PyLight, A Common Architecture for Composable Workflows to Facilitate Data Acquisition and Analysis at NSLS-II

Dantong Yu

Yong Chu, Wah-Keat Lee, Wei Xu, Li Li, Xi Deng
Hanfei Yan, Xiaojing Huang, Karen Yu-Chen Chen
Cheng Chang, Yulin Huang (students)

Outlines

- Science Needs
- Workflow Requirements
- Our Approach
- Scientific Tasks
- Pipeline status and progress
- Workflow framework
- Web interfaces implemented and integrated into the framework
- Demo

Objectives

- Build an integrated workflow system to serve more than one beamline: expand the available workflow prototype to six early beamlines and all beamlines at NSLS-II eventually
- Provide a suite of modules with high performance computing support, machine learning and pattern recognition, and visualization

What Beamlines need?

- Improved Data (Image) Analysis algorithm (tomography)
- Preliminary data processing/analysis of the incoming data
- Real time or near real time visualization of data for QA:
  - Raw data:
    - 2D area detectors, Spectrum from multi-element SDD
  - Constructed data:
    - Integrated intensity over regions of interest, phase gradients per pixel, etc
  - Time resolved or spatial data in 3D
- Complex multi-model Data Analysis/Comparison/Correlation of the incoming data with the other data
  - XRF image, SEM image, 3D data, Photograph, Light microscopy image,
  - Correlation, algebraic operations on images, projection, integration
- In-Situ Analysis must guide experiment decisions and help determine the goodness of data

Additional Requirements

- A system for tracking of data provenance and version as they are analyzed/created by both users and system pipelines.
- A system for tracking the paths that users take through the system (e.g., the series of data and parametric inputs, commands, outputs and manual notation a user makes, in temporal order, as they use the system – Experimental Notebook);
- The ability of beamline users/developers to easily add modules that works on known data types;
- The ability to access serious computational power and storage resources

What can workflow provide?

- Provide a framework to integrate the building blocks of data processing and analysis
- Create a generic web-based graphical user interface for scientists to manage and control data collection integration and subsequent analysis
- Support experiment (measurement) pipeline creation, execution, sharing, reuse and provenance
- Integrate highly scalable parallelism supports (Multi-core, GPU clusters, cloud computing, and cluster computing supports) for data transformation and analysis

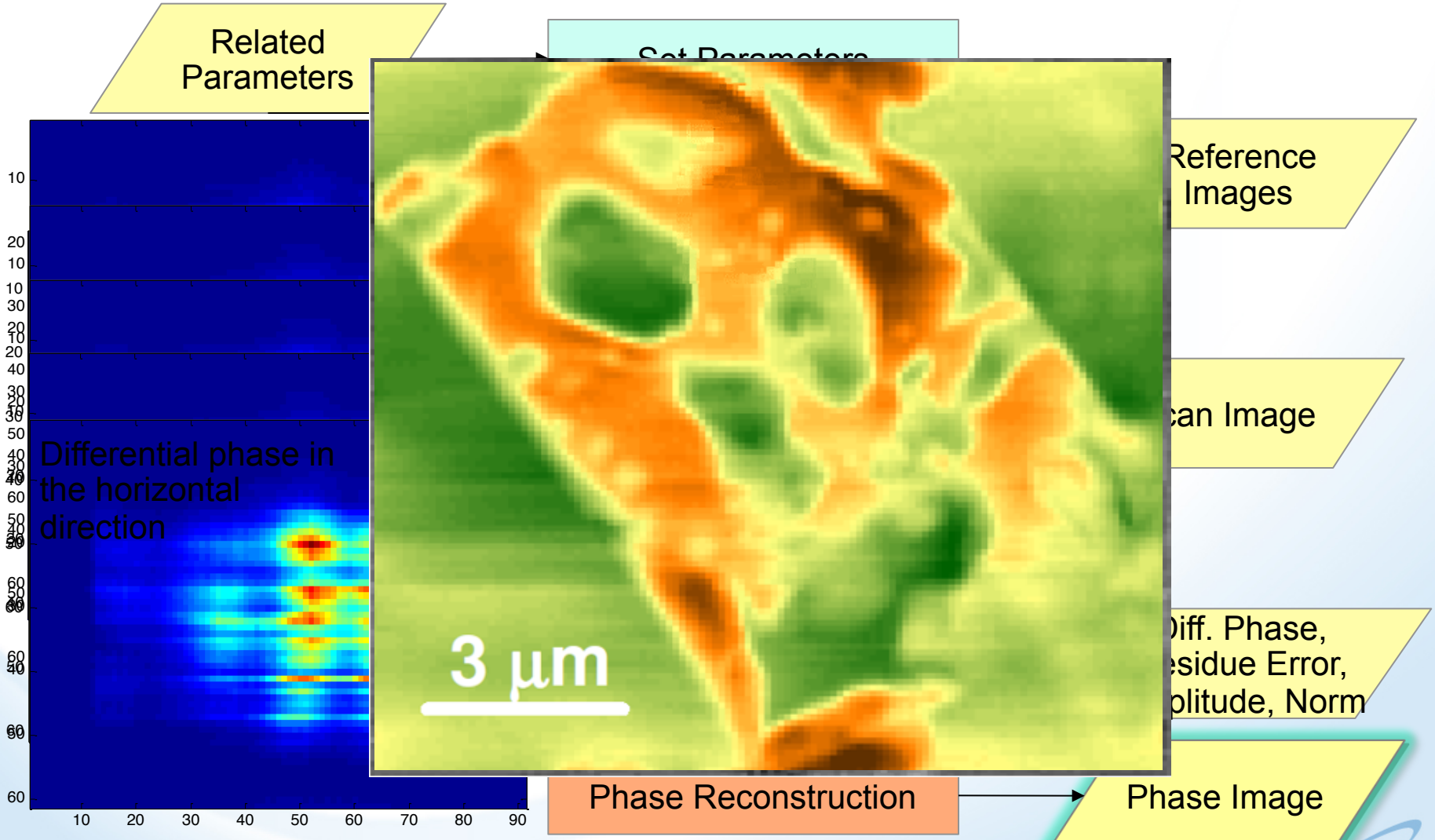
Approach

- Individual Process Pipelines ➔ General Framework/Workflow ➔ Common/Sharable/HPC based Module/tool library ➔ Common API for tool extension
- Start with beamline specific requirements and manually curated processing pipeline to integrate existing in-house analysis tools
    - DPC
    - MAPS (from APS)
    - Reciprocal space coordinate transformation
    - FXI tomography
- Build data analysis pipeline framework around the analysis tools
- Common Module Libraries
    - CDI/Ptychography (FFTW package)
    - Tomography
    - Visualization
    - etc.

