Center for Transformative Science
ShanghaiTech University
大科学中心

第二十一届全国科学计算与信息化会议

X射线自由电子激光装置科学数据管理软件建设与展望

报告人：雷 蕾

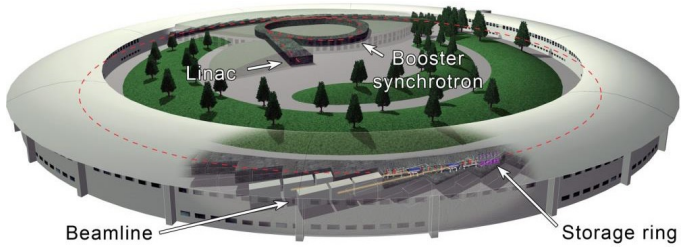
2025年8月26日

上海科技大学
ShanghaiTech University

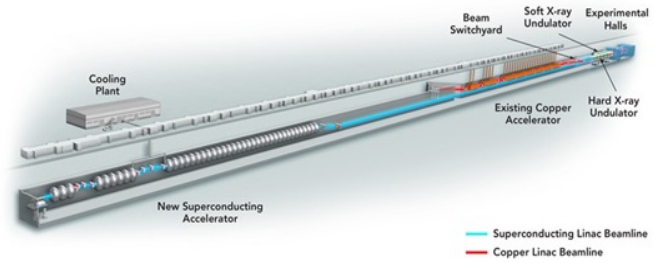
目 录

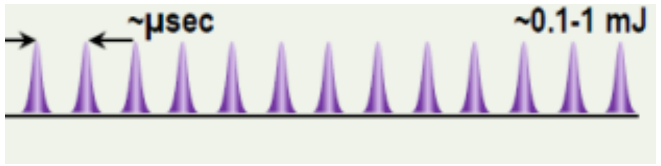
先进光源装置

Synchrotron Radiation
同步辐射光源

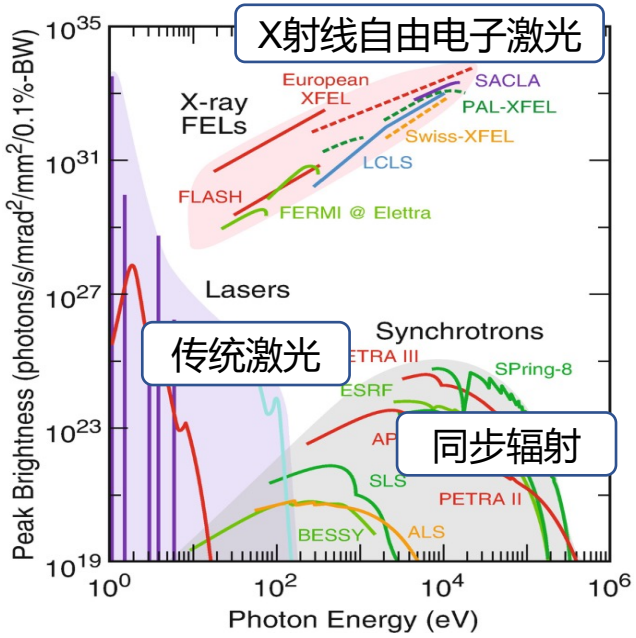


X-Ray Free Electron Laser
X射线自由电子激光装置

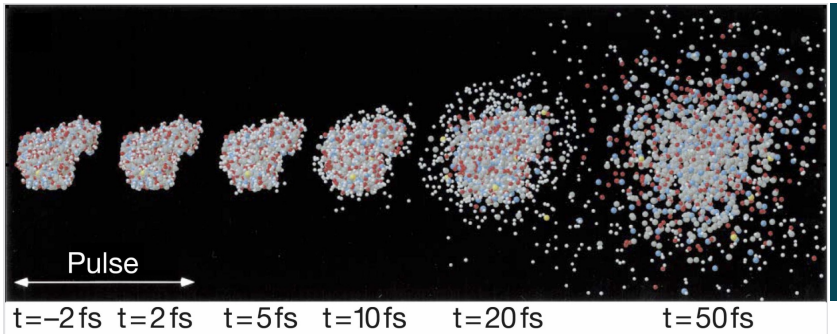




高亮度、短脉冲、强相干



X射线自由电子激光与第三代同步辐射光源峰值亮度的对比



Nature **406**, 752–757 (2000)

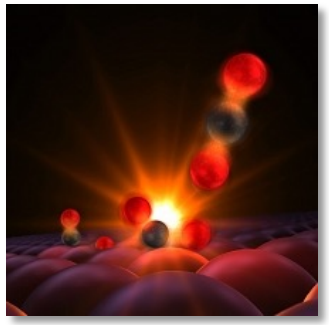
High Brightness + Ultrafast
Detection Before Destruction

S. Boutet et al., X-ray Free Electron Lasers: A Revolution in Structural Biology, Springer, Switzerland, 2018

XFEL可探索超微空间和超快时间尺度

□ 探测超小结构、捕获超快过程的强有力工具，促进学科前沿发展

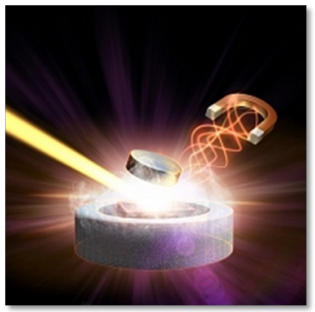
化学



化学反应动态过程跟踪（成键、断键、中间体）
例如：跟踪化学键形成过程

Science 347, 978–982 (2015)

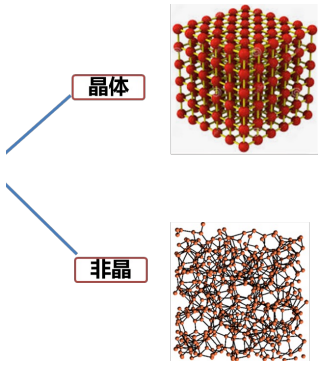
物理



观测复杂体系的电子结构动态过程
例如：高温超导的机理

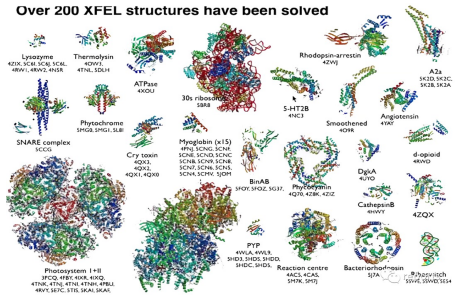
Science 357, 71 (2017)

