The scientific data analysis software framework for HEPS

Yu Hu (on behalf of HEPSCC)

Institute of High Energy Physics, CAS

2024/07/19
ICHEP @ Prague

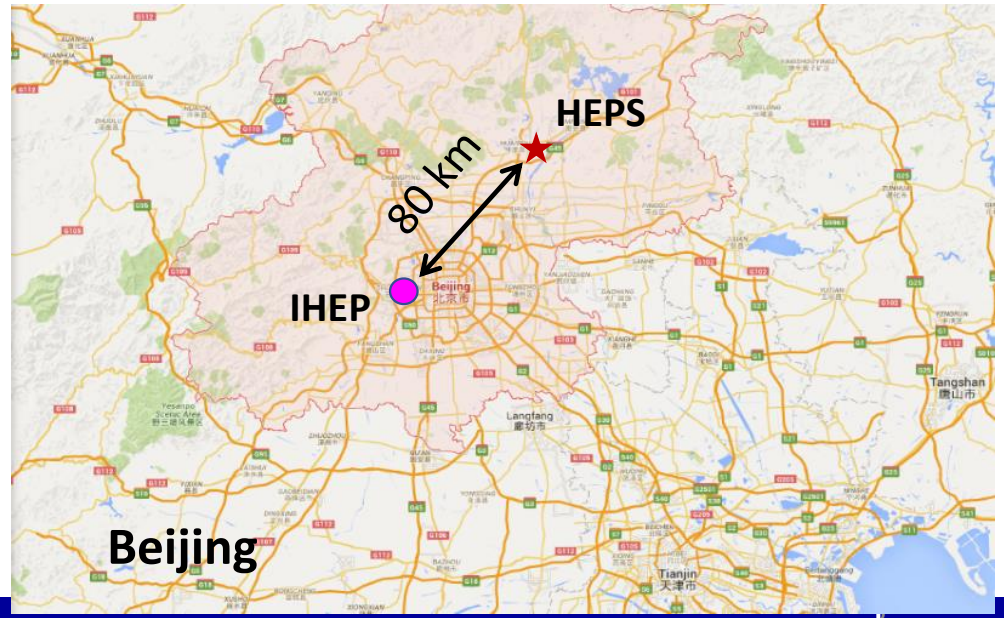
Outline

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

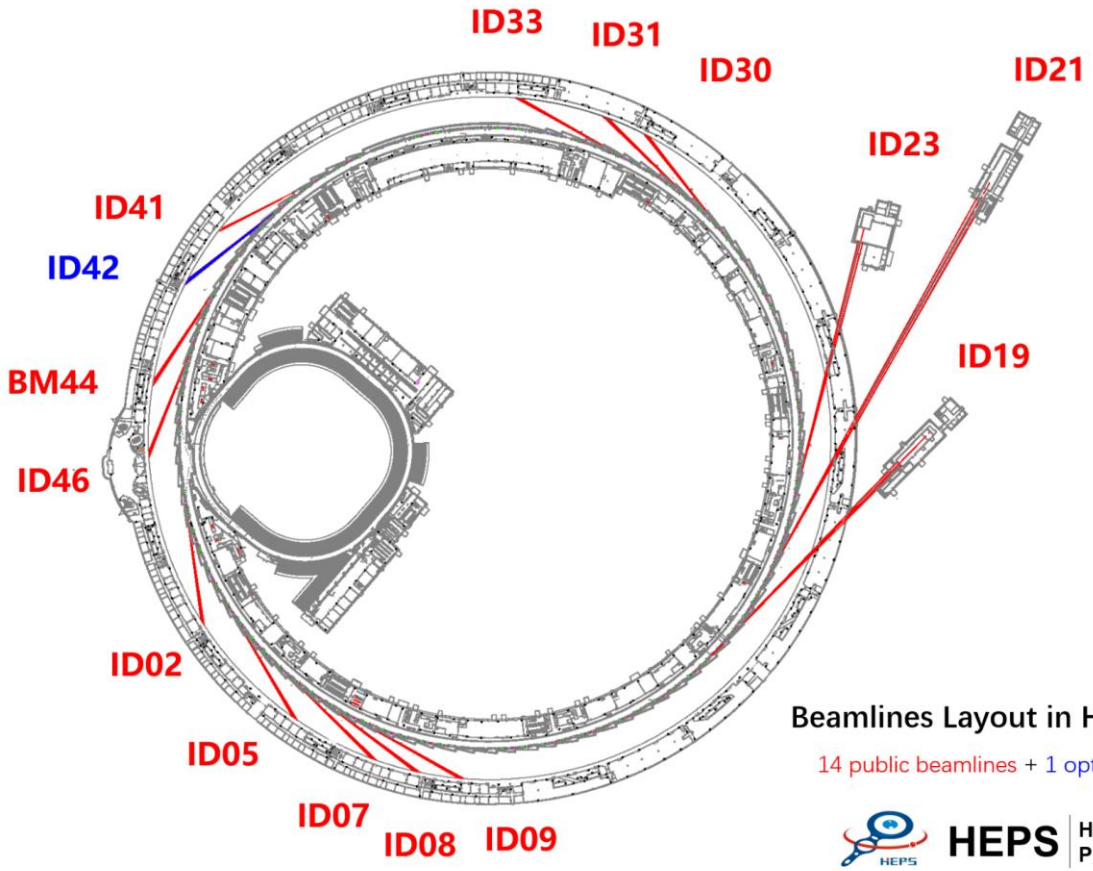
High Energy Photon Source (HEPS)

- 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 in the mid-2019
- 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/mm^2/mrad^2/0.1%BW	>1x10^{22}
Beam current	mA	200
Injection		Top-up



Beamlines in HEPS phase I



Beamlines Layout in HEPS phase I

14 public beamlines + 1 optics test beamline

HEPS HIGH ENERGY PHOTON SOURCE

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

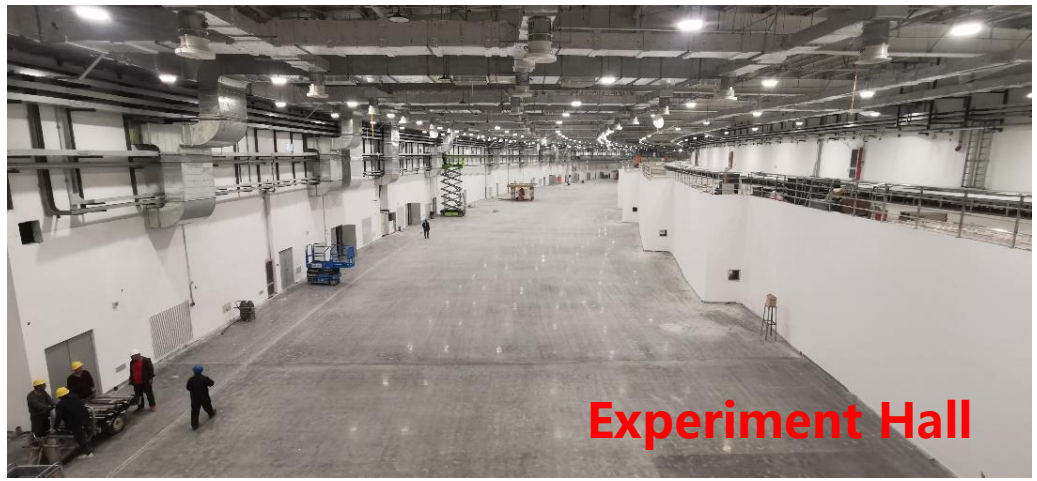
ID33-Hard X-Ray High Resolution Spectroscopy Beamline

