SHINE

Data-Intensive Infrastructure for SHINE: From High-Speed DAQ to Standardized and AI-Ready Scientific Data

Ping Huai

ShanghaiTech University

Shanghai Advanced Research Institute, Chinese Academy of Sciences

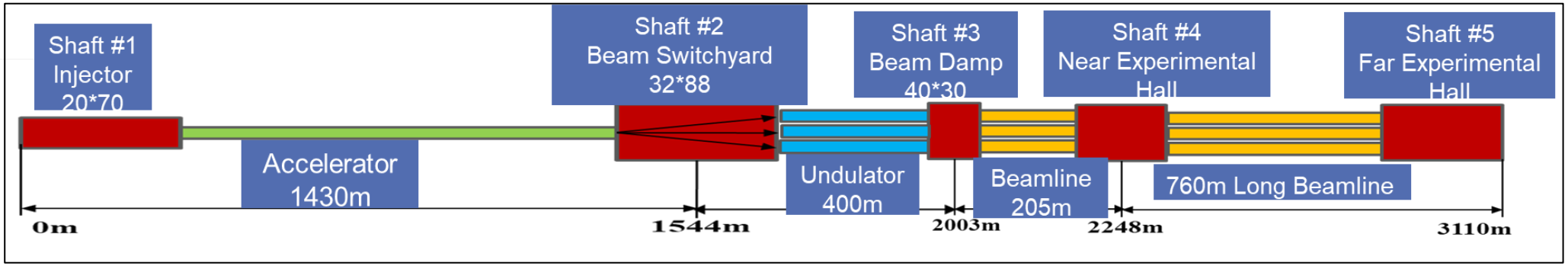
October 15, 2026

Overview of SHINE

SHINE

Shanghai HIgh repetitioN rate XFEL and Extreme light facility （2018.4 ground breaking）

- SC 8GeV linac, 2 undulator lines, photons from 0.2-15 keV
- 2 X-ray beamlines, 6 endstations, 100 PW laser facility

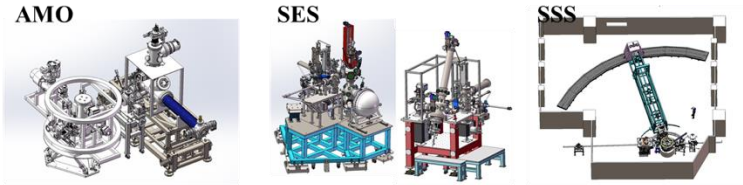


FEL-I Hard X-ray Endstation

- **HSS:** Hard X-ray Scattering and Spectroscopy Endstation
- **CDS:** Coherent Diffraction Endstation for Single Molecules and Particles
- **SEL:** Station of Extreme Light

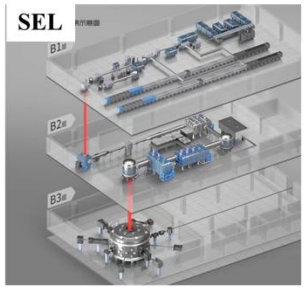
FEL-II Soft X-ray Endstation

- **AMO:** Atomic, Molecular, and Optical Science Endstation
- **SES:** Spectrometer for Electronic Structure
- **SSS:** Soft X-ray Scattering and Spectroscopy Endstation





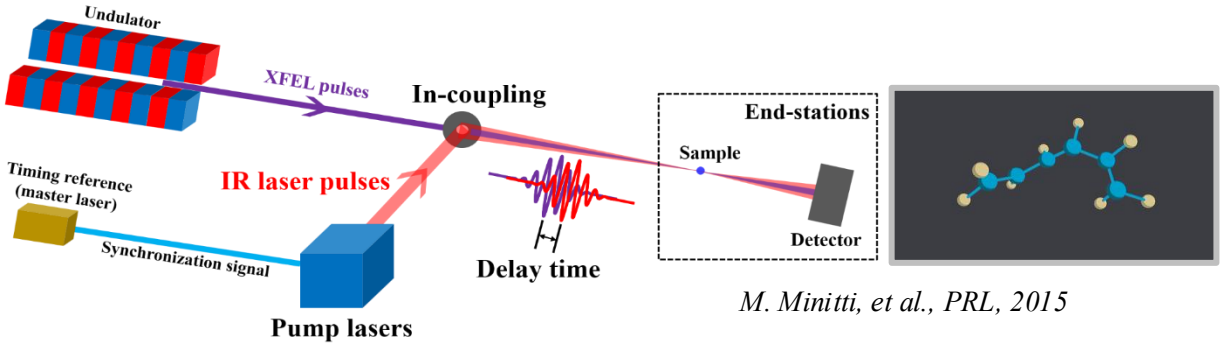




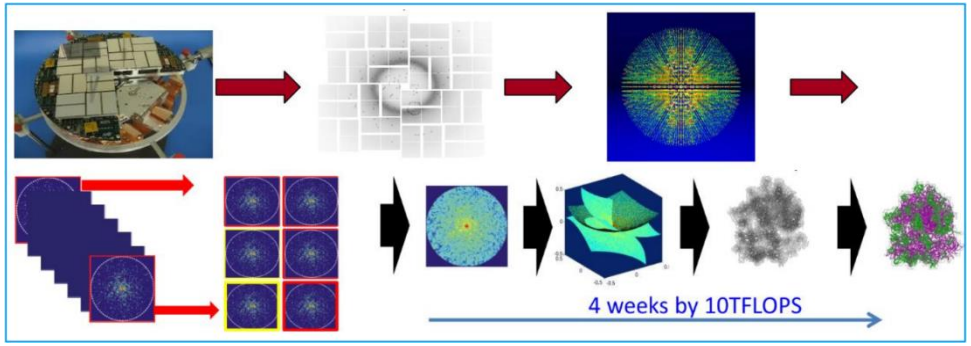
Challenge of Data Science in SHINE

As high-repetition rate XFEL, SHINE poses significant challenge to data acquisition, data storage, data analysis, data transfer, data management.

- 1MHz repetition rate
- High brightness (10^{13} ph/pulse)
- Pump-probe experiment (~10fs)
- 100GB/s peak data throughput
- 100PB/y data volume



M. Minitti, et al., PRL, 2015

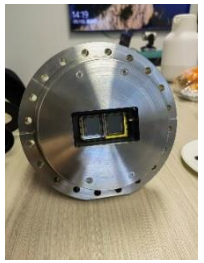


High Frame-Rate Pixel Area Detectors

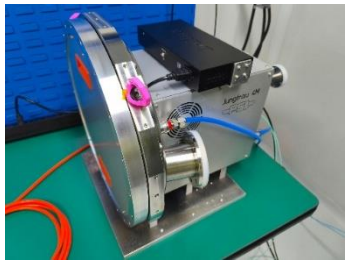
SHINE

3 × JF4M
1 × JF1M
4 × JF500K
Imaging&Crystallography

CMOS
customized dual-sensor detector for SSS



Dhyana XF400BSl-DS



JF4M



JF500K

JUNGFRAU detectors ： high dynamic range and high frame rate required for FEL-I CDS and HSS.
ML camera：FEL-I, FEL-II Photon Diagnostics

Endstation	Detector	Pixel	Frame Rate	Data Rate
CDS	Hybrid (JUNGFRAU)	4M	2.2 kHz	17.6GB/s
HSS	Hybrid (JUNGFRAU)	500K	2.2 kHz	2.2GB/s
SSS	CMOS	16M	40 Hz	0.64GB/s
Diagnostics	CMOS	1D 16k	300 kHz	9.6GB/s

