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

Kyle Chard
University of Chicago
Argonne National Laboratory

chard@uchicago.edu

THE UNIVERSITY OF CHICAGO

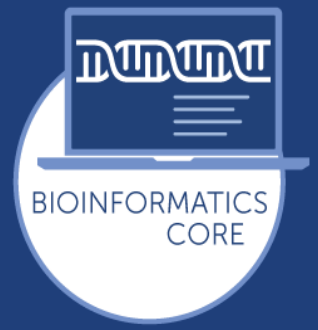
Argonne NATIONAL LABORATORY

globus

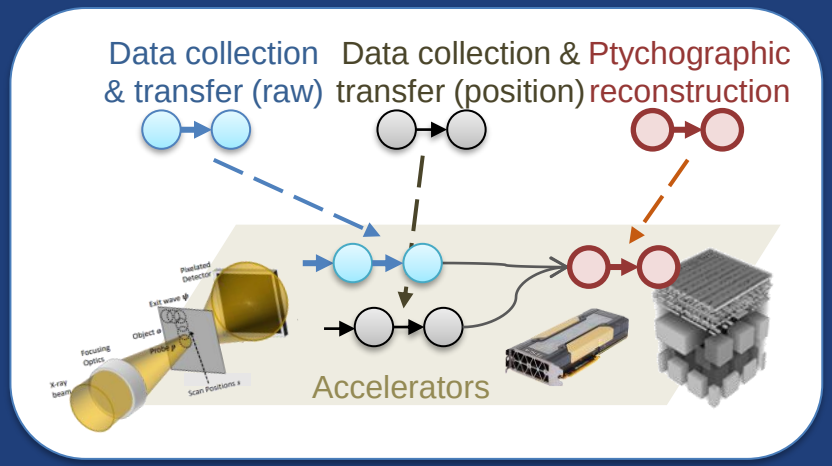
Towards self-driving instruments

- **Exponential growth in the rate that instruments perform measurements**
  - Data generated at GB+/s and 100+TB/day
- **Analysis requires significant online computing capacity**
  - HPC resources, GPUs, AI accelerators
- **Computation (and AI) can steer experiments**
  - Workflows involve both humans and machines

→ We need new methods to automate these workflows and coordinate actions and resources across experiment and compute environments



Example instrument patterns

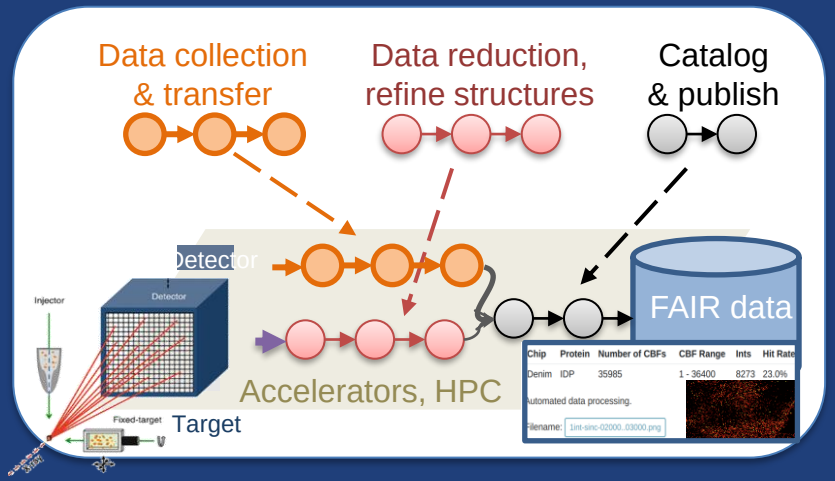


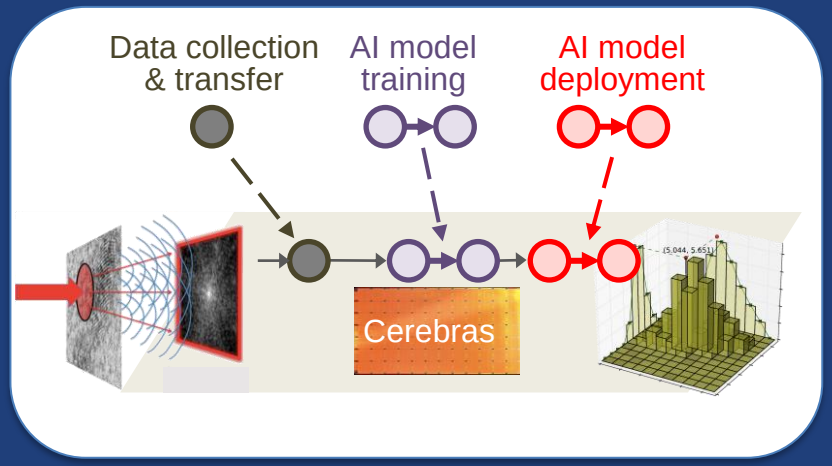
Ptychography

Reconstruction algorithms applied to data as they are acquired

Serial Crystallography

Data are reduced and made available to scientists and collaborators



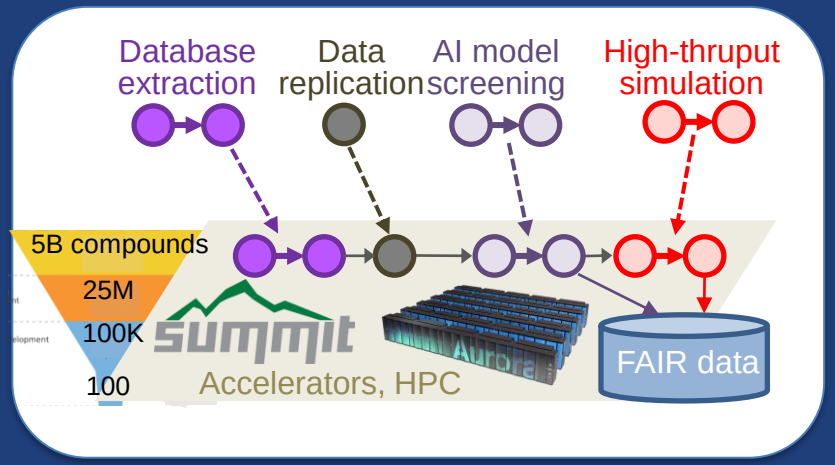


HEDM

Models trained on acquired data and deployed at the edge for fast inference

COVID

Data replicated across sites to apply AI screening methods and AI-guided simulation



These patterns highlight diverse automation needs

- **Support various *actions***
  - Transfer, compute, ingest in a search index, associate a persistent identifier, modify access permissions
- **Robust orchestration spanning several locations**
  - Enable remote control of actions in different places
- **Authentication/authorization model to provide secure management of remote operations across the computing continuum**

Globus automation capabilities

Timer Service

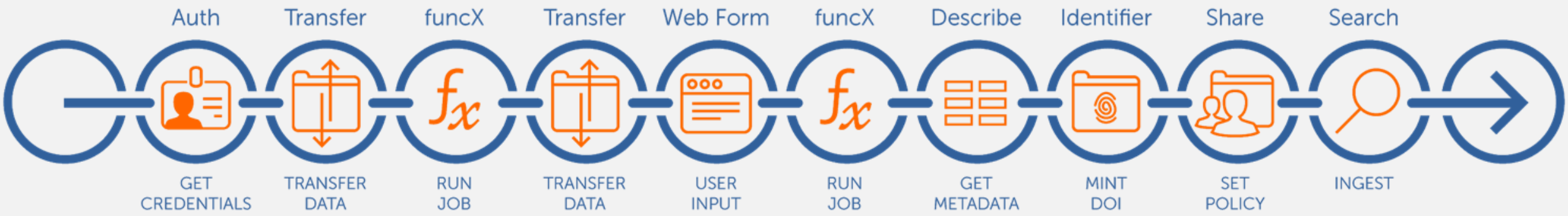
Scheduled and recurring transfers (*a.k.a. Globus cron*)

