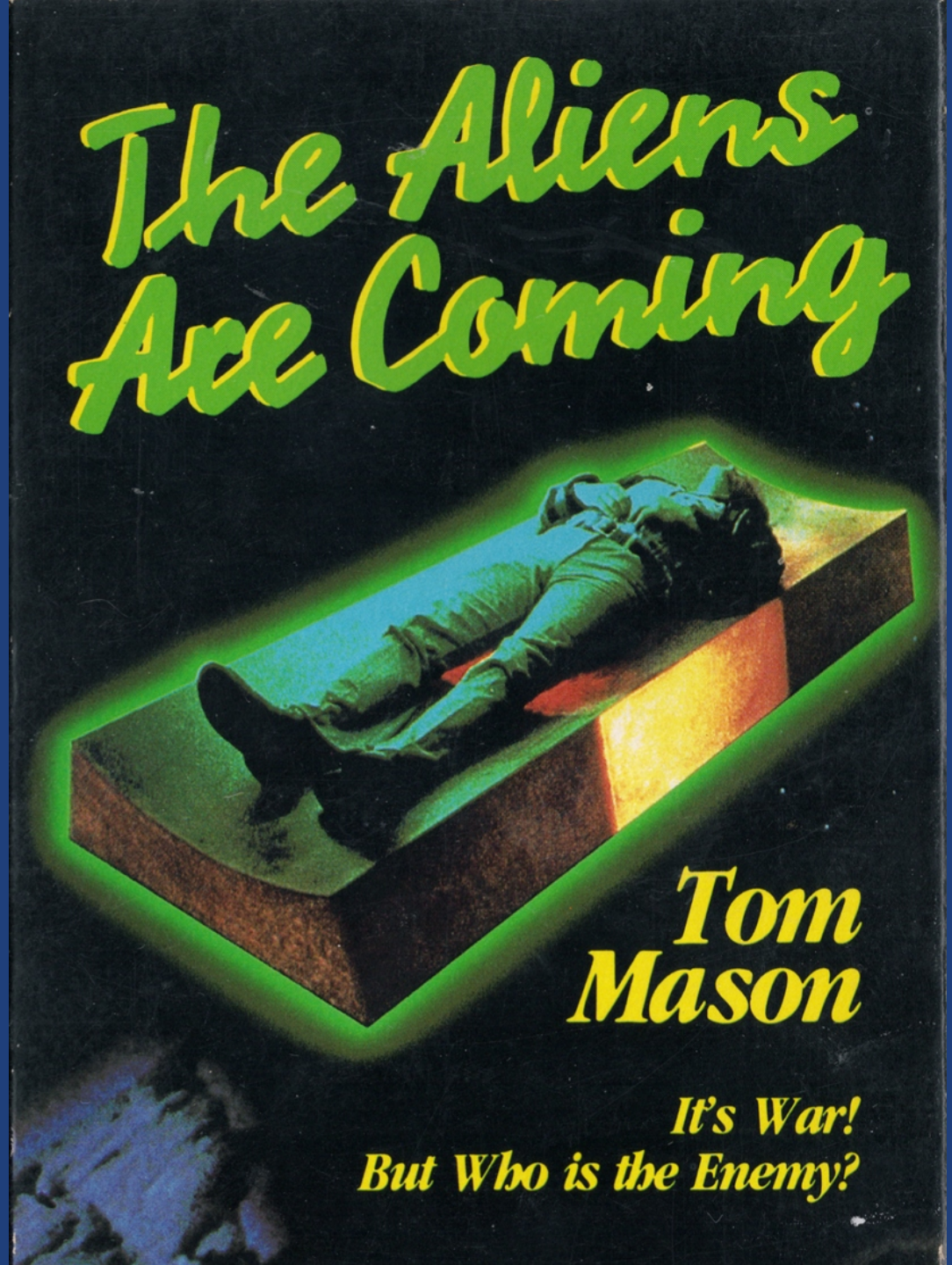
Scalable Data Management for National Facilities Using the Modern Research Data Portal

Vas Vasiliadis
Rachana Ananthakrishnan

SC20, November 2020

Circa 1980…



Circa 2020: The Instruments are Coming!

NEW YORK GENOME CENTER®

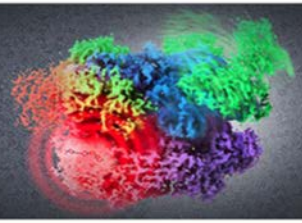
ADVANCED LIGHT SOURCE

NIH Intramural Sequencing Center

PURDUE SCIENCE

Purdue Cryo-EM Facility

High Resolution Cryo-EM



National Cryo-Electron Microscopy Facility

globus
a uchicago non-profit service

I Want To... Pricing Resources Support About Log In

Transfer your data.

Gigabytes, terabytes, petabytes—research data is large and distributed. Globus lets you efficiently, securely, and reliably transfer data directly between systems separated by an office wall or an ocean. Focus on your research and offload your data transfer headaches to Globus.

LEARN MORE

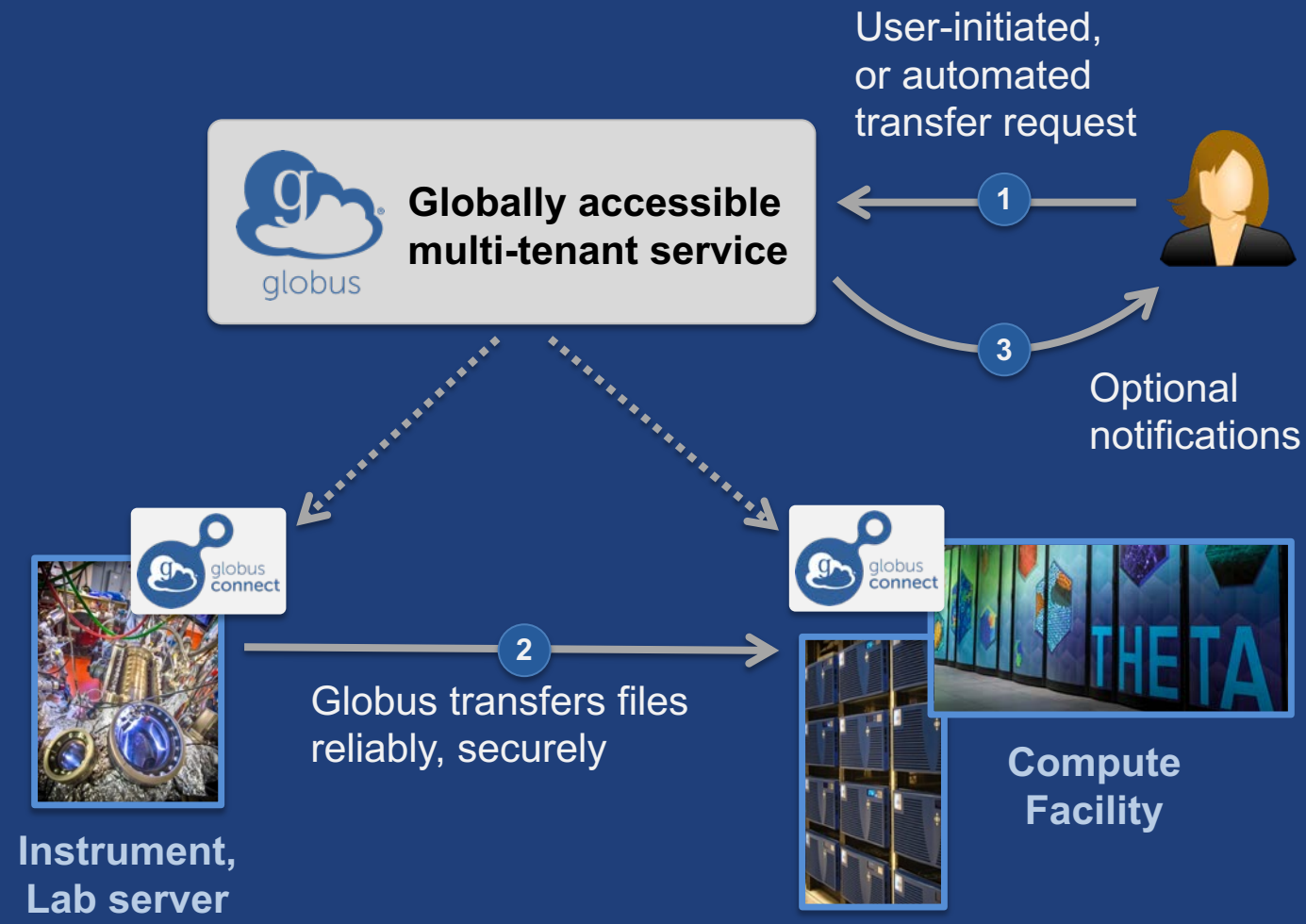
TRANSFER DATA NOW

CONNECT YOUR SYSTEM TO GLOBUS

1,009,341,031,957 MB TRANSFERRED

Fast, reliable file transfer …from any to any system

- **Fire-and-forget transfers**
- **Optimized speed**
- **Assured reliability**
- **Unified view of storage**
- **Browser, REST API, CLI**



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**

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

1

globus

Globally accessible multi-tenant service

Globus controls access to shared files on existing storage

On-prem or public cloud storage

2

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

Laptop, server, compute facility

Point-and-click scales only up to a point!