材料



非晶材料结构解析
测量非晶材料原子位置及纳米尺度相变过程

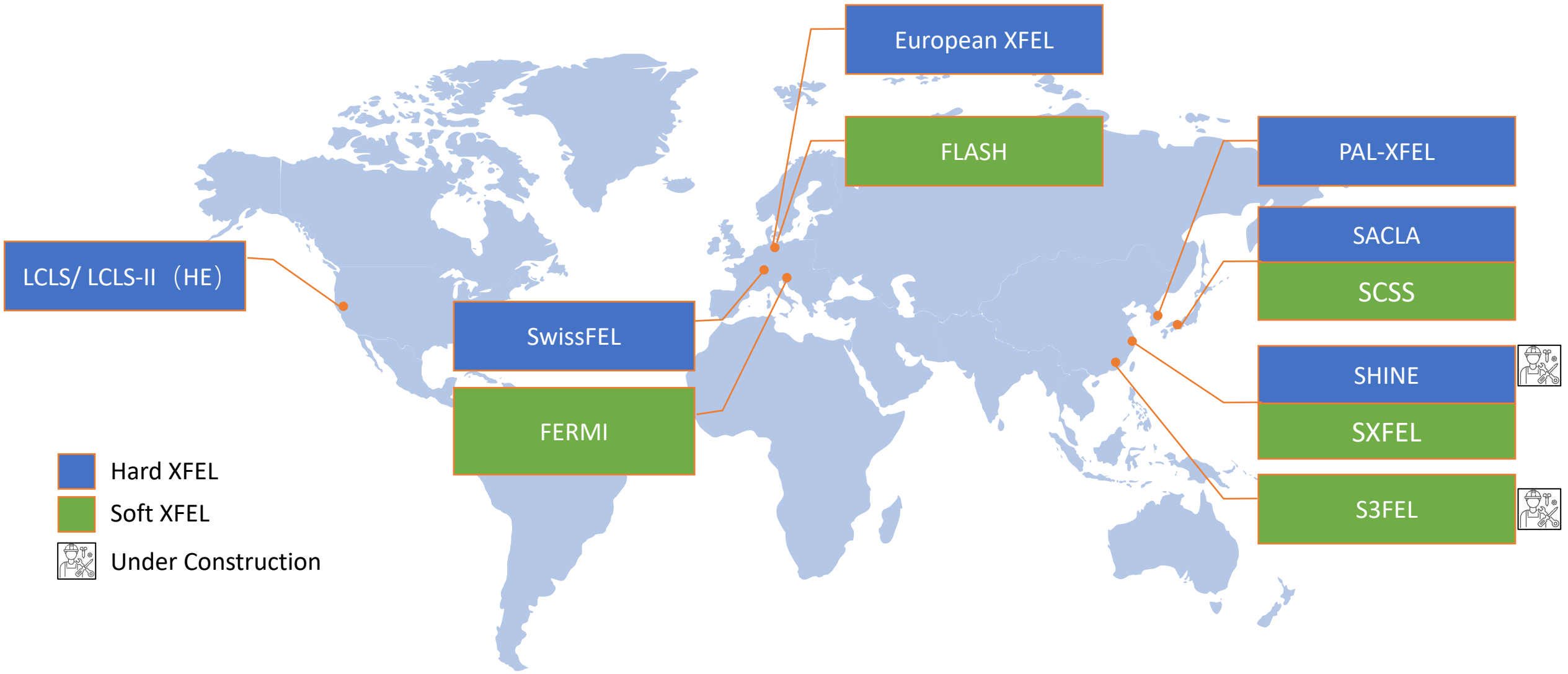
生命科学



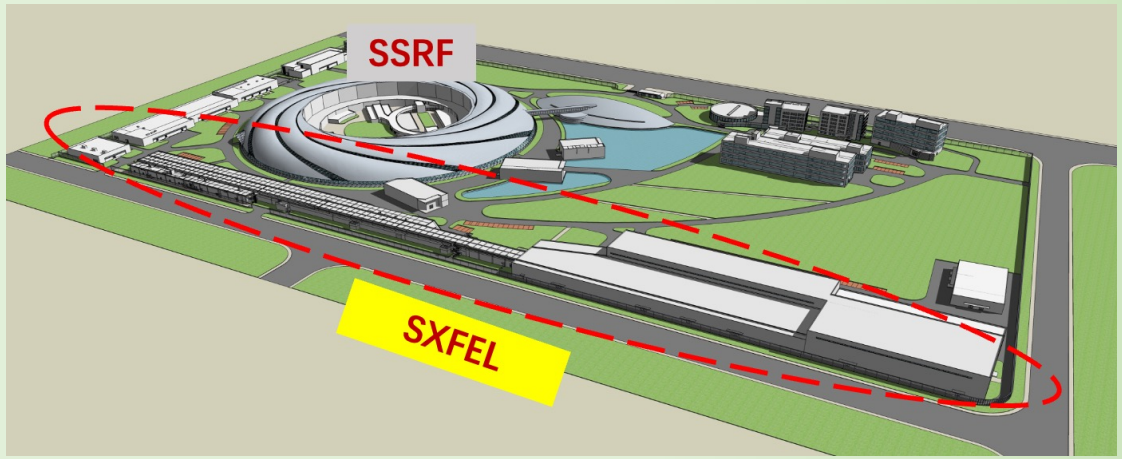
原位观测生命体的动态过程
例：蛋白质晶体动态结构解析

Courtesy of H. Chapman, DESY

XFEL装置全球布局


European XFEL
FLASH
PAL-XFEL
SACLA
SCSS
LCLS/ LCLS-II （HE）
SwissFEL
FERMI
SHINE
SXFEL
S3FEL
Hard XFEL
Soft XFEL
Under Construction

SXFEL & SHINE



Shanghai Soft X-ray Free Electron Laser Facility

上海软X射线自由电子激光装置（SXFEL）

2022年建成，2023年开放运行

与SSRF形成集成布局，功能与优势互补

最高光子能量为620eV，具有独特的科学应用前景：

（1）水窗全覆盖（2.3-4.4 nm，生命科学等）

（2）覆盖6.5-13.5nm的全相干光源和超快光源

Shanghai HIgh repetitioN rate XFEL and Extreme light facility

硬X射线自由电子激光装置（SHINE）

2018年开工建设，计划2027年建成

装置总长3.1km，埋深地下29m

电子束能量8GeV，重复频率1MHz，光子能量0.4-25keV

nm级高分辨率成像、先进结构解析、飞秒超快过程探索等尖端的科学研究手段，可拍分子电影

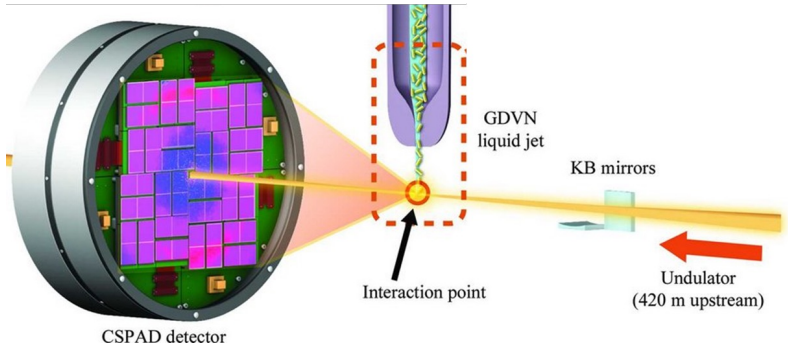
XFEL面临的数据挑战

机遇

- 高亮度+短脉冲——样品破坏前成像、活细胞
- 强相干—— 相干衍射成像、研究非晶态
- 超高空间分辨率 —— ~ 10^{-10} 米分辨率
- 超高时间分辨率 —— ~ 10^{-15} 米分辨率
- 波长范围覆盖宽、连续可调谐范围广

挑战

- 数据通量大 —— ~百GB/s
- 数据总量大 —— ~十PB/年
- 数据价值密度低 —— ~1%命中率
- 数据处理复杂 ——种类和来源多样化（控制、诊断、定时、探测器…）；在线+离线数据处理

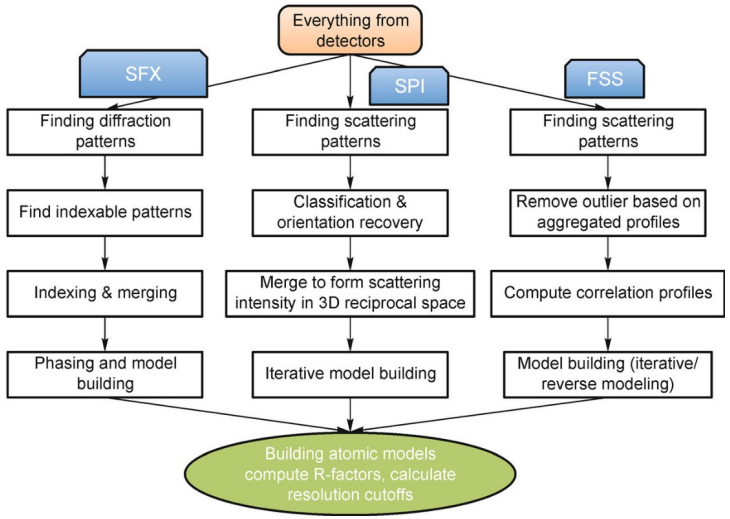


样品溶液高速喷出——> 样品随机取向——>解析一个结构需要百万级的原始衍射图样

Run Number	Number of frames in XTC file	Number of hits by diffusion map embedding	Number of hits by principal component analysis	Number of hits by manifold embedding
182	97,733	15	2	0
183	169,227	3	0	0
184	136,624	1	0	0
185	77,079	7	4	0
186	453,285	2,048	1,121	4,759
188	322,097	1,171	599	2,885
190	265,585	1,498	838	4,175
191	293,284	1,718	944	4,427
192	224,583	1,279	656	3,209
193	329,049	2,525	1,370	6,781
194	214,891	1,661	897	4,555
196	204,236	1,436	818	3,792
197	188,895	1,410	743	2,967
Total	2,976,568	14,772	7,992	37,550

PR772噬菌体，共采集原始图像**2,976,568张，约6TB，最高命中率为1.26%**

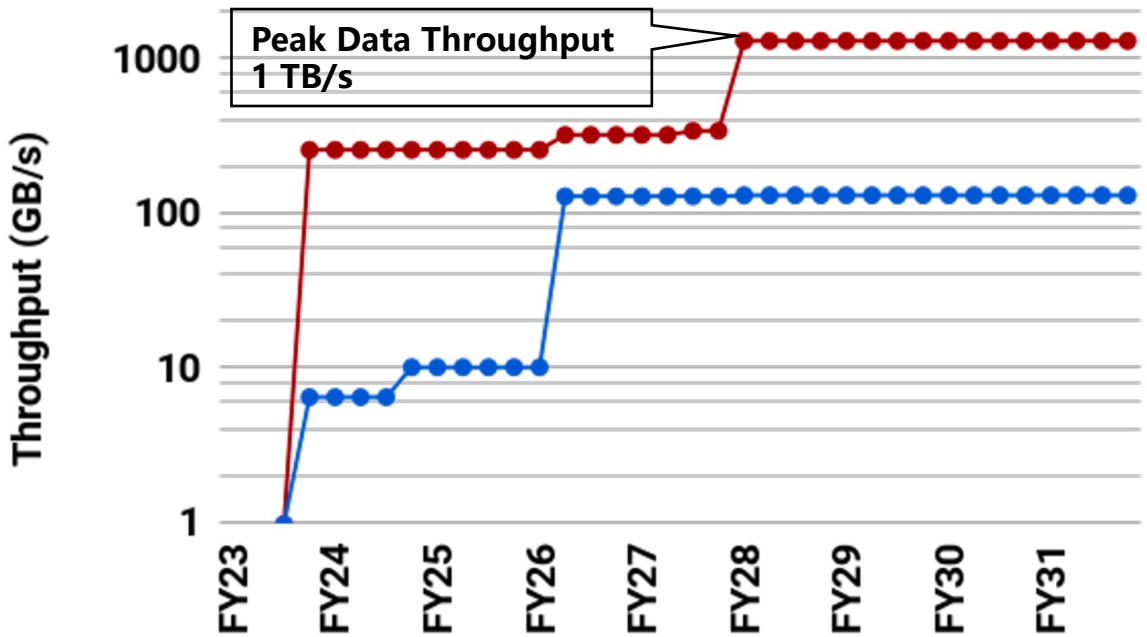
H.K.N. Reddy et al., Scientific Data 4, 170079 (2017)



Quantitative Biology, 4 159–176 (2016)

