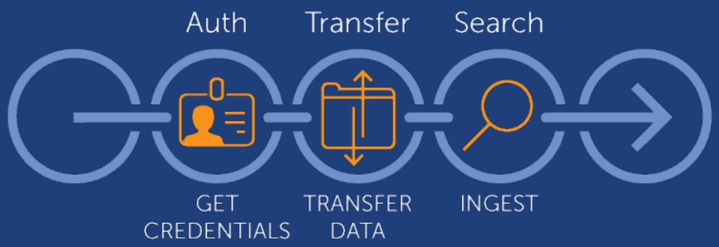


Unified data access

Publication & discovery

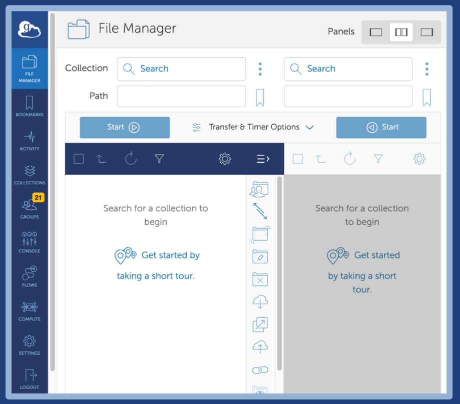


Managed remote execution



Reliable automation

Platform-as-a-Service



Search for a collection to begin

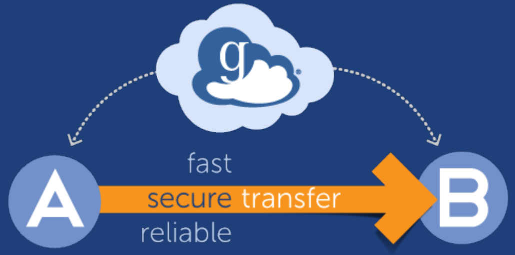
Get started by taking a short tour.

Search for a collection to begin

Get started by taking a short tour.

Software-as-a-Service

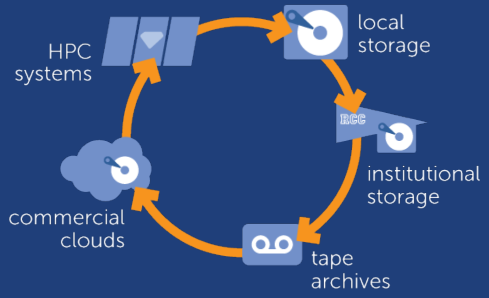
Making data FAIR by default



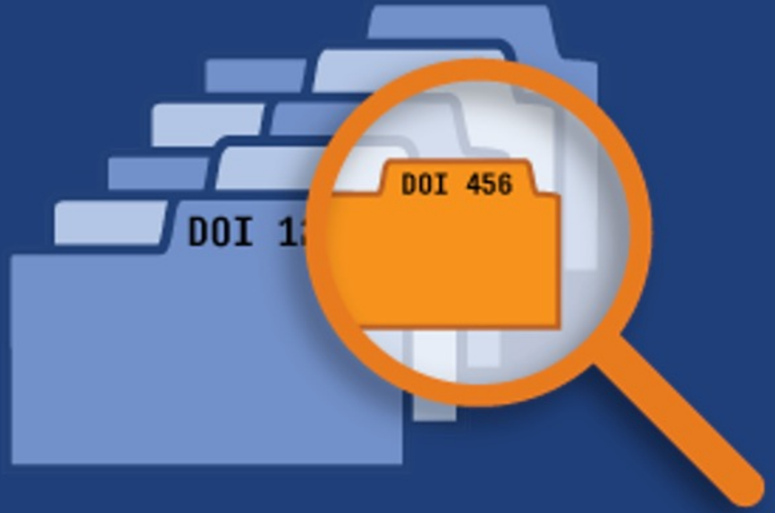
Managed transfer & sync

collaborators
colleagues

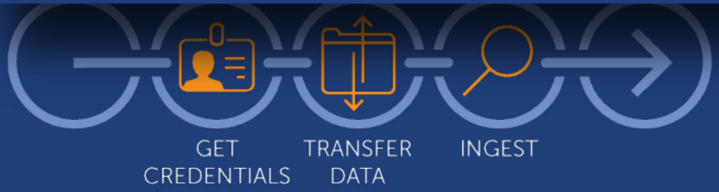
Collaborative data sharing



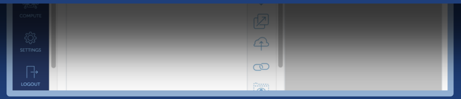
Unified data access



Publication & discovery



Reliable automation



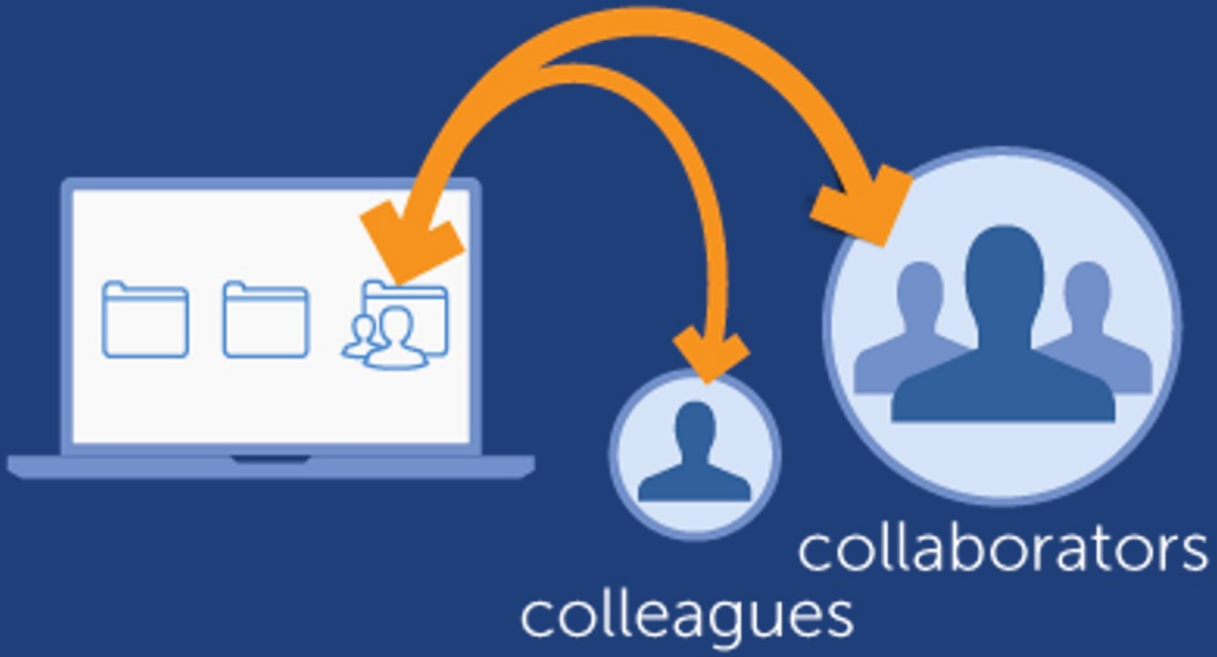
Software-as-a-Service

Globus Search: Data description and discovery

- Metadata store with fine-grained visibility controls
- Schema agnostic
- DOI minting via DataCite
- Simple search using URL queries
- Complex search using filters and facets

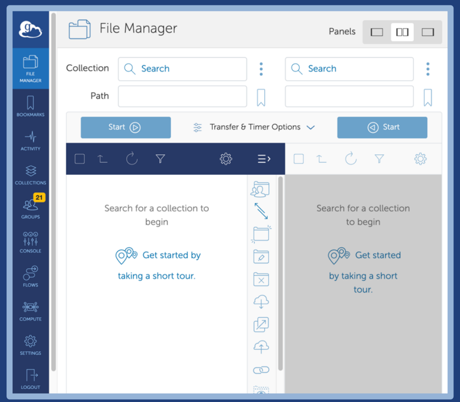
docs.globus.org/api/search

Making data more FAIR by default



Collaborative data sharing

Platform-as-a-Service



File Manager

Panels

Collection

Search

Path

Start

Transfer & Timer Options

Start

Search for a collection to begin

Get started by taking a short tour.

Search for a collection to begin

Get started by taking a short tour.