Globus Flows service

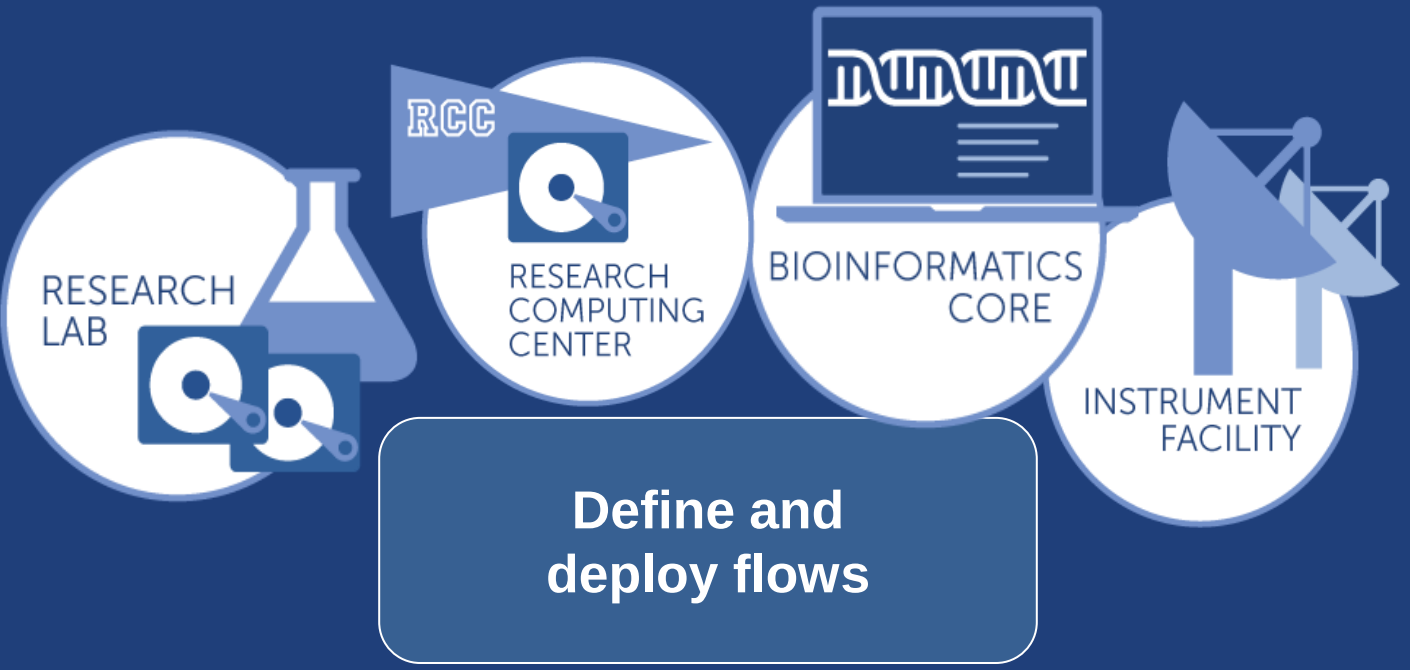
Comprehensive task (data and compute) orchestration with human in the loop interactions

