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

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

高能所计算中心
IHEP Computing Center

HEPS scientific computing system design for interactive data analysis scenarios

Qingbao Hu

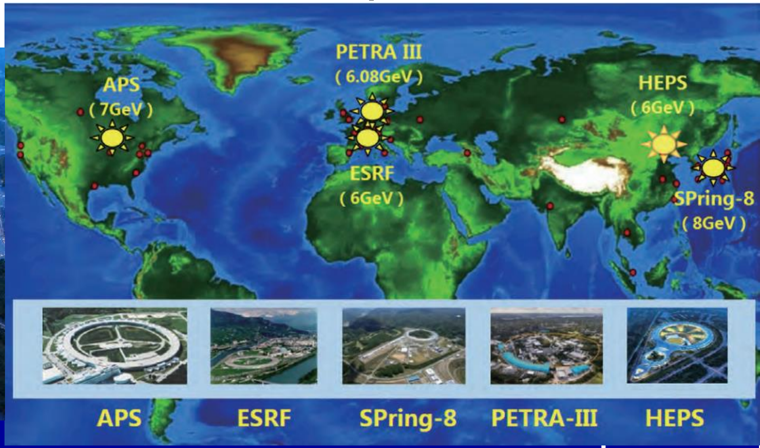
IHEP-CC HEPS-CC

Outline

1. Introduction to HEPS
2. Challenges in HEPS Data Analysis
3. Computing System Design
    1. Computing Analysis Service
    2. User Management
    3. Data Access Methods
    4. Resource allocation
4. Summary and Outlook

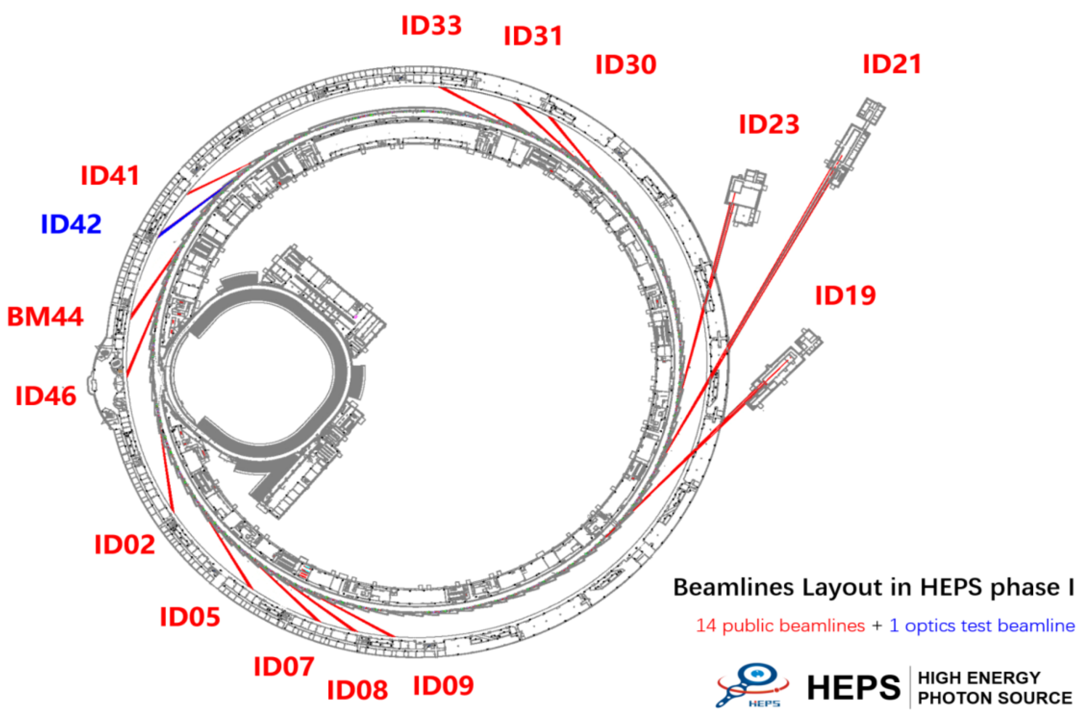
HEPS Introduction

- The High Energy Photon Source (HEPS), located at Beijing HuaiRou, is an advanced public platform for multidisciplinary innovation research and high-tech development, as well as understanding many scientific questions in the fields of physics, chemistry, biology, etc.
- At HEPS, researchers will be able to observe the complex samples with more sensitive, finer, faster experimental tools, under condition close to the actual working environment. Therefore, researchers will be able to obtain the multidimensional, real-time, in-situ characterization of sample structure, as well as the dynamic evolution processes. HEPS can help researchers understand matter more accurately at the level of molecules, atoms, electrons and spin and in the dimensions of space, time and energy.