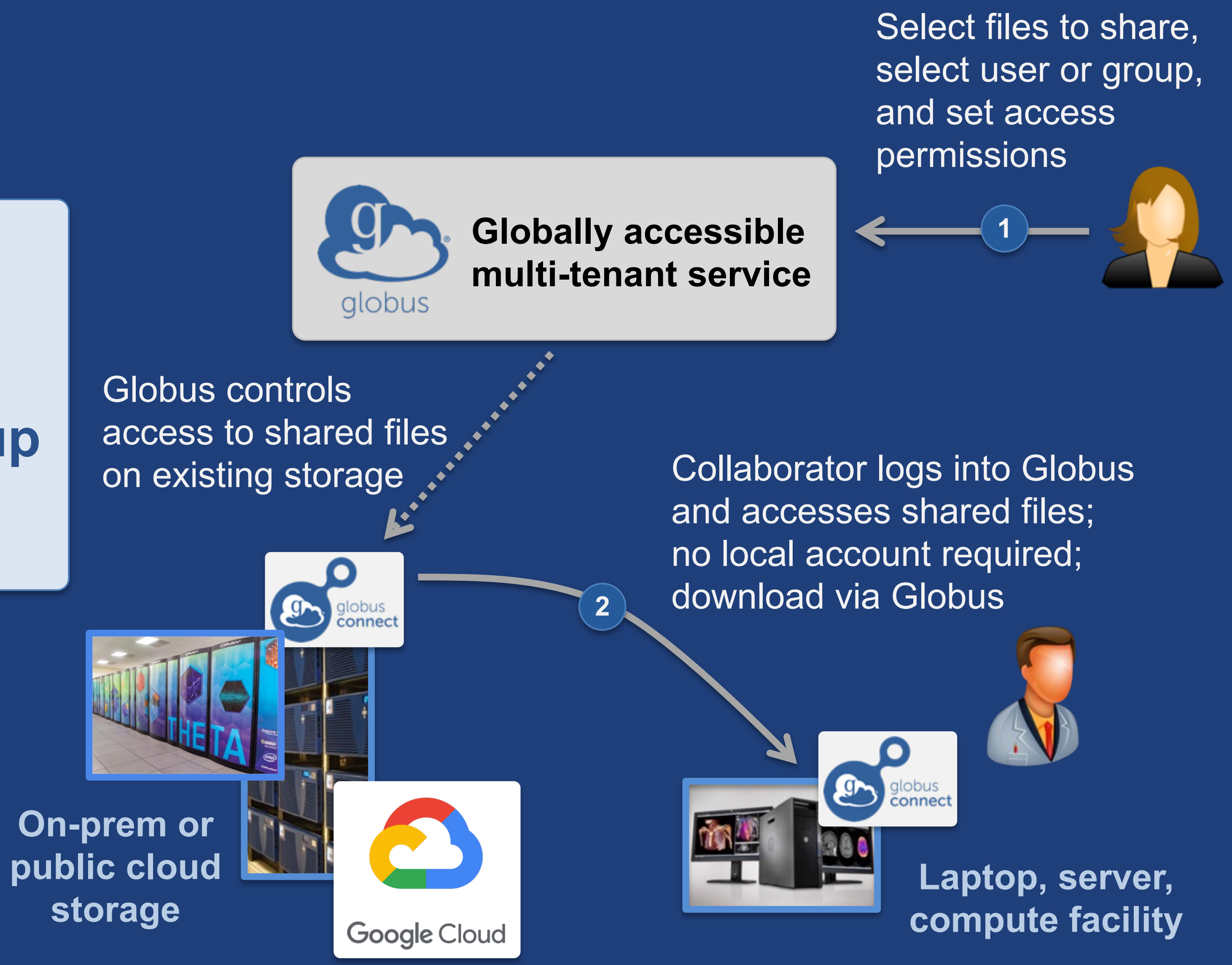
Software-as-a-Service

Unified data access

Reliable automation

Secure data sharing …from any storage

- Fine-grained access control
- Storage system “overlay”
- Share with identity/email/group
- No data staging required



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

Globally accessible multi-tenant service

Globus controls access to shared files on existing storage

Collaborator logs into Globus and accesses shared files; no local account required; download via Globus

On-prem or public cloud storage

Laptop, server, compute facility

Distinct access policies may be applied to

Data *and* Metadata

Support for managing protected data

Security controls

→ NIST 800-53

→ NIST 800-171

Restricted data handling

→ PHI, PII, CUI

→ Compliant data sharing

BAA w/Uchicago

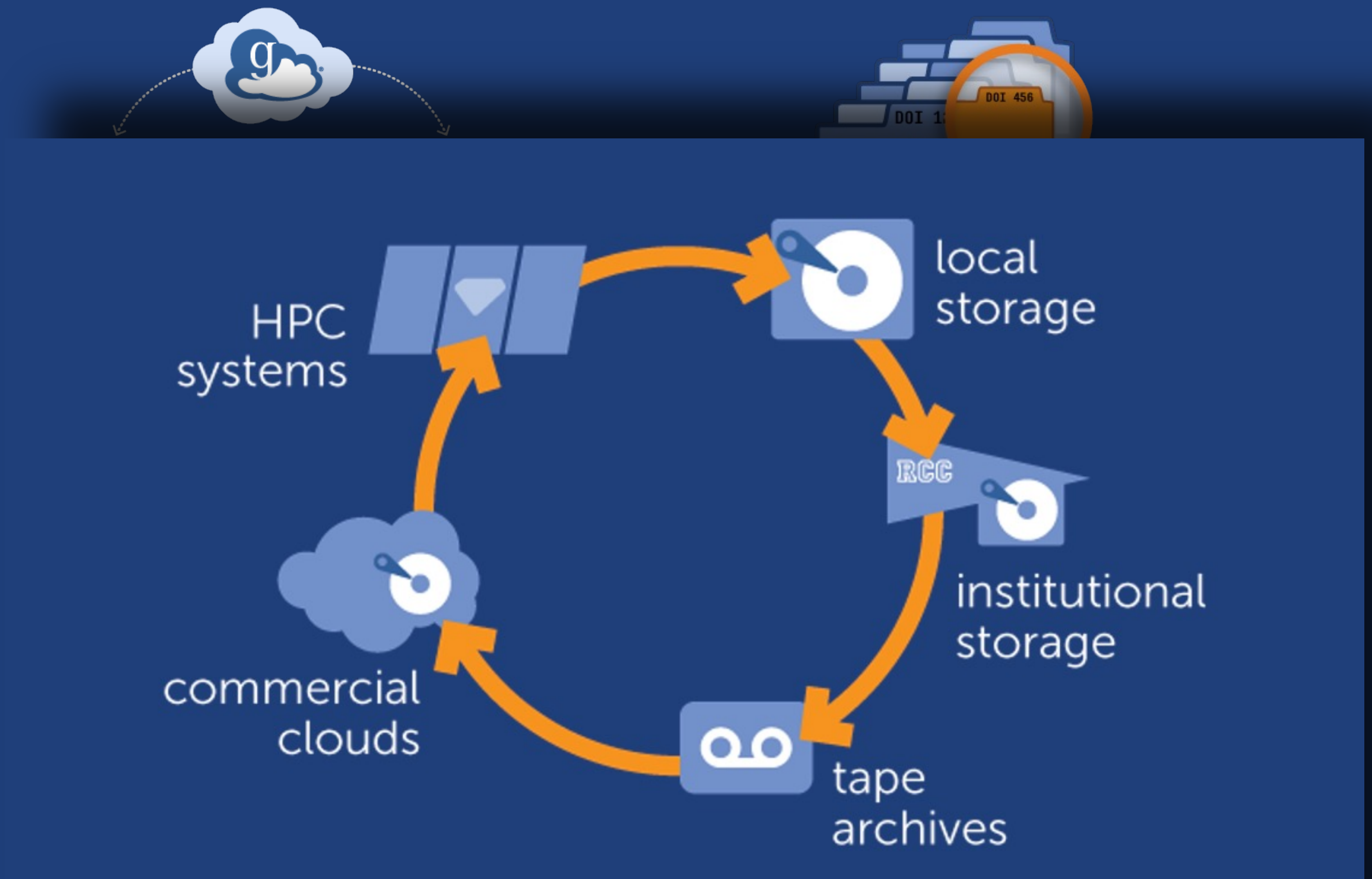
→ UChicago BAA with Amazon

Not a tradeoff you need to make…

Access
Sharing
Collaboration

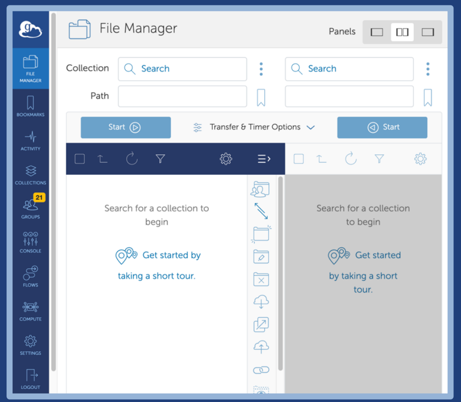
Security
Privacy
Compliance

Making data more FAIR by default



Unified data access

Platform-as-a-Service



Collection: Search

Path:

Search for a collection to begin

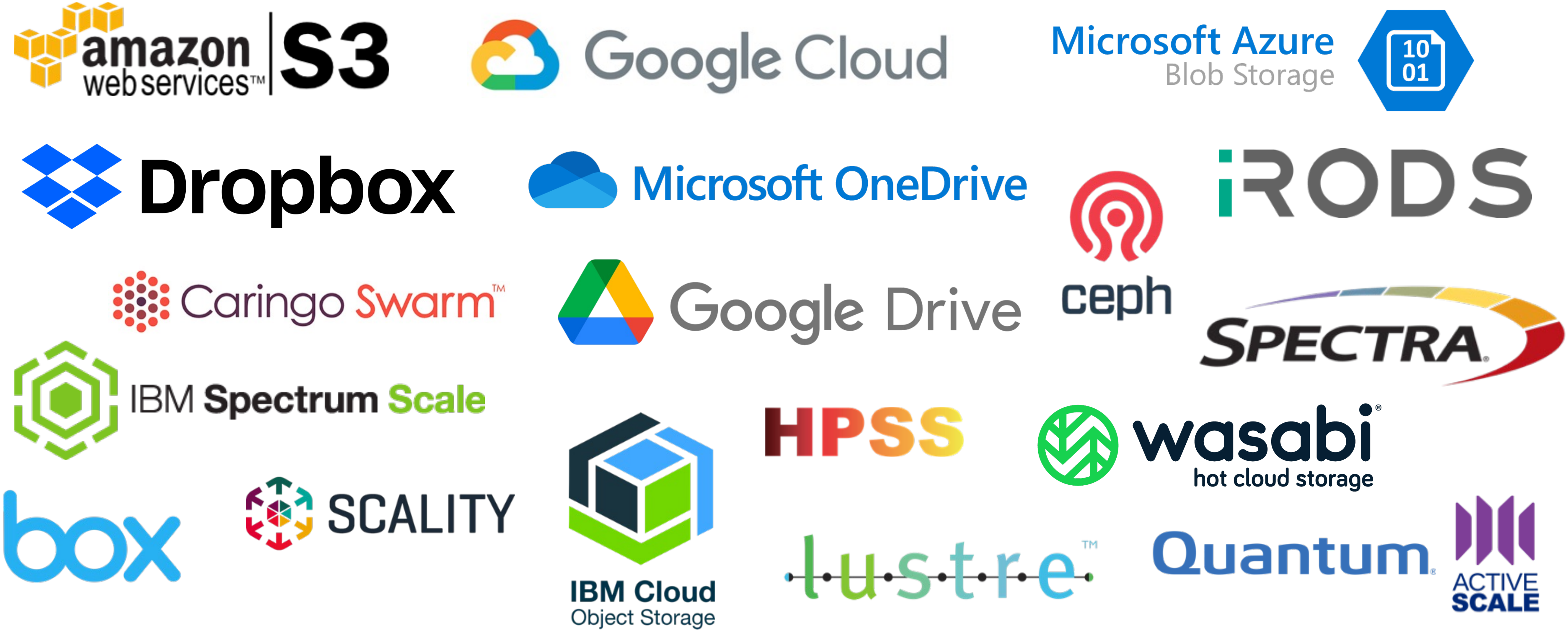
Get started by taking a short tour.

Search for a collection to begin

Get started by taking a short tour.

Software-as-a-Service

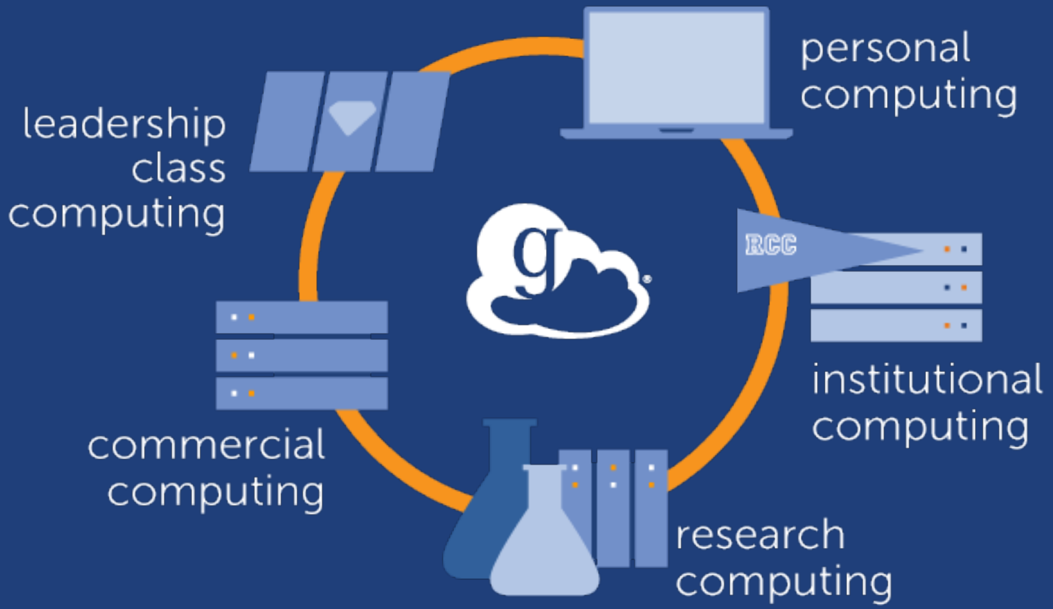
Globus supports diverse storage systems



Enabling reuse via remote computation

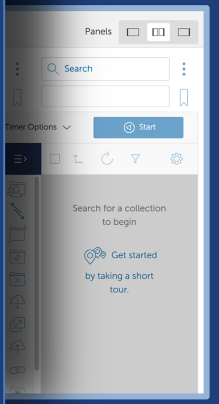
Managed transfer &

Collaborative data sh

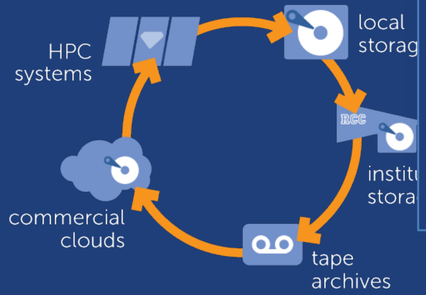


Managed remote execution

-a-Service


Panels
Search
Timer Options
Start
Search for a collection to begin
Get started by taking a short tour.

