Status of IHEP Grid Data Management System

Xuantong Zhang, on behalf of Grid Computing Group

Computing Center, Institute of High Energy Physics, CAS

Introduction

CC-IHEP, CAS:

- The first and largest Grid Site in Mainland China,
  - Serving both Chinese located and WLCG experiments.
  - 58K CPU cores, 250 GPUs,
  - 97.4 PB disk storage, 80 PB tape storage.
- Grid Data Management system of IHEP:
  - Mainly serves IHEP driven experiments.
  - Service in running:
    - JUNO, >3.0 PB/y raw & simulation data.
    - HERD, ~45.5 PB/10y data.
  - Service in construction or plan:
    - CEPC, under construction.
    - HXMT, plan to use Rucio for Grid data management.
    - LHAASO, plan to accept distributed computing & storage resource.

Chinese located or IHEP driven experiments

BESIII (Beijing Spectrometer III at BEPCII)

JUNO (Jiangmeng Underground Neutrino Observatory)

HXMT (Hard X-Ray Moderate Telescope)

CSNS (China Spallation Neutron Source)

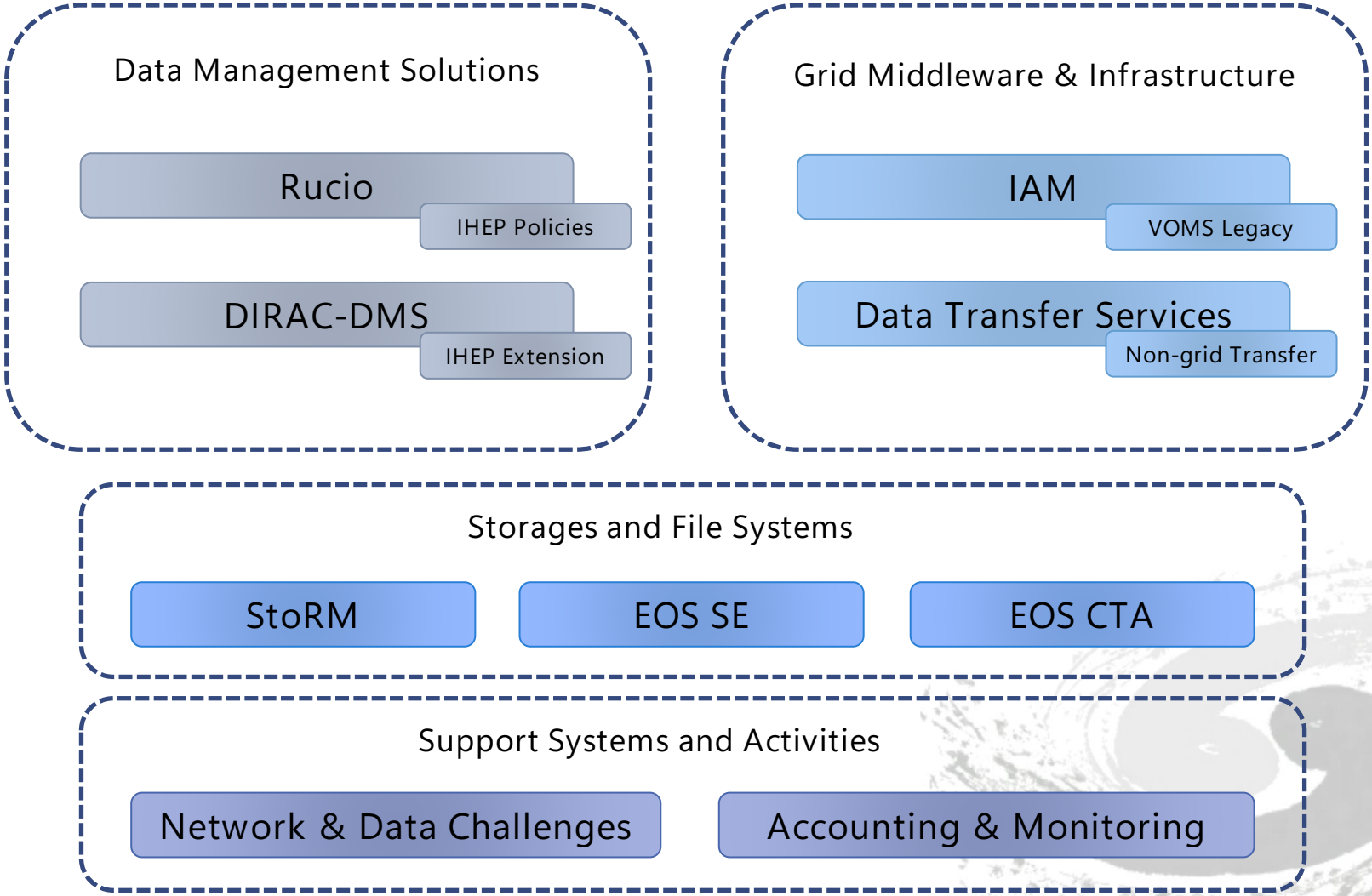
LHAASO (Large High Altitude Air Shower Observatory)

