中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

国家高能物理科学数据中心
National HEP Data Center

高能所計算中心
IHEP Computing Center

HEPS数据处理软件框架Daisy的进展及规划

胡誉 (huyu@ihep.ac.cn，代表HEPSCC)

中国科学院高能物理研究所

DAISY
Data Analysis and Software Integration System

HEPS

第二十届（2023）第二十届全国科学计算与信息化会议

Outline

1. 背景介绍
2. 科学数据与软件系统的需求与挑战
3. 软件框架的架构与设计
4. 软件框架的开发进展
5. 总结

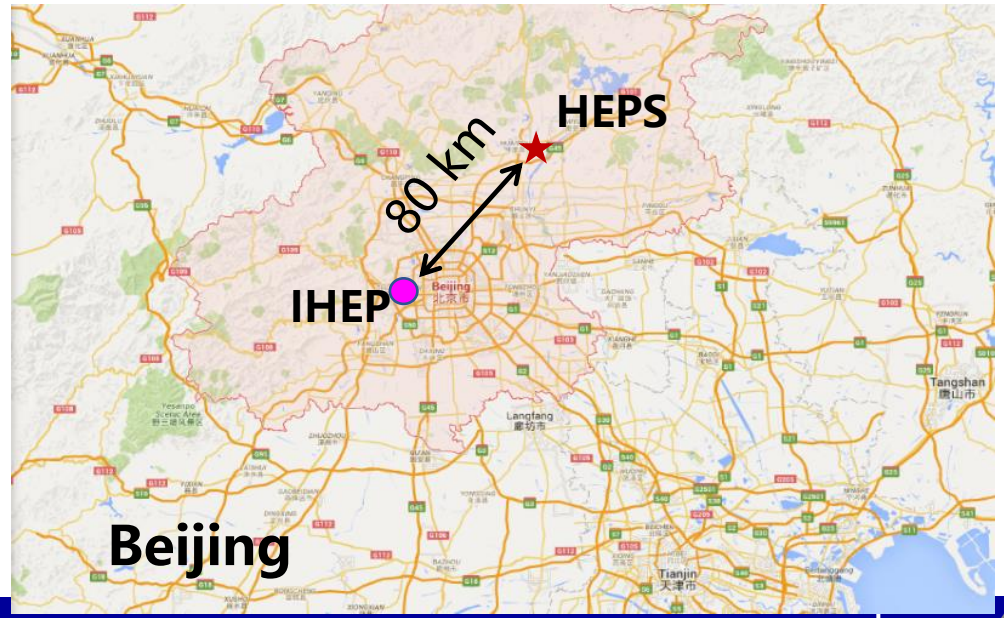
Outline

1. 背景介绍
2. 科学数据与软件系统的需求与挑战
3. 软件框架的架构与设计
4. 软件框架的开发进展
5. 总结

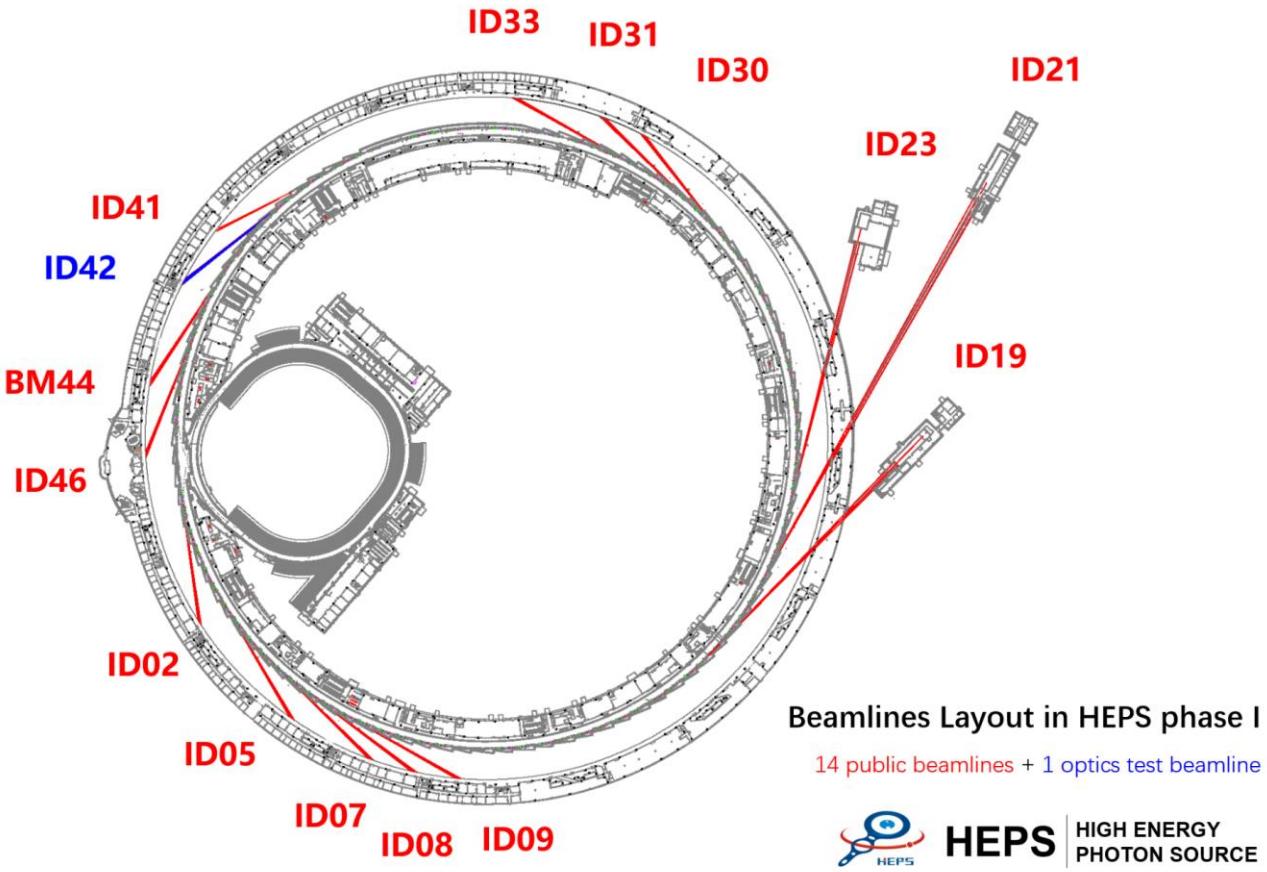
高能同步辐射光源（HEPS）

- 第四代同步辐射光源 — 高能量，高亮度，低发射度
- 将与世界上正在运行的美国 APS、欧洲 ESRF、日本 Spring 8 、德国 PETRA III 一起，构成世界五大高能同步辐射光源
- 位于北京怀柔科学城，核心装置，距离中科院高能所 80 km
- 建设周期6.5年，2018年底开始建设，将于2025年底验收投入运行

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



HEPS 一期光束线站

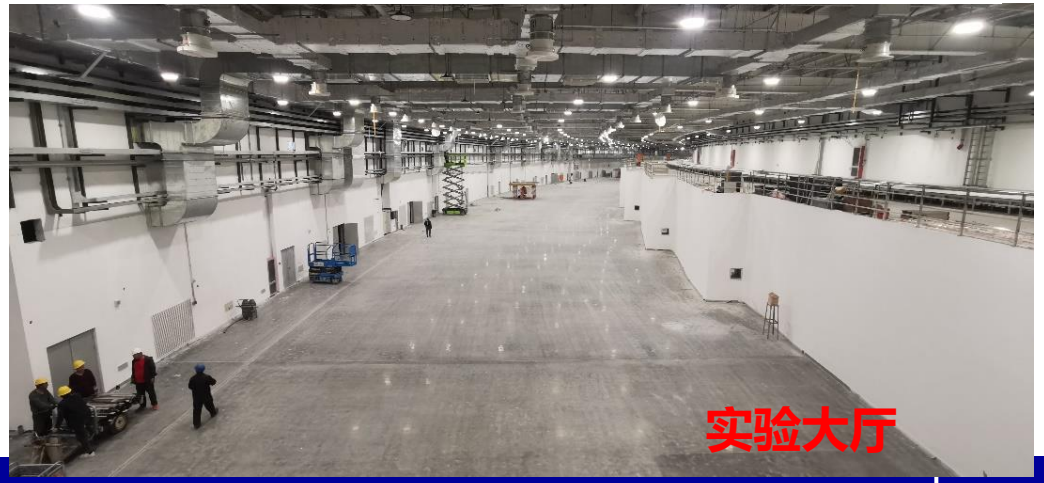


生物大分子微晶衍射线站	-ID02	ID30-	X射线显微成像线站
低维结构探针线站	-ID05	ID31-	高压线站
工程材料线站	-ID07	ID33-	高分辨谱学线站
粉光小角散射线站	-ID08	ID41-	高分辨纳米电子结构线站
硬X射线相干散射线站	-ID09	ID42-	光学测试线
硬X射线纳米探针线站	-ID19	BM44-	通用环境谱学线站
硬X射线成像线站	-ID21	ID46-	X射线吸收谱学线站
结构动力学线站	-ID23		辅助实验室