Software-as-a-Service

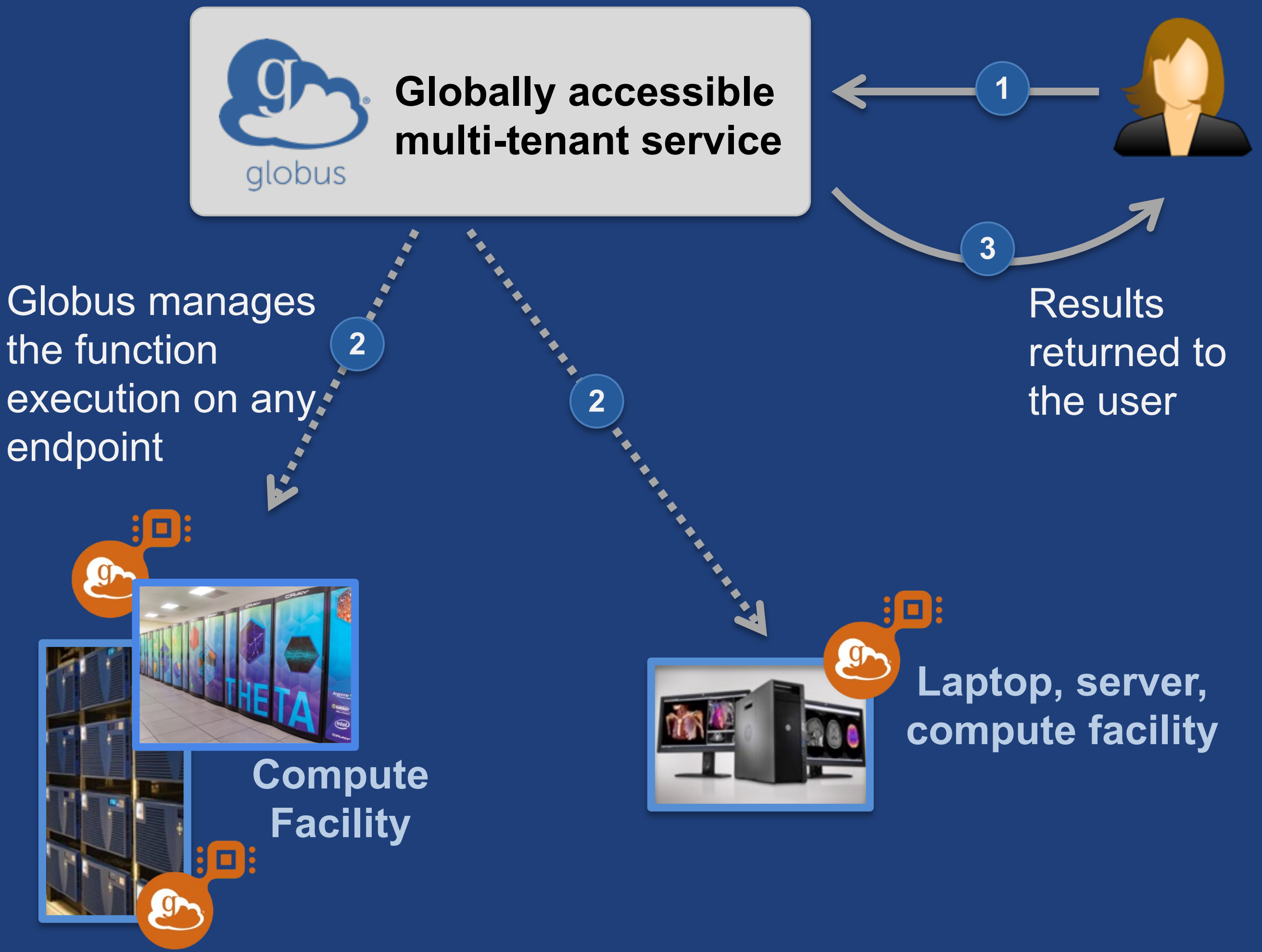


Unified data access

Reliable automation

Globus Compute: Function-as-a-Service …on any system

- Fire and forget function execution
- Uniform interface to diverse compute resources
- …from a laptop to a supercomputer



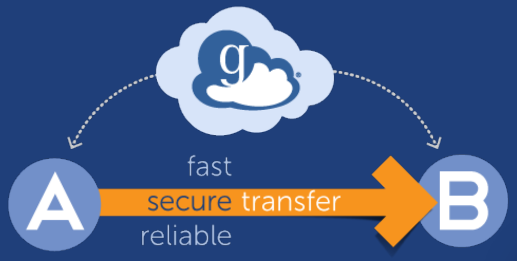
User submits a function to be run on compute endpoints

Globus manages the function execution on any endpoint

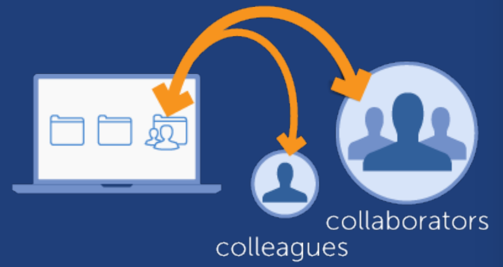
Results returned to the user

globus
Globally accessible multi-tenant service
1
2
2
3
THETA
Compute Facility
Laptop, server, compute facility

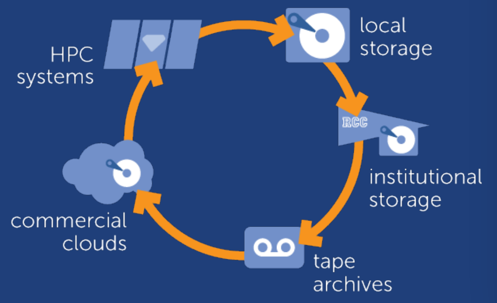
Scaling FAIRness in the research enterprise



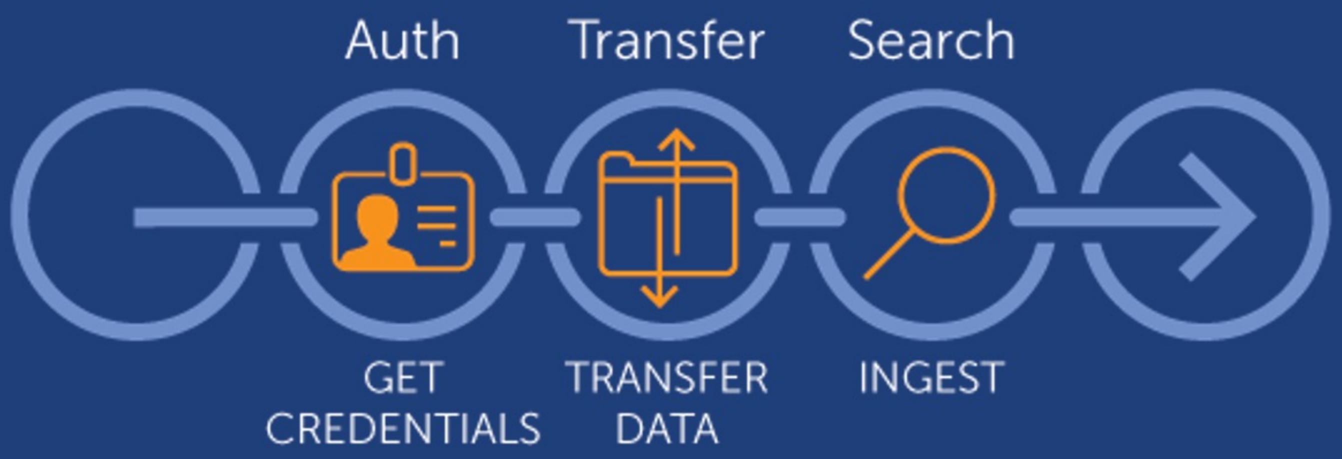
Managed transfer & sync



Collaborative data sharing



Unified data access

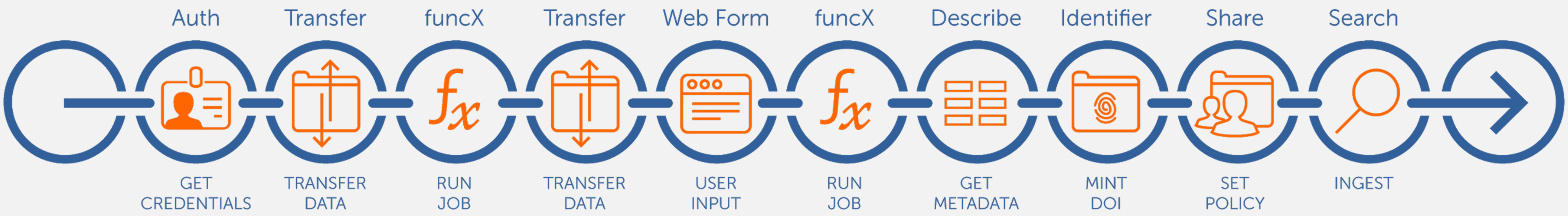


Reliable automation

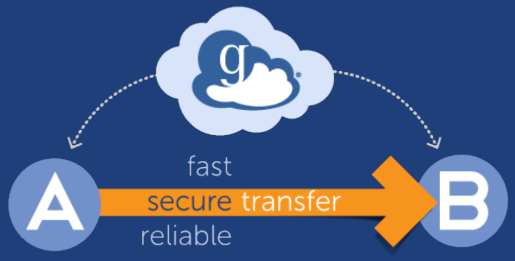
Globus Flows: Reliable, secure task orchestration