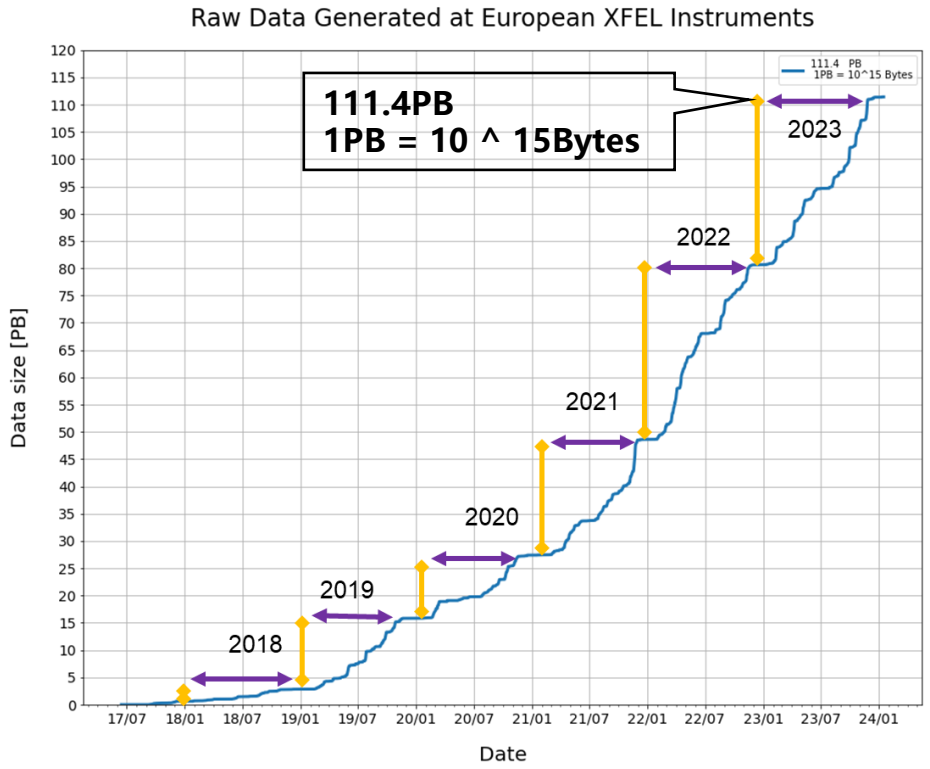
XFEL面临的数据挑战

Data Throughput @LCLS



Jana Thayer for the LCLS Data Systems Team "Data Systems Update" UEC Meeting 2021

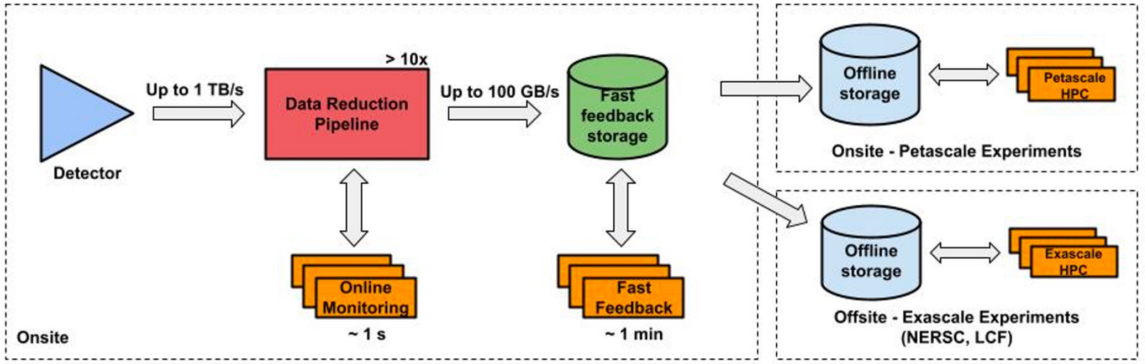
Raw Data Volume @EuropeanXFEL

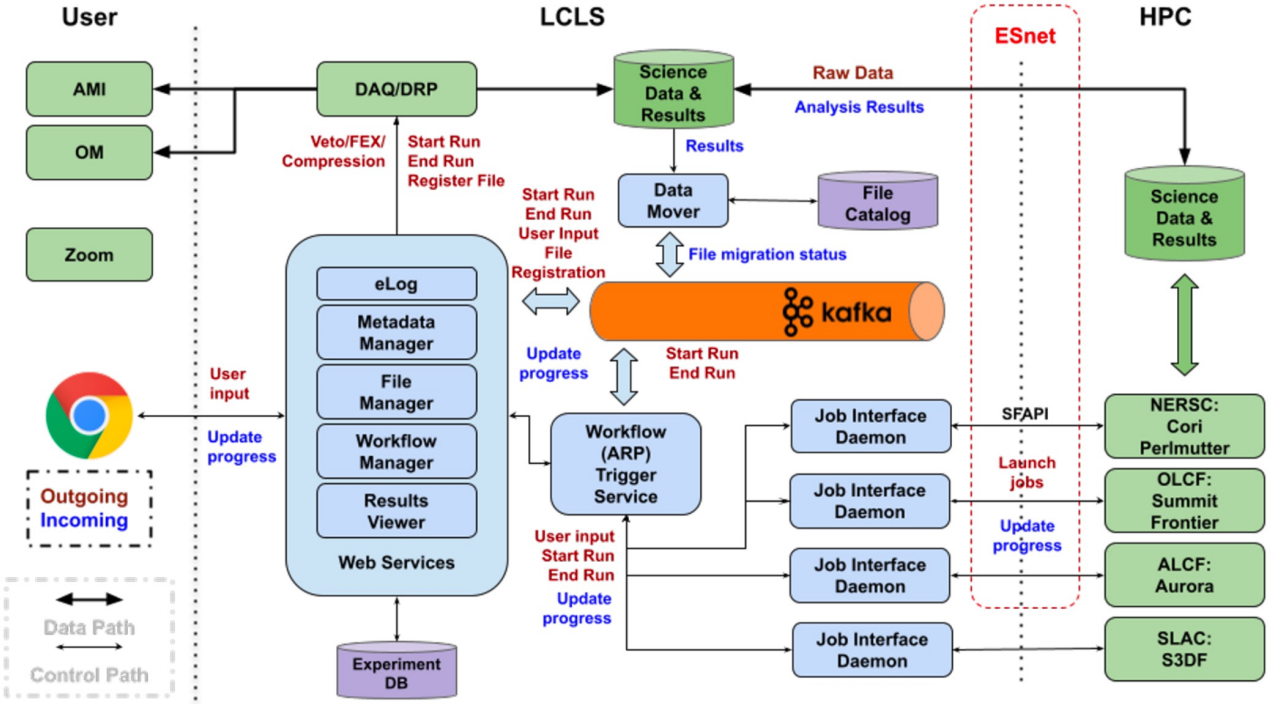


Krzysztof Wrona "Scientific Data Policy of the European XFEL" European XFEL Users' Meeting2024

美国LCLS/LCLS-II

- LCLS脉冲频率 120 Hz，数据传输速率 5 GB/s
- LCLS-II脉冲频率提升至 1 MHz，数据传输速率 > 100GB/s





LCLS数据管理系统

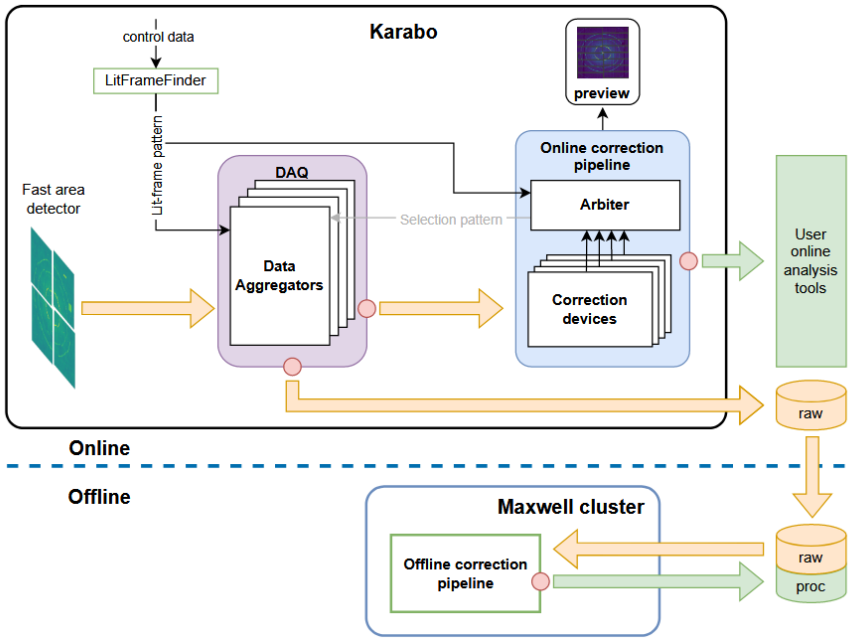
实现LCLS实验数据生命周期管理，注重数据完整性和提供安全访问，支持高重频装置的极端需求：

- Data Access portal：WEB界面允许用户从任何地方浏览、搜索和检索实验数据和元数据。
- Data Storage & DataMover：多层次存储基础设施提供短期和长期数据存储，和跨存储的自动数据移动。
- ML用于数据管理：数据智能分类和标记、数据生命周期预测、基于ML的数据压缩