Data volume of HEPS

- 14 beamlines is built in HEPS, first phase
- 20+PB raw experimental data per month



Beamlines	Burst output (TB/day)	Average output (TB/day)
B1 Engineering Materials Beamline	770.00	20.00
B2 Hard X-ray Multi-analytical Nanoprobe Beamline	2300.00	340.00
B3 Structural Dynamics Beamline	8.00	3.00
B4 Hard X-ray Coherent Scattering Beamline	180.00	90.00
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	98.00	85.60
B8 X-ray Absorption Spectroscopy Beamline	80.00	10.00
B9 Low-Dimension Structure Probe Beamline	20.00	5.00
BA Biological Macromolecule Microfocus Beamline	35.00	10.00
BB pink SAXS	40.00	10.00
BC Nanoscale Electronic Structure Spectroscopy Beamline	10.00	1.00
BD Tender X-ray beamline	40.00	10.00
BE Transmission X-ray Microscope Beamline	99.36	60.20
BF Test beamline	1000.00	60.00
Total average:		720TB/Day

Motivation

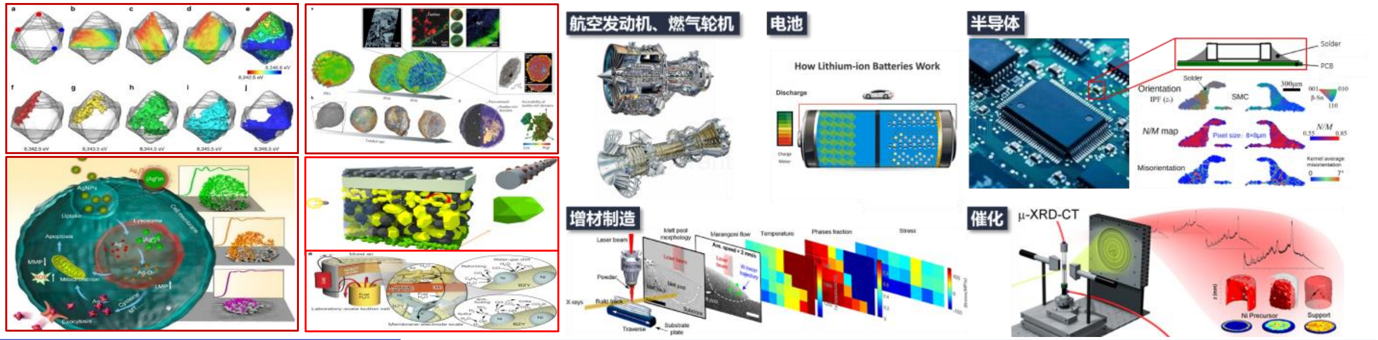
- Traditional model - Rarely using the platform computing resources
  - Storage of the raw data.
  - Download the experiment data
  - Analysis with personal computing resources.
- HEPS - brings unprecedented storage and computing challenges
  - The data scale and throughput generated by advanced light sources is very huge.

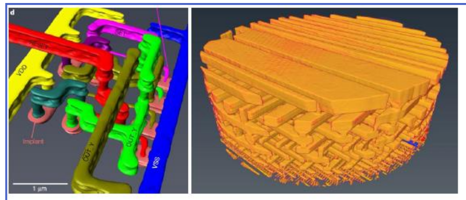
- With such a huge amount of data, how to reduce the movement of large-capacity experimental data, provide the same data analysis environment as the user's personal computer, and implement the data generation and data analysis processes online, which is a new idea for modern light sources platform to provide services to users.

Data Analysis Requirements

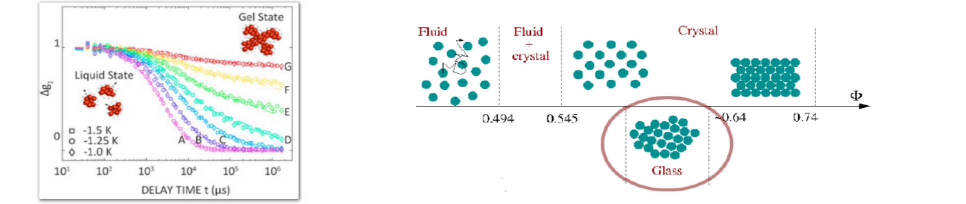
- HEPS includes three methodologies: imaging, diffraction/scattering, and spectroscopy.
- Data analysis scenarios involving: CT reconstruction, image segmentation, model training, diffraction data reconstruction, absorption spectrum data processing, etc.
- HEPS has a large user base, and a large amount of experimental data, and access must be strictly controlled.

- Complex methodologies
- Diverse analysis tools
- Differentiated environments
- Strict ACL management





The 3D imaging of integrated circuits, providing the direct feedback between design and manufacturing processes, the relationships of microstructure and function.
Mirko Holler et al., Nature ***543***,*402(2017)*



The dynamics of colloidal nanoparticles, providing the mechanisms of phase transition such as supercooled liquids, glass transition, gel formation and dissolution, crystallization, jamming; improving the dynamics model systems such as hard sphere model. *Q. Zhang et al. Phys. Rev. Lett.,* ***119****, 178006(2017); P. Pusey et al., Nature* ***320****, 340(1986); D. Orsi et al., Physical Review E, 85, 011402(2012)*

Goal

- The HEPS computing system
  - Provide the most suitable computing service solutions based on users' different usage patterns and different experimental needs.
  - Provide rapid deployment of resources at the minute level based on the user's operating system and computing power needs.

