

Gladier: An Architecture to Enable Modular Automation of Data Capture, Storage, and Analysis at Experimental Facilities

Ryan Chard, Nickolaus Saint, David Kelly, Rachana Ananthakrishnan, Kyle Chard, Tyler Skluzacek, Rick Wagner, Suresh Narayanan, Darren Sherrell, Nicholas Schwarz, Ben Blaiszik, Ian Foster

Contact: bblaiszik@anl.gov, foster@anl.gov

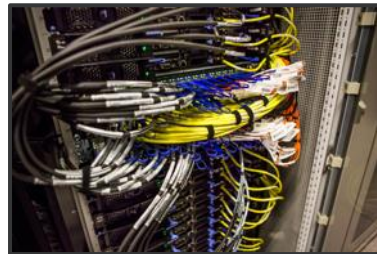
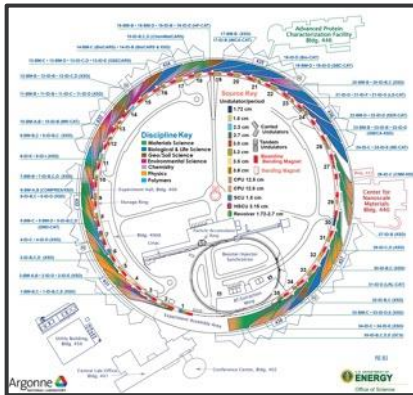
Problem Statement

- *Data rates and heterogeneity increasing*
- *A continuum of computing resources are available*
- *Differing workflows across beamlines*

Computing Continuum



Argonne Advanced Photon Source

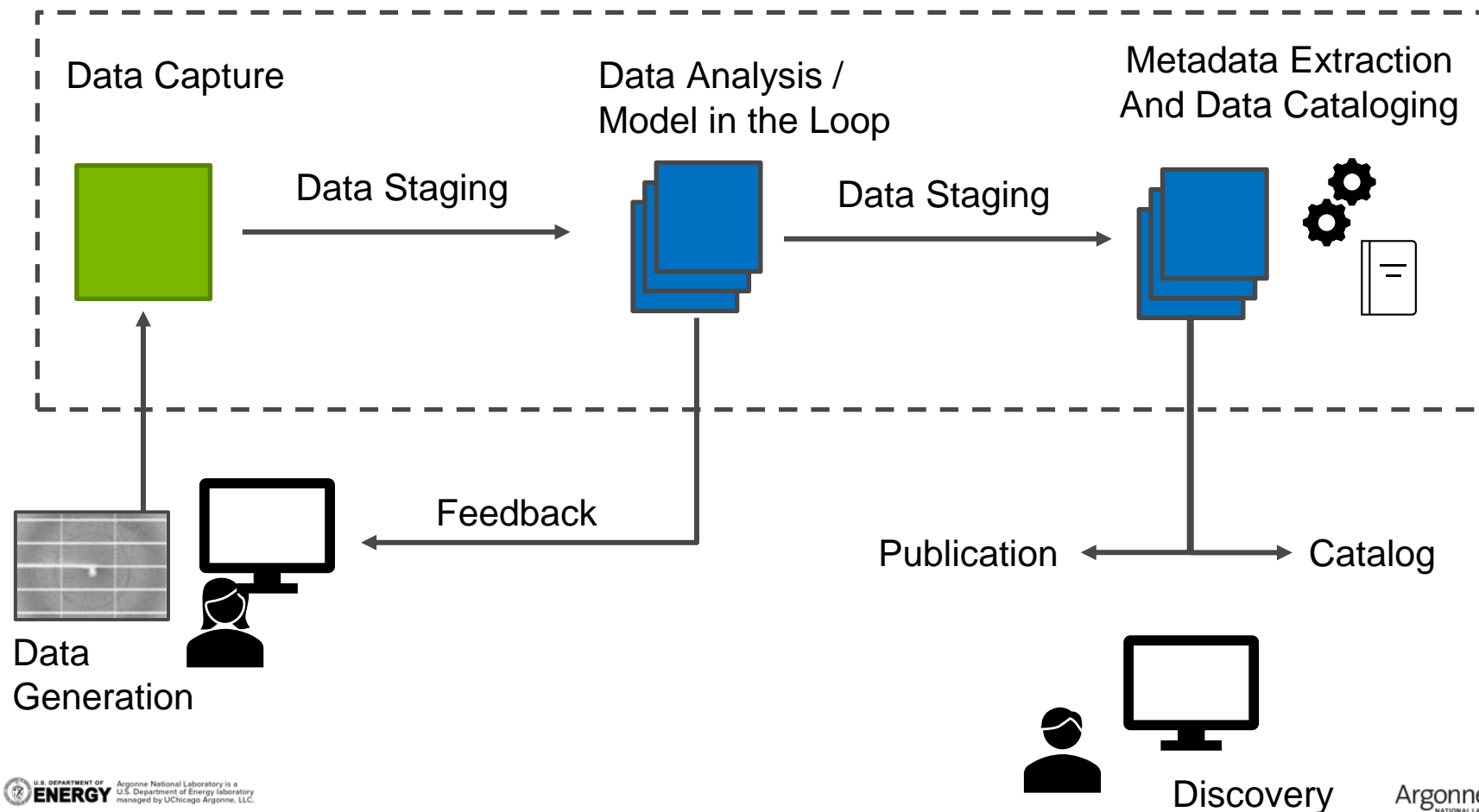


Gladier

Gladier: The Globus Architecture for Data-Intensive Experimental Research

- Build interactive and automated research flows with composable and modular service components and software
- Encourage FAIR (Findable, Accessible, Interoperable, Reusable) data principles for complex and distributed research flows
- Simplify the connection between experimental facilities and computing facilities (e.g., Leadership Computing Facilities, local clusters, etc.)

