

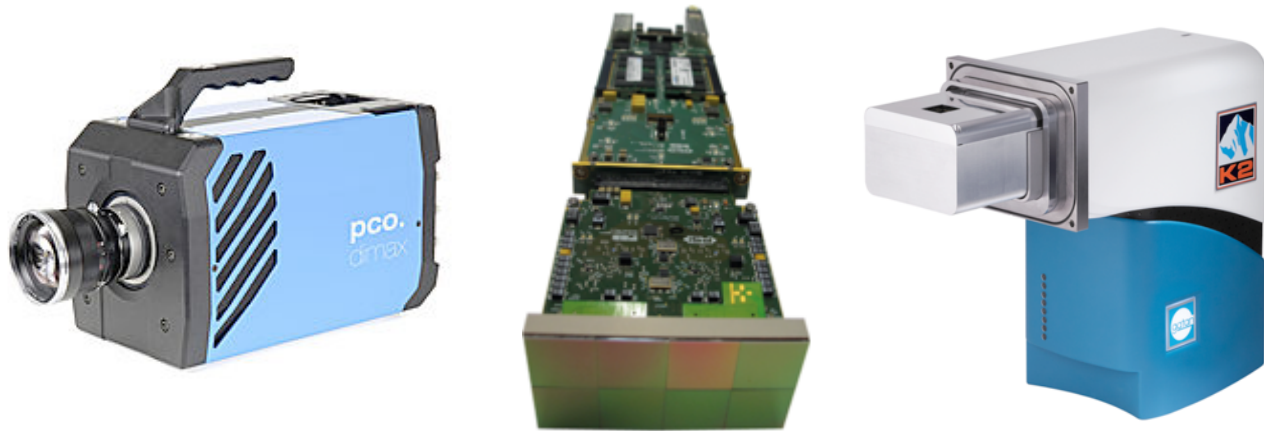
Data Management and Automation at the APS

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

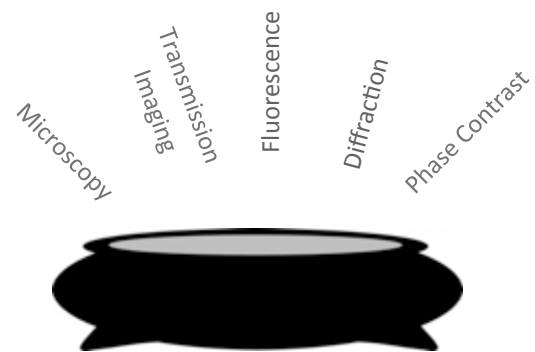
Detectors



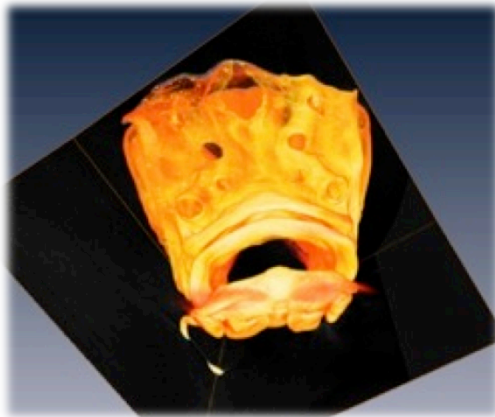
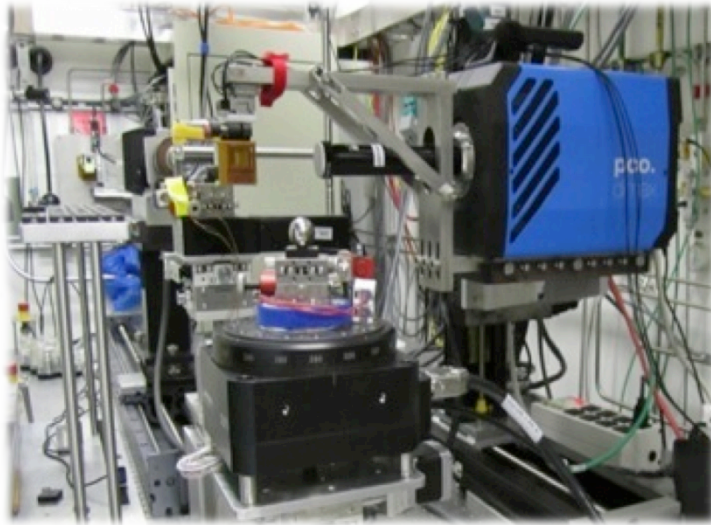
Experiment Automation



