The Design and Development of the scientific data and software for High Energy Photon Source in China

Yu Hu, Hao Hu, Fazhi Qi (On behalf of HEPSCC)

Institute of High Energy Physics, CAS

Outline

1. HEPS Introduction
2. Demand and Challenges of scientific data and software system
3. The architecture and design of the framework
4. The status of the system
5. Summary

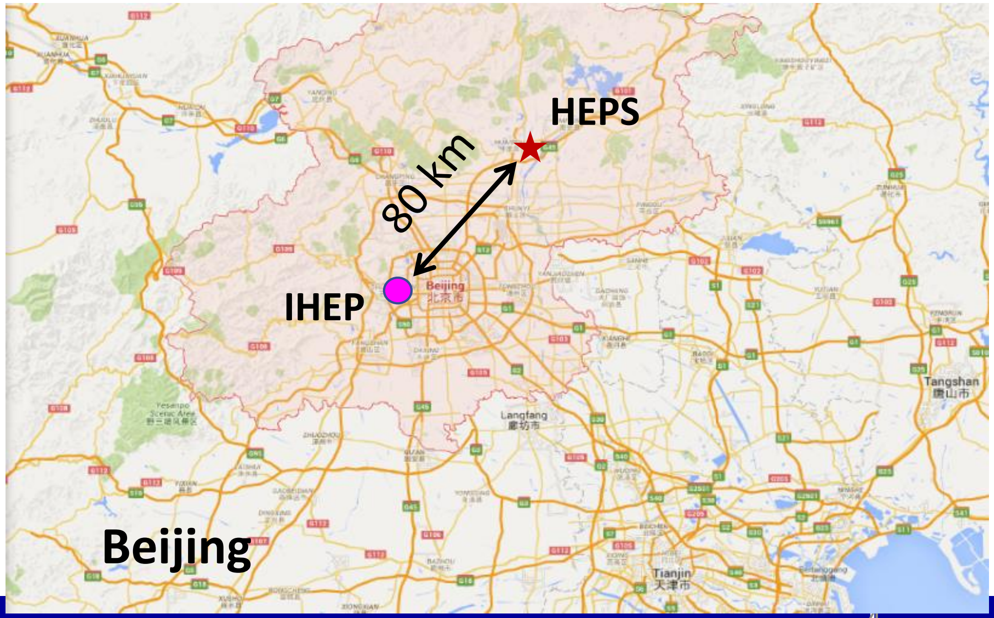
Outline

1. HEPS Introduction
2. Demand and Challenges of scientific data and software system
3. The architecture and design of the framework
4. The status of the system
5. Summary

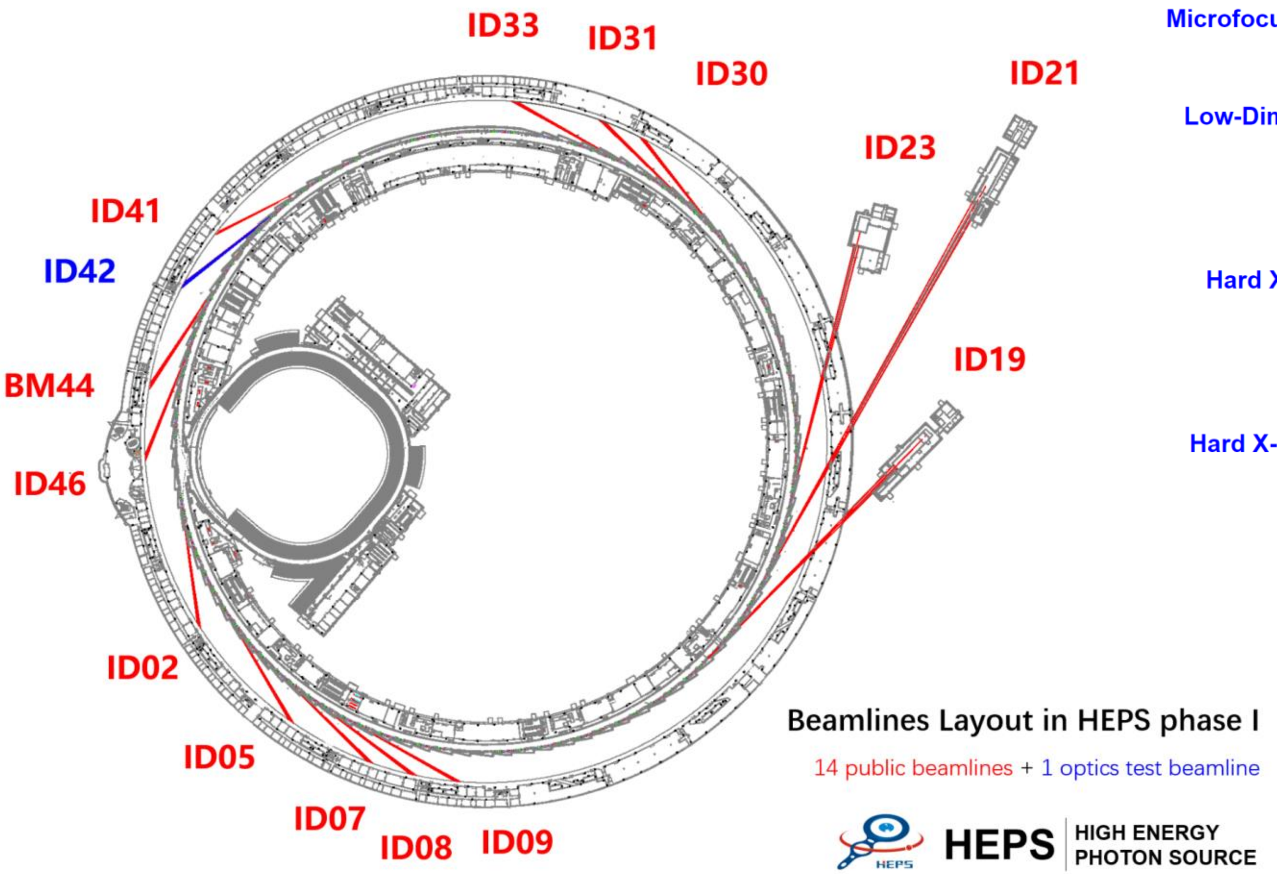
High Energy Photon Source (HEPS)

- New light source in China — High energy, high brightness
- Located in Beijing - about 80KM from IHEP
- Officially approved in Dec. 2017
- The construction was started at the end of 2018
- The whole project will be finished in mid-2025

Main parameters	Unit	Value
Beam energy	GeV	6
Circumference	m	1360.4
Emittance	pm·rad	< 60
Brightness	phs/s/mm²/mrad²/0.1%BW	>1x10²²
Beam current	mA	200
Injection		Top-up



Beamlines in HEPS phase I



Microfocusing X-Ray Protein Crystallography-ID02 Beamline

Low-Dimensional Structure Probe Beamline-ID05

Engineering Materials Beamline-ID07

Hard X-Ray Coherent Scattering Beamline-ID09

Pink Beam SAXS Beamline-ID08

Hard X-Ray Nanoprobe Multimodal Imaging-ID19 Beamline

Hard X-Ray Imaging Beamline-ID21

Structural Dynamics Beamline-ID23

ID30-Transmission X-Ray Microscopic Beamline

ID31-High Pressure Beamline

ID33-Hard X-Ray High Resolution Spectroscopy Beamline

BM44-Tender X-Ray Beamline

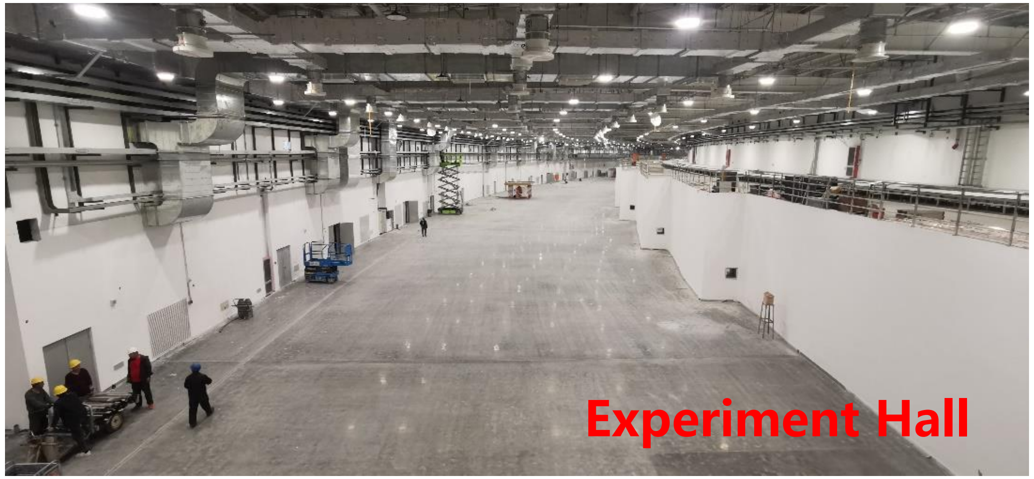
ID41-High Resolution Nanoscale Electronic Structure Spectroscopy Beamline