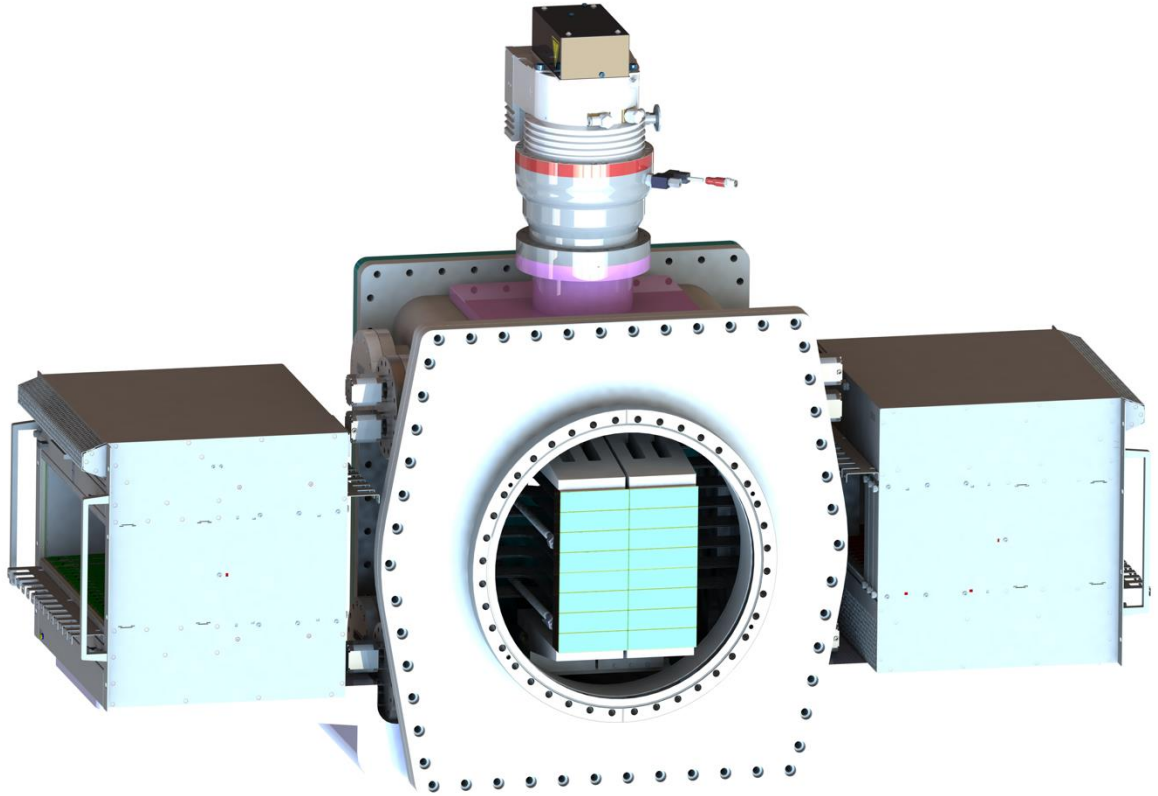
High Frame-Rate Detector R&D: STARLIGHT

SHINE

STARLIGHT : *SemiconducTor Array detectoR with Large dynamIc ranGe and cHarge inTegrating readout*

Parameters	STARLIGHT0/SL0	STARLIGHT1/SL1
Operation Mode	Charge Integrating, CW	Charge Integrating, CW
Readout Chip / ASIC	HYLITE 200F	HYLITE 100F
Frame Rate	≥ 1k Hz	≥ 10k Hz
1 photon sensitive	Yes， @ 12keV	Yes， @ 12keV
Dynamic Range	0, 1 – 10^4 ph./pixel/pulse @ 12keV	0, 1 – 10^4 ph./pixel/pulse @ 12keV
Energy Range	5keV – 25keV	5keV – 25keV
Quantum Efficiency	≥ 80% @ 8keV	≥ 80% @ 8keV
ASIC Pixel Size	≤ 200μm × 200μm	≤ 100μm × 100μm
Module Pixel Array	(64×64)×2×8 ≈ 65k	(128×128)×2×8 ≈ 256k
Data Throughput	≥ 135 MB/s	≥ 5 GB/s
Vacuum Compatible	≤ 10^{-6} mBar	≤ 10^{-6} mBar

SL0 / SL1 target specifications



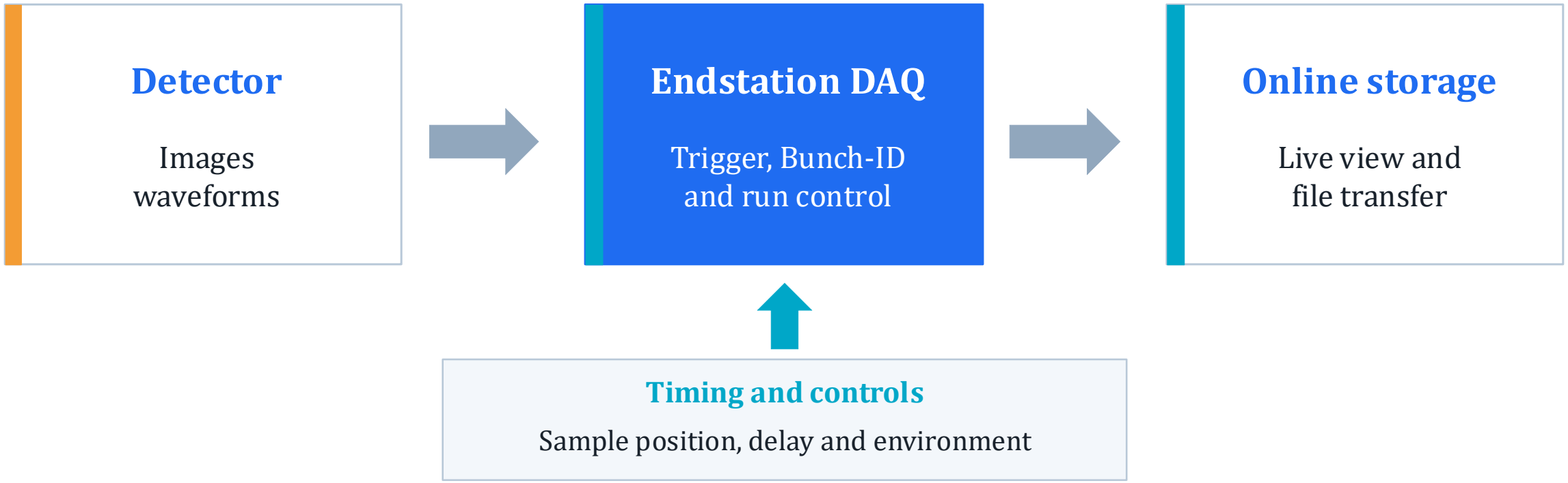
Detector concept

SL0 ≥1 kHz frame rate and ≥135 MB/s throughput

SL1 ≥10 kHz frame rate and ≥5 GB/s throughput

Endstation Data Acquisition

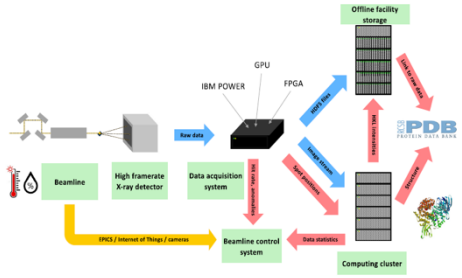
Detector frames are recorded with timing and instrument settings.

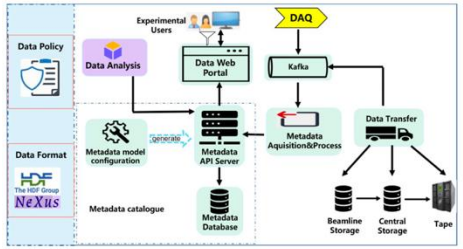


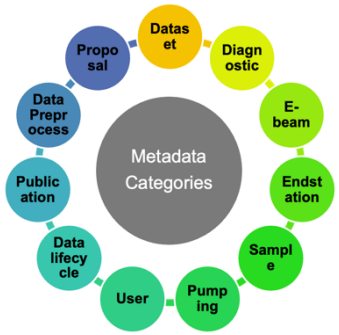
6

From DAQ to AI-Ready Scientific Data

Reliable high-speed data flow, standardized scientific data, and AI-ready data foundations.









High-Speed DAQ → **Integrated Data Infrastructure** → **Standardized Scientific Data** → **AI-Ready Scientific Data**

Capture
Timing, Bunch-ID, high-rate detectors and distributed DAQ

Preserve
Online/offline storage, archive and data lifecycle management

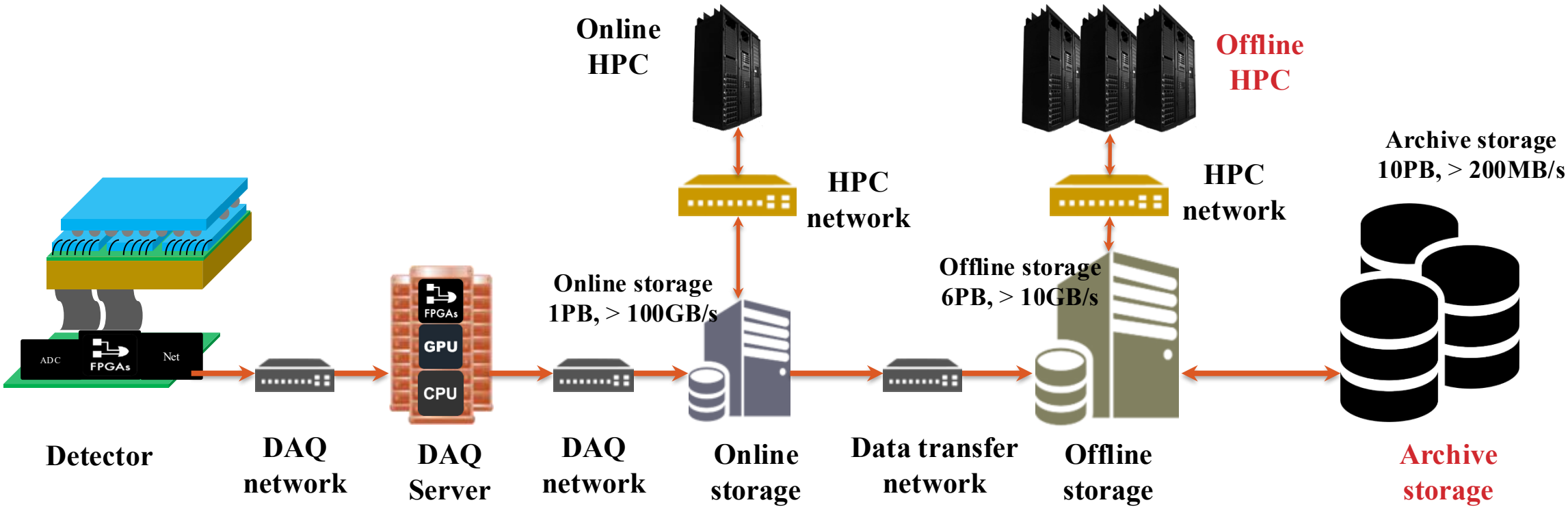
Interpret
Metadata, semantics and cross-system association