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

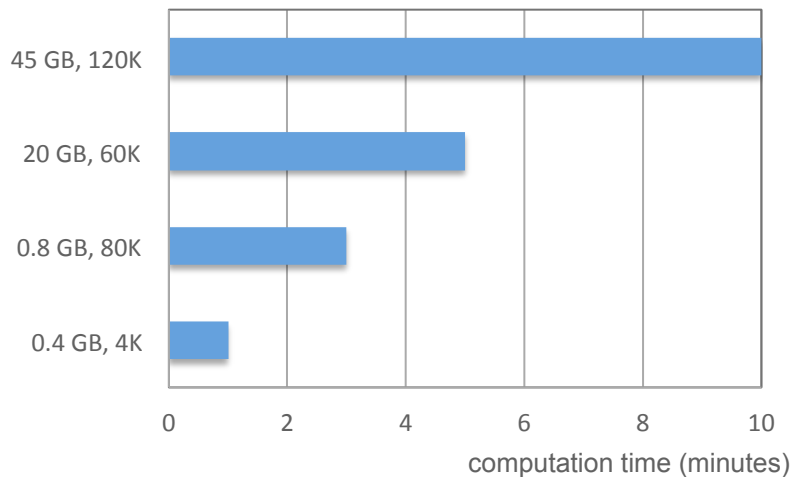
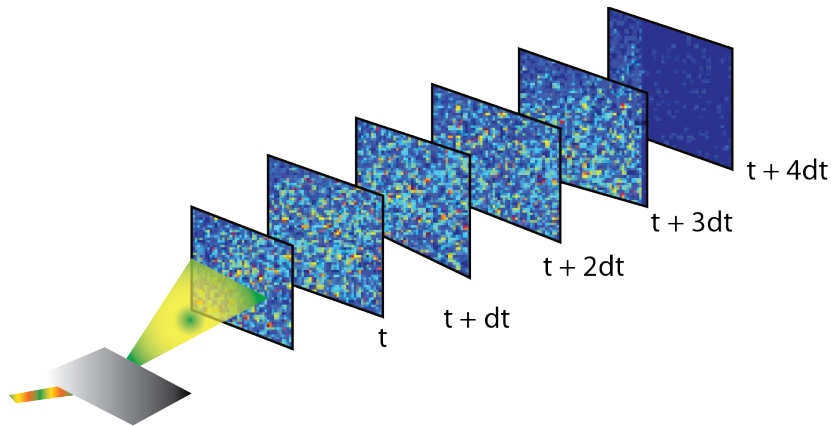
Tomography Dataset

- 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

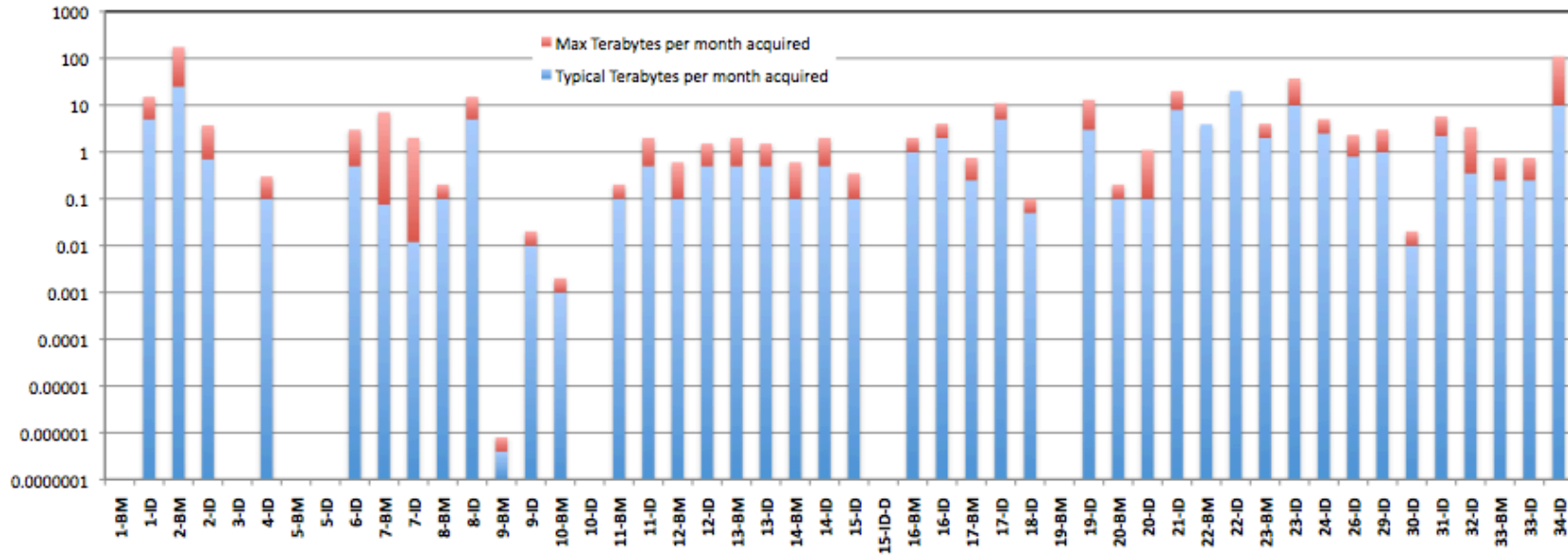
- 2D images over time
- A few minutes to collect
- Varies from 0.5GB to 50GB

