# Distributed Data Management System at IHEP

Xuantong Zhang, Xiaomei Zhang

IHEP, CAS

# Outline

## Introduction

- CC-IHEP, CAS and its serving experiments
- Overview of IHEP distributed computing system

## Status of IHEP distributed data system

- DIRAC Data Management System (BESIII, JUNO, CEPC)
- Rucio (JUNO, HERD)

## Infrastructures for IHEP distributed data

- Storage: StoRM, EOS, EOSCTA (JUNO, HERD, CEPC)
- Grid Midware: IAM and token-based TPC
- IHEP developed components: TPC monitoring, etc.

## Summary

# Introduction

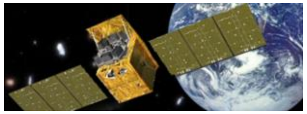
## CC-IHEP, CAS:

- The **first and largest** Grid Site in Mainland China,
  - Serving both Chinese located and WLCG experiments.
  - >50 K CPU cores, 210 GPUs,
  - >75 PB disk storage, >50 tape storage.
- Distributed Data Management system of IHEP serves:
  - BESIII,
    - A running spectrometer at BEPCII.
  - JUNO,
    - A neutrino observatory located in the south of China,
    - 2.4 PB raw data, 0.6 PB simulation data per year.
  - HERD,
    - A high energy cosmic detector on China Space Station,
    - ~1 PB data per year.
  - CEPC,
    - A planning circular electron positron collider.

## Chinese located or IHEP driven experiments

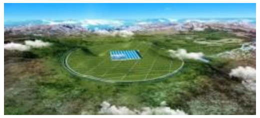
BESIII (Beijing Spectrometer III at BEPCII)

JUNO (Jiangmeng Underground Neutrino Observatory)

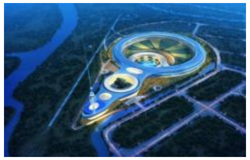


HXMT (Hard X-Ray Moderate Telescope)

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)