Reuse
Multimodal fusion, AI-assisted analysis and scientific agents

Integrated Data System

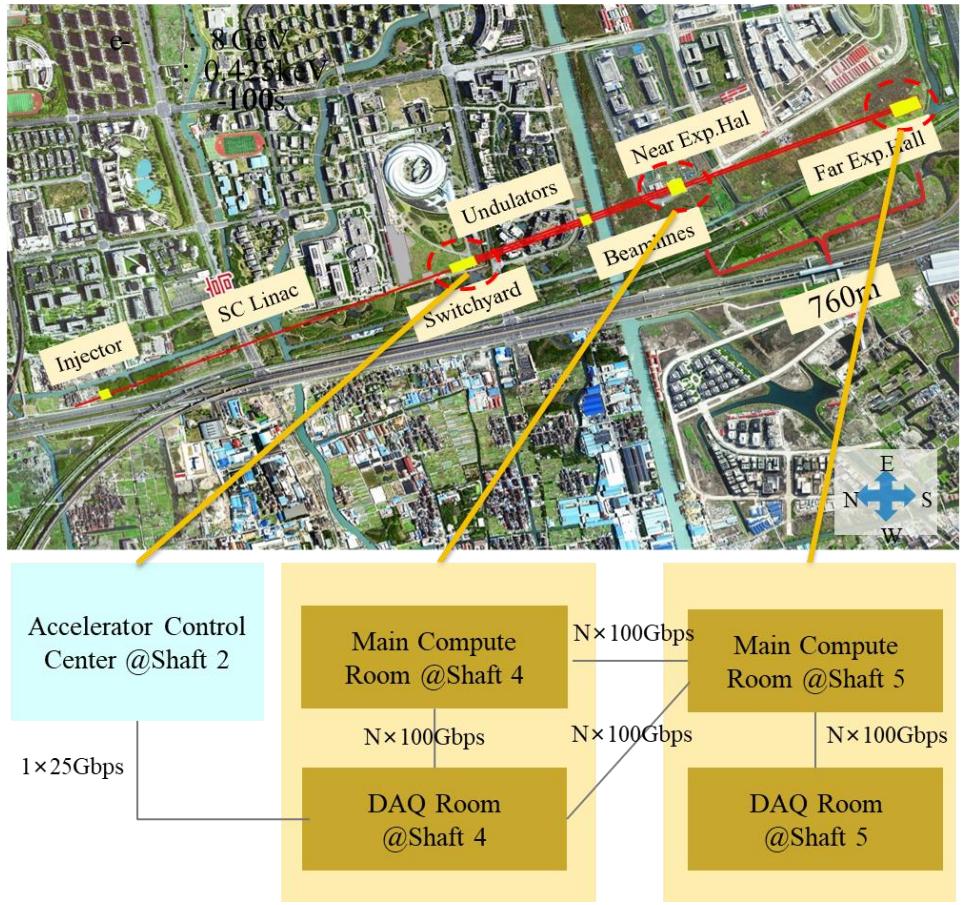
SHINE

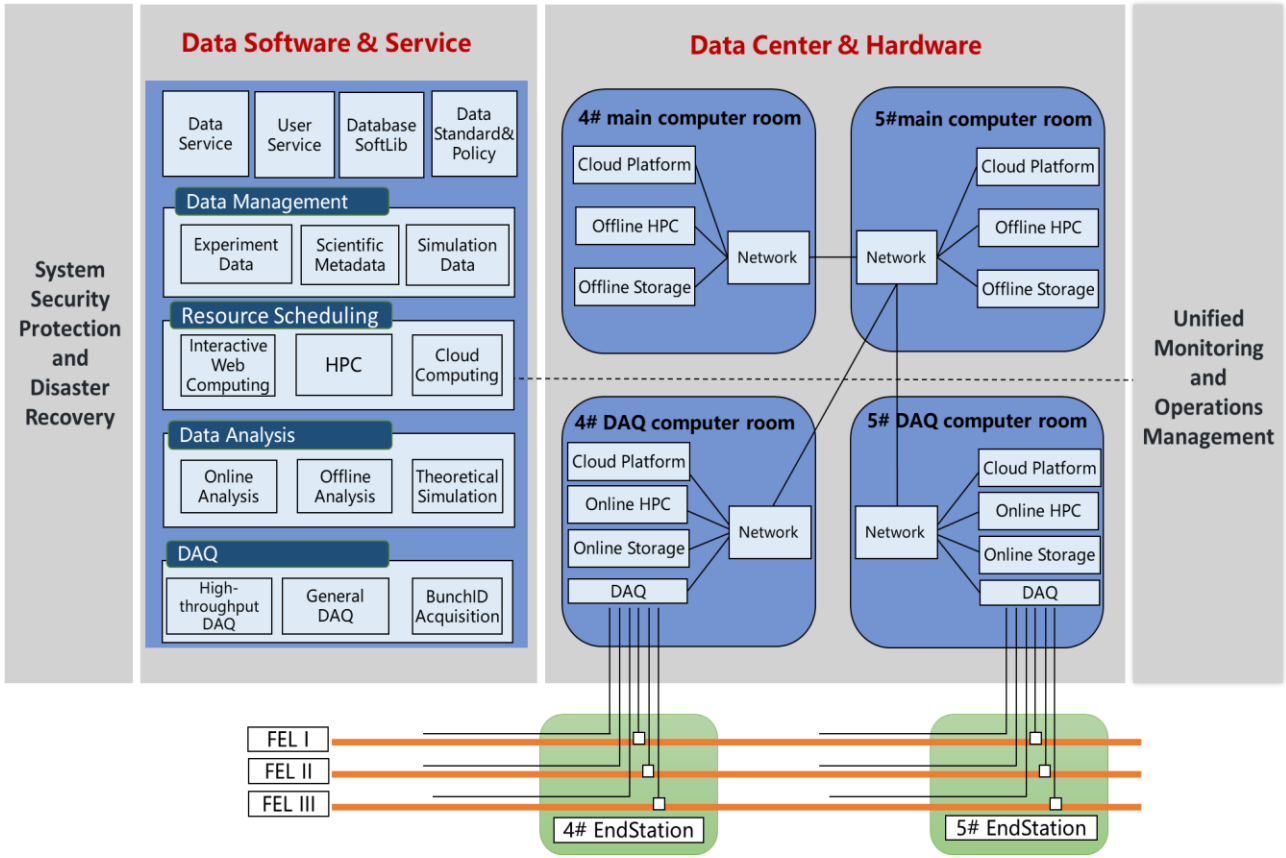
SHINE is validating the full scientific-data lifecycle from DAQ through online, offline and archive storage in stages for FEL-II commissioning.



Overview of SHINE Data Infrastructure

IT infrastructure for data acquisition, online data analysis, real-time monitoring, massive data storage and management, data archiving, and remote access.

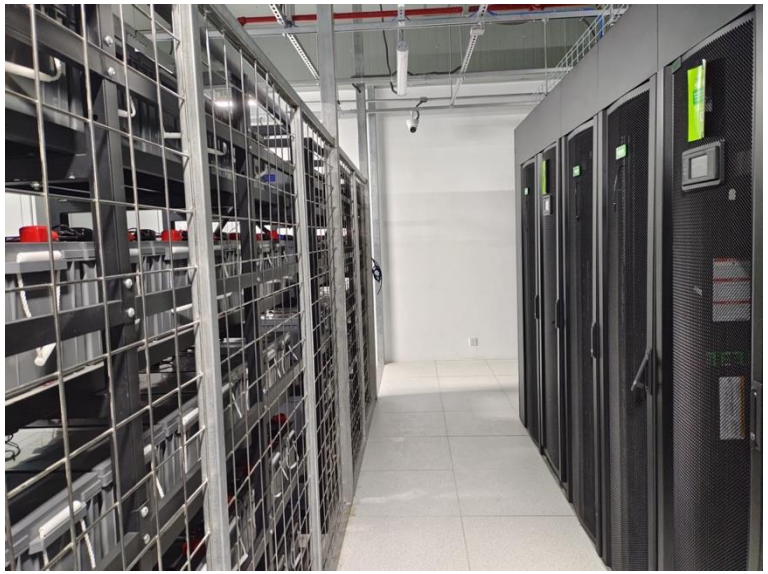




Data Center of SHINE BE in Phase-I

- **Storage**: 18 PB (2*1PB SSD, 6PB HDD, 10PB Tape)
- **CPU nodes**: 30*CPU nodes, 120 TFlops
- **GPU nodes:** 1*GPU node, 6*GPU cards, ≥100TFlops