- It is of great significance to build a HEPS scientific data computing system, provide users with computing resources and analysis environment during and after the experiment, and help users discover one-stop scientific research activities.

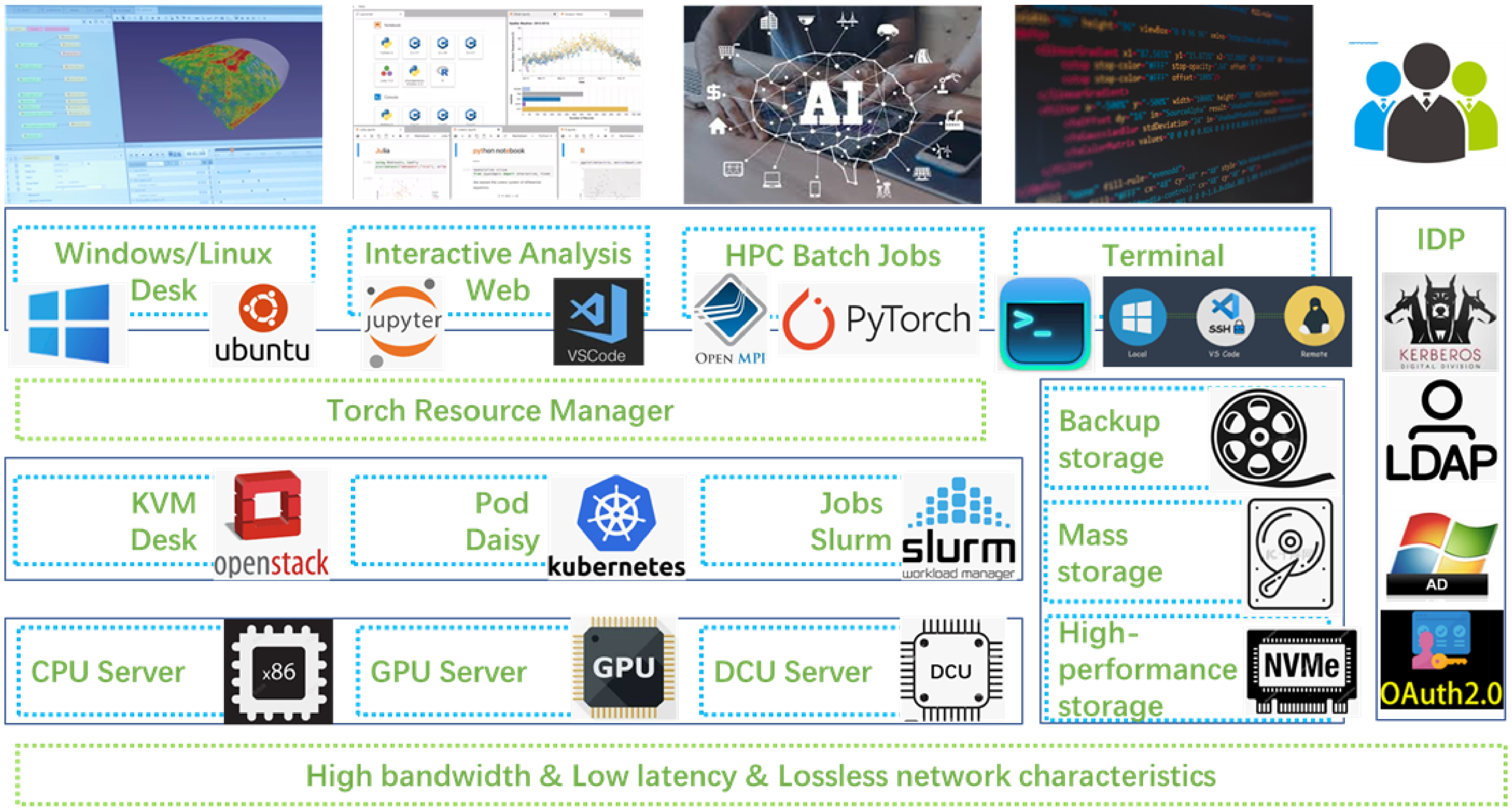
Framework

Torch: the HEPS scientific computing platform

Computing service types

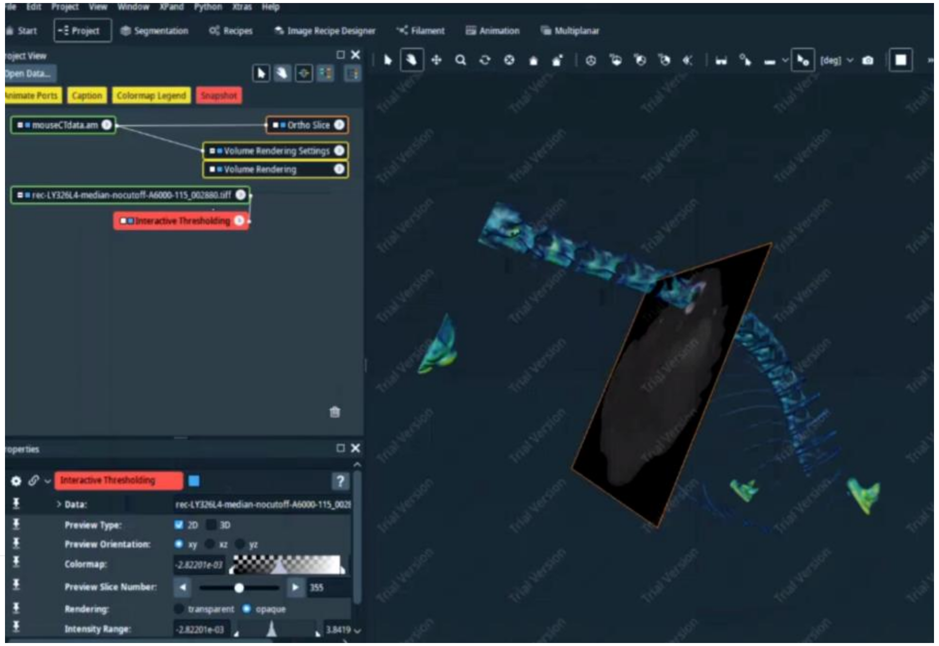
Analysis environment integration

Heterogeneous resource management

