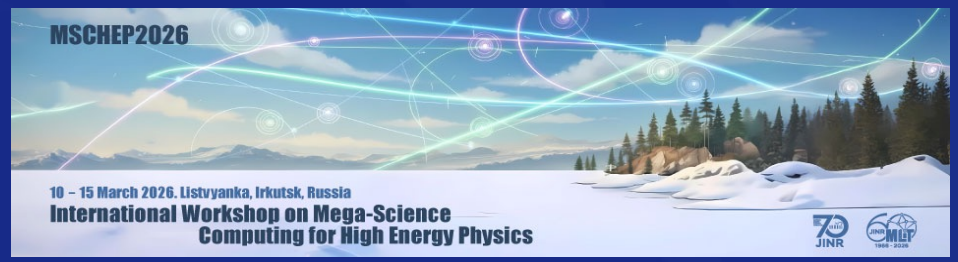
国家高能物理科学数据中心
National HEP Science Data Center

高能所計算中心
IHEP Computing Center

IHEP Data & Computing Status

Fazhi QI, Zhengde Zhang, Hao Hu, Yu Hu
qfz@ihep.ac.cn, Fazhi.Qi@cern.ch
Computing Center, IHEP
National HEP Science Data Center



Outline

1 Brief Introduction to IHEPCC

2 Data & Computing Status

3 Future Plan & Summary

Large Science Facilities in IHEP



Data & Computing Resource Requirements

Research Areas	Experiment	Needs Characteristics	Storage (per year)	Computing	Location	Status
Particle Physics	BESIII	dominated by high-throughput computing(HTC), primarily using batch jobs	1 PB	20,000 CPU Cores	Beijing	Running
	JUNO		2 PB	10,000 CPU Cores	Jiangmen, Guangdong	Running
	CEPC		10 PB	10,000 CPU Cores	TBD	Pre-research
	LHC		10 PB	5,000 CPU Cores	WLCG Tier-2 Sites	Running
Astrophysics	LHAASO		10 PB	15,000 CPU Cores	Daocheng, Sichuan	Running
	HXMT/GECAM/AliCPT		500 TB	5,000 CPU Cores	Beijing	Running
	HERD		3 PB	2,000 CPU Cores	Beijing	Pre-research
	eXTP		400 TB	2,000 CPU Cores	Beijing	Pre-research
Interdisciplinary	CSNS		1 PB	5,000 CPU Cores	Dongguan, Guangdong	Running
	SPS		>300 PB	>100,000 CPU Cores	Dongguan, Guangdong	Pre-research
	HEPS		>300 PB	>100,000 CPU Cores	Huairou, Beijing	Running
Theoretical Physics	Lattice QCD	dominated by high performance computing (HPC), primarily utilizing multi-core parallelization and GPUs	1 PB	>1P Flops (double)	Beijing	Running
Accelerator Design	Accelerator Design		1 PB	>1P Flops (double)	Beijing	Running

Resources also be required by International Cooperation

- Play an active role in many international cooperative projects:

  ATLAS, CMS, LHCb, AMS, Belle-II, PANDA, EXO, ILC, etc.

- Host many international large science projects:

  BESIII, JUNO, LHAASO, CEPC, HERD, eXTP, etc.























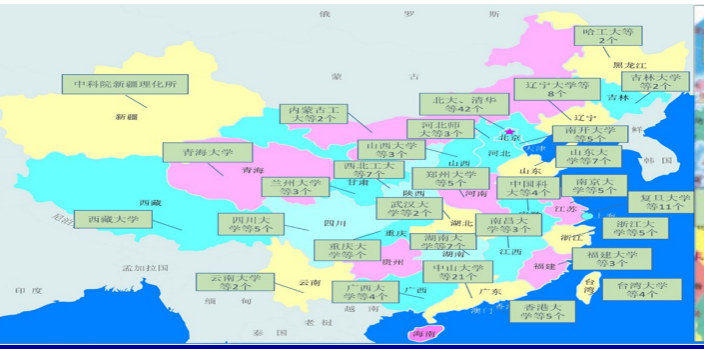




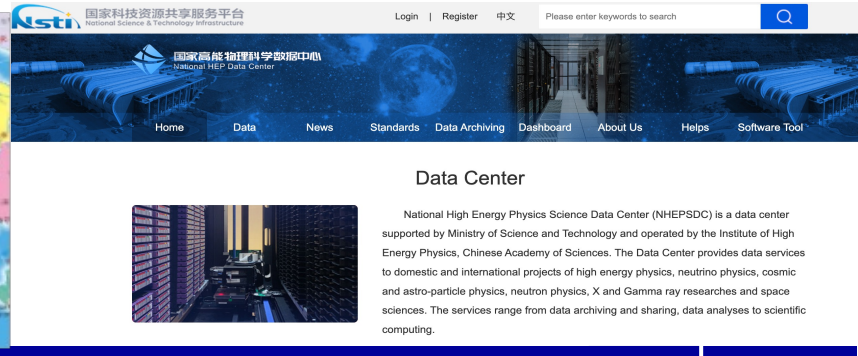
National HEP Science Data Center(NHEPSDC)

- Funded by the MoST
  - Constructed and operated by the Computing Center, IHEP
- The platform of science data curation and sharing for HEP
  - Development of technologies and policies for long-term preservation, open access, security, …
  - Collaboration with NADC (National Astronomical Data Center) and NSSDC (National Space Science Data Center) to facilitate multimodal data fusion.

🏆 **Acknowledged by MoST as the top-level one of the national data centers**







Missions for IHEP-CC

- **Provide Data and Computing Services** for HEP experiments
  - Facilities
  - Network & security
  - Storage & Computing
  - Data, AI & Scientific Software
- Launch **Research & Development in Information Technologies** related to HEP
  - High Performance Network, Computing and Storage technologies
  - Scientific Software for Advanced Light Sources
  - IT systems for advanced information processing

Team & Research

- Human resources of IHEP-CC (~110 member)
  - ~60 FTEs, 20 temporary staff, 6 post-docs, 30 Master's and Ph.D. students.
- Research fields & Missions

Computing and Storage	Network and Cyber security	IT Services	Scientific Software	Innovation
• High Performance Computing • High Throughput Computing • Grid/Cloud computing • Distributed storage	• Datacenter and campus network • Dedicated link for remote experiments • International network collaboration	• Database technology and application • Conferencing Technology • Institutional management tool	• Open data and open science • Scientific data management • Scientific software framework	• Big data • AI for science • Quantum computing

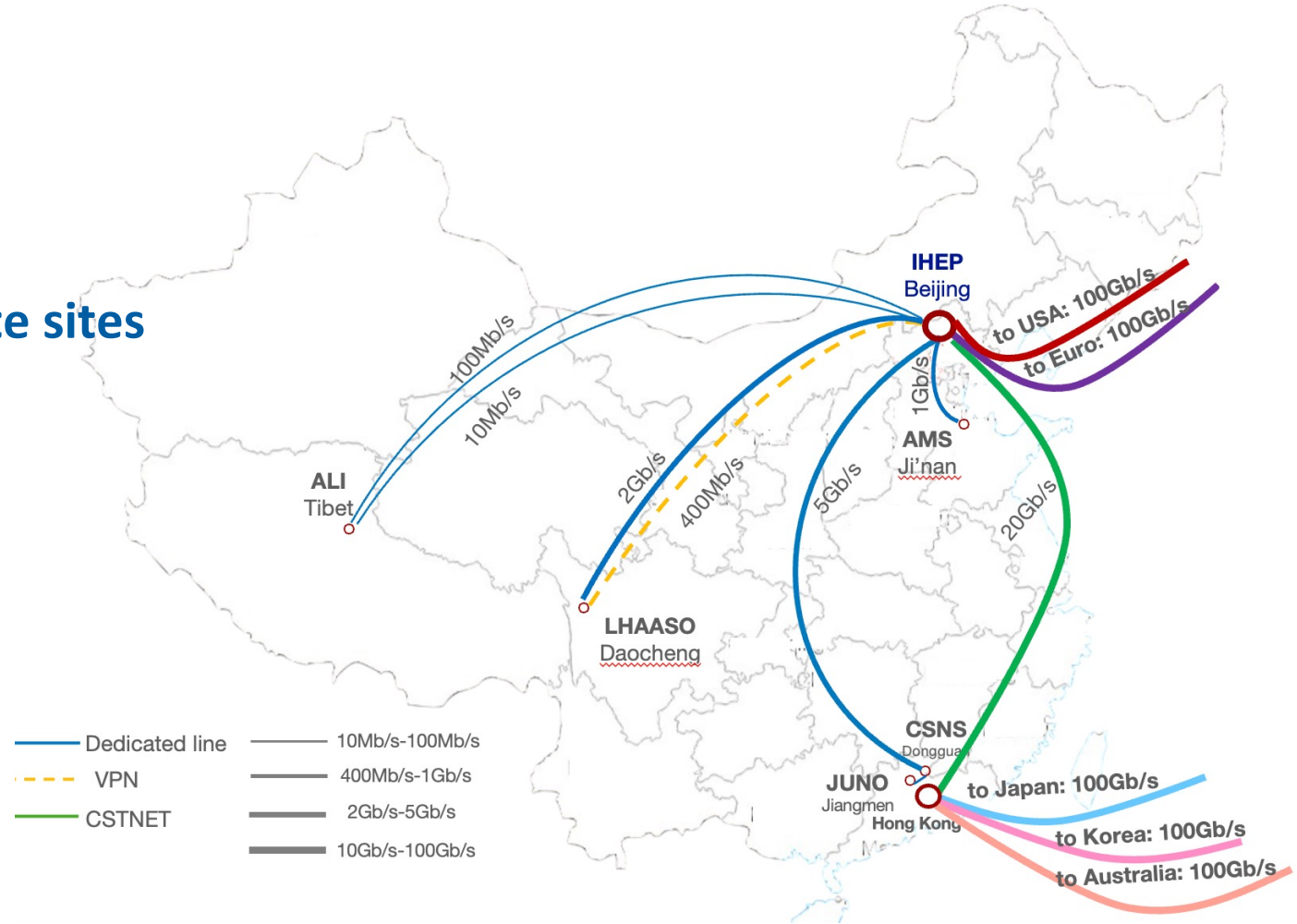
One Platform & Multi Centers



Through the High Energy Physics Open Connections (HEPOC) network and Open Data Middleware (ODM), built a high-energy physics data service platform integrating scientific data, heterogeneous computing power, and scientific software. The platform provides service capabilities of over 100,000 CPU cores, hundreds of GPU cards, over 100 PB of disk storage, and nearly 200 PB of tape storage.