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



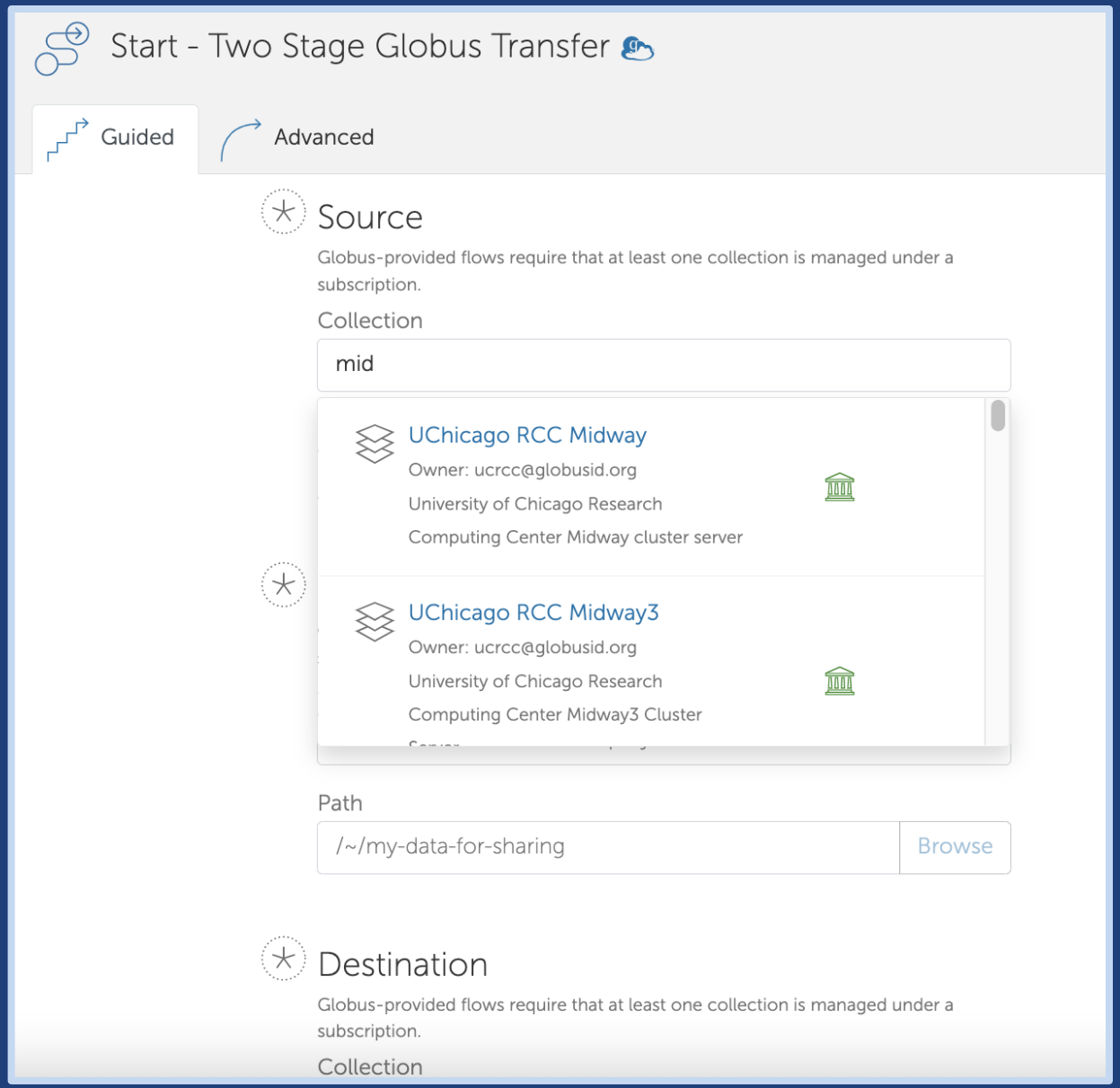
Create and deploy flows

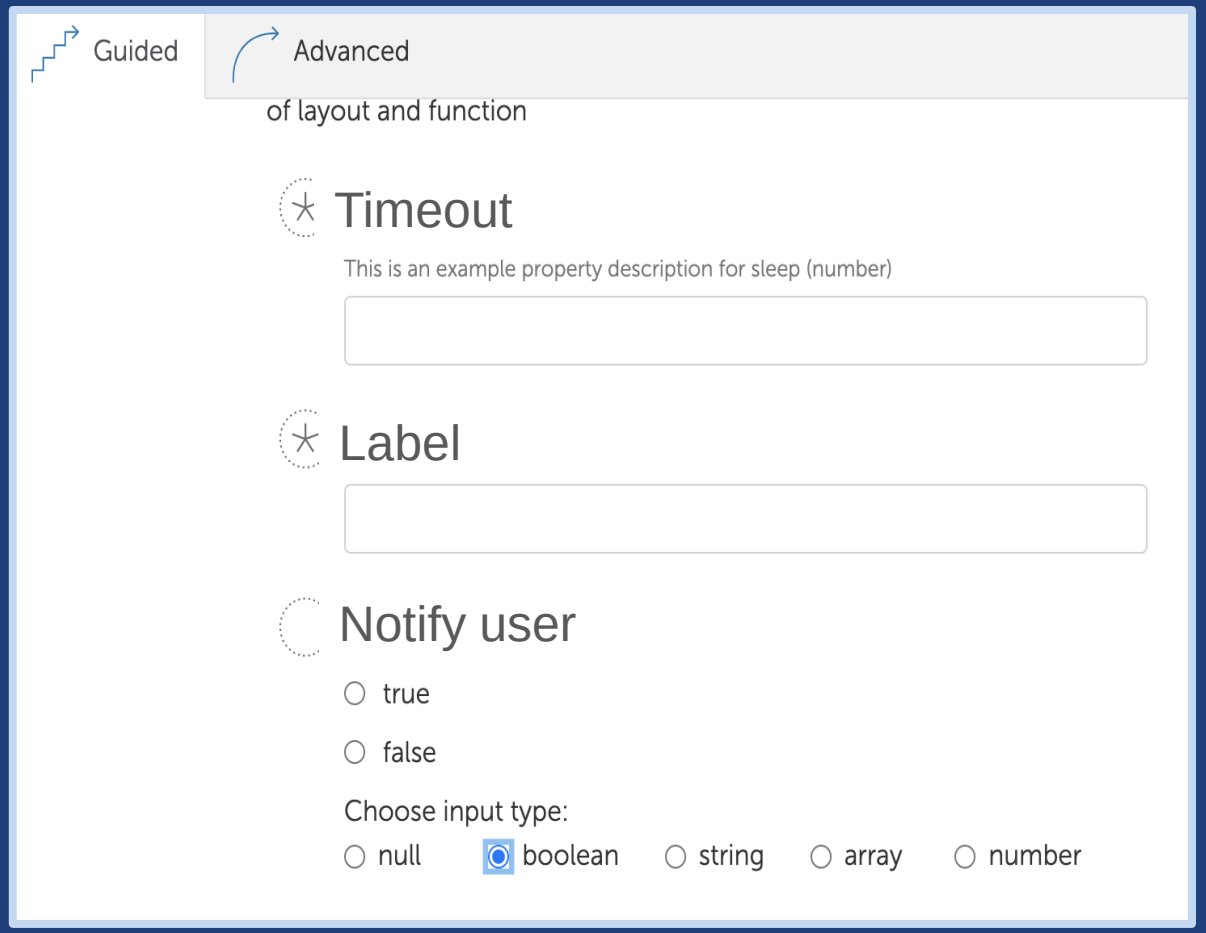


- **Use declarative language (JSON or YAML)**
- **Set input schema**
- **Set policy for use of the flow**

```
{
  "States": {
    "SetPermission": {
      "End": true,
      "Type": "Action",
      "Comment": "Grant read permission on the data to a Globus user or group",
      "ActionUrl": "https://actions.automate.globus.org/transfer/set_permission",
      "Parameters": {
        "path.$": "$.input.destination.path",
        "operation": "CREATE",
        "permissions": "r",
        "principal.$": "$.input.principal_identifier",
        "endpoint_id.$": "$.input.destination.id",
        "principal_type.$": "$.input.principal_type"
      },
      "ResultPath": "$.SetPermission"
    },
    "TransferFiles": {
      "Next": "SetPermission",
      "Type": "Action",
      "Comment": "Transfer to a guest collection",
      "WaitTime": 60,
      "ActionUrl": "https://actions.automate.globus.org/transfer/transfer",
      "Parameters": {
        "transfer_items": [
          {
            "recursive.$": "$.input.recursive_tx",
            "source_path.$": "$.input.source.path",
            "destination_path.$": "$.input.destination.path"
          }
        ],
        "source_endpoint_id.$": "$.input.source.id",
        "destination_endpoint_id.$": "$.input.destination.id"
      },
      "ResultPath": "$.TransferFiles"
    }
  },
  "Comment": "Transfer files to a guest collection and set access permissions",
  "StartAt": "TransferFiles"
}
```

