Distributed Computing Status at IHEP, CAS

Xuantong Zhang, On behalf of Distributed Computing Group

Computing Center of IHEP, CAS

Outline

Introduction

Grid infrastructures and services at IHEP

Distributed computing system for HERD experiment (As an example)

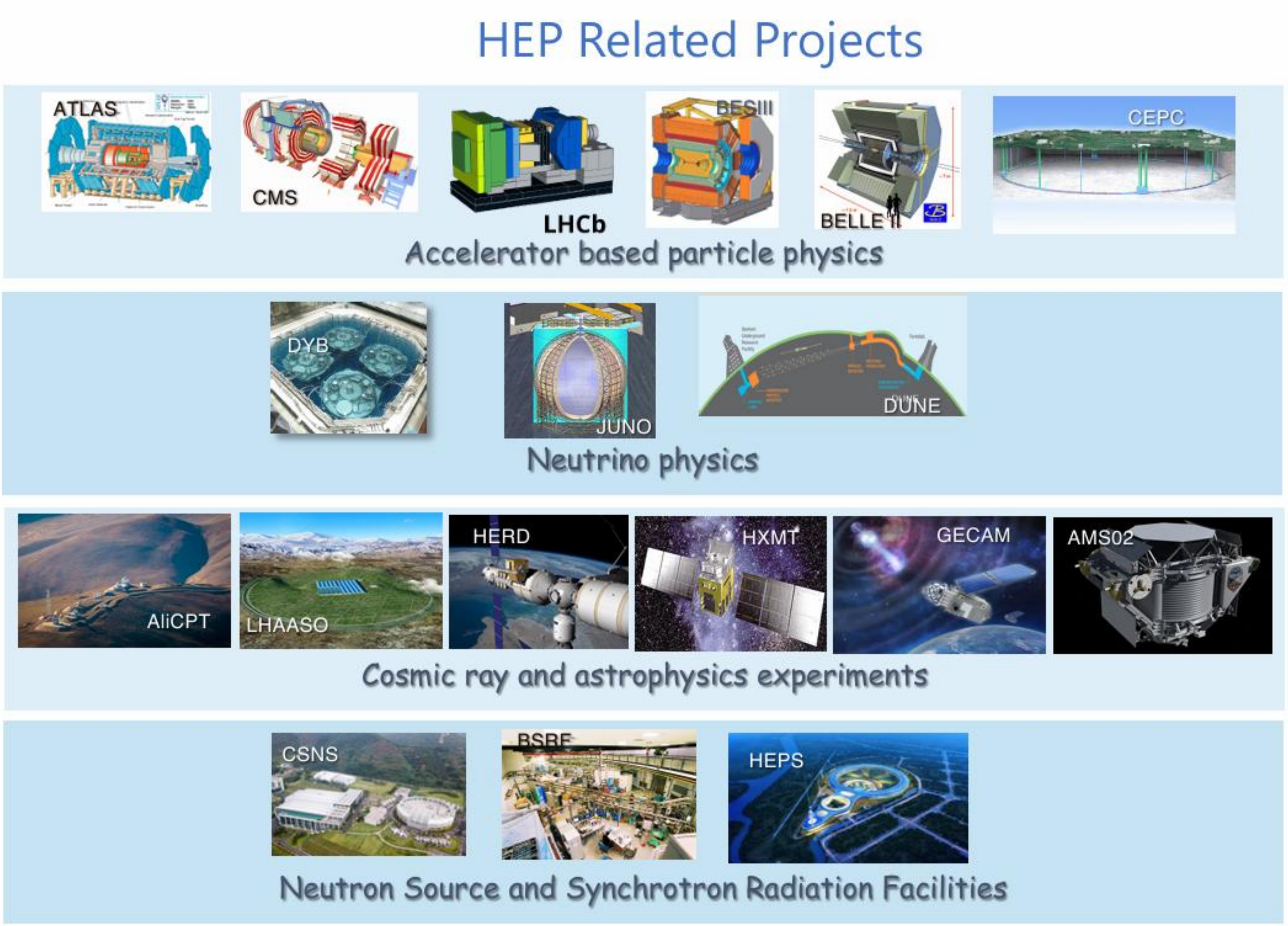
Summary

Brief Introduction to IHEP

Institute of High Energy Physics, Chinese Academic of Sciences

The largest fundamental research center in China.

- Experimental particle physics
- Theoretical particle physics
- Astrophysics and cosmic-rays
- Accelerator technology and applications
- Synchrotron radiation and applications
- Nuclear analysis technique
- Computing and network application


HEP Related Projects
ATLAS
CMS
LHCb
BESIII
BELLE II
CEPC
Accelerator based particle physics
DYB
JUNO
DUNE
Neutrino physics
AliCPT
LHAASO
HERD
HXMT
GECAM
AMS02
Cosmic ray and astrophysics experiments
CSNS
BSRF
HEPS
Neutron Source and Synchrotron Radiation Facilities

IHEP Computing Center

101k CPU cores, 400 GPU cards to for more than 10 exp.

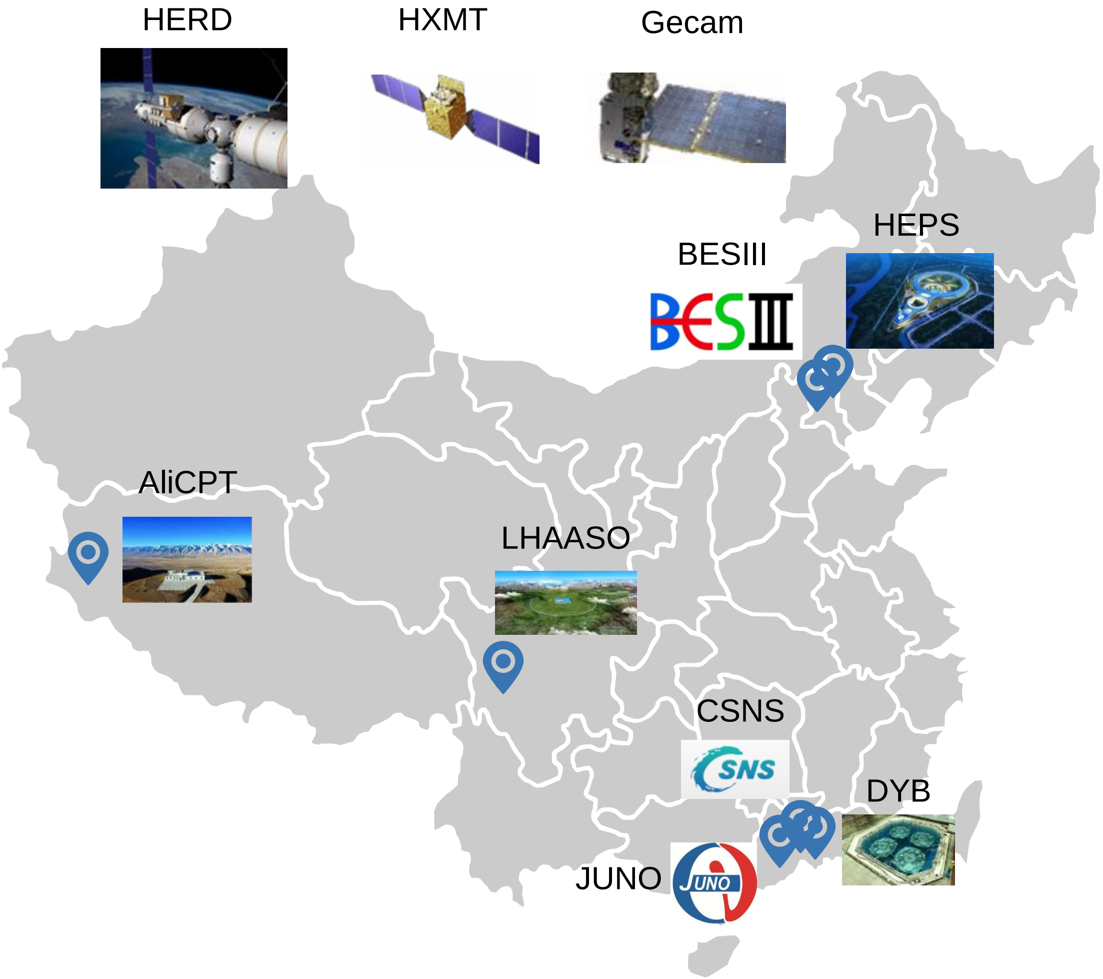
- HTC cluster (~48k CPU cores)
- HPC cluster (~43k CPU cores + 400 GPU Cards)
- Distributed computing, WLCG, Grid etc. (~10k cores at IHEP)

142 PB disk storage, 88 PB tape EOS+CTA storage

- Lustre (41 PB, POSIX, XRootD)
- EOS (101 PB, XRootD)

Network

- 100Gb/s uplink to GEANT
- 800 Gb/s backbone bandwidth inside IHEPCC
- 1.2 Tb/s backbone bandwidth inside HEPSCC



Distributed Computing at IHEP

Distributed computing services at IHEP is mainly for:

- **To develop and deploy services to take advantage of all distributed resources.**
- **To support IHEP experiments, especially the international collaborated experiments with Grid computing resources.**
- **To meet huge computing requirements from experiments that local computing cannot handle.**

Supported experiments:

- **JUNO, Jiangmen Underground Neutrino Observatory**
  - A multi-purpose of neutrino observatory
  - Will start physics data taking since this month (July 2025).
- **HERD, High Energy Radiation Detection Facility**
  - A space particle astrophysics experiments in the Chinese Space Station
  - Will lanch to space station since 2027.
- **CEPC, Circular Electron Positron Collider**
  - A next generation collider for future
  - At the very beginning, everything is in design.