BM44-Tender X-Ray Beamline

ID41-High Resolution Nanoscale Electronic Structure Spectroscopy Beamline

ID42-Optics Test Beamline

ID46-X-Ray Absorption Spectroscopy Beamline

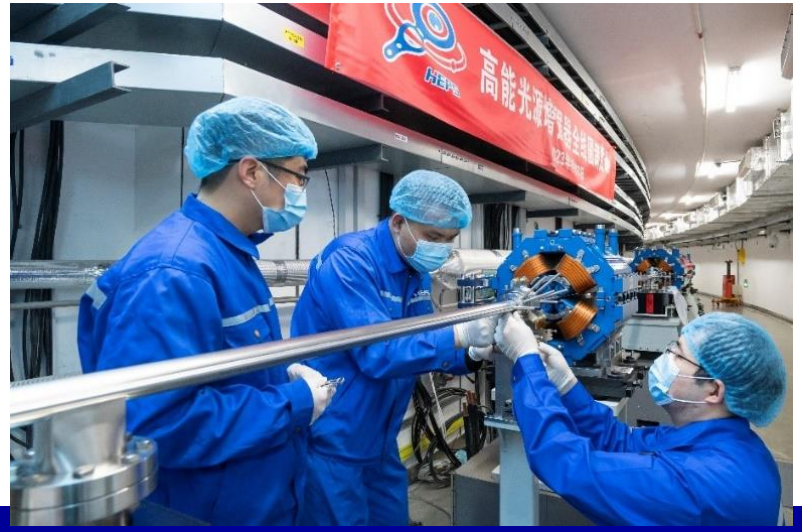


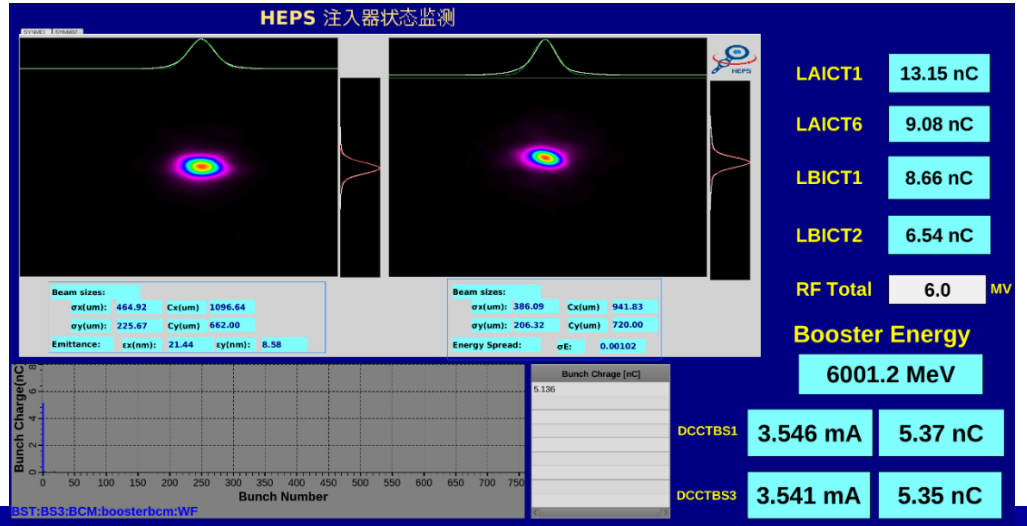
14 public beamlines + 1 optics test beamline in Phase I

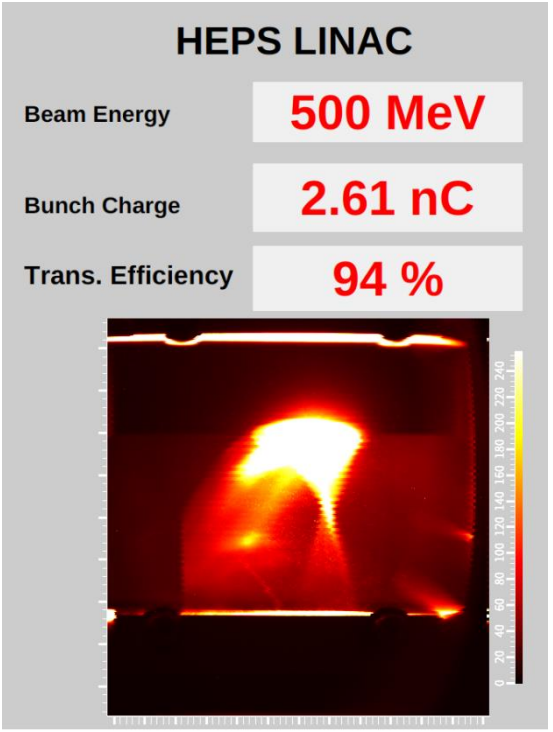
Can accommodate over 90 beamlines in total

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
- 2023.03, HEPS achieved the first electron beam accelerated to 500 MeV
- 2023.11, Electron beam ramped up to 6 GeV
- 2024.07, HEPS storage ring installation completed
- 1st SR X-ray to be emitted in the near future





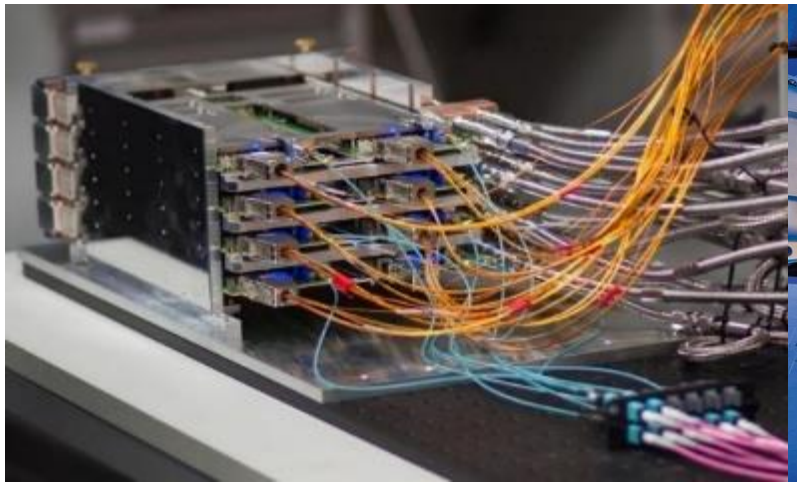


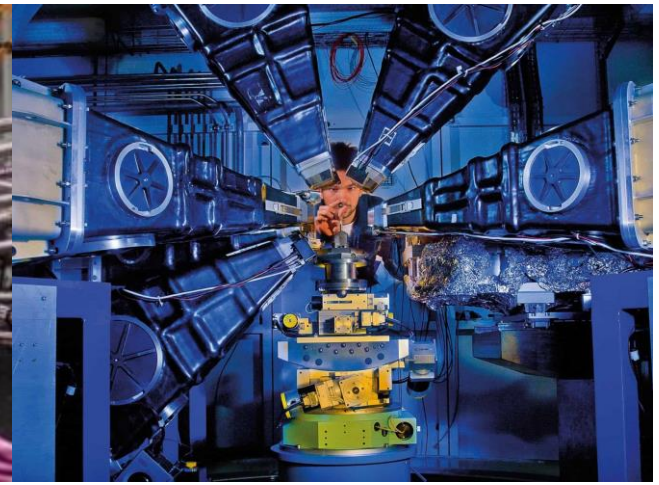
Outline

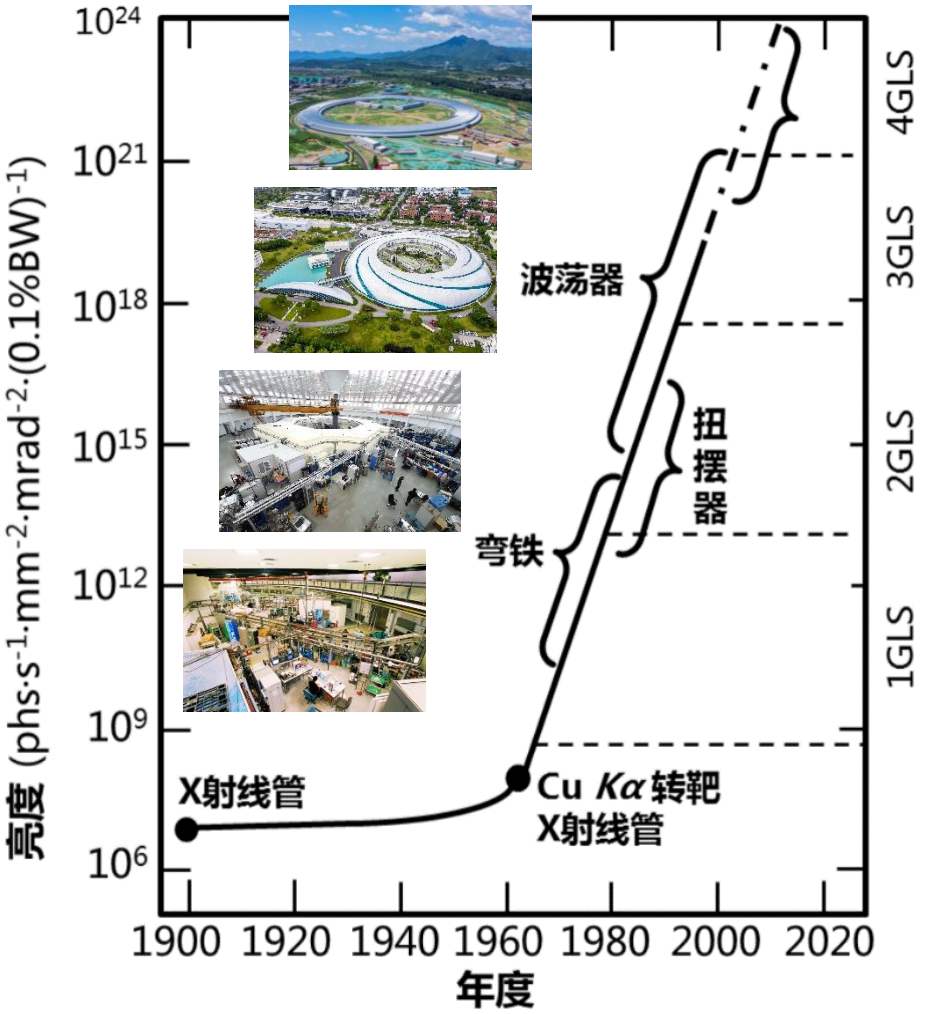
1. Introduction
2. **Demand and Challenges of scientific data and software system**
3. The architecture and design of the framework
4. The progress of the framework
5. Summary