Individual Scientific Tasks in progress

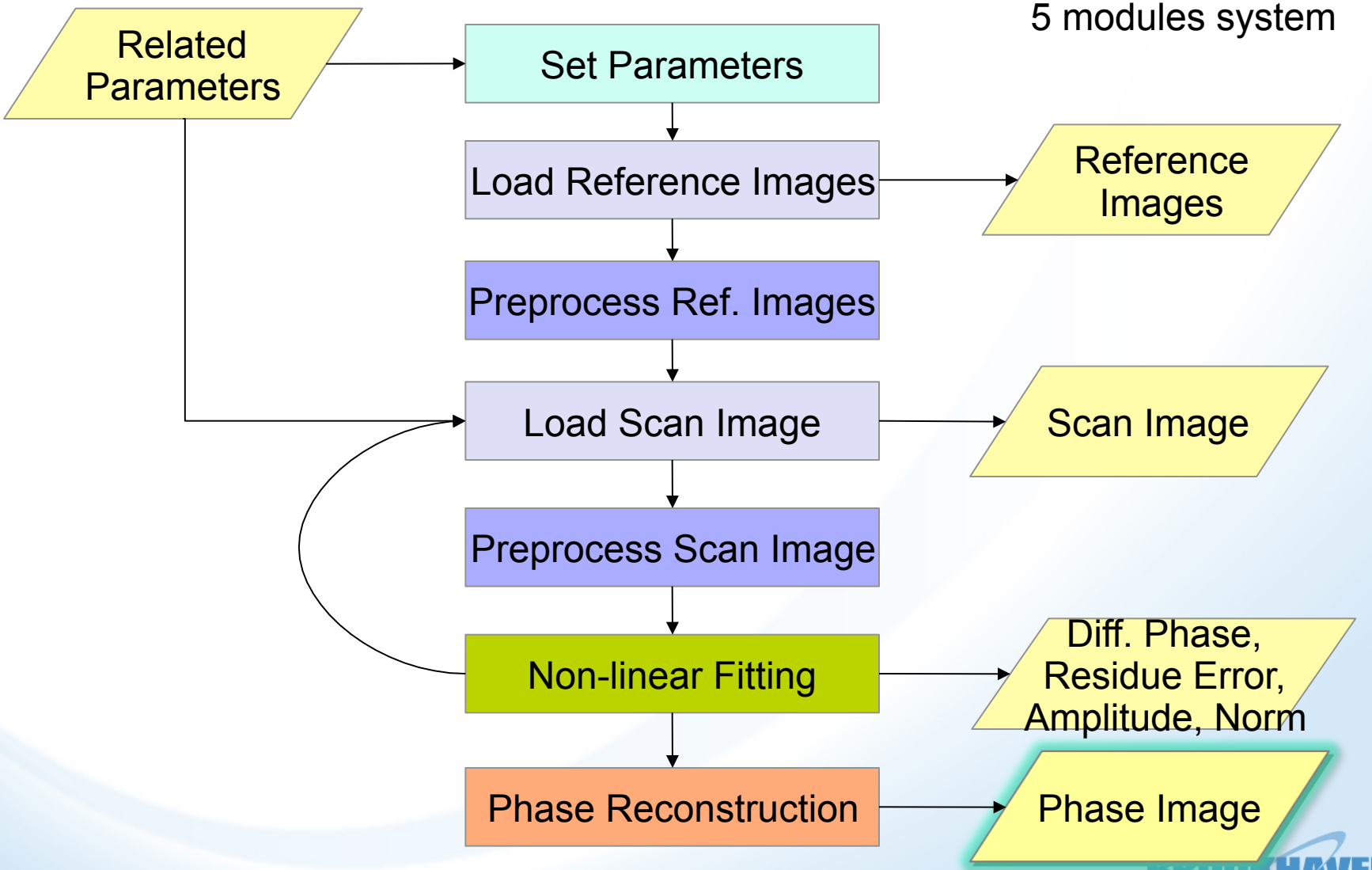
- Differential Phase Contrast Imaging (complete):
  - Straightforward & in-house expertise
- Tomography and 3D rendering (near completion) (with FXI W.K. Lee's team)
  - Benchmark using exiting data set
  - Common HDF data file structure as APS and ALS
- 2D, 3D Spectroscopy Image
- 2D image registration and workflow for TXM
  - Collaborate with W.K. Lee's team
- 2D and 3D XRF imaging
  - Receive MAPS written in python from the APS
  - Incorporate into workflow framework
  - Benchmark at APS 2-ID-E
- 2D and 3D Ptychography pipelines
  - Collaborate with the APS, ALS, and SLS
  - Benchmark at APS 34-ID-C and/or SLS c-SAX

Workflow


Related Parameters
Set Parameters
Reference Images
can Image
Differential phase in the horizontal direction
3 µm
Diff. Phase, esidue Error, plitude, Norm
Phase Reconstruction
Phase Image

Workflow

5 modules system

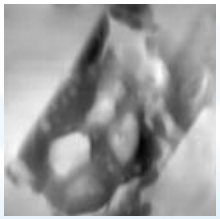


Differential Phase Contrast Image (DPC)

- DPC implemented with various programming languages on a single computer, submitted to workshop

	Test 1	Test 2	Test 3	Test 4	Average
Python	3449.769	3777.162	3684.838	3704.820	3654.147
Matlab	577.764	555.970	570.677	566.742	567.788
C++ ASA047	53.320	51.096	52.334	50.900	51.913
C++ GSL	44.550	44.080	43.063	42.952	43.661

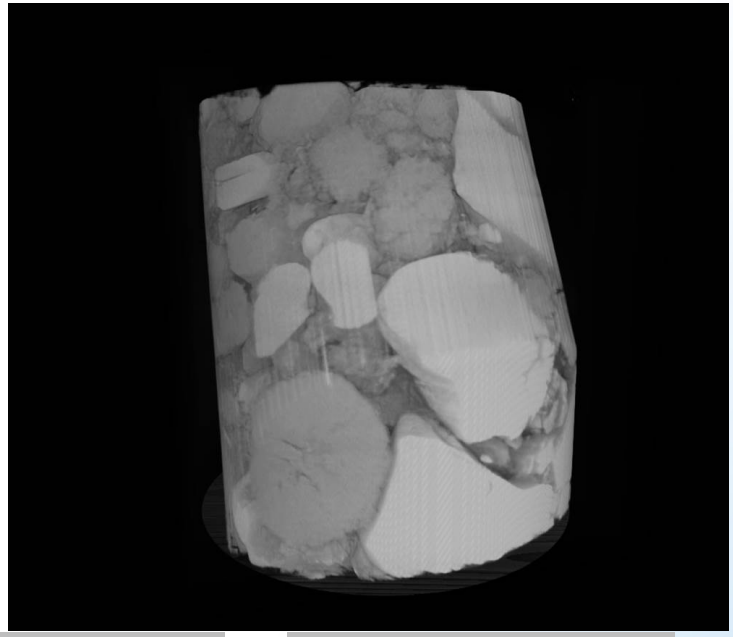
- Multi-core cluster computing
  - Partition the fitting for all pixels to a number of groups
  - One group of pixels are computed on one core independently in parallel with other groups
  - (MPI) Message passing for task coordination

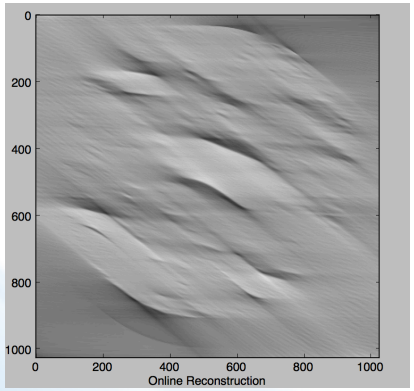


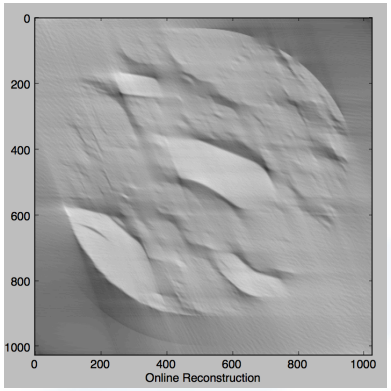
Multi-core (Python): 37s
~100 times faster

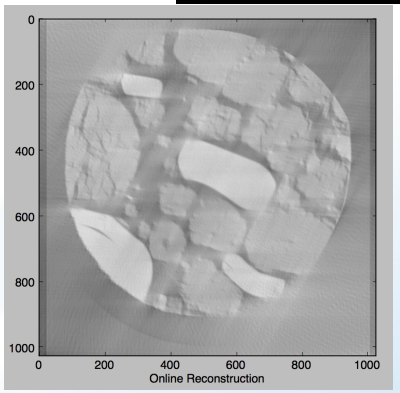
Example: Parallel computing for Tomography (Approach 1)