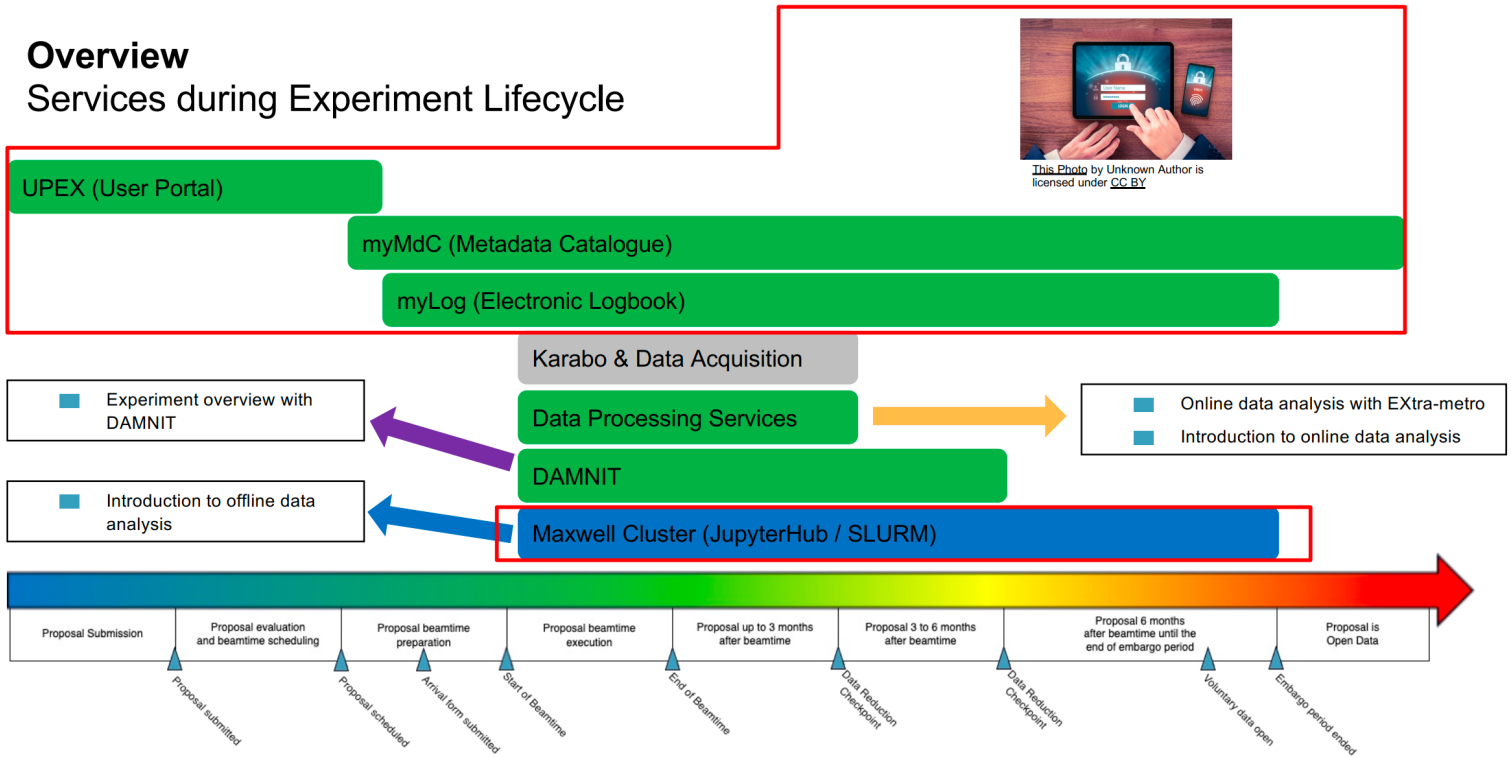
EPJ Web of Conferences 295, 13002 (2024)

European XFEL

- 系统架构：在线层（快速缓存，匹配数据生成速率15GB/s）、高性能存储（用于实验期间和实验后的数据处理，1Tb/s链路连接）、大容量存储（扩展存储容量，中期数据访问）、磁带存档（数据安全性、长期存档）
- 科学数据策略：定义实验所有参与方的数据权利和责任，制定数据管理规则
- 元数据目录服务：myMdc，基于关系型数据库构建，还作为与其他服务集成的枢纽

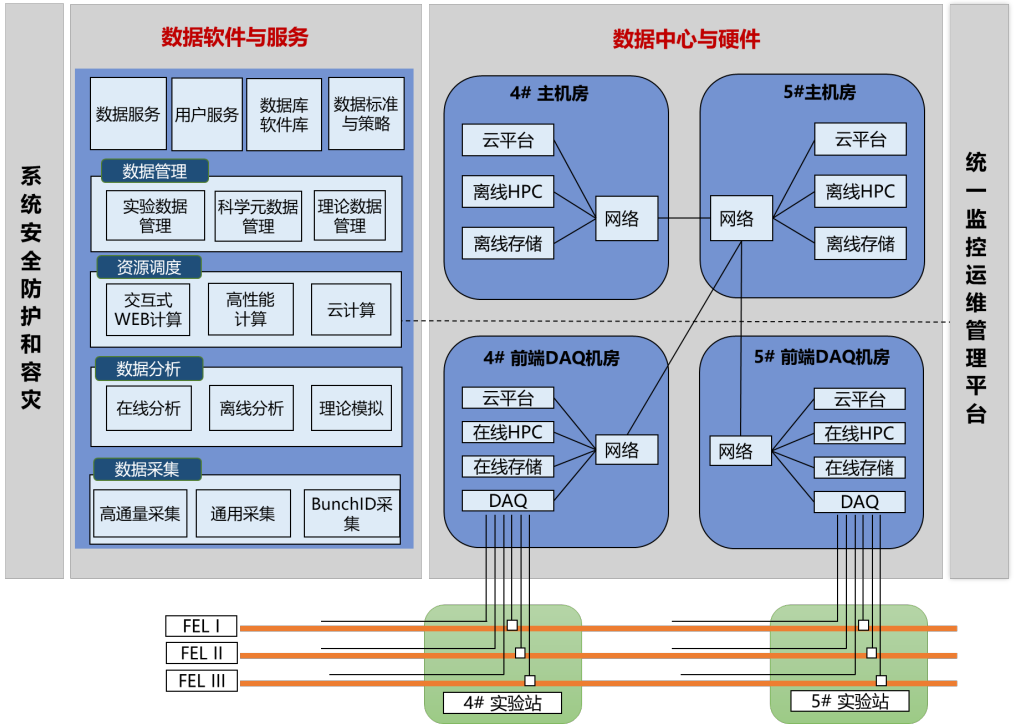


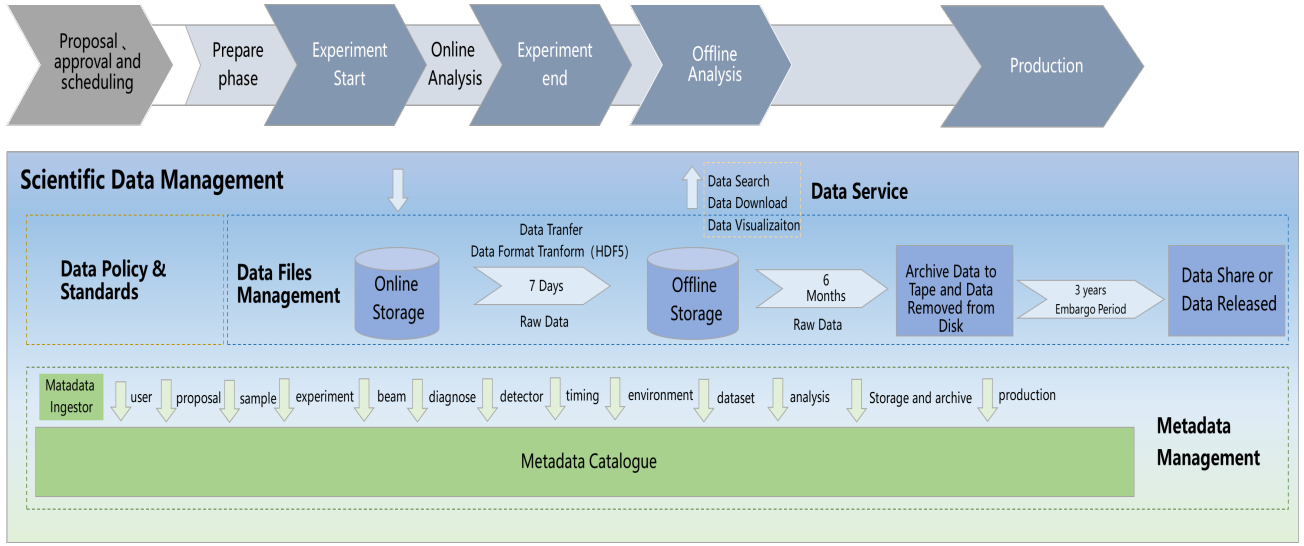
Data infrastructure



SHINE

- SHINE数据通量初期可达几十GB/s（10KHz），最终可达到TB/s（1MHz），年数据量上百PB
- 将提供数据采集、数据处理分析、数据传输、海量数据存储和管理、数据归档和长期访问所需的高性能硬件系统和软件系统



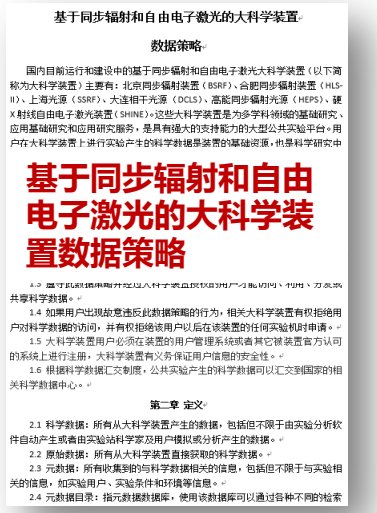


实验数据全生命周期管理

对海量实验数据进行分级处理和统一管理，提升数据利用效率

科学数据策略

- 提供装置用户数据管理和访问服务
- **遵循FAIR原则**（Findability，Accessibility，Interoperability，and Reuse）
- 明确设施/机构和用户的责任和义务
- 指导SHINE数据管理的设计和实施


基于同步辐射和自由电子激光的大科学装置

数据策略

国内目前运行和建设中的基于同步辐射和自由电子激光大科学装置（以下简称为大科学装置）主要有：北京同步辐射装置（BSRF）、合肥同步辐射装置（HLS-II）、上海光源（SSRF）、大连相干光源（DCLS）、高能同步辐射光源（HEPS）、硬X射线自由电子激光装置（SHINE）。这些大科学装置是为多学科领域的基础研究、应用基础研究和应用研究服务，是具有强大的支持能力的大型公共实验平台。用户在大科学装置上进行实验产生的科学数据是装置的基础资源，也是科学研究中

