NEXUS:
INTEGRATED RESEARCH INFRASTRUCTURE
AT ARGONNE

Tom Uram, turam@anl.gov - Argonne Leadership Computing Facility

Hands-on Workshop, October 2023

U.S. DEPARTMENT OF ENERGY

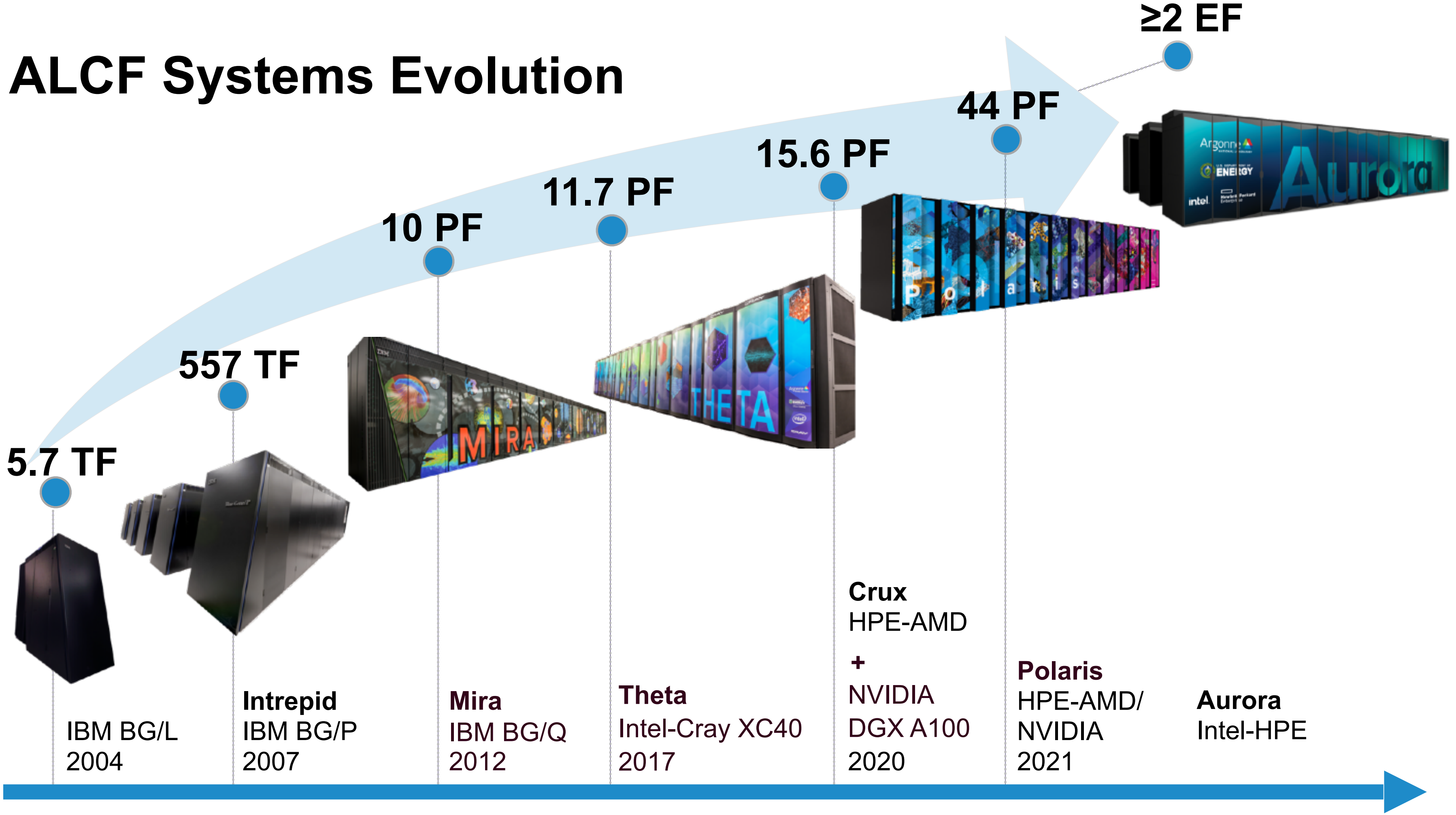
Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

Argonne NATIONAL LABORATORY

Polaris

Argonne
NATIONAL LABORATORY
U.S. DEPARTMENT OF
ENERGY
intel
Hewlett Packard
Enterprise
Aurora

ALCF Systems Evolution



5.7 TF

IBM BG/L
2004

557 TF

Intrepid
IBM BG/P
2007

10 PF

Mira
IBM BG/Q
2012

11.7 PF

Theta
Intel-Cray XC40
2017

15.6 PF

Crux
HPE-AMD
+
NVIDIA
DGX A100
2020

44 PF

Polaris
HPE-AMD/
NVIDIA
2021

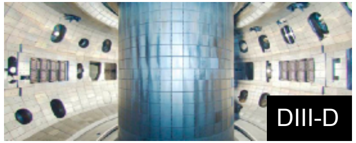
≥2 EF

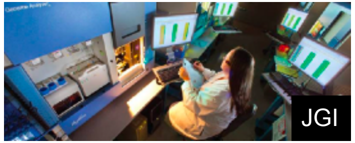
Aurora
Intel-HPE

DOE EXPERIMENTAL USER FACILITIES

- DOE operates 24 experimental user facilities
- Similar to the computing facilities, some of them are undergoing upgrades
- Their data rate and their computing needs will increase accordingly

FNAL AC


DIII-D

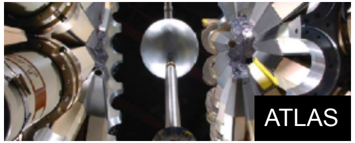

JGI

NSTX-U

TMF

SNS

CNMS


ATLAS

RHIC

NSLS-II

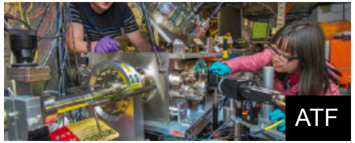
CNM

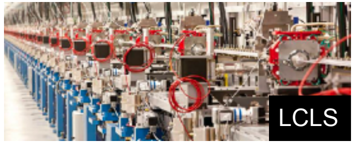
SSRL

EMSL

HFIR

CFN


ATF


LCLS

FRIB

CEBAF

ARM

CINT

FACET-II

APS

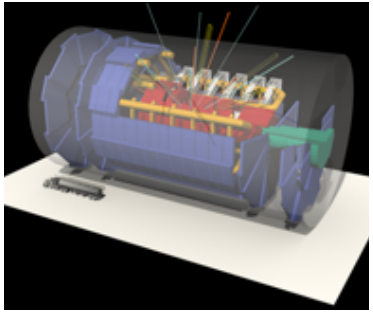
LCLS

ALCF SUPPORT FOR EXPERIMENTAL SCIENTIFIC COMPUTING

Allocation programs: Director's Discretionary, Early Science Program, ALCF Data Science Program

Technologies: Balsam Workflows, Cobalt Scheduler, Globus Transfer

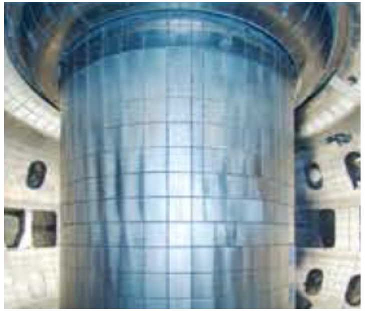
HEP - ATLAS



2015-16

Hundreds of millions of core hours of simulation and analysis for ATLAS (DD, ADSP, ESP)

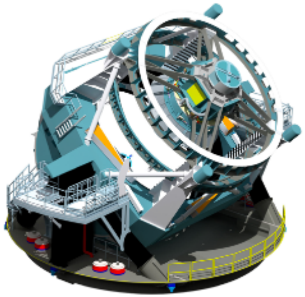
FES - DIII-D



2017

Near real-time analysis of DIII-D fusion experiment data, powered by Balsam workflows (DD); deep learning for fusion (ESP)

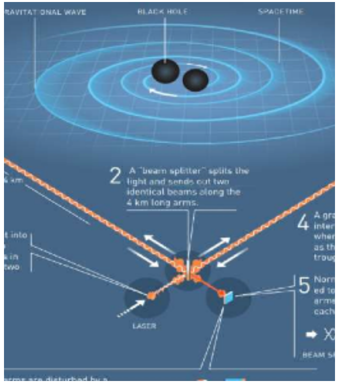
HEP/NP



2017-18

Simulation and analysis of telescope images for LSST-DESC (ADSP)

HEP - LIGO



2018-19

Deep learning for gravitational wave detection with LIGO (ADSP)

ALCF SUPPORT FOR EXPERIMENTAL SCIENTIFIC COMPUTING

Allocation programs: Director's Discretionary, Early Science Program, ALCF Data Science Program

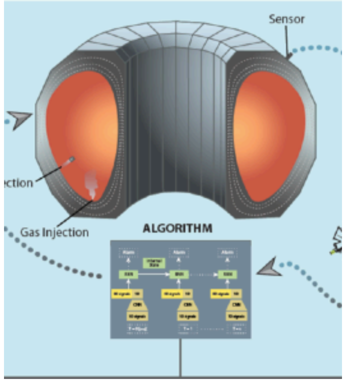
Technologies: Balsam Workflows, Cobalt Scheduler, Globus Transfer