Reduction

- Multi-tau autocorrelation

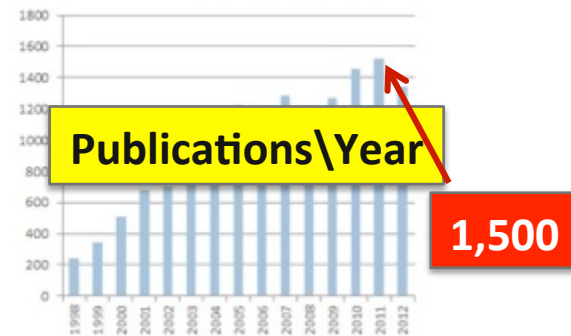
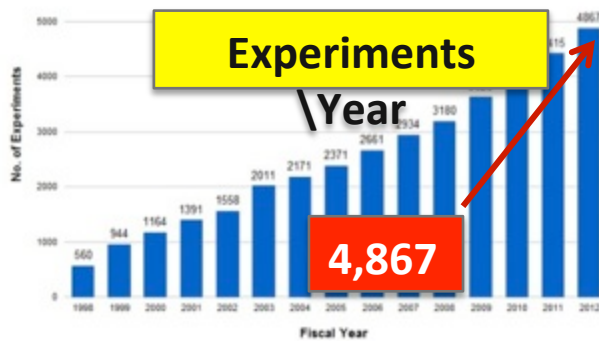
APS Average Data Rate

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

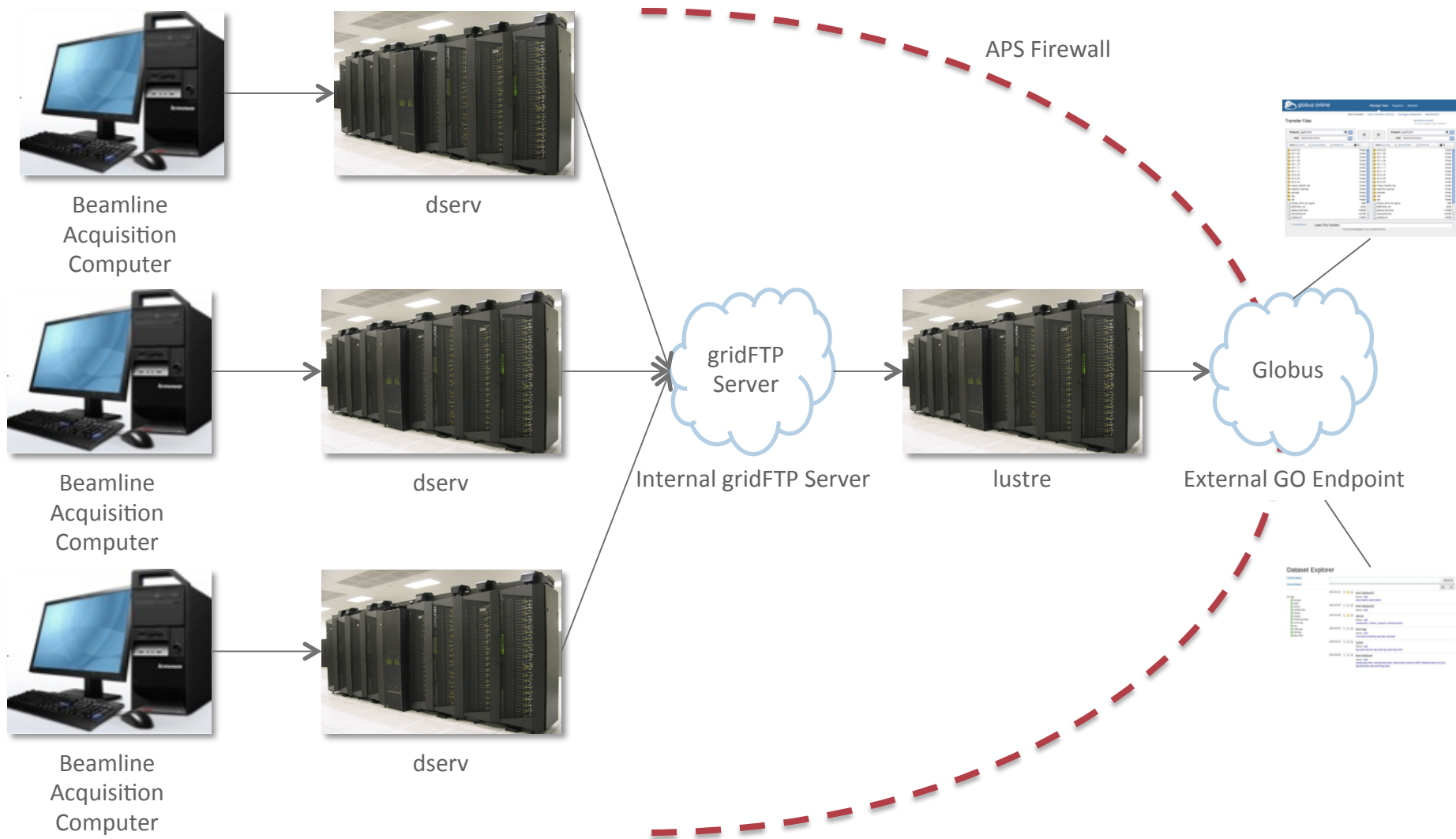
- Store detector data temporarily
- Reduce (analyze) data after acquisition
- Keep data for some moderate amount of time
- Facilitate ability of users to take their data home