基于同步辐射和自由电子激光的大科学装置数据策略

1.4 如果用户出现故意违反此数据策略的行为，相关大科学装置有权拒绝用户对科学数据的访问，并有权拒绝该用户以后在该装置的任何实验机时申请。

1.5 大科学装置用户必须在装置的用户管理系统或者其它被装置官方认可的系统上进行注册，大科学装置有义务保证用户信息的安全性。

1.6 根据科学数据汇交制度，公共实验产生的科学数据可以汇交到国家的相关科学数据中心。

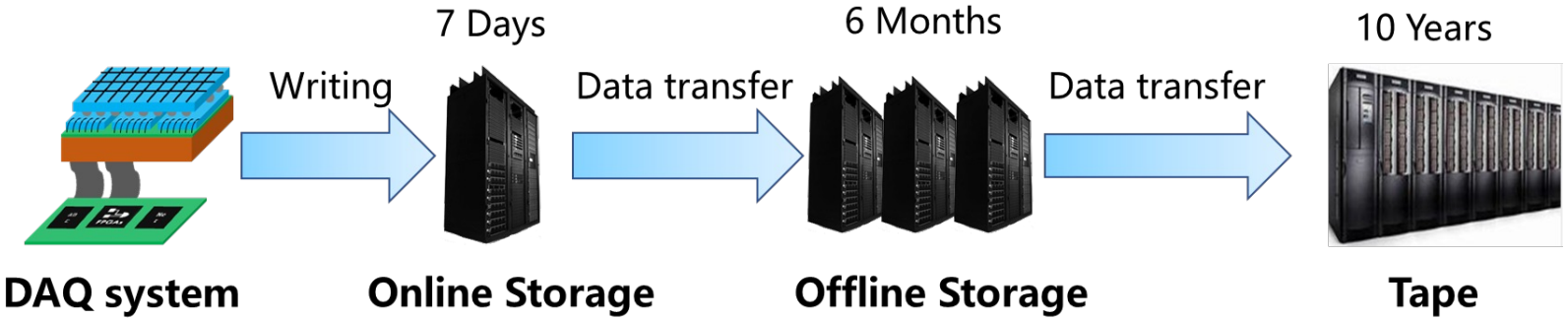
第二章 定义

2.1 科学数据：所有从大科学装置产生的数据，包括但不限于由实验分析软件自动产生或者由实验站科学家及用户模拟或分析产生的数据。

2.2 原始数据：所有从大科学装置直接获取的科学数据。

2.3 元数据：所有收集到的与科学数据相关的信息，包括但不限于与实验相关的信息，如实验用户、实验条件和环境等信息。

2.4 元数据目录：指元数据数据库，使用该数据库可以通过各种不同的检索

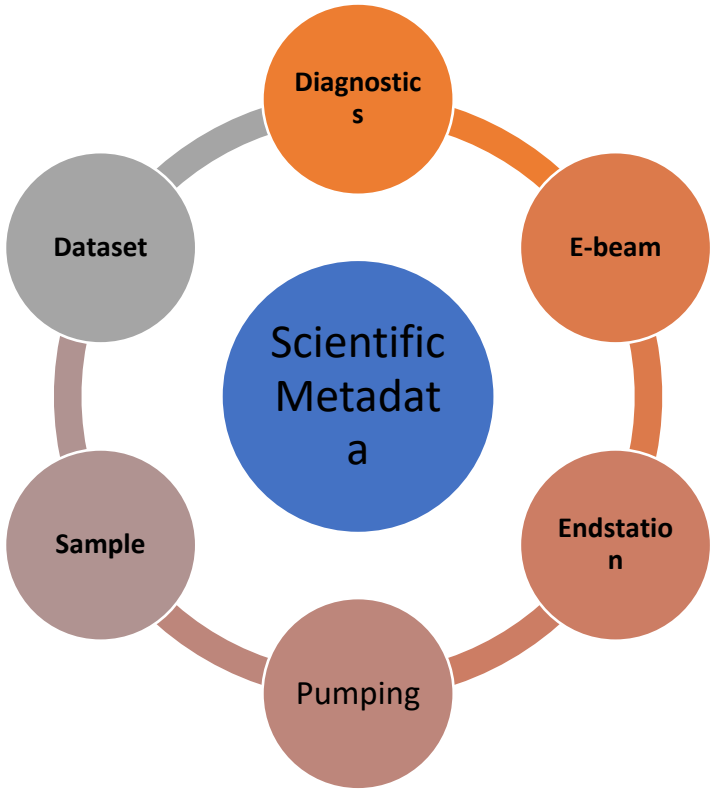


BSRF、HEPS、SHINE、HALF

元数据标准

- **高能所、网络中心、上科大、中科大牵头**制定**国内首个**光源类大科学装置实验数据元数据标准，多年来共同合作推动光源科学数据管理、利用和共享。**当前已经国标立项。**
- **装置自身根据科学特点扩展核心科学元数据字段**

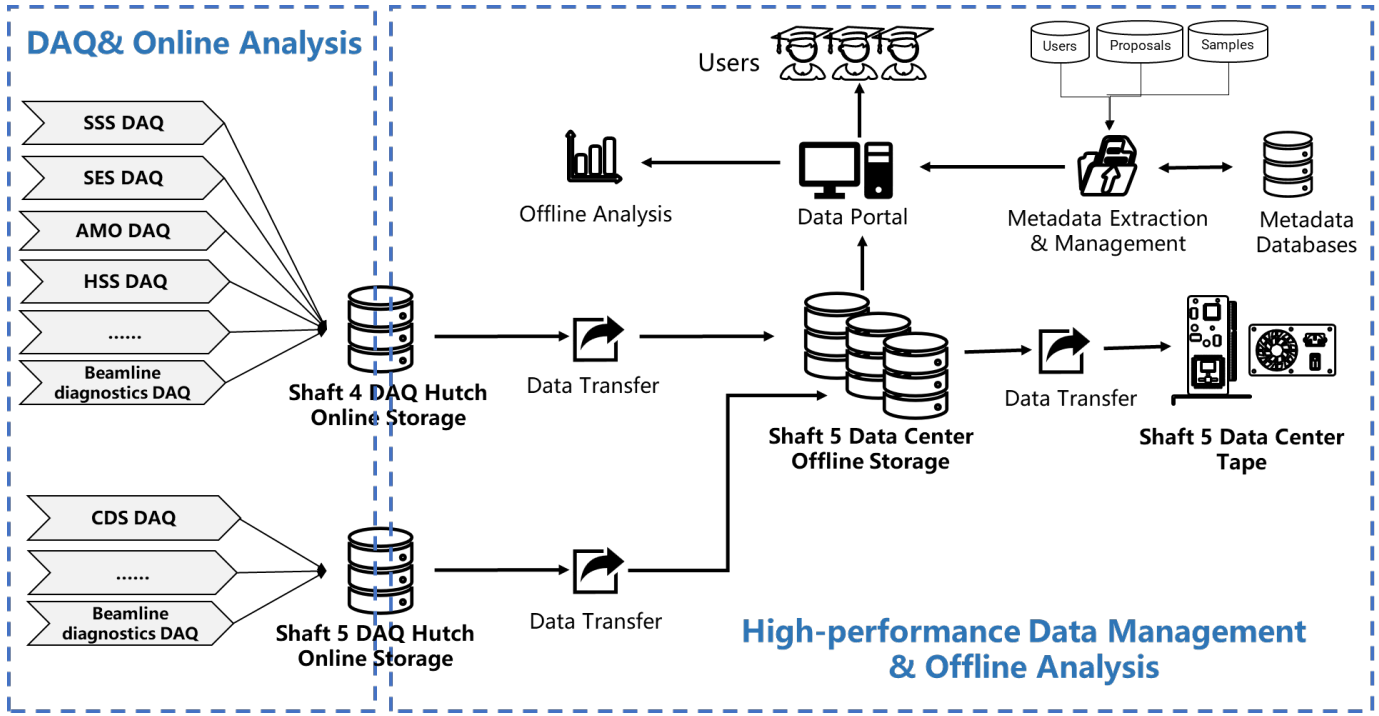
《重大科技基础设施平台 光子于射线实验数据元数据》2024年12月通过国标委答辩，于2025年4月立项



SHINE科学元数据构成

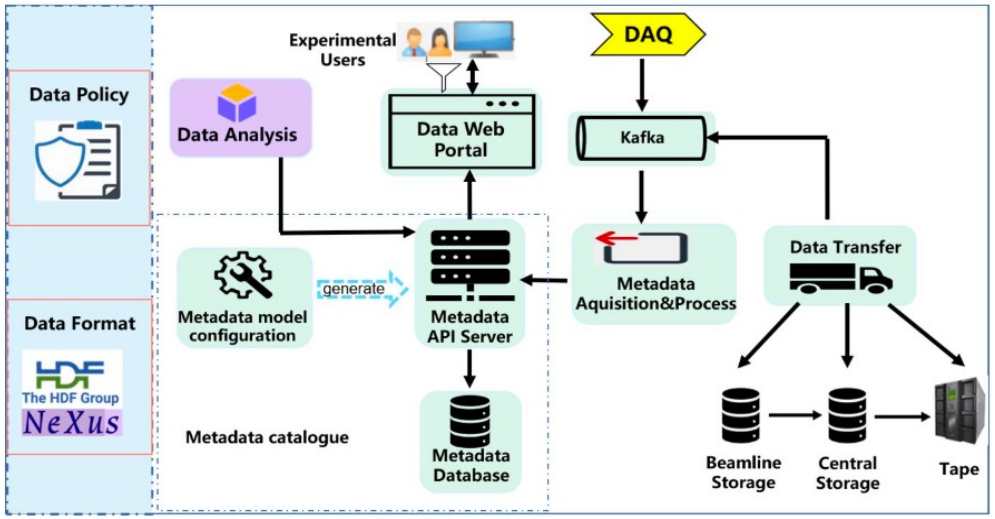
数据管理软件需求

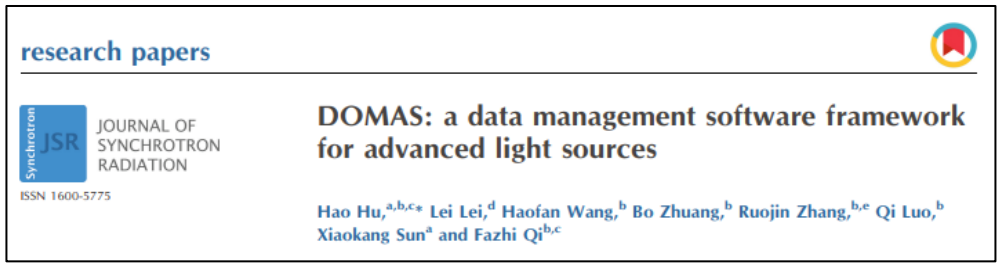
- **实验数据管理软件：对内承接数据采集系统，对外面向实验用户**
- **功能需求：**
  - **数据传输**：实验数据文件的自动化迁移和校验（在线存储 → 离线存储 → 磁带归档）
  - **元数据管理**：元数据目录和模型设计、元数据存储和访问
  - **元数据提取和整合**：科学元数据、管理元数据（消息系统、文件、数据库）
  - **数据服务**：为用户提供数据检索、在线查看、数据下载、数据授权、实验日志查看等功能