DAQ server room at shaft 4



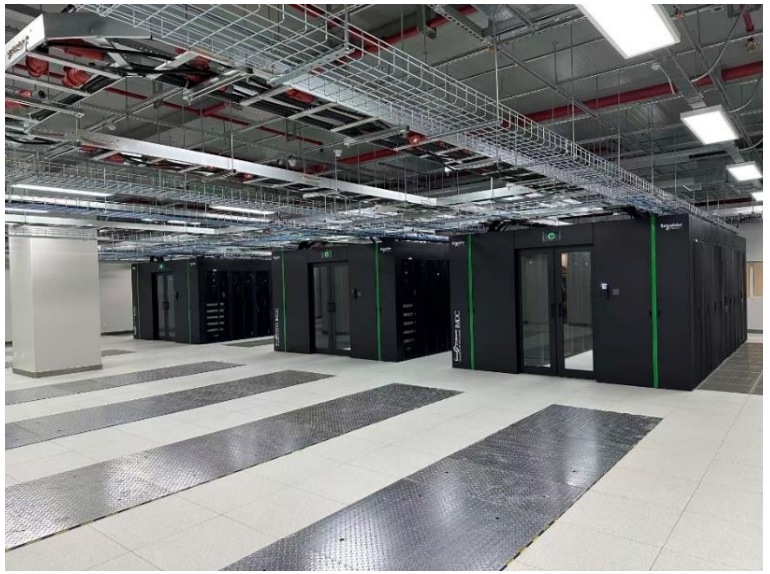
14 racks
120kW power supply

DAQ server room at shaft 5



14 racks
120kW power supply

Main server room at shaft 5



42 of 84 racks
2000kW power supply

SHINE BE Network

High Throughput

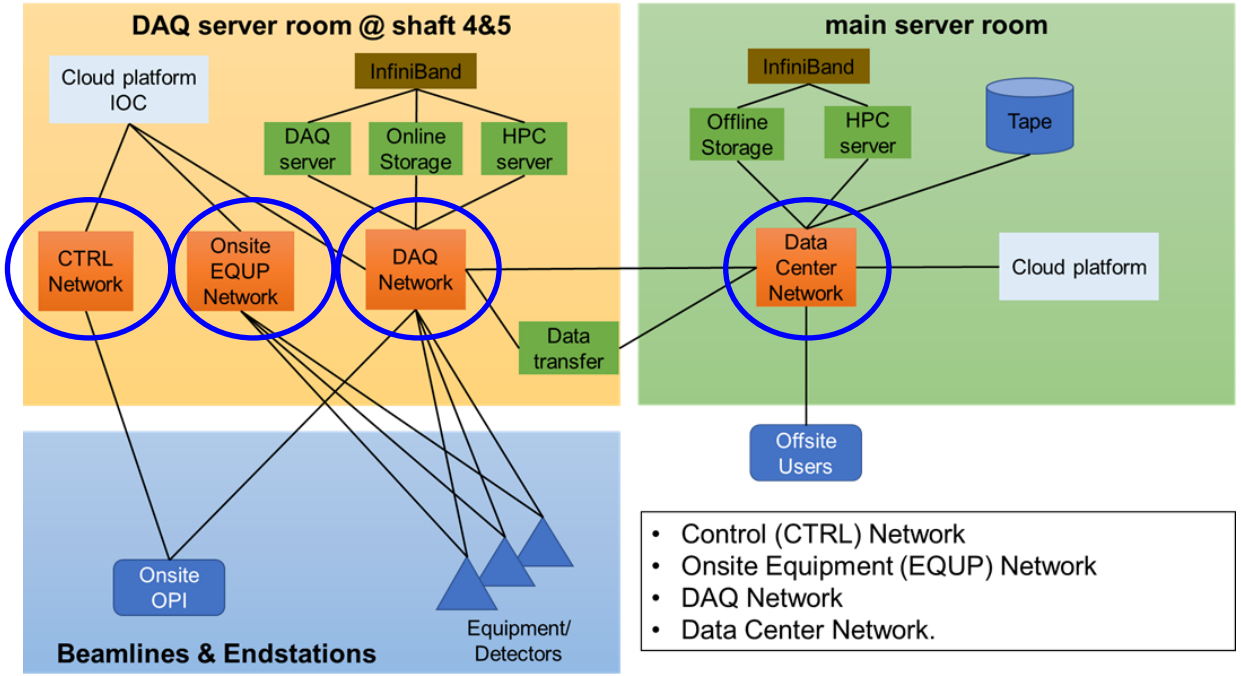
- Up to 800Gbps bandwidth between end-stations and storage
- Up to 800Gbps bandwidth between server rooms

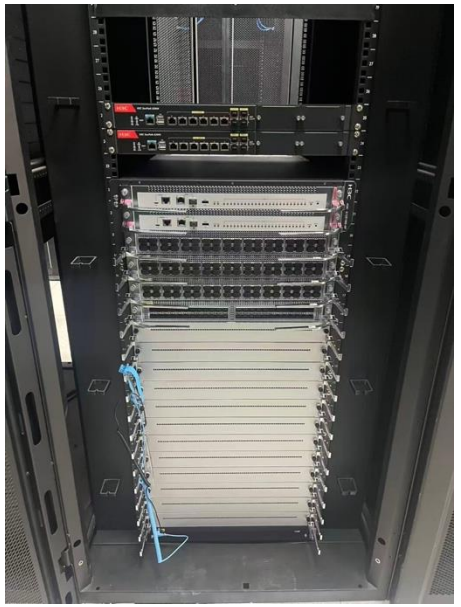
Low Latency

- average latency ≤**0.03 ms** between servers via Ether network
- average latency ≤**1.2 μs** between servers via InfinBand Network

High Availability

- Core switches use **M-LAG** for high availability

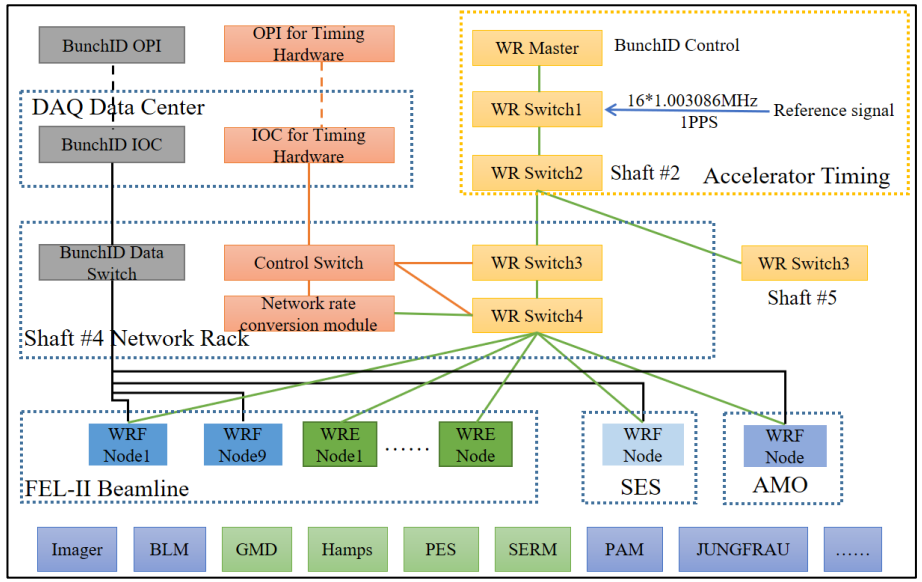






Timing System

SHINE Timing system **provides synchronization between accelerator and beamline systems**



Parameters	Value
Maximum timing jitter w.r.t clock	30ps rms
Maximum trigger rate	~1MHz
Width range, step	10ns-10ms,10ns
Delay range, step	50ps - 1s, 50ps