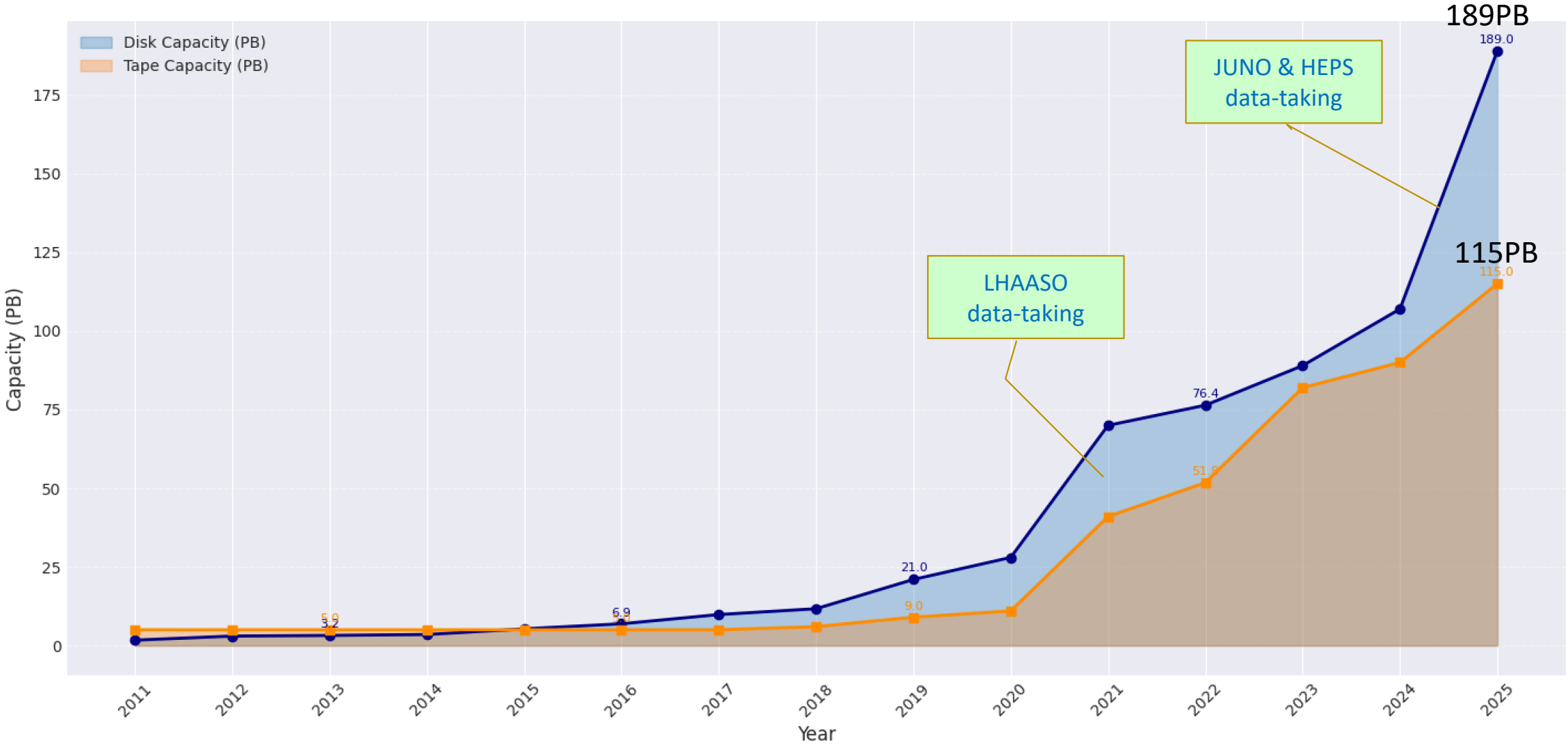
Network connections

- **Internet connections**
  - LHCONE from 2019, LHCOPN from 2023
  - 100Gbs to Internet
- **Dedicated links between IHEP and remote sites**
  - HEPS-IHEP: 100 Gb/s
  - CSNS-IHEP: 20 Gb/s
  - LHAASO-IHEP: 2 Gb/s
  - JUNO-IHEP: 155 Mb/s (2 Gb/s in 2024)
  - AMS-IHEP: 1 Gb/s
  - ALi-IHEP: 110 Mb/s

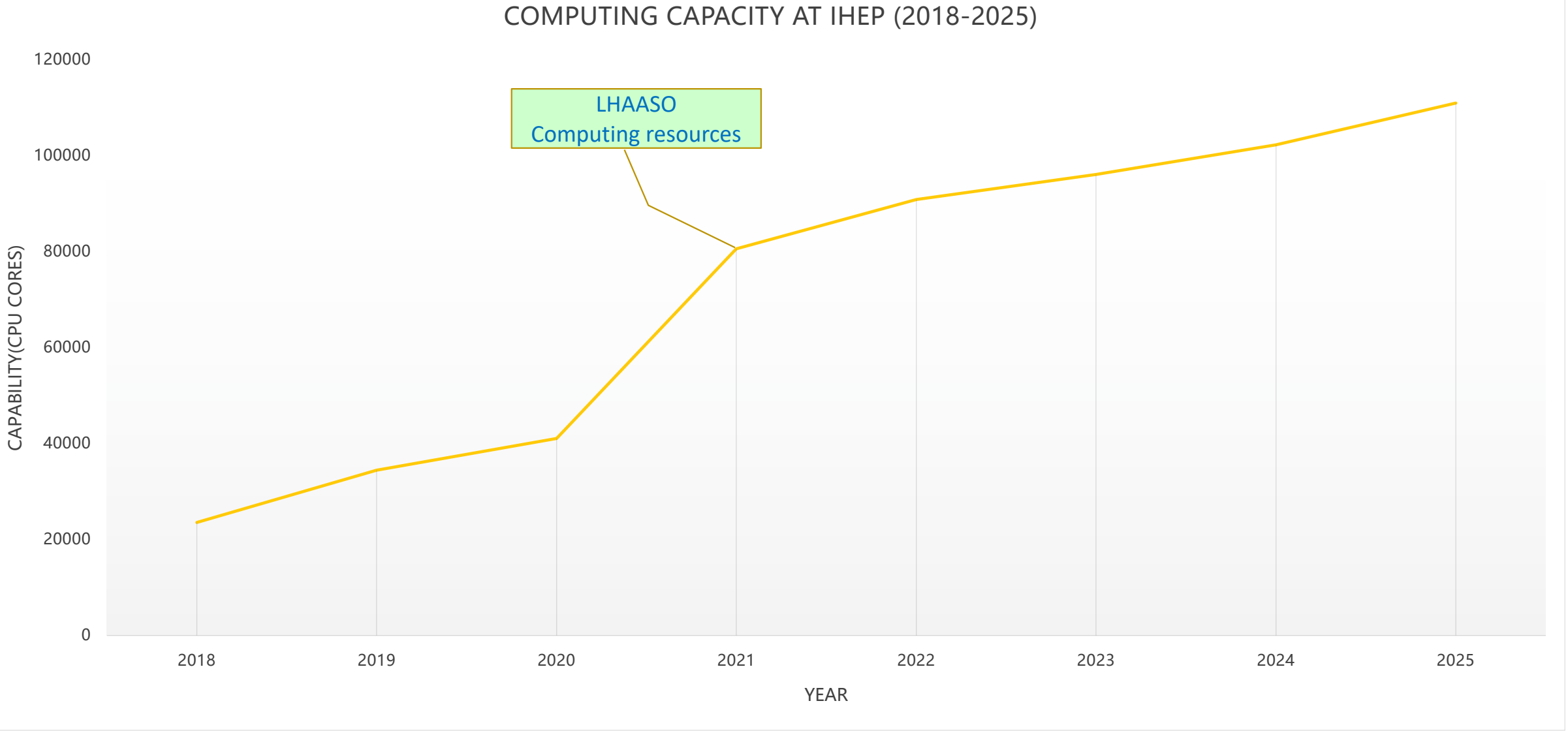


- **The best performance international network for science in China**

Data

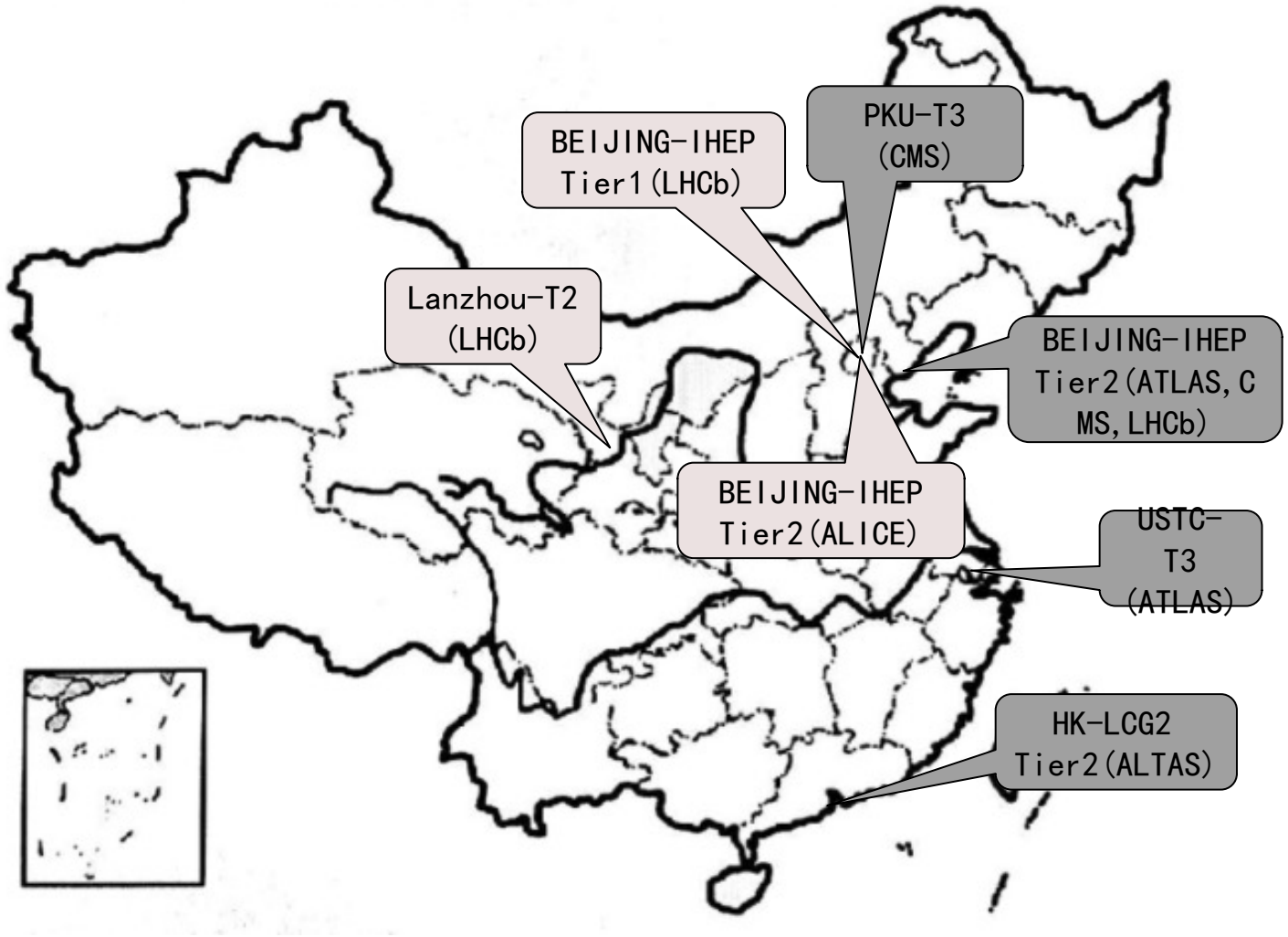

Disk Capacity (PB)
Tape Capacity (PB)
Capacity (PB)
Year
0
25
50
75
100
125
150
175
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
189PB
189.0
115PB
115.0
JUNO & HEPS
data-taking
LHAASO
data-taking
76.4
51.6
21.0
9.0
6.9
5.9
3.2