Automated instrument data egress/distribution

- **Reliable, near-real time data access**
- **Automatically set policy based permissions**
- **Self-service access control, management**
- **Federated login for frictionless data access**

Cryo EM
Lightsheet
Sequencer
ALS/APS
….

--/cohort045
--/cohort096
--/cohort127

Local policy store

Local system download

Remote analysis, visualization

Advanced Photon Source

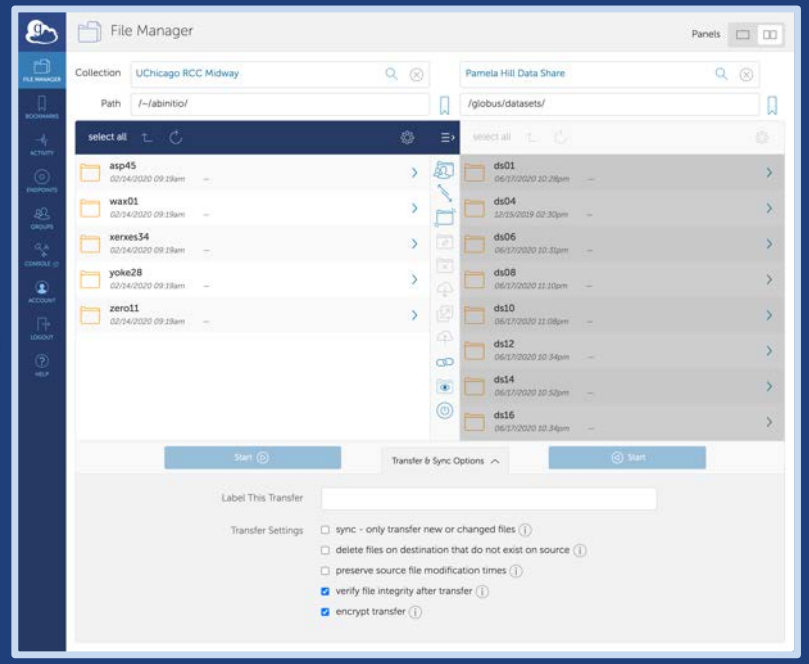
- 2-D, 3-D imaging
- 2016: ~112TB/month
- 100x – 1,000x growth

Argonne NATIONAL LABORATORY

Use(r)-appropriate interfaces

Globus service

Web



CLI

```
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.
```

Platform (RESTful APIs)