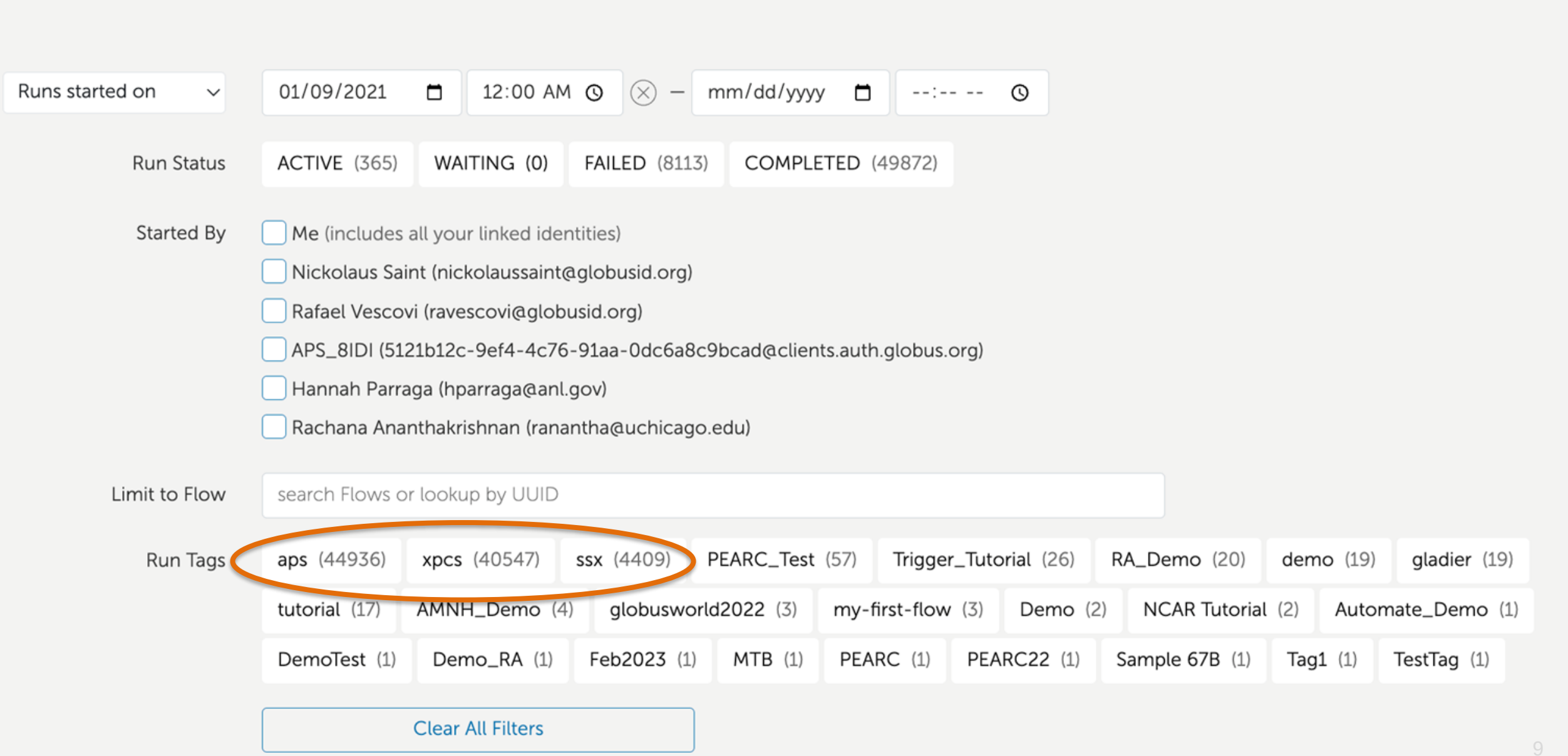
Start flows: Guided input





Dynamic forms generated from input schema

Managing runs at scale

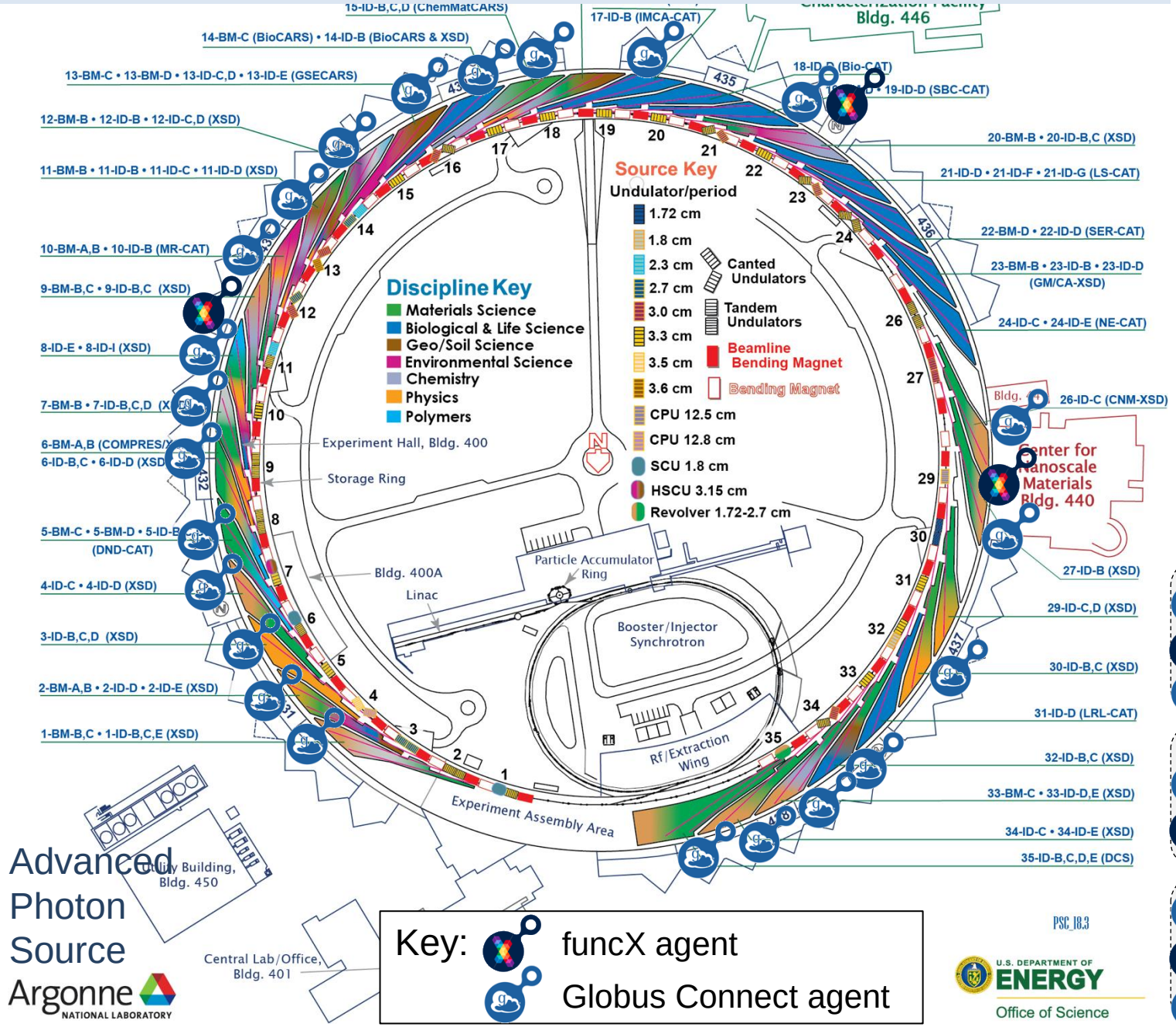


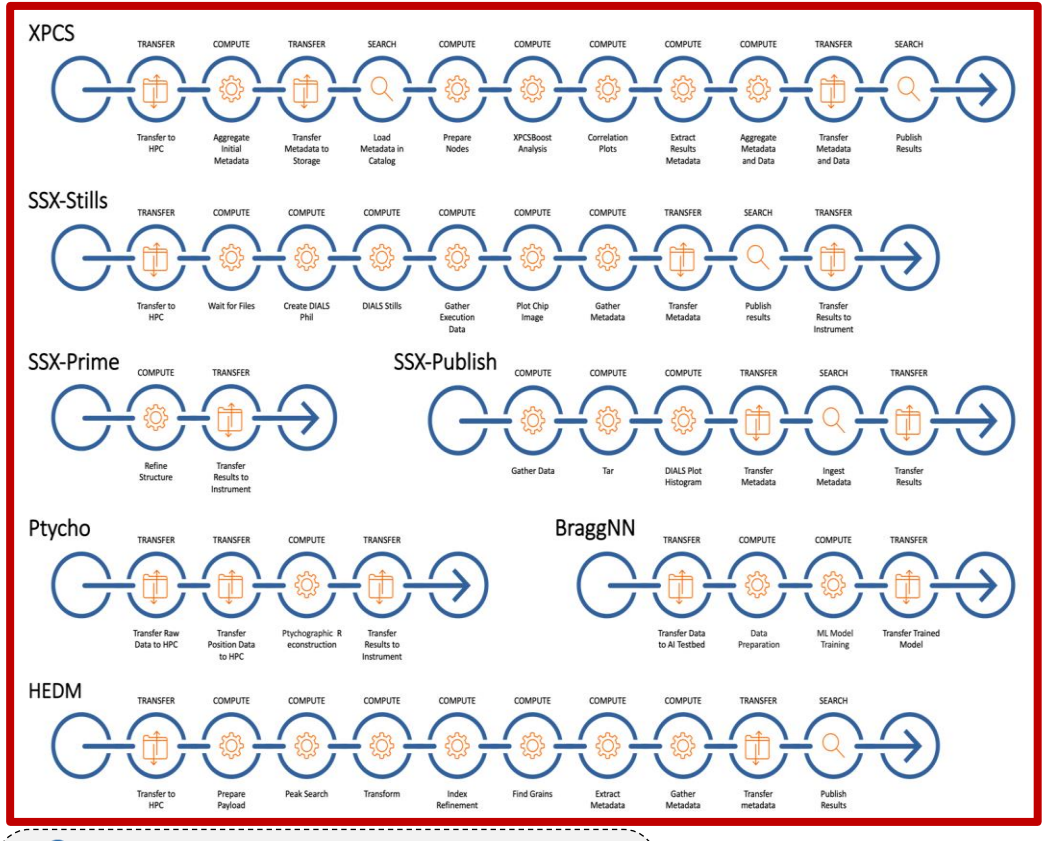
Running flows across the computing continuum requires a universal data and compute fabric

Globus Auth:

standards compliant identity and access management platform

Endpoints at experimental facilities





Polaris

Theta

Eagle store

Argonne Leadership Computing Facility

Bebop Cluster

Laboratory Computing Research Center

Orthros Cluster

APS DM system

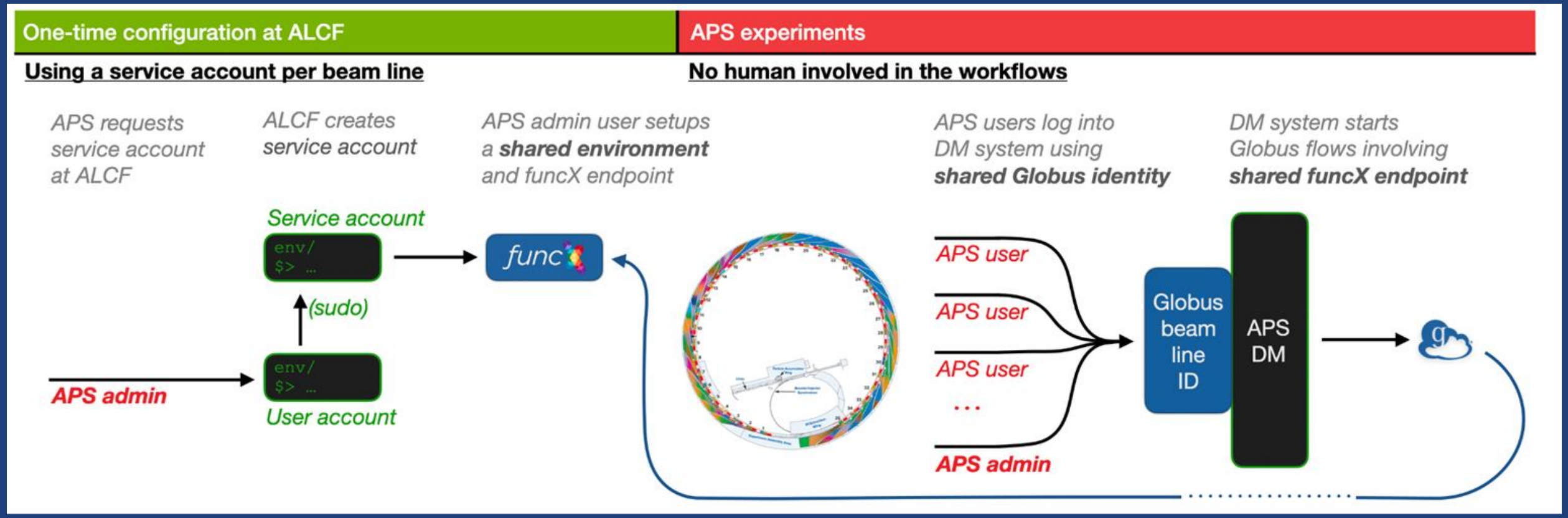
APS Computing

Globus-accessible storage and computing (10,000s of systems)