## International collaborated experiments

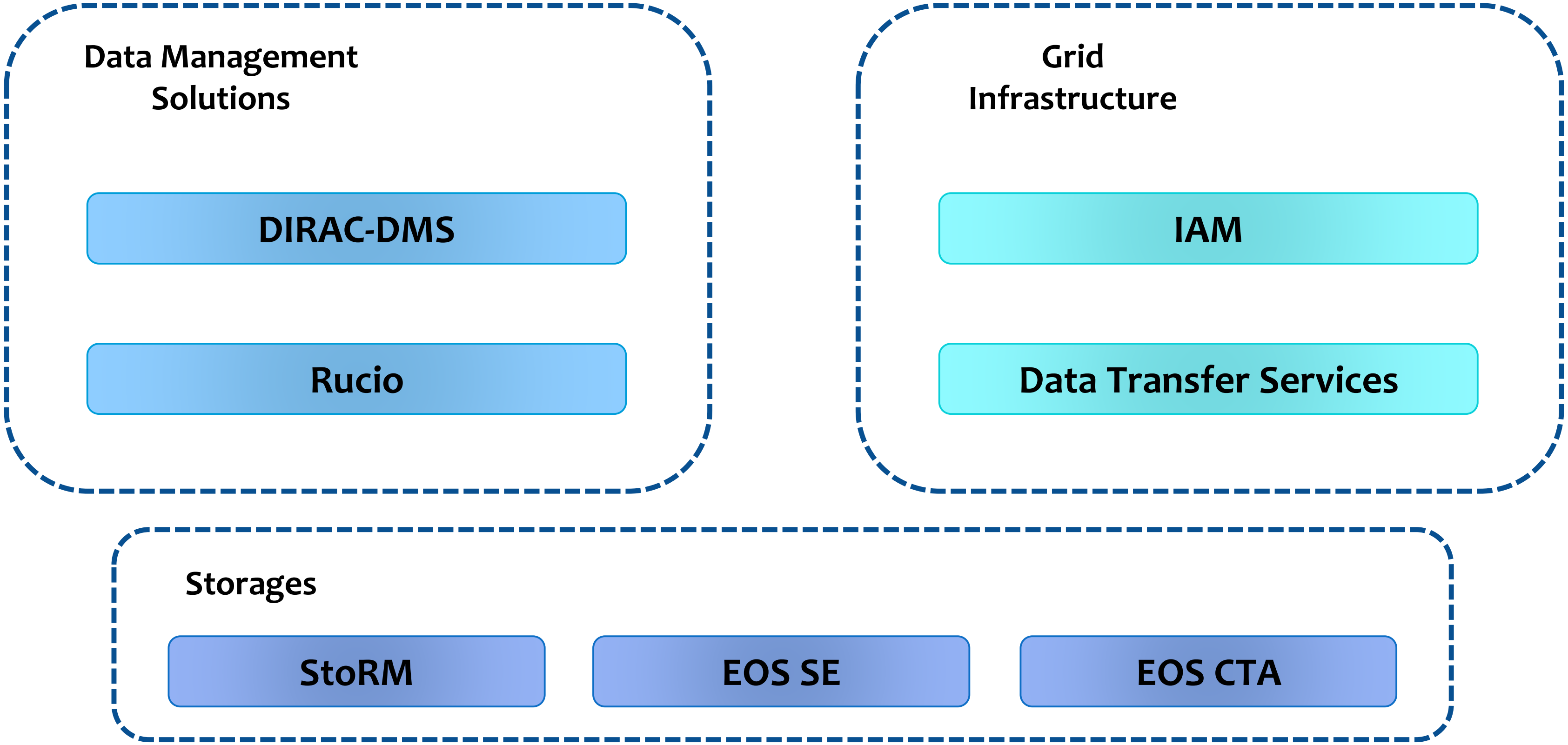
CMS

ATLAS EXPERIMENT

LHCb

Belle II

# System Overview



Data Management Solutions

DIRAC-DMS

Rucio

Grid Infrastructure

IAM

Data Transfer Services

Storages

StoRM

EOS SE

EOS CTA

# DIRAC-DMS Based Solution

## DIRAC-DMS:

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

## DIRAC-DFC:

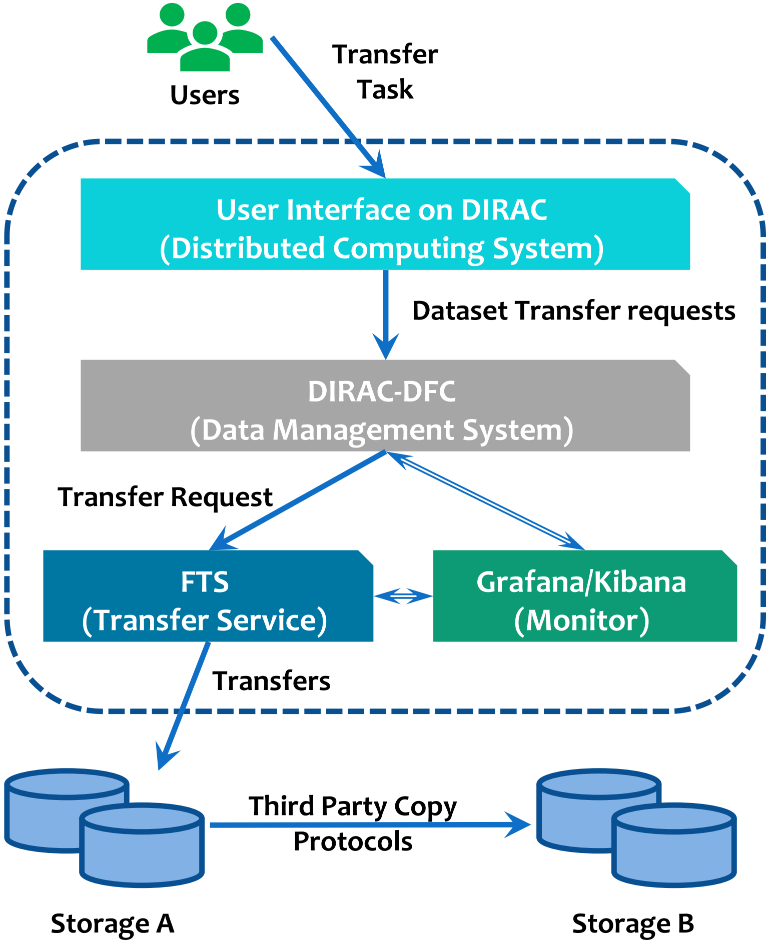
- Provides **global data view**,
- Supports dataset management,
- Manages transfers based on **datasets and metadata**.

## Data Transfer Service, FTS:

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

## Third-party-copy Protocols:

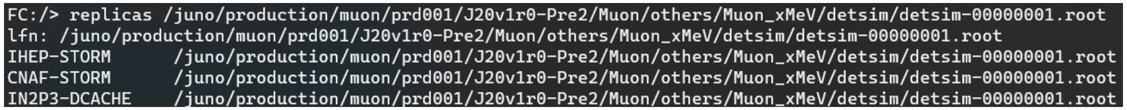
- **Xrootd, Webdav.**



# Data Management

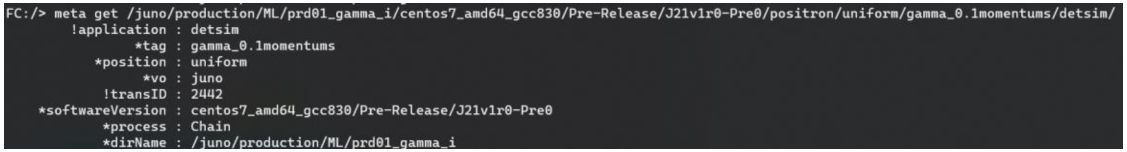
## Features of DIRAC-DMS:

- Linux-like directories.
- Logic and physical view of global data.



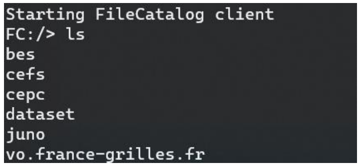
```
FC:/> replicas /juno/production/muon/prd001/J20v1r0-Pre2/Muon/others/Muon_xMeV/detsim/detsim-00000001.root
lfn: /juno/production/muon/prd001/J20v1r0-Pre2/Muon/others/Muon_xMeV/detsim/detsim-00000001.root
IHEP-STORM      /juno/production/muon/prd001/J20v1r0-Pre2/Muon/others/Muon_xMeV/detsim/detsim-00000001.root
CNAF-STORM      /juno/production/muon/prd001/J20v1r0-Pre2/Muon/others/Muon_xMeV/detsim/detsim-00000001.root
IN2P3-DCACHE    /juno/production/muon/prd001/J20v1r0-Pre2/Muon/others/Muon_xMeV/detsim/detsim-00000001.root
```

- Datasets, directories, files can set metadata.
- Data can be queried by metadata.

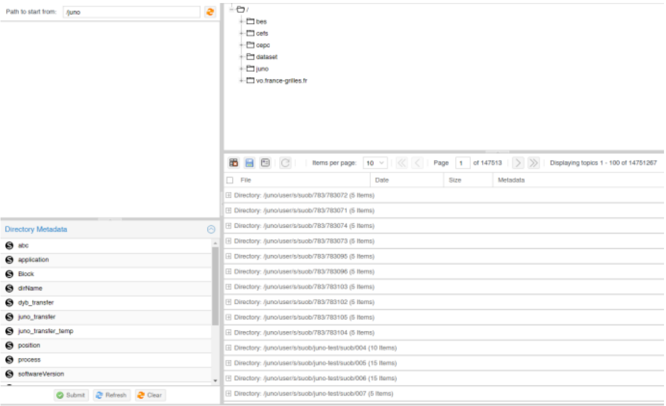


```
FC:/> meta get /juno/production/ML/prd01_gamma_i/centos7_amd64_gcc830/Pre-Release/J21v1r0-Pre0/positron/uniform/gamma_0.1momentums/detsim/
        !application : detsim
                *tag : gamma_0.1momentums
           *position : uniform
                 *vo : juno
            !transID : 2442
    *softwareVersion : centos7_amd64_gcc830/Pre-Release/J21v1r0-Pre0
            *process : Chain
            *dirName : /juno/production/ML/prd01_gamma_i
```

- Command line and Web UI is supported.



```
Starting FileCatalog client
FC:/> ls
bes
cefs
cepc
dataset
juno
vo.france-grilles.fr
```



Path to start from: /juno

Items per page: 10 | Page 1 of 147513 | Displaying topics 1 - 100 of 14751267

| File | Date | Size | Metadata |
|---|---|---|---|
| Directory: /juno/user/s/suob/783/783072 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783071 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783074 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783073 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783095 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783096 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783103 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783102 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783105 (5 Items) | | | |
| Directory: /juno/user/s/suob/783/783104 (5 Items) | | | |
| Directory: /juno/user/s/suob/juno-test/suob/004 (10 Items) | | | |
| Directory: /juno/user/s/suob/juno-test/suob/005 (15 Items) | | | |
| Directory: /juno/user/s/suob/juno-test/suob/006 (15 Items) | | | |
| Directory: /juno/user/s/suob/juno-test/suob/007 (5 Items) | | | |