HEPS (High Energy Photon Source)

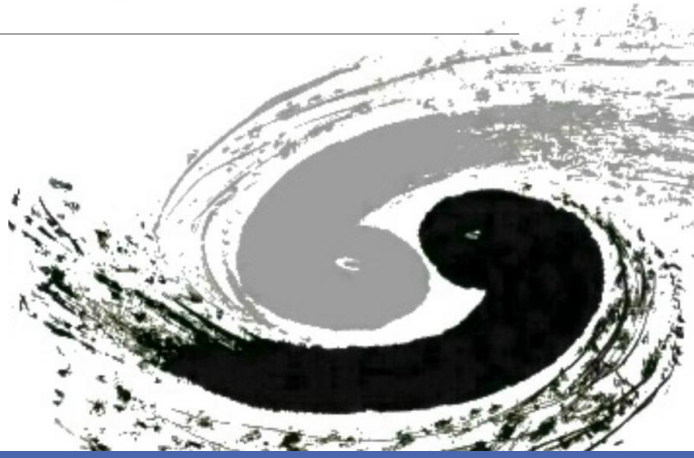
HERD (High Energy Cosmic Radiation Detection)

CEPC (Circular Electron Positron Collider)

System Overview



Data Management Solutions



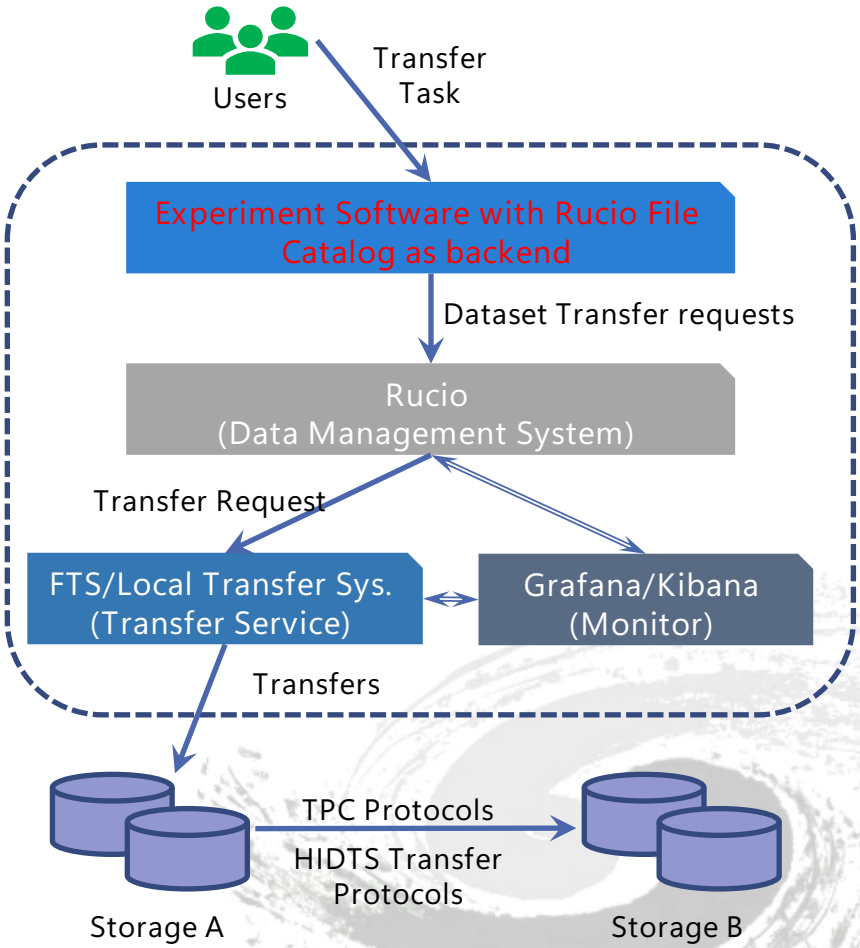
Rucio Based Solution

Rucio Data Management system:

- Supports HERD experiment,
- Preparing for HXMT, LHAASO, etc..

We developed Rucio solution under our experiments needs:

- Deeply integrated to different experiment software, work as a backend service behind it.
- Customized data logic catalogs for different experiment data structure.
- Developed experiment users-oriented APIs which is developed for data access.
- Highly involved with local data transfer system (HIDTS).



Rucio Policies and Plugins at IHEP

Rucio DID policies,

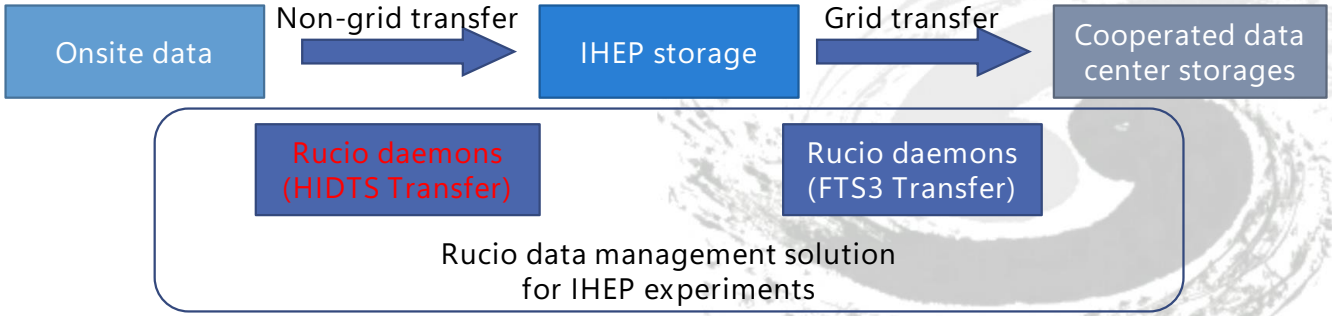
- Make data logic name closer to local data.
- Use “Prefix + local file” name style.
- Distinguish files and datasets by name.

```
+-------------------------------------------------------------------------------+-------------------+
| 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      |
+-------------------------------------------------------------------------------+-------------------+
```

Rucio DID	Rucio DID policy for HERD experiments
Name	Linux-like directory and file path
Scope	Defined as data status in data flow
Dataset	Collection of all Files in a directory
Container	Collection of all sub-directories (=datasets) in a directory

IHEP HIDTS Plugins for Rucio:

- IHEP HIDTS is a non-grid data transfer service for local storage sites.
- Similar to FTS3 but not using grid protocols and certificates.
- Share same APIs in Rucio transfer plugins and could be easily replaced by others.
- Under developing.