Grid Infrastructure and Services at IHEP

Computing and Storage

Computing:

- HTCondor for local computing, Condor CE for Grid computing.
- Local grid computing resources are split.

Disk Storage:

- Lustre file system and EOS are both used for local storage.
- Lustre storage could be accessed by grid user via a Xrootd service.
- EOS supplied Grid user access with it original Xrootd supports.

Tape storage:

- Only EOS-CTA are available, for both grid and local.

Computing Type	Batch System	Resource
Local	HTCondor	Dedicated
Grid	HTCondor CE	Dedicated

Disk Storage Type	File System	Resource
Local	Lustre	Shared with Grid
	EOS	
Grid	Lustre + Xrootd	Shared with Local
	EOS	

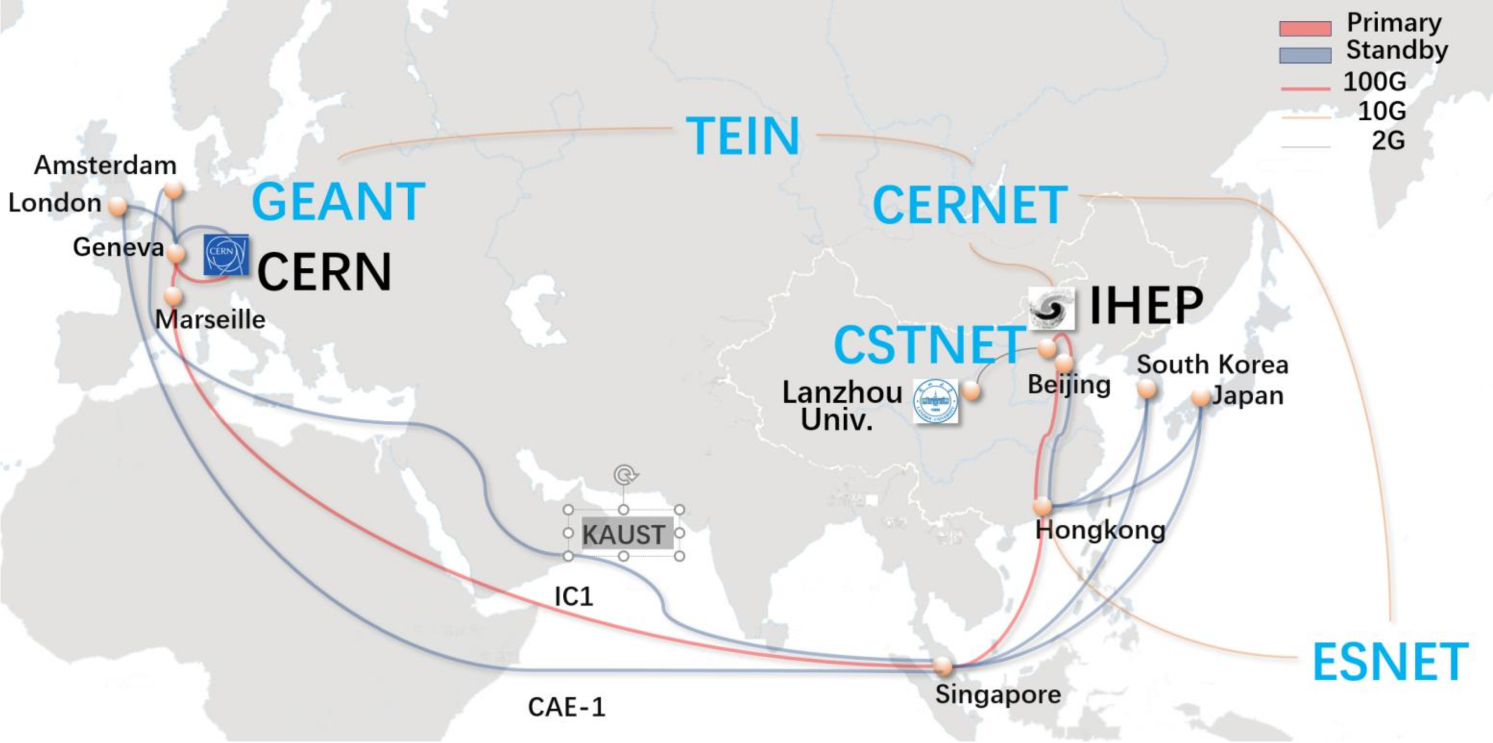
Network Status

100 Gbps shared bandwidth to GEANT for WLCG experiment.

- Sadly IHEP doesn’t have direct network link to JINR.
- All data communication between IHEP and JINR need to pass through GEANT.

Inside data center network,

- Backbone bandwidth,
  - 800 Gbps in total,
  - ~50 Gbps average bandwidth,
- Access switch (SW) status,
  - 100 Gbps port: 432,
  - 25 Gbps port: 1440,
  - 10 Gbps port: 864.



From VOMS-Proxy to WLCG-Token

WLCG steps to phase out VOMS and move to Indigo-IAM as credential service,

- JUNO experiment has legacy VOMS services, and is moving to Indigo-IAM this year.
- HERD/CEPC started from Indigo-IAM at the beginning.

VOMS-importer,

- The legacy VOMS services are imported to Indigo-IAM by VOMS-importer.

Token support from Grid services,

- HTCondor CE at IHEP started to support token-based jobs since 2024, and at present, all JUNO’ s grid services are submitted by Token.
- Storages at IHEP started token support from 2023. TPC transfer and data access in next slide.

Experiment	Legacy VOMS Service	IAM-VOMS	IAM-Token
JUNO	Yes, still working, replicated instances at IHEP, JINR	Yes, synchronized with legacy VOMS	Yes
HERD	No	Yes, supporting for legacy storages	Yes
CEPC	No	No	Yes

Third Party Copy Support For JUNO

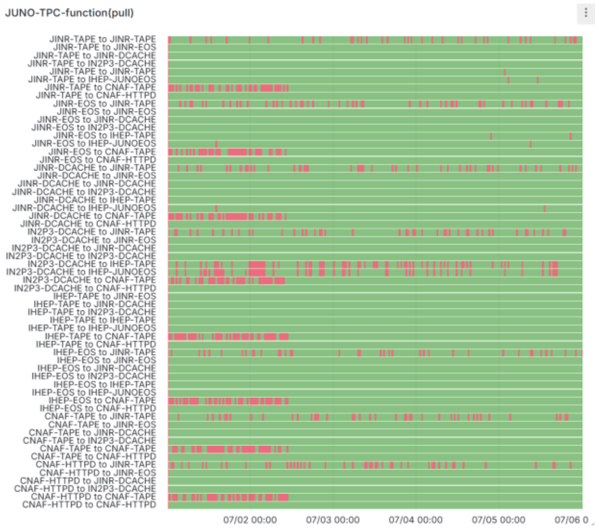
Following the WLCG TPC development, JUNO experiment and its sites,

- Started fully supporting HTTP-TPC since 2023.
- HTTP protocol works as main protocol, providing both data access service and token-based TPC transfer. Supporting Tokens and Macaroon.
- Xrootd protocol works as back-up protocol.

TPC Active Probing System,

- Active probing JUNO TPC function and speed.
- Tests executed by Gfal2 tools, results collected and shown in Elasticsearch-Grafana.
- Function tests: Upload/download, list, remove test in every 30 minutes.
- TPC mode tests: pull/push/streamed mode test in ever 30 minutes.
- Transfer performance tests in ever 2 hours.