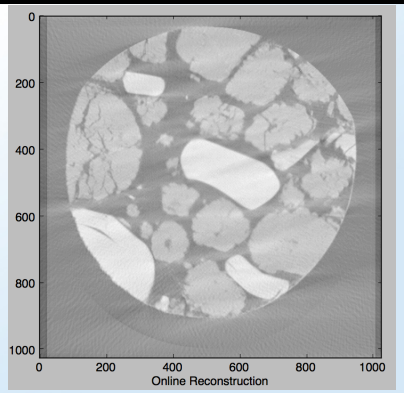
- Filtered backprojection (FBP):
  - Compute filtered projection and “smear” the projections back to the object space independently in parallel
  - One more step to accumulate all the back projected images
  - Based on MPI (OpenMPI)











With more projections back-projected

Speed Up Factor

Initialization overhead (HDF5 reading from disks) is significant

- Projection size: 1026*200
- Projection number: 360
- 3D data size: 1026*1026*200

- Program input and output as hdf5 files

- Time performance:
- Single core: 4556s
- Multi core (9*16=144): 198s
- Speedup: ~ 23 times

- Projection size: 1026*20
- Projection number: 360
- 3D data size: 1026*1026*20

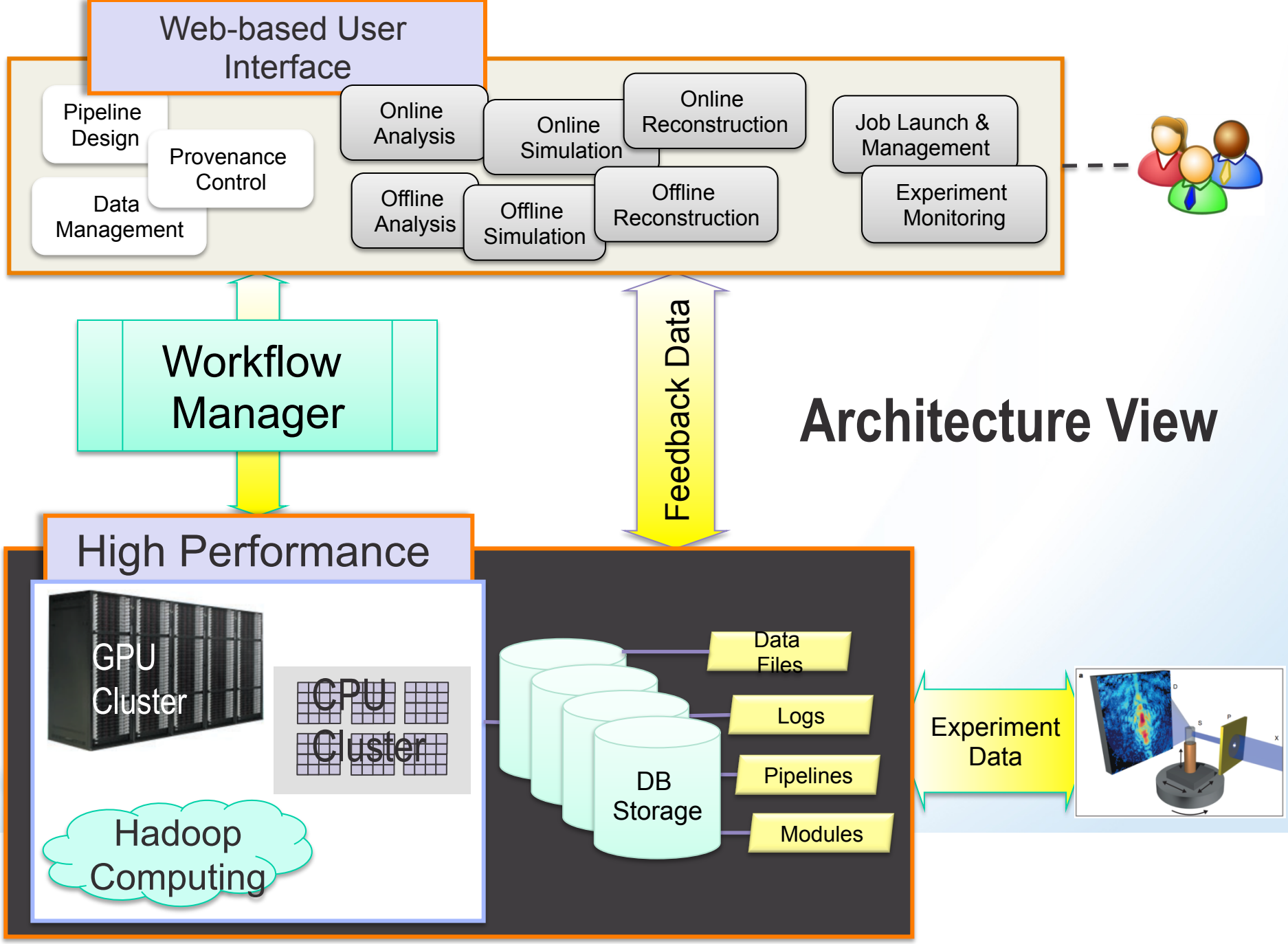
- Program input and output as hdf5 files

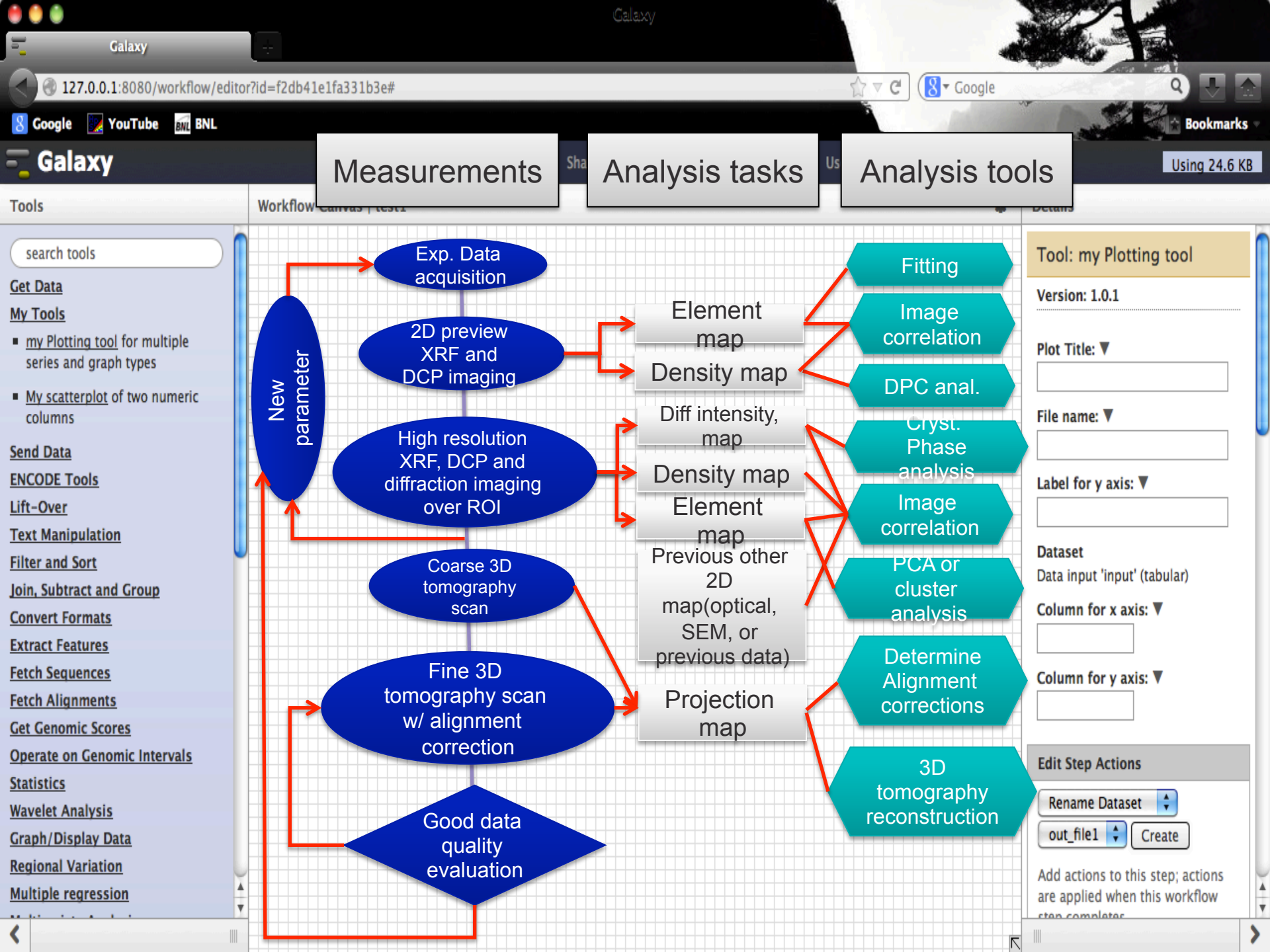
- Time performance:
- Single core: 214s
- Multi core (9*16=144): 45s
- Speedup: ~ 5 times