ID42-Optics Test Beamline

ID46-X-Ray Absorption Spectroscopy Beamline

14 public beamlines + 1 optics test beamline in Phase I

Can accommodate over 90 beamlines in total

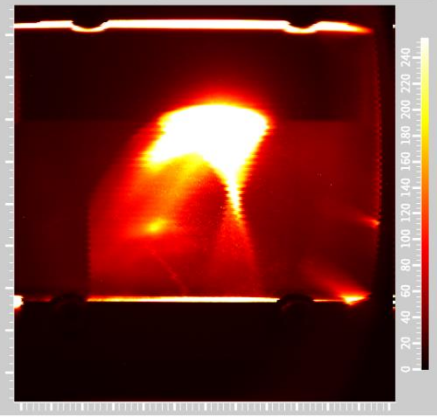


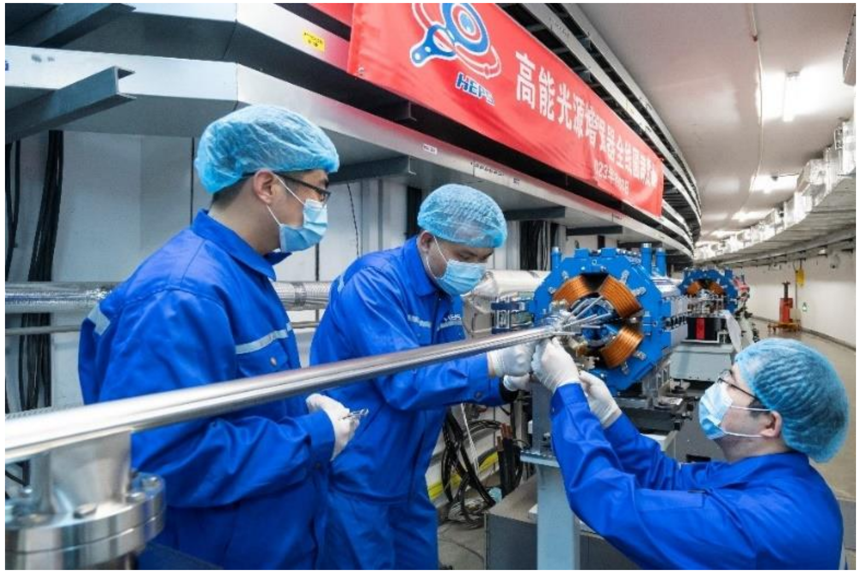
Progress of the HEPS project

- The construction of the civil structure completed. Now at the stage of equipment installation
- 2023.01, HEPS booster installation completed
- 2023.02, Start installation of storage ring
- 2023.03, HEPS achieved the first electron beam accelerated to 500 MeV.

HEPS LINAC

Beam Energy	500 MeV
Bunch Charge	2.61 nC
Trans. Efficiency	94 %





Outline

1. HEPS Introduction
2. **Demand and Challenges of scientific data and software system**
3. The architecture and design of the framework
4. The status of the system
5. Summary

Data Challenges @HEPS

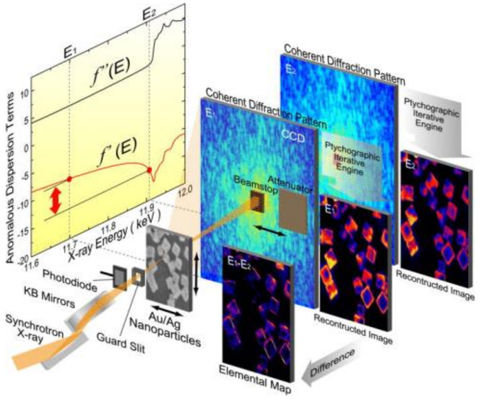
- Increased source brightness
  - More raw data in greater detail and less time
- X-ray detector capabilities constantly improving:
  - Increased dynamic range, faster readout rates, larger pixel arrays
  - Bigger frames, higher frame rates => more raw data
- >200PB raw data per year for Phase I (15 beamlines)
- More than 90 beamlines volume in total

Data volume of HEPS Beamlines:

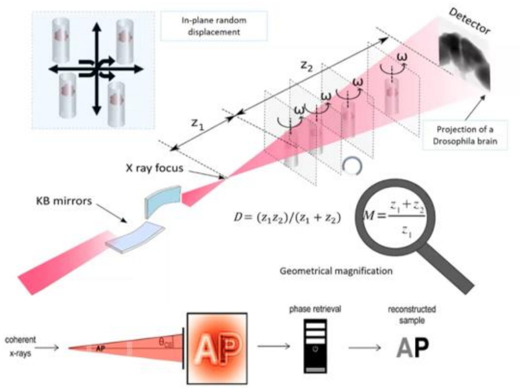
Beamlines	Burst output (TB/day)	Average output (TB/day)
B1 Engineering Materials Beamline	600.00	200.00
B2 Hard X-ray Multi-analytical Nanoprobe (HXMAN) Beamline	500.00	200.00
B3 Structural Dynamics Beamline	8.00	3.00
B4 Hard X-ray Coherent Scattering Beamline	10.00	3.00
B5 Hard X-ray High Energy Resolution Spectroscopy Beamline	10.00	1.00
B6 High Pressure Beamline	2.00	1.00
B7 Hard X-Ray Imaging Beamline	1000.00	250.00
B8 X-ray Absorption Spectroscopy Beamline	80.00	10.00
B9 Low-Dimension Structure Probe (LODISP) Beamline	20.00	5.00
BA Biological Macromolecule Microfocus Beamline	35.00	10.00
BB pink SAXS	400.00	50.00
BC High Res. Nanoscale Electronic Structure Spectroscopy Beamline	1.00	0.20
BD Tender X-ray beamline	10.00	1.00
BE Transmission X-ray Microscope Beamline	25.00	11.20
BF Test beamline	1000.00	60.00
Total average:		**805**

Data Challenges @HEPS

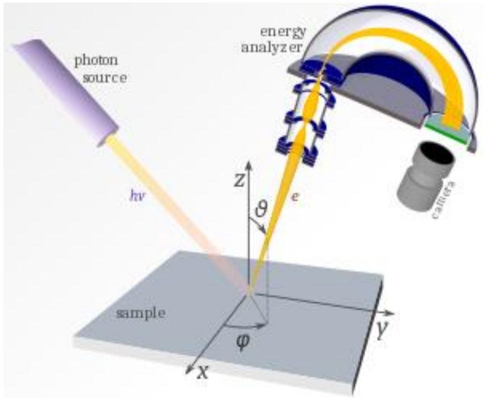
- New and more complex experiments
- Multi-modal experiments that combine data from multiple samples, techniques, and facilities
- In situ and in operando experiments require real-time feedback and autonomous control
- Data throughput and volume vary greatly with beamlines and scientific goals