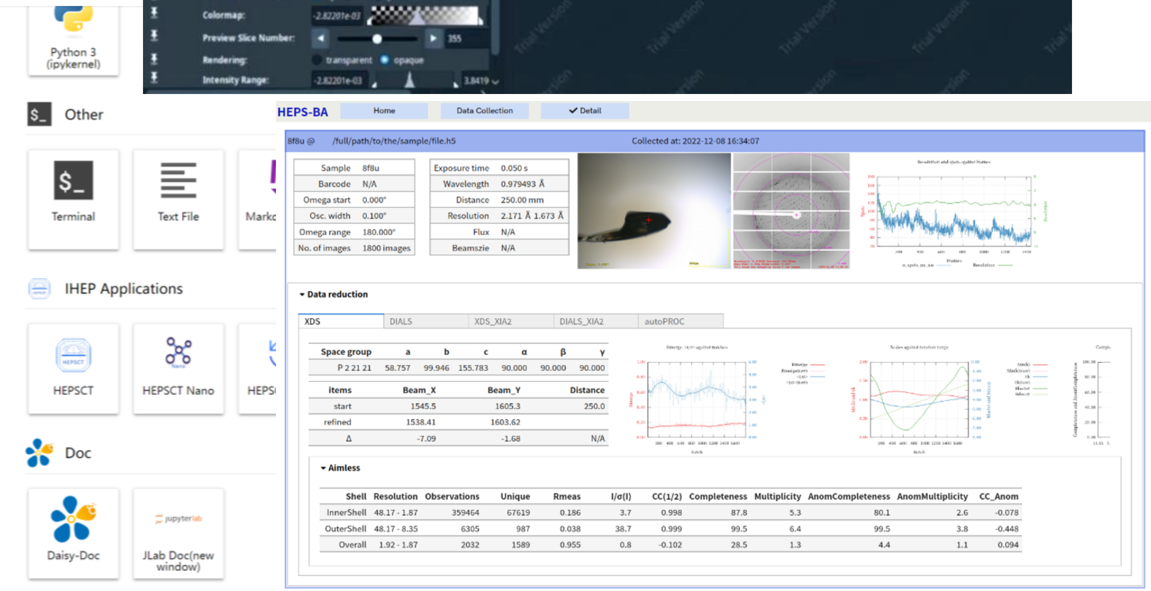


Service scenarios

- Use Image analysis software to segment each layer of experimental data based on the Windows or Ubuntu desktop environment, and use the mouse to rotate at different angles to view the 3D imaging effect.

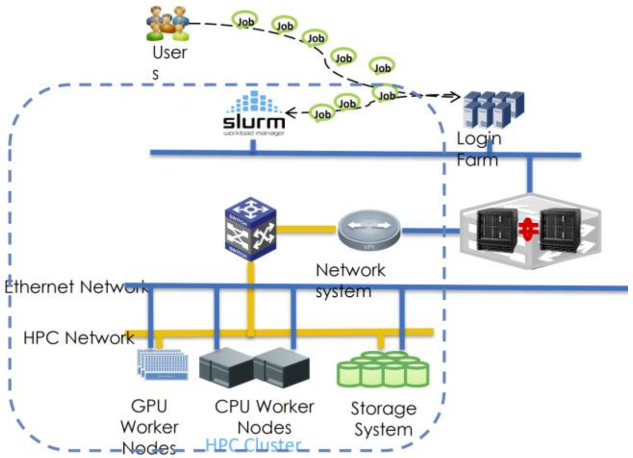


- Based on the JupyterLab and combined with the GUI model of Daisy software framework, the interactive data reconstruction requirements for experimental data are completed.