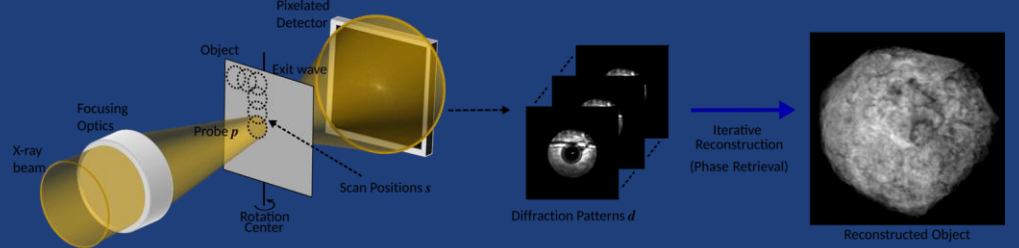
Portal server

amazon web services

Integrating APS and ALCF

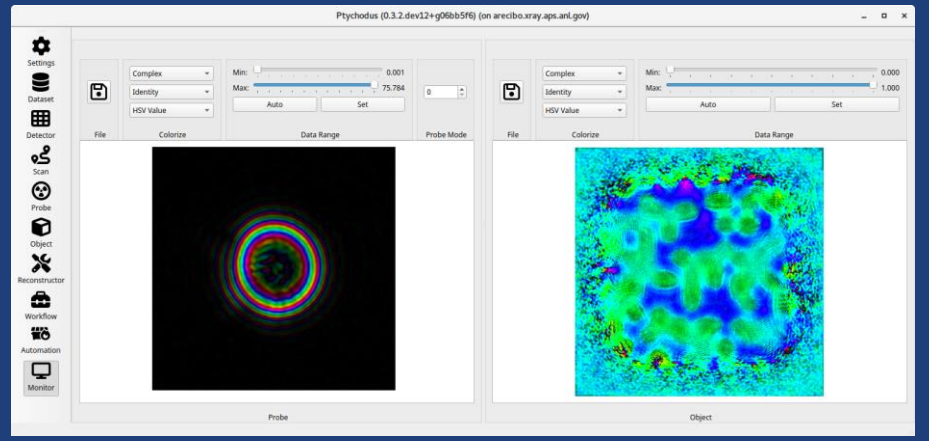


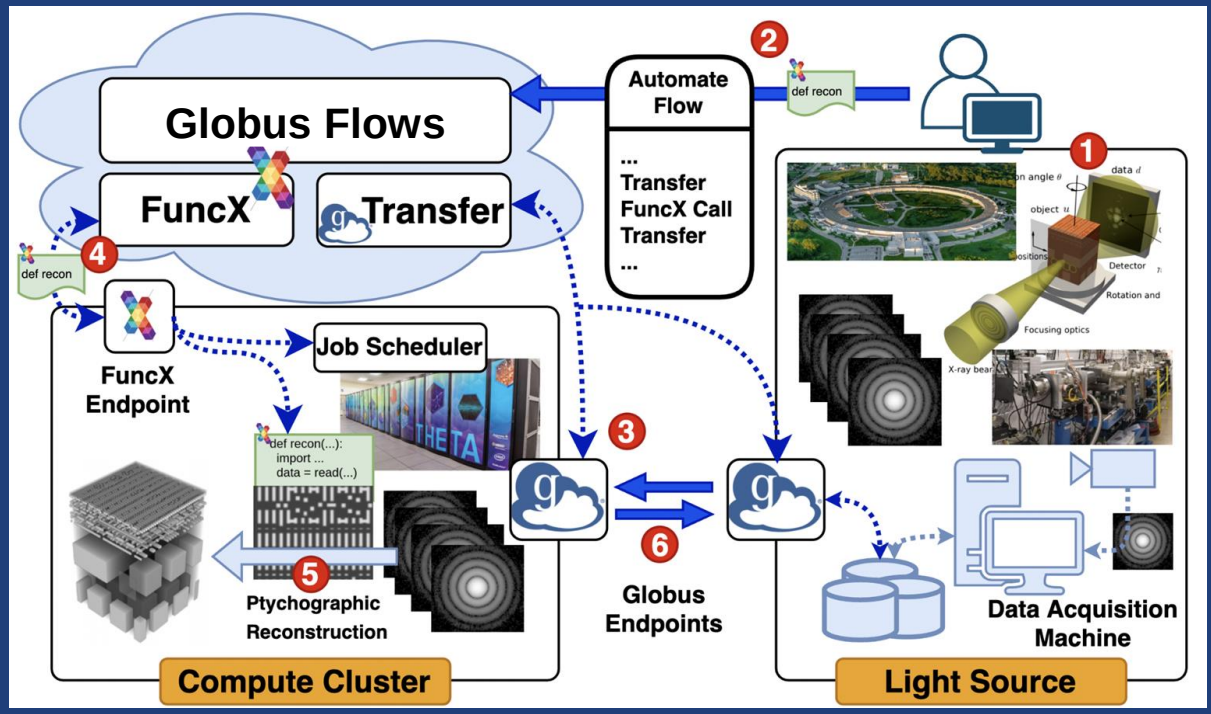
Ptychography at 26ID



- **Ptychography is a computational microscopy technique for reconstructing the complex-valued transmission function of an object**
- **Diffraction patterns are recorded at many overlapping scan positions**
- **Flow:**
  - Ptychodus monitors the local filesystem to trigger flows as data are collected
  - Scans are reconstructed at ALCF using on-demand queue
  - Results returned to APS where Ptychodus loads them for visualization
  - Users can customize where the compute needs to run via flow configuration

Full automation using service accounts; on-demand queue for timely runs

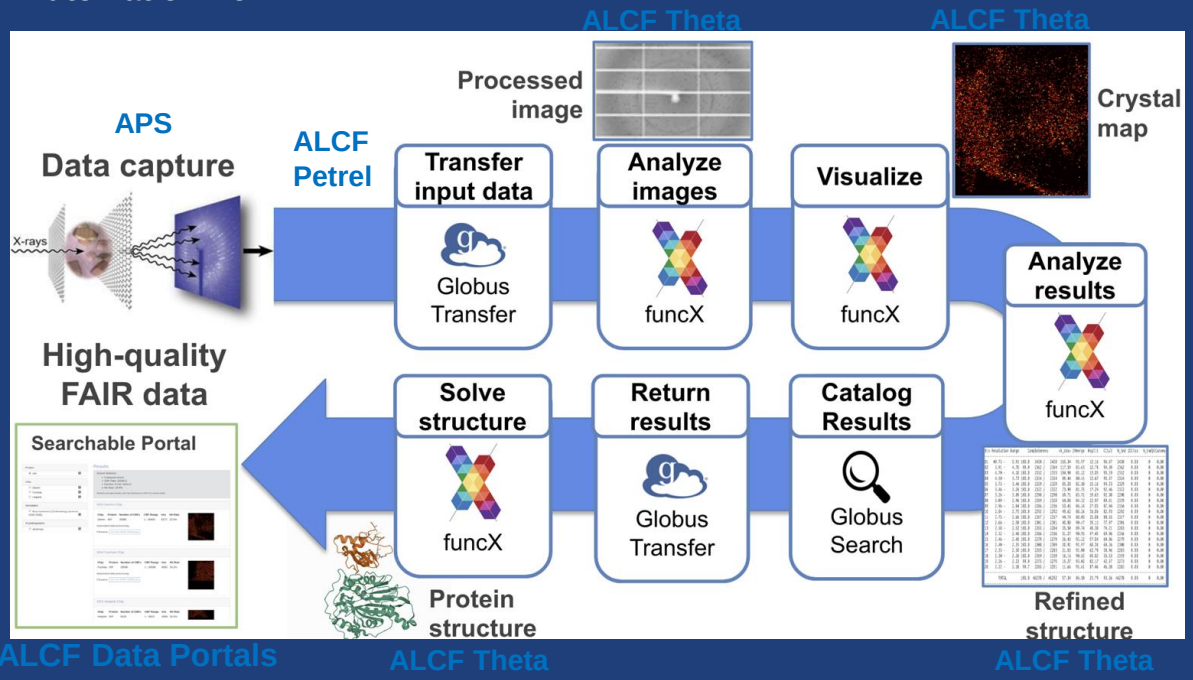


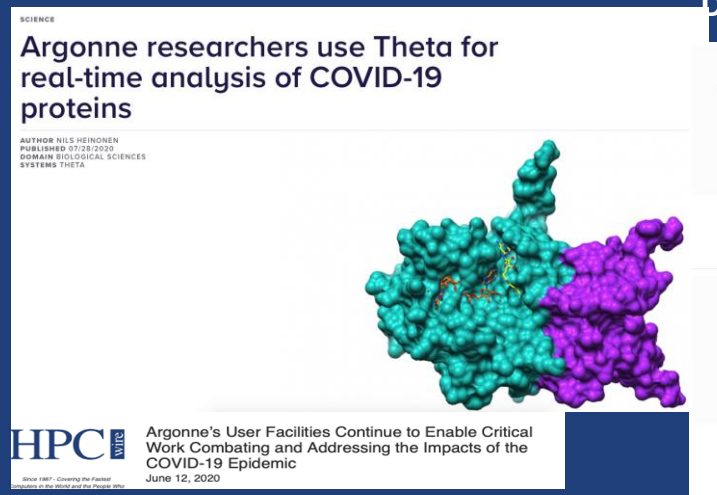


Solving Protein Structures an Order of Magnitude Faster

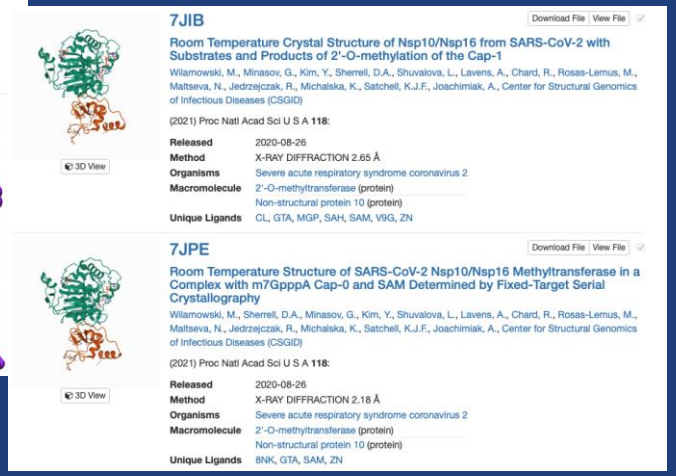
- **Flow:** collect data, analyze and visualize the data, solve protein structure and load results into a searchable portal for real-time feedback
- **Achieved >order of magnitude speed up** in time to solution of protein structures at APS
- **Leveraged unique DOE facilities** at Advanced Photon Source (SBC Sector 19) and ALCF (Theta/ ThetaGPU, Petrel, and Data Portals)