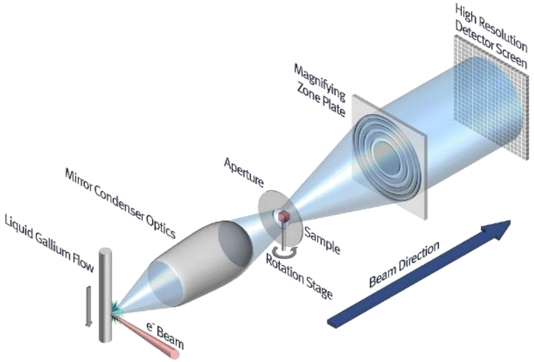
Fluorescence mapping



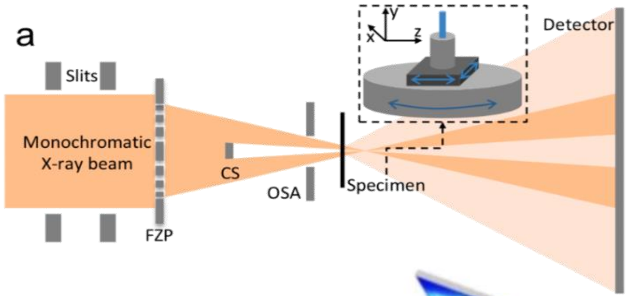
Nanoholotomography



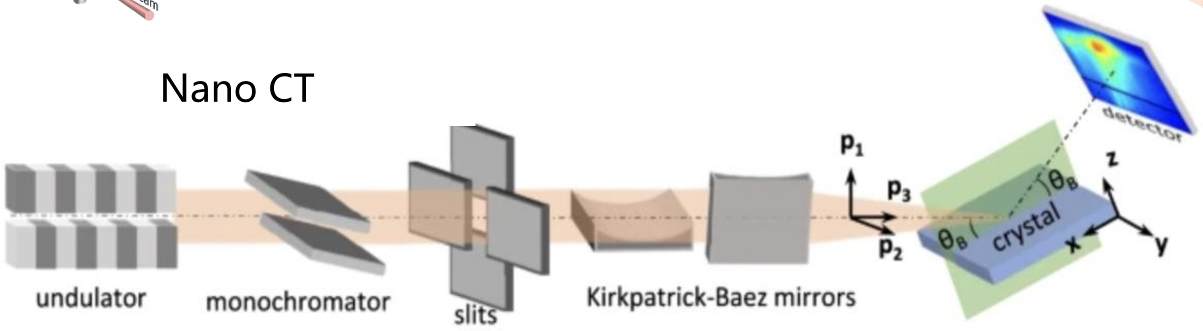
ARPES



Nano CT



Ptychography CT



Bragg ptychography

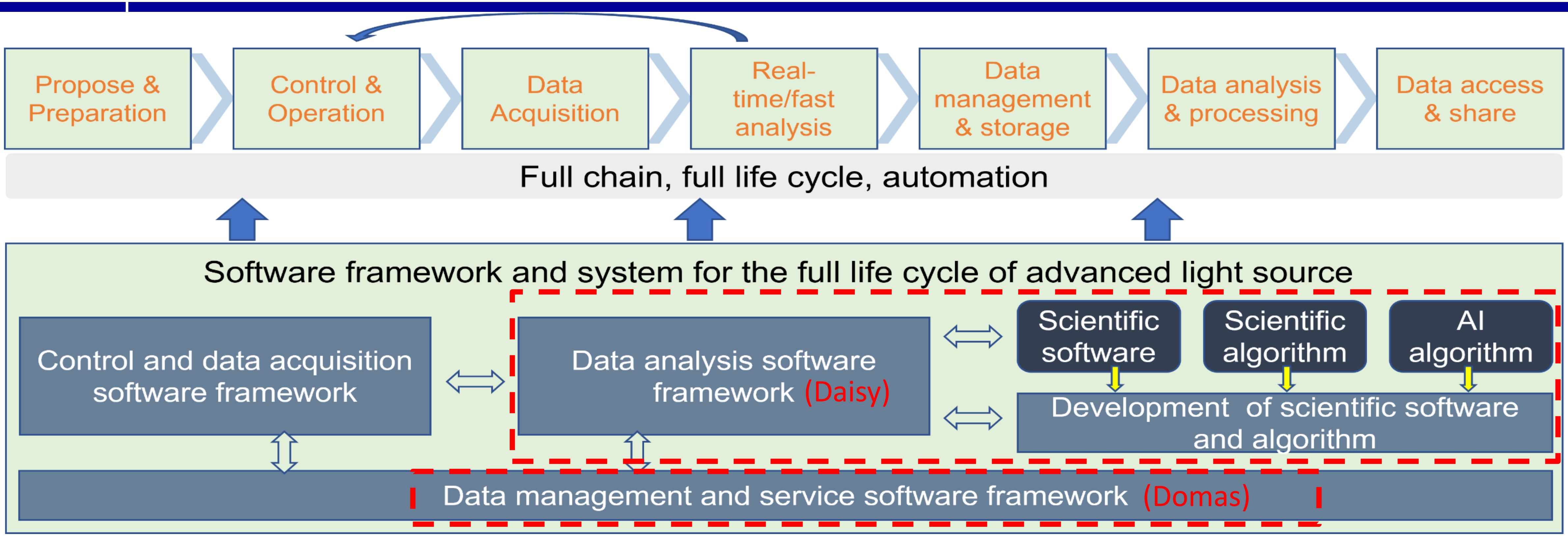
Data Challenges @HEPS

- Analysis and management of large datasets at synchrotron photon sources is becoming progressively more challenging
- Development and integration of advanced analysis and management tools is needed
  - Provide storage, organization and management of massive scientific data
  - During the experiment, provide real-time analysis and fast feedback to guide the experiment steering and optimize the data acquisition
  - After the experiment, process the massive offline data, accelerate the scientific discovery
  - Provide the scalable distributed heterogeneous computing power, meet the diverse computing requirements of different scientific goals

Outline

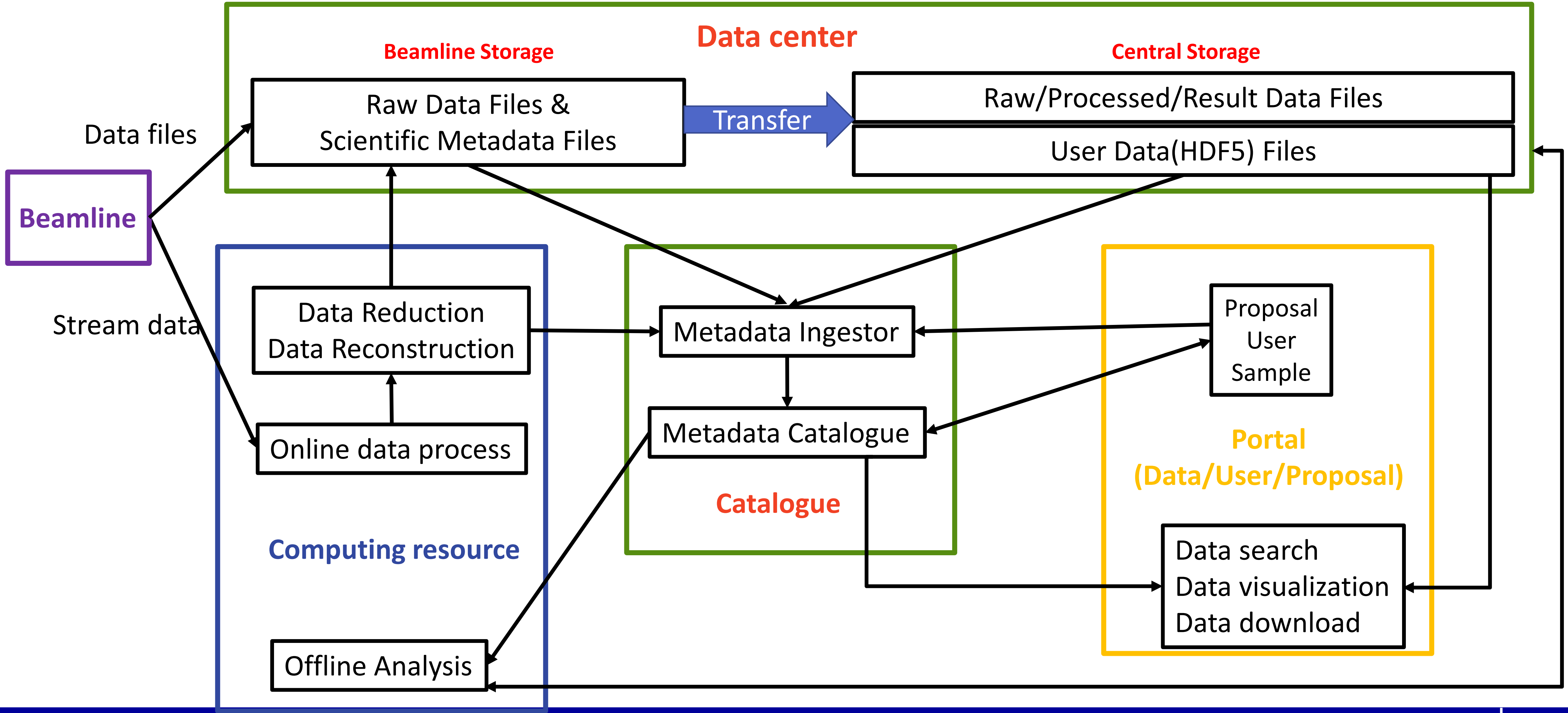
1. HEPS Introduction
2. Demand and Challenges of scientific data and software system
3. **The architecture and design of the framework**
4. The status of the system
5. Summary

Full data lifecycle software system



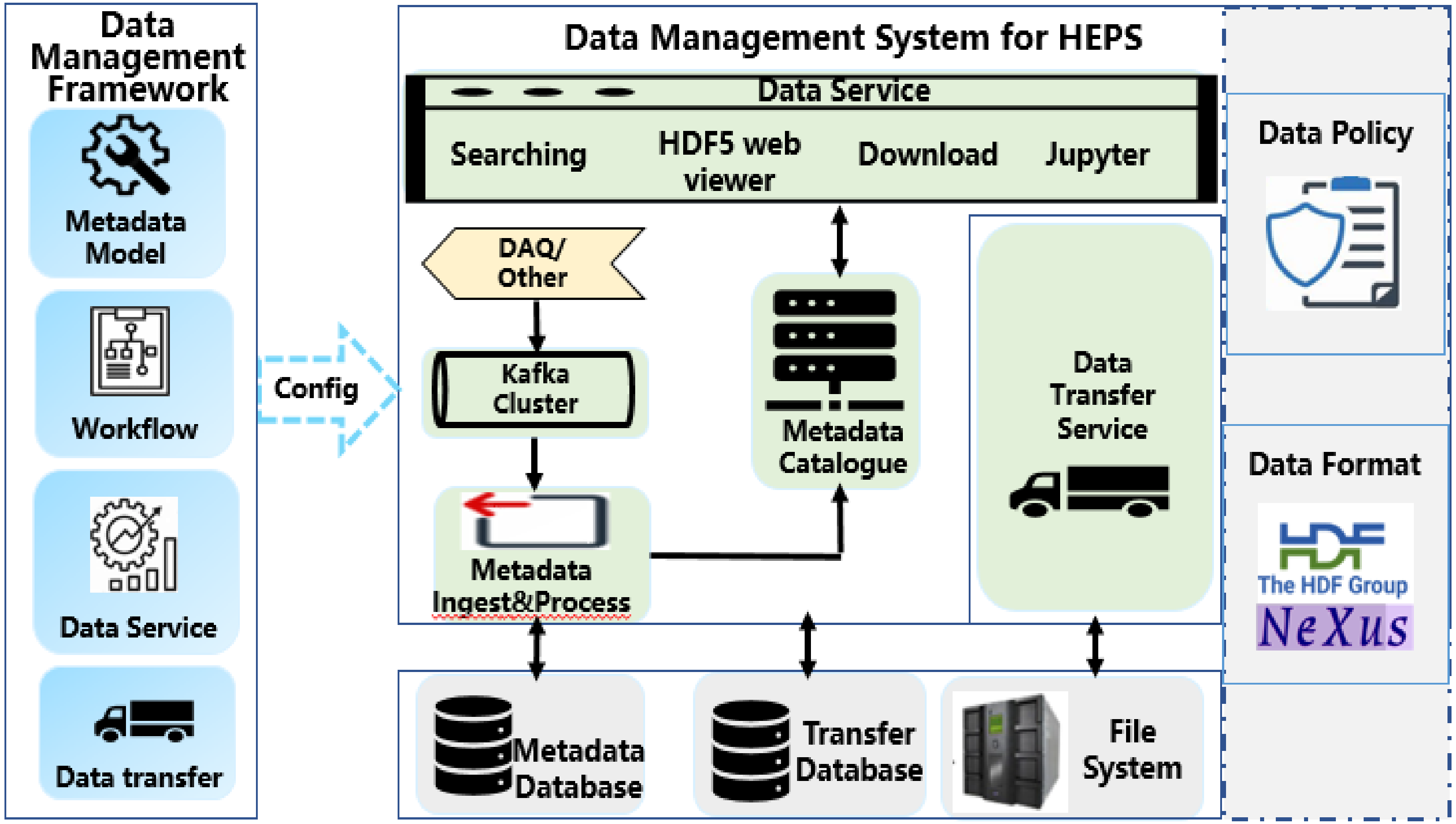
- Software framework and system for the full data life cycle of advanced photon source
- Promoting the intelligence and automation of the full lifecycle of photon source experiments
- Implement the tracking and management of scientific data throughout the full lifecycle
- Support the development of new advanced data analysis methods and software, as well as the integration of existing algorithm and software into the framework.