- 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

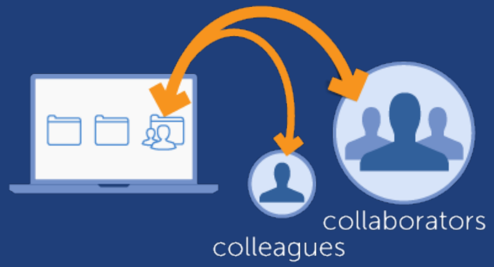


Extending FAIRness beyond the institution



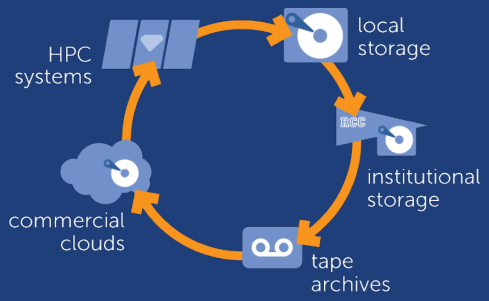
Managed transfer & sync

Publicatio

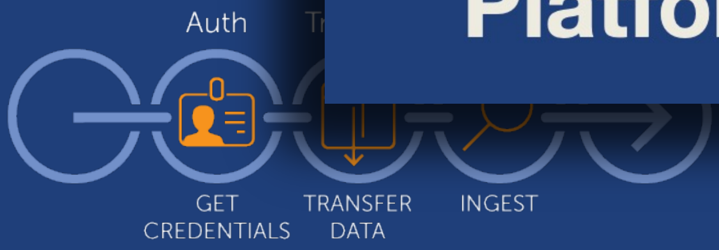


Collaborative data sharing

Managed re

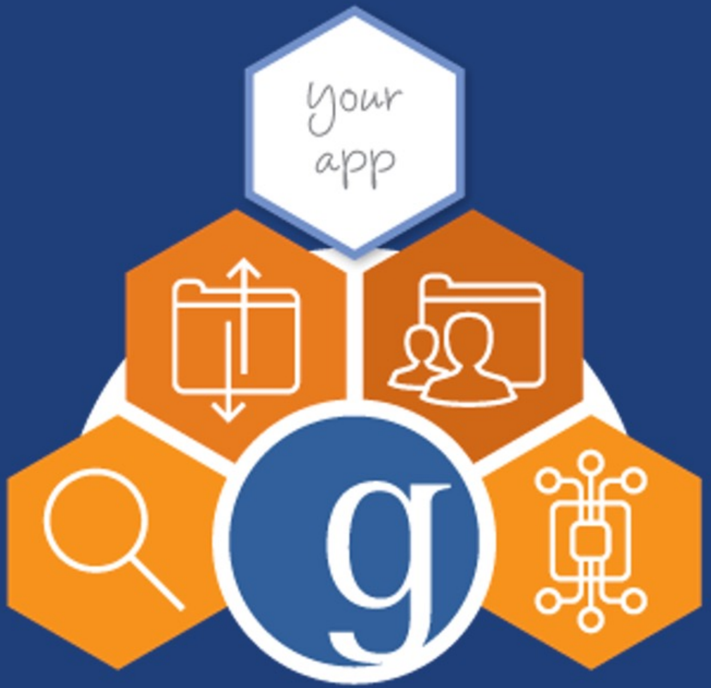


Unified data access



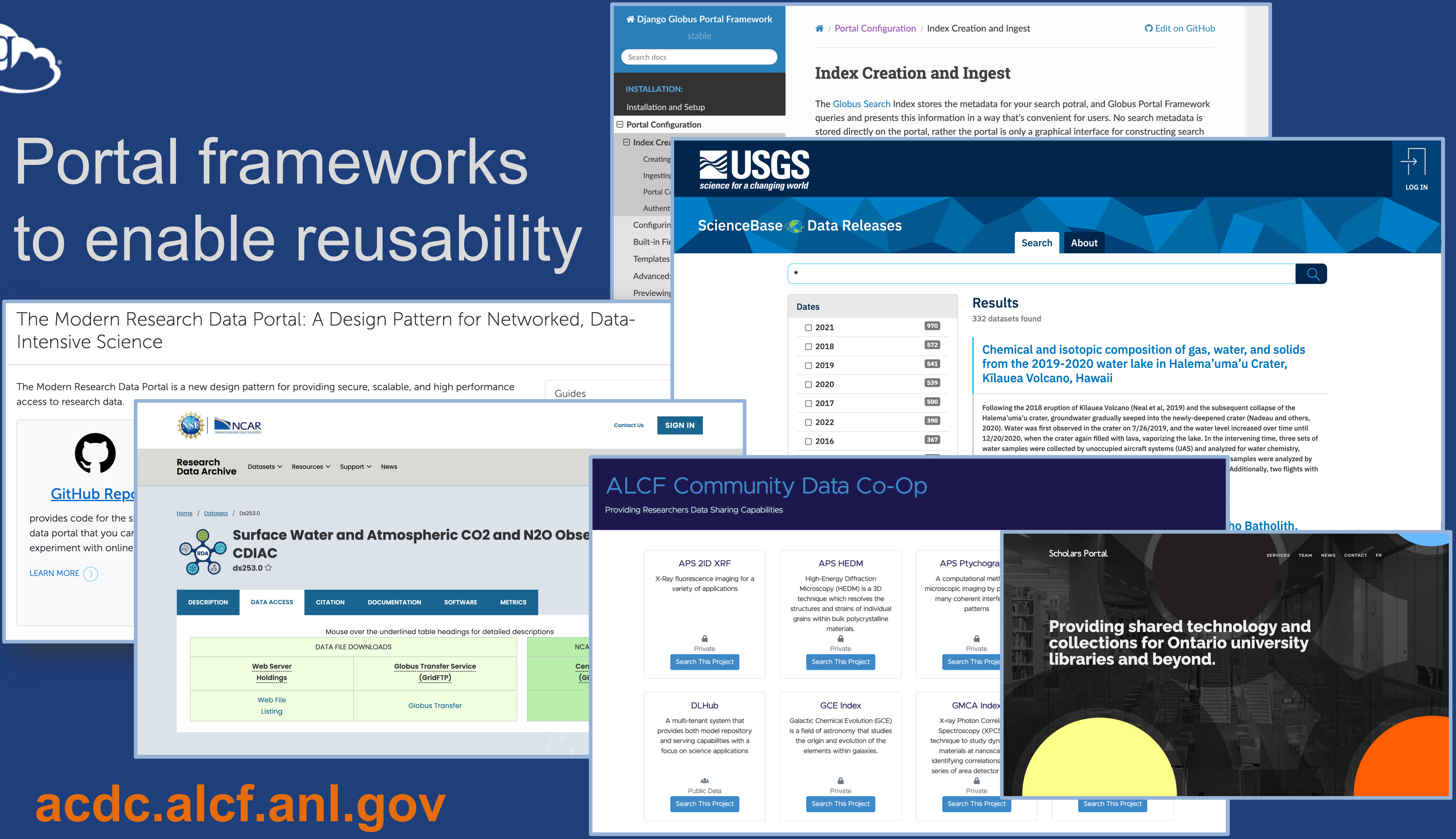
Reliable automation

Software-as-a-Service



Platform-as-a-Service

Portal frameworks to enable reusability



Django Globus Portal Framework

stable

Search docs

INSTALLATION:

Installation and Setup

Portal Configuration

Index Crea

Creating

Ingesting

Portal C

Authent

Configurin

Built-in Fie

Templates

Advanced:

Previewing

Portal Configuration / Index Creation and Ingest

Edit on GitHub

Index Creation and Ingest

The Globus Search Index stores the metadata for your search potral, and Globus Portal Framework queries and presents this information in a way that's convenient for users. No search metadata is stored directly on the portal, rather the portal is only a graphical interface for constructing search

USGS

science for a changing world

LOG IN

ScienceBase Data Releases

Search About

*

Dates

□ 2021	970
□ 2018	572
□ 2019	541
□ 2020	539
□ 2017	500
□ 2022	390
□ 2016	367

Results

332 datasets found

Chemical and isotopic composition of gas, water, and solids from the 2019-2020 water lake in Halema'uma'u Crater, Kīlauea Volcano, Hawaii

Following the 2018 eruption of Kīlauea Volcano (Neal et al, 2019) and the subsequent collapse of the Halema'uma'u crater, groundwater gradually seeped into the newly-deepened crater (Nadeau and others, 2020). Water was first observed in the crater on 7/26/2019, and the water level increased over time until 12/20/2020, when the crater again filled with lava, vaporizing the lake. In the intervening time, three sets of water samples were collected by unoccupied aircraft systems (UAS) and analyzed for water chemistry,

samples were analyzed by

Additionally, two flights with

ho Batholith.

The Modern Research Data Portal: A Design Pattern for Networked, Data-Intensive Science

The Modern Research Data Portal is a new design pattern for providing secure, scalable, and high performance access to research data.

Guides

GitHub Repo

provides code for the s data portal that you car experiment with online

LEARN MORE

NSF NCAR

Contact Us

SIGN IN

Research Data Archive

Datasets Resources Support News

Home / Datasets / Ds253.0

Surface Water and Atmospheric CO2 and N2O Obse CDIAC

ds253.0 ☆

DESCRIPTION | DATA ACCESS | CITATION | DOCUMENTATION | SOFTWARE | METRICS

Mouse over the underlined table headings for detailed descriptions

DATA FILE DOWNLOADS	
Web Server Holdings	Globus Transfer Service (GridFTP)
Web File Listing	Globus Transfer

NCA

Cen (Gl

ALCF Community Data Co-Op

Providing Researchers Data Sharing Capabilities

APS 2ID XRF

X-Ray fluorescence imaging for a variety of applications

Private

Search This Project

APS HEDM

High-Energy Diffraction Microscopy (HEDM) is a 3D technique which resolves the structures and strains of individual grains within bulk polycrystalline materials.

Private

Search This Project

APS Ptychogra

A computational met microscopic imaging by p many coherent interfe patterns

Private

Search This Proje

DLHub

A multi-tenant system that provides both model repository and serving capabilities with a focus on science applications

Public Data

Search This Project

GCE Index

Galactic Chemical Evolution (GCE) is a field of astronomy that studies the origin and evolution of the elements within galaxies.