Data Challenges @HEPS

- 10^2~10^3 higher brightness than 3rd gen. SR sources
  - More raw data in greater detail and less time
- Detector capabilities constantly improving:
  - Increased dynamic range, faster readout rates, larger pixel arrays（e.g. 32bits，20KHz，30kx30k）
  - Bigger frames, higher frame rates => more raw data



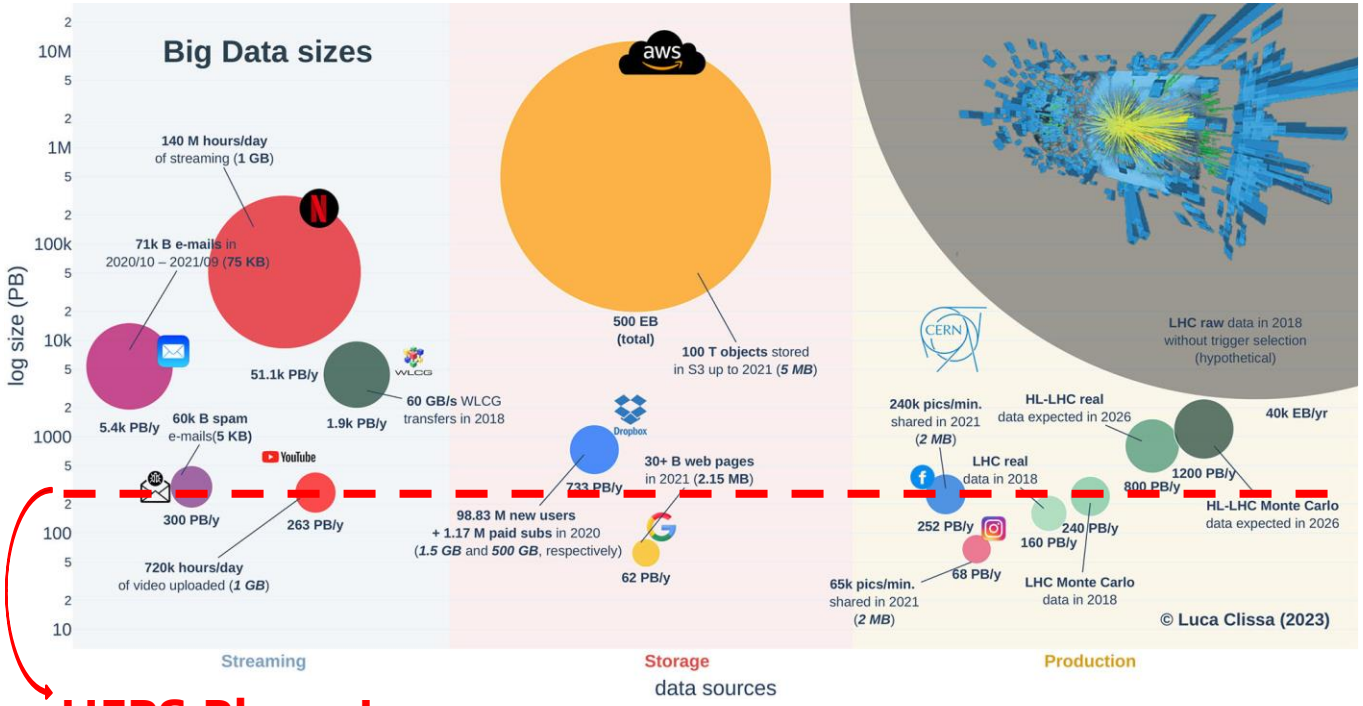




Development of synchrotron radiation light source

Data Challenges @HEPS

- >200PB raw data per year for HEPS Phase I (15 beamlines)
- More than 90 beamlines volume in total
- Data volume will soon reach the EB scale