An Example Research Flow



Software and Services



APS Data Management

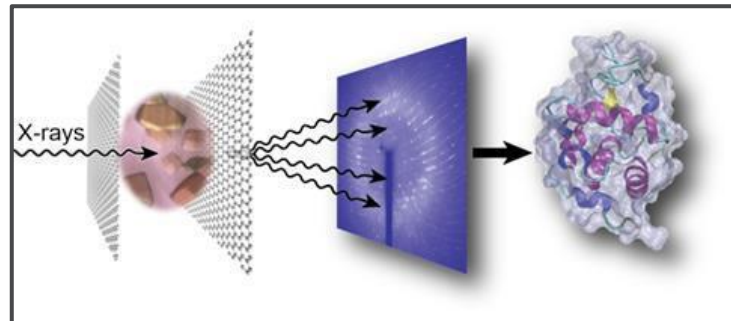


Many others...

Applications at the Advanced Photon Source

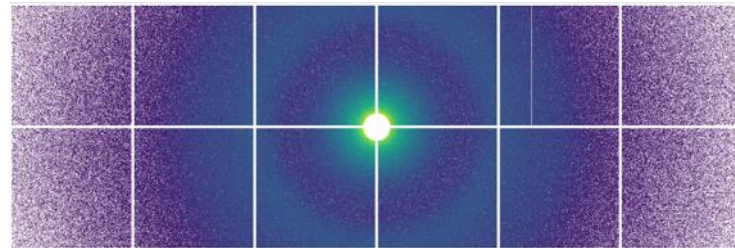
Serial X-Ray Crystallography (SSX)

High-throughput determination of complex protein structures at near room temperatures



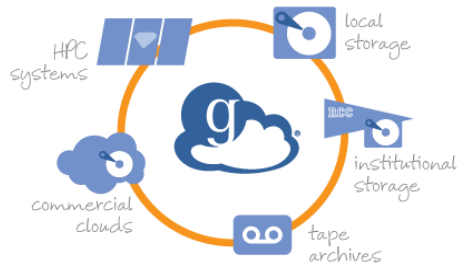
X-Ray Photon Correlated Spectroscopy (XPCS)

Powerful probe of microstructural dynamics in materials

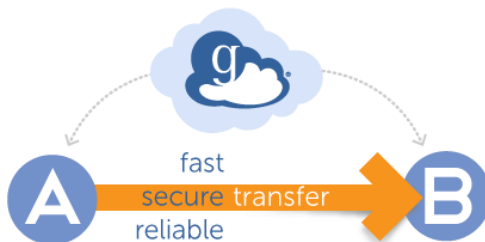


Globus – Research Data Management Platform

Unified Data Access



Data Transfer



Platform as a Service

Auth
Transfer
Share
Search
...



1960 most shared endpoints at a single institution	987+ PB moved	119 billion files processed	1866 active server endpoints
115 subscribers	149,000 total users	80 countries where Globus is used	
23,818 active personal endpoints	745 identity providers	2.9 PB largest single transfer to date	6764 active shared endpoints
			99.9% availability

See DOE Data
Days talk from
Vas Vasiliadis
for more details

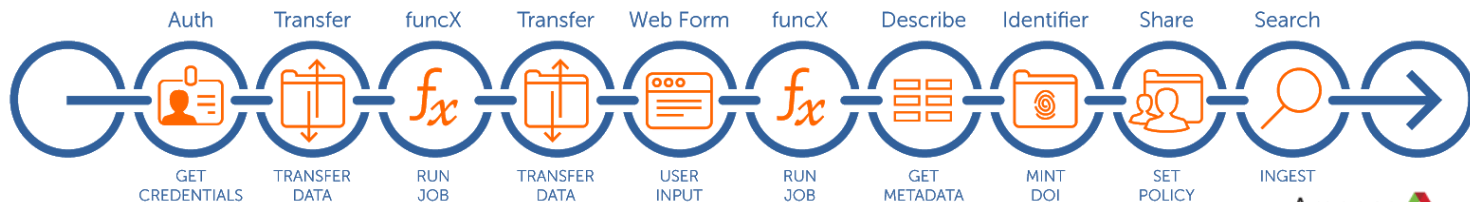
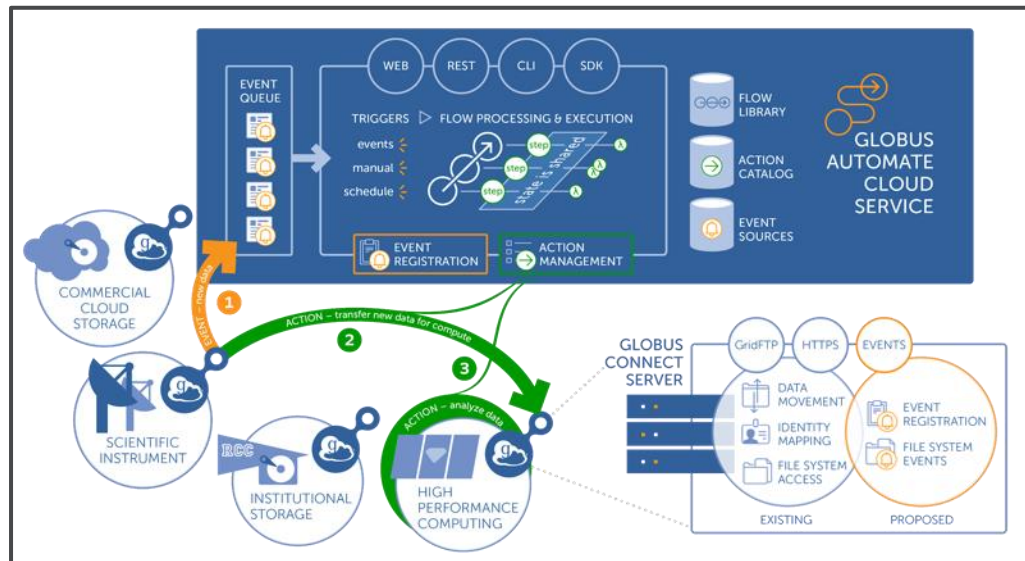
<https://www.globus.org>