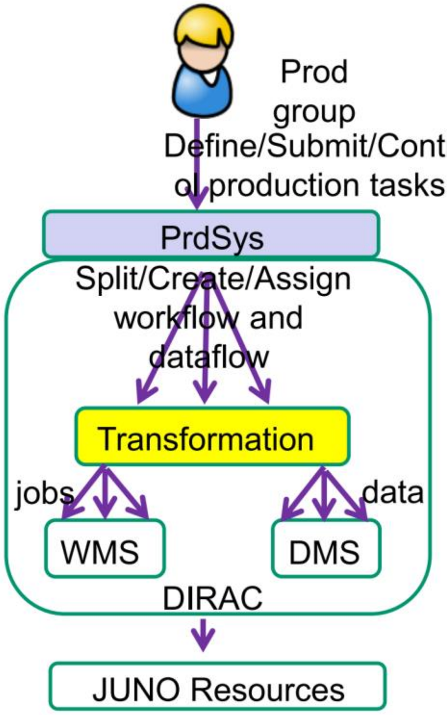
# MC Data Production

## We developed ProdSys for experiments at IHEP:

- For massive MC production tasks.
- Automatically create and manage workflow.
- Multi experiments production software merged.

## DIRAC-DMS in ProdSys:

- Fully integrated for managing data flows.
  - Input data management,
  - Produced data register and transfer,
  - Multi-sites data replication.