Computing


COMPUTING CAPACITY AT IHEP (2018-2025)
120000
100000
80000
60000
40000
20000
0
CAPABILITY(CPU CORES)
LHAASO
Computing resources
2018
2019
2020
2021
2022
2023
2024
2025
YEAR

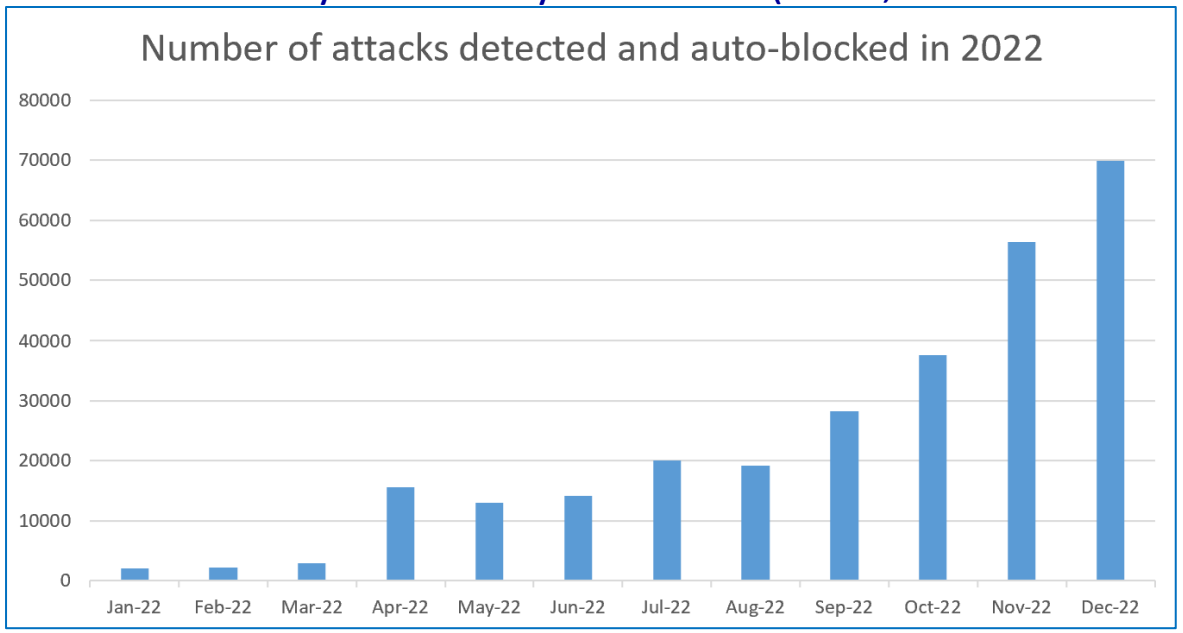
WLCG Sites in China

- Tier-2 sites
  - BEIJING-IHEP (ATLAS, CMS, ALICE)
  - Tier-2: LZU-T2 (LHCb)
  - HK-CUHK(ATLAS)
- Tier-1 Site
  - BEIJING-IHEP (LHCb)
  - 3200CPU Cores
  - 12.5 PB Disk
  - 22 PB Tape



Cyber-security Assurance

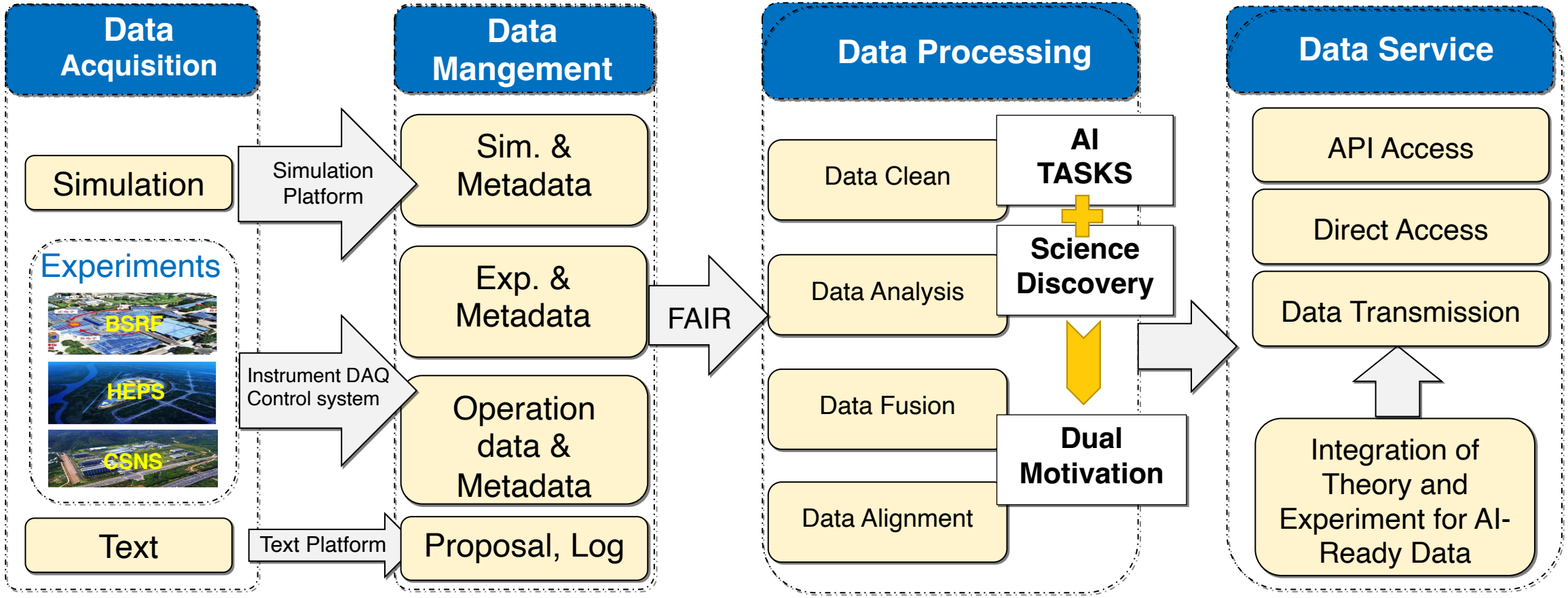
- Self-developed and deployed Security Operation Center (SOC), which provides:
  - Threat discovery → warning → response → defense automatically
  - Guarantee the network, data and network security of IHEP
  - Provide the intelligence and knowledge sharing with others
- IHEP SOC has also been promoted for deployment in other large scientific facilities and scientific data centers in China
- Asia-Pacific cybersecurity federation(RING, IHEP-KEK-ASGC) and cybersecurity cooperation with CERN





Data System

- The architecture design of data system (from the perspective of full lifecycle of data)



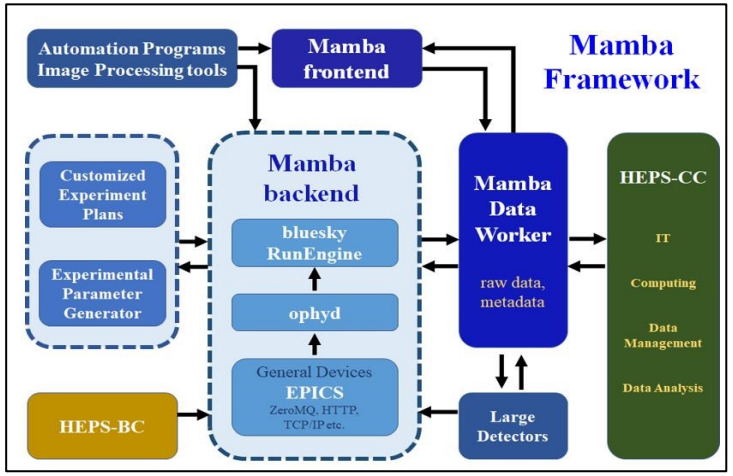
Software & Tools
Data Strategy, Metadata Standards, Data Format

Data Software

- Foundational framework/software covering the full lifecycle of data from generation to utilization.

Experimental data acquisition and processing

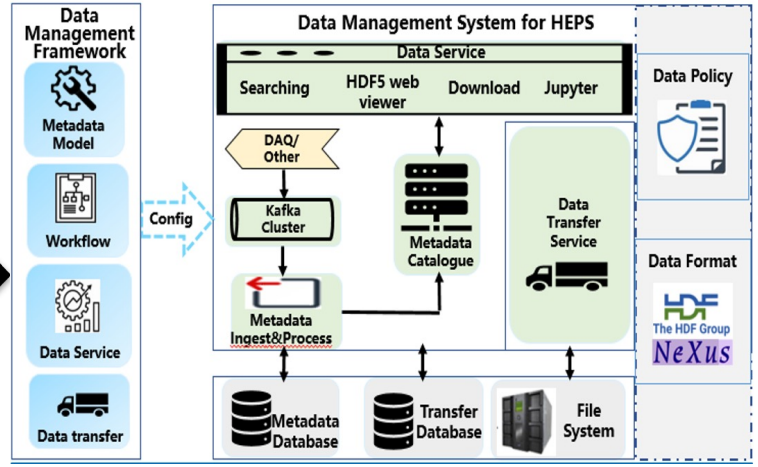
Mamba Framework



- Complex experimental process control
- E-scale intelligent big data acquisition
- Efficient online analysis

Data storage and management

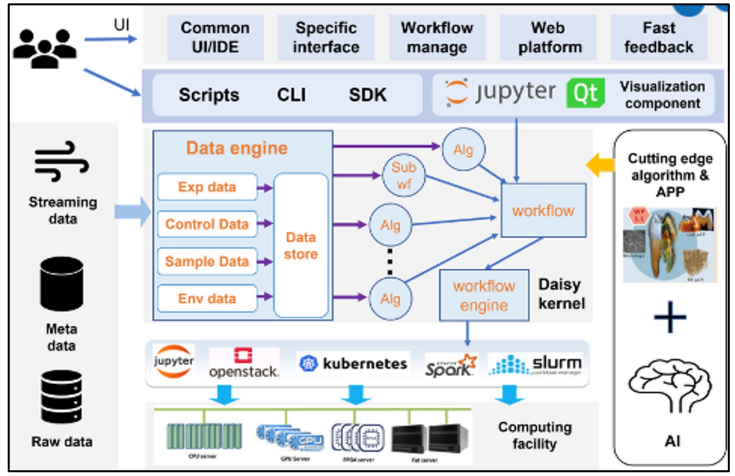
DOMAS Framework



- E-scale big data storage
- Complex metadata management
- Standardized data formats