科学数据管理软件框架DOMAS+定制开发

- **科学数据管理软件框架DOMAS：**
  - 高能所自研，用于装置科学实验数据的自动化组织、传输、存储和分发
  - 提供装置数据管理所需的通用基础模块和通用接口：元数据目录、元数据提取、数据传输、数据服务
  - 低人力成本：基于DOMAS框架，装置可通过少量开发快速建立自己的数据管理系统




research papers

JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

DOMAS: a data management software framework for advanced light sources

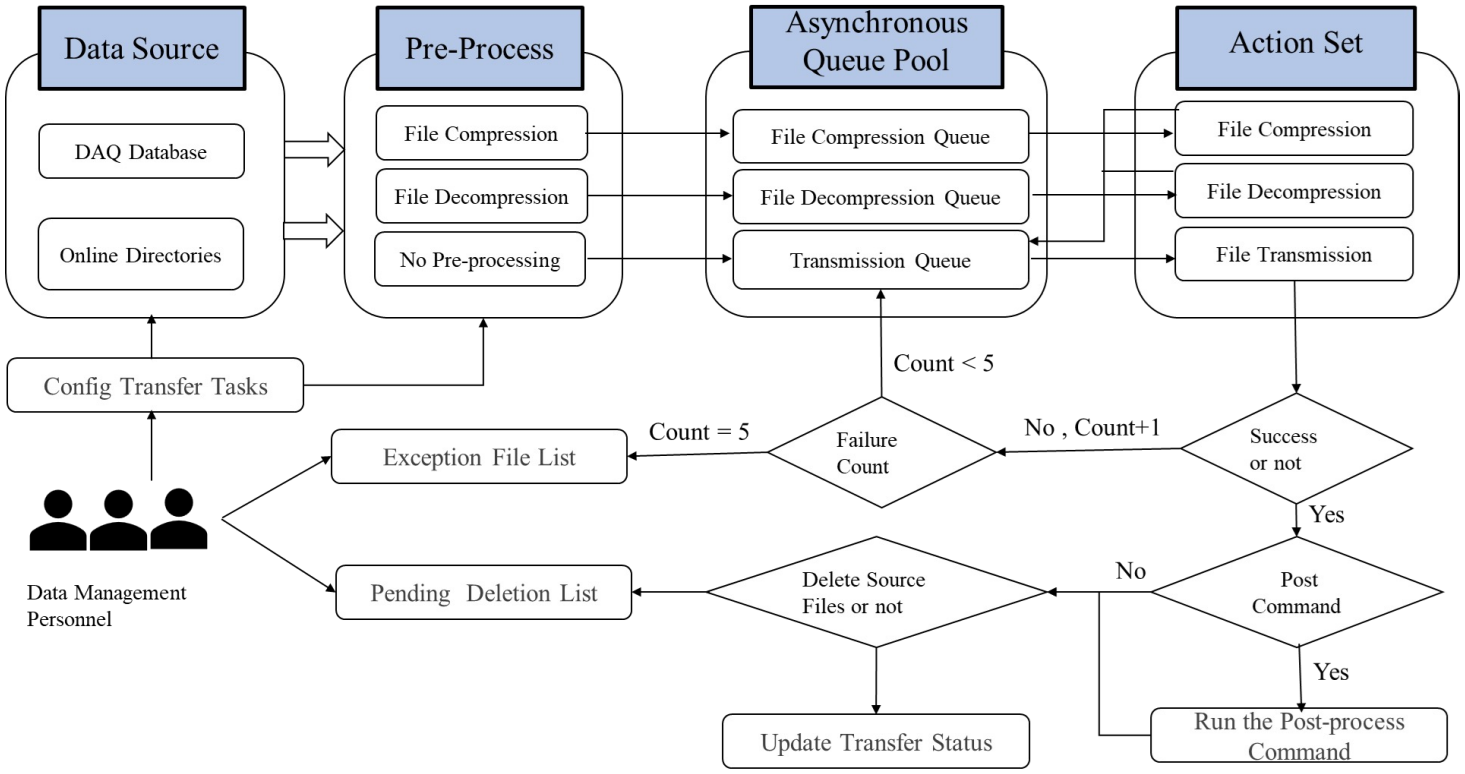
Hao Hu,[a,b,c]* Lei Lei,[d] Haofan Wang,[b] Bo Zhuang,[b] Ruojin Zhang,[b,e] Qi Luo,[b] Xiaokang Sun[a] and Fazhi Qi[b,c]

- **与SciCAT相比：**
  - + 灵活的元数据模型配置管理
  - + 集成工作流模块
  - + 可配置的数据传输
  - + 丰富的标准化API

- **当前已在国内BSRF、HEPS和散裂中子源等装置中应用。**

数据传输

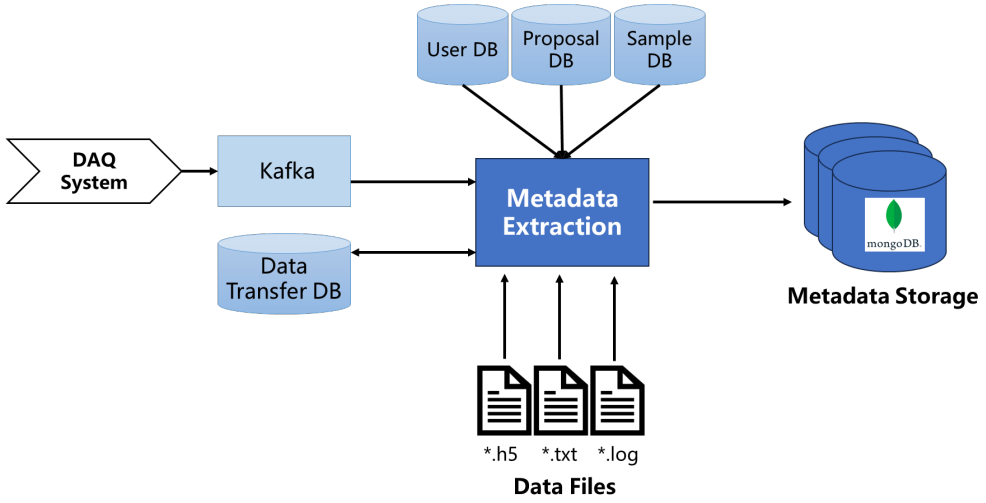
- 用于实验数据文件自动化迁移传输和传输任务可视化配置管理
- 支持多种方式获取传输源文件、支持多种传输工具，可同时支持多个实验传输任务
- 支持文件校验、支持传输日志记录和异常文件人工干预.



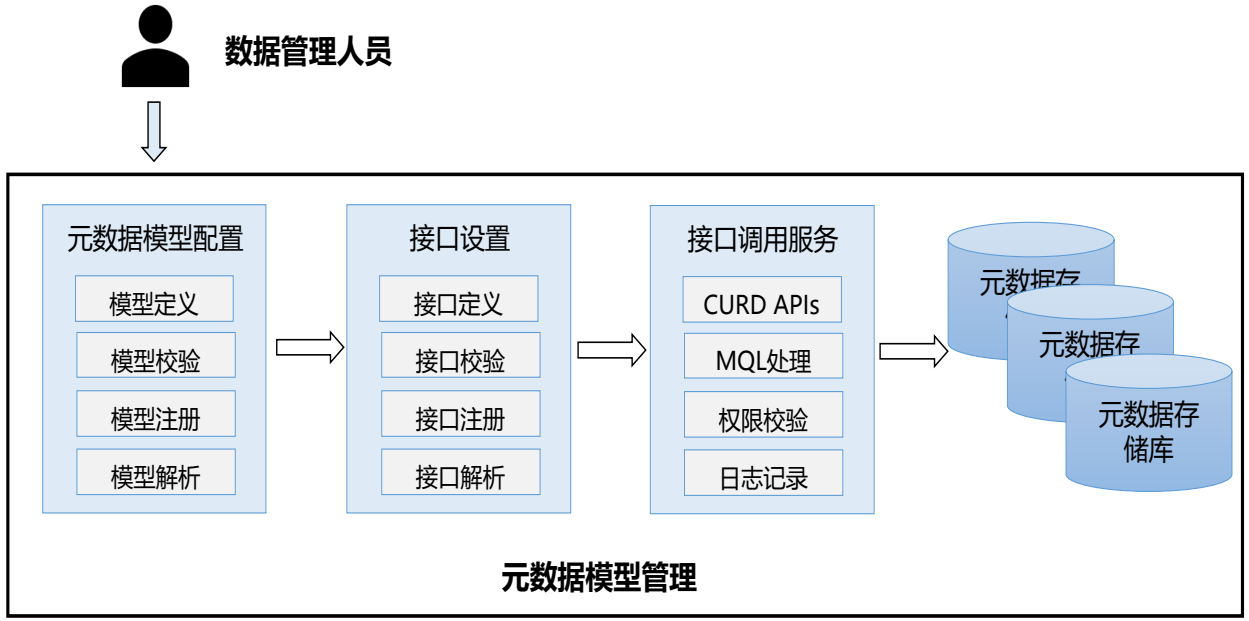
为应对高数据吞吐量，采用消息队列来解决传输任务的积压问题。同时利用多个服务器节点和多线程执行传输任务，实现集群化和多线程传输，确保传输效率。