Prod group Define/Submit/Control production tasks

PrdSys

Split/Create/Assign workflow and dataflow

Transformation

jobs

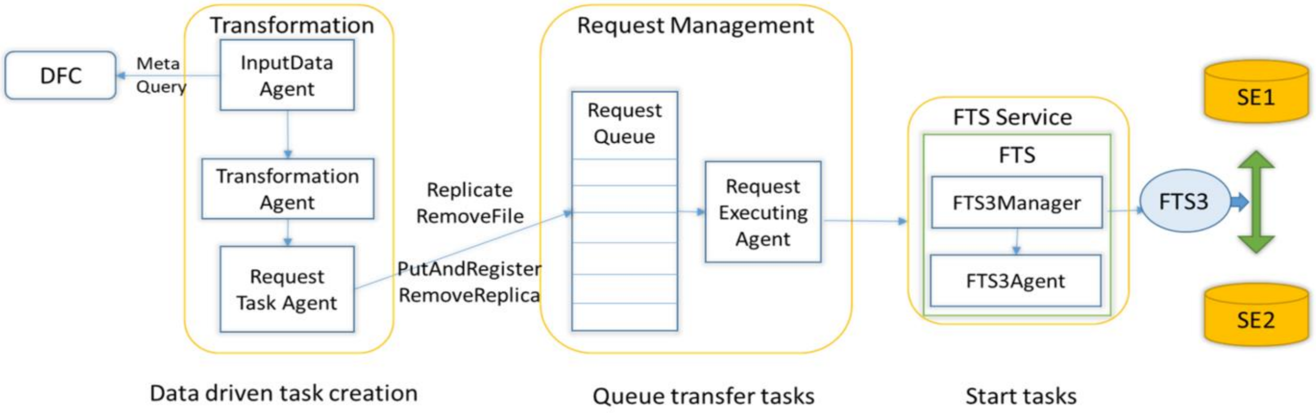
data

WMS

DMS

DIRAC

JUNO Resources



DFC

Meta Query

Transformation

InputData Agent

Transformation Agent

Request Task Agent

Replicate RemoveFile

PutAndRegister RemoveReplica

Request Management

Request Queue

Request Executing Agent

FTS Service

FTS

FTS3Manager

FTS3Agent

FTS3

SE1

SE2

Data driven task creation

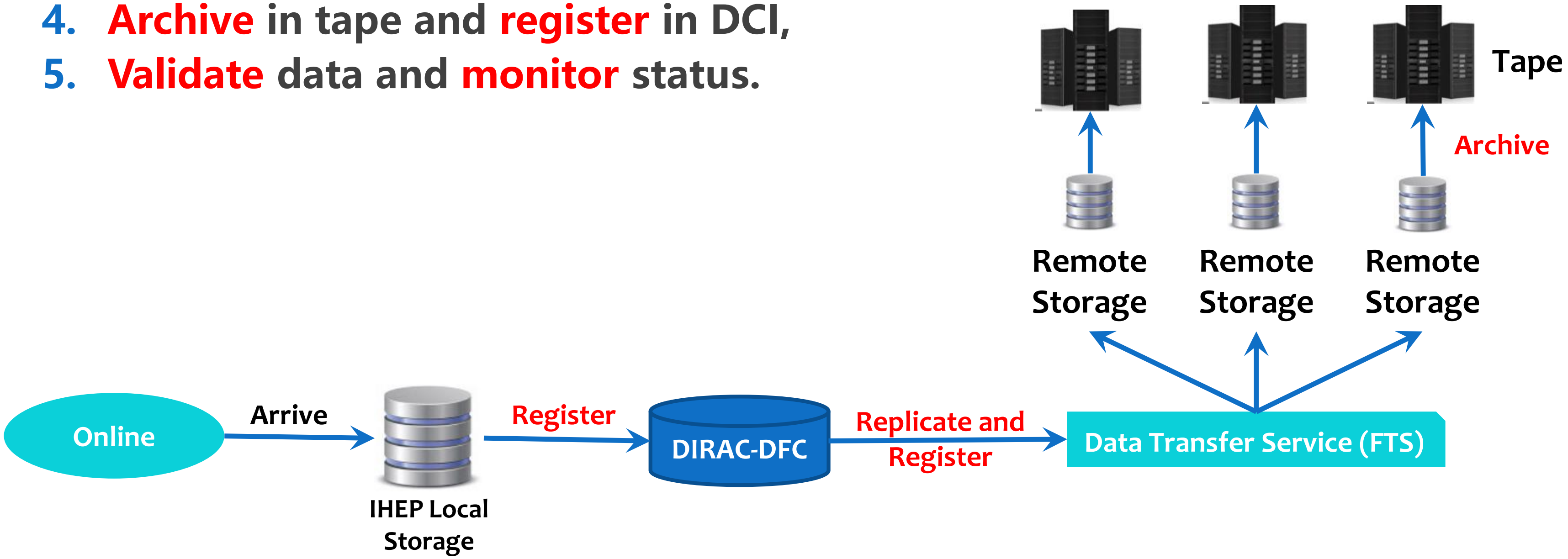
Queue transfer tasks

Start tasks

# Raw Data Flow

## Supports raw data transfer and archive.

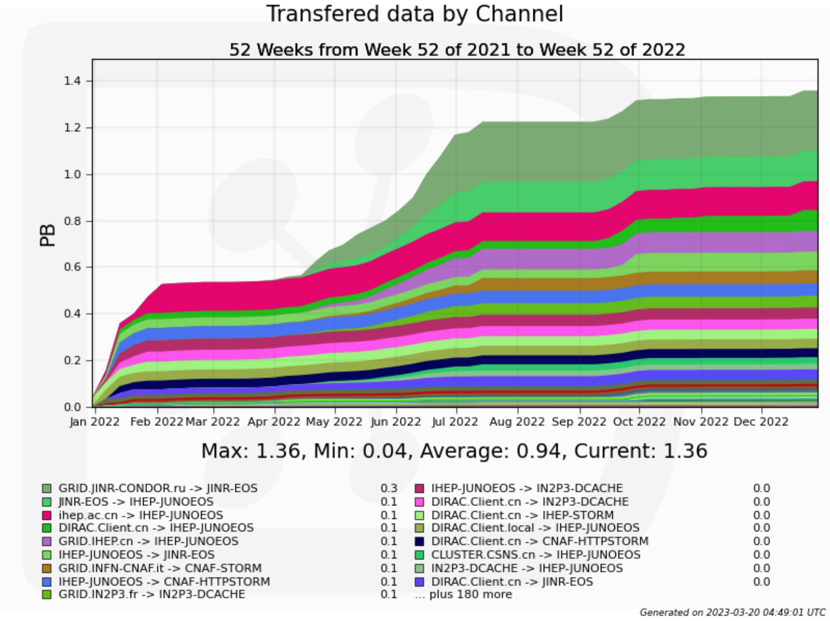
1. **Receive** data process trigger when data arriving at IHEP local storage,
   - Trigger could be a message queue, a new database record, an active file probing, etc., mainly based on experiment design.
2. **Register** data from IHEP local storage to DIRAC-DFC,
3. **Replicate** data from IHEP to cooperated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

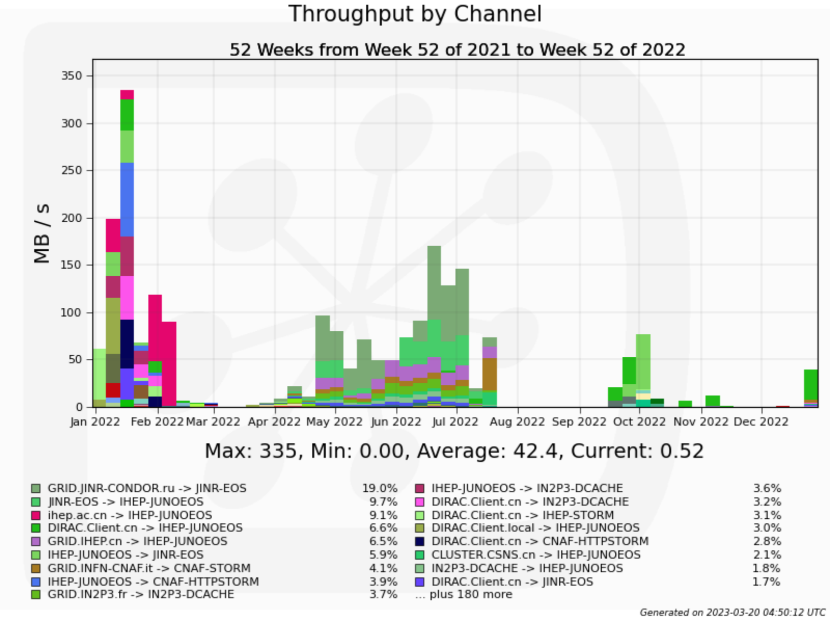


# General Usage in 2022

## For our system in 2022:

- Large transfers has been done with high quality and good speed.
- Total 1.4 PB data transferred by DIRAC-DMS.
- Total 1 PB and 4 Million files registered and managed by DIRAC-DFC.





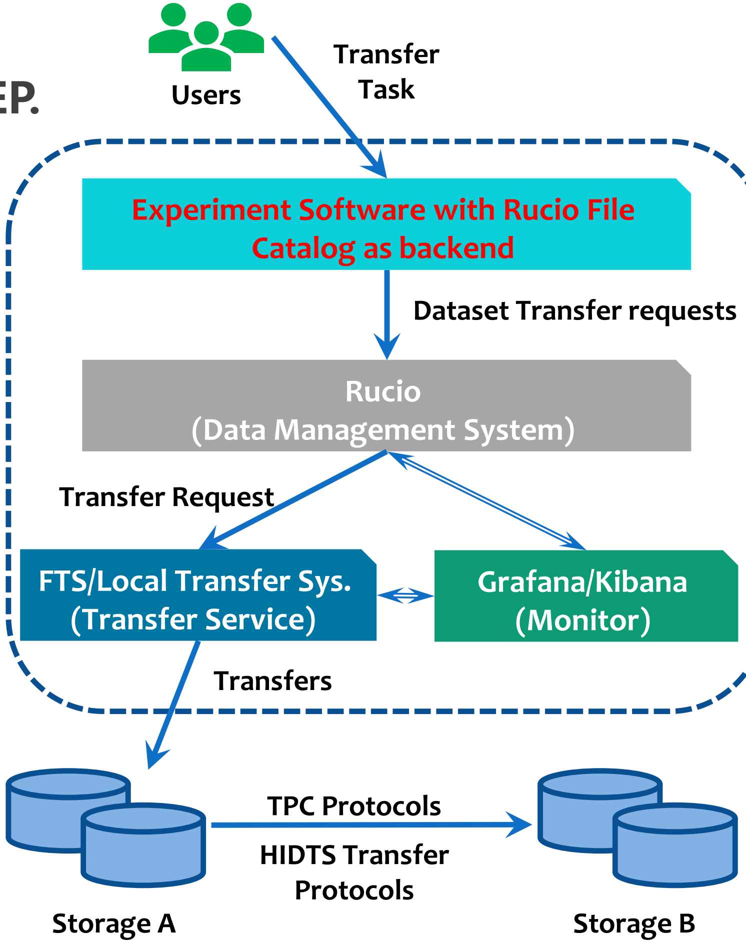
# Rucio Based Solution

## Rucio Data Management system:

- Supports HERD experiment,
- Preparing for future experiments of IHEP.