Data analysis and Processing

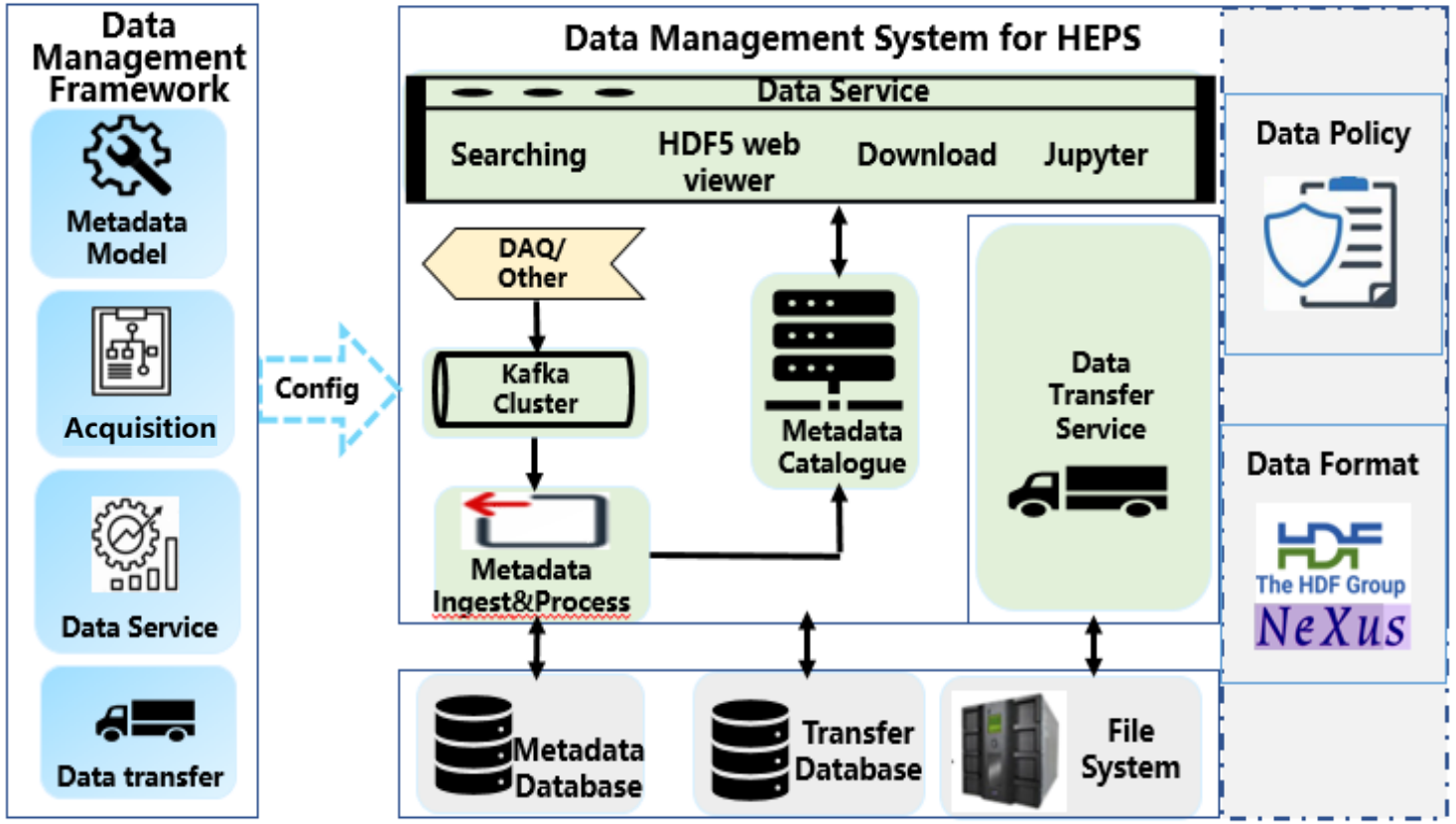
Daisy Framework



- Integration of multiple methodologies and AI algorithms
- High-throughput data I/O
- Domain-specific software and workflow

DOMAS Framework

- Core modules
  - ✓ Metadata catalogue
  - ✓ Data acquisition
  - ✓ Data transfer
  - ✓ Data service
- MongoDB: Metadata Flexible and extensible
- Open Source: Customize APIs and connect with other service systems
- By simply modifying the configuration, DOMAS can be used to build data management system for facilities quickly



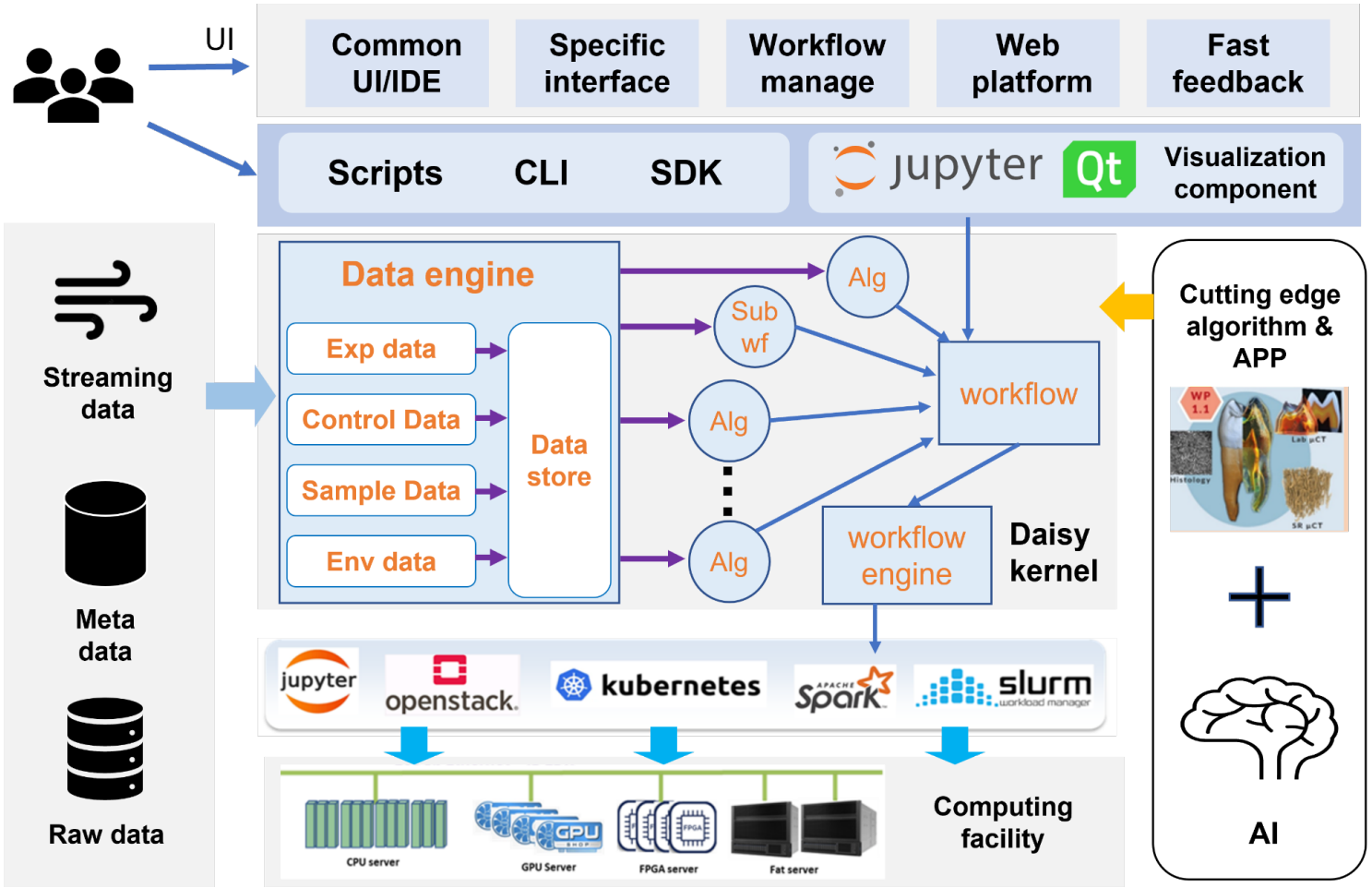
Apply to several facilities in China: BSRF, CSNS, HEPS, SHINE

DOI: 10.1107/S1600577524000043

*Hao Hu *, Leilei , Haofan Wang , Bo Zhuang , Ruojin Zhang, Qi Luo , Xiaokang Sun , Fazhi Qi*

Daisy Framework

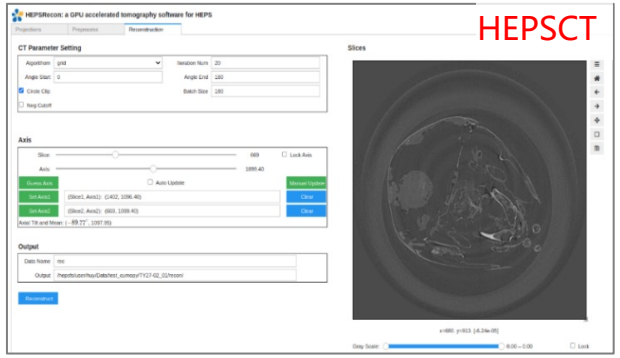
Domain-specific applications built on the framework, plus a general workflow-orchestration system for flexible data-processing needs



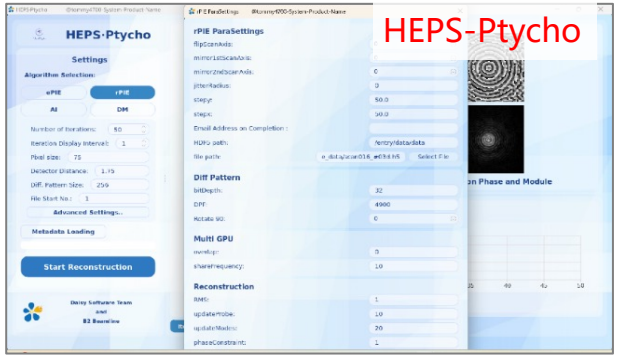
Enabling technology modules for next-generation light-source data processing

- ✓ Data-object management for high-throughput I/O, multimodal parsing, and multi-source ingestion
- ✓ Elastic, heterogeneous compute-cluster capacity for low-latency processing across scales and throughputs
- ✓ User software interfaces and developer environment support for integrating and advancing domain methodologies