BES/OTHER

2018-

3D Reconstruction of mouse brain from electron microscopy images at Argonne (DD, ADSP, ESP)

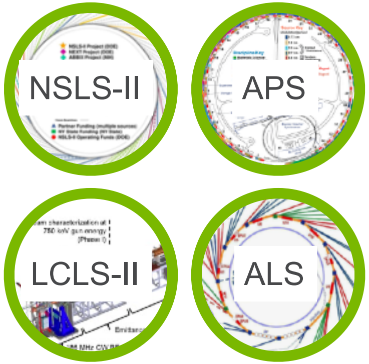
FES



2018-

Hyperparameter tuning at exascale to predict and mitigate disruption events (ESP)

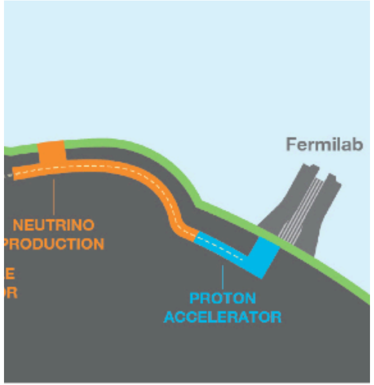
BES - LS



2019-

Online analysis of experiment data from multiple light sources using near-real-time queue on Theta (ADSP)

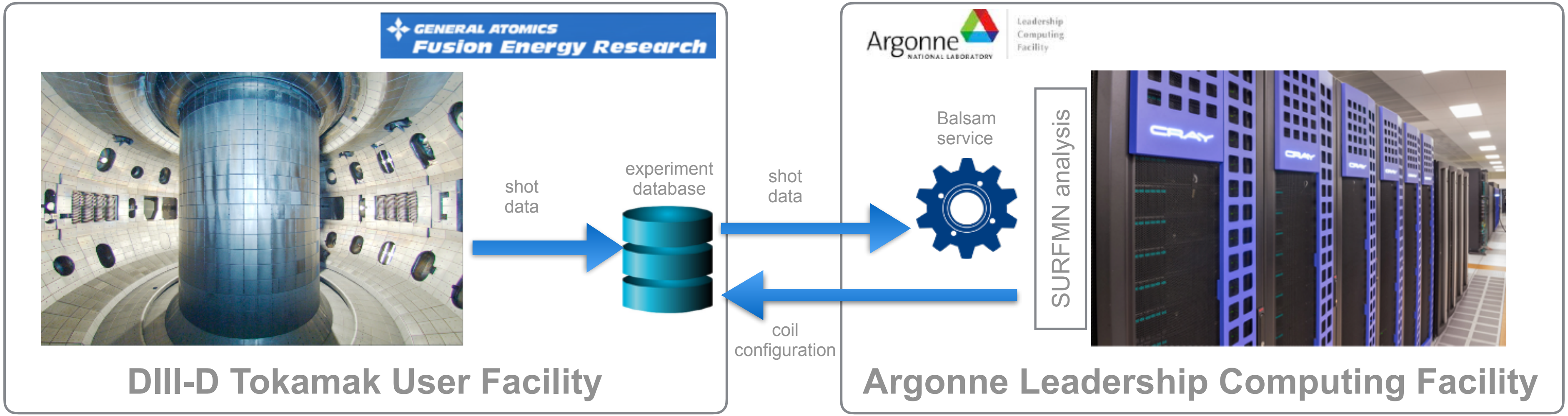
HEP - SBND



2019-

Large-scale simulation and deep learning for SBND/ DUNE (DD, ADSP)

Automatic Between-shot Analysis of DIII-D Experimental Data



GENERAL ATOMICS Fusion Energy Research

shot data

experiment database

shot data

coil configuration

Argonne NATIONAL LABORATORY Leadership Computing Facility

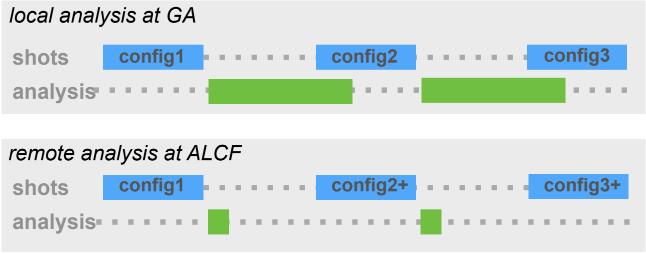
Balsam service

SURFMN analysis

DIII-D Tokamak User Facility

Argonne Leadership Computing Facility

- Scientists configure experimental “shots” every 20 minutes
  - A shot is an attempt to magnetically confine high temperature plasma
  - The timing/current of magnetic coils are configured to control the plasma during a disruption to avoid damage to the containing vessel (applicable to DIII-D and future reactors)
  - Analyses indicate how to optimize coil configuration for confinement
- Each shot triggers an automatic, near real-time analysis job at ALCF
- GA scientists integrate analysis results into configuration for next shot
- Analysis at ALCF enables more complex analyses (16x resolution) to be completed faster, improving the accuracy of results and allowing analyses to inform every shot instead of



local analysis at GA

shots config1 config2 config3

analysis

remote analysis at ALCF

shots config1 config2+ config3+

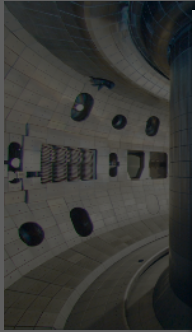
analysis

Faster analysis time allows analysis results to be integrated into magnet configuration for subsequent shots. Higher resolution analyses improve configuration accuracy.

M. Kostuk, T. Uram, et al, 2nd IAEA Technical Meeting on Fusion Data Processing, Validation, and Analysis, 2018

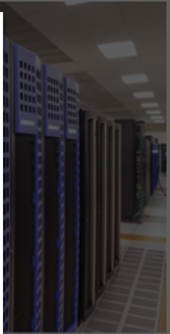
Automatic Between-shot Analysis of DIII-D Experimental Data

GENERAL ATOMICS
Fusion Energy Research

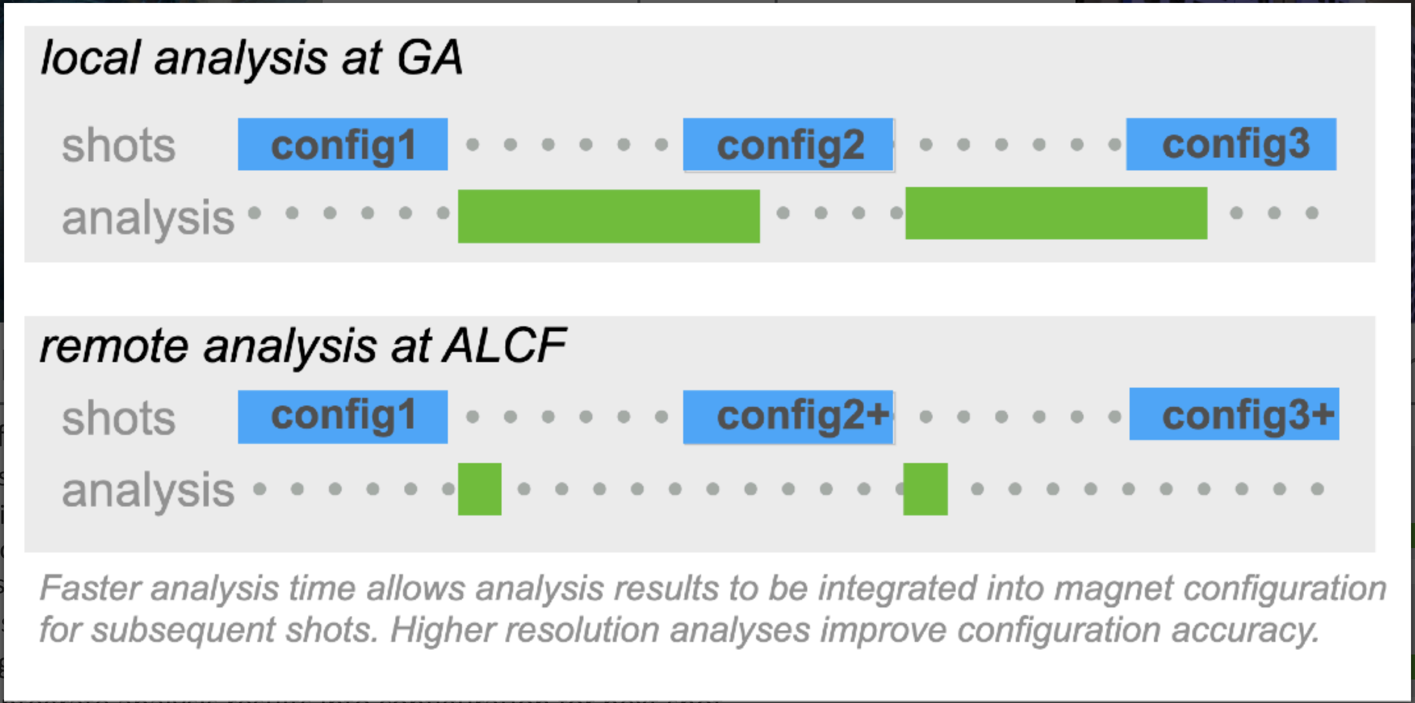


DIII-

Argonne NATIONAL LABORATORY | Leadership Computing Facility

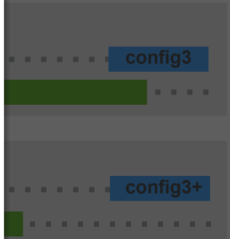


ng Facility



Faster analysis time allows analysis results to be integrated into magnet configuration for subsequent shots. Higher resolution analyses improve configuration accuracy.