JUNO Data Center	Storage System	Access Protocols	Token-based TPC Support	Available Tokens
CNAF	StoRM	HTTP, Xrootd, SRM	Yes, since 2021	WLCG-Token, Macaroon
IHEP	EOS/Xrootd	HTTP, Xrootd	Yes, since 2022	WLCG-Token, Macaroon
CC-IN2P3	dCache	HTTP, Xrootd	Yes, since 2022	WLCG-Token, Macaroon
JINR	dCache	HTTP, Xrootd	Yes, since 2023	WLCG-Token, Macaroon



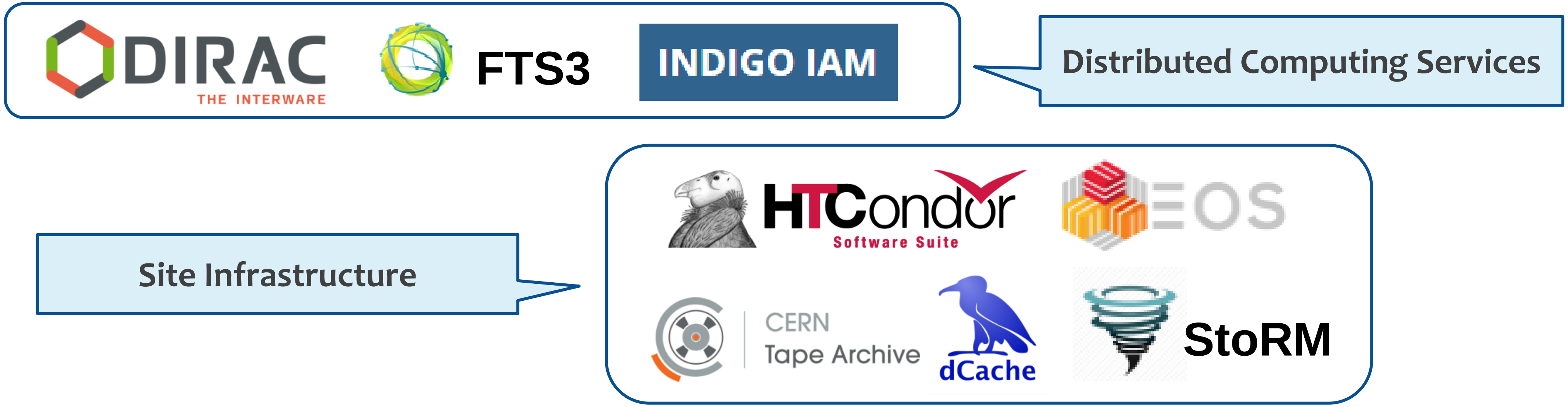
Distributed Computing Monitoring

Hundreds of metrics to monitor for each experiments:

- Passive collection: Computing site and storage site status and accounting, services status, etc.
- Active probing: Basic functions and performances of sites, which are not provided by sites.

A monitoring system with full monitoring data collection and probing are developed:

- Design based on workflow application.
- Refer to Xiao Han’ s talk at https://indico.jinr.ru/event/5170/contributions/31709/.



Distributed Computing Services

Site Infrastructure

Distributed Computing System -- DIRAC

DIRAC: Distributed Infrastructure with Remote Agent Control

- First developed for LHCb. Widely used in Belle2, ILC, CTA.

Middle layer between users and resources.

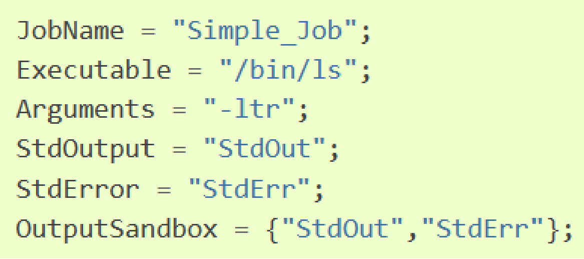
- User interface: API, REST, Web, CLI.
- Computing management: Grid, Cloud, Cluster.

Job management in DIRAC:

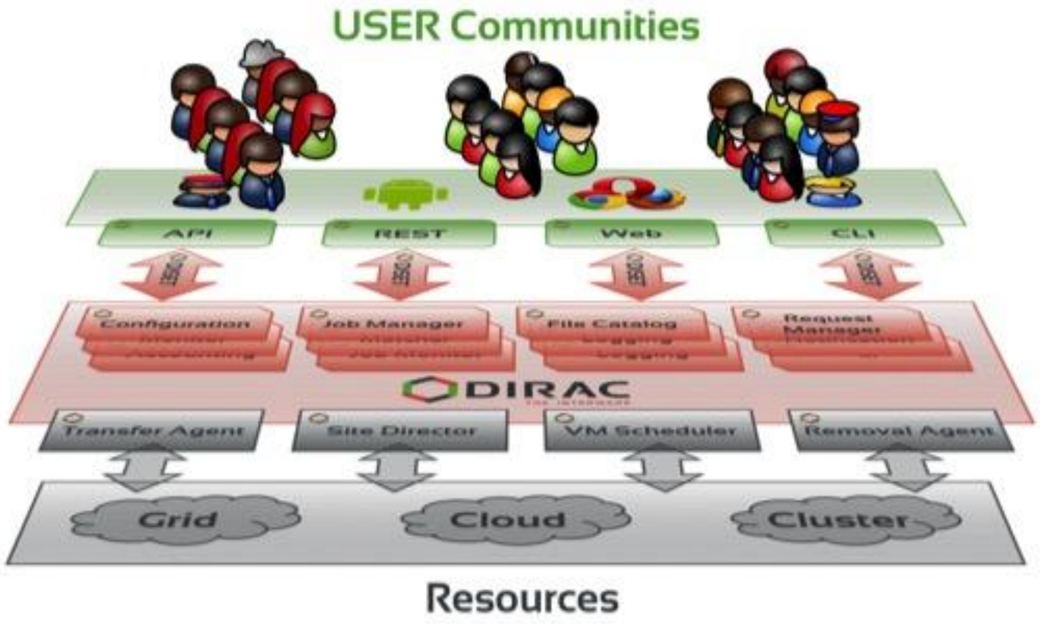
- Job submit: JDL job.
- Job schedule: Pilot job.

Will or already in

- JUNO, HERD, CEPC.



```
JobName = "Simple_Job";
Executable = "/bin/ls";
Arguments = "-ltr";
StdOutput = "StdOut";
StdError = "StdErr";
OutputSandbox = {"StdOut","StdErr"};
```



Data Management System -- RUCIO

Rucio system,

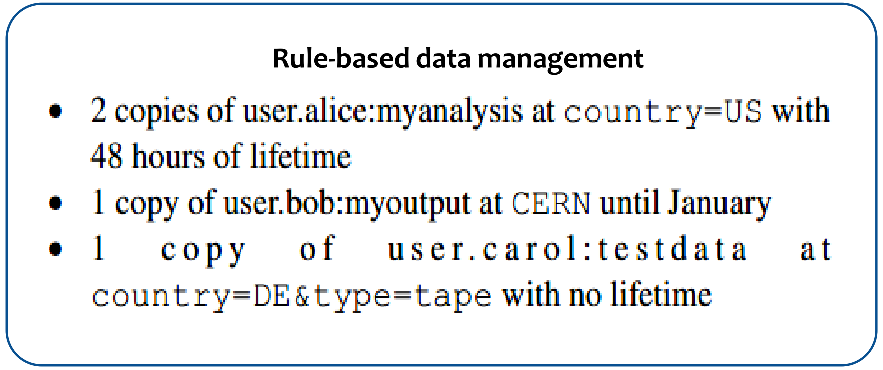
- To unified data distribution and management for heterogeneous storage systems across data-intensive scientific collaborations.
- Originally designed for the ATLAS experiment.
- Now widely adopted in high-energy physics, astronomy, biology, and other scientific fields.

Compare with other tranditional data management system,

- High Scalability
- Extreme Extensibility
- Open-source & Developer-Friendly
- Asynchronous Big Data Processing
- Rule-based Data Management

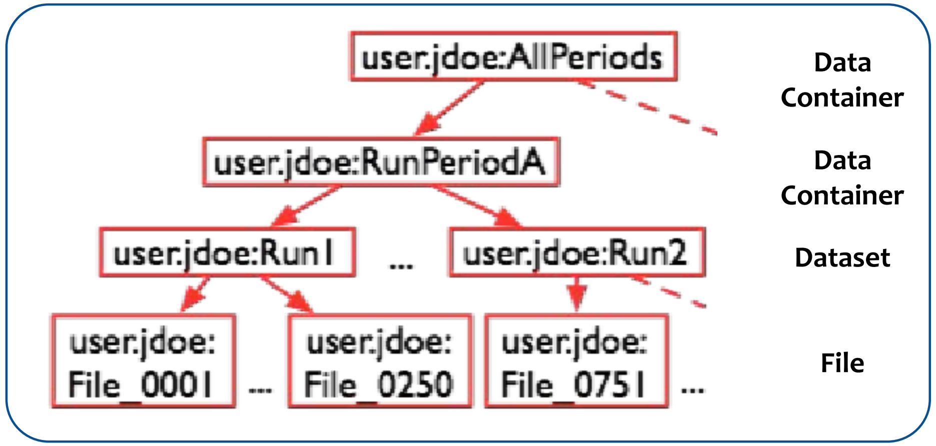
Will or already in

- HERD, CEPC.