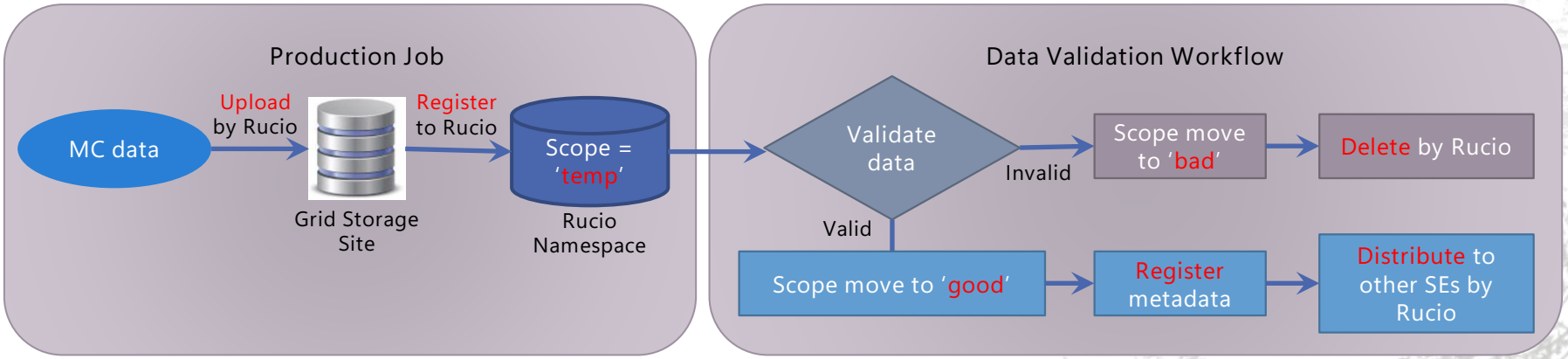


Application: MC Data Flow for HERD

A user-oriented API is developed for HERD experiment:

- Provide job environment for user and production jobs to management data in Rucio.
- HERD MC data flow.
  - Rucio Scope is defined as data status in data flow: 'Temp', 'good', 'bad'.

1. Production jobs generate data, upload it to SE by Rucio.
2. Data is registered as scope of 'temp' for further validation.
3. Data validation workflow uses APIs to validate whether data are good.
4. If good, move scope to 'good', then provide it to metadata registering and distribute data to other storage by Rucio.
5. If not good, move scope to 'bad' scope, waiting for deleting by Rucio.



DIRAC-DMS Based Solution

DIRAC-DMS:

- Present Data Management System:
- DIRAC Data File Catalog (DIRAC-DFC),
- Supports BESIII, JUNO, CEPC.

DIRAC-DFC:

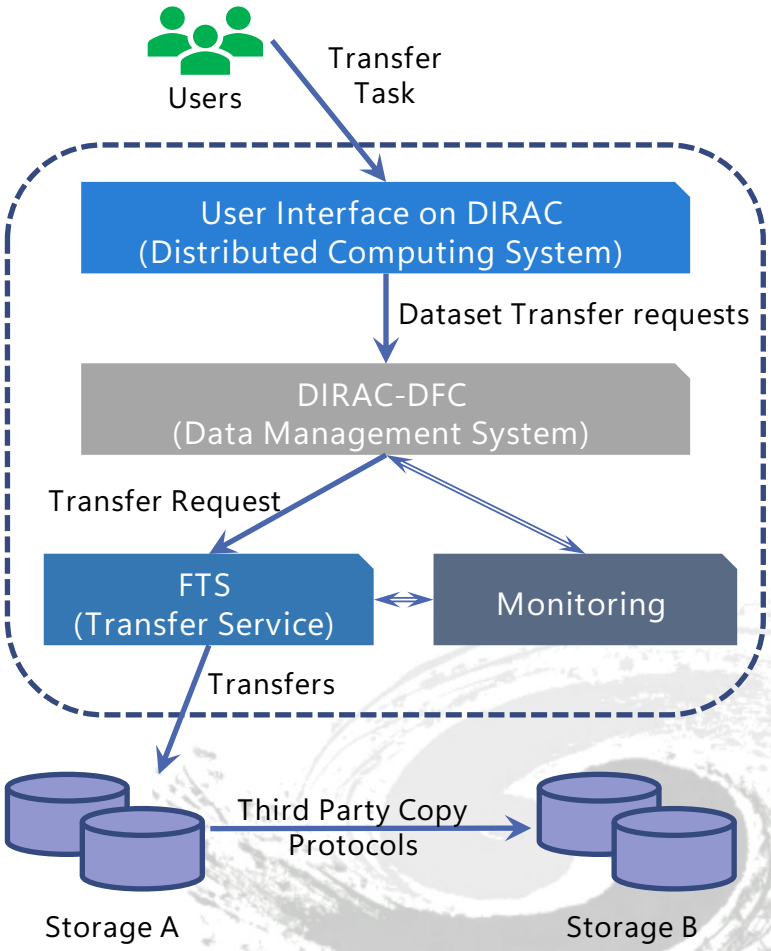
- Provides global data view,
- Supports dataset management,
- Manages transfers based on datasets and metadata.