Constraints

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



Architecture



Workflow Pipeline



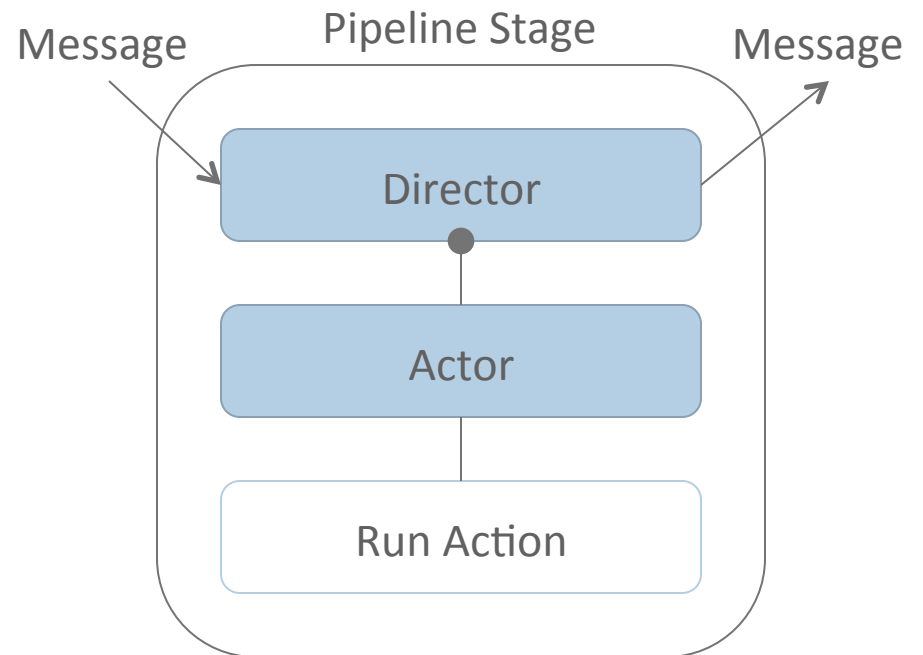
- Series of actors connected by message queues
- Stage
 - Acquisition
 - Run data analysis
 - Transfer files
 - Many languages: C++, Java, Python
- Message queues
 - Pass messages from one actor to the next
 - JMS message queues (ActiveMQ)