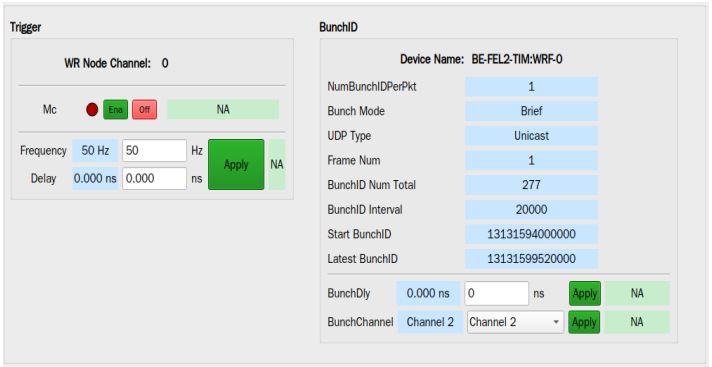




Shaft #4



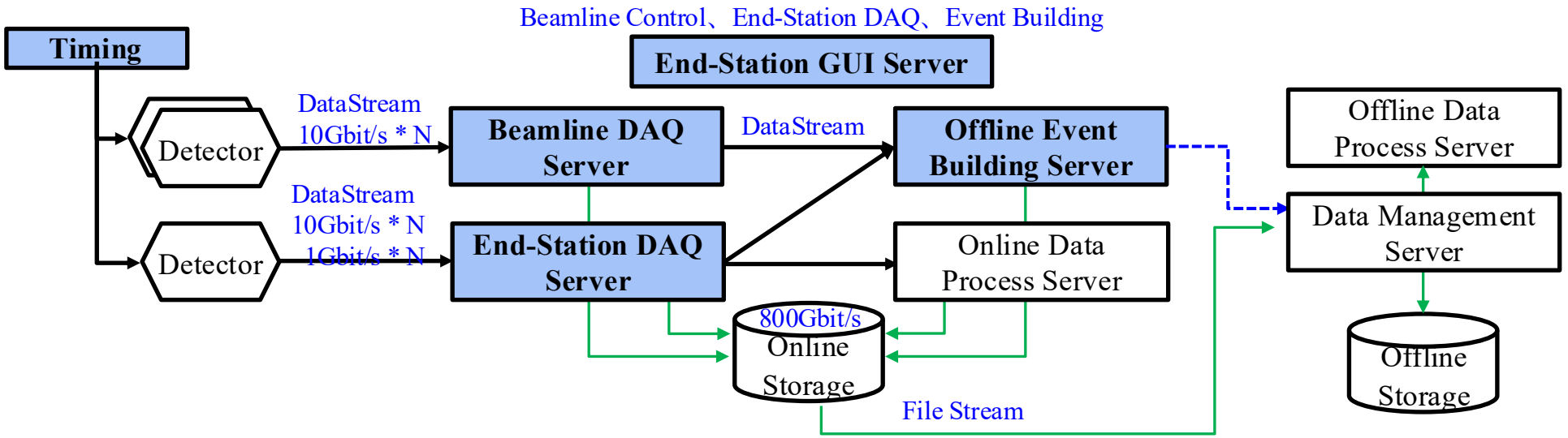
Jitter (Level 2 and Level 4) <10ps

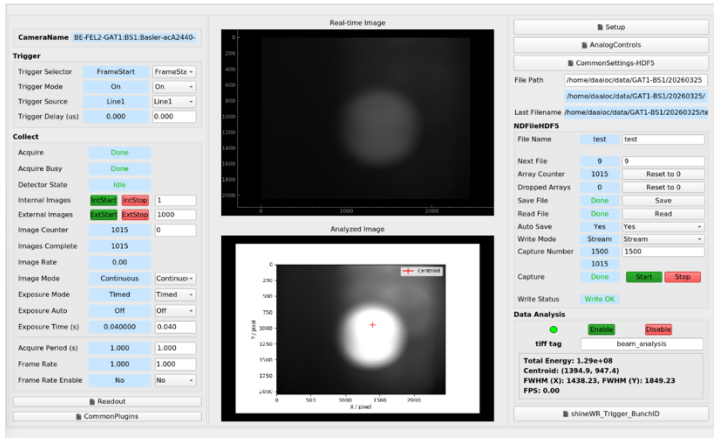


User GUI with Trigger and Bunch-ID

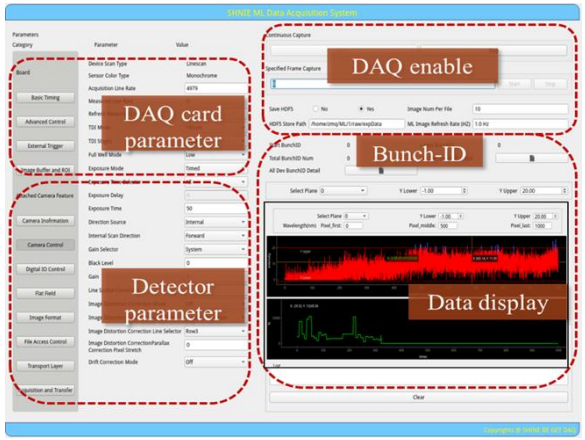
DAQ & Event-Level Data Integration

Timing and Bunch-ID align detector frames, beam diagnostics and control data into traceable events

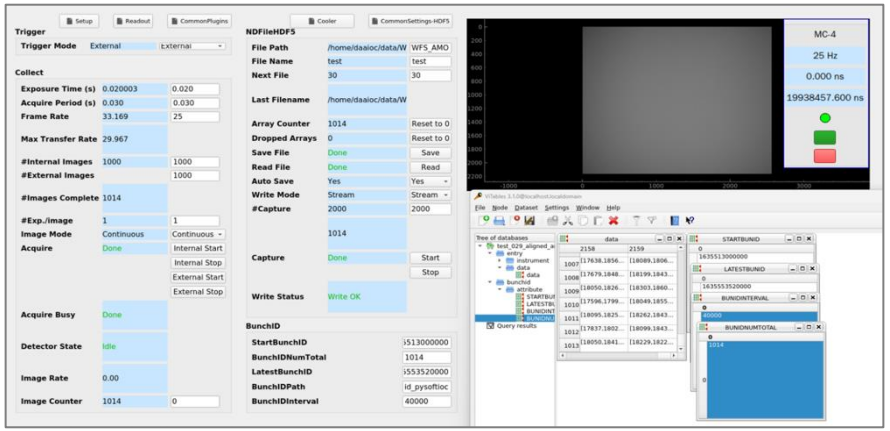




OPI of Imager



OPI of PAM



WFS DAQ GUI

From Raw Data to Standardized Scientific Data

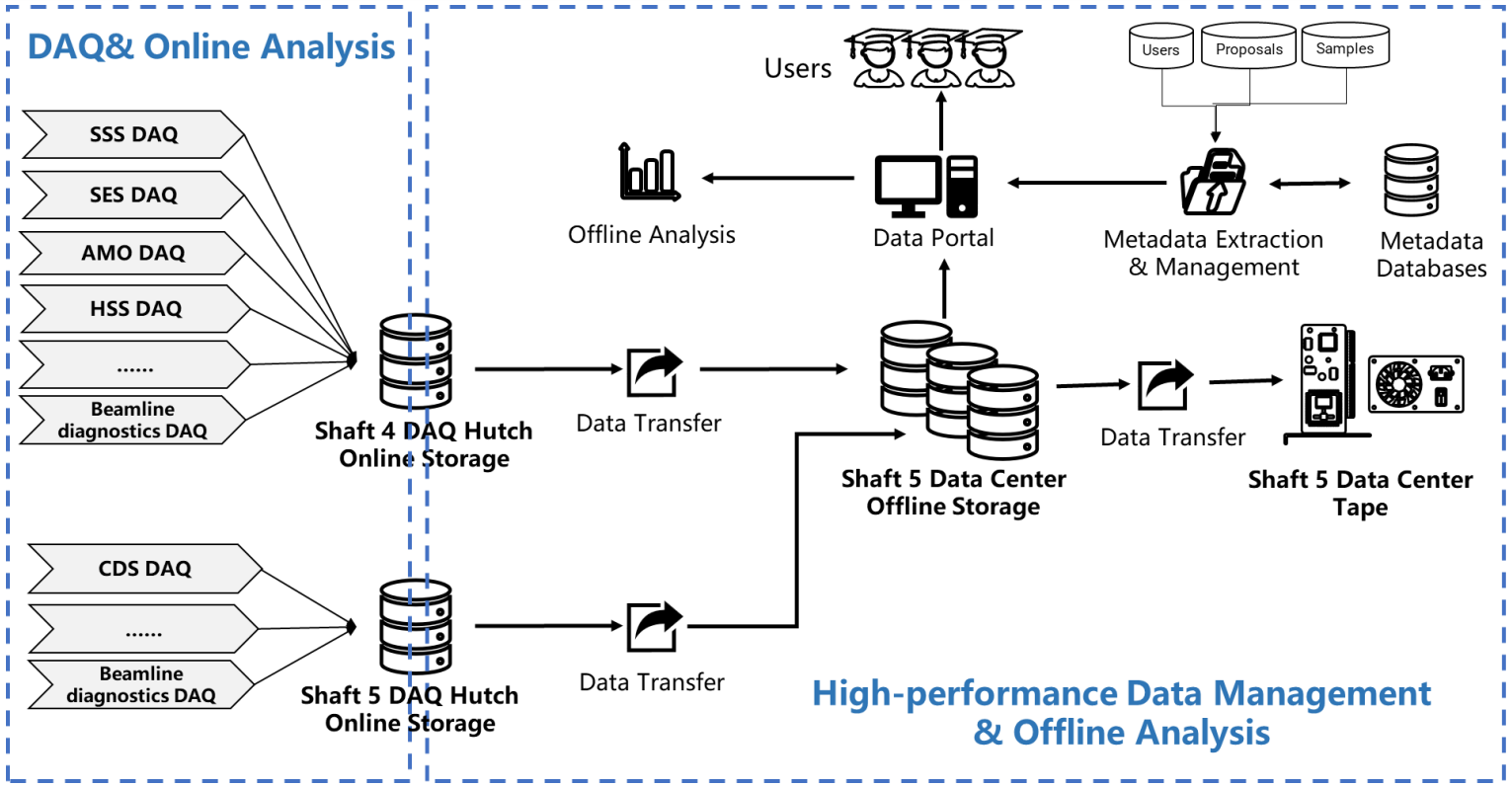
Standardization is the bridge between high-rate DAQ and AI-ready scientific data.