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

Automatically tag, share, notify, and distribute data

DMagic a Globus implementation at the APS

http://dmagic.readthedocs.org

TomoPy

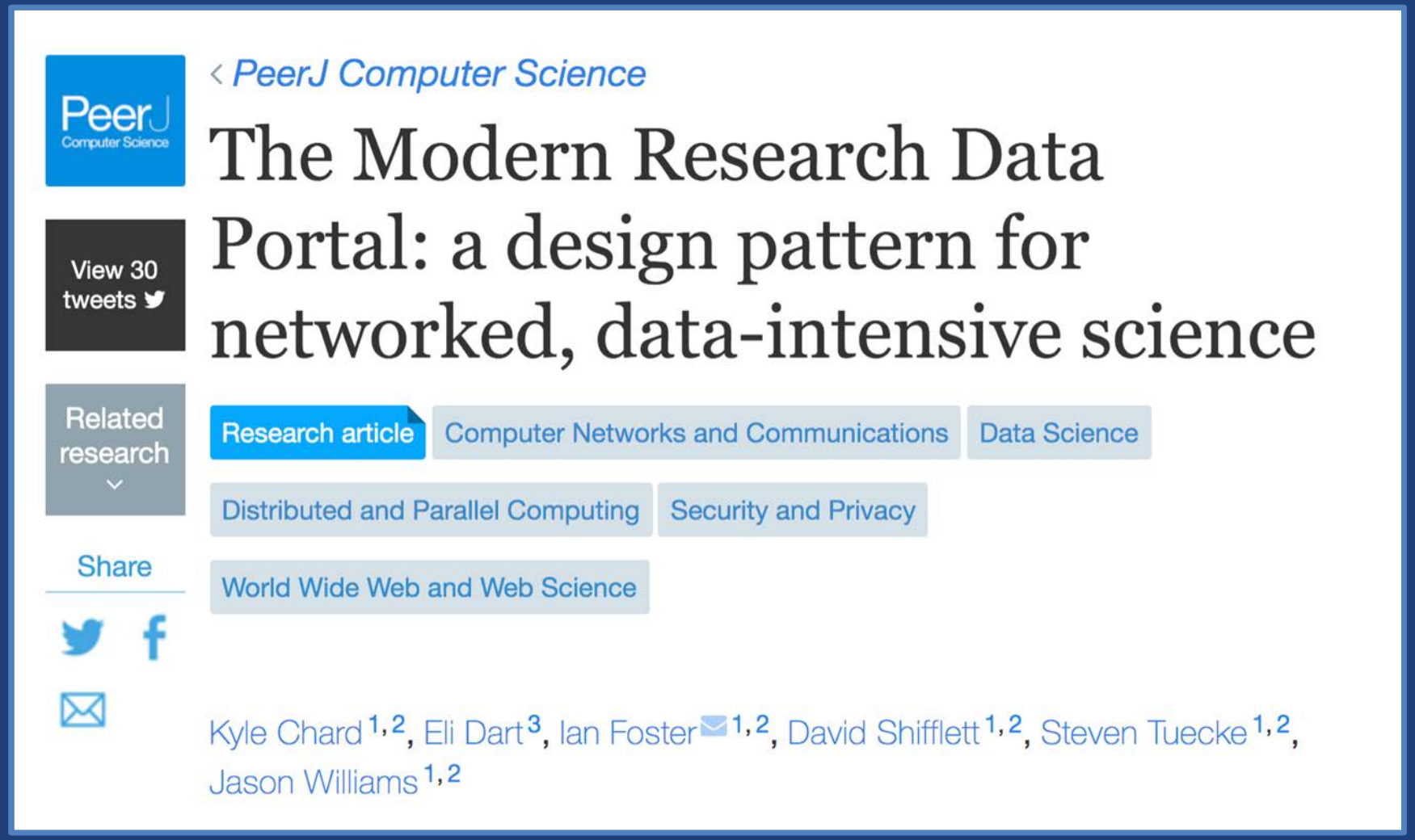
Tomographic reconstruction in Python

Doga Gursoy

http://tompy.readthedocs.org

Bespoke solutions only get us so far…

MRDP to the rescue

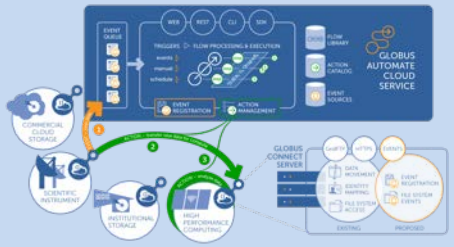


Source: peerj.com/articles/cs-144

MRDP: Key elements

Globus Platform

Secure, reliable data orchestration



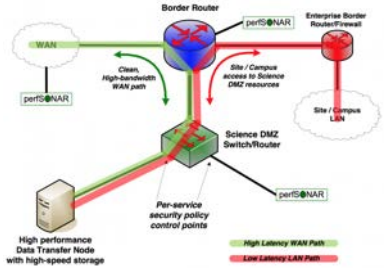
Globus Connect

Storage system access

globus connect

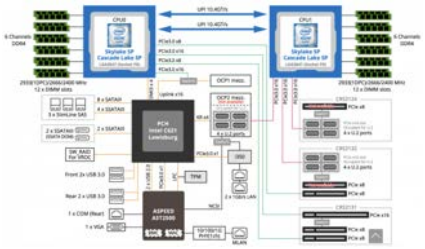
Science DMZ

Fast, clean data path



Data Transfer Node

Purpose-built data mover



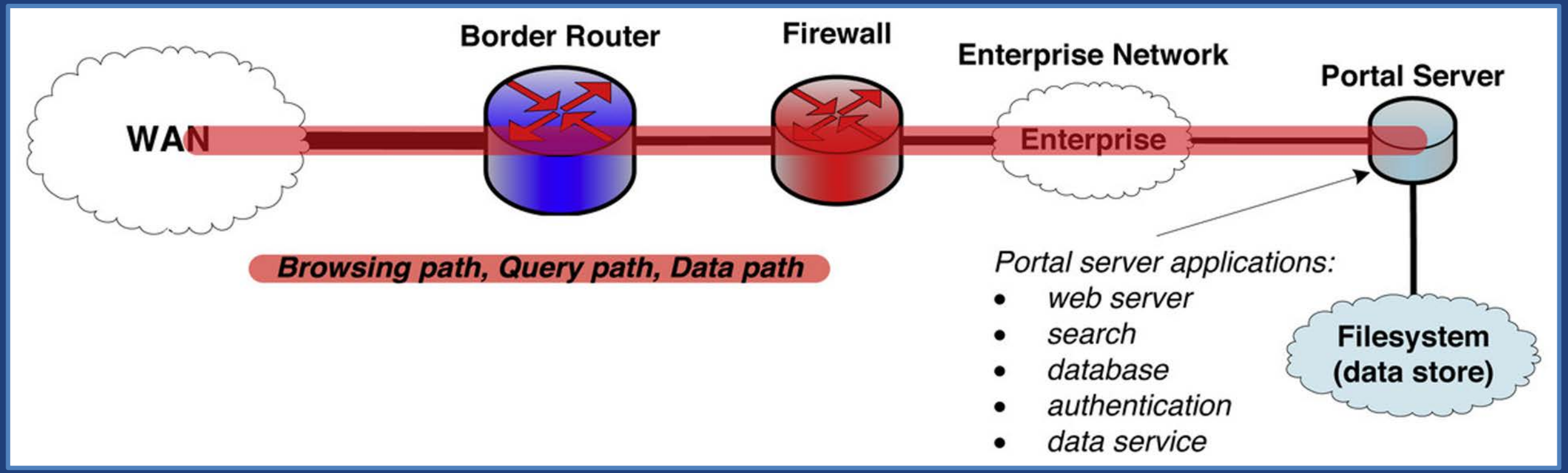
What's wrong with my LRDP?

"If you're doing something the same way you have been doing it for ten years, the chances are you are doing it wrong."

– Charles Kettering

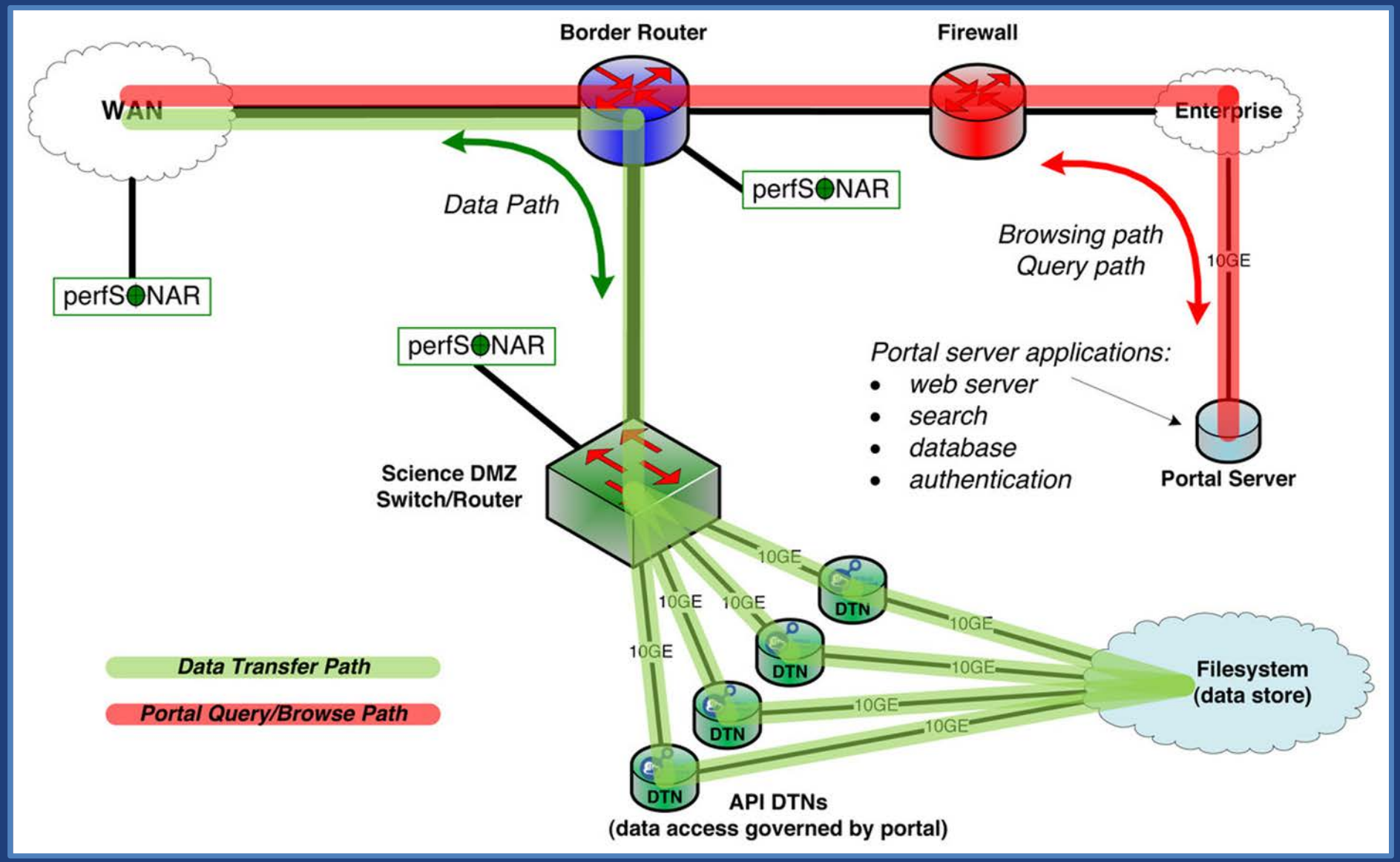
Source: wikipedia.org/wiki/Charles_F._Kettering