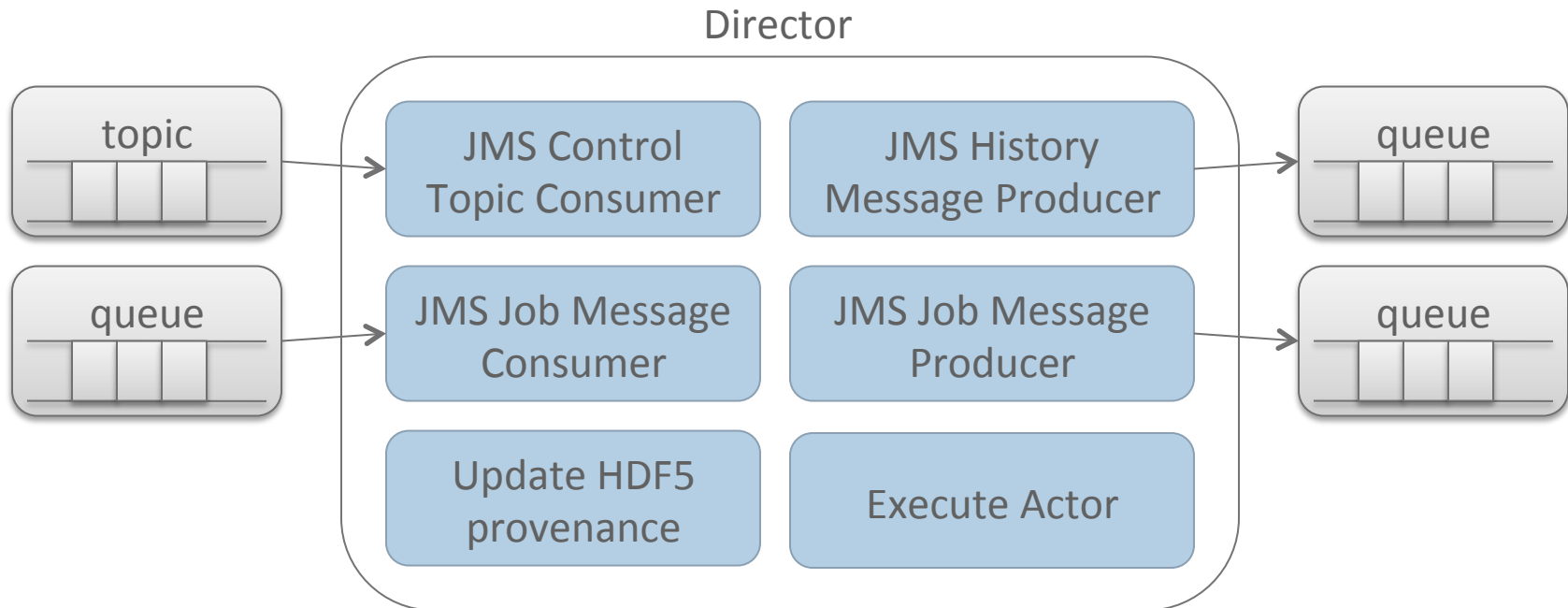
Pipeline Stages

Stages are composed of two separate classes:

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



Director

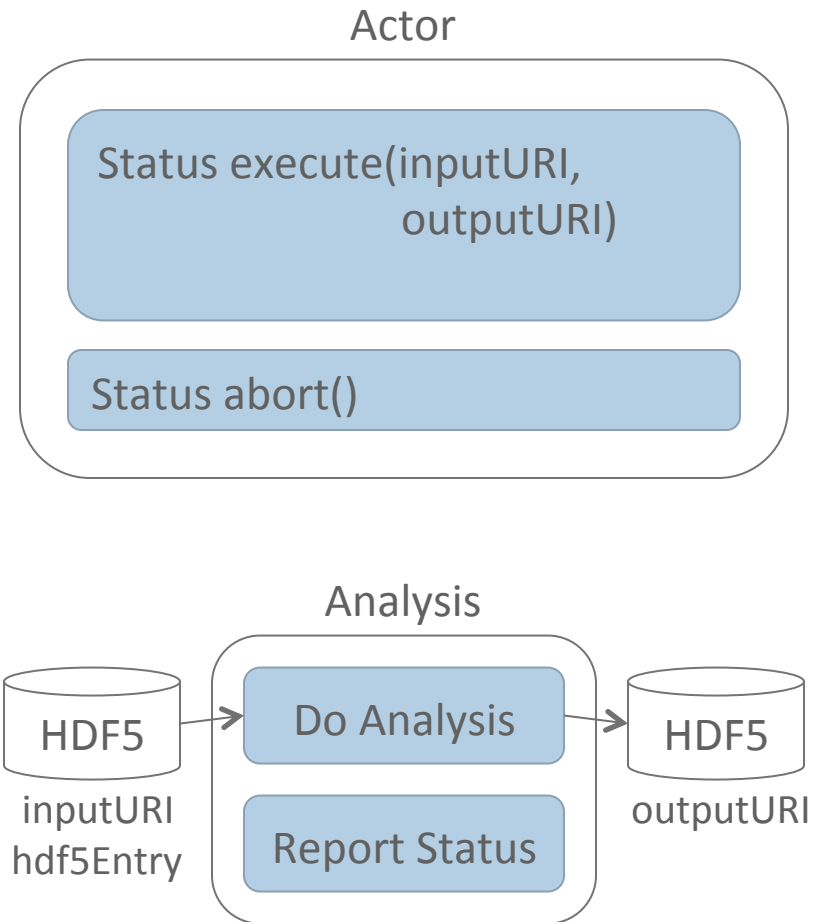


- Broker interface
- Handle job messages
- Control messages
- Update history
- Maintain provenance
- Execute Actor