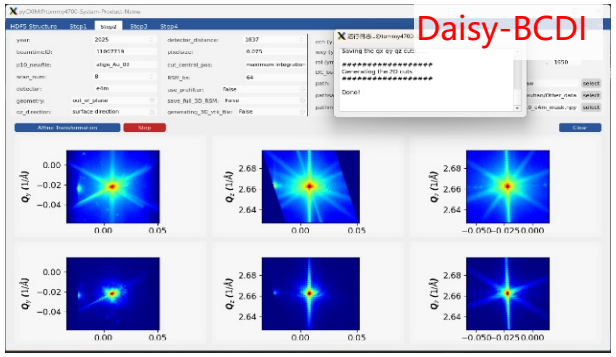
Daisy Scientific Applications for BLs

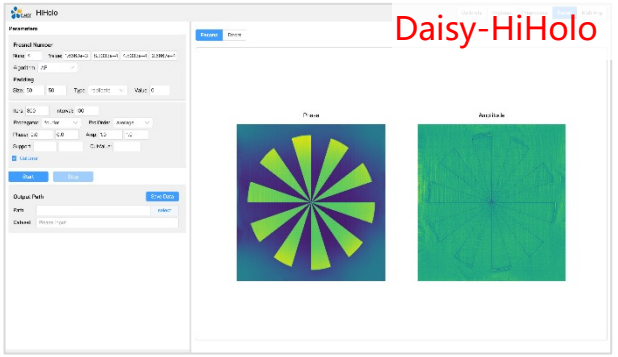


HEPSCT



HEPS-Ptycho

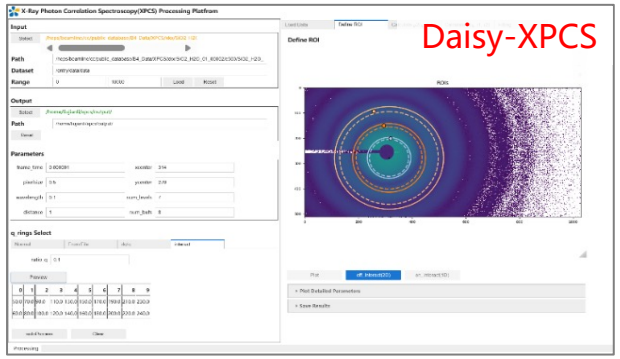

Daisy-BCDI



Daisy-HiHolo



Daisy-PDF



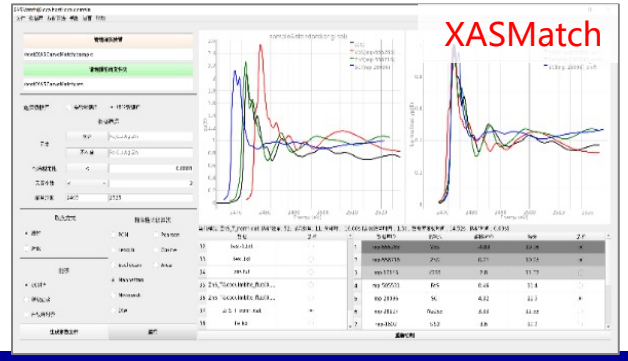
Daisy-XPCS

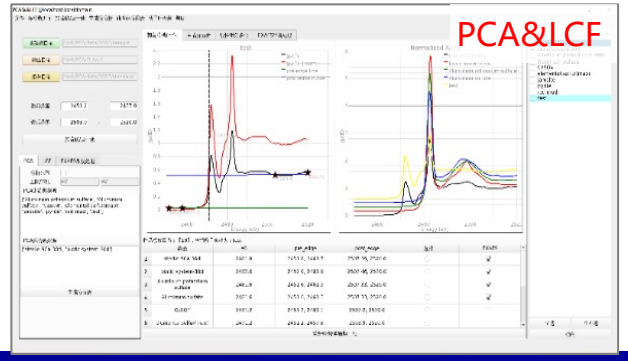


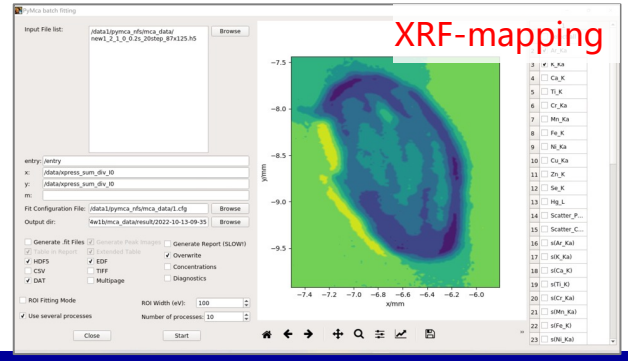
Daisy-SAXS



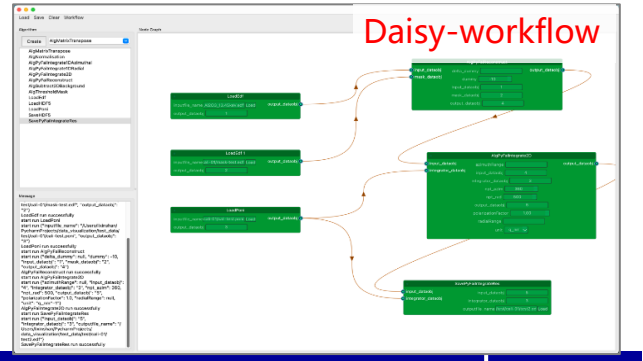
Daisy-BMX


XASMatch

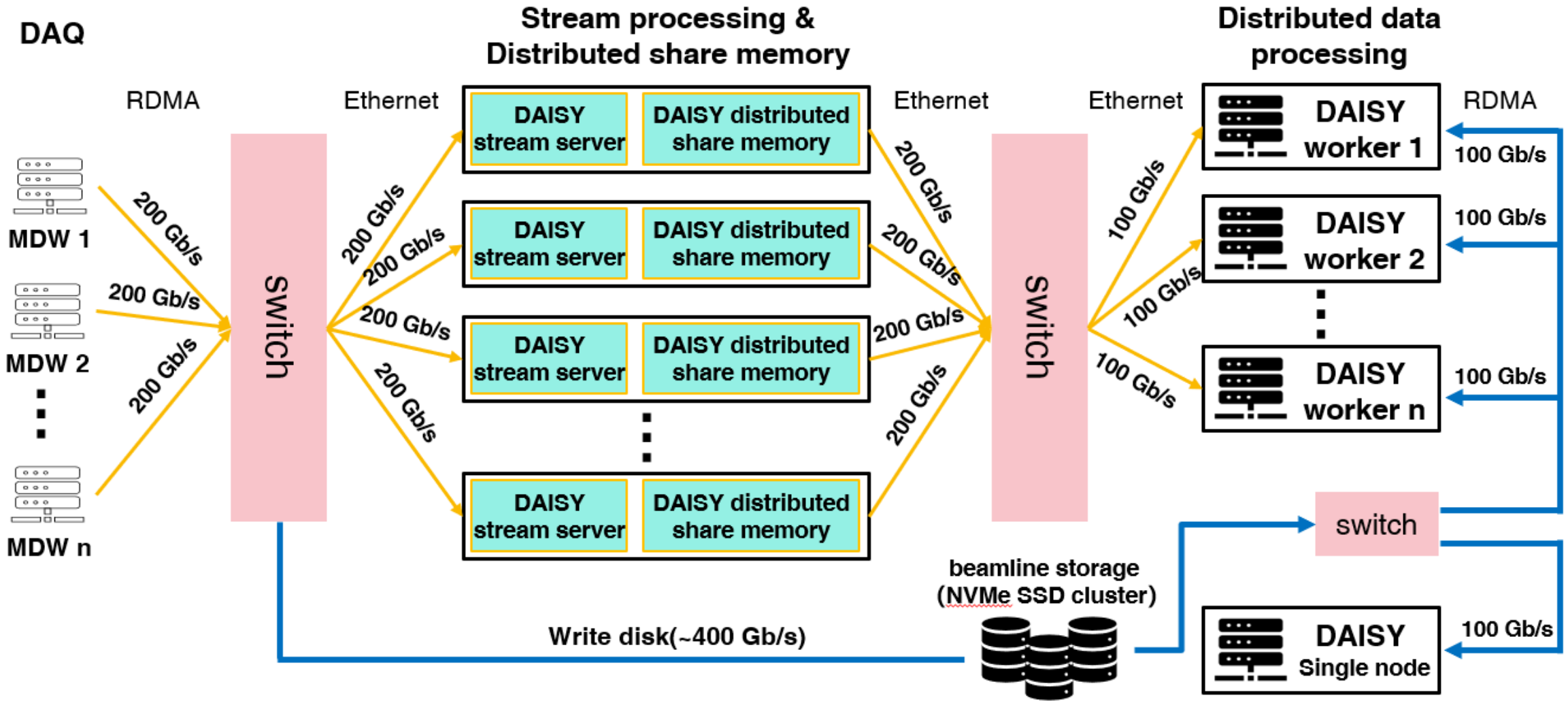

PCA&LCF



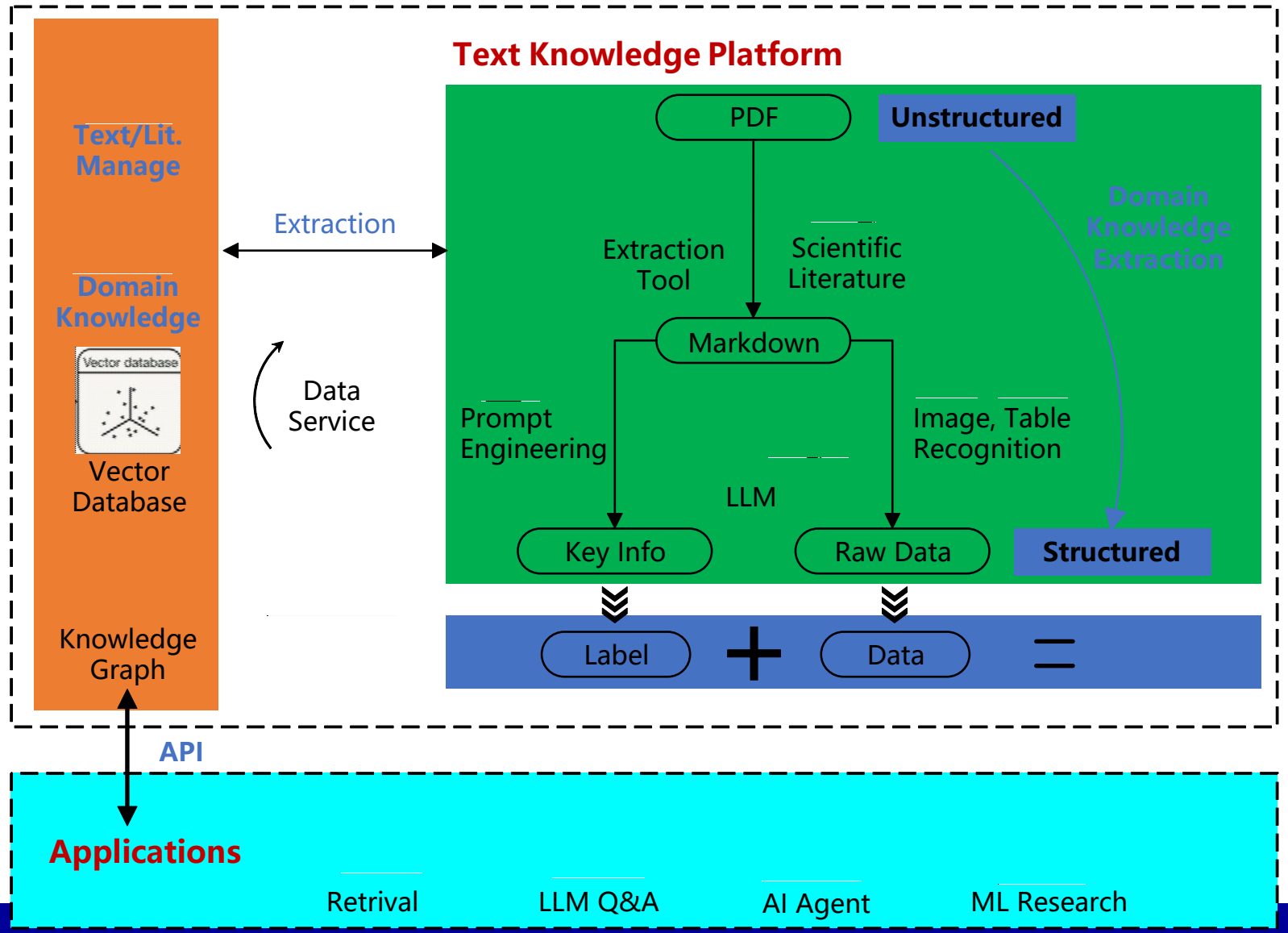
XRF-mapping


Daisy-workflow

Streaming and distributed system based on Daisy for BLs



Text & Knowledge Platform (Developing)