元数据提取和目录管理

- 元数据提取：
  - 实现科学元数据和管理元数据的自动提取、解析、校验和入库
  - 基于内存数据库的缓存机制和消息队列的异步通信机制，解耦消息消费逻辑，提高消息处理效率和系统的可扩展性
- 元数据目录管理：
  - 实现装置元数据模型的灵活配置、元数据接口的规范使用和元数据的高可靠存储
  - 提供可视化界面支持元数据接口配置，实现元数据接口自动化生成，以供其他应用程序使用

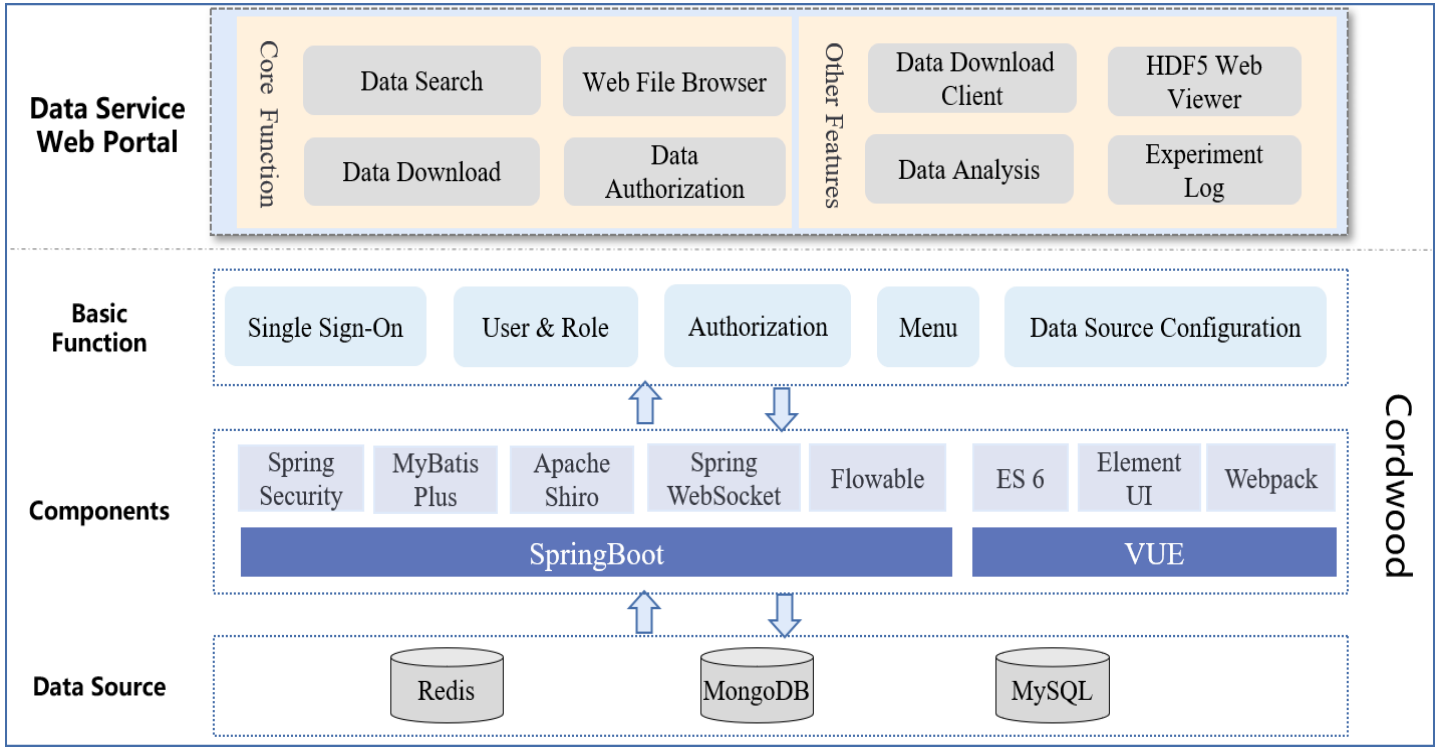


元数据提取器



数据服务

- 提供数据检索、数据在线查看、数据下载、数据授权、数据分析和实验日志查看等功能
- 满足实验用户对数据利用、分析、共享等不同维度的数据服务需求



数据服务软件架构

- 前后端分离架构
- 基于高能所低代码平台Cordwood实现后台管理功能，如用户认证、授权、角色管理等功能
- 在此基础上开发了前端页面，包括数据检索、数据下载、文件在线预览和数据授权等功能

HuHao et al., J. Synchrotron Rad. (2024)

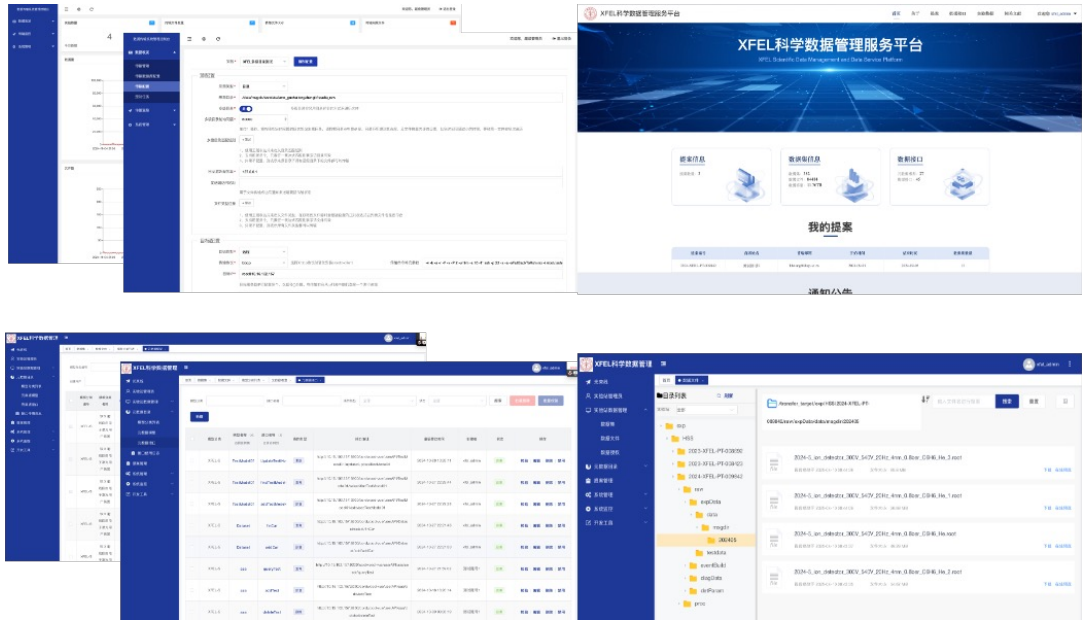
数据管理软件定制化开发和测试

定制化开发

- 支持两种与DAQ系统对接方式：Kafka消息监听 & 目录轮询
- 定义SHINE实验数据元数据模型、开发元数据提取器
- 数据传输增加bbcp 工具支持
- web UI修改，接入本地统一认证(AD、OAuth)
- 增加各模块运行状态监控和自动部署功能

diagnostic	e-beam	endstation	sample	dataset	proposal
BLM-1	AVE-POWER	FEL	DESCRIPTION	RUNID	PROPOSALID
GMD	ENERGY	ENDSTATION	NAME	RUNTYPE	PROPOSALNAME
HAMPS-1	FREQUENCY	PUMPING-DELAY	SIZE	FILELIST	PROPOSALTYPE
IMAGER-13	PULSE-LOSS	PUMPING-ENERGY	TYPE	EVENTSNUM	PROPOSALABSTRACT
PAM-2	PULSE-Q	PUMPING-SHIFT		FILESNUM	PIACCOUNTINFO
PAM-4	UNDULATOR-GAP	KB-PARAMETERS		ONLINEPATH	STATE
PAM-RESULT	UNDULATOR-VAC	DETECTOR-VOL		OFFLINEPATH	DEADLINE
		CURRENT		TAPEPATH	LOCALCONTACT
		EXTRAFIELD-VOL		SIZE	ROOTPATH
				DATASETLIFECYCLE	...

