Data Management and Automation at the APS

2013 Three-Way Meeting

1 August 2013

Nicholas Schwarz

Principal Computer Scientist / Group Leader

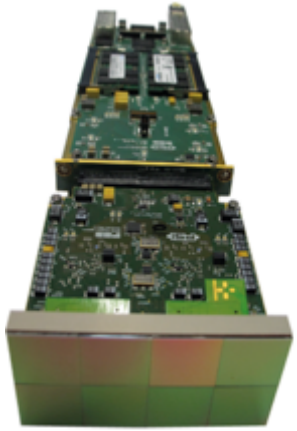
Software Services Group

APS Engineering Support Division (AES)

Advanced Photon Source (APS)

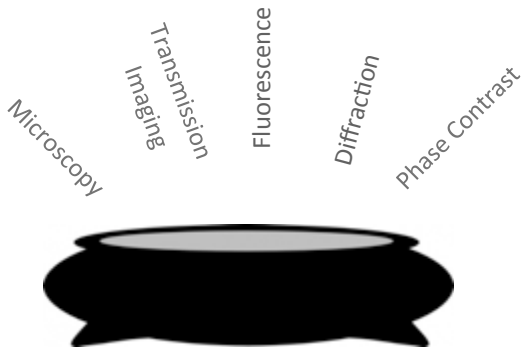
Increasing Amounts and Complexity of Data

Detectors

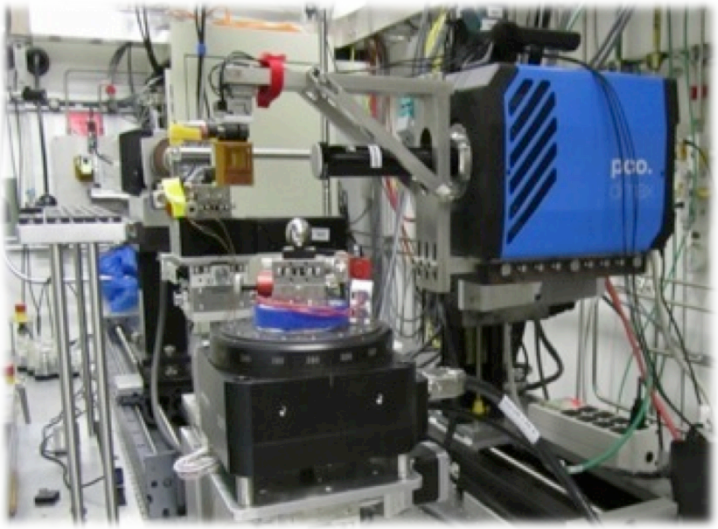


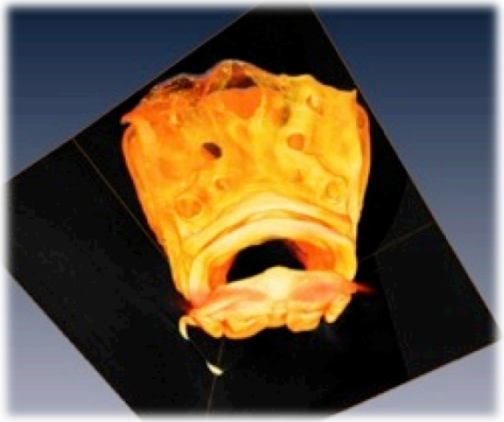
Experiment Automation

Techniques



Fast Micro-Tomography





Keith Chang (Penn State College of Medicine)

Beamline and Detector

- 2-BM
- pco.dimax (36GB of onboard RAM)
- 2K X 2K @ 16-bits

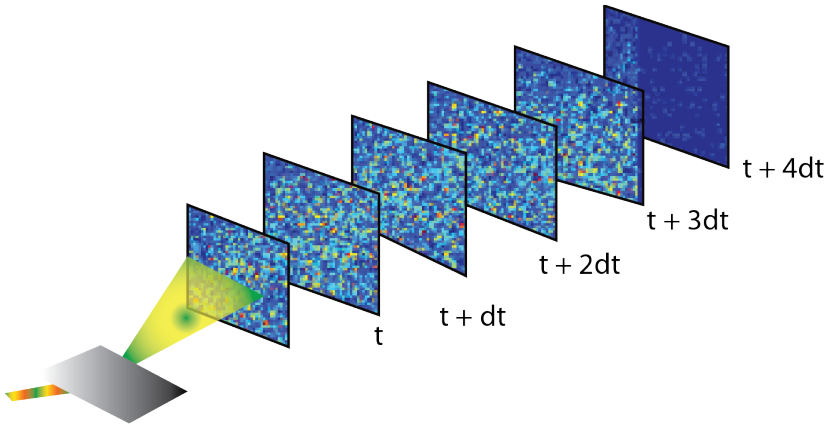
Tomography Dataset

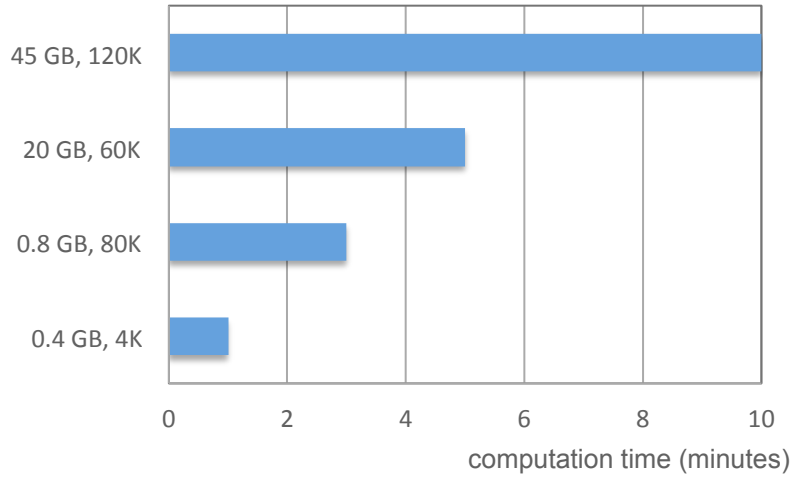
- Projections around 180 degrees
- ~1,500 FPS
- A few seconds of data collection
- ~1 minute readout
- **Up to 11GB raw data per minute**