- Scientists conf
  - A shot is
  - The timi disruptio reactors
  - Analyses
- Each shot trigg
- GA scientists integrate analysis results into configuration for next shot
- Analysis at ALCF enables more complex analyses (16x resolution) to be completed faster, improving the accuracy of results and allowing analyses to inform every shot instead of

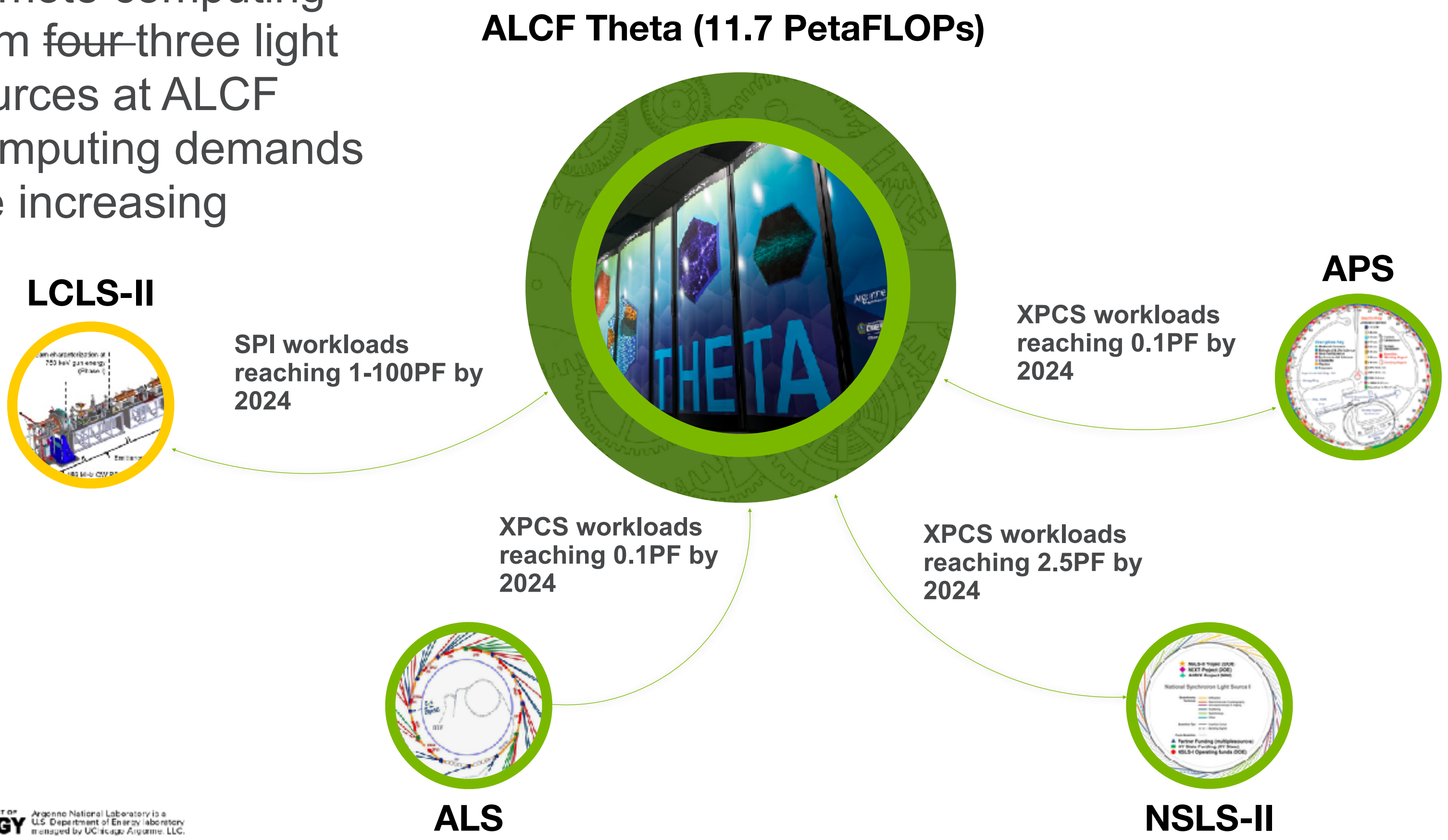


Faster analysis time allows analysis results to be integrated into magnet configuration for subsequent shots. Higher resolution analyses improve configuration accuracy.

M. Kostuk, T. Uram, et al, 2nd IAEA Technical Meeting on Fusion Data Processing, Validation, and Analysis, 2018

ALCF SUPPORT OF LIGHT SOURCE COMPUTING

- Remote computing from ~~four~~ three light sources at ALCF
- Computing demands are increasing



ALCF Theta (11.7 PetaFLOPs)

LCLS-II

SPI workloads reaching 1-100PF by 2024

APS

XPCS workloads reaching 0.1PF by 2024

ALS

XPCS workloads reaching 0.1PF by 2024

NSLS-II

XPCS workloads reaching 2.5PF by 2024

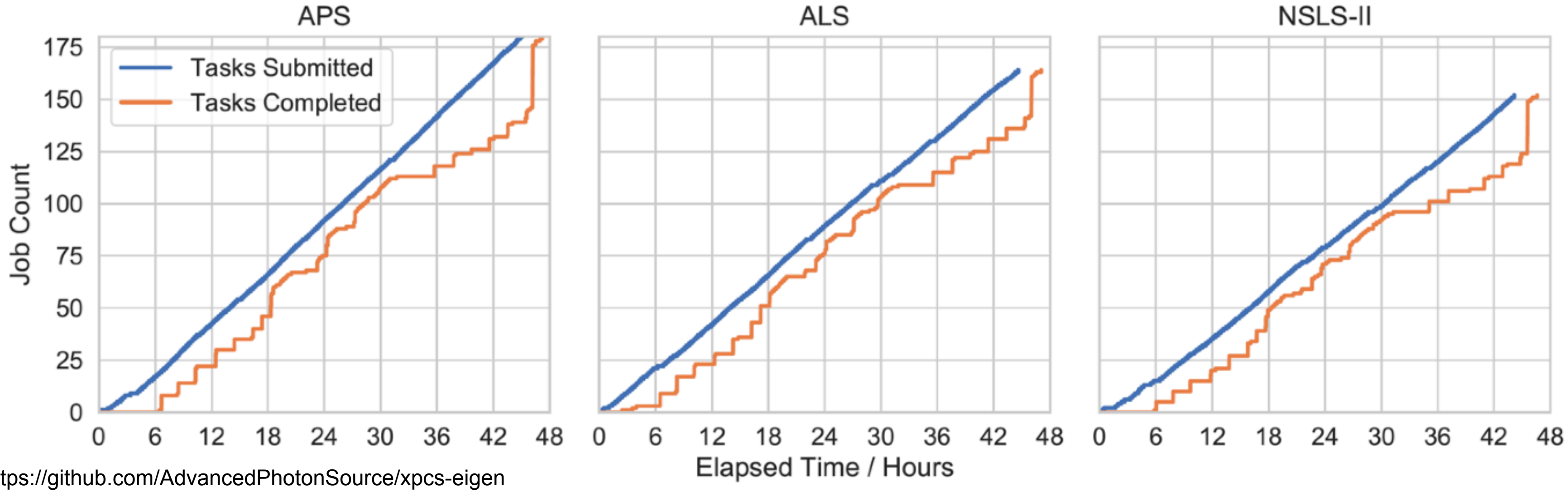
NEAR REAL-TIME PROCESSING OF WORKLOADS FROM THREE LIGHT SOURCES AT ALCF

Experiment

- Transfer 40GB input dataset from APS, ALS, NSLS-II
- Analyze data in near real-time with XPCS-Eigen* using **backfill queue** on Theta
- Transfer results to originating light source

Results

- Continuously executed for **48+ hours**
- Transferred 23TB input data from APS/ALS/NSLS-II to ALCF
- Analyzed 500+ datasets
- Transferred 179GB output data from ALCF to APS/ALS/NSLS-II