Data Transfer Service, FTS:

- Manages large-scale data transfer,
- Robust, mature, popular.

Third-party-copy Protocols:

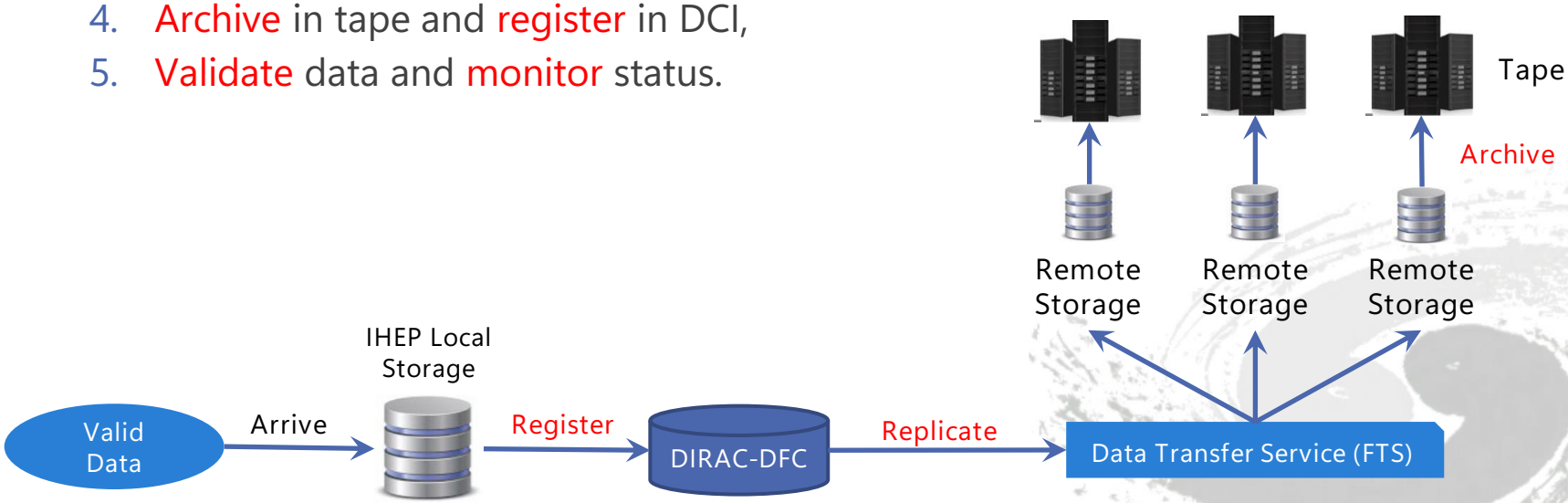
- Xrootd, WebDAV (X509 certificate).



Application: MC & Raw Data Transfer in JUNO

Supports production & raw data distribute and archive.

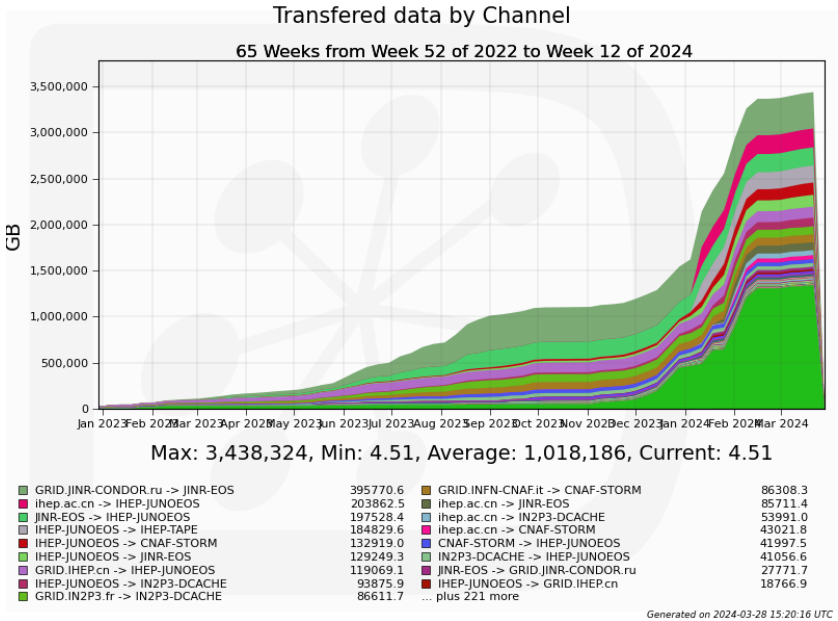
1. Receive data process trigger when data arriving at IHEP local storage,
   - Trigger could be a message queue, a new database record, an active file probing, etc., mainly based on experiment design.
2. Register data from IHEP local storage to DIRAC-DFC,
   - Data validation is ahead of registering to DIRAC-DFC.
3. Replicate data from IHEP to cooperated data centers disk and then register,
   - By FTS3 with TPC transfer protocol.
4. Archive in tape and register in DCI,
5. Validate data and monitor status.