Parallel computing for Tomography (Approach 2): In progress

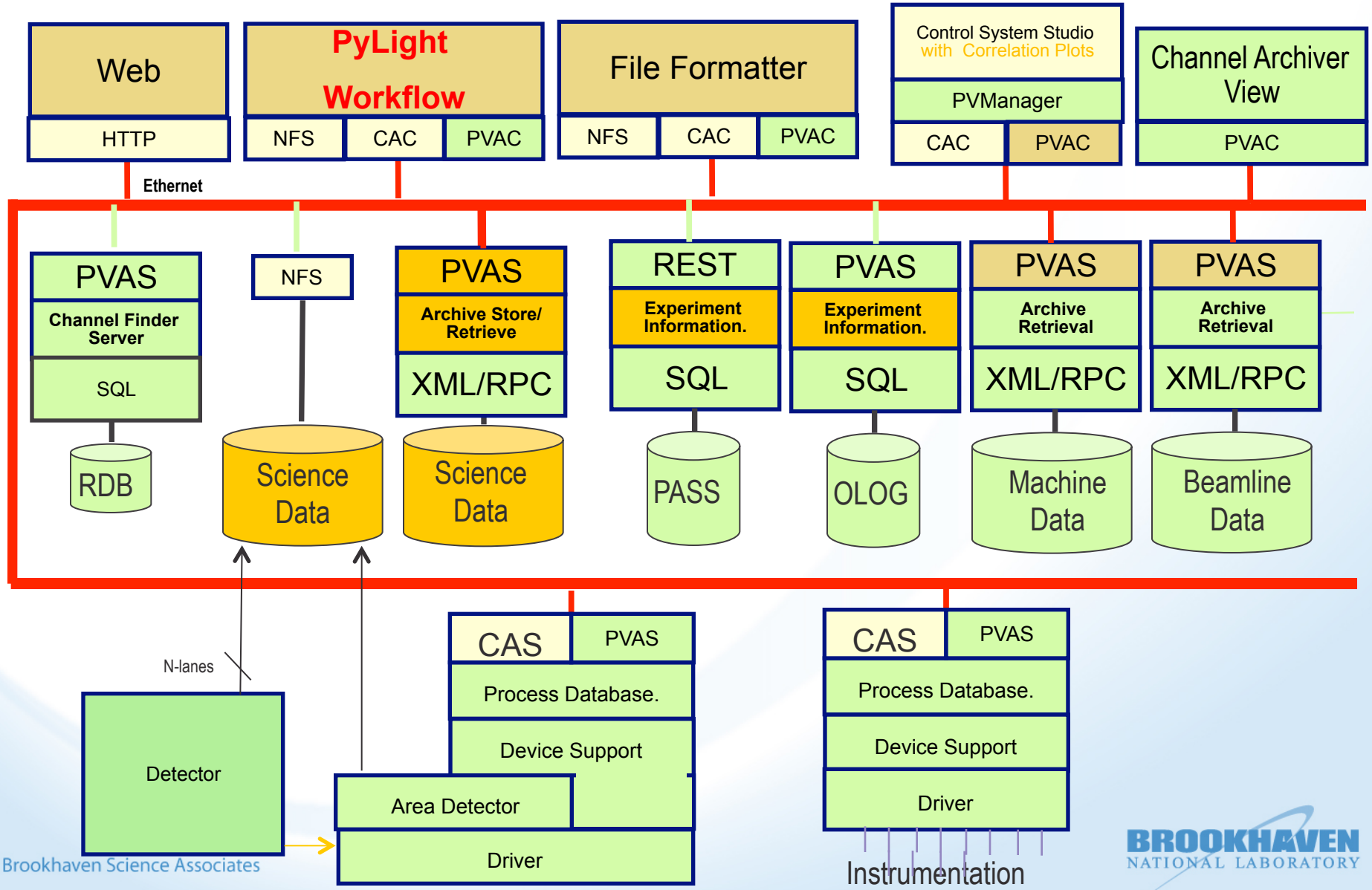
- SIRT, ART, SART, for example: Mark Rivers's tomoRecon: http://cars9.uchicago.edu/software/epics/tomoRecon.html
- Computation intensive
- Iterative reconstruction:
  - The parallelism is within each iteration
  - Run the following in each core in parallel:
    - Forward projection from a specific angle
    - Correction
    - Backward projection from the same angle
  - Accumulate and update pixel before next iteration
- Performs better when with limited angle or missing angle, or noisy projections
- Need to re-implemented with MPI (OpenMPI)
- Or GPGPU