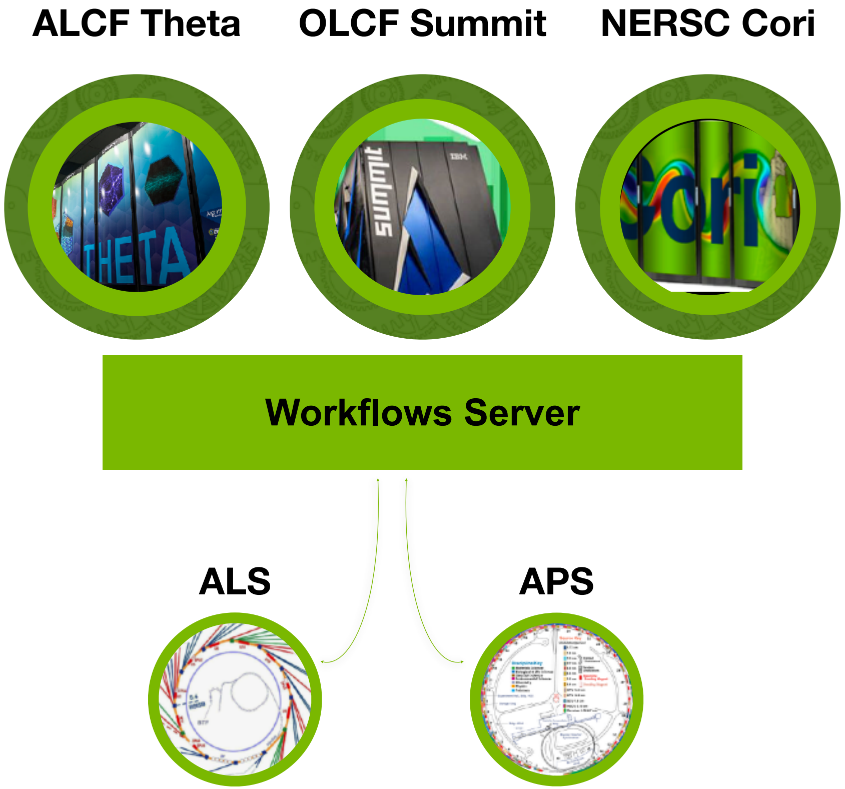


* https://github.com/AdvancedPhotonSource/xpcs-eigen

XPCS WORKFLOW BETWEEN TWO LIGHT SOURCES AND THREE COMPUTE FACILITIES

Balsam enabled analysis of data originating at two light souces (APS and ALS) to run at three DOE compute facilities (ALCF, OLCF, and NERSC)

- A Balsam Site agent runs at each compute facility, interfacing closely with the local job scheduler.
- The Balsam Workflows Server exposes a **public API** used to define jobs to run at each Site.
- XPCS jobs were injected from the light sources, targeting one of the compute facilities. Job throughput was monitored to **intelligently direct subsequent job submissions**.
- Data transfers were conducted using Globus, using a batched approach to optimize throughput.
- Job failures were synthesized to demonstrate Balsam's ability to recover by redirecting jobs to other sites.



ALCF Theta

OLCF Summit

NERSC Cori

Workflows Server

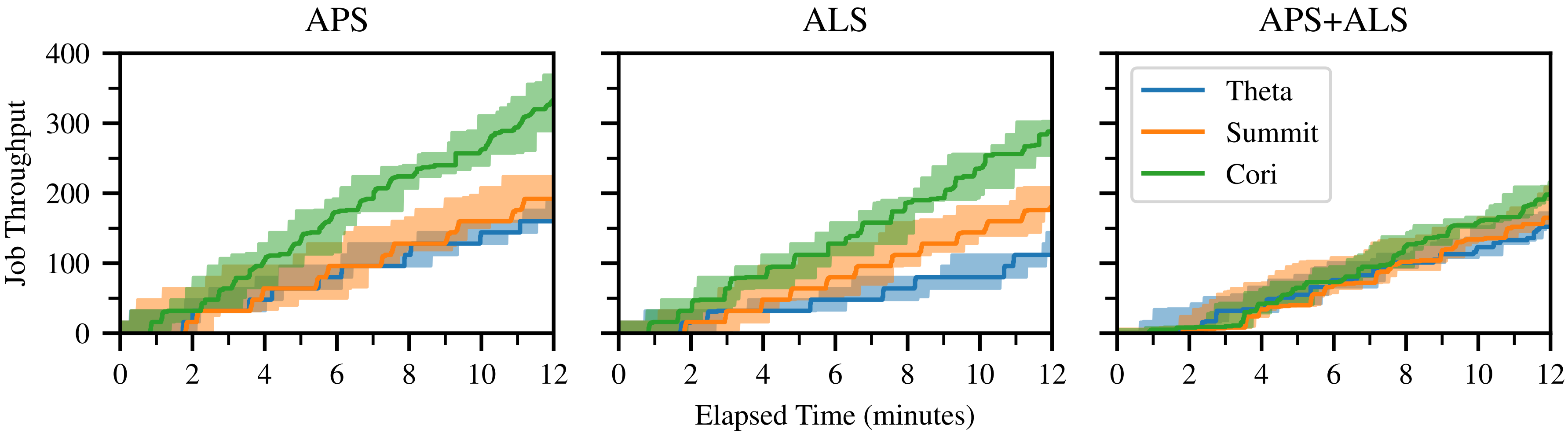
ALS

APS

Toward Real-time Analysis of Experimental Science Workloads on Geographically Distributed Supercomputers, Salim et al

Automated analysis between two light sources and three compute facilities

- Jobs and data transfers scheduled with single Python call
- XPCS dataset (878MB) transferred from APS to Theta/Summit/Cori (left), ALS to Theta/Summit/Cori, and APS/ALS to Theta/Summit/Cori
- Pool of nodes maintained for fast injection of jobs on arrival, and immediate return of results to originating site



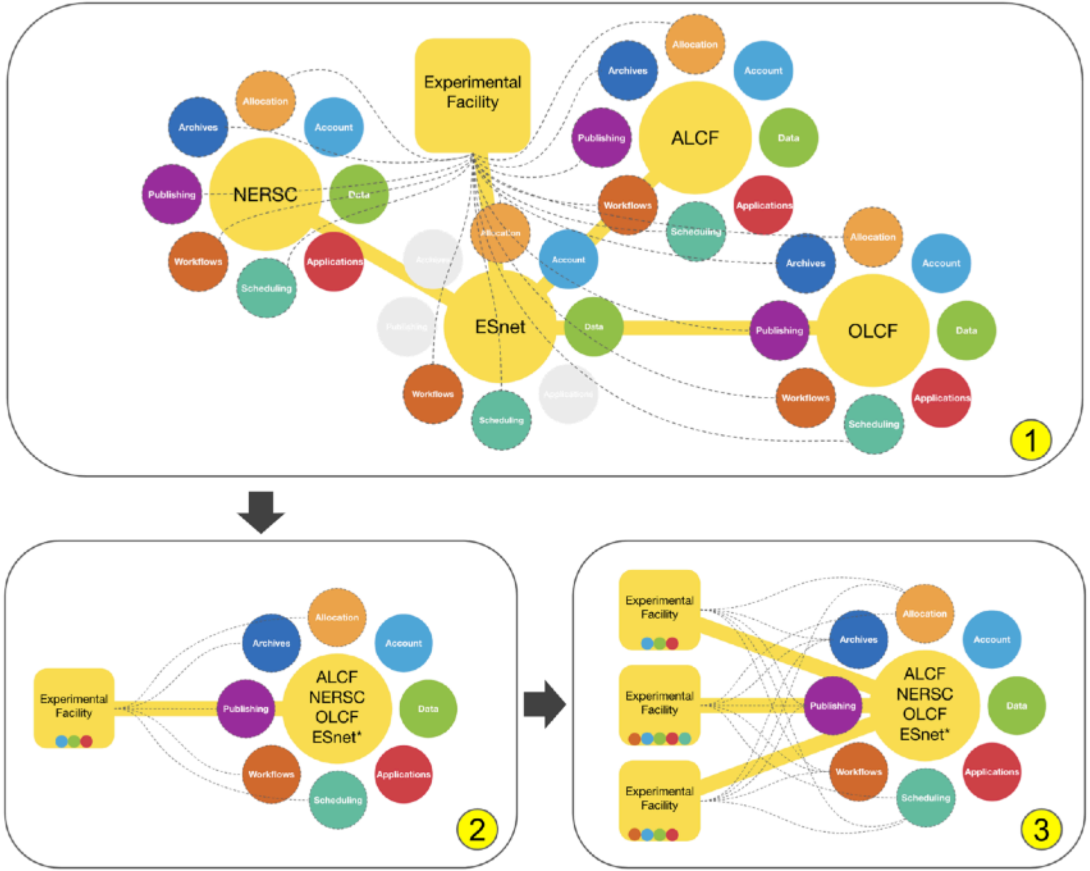
IRI White Paper March 2021

Depiction of the integration of experimental facilities with computational facilities, across the range of services provided, in contrast with the one-to-one approach required today.

1. Today, an experimental facility must arrange separate bespoke interactions with individual HPC/HPN facilities.

2. A future paradigm with common interfaces could simplify integration of an experimental facility with multiple HPC/HPN facilities.

3. In turn, these common interfaces could support expansion and integration **across multiple experimental facilities and HPC/HPN facilities**.