- **Core functions:**
  - ✓ **Text, Knowledge Management**
  - ✓ **Knowledge Extraction**
  - ✓ Vector Database (in dev)
  - ✓ Knowledge Graph (in dev)
- **Text for Multiple Domains:**
  - ✓ **Scientific Literature**
  - ✓ Proposal, Log from Experiments
  - ✓ Documents for Data Analysis
  - ✓ TEXT/PDF/Image/PPT/…
- **Applications for domain knowledge extraction:**
  - ✓ **Shale Research Literature Dataset**
  - ✓ **Neutron Research Literature Dataset**
  - ✓ Sensor Research Literature Dataset

Data Agent (Developing)

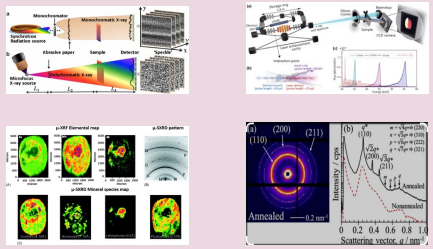
Data Agent Utilizing AI agent technology to process and supply data from HEPS&CSNS.

Sensor & Actuator

Data Management

Experimental data | Metadata

Spectroscopy, Diffraction, Scattering, Imaging, **etc.**



Processed data | Simulation data

True label for AI task

AI-Ready Dataset

Memory

- RAG & Knowledge Graph
- For beamline (HEPS), instrument (CSNS)
- For experimental data analysis

RAG Knowledge Graph



① Update
- Proposal
- Experiment
- Materials
- etc.



LLM

② Task identification
- Reconstruction, Peak finding
- Denoising, Compression
- Segmentation, etc.

API

⑥ Data for multi-objective research

Sensor

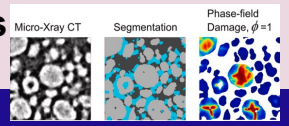
③ Data Retrieval
- Format Conversion
- Execute commands

Actuator

④ Automatic processing
- Execute tools or code

⑤ Verification and Storage

Analysis Platform
- XAFS Tools
- XRD Tools
- PDF Tools



Simulation Platform
- Neutron Scattering
- First-principles calculation
- ……

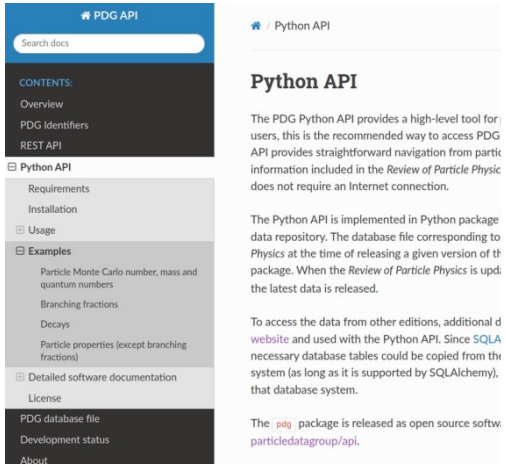
Data Service & Provisioning

- **Data Service Portal**/Data Fusion Portal/Dataset Module in AI Platform for **Experiment User & Domain Scientist**
  - Accessible via both intranet and internet.
  - Providing data retrieval, high-speed downloads, file preview, e-logs.

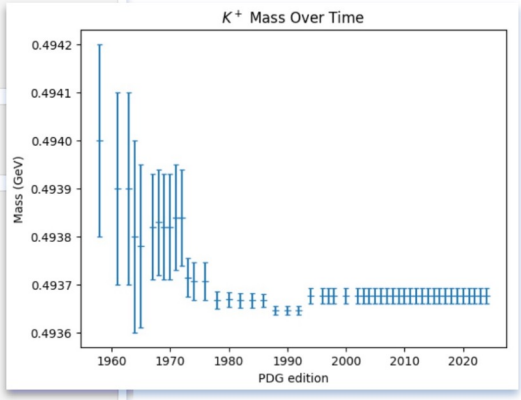


- **RESTful API for Data Researcher & AI Agent**
  - REST API
    - Download Metadata with JSON format
    - Sporadic, limited access
    - Python API
  - APIs for accessing facility data
    - Data loader for AI
    - Local storage
  - Database File
    - Access data entity file
    - Low-level access

Retrieval, Download, Preview

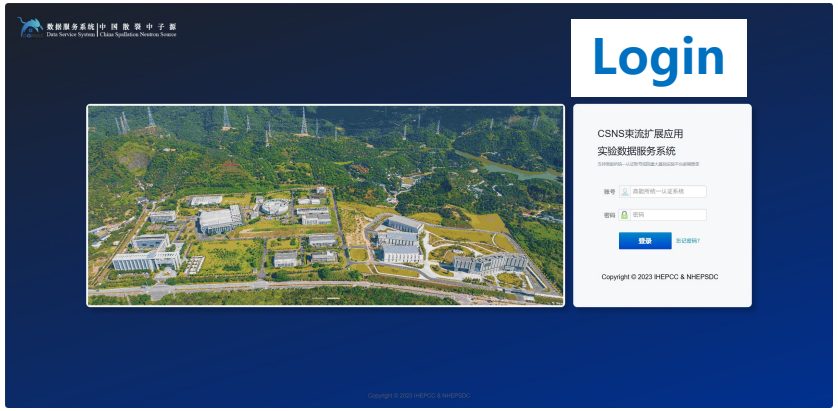


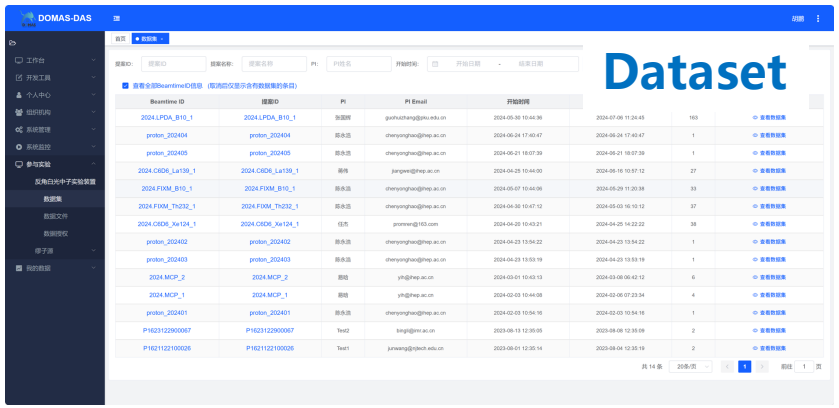

```
[1]: import pdg
     import matplotlib.pyplot as plt
     import numpy as np

[2]: # Use a PDG database file with historical data
     api = pdg.connect("sqlite:///pdgall-2024-v0.1.2.sqlite")

[3]: xs, ys, yerrs = [], [], []
     for edition in api.editions:
         p = api.get_particle_by_name('K+', edition=edition)
         if p.has_mass_entry and p.mass is not None:
             xs.append(int(edition))
             ys.append(p.mass)
             yerrs.append(p.mass_error)

     plt.errorbar(xs, ys, yerrs, marker='_', capsize=2, linestyle='')
     plt.xlabel('PDG edition')
     plt.ylabel('Mass (GeV)')
     plt.title('$K^+$ Mass Over Time');
```

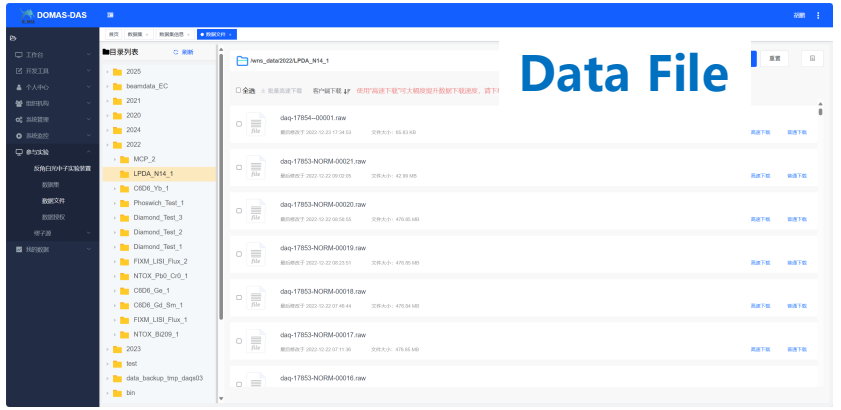


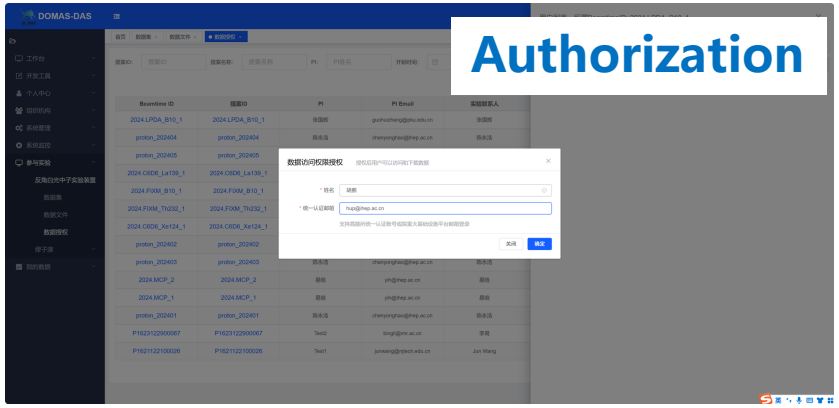
Diverse data services and flexible provisioning methods to meet needs across various scenarios and complexity levels.