Private

Search This Project

GMCA Index

X-ray Photon Correl Spectroscopy (XPCS technique to study dyn materials at nanosca identifying correlations series of area detector

Private

Search This Project

Search This Project

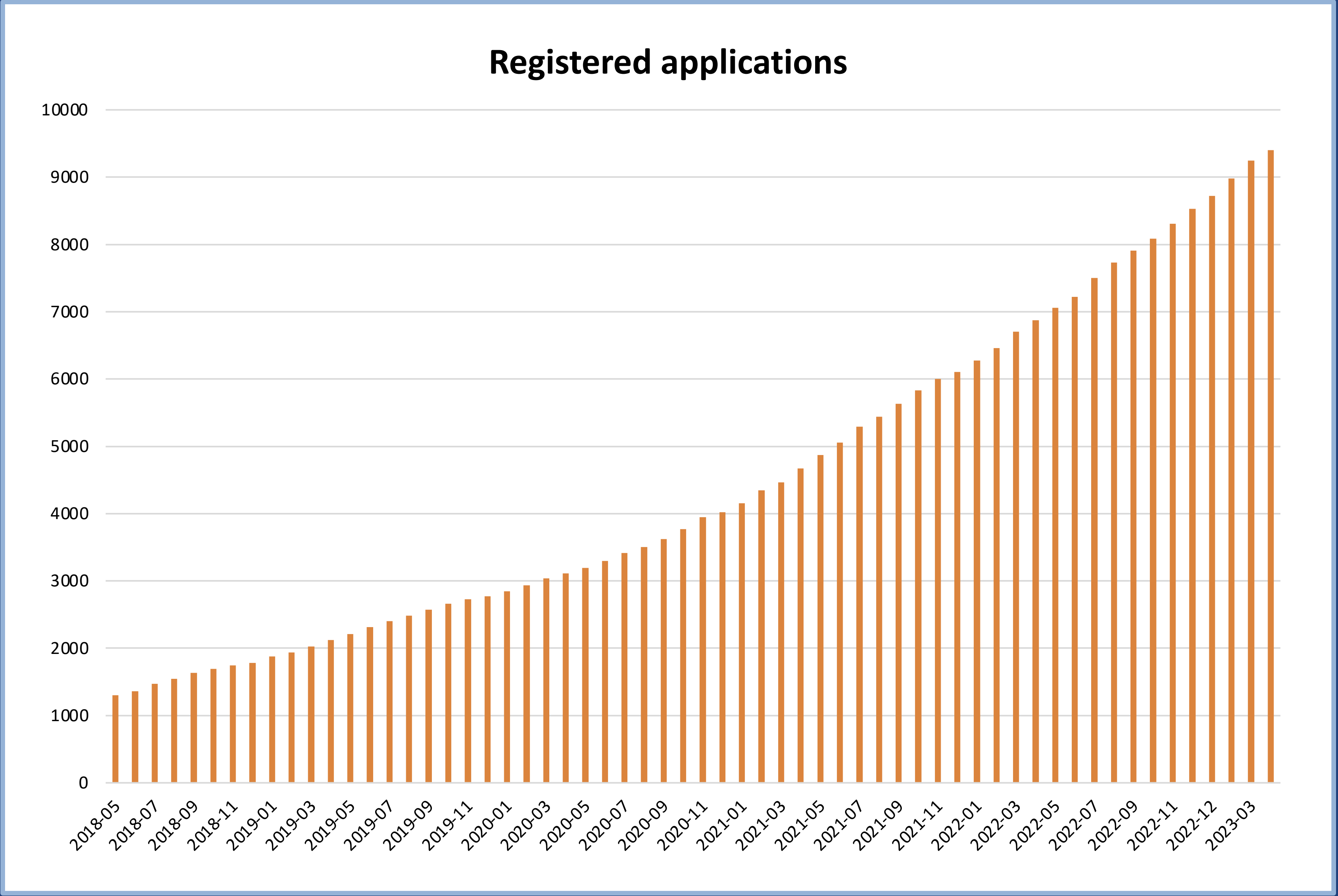
Scholars Portal

SERVICES TEAM NEWS CONTACT FR

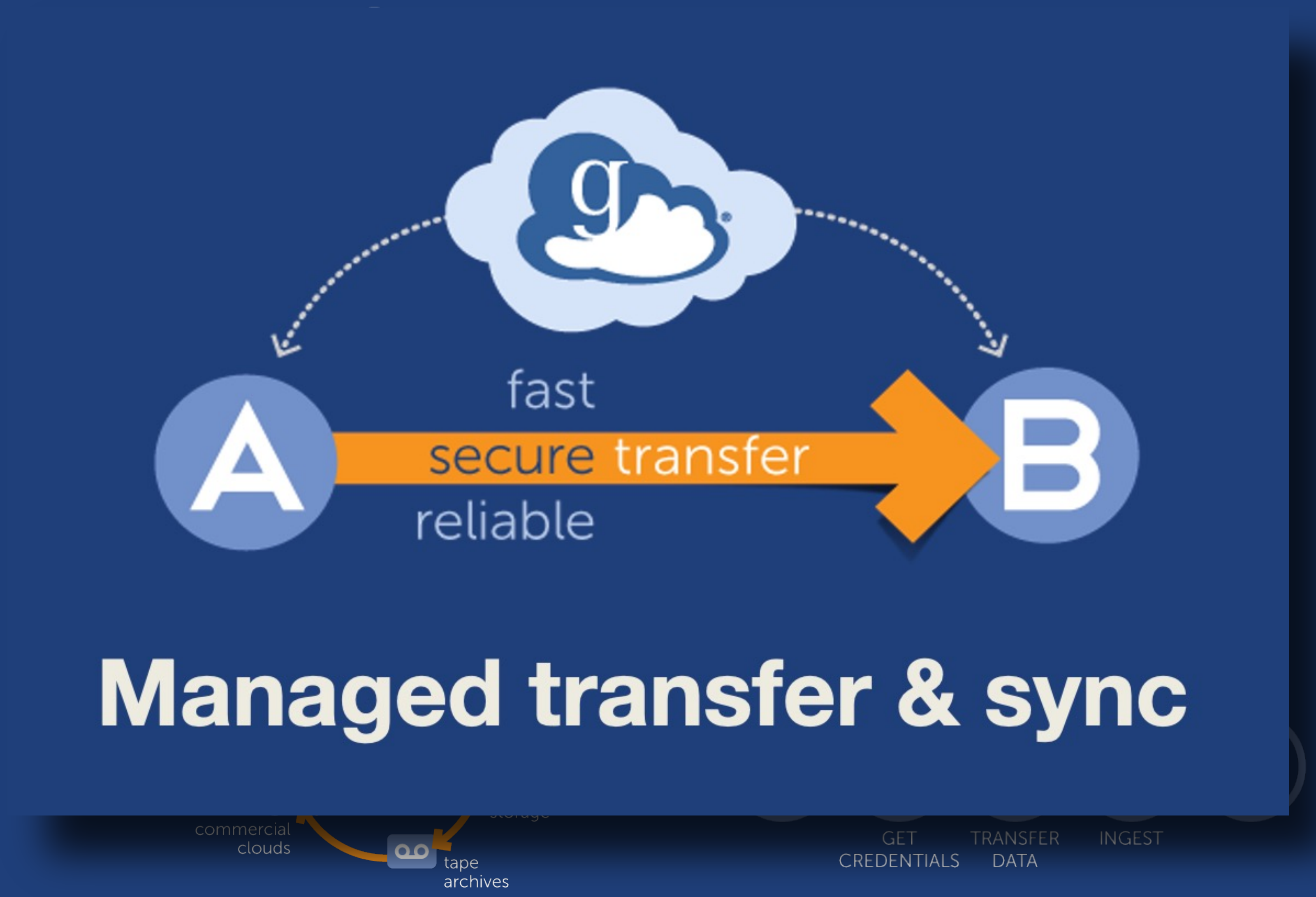
Providing shared technology and collections for Ontario university libraries and beyond.

acdc.alcf.anl.gov

A growing ecosystem



Data FAIRness through data mobility

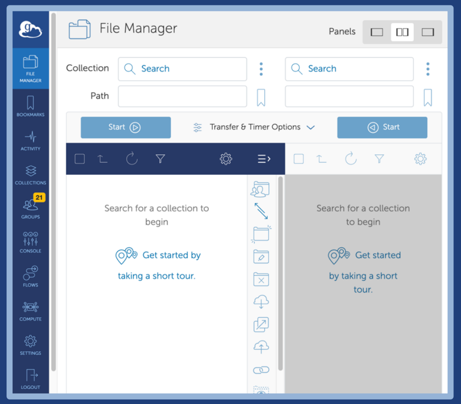


Managed transfer & sync

Unified data access

Reliable automation

Platform-as-a-Service



Search for a collection to begin

Get started by taking a short tour.