Architecture View

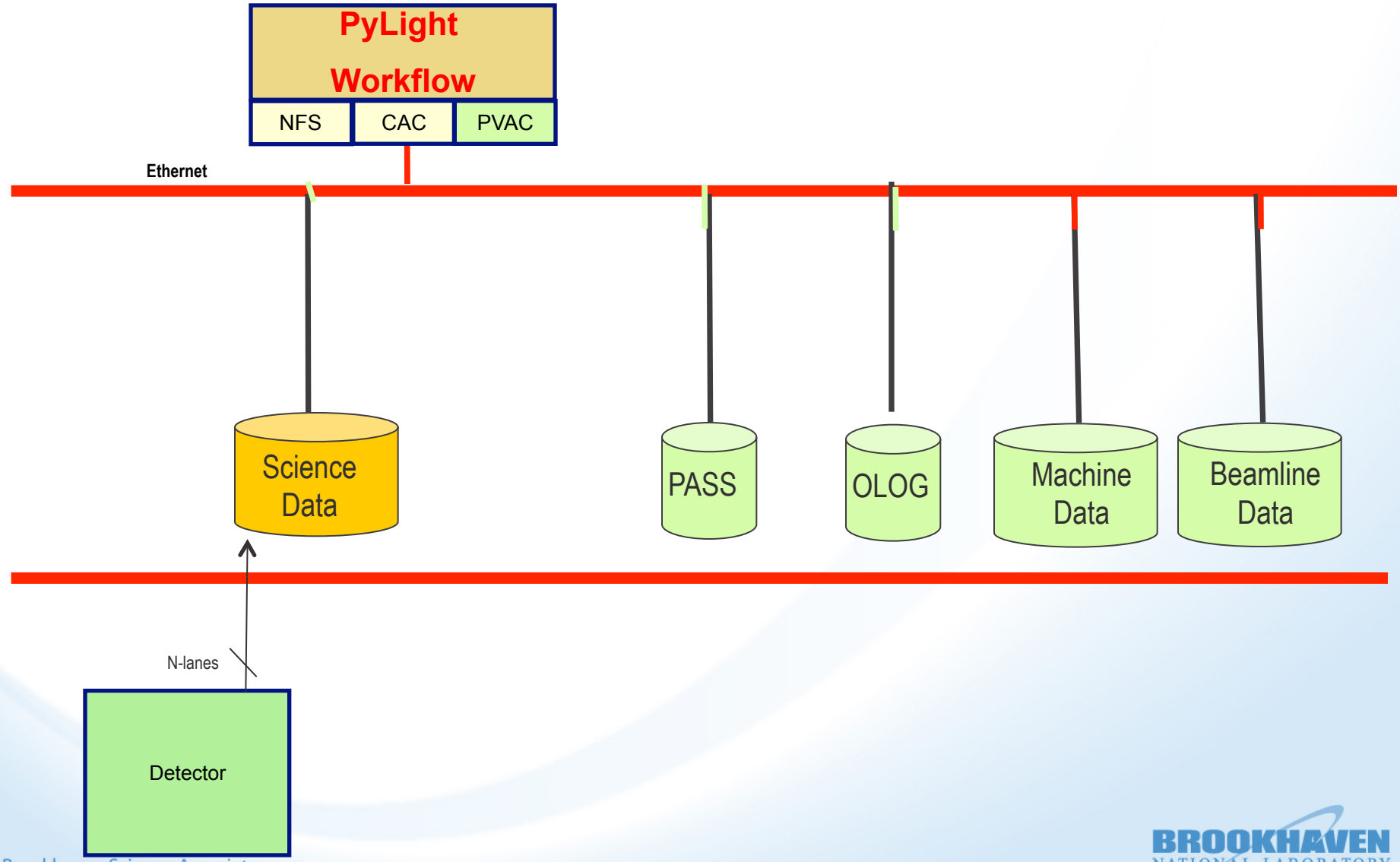



Galaxy
127.0.0.1:8080/workflow/editor?id=f2db41e1fa331b3e#
Google
YouTube
BNL
Bookmarks
Galaxy
Using 24.6 KB
Measurements
Analysis tasks
Analysis tools
Tools
Workflow
search tools
Get Data
My Tools
my Plotting tool for multiple series and graph types
My scatterplot of two numeric columns
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
New parameter
Exp. Data acquisition
2D preview XRF and DCP imaging
High resolution XRF, DCP and diffraction imaging over ROI
Coarse 3D tomography scan
Fine 3D tomography scan w/ alignment correction
Good data quality evaluation
Element map
Density map
Diff intensity, map
Density map
Element map
Previous other 2D map(optical, SEM, or previous data)
Projection map
Fitting
Image correlation
DPC anal.
Cryst. Phase analysis
Image correlation
PCA or cluster analysis
Determine Alignment corrections
3D tomography reconstruction
Tool: my Plotting tool
Version: 1.0.1
Plot Title:
File name:
Label for y axis:
Dataset
Data input 'input' (tabular)
Column for x axis:
Column for y axis:
Edit Step Actions
Rename Dataset
out_file1
Create
Add actions to this step; actions are applied when this workflow

Integration with Experimental Data Architecture



Integration with Experimental Data Architecture



PyLight: Workflow System for NSLSII

PYLIGHT | ANALYZE DATA | WORKFLOW | VISUALIZATION | ACCOUNT | HELP | keithdeng@gmail.com | Using 35.7 M

Control panel

User information

Tools Panel

Tools

Search tools

Search

Early beamlines

HXN

SRX

CSX-1

CSX-2

CHX

XPD

IXS

Additional Beamlines

FXI

History

Operation history Panel

About the National Synchrotron Light Source II

Though the current NSLS has been continually updated since its commissioning in 1982, today the practical limits of machine performance have been reached. Meeting the critical scientific challenges of our energy future will require advanced new capabilities that NSLS-II will uniquely provide.

NSLS-II will be a new state-of-the-art, medium-energy electron storage ring (3 billion electron-volts) designed to deliver world-leading intensity and brightness, and will produce x-rays more than 10,000 times brighter than the current NSLS. The superlative character and combination of capabilities will have broad impact on a wide range of disciplines and scientific initiatives, including the National Institutes of Health, structural genomics initiative, DOE, Genomics GTL initiative, and the federal nanoscience initiative.

The facility will be a key resource for researchers at Brookhaven, Center for Functional Nanomaterials, allowing for analysis of new materials that are expected to transform the nation, energy future. Construction of the NSLS-II ring building began in March 2009. If plans are carried through as proposed, the new facility will begin operating in 2015.

Main Panel

Brookhaven National Lab Copyright © 2013 Tweet

Tools

Search tools

Early Beamlines

HXN

Upload data

DPC

Tomography

SRX

CSX-1

CSX-2

CHX

XPD

IXS

Additional Beamlines

FXI

Tomography

Input file:

Dataset missing? See TIP below.

Start angle:

Dataset missing? See TIP below.

End angle:

Dataset missing? See TIP below.

Execute

History

Unnamed History

2. DPC.tiff

1. tomography.input

Tools

Search tools

Early Beamlines

HXN

Upload data

DPC

Tomography

SRX

CSX-1

CSX-2

CHX

XPD

IXS

Additional Beamlines

FXI

Tool: Tomography

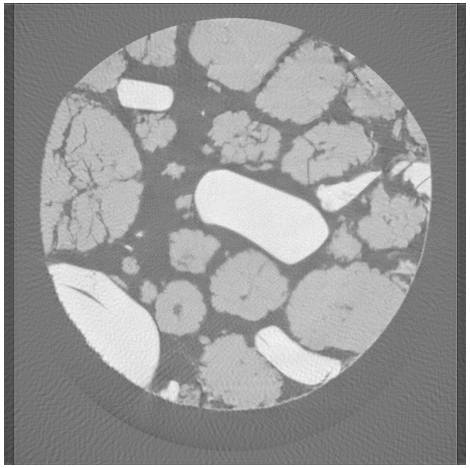
Rerun

Name:	tomography 1
Created:	Jul 23, 2013
File size:	0 bytes
Format:	txt

Input Parameter	Value
Start angle	0
End angle	180

Output

Image 1:



Notes:

This is a note.