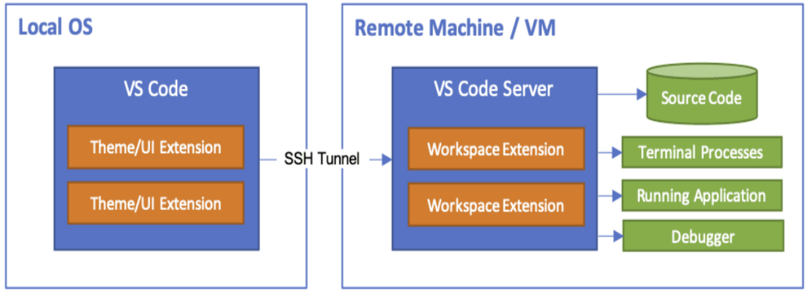


Service scenarios

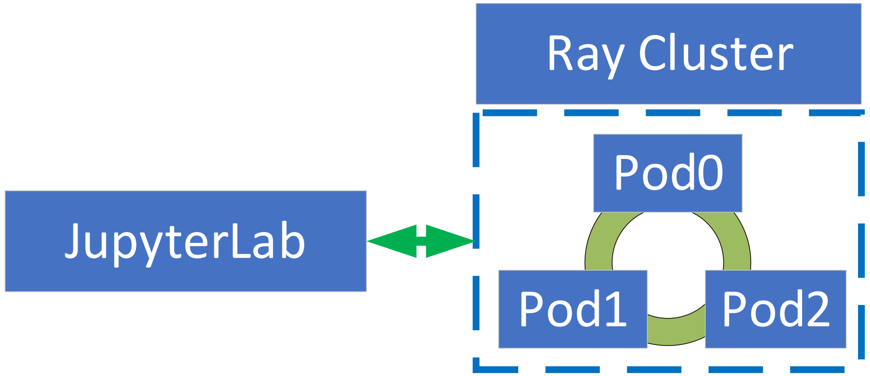
- Implement terminal command line interactive mode based on web pages and submit HPC jobs to the slurm cluster.



- Provides ssh access by ssh-key, users can interact with files and folders anywhere on the computing environment by vscode remote ssh extension.



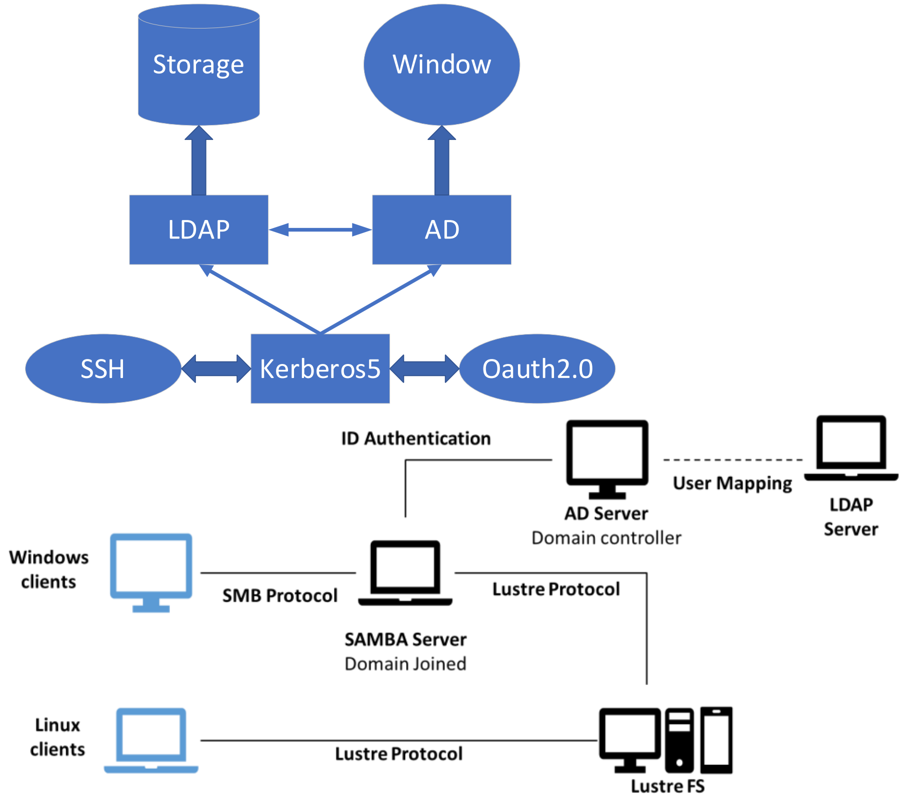
- Customize a small cluster of several nodes, log in to each other with ssh-key, and build a spark cluster or Ray cluster to achieve parallel analysis.