AL Allocations

AC Accounts

DA Data

AP Applications

SC Scheduling

WF Workflows

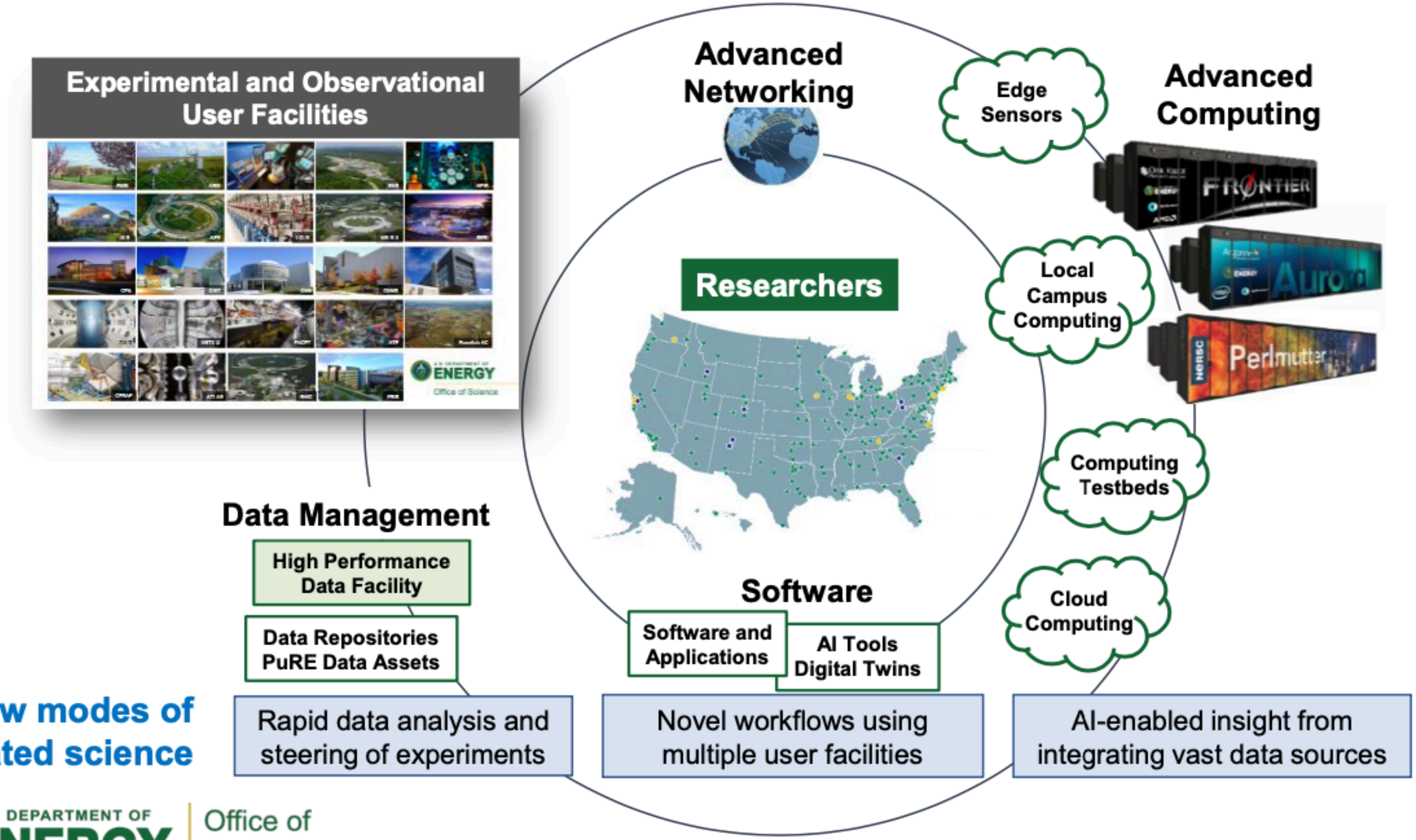
PB Publication

AR Archiving

Toward a Seamless Integration of Computing, Experimental, and Observational
Science Facilities: A Blueprint to Accelerate Discovery
(www.osti.gov/servlets/purl/1863562)

DOE's Integrated Research Infrastructure (IRI) Vision:

To empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation



Experimental and Observational User Facilities

Advanced Networking

Advanced Computing

Edge Sensors

Local Campus Computing

Computing Testbeds

Cloud Computing

Researchers

Data Management

High Performance Data Facility

Data Repositories PuRE Data Assets

Software

Software and Applications

AI Tools Digital Twins

New modes of integrated science

Rapid data analysis and steering of experiments

Novel workflows using multiple user facilities

AI-enabled insight from integrating vast data sources

U.S. DEPARTMENT OF ENERGY | Office of Science

Slide from Ben Brown, DOE, ASCAC, June 2023

IRI BLUEPRINT ACTIVITY REPORT (2023)

U.S. DEPARTMENT OF ENERGY | Office of Science

THE DOE OFFICE OF SCIENCE

Integrated Research Infrastructure Architecture Blueprint Activity

FINAL REPORT

2023

IRI Science Patterns (3)

Time-sensitive pattern has *urgency*, requiring real-time or end-to-end performance with high reliability, e.g., for timely decision-making, experiment steering, and virtual proximity.

Data integration-intensive pattern requires combining and analyzing data from multiple sources, e.g., sites, experiments, and/or computational runs.

Long-term campaign pattern requires sustained access to resources over a long period to accomplish a well-defined objective.

IRI Practice Areas (6)

User experience practice will ensure relentless attention to user perspectives and needs through requirements gathering, user-centric (co)-design, continuous feedback, and other means.

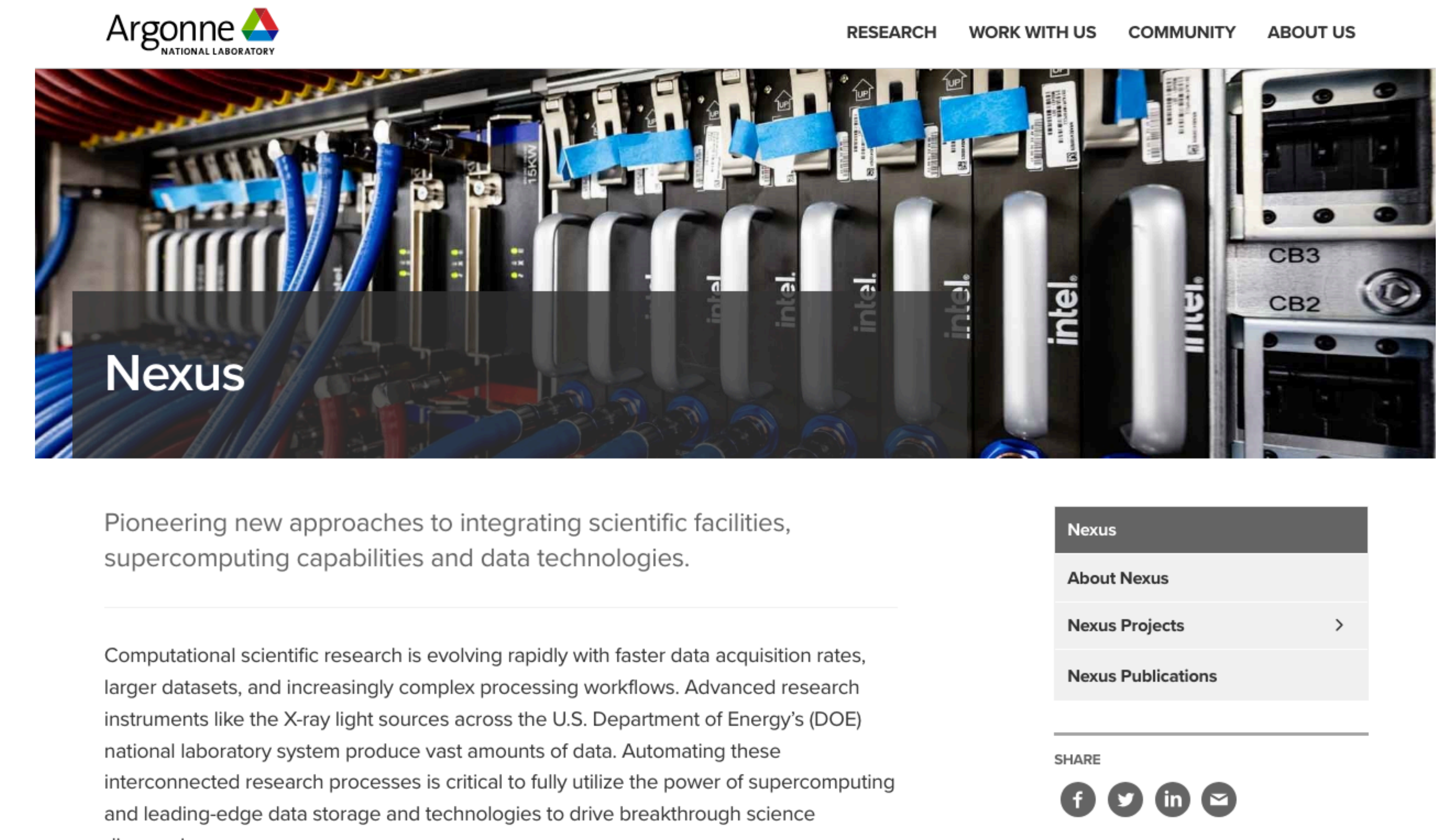
Resource co-operations practice is focused on creating new modes of cooperation, collaboration, co-scheduling, and joint planning across facilities and DOE programs.

Cybersecurity and federated access practice is focused on creating novel solutions that enable seamless scientific collaboration within a secure and trusted IRI ecosystem.

Workflows, interfaces, and automation practice is focused on creating novel solutions that facilitate the dynamic assembly of components across facilities into end-to-end IRI pipelines.

Scientific data life cycle practice is focused on ensuring that users can manage their data and metadata across facilities from inception to curation, archiving, dissemination, and publication.