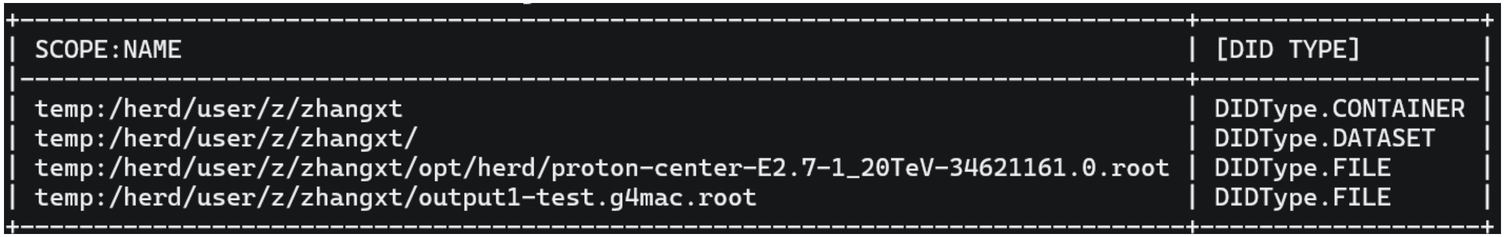
## We developed Rucio solution under our experiments needs:

- 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). (see page 12.)