User management

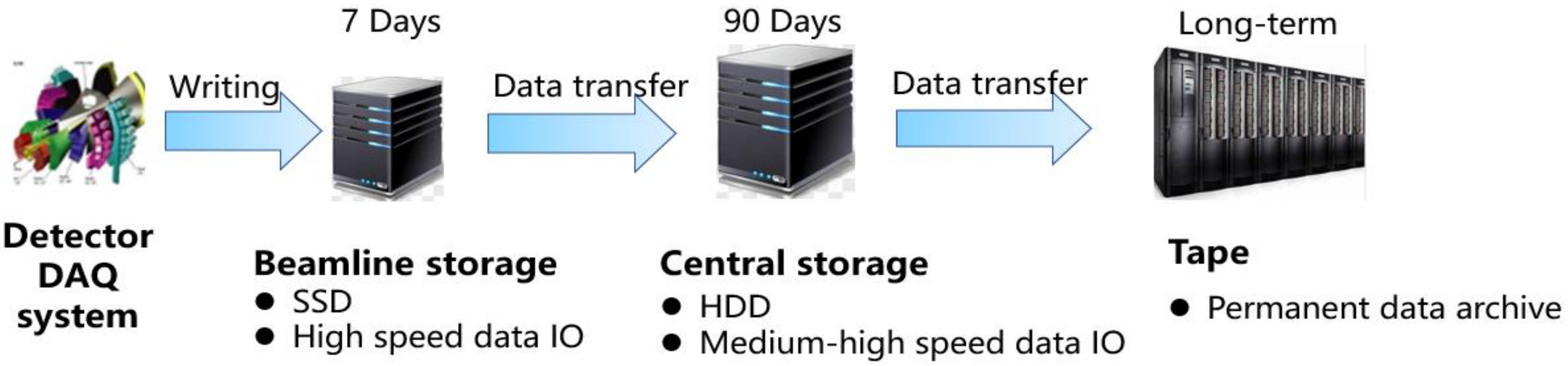
- These rich data analysis scenarios of HEPS lead to users accessing experimental data in a variety of ways.
- Maintaining the consistency of user identity and data access permissions in different scenarios are necessary.

- Associate user accounts and user identities to create secure experimental data access patterns.
- Accounts with the same identity have the same data access rights in different scenarios
- The data generated by accounts in respective scenarios also have the same ACL attributes in other scenarios.



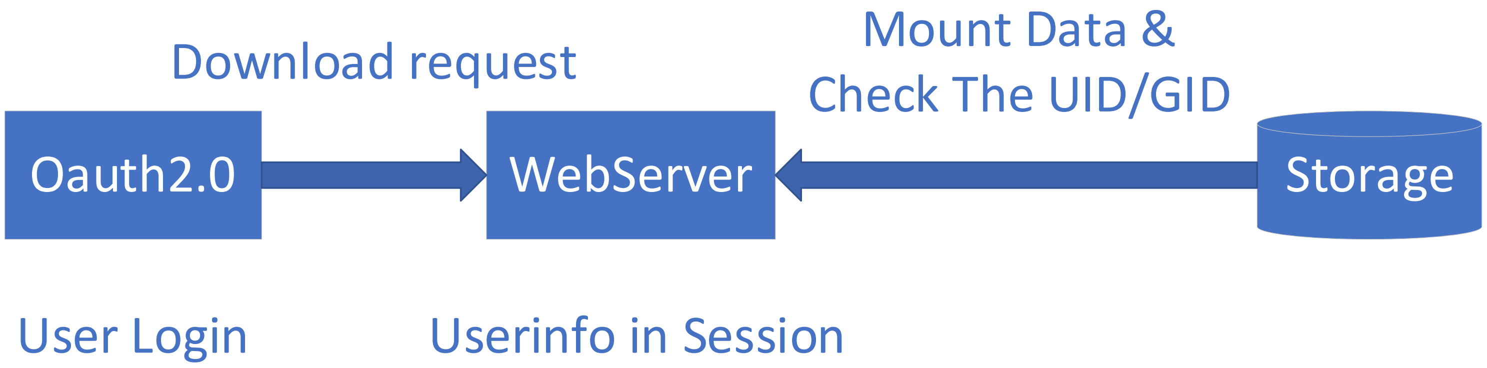
Data management

- After the HEPS experimental data is generated from the DAQ module, it is written to the POSIX storage system and the ACL corresponding to the user list is set.
- Permissions are retained during the data migration process to ensure the manageability of users downloading data from the website and accessing data from the computing environment.