一期将建设14条公共光束线站 + 1 条光学测试线站

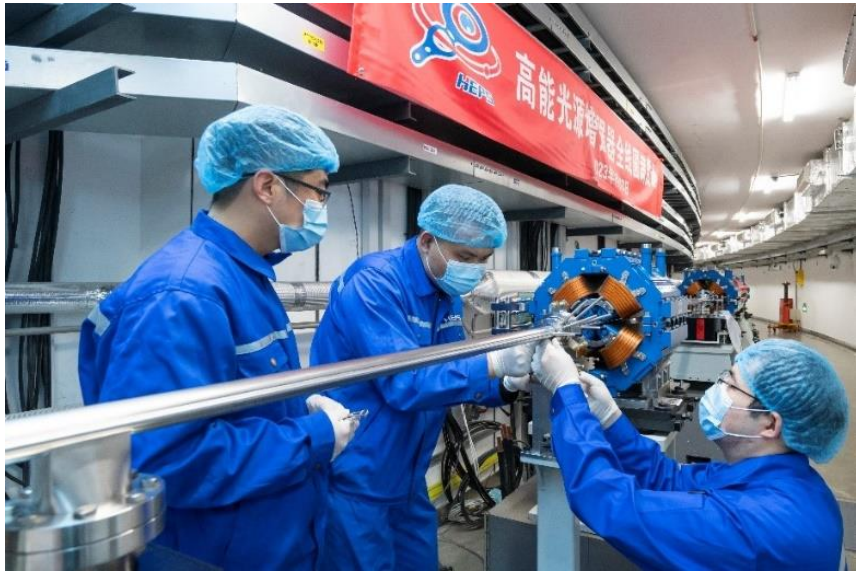
涉及成像、衍射/散射、谱学等学科

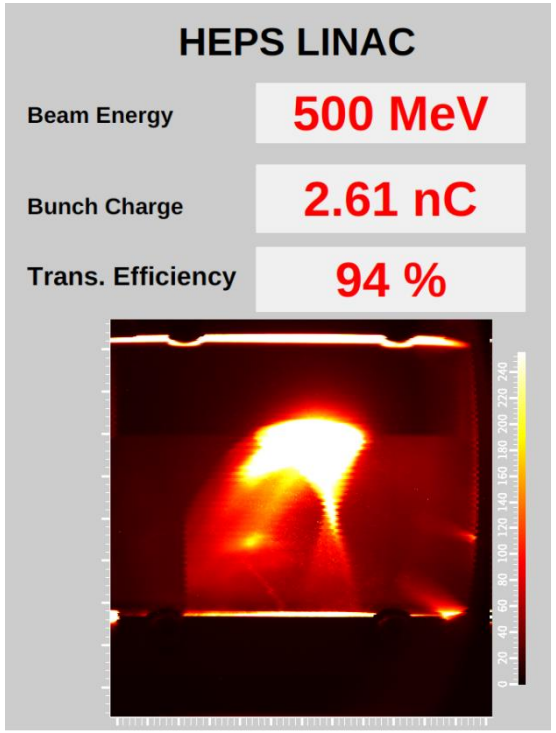
HEPS 总共可以容纳超过 90 条线站



HEPS 进展

- 目前已经完成全部土建结构施工，进入设备安装阶段
- 2023.01, HEPS 增强器全线贯通
- 2023.02, 启动储存环隧道设备安装
- 2023.03, HEPS直线加速器满能量出束，成功将第一束电子束加速到500 MeV





Outline

1. 背景介绍
2. **科学数据与软件系统的需求与挑战**
3. 软件框架的架构与设计
4. 软件框架的开发进展
5. 总结

Data Challenges @HEPS

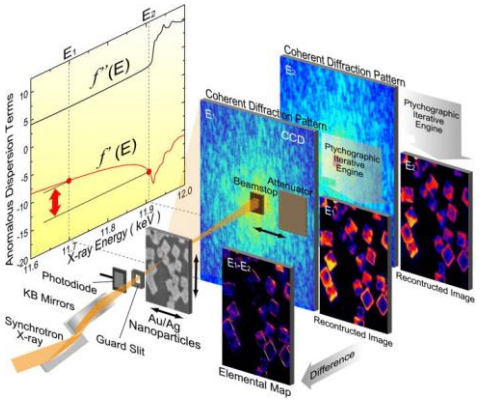
- 亮度相对三代光源提升了 2-3 个量级
  - 在更短的时间内产生更加海量的具有更多细节信息的原始数据
- X 射线探测器能力不断提高:
  - 更宽的动态范围，更快的读出速率，更大的像素阵列
  - 更大的帧数，更高的帧率=更多的原始数据
- HEPS 一期(15个线站)数据产生率接近PB/天，每年将产生超过200PB的数据
- HEPS 总共能容纳超过90条线站，更多的数据

Data volume of HEPS Beamlines:

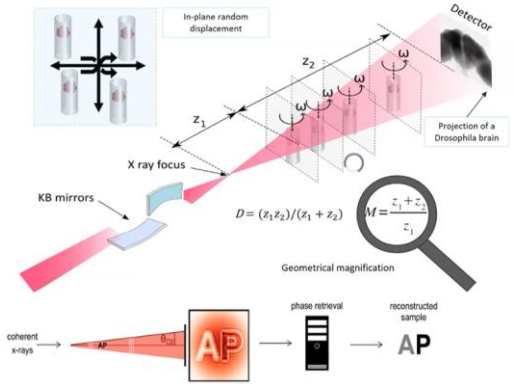
Beamlines	Burst output (TB/day)	Average output (TB/day)
B1 Engineering Materials Beamline	600.00	200.00
B2 Hard X-ray Multi-analytical Nanoprobe (HXMAN) Beamline	500.00	200.00
B3 Structural Dynamics Beamline	8.00	3.00
B4 Hard X-ray Coherent Scattering Beamline	10.00	3.00
B5 Hard X-ray High Energy Resolution 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

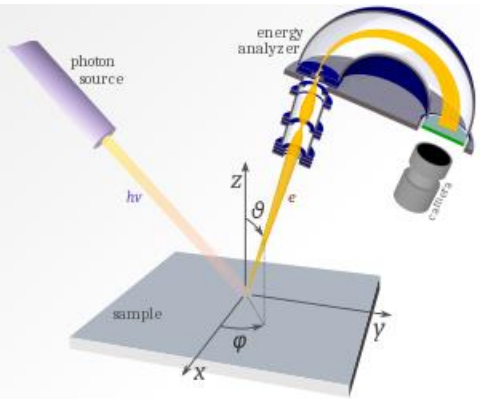
- X射线光源技术的发展，不断催生出更加复杂的新方法、新技术以及新研究领域，需要新的学科软件及算法
- 多模态实验需要结合多个样本、技术和设备的数据
- 原位和动态加载实验需要实时反馈和自主控制
- 不同的光束线站以及科学目标，其实验数据通量和容量存在巨大差异