Reconstruction

- Takes ~5 minutes
- Reconstructed data ~22GB

X-ray Photon Correlation Spectroscopy





Beamline

- 8-ID
- Kinetics measurements

XPCS Data

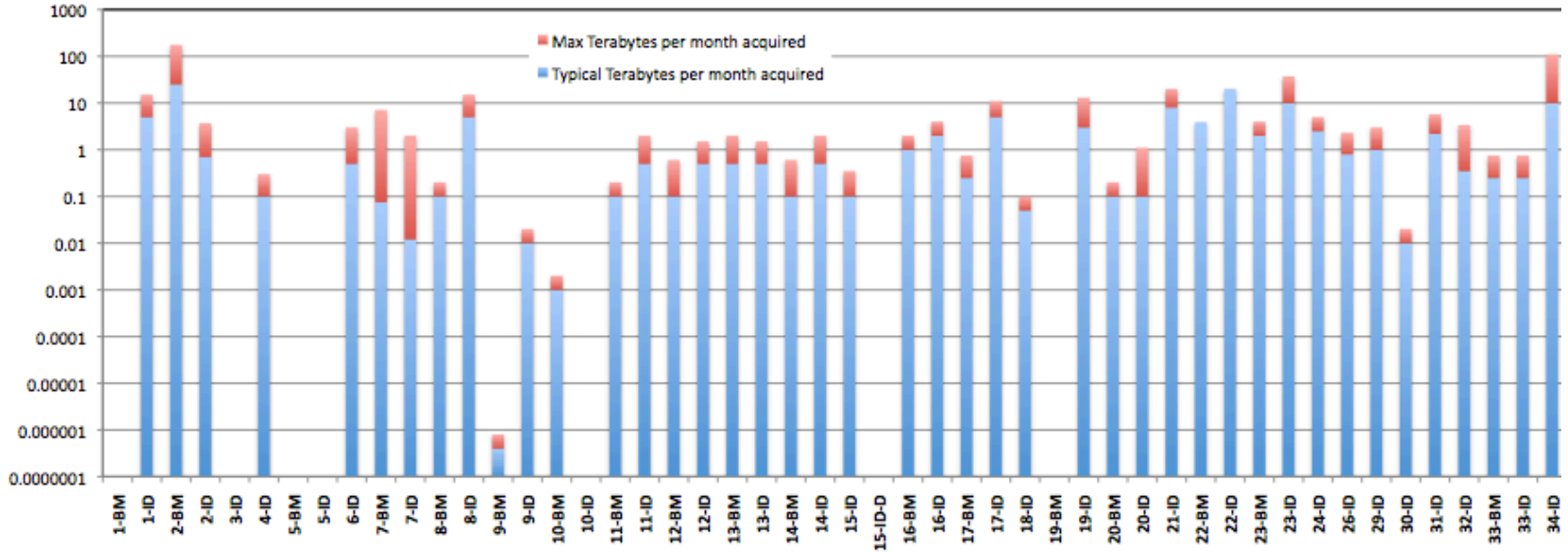
- 2D images over time
- A few minutes to collect
- Varies from 0.5GB to 50GB

Reduction

- Multi-tau autocorrelation

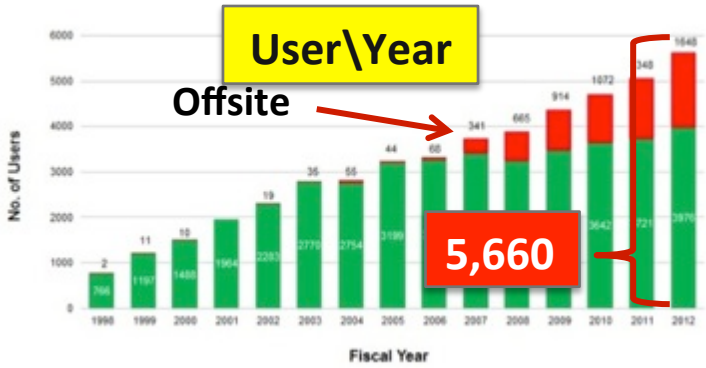
APS Average Data Rate

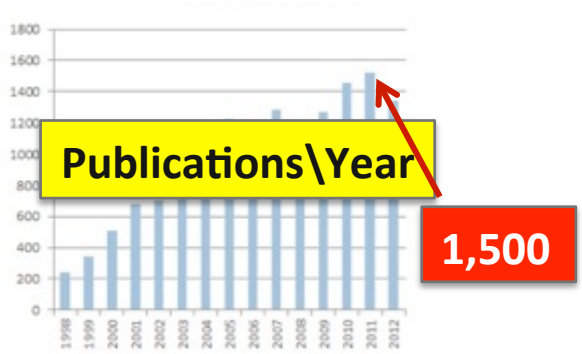
Cumulative typical data volume at present: 112 TB/month
Cumulative maximum data volume: 368 TB/month