Data Flow of HEPS



Data management framework

- Common function modules of data management
  - ✓ Metadata Model
  - ✓ Workflow
  - ✓ Data service
  - ✓ Data transfer
- Extensible and standard interface
- Be able to build data management system suitable for facilities/beamlines quickly



HEPS Data Policy

The ownership, curation, archiving and access to scientific data and metadata

- Recommend providing at least 3 months disk storage and permanent tape archive (depends on final funding)
- Provide permanent storage for raw data
- Provide temporary storage for processed data, calibration data and result data
- Each dataset will have a unique persistent identifier(CSTR/PID21/doi)
- Experimental teams have sole access to the data during the embargo period.
- After the embargo, the data will be released with open access to any registered users of the HEPS data portal.

A draft version of ***The Data Policy for HEPS*** is finished, which will be discussed and approved by the HEPS council.

Reference:

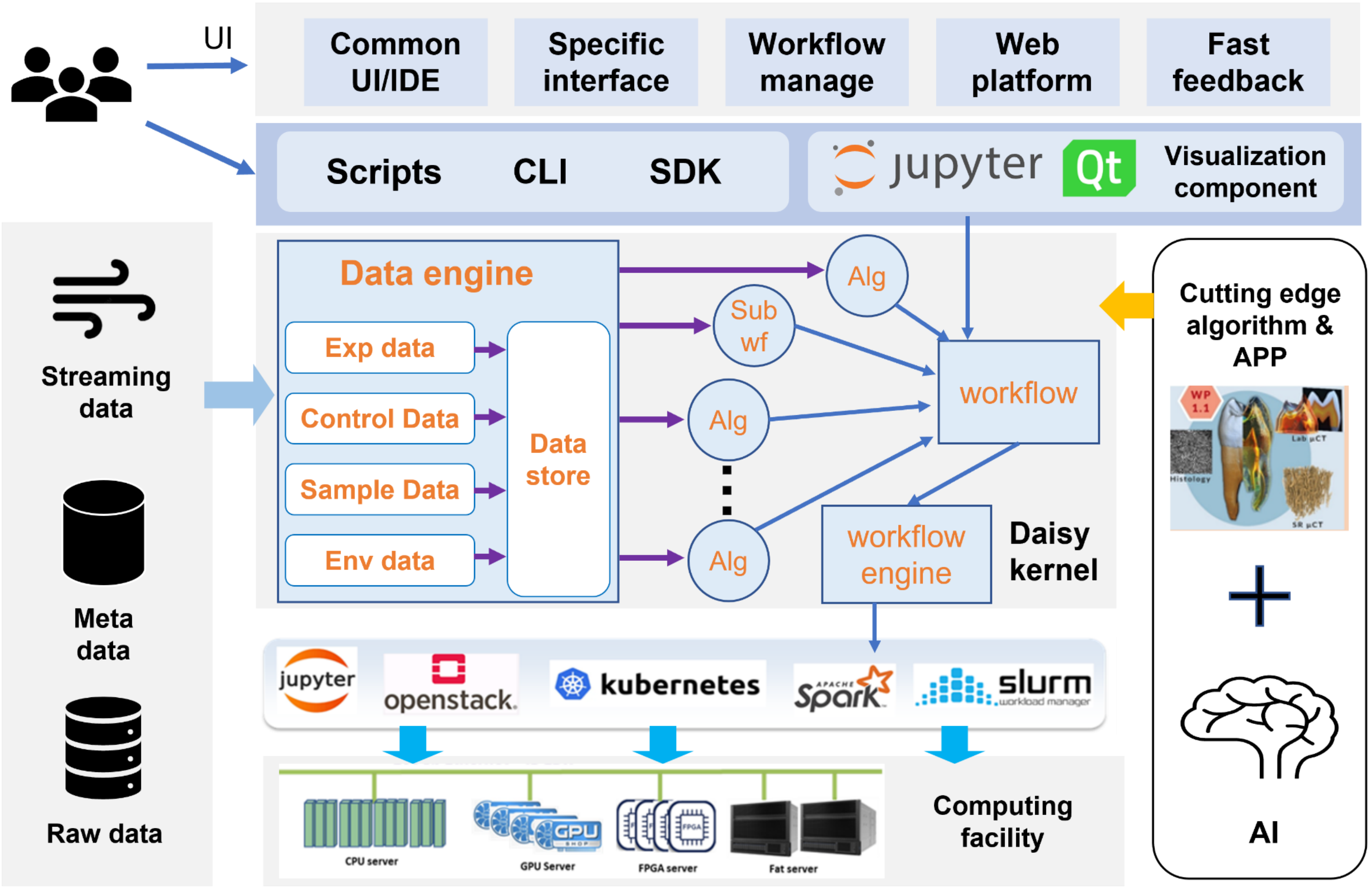
http://pan-data.eu/sites/pan-data.eu/files/PaN-data-D2-1.pdf

https://in.xfel.eu/upex/docs/upex-scientific-data-policy.pdf

Metadata items to cataloging & Acquisition

Metadata	Metadata Items	From
◆ **Administrative Metadata**	• Proposal Info, User Info, Exp type, Beamline…	Proposal system, User service system, Transfer system, Storage, Analysis system
	• Data type: raw data, processed data, simulated data, calibration data	
	• Dataset： PID, Path, Data file list, file size, checksum…	
	• Status: disk/tape, transfer status, transfer check value	
	• Analysis software, update time…	
◆ **Scientific Metadata**	• Sample Info	Sample database, Proposal system,
	• Exp environment params：voltage、magnetic field、electric field… Detector Info: scan, x-ray exposure params…	DAQ system, Control system
	• E-log	E-log System

Data analysis software framework—Daisy



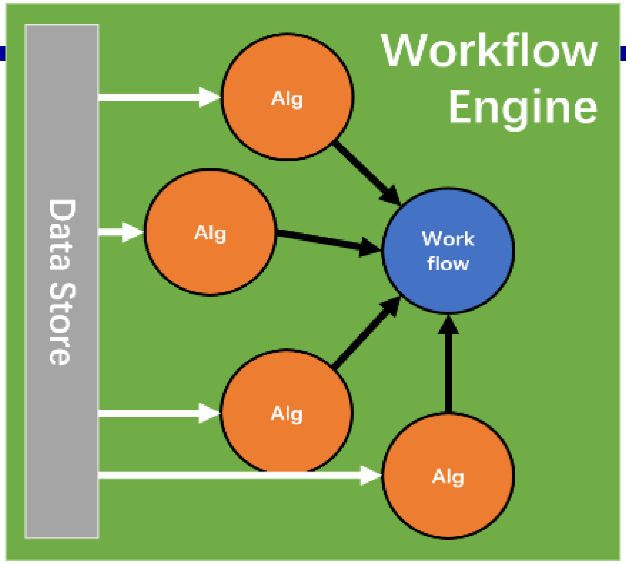
- Kernel of the framework
- Derivative technology modules to meet the data processing requirements of new generation photon sources
  - Data object management module for high-throughput data I/O, multimodal data exchange, and multi-source data access.
  - Scalable cluster computing power support for data processing with different scales, different throughputs, and low latency
  - Interface and developing environment for scientific software integration and development
- Domain specific App and flexible general workflow management system based on the framework

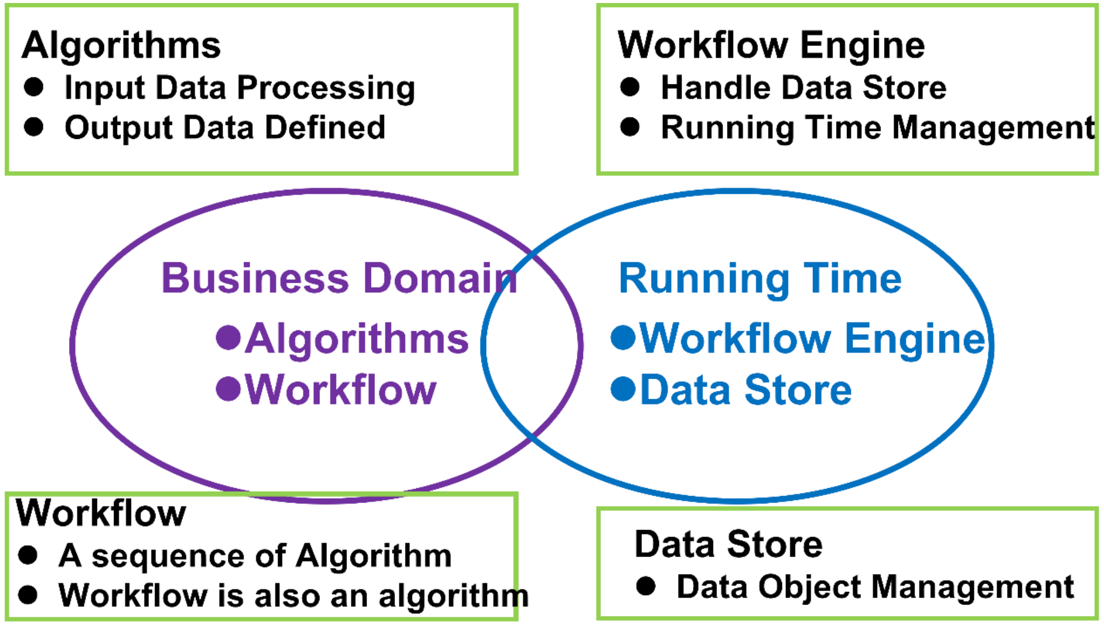
Kernel of the Daisy framework

Extract domain models independent from technology, and establish relationships between models to form a domain architecture