History

Unnamed History

2. DPC.tiff

1. tomography.input

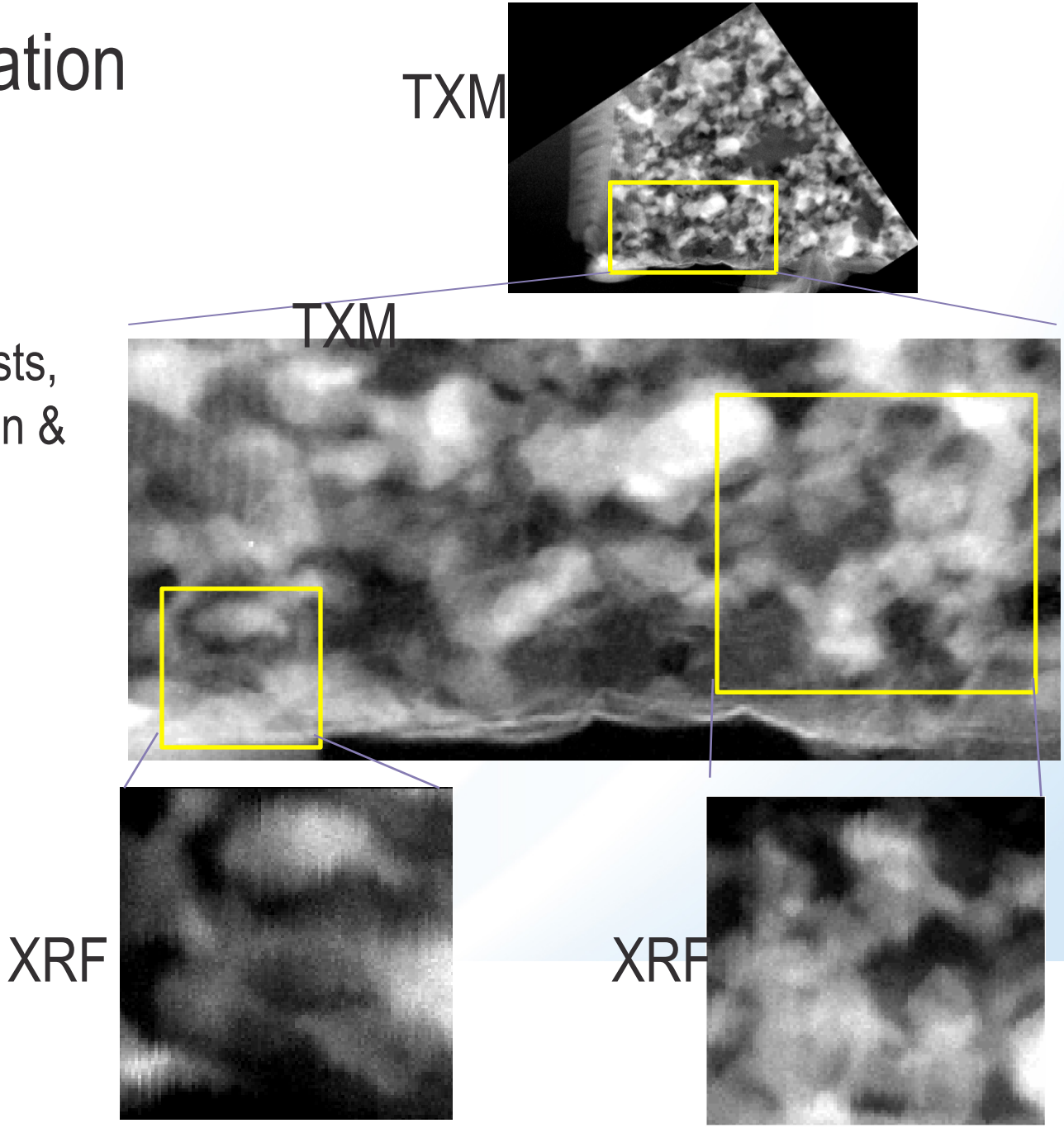
Summary

- Prototype Framework is deployed in a computing cluster
- Web service based architectures to allow users to enter parameters
- HPC must be incorporated into the workflow design to handle real time requirements
  - Learning curve (MPI, GPU, co processor) in the science domain and parallel programming
- Need to co-design with data management component
- Start work with six early beamlines at NSLS-II to collect user requirements

Backup Slides

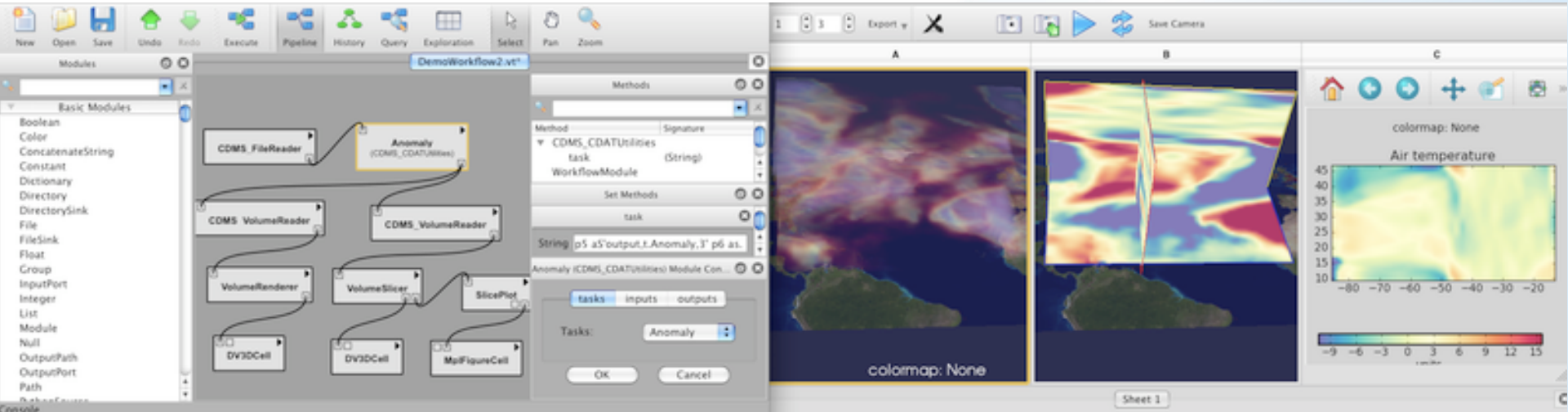
Image Registration

- Must account for difference in contrasts, in addition to rotation & translation



VisTrails

- Strength is to manage and maintain provenance for reproduce, validate and share scientific results and provide strong graphics capabilities.
- Support exploratory tasks for visualization and data mining.
- Allow user define data format, data input, and output.
- Tools can be easily integrated into the repository (tool shed).
- Workflow Engine:
  - Support local host and batch queue execution
  - Parallel and remote execution is in development

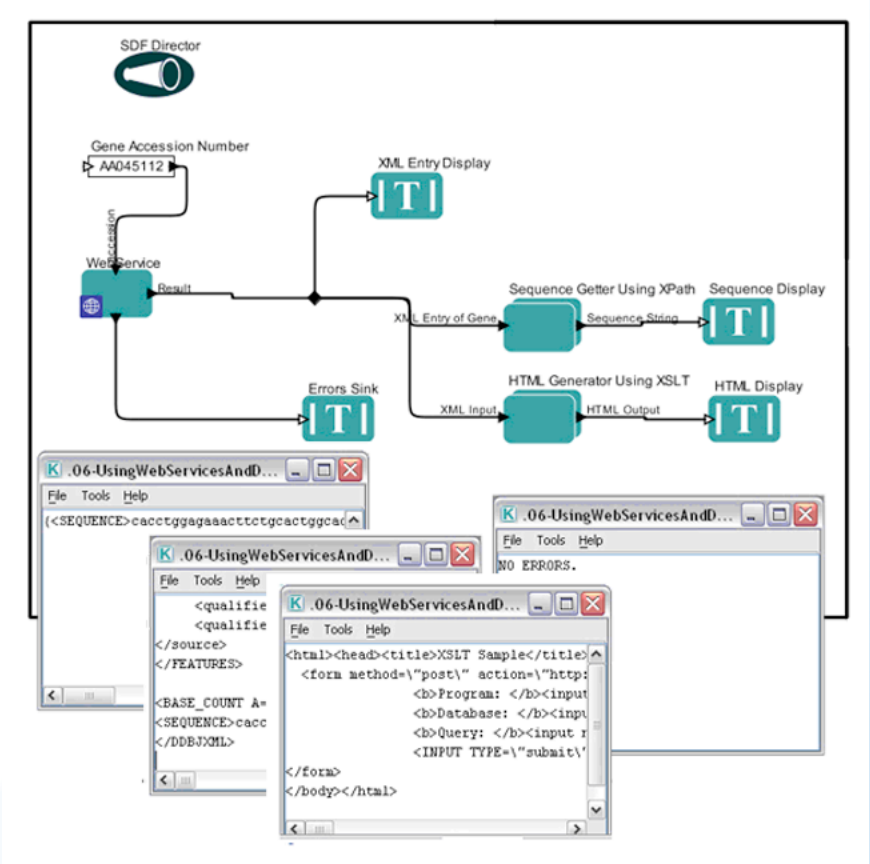


Kepler

Better support in large scale storage

- Support semantics, metadata based data accessing
- Support Meta data and support
- Support Grid Actions (Globus Job and GridFTP)
- Support SRB (Storage Resource Broker)