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

# Features of Rucio at IHEP

## Rucio DID customization, to make data logic name closer to local data. (see page 15.)



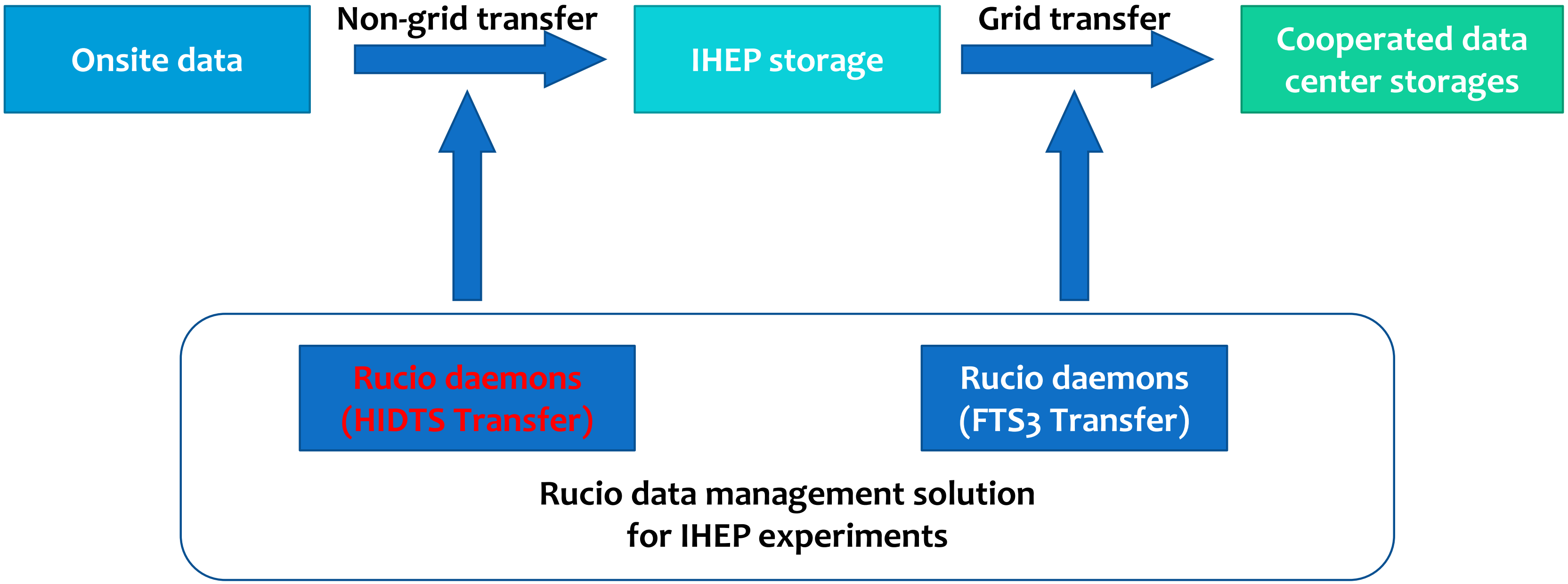
```
+------------------------------------------------------------------------------+--------------------+
| SCOPE:NAME                                                                   | [DID TYPE]         |
|------------------------------------------------------------------------------+--------------------|
| 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 |

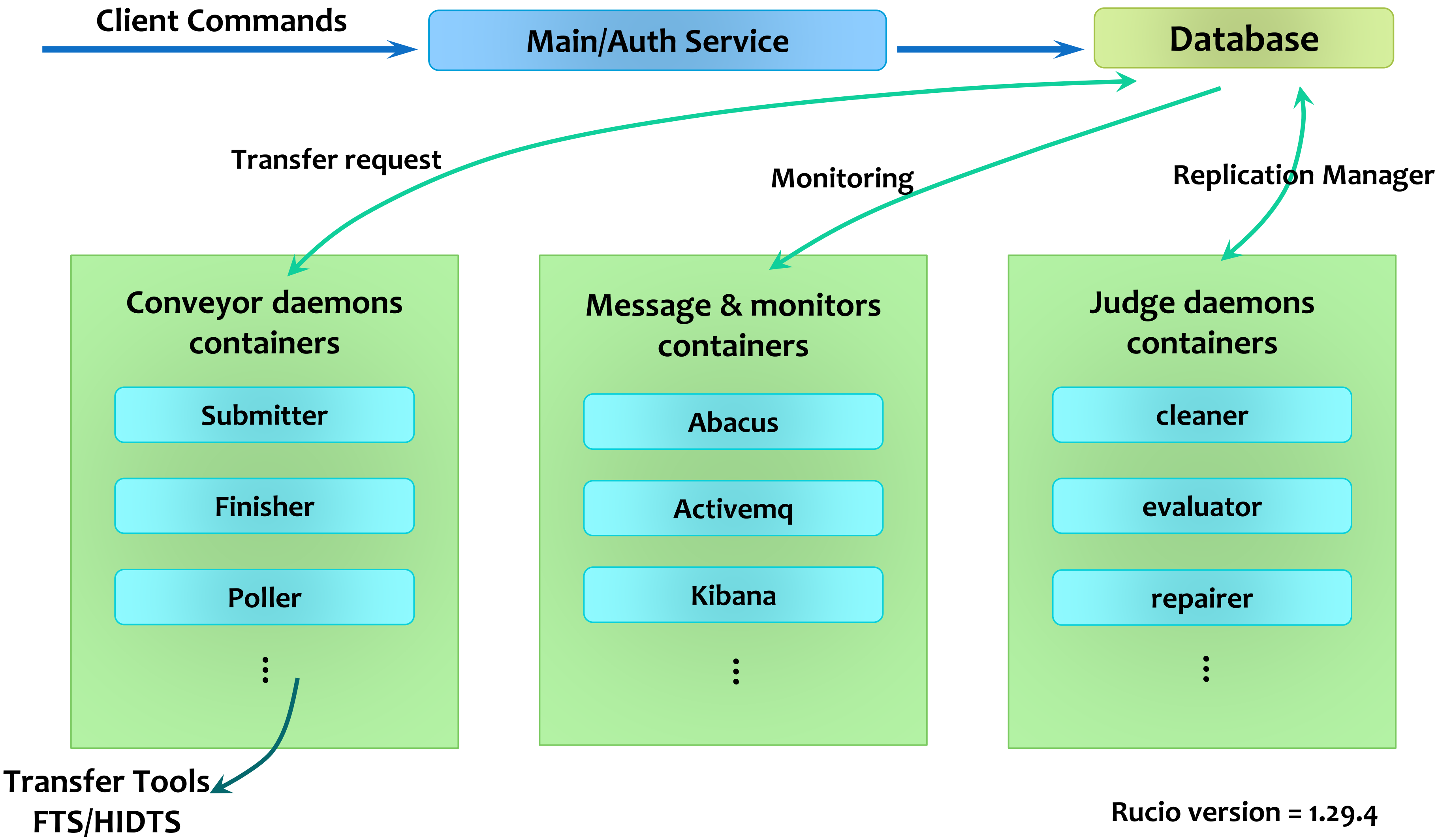
# Features of Rucio at IHEP

## IHEP HIDTS Plugins for rucio daemons of data transfers:

- IHEP HIDTS is a non-grid data transfer services for IHEP storage site.
- Similar to FTS3 but not using grid protocols and certificates.
- To manage pre-transfer between experiment location to IHEP grid storage elements.
- Still in developing.


Onsite data
Non-grid transfer
IHEP storage
Grid transfer
Cooperated data center storages
Rucio daemons (HIDTS Transfer)
Rucio daemons (FTS3 Transfer)
Rucio data management solution for IHEP experiments

# Rucio Services Structures at IHEP

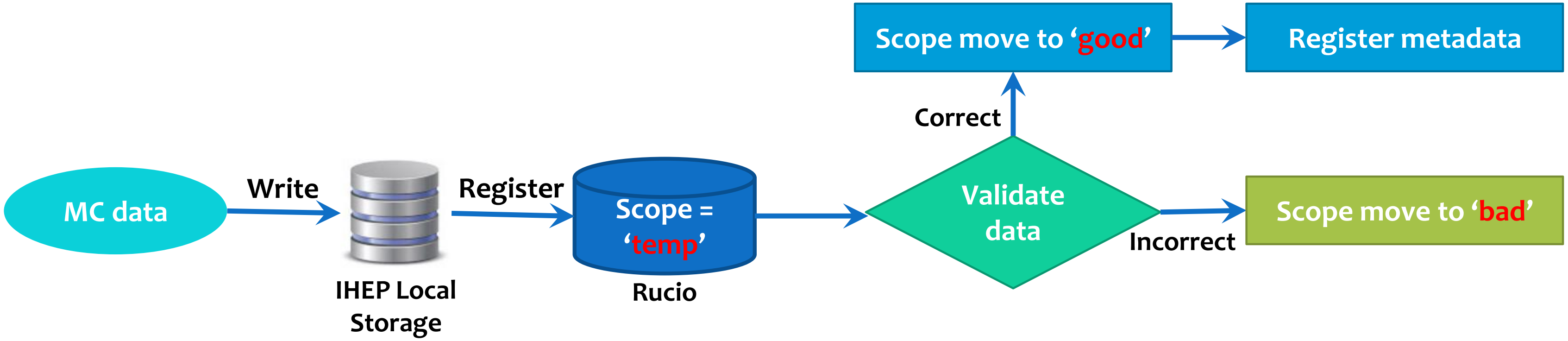


Rucio version = 1.29.4

# MC Data Flow for HERD

## A user-oriented API is developed for HERD experiment:

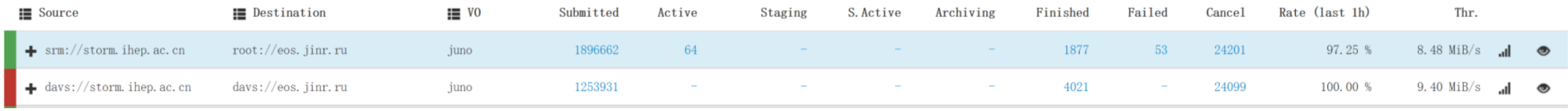
- **Rucio Scope is defined as data status in data flow.**
  - ‘Temp’ , ‘good’ , ‘bad’ .
- **Example: MC data flow.**
  1. Register all raw MC data to ‘temp’ scope,
  2. Data validation program use APIs to validated whether data are good.