Four core modules are provided:

- **Algorithm:** The smallest unit in framework, defining the domain model, basic data processing module, support integration of third-party libraries.
- **Workflow:** Defines the domain architecture, execute processing tasks by calling a series of algorithms, supporting nesting.
- **Workflow Engine:** Manages the runtime environment and the distribution of the algorithm modules. Uncouple the process task from the computing environment.
- **Datastore:** Manages the creation and transmission of data objects between algorithms.

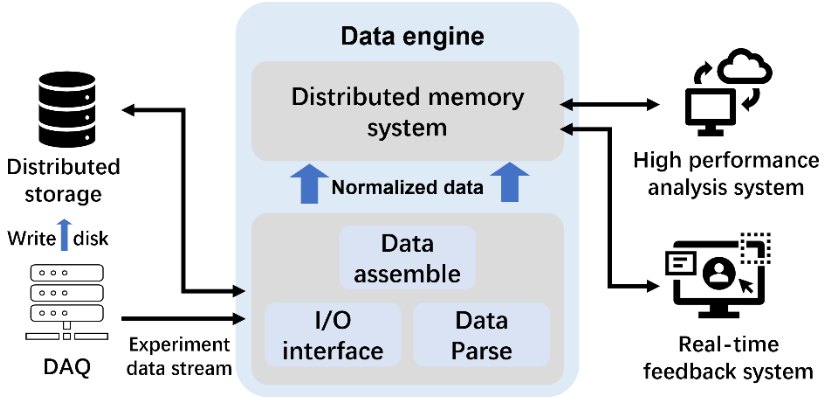




High performance modules

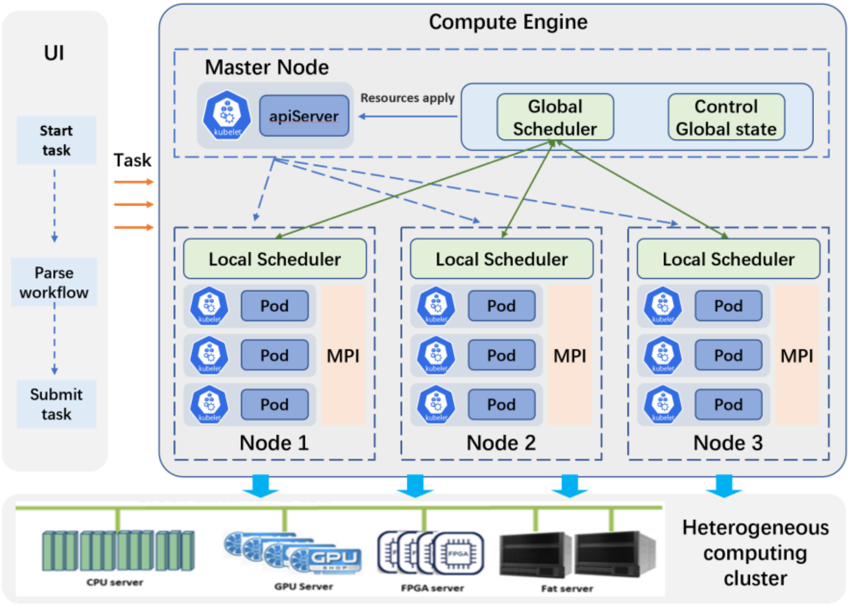
■ Data object management

- ☐ Unified I/O interface to shield the difference of underlying architecture and data structure
- ☐ Support the I/O of stream data besides disk file, for real time, high throughput data process
- ☐ Employ asynchronous parallel, distributed memory, adaptive storage parameters and compression to optimize the I/O



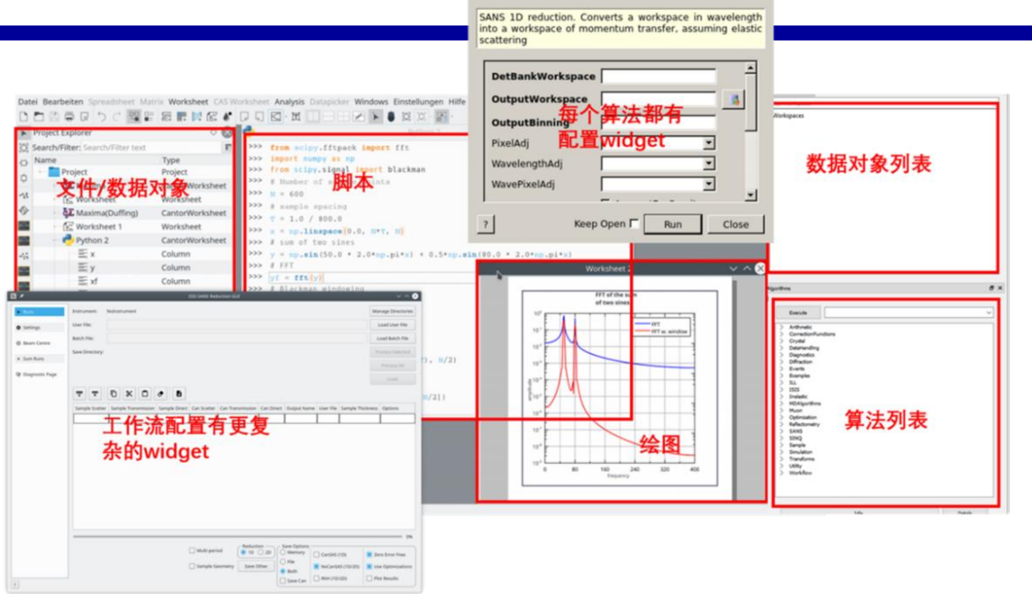
■ Heterogeneous distributed computing power support

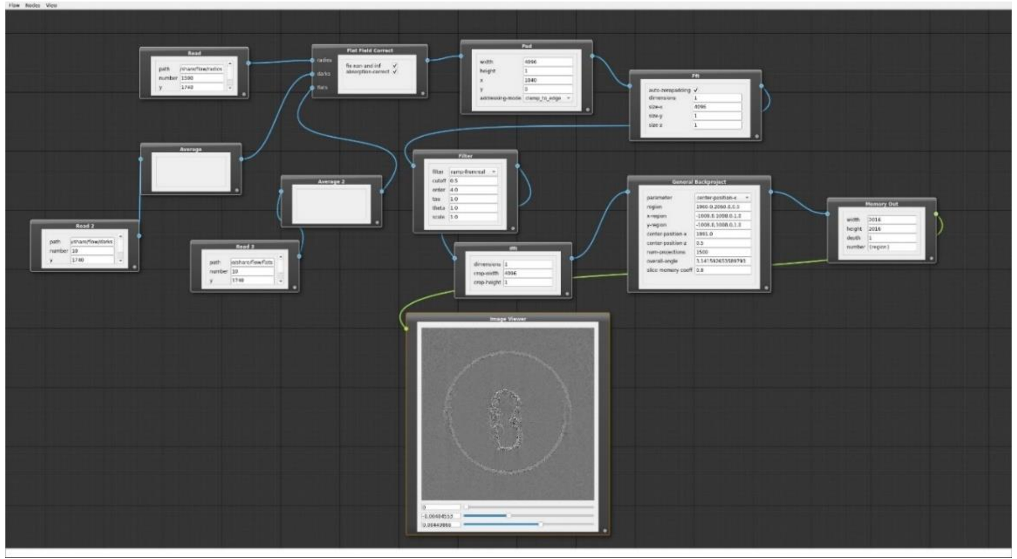
- ☐ High performance numerical analysis computing library to speed up the computing hot spots
- ☐ Provide a unified flexible programming interface API for computing models, to reduce the complexity of parallel programming
- ☐ Distributed computing task scheduler to achieve better efficiency



User interface for scientific applications

- Variety of user interfaces to support different scientific applications
  - Data visualization interface
  - Integrated development environment interface
  - Specific scientific interface
  - Web data analysis platform
  - Script and command line interfaces
- Provide a variety of reusable common widgets, support secondary development
  - Widgets for analysis, drawing, browsing, configuration, object list……
- Workflow management system, for flexible and general data process task
  - App and Web GUI, support interactive workflow creating, import, export and operation monitoring
  - Follow the Common Workflow Language(CWL) standard