Globus Automate



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

- Flows are comprised of steps
- Steps can be initiated by triggers (e.g. schedule, or events)
- Flows propagate state, and can contain loops, conditionals, and built in fault tolerance



funcX – Serverless Computing

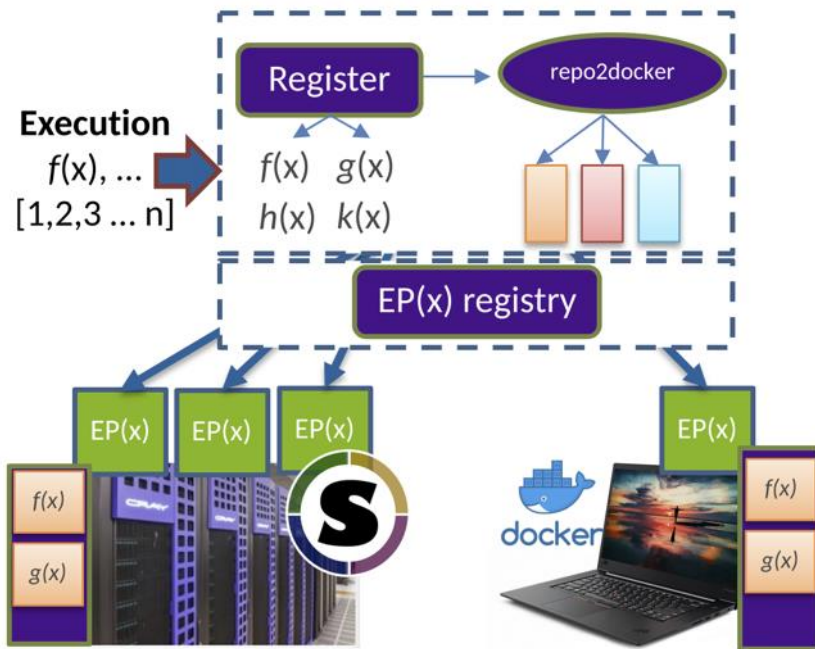
- Turn any machine into a *function* serving *endpoint*
- Remove barriers to using diverse and distributed infrastructure

Functions:

- Register once, run anywhere
- Create containers for encapsulation
- Authn/z for execution and sharing

Endpoints:

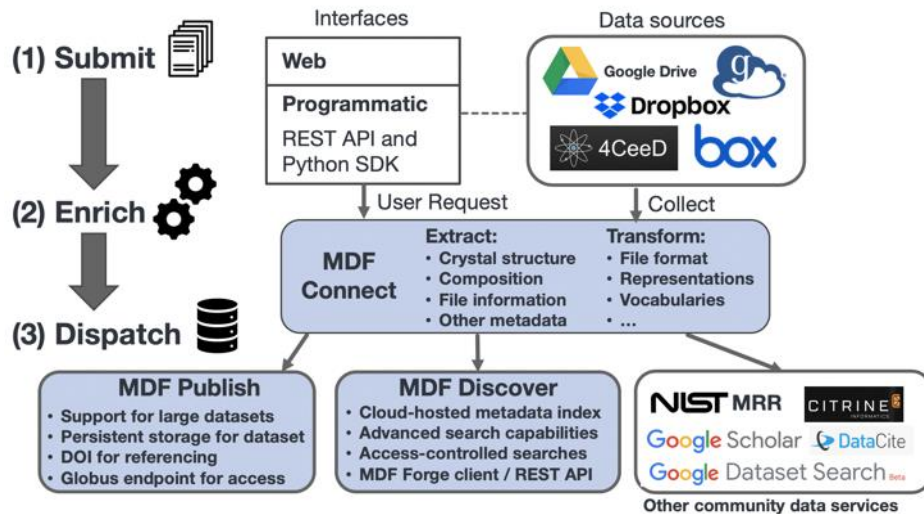
- Lightweight agent that can be deployed by users
- Abstracts underlying resource and elastically scales to demand



The Materials Data Facility



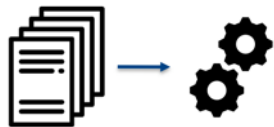
- Facilitate data publication and discovery
 - Move data to long term storage,
 - Mint permanent identifiers,
 - Extract and register metadata
 - Dispatch data and metadata to other services