  3. If good, move scope to ‘good’ , then provide it to metadata registering.
  4. If not good, move scope to ‘bad’ scope, waiting for deletion.



# Transfer Operation in 2022

## JUNO data transfer mission:

- IHEP StoRM -> JINR EOS, ~70 TB, ~10 Million files,
- Suffer from small size file, max speed ~20 MB/s,
- File register speed enhanced to ~90,000 files/s,
- Rucio worked smooth and stable, after the well tuning of data policy and configuration,

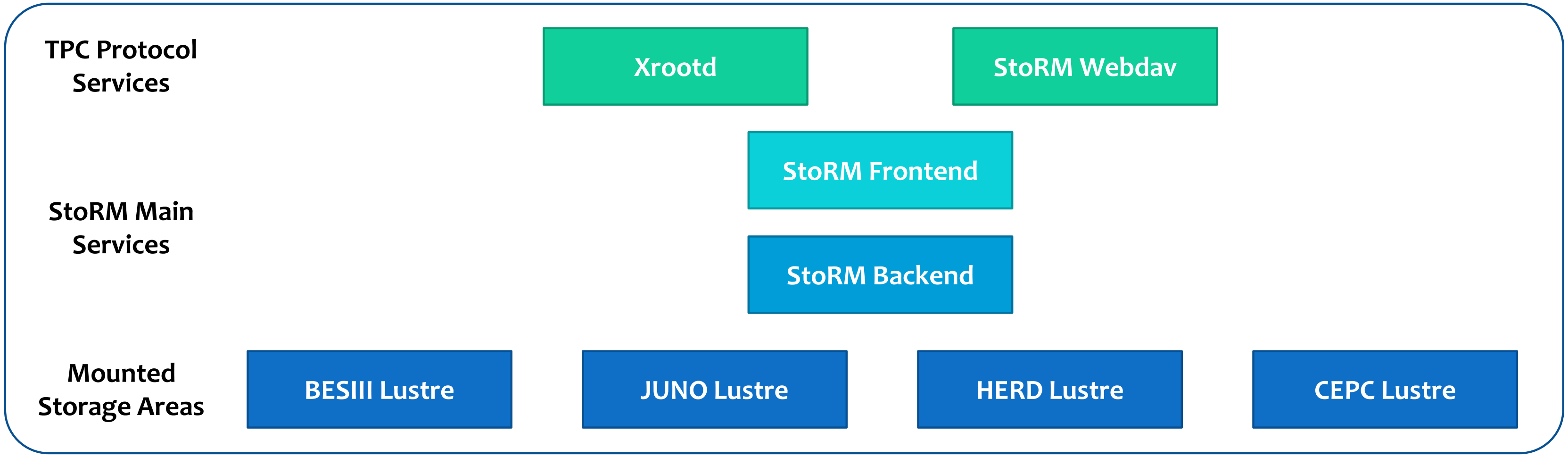


| Source | Destination | VO | Submitted | Active | Staging | S.Active | Archiving | Finished | Failed | Cancel | Rate (last 1h) | Thr. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| srm://storm.ihep.ac.cn | root://eos.jinr.ru | juno | 1896662 | 64 | - | - | - | 1877 | 53 | 24201 | 97.25 % | 8.48 MiB/s |
| davs://storm.ihep.ac.cn | davs://eos.jinr.ru | juno | 1253931 | - | - | - | - | 4021 | - | 24099 | 100.00 % | 9.40 MiB/s |

# StoRM

## StoRM system at IHEP serves BESIII, JUNO, HERD, CEPC.

- Mounted Lustre file system in backend.
- Supported TPC protocols: Xrootd, Webdav.
- Supported authorization, IAM-token and VOMS credential,
- Fine-grained authorization by WLCG JWT scopes and VOMS roles.
- Version = 1.11.21



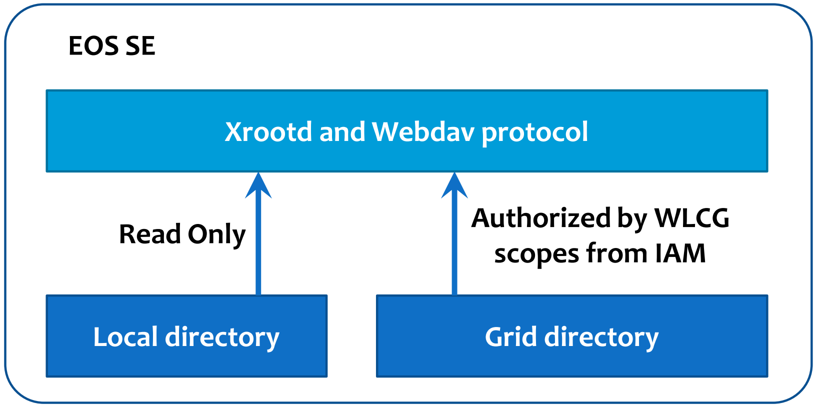
# EOS SE and EOS-CTA

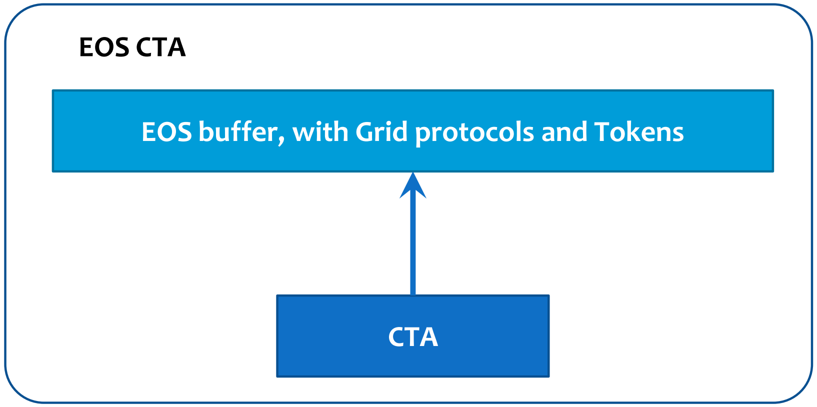
## EOS SE at IHEP serves JUNO, HERD.

- Only one EOS instance for one experiment.
- Grid and local directory is separated.
- Grid users is mapped by WLCG scopes from IAM.

## EOS-CTA at IHEP serves JUNO.

- CTA is the tape system behind EOS buffer.
- Also support Grid Protocols with WLCG scopes in EOS buffer.