Rule-based data management

- 2 copies of user.alice:myanalysis at `country=US` with 48 hours of lifetime
- 1 copy of user.bob:myoutput at `CERN` until January
- 1 copy of `user.carol:testdata` at `country=DE&type=tape` with no lifetime



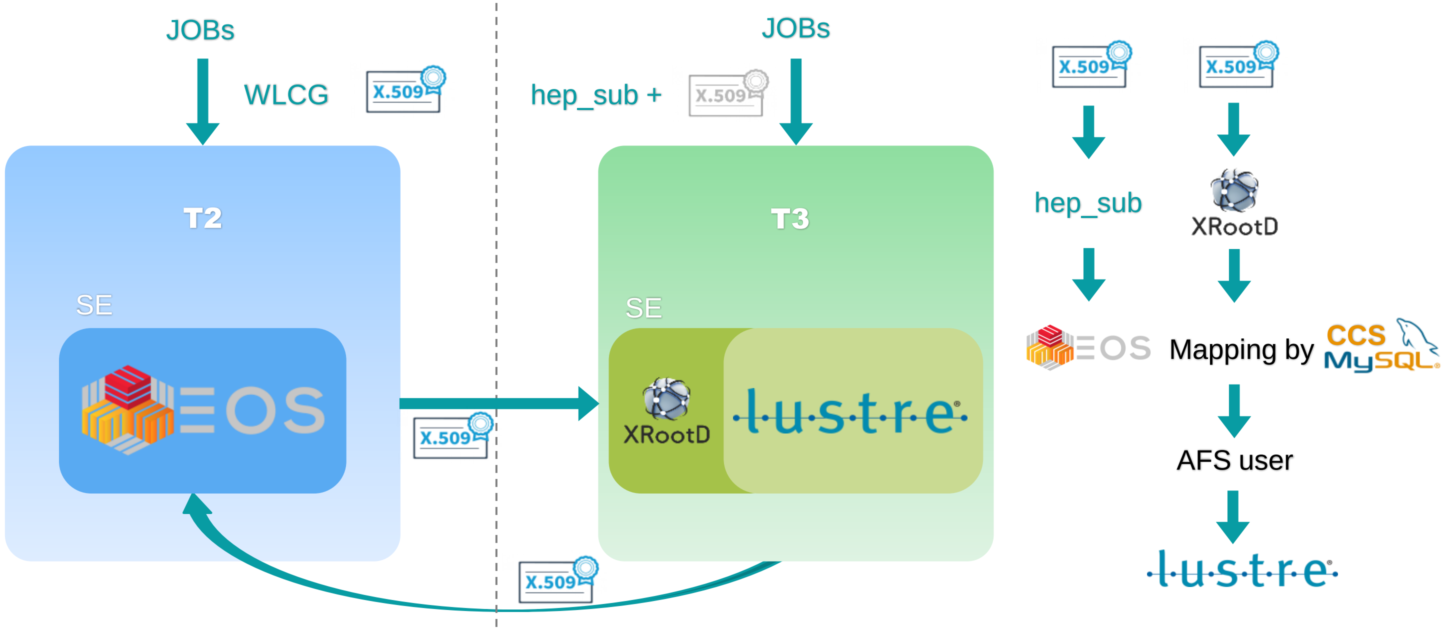
Some Tools…

Some tools were developed for better using Grid services at IHEP.

✓**Grid plugin of HepSub:** A command line tools for user submit local jobs with grid credential at IHEP.

✓**ihepPyIAM:** A python command line and package for manage IAM user data at IHEP.

✓**WLCG Tier2 computing and Local computing bi-directory data accessibility:** A simple framework for LHC user to access Tier2 data in local jobs and access local data in Tier2 jobs at IHEP.



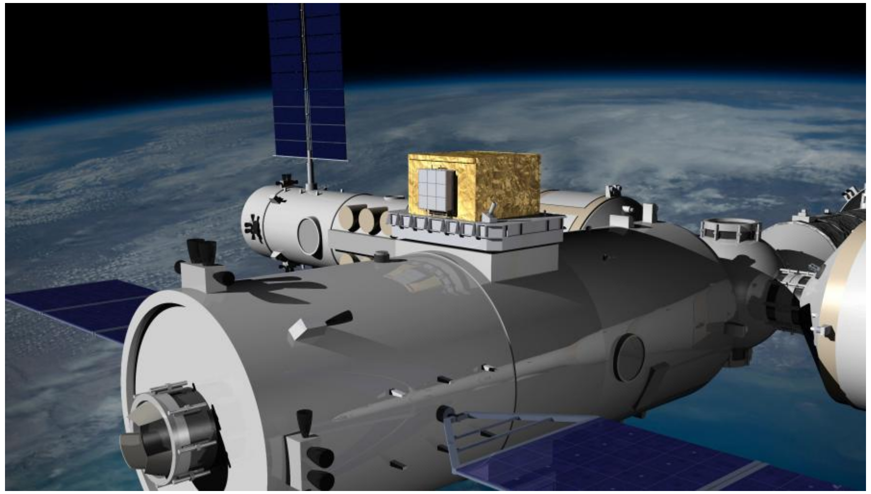
Distributed computing system for HERD experiment

HERD Experiment

High Energy cosmic-Radiation Detection facility (HERD),

- A space particle astrophysics experiments, will run in the Chinese Space Station for >10 years since 2027.
- HERD is a international collaborated experiment with around 47 institutes, labs and universities from China, Italy, Switzerland, Spain and Sweden.

Science Goal	Type	Contribution to Physics	Methods
Precision measurement of cosmic ray electron flux and dark matter search	Core	Key contribution to solve one of the most important puzzle for astronomy and physics: dark matter	Precision flus measurement of high energy electron and gamma.
Origin, acceleration and propagation of cosmic rays	Core	Key contribution to the origin of cosmic rays	Measurement of cosmic ray nuclei up to Z=28 to the highest energy
High energy gramma rays all-sky survey and monitoring		Search and identify gamma ray source , understand the physics of extreme conditions in the universe; search for new physical signals	Wide energy range, High precision measurement of Gamma rays



HERD Computing Requirement

Storage resources:

- >30PB in 5 years,
- >90PB in 10 years.

Computing resources:

- ~7500 CPU cores in 5 years,
- ~16000 CPU cores in 10 years.

Network and data transfer:

- 10-100 Gbps.

Data type	Data size（PB）			Computing（CPU Core）		
	5 years	10 years	Site	5 year	10 year	Site
Flight Data	2	6	T0, T1	-	-	T0
Standard Reconstruction	2.5	7.5	T0, T1	200	400	T0
Data transmission control system	1	2	T0	300	600	T0
PassN reconstruction	5（2 version）	15（2 version）	T0, T1	1000	3000	T0
Simulation data	5	15	T0, T1	4000	8000	T0
Analysis Data	2	4	T1	2000	4000	T1
Summary	15.5+16.5	45.5+47.5		7500	16000	

Data processing challenges:

- Need to distribute RAW data from China Space Station to CN and EU data centers.
- Need to schedule multiple data process tasks among CN and EU data centers.
- Need to provide uniform user authentication and resources permission management system.

So, we need to build infrastructure for HERD computing.

3 Layers of HERD Computing

User Interface:

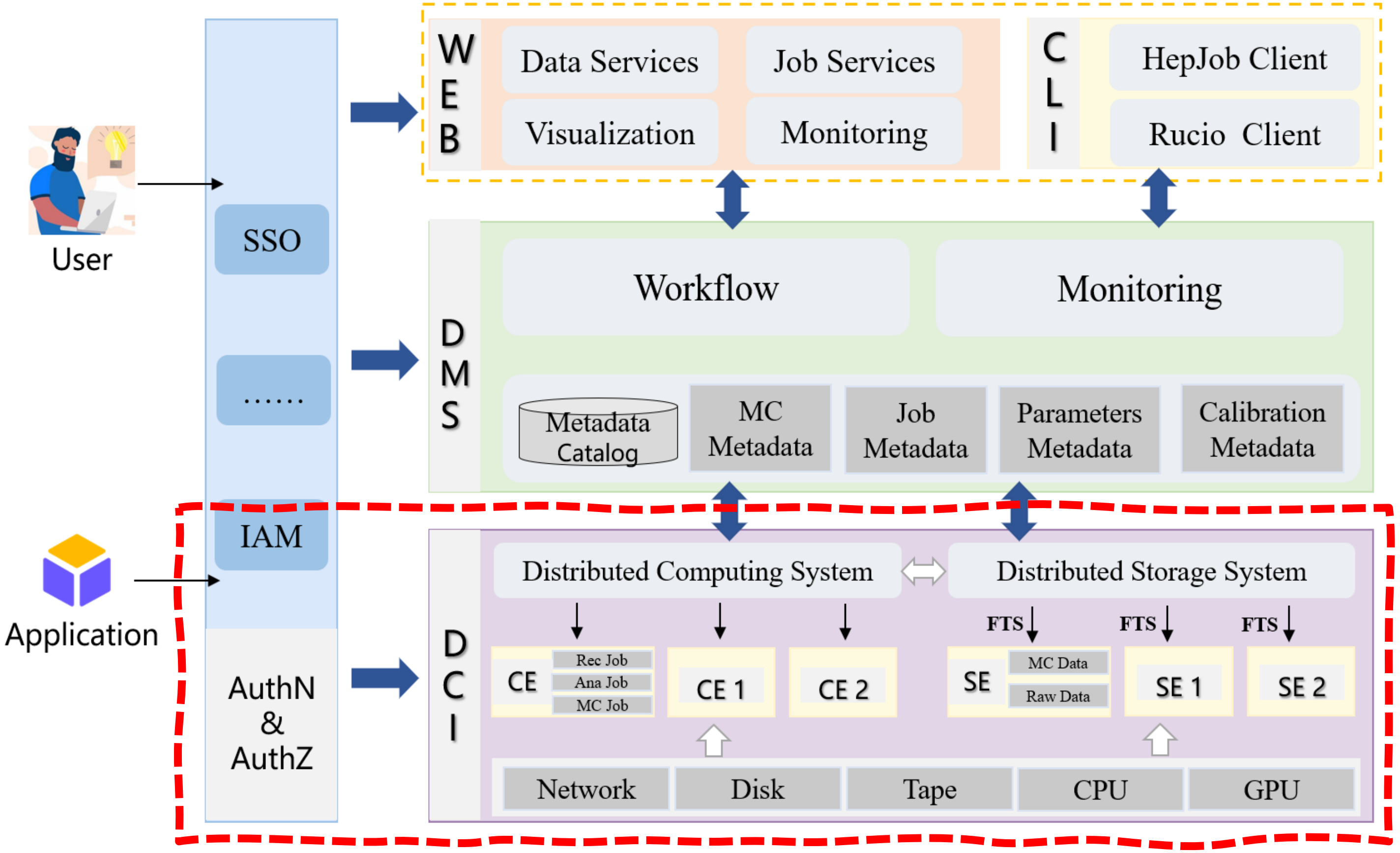
- Provide web UI and CLI for user.
- Trigger data process, analysis, monitoring task by user.