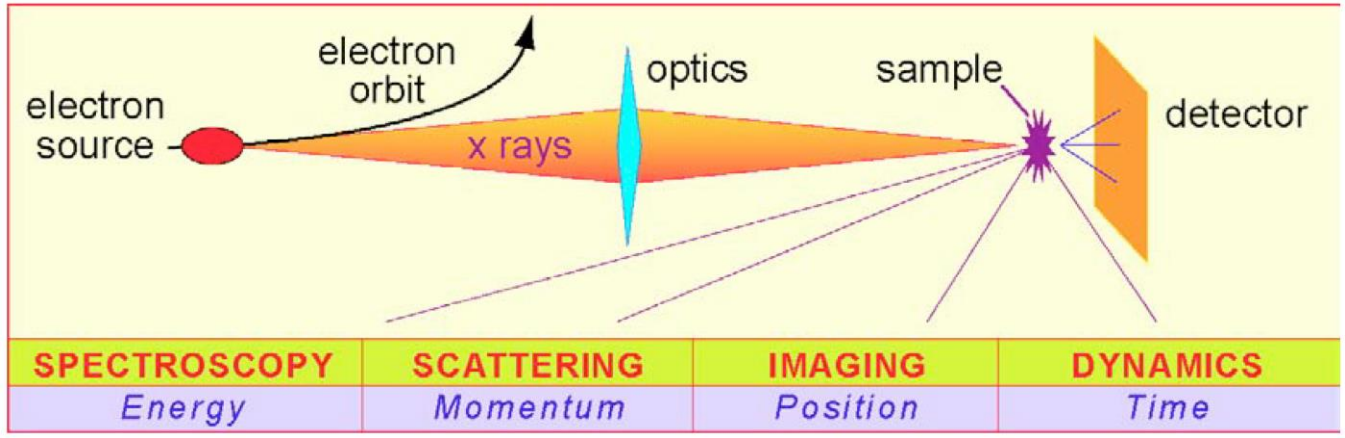
HEPS Phase I

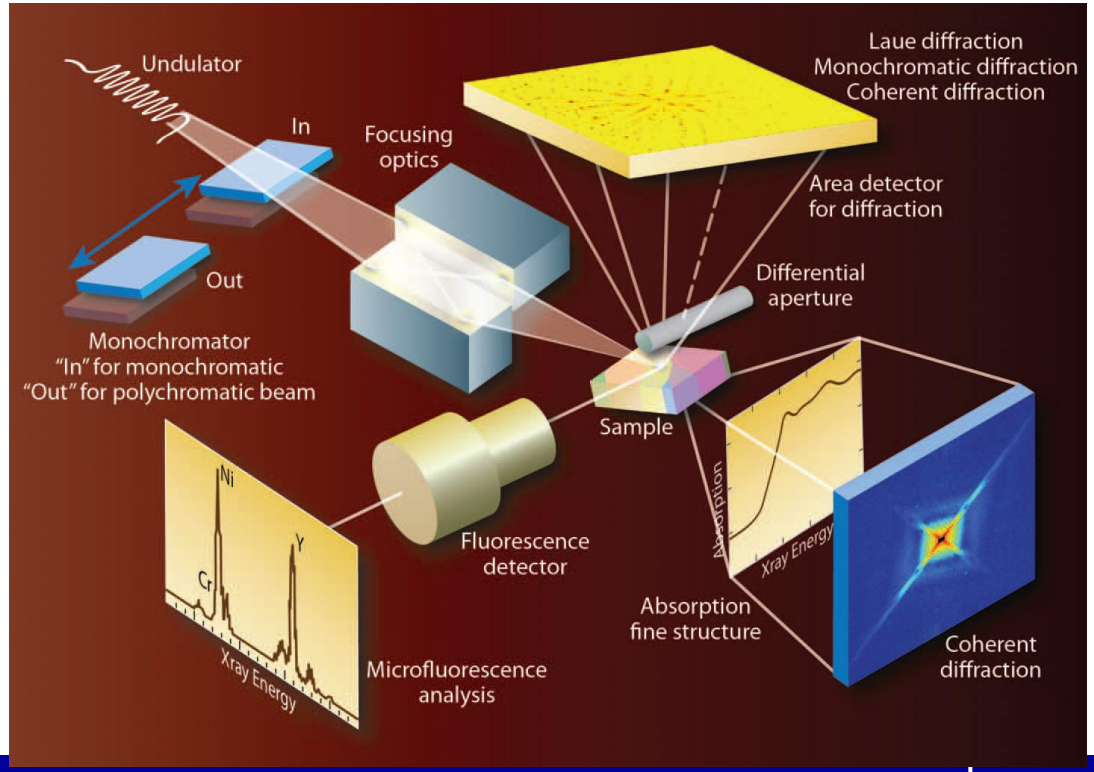
Data volume of HEPS Phase I Beamlines:

Beamlines	Burst output (TB/day)	Average output (TB/day)
Engineering Materials	600.00	200.00
Hard X-ray Multi-analytical Nanoprobe	500.00	200.00
Structural Dynamics	8.00	3.00
Hard X-ray Coherent Scattering	10.00	3.00
Hard X-ray High Energy Resolution Spec.	10.00	1.00
High Pressure	2.00	1.00
Hard X-Ray Imaging	1000.00	250.00
X-ray Absorption Spectroscopy	80.00	10.00
Low-Dimension Structure Probe	20.00	5.00
Biological Macromolecule Microfocus	35.00	10.00
pink SAXS	400.00	50.00
High Res. Nanoscale Elec. Struc. Spec.	1.00	0.20
Tender X-ray beamline	10.00	1.00
Transmission X-ray Microscope	25.00	11.20
Test beamline	1000.00	60.00
Total average:		**805**

Data Challenges @HEPS

- Multi-disciplinary, e.g. spectroscopy, imaging, diffraction & scattering ……
- New and more complex experiments, new algorithm and tools
  - 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
- New users from a wide range of domains and backgrounds





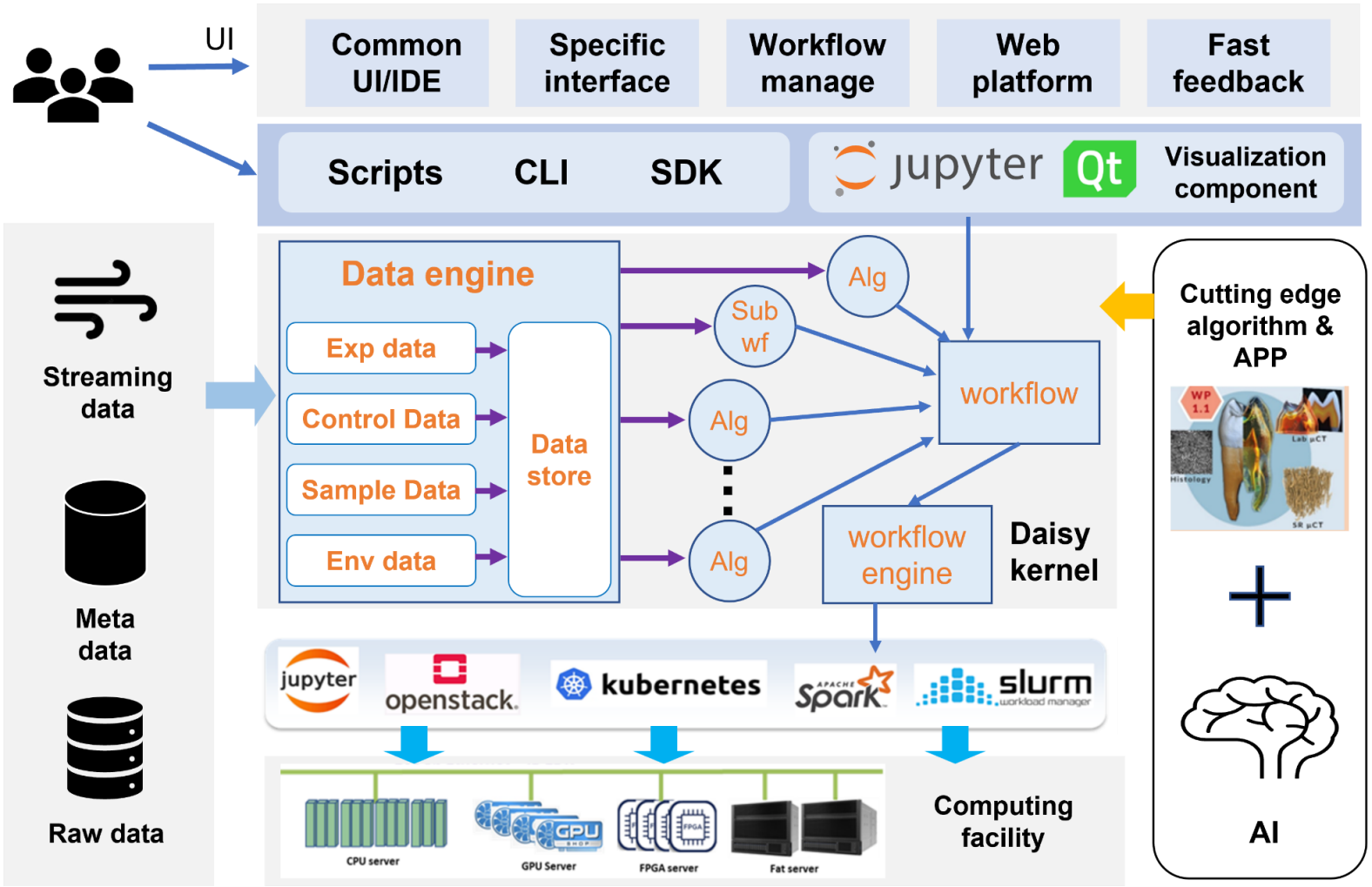
Data Challenges @HEPS

- Analysis and management of large datasets at advanced SR 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. Introduction
2. Demand and Challenges of scientific data and software system
3. **The architecture and design of the framework**
4. The progress of the framework
5. Summary