F. De Carlo (ANL)

Increasing User Base





Goals / Constraints

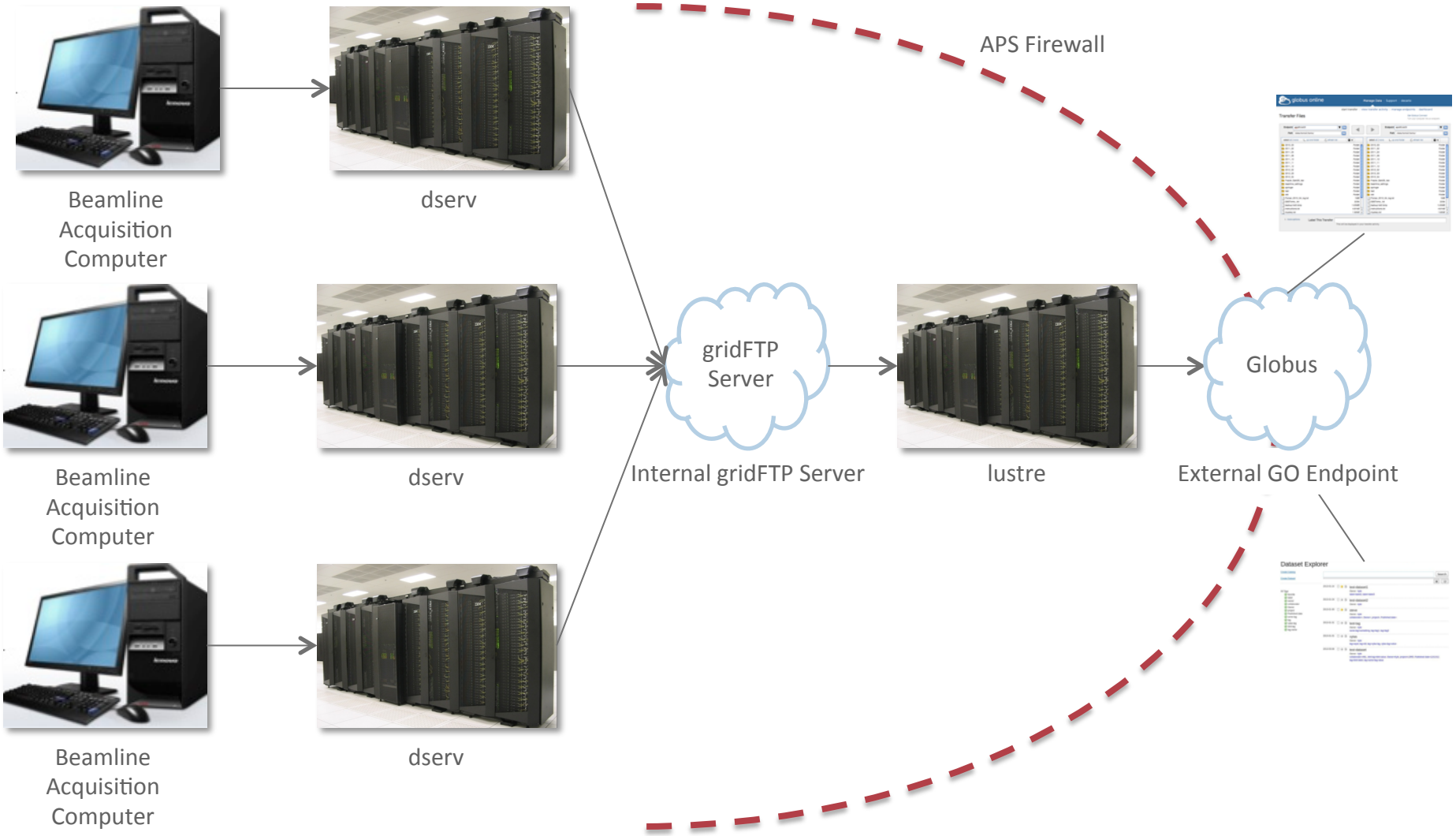
Goals

- Store detector data temporarily
- Reduce (analyze) data after acquisition
- Keep data for some moderate amount of time
- Facilitate ability of users to take their data home

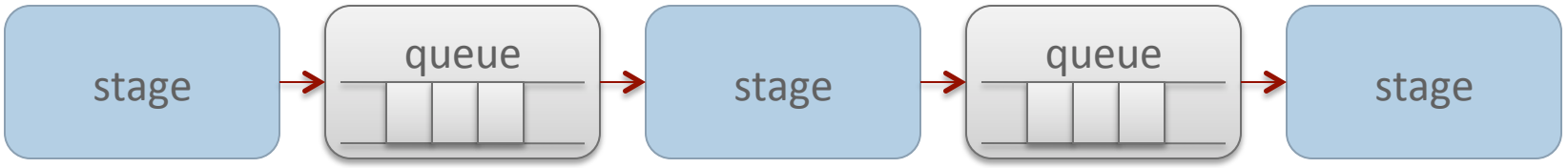
Constraints

- Acquisition must be as robust as possible
- Institution policy (e.g. firewalls, ownership, etc.)