Portable/scalable solutions practice is focused on ensuring that transitions can be made across heterogeneous facilities (portability) and from smaller to larger resources (scalability).



Argonne NATIONAL LABORATORY

RESEARCH WORK WITH US COMMUNITY ABOUT US

Nexus

Pioneering new approaches to integrating scientific facilities, supercomputing capabilities and data technologies.

Computational scientific research is evolving rapidly with faster data acquisition rates, larger datasets, and increasingly complex processing workflows. Advanced research instruments like the X-ray light sources across the U.S. Department of Energy's (DOE) national laboratory system produce vast amounts of data. Automating these interconnected research processes is critical to fully utilize the power of supercomputing and leading-edge data storage and technologies to drive breakthrough science

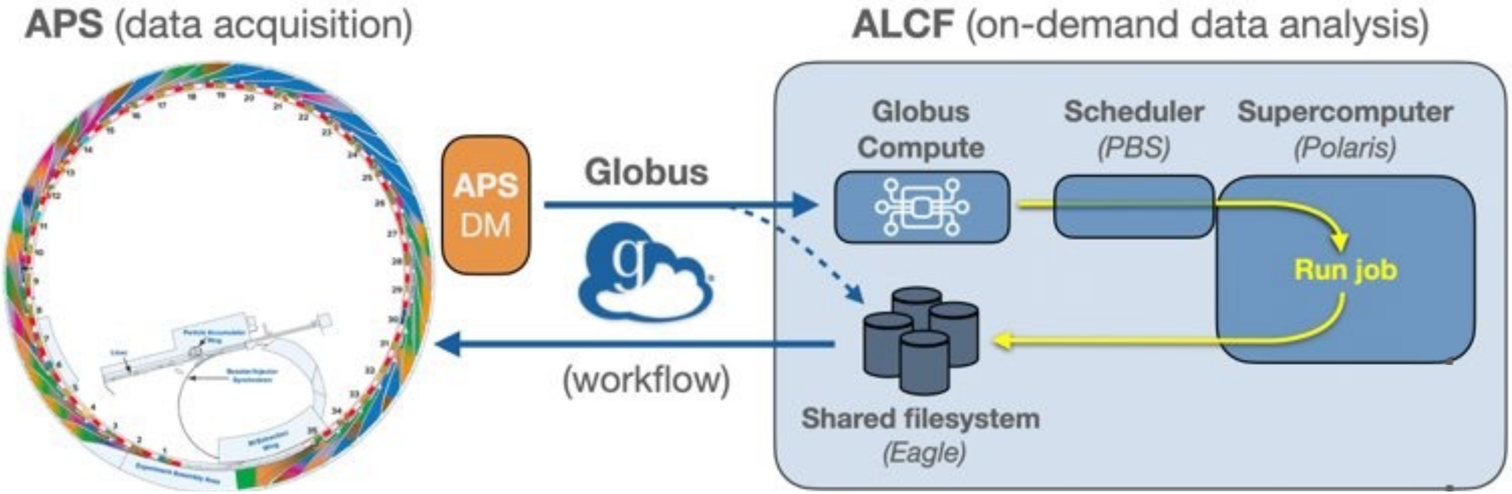
ARGONNE NEXUS

- The **Nexus** initiative at Argonne enables experimental facilities to leverage supercomputing facilities for experiment-time data analysis
- Nexus goals align with the DOE vision for integrated research infrastructure (IRI)
  - Simplify access to DOE computing resources (accounts, job submission)
  - Automate data transfer and workflows triggered by experiment events
  - Provide robust, generalizable solutions, not tailored to a single experiment/facility
  - Support interactive inspection and provenance of data and derived products
  - Publish data for community access, citation, and archiving

ARGONNE NEXUS SERVICES

- Demand Queue
    - Deployed now on Polaris: reduces queue wait time for experiment-time analysis; backfilled by preemptable jobs
- Service Accounts
    - Provides experiment-specific accounts for running analyses in a controlled environment
- Eagle Data Sharing (100PB filesystem)
    - Users can define collections of data to share publicly or with designated Globus users, without an ALCF account
    - Leveraged in ALCF-HEP cosmology data sharing portal
- Globus Infrastructure
    - Distributed compute, transfer, and web-based monitoring
- ALCF Community Data Cooperative (ACDC)
    - Layer atop Eagle data sharing to enable more sophisticated metadata-based navigation and search
    - Metadata extraction and capture services
- Dedicated Web Applications
    - Foundation for user-driven access to data and analysis
    - Developing in collaborations with APS and Argonne-HEP for extension to other science areas

ARGONNE NEXUS: LIGHTSOURCE AUTOMATION



- Integration with the data management (DM) system at APS allows workflow to begin as soon as data is taken
- Workflow moves data from the APS beamline to ALCF and submits job to demand queue on Polaris
- Results are written to Eagle, where they're reachable via Jupyter, and also returned to APS for evaluation