Legacy Research Data Portal architecture



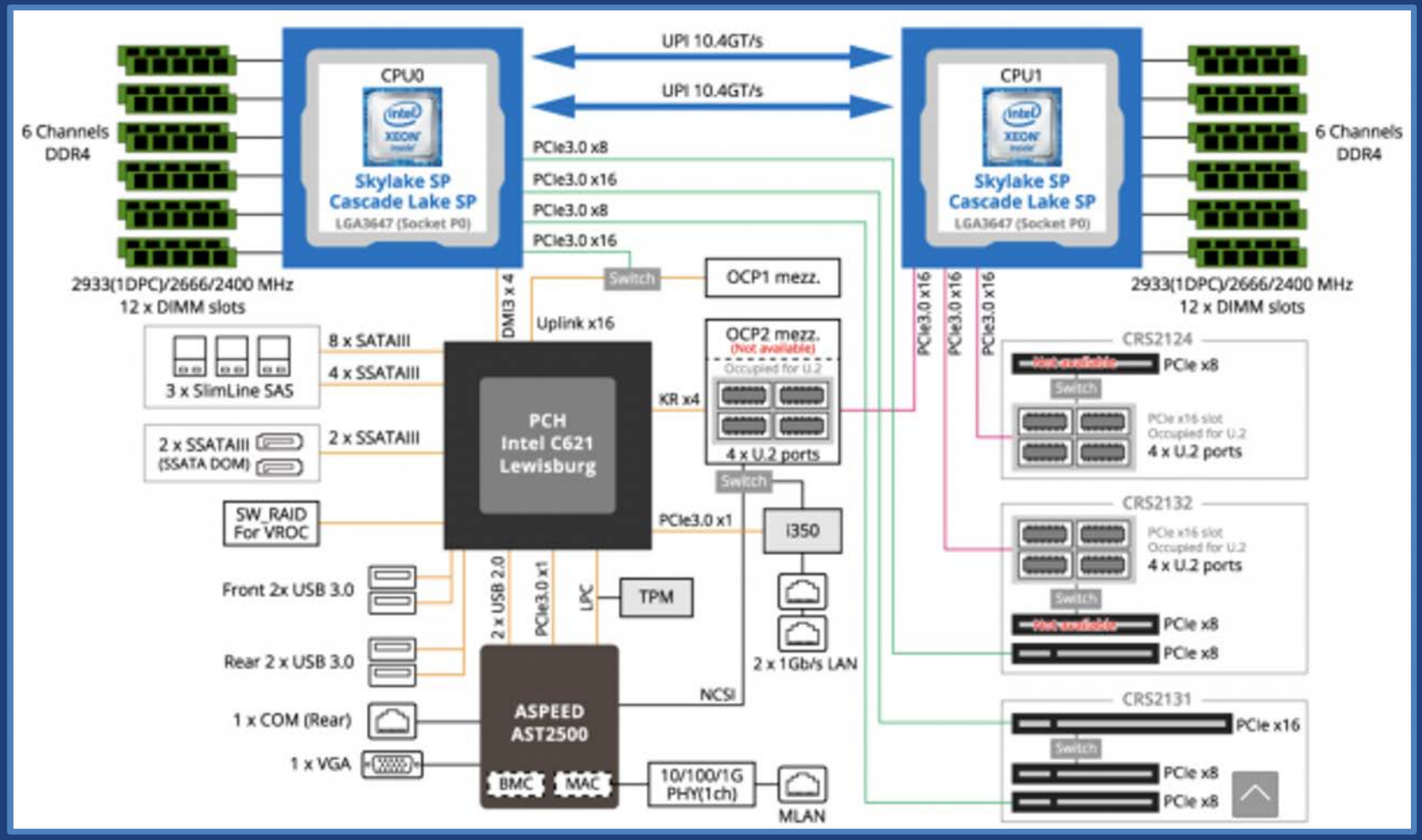
Source: ESnet Science Engagement team

MRDP network architecture: Science DMZ



Source: ESnet Science Engagement team

The Data Transfer Node



Source: ESnet Science Engagement team

fasterdata.es.net/science-dmz/DTN/reference-implementation

…makes diverse storage systems accessible via Globus

Globus Connectors

Coming soon…

iRODS

globus connect

Community Developed

Globus Developed

Microsoft Azure Blob Storage

OneDrive

Dropbox

Requirements for instrument data orchestration

- **Authentication and Authorization**
- **Automated data ingest**
- **Data and compute orchestration**
- **Data description and discovery**

Enabling Globus platform services

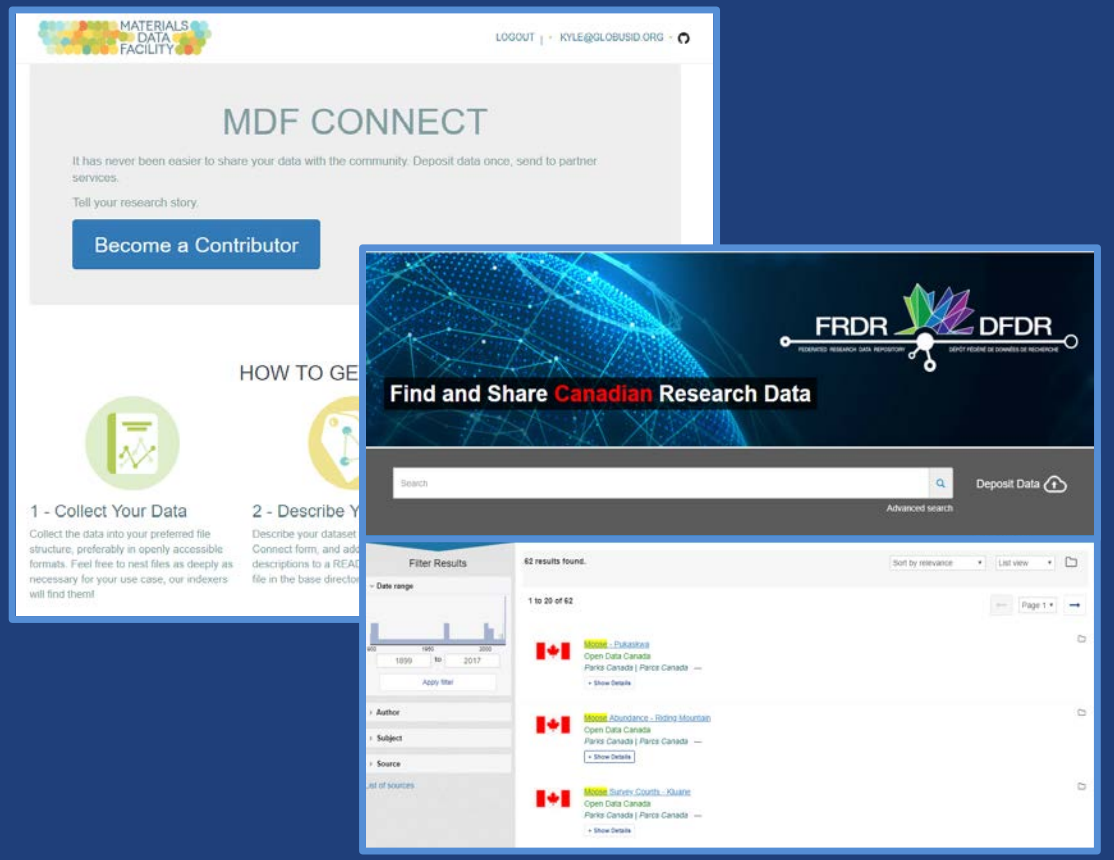
- **Identity and Access Management (IAM): Auth, Groups**
- **Data Services: Connect, Transfer, Manifest***
- **Search**
- **Identifiers (collaboration with DataCite)**
- **Automate***

Globus Auth: Foundational IAM service

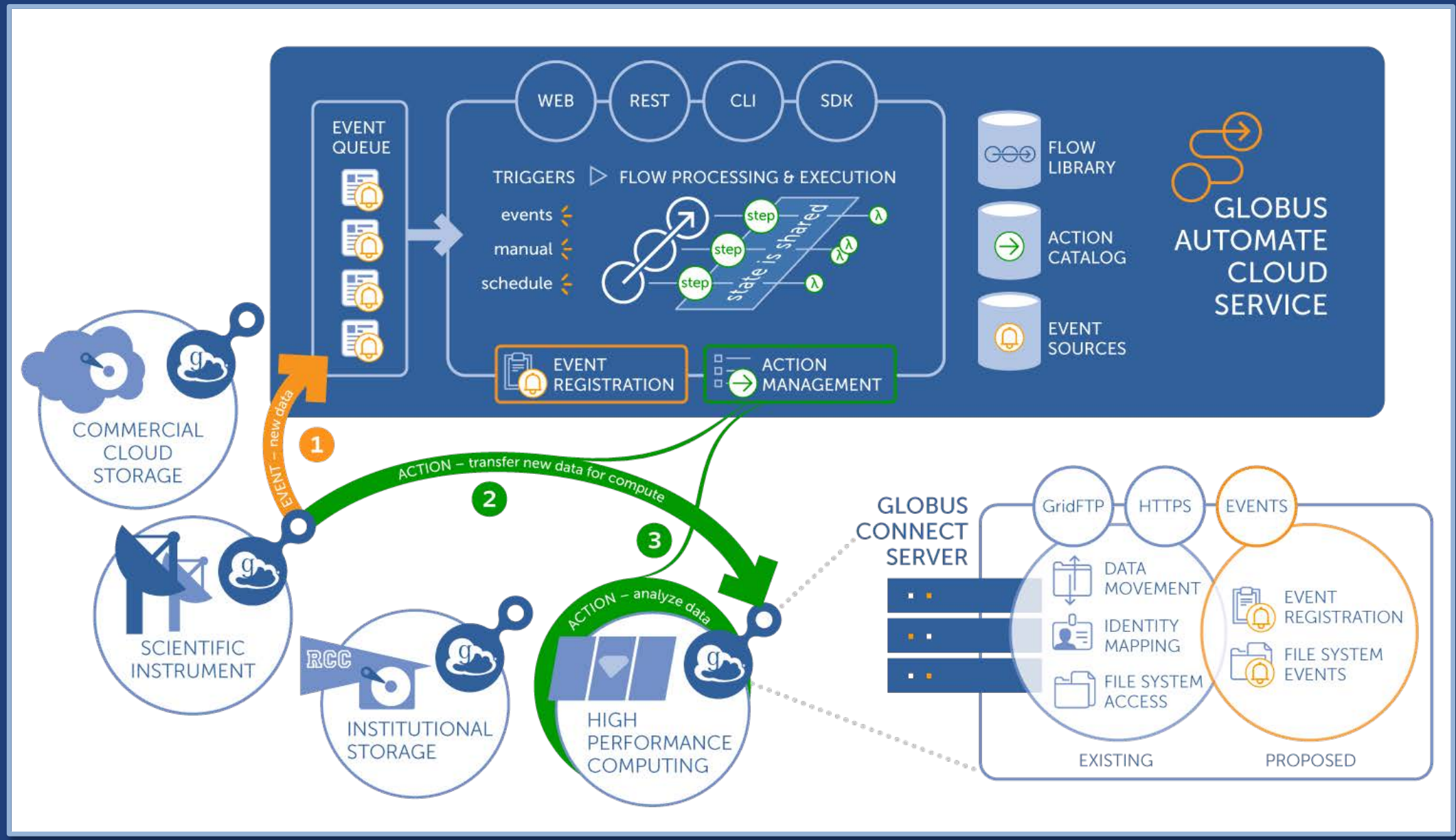
- **Enables login for diverse app ecosystem**
- **Protects REST API communications between and among apps and services**
- **Use existing identities: 1,000+ trusted IdPs & growing**
- **Employs least privileges security model**
- **Uses OAuth2 and OpenID Connect standards**
- **Programming language and framework agnostic**