Architecture



Workflow Pipeline

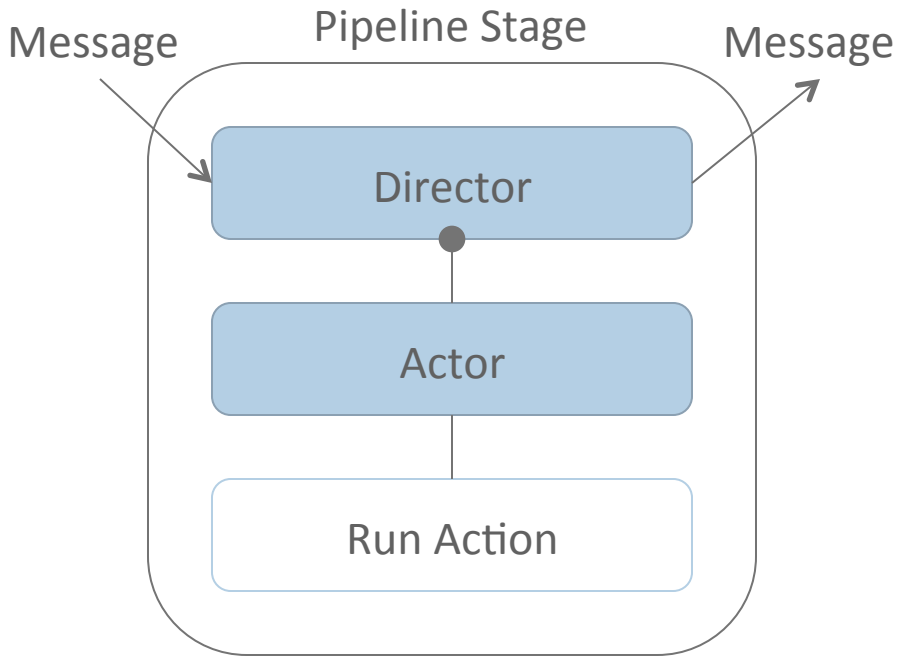


- Series of actors connected by message queues
- Stage
  - Acquisition
  - Run data analysis
  - Transfer files
  - Many languages: C++, Java, Python
- Message queues
  - Pass messages from one actor to the next
  - JMS message queues (ActiveMQ)

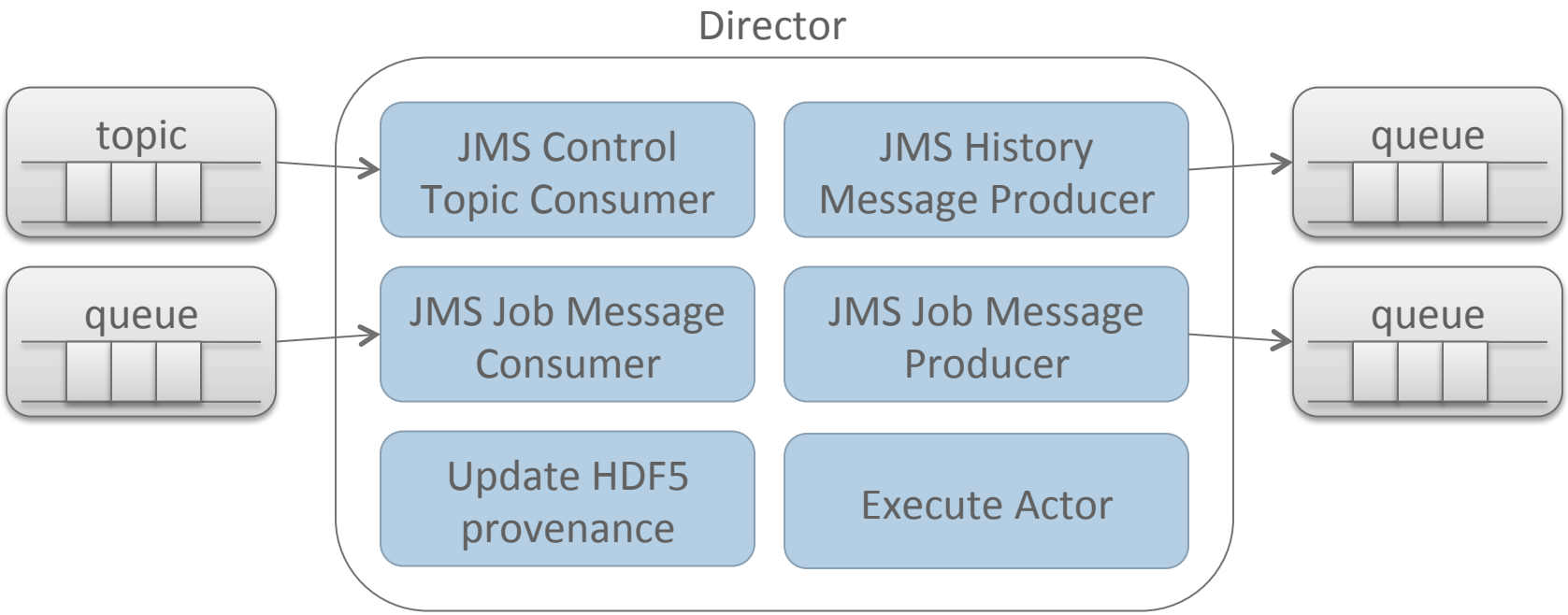
Pipeline Stages

Stages are composed of two separate classes:

- Director
  - interface with the message broker
- Actor
  - run scan / analysis / file transfer
  - save results
  - report status