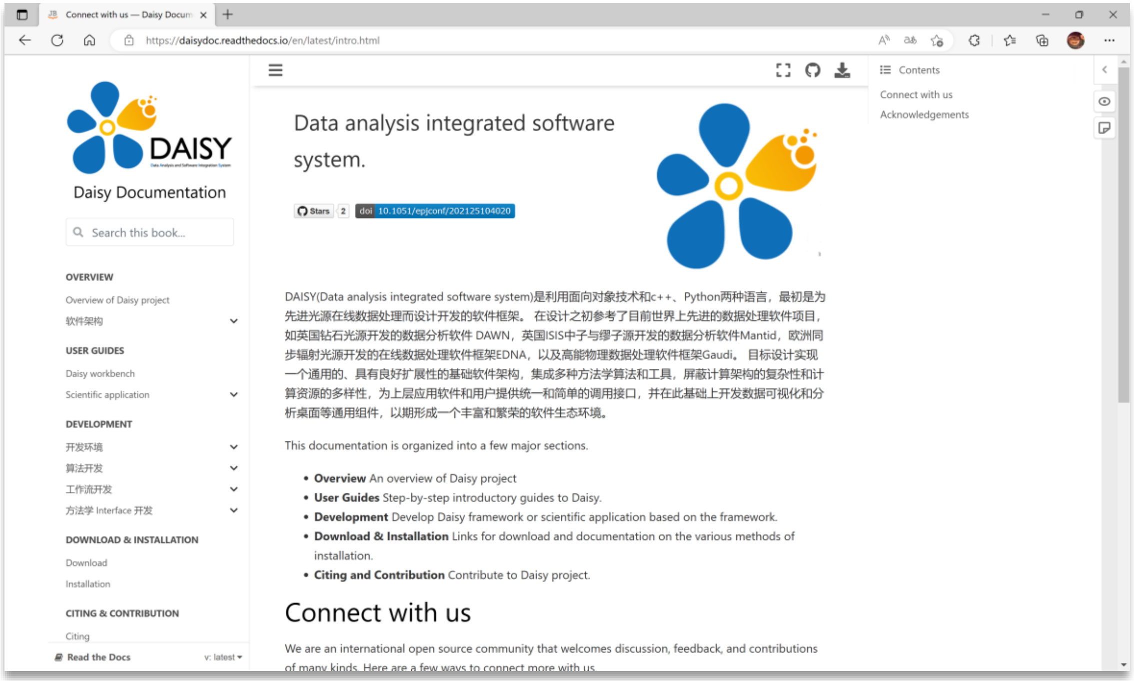


Outline

1. HEPS Introduction
2. Demand and Challenges of scientific data and software system
3. The architecture and design of the framework
4. **The status of the system**
5. Summary

Daisy framework

- The basic software framework of general scientific data processing Daisy is designed and implemented
- Four types of basic interfaces are provided：
  - Algorithm and workflow module: Implement domain model and processing task
  - Workflow engine and datastore module: Manage the software runtime environment and data objects
- Provide several types GUI: general-purpose GUI, domain specific GUI, user development IDE, web platform
- Open source, user documentation provided
- Integrated several scientific software and algorithms, developed several domain specific GUI

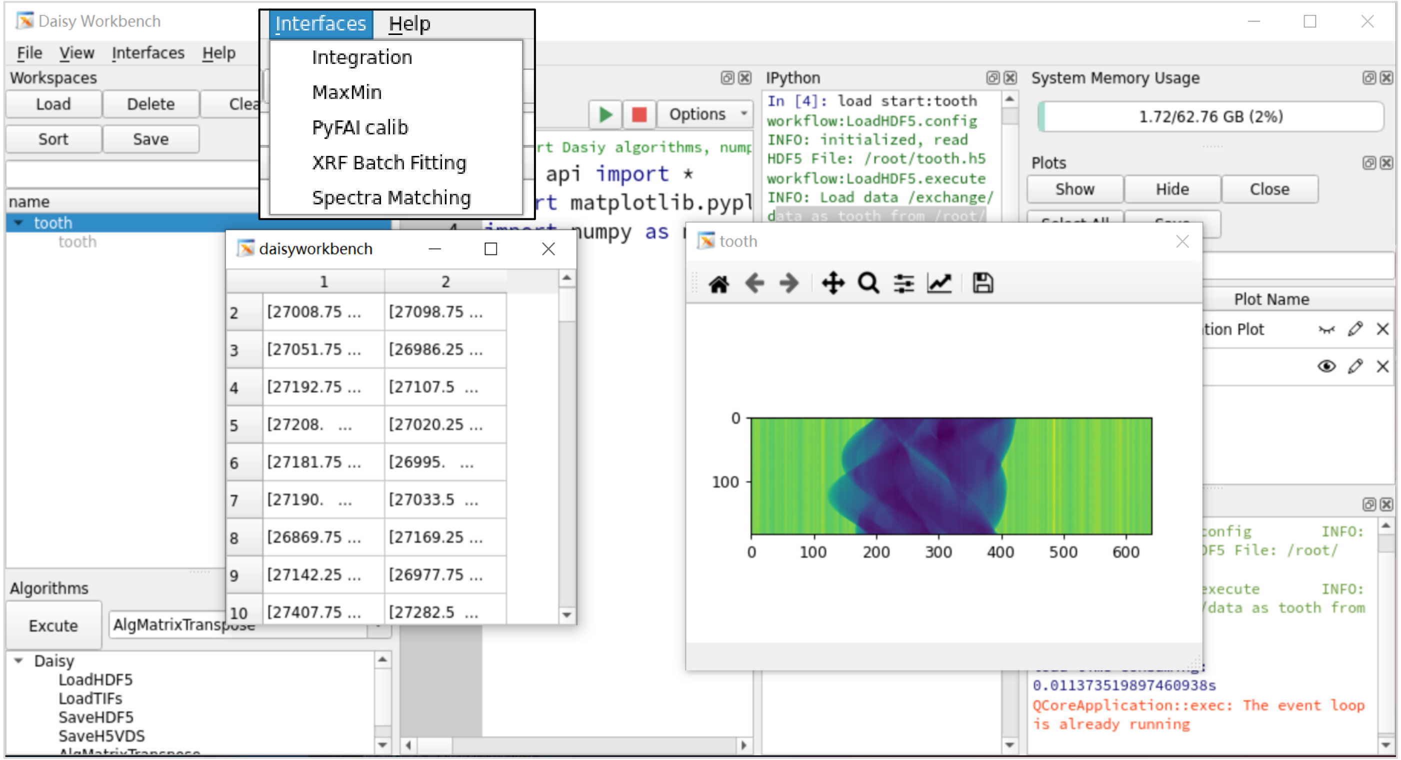


User documentation：

https://daisydoc.readthedocs.io

Yu Hu et al. EPJ Web of Conferences 251, 04020 (2021).

Daisy graphical user interface



Daisy workbench:

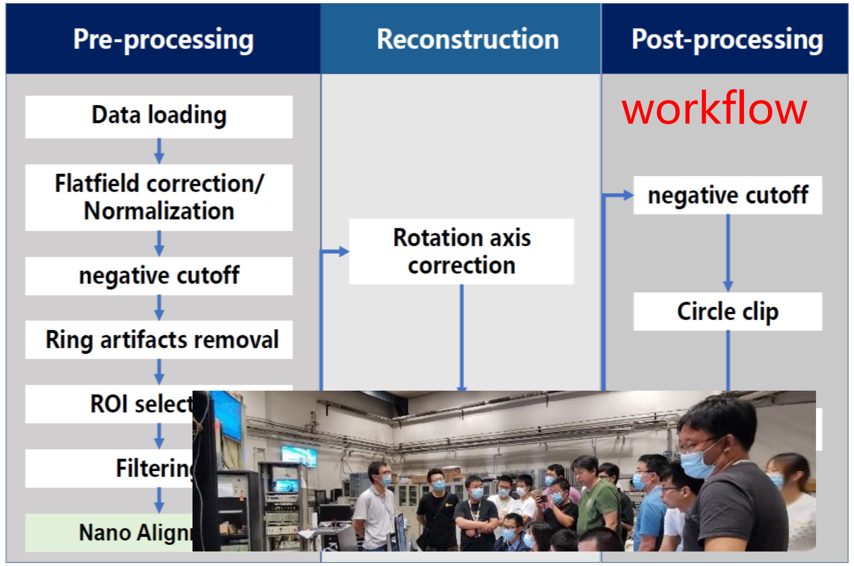
- General-purpose GUI based on PyQt5
- Include data object list, algorithm list, data view/visualization, and IDE for developers
- Interfaces of custom GUIs for a variety of scientific techniques

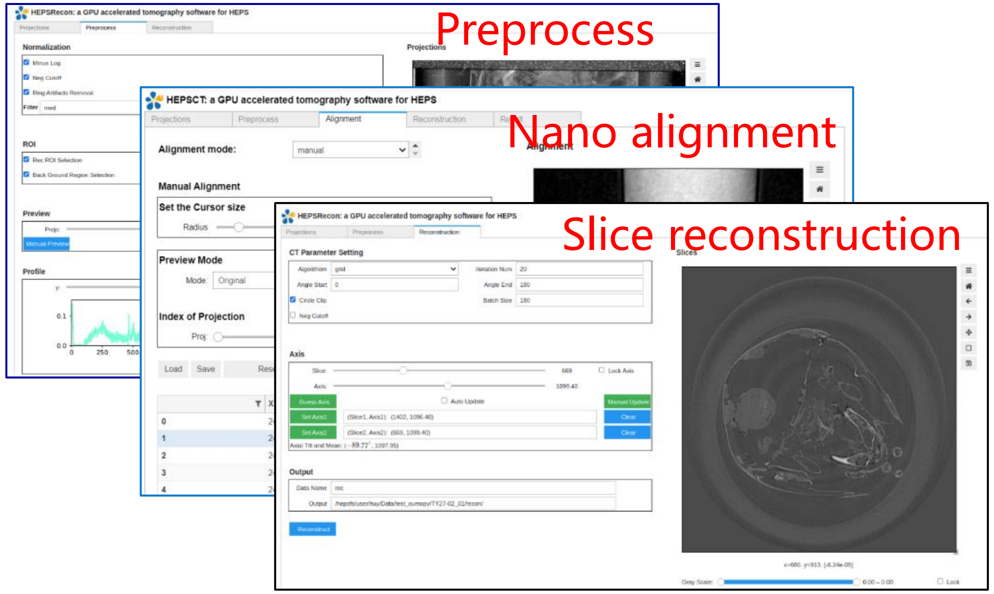
Web data analysis platform:

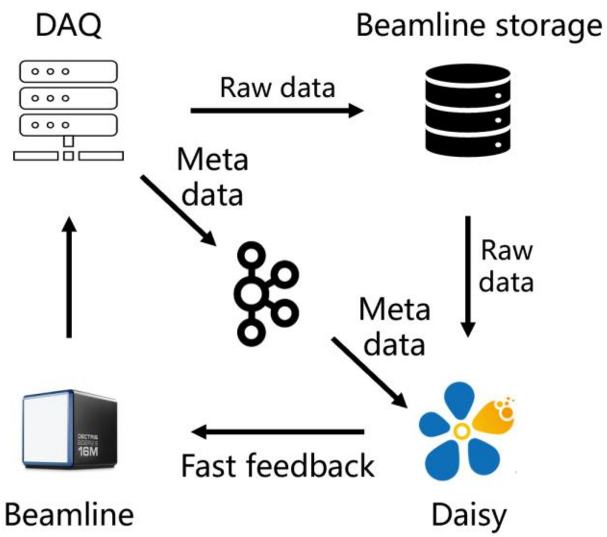
- Based on the jupyterlab ecosystem
- Container encapsulates the computing environment
- Scalable computing resource
- Terminal and web scientific APP

Web based application for X-ray CT

- Web-based interactive data processing interface, integrates self-developed software HEPSCT
- Implement the reconstruction of micro CT and nano CT with different data formats (HDF5， tiff)
- Deployed on the Web data analysis platform， will support multiple beamlines of HEPS
- Validated on the BSRF 3W1A test bed. Implemented the automated data processing pipeline with DAQ system and HEPS-B7 beamline

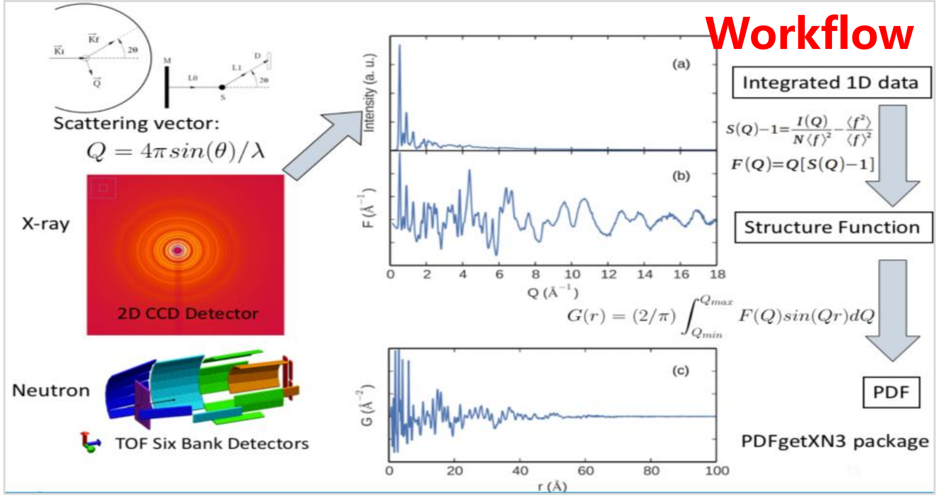


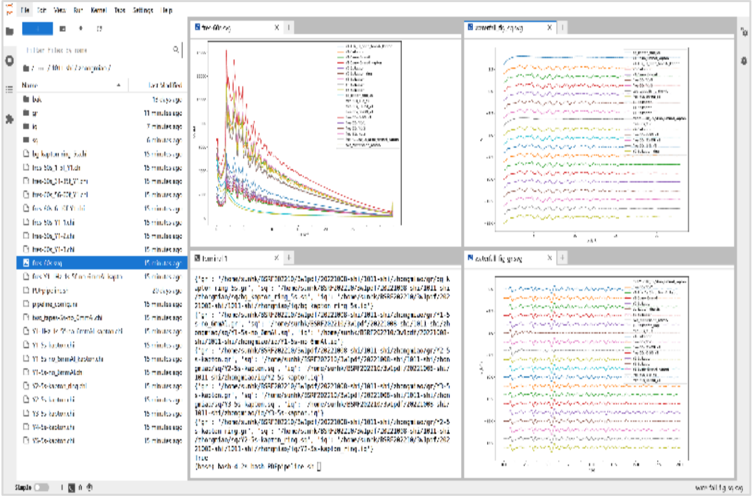




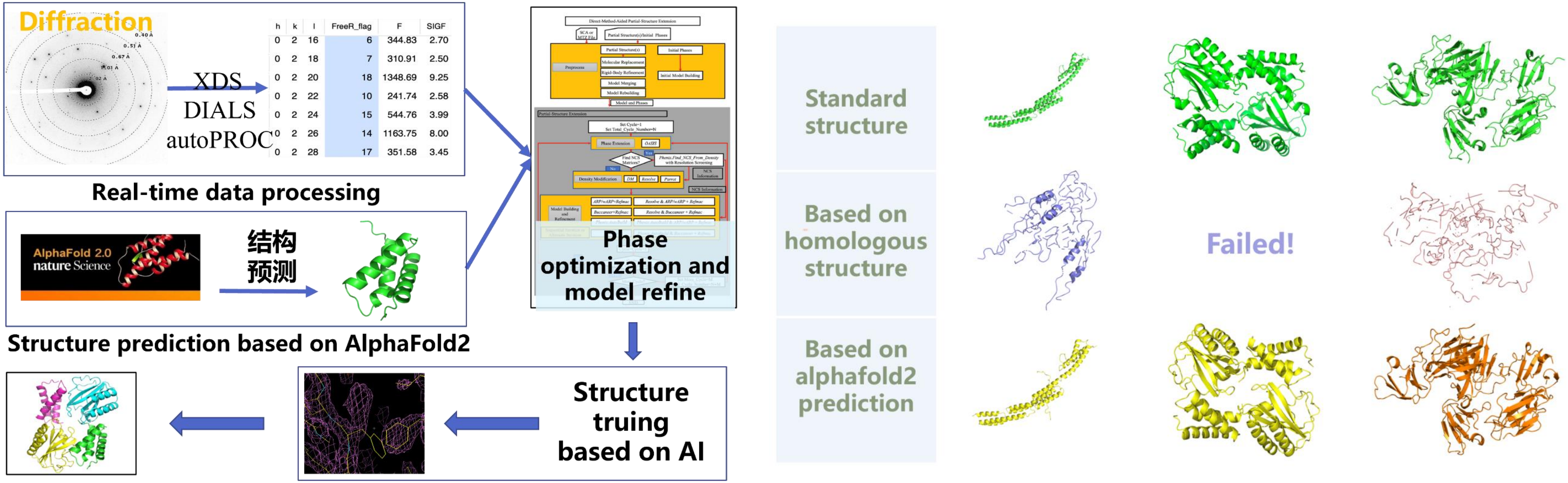
Application for Pair distribution function(PDF)

- Serve for total scattering experiment
- Developed PDFHEPS python package, integrated several scientific software, such as PyFai and PDFgetX3
- Implement a pipeline from raw data to PDF， include background reduction, masking, azimuth integration and PDF data transform
- A web-based GUI also provided for interaction and data visualization
- More function will provide in the future, such as Similarity Mapping, Structure Mining via machine learning





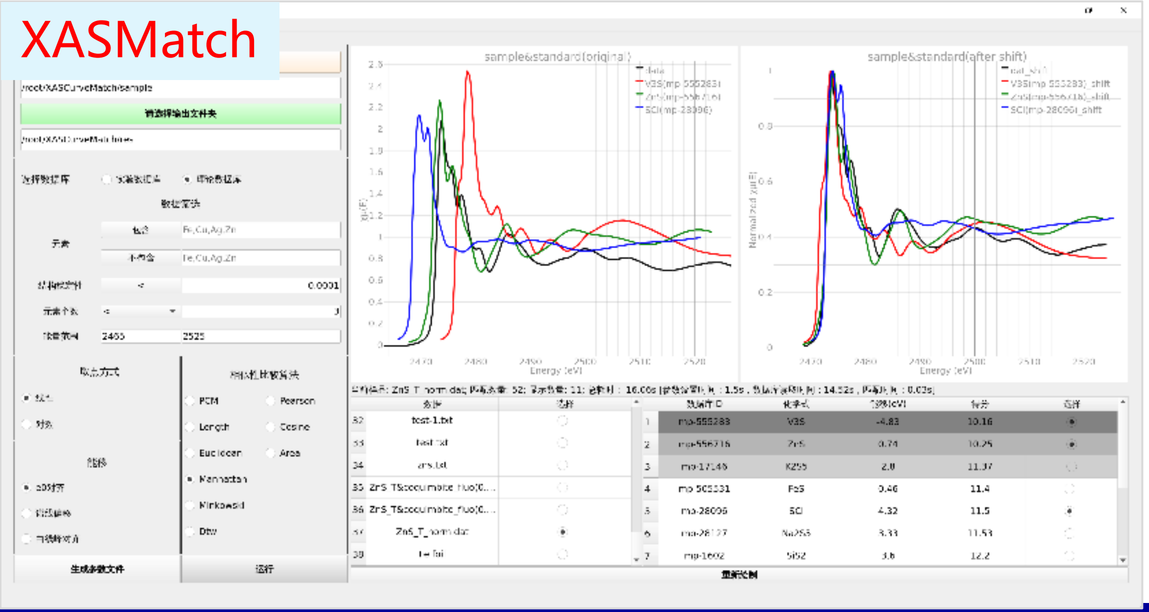
AI-based application for biological macromolecule