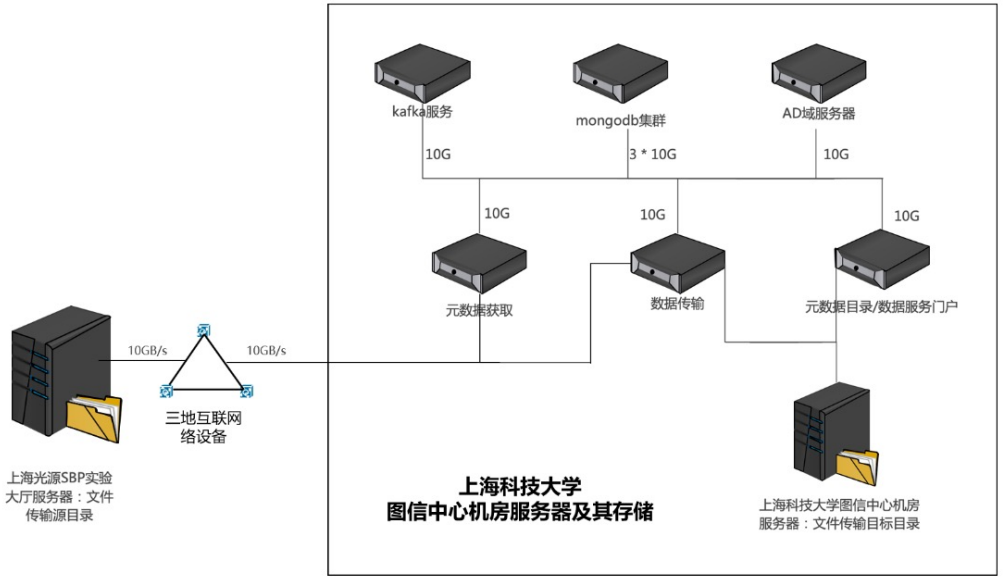
FEL-II 元数据示例



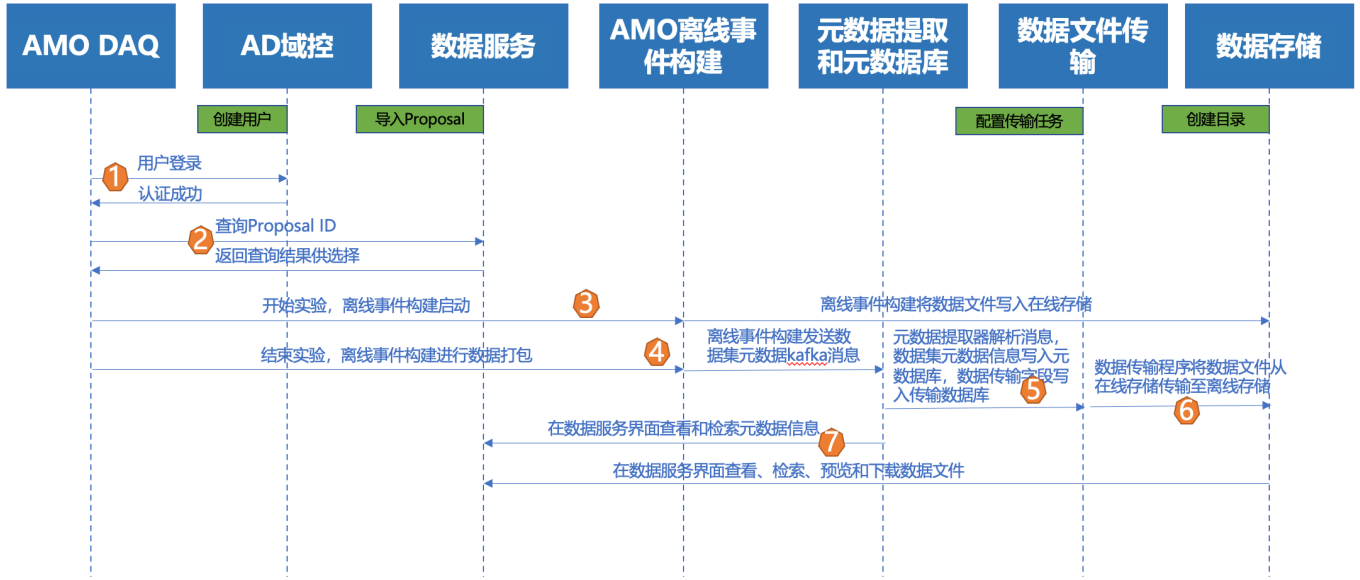
界面展示：数据传输管理、元数据模型和接口管理、数据服务门户

数据管理软件定制化开发和测试

- 于2024年10月完成上科大本地部署，并基于上海光源-蛋白质中心-上海科技大学三地互联网络对软件功能和工作流程开展测试
- 目前正在针对SHINE实验站数据特点开展本地化开发和测试

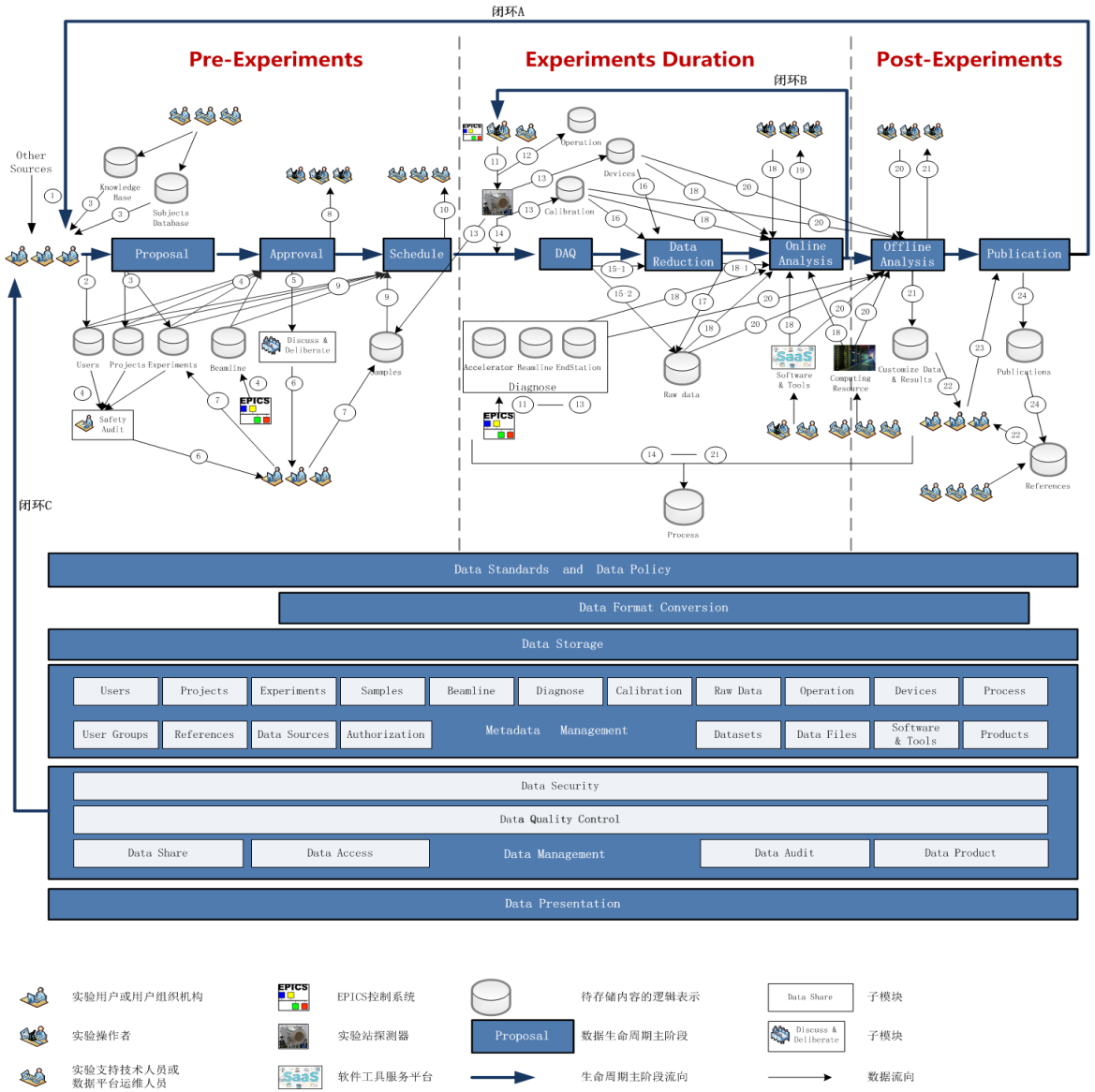


基于三地互联网络的数据管理系统软件工作流程和功能测试



AMO实验站数据管理软件与DAQ软件联动时序图

总结和展望



基于全生命周期的科学数据管理与分析系统

序号	组成	描述
1	Proposal	用户在统一认证系统注册认证后，登录门户网站，提交实验相关资料，包括项目相关信息和实验计划，并申请预约机时。
2	Approval	束线站科研技术人员对实验安全进行审计，并根据束线站运行情况和用户需求安排机时，复杂情况需多位科学家共同商议，对用户的申请进行讨论与评议。用户根据反馈结果准备样品，并在门户中提交样品的详细信息。同时安排实验操作者进行实验操作培训。
3	Schedule	束线站科研技术人员根据上一阶段的结果和用户提交的信息，确认用户机时，并安排相应的实验技术支持人员。
4	DAQ	实验开始，采集探测器输出数据。若实验不需要数据约简，则将实验原始数据保存下来，供后续在线分析。
5	Data Reduction	对于需要数据约简的实验，根据用户设置的约简比重进行数据约简，约简后的原始数据保存下来，供后续分析。
6	Online Analysis	用户根据保存下来的原始数据、装置诊断信息、探测器刻度数据以及数据集采时探测器的状态，利用软件平台和超算平台进行在线分析，得到快速反馈结果。
7	Offline Analysis	用户根据保存下来的原始数据、装置诊断信息、探测器刻度数据以及数据集采时探测器的状态，利用软件平台和超算平台进行离线分析，得到结果。
8	Publication	用户根据分析结果编写论文，投稿发表。
9	Data Standards and Data Policy	1-8 的过程中遵循相应的数据标准和政策。
10	Data Format Conversion	1-8 的过程中部分数据需要进行格式转换，保证存储下来的是平台定义的标准格式。
11	Data Storage	1-8 的过程中产生的所有数据都需要存储。
12	Metadata Management	1-8 的过程中的元数据管理。
13	Data Management	1-8 的过程中的科研数据管理，包括科研数据安全管理、科研数据质量管控、科研数据共享管理、科研数据访问控制、科研数据合规审计、科研成果产出管理。
14	Data Presentation	1-8 的过程中的数据展现。
15	闭环A	8 Publication 阶段产生的文献可以作为其他科研项目或实验的来源。
16	闭环B	根据 6 Online Analysis 在线分析结果，可能需要重复多次实验过程。
17	闭环C	根据 13 Data Management 的统计结果，会给用户或用户组织反馈。

总结和展望


工程调试与应用
软线实验数据管理软件工程调试与应用
硬线实验数据管理软件工程调试与应用
压力测试与技术优化
标准化管理
元数据标准
数据质量评估
数据安全管理
基于LLM的元数据语义检索
基于AI的数据质量校验
基于AI的数据压缩
数据分析工作流建立和算法集成
AI技术应用

感谢聆听！

雷 蕾
上海科技大学
2025年8月26日