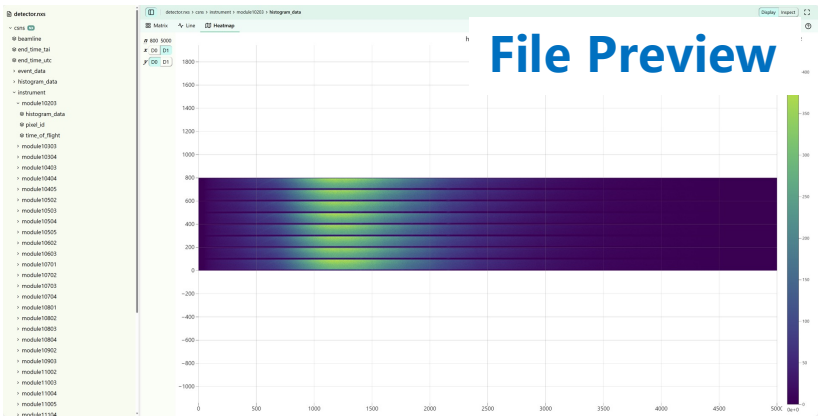
Data Service Portal

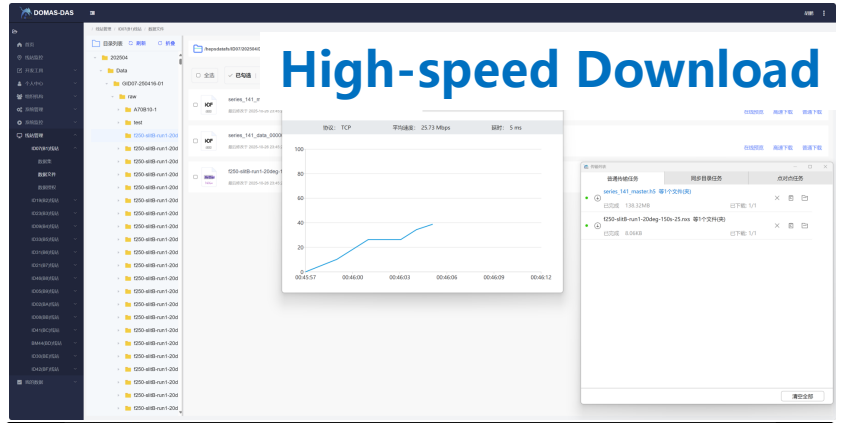






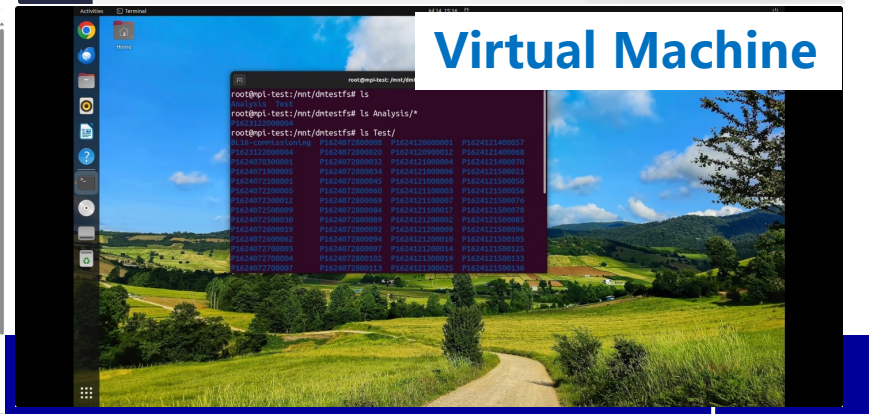






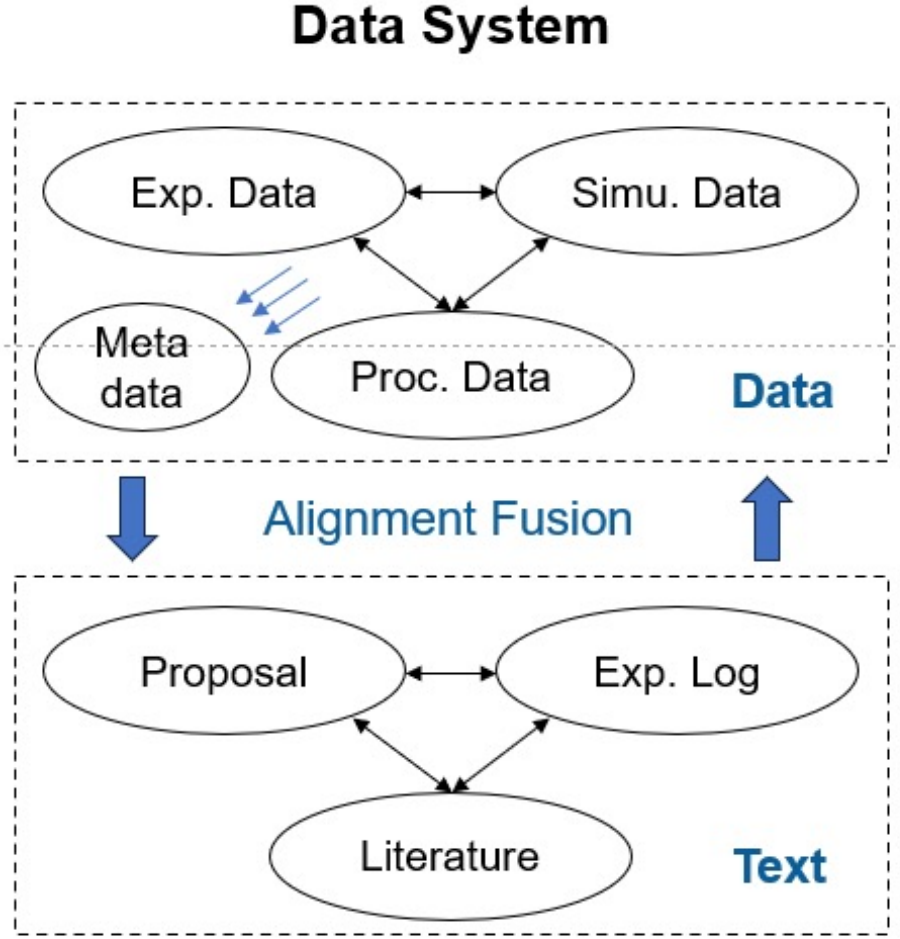


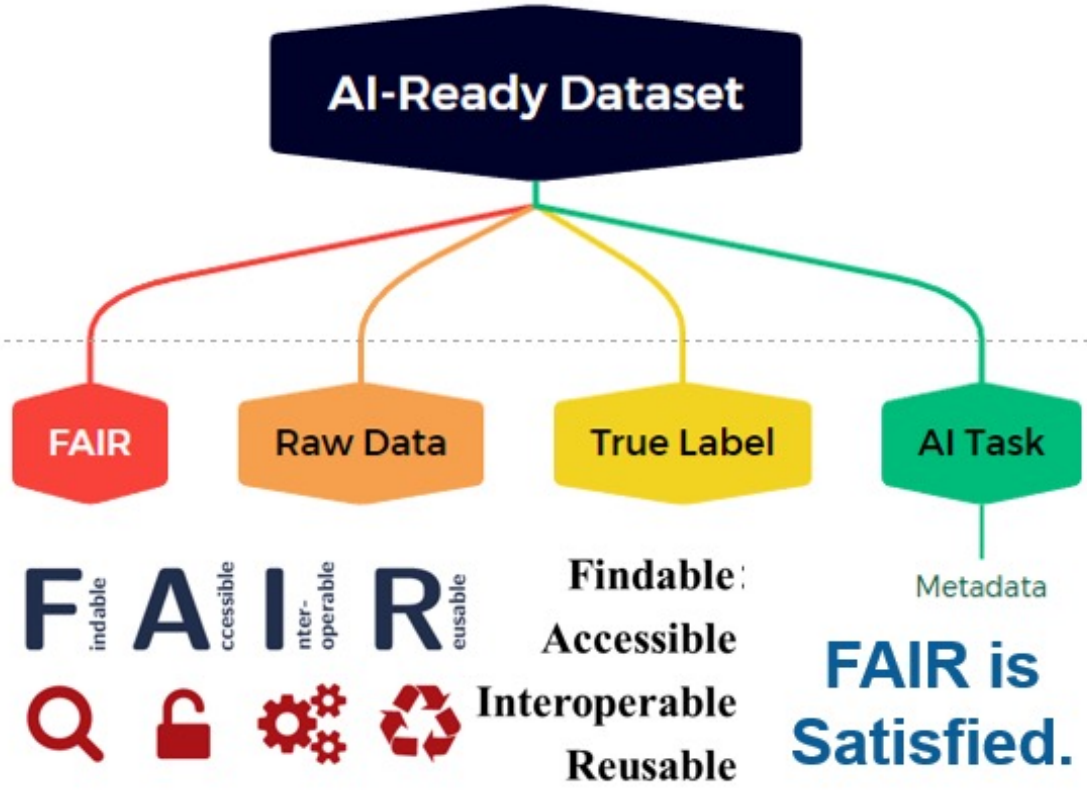




AI-Ready Datasets

- **AI-Ready domain datasets: Ready-to-use for AI training/testing for domain research.**





Next key focus: Constructing true labels based on AI tasks.