CH_{MaD}

NIST

Extract Information from Collected Files



Index Dataset
Contents

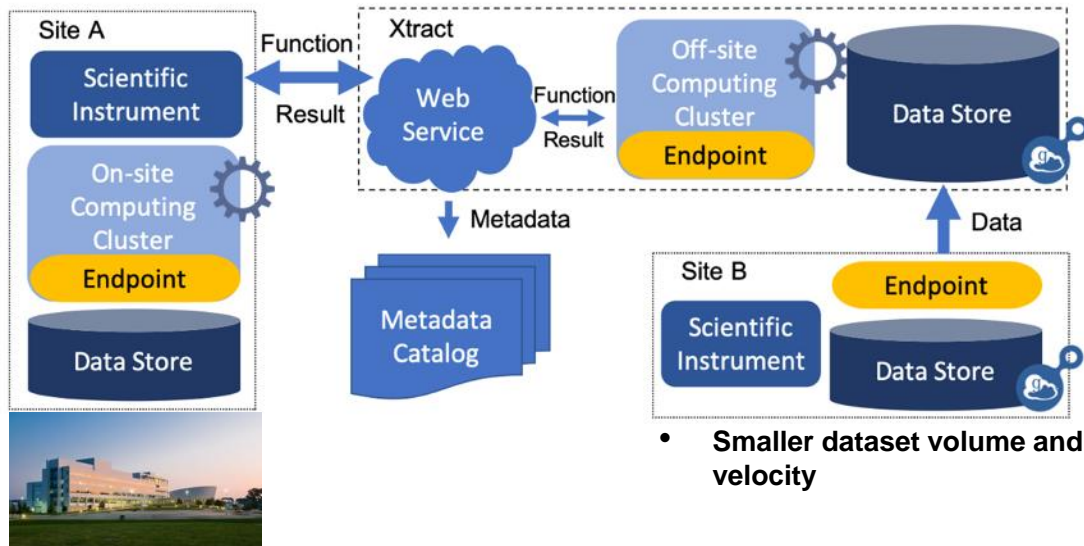
Enable Richer
Queries

Reduce User
Burden

Community for
Extractors

Xtract Service (currently internal use only)

- Modular domain-specific metadata extractors
- Smart data collection and movement
- Distributed architecture



- Smaller dataset volume and velocity

- User Facilities
- Data Sensitivities
- Large Data Sources

And Much More...