One-time configuration at ALCF | **APS experiments**

Using user accounts | **Humans involved in the workflows**

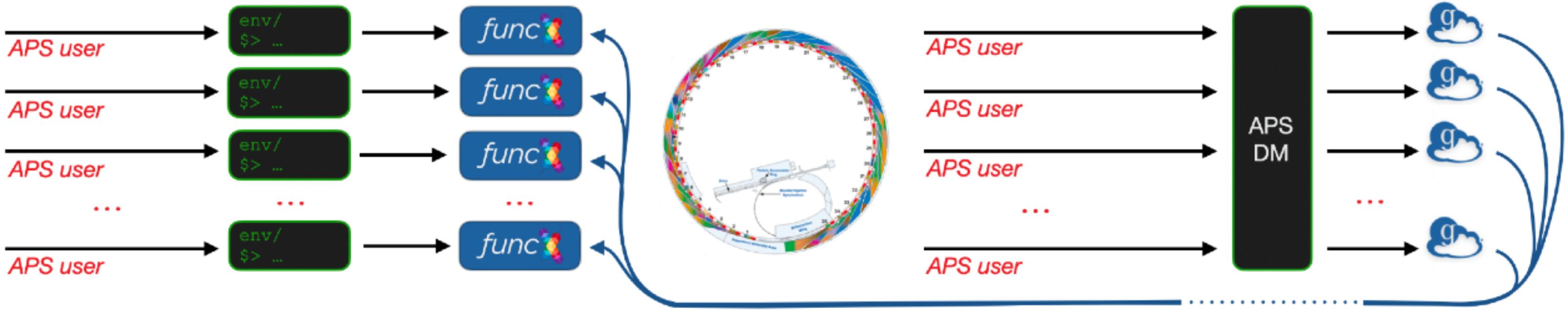
APS requests user accounts at ALCF

ALCF creates user accounts

*APS users setup their **personal environment** and funcX endpoint*

*APS users log into DM system using **personal credentials***

*DM system starts Globus flows involving **specific funcX endpoints***



One-time configuration at ALCF | **APS experiments**

Using a service account per beam line | **No human involved in the workflows**

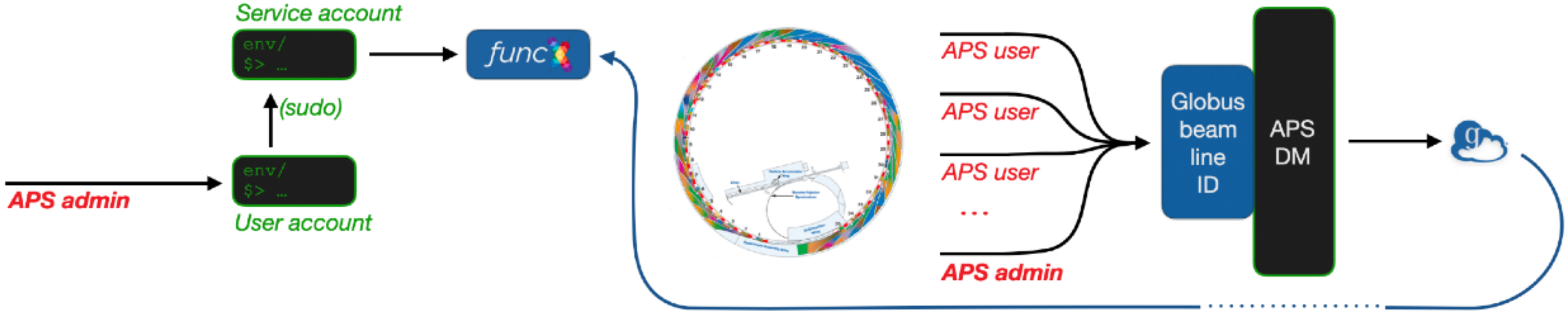
APS requests service account at ALCF

ALCF creates service account

*APS admin user setups a **shared environment** and funcX endpoint*

*APS users log into DM system using **shared Globus identity***

*DM system starts Globus flows involving **shared funcX endpoint***



NEXUS SUPPORTS SEVERAL APS BEAMLINES

Several APS beamlines have run production experiments with Globus Compute

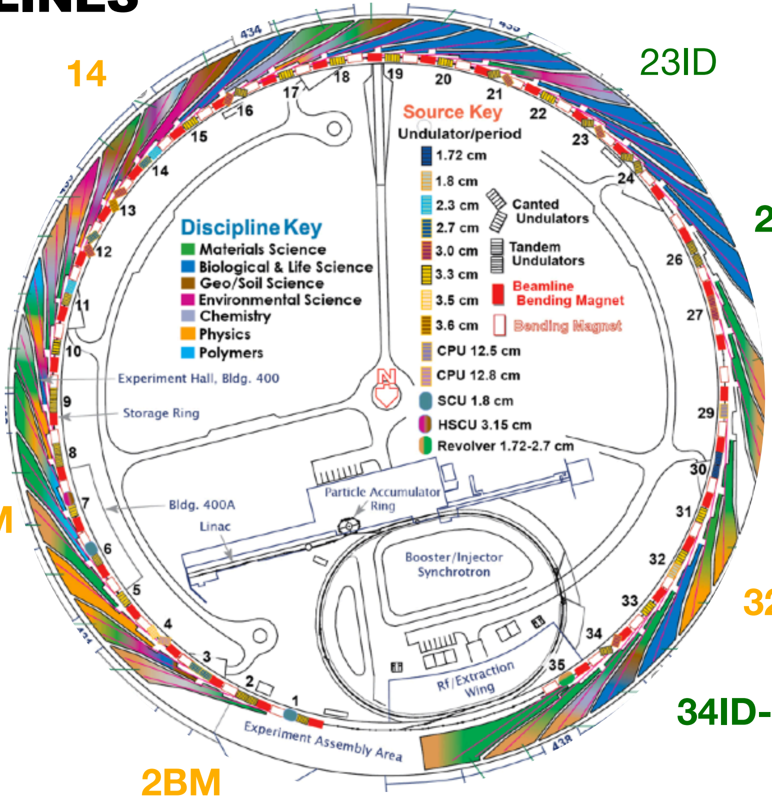
- 8ID-I: X-ray Protein Cystal Spectroscopy (XPCS)
- 19ID: Serial Crystallography (SSX) at the structural biology center (SBC), using DIALS for crystallography analysis
- 1ID: High energy diffraction microscopy using MIDAS for tomography and ptychography
- 34ID-E: Laue Micro-diffraction spectroscopy
- 26ID: Ptychography with Ptychodus and Tike

Other beamlines are running but have not yet run in production

- 23ID: GM/CA using crystfel toolset (+AlphaFold)
- 2ID-D: X-ray fluorescence image processing using XRF-Maps

Ramping up

- 14: BioCARS using crystfel
- 2BM, 7BM, 32ID: Tomography using tomocupy



19ID

23ID

14

26ID

8ID-I

7BM

32ID

34ID-E

2BM

2ID-D

1ID

running, in production

running, ready for production

starting up

Leveraging Globus Tools

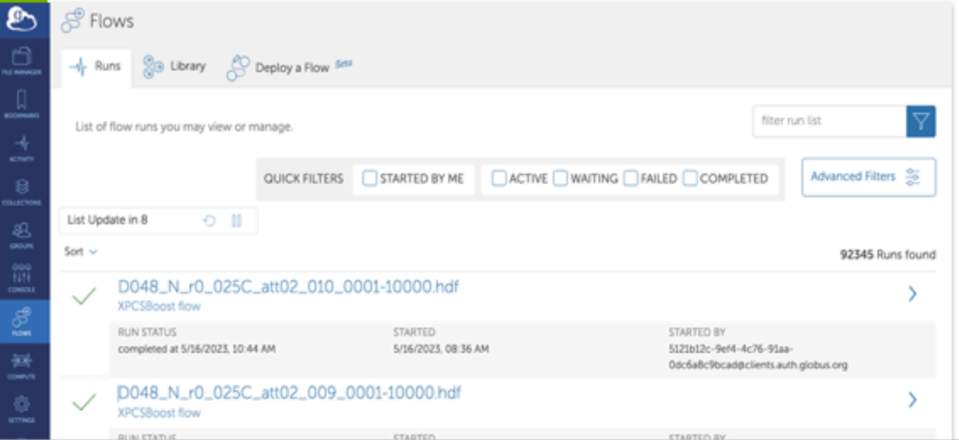
Compute: A managed service that implements a universal computing fabric

Local **Transfer** and **Compute** agents provide global footprint for actions

Auth provides federated identities and distributed authorization with delegation

Flows orchestrates actions across the computing continuum

UIs to monitor, manage, inspect it all



List of flow runs you may view or manage.

List Update in 8

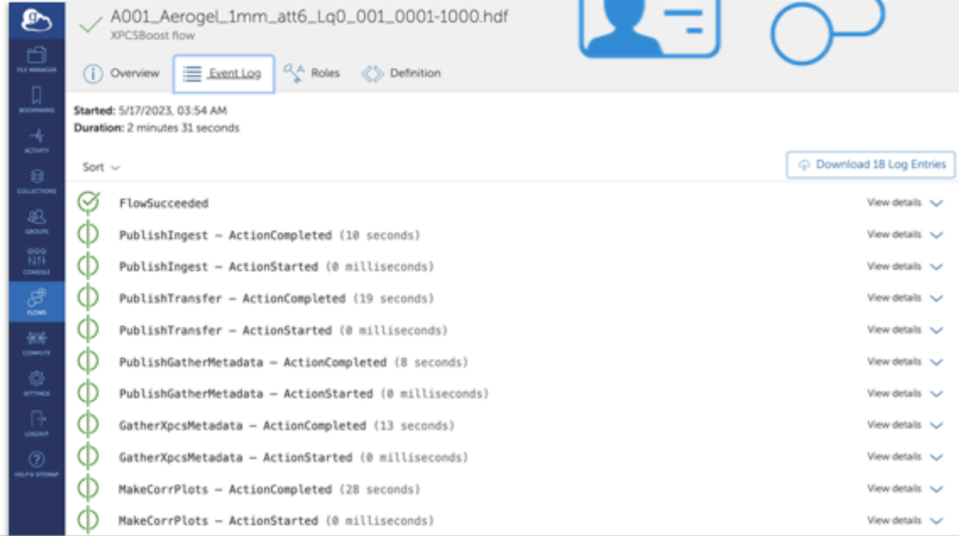
92345 Runs found

✓ D048_N_r0_025C_att02_010_0001-10000.hdf
XPCSBoost flow

RUN STATUS	STARTED	STARTED BY
completed at 5/16/2023, 10:44 AM	5/16/2023, 08:36 AM	5121b12c-9ef4-4c76-91aa-0dc6a8c9bcad@clients.auth.globus.org