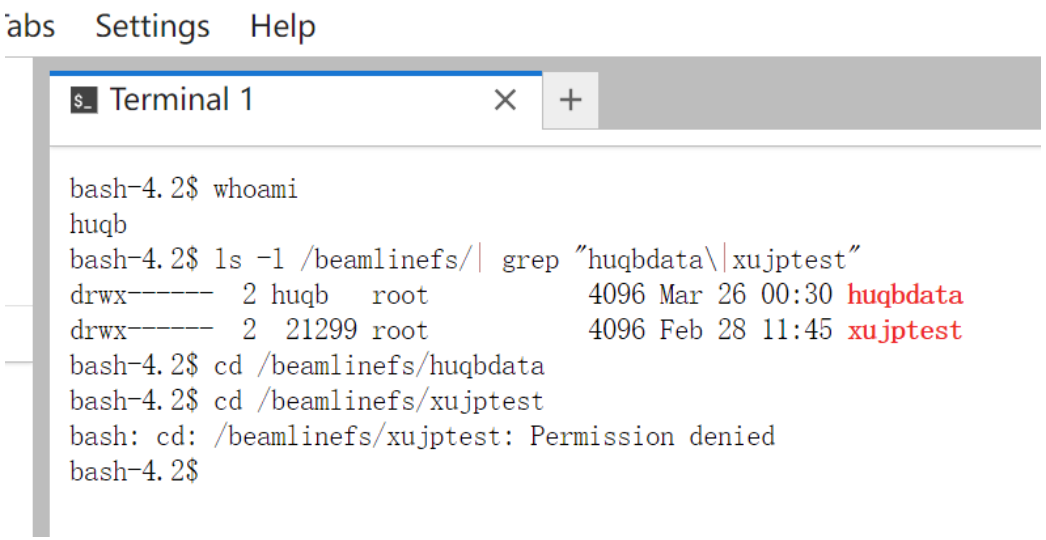
Data access

- User can access data with the following three modes:
  - Download data from the web site.
    - Users can only see the authorized data list on the page.
    - check user info from the web session, and check the acl of object file contains the current user.

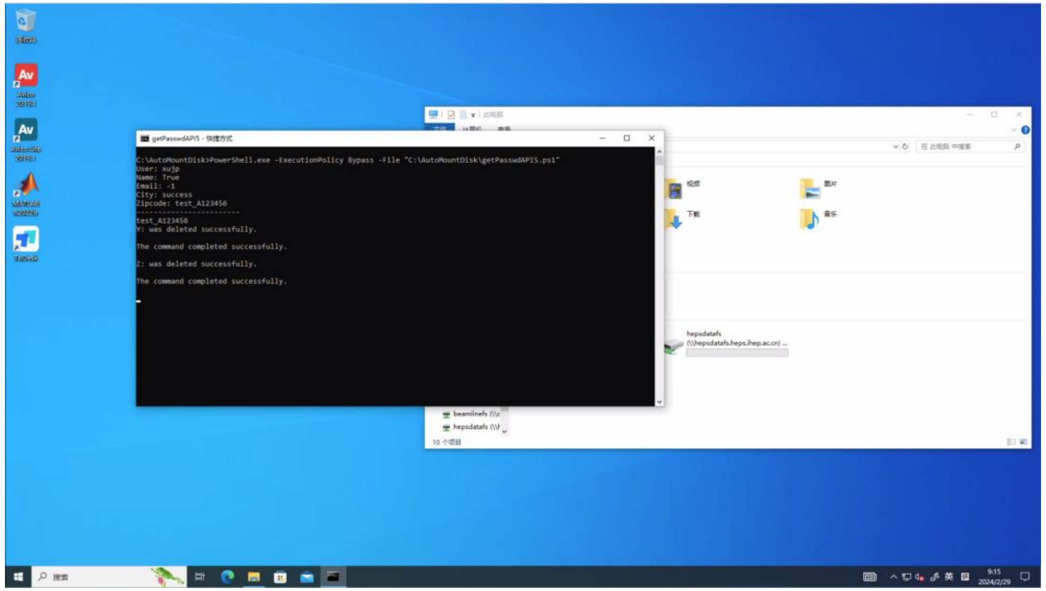


Data access

- Use the jupyterLab tool to analyze the data.
  - Restrict the user identity of the user's operating system session, and the user directly accesses the data using ACL with POSIX semantics.



- Use virtual cloud desktops to analyze the data.
  - Map the Microsoft Windows security identifiers (SIDs) to the Linux user/group identifiers (UID/GID). SMB/CIFS



IO performance

- During the HEPS user experiment, it is necessary to quickly analyze the experimental data of the sample to optimize the control parameters of the detector.
- After the experiment, it is necessary to shorten the data analysis time and obtain the experimental results quickly.

- Data transmission network: RoCE - high bandwidth, low latency network
- Data transfer method: memory to memory, data flow, avoid writing to disk.
- Separate data transmission: Containers/Nodes both have TCP/RoCE dual network
- Data access mode: pre-read data & HDF5 & RDMA & Nvidia-GDS
- Multi-tier storage media: All-NVME array & Disk storage
- Data Parallel Analysis: Parallel Data Access

Resource allocation

- In order to meet the desktop analysis environment during user experiments, we started 14 VMs on 6 physical machines and equipped them with A5000 GPU cards according to analysis needs. (Static & Simple)
- For HPC computing needs, 4 * A800 GPU servers are provided, and slurm jobs are queued to use these resources. (Queue & Simple)
- For jupyterlab data analysis needs, k8s manages 8 physical nodes and provides computing resources. (Dynamic & Complex)
    - In order to prevent users from continuing to occupy resources, we wrapped jupyterhubapi to form torchspawnerAPI, which can meet richer control needs.