Automation flow





Deposited results in open repositories

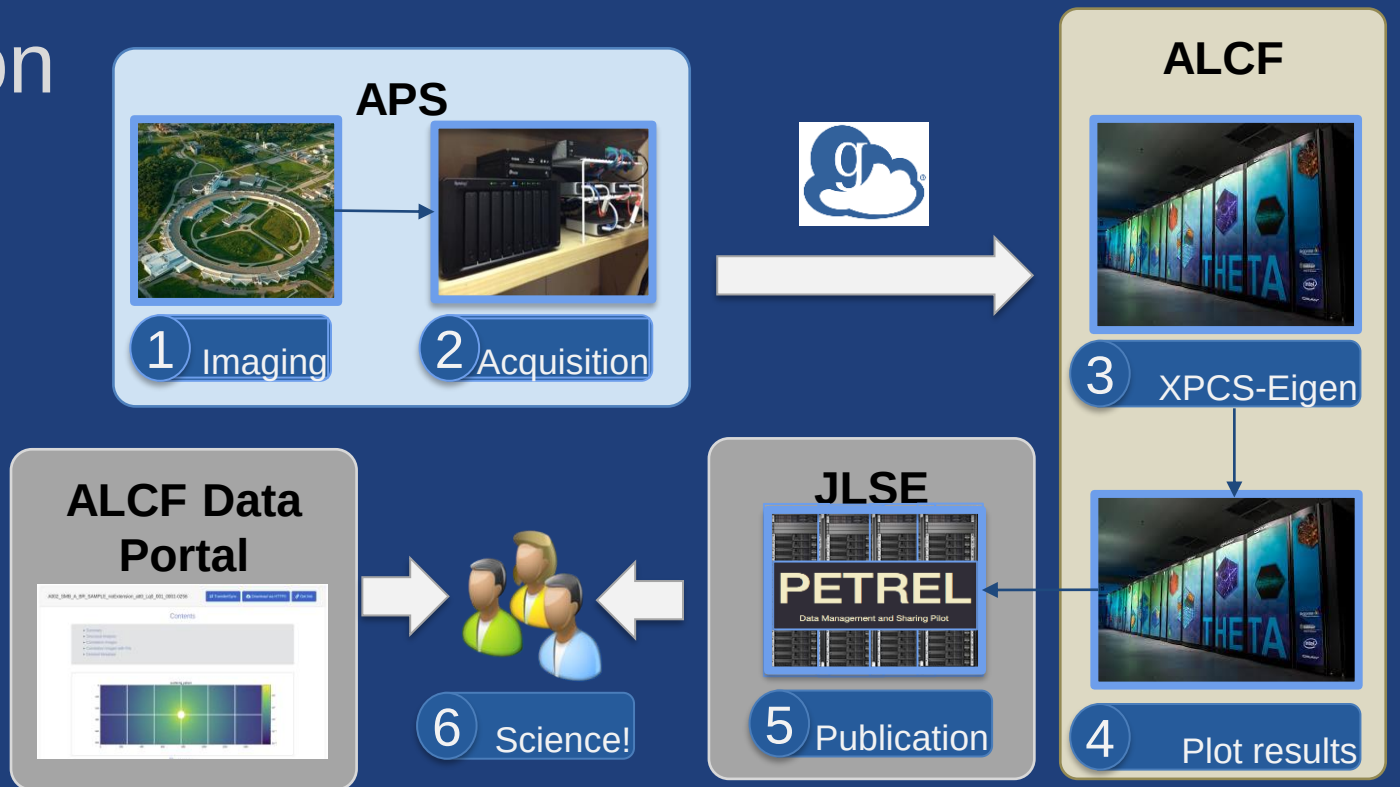


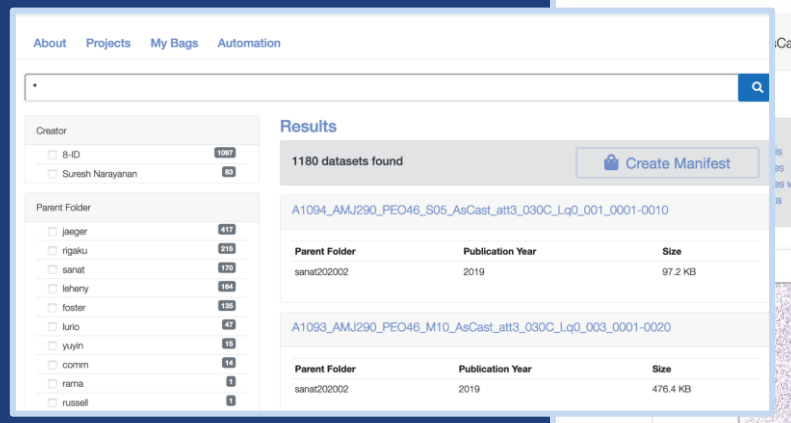
> *"These data services have taken the time to solve a structure from weeks to days and now to hours"*
>
> Darren Sherrell, SBC beamline scientist APS Sector 19

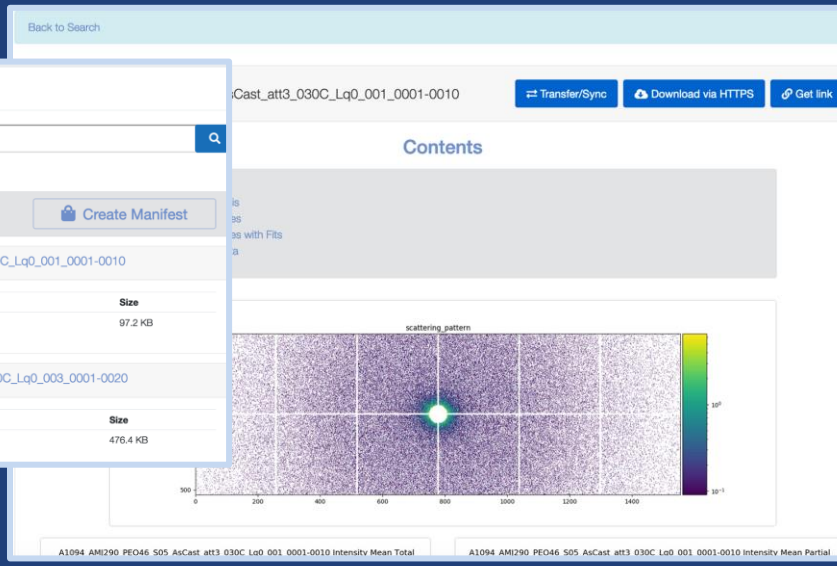
(R. Chard, Vescovi, Foster, Blaiszik, Sherrell, Joachimiak, et al.)

X-ray Photon Correlation Spectroscopy (XPCS)

- **XPCS studies dynamical properties of materials by recording speckle patterns over time, constructing a time correlation function, and measuring processes of interest (e.g., diffusion)**
- **Flows:**
  - Integrated with APS's Data Management system to automatically invoke flows
  - Data analyzed and published to a searchable ALCF portal
  - HTTPS-enabled portals to visualize results
  - Reprocessing capabilities in portal to invoke flows