Standardized structures
Common data organization and formats

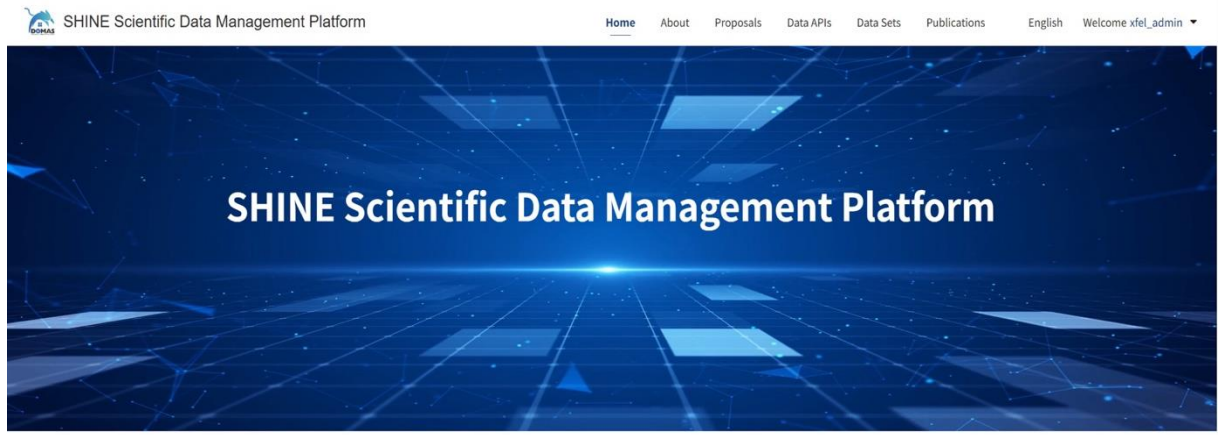
Scientific metadata
Proposal, sample, beam, detector, diagnostics and environment descriptors

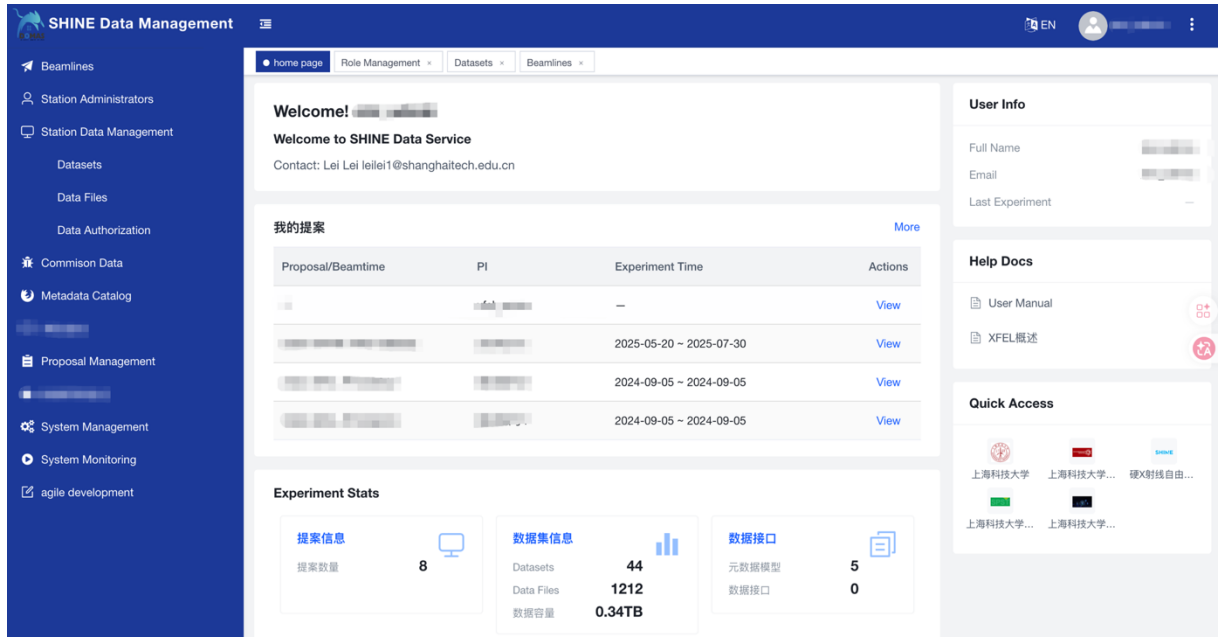
Scientific semantics
Experimental context that explains how and why data were produced

Machine interfaces
APIs and services for search, preview, retrieval and reuse



Data Service Portal





2026 New Features:

+ Beam tuning data management:
  diagnostic data, vacuum and mechanical data
+ Full UI internationalization
+ Data permission monitoring:
  GPFS permissions, authorization permissions
+ Message consumption and lag monitoring
+ MCP tool integration：
  exposing data services as MCP tools

* Based on DOMAS of IEHP

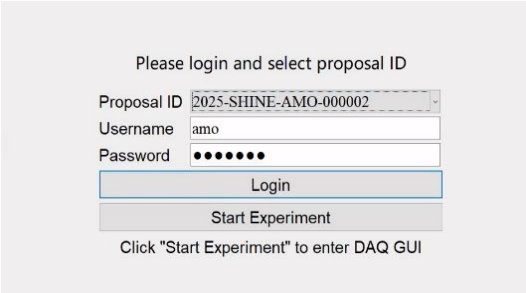
Data Policies

- Follow the **FAIR** (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) guiding principles.
- Cached data from the DAQ system will be retained in **online storage** for less than **7 days**.
- Data migrated from online storage will be preserved in **offline storage** for up to **6 months**.
- Archived raw data （possibly after reduction） will be stored for more than **10 years** to ensure long-term availability.
- All data will be made **publicly accessible** after the embargo period.

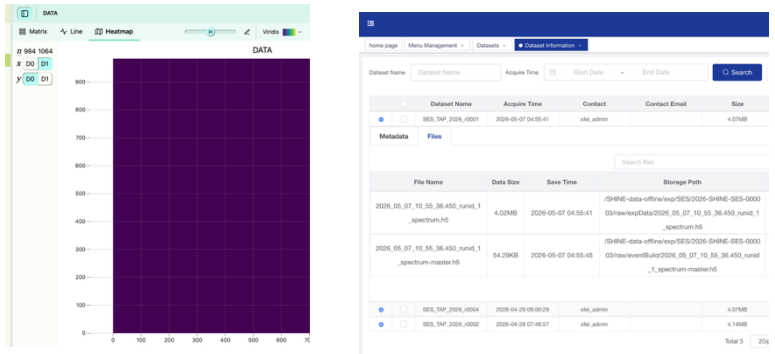
Raw data	Reserved (PB)	Day one (PB)	Duration	Storage medium
Online storage	10	2	<7days	SSD
Offline storage	100	6	6 mouths	HDD
Archiving storage	500	10	>10 years	Tape

End-to-End Workflow Validation

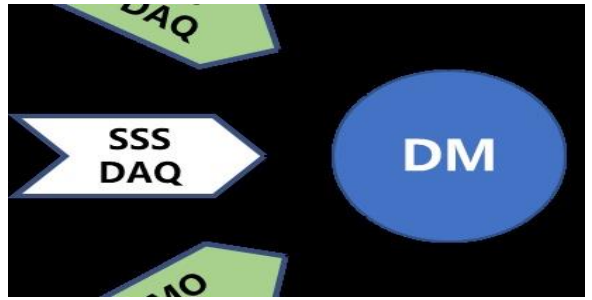
Joint commissioning of Data management with the DAQ system has been carried out in SES, and has been validated with the workflow.