Transfer



Delete



ACLs



funcX



DLHub



Identifier



User Form



Notification



Xtract



Web Form



Ingest



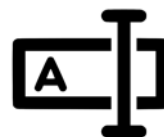
Expression
Evaluation



Search



Describe



An Example Research Flow

- Auth, Transfer, Search, Automate



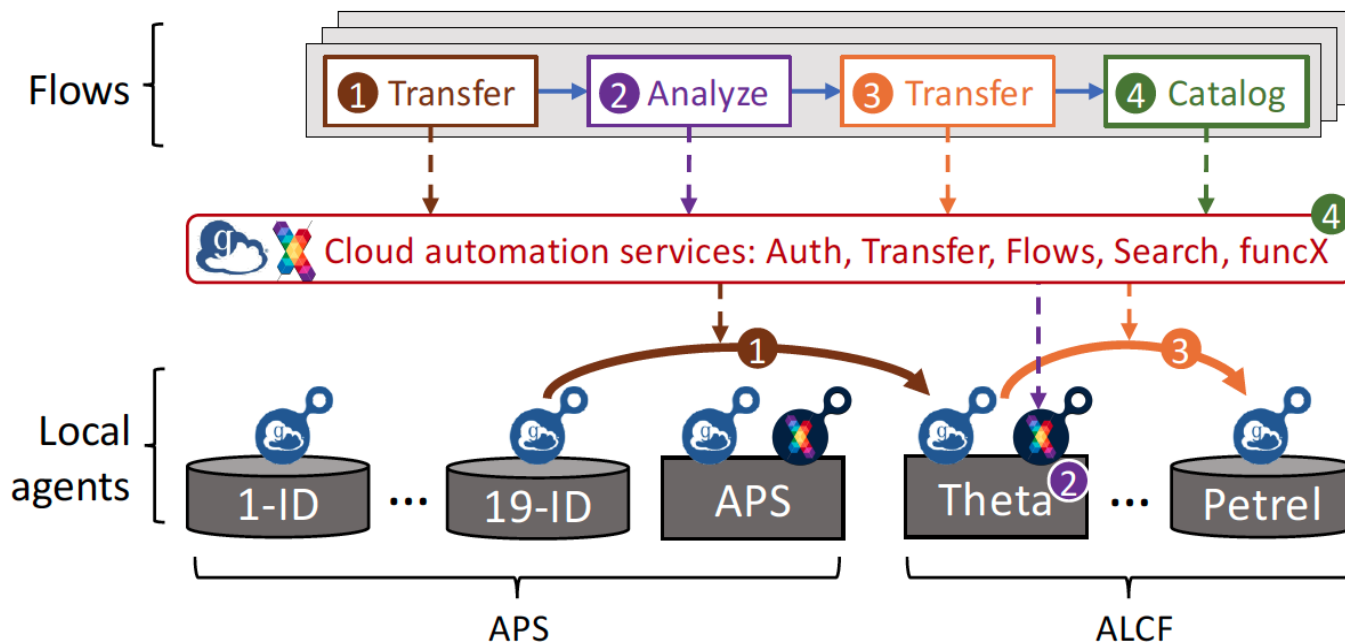
APS DM

- Data capture and processing

- Data publication, discovery, metadata extraction



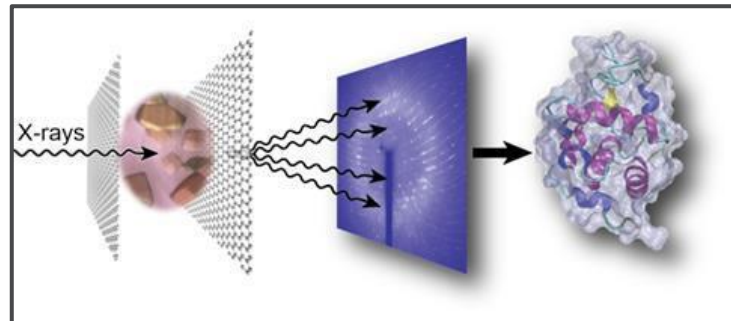
- Distributed computing



Applications at the Advanced Photon Source

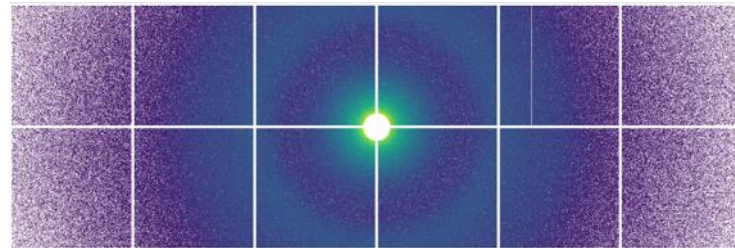
Serial X-Ray Crystallography (SSX)

High-throughput determination of complex protein structures at near room temperatures



X-Ray Photon Correlated Spectroscopy (XPCS)

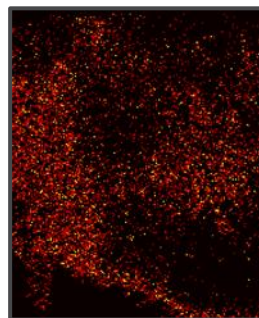
Powerful probe of microstructural dynamics in materials



Enabling Serial Crystallography (SSX) at Scale

Connecting light sources and leadership computing facilities to enable new science

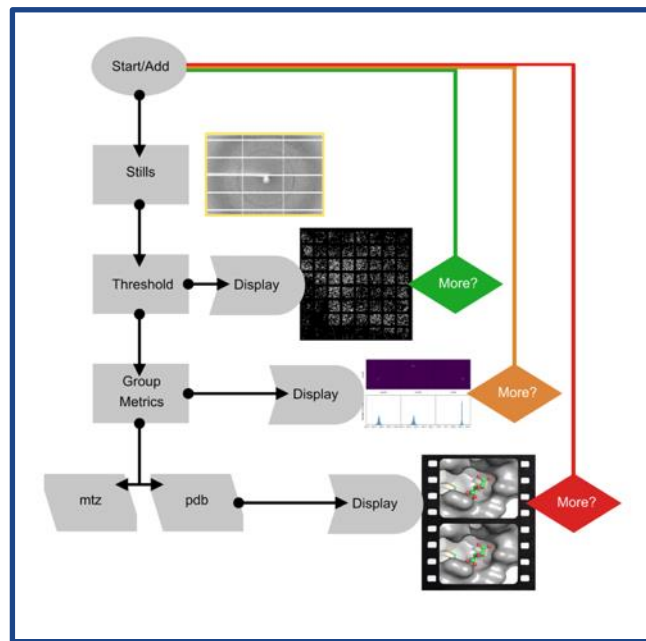
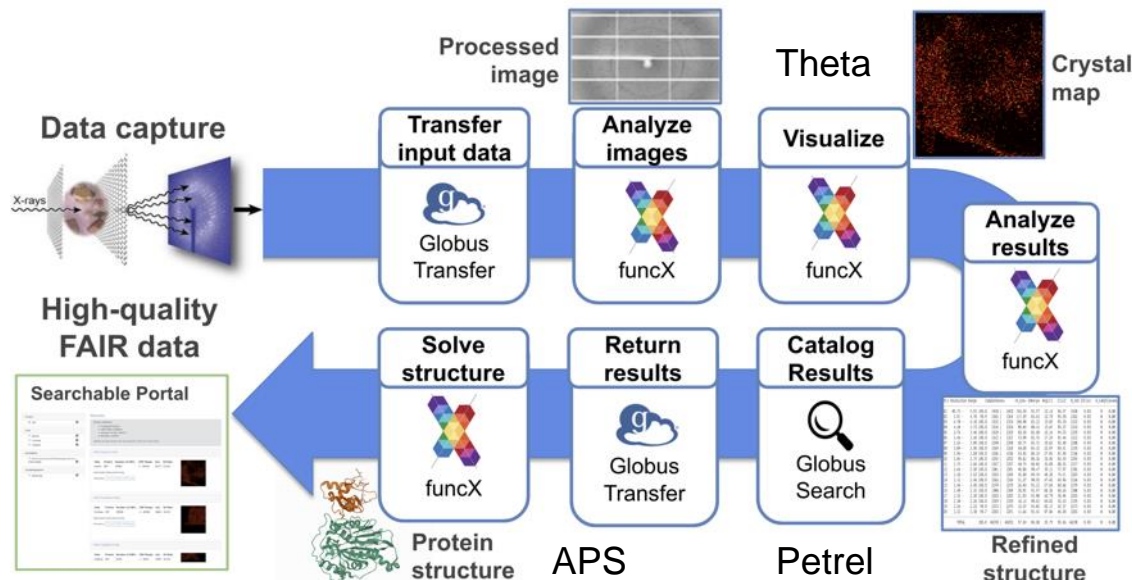
- Perform serial imaging of chips with thousands of embedded protein crystals
- Run quality control on initial set (e.g., first 1000) and report failures
- Analyze batches of images as collected
- Report statistics and and summary images during experiment
- Return crystal structure to scientist