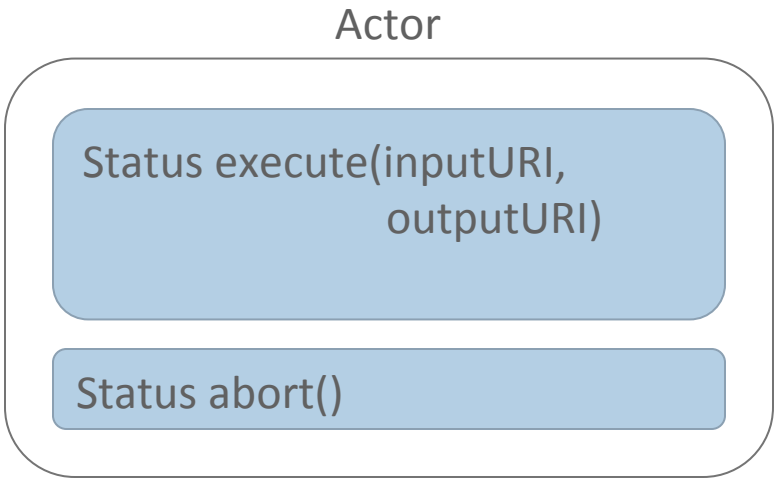
Director



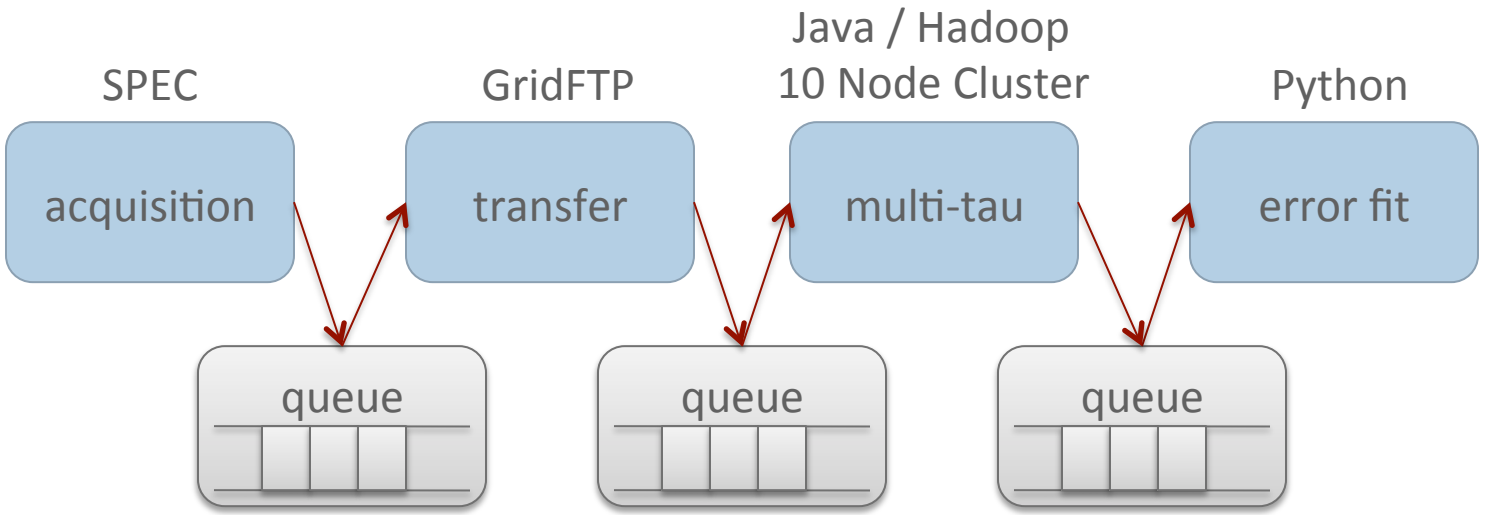
- Broker interface
- Handle job messages
- Control messages
- Update history
- Maintain provenance
- Execute Actor

Actor

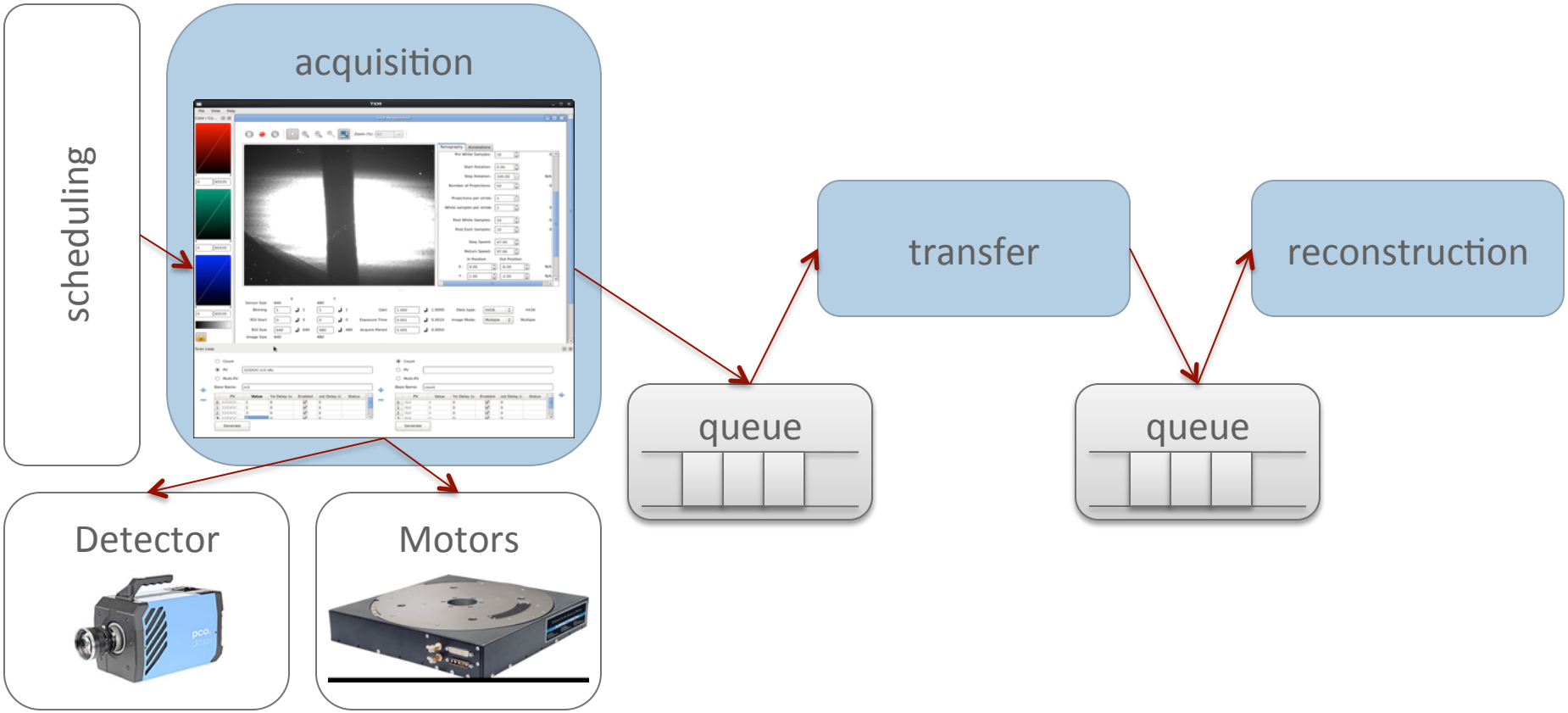
- Actor
  - Called by Director
  - Abstract interface for concrete implementation
  - Execute job / analysis
  - Report status
- Analysis
  - External application
  - Internal code
  - Read input data from HDF5
  - Write output data to HDF5