experiment login / proposal selection

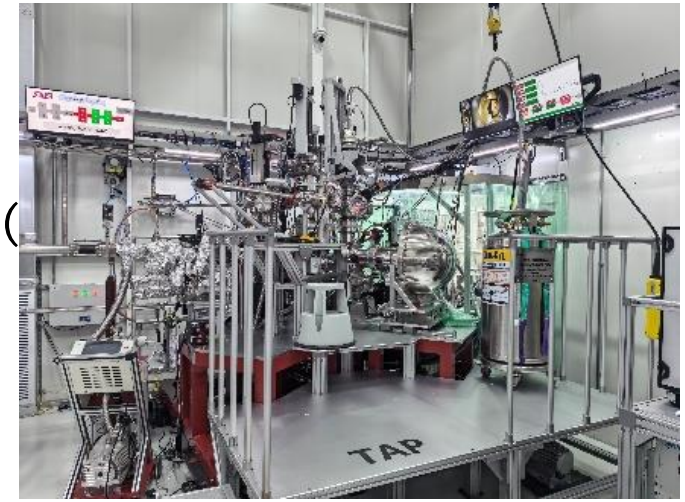


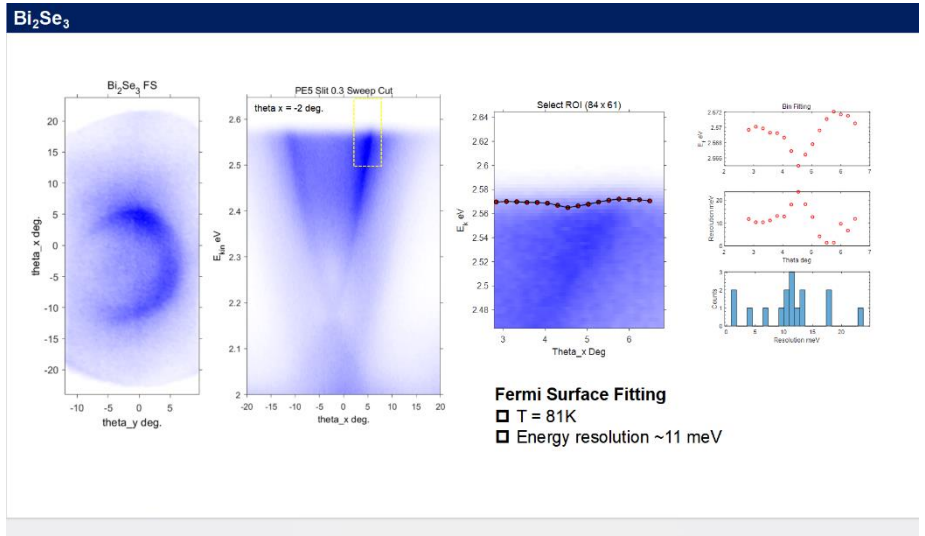
SES-TAP DAQ



data-management interface

SES endstation - Ultrafast electronic dynamics subsystem（TAP）

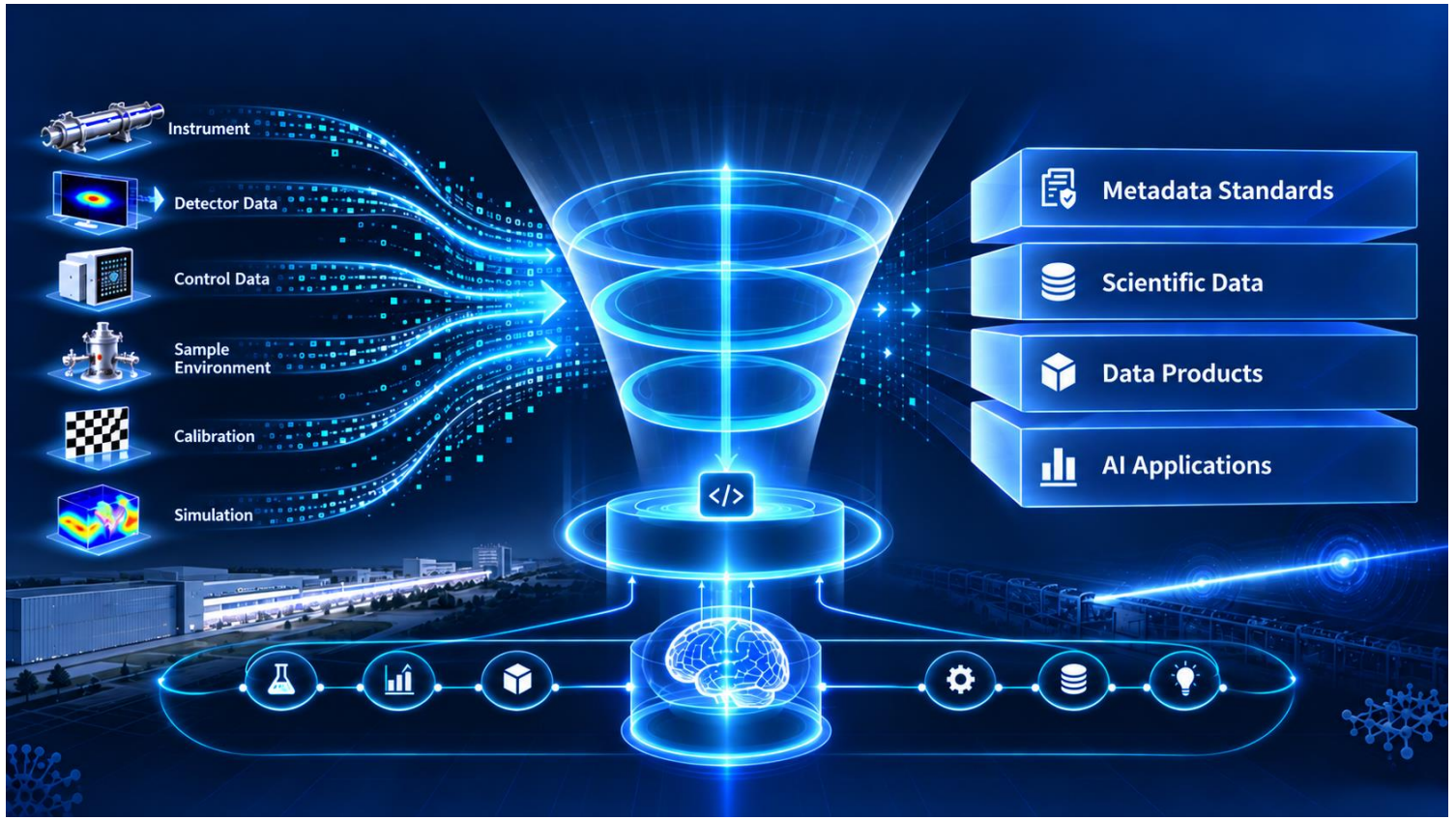




Scientific Data Association Across Systems

Hierarchical association makes each experiment traceable from the scientific objective to the single-pulse data.

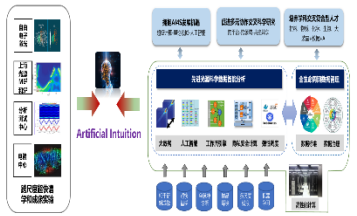
Proposal — **Run** — **Pulse / Event** — **Dataset**

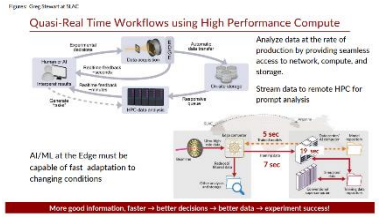


Opportunity of AI for Science in XFEL

SHINE

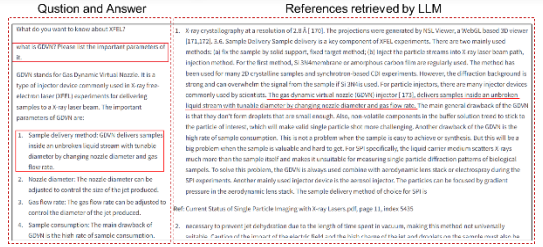
Integration of simulation, control, and experimental data, towards "artificial intuition"





Design of Experiment Based on LLM, Literature Database, and Knowledge Graph





Real-Time Analysis

Design of Experiment

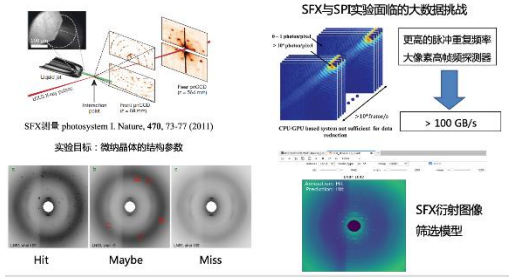
AI4XFEL

Data Processing

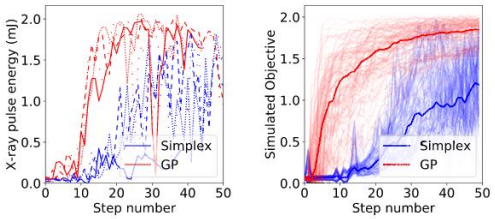
Machine Operation

Diagnostics

Data Screening & Reduction Based on Deep Learning



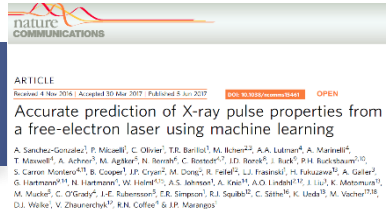
Intelligent Beam Tuning Based on Digital Twin & Deep Learning

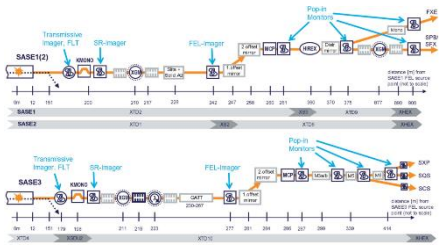


(a) Live 12 quadrupole optimization

(b) Simulated 12 quadrupole optimization

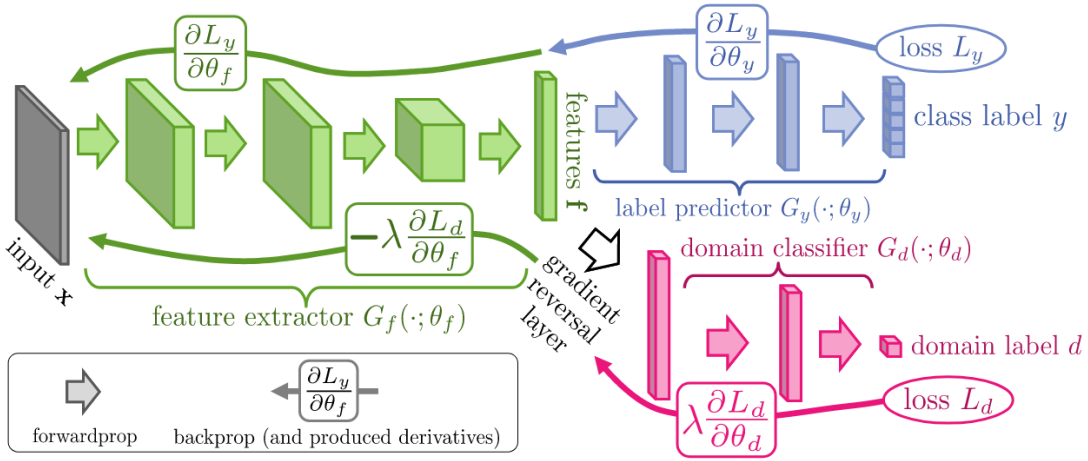
Single-Pulse Diagnostics Based on Machine Learning