Example sample map

SSX Crystallography - Automation in Depth

- Plugs into the APS data management system (DM)
- Creates inputs to Globus Automate, reliably batches files together automatically
- Invokes a flow to move data to ALCF, perform analysis, catalog results
- Integrates with ALCF portal allowing users to monitor experiments and reprocess data



Portal

With Andrzej Joachimiak, Darren Sherrell et al. APS Sector 19

SSX Search Index

Facilitate easy discovery and domain-specific interactions with data

- Integrates with Globus Search for back end
- Users publish metadata, figures, and files (int list, phil file, etc.)
- Users can: facet on relevant info, compute aggregate stats, see latest results (automatically updated during experiments), trigger reprocessing

Protein

☒ nsp10nsp16

5

Chip

☐ Haast1

1

☐ Haast2

1

☐ Inglewood

1

☐ Kinloch1

1

☐ Levin

1

Date Created

☐ May 12 2020

2

☐ May 13 2020

4

☐ May 30 2020

2

☒ May 31 2020

5

☐ Jun 02 2020

1

Results

Search Statistics

- 5 datasets found
- CBF Files: 19750.0
- Number of Ints: 2330.0

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

SSX Levin Chip

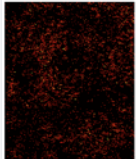
Chip	Protein	Number of CBFs	CBF Range	Ints
Levin	nsp10nsp16	37400	1 - 37400	5839

Automated data processing.

Int Files: [Levin_ints.txt](#)

Full Size Image Preview: [composite.png](#)

Updated: May 31, 2020, 12:18 p.m.



SSX Kinloch1 Chip

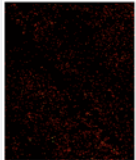
Chip	Protein	Number of CBFs	CBF Range	Ints
Kinloch1	nsp10nsp16	37400	1 - 37400	2455

Automated data processing.

Int Files: [Kinloch1_ints.txt](#)

Full Size Image Preview: [composite.png](#)

Updated: May 31, 2020, 11:55 a.m.



SSX Portal: Search data, track processing progress, and more

Gladier in the DOE COVID-19 Fight

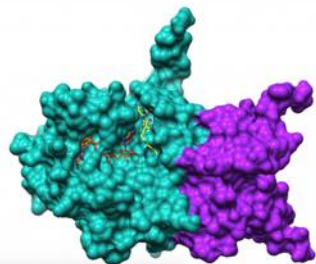
“These data services have taken the time to solve a structure from weeks to days and now to hours”

Darren Sherrell, SBC beamline scientist APS Sector 19

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



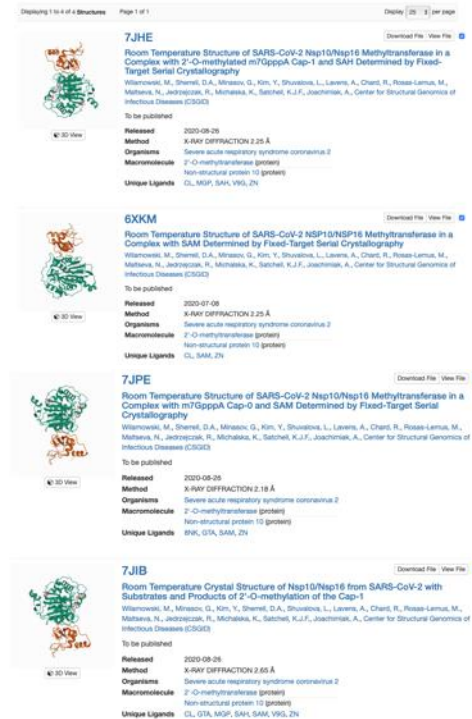
HPC wire

Since 1967 - Covering the Fastest
Computers in the World and the People Who

Argonne's User Facilities Continue to Enable Critical Work Combating and Addressing the Impacts of the COVID-19 Epidemic
June 12, 2020

ALCF + APS capabilities were used to determine the room temperature structure of 2 viral surface proteins

Next steps:
Develop methods paper, continue running flow



Displaying 1 to 4 of 4 Structures Page 1 of 1

7JHE
Room Temperature Structure of SARS-CoV-2 Nsp10/Nsp16 Methyltransferase in a Complex with 2'-O-methylated m7GpppA Cap-1 and SAM Determined by Fixed-Target Serial Crystallography
Wilmarowski, M., Shemelt, D.A., Miranek, G., Kim, Y., Shustrova, L., Lavers, A., Chant, R., Rose-Lemus, M., Mathews, N., Jadczyk, R., Mohabak, K., Satchel, K.J.F., Joachimiak, A., Center for Structural Genomics of Infectious Diseases (CSGI)
To be published
Released: 2020-08-28
Method: X-RAY DIFFRACTION 2.25 Å
Organisms: Severe acute respiratory syndrome coronavirus 2
Macromolecule: 2'-O-methyltransferase (protein)
Unique Ligands: CL, NSP, SAM, V95, ZN

6XKM
Room Temperature Structure of SARS-CoV-2 Nsp10/Nsp16 Methyltransferase in a Complex with SAM Determined by Fixed-Target Serial Crystallography
Wilmarowski, M., Shemelt, D.A., Miranek, G., Kim, Y., Shustrova, L., Lavers, A., Chant, R., Rose-Lemus, M., Mathews, N., Jadczyk, R., Mohabak, K., Satchel, K.J.F., Joachimiak, A., Center for Structural Genomics of Infectious Diseases (CSGI)
To be published
Released: 2020-07-08
Method: X-RAY DIFFRACTION 2.25 Å
Organisms: Severe acute respiratory syndrome coronavirus 2
Macromolecule: 2'-O-methyltransferase (protein)
Unique Ligands: CL, SAM, ZN