Data Management System (DMS):

- Manage data processing tasks.
- Provides metadata database.

Distributed Computing Infrastructure (DCI):

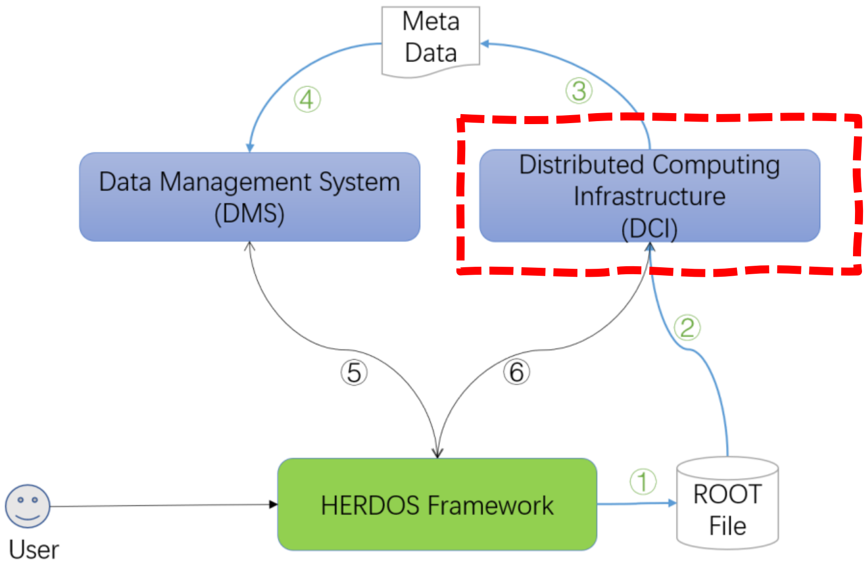
- Manage computing/storage resources.
- Executing data processing tasks.



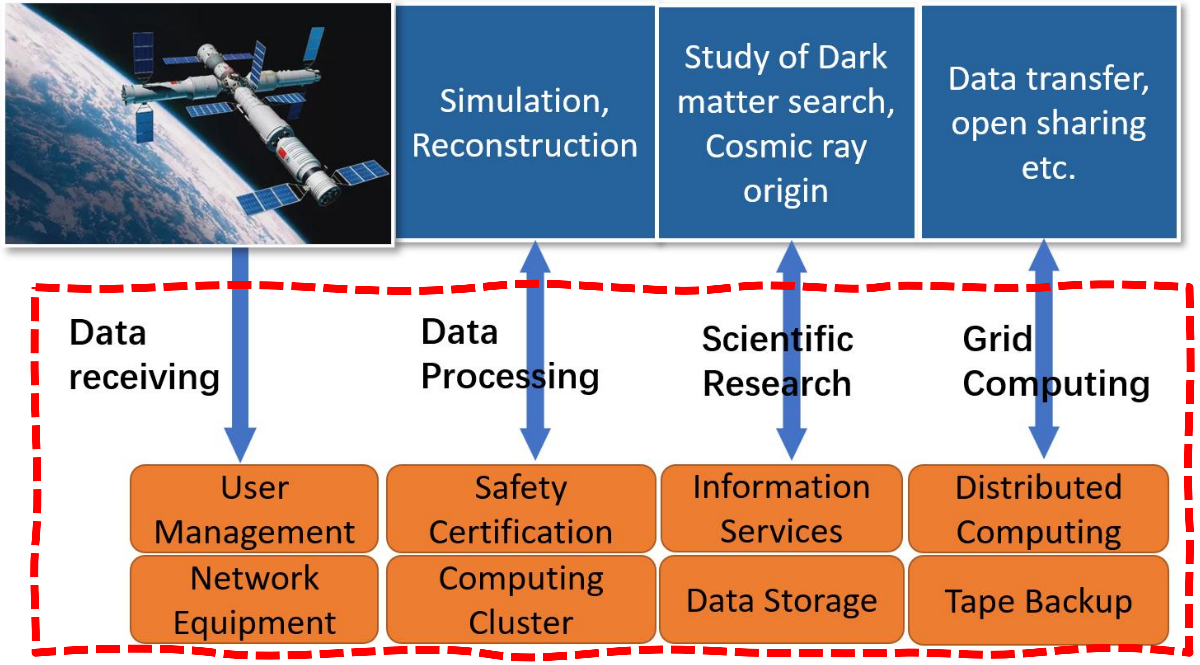
Distributed Computing Infrastructure

HERD DCI is a distributed computing system for,

- **Data processing** -> Distributed computing system
- **Data access** -> Distributed data management system
- **Data distribution** -> Network and data transfer system
- **Data privilege management** -> Authentication and authorization system



Meta Data
④
③
Data Management System (DMS)
Distributed Computing Infrastructure (DCI)
②
⑤
⑥
User
HERDOS Framework
①
ROOT File



Scientific Data Platform

Computing Model of HERD

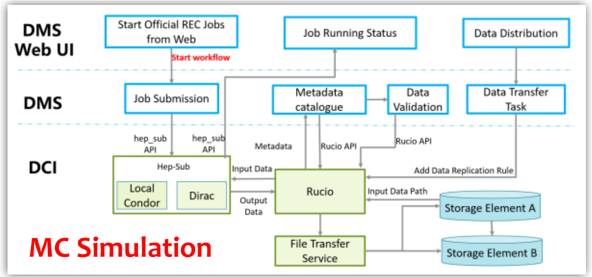
DCI supports all data processing with computing and storage resources.

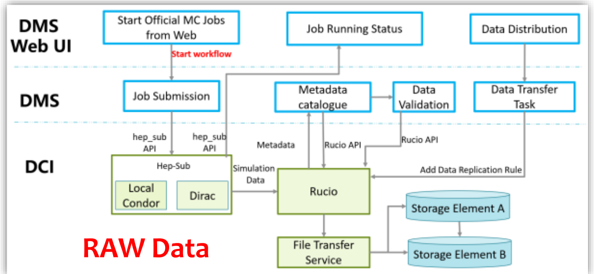
Multiple data process workflows,

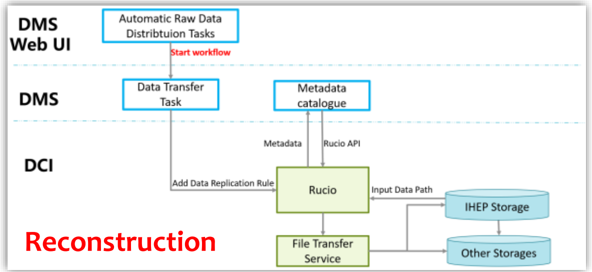
- MC Simulation,
- RAW Data,
- Reconstruction,
- User Analysis.

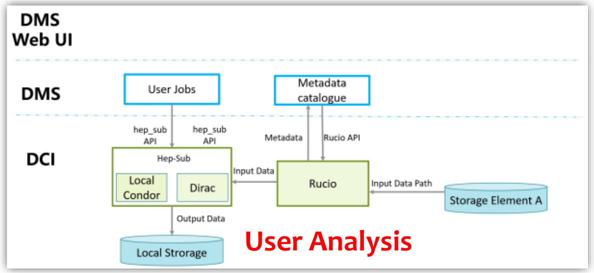
Tier model for data process,

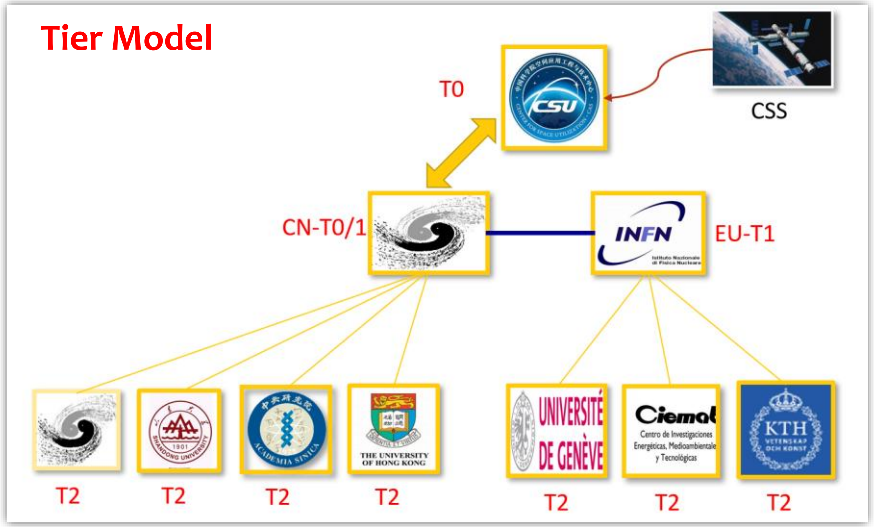
- **Tier-0 sites: Central site.**
  - CSU: Raw data acquisition from Space Station.
  - IHEP: All types data storage and data distribution source.
- **Tier-1 site: Regional center site**
  - IHEP(CN-T1)/INFN(EU-T1): SIM and REC data storage, computing resources.
- **Tier-2 site: simulation data processing**











Subsystems in DCI

Distributed Computing,

- Manage distributed computing resources and execute computing tasks.