Use Case: XPCS



Use Case: Tomography

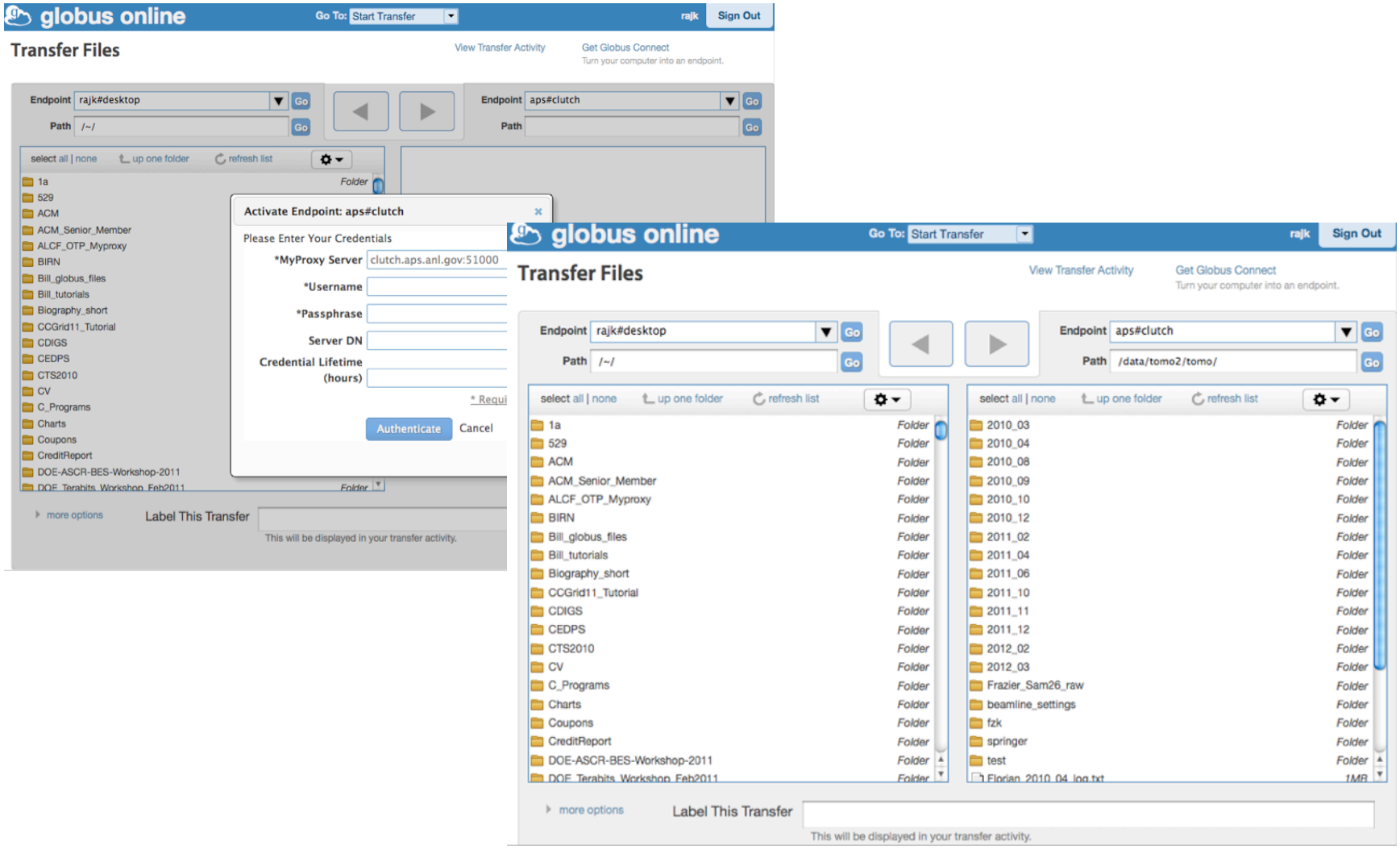


User Access to Data

Globus Online

- Provides access to APS data from the outside
- Used by many user facilities and supercomputing centers
- Optimized bandwidth utilization for faster transfers
- www.globusonline.org