Globus Search service

- **Scalable service for research data discovery**
- **Schema agnostic**
- **Fine grained access control**
- **Robust search capabilities**
  - Plain text
  - Facets
  - Rich query language



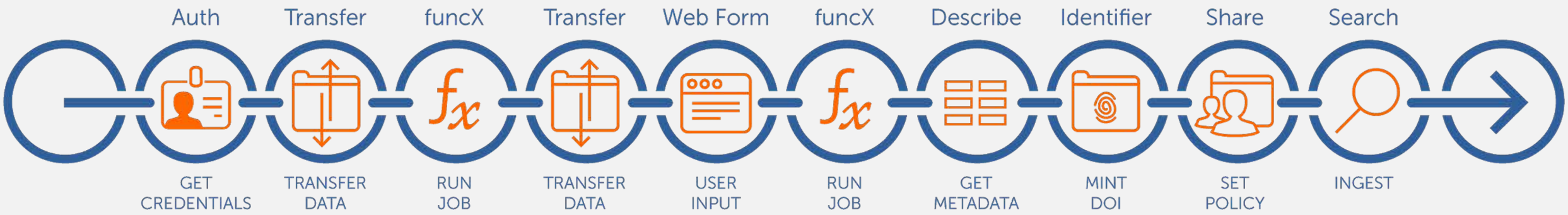
Globus automation platform



Globus Automate

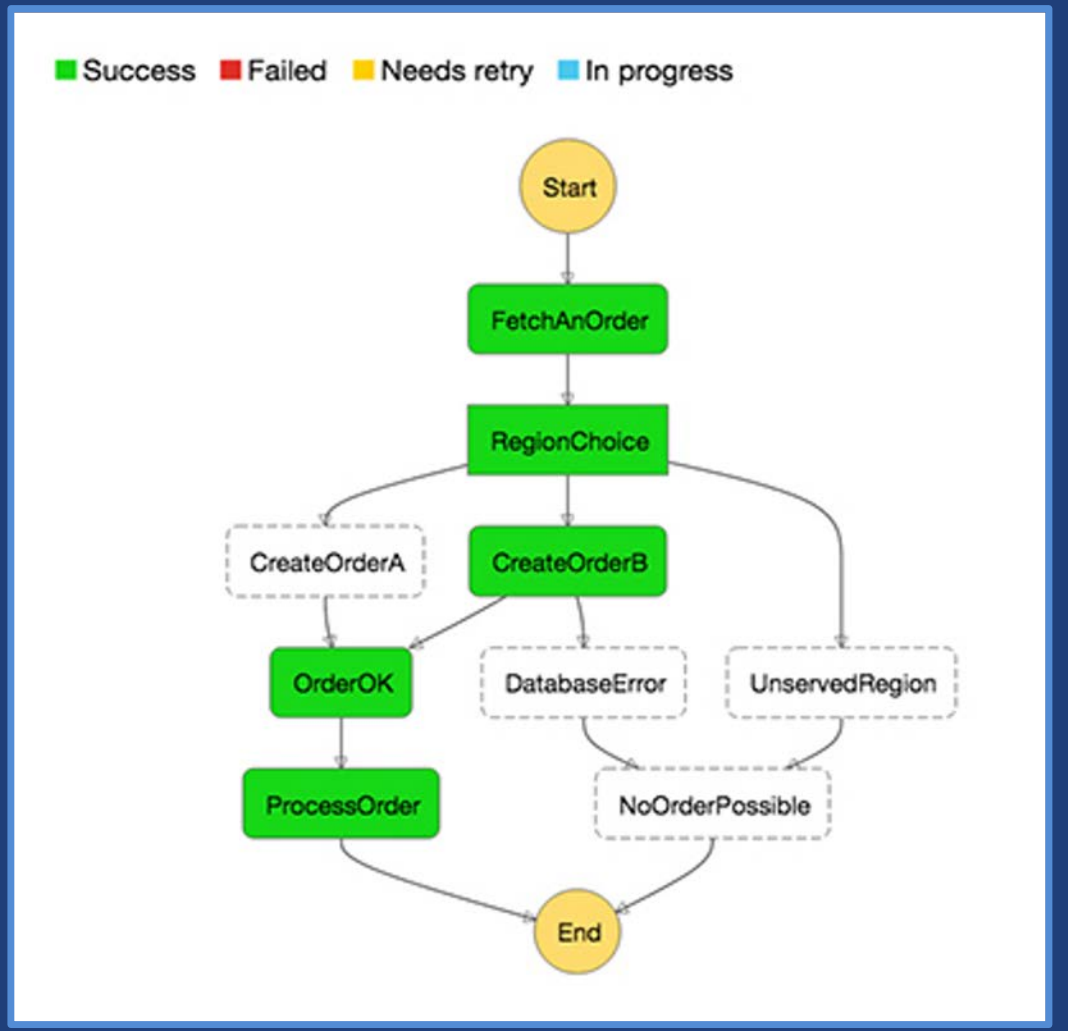
A platform service for defining, applying, and sharing distributed research automation **flows**

- **Triggers** start flows based on subscribed events
- Flows call **Action Providers** to perform tasks



Globus automation architecture

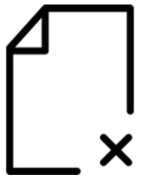
- **Built on AWS Step Functions**
  - JSON-based state machine language
  - Conditions, loops, fault tolerance, etc.
  - Propagates state through the flow
- **Standardized API for integrating custom event and action services**
  - Actions: synchronous or asynchronous
  - Custom Web forms prompt for user input
- **Actions secured with Globus Auth**