AI R&D for XFEL Data

SFX DATA REDUCTION

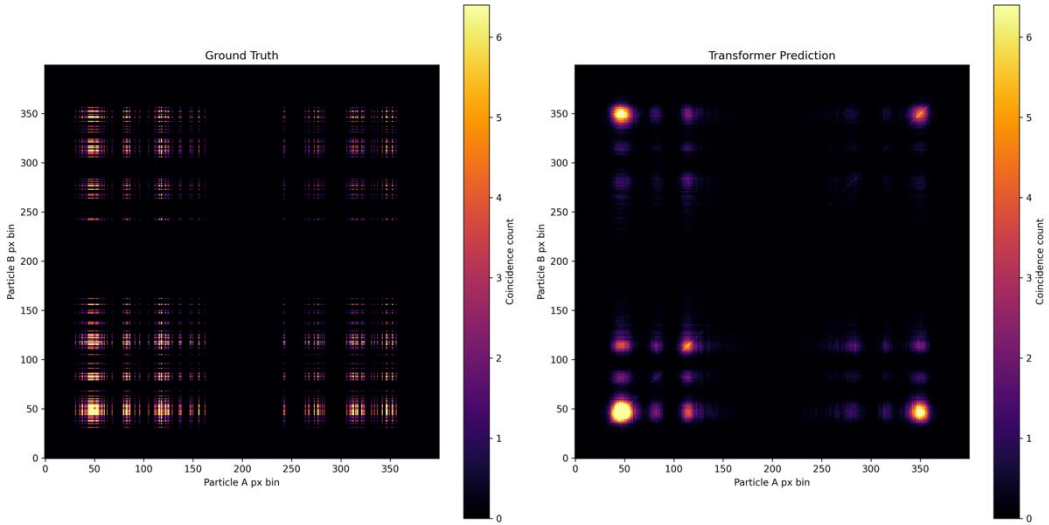


1/6 training-label volume

80× faster inference

Online reduction software has been developed for SFX data screening.

AMO EXPERIMENTAL ANALYSIS



Measured structure → **Model reconstruction**

A Transformer reproduces the main structure of the covariance map, supporting faster analysis of sparse experimental signals.

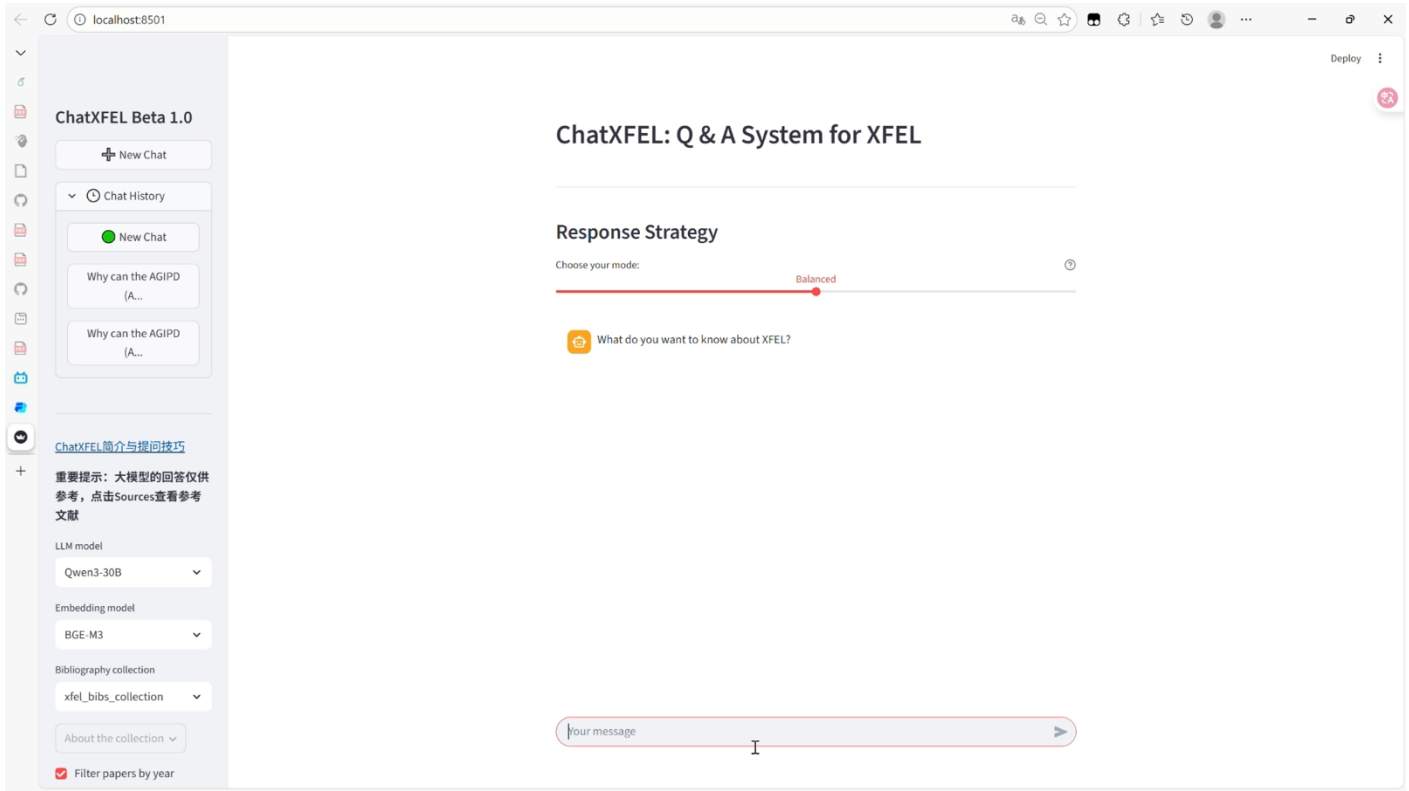
SFX result: Xie et al., Optics Express 31, 32909–32924 (2023).

AI-enabled Scientific Services

connect models with trusted knowledge, facility state and user workflows.

ChatXFEL

Retrieval-augmented Q&A prototype

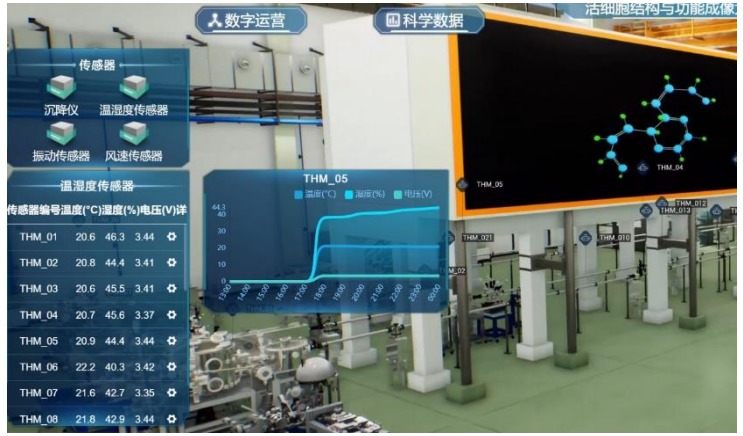


3,675 XFEL papers

+ long-form facility design reports in the retrieval corpus

Digital-twin prototype

Related facility R&D · SXFEL + SBP



- 3D facility reconstruction
- Environmental sensor streams
- Video fused with the virtual scene

Prototype status: useful integration path, not yet a routine experimental service.

Collaboration is Very Important!

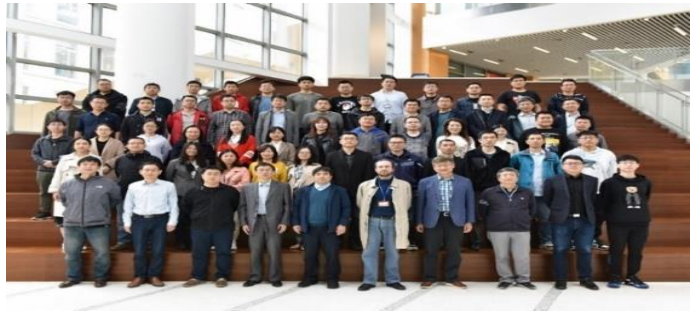
SHINE

Data and Software Workshops for advanced light source and neutron source in China

1st ShanghaiTech （2018 Shanghai）



2nd ShanghaiTech （2021 Shanghai）



3rd IHEP （2023 Beijing）



4th SARI （2024 Shanghai）



5th USTC （2025 Hefei）



6th IHEP （2026 Beijing）



Acknowledge

Collaboration with DESY, PSI, HEPS, SSRF, HALF, CSNS.

My colleagues of Department of General Experiment Technology (GET)

DAQ Group

DM Group


P. Huai Department Head


X.F. Zhang


X. Hai


Y.R. Ying


C.C. Yin Leader of DAQ


D. Jiang


W.J.Shi Leader of DM


L. Lei


W.J. Zhang


W. Xu


Y. Fang


J.T. Zheng

Summary

- The infrastructure for DAQ, storage, analysis, and data management has been established to meet the big data challenge of SHINE.
- SHINE is establishing reliable high-rate data acquisition and event-level synchronization system for commissioning, toward standardized, reusable and AI-ready scientific data.
- Standardized and interconnected scientific data provide the foundation for multimodal fusion, AI-assisted analysis and scientific agents.





Thank you for your attention!