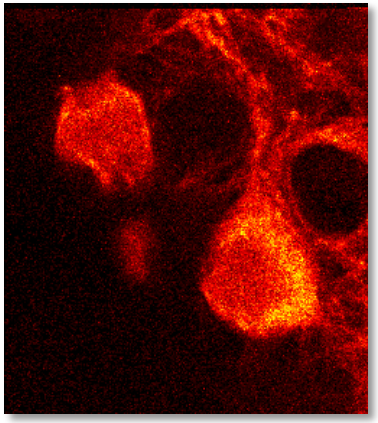
Globus Online



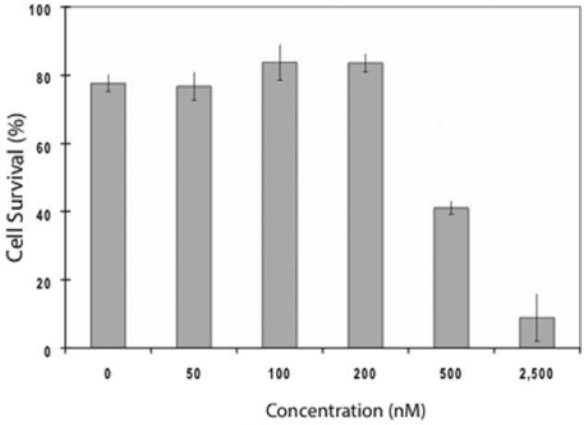
User / Scientist Perspective

Publication

- Multiple figures
- Different types of data



Laboratory Microscope Data

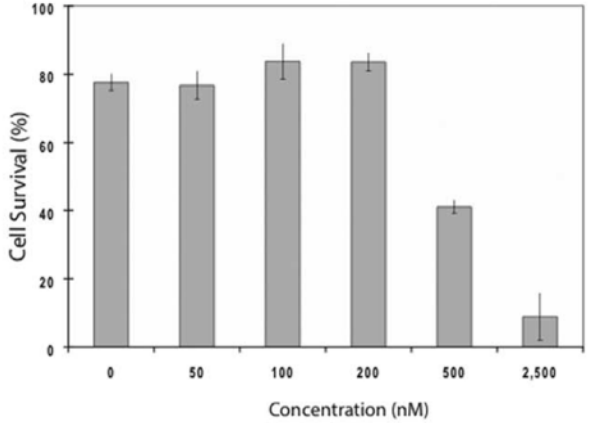


Synchrotron Derived Data

User / Scientist Perspective

Even a single figure with synchrotron data may have data from multiple facilities.