Open Dr.Sai

The foundation of above agents

Opensourced:
https://github.com/hepai-lab/drsai

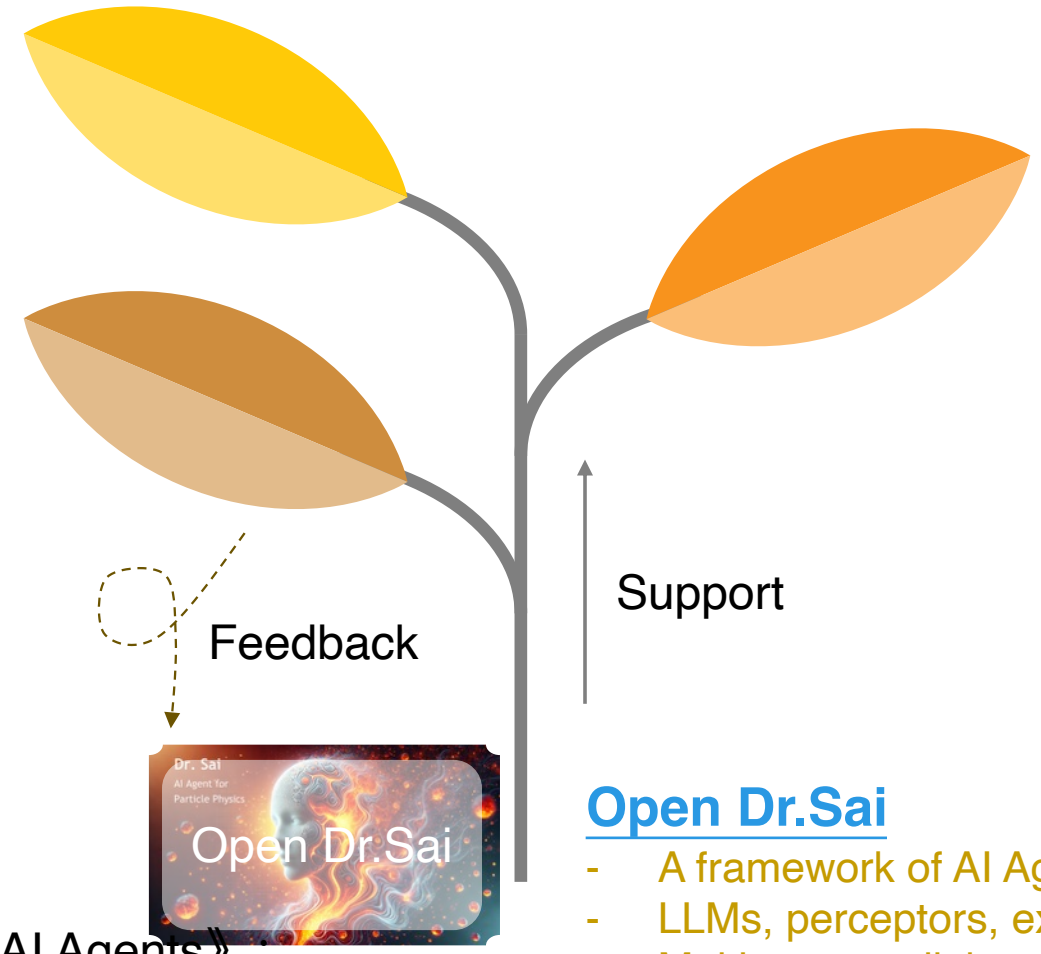
03 More Agents

...

01 Dr.Sai BESIII

Physics analysis agent

- Enhance physical abilities
- Perceive BESIII data
- Equipped with BOSS tools



02 DORA Data Agent

Light source data agent

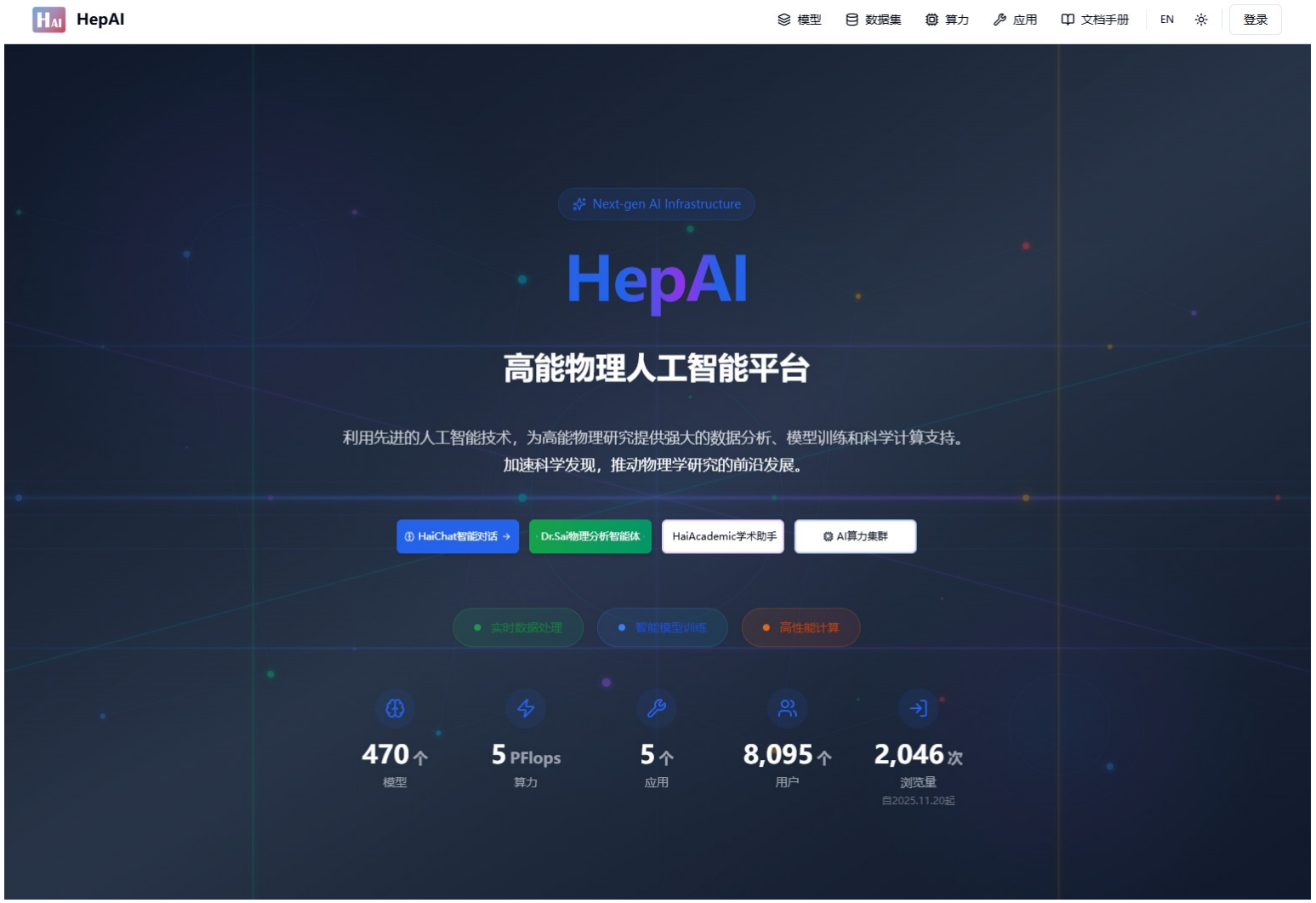
- Enhance data capability
- Perceived light source data
- Equipped with subject methodology tools

《Practical Guide to Building AI Agents》:
https://note.ihep.ac.cn/s/_vSnuf5Uh

Open Dr.Sai

- A framework of AI Agent for science.
- LLMs, perceptors, executors, and memory
- Multi agent collaborative logic
- Permission verification and human-computer interaction

AI Platform





https://ai.ihep.ac.cn

Providing
“Data-Computing-Models-Tools”
for HEP Users

Collaboration

- Institutes















- Open source collaborations and contributions













- Industry



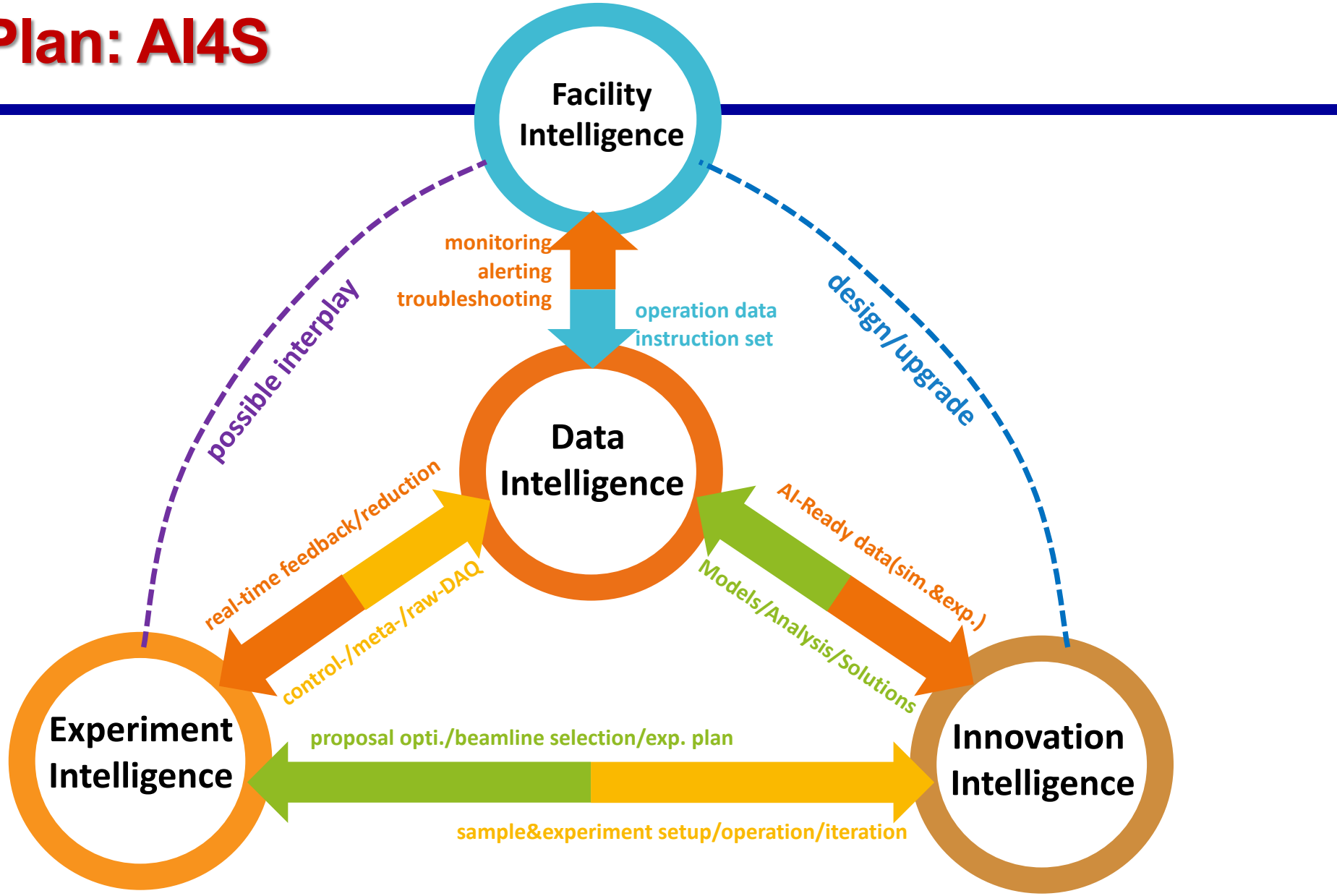







- Participation in international conferences ~60 times/year
  - CHEP, ACAT, HEPiX, ISGC, AFAD
- Hosting national/international conferences and workshops ~4 times/year
- Academic visitors
  - FNAL, CERN, LBNL, IN2P3, JINR, SIB, MTA, NTU, CHTC

Future Plan: AI4S



Future Plan: New Generation Computing Infra.



Strategic Goals

- Centralized Resource Management & Multi-dimensional Collaboration
- Interoperability & Compatibility Assurance

One-click HEP-AI Services

Aggregation

Models

Data

Computing

Network

Storage

Security

The 3rd Resources

Integration & Security

- External Infrastructure Convergence
- Comprehensive Security Coverage for All Layers

Summary

- Data, computing and software(AI included) are important
- We have had good progress so far
- But still facing many challenges and looking forward to closed collaboration
  - Computing, data , software & AI
  - Post-doctors and Staffs ……

Thanks