The scientific data analysis software framework for HEPS

Yu Hu

Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China

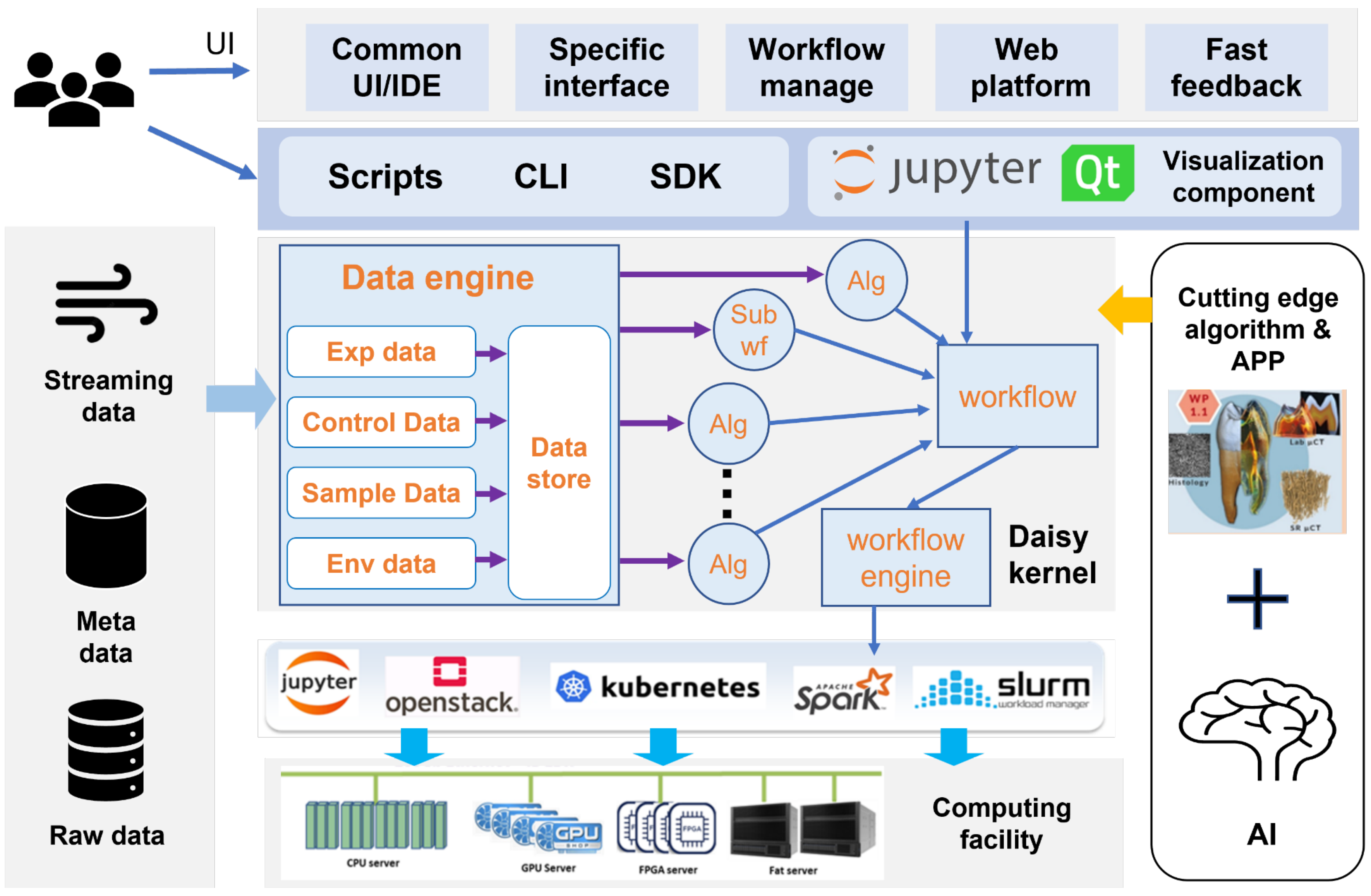
Introduction

- The construction and upgrading of advanced light source have led to the explosion of scientific data
- Scientific discovery under the big data requires advanced mathematics and algorithms
- Efficient storage and processing requires large-scale data management and computing systems, well-designed workflows, and complex workload management software
- A basic scientific data analysis software framework DAISY was designed and developed
- Daisy provides common support for large scale data processing of multi-methodology