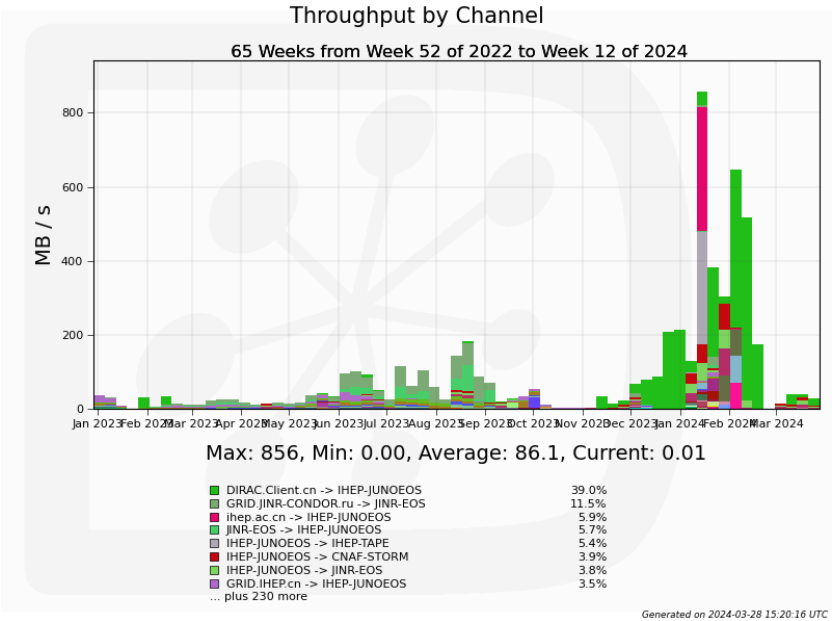


General Usage

For our system in 2023 & 2024:

- DIRAC-DFC is in mass production, and Rucio is in pre-production study.
- A data challenge for JUNO data distribution between China and Europe has been done while WLCG DC24.
- Total 3.44 PB data transferred.
- Total 0.77 PB and 11.9 Million files registered and managed.