Automation Action Providers

Transfer

Delete



ACLs

funcX

DLHub

Identifier

User Form

Notification

Xtract

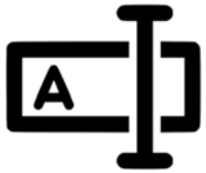
Web Form

Ingest

Expression Evaluation

Search

Describe

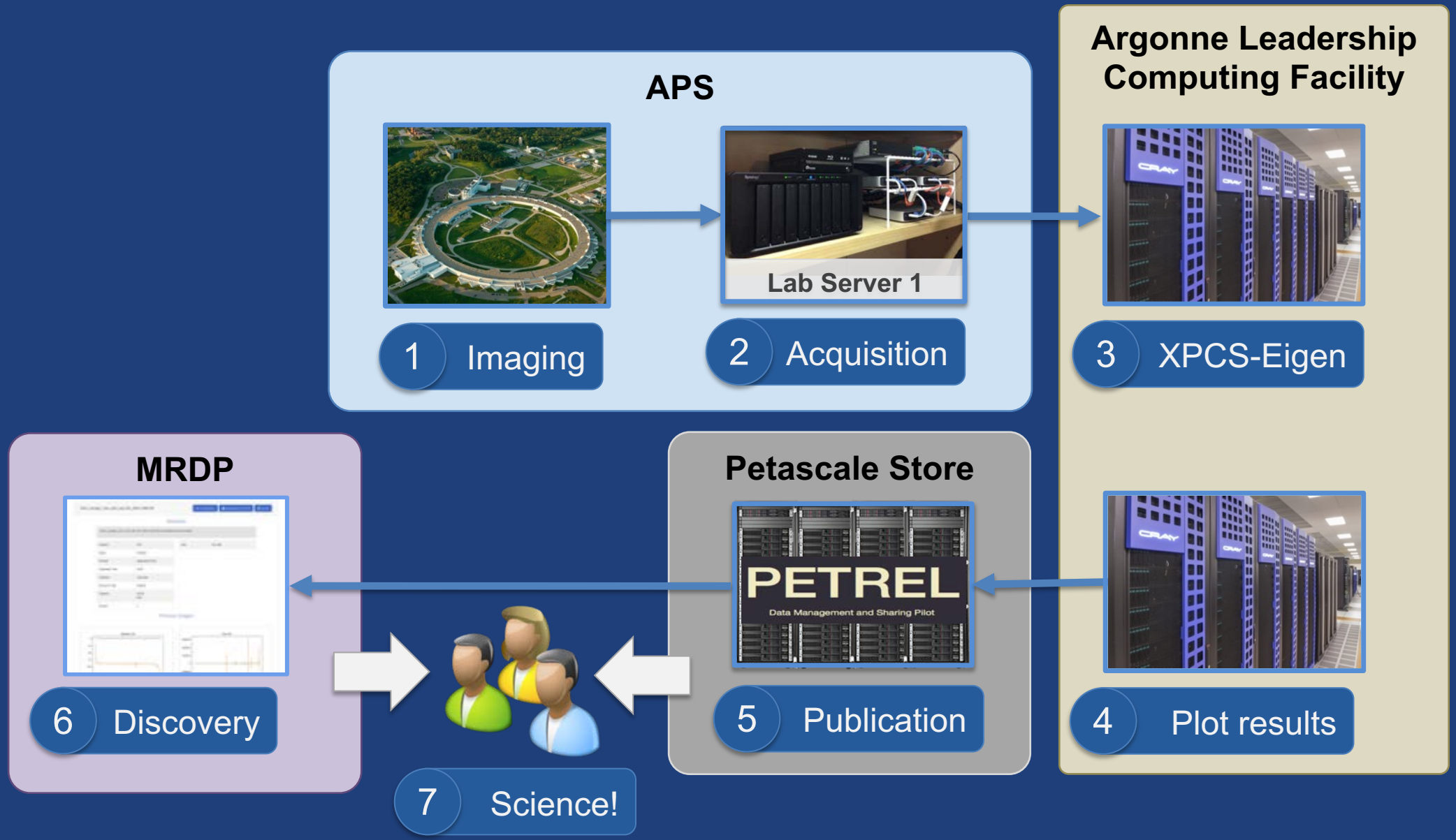


Globus action providers

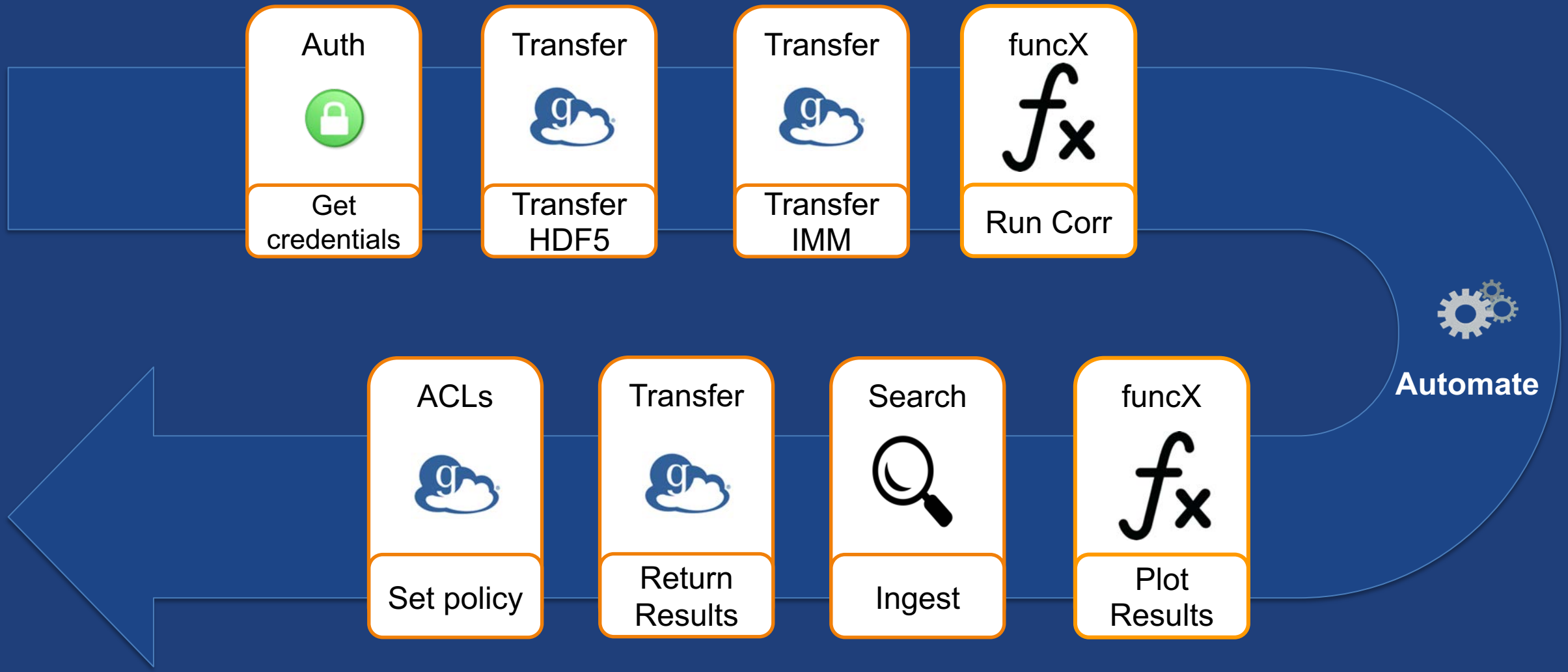
Custom action providers

Automation at the Advanced Photon Source and ALCF

XPCS: X-ray Photon Correlation Spectroscopy



Automating XPCS



Neurocartography

APS

1 Imaging

2 Acquisition

3 Pre-processing

Argonne Leadership Computing Facility

4 Preview/Center

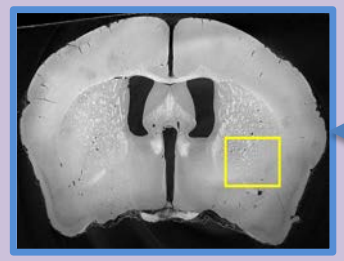
5 User validation

6 Reconstruction

Argonne JLSE

7 Publication

UChicago

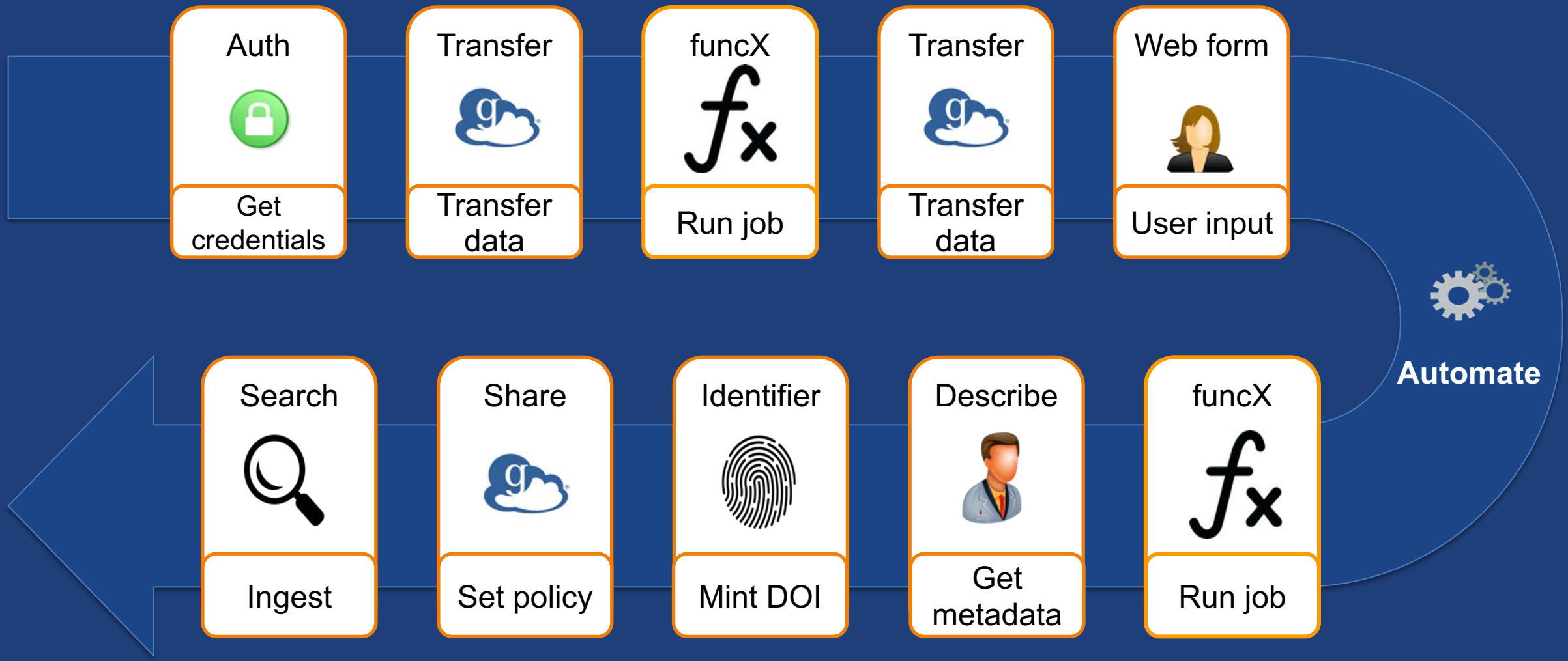


8 Visualization

9 Science!