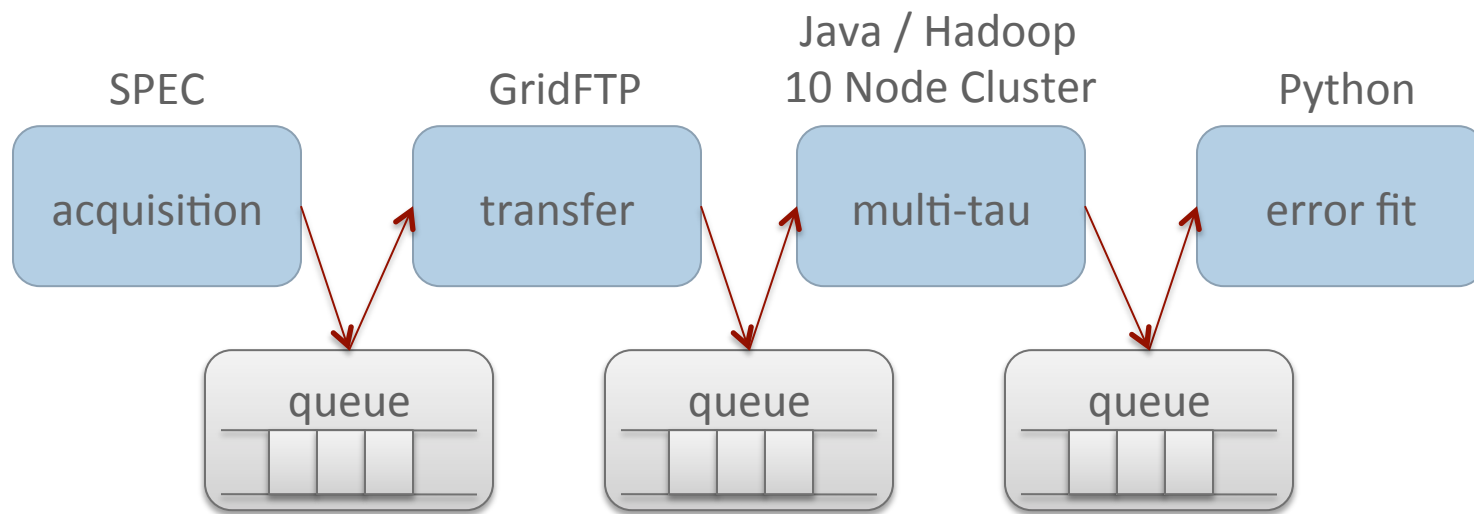
Actor

- Actor
 - Called by Director
 - Abstract interface for concrete implementation
 - Execute job / analysis
 - Report status