Resource allocation

Example Value | Schema

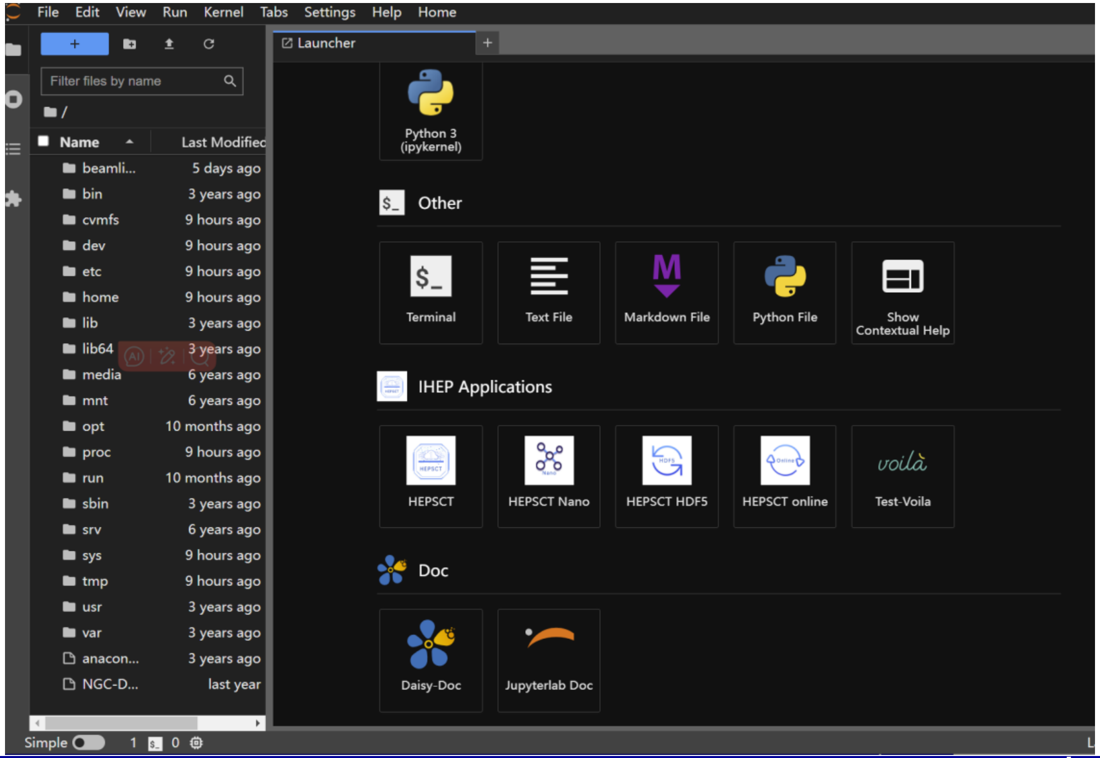

```
{
  "username": "string",
  "sourceip": "string",
  "description": "string",
  "podname": "default",
  "userimage": "dockerhub.ihep.ac.cn/opendata/cumopy:db61750",
  "ncores": 48,
  "memory": "200G",
  "sshpubkeys": "ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQDW1SRb
RrlekTTgLTelRlgNiAEX/c0rafS2k7vnoJ0QZV7HQ84e+rcYzuNpzbG1gjY2x
  "extra_resource": {
    "nvidia.com/gpu": 1
  },
  "single_cmd": "start.sh",
  "single_cmd_args": "jupyterhub-singleuser;--ip=0.0.0.0;--po
  "scriptenv": "",
  "expireruntime": 86400
}
```


default

POST	/users/	Create User
GET	/users/	Read Users
GET	/users/count	Read Users
GET	/users/{username}	Read Item
POST	/userpod/	Create Pod By User
DELETE	/userpod/	Delete Pod By User
PATCH	/userpod/	Update Pod By User
GET	/refreshuserspod/	Get Pods Sync Db
GET	/userpod/{username}	Get Pods By User
GET	/uniqueipinfo/	Get Uniqueipinfo
PUT	/uniqueipinfo/	Update Uniqueipinfo
POST	/uniqueipinfo/	Create Uniqueipinfo

Resource allocation

- Each of the 14 beamlines has a computer placed in their operation room, corresponding to 14 IP addresses. When users from these IPs start jupyterlab, we will release the container resources previously requested by this IP.



Summary and Outlook

- The HEPS scientific computing system realizes centralized management and unified scheduling of heterogeneous computing resources, and provides diversified computing service solutions according to different usage patterns.
- In the future, it will be further combined with the software framework to promote parallel computing and data access efficiency; it will also adapt to more accelerator card resource management. Combined with application requirements, optimize resource scheduling algorithms and improve resource usage efficiency. Improve better experimental data analysis services.

Thanks for your attention!

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

HEPS-CC