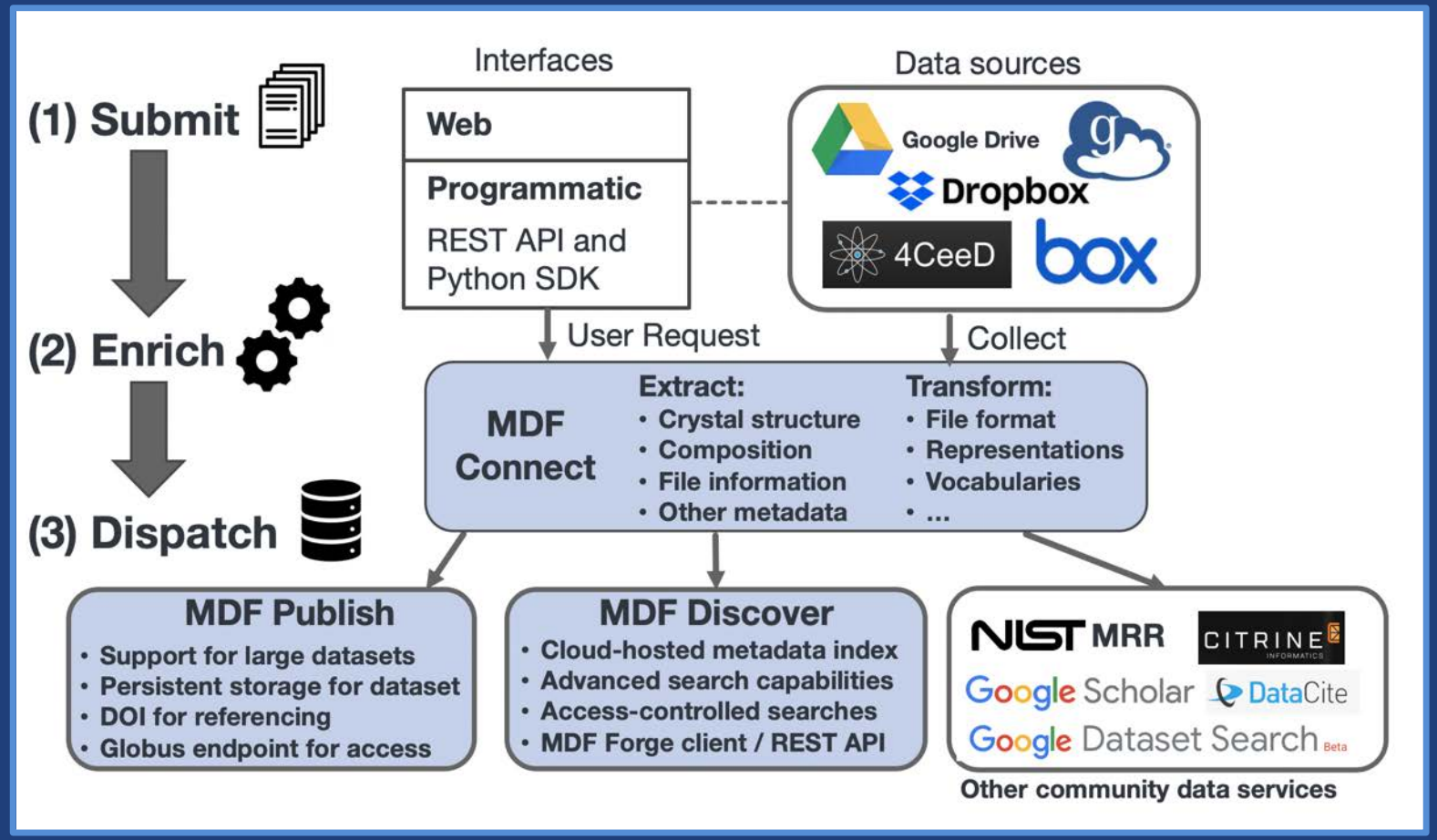
Neuroanatomy reconstruction pipeline

Automating neurocartography



Materials Data Facility

- Accept data from many locations with flexible interfaces
- Index dataset contents in science-aware ways
- Dispatch data to the community
- Using Globus Automate to simplify building of composable flows

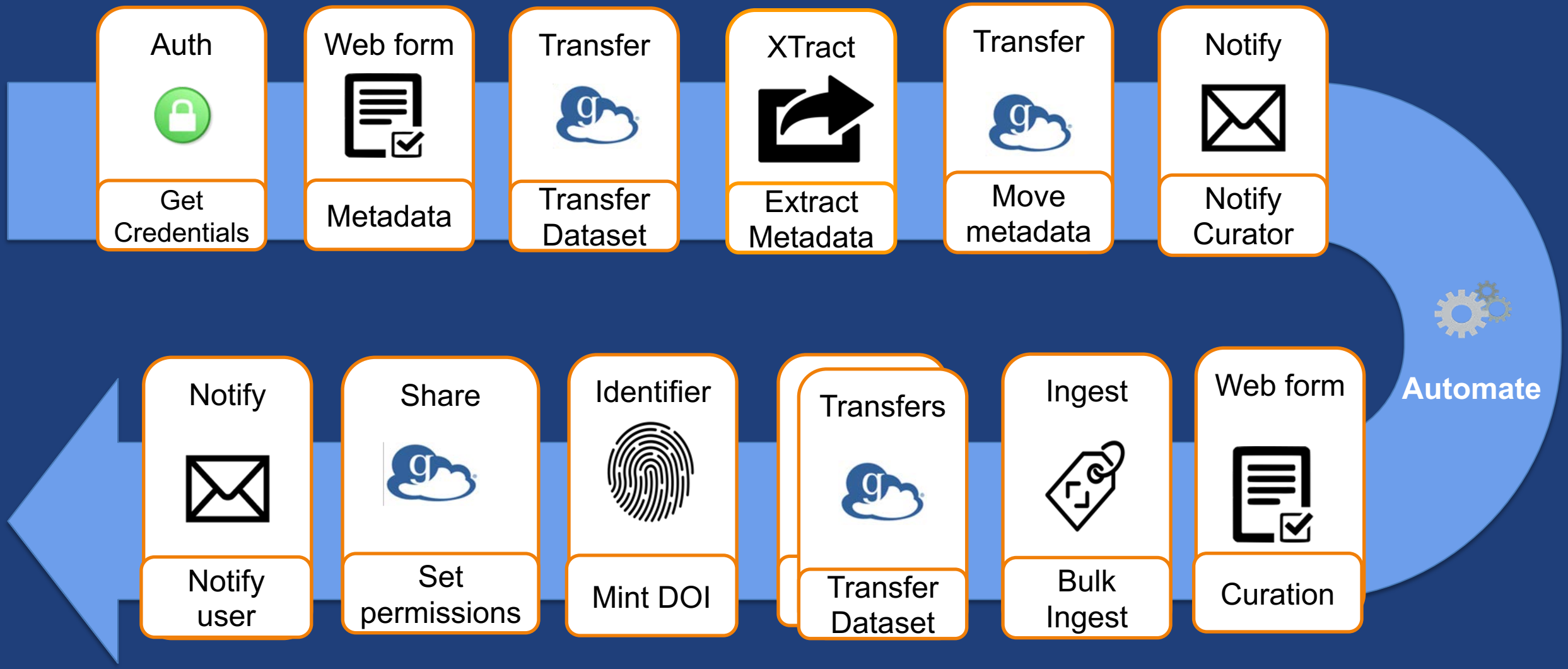


> 35 TB of data

> 400 datasets

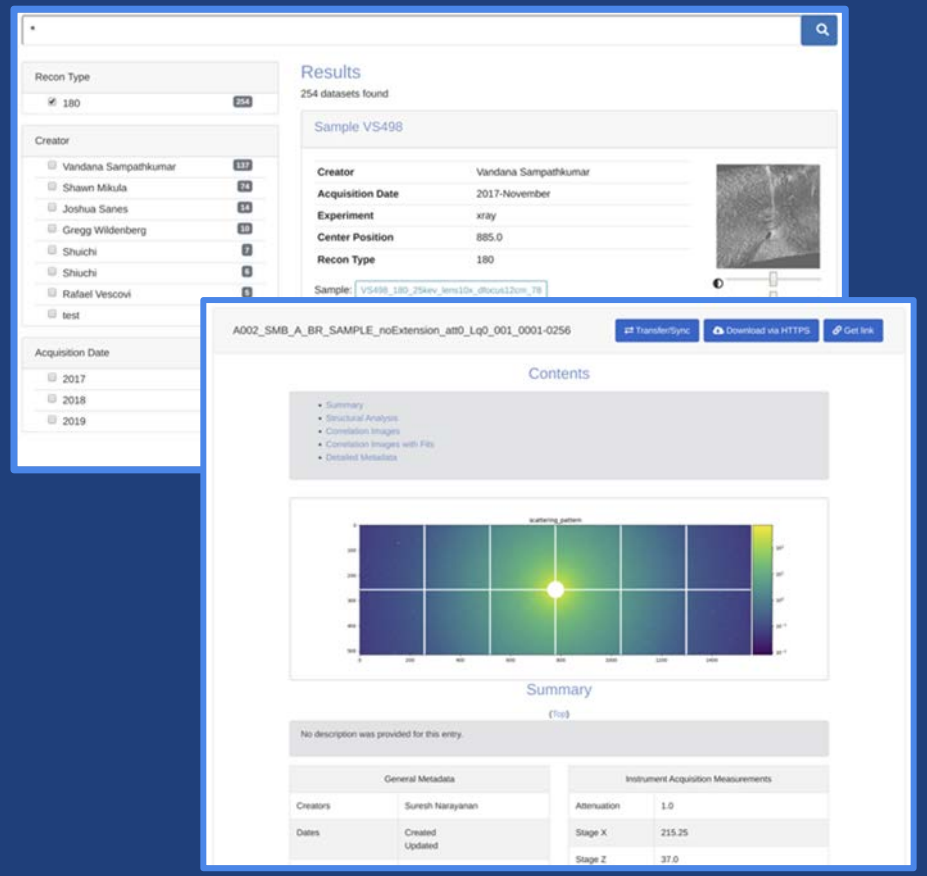
> 320 authors

MDF Data Publication Automation



Petrel data services: petreldata.net

- **Data service providing simple, self-managed project-based data and sharing capabilities**
- **Flexible user-managed search index and discovery portal**
- **Employs the MRDP design pattern**



petreldata.net walkthrough...