Distributed Storage,

- Manage distributed storage and storage raw and processed data.

Data Transfer and network,

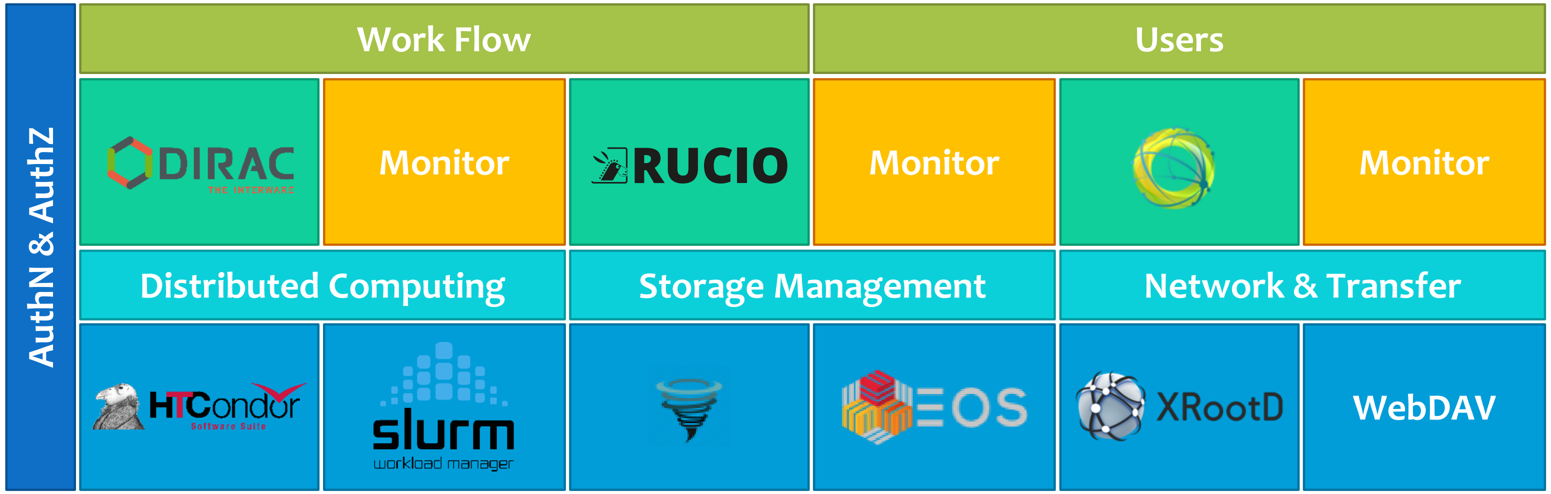
- Manage data transfer tasks.

AuthN & AuthZ,

- Manage user permissions in data processing.

Resources,

- Computing clusters and disk/tape storage in each data centers.



AuthN & AuthZ

Work Flow

Users

Monitor

Monitor

Monitor

Distributed Computing

Storage Management

Network & Transfer

WebDAV

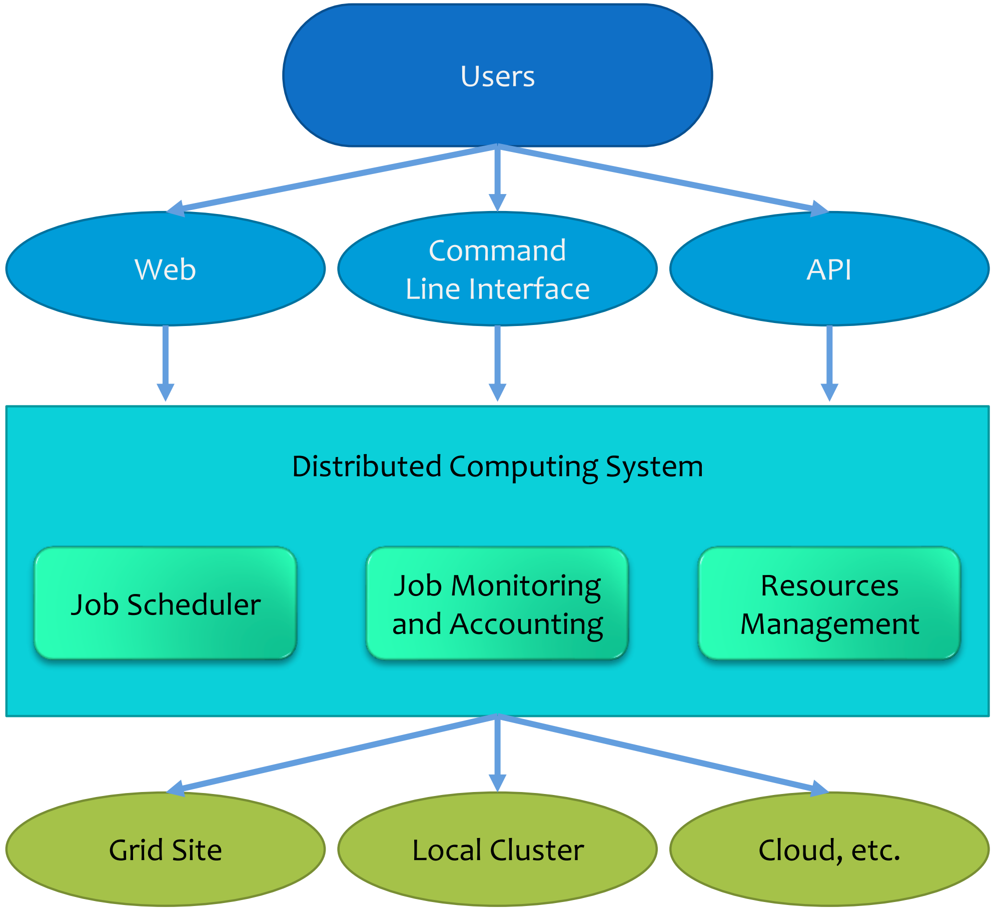
Computing Management Design

To users:

- **Unify computing sites** with heterogeneous computing systems.
- Allow to use HTCondor, Slurm, cloud computing, supercomputing, local cluster, etc. **based on data processing tasks**.
- Supply **unified job management interface**.
- By Web, Command line interface and APIs.

To computing sites:

- **Schedule jobs** in computing resources.
- **Optimize jobs distribution** among sites.
- Monitoring computing resources status.
- Generate site reports and accounting sites usage.



Structure of Computing System

"One entrance, all computing tasks"

Distributed computing system

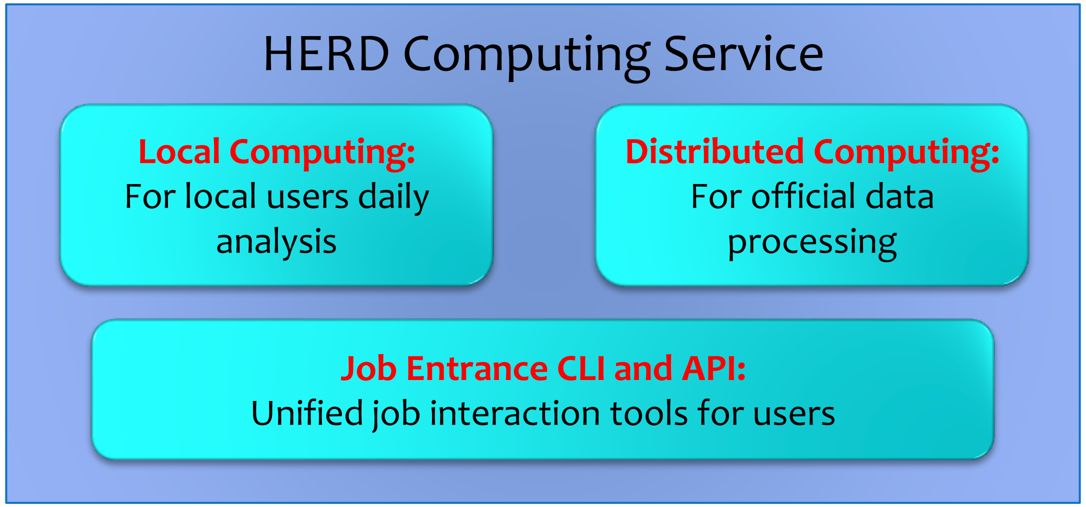
- Distributed computing resources around the world.
- For **official data processing** and special tasks.

Local computing system

- Local computing resources in each data center.
- For **local users analysis** tasks.

Unified job entrance tools

- Provide a unified job entrance tool for all type of data processing and analysis jobs.