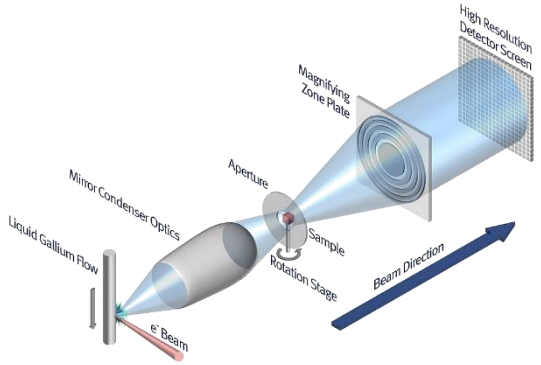
Fluorescence mapping



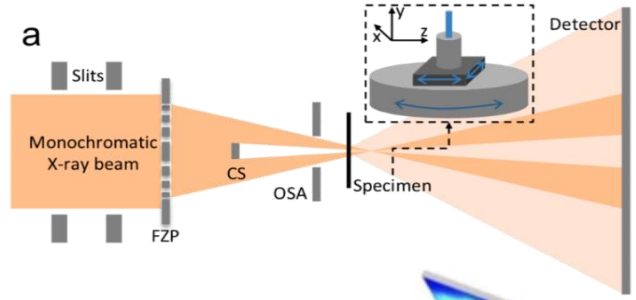
Nanoholotomography



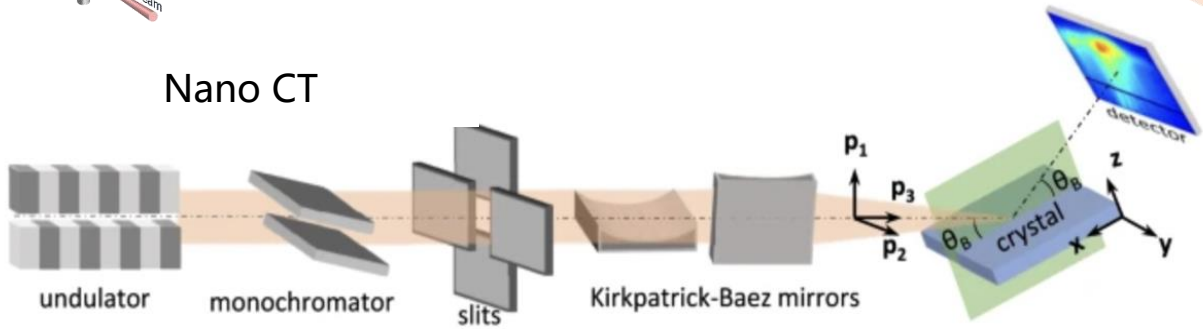
ARPES



Nano CT



Ptychography CT



Bragg ptychography

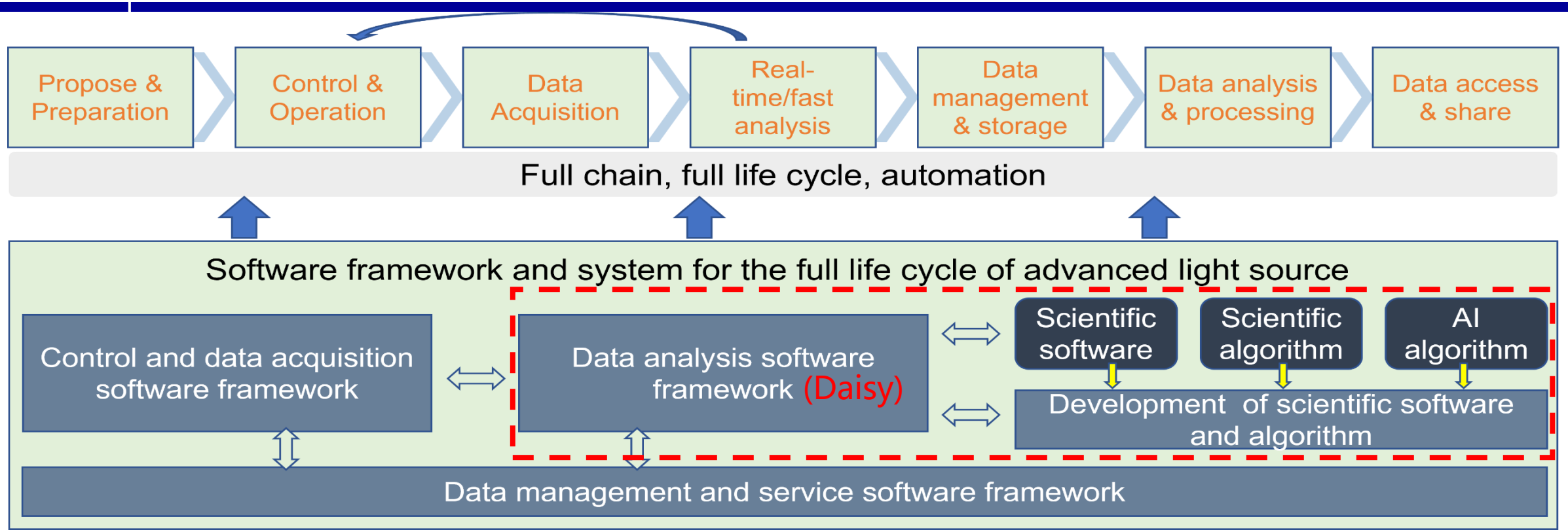
Data Challenges @HEPS

- 先进光源上大规模科学数据的分析和管理变得越来越具有挑战性
- 需要开发和集成先进的分析和管理工具
  - 提供海量科学数据的存储、组织和管理
  - 为方法学软件和算法的多样化发展提供通用的底层软件框架支持
  - 在实验过程中，进行实时数据分析和快速反馈，提供决策指导和修正实验过程，并优化数据采集
  - 在实验结束后，处理海量多模态数据，帮助用户快速完成实验数据分析、获取科研成果，加速科学发现
  - 提供可伸缩的分布式异构算力支持，满足不同科学目标不同规模的计算分析需求

Outline

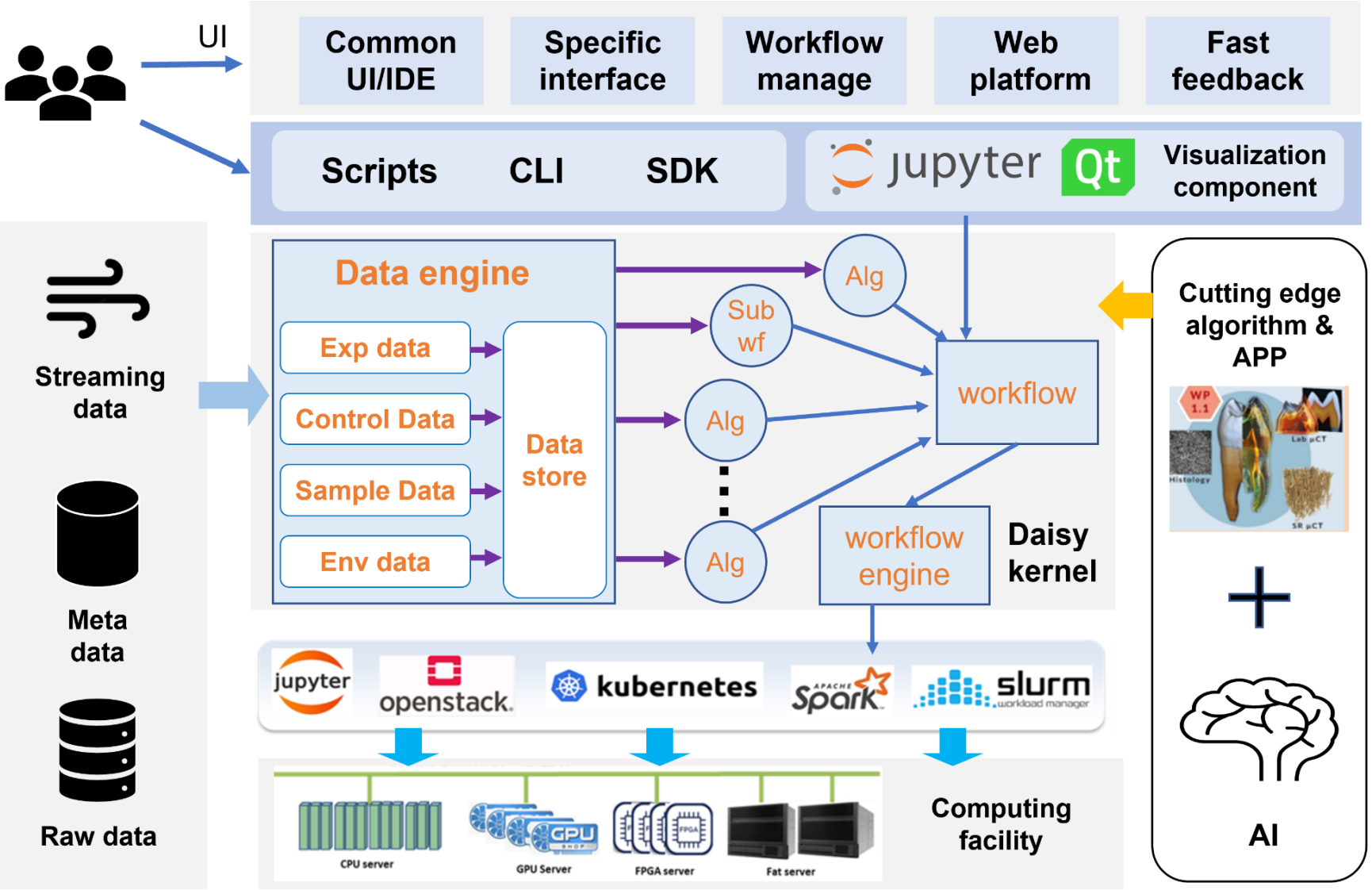
1. 背景介绍
2. 科学数据与软件系统的需求与挑战
3. **软件框架的架构与设计**
4. 软件框架的开发进展
5. 总结

光源全生命周期的软件系统



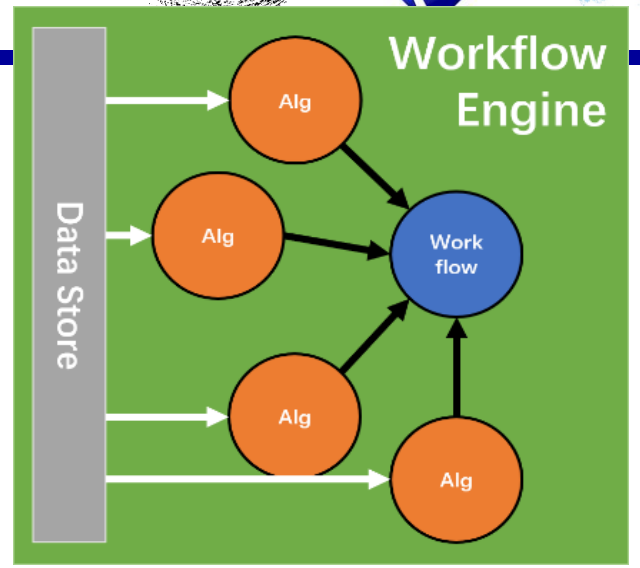
- 设计了先进光源全链条、全生命周期的软件框架和系统
- 促进先进光源软件系统全流程的智能化、自动化，实现科学数据全生命周期的跟踪和管理
- 支持发展光束线站数据分析新方法、新软件，以及已有方法学和软件的标准化框架集成
- 建立有影响力的开源软件社区，吸引并支持潜在的社区贡献者，建立大科学装置全生命周期、多设施协作、多学科融合的软件生态环境