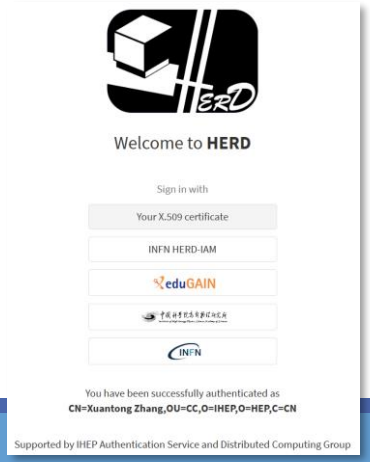
Grid Middleware & Storages

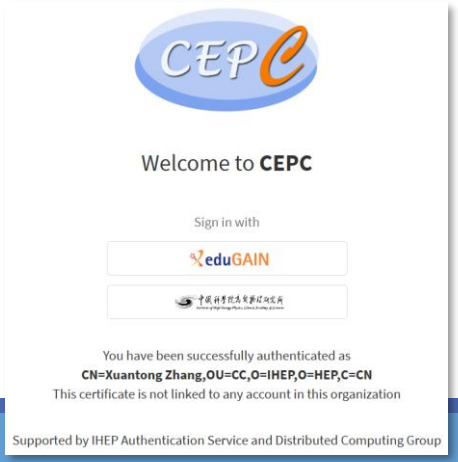
IAM

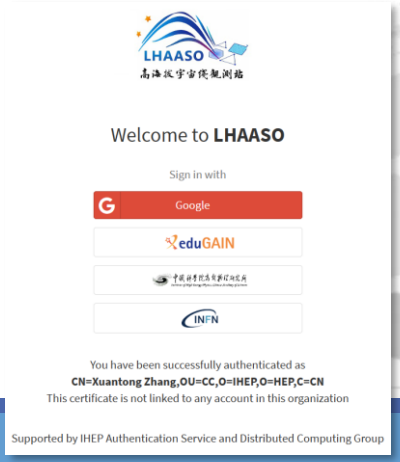
IHEP started to supply IAM services for local experiments since 2022.

There are 3 instances at IHEP, all in v1.8.3:

- HERD-IAM, since 2022:
  - Support Sci-token and VOMS-proxy for IHEP's grid resources authorization.
  - Support eduGain, INFN-SSO and IHEP-SSO for authentication.
- CEPC-IAM, since 2023:
  - Successor for CEPC-VOMS, now working together with CEPC-VOMS with VOMS-importer.
  - Support VOMS-proxy for authorization, support eduGain and IHEP-SSO for authentication.
- LHAASO-IAM, since 2024:
  - Testbed for explore further usages in LHAASO experiments.
  - Plan to use pure Sci-token for authorization.







Storages

StoRM, EOS, and EOS-CTA are the main storage systems for Grid data at IHEP.

- All support Xrootd and WebDAV TPC protocol with Sci-token and VOMS-proxy.

StoRM:

- Use Lustre as backend system, share same Lustre with local storage.
- Serve JUNO since 2018 and HERD since 2022, with ~450 TB data managed.

EOS:

- Shared with local storage, the main data storage.
- Serve JUNO since 2020, with 4 PB space, 2.26 PB used, 0.77 PB managed by Grid.

EOS-CTA:

- Tape storage, firstly used on Grid for JUNO since 2023.
- With 0.51 PB used and managed by Grid.

Supporting Services & Activities

JUNO Network Data Challenge

JUNO data transfer background:

- 3 PB/year (~102.1 MB/s) data distribution from IHEP to [CNAF, IN2P3, JINR].
- Network upgrade: IHEP to GEANT upgraded from 10 Gbps to 100 Gbps on 22nd Sept 2023.
- WLCG started DC24 in Feb 2024, would occupy lots of network bandwidth.

Goals for challenges:

1. To learn network and Grid infrastructures limit capability.
2. To optimize network and data transfer performance.
3. To develop transfer status monitoring and visualization.

What was done in challenges,

1. Mimic data transfer challenges with Grid tools (FTS3 + VOMS).
2. Challenge tools development and transfer results visualization.
3. Network and Grid infrastructure optimization before and after challenges.

Pre-Challenge

We had 11 times pre-transfers before we started Data challenge.

Experiences and knowledges we got from pre-challenge:

1. IHEP<->JINR transfer is still not good, 3 PB/y (~102 MB/s) throughput still get stuck and high failure rate.
2. Mimic data transfer between IHEP and [CNAF, JINR] with 12 PB/y (~512 MB/s) throughput is still good enough.
3. The storage I/O could possibly a bottle neck in mimic data transfer.
4. Size per file of mimic data estimation to 2-5 GB.
5. Transfer retry by FTS3 is helpful for data transfer failure.
6. FTS3 service max active streams has limitation.