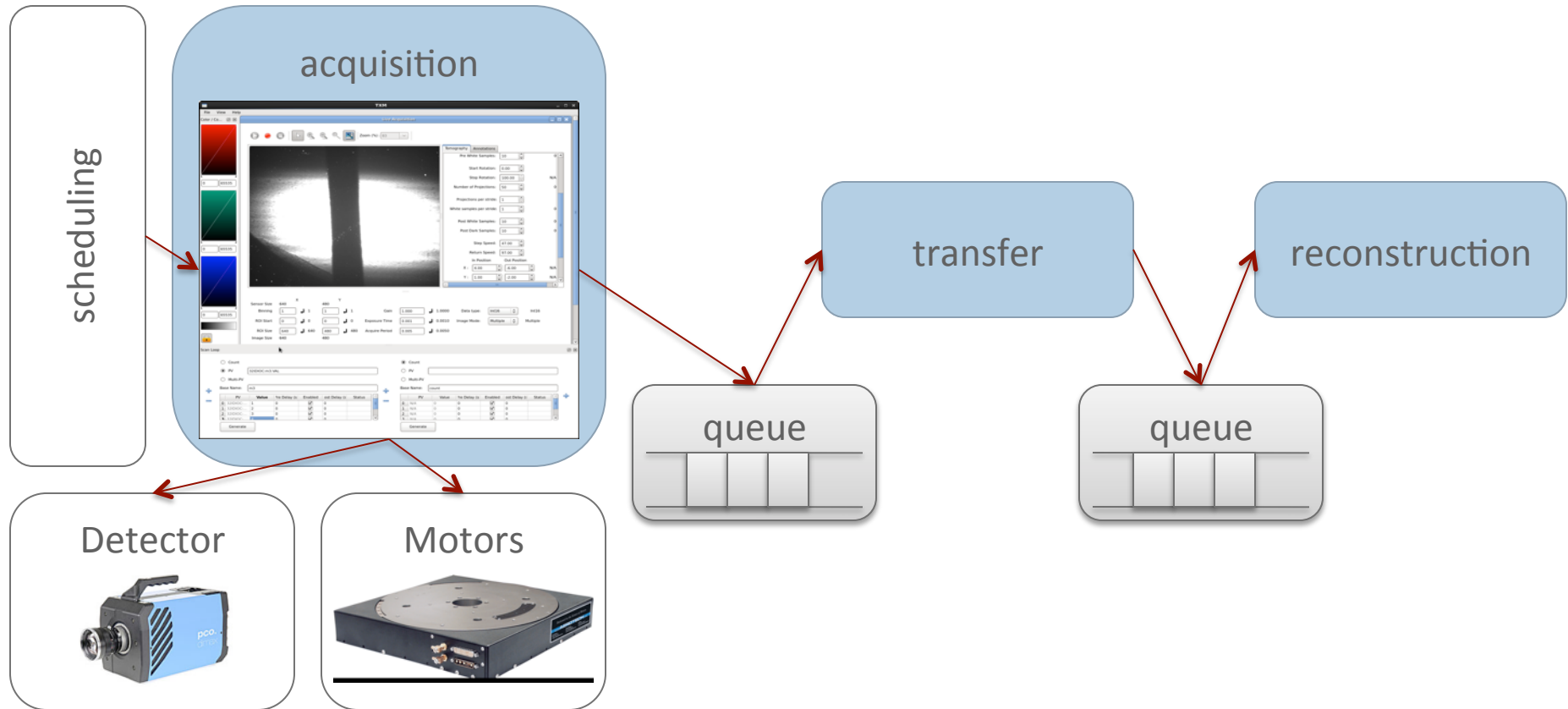
- Analysis
 - External application
 - Internal code
 - Read input data from HDF5
 - Write output data to HDF5



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

The image displays two screenshots of the Globus Online web interface. The top screenshot shows the 'Transfer Files' page with an 'Activate Endpoint: aps#clutch' dialog box overlaid. The dialog box prompts the user to enter credentials for the proxy server 'clutch.aps.anl.gov:51000', including fields for Username, Passphrase, Server DN, and Credential Lifetime (hours). The bottom screenshot shows the same 'Transfer Files' page, but with a file list for the 'aps#clutch' endpoint. The file list includes folders for years 2010 and 2011, and a file named 'Florian_2010_04_10a.txt' (1MB).

Top Screenshot: Transfer Files Page

- Endpoint: rajk#desktop
- Path: /~/
- Endpoint: aps#clutch
- Path: /~/
- Dialog Box: Activate Endpoint: aps#clutch
- MyProxy Server: clutch.aps.anl.gov:51000
- Fields: *Username, *Passphrase, Server DN, Credential Lifetime (hours)
- Buttons: Authenticate, Cancel

Bottom Screenshot: Transfer Files Page

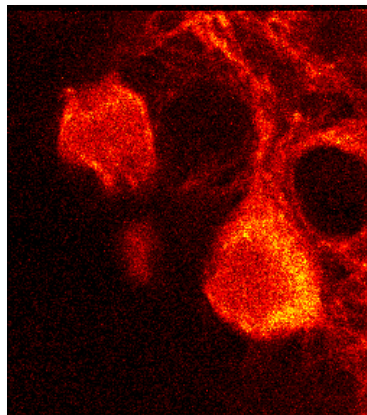
- Endpoint: rajk#desktop
- Path: /~/
- Endpoint: aps#clutch
- Path: /data/tomo2/tomo/
- File List:

Item	Type
1a	Folder
529	Folder
ACM	Folder
ACM_Senior_Member	Folder
ALCF_OTP_Myproxy	Folder
BIRN	Folder
Bill_globus_files	Folder
Bill_tutorials	Folder
Biography_short	Folder
CCGrid11_Tutorial	Folder
CDIGS	Folder
CEDPS	Folder
CTS2010	Folder
CV	Folder
C_Programs	Folder
Charts	Folder
Coupons	Folder
CreditReport	Folder
DOE-ASCR-BES-Workshop-2011	Folder
DOE_Terabits_Workshop_Feb2011	Folder
2010_03	Folder
2010_04	Folder
2010_08	Folder
2010_09	Folder
2010_10	Folder
2010_12	Folder
2011_02	Folder
2011_04	Folder
2011_06	Folder
2011_10	Folder
2011_11	Folder
2011_12	Folder
2012_02	Folder
2012_03	Folder
Frazier_Sam26_raw	Folder
beamline_settings	Folder
fzk	Folder
springer	Folder
test	Folder
Florian_2010_04_10a.txt	1MB