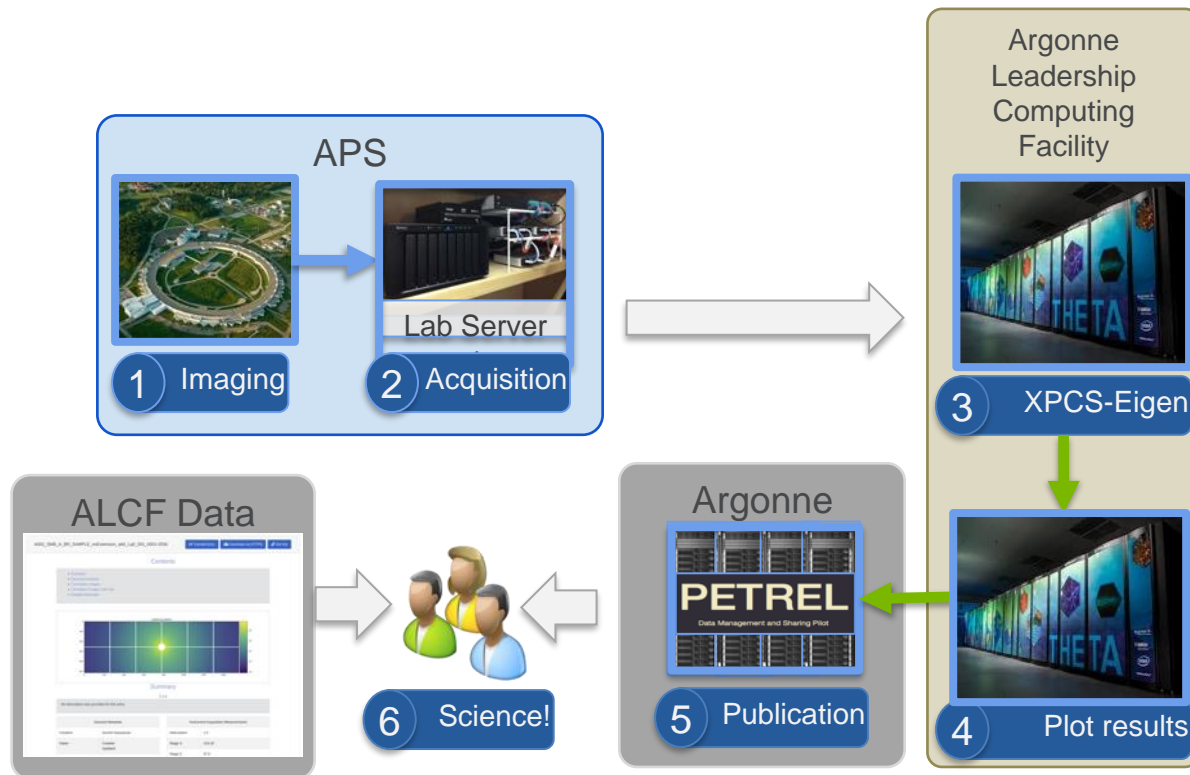
7JPE
Room Temperature Structure of SARS-CoV-2 Nsp10/Nsp16 Methyltransferase in a Complex with m7GpppA Cap-0 and SAM Determined by Fixed-Target Serial Crystallography
Wilmarowski, M., Shemelt, D.A., Miranek, G., Kim, Y., Shustrova, L., Lavers, A., Chant, R., Rose-Lemus, M., Mathews, N., Jadczyk, R., Mohabak, K., Satchel, K.J.F., Joachimiak, A., Center for Structural Genomics of Infectious Diseases (CSGI)
To be published
Released: 2020-08-28
Method: X-RAY DIFFRACTION 2.18 Å
Organisms: Severe acute respiratory syndrome coronavirus 2
Macromolecule: 2'-O-methyltransferase (protein)
Unique Ligands: BHK, GTA, SAM, ZN

7JIB
Room Temperature Crystal Structure of Nsp10/Nsp16 from SARS-CoV-2 with Substrates and Products of 2'-O-methylation of the Cap-1
Wilmarowski, M., Miranek, G., Kim, Y., Shemelt, D.A., Shustrova, L., Lavers, A., Chant, R., Rose-Lemus, M., Mathews, N., Jadczyk, R., Mohabak, K., Satchel, K.J.F., Joachimiak, A., Center for Structural Genomics of Infectious Diseases (CSGI)
To be published
Released: 2020-08-28
Method: X-RAY DIFFRACTION 2.65 Å
Organisms: Severe acute respiratory syndrome coronavirus 2
Macromolecule: 2'-O-methyltransferase (protein)
Unique Ligands: CL, GTA, MGP, SAM, V95, ZN

4 structures available in PDB – Scientific paper forthcoming

Gladier - Automated XPCS Flow

- Automate flows stage data to ALCF for on-demand analysis and publication
- Metadata and plots are dynamically extracted and published into a search catalog
- Scientists can select datasets and initiate flows to perform batch analysis tasks, publish to MDF



With Nicholas Schwarz, and Suresh Narayanan et al. APS Sector 8-ID

XPCS Data Processing

Facilitate easy discovery and domain-specific interactions with data

- Users gather collections of data using Globus Search
- Collections can be published to MDF, transferred to a new location, or re-processed through Theta
- Collections can be tailored for each Index. Currently used by ExaLearn and XPCS.