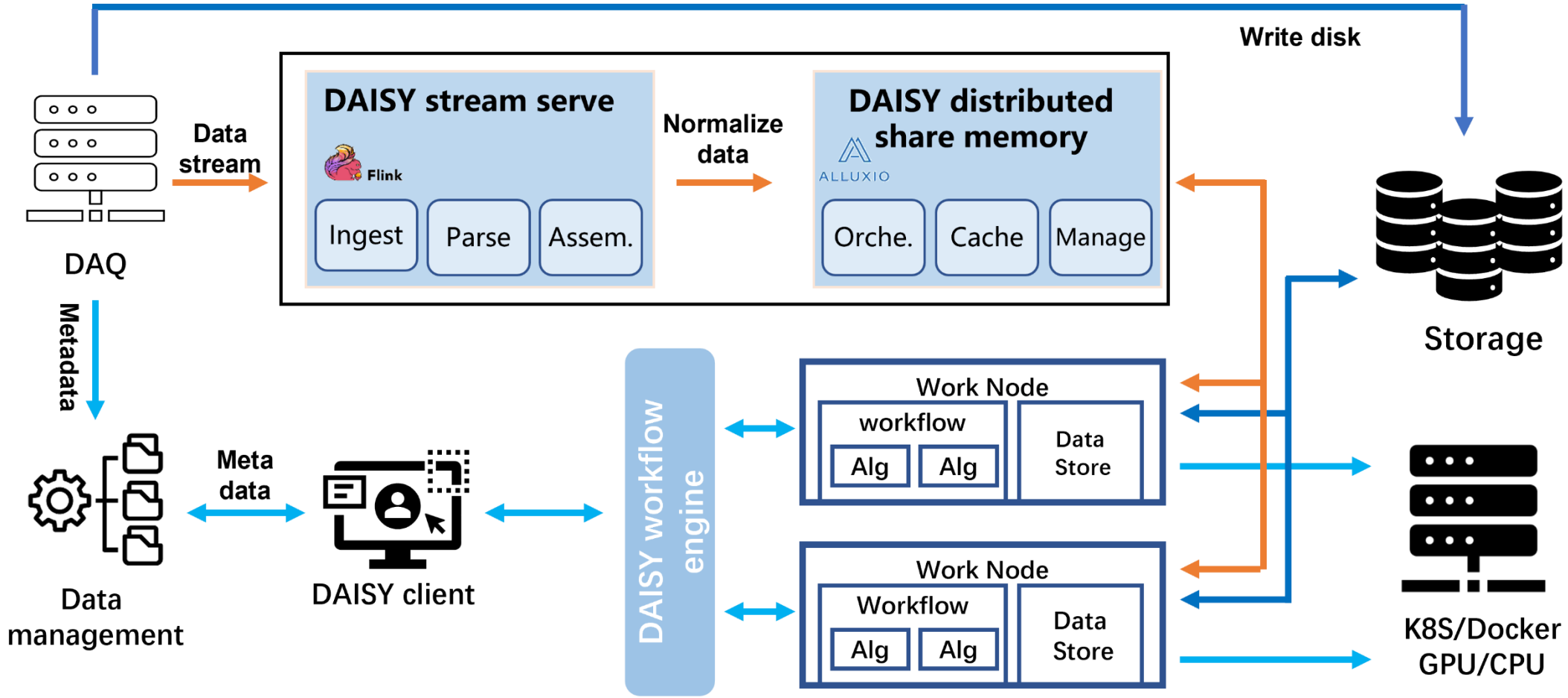
Data analysis software framework—Daisy



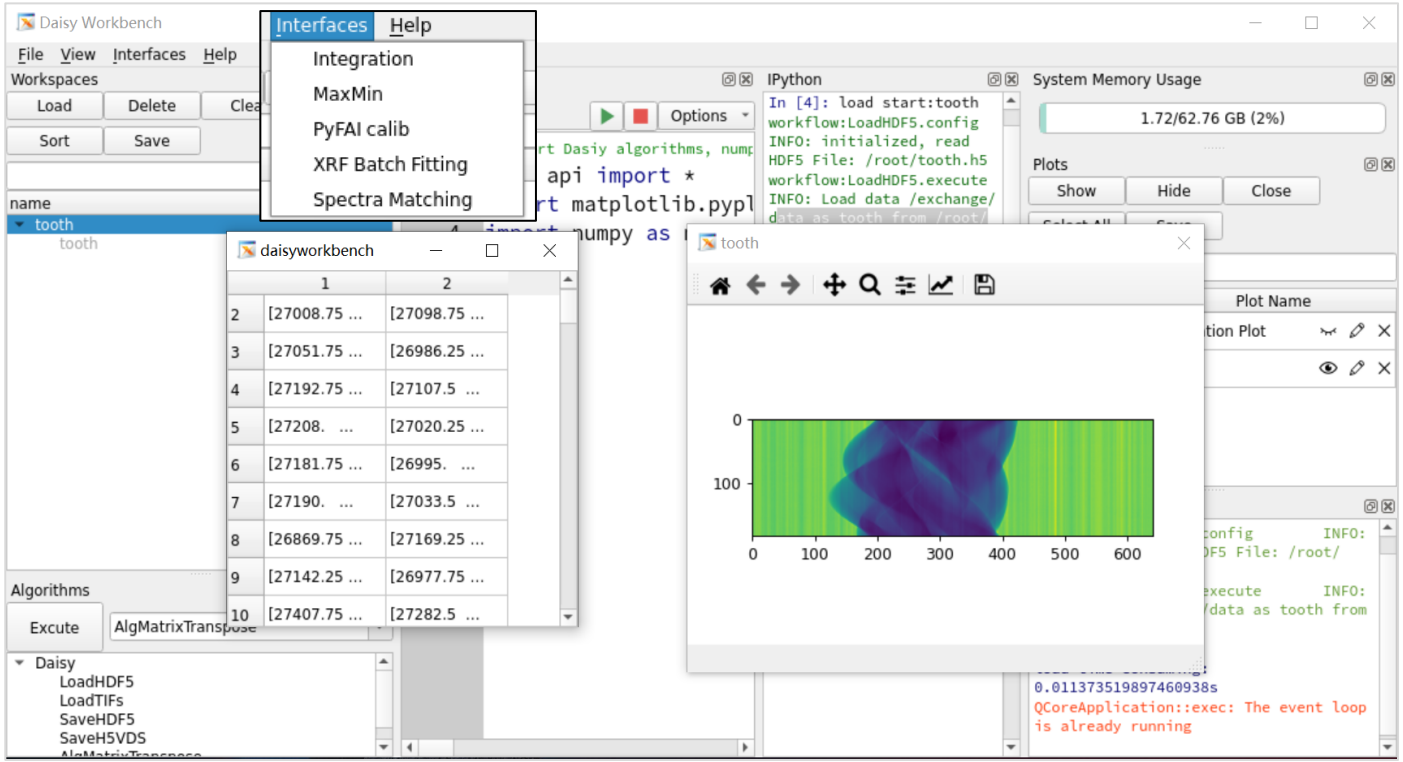
- Kernel of the framework
- Derivative modules to meet the requirements of advanced SR sources
  - Data object management module for high-throughput data I/O, multimodal data exchange, and multi-source data access
  - Scalable cluster computing power support for data processing with different scales, 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

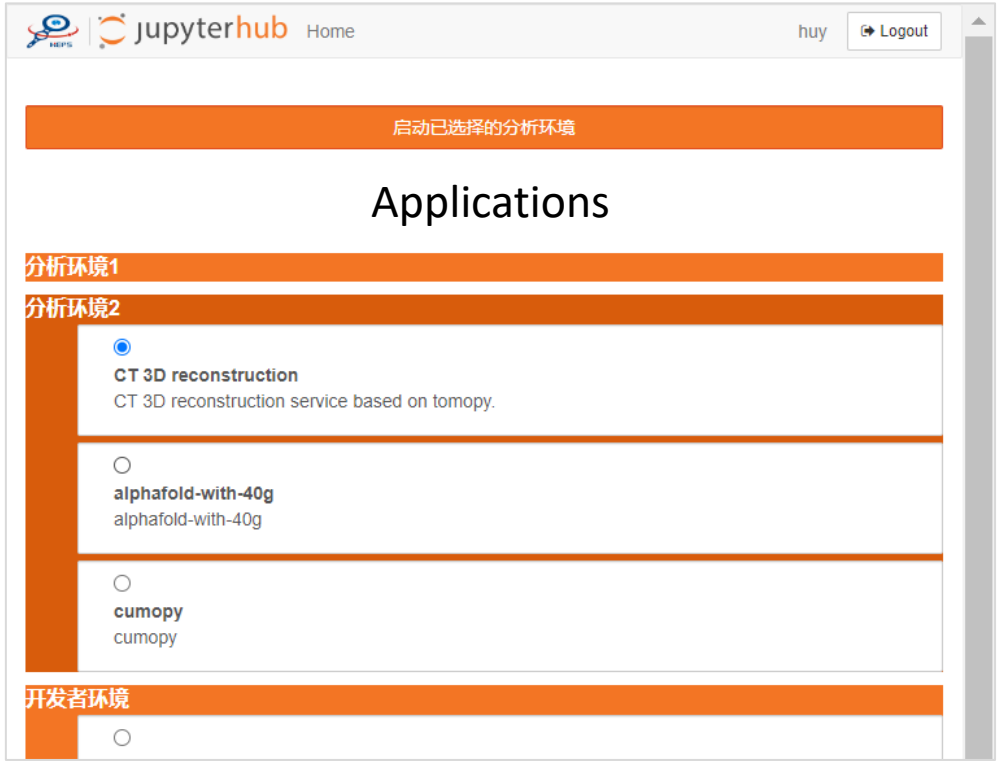
Daisy data processing flow

- Supports stream processing and batch processing
- Supports interactive processing via GUI and automated processing