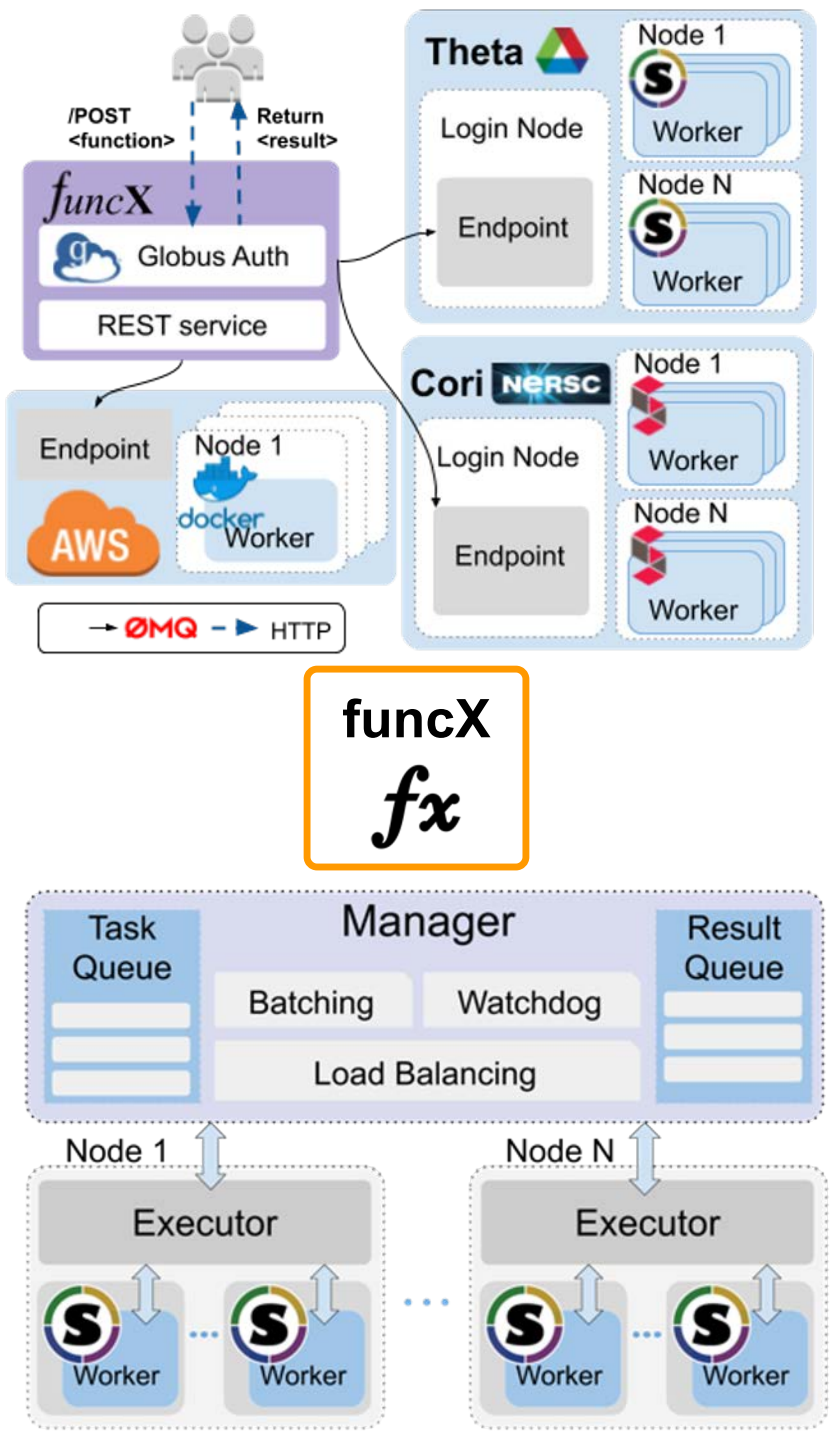
Let’s see what’s in the lab…

funcX action provider

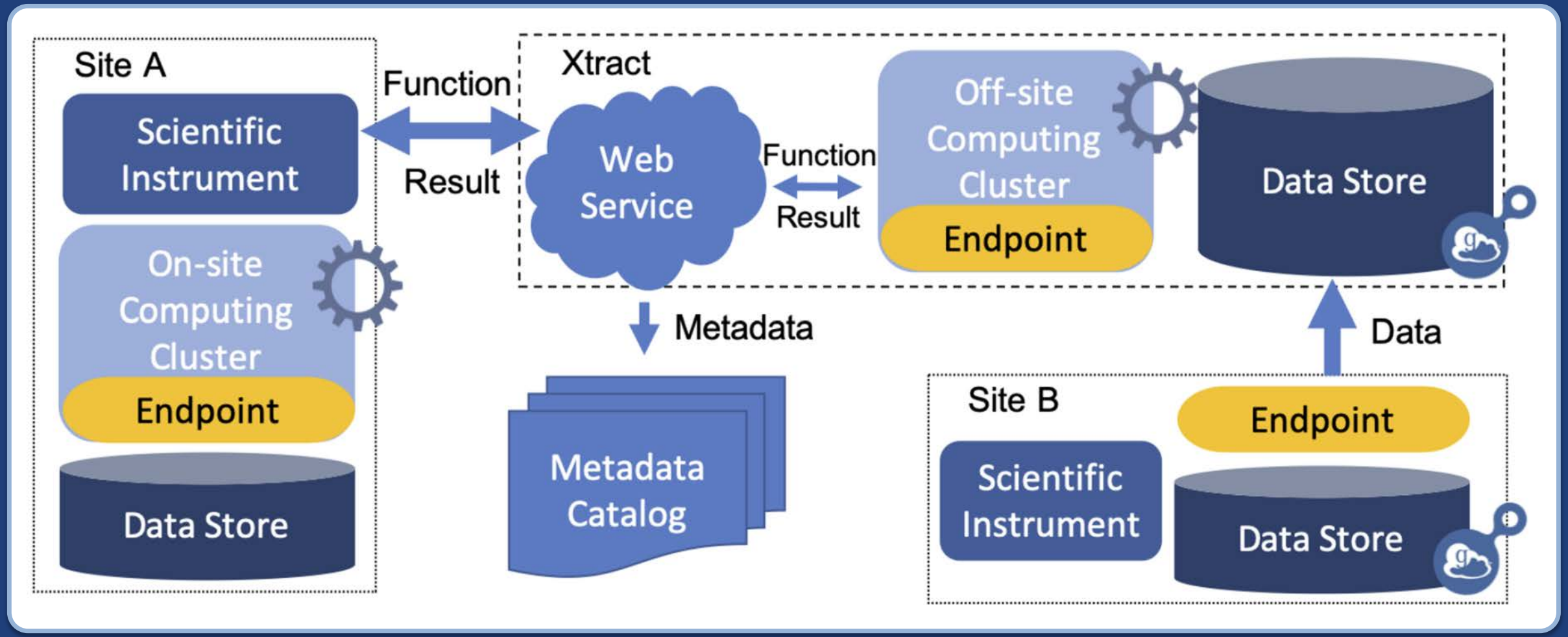
funcX: FaaS platform for HPC

- funcX endpoints deployed at resources
- Service routes requests to endpoints
- Parsl acquires resources
- Singularity containers run functions
- Globus Auth secures communication

globus labs



Xtract: Rich metadata on-demand



globus labs

Globus is …

a non-profit, production grade 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

Thank you, funders...

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

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

Argonne
NATIONAL LABORATORY

Thank you, subscribers!

app.globus.org

docs.globus.org

globus.org/**connectors**

globus.org/**subscriptions**

outreach@globus.org

support@globus.org