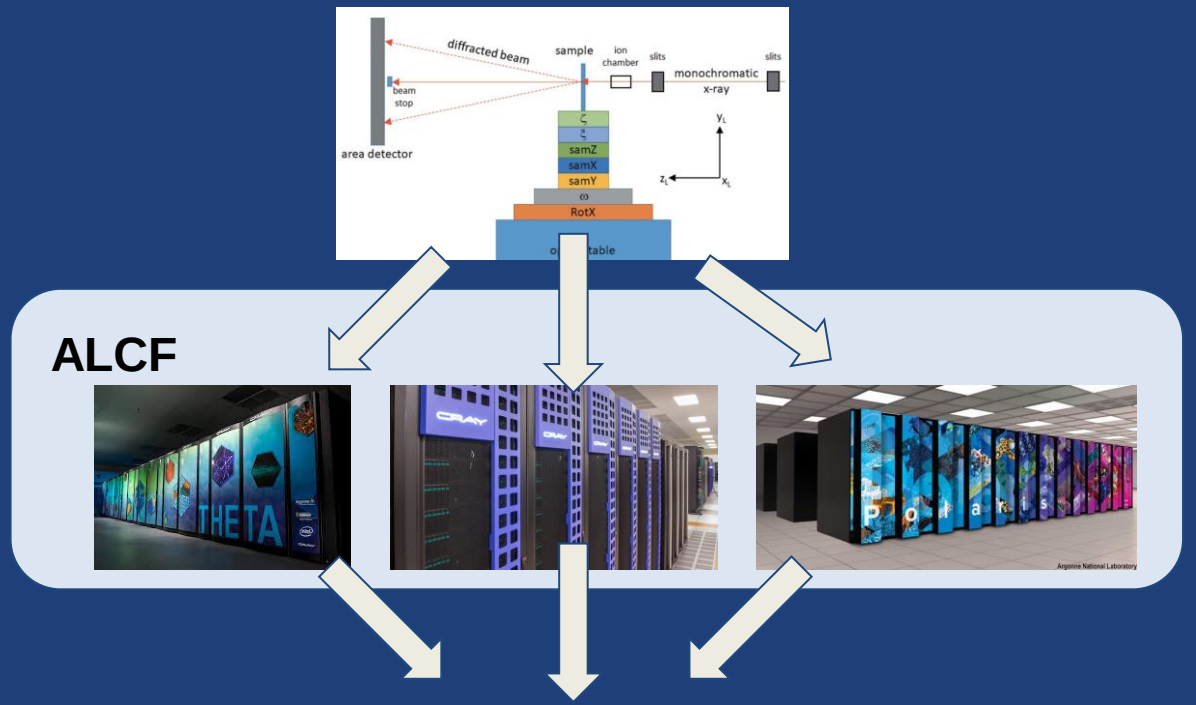


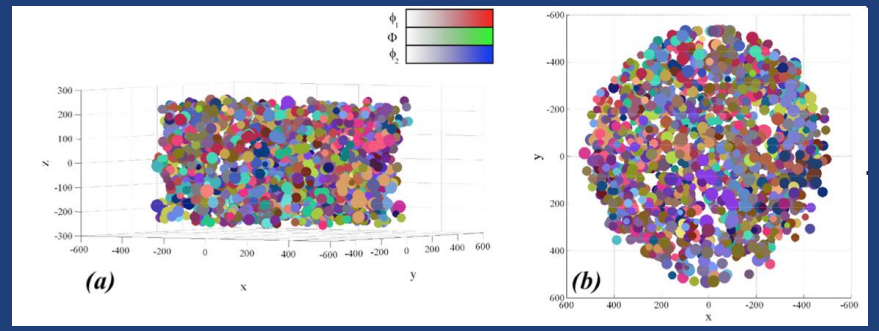




High-Energy X-ray Diffraction Microscopy (HEDM)

- **HEDM combines imaging and crystallography to characterize polycrystalline microstructure in 3D under various in situ thermo-mechanical conditions**
- **Flow**
  - Select analysis to run at APS Orthros, ALCF ThetaGPU/Cooley
  - Globus Transfer data
  - Deploy containers with MIDAS software to perform tasks
  - Results assembled and returned to APS user
  - Mechanism for users to run analysis at home institute

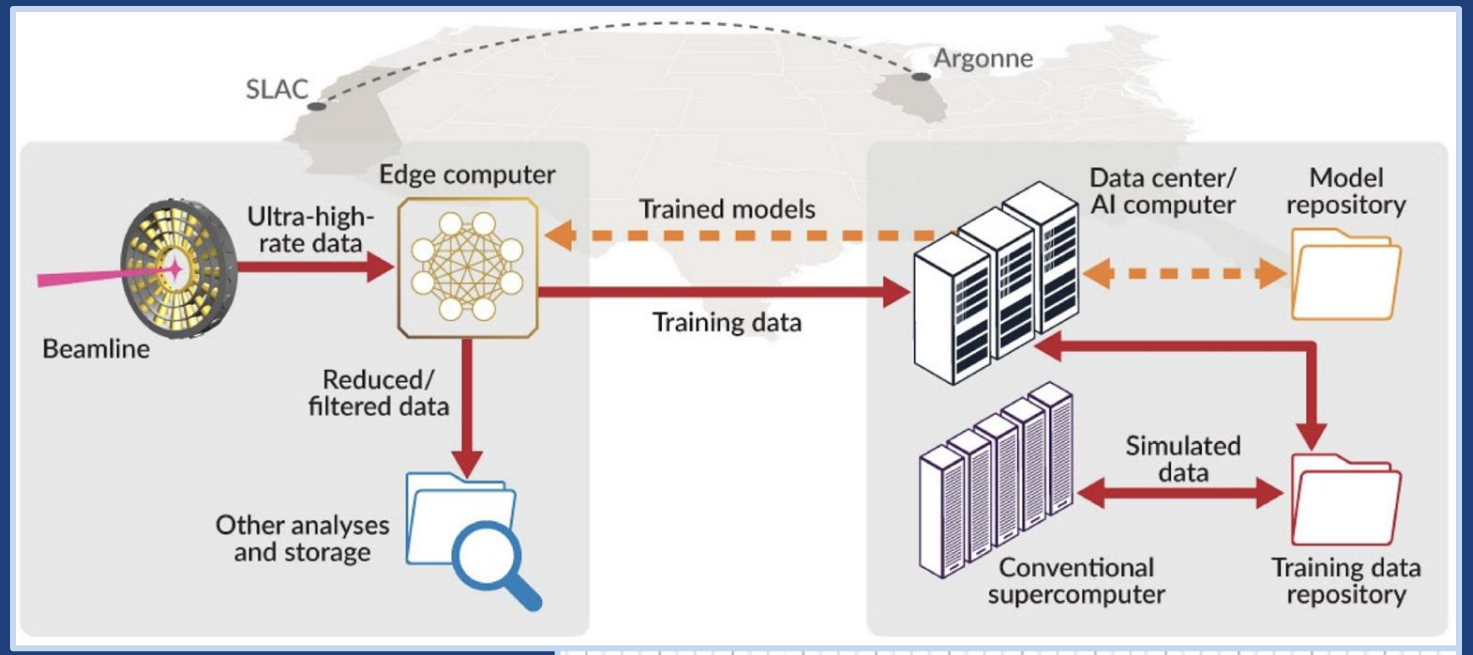


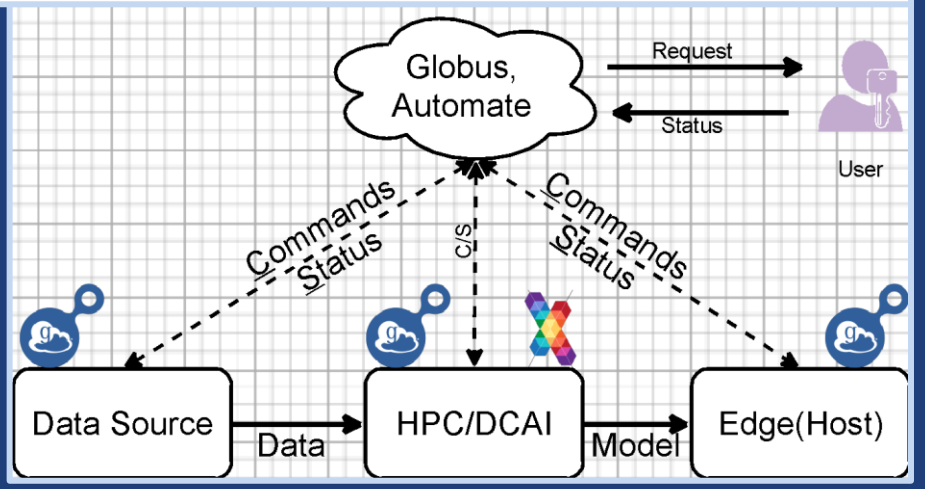


Hemant Sharma, et al.

Rapid Training of Deep Neural Networks using Remote Resources

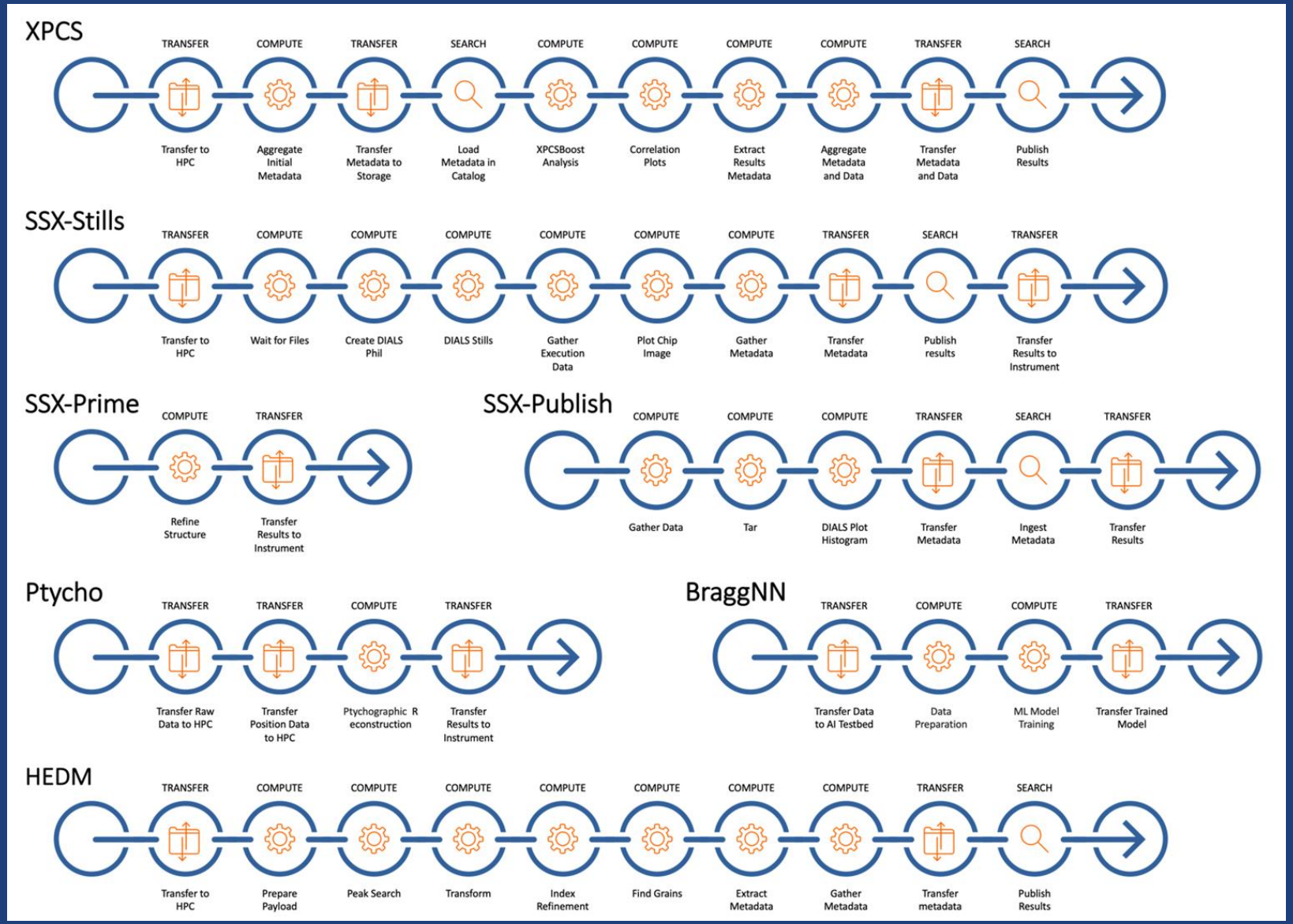
- **HEDM workflow that deploys BragNN at the edge for real-time diffraction peak analysis (e.g., for experiment steering and anomaly detection)**
- **Tight coupling with simulation and training with real-time data**
- **Flow:**
  - Globus to rapidly move data for training
  - funcX for simulation and model training
  - Globus to move models to the edge