数据处理软件框架总体架构



- 数据处理软件框架核心
- 满足新一代光源数据处理需求的衍生技术模块
  - 应对高通量数据I/O、多模态数据解析、多源数据接入的数据对象管理
  - 应对不同规模、不同通量、低延迟数据处理需求的弹性异构计算集群算力支持
  - 服务于学科方法学软件集成和发展的用户软件接口和软件开发环境支持
- 基于软件框架的学科专用应用软件以及针对灵活数据处理需求的通用工作流编排系统

数据处理软件框架核心

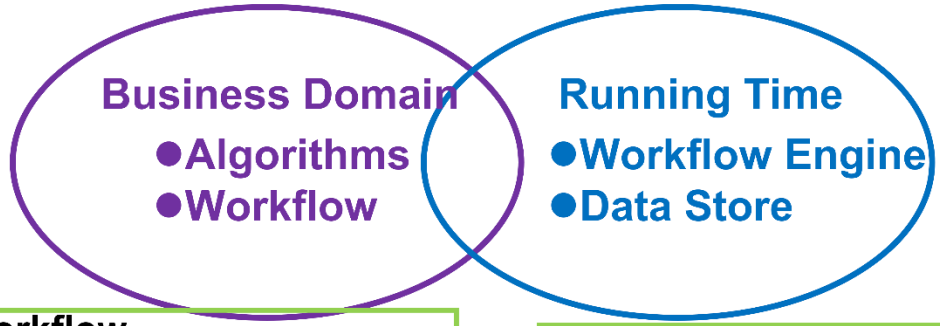


- 遵循**领域驱动设计**的概念，从领域知识中提取出与实现技术无关的**领域模型**，保证业务抽象性和独立性，并建立领域模型之间的关系，形成**领域架构**。
- 一个**领域模型**对应开发一个或多个**算法或子工作流**，对多个算法或子工作流的**有序调用**对应**领域框架的实现**。

- **算法**：框架中的最小单元，定义领域模型，具体的数据处理模块，支持第三方程序库集成。同一个算法可以有多个实体实现，相互替代。
- **工作流**：定义领域架构，通过调用一系列算法完成特定的处理分析任务，支持嵌套。
- **工作流引擎**：根据计算环境提供的计算资源，在运行时判断算法的具体执行实体，同时管理算法模块的并行分布式执行。解除了业务流程和计算环境之间的耦合。
- **数据仓库**：管理算法之间数据对象的创建和传递，使业务代码不用考虑数据对象的管理问题。

Algorithms
- Input Data Processing
- Output Data Defined

Workflow Engine
- Handle Data Store
- Running Time Management



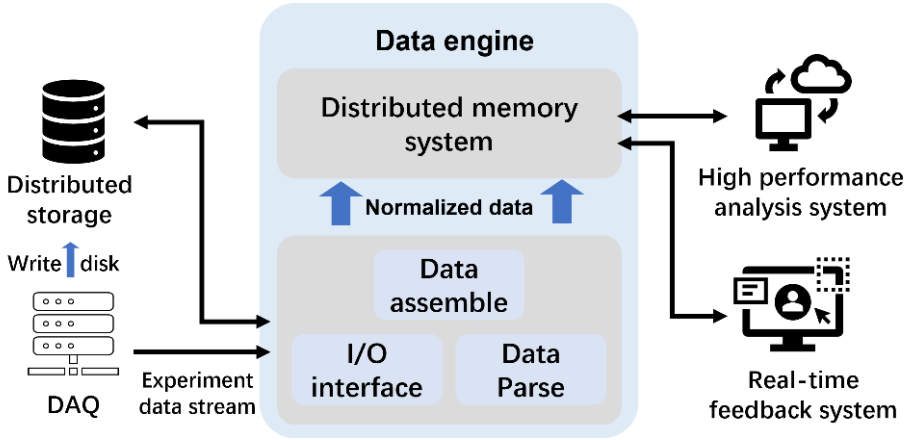
Workflow
- A sequence of Algorithm
- Workflow is also an algorithm

Data Store
- Data Object Management

高性能模块

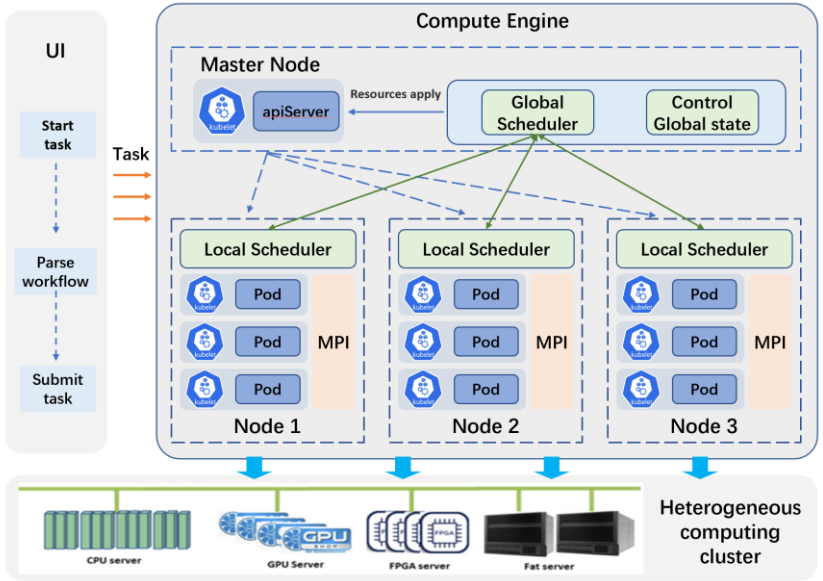
数据对象管理

- 提供统一的读写接口，屏蔽底层架构，光源数据格式，数据来源的差异性
- 除磁盘文件外，也将支持流数据的读写，实现实时、高通量的在线数据处理
- 采用异步并行、分布式内存、自适应存储参数和数据压缩等方法优化数据I/O



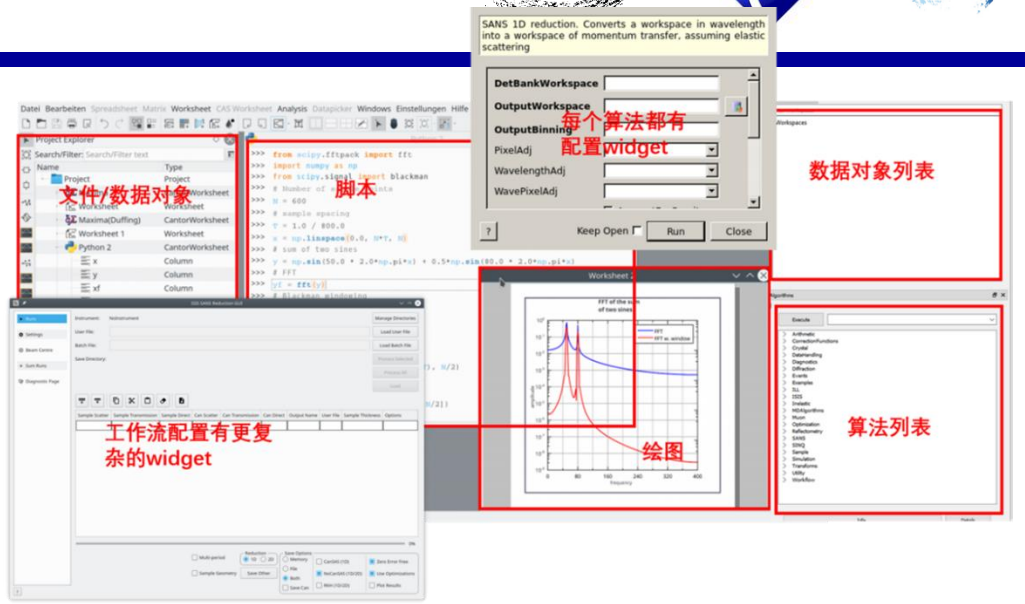
异构分布式算力支持

- 充分利用先进计算基础设施
- 针对计算热点，形成高性能数值分析计算库
- 提供统一灵活的并行编程接口API，屏蔽底层硬件资源的体系结构差异，降低数据分析软件的并行编程复杂度
- 提供分布式计算任务调度器，保证分布式任务的高可用和高效率



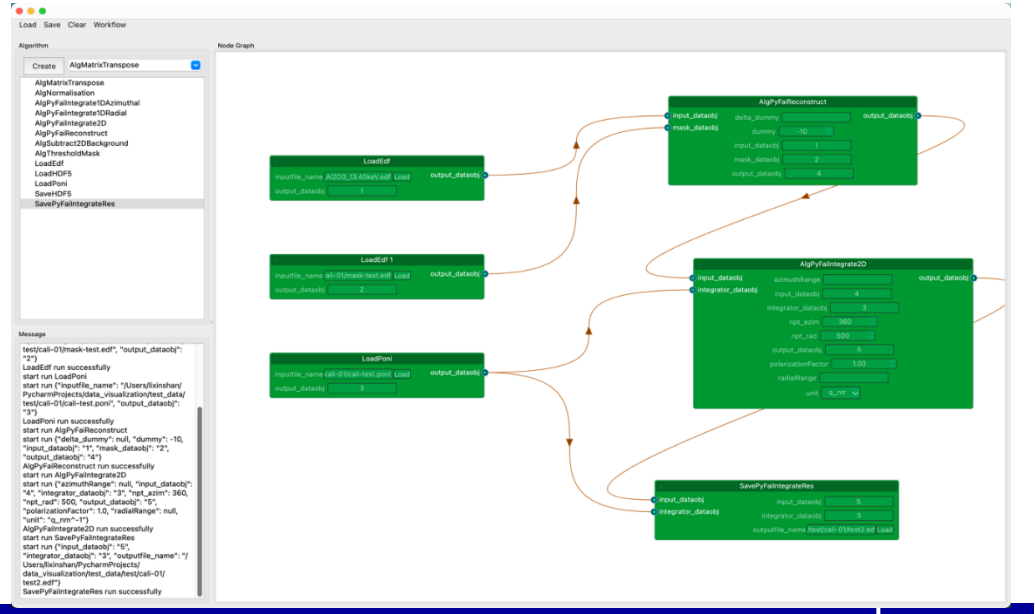
面向学科应用的用户软件接口

- **提供多种形式的用户软件接口，支持不同场景的方法学应用**
  - 数据可视化界面
  - 集成开发环境界面
  - 方法学接口界面
  - Web 数据分析平台
  - 脚本和命令行接口