HERD Computing Service

Local Computing:
For local users daily analysis

Distributed Computing:
For official data processing

Job Entrance CLI and API:
Unified job interaction tools for users

Local Computing – HTC/HPC

Type of HERD computing jobs includes,

- **Single-core job or multi-core job within one node**: simulation, reconstruction, analysis,
- **Multi-core job on multi node or GPU job**: part of reconstruction, AI training.

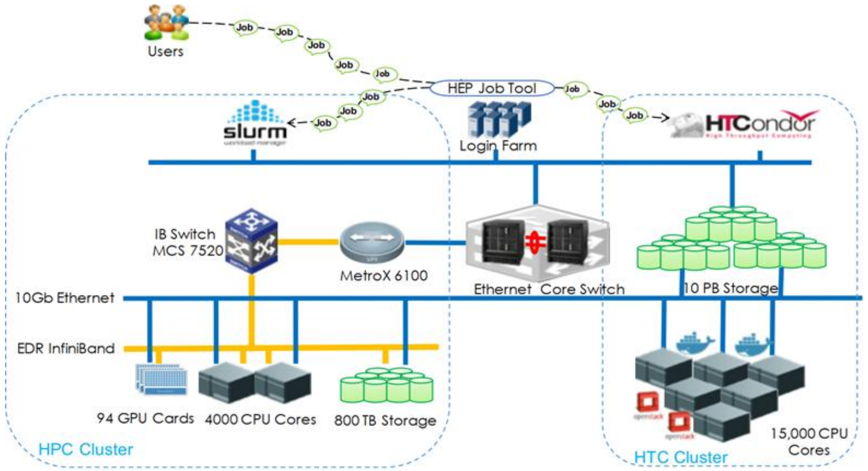
High throughput computing: single-core or multi-core within one node,

- HERD **data processing** is a typical high throughput computing with hundreds of thousand jobs,
- HERD HTC cluster is **based on HTCondor** which is a open-source high throughput computing software suite.

High performance computing: big multi-core job or GPU job

- Part of HERD **reconstruction data processing** is using AI to driven,
- HERD HPC cluster is **based on Slurm**.

dHTC for share resources between HTC and HPC



Job Entrance – HepSub Tools

One Job Entrance is a job APIs, based on HepSub tools:

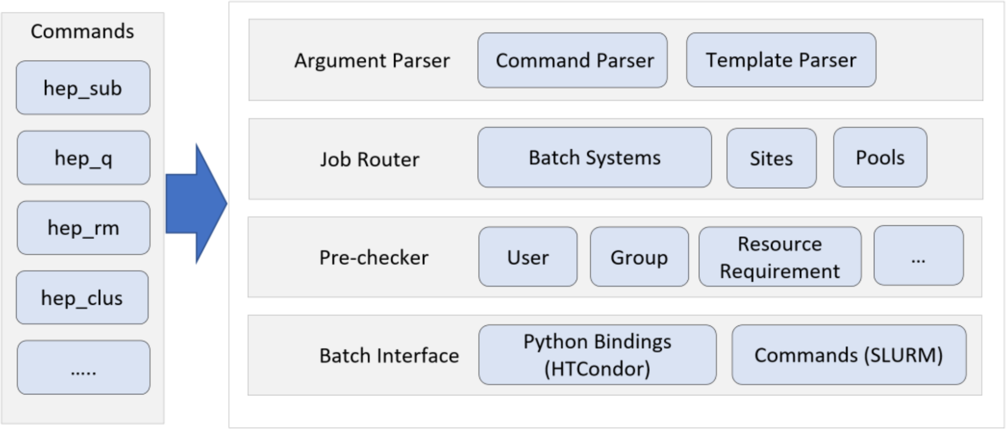
- Support **unified job submission endpoint and interface**, no matter of user jobs or production jobs.
- Support **both Grid jobs (DIRAC jobs) and cluster jobs** submission.
- Support **Grid data access and management**.
- Flexible enough for **integrating to other job services**, such as HERD data production system service, authentication system, monitoring and accounting system, common user interfaces…

HepSub is a job submission tool developed by IHEP,

- Already unified HTCondor/Slurm cluster job operations: submission, query, remove, etc.

HepSub is under development to extent to submit DIRAC jobs.

- To **support DIRAC JDL** format translation,
- To **support IAM** with X.509 certificate and Sci-Token authentication.



Storage Management Design

HERD Grid storage management,

- To **produce and distribute data** from distributed computing and storage sites.
- To **manage distributed data access quests** from DMS system.
- Based on **Rucio system**, a popular grid data management system in HEP.

Storage management services manages data production,

- **Raw data distribution**, from IHEP to Chinese and European Storage Sites.
- **Processed data distribution**, replicate among Tier1 and Tier2 sites.
- **Official data operation**: adding, deleting, modifying, querying in distributed storage sites.

For normal users,

- Supply data access in developed HERD Software APIs and CLI command, normal user could get official processed data by HERD Software.

Structure of Storage Management

“One API, all data management tasks”

Storage Management System:

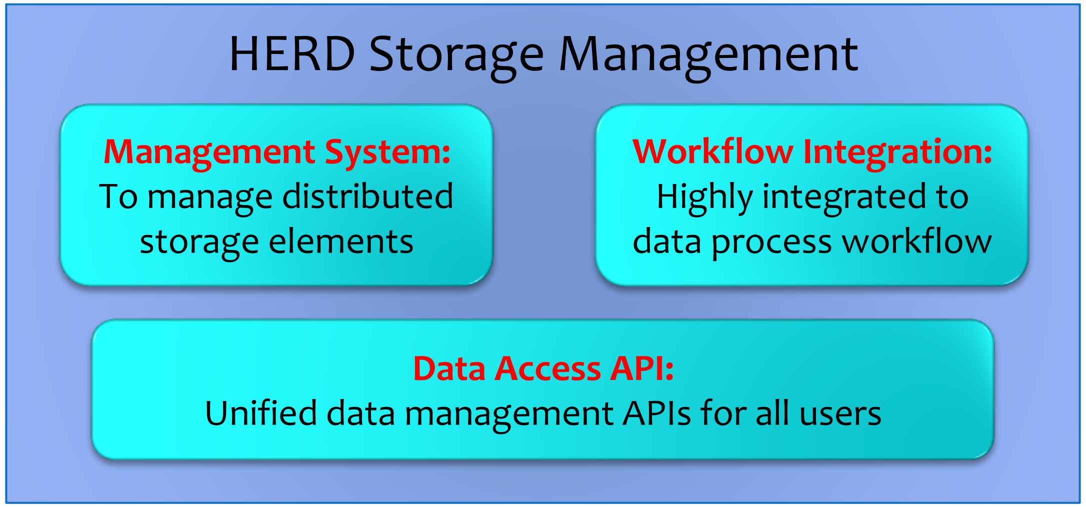
- **Based on Rucio** to manage distributed disk and tape storages,
- Develop **HERD customized file catalog policy** and **HERD Rucio plugins**.

Workflow Integration:

- Highly **integrated to HERD software** and data process workflow.
- Allow to be used in all tasks.

Data Operation API:

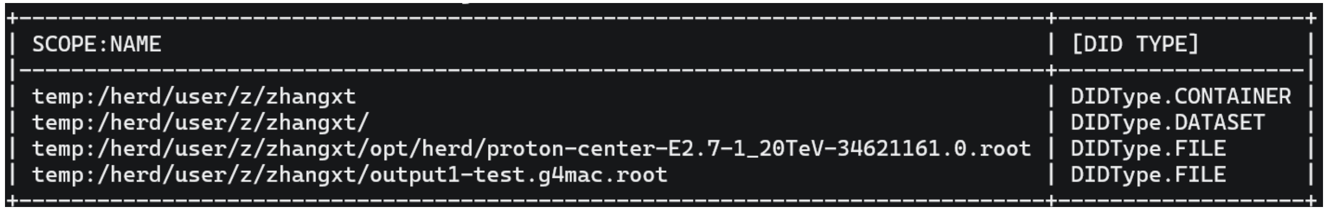
- Develop a **API for data operation** for both users and workflow tasks.


HERD Storage Management
Management System:
To manage distributed storage elements
Workflow Integration:
Highly integrated to data process workflow
Data Access API:
Unified data management APIs for all users

Distributed Storage Management – Rucio

Rucio with customized grid data file catalog namespace,

- **To make data logic name closer to local data, follow normal POSIX rules, has 3 types:**
  - Data container, to contain other containers and datasets,
  - Dataset, to collect files,
  - Data file, basically ROOT files.
- **Scopes are working as data status zones, so data types could be distinguished by its name,**
  - Temp, Valid, Corrupt, etc.