High Energy Photon Source (HEPS)

- New 4th generation light source in China with high energy and high brightness
- Located in Huairou Beijing
- The construction was started in the middle of 2019
- The whole project will be finished in middle of 2025
- There will be 14 public beamlines and 1 optics test beamline in Phase I
- The installation of linac, booster and storage ring has been completed

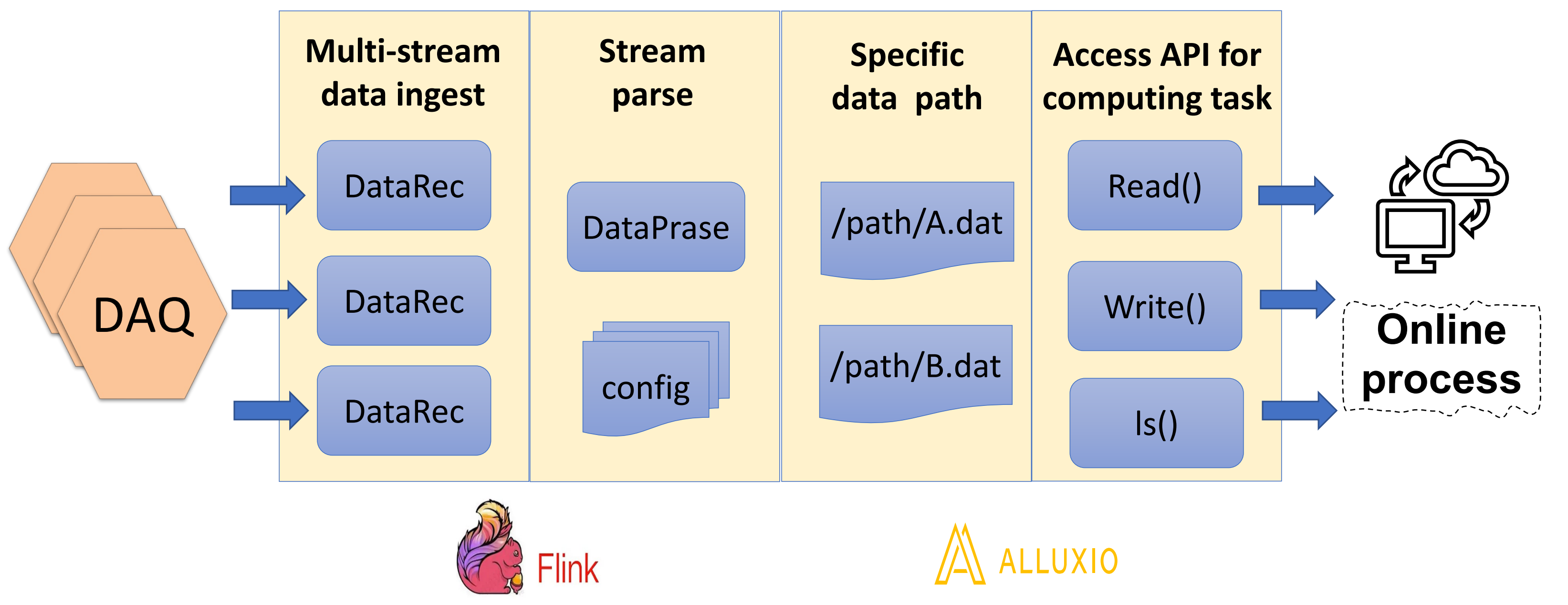
Architecture of DAISY



- 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

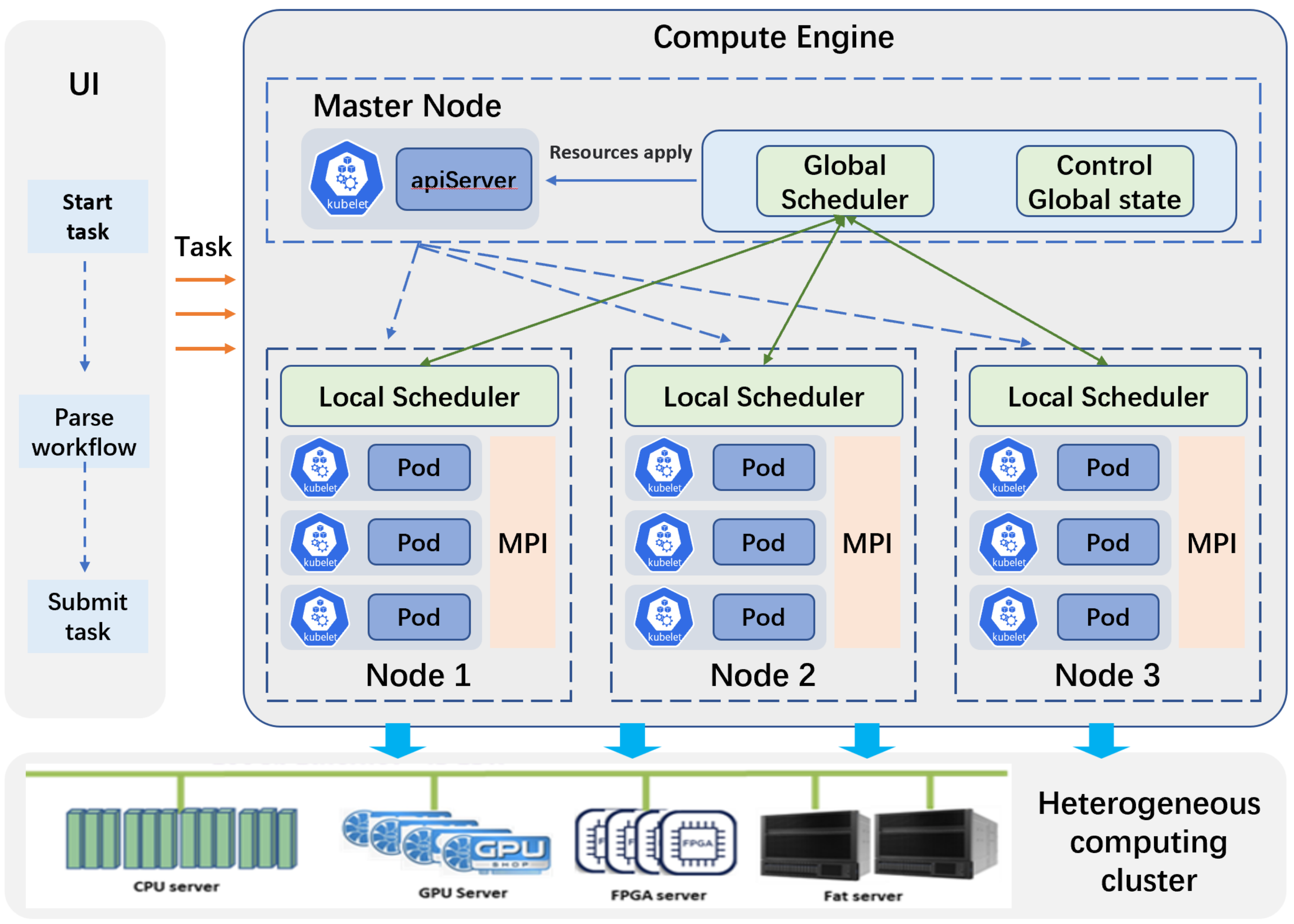
Daisy I/O module

- 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

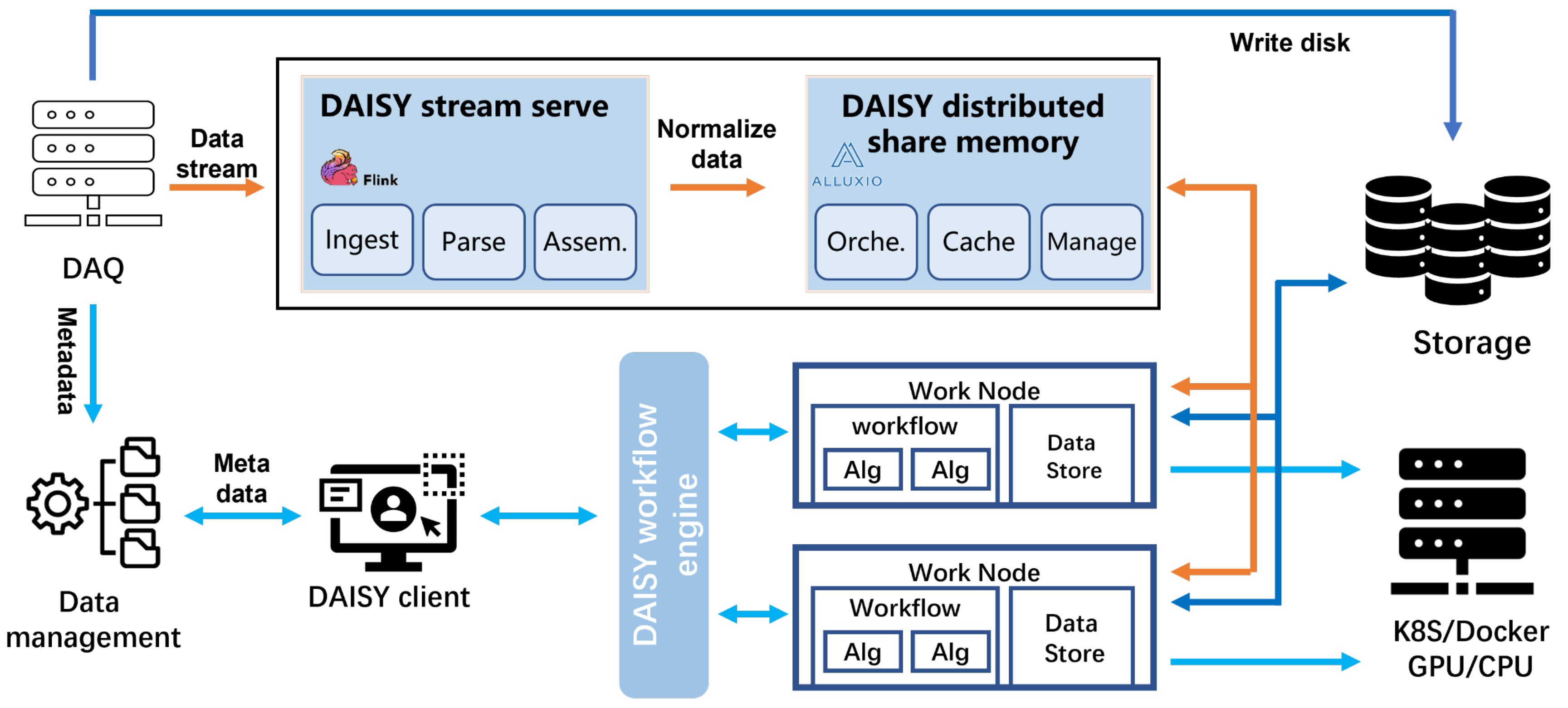


Daisy distributed computing engine

- 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



Daisy data processing flow



1.Hu Yu et al. “Daisy: Data analysis integrated software system for X-ray experiments.” *EPJ Web of Conferences* : 251, 04020 (2021). 2.Zhibin Liu et al. “Evolution of the HEPS Jupyter-based remote data analysis System.” *EPJ Web of Conferences* : 251, 02046 (2021).