- **提供多种可复用的图形化界面常见控件，支持各实验站在此基础上进行二次开发**
  - 分析绘图桌面，算法配置控件，工作流配置控件，算法/工作流对象列表，数据对象列表，IDE……

- **用于灵活通用数据处理任务的 workflow 管理系统**
  - 前端提供App和Web端的图形界面，支持交互式的工作流编排、导入、导出和运行监控等
  - 遵循 Common Workflow Language(CWL) 标准，支持CWL的解析和生成



开发用户者支持

版本控制

- Git 控制版本，Gitlab 托管项目代码、连接CI/CD

运行环境

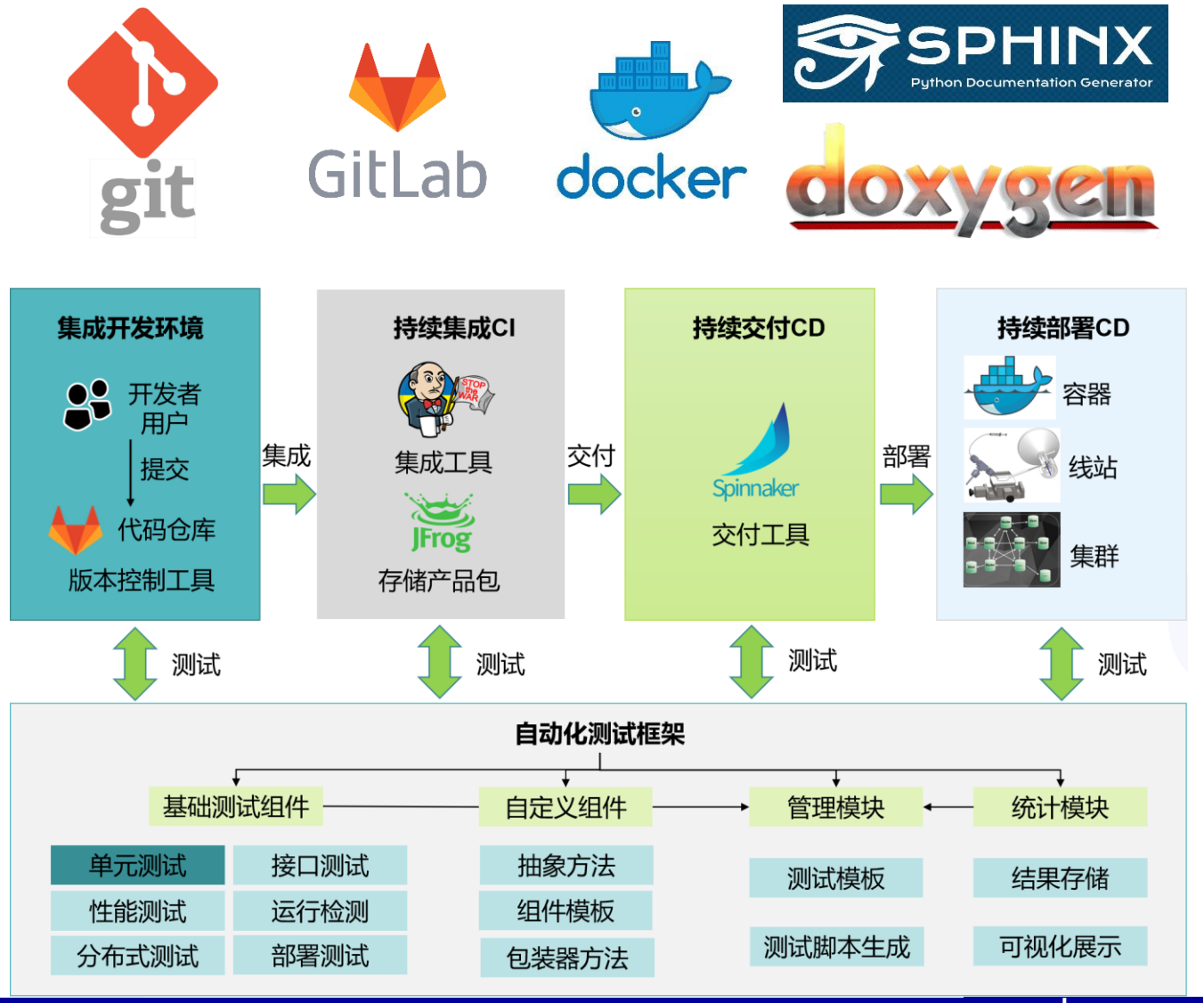
- 容器封装基础运行环境，CVMFS 部署编译好的软件

文档指南

- 用户使用文档，开发者开放指南
- Jupyterbook，Sphinx，readthedocs
- Doxygen 根据注释自动生成代码文档

软件开发的可持续集成、部署(CI/CD)和测试框架

- 支持软件的自动化集成、编译、测试和部署，降低开发门槛
- 实现软件开发过程的标准化、自动化和智能化

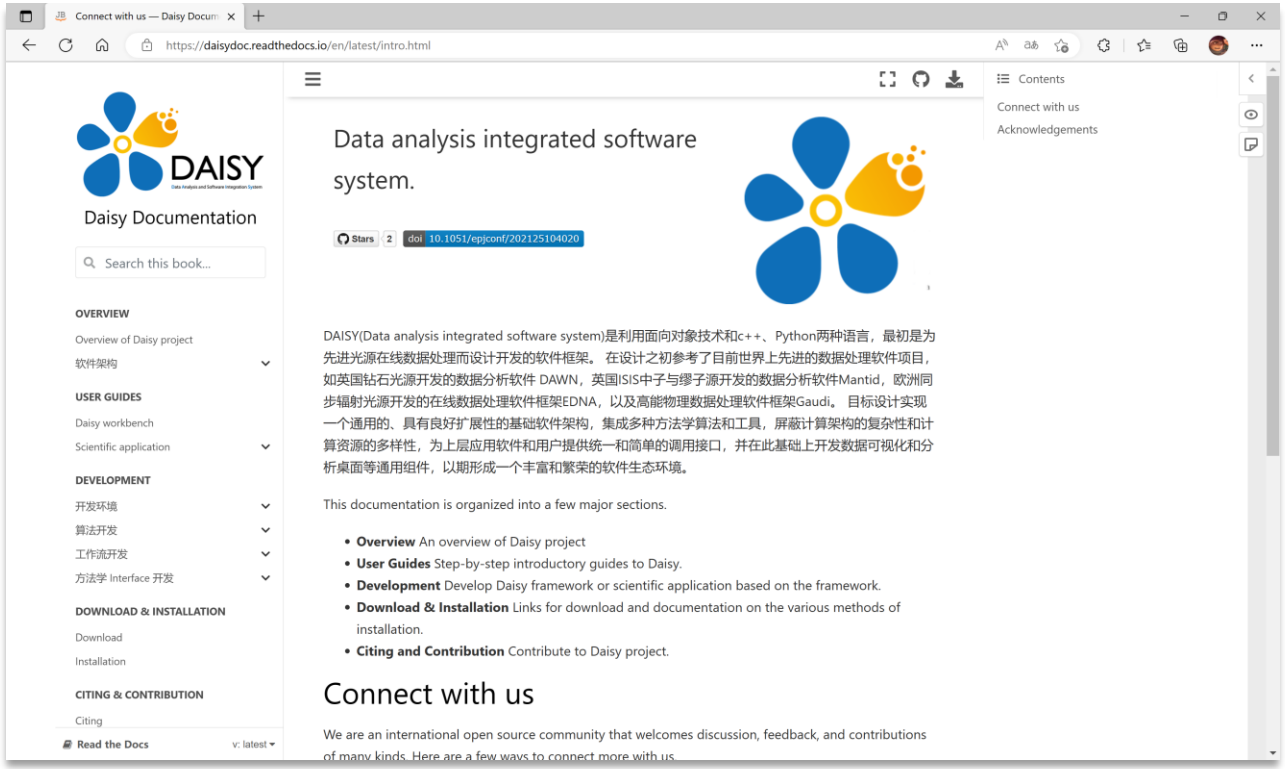


Outline

1. 背景介绍
2. 科学数据与软件系统的需求与挑战
3. 软件框架的架构与设计
4. **软件框架的开发进展**
5. 总结

Daisy framework

- 设计实现了通用科学数据处理基础软件框架 Daisy（Data analysis integrated software system）
- 向外提供了四类基础编程接口：
  - 算法模块和工作流模块：实现领域方法学模型和具体业务逻辑
  - 工作流引擎模块和数据仓库模块：管理软件运行时环境和数据对象
- 多种图形化应用接口：通过用户界面，领域方法学集成接口，web 应用接口、用户开发 IDE
- 代码和容器镜像开源，发布了网页版用户文档，建立了丰富的内部知识库
- 集成了十几个学科方法学软件与算法，开发了多个学科方法学应用