Daisy GUI





Workbench client

- General-purpose GUI based on PyQt5
- Include data object list, algorithm list, data viewer tools, and IDE for developers
- Interfaces of customized GUIs for scientific applications

Web data analysis platform

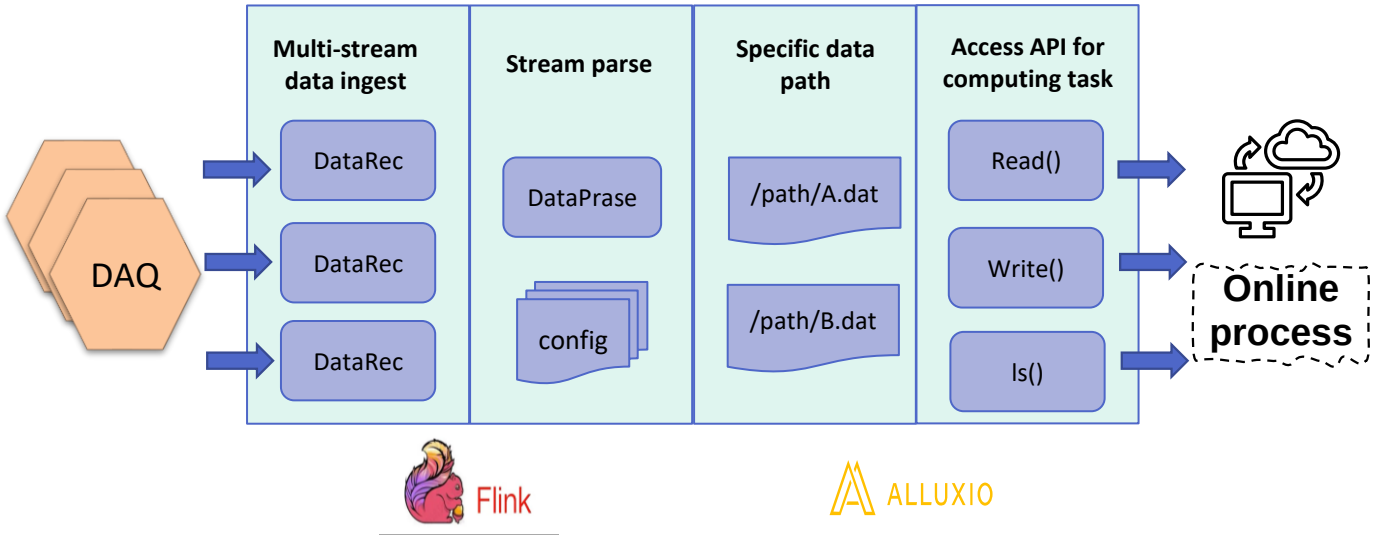
- Based on JupyterLab
- Advantageous for remote data access and analysis
- Terminal and dedicated scientific App

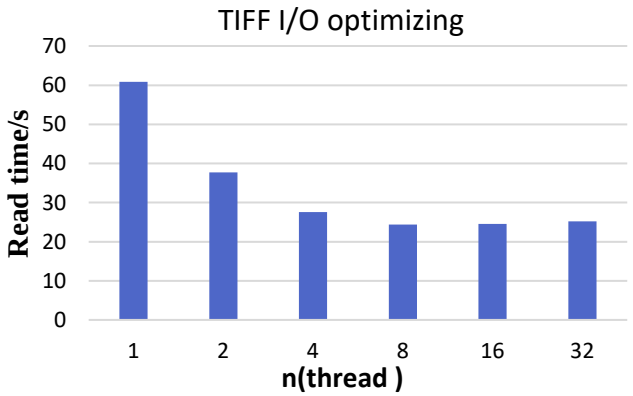
Outline

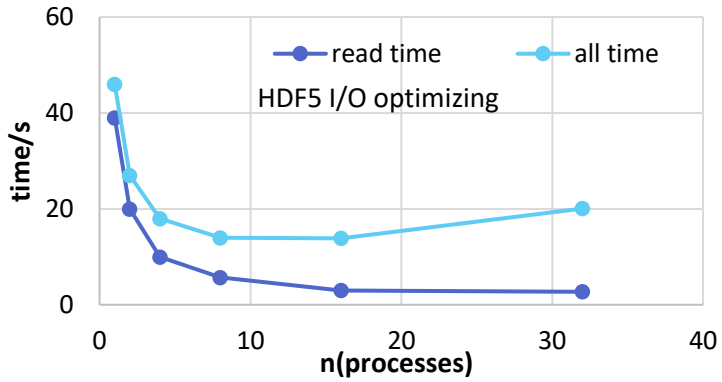
1. Introduction
2. Demand and Challenges of scientific data and software system
3. The architecture and design of the framework
4. **The progress of the framework**
5. Summary

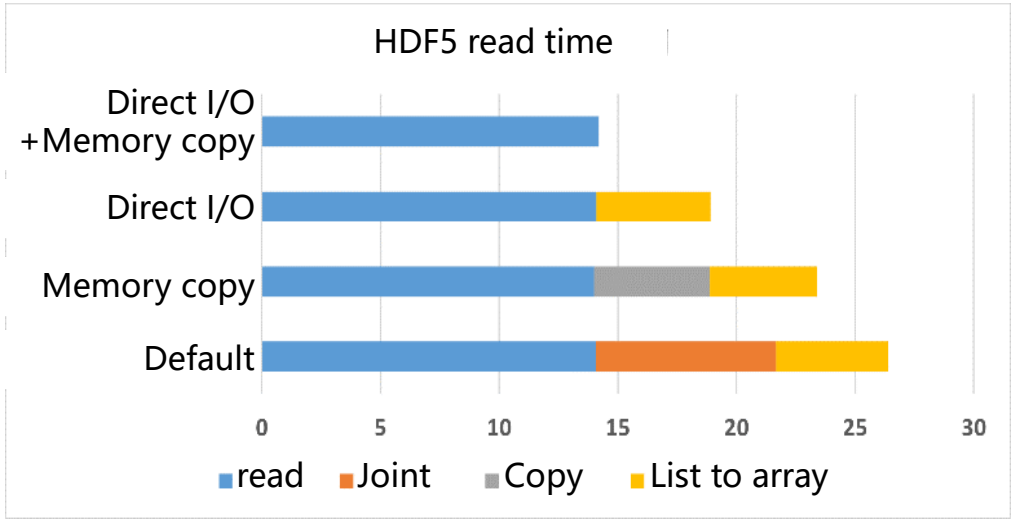
Daisy I/O module

- Designed a unified I/O interface to shield the difference of underlying architecture and data structure
- Data I/O optimization
  - HDF5 parallel I/O based on multi-process. Memory copy, asynchronous I/O, direct I/O also employed
  - TIFF parallel I/O based on multi-thread
- R&D on stream data processing is under way