Transfer Status

xpcs-4bags-april	i
concierge-test-upgrade	i
Status	SUCCEEDED
Date Started	Feb. 12, 2020, 9:52 a.m.
Source	petrel@xpcs
Destination	Globus Tutorial Endpoint 1
Files Transferred	12
Total Data Transferred	892 bytes
Remove	
concierge-upgrade-test	i

My Bags

aerogel-samples [Info](#) [Transfer](#) [Reprocess](#) [Delete](#)

aerogel-samples

Date Created	Aug. 31, 2020, 3:01 p.m.	Files	1312
Index	xpcs	Total Aggregated Size	34.7 GB
ID	minid.test:199NbJQ3rcS		
Created By	https://petreldata.net		
Original Query	Search URL		

MDF

Submission Title

aerogel-samples

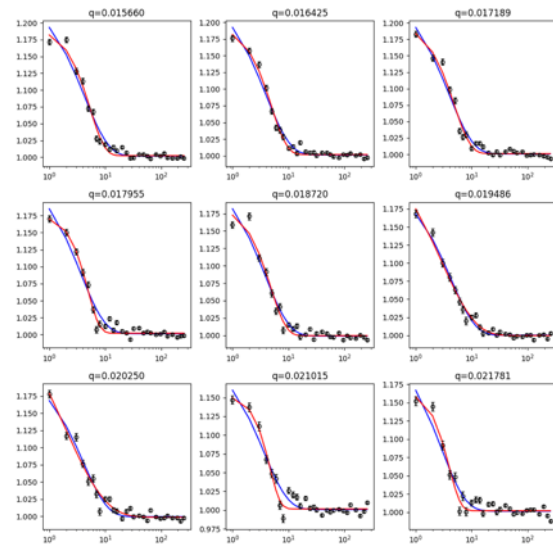
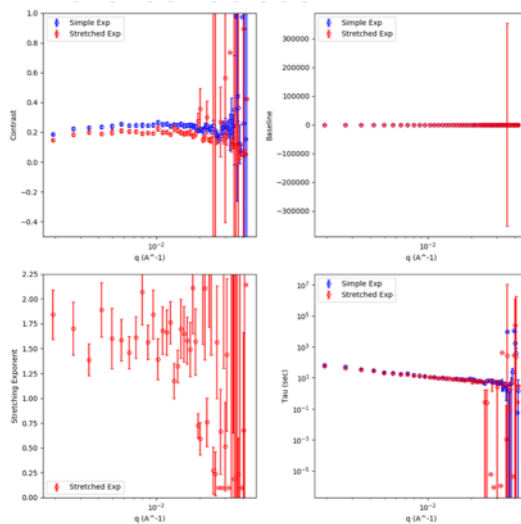
A title for this Submission

[Publish to MDF](#)

XPCS Detail View: Move data, reprocess data, or publish data to the community

Gladier - Automated XPCS Flow

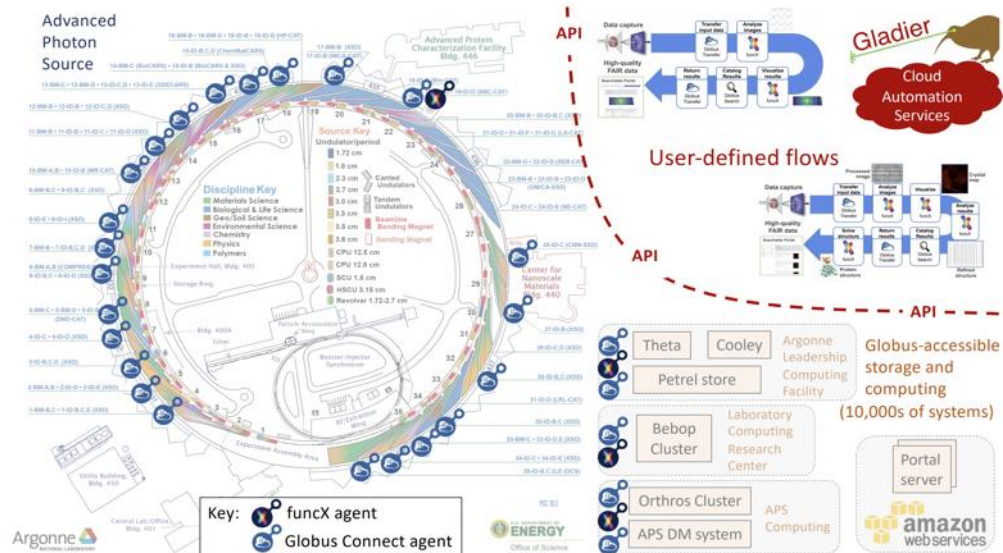
- Automate flows stage data to ALCF for on-demand analysis and publication
- Metadata and plots are dynamically extracted and published into a search catalog
- Scientists can select datasets and initiate flows to perform batch analysis tasks, publish to MDF



With Nicholas Schwarz, and Suresh Narayanan et al. APS Sector 8-ID

Next Steps

- Continue to generalize the software components for wider release
- Integrate more services
- Gather use cases, and deploy these capabilities across the Advanced Photon Source (APS)
- Engage partners beyond APS



Thank You!

Gladier

Argonne Leadership Computing Facility

The **Advanced Photon Source**

a U.S. Department of Energy Office of Science User Facility



U.S. DEPARTMENT OF
ENERGY

Contact: Ben Blaiszik (bblaiszik@anl.gov)



<https://www.funcx.org>



CHMaD

NIST

<https://www.materialsdatafacility.org>



U.S. DEPARTMENT OF
ENERGY

<https://www.dlhub.org>