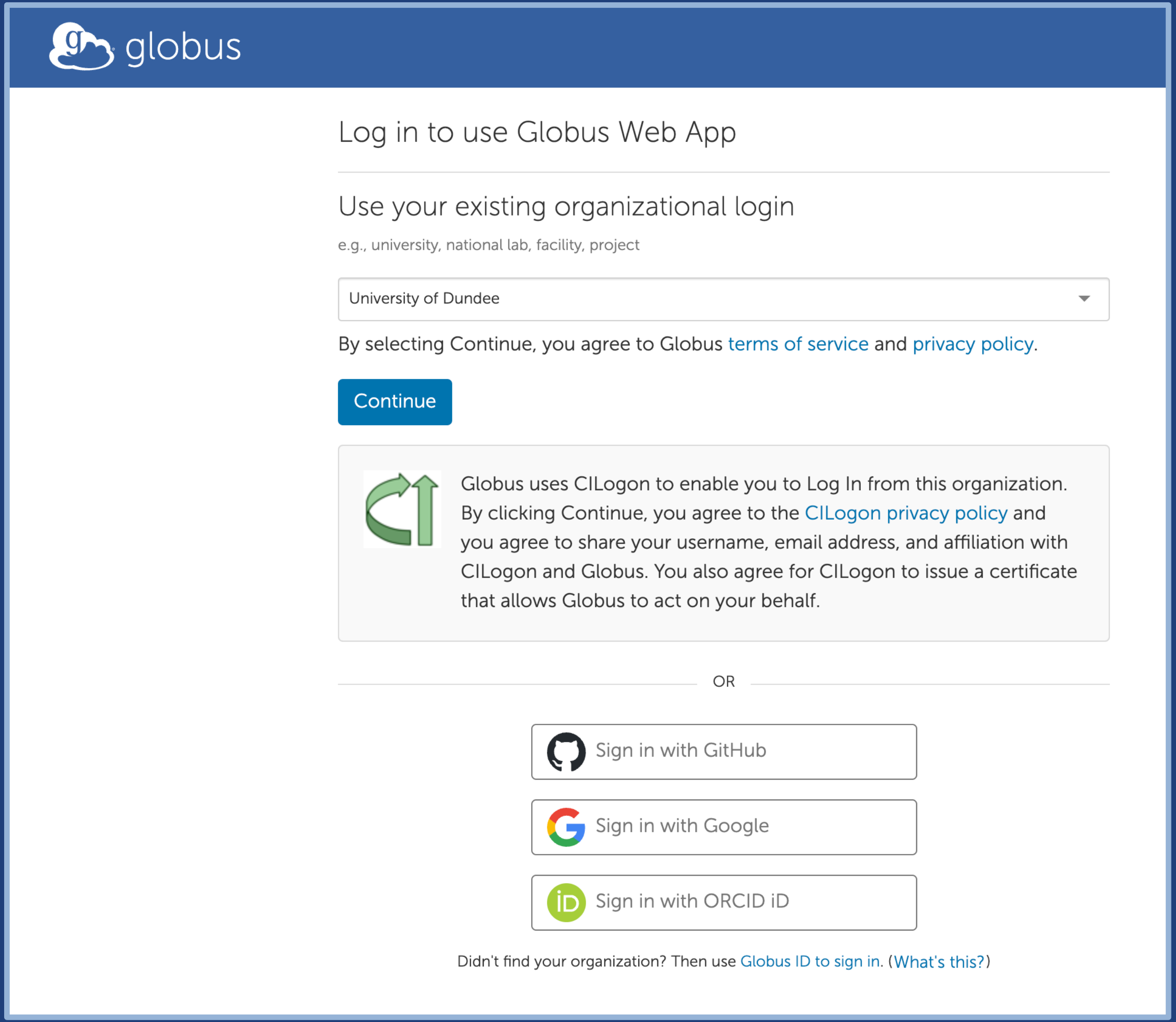
Search for a collection to begin

Get started by taking a short tour.

Software-as-a-Service

Readily accessible by researchers everywhere

- Federated login for 1,800+ institutions
- Access via ORCID ID for millions of PIs
- Open to the world via Google, GitHub, …



Log in to use Globus Web App

Use your existing organizational login

e.g., university, national lab, facility, project

University of Dundee

By selecting Continue, you agree to Globus terms of service and privacy policy.

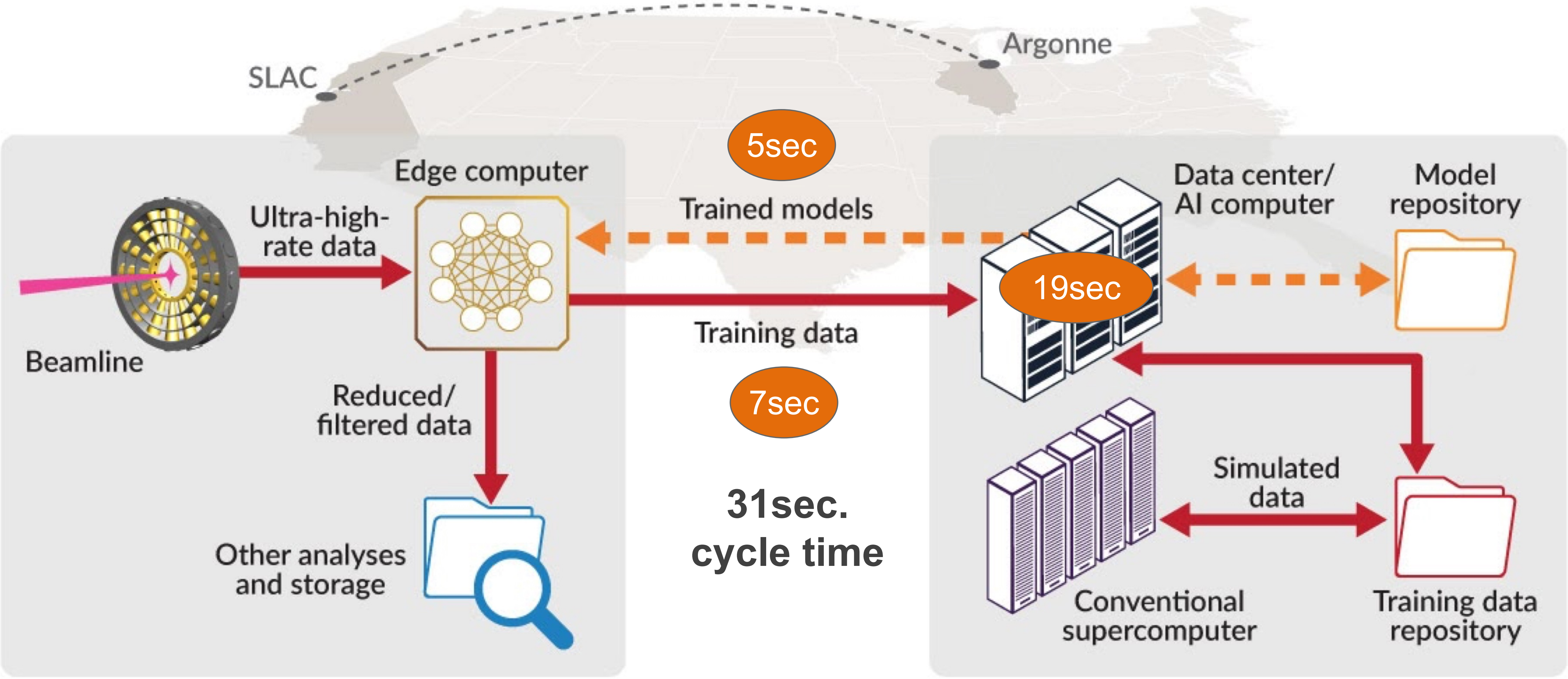
Globus uses CILogon to enable you to Log In from this organization. By clicking Continue, you agree to the CILogon privacy policy and you agree to share your username, email address, and affiliation with CILogon and Globus. You also agree for CILogon to issue a certificate that allows Globus to act on your behalf.

Didn't find your organization? Then use Globus ID to sign in. (What's this?)

Large-scale automation is increasingly enabling intelligent analysis and experiment steering

What keeps you up at night?