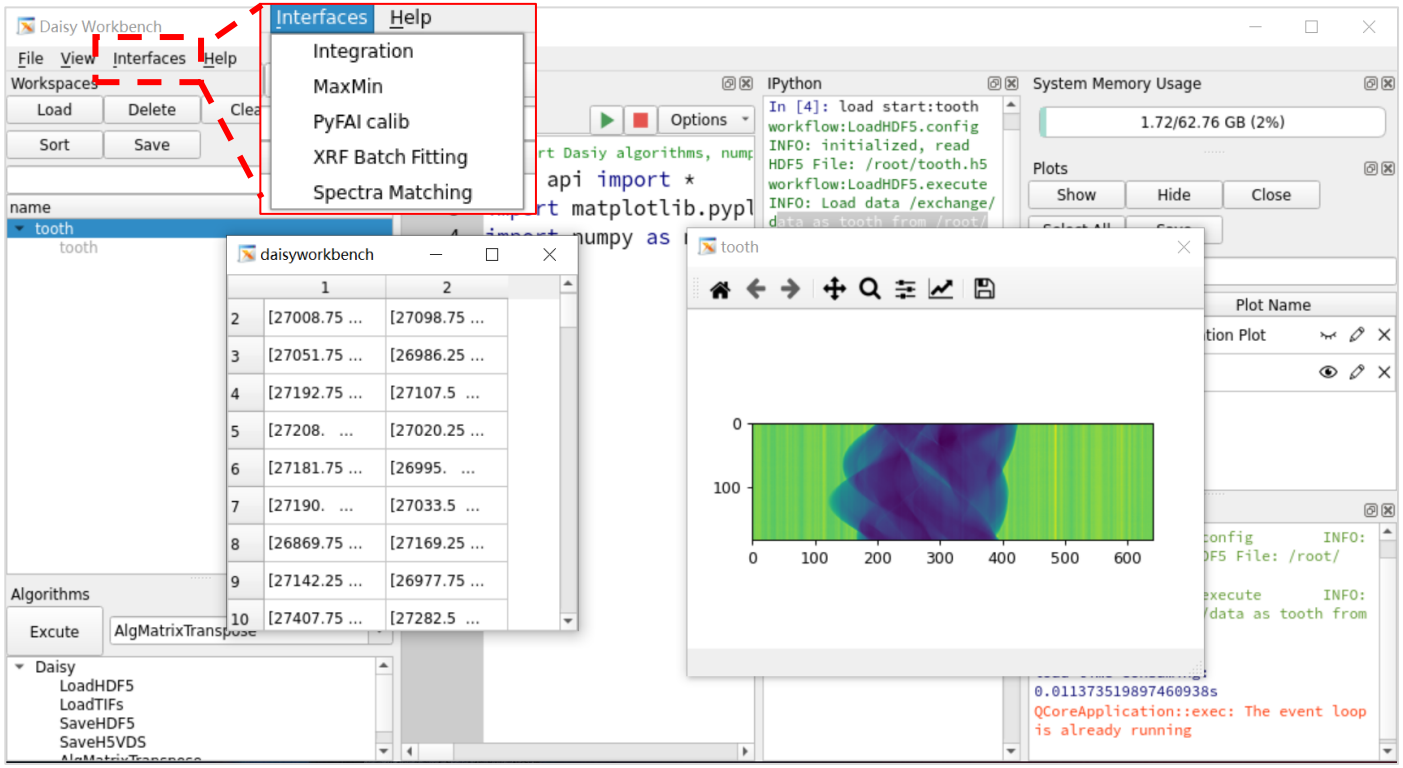


User documentation：

https://daisydoc.ihep.ac.cn/

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

Daisy graphical user interface



Daisy workbench:

- 通用用户界面，基于 PyQt5
- 包含数据对象列表，算法列表，数据展示/可视化，Log信息，系统内存监控，以及提供给开发者的 IDE 等模块
- 面向多样化学科方法学的专用用户界面接口

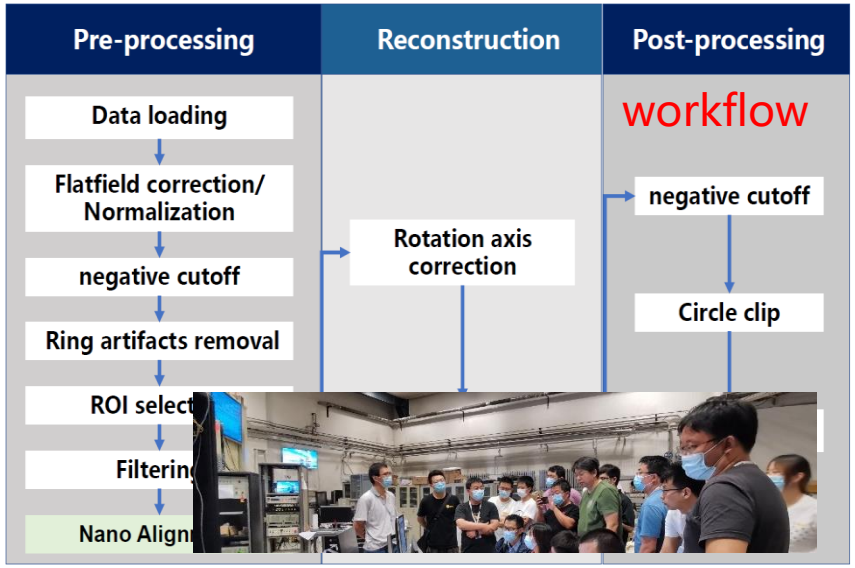
Web data analysis platform:

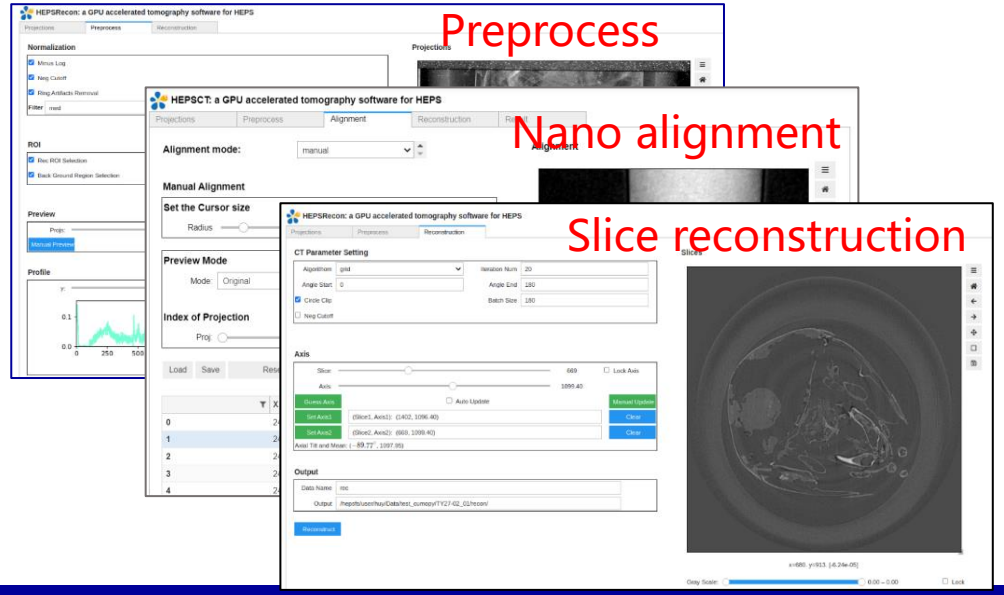
- 基于 Jupyterlab 生态，开箱即用
- 利用容器技术封装开发和运行时环境
- 通过K8s提供弹性可伸缩的计算资源
- 终端+用户友好界面，适合不同专业程度的用户

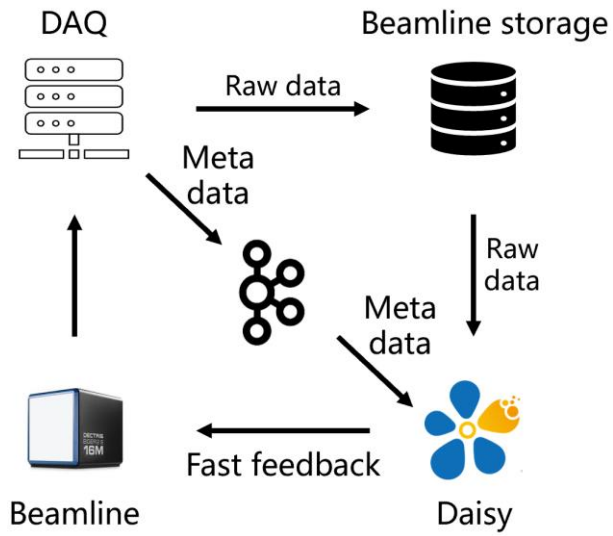
Web based application for X-ray CT

- 集成了多个CT数据处理软件：Tomopy，UFO，及HEPS-BE自研的HEPSCT，基于 GPU 硬件加速
- 基于 WEB 的用户交互式应用，支持多种数据格式的显微CT和纳米CT的断层扫描重建
- 部署在交互式计算平台上，支持多个线站(HEPS-B1，B2，B4，B7，BC，BE)的用户进行测试和使用
- 在BSRF 3W1A测试床上与HEPS-B7，DAQ 系统联合验证了基于文件的自动化数据处理流程
- 未来将提供基于流的数据处理，分布式并行化支持，相位恢复等更多功能