- Ready to serve JUNO.


EOS SE
Xrootd and Webdav protocol
Read Only
Authorized by WLCG scopes from IAM
Local directory
Grid directory



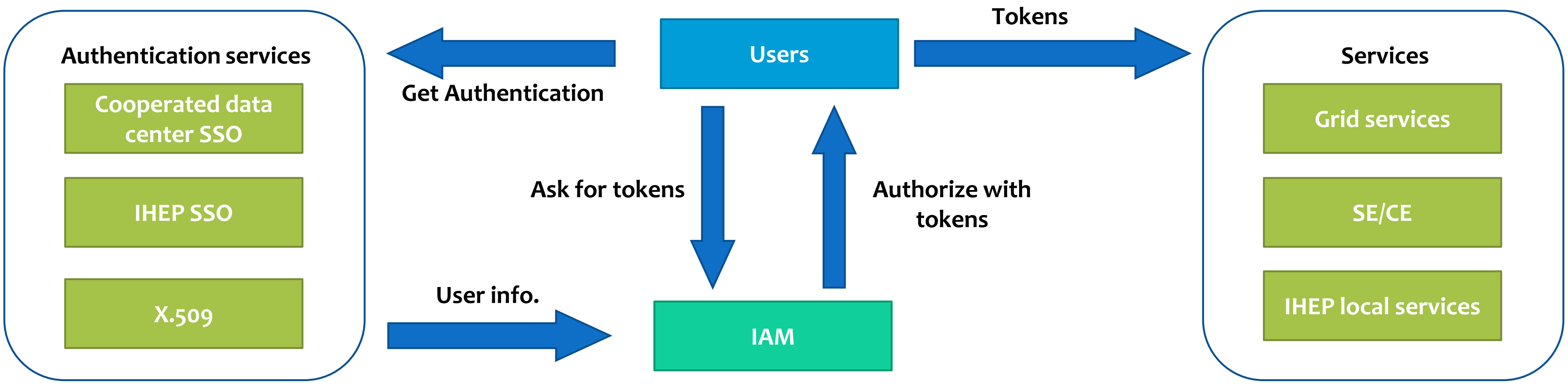
# IAM Service

## IAM at IHEP serves HERD, CEPC.

- Multi-authentication: IHEP SSO, X.509, third party IdPs.
- Multi-authorization: VOMS credential, Sci-tokens.

## Supported local services at IHEP.

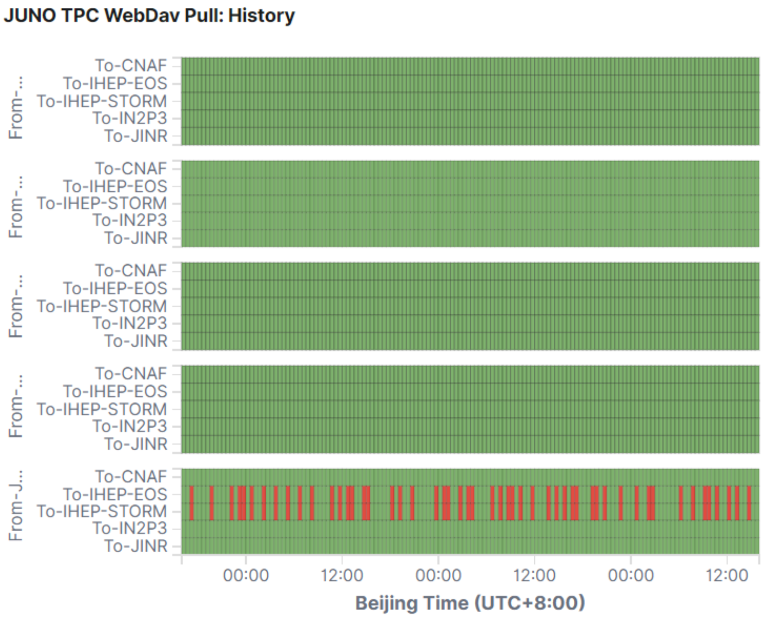
- Some local storages and computing resources at IHEP.
- Grid services: Rucio, FTS3.
- Scientific services at IHEP: Indico, etc.

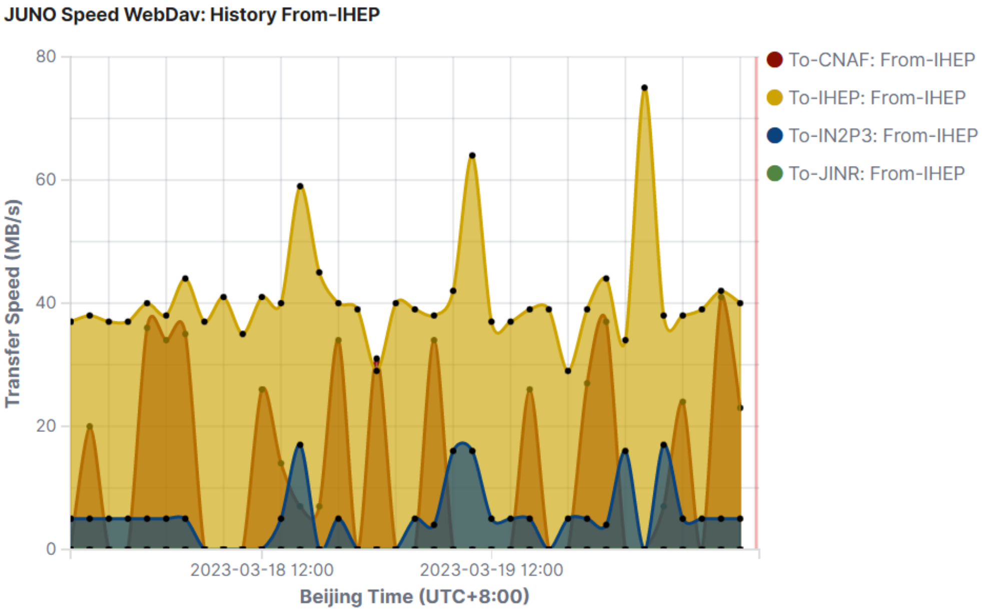


# TPC Active Probing System

## Active Probing system for JUNO & HERD Third Party Copy (TPC) function and speed.

- Tests executed by Gfal2 tools, results collected and shown in Elasticsearch-Kibana.
- 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.





# Summary

## Distributed data management system at IHEP,

- To support international experiments located in China or driven by IHEP, including BESIII, JUNO, HERD, CEPC.
- At present, DIRAC-DMS is in production and proved to be good.
- Rucio for HERD, JUNO, CEPC with deep customization is in development.

## Developments for more Grid service based on experiments.

- IAM with sci-tokens for storage and IHEP services.
- TPC active probing system for experiments.

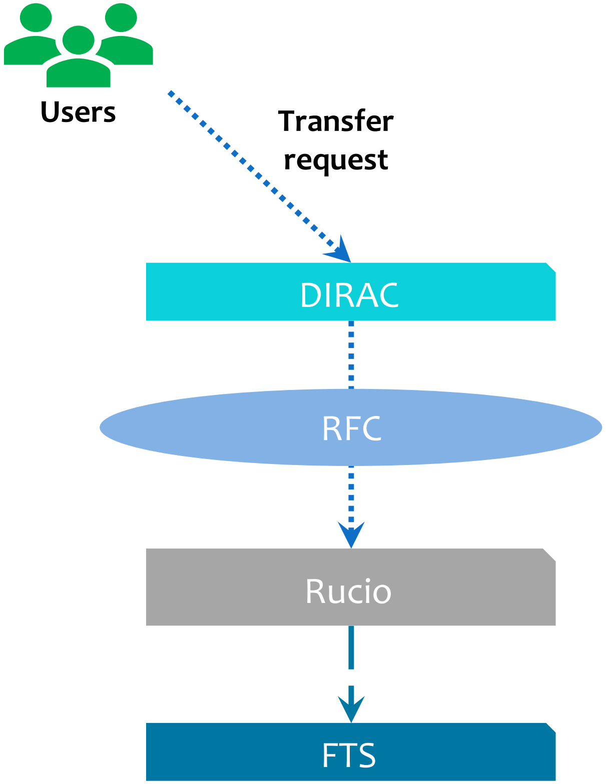
Thank you!

# backup

# Backup DMS for JUNO

## Works as the backup DMS for JUNO experiment.

- Based on Rucio File Catalog (RFC) Component in DIRAC.
- Compatible with the present DIRAC-DFC data namespace policy.


Users
Transfer request
DIRAC
RFC
Rucio
FTS