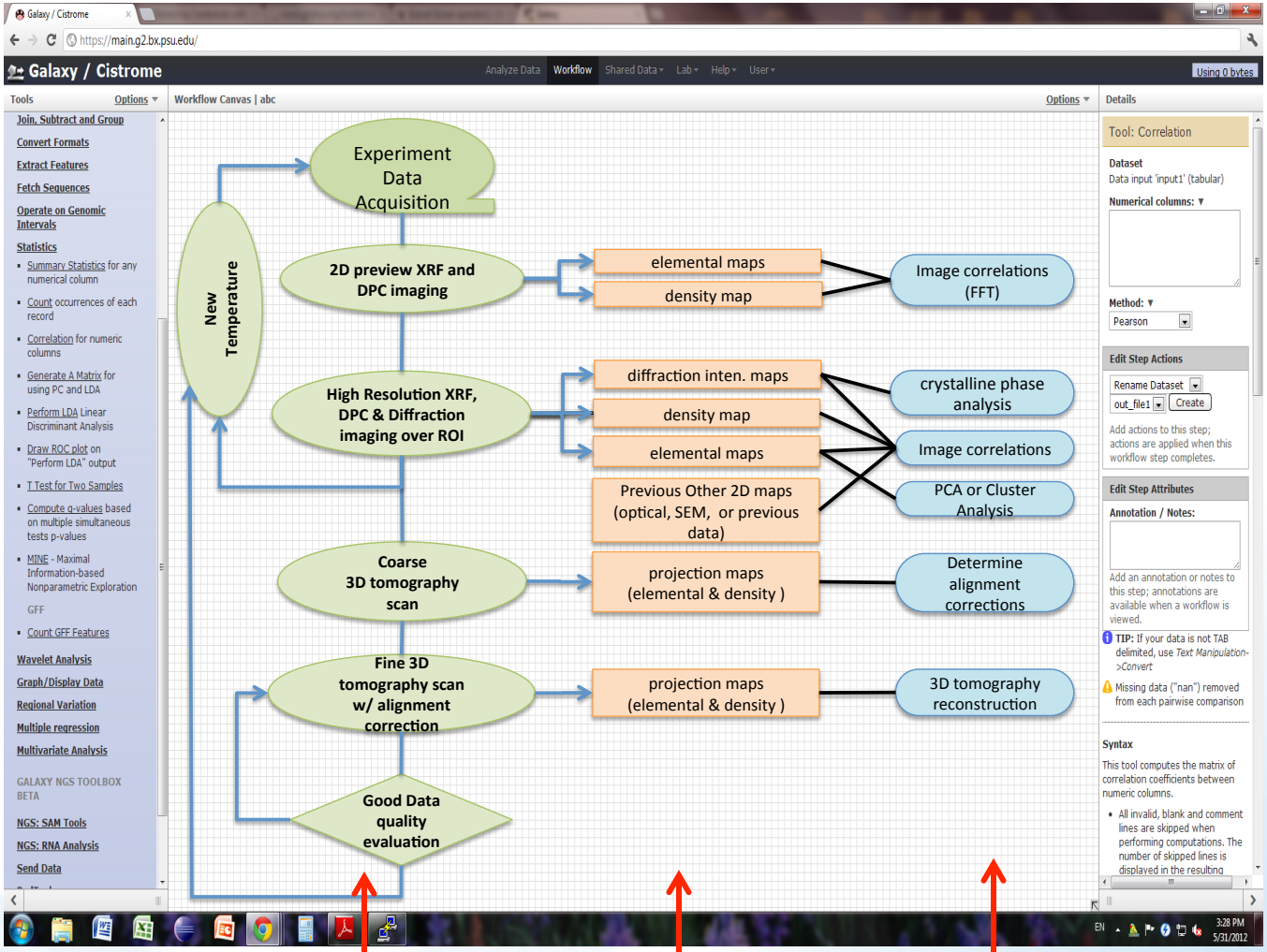
Can be extended to support real-time data streaming and on-the-fly analysis of image stream



Other Workflow System

- Pegasus, etc…
- Are these tools capable handing the data challenge of NSLS-II? Determined by
  - Data Volumes and Streaming Throughput
  - Number of tasks and complexity of workflows
  - Capacity and Capability of Analysis tools within the workflow
  - Available Computing Facilities