原始数据	重建时间
2048×2048×1442(12GB)	1min
6144×4400×1500(76GB)	10mins
10668×4402×7200(630GB)	2h

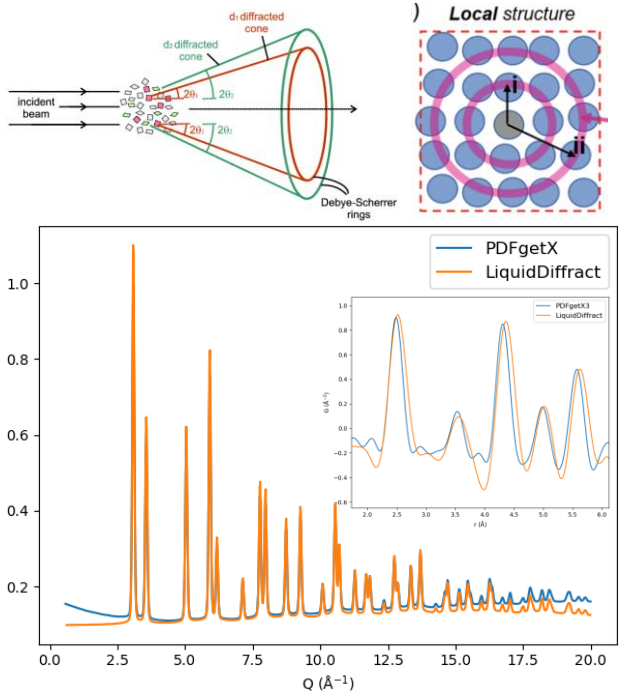


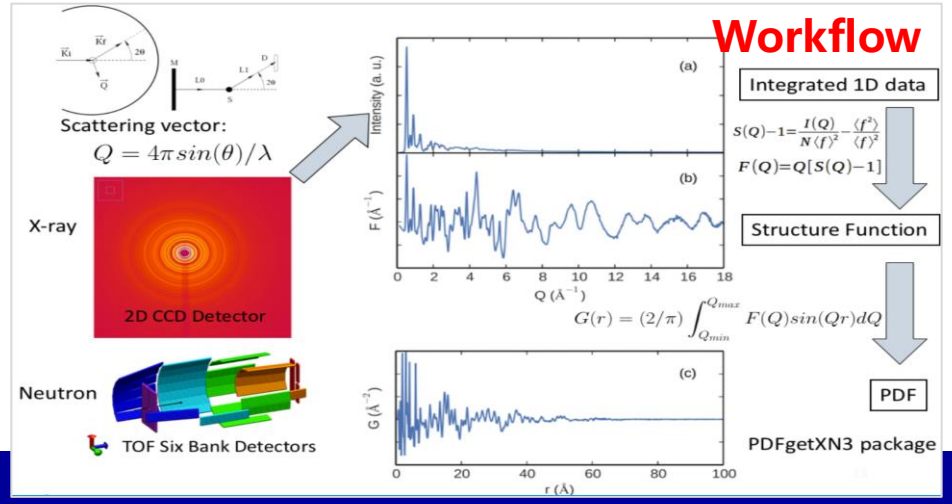


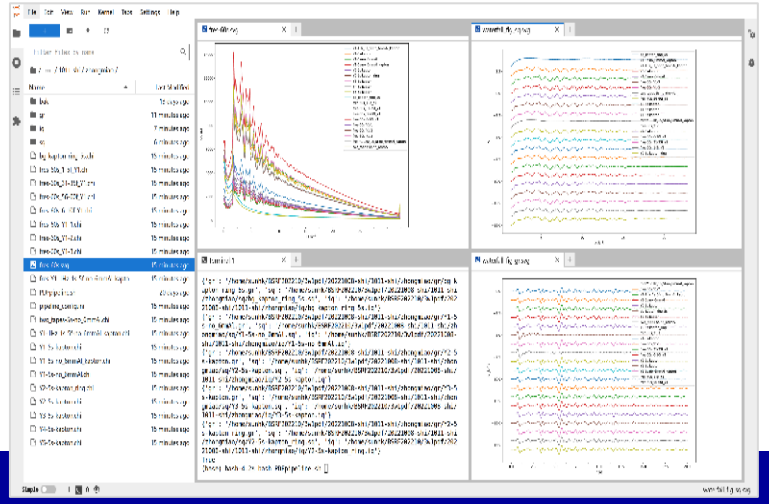


Application for Pair distribution function(PDF)

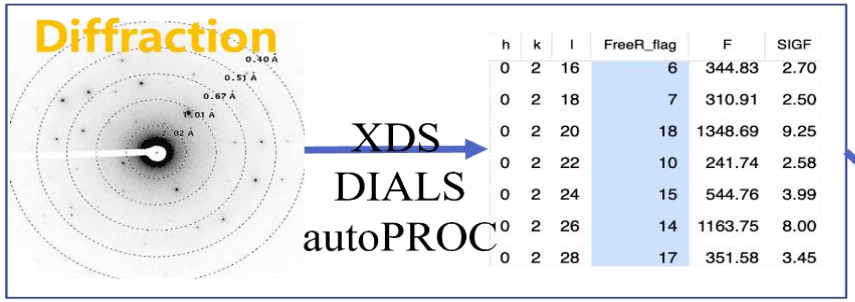
- 面向全散射实验，直接揭示实空间材料中原子间的距离
- 集成了 PyFai， PDFgetX3，liquiddiffract 等科学软件， 形成 PDFHEPS 包
- 基于 web 的用户交互式 PDF 数据处理应用，提供了从衍射数据到 PDF（包括背景扣除、masking、方位角积分、PDF 转换和结果可视化等）的完整 pipeline
- 权威 PDF 软件 PDFgetX3 需要申请 license，开源软件仅可用于快视，精细化分析还需要进一步优化
- 基于 Daisy 框架，合作发展具有自主知识产权的 PDF 软件







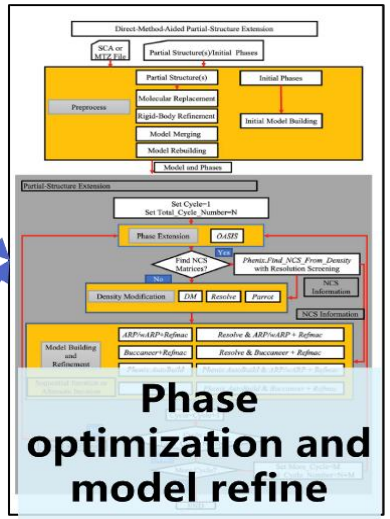
AI-based application for biological macromolecules

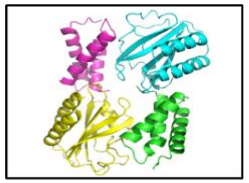


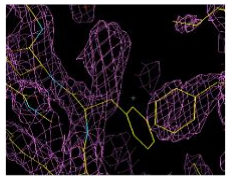
h	k	l	FreeR_flag	F	SIGF
0	2	16	6	344.83	2.70
0	2	18	7	310.91	2.50
0	2	20	18	1348.69	9.25
0	2	22	10	241.74	2.58
0	2	24	15	544.76	3.99
0	2	26	14	1163.75	8.00
0	2	28	17	351.58	3.45

Real-time data processing

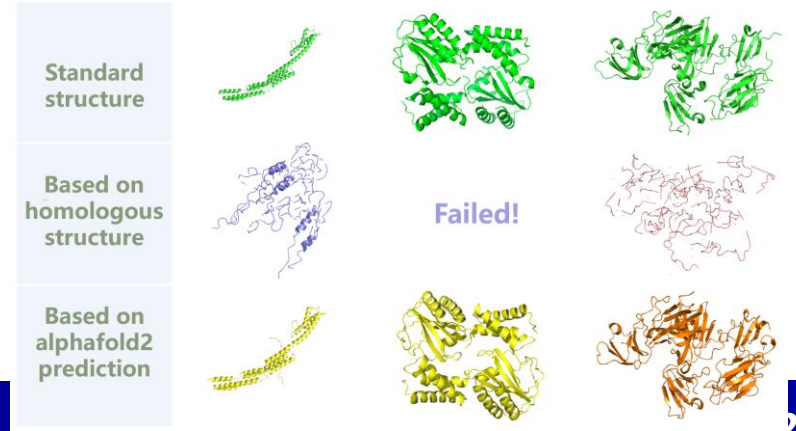
Structure prediction based on AlphaFold2