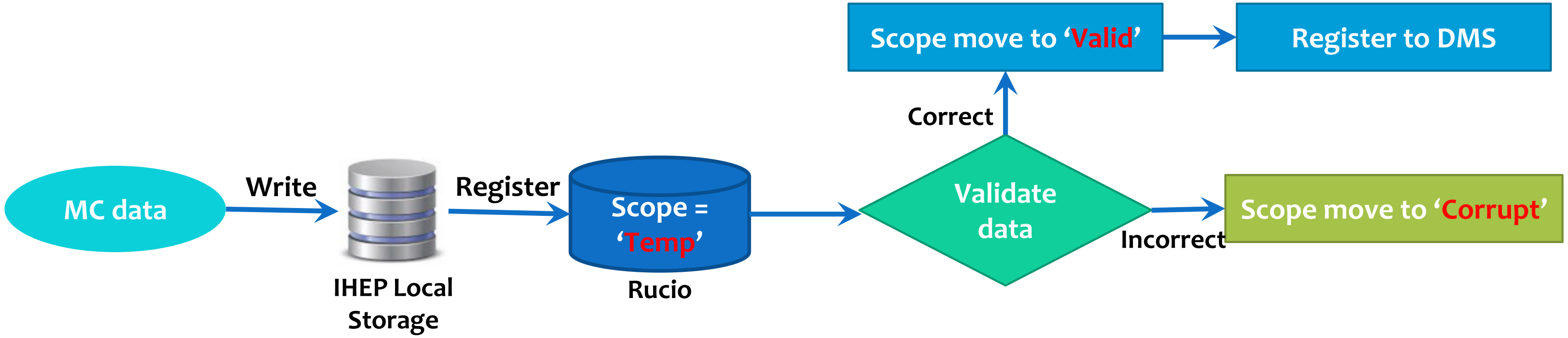
```
+-------------------------------------------------------------------------------+-------------------+
| SCOPE:NAME                                                                    | [DID TYPE]        |
|-------------------------------------------------------------------------------+-------------------|
| temp:/herd/user/z/zhangxt                                                     | DIDType.CONTAINER |
| temp:/herd/user/z/zhangxt/                                                    | DIDType.DATASET   |
| temp:/herd/user/z/zhangxt/opt/herd/proton-center-E2.7-1_20TeV-34621161.0.root | DIDType.FILE      |
| temp:/herd/user/z/zhangxt/output1-test.g4mac.root                             | DIDType.FILE      |
+-------------------------------------------------------------------------------+-------------------+
```

Namespace Component	HERD Namespace Policy
Name	Linux-like directory and file path
Scope	Defined as data status in data flow(Temp, Valid, Corrupt)
Dataset	Collection of all Files in a directory
Container	Collection of all sub-directories (=datasets) in a directory

HERD Dataflow Integration

HERD data processing are all based on workflow,

- **For Workflow system:**
  - Synchronize file catalogs from Rucio.
  - Trigger data processing jobs. In these jobs, data are downloaded/uploaded by Rucio.
  - Trigger data validation and data distribution in Rucio.
- **MC dataflow as an example:**
  - Register all raw MC data to 'Temp' scope,
  - Data validation program use APIs to validate whether data are good.
  - If good, move scope to 'Valid' , then provide it to metadata registering.
  - If not good, move scope to 'Corrupt' scope, waiting for deletion.



HERD Data Operation API

We are developing a HERD workflow-oriented API,

- For both experiment software data access in jobs and workflow.
- Merged to HERD software and workflow system.

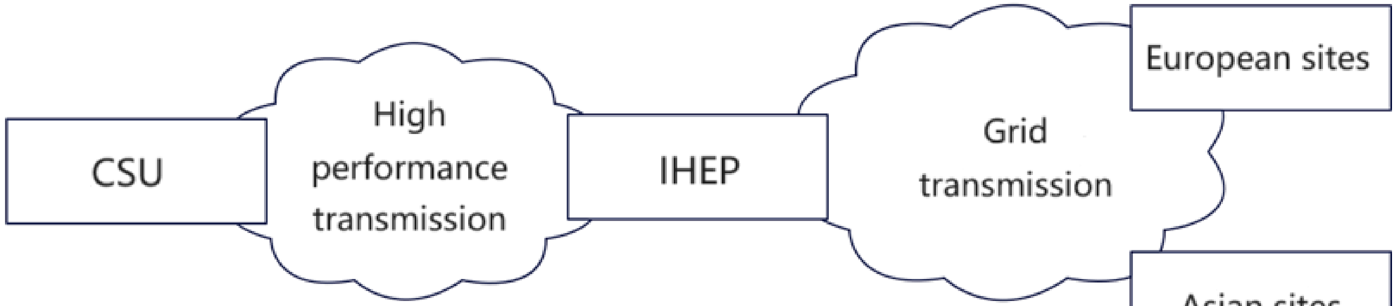
API can provide methods for:

- **Formatted metadata methods** for workflow system, keys includes:
  - Production batch, log file path, job finished time, etc.
  - Which could not got directly from remote jobs.
- **Method not directly provided from Rucio** commands:
  - Scope modification.
  - File removal.
  - Batch files upload with divided backend jobs or submit to local computing cluster.
  - Automatic container creation based on ‘HERD’ policy.
- **Some daemons**:
  - Automatic account synchronizer from IHEP-SSO and HERD-IAM.
  - Automatic register and rules creation (under development).
- Other common Rucio methods but packaged in a better model for HERD production.

Network and Data Transfer

Network link should be established between HERD data centers,

- A high-performance data transmission network of **no less than 10Gbps** from CSU to IHEP.
- Network between IHEP and other European sites share a **100 Gbps link by LHCONE**.
- IHEP and other Chinese sites shared **100 Gbps link by CSTNet and CERNet**.



10 Gbps
dedicated network

100 Gbps shared by LHCONE

100 Gbps shared by CSTNet
100 Gbps shared by CERNet

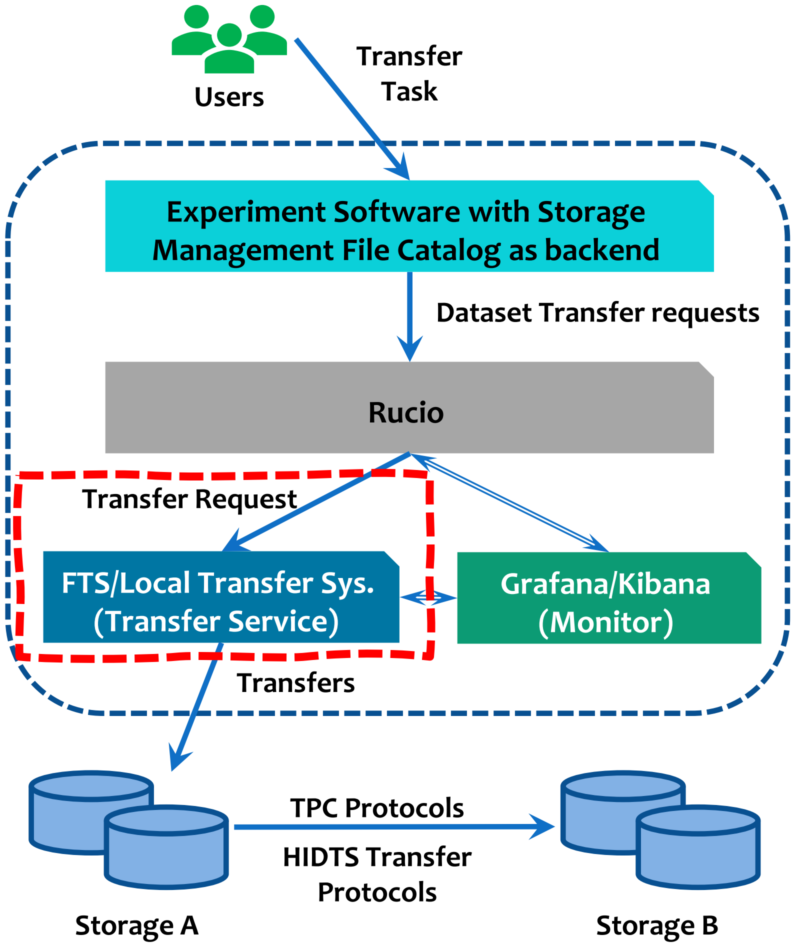
Rucio Transfer Plugins

We are also developing some Rucio Transfer Plugins for IHEP local data transfer system (HiDTS).

- **Working as another FTS plugins** but for HiDTS, an IHEP self-developed data transfer system.
  - HiDTS uses a commercial data transfer system as backend,
  - But IHEP-CC developed a RESTful API for HiDTS, allowing user submits local storage data transfer.
- **We are developing the plugins with allow Rucio use HiDTS as transfer system.**

So that we could support more local storages,

- **IHEP has lots of storage sites not supporting normal protocol such as Xrootd or WebDAV…**
- **Serving for future non-WLCG type experiments or big science devices.**


Users
Transfer Task
Experiment Software with Storage Management File Catalog as backend
Dataset Transfer requests
Rucio
Transfer Request
FTS/Local Transfer Sys. (Transfer Service)
Grafana/Kibana (Monitor)
Transfers
TPC Protocols
HIDTS Transfer Protocols
Storage A
Storage B

Summary

At IHEP, distributed computing services are developed and deployed for JUNO/HERD/CEPC experiments.

We are following the up-to-date Grid technique and developing customized services at IHEP.

HERD experiment plan to processed 90PB dat in next 10 years since 2027.

- "One entrance, all computing tasks"
- "One API, all data management tasks"

The distributed computing services at IHEP are extending,

- To support more experiments in the future.
- To engage more local system and technique into IHEP distributed computing system.
- To apply distributed computing technique and services to other local experiments.

Thanks for your attention