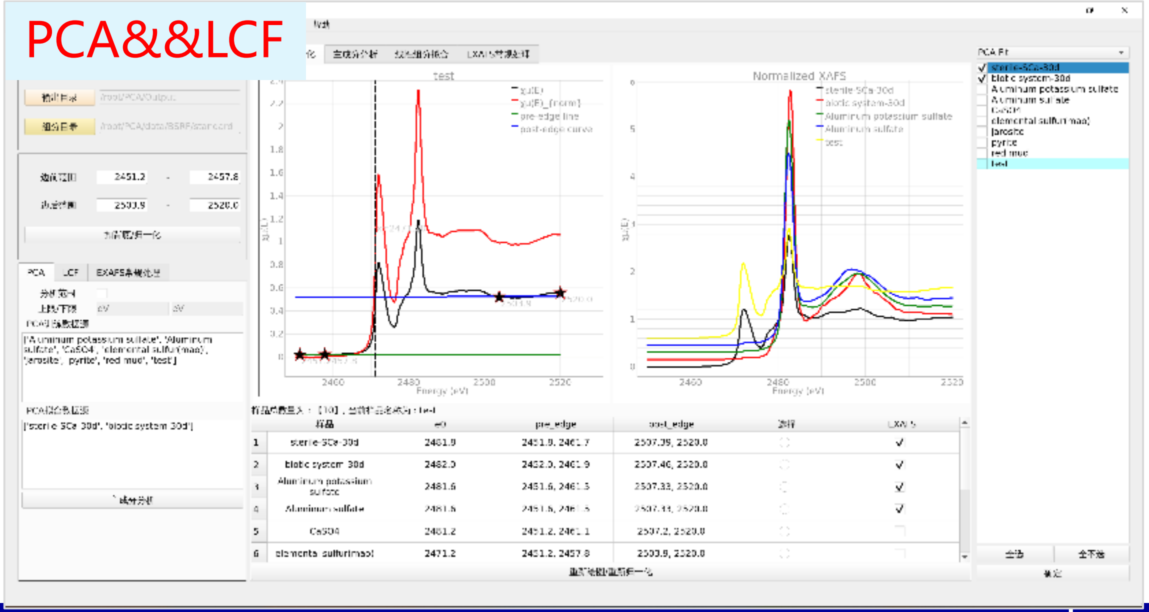


- Automatic pipeline based on direct method, structure prediction and AI
- Include 3 modules: data processing, structure prediction, model building
- Module of structure truing based on AI will be added in the future
- Based on alphafold2, the success rate and accuracy of macromolecular structure reconstruction get improved

Applications for X-ray absorption spectroscopy

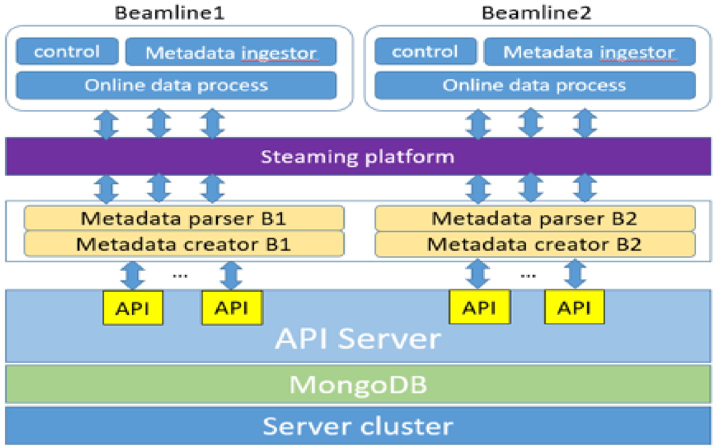
- Based on PyQt5, for spectroscopy components analysis
- XASMatch
  - Fast matching of experimental spectroscopy from database
  - Integrate multiple matching and energy shift methods, support multiple input databases
- PCA&&LCF
  - Spectroscopy components analysis via PCA and LCF method
  - Automatic pipeline, batch processing, multi-standard spectroscopy input



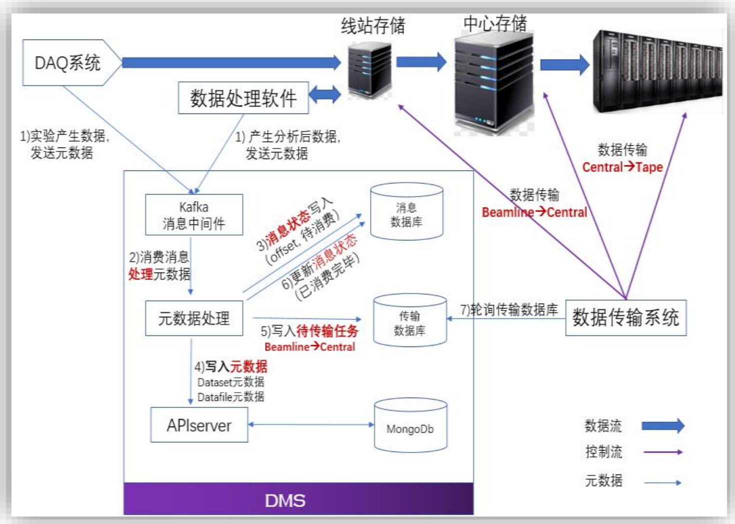


The progress of data management

1. **Finished the core function modules**
   - ✓ Metadata catalogue
   - ✓ Metadata ingestion
   - ✓ Interfaces with other systems: control system, transfer module, storage system, analysis system
   - ✓ Automatic data management when using Hierarchical storage policy (**beamline storage → central storage → tape**)
2. **Finished the design of data management scheme when network interrupts**
   - ✓ when network interrupts, metadata and data are saved to local disk
   - ✓ After the network recovers, metadata will be sent to be catalogued
3. **HEPS data format design**
   - ✓ Designed and released data format for 5 beamlines



Metadata catalogue



Automatic data management flow

HEPS CC system integration/Test bed/Production

Set up testbed, integrate full data lifecycle software systems to verify the system interfaces, run in the real experimental environment, move to production gradually.

1

Oct, 2020, BSRF 1W1A

Simple verification of the data management system

- Network bandwidth is 1Gb/s
- Beamline storage: **2TB** NAS, Dell EMC NX3240, NFS file system
- Central storage: **80TB** disk array, Lustre file system
- Metadata ingest, catalogue, data transfer, data service

2

July, 2021, BSRF-3W1 test beamline

- Network bandwidth updated to 10Gb/s
- Beamline storage & Central storage: **80TB** disk array, Lustre file system
- Integrate DAQ system, data management system, analysis software framework, computing cluster

3

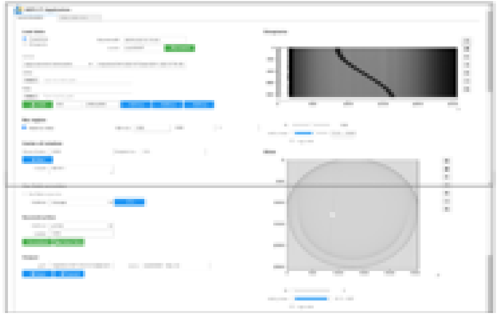
June, 2022, BSRF 4W1B

Running in production environment

- Network bandwidth updated to 25Gb/s
- Beamline storage: Huawei Ocean Store 9950
- Central storage: 80TB disk array, Lustre file system
- **Follow real experiment process, provide Pymca to do analyzing**

Data acquisition

Analysis framework Interface



CT reconstruction

Integration test at BSRF

Outline

1. HEPS Introduction
2. Demand and Challenges of scientific data and software system
3. The architecture and design of the framework
4. The status of the system
5. **Summary**

Summary

- The system design has been finished
- Cooperation with other facilities and community is ongoing
- The basic framework has been stable and tested on the test bed
- Based on the framework, scientific software integration and application development are ongoing
- The development of scientific software ecosystem also needs the support and participation of user community

Thanks！

Back up

Tasks & Goals of Data Management

- **Data policy and Data Format**
  - The ownership, curation, archiving and access to scientific data and metadata
  - HDF5 is chosen as the standard data file format, follows NeXus conventions
- **Metadata catalogue**
  - Support the management of the whole scientific data lifecycle
  - Hierarchical storage: beamline storage → central storage → tape
  - Catalogue and Catalogue and provide application to metadata
- **Metadata acquisition**
  - Ingest metadata from other sub-systems(DAQ, transfer, storage, analysis)
- **Data transfer**
  - Transfer all the data between beamline storage, central storage and Tape
  - Interact with metadata catalogue when the data storage status changed
- **Data service**
  - Provide a web-based GUI for user to search, access, download, analysis data

国家高能物理科学数据中心
National HEP Science Data Center

高能所计算中心
IHEP Computing Center

HEPS-CC