“Smart” instruments are emerging



31sec. cycle time

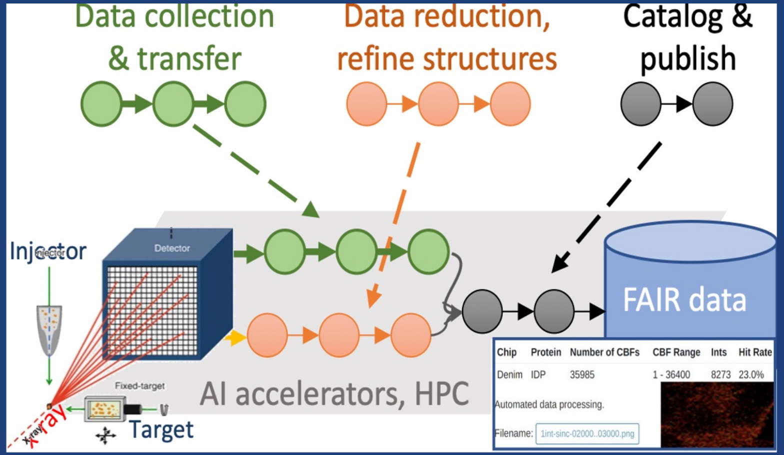
Current research will increase access to advanced instrument automation capabilities

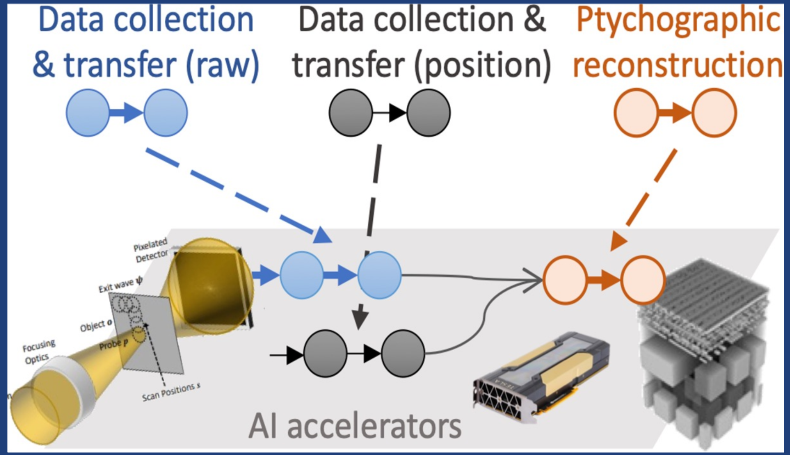
Braid (2020-2023) developed methods and tools for defining, running, and scheduling flows that link instruments, computers, data repositories, people

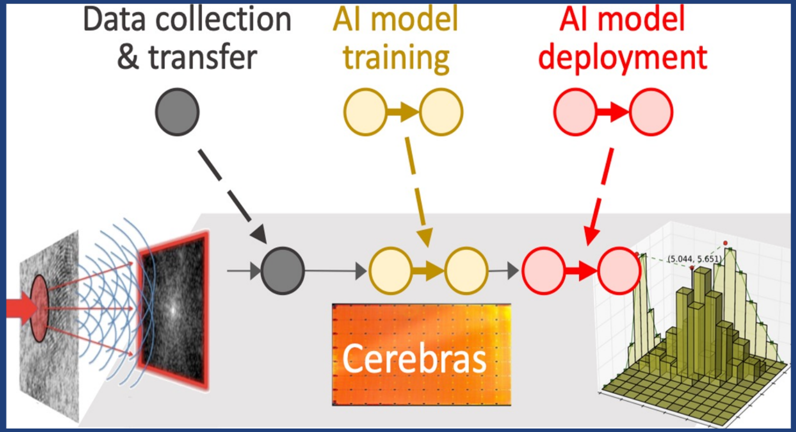
- **Method and tools:** Globus Flows for representing flows, Braid for scheduling sets of flows, HPC DL training algorithms, data pub methods, etc.
- **Deployment and application:** APS, LCLS, microscopes, CryoEM, etc.; ALCF and other compute
- **Online experimental data analysis, experiment steering, data ingest pipelines, climate data analysis, etc.**
- **Operational experience:** 10,000s of flow runs, 100s TB data processed, 10,000s node hours
- **Science impact:** Determination of Covid protein structures, online high energy diffraction microscopy analysis, etc.

https://doi.org/10.1016/j.patter.2022.100606

Blaiszik, Chard, Chard, Foster, Huerta, Nicolae, Vescovi, Wozniak


Data collection & transfer
Data reduction, refine structures
Catalog & publish
Injector
Detector
FAIR data
Fixed-target
X-ray
Target
AI accelerators, HPC
Chip Protein Number of CBFs CBF Range Ints Hit Rate
Denim IDP 35985 1 - 36400 8273 23.0%
Automated data processing.
Filename:


Data collection & transfer (raw)
Data collection & transfer (position)
Ptychographic reconstruction
Pixelated Detector
Exit wave ψ
Object o
Probe P
Focusing Optics
Scan Positions s
AI accelerators

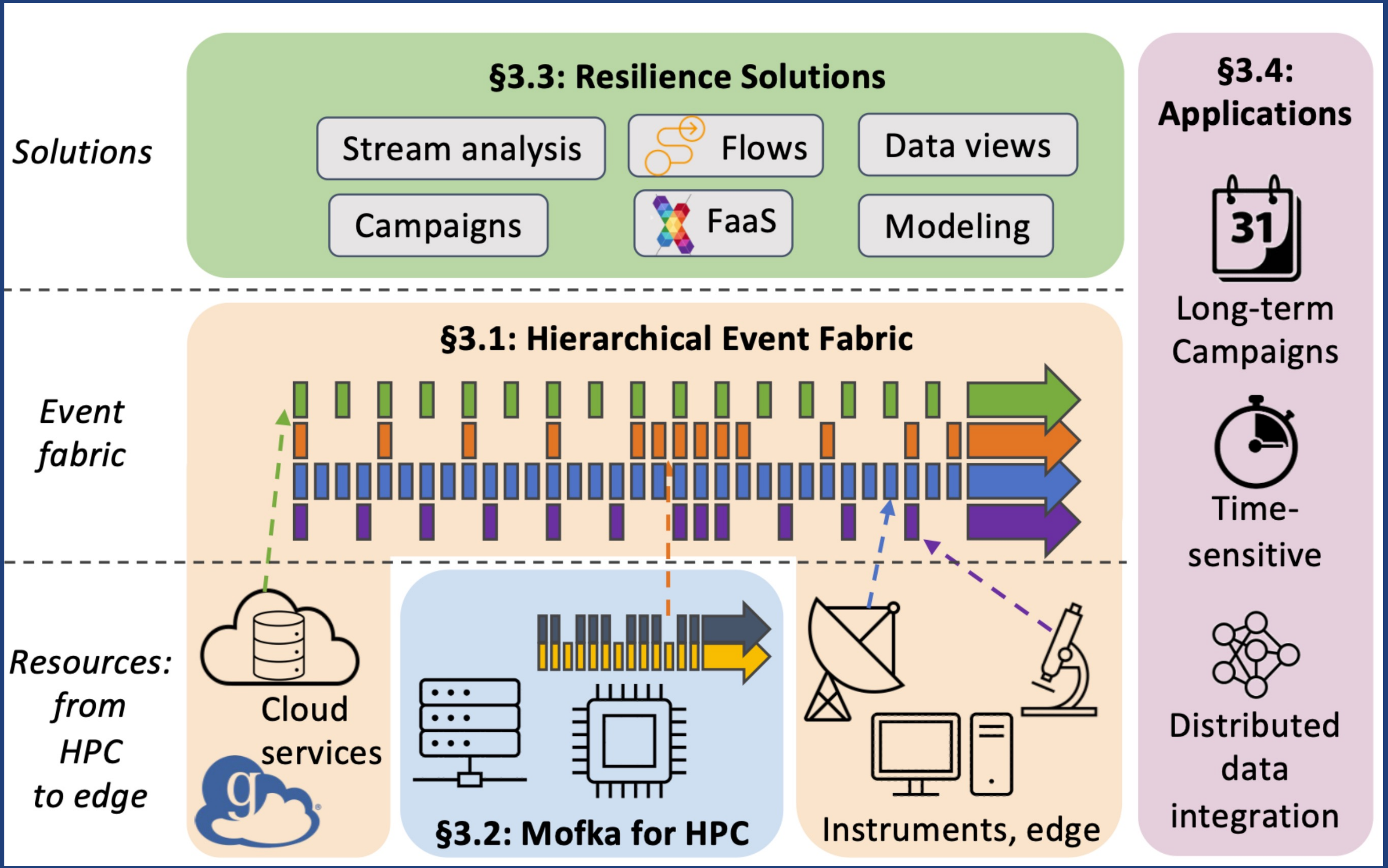


Diaspora (2023-2028) will develop a hierarchical event fabric and resilience solutions that will be applied to distributed applications involving long-term campaigns, time-sensitive analysis, and distributed data integration

- **Argonne, ORNL, SLAC, Texas Tech University**
- **Event fabric supports eventing, monitoring, and resilience across institutions/facilities**
- **Resilient distributed data structures can be deployed across diverse resources**

https://diaspora-project.github.io

Foster, Rao, Thayer, Corsi



§3.3: Resilience Solutions
Solutions
Stream analysis
Flows
Data views
Campaigns
FaaS
Modeling
§3.4: Applications
Long-term Campaigns
Time-sensitive
Distributed data integration
§3.1: Hierarchical Event Fabric
Event fabric
Resources: from HPC to edge
Cloud services
§3.2: Mofka for HPC
Instruments, edge
31