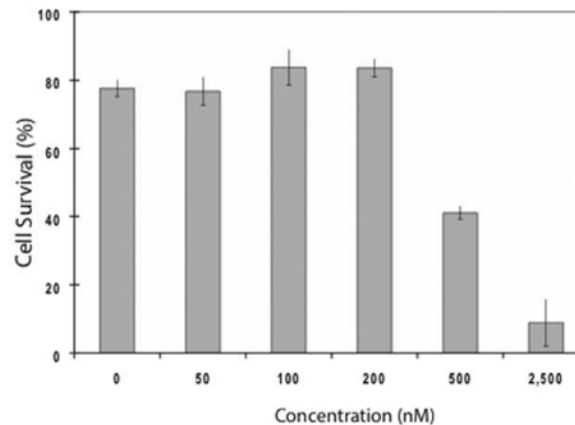
User / Scientist Perspective

Publication

- Multiple figures
- Different types of data



Laboratory Microscope Data



Synchrotron Derived Data



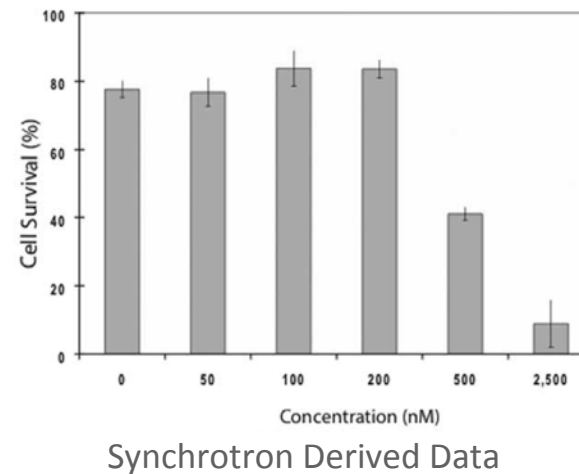
User / Scientist Perspective



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



Normalize Intensity
Cell Finding Algorithm
Data Fusion



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



Globus Online Catalog

The screenshot displays the Globus Online interface. At the top, the header includes the Globus logo and the text "globus online". Navigation links for "Manage Data", "Groups", "Support", and "faisal" are present. Below the header, there are links for "manage datasets", "start transfer", "view transfer activity", "manage endpoints", and "dashboard".

The main content area shows a search bar with a "Search" button and a "Create Dataset" button. A dropdown menu is set to "AES" with a "Create Catalog" link below it. A "Filter by Annotation" option is also visible.

The primary focus is on a dataset entry for the file "control_cds_ps75_180C_Fq2_003_0001-20481.hdf", dated 2013-07-30. The entry includes an owner "faisal" and a "label:" field. A modal window titled "Edit Tags | Add Tags" is open, showing a list of metadata fields with their current values and edit icons:

Field Name	Value
specscan_data_number	180
normalization_method	TRANSMITTED
specfile	simon20120627
output_file_local	/data/exchange/file1
output_data	/exchange1
kinetics	DISABLED
flatfield_enabled	DISABLED
file_mode	MULTI
blemish_enabled	ENABLED
compression	ENABLED

Below the modal, another dataset entry is partially visible: "R056_G14_p01_30C_PICCD_Sq1_0001-0505.hdf", dated 2013-07-30, owned by "faisal".

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



Globus Online Catalog

The screenshot displays the Globus Online Catalog interface. At the top, the header includes the Globus logo and the text 'globus online', along with navigation links for 'Manage Data', 'Groups', 'Support', and 'faisal'. Below the header, there are links for 'manage datasets', 'start transfer', 'view transfer activity', 'manage endpoints', and 'dashboard'. On the left side, there is a 'Catalog' dropdown menu set to 'AES' and a 'Filter by Annotation' section. The main content area shows a search bar and a 'Create Dataset' button. A dataset entry is visible with the filename 'control_cds_ps75_180C_Fq2_003_0001-20481.hdf', dated '2013-07-30', and owned by 'faisal'. An 'Edit Tags' modal is open over this dataset, showing a list of tags with their values and edit icons:

Tag Name	Value
specscan_data_number	180
normalization_method	TRANSMITTED
specfile	simon20120627
output_file_local	/data/exchange/file1
output_data	exchange1
kinetics	DISABLED
flatfield_enabled	DISABLED
file_mode	MULTI
blemish_enabled	ENABLED
compression	ENABLED

Below the modal, another dataset entry is partially visible: 'R056_G14_p01_30C_PICCD_Sq1_0001-0505.hdf', dated '2013-07-30', and owned by 'faisal'.

- 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!