Challenge 2024

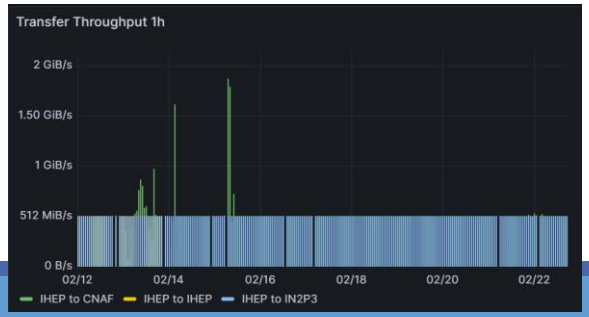
JUNO and WLCG DC24 were carried out on same period since 12th Feb 2024.

Transfer design:

- Only IHEP->[CNAF, IN2P3] directions were challenged.
- Stress challenge by 12 PB/y (~512 MB/s) mimic data throughput (4x of real data throughput).
- 3 time retries for failure transfer file.

Results:

- IHEP->CNAF met some connection failure and triggered re-transfer.
  - Known issue, CNAF StoRM met some running instability during LHC DC24, JUNO got impacted. Came back to normal since 16th Feb.
- IHEP->IN2P3 transfer worked well.
- 12 PB/y (~512 MB/s) is OK and a 24 PB/y (1 GB/s) nearly reach to limit.



Impacted by StoRM instability during LHC DC24

Monitoring Platform

At IHEP, Grid data management is lack in monitoring.

Metrics need to be monitored:

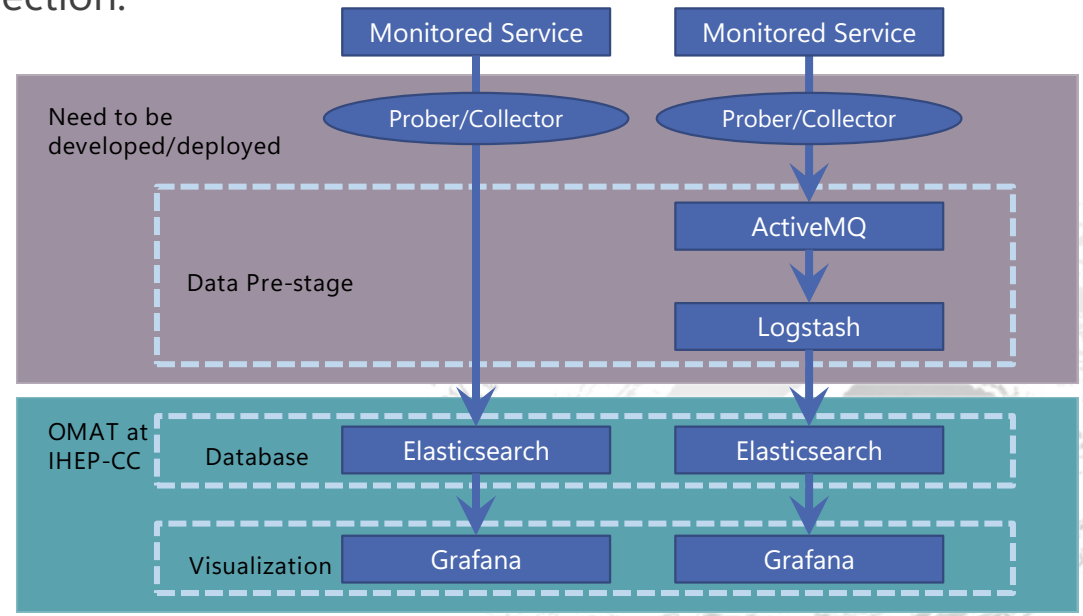
- FTS3 transfer status.
- Site Availability Monitor (SAM) test for all SEs.
- SEs' used and pledged space accounting.

Techs and systems:

- Developed probes for logs collection.
- ELK package + Grafana.

Started since Oct 2023.

- Still in developing.
- Now ready:
  - FTS3 transfer status,
  - TPC Monitoring,
  - Part of sites storage accounting.



Summary

Summary

1. At IHEP, Grid Data management based on solutions of Rucio and DIRAC-DMS. For both of them, we developed experiment-oriented plugins and extensions.
2. IAM and TPC protocols in storages are in more and more use for Grid data management at IHEP.
3. We had a data challenge for JUNO at the same time of WLCG DC24. The upgraded 100 Gbps network showed good quality in challenges.
4. A monitoring platform is in developing for Grid data management at IHEP.

Thank you!