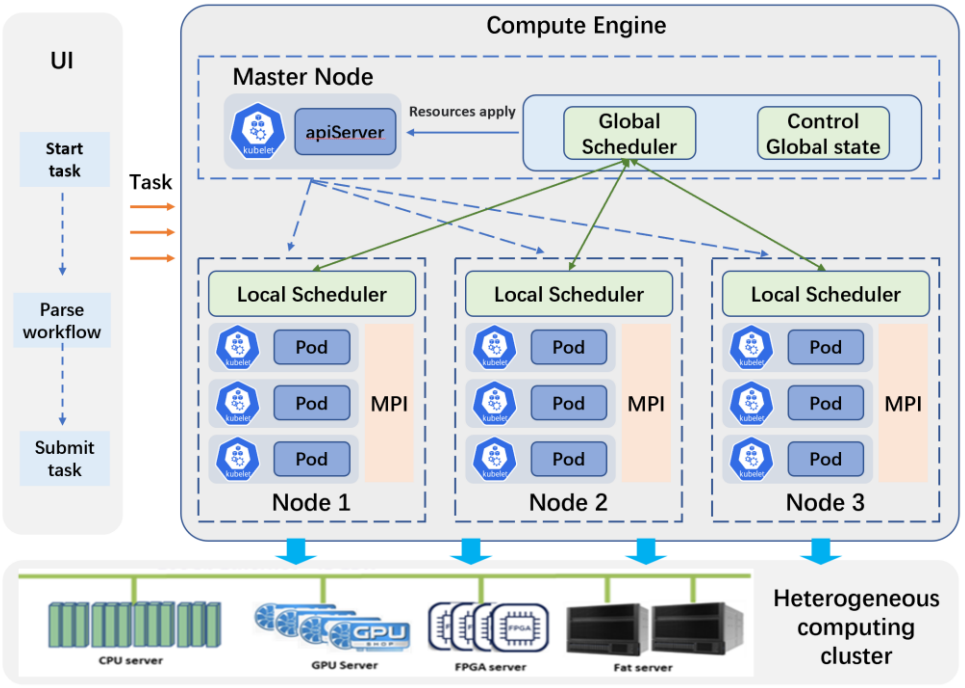




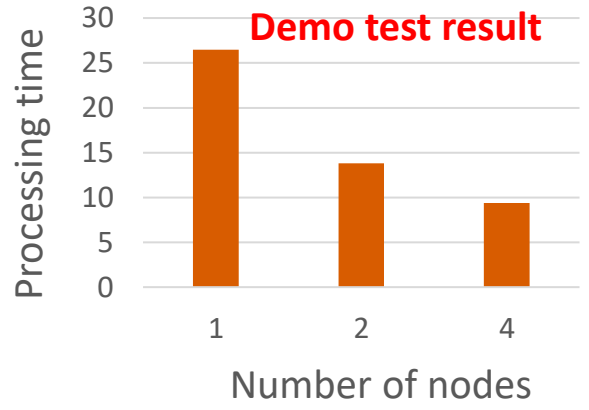


Daisy distributed computing engine

- A single dataset of HEPS imaging experiment will reach the TB scale
- Scientists expect data processing time at the scale of DAQ time
- A distributed data processing system is developing
- Support heterogeneous distributed computing power
- Provide a unified flexible programming interface API for computing models, to reduce the complexity of parallel programming



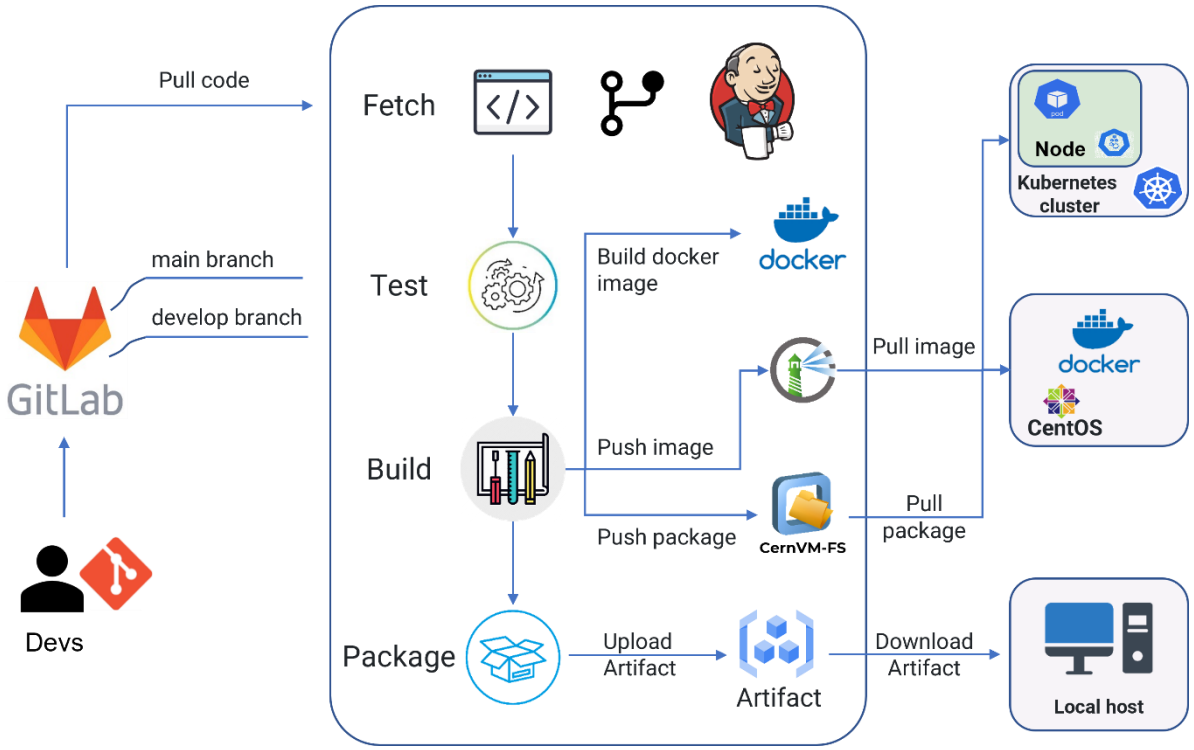
Mode	Detector pixel	Frame rate	Projections number	Data rate	Dataset (TB)	DAQ time	Daily data (TB/d)	Annual data (PB/y)
Powder CT	6k×6k	19fps@16bit	6k	1.08 GB/s	0.432	6.3 min.	78	9.4
High voxel CT	28kx10k	2pfs@16bit	28k	1.1 GB/s	15.68	240 min.	87	10.4
Fast CT	5k×4k	595fps@8bit	5k	1.7 GB/s	0.1	1 min.	98	5.9

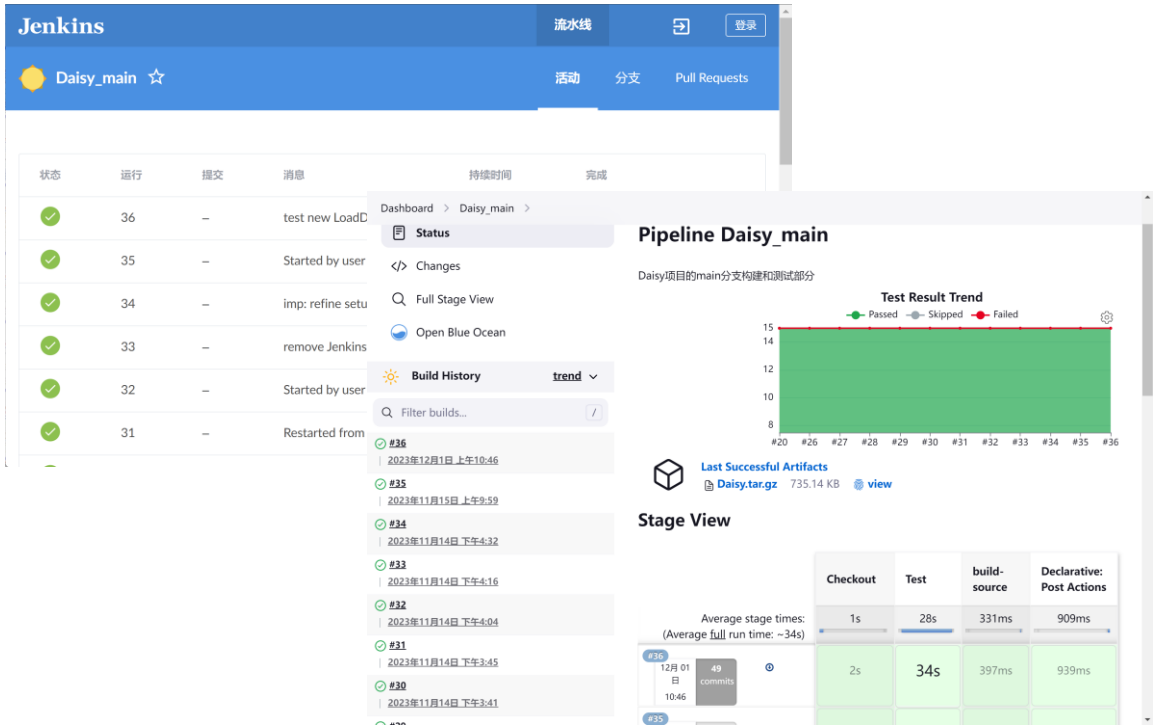


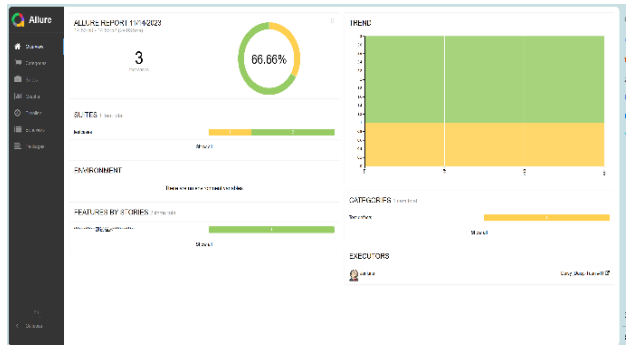
Developer user support

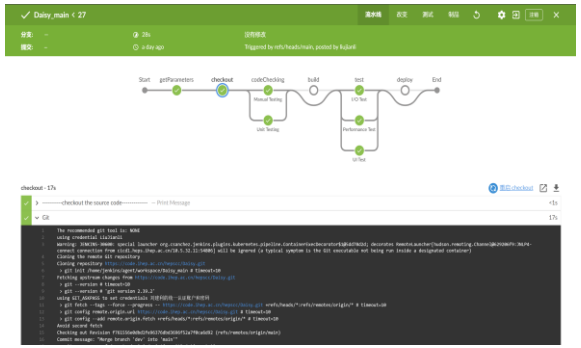
Continuous integration/delivery/deployment system for software development