Normalize Intensity
Cell Finding Algorithm
Data Fusion

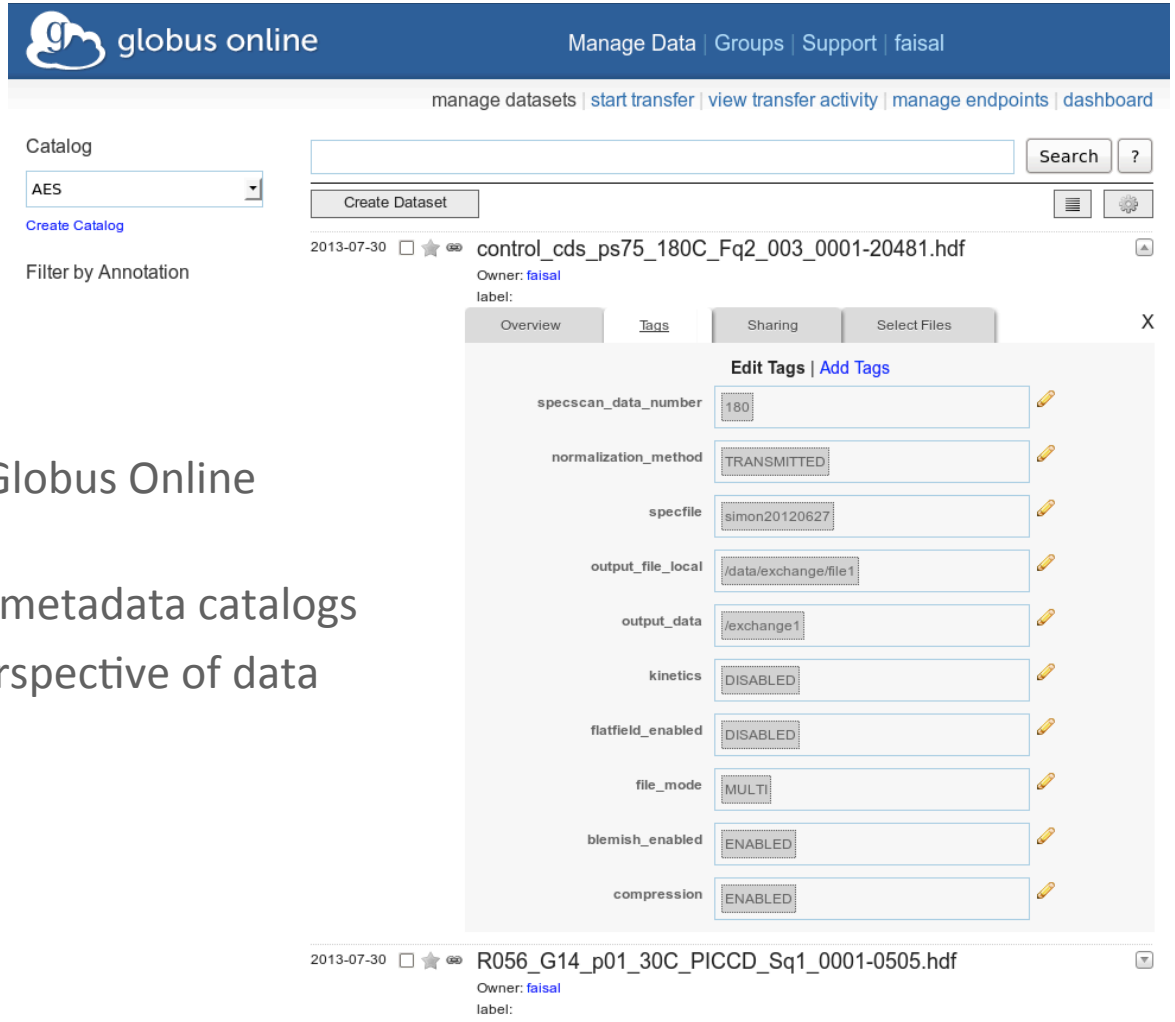


Synchrotron Derived Data

Process of analyzing data generates new knowledge and data (and metadata).

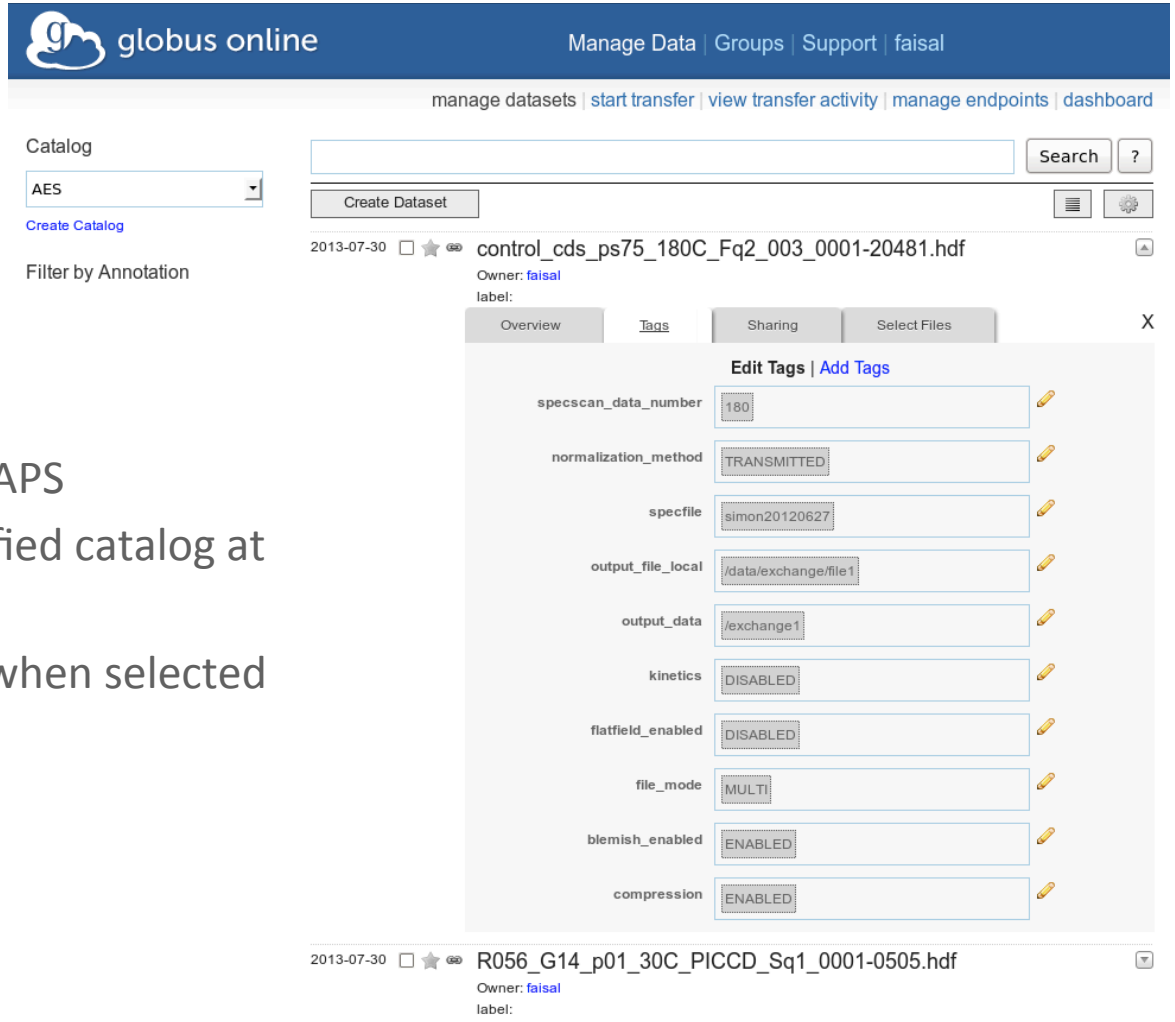
Globus Online Catalog

- Bridge across multiple Globus Online endpoints
- User builds meaningful metadata catalogs
- Not a facility-centric perspective of data



Globus Online Catalog

- Pilot techniques at the APS
- Push data to user specified catalog at end of acquisition
- Read data to a catalog when selected



Acknowledgements

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

This work is supported by the U.S. Department of Energy, Basic Energy Sciences, Office of Science, under contract # DE-AC02-06CH11357.

Thank you!