Zhengchun Liu, Jana Thayar, et al.

Production flows linking instruments and computation



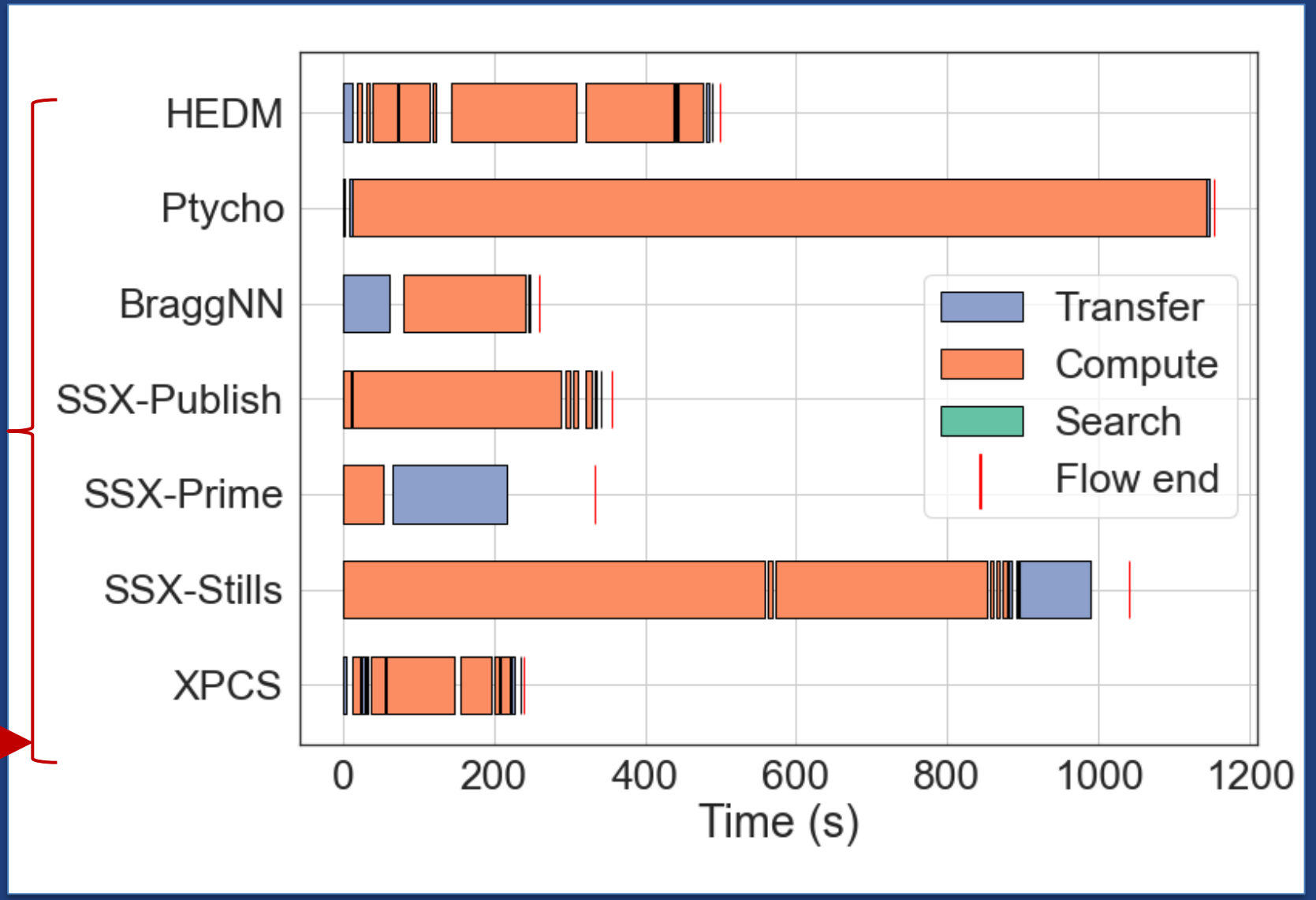
R. Vescovi et al.,
https://doi.org/10.1016/j.patter.2022.100606

Flows span spatial and temporal ranges

Reliable flow orchestration across resources

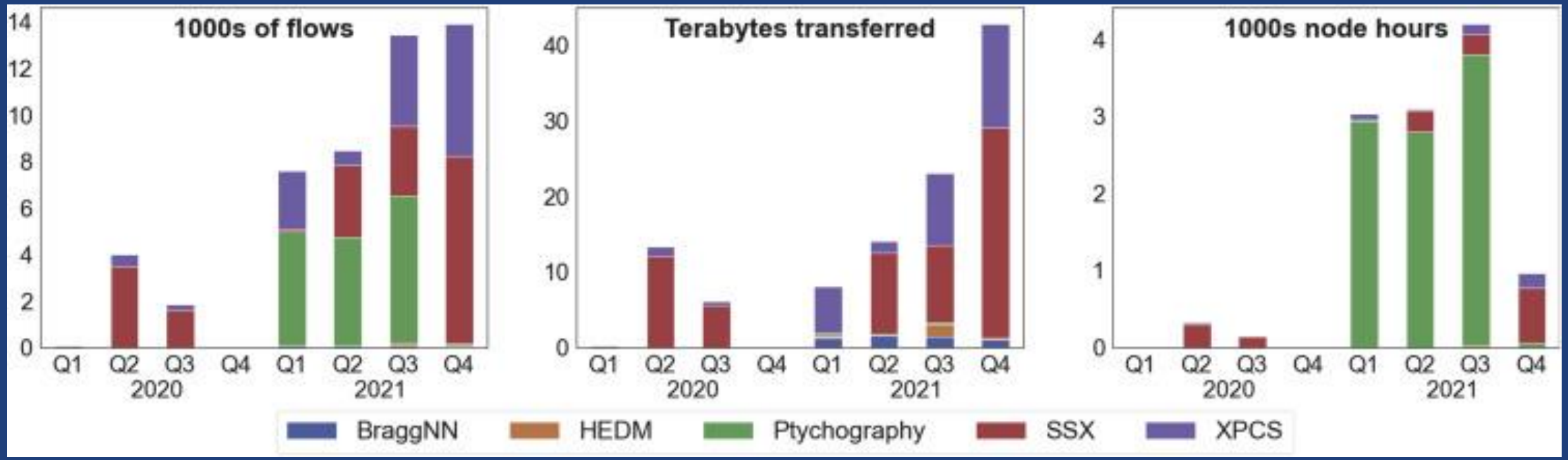
Functions executed in various locations: at a beamline, local server, cluster, cloud

Execution times at the Argonne Leadership Computing Facility



R. Vescovi et al., https://doi.org/10.1016/j.patter.2022.100606

Flows are increasingly critical to APS science



R. Vescovi et al.,
https://doi.org/10.1016/j.patter.2022.100606

Globus documentation: docs.globus.org

YouTube: youtube.com/GlobusOnline

Helpdesk: support@globus.org

Patterns

CellPress
OPEN ACCESS

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

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 subscribers