- CI/CD platform, implement full-process automation of code integration -> testing -> building -> deployment for multi-developer
- Simplify and accelerate the software development lifecycle









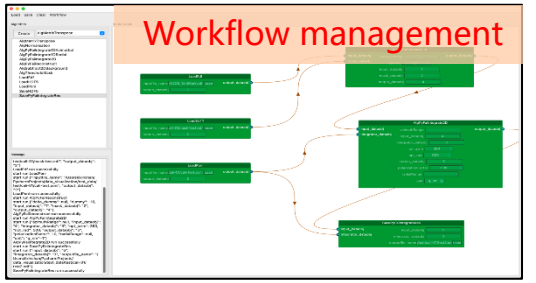
Daisy applications for synchrotron radiation

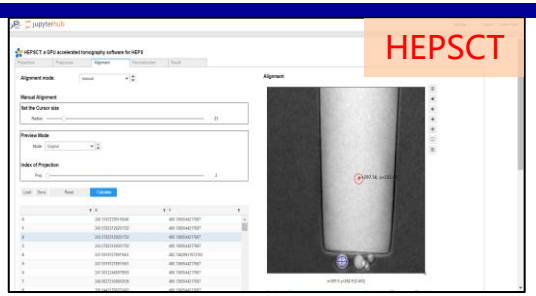
Serving multiple disciplines: imaging, diffraction/scattering, spectroscopy ……

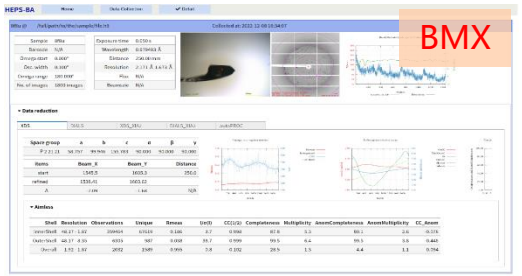
- Daisy-BMX for Biological Macromolecule Crystallography
- XRF-mapping for fluorescence spectrum batch processing
- Hepsptycho for ptychography phase retrieval
- Daisy-PDF for pair distribution function
- XASMatch and PCA&LCF for X-ray absorption spectrometry component analysis
- HEPSCT for X-ray tomography
- Daisy workflow management sys.

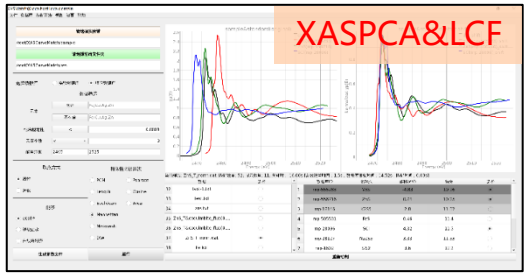
More under development:

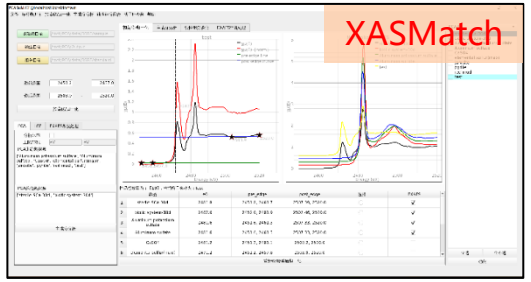
- SAXS
- Holotomography
- XPCS
- Bragg CDI
- ……

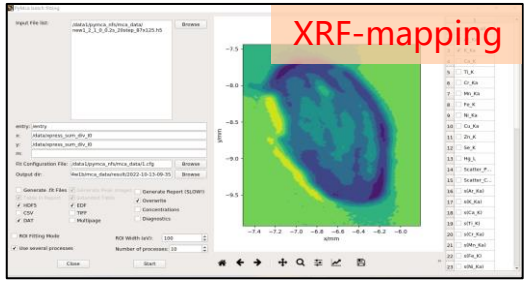


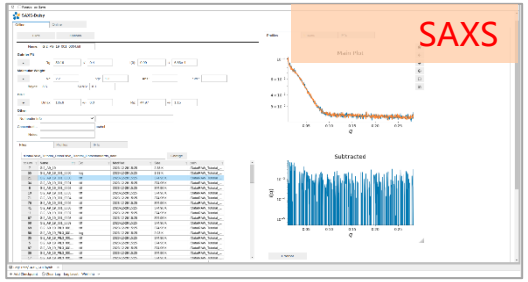












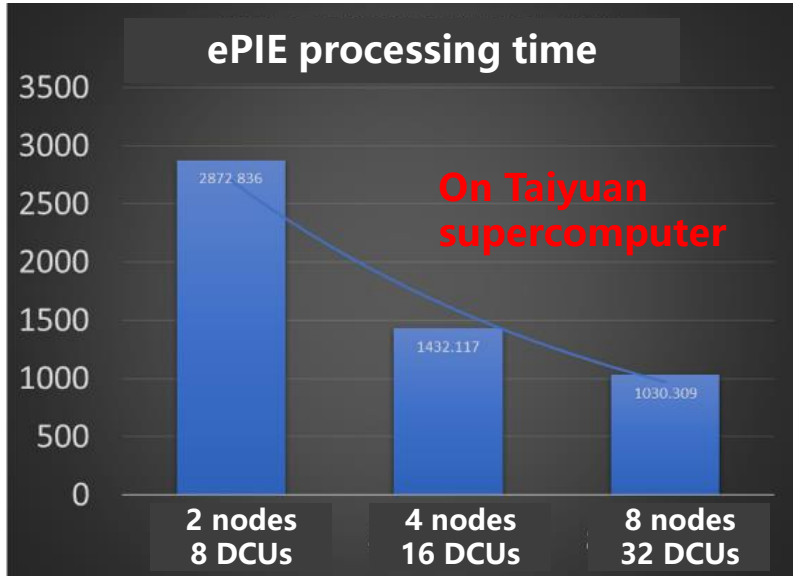
Heterogeneous computing support

Migration scientific software on ROCm platform (AMD GPUs)

- **X-ray tomograpgy**: ASTRA, UFO and Tomocupy migrated on Sugon DCU Z100
- **X-ray Ptychography:** ePIE, DM migrated on Sugon DCU Z100，and deployed on Taiyuan supercomputer(Sugon DCU)
- **Deep learning:** W1-Net, migrated on Sugon DCU Z100 and HUAWEI Ascend 910a

Will be integrated into the framework Daisy

Software	UFO	ASTRA	Tomocupy	ePie，DM	W1-Net
Platform	Sugon DCU Z100	Sugon DCU Z100	Sugon DCU Z100	Sugon DCU Z100	Sugon DCU Z100、HUAWEI Ascend 910a
Performance	~ A100	—	—	~25% A100	DCU ~ 66% A100，Ascend: unoptimized



Outline

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

Summary

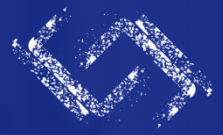
- Scientific data analysis framework：Daisy
- Daisy has been applied to light sources and other facilities
- Optimization on stream data processing and distributed computing is in progress
- Serval scientific applications have been developed, more are on the way
- Hope for more cooperation with other facilities and communities

https://daisy.ihep.ac.cn

Thank you for your attention!

国家高能物理科学数据中心

National HEP Science Data Center



高能所計算中心

IHEP Computing Center

HEPS-CC