✓ D048_N_r0_025C_att02_009_0001-10000.hdf
XPCSBoost flow

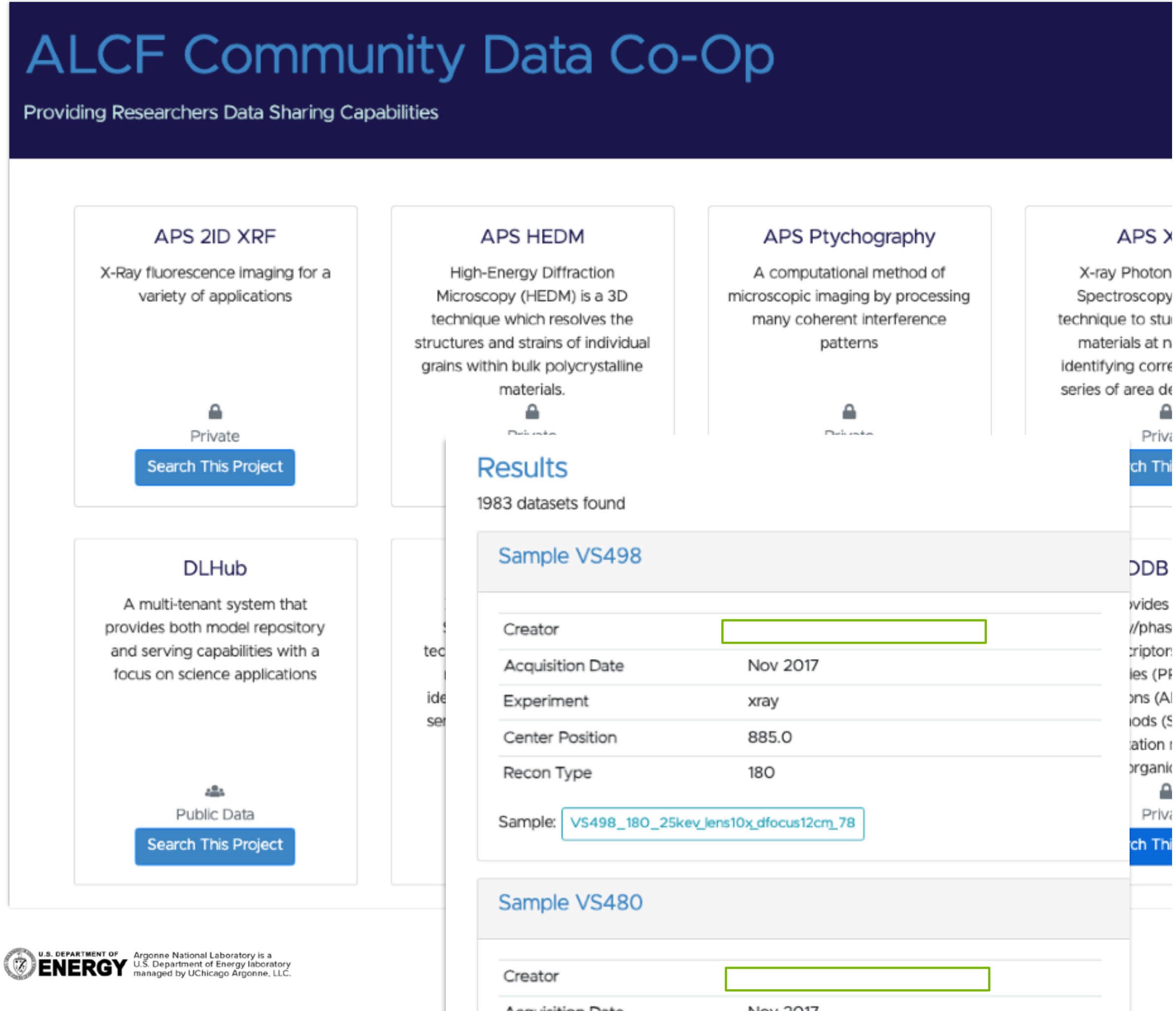
RUN STATUS STARTED STARTED BY



✓ A001_Aerogel_1mm_att6_Lq0_001_0001-1000.hdf
XPCSBoost flow

Started: 5/17/2023, 03:54 AM
Duration: 2 minutes 31 seconds

```
FlowSucceeded                                                  View details
PublishIngest – ActionCompleted (10 seconds)                   View details
PublishIngest – ActionStarted (0 milliseconds)                 View details
PublishTransfer – ActionCompleted (19 seconds)                 View details
PublishTransfer – ActionStarted (0 milliseconds)               View details
PublishGatherMetadata – ActionCompleted (8 seconds)            View details
PublishGatherMetadata – ActionStarted (0 milliseconds)         View details
GatherXpcsMetadata – ActionCompleted (13 seconds)              View details
GatherXpcsMetadata – ActionStarted (0 milliseconds)            View details
MakeCorrPlots – ActionCompleted (28 seconds)                   View details
MakeCorrPlots – ActionStarted (0 milliseconds)                 View details
```



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.

APS Ptychography

A computational method of microscopic imaging by processing many coherent interference patterns

APS X

X-ray Photon
Spectroscopy
technique to stu
materials at n
identifying corre
series of area de

Priva

ch Thi

DLHub

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

Public Data

Search This Project

DDB

Results

1983 datasets found

Sample VS498

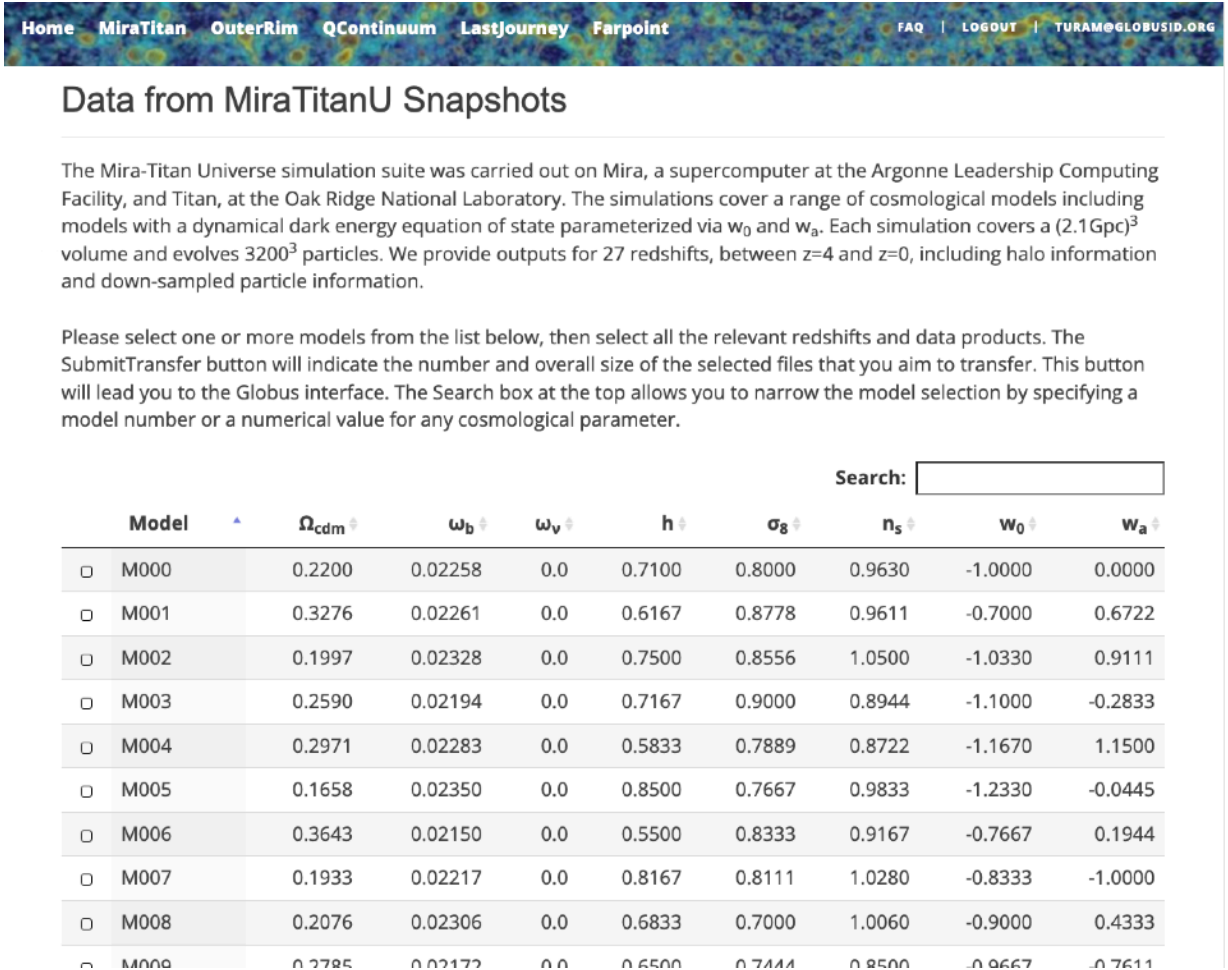
Creator	
Acquisition Date	Nov 2017
Experiment	xray
Center Position	885.0
Recon Type	180

Sample: VS498_180_25kev_lens10x_dfocus12cm_78

Sample VS480

Creator	

- ACDC provides data publication interface backed by Eagle filesystem
- Data tagged with application metadata to facilitate searching
- Data-specific plots and visualizations can be included via templates
- One-click access to transfer datasets via Globus



Data from MiraTitanU Snapshots

The Mira-Titan Universe simulation suite was carried out on Mira, a supercomputer at the Argonne Leadership Computing Facility, and Titan, at the Oak Ridge National Laboratory. The simulations cover a range of cosmological models including models with a dynamical dark energy equation of state parameterized via w_0 and w_a. Each simulation covers a $(2.1\text{Gpc})^3$ volume and evolves 3200^3 particles. We provide outputs for 27 redshifts, between z=4 and z=0, including halo information and down-sampled particle information.

Please select one or more models from the list below, then select all the relevant redshifts and data products. The SubmitTransfer button will indicate the number and overall size of the selected files that you aim to transfer. This button will lead you to the Globus interface. The Search box at the top allows you to narrow the model selection by specifying a model number or a numerical value for any cosmological parameter.

Search:

	Model	Ω_{cdm}	ω_b	ω_ν	h	σ_8	n_s	w_0	w_a
□	M000	0.2200	0.02258	0.0	0.7100	0.8000	0.9630	-1.0000	0.0000
□	M001	0.3276	0.02261	0.0	0.6167	0.8778	0.9611	-0.7000	0.6722
□	M002	0.1997	0.02328	0.0	0.7500	0.8556	1.0500	-1.0330	0.9111
□	M003	0.2590	0.02194	0.0	0.7167	0.9000	0.8944	-1.1000	-0.2833
□	M004	0.2971	0.02283	0.0	0.5833	0.7889	0.8722	-1.1670	1.1500
□	M005	0.1658	0.02350	0.0	0.8500	0.7667	0.9833	-1.2330	-0.0445
□	M006	0.3643	0.02150	0.0	0.5500	0.8333	0.9167	-0.7667	0.1944
□	M007	0.1933	0.02217	0.0	0.8167	0.8111	1.0280	-0.8333	-1.0000
□	M008	0.2076	0.02306	0.0	0.6833	0.7000	1.0060	-0.9000	0.4333
□	M009	0.2785	0.02172	0.0	0.6500	0.7444	0.8500	-0.9667	-0.7611

- Nexus infrastructure aims to include support for project-developed data applications
- Data catalogs resident on Eagle filesystem, searchable via user-provided metadata
- Community analysis of data supported via back-end integration of workflows on Polaris
- Reusable components can be deployed and customized for future science domains

IRI BLUEPRINT ACTIVITY REPORT (2023)

U.S. DEPARTMENT OF ENERGY | Office of Science

THE DOE OFFICE OF SCIENCE

Integrated Research Infrastructure Architecture Blueprint Activity

FINAL REPORT

2023

Questions...

- about Nexus? We are looking to expand into new experimental and observational science areas, and would like to talk with you.
- about IRI? This DOE-led effort is formulating what future infrastructure should look like, and will be better informed if you describe your science case.