Concept for Computational Workflow for HXN at NSLS-II



measurements

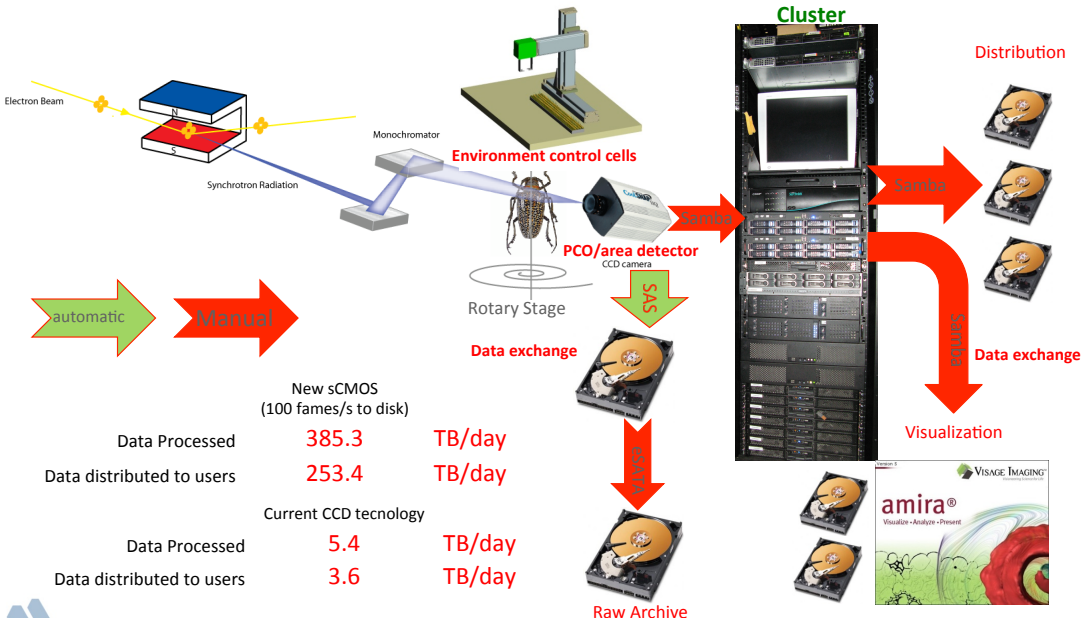
Type of analysis tasks

Type of tools for the analysis

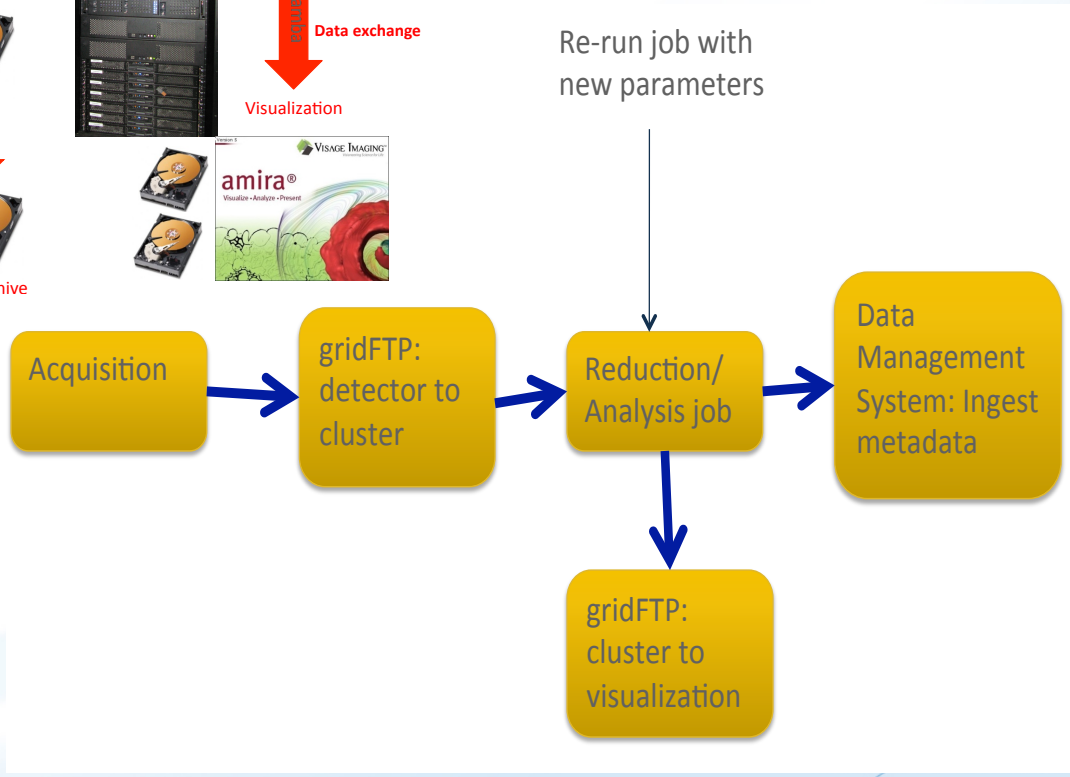
Design Considerations

- Workflow software design, data analysis tools, data management, computing and storage facility planning need to be coordinated together.
- Interface design between them are planned in advance.
- Workflow system will be the portal for users to manage and interact various Beamline experiments, on-line and off-line resources.
  - Consistent for on-the-fly data processing at beamline and post-experiment analysis at their home institutions
- Beamline scientists will be the central stakeholder.
  - Yong
  - Wah-Keat

Start with simple prototype and build on top of success



New sCMOS (100 fames/s to disk)		
Data Processed	385.3	TB/day
Data distributed to users	253.4	TB/day
Current CCD tecnology		
Data Processed	5.4	TB/day
Data distributed to users	3.6	TB/day

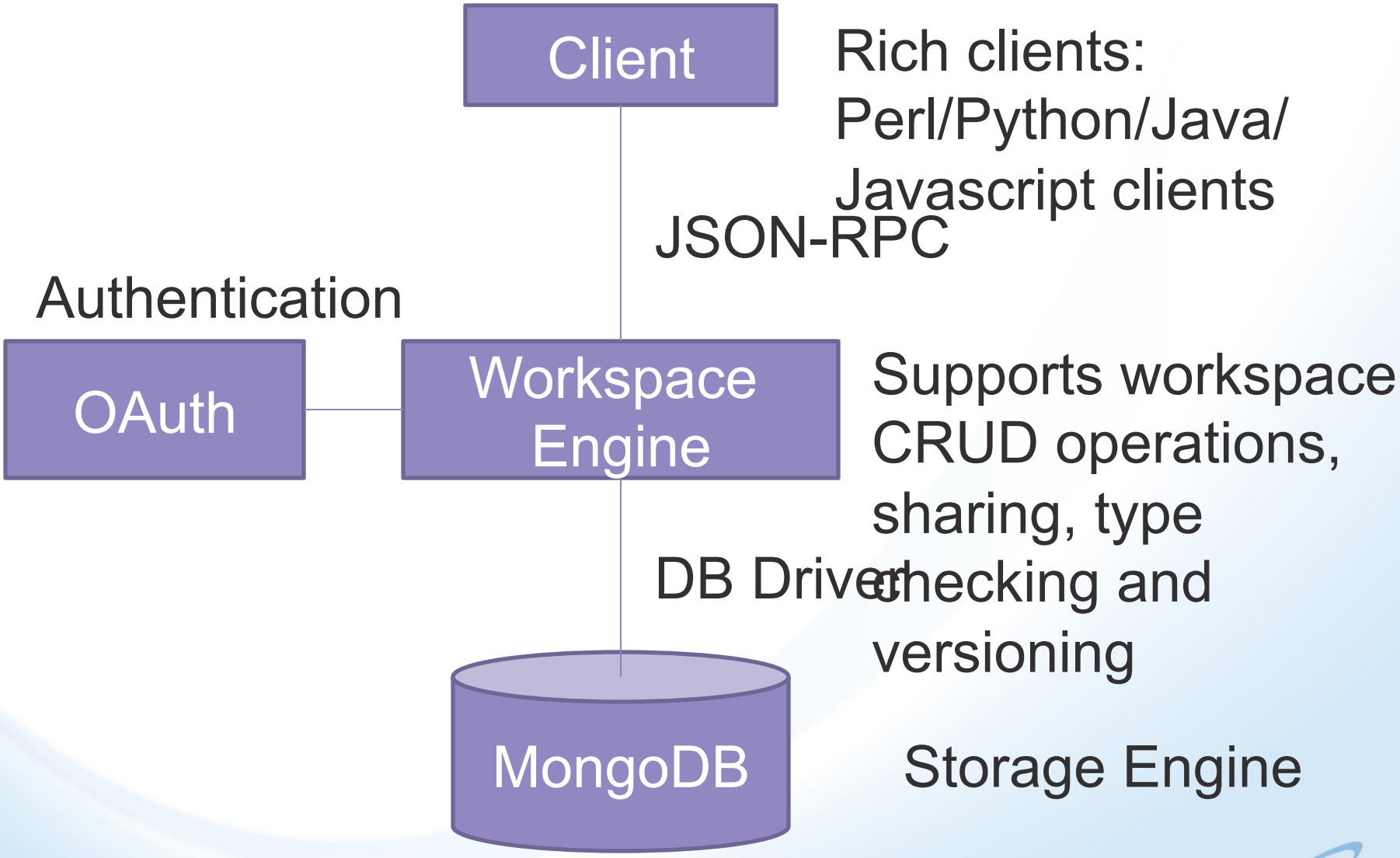


Four Joint Openings between Photon Science and Computational Science

- Postdoctoral Research Associate (RA-1) – Data Analysis Tool Development – LDRD Job ID: 16270 (joint)-Under Yong
- Research Associate - Multidimensional x-ray image analysis (2 positions)-Under Wah-keat
- Advanced Applications Engineer / Software Engineer (I7) (RM6768) (joint) Under Dantong and Yong
- Promising Candidates were being interviewed and we made offer to one opening already.
- We plan to hold a kick-off meeting in later January/ 2013 once the job candidates on-board

Beamlines that expressed interests to our project

- Hard X-RAY NANO-PROBE
- Full-Field X-ray Imaging Beamline
- Protein Crystallography
- SRX: Submicron Resolution X-ray Spectroscopy
- What are needed to ensure the success:
  - Coordinator/Project Manager
  - Science User Committee (Science leads)
  - Scientific Data Analysis Tool Developer and Integrator (Post-docs)
  - Workflow Architect
  - Software Engineer
  - System Administrator/Architect (Data Storage/ Management/Transfer/System virtualization)


Client
Rich clients: Perl/Python/Java/ Javascript clients
JSON-RPC
Authentication
OAuth
Workspace Engine
Supports workspace CRUD operations, sharing, type checking and versioning
DB Driver
MongoDB
Storage Engine