Structure truing based on AI

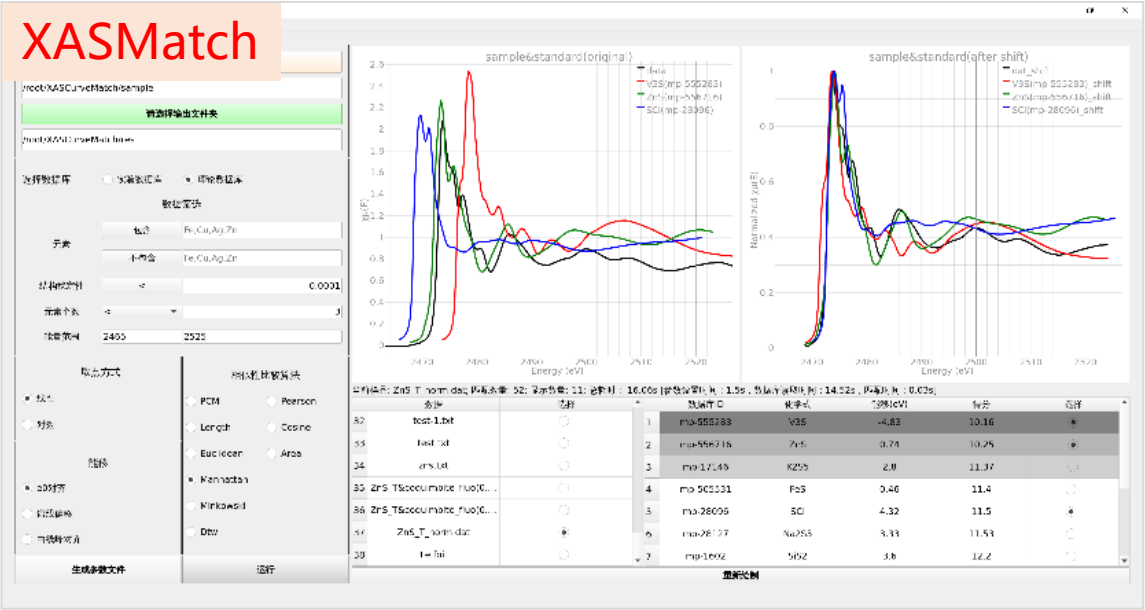


- 从衍射数据到生物大分子结构的全自动流水线
- 可用模块: 数据处理，结构预测，模型构建。将增加基于AI的结构精修模块
- 基于 web 可视化用户操作界面，提供数据处理状态监控和结果查询
- 结合传统算法与人工智能算法，提高了大分子结构解析的成功率和精确性

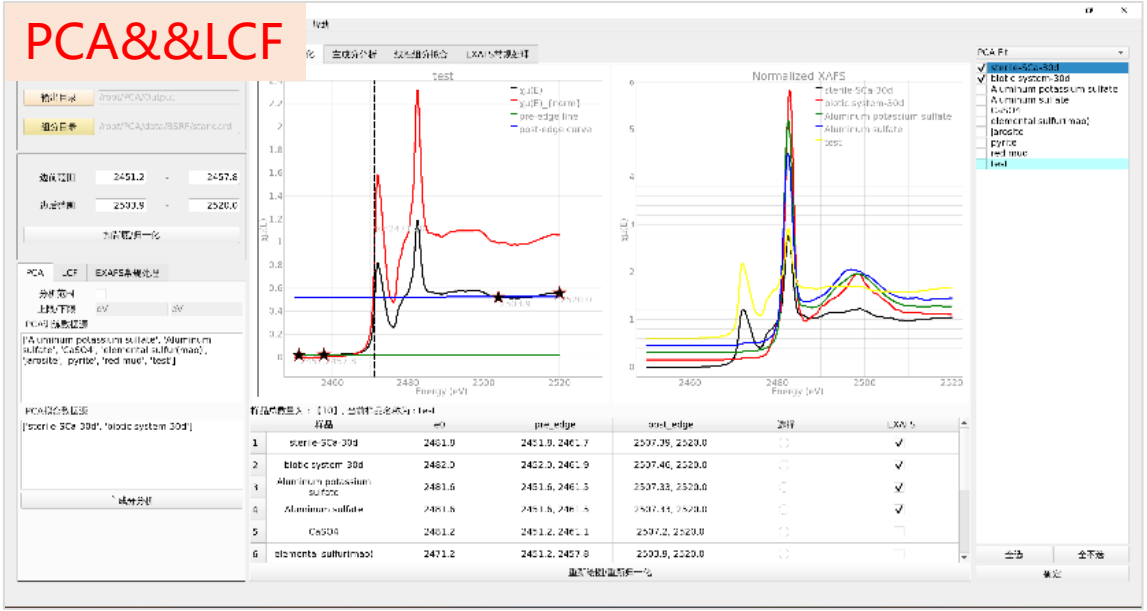
Applications for X-ray absorption spectroscopy

- 集成了多种谱学分析算法，包括数据预处理，谱线匹配，主成分分析，线性联合拟合等
- 开发了吸收谱线站谱线匹配应用XASMatch， PCA&&LCF谱线组分分析应用
  - XASMatch：对实验谱在数据库中进行快速匹配，集成多种匹配算法、能移方法，支持多种数据库输入
  - PCA&&LCF：适用于催化、电池反应中相变的成分分析，支持自动化批处理、多标准谱输入
- 计划与PAPS平台合作发展国内本领域的谱学实验数据库，充分利用光源科学数据资源

XASMatch



PCA&&LCF



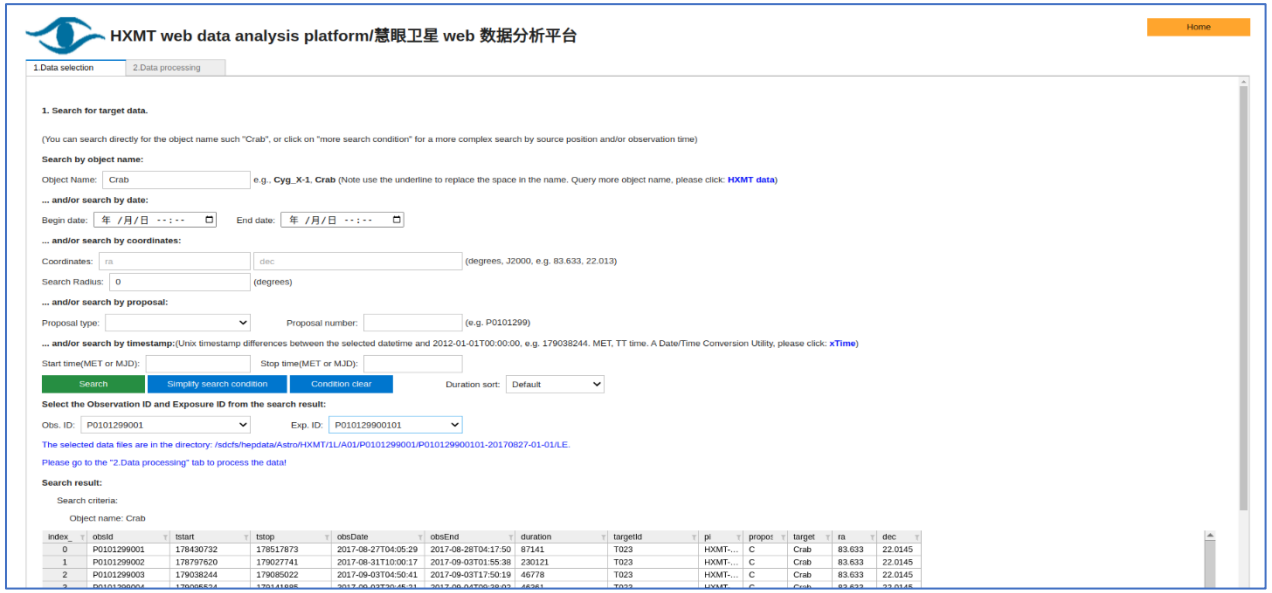
Daisy 在空间天文学的应用

可能的应用场景

- 数据处理，数据分析，数据产品生成
- 探测器模拟，观测模拟
- 整合现有软件资源，形成共性软件包

HXMT web 数据处理平台

- 基于 jupyterlab 的 HXMT web 数据分析平台
- 通过 web 浏览器为用户提供数据处理环境和服务



https://sdccompute.ihep.ac.cn/

Svom 数据产品生成软件集成

- 尝试将软件框架应用于 **Svom** 卫星二级数据产品生成
- 完成了 fits 数据文件读写算法的集成，部分数据产品生成算法

eXTP 卫星数据处理软件集成

- 已经完成数据模板生成，数据分割等算法的集成

Outline

1. 背景介绍
2. 科学数据与软件系统的需求与挑战
3. 软件框架的架构与设计
4. 软件框架的开发进展
5. **总结**

总结

- 四代光源的建成以及探测器技术的发展，科学数据爆发式增长，带来了新的机遇与挑战
- 大数据场景下的科学发现，需要先进的数学及算法
- 多样化的应用场景及数据结构，使得学科软件及算法的发展需要底层软件框架支持
- 围绕HEPS需求，规划、设计了面向先进光源的科学数据处理分析软件框架
- 基础框架已经形成并验证，围绕软件框架，学科方法学开发在持续进行中
- 科学软件生态的发展还需要用户社群的支持与参与

https://daisydoc.